

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

Genetic and Bioinformatic Approaches To Identify Polymorphic Modulators of Transcription Factor Binding and Disease Phenotypes Including HIV-1 Viremia

Permalink

<https://escholarship.org/uc/item/4v34t8vt>

Author

Williamson, David Wayne

Publication Date

2008-04-02

Peer reviewed|Thesis/dissertation

Genetic and Bioinformatic Approaches To Identify Polymorphic Modulators of
Transcription Factor Binding and Disease Phenotypes Including HIV-1 Viremia

by

David Wayne Williamson

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Graduate Program in Biological and Medical Informatics (BMI)

Integrative Program in Quantitative Biology (iPQB)

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Copyright 2008

by

David Wayne Williamson

Dedication and Acknowledgments

1.1 Publication Reprints

The text of this dissertation/thesis contains a reprint of material as it appears in:

Hunt PW, Harrigan PR, Huang W, Bates M, Williamson DW, McCune JM, Price RW, Spudich SS, Lampiris H, Hoh R, Leigler T, Martin JN, Deeks SG.

Prevalence of CXCR4 tropism among antiretroviral-treated HIV-1-infected patients with detectable viremia.

J Infect Dis. 2006 Oct 1;194(7):926-30. Epub 2006 Aug 29. PMID: 16960780

The co-author listed in this publication participated by collecting the primary CCR5 genotypes (page 345)

The text of this dissertation/thesis contains a reprint of material as it appears in:

Hodoglugil U, Tanyolaç S, Williamson DW, Huang Y, Mahley RW.

Apolipoprotein A-V: a potential modulator of plasma triglyceride levels in Turks.

J Lipid Res. 2006 Jan;47(1):144-53. Epub 2005 Oct 28. PMID: 16258166

The co-author listed in this publication participated by collecting the primary genotypes, calculating the association statistics, and co-writing the manuscript (page 350).

The text of this dissertation/thesis contains a reprint of material as it appears in:

Hodoğlugil U, Williamson DW, Huang Y, Mahley RW.

An interaction between the TaqIB polymorphism of cholesterol ester transfer protein and smoking is associated with changes in plasma high-density lipoprotein cholesterol levels in Turks.

Clin Genet. 2005 Aug;68(2):118-27. PMID: 15996208

The co-author listed in this publication participated by collecting the primary genotypes, calculating the association statistics, and co-writing the manuscript (page 360).

The text of this dissertation/thesis contains a reprint of material as it appears in:

Hodoğlugil U, Williamson DW, Huang Y, Mahley RW.

Common polymorphisms of ATP binding cassette transporter A1, including a functional promoter polymorphism, associated with plasma high density lipoprotein cholesterol levels in Turks.

Atherosclerosis. 2005 Dec;183(2):199-212. Epub 2005 Jun 2. PMID: 15935359

The co-author listed in this publication participated by collecting the primary genotypes, calculating the association statistics, and co-writing the manuscript (page 370).

The text of this dissertation/thesis contains a reprint of material as it appears in:

International Congress Series

Volume 1262 , May 2004, Pages 193-199

Atherosclerosis XIII. Proceedings of the 13th International Atherosclerosis Symposium

Low HDL-C: lessons learned from the Turkish Heart Study

U. Hodolugil, D. Williamson and R. W. Mahley

The co-author listed in this publication participated by collecting the primary genotypes, calculating the association statistics for the ABCA1 and CETP genes (page 384).

The text of this dissertation/thesis contains a reprint of material as it appears in:

Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E.

MATCH: A tool for searching transcription factor binding sites in DNA sequences.

Nucleic Acids Res. 2003 Jul 1;31(13):3576-9.

PMID: 12824369

This was taken without permission (page 407).

The text of this dissertation/thesis contains a reprint of material as it appears in:

**Inoue M, Takata H, Ikeda Y, Suehiro T, Inada S, Osaki F, Aii K, Kumon Y,
Hashimoto K.**

**A promoter polymorphism of the alpha2-HS glycoprotein gene is associated with
its transcriptional activity.**

Diabetes Res Clin Pract. 2008 Jan;79(1):164-70. Epub 2007 Sep 24.

PMID: 17889958 [PubMed - indexed for MEDLINE]

This was taken without permission (page 407).

The text of this dissertation/thesis contains a screenshot reprint of material provided by
the BIOBASE MATCH program version 10.2 [1].

<http://10.1.4.248:8080/cgi-bin/biobase/transfac/10.2/bin/start.cgi>

This was taken without permission (page 401).

The text of this dissertation/thesis contains a screenshot reprint of material as it appears

in: http://www.broad.mit.edu/mpg/haploview/haploview_doc.pdf

This was taken without permission (page 415).

1.2 Personal Acknowledgments

I dedicate the following work to my lovely wife Michelle, my parents Sandy and Bob, and my sister Pamela. Michelle, you are my best friend. I care passionately about you. We have both worked very hard to arrive to this transitional period of our lives. I hope that in our journey ahead I may inspire you as much as you have inspired in me. I cherish you. To my parents and sister. I thank you for supporting my non-linear career path and personal explorations. You have taught me to be inquisitive, ambitious, and self-reliant and this has come at the cost of allowing me to travel and to live at a distance from you. I care about you all deeply.

[Robert W. Mahley](#) (President of the J. David Gladstone Institutes, Thesis Committee Member, Rotation Advisor, Art of Lecturing Mentor) As my primary graduate advisor, mentor and friend, Dr. Mahley has taught me to think critically and to ask questions. He has fostered my career development and I am greatly indebted to him. I was a Research Associate in for him during the 18 months prior to my enrollment into this Ph.D. graduate program.

[Bruce Conklin](#) (Senior Investigator GICD, Thesis Committee Member, Rotation Advisor, Oral Thesis Committee Member) - has helped me to think about how Delta-MATCH fits into the world of computational resources. He was my faculty sponsor for the BMI program.

[Joseph “Mike” McCune](#) (President of the SFGH Department of Experimental Medicine, Thesis Committee Member) - introduced me to the world of HIV-1 biology and immunology and lead the charge into the investigation of TLR9, IRF5, and CCR5. He is

a very patient and supportive man. My first year rotation project in his lab developed into AIM 2.

Ugur Hodoglugil (Post Doctoral Fellow in the Mahley Lab, GICD) - collaborated with me for almost 6 years in the Mahley genetics lab where as part of the Turkish Heart Study, we've investigated many genes (LIPC, ApoA5, MTP, & CETP) for their association with dyslipidemia, hypertriglyceridemia, and the Metabolic Syndrome. Ugur is an excellent biologist and a good friend.

Mark Segal (BMI faculty, UCSF Statistician, Thesis Committee Chair)- chaired my thesis proposal committee and offered keen advice on study design and statistical issues. He was very supportive of my many questions.

Ru-Fang Yeh (BMI faculty, UCSF Statistician, Thesis Committee Member) - offered great feedback on the Delta-MATCH project, and has participated in the glioma and SNPLogic projects.

Sergio Baranzini (BMI faculty, Dept. of Neurology, Rotation Mentor) - advised me in a rotation project where I tried to identify genes associating with allele-specific expression and multiple sclerosis. Sergio is validating some Delta-MATCH predictions in a second round genotyping association study studying multiple sclerosis.

Jorge Oksenberg (BMI faculty, Dept. of Neurology, Rotation Mentor) - advised me in a rotation project trying to identify genes associating with allele-specific expression and multiple sclerosis.

Doug Nixon (GIVI Faculty, Art of Lecturing Mentor) - helped me to hone my presentation skills. He also helped me to interpret the TLR9 data during the earlier part of the AIM 2 investigation.

[Yadong Huang](#) (GICD Faculty, Scientific Mentor) - was one of my primary advisors and mentors during my first years at the Gladstone. The Huang and Mahley groups shared lab meetings for about more than 4 years. He and is an excellent molecular biologist and incredibly intelligent.

[Chris Barker](#) (Gladstone Genomics Core Director) - has offered wonderful advise on numerous scientific projects and career development, and is an expert high-throughput genomics technologist.

Alex Pico (Post Doctoral Fellow in the Conklin Lab) - invited me into the SNPLogic project, and was the strongest Beta user of Delta-MATCH. Alex is an excellent computational biologist and a computer wizard.

Nathan Salomonis (Graduate Student in the Conklin Lab) - a Beta user of Delta-MATCH, and excellent computational biologist. Nathan and I enjoyed co-instructing a bioinformatics lesson for visiting students during the National Student Leadership Conferencev(http://conklinwolf.ucsf.edu/informatics/gene_hunt.html).

Alex Zambon (Post Doctoral Fellow in the Conklin Lab) - brainstormed the Delta-MATCH project and helped to validate the JPH2 EMSA shifts (Figure page 247).

Steven Deeks (UCSF faculty, HIV Clinician at the SFGH) - instructed me on the context of TLR9 and IRF5 in the field of HIV biology and inflammation, and classified HIV-1 patients based on viremia levels.

Peter Hunt (Attending Physician, Positive Health Practice at the SFGH) - invited me to co-author a publication describing the CCR5 genotype results (Chapter 3).

Jeff Martin (UCSF Epidemiologist) - evaluated my case/control study design and classified HIV-1 patients based on viremia levels.

Richard Hecht (San Francisco HIV Cohort Clinician Director) - provided a cohort HIV-positive biological samples.

Susan Buchbinder (San Francisco City Cohort Director) - provided some San Francisco City Cohort HIV positive biological samples.

David Bangsberg (San Francisco City Cohort Director) - provided some San Francisco City Cohort HIV positive biological samples.

Bruce Walker and **Florencia Pererya** (Investigators at Mass. General Hospital) - provided many HIV-positive biological samples of elite suppressors.

Spyros Kalams (Vanderbilt University Infections Disease Clinician) - provided some HIV-positive biological samples of elite suppressors.

Esper Kallas (Sao Paulo Brazil, HIV & Infectious Disease Clinician) - provided some HIV-positive biological samples from Brazil, and hosted my stay in Sao Paulo.

Hugo Barbosa (Sao Paulo HIV Clinician and visiting researcher) - is a wonderful and spirited HIV clinician in Brazil who shared with me his interests in HIV and HTLV-1 pathology.

Mimi Zeiger (Art of Lecturing Mentor) - enhanced my presentation and writing skills.

Sylvia Richmond (Administrator to the President) - coordinated the Mahley lab and was my fairy godmother. She shows wonderful maternally care and has helped me to stay on course and out of harms way.

Margaret Wrench (UCSF Faculty) - coordinates the SNPLogic and glioma projects.

Tim Beaumont (Post Doctoral Fellow in the McCune Lab) - identified TLR9 as a candidate gene for HIV pathogenesis. He invited me into my rotation in the McCune lab and is a wickedly sharp footballer.

Mark Schwenecker (Post-doctoral Fellow in the McCune Lab) - conducted some western blots using a TLR9 isoform-specific antibody I designed.

Sunil Emu-Parikh (Epidemiology graduate student at Berkeley) - genotyped four TLR9 SNPs in African children attempting to show correlation with genotype and malarial resistance.

Mary Beth Hanley - (Senior Research Associate in the McCune Lab) - extracted DNA many HIV positive PMBC samples. She helped to coordinate the early HIV cohort collection.

Sophie Stephenson (Research Associate in the Core Lab) - extracted DNA many HIV positive PMBC samples.

Christophe Kreis (GIVI Senior Research Associate) - taught me how to use the ABI 7700 Taqman machine.

Katie Pollard (Faculty at UC Davis) - spearheaded the HAR project and is following up on some human- and chimpanzee-specific Delta-MATCH predictions through collaborations with the Guillemot lab.

The following work was conducted at

The J. David Gladstone Institutes

Gladstone Institute of Cardiovascular Disease (GICD)

Gladstone Institute of Immunology and Virology (GIVI)

1650 Owens Street San Francisco, CA 94158

Genetic and Bioinformatic Approaches To Identify Polymorphic Modulators of Transcription Factor Binding and Disease Phenotypes Including HIV-1 Viremia

David Wayne Williamson

Abstract

(PROBLEM) The overall goal of this thesis is to identify polymorphic alleles that associate with elevated risk and disease progression. Two different approaches were used to achieve this goal. (METHODS AIM 1) A database resource called Delta-MATCH was created using a predictive computational approach. The aim of the Delta-MATCH program is to identify human polymorphic variants that may create allele-specific transcription factor binding sites. In this version (v 1.0) 4,547,844 high-value candidate polymorphisms have been scored and ranked by the Delta-MATCH algorithm. These polymorphisms were either positioned within a 10,000 base pair window of a refSeq gene, or located within a region of high conservation in the human genome. The major and minor alleles for each of these 4.5 million polymorphisms were independently evaluated by the MATCH algorithm against a library of 550 known transcription factor binding site motifs (BIOBASE TRANSFAC v10.2) to determine the “highest MATCH scores” for each allele and transcription factor pair. (CONCLUSIONS AIM 1) The ranked list of Delta-MATCH predictions for each transcription factor binding site (matrix name) can be queried online at <http://deltamatch.org>. Predictions have been ranked in descending order of importance by a statistic called the “Delta-MATCH potential score”, which reflects the potential of a polymorphism to create an allele-specific transcription factor binding site. (METHODS AIM 2) The common genotypes and haplotypes of four candidate genes (CCR5, TLR9, IRF5, APOE) were investigated for their association with the phenotype of HIV-1 viremia levels in a population of HIV-infected Americans

primarily derived from the San Francisco SCOPE cohort. (CONCLUSIONS AIM 2) TLR9 and IRF5 variants associated with HIV viremia levels in White Americans. Additionally, individuals infected with HIV should try to avoid chronic inflammation, which means avoiding other viral and bacteria coinfections, traumas, and other behaviors that promote a chronic inflammatory state. Furthermore, the magnitude of TLR9- and IRF5-dependant inflammatory responses during the acute phase of HIV-1 infection may partially determine the viremia level of chronic infection (CVL classification).

Table of Contents

Dedication and Acknowledgments	iii
1.1 Publication Reprints	iii
1.2 Personal Acknowledgments	vi
Abstract	xiii
Table of Contents	xv
List of Tables	xxiii
List of Figures	xxiv
List of Equations	xxx
Introduction	1
1.3 AIM 1: Delta-MATCH: A Computational Survey	3
1.4 AIM 2: A Genetic Survey of Genetic Modulators of HIV-1 Viremia	4
Chapter 1: Delta-MATCH - A Computational Survey	5
1.5 Delta-MATCH Overview	5
1.5.1 The Aim of the Delta-MATCH Program	5
1.5.2 Transcription Factor Binding Affinity May Be Correlated with the Level of mRNA Expression and Associated with Some Human Diseases (Δ binding affinity \approx Δ expression)	7
1.5.3 The Delta-MATCH Hypothesis	9
1.6 Computational Survey	10
1.6.1 What is Delta-MATCH Query Tool?	10
1.6.2 The Delta-MATCH Query Tool (DMQT) Website Address	10
1.6.3 DMQT Overview	11
1.6.4 Building a Workstation	12
1.6.5 Computing Time	12
1.6.6 How the Delta-MATCH Query Tool Was Constructed	13
1.6.7 SNP Identification and Selection	13
1.6.8 What is a Transcription Factor Binding Site Matrix?	15
1.6.8.1 Definition - matrix (transcription factor binding site matrix)	15
1.6.9 Transcription Factor Binding Site Matrix (percentage count)	16
1.6.10 Transcription Factor Binding Site Matrix (after eigenvector multiplication)	16
1.6.10.1 Definition - information eigenvector	17
1.6.11 What is a MATCH Score?	17
1.6.11.1 Definition - MATCH score	17
1.6.12 Three Genetic Models of Human Disease Paired with High or Low Levels of mRNA Transcription	18
1.6.12.1 Definition - phenotype case 1 (low mRNA transcription = disease)	18
1.6.12.2 Definition - phenotype case 2 (high mRNA transcription = disease)	18
1.6.12.3 Definition - genetic model 1 - (dominant model)	19
1.6.12.4 Definition - genetic model 2 - (co-dominant model)	19
1.6.12.5 Definition - genetic model 3 - (recessive model)	19
1.6.13 Hardy-Weinberg Expectation Equations	19
1.6.13.1 Definition - Hardy-Weinberg Expectation (HWE)	19
1.6.14 The Predicted Genotype Frequencies of Three Genetic Models Paired with High or Low Levels of mRNA Transcription	21
1.6.15 The Predicted Phenotype Frequencies of Three Genetic Models Paired with High or Low Levels of mRNA Transcription	22

1.6.16	A Large Difference in MATCH Score May Correlate with a Large Difference in Transcription Factor Binding Affinity	26
1.6.17	Can a Large Delta-MATCH Score Identify a Genetic Locus Associated with Human Disease?	29
1.7	The Delta-MATCH Method (Predicting Which Polymorphisms May Create Allele-Specific Binding Sites)	30
1.7.1	What is Biological Relevance?	30
1.7.1.1	Definition - biological relevance	30
1.7.2	What is a "Delta-MATCH Potential Score (potential)?	30
1.7.2.1	Definition - potential (Delta-MATCH Potential Score)	30
1.7.2.2	Warning - The "Delta-MATCH potential score" is informative, but not sufficient	30
1.7.3	The Threshold of Biological Relevance Is Estimated By the False Positive Threshold Cutoff Score	31
1.7.3.1	Definition - cutoff threshold of biological relevance	31
1.7.3.2	Definition - false positive cutoff score (FP)	31
1.7.4	The False Positive (FP) Cutoff Is Not Correlated with Matrix Length	33
1.7.4.1	Definition - model	34
1.7.4.2	Definition - biological relevance of a MATCH score (brm)	34
1.7.5	How is Biological Relevance Calculated?	35
1.7.6	Calculating the "absolute percent difference" in allelic MATCH scores and the "Delta-MATCH potential score"	36
1.7.6.1	Definition - mean MATCH score	36
1.7.6.2	Definition - larger polymorphism MATCH score (m_max)	37
1.7.6.3	Definition - smaller polymorphism MATCH score (m_min)	37
1.7.6.4	Definition - absolute difference in MATCH score (m_dif)	37
1.7.6.5	Definition - absolute percent difference in MATCH score (m_per)	38
1.7.7	How is a Delta-MATCH Potential Score for a Polymorphism Calculated?	38
1.7.8	Ranking Delta-MATCH Results (by potential, (max (m1, m2)), m_per)	39
1.7.9	Calculating Example Potential Scores (Estimation Model 2)	39
1.7.10	The Delta-MATCH Estimation Model Is Linear (used in version 1.0)	46
1.7.11	What Level of Potential Score Is Considered Significant?	47
1.7.12	Future Versions of Delta-MATCH May Use Higher Order Models That May Reduce Type-1 Errors (False Positives)	48
1.7.12.1	Calculating Example Potential Scores (Estimation Model 2)	50
1.7.13	Comparison of Estimation Model 1 and Estimation Model 2 Ranked Examples	52
1.7.14	Estimation Model 1 ranked examples	52
1.7.15	Estimation Model 2 ranked examples	52
1.7.15.1	Definition - biological relevance of a polymorphic site (brps)	54
1.7.16	Viewing a Ranked Set of Delta-MATCH Potential Scores Graphically	55
1.7.17	How to Calculate the Rareness of a Single Delta-MATCH Result	64
1.7.18	Caveats of the Delta-MATCH Method	64
1.7.18.1	Warning: Do Not Compare Absolute Potential Scores Across Different TFBS Matrixes	64
1.8	The Delta-MATCH Algorithm	66
1.8.1	Polymorphism Selection	66
1.8.2	Polymorphism Exclusions	67
1.8.3	Creating Double-Stranded DNA Allele Sequences	67
1.8.4	Computing the Highest MATCH Scores	70
1.8.5	Recording Delta-MATCH Scores	71
1.8.5.1	Definition - s1 and s2	71
1.8.5.2	Definition - p1 and p2	71
1.8.5.3	Definition - m1 and m2	71
1.8.6	Identifying the Highest MATCH Score for an Allele (Exhaustive Search)	72
1.8.7	Why Was a 61 Base Pair Length of Sequence Chosen?	72

1.9 The Delta-MATCH Database.....	75
1.9.1 How Many Results Are In the Delta-MATCH Database?	75
1.9.1.1 Definition - Delta-MATCH hit or result.....	75
1.9.2 No Correlation Between Matrix Length and Number of Delta-MATCH Hits.....	76
1.9.3 The Delta-MATCH Query Tool Search Engine (version 1.0)	81
1.9.4 Creating a Delta-MATCH Query.....	81
1.9.5 Creating a Query Using the Delta-MATCH Query Tool.....	83
1.9.6 Easy Mode vs. Expert Mode.....	87
1.9.7 Easy Mode Selections.....	87
1.9.8 Expert Mode Additional Selections	88
1.10 Easy Mode.....	97
1.10.1 STEP 1 - Select Matrix Names.....	97
1.10.1.1 Primary Matrix Selection Button 1 - Single Transcription Factor Matrix Name.....	98
1.10.1.2 Primary Matrix Selection Button 2 - List of Transcription Factor Matrix Names	99
1.10.1.3 Primary Matrix Selection Button 3 - Transcription Factor Name.....	99
1.10.1.4 Primary Matrix Selection Button 4 - Tissue-Specific Transcription Factor Names	100
1.10.1.5 Table - Tissues Types in the Delta-MATCH Query Tool	100
1.10.1.6 Primary Matrix Selection Button 5 - All Transcription Factor Matrix Names .	101
1.10.2 STEP 2 -Add Restriction Criteria.....	101
1.10.2.1 Warning - Please Read About Each Restriction Criteria Before Checking Everything in Sight.....	102
1.10.2.2 Minimum Potential Score	102
1.10.2.3 Warning - Don't compare the potential scores between different matrix names	103
1.10.2.4 Selecting the best Minimum Potential Score Value (potential >= 0.3).....	103
1.10.2.5 Top Most Significant Hits	104
1.10.2.6 Matrix Quality	104
1.10.2.7 Sort Results Table	105
1.10.2.8 Search By rsnumbers.....	106
1.10.2.9 Uploading a List of rsnumbers	106
1.10.2.10 rsnumber Window.....	107
1.10.2.11 Search By Gene Names	107
1.10.2.12 Search for Gene Without Returning Results' (MOCK SEARCH).....	108
1.10.2.13 What Happens When a Gene Name has Multiple Transcripts?	109
1.10.3 STEP 3 - Submit (press the submit button).....	110
1.10.3.1 Hint - Opening Your Output Results Page in a New Tab (right click option). 110	
1.10.3.2 A Delta-MATCH Query May Take Seconds or Minutes (up to tens of minutes)	111
1.10.4 A Successful Delta-MATCH Run Creates 5 Output Files.....	112
1.10.5 Viewing Delta-MATCH Data as UCSC Genome Browser Tracks	118
1.10.6 Description of the Delta-MATCH UCSC Tracks.....	119
1.10.6.1 Definition - rsnumber_A1	120
1.10.6.2 Definition - rsnumber_A2	120
1.10.6.3 Definition - rsnumber_P	120
1.11 Delta-MATCH Examples (Easy Mode).....	121
1.11.1 Delta-MATCH Proof of Principle Example - AHSG rs2248690.....	122
1.11.1.1 Example OMIM Links for ASHG	136
1.11.2 Delta-MATCH Query Examples (Easy Mode).....	137
1.11.3 Example 1 - Single Transcription Factor Matrix Name (the default submission)..	137
1.11.4 Figure - Example 1 Results Table.....	138
1.11.4.1 Definition - hit (Delta-MATCH rsnumber row in the query result table)	138
1.11.4.2 Definition - rsnumber (dbSNP accession).....	139
1.11.4.3 Definition - chrom (chromosome).....	139

1.11.4.4	Definition - chromStart (polymorphism starting base position).....	139
1.11.4.5	Definition - factor (transcription factor name)	139
1.11.4.6	Definition - mat_id (matrix name)	139
1.11.4.7	Definition - potential (Delta-MATCH Potential Score).....	140
1.11.4.8	Definition - threshold (FP = false positive cutoff threshold).....	140
1.11.4.9	Definition - m1 (highest MATCH score for allele 1).....	140
1.11.4.10	Definition - m2 (highest MATCH score for allele 2).....	140
1.11.4.11	Definition - m_per (absolute percent difference in MATCH score)	141
1.11.4.12	Definition - rank.....	141
1.11.4.13	Definition - p1_window (UCSC position window of the highest allele 1 MATCH score).....	142
1.11.4.14	Definition - pubmed (link to PubMed citations).....	142
1.11.5	Example 2 - List of Transcription Factor Matrix Names	144
1.11.6	Example 3 - Transcription Factor Name	145
1.11.7	Example 4 - Tissue-Specific Transcription Factor Names	146
1.11.8	Example 5 - Top Most Significant Hits (unchecked).....	147
1.11.9	Example 6 - Minimum Potential Score (unchecked).....	148
1.11.10	Example 7- Error 1 - no matrixes passed your selected criteria	150
1.11.11	Example 8 - Error 2 - more than 1,500 results returned.....	151
1.11.12	Example 9 - Error 3 - no rsnumbers were found.....	152
1.11.13	Example 10 - Searching by rsnumbers and Sorting By Chromosomal Position	153
1.11.14	Example 11 - Using the “rsnumber Window” checkbox	155
1.11.15	Example 12 - Uploading a File of rsnumbers	157
1.11.16	Example 13 - ‘Search By Gene Names’ Without Returning Results (mock search when unsure of true gene names)	159
1.11.17	Example 14 - ‘Search By Gene Names’ (includes using the “Gene Window” sub- checkbox).....	163
1.11.18	Example 15 - Error 4 - no rsnumbers were found in the select gene names (bad gene name submission).....	166
1.11.19	Example 16 - Error 6 - more than 5 gene names were submitted.....	168
1.12	Delta-MATCH Query Examples (Expert Mode).....	170
1.12.1	Show the Matrix Details	170
1.12.1.1	Definition - factor_description (expanded factor name).....	170
1.12.1.2	Definition - count_ge_potential (count of hits with a potential score greater than or equal to this potential score)	170
1.12.1.3	Definition - mat_count (number of hits in the database for this matrix)	170
1.12.1.4	Definition - rareness (rareness of a potential score).....	171
1.12.1.5	Definition - qual (quality of a matrix).....	171
1.12.1.6	Definition - mat_len (matrix length)	171
1.12.2	Example 17 - Show the Matrix Details.....	172
1.12.3	Minimum Matrix Length.....	173
1.12.4	Example 18 - ‘Minimum Matrix Length’ sub-checkbox	174
1.12.5	Show the Position Details	175
1.12.5.1	Definition - p2_window (UCSC position window of the highest allele 2 MATCH score)	175
1.12.5.2	Definition - p1 (position offset of highest allele 1 MATCH score).....	175
1.12.5.3	Definition - p2 (position offset of highest allele 2 MATCH score).....	176
1.12.5.4	Definition - s1 (strand of highest allele 1 MATCH score).....	176
1.12.5.5	Definition - s2 (strand of highest allele 2 MATCH score).....	176
1.12.6	Example 19 - ‘Show the Position Details’	176
1.12.7	Chromosome	182
1.12.8	Position Range	182
1.12.9	Example 20 - Restricting By Chromosome and Position Range	182
1.12.10	Strand.....	184
1.12.11	Example 21 - Strand	184

1.12.12	Genomic Regions	185
1.12.12.1	Definition - up10k	185
1.12.12.2	Definition - phastconsElements17way	185
1.12.12.3	Definition - utr5	185
1.12.12.4	Definition - coding	185
1.12.12.5	Definition - down10k	186
1.12.12.6	Definition - exons	186
1.12.12.7	Definition - introns	186
1.12.12.8	Definition - utr3	186
1.12.12.9	Definition - all	186
1.12.13	Example 22 - Genomic Regions	186
1.12.14	Bonferonni Correction	188
1.12.15	Example 23 - Bonferonni	190
1.12.16	Minimum Number of Delta-MATCH Hits	191
1.12.16.1	Definition - number_hits	192
1.12.17	Example 24 - Minimum Total Number of Delta-MATCH Hits	192
1.12.18	Hugo Names	194
1.12.18.1	Definition - hugo_name	195
1.12.19	Example 25 - HUGO Names	195
1.12.20	Reflink	197
1.12.20.1	Definition - reflink_product	198
1.12.20.2	Definition - reflink_mrnaAcc	198
1.12.20.3	Definition - reflink_protAcc	198
1.12.20.4	Definition - reflink_name	198
1.12.20.5	Definition - reflink_prodName	198
1.12.20.6	Definition - reflink_locusLinkId	198
1.12.20.7	Definition - reflink_omimId	199
1.12.21	Example 26 - Reflink	199
1.12.22	Distance from txStart or cdStart	201
1.12.22.1	Definition - dist_from_ref (distance from reference)	201
1.12.22.2	Definition - dist_from_tx (distance from transcription start site)	201
1.12.22.3	Definition - dist_from_cds (distance from coding start site)	202
1.12.23	Example 27 - Distance From txStart or cdStart	202
1.12.24	Gene Ontology	204
1.12.24.1	Definition - go_names (gene ontology names)	204
1.12.24.2	Definition - go_number (gene ontology number)	204
1.12.25	Example 28 - Gene Ontology	204
1.12.26	Affymetrix	206
1.12.27	Example 29 - Affymetrix	207
1.12.28	Using the HapMap Database to Find Other rsnumbers in Strong Linkage Disequilibrium with Polymorphisms on an Affymetrix SNPchip	208
1.12.28.1	Definition - linkage disequilibrium (LD)	209
1.12.28.2	Definition - rsquare (r^2 linkage disequilibrium value)	209
1.12.28.3	Definition - dprime (D' linkage disequilibrium value)	209
1.12.29	Example 30 - Affymetrix with Linkage Disequilibrium	211
1.12.29.1	Definition - name_affy	211
1.12.30	Illumina	213
1.12.31	Example 31 - Illumina	215
1.12.32	Example 32 - Illumina with Linkage Disequilibrium	216
1.12.32.1	Definition - name_illumina	216
1.12.33	Example 33 - Affymetrix and Illumina (all checkboxes checked)	217
1.12.34	HapMap	219
1.12.34.1	Definition - ld_name	220
1.12.34.2	Definition - ld_name_affy	220
1.12.34.3	Definition - ld_name_illumina	220
1.12.34.4	Definition - ld_lod	220

1.12.34.5	Definition - Id_dprime	221
1.12.34.6	Definition - Id_rsquare	221
1.12.34.7	Definition - Id_pos_dif.....	221
1.12.34.8	Definition - Id_pos1_hg17	221
1.12.34.9	Definition - Id_pos2_hg17	221
1.12.34.10	Definition - Id_fbin	221
1.12.35	Example 34 - Affymetrix with HapMap.....	222
1.12.36	Example 35 - Affymetrix with HapMap (with Minimum Total Number of Delta-MATCH Hits).....	225
1.12.37	HIV-1 Candidate Genes.....	226
1.12.38	Example 36 - HIV-1 Candidate Genes	227
1.12.38.1	Definition - log P-value (-logp)	227
1.12.39	Copy Number Variation.....	228
1.12.40	Example 37 - Copy Number Variation	228
1.12.41	PReMod Modules.....	229
1.12.42	Example 38 - PReMod Modules	230
1.12.43	UCSC rsnumber Details	232
1.12.43.1	Warning - Using the “and” buttons will greatly increase computation time.	232
1.12.43.2	Definition - reference base at the UCSC Browser (refUCSC).....	232
1.12.43.3	Definition - reference base at NCBI (refNCBI)	232
1.12.43.4	Definition - the observed alleles at this rsnumber (observed).....	233
1.12.43.5	Definition - rsnumber strand (strand)	233
1.12.43.6	Definition - Validation Types (validtype)	233
1.12.43.7	Definition - Function Types (functype)	233
1.12.43.8	Definition - Locations Types (loctype).....	233
1.12.43.9	Definition - Molecular Types (moltype)	233
1.12.43.10	Definition - Average Heterozygosity (avHet)	233
1.12.43.11	Definition - Average Heterozygosity (avHetSE)	233
1.12.44	Example 39 - UCSC rsnumber Details	234
1.12.45	Example 40 - NF-kB (rs5743836, rs6031444, rs28431981)	237
1.13	Predicting Modulators of NF-kB-dependent Transcription.....	239
1.13.1	Junctophilin 2 (JPH2) rs6031444 G>T	239
1.13.2	Toll-like receptor 9 (TLR9) rs5743836 T>C.....	243
1.13.3	Kynurenine 3-monooxygenase (KMO) rs28431981 A>G.....	245
1.13.4	Validating the Delta-MATCH NF-kB Predictions.....	246
1.14	Validating Other Delta-MATCH Predictions.	248
1.15	Using Delta-MATCH To Identify Species-Specific Transcription Factor Binding Sites Though Comparative Genomics	249
1.15.1	Background.....	249
1.15.2	Method	249
1.15.3	Results (HAR152/PAX6).....	250
1.15.4	Discussion.....	253
1.16	Conclusions for AIM 1	254
Chapter 2: A Genetic Survey of Genetic Modulators of HIV-1 Viremia		255
1.17	Background.....	255
1.18	Methods Genetic Survey	259
1.18.1	CVL Classification (viremia level).....	259
1.18.2	Study Design (a genotype and haplotype analysis of 11 polymorphisms).....	263
1.18.3	Results (genotype data).....	268
1.19	CCR5 - Chemokine Receptor 5	289
1.19.1	CCR5 Background	289

1.19.2	CCR5 Results	292
1.20	TLR9 - Toll-Like Receptor 9.....	293
1.20.1	TLR9 Background	293
1.20.1.1	Confirmed Associations	294
1.20.1.2	Failed Associations.....	294
1.20.2	TLR9 Method	294
1.20.3	TLR9 Results	303
1.21	IRF5 - Interferon Regulatory Fragment 5.....	312
1.21.1	IRF5 Background.....	312
1.21.2	IRF5 Results	323
1.22	APOE - Apolipoprotein E.....	332
1.22.1	APOE Background	332
1.22.2	APOE Results	335
1.23	Conclusions for AIM 2	342
1.23.1	CCR5 Conclusions	342
1.23.2	TLR9 Conclusions	343
1.23.3	IRF5 Conclusions	344
1.23.4	APOE Conclusions	344
	<i>Chapter 3: Prevalence of CXCR4 tropism among antiretroviral-treated HIV-1-infected patients with detectable viremia</i>	345
	<i>Chapter 4: Apolipoprotein A-V: a potential modulator of plasma triglyceride levels in Turks</i>	350
	<i>Chapter 5: An interaction between the TaqIB polymorphism of cholesterol ester transfer protein and smoking is associated with changes in plasma high-density lipoprotein cholesterol levels in Turks</i>	360
	<i>Chapter 6: Common polymorphisms of ATP binding cassette transporter A1, including a functional promoter polymorphism, associated with plasma high density lipoprotein cholesterol levels in Turks</i>	370
	<i>Chapter 7: Low HDL-C: lessons learned from the Turkish Heart Study</i>	384
	<i>Thesis Discussion</i>	391
1.23.5	Future Technology	391
1.23.6	Controlling for Ethnicity	392
1.23.7	Studying Rare Phenotypes	393
1.23.8	Transitioning	394
	<i>Bibliography</i>	396
	<i>Appendices.....</i>	401
1.24	AIM 1 Extras (Delta-MATCH)	401
1.24.1	Delta-MATCH MYSQL Databases and Tables	408
1.24.2	The Delta-MATCH XML DTD	408
1.24.3	List of Delta-MATCH Errors.....	408
	AIM 2 Extras (A Genetic Survey)	415
1.25	Other Software by David W. Williamson	416
1.25.1	What Color Eyes Would Your Children Have? (flash)	416
1.25.2	What Color Eyes Would Your Children Have? (html).....	416
1.25.3	SNP Enzyme Finder.....	416
1.25.4	Haplotype Mapper	416

1.26	Ph.D. Thesis Defense (February 06, 2008)	418
1.26.1	Seminar Announcement	418
	<i>UCSF Library Release</i>	420

List of Tables

Table 1 Predicted Genotype Frequencies of Three Genetic Models.....	21
Table 2 Predicted Phenotype Frequencies of Three Genetic Models	22
Table 3 Distribution of Delta-MATCH Hits for Matrix Name V\$NFKB_Q6.....	62
Table 4 Distribution of Delta-MATCH Hits and Counts for High and Low Quality Matrixes	77
Table 5 Cohorts Genotyped for CCR5, TLR9, IRF5 and APOE Polymorphisms.....	261
Table 6 Genotyping Conditions (TLR9, CCR5, IRF5, APOE)	265
Table 7 PCR Conditions for Genotyping (RFLP).....	266
Table 8 PCR Primer Sequences.....	267
Table 9 Genotype Counts Test1.....	269
Table 10 Genotype Counts Test2.....	270
Table 11 Genotype Counts Test3.....	271
Table 12 Genotype Counts of Other Non-HIV Positive Populations.....	272
Table 13 Genotype Frequencies Test1.....	273
Table 14 Genotype Frequencies Test2.....	274
Table 15 Genotype Frequencies Test3.....	275
Table 16 Genotype Frequencies of Other Non-HIV Positive Populations.....	276
Table 17 Allele Frequencies Test1	277
Table 18 Allele Frequencies Test2	278
Table 19 Allele Frequencies Test3	279
Table 20 Allele Frequencies of Other Non-HIV Positive Populations	280
Table 21 Number of Samples Per Cohort Test1	281
Table 22 Number of Samples Per Cohort Test2.....	282
Table 23 Number of Samples Per Cohort Test3.....	283
Table 24 Number of Samples Per Cohort of Other Non-HIV Positive Populations.....	284
Table 25 Haploview Chi-Square Permuted-p Values Test1	285
Table 26 Haploview Chi-Square Permuted-p Values Test2.....	286
Table 27 Haploview Chi-Square Permuted-p Values Test3.....	287
Table 28 Haploview Chi-Square Permuted-p Values of Other Non-HIV Positive Populations	288
Table 29 PCR Conditions (TLR9 extended).....	302
Table 30 TLR9 Haplotypes Test1 (CVL-1/2/3 vs CVL-4)	305
Table 31 TLR9 Haplotypes Test2 (CVL-1/2 vs CVL-4)	306
Table 32 TLR9 Haplotypes Test3 (CVL-1 vs CVL-4)	307
Table 33 IRF5 Haplotypes Test1 (CVL-1/2/3 vs CVL-4)	325
Table 34 IRF5 Haplotypes Test2 (CVL-1/2 vs CVL-4)	326
Table 35 IRF5 Haplotypes Test3 (CVL-1 vs CVL-4)	327
Table 36 APOE Haplotypes Test1 (CVL-1/2/3 vs CVL-4)	336
Table 37 APOE Haplotypes Test2 (CVL-1/2 vs CVL-4)	337
Table 38 APOE Haplotypes Test3 (CVL-1 vs CVL-4)	338
Table 39 Delta-MATCH Tissue Types.....	404
Table 40 NF-kB TFBS Matrixes Used by Delta-MATCH.....	404
Table 41 Distribution of Potential Scores (dif_z) for NF-kB TFBS Matrixes.....	404
Table 42 List of 351 Transcription Factors.....	405
Table 43 List of 584 Matrix Names.....	405
Table 44 Distribution of Polymorphisms in the human genome (hg18.snp126).....	406

List of Figures

Figure 1 A Polymorphism May Create an Allele-specific Transcription Factor Binding Site	6
Figure 2 Transcription Factor Binding Affinity May Positively Correlate with Level of mRNA Expression	8
Figure 3 Transcription Factor Binding Affinity May Negatively Correlate with Level of mRNA Expression	9
Figure 4 Transcription Factor Binding Site Matrix	16
Figure 5 Transcription Factor Binding Site Matrix After Eigenvalue Correction	17
Figure 6 Phenotype Frequencies Case1 /Model Dominant	23
Figure 7 Phenotype Frequencies Case1 /Model Co-Dominant	23
Figure 8 Phenotype Frequencies Case1 /Model Recessive	24
Figure 9 Phenotype Frequencies Case2 /Model Dominant	24
Figure 10 Phenotype Frequencies Case2 /Model Co-Dominant	25
Figure 11 Phenotype Frequencies Case2 /Model Recessive	25
Figure 12 Density plot of allelic MATCH scores for 4,547,844 polymorphisms using the NF-kappaB Matrix V\$NFKB_Q6	27
Figure 13 The FP Threshold Cutoff Represents the Minimum MATCH Score Required to Recruit a Transcription Factor to a Sequence	33
Figure 14 False Positive Cutoff Score vs. Matrix length	34
Figure 15 Estimation Model 1 - a linear estimation curve	35
Figure 16 Count Versus Mean MATCH Score (V\$NFKB_Q6, n = 4,547,844)	43
Figure 17 Histogram of MATCH scores for 4,547,844 polymorphisms using the NF-kappaB Matrix V\$NFKB_Q6	44
Figure 18 Delta-MATCH estimates the biological relevance of a MATCH score with a linear model that approximates transcription factor binding affinity	45
Figure 19 Future Alternative Delta-MATCH Models May Use Exponential Estimation Curves	48
Figure 20 Estimation Model 2 - an exponential estimation curve	49
Figure 21 Density Plot of the Allelic MATCH Scores for 4,547,844 Polymorphisms (NF-kB)	56
Figure 22 Count Versus Biological Relevance of a MATCH Score	57
Figure 23 Absolute Difference in MATCH Score vs. Larger MATCH Score of a Polymorphism (the ranked distribution)	58
Figure 24 Potential Score Versus Biological Relevance of a Polymorphic Site	60
Figure 25 Potential Score Versus Absolute Percent Difference in MATCH Score	61
Figure 26 Rank versus potential score versus absolute percent difference in MATCH score for 950 high-value polymorphisms (3-D plot)	63
Figure 27 Location of SNPs Evaluated by Delta-MATCH	66
Figure 28 The 61 Base Pairs of DNA Sequence Surrounding rs6013444 in the UCSC Genome Browser (Mar. 2006 Assembly)	68
Figure 29 The DAS DNA Sequence Retrieval Web Tool (retrieving the 61 bp sequence surrounding rs6013444)	69
Figure 30 Determining the highest MATCH scores for a pair of alleles	73
Figure 31 Number of Delta-MATCH Results vs. Matrix Length for 4.5 Million Hits	78
Figure 32 Count of Matrixes vs. Matrix Length For High and Low Quality Matrixes	78
Figure 33 The Delta-MATCH Website, http://deltamatch.org	79
Figure 34 The Delta-MATCH website hosts tutorials, examples, and downloadable data tables.	80
Figure 35 List of selectable parameters at the Delta-MATCH website	82

Figure 36 The number of Delta-MATCH hits returned is dependent on the parameters selected.....	84
Figure 37 Number of Hits Returned vs. Parameter (Description)	85
Figure 38 Delta-MATCH Returns the Intersection of Restriction Criteria.....	86
Figure 39 Delta-MATCH Easy Mode Input Page	89
Figure 40 Additional Parameter Fields Included in the Expert Mode.....	91
Figure 41 STEP 1 - Select Matrix Names	98
Figure 42 Minimum Potential Score Input.....	104
Figure 43 Top Most Significant Hits.....	104
Figure 44 Matrix Quality Input.....	105
Figure 45 Sort Results Table Input.....	105
Figure 46 Sorting Selections	105
Figure 47 Search By rsnumbers.....	106
Figure 48 Search By Gene Names.....	107
Figure 49 TLR9 Isoforms.....	109
Figure 50 The Delta-MATCH Output Results Page	110
Figure 51 Download and save Delta-MATCH results as HTML, XML or TXT files	112
Figure 52 Right Click a Web Link To Download a Temporary Result Table or Log File (Firefox).....	113
Figure 53 Downloadable File of the Results Table (DM_*_table.html)	113
Figure 54 Downloadable File of the Results Table (DM_*_table.txt).....	114
Figure 55 Downloadable File of the Results Table (DM_*_table.xml) (viewed in text program)	115
Figure 56 Downloadable File of the Results Table (DM_*_table.xml) (viewed in text web browser)	116
Figure 57 Downloadable Log File (DM_*_log.html).....	117
Figure 58 Delta-MATCH Data Can Be Visualized as a Custom Track in the UCSC Genome Browser	119
Figure 59 The AHSG -799T Allele Has a Higher Affinity for the AP-1 Transcription Factor Than -779 A	123
Figure 60 Input Parameters for the Proof of Principle Example	125
Figure 61 AHSG rs2248690 A>T Delta-MATCH Scores.....	132
Figure 62 Pressing the p1_window Button (chr3:187812781-187812789).....	133
Figure 63 Density Plot of the Allelic MATCH Scores for 4,547,844 Polymorphisms (AP-1)	134
Figure 64 AHSG rs2248690 A>T Ranks 747 th for AP-1 TFBS Matrix (V\$AP1_C).....	135
Figure 65 AHSG rs2248690 Is in Linkage Disequilibrium with Other SNPs that Associate with Type 2 Diabetes.....	136
Figure 66 Input Parameters for Example 1	137
Figure 67 ASHG rs22486890 A>T Proof of Concept PubMed link (pubmed).....	142
Figure 68 rs3093317 Hyperlink to the UCSC Human Genome Browser (hg18.snp126)	143
Figure 69 Input Parameters for Example 2	144
Figure 70 Input Parameters for Example 3	145
Figure 71 Input Parameters for Example 4	146
Figure 72 Input Parameters for Example 5	147
Figure 73 Input Parameters for Example 6	148
Figure 74 Input Parameters for Example 7	150
Figure 75 Input Parameters for Example 8	151
Figure 76 Input Parameters for Example 9	152
Figure 77 Input Parameters for Example 10	153

Figure 78 Input Parameters for Example 11	155
Figure 79 Input Parameters for Example 12	157
Figure 80 Input Parameters for Example 13	160
Figure 81 Example Entry Found By a Mock Gene Name Search	161
Figure 82 Summary of Gene Names Found	161
Figure 83 Summary of Gene Names Not Found.....	162
Figure 84 Input Parameters for Example 14	163
Figure 85 Summary of rsnumbers found in Gene Names	165
Figure 86 Summary of rsnumbers found in Gene Names (bad_gene_name)	165
Figure 87 Input Parameters for Example 15	166
Figure 88 Input Parameters for Example 16	168
Figure 89 Summary of the First 5 Submitted Gene Names.....	169
Figure 90 Input Parameters for Example 17	172
Figure 91 Output Results Showing the Matrix Details (sorted).....	173
Figure 92 Input Parameters for Example 18	174
Figure 93 Input Parameters for Example 19	177
Figure 94 UCSC Browser Example 19 (rs6031444)	179
Figure 95 UCSC Browser Example 19 (rs1680789)	180
Figure 96 UCSC Browser Example 19 (rs2104240)	181
Figure 97 Input Parameters for Example 20	182
Figure 98 Input Parameters for Example 21	184
Figure 99 Input Parameters for Example 22	186
Figure 100 Example 22a (button set to 'or')	187
Figure 101 Example 22b (button set to 'and')	188
Figure 102 Input Parameters for Example 23	190
Figure 103 Bonferonni - Adjusted Rareness (bonferonni)	191
Figure 104 Input Parameters for Example 24	192
Figure 105 Example 24A Results Table	194
Figure 106 Example 24B Results Table	194
Figure 107 Input Parameters for Example 25	195
Figure 108 Example 25 Results Table.....	197
Figure 109 Input Parameters for Example 26	199
Figure 110 Example 26 Results Table.....	200
Figure 111 Input Parameters for Example 27	202
Figure 112 Example 27 Results Table.....	203
Figure 113 Input Parameters for Example 28	205
Figure 114 Example 28 Results Table.....	206
Figure 115 Input Parameters for Example 29	207
Figure 116 Population / Linkage Disequilibrium rsquare Pairs for the Affymetrix 500k SNPchip	210
Figure 117 Input Parameters for Example 30	211
Figure 118 Example 30 Results Table.....	213
Figure 119 Population / Linkage Disequilibrium rsquare Pairs for the Illumina 550k SNPchip	214
Figure 120 Input Parameters for Example 31	215
Figure 121 Input Parameters for Example 32	216
Figure 122 Input Parameters for Example 33	218
Figure 123 Example 33 Results Table.....	219
Figure 124 Input Parameters for Example 34	222
Figure 125 Example 34 Results Table.....	224
Figure 126 Input Parameters for Example 35	225

Figure 127 Example 35 Results Table (partial).....	226
Figure 128 Input Parameters for Example 36	227
Figure 129 Input Parameters for Example 37	229
Figure 130 Input Parameters for Example 38	230
Figure 131 Example 38 PReMod Modules Summary (report.html)	231
Figure 132 Input Parameters for Example 39	234
Figure 133 Example 39 Results Table (partial).....	236
Figure 134 Input Parameters for Example 40	237
Figure 135 JPH2 rs6031444, TLR9 rs5733836, and KMO rs28431981	239
Figure 136 JPH2 rs6031444 G>T in the UCSF Browser (zoom in).....	241
Figure 137 JPH2 rs6031444 G>T in the UCSF Browser (zoom out).....	242
Figure 138 TLR9 rs5743836 T>C May Create An Allele-specific NF-kB Binding Site ..	244
Figure 139 TLR9 rs5743836 T>C in the UCSF Browser.....	244
Figure 140 KMO rs28431981 A>G in the UCSF Browser.....	246
Figure 141 EMSA for JPH2 rs6031444 G>T and TLR9 rs5743836 T>C.....	247
Figure 142 Human-specific and Chimpanzee-specific Delta-MATCH Predictions.....	250
Figure 143 Location of HAR152 predicts the human allele will recruit PAX6.....	252
Figure 144 Neurogenin-2 in UCSC Browser	252
Figure 145 A Lower Baseline Level of HIV Viremia Is Predictive of Longer Survival....	258
Figure 146 NF-kB May Enhance HIV Retroviral Gene Expression and Replication	258
Figure 147 HIV-1 CVL Classification Scheme	262
Figure 148 Statistical Tests (chi-squared).....	264
Figure 149 CCR5 (rs333 ins32>del32)	290
Figure 150 Genotyping the CCR5 del32 (rs333) Polymorphism	291
Figure 151 Dendritic cells respond through TLR3/7/8/9 [70].....	296
Figure 152 Toll-like Receptor Signaling [54]	297
Figure 153 TLR9 (rs352140 G>A, rs352139G>A, rs5743836 T>C, rs187084 T>C).....	298
Figure 154 TLR9A/B/C Transcripts Have Variable Signaling Activity.....	299
Figure 155 RFLP Agarose Gel Photos for Four TLR9 SNPs.....	300
Figure 156 Sequencing Chromatograms of Four TLR9 SNPs	300
Figure 157 Resequencing the TLR9 Locus (8,000 bp).....	301
Figure 158 TLR9 rs352139 G>A and rs352140 A>G Associated with Higher HIV Viremia in White Americans	308
Figure 159 TLR9 Haplotype 1 Associated with Higher Viremia in White Americans	308
Figure 160 Linkage Disequilibrium (D') for Four TLR9 SNPs and One CCR5 In/Del in African American Test1	309
Figure 161 Linkage Disequilibrium (R-squared) for Four TLR9 SNPs and One CCR5 In/Del in African American Test1	309
Figure 162 Linkage Disequilibrium (D') for Four TLR9 SNPs and One CCR5 In/Del in White American Test1	310
Figure 163 Linkage Disequilibrium (R-squared) for Four TLR9 SNPs and One CCR5 In/Del in White American Test1	310
Figure 164 Linkage Disequilibrium (D') for Four TLR9 SNPs and One CCR5 In/Del in His/Lat American Test1	311
Figure 165 Linkage Disequilibrium (R-squared) for Four TLR9 SNPs and One CCR5 In/Del in His/Lat American Test1	311
Figure 166 IRF5 (rs2004640 T>G, rs2070197 T>C, rs10954213 A>G, rs2280714 T>C)	314
Figure 167 IRF5 mRNA variant shown in the UCSC Genome Browser	315
Figure 168 IRF5 Sequencher alignment of 11 mRNA variants (part 1)	316
Figure 169 IRF5 Sequencher alignment of 11 mRNA variants (part 3)	316

Figure 170 IRF5 Sequencher alignment of 11 mRNA variants (part 3)	316
Figure 171 Alignment of five human IRF5 transcripts with one mouse and one cow transcript.	316
Figure 172 IRF5 Sequencher alignment of mRNA variant 1	317
Figure 173 IRF5 Sequencher alignment of mRNA variant 2	317
Figure 174 IRF5 Sequencher alignment of mRNA variant 3	318
Figure 175 IRF5 Sequencher alignment of mRNA variant 4	318
Figure 176 IRF5 Sequencher alignment of mRNA variant 5	318
Figure 177 IRF5 Sequencher alignment of mRNA variant 6	319
Figure 178 IRF5 Sequencher alignment of mRNA variant 7	319
Figure 179 IRF5 Sequencher alignment of mRNA variant 8	319
Figure 180 IRF5 Sequencher alignment of mRNA variant 9	320
Figure 181 IRF5 Sequencher alignment of mRNA variant 10	320
Figure 182 IRF5 Sequencher alignment of mRNA variant 11	320
Figure 183 Sequencher alignment legend	321
Figure 184 Genotyping Four IRF5 SNPs with Taqman Assays-on-Demand	321
Figure 185 IRF5 Is a Critical Switch Regulating Inflammation and Autoimmunity and is Associated with Lupus.....	322
Figure 186 IRF5 Haplotypes in White Americans	328
Figure 187 Linkage Disequilibrium (D') for Four IRF5 SNPs in African American Test1	329
Figure 188 Linkage Disequilibrium (R -squared) for Four IRF5 SNPs in African American Test1.....	329
Figure 189 Linkage Disequilibrium (D') for Four IRF5 SNPs in White American Test1.....	330
Figure 190 Linkage Disequilibrium (R -squared) for Four IRF5 SNPs in African American Test1.....	330
Figure 191 Linkage Disequilibrium (D') for Four IRF5 SNPs in His/Lat American Test1	331
Figure 192 Linkage Disequilibrium (R -squared) for Four IRF5 SNPs in His/Lat American Test1.....	331
Figure 193 APOE (rs429358 T>C, rs7412 C>T).....	333
Figure 194 Genotyping APOE (E2, E3 and E4).....	334
Figure 195 Linkage Disequilibrium (D') for Two APOE SNPs in African American Test1	339
Figure 196 Linkage Disequilibrium (R -squared) for Two APOE SNPs in African American Test1.....	339
Figure 197 Linkage Disequilibrium (D') for Two APOE SNPs in White American Test1	340
Figure 198 Linkage Disequilibrium (R -squared) for Two APOE SNPs in White American Test1.....	340
Figure 199 Linkage Disequilibrium (D') for Two APOE SNPs in His/Lat American Test1	341
Figure 200 Linkage Disequilibrium (R -squared) for Two APOE SNPs in His/Lat American Test1.....	341
Figure 201 The BIOBASE MATCH Program Version 10.2.....	401
Figure 202 How to Calculate a MATCH Score [1].....	402
Figure 203 MATCH Score Calculation [1].....	403
Figure 204 Architectural Diagram for the Delta-MATCH Query Tool (DMQT)	407
Figure 205 Delta-MATCH Error 1 - no matrixes passed your selected criteria	409
Figure 206 Delta-MATCH Error 2- more than 1,500 rnumbers passed your selected criteria	409

<i>Figure 207 Delta-MATCH Error 3 - no rsnumbers were found that passed your selected criteria</i>	<i>410</i>
<i>Figure 208 Delta-MATCH Error 4- no rsnumbers were found in the select gene names</i>	<i>411</i>
<i>Figure 209 Delta-MATCH Error 5 - could not connect to database</i>	<i>411</i>
<i>Figure 210 Delta-MATCH Error 6 - more than 5 gene names were submitted</i>	<i>412</i>
<i>Figure 211 Delta-MATCH Error 7 - no gene names were found</i>	<i>412</i>
<i>Figure 212 Delta-MATCH Error 8 - rsnumber file was not uploaded properly</i>	<i>413</i>
<i>Figure 213 Delta-MATCH Error 9 - no premod modules were found.....</i>	<i>413</i>
<i>Figure 214 Delta-MATCH Graphic Motif</i>	<i>414</i>
<i>Figure 215 Delta-MATCH Resources (Graphics).....</i>	<i>414</i>
<i>Figure 216 Haploview Linkage Disequilibrium Legend.....</i>	<i>415</i>
<i>Figure 217 The DNA Degenerate Alphabet</i>	<i>415</i>
<i>Figure 218 David W. Williamson's Contact and Business Card</i>	<i>417</i>
<i>Figure 219 Joseph "Mike" McCune, Bruce Conklin, David Williamson, Robert Mahley</i>	<i>419</i>

List of Equations

Equation 1 - Expected Frequency of Homozygous Carriers of Allele 1 ($Freq_{A1/A1}$)	20
Equation 2 - Expected Frequency of Heterozygous Carriers ($Freq_{A1/A2}$).....	20
Equation 3 - Expected Frequency of Homozygous Carriers of Allele 2 ($Freq_{A2/A2}$)	20
Equation 4 - mean MATCH score.....	37
Equation 5 - larger polymorphism MATCH score (m_{max}).....	37
Equation 6 - smaller polymorphism MATCH score (m_{min})	37
Equation 7 - absolute difference in MATCH score (m_{dif})	37
Equation 8 - Delta-MATCH potential score (potential)	38
Equation 9 - absolute percent difference in MATCH score (m_{per})	38
Equation 10 - Delta-MATCH potential score (potential)	38
Equation 11 - Slope of Linear Estimation Curve (slope).....	47
Equation 12 - Biological Relevance of a Polymorphic Site (brps).....	54
Equation 13 - Rareness of a Hit ($HIT_{rareness}$).....	64
Equation 14 - Number of Calculations on the Plus Strand ($Number_{plus}$)	70
Equation 15 - Number of Calculations on the Plus Strand ($Number_{minus}$)	70
Equation 16 - Number of Calculations Required to Find Highest Match ($Number_{total}$).....	70
Equation 17 - Number of MATCH scores calculated.....	74
Equation 18 - Number of highest MATCH scores recorded into the Delta-MATCH database	74
Equation 19 - rareness of a potential score (rareness)	171
Equation 20 - Bonferonni-adjusted rareness (bonferonni).....	188

Introduction

The overall goal of this thesis is to identify polymorphic alleles that associate with elevated risk and disease progression. To achieve this goal I've used two different approaches. In AIM 1 I'll present a predictive approach, describing a resource I've developed called Delta-MATCH. Delta-MATCH helps to identify and predict which human single nucleotide polymorphisms (SNPs) are likely to create allele-specific transcription factor binding sites, and is an example of a prospective computational survey. In AIM 2 I'll present the results of a classical genetic survey, using a candidate gene approach, to investigate the polymorphisms in four candidate genes, for their association with the phenotype of HIV-1 viremia.

Human polymorphic variation may contribute to pathogenesis and disease phenotypes by modulating molecular processes such as gene transcription, mRNA processing, and protein modification, structure and function. Although much of the public effort has picked the lowest hanging fruit by identifying polymorphisms that code for dramatic nonsynonymous amino acid substitutions, it may be important to identify the genetic modulators of gene transcription, and those variants that modulate the magnitude of inflammatory responses [2].

A subset of human polymorphisms may modulate transcription factor binding affinity and gene transcription by altering with consensus sequence of a transcription factor binding site at a position in the promoter if a gene and proximal to its transcriptional start site. Resources like [rVISTA 2.0](#) have attempted to map the genome-wide distribution of transcription factor binding sites by using pattern matching approaches and comparative genomics to identify conserved non-coding regulatory regions [3-5]. Although some portal sites allow users to query public databases to identify polymorphisms positioned

within these regulatory regions [6-9], existing tools don't have robust ranking methods that incorporate orthogonal data types that allow users to identify and predict which genetic variants will associate with human diseases [5, 10, 11].

I have constructed a novel resource and database called Delta-MATCH that predicts if a polymorphism may promote an allele-specific transcription factor binding recruitment event. This tool was developed as an extension of an existing tool called the [MATCH](#) program, which predicts quantitatively how well a transcription factor will bind to a given genetic sequence [1]. MATCH scores were calculated for 4.5 million pairs of SNP alleles, and ranked by their importance.

The Delta-MATCH database has been used to identify lists of candidate SNPs that are being investigated for their association to a number of disease phenotypes including autoimmunity, multiple sclerosis, dyslipidemia, hypertriglyceridemia, Alzheimer's disease and HIV/AIDS progression. Furthermore, because homologous genomic sequences of two distinct organisms can be aligned, Delta-MATCH has been used to identify and predict species-specific transcription factor binding sites. Specifically, when the human and chimpanzee genomes were aligned, a relative polymorphism in the Neurogenin-2 gene was identified that may create a PAX-6 transcription factor binding site in chimpanzees (and non-human vertebrates), but not humans (page 249).

Strong associations between polymorphic variants and disease phenotypes have been identified, and may be important to identify those that contribute to the pathologies of multiple disorders. The apolipoprotein E ([APOE](#)) epsilon 4 ($\epsilon 4$) allele, for example, is associated with increased risk of cardiovascular disease (CVD), Alzheimer's disease ([AD](#)), and HIV-related dementia [12-14]. Transgenic mice expressing the human $\epsilon 4$

protein are used as a model of AD [15], and mice deficient in apoE have elevated lipid levels, and are used as a proinflammatory model for studying atherosclerosis [15, 16]. Other strong associations have been identified between variants of interferon regulatory factor 5 ([IRF5](#)) and risk of developing systemic erythematosus lupus ([SLE](#)) [17-19], and between a haplotype of Toll-like receptor 9 ([TLR9](#)) and the rate of CD4+ T cell loss during HIV-1 infection [20].

However, not all polymorphic variants are associated with detrimental phenotypes. Indeed, some variants protect against viral and bacterial infection [21, 22]. For example, the del32 allele of chemokine receptor 5 ([CCR5](#)), a seven transmembrane protein expressed by T cells and macrophages, and coreceptor for the human immunodeficiency virus type 1 ([HIV-1](#)), confers protection against [HIV-1](#) infection.

Because polymorphisms in [CCR5](#), [TLR9](#), [IRF5](#), and [APOE](#) have been associated with multiple phenotypes that are mediated by inflammation, and because HIV/AIDS infection is modulated by the innate inflammatory response, the common genetic variants in these four genes were investigated for their association to HIV/AIDS viremia levels as a surrogate marker of risk of disease progression.

***1.3* AIM 1: Delta-MATCH: A Computational Survey**

To conduct a computational survey of the database of human single nucleotide polymorphisms (SNPs) to identify and rank prioritize polymorphisms that may associate with allele-specific transcription factor (TF) recruitment. Transcription factor binding site (TFBS) matrixes provided by the [BIOBASE TRANSFAC](#) database were pattern matched

against human genome sequence to derive quantitative scores reflecting allele-specific transcription factor binding affinity. Delta-MATCH is a web-based tool (<http://deltamatch.org>) providing the scientific community the ability to identify lists of high-value candidate SNPs based on a number of independent selectable criteria. These candidate SNPs may modulate transcription factor binding and associate with both allele-specific gene expression and phenotypic disease.

1.4 AIM 2: A Genetic Survey of Genetic Modulators of HIV-1 Viremia

To conduct a genetic survey of four genes ([CCR5](#), [TLR9](#), [IRF5](#), [APOE](#)) to identify associations between genotype/haplotype frequencies and [HIV-1](#) viremia levels, in a population of [HIV-1](#)-infected Americans primarily derived from the San Francisco SCOPE cohort. The HIV-1 cohort collection was coordinated by Mike McCune at the Gladstone Institute of Virology and Immunology ([GIVI](#)), and by Stephen Deeks and Jeff Martin at the San Francisco General Hospital.

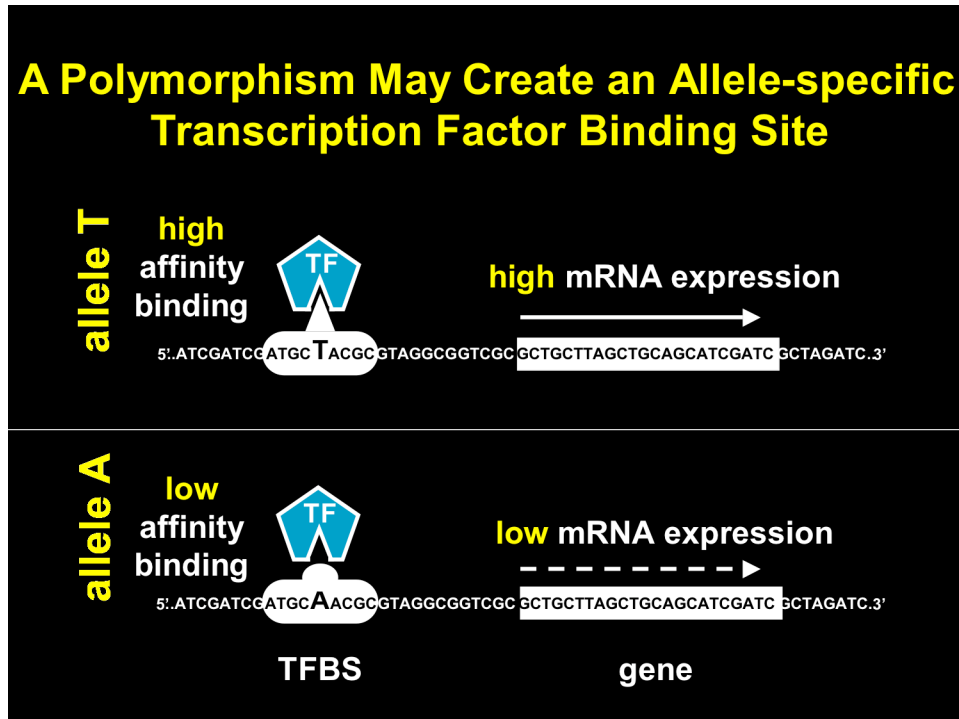
Chapter 1: Delta-MATCH - A Computational Survey

1.5 Delta-MATCH Overview

1.5.1 The Aim of the Delta-MATCH Program

The aim of the Delta-MATCH program is to identify those human polymorphic variants differ greatly in their predicted transcription factor binding affinity (difference in MATCH score = Δ - MATCH). In this version of the Delta-MATCH database (version 1.0), 4,547,844 high value candidate polymorphisms have been scored and ranked by the Delta-MATCH algorithm to determine their “potential” to create an allele-specific transcription factor binding site. These high-value polymorphisms were either positioned within a 10,000 base pair window (10k upstream + gene + 10k downstream) of any [refSeq](#) gene (UCSC browser table hg18.refGene.name2), or positioned within a region of high conservation in the human genome (UCSC browser hg18. phastCons17way) (Table 41 page 406). The major and minor alleles for each of these 4.5 million polymorphisms were independently evaluated by the MATCH algorithm [1] (Figure 201) against a library of 550 known transcription factor binding site sequences (Table 43, page 405) to determine the “highest MATCH scores” for each allele and transcription factor pair. A ranked list of polymorphisms was then determined for each of the 550 transcription factor binding sites (matrix names) and catalogued in the Delta-MATCH database. These polymorphisms have been ranked by a statistic called the “Delta-MATCH potential score”, which reflects the “potential” of a polymorphism to create an allele-specific transcription factor binding site (page 30).

Figure 1 A Polymorphism May Create an Allele-specific Transcription Factor Binding Site



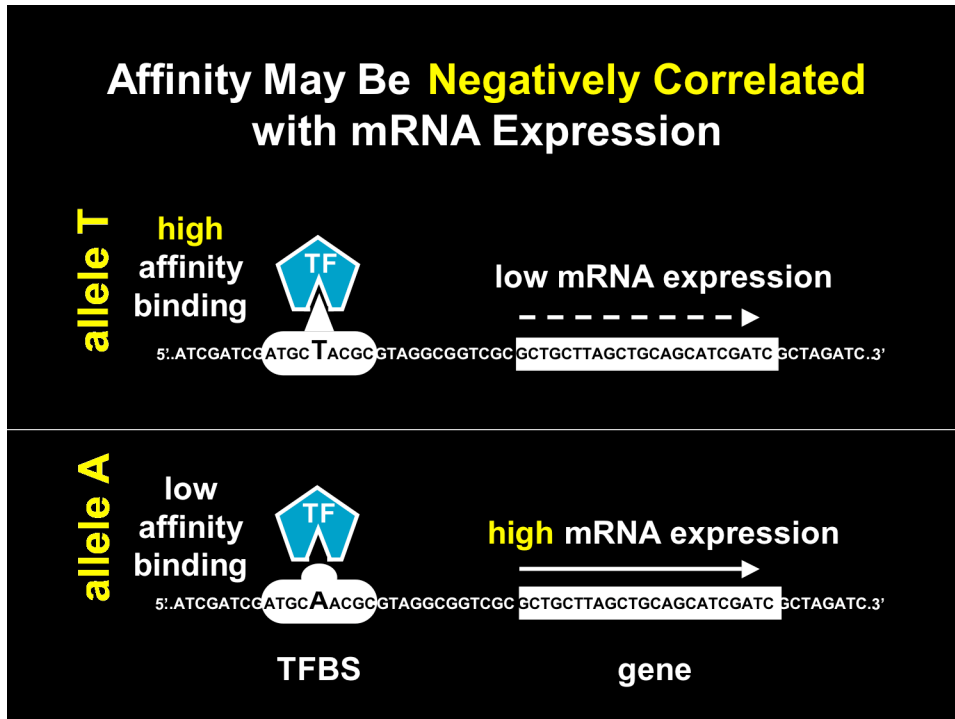
An allele-specific transcription factor binding site (TFBS) is created when a polymorphism is positioned in a regulatory element proximal to a given gene and results in the two alleles having very different affinities for the relevant transcription factor (TF). Figure 1 shows a transcription factor in blue, which has a high affinity for a nucleotide sequence (allele T), upstream of gene that transcribes high levels of mRNA. In this example, there is strong specificity, between the transcription factor and the transcription factor binding site, and this correlates with high gene expression. However, if for example, a mutation is induced that converts the T allele, to an A allele, the A allele creates a lower affinity binding site for the transcription factor. In this instance the difference in TF binding affinity, between the two alleles and the transcription factor, correlates with a difference mRNA expression.

1.5.2 Transcription Factor Binding Affinity May Be Correlated with the Level of mRNA Expression and Associated with Some Human Diseases (Δ binding affinity $\approx \Delta$ expression)

When a polymorphism creates an allele-specific transcription factor binding site, it may cause human disease by causing a dysregulation in transcription factor binding and mRNA transcription. Therefore, it is important to try to identify those polymorphisms that have strong differences in binding affinity between their alleles and a TF.

Two cases may be considered. In case 1 (Figure 2), binding affinity may be positively correlated with mRNA expression, when high affinity binding correlates to high mRNA expression, and low affinity binding correlates to low mRNA expression. In case 2 (Figure 3), binding affinity may be negatively correlated with mRNA expression, when high affinity binding correlates to low mRNA expression, and low affinity binding correlates to high mRNA expression.

Figure 3 Transcription Factor Binding Affinity May Negatively Correlate with Level of mRNA Expression



Although there is currently no direct way to know, if a transcription factor will positively or negatively correlate to expression, what can be said is, that a strong differences in transcription factor binding, may correlate to differences in gene expression and human disease phenotypes.

1.5.3 The Delta-MATCH Hypothesis

The Delta-MATCH method hypothesizes a difference in mRNA expression may be correlated a difference in transcription factor binding affinity between a pair of polymorphic alleles, and these differences may be associated with allele-specific gene expression and some disease phenotypes.

1.6 Computational Survey

Exactly 4,547,844 priority SNPs were computationally surveyed and prioritized by using the *delta_match.py* and *prioritize_results.py* python scripts and all scores, and associated data were accumulated into a single Delta-MATCH databases (DBDM).

1.6.1 What is Delta-MATCH Query Tool?

The Delta-MATCH Query tool is a web-based tool used to identify and rank polymorphisms in the database of human single nucleotide polymorphisms ([dbSNP](#) rsnumbers) for their potential to create allele-specific transcription factor binding sites (TFBS). It should be noted the results from this tool are returned in descending order of their ranked importance.

It is predicted that in most cases, a polymorphism is unlikely to create an allele-specific binding site. However, in the most extreme case, the nucleotide sequence of one allele may be determined to be perfect match to the sequence of a known transcription factor binding site, while the other allele is determined to be a complete “mis-match” binding site. These may be cases where a transcription factor may bind to only one of the two polymorphic allele sequences, and may be cases where the two alleles associate with variable gene expression and disease phenotypes.

1.6.2 The Delta-MATCH Query Tool (DMQT) Website Address

<http://deltamatch.org/>

1.6.3 DMQT Overview

A Delta-MATCH query tool (DMQT) was built to query the ranked results of the Delta-MATCH database (DBDM). Starting with the entire list of more than 4.5 million human polymorphisms, the query tool functions as a filtering engine, returning only those results that meet a list of criteria. A diverse number of complexly layered queries can be formulated simply by manipulating a series of radio buttons, check boxes, and drop-down menus before submitting the structured query language (SQL) request and awaiting the resultant pages. It is possible to search the DBDM by SNP accession numbers (rsnumbers), gene names, chromosomal positions, transcription factor binding site (TFBS) matrix names, transcription factor (TF) names, and/or by tissue types (where a list of associated transcription factors is known to be highly expressed) (Table 39 page 404). SNPs can be identified (for example) that have a minimum potential score (potential), or a minimum frequency of heterozygosity, or are located near genes associated with a specified gene ontology (GO) term, within a particular genomic region (Table 44 page 406), within a specified distance of transcriptional or translational start site, within a region of high conservation, and/or within a region of high TFBS density ([PReMod](#)). When filtering for SNPs present on [Affymetrix](#) or [Illumina](#) genotyping chips, the DMQT will optionally return additional significant results for all SNPs known to be in strong linkage disequilibrium (LD) with SNPs on the genotyping chips as identified by cross-referencing LD tables in the [HapMap](#) database. Because LD values are ethnicity-specific, the DMQT includes LD values calculated for four [HapMap](#) populations; European (CEU), African (YRI), Japanese (JPT), and Chinese (CHB) [23]. Results may further be filtered to include only results from TFBS matrixes that are of 'high' quality or of a minimum matrix length, and the resultant webpage may be sorted by a number of different methods for best viewing.

The DMQT has been designed for flexibility. Each button, box, and drop-down menu acts as an additional independent query layer, and it should therefore be possible to remodel and expand its capability to incorporate and cross-reference additional bioinformatics resources in the future with minimal effort. Examples of the resultant web pages are shown (Figure 50 page 110).

1.6.4 Building a Workstation

To construct an informatics workstation, I assembled a machine that has two 2.3-Ghz processors, 8 Gb of memory, and more than two terabytes of storage and running the gen[24]machine I installed a [mirrored](#) copy of the [UCSC genome browser](#) complete with all genomic MySQL data tables and an academic licensed copy of the [Transcription Factor](#) database ([TRANSFAC](#) version 10.2), distributed by [BIOBASE](#) [3]. Exactly 584 vertebrate Transcription Factor Binding Site (TFBS) matrixes (two-dimensional Position Specific Scoring Matrixes representing transcription factor-specific nucleotide binding site sequence) were co-opted from the [MATCH](#) program [1]. Additional resources ([HapMap](#), [Affymetrix](#), [Illumina](#), [PReMod](#), [Gene Ontology](#)) were adapted and installed as accessory databases [23, 25, 26].

1.6.5 Computing Time

The Delta-MATCH script computed at a maximum of 18 polymorphisms per minute on the Linux workstation. Surveying the list of 4,547,844 priority SNPs was calculated to require a minimum of 175 CPU days. Therefore, the bulk of the workload was distributed over 8 additional computers (G5 Macintosh OS 10.4), each running independent lists of

SNPs. These nodes computed autonomously except for the genomic sequence retrieval step, which was accomplished by using the [DAS sequence retrieval web server](#) function on the Linux machine. After network interruptions and administrative solutions, I estimated that more than 5 weeks of 24-hour computation on 9 machines was needed to survey the 4.5 million priority SNPs using the *delta_match.py* script. Subsequently, the *prioritize_results.py* script took 7 days to completely rank and prioritize the results for 550 separate TFBS matrixes.

1.6.6 How the Delta-MATCH Query Tool Was Constructed

Transcription factor binding site matrixes (mat_id) provided by the [BIOBASE TRANSFAC](#) [27, 28] database were pattern matched against human genome sequence using the [MATCH](#) algorithm [1] to derive quantitative scores (potential scores) reflecting allele-specific transcription factor binding affinity. The Delta-MATCH Query Tool integrates data from many external bioinformatics databases ([UCSC human genome browser](#), [HapMap](#), [Affymetrix](#), [Illumina](#), [PReMod](#), [Gene Ontology](#), [Database of Genomic Variants](#), [Database of HIV-1 Candidate Genes](#), [Database of Alzheimer's Disease Candidate Genes](#), [Database of HIV-1 Cohorts](#)) and may be used to produce a filtered list of high-value candidate SNP targets that may be associated with allele-specific transcription factor binding events.

1.6.7 SNP Identification and Selection

All 11,647,909 distinct polymorphisms from the UCSC March 2006 human genome database (UCSC browser table hg18.snp126) were classified as belonging to one or more genomic positions relative to all 'knownGene' (UCSC browser table

hg18.knownGene) reference sequences (Table 44 page 406). Polymorphisms located in the following positions were identified and prioritized for computation:

- within the 10k upstream sequence flanking a knownGene sequence
- within the 10k downstream sequence flanking a knownGene sequence
- within an the 5'UTR of a knownGene sequence
- within an the 3'UTR of a knownGene sequence
- within an knownGene sequence exon
- within an knownGene sequence intron
- within a region of strong conservation
- within a cpgisland
- within a region with high regulatory potential

A prioritized list 4,547,844 biallelic single nucleotide polymorphisms (SNPs) was constructed for primary analysis and included all SNPs positioned within 10 kb of any 'knownGene' sequence, within a region of strong conservation (UCSC browser table hg18. phastconsElements17way), or present on a known human [Affymetrix](#), or [Illumina](#) genotyping SNPchip.

Polymorphisms positioned within regions of insertion/deletions, simple repeats, or microsatellites were excluded. Also excluded were polymorphisms with more than two allele states and those that mapped to more than one chromosomal position.

1.6.8 What is a Transcription Factor Binding Site Matrix?

The Delta-MATCH algorithm uses the scoring method and the many of the transcription factor binding site matrixes (n = 550) originally derived from the BIOBASE [MATCH](#) program [1].

1.6.8.1 Definition - matrix (transcription factor binding site matrix)

A transcription factor binding matrix is a two-dimensional mathematical representation of what a transcription factor binding site looks like in nucleotide sequence space.

For each base position of a matrix, there are weights attributed for each the four possible deoxy-ribonucleic-nucleotide bases (A, C, G, T) that reflect the specificity of a given base at each position. The weights of the matrix are lowest when the nucleotide diversity at that position is highest (very little specificity), and are highest when the nucleotide diversity at that position is lowest (very high specificity). The weight values for the 550 TRANSFAC matrixes were created by:

- aligning the promoters of genes known to be responsive to a known transcription factor
- identifying small, but conserved motifs in these gene promoters (the transcription factor binding site sequence)
- summing up the number of times each of the four (A,C,G,T) bases are present in at each of the transcription factor binding site positions into a 2-dimensional position-specific base counting matrix (base count versus base position)
- converting the resulting scores into a percentage count matrix (normalize to 100)
(Figure page 16)

- multiplying the 2-dimensional position-specific base counting matrix by an information eigenvector that represents the nucleotide diversity at each position of the matrix (Figure page 17) [1].

1.6.9 Transcription Factor Binding Site Matrix (percentage count)

Figure 4 displays an example transcription factor binding site matrix prior to the multiplication of the information eigenvector. This matrix represents the base specificity of 6 base positions which have been normalized to 100 percent (a percentage count matrix). [base position specificity: ((3=4) > 5 > 2 > 1 > 6)]

Figure 4 Transcription Factor Binding Site Matrix

base \ position	1	2	3	4	5	6
A	0	0	100	0	0	25
C	50	75	0	0	25	25
G	50	10	0	0	0	25
T	0	15	0	100	75	25
consensus	S	C	A	T	T	N
specificity	**	***	*****	*****	****	*
eigenvector	530	526	599	599	543	461

1.6.10 Transcription Factor Binding Site Matrix (after eigenvector multiplication)

The matrixes use by Delta-MATCH (and the BIOBASE MATCH program) have been multiplied through by an information eigenvector to attribute more importance to the most informative TFBS base positions. Figure 5 displays an example TFBS matrix that

has been multiplied through by an information eigenvector and normalized to 100 percent. Notice the newly adjusted weights for the least specific base positions are relatively lower than their corresponding weights before the eigenvector multiplication.

Figure 5 Transcription Factor Binding Site Matrix After Eigenvalue Correction

base \ position	1	2	3	4	5	6
A	0	0	100	0	0	19.2
C	44.2	71.4	0	0	19.2	19.2
G	44.2	6.2	0	0	0	19.2
T	0	10.3	0	100	71.4	19.2

1.6.10.1 Definition - information eigenvector

This is a weighted vector the same length as a matrix that describes the nucleotide diversity across every base position and is an estimator of base specificity.

1.6.11 What is a MATCH Score?

1.6.11.1 Definition - MATCH score

This is a statistic that reflects the sequence identity between a given DNA sequence and a transcription factor **matrix**.

The mathematical definitions of the MATCH score and eigenvector definition are described in the original MATCH publication [1] (Figure page 402).

1.6.12 Three Genetic Models of Human Disease Paired with High or Low Levels of mRNA Transcription

Delta-MATCH has been created to help to identify human polymorphisms that associate (and potentially cause) allele-specific gene transcription and human disease. In Figure 2 and Figure 3, two SNP alleles induce transcription factor binding with different levels of affinity.

When the normal mRNA transcription is dependent on high affinity TF binding, it might be the case that an allele that creates a low affinity binding site might be associated with phenotypes caused by lower levels of gene transcription.

1.6.12.1 Definition - phenotype case 1 (low mRNA transcription = disease)

In this case, high mRNA transcription is associated with a normal state, and low mRNA transcription is associated with a diseased state.

When a transcription factor binds to a promoter with high affinity and acts as a suppressor, it might be the case that normal low levels mRNA transcription is dependent on high affinity TF binding, and that an allele that creates a low affinity binding site might be associated with phenotypes caused by higher levels of gene transcription.

1.6.12.2 Definition - phenotype case 2 (high mRNA transcription = disease)

In this case, low mRNA transcription is associated with a normal state, and high mRNA transcription is associated with a diseased state.

There are three genetic models that may be paired with two above two phenotype cases when associating genetic markers with a given phenotype.

1.6.12.3 Definition - genetic model 1 - (dominant model)

The disease state allele is dominant over the normal state allele. In this model a single copy of the disease allele is sufficient to create the disease phenotype. The mRNA transcription levels may be either high or low.

1.6.12.4 Definition - genetic model 2 - (co-dominant model)

The disease state allele is co-dominant with the normal state allele. In this model, there is a dose-dependence correlation of the disease allele with the level of mRNA transcription. The mRNA transcription levels may be high, medium or low.

1.6.12.5 Definition - genetic model 3 - (recessive model)

The disease is recessive to the normal state allele. In this model two copies of the disease allele is required to create the disease phenotype. The mRNA transcription levels may be either high or low.

1.6.13 Hardy-Weinberg Expectation Equations

1.6.13.1 Definition - Hardy-Weinberg Expectation (HWE)

The sum of the allele frequencies for any given pair of alleles (A1 and A2) must equal one, and the sum of all homozygous and heterozygous genotype frequencies must equal one.

$$(A1 \text{ Freq}) + (A2 \text{ Freq}) = 1$$

and

$$\text{Freq} (A1 / A1) + \text{Freq} (A1 / A2) + \text{Freq} (A2 / A2) = 1$$

Equation 1 - Expected Frequency of Homozygous Carriers of Allele 1 ($\text{Freq}_{A1/A1}$)

$$\text{Freq}_{A1/A1} (A1 / A1) = (A1 \text{ Freq}) * (A1 \text{ Freq})$$

Equation 2 - Expected Frequency of Heterozygous Carriers ($\text{Freq}_{A1/A2}$)

$$\text{Freq}_{A1/A2} (A1 / A2) = 2 * (A1 \text{ Freq}) * (A2 \text{ Freq})$$

Equation 3 - Expected Frequency of Homozygous Carriers of Allele 2 ($\text{Freq}_{A2/A2}$)

$$\text{Freq}_{A2/A2} (A2 / A2) = (A2 \text{ Freq}) * (A2 \text{ Freq})$$

1.6.14 The Predicted Genotype Frequencies of Three Genetic Models Paired with High or Low Levels of mRNA Transcription

For each of the Phenotype Case / Genetic Model Pairs, it is possible to predict the expected genotype frequencies [Freq(A1/A1), Freq(A1/A2), Freq(A2/A2)] for any pair of allele frequencies (A1 Freq, A2 Freq) using the Hardy-Weinberg expectation equations (page 20).

Table 1 Predicted Genotype Frequencies of Three Genetic Models

			A1 Freq	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
			A2 Freq	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0
Case / Model	Genotype	Phenotype												
1 Case 1 / Dominant	A1 / A1	high	Freq (A1 / A1)	0.00	0.01	0.04	0.09	0.16	0.25	0.36	0.49	0.64	0.81	1.00
A1 = high = normal	A1 / A2	low	Freq (A1 / A2)	0.00	0.18	0.32	0.42	0.48	0.50	0.48	0.42	0.32	0.18	0.00
A2 = low = disease	A2 / A2	low	Freq (A2 / A2)	1.00	0.81	0.64	0.49	0.36	0.25	0.16	0.09	0.04	0.01	0.00
2 Case 1 / Co-Dominant	A1 / A1	high	Freq (A1 / A1)	0.00	0.01	0.04	0.09	0.16	0.25	0.36	0.49	0.64	0.81	1.00
A1 = high = normal	A1 / A2	medium	Freq (A1 / A2)	0.00	0.18	0.32	0.42	0.48	0.50	0.48	0.42	0.32	0.18	0.00
A2 = low = disease	A2 / A2	low	Freq (A2 / A2)	1.00	0.81	0.64	0.49	0.36	0.25	0.16	0.09	0.04	0.01	0.00
3 Case 1 / Recessive	A1 / A1	high	Freq (A1 / A1)	0.00	0.01	0.04	0.09	0.16	0.25	0.36	0.49	0.64	0.81	1.00
A1 = high = normal	A1 / A2	high	Freq (A1 / A2)	0.00	0.18	0.32	0.42	0.48	0.50	0.48	0.42	0.32	0.18	0.00
A2 = low = disease	A2 / A2	low	Freq (A2 / A2)	1.00	0.81	0.64	0.49	0.36	0.25	0.16	0.09	0.04	0.01	0.00
4 Case 2 / Dominant	A1 / A1	high	Freq (A1 / A1)	0.00	0.01	0.04	0.09	0.16	0.25	0.36	0.49	0.64	0.81	1.00
A1 = high = disease	A1 / A2	high	Freq (A1 / A2)	0.00	0.18	0.32	0.42	0.48	0.50	0.48	0.42	0.32	0.18	0.00
A2 = low = normal	A2 / A2	low	Freq (A2 / A2)	1.00	0.81	0.64	0.49	0.36	0.25	0.16	0.09	0.04	0.01	0.00
5 Case 2 / Co-Dominant	A1 / A1	high	Freq (A1 / A1)	0.00	0.01	0.04	0.09	0.16	0.25	0.36	0.49	0.64	0.81	1.00
A1 = high = disease	A1 / A2	medium	Freq (A1 / A2)	0.00	0.18	0.32	0.42	0.48	0.50	0.48	0.42	0.32	0.18	0.00
A2 = low = normal	A2 / A2	low	Freq (A2 / A2)	1.00	0.81	0.64	0.49	0.36	0.25	0.16	0.09	0.04	0.01	0.00
6 Case 2 / Recessive	A1 / A1	high	Freq (A1 / A1)	0.00	0.01	0.04	0.09	0.16	0.25	0.36	0.49	0.64	0.81	1.00
A1 = high = disease	A1 / A2	low	Freq (A1 / A2)	0.00	0.18	0.32	0.42	0.48	0.50	0.48	0.42	0.32	0.18	0.00
A2 = low = normal	A2 / A2	low	Freq (A2 / A2)	1.00	0.81	0.64	0.49	0.36	0.25	0.16	0.09	0.04	0.01	0.00

1.6.15 The Predicted Phenotype Frequencies of Three Genetic Models Paired with High or Low Levels of mRNA Transcription

For each of the Phenotype Case / Genetic Model Pairs, it is possible to predict the phenotype frequencies [Freq(high), Freq(medium), Freq(low)] for any pair of allele frequencies (A1 Freq, A2 Freq) by summing up the expected genotype frequencies when grouping by identical phenotype (high, medium, and low).

Table 2 Predicted Phenotype Frequencies of Three Genetic Models

			A1 Freq	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	
			A2 Freq	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0	
Case / Model	Genotype	Phenotype													
1 Case 1 / Dominant	A1 / A1	high	freq (high)	0.00	0.01	0.04	0.09	0.16	0.25	0.36	0.49	0.64	0.81	1.00	
	A1 = high = normal	A1 / A2	medium	freq (medium)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
	A2 = low = disease	A2 / A2	low	freq (low)	1.00	0.99	0.96	0.91	0.84	0.75	0.64	0.51	0.36	0.19	0.00
2 Case 1 / Co-Dominant	A1 / A1	high	freq (high)	0.00	0.01	0.04	0.09	0.16	0.25	0.36	0.49	0.64	0.81	1.00	
	A1 = high = normal	A1 / A2	medium	freq (medium)	0.00	0.18	0.32	0.42	0.48	0.50	0.48	0.42	0.32	0.18	0.00
	A2 = low = disease	A2 / A2	low	freq (low)	1.00	0.81	0.64	0.49	0.36	0.25	0.16	0.09	0.04	0.01	0.00
3 Case 1 / Recessive	A1 / A1	high	freq (high)	0.00	0.19	0.36	0.51	0.64	0.75	0.84	0.91	0.96	0.99	1.00	
	A1 = high = normal	A1 / A2	medium	freq (medium)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
	A2 = low = disease	A2 / A2	low	freq (low)	1.00	0.81	0.64	0.49	0.36	0.25	0.16	0.09	0.04	0.01	0.00
4 Case 2 / Dominant	A1 / A1	high	freq (high)	0.00	0.19	0.36	0.51	0.64	0.75	0.84	0.91	0.96	0.99	1.00	
	A1 = high = disease	A1 / A2	medium	freq (medium)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
	A2 = low = normal	A2 / A2	low	freq (low)	1.00	0.81	0.64	0.49	0.36	0.25	0.16	0.09	0.04	0.01	0.00
5 Case 2 / Co-Dominant	A1 / A1	high	freq (high)	0.00	0.01	0.04	0.09	0.16	0.25	0.36	0.49	0.64	0.81	1.00	
	A1 = high = disease	A1 / A2	medium	freq (medium)	0.00	0.18	0.32	0.42	0.48	0.50	0.48	0.42	0.32	0.18	0.00
	A2 = low = normal	A2 / A2	low	freq (low)	1.00	0.81	0.64	0.49	0.36	0.25	0.16	0.09	0.04	0.01	0.00
6 Case 2 / Recessive	A1 / A1	high	freq (high)	0.00	0.01	0.04	0.09	0.16	0.25	0.36	0.49	0.64	0.81	1.00	
	A1 = high = disease	A1 / A2	medium	freq (medium)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
	A2 = low = normal	A2 / A2	low	freq (low)	1.00	0.99	0.96	0.91	0.84	0.75	0.64	0.51	0.36	0.19	0.00

Figure 6 Phenotype Frequencies Case1 /Model Dominant

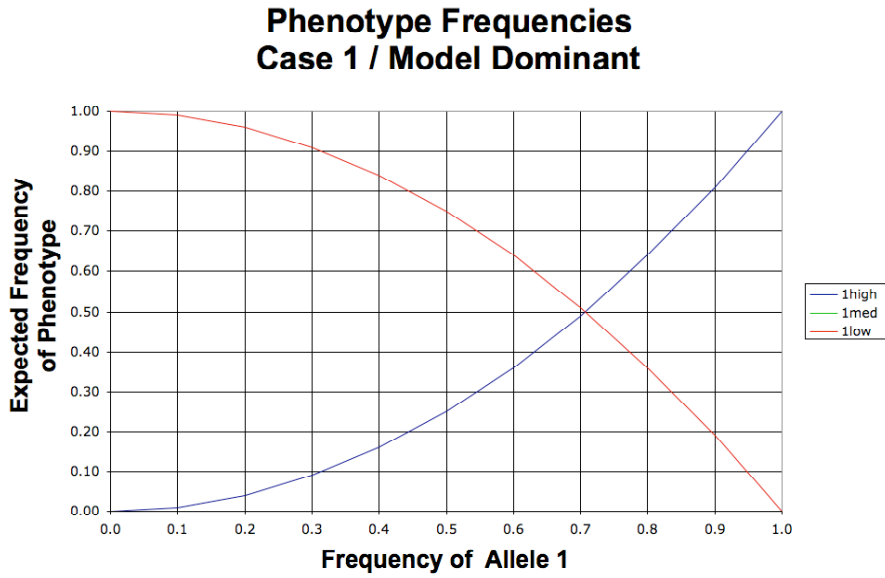


Figure 7 Phenotype Frequencies Case1 /Model Co-Dominant

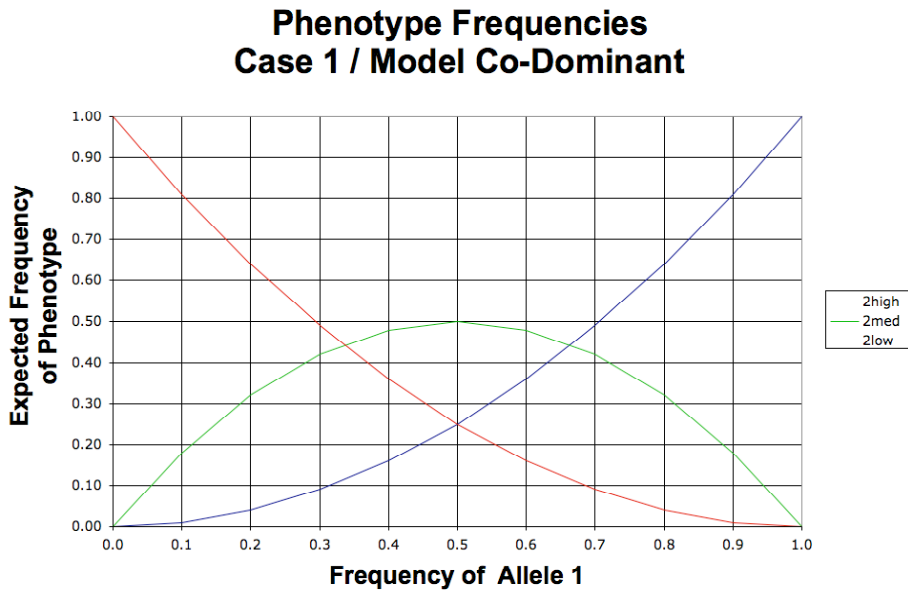


Figure 8 Phenotype Frequencies Case1 /Model Recessive

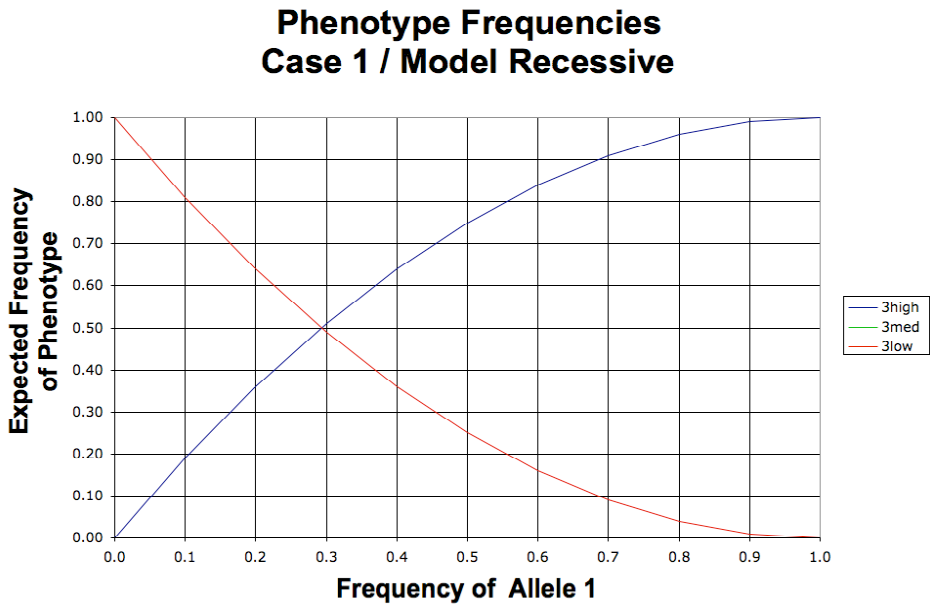


Figure 9 Phenotype Frequencies Case2 /Model Dominant

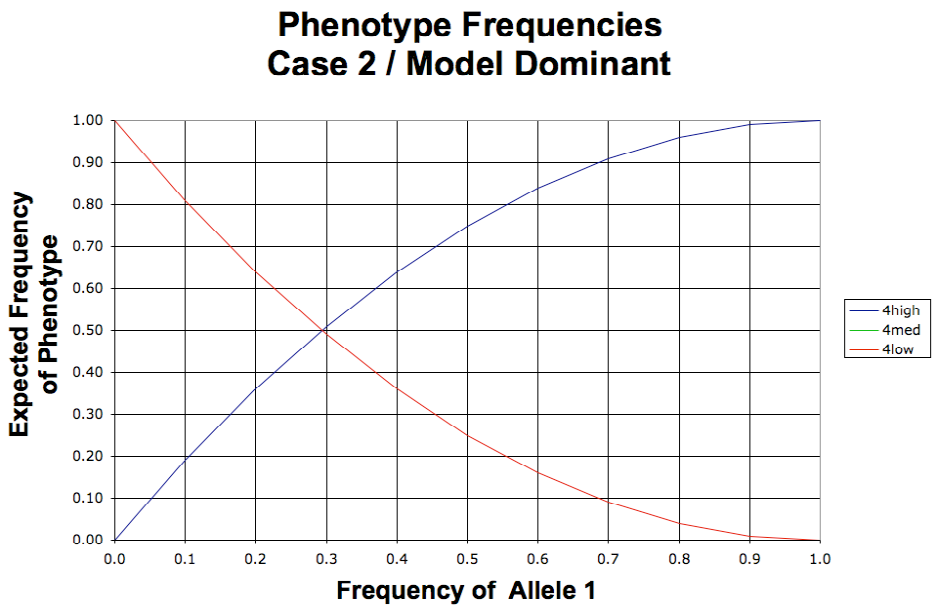


Figure 10 Phenotype Frequencies Case2 /Model Co-Dominant

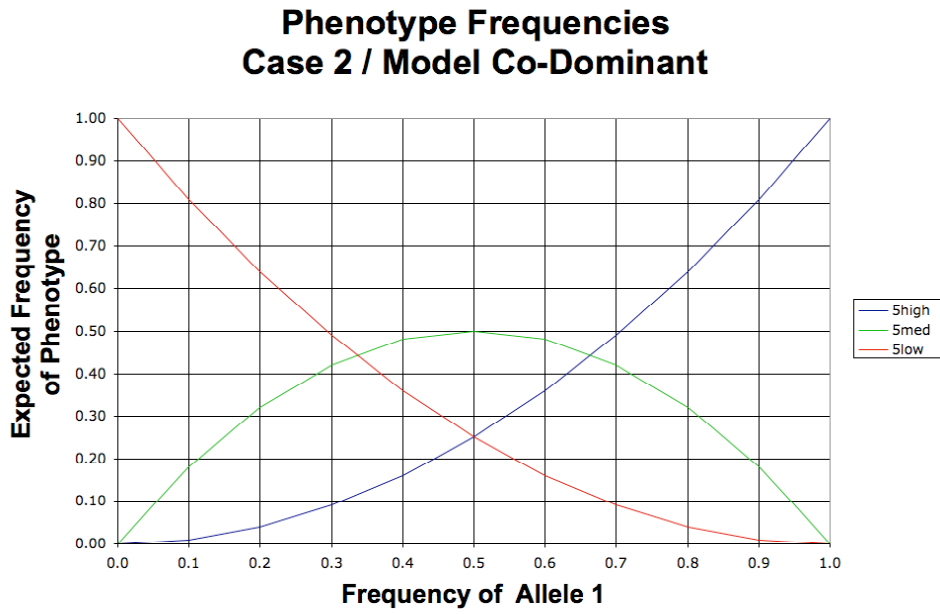
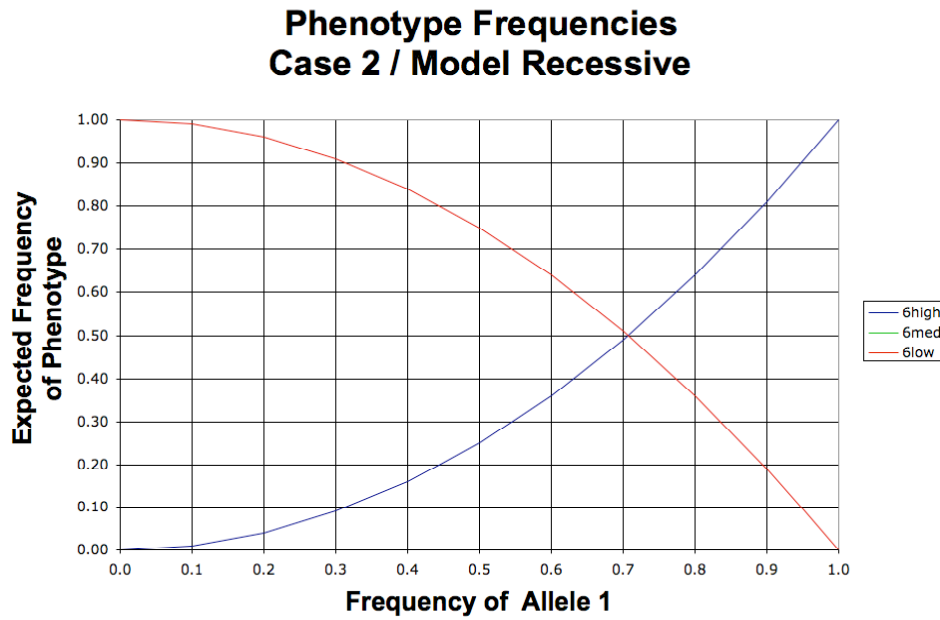


Figure 11 Phenotype Frequencies Case2 /Model Recessive

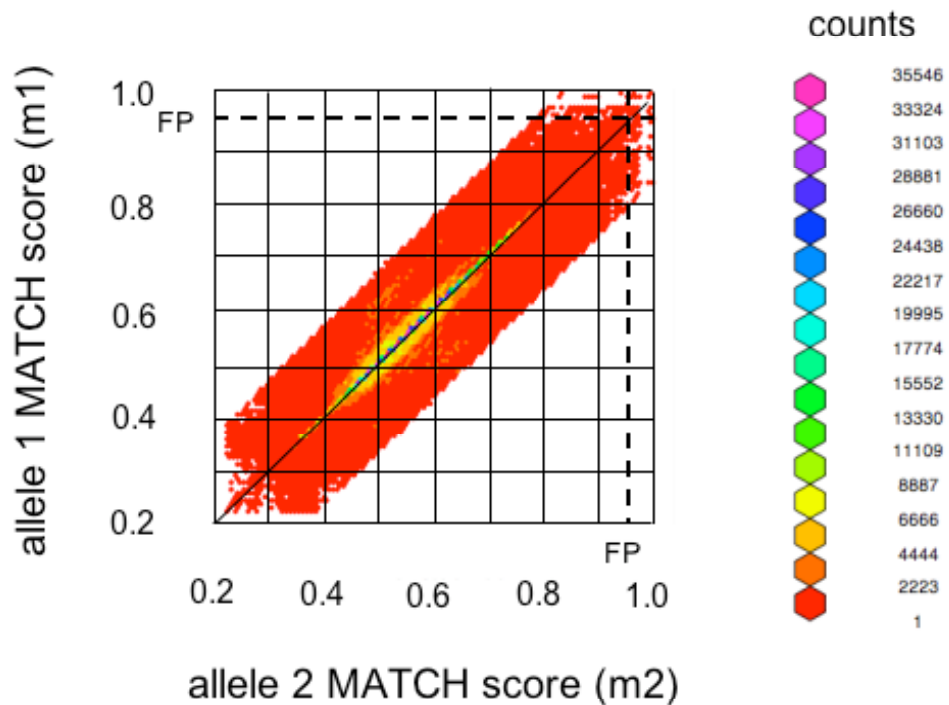


1.6.16 A Large Difference in MATCH Score May Correlate with a Large Difference in Transcription Factor Binding Affinity

For a given submitted DNA sequence (**a**) the MATCH [1] program predicts if a transcription factor will bind to a given genetic sequence by calculating a statistic called a “MATCH score” (**m**). This MATCH score represents an identity score between a sequence and a transcription factor binding site consensus sequence. Therefore it follows that two alleles (**a1** and **a2**) of a given polymorphism (the DNA sequence flanking the position of a polymorphic locus) may be separately evaluated by the MATCH algorithm, and a corresponding MATCH score for each allele (**m1** and **m2**) may be calculated (Figure page 402). It may be concluded that if the polymorphic variant is positioned within a regulatory region in the human genome, and the MATCH scores for these two alleles differ greatly, the polymorphism may associate with allele-specific gene expression.

Figure 12 Density plot of allelic MATCH scores for 4,547,844 polymorphisms using the NF-kappaB Matrix V\$NFKB_Q6

Density Plot of Allelic MATCH Scores (m1 and m2) for 4,547,844 Polymorphisms Using the NF-kappaB Matrix V\$NFKB_Q6



[950 polymorphisms have (m1 and/or m2) \geq FP, where FP = 0.955]

This is a density plot of the distribution of the allelic MATCH scores for 4,547,844 polymorphisms using the NF-kappaB transcription factor binding site matrix V\$NFKB_Q6. Most polymorphisms have small differences between their allele 1 (m1) and allele 2 (m2) MATCH scores. The dotted lines (FP = false positive cutoff threshold) represent the minimum MATCH score required to initiate transcription factor binding for the specified matrix.

The 950 polymorphisms having a MATCH score (m_1 and/or m_2) greater than or equal to the false positive cutoff threshold score (FP = 0.955) were ranked by the Delta-MATCH algorithm to identify those polymorphisms with the highest potential to create an allele-specific transcription factor binding site.

Those polymorphisms with large differences between their allelic MATCH scores (furthest from the line where $m_1 = m_2$), where either m_1 or m_2 is equal to 1.0 ranked highest.

1.6.17 Can a Large Delta-MATCH Score Identify a Genetic Locus Associated with Human Disease?

It is hypothesized that the difference between two allelic MATCH scores may correspond to a difference in transcription factor binding affinity. Furthermore, by calculating Delta-MATCH scores for every known transcription factor, for all known human polymorphisms, it may be possible to predict which polymorphisms may associate with human diseases characterized by irregular levels of mRNA expression by ranking these predictions in descending order of importance (descending order of their Delta-MATCH potential score). Once ranked, a Delta-MATCH Query Tool may be used to allow users to filter/search through the ranked predictions to identify novel candidate polymorphisms that may be good future targets of gene therapy.

1.7 The Delta-MATCH Method (Predicting Which Polymorphisms May Create Allele-Specific Binding Sites)

1.7.1 What is Biological Relevance?

1.7.1.1 Definition - biological relevance

A DNA sequence containing a nucleotide motif that may attract and bind with a transcription factor is considered “biologically relevant”, and a sequence that can’t is considered “biologically irrelevant”.

1.7.2 What is a “Delta-MATCH Potential Score (potential)?

1.7.2.1 Definition - potential (Delta-MATCH Potential Score)

The “Delta-MATCH Potential Score” is a statistic that reflects the absolute difference in biological relevance between two to polymorphic alleles. This is the primary ranking statistic in the Delta-MATCH database. A potential score may range from 0.0 to 1.0. Polymorphisms with high potential scores may be considered candidate polymorphism for human diseases that are characterized by dysregulation in mRNA gene expression.

1.7.2.2 Warning - The “Delta-MATCH potential score” is informative, but not sufficient

The “Delta-MATCH potential score” is informative, not by itself sufficient to predict whether or not a polymorphic site will have a biological affect on transcription factor binding. Other characteristics of the polymorphism must be considered in conjunction

with a Delta-MATCH potential score to determine if a candidate polymorphism may associate with a disease phenotype. Specifically, it is important to consider if a polymorphism is located in a potential regulatory region. Although a polymorphism may affect gene expression when located in an enhancer region distal to a set of genes, it is more likely that it may associate with variable levels of gene expression when it is located in a promoter region immediately upstream of a gene, near a transcriptional start site, or near an mRNA splicing junction.

1.7.3 The Threshold of Biological Relevance Is Estimated By the False Positive Threshold Cutoff Score

The Delta-MATCH method uses 550 TFBS matrixes defined by the in the BIOBASE MATCH program that represent 550 different vertebrate transcription factor binding site consensus sequences [1].

1.7.3.1 Definition - cutoff threshold of biological relevance

This is the minimum value of a MATCH score that is biological relevant for a given TFBS matrix.

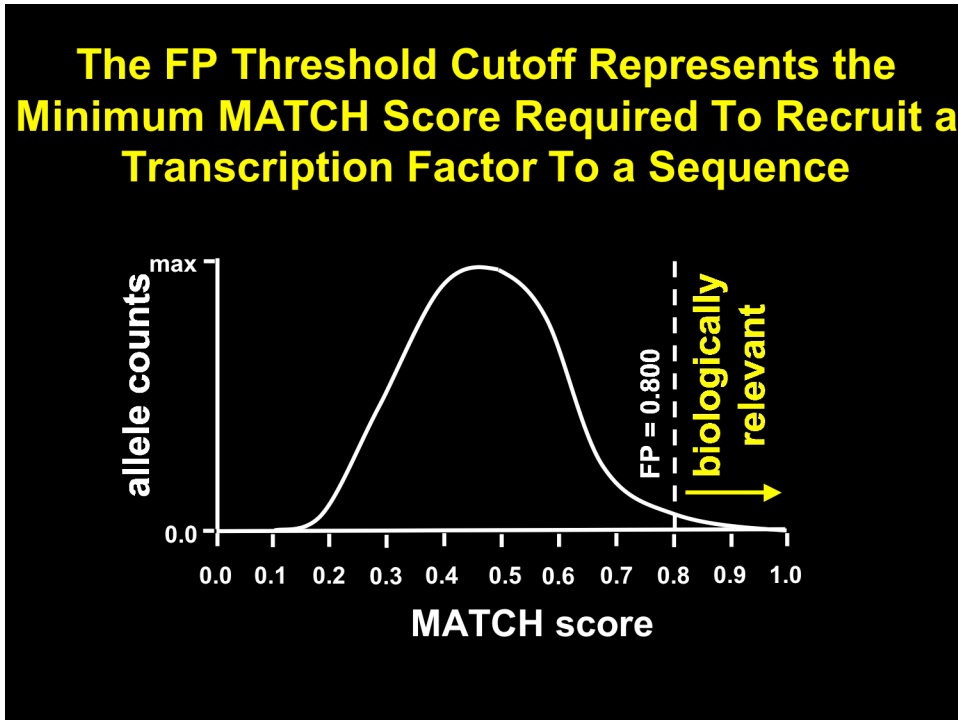
1.7.3.2 Definition - false positive cutoff score (FP)

This is an estimation of the “cutoff threshold of biological relevance” that was determined by the BIOBASE team. Each of the 550 TFBS matrixes has a unique FP cutoff threshold score.

The BIOBASE team has empirically determined a minimum false positive threshold cutoff score (FP) that represents the minimal score required to induce the first moment of transcription factor binding. BIOBASE has estimated a unique FP cutoff for each of the 550 matrixes used by Delta-MATCH [1]. Note that 34 of the 584 vertebrate matrixes provided by MATCH version 10.2 did not have a FP statistic estimated, and these matrixes have been removed from consideration in this Delta-MATCH release.

If a DNA sequence is compared to a TFBS matrix and scored using the MATCH algorithm, and its “highest MATCH score” is greater than or equal to the FP cutoff score, the sequence is considered to be “biologically relevant”. However, if the “highest MATCH score” is less than the FP cutoff score, the sequence is considered to be “biologically irrelevant”. In the following figure the FP cutoff threshold for the example MATCH score distribution is 0.8.

Figure 13 The FP Threshold Cutoff Represents the Minimum MATCH Score Required to Recruit a Transcription Factor to a Sequence.

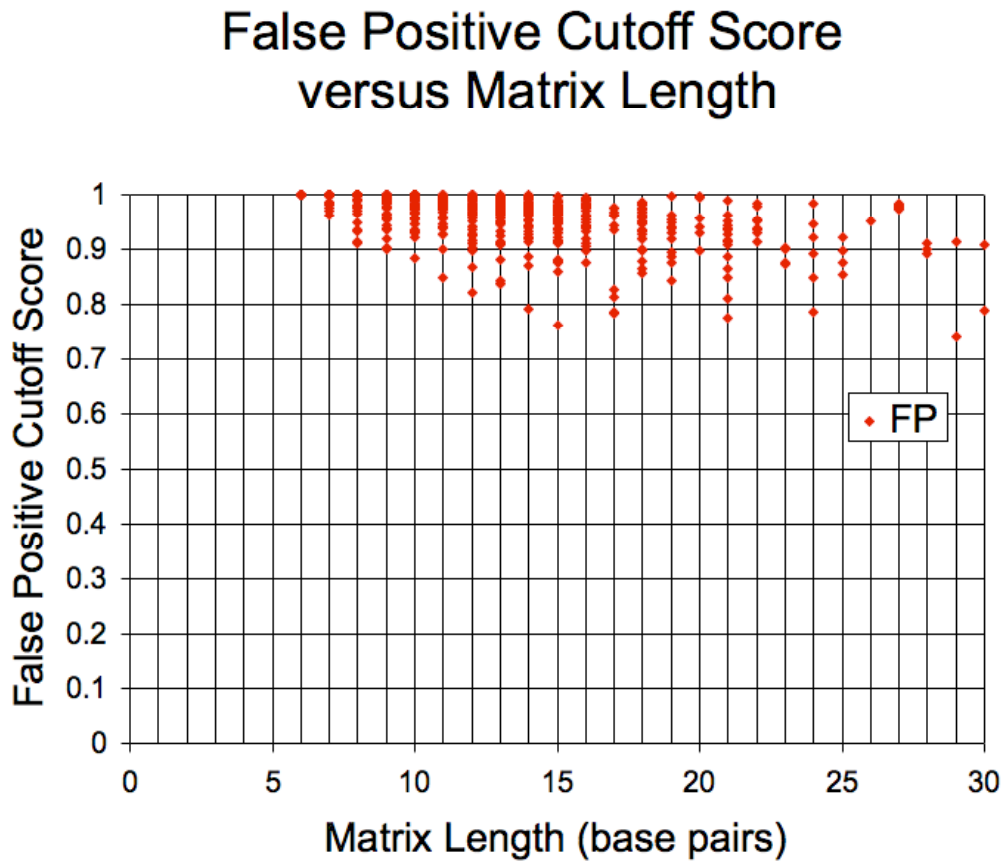


For a given matrix it is possible to show the complete distribution of every polymorphic allele by graphing the number of allele counts (y-axis) per MATCH score (x-axis). It is then possible to calculate the “biological relevance” for a given MATCH score if the threshold cutoff score for that matrix is known (page 35).

1.7.4 The False Positive (FP) Cutoff Is Not Correlated with Matrix Length

A graph of the False Positive Cutoff Score versus the Matrix Length for the 550 BIOBASE transcription factor matrixes has a low correlation coefficient (-0.493). See Figure 32 on page 78 for the distribution of TFBS matrix lengths.

Figure 14 False Positive Cutoff Score vs. Matrix length



1.7.4.1 Definition - model

This is a mathematical approximation that Delta-MATCH uses to calculate an estimation of “biological relevance of a MATCH score” (brm)

1.7.4.2 Definition - biological relevance of a MATCH score (brm)

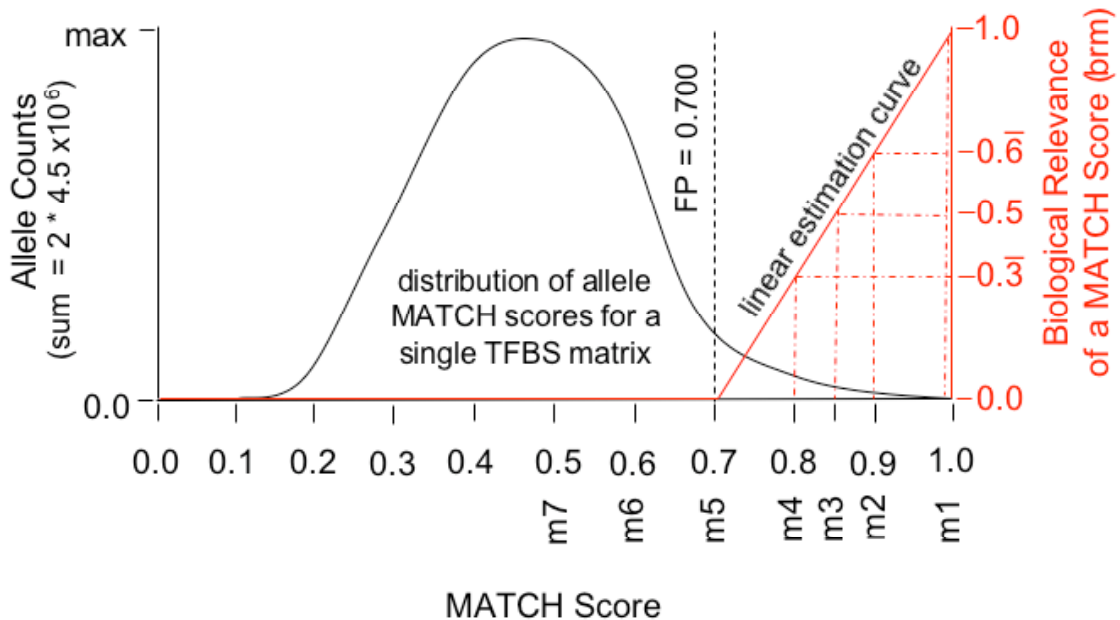
The “biological relevance of a MATCH score” can be estimated from a linear model (Estimation Model 1 page 35) overlaid on the complete distribution of MATCH scores for a set of alleles if the “minimum cutoff threshold of biological relevance” has been estimated (FP).

1.7.5 How is Biological Relevance Calculated?

In a hypothetical example distribution of MATCH scores (Estimation Model 1), the false positive threshold (FP) is 0.700. For a given set of allele MATCH scores ($m_1 = 1.0$, $m_2 = 0.9$, $m_3 = 0.85$, $m_4 = 0.8$, $m_5 = 0.7$, $m_6 = 0.6$), it is possible to correlate an associated “biological relevance of a MATCH score” (brm) by estimating from a “linear curve” starting from the x-axis at the point of the false positive threshold ($x = 0.7$, $y = 0.0$) up to the point at which the biological relevance score is maximum and reflective of the optimal transcription factor binding site consensus motif ($x = 1.0$, $y = 1.0$).

Figure 15 Estimation Model 1 - a linear estimation curve

Estimation Model 1 - The Potential Score Is the Absolute Difference in Biological Relevance Between Two Polymorphic Alleles Using a Linear Curve



Note that the false positive cutoff value for this matrix is fairly low (FP = 0.700) and the slope of the linear estimation curve is fairly small (~ 1.0 brm / 0.3 MATCH score units).

The “biological relevance of a MATCH score” (**brm**) for these hypothetical alleles using Model 1 are as follows:

brm1 (m1 = 1.0) =	1.000
brm2 (m2 = 0.9) =	0.666
brm3 (m3 = 0.85) =	0.500
brm4 (m4 = 0.8) =	0.333
brm5 (m5 = 0.7) =	0.000
brm6 (m6 = 0.6) =	0.000
brm7 (m7 = 0.5) =	0.000

Note that in Estimation Model 1, there may be many false positive predictions for alleles with MATCH scores greater than 0.7 and less than 0.8 (below the red linear curve within the region nearest the estimated threshold cutoff score).

1.7.6 Calculating the “absolute percent difference” in allelic MATCH scores and the “Delta-MATCH potential score”

1.7.6.1 Definition - mean MATCH score

The “**mean MATCH score**” is the average of the two allelic MATCH scores (Equation 7).

Equation 4 - mean MATCH score

$$\text{mean}(m1, m2) = ((m1 + m2) / 2)$$

1.7.6.2 Definition - larger polymorphism MATCH score (m_max)

The “larger polymorphism MATCH score is the greater of m1 and m2.

Equation 5 - larger polymorphism MATCH score (m_max)

$$m_max(m1, m2) = \max(m1, m2)$$

1.7.6.3 Definition - smaller polymorphism MATCH score (m_min)

The “larger polymorphism MATCH score is the lesser of m1 and m2.

Equation 6 - smaller polymorphism MATCH score (m_min)

$$m_max(m1, m2) = \min(m1, m2)$$

1.7.6.4 Definition - absolute difference in MATCH score (m_dif)

The “**absolute difference in MATCH score**” is the absolute difference between the highest MATCH scores for allele 1 and allele 2 (Equation 7).

Equation 7 - absolute difference in MATCH score (m_dif)

$$m_dif(m1, m2) = \text{abs}(m1 - m2)$$

The “**Delta-MATCH potential score**” for a polymorphism is calculated as the absolute difference in biological relevance between a pair of MATCH scores (Equation 8).

Equation 8 - Delta-MATCH potential score (potential)

$$\text{potential} (m1, m2) = \text{abs}(brm1 - bmr2)$$

1.7.6.5 Definition - absolute percent difference in MATCH score (m_per)

The “absolute percent difference in MATCH score” is calculated by multiplying the absolute difference between the MATCH scores between two alleles by 100, and dividing the product by the larger of the MATCH scores (Equation 9).

Equation 9 - absolute percent difference in MATCH score (m_per)

$$m_per (m1, m2) = (100*\text{abs} (m1 - m2))/(\text{max}(m1,m2))$$

1.7.7 How is a Delta-MATCH Potential Score for a Polymorphism Calculated?

The “**Delta-MATCH potential score**” for a polymorphism with two alleles a1 and a2, can be calculated by determining the absolute difference between brm1 and brm2 (Equation 10).

Equation 10 - Delta-MATCH potential score (potential)

$$\text{potential} (m1, m2) = \text{abs} (brm1 - bmr2)$$

1.7.8 Ranking Delta-MATCH Results (by potential, (max (m1, m2)), m_per)

After the potential scores for the set of SNPs were calculated (*prioritize_results.py*), a second python script (*prioritize_results.py*) ranked these SNPs by “descending order of importance” for each of the 550 TFBS matrixes (this is the order returned by the Delta-MATCH Query Tool). SNPs were ranked by sorting **firstly** by descending order by their Delta-MATCH potential scores (**potential**), and **secondly** by descending order of their percent difference in MATCH scores (**m_per**), and **thirdly** by descending order of their largest MATCH score (**max (m1, m2)**),

1.7.9 Calculating Example Potential Scores (Estimation Model 2)

These examples are ranked in descending order of importance.

$$\begin{aligned} \text{example 1 - } & \text{potential (m1, m7)} = \text{abs (1.000 - 0.000)} = 1.000 \\ & \text{m_per (m1, m7)} = (100*\text{abs}(1.0 - 0.5))/1.0 = 50 \% \end{aligned}$$

$$\begin{aligned} \text{example 2 - } & \text{potential (m1, m6)} = \text{abs (1.000 - 0.000)} = 1.000 \\ & \text{m_per (m1, m6)} = (100*\text{abs}(1.0 - 0.6))/1.0 = 40 \% \end{aligned}$$

$$\begin{aligned} \text{example 3 - } & \text{potential (m1, m5)} = \text{abs (1.000 - 0.000)} = 1.000 \\ & \text{m_per (m1, m5)} = (100*\text{abs}(1.0 - 0.7))/1.0 = 30 \% \end{aligned}$$

$$\begin{aligned} \text{example 4 - } & \text{potential (m1, m4)} = \text{abs (1.000 - 0.333)} = 0.666 \\ & \text{m_per (m1, m4)} = (100*\text{abs}(1.0 - 0.8))/1.0 = 20 \% \end{aligned}$$

$$\text{example 5 - } \text{potential (m3, m6)} = \text{abs (0.500 - 0.000)} = 0.500$$

$$m_per (m3, m6) = (100*abs(0.85 - 0.6))/0.85 = 29.4 \%$$

example 6 - $potential (m3, m5) = abs (0.500 - 0.000) = 0.500$

$$m_per (m3, m5) = (100*abs(0.85 - 0.7))/0.85 = 17.6 \%$$

example 7 - $potential (m1, m3) = abs (1.000 - 0.500) = 0.500$

$$m_per (m1, m3) = (100*abs(1.0 - 0.85))/1.0 = 15 \%$$

example 8 - $potential (m4, m7) = abs (0.333 - 0.000) = 0.333$

$$m_per (m4, m7) = (100*abs(0.8 - 0.5))/0.8 = 37.5\%$$

example 9 - $potential (m4, m5) = abs (0.333 - 0.000) = 0.333$

$$m_per (m4, m5) = (100*abs(0.8 - 0.7))/0.8 = 12.5 \%$$

example 10 - $potential (m2, m4) = abs (0.666 - 0.333) = 0.333$

$$m_per (m2, m4) = (100*abs(0.9 - 0.8))/0.9 = 11.1 \%$$

example 11 - $potential (m1, m2) = abs (1.000 - 0.666) = 0.333$

$$m_per (m1, m2) = (100*abs(1.0 - 0.9))/1.0 = 10.0 \%$$

example 12 - $potential (m5, m7) = abs (0.000 - 0.000) = 0.000$

$$m_per (m5, m7) = (100*abs(0.7 - 0.5))/0.7 = 28.6 \%$$

example 13 - $potential (m6, m7) = abs (0.000 - 0.000) = 0.000$

$$m_per (m6, m7) = (100*abs(0.6 - 0.5))/0.6 = 16.7 \%$$

example 14 - potential (m5, m6) = abs (0.000 - 0.000) = 0.000

m_per (m5, m6) = (100*abs(0.7 - 0.6))/1.0 = 14.3 %

example 15 - potential (m1, m1) = abs (1.000 - 1.000) = 0.000

m_per (m1, m1) = (100*abs(1.0 - 1.0))/1.0 = 0.0 %

example 16 - potential (m3, m3) = abs (0.850 - 0.850) = 0.000

m_per (m3, m3) = (100*abs(0.85 - 0.85))/0.85 = 0.0 %

example 17 - potential (m5, m5) = abs (0.000 - 0.000) = 0.000

m_per (m5, m5) = (100*abs(0.7 - 0.7))/0.7 = 0.0 %

example 18 - potential (m7, m7) = abs (0.000 - 0.000) = 0.000

m_per (m7, m7) = (100*abs(0.5 - 0.5))/0.5 = 0.0 %

Note that in the above examples the maximum potential score of 1.0 is found when one allele has a biological relevance of 1.0 (MATCH score = 1.0) when the other allele has a biological relevance of 0.0 (MATCH score <= FP). A potential score of 0.0 is found when both alleles have MATCH scores less than or equal to the FP cutoff, and when the allelic MATCH scores are equal whether high or low (m1 = m2).

Interestingly, because the estimation curve is linear, it is possible to create the **same** “**potential**” **score** for more than one combination of allele pairs [Estimation Model 1 ranked examples (1 = 2 = 3) > 4 > (5 = 6 =7) > (8 = 9 = 10 = 11) > (12 = 13 = 14 = 15 =

16 = 17 = 18)]. Note that example 5 is ranked higher than example 6 because the results are sorted by descending **m_per** of before descending order of **max(m1, m2)**.

The aim of Delta-MATCH is to identify those polymorphisms that are biologically relevant (large potential) and have very different allelic MATCH scores (a large m_per). For a given TFBS matrix, most polymorphisms are “biologically irrelevant” (Definition page 30) and have a potential score equal to 0.0 because both allelic MATCH scores (m1 and m2) are less than the false positive cutoff. Conversely, very few polymorphisms are “biologically relevant” (Definition page 30). For example, out of the 4.5 million SNPs evaluated by Delta-MATCH using the V\$NFKB_Q6 TFBS matrix, only 878 SNPs were biologically relevant and had potential scores greater than 0.0 (Table 3 page 62).

Figure 16 Count Versus Mean MATCH Score (V\$NFKB_Q6, n = 4,547,844)

This is a true distribution of MATCH scores for 4,547,844 polymorphisms using the NFKB transcription factor binding site matrix V\$NFKB_Q6. Note that the false positive cutoff value for this matrix is fairly high (FP = 0.955) and the slope of the linear estimation curve is fairly large (~ 1.0 brm / 0.045 mean MATCH score units).

Count vs. Mean MATCH Score (V\$NFKB_Q6; n = 4,547,848)

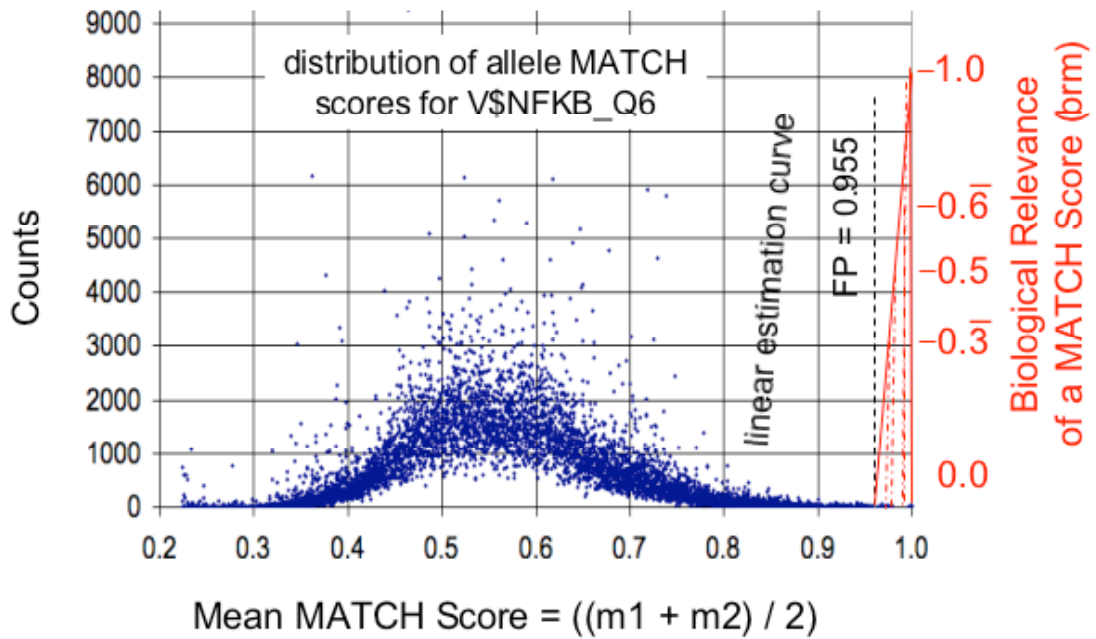


Figure 17 Histogram of MATCH scores for 4,547,844 polymorphisms using the NF-kappaB Matrix V\$NFKB_Q6

This is a histogram for the MATCH scores of 4,547,844 polymorphisms using the NFKB transcription factor binding site matrix V\$NFKB_Q6 (each block equals 5%). The dotted blue curve represents the distribution of MATCH scores for the UCSC reference allele (m1) and the dotted green curve represents the distribution of MATCH scores for the alternate allele (m2). The false positive cutoff threshold value for this matrix is (FP) is 0.955 and the slope of the linear estimation curve is fairly large (1.0 brm / 0.045 MATCH score units).

Histogram of MATCH Scores for 4,547,844 Polymorphisms Using the NF-kappaB Matrix V\$NFKB_Q6

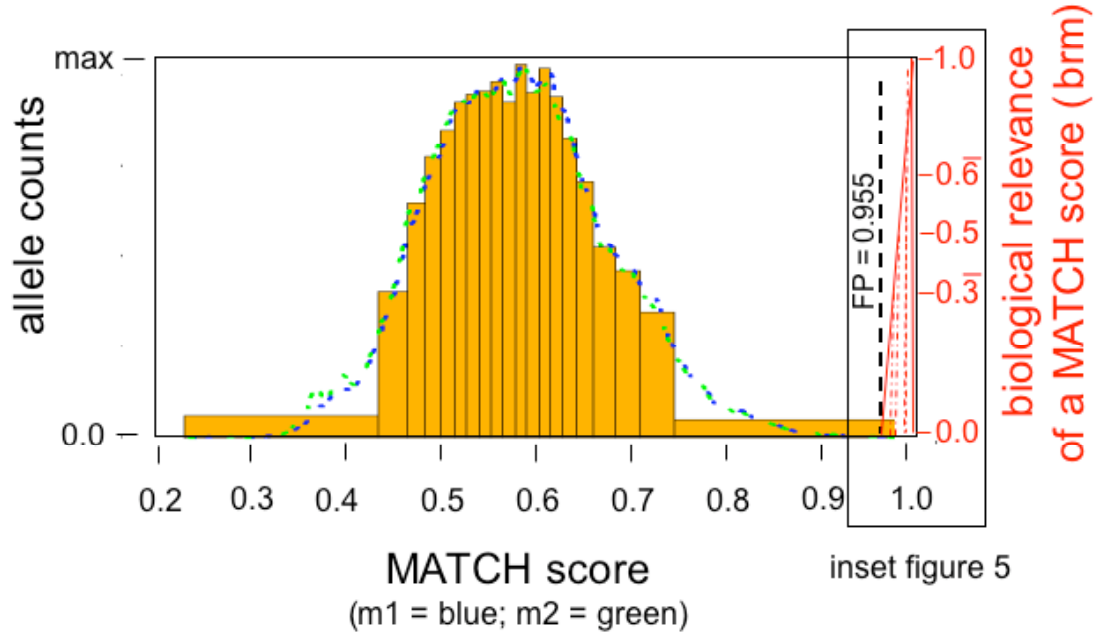
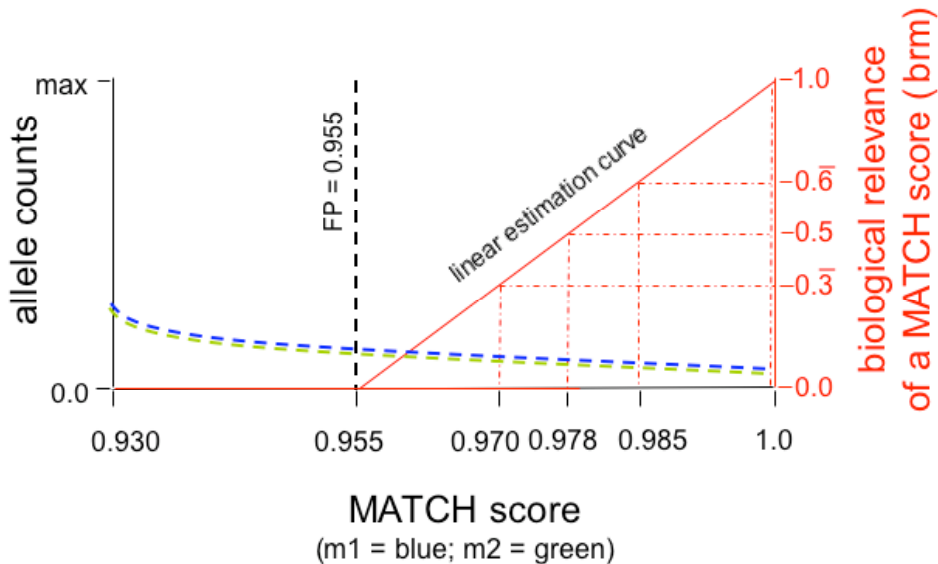


Figure 18 Delta-MATCH estimates the biological relevance of a MATCH score with a linear model that approximates transcription factor binding affinity

Sequences are predicted to have transcription factor binding affinity proportional to the degree that a given MATCH score is above a false positive cutoff threshold (FP), a score representing the minimum MATCH score required to recruit a transcription factor to a double-stranded nucleotide sequence. Delta-MATCH uses the cutoff threshold values provided by the BIOBASE TRANSFAC database version 10.2.

Sequences with a MATCH score of 1.0 are predicted to have the strongest transcription factor binding affinity and the highest biological relevance. Sequences with MATCH scores less than or equal to the minimum cutoff (FP) are predicted to have no transcription factor binding affinity and no biological relevance.

Delta-MATCH Estimates the Biological Relevance of a MATCH Score with a Linear Model that Approximates Transcription Factor Binding Affinity



1.7.10 The Delta-MATCH Estimation Model Is Linear (used in version 1.0)

The currently employed Delta-MATCH estimation model is linear and is useful as a good first approximation for estimating the biological relevance of a MATCH score (**brm**) (Definition page 34). It is not yet known if other models might improve reduce the number of false positives and false negatives, and it is anticipated that Delta-MATCH will continue to improve its model as more accurate transcription factor binding sites definitions are defined, and as the molecular conditions needed for proper transcriptional regulation are better understood.

In the linear estimation model, those alleles with MATCH scores less than the false positive score (FP) are considered biological irrelevant ($brm = 0.0$) and are not predicted recruit transcription factors to bind. Those alleles that have MATCH scores greater than or equal to the false positive cutoff are predicted to recruit transcription factors with an affinity proportional to the increase in MATCH score above the FP cutoff value to the point of its maximum.

The linear estimation curve is drawn starting from the x-axis at point of the FP cutoff score ($m = 0.0$; $brm = 0.0$) and extending up to it maximum point at which maximal binding is predicted ($m = 0.0$; $brm = 1.0$).

The linear estimation curve has a slope of zero for MATCH scores greater than zero but less than or equal to the FP cutoff. The slope of the linear estimation curve for MATCH score values greater than the FP can be calculated (Equation 11).

Equation 11 - Slope of Linear Estimation Curve (slope)

$$\text{slope} = \frac{(1.0 \text{ biological relevance of a MATCH score})}{(1.0 \text{ MATCH score} - (\text{FP cutoff value MATCH score}))}$$

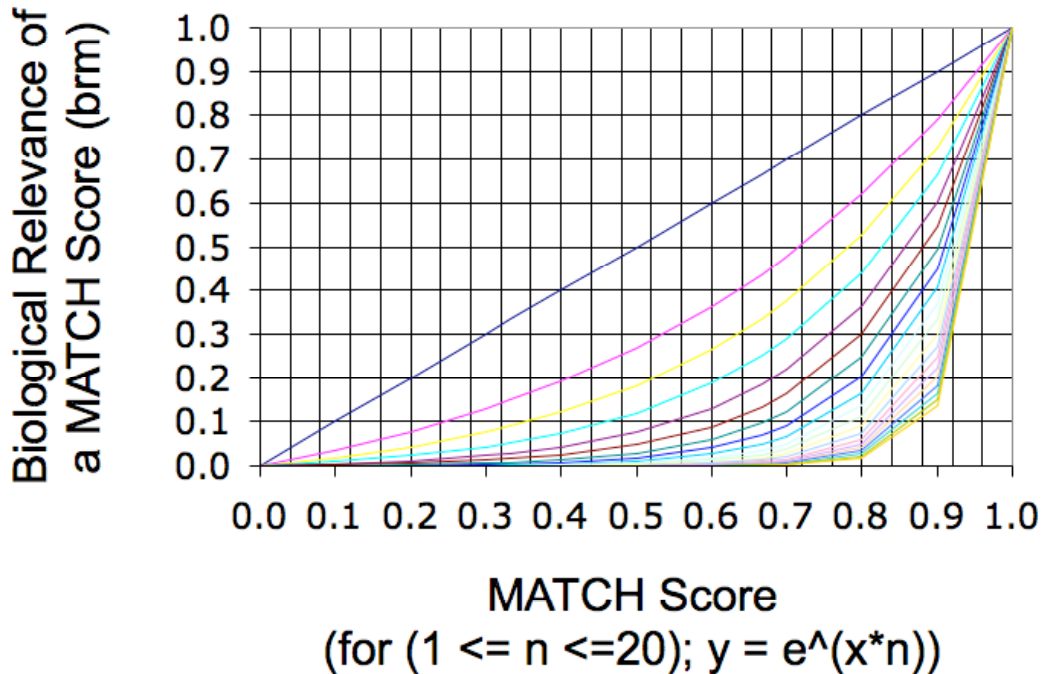
The exact biological relevance of any MATCH score (**m**) can be found by identifying the MATCH score on the x-axis, tracing a line vertically from to the point of intersection with the linear estimation curve, and then tracing horizontally to the right to its corresponding **brm** value.

1.7.11 What Level of Potential Score Is Considered Significant?

For a given TFBS matrix, polymorphisms with the highest potential scores should be considered the most likely to promote allele-specific TF binding. For now, it is recommended that users focus primarily on polymorphisms with potential scores greater than or equal to 0.3, and to consider polymorphisms with lower potential scores secondarily. Ignoring results below the 0.3 cutoff may greatly reduce the number of false positive predictions.

Figure 19 Future Alternative Delta-MATCH Models May Use Exponential Estimation Curves

Alternative Delta-MATCH Estimation Models May Use Exponential Curves

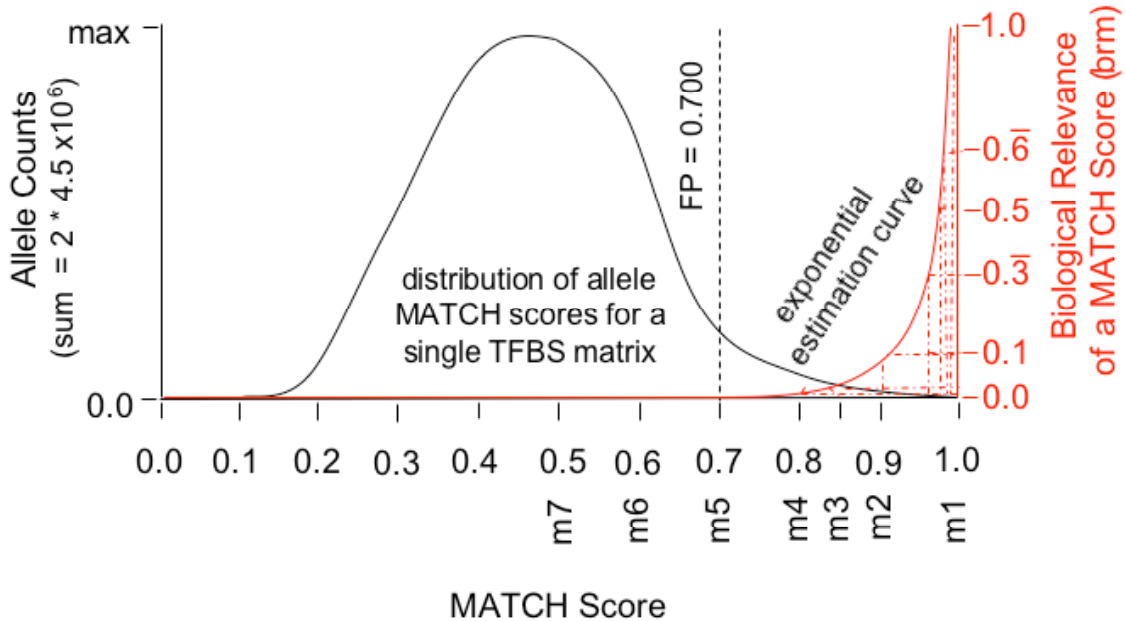


1.7.12 Future Versions of Delta-MATCH May Use Higher Order Models That May Reduce Type-1 Errors (False Positives)

Future versions of Delta-MATCH may allow users to select from a variety of higher order estimation models when ranking results. It is expected that the largest number of false positive biologically relevant predictions occur at or near the estimated threshold cutoff (FP). Using exponential estimation models might reduce the number of Type-1 errors by more conservatively estimating the biological relevance of a MATCH score (brm) for those MATCH scores at or near the estimated false positive cutoff.

Figure 20 Estimation Model 2 - an exponential estimation curve

Estimation Model 2 - The Potential Score Is the Absolute Difference in Biological Relevance Between Two Polymorphic Alleles Using an Exponential Curve



In the figure of Estimation Model 2, the majority of the false positive are likely located where the MATCH score is greater than 0.7 and less than 0.8, and below the red exponential curve. When compared with the linear estimation model (Estimation Model 1 page 35), it is evident that using an exponential model (Estimation Model 2) might have the relative effect of reducing the number of the false positive predictions in the Delta-MATCH database. A higher order model would more conservatively estimate the relationship between a MATCH score and its biological relevance (where $\Delta y > \Delta x$; $FP < x < 1.0$), at the risk of losing some important predictions through a type-2 error.

The “biological relevance of a MATCH score” (**brm**) for each of the hypothetical example alleles (m1 - m7) using Estimation Model 2 is as follows:

$$\text{brm1 (m1 = 1.0) = 1.000}$$

$$\text{brm2 (m2 = 0.9) = 0.100}$$

$$\text{brm3 (m3 = 0.85) = 0.010}$$

$$\text{brm4 (m4 = 0.8) = 0.001}$$

$$\text{brm5 (m5 = 0.7) = 0.000}$$

$$\text{brm6 (m6 = 0.6) = 0.000}$$

$$\text{brm7 (m7 = 0.5) = 0.000}$$

1.7.12.1 Calculating Example Potential Scores (Estimation Model 2)

$$\text{example 1 - potential (m1, m7) = abs (1.000 - 0.000) = 1.000}$$

$$\text{m_per (m1, m7) = (100*abs(1.0 - 0.5))/1.0 = 50 \%}$$

$$\text{example 2 - potential (m1, m6) = abs (1.000 - 0.000) = 1.000}$$

$$\text{m_per (m1, m6) = (100*abs(1.0 - 0.6))/1.0 = 40 \%}$$

$$\text{example 3 - potential (m1, m5) = abs (1.000 - 0.000) = 1.000}$$

$$\text{m_per (m1, m5) = (100*abs(1.0 - 0.7))/1.0 = 30 \%}$$

$$\text{example 4 - potential (m1, m4) = abs (1.000 - 0.001) = 0.999}$$

$$\text{m_per (m1, m4) = (100*abs(1.0 - 0.8))/1.0 = 20 \%}$$

$$\text{example 7 - potential (m1, m3) = abs (1.000 - 0.010) = 0.990}$$

$$\text{m_per (m1, m3) = (100*abs(1.0 - 0.85))/1.0 = 15 \%}$$

example 11 - potential (m1, m2) = abs (1.000 - 0.100) = 0.900
m_per (m1, m2) = (100*abs(1.0 - 0.9))/1.0 = 10.0 %

example 10 - potential (m2, m4) = abs (0.100 - 0.001) = 0.099
m_per (m2, m4) = (100*abs(0.9 - 0.8))/0.9 = 11.1 %

example 5 - potential (m3, m6) = abs (0.010 - 0.000) = 0.010
m_per (m3, m6) = (100*abs(0.85 - 0.6))/0.85 = 29.4 %

example 6 - potential (m3, m5) = abs (0.010 - 0.000) = 0.010
m_per (m3, m5) = (100*abs(0.85 - 0.7))/0.85 = 17.6 %

example 8 - potential (m4, m7) = abs (0.001 - 0.000) = 0.001
m_per (m4, m7) = (100*abs(0.8 - 0.5))/0.8 = 37.5%

example 9 - potential (m4, m5) = abs (0.001 - 0.000) = 0.001
m_per (m4, m5) = (100*abs(0.8 - 0.7))/0.8 = 12.5 %

example 12 - potential (m5, m7) = abs (0.000 - 0.000) = 0.000
m_per (m5, m7) = (100*abs(0.7 - 0.5))/0.7 = 28.6 %

example 13 - potential (m6, m7) = abs (0.000 - 0.000) = 0.000
m_per (m6, m7) = (100*abs(0.6 - 0.5))/0.6 = 16.7 %

example 14 - potential (m5, m6) = abs (0.000 - 0.000) = 0.000

$$m_per (m5, m6) = (100*abs(0.7 - 0.6))/1.0 = 14.3 \%$$

example 15 - potential (m1, m1) = abs (1.000 - 1.000) = 0.000

$$m_per (m1, m1) = (100*abs(1.0 - 1.0))/1.0 = 0.0 \%$$

example 16 - potential (m3, m3) = abs (0.010 - 0.010) = 0.000

$$m_per (m3, m3) = (100*abs(0.85 - 0.85))/0.85 = 0.0 \%$$

example 17 - potential (m5, m5) = abs (0.000 - 0.000) = 0.000

$$m_per (m5, m5) = (100*abs(0.7 - 0.7))/0.7 = 0.0 \%$$

example 18 - potential (m7, m7) = abs (0.000 - 0.000) = 0.000

$$m_per (m7, m7) = (100*abs(0.5 - 0.5))/0.5 = 0.0 \%$$

1.7.13 Comparison of Estimation Model 1 and Estimation Model 2 Ranked

Examples

1.7.14 Estimation Model 1 ranked examples

Examples are ranked from left to right.

$$(1 = 2 = 3) > 4 > (5 = 6 = 7) > (8 = 9 = 10 = 11) > (12 = 13 = 14 = 15 = 16 = 17 = 18)$$

1.7.15 Estimation Model 2 ranked examples

Examples are ranked from left to right.

$$(1 = 2 = 3) > 4 > 7 > 11 > 10 > (5 = 6) > (8 = 9) > (12 = 13 = 14 = 15 = 16 = 17 = 18)$$

Re-ranking the previous examples using Estimation Model 2 promoted examples **7**, **11** and **10** (highlighted in red) in rank because their “potential” scores are much higher in the exponential model when compared with the linear model. Although Estimation Models 1 and 2 rank some of the example pairs differently, they both elevate the polymorphisms with the largest potential scores to the top.

1.7.15.1 Definition - biological relevance of a polymorphic site (brps)

This is a logarithmic transformation of the maximum of a polymorphisms brm1 and brm2, and reflects how biologically relevant a polymorphic site is.

Equation 12 - Biological Relevance of a Polymorphic Site (brps)

$$\text{brps} = -\log(1.0000001 - \max(\text{brm1}, \text{brm2}))$$

The “biological relevance of a polymorphic site” is calculated using the larger of the two allelic biological relevance scores (brm1 or brm2). This log transformation helps to visually separate the highest ranking polymorphisms from the majority so the distribution of millions of results can be displayed simultaneously without obscuring the most interesting results (Figure 24 page 60)

1.7.16 Viewing a Ranked Set of Delta-MATCH Potential Scores Graphically

It is the purpose of the Delta-MATCH algorithm to rank a list of polymorphisms by their order of importance. There are several ways to view the distribution of millions of ranked Delta-MATCH scores simultaneously. Each viewing method has the advantage of either displaying very large distributions of polymorphisms, or displaying only those biologically relevant polymorphisms ranked by descending order of their importance. The following graphs/plots show the results for 4,547,844 polymorphisms that have been searched by Delta-MATCH using the NF-kB TFBS matrix V\$NFKB_Q6 (FP threshold cutoff = 0.955).

1

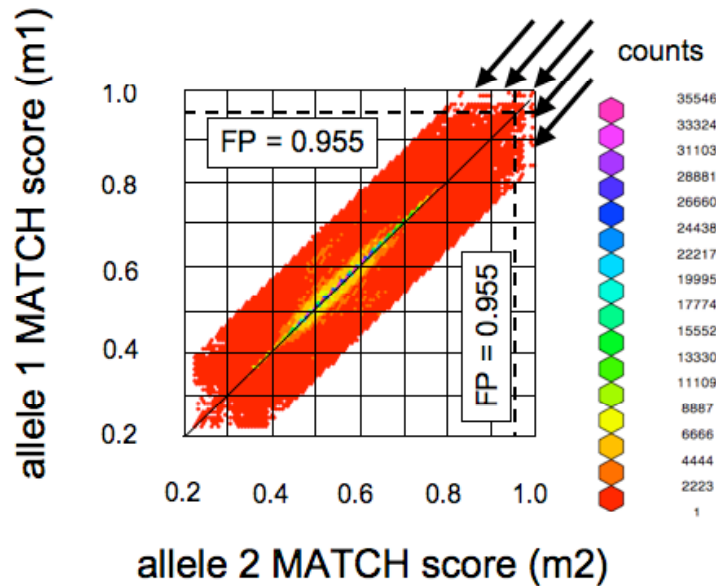
- Density Plot of the Allelic MATCH Scores for 4,547,844 Polymorphisms (NF-kB) (Figure 21 page 56)
- Count Versus Biological Relevance of a MATCH Score (Figure 22 page 57)
- Absolute Difference in MATCH Score vs. Larger MATCH Score of a Polymorphism (Figure 23 page 58)
- Potential Score Versus Biological Relevance of a Polymorphic Site (Figure 24 page 60)
- Potential Score Versus Absolute Percent Difference in MATCH Score (Figure 25 page 61)
- Rank versus potential score versus absolute percent difference in MATCH score for 950 high-value polymorphisms (3-D plot) (Figure 26 page 61)

¹ **Note:** the “Potential Score Versus Absolute Percent Difference in MATCH Score” and the “Rank versus potential score versus absolute percent difference in MATCH score for 950 high-value polymorphisms (3-D plot)” best visually separate polymorphisms with equivalent potential scores.

Figure 21 Density Plot of the Allelic MATCH Scores for 4,547,844 Polymorphisms (NF-kB)

This is a plot of the allelic MATCH scores for 4.5 million polymorphisms. For every polymorphism the highest MATCH score for allele 1 (y-axis) is plotted versus the highest MATCH score for allele 2 (x-axis).

Density Plot of the Allelic MATCH Scores for 4,547,844 Polymorphisms (NF-kB)
(V\$NFKB_Q6; FP = 0.955)



Each pixel of this plot represents the number of polymorphisms with a particular combination of allele 1 and allele 2 MATCH scores. The number of polymorphisms at each pixel is color-coated by the density of counts using the heat map on the right. Dotted lines representing the NF-kB cutoff threshold score (FP = 0.955) are shown. Only 950 of 4.5 million polymorphisms are positioned either above the dotted line in the y-axis, or to the right of the dotted line in the x-axis, and are cases where at least one of the two alleles has a MATCH score greater than or equal to the threshold (those red

points positioned under the black arrows). It can be seen in this graph that relatively few (409) of the 4.5 million searched polymorphisms have potential scores greater than 0.3 (Figure page 27). It is the purpose of the Delta-MATCH algorithm to rank these 950 NF-kB results by their order of importance.

Figure 22 Count Versus Biological Relevance of a MATCH Score

This plot shows the cumulative count of polymorphisms having less than or equal to a particular ‘biological relevance of a MATCH score’ (brm) (Definition page 34). The ‘larger polymorphism MATCH score’ (m_max) (Equation 5 page 20) for each polymorphism is plotted on the x-axis.

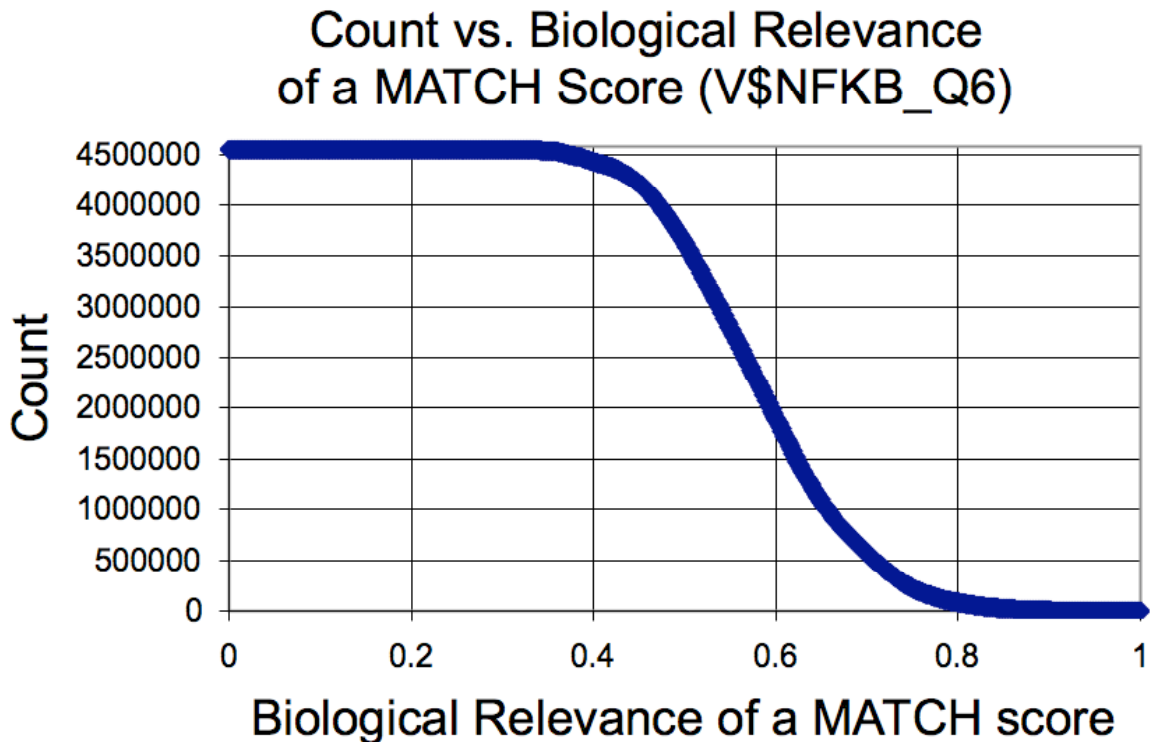
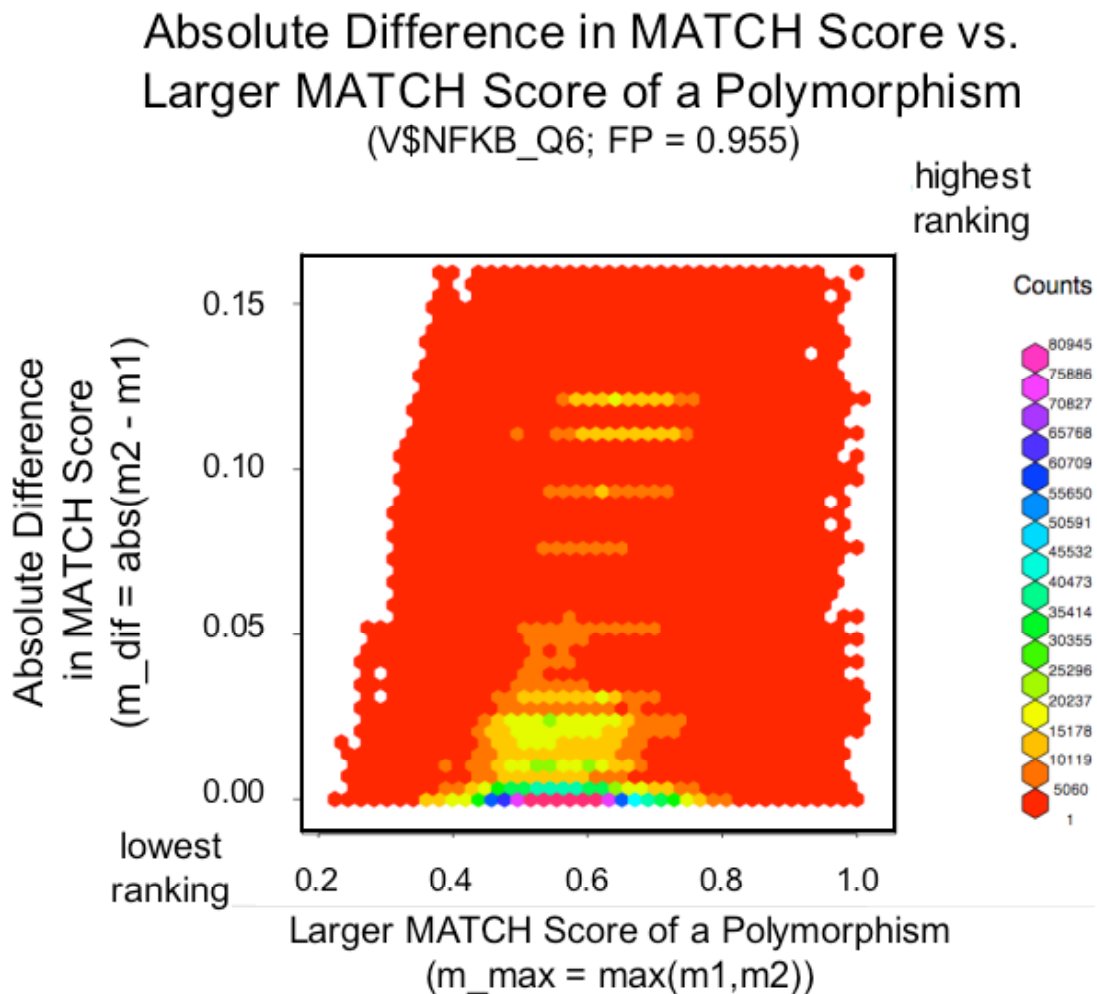


Figure 23 Absolute Difference in MATCH Score vs. Larger MATCH Score of a Polymorphism (the ranked distribution)

This is the ranked distribution for 4,547,844 polymorphisms that have been evaluated with the NF-kB TFBS matrix (V\$NFKB_Q6). Results are ranked in descending order of importance firstly from top to bottom, and secondly from right to left. This graph has the advantage that it shows the ranked distribution of all polymorphisms regardless of their potential score. The “absolute difference in MATCH score” is plotted (y-axis), versus the “larger polymorphism MATCH score” (m_max) (Equation 5 page 20) (x-axis).



This color density map shows the bulk of these polymorphisms are “biological irrelevant” (less than $FP = 0.955$ on the x-axis) (page 30) and most have very low differences in MATCH score (low on the y-axis). The Delta-MATCH Query Tool will return results in descending order of importance, from the highest ranking polymorphisms (upper right quadrant) to the lowest ranking polymorphisms (lower left quadrant).

Figure 24 Potential Score Versus Biological Relevance of a Polymorphic Site

This is the ranked distribution for 950 “biologically relevant” (Definition page 30) polymorphisms that have been evaluated with the NF-kB TFBS matrix (V\$NFKB_Q6). The values in this figure are left-bounded by the equation: $(y = -\log(1.0000001 - X))$. These results are ranked by descending order of importance, **firstly** from top to bottom, and **secondly** from right to left. This graph has the advantage that it only shows the distribution of scores for polymorphisms with an allelic MATCH score (m1 and/or m2) greater than the false positive cutoff score (FP = 0.955). It has the affect of removing “biologically irrelevant” polymorphisms from consideration.

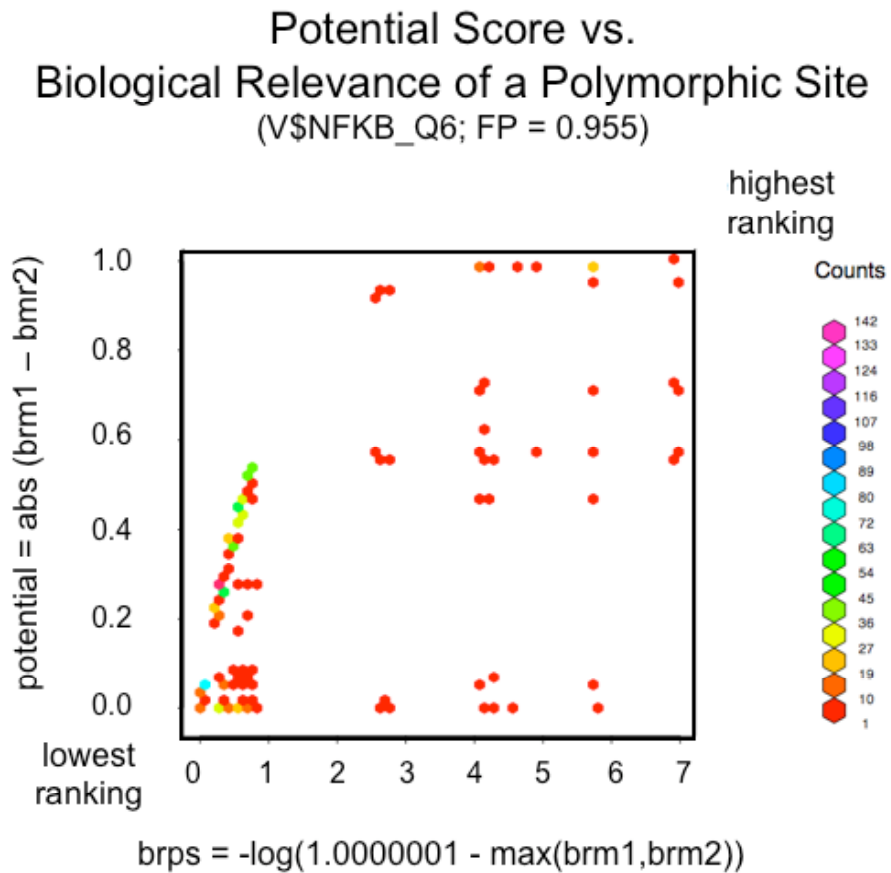


Figure 25 Potential Score Versus Absolute Percent Difference in MATCH Score

These results are ranked in order firstly from top to bottom, and secondly from right to left. This graph has the advantage that it only shows the distribution of scores for polymorphisms (n = 950) with allelic MATCH scores (m1 and/or m2) greater than the false positive cutoff score (FP = 0.955). Note the largest value of m_per is less than or equal to 17 because this is the maximum effect a single base change can have for the V\$NFKB_Q6 matrix (having a polymorphic base aligning to matrix base position 4 or 6).

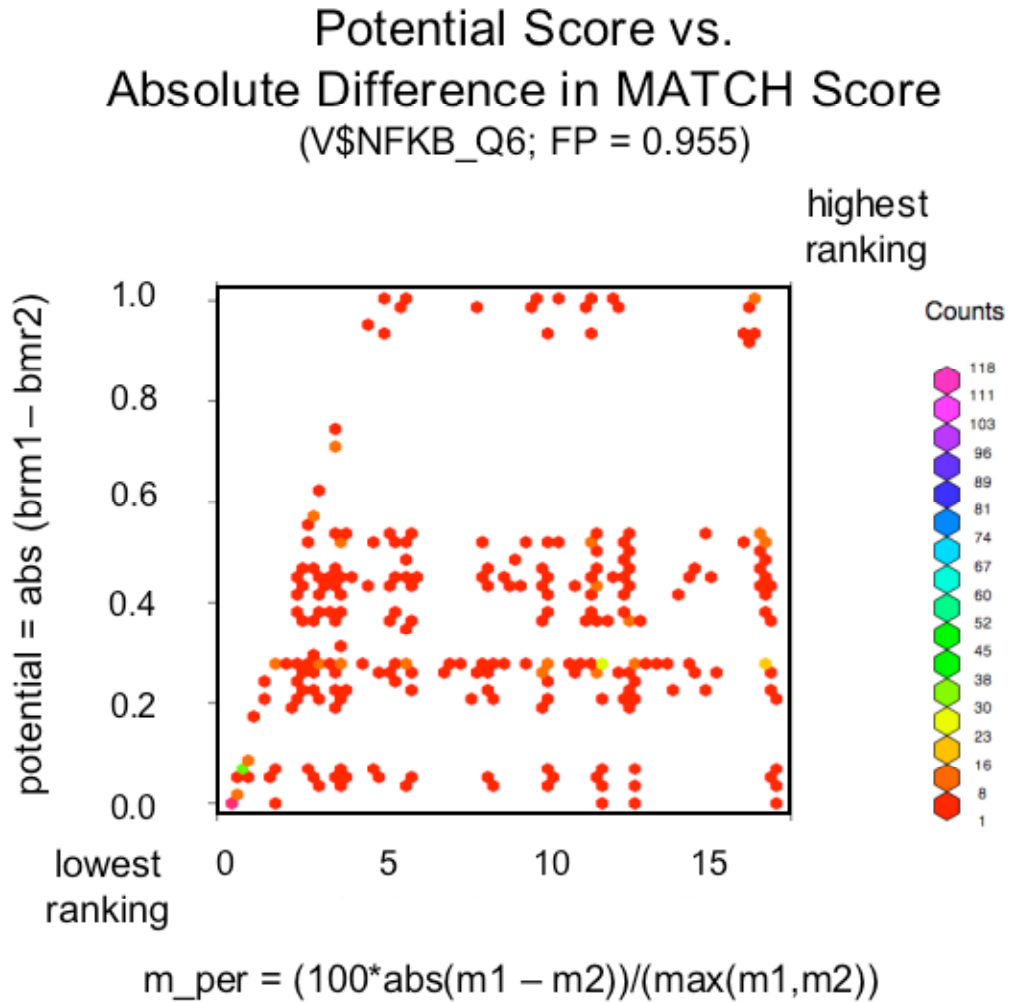


Table 3 Distribution of Delta-MATCH Hits for Matrix Name V\$NFKB_Q6

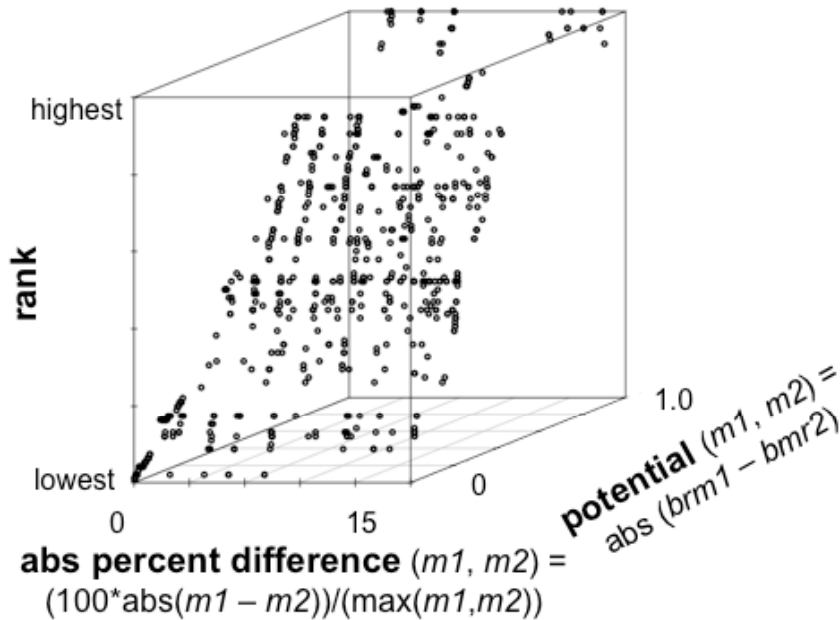
This is the distribution of the 950 "biologically relevant" polymorphisms binned by **potential** scores (Definition page 30) into 0.1 intervals

potential	number of hits >= potential	percent of hits >= potential
0.0	950	100.0
0.1	671	70.6
0.2	671	70.6
0.3	409	43.1
0.4	333	35.1
0.5	179	18.8
0.6	72	7.6
0.7	71	7.5
0.8	61	6.4
0.9	61	6.4
1.0	7	0.7

Figure 26 Rank versus potential score versus absolute percent difference in MATCH score for 950 high-value polymorphisms (3-D plot)

This is a 3-dimensional plot (**rank versus potential score versus absolute percent difference in MATCH score**) for the 950 “biologically relevant” polymorphisms having allelic MATCH scores ($m1$ and/or $m2$) greater than or equal to the V\$NFKB_Q6 false positive cutoff score (FP = 0.955). Those polymorphisms having both a large potential score and a large absolute percent difference in allelic MATCH scores are ranked highest and have the strongest potential to create an allele-specific transcription factor binding site.

**Rank vs. Potential Score vs.
Absolute Percent Difference in MATCH Score
(V\$NFKB_Q6; FP = 0.955)**



1.7.17 How to Calculate the Rareness of a Single Delta-MATCH Result

It is possible estimate the importance of a single Delta-MATCH result by examining the complete distribution of potential scores and calculating a value describing the rareness of an event. The rareness of a result ($HIT_{rareness}$) is the quotient of the number results with a potential score greater than or equal to the specified polymorphism's potential score, divided by the total number of polymorphisms searched.

Equation 13 - Rareness of a Hit ($HIT_{rareness}$)

$$HIT_{rareness} = \frac{(\text{\# of polymorphisms where potential} \geq X_{potential})}{(\text{total \# of polymorphisms searched})}$$

1.7.18 Caveats of the Delta-MATCH Method

It is important to remember that **MATCH** scores, and Delta-MATCH **potential** scores **should not** be directly compared across different TFBS matrixes (mat_ids) because the score distributions for each matrix is unique.

1.7.18.1 Warning: Do Not Compare Absolute Potential Scores Across Different TFBS Matrixes

As described in the original TRANSFAC BIOBASE publications, the distribution of MATCH scores for a given matrix name (mat_id) is dependent on the matrix length (mat_len) and its positional nucleotide diversity. These TFBS position-specific scoring matrixes were created by aligning the DNA sequences of the promoters of many genes that are known to be responsive to a given transcription factor, and then characterizing

the small and highly conserved nucleotide motifs common to the aligned ensemble set. It follows that Delta-MATCH potential scores can only be as accurate as a transcription factor binding site matrix represents a true binding site sequence.

The minimum threshold score for many of the BIOBASE TRANSFAC matrixes have been empirically determined and are presented in Delta-MATCH as the BIOBASE false positive threshold cutoff score (FP). It is recognized that the Delta-MATCH potential scores are highly dependent on the proper estimation of the false positive cutoff scores. It may be the case that a FP score that underestimates the true biologically relevant cutoff might cause a Type -1 error (the enrichment of False Positive predictions in the Delta-MATCH database). Contrariwise, a FP score that overestimates the true biologically relevant cutoff might exclude important polymorphisms from further consideration by a Type-2 error (False Negatives).

Note that in this version of Delta-MATCH, only 550 of the 584 vertebrate BIOBASE TFBS matrixes have their FP cutoff estimated because BIOBASE failed to provide the remaining 34 FP scores with the TRANSFAC database (version 10.2). Predictions for these 34 TFBS matrixes are inaccessible by the Delta-MATCH Query tool, but may become available in a future release if an estimation of the minimum cutoff value of biological relevance can be estimated.

1.8 The Delta-MATCH Algorithm

1.8.1 Polymorphism Selection

In this version of the Delta-MATCH database (version 1.0), 4,547,844 high value candidate polymorphisms (UCSC browser table hg18.snp126.name) have been scored and ranked by the Delta-MATCH algorithm to determine their “potential” to create an allele-specific transcription factor binding site. These high-value polymorphisms were selected if they were either positioned within a 10,000 base pair window (10k upstream + gene + 10k downstream) of any [refSeq](#) gene (UCSC browser table hg18.refGene.name2), or positioned within a region of high conservation anywhere in the human genome (UCSC browser hg18.phastCons17way).

Figure 27 Location of SNPs Evaluated by Delta-MATCH

Location of Single Nucleotide Polymorphisms (SNPs) Evaluated by Delta-MATCH	
region	count
10kb-promoter	647,311
5'-UTR	16,376
exons	212,764
introns	3,415,853
3'-UTR	84,503
10kb-downstream	648,916
conserved	397,802
total evaluated	4,547,844

1.8.2 Polymorphism Exclusions

Polymorphisms were excluded from consideration if they:

- mapped to more than one chromosomal position (mapped ambiguously)
- were not a biallelic nucleotide polymorphism (had more than two allele states)
- were positioned in a microsatellite region
- were positioned in a region of simple repeats (low complexity)
- were positioned in a region of a large insertion/deletion

1.8.3 Creating Double-Stranded DNA Allele Sequences

During the scoring algorithm, the 61 base pairs of double-stranded DNA sequence surrounding each polymorphism was retrieved from, and oriented to, the plus strand of the UCSC human genome (build 36) using a [DAS](#) DNA sequence retrieval web tool. For example this URL:

<http://genome.ucsc.edu/cgi-bin/das/hg18/dna?segment=chr20:50396817,50396877>)

will retrieve the 61 bases surrounding the SNP rs6013444.

Figure 28 The 61 Base Pairs of DNA Sequence Surrounding rs6013444 in the UCSC Genome Browser (Mar. 2006 Assembly)

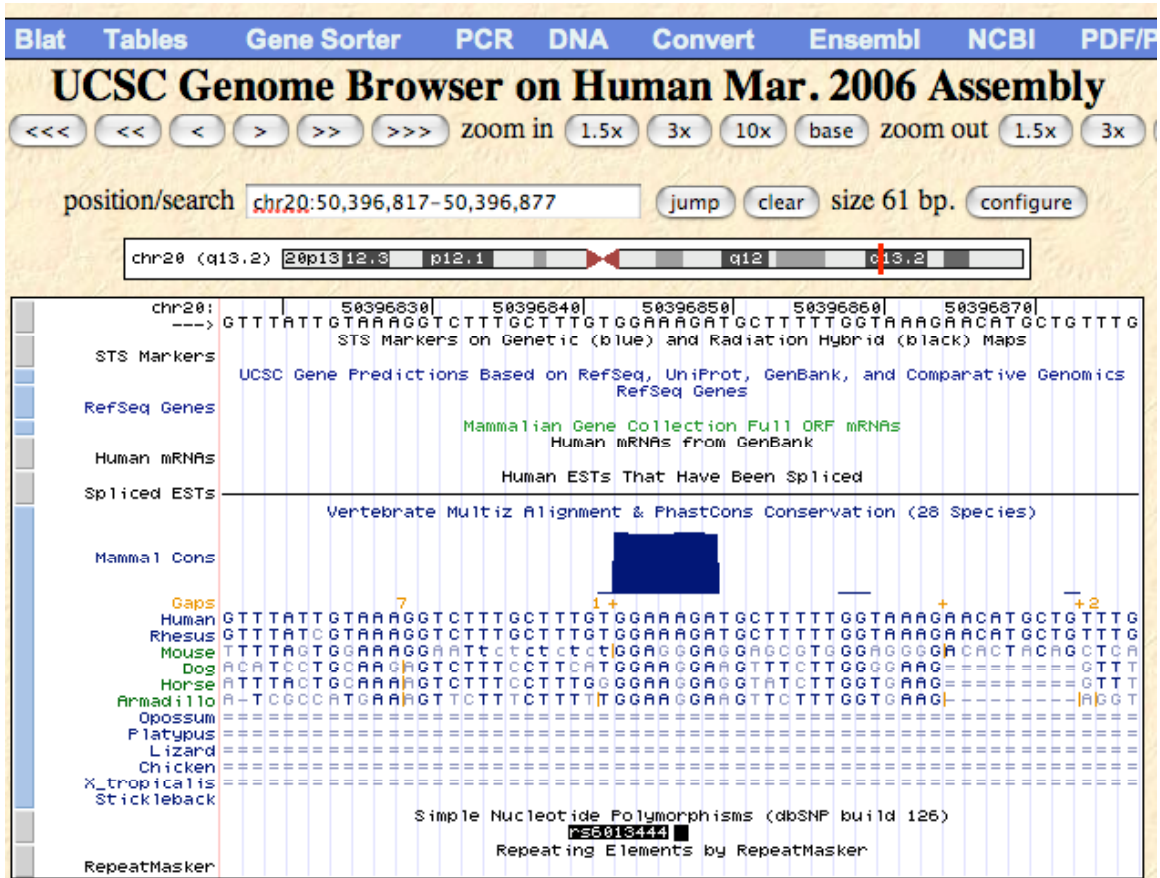
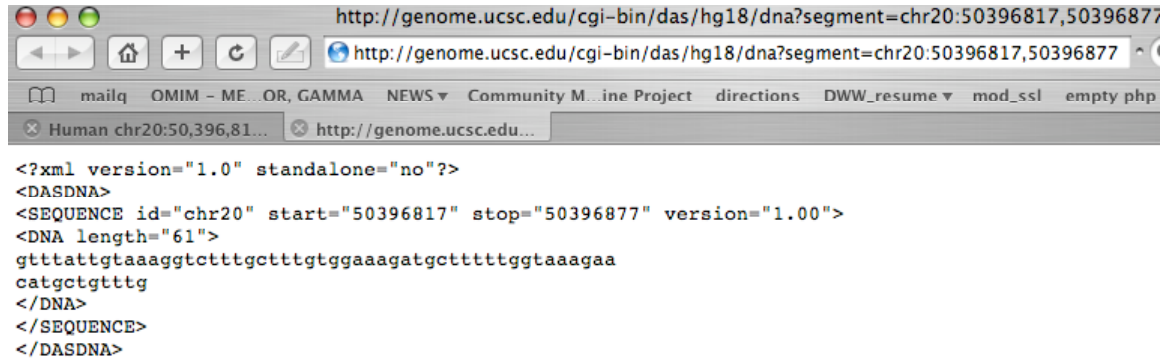


Figure 29 The DAS DNA Sequence Retrieval Web Tool (retrieving the 61 bp sequence surrounding rs6013444)



```
<?xml version="1.0" standalone="no"?>
<DASDNA>
<SEQUENCE id="chr20" start="50396817" stop="50396877" version="1.00">
<DNA length="61">
gtttattgtaaaaggtctttgctttgtggaagatgctttttggtaaagaa
catgctgtttg
</DNA>
</SEQUENCE>
</DASDNA>
```

After retrieving the dsDNA sequence, two alternate 61 base pair double-stranded allelic sequences were created, one for allele 1 and another for allele 2. In the Delta-MATCH database allele 1 is always the base referenced by the UCSC genome browser (hg18.snp126.refUCSC) and is commonly referred to as the major allele. Allele 2 is the alternately observed allele at the polymorphic site. Please note that the double-stranded DNA sequences representing allele 1 and allele 2 were always created and polarized to the plus sense strand of the genome. In other words, care was taken to make sure that when an **rsnumber** (hg18.snp126.name) was identified on the minus strand (hg18.snp126.strand = "-") of a human chromosome, the sequence of allele 1 reflected the plus sense strand of the human genome centered around the polymorphic allele, and the base on the plus strand was the "reverse complement" of the UCSC (hg18.snp126.refUCSC) reference base.

1.8.4 Computing the Highest MATCH Scores

The “highest calculated MATCH score” for each polymorphism allele, for each transcription factor, was identified. For each polymorphic allele, a separate MATCH score was calculated after aligning each position of each matrix with the position of a polymorphic allele sequence, along both the plus sense and minus sense strands. (Figure 30 page 73).

In order to minimize the computational effort, MATCH scores were calculated only at those positions where the matrix overlapped the polymorphic base. Thus identifying the “highest MATCH” score required calculating from as few as 12 ($\text{mat_len} = 6$) to as many as 60 ($\text{mat_len} = 30$) independent MATCH scores per SNP allele, depending on the length of a TFBS matrix.

Equation 14 - Number of Calculations on the Plus Strand ($\text{Number}_{\text{plus}}$)

$$\text{Number}_{\text{plus}} = \text{length of matrix}$$

Equation 15 - Number of Calculations on the Minus Strand ($\text{Number}_{\text{minus}}$)

$$\text{Number}_{\text{minus}} = \text{length of matrix}$$

Equation 16 - Number of Calculations Required to Find Highest Match ($\text{Number}_{\text{total}}$)

$$\text{Number}_{\text{total}} = 2 \times \text{length of matrix}$$

When the matrix length is 6:

$$\text{Number}_{\text{total}} = (2 \times 6) = 12 \text{ (fewest)}$$

When the matrix length is 30:

$$\text{Number}_{\text{total}} = (2 \times 30) = 60 \text{ (most)}$$

1.8.5 Recording Delta-MATCH Scores

The “strand” (s1 and s2), the “relative offset position” (p1 and p2), and the magnitude of the “highest calculated MATCH score” (m1 and m2) for each polymorphic allele, for each TFBS matrix, were recorded into the Delta-MATCH database.

1.8.5.1 Definition - s1 and s2

This is the strand (“+” or “-”) along where a TFBS matrix had its “highest calculated MATCH score” (m1 and m2) for allele 1 and allele 2.

1.8.5.2 Definition - p1 and p2

This is the **relative offset position** of the “highest calculated MATCH score” for allele 1 and allele 2. This is the leftmost position of the TFBS matrix alignment relative to the position of the polymorphic base after aligning a matrix to the plus sense strand of the human genome build 36, March 2006.

1.8.5.3 Definition - m1 and m2

This is the magnitude of the “highest calculated MATCH score” for allele 1 and allele 2. The range of a m1 and m2 are from 0.0 to 1.0 (Figure 202 page 402).

1.8.6 Identifying the Highest MATCH Score for an Allele (Exhaustive Search)

In the **first half** of the exhaustive search, the **first iteration** MATCH score was calculated for a given transcription factor matrix by aligning the leftmost position of the matrix with a position on the **plus** sense strand of the 61-mer so that the last (rightmost) position of the matrix overlaid the exact position of the polymorphic allele (Figure 30 page 73). For the **second iteration**, the matrix was repositioned one base position to the right so that the last position of the matrix aligned on the **plus** sense strand of the 61-mer exactly one base to the right of the polymorphic allele, and the MATCH score for this second iteration was recalculated. **Subsequent iterations** calculated MATCH scores after repositioning the matrix consecutively one base to the right on the **plus** sense strand. The first half of the search concluded after calculating the MATCH score where the first position of the matrix aligned on the **plus** sense strand exactly to the position of the polymorphic base. The **second half** of the exhaustive search followed exactly like the first half of the search, except that all MATCH scores were calculated after aligning to the sequence at positions relative to the **minus** sense strand of the 61-mer double-stranded DNA alleles (the reverse complement sequence of the plus sense strand).

1.8.7 Why Was a 61 Base Pair Length of Sequence Chosen?

The 61 base pair length of alleles was chosen specifically to allow an exhaustive search by the longest vertebrate transcription factor matrix. The longest matrix in the TRANSFAC database is 30 base pairs long (mat_id = V\$HOX13_01, and V\$PAX4_04). Retrieving the 30 base pairs upstream of the leftmost, and downstream of the rightmost positions of a polymorphic site (relative to the plus strand) assured that every position of

1.9 The Delta-MATCH Database

The Delta-MATCH database is a collection of MySQL database tables that can be cross-referenced with the Delta-MATCH Query Tool. Some of these tables have been adopted from public resources such as [UCSC Genome Browser](#). Others have been developed from supplementary data tables from published literature. The details of the Delta-MATCH database architecture, and scripts that may reconstruct some of these accessory resources are available in the Appendix (page 408).

1.9.1 How Many Results Are In the Delta-MATCH Database?

Exactly 4,547,844 polymorphisms have been searched against 550 (high and low quality) vertebrate transcription factor matrixes. From these searches, there were 6,206,823 “**Delta-MATCH hits**” identified and recorded. All of these calculated scores are accessible with the Delta-MATCH Query Tool.

1.9.1.1 Definition - Delta-MATCH hit or result

A result is any instance when a polymorphism is scored against a transcription factor matrix using the Delta-MATCH algorithm and at least one of the two allelic MATCH scores (m1 or m2) is greater than or equal to the false positive threshold cutoff (FP page 31) for that matrix. Each of the 550 TFBS matrixes has a unique list of ranked results.

It is noteworthy that there are a disproportionate number of **hits** in the Delta-MATCH Database derived from the 183 “low quality” matrixes (hits = 4,682,078) when compared with the 367 “high quality” matrixes (hits = 1,524,745). This means that 75.4% of the

Delta-MATCH results can be filtered away by requiring “high quality” matrix results (page 104).

1.9.2 No Correlation Between Matrix Length and Number of Delta-MATCH Hits

There does not appear to be a strong correlation between TFBS matrix length and the number of Delta-MATCH hits (correlation coeff = -0.67). There are relatively few hits for matrixes longer than 26 base pairs (with the exception where `mat_len` \geq 29 and the quality is “low”). More generally, it can be said that there are proportionally more hits for shorter length low quality matrixes, relative to the high quality matrixes (Figure 32 page 78).

Table 4 Distribution of Delta-MATCH Hits and Counts for High and Low Quality

Matrixes

	num_hits	num_hits	num_hits	count	count	count
quality	all	high	low	all	high	low
sum	6,206,823	1,524,745	4,682,078	550	183	367
mat_len						
6	129,757	0	129,757	8	8	0
7	440,453	26,283	414,170	17	15	2
8	466,381	52,354	414,027	43	21	22
9	458,683	12,591	446,092	36	24	12
10	337,392	90,224	247,168	58	25	33
11	360,331	48,267	312,064	42	17	25
12	569,356	265,247	304,109	61	15	46
13	548,642	85,675	462,967	49	18	31
14	435,482	156,574	278,908	53	9	44
15	753,734	198,551	555,183	41	9	32
16	108,030	69,307	38,723	37	3	34
17	319,615	209,654	109,961	12	3	9
18	445,874	112,437	333,437	25	6	19
19	158,288	10,792	147,496	12	4	8
20	1,185	1,185	0	6	0	6
21	283,172	83,238	199,934	15	4	11
22	2,852	2,852	0	8	0	8
23	118,914	21,126	97,788	4	1	3
24	46,566	46,566	0	6	0	6
25	1,112	1,112	0	4	0	4
26	0	0	0	1	0	1
27	803	803	0	5	0	5
28	595	595	0	3	0	3
29	28,981	28,981	0	2	0	2
30	190,625	331	190,294	2	1	1
corr. coef.	-0.674442	-0.345665	-0.697968			

Figure 31 Number of Delta-MATCH Results vs. Matrix Length for 4.5 Million Hits

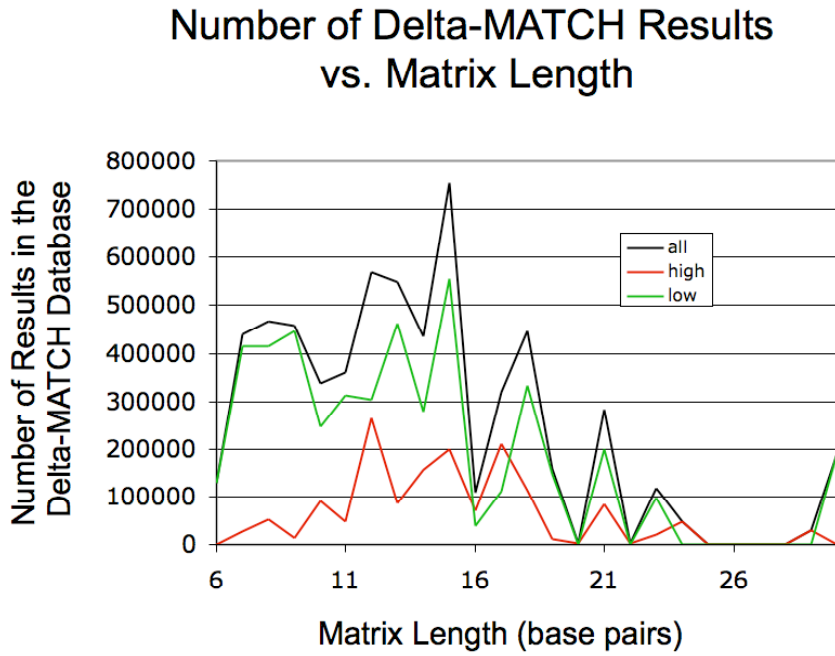


Figure 32 Count of Matrixes vs. Matrix Length For High and Low Quality Matrixes

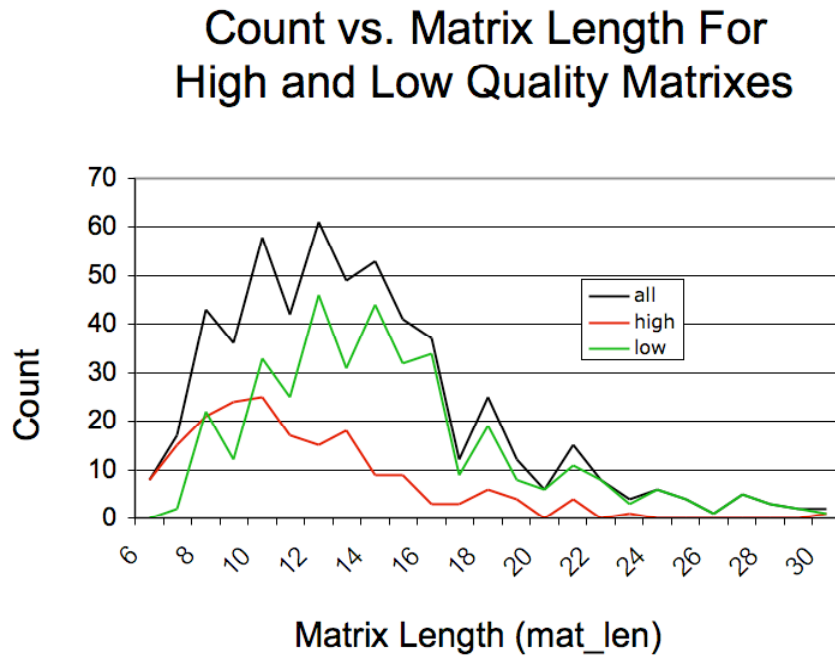


Figure 33 The Delta-MATCH Website, <http://deltamatch.org>

At the Delta-MATCH website (<http://deltamatch.org>) it is possible to query the Delta-MATCH database to identify lists of polymorphisms that are predicted to create allele-specific transcription factor binding sites. Online queries may be submitted using a series of radio buttons, drop-down menus, and text fields.

Identify Lists of Allele-Specific Transcription Factor Binding Sites at <http://deltamatch.org> (Search by Transcription Factors, rsnumbers, and Gene Names)

Delta-MATCH™ HOME ABOUT EASY MODE **EXPERT MODE** TUTORIAL DOWNLOADS AUTHOR

Expert Mode

STEP 1 - SELECT MATRIX NAMES

- 1 - SINGLE MATRIX NAME
- 2 - LIST OF MATRIX NAMES
- 3 - FACTOR NAME
- 4 - TISSUE-SPECIFIC NAMES
- 5 - ALL TF NAMES

STEP 2 - ADD CRITERIA

- MIN POTENTIAL SCORE
- TOP MOST SIGNIFICANT HITS
- MATRIX QUALITY
- SORT RESULTS TABLE
- SEARCH BY RSNUMBERS
- SEARCH BY GENE NAMES
- MATRIX DETAILS
- POSITION DETAILS
- CHROMOSOME
- POSITION RANGE
- STRAND
- GENOMIC REGIONS
- BONFERRONI CORRECTION

STEP 1 - Select Matrix Names

Choose one of the five 'Primary Matrix Selection Buttons' (left), and adjust (right). View the details of the '500 Vertebrate Transcription Factor Matrix'.

1 - Single Transcription Factor Matrix Name

Select a 'Single Transcription Factor Matrix Name' (n=550)

VSNFKB_Q6 (950) (mat_id)

2 - List of Transcription Factor Matrix Names

Hand-type a comma-separated 'List of Transcription Factor Matrix Names' (n=550)

VSNFKB_Q6,VSNFKB_C (mat_id) (1024 chars)

3 - Transcription Factor Name

Select transcription factor matrix names by a 'Transcription Factor Name' (n=550)

NF-kappaB (factor)

4 - Tissue-Specific Transcription Factor Names

Select transcription factor matrix names by a 'Tissue Type' (n=550)

immune_cell_specific (n=113)

5 - All Transcription Factor Matrix Names

Select all matrix names in the database (n=550)

Search By rsnumbers

Either limit results by a comma-separated list of dbSNP rsnumbers (1024 max)

rs5743836,rs6031444 rsnumbers

Or upload list of rsnumbers in a plain text file

'rsnumber filename' download example file
(one rsnumber per row, 10,000 rsnumbers max)

NOTE - The 'rsnumber' text field will be ignored if an 'rsnumber filename' is specified

Search additional bases upstream and downstream of specified rsnumber

'rsnumber Window'

2000 Include other rsnumbers within this many bases

Search By Gene Names

'Search for gene without returning results' (MOCK SEARCH) [REF1, 1]

UNCHECK this box to perform 'real' search after your gene names are verified

Limit results by a comma-separated list of 'Gene Names' (exact text match)

JPH2,PLAT,TLR9 'Gene Name'

refGene (name2, name) 'UCSC hg18 Table Name'

name2 'Field Name' download help file

Search more bases upstream and downstream of specified genes

'Gene Window'

2000 Select the number of additional bases

Figure 34 The Delta-MATCH website hosts tutorials, examples, and downloadable data tables.

Users may choose to view Delta-MATCH tables as a custom track in the UCSC human genome browser.

The Delta-MATCH Website Hosts Tutorials, Examples and Downloadable Data Tables

The screenshot displays the Delta-MATCH website interface, organized into four main columns:

- Tutorial:** Contains a section titled "The Delta-MATCH Tutorial" with a "Download the Delta-MATCH Tutorial" button. Below it is a "Quick Start Instructions" section with three steps: "STEP 1 - Select Matrix", "STEP 2 - Add Rest of Parameters", and "STEP 3 - Press the Run Button".
- Examples:** Features a "Download the complete description" link. It lists "Over 40 'Delta-MATCH Example' tables that may be created with the Delta-MATCH tool, each with a unique parameter." It also provides a list of output files: log.html, report.html, table.html, table.txt, and table.xml. A "Proof of Principle Example - alpha2-Heremans" section is also visible, describing a specific example.
- Downloads:** Includes sections for "UCSC Browser Tracks" (with a link to "Learn to create a custom 'UCSC Genome Browser Track'"), "Databases" (with a link to "A list of all of the the databases and tables used in the Delta-MATCH tool. See the 'Databases Page.'"), "Scripts" (with a link to "A selection of scripts are available to download and use to create, populate and access the Delta-MATCH database tables."), and "Tutorial" (with a link to "The tutorial details the methodology of the query tool. It includes a 'Quick Start' section that lists the parameters of the query tool. It includes a 'Download the Tutorial' link that may want to create and submit.").
- UCSC Browser Tracks:** Shows a list of tracks under the heading "A". The tracks listed include: VSACAAT_B (0), VSAFP1_Q6 (0), VSAHR_01 (0), VSAHR_Q5 (0), VSAHRARNT_01 (0), VSAHRARNT_02 (0), VSAHRHIF_Q6 (0), VSAIRE_01 (0), VSAIRE_02 (0), VSALPHACP1_01 (0), VSALX4_01 (0), VSAMEF2_Q6 (0), VSAML_Q6 (0), VSAML1_01 (0), VSAML1_Q6 (0), VSAP1_01 (0), VSAP1_C (1321), VSAP1_Q2 (3855), VSAP1_Q2_01 (15446), VSAP1_Q4 (6986), and VSAP1_Q4_01 (10689).

1.9.3 The Delta-MATCH Query Tool Search Engine (version 1.0)

The Delta-MATCH Query Tool is a PHP web-based tool that allows users to identify from a database of over 4.5 million human SNPs, those polymorphisms (**rsnumbers**) that have a strong “potential” to create an allele-specific transcription factor binding site.

1.9.4 Creating a Delta-MATCH Query

Users may create a query by selecting the appropriate **radio buttons, check boxes and drop-down menus** to select the best constellation of parameters before searching the database by pressing of the “Submit” button.

Figure 35 List of selectable parameters at the Delta-MATCH website

Delta-MATCH Query Tool Parameters

Transcription Factor	Optional Parameters	
1 - SINGLE MATRIX NAME	MIN POTENTIAL SCORE	MINIMUM TOTAL NUMBER
2 - LIST OF MATRIX NAMES	TOP MOST SIGNIFICANT HITS	HUGO NAMES
3 - FACTOR NAME	MATRIX QUALITY	REFLINK
4 - TISSUE-SPECIFIC NAMES	SORT RESULTS TABLE	DISTANCE DETAILS
5 - ALL TF NAMES	SEARCH BY RSNUMBERS	GENE ONTOLOGY
	SEARCH BY GENE NAMES	AFFYMETRIX
	MATRIX DETAILS	ILLUMINA
	POSITION DETAILS	HAPMAP
	CHROMOSOME	HIV-1 CANDIDATE GENES
	POSITION RANGE	COPY NUMBER VARIATION
	STRAND	PREMOD MODULES
	GENOMIC REGIONS	UCSC RSNUMBER DETAILS
	BONFERONNI CORRECTION	

1.9.5 Creating a Query Using the Delta-MATCH Query Tool

Three steps are required to create a query using the Delta-MATCH Query Tool

- STEP 1 - Select Matrix Names
- STEP 2 - Add Restriction Criteria
- STEP 3 - Press the Submit Button

By default, the Delta-MATCH Query Tool will attempt to return the complete list of polymorphisms (rsnumbers) for each of the selected transcription factor matrix names in STEP 1.

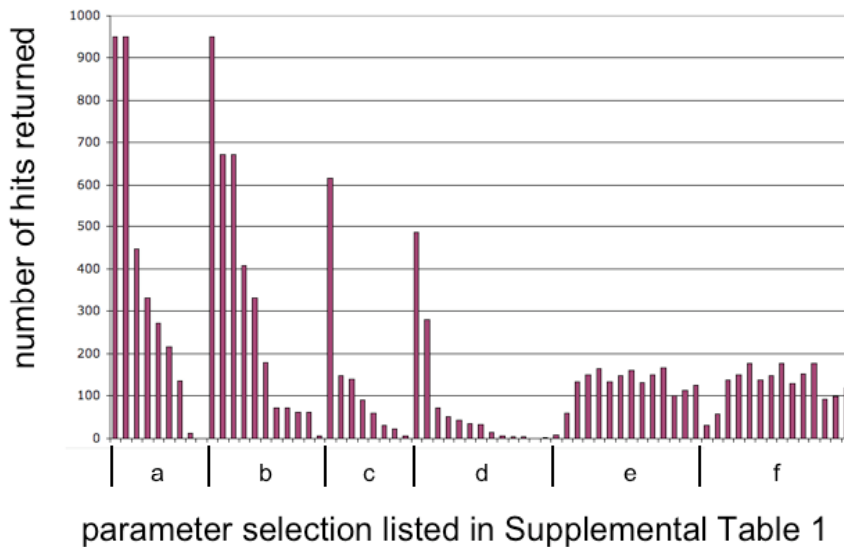
In STEP 2, users may select from a list of greater than 20 additional selection criteria using radio buttons, drop down menus, and text fields. Only those rsnumbers that satisfy the complete set of criteria will be returned.

For example, a maximum of 950 rsnumbers may be returned for the NK-kB (V\$NFKB_Q6) transcription factor matrix. In other words, after searching 4.5 million polymorphisms, only 950 of these were considered “biologically relevant” and may have the “potential” create allele-specific NF-kB binding sites (page 55). The following figure (Figure 36 page 84) and table (Figure 37 page 85) show the number of polymorphisms that will be returned when additional parameters are selected using the V\$NFKB_Q6 matrix.

Figure 36 The number of Delta-MATCH hits returned is dependent on the parameters selected

Users may search the Delta-MATCH database and identify polymorphisms with allelic MATCH scores (m1 and/or m2) greater than or equal to the matrix cutoff threshold using up to 550 vertebrate transcription factor binding site matrixes. When additional parameters are selected, only those rsnumbers that satisfy all of the selected criteria (the intersection) are returned. The Delta-MATCH Query Tool will return up to 950 rsnumbers for the NFKB transcription factor matrix (V\$NFKB_Q6) or a subset of these depending on what parameters are selected: (a = Table 1 rows 1 through 9) minor allele frequency; (b = rows 10 through 20) potential score; (c = rows 21 through 28) polymorphism location; (d = rows 29 through 40) assorted criteria; (e = rows 41 through 55) Affymetrix; (f = 56 through 69) Illumina (see Supplemental Table 1 for a detailed description of each set of lettered parameters).²

The number of Delta-MATCH Hits Returned Is Dependent on the Parameters Selected



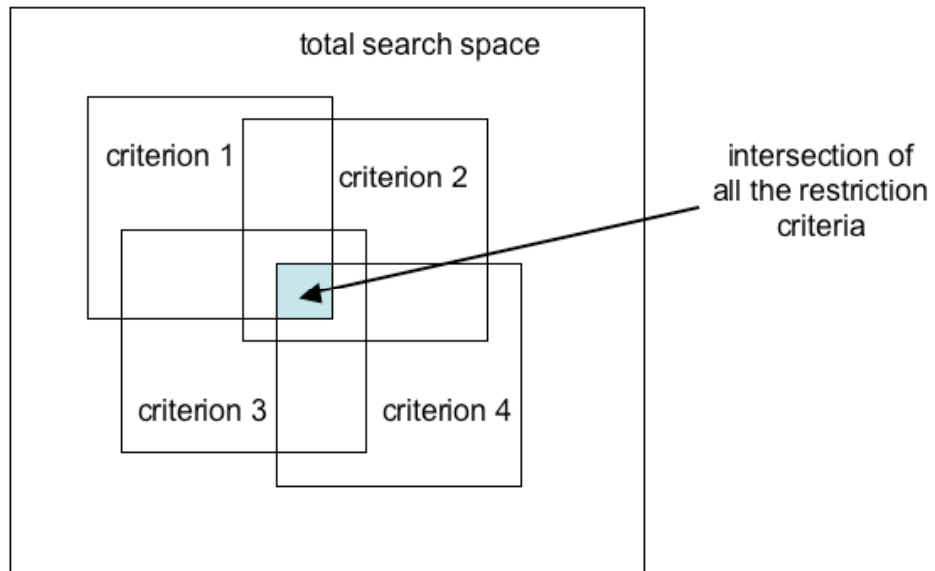
² Supplemental Table 1 (Figure 37 page 85)

Figure 37 Number of Hits Returned vs. Parameter (Description)

Order	Parameter (Description)	Number
1	total number of hits (rsnumbers) for the matrix V\$NFKB_Q6	950
2	with a MAF >= 0.0	950
3	with a MAF >= 0.01	447
4	with a MAF >= 0.1	334
5	with a MAF >= 0.2	273
6	with a MAF >= 0.3	217
7	with a MAF >= 0.4	138
8	with a MAF >= 0.5	12
9	with a MAF >= 0.6	1
10	with potential >= 0.0	950
11	with potential >= 0.1	671
12	with potential >= 0.2	671
13	with potential >= 0.3	409
14	with potential >= 0.4	333
15	with potential >= 0.5	179
16	with potential >= 0.6	72
17	with potential >= 0.7	71
18	with potential >= 0.8	61
19	with potential >= 0.9	61
20	with potential >= 1.0	7
21	located in an intron	615
22	located within 10kb downstream of a gene	150
23	located within 10kb upstream of a gene	142
24	located in a region of conservation (phastcons17)	89
25	located in an exon	59
26	located in a coding region	30
27	located in an 3' untranslated region	23
28	located in an 5' untranslated region	6
29	associated with a HUGO gene name (REF)	486
30	located in a region of known copy number variation (REF)	281
31	located next to a HUGO gene that has a Gene Ontology term 'transcription' (REF)	71
32	located in a PReMod module region (REF)	52
33	located on chromosome 8	42
34	located within 2000 bases of a known transcriptional start site, and 2000 bases of a known translational start site	35
35	located next to a HUGO name that has a term 'kinase'	33
36	associated as a candidate polymorphism for HIV-1 progression	15
37	number of hits with a Bonferonni-adjusted rareness <= 0.005	7
38	the top 5 highest ranked results	5
39	located within a 2000 bp window around the genes TLR9, JPH2 or PLAT	4
40	located between base pair 128,100,000 and 128,700,000	1
41	located on an Affymetrix 10k SNP CHIP	2
42	located on an Affymetrix 100k SNP CHIP	9
43	located on an Affymetrix 500k SNP CHIP	59
44	located on, or in LD (pop =European; D' = 1.0; r ² = 1.0) with a rsnumber on, the Affymetrix 500k SNP CHIP	134
45	located on, or in LD (pop =European; D' = 1.0; r ² = 0.9) with a rsnumber on, the Affymetrix 500k SNP CHIP	152
46	located on, or in LD (pop =European; D' = 1.0; r ² = 0.8) with a rsnumber on, the Affymetrix 500k SNP CHIP	166
47	located on, or in LD (pop =Chinese; D' = 1.0; r ² = 1.0) with a rsnumber on, the Affymetrix 500k SNP CHIP	134
48	located on, or in LD (pop =Chinese; D' = 1.0; r ² = 0.9) with a rsnumber on, the Affymetrix 500k SNP CHIP	150
49	located on, or in LD (pop =Chinese; D' = 1.0; r ² = 0.8) with a rsnumber on, the Affymetrix 500k SNP CHIP	162
50	located on, or in LD (pop =Japanese; D' = 1.0; r ² = 1.0) with a rsnumber on, the Affymetrix 500k SNP CHIP	132
51	located on, or in LD (pop =Japanese; D' = 1.0; r ² = 0.9) with a rsnumber on, the Affymetrix 500k SNP CHIP	152
52	located on, or in LD (pop =Japanese; D' = 1.0; r ² = 0.8) with a rsnumber on, the Affymetrix 500k SNP CHIP	167
53	located on, or in LD (pop =African; D' = 1.0; r ² = 1.0) with a rsnumber on, the Affymetrix 500k SNP CHIP	101
54	located on, or in LD (pop =African; D' = 1.0; r ² = 0.9) with a rsnumber on, the Affymetrix 500k SNP CHIP	112
55	located on, or in LD (pop =African; D' = 1.0; r ² = 0.8) with a rsnumber on, the Affymetrix 500k SNP CHIP	127
56	located on the Illumina Hap300 SNP CHIP	31
57	located on the Illumina Hap550 SNP CHIP	57
58	located on, or in LD (pop =European; D' = 1.0; r ² = 1.0) with a rsnumber on, the Illumina Hap550 SNP CHIP	140
59	located on, or in LD (pop =European; D' = 1.0; r ² = 0.9) with a rsnumber on, the Illumina Hap550 SNP CHIP	152
60	located on, or in LD (pop =European; D' = 1.0; r ² = 0.8) with a rsnumber on, the Illumina Hap550 SNP CHIP	177
61	located on, or in LD (pop =Chinese; D' = 1.0; r ² = 1.0) with a rsnumber on, the Illumina Hap550 SNP CHIP	140
62	located on, or in LD (pop =Chinese; D' = 1.0; r ² = 0.9) with a rsnumber on, the Illumina Hap550 SNP CHIP	150
63	located on, or in LD (pop =Chinese; D' = 1.0; r ² = 0.8) with a rsnumber on, the Illumina Hap550 SNP CHIP	178
64	located on, or in LD (pop =Japanese; D' = 1.0; r ² = 1.0) with a rsnumber on, the Illumina Hap550 SNP CHIP	131
65	located on, or in LD (pop =Japanese; D' = 1.0; r ² = 0.9) with a rsnumber on, the Illumina Hap550 SNP CHIP	154
66	located on, or in LD (pop =Japanese; D' = 1.0; r ² = 0.8) with a rsnumber on, the Illumina Hap550 SNP CHIP	177
67	located on, or in LD (pop =African; D' = 1.0; r ² = 1.0) with a rsnumber on, the Illumina Hap550 SNP CHIP	93
68	located on, or in LD (pop =African; D' = 1.0; r ² = 0.9) with a rsnumber on, the Illumina Hap550 SNP CHIP	98
69	located on, or in LD (pop =African; D' = 1.0; r ² = 0.8) with a rsnumber on, the Illumina Hap550 SNP CHIP	121

Figure 38 Delta-MATCH Returns the Intersection of Restriction Criteria

The Delta-MATCH Query Tool Returns
Only the Results that Meet the
Intersection of All the Restriction Criteria



1.9.6 Easy Mode vs. Expert Mode

The Delta-MATCH Query Tool can be run in [Easy Mode](#), or [Expert Mode](#). Easy Mode provides enough parameters for the basic user, while the Expert Mode allows the user to ask very complicated queries using an expanded set of additional selection criteria.

1.9.7 Easy Mode Selections

- Single Transcription Factor Matrix Name
- List of Transcription Factor Matrix Names
- Transcription Factor Name
- Tissue-Specific Transcription Factor Names
- All Transcription Factor Names
- Minimum Potential Score
- Top Most Significant Hits
- Matrix Quality
- Sort Results Table
- Search By rsnumbers
- Search By Gene Names

1.9.8 Expert Mode Additional Selections

- Show the Matrix Details
- Show the Position Details
- Chromosome
- Position Range
- Strand
- Genomic Regions
- Bonferonni Correction
- Minimum Total Number of Hits
- HUGO Names
- Reflink
- Distance From txStart or cdStart
- Gene Ontology
- Affymetrix
- Illumina
- HapMap
- HIV-1 Candidate Genes
- Copy Number Variation
- PReMod Modules
- UCSC rsnumber Details

Figure 39 Delta-MATCH Easy Mode Input Page

STEP 1 - Select Matrix Names

Choose one of the five 'Primary Matrix Selection Buttons' (left), and adjust its corresponding drop-down menu or text field (right). View the details of the '550 Vertebrate Transcription Factor Matrix Names' [REF]

[Goto submit](#)

1 - Single Transcription Factor Matrix Name

Select a 'Single Transcription Factor Matrix Name' (n=550)
 (mat_id)

2 - List of Transcription Factor Matrix Names

Hand-type a comma-separated 'List of Transcription Factor Matrixes Names'
 (mat_id) (1024 chars)

3 - Transcription Factor Name

Select transcription factor matrix names by a 'Transcription Factor Name' (n=351)
 (factor)

4 - Tissue-Specific Transcription Factor Names

Select transcription factor matrix names by a 'Tissue Type'

5 - All Transcription Factor Matrix Names

Select all matrix names in the database (n=550)

STEP 2 - Add Restriction Criteria

Select additional Primary Boxes (left), and adjust the parameters of those checked boxes (right).
NOTE - Parameters will only be active if the primary box on the left is checked.

[Back to top](#)

Minimum Potential Score

Select a 'Minimum Potential Score' (potential)
 (0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix

Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (**qual**) [[REF](#)]

('high = 1', 'low = 0')

Sort Results Table

Sort the results table by

(asc = ascending, desc = descending)

[Back to top](#)

Search By rsnumbers

Either limit results by a comma-separated list of dbSNP rsnumbers (1024 chars max)

[[REF1](#), [REF2](#)]

'rsnumbers'

Or upload list of rsnumbers in a plain text file

'rsnumber filename' [download example file](#)

(one rsnumber per row, 10,000 rsnumbers max)

NOTE - The 'rsnumber' text field will be ignored if an 'rsnumber filename' is uploaded

Search additional bases upstream and downstream of specified rsnumbers

'rsnumber Window'

Include other rsnumbers within this many bases

[Back to top](#)

Search By Gene Names

'Search for gene without returning results' (*MOCK SEARCH*) [[REF1](#), [REF2](#)]

UNCHECK this box to perform 'real' search after your gene names are verified

Limit results by a comma-separated list of 'Gene Names' (exact text match, 5 max)

'Gene Names'

'UCSC hg18 Table Name'

'Field Name' [download help file](#)

Search more bases upstream and downstream of specified genes

'Gene Window'

Select the number of additional bases

STEP 3 - Press Submit

[Back to top](#)

Submit (a maximum of 1,500 results will be returned)

Figure 40 Additional Parameter Fields Included in the Expert Mode

Show the Matrix Details

[Back to top](#)

Show the matrix details [\[REF\]](#)

([count_ge_potential](#), [mat_count](#), [frequency](#), [factor](#), [factor_description](#), [qual](#), [mat_len](#))

Minimum Matrix Length

Limit searches to those matrixes with minimum length ([mat_len](#))

([mat_len](#) >= x)

Show the Position Details

Show the position and strand details [\[REF\]](#)

([p1_window](#), [p2_window](#), [p1](#), [p2](#), [s1](#), [s2](#))

Chromosome

[Back to top](#)

Limit results to a chromosome [\[REF\]](#)

([chrom](#))

Position Range

Limit results between two positions [\[REF\]](#)

Enter lowest base ([chrStart](#) >= x)

Enter highest base ([chrStart](#) <= x)

Strand

Limit matrix hits to a DNA strand [\[REF\]](#)

([strand](#))

[Back to top](#)

Genomic Regions

Limit results to include rsnumbers positioned in these genomic regions of refSeq genes [\[REF\]](#)

- up10k** (647,311)
- phastconsElements17way** (397,802)
- utr5** (16,376)
- coding** (113,832)
- down10k** (648,916)
- exons** (212,764)
- introns** (3,415,853)
- utr3** (84,503)
- all** (11,647,909)

or and ("and" IS VERY SLOW!)

[Back to top](#)

Bonferonni Correction

Limit results by 'Minimum Bonferonni-Adjusted Rareness' (**bonferonni**)

bonferonni = rareness*(number of returned hits)

(**bonferonni** <= x)

NOTE - must have 'Matrix Details' checked to see this column

Minimum Total Number of Hits

Limit results to rsnumbers with a minimum 'total number of hits'

This is the sum number of hits for an rsnumber in the database

number_hits >= x)

HUGO Names

Show the HUGO names of the genes associated with each rsnumber [\[REF1, REF2\]](#)

(hugo_name)

Limit results to rsnumbers next to known HUGO genes

Download the rsnumber to hugo name file

(WARNING 48.4 Mb, right-click and 'download file') [SNP-Genes_HUGO.txt](#)

[Back to top](#)

Reflink

Show refLink Details [\[REF\]](#)

(reflink_mrnaAcc, reflink_protAcc, reflink_name, reflink_prodName, reflink_locusLinkId, reflink_omimId)

Limit results with text matching the **hg18.reflink_product**

Distance From txStart or cdStart

Show the distance details [\[REF\]](#)
([dist_from_ref](#), [dist_from_tx](#), [dist_from_cds](#))

Include this many bases upstream/downstream of selected genes
([dist_from_ref](#))

Absolute minimum distance from any 'Transcriptional' start
([dist_from_tx](#))

Absolute minimum distance from any 'Translational' start
([dist_from_cds](#))

[Back to top](#)

Gene Ontology

Show gene ontology details [\[REF\]](#)
([go_names](#), [go_number](#))

Limit to text matching a 'Gene Ontology' term ([go_names](#))

Download the rsnumber to HUGO name file
(WARNING 352 Mb, right-click and 'download file') [SNP-Genes_GO.txt](#)

[Back to top](#)

Affymetrix

Limit results to rsnumbers on an Affymetrix SNP-CHIP [\[REF1\]](#), [\[REF2\]](#)

Include SNPs in strong LD with those on the Affymetrix 500k SNP-chip ([name_affy](#))

(LD = *linkage disequilibrium*)

Illumina

Limit results to rsnumbers on an Illumina SNP-chip [\[REF\]](#)

Include SNPs in strong LD with those on the ILMN_HumanHap550 SNP-chip

(LD = *linkage disequilibrium*)

[Back to top](#)

HapMap

Include other SNPs in strong linkage disequilibrium [\[REF\]](#)
(`ld_name`, `ld_name_affe`, `ld_name_illumina`, `ld_lod`, `ld_dprime`, `ld_rsquare`, `ld_pos_dif`,
`ld_pos1_hg17`, `ld_pos2_hg17`, `ld_fbin`)

The following requirements will be met

HapMap population
(CEU = Caucassian, YRI = African, JPT = Japanese, CHB = Chinese)

(`ld_dprime` LD >= x)

(`ld_rsquare` LD >= x)

(`ld_lod` LD >= x)

View HapMap details

You must check this box to show these parameters, otherwise they will be hidden

[Back to top](#)

HIV-1 Candidate Genes

Limit results to those from the 'Database of HIV-1 Candidate Genes'
where an rsnumber had an significance greater than or equal to a (`-logp`) value [\[REF\]](#)

(`-logp` >= x)

Copy Number Variation

Limit results to those 'within' a region of 'Copy Number Variation' (CNV)
as described in the 'Database of Humman Genomic Variants' (hg18.v2) [\[REF1\]](#), [\[REF2\]](#)

PReMod Modules

Limit to rsnumbers positioned within 'PReMod Modules' [\[REF\]](#)

List of comma-separated 'factor' or 'module_matrix' Names

(input 5 terms max)

and or

Select your 'factor' or 'module_matrix' names from this [PReMod key](#)

NOTE - there are 123,510 modules in the 'human_module_database' mapped to 'hg17.snp125'

[Back to top](#)

UCSC rsnumber Details

Show the rsnumber details from UCSC hg18.snp126 Table ([avHet](#), [avHetSE](#), [refUCSC](#), [refNCBI](#)) [[REF1](#), [REF2](#)]

Select Minimum Average Heterozygosity Cutoff ([avHet](#))

(0 <= [avHet](#) <= 1.0)

Select 'Validation Types' ([valid](#))

by-2hit-2allele (1,692,687)

by-cluster (1,154,345)

by-frequency (1,933,537)

by-submitter (214,482)

by-hapmap (9)

unknown (1,755,067)

and or

Select 'Function Types' ([func](#))

[Back to top](#)

locus (211,913)

coding (90,767)

coding-synon (40,422)

coding-nonsynon (50,572)

untranslated (92,688)

intron (2,848,608)

splice-site (678)

cds-reference (0)

unknown (1,364,457)

and or

Select 'Location Types' ([loctype](#))

[Back to top](#)

exact (4,784,820)

range (13,202)

between (4,866)

rangeInsertion (2,909)

rangeSubstitution (251)

rangeDeletion (4,866)

unknown (0)

and or

Select 'Molecular Types' (**moltype**)

[Back to top](#)

genomic (4,493,416)

cDNA (54,425)

unknown (0)

and or

1.10 Easy Mode

1.10.1 STEP 1 - Select Matrix Names

The “Delta-MATCH Hits” are internally organized as tables of ranked results in a MySQL database. Each transcription factor binding site matrix name (*mat_id*) has its own table of results (*rsnumbers*) that are ranked in descending order by the magnitude of their “Delta-MATCH *potential* score” (Definition page 30).

Users must select one of the five primary matrix selection buttons: (1) “Single Transcription Factor Matrix Name”, (2) “List of Transcription Factor Matrix Names”, (3) “Transcription Factor Name”, (4) “Tissue-Specific Transcription Factor Names”, (5) “All Transcription Factor Matrix Names”. These five primary matrix selections determine which of the 550 BIOBASE TRANSFAC matrixes will be included for the given query (database version 10.2). These Matrixes can be of “high” or “low” quality. A list of these 550 matrixes may be downloaded at the top of the Easy or Expert Mode web page (click the link that says “550 Vertebrate Transcription Factor Matrix Names” to download a file called “550_matrixes.txt”).

Figure 41 STEP 1 - Select Matrix Names

STEP 1 - Select Matrix Names

[Goto submit](#)

Choose one of the five 'Primary Matrix Selection Buttons' (left), and adjust its corresponding drop-down menu or text field (right). View the details of the '[550 Vertebrate Transcription Factor Matrix Names](#)' [[REF](#)]

1 - Single Transcription Factor Matrix Name

Select a 'Single Transcription Factor Matrix Name' (n=550)
 (mat_id)

2 - List of Transcription Factor Matrix Names

Hand-type a comma-separated 'List of Transcription Factor Matrixes Names'
 (mat_id) (1024 chars)

3 - Transcription Factor Name

Select transcription factor matrix names by a 'Transcription Factor Name' (n=351)
 (factor)

4 - Tissue-Specific Transcription Factor Names

Select transcription factor matrix names by a 'Tissue Type'

5 - All Transcription Factor Matrix Names

Select all matrix names in the database (n=550)

1.10.1.1 Primary Matrix Selection Button 1 - Single Transcription Factor Matrix Name

When the “**Single Transcription Factor Matrix Name**” radio button is selected (the default), the user will choose one from the list of corresponding 550 matrix names (mat_id) provided by the BIOBASE group. The Delta-MATCH Query Tool can quickly identify those polymorphisms where the “highest calculated MATCH scores” are greater than or equal to the transcription factor matrix false positive (**FP**) cutoff score (see the

MATCH publication for a description of the False Positive cutoff scores [1]). The number of “biologically relevant” (page 30) polymorphisms for each TFBS matrix is labeled in parentheses after the “*mat_id*” name. For example, when 4.5 million polymorphisms were calculated against the “V\$NFKB_Q6” transcription factor matrix, there were 950 “Delta-MATCH Hits” recorded into the database [V\$NFKB_Q6 (950)]. Each of these 950 rsnumbers had at least one allelic MATCH score (*m1* or *m2*) greater than or equal to the false positive cutoff (0.955). Note that some transcription factor matrix names have literally tens of thousands of rsnumber hits associated with them. Others have none. The number of hits identified for a given matrix was dependent on both the length of the matrix (in base pairs), and the diversity of its position-specific scoring matrix (nucleotide position-specific probability distribution).

1.10.1.2 Primary Matrix Selection Button 2 - List of Transcription Factor Matrix

Names

When the “**List of Transcription Factor Matrix Names**” radio button is selected, the user may type a comma separated list of transcription factor matrix names (*mat_id*). The number of the hand-typed characters must be less than or equal to 1024 characters, and each matrix name must be an **exact match** for those matrix names (*mat_id*) listed in the file “550_matrixes.txt”.

1.10.1.3 Primary Matrix Selection Button 3 - Transcription Factor Name

When the “**Transcription Factor Name**” radio button is selected, all of the matrix names (*mat_id*) corresponding to the selected “Transcription Factor Name” will be included in the search. There are 351 transcription factor names to select from (Table 42 page 405). One or more matrixes may belong to a single transcription factor name. For example, “NF-kappaB” has six “high quality” transcription factor matrix names

associated with it (V\$NFKAPPAB_01, V\$NFKAPPAB50_01, V\$NFKAPPAB65_01, V\$NFKB_C, V\$NFKB_Q6, and V\$NFKB_Q6_01).

1.10.1.4 Primary Matrix Selection Button 4 - Tissue-Specific Transcription Factor

Names

When the “**Tissue-Specific Transcription Factor Names**” radio button is selected, the transcription factor matrixes derived from the corresponding “Tissue Type” drop-down menu will be included in the search. A given transcription factor matrix may belong to more than one tissue type.

1.10.1.5 Table - Tissues Types in the Delta-MATCH Query Tool

- glioma_specific (n=145)
- immune_cell_specific (n=113)
- adipocyte_specific (n=68)
- cell_cycle_specific (n=85)
- liver_specific (n=112)
- lung_specific (n=59)
- muscle_specific (n=58)
- nerve_system_specific (n=158)
- pancreatic_beta_cell_specific (n=80)
- pituitary_specific (n=62)
- vertebrate_non_redundant (n=145)
- nerve_and_immune_cell_specific (n=213)

All of these tissue type groupings were provided by the BIOBASE team with the exception of the “glioma” selection, which was created by Alex Pico in the Conklin lab at

the Gladstone Institute of Cardiovascular Disease. The number of matrixes belonging to the each tissue type is labeled in parentheses. For example, the list of “glioma_specific” matrixes is comprised of 145 different TFBS matrixes.

1.10.1.6 Primary Matrix Selection Button 5 - All Transcription Factor Matrix Names

When the “**All Transcription Factor Matrix Names**” radio button is selected, every one of the 550 matrix names in the Delta-MATCH database will be searched.

1.10.2 STEP 2 -Add Restriction Criteria

There are many additional restriction criteria that users may want to include in a query. Each additional criterion (header names) may be selected by checking the appropriate “**header name checkbox**” () on the far left of the input page. When a given header name is selected, the additional sub-selections to the right of a header name checkbox (drop down menus, text fields, upload buttons, and internal checkboxes) become activated and are included in the query. In this way it is easy to create very detailed and complicated queries with a few simple selection.

During a query, if every header name checkbox is left unchecked, the Delta-MATCH tool will try to return the list of every rsnumber for every transcription factor matrix that is selected. Each header name is internally treated as an independent selection criterion, and once checked, only the rsnumbers that meet (the intersection) of all of the criteria will be returned in the results page (Figure page 86).

The default query searches only a single transcription factor matrix (V\$NFKB_Q6), and has the “Minimum Potential Score”, “top Most Significant Hits”, and “Matrix Quality”

header names checked and set to “0.8”, “5”, and “high” respectively. This default constellation of parameters creates a fairly restrictive search for against a single matrix, and can return a result almost instantly (Example 1 page 137).

1.10.2.1 Warning - Please Read About Each Restriction Criteria Before Checking Everything in Sight

1.10.2.2 Minimum Potential Score

When the “**Minimum Potential Score**” checkbox is checked, only those polymorphisms with a “Delta-MATCH potential score” (*potential*) greater or equal to the corresponding value will be returned (page 30). The minimum potential score may range from 0.0 to 1.0. A high potential score predicts that the two alleles for the corresponding polymorphisms have strong differences in transcription factor binding affinity. A potential score of 1.0 is calculated when one of the polymorphisms alleles creates a transcription factor binding site that matches the optimal binding site motif defined by the corresponding transcription factor matrix (MATCH score equals 1.0), while the other allele creates a binding site with a MATCH score less than or equal to the matrix name’s false positive (FP) cutoff.

It is expected that when comparing the potential score between two separate rsnumbers for the same transcription factor matrix name, the polymorphism with the larger potential score is predicted to have a larger difference in transcription factor binding affinity between its two alleles.

1.10.2.3 Warning - Don't compare the potential scores between different matrix names

It is not appropriate to directly compare the potential scores derived from different matrix names (mat_ids) because the distribution of Delta-MATCH hits for a given matrix is dependent on the length and accuracy of the matrix's probability distribution, and the accuracy at which its false positive threshold has been empirically estimated.

1.10.2.4 Selecting the best Minimum Potential Score Value (potential >= 0.3)

It is recommended starting your first queries with a relatively high "minimum potential score" cutoff ($0.8 \leq \text{potential} \leq 1.0$). Queries with high minimum potential score cutoffs will be the most stringent and will result in and shorter lists results. However, if more results are desired, it is possible to increase the number of hits returned for a given matrix name by lowering the "Minimum Potential Score" value. It is noteworthy that many of the rnumber hits with lowest potential scores ($\text{potential} \leq 0.3$) may be false positive predictions. For this reason it is generally recommended not to decrease the potential cutoff below 0.3. Keeping the potential cutoff higher than 0.3 should have the effect of filtering away roughly half of the lowest scoring Delta-MATCH hits for most matrixes. For example, the distribution of Delta-MATCH hits for the V\$NFKB_Q6 matrix shows that roughly half (56.9%) of its hits have potential scores less than or equal to 0.3 (Table page 62). If every result for a given matrix is wanted (the number in the parentheses in STEP 1 selection 1), simply uncheck the "**Minimum Potential Score**" box. This will force the DMQT to return all the hits for each matrix name (or up to the number selected by the "Maximum Returned rnumbers" check box) selected regardless of their potential score.

Figure 42 Minimum Potential Score Input

Minimum Potential Score

[Back to top](#)

Select a 'Minimum Potential Score' (potential)

0.80 (0 <= x <= 1.0)

1.10.2.5 Top Most Significant Hits

When the “**Top most Significant Hits**” checkbox is checked, this maximum number of polymorphisms will be returned for each of the matrix names passing the primary selection criteria in STEP 1. This value may range from 1 up to 1500. Note however, the maximum total number of results returned by the DMQT will be 1500 regardless of other parameters chosen.

Figure 43 Top Most Significant Hits

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix

5 Limit the number of polymorphisms searched per matrix

1.10.2.6 Matrix Quality

When the “**Matrix Quality**” checkbox is checked the matrixes selected in the primary matrix selection are filtered to only include the specified “high” or “low” quality matrixes as defined by BIOBASE TRANSFAC. A listing of the quality of each matrix can be found in the file “500_matrixes.txt” where a “1” is equivalent to “high” quality, and a “0” is equivalent to “low” quality. If this is left unchecked, both “high” and “low” quality matrix results will be returned.

Figure 44 Matrix Quality Input

Matrix Quality

Limit results by matrix 'Quality Type' (qual) [REF]
high (high = 1, low = 0)

1.10.2.7 Sort Results Table

By default, results will be grouped by matrix name, and returned in descending order of their potential score (rsnumbers with the largest “potential” scores are returned first).

However, when the “**Sort Results Table**” checkbox is checked, it is possible to sort the final results table in a number of ways. The results may be sorted by a set of descending (desc) and ascending (asc) parameter values. You may consider sorting the results by chromosomal position (a), by rsnumber (b), or by a descending value of their potential scores by selecting (c).

Figure 45 Sort Results Table Input

Sort Results Table

Sort the results table by
chrom asc, position asc (a)
(asc = ascending, desc = descending)

Figure 46 Sorting Selections

Sort the results table by

- chrom asc, position asc (a)
- chrom asc, position asc (a)
- name asc, matrix asc (b)
- potential desc, m_per desc, m1m2 asc, matrix asc (c)

1.10.2.8 Search By rsnumbers

When the “**Search By rsnumbers**” check box is checked, Delta-MATCH filters results to only include those listed in the corresponding “rsnumbers” text field. The typed list of comma-separated rsnumbers (dbSNP accession numbers) must match exactly those rsnumbers listed in the UCSC genome browser (UCSC hg18.snp126.name). A maximum of 1,024 characters may be typed into the “rsnumbers” text field.

Figure 47 Search By rsnumbers

Search By rsnumbers [Back to top](#)

Either limit results by a comma-separated list of dbSNP rsnumbers (1024 chars max)
[REF1, REF2]
rs5743836, rs6031444 'rsnumbers'

Or upload list of rsnumbers in a plain text file
Browse...
'rsnumber filename' [download example file](#)
(one rsnumber per row, 10,000 rsnumbers max)

NOTE - The 'rsnumber' text field will be ignored if an 'rsnumber filename' is uploaded

Search additional bases upstream and downstream of specified rsnumbers
 'rsnumber Window'
2000 Include other rsnumbers within this many bases

1.10.2.9 Uploading a List of rsnumbers

Alternatively, users may choose to upload a simple text file containing a list of up to 10,000 rsnumbers by selecting the “Choose File” button. The text file should contain only the list of rsnumbers, one per line. If using MS-Word to create the upload file, be sure to save file as the type “MS-DOS Text”. An “example file” created using the UNIX “vi” editor is provided as a reference and may be viewed by clicking on its link. Note that when a

file is uploaded, the rsnumbers typed in the rsnumber text field are ignored. If an upload file is not 'text/plain', Error 8 will be returned (Figure 212 page 413).

1.10.2.10 rsnumber Window

When both the **“Search By rsnumbers”** and the **“rsnumber Window”** check boxes are checked, a positional window around every submitted rsnumber is searched to identify and return other polymorphisms that pass the remaining criteria. The length of the SNP window can be selected to include up to a maximum of 30,000 base pairs upstream and downstream of every submitted rsnumber.

1.10.2.11 Search By Gene Names

After the primary matrix button has been selected, the user may choose to limit the results to include those polymorphisms that are located “within” or in near proximity to a comma-separated list of gene names. When the **“Search by Genes”** check box is checked, the typed gene names are text matched against the corresponding “UCSC hg18 Table”. Names (or accessions) may match the “name2”, “name”, or “proteinID” fields in the following tables: “hg18.refGene”, “hg18.geneid”, “hg18.mgcGenes” and “hg18.knownGene”. The submitted “Gene Names” should be an exact match to the corresponding table field. Be sure to choose the appropriate “Name Type” for the corresponding “UCSC hg18 Table” (download the “gene_name_name2.txt” help file to see some examples). A maximum of 1,024 characters may be typed into the “Gene Names” text field.

Figure 48 Search By Gene Names

[Back to top](#)

Search By Gene Names

'Search for gene without returning results' (MOCK SEARCH) [REF1, REF2]
UNCHECK this box to perform 'real' search after your gene names are verified

Limit results by a comma-separated list of 'Gene Names' (exact text match, 5 max)

JPH2, PLAT, TLR9 'Gene Names'

refGene (name2, name) 'UCSC hg18 Table Name'

name2 'Field Name' [download help file](#)

Search more bases upstream and downstream of specified genes

'Gene Window'

2000 Select the number of additional bases

1.10.2.12 Search for Gene Without Returning Results' (MOCK SEARCH)

By default, when you select the “Search By Gene Names” checkbox, the “search for gene without returning results” checkbox is checked. Pressing submit starts a mock run that produces no output table after verifying the existence of the typed gene names and displaying their chromosomal position and associated accession numbers. This type of quick search should be performed before searching for a new gene, particularly when the exact spelling of a gene name (abbreviation) is unknown. After the results are returned, you may press the backward arrow in your gene browser to return to the input page, uncheck the “search for gene without returning results” checkbox, and start your search for real. If your gene name is not found, back up and try another set of gene names.

When the “**Search By Gene Names**” check box and the “**Gene Window**” check box are both checked, all polymorphisms located given distance upstream or downstream of the designated genes will be included in the search. The length of the Gene Window can include up to 30,000 bp upstream and downstream of each gene in the “Gene Names” list. For example, when gene name “TLR9” is selected and matched against the name2

field of the hg18.refGene table and the “Gene Window is set to the default of 2,000 base pairs, Delta-MATCH will search for all polymorphisms on human chromosome 3 between base pairs 52228272 and 52236585. This search may identify additional polymorphisms 2,000 bases upstream and downstream of the TLR9 gene.

1.10.2.13 What Happens When a Gene Name has Multiple Transcripts?

If a specified gene name has more than one entry on the same chromosome in the specified UCSC hg18 table, the leftmost and rightmost positions of the set of transcripts will be used to designate the genes position. For example, when gene name “TLR9” is selected and matched against the name2 field of the refGene table, two matches are found. TLR9 corresponds to the name fields “NM_138688” and “NM_017442”. Delta-MATCH summarizes that TLR9 is positioned on the negative strand of human chromosome 3, has a leftmost position equal to base pair 52230137 and a rightmost position equal to base pair 52235219, and would search for all polymorphisms within this range.

Figure 49 TLR9 Isoforms

TLR9 - Entry 1	TLR9 - Entry 2
name2: TLR9	name2: TLR9
name: NM_138688	name: NM_017442
strand: -	strand: -
chrom: chr3	chrom: chr3
txStart: 52230137	txStart: 52230137
txEnd: 52233247	txEnd: 52235219
distance: 2000	distance: 2000
new_start: 52228137	new_start: 52228137
new_end: 52235247	new_end: 52237219
min_txStart: 52228137	min_txStart: 52228137
max_txEnd: 52235247	max_txEnd: 52237219

1.10.3 STEP 3 - Submit (press the submit button)

STEP 3 - Press Submit

[Back to top](#)

Submit (a maximum of 1,500 results will be returned)

When you have configured your query, press the “Submit” button to initiate your search.

1.10.3.1 Hint - Opening Your Output Results Page in a New Tab (right click option)

By default, a Delta-MATCH open the results page in either new tab or a new browser tab or in a new browser window. This function allows the user to keep the original Delta-MATCH input page available for subsequent modification. In this way, it is possible to compare sets of results quickly by submitting slightly different constellations of parameters.

Figure 50 The Delta-MATCH Output Results Page

The screenshot shows the Delta-MATCH Results page. At the top, there is a navigation bar with the Delta-MATCH logo and links for HOME, ABOUT, EASY MODE, EXPERT MODE, TUTORIAL, DOWNLOADS, and AUTHOR. Below the navigation bar, the word "Results" is displayed in a large, blue font. Underneath "Results", there are two columns of links. The left column contains links to various output files: DM_1464916942_table.html, DM_1464916942_table.txt, DM_1464916942_table.xml, DM_1464916942_report.html, and DM_1464916942_log.html. The right column contains information about the search: DM_result_1464916942, October 22, 2007, 11:28 am, run time: 1 s, 1 matrix name was searched, and There were 5 'Delta-MATCH hits' returned. Below this information is a table with 14 columns: hit, rsnumber, chrom, chromStart, factor, mat_id, potential, threshold, m1, m2, m_per, rank, p1_window, name, and hit. The table contains 5 rows of results, each with a hit number from 1 to 5. The first row is: 1, rs3093317, chr16, 27351578, NF-kappaB, V\$NFKB_Q6, 1, 0.955, 0.8377, 1.0000, 16.23, 7, chr16:27351571-27351584, rs3093317, 1. The second row is: 2, rs8030978, chr15, 64651526, NF-kappaB, V\$NFKB_Q6, 1, 0.955, 0.8377, 1.0000, 16.23, 2, chr15:64651519-64651532, rs8030978, 2. The third row is: 3, rs1775044, chr1, 7418248, NF-kappaB, V\$NFKB_Q6, 1, 0.955, 1.0000, 0.8795, 12.05, 4, chr1:7418246-7418259, rs1775044, 3. The fourth row is: 4, rs7296179, chr12, 100126053, NF-kappaB, V\$NFKB_Q6, 1, 0.955, 0.8795, 1.0000, 12.05, 6, chr12:100126043-100126056, rs7296179, 4. The fifth row is: 5, rs6031444, chr20, 42249151, NF-kappaB, V\$NFKB_Q6, 1, 0.955, 1.0000, 0.8895, 11.05, 1, chr20:42249149-42249162, rs6031444, 5.

hit	rsnumber	chrom	chromStart	factor	mat_id	potential	threshold	m1	m2	m_per	rank	p1_window	name	hit
1	rs3093317	chr16	27351578	NF-kappaB	V\$NFKB_Q6	1	0.955	0.8377	1.0000	16.23	7	chr16:27351571-27351584	rs3093317	1
2	rs8030978	chr15	64651526	NF-kappaB	V\$NFKB_Q6	1	0.955	0.8377	1.0000	16.23	2	chr15:64651519-64651532	rs8030978	2
3	rs1775044	chr1	7418248	NF-kappaB	V\$NFKB_Q6	1	0.955	1.0000	0.8795	12.05	4	chr1:7418246-7418259	rs1775044	3
4	rs7296179	chr12	100126053	NF-kappaB	V\$NFKB_Q6	1	0.955	0.8795	1.0000	12.05	6	chr12:100126043-100126056	rs7296179	4
5	rs6031444	chr20	42249151	NF-kappaB	V\$NFKB_Q6	1	0.955	1.0000	0.8895	11.05	1	chr20:42249149-42249162	rs6031444	5

1.10.3.2 A Delta-MATCH Query May Take Seconds or Minutes (up to tens of minutes)

The default submission should only take a few seconds to return its results to the web browser. Most moderate level queries take no more than 5 or 8 minutes to complete. Be patient, and please do not submit multiple complex queries simultaneously. If the browser fails to return a result page after a reasonable period of time, stop the job by quitting the browser window. (Delta-MATCH has been tested primarily with [Firefox](#) and Safari Browsers).

Complex queries that search many transcription factor matrixes (550 high and low quality matrixes) while combining computationally intensive restriction criteria (Hugo Name, Gene Ontology, Bonferonni Correction, Distance From txStart or cdStart, HapMap, Affymetrix, Illumina HIV-1 Candidate Genes, Copy Number Variation, PReMod) have a higher risk of bogging down. Theoretically, a query could try to return the entire list of database hits at once, but some safeguards are in place to try to shut down runaway processes.

Presently there is a maximum 60-minute limit to the amount of time the Delta-MATCH Query Tool PHP script is allowed to run before returning a PHP timeout error. However, it should be realized that no matter how complicated a user makes a query, a maximum number of 1,500 results can be returned per query.

As more users start to submit users, this maximum number of returned hits may be adjusted to allow all users optimal performance.

1.10.4 A Successful Delta-MATCH Run Creates 5 Output Files

Figure 51 Download and save Delta-MATCH results as HTML, XML or TXT files

Every Delta-MATCH query generates 5 separate output files that allow users to save the "Delta-MATCH Results Table" in three different file formats (table.html, table.xml and/or table.txt), to save a brief report of the query (report.html), and a log file that allows users to replicate the exact same query at a later time point with the click of a button (log.html).

Download and Save Delta-MATCH Results as HTML, XML or TXT Files

The screenshot displays the Delta-MATCH results interface. On the left, a table lists hits with columns for hit number, chromosome, start position, factor, motif ID, potential, threshold, n1, n2, n_per, rank, factor description, count, potential, motif count, and rerness. Three hits are highlighted in red. On the right, an XML snippet shows the structure of the data, including fields for hit ID, chromosome, start position, factor, motif ID, potential, threshold, n1, n2, n_per, rank, factor description, count, potential, motif count, and rerness.

hit	name	chrom	chromStart	mot_id	dif_z	threshold	n1	n2	n_per	rank	p1_window
1	rs3093317	chr16	27351578	V\$NFKB_Q6	1	0.955	0.8377	1.0000	16.23	7	chr16:27351571-27351584
2	rs8030978	chr15	64651526	V\$NFKB_Q6	1	0.955	0.8377	1.0000	16.23	2	chr15:64651519-64651532
3	rs6031444	chr20	42249151	V\$NFKB_Q6	1	0.955	1.0000	0.8895	11.05	1	chr20:42249149-42249162

Rerun this exact Delta-MATCH query using all of the parameters listed below at <http://deltamatch.org>

After viewing the Delta-MATCH Query Results in the web browser, users may want to save copies of the results table (table.html, table.txt, table.xml), and a list of the parameters that were selected for the present search (log.html). At the top left of the output results page are links that allow the user to download the files for the current query. These files are removed from the server every 24-hours. Users may right-click one of these web links with your mouse and select to “download” the linked file.

Figure 52 Right Click a Web Link To Download a Temporary Result Table or Log File (Firefox)

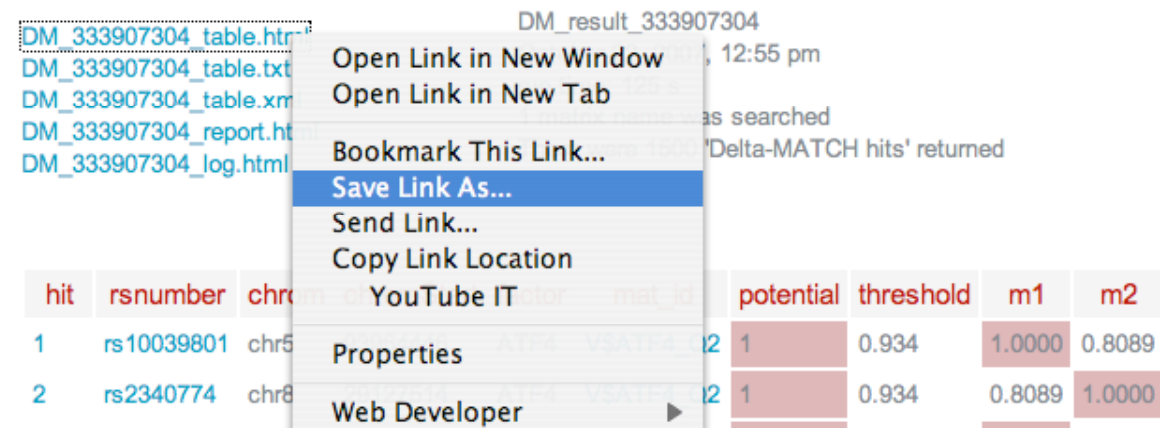


Figure 53 Downloadable File of the Results Table (DM_*_table.html)

hit	rsnumber	chrom	chromStart	factor	mat_id	potential	threshold	m1	m2	m_per	rank	p1_window	name	hit
1	rs3093317	chr16	27351578	NF-kappaB	VSNFKB_Q6	1	0.955	0.8377	1.0000	16.23	7	chr16:27351571-27351584	rs3093317	1
2	rs8030978	chr15	64651526	NF-kappaB	VSNFKB_Q6	1	0.955	0.8377	1.0000	16.23	2	chr15:64651519-64651532	rs8030978	2
3	rs1775044	chr1	7418248	NF-kappaB	VSNFKB_Q6	1	0.955	1.0000	0.8795	12.05	4	chr1:7418246-7418259	rs1775044	3
4	rs7296179	chr12	100126053	NF-kappaB	VSNFKB_Q6	1	0.955	0.8795	1.0000	12.05	6	chr12:100126043-100126056	rs7296179	4
5	rs6031444	chr20	42249151	NF-kappaB	VSNFKB_Q6	1	0.955	1.0000	0.8895	11.05	1	chr20:42249149-42249162	rs6031444	5

This is a hypertext markup text file (html) of your results that once downloaded, can be viewed by opening it up with your favorite web browser. Users may want to save this file because it preserves active hyperlinks (blue underlined links) to other outside resources

that may want to be investigated later. Users also may want to save this file to review the color intensities of the hits (potential, m1, m2, m_per).

Figure 54 Downloadable File of the Results Table (DM *_*_table.txt)

```

hit      rsnumber      chrom  chromStart  factor  mat_id  potential  threshold  m1  m2  m_per  rank
pl_window name      hit
1      rs3093317      chr16  27351578    7      V$NFKB_Q6    1      0.955  0.8377  1.0000  16.23  7
chr16:27351571-27351584 rs3093317 1
2      rs8030978      chr15  64651526    7      V$NFKB_Q6    1      0.955  0.8377  1.0000  16.23  2
chr15:64651519-64651532 rs8030978 2
3      rs1775044      chr1   7418248     7      V$NFKB_Q6    1      0.955  1.0000  0.8795  12.05  4      chr1:7418246-7418259
rs1775044 3
4      rs7296179      chr12  100126053   7      V$NFKB_Q6    1      0.955  0.8795  1.0000  12.05  6
chr12:100126043-100126056 rs7296179 4
5      rs6031444      chr20  42249151    7      V$NFKB_Q6    1      0.955  1.0000  0.8895  11.05  1
chr20:42249149-42249162 rs6031444 5

```

This is a simple text file (txt) of your results table. All html markups have been removed from this file (no embedded links). This file is tab separated and can be downloaded and opened up in a spreadsheet program. Some users may prefer to save this file in order to further sort and filter the results in a program like MS Excel.

Figure 55 Downloadable File of the Results Table (DM *_table.xml) (viewed in text program)

```

<results>
  <result>
    <hit>1</hit>
    <rsnumber>rs3093317</rsnumber>
    <chrom>chr16</chrom>
    <chromStart>27351578</chromStart>
    <factor>NF-kappaB</factor>
    <mat_id>V$NFKB_Q6</mat_id>
    <potential>1</potential>
    <threshold>0.955</threshold>
    <m1>0.8377</m1>
    <m2>1.0000</m2>
    <m_per>16.23</m_per>
    <rank>7</rank>
    <p1_window>chr16:27351571-27351584</p1_window>
  </result>

  <result>
    <hit>2</hit>
    <rsnumber>rs8030978</rsnumber>
    <chrom>chr15</chrom>
    <chromStart>64651526</chromStart>
    <factor>NF-kappaB</factor>
    <mat_id>V$NFKB_Q6</mat_id>
    <potential>1</potential>
    <threshold>0.955</threshold>
    <m1>0.8377</m1>
    <m2>1.0000</m2>
    <m_per>16.23</m_per>
    <rank>2</rank>
    <p1_window>chr15:64651519-64651532</p1_window>
  </result>

  (two results missing here)

  <result>
    <hit>5</hit>
    <rsnumber>rs5031444</rsnumber>
    <chrom>chr20</chrom>
    <chromStart>42249151</chromStart>
    <factor>NF-kappaB</factor>
    <mat_id>V$NFKB_Q6</mat_id>
    <potential>1</potential>
    <threshold>0.955</threshold>
    <m1>1.0000</m1>
    <m2>0.8895</m2>
    <m_per>11.05</m_per>
    <rank>1</rank>
    <p1_window>chr20:42249149-42249162</p1_window>
  </result>
</results>

```

This is an extensible markup language (XML) file of the results table. Each cell value in the results table is marked up with a pair of embedded tags (<example_tag_name> example_tag_value </example_tag_name>). Similarly each resultant row is enclosed in a tag called “result”, and the entire file is enclosed with a tag called “results”. This file has

an “.xml” extension in its name that if opened in a web browser will cause the file to be stripped of all of its XML tags as shown in the above figure. The file may be downloaded and opened with a text editor to see the tags clearly.

Figure 56 Downloadable File of the Results Table (DM *_table.xml) (viewed in text web browser)

```
1 rs3093317 chr16 27351578 NF-kappaB V$NFKB_Q6 1 0.955 0.8377 1.0000 16.23 7 chr16:27351571-27351584 2
rs8030978 chr15 64651526 NF-kappaB V$NFKB_Q6 1 0.955 0.8377 1.0000 16.23 2 chr15:64651519-64651532 3
rs1775044 chr1 7418248 NF-kappaB V$NFKB_Q6 1 0.955 1.0000 0.8795 12.05 4 chr1:7418246-7418259 4 rs7296179
chr12 100126053 NF-kappaB V$NFKB_Q6 1 0.955 0.8795 1.0000 12.05 6 chr12:100126043-100126056 5 rs6031444
chr20 42249151 NF-kappaB V$NFKB_Q6 1 0.955 1.0000 0.8895 11.05 1 chr20:42249149-42249162
```

Delta-MATCH provides users may want to use the downloaded XML files to parse out and save a limited number of column field types, and may build their own parsers by downloading the XML partner DTD file (dm2_results.dtd) (page 408).

Figure 57 Downloadable Log File (DM_*_log.html)

This html file shows of all of the input parameters for the present query. It is recommended that users download and save this log file as a record of an interesting search because it provides the ability to repeat the same search at another time point without using the standard input page. Pressing the submit button will reinitiate a new query using the exact same set of parameters at <http://deltamatch.org> if you are connected to the internet. (Presently, log files for searches having the “Search By rsnumbers” header name checked will produce Error 3 when resubmitted, page 410).

```

Rerun this exact Delta-MATCH query using all of the parameters listed below at http://deltamatch.org

Input Parameters
filename = DM_result_283836056_log.html
date_of_run = 20071021224452
script_address = http://ipx1.localMozzila/5.0 (Macintosh; U; PPC Mac OS X; en) AppleWebKit/522.11.1 (KHTML, like Gecko) Version/3.0.3 Safari/522.12.1
Delta_MATCH_version = 1.0
filename_rsnumbers =
number_of_hits = 5
run_time = 1
dm_error =
select_tf_RadioGroup1 = select_by_matrixes
matrix = V&NFKB_Q6
matrix_names = V&NFKB_Q6,V&NFKB_C
transcription_factor_name = NF-kappaB
tissue = immune_cell_specific
dif_z_box = true
difz = 0.80
limf_box = true
limit_max_per_matrix = 5
q_box = true
qual = high
order_letter = a
rsnumber = rs5743836,rs6031444
distance_from_snp = 2000
gene_box = true
gene = JPH2,PLAT,TLR9
gene_db = refGene
name_type = name2
distance_from_gene_box = true
distance = 2000
mat = 12
chromosome = chr8
base_start = 128100000
base_end = 128700000
snp_strand = 1
region_up10k_box = true
region_phastconsElements17way_box = true
region_RadioGroup1 = region_or
pvalue_adjusted_cutoff = 0.005
number_hits = 1
reflink_text = kinase
reference = 1
distance_from_tx = 2000
distance_from_cds = 2000
mygo = transcription
affy = 250k_all
pop1 = CEU
pop2 = VRI
id_affy_500k_db = id_affy_ceu_1_0
illumina_list = TLMN_HumanHap550_SNPlist
id_illumina_550_db = id_illumina_550_ceu_1_0
id = CEU
id_dprime_min = 1.00
id_rsquare_min = 0.80
id lod_min = 18
id_details_box = true
chavi_logp = 0.5
dgv_selection = within
premod = include
premod_names_box = true
premod_names = M00769,M00701
premod_RadioGroup1 = premod_and
Het = 0.01
valid_by_2hit_Zallele_box = true
valid_RadioGroup1 = valid_or
function_locus_box = true
function_RadioGroup1 = function_or
loc_exact_box = true
loc_RadioGroup1 = loc_or
mol_genomic_box = true
mol_RadioGroup1 = mol_or
cutoff = FP
regpotential7x = include
cpgislands = include
repeatsmasker = include
simperepeats = exclude
microsatellite = exclude
database_name = 1
oligo_length = 30

```

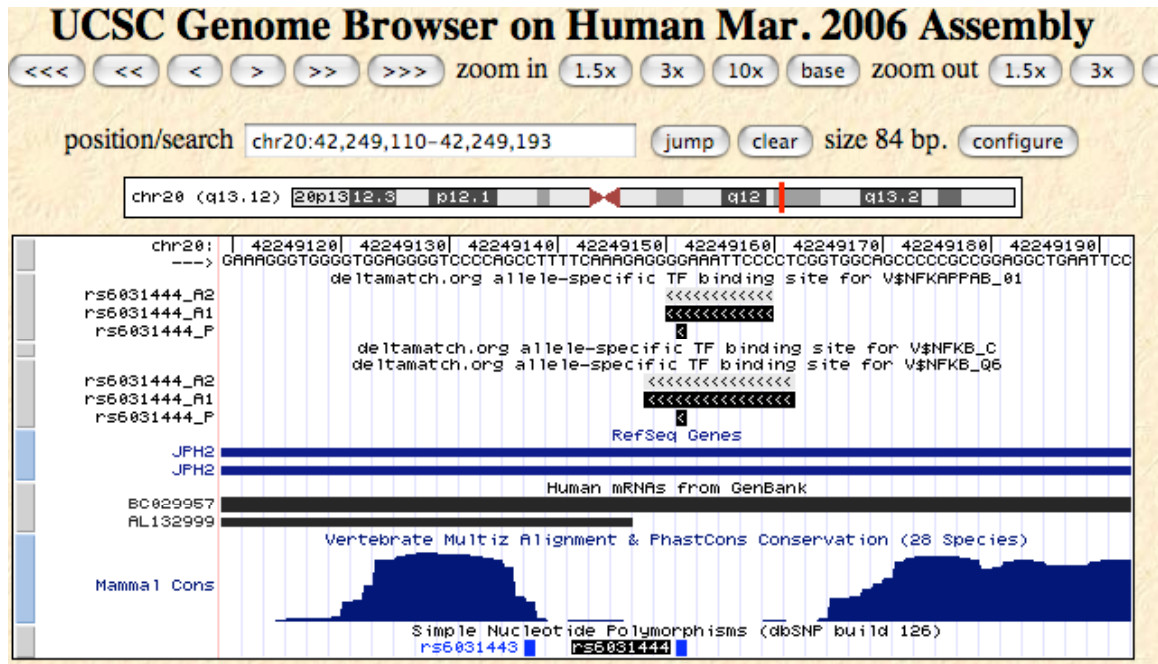
1.10.5 Viewing Delta-MATCH Data as UCSC Genome Browser Tracks

It is possible to visually view the position and details of Delta-MATCH results by viewing their potential scores as they align to the human genome (hg18) in the UCSC Genome Browser (Figure 58 page 119). To do this, perform the following:

- download the appropriate tracks from the Delta-MATCH UCSC Browser Tracks downloads page
- read and follow the instructions on the UCSC [‘Displaying and Managing Custom Tracks’](#) web page to install the Delta-MATCH track files for any wanted matrix names (examples V\$NFKB_Q6, V\$NFKB_C, V\$NFKAPPAB_01)
- open a new UCSC Genome Browser [window](#)
- turn the “SNPs(126)” UCSC browser track to “pack” (located under “Variation and Repeats”)
- turn the “Conservation” UCSC browser track to “squish” (located under “Comparative Genomics”)
- turn the “Human mRNAs” UCSC browser track to “pack” (located under “mRNA and EST Tracks”)
- press the “refresh” button
- type an rsnumber into the ‘position/search’ window in the browser and press enter (example rs6031444)
- click the link under “Simple Nucleotide Polymorphisms (dbSNP build 126) (snp126)”
- zoom in to ‘base’ by pressing the ‘base’ button

If you followed the above instructions, you should see this:

Figure 58 Delta-MATCH Data Can Be Visualized as a Custom Track in the UCSC Genome Browser



1.10.6 Description of the Delta-MATCH UCSC Tracks

The following examples use these three uploaded Delta-MATCH files:

- dm_track_V\$NFKAPPAB_01.txt
- dm_track_V\$NFKB_C.txt
- dm_track_V\$NFKB_Q6.txt

Each “Delta-MATCH hit” for a given transcription factor matrix can be visualized in the UCSC browser as a series of three track entries listed under a track called “deltamatch.org allele-specific TF binding site for (**mat_id**)”.

1.10.6.1 Definition - rsnumber_A1

This track shows the position and magnitude of the highest MATCH score for allele 1 (**m1**). The arrows show the strand of the match (forward arrows = "+ strand"; reverse arrows = "- strand"). The shade of the arrows is proportional to the magnitude of the MATCH score (lightest = m1 = 0.0; darkest = m1 = 1.0).

1.10.6.2 Definition - rsnumber_A2

This track shows the position and magnitude of the highest MATCH score for allele 2 (**m2**). The arrows show the strand of the match (forward arrows = "+ strand"; reverse arrows = "- strand"). The shade of the arrows is proportional to the magnitude of the MATCH score (lightest = m2 = 0.0; darkest = m2 = 1.0).

1.10.6.3 Definition - rsnumber_P

This track shows the position of the rsnumber. The shade of the arrows is proportional to the magnitude of the Delta-MATCH potential score (**potential**) (lightest = potential = 0.0; darkest = potential = 1.0).

To learn more about viewing the Delta-MATCH results in the UCSC track, see "Example 20 - Restricting By Chromosome and Position Range" (page 182).

1.11 Delta-MATCH Examples (Easy Mode)

This document details the 40 examples found at the **Delta-MATCH > Tutorial > Examples** webpage (<http://deltamatch.org>).

These examples display the wide variety of types of queries that can be created with the Delta-MATCH Query Tool (Easy Mode and Expert Mode).

'Delta-MATCH Examples' have been created to demonstrate the diversity of queries that may be created with the Delta-MATCH Query Tool. These examples showcase the usefulness of each selection parameter.

Each Delta-MATCH Query generates up to 5 separate output files that allow you to save the "Delta-MATCH Results Table" in three different file formats (table), to view a brief report of the query results (report), and to replicate the exact query at a later time point at the click of a button (log). The function of these files has been previously described (page 112).

- table.log
- report.html
- table.html
- table.txt
- table.xml

Every example highlights a single (or combinations of) function(s) that can be used to return lists of polymorphisms that have a strong “potential” to create an allele-specific

transcription factor binding site. With the exception of the “Delta-MATCH Proof of Principle Example - AHSG rs2248690”, (page 122), these examples are generally ordered by their complexity, starting with the simplest to the most complex.

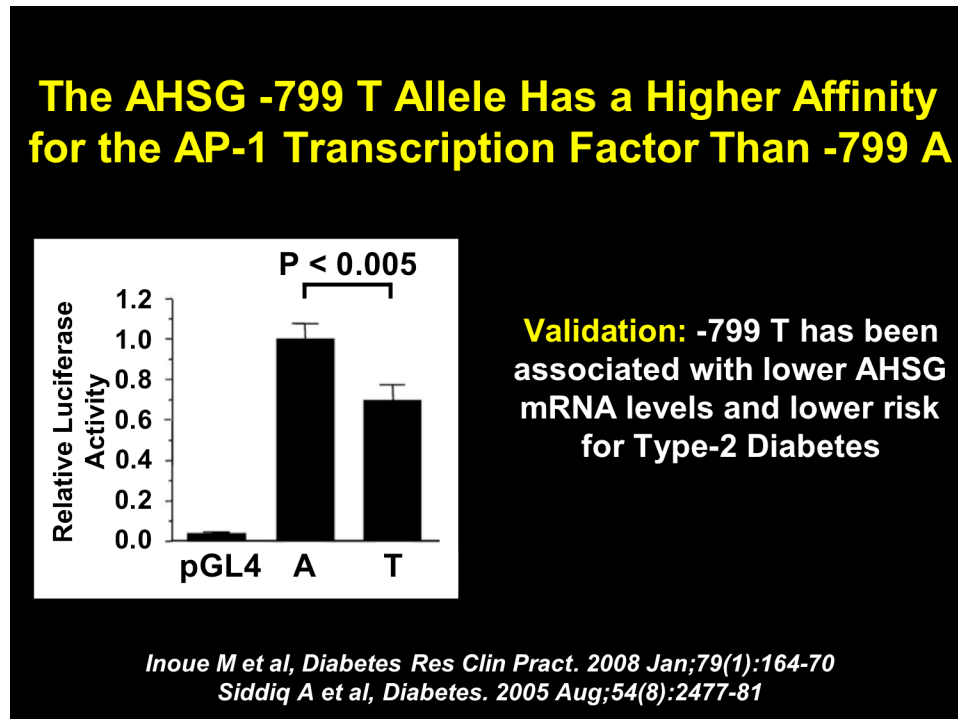
A detailed description of the Delta-MATCH Algorithm, Database and Query tool can be downloaded in the Delta-MATCH tutorial pdf at **Delta-MATCH > Tutorial**.

1.11.1 Delta-MATCH Proof of Principle Example - AHSG rs2248690

This is the proof of principle example for the Delta-MATCH Database and Query Tool.

In this example the Delta-MATCH Query Tool identifies rs2248690 as an A>T polymorphism that is 799 base pairs upstream of the alpha2-Heremans-Schmid glycoprotein (AHSG) gene. It has been shown that this polymorphism is located in a binding site for the AP-1 transcription factor and that the -799T allele is associated with increased AP-1 affinity, decreased AHGS mRNA expression relative to the -799A allele and is associated with Type 2 Diabetes (Figure 67 page 142) [29, 30].

Figure 59 The AHSG -799T Allele Has a Higher Affinity for the AP-1 Transcription Factor Than -779 A³



This query will return from the Delta-MATCH database a list of up to 'five' (Top Most Significant Hits) biallelic polymorphisms (rsnumber) that have the potential to create an allele-specific transcription factor binding site for any 'high quality' (Matrix Quality) matrix that has a matrix length at least 8 base pairs long (Show the Matrix Details >= 8).

³ This bar chart was borrowed and adapted without permission from Figure 1 Reference 29. Inoue, M., et al., *A promoter polymorphism of the alpha2-HS glycoprotein gene is associated with its transcriptional activity*. *Diabetes Res Clin Pract*, 2007.

The list of returned polymorphisms:

- must have a potential score ≥ 0.3
- must be located within 2000 base pairs of a gene named 'AHSG'
- must be located on chromosome 3
- must be located between base pairs 18780271 and 18781271
- must have either m1 or m2 align to the plus (Watson) strand of DNA
- must be located in a refSeq 10kb upstream region, 10kb downstream region, 5' untranslated region, 3' untranslated region, intronic region, exonic region, coding region, **'or'** region of strong conservation (phastcons17).
- must have a Bonferonni-adjusted rareness ≤ 0.05
- must be associated with a gene in the HUGO database that has a gene ontology term that matches the term 'insulin'
- must be located within 2000 base pairs of a 'transcriptional' start site
- must be located within 2000 base pairs of a 'translational' start site
- must have an average heterozygosity frequency ≥ 0.3
- must have a 'Validation Type' matching the terms 'by-2hit-2allele', 'by-cluster' **and** 'by-frequency'
- must have a 'Function Type' matching the term 'locus'
- must have a 'Location Type' matching the term 'exact'
- must have a 'Molecular Type' matching the term 'genomic'

Additionally the results:

- will display the details of the matrix that matched the position of the polymorphism
- will calculate and display the rareness of each result
- will be sorted by chromosomal position

- will display the magnitude, strand and position offset of each highest MATCH score for each allele (m1, m2, p1, p2, s1, s2)
- will display the 'observed' polymorphic alleles and specify which allele is referenced by the UCSC and NCBI genome alignments
- will display any associated HUGO genes
- will display clickable hyperlinks for NCBI entries for mRNA and protein sequences, and the Gene, OMIM, and LocusLink entries for the associated HUGO genes
- will display hyperlinks to the UCSC Genome Browser showing the exact position of each allelic MATCH score (m1 and m2)
- will display clickable hyperlinks to dbSNP for each returned rsnumber
- will display a clickable hyperlink to [PubMed](#) citations matching the returned rsnumber
- will display other polymorphisms that are in strong linkage disequilibrium with every returned hit as determined by HapMap (population = Japanese; $D' = 1.0$; $r^2 = 0.3$; $LOD \geq 2$)
- will display the distance the rsnumber is from a known transcriptional and translational start site

Figure 60 Input Parameters for the Proof of Principle Example

STEP 1 - (5) All Transcription Factor Matrix Names

STEP 2 - Minimum Potential Score (checked) = "0.3"

STEP 2 - Top Most Significant Hits (checked) = "5"

STEP 2 - Matrix Quality (checked) = "high"

STEP 2 - Sort Results Table (checked) = "chrom asc, position asc (a)"

STEP 2 - Search By Gene Names (checked); Search for gene without returning results (unchecked), Gene Names = "AHSG"; UCSC hg18 Table Name = "refGene"; Field Name = "name2"; Gene Window (checked) = "2000"

STEP 2 - Show the Matrix Details (checked); Minimum Matrix Length (checked) = "8"

STEP 2 - Show the Position Details (checked)

STEP 2 - Chromosome (checked) = "3"

STEP 2 - Position Range (checked) ; lowest base ="187802781"; highest base = "187812781"

STEP 2 - Strand (checked) = "+"

STEP 2 - Genomic Regions (checked) = "up10k; phastconsElements17way; utr5; coding; down10k; exons; introns; utr3"; "or"

STEP 2 - Bonferonni Correction (checked) = "0.05"

STEP 2 - Minimum Total Number of Delta-MATCH Hits (checked) = "1"

STEP 2 - Hugo Names (checked); Limit results to rsnumbers next to known HUGO_GENES (checked)

STEP 2 - Reflink (checked); Limit results with text matching the hg18.reflink_product (checked) = "glycoprotein"

STEP 2 - Distance From txStart or cdStart (checked); = ("1", "2000", "2000")

STEP 2 - Gene Ontology (checked); Limit to text matching a Gene Ontology term (checked) = "insulin"

STEP 2 - HapMap (checked); HapMap population = "JPT Japanese"; Id_prime >= "1.00"; Id_square >= "0.3"; Id_lod >= "2"; View HapMap Details (checked)

STEP 2 - UCSC rsnumber Details (checked); "Select Minimum Average Heterozygosity Cutoff (checked) = 0.3; "Select 'Validation Types'" = "by-2hit-2allele; by-cluster; by-frequency" (checked/and); "Select 'Function Types'" = "locus" (checked/or); "Select

'Location Types' = "exact" (checked/or); "Select 'Molecular Types'" = "genomics"
(checked/or)

5 - All Transcription Factor Matrix Names

Select all matrix names in the database (n=550)

Minimum Potential Score

[Back to top](#)

Select a 'Minimum Potential Score' (**potential**)

(0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix

Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (**qual**) [[REF](#)]

('high = 1', 'low = 0')

Sort Results Table

Sort the results table by

(asc = ascending, desc = descending)

Search By Gene Names

[Back to top](#)

'Search for gene without returning results' (*MOCK SEARCH*) [[REF1](#), [REF2](#)]

UNCHECK this box to perform 'real' search after your gene names are verified

Limit results by a comma-separated list of 'Gene Names' (exact text match, 5 max)

'Gene Names'

'UCSC hg18 Table Name'

'Field Name' [download help file](#)

Search more bases upstream and downstream of specified genes

'Gene Window'

Select the number of additional bases

Show the Matrix Details

[Back to top](#)

Show the matrix details [\[REF\]](#)

([count_le_potential](#), [mat_count](#), [frequency](#), [factor](#), [factor_description](#), [qual](#), [mat_len](#))

Minimum Matrix Length

Limit searches to those matrixes with minimum length ([mat_len](#))

([mat_len](#) >= x)

Show the Position Details

Show the position and strand details [\[REF\]](#)

([p1_window](#), [p2_window](#), [p1](#), [p2](#), [s1](#), [s2](#))

Chromosome

[Back to top](#)

Limit results to a chromosome [\[REF\]](#)

([chrom](#))

Position Range

Limit results between two positions [\[REF\]](#)

Enter lowest base ([chrStart](#) >= x)

Enter highest base ([chrStart](#) <= x)

Strand

Limit matrix hits to a DNA strand [\[REF\]](#)

([strand](#))

Genomic Regions

[Back to top](#)

Limit results to include rsnumbers positioned in these genomic regions of refSeq genes [\[REF\]](#)

[up10k](#) (647,311)

[phastconsElements17way](#) (397,802)

[utr5](#) (16,376)

[coding](#) (113,832)

[down10k](#) (648,916)

[exons](#) (212,764)

[introns](#) (3,415,853)

[utr3](#) (84,503)

or and ("[and](#)" IS VERY SLOW!)

[Back to top](#)

Bonferonni Correction

Limit results by 'Minimum Bonferonni-Adjusted Rareness' (**bonferonni**)
(**bonferonni** = rareness*(number of returned hits))

(**bonferonni** <= x)

NOTE - must have 'Matrix Details' checked to see this column

Minimum Total Number of Hits

Limit results to rsnumbers with a minimum 'total number of hits'

This is the sum number of hits for an rsnumber in the database

(**number_hits** >= x)

HUGO Names

Show the HUGO names of the genes associated with each rsnumber [[REF1](#), [REF2](#)]
(**hugo_name**)

Limit results to rsnumbers next to known HUGO genes

Download the rsnumber to hugo name file

(*WARNING 48.4 Mb, right-click and 'download file'*) [SNP-Genes_HUGO.txt](#)

[Back to top](#)

Reflink

Show reflink Details [[REF](#)]

(**reflink_mmaAcc**, **reflink_protAcc**, **reflink_name**, **reflink_prodName**, **reflink_locusLinkId**,
reflink_omimId)

Limit results with text matching the **hg18.reflink_product**

Distance From txStart or cdStart

Show the distance details [[REF](#)]

(**dist_from_ref**, **dist_from_tx**, **dist_from_cds**)

Include this many bases upstream/downstream of selected genes

(**dist_from_ref**)

Absolute minimum distance from any 'Transcriptional' start

(**dist_from_tx**)

Absolute minimum distance from any 'Translational' start

(**dist_from_cds**)

[Back to top](#)

Gene Ontology

Show gene ontology details [REF]
([go_names](#), [go_number](#))

Limit to text matching a 'Gene Ontology' term ([go_names](#))

Download the rsnumber to HUGO name file
(WARNING 352 Mb, right-click and 'download file') [SNP-Genes_GO.txt](#)

[Back to top](#)

HapMap

Include other SNPs in strong linkage disequilibrium [REF]
([ld_name](#), [ld_name_affy](#), [ld_name_illumina](#), [ld_lod](#), [ld_dprime](#), [ld_rsquare](#), [ld_pos_dif](#),
[ld_pos1_hg17](#), [ld_pos2_hg17](#), [ld_fbin](#))

The following requirements will be met

HapMap population

(CEU = Caucassian, YRI = African, JPT = Japanese, CHB = Chinese)

([ld_dprime](#) LD >= x)

([ld_rsquare](#) LD >= x)

([ld_lod](#) LD >= x)

View HapMap details

You must check this box to show these parameters, otherwise they will be hidden

[Back to top](#)

UCSC rsnumber Details

Show the rsnumber details from UCSC hg18.snp126 Table ([avHet](#), [avHetSE](#), [refUCSC](#),
[refNCBI](#)) [REF1, REF2]

Select Miminum Average Heterozygosity Cutoff ([avHet](#))

(0 <= [avHet](#) <= 1.0)

Select 'Validation Types' ([valid](#))

by-2hit-2allele (1,692,687)

by-cluster (1,154,345)

by-frequency (1,933,537)

by-submitter (214,482)

by-hapmap (9)

unknown (1,755,067)

and or

Select 'Function Types' (**func**)

[Back to top](#)

- locus (211,913)
- coding (90,767)
- coding-synon (40,422)
- coding-nonsynon (50,572)
- untranslated (92,688)
- intron (2,848,608)
- splice-site (678)
- cds-reference (0)
- unknown (1,364,457)

and or

Select 'Location Types' (**loctype**)

[Back to top](#)

- exact (4,784,820)
- range (13,202)
- between (4,866)
- rangeInsertion (2,909)
- rangeSubstitution (251)
- rangeDeletion (4,866)
- unknown (0)

and or

Select 'Molecular Types' (**moltype**)

[Back to top](#)

- genomic (4,493,416)
- cDNA (54,425)
- unknown (0)

and or

This result details the match of rs2248690 with V\$AP1_C, a high quality matrix that defines the binding site motif for the AP-1 heterodimeric (c-Fos, c-Jun) transcription factor.

Notice that the only rs2248690 is identified in this very restrictive query. As specified in the results table rs2248690 is located on chromosome 3 at base position 187812781 and is polymorphic for the A and T nucleotides (observed) at a position 799 base pairs

upstream of a transcriptional start site (dist_from_tx), and 843 base pairs upstream of a translational start site (dist_from_cds) for the alpha2-Heremans-Schmid-glycoprotein (AHSG) gene (NM_001622) in its upstream region (up10k). It is shown that AHSG has 10 gene ontology terms associated with it including one specified as with the phrase “negative regulation of **insulin** receptor signaling pathway”.

Figure 61 AHSG rs2248690 A>T Delta-MATCH Scores

The -799T allele is the reference base in both the UCSC and NCBI genome alignment. In this example the AHSG -799T allele (a1) has an optimal match score (m1 = 1.0) for the V\$AP1_C matrix, whereas the -799A allele (a2) has a much lower MATCH score (m2 = 0.8073), one that is clearly below the matrix-specific cutoff threshold (false positive cutoff = 0.998).

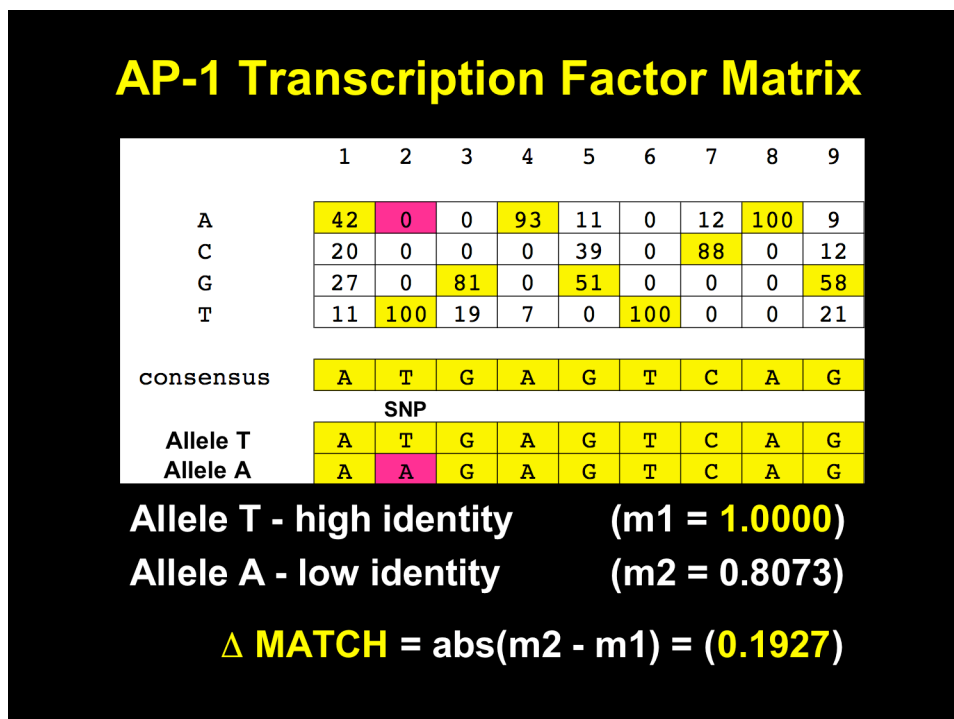


Figure 62 Pressing the p1_window Button (chr3:187812781-187812789)

The 9 bases of the V\$AP1_C matrix aligned with both of the -799 A and T alleles from base 187812781 to 187812789 along the plus stand (strand = '+') (s1 = s2 = "+") of chromosome 3. Thus the polymorphic base aligned with the second base position of the transcription factor matrix (p1 = p2 = -1).

UCSC Genome Browser on Human Mar. 2006 Assembly

<<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chr3:187,812,781-187,812,789 jump clear size 9 bp. configure

chr3 (q27.3) [chromosome ideogram]

chr3: ---->	A	T	G	A	G	T	C	A	G
	deltamatch.org allele-specific TF binding site for V\$NFKB_Q6								
	Human mRNAs from GenBank								
	Vertebrate Multiz Alignment & PhastCons Conservation (28 Species)								
Mammal Cons	[conservation bars]								
rs2248698	Simple Nucleotide Polymorphisms (dbSNP build 126)								

move start < 2.0 > Click on a feature for details. Click on base position to zoom in around cursor. Click gray/blue bars on left for track options and descriptions. move end < 2.0 >

default tracks hide all manage custom tracks configure reverse refresh

Figure 63 Density Plot of the Allelic MATCH Scores for 4,547,844 Polymorphisms (AP-1)

This is a density plot of the distribution of the allelic MATCH scores for 4,547,844 polymorphisms using the AP-1 transcription factor binding site matrix V\$AP1_C. Most polymorphisms have small differences between their allele 1 (m1) and allele 2 (m2) MATCH scores. The dotted lines represent the minimum MATCH score (FP = 0.998) required to initiate transcription factor binding for the specified matrix.

The 1,321 polymorphisms having a MATCH score (m1 and/or m2) greater than or equal to the false positive cutoff threshold score (FP = 0.998) were ranked by the Delta-MATCH algorithm to identify those polymorphisms with the highest potential to create an allele-specific AP-1 binding site.

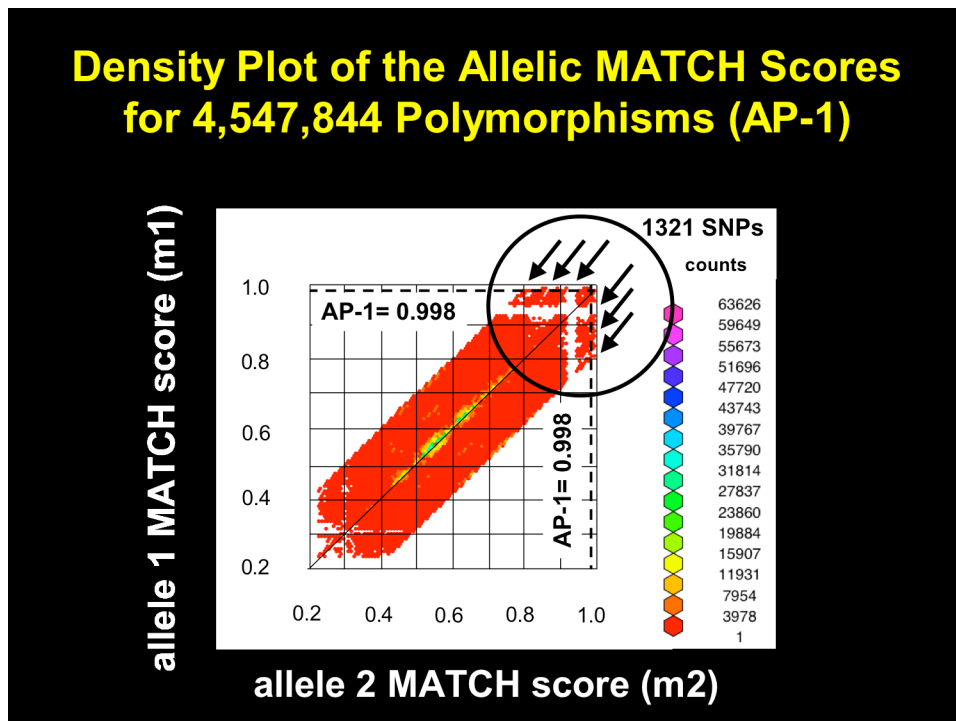


Figure 64 AHSG rs2248690 A>T Ranks 747th for AP-1 TFBS Matrix (V\$AP1_C)

The “potential” score for this polymorphism/matrix pair is 1.0. Only 747 other polymorphisms in the Delta-MATCH database have a potential score of this magnitude or greater for the V\$AP1_C matrix. This gives this polymorphism a rareness equal to 1.6425×10^{-4} (rareness = 747/4,547,844).

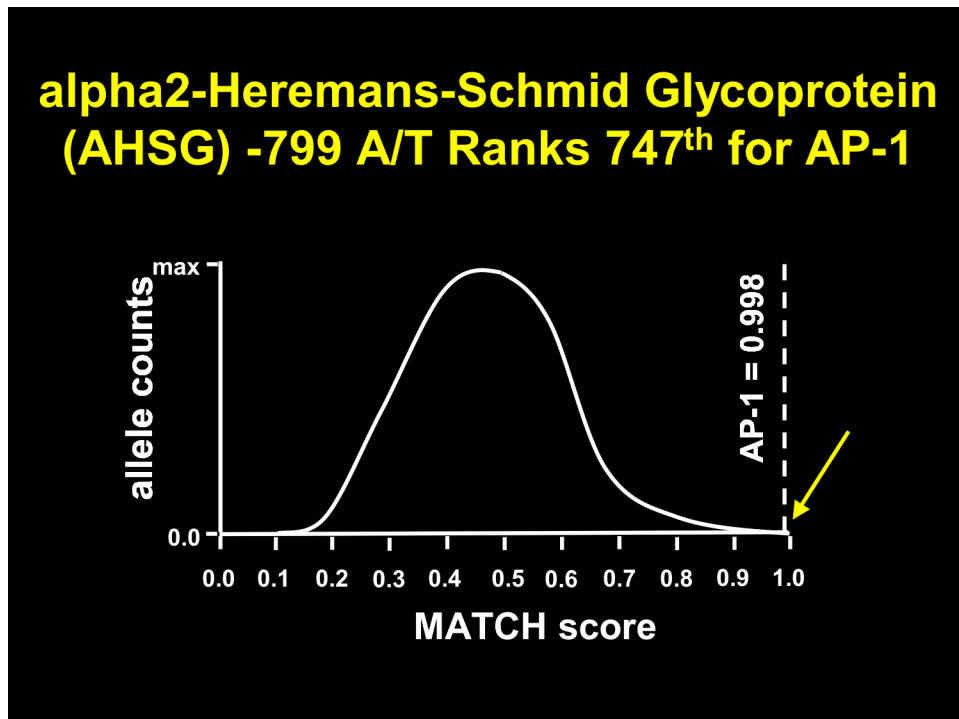


Figure 65 AHSG rs2248690 Is in Linkage Disequilibrium with Other SNPs that Associate with Type 2 Diabetes

It is noteworthy Delta-MATCH identifies other polymorphisms that are in strong linkage disequilibrium (LD) with rs2248690 in the Japanese population. Of these rs2077119, rs4917 and rs4918 have been previously associated with AHSG mRNA levels in a Japanese population or associated with Type 2 Diabetes in a population of French Caucasians [29, 30]. Furthermore, the 3q27 region was recently associated with Type 2 Diabetes in an independent genome wide association study [31].

hit	rsnumber	chrom	chromStart	number_hits	ld_name	ld_number_hits	ld_ld	ld_dprime	ld_rsquare	ld_pos_dif	ld_pos1_hg17	ld_pos2_hg17	ld_fbin
1	rs2248690	chr3	187812781	7	(1)rs2077119	(1)0 (2)0 (3)1	2.50	1.000	0.1730	374 2483	187812790	187813164	1878
					(2)rs2593813	(4)0 (5)0 (6)0	5.03	1.000	0.4740	5853 6088	187812790	187815273	1878
					(3)rs2070633	(7)0 (8)1 (9)3	3.29	1.000	0.2180	7625 7739	187812790	187818643	1878
					(4)rs2070635		3.43	1.000	0.2310	8047 8145	187812790	187818878	1878
					(5)rs4917		6.48	1.000	0.5730	8294	187812790	187820415	1878
					(6)rs2518136		2.81	1.000	0.2020		187812790	187820529	1878
					(7)rs1900618		6.54	1.000	0.6100		187812790	187820837	1878
					(8)rs1029353		3.06	1.000	0.1880		187812790	187820935	1878
					(9)rs4918		6.48	1.000	0.5730		187812790	187821084	1878

1.11.1.1 Example OMIM Links for ASHG

ALPHA-2-HS-GLYCOPROTEIN

<http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=138680>

JUN

<http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=165160>

FOS

<http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=164810>

1.11.2 Delta-MATCH Query Examples (Easy Mode)

In the following Delta-MATCH examples, assume that all header name checkboxes are in the default condition, unless otherwise specified. Please refresh your browser window to clear any previous settings before beginning each example (NOTE: The Firefox browser will not reset the button and box settings pressing the refresh button. To refresh the page in Firefox, close the current window and open another).

1.11.3 Example 1 - Single Transcription Factor Matrix Name (the default submission)

This example returns the top 5 hits for the specified “high” quality NF-kB transcription factor matrix (V\$NFKB_Q6) where the “potential” score is greater than or equal to 0.8.

Figure 66 Input Parameters for Example 1

STEP 1 - (1) Single Transcription Factor Matrix Name = “V\$NFKB_Q6 (950)”

STEP 2 - Minimum Potential Score (checked) = “0.8”

STEP 2 - Top Most Significant Hits (checked) = “5”

STEP 2 - Matrix Quality (checked) = “high”

1 - Single Transcription Factor Matrix Name

[Goto submit](#)

Select a 'Single Transcription Factor Matrix Name' (n=550)
 (mat_id)

Minimum Potential Score

[Back to top](#)

Select a 'Minimum Potential Score' (potential)
 (0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix
 Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (qual) [REF]
 ('high = 1', 'low = 0')

The Delta-MATCH results table for Example 1 has 14 unique columns: **hit**, **rsnumber**, **chrom**, **chromStart**, **mat_id**, **factor**, **potential**, **threshold**, **m1**, **m2**, **m_per**, **rank**, **p1_window**, and **pubmed**. In this example there are 5 rsnumbers returned when the searched against the single transcription factor matrix named “V\$NFkB_Q6”. Note that only the first 5 rsnumbers with the highest potential scores are returned even though there are 950 that could have been returned for this matrix name (mat_id). This is because the “Top Most Significant Hits” is set to a maximum limit of 5.

1.11.4 Figure - Example 1 Results Table

hit	rsnumber	chrom	chromStart	factor	mat_id	potential	threshold	m1	m2	m_per	rank	p1_window	pubmed	hit
1	rs3093317	chr16	27351578	NF-kappaB	V\$NFkB_Q6	1	0.955	0.8377	1.0000	16.23	7	chr16:27351571-27351584	rs3093317	1
2	rs8030978	chr15	64651526	NF-kappaB	V\$NFkB_Q6	1	0.955	0.8377	1.0000	16.23	2	chr15:64651519-64651532	rs8030978	2
3	rs1775044	chr1	7418248	NF-kappaB	V\$NFkB_Q6	1	0.955	1.0000	0.8795	12.05	4	chr1:7418246-7418259	rs1775044	3
4	rs7296179	chr12	100126053	NF-kappaB	V\$NFkB_Q6	1	0.955	0.8795	1.0000	12.05	6	chr12:100126043-100126056	rs7296179	4
5	rs6031444	chr20	42249151	NF-kappaB	V\$NFkB_Q6	1	0.955	1.0000	0.8895	11.05	1	chr20:42249149-42249162	rs6031444	5

1.11.4.1 Definition - hit (Delta-MATCH rsnumber row in the query result table)

This is any rsnumber row returned to the browser by a successful Delta-MATCH query. In the results table, the number under the “hit” column is simply an ascending row number, one for each returned rsnumber result. This number has no particular meaning

but can be changed by using the sorting function (“Sort Results Table” checkbox in Expert Mode). By default the highest number in this column could be 1,500 (see Figure 206 Delta-MATCH Error 2- more than 1,500 rsnumbers passed your selected criteria, page 409). Every number in the hit column is hyperlinked to the dbSNP entry for the corresponding rsnumber (example [rs3093317](#)).

1.11.4.2 Definition - rsnumber (dbSNP accession)

This is the dbSNP accession number (hg18.snp126.name). Every rsnumber in the results table is hyperlinked to its entry in the UCSC Genome Browser for the human genome March, 2006 (Figure 85 page 143).

1.11.4.3 Definition - chrom (chromosome)

This is the chromosome that a polymorphism is mapped to (hg18.snp126.chrom).

1.11.4.4 Definition - chromStart (polymorphism starting base position)

This is the leftmost base position of a polymorphism relative to the plus sense strand of the chromosome (hg18.snp126.chromStart).

1.11.4.5 Definition - factor (transcription factor name)

This is the name of the transcription factor in the BIOBASE TRANSFAC database (dm2_5_million.matrix_tf10_2.factor).

1.11.4.6 Definition - mat_id (matrix name)

This is the matrix name as specified by the BIOBASE TRANSFAC database (dm2_5_million.matrix_tf10_2.mat_id). The hyperlink for this value will take you to the

[local copy](#) of the BIOBASE TRANSFAC database where the details for the given matrix can be seen. This page is only accessible from within the J. David Gladstone Institutes.

1.11.4.7 Definition - potential (Delta-MATCH Potential Score)

The Delta-MATCH Potential Score for a given polymorphism for a given transcription factor matrix is the absolute difference in biological relevance between two to polymorphic alleles (Definition page 30).

1.11.4.8 Definition - threshold (FP = false positive cutoff threshold)

This is the false positive threshold (FP) specified by the BIOBASE TRANSFAC database. Note that potential cells are colored white if the MATCH score is 0.0, colored darkest red when the potential is equal to 1.0, or colored proportionally red for intermediate values (dm2_5_million. matrix_tf10_2.FP) (Definition page 31).

1.11.4.9 Definition - m1 (highest MATCH score for allele 1)

This is the magnitude of the “highest calculated BIOBASE MATCH score” for allele 1 (Definition page 71). The m1 and m2 cells are colored white if the allelic MATCH score (m1 or m2) is less than or equal to the matrix’s false positive (FP) cutoff threshold. These cells may be colored proportionally when if the MATCH score (m1 or m2) is greater than the false positive cutoff, but less than 1.0, and colored darkest red when the MATCH score is 1.0.

1.11.4.10 Definition - m2 (highest MATCH score for allele 2)

This is the magnitude of the “highest calculated BIOBASE MATCH score” for allele 2 (Definition page 71).

1.11.4.11 Definition - **m_per** (absolute percent difference in MATCH score)

This is the percent difference between the in BIOBASE MATCH scores for allele 1 and allele 2. (Equation 9 page 38) When two different rsnumbers have the same potential score, but different **m_per** values, the rsnumber with the larger **m_per** value might be considered more probable to create an allele-specific transcription factor binding site.

This is the case when two different rsnumbers each have the same larger polymorphism MATCH score (**m_max**) (Definition page 37), but different values for the smaller polymorphism MATCH score (**m_min**) (Definition page 37). In these cases, the minimum MATCH score for both rsnumbers are less than or equal to the false positive (FP) cutoff, but different from each other. Although the potential scores are the same for two rsnumbers like these, the biological relevance of the two polymorphisms may be different, particularly if the false positive cutoff value provided by the BIOBASE team underestimates the true cutoff threshold (See Equation 8, page 38).

1.11.4.12 Definition - **rank**

The rank is the actual order this polymorphism **rsnumber** is listed internally in the Delta-MATCH database table for the corresponding **mat_id**. The lowest numbered ranks will have the highest **potential** scores. It should be noted that two or more rsnumbers with the same **potential** and **m_per** values will have different ranks out of necessity, but may be considered equally important. It is suggested to sort equivalent potential scores secondarily by their “**m_per**” value. Of those rsnumbers with equivalent potential scores, the ones with the larger percentage difference between the allele 1 and allele 2 MATCH scores are may be ranked higher (page 39).


1.11.4.13 Definition - p1_window (UCSC position window of the highest allele 1 MATCH score)

This is a link to the physical position best MATCH of the corresponding mat_id to allele 1 in the UCSC genome browser (human March 2006 Assembly, hg18) (see Example 19, Figure 94 page 179). The number of bases shown in the linked page should match the exact length (mat_len) and position of the highest scoring matrix (mat_id) alignment. If the UCSC Browser track called “SNPs(126)” is set to (dense, squish, pack, full) in the genome browser, the physical position of the polymorphism will be visible within the p1_window.

1.11.4.14 Definition - pubmed (link to PubMed citations)

This is a hyperlink link to any pubmed literature citations for the associated rsnumber.

Figure 67 ASHG rs22486890 A>T Proof of Concept PubMed link (pubmed)

- 1: [Inoue M, Takata H, Ikeda Y, Suehiro T, Inada S, Osaki F, Arai K, Kumon Y, Hashimoto K.](#) Related Article
 A promoter polymorphism of the alpha2-HS glycoprotein gene is associated with its transcriptional activity.
Diabetes Res Clin Pract. 2008 Jan;79(1):164-70. Epub 2007 Sep 24.
PMID: 17889958 [PubMed - indexed for MEDLINE]


- 2: [Siddiq A, Lepretre F, Hercberg S, Froguel P, Gibson F.](#) Related Article
 A synonymous coding polymorphism in the alpha2-Heremans-schmid glycoprotein gene is associated with type 2 diabetes in French Caucasians.
Diabetes. 2005 Aug;54(8):2477-81.
PMID: 16046317 [PubMed - indexed for MEDLINE]

Figure 68 rs3093317 Hyperlink to the UCSC Human Genome Browser
(hg18.snp126)

Simple Nucleotide Polymorphisms (dbSNP build 126) (snp126)
[rs3093317 at chr16:27351329-27351829](#)

SNPs from the CEU Population (hapmapSnpsCEU)
[rs3093317 at chr16:27351329-27351829](#)

SNPs from the CHB Population (hapmapSnpsCHB)
[rs3093317 at chr16:27351329-27351829](#)

SNPs from the JPT Population (hapmapSnpsJPT)
[rs3093317 at chr16:27351329-27351829](#)

SNPs from the YRI Population (hapmapSnpsYRI)
[rs3093317 at chr16:27351329-27351829](#)

Orthologous Alleles from Chimp (panTro2) (hapmapAllelesChimp)
[rs3093317 at chr16:27351329-27351829](#)

Orthologous Alleles from Macaque (rheMac2) (hapmapAllelesMacaque)
[rs3093317 at chr16:27351329-27351829](#)

Illumina Human Hap 550v3 (snpArrayIllumina550)
[rs3093317 at chr16:27351329-27351829](#)

Illumina Human Hap 650v3 (snpArrayIllumina650)
[rs3093317 at chr16:27351329-27351829](#)

1.11.5 Example 2 - List of Transcription Factor Matrix Names

This example returns the top 5 hits for each of the two specified “high” quality transcription factor matrixes (V\$NFKB_Q6, V\$NFKB_C) where the potential score is greater than or equal to 0.8. A total of 10 results are returned.

Figure 69 Input Parameters for Example 2

STEP 1 - (2) List of Transcription Factor Matrix Names = “V\$NFKB_Q6, V\$NFKB_C”

STEP 2 - Minimum Potential Score (checked) = “0.8”

STEP 2 - Top Most Significant Hits (checked) = “5”

STEP 2 - Matrix Quality (checked) = “high”

2 - List of Transcription Factor Matrix Names

Hand-type a comma-separated 'List of Transcription Factor Matrixes Names'
 (mat_id) (1024 chars)

Minimum Potential Score

[Back to top](#)

Select a 'Minimum Potential Score' (potential)
 (0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix
 Limit the number of polymorphisms seached per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (qual) [REF]
 ('high = 1', 'low = 0')

1.11.6 Example 3 - Transcription Factor Name

This example returns 25 hits, the top 5 hits for each of the six “high” quality NF-kappaB matrixes where their potential score is greater than or equal to 0.8. Note that no results for the matrix “V\$NF-KB50_01” were returned because none had potential scores greater than or equal to 0.8.

Figure 70 Input Parameters for Example 3

STEP 1 - (3) Transcription Factor Name = “NK-kappaB”

STEP 2 - Minimum Potential Score (checked) = “0.8”

STEP 2 - Top Most Significant Hits (checked) = “5”

STEP 2 - Matrix Quality (checked) = “high”

3 - Transcription Factor Name

Select transcription factor matrix names by a 'Transcription Factor Name' (n=351)

(factor)

Minimum Potential Score

[Back to top](#)

Select a 'Minimum Potential Score' (potential)

(0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix

Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (qual) [REF]

('high = 1', 'low = 0')

1.11.7 Example 4 - Tissue-Specific Transcription Factor Names

This example returns 212 hits, after searching the 54 “high” quality matrixes defined in the list for the “immune cell-specific” tissue type where the potential score is greater than or equal to 0.8. There were 59 other immune cell-specific transcription factor matrixes that were excluded from the search because they were “low” quality. The results are ordered alphabetically by mat_id.

Figure 71 Input Parameters for Example 4

STEP 1 - (4) Tissue-Specific Transcription Factor Names = “immune_cell_specific (n=113)”

STEP 2 - Minimum Potential Score (checked) = “0.8”

STEP 2 - Top Most Significant Hits (checked) = “5”

STEP 2 - Matrix Quality (checked) = “high”

4 - Tissue-Specific Transcription Factor Names

Select transcription factor matrix names by a 'Tissue Type'

Minimum Potential Score

[Back to top](#)

Select a 'Minimum Potential Score' (**potential**)

(0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix

Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (qual) [REF]

('high = 1', 'low = 0')

1.11.8 Example 5 - Top Most Significant Hits (unchecked)

This is like Example 1, except all (61) V\$NFKB_Q6 hits with a potential greater than or equal to 0.8 are returned because the “Top Most Significant Hits” is unchecked.

Figure 72 Input Parameters for Example 5

STEP 1 - (1) Single Transcription Factor Matrix Name = “V\$NFKB_Q6 (950)”

STEP 2 - Minimum Potential Score (checked) = “0.8”

STEP 2 - Top Most Significant Hits (**unchecked**)

STEP 2 - Matrix Quality (checked) = “high”

1 - Single Transcription Factor Matrix Name

[Goto submit](#)

Select a 'Single Transcription Factor Matrix Name' (n=550)

(mat_id)

Minimum Potential Score

[Back to top](#)

Select a 'Minimum Potential Score' (potential)

(0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix

Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (qual) [REF]

high ▾ ('high = 1', 'low = 0')

1.11.9 Example 6 - Minimum Potential Score (unchecked)

This is like Example 5 except “Minimum Potential Score” is unchecked. There are 950 polymorphisms returned for the single transcription factor matrix named “V\$NFKB_Q6”. The same 950 could have been returned if the “Minimum Potential Score” was left checked but set to “0.0”. Notice the change in color intensities in the “potential” column as you scroll down through the list. At the bottom of the list are those rsnumbers with the lowest potential scores. Interestingly, the highest MATCH scores (m1 and m2) for these polymorphisms may be either high or low.

Figure 73 Input Parameters for Example 6

STEP 1 - (1) Single Transcription Factor Matrix Name = “V\$NFKB_Q6 (950)”

STEP 2 - Minimum Potential Score (**unchecked**)”

STEP 2 - Top Most Significant Hits (**unchecked**)

STEP 2 - Matrix Quality (checked) = “high”

1 - Single Transcription Factor Matrix Name

[Goto submit](#)

Select a 'Single Transcription Factor Matrix Name' (n=550)

V\$NFKB_Q6 (950) ▾ (mat_id)

Minimum Potential Score

[Back to top](#)

Select a 'Minimum Potential Score' (potential)

0.80 ▾ (0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix

Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (**qual**) [[REF](#)]

('high = 1', 'low = 0')

1.11.10 Example 7- Error 1 - no matrixes passed your selected criteria

This is like Example 1 (Figure 205 page 409), except the “Matrix Quality is set to “low”.

This example will return Error 1 because the only matrix name (V\$NFKB_Q6) selected is actually a “high quality” matrix and does not pass the “Matrix Quality” (low) requirement (see page 408 for the “List of Delta-MATCH Errors” in the Appendix).

Figure 74 Input Parameters for Example 7

STEP 1 - (1) Single Transcription Factor Matrix Name = “V\$NFKB_Q6 (950)”

STEP 2 - Minimum Potential Score (checked) = “0.8”

STEP 2 - Top Most Significant Hits (checked) = “5”

STEP 2 - Matrix Quality (checked) = “low”

1 - Single Transcription Factor Matrix Name

[Goto submit](#)

Select a 'Single Transcription Factor Matrix Name' (n=550)

V\$NFKB_Q6 (950) (mat_id)

Minimum Potential Score

[Back to top](#)

Select a 'Minimum Potential Score' (potential)

0.80 (0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix

5 Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (qual) [REF]

low ('high = 1', 'low = 0')

1.11.11 Example 8 - Error 2 - more than 1,500 results returned

When the single transcription factor matrix named “V\$ATF4_Q2” selected in STEP 1, and no other boxes are selected in STEP 2, **Error 2** (Figure 206 page 409) is returned preceding a results table of that has **only the first 1,500 results**. Note this example may take a couple of minutes to run.

Figure 75 Input Parameters for Example 8

STEP 1 - (1) Single Transcription Factor Matrix Name = “V\$ATF4_Q2 (1617)”

STEP 2 - Minimum Potential Score (**unchecked**)”

STEP 2 - Top Most Significant Hits (**unchecked**)

STEP 2 - Matrix Quality (checked) = “high”

1 - Single Transcription Factor Matrix Name

[Goto submit](#)

Select a 'Single Transcription Factor Matrix Name' (n=550)

V\$ATF4_Q2 (1617) (mat_id)

Minimum Potential Score

[Back to top](#)

Select a 'Minimum Potential Score' (potential)

0.80 (0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rnumbers per selected matrix

5 Limit the number of polymorphisms seached per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (qual) [REF]

high ('high = 1', 'low = 0')

1.11.12 Example 9 - Error 3 - no rsnumbers were found

When the single transcription factor matrix named “V\$ACAAT_B (0)” is selected in STEP 1, and no other boxes are selected in STEP 2, **Error 3** (Figure 207 page 410) is returned because there are no rsnumber results associated with the selected set of matrixes in the Delta-MATCH database. This could have predicted because this matrix name has a zero in parentheses next it in the drop down menu of STEP 1 -1.

Figure 76 Input Parameters for Example 9

STEP 1 - (1) Single Transcription Factor Matrix Name = “V\$ACAAT_B (0)”

STEP 2 - Minimum Potential Score (checked) = “0.8”

STEP 2 - Top Most Significant Hits (checked) = “5”

STEP 2 - Matrix Quality (checked) = “high”

1 - Single Transcription Factor Matrix Name

[Goto submit](#)

Select a 'Single Transcription Factor Matrix Name' (n=550)

V\$ACAAT_B (0) (mat_id)

Minimum Potential Score

[Back to top](#)

Select a 'Minimum Potential Score' (potential)

0.80 (0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix

5 Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (qual) [REF]

high ('high = 1', 'low = 0')

1.11.13 Example 10 - Searching by rsnumbers and Sorting By Chromosomal Position

This example will search for all “Delta-MATCH hits” predictions specific to the two polymorphisms rs5743836 and rs6031444. After searching all 550 matrixes 12 results were found, 8 for rs6031444, and 4 for rs5743836. These results are ordered by base position, thus placing all the results for a given rsnumber next to each other. If the ‘Matrix Quality’ had been checked and set to “high” in this example, there would have been 9 results after searching only the 367 high quality matrixes. Rerunning this query using the log file will cause Error 3 (Figure 207 page 410).

Figure 77 Input Parameters for Example 10

STEP 1 - (5) All Transcription Factor Matrix Names

STEP 2 - Minimum Potential Score (**unchecked**)

STEP 2 - Top Most Significant Hits (checked) = “5”

STEP 2 - Matrix Quality (**unchecked**)

STEP 2 - Sort Results Table (**checked**) = “chrom asc, position asc (a)”

STEP 2 - Search By rsnumbers (**checked**); rsnumbers = “rs5743836, rs6031444”;

rsnumber Window (unchecked)

5 - All Transcription Factor Matrix Names

Select all matrix names in the database (n=550)

Minimum Potential Score

[Back to top](#)

Select a 'Minimum Potential Score' (potential)

0.80 (0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix

Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (**qual**) [[REF](#)]

('high = 1', 'low = 0')

Sort Results Table

Sort the results table by

(asc = ascending, desc = descending)

[Back to top](#)

Search By rsnumbers

Either limit results by a comma-separated list of dbSNP rsnumbers (1024 chars max)

[[REF1](#), [REF2](#)]

'rsnumbers'

Or upload list of rsnumbers in a plain text file

'rsnumber filename' [download example file](#)

(one rsnumber per row, 10,000 rsnumbers max)

NOTE - The 'rsnumber' text field will be ignored if an 'rsnumber filename' is uploaded

Search additional bases upstream and downstream of specified rsnumbers

'rsnumber Window'

Include other rsnumbers within this many bases

1.11.14 Example 11 - Using the “rsnumber Window” checkbox

This example will search for all “high” quality matrixes for any polymorphisms that are located within 2000 base pairs upstream or downstream of the two selected polymorphisms (rs5743836 and rs6031444). After searching 367 matrixes. There are 15 results found for a total of 7 distinct rsnumbers (the two specified, plus 5 found by proximity). Rerunning this query using the log file will cause Error 3 (Figure 207 page 410).

Figure 78 Input Parameters for Example 11

STEP 1 - (5) All Transcription Factor Matrix Names

STEP 2 - Minimum Potential Score (**unchecked**)

STEP 2 - Top Most Significant Hits (**unchecked**)

STEP 2 - Matrix Quality (checked) = “high”

STEP 2 - Sort Results Table (**checked**) = “chrom asc, position asc (a)”

STEP 2 - Search By rsnumbers (**checked**); rsnumbers = “rs5743836, rs6031444”;

rsnumber Window (**checked**) = “2000”

5 - All Transcription Factor Matrix Names

Select all matrix names in the database (n=550)

Minimum Potential Score

[Back to top](#)

Select a 'Minimum Potential Score' (**potential**)

(0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix

Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (**qual**) [[REF](#)]

('high = 1', 'low = 0')

Sort Results Table

Sort the results table by

(asc = ascending, desc = descending)

Search By rsnumbers

[Back to top](#)

Either limit results by a comma-separated list of dbSNP rsnumbers (1024 chars max)

[[REF1](#), [REF2](#)]

'rsnumbers'

Or upload list of rsnumbers in a plain text file

'rsnumber filename' [download example file](#)

(one rsnumber per row, 10,000 rsnumbers max)

NOTE - The 'rsnumber' text field will be ignored if an 'rsnumber filename' is uploaded

Search additional bases upstream and downstream of specified rsnumbers

'rsnumber Window'

Include other rsnumbers within this many bases

156

1.11.15 Example 12 - Uploading a File of rsnumbers

This example will search all “NF-kappaB”-related hits for the 10 rsnumbers that are uploaded from the downloadable example file (“test_10.txt”). After searching the 6 NF-kappaB matrixes, 32 results are found and ordered by position, thus placing all the results for a given rsnumber next to each other. Rerunning this query using the log file will cause Error 3 (Figure 207 page 410).

Figure 79 Input Parameters for Example 12

STEP 1 - (3) Transcription Factor Name “NF-kappaB”

STEP 2 - Minimum Potential Score (**unchecked**)

STEP 2 - Top Most Significant Hits (**unchecked**)

STEP 2 - Matrix Quality (**unchecked**)

STEP 2 - Sort Results Table (**checked**) = “chrom asc, position asc (a)”

STEP 2 - Search By rsnumbers (checked); uploaded filename = “test_10.txt”

3 - Transcription Factor Name

Select transcription factor matrix names by a 'Transcription Factor Name' (n=351)

NF-kappaB (factor)

Minimum Potential Score

[Back to top](#)

Select a 'Minimum Potential Score' (potential)

0.80 (0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix

5 Limit the number of polymorphisms seached per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (**qual**) [[REF](#)]

('high = 1', 'low = 0')

Sort Results Table

Sort the results table by

(asc = ascending, desc = descending)

Search By rsnumbers

[Back to top](#)

Either limit results by a comma-separated list of dbSNP rsnumbers (1024 chars max)

[[REF1](#), [REF2](#)]

'rsnumbers'

Or upload list of rsnumbers in a plain text file

'rsnumber filename' [download example file](#)

(one rsnumber per row, 10,000 rsnumbers max)

NOTE - The 'rsnumber' text field will be ignored if an 'rsnumber filename' is uploaded

Search additional bases upstream and downstream of specified rsnumbers

'rsnumber Window'

Include other rsnumbers within this many bases

1.11.16 Example 13 - 'Search By Gene Names' Without Returning Results (mock search when unsure of true gene names)

In this example, each specified gene name is searched to see if it exists. For every gene name that is found, every associated entry is printed showing the corresponding field names (**name**, **name2**, **proteinID**), the chromosome (**chrom**) the gene is located on, the strand (**strand**) of DNA that the gene is transcribed on, and the associated transcript's starting and end positions (**txStart**, **txEnd**). Multiple entries may be found for a single gene name if more than one transcript is associated with the gene. In this example, two entries are found for JPH2 (NM_020433, NM_175913), 3 entries are found for PLAT (M_000930, NM_000931, NM_033011), and 2 entries are found for TLR9 (NM_138688, NM_017442). For each gene, all of the entries are compared to identify the leftmost and rightmost base positions of all of the transcripts relative to the plus sense DNA strand. Once these minimum (**min_txStart**) and maximum (**max_txEnd**) positions have been identified, all of the rsnumbers within each gene window is found and returned if they pass the remaining input criteria. When the "Gene Window" box is checked and set to 2000, the minimum and maximum base positions are extended to include the additional number of base pairs (**new_start**, **new_end**), there by increasing the gene window by a total of 4000 bases (2000 upstream and downstream) for each gene. After the entries are printed, a short summary of the search results are presented that show which of the gene names were found (Figure 82 page 161). If a "bad gene name" is submitted, or if there is a mismatch between the "UCSC hg18 Table Name" and the "Field Name" combination submitted, you may receive a notice of which gene names were, and were not found (Figure 83 page 162). When no GENE NAMES are found, Error 7 is returned (Figure 211 page 412).

Figure 80 Input Parameters for Example 13

STEP 1 - (5) All Transcription Factor Matrix Names

STEP 2 - Minimum Potential Score (checked) = "0.35"

STEP 2 - Top Most Significant Hits (**unchecked**)

STEP 2 - Matrix Quality (checked) = "high"

STEP 2 - Sort Results Table (**checked**) = "chrom asc, position asc (a)"

STEP 2 - Search By Gene Names (**checked**); **Search for gene without returning results (checked)**; Gene Names = "JPH2, PLAT, TLR9"; UCSC hg18 Table Name = "refGene"; Field Name = "name2"; Gene Window (checked) = "2000"

5 - All Transcription Factor Matrix Names

Select all matrix names in the database (n=550)

Minimum Potential Score

[Back to top](#)

Select a 'Minimum Potential Score' (**potential**)

(0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix

Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (**qual**) [REF]

('high = 1', 'low = 0')

Sort Results Table

Sort the results table by

(asc = ascending, desc = descending)

Search By Gene Names

'Search for gene without returning results' (*MOCK SEARCH*) [[REF1](#), [REF2](#)]
UNCHECK this box to perform 'real' search after your gene names are verified

Limit results by a comma-separated list of 'Gene Names' (exact text match, 5 max)

JPH2, PLAT, TLR9 'Gene Names'

refGene (name2, name) 'UCSC hg18 Table Name'

name2 'Field Name' [download help file](#)

Search more bases upstream and downstream of specified genes

'Gene Window'

2000 Select the number of additional bases

Figure 81 Example Entry Found By a Mock Gene Name Search

JPH2 - Entry 1

name2: [JPH2](#)
 name: [NM_020433](#)
 strand: -
 chrom: chr20
 txStart: 42173750
 txEnd: 42249632
 distance: 2000
 new_start: 42171750
 new_end: 42251632
 min_txStart: 42171750
 max_txEnd: 42251632

JPH2 - Entry 2

name2: [JPH2](#)
 name: [NM_175913](#)
 strand: -
 chrom: chr20
 txStart: 42238870
 txEnd: 42249632
 distance: 2000
 new_start: 42236870
 new_end: 42251632
 min_txStart: 42171750
 max_txEnd: 42251632

PLAT - Entry 1

name2: [PLAT](#)
 name: [NM_000930](#)
 strand: -
 chrom: chr8
 txStart: 42151911
 txEnd: 42184351
 distance: 2000
 new_start: 42149911
 new_end: 42186351
 min_txStart: 42149911
 max_txEnd: 42186351

PLAT - Entry 2

name2: [PLAT](#)
 name: [NM_000931](#)
 strand: -
 chrom: chr8
 txStart: 42151911
 txEnd: 42184351
 distance: 2000
 new_start: 42149911
 new_end: 42186351
 min_txStart: 42149911
 max_txEnd: 42186351

PLAT - Entry 3

name2: [PLAT](#)
 name: [NM_033011](#)
 strand: -
 chrom: chr8
 txStart: 42151911
 txEnd: 42184351
 distance: 2000
 new_start: 42149911
 new_end: 42186351
 min_txStart: 42149911
 max_txEnd: 42186351

TLR9 - Entry 1

name2: [TLR9](#)
 name: [NM_138688](#)
 strand: -
 chrom: chr3
 txStart: 52230137
 txEnd: 52233247
 distance: 2000
 new_start: 52228137
 new_end: 52235247
 min_txStart: 52228137
 max_txEnd: 52235247

TLR9 - Entry 2

name2: [TLR9](#)
 name: [NM_017442](#)
 strand: -
 chrom: chr3
 txStart: 52230137
 txEnd: 52235219
 distance: 2000
 new_start: 52228137
 new_end: 52237219
 min_txStart: 52228137
 max_txEnd: 52237219

Figure 82 Summary of Gene Names Found

These GENE NAMES were found (name):
NM_020433,NM_175913,NM_000930,NM_000931,NM_033011,NM_138688,NM_017442

These GENE NAMES found (name2):
JPH2,PLAT,TLR9

These GENE NAMES were NOT FOUND:

You can now go back to the input page and submit these gene names (name/name2)

Don't forget to uncheck the 'Search for Gene Without Returning Results' (MOCK SEARCH) box before resubmitting

Figure 83 Summary of Gene Names Not Found

These GENE NAMES were found (name):
NM_020433,NM_175913,NM_000930,NM_000931,NM_033011

These GENE NAMES found (name2):
JPH2,PLAT

These GENE NAMES were NOT FOUND:
bad_gene_name

1.11.17 Example 14 - 'Search By Gene Names' (includes using the “Gene Window” sub-checkbox)

This is like Example 13 except the “Search for gene without returning results” box is unchecked. It searches for any result located within three specified genes (JPH2, PLAT, TLR9) that have a Minimum Potential Score greater than or equal to 0.35, for any “high” quality matrix. The 40 results are sorted by chromosomal position. During the search, the position of each typed gene name is compared with the field names in the refGene database (hg18.refGene.name2) to identify the leftmost and rightmost positions of their mRNA transcripts (relative to the plus strand). Once the leftmost and rightmost positions have been found, these are extended by 2000 base pairs upstream and downstream to bracket a chromosomal window in which to search for rsnumbers with potential scores greater than or equal to 0.35. The output page includes a short summary of the number of rsnumbers found for each gene name (Figure 85 page 165). There are 875 rsnumbers are identified within the windows corresponding to the gene loci (these were found including the 2000 base pair upstream and downstream extensions). Specifically, 481, 358, and 36 rsnumbers are found in the gene windows for JPH2, PLAT and TLR9 respectively. When this example is rerun looking for refGenes that include a mistyped gene name (JPH2, PLAT, bad_gene_name), 36 hits are returned. Specifically, the 839 rsnumbers from the JPH2 and PLAT genes are still found, but none are found for “bad_gene_name” (Example 14B) (Figure 83 page 162).

Figure 84 Input Parameters for Example 14

STEP 1 - (5) All Transcription Factor Matrix Names

STEP 2 - Minimum Potential Score (checked) = “**0.35**”

STEP 2 - Top Most Significant Hits (**unchecked**)

STEP 2 - Matrix Quality (checked) = "high"

STEP 2 - Sort Results Table (**checked**) = "chrom asc, position asc (a)"

STEP 2 - Search By Gene Names (**checked**); **Search for gene without returning results (unchecked)**, Gene Names = "JPH2, PLAT, TLR9"; UCSC hg18 Table Name = "refGene"; Field Name = "name2"; Gene Window (checked) = "2000"

5 - All Transcription Factor Matrix Names

Select all matrix names in the database (n=550)

Minimum Potential Score

[Back to top](#)

Select a 'Minimum Potential Score' (**potential**)

(0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix

Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (**qual**) [[REF](#)]

('high = 1', 'low = 0')

Sort Results Table

Sort the results table by

(asc = ascending, desc = descending)

Search By Gene Names

[Back to top](#)

'Search for gene without returning results' (*MOCK SEARCH*) [[REF1](#), [REF2](#)]
UNCHECK this box to perform 'real' search after your gene names are verified

Limit results by a comma-separated list of 'Gene Names' (exact text match, 5 max)

'Gene Names'

'UCSC hg18 Table Name'

'Field Name' [download help file](#)

Search more bases upstream and downstream of specified genes

'Gene Window'

Select the number of additional bases

Figure 85 Summary of rsnumbers found in Gene Names

There were 40 'Delta-MATCH hits' returned

Summary - [JPH2,PLAT,TLR9](#)

There were no rsnumbers found in hg18.snp126 for this gene name

Summary - [JPH2](#)

There were 481 rsnumbers found in hg18.snp126 on chromosome chr8 between base positions 42171750 and 42251632

Summary - [PLAT](#)

There were 358 rsnumbers found in hg18.snp126 on chromosome chr8 between base positions 42149911 and 42186351

Summary - [TLR9](#)

There were 36 rsnumbers found in hg18.snp126 on chromosome chr8 between base positions 52228137 and 52237219

Summary - Genes Names

There were 875 total rsnumbers found

Figure 86 Summary of rsnumbers found in Gene Names (bad_gene_name)

There were 36 'Delta-MATCH hits' returned

Summary - [JPH2,PLAT,bad_gene_name](#)

There were no rsnumbers found in hg18.snp126 for this gene name

Summary - [JPH2](#)

There were 481 rsnumbers found in hg18.snp126 on chromosome chr8 between base positions 42171750 and 42251632

Summary - [PLAT](#)

There were 358 rsnumbers found in hg18.snp126 on chromosome chr8 between base positions 42149911 and 42186351

Summary - [bad_gene_name](#)

There were no rsnumbers found in hg18.snp126 for this gene name

Summary - Genes Names

There were 839 total rsnumbers found

1.11.18 Example 15 - Error 4 - no rsnumbers were found in the select gene names (bad gene name submission)

This example uses the default conditions (Example 1) with the addition of having the “Search by Gene Names” checkbox checked. Error 4 (Figure 208 page 411) is produced because the user has typed a gene name (“bad_gene_name”) that doesn’t exist in the UCSC hg18.refGene table. This error can be avoided by properly testing to see if “bad_gene_name” existed by preceding this query with a mock search (see Example 13).

Figure 87 Input Parameters for Example 15

STEP 1 - (1) Single Transcription Factor Matrix Name = “V\$NFKB_Q6 (950)”

STEP 2 - Minimum Potential Score (checked) = “0.8”

STEP 2 - Top Most Significant Hits (checked) = “5”

STEP 2 - Matrix Quality (checked) = “high”

STEP 2 - Search By Gene Names (**checked**); **Search for gene without returning results (unchecked)**; Gene Names = “bad_gene_name”; UCSC hg18 Table Name = “refGene”; Field Name = “name2”; Gene Window (checked) = “2000”

1 - Single Transcription Factor Matrix Name

[Goto submit](#)

Select a 'Single Transcription Factor Matrix Name' (n=550)
 (mat_id)

Minimum Potential Score

[Back to top](#)

Select a 'Minimum Potential Score' (potential)
 (0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix

Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (**qual**) [[REF](#)]

('high = 1', 'low = 0')

[Back to top](#)

Search By Gene Names

'Search for gene without returning results' (*MOCK SEARCH*) [[REF1](#), [REF2](#)]

UNCHECK this box to perform 'real' search after your gene names are verified

Limit results by a comma-separated list of 'Gene Names' (exact text match, 5 max)

'Gene Names'

'UCSC hg18 Table Name'

'Field Name' [download help file](#)

Search more bases upstream and downstream of specified genes

'Gene Window'

Select the number of additional bases

1.11.19 Example 16 - Error 6 - more than 5 gene names were submitted

In this example only 5 rsnumbers are returned, one for each of the first five genes in the submitted list. This query searches up to ten hits for a list of 7 submitted genes, and receives Error 6 (Figure 210 page 412). This error is not critical, but it warns the user that the maximum number of genes permitted to be submitted per query has been exceeded (max = 5).

Figure 88 Input Parameters for Example 16

STEP 1 - (1) Single Transcription Factor Matrix Name = "V\$NFKB_Q6 (950)"

STEP 2 - Minimum Potential Score (checked) = "0.8"

STEP 2 - Top Most Significant Hits (checked) = "10"

STEP 2 - Matrix Quality (checked) = "high"

STEP 2 - Search By Gene Names (checked); Search for Gene without returning results (unchecked), Gene Names = "JPH2, IL21R, CAMTA1, SLC5A8, RGS6, DOCK1, RXRG"; UCSC hg18 Table Name = "refGene"; Field Name = "name2"; Gene Window (unchecked) = "2000"

1 - Single Transcription Factor Matrix Name

[Goto submit](#)

Select a 'Single Transcription Factor Matrix Name' (n=550)

V\$NFKB_Q6 (950) (mat_id)

Minimum Potential Score

[Back to top](#)

Select a 'Minimum Potential Score' (potential)

0.80 (0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix

Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (**qual**) [[REF](#)]

('high = 1', 'low = 0')

Search By Gene Names

[Back to top](#)

'Search for gene without returning results' (*MOCK SEARCH*) [[REF1](#), [REF2](#)]
UNCHECK this box to perform 'real' search after your gene names are verified

Limit results by a comma-separated list of 'Gene Names' (exact text match, 5 max)

'Gene Names'

'UCSC hg18 Table Name'

'Field Name' [download help file](#)

Search more bases upstream and downstream of specified genes

'Gene Window'

Select the number of additional bases

Figure 89 Summary of the First 5 Submitted Gene Names

There were 5 'Delta-MATCH hits' returned

Summary - [JPH2,IL21R,CAMTA1,SLC5A8,RGS6,DOCK1,RXRG](#)

There were no rsnumbers found in hg18.snp126 for this gene name

ERROR 6 - more than 5 gene names were submitted

Summary - [JPH2](#)

There were 481 rsnumbers found in hg18.snp126 on chromosome chr8 between base positions 42171750 and 42251632

Summary - [IL21R](#)

There were 291 rsnumbers found in hg18.snp126 on chromosome chr8 between base positions 27319223 and 27371616

Summary - [CAMTA1](#)

There were 3573 rsnumbers found in hg18.snp126 on chromosome chr8 between base positions 6765970 and 7754350

Summary - [SLC5A8](#)

There were 173 rsnumbers found in hg18.snp126 on chromosome chr8 between base positions 100071408 and 100130120

Summary - [RGS6](#)

There were 2806 rsnumbers found in hg18.snp126 on chromosome chr8 between base positions 71467585 and 72102407

Summary - Genes Names

There were 7324 total rsnumbers found

1.12 Delta-MATCH Query Examples (Expert Mode)

1.12.1 Show the Matrix Details

When the “**Show the Matrix Details**” checkbox is checked (Example 17), extra columns are presented in the results table that detail the each particular hit. Included in the output are the columns **factor_description**, **count_ge_potential**, **mat_count**, **rareness**, **qual**, and **mat_len**. Some of the parameters may be viewed by downloading an associated text file (Table 43 page 405).

1.12.1.1 Definition - **factor_description** (expanded factor name)

This is slightly different and perhaps longer description of the transcription factor name.

1.12.1.2 Definition - **count_ge_potential** (count of hits with a potential score greater than or equal to this potential score)

This is the number of other hits for this **mat_id** that have a **potential** score greater than or equal to the potential score for this **rsnumber**. For a given **mat_id**, all **rsnumbers** with the equivalent values for “**count_ge_potential**” may be considered equally important, but those that also have the higher **m_per** are ranked higher.

1.12.1.3 Definition - **mat_count** (number of hits in the database for this matrix)

This is the total number of “biologically relevant” polymorphisms for this **mat_id** in the database. This is the number of **rsnumbers** with at least one allelic **MATCH** score greater than or equal to the corresponding matrix’s **threshold** score.

1.12.1.4 Definition - rareness (rareness of a potential score)

This is a measurement of how many other rsnumbers have a potential score greater or equal to the corresponding potential score in this database. Rareness is calculated by dividing the number of rsnumbers with a potential score less than or equal to the corresponding rsnumber's potential score by the total number of polymorphisms in the database. The lower the frequency is, the more rare the event is. However, a very low rareness value doesn't guarantee that the difference in potential score will be biologically relevant.

Equation 19 - rareness of a potential score (rareness)

$$\text{rareness} = \text{count_ge_potential} / 4,547,844$$

1.12.1.5 Definition - qual (quality of a matrix)

This is the quality of the transcription factor matrix as defined in the BIOBASE TRANSFAC database. A "high" quality matrix is equivalent to 1, and a "low" quality matrix is equivalent to 0. The quality of each mat_id is described in the linked file called "550_matrixes.txt" (Table 43 page 405).

1.12.1.6 Definition - mat_len (matrix length)

This is the length of the transcription factor binding site matrix in number of base pairs. The matrix length of each mat_id is described in the linked file called "550_matrixes.txt" (Table 43 page 405).

1.12.2 Example 17 - Show the Matrix Details

This result will show is like Example 1 (the default) except the matrix details are shown.

Figure 90 Input Parameters for Example 17

STEP 1 - (1) Single Transcription Factor Matrix Name = "V\$NFKB_Q6 (950)"

STEP 2 - Minimum Potential Score (checked) = "0.8"

STEP 2 - Top Most Significant Hits (checked) = "5"

STEP 2 - Matrix Quality (checked) = "high"

STEP 2 - Show the Matrix Details (checked)

1 - Single Transcription Factor Matrix Name [Goto submit](#)

Select a 'Single Transcription Factor Matrix Name' (n=550)
 (mat_id)

Minimum Potential Score [Back to top](#)

Select a 'Minimum Potential Score' (potential)
 (0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rnumbers per selected matrix
 Limit the number of polymorphisms seached per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (qual) [REF]
 ("high = 1", "low = 0")

Sort Results Table

Sort the results table by

(asc = ascending, desc = descending)

Show the Matrix Details

[Back to top](#)

Show the matrix details [[REF](#)]

(count_le_potential, mat_count, frequency, factor, factor_description, qual, mat_len)

Minimum Matrix Length

Limit searches to those matrixes with minimum length (mat_len)

12 (mat_len >= x)

Figure 91 Output Results Showing the Matrix Details (sorted)

hit	rsnumber	chrom	chromStart	factor	mat_id	potential	threshold	m1	m2	m_per	rank	factor_description	count_ge_potential	mat_count	rareness	qual	mat_len	p1_window	pubmed	hit
1	rs3093317	chr16	27351578	NF-kappaB	VS/NFKB_D6	1	0.955	0.8377	1.0000	16.23	7	NF-kappaB	7	950	1.5392e-6	1	14	chr16:27351571-27351584	rs3093317	1
2	rs8030978	chr15	64651526	NF-kappaB	VS/NFKB_D6	1	0.955	0.8377	1.0000	16.23	2	NF-kappaB	7	950	1.5392e-6	1	14	chr15:64651519-64651532	rs8030978	2
3	rs1775044	chr1	7418248	NF-kappaB	VS/NFKB_D6	1	0.955	1.0000	0.8795	12.05	4	NF-kappaB	7	950	1.5392e-6	1	14	chr1:7418246-7418259	rs1775044	3
4	rs7296179	chr17	100126053	NF-kappaB	VS/NFKB_D6	1	0.955	0.8795	1.0000	12.05	6	NF-kappaB	7	950	1.5392e-6	1	14	chr17:100126043-100126056	rs7296179	4
5	rs6031444	chr20	42249151	NF-kappaB	VS/NFKB_D6	1	0.955	1.0000	0.8895	11.05	1	NF-kappaB	7	950	1.5392e-6	1	14	chr20:42249149-42249162	rs6031444	5

1.12.3 Minimum Matrix Length

When both the “**Show the Matrix Details**” checkbox and the “**Minimum Matrix Length**” sub -checkbox is checked (Example 18), it is possible results to those matrixes that have minimum length (**mat_len**). The mat_len for the 550 matrixes range from 6 to 30 (Table 4 page 77).

1.12.4 Example 18 - 'Minimum Matrix Length' sub-checkbox

This is like Example 3 except the “Minimum Matrix Length” sub-checkbox is checked and set to 12 base pairs. Adding this filter limits the selection of matrixes to those that are at least 12 base pairs long and has the effect of excluding three matrixes that were each only 10 base pairs long (V\$NFKAPPAB50_01,V\$NFKAPPAB65_01, V\$NFKAPPAB_01). Only 15 results are returned.

Figure 92 Input Parameters for Example 18

STEP 1 - (3) Transcription Factor Name = “NK-kappaB”

STEP 2 - Minimum Potential Score (checked) = “0.8”

STEP 2 - Top Most Significant Hits (checked) = “5”

STEP 2 - Matrix Quality (checked) = “high”

STEP 2 - Show the Matrix Details (checked); Minimum Matrix Length (checked) = “12”

3 - Transcription Factor Name

Select transcription factor matrix names by a 'Transcription Factor Name' (n=351)

NF-kappaB (factor)

Minimum Potential Score

[Back to top](#)

Select a 'Minimum Potential Score' (potential)

0.80 (0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rnumbers per selected matrix

5 Limit the number of polymorphisms seached per matrix

Show the Matrix Details

[Back to top](#)

Show the matrix details [\[REF\]](#)

([count_le_potential](#), [mat_count](#), [frequency](#), [factor](#), [factor_description](#), [qual](#), [mat_len](#))

Minimum Matrix Length

Limit searches to those matrixes with minimum length ([mat_len](#))

([mat_len](#) >= x)

1.12.5 Show the Position Details

When the “Show the Position Details” checkbox is checked (Example 19, page 176), additional columns (**p2_window**, **p1**, **p2**, **s1**, and **s2**) are presented in the results table that detail the position and strand of each matrix MATCH for each allele of the given rsnumber.

1.12.5.1 Definition - p2_window (UCSC position window of the highest allele 2 MATCH score)

This is a link to the physical position best MATCH of the corresponding mat_id to allele 2 in the UCSC genome browser.

1.12.5.2 Definition - p1 (position offset of highest allele 1 MATCH score)

This is the position offset of the leftmost border (along the plus strand of DNA) of the mat_id match relative to the position of the rsnumber (chromStart) for allele 1. If the best MATCH of allele 1 is located on the plus (+) strand, the **first** position of the matrix match (leftmost border) will be equal to the sum of the position of the rsnumber and the position offset (leftmost border = chromStart + p1) and the **last** position of the matrix match (rightmost border) will be X bp downstream of the leftmost border, where X is the length of the matrix (mat_len). If the best MATCH of allele1 is on the minus (-), the **last** position

of the matrix match (leftmost border) will be equal to the sum of the position of the rsnumber and the position offset (leftmost border = chromStart + p1) and the **first** position of the matrix match (rightmost border) will be X bp downstream of the leftmost border, where X is the length of the matrix (mat_len).

1.12.5.3 Definition - p2 (position offset of highest allele 2 MATCH score)

This is the position offset of the leftmost border (along the plus strand of DNA) of the mat_id match relative to the position of the rsnumber (chromStart) for allele 2.

1.12.5.4 Definition - s1 (strand of highest allele 1 MATCH score)

This is the strand of the best MATCH for the mat_id with allele 1 (plus strand = "+", minus strand = "-").

1.12.5.5 Definition - s2 (strand of highest allele 2 MATCH score)

This is the strand of the best MATCH for the mat_id with allele 2 (plus strand = "+", minus strand = "-").

1.12.6 Example 19 - 'Show the Position Details'

In this example, 12 results are returned. Notice that hit 7 (rsnumber rs6031444) has MATCH (m1 and m2) aligning with the "V\$NFKB_Q6" matrix along 14 base pairs of chromosome 20, starting at base 42,249,149 and extending to base 42,249,162. For these alleles, the "V\$NFKB_Q6" matrix matches on the minus strand (-) of the chromosome with its leftmost border of the match aligning to a position that is 3 base pairs offset to the left of the position (-3) of the rsnumber's position (chromStart) relative to the plus strand. In other words, because rsnumber rs6031444 is at position

42,249,152 on chromosome 20, we know that the “V\$NFKB_Q6” matrix should have its best match when aligned on the negative strand of the chromosome so that the first position of the matrix aligns with position 42,249,162, and the last position of the matrix aligns with position 42,249,149. If the “dm_track_V\$NFKB_C.txt” file has been downloaded from the downloads page (**Delta-MATCH > Downloads > UCSC Browser Tracks**), and uploaded into the UCSC Genome Browser as a Custom Track, it is possible to view the result for rs6031444 in detail. The exact position of the aligned match can be found in the UCSC Genome Browser by clicking on the hyperlink of the p1_window or p2_window for each allele respectively (Figure 94 page 179). If your browser UCSC browser doesn't show the specified rsnumber, be sure to check to make sure the “SNPs (126)” track is set to “pack” under the “Variation and Repeats Section”. For more details, please see the section called “Uploading Delta-MATCH Data as UCSC Browser Tracks” see (page 118). Rerunning this query using the log file will cause Error 3 (Figure 207 page 410).

Figure 93 Input Parameters for Example 19

STEP 1 - (3) Transcription Factor Name = “NF-kappaB”

STEP 2 - Minimum Potential Score (**unchecked**)

STEP 2 - Top Most Significant Hits (**unchecked**)

STEP 2 - Matrix Quality (**unchecked**)

STEP 2 - Sort Results Table (checked) = “chrom asc, position asc (a)”

STEP 2 - Search By rsnumbers (**checked**); rsnumbers = “**rs1680789, rs6031444, rs2104240**”; ‘rsnumber Window’ (**unchecked**)

STEP 2 - Show the Position Details (**checked**)

3 - Transcription Factor Name

Select transcription factor matrix names by a 'Transcription Factor Name' (n=351)

NF-kappaB (factor)

Minimum Potential Score

[Back to top](#)

Select a 'Minimum Potential Score' (potential)

0.80 (0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix

5 Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (qual) [REF]

high ('high = 1', 'low = 0')

Sort Results Table

Sort the results table by

chrom asc, position asc (a)

(asc = ascending, desc = descending)

Search By rsnumbers

[Back to top](#)

Either limit results by a comma-separated list of dbSNP rsnumbers (1024 chars max)

[REF1, REF2]

rs1680789, rs6031444, rs2104240 'rsnumbers'

Or upload list of rsnumbers in a plain text file

'rsnumber filename' [download example file](#)

(one rsnumber per row, 10,000 rsnumbers max)

NOTE - The 'rsnumber' text field will be ignored if an 'rsnumber filename' is uploaded

Search additional bases upstream and downstream of specified rsnumbers

'rsnumber Window'

2000 Include other rsnumbers within this many bases

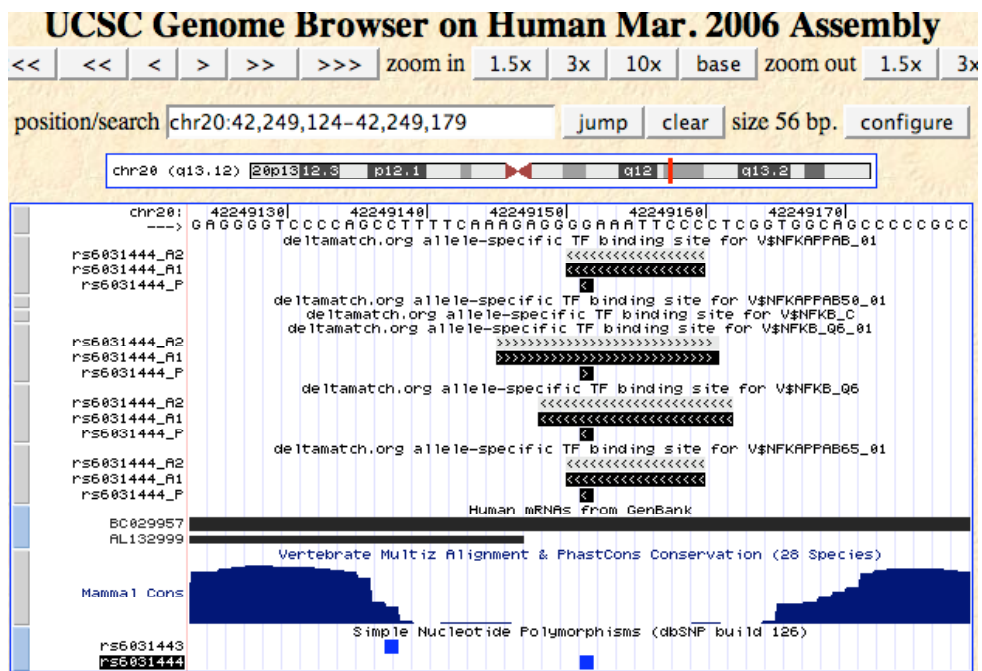
Show the Position Details

Show the position and strand details [REF]

(p1_window, p2_window, p1, p2, s1, s2)

Figure 94 UCSC Browser Example 19 (rs6031444)

In this example, there are four Delta-MATCH hits for rs6031444, one for V\$NFKAPPAB_01, V\$NFKB_Q6, V\$NFKAPPAB65_01, and V\$NFKB_Q6_01.

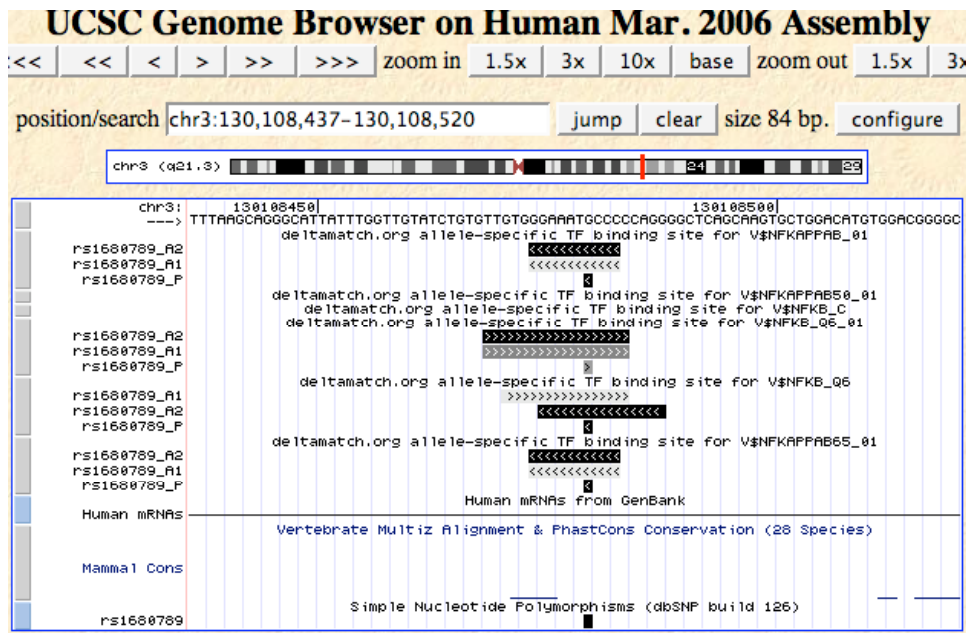


These highest MATCH scores align to the minus strand of chromosome 20 (reverse arrows). The polymorphic base aligns to the second to last matrix base position for matrix V\$NFKAPPAB_01, and the fourth from last base position for matrix V\$NFKB_Q6. The potential scores for these hits are high (rs6031444_P is very dark, and rs6031444_A1 and rs6031444_A2 are very opposite intensity). Allele 1 has higher sequence identity to each matrix than does allele 2 (is darker). It is more likely that NF-kB would bind better to the DNA sequence spanning the rs6031444 major allele (allele 1), than its minor allele (allele 2). Notice the hit for V\$NFKAPPAB50_01 is on the forward strand. It appears that rs6031444 is located only 5 base pairs upstream of an alternative mRNA initiation site. It could be hypothesized mRNA expression of transcript

AL132999 may be NFKB-dependent and correlated to the rs6031444 genotype. However, the dbSNP database suggests the rs6031444 minor allele has a very low population frequency, if it truly exists at all in the human population.

Figure 95 UCSC Browser Example 19 (rs1680789)

It is sometimes possible to find that the two alleles of a given rsnumber do not have their best respective MATCH score for a given transcription factor matrix aligning on the same strand or with the same offset position along the chromosome. This may commonly occur when the MATCH score (m1 and m2) are very different. For example when rs1680789 is matched against the “V\$NFKB_Q6” matrix (hit 11), allele 1 matches on the plus strand (p1 = -9), and allele2 matches on the minus strand (p2 = -5). This effect may be more frequently seen when the polymorphism is a type of insertion or deletion

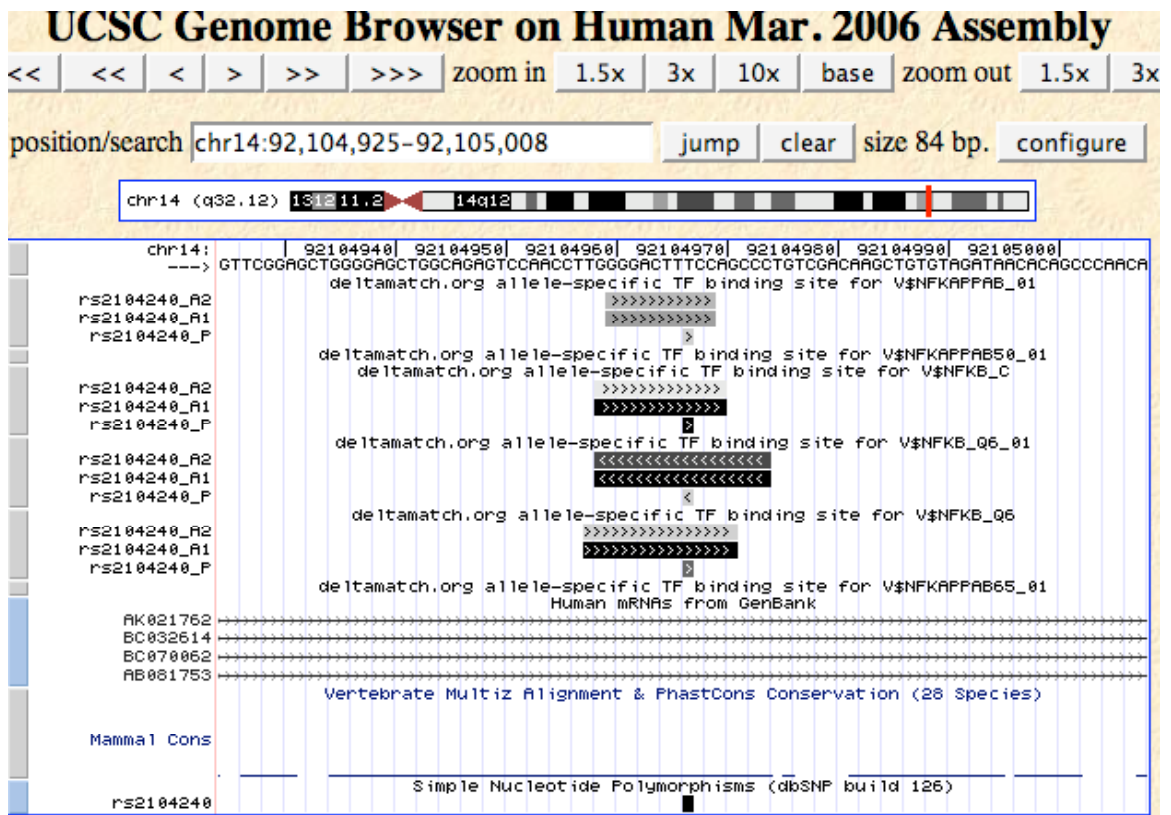


In this example the highest MATCH score for rs1680789, with matrix V\$NFKB_Q6 aligned on opposite strands (rs1680789_A1 has forward arrows, rs1680789_A2 has

reverse arrows). The potential score for these hits is high (rs6013444_P is very dark, and rs6031444_A1 and rs6031444_A2 are very opposite intensity).

Figure 96 UCSC Browser Example 19 (rs2104240)

In this example the highest MATCH score for rs2104240, with matrixes V\$NFKAPPAB_01, V\$NFKB_C, V\$NFKB_Q6, and V\$NFKB_Q6_01 are shown. The intensity of the potential score track (rsnumber_P) for these hits are proportional to the difference in intensity between the MATCH score tracks (rsnumber_A1 and rsnumber_A).



1.12.7 Chromosome

When the “**Chromosome**” checkbox is checked, the results will be filtered to include those from a single chromosome (1-22, X, Y).

1.12.8 Position Range

Users may require to only return results from a limited range of base pair positions.

When the “**Position Range**” checkbox is checked, only those polymorphisms that are positioned within the specified lowest and highest chromosomal base position will be returned.

1.12.9 Example 20 - Restricting By Chromosome and Position Range

This example returns 12 rsnumbers positioned on chromosome 8 between base pair 128,100,000 and 128,700,000 that have potential scores greater than or equal to 0.3 for any of the 384 “high” quality matrixes. The results are sorted by chromosomal position.

Figure 97 Input Parameters for Example 20

STEP 1 - (5) All Transcription Factor Matrix Names

STEP 2 - Minimum Potential Score (checked) = “0.3”

STEP 2 - Top Most Significant Hits (checked) = “5”

STEP 2 - Matrix Quality (checked) = “high”

STEP 2 - Sorted Results Table (**checked**) = “chrom asc, position asc (a)”

STEP 2 - Chromosome (**checked**) = “8”

STEP 2 - Show the Position Details (**checked**)

STEP 2 - Position Range (**checked**); lowest base = "128100000"; highest base = "128700000"

5 - All Transcription Factor Matrix Names

Select all matrix names in the database (n=550)

Minimum Potential Score

[Back to top](#)

Select a 'Minimum Potential Score' (**potential**)

(0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rnumbers per selected matrix

Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (**qual**) [[REF](#)]

('high = 1', 'low = 0')

Sort Results Table

Sort the results table by

(asc = ascending, desc = descending)

Show the Position Details

Show the position and strand details [[REF](#)]

(p1_window, p2_window, p1, p2, s1, s2)

Chromosome

[Back to top](#)

Limit results to a chromosome [[REF](#)]

(**chrom**)

Position Range

Limit results between two positions [\[REF\]](#)

Enter lowest base (chrStart >= x)

Enter highest base (chrStart <= x)

1.12.10 Strand

When the "Strand" checkbox is checked it is possible to limit results to those that where the highest MATCH scores align to the "plus" or "minus" strand of the DNA.

1.12.11 Example 21 - Strand

This is like Example 19 except only hits with a MATCH aligning to the plus strand of the genome (where s1 and/or s2 = "+") are returned. Notice that the hit for rs1680789 with V\$NFKB_Q6 is included because the match for allele1 is on the plus strand (s1 = "+"). There are 6 results found. Rerunning this query using the log file will cause Error 3 (Figure 207 page 410).

Figure 98 Input Parameters for Example 21

Same as Example 19 **plus:**

STEP 2 - Strand (checked) = "+"

Strand

Limit matrix hits to a DNA strand [\[REF\]](#)

(strand)

1.12.12 Genomic Regions

It is possible to restrict results to those polymorphisms that are located within specific genomic regions. Lists have been curated to identify all of the human polymorphisms (UCSC browser hg18.snp126.name) that are located within 9 genomic regions. The number of polymorphisms that have been identified in each of these regions is listed in parentheses next to the region name. During a query, a column for each of these regions is shown in the output results table stating if the rsnumber is located within that region (“yes”) or not (“-“). Users may select using the “or” (union) or the “and” (intersection) buttons. Using the “and” button tends to increase the run time considerably (6 - 10 minutes), as does combining many of these regions in a single query.

1.12.12.1 Definition - up10k

A list of polymorphisms positioned within 10,000 base pairs upstream of any refGene transcript.

1.12.12.2 Definition - phastconsElements17way

A list of polymorphisms positioned within 10,000 base pairs upstream of any refGene transcript. (in a region under track hg18.phastConsElements17way)

1.12.12.3 Definition - utr5

A list of polymorphisms positioned within any 5 prime untranslated region of a refGene transcript.

1.12.12.4 Definition - coding

A list of polymorphisms positioned within any coding region of a refGene transcript.

1.12.12.5 Definition - down10k

A list of polymorphisms positioned within 10,000 base pairs downstream of any refGene transcript

1.12.12.6 Definition - exons

A list of polymorphisms positioned within any exons of any refGene transcript

1.12.12.7 Definition - introns

A list of polymorphisms positioned within any intron of any refGene transcript

1.12.12.8 Definition - utr3

A list of polymorphisms positioned within any 3 prime untranslated region of a refGene transcript.

1.12.12.9 Definition - all

A list of all polymorphisms in the above listed regions

1.12.13 Example 22 - Genomic Regions

This example identifies the first five Delta-MATCH hits for the transcription factor matrix V\$NFKB_Q6, where the polymorphism is located with any a region 10,000 bases upstream of any refGene transcript, or located in a region of high conservation.

Figure 99 Input Parameters for Example 22

STEP 1 - (1) Single Transcription Factor Matrix Name = "V\$NFKB_Q6 (950)"

STEP 2 - Minimum Potential Score (unchecked)

STEP 2 - Top Most Significant Hits (checked) = "5"

STEP 2 - Matrix Quality (checked) = "high"

STEP 2 - Genomic Regions (checked) = "up10k; phastconsElements17way"; or

1 - Single Transcription Factor Matrix Name

Select a 'Single Transcription Factor Matrix Name' (n=550)

V\$NFKB_Q6 (950) (mat_id)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix

5 Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (qual) [REF]

high ('high = 1', 'low = 0')

Genomic Regions

[Back to top](#)

Limit results to include rsnumbers positioned in these genomic regions of refSeq genes [REF]

- up10k (647,311)
- phastconsElements17way (397,802)
- utr5 (16,376)
- coding (113,832)
- down10k (648,916)
- exons (212,764)
- introns (3,415,853)
- utr3 (84,503)

or and ("and" IS VERY SLOW!)

Figure 100 Example 22a (button set to 'or')

hit	rsnumber	chrom	chromStart	factor	mat_id	potential	threshold	m1	m2	m_per	rank	up10k	phastconsElements17way	p1_window	pubmed	hit
1	rs8030978	chr15	64651526	NF-kappaB	V\$NFKB_Q6	1	0.955	0.8377	1.0000	16.23	2	yes	-	chr15:64651519-64651532	rs8030978	1
2	rs4742069	chr9	5117460	NF-kappaB	V\$NFKB_Q6	0.9978	0.955	0.8895	0.9999	11.0411	12	-	yes	chr9:5117452-5117465	rs4742069	2
3	rs2773469	chr10	115788884	NF-kappaB	V\$NFKB_Q6	0.9978	0.955	0.8895	0.9999	11.0411	10	yes	-	chr10:115788883-115788896	rs2773469	3
4	rs539846	chr15	38185227	NF-kappaB	V\$NFKB_Q6	0.9978	0.955	0.9999	0.8895	11.0411	11	-	yes	chr15:38185225-38185238	rs539846	4
5	rs12095513	chr1	24746446	NF-kappaB	V\$NFKB_Q6	0.9978	0.955	0.9999	0.9051	9.4809	8	yes	-	chr1:24746445-24746458	rs12095513	5

Figure 101 Example 22b (button set to 'and')

hit	rsnumber	chrom	chromStart	factor	mat_id	potential	threshold	m1	m2	m_per	rank	up10k	phastconsElements17way	p1_window	pubmed	hit
1	rs12449	chr19	55126008	NF-kappaB	V\$NFKB_Q6	0.4267	0.955	0.9742	0.8638	11.3324	303	yes	yes	chr19:55126000-55126013	rs12449	1
2	rs5986945	chrX	152891433	NF-kappaB	V\$NFKB_Q6	0.2644	0.955	0.8465	0.9669	12.4522	577	yes	yes	chrX:152891432-152891445	rs5986945	2
3	rs2549601	chr16	68765338	NF-kappaB	V\$NFKB_Q6	0.2644	0.955	0.8722	0.9669	9.7942	576	yes	yes	chr16:68765328-68765341	rs2549601	3
4	rs7787483	chr7	15701171	NF-kappaB	V\$NFKB_Q6	0.26	0.955	0.8719	0.9667	9.8066	613	yes	yes	chr7:15701161-15701174	rs7787483	4
5	rs4266315	chr4	123082254	NF-kappaB	V\$NFKB_Q6	0.0622	0.955	0.8373	0.9578	12.5809	693	yes	yes	chr4:123082252-123082265	rs4266315	5

1.12.14 Bonferonni Correction

When the “**Bonferonni Correction**” checkbox is checked, the results are limited to those rsnumbers that have a “Bonferonni-adjusted rareness” value (**bonferonni**) less than or equal to the user-defined minimum. The Bonferonni-adjusted rareness is calculated by multiplying the **rareness** of the rsnumber by the total number all rsnumbers (hits) that passed the cumulative selection criteria.

Equation 20 - Bonferonni-adjusted rareness (bonferonni)

$$\text{bonferonni} = \text{rareness} * (\text{total number of results returned prior to adjustment})$$

For example, if the “V\$NFKB_Q6” mat_id is selected as a “Single Matrix” search, and only the “Matrix Details” and the “Bonferonni Correction” boxes are checked with the “Minimum Bonferonni-adjusted rareness” set to “0.005” (the default), only those 7 of the 950 “V\$NFKB_Q6” results are returned. This Bonferonni Correction may be useful when examining large genotyping dataset to determine which of a large list of polymorphisms

may have a very strong, although rare, potential to alter transcription factor binding after adjusting for multiple testing. Note that the “Matrix Details” box must also be checked to see the Bonferonni-adjusted frequency (**Bonferonni**) in the results table (Figure 103 page 191).

1.12.15 Example 23 - Bonferonni

This example will return 7 hits for the V\$NFKB_Q6 matrix where the Bonferonni-adjusted rareness is less than or equal to 0.005. The output includes the columns specific to the Matrix Details checkbox, plus the addition of a bonferonni column (**bonferonni**).

Figure 102 Input Parameters for Example 23

STEP 1 - (1) Single Transcription Factor Matrix Name = "V\$NFKB_Q6 (950)"

STEP 2 - Minimum Potential Score (unchecked)

STEP 2 - Top Most Significant Hits (unchecked)

STEP 2 - Matrix Quality (checked) = "high"

STEP 2 - Show the Matrix Details (checked)

STEP 2 - Bonferonni Correction (checked) = "0.005"

1 - Single Transcription Factor Matrix Name

Select a Single Transcription Factor Matrix Name (n=550)
V\$NFKB_Q6 (950) (mat_id)

[Back to top](#)

Minimum Potential Score

Select a 'Minimum Potential Score' (potential)
0.80 (0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix
5 Limit the number of polymorphisms seached per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (qual) [REF]
high ('high = 1', 'low = 0')

Show the Matrix Details

Show Matrix Details

(count_le_potential, mat_count, frequency, factor, factor_description, qual, mat_len)

Minimum Matrix Length

Limit searches to those matrixes with minimum length (mat_len)

12 (mat_len >= x)

Bonferonni Correction

Limit Results by 'Minimum Bonferonni-Adjusted rareness' (bonferonni)

[bonferonni = rareness*(number of returned hits)]

0.005 (bonferonni <= x)

NOTE - must have 'Matrix Details' checked to see this column

Figure 103 Bonferonni - Adjusted Rareness (bonferonni)

hit	rsnumber	chrom	chromStart	factor	mat_id	potential	threshold	m1	m2	m_per	rank	factor_description	count_ge_potential	mat_count	rareness	bonferonni	qual	mat_len	p1_window	pubmed	hit
1	rs3093317	chr16	27351578	NF-kappaB	V\$NFkB_G6	1	0.955	0.8377	1.0000	16.23	7	NF-kappaB	7	950	1.5392e-6	1.4622e-3	1	14	chr16:27351571-27351584	rs3093317	1
2	rs8030978	chr15	64651526	NF-kappaB	V\$NFkB_G6	1	0.955	0.8377	1.0000	16.23	2	NF-kappaB	7	950	1.5392e-6	1.4622e-3	1	14	chr15:64651519-64651532	rs8030978	2
3	rs1775044	chr1	7418248	NF-kappaB	V\$NFkB_G6	1	0.955	1.9000	0.8795	12.05	4	NF-kappaB	7	950	1.5392e-6	1.4622e-3	1	14	chr1:7418246-7418259	rs1775044	3
4	rs7296179	chr12	100126053	NF-kappaB	V\$NFkB_G6	1	0.955	0.8795	1.0000	12.05	6	NF-kappaB	7	950	1.5392e-6	1.4622e-3	1	14	chr12:100126043-100126056	rs7296179	4
5	rs6031444	chr20	42249151	NF-kappaB	V\$NFkB_G6	1	0.955	1.0000	0.8895	11.05	1	NF-kappaB	7	950	1.5392e-6	1.4622e-3	1	14	chr20:42249149-42249162	rs6031444	5
6	rs12090552	chr1	146194575	NF-kappaB	V\$NFkB_G6	1	0.955	1.9000	0.9052	9.48	3	NF-kappaB	7	950	1.5392e-6	1.4622e-3	1	14	chr1:146194565-146194578	rs12090552	6
7	rs2283412	chr14	71728984	NF-kappaB	V\$NFkB_G6	1	0.955	0.9467	1.0000	5.33	5	NF-kappaB	7	950	1.5392e-6	1.4622e-3	1	14	chr14:71728980-71728993	rs2283412	7

1.12.16 Minimum Number of Delta-MATCH Hits

When the “Minimum Number of Delta-MATCH Hits” checkbox is checked results may be limited to include those polymorphisms that have minimum total number of hits in the Delta-MATCH database. Theoretically, since there are 550 matrixes in the database, a single rsnumber could have up to 550 total numbers of hits. When the “Min Total Number of Delta-MATCH Hits” box is checked an additional column of results is returned showing how many hits each remaining rsnumber has in the database.

1.12.16.1 Definition - number_hits

This is the total number of hits in the Delta-MATCH database for this rsnumber.

Theoretical max number possible for this variable is 550 (one hit per matrix).

1.12.17 Example 24 - Minimum Total Number of Delta-MATCH Hits

In this example all hits for rs5743836 and rs6031444 and sorted by “chrom asc, position asc”. Polymorphism rs6031444 returns 8 hits, and rs5743836 returns 4 hits (Figure 24A).

When this same query is rerun with the minimum number hits box checked and set to 5, only the 8 results for rs6031444 are returned because the total number of hits for rs5743836 in the Delta-MATCH database is only 4, and is less than the selected minimum (Figure 106 Example 24B Results Table, page 194). Rerunning this query using the log file will cause Error 3 (Figure 207 page 410).

Figure 104 Input Parameters for Example 24

STEP 1 - (5) All Transcription Factor Matrix Names

STEP 2 - Minimum Potential Score (unchecked)

STEP 2 - Top Most Significant Hits (unchecked)

STEP 2 - Matrix Quality (unchecked)

STEP 2 - Sorted Results Table (checked) = “chrom asc, position asc (a)”

STEP 2 - Search By rsnumbers (checked) = “rs5743836, rs6031444”

(example 24B adds the following parameter)

STEP 2 - Minimum Total Number of Delta-MATCH Hits (checked) = “5”

5 - All Transcription Factor Matrix Names

Select all matrix names in the database (n=550)

[Back to top](#)

Minimum Potential Score

Select a 'Minimum Potential Score' (**potential**)

(0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix

Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (**qual**) [[REF](#)]

('high = 1', 'low = 0')

Sort Results Table

Sort the results table by

(asc = ascending, desc = descending)

[Back to top](#)

Search By rsnumbers

Either limit results by a comma-separated list of dbSNP rsnumbers (1024 chars max)

[\[REF1, REF2\]](#)

'rsnumbers'

Or upload list of rsnumbers in a plain text file

'rsnumber filename' [download example file](#)

(one rsnumber per row, 10,000 rsnumbers max)

NOTE - The 'rsnumber' text field will be ignored if an 'rsnumber filename' is uploaded

Search additional bases upstream and downstream of specified rsnumbers

'rsnumber Window'

Include other rsnumbers within this many bases

Minimum Total Number of Hits

Limit results to rsnumbers with a minimum 'total number of hits'

This is the sum number of hits for an rsnumber in the database

(**number_hits** >= x)

Figure 105 Example 24A Results Table

hit	rsnumber	chrom	chromStart	factor	mat_id	potential	threshold	m1	m2	m_per	rank	p1_window	pubmed	hit
1	rs6031444	chr20	42249151	c-Rel	V\$CREL_Q1	0.5875	0.976	0.9901	0.8516	13.9885	657	chr20:42249151-42249160	rs6031444	1
2	rs6031444	chr20	42249151	HMG1Y	V\$HMG1Y_Q3	0.0698	0.914	0.9200	0.9140	0.6522	81098	chr20:42249148-42249162	rs6031444	2
3	rs6031444	chr20	42249151	HMG	V\$HMG1Y_Q6	0	1.000	1.0000	0.9672	3.28	2655	chr20:42249151-42249157	rs6031444	3
4	rs6031444	chr20	42249151	MZF1	V\$MZF1_Q2	0.562	0.908	0.9597	0.8530	11.1181	3125	chr20:42249143-42249155	rs6031444	4
5	rs6031444	chr20	42249151	NF-kappaB	V\$NFKAPPAB65_Q1	1	0.991	1.0000	0.8523	14.77	9	chr20:42249151-42249160	rs6031444	5
6	rs6031444	chr20	42249151	NF-kappaB	V\$NFKAPPAB_Q1	1	0.984	1.0000	0.8975	10.25	15	chr20:42249151-42249160	rs6031444	6
7	rs6031444	chr20	42249151	NF-kappaB	V\$NFKB_Q6	1	0.955	1.0000	0.8895	11.05	1	chr20:42249149-42249162	rs6031444	7
8	rs6031444	chr20	42249151	NF-kappaB	V\$NFKB_Q6_Q1	0.9492	0.876	0.9985	0.8808	11.7877	117	chr20:42249146-42249161	rs6031444	8
9	rs5743836	chr3	52235821	c-Rel	V\$CREL_Q1	0.5333	0.976	0.8667	0.9888	12.3483	1406	chr3:52235821-52235830	rs5743836	9
10	rs5743836	chr3	52235821	RFX1	V\$EFC_Q6	0.0663	0.792	0.6930	0.8058	13.9985	27504	chr3:52235814-52235827	rs5743836	10
11	rs5743836	chr3	52235821	NF-kappaB	V\$NFKB_Q6	0.4133	0.955	0.8531	0.9736	12.3767	331	chr3:52235819-52235832	rs5743836	11
12	rs5743836	chr3	52235821	NF-kappaB	V\$NFKB_Q6_Q1	0.9339	0.876	0.8781	0.9939	11.6511	155	chr3:52235816-52235831	rs5743836	12

Figure 106 Example 24B Results Table

hit	rsnumber	chrom	chromStart	number_hits	factor	mat_id	potential	threshold	m1	m2	m_per	rank	p1_window	pubmed	hit
1	rs6031444	chr20	42249151	8	c-Rel	V\$CREL_Q1	0.5875	0.976	0.9901	0.8516	13.9885	657	chr20:42249151-42249160	rs6031444	1
2	rs6031444	chr20	42249151	8	HMG1Y	V\$HMG1Y_Q3	0.0698	0.914	0.9200	0.9140	0.6522	81098	chr20:42249148-42249162	rs6031444	2
3	rs6031444	chr20	42249151	8	HMG	V\$HMG1Y_Q6	0	1.000	1.0000	0.9672	3.28	2655	chr20:42249151-42249157	rs6031444	3
4	rs6031444	chr20	42249151	8	MZF1	V\$MZF1_Q2	0.562	0.908	0.9597	0.8530	11.1181	3125	chr20:42249143-42249155	rs6031444	4
5	rs6031444	chr20	42249151	8	NF-kappaB	V\$NFKAPPAB65_Q1	1	0.991	1.0000	0.8523	14.77	9	chr20:42249151-42249160	rs6031444	5
6	rs6031444	chr20	42249151	8	NF-kappaB	V\$NFKAPPAB_Q1	1	0.984	1.0000	0.8975	10.25	15	chr20:42249151-42249160	rs6031444	6
7	rs6031444	chr20	42249151	8	NF-kappaB	V\$NFKB_Q6	1	0.955	1.0000	0.8895	11.05	1	chr20:42249149-42249162	rs6031444	7
8	rs6031444	chr20	42249151	8	NF-kappaB	V\$NFKB_Q6_Q1	0.9492	0.876	0.9985	0.8808	11.7877	117	chr20:42249146-42249161	rs6031444	8

1.12.18 Hugo Names

When the “**Hugo Names**” checkbox is checked, each rsnumber is searched to see if it is associated (is located within or next to) a known transcript. If that transcript is annotated in the HUGO database, the HUGO name will be shown in the results table. It is sometimes convenient to **view only** those rsnumbers that associate with annotated HUGO names when perusing a results table looking for interesting candidate genes. Therefore, it is possible to limit the results table to only include those rsnumbers with a that are located within or near to an annotated HUGO name, by checking the internal check box (“Limit results to rsnumbers next to known HUGO_GENES”).

1.12.18.1 Definition - hugo_name

This is the gene name abbreviation in the HUGO database (UCSC genome browser hg18.refGene.name2) that is associated with this corresponding rsnumber. A single rsnumber may associate with multiple HUGO names and if it is, they will be numerically listed.

A complete list of the associations between the SNP to HUGO names can be downloaded in the file “SNP-Genes_HUGO.txt” (48.4 Mb) from the website (**Delta-MATCH > Downloads**). Note that 2,890,665 rsnumber to HUGO name association exist.

1.12.19 Example 25 - HUGO Names

In this example, three rsnumbers are search against a single matrix. Because the Hugo Names checkbox is checked, an additional column (hugo_name) is found in the results table. Note that rsnumber rs2305917 is associated with four HUGO names (TBC1D17, AKT1S1, IL4I1, and NUP62), rs6031444 is associated with one HUGO name (JPH2), and rs8030978 is not associated with a HUGO name. If this example had been re-run with the internal (“Limit results to rsnumbers next to known HUGO genes”) box checked, only the results for rs2305917 and rs6031444 would have been returned. The hugo_name hyperlinks to the UCSC browser entry for the corresponding HUGO name. Rerunning this query using the log file will cause Error 3 (Figure 207 page 410).

Figure 107 Input Parameters for Example 25

STEP 1 - (1) Single Transcription Factor Matrix Name = “V\$NFkB_Q6 (950)”

STEP 2 - Minimum Potential Score (**unchecked**)

STEP 2 - Top Most Significant Hits (**unchecked**)

STEP 2 - Matrix Quality (checked) = "high"

STEP 2 - Sorted Results Table (**checked**) = "chrom asc, position asc (a)"

STEP 2 - Search By rsnumbers (**checked**) = "rs6031444, rs8030978, rs2305917"

STEP 2 - Hugo Names (**checked**); Limit results to rsnumbers next to known

HUGO_GENES (unchecked)

1 - Single Transcription Factor Matrix Name

Select a 'Single Transcription Factor Matrix Name' (n=550)

V\$NFKB_Q6 (950) (mat_id)

[Back to top](#)

Minimum Potential Score

Select a 'Minimum Potential Score' (**potential**)

0.80 (0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix

5 Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (**qual**) [REF]

high ('high = 1', 'low = 0')

Sort Results Table

Sort the results table by

chrom asc, position asc (a)

(asc = ascending, desc = descending)

[Back to top](#)

Search By rsnumbers

Either limit results by a comma-separated list of dbSNP rsnumbers (1024 chars max)

[\[REF1, REF2\]](#)

'rsnumbers'

Or upload list of rsnumbers in a plain text file

'rsnumber filename' [download example file](#)

(one rsnumber per row, 10,000 rsnumbers max)

NOTE - The 'rsnumber' text field will be ignored if an 'rsnumber filename' is uploaded

Search additional bases upstream and downstream of specified rsnumbers

'rsnumber Window'

Include other rsnumbers within this many bases

HUGO Names

Show the HUGO names of the genes associated with each rsnumber [\[REF1, REF2\]](#)

([hugo_name](#))

Limit results to rsnumbers next to known HUGO genes

Download the rsnumber to hugo name file

(WARNING 48.4 Mb, right-click and 'download file') [SNP-Genes_HUGO.txt](#)

Figure 108 Example 25 Results Table

hit	rsnumber	chrom	chromStart	factor	mat_id	potential	threshold	m1	m2	m_per	rank	hugo_name	p1_window	pubmed	hit
1	rs8030978	chr15	64651526	NF-kappaB	V\$NFKB_Q6	1	0.955	0.8377	1.0000	16.23	2		chr15:64651519-64651532	rs8030978	1
2	rs2305917	chr19	55083110	NF-kappaB	V\$NFKB_Q6	0.7178	0.955	0.9676	0.9999	3.2303	68	(1)TBC1D17 (2)AKT1S1 (3)L411 (4)NUP62	chr19:55083104-55083117	rs2305917	2
3	rs6031444	chr20	42249151	NF-kappaB	V\$NFKB_Q6	1	0.955	1.0000	0.8895	11.05	1	(1)JPH2	chr20:42249149-42249162	rs6031444	3

1.12.20 Reflink

When the “**Reflink**” checkbox is checked, 7 additional columns are presented in the results table that describe other information about the HUGO genes associating with each rsnumber.

1.12.20.1 Definition - reflink_product

This field further describes the HUGO name. (UCSC genome browser track hg18.refLink.product)

1.12.20.2 Definition - reflink_mrnaAcc

This field links to the NCBI Entrez [Nucleotide](#) entries and may include non-human entries. (UCSC genome browser track hg18.refLink.mrnaAcc)

1.12.20.3 Definition - reflink_protAcc

This field links to the NCBI Entrez [Protein](#) entries. (UCSC genome browser track hg18.refLink.protAcc)

1.12.20.4 Definition - reflink_name

This field links to the [NCBI](#) Entrez [Gene](#) entries. These reflink_names may be duplicated. (UCSC genome browser track hg18.refLink.name)

1.12.20.5 Definition - reflink_prodName

(UCSC genome browser track hg18.refLink.prodName)

1.12.20.6 Definition - reflink_locusLinkId

This field links to the NCBI Entrez [Gene](#) entries for homologous genes in other non-human organisms. These reflink_locusLinkId may be duplicated. (UCSC genome browser track hg18.refLink.locusLinkId)

1.12.20.7 Definition - reflink_omimId

This field links to the NCBI Entrez [OMIM](#) (Online Mendelian Inheritance in Man) database. These reflink_omimId may be duplicated. (UCSC genome browser track hg18.refLink.omimId)

It is additionally possible to limit results to only include rsnumbers associated with particular text names by checking the additional internal “Limit results with text matching the hg18.reflink_product” box. If the submitted text (1,024 characters maximum) **is not** an exact match for a term in the associated product (hg18.refLink.product), the rsnumber will be excluded, and those rsnumbers **with** and exact match will be returned in the results table.

1.12.21 Example 26 - Reflink

This example will search for the top 5 hits for the V\$NFKB_Q6 matrix that are associated with hugo name gene that has the word “kinase” in its HUGO name annotation (hg18.reflink_product), and also having a potential greater than or equal to 0.30.

Figure 109 Input Parameters for Example 26

STEP 1 - (1) Single Transcription Factor Matrix Name = “V\$NFKB_Q6 (950)”

STEP 2 - Minimum Potential Score (checked) = “**0.30**”

STEP 2 - Top Most Significant Hits (checked) = “5”

STEP 2 - Matrix Quality (checked) = “high”

STEP 2 - Hugo Names (**checked**); Limit results to rsnumbers next to known HUGO-GENES (unchecked)

STEP 2 - Reflink (**checked**); Limit results with text matching the hg18.reflink_product (**checked**) = “kinase”

1 - Single Transcription Factor Matrix Name

Select a 'Single Transcription Factor Matrix Name' (n=550)

V\$NFKB_Q6 (950) (mat_id)

Minimum Potential Score

[Back to top](#)

Select a 'Minimum Potential Score' (potential)

0.30 (0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix

5 Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (qual) [REF]

high ('high = 1', 'low = 0')

HUGO Names

Show the HUGO names of the genes associated with each rsnumber [REF1, REF2]
(hugo_name)

Limit results to rsnumbers next to known HUGO genes

Download the rsnumber to hugo name file

(WARNING 48.4 Mb, right-click and 'download file') [SNP-Genes_HUGO.txt](#)

Reflink

[Back to top](#)

Show reflink Details [REF]

(reflink_mrnaAcc, reflink_protAcc, reflink_name, reflink_prodName, reflink_locusLinkId,
reflink_omimId)

Limit results with text matching the hg18.reflink_product

kinase

Figure 110 Example 26 Results Table

id	rsnumber	chrom	chromStart	factor	max_id	posdist	threshold	r1	r2	r_start	end	hugo_name	tx_start	ntfink_procc	ntfink_rnaAcc	ntfink_proccAcc	ntfink_name	ntfink_posname	ntfink_locuslink	ntfink_cdnlink	pubmed	hit		
1	rs174288	chr9	5117480	FE-4updown	V2F_H2_O8	0.3573	0.205	0.205	0.205	11,2411	12	13ANK2	chr9:5117420-5117480	13ANK2	13ANK2	13ANK2	13ANK2	13ANK2	13ANK2	13ANK2	13ANK2	13ANK2	13ANK2	
2	rs381888	chr10	8188200	FE-4updown	V2F_H2_O8	0.3873	0.288	0.288	0.288	11,081	88	13FKF1	chr10:8178000-8188200	13FKF1	13FKF1	13FKF1	13FKF1	13FKF1	13FKF1	13FKF1	13FKF1	13FKF1	13FKF1	
3	rs178194	chr2	2002001	FE-4updown	V2F_H2_O8	0.3711	0.205	0.2142	0.205	2,5108	74	13ALK	chr2:2002000-2002000	13ALK	13ALK	13ALK	13ALK	13ALK	13ALK	13ALK	13ALK	13ALK	13ALK	
4	rs228207	chr11	4722200	FE-4updown	V2F_H2_O8	0.3573	0.205	0.2183	0.212	10,3742	80	13AKO2	chr11:4722200-4722200	13AKO2	13AKO2	13AKO2	13AKO2	13AKO2	13AKO2	13AKO2	13AKO2	13AKO2	13AKO2	13AKO2
5	rs1007528	chr2	18334074	FE-4updown	V2F_H2_O8	0.3573	0.205	0.2470	0.212	9,2328	104	13YF30	chr2:18334000-18334075	13YF30	13YF30	13YF30	13YF30	13YF30	13YF30	13YF30	13YF30	13YF30	13YF30	13YF30

1.12.22 Distance from txStart or cdStart

When the “Distance from txStart of cdStart” checkbox is checked, it is possible to view how far each rsnumber is from the transcriptional and coding start site for each of its associated hugo names. When checked, three additional columns appear in the results page. Multiple values for these variables are enumerated.

1.12.22.1 Definition - dist_from_ref (distance from reference)

This is the distance in base pairs that this rsnumber is from the input reference base position. By default the input reference base position equals 1 (so by default; dist_from_ref = chromStart - 1).

1.12.22.2 Definition - dist_from_tx (distance from transcription start site)

This is the distance in base pairs that the rsnumber is from the transcriptional site start for any associated HUGO name transcript.

1.12.22.3 Definition - dist_from_cds (distance from coding start site)

This is the distance in base pairs that the rsnumber is from the coding site start for any associated HUGO name transcript.

When the internal sub-checkboxes are checked, results will be restricted to only those rsnumbers that are positioned within the specified distances from an associated HUGO name transcriptional and/or coding start site. For those rsnumbers with multiple transcripts associated with it, an rsnumber must be positioned within the specified number of bases **for at least one** start site to pass the criteria.

1.12.23 Example 27 - Distance From txStart or cdStart

This example will try to return up to five hits for the V\$NFKB_Q6 matrix that have potential scores greater than or equal to 0.8 for rsnumbers located within 2000 base pairs of both a known transcriptional, and a coding start site. Only 3 rsnumbers are returned. These rsnumbers associate with the HUGO names JPH2, BMF, and EIF4G2.

Figure 111 Input Parameters for Example 27

STEP 1 - (1) Single Transcription Factor Matrix Name = "V\$NFKB_Q6 (950)"

STEP 2 - Minimum Potential Score (checked) = "0.80"

STEP 2 - Top Most Significant Hits (checked) = "5"

STEP 2 - Matrix Quality (checked) = "high"

STEP 2 - Hugo Names (**checked**); Limit results to rsnumbers next to known HUGO-GENES (unchecked)

STEP 2 - Distance From txStart or cdStart (**checked**); = ("1", "2000", "2000")

1 - Single Transcription Factor Matrix Name

Select a 'Single Transcription Factor Matrix Name' (n=550)

V\$NFKB_Q6 (950) (mat_id)

Minimum Potential Score

[Back to top](#)

Select a 'Minimum Potential Score' (potential)

0.80 (0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix

5 Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (qual) [REF]

high ('high = 1', 'low = 0')

HUGO Names

Show the HUGO names of the genes associated with each rsnumber [REF1, REF2]
(hugo_name)

Limit results to rsnumbers next to known HUGO genes

Download the rsnumber to hugo name file

(WARNING 48.4 Mb, right-click and 'download file') [SNP-Genes_HUGO.txt](#)

Distance From txStart or cdStart

Show the distance details [REF]

(dist_from_ref, dist_from_tx, dist_from_cds)

1 Include this many bases upstream/downstream of selected genes
(dist_from_ref)

2000 Absolute minimum distance from any 'Transcriptional' start
(dist_from_tx)

2000 Absolute minimum distance from any 'Translational' start
(dist_from_cds)

Figure 112 Example 27 Results Table

hit	rsnumber	chrom	chromStart	factor	mat_id	potential	threshold	m1	m2	m_per	rank	hugo_name	dist_from_ref	dist_from_tx	dist_from_cds	p1_window	pubmed	hit
1	rs6031444	chr20	42249151	NF-kappaB	VSNFKB_Q6	1	0.955	1.0000	0.8895	11.05	1	(1)JPH2	42249150	(1)481 (2)481	(1)-392 (2)-392	chr20:42249149-42249162	rs6031444	1
2	rs539846	chr15	38185227	NF-kappaB	VSNFKB_Q6	0.9978	0.955	0.9999	0.8895	11.0411	11	(1)BMF	38185226	(1)352 (2)352 (3)704 (4)3140	(1)352 (2)352 (3)352 (4)352	chr15:38185225-38185238	rs539846	2
3	rs16908327	chr11	10785681	NF-kappaB	VSNFKB_Q6	0.9244	0.955	0.9966	0.8343	16.2854	60	(1)EIF4G2	10785680	(1)1477 (2)1477	(1)-263 (2)-263	chr11:10785672-10785685	rs16908327	3

1.12.24 Gene Ontology

When the “**Gene Ontology**” checkbox is checked, two additional columns appear in the results table that detail any gene ontology names and accession numbers associated with a given HUGO name. If the internal checkbox (“Limit to text matching a ‘Gene Ontology’ term”) is checked, only those rsnumbers with go_names matching the submitted text will be returned. It is useful to have the Hugo Names checkbox checked when examining Gene Ontology results. Hyperlinks go to the associated AmiGO database entry (<http://amigo.geneontology.org>).

1.12.24.1 Definition - go_names (gene ontology names)

These text descriptions of all of the gene ontology names associated with the corresponding HUGO name.

1.12.24.2 Definition - go_number (gene ontology number)

These are accession numbers for the gene ontology names associated with the corresponding HUGO name.

1.12.25 Example 28 - Gene Ontology

This example returns all results for VSNFKB_Q6, with a potential greater than or equal to 0.8 where the HUGO name for the rsnumber has a gene ontology term matching the text “transcription”. Results are found for three HUGO names (NR3C1, RXRG, and TEAD1).

Figure 113 Input Parameters for Example 28

STEP 1 - (1) Single Transcription Factor Matrix Name = “V\$NFKB_Q6 (950)”

STEP 2 - Minimum Potential Score (checked) = “0.80”

STEP 2 - Top Most Significant Hits (checked) = “5”

STEP 2 - Matrix Quality (checked) = “high”

STEP 2 - Hugo Names (**checked**); Limit results to rsnumbers next to known HUGO-GENES (unchecked)

STEP 2 - Gene Ontology (**checked**); Limit to text matching a Gene Ontology term (checked) = “**transcription**”

1 - Single Transcription Factor Matrix Name

Select a 'Single Transcription Factor Matrix Name' (n=550)
 (mat_id)

Minimum Potential Score

[Back to top](#)

Select a 'Minimum Potential Score' (potential)
 (0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix
 Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (qual) [REF]
 ('high = 1', 'low = 0')

Gene Ontology

[Back to top](#)

Show gene ontology details [\[REF\]](#)
 (go_names, go_number)

Limit to text matching a 'Gene Ontology' term (go_names)

transcription

Download the rsnumber to HUGO name file
 (WARNING 352 Mb, right-click and 'download file') [SNP-Genes_GO.txt](#)

Figure 114 Example 28 Results Table

hit	rsnumber	chrom	chromStart	factor	mac_id	potential	threshold	m1	m2	m_per	rank	hugo_name	go_names	go_number	pt_window	pubmed	hit
1	rs1002282	chr5	14281442	FF-4aggas	V3NF_HB_Q8	0.2273	0.265	0.2222	0.2278	18.2218	93	11K65C1	[1] transcriptional motif [2] transcription from RNA polymerase II promoter [3] transcription factor activity [4] transcriptional response [5] cis-act binding [5] signal transduction [7] stem binding [8] glucose-6-phos phosphate activity [9] nucleus [10] regulation of transcription, DNA-dependent [11] methion binding [12] cytoplasm	[1] 0002750 [2] 0002800 [3] 0002700 [4] 0002294 [5] 0002428 [6] 0002718 [7] 0002515 [8] 0002428 [9] 0002894 [10] 0002525 [11] 0002422 [12] 0002717	chr5:14281441-14281443	rs1002282	1
2	rs1003323	chr1	18287313	FF-4aggas	V3NF_HB_Q8	0.2273	0.265	0.2222	0.21794	12.2212	91	11K2492	[1] regulation of transcription, DNA-dependent [2] cis-act binding [3] nucleus [4] methion binding [5] cis-act histone receptor activity [6] histone-H receptor activity [7] transcription factor activity [8] transcription	[1] 0002525 [2] 0002428 [3] 0002294 [4] 0002422 [5] 0002717 [6] 0002428 [7] 0002718 [8] 0002525	chr1:18287311-18287329	rs1003323	2
3	rs1922294	chr11	12833223	FF-4aggas	V3NF_HB_Q8	0.2273	0.265	0.2222	0.2212	15.2148	98	11Y7CAD1	[1] regulation of transcription, DNA-dependent [2] transcription factor activity [3] transcription binding [4] transcription [5] nucleus	[1] 0002525 [2] 0002700 [3] 0002515 [4] 0002525 [5] 0002294	chr11:12833212-12833225	rs1922294	3

1.12.26 Affymetrix

When the Affymetrix checkbox is checked, it is possible to restrict results to those rsnumber that are present on an Affymetrix genotyping chip (SNPchip). It is possible to select from a number of chip platforms including the10k, 100k and 500k chips:

- 500k_all (492,555)
- 250k_nsp (257,877)
- 250k_sty (234,678)
- 10k_all (11,383)
- 10k_xba131 (10,009)
- 10k_xba142 (11,316)
- 100k_all (115,117)
- 50k_hind240 (56,726)
- 50k_xba240 (58,391)
- all_affy_snps< (583,396)

The number of Delta-MATCH SNPs on each platform is listed on each SNPchip is listed in within the parentheses.

1.12.27 Example 29 - Affymetrix

This example returns all results for V\$NFKB_Q6, with a potential greater than or equal to 0.8 where the rsnumber is present on the Affymetrix 500k SNPchip. Three rsnumbers are returned (rs3093317, rs6481864, rs6036746)

Figure 115 Input Parameters for Example 29

STEP 1 - (1) Single Transcription Factor Matrix Name = "V\$NFKB_Q6 (950)"

STEP 2 - Minimum Potential Score (checked) = "0.80"

STEP 2 - Top Most Significant Hits (unchecked)

STEP 2 - Matrix Quality (checked) = "high"

STEP 2 - Affymetrix (checked) = "500k_all (492,555)

1 - Single Transcription Factor Matrix Name

Select a 'Single Transcription Factor Matrix Name' (n=550)
 (mat_id)

Minimum Potential Score

[Back to top](#)

Select a 'Minimum Potential Score' (potential)
 (0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix
 Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (qual) [REF]
 ('high = 1', 'low = 0')

Affymetrix

[Back to top](#)

Limit results to rsnumbers on an Affymetrix SNP-CHIP [REF1, REF2]

Include SNPs in strong LD with those on the Affymetrix 500k SNP-chip (name_affy)

(LD = linkage disequilibrium)

1.12.28 Using the HapMap Database to Find Other rsnumbers in Strong Linkage Disequilibrium with Polymorphisms on an Affymetrix SNPchip

Sometimes a high-throughput genotyping survey will identify a number of polymorphisms that associate with a particular phenotype (variable mRNA expression levels), but the markers themselves can't be linked back to a biological affect using traditional

techniques of molecular biology (promoter reporter assays). In these cases, it may be useful to look for other polymorphisms in strong “linkage disequilibrium” with the associated ones. It is possible to use the HapMap database to do this.

When the “Include SNPs in strong LD with those on the Affymetrix 500k SNP-chip” checkbox is checked, it is possible to search for results that are not present on the specified Affymetrix SNPchip, but are in strong “linkage disequilibrium” with those that are. In this way it is possible to identify markers that are not on the common genotyping platforms that may have a biological affect (an ability to create allele-specific transcription factor binding site).

1.12.28.1 Definition - linkage disequilibrium (LD)

This is a term that describes how much recombination has occurred between two markers positioned on the same chromosome. A pair of markers will have high linkage disequilibrium and be inherited together when the distance between them is small, when the time between the origination of the first and second marker was small, and when the time of the origination of the second marker is recent

(http://www.hapmap.org/gbrowse_help.html).

1.12.28.2 Definition - rsquare (r^2 linkage disequilibrium value)

This is a statistic that reflects the degree of linkage disequilibrium between two genetic markers [32].

1.12.28.3 Definition - dprime (D' linkage disequilibrium value)

This is a statistic that reflects the degree of linkage disequilibrium between two genetic markers [33].

Thus far 12 pre-tabulated lists of polymorphisms have been created using three minimum rsquare cutoff values (1.0, 0.9, or 0.8) paired with one of four separate ethnic populations (CEU, YRI, JPT, and CHB).

- CEU = Caucasian / European
- YRI = Yoruba / African
- JPT = Japanese
- CHB = Chinese

Figure 116 Population / Linkage Disequilibrium rsquare Pairs for the Affymetrix 500k SNPchip

This is a list of pre-tabulated LD lists. The number of polymorphisms found in each is found inside the parentheses. Notice that there are higher numbers of polymorphisms on the lists with lower rsquare cutoff values. Also notice that the African population has the lowest amount of linkage disequilibrium (smallest lists) when compared to the other HapMap populations. In effect, by checking the additional checkbox under the Affymetrix section it is possible to search the Delta-MATCH database for more than 2.8-fold the number of polymorphisms on the original 500k SNPchip (maximum fold increase is for CEU @ 0.8 = $1,396,609 / 492,555 = 2.84$)

```

✓ CEU rsquare >= 1.0 (1,058,667)
  CEU rsquare >= 0.9 (1,240,120)
  CEU rsquare >= 0.8 (1,396,609)
  YRI rsquare >= 1.0 (786,352)
  YRI rsquare >= 0.9 (898,884)
  YRI rsquare >= 0.8 (1,022,791)
  CHB rsquare >= 1.0 (1,084,004)
  CHB rsquare >= 0.9 (1,233,628)
  CHB rsquare >= 0.8 (1,375,518)
  JPT rsquare >= 1.0 (1,094,046)
  JPT rsquare >= 0.9 (1,233,192)
  JPT rsquare >= 0.8 (1,370,540)

```

1.12.29 Example 30 - Affymetrix with Linkage Disequilibrium

This example returns all results for V\$NFKB_Q6, with a potential greater than or equal to 0.8 for any rsnumber present on the Affymetrix 500k SNPchip, and for any rsnumber in strong linkage disequilibrium with any marker present on the Affymetrix 500k SNPchip. In addition to the three rsnumbers identified in Example 29, this example returns an additional 8. Note one additional column (name_affy) is shown in the results table for this example.

1.12.29.1 Definition - name_affy

This is a description of whether or not this rsnumber is “present” or “absent” from the Affymetrix 500k SNPchip.

Figure 117 Input Parameters for Example 30

STEP 1 - (1) Single Transcription Factor Matrix Name = “V\$NFKB_Q6 (950)”

STEP 2 - Minimum Potential Score (checked) = “0.80”

STEP 2 - Top Most Significant Hits (unchecked)

STEP 2 - Matrix Quality (checked) = “high”

STEP 2 - Affymetrix (**checked**) = “500k_all (492,555)”; Include SNPs in strong LD with those on the AFFY 500k SNP-CHIP (**checked**) = “CEU rsquare >= 1.0 (1,058,667)”

1 - Single Transcription Factor Matrix Name

Select a 'Single Transcription Factor Matrix Name' (n=550)
 (mat_id)

Minimum Potential Score

[Back to top](#)

Select a 'Minimum Potential Score' (potential)
 (0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix
 Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (qual) [REF]
 ('high = 1', 'low = 0')

[Back to top](#)

Affymetrix

Limit results to rsnumbers on an Affymetrix SNP-CHIP [REF1, REF2]

Include SNPs in strong LD with those on the Affymetrix 500k SNP-chip (name_affy)

(LD = linkage disequilibrium)

Figure 118 Example 30 Results Table

hit	rsnumber	chrom	chromStart	name_affy	factor	mat_id	potential	threshold	m1	m2	m_per	rank	p1_window	pubmed	hit
1	rs3093317	chr16	27351578	PRESENT	NF-kappaB	VSNFKB_Q6	1	0.955	0.8377	1.0000	16.23	7	chr16:27351571-27351584	rs3093317	1
2	rs2283412	chr14	71728984	ABSENT	NF-kappaB	VSNFKB_Q6	1	0.955	0.9467	1.0000	5.33	5	chr14:71728980-71728993	rs2283412	2
3	rs7380665	chr5	121731207	ABSENT	NF-kappaB	VSNFKB_Q6	0.9978	0.955	0.9999	0.8376	16.2316	14	chr5:121731205-121731218	rs7380665	3
4	rs4074910	chr6	21228732	ABSENT	NF-kappaB	VSNFKB_Q6	0.9978	0.955	0.8376	0.9999	16.2316	13	chr6:21228723-21228736	rs4074910	4
5	rs6481864	chr10	33968031	PRESENT	NF-kappaB	VSNFKB_Q6	0.9978	0.955	0.9999	0.8376	16.2316	21	chr10:33968024-33968037	rs6481864	5
6	rs6036746	chr20	24284794	PRESENT	NF-kappaB	VSNFKB_Q6	0.9978	0.955	0.9999	0.9051	9.4809	22	chr20:24284784-24284797	rs6036746	6
7	rs4348296	chr6	162392452	ABSENT	NF-kappaB	VSNFKB_Q6	0.9933	0.955	0.9997	0.8892	11.0533	34	chr6:162392442-162392455	rs4348296	7
8	rs17013128	chr3	23376599	ABSENT	NF-kappaB	VSNFKB_Q6	0.9844	0.955	0.8370	0.9993	16.2414	46	chr3:23376592-23376605	rs17013128	8
9	rs6975	chr6	131197794	ABSENT	NF-kappaB	VSNFKB_Q6	0.9844	0.955	0.9993	0.8370	16.2414	49	chr6:131197792-131197805	rs6975	9
10	rs2275128	chr10	13418450	ABSENT	NF-kappaB	VSNFKB_Q6	0.9844	0.955	0.8370	0.9993	16.2414	51	chr10:13418448-13418461	rs2275128	10
11	rs11154234	chr6	124738177	ABSENT	NF-kappaB	VSNFKB_Q6	0.9533	0.955	0.9571	1.0000	4.29	53	chr6:124738169-124738182	rs11154234	11

1.12.30 Illumina

When the Illumina checkbox is checked, it is possible to restrict results to those rsnumbers that are present on an Illumina genotyping SNPchips (ILMN_HumanHap550 and ILMN_HumanHap300). The number of the polymorphisms on each SNPchip is noted in the parentheses. As with the Affymetrix box, it is possible to identify Delta-MATCH hits for polymorphisms that are not present on the Illumina 550k chip, but in strong linkage disequilibrium with those markers that are present on the 550k chip by checking the “Include SNPs in strong LD with those on the ILMN_humanHap550 SNP-CHIP” box. By checking the additional checkbox under the Illumina section it is possible to search the Delta-MATCH database for more than 2.9-fold the number of polymorphisms on the original 500k SNPchip (maximum fold increase is for CEU @ 0.8 = 1,639,617 / 555,174 = 2.95)

Figure 119 Population / Linkage Disequilibrium rsquare Pairs for the Illumina 550k

SNPchip

✓ CEU rsquare \geq 1.0 (1,185,043)
CEU rsquare \geq 0.9 (1,426,554)
CEU rsquare \geq 0.8 (1,639,617)
YRI rsquare \geq 1.0 (850,292)
YRI rsquare \geq 0.9 (986,091)
YRI rsquare \geq 0.8 (1,144,634)
CHB rsquare \geq 1.0 (1,235,505)
CHB rsquare \geq 0.9 (1,359,939)
CHB rsquare \geq 0.8 (1,599,238)
JPT rsquare \geq 1.0 (1,253,266)
JPT rsquare \geq 0.9 (1,427,252)
JPT rsquare \geq 0.8 (1,593,841)

1.12.31 Example 31 - Illumina

This example returns all results for V\$NFKB_Q6, with a potential greater than or equal to 0.8 where the rsnumber is present on the Illumina 550k SNPchip. Three rsnumbers are returned (rs3093317, rs10800098, rs6036746). (Note, only two of these three are the same as the three returned in the first Affymetrix example.)

Figure 120 Input Parameters for Example 31

STEP 1 - (1) Single Transcription Factor Matrix Name = "V\$NFKB_Q6 (950)"

STEP 2 - Minimum Potential Score (checked) = "0.80"

STEP 2 - Top Most Significant Hits (unchecked)

STEP 2 - Matrix Quality (checked) = "high"

STEP 2 - Illumina (**checked**) = "ILMN_HumanHap550_SNPlist (555,174)"; Include SNPs in strong LD with those on the ILMN_HumanHap550 SNP-CHIP (unchecked)

1 - Single Transcription Factor Matrix Name

Select a Single Transcription Factor Matrix Name (n=550)

V\$NFKB_Q6 (950) (mat_id)

1 - Single Transcription Factor Matrix Name

Select a 'Single Transcription Factor Matrix Name' (n=550)

V\$NFKB_Q6 (950) (mat_id)

Minimum Potential Score

[Back to top](#)

Select a 'Minimum Potential Score' (potential)

0.80 (0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix

5 Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (**qual**) [REF]

('high = 1', 'low = 0')

Illumina

Limit results to rsnumbers on an Illumina SNP-chip [REF]

Include SNPs in strong LD with those on the ILMN_HumanHap550 SNP-chip

(LD = linkage disequilibrium)

1.12.32 Example 32 - Illumina with Linkage Disequilibrium

This example returns all results for V\$NFKB_Q6, with a potential greater than or equal to 0.8 for any rsnumber present on the Illumina 550k SNPchip, and for any rsnumber in strong linkage disequilibrium with any marker present on the Illumina 550k SNPchip. In addition to the three rsnumbers identified in Example 31, this example returns an additional 8. Note one additional column (name_illumina) is shown in the results table for this example.

1.12.32.1 Definition - name_illumina

This is a description of whether or not this rsnumber is “present” or “absent” from the Illumina 550k SNPchip.

Figure 121 Input Parameters for Example 32

STEP 1 - (1) Single Transcription Factor Matrix Name = “V\$NFKB_Q6”

STEP 2 - Minimum Potential Score (checked) “0.8”

STEP 2 - Top Most Significant Hits (**unchecked**)

STEP 2 - Matrix Quality (checked) = "high"

STEP 2 - Illumina (**checked**) = "ILMN_HumanHap550_SNPlist (555,174)"; Include SNPs in strong LD with those on the ILMN_HumanHap550 SNP-CHIP (**checked**) = "CEU rsquare \geq 1.0 (1,185,043)"

1 - Single Transcription Factor Matrix Name

Select a 'Single Transcription Factor Matrix Name' (n=550)

V\$NFKB_Q6 (950) (mat_id)

Minimum Potential Score

[Back to top](#)

Select a 'Minimum Potential Score' (potential)

0.80 (0 \leq x \leq 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix

5 Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (qual) [REF]

high ('high = 1', 'low = 0')

Illumina

Limit results to rsnumbers on an Illumina SNP-chip [REF]

ILMN_HumanHap550_SNPlist (555,174)

Include SNPs in strong LD with those on the ILMN_HumanHap550 SNP-chip

CEU rsquare \geq 1.0 (1,185,043)

(LD = linkage disequilibrium)

1.12.33 Example 33 - Affymetrix and Illumina (all checkboxes checked)

This example returns all results for V\$NFKB_Q6, with a potential greater than or equal to 0.8 for any rsnumber present on or in strong disequilibrium with markers on the

Affymetrix 500k and Illumina 550k SNPchip. Six rsnumbers are returned. The status of whether or not each polymorphism is “present” or “absent” on each SNPchip is noted (name_affy and name_illumina).

Figure 122 Input Parameters for Example 33

STEP 1 - (1) Single Transcription Factor Matrix Name = “V\$NFKB_Q6 (950)”

STEP 2 - Minimum Potential Score (checked) = “0.80”

STEP 2 - Top Most Significant Hits (**unchecked**)

STEP 2 - Matrix Quality (checked) = “high”

STEP 2 - Affymetrix (**checked**) = “**500k_all (492,555)**”; Include SNPs in strong LD with those on the AFFY 500k SNP-CHIP (**checked**) = “CEU rsquare \geq 1.0 (1,058,667)”

STEP 2 - Illumina (**checked**) = “**ILMN_HumanHap550_SNPlist (555,174)**”; Include SNPs in strong LD with those on the ILMN_HumanHap550 SNP-CHIP (**checked**) = “**CEU rsquare \geq 1.0 (1,185,043)**”

1 - Single Transcription Factor Matrix Name

Select a 'Single Transcription Factor Matrix Name' (n=550)

V\$NFKB_Q6 (950) (mat_id)

[Back to top](#)

Minimum Potential Score

Select a 'Minimum Potential Score' (potential)

0.80 (0 \leq x \leq 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix

5 Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (**qual**) [REF]

('high = 1', 'low = 0')

[Back to top](#)

Affymetrix

Limit results to rsnumbers on an Affymetrix SNP-CHIP [REF1, REF2]

Include SNPs in strong LD with those on the Affymetrix 500k SNP-chip (**name_affy**)

(LD = linkage disequilibrium)

Illumina

Limit results to rsnumbers on an Illumina SNP-chip [REF]

Include SNPs in strong LD with those on the ILMN_HumanHap550 SNP-chip

(LD = linkage disequilibrium)

Figure 123 Example 33 Results Table

hit	rsnumber	chrom	chromStart	name_affy	name_illumina	factor	mat_id	potential	threshold	m1	m2	m_per	rank	p1_window	pubmed	hit
1	rs3093317	chr16	27351578	PRESENT	PRESENT	NF-kappaB	V\$NFKB_Q6	1	0.955	0.8377	1.0000	16.23	7	chr16:27351571-27351584	rs3093317	1
2	rs2283412	chr14	71728984	ABSENT	ABSENT	NF-kappaB	V\$NFKB_Q6	1	0.955	0.9467	1.0000	5.33	5	chr14:71728980-71728993	rs2283412	2
3	rs6036746	chr20	24284794	PRESENT	PRESENT	NF-kappaB	V\$NFKB_Q6	0.9978	0.955	0.9999	0.9051	9.4809	22	chr20:24284784-24284797	rs6036746	3
4	rs4348296	chr6	162392452	ABSENT	ABSENT	NF-kappaB	V\$NFKB_Q6	0.9933	0.955	0.9997	0.8892	11.0533	34	chr6:162392442-162392455	rs4348296	4
5	rs2275128	chr10	13418450	ABSENT	ABSENT	NF-kappaB	V\$NFKB_Q6	0.9844	0.955	0.8370	0.9993	16.2414	51	chr10:13418448-13418461	rs2275128	5
6	rs11154234	chr6	124738177	ABSENT	ABSENT	NF-kappaB	V\$NFKB_Q6	0.9533	0.955	0.9571	1.0000	4.29	53	chr6:124738169-124738182	rs11154234	6

1.12.34 HapMap

Although it is possible to search the Delta-MATCH database for more than 2.8-fold and 2.9-fold the number of markers present on the Affymetrix 500k and Illumina 550k SNPchip (respectively) by looking for markers that are strong linkage disequilibrium to

those present on a given genotyping platform (examples 29 through 33), it is also very helpful to visualize the linkage relationship between markers in a results table. If you were to rerun examples 29 through 33 with the HapMap checkbox checked, there would be many additional columns returned in the results table that are derived from the HapMap database. These columns detail the linkage disequilibrium statistics between pairs of markers. Furthermore, it is possible to restrict results to those with a minimum r^2 , D' , r^2 linkage disequilibrium statistic, and a minimum likelihood odds score (LOD) (see the <http://www.hapmap.org> website for the details).

1.12.34.1 Definition - Id_name

This is the name of a polymorphism in strong LD with the given **rsnumber**. There may be multiple Id_names for each rsnumber. Clicking the hyperlink will open the UCSC Genome Browser link for the Id_name.

1.12.34.2 Definition - Id_name_affy

This is a statement of whether the Id_name is “present” or “absent” on the Affymetrix 500k SNPchip.

1.12.34.3 Definition - Id_name_illumina

This is a statement of whether the Id_name is “present” or “absent” on the Illumina 550k SNPchip.

1.12.34.4 Definition - Id_lod

This is the likelihood odds ratio statistic value for the LD between the **rsnumber** and **Id_name**.

1.12.34.5 Definition - ld_dprime

This is the dprime statistic value for the LD between the **rsnumber** and **ld_name**.

1.12.34.6 Definition - ld_rsquare

This is the rsquare statistic value for the LD between the **rsnumber** and **ld_name**.

1.12.34.7 Definition - ld_pos_dif

This is the number of base pairs between the **rsnumber** and **ld_name** (hg17).

1.12.34.8 Definition - ld_pos1_hg17

This is the base position of the **rsnumber** in the human genome (hg17).

1.12.34.9 Definition - ld_pos2_hg17

This is the base position of the **ld_name** in the human genome (hg17).

1.12.34.10 Definition - ld_fbin

This is a binning value for this LD pair.

Note: The actual search for markers in strong linkage disequilibrium with each rsnumber is done after all of the primary hits have been found. Therefore it is possible to use this HapMap function in conjunction with any one the previous examples. When no ld_names are found for a given rsnumber, the above 9 parameters are left blank.

1.12.35 Example 34 - Affymetrix with HapMap

This example returns the same 11 results as from example 30. However, additional columns in the results table detail other polymorphisms in the HapMap database (ld_name) that are in strong linkage disequilibrium with the each rsnumber. The number of ld_names returned for each rsnumber can be controlled (only those ld_names with an associated ld_square, ld_square, and ld_lod greater than or equal to the input parameters will be returned). In this example, rs6036746 is in perfect linkage disequilibrium (ld_prime = 1.0; ld_square = 1.0) with 9 polymorphisms (rs6049622, rs6036747, rs6036748, rs6114672, 6049623, rs6049624, rs1474735, rs1474734, and rs2143508). It might be expected associations between each of these 9 polymorphisms and a given phenotype might be equivalent (because of the strong linkage disequilibrium between them). However, rs6036746 is the only one of these ten polymorphisms predicted by Delta-MATCH to have a high potential (potential = 0.9978) to create an allele-specific transcription factor binding site for matrix V\$NFKB_Q6.

Figure 124 Input Parameters for Example 34

STEP 1 - (1) Single Transcription Factor Matrix Name = "V\$NFKB_Q6 (950)"

STEP 2 - Minimum Potential Score (checked) = "0.80"

STEP 2 - Top Most Significant Hits (**unchecked**)

STEP 2 - Matrix Quality (checked) = "high"

STEP 2 - Affymetrix (**checked**) = "500k_all (492,555)"; Include SNPs in strong LD with those on the AFFY 500k SNP-CHIP (**checked**) = "CEU rsquare >= 1.0 (1,058,667)"

STEP 2 - HapMap (**checked**); HapMap population = "CEU European"; ld_prime >= "1.00"; ld_square >= "0.8"; ld_lod >= "18"; View HapMap Details (**checked**)

1 - Single Transcription Factor Matrix Name

Select a 'Single Transcription Factor Matrix Name' (n=550)

V\$NFKB_Q6 (950) (mat_id)

[Back to top](#)

Minimum Potential Score

Select a 'Minimum Potential Score' (potential)

0.80 (0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix

5 Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (qual) [REF]

high ('high = 1', 'low = 0')

[Back to top](#)

Affymetrix

Limit results to rsnumbers on an Affymetrix SNP-CHIP [REF1, REF2]

500k_all (492,555)

Include SNPs in strong LD with those on the Affymetrix 500k SNP-chip (name_affy)

CEU rsquare >= 1.0 (1,058,667)

(LD = linkage disequilibrium)

[Back to top](#)

HapMap

Include other SNPs in strong linkage disequilibrium [REF]

(ld_name, ld_name_affy, ld_name_illumina, ld_lod, ld_dprime, ld_rsquare, ld_pos_dif, ld_pos1_hg17, ld_pos2_hg17, ld_fbin)

The following requirements will be met

CEU European HapMap population

(CEU = Caucassian, YRI = African, JPT = Japanese, CHB = Chinese)

1.00 (ld_dprime LD >= x)

0.80 (ld_rsquare LD >= x)

18 (ld_lod LD >= x)

View HapMap details

You must check this box to show these parameters, otherwise they will be hidden

Figure 125 Example 34 Results Table

PK	rsznumber	chrom	chromStart	name_spy	is_name	is_name_spy	is_loo	is_optim	is_savare	is_pos_diff	is_pos_1_hg17	is_pos_2_hg17	is_rbin	factor	mat_lo	posendal	phozhok	m1	m2	m_per	rank	p1_window	pubmed	PK		
1	rs022017	chr18	21201573	PRRS52NF										FE-AppoB	V24 NR_OB	1	0.205	0.2077	1.0000	10.23	7	chr18:21201573-21201574	rs022017	1		
2	rs022012	chr14	11220094	AB542NF [rs1818101]	AB542NF AB542NF	AB542NF AB542NF	21.27 21.27	1.0000	1.0000	1.0000	412 1244	11220093 11220093	11220097 11220097	111 111	FE-AppoB	V24 NR_OB	1	0.200	0.2047	1.0000	5.32	3	chr14:11220093-11220093	rs022012	2	
3	rs130000	chr3	121191207	AB542NF										FE-AppoB	V24 NR_OB	0.2010	0.2000	0.2010	10.2310	14	chr3:121191205-121191210	rs130000	3			
4	rs0494310	chr8	21220192	AB542NF	[rs0494310] [rs0494310]	PRRS52NF PRRS52NF	20.20 20.34	1.0000	1.0000	0.2000	1.0000	11920 19202	21220190 21220192	21240192 21241920	212 212	FE-AppoB	V24 NR_OB	0.2010	0.2010	0.2000	10.2310	10	chr8:21220190-21220190	rs0494310	4	
5	rs0431204	chr10	32200001	PRRS52NF										FE-AppoB	V24 NR_OB	0.2010	0.2000	0.2010	10.2310	21	chr10:32200000-32200001	rs0431204	5			
8	rs0220140	chr20	24204194	PRRS52NF	[rs0220140]	AB542NF	21.20	1.0000	1.0000	1.0000	3000 1900	24204190 24204190	24200000 24200001	242 242	FE-AppoB	V24 NR_OB	0.2010	0.2000	0.2000	5.4000	22	chr20:24204190-24204191	rs0220140	8		
				AB542NF	[rs0220140]	AB542NF	19.70	1.0000	1.0000	1.0000	1800 1800	24204190 24204190	24200000 24200001	242 242												
				PRRS52NF	[rs0220140]	PRRS52NF	21.20	1.0000	1.0000	1.0000	1000 1000	24204190 24204190	24200000 24200001	242 242												
				PRRS52NF	[rs0220140]	PRRS52NF	21.14	1.0000	1.0000	1.0000	3000 2040	24204190 24204190	24200000 24200001	242 242												
				AB542NF	[rs0220140]	AB542NF	19.40	1.0000	1.0000	1.0000	2000	24204190	24200000	242												
				PRRS52NF	[rs0220140]	PRRS52NF	19.31																			
PRRS52NF	[rs0220140]	PRRS52NF	21.20																							
PRRS52NF	[rs0220140]	PRRS52NF	21.20																							
AB542NF	[rs0220140]	AB542NF	19.41																							
7	rs043200	chr8	10202040	AB542NF										FE-AppoB	V24 NR_OB	0.2000	0.2000	0.2000	11.0000	04	chr8:10202040-10202040	rs043200	7			
9	rs11019120	chr3	30010000	AB542NF										FE-AppoB	V24 NR_OB	0.2004	0.2000	0.2010	10.2414	40	chr3:30010000-30010000	rs11019120	9			
9	rs00170	chr8	191191704	AB542NF	[rs00170]	AB542NF	20.02	1.0000	0.2010	0.2000	191191700	191220007	1911	FE-AppoB	V24 NR_OB	0.2004	0.2000	0.2000	10.2414	40	chr8:191191700-191191700	rs00170	9			
10	rs0270120	chr10	19410400	AB542NF	[rs0270120]	AB542NF	20.14	1.0000	1.0000	1.0000	240 1600	19410401 19410401	19410000 19410000	194 194	FE-AppoB	V24 NR_OB	0.2004	0.2000	0.2010	10.2414	31	chr10:19410400-19410401	rs0270120	10		
				AB542NF	[rs0270120]	AB542NF	22.20	1.0000	1.0000	0.2000	4000 4101	19410401 19410401	19421000 19422010	194 194												
				AB542NF	[rs0270120]	AB542NF	21.01	1.0000	0.2000	0.2000	3000	19410401	19421000	194												
				AB542NF	[rs0270120]	AB542NF	21.01																			
				AB542NF	[rs0270120]	AB542NF	21.01																			
11	rs11194204	chr8	124120117	AB542NF	[rs11194204] [rs11194204]	PRRS52NF AB542NF	22.00 21.12	1.0000	1.0000	0.2110	0.2000	0.2000	2000	124120110	124140007	1241	FE-AppoB	V24 NR_OB	0.2000	0.2011	1.0000	4.20	50	chr8:124120100-124120102	rs11194204	11

1.12.36 Example 35 - Affymetrix with HapMap (with Minimum Total Number of Delta-MATCH Hits)

This example uses the same parameters as example 34 plus the addition of checking “Minimum Total Number of Delta-MATCH Hits”. It returns all 11 of the hits found in examples 30 and 34. In the results table are two columns (number_hits and Id_number_hits) that show the “Total number of Delta-MATCH Hits for each rsnumber and Id_name respectively. In this example rs6036746 has four hits (number_hits = 4). Notice that rs1474734 (Id_name (8) for rs6036746) has 3 total hits in the Delta-MATCH database. If the “Minimum Total Number of Delta-MATCH Hits” had been set to a higher number (7), only four of the 11 rsnumbers would have been returned (where number_hits >= 7).

Figure 126 Input Parameters for Example 35

(Same as Example 34 plus the following)

STEP 2 - Minimum Total Number of Delta-MATCH Hits (checked) = “1”

Minimum Total Number of Hits

Limit results to rsnumbers with a minimum 'total number of hits'
This is the sum number of hits for an rsnumber in the database
(number_hits >= x)

Figure 127 Example 35 Results Table (partial)

hit	rsnumber	chrom	chromStart	number_hits	name_affy	ld_name	ld_number_hits	ld_name_affy
1	rs3093317	chr16	27351578	5	PRESENT			
2	rs2283412	chr14	71728984	7	ABSENT	(1)rs2283411 (2)rs7161011	(1)? (2)1	ABSENT PRESENT
3	rs7380665	chr5	121731207	7	ABSENT			
4	rs4074910	chr6	21228732	8	ABSENT	(1)rs9460598 (2)rs9465970	(1)1 (2)?	PRESENT PRESENT
5	rs6481864	chr10	33968031	8	PRESENT			
6	rs6036746	chr20	24284794	4	PRESENT	(1)rs6049622 (2)rs6036747 (3)rs6036748 (4)rs6114672 (5)rs6049623 (6)rs6049624 (7)rs1474735 (8)rs1474734 (9)rs2143508	(1)? (2)? (3)2 (4)1 (5)? (6)? (7)1 (8)3 (9)?	ABSENT ABSENT PRESENT PRESENT ABSENT PRESENT PRESENT PRESENT ABSENT
7	rs4348296	chr6	162392452	5	ABSENT			
8	rs17013128	chr3	23376599	5	ABSENT			
9	rs6975	chr6	131197794	6	ABSENT	(1)rs915171	(1)?	ABSENT
10	rs2275128	chr10	13418450	5	ABSENT	(1)rs7901303 (2)rs3802583 (3)rs10752298 (4)rs6602648 (5)rs1005089	(1)? (2)1 (3)3 (4)1 (5)1	ABSENT ABSENT ABSENT ABSENT ABSENT
11	rs11154234	chr6	124738177	5	ABSENT	(1)rs10457447 (2)rs12195624	(1)2 (2)?	PRESENT ABSENT

1.12.37 HIV-1 Candidate Genes

It is possible to restrict results to include rsnumbers that have been investigate for their association to HIV-1 by checking the box for “HIV-1 Candidate Genes”. Data have been adapted from a whole genome association study of major determinants for host control of HIV-1 [34].

1.12.38 Example 36 - HIV-1 Candidate Genes

This example will return rsnumbers for the “high quality” NF-kappaB matrixes that have potential scores greater than or equal to 0.5, and have HIV-1 log P-values (-logp) that are greater than or equal to 1.0. Exactly 12 rsnumbers are returned. Using this checkbox will make the output include a column (-logp) that describes the association study P-value significance.

1.12.38.1 Definition - log P-value (-logp)

This is the log-transformed p-value for the significance of this rsnumber from the genome-wide association study (Fellay et al.) [34].

Figure 128 Input Parameters for Example 36

STEP 1 - (3) Transcription Factor Name = “NF-kappaB”

STEP 2 - Minimum Potential Score (**checked**) = “0.3”

STEP 2 - Top Most Significant Hits (**unchecked**)

STEP 2 - Matrix Quality (checked) = “high”

STEP 2 - Sort Results Table (**checked**) = “chrom asc, position asc (a)”

STEP 2 - HIV-1 Candidate Genes (**checked**); (**-logp = “1.0”**)

3 - Transcription Factor Name

Select transcription factor matrix names by a 'Transcription Factor Name' (n=351)

NF-kappaB (factor)

Minimum Potential Score

[Back to top](#)

Select a 'Minimum Potential Score' (potential)

0.50 (0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix
 Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (qual) [REF]
 ('high = 1', 'low = 0')

Sort Results Table

Sort the results table by

(asc = ascending, desc = descending)

[Back to top](#)

HIV-1 Candidate Genes

Limit results to those from the 'Database of HIV-1 Candidate Genes'
where an rsnumber had an significance greater than or equal to a (-logp) value [REF]
 (-logp >= x)

1.12.39 Copy Number Variation

Some regions of the genome are known to be associated with copy number variation. It is possible to search the Delta-MATCH database for only rsnumbers that are positioned in regions of copy number variation as specified in the “Database of Genomic Variants” (<http://projects.tcag.ca/variation/>) [35, 36].

1.12.40 Example 37 - Copy Number Variation

This example will search for all “V\$NFKB_Q6” rsnumbers that have potential scores greater than or equal to 0.8 that are located in a region of copy number variation. Exactly 14 results are returned.

Figure 129 Input Parameters for Example 37

STEP 1 - (1) Single Transcription Factor Matrix Name = “V\$NFKB_Q6 (950)”

STEP 2 - Minimum Potential Score (checked) = “0.8”

STEP 2 - Top Most Significant Hits (**unchecked**)

STEP 2 - Matrix Quality (checked) = “high”

STEP 2 - Copy Number Variation (**checked**)

1 - Single Transcription Factor Matrix Name

Select a 'Single Transcription Factor Matrix Name' (n=550)

V\$NFKB_Q6 (950) (mat_id)

Minimum Potential Score

[Back to top](#)

Select a 'Minimum Potential Score' (potential)

0.80 (0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix

5 Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (qual) [REF]

high ('high = 1', 'low = 0')

Copy Number Variation

Limit results to those 'within' a region of 'Copy Number Variation' (CNV)
as described in the 'Database of Human Genomic Variants' (hg18.v2) [REF1, REF2]

1.12.41 PReMod Modules

When the “PReMod” checkbox is checked, it is possible to require that returned rsnumber hits are located within regulatory regions called PReMod Modules [24, 25]. If

no modules are found that fit the specified criteria, Error 9 will be returned (Figure 213 page 413). Please view the “PReMod key” text file to learn identify the relationships between the “FACTOR”, “MODULE_MATRIX”, and “MAT_ID” names. There are currently 123,510 PReMod modules defined and mapped to the human genome SNP database (UCSC table hg17.snp125).

1.12.42 Example 38 - PReMod Modules

This search returns the list of Delta-MATCH predictions for polymorphisms that are located within regulatory regions called PReMod modules for all TFBS matrixes. Only polymorphisms located with PReMod modules tagged for both NF-kappaB (“M00769” = “V\$NFKB_Q6”) and SMAD-3 (“M00701” = “V\$SMAD3_Q6”) are considered. The report.html file (Figure 131 page 231) shows that 22 results are returned and located within 8 PReMod modules for 14 unique rsnumbers.

Figure 130 Input Parameters for Example 38

STEP 1 - (5) All Transcription Factor Matrix Names

STEP 2 - Minimum Potential Score (unchecked)

STEP 2 - Top Most Significant Hits (unchecked)

STEP 2 - Matrix Quality (unchecked)

STEP 2 - Sorted Results Table (checked) = “chrom asc, position asc (a)”

STEP 2 - PReMod Modules (checked); input 5 terms max = “M00769, M00701”; “and”

5 - All Transcription Factor Matrix Names

Select all matrix names in the database (n=550)

Minimum Potential Score

Select a 'Minimum Potential Score' (**potential**)

(0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix

Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (**qual**) [[REF](#)]

('high = 1', 'low = 0')

Sort Results Table

Sort the results table by

(asc = ascending, desc = descending)

PReMod Modules

Limit to rsnumbers positioned within 'PReMod Modules' [[REF](#)]

(input 5 terms max)

and or

Select your 'factor' or 'module_matrix' names from this [PReMod key](#)

NOTE - there are 123,510 modules in the 'human_module_database' mapped to 'hg17.snp125'

Figure 131 Example 38 PReMod Modules Summary (report.html)

There were 22 'Delta-MATCH hits' returned

```
SELECT * FROM premod.human_module_tab WHERE ( tag1 like '%M00769%' OR tag2 like '%M00769%' OR tag3 like '%M00769%' OR tag4 like '%M00769%' OR tag5 like '%M00769%') AND ( tag1 like '%M00701%' OR tag2 like '%M00701%' OR tag3 like '%M00701%' OR tag4 like '%M00701%' OR tag5 like '%M00701%');
```

8 PReMod Modules were found:

mod013019, mod048393, mod057861, mod070751, mod074350, mod080327, mod092836, mod114214

14 rsnumbers were found in 8 PReMod Modules

1.12.43 UCSC rsnumber Details

When the “UCSC rsnumber Details” checkbox is checked, it is possible to restrict results to include other selection criteria that are specific human SNP database (hg18.snp126). The reference base for the UCSC browser (**refUCSC**) and NCBI database (**refNCBI**) for the two observed alleles (**observed**), and the strand of the rsnumber (**strand**) can be viewed. Note: the reference base at UCSC will be the reverse complement of one of the two observed alleles when the “strand” is “-“. It is possible require a certain minimum average heterozygosity (**avHet**), or limit results by “Validation” (**validtype**), “Function” (**functype**), “Location” (**loctype**), and “Molecular” (**moltype**) type [37, 38]. It is possible to us the “and” / “or” buttons below the selections to control if the results are the intersection or union of the selected criteria. The number of polymorphisms categorized by each parameter is listed in parentheses [by-2hit-2allele (1,692,687)].

1.12.43.1 Warning - Using the “and” buttons will greatly increase computation time.

1.12.43.2 Definition - reference base at the UCSC Browser (refUCSC)

This is the reference base displayed on the “plus” strand of the human genome browser (hg18.snp126.refUCSC).

1.12.43.3 Definition - reference base at NCBI (refNCBI)

This is the reference base displayed on the “plus” strand of the human genome browser (hg18.snp126.refNCBI).

1.12.43.4 Definition - the observed alleles at this rsnumber (observed)

These are the two alleles for this rsnumber (hg18.snp126.observed).

1.12.43.5 Definition - rsnumber strand (strand)

This is the strand (“+” or “-“) of the rsnumber (hg18.snp126.strand).

1.12.43.6 Definition - Validation Types (validtype)

This is the average heterozygosity of the rsnumber (hg18.snp126.valid).

1.12.43.7 Definition - Function Types (functype)

This is the average heterozygosity of the rsnumber (hg18.snp126.func).

1.12.43.8 Definition - Locations Types (loctype)

This is the average heterozygosity of the rsnumber (hg18.snp126.loctype).

1.12.43.9 Definition - Molecular Types (moltype)

This is the average heterozygosity of the rsnumber (hg18.snp126.moltype).

1.12.43.10 Definition - Average Heterozygosity (avHet)

This is the average heterozygosity of the rsnumber (hg18.snp126.avHet).

1.12.43.11 Definition - Average Heterozygosity (avHetSE)

This is the standard error of the average heterozygosity of the rsnumber (hg18.snp126.avHetSE).

1.12.44 Example 39 - UCSC rsnumber Details

This example will search for “V\$_NFkB_Q6” rsnumbers that have a minimum average heterozygosity (avHet) greater than or equal to 0.05, that have a “by-2hit-2allele” valid type (validtype), a “locus” function type (functype), an “exact” location type (loctype), and a “genomic” molecular type (moltype). Exactly 13 rsnumbers are returned.

Figure 132 Input Parameters for Example 39

STEP 1 - (1) Single Transcription Factor Matrix Name = “V\$NFkB_Q6 (950)”

STEP 2 - Minimum Potential Score (unchecked)

STEP 2 - Top Most Significant Hits (unchecked)

STEP 2 - Matrix Quality (checked) = “high”

STEP 2 - UCSC rsnumber Details (**checked**); “Select Minimum Average Heterozygosity

Cutoff = “0.05” (**checked**); “Select ‘Validation’ Types” (**checked/by-2hit-2allele/or**);

“Select ‘Function Types” (**checked/locus/or**); Select ‘Location Types”

(**checked/exact/or**); “Select ‘Molecular Types” (**checked/genomic/or**)

1 - Single Transcription Factor Matrix Name

Select a 'Single Transcription Factor Matrix Name' (n=550)

V\$NFkB_Q6 (950) (mat_id)

Minimum Potential Score

[Back to top](#)

Select a 'Minimum Potential Score' (potential)

0.80 (0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix

5 Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (**qual**) [REF]

high (high = 1, low = 0)

UCSC rsnumber Details

[Back to top](#)

Show the rsnumber details from UCSC hg18.snp126 Table ([avHet](#), [avHetSE](#), [refUCSC](#), [refNCBI](#)) [[REF1](#), [REF2](#)]

Select Minimum Average Heterozygosity Cutoff (**avHet**)

0.05 (0 <= **avHet** <= 1.0)

Select 'Validation Types' (**validtype**)

- by-2hit-2allele (1,692,687)
- by-cluster (1,154,345)
- by-frequency (1,933,537)
- by-submitter (214,482)
- by-hapmap (9)
- unknown (1,755,067)

and or

Select 'Function Types' (**functype**)

[Back to top](#)

- locus (211,913)
- coding (90,767)
- coding-synon (40,422)
- coding-nonsynon (50,572)
- untranslated (92,688)
- intron (2,848,608)
- splice-site (678)
- cds-reference (0)
- unknown (1,364,457)

and or

Select 'Location Types' (**loctype**)

[Back to top](#)

- exact (4,784,820)
- range (13,202)
- between (4,866)
- rangeInsertion (2,909)
- rangeSubstitution (251)
- rangeDeletion (4,866)
- unknown (0)

and or

Select 'Molecular Types' (**moltype**)

[Back to top](#)

genomic (4,493,416)

cDNA (54,425)

unknown (0)

and or

Figure 133 Example 39 Results Table (partial)

refUCSC	refNCBI	observed	strand	validtype	funcntype	loctype	moltype	avHet	avHetSE	pubmed	hit
T	T	A/T	+	by-frequency,by-2hit-2allele	locus	exact	genomic	0.299444	0.245062	rs7928331	1
C	C	C/T	+	by-frequency,by-2hit-2allele	locus,intron	exact	genomic	0.398858	0.200852	rs4075655	2
G	G	A/G	+	by-frequency,by-2hit-2allele	locus	exact	genomic	0.20355	0.245647	rs9898132	3
T	T	C/T	+	by-cluster,by-frequency,by-2hit-2allele	locus	exact	genomic	0.42	0.183303	rs7294536	4
G	G	C/G	+	by-frequency,by-2hit-2allele	locus,intron	exact	genomic	0.496172	0.0435822	rs669340	5
A	A	A/G	+	by-cluster,by-frequency,by-2hit-2allele	locus,intron	exact	genomic	0.437045	0.165874	rs11810295	6
G	G	C/G	+	by-cluster,by-frequency,by-submitter,by-2hit-2allele	locus	exact	genomic	0.339846	0.233297	rs1483979	7
C	C	C/G	+	by-cluster,by-frequency,by-2hit-2allele	locus	exact	genomic	0.408383	0.193429	rs3753444	8
G	G	C/T	-	by-frequency,by-2hit-2allele	locus	exact	genomic	0.132228	0.220521	rs322107	9
C	C	C/T	+	by-cluster,by-frequency,by-2hit-2allele	locus,intron	exact	genomic	0.197846	0.244499	rs11928674	10
A	A	A/G	+	by-cluster,by-frequency,by-2hit-2allele	locus	exact	genomic	0.210435	0.246849	rs2984920	11
G	C	C/T	-	by-cluster,by-frequency,by-2hit-2allele	locus	exact	genomic	0.388878	0.207877	rs3853419	12
G	G	A/G	+	by-cluster,by-frequency,by-submitter,by-2hit-2allele	locus	exact	genomic	0.357678	0.225622	rs1800686	13

1.12.45 Example 40 - NF-kB (rs5743836, rs6031444, rs28431981)

This example will search for all NF-kB hits for the three specified rsnumbers. Ten hits are returned; rs28431981 has four, rs6031444 has four and rs5743836 has three. These results are detailed in the following section (page 239).

Figure 134 Input Parameters for Example 40

STEP 1 - (3) Transcription Factor Name = "NF-kappaB" STEP 2 - Minimum Potential Score (unchecked)

STEP 2 - Top Most Significant Hits (unchecked)

STEP 2 - Matrix Quality (checked) = "high"

STEP 2 - Sort Results Table (checked) = "chrom asc, position asc (a)"

STEP 2 - Search By rsnumbers (checked); rsnumbers = "rs5743836, rs6031444, rs28431981"; rsnumber Window (unchecked)

3 - Transcription Factor Name

Select transcription factor matrix names by a 'Transcription Factor Name' (n=351)

(factor)

Minimum Potential Score

[Back to top](#)

Select a 'Minimum Potential Score' (potential)

(0 <= x <= 1.0)

Top Most Significant Hits

Select the maximum number of returned rsnumbers per selected matrix

Limit the number of polymorphisms searched per matrix

Matrix Quality

Limit results by matrix 'Quality Type' (qual) [REF]

('high = 1', 'low = 0')

Sort Results Table

Sort the results table by

(asc = ascending, desc = descending)

[Back to top](#)

Search By rsnumbers

Either limit results by a comma-separated list of dbSNP rsnumbers (1024 chars max)

[\[REF1, REF2\]](#)

'rsnumbers'

Or upload list of rsnumbers in a plain text file

'rsnumber filename' [download example file](#)

(one rsnumber per row, 10,000 rsnumbers max)

NOTE - The 'rsnumber' text field will be ignored if an 'rsnumber filename' is uploaded

Search additional bases upstream and downstream of specified rsnumbers

'rsnumber Window'

Include other rsnumbers within this many bases

Show the Matrix Details

[Back to top](#)

Show the matrix details [\[REF\]](#)

([factor_description](#), [count_ge_potential](#), [mat_count](#), [rareness](#), [qual](#), [mat_len](#))

Minimum Matrix Length

Limit searches to those matrixes with minimum length ([mat_len](#))

([mat_len](#) >= x)

Show the Position Details

Show the position and strand details [\[REF\]](#)

([p1_window](#), [p2_window](#), [p1](#), [p2](#), [s1](#), [s2](#))

HUGO Names

Show the HUGO names of the genes associated with each rsnumber [\[REF1, REF2\]](#)

([hugo_name](#))

Limit results to rsnumbers next to known HUGO genes

Download the rsnumber to hugo name file

(*WARNING 48.4 Mb, right-click and 'download file'*) [SNP-Genes_HUGO.txt](#)

1.13 Predicting Modulators of NF-κB-dependent Transcription

Six different [NF-κB](#) TFBS matrixes can be searched with Delta-MATCH (Table 40 page 404). The distribution of the 4.5 million potential scores for the NF-κB TFBS matrixes is shown (Table 41 page 404). Exactly 950 SNPs were identified with at least one allelic MATCH score greater than or equal to the FP threshold score of 0.955 when matched against the V\$NFKB_Q6 TFBS matrix (Example 6 page 148).

Specifically, three SNPs with strong potential scores for [NF-κB](#) are noteworthy [Junctophillin 2 ([JPH2](#)) [rs6031444](#) G>T, Toll-like receptor 9 ([TLR9](#)) [rs5743836](#) T>C, and kynurenine 3-monooxygenase ([KMO](#)) [rs28431981](#) A>G] (Example 40).

Figure 135 JPH2 rs6031444, TLR9 rs5743836, and KMO rs28431981

NF-κB is Predicted to bind specifically to the JPH2 rs6031444 G>T major allele, the TLR9 rs5743836 T>C and KMO rs28431981 A>G minor alleles. This is the output file for Example 40.

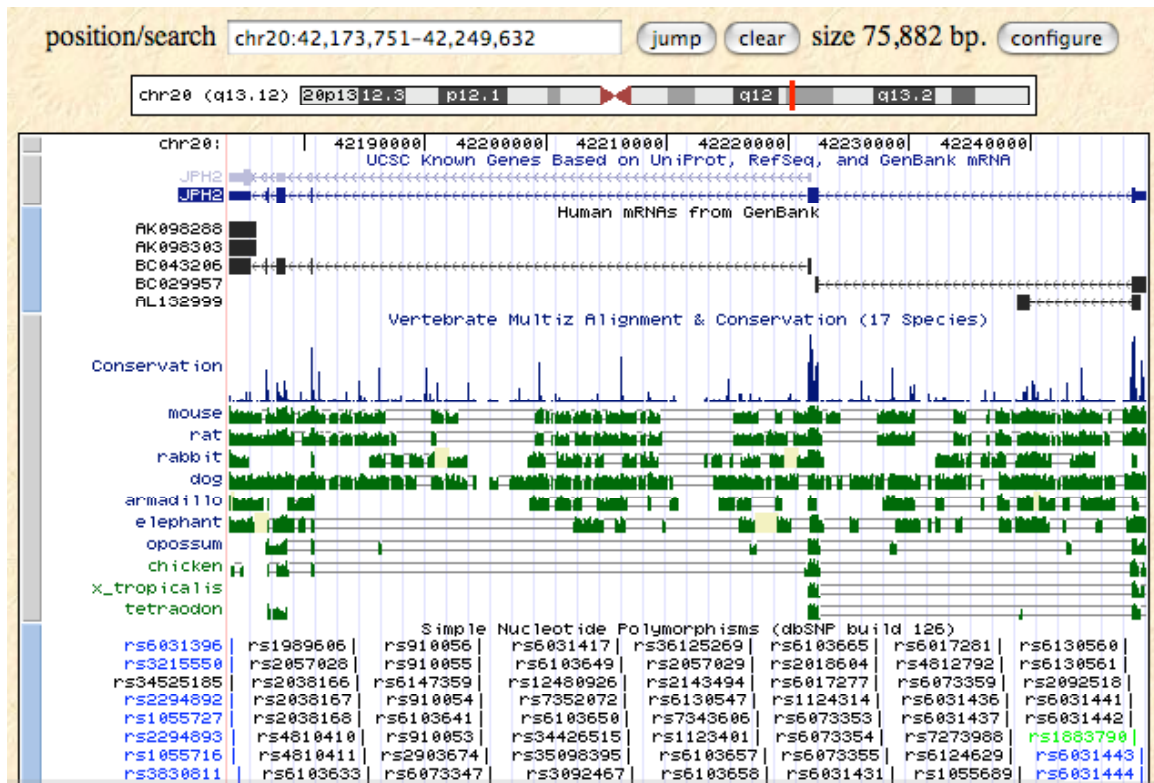
ht	rsnumber	chrom	chromStart	factor	mat_id	potential	threshold	m1	m2	m_per	rank	factor_description	count_ge_potential	mat_count	rareness	qual	mat_len	hugo_name	dist_from_ref	dist_from_tx	dist_from_cds	p1_window	p2_window	a1	a2	p1	p2	s1	s2	pubmed	ht
1	rs6031444	chr20	42249151	NF-κB_Q6	V\$NFKB_Q6_1	0.991	0.955	0.991	0.991	24	24	NF-κB_Q6 (Q6)	167	167	2.6E-29-0	1	10	JPH2	42249150	42249150	42249150	(1)none	(1)none	G	T	-	-	-	-	rs6031444	1
2	rs28431981	chr1	23977876	NF-κB_Q6	V\$NFKB_Q6_1	0.984	0.955	0.955	0.955	40	40	NF-κB_Q6	153	505	3.362E-5	1	10	TLR9	23977875	23977875	23977875	(1)none	(1)none	A	G	-	-	-	-	rs28431981	2
3	rs28431981	chr1	23977876	NF-κB_Q6	V\$NFKB_Q6_1	0.984	0.955	0.955	0.955	41	41	NF-κB_Q6	51	350	1.121E-4	1	14	KMO	23977875	23977875	23977875	(1)none	(1)none	A	G	-	-	-	-	rs28431981	3
4	rs28431981	chr1	23977876	NF-κB_Q6	V\$NFKB_Q6_1	0.939	0.955	0.939	0.939	1144	1177	NF-κB_Q6	1377	24160	2.589E-4	1	16		23977875	23977875	23977875	(1)none	(1)none	A	G	-	-	-	-	rs28431981	4
5	rs6031444	chr20	42249151	NF-κB_Q6	V\$NFKB_Q6_1	0.991	0.955	0.991	0.991	24	24	NF-κB_Q6 (Q6)	167	167	3.672E-5	1	10	JPH2	42249150	42249150	42249150	(1)none	(1)none	G	T	-	-	-	-	rs6031444	5
6	rs6031444	chr20	42249151	NF-κB_Q6	V\$NFKB_Q6_1	0.984	0.955	0.979	0.979	25	25	NF-κB_Q6	153	505	3.362E-5	1	10	JPH2	42249150	42249150	42249150	(1)none	(1)none	G	T	-	-	-	-	rs6031444	6
7	rs6031444	chr20	42249151	NF-κB_Q6	V\$NFKB_Q6_1	0.955	0.955	0.955	0.955	1	1	NF-κB_Q6	7	505	1.552E-4	1	14	JPH2	42249150	42249150	42249150	(1)none	(1)none	G	T	-	-	-	-	rs6031444	7
8	rs6031444	chr20	42249151	NF-κB_Q6	V\$NFKB_Q6_1	0.955	0.955	0.955	0.955	1117	1117	NF-κB_Q6	126	24160	2.592E-4	1	14	JPH2	42249150	42249150	42249150	(1)none	(1)none	G	T	-	-	-	-	rs6031444	8
9	rs5743836	chr3	52235821	NF-κB_Q6	V\$NFKB_Q6_1	0.973	0.955	0.973	0.973	331	331	NF-κB_Q6	332	350	7.322E-5	1	14	TLR9	52235820	52235820	52235820	(1)none	(1)none	A	G	-	-	-	-	rs5743836	9
10	rs5743836	chr3	52235821	NF-κB_Q6	V\$NFKB_Q6_1	0.973	0.955	0.973	0.973	158	158	NF-κB_Q6	170	24160	3.789E-5	1	16	TLR9	52235820	52235820	52235820	(1)none	(1)none	A	G	-	-	-	-	rs5743836	10

1.13.1 Junctophillin 2 (JPH2) rs6031444 G>T

Only 7 polymorphisms out of 4,547,844 ranked as high as [JPH2 rs6031444](#) when matched against the V\$NFKB_Q6 TFBS matrix (potential =1.0). [JPH2](#) has multiple transcriptional start sites and transcriptional isoforms of different lengths (Figure 137 page 242). One of these isoforms is expressed highly in the right ventricle of heart

muscle, is an essential component of the junctional complexes between the plasma membrane and the endoplasmic/sarcoplasmic reticulum, and participates in Ca^{2+} homeostasis in myocytes [39, 40]. [JPH2](#) is downregulated in hypertrophic and dilated cardiomyopathies, and [JPH2](#)-deficient mice die during embryogenesis [40, 41]. It may be an important finding that [rs6031444](#) is located only 5 bp upstream from one of the transcriptional start sites, isoforms (AL132999) (Figure 136 page 241). It is plausible that in the heart, a tissue where [NF-kB](#) is highly expressed, the [rs6031444](#) G allele (m1 = 1.0) may be more likely than the T allele (m2 = 0.8895) to recruit [NF-kB](#) and drive the expression of the shorter [JPH2](#) isoform. Although the frequency this polymorphism in the general human population is unknown, I hypothesize [rs6031444](#) should be considered an important candidate SNP for cardiomyopathies and heart arrhythmias.

Figure 137 JPH2 rs6031444 G>T in the UCSF Browser (zoom out)



1.13.2 Toll-like receptor 9 (TLR9) rs5743836 T>C

The [TLR9 rs5743836](#) minor C allele ($m_2 = 0.974$) is predicted to better recruit [NF- \$\kappa\$ B](#) than the major T allele ($m_1 = 0.853$). Only 333 other SNPs in the database ranked greater than or equal to the TLR9 rs5743836 potential score (potential ≥ 0.4133) for the V\$NF κ B_Q6 TFBS matrix (Figure 138 page 244). I investigated this polymorphism for association with HIV-1 viremia levels by genotyping rs5743836 in an HIV-positive cohort (see AIM 2, TLR9 - Toll-Like Receptor 9 page 293). Although the [TLR9 rs5743836](#) C allele by itself did not significantly associate with elevated viremia in the White-Americans, TLR9 haplotype 1 ([rs352140](#) G, [rs352139](#) A, [rs5743836](#) T, [rs187084](#) T) did associate significantly with elevated viremia in the white HIV-1-infected population. It remains to be determined if the association between the [TLR9](#) haplotype 1 with viremia is [NF- \$\kappa\$ B](#)-dependent.

Figure 138 TLR9 rs5743836 T>C May Create An Allele-specific NF-kB Binding Site

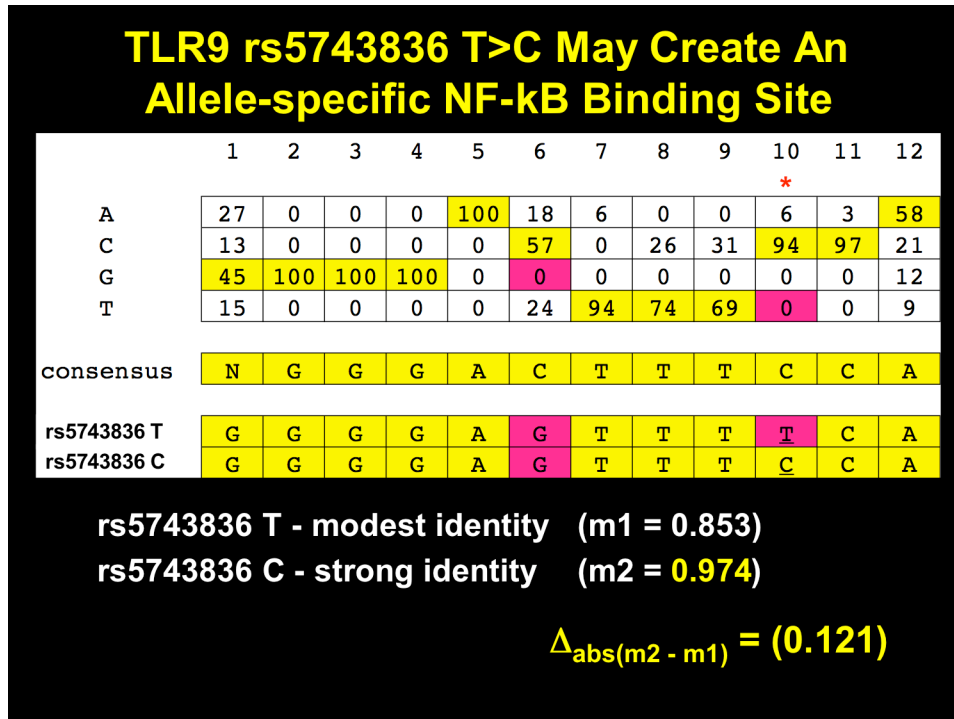
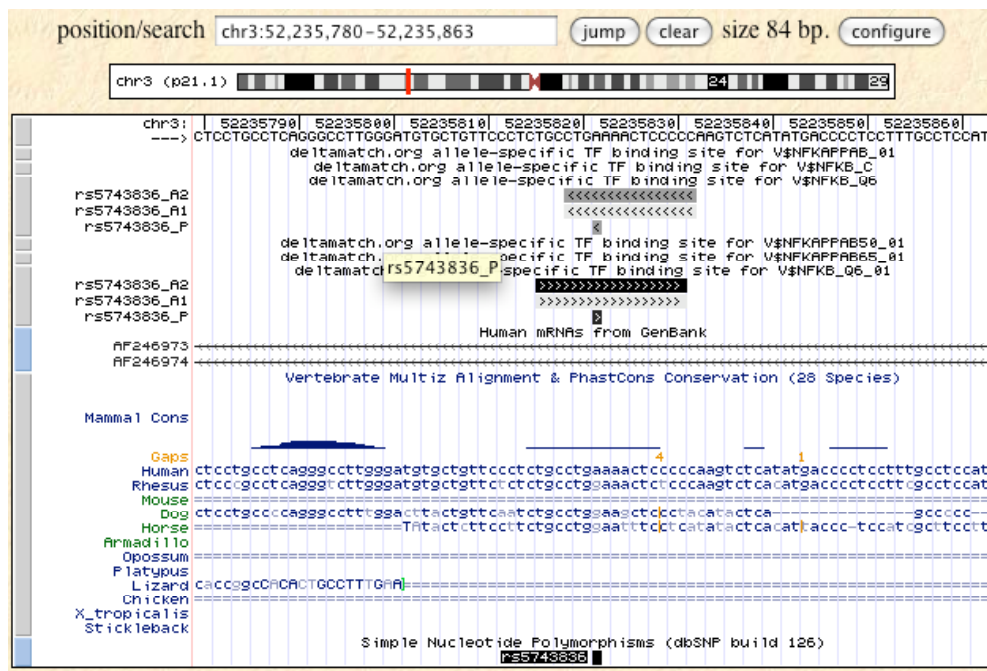


Figure 139 TLR9 rs5743836 T>C in the UCSF Browser

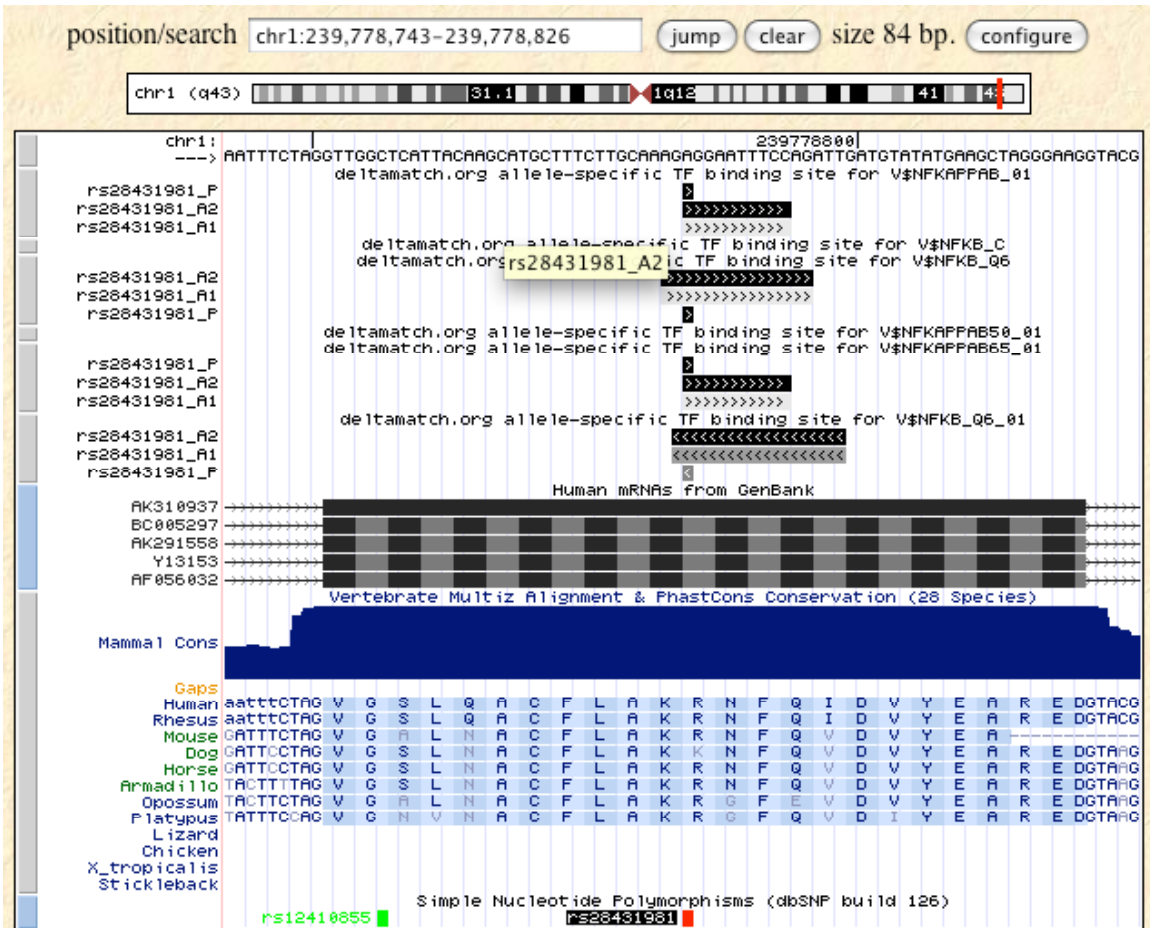


1.13.3 Kynurenine 3-monooxygenase (KMO) rs28431981 A>G

Delta-MATCH predicted that [KMO rs28431981](#) A>G may bind [NF- \$\kappa\$ B](#) in an allele-specific manner to modulate the efficiency of catalyzing the hydroxylation of L-kynurenine to form L-3-hydroxykynurenine. Only 333 other SNPs in the database ranked greater than or equal to the [KMO rs28431981](#) potential score (potential ≥ 0.984) for the V\$NF κ B_Q6 TFBS matrix. Loss-of-function mutation screens in yeast identified [KMO](#) as a suppressor of huntingtin protein ([Htt](#)) toxicity, and [rs28431981](#) may be considered a candidate SNP associated with Huntington disease [42, 43].

It is noteworthy that rs28431981 is located in KMO exon at the place of a non-synonymous substitution ([Arg30Gly](#)). However, after genotyping 240 individuals with RFLP and Taqman assays in sample population of Turks, I have failed to identify the higher scoring minor allele. I conclude the rs28431981 polymorphism is very rare, if it exists at all in the Turkish population. It is possible the higher scoring G allele ($m2 = 0.9844$) may not truly exist in humans and may be a sequencing artifact. Even so, there are plans to genotype this and other Delta-MATCH predicted candidate SNPs in a Huntington disease cohort through collaboration with Paul Muchowski and Daniel Zwilling in the Gladstone Institute of Neurological Disease ([GIN](#)D).

Figure 140 KMO) rs28431981 A>G in the UCSF Browser

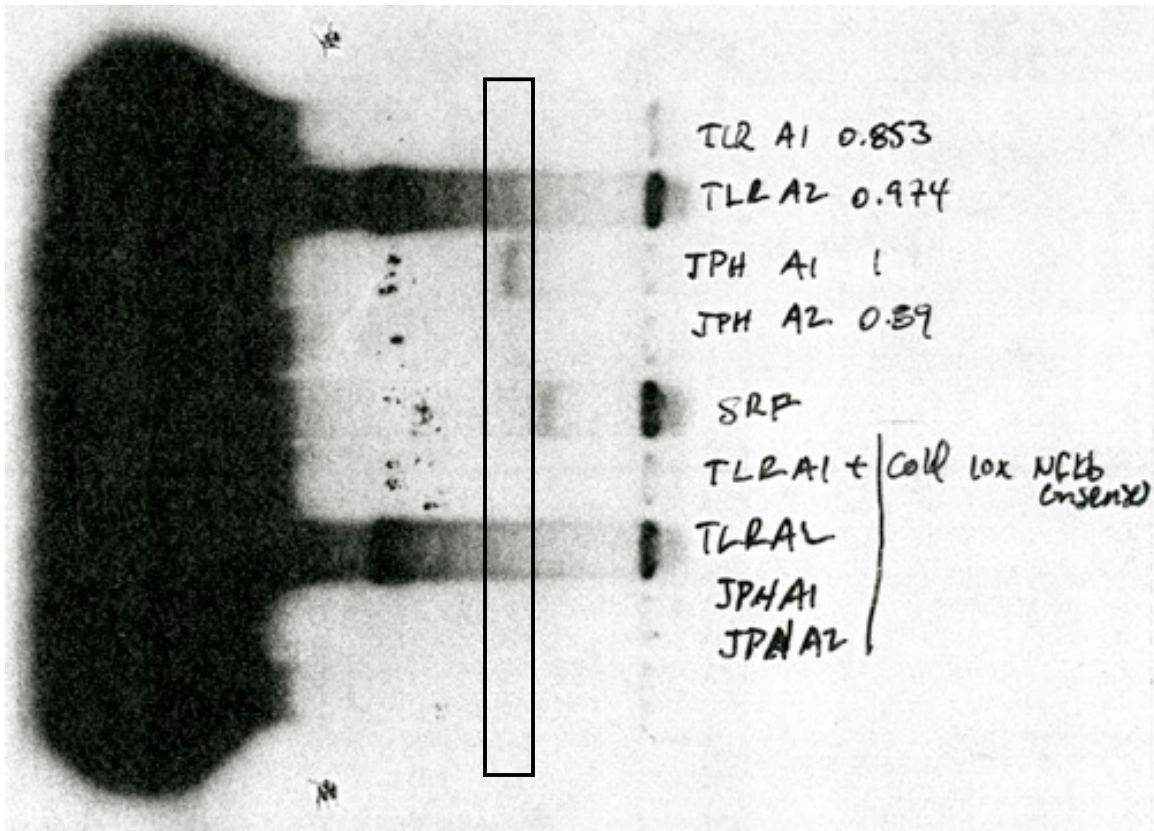


1.13.4 Validating the Delta-MATCH NF-kB Predictions

The molecular validation of the Delta-MATCH predictions has begun with the collaboration of Alex Zamboni ([GICD](#)). Preliminary studies with electrophoretic mobility shift assays (EMSA) showed that radiolabeled double-stranded oligos corresponding to the [TLR9](#) and [JPH2](#) alleles bind to nuclear extract isolated from mouse heart muscle (a tissue rich in [NF-kB](#)). The strength of binding correlated with the predicted MATCH score, and the difference in binding between a pair of alleles correlated with the Delta-

MATCH potential score (Figure 141). Future EMSA and luciferase reporter assays may validate other important [NF-kB](#) Delta-MATCH predictions.

Figure 141 EMSA for JPH2 rs6031444 G>T and TLR9 rs5743836 T>C.



Nuclear extract rich in NF-kB and SRF was isolated from mouse heart tissue and incubated with p-32 radiolabeled double-stranded-oligos specific for the JPH2 and TLR9 major (A1) and minor alleles (A2) (Lanes 1-4). The allelic MATCH scores for TLR9 (m1 = 0.853, m2 = 0.974 and JPH2 (m1 = 1.0, m2 = 0.89) were retrieved from the database. Delta-MATCH calculated TLR9 rs5743836 T>C and JPH2 rs6031444 G>T had potential scores equal to 0.4133 1.0, respectively. A single stronger band is present with the JPH2 major G allele (A1), and a partial band is present (with smear) with the TLR9 minor C

allele (A2) oligos (inside the black rectangle). These bands were absent when competed with 10x unlabelled oligo (Lanes 5-9). A single band is seen for oligo specific for SRF at a position different from the NF-kB band (Lane 5). Note: this gel is considered preliminary data and was produced by Alex Zambon ([GICD](#)).

1.14 Validating Other Delta-MATCH Predictions.

Delta-MATCH has been used to produce lists of candidate SNPs and gene targets involved with various biological pathways and disease phenotypes. These predictions may be validated through classical genetic investigation with various collaborators. Delta-MATCH predictions have been used to investigate a diversity of disease phenotypes, including aortic valve calcification (Vishal Nigam, [GICD](#)), multiple sclerosis (Sergio Baranzini and Jorge Oksengerg, [UCSF Multiple Sclerosis Center](#)), cardiomyopathy (Alex Zambon, [GICD](#)), cardiovascular disease and dyslipidemia (Ugur Hodoglugil, [GICD](#)), non-insulin-dependent diabetes mellitus (Sinan Tanyolac, [UCSF Diabetes Center](#)), tryptophan metabolism and Huntington disease (Paul Muchowski and Daniel Zwilling, [GIND](#)), and gliomagenesis (Alex Pico, [GICD](#), Ru-Fang Yeh, [CBMB](#), and Margaret Wrensch, [UCSF Dept. of Neurological Surgery](#)).

1.15 Using Delta-MATCH To Identify Species-Specific Transcription Factor Binding Sites Through Comparative Genomics

1.15.1 Background

Delta-MATCH can be used to identify species-specific transcription factor binding sites by comparing MATCH scores between humans and chimpanzees. Sequence were aligned across the human accelerated regions (HARs) defined by Katie Pollard in 2006 [44, 45].

1.15.2 Method

Exactly 126 HARs were investigated with Delta-MATCH to identify human- and chimpanzee-specific transcription factor binding sites. By aligning these genomes at each of these HAR regions, it was found that 696 base positions differed between human and chimpanzee [44, 45]. Each relative polymorphism was submitted through the Delta-MATCH algorithm using the *delta_match.py* script. Results for the human/chimpanzee comparison were curated, prioritized, and organized into a separate Delta-MATCH-HAR database. Exactly 264 HARs had a human or chimpanzee allele with a MATCH score greater than or equal to a TFBS matrix FP threshold score, and 64 of these had potential scores greater than or equal to 0.25. The subset of 11 “nerve system specific” results are detailed (Figure 142 page250).

Figure 142 Human-specific and Chimpanzee-specific Delta-MATCH Predictions

Query Delta-Match Results

Number of matrices searched: 98
There were 20 hits in a database containing 696 SNPs

hit	name	chrom	chromStart	diff_s	threshold	m1	m2	m_per	rank	factor	factor_description	mat_id	qual	mat_len	p1_window	p2_window	a1	a2	db1	chrom1	chromStart1	db2	chrom2	chromStart2
1	HMR65_hu18_chr9_20794799_panTco2_chr9_21289927	chr9	20794799	1	0.985	1.0000	0.9843	1.2	0	STATs	signal transducers and activators of transcription	VSTAT_01	1	9	chr9:20794797-20794805	chr9:21289924-21289931	G	T	hg18	chr9	20794799	panTco2	chr9	21289927
4	HMR178_hu18_chr2_181419988_panTco2_chr2_18592603	chr2	181419988	0.5228	0.898	0.8970	0.9513	5.2	0	PIRX		VSPRX_Q2	1	12	chr2:181419981-181419992	chr2:185926396-185926407	G	A	hg18	chr2	181419988	panTco2	chr2	18592603
1	HMR123_hu18_chr6_113654385_panTco2_chr6_115982715	chr6	113654385	0.5036	0.776	0.7568	0.8888	14.2	0	Pax-6	Pax-6	VSPAX6_01	1	21	chr6:113654386-113654409	chr6:115982716-115982736	A	G	hg18	chr6	113654385	panTco2	chr6	115982715
4	HMR153_hu18_chr6_113654395_panTco2_chr6_115982725	chr6	113654395	0.5036	0.776	0.7568	0.8888	14.2	0	Pax-6	Pax-6	VSPAX6_01	1	21	chr6:113654396-113654409	chr6:115982726-115982736	A	G	hg18	chr6	113654395	panTco2	chr6	115982725
5	HMR123_hu18_chr6_113654401_panTco2_chr6_115982731	chr6	113654401	0.5036	0.776	0.7568	0.8888	14.2	0	Pax-6	Pax-6	VSPAX6_01	1	21	chr6:113654402-113654409	chr6:115982732-115982736	A	G	hg18	chr6	113654401	panTco2	chr6	115982731
4	HMR160_hu18_chr7_26426698_panTco2_chr7_26753007	chr7	26426698	0.4118	0.898	0.6590	0.9400	29.2	0	PIRX		VSPRX_Q2	1	12	chr7:26426698-26426701	chr7:26753009-26753010	A	G	hg18	chr7	26426698	panTco2	chr7	26753007
2	HMR160_hu18_chr7_26426700_panTco2_chr7_26753009	chr7	26426700	0.4118	0.898	0.6590	0.9400	29.2	0	PIRX		VSPRX_Q2	1	12	chr7:26426698-26426701	chr7:26753009-26753010	T	C	hg18	chr7	26426700	panTco2	chr7	26753009
4	HMR29_hu18_chr5_3831333_panTco2_chr5_3855302	chr5	3831333	0.3328	0.898	0.7735	0.9320	17.2	0	PIRX		VSPRX_Q3	1	12	chr5:3831327-3831333	chr5:3855306-3855307	T	C	hg18	chr5	3831333	panTco2	chr5	3855302
2	HMR29_hu18_chr5_3831334_panTco2_chr5_3855303	chr5	3831334	0.3333	0.898	0.7735	0.9320	17.2	0	PIRX		VSPRX_Q3	1	12	chr5:3831327-3831333	chr5:3855306-3855307	A	G	hg18	chr5	3831334	panTco2	chr5	3855303
10	HMR125_hu18_chr13_7129798_panTco2_chr13_71981467	chr13	7129798	0.2084	0.902	0.8475	0.9219	9.2	0	MZF-2		VSMZF2_Q6_01	1	12	chr13:7129781-7129792	chr13:71981461-71981467	A	T	hg18	chr13	7129798	panTco2	chr13	71981467
11	HMR9_hu18_chr20_4095908_panTco2_chr20_40201007	chr20	4095908	0.2526	0.962	0.9793	0.9697	0.2	0	Oct-1	octamer-binding factor 1	VSOCT1_06	1	14	chr20:4095909-4095910	chr20:40201004-40201014	G	T	hg18	chr20	4095908	panTco2	chr20	40201007

There were 11 ‘nerve system specific’ Delta-MATCH predictions. Hyperlinks embedded in the Delta-MATCH-HAR resultant pages exist and link to the position of the HAR sites in the human and chimpanzee genomes.

1.15.3 Results (HAR152/PAX6)

Delta-MATCH predicted PAX6, a transcription factor expressed in the brain will bind to the 3’ UTR of Neurogenin-2 ([NEUROG2](#)) in chimps with modest affinity, but not in humans (

Figure 143 page 252). NEUROG2 is expressed on the reverse strand of chromosome 4 in humans. Three bases differ between the human and chimpanzee at the HAR152 locus, however all three chimpanzee alleles are generally conserved across other vertebrates (Figure page 252). The MATCH scores for the chimpanzee and human alleles for the V\$PAX6_01 TFBS matrix were 0.8888 and 0.7568, respectively (FP threshold = 0.776). The potential score for this [PAX6](#) binding site was 0.5036. These results suggest there may be a moderately strong [PAX6](#) TFBS in HAR152 in most vertebrates but this site may have been uniquely lost in humans.

1.15.4 Discussion

This prediction is interesting for a couple of reasons. Firstly, Neurogenin 2 is expressed in distinct progenitor populations in the central and peripheral nervous systems during mouse neurogenesis, and is essential for the determination of some precursor sensory neurons [46, 47]. And secondly, and perhaps more importantly, the Guillemot lab has proven that Neurogenin 2 is both responsive to and a regulator of PAX6 [48]. To summarize, the interaction between PAX6 and Neurogenin 2 is well documented. Delta-MATCH predicts PAX6 will bind Neurogenin 2 with higher affinity in the chimpanzee and other non-human vertebrates, than in humans. Because PAX6 is known to act a repressor of transcription, it may be that PAX6 has a repressive role in regulating Neurogenin 2 in the chimp, but not in humans where the binding affinity is predicted to be less.

This then begs the question, have novel changes in the human Neurogenin 2 sequence removed a PAX6 binding site, thus altering Neurogenin 2 gene expression in humans relative to chimpanzees? And could this have contributed to differences in brain morphology, function and intelligence between these species? This hypothesis is now under investigation through collaboration with the [Pollard lab](#) at UC Davis, and with the [Guillemot lab](#) at the National institute for Medical Research in the United Kingdom.

1.16 Conclusions for AIM 1

My conclusions for AIM 1 and Delta-MATCH are:

Firstly, 4.5 million human SNPs have been matched against 550 transcription factor binding site matrixes, and these results may be investigated online at <http://deltamatch.org>.

Secondly, Delta-MATCH is extensible and can be used to identify species-specific transcription factor binding sites. Because all of the selectable criteria and orthogonal data in Delta-MATCH are truly independent from each other, it is possible to add new data resources to the tool as they are developed. For example, in the future, it may be possible to add a data set describing the methylation state of a chromosome, and then use it as a selectable criterion during a query.

Thirdly, this resource was built on the premise of finding allele-specific transcription factor binding sites, by creating a differential score reflecting how well two polymorphic alleles pattern match to a 2-dimensional matrix. It may be reasonable to consider substituting the library of definitions to include other known genomic sequence motifs, such as those for microRNA binding sites, or splicing junction sites, and then to calculate a differential score that may identify SNPs that may modulate these other molecular mechanisms.

Chapter 2: A Genetic Survey of Genetic Modulators of HIV-1 Viremia

1.17 Background

The human immunodeficiency virus (HIV-1) is an RNA retrovirus that infects CD4+ target cells coexpressing the coreceptors CCR5 and/or CXCR4. These are primarily thymus-derived lymphocytes (T-cells), but include other cells as well. HIV causes AIDS, the acquired immunodeficiency syndrome.

The HIV/AIDS epidemic has cost an enormous amount in both lives and dollars. It is estimated there are 40 million people infected with HIV worldwide and estimated there were 2.9 million HIV/AIDS related deaths, and 4.3 million new infections in 2006 alone. More than 22 billion dollars will be spent combating the epidemic this year. In January of 2008 during the state of the union address, President Bush suggested investing an additional 30 billion dollars over the next 5 years to combat this disease.

Immediately following HIV exposure during “acute” infection, there is a spike in viral replication, that is controlled to a baseline level during chronic infection by a functioning immune system. Over time, however, as the CD4+ T-cells are depleted, the immune system weakens allowing the virus to escape, and to replicate as the host progresses to an eventual HIV/AIDS-related death.

Interestingly, the level of viremia at baseline may vary from person to person. It is important to recognize that a lower level of HIV viremia at baseline is predictive of a longer survival time, even after the CD4+ lymphocyte count is considered (Figure 145 page 258) [49]. In fact, a small percentage (< 1 %) of HIV-1-infected individuals can

control viremia at or under levels of detection without highly active antiretroviral treatment (HAART). These individuals are called HIV controllers [50]. It is the purpose of AIM 2, to identify the biological reasons for this.

It is important to consider how variant alleles in [CCR5](#), [TLR9](#), [IRF5](#), and [APOE](#) may associate with [HIV-1](#) viremia. TLR9 and IRF5 are components of the innate immune system, and have the job of recognizing foreign invasion, and signaling the appropriate inflammatory response. Viral and bacterial pathogen-associated molecular patterns (PAMPs), such as lipopolysaccharides (LPS) and unmethylated-deoxyoligonucleotides (CpGs), are recognized by toll-like receptors ([TLR4](#) and [TLR9](#), respectively). TLR9 is highly expressed in plasmacytoid dendritic cells in the blood, and when stimulated during viral and bacterial infections, transduces a signal through IRF5, resulting in the nuclear translocation of NF-kappaB, and in the upregulation of interferon- and cytokine-dependant inflammation (Type 1 IFN, TNF) (Figure 151 page 297, and Figure 152 page 297). Evidence suggests that [NF-kB](#)-dependent inflammation may contribute to the progression of HIV/AIDS as it does with other autoimmune/autoinflammatory diseases. Polymorphic variants that markedly elevate inflammatory signaling may increase the risk of developing autoinflammatory diseases and [HIV-1](#) progression [18, 51-54].

Unfortunately, it has been shown that HIV may use the activation of NF-kB, and inflammation, to enhance its own gene expression, and replication (Figure 146 page 258) [53]. In the context of HIV, chronic inflammation may be a risk factor for disease progression. I hypothesized that if strong signaling responses through TLR9 and IRF5 may mediate HIV replication through an NF-kB-dependent mechanism, genetic variants within the TLR9 and IRF5 locus may associate with HIV viremia levels and disease progression. Additionally, because CCR5 is a coreceptor with a known allelic association

with HIV infection, it too should be investigated. Furthermore, if a general and chronic inflammation is associated with multiple different disease phenotypes, the APOE epsilon 4 ($\epsilon 4$) isoform may also associate with the level of HIV viremia during chronic infection.

It will be my conclusion that genetic variation in Toll-like Receptor 9, and Interferon Responsive Factor 5, associates with the level of HIV viremia at baseline. It will also be my conclusion that HIV-infected individuals should avoid high levels of chronic inflammation.

Figure 145 A Lower Baseline Level of HIV Viremia Is Predictive of Longer Survival

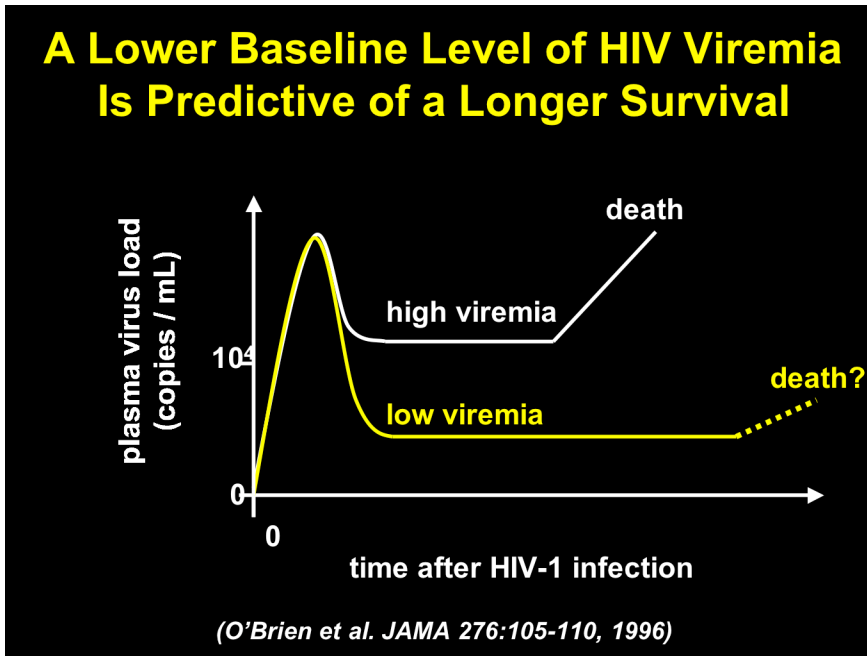
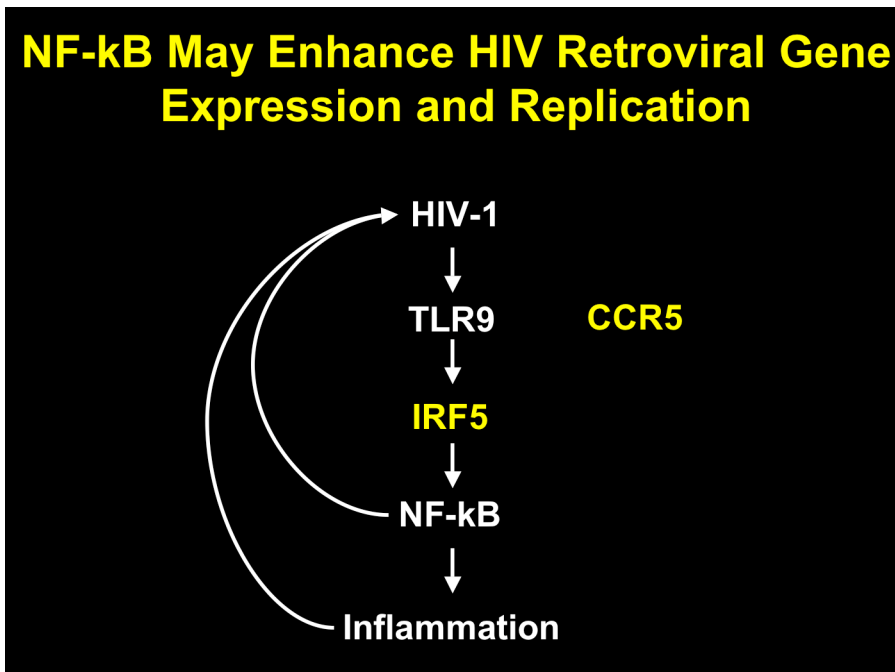


Figure 146 NF- κ B May Enhance HIV Retroviral Gene Expression and Replication



1.18 Methods Genetic Survey

1.18.1 CVL Classification (viremia level)

A cohort of over 1000 Blood and DNA samples was collected from HIV-infected, and non-infected individuals. Samples were contributed by clinicians from San Francisco, Boston, Tennessee, Brazil, Uganda, and Turkey (Table 5 page 261)⁴. These participants provided informed consent, were self-described as 'African-American' (*afam*), 'White-American' (*white*), or 'Hispanic/Latino-American' (*hislat*), and were predominantly male. Genomic DNA was isolated from the thawed blood samples using the Qiagen QIAamp DNA Blood Mini Kit protocol (2003). Participants were categorized according to levels of HIV-1 viremia (CVL status), averaged over multiple time points, according to a phenotypic scheme (version 07/27/05) developed by Steve Deeks and Jeff Martin (Figure 147 page 262)⁵.

Individuals with viremia levels greater than 10,000 copies of HIV-1 RNA per mL of plasma were classified as noncontrollers (group 4 = CVL-4). Individuals who had not received HAART over the previous 12 months while maintaining viremia levels less than 10,000 copies of HIV-1 RNA were classified as controllers. The controllers were further subdivided using lower threshold levels of viremia. Individuals with levels less than 10,000 copies but greater than 2,000 copies were defined as group 3 (CVL-3), individuals with levels less than 2,000 copies but higher than the level of detection (e.g., > 75 copies/mL by bDNA v.3 or > 50 by PCR ultra-sensitive v1.5) were defined as group

⁴ Note: Samples were collected to enrich for the presence of "elite controllers"

⁵ Note: all of the HIV cohort samples were classified by Steve Deeks and Jeff Martin.

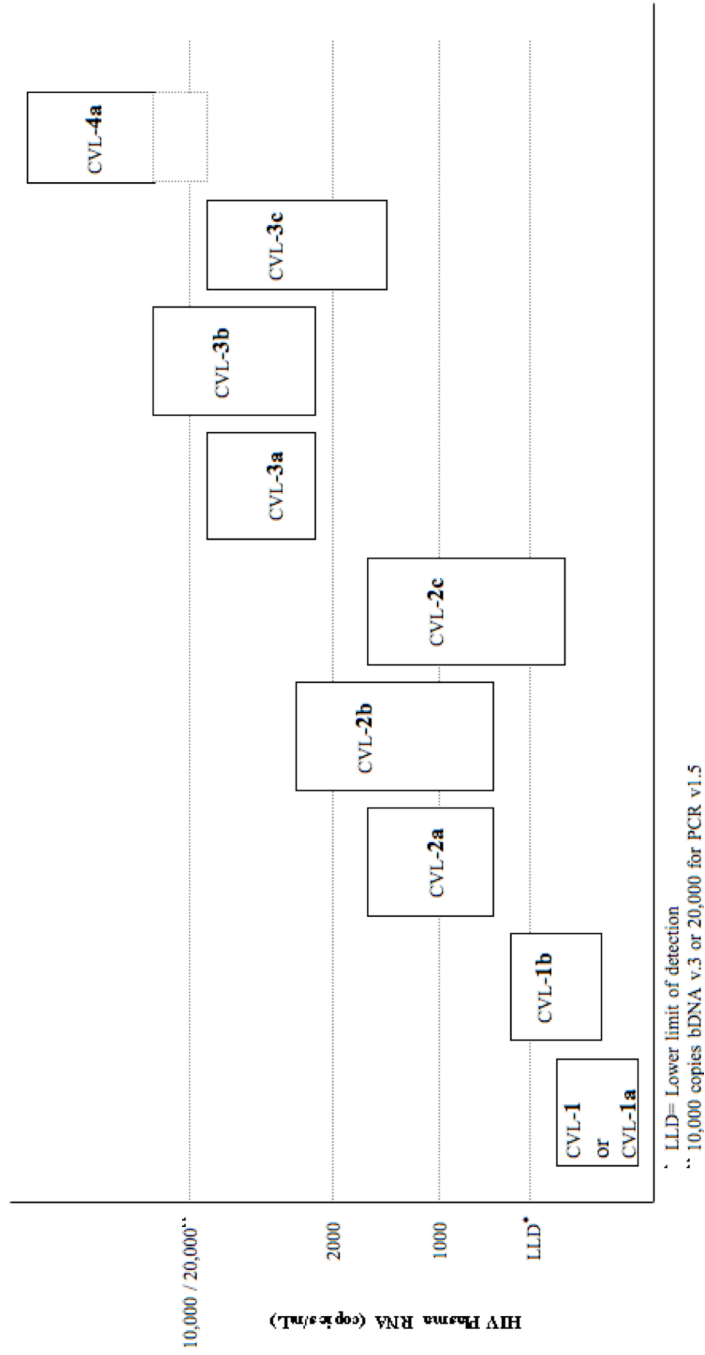
2 (CVL-2), and individuals with undetectable levels of viremia were defined as group 1 (CVL-1) and were termed 'elite controllers'.

Table 5 Cohorts Genotyped for CCR5, TLR9, IRF5 and APOE Polymorphisms.

cohort	cohort_long	contact	institute	city	state	size	description
1	SCOPE cohort	Steve Deeks & Jeff Martin	San Francisco General Hospital	San Francisco	California	528	HIV positive
2	BUCHHINDER San Francisco City cohort	Susan Buchbinder	San Francisco General Hospital	San Francisco	California	35	HIV positive
3	WALKER Walker cohort	Bruce Walker	Massachusetts General Hospital	Boston	Massachusetts	31	HIV positive
4	VANDERBILT Vanderbilt cohort	Spyros Kalans	Partners AIDS Research Center	Nashville	Tennessee	12	HIV positive
5	BANGSBERG Bangsberg cohort	David Bangsberg	Infectious Diseases Division, Federal University of Sao Paulo	San Francisco	California	5	HIV positive
6	SAO PAULO Sao Paulo cohort	Esper Kallias	Infectious Diseases Division, Federal University of Sao Paulo	Sao Paulo (Brazil)	California	82	HIV positive
7	HECHT Hecht cohort	Rick Hecht		San Francisco	California	70	HIV Exposed Seronegative
8	THIS Turkish Heart Study cohort	Robert W. Mahley	The J. David Gladstone Institutes	San Francisco	California	58	Epidemiological survey for cardiovascular disease in the Turkish Population, HIV status unknown
9	WELLS FARGO Wells Fargo cohort	Robert W. Mahley	The J. David Gladstone Institutes	San Francisco	California	178	Epidemiological survey for cardiovascular disease in the San Francisco Population of Wells Fargo Bank Employees, HIV status unknown
10	CONTROLS lab controls	David Williamson	The J. David Gladstone Institutes	San Francisco	California	6	HIV unknown labmate controls
11	AFRICAN CHILDREN AFRICAN CHILDREN	Sunil Emu-Parkhi	San Francisco General Hospital	San Francisco	California	307	Malaria study cohort, HIV unknown

Figure 147 HIV-1 CVL Classification Scheme

**Graphical Depiction of the Virologic Classification Scheme for Chronically Infected Persons
(CVL-xx Series)
Version 07/27/05**



1.18.2 Study Design (a genotype and haplotype analysis of 11 polymorphisms)

I used Taqman allele discrimination and RFLP assays to genotype the HIV-1-infected and uninfected control groups for one polymorphism in [CCR5 \(rs333\)](#), four in [TLR9 \(rs352140, rs352139, rs5743836, rs187084\)](#), four in [IRF5 \(rs2004640, rs2070197, rs10954213, rs2280714\)](#), and two in [APOE \(rs429358, rs7412\)](#). Restriction fragment length polymorphism electrophoresis (RFLP) and Taqman allele discrimination assays were developed (Table 6 page 265, Table 7 page 266) [12, 17-19, 55]. Primers for sequencing and genotyping these loci were designed and ordered (Table 7 page 266). After genotyping these eleven polymorphisms (Figures pages 290, 298, 314 and 333), their frequencies were compared using [HAPLOVIEW](#) v. 3.32 software [56] to identify polymorphic alleles and haplotypes that were significantly associated with HIV-1 viremia levels

Three statistical comparisons were formulated as part of a case and control study design (Figure 148 page 264). Groups 1, 2 and 3 were considered the “cases”, and were comprised with individuals having the ‘lowest’ levels of virus. Group 4 was considered the “controls” and was comprised with individuals having the ‘highest’ levels of virus. It is the purpose of this survey to identify the genes and genetic variations that may be enriched the group 1 (the elite controllers) when compared to the noncontrollers, group 4. Statistical **Test 1** compared the frequency of genotypes and haplotypes in group 4 with the combined groups 1,2, and 3. Statistical **Test 2** compared group 4 to groups 1 and 2. Statistical **Test 3** compared group 4 with only the elite controllers, group 1. (**test1 = CVL-1/2/3 vs. CVL-4; test2 = CVL-1/2 vs. CVL-4; test3 = CVL-1 vs. CVL-4**).

Although I was most interested in identify genetic markers that have different frequencies in groups 1 and 4, I created these three tests so that I could increase the sample size of my to increase the power of my statistical comparisons.

Figure 148 Statistical Tests (chi-squared)

The number of *white* and *afam* individuals in each viremia group are shown.

Statistical Tests (chi-squared)			
Group	HIV RNA copies/mL	Count	
		White	Af.Am.
4	$10,000 \leq x$	202	86
3	$2,000 \leq x < 10,000$	18	16
2	$1,000 \leq x < 2,000$	31	27
1	$x < 1,000$	47	42
Tests			
Tests		White	Af.Am.
Test 1	4 vs 1,2,3	96 / 202	85 / 86
Test 2	4 vs 1,2	78 / 202	69 / 86
Test 3	4 vs 1	47 / 202	42 / 86

Table 6 Genotyping Conditions (TLR9, CCR5, IRF5, APOE)

number	1	2	3	4	5	
Gene	CCR5	TLR9	TLR9	TLR9	TLR9	
rs	rs333	rs352140	rs352139	rs5743836	rs187084	
position hg18.snp126	chr03 463889950	chr03 52231736	chr03 52233411	chr03 52235821	chr03 52236070	
name	ccr5_indel32	tlr9_2848_G>A	tlr9_1174_G>A	tlr9_p1237_T>C	tlr9_p1486_T>C	
taqman genotype kit		C 2301954_20	C 2301953_10		C 2301952_10	
PCR condition	CCR5	TG5840	TG6040	TG5840	TG5840	
p1	CCR5_01	tlr9_06	tlr9_10	tlr9_14	tlr9_01	
p2	CCR5_02	tlr9_07	tlr9_11	tlr9_02	tlr9_02	
amplicon	241	681	105	307	648	
enzyme		BstUI	Avall	ScrFI	MseI	
buffer		2	4	4	2	
temp		60	37	37	37	
BSA		no	no	no	yes	
1	241	365, 316	75, 30	186,75,38,8	332, 153, 83, 80	
2	241, 209	681, 365, 316	105, 75, 30	186, 138, 75, 48, 38, 8	332, 233, 153, 83, 80	
3	209	681	105	138, 75, 48, 38, 8	332, 233, 83	
loading gel	xylene cyanol	bromphenol blue	xylene cyanol	xylene cyanol	xylene cyanol	
agarose gel	3%	2%	3%	3%	3%	
number	6	7	8	9	10	11
Gene	IRF5	IRF5	IRF5	IRF5	APOE	APOE
rs	rs2004640	rs2070197	rs10954213	rs2280714	rs429358	rs7412
position hg18.snp126	chr07 128365536	chr07 128376235	chr07 128376662	chr07 128381961	chr19 50103780	chr19 50103918
name	irf5a_T>G	irf5d_T>C	irf5c_A>G	irf5b_T>C	apoe_Cys112Arg	apoe_Arg158Cys
taqman genotype kit	C 9491614_10	C 2691236_10		C 2691243_1	C 3084793_20	C 904973_10

Table 7 PCR Conditions for Genotyping (RFLP)

TG5840	temp C	time
1	95	10 min
2	94	45 sec
3	58	45 sec
4	72	45 sec
5	repeat 39x to step 2	
6	end	
TG6040	temp C	time
1	95	10 min
2	94	45 sec
3	60	45 sec
4	72	45 sec
5	repeat 39x to step 2	
6	end	
CCR5	temp C	time
1	94	3 min
2	94	75 s
3	58	60 s
4	72	60 s
5	repeat 34x to step 2	
6	end	

Table 8 PCR Primer Sequences

1	name	order sequece	Direction
2	ccr5_01	TCAAAAAGAAGGTCTTCATTACACC	F
3	ccr5_02	AGCCCGAGAAGAGAAAATAACAATC	R
4	tlr9_01	GCCATGATACCACCCAGAGT	F
5	tlr9_02	TCAAAGCCACAGTCCACAGA	R
6	tlr9_03	GTGGATGGAGGGAATGAATG	F
7	tlr9_05	ACTCTGGGTGGTATCATGGC	R
8	tlr9_06	AGATGGAGGGGAGAAGGTCT	F
9	tlr9_07	AAGGCCAGGTAATTGTCACG	R
10	tlr9_08	CCCTGTTGAGAGGGTGACAT	F
11	tlr9_10	AGGGCTGTGTGAGTGGCCGGCCCCAGGTC	R
12	tlr9_11	CTTCTGCAGGTAGGGCTTGGAGAGAGG	F
13	tlr9_12	CACACACCTGGCCTCTAGGA	F
14	tlr9_13	CTTCCCAGGATATCCCCTTC	R
15	tlr9_14	CATAGACCAGGCAAGGAGC	F
16	tlr9_21	ACCTTTGGCCACAAGAAGTG	F
17	tlr9_22	CACAGTGTGGCAAAAACGAC	R
18	tlr9_23	TCCCATGGCCTTTTGTAGTC	F
19	tlr9_24	ACCTGGGAGCCAATGTTTC	R
20	tlr9_25	GCTACTGAGTGGGCACTGCT	F
21	tlr9_26	CCTGCTTGCAGTTGACTGTG	R
22	tlr9_28	AGGCACCATCTCCAGAGTTC	R
23	tlr9_29	TGTGGAGGAGGAGGTCTTGT	F
24	tlr9_30	ACAACCCGTCACTGTTGCTT	R
25	tlr9_31	CCAGACCCTCTGGAGAAGC	F
26	tlr9_32	TTCAGGCAGGTGTGAGAGTC	R
27	tlr9_33	TCTGACCCATAAGGCAAAGG	F
28	tlr9_34	CAGGTGGGCAAAGTCAGAAT	R
29	tlr9_35	AACTGCAACTGGCTGTTCTT	F
30	tlr9_36	CTCACTCAGGTCCAGCACTC	R
31	tlr9_37	TCACTTCCCCAGCTACATC	F
32	tlr9_38	TCGTGGTAGAGGTCCAGCTT	R
33	tlr9_39	TTCACCTTGGATCTGTCACG	F
34	tlr9_40	TAGTATTTGCAGGGCACTCG	R
35	tlr9_41	ATGTCAGCTGCAACAGCATC	F
36	tlr9_42	GTCAGGGCTCAGGATCACC	R
37	tlr9_43	CTGGCAAAACCCTCTTTGAG	F
38	tlr9_44	GTAGGCGAGGGAGACAGACA	R
39	tlr9_45	CTGGATGTGGTCTTGGTCCT	F
40	tlr9_46	CTGAGACCAGCCTAGGCAAC	R
41	tlr9_47	ATGGGGACGGTGGGCTGTGGG	F
42	tlr9_48	GGGGCTCCTAGAGGCCAGGTG	R
43	tlr9_49	CCTGAGGCAGGAGAATGCCC	F
44	tlr9_50	CCCCTAGAGGGAAGTGTCA	R
45	tlr9_51	TTAAACGCGTACTTGTGCTTGGCCCTGAGA	F
46	tlr9_52	CAAGAAGCTTTCTCCACATTCAGAGCC	R
47	tlr9_53	GAGACGGAGTTTCGCTCTTGT	F
48	tlr9_54	GGGCTTCGGCTCTGAAGTCTTC	R
49	tlr9_55	CCTGAGCAAGGTGGCAGCTGGC	R
50	tlr9_56	GGAGGTCTTGTTCCGGAAGA	F
51	tlr9_57	TGGTGAAGTCAACTGGCTGTTT	F
52	tlr9_58	AGGGCGACCCCTTGTAGTCTG	F
53	tlr9_59	ATGGGTTTCTGCCGACGCGCCCTG	F
54	tlr9_60	ATGGCAGCACCCCGTGGCAATG	F
55	tlr9_61	ATGCTCTACTCCAGCTGCAAGAG	F
56	tlr9_62	TTGGCCCGTGGGTCCTGGC	R
57	tlr9_63	TLR9_63_p1237_T	F
58	tlr9_64	TLR9_64_p1237_C	F
59	tlr9_65	TLR9_65_p1237_T	R
60	tlr9_66	TLR9_66_p1237_C	R

1.18.3 Results (genotype data)

The genotype counts, genotype frequencies, allele frequencies, number of samples per cohort, and chi-square permuted p-value significances (10,000 permutations) for these polymorphisms are shown (Tables pages 268 - 288). These tables describe significant allele associations with HIV-1 viremia levels after sub-stratification by ethnicity.

Haplotypes for [TLR9](#), [IRF5](#), and [APOE](#) were constructed and similarly associated with HIV-1 viremia levels. Significant haplotype associations are described in Tables (pages 305, 306, 307 ([TLR9](#)); pages 325, 326, 327 ([IRF5](#)); and pages 336, 337, 338 ([APOE](#)).

Significances ($p \leq 0.05$) are highlighted in pink and trends ($0.05 < p \leq 0.10$) highlighted in light blue.

Table 9 Genotype Counts Test1

					CCR5_del_32	2848_A>G_hiv03a	1174_G>A_hiv03a	p1237_T>C_hiv03a	p1486_T>C_hiv03a	IRF5a_T>G	IRF5d_T>C	IRF5c_A>G	IRF5b_T>C	apoe_T>C_Cys112Arg_snp2	apoe_C>T_Arg158Cys_snp1c
					rs333	rs352140	rs352139	rs5743836	rs187084	rs2004640	rs2070197	rs10954213	rs2280714	rs429358	rs7412
TEST1	CVL 4	AF_AM	CONTOL	1	86	3	30	30	46	25	80	23	31	55	67
		AF_AM	CONTOL	2	0	44	41	39	33	32	2	40	37	23	12
		AF_AM	CONTOL	3	0	39	15	17	6	24	0	18	14	4	3
	CVL 1/2/3	AF_AM	CASE	1	80	11	32	36	44	8	74	16	20	46	58
		AF_AM	CASE	2	5	36	35	37	29	38	2	41	43	27	17
		AF_AM	CASE	3	0	38	18	12	12	27	0	19	13	3	0
			p		0.025	NS	NS	NS	NS	0.025	NS	NS	NS	NS	NS
		3x2	chi		5.211					9.057					
TEST1	CVL 4	WHITE	CONTOL	1	154	56	54	148	80	43	142	65	79	136	148
		WHITE	CONTOL	2	48	99	100	49	88	97	50	101	95	52	41
		WHITE	CONTOL	3	0	46	48	5	33	53	1	23	18	5	4
	CVL 1/2/3	WHITE	CASE	1	70	38	40	59	26	27	75	38	45	62	83
		WHITE	CASE	2	26	46	46	32	49	46	14	44	40	27	10
		WHITE	CASE	3	0	12	9	4	21	18	2	10	7	3	0
			p		NS	0.05	0.01	NS	NS	NS	NS	NS	NS	NS	0.05
		3x2	chi			6.433	11.71								7.027
TEST1	CVL 4	HIS_LAT	CONTOL	1	28	11	13	23	14	8	27	9	9	24	29
		HIS_LAT	CONTOL	2	5	14	13	10	10	14	3	16	17	9	4
		HIS_LAT	CONTOL	3	0	8	7	0	9	11	3	7	7	0	0
	CVL 1/2/3	HIS_LAT	CASE	1	15	3	3	16	7	5	15	7	8	13	15
		HIS_LAT	CASE	2	3	9	9	2	9	8	2	8	7	2	2
		HIS_LAT	CASE	3	0	6	6	0	2	4	0	2	2	2	1
			p		NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
		3x2	chi												

Table 10 Genotype Counts Test2

					rs333	CCR5_del_32	rs352140	2848_A>G_hiv03a	rs352139	1174_G>A_hiv03a	rs5743836	p1237_T>C_hiv03a	rs187084	p1486_T>C_hiv03a	rs2004640	IRF5a_T>G	rs2070197	IRF5d_T>C	rs10954213	IRF5c_A>G	rs2280714	IRF5b_T>C	rs429358	apoe_T>C_Cys112Arg_snp2	rs7412	apoe_C>T_Arg158Cys_snp1c
TEST2	CVL 4	AF_AM	CONTOL	1	86	3	30	30	46	25	80	23	31	55	67											
		AF_AM	CONTOL	2	0	44	41	39	33	32	2	40	37	23	12											
		AF_AM	CONTOL	3	0	39	15	17	6	24	0	18	14	4	3											
	CVL 1/2	AF_AM	CASE	1	65	11	26	30	35	7	61	14	16	37	48											
		AF_AM	CASE	2	4	27	28	29	24	32	2	32	36	24	14											
		AF_AM	CASE	3	0	31	15	10	10	22	0	17	11	2	0											
			p		0.025	0.025	NS	NS	NS	0.025	NS	NS	NS	NS	NS											
		3x2	chi		5.118	7.785				7.545																
TEST2	CVL 4	WHITE	CONTOL	1	154	56	54	148	80	43	142	65	79	136	148											
		WHITE	CONTOL	2	48	99	100	49	88	97	50	101	95	52	41											
		WHITE	CONTOL	3	0	46	48	5	33	53	1	23	18	5	4											
	CVL 1/2	WHITE	CASE	1	59	29	30	47	25	21	61	29	34	48	69											
		WHITE	CASE	2	19	39	39	26	38	38	11	37	35	24	6											
		WHITE	CASE	3	0	10	8	4	15	15	2	9	6	3	0											
			p		NS	NS	0.025	NS	NS	NS	NS	NS	NS	NS	0.025											
		3x2	chi				7.751								8.14											
TEST2	CVL 4	HIS_LAT	CONTOL	1	28	11	13	23	14	8	27	9	9	24	29											
		HIS_LAT	CONTOL	2	5	14	13	10	10	14	3	16	17	9	4											
		HIS_LAT	CONTOL	3	0	8	7	0	9	11	3	7	7	0	0											
	CVL 1/2	HIS_LAT	CASE	1	15	3	3	16	7	5	15	7	8	13	15											
		HIS_LAT	CASE	2	3	9	9	2	9	8	2	8	7	2	2											
		HIS_LAT	CASE	3	0	6	6	0	2	4	0	2	2	1												
			p		NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS											
		3x2	chi																							

Table 11 Genotype Counts Test3

					CCR5_del_32	2848_A>G_hiv03a	1174_G>A_hiv03a	p1237_T>C_hiv03a	p1486_T>C_hiv03a	IRF5a_T>G	IRF5d_T>C	IRF5c_A>G	IRF5b_T>C	apoe_T>C_Cys112Arg_snp2	apoe_C>T_Arg158Cys_snp1c
					rs333	rs352140	rs352139	rs5743836	rs187084	rs2004640	rs2070197	rs10954213	rs2280714	rs429358	rs7412
TEST3	CVL 4	AF_AM	CONTOL	1	86	3	30	30	46	25	80	23	31	55	67
		AF_AM	CONTOL	2	0	44	41	39	33	32	2	40	37	23	12
		AF_AM	CONTOL	3	0	39	15	17	6	24	0	18	14	4	3
	CVL 1	AF_AM	CASE	1	39	7	13	19	24	4	39	9	10	25	33
		AF_AM	CASE	2	3	14	18	18	14	24	2	24	26	15	8
		AF_AM	CASE	3	0	21	11	5	4	11	0	8	5	1	0
			p		0.025	0.025	NS	NS	NS	0.025	NS	NS	NS	NS	NS
		3x2	chi		6.29	8.383				7.383					
TEST3	CVL 4	WHITE	CONTOL	1	154	56	54	148	80	43	142	65	79	136	148
		WHITE	CONTOL	2	48	99	100	49	88	97	50	101	95	52	41
		WHITE	CONTOL	3	0	46	48	5	33	53	1	23	18	5	4
	CVL 1	WHITE	CASE	1	37	15	16	26	16	16	38	22	24	32	40
		WHITE	CASE	2	10	25	25	17	25	21	6	21	20	13	5
		WHITE	CASE	3	0	7	5	3	6	9	2	4	3	2	0
			p		NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
		3x2	chi												
TEST3	CVL 4	HIS_LAT	CONTOL	1	28	11	13	23	14	8	27	9	9	24	29
		HIS_LAT	CONTOL	2	5	14	13	10	10	14	3	16	17	9	4
		HIS_LAT	CONTOL	3	0	8	7	0	9	11	3	7	7	0	0
	CVL 1	HIS_LAT	CASE	1	10	1	1	10	5	3	9	5	5	6	9
		HIS_LAT	CASE	2	1	6	6	1	5	5	1	4	4	2	2
		HIS_LAT	CASE	3	0	4	4	0	1	2	0	1	1	2	0
			p		NS	NS	NS	NS	NS	NS	NS	NS	NS	0.05	NS
		3x2	chi											6.72	

Table 12 Genotype Counts of Other Non-HIV Positive Populations

				rs333	CCR5_del_32	rs352140	2848_A>G_hiv03a	rs352139	1174_G>A_hiv03a	rs5743836	p1237_T>C_hiv03a	rs187084	p1486_T>C_hiv03a	rs2004640	IRF5a_T>G	rs2070197	IRF5d_T>C	rs10954213	IRF5c_A>G	rs2280714	IRF5b_T>C	rs429358	apoe_T>C_Cys112Arg_snp2	rs7412	apoe_C>T_Arg158Cys_snp1c	
		TURK	1	64	20	23	62	23															23	23		
		TURK	2	3	38	34	13	34															34	34		
		TURK	3	1	20	21	2	11															11	11		
		SF_WHITE	1	119	52	54	116	55								113	61	67					55	55		
		SF_WHITE	2	36	68	69	38	73								35	76	71					73	73		
		SF_WHITE	3	3	37	35	4	30								3	14	13					30	30		
		SF_AF_AM	1	48	9	17	23	20								31	7	9					20	20		
		SF_AF_AM	2	1	15	19	19	21								1	18	18					21	21		
		SF_AF_AM	3	0	25	12	6	4								0	7	5					4	4		
		AFRICAN	1		23	106	121	167															167	167		
		AFRICAN	2		137	151	132	106															106	106		
		AFRICAN	3		144	48	51	29															29	29		

Table 13 Genotype Frequencies Test1

					rs333	rs352140	rs352139	rs5743836	rs187084	rs2004640	rs2070197	rs10954213	rs2280714	rs429358	rs7412
					CCR5_del_32	2848_A>G_hiv03a	1174_G>A_hiv03a	p1237_T>C_hiv03a	p1486_T>C_hiv03a	IRF5a_T>G	IRF5d_T>C	IRF5c_A>G	IRF5b_T>C	apoe_T>C_Cys112Arg_snp2	apoe_C>T_Arg158Cys_snp1c
TEST1	CVL 4	AF_AM	CONTOL	1	100.0	3.5	34.9	34.9	54.1	30.9	97.6	28.4	37.8	67.1	81.7
		AF_AM	CONTOL	2	0.0	51.2	47.7	45.3	38.8	39.5	2.4	49.4	45.1	28.0	14.6
		AF_AM	CONTOL	3	0.0	45.3	17.4	19.8	7.1	29.6	0.0	22.2	17.1	4.9	3.7
	CVL 1/2/3	AF_AM	CASE	1	94.1	12.9	37.6	42.4	51.8	11.0	97.4	21.1	26.3	60.5	77.3
		AF_AM	CASE	2	5.9	42.4	41.2	43.5	34.1	52.1	2.6	53.9	56.6	35.5	22.7
		AF_AM	CASE	3	0.0	44.7	21.2	14.1	14.1	37.0	0.0	25.0	17.1	3.9	0.0
TEST1	CVL 4	WHITE	CONTOL	1	76.2	27.9	26.7	73.3	39.8	22.3	73.6	34.4	41.1	70.5	76.7
		WHITE	CONTOL	2	23.8	49.3	49.5	24.3	43.8	50.3	25.9	53.4	49.5	26.9	21.2
		WHITE	CONTOL	3	0.0	22.9	23.8	2.5	16.4	27.5	0.5	12.2	9.4	2.6	2.1
	CVL 1/2/3	WHITE	CASE	1	72.9	39.6	42.1	62.1	27.1	29.7	82.4	41.3	48.9	67.4	89.2
		WHITE	CASE	2	27.1	47.9	48.4	33.7	51.0	50.5	15.4	47.8	43.5	29.3	10.8
		WHITE	CASE	3	0.0	12.5	9.5	4.2	21.9	19.8	2.2	10.9	7.6	3.3	0.0
TEST1	CVL 4	HIS_LAT	CONTOL	1	84.8	33.3	39.4	69.7	42.4	24.2	81.8	28.1	27.3	72.7	87.9
		HIS_LAT	CONTOL	2	15.2	42.4	39.4	30.3	30.3	42.4	9.1	50.0	51.5	27.3	12.1
		HIS_LAT	CONTOL	3	0.0	24.2	21.2	0.0	27.3	33.3	9.1	21.9	21.2	0.0	0.0
	CVL 1/2/3	HIS_LAT	CASE	1	83.3	16.7	16.7	88.9	38.9	29.4	88.2	41.2	47.1	76.5	83.3
		HIS_LAT	CASE	2	16.7	50.0	50.0	11.1	50.0	47.1	11.8	47.1	41.2	11.8	11.1
		HIS_LAT	CASE	3	0.0	33.3	33.3	0.0	11.1	23.5	0.0	11.8	11.8	11.8	5.6

Table 14 Genotype Frequencies Test2

					rs333	rs352140	rs352139	rs5743836	rs187084	rs2004640	rs2070197	rs10954213	rs2280714	rs429358	rs7412
					CCR5_del_32	2848_A>G_hiv03a	1174_G>A_hiv03a	p1237_T>C_hiv03a	p1486_T>C_hiv03a	IRF5a_T>G	IRF5d_T>C	IRF5c_A>G	IRF5b_T>C	apoe_T>C_Cys112Arg_snp2	apoe_C>T_Arg158Cys_snp1c
TEST2	CVL 4	AF_AM	CONTOL	1	100.0	3.5	34.9	34.9	54.1	30.9	97.6	28.4	37.8	67.1	81.7
		AF_AM	CONTOL	2	0.0	51.2	47.7	45.3	38.8	39.5	2.4	49.4	45.1	28.0	14.6
		AF_AM	CONTOL	3	0.0	45.3	17.4	19.8	7.1	29.6	0.0	22.2	17.1	4.9	3.7
	CVL 1/2	AF_AM	CASE	1	94.2	15.9	37.7	43.5	50.7	11.5	96.8	22.2	25.4	58.7	77.4
		AF_AM	CASE	2	5.8	39.1	40.6	42.0	34.8	52.5	3.2	50.8	57.1	38.1	22.6
		AF_AM	CASE	3	0.0	44.9	21.7	14.5	14.5	36.1	0.0	27.0	17.5	3.2	0.0
TEST2	CVL 4	WHITE	CONTOL	1	76.2	27.9	26.7	73.3	39.8	22.3	73.6	34.4	41.1	70.5	76.7
		WHITE	CONTOL	2	23.8	49.3	49.5	24.3	43.8	50.3	25.9	53.4	49.5	26.9	21.2
		WHITE	CONTOL	3	0.0	22.9	23.8	2.5	16.4	27.5	0.5	12.2	9.4	2.6	2.1
	CVL 1/2	WHITE	CASE	1	75.6	37.2	39.0	61.0	32.1	28.4	82.4	38.7	45.3	64.0	92.0
		WHITE	CASE	2	24.4	50.0	50.6	33.8	48.7	51.4	14.9	49.3	46.7	32.0	8.0
		WHITE	CASE	3	0.0	12.8	10.4	5.2	19.2	20.3	2.7	12.0	8.0	4.0	0.0
TEST2	CVL 4	HIS_LAT	CONTOL	1	84.8	33.3	39.4	69.7	42.4	24.2	81.8	28.1	27.3	72.7	87.9
		HIS_LAT	CONTOL	2	15.2	42.4	39.4	30.3	30.3	42.4	9.1	50.0	51.5	27.3	12.1
		HIS_LAT	CONTOL	3	0.0	24.2	21.2	0.0	27.3	33.3	9.1	21.9	21.2	0.0	0.0
	CVL 1/2	HIS_LAT	CASE	1	83.3	16.7	16.7	88.9	38.9	29.4	88.2	41.2	47.1	76.5	83.3
		HIS_LAT	CASE	2	16.7	50.0	50.0	11.1	50.0	47.1	11.8	47.1	41.2	11.8	11.1
		HIS_LAT	CASE	3	0.0	33.3	33.3	0.0	11.1	23.5	0.0	11.8	11.8	11.8	5.6

Table 15 Genotype Frequencies Test3

					rs333	rs352140	rs352139	rs5743836	rs187084	rs2004640	rs2070197	rs10954213	rs2280714	rs429358	rs7412
					CCR5_del_32	2848_A>G_hiv03a	1174_G>A_hiv03a	p1237_T>C_hiv03a	p1486_T>C_hiv03a	IRF5a_T>G	IRF5d_T>C	IRF5c_A>G	IRF5b_T>C	apoe_T>C_Cys112Arg_snp2	apoe_C>T_Arg158Cys_snp1c
TEST3	CVL 4	AF_AM	CONTOL	1	100.0	3.5	34.9	34.9	54.1	30.9	97.6	28.4	37.8	67.1	81.7
		AF_AM	CONTOL	2	0.0	51.2	47.7	45.3	38.8	39.5	2.4	49.4	45.1	28.0	14.6
		AF_AM	CONTOL	3	0.0	45.3	17.4	19.8	7.1	29.6	0.0	22.2	17.1	4.9	3.7
	CVL 1	AF_AM	CASE	1	92.9	16.7	31.0	45.2	57.1	10.3	95.1	22.0	24.4	61.0	80.5
		AF_AM	CASE	2	7.1	33.3	42.9	42.9	33.3	61.5	4.9	58.5	63.4	36.6	19.5
		AF_AM	CASE	3	0.0	50.0	26.2	11.9	9.5	28.2	0.0	19.5	12.2	2.4	0.0
TEST3	CVL 4	WHITE	CONTOL	1	76.2	27.9	26.7	73.3	39.8	22.3	73.6	34.4	41.1	70.5	76.7
		WHITE	CONTOL	2	23.8	49.3	49.5	24.3	43.8	50.3	25.9	53.4	49.5	26.9	21.2
		WHITE	CONTOL	3	0.0	22.9	23.8	2.5	16.4	27.5	0.5	12.2	9.4	2.6	2.1
	CVL 1	WHITE	CASE	1	78.7	31.9	34.8	56.5	34.0	34.8	82.6	46.8	51.1	68.1	88.9
		WHITE	CASE	2	21.3	53.2	54.3	37.0	53.2	45.7	13.0	44.7	42.6	27.7	11.1
		WHITE	CASE	3	0.0	14.9	10.9	6.5	12.8	19.6	4.3	8.5	6.4	4.3	0.0
TEST3	CVL 4	HIS_LAT	CONTOL	1	84.8	33.3	39.4	69.7	42.4	24.2	81.8	28.1	27.3	72.7	87.9
		HIS_LAT	CONTOL	2	15.2	42.4	39.4	30.3	30.3	42.4	9.1	50.0	51.5	27.3	12.1
		HIS_LAT	CONTOL	3	0.0	24.2	21.2	0.0	27.3	33.3	9.1	21.9	21.2	0.0	0.0
	CVL 1	HIS_LAT	CASE	1	90.9	9.1	9.1	90.9	45.5	30.0	90.0	50.0	50.0	60.0	81.8
		HIS_LAT	CASE	2	9.1	54.5	54.5	9.1	45.5	50.0	10.0	40.0	40.0	20.0	18.2
		HIS_LAT	CASE	3	0.0	36.4	36.4	0.0	9.1	20.0	0.0	10.0	10.0	20.0	0.0

Table 16 Genotype Frequencies of Other Non-HIV Positive Populations

					rs333	rs352140	rs352139	rs5743836	rs187084	rs2004640	rs2070197	rs10954213	rs2280714	rs429358	rs7412				
					CCR5_del_32	2848_A>G_hiv03a	1174_G>A_hiv03a	p1237_T>C_hiv03a	p1486_T>C_hiv03a	IRF5a_T>G	IRF5d_T>C	IRF5c_A>G	IRF5b_T>C	apoe_T>C_Cys112Arg_snp2	apoe_C>T_Arg158Cys_snp1c				
		TURK	1		94.1	25.6	29.5	80.5	33.8									33.8	33.8
		TURK	2		4.4	48.7	43.6	16.9	50.0									50.0	50.0
		TURK	3		1.5	25.6	26.9	2.6	16.2									16.2	16.2
		SF_WHITE	1		75.3	33.1	34.2	73.4	34.8		74.8	40.4	44.4	34.8	34.8				
		SF_WHITE	2		22.8	43.3	43.7	24.1	46.2		23.2	50.3	47.0	46.2	46.2				
		SF_WHITE	3		1.9	23.6	22.2	2.5	19.0		2.0	9.3	8.6	19.0	19.0				
		SF_AF_AM	1		98.0	18.4	35.4	47.9	44.4		96.9	21.9	28.1	44.4	44.4				
		SF_AF_AM	2		2.0	30.6	39.6	39.6	46.7		3.1	56.3	56.3	46.7	46.7				
		SF_AF_AM	3		0.0	51.0	25.0	12.5	8.9		0.0	21.9	15.6	8.9	8.9				
		AFRICAN	1			7.6	34.8	39.8	55.3									55.3	55.3
		AFRICAN	2			45.1	49.5	43.4	35.1									35.1	35.1
		AFRICAN	3			47.4	15.7	16.8	9.6									9.6	9.6

Table 17 Allele Frequencies Test1

					rs333	rs352140	rs352139	rs5743836	rs187084	rs2004640	rs2070197	rs10954213	rs2280714	rs429358	rs7412
					CCR5_del_32	2848_A>G_hiv03a	1174_G>A_hiv03a	p1237_T>C_hiv03a	p1486_T>C_hiv03a	IRF5a_T>G	IRF5d_T>C	IRF5c_A>G	IRF5b_T>C	apoe_T>C_Cys112Arg_snp2	apoe_C>T_Arg158Cys_snp1c
TEST1	CVL 4	AF_AM	CONTOL	MAJOR	100.0	29.1	58.7	57.6	73.5	50.6	98.8	53.1	60.4	81.1	89.0
		AF_AM	CONTOL												
		AF_AM	CONTOL	MINOR	0.0	70.9	41.3	42.4	26.5	49.4	1.2	46.9	39.6	18.9	11.0
	CVL 1/2/3	AF_AM	CASE	MAJOR	97.1	34.1	58.2	64.1	68.8	37.0	98.7	48.0	54.6	78.3	88.7
		AF_AM	CASE												
		AF_AM	CASE	MINOR	2.9	65.9	41.8	35.9	31.2	63.0	1.3	52.0	45.4	21.7	11.3
				p	0.025	NS	NS	NS	NS	0.025	NS	NS	NS	NS	NS
				%dif	100.0	-7.1	1.2	-15.5	15.1	21.6	7.3	9.7	12.7	12.9	3.2
TEST1	CVL 4	WHITE	CONTOL	MAJOR	88.1	52.5	51.5	85.4	61.7	47.4	86.5	61.1	65.9	83.9	87.3
		WHITE	CONTOL												
		WHITE	CONTOL	MINOR	11.9	47.5	48.5	14.6	38.3	52.6	13.5	38.9	34.1	16.1	12.7
	CVL 1/2/3	WHITE	CASE	MAJOR	86.5	63.5	66.3	78.9	52.6	54.9	90.1	65.2	70.7	82.1	94.6
		WHITE	CASE												
		WHITE	CASE	MINOR	13.5	36.5	33.7	21.1	47.4	45.1	9.9	34.8	29.3	17.9	5.4
				p	NS	0.05	0.01	NS	NS	NS	NS	NS	NS	NS	0.05
				%dif	12.3	-23.3	-30.6	30.6	19.2	-14.3	-26.6	-10.6	-14.0	10.4	-57.6
TEST1	CVL 4	HIS_LAT	CONTOL	MAJOR	92.4	54.5	59.1	84.8	57.6	45.5	86.4	53.1	53.0	86.4	93.9
		HIS_LAT	CONTOL												
		HIS_LAT	CONTOL	MINOR	7.6	45.5	40.9	15.2	42.4	54.5	13.6	46.9	47.0	13.6	6.1
	CVL 1/2/3	HIS_LAT	CASE	MAJOR	91.7	41.7	41.7	94.4	63.9	52.9	94.1	64.7	67.6	82.4	88.9
		HIS_LAT	CASE												
		HIS_LAT	CASE	MINOR	8.3	58.3	58.3	5.6	36.1	47.1	5.9	35.3	32.4	17.6	11.1
				p	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
				%dif	9.1	22.1	29.9	-63.3	-14.9	-13.7	-56.9	-24.7	-31.1	22.7	45.5

Table 18 Allele Frequencies Test2

					rs333	rs352140	rs352139	rs5743836	rs187084	rs2004640	rs2070197	rs10954213	rs2280714	rs429358	rs7412
					CCR5_del_32	2848_A>G_hiv03a	1174_G>A_hiv03a	p1237_T>C_hiv03a	p1486_T>C_hiv03a	IRF5a_T>G	IRF5d_T>C	IRF5c_A>G	IRF5b_T>C	apoe_T>C_Cys112Arg_snp2	apoe_C>T_Arg158Cys_snp1c
TEST2	CVL 4	AF_AM	CONTOL	MAJOR	100.0	29.1	58.7	57.6	73.5	50.6	98.8	53.1	60.4	81.1	89.0
		AF_AM	CONTOL												
		AF_AM	CONTOL	MINOR	0.0	70.9	41.3	42.4	26.5	49.4	1.2	46.9	39.6	18.9	11.0
	CVL 1/2	AF_AM	CASE	MAJOR	97.1	35.5	58.0	64.5	68.1	37.7	98.4	47.6	54.0	77.8	88.7
		AF_AM	CASE												
		AF_AM	CASE	MINOR	2.9	64.5	42.0	35.5	31.9	62.3	1.6	52.4	46.0	22.2	11.3
				p	0.025	0.025	NS	NS	NS	0.025	NS	NS	NS	NS	NS
				%dif	100.0	-9.1	1.8	-16.3	17.0	20.7	23.2	10.4	13.9	14.9	2.8
TEST2	CVL 4	WHITE	CONTOL	MAJOR	88.1	52.5	51.5	85.4	61.7	47.4	86.5	61.1	65.9	83.9	87.3
		WHITE	CONTOL												
		WHITE	CONTOL	MINOR	11.9	47.5	48.5	14.6	38.3	52.6	13.5	38.9	34.1	16.1	12.7
	CVL 1/2	WHITE	CASE	MAJOR	87.8	62.2	64.3	77.9	56.4	54.1	89.9	63.3	68.7	80.0	96.0
		WHITE	CASE												
		WHITE	CASE	MINOR	12.2	37.8	35.7	22.1	43.6	45.9	10.1	36.7	31.3	20.0	4.0
				p	NS	NS	0.025	NS	NS	NS	NS	NS	NS	NS	0.025
				%dif	2.4	-20.4	-26.4	33.9	12.1	-12.6	-24.8	-5.7	-8.2	19.7	-68.5
TEST2	CVL 4	HIS_LAT	CONTOL	MAJOR	92.4	54.5	59.1	84.8	57.6	45.5	86.4	53.1	53.0	86.4	93.9
		HIS_LAT	CONTOL												
		HIS_LAT	CONTOL	MINOR	7.6	45.5	40.9	15.2	42.4	54.5	13.6	46.9	47.0	13.6	6.1
	CVL 1/2	HIS_LAT	CASE	MAJOR	91.7	41.7	41.7	94.4	63.9	52.9	94.1	64.7	67.6	82.4	88.9
		HIS_LAT	CASE												
		HIS_LAT	CASE	MINOR	8.3	58.3	58.3	5.6	36.1	47.1	5.9	35.3	32.4	17.6	11.1
				p	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
				%dif	9.1	22.1	29.9	-63.3	-14.9	-13.7	-56.9	-24.7	-31.1	22.7	45.5

Table 19 Allele Frequencies Test3

					rs333	rs352140	rs352139	rs5743836	rs187084	rs2004640	rs2070197	rs10954213	rs2280714	rs429358	rs7412
					CCR5_del_32	2848_A>G_hiv03a	1174_G>A_hiv03a	p1237_T>C_hiv03a	p1486_T>C_hiv03a	IRF5a_T>G	IRF5d_T>C	IRF5c_A>G	IRF5b_T>C	apoe_T>C_Cys112Arg_snp2	apoe_C>T_Arg158Cys_snp1c
TEST3	CVL 4	AF_AM	CONTOL	MAJOR	100.0	29.1	58.7	57.6	73.5	50.6	98.8	53.1	60.4	81.1	89.0
		AF_AM	CONTOL												
		AF_AM	CONTOL	MINOR	0.0	70.9	41.3	42.4	26.5	49.4	1.2	46.9	39.6	18.9	11.0
	CVL 1	AF_AM	CASE	MAJOR	96.4	33.3	52.4	66.7	73.8	41.0	97.6	51.2	56.1	79.3	90.2
		AF_AM	CASE												
		AF_AM	CASE	MINOR	3.6	66.7	47.6	33.3	26.2	59.0	2.4	48.8	43.9	20.7	9.8
				p	0.025	0.025	NS	NS	NS	0.025	NS	NS	NS	NS	NS
				%dif	100.0	-6.0	13.3	-21.5	-1.1	16.3	50.0	3.8	9.7	8.8	-11.1
TEST3	CVL 4	WHITE	CONTOL	MAJOR	88.1	52.5	51.5	85.4	61.7	47.4	86.5	61.1	65.9	83.9	87.3
		WHITE	CONTOL												
		WHITE	CONTOL	MINOR	11.9	47.5	48.5	14.6	38.3	52.6	13.5	38.9	34.1	16.1	12.7
	CVL 1	WHITE	CASE	MAJOR	89.4	58.5	62.0	75.0	60.6	57.6	89.1	69.1	72.3	81.9	94.4
		WHITE	CASE												
		WHITE	CASE	MINOR	10.6	41.5	38.0	25.0	39.4	42.4	10.9	30.9	27.7	18.1	5.6
				p	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
				%dif	-10.5	-12.7	-21.6	41.6	2.7	-19.4	-19.3	-20.7	-18.9	11.2	-56.2
TEST3	CVL 4	HIS_LAT	CONTOL	MAJOR	92.4	54.5	59.1	84.8	57.6	45.5	86.4	53.1	53.0	86.4	93.9
		HIS_LAT	CONTOL												
		HIS_LAT	CONTOL	MINOR	7.6	45.5	40.9	15.2	42.4	54.5	13.6	46.9	47.0	13.6	6.1
	CVL 1	HIS_LAT	CASE	MAJOR	95.5	36.4	36.4	95.5	68.2	55.0	95.0	70.0	70.0	70.0	90.9
		HIS_LAT	CASE												
		HIS_LAT	CASE	MINOR	4.5	63.6	63.6	4.5	31.8	45.0	5.0	30.0	30.0	30.0	9.1
				p	NS	NS	NS	NS	NS	NS	NS	NS	NS	0.05	NS
				%dif	-40.0	28.6	35.7	-70.0	-25.0	-17.5	-63.3	-36.0	-36.1	54.5	33.3

Table 21 Number of Samples Per Cohort Test1

					rs333	CCR5_del1_32	rs352140	2848_A>G_hiv03a	rs352139	1174_G>A_hiv03a	rs5743836	p1237_T>C_hiv03a	rs187084	p1486_T>C_hiv03a	rs2004640	IRF5a_T>G	rs2070197	IRF5d_T>C	rs10954213	IRF5c_A>G	rs2280714	IRF5b_T>C	rs429358	apoe_T>C_Cys112Arg_snp2	rs7412	apoe_C>T_Arg158Cys_snp1c
TEST1	CVL 4	AF_AM	CONTOL	COUNT	86	86	86	86	86	85	81	82	81	82	81	82	82	82	82	82	82	82	82	82	82	82
		AF_AM	CONTOL																							
		AF_AM	CONTOL																							
	CVL 1/2/3	AF_AM	CASE	COUNT	85	85	85	85	85	85	73	76	76	76	76	76	76	76	76	76	76	76	76	76	75	75
		AF_AM	CASE																							
		AF_AM	CASE																							
TEST1	CVL 4	WHITE	CONTOL	COUNT	202	201	202	202	201	193	193	189	192	193	193	189	192	193	193	193	193	193	193	193	193	
		WHITE	CONTOL																							
		WHITE	CONTOL																							
	CVL 1/2/3	WHITE	CASE	COUNT	96	96	95	95	96	91	91	92	92	92	92	92	92	92	92	92	92	92	92	92	93	
		WHITE	CASE																							
		WHITE	CASE																							
TEST1	CVL 4	HIS_LAT	CONTOL	COUNT	33	33	33	33	33	33	33	33	32	33	33	32	33	33	33	33	33	33	33	33	33	
		HIS_LAT	CONTOL																							
		HIS_LAT	CONTOL																							
	CVL 1/2/3	HIS_LAT	CASE	COUNT	18	18	18	18	18	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	18	
		HIS_LAT	CASE																							
		HIS_LAT	CASE																							

Table 22 Number of Samples Per Cohort Test2

					rs333	CCR5_del_32	rs352140	2848_A>G_hiv03a	rs352139	1174_G>A_hiv03a	rs5743836	p1237_T>C_hiv03a	rs187084	p1486_T>C_hiv03a	rs2004640	IRF5a_T>G	rs2070197	IRF5d_T>C	rs10954213	IRF5c_A>G	rs2280714	IRF5b_T>C	rs429358	apoe_T>C_Cys112Arg_snp2	rs7412	apoe_C>T_Arg158Cys_snp1c
TEST2	CVL 4	AF_AM	CONTOL	COUNT	86	86	86	86	86	85	81	82	81	82	81	82	82	82	82	82	82	82	82	82	82	82
		AF_AM	CONTOL																							
		AF_AM	CONTOL																							
	CVL 1/2	AF_AM	CASE	COUNT	69	69	69	69	69	69	61	63	63	63	63	63	63	63	63	63	63	63	63	63	63	62
		AF_AM	CASE																							
		AF_AM	CASE																							
TEST2	CVL 4	WHITE	CONTOL	COUNT	202	201	202	202	201	193	193	189	192	193	193	189	192	193	193	193	193	193	193	193	193	193
		WHITE	CONTOL																							
		WHITE	CONTOL																							
	CVL 1/2	WHITE	CASE	COUNT	78	78	77	77	78	74	74	75	75	75	75	75	75	75	75	75	75	75	75	75	75	75
		WHITE	CASE																							
		WHITE	CASE																							
TEST2	CVL 4	HIS_LAT	CONTOL	COUNT	33	33	33	33	33	33	33	32	33	33	32	33	33	33	32	33	33	33	33	33	33	33
		HIS_LAT	CONTOL																							
		HIS_LAT	CONTOL																							
	CVL 1/2	HIS_LAT	CASE	COUNT	18	18	18	18	18	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	18
		HIS_LAT	CASE																							
		HIS_LAT	CASE																							

Table 23 Number of Samples Per Cohort Test3

					rs333	CCR5_del1_32	rs352140	2848_A>G_hiv03a	rs352139	1174_G>A_hiv03a	rs5743836	p1237_T>C_hiv03a	rs187084	p1486_T>C_hiv03a	rs2004640	IRF5a_T>G	rs2070197	IRF5d_T>C	rs10954213	IRF5c_A>G	rs2280714	IRF5b_T>C	rs429358	apoe_T>C_Cys112Arg_snp2	rs7412	apoe_C>T_Arg158Cys_snp1c
TEST3	CVL 4	AF_AM	CONTOL	COUNT	86	86	86	86	86	85	81	82	81	82	81	82	82	82	82	82	82	82	82	82	82	82
		AF_AM	CONTOL																							
		AF_AM	CONTOL																							
	CVL 1	AF_AM	CASE	COUNT	42	42	42	42	42	42	39	41	41	41	41	41	41	41	41	41	41	41	41	41	41	41
		AF_AM	CASE																							
		AF_AM	CASE																							
TEST3	CVL 4	WHITE	CONTOL	COUNT	202	201	202	202	201	193	193	189	192	193	193	193	193	193	193	193	193	193	193	193	193	193
		WHITE	CONTOL																							
		WHITE	CONTOL																							
	CVL 1	WHITE	CASE	COUNT	47	47	46	46	47	46	46	47	47	47	47	47	47	47	47	47	47	47	47	47	47	45
		WHITE	CASE																							
		WHITE	CASE																							
TEST3	CVL 4	HIS_LAT	CONTOL	COUNT	33	33	33	33	33	33	33	33	32	33	33	33	33	33	32	33	33	33	33	33	33	33
		HIS_LAT	CONTOL																							
		HIS_LAT	CONTOL																							
	CVL 1	HIS_LAT	CASE	COUNT	11	11	11	11	11	11	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	11
		HIS_LAT	CASE																							
		HIS_LAT	CASE																							

Table 25 Haploview Chi-Square Permuted-p Values Test1

					rs333	rs352140	rs352139	rs5743836	rs187084	rs2004640	rs2070197	rs10954213	rs2280714	rs429358	rs7412
					CCRE5_del_32	2848_A>G_hiv03a	1174_G>A_hiv03a	p1237_T>C_hiv03a	p1486_T>C_hiv03a	IRF5a_T>G	IRF5d_T>C	IRF5c_A>G	IRF5b_T>C	apoe_T>C_Cys112Arg_snp2	apoe_C>T_Arg158Cys_snp1c
TEST1	CVL 4	AF_AM AF_AM AF_AM	CONTOL CONTOL CONTOL	MAJOR CONTOL MINOR	172	50	101	99	125	82	162	86	99	133	146
					0	122	71	73	45	80	2	76	65	31	18
	CVL 1/2/3	AF_AM AF_AM AF_AM	CASE CASE CASE	MAJOR CASE MINOR	165	58	99	109	117	54	150	73	83	119	133
					5	112	71	61	53	92	2	79	69	33	17
				p	0.0235	0.3153	0.9274	0.2141	0.3381	0.0162	0.9390	0.3701	0.3005	0.5349	0.9199
				2x2 chi	5.1340	1.0080	0.0080	1.5440	0.9180	5.7860	0.0006	0.8030	1.0720	0.3850	0.0100
				perm-p	0.0261	0.9401	1.0000	0.8535	0.9598	0.0710	1.0000	1.0000	0.8685	0.8420	0.9921
TEST1	CVL 4	WHITE WHITE WHITE	CONTOL CONTOL CONTOL	MAJOR CONTOL MINOR	356	211	208	345	248	183	334	231	253	324	337
					48	191	196	59	154	203	52	147	131	62	49
	CVL 1/2/3	WHITE WHITE WHITE	CASE CASE CASE	MAJOR CASE MINOR	166	122	126	150	101	100	164	120	130	151	176
					26	70	64	40	91	82	18	64	54	33	10
				p	0.5674	0.0111	0.0007	0.0492	0.0354	0.0937	0.2256	0.3455	0.2566	0.5749	0.0070
				2x2 chi	0.3300	6.4460	11.5480	3.8690	4.4280	2.8090	1.4680	0.8900	1.2870	0.3150	7.2660
				perm-p	0.5674	0.0357	0.0026	0.2320	0.1125	0.3992	0.7164	0.8821	0.7753	0.8463	0.0216
TEST1	CVL 4	HIS_LAT HIS_LAT HIS_LAT	CONTOL CONTOL CONTOL	MAJOR CONTOL MINOR	61	36	39	56	38	30	57	34	35	57	62
					5	30	27	10	28	36	9	30	31	9	4
	CVL 1/2/3	HIS_LAT HIS_LAT HIS_LAT	CASE CASE CASE	MAJOR CASE MINOR	33	15	15	34	23	18	32	22	23	28	32
					3	21	21	2	13	16	2	12	11	6	4
				p	0.8919	0.2138	0.0920	0.1506	0.5343	0.4778	0.2404	0.2701	0.1606	0.5947	0.3646
				2x2 chi	0.0180	1.5450	2.8390	2.0660	0.3860	0.5040	1.3780	1.2160	1.9680	0.2830	0.8220
				perm-p	1.0000	0.7971	0.4287	0.5570	1.0000	0.9970	0.7955	0.8617	0.6659	0.9074	0.7487

Table 26 Haploview Chi-Square Permuted-p Values Test2

					rs333	rs352140	rs352139	rs5743836	rs187084	rs2004640	rs2070197	rs10954213	rs2280714	rs429358	rs7412
					CCRE5_del_32	2848_A>G_hiv03a	1174_G>A_hiv03a	p1237_T>C_hiv03a	p1486_T>C_hiv03a	IRF5a_T>G	IRF5d_T>C	IRF5c_A>G	IRF5b_T>C	apoe_T>C_Cys112Arg_snp2	apoe_C>T_Arg158Cys_snp1c
TEST2	CVL 4	AF_AM AF_AM AF_AM	CONTOL CONTOL CONTOL	MAJOR CONTOL MINOR	172	50	101	99	125	82	162	86	99	133	146
					0	122	71	73	45	80	2	76	65	31	18
	CVL 1/2	AF_AM AF_AM AF_AM	CASE CASE CASE	MAJOR CASE MINOR	134	49	80	89	94	46	124	60	68	98	110
					4	89	58	49	44	76	2	66	58	28	14
				p	0.0246	0.2270	0.8941	0.2142	0.2973	0.0304	0.7901	0.3572	0.2745	0.4864	0.9329
				2x2 chi	5.0510	1.4600	0.0180	1.5430	1.0860	4.6870	0.0710	0.8480	1.1940	0.4850	0.0070
				perm-p	0.0379	0.8176	1.0000	0.7884	0.9282	0.1766	1.0000	0.9481	0.8949	0.7990	1.0000
TEST2	CVL 4	WHITE WHITE WHITE	CONTOL CONTOL CONTOL	MAJOR CONTOL MINOR	356	211	208	345	248	183	334	231	253	324	337
					48	191	196	59	154	203	52	147	131	62	49
	CVL 1/2	WHITE WHITE WHITE	CASE CASE CASE	MAJOR CASE MINOR	137	97	99	120	88	80	133	95	103	120	144
					19	59	55	34	68	68	15	55	47	30	6
				p	0.9223	0.0388	0.0066	0.0342	0.2527	0.1692	0.2975	0.6356	0.5400	0.2777	0.0029
				2x2 chi	0.0100	4.2690	7.3820	4.4840	1.3080	1.8900	1.0850	0.2250	0.3750	1.1780	8.8680
				perm-p	1.0000	0.1804	0.0213	0.1679	0.6902	0.6716	0.8746	0.9982	0.9855	0.5381	0.0103
TEST2	CVL 4	HIS_LAT HIS_LAT HIS_LAT	CONTOL CONTOL CONTOL	MAJOR CONTOL MINOR	61	36	39	56	38	30	57	34	35	57	62
					5	30	27	10	28	36	9	30	31	9	4
	CVL 1/2	HIS_LAT HIS_LAT HIS_LAT	CASE CASE CASE	MAJOR CASE MINOR	33	15	15	34	23	18	32	22	23	28	32
					3	21	21	2	13	16	2	12	11	6	4
				p	0.8918	0.2138	0.0920	0.1506	0.5343	0.4778	0.2404	0.2701	0.1606	0.5947	0.3646
				2x2 chi	0.0180	1.5450	2.8390	2.0660	0.3860	0.5040	1.3780	1.2160	1.9680	0.2830	0.8220
				perm-p	1.0000	0.7988	0.4265	0.5565	1.0000	0.9970	0.7955	0.8617	0.6659	0.9134	0.7498

Table 27 Haploview Chi-Square Permuted-p Values Test3

					rs333	rs352140	rs352139	rs5743836	rs187084	rs2004640	rs2070197	rs10954213	rs2280714	rs429358	rs7412
					CCRE_del_32	2848_A>G_hiv03a	1174_G>A_hiv03a	p1237_T>C_hiv03a	p1486_T>C_hiv03a	IRF5a_T>G	IRF5d_T>C	IRF5c_A>G	IRF5b_T>C	apoe_T>C_Cys112Arg_snp2	apoe_C>T_Arg158Cys_snp1c
TEST3	CVL 4	AF_AM AF_AM AF_AM	CONTOL CONTOL CONTOL	MAJOR CONTOL MINOR	172	50	101	99	125	82	162	86	99	133	146
					0	122	71	73	45	80	2	76	65	31	18
	CVL 1	AF_AM AF_AM AF_AM	CASE CASE CASE	MAJOR CASE MINOR	81	28	44	56	62	32	80	42	46	65	74
					3	56	40	28	22	46	2	40	36	17	8
				p	0.0127	0.4865	0.3365	0.1615	0.9620	0.1634	0.4759	0.7827	0.5212	0.7329	0.7693
				2x2 chi	6.2160	0.4840	0.9240	1.9600	0.0020	1.9420	0.5080	0.0760	0.4120	0.1160	0.0860
				perm-p	0.0361	0.9950	0.9392	0.6970	1.0000	0.6755	0.9862	1.0000	1.0000	0.9580	0.9780
TEST3	CVL 4	WHITE WHITE WHITE	CONTOL CONTOL CONTOL	MAJOR CONTOL MINOR	356	211	208	345	248	183	334	231	253	324	337
					48	191	196	59	154	203	52	147	131	62	49
	CVL 1	WHITE WHITE WHITE	CASE CASE CASE	MAJOR CASE MINOR	84	55	57	69	57	53	82	65	68	77	85
					10	39	35	23	37	39	10	29	26	17	5
				p	0.7351	0.2918	0.0692	0.0154	0.8502	0.0787	0.5044	0.1493	0.2323	0.6535	0.0424
				2x2 chi	0.1140	1.1110	3.3020	5.8690	0.0360	3.0920	0.4460	2.0800	1.4270	0.2250	4.1180
				perm-p	0.8471	0.7922	0.3655	0.0814	1.0000	0.3749	0.9920	0.5721	0.7490	0.9177	0.1050
TEST3	CVL 4	HIS_LAT HIS_LAT HIS_LAT	CONTOL CONTOL CONTOL	MAJOR CONTOL MINOR	61	36	39	56	38	30	57	34	35	57	62
					5	30	27	10	28	36	9	30	31	9	4
	CVL 1	HIS_LAT HIS_LAT HIS_LAT	CASE CASE CASE	MAJOR CASE MINOR	21	8	8	21	15	11	19	14	14	14	20
					1	14	14	1	7	9	1	6	6	6	2
				p	0.6253	0.1396	0.0642	0.1927	0.3787	0.4540	0.2912	0.1832	0.1793	0.0911	0.6523
				2x2 chi	0.2380	2.1820	3.4250	1.6970	0.7750	0.5610	1.1140	1.7720	1.8030	2.8540	0.2380
				perm-p	1.0000	0.6049	0.3092	0.7088	0.8487	1.0000	0.8797	0.6707	0.6438	0.1864	1.0000

***1.19* CCR5 - Chemokine Receptor 5**

1.19.1 CCR5 Background

Chemokine Receptor 5 (CCR5) is transcribed on the forward strand of chromosome 3 and has a 32 base pair insertion/deletion polymorphism called del32 in its third exon that is easily genotyped by PCR amplification, electrophoresis (Figure 150 page 291). This polymorphism was genotyped for three reasons. Firstly, it is known that CCR5 is a known HIV coreceptor and its expression helps the virus to infect target cells [21]. Secondly, it is known the del32 allele is associated with protection against both HIV and bacterial infection [22]. Thirdly, it is known that CCR5 is located within 6 megabases along the same chromosome as TLR9. Because del32 has a known association with HIV infection and is on the same chromosome near TLR9, it was important to prove there is no linkage disequilibrium between the CCR5 and TLR9 loci, so as to control for the affects of the del32 allele before evaluating the TLR9 locus.

Figure 149 CCR5 (rs333 ins32>del32)

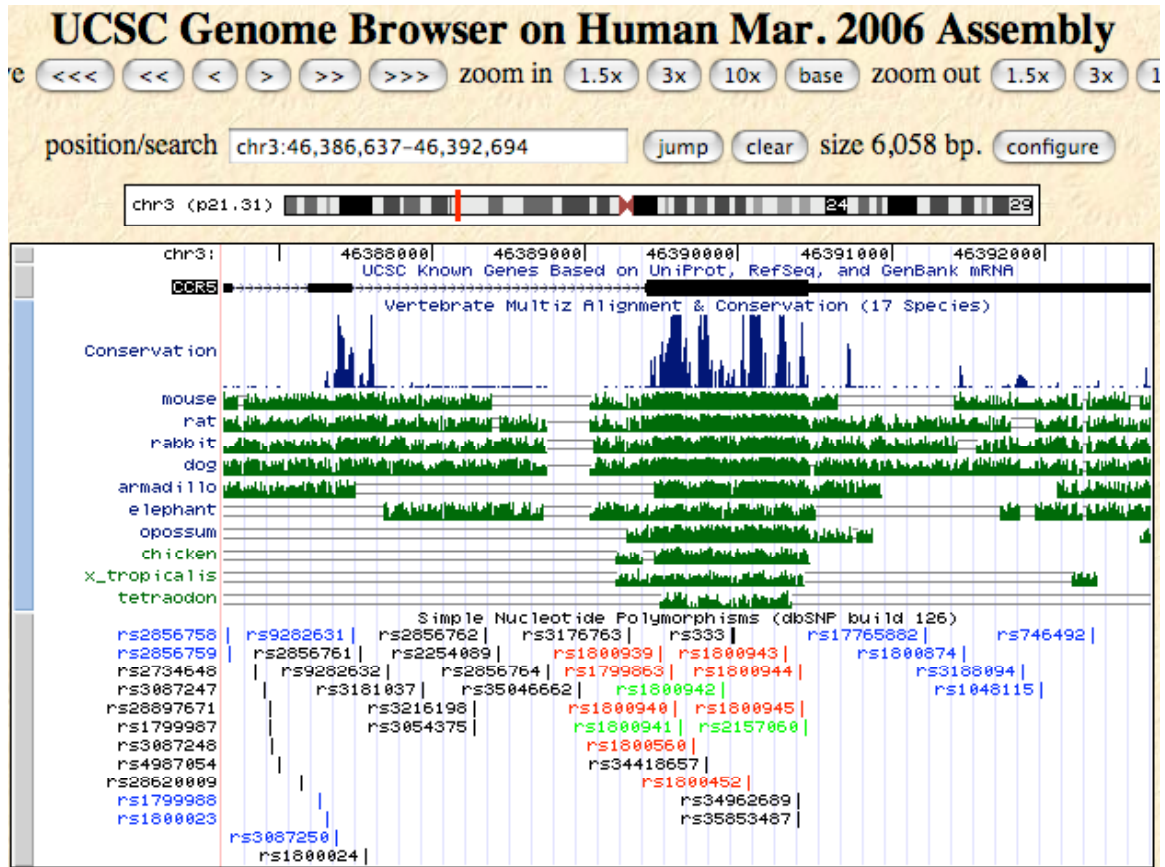
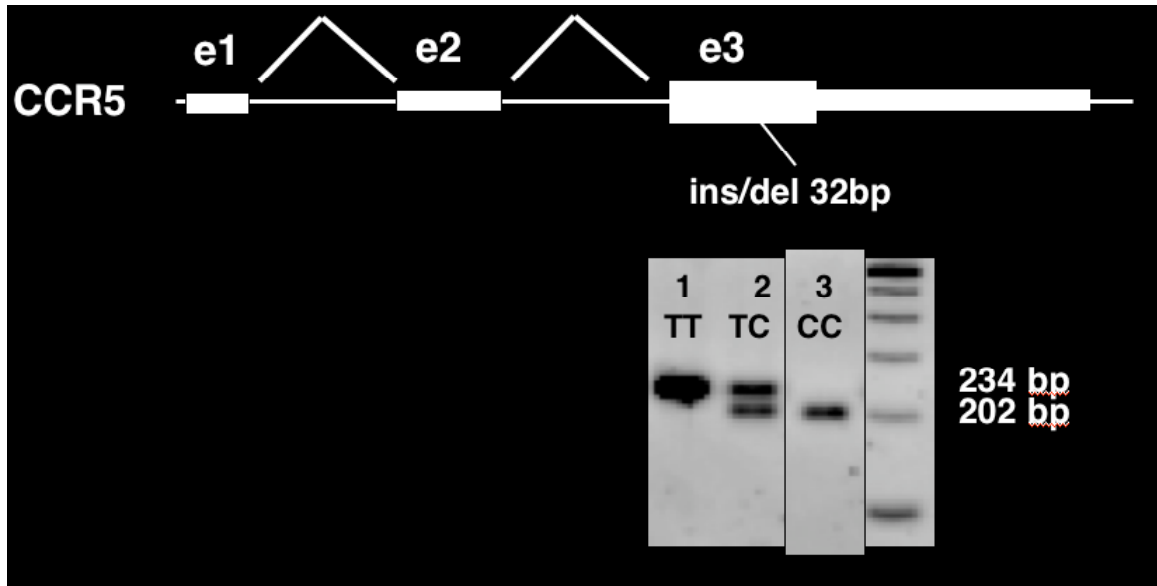


Figure 150 Genotyping the CCR5 del32 (rs333) Polymorphism



1.19.2 CCR5 Results

The *afam* controllers (CVL-1/2/3) had a significantly higher frequency of the [CCR5 rs333](#) del32 allele than the *afam* CVL-4 noncontrollers in all three tests (test1 permuted-p \leq 0.0261; test2 permuted-p \leq 0.0379; test3 permuted-p \leq 0.0361) (Tables pages 277 - 280, 285 - 288). Individuals with two copies of the del32 allele are highly protected against HIV-1 infection, and as expected, no individuals were found to be homozygous for the [CCR5](#) del32 allele in the HIV-1-infected group [21]. I have published a detailed summary of the frequency of the [CCR5](#) del32 allele in the SCOPE cohort in a paper describing the enrichment of dual/mixed/X4-tropic viruses among HIV-1-infected individuals treated with Highly Active Antiretroviral Therapy (HAART) [57]. No associations for the del32 were found in the *white* and *hislat* populations.

1.20 TLR9 - Toll-Like Receptor 9

1.20.1 TLR9 Background

Toll-like Receptor 9 (TLR9) is a protein receptor that is highly expressed on the surface of the endoplasmic reticulum of plasmacytoid dendritic cells and macrophages, and is recruited to the early endosome and a tubular lysosomal compartment upon stimulation with its agonist, unmethylated CpG oligonucleotide [58]. As part of the innate immune response, TLR9 functions to recognize foreign invasion by recognizing the CpG pathogen associated molecular pattern (PAMP) that is common to many bacterial and viral genomes. Strong signaling through TLR9 in dendritic cells leads to the transcription of [NF-κB](#)-dependent inflammatory cytokines (Type 1 IFN, TNF) (Figure 151 page 297, and Figure 152 page 297). Interestingly, mice deficient in TLR9 (TLR9^{-/-} knockouts) are resistant to lethal doses of CpG agonists [59, 60].

TLR9 has different transcriptional isoforms (Figure 154 page 299). TLR9A translates a 1,032 residue protein translated from a transcript that splices a single methionine (exon 1) to exon 2. The TLR9B isoform is 57 amino acids shorter than TLR9A and has been associated with weaker TLR9 signaling than the TLR9A.

TLR9 polymorphisms have been associated with a number of disease phenotypes. Four common tagging-SNPs [[rs352140](#) (2848 A>G or 1635 A>G), [rs352139](#) (1174 G>A), [rs5743836](#) (p1237 T>C), [rs187084](#) (p1486 T>C)] have been identified that define the common TLR9 haplotypes in an association study that failed to correlate these genetic variants with an asthma phenotype (Figure 155 page 6) [55]. It was further shown the frequency of three of these SNPs were ethnicity-specific (*white*, *afam*, and *hislat*).

1.20.1.1 Confirmed Associations

These same four TLR9 polymorphisms have been investigated for their association with the rate of CD4+ T-cell loss in a cohort of HIV-infected 12,000 Swiss [20]. TLR9 rs352139 A>G and rs352140 G>A were each associated with HIV/AIDS rapid progression ($P \leq 0.0008$). In a very recent study 1237 T>C was associated with increased risk of asthma in a cohort of Tunisian children [61], associated with susceptibility to pulmonary aspergillosis [62], and associate with increased risk of atopic eczema (AE) in two panels of families as well as in a cohort of unrelated adults [63].

1.20.1.2 Failed Associations

However, many studies have failed to show association of TLR9 to disease phenotypes. TLR9 1237 T>C could not be associated with multiple sclerosis in a Portuguese population [64], or with predisposition to severe malaria in African children [65], or with Crohn's disease in a New Zealand Caucasian cohort [66]. In two other studies, TLR9 polymorphisms could not be associated with systemic lupus erythematosus in a cohort of family trios [67], or in a case-control study of Caucasian women [68]. Furthermore, SNPs in TLR9 could not be associated with Behçet's disease in Japanese patients [69].

1.20.2 TLR9 Method

To explore the possibility that polymorphisms in TLR9 might associate with HIV/AIDS viremia levels, four common TLR9 SNPs identified by Lazarus [55] were genotyped 237 in the HIV-infected cohort. In a preliminary survey, 237 individuals were genotyped using RFLP assays (Figure 155 page 300). In order to be sure that other common or rare SNPs weren't overlooked in this study, a the TLR9 locus was resequenced on both the

forward and reverse strand at the TLR9 locus across 8,000 bases. From the genotypes of the 237 unrelated individuals, the frequencies of the most common haplotypes of TLR9 were calculated using Haploview software. Eleven individuals representing all of the major haplotypes were identified and resequenced. It was hypothesized other potentially rare SNPs may be identified by sequencing DNA samples from individuals representing all of the major haplotypes. In total 12 amplicons were resequenced spanning the 5,000 base pair gene TLR9 transcript and the 3,000 base pair promoter region (Figure 157 page 301). A total of 10 polymorphisms were identified at the TLR9 locus (Figure 156 page 300, Table 29 page 302). However, after genotyping these it was found that the four tagging-SNPs identified by Lazarus were sufficient to identify the major TLR9 haplotypes in the HIV cohort.

Figure 151 Dendritic cells respond through TLR3/7/8/9 [70]

Dendritic cells may be stimulated through TLR3, TLR7, TLR8 and TLR9. Stimulation of TLR9 by unmethylated CpG oligonucleotides is mediated by both an IRF5-dependent and IRF5-independent pathways which both result in the increased production of proinflammatory cytokines.

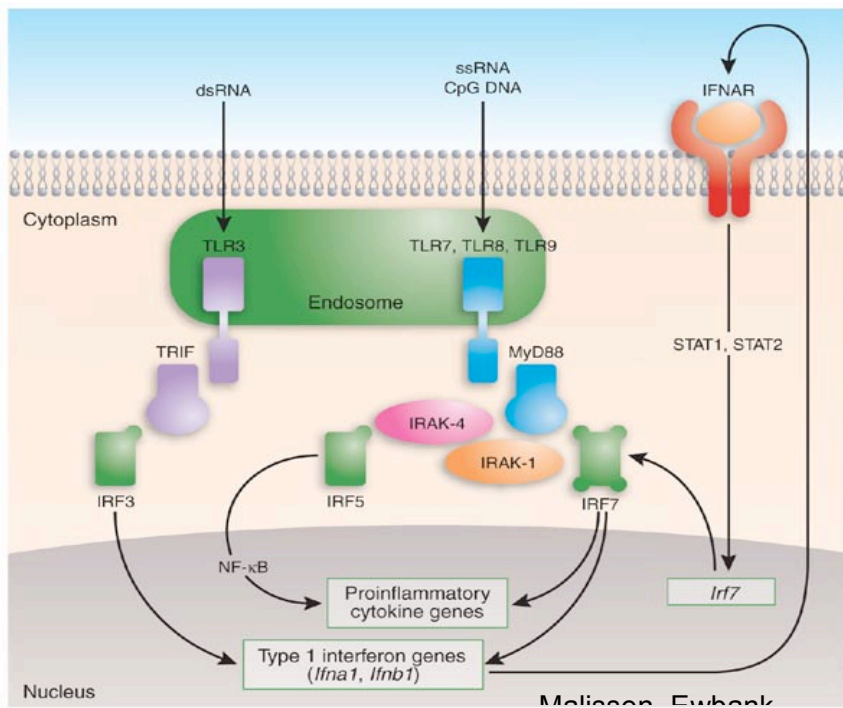
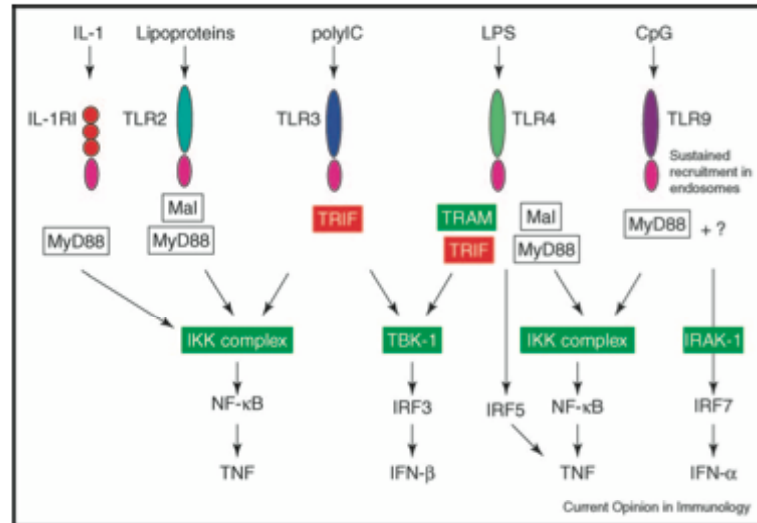


Figure 152 Toll-like Receptor Signaling [54]

Toll-like Receptor signaling ism IRF5 leading to increases of NF-κB and the production Type 1 IFN and TNF.

Figure 1



Different TLRs signal via different combinations of adapters. IL-1RI activates the NF-κB pathway via MyD88. TLR2 requires Mal to enable MyD88 recruitment. TLR9 also signals via MyD88, but causes a sustained recruitment of MyD88 to endosomes for IRF7 activation to occur via IRAK-1. TLR3 recruits Trif, which leads to NF-κB and IRF3 activation. Finally, TLR4 activates NF-κB via Mal and MyD88, and can also trigger IRF3 activation via Tram and Trif. Several TLRs also activate IRF5 (shown here for TLR4) which is involved in induction of genes such as that encoding TNF.

Figure 153 TLR9 (rs352140 G>A, rs352139G>A, rs5743836 T>C, rs187084 T>C)

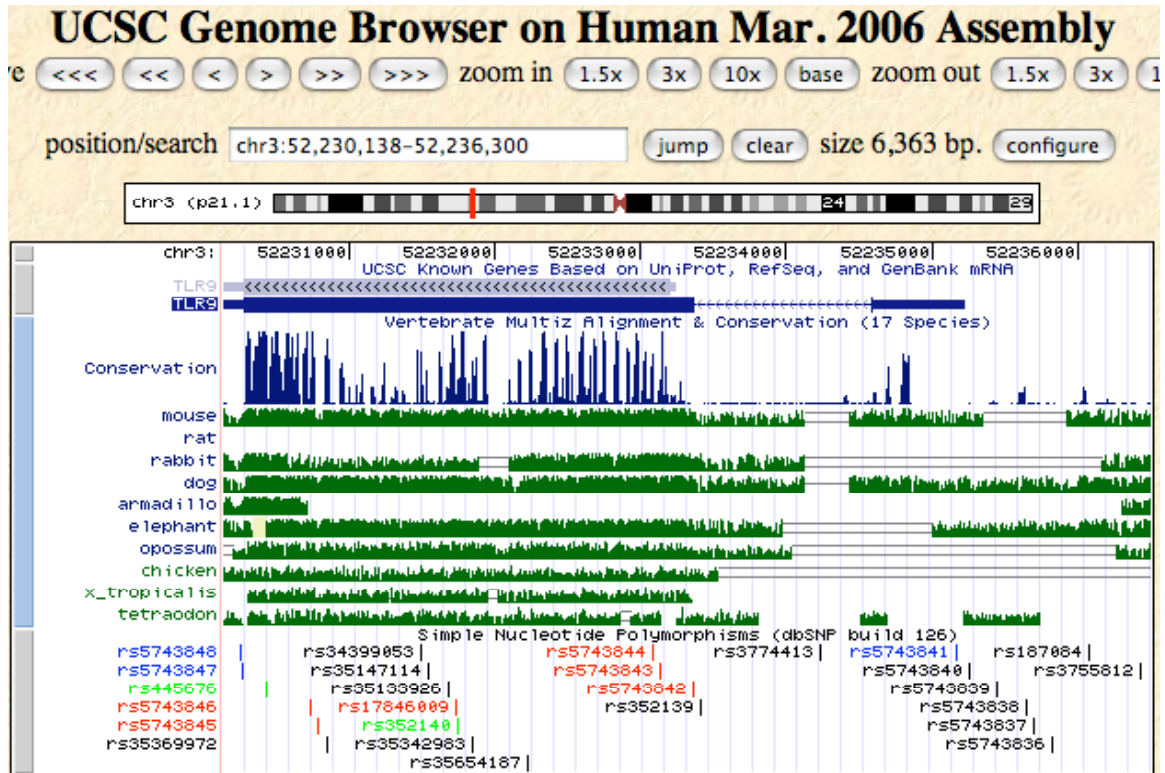


Figure 154 TLR9A/B/C Transcripts Have Variable Signaling Activity

Three TLR9 transcripts can be found in the NCBI database (TLR9A, TLR9B AND TLR9C). The TLR9A transcript is the reference transcript. The TLR9B transcript initiates translation at a secondary methionine resulting in a protein isoform that is 57 amino acids shorter than the TLR9A isoform. The TLR9C isoform may be hypothetical.

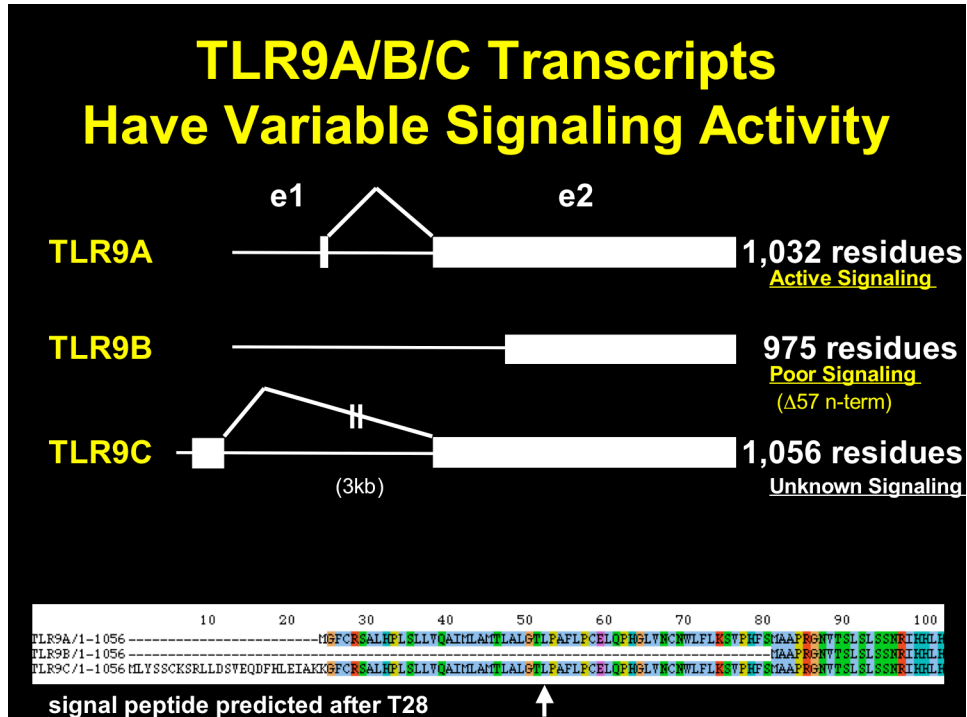


Figure 155 RFLP Agarose Gel Photos for Four TLR9 SNPs

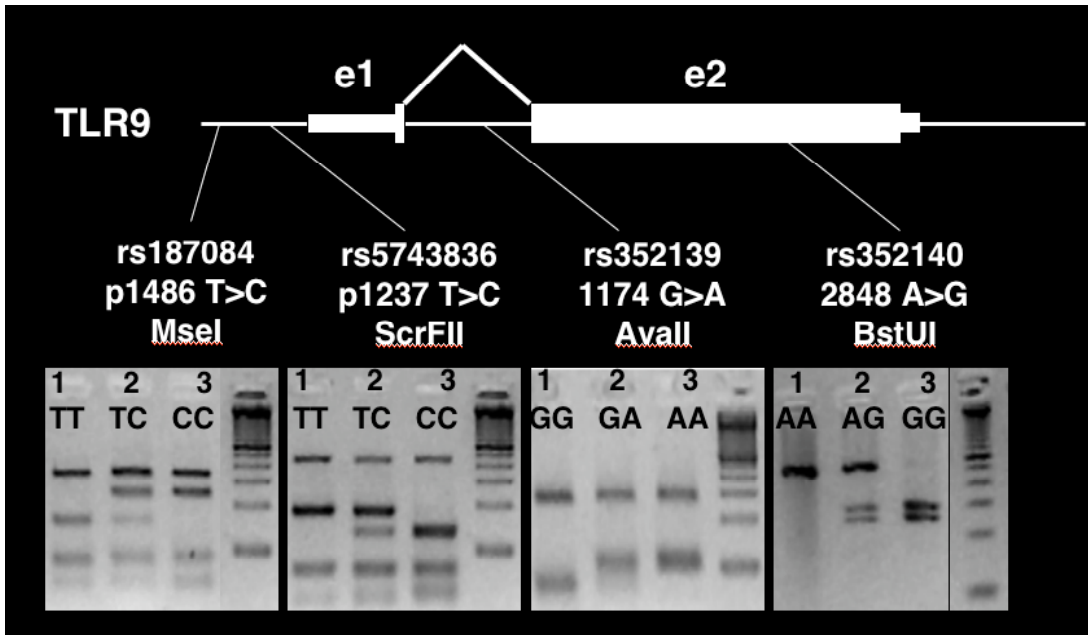


Figure 156 Sequencing Chromatograms of Four TLR9 SNPs

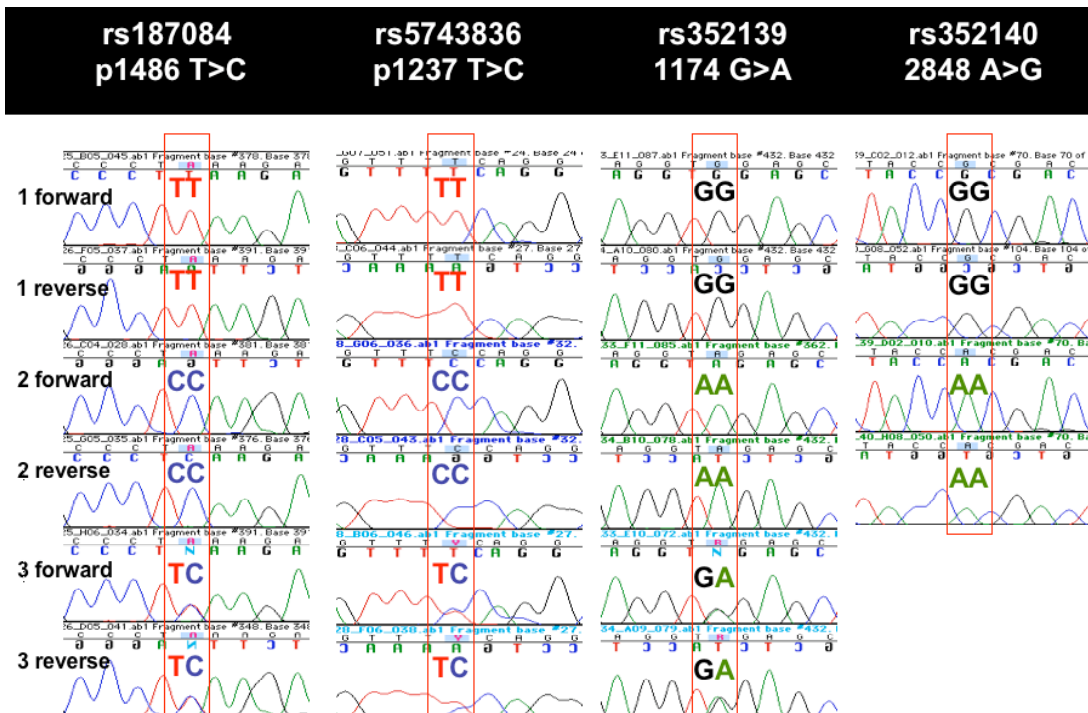
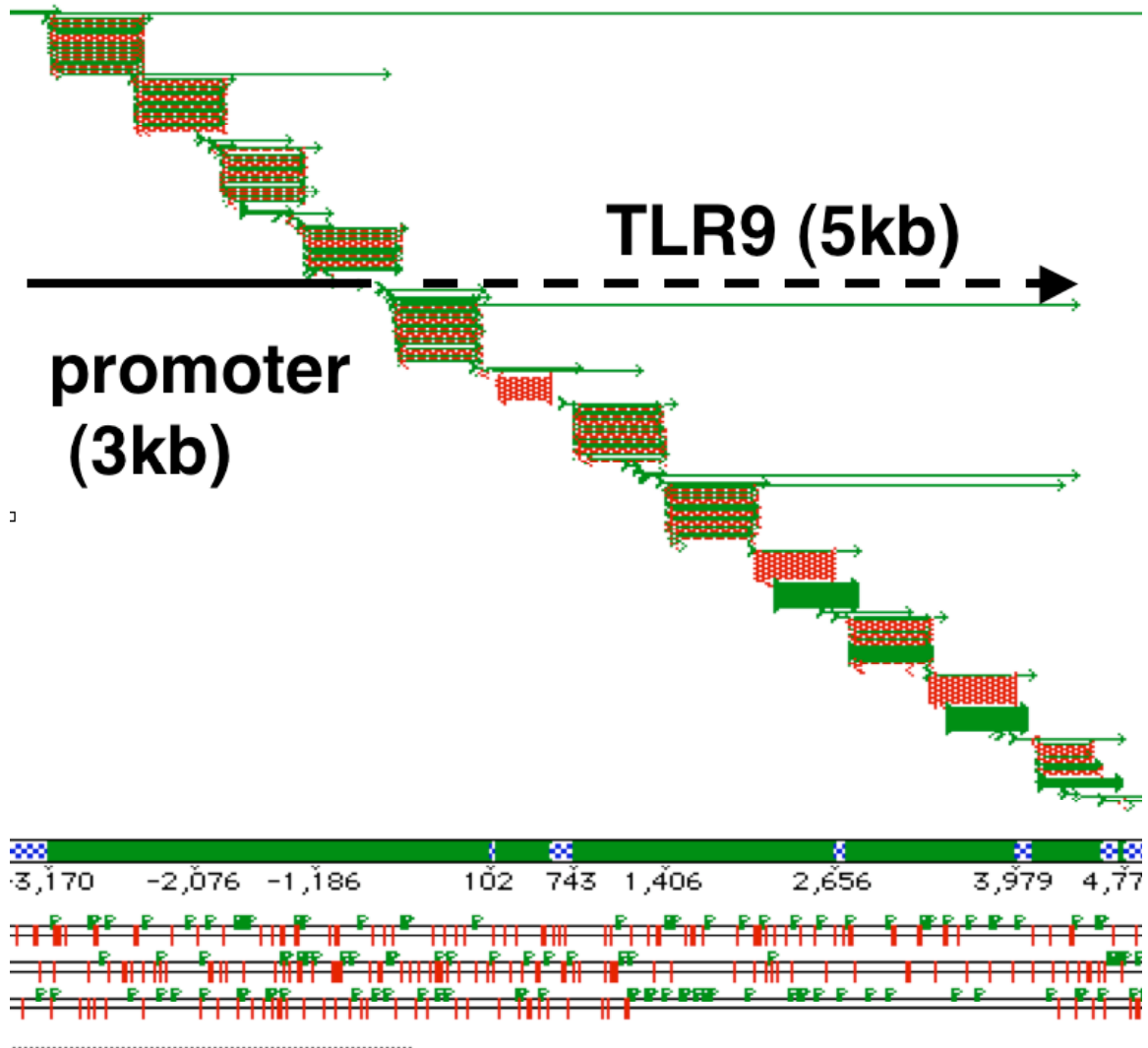


Figure 157 Resequencing the TLR9 Locus (8,000 bp)



1.20.3 TLR9 Results

White controllers (CVL-1/2/3) had significantly higher frequencies of the [TLR9 rs352140](#) A>G allele (permuted-p \leq 0.0357) and the [TLR9 rs352139](#) G>A alleles (permuted-p \leq 0.0026) than the white noncontrollers (CLV-4) (test1) [(Tables pages 277 - 280, 285 - 288) (Figure 158 page 300)]. These polymorphisms correlated with a significantly lower frequency of TLR9 haplotype 1 in the white controllers versus the noncontrollers during test1 (permuted-p \leq 0.0082) (Figure 159 page 308). This association was reduced to a trend during test2 (permuted-p \leq 0.0521) and was lost during test3 as the number of noncontrollers was reduced (CVL-1/2/3 n = 96, CVL-1/2 n = 78, CVL-1 n = 47) (Tables pages 305, 306, and 307). Although it did not reach statistical significance, there was a trend of enrichment of TLR9 haplotype 4 in the *afam* noncontrollers (CVL-4) in test1 (permuted-p \leq 0.0938) and test2 (permuted-p \leq 0.0757).

The *rsquare* and *dprime* (Definitions page 209) linkage disequilibrium values across the TLR9 locus, and between the TLR9 and CCR5 loci were calculated (Figures pages 309 - 311). No linkage disequilibrium was detected between the TLR9 and CCR5 loci.

In White Americans, the frequency of the TLR9 rs5743836 1237 T>C allele was enriched in the lower HIV viremia (CVL-1/2/3) group when compared with the high viremia group (CVL-4). However, the association did not sustain significance using the Haploview chi-square permutation test (tables pages 285 - 287). It may be the case that study was underpowered to detect the association of rs5743836 because the numbers of individuals in the low viremia group was too small. It may be the case that if this study were replicated in a larger cohort the 1237 T>C may associate.

It should be noted in the *whites*, the rs5743836 T>C allele was in strong linkage disequilibrium with rs352139 G>A and rs352140 A>G (Figure 162 page 310), and that rs5743836 C allele is not on TLR9 haplotype 1 (Figure 159 page 308).

It is plausible that TLR9 rs5743836 T>C may associate with risk of asthma [61], susceptibility to pulmonary aspergillosis [62], and with increased risk of atopic eczema [63] through an NF-κB-dependent mechanism. Delta-MATCH predicted the rs5743836 T>C polymorphism has a strong potential to create an allele-specific NF-κB binding site (Figure 138 page 244). If the minor rs5743836 1237 C allele has a stronger predicted binding affinity to the NF-κB transcription factor, it is plausible the TLR9 expression levels for individuals homozygous for the major (T) and minor (C) alleles may differ.

It was also predicted the TLR9A and TLR9B isoforms may differ in that the longer isoform may have a signal sequence in its N-terminus at position T28. If a signal sequence is required for proper TLR9 processing, it is possible that an isoform lacking the signal sequence may not function properly and associate with lower signaling efficiency. It would be interesting to investigate which TLR9 genotypes and haplotypes associate with TLR9A and TLR9B isoform production.

Table 30 TLR9 Haplotypes Test1 (CVL-1/2/3 vs CVL-4)

test1	CVL4 vs CVL1/2/3		hap1	hap2	hap3	hap4	hap5	hap6	hap7	hap8	
			1	2	3	4	5	6	7	8	
2848 G>A hiv03a	rs352140		G	A	A	G	G	G	A	G	
1174 G>A hiv03a	rs352139		A	G	G	G	G	G	G	A	
p1237 T>C hiv03a	rs5743836		T	T	C	C	T	C	T	C	
p1486 T>C hiv03a	rs187084		T	C	T	T	C	C	T	C	
afam											
	control	hap	67.9	26.3	20.4	35.0	1.4	14.6	3.3	3.0	
	control	non-hap	104.1	145.7	151.6	137.0	170.6	157.4	168.7	169.0	86.0
	case	hap	68.9	30.5	24.8	18.1	3.2	17.1	2.7	1.0	
	case	non-hap	101.1	139.5	145.2	151.9	166.8	152.9	167.3	169.0	85.0
	ave	freq	0.400	0.166	0.132	0.155	0.013	0.093	0.018	0.012	0.989
	control	freq	0.395	0.153	0.118	0.203	0.008	0.085	0.019	0.018	0.999
	case	freq	0.405	0.180	0.146	0.107	0.019	0.101	0.016	0.006	0.980
	dif	case-control	0.0100	0.0270	0.0280	-0.0960	0.0110	0.0160	-0.0030	-0.0120	
		p	0.8475	0.5067	0.4555	0.0135	0.3730	0.6237	0.8024	0.2337	
	2x2	chi	0.0370	0.4410	0.5570	6.1070	0.7940	0.2410	0.0630	0.9740	
		perm-p	1.0000	1.0000	1.0000	0.0938	0.9936	1.0000	1.0000	0.9431	
white											
	control	hap	189.9	148.9	57.1						
	control	non-hap	214.1	255.1	346.9						202.0
	case	hap	65.0	84.0	34.0						
	case	non-hap	127.0	108.0	158.0						96.0
	ave	freq	0.428	0.391	0.153						0.972
	control	freq	0.470	0.369	0.141						0.980
	case	freq	0.339	0.437	0.177						0.953
	dif	control-case	-0.1310	0.0680	0.0360	0.0000	0.0000	0.0000	0.0000	0.0000	
		p	0.0024	0.1077	0.2591						
	2x2	chi	9.1970	2.5880	1.2740						
		perm-p	0.0082	0.3581	0.7290						
hislat											
	control	hap	27.0	27.0	7.0	2.0		1.0	2.0		
	control	non-hap	39.0	39.0	59.0	64.0		65.0	64.0		33.0
	case	hap	21.0	13.0	2.0	0.0		0.0	0.0		
	case	non-hap	15.0	23.0	34.0	36.0		36.0	36.0		18.0
	ave	freq	0.471	0.392	0.088	0.019		0.010	0.020		1.000
	control	freq	0.409	0.409	0.106	0.030		0.016	0.031		1.001
	case	freq	0.583	0.361	0.056	0.000		0.000	0.000		1.000
	dif	control-case	0.1740	-0.0480	-0.0500	-0.0300	0.0000	-0.0160	-0.0310	0.0000	
		p	0.0920	0.6380	0.3901	0.2947		0.4522	0.2884		
	2x2	chi	2.8390	0.2210	0.7390	1.0980		0.5650	1.1270		
		perm-p	0.4287	1.0000	0.9591	0.9060		0.9842	0.8653		

Table 31 TLR9 Haplotypes Test2 (CVL-1/2 vs CVL-4)

test2	CVL4 vs CVL1/2		hap1	hap2	hap3	hap4	hap5	hap6	hap7	hap8		
			1	2	3	4	5	6	7	8		
2848 G>A hiv03a	rs352140		G	A	A	G	G	G	A	G		
1174 G>A hiv03a	rs352139		A	G	G	G	G	G	G	A		
p1237 T>C hiv03a	rs5743836		T	T	C	C	T	C	T	C		
p1486 T>C hiv03a	rs187084		T	C	T	T	C	C	T	C		
afam												
	control	hap	67.9	26.3	20.4	35.0	1.4	14.6	3.3	3.0		
	control	non-hap	104.1	145.7	151.6	137.0	170.6	157.4	168.7	169.0		86.0
	case	hap	55.9	25.6	21.4	13.6	3.2	13.0	2.0	1.0		
	case	non-hap	82.1	112.4	116.6	124.4	134.8	125.0	136.0	137.0		69.0
	ave	freq	0.399	0.167	0.135	0.157	0.015	0.089	0.017	0.013		0.992
	control	freq	0.395	0.153	0.119	0.203	0.008	0.085	0.019	0.017		0.999
	case	freq	0.405	0.186	0.155	0.099	0.023	0.094	0.015	0.007		0.984
	dif	case-control	0.0100	0.0330	0.0360	-0.1040	0.0150	0.0090	-0.0040	-0.0100		
		p	0.8579	0.4391	0.3539	0.0116	0.2775	0.7698	0.7461	0.4313		
	2x2	chi	0.0320	0.5990	0.8590	6.7000	1.1790	0.0860	0.1050	0.6190		
		perm-p	1.0000	0.9992	0.9809	0.0757	0.9066	1.0000	1.0000	0.9984		
white												
	control	hap	189.9	148.9	57.1							
	control	non-hap	214.1	255.1	346.9							202.0
	case	hap	56.0	63.0	30.0							
	case	non-hap	100.0	93.0	126.0							78.0
	ave	freq	0.439	0.378	0.155							0.972
	control	freq	0.470	0.369	0.141							0.980
	case	freq	0.359	0.404	0.192							0.955
	dif	control-case	-0.1110	0.0350	0.0510	0.0000	0.0000	0.0000	0.0000	0.0000		
		p	0.0176	0.4417	0.1371							
	2x2	chi	5.6400	0.5920	2.2100							
		perm-p	0.0521	0.9311	0.5600							
hislat												
	control	hap	27.0	27.0	7.0	2.0		1.0	2.0			
	control	non-hap	39.0	39.0	59.0	64.0		65.0	64.0			33.0
	case	hap	21.0	13.0	2.0	0.0		0.0	0.0			
	case	non-hap	15.0	23.0	34.0	36.0		36.0	36.0			18.0
	ave	freq	0.471	0.392	0.088	0.019		0.010	0.020			1.000
	control	freq	0.409	0.409	0.106	0.030		0.016	0.031			1.001
	case	freq	0.583	0.361	0.056	0.000		0.000	0.000			1.000
	dif	control-case	0.1740	-0.0480	-0.0500	-0.0300	0.0000	-0.0160	-0.0310	0.0000		
		p	0.0920	0.6380	0.3901	0.2947		0.4522	0.2884			
	2x2	chi	2.8390	0.2210	0.7390	1.0980		0.5650	1.1270			
		perm-p	0.4265	1.0000	0.9609	0.9066		0.9848	0.8671			

Table 32 TLR9 Haplotypes Test3 (CVL-1 vs CVL-4)

test3	CVL4 vs CVL1		hap1	hap2	hap3	hap4	hap5	hap6	hap7	hap8	
			1	2	3	4	5	6	7	8	
2848 G>A_hiv03a	rs352140		G	A	A	G	G	G	A	G	
1174 G>A_hiv03a	rs352139		A	G	G	G	G	G	G	A	
p1237 T>C_hiv03a	rs5743836		T	T	C	C	T	C	T	C	
p1486 T>C_hiv03a	rs187084		T	C	T	T	C	C	T	C	
afam											
	control	hap	68.0	26.5	20.2	35.2		14.6	3.3	3.0	
	control	non-hap	104.0	145.5	151.8	136.8		157.4	168.7	169.0	86.0
	case	hap	39.0	13.9	13.2	7.7		6.0	0.9	1.0	
	case	non-hap	45.0	70.1	70.8	76.3		78.0	83.1	83.0	42.0
	ave	freq	0.418	0.158	0.131	0.168		0.081	0.016	0.016	0.988
	control	freq	0.395	0.154	0.118	0.204		0.085	0.019	0.018	0.993
	case	freq	0.464	0.166	0.157	0.092		0.072	0.010	0.012	0.973
	dif	case-control	0.0690	0.0120	0.0390	-0.1120	0.0000	-0.0130	-0.0090	-0.0060	
		p	0.2940	0.8103	0.3756	0.0237		0.7210	0.6007	0.7398	
	2x2	chi	1.1010	0.0580	0.7850	5.1130		0.1280	0.2740	0.1100	
		perm-p	0.9181	1.0000	0.9697	0.1484		1.0000	0.9995	1.0000	
white											
	control	hap	189.9	148.9	57.1						
	control	non-hap	214.1	255.1	346.9						202.0
	case	hap	36.0	32.0	19.0						
	case	non-hap	58.0	62.0	75.0						47.0
	ave	freq	0.454	0.363	0.153						0.970
	control	freq	0.470	0.369	0.141						0.980
	case	freq	0.383	0.340	0.202						0.925
	dif	control-case	-0.0870	-0.0290	0.0610	0.0000	0.0000	0.0000	0.0000	0.0000	
		p	0.1266	0.6084	0.1425						
	2x2	chi	2.3340	0.2630	2.1510						
		perm-p	0.4647	1.0000	0.4846						
hislat											
	control	hap	27.0	27.0	7.0	2.0		1.0	2.0		
	control	non-hap	39.0	39.0	59.0	64.0		65.0	64.0		33.0
	case	hap	14.0	7.0	1.0	0.0		0.0	0.0		
	case	non-hap	8.0	15.0	21.0	22.0		22.0	22.0		11.0
	ave	freq	0.466	0.386	0.091	0.022		0.012	0.023		1.000
	control	freq	0.409	0.409	0.106	0.030		0.016	0.031		1.001
	case	freq	0.636	0.318	0.045	0.000		0.000	0.000		0.999
	dif	control-case	0.2270	-0.0910	-0.0610	-0.0300	0.0000	-0.0160	-0.0310	0.0000	
		p	0.0642	0.4504	0.3918	0.4125		0.5555	0.4052		
	2x2	chi	3.4250	0.5700	0.7330	0.6720		0.3480	0.6930		
		perm-p	0.3092	1.0000	0.9216	0.9868		1.0000	0.9675		

Figure 158 TLR9 rs352139 G>A and rs352140 A>G Associated with Higher HIV Viremia in White Americans

**TLR9 rs352139 G>A and rs352140 A>G
Associated with Higher HIV Viremia
in White Americans**

	frequencies		
	high	low	perm-p
rs352139 G/A	.48	.34	.01
rs352140 A/G	.48	.37	.05

Figure 159 TLR9 Haplotype 1 Associated with Higher Viremia in White Americans

**TLR9 Haplotype 1 Associated with
Higher Viremia in White Americans**

Haplotype	rs187084 T/C	5743836 T/C	rs352139 G/A	rs352140 A/G	frequencies		
					high	low	perm-p
1	T	T	A	G	.470	.339	.008
2	C	T	G	A	.369	.437	
3	T	C	G	A	.141	.177	
count					202	96	

Figure 160 Linkage Disequilibrium (D') for Four TLR9 SNPs and One CCR5 In/Del in African American Test1

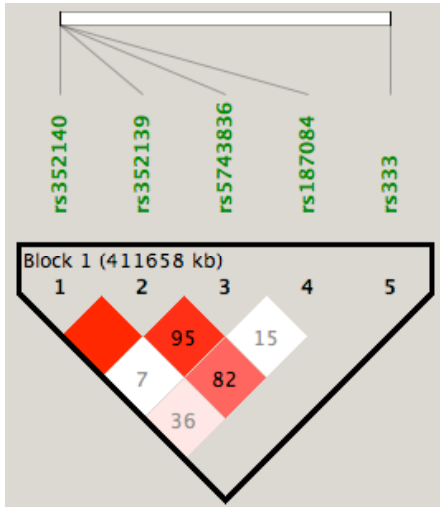


Figure 161 Linkage Disequilibrium (R-squared) for Four TLR9 SNPs and One CCR5 In/Del in African American Test1

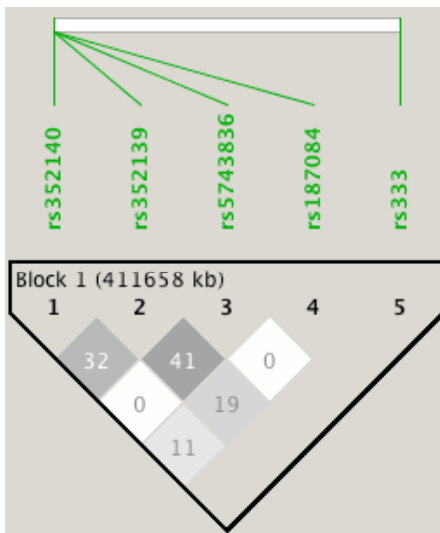


Figure 162 Linkage Disequilibrium (D') for Four TLR9 SNPs and One CCR5 In/Del in White American Test1

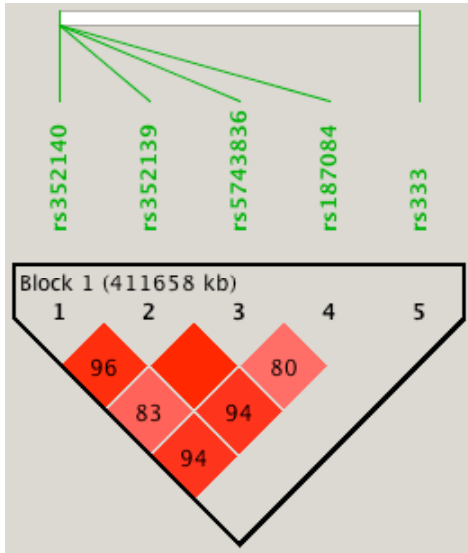


Figure 163 Linkage Disequilibrium (R-squared) for Four TLR9 SNPs and One CCR5 In/Del in White American Test1

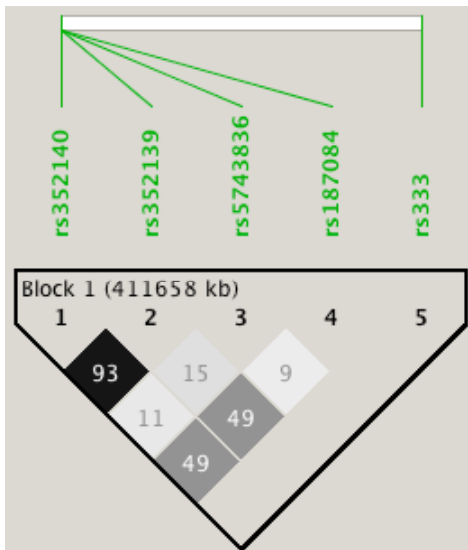


Figure 164 Linkage Disequilibrium (D') for Four TLR9 SNPs and One CCR5 In/Del in His/Lat American Test1

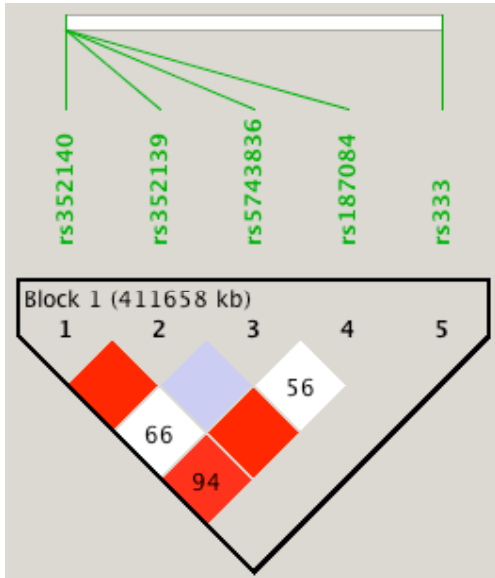
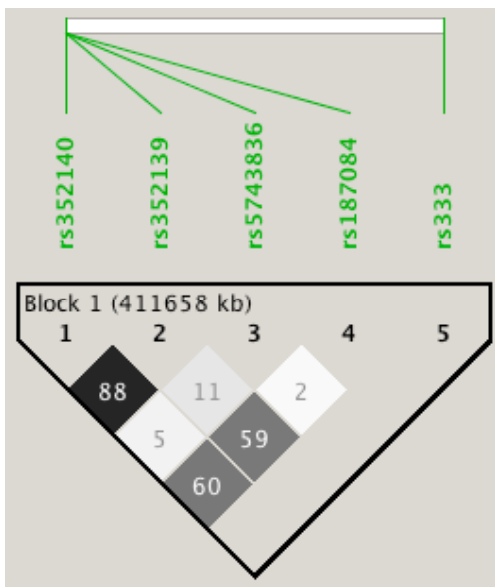


Figure 165 Linkage Disequilibrium (R-squared) for Four TLR9 SNPs and One CCR5 In/Del in His/Lat American Test1



1.21 IRF5 - Interferon Regulatory Fragment 5

1.21.1 IRF5 Background

Interferon responsive factor 5 (IRF5) is a component of the innate immune response and is an important modulator of NF- κ B-dependent interferon cytokine production (Figure 186 page 322). A genomic scan identified the [IRF5](#) locus is associate with variable gene expression [71]. At least 11 [IRF5](#) mRNA transcriptional variants exist, and at least three different transcriptional start sites have been defined (exon1a, exon1b, exon1c) (Figures pages 316 - 320) [17, 72].

Four important IRF5 tagging-SNPs have been identified [([rs2004640](#) T>G, [rs2070197](#) T>C, [rs10954213](#) A>G, and [rs2280714](#) T>C)] and many [IRF5](#) polymorphism associations have been described. The [rs2004640](#) T>G (T) allele is associated with high expression and increased production of transcripts starting at exon1b. The [rs10954213](#) A>G (A) allele produces an early polyadenylation site in the [IRF5](#) 3'UTR, and transcripts with this A allele are shorter, have a longer half-life, and produce 5-fold more protein than transcripts with the G allele [17-19]. The [rs2280714](#) T>C (T) allele has also been associated with high expression levels of IRF5 and systemic lupus erythematosus ([SLE](#)) [17]. Two insertion/deletion (indel) polymorphisms have been described in IRF5 near exons 6 and 7 [19]. The longer 30-bp indel polymorphism removes a 10-amino acid PEST domain in the deleted form. A similar PEST domain in [I \$\kappa\$ B \$\alpha\$](#) , an inhibitor of kappa light chain gene enhancer in B cells, is critical for its calpain-dependent degradation [73]. It has been suggested that the PEST domain in [IRF5](#) may modulate protein stability, but this hypothesis has not been proven [17, 19].

Recent publications have demonstrated that an [IRF5](#) haplotype with [rs2004640](#) T, [rs10954213](#) A, and [rs2280714](#) T is associated with an increased risk [SLE](#) in populations of European-Caucasians and Indo-Pakistanis as shown with a transmission disequilibrium test [18]. Furthermore, a haplotype with [IRF5 rs2004640](#) T, [rs2070197](#) C, and [rs10954213](#) A (haplotype 1) is also associated with [SLE](#), while haplotypes with the [rs2070197](#) T allele (haplotypes 2 and 5) were associated with the 30-bp deletion and the absence of the PEST domain in the [IRF5](#) protein [19]. Finally, the absence of the PEST domain in haplotype 5 is associated with protection against [SLE](#).

Figure 166 IRF5 (rs2004640 T>G, rs2070197 T>C, rs10954213 A>G, rs2280714 T>C)

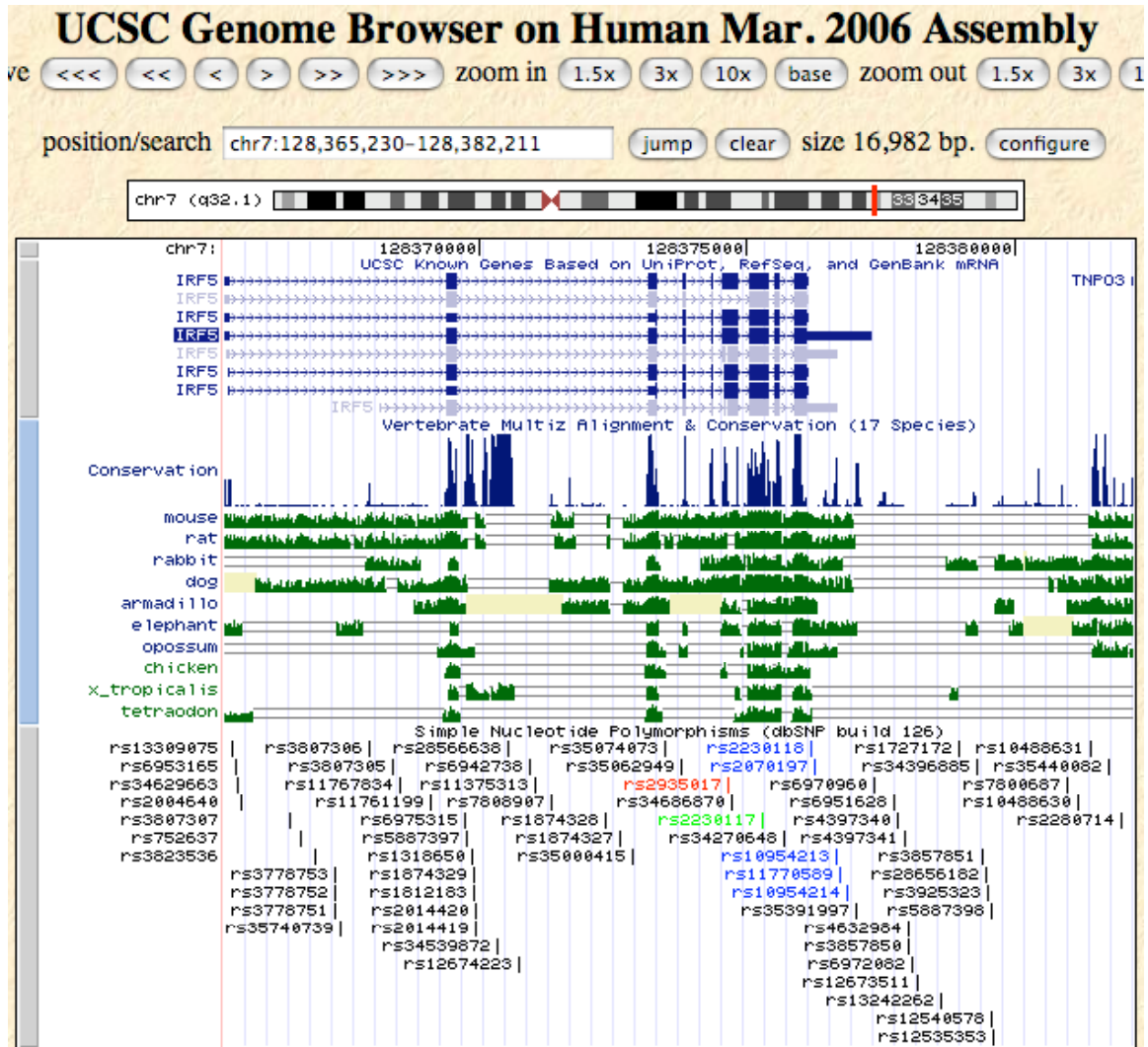


Figure 167 IRF5 mRNA variant shown in the UCSC Genome Browser

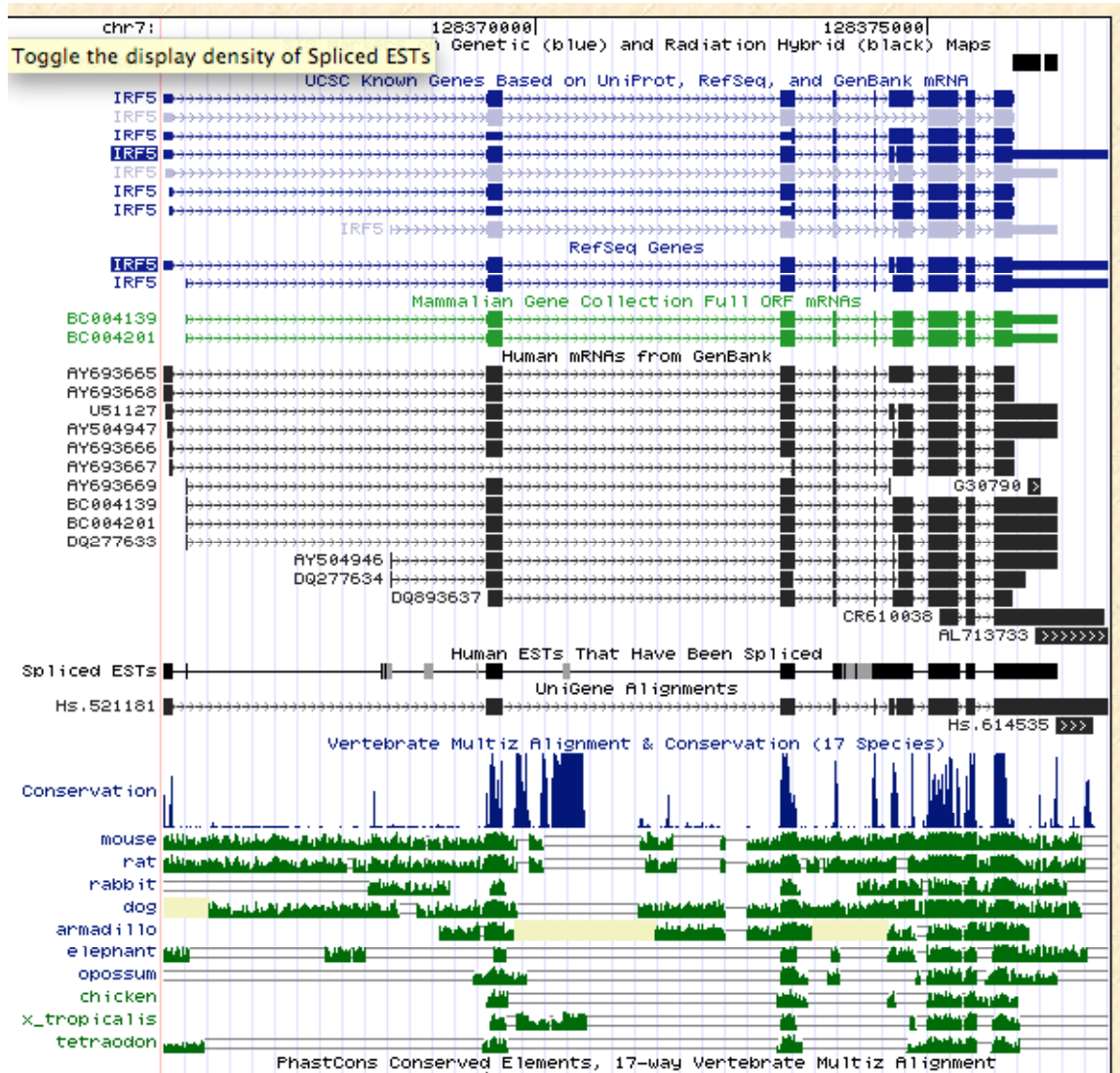


Figure 168 IRF5 Sequencher alignment of 11 mRNA variants (part 1)



Figure 169 IRF5 Sequencher alignment of 11 mRNA variants (part 3)

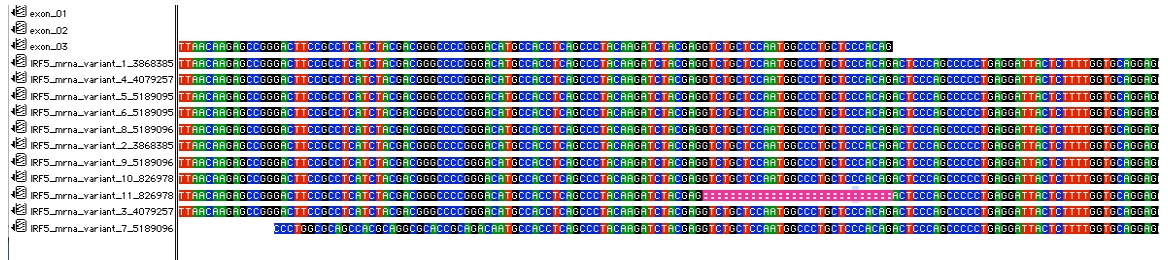


Figure 170 IRF5 Sequencher alignment of 11 mRNA variants (part 3)

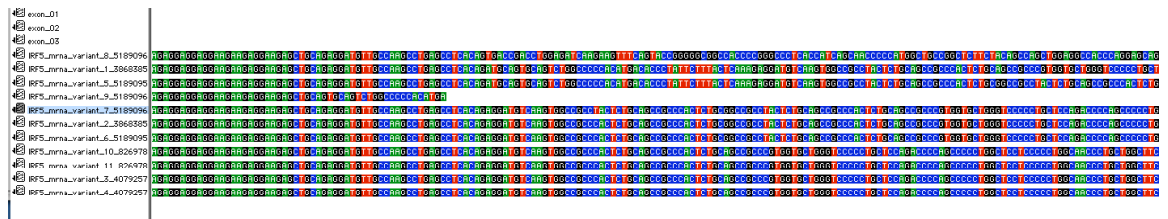


Figure 171 Alignment of five human IRF5 transcripts with one mouse and one cow transcript.

```

gi|40792576|gb|AA90325.1|      QRMLPSLSLT-----EDVKWPPTLQP-----PTL 174
gi|119604111|gb|EAW83705.1|   QRMLPSLSLT-----EDVKWPPTLQPPTLRPPTLQPPTL 184
gi|119604109|gb|EAW83703.1|   QRMLPSLSLTDVAVQSGPHMTPYSLKEDVKWPPTLQPPTLRPPTLQPPTL 200
gi|4504727|ref|NP_002191.1|   QRMLPSLSLTDVAVQSGPHMTPYSLKEDVKWPPTLQP-----PTL 190
gi|119604110|gb|EAW83704.1|   -----
gi|6754368|ref|NP_036187.1|   QRMLPGLSITEPALGPPNAPYSLPKEDTKWP-----PAL 184
gi|78365289|ref|NP_001030542.1| QRMLPGLSITEAVQPGPAMAPYSLPKEDVKWP-----PTL 180

```

Human transcripts show two indel polymorphisms (one 48 bp and one 30 bp) in the IRF5 exon6. Alignment with the mouse (NP_036187) and cow (NP_001030542) show the 30 bp indel coding for a PEST domain is absent other vertebrates, the 48 bp indel however, is present.

Figure 172 IRF5 Sequencer alignment of mRNA variant 1

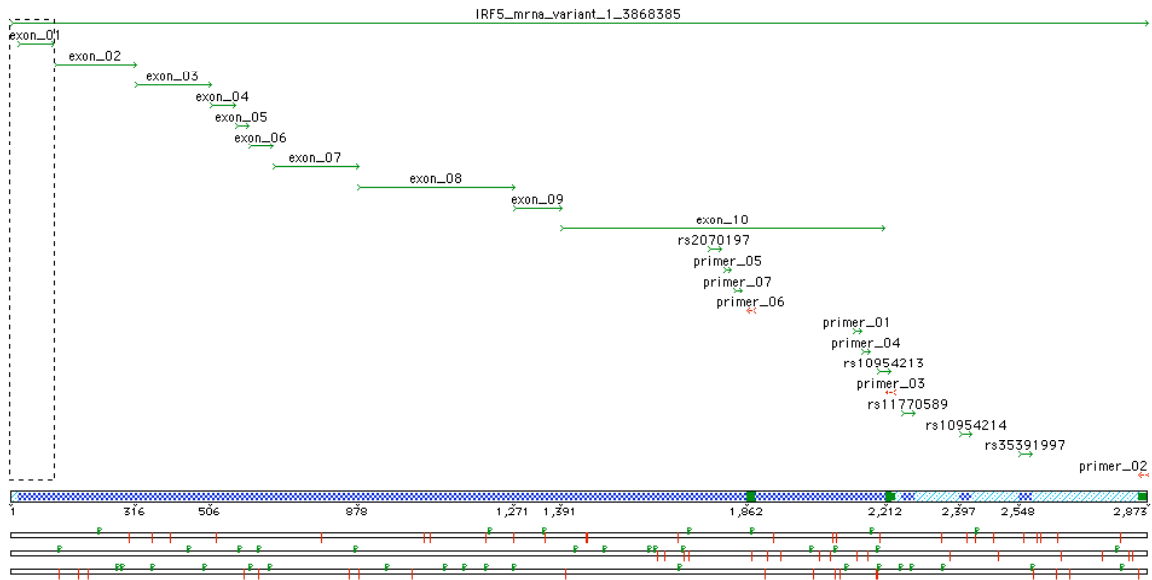


Figure 173 IRF5 Sequencer alignment of mRNA variant 2

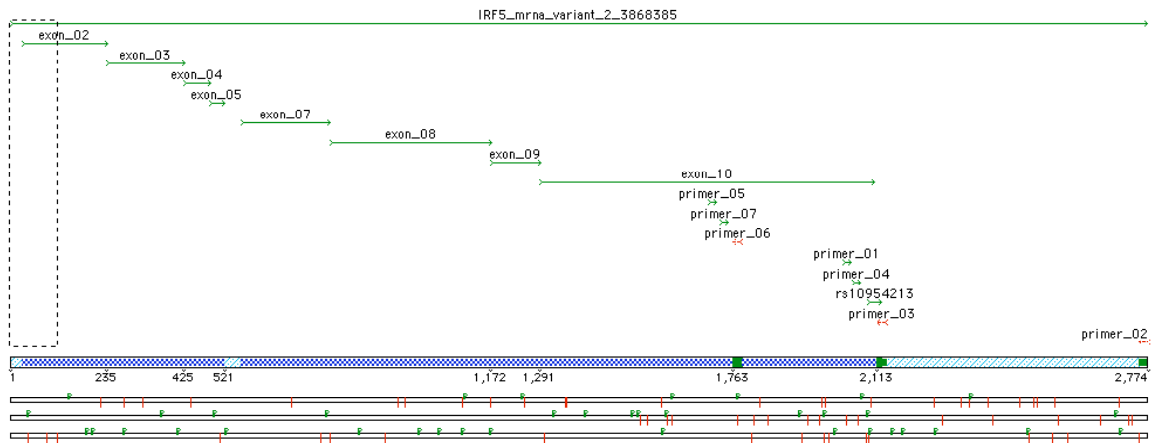


Figure 174 IRF5 Sequencer alignment of mRNA variant 3

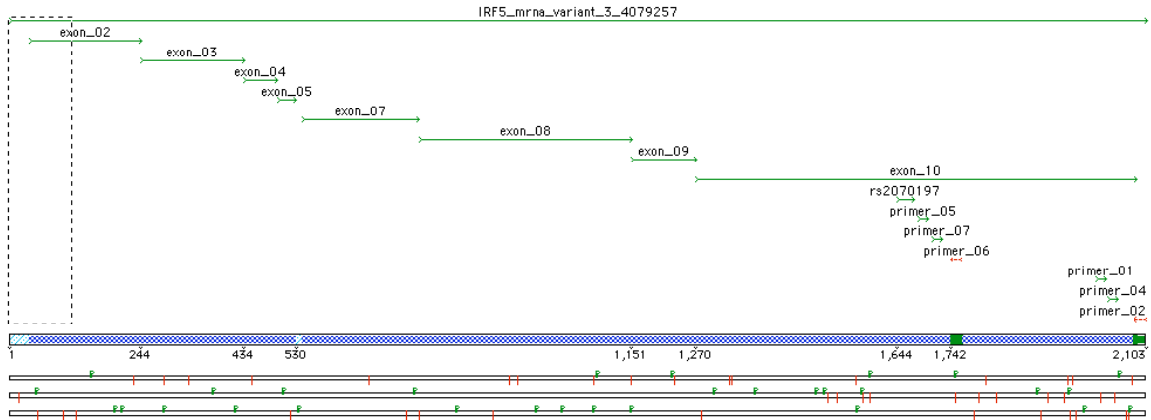


Figure 175 IRF5 Sequencer alignment of mRNA variant 4

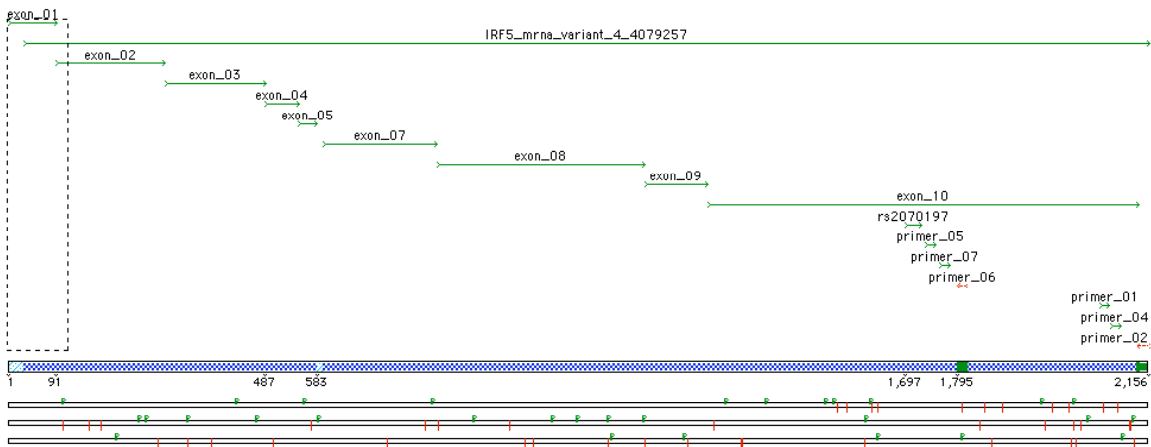


Figure 176 IRF5 Sequencer alignment of mRNA variant 5

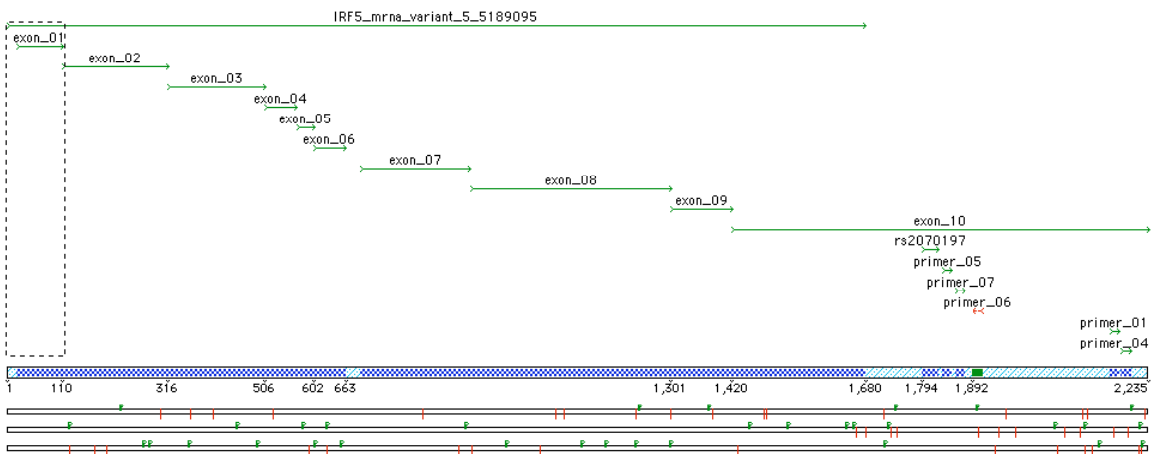


Figure 177 IRF5 Sequencer alignment of mRNA variant 6

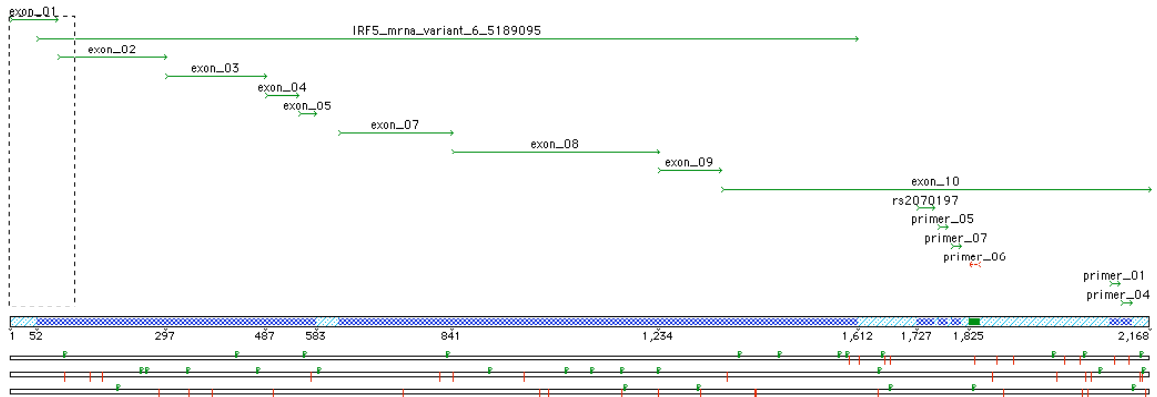


Figure 178 IRF5 Sequencer alignment of mRNA variant 7

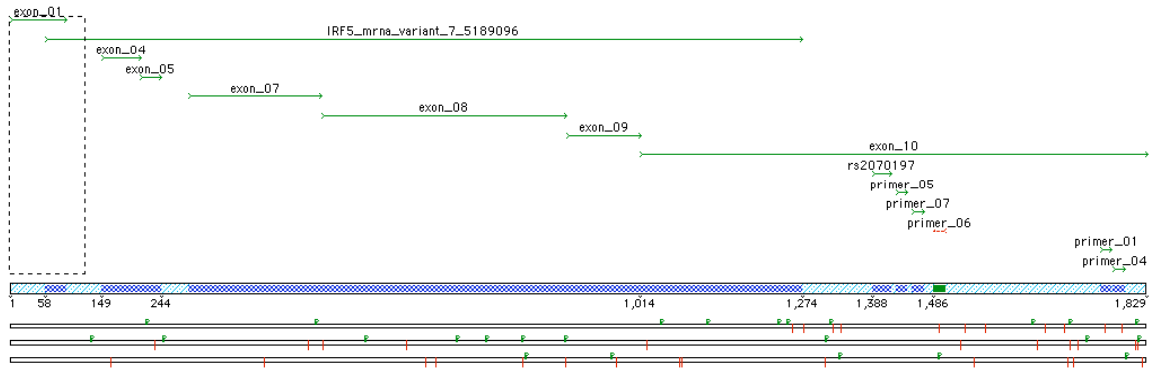


Figure 179 IRF5 Sequencer alignment of mRNA variant 8

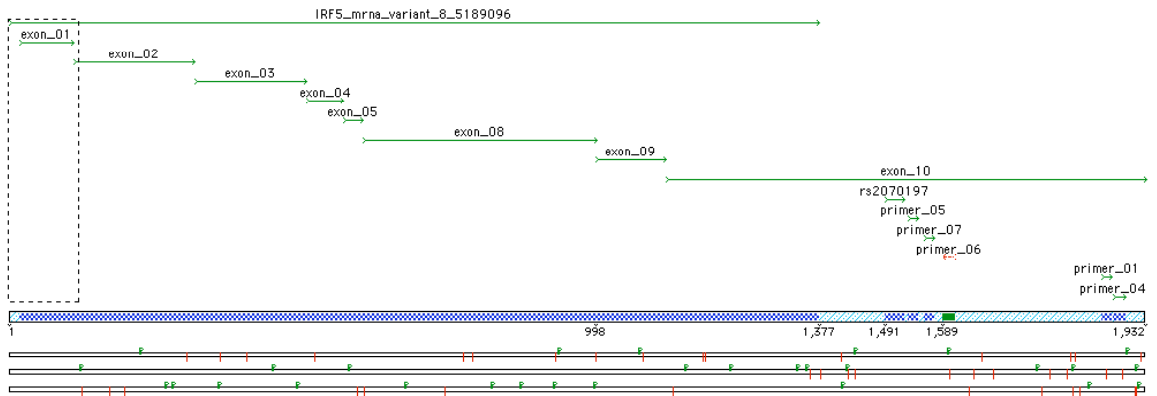


Figure 180 IRF5 Sequencher alignment of mRNA variant 9

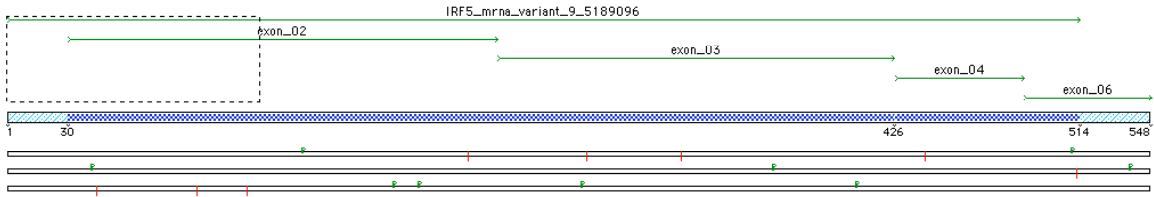


Figure 181 IRF5 Sequencher alignment of mRNA variant 10

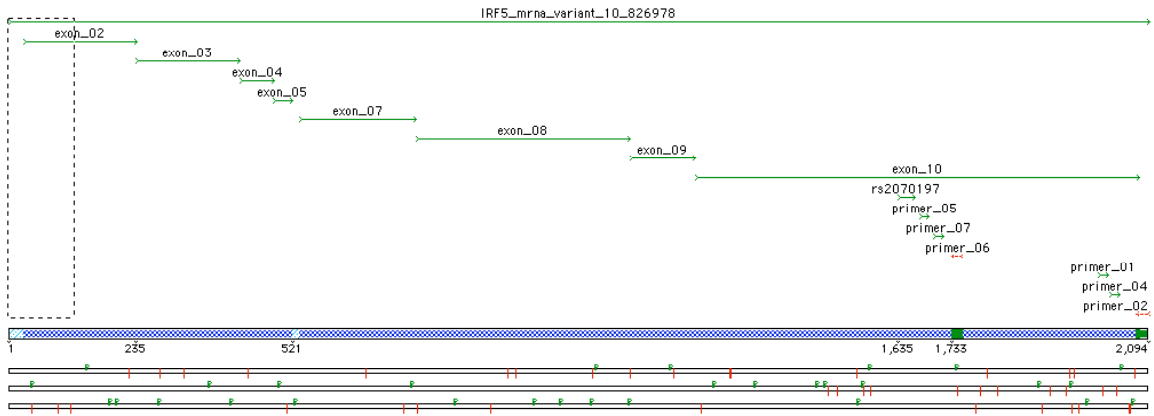


Figure 182 IRF5 Sequencher alignment of mRNA variant 11

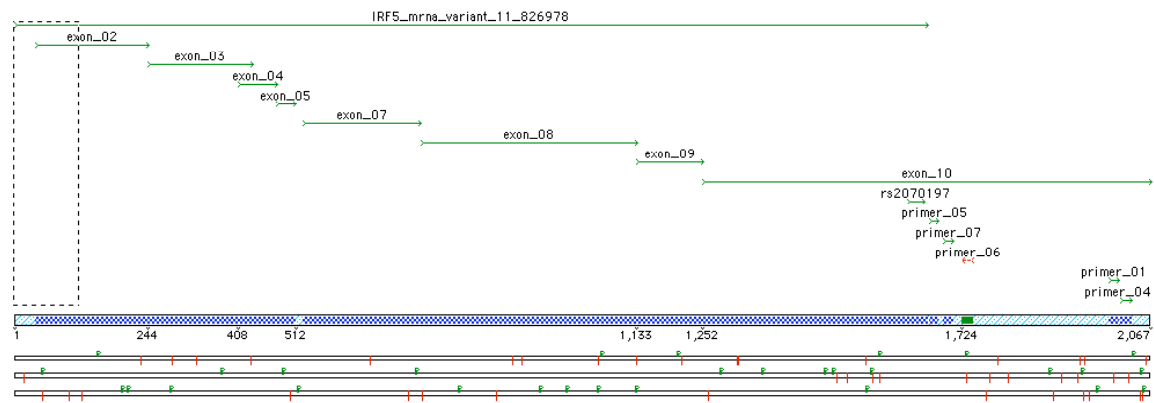


Figure 183 Sequencher alignment legend



Figure 184 Genotyping Four IRF5 SNPs with Taqman Assays-on-Demand

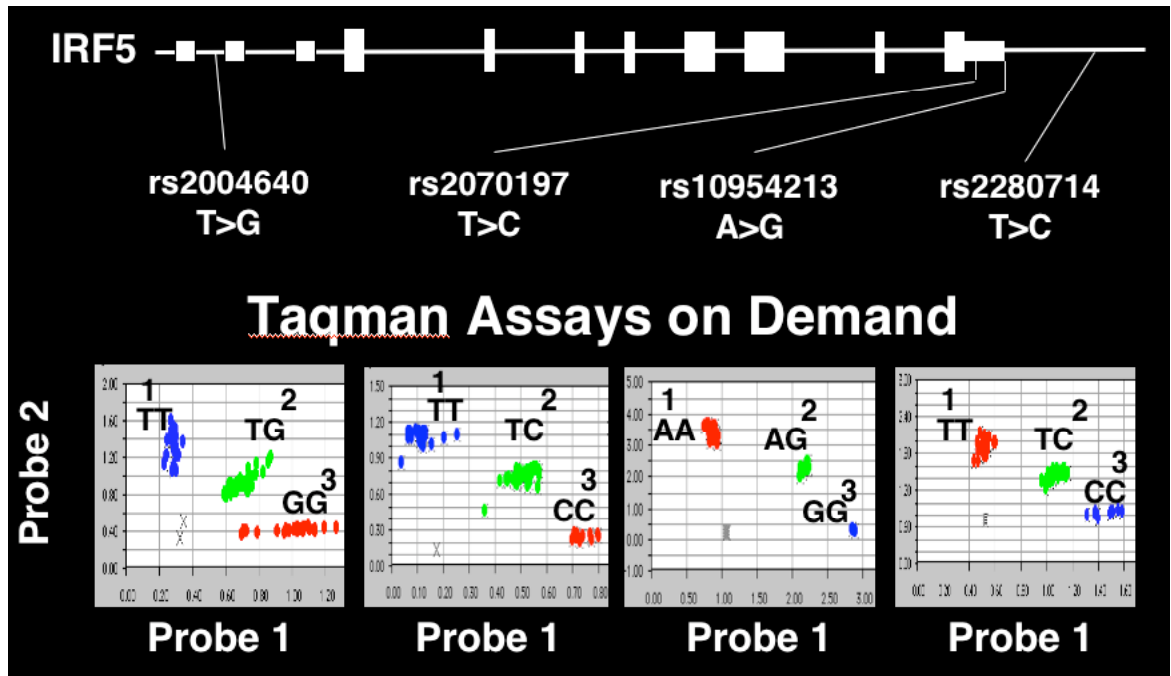
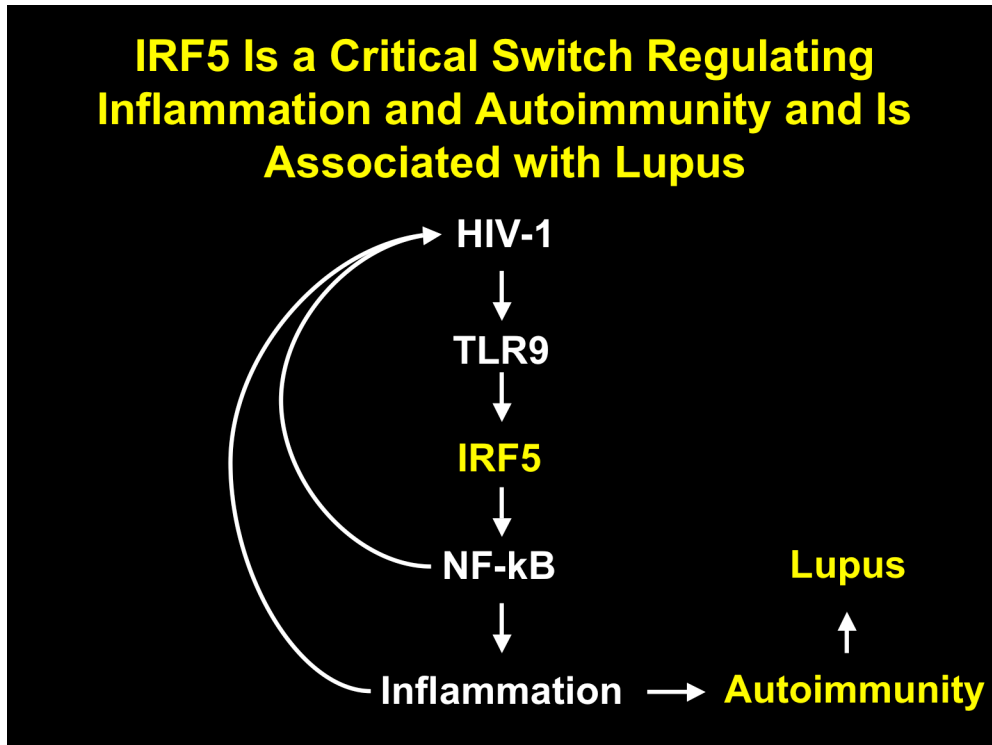


Figure 185 IRF5 Is a Critical Switch Regulating Inflammation and Autoimmunity and is Associated with Lupus



1.21.2 IRF5 Results

The four common IRF5 polymorphisms were genotyped and the frequencies of the common IRF5 haplotypes were estimated (Figure 185 page 321). The *rsquare* and *dprime* (Definitions page) linkage disequilibrium values across the IRF5 locus were calculated (Figures pages 329 - 331).

[IRF5](#) Haplotype 2 was significantly enriched in the *white* controllers during test1 (permuted-p \leq 0.0153) and test3 (permuted-p \leq 0.0174), and a trend was seen in the white test2 (permuted-p \leq 0.0544) (Tables pages 325, 326, and 327) (Figure 186 page 328). None of the IRF5 SNPs showed a direct association with HIV viremia in the cohort. However, the [IRF5 rs2004640](#) T>G allele showed a trend toward an association with higher viremia levels in the *afam* controllers in test1 (permuted-p \leq 0.071) (Tables pages 277 and 285).

Using the four [IRF5](#) tagging-SNPs, I identified five haplotypes in the white HIV-1-infected population (freq \geq 0.01) (Tables pages 325 - 327). [IRF5](#) haplotype 2 was associated with low level of viremia. Interestingly, all the SNP alleles in [IRF5](#) haplotype 2 are present in the genomes of chimpanzees and other vertebrates, and published protein sequences for the cow and mouse are missing the PEST domain (Figure 171 page 316). I hypothesize that [IRF5](#) haplotype 2 (with the 30-bp PEST domain deleted) is the ancestral IRF5 haplotype and that a recent polymorphism at [rs2070197](#) (C allele) produced a variant transcript (haplotype 1). [IRF5](#) haplotype 1 starts transcription at exon1b ([rs2004640](#) T), undergoes early polyadenylation, and has a long mRNA half-life ([rs10954213](#) A). It is highly expressed ([rs2004640](#) T and [rs2280714](#) T) and perhaps gained a novel PEST domain that changed the stability and life-span of the associated

[IRF5](#) protein. A highly expressed, stable, and long-lived [IRF5](#) transcript might associate with elevated sensitivity to TLR agonists and perhaps chronic inflammatory states during viral and bacterial infection. An inability to clear [IRF5](#) levels after a punctuated [NF-kB](#)-dependent inflammatory response might potentiate other autoimmune/autoinflammatory diseases and contribute to HIV-1 progression. In fact, the frequency of [IRF5](#) haplotype 1 was higher in white and *hislal* noncontrollers (CVL-4) than in the controllers, although this did not reach a level of significance (Tables pages 325 - 327).

I proposed to investigate the [IRF5](#) mRNA diversity in HIV-1-infected whites homozygous for [IRF5](#) haplotype 1 or haplotype 2, with the goal of distinguishing how the presence (or absence) of the two [IRF5](#) indel polymorphisms correlate with mRNA and protein stability and levels of HIV-1 viremia. [IRF5](#) haplotypes 1 and 2 have been associated with high expression of [IRF5](#) mRNA and with more stable [IRF5](#) protein than proteins coded by haplotypes 1, 3 and 4 [18, 73]. However, it is unclear how the presence or absence of the PEST domain may associate with sensitivity to TLR agonists. Haplotype 1 or 2 may associate with an inability to produce strong [NF-kB](#)-dependent inflammatory responses and a generally lower inflammatory state. I further hypothesize that [IRF5](#) haplotype 1 and 2 may associate with different magnitudes of [NF-kB](#)-dependent inflammatory responses, and may be associated with an increased risk of diseases such as inflammatory bowel disease ([IBD](#)), multiple sclerosis ([MS](#)), and Alzheimer's disease ([AD](#)), as well as HIV-related dementia.

Table 33 IRF5 Haplotypes Test1 (CVL-1/2/3 vs CVL-4)

test1	CVL4 vs CVL1/2/3		hap1	hap2	hap3	hap4	hap5	
			1	2	3	4	5	
IRF5a_T>G	rs2004640		T	T	T	G	G	
IRF5d_T>C	rs2070197		C	T	T	T	T	
IRF5c_A>G	rs10954213		A	A	G	G	A	
IRF5b_T>C	rs2280714		T	T	T	C	T	
afam								
	control	hap	2.0	70.3	7.4	60.7	14.5	
	control	non-hap	162.0	93.7	156.6	103.3	149.5	82.0
	case	hap	2.0	49.2	4.0	67.6	21.7	
	case	non-hap	150.0	102.8	148.0	84.4	130.3	76.0
	ave	freq	0.013	0.378	0.036	0.406	0.115	0.948
	control	freq	0.012	0.429	0.045	0.370	0.088	0.944
	case	freq	0.013	0.324	0.026	0.444	0.143	0.950
	dif	case-control	0.0010	-0.1050	-0.0185	0.0740	0.0550	
		p	0.9390	0.0544	0.3741	0.1771	0.1261	
	2x2	chi	0.0060	3.6990	0.7900	1.8220	2.3390	
		perm-p	1.0000	0.3259	1.0000	0.6340	0.5171	
white								
	control	hap	52.0	107.2	23.8	125.6	68.3	
	control	non-hap	334.0	278.8	362.2	260.4	317.7	193.0
	case	hap	18.3	73.8	9.6	53.8	27.8	
	case	non-hap	165.7	110.2	174.4	130.2	156.2	92.0
	ave	freq	0.123	0.317	0.059	0.315	0.169	0.983
	control	freq	0.135	0.278	0.062	0.325	0.177	0.977
	case	freq	0.099	0.401	0.052	0.293	0.151	0.996
	dif	control-case	-0.0360	0.1230	-0.0100	-0.0320	-0.0260	0.019
		p	0.2297	0.0031	0.6429	0.4325	0.4355	
	2x2	chi	1.4430	8.7710	0.2150	0.6160	0.6080	
		perm-p	0.7244	0.0153	0.9961	0.9409	0.9422	
hislat								
	control	hap	8.5	19.5	2.0	28.9	5.0	
	control	non-hap	57.5	46.5	64.0	37.1	61.0	33.0
	case	hap	2.0	15.3	0.7	11.0	4.7	
	case	non-hap	32.0	18.7	33.3	23.0	29.3	17.0
	ave	freq	0.105	0.348	0.027	0.399	0.096	0.975
	control	freq	0.128	0.296	0.031	0.438	0.075	0.968
	case	freq	0.059	0.451	0.020	0.323	0.137	0.990
	dif	control-case	-0.0690	0.1550	-0.0110	-0.1150	0.0620	
		p	0.2838	0.1230	0.7473	0.2650	0.3206	
	2x2	chi	1.1490	2.3790	0.1040	1.2420	0.9870	
		perm-p	0.8628	0.5625	0.9994	0.8566	0.9018	

Table 34 IRF5 Haplotypes Test2 (CVL-1/2 vs CVL-4)

test2	CVL4 vs CVL1/2		hap1	hap2	hap3	hap4	hap5	
			1	2	3	4	5	
IRF5a T>G	rs2004640		T	T	T	G	G	
IRF5d T>C	rs2070197		C	T	T	T	T	
IRF5c A>G	rs10954213		A	A	G	G	A	
IRF5b T>C	rs2280714		T	T	T	C	T	
afam								
	control	hap	2.0	70.4	7.3	60.6	4.9	
	control	non-hap	162.0	93.6	156.7	103.4	159.1	82.0
	case	hap	2.0	41.5	2.5	56.6	5.6	
	case	non-hap	124.0	84.5	123.5	69.4	120.4	63.0
	ave	freq	0.014	0.386	0.033	0.404	0.106	0.943
	control	freq	0.012	0.429	0.044	0.370	0.088	0.943
	case	freq	0.016	0.329	0.019	0.449	0.131	0.944
	dif	case-control	0.0040	-0.1000	-0.0250	0.0790	0.0430	
		p	0.7901	0.0839	0.2455	0.1717	0.2430	
	2x2	chi	0.0710	2.9870	1.3480	1.8680	1.3630	
		perm-p	1.0000	0.4247	0.8084	0.6459	0.8039	
white								
	control	hap	52.0	107.1	23.9	125.5	68.3	
	control	non-hap	334.0	278.9	362.1	260.5	317.7	193.0
	case	hap	15.3	58.6	7.8	46.9	21.0	
	case	non-hap	134.7	91.4	142.2	103.1	129.0	75.0
	ave	freq	0.126	0.309	0.059	0.322	0.167	0.983
	control	freq	0.135	0.278	0.062	0.325	0.177	0.977
	case	freq	0.102	0.391	0.052	0.312	0.140	0.997
	dif	control-case	-0.0330	0.1130	-0.0100	-0.0130	-0.0370	0.020
		p	0.3035	0.0109	0.6545	0.7748	0.2981	
	2x2	chi	1.0590	6.4790	0.2000	0.0820	1.0830	
		perm-p	0.8766	0.0544	0.9984	1.0000	0.8747	
hislat								
	control	hap	8.5	19.5	2.0	28.9	5.0	
	control	non-hap	57.5	46.5	64.0	37.1	61.0	33.0
	case	hap	2.0	15.3	0.7	11.0	4.7	
	case	non-hap	32.0	18.7	33.3	23.0	29.3	17.0
	ave	freq	0.105	0.348	0.027	0.399	0.096	0.975
	control	freq	0.128	0.296	0.031	0.438	0.075	0.968
	case	freq	0.059	0.451	0.020	0.323	0.137	0.990
	dif	control-case	-0.0690	0.1550	-0.0110	-0.1150	0.0620	
		p	0.2838	0.1230	0.7473	0.2650	0.3206	
	2x2	chi	1.1490	2.3790	0.1040	1.2420	0.9870	
		perm-p	0.8628	0.5625	0.9994	0.8566	0.9018	

Table 35 IRF5 Haplotypes Test3 (CVL-1 vs CVL-4)

test3	CVL4 vs CVL1		hap1	hap2	hap3	hap4	hap5	
			1	2	3	4	5	
IRF5a_T>G	rs2004640		T	T	T	G	G	
IRF5d_T>C	rs2070197		C	T	T	T	T	
IRF5c_A>G	rs10954213		A	A	G	G	A	
IRF5b_T>C	rs2280714		T	T	T	C	T	
afam								
	control	hap	2.0	70.4	7.5	60.7	14.4	
	control	non-hap	162.0	93.6	156.5	103.3	149.6	82.0
	case	hap	2.0	29.6	1.5	35.7	10.4	
	case	non-hap	80.0	52.4	80.5	46.3	71.6	41.0
	ave	freq	0.016	0.406	0.037	0.392	0.101	0.952
	control	freq	0.020	0.429	0.046	0.370	0.088	0.953
	case	freq	0.003	0.361	0.019	0.435	0.126	0.944
	dif	case-control	-0.0170	-0.0680	-0.0270	0.0650	0.0380	
		p	0.4759	0.3059	0.2929	0.3226	0.3441	
	2x2	chi	0.5080	1.0480	1.1060	0.9780	0.8950	
		perm-p	0.9862	0.9130	0.9082	0.9232	0.9313	
white								
	control	hap	52.0	107.2	23.8	125.5	68.3	
	control	non-hap	334.0	278.8	362.2	260.5	317.7	193.0
	case	hap	10.3	41.3	3.1	25.9	13.4	
	case	non-hap	83.7	52.7	90.9	68.1	80.6	47.0
	ave	freq	0.130	0.309	0.056	0.315	0.170	0.980
	control	freq	0.135	0.278	0.062	0.325	0.177	0.977
	case	freq	0.110	0.439	0.033	0.276	0.142	1.000
	dif	control-case	-0.0250	0.1610	-0.0290	-0.0490	-0.0350	0.023
		p	0.5144	0.0024	0.2779	0.3533	0.4206	
	2x2	chi	0.4250	9.1900	1.1780	0.8620	0.6490	
		perm-p	0.9920	0.0174	0.8503	0.9565	0.9691	
hislat								
	control	hap	8.5	19.5	2.0	28.9	5.0	
	control	non-hap	57.5	46.5	64.0	37.1	61.0	33.0
	case	hap	1.0	10.0	0.0	6.0	3.0	
	case	non-hap	19.0	10.0	20.0	14.0	17.0	10.0
	ave	freq	0.110	0.342	0.024	0.406	0.093	0.975
	control	freq	0.129	0.295	0.031	0.438	0.076	0.969
	case	freq	0.050	0.500	0.000	0.300	0.150	1.000
	dif	control-case	-0.0790	0.2050	-0.0310	-0.1380	0.0740	
		p	0.3247	0.0911	0.4346	0.2687	0.3166	
	2x2	chi	0.9700	2.8550	0.6100	1.2230	1.0030	
		perm-p	0.9076	0.4830	0.9703	0.8460	0.9031	

Figure 186 IRF5 Haplotypes in White Americans

IRF5 Haplotypes in White Americans									
Haplotype	rs2004640 T>G	rs2070197 T>C	rs10954213 A>G	rs2280714 T>C	frequencies				perm-p
					Group 4	Group 1,2,3	Group 1,2	Group 1	
1	T	C	A	T	.135	.099	.102	.110	
2	T	T	A	T	.278	.401*	.391	.439*	.017
3	T	T	G	T	.062	.052	.052	.033	
4	G	T	G	C	.325	.293	.312	.276	
5	G	T	A	T	.177	.151	.140	.132	
count					193	92	78	47	

Figure 187 Linkage Disequilibrium (D') for Four IRF5 SNPs in African American Test1

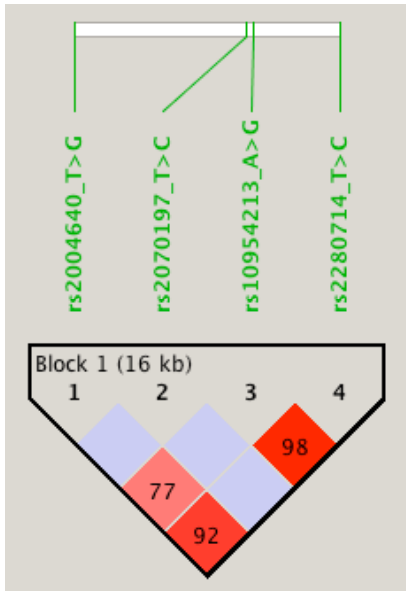


Figure 188 Linkage Disequilibrium (R-squared) for Four IRF5 SNPs in African American Test1

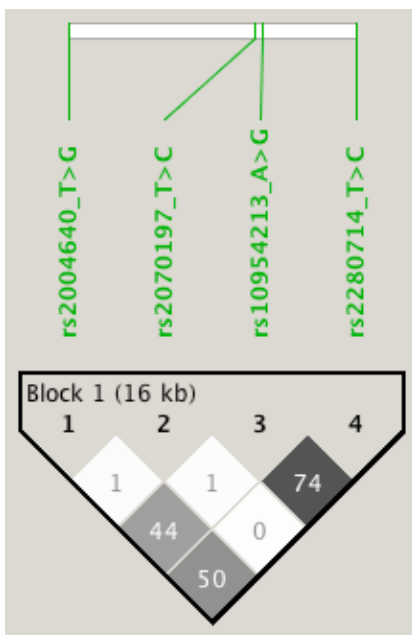


Figure 189 Linkage Disequilibrium (D') for Four IRF5 SNPs in White American Test1

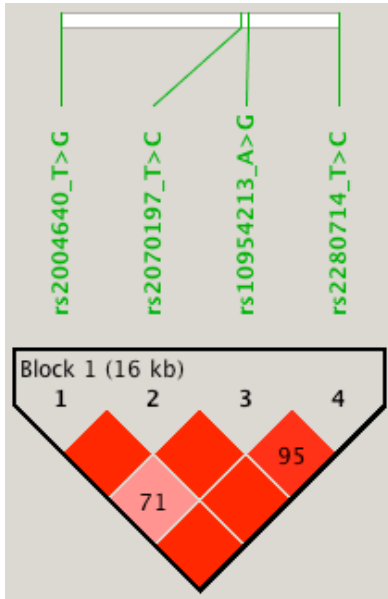


Figure 190 Linkage Disequilibrium (R-squared) for Four IRF5 SNPs in African American Test1

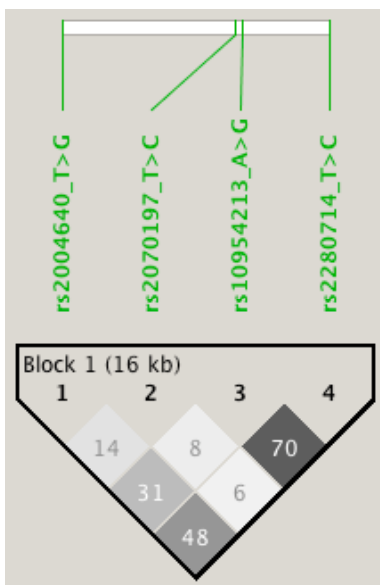


Figure 191 Linkage Disequilibrium (D') for Four IRF5 SNPs in His/Lat American Test1

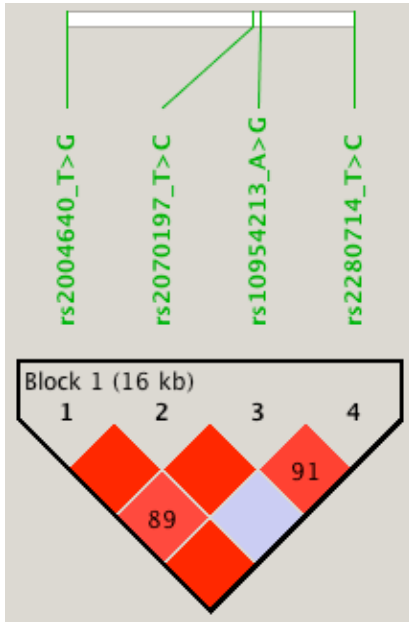
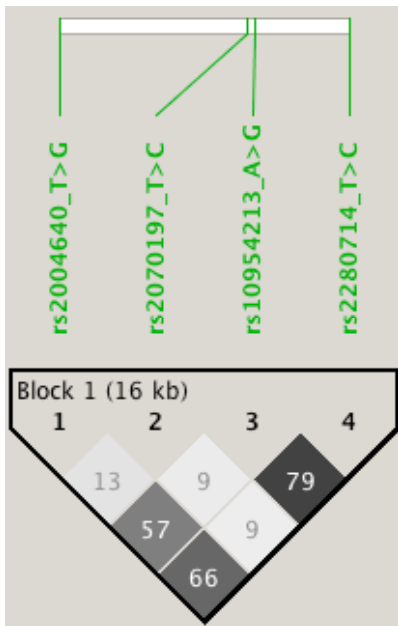


Figure 192 Linkage Disequilibrium (R-squared) for Four IRF5 SNPs in His/Lat American Test1



1.22 APOE - Apolipoprotein E

1.22.1 APOE Background

Apolipoprotein E ([APOE](#)) is a major apoprotein of the chylomicron and facilitates the clearance of chylomicron and very low density lipoprotein remnants from the circulation back to the liver. There are three major isoforms of APOE protein, [ϵ 2, ϵ 3, and ϵ 4]] which can be visualized by isoelectric focusing [74] or predicted by genotyping two common (SNPs rs429358 T>C, rs7412 C>T) (Figure 194 page 334). APOE ϵ 3 is considered the wild-type allele and is the most abundant.

Strong associations between the APOE alleles and the pathologies of multiple disorders have been identified. The ϵ 4 allele, for example, is associated with increased risk of cardiovascular disease (CVD), Alzheimer's disease ([AD](#)), and HIV-related dementia [12-14]. Transgenic mice expressing the human ϵ 4 protein are used as a model of AD [15], and mice deficient in apoE have elevated lipid levels, and are used as a proinflammatory model for studying atherosclerosis [15, 16]. Furthermore, homozygosity of ϵ 2 is associated with dysbetalipoproteinemia [75].

APOE has also been associated with HIV/AIDS-related pathologies. It has been shown APOE variants contribute to an unfavorable lipid profile and can lead to severe hyperlipidemia in HIV-infected individuals on antiretroviral therapy [76]. In a separate study it was shown the ϵ 4 associated with HIV-associated dementia (HAD) in an aging cohort of Hawaiians after controlling for age and diabetes status [77]. In unpublished work from Trevor Burt ([GIVI](#)) has shown that the [APOE](#) epsilon 4 allele is associated with increased HIV-1 fusion. Because of these findings, it was reasonable to hypothesize

that $\epsilon 4$ may associate with HIV viremia levels in my cohort. So I set out to investigate by genotyping the rs429358 T>C and rs7412 C>T polymorphisms using taqman assays (Figure 194 page 334).

Figure 193 APOE (rs429358 T>C, rs7412 C>T)

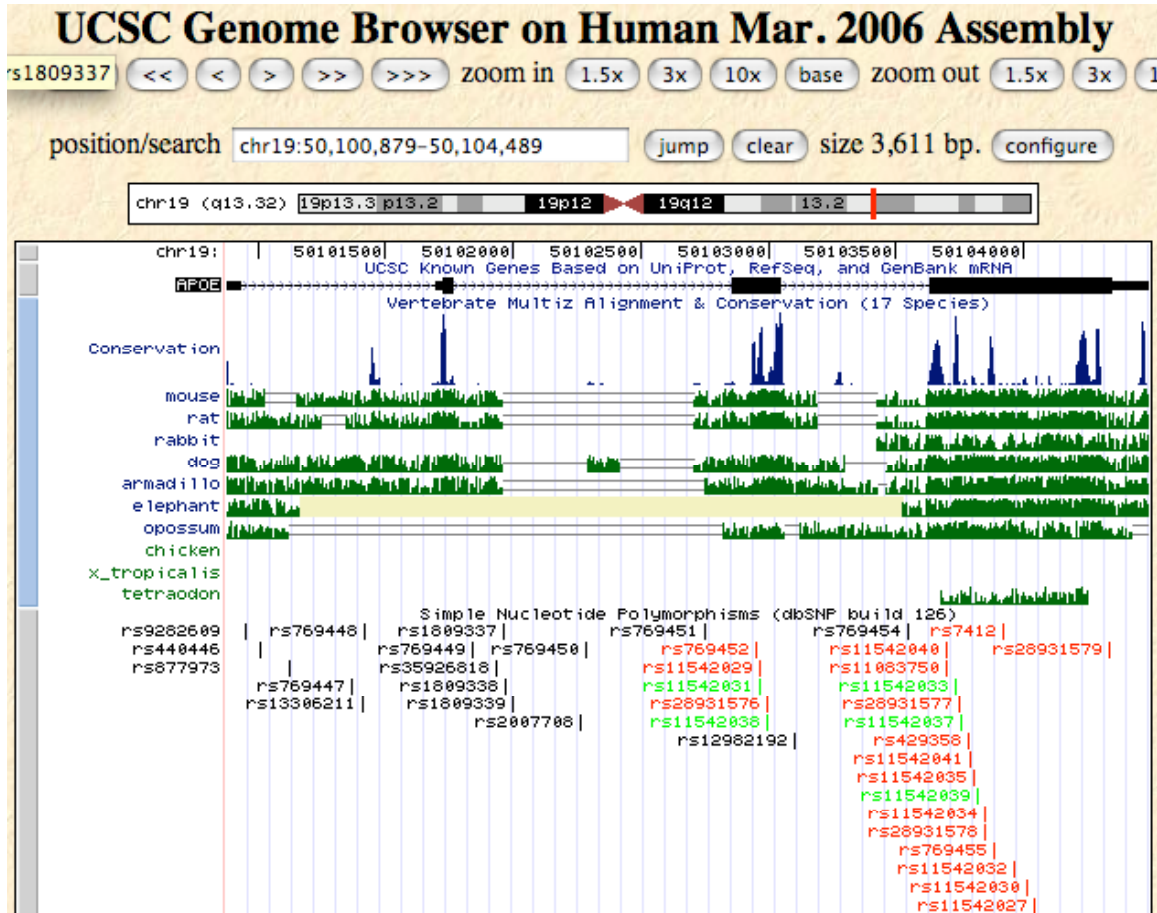
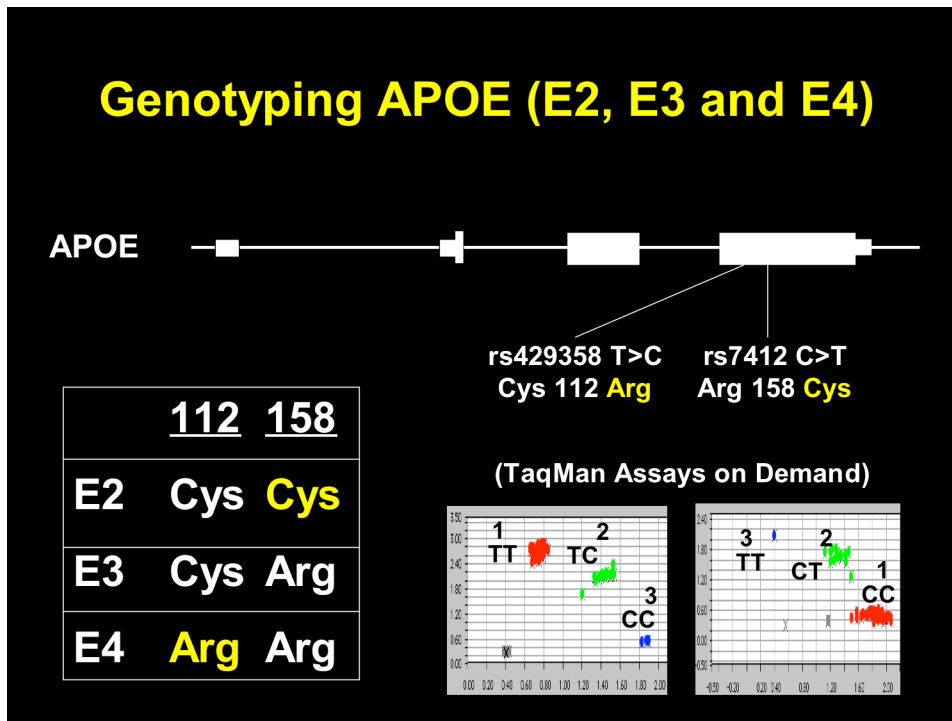


Figure 194 Genotyping APOE (E2, E3 and E4)



1.22.2 APOE Results

The [APOE rs7412](#) minor T allele was significantly enriched in the white non-controller group (CVL-4) during test1 (permuted-p ≤ 0.0216) and test2 (permuted-p ≤ 0.0103) (Tables pages 277 - 280, 285 - 288). The *white* noncontrollers were enriched in the [rs7412](#) T allele and the [APOE](#) $\epsilon 2$ allele in during test1 (permuted-p ≤ 0.0216) and test2 (permuted-p ≤ 0.0103) (Tables pages 336 - 338). This may be the first report of an association between the [APOE](#) $\epsilon 2$ allele and increased viremia in HIV-1-infected adults. No association was identified between $\epsilon 4$ and viremia. The *rsquare* and *dprime* (Definitions page 209) linkage disequilibrium values across the APOE locus were calculated (Figures pages 339 - 341).

Table 36 APOE Haplotypes Test1 (CVL-1/2/3 vs CVL-4)

test1	CVL4 vs CVL1/2/3		hap1	hap2	hap3	hap4	
			E3	E4	E2		
			1	2	3	4	
APOE_T>C	rs429358		T	C	T	C	
APOE_C>T	rs7412		C	C	T	T	
afam	control	hap	115.0	31.0	18.0		
	control	non-hap	49.0	133.0	146.0		82.0
	case	hap	101.9	33.0	17.1		
	case	non-hap	50.1	119.0	134.9		76.0
	ave	freq	0.686	0.203	0.111		
	control	freq	0.701	0.189	0.110		
	case	freq	0.670	0.217	0.113		
	dif	case-control	-0.0310	0.0280	0.0030		0.0000
		p	0.5518	0.5349	0.9324		
	2x2	chi	0.3540	0.3850	0.0070		
		perm-p	0.8533	0.9921	0.8420		
white	control	hap	275.0	62.0	49.0		
	control	non-hap	111.0	324.0	337.0		193.0
	case	hap	142.6	33.4	10.0		
	case	non-hap	43.4	152.6	176.0		93.0
	ave	freq	0.730	0.167	0.103		
	control	freq	0.712	0.161	0.127		
	case	freq	0.767	0.179	0.054		
	dif	control-case	0.0550	0.0180	-0.0730		0.0000
		p	0.1699	0.5721	0.0070		
	2x2	chi	1.8840	0.3190	7.2660		
		perm-p	0.3705	0.8334	0.0216		
hislat	control	hap	53.0	9.0	4.0		
	control	non-hap	13.0	57.0	62.0		33.0
	case	hap	25.7	6.3	4.0		
	case	non-hap	10.3	29.7	32.0		18.0
	ave	freq	0.771	0.150	0.078		
	control	freq	0.803	0.136	0.061		
	case	freq	0.713	0.176	0.111		
	dif	control-case	-0.0900	0.0400	0.0500		0.0000
		p	0.3017	0.5950	0.3646		
	2x2	chi	1.0670	0.2830	0.8220		
		perm-p	0.5862	0.9074	0.7487		

Table 37 APOE Haplotypes Test2 (CVL-1/2 vs CVL-4)

test2	CVL4 vs CVL1/2		hap1	hap2	hap3	hap4	
			E3	E4	E2		
			1	2	3	4	
APOE_T>C	rs429358		T	C	T	C	
APOE_C>T	rs7412		C	C	T	T	
afam							
	control	hap	115.0	31.0	18.0		
	control	non-hap	49.0	133.0	146.0		82.0
	case	hap	83.9	28.0	14.1		
	case	non-hap	42.1	98.0	111.9		63.0
	ave	freq	0.686	0.203	0.111		
	control	freq	0.701	0.189	0.110		
	case	freq	0.666	0.222	0.112		
	dif	case-control	-0.0350	0.0330	0.0020	0.0000	
		p	0.5167	0.4864	0.9473		
	2x2	chi	0.4200	0.4850	0.0040		
		perm-p	0.8299	0.7990	1.0000		
white							
	control	hap	275.0	62.0	49.0		
	control	non-hap	111.0	324.0	337.0		193.0
	case	hap	114.0	30.0	6.0		
	case	non-hap	36.0	120.0	144.0		75.0
	ave	freq	0.726	0.172	0.103		
	control	freq	0.712	0.161	0.127		
	case	freq	0.760	0.200	0.040		
	dif	control-case	0.0480	0.0390	-0.0870	0.0000	
		p	0.2678	0.2777	0.0029		
	2x2	chi	1.2280	1.1780	8.8680		
		perm-p	0.5050	0.5381	0.0103		
hislat							
	control	hap	53.0	9.0	4.0		
	control	non-hap	13.0	57.0	62.0		33.0
	case	hap	25.7	6.3	4.0		
	case	non-hap	10.3	29.7	32.0		18.0
	ave	freq	0.771	0.150	0.078		
	control	freq	0.803	0.136	0.061		
	case	freq	0.713	0.176	0.111		
	dif	control-case	-0.0900	0.0400	0.0500	0.0000	
		p	0.3017	0.5950	0.3646		
	2x2	chi	1.0670	0.2830	0.8220		
		perm-p	0.5855	0.9134	0.7498		

Table 38 APOE Haplotypes Test3 (CVL-1 vs CVL-4)

test3	CVL4 vs CVL1		hap1	hap2	hap3	hap4	
			E3	E4	E2		
			1	2	3	4	
APOE_T>C	rs429358		T	C	T	C	
APOE_C>T	rs7412		C	C	T	T	
afam	control	hap	115.0	31.0	18.0		
	control	non-hap	49.0	133.0	146.0		82.0
	case	hap	57.0	17.0	8.0		
	case	non-hap	25.0	65.0	74.0		41.0
	ave	freq	0.699	0.195	0.106		
	control	freq	0.701	0.189	0.110		
	case	freq	0.695	0.207	0.098		
	dif	case-control	-0.0060	0.0180	-0.0120	0.0000	
		p	0.9217	0.7329	0.7693		
	2x2	chi	0.0100	0.1160	0.0860		
		perm-p	1.0000	0.9580	0.9780		
white	control	hap	275.0	62.0	49.0		
	control	non-hap	111.0	324.0	337.0		193.0
	case	hap	72.0	17.0	5.0		
	case	non-hap	22.0	77.0	89.0		47.0
	ave	freq	0.723	0.165	0.113		
	control	freq	0.712	0.161	0.127		
	case	freq	0.766	0.181	0.053		
	dif	control-case	0.0540	0.0200	-0.0740	0.0000	
		p	0.2985	0.6353	0.0424		
	2x2	chi	1.0810	0.2250	4.1180		
		perm-p	0.5691	0.9177	0.1050		
hislat	control	hap	53.0	9.0	4.0		
	control	non-hap	13.0	57.0	62.0		33.0
	case	hap	13.6	6.4	2.0		
	case	non-hap	8.4	15.6	20.0		11.0
	ave	freq	0.757	0.175	0.068		
	control	freq	0.803	0.136	0.061		
	case	freq	0.619	0.290	0.091		
	dif	control-case	-0.1840	0.1540	0.0300	0.0000	
		p	0.0818	0.1008	0.6253		
	2x2	chi	3.0280	2.6930	0.2380		
		perm-p	0.1596	0.1956	1.0000		

Figure 195 Linkage Disequilibrium (D') for Two APOE SNPs in African American Test1

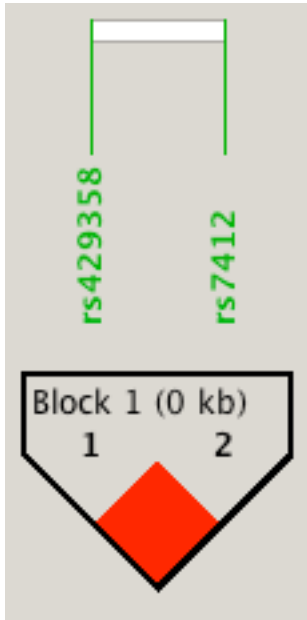


Figure 196 Linkage Disequilibrium (R-squared) for Two APOE SNPs in African American Test1

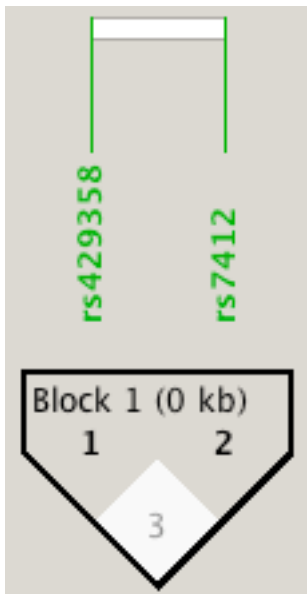


Figure 197 Linkage Disequilibrium (D') for Two APOE SNPs in White American Test1

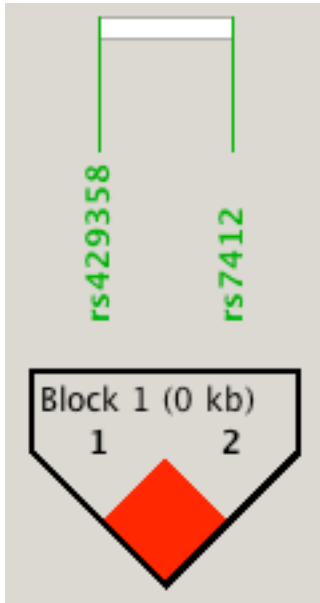


Figure 198 Linkage Disequilibrium (R-squared) for Two APOE SNPs in White American Test1

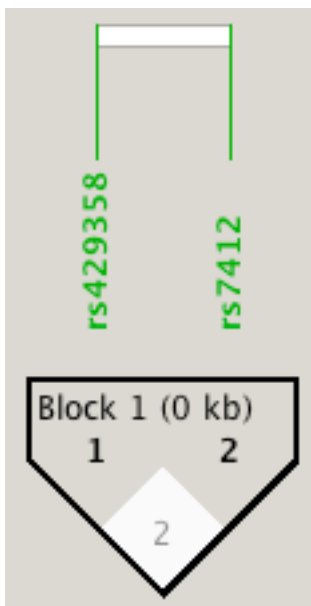


Figure 199 Linkage Disequilibrium (D') for Two APOE SNPs in His/Lat American Test1

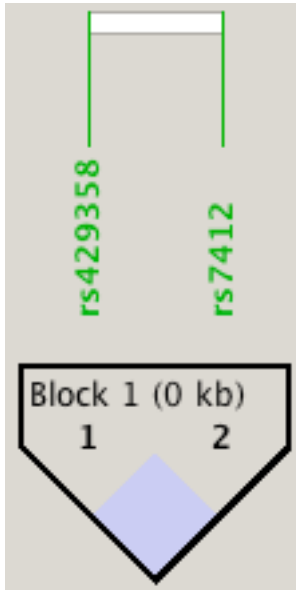
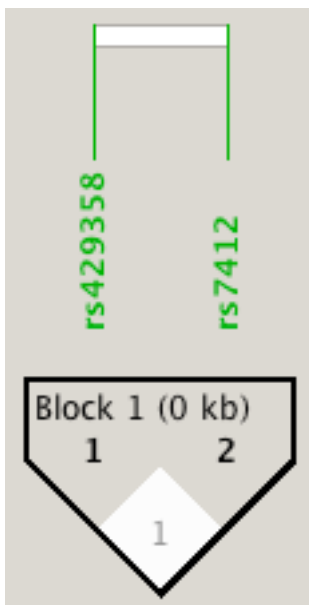


Figure 200 Linkage Disequilibrium (R-squared) for Two APOE SNPs in His/Lat American Test1



1.23 Conclusions for AIM 2

As a general summary for AIM 2 I conclude, TLR9 and IRF5 variants associated with HIV viremia levels in White Americans. Additionally, individuals infected with HIV should try to avoid chronic inflammation, which means avoiding other viral and bacteria coinfections, traumas, and other behaviors that promote a chronic inflammatory state. Furthermore, the magnitude of TLR9- and IRF5-dependant inflammatory responses during the acute phase of HIV infection may partially determine the viremia level of chronic infection (CVL classification).

1.23.1 CCR5 Conclusions

Conclusion 1

CCR5 was not in linkage disequilibrium with TLR9, therefore the CCR5 and TLR9 loci can be treated independently

Conclusion 2

The CCR5 del32 allele associated with lower HIV viremia in African-Americans. This finding is reasonable. Individuals with a single copy of the del32 allele are partially protected from HIV infection, and would be expected to have lower viremia levels. This finding is also interesting, because del32 is usually absent in African populations, so this may be considered evidence of admixture

Conclusion 3

CCR5 del32 homozygotes were not found in the HIV cohorts. This was expected because having a single copy of the del32 allele reduces the risk of HIV infection, and

having 2 copies makes individuals highly resistant to infection. It follows that someone highly resistant to infection would never have been infected in the first place, nor made it into our HIV-infected cohort.

Conclusion 3

The frequencies of the CCR5 genotype in the HIV-1 cohort were published (Chapter 3: Prevalence of CXCR4 tropism among antiretroviral-treated HIV-1-infected patients with detectable viremia, page 345) [57].

1.23.2 TLR9 Conclusions

Conclusion 1

The minor alleles for two polymorphisms (TLR9 rs352140 G>A, rs352139G>A), and TLR9 haplotype 1 associated with higher HIV viremia, in White Americans. These are same markers have been previously associated with higher rates of CD4 positive T-cell loss, in a population of 12,000 HIV-infected Swiss [20]. In Swiss adults, the [TLR9 rs352140](#) A>G ($p \leq 0.0005$) and the TLR9 [rs352139](#) G>A alleles ($p \leq 0.0007$) were significantly more frequent in a group of 'rapid progressors' as classified by measuring the loss of CD4+ T cells over time [20]. Moreover, [TLR9](#) haplotype 1 ([rs5743836](#) T>C, [rs352139](#) G>A, [rs352140](#) A>G) was also more frequent in the rapid progressors than in the controls ($p \leq 0.001$). I conclude, [TLR9](#) haplotype 1 is associated with both a higher level of HIV viremia (present study) and a more rapid loss of CD4⁺ T cells (previous study) in HIV-infected individuals Caucasians.

Conclusion 2

This is an independent validation of the association between TLR9 and an HIV/AIDS phenotype.

1.23.3 IRF5 Conclusions

Conclusion 1

IRF5 haplotype 2 associated with lower HIV viremia in White Americans.

Conclusion 2

IRF5 haplotypes 1 and 2 may have different associations to HIV viremia levels.

Additionally, these haplotypes differ by an insertion/deletion polymorphism in exon 6, an indel that may cause the mature IRF5 protein to have, or to lack a 10 amino acid PEST domain.

Conclusion 3

The magnitude of an IRF5-mediated innate immune response, and inflammation, may determine the survival time during HIV infection.

1.23.4 APOE Conclusions

Conclusion 1

The APOE epsilon 2 allele (E2) associated with higher viremia in White Americans.

Chapter 3: Prevalence of CXCR4 tropism among antiretroviral-treated HIV-1-infected patients with detectable viremia

BRIEF REPORT

Prevalence of CXCR4 Tropism among Antiretroviral-Treated HIV-1-Infected Patients with Detectable Viremia

Peter W. Hunt,¹ P. Richard Harrigan,^{10,11} Wei Huang,⁹ Michael Bates,⁹ David W. Williamson,^{3,8} Joseph M. McCune,³ Richard W. Price,² Serena S. Spudich,² Harry Lampiris,⁹ Rebecca Hoh,¹ Teri Leigler,¹ Jeffrey N. Martin,^{1,6,7} and Steven G. Deeks¹

¹Positive Health Program and ²Department of Neurology, San Francisco General Hospital, ³Graduate Program in Biological and Medical Informatics, ⁴Division of Experimental Medicine, Department of Internal Medicine, ⁵San Francisco Veterans Affairs Medical Center, ⁶Center for AIDS Prevention Studies, and ⁷Department of Epidemiology and Biostatistics, University of California, San Francisco, ⁸Gladstone Institute of Virology and Immunology, San Francisco, and ⁹Monogram Biosciences, Inc., South San Francisco, California; ¹⁰British Columbia Centre for Excellence in HIV/AIDS, Providence Health Care, and ¹¹Faculty of Medicine, University of British Columbia, Vancouver, Canada

Although CXCR4-tropic viruses are relatively uncommon among untreated human immunodeficiency virus (HIV)-infected individuals except during advanced immunodeficiency, the prevalence of CXCR4-tropic viruses among treated patients with detectable viremia is unknown. To address this issue, viral coreceptor usage was measured with a single-cycle recombinant-virus phenotypic entry assay in treatment-naïve and treated HIV-infected participants with detectable viremia sampled from 2 clinic-based cohorts. Of 182 treated participants, 75 (41%) harbored dual/mixed or X4-tropic viruses, compared with 178 (18%) of the 976 treatment-naïve participants ($P < .001$). This difference remained significant after adjustment for CD4⁺ T cell count and CCR5 Δ32 genotype. Enrichment for dual/mixed/X4-tropic viruses among treated participants was largely but incompletely explained by lower pretreatment nadir CD4⁺ T cell counts. CCR5 inhibitors may thus be best strategically used before salvage therapy and before significant CD4⁺ T cell depletion.

With CCR5 inhibitors, a new class of antiretroviral medications targeting HIV-1 entry, in phase 3 clinical trials, there is renewed

interest in the prevalence of viruses capable of using the alternative CXCR4 coreceptor (X4) for entry. Among untreated patients, R5-tropic viruses predominate during most stages of HIV infection [1], accounting for 82% of isolates from treatment-naïve patients initiating antiretroviral therapy in a recent population-based study [2]. However, the prevalence of viruses that can use the CXCR4 coreceptor for entry approaches almost 50% among untreated individuals with advanced immunodeficiency [1, 2].

Comparatively little is known about the prevalence of X4-tropic viruses in antiretroviral-treated patients with detectable viremia. Because CCR5 inhibitors are likely to be initially used in treated patients with resistance to currently available antiretroviral medications, it is important to assess the prevalence of CXCR4 tropism in this population. There are several reasons to believe that antiretroviral therapy might alter the prevalence of X4-tropic viruses. First, partially suppressive therapy may lead to an increase in HIV-specific T cell responses [3]. Because X4-tropic variants may be more susceptible to cytotoxic T cell responses than R5-tropic viruses [4], increases in HIV-specific T cell responses during partial treatment-mediated viral suppression might select against X4-tropic viruses. Second, antiretroviral therapy reduces CCR5 expression on T cells, presumably as a consequence of reductions in T cell activation [5, 6], potentially selecting for X4-tropic viruses [2]. Last, certain antiretroviral drugs may preferentially select for one virus population, either because of enhanced activity against X4 viruses (as has been suggested for enfuvirtide [7]) or because of suboptimal drug metabolism in the cellular reservoirs for X4 viruses (as has been suggested for zidovudine [8]). To assess the potential impact that partial treatment-mediated viral suppression has on the prevalence of CXCR4 tropism, we compared the prevalence of X4-tropic viruses between treatment-naïve participants and treated participants with detectable plasma HIV RNA levels.

Participants and methods. Antiretroviral-treated partici-

Presented in part: 13th Conference on Retroviruses and Opportunistic Infections, Denver, 5–8 February 2006 (abstract 43).

Potential conflicts of interest: M.B. and W.H. are employees of Monogram, Inc. S.G.D. has received honoraria from Monogram Biosciences. For all other authors, no conflicts are reported.

Financial support: University-Wide AIDS Research Program (grant CC99-SF-001); University of California, San Francisco/Gladstone Institute of Virology and Immunology Center for AIDS Research (grants P30 AI27763 and P30 MH59037); National Institutes of Health (NIH), grants R01 AI52745, R37 AI40312, NS 37660, and K23 AI65244; General Clinical Research Center at San Francisco General Hospital (grant 5-M01-RR00083-37); Center for AIDS Prevention Studies (grant P30 MH62246); Burroughs Wellcome Fund (Clinical Scientist Award in Translational Research to J.M.M.); NIH Director's Pioneer Award Program, NIH Roadmap for Medical Research (grant DPI 0D00329 to J.M.M.).

Received 15 March 2006; accepted 17 May 2006; electronically published 29 August 2006.
Reprints or correspondence: Dr. Peter W. Hunt, Positive Health Program, San Francisco General Hospital, Bldg. 80, Ward 84, 935 Potrero Ave., San Francisco, CA 94110 (phunt@pdp.ucsf.edu).

The Journal of Infectious Diseases 2006;194:326–30
© 2006 by the Infectious Diseases Society of America. All rights reserved.
0022-1899/2006/19407-0003\$15.00

pants were sampled from the Study of the Consequences of the Protease Inhibitor Era (SCOPE), an ongoing clinic-based cohort of >600 chronically HIV-infected patients in San Francisco. Participants are seen every 4 months, at which time an extensive evaluation is performed and biologic specimens are obtained. Additional treated participants were sampled from a previously reported cross-sectional study comparing tropism of plasma and cerebrospinal fluid viral isolates [9]. Treated participants were eligible for the current analysis if they were receiving a stable antiretroviral regimen for >4 months and had detectable plasma HIV RNA levels (>50 copies/mL). Participants with exposure to coreceptor antagonists were excluded. Antiretroviral-naïve participants were sampled, before starting antiretroviral therapy, from the Highly Active Antiretroviral Therapy Observational Medical Evaluation and Research (HOMER) cohort, a population-based cohort of HIV-1-infected patients within the British Columbia Centre for Excellence in HIV/AIDS network. Coreceptor tropism was measured, in all of these antiretroviral-naïve participants with detectable plasma HIV RNA levels, as reported elsewhere [2]. Informed consent was obtained from all participants, and ethics approval was obtained from the ethics boards of each institution.

Plasma HIV RNA levels were determined by the branched DNA (bDNA) amplification technique (Quantiplex HIV RNA, version 3.0; Chiron) for SCOPE participants and by polymerase chain reaction (PCR) (Cobas AmpliCor HIV-1 Monitor Test, version 1.5; Roche Diagnostics) for HOMER participants. For treated participants, the pretreatment nadir CD4⁺ T cell count was the lowest self-reported value before the initiation of the current regimen. For treated participants, resistance to protease inhibitors and reverse transcriptase inhibitors was measured within 6 months of the coreceptor tropism measurement, using the TRUGENE HIV-1 Genotyping Kit (Bayer HealthCare Diagnostics); was analyzed with OpenGene software (Visible Genetics); and was interpreted on the basis of 2004 International AIDS Society–USA guidelines [10].

CCR5 Δ32 genotype was determined by extracting DNA from whole blood or peripheral-blood mononuclear cells by use of the QIAamp Blood Midi Kit (Qiagen). Extracted DNA was amplified in a single round of PCR using primers flanking the *CCR5* Δ32 region (forward, 5'-TCAAAAAGAAGTCTTCA-TTACACC-3'; reverse, 5'-AGCCCAGAAGAGAAAATAACA-ATC-3'). PCR products were visualized by electrophoresis on a 3% agarose gel and classified on the basis of fragment length (241 bp for wild type and 209 bp for the *CCR5* Δ32 allele).

The PhenoSense HIV-1 coreceptor use [9]. Briefly, participant-derived *env* DNA (gp160) was amplified by PCR from plasma isolates and ligated into pCXAS expression vectors. A replication-defective retroviral vector containing a luciferase expression cassette inserted within the *env* gene was used to co-

transfect human embryonic kidney cell cultures with the sample plasmid DNA. Recombinant viruses were harvested after 48 h and were assessed for their ability to infect cells expressing CCR5 or CXCR4 by measuring luciferase activity in the presence of coreceptor-specific inhibitors. The PhenoSense assay classifies isolates as R5-, X4-, or dual and/or mixed-tropic virus.

Plasma HIV RNA levels obtained by use of Cobas AmpliCor (version 1.5) were converted to bDNA (Quantiplex HIV RNA, version 3.0) equivalents by subtracting 0.3 log₁₀ copies/mL [11]. For analysis, tropism was dichotomized as either R5- or dual/mixed/X4-tropic, given the small number of purely X4-tropic isolates. Among treated participants, the treatment-mediated change in CD4⁺ T cell count was defined as the difference between the current CD4⁺ T cell count (at the time of tropism measurement) and the pretreatment nadir CD4⁺ T cell count. Factors associated with tropism were assessed in unadjusted analyses with Fisher's exact tests and in adjusted multivariable logistic regression or stratified analyses. Both current and pretreatment nadir CD4⁺ T cell counts, plasma HIV RNA levels, and *CCR5* Δ32 genotype were considered as potential confounding factors/mediators. Backward stepwise model selection was used for multivariable logistic regression models, retaining factors that altered the association between treatment status and tropism by at least 10%.

Results. Compared with the 976 treatment-naïve participants, the 182 treated participants had lower plasma HIV RNA levels and pretreatment nadir CD4⁺ T cell counts, but the majority in each group had a current CD4⁺ T cell count >250 cells/mm³ (table 1). A similar percentage of participants in each group were heterozygous for the *CCR5* Δ32 mutation. Most of the treated participants were receiving a protease inhibitor-based regimen and had a moderate number of drug-resistance mutations. Only 10% of treated participants were receiving the HIV-entry inhibitor enfuvirtide.

Compared with treatment-naïve participants, a higher percentage of treated participants were harboring dual/mixed/X4-tropic viruses (41% vs. 18%; $P < .001$). Only 1 participant in each group apparently harbored a purely X4-tropic virus population, so these were included with participants harboring dual/mixed-tropic viruses in subsequent analyses. Although the prevalence of dual/mixed/X4 tropism was higher at lower CD4⁺ T cell counts ($P < .001$ for trend), the treated participants had a higher prevalence of dual/mixed/X4 tropism at any given CD4⁺ T cell count ($P < .05$ within each CD4⁺ T cell count stratum) (figure 1A).

Among 150 participants heterozygous for the *CCR5* Δ32 mutation, 48 (32%) harbored dual/mixed/X4-tropic virus, compared with only 193 (20%) of 976 participants without this mutation ($P = .001$). In unadjusted analyses, higher plasma HIV RNA levels were associated with dual/mixed/X4 tropism among treatment-naïve participants ($P < .001$) and treated par-

Table 1. Characteristics of 1158 chronically HIV-infected participants with detectable viremia.

Characteristic	Treatment-naive participants (n = 976)	Treated participants (n = 182)
Age, years	37 (32–44)	45 (41–52)
Male, no. (%)	835 (86)	159 (87)
CD4 ⁺ T cell count, cells/mm ³	260 (120–420)	258 (134–365)
Nadir CD4 ⁺ T cell count, cells/mm ³	260 (120–420)	60 (17–176)
Treatment-mediated change in CD4 ⁺ T cell count, ^a cells/mm ³	...	+153 (57–273)
Plasma HIV RNA level, copies/mL	4.8 (4.4–5.2)	3.6 (3.1–4.3)
<i>CCR5</i> Δ32 heterozygous, ^b no. (%)	128 (13)	22 (14)
Currently used antiretroviral medications, no. (%)		
Zidovudine or stavudine	...	123 (68)
Didanosine	...	28 (15)
Nonnucleoside reverse transcriptase inhibitors	...	39 (21)
Protease inhibitors	...	139 (76)
Enfuvirtide (T-20)	...	19 (10)
Drug-resistance mutations by class, ^c no.		
Nucleoside reverse transcriptase inhibitors	...	4 (1–7)
Nonnucleoside reverse transcriptase inhibitors	...	1 (0–1)
Protease inhibitors, minor mutations	...	4 (0–6)
Protease inhibitors, major mutations	...	1 (0–3)

NOTE. Data are median (interquartile range) of values, unless otherwise indicated.

^a The treatment-mediated change in CD4⁺ T cell count was defined as the difference between the current and pretreatment nadir CD4⁺ T cell counts.

^b *CCR5* genotyping was available for 964 of 972 treatment-naive and 161 of 182 treated participants.

^c Drug-resistance data were available within 6 months of the tropism measurement for 152 of the 182 treated participants. Drug-resistance mutations were defined on the basis of International AIDS Society–USA guidelines (October 2004). Minor protease inhibitor mutations were considered to be present only in the setting of major protease inhibitor mutations.

participants ($P = .05$). However, among either treatment-naive or treated participants, there was no longer evidence for an independent association between plasma HIV RNA levels and coreceptor tropism after adjusting for CD4⁺ T cell count ($P > .85$, for each association). Even after adjusting for current CD4⁺ T cell count and *CCR5* Δ32 genotype, antiretroviral-therapy use continued to be associated with 4-fold increased odds of dual/mixed/X4 tropism (95% confidence interval, 2.7-fold to 5.8-fold; $P < .001$).

Lower pretreatment nadir CD4⁺ T cell counts were also associated with a higher prevalence of dual/mixed/X4 tropism for both treatment-naive ($P < .001$ for trend) and treated participants ($P = .03$ for trend) (figure 1B). After stratifying by pretreatment nadir CD4⁺ T cell count, there was no longer any evidence for an independent association between antiretroviral therapy and dual/mixed/X4 tropism among those with pretreatment nadir CD4⁺ T cell counts in the lowest 3 quartiles (≤ 176 cells/mm³) ($P > .47$ within each stratum). Thus, the enrichment for dual/mixed/X4 tropism among the majority of treated participants in our sample was largely explained by low pretreatment nadir CD4⁺ T cell counts, even though the nadir

had occurred a median of 5 years earlier (interquartile range [IQR], 1–7 years) and CD4⁺ T cell counts had increased by a median of 148 cells/mm³ (IQR, +79 to +257 cells/mm³) during that time. However, among those with pretreatment nadir CD4⁺ T cell counts in the highest quartile (>176 cells/mm³), treatment continued to be associated with a higher prevalence of dual/mixed/X4 tropism, independent of the pretreatment nadir CD4⁺ T cell count ($P < .001$). This difference remained significant even if the analysis was restricted to those with nadirs <350 cells/mm³ ($P < .001$).

We next assessed whether the enrichment for dual/mixed/X4-tropic viruses among treated participants might be explained by exposure to specific drugs. Prior studies indicated that thymidine analogs (zidovudine or stavudine) select for syncytium-inducing (and presumably CXCR4-using) viruses in vivo, because these drugs are less likely to be phosphorylated in the cellular reservoirs for these viruses [8]. However, among the 182 antiretroviral-treated participants, we found no evidence for an association between current thymidine-analog use and tropism ($P = .63$). Enfuvirtide may also interfere with gp120 binding to CXCR4 but not CCR5, potentially selecting

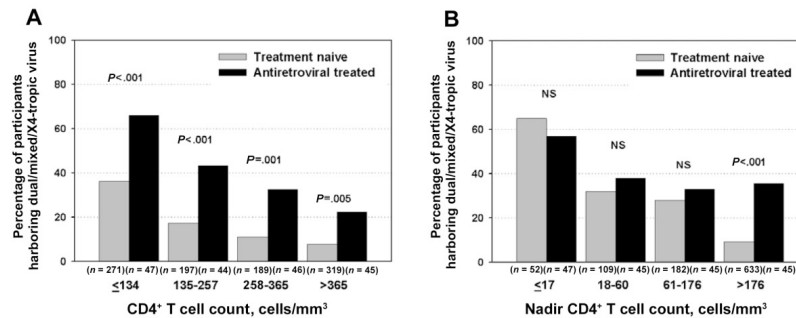


Figure 1. Prevalence of dual/mixed/X4 tropism by treatment status and either current or pretreatment nadir CD4⁺ T cell count. The percentage of participants harboring dual/mixed or X4 tropic viruses is plotted by quartiles of current CD4⁺ T cell count and pretreatment nadir CD4⁺ T cell count among 976 treatment-naive (gray bars) and 182 treated (black bars) participants with chronic HIV infection. *A*, Higher prevalence of dual/mixed/X4 tropism at lower current CD4⁺ T cell counts ($P<.001$ for trend) and higher prevalence of dual/mixed/X4 tropism for treated participants at any given current CD4⁺ T cell count ($P<.01$ within each CD4⁺ T cell count stratum). *B*, Association of lower pretreatment nadir CD4⁺ T cell counts with a higher prevalence of dual/mixed/X4 tropism for both treatment-naive ($P<.001$ for trend) and treated participants ($P=.03$ for trend). Among participants with CD4⁺ T cell count nadirs in the lowest 3 quartiles (≤ 176 cells/mm³), there was no evidence for an association between treatment status and dual/mixed/X4 tropism after controlling for pretreatment nadir CD4⁺ T cell count. However, among participants with pretreatment nadir CD4⁺ T cell counts >176 cells/mm³, treatment was independently associated with a higher prevalence of dual/mixed/X4 tropism ($P<.001$). $P<.001$ for interaction. NS, not significant.

for R5-tropic variants [7]. Although underpowered, our study provided no evidence for an association between enfuvirtide use and tropism in our treated participants ($P=.63$).

Discussion. Although much is known about the prevalence of CXCR4 tropism among patients with untreated HIV-1 infection, comparatively little is known about the role that CXCR4 tropism plays in treated patients with drug-resistant viremia. In the present study, we demonstrate that treated participants with detectable viremia, regardless of antiretroviral regimen, have 4-fold greater odds of dual/mixed/X4 tropism than treatment-naive participants, independent of CD4⁺ T cell count and *CCR5* $\Delta 32$ genotype. Although most of this enrichment for dual/mixed/X4-tropic viruses appears to be explained by lower pretreatment nadir CD4⁺ T cell counts, X4-tropic variants apparently persist despite treatment-mediated restoration of peripheral CD4⁺ T cell counts.

Our estimate of an unusually high prevalence of dual/mixed/X4 tropic viruses among treated participants with detectable viremia in the present study is consistent with other recent reports. For example, one-third to one-half of treatment-experienced patients enrolling in clinical trials harbor dual/mixed/X4-tropic viruses [12, 13]. Also, despite previous reports of rare CXCR4 tropism among untreated individuals infected with HIV-1 subtype C, 50% of treated patients infected with drug-resistant HIV-1 subtype C harbored viruses capable of using CXCR4 for entry in one small study [14].

It remains unclear why treated patients with drug-resistant

viremia remain enriched for dual/mixed/X4-tropic viruses despite treatment-mediated gains in peripheral CD4⁺ T cell counts. The association between *CCR5* $\Delta 32$ heterozygosity and dual/mixed/X4 tropism suggests a potential role of *CCR5*⁺ target cell availability [2]. Persistent loss of *CCR5*⁺ T cell targets during antiretroviral therapy might be explained by poor reconstitution of gut-associated lymphoid tissue, where most *CCR5*⁺ T cell targets reside, despite robust peripheral treatment-mediated CD4⁺ T cell gains [15]. Alternatively, treatment-mediated viral suppression may cause declines in *CCR5* expression [5, 6], potentially explaining the enrichment for dual/mixed/X4-tropic viruses observed among participants with high pretreatment nadir CD4⁺ T cell counts.

This study has some limitations that deserve comment. First, because the tropism measurement used in the present study cannot distinguish between mixtures and dual-tropic viral populations or quantify the relative proportion of X4-tropic viruses within a given dual/mixed sample, it is possible that the high prevalence of dual/mixed-tropic viruses in treated participants is driven by participants with clinically insignificant amounts of X4-tropic virus replication. However, luciferase activity in CXCR4-expressing cells was comparable among dual/mixed-tropic isolates from both groups (data not shown). In addition, because this study was cross-sectional and potentially susceptible to selection bias, we can only speculate as to the timing and causes of tropism switches among treated participants. Longitudinal studies of participants initiating antiretroviral

therapy before the onset of virologic failure will be necessary to specifically address these issues.

In summary, we have observed an unexpectedly high prevalence of dual/mixed/X4 tropism among treated participants with detectable viremia. Consequently, treated participants with drug resistance may be less likely than treatment-naïve participants to achieve viral suppression while receiving CCR5 inhibitors. If these results are corroborated by other studies, CCR5 inhibitors may be better strategically used before salvage therapy and before advanced immunodeficiency.

References

1. Schuitemaker H, Koot M, Kootstra NA, et al. Biological phenotype of human immunodeficiency virus type 1 clones at different stages of infection: progression of disease is associated with a shift from monocyto-tropic to T-cell-tropic virus population. *J Virol* **1992**; 66:1354–60.
2. Brumme ZL, Goodrich J, Mayer HB, et al. Molecular and clinical epidemiology of CXCR4-using HIV-1 in a large population of anti-retroviral-naïve individuals. *J Infect Dis* **2005**; 192:466–74.
3. Deeks SG, Martin JN, Sinclair E, et al. Strong cell-mediated immune responses are associated with the maintenance of low-level viremia in antiretroviral-treated individuals with drug-resistant human immunodeficiency virus type 1. *J Infect Dis* **2004**; 189:312–21.
4. Harouse JM, Buckner C, Gettie A, et al. CD8+ T cell-mediated CXC chemokine receptor 4-similar/human immunodeficiency virus suppression in dually infected rhesus macaques. *Proc Natl Acad Sci USA* **2003**; 100:10977–82.
5. Andersson J, Fehniger TE, Patterson BK, et al. Early reduction of immune activation in lymphoid tissue following highly active HIV therapy. *AIDS* **1998**; 12:F123–9.
6. Giovannetti A, Ensofi F, Mazzetta F, et al. CCR5 and CXCR4 chemokine receptor expression and beta-chemokine production during early T cell repopulation induced by highly active anti-retroviral therapy. *Clin Exp Immunol* **1999**; 118:87–94.
7. Yuan W, Craig S, Si Z, Farzan M, Sodroski J. CD4-induced T-20 binding to human immunodeficiency virus type 1 gp120 blocks interaction with the CXCR4 coreceptor. *J Virol* **2004**; 78:5448–57.
8. Boucher CA, Lange JM, Miedema FF, et al. HIV-1 biological phenotype and the development of zidovudine resistance in relation to disease progression in asymptomatic individuals during treatment. *AIDS* **1992**; 6:1259–64.
9. Spudich SS, Huang W, Nilsson AC, et al. HIV-1 chemokine coreceptor utilization in paired cerebrospinal fluid and plasma samples: a survey of subjects with viremia. *J Infect Dis* **2005**; 191:890–8.
10. Johnson VA, Brun-Vezinet F, Clotet B, et al. Update of the drug resistance mutations in HIV-1: 2004. *Top HIV Med* **2004**; 12:119–24.
11. Elbeik T, Alvord WG, Trichavaroj R, et al. Comparative analysis of HIV-1 viral load assays on subtype quantification: Bayer Versant HIV-1 RNA 3.0 versus Roche Amplicor HIV-1 Monitor version 1.5. *J Acquir Immune Defic Syndr* **2002**; 29:330–9.
12. Demarest JF, Bonny T, Vavro CL. HIV-1 co-receptor tropism in treatment naïve and experienced subjects [abstract H-1136]. In: Program and abstracts of the 44th Interscience Conference on Antimicrobial Agents and Chemotherapy (Washington, DC). Herndon, VA: ASM Press, **2004**:300.
13. Wilkin T, Su Z, Kuritzkes D, et al. Co-receptor tropism in patients screening for ACTG 5211, a phase 2 study of Vicriviroc, a CCR5 inhibitor [abstract 655]. In: Program and abstracts of the 13th Conference on Retroviruses and Opportunistic Infections (Denver). Alexandria, VA: Foundation for Retrovirology and Human Health, **2006**:283.
14. Johnston ER, Zijenah LS, Mutetwa S, Kantor R, Kittinunvorakoon C, Katzenstein DA. High frequency of syncytium-inducing and CXCR4-tropic viruses among human immunodeficiency virus type 1 subtype C-infected patients receiving antiretroviral treatment. *J Virol* **2003**; 77:7682–8.
15. Guadalupe M, Reay E, Sankaran S, et al. Severe CD4+ T-cell depletion in gut lymphoid tissue during primary human immunodeficiency virus type 1 infection and substantial delay in restoration following highly active antiretroviral therapy. *J Virol* **2003**; 77:11708–17.

Chapter 4: Apolipoprotein A-V: a potential modulator of plasma triglyceride levels in Turks

Apolipoprotein A-V: a potential modulator of plasma triglyceride levels in Turks[□]

Uğur Hodoğlugil,^{*,†} Sinan Tanyolac,^{*,§} David W. Williamson,^{*,**} Yadong Huang,^{*,††} and Robert W. Mahley^{1,*†,††,§§}

Gladstone Institute of Cardiovascular Disease,^{*} Cardiovascular Research Institute,[†] Graduate Program in Biological and Medical Informatics,^{**} and Departments of Pathology^{††} and Medicine,^{§§} University of California, San Francisco, San Francisco, CA; and Faculty of Medicine,[§] Department of Internal Medicine, Division of Endocrinology and Metabolism, Istanbul University, Istanbul, Turkey

Abstract The apolipoprotein A-V gene (*APOA5*) plays an important role in determining plasma triglyceride levels. We studied the effects of *APOA5* polymorphisms on plasma triglyceride levels in Turks, a population with low levels of HDL cholesterol and a high prevalence of coronary artery disease. We found 15 polymorphisms, three of which were novel. Seven haplotype-tagging single nucleotide polymorphisms (SNPs) were chosen and genotyped in ~3,000 subjects. The rare alleles of the -1464T>C, -1131T>C, S19W, and 1259T>C SNPs were significantly associated with increased triglyceride levels (19–86 mg/dl; $P < 0.05$) and had clear gene-dose effects. Haplotype analysis of the nine common *APOA5* haplotypes revealed significant effects on triglyceride levels ($P < 0.001$). Detailed analysis of haplotypes clearly showed that the -1464T>C polymorphism had no effect by itself but was a marker for the -1131T>C, S19W, and 1259T>C polymorphisms. The -1131T>C and 1259T>C polymorphisms were in a strong but incomplete linkage disequilibrium and appeared to have independent effects. Thus, the *APOA5*-1131T>C, S19W, and 1259T>C rare alleles were associated with significant increases in plasma triglyceride levels. At least one of these alleles was present in ~40% of the Turks. Similar associations were observed for -1131T>C and S19W in white Americans living in San Francisco, California.—Hodoğlugil, U., S. Tanyolac, D. W. Williamson, Y. Huang, and R. W. Mahley. Apolipoprotein A-V: a potential modulator of plasma triglyceride levels in Turks. *J. Lipid Res.* 2006. 47: 144–153.

Supplementary key words Turkish population • polymorphism • haplotype • high density lipoprotein cholesterol

Atherogenic dyslipidemia, including hypertriglyceridemia, is a risk factor for coronary artery disease (CAD) (1, 2). Family and twin studies have shown that triglyceride levels are controlled by genetic factors, although heritability estimates vary widely (3–5). Recently, the multina-

tional Genetic Epidemiology of Metabolic Syndrome project (6) conducted a genome scan for atherogenic dyslipidemia and found significant evidence for linkage to triglyceride levels near the apolipoprotein A-V gene (*APOA5*), on chromosome 11q22, only in Turkish families (7). ApoA-V is an important regulator of plasma triglyceride levels (8, 9). Triglyceride levels are 4-fold higher in *Apoa5* knockout mice and significantly lower in transgenic mice (8) or in adenovirus-treated mice expressing human *APOA5* (9) than in wild-type mice. ApoA5 may decrease plasma triglyceride levels by increasing lipoprotein lipase activity (10, 11) and reducing hepatic levels of very low density lipoprotein triglyceride (11).

Several single nucleotide polymorphisms (SNPs) within the *APOA5* locus (-1131T>C, -3A>G, S19W, IVS3+476G>A, 1259T>C, and G185C) have been identified, and their rare alleles are associated with increased plasma triglyceride levels in different populations (8, 12–22). The -1131T>C, -3A>G, IVS3+476G>A, and 1259T>C SNPs (haplotype *APOA5**2) were in almost complete linkage disequilibrium (LD) in European populations (17, 20). Therefore, any one of these polymorphisms might serve as a marker for the others in these populations. The frequencies of the rare alleles of -1131T>C and S19W vary greatly among populations (8, 12–22). The plasma triglyceride increase associated with these rare alleles also varies, ranging from no association (20, 22) to 69% higher triglyceride levels in CC than in TT subjects with the -1131T>C polymorphism (16) and from no association (23, 24) to 20–30% higher triglyceride levels in SW than in SS subjects with the S19W polymorphism (20).

Abbreviations: *APOA5*, apolipoprotein A-V gene; BMI, body mass index; CAD, coronary artery disease; HDL-C, high density lipoprotein cholesterol; htSNP, haplotype-tagging single nucleotide polymorphism; LD, linkage disequilibrium; SNP, single nucleotide polymorphism; THS, Turkish Heart Study; UTR, untranslated region.

¹ To whom correspondence should be addressed.

e-mail: rmahley@gladstone.ucsf.edu

[□] The online version of this article (available at <http://www.jlr.org>) contains three additional tables.

Manuscript received 3 August 2005 and in revised form 20 September 2005.

Published, JLR Papers in Press, October 28, 2005.

DOI 10.1194/jlr.M500343-JLR200

Copyright © 2006 by the American Society for Biochemistry and Molecular Biology, Inc.

This article is available online at <http://www.jlr.org>

Haplotype analysis in European populations identified three common haplotypes, two of which, uniquely described by the rare alleles -1131T>C and S19W, are associated with higher triglyceride levels than the most common haplotype (18, 20, 21). Although haplotype structure and distributions were different in Chinese (15), African-Americans (22), and three different Singaporean populations (16), significant haplotype-triglyceride associations were identified.

APOA5 SNPs have also been associated with reduced high density lipoprotein cholesterol (HDL-C; -1131T>C, -3A>G, and IVS3+476A>G) (16), decreased LDL cholesterol size (1259T>C and -3A>G) (13), and increased numbers of remnant-like particles (-1131T>C and S19W) (17). The -1131T>C SNP was more frequent in CAD patients (25). Both the -1131T>C and S19W SNPs were associated with cardiovascular events (17) but not with coronary artery diameter (23).

In this study, we explored the association between *APOA5* sequence variations and plasma triglyceride levels in >3,000 participants in the Turkish Heart Study (THS), a large, cross-sectional epidemiological survey of the Turkish population (26). The *APOA5* gene was sequenced to detect polymorphisms, haplotype-tagging single nucleotide polymorphisms (htSNPs) were genotyped, and these SNPs and haplotypes were associated with significantly increased levels of triglycerides.

MATERIALS AND METHODS

Study population and biochemical analyses

The primary study population consisted of 3,020 subjects randomly selected from the THS (26). A second cohort of 802 self-reported white American bank employees from a broad range of socioeconomic levels was used for some assays (27). Detailed biodata and blood samples obtained after an overnight fast were collected for each subject. Plasma lipids were measured as described (26). The protocols were approved by the Committee on Human Research of the University of California, San Francisco, and were in accordance with the Helsinki Declaration. Subjects who were taking lipid-lowering medication, had a history of diabetes mellitus, or had a plasma triglyceride level > 800 mg/dl were excluded.

Detection of *APOA5* polymorphisms

Primers were designed to amplify across the *APOA5* promoter, the 5' untranslated region (UTR), and all exons, including intron/exon splicing boundaries when possible. DNA from 23 subjects (13 THS participants and 10 white Americans) was sequenced to identify polymorphisms in *APOA5*. DNA sequences were aligned and analyzed with Sequencher DNA analysis software (Gene Codes, Ann Arbor, MI).

Genotyping

After amplification by polymerase chain reaction, each polymorphism was genotyped by restriction fragment length polymorphism, digesting the primary amplification with restriction endonucleases and separating the resulting fragments with 1-3% agarose gels. The conditions of all assays are described in supplementary Table 1.

Statistics and data analysis

Data were analyzed with SPSS 10.0, Microsoft Access, and Excel. Associations between genotypes, lipids, and other parameters were analyzed separately for males and females. Lipid levels are expressed in mg/dl, and all values are reported as means \pm SD. Mean values were compared with the *t*-test according to genotype or haplotype; $P < 0.05$ (two-tailed) was considered significant. Because triglyceride levels were not normally distributed, log-transformed values were used for statistical comparison; untransformed mean values are reported here. Analysis of covariance was used to construct a model to explain the variation in triglyceride levels and the overall effect of haplotype on plasma triglyceride levels. Body mass index (BMI), age, smoking, and alcohol consumption were included as covariates, and genotype score was included as a fixed factor in the model (GLM Univariate, SPSS 10.0). The proportion of variation in plasma triglyceride level from each SNP or haplotype was estimated from partial regression coefficients (28). Chi-square analysis was used to test differences between the observed and expected frequencies of alleles (assuming a Hardy-Weinberg equilibrium) and to compare genotype, allele, or haplotype frequencies after stratification by age- and gender-adjusted triglyceride percentiles ($\leq 20^{\text{th}}$ and $\geq 80^{\text{th}}$).

The expectation-maximization algorithm was used to estimate the maximum-likelihood haplotype frequencies from multilocus genotypic data without known gametic phase (Arlequin software, version 2.00) (29). All subjects with missing genotype data were excluded during haplotype prediction. Haplotypes that could be unambiguously attributed to individuals were further analyzed for associations with lipid and demographic data. The LD between polymorphisms was similarly calculated with Arlequin (29) and expressed in terms of $D' = D/D_{\text{max}}$ or D/D_{min} (30).

RESULTS

Population characteristics

Demographic and biochemical characteristics of 3,020 THS participants are presented in Table 1. Both males and females had low plasma HDL-C levels and high total cholesterol/HDL-C ratios. Detailed analyses of the THS data

TABLE 1. Demographic and biochemical characteristics of Turkish Heart Study participants (n = 3,020)

Variable	Males (n = 1,661)	Females (n = 1,359)	P
Age (years)	42 \pm 13	42 \pm 15	NS
Body mass index (kg/m ²)	26.1 \pm 3.9	26.6 \pm 5.4	<0.05
HDL cholesterol (mg/dl)	35.8 \pm 7.5	41.2 \pm 9	<0.001
Total cholesterol (mg/dl)	184 \pm 45	183 \pm 42	NS
LDL cholesterol (mg/dl)	126 \pm 41	116 \pm 39	<0.05
Triglycerides (mg/dl)	153 \pm 107	110 \pm 70	<0.001
Total cholesterol/HDL cholesterol ratio	5.8 \pm 2.9	4.5 \pm 1.4	<0.01
Systolic blood pressure (mm Hg)	125 \pm 23	122 \pm 21	NS
Diastolic blood pressure (mm Hg)	82 \pm 14	81 \pm 13	NS
Consumption of alcohol (%) ^a	29.9	5.5	<0.001
Cigarette smoking (%) ^b	56.7	24.1	<0.001

Values are means \pm SD or percentages. Means were compared by *t*-test, and percentages were analyzed by chi-square test.

^a One or more drinks per week.

^b One or more cigarettes per day.

TABLE 2. Description and frequency of *APOA5* polymorphisms in Turks

Polymorphic Site ^a	Nucleotide Change	Location in the Gene	Location on Chromosome 11 ^b	Rare Allele	Number ^c	SNP Identifier	Reference ^d
				%			
-1464T>C	T/C	Promoter	-1,456	29.0	1,574/1,304	rs10750097	—
-1275G>A	G/A	Promoter	-1,267	9.4	129/106	rs17120035	—
-1131T>C	T/C	Promoter	-1,123	12.8	1,601/1,302	rs662799	8
-1099C>T	C/T	Promoter	-1,091	10.3	1,505/1,181	rs1729411	16
-1021G>A	G/A	Promoter	-1,012	5.8	1,596/1,288		New
-3A>G	A/G	5' UTR	6	13.9	230/188	rs651821	8
C56G (S19W)	C/G	Exon 3	178	5.6	1,633/1,334	rs3135506	20
I432A (I44I)	C/A	Exon 3	254	6.1	130/107	rs12287066	20
IVS3 + 476G>A	G/A	Intron 3	759	13.6	176/144	rs2072560	8
G457A (V153M)	G/A	Exon 4	1,097	4.6	1,502/1,162	rs3135507	15, 16
G553T (G185C)	G/T	Exon 4	1,193	0.6	201/288	rs2075291	15, 16
1177C>T	C/T	3' UTR	1,817	4.5	333/272		15
1259T>C	T/C	3' UTR	1,899	14.6	1,634/1,327	rs2266788	8
1387-1388delAG	(AG)	3' UTR	2,027-2,028	~4.6	13 ^e		New
1495T>C	T/C	3' UTR	2,135	4.8	136/111		New

APOA5, apolipoprotein A-V gene; SNP, single nucleotide polymorphism; UTR, untranslated region.
^aRelative to ATG start, reference sequence AAS68229.1. Synonymous and nonsynonymous changes and their locations are shown in parentheses.
^b(+) strand ENSEMBL, NCBI build 35, Ch11, 116165297:116167794:1.
^cNumber of males/females genotyped by restriction fragment length polymorphism.
^dFirst publication of the particular polymorphism.
^eVariation frequency determined by direct sequencing.

have been reported (26, 31, 32). It is noteworthy that low plasma HDL-C levels were found to increase the relative risk for CAD, and the plasma total cholesterol/HDL-C ratio was found to be an independent predictor of coronary events in Turks (33, 34).

***APOA5* polymorphisms**

Fifteen SNPs with rare allelic frequencies from <1% to 29% were identified (Table 2). Five SNPs were in the promoter region, including the novel -1021G>A, and one in the 5' UTR (-3A>G). Four SNPs were in the coding sequence: three were nonsynonymous (S19W, V153M, and G185C) and one was synonymous (I44I). Four SNPs were in the 3' UTR: two were novel (1387-1388delAG and 1495T>C) and two were published previously (1177C>T and 1259T>C). The IVS3+476G>A intronic SNP was also identified previously.

LD for *APOA5*

The LD between polymorphic sites was calculated using unphased genotypes from 14 SNPs from 215 randomly

chosen unrelated Turkish subjects (Table 3; see supplementary Table II). Three clusters of *APOA5* polymorphic sites were in strong LD: -3A>G, IVS3+476G>A, and 1259T>C; S19W and I44I; and V153M, 1177C>T, and 1495T>C. 1259T>C, S19W, and V153M were chosen as markers for their clusters. -1275G>A was exclusively on one haplotype and in LD with -1464T>C. Seven htSNPs (-1464T>C, -1131T>C, -1099C>T, -1021G>A, S19W, V153M, and 1259T>C) were selected to assess the association between *APOA5* polymorphisms and plasma triglyceride levels in ~3,000 Turkish subjects.

The initial sequencing results suggested that the 1387-1388delAG variant was completely linked to the V153M, 1177C>T, and 1495T>C variants. This linkage was further supported by sequencing three additional 153MM subjects. Because it was in LD with a marker for V153M, the 1387-1388delAG variant was not analyzed further.

The frequency of the rare G185C allele, which is significantly associated with high triglyceride levels in the Chinese population (15), was 0.6% (n = 487) in the Turkish population. Only five Turkish males and one female

TABLE 3. Common *APOA5* haplotypes in a random Turkish population

Haplotype	-1464T>C	-1275G>A	-1131T>C	-1099C>T	-1021G>A	-3A>G ^a	S19W ^b	I44I ^b	IVS3+476G>A ^a	V153M ^c	G185C	1177C>T ^c	1259T>C ^c	1495T>C ^c
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	2	1	2	1	1	2	1	1	2	1	1	1	2	1
3	1	1	1	2	1	1	1	1	1	1	1	1	1	1
4	2	2	1	1	1	1	1	1	1	1	1	1	1	1
5	1	1	1	1	2	1	1	1	1	1	1	1	1	1
6	2	1	1	1	1	1	2	2	1	1	1	1	1	1
7	1	1	1	1	1	1	1	1	1	2	1	2	1	2
8	2	1	1	1	1	2	1	1	2	1	1	1	2	1
9	1	1	2	1	1	1	1	1	1	1	1	1	1	1

1, common allele; 2, rare allele.
^a-3A>G, IVS3+476G>A, and 1259T>C are in strong linkage disequilibrium (LD).
^bS19W and I44I are in strong LD.
^cV153M, 1177C>T, and 1495T>C are in strong LD.

TABLE 4. *APOA5* SNPs and plasma triglyceride levels in a random Turkish population

SNP	AA	AB	BB	<i>P</i> (AA vs. AB)	<i>P</i> (AA vs. BB)	Difference (BB – AA) ^a	
	mg/dl	mg/dl	mg/dl			mg/dl	%
Males							
-1464T>C	139 ± 94 (774)	159 ± 108 (688)	187 ± 144 (112)	<0.001	<0.001	48	35
-1131T>C	144 ± 99 (1,220)	170 ± 117 (357)	230 ± 148 (24)	<0.001	<0.001	86	60
-1099C>T	151 ± 104 (1,198)	148 ± 104 (298)	161 ± 97 (9)	NS	NS		
-1021G>A	152 ± 107 (1,431)	147 ± 87 (158)	145 ± 111 (7)	NS	NS		
S19W	148 ± 102 (1,438)	187 ± 132 (193)	191 ± 6 (2)	<0.001	—	39	26 ^a
V153M	151 ± 105 (1,364)	156 ± 111 (134)	104 ± 51 (4)	NS	NS		
1259T>C	149 ± 104 (1,211)	166 ± 114 (390)	195 ± 139 (33)	<0.005	<0.05	46	31
Females							
-1464T>C	101 ± 59 (649)	115 ± 78 (551)	122 ± 98 (104)	<0.001	<0.03	21	21
-1131T>C	104 ± 64 (983)	121 ± 87 (298)	135 ± 107 (21)	<0.002	<0.05	31	30
-1099C>T	109 ± 73 (947)	102 ± 70 (229)	120 ± 61 (5)	NS	NS		
-1021G>A	110 ± 73 (1,137)	103 ± 56 (141)	98 ± 49 (10)	NS	NS		
S19W	108 ± 69 (1,198)	127 ± 95 (135)	96 (1)	<0.001	—	19	18 ^a
V153M	107 ± 71 (1,062)	117 ± 79 (97)	82 ± 14 (3)	NS	NS		
1259T>C	104 ± 64 (961)	124 ± 86 (326)	150 ± 118 (40)	<0.001	<0.002	46	44

A, common allele; B, rare allele. Values shown are means ± SD. Number of subjects is shown in parentheses.
^aFor S19W, the difference is between AB and AA.

with 185GC heterozygosity were identified, and all had very high plasma triglyceride levels (372 ± 250 mg/dl for males, 499 mg/dl for the female). However, because of its low frequency, this SNP is unlikely to have a significant impact in the Turkish population.

***APOA5* htSNPs and plasma triglyceride levels**

The seven htSNPs and their associations with plasma triglyceride levels are presented in Table 4, where A denotes common alleles and B denotes rare alleles. In both males and females, triglyceride levels were significantly higher in AB and BB subjects (-1464T>C, -1131T>C, and 1259T>C) than in AA subjects (*P* < 0.005). All three of these SNPs had clear gene-dose effects. Additionally, S19W subjects had significantly higher triglyceride levels than those with 19SS (*P* < 0.001), and this effect was more prominent in males. There were too few 19WW subjects for

statistical analysis. In males, the -1131T>C SNP had the greatest effect on plasma triglycerides; the difference between the BB and AA genotypes was 86 mg/dl (60%). Interestingly, the rare -1131T>C allele had a much greater effect in males than in females (60% vs. 30% increase). The 1259T>C polymorphism had the largest impact in females: the triglyceride level was 46 mg/dl (44%) greater in the BB group than in the AA group.

Because clear gene-dose effects were observed (Table 4), both AB and BB individuals were combined into a group of B allele carriers (AB + BB), and both allele and genotype frequency distributions were determined for subjects with triglyceride levels in the ≤20th and ≥80th percentiles (Table 5). The B allele and B allele carriers were significantly more frequent in the ≥80th percentile than in the ≤20th percentile groups in both males and females with -1464T>C, -1131T>C, S19W, and 1259T>C (*P* < 0.05),

TABLE 5. Frequencies of rare allele (B) and rare allele carriers (AB + BB) in men and women with *APOA5* SNPs who are in the ≤20th and ≥80th percentiles of triglyceride levels

SNP	B			<i>P</i> (≤20 th vs. ≥80 th Percentile)	AB + BB			<i>P</i> (≤20 th vs. ≥80 th Percentile)
	All	≤20 th Percentile	≥80 th Percentile		All	≤20 th Percentile	≥80 th Percentile	
Males								
-1464T>C	0.290	0.252	0.344	<0.001	0.508 (800)	0.434 (118)	0.590 (233)	<0.001
-1131T>C	0.126	0.090	0.163	<0.001	0.238 (381)	0.176 (48)	0.304 (121)	<0.001
-1099C>T	0.105	0.113	0.093	—	0.204 (307)	0.218 (57)	0.179 (68)	—
-1021G>A	0.054	0.051	0.054	—	0.103 (165)	0.096 (26)	0.103 (41)	—
S19W	0.060	0.048	0.090	<0.005	0.119 (195)	0.096 (27)	0.177 (75)	<0.01
V153M	0.047	0.051	0.043	—	0.092 (138)	0.094 (24)	0.087 (33)	—
1259T>C	0.140	0.111	0.165	<0.01	0.259 (423)	0.208 (57)	0.304 (130)	<0.01
Females								
-1464T>C	0.291	0.235	0.326	<0.001	0.502 (655)	0.418 (104)	0.559 (170)	<0.001
-1131T>C	0.131	0.102	0.158	<0.01	0.245 (319)	0.201 (50)	0.292 (88)	<0.02
-1099C>T	0.101	0.093	0.090	—	0.198 (234)	0.186 (44)	0.172 (47)	—
-1021G>A	0.063	0.075	0.060	—	0.117 (151)	0.137 (34)	0.110 (33)	—
S19W	0.051	0.032	0.060	<0.03	0.102 (136)	0.063 (16)	0.120 (38)	<0.05
V153M	0.044	0.033	0.056	—	0.086 (100)	0.066 (16)	0.112 (30)	—
1259T>C	0.153	0.110	0.202	<0.001	0.276 (366)	0.212 (53)	0.351 (112)	<0.001

Percentages were analyzed by chi-square (2×2) test. Number of subjects is shown in parentheses.

TABLE 6. Additive effects of *APOA5* SNPs on plasma triglyceride levels

S19W	-1131T>C	Males	Increase ^a	Females	Increase ^a
SS	TT	138 ± 93 (1,041)	—	102 ± 58 (852)	—
	TC	165 ± 111 (332)	27	119 ± 86 (279)	17
	CC	228 ± 151 (23)	90	138 ± 110 (20)	36
SW	TT	181 ± 125 ^b (158)	43	124 ± 94 ^b (113)	22
	TC	218 ± 112 ^c (20)	80	129 ± 106 (15)	27
S19W	1259T>C	Males	Increase ^d	Females	Increase ^d
SS	TT	142 ± 96 (1,028)	—	102 ± 58 (826)	—
	TC	162 ± 112 (364)	20	122 ± 84 (310)	20
	CC	195 ± 139 (33)	53	150 ± 118 (40)	48
SW	TT	187 ± 133 ^e (164)	45	125 ± 93 ^e (121)	23
	TC	221 ± 109 ^f (22)	79	146 ± 119 (13)	44

Values shown are mg/dl, means ± SD. Number of subjects is shown in parentheses.

^aTriglyceride increase relative to 19SS/-1131TT.

^b*P* < 0.005 versus 19SS/-1131TT.

^c*P* < 0.005 versus 19SS/-1131TC.

^dTriglyceride increase relative to 19SS/1259TT.

^e*P* < 0.05 versus 19SS/1259TT.

^f*P* < 0.05 versus 19SS/1259TC.

further supporting the association of these SNPs with increased triglycerides.

Two SNP pairs, S19W/-1131T>C and S19W/1259T>C, had significant additive and independent effects on triglyceride levels (Table 6). -1131T>C and 1259T>C were each associated significantly with increased triglyceride levels in 19SS homozygous males and females. The largest effects were a 90 mg/dl difference between the 19SS/-1131CC and 19SS/-1131TT genotypes in males and a 48 mg/dl difference between the 19SS/1259CC and 19SS/1259TT genotypes in females. There were too few S19W/-1131CC and S19W/1259CC subjects for statistical analysis (data not shown). Notably, for both SNP pairs examined, double heterozygotes always had higher triglyceride levels than single heterozygotes. The -1099C>T, -1021G>A, and V153M SNPs were not associated with plasma triglyceride levels.

APOA5 haplotypes and plasma triglyceride levels

The nine most common *APOA5* haplotypes (frequency > 1.0%) accounted for 96.0% of all 36 predicted haplo-

types (Table 7). Plasma triglyceride levels associated with haplotype 1, the most frequent haplotype possessing the common alleles for all seven htSNPs, were compared with the mean triglyceride levels for the other haplotypes (Table 7). Haplotype 2, characterized by the rare alleles for -1464T>C, -1131T>C, and 1259T>C, was associated with significantly higher triglyceride levels in both males and females than haplotype 1. Notably, the rare -1464T>C SNP occurred in isolation on haplotype 4, and its triglyceride level was not different from that associated with haplotype 1, suggesting that -1464T>C by itself had no effect. However, haplotype 6, which possessed the rare alleles of -1464T>C and S19W, was associated with higher triglyceride levels than haplotype 1 in both males and females. The triglyceride levels associated with haplotypes 2 and 6 were significantly higher in males than in females (haplotype 2, 22% vs. 14%; haplotype 6, 35% vs. 17%). Additionally, haplotype 9, with the -1131T>C rare allele in isolation, was associated with higher triglyceride levels than haplotype 1 in males only [27 mg/dl

TABLE 7. Plasma triglyceride levels of common haplotypes of *APOA5* and their frequencies in a random Turkish population

Haplotype	Frequency	Males ^d		Increase versus Haplotype 1		Females ^d		Increase versus Haplotype 1		-1464T>C -1131T>C -1099C>T -1021G>A S19W V153M 1259T>C						
		mg/dl		mg/dl	%	mg/dl		mg/dl	%							
1	0.481	143 ± 98 (1,341)	—	—	—	109 ± 75 (1,104)	—	—	—	—	—	—	—	—	—	—
2	0.101	174 ± 124 ^a (270)	31	22	124 ± 90 ^a (243)	15	14	2	2	1	1	1	1	1	1	2
3	0.101	151 ± 106 (290)			100 ± 57 (218)			1	1	2	1	1	1	1	1	1
4	0.104	144 ± 101 (312)			100 ± 58 (211)			2	1	1	1	1	1	1	1	1
5	0.051	145 ± 86 (140)			101 ± 57 (120)			1	1	1	2	1	1	1	1	1
6	0.050	193 ± 131 ^a (142)	50	35	127 ± 90 ^a (111)	18	17	2	1	1	1	2	1	1	1	1
7	0.037	154 ± 114 (112)			118 ± 78 (79)			1	1	1	1	1	1	2	1	1
8	0.021	138 ± 90 (56)			127 ± 97 ^{b,c} (51)			2	1	1	1	1	1	1	1	2
9	0.015	170 ± 80 ^a (49)	27	19	107 ± 88 (25)			1	2	1	1	1	1	1	1	1
Sum	0.960															

1, common allele; 2, rare allele. Number of subjects is shown in parentheses.

^a*P* < 0.05 versus haplotype 1 (*t*-test).

^b*P* < 0.09 versus haplotype 1 (*t*-test).

^c*P* < 0.05 versus haplotypes 3, 4, and 5 (*t*-test).

^dValues shown are means ± SD.

(19%); $P < 0.05$). However, haplotype 8, containing the -1464T>C and 1259T>C rare alleles, was associated with higher triglyceride levels than haplotype 1 in females [18 mg/dl (17%); $P = 0.09$]. The -1099C>T SNP was found only on haplotype 3, -1021G>A only on haplotype 5, and V153M only on haplotype 7. These haplotypes were not associated with differences in plasma triglyceride levels.

Haplotypes 2 and 6 were 1.6- to 2.1-fold more frequent in the $\geq 80^{\text{th}}$ than in the $\leq 20^{\text{th}}$ percentile group in both sexes (Table 8). Although haplotype 9 was >2-fold more frequent in the $\geq 80^{\text{th}}$ percentile group in males, the difference was not statistically significant, possibly because of the low number of subjects tested. However, when triglyceride tertiles were used, haplotype 9 was significantly more frequent in the $\geq 67^{\text{th}}$ percentile than in the $\leq 33^{\text{rd}}$ percentile [2.8% (n = 29) vs. 0.9% (n = 7); $P < 0.01$]. In females, haplotype 8 was more common in the $\geq 80^{\text{th}}$ than in the $\leq 20^{\text{th}}$ percentile (Table 8). These findings further substantiate the association between the -1131T>C, S19W, and 1259T>C SNPs and increased plasma triglyceride levels in Turks.

To assess the additive effects of haplotypes, we examined the mean triglyceride values of subjects with haplotype pairs 1-1, 1-2, 1-6, 2-2, and 2-6 (other haplotype pairs were too infrequent to analyze). In males, triglyceride levels were higher in those with haplotype pairs 2-2 (252 \pm 190 mg/dl; n = 8) and 2-6 (261 \pm 138 mg/dl; n = 8) than in those with haplotype pair 1-1 (133 \pm 89 mg/dl; n = 288), 1-2 (150 \pm 114 mg/dl; n = 155), or 1-6 (185 \pm 131 mg/dl; n = 90). In females, triglyceride levels were significantly higher in those with haplotype pair 2-2 (172 \pm 140 mg/dl; n = 10) than in those with haplotype pair 1-1 (108 \pm 74 mg/dl; n = 262) or 1-2 (128 \pm 87 mg/dl; n =

127). Also in females, haplotype pair 2-6 (128 \pm 101 mg/dl; n = 9) was associated with higher triglyceride levels than haplotype pair 1-1. These findings suggest that *APOA5* haplotypes 2 and 6 had additive effects, particularly in males.

In addition to *t*-test comparisons, analysis of covariance (covariates were HDL-C, age, BMI, smoking, and alcohol consumption) confirmed the significance of the SNP and haplotype effects on triglyceride levels (see supplementary Table III). Bonferroni post hoc analysis showed that this significance originated principally from haplotypes 2, 6, and 9 in males and from haplotypes 2, 6, and 8 in females.

White American study population and the -1464T>C SNP

Haplotype analysis in the Turks suggested that the -1464T>C SNP was a marker for the -1131T>C, S19W, and 1259T>C SNPs and that the associated phenotype seen with the -1464T>C SNP derived from the strong LD between -1464T>C and these other three SNPs (Tables 7, 8; see supplementary Table II). To confirm this phenomenon in another population, we analyzed the distribution of the -1464T>C, -1131T>C, S19W, and 1259T>C SNPs in 802 self-reported white non-Hispanic Americans. Initial analysis showed that the -1131T>C and 1259T>C SNPs were almost in complete LD in white Americans (only 3 of 228 paired genotypes were different; $D' = 0.935$), as in other European populations; therefore, 1259T>C was not genotyped further. In contrast, the -1131T>C and 1259T>C SNPs were not as strongly linked in Turks ($D' = 0.698$; see supplementary Table II). The rare allele frequencies for the -1464T>C, -1131T>C, and S19W SNPs were 19.0, 5.9, and 6.0%, respectively, in white Americans and 29.0, 12.8, and 5.6%, respectively, in Turks (Table 2). Triglyceride levels were significantly higher in AB and BB subjects with -1464T>C and in AB subjects with both -1131T>C and S19W than in AA subjects ($P < 0.05$) (Table 9). Haplotype analysis suggested LD between -1464T>C and the -1131T>C and S19W SNPs, and that -1464T>C might be a marker for these other SNPs in white Americans as in Turks. The rare -1464T>C SNP occurred in isolation on haplotype X, and the triglyceride level associated with this haplotype was not different from that associated with haplotype W (Table 9), suggesting that -1464T>C by itself had no effect.

DISCUSSION

This study shows that three common *APOA5* SNPs (-1131T>C, S19W, and 1259T>C) and the haplotypes formed with seven *APOA5* htSNPs were significantly associated with increased plasma triglyceride levels in Turks, regardless of sex. No other associations with lipid parameters (HDL-C, LDL, total cholesterol, or total cholesterol/HDL-C ratio) were found. The rare SNP alleles were significantly more frequent in subjects with the highest plasma triglyceride levels ($\geq 80^{\text{th}}$ percentile) than in those with the lowest levels ($\leq 20^{\text{th}}$ percentile). The effects of S19W and of -1131T>C and 1259T>C were indepen-

TABLE 8. Frequency comparison of common haplotypes of *APOA5* between the $\leq 20^{\text{th}}$ and $\geq 80^{\text{th}}$ percentiles of triglyceride

Haplotype	All Groups	Triglyceride Subgroups		P ($\leq 20^{\text{th}}$ vs. $\geq 80^{\text{th}}$ Percentile)
		$\leq 20^{\text{th}}$ Percentile	$\geq 80^{\text{th}}$ Percentile	
Males				
1	0.474	0.516 (254)	0.439 (316)	0.01
2	0.096	0.069 (34)	0.125 (90)	0.02
3	0.103	0.108 (53)	0.093 (67)	NS
4	0.110	0.118 (58)	0.103 (74)	NS
5	0.050	0.045 (22)	0.050 (36)	NS
6	0.050	0.036 (18)	0.079 (57)	0.004
7	0.039	0.042 (21)	0.038 (28)	NS
8	0.020	0.016 (8)	0.015 (11)	NS
9	0.017	0.008 (4)	0.018 (13)	NS
Females				
1	0.491	0.543 (246)	0.491 (260)	NS
2	0.108	0.084 (38)	0.132 (70)	0.02
3	0.097	0.090 (41)	0.076 (40)	NS
4	0.094	0.102 (46)	0.068 (36)	NS
5	0.054	0.066 (30)	0.047 (25)	NS
6	0.049	0.029 (13)	0.059 (31)	0.035
7	0.036	0.024 (11)	0.049 (26)	NS
8	0.023	0.009 (4)	0.032 (17)	0.021
9	0.012	0.013 (6)	0.008 (4)	NS

Values shown are frequencies. Number of subjects is shown in parentheses. Percentages were analyzed by chi-square test.

TABLE 9. Plasma triglyceride levels of SNPs and haplotypes of *APOA5* in a white American population

SNP	AA	AB	BB	Difference (AB – AA)		
Males						
-1464T>C	132 ± 77 (160)	175 ± 115 ^a (65)	170 ± 139 ^a (7)	43		
-1131T>C	139 ± 93 (208)	162 ± 82 ^a (21)	205 ± 42 (2)	23		
S19W	138 ± 83 (191)	164 ± 88 ^a (23)	459 ± 300 (2)	26		
Females						
-1464T>C	116 ± 67 (364)	131 ± 87 ^a (176)	137 ± 107 ^a (23)	15		
-1131T>C	120 ± 76 (505)	139 ± 78 ^a (63)	211 ± 78 (3)	19		
S19W	121 ± 77 (500)	138 ± 75 ^a (58)	150 ± 105 (4)	17		
Haplotype	Frequency	Males	Females	-1464T>C	-1131T>C	S19W
W	0.76	132 ± 73 (335)	117 ± 69 (844)	0	0	0
X	0.12	135 ± 75 (30)	122 ± 98 (107)	1	0	0
Y	0.05	174 ± 96 ^b (21)	132 ± 62 ^b (51)	1	0	1
Z	0.04	168 ± 89 ^b (16)	132 ± 70 ^b (50)	1	1	0

A, common allele; B, rare allele. Values shown are mg/dl, means ± SD. Number of subjects is shown in parentheses.

^a*P* < 0.05 versus AA.

^b*P* < 0.05 versus haplotype W.

dent of each other and additive, each showing a dose-dependent association with phenotype. Turks have low HDL-C levels (26, 31–34), and the inverse relationship between plasma HDL-C and triglyceride levels is well established (35, 36). When plasma triglyceride levels were adjusted for covariates (HDL-C, age, BMI, smoking, and alcohol consumption), the -1131T>C, S19W, and 1259T>C SNPs and haplotypes were significantly associated with increased plasma triglyceride levels, suggesting that the primary associations were between these polymorphisms and triglyceride levels.

The rare allele of -1464T>C was associated with increased plasma triglyceride levels in Turks. However, the mean triglyceride level for the haplotype with the rare -1464T>C allele in isolation (haplotype 4; Table 7) was not significantly different from that for the most frequent haplotype (haplotype 1; Table 7), suggesting that the association between -1464T>C and triglyceride level was primarily attributable to -1131T>C, 1259T>C, and S19W. Additionally, the frequency of haplotype 4 was not higher in the ≥80th than in the ≤20th percentile group. These results suggest that -1464T>C might only be a marker for those three SNPs and not a direct modulator of triglyceride levels. The same conclusion was reached in analyzing the effect of -1464T>C in white Americans (Table 9).

Two *APOA5* SNPs, -1131T>C and S19W, have been studied extensively, and their rare allele frequencies vary greatly among populations. The frequency of the -1131T>C rare C allele was 27–37% in East Asians (12–14, 16, 19), 13–16% in Hispanics (12, 20), 6–9% in African-Americans and western Europeans (or their descendants) (8, 17, 20, 21), and 12.8% in Turks. The allele frequency of the S19W SNP was very rare (<0.1%) in Chinese (16) and Japanese (13), 4–8% in African-Americans and western Europeans (17, 18, 20, 21), 15% in Hispanics (20), and 5.6% in Turks. More importantly, the high plasma triglyceride levels associated with these rare alleles also vary among populations. Japanese (19) and Malay (16) homozygotes for the -1131T>C rare allele had 10% and 69% higher triglyceride levels, respectively, than homozygotes for the common

allele. Intermediate increases in triglycerides have been associated with this polymorphism in other populations (12, 16, 17, 20). The impact of the -1131T>C rare allele on plasma triglyceride levels was comparatively high in Turks. The -1131T>C subjects had a 60% (86 mg/dl) increase in triglyceride levels in Turkish males and a 30% (31 mg/dl) increase in Turkish females. Compared with 19SS, 19SW was associated with 8–16% higher plasma triglyceride levels among Caucasians (21) and 20–30% higher levels in African-Americans (20) and with 26% (39 mg/dl) higher levels in Turkish males and 18% (19 mg/dl) higher levels in Turkish females. Interestingly, associations were not found for -1131T>C in African-American males or females (20, 22) or for S19W in the LOCAT study (24) or in African-American males or white females from the CARDIA study (22) or in CAD patients from Vancouver, Canada (23). However, the associations of -1131T>C and S19W with triglyceride levels in Turks are significant and some of the highest reported for the *APOA5* locus. In our analysis of non-Hispanic white Americans, the alleles had effects similar to those reported for Caucasian and European populations.

The 1259T>C SNP has also been investigated for its association with triglyceride levels. It was associated with 37% higher triglyceride levels in a Japanese-American population (13) and with 33–53% higher levels in three Singaporean populations (16). We found similar increases of 31% (46 mg/dl) in Turkish males and 44% (46 mg/dl) in Turkish females.

The -1131T>C, S19W, and 1259T>C polymorphisms explained 18.6, 10.7, and 8.6% of the variance in triglyceride levels, respectively, in Turkish males and 9.3, 3.8, and 12.5%, respectively, in Turkish females. The magnitudes of these variances are consistent with the higher percentage increase in triglyceride levels associated with both -1131T>C and S19W and with the lower percentage increase associated with 1259T>C in Turkish males (Table 4). Previously, we showed that gender has a much greater effect on HDL-C levels in Turks, especially males, than in other populations (32). Similarly, the combined

effect of the nine common *APOA5* haplotypes explained 16.2% of the variance in triglyceride levels in Turkish males and 12.8% in females, and the percentage increases associated with haplotypes 2 and 6 were higher in males (Table 7). These results suggest that gender-specific influences may interact with these polymorphisms to modulate triglyceride levels in Turks.

In European populations, four SNPs (-1131T>C, -3A>G, IVS3+476G>A, and 1259T>C) constituted a single haplotype (17, 20). However, in Turks, three Singaporean populations, and African-Americans, the *APOA5* haplotype structure was more complex (16, 22). -1131T>C was in strong, but not complete, LD with the three other SNPs in Turks and Singaporeans (16), and 1259T>C was very rare in African-Americans (<0.001%) (22). Haplotypes containing both the -1131T>C and 1259T>C rare alleles modulated triglyceride levels in Turks, and the effect of these SNPs may be independent of each other and gender-specific, because triglyceride increase was associated with haplotype 9 (-1131T>C in isolation) only in males and with haplotype 8 (1259T>C in isolation) only in females. On the other hand, -1131T>C was not associated with triglyceride levels in African-Americans, in whom 1259T>C is extremely rare (22). Association studies, including studies of *APOA5*, have shown gender differences in lipid metabolism (37–40), but the mechanism is not fully understood. Functional studies should be conducted to determine how the -1131T>C and 1259T>C SNPs modulate triglycerides.

The G185C SNP, with an allelic frequency of 4.2%, was significantly associated with increased triglyceride levels in a Chinese population (15) but was extremely rare or absent in Caucasians (20, 41). Although G185C was very rare in the Turkish population (0.6% allelic frequency), all six GC heterozygotes had very high plasma triglyceride levels.

Plasma triglyceride levels were decreased significantly by overexpression of *APOA5* (8, 9) and increased significantly in *Apoa5* knockout mice (8). Because *APOA5* polymorphisms have been associated with high plasma triglyceride levels, SNP-associated increases may reflect the impaired function of apoA-V. In HepG2 cells, the W19-encoded signal peptide was secreted into the medium at significantly lower levels than the S19-encoded signal peptide (42). Potentially, the -1131T>C, -3A>G, and 1259T>C SNPs may also affect the function of *APOA5*. -1131T>C is located in the promoter region and may alter *APOA5* expression, and 1259T>C, located in the 3' UTR, might affect the stability of *APOA5* mRNA. Alternatively, 1259T>C, which is in complete LD in Turks, may be a marker for -3A>G; however, expression assays did not support a biological function for -3A>G (42). Although in vitro studies did not show individual effects of these three SNPs, cooperative effects cannot be excluded. Except for the two studies of African-American males and females in whom the 1259T>C SNP was very rare (20, 22), the -1131T>C SNP was shown to be associated with increased plasma triglyceride levels in several studies (8, 12–14, 16–22) and supports the idea of cooperation between *APOA5* SNPs.

APOA5 is located downstream of the *APOA1/C3/A4* gene cluster in a small 60 kb region on human chromo-

some 11. *APOA1* variants are primarily associated with altered HDL-C levels (43, 44) and *APOC3* variants with altered triglyceride levels (43–45). Transgenic and knock-out studies suggest that *APOA5* and *APOC3* independently influence plasma triglyceride levels in an opposite manner (46). A recent study in Caucasians suggested a high degree of LD across the entire gene cluster; nevertheless, *APOA5* was separated from the other apolipoprotein genes by a region of low LD (47). Additionally, some individual *APOA5* SNPs (haplotype *APOA5**2) were in strong LD with *APOC3* SNPs, whereas S19W exerted its effect on triglyceride levels independently of *APOC3* SNPs (47). The structure of the *APOA1/C3/A4/A5* cluster and its association with triglyceride levels should be examined in other populations.

Hypertriglyceridemia is an independent risk factor for CAD (1, 2). For every 1 mmol/l (~88.5 mg/dl) increase in plasma triglycerides, the risk of CAD was increased significantly by 14% in males and 37% in females after adjustment for HDL-C and other factors (48). In Turks, the *APOA5* SNP-associated triglyceride increase was 19–86 mg/dl, depending on sex and the polymorphism, in a population in which ~40% carry at least one rare allele of -1131T>C, S19W, or 1259T>C. The magnitude of the change in triglyceride levels and the relatively high frequencies of these rare *APOA5* alleles are important considerations in assessing the risk of CAD in Turks, particularly those with low plasma HDL-C levels. ■

The authors are indebted to their associates at the American Hospital, Istanbul, especially Guy Pépin, Sibel Tanir, Judy Dawson-Pépin, and Linda L. Mahley in the Gladstone Institute (Istanbul) and Dr. K. Erhan Palaoglu at the American Hospital (Istanbul). The authors thank Sylvia Richmond for manuscript preparation and Stephen Ordway and Gary Howard for editorial assistance. The authors acknowledge the generous support of the American Hospital, especially Mr. George Rountree, and the J. David Gladstone Institutes. This work was supported in part by Grants R01 HL-71027 and R01 HL-64162 from the National Institutes of Health.

REFERENCES

1. Murray, C. J. L., and A. D. Lopez. 1997. Mortality by cause for eight regions of the world. Global Burden of Disease Study. *Lancet*. 349: 1269–1276.
2. Peto, R., A. D. Lopez, J. Boreham, M. Thun, C. Heath, Jr., and R. Doll. 1996. Mortality from smoking worldwide. *Br. Med. Bull.* 52: 12–21.
3. Brenn, T. 1994. Genetic and environmental effects on coronary heart disease risk factors in Northern Norway. The cardiovascular disease study in Finnmark. *Ann. Hum. Genet.* 58: 369–379.
4. Heller, D. A., U. de Faire, N. L. Pedersen, G. Dahlén, and G. E. McClearn. 1993. Genetic and environmental influences on serum lipid levels in twins. *N. Engl. J. Med.* 328: 1150–1156.
5. Pérusse, L., T. Rice, J. P. Després, J. Bergeron, M. A. Province, J. Gagnon, A. S. Leon, D. C. Rao, J. S. Skinner, J. H. Wilmore, et al. 1997. Familial resemblance of plasma lipids, lipoproteins and postheparin lipoprotein and hepatic lipases in the HERITAGE family study. *Atheroscler. Thromb. Vasc. Biol.* 17: 3263–3269.
6. Wyszynski, D. F., D. M. Waterworth, P. J. Barter, J. Cohen, Y. A. Kesäniemi, R. W. Mahley, R. McPherson, G. Waeber, T. P. Bersot, S.

- S. Sharma, et al. 2005. Relation between atherogenic dyslipidemia and the Adult Treatment Program-III definition of metabolic syndrome (Genetic Epidemiology of Metabolic Syndrome Project). *Am. J. Cardiol.* **95**: 194–198.
7. Yu, Y., D. F. Wyszynski, D. M. Waterworth, S. D. Wilton, P. J. Barter, Y. A. Kesäniemi, R. W. Mahley, R. McPherson, G. Waeber, T. P. Bersot, et al. 2005. Multiple QTLs influencing triglyceride and HDL and total cholesterol levels identified in families with atherogenic dyslipidemia. *J. Lipid Res.* **46**: 2202–2213.
8. Pennacchio, L. A., M. Olivier, J. A. Hubacek, J. C. Cohen, D. R. Cox, J.-C. Fruchart, R. M. Krauss, and E. M. Rubin. 2001. An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science*. **294**: 169–173.
9. van der Vliet, H. N., F. G. Schaap, J. H. M. Levels, R. Ottenhoff, N. Loofje, J. G. Wesseling, A. K. Groen, and R. A. F. M. Chamuleau. 2002. Adenoviral overexpression of apolipoprotein A-V reduces serum levels of triglycerides and cholesterol in mice. *Biochem. Biophys. Res. Commun.* **295**: 1156–1159.
10. Fruchart-Najib, J., E. Baugé, L.-S. Niculescu, T. Pham, B. Thomas, C. Rommens, Z. Majd, B. Brewer, L. A. Pennacchio, and J.-C. Fruchart. 2004. Mechanism of triglyceride lowering in mice expressing human apolipoprotein A5. *Biochem. Biophys. Res. Commun.* **319**: 397–404.
11. Schaap, F. G., P. C. N. Rensen, P. J. Voshol, C. Vriens, H. N. van der Vliet, R. A. F. M. Chamuleau, L. M. Havekes, A. K. Groen, and K. W. van Dijk. 2004. ApoAV reduces plasma triglycerides by inhibiting very low density lipoprotein-triglyceride (VLDL-TG) production and stimulating lipoprotein lipase-mediated VLDL-TG hydrolysis. *J. Biol. Chem.* **279**: 27941–27947.
12. Aouizerat, B. E., M. Kulkarni, D. Heilbron, D. Drown, S. Raskin, C. R. Pullinger, M. J. Malloy, and J. P. Kane. 2003. Genetic analysis of a polymorphism in the human apoA-V gene: effect on plasma lipids. *J. Lipid Res.* **44**: 1167–1173.
13. Austin, M. A., P. J. Talmud, F. M. Farin, D. A. Nickerson, K. L. Edwards, D. Leonetti, M. J. McNeely, H.-M. Viernes, S. E. Humphries, and W. Y. Fujimoto. 2004. Association of apolipoprotein A5 variants with LDL particle size and triglyceride in Japanese Americans. *Biochim. Biophys. Acta.* **1688**: 1–9.
14. Baum, L., B. Tomlinson, and G. N. Thomas. 2003. APOA5-1131T>C polymorphism is associated with triglyceride levels in Chinese men. *Clin. Genet.* **63**: 377–379.
15. Kao, J.-T., H.-C. Wen, K.-L. Chien, H.-C. Hsu, and S.-W. Lin. 2003. A novel genetic variant in the apolipoprotein A5 gene is associated with hypertriglyceridemia. *Hum. Mol. Genet.* **12**: 2533–2539.
16. Lai, C.-Q., E.-S. Tai, C. E. Tan, J. Cutter, S. K. Chew, Y.-P. Zhu, X. Adiconis, and J. M. Ordovas. 2003. The APOA5 locus is a strong determinant of plasma triglyceride concentrations across ethnic groups in Singapore. *J. Lipid Res.* **44**: 2365–2373.
17. Lai, C.-Q., S. Demissie, L. A. Cupples, Y. Zhu, X. Adiconis, L. D. Parnell, D. Corella, and J. M. Ordovas. 2004. Influence of the APOA5 locus on plasma triglyceride, lipoprotein subclasses, and CVD risk in the Framingham Heart Study. *J. Lipid Res.* **45**: 2096–2105.
18. Martin, S., V. Nicaud, S. E. Humphries, and P. J. Talmud. 2003. Contribution of APOA5 gene variants to plasma triglyceride determination and to the response to both fat and glucose tolerance challenges. *Biochim. Biophys. Acta.* **1637**: 217–225.
19. Nabika, T., S. Nasreen, S. Kobayashi, and J. Masuda. 2002. The genetic effect of the apoprotein AV gene on the serum triglyceride level in Japanese. *Atherosclerosis.* **165**: 201–204.
20. Pennacchio, L. A., M. Olivier, J. A. Hubacek, R. M. Krauss, E. M. Rubin, and J. C. Cohen. 2002. Two independent apolipoprotein A5 haplotypes influence human plasma triglyceride levels. *Hum. Mol. Genet.* **11**: 3031–3038.
21. Talmud, P. J., E. Howe, S. Martin, M. Olivier, G. J. Miller, E. M. Rubin, L. A. Pennacchio, and S. E. Humphries. 2002. Relative contribution of variation within the APOC3/A4/A5 gene cluster in determining plasma triglycerides. *Hum. Mol. Genet.* **11**: 3039–3046.
22. Klos, K. L. E., S. Hamon, A. G. Clark, E. Boerwinkle, K. Liu, and C. F. Sing. 2005. APOA5 polymorphisms influence plasma triglycerides in young, healthy African Americans and whites of the CARDIA Study. *J. Lipid Res.* **46**: 564–570.
23. Lee, K. W. J., A. F. Ayyobi, J. J. Frohlich, and J. S. Hill. 2004. APOA5 gene polymorphism modulates levels of triglyceride, HDL cholesterol and FER_{LDL}, but is not a risk factor for coronary artery disease. *Atherosclerosis.* **176**: 165–172.
24. Talmud, P. J., S. Martin, M.-R. Taskinen, M. H. Frick, M. S. Nieminen, Y. A. Kesäniemi, A. Pasternack, S. E. Humphries, and M. Syväne. 2004. APOA5 gene variants, lipoprotein particle distribution, and progression of coronary heart disease: results from the LOCAT study. *J. Lipid Res.* **45**: 750–756.
25. Szalai, C., M. Keszei, J. Duba, Z. Prohászka, G. T. Kozma, A. Császár, S. Balogh, Z. Almásy, G. Fust, and A. Czinner. 2004. Polymorphism in the promoter region of the apolipoprotein A5 gene is associated with an increased susceptibility for coronary artery disease. *Atherosclerosis.* **173**: 109–114.
26. Mahley, R. W., K. E. Palaoğlu, Z. Atak, J. Dawson-Pepin, A.-M. Langlois, V. Cheung, H. Onat, P. Fulks, L. L. Mahley, F. Vakar, et al. 1995. Turkish Heart Study. Lipids, lipoproteins, and apolipoproteins. *J. Lipid Res.* **36**: 839–859.
27. Bersot, T. P., S. J. Russell, S. R. Thatcher, N. K. Pomernacki, R. W. Mahley, K. H. Weisgraber, T. L. Innerarity, and C. S. Fox. 1993. A unique haplotype of the apolipoprotein B-100 allele associated with familial defective apolipoprotein B-100 in a Chinese man discovered during a study of the prevalence of this disorder. *J. Lipid Res.* **34**: 1149–1154.
28. Corbex, M., O. Poirier, F. Fumeron, D. Betoulle, A. Evans, J. B. Ruidavets, D. Arveiler, G. Luc, L. Tiret, and F. Cambien. 2000. Extensive association analysis between the CETP gene and coronary heart disease phenotypes reveals several putative functional polymorphisms and gene-environment interaction. *Genet. Epidemiol.* **19**: 64–80.
29. Schneider, S., D. Roessli, and L. Excoffier. 2000. Arlequin, Ver. 2.000: A Software for Population Genetics Data Analysis. (Genetics and Biometry Laboratory, University of Geneva, Switzerland, Accessed November 4, 2005 at <http://lgb.unige.ch/arlequin/>)
30. Thompson, E. A., S. Deeb, D. Walker, and A. G. Motulsky. 1988. The detection of linkage disequilibrium between closely linked markers: RFLPs at the AI-CIII apolipoprotein genes. *Am. J. Hum. Genet.* **42**: 113–124.
31. Mahley, R. W., J. Pépin, K. E. Palaoğlu, M. J. Malloy, J. P. Kane, and T. P. Bersot. 2000. Low levels of high density lipoproteins in Turks, a population with elevated hepatic lipase: high density lipoprotein characterization and gender-specific effects of apolipoprotein E genotype. *J. Lipid Res.* **41**: 1290–1301.
32. Mahley, R. W., P. Arslan, G. Pekcan, G. M. Pépin, A. Agaçdiken, N. Karaagaoglu, N. Rakıcıoğlu, B. Nursal, P. Dayanikli, K. E. Palaoğlu, et al. 2001. Plasma lipids in Turkish children: impact of puberty, socioeconomic status, and nutrition on plasma cholesterol and HDL. *J. Lipid Res.* **42**: 1996–2006.
33. Onat, A. 2001. Risk factors and cardiovascular disease in Turkey. *Atherosclerosis.* **156**: 1–10.
34. Onat, A. 2004. Lipids, lipoproteins and apolipoproteins among Turks, and impact on coronary heart disease. *Anadolu Kardiyol. Derg.* **4**: 236–245.
35. Assmann, G., and H. Schulte. 1992. Relation of high-density lipoprotein cholesterol and triglycerides to incidence of atherosclerotic coronary artery disease (the PROCAM experience). *Am. J. Cardiol.* **70**: 733–737.
36. Burchfiel, C. M., A. Laws, R. Benfante, R. J. Goldberg, L. J. Hwang, D. Chiu, B. L. Rodriguez, J. D. Curb, and D. S. Sharp. 1995. Combined effects of HDL cholesterol, triglyceride, and total cholesterol concentrations on 18-year risk of atherosclerotic disease. *Circulation.* **92**: 1430–1436.
37. Couture, P., J. D. Otvos, L. A. Cupples, P. W. F. Wilson, E. J. Schaefer, and J. M. Ordovas. 1999. Association of the A-204C polymorphism in the cholesterol 7 α -hydroxylase gene with variations in plasma low density lipoprotein cholesterol levels in the Framingham Offspring Study. *J. Lipid Res.* **40**: 1883–1889.
38. Hodoglugil, U., D. W. Williamson, Y. Huang, and R. W. Mahley. 2005. Common polymorphisms of ATP binding cassette transporter A1, including a functional promoter polymorphism, associated with plasma high density lipoprotein cholesterol levels in Turks. *Atherosclerosis.* **183**: 199–212.
39. Ludwig, E. H., R. W. Mahley, E. Palaoğlu, S. Özbayraktı, M. E. Balestra, I. B. Borecki, T. L. Innerarity, and R. V. Farese, Jr. 2002. DGAT1 promoter polymorphism associated with alterations in body mass index, high density lipoprotein levels and blood pressure in Turkish women. *Clin. Genet.* **62**: 68–73.
40. Ordovas, J. M. 2000. Lipoprotein lipase genetic variation and gender-specific ischemic cerebrovascular disease risk. *Nutr. Rev.* **58**: 315–323.
41. Hubáček, J. A., V. Adámková, R. Ceska, R. Poledne, A. Horánek, and M. Vráblík. 2004. New variants in the apolipoprotein AV gene

- in individuals with extreme triglyceride levels. *Physiol. Res.* **53**: 225–228.
42. Talmud, P. J., J. Palmen, W. Putt, L. Lins, and S. E. Humphries. 2005. Determination of the functionality of common *APOA5* polymorphisms. *J. Biol. Chem.* **280**: 28215–28220.
43. Groenendijk, M., R. M. Cantor, T. W. A. de Bruin, and G. M. Dallinga-Thie. 2001. The apoA1-CIII-AIV gene cluster. *Atherosclerosis*. **157**: 1–11.
44. Kim, J. Q., J. Song, Y. B. Park, and S. H. Hong. 1998. Molecular bases of coronary heart disease in Koreans. *J. Korean Med. Sci.* **13**: 1–15.
45. van Dijk, K. W., P. C. N. Rensen, P. J. Voshol, and L. M. Havekes. 2004. The role and mode of action of apolipoproteins CIII and AV: synergistic actors in triglyceride metabolism? *Curr. Opin. Lipidol.* **15**: 239–246.
46. Baroukh, N., E. Bauge, J. Akiyama, J. Chang, V. Afzal, J. C. Fruchart, E. M. Rubin, J. Fruchart-Najib, and L. A. Pennacchio. 2004. Analysis of apolipoprotein A5, C3, and plasma triglyceride concentrations in genetically engineered mice. *Arterioscler. Thromb. Vasc. Biol.* **24**: 1297–1302.
47. Olivier, M., X. Wang, R. Cole, B. Gau, J. Kim, E. M. Rubin, and L. A. Pennacchio. 2004. Haplotype analysis of the apolipoprotein gene cluster on human chromosome 11. *Genomics*. **83**: 912–923.
48. Austin, M. A., J. E. Hokanson, and K. L. Edwards. 1998. Hypertriglyceridemia as a cardiovascular risk factor. *Am. J. Cardiol.* **81** (Suppl.): 7B–12B.

Chapter 5: An interaction between the *TaqIB* polymorphism of cholesterol ester transfer protein and smoking is associated with changes in plasma high-density lipoprotein cholesterol levels in Turks

Clin Genet 2005; 68: 118–127
Printed in Singapore. All rights reserved

Copyright © Blackwell Munksgaard 2005
CLINICAL GENETICS
doi:10.1111/j.1399-0004.2005.00467.x

Original Article

An interaction between the *TaqIB* polymorphism of cholesterol ester transfer protein and smoking is associated with changes in plasma high-density lipoprotein cholesterol levels in Turks

Hodoğlugil U, Williamson DW, Huang Y, Mahley RW. An interaction between the *TaqIB* polymorphism of cholesterol ester transfer protein and smoking is associated with changes in plasma high-density lipoprotein cholesterol levels in Turks. *Clin Genet* 2005; 68: 118–127. © Blackwell Munksgaard, 2005

Low levels of high-density lipoprotein cholesterol (HDL-C) are an independent risk factor for atherosclerosis. We investigated the effects of the *TaqIB* polymorphism of cholesterol ester transfer protein (CETP) on CETP activity and plasma HDL-C levels in random nondiabetic and self-reported diabetic subjects in a population with very low HDL-C levels. The rare B2B2 genotype was associated with significantly higher HDL-C levels and lower CETP activity in random subjects and with higher HDL-C in diabetic subjects. After stratification of random subjects by smoking status, the common B1B1 genotype was associated with lower HDL-C levels than the B2B2 genotype. Although smoking was associated with lower HDL-C, especially in men, HDL-C levels between smokers and nonsmokers were not different in subjects with the B1B2 or B2B2 genotypes. However, smoking (20+ cigarettes/day) was associated with a marked reduction in HDL-C in the B1B1 subjects. The B1B1/smoking interaction was not reflected in a difference in CETP activity. High triglycerides and elevated body mass index (BMI) lower HDL-C. The B2B2 genotype was associated with the highest HDL-C levels, and these levels were significantly lower in the hypertriglyceridemic subjects (\geq 50th percentile). The lowest HDL-C levels were seen in hypertriglyceridemic subjects with the B1B1 genotype. Although BMI (\geq 50th vs $<$ 50th percentile) did not affect HDL-C in B2B2 subjects, a high BMI was associated with markedly lower HDL-C in B1B1 subjects. Thus, HDL-C levels in Turks may be modulated by an interaction between the CETP *TaqIB* polymorphism and smoking, as well as an interaction with hypertriglyceridemia and BMI.

**U Hodoğlugil^{a,b},
DW Williamson^e, Y Huang^{a,b,c}
and RW Mahley^{a,b,c,d}**

^aGladstone Institute of Cardiovascular Disease, ^bCardiovascular Research Institute, ^cDepartment of Pathology, and ^dDepartment of Medicine, University of California, San Francisco, CA, USA, ^eGraduate Program in Biological and Medical Informatics

Key words: body mass index – cholesterol ester transfer protein – diabetes – HDL cholesterol – polymorphism – *TaqIB* – smoking – triglyceride – Turkish population

Corresponding author: Robert W. Mahley, MD, PhD, Gladstone Institute of Cardiovascular Disease, 1650 Owens Street, San Francisco, CA 94158, USA. Tel.: +1 415 734 2062; fax: +1 415 355 0820; e-mail: rmahley@gladstone.ucsf.edu

Received 14 January 2003, revised and accepted for publication 5 April 2005

Coronary heart disease is the leading cause of death worldwide (1), and altered lipoprotein levels are pivotal risk factors for atherosclerosis (2, 3). In particular, low levels of high-density lipoprotein cholesterol (HDL-C) are a major independent risk factor for atherosclerosis (4, 5). Smoking, another major risk factor, has a direct effect on plasma lipids and lipoproteins

and potentially on atherogenesis and thrombosis (6, 7).

HDL metabolism is significantly influenced by cholesterol ester transfer protein (CETP), which facilitates the transfer of cholesteryl esters from HDL to low-density lipoproteins (LDLs) and very low density lipoproteins (VLDLs) in exchange for triglycerides, thereby decreasing

CETP *TaqIB* genotype, smoking, and HDL-C

the levels of protective HDL-C and increasing the levels of pro-atherogenic LDL-C (8). Patients with mutations in CETP that reduce its activity have abnormally high plasma HDL-C levels (9–11). Although these mutations are rare, CETP polymorphisms can also influence plasma lipoprotein concentrations. The most studied CETP polymorphism is *TaqIB*, a silent base change in the first intron. The rare allele B2 is associated with increased HDL-C levels and decreased CETP activity and levels in most populations studied (12–17) but not all (18, 19). Similar findings have been reported in most (17, 20, 21) but not all (22) studies in diabetic patients. In addition, the association of the *TaqIB* polymorphism with plasma HDL-C is highly influenced by environmental factors, such as alcohol consumption (13, 15) and smoking (23–26).

Studies of the effects of smoking on CETP activity have yielded conflicting results. CETP activity was higher (27, 28) or lower (29, 30) in smokers or the same as in nonsmokers (31, 32). In this study, we investigated the effects of interactions between the *TaqIB* polymorphism and environmental factors, particularly smoking, on plasma HDL-C levels in self-reported diabetic and random nondiabetic subjects from the Turkish Heart Study (THS) (33). This large-scale epidemiological survey of more than 9000 volunteers from six regions of Turkey showed that Turks have very low levels of plasma HDL-C and a high prevalence of smoking, making this population ideal for studying interactions between smoking and genes that influence HDL-C levels (33–36).

Methods

Study population

The study population consisted of 2011 subjects randomly selected from nondiabetic subjects and 187 subjects with self-reported adult-onset diabetes who participated in the THS (33). Detailed biodata were obtained from each participant. The protocol was approved by the Committee on Human Research of the University of California, San Francisco, and was in accordance with the Helsinki Declaration. Subjects who were taking any lipid-lowering medication were excluded.

Biochemical analyses

Blood samples were obtained after an overnight fast. Total cholesterol and triglyceride levels were

determined by enzymatic colorimetric methods. HDL-C levels were determined with the CHOD-PAP method after phosphotungstic acid – magnesium precipitation of VLDLs and LDLs (33). LDL-C was calculated by the Friedewald formula (37) for participants with triglyceride levels < 400 mg/dl. CETP activity, previously measured in a subgroup of the study population, is expressed as percent cholesteryl ester transfer (36).

Genotyping

CETP genotyping was performed as described (15); its accuracy was evaluated by randomly inserting duplicate DNA samples in the assays (approximately 6% replication). Genotyping discrepancies were found in less than 1% of the samples and were resolved by rescoring or eliminating the data.

Data analysis

The data were analyzed with Microsoft Access, Excel, and spss 10.0. Associations between genotypes, lipids and other parameters were analyzed separately for males and females. Allele frequencies were calculated by the gene-counting method. All values are reported as mean \pm SD. As triglyceride was not normally distributed, log-transformed values were used for statistical comparison. Mean values were compared by two-tailed *t* test according to genotype; $p < 0.05$ was considered significant. Subjects were categorized by smoking status (nonsmokers, 1–19 cigarettes/day, 20+ cigarettes/day) and alcohol consumption (nondrinkers, 1–5 drinks/week, > 5 drinks/week). Univariate analysis of variance was used to construct the model to explain the variation in HDL-C levels. Plasma triglyceride levels, body mass index (BMI), smoking, and alcohol consumption were included as covariates, and genotype score was included as a fixed factor in the model (GLM Univariate, spss 10.0). χ^2 analysis was used to test differences between the observed and expected frequencies of alleles (assuming a Hardy–Weinberg equilibrium) and to compare genotype or allele frequency distribution after stratification by HDL-C levels.

Results

Population characteristics

The demographic and biochemical characteristics of the study subjects are shown in Table 1.

Table 1. Demographic and biochemical characteristics of random and diabetic subjects by sex

Characteristics	Random subjects		p	Diabetic subjects		p
	Females (n = 792)	Males (n = 1219)		Females (n = 66)	Males (n = 121)	
Age (years)	42 ± 14	41 ± 12	NS	54 ± 13	52 ± 12	NS
BMI (kg/m ²)	26.1 ± 5.3	25.8 ± 3.8	NS	31.7 ± 6.9	26.9 ± 3.8	< 0.001
HDL-C (mg/dl)	41 ± 8	35 ± 7	< 0.001	38 ± 8	36 ± 7	NS
Total cholesterol (mg/dl)	184 ± 48	190 ± 45	< 0.05	199 ± 51	200 ± 48	NS
LDL-C (mg/dl)	120 ± 41	126 ± 40	< 0.05	125 ± 41	127 ± 39	NS
Triglycerides (mg/dl)	117 ± 73	150 ± 92	< 0.001	201 ± 110	194 ± 126	NS
TC/HDL-C ratio	4.7 ± 1.6	5.6 ± 1.6	< 0.001	5.4 ± 1.6	5.8 ± 1.9	NS
SBP (mm Hg)	126 ± 24	125 ± 21	NS	149 ± 25	137 ± 23	< 0.01
DBP (mm Hg)	81 ± 16	82 ± 14	NS	91 ± 16	86 ± 15	< 0.05
Alcohol consumption (%) ^a	6.1	27.7	< 0.001	5.1	33.3	< 0.05
Cigarette smoking (%) ^b	26.5	57.4	< 0.001	19.7	48.8	< 0.05

BMI, body mass index; DBP, diastolic blood pressure; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; NS, not significant; SBP, systolic blood pressure; TC, total cholesterol.

Values are mean ± SD or percentages. Means were compared by *t* test, and percentages were analyzed by χ^2 test.

^aOne or more drinks/week.

^bOne or more cigarettes/day.

Plasma HDL-C levels were very low, and total cholesterol/HDL-C ratios were high in both men and women in the random and diabetic groups. Even though plasma total cholesterol and LDL-C levels were not excessively elevated, the low HDL-C levels could represent a significant risk factor for heart disease in the Turkish population (33, 35). Detailed analyses of the THS data have been reported (33, 34, 36, 38).

Association of the *TaqIB* polymorphism with plasma HDL-C and CETP activity

In the random population, the CETP *TaqIB* polymorphism was associated with altered plasma HDL-C levels in both men and women. The B2B2 genotype was associated with 8–9% higher HDL-C than the B1B1 genotype ($p < 0.001$, Table 2). The frequency of the rare B2 allele was 44.4% (frequency varied from 41.6 to 46.2% among regions). Stratification by HDL-C level confirmed that the B2B2 genotype and the B2 allele were significantly more common in random subjects with high HDL-C levels than in those with low HDL-C levels (Fig. 1).

In subjects with adult-onset diabetes mellitus, the B2B2 genotype was also associated with significantly higher levels of HDL-C (Table 2). The B2 allele was more frequent in those with high HDL-C than with low HDL-C (males: 50.0 vs 36.7%, females: 50.0 vs 38.9%).

Plasma CETP activity (percent cholesteryl ester transfer) was measured in a subset of the random Turkish population (36). These values were 30.7 ± 7.4 (n = 34) in B1B1, 28.5 ± 7.0 (n = 66)

in B1B2, and 22.4 ± 5.7 (n = 26) in B2B2 women and 27.9 ± 7.2 (n = 32) in B1B1, 27.1 ± 8.4 (n = 76) in B1B2, and 22.9 ± 6.3 (n = 21) in B2B2 men. In both sexes, plasma CETP activity was significantly lower in B2B2 subjects ($p < 0.05$).

The effects of plasma triglyceride levels and BMI on HDL-C levels in association with the *TaqIB* polymorphism were examined in the random population (Table 2). As expected, high triglyceride levels and increased BMI were associated with lower HDL-C levels. The HDL-C levels were lowest in B1B1 subjects with the highest triglycerides and BMI (≥ 50 th percentile) and highest in B2B2 subjects with the lowest triglyceride levels; however, HDL-C levels were lower in the B2B2 subjects with high triglycerides. BMI did not affect HDL-C levels in B2B2 subjects (Table 2).

CETP activity did not differ between the < 50 th and ≥ 50 th percentiles for triglycerides in both genders [males: < 50 th, 25.6 ± 7.3 (n = 50), ≥ 50 th, 27.9 ± 8.2 (n = 79); females: < 50 th, 27.0 ± 7.1 (n = 53), ≥ 50 th, 28.9 ± 8.5 (n = 70)]. Likewise, CETP activity was not different with respect to BMI [males: < 50 th, 26.4 ± 8.2 (n = 55), ≥ 50 th, 26.9 ± 7.8 (n = 79); females: < 50 th, 28.7 ± 8.5 (n = 78), ≥ 50 th, 27.8 ± 6.9 (n = 50)].

After controlling for sex, triglyceride level, BMI, smoking, alcohol consumption, and region from which samples were obtained, the *TaqIB* polymorphism was still significantly associated with HDL-C levels ($p < 0.05$) and explained 7.6% of the variation in HDL-C in the random population and 5.2% in the diabetic group. In an

Table 2. CETP TaqIB polymorphism and mean HDL-C levels in random and diabetic Turkish subjects

	Mean HDL-C (mg/dl ± SD)				p	% increase in HDL-C	
	All	B1B1	B1B2	B2B2		B1B1 vs B2B2	B2B2 vs B1B1
Women							
Random (all)	41.0 ± 8.1 (792)	39.9 ± 7.4 (247)	40.9 ± 7.7 (390)	43.0 ± 9.6 (155)	< 0.001		8.0
Triglycerides (percentile)							
< 50th	43.2 ± 8.4 (364)	42.5 ± 7.2 (124)	42.5 ± 7.8 (174)	46.2 ± 11.1 (66)	< 0.001		8.6
≥ 50th	39.0 ± 7.0 (421)	37.2 ± 6.6 (121)	39.4 ± 6.9 (213)	40.4 ± 7.3 (87)	< 0.001		8.7
p	< 0.001	< 0.001	< 0.001	< 0.001			
BMI (percentile)							
< 50th	42.7 ± 8.7 (390)	41.9 ± 7.8 (119)	42.8 ± 8.3 (185)	43.7 ± 11.0 (76)	NS		4.2
≥ 50th	39.2 ± 6.9 (390)	37.8 ± 6.5 (124)	38.9 ± 6.4 (190)	42.1 ± 8.0 (76)	< 0.001		11.2
p	< 0.001	< 0.001	< 0.001	NS			
Diabetic	38.0 ± 8.4 (66)	36.9 ± 7.3 (23)	37.5 ± 9.0 (31)	41.3 ± 9.1 (12)	< 0.05		11.9
Men							
Random (all)	35.2 ± 6.5 (1219)	34.1 ± 5.5 (378)	35.0 ± 6.6 (594)	37.2 ± 7.5 (246)	< 0.005		9.1
Triglycerides (percentile)							
< 50th	37.0 ± 6.8 (529)	35.5 ± 5.4 (152)	36.7 ± 6.6 (260)	39.7 ± 8.2 (117)	< 0.005		11.6
≥ 50th	33.0 ± 5.9 (686)	33.0 ± 5.3 (226)	33.7 ± 6.2 (332)	34.8 ± 6.0 (128)	< 0.05		5.3
p	< 0.001	< 0.001	< 0.001	< 0.001			
BMI (percentile)							
< 50th	36.2 ± 6.5 (592)	35.3 ± 5.7 (183)	36.1 ± 6.7 (284)	37.9 ± 6.8 (125)	< 0.005		7.3
≥ 50th	34.2 ± 6.4 (595)	32.9 ± 4.9 (187)	34.2 ± 6.3 (290)	36.4 ± 8.1 (118)	< 0.001		10.5
p	< 0.001	< 0.001	< 0.001	NS			
Diabetic	36.1 ± 8.2 (121)	34.6 ± 7.6 (39)	36.2 ± 8.4 (57)	38.0 ± 8.3 (25)	< 0.05		9.8

BMI, body mass index; CETP, cholesteryl ester transfer protein; HDL-C, high-density lipoprotein cholesterol; NS, not significant. p values were determined by t test. Values in parentheses are numbers of subjects.

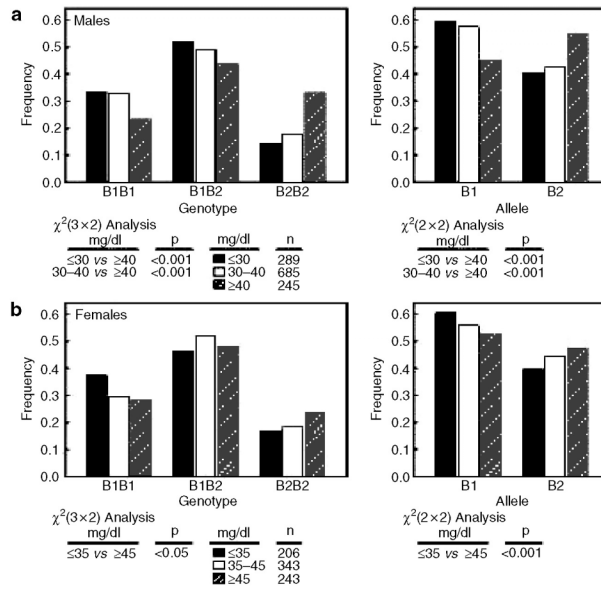


Fig. 1. Genotypic (left) and allelic (right) frequencies of cholesterol ester transfer protein *TaqIB* polymorphisms in males (a) and females (b) with different high-density lipoprotein cholesterol (HDL-C) levels (mg/dl). Males and females were separately grouped according to HDL-C level (low, medium, and high), and the genotype and allele frequencies were evaluated by χ^2 analysis. The results are given below each panel.

effort to determine whether any of the significant sources of variables interact with the *TaqIB* genotype, interaction variables were input in the statistical model for the random population. There was a significant interaction between smoking and the *TaqIB* genotype ($p < 0.05$) on HDL-C levels, whereas no interaction was found between region and the *TaqIB* genotype or between region and smoking (with or without the *TaqIB* genotype in the model as a covariate). Separate covariate analyses of males and females yielded similar results (data not shown).

Interaction between lifestyle factors and the *TaqIB* polymorphism: effect on HDL-C levels in random subjects

Smoking was associated with a statistically significant decrease in plasma HDL-C levels in males (Table 3). In both men and women, HDL-C levels were significantly higher in B2B2 than B1B1 subjects in all groups of smokers and nonsmokers (Table 3). This increase in HDL-C B2B2 subjects was much greater in men and women who smoked 20+ cigarettes/day. Interestingly, in both male and female B1B1 subjects, HDL-C levels were significantly reduced in those who smoked 20+ cigarettes/day ($p < 0.005$). HDL-C levels did not differ

between smokers and nonsmokers with either the B1B2 or B2B2 genotype (carriers of B2 allele). Thus, the B1B1 genotype appears to interact with smoking to lower plasma HDL-C levels, while the B1B2 and the B2B2 subjects were resistant to smoking-induced decreases in HDL-C levels.

CETP activity did not differ in men and women (26.6 ± 7.9 , $n = 129$, vs 27.9 ± 7.4 , $n = 126$, respectively) or between males and females within a specific *TaqIB* genotype. Therefore, the CETP activity data for both sexes were combined to improve statistical power. CETP activity did not differ according to smoking status alone (Table 3). However when stratified by smoking status and *TaqIB* genotype, CETP activity tended to be lower only among B2B2 subjects in the two groups of smokers. There was a statistical trend in CETP activity between nonsmokers and the 1-19 cigarettes/day group in the B2B2 subjects ($p = 0.063$, Table 3). When data from all smokers were combined, the CETP activity was significantly lower in smokers than nonsmokers only in B2B2 subjects ($p < 0.05$). Thus, an interaction between the *TaqIB* genotype and smoking appears to affect plasma CETP activity.

Alcohol consumption was associated with significantly increased plasma HDL-C levels in

CETP *TaqIB* genotype, smoking, and HDL-C

Table 3. Interaction between CETP *TaqIB* polymorphism and smoking on HDL-C levels and CETP activity in a random Turkish population

Cigarettes/day	All	B1B1	B1B2	B2B2	p		% increase in HDL-C
					B1B1 vs B2B2	B2B2 vs B1B1	
HDL-C (mg/dl ± SD)							
Females							
0	40.9 ± 7.8 (559)	40.2 ± 7.3 (180)	40.8 ± 7.8 (268)	42.3 ± 8.5 (111)	< 0.05		5.2
1-19	41.2 ± 8.2 (168)	40.3 ± 8.2 (48)	40.9 ± 7.3 (90)	43.4 ± 10.2 (30)	< 0.05		7.7
20+	40.9 ± 9.8 (65)	35.5 ± 4.0 (19)	41.3 ± 7.4 (32)	43.6 ± 10.2 (14)	< 0.05		23.2
p (0 vs 20+)	NS	< 0.005	NS	NS			
Males							
0	36.2 ± 6.6 (509)	35.1 ± 5.7 (162)	35.8 ± 6.8 (241)	37.1 ± 6.9 (106)	< 0.05		5.7
1-19	34.9 ± 6.2 (335)	34.0 ± 4.6 (101)	34.5 ± 6.5 (173)	37.5 ± 7.1 (61)	< 0.001		10.3
20+	34.9 ± 6.7 (375)	32.8 ± 5.5 (116)	34.5 ± 6.1 (180)	37.0 ± 8.7 (79)	< 0.001		12.8
p (0 vs 20+)	0.01	0.005	NS	NS			
CETP activity (% transfer ± SD)							
Females + Males							
All	27.4 ± 7.9 (255)	29.8 ± 8.0 (66)	27.7 ± 7.8 (142)	23.1 ± 6.4 (47)	< 0.01		-22.7
0	27.9 ± 7.9 (130)	29.5 ± 8.0 (36)	28.3 ± 8.2 (72)	24.4 ± 6.3 (22)	< 0.05		-17.3
1-19	26.0 ± 7.1 (106)	28.8 ± 6.6 (25)	26.8 ± 7.1 (58)	21.6 ± 5.4 (23)	< 0.05		-25.0
20+	28.3 ± 8.6 (19)	30.9 ± 7.6 (5)	29.0 ± 8.2 (12)	17.4 ± 9.3 (2)	^a		-43.0
p (0 vs 1-19)	NS	NS	NS	0.063			

CETP, cholesterol ester transfer protein; HDL-C, high-density lipoprotein cholesterol; NS, not significant. Values in parentheses are numbers of subjects. p values were determined by *t* test. ^an values was too low for statistical comparisons.

males only ($p < 0.01$, data not shown). However, there was no interaction between alcohol consumption and the *TaqIB* genotype on HDL-C levels.

Discussion

The main lipid characteristic of Turks is extremely low levels of plasma HDL-C (33, 36, 38). Low levels of HDL-C have also been found in Turks living in Germany, the Netherlands, and the United States (36, 39, 40), suggesting that genetics plays a significant role. This important finding makes Turks an ideal population for studying genes that influence HDL-C levels. More than 70% of Turkish males and 50% of Turkish females have HDL-C < 40 mg/dl (33). In the present study, we analyzed the effect of the CETP *TaqIB* polymorphism on HDL-C levels and its interaction with environmental factors, particularly smoking, which reduces plasma HDL-C levels (6, 7).

This study shows that the B2B2 genotype of the CETP *TaqIB* polymorphism is associated with increased plasma HDL-C levels and decreased CETP activity in Turkish men and women. The association between the CETP *TaqIB* polymorphism and HDL-C (13, 14, 26, 41-44) and CETP activity (14, 16, 43) has been reported. While most studies have confirmed this

association, some have not (18, 19). In our study, the HDL-C levels were 8-9% higher in B2B2 than B1B1 subjects, consistent with previous reports (13, 14, 26, 41-44). Likewise, the higher HDL-C levels in the B2B2 subjects were associated with 18% of lower CETP activity in men and 27% of lower CETP activity in women.

The low HDL-C levels in Turks are not associated with a high prevalence of diabetes mellitus or insulin resistance. In a large cross-sectional study in Turkey, the prevalence of diabetes was 7.2% (45), and data from the THS suggest that it may be lower (33). The B2B2 genotype was also associated with elevated HDL-C levels in diabetic Turkish subjects, as previously reported (17). Elevated HDL-C and reduced CETP activity or mass in association with the *TaqI* B2B2 genotype have been shown in two studies of diabetics (20, 21). However, another study showed an association between HDL-C levels and *TaqIB* in diabetic males but not females (46), and yet another showed no association (22).

The *TaqIB* genotype has been associated with plasma HDL-C levels in both smokers and nonsmokers (47, 48) and smokers only (24-26). The B2B2 genotype has been associated with reduced risk of coronary heart disease (14, 23, 44, 49) but only in nonsmokers (23). In Turks, the B2B2 genotype clearly affected HDL-C levels in both smokers and nonsmokers. Interestingly, the HDL-C-lowering effect of smoking was most pronounced in

B1B1 subjects; B2B2 subjects appeared to be protected. Among B1B1 subjects, HDL-C levels were lower in smokers (20+ cigarettes/day) than in nonsmokers (7% in males and 13% in females).

Many epidemiological studies have shown that smoking affects lipoprotein profiles (6, 7). In a meta-analysis, heavy smokers have, on average, 9% lower HDL-C levels than matched nonsmokers (50). However, studies of the effects of smoking have yielded inconsistent results, with CETP activity being either higher (27, 28) or lower (29, 30) in smokers. This discrepancy may reflect differences in population-specific characteristics, environmental factors, selection criteria, and sample sizes. Although we and others (31, 32) observed no differences in CETP activity between smokers and nonsmokers, the *TaqIB* genotype appeared to interact with smoking to affect CETP activity, which was significantly lower in smokers with the B2B2 genotype. This finding may help explain why the HDL-C-lowering effect of smoking was seen only in the B1B1 group and not in the B1B2 and B2B2 groups.

The interaction between the CETP *TaqIB* polymorphism and smoking may be especially important in the Turkish population, where the prevalence of smoking has increased over the last 20 years (51). More than one half of males (up to 70%) and one quarter (up to 43%) of females smoke (33). The B1B1 genotype, found in about 31% of the population, was associated with low HDL-C (and high CETP activity), and the HDL-C-lowering effect was magnified in smokers. At least a component of low HDL-C in Turks may reflect the interaction between the CETP *TaqIB* polymorphism and smoking.

The HDL-C levels were highest in B2B2 subjects with low triglyceride and low BMI. Likewise, in a Taiwanese Chinese population, HDL-C was highest in those B2B2 subjects with low BMI ($\leq 26 \text{ kg/m}^2$) and low triglyceride ($\leq 150 \text{ mg/dl}$) (42). This association was modified by triglyceride levels (lower HDL-C associated with higher triglycerides) but not by BMI (high vs low). The lowest HDL-C levels were seen in males and females with the highest triglycerides and BMI in the B1B1 subjects. As CETP activity is not different in individuals with low and high triglycerides in this or other studies (52, 53), the interaction of plasma triglycerides and CETP polymorphism on HDL-C may be independent of CETP activity. On the other hand, CETP activity has been shown to be increased with obesity (54). Additional studies are required to understand the interactions among plasma triglycerides, obesity, HDL-C, CETP activity, and CETP *TaqIB* polymorphism.

Other environmental, lifestyle, and genetic factors also modulate the effects of the *TaqIB* polymorphism, including alcohol consumption in some (13, 15) but not all (14, 16, 47) studies, and apolipoprotein E genotype, which had an effect in one study in children (55) but not in others (56, 57). In our analysis, interaction between the CETP *TaqIB* polymorphism and alcohol consumption or apolipoprotein E genotype had no effect on HDL-C levels.

The *TaqIB* polymorphism is in strong linkage disequilibrium with G-971A and C-629A (12, 13, 58-61). The rare -971A and -629A alleles are associated with lower CETP mass (12, 13, 58, 59) and higher HDL-C levels (12, 13, 58, 59, 62). Furthermore, -629A has lower transcriptional activity *in vitro* than -629C (58). Thus, the *TaqIB* polymorphism may be a marker for the C-629A promoter polymorphism (61).

The possible role of CETP in atherogenesis and the potential antiatherogenic effects of inhibiting CETP activity (63, 64) have led to the development of CETP inhibitors, two of which were recently tested in humans. Both JTT-705 (65) and torcetrapib (66, 67) significantly increased HDL-C and decreased LDL-C. Studies in Japan, where there is an increased incidence of marked CETP deficiency, suggest that CETP deficiency is atherogenic (67). However, partial inhibition of CETP may not result in an atherogenic lipid profile (68), as residual CETP activity may prevent the accumulation of very large abnormal HDL and LDL particles characteristic of Japanese patients (67). B2B2 subjects may have an optimal lower level of CETP activity, allowing HDL-C to remain high even under conditions that might otherwise cause HDL-C to be lower. The approximately 20% reduction in CETP activity in B2B2 subjects we observed may protect against the HDL-C-lowering effect of smoking. Alternatively, in B1B1 subjects with higher CETP activity, HDL-C may be more susceptible to modulating factors such as smoking. Further studies are necessary to determine whether pharmacological inhibitors of CETP reduce the risk of atherosclerosis and whether there is an optimal level of CETP activity that modulates its effects on HDL-C levels and other coronary heart disease risk factors.

In summary, the common *TaqIB* polymorphism of CETP was associated with altered plasma lipid levels in randomly selected and diabetic Turkish subjects from the THS. The B2B2 genotype of the *TaqIB* polymorphism was associated with high plasma HDL-C levels and appeared to protect against the HDL-C-lowering effects of smoking. The more common B1B1 genotype was associated

with significantly lower HDL-C levels in smokers. The B2B2 genotype was associated with a 5.2–23.2% increase in plasma HDL-C, depending on smoking status, and the B1B1 genotype was associated with lower HDL-C levels. Triglyceride levels and BMI also interacted with the polymorphism and altered HDL-C levels. These observations may be significant in assessing the risk of coronary artery disease in Turks, as a 1% increase in plasma HDL-C level is associated with a 2–3% decrease in cardiovascular morbidity and mortality (69).

Acknowledgements

We are indebted to our associates at the American Hospital, Istanbul, especially Guy M. Pépin, Sibel Tanir, and Linda L. Mahley in the Gladstone Institute (Istanbul), and Dr K. Erhan Palaoglu at the American Hospital (Istanbul). We thank Sylvia Richmond and Jennifer Polizzotto for manuscript preparation and Stephen Ordway and Gary Howard for editorial assistance. We acknowledge the support of the American Hospital, especially Mr George Rountree and the J. David Gladstone Institutes. This work was supported in part by grants HL71027 and HL64162 from the National Institutes of Health and by grant 12FT-0226 from the California Tobacco-Related Disease Research Program. Preliminary data were presented at the XIIIth International Symposium on Atherosclerosis in Kyoto, Japan (September 2003).

References

1. Murray CJL, Lopez AD. Mortality by cause for eight regions of the world: Global Burden of Disease study. *Lancet* 1997; 349: 1269–1276.
2. van Lennep JER, Westerveld HT, Erkelens DW et al. Risk factors for coronary heart disease: implications of gender. *Cardiovasc Res* 2002; 53: 538–549.
3. LaRosa JC. Triglycerides and coronary risk in women and the elderly. *Arch Intern Med* 1997; 157: 961–968.
4. Gordon DJ, Probstfield JL, Garrison RJ et al. High-density lipoprotein cholesterol and cardiovascular disease. Four prospective American studies. *Circulation* 1989; 79: 8–15.
5. Jacobs DR Jr, Mebane IL, Bangdiwala SI et al. High density lipoprotein cholesterol as a predictor of cardiovascular disease mortality in men and women: the follow-up study of the Lipid Research Clinics Prevalence Study. *Am J Epidemiol* 1990; 131: 32–47.
6. Bolego C, Poli A, Paoletti R. Smoking and gender. *Cardiovasc Res* 2002; 53: 568–576.
7. Cullen P, Schulte H, Assmann G. Smoking, lipoproteins and coronary heart disease risk. Data from the Münster Heart Study (PROCAM). *Eur Heart J* 1998; 19: 1632–1641.
8. Bruce C, Chouinard RA Jr, Tall AR. Plasma lipid transfer proteins, high-density lipoproteins, and reverse cholesterol transport. *Annu Rev Nutr* 1998; 18: 297–330.
9. Takegoshi T, Haba T, Kitoh C et al. Compound heterozygote of cholesteryl-ester transfer protein deficiency in a patient with hyperalphalipoproteinemia. *Atherosclerosis* 1992; 96: 83–85.

10. Inazu A, Jiang X-C, Haraki T et al. Genetic cholesteryl ester transfer protein deficiency caused by two prevalent mutations as a major determinant of increased levels of high density lipoprotein cholesterol. *J Clin Invest* 1994; 94: 1872–1882.
11. Hill SA, Nazir DJ, Jayaratne P et al. Mutations in cholesteryl ester transfer protein and hepatic lipase in a North American population. *Clin Biochem* 1997; 30: 413–418.
12. Le Goff W, Guerin M, Nicaud V et al. A novel cholesteryl ester transfer protein promoter polymorphism (–971G/A) associated with plasma high-density lipoprotein cholesterol levels. Interaction with the *TaqIB* and –629C/A polymorphisms. *Atherosclerosis* 2002; 161: 269–279.
13. Corbex M, Poirier O, Fumeron F et al. Extensive association analysis between the CETP gene and coronary heart disease phenotypes reveals several putative functional polymorphisms and gene-environment interaction. *Genet Epidemiol* 2000; 19: 64–80.
14. Ordovas JM, Cupples LA, Corella D et al. Association of cholesteryl ester transfer protein *TaqIB* polymorphism with variations in lipoprotein subclasses and coronary heart disease risk. The Framingham Study. *Arterioscler Thromb Vasc Biol* 2000; 20: 1323–1329.
15. Fumeron F, Betoulle D, Luc G et al. Alcohol intake modulates the effect of a polymorphism of the cholesteryl ester transfer protein gene on plasma high density lipoprotein and the risk of myocardial infarction. *J Clin Invest* 1995; 96: 1664–1671.
16. Kauma H, Savolainen MJ, Heikkilä R et al. Sex difference in the regulation of plasma high density lipoprotein cholesterol by genetic and environmental factors. *Hum Genet* 1996; 97: 156–162.
17. Yilmaz H, Agachan B, Karaali ZE et al. *TaqIB* polymorphism of CETP gene on lipid abnormalities in patients with type II diabetes mellitus. *Int J Mol Med* 2004; 13: 889–893.
18. Mitchell RJ, Earl L, Williams J et al. Polymorphisms of the gene coding for the cholesteryl ester transfer protein and plasma lipid levels in Italian and Greek migrants to Australia. *Hum Biol* 1994; 66: 13–25.
19. Tenkanen H, Koskinen P, Kontula K et al. Polymorphisms of the gene encoding cholesterol ester transfer protein and serum lipoprotein levels in subjects with and without coronary heart disease. *Hum Genet* 1991; 87: 574–578.
20. Kawasaki I, Tahara H, Emoto M et al. Relationship between *TaqIB* cholesteryl ester transfer protein gene polymorphism and macrovascular complications in Japanese patients with type 2 diabetes. *Diabetes* 2002; 51: 871–874.
21. van Venrooij FV, Stolk RP, Banga J-D et al. Common cholesteryl ester transfer protein gene polymorphisms and the effect of atorvastatin therapy in type 2 diabetes. *Diabetes Care* 2003; 26: 1216–1223.
22. Meguro S, Takei I, Murata M et al. Cholesteryl ester transfer protein polymorphism associated with macroangiopathy in Japanese patients with type 2 diabetes. *Atherosclerosis* 2001; 156: 151–156.
23. Freeman DJ, Samani NJ, Wilson V et al. A polymorphism of the cholesteryl ester transfer protein gene predicts cardiovascular events in non-smokers in the West of Scotland Coronary Prevention Study. *Eur Heart J* 2003; 24: 1833–1842.
24. Park K-W, Choi J-H, Kim H-K et al. The association of cholesteryl ester transfer protein polymorphism with high-density lipoprotein cholesterol and coronary artery disease in Koreans. *Clin Genet* 2003; 63: 31–38.
25. Hannuksela ML, Liinamaa MJ, Kesäniemi YA et al. Relation of polymorphisms in the cholesteryl ester transfer protein gene to transfer protein activity and plasma

- lipoprotein levels in alcohol drinkers. *Atherosclerosis* 1994; 110: 35-44.
26. Freeman DJ, Griffin BA, Holmes AP et al. Regulation of plasma HDL cholesterol and subfraction distribution by genetic and environmental factors. Associations between the *TaqI* B RFLP in the CETP gene and smoking and obesity. *Arterioscler Thromb Vasc Biol* 1994; 14: 336-344.
 27. Dullaart RPF, Groener JEM, Dikkeschei BD et al. Elevated cholesteryl ester transfer protein activity in IDDM men who smoke. Possible factor for unfavorable lipoprotein profile. *Diabetes Care* 1991; 14: 338-341.
 28. Dullaart RPF, Hoogenberg K, Dikkeschei BD et al. Higher plasma lipid transfer protein activities and unfavorable lipoprotein changes in cigarette-smoking men. *Arterioscler Thromb* 1994; 14: 1581-1585.
 29. Mero N, Van Tol A, Scheek LM et al. Decreased postprandial high density lipoprotein cholesterol and apolipoproteins A-I and E in normolipidemic smoking men: relations with lipid transfer proteins and LCAT activities. *J Lipid Res* 1998; 39: 1493-1502.
 30. Kinoshita M, Teramoto T, Shimazu N et al. CETP is a determinant of serum LDL-cholesterol but not HDL-cholesterol in healthy Japanese. *Atherosclerosis* 1996; 120: 75-82.
 31. Freeman DJ, Caslake MJ, Griffin BA et al. The effect of smoking on post-heparin lipoprotein and hepatic lipase, cholesteryl ester transfer protein and lecithin:cholesterol acyl transferase activities in human plasma. *Eur J Clin Invest* 1998; 28: 584-591.
 32. Ito T, Nishiwaki M, Ishikawa T et al. CETP and LCAT activities are unrelated to smoking and moderate alcohol consumption in healthy normolipidemic men. *Jpn Circ J* 1995; 59: 541-546.
 33. Mahley RW, Palaoglu KE, Atak Z et al. Turkish Heart Study: lipids, lipoproteins, and apolipoproteins. *J Lipid Res* 1995; 36: 839-859.
 34. Mahley RW, Arslan P, Pekcan G et al. Plasma lipids in Turkish children: impact of puberty, socioeconomic status, and nutrition on plasma cholesterol and HDL. *J Lipid Res* 2001; 42: 1996-2006.
 35. Onat A. Risk factors and cardiovascular disease in Turkey. *Atherosclerosis* 2001; 156: 1-10.
 36. Bersot TP, Vega GL, Grundy SM et al. Elevated hepatic lipase activity and low levels of high density lipoprotein in a normotriglyceridemic, nonobese Turkish population. *J Lipid Res* 1999; 40: 432-438.
 37. Friedewald WT, Levy RI, Fredrickson DS. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clin Chem* 1972; 18: 499-502.
 38. Mahley RW, P epin J, Palaoglu KE et al. Low levels of high density lipoproteins in Turks, a population with elevated hepatic lipase: high density lipoprotein characterization and gender-specific effects of apolipoprotein E genotype. *J Lipid Res* 2000; 41: 1290-1301.
 39. Koyucu B, Kara T, Camlidag O et al. Risk factors for cardiovascular diseases in Turks in Amsterdam and in Ankara. [Article in Dutch]. *Ned Tijdschr Geneesk* 1997; 141: 882-888.
 40. L uttmann S, von Eckardstein A, Wei W et al. Electrophoretic screening for genetic variation in apolipoprotein C-III. Identification of a novel apoC-III variant, apoC-III (Asp45→Asn), in a Turkish patient. *J Lipid Res* 1994; 35: 1431-1440.
 41. Kondo I, Berg K, Drayna D et al. DNA polymorphism at the locus for human cholesteryl ester transfer protein (CETP) is associated with high density lipoprotein cholesterol and apolipoprotein levels. *Clin Genet* 1989; 35: 49-56.
 42. Hsu L-A, Ko Y-L, Hsu K-H et al. Genetic variations in the cholesteryl ester transfer protein gene and high density lipoprotein cholesterol levels in Taiwanese Chinese. *Hum Genet* 2002; 110: 57-63.
 43. Kuivenhoven JA, Jukema JW, Zwinderman AH et al. The role of a common variant of the cholesteryl ester transfer protein gene in the progression of coronary atherosclerosis. *N Engl J Med* 1998; 338: 86-93.
 44. Brousseau ME, O'Connor JJ Jr, Ordovas JM et al. Cholesteryl ester transfer protein *TaqI* B2B2 genotype is associated with higher HDL cholesterol levels and lower risk of coronary heart disease end points in men with HDL deficiency. *Veterans Affairs HDL Cholesterol Intervention Trial. Arterioscler Thromb Vasc Biol* 2002; 22: 1148-1154.
 45. Satman I, Yilmaz T, Seng ul A et al. Population-based study of diabetes and risk characteristics in Turkey. Results of the Turkish Diabetes Epidemiology Study (TURDEP). *Diabetes Care* 2002; 25: 1551-1556.
 46. Durlach A, Clavel C, Girard-Globa A et al. Sex-dependent association of a genetic polymorphism of cholesteryl ester transfer protein with high-density lipoprotein cholesterol and macrovascular pathology in type II diabetic patients. *J Clin Endocrinol Metab* 1999; 84: 3656-3659.
 47. Talmud PJ, Hawe E, Robertson K et al. Genetic and environmental determinants of plasma high density lipoprotein cholesterol and apolipoprotein AI concentrations in healthy middle-aged men. *Ann Hum Genet* 2002; 66: 111-124.
 48. Corella D, S aiz C, Guill en M et al. Association of *TaqIB* polymorphism in the cholesteryl ester transfer protein gene with plasma lipid levels in a healthy Spanish population. *Atherosclerosis* 2000; 152: 367-376.
 49. de Grooth GJ, Smilde TJ, van Wissen S et al. The relationship between cholesteryl ester transfer protein levels and risk factor profile in patients with familial hypercholesterolemia. *Atherosclerosis* 2004; 173: 261-267.
 50. Craig WY, Palomaki GE, Haddow JE. Cigarette smoking and serum lipid and lipoprotein concentrations: an analysis of published data. *BMJ* 1989; 298: 784-788.
 51. Shafey O, Dolwick S, Guindon GE (eds). In: *Tobacco control country profiles - Turkey*. Atlanta, GA: American Cancer Society, Inc., World Health Organization, and International Union Against Cancer, 2003. Retrieved from http://www.cancer.org/docroot/PRO/content/PRO_1-1_Tobacco_Control_Country_Profiles.asp. Accessed on July 19, 2004, p. 402.
 52. Mann CJ, Yen FT, Grant AM et al. Mechanism of plasma cholesteryl ester transfer in hypertriglyceridemia. *J Clin Invest* 1991; 88: 2059-2066.
 53. Foger B, Ritsch A, Doblinger A et al. Relationship of plasma cholesteryl ester transfer protein to HDL cholesterol. Studies in normotriglyceridemia and moderate hypertriglyceridemia. *Arterioscler Thromb Vasc Biol* 1996; 16: 1430-1436.
 54. Arai T, Yamashita S, Hirano K-I et al. Increased plasma cholesteryl ester transfer protein in obese subjects. A possible mechanism for the reduction of serum HDL cholesterol levels in obesity. *Arterioscler Thromb* 1994; 14: 1129-1136.
 55. Rump P, Mensink RP, Hornstra G. Interaction between a common variant of the cholesteryl ester transfer protein gene and the apolipoprotein E polymorphism: effects on plasma lipids and lipoproteins in a cohort of 7-year-old children. *Nutr Metab Cardiovasc Dis* 2002; 12: 317-324.
 56. Knoblauch H, Bauerfeind A, Kr ahenb uhl C et al. Common haplotypes in five genes influence genetic variance of LDL and HDL cholesterol in the general population. *Hum Mol Genet* 2002; 11: 1477-1485.

CETP *TaqIB* genotype, smoking, and HDL-C

57. Skoglund-Andersson C, Ehrenborg E, Fisher RM et al. Influence of common variants in the CETP, LPL, HL and APO E genes on LDL heterogeneity in healthy, middle-aged men. *Atherosclerosis* 2003; 167: 311-317.
58. Dacet C, Poirier O, Cambien F et al. New functional promoter polymorphism, CETP/-629, in cholesteryl ester transfer protein (CETP) gene related to CETP mass and high density lipoprotein cholesterol levels. Role of Sp1/Sp3 in transcriptional regulation. *Arterioscler Thromb Vasc Biol* 2000; 20: 507-515.
59. Lu H, Inazu A, Moriyama Y et al. Haplotype analyses of cholesteryl ester transfer protein gene promoter: a clue to an unsolved mystery of *TaqIB* polymorphism. *J Mol Med* 2003; 81: 246-255.
60. Tai ES, Ordovas JM, Corella D et al. The *TaqIB* and -629C>A polymorphisms at the cholesteryl ester transfer protein locus: associations with lipid levels in a multiethnic population. The 1998 Singapore National Health Survey. *Clin Genet* 2003; 63: 19-30.
61. Klerkx AHM, Tanck MWT, Kastelein JJP et al. Haplotype analysis of the CETP gene: not *TaqIB*, but the closely linked -629C>A polymorphism and a novel promoter variant are independently associated with CETP concentration. *Hum Mol Genet* 2003; 12: 111-123.
62. Kakko S, Tamminen M, Päiväsalo M et al. Variation at the cholesteryl ester transfer protein gene in relation to plasma high density lipoproteins cholesterol levels and carotid intima-media thickness. *Eur J Clin Invest* 2001; 31: 593-602.
63. Barter PJ, Brewer HB Jr, Chapman MJ et al. Cholesteryl ester transfer protein. A novel target for raising HDL and inhibiting atherosclerosis. *Arterioscler Thromb Vasc Biol* 2003; 23: 160-167.
64. van der Steeg WA, Kuivenhoven JA, Klerkx AH et al. Role of CETP inhibitors in the treatment of dyslipidemia. *Curr Opin Lipidol* 2004; 15: 631-636.
65. de Grooth GJ, Kuivenhoven JA, Stalenhoef AFH et al. Efficacy and safety of a novel cholesteryl ester transfer protein inhibitor, JTT-705, in humans. A randomized phase II dose-response study. *Circulation* 2002; 105: 2159-2165.
66. Brousseau ME, Schaefer EJ, Wolfe ML et al. Effects of an inhibitor of cholesteryl ester transfer protein on HDL cholesterol. *N Engl J Med* 2004; 350: 1505-1515.
67. Clark RW, Sutfin TA, Ruggeri RB et al. Raising high-density lipoprotein in humans through inhibition of cholesteryl ester transfer protein: an initial multidose study of torcetrapib. *Arterioscler Thromb Vasc Biol* 2004; 24: 490-497.
68. Brewer HB Jr. High-density lipoproteins: a new potential therapeutic target for the prevention of cardiovascular disease. *Arterioscler Thromb Vasc Biol* 2004; 24: 387-391.
69. Manninen V, Elo MO, Frick MH et al. Lipid alterations and decline in the incidence of coronary heart disease in the Helsinki Heart Study. *JAMA* 1988; 260: 641-651.

Supplementary material

The following supplementary material is available for this article online: **Table S1.** Stratified HDL-C (mg/dl ± SD) data with common apoE and CETP *TaqIB* polymorphisms in a random Turkish population.

Chapter 6: Common polymorphisms of ATP binding cassette transporter A1, including a functional promoter polymorphism, associated with plasma high density lipoprotein cholesterol levels in Turks



Atherosclerosis 183 (2005) 199–212

ATHEROSCLEROSIS

www.elsevier.com/locate/atherosclerosis

Common polymorphisms of ATP binding cassette transporter A1, including a functional promoter polymorphism, associated with plasma high density lipoprotein cholesterol levels in Turks

Uğur Hodoğlugil^{a,d}, David W. Williamson^{a,e}, Yadong Huang^{a,b,d}, Robert W. Mahley^{a,b,c,d,*}

^a Gladstone Institute of Cardiovascular Disease, 1650 Owens Street, San Francisco, CA 94158, USA

^b Department of Pathology, University of California, San Francisco, CA, USA

^c Department of Medicine, University of California, San Francisco, CA, USA

^d Cardiovascular Research Institute, University of California, San Francisco, CA, USA

^e Graduate Program in Biological and Medical Informatics, University of California, San Francisco, CA, USA

Received 9 July 2004; received in revised form 7 February 2005; accepted 1 March 2005

Available online 2 June 2005

Abstract

The role of high levels of high density lipoprotein cholesterol (HDL-C) in protection against development of atherosclerosis is generally attributed to its role in reverse cholesterol transport, and the ATP binding cassette transporter A1 (ABCA1) is a key element of this process. We examined polymorphisms in ABCA1 in Turks, a population characterized by very low HDL-C levels. We discovered 36 variations in ABCA1 and genotyped informative polymorphisms in over 2300 subjects. The rare alleles of C-14T and V771M polymorphisms were associated with higher HDL-C levels in men and, in combination with the rare alleles of R219K and I883M, respectively, with higher HDL-C in both sexes. Rare alleles of the C-14T and V771M polymorphisms were more frequent in the high HDL-C (≥ 40 mg/dl) than in the low HDL-C group (≤ 30 mg/dl) in men ($P < 0.05$). Moreover, the T allele of C-14T had more *in vitro* transcriptional activity than the C allele (20–88%), depending on the cell line ($P < 0.05$), suggesting its functionality. Haplotype construction and haplotype association with phenotype were performed in the promoter and coding region of ABCA1 separately. Analysis of the promoter haplotype block supported the association with the C-14T polymorphism. The C-14T and R219K polymorphisms were on different haplotype blocks. Analysis of the coding region structure revealed that the rare M allele of V771M was distributed predominantly among three common haplotypes, but the sum of their frequencies comprise only two-thirds of the frequency of the M allele. The rare alleles of the V771M and the I883M polymorphisms do not exist together on any of the common haplotypes. In conclusion, we describe a functional promoter polymorphism (C-14T) and a coding sequence variant (V771M) of ABCA1 and their interactions with two other variants (R219K and I883M) on plasma HDL-C levels in Turks.
© 2005 Elsevier Ireland Ltd. All rights reserved.

Keywords: ATP binding cassette transporter A1; HDL cholesterol; Triglyceride; Turkish population; Polymorphism

1. Introduction

Atherosclerotic cardiovascular disease is a leading cause of death worldwide [1], and a low level of high density lipoprotein cholesterol (HDL-C) is a major independent risk factor for atherosclerosis [2,3]. The protective role of HDL-C is generally attributed to its participation in reverse cholesterol transport, a process in which excess cholesterol is

transported from peripheral cells to HDL particles for delivery to the liver and excretion. The ATP binding cassette transporter A1 (ABCA1) participates in apolipoprotein-mediated efflux of cholesterol and phospholipid from peripheral cells, especially macrophages, that is crucial for the initial step of reverse cholesterol transport. The identification of mutations in the ABCA1 gene in patients with Tangier disease, who have very low HDL-C, elevated triglyceride levels, and increased risk of premature coronary atherosclerosis, suggested a major role for ABCA1 in regulating plasma HDL-C levels [4–8]. Some common polymorphisms of ABCA1, including R219K

* Corresponding author. Tel.: +1 415 734 2000; fax: +1 415 355 0820.
E-mail address: rmahley@gladstone.ucsf.edu (R.W. Mahley).

[9] and I883M [9–11], are associated with elevated HDL-C levels, although not in all studies [10,12–17]. Interestingly, common polymorphisms of ABCA1 may significantly alter the severity of atherosclerosis, apparently without influencing plasma lipid levels [12,13,15,18,19]. Recently, it was shown that some polymorphisms of ABCA1 were associated with increases (V771M and V825I) or decreases (R1587K) in HDL-C in women and with some consistent trends in men in a large random Danish population [17].

Haplotype analysis of the ABCA1 gene with respect to plasma HDL-C levels [20,21] and plasma apolipoprotein (apo) A1 levels and myocardial infarction [19] have been reported. Haplotype analysis revealed that ABCA1 accounted for about 10% of HDL-C variation [21], but no haplotype effect on apoA1 variability or on the risk of myocardial infarction was detected [21]. Certain haplotypes were more frequent among coronary artery disease (CAD) patients than controls in the Malay population, but not in the Chinese and Indian populations [20].

Although cardiovascular risk factor profiles and the frequency of coronary events differ by gender, the mechanisms for the differences remain to be resolved. Gender differences were observed in other lipid-related association studies [22–27], suggesting that gender-related mechanisms or factors might interact differently with the variants of a particular gene. This phenomenon was also observed with ABCA1. In one study, where males and females were analyzed separately, R219K and I883M were associated with elevated HDL-C levels in females only [9].

We hypothesized that polymorphisms in ABCA1 are important for determining plasma lipid levels and that gender may modulate the role of these polymorphisms. To test this hypothesis, we studied male and female subjects from the Turkish Heart Study [28], a large, random epidemiological survey of the Turkish population. The main characteristic of this population is a very low level of plasma HDL-C, making this an ideal population in which to study genes that influence HDL-C levels [28–31]. We screened the promoter region and the exons and exon/intron splice junctions of ABCA1 with denaturing high-performance liquid chromatography (dHPLC) to detect polymorphisms, and subsequently analyzed their associations with plasma lipid levels.

2. Methods

2.1. Study population

The study population ($n = 2700$) was randomly selected from the Turkish Heart Study database of more than 9000 volunteers from six regions of Turkey [28]. Detailed biobata were obtained from each participant. The protocol was approved by the Committee on Human Research of the University of California, San Francisco, and was in accordance with the Helsinki Declaration. Subjects who were taking any lipid-lowering medication or had a history of diabetes mellitus were excluded.

2.2. Biochemical analyses

Blood samples were obtained after an overnight fast. Total cholesterol and triglyceride levels were determined by enzymatic colorimetric methods, and the HDL-C levels were determined with the CHOD-PAP method with precipitation of very low density lipoproteins and low density lipoproteins (LDL) [28]. LDL cholesterol (LDL-C) was calculated by the Friedewald formula [32] for participants with triglyceride levels <400 mg/dl. Plasma total apoA1 levels were measured with Hydragel ApoA1 kits (Sebia, Norcross, GA, USA) in a subset of the study population [33].

2.3. Detection of polymorphisms by dHPLC

DNA was screened to identify variations in the ABCA1 gene among subjects whose HDL-C levels were in the lowest and highest fifth percentiles. These DNAs were randomly plated and screened ($n = 95$ –240). Primers were designed to amplify the ABCA1 promoter, the 5' untranslated region, and all exons, including intron/exon splicing boundaries if possible. The amplified DNA was denatured and slowly reannealed to form homo- and heteroduplex DNA. Subjects who were heterozygous in any region on the amplified product formed heteroduplex DNA. The amplified DNA (10–15 μ l) was loaded onto the dHPLC apparatus (WAVE DNA fragment analysis system, Transgenomic, Omaha, NE) and run under conditions determined by the WAVE software for dHPLC for the given DNA sequence. Representative genomic DNA samples that displayed heterozygous profiles were sequenced to confirm the mutation or polymorphism. DNA sequences were aligned and analyzed with Sequencher DNA analysis software (Gene Codes, Ann Arbor, MI, USA). Because not every heterozygous profile was sequenced, it is possible that some single nucleotide polymorphisms were not discovered using this method. No other method was used to detect polymorphisms in this study.

2.4. Genotyping

After polymerase chain reaction amplification, each polymorphism was genotyped by restriction fragment length polymorphism or allele-specific oligonucleotide hybridization [34]. The conditions of all assays are described in Supplemental Table I. The accuracy of the genotyping was evaluated by randomly inserting duplicate DNA samples in the assays ($\sim 6\%$ replication). Genotyping discrepancies were found in less than 1% of the samples and were resolved by rescoring or eliminating the data.

2.5. Cloning the ABCA1 promoter into a reporter vector

Although multiple transcriptional start sites have been suggested for ABCA1, the base numbering used in this study is relative to the transcriptional start in the published sequence by Santamarina-Fojo et al. (AF275948) [35].

Genomic DNAs from homozygotes (CC or TT) for the polymorphism at position –14 were amplified (forward primer: 5'-CCATTACCCAGAGGACTGTC-3'; reverse primers: ACTGGCTAGCGTTTTGCGGGACTAGTTC-3' for CC subjects or 5'-ACTGGCTAGCGTTT-TTGCCGAGACTAGTTC-3' for TT subjects) and double digested with *SacI* and *NheI* (–473 and –2). The resulting 471–base pair DNA fragment was ligated to a pGL3-Basic vector (Promega, Madison, WI, USA) predigested with both *SacI* and *NheI*. The ABCA1 promoter and 5' untranslated region were inserted immediately upstream of the transcriptional start site of the reporter vector, exchanging the ABCA1 start site for the luciferase start site. Positive clones were selected, and the integrity of inserts and vector sequences surrounding the ligation sites were confirmed by DNA sequencing. The only difference between the two constructs was a T or C at position –14.

2.6. Cell culture and transfection

The ABCA1 promoter/luciferase construct was successfully transfected into three cell lines: human hepatoma (HepG2), green monkey kidney (COS-7), and Chinese hamster ovary (CHO). CHO is a commonly used cell line in transfection studies because of its high efficiency. The liver and kidney cell lines have high ABCA1 mRNA expression in mice [36], suggesting these lines are suitable to test the promoter activity of ABCA1. The cells were plated in 24-well plates for 24 h until they reached 70–80% confluence and were then transfected with 500 ng of ABCA1 promoter/luciferase plasmid, 10 ng of control *Renilla* luciferase plasmid, and Lipofectamine Plus (Invitrogen, Carlsbad, CA, USA). The cells were harvested 24 h later and assayed with the dual-luciferase reporter assay system (Promega, Madison, WI, USA) in a luminometer (1450 Microbeta, Perkin-Elmer, Boston, MA, USA). The protein concentration in each well was determined with Micro BCA reagents (Pierce, Rockford, IL, USA).

2.7. Statistics and data analysis

Data were analyzed with SPSS 10.0, Microsoft Access, and Excel. Associations between genotypes, lipids, and other parameters were analyzed separately for males and females. Lipid levels are expressed in milligrams per deciliter, and all values are reported as mean \pm S.D. Since triglyceride levels were not normally distributed, log-transformed values were used for statistical comparison, and untransformed mean values are reported in the text. Mean values were compared with the *t* test according to genotype, and two-tailed $P < 0.05$ was considered significant. Analysis of covariance was used to construct the model to explain the variation in HDL-C levels. Plasma triglyceride levels, body mass index (BMI), smoking, and alcohol consumption were included as covariates, and genotype score was included as a fixed factor in the model (GLM Univariate, SPSS 10.0). The –14C ver-

sus –14T ABCA1 promoter activity in a luciferase reporter assay was compared with the Mann–Whitney *U* test. Chi-square (χ^2) analysis was used to test differences between the observed and expected frequencies of alleles (assuming a Hardy–Weinberg equilibrium), to test differences in percentages between males and females, and to compare genotype, allele, or haplotype frequencies after stratification by HDL-C levels.

The expectation-maximization algorithm was used to estimate the maximum-likelihood haplotype frequencies from multilocus genotypic data without known gametic phase (Arlequin software, Version 2.00) [37]. All subjects with one or more missing genotypes were excluded for haplotype construction. The haplotypes that were assigned unambiguously to subjects were further analyzed. The linkage disequilibrium (LD) between polymorphisms was similarly calculated with the same software [37] and expressed in terms of $D' = D/D_{\max}$ or D/D_{\min} [38].

3. Results

3.1. Population characteristics

The demographic and biochemical characteristics and the apoE genotypes of 2700 randomly selected Turkish Heart Study participants are presented in Table 1. Both men and women had very low plasma HDL-C levels and high total cholesterol/HDL-C ratios, as reported [28]. Even though plasma total cholesterol and LDL-C levels were not exces-

Table 1
Demographic, biochemical, and apoE genotypic characteristics of Turkish Heart Study participants according to sex ($n = 2700$)^a

	Females ($n = 1149$)	Males ($n = 1551$)	<i>P</i>
Age (years)	41 \pm 14	41 \pm 12	NS
BMI (kg/m ²)	26.1 \pm 5.2	25.7 \pm 3.8	<0.05
HDL-C (mg/dl)	41 \pm 9	35 \pm 7	<0.001
Total cholesterol (mg/dl)	182 \pm 46	187 \pm 44	<0.05
LDL-C (mg/dl)	119 \pm 40	123 \pm 39	<0.05
Triglycerides (mg/dl)	111 \pm 73	145 \pm 92	<0.001
Total cholesterol/HDL-C ratio	4.6 \pm 1.5	5.5 \pm 1.7	<0.001
Systolic blood pressure (mm Hg)	126 \pm 24	124 \pm 20	NS
Diastolic blood pressure (mm Hg)	81 \pm 15	81 \pm 13	NS
ApoE alleles			
ε2 (%)	8.5	8.0	
ε3 (%)	84.0	84.1	
ε4 (%)	7.5	7.9	
Alcohol consumption (%) ^b	6.1	27.7	<0.001
Cigarette smoking (%) ^c	26.5	57.4	<0.001

^a Values are mean \pm S.D. or percentages. Means were compared by *t* test, and percentages were analyzed by χ^2 test. NS, not significant.

^b One or more drinks per week.

^c One or more cigarettes per day.

sively elevated, low HDL-C with or without mildly elevated triglyceride levels could represent a significant risk factor for heart disease in the Turkish population [28,30]. Detailed analyses of the Turkish Heart Study data have been reported [28,29,31,33].

3.2. Identification of ABCA1 polymorphisms

In a survey of DNA samples from Turks with HDL-C levels in the lowest and highest fifth percentiles, 36 poly-

morphisms in the ABCA1 gene were identified by dHPLC and verified by sequencing (Table 2). Six polymorphisms were in the 5' untranslated and promoter regions. In the initial screening, it was found that the C allele of G-407C polymorphism was strong in LD with the C allele of T-564C and was not associated with HDL-C ($n=443$, Table 3) and therefore not further genotyped. Four polymorphisms (T-564C, G-99C, C-14T, and InsG 319) were genotyped in a larger number of subjects ($n=1996-2332$). Only the C-14T polymorphism was highly informative. The G-803A

Table 2

ABCA1 polymorphisms

Nucleotide change ^a	Carrier of rare allele (%)	Rare allele (%)	n^b	References ^c		
Frequency in 5' and promoter regions						
G-803A	~10		92/141	[19]		
T-564C	72.7	47.7	728/1268	[18]		
G-407C	62.5	40.7	220/233	[18]		
G-99C	42.2	24.2	875/1130	[46]		
C-14T	61.2	37.7	916/1416	[46]		
InsG 319	26.6	14.2	848/1288	[46]		
Nucleotide change ^d	Amino acid change	Exon	Carrier of rare allele (%)	Rare allele (%)	n^b	References ^e
Frequency in coding sequence						
Nonsynonymous						
G(70943)A	R219K	7	62.3	38.5	996/1466	db_2230806
G(102555)A	V771M	16	10.0	5.1	981/1477	db_2066718
G(103777)A	V825I	17	12.3	6.2	960/1145	db_4149312
A(105057)G	I883M	18	38.1	21.8	1084/1448	db_4149313
G(112177)C	E1172D	24	9.0	4.6	1001/1237	[12,13]
A(116887)G	Q1328R	28	0.003 ^e	0.0015	172/156	NR
G(129004)A	R1587K	35	55.1	33.0	937/1351	db_2230808
T(133402)C	Y1767H	39	~1.0 ^e		40/55	NR
G(133420)A	V1773M	39	~1.0 ^e		40/55	NR
Synonymous						
C(100538)A	I620I	15	~4.0		40/55	NR
C(109469)T	V990V	21	~3.0		87/90	[17]
T(109861)G	V1053V	22	~1.0		90/90	[12]
C(109868)T	L1056L	22	~5.0		90/90	NR
C(109906)T	R1068R	22	~3.0		90/90	NR
A(113280)G	E1211E	25	~4.0		40/55	[17]
A(116879)G	T1325T	28	~1.0		40/55	NR
T(137043)C	Y1921Y	43	~1.0		40/55	NR
Nucleotide change ^d	Intron	Carrier of rare allele (%)	Intronic location	n^b	References ^e	
Frequency in noncoding sequence						
G(23816)A	1	~1.0	11 bp 5' exon 1b	94	NR	
G(23819)C	1	42	8 bp 5' exon 1b	94	NR	
A(22997)T	1	46	90 bp 5' exon 1d	91	NR	
A(23004)G	1	~1.0	83 bp 5' exon 1d	91	NR	
G(23058)C	1	46	29 bp 5' exon 1d	91	NR	
G(40504)A	3	2.6	26 bp 3' of exon 3	192	NR	
C(45217)T	4	0.7	64 bp 3' of exon 4	142	NR	
T(98628)A	14	35-40	24 bp 3' of exon 14	190	db_4743763	
C(100332)T	14	4.3	59 bp 5' of exon 15	90	db_2066717	
C(108020)T	19	0.7	3 bp 5' of exon 20	144	NR	
DelTTT(134503-6)	39	7.0	20-23 bp 5' of exon 40	90	NR	
C(142026)T	46	8.6	34 bp 5' of exon 47	116	NR	
A(142751)G	48	15.9	13 bp 3' of exon 48	107	NR	

^a Relative to transcriptional start.

^b Female/male.

^c Reference or single-nucleotide polymorphism database number; NR, not previously reported.

^d AF275948 (accession number of reference ABCA1 sequence).

^e Observed in very low frequency and no further analyses performed.

Table 3
ABCA1 polymorphisms and mean plasma HDL-C levels (mg/dl \pm S.D.) in a random Turkish population

	Females			Males		
	AA ^a	AB	BB	AA ^a	AB	BB
Promoter and 5' region						
T-564C	40.4 \pm 8.4 (205)	41.0 \pm 7.9 (365)	39.9 \pm 7.6 (158)	35.6 \pm 7.4 (339)	35.5 \pm 6.5 (635)	34.9 \pm 6.5 (294)
G-407C	41.3 \pm 11.2 (82)	40.7 \pm 7.7 (101)	40.7 \pm 12.2 (37)	35.7 \pm 6.9 (88)	35.3 \pm 6.7 (96)	35.4 \pm 7.4 (49)
G-99C	40.7 \pm 7.5 (505)	41.0 \pm 7.2 (317)	41.3 \pm 8.4 (53)	35.1 \pm 6.5 (654)	35.0 \pm 6.3 (405)	34.9 \pm 6.5 (71)
C-14T	41.3 \pm 9.0 (361)	41.0 \pm 9.0 (417)	41.0 \pm 9.4 (138)	34.8 \pm 7.0^b (547)	35.5 \pm 7.5 (675)	36.7 \pm 8.1^b (194)
InsG 319	41.3 \pm 8.1 (641)	40.5 \pm 7.7 (193)	43.1 \pm 8.0 (14)	35.3 \pm 6.4 (933)	34.7 \pm 6.4 (332)	34.5 \pm 6.5 (23)
Nonsynonymous						
R219K	41.2 \pm 9.4 (364)	40.8 \pm 8.6 (480)	41.2 \pm 10 (152)	35.2 \pm 7.3 (574)	35.3 \pm 7.3 (688)	35.2 \pm 7.6 (204)
V771M	40.9 \pm 9.2 (896)	42.8 \pm 9.3 (82)	38.7 \pm 10 (3)	35.1 \pm 7.2^c (1330)	37.1 \pm 8.0^b (144)	45.7 \pm 10.7 (3)
V825I	40.6 \pm 9.4 (842)	38.9 \pm 8.5 (117)	42 (1)	35.8 \pm 8.7 (1005)	37.1 \pm 9.5 (140)	(–)
I883M	41.1 \pm 9.5 (643)	40.8 \pm 8.4 (372)	41.4 \pm 9.1 (69)	35.0 \pm 7.0 (922)	35.7 \pm 8.0 (457)	35.6 \pm 7.4 (69)
E1172D	40.5 \pm 8.8 (907)	40.5 \pm 7.6 (93)	29 (1)	34.6 \pm 7.3 (1129)	35.4 \pm 7.6 (105)	30.7 \pm 8.7 (3)
R1587K	41.1 \pm 9.4 (410)	40.8 \pm 9.6 (433)	41.0 \pm 10.1 (94)	35.9 \pm 8.1 (617)	35.1 \pm 7.6 (579)	35.3 \pm 8.6 (155)

Numbers of subjects are shown in parenthesis. Significance was determined by *t* test.

^a A, common allele; B, rare allele.

^b AA vs. BB, *P* < 0.02.

^c AA vs. AB, *P* < 0.01.

polymorphism was evaluated by dHPLC only and showed no differences in allele frequencies between the lowest and highest HDL-C groups. It was not evaluated further.

Seventeen of the ABCA1 polymorphisms were in the coding sequence (Table 2). Nine were nonsynonymous; the remaining eight were synonymous, did not change the protein sequence, and were not analyzed further. Informative associations between HDL-C levels and the R219K, V771M, and I883M polymorphisms will be discussed.

Recently three new exons (1a, 1b, and 1c) and three LXR response elements in intron 1 of ABCA1 were described [39]. When we screened 186 subjects, no polymorphisms were found in exons 1a, 1b, or 1c, or in the LXR response elements.

For amplification of genomic DNA, amplicons were constructed with intron/exon splice boundaries when possible. Thirteen intronic polymorphisms were found (Table 2). Except for an intron 19 polymorphism, which was three base pairs 5' of exon 20, the other intronic polymorphisms were 8–90 base pairs away from the splice boundaries. Because of the rarity of the intron 19 variation (1 of 144 subjects) and the distance of other polymorphisms from splice boundaries, no further analyses were performed.

For all the polymorphisms described, no significant differences in the frequencies of rare alleles were observed between males and females. The distribution of alleles was consistent with Hardy–Weinberg equilibrium for all the polymorphisms genotyped.

3.3. Association of the ABCA1 C-14T polymorphism with plasma lipid levels in a random Turkish population

The association of the C-14T polymorphism with different phenotypic variables was analyzed in a random group of 1416 males and 916 females. The -14T allele, which occurs

with a frequency of ~37.7%, was associated with significantly higher HDL-C (CC versus TT, *P* < 0.02, Table 3) and lower triglyceride levels in males only (CC, 149 \pm 89 versus TT, 128 \pm 84, *P* < 0.02). Because plasma triglyceride levels are inversely associated with plasma HDL-C levels [40–42], changes in triglyceride levels may act as a confounding factor for the association of the C-14T polymorphism with HDL-C. In a covariate analysis of the effects of triglyceride, BMI, smoking, and alcohol consumption on HDL-C levels, the -14T association with elevated HDL-C remained significant (*P* < 0.05); using HDL-C (and BMI, smoking, or alcohol consumption) as a covariate for triglyceride levels resulted in a loss of significance. Thus, the C-14T polymorphism appears to be associated primarily with plasma HDL-C levels in Turkish males.

To further confirm the association of HDL-C with the C-14T polymorphism in males, data were stratified by HDL-C level (≤ 30 , 30–40, ≥ 40 mg/dl) and analyzed for genotype distribution and allele frequency. The rare TT genotype or T allele was more frequent in subjects with high HDL-C, and the CC genotype or C allele was more frequent in those with low HDL-C (Fig. 1, *P* < 0.05). To perform an internal intrapopulation control to confirm the validity of the C-14T associations in males, the entire data set was randomly divided into two groups, and the associations were re-examined (Supplemental Table II). The association of the -14TT genotype and -14T allele frequency with elevated HDL-C levels was confirmed in both groups, demonstrating the reproducibility of the results in a large random population of Turks.

In a subset of the random Turkish population [33], plasma apoA1 mean levels were slightly, but not significantly, higher in subjects with the -14TT genotype than in those with the -14CC genotype [112.1 \pm 23.8 mg/dl (*n* = 15) versus 102.9 \pm 20.9 mg/dl (*n* = 47)].

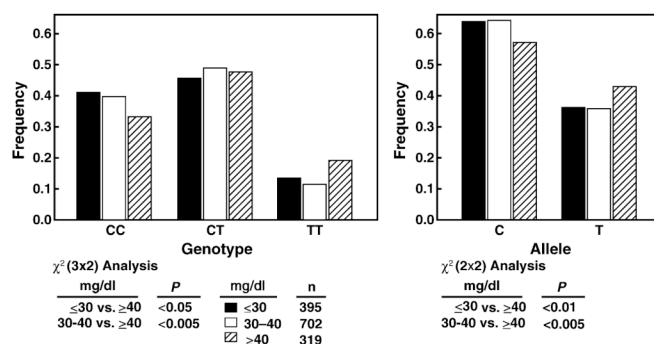


Fig. 1. Genotypic (left) and allelic (right) frequencies of the ABCA1 C-14T polymorphism in males with low (≤ 30 mg/dl), moderate (30–40 mg/dl), or high (≥ 40 mg/dl) levels of HDL-C. The frequencies were evaluated by χ^2 analysis.

Stratification by age (tertiles), BMI (tertiles), smoking (nonsmokers, 1–19 cigarettes/day, 20+ cigarettes/day), alcohol consumption (non-drinkers, 1–5 drinks/week, >5 drinks/week), and apoE polymorphism did not alter the association of C-14T with plasma HDL-C (data not shown).

3.4. Biological activity of ABCA1 C-14T polymorphism

To determine if the C-14T polymorphism influenced ABCA1 expression, we cloned the ABCA1 promoter polymorphisms into a luciferase reporter vector (pGL3-Basic). Expression of the -14C and -14T ABCA1-luciferase constructs was examined in COS-7, CHO, and HepG2 cells cotransfected with a *Renilla* luciferase plasmid. The pGL3-Basic vector without an insert exhibited very little activity. The vector/insert constructs showed several-fold increases in luciferase activity compared with pGL3-Basic vector (100–150-fold in COS-7, 400–700-fold in CHO, and 90–120-fold in HepG2 cell lines). Data were normalized to both *Renilla* luciferase and protein concentration. In each cell line, the -14T construct expressed 20–88% more luciferase than the -14C construct ($P < 0.05$, Fig. 2), suggesting that the rare T allele enhances transcriptional activity more than the C allele.

3.5. Nonsynonymous polymorphisms in the coding sequence of ABCA1

Six polymorphisms in the coding sequence of ABCA1 that occurred at a significant allelic frequency ($>1\%$) were examined for their effects on HDL-C (Table 3). Of these, only the V771M polymorphism appeared to be associated with altered HDL-C levels in males (Table 3). Since the V771M polymorphism was present in the Turkish population with an allelic frequency of 5.1% ($n = 2458$), the rare 771MM homozygous genotype (BB) occurred infrequently (three males and three females). However, males with the VM genotype (AB) had significantly higher HDL-C ($P < 0.01$, Table 3) and

significantly lower triglyceride levels (129 ± 76 mg/dl versus 147 ± 94 mg/dl, $P < 0.02$) than those with the VV genotype (AA). Analysis of covariance revealed that the V771M polymorphism was primarily associated with plasma HDL-C, not with triglycerides, in Turkish males ($P < 0.05$). After stratification by HDL-C levels, the VM + MM genotype and the M allele were more frequent in males with high HDL-C, whereas the VV genotype and V allele were more frequent in those with low HDL-C (Fig. 3). The V771M polymorphism had no significant effect on HDL-C levels in females (Table 3). Further stratification of the V771M polymorphism by age, BMI, smoking, alcohol consumption, and apoE genotype yielded no additional information. The association of the V771M polymorphism with HDL-C levels was confirmed in the randomly divided subpopulations of Turkish males (Supplemental Table II).

Plasma apoA1 levels in VM + MM males (113.3 ± 19.8 mg/dl, $n = 14$) were higher than in VV subjects

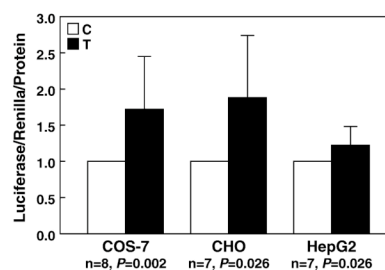


Fig. 2. In vitro, the ABCA1 promoter with -14T had significantly higher activity than the promoter with -14C. The ABCA1 promoters from -473 to -2 (relative to transcriptional start) were cloned into a luciferase reporter vector. Both constructs were identical except for a single base change, a C or T, at position -14. HepG2, COS-7, and CHO cells were transfected with either the C or T allele containing a promoter vector and a control *Renilla* luciferase plasmid.

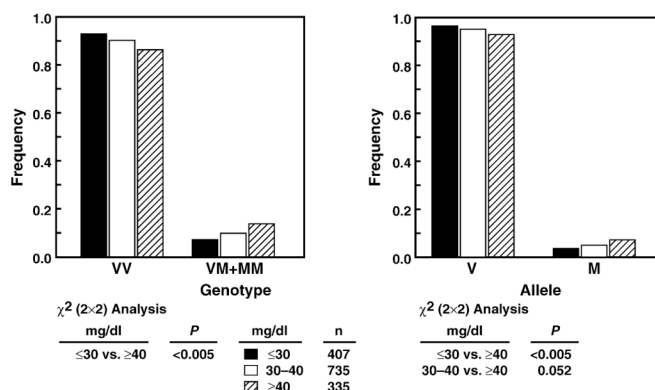


Fig. 3. Genotypic (left) and allelic (right) frequencies of the ABCA1 V771M polymorphism in males with low (≤ 30 mg/dl), moderate (30–40 mg/dl), or high (≥ 40 mg/dl) levels of HDL-C. Samples from three MM subjects were combined with VM subjects for analysis. The frequencies were evaluated by χ^2 analysis.

(104.6 ± 19.9 mg/dl, $n = 113$), but the difference did not reach statistical significance.

3.6. Effect of combined ABCA1 polymorphisms on HDL-C levels in a random Turkish population

The R219K and I883M polymorphisms have been shown to affect lipid metabolism or CAD risk [9,10,12,15,43]. Both were common in the random Turkish population (rare alleles, 38.5 and 21.8%, respectively; Table 2), but neither was separately associated with altered HDL-C levels (Table 3). In contrast, the combinations of C-14T with R219K and of V771M with I883M were associated with altered HDL-C levels.

As shown in Table 4, subjects were stratified according to their R219K genotype (RR, RK, or KK) versus C-14T (CC, CT, or TT). Although the -14TT genotype was by itself associated with HDL-C in men only, the 219KK and -14TT double homozygous genotype combination was significantly associated with elevated HDL-C levels in both genders.

When the data for Turkish females with the 219KK genotype were stratified by HDL-C level and C-14T polymorphism, the -14T allele was significantly less frequent in the low HDL-C (≤ 35 mg/dl) than in the high HDL-C group (≥ 45 mg/dl) (27.1% versus 43.8%, $P < 0.025$; Fig. 4). The 219KK and the -14TT genotypes also tended to be less frequent in females in the low HDL-C group ($P = 0.054$, Fig. 4). Similarly, among males with the 219KK genotype, the -14T allele was less frequent in the low HDL-C (≤ 30 mg/dl) than in the high HDL-C group (≥ 40 mg/dl) (31.3% versus 48.5%, $P < 0.04$). The 219KK and -14TT genotypes also tended to be less frequent in males with low HDL-C ($P = 0.069$).

To assess the combined effect of the V771M and I883M polymorphisms on HDL-C levels, subjects were stratified by the I883M genotype (II, IM, or MM) versus V771M. Among

Turkish females with 883IM, those with 771VM had significantly higher levels of HDL-C than subjects with 771VV (Table 4, $P < 0.05$). Likewise, males with either 883II or 883IM plus 771VM had significantly higher levels of HDL-C than subjects with 771VV (Table 4, $P < 0.05$).

3.7. Separate haplotype blocks of ABCA1

In order to assess whether the length of the ABCA1 locus could be treated as a single haplotype block, haplotypes were constructed and significant LDs between polymorphisms were calculated using 10 common ABCA1 polymorphisms

Table 4
Interactive effects of ABCA1 polymorphisms on mean plasma HDL-C levels (mg/dl ± S.D.)

R219K	C-14T	Females	Males
RR	CC	41.4 ± 8.5 (128)	34.7 ± 6.7 (207)
	CT	41.6 ± 9.8 (150)	35.7 ± 7.6 (253)
	TT	41.3 ± 8.9 (55)	36.3 ± 7.8 (84)
RK	CC	41.8 ± 8.3 (169)	34.8 ± 7.4 (261)
	CT	39.9 ± 8.3 (203)	35.5 ± 6.8 (309)
	TT	39.5 ± 9.0 (60)	36.8 ± 8.5 (84)
KK	CC	39.8 ± 11.2 (60)	34.0 ± 6.2 (69)
	CT	41.1 ± 9.0 (60)	35.7 ± 8.4 (102)
	TT	44.7 ± 9.8^a (22)	37.1 ± 7.8^a (22)
I883M	V771M		
II	VV	40.9 ± 9.8 (499)	34.8 ± 6.8 (797)
	VM	41.4 ± 8.3 (74)	36.7 ± 8.3^b (112)
IM	VV	40.3 ± 8.0 (313)	35.1 ± 8.1 (427)
	VM	44.2 ± 11.5^b (17)	37.3 ± 7.4^b (26)
MM	VV	41.6 ± 9.2 (60)	35.6 ± 7.5 (67)

Numbers of subjects are shown in parenthesis.

^a CC vs. TT, $P < 0.05$ by t test.
^b VV vs. VM, $P < 0.05$ by t test.

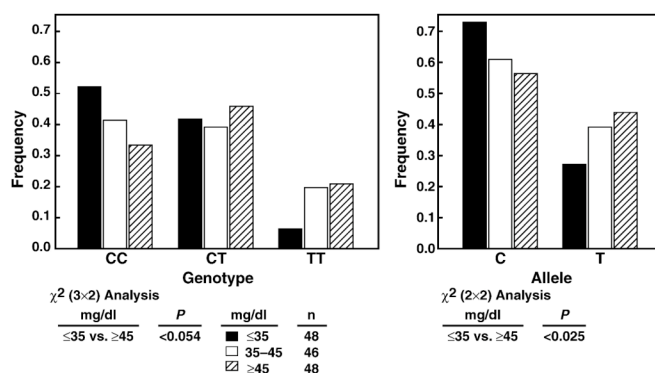


Fig. 4. Genotypic (left) and allelic (right) frequencies of the ABCA1 C–14T polymorphism in females with the 219KK genotype and low (≤ 35 mg/dl), moderate (35–45 mg/dl), or high (≥ 45 mg/dl) HDL-C. The frequencies were evaluated by χ^2 analysis.

that were genotyped in about 2000 subjects (Supplemental Table III and Table 5). Twenty-six haplotypes with frequencies of 1.0–13.7% accounted for 78.4% of all haplotypes. The remaining 98 haplotypes (sum = 21.6%) had frequencies $< 1\%$. Many of the rare single nucleotide polymorphisms, such as V771M, were not well represented on the 26 common haplotypes and therefore existed only on the remaining rare haplotypes.

Unfortunately, the attempt to map haplotype combinations to subjects failed because after constructing all pairwise combinations of common haplotypes, the probability distribution for the 351 ($N = 26$, $D = N \times (N + 1) / 2 = 351$) predicted diplotypes prevented the accurate attribution of specific pairs of haplotypes, over other pairs, to a single subject when more than one pair of haplotypes corresponded to a subject's genotype. In effect, over three quarters of the sampling data were not informative with this method (e.g., haplotypes 7 + 25 will map to the same genotype as haplotypes 5 + 4, Supplemental Table III). Furthermore, LD was found to be relatively weak across the entire ABCA1 locus (Table 5) and this is likely the consequence of the large size of the ABCA1 gene sequence (almost 150 kbp). Consequently, a haplotype-to-phenotype association could not be conducted when treating the ABCA1 gene as a single haplotype block. However, it was possible

to separate the promoter and coding regions of the ABCA1 gene into independent haplotype blocks and associate them with phenotype [17,19,21].

3.8. Haplotype structure and haplotype–phenotype association of polymorphisms in the promoter region of ABCA1

Four polymorphisms (T–564C, G–99C, C–14T, and InsG 319) were used to construct the haplotype structure of the promoter region (Table 6). Eight haplotypes with a frequency $> 1\%$ accounted for 98.9% of all haplotypes. Eighty-two percent of the diplotypes mapped to unique genotypes, defined by a single pair of haplotypes. The remaining diplotypes that could be constructed with more than one pair of haplotypes were not included in further analyses. Mean HDL-C haplotype values in men and women are presented in Table 7. In males, the mean HDL-C level for haplotype 1, having exclusively the rare allele at C–14T, was slightly higher than that for haplotype 4, a haplotype with all common alleles; however, this did not reach statistical significance. For all haplotypes, analysis of covariance (triglyceride, BMI, smoking, and alcohol consumption) revealed no significant haplotype effect on HDL-C levels ($P > 0.05$). However, further analysis

Table 5
Allele frequency and significant ($P < 0.01$) pair-wise linkage disequilibrium coefficients between the ABCA1 polymorphisms in a random Turkish population

Position	Allele %	T–564C	G–99C	C–14T	InsG 319	R219K	V771M	V825I	I883M	E1172D	R1587K
T–564C	52.3/47.7	–									
G–99C	75.8/24.2	0.36	–								
C–14T	62.3/37.7	–0.96	–0.98	–							
InsG 319	85.8/14.2	–	–	–	–						
R219K	61.5/38.5	–	–	–	–	–					
V771M	94.9/5.1	–	–	–	0.99	–	–				
V825I	93.8/6.2	–0.40	–0.69	–	–	–0.76	–	–			
I883M	78.2/21.8	–	–0.42	–	–	0.20	–0.79	0.91	–		
E1172D	95.4/4.6	–	–	–	–	–	0.34	–	–	–	
R1587K	67.0/33.0	–	–	–0.23	–	–	0.56	–	–	0.87	–

Table 6
Common haplotypes of promoter and coding regions of ABCA1 gene in a random Turkish population

Promoter region haplotypes	%	T-564C	G-99C	C-14T	InsG 319
1	31.7	0	0	1	0
2	20.1	1	1	0	0
3	19.9	1	0	0	0
4	12.7	0	0	0	0
5	4.6	0	0	1	1
6	4.2	1	1	0	1
7	3.2	1	0	0	1
8	2.5	0	0	0	1
Sum	98.9				

Coding region haplotypes	%	R219K	V771M	V825I	I883M	E1172D	R1587K
1	39.3	0	0	0	0	0	0
2	12.6	1	0	0	0	0	0
3	12.0	0	0	0	0	0	1
4	8.7	1	0	0	1	0	0
5	8.3	1	0	0	0	0	1
6	4.0	0	0	1	1	0	0
7	3.4	0	0	0	0	1	1
8	1.6	1	0	0	1	0	1
9	1.3	1	1	0	0	0	0
10	1.3	0	1	0	0	1	1
11	1.3	0	0	1	1	0	1
12	1.1	1	1	0	0	0	1
Sum	95.1						

0: common allele; 1: rare allele.

of subjects with both of their haplotypes the same (homozygosity) revealed that males with haplotype 1 homozygosity had significantly higher HDL-C levels than those with haplotype 4 homozygosity (36.4 ± 6.8 [$n = 176$] versus 33.8 ± 4.3 [$n = 26$], $P < 0.02$). This result confirmed and extended the re-

sult from single-locus analysis for the C-14T polymorphism on HDL-C levels.

After stratification of the data (Table 7) by HDL-C level (≤ 30 , $30-40$, ≥ 40 mg/dl for males and ≤ 35 , $35-45$, ≥ 45 mg/dl for females) and analysis for haplotype frequen-

Table 7
Mean plasma HDL-C levels associated with common haplotypes of ABCA1 in the 5' and promoter regions and their frequencies in a random Turkish population

Haplotype	Mean \pm S.D. (n)	All groups (%)	HDL-C subgroups (%) (n)			P (≤ 30 mg/dl vs. ≥ 40 mg/dl)
			≤ 30 mg/dl	>30 and <40 mg/dl	≥ 40 mg/dl	
Males						
1	36.2 \pm 7.9 (549)	28.2	25.1 (109)	27.7 (309)	33.0 (131)	<0.02
2	34.9 \pm 6.4 (355)	18.2	20.2 (88)	18.0 (200)	16.9 (67)	NS
3	35.2 \pm 6.2 (352)	18.1	17.9 (78)	18.7 (208)	16.6 (66)	NS
4	35.0 \pm 7.0 (205)	10.5	10.6 (46)	10.9 (121)	9.6 (38)	NS
5	35.8 \pm 6.6 (43)	2.2	3.0 (13)	1.8 (20)	2.5 (10)	NS
6	35.2 \pm 6.1 (27)	1.4	1.2 (5)	1.5 (17)	1.3 (5)	NS
7	35.3 \pm 5.8 (25)	1.3	1.4 (6)	1.5 (17)	0.5 (2)	NS
8	34.2 \pm 9.8 (20)	1.0	1.6 (7)	0.9 (10)	0.8 (3)	NS
Females						
1	40.6 \pm 7.7 (325)	29.4	29.4 (90)	30.2 (144)	28.1 (91)	NS
2	41.0 \pm 7.8 (194)	17.5	18.0 (55)	15.9 (76)	19.5 (63)	NS
3	40.3 \pm 7.5 (188)	17.0	17.6 (54)	17.6 (84)	15.5 (50)	NS
4	41.4 \pm 8.1 (134)	12.1	10.1 (31)	12.8 (61)	13.0 (42)	NS
5	40.6 \pm 7.4 (25)	2.3	3.3 (10)	1.5 (7)	2.5 (8)	NS
6	40.8 \pm 7.6 (20)	1.8	2.0 (6)	1.9 (9)	1.5 (5)	NS
7	40.8 \pm 4.8 (8)	0.7	0.3 (1)	0.8 (4)	0.9 (3)	NS
8	34.5 \pm 2.1 (2)	0.2	0.3 (1)	0.3 (1)	0.0 (0)	NS

Percentages were analyzed by χ^2 test. NS, not significant.

cies, there was a statistically significant enrichment of haplotype 1 in males in the ≥ 40 mg/dl group ($\chi^2 = 6.5$, $P = 0.011$, Table 6). This further supported the association with the C-14T polymorphism.

We observed an interaction between the C-14T and R219K polymorphisms (Table 4). In an effort to construct haplotypes including these two polymorphisms, a haplotype block including five polymorphisms (T-564C, G-99C, C-14T, InsG 319, and R219K) was used for mapping. Fifteen haplotypes with a frequency $>1\%$ accounted about 96% of all haplotypes. However, over 44% of the diplotypes could not be discretely mapped to individual Turks. Therefore, haplotype-to-phenotype associations containing the C-14T and R219K polymorphisms within the same block could not be conducted.

3.9. Haplotype structure and haplotype-phenotype association of polymorphisms in the coding region of ABCA1

When the haplotype structure of the coding region was constructed using six polymorphisms (R219K, V771M, V825I, I883M, E1172D, and R1587K; Table 6), 12 haplotypes with a frequency $>1\%$ accounted for 95.1% of all haplotypes. About 81% of the predicted diplotypes could be

discretely mapped to unique genotypes, and were included for analysis. Plasma HDL-C levels did not differ between haplotype groups in this block (Table 8). Stratification by HDL-C levels yielded no additional information. The rare allele frequency of the V771M polymorphism, which was associated with high HDL-C in males, was 5.1% (Table 2). The M allele was primarily on haplotypes 9, 10, and 12, and the sum of their frequencies was 3.7%. Those remaining must be distributed among the rare ($<1\%$) haplotypes, and this distribution may explain the lack of association for the V771M on haplotype analysis.

We observed an interaction between the V771M and I883M polymorphisms (Table 4). In both males and females with 883IM, those with 771VM had significantly higher levels of HDL-C than subjects with 771VV (Table 4, $P < 0.05$). The rare alleles for these polymorphisms were never found together on common haplotypes (Tables 5 and 6). Therefore, their interaction could not be measured by haplotype association.

3.10. LD of ABCA1 polymorphisms

A few significant LDs were identified among the polymorphisms of ABCA1. Significant LD coefficients ($\pm D'$) and allele frequencies are shown in Table 5. A strong positive LD

Table 8
Mean plasma HDL-C levels of common haplotypes of ABCA1 in the coding region and their frequencies in a random Turkish population

Haplotype	Mean \pm S.D. (n)	All groups (%)	HDL-C subgroups (%) (n)			P (≤ 30 mg/dl vs. ≥ 40 mg/dl)
			≤ 30 mg/dl	>30 and <40 mg/dl	≥ 40 mg/dl	
Males						
1	35.0 \pm 6.9 (670)	35.1	38.2 (187)	32.8 (341)	37.1 (142)	NS
2	35.1 \pm 7.2 (198)	10.4	10.4 (51)	10.5 (109)	9.9 (38)	NS
3	34.7 \pm 6.6 (198)	10.4	9.0 (44)	11.8 (123)	8.1 (31)	NS
4	35.7 \pm 8.8 (130)	6.8	5.7 (28)	6.5 (67)	9.1 (35)	NS
5	34.5 \pm 6.7 (82)	4.3	3.9 (19)	4.9 (51)	3.1 (12)	NS
6	34.8 \pm 7.3 (32)	1.7	1.4 (7)	1.8 (19)	1.6 (6)	NS
7	35.5 \pm 7.5 (56)	2.9	2.9 (14)	2.8 (29)	3.4 (13)	NS
8	34.6 \pm 5.2 (34)	1.8	1.4 (7)	2.1 (22)	1.3 (5)	NS
9	34.1 \pm 6.6 (18)	0.9	1.0 (5)	1.1 (11)	0.5 (2)	NS
10	34.2 \pm 6.9 (32)	1.7	2.2 (11)	1.5 (16)	1.3 (5)	NS
11	36.2 \pm 8.9 (20)	1.0	0.8 (4)	1.1 (11)	1.3 (5)	NS
12	35.5 \pm 10.2 (14)	0.7	0.6 (3)	0.8 (8)	0.8 (3)	NS
Females						
1	40.7 \pm 9.3 (412)	34.7	34.7 (129)	35.8 (174)	32.9 (109)	NS
2	41.4 \pm 10.8 (104)	8.7	9.2 (34)	7.4 (36)	10.3 (34)	NS
3	39.7 \pm 7.7 (125)	10.5	10.5 (39)	10.3 (50)	10.9 (36)	NS
4	40.2 \pm 8.2 (91)	7.7	6.7 (25)	8.8 (43)	6.9 (23)	NS
5	39.1 \pm 10.6 (44)	3.7	4.6 (17)	3.1 (15)	3.6 (12)	NS
6	41.5 \pm 9.3 (26)	2.2	1.9 (7)	1.9 (9)	3.0 (10)	NS
7	39.2 \pm 7.1 (33)	2.8	3.2 (12)	2.3 (11)	3.0 (10)	NS
8	39.4 \pm 6.7 (27)	2.3	1.9 (7)	3.1 (15)	1.5 (5)	NS
9	41.5 \pm 8.2 (12)	1.0	0.8 (3)	1.0 (5)	1.2 (4)	NS
10	41.1 \pm 7.5 (20)	1.7	1.9 (7)	1.2 (6)	2.1 (7)	NS
11	40.6 \pm 8.5 (14)	1.2	1.1 (4)	1.4 (7)	0.9 (3)	NS
12	40.0 \pm 7.1 (11)	0.9	0.8 (3)	1.0 (5)	0.9 (3)	NS

Percentages were analyzed by χ^2 test. NS, not significant.

(rare allele in LD with rare allele) was observed between pairs of V771M and InsG 319, V825I and I883M, and R1587K and E1172D. There was negative LD (rare allele to common allele) between I883M and V771M as seen in Table 4. The R219K, C-14T, and V771M polymorphisms were not in LD in the Turkish population.

4. Discussion

Turks represent an ideal population for studying genes that influence HDL-C levels. Turks have extremely low levels of plasma HDL-C that appear to be, in part, of genetic origin [28,31,33]. More than 70% of Turkish men and 50% of Turkish women have HDL-C <40 mg/dl.

The cell-surface protein ABCA1, which controls the delivery of cholesterol and phospholipids from cells to form plasma HDL, has attracted significant attention because mutations in the ABCA1 gene have been linked to low or absent HDL-C in Tangier disease [4–8]. In the present study, we examined polymorphisms in ABCA1. To increase statistical power and reduce the risk of false associations, we used a large sample size; more than 2300 subjects were examined for informative polymorphisms. We used two approaches to analyze the association between common variants of ABCA1 and plasma HDL-C levels: single-locus analysis and haplotype analysis. Single-locus analysis is very helpful for assessing the combined effect of variants that are not in the same haplotype block. The haplotype analysis may provide valuable information if particular combinations of nucleotides are on the same haplotype. When a haplotype block contains a large number of haplotype alleles, it is difficult to correctly map the haplotype pairs to specific subjects if multiple haplotype pairs correspond to the same genotype. A solution is to assign the most probable haplotype pair to these subjects [19,44] or to exclude those who cannot be discretely mapped. Excluding a portion of the subjects from analysis may diminish the power of the study, but assigning the most probable haplotypes to particular individuals may lead to biased conclusions. We chose not to include ambiguously predicted subjects in our haplotype–phenotype analysis.

Another problem with haplotype analysis may arise if a genotype/phenotype association is only seen in homozygous recessive subjects. During a haplotype analysis, when the mean phenotype per haplotype allele is calculated, a haplotype allele may not show association with a phenotype even if the single locus homozygous genotype did show association. This is because, when the mean phenotype per haplotype is calculated, the phenotypes of the heterozygous subjects are included in the calculation, diluting the significance of the homozygous recessive association.

In this study, a functional promoter polymorphism, C-14T, was associated with elevated HDL-C in men. Furthermore, the -14TT genotype in combination with the 219KK genotype was associated with elevated HDL-C in both sexes. The 771VM genotype was associated with elevated HDL-C

by itself in men and, in combination with the 883IM genotype, associated with elevated HDL-C in both sexes.

The -14TT genotype was associated with higher HDL-C than the -14CC genotype (5.5% increase), and the -14T allele was significantly more frequent in the high HDL-C group. The C-14T polymorphism significantly affected transcriptional activity. In a reporter gene assay, the promoter containing the -14T allele displayed higher activity (20–88%) than the -14C allele. This functional polymorphism might protect against atherosclerosis in the subset of the population possessing the -14T allele because overexpression of human ABCA1 in mice increases HDL-C and apoA1 levels (an antiatherogenic profile) [45]. In our study, plasma apoA1 levels were slightly higher in -14TT than in -14CC subjects. On the other hand, low levels of ABCA1 expression with the common -14C allele were associated with low HDL-C and apoA1 levels.

Haplotype analysis was performed in the promoter and coding regions separately, an approach used by others [19,21]. In the analysis of the haplotype block of the promoter region, the mean HDL-C levels associated with haplotype 1 (possessing -14T variant only) was slightly higher than that associated with haplotype 4 (possessing all common alleles). The frequency of haplotype 1 was significantly higher in men with high HDL-C. Further analysis of subjects with both haplotypes the same revealed that the HDL-C levels of haplotype 1 were significantly higher than haplotype 4. We did not include ambiguously constructed diplotypes in our analysis; because we analyzed over 80% of haplotype–phenotype bi- data, we do not expect a major effect on the results. These results supported an association of HDL-C levels with the C-14T polymorphism in our study.

The -14T allele, found in 37.7% of the random Turkish population, occurs in 32–35% of Chinese, Malays, and Indians in Singapore [20], 38% of a random U.S. population [46], and 13.8% of Dutch men with proven CAD [18]. Consistent with our data, Chinese subjects with the -14T allele had higher HDL-C [20]. In neither Malay [20] nor Danish [17] subjects was there allelic differences between healthy controls and cardiovascular patients [20] or between low and high HDL-C groups [17]. On the other hand, Dutch men with CAD [18] and Indians with CAD [20] had an overrepresentation of the T allele. (In the study of the Dutch CAD patients, the polymorphism was referred to as C69T due to a difference in the sequence numbering used.) One might expect that the -14T allele would be underrepresented in CAD patients. Haplotype analysis showed that ABCA1 was a significant source of plasma HDL-C variation in German families [21], whereas no haplotype effect was observed on apoA1 variability or on the myocardial infarction risk in the ECTIM study participants [19]. However, haplotype mapping to distinguish the effect of C-14T on other CAD risk factors and to unravel the differences between the associations found in the populations studied and combined effect(s) among polymorphisms might be valuable in considering differences in HDL-C levels and CAD risk in different populations.

Gender differences were detected in cardiovascular risk factors and in the frequency of coronary events [47,48]. Interestingly, gender differences in the phenotypic expression of gene variants were also detected in lipid metabolism-related association studies [9,22–27]. It is important to mention that C–14T was not directly associated with altered HDL-C levels in females. This difference is likely to be complex and multifactorial, and it may involve sex hormones. We, and others, have previously suggested that gender and sex hormone levels may play a significant role in modulating HDL-C levels in Turks [29,49]. The gender differences in the phenotypic expression of ABCA1 variants could be related, at least in part, to environmental factors. In our study population males smoked more and consumed more alcohol than females. In the analysis of covariance, where the sources of variance on plasma HDL-C levels were examined, introduction of interaction variables (smoking or alcohol consumption) \times (C–14T or V771M) to the model showed that neither parameter modulated ($P > 0.05$) the effect of polymorphisms on HDL-C levels in males or females or when data for both genders were pooled. However, we are able to see an association with elevated HDL-C in females through the combined effect between C–14T and R219K. In a random sample of individuals, HDL-C levels were significantly higher in females and males with the 219KK and –14TT genotypes (Table 4). We tried to examine the combined effect of the C–14T and R219K polymorphisms by examining the haplotypes that contain these two rare allele genotypes; unfortunately, haplotype construction was complicated by an inability to predict a unique set of diplotypes. This prevented further statistical association of haplotypes that include both the C–14T and the R219K polymorphisms in the same haplotype block.

R219K by itself did not affect plasma lipid levels in Turks and Danes [17]. In a small Finnish study [9], however, the 219K allele was associated with elevated HDL-C in females, but not in males. This rare allele was also associated with decreased severity of atherosclerosis [12,15] and a reduction in coronary events [12] and was significantly less frequent in CAD patients than in subjects without CAD [19,43]. These protective effects were independent of plasma HDL-C levels in those studies [12,15,43]. On the other hand, the R219K polymorphism did not affect the severity of coronary atherosclerosis in the Veterans Administration HDL Cholesterol Intervention Trial [13] or in Japanese CAD patients [14]. In Turks, the interaction of C–14T and R219K may play a role in altering plasma HDL-C levels and possibly CAD risk.

In Turkish males but not females, the M allele of the V771M polymorphism was associated with high plasma HDL-C levels. After stratification by HDL-C levels, the M allele was significantly more frequent in the high than in the low HDL-C group, further supporting an association with high plasma HDL-C levels. The 771VM genotype was associated with high HDL-C in both males and females when it occurred in combination with the 1M genotype of the I883M polymorphism. Haplotype analysis of the coding region showed that 771M existed on haplotypes 9, 10, and 12 (frequencies of 1.3,

1.3, and 1.1%, respectively; Table 5). The total frequency of the 771M allele in this study was 5.1%, showing that over one quarter of the haplotypes containing the M allele exist at frequencies that are too low to test. Since 771M and 883M were not discovered on the same common haplotypes in the Turkish and European populations [17,50], their combined effect could not be confirmed by haplotype analysis.

The V771M polymorphism was associated with high HDL-C in females and consistent trends in males [17], whereas male subjects with the 771M allele had decreased focal atherosclerosis without alterations in plasma lipid levels [12].

The I883M polymorphism alone did not appear to alter HDL-C levels in Turks. However, Canadian Inuits with 883IM or MM have significantly higher HDL-C than those with the 883II genotype [10]. Similar findings have been reported in a Japanese population [11] and in Finnish females, but not males [9]. In contrast, the M allele of the I883M polymorphism was associated with increased progression of atherosclerosis [12]. The combination of V771M and I883M alone or with other factors may modulate plasma HDL-C levels and CAD risk. Additional studies are needed to assess the importance of those interactions in different ethnic populations.

The association with V825I and R1587K on plasma HDL-C levels were found in a large general Danish population [17]; however, it was not observed in Dutch men with CAD [18] or in our population.

In summary, the common polymorphisms of ABCA1 were associated with altered plasma lipid levels in a large random Turkish population. A functional promoter polymorphism, C–14T, and a coding sequence polymorphism, V771M, in the ABCA1 gene appeared to affect HDL-C levels in Turks and these results could be replicated when our entire data set was randomly divided in two different subsets (Supplemental Table II). Two combinations of rare alleles—C–14T with R219K and V771M with I883M—were associated with high HDL-C in both males and females. The four polymorphisms of ABCA1 described here, representing the rare alleles, were associated with a 6–9% increase in plasma HDL-C; conversely, the common alleles correlated with lower HDL-C. These observations may be significant in assessing the risk of CAD in Turks since a 1% change in plasma HDL-C is associated with a 2–3% inverse association with cardiovascular morbidity and mortality [51].

Acknowledgments

We are indebted to our associates at the American Hospital, Istanbul, especially Guy Pépin, Sibel Tanir, Judy Dawson-Pépin, and Linda L. Mahley at the Gladstone Institute (Istanbul), and Dr. K. Erhan Palaoğlu at the American Hospital (Istanbul). The authors gratefully acknowledge the technical assistance of and helpful discussions with Dr. Erwin Ludwig. We thank Sylvia Richmond and Jennifer

Polizzotto for manuscript preparation and Stephen Ordway and Gary Howard for editorial assistance. We acknowledge the generous support of the American Hospital, especially Mr. George Rountree, and the J. David Gladstone Institutes. This work was supported in part by R01 grant HL71027 from the National Institutes of Health. Preliminary data were presented at the meeting of the International Society of Atherosclerosis in Kyoto, Japan (September 2003).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.atherosclerosis.2005.03.004.

References

- [1] Murray CJL, Lopez AD. Mortality by cause for eight regions of the world: Global Burden of Disease study. *Lancet* 1997;349:1269–76.
- [2] Gordon DJ, Probstfield JL, Garrison RJ, et al. High-density lipoprotein cholesterol and cardiovascular disease. Four prospective American studies. *Circulation* 1989;79:8–15.
- [3] Jacobs Jr DR, Mebane IL, Bangdiwala SI, Criqui MH, Tyroler HA. High density lipoprotein cholesterol as a predictor of cardiovascular disease mortality in men and women: the follow-up study of the Lipid Research Clinics Prevalence study. *Am J Epidemiol* 1990;131:32–47.
- [4] Miller M, Rhyne J, Hamlette S, Birnbaum J, Rodriguez A. Genetics of HDL regulation in humans. *Curr Opin Lipidol* 2003;14:273–9.
- [5] Singaraja RR, Brunham LR, Visscher H, Kastelein JJP, Hayden MR. Efflux and atherosclerosis. The clinical and biochemical impact of variations in the ABCA1 gene. *Arterioscler Thromb Vasc Biol* 2003;23:1322–32.
- [6] Brewer Jr HB, Santamarina-Fojo S. Clinical significance of high-density lipoproteins and the development of atherosclerosis: focus on the role of the adenosine triphosphate-binding cassette protein A1 transporter. *Am J Cardiol* 2003;92(Suppl.):10K–6K.
- [7] Brousseau ME. ATP-binding cassette transporter A1, fatty acids, and cholesterol absorption. *Curr Opin Lipidol* 2003;14:35–40.
- [8] Schmitz G, Buechler C. ABCA1: Regulation, trafficking and association with heteromeric proteins. *Ann Med* 2002;34:334–47.
- [9] Kakko S, Kelloniemi J, von Rohr P, et al. ATP-binding cassette transporter A1 locus is not a major determinant of HDL-C levels in a population at high risk for coronary heart disease. *Atherosclerosis* 2003;166:285–90.
- [10] Wang J, Burnett JR, Near S, et al. Common and rare ABCA1 variants affecting plasma HDL cholesterol. *Arterioscler Thromb Vasc Biol* 2000;20:1983–9.
- [11] Harada T, Imai Y, Nojiri T, et al. A common Ile 823 Met variant of ATP-binding cassette transporter A1 gene (ABCA1) alters high density lipoprotein cholesterol level in Japanese population. *Atherosclerosis* 2003;169:105–12.
- [12] Clee SM, Zwinderman AH, Engert JC, et al. Common genetic variation in ABCA1 is associated with altered lipoprotein levels and a modified risk for coronary artery disease. *Circulation* 2001;103:1198–205.
- [13] Brousseau ME, Bodzioch M, Schaefer EJ, et al. Common variants in the gene encoding ATP-binding cassette transporter 1 in men with low HDL cholesterol levels and coronary heart disease. *Atherosclerosis* 2001;154:607–11.
- [14] Takagi S, Iwai N, Miyazaki S, Nonogi H, Goto Y. Relationship between ABCA1 genetic variation and HDL cholesterol level in subjects with ischemic heart diseases in Japanese. *Thromb Haemost* 2002;88:369–70.
- [15] Sandhofer AD, Iglseider B, Kaser S, et al. The common R219K variant in the ABCA1-gene reduces the risk of carotid atherosclerosis independently of high density lipoprotein cholesterol concentration in younger men. *Circulation* 2002;106:II-252 (abstr.).
- [16] Srinivasan SR, Li S, Chen W, Boerwinkle E, Berenson GS. R219K polymorphism of the ABCA1 gene and its modulation of the variations in serum high-density lipoprotein cholesterol and triglycerides related to age and adiposity in white versus black young adults. The Bogalusa heart study. *Metabolism* 2003;52:930–4.
- [17] Frikke-Schmidt R, Nordestgaard BG, Jensen GB, Tybjaerg-Hansen A. Genetic variation in ABC transporter A1 contributes to HDL cholesterol in the general population. *J Clin Invest* 2004;114:1343–53.
- [18] Zwarts KY, Ctee SM, Zwinderman AH, et al. ABCA1 regulatory variants influence coronary artery disease independent of effects on plasma lipid levels. *Clin Genet* 2002;61:115–25.
- [19] Tregouet D-A, Ricard S, Nicaud V, et al. In-depth haplotype analysis of ABCA1 gene polymorphisms in relation to plasma apoA1 levels and myocardial infarction. *Arterioscler Thromb Vasc Biol* 2004;24:775–81.
- [20] Tan JH-H, Low P-S, Tan Y-S, et al. ABCA1 gene polymorphisms and their associations with coronary artery disease and plasma lipids in males from three ethnic populations in Singapore. *Hum Genet* 2003;113:106–17.
- [21] Knoblauch H, Bauerfeind A, Toliat MR, et al. Haplotypes and SNPs in 13 lipid-relevant genes explain most of the genetic variance in high-density lipoprotein and low-density lipoprotein cholesterol. *Hum Mol Genet* 2004;13:993–1004.
- [22] Nofer J-R, von Eckardstein A, Wiebusch H, et al. Screening for naturally occurring apolipoprotein A-I variants: apo A-I(Δ K107) is associated with low HDL-cholesterol levels in men but not in women. *Hum Genet* 1995;96:177–82.
- [23] Couture P, Otvos JD, Cupples LA, et al. Association of the A-204C polymorphism in the cholesterol 7 α -hydroxylase gene with variations in plasma low density lipoprotein cholesterol levels in the Framingham Offspring Study. *J Lipid Res* 1999;40:1883–9.
- [24] Ludwig EH, Mahley RW, Palaoglu E, et al. DGAT1 promoter polymorphism associated with alterations in body mass index, high density lipoprotein levels and blood pressure in Turkish women. *Clin Genet* 2002;62:68–73.
- [25] Bauerfeind A, Knoblauch H, Schuster H, Luft FC, Reich JG. Single nucleotide polymorphism haplotypes in the cholesteryl-ester transfer protein (CETP) gene and lipid phenotypes. *Hum Hered* 2002;54:166–73.
- [26] Kauma H, Savolainen MJ, Heikkilä R, et al. Sex difference in the regulation of plasma high density lipoprotein cholesterol by genetic and environmental factors. *Hum Genet* 1996;97:156–62.
- [27] Ordovas JM. Lipoprotein lipase genetic variation and gender-specific ischemic cerebrovascular disease risk. *Nutr Rev* 2000;58:315–23.
- [28] Mahley RW, Palaoglu KE, Atak Z, et al. Turkish Heart Study: lipids, lipoproteins, and apolipoproteins. *J Lipid Res* 1995;36:839–59.
- [29] Mahley RW, Arslan P, Pekcan G, et al. Plasma lipids in Turkish children: impact of puberty, socioeconomic status, and nutrition on plasma cholesterol and HDL. *J Lipid Res* 2001;42:1996–2006.
- [30] Onat A. Risk factors and cardiovascular disease in Turkey. *Atherosclerosis* 2001;156:1–10.
- [31] Bersot TP, Vega GL, Grundy SM, et al. Elevated hepatic lipase activity and low levels of high density lipoprotein in a normotriglyceridemic, nonobese Turkish population. *J Lipid Res* 1999;40:432–8.
- [32] Friedewald WT, Levy RI, Fredrickson DS. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clin Chem* 1972;18:499–502.
- [33] Mahley RW, Pépin J, Palaoglu KE, et al. Low levels of high density lipoproteins in Turks, a population with elevated hepatic lipase:

- high density lipoprotein characterization and gender-specific effects of apolipoprotein E genotype. *J Lipid Res* 2000;41:1290–301.
- [34] Ludwig EH, Hopkins PN, Allen A, et al. Association of genetic variations in apolipoprotein B with hypercholesterolemia, coronary artery disease, and receptor binding of low density lipoproteins. *J Lipid Res* 1997;38:1361–73.
- [35] Santamarina-Fojo S, Peterson K, Knapper C, et al. Complete genomic sequence of the human ABCA1 gene: Analysis of the human and mouse ATP-binding cassette A promoter. *Proc Natl Acad Sci USA* 2000;97:7987–92.
- [36] Wellington CL, Walker EKY, Suarez A, et al. ABCA1 mRNA and protein distribution patterns predict multiple different roles and levels of regulation. *Lab Invest* 2002;82:273–83.
- [37] Schneider S, Roessli D, Excoffier L. Arlequin, Version 2.000: a software for Population Genetics Data Analysis, Genetics and Biometry Laboratory, University of Geneva, Switzerland. Available at: <http://lgb.unige.ch/arlequin/>.
- [38] Thompson EA, Deeb S, Walker D, Motulsky AG. The detection of linkage disequilibrium between closely linked markers: RFLPs at the AI-CIII apolipoprotein genes. *Am J Hum Genet* 1988;42:113–24.
- [39] Singaraja RR, Bocher V, James ER, et al. Human ABCA1 BAC transgenic mice show increased high density lipoprotein cholesterol and apoAI-dependent efflux stimulated by an internal promoter containing liver X receptor response elements in intron 1. *J Biol Chem* 2001;276:33969–79.
- [40] Burchfiel CM, Laws A, Benfante R, et al. Combined effects of HDL cholesterol, triglyceride, and total cholesterol concentrations on 18-year risk of atherosclerotic disease. *Circulation* 1995;92:1430–6.
- [41] Mahaney MC, Blangero J, Rainwater DL, et al. A major locus influencing plasma high-density lipoprotein cholesterol levels in the San Antonio Family Heart Study: segregation and linkage analyses. *Arterioscler Thromb Vasc Biol* 1995;15:1730–9.
- [42] Assmann G, Schulte H. Relation of high-density lipoprotein cholesterol and triglycerides to incidence of atherosclerotic coronary artery disease (the PROCAM experience). *Am J Cardiol* 1992;70:733–7.
- [43] Cenarro A, Artieda M, Castillo S, et al. A common variant in the ABCA1 gene is associated with a lower risk for premature coronary heart disease in familial hypercholesterolaemia. *J Med Genet* 2003;40:163–8.
- [44] Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 2001;68:978–89.
- [45] Joyce CW, Amar MIA, Lambert G, et al. The ATP binding cassette transporter A1 (ABCA1) modulates the development of aortic atherosclerosis in C57BL/6 and apoE-knockout mice. *Proc Natl Acad Sci USA* 2002;99:407–12.
- [46] Pullinger CR, Hakamata H, Duchateau PN, et al. Analysis of hABCA1 gene 5' end: additional peptide sequence, promoter region, and four polymorphisms. *Biochem Biophys Res Commun* 2000;271:451–5.
- [47] van Lennep JER, Westerveld HT, Erkelens DW, van der Wall EE. Risk factors for coronary heart disease: implications of gender. *Cardiovasc Res* 2002;53:538–49.
- [48] LaRosa JC. Triglycerides and coronary risk in women and the elderly. *Arch Intern Med* 1997;157:961–8.
- [49] Hergeng G, Schulte H, Assmann G, von Eckardstein A. Associations of obesity markers, insulin, and sex hormones with HDL-cholesterol levels in Turkish and German individuals. *Atherosclerosis* 1999;145:147–56.
- [50] Katzov H, Chalmers K, Palmgren J, et al. Genetic variants of ABCA1 modify Alzheimer disease risk and quantitative traits related to β -amyloid metabolism. *Hum Mutat* 2004;23:358–67.
- [51] Manninen V, Elo MO, Frick MH, et al. Lipid alterations and decline in the incidence of coronary heart disease in the Helsinki Heart Study. *J Am Med Assoc* 1988;260:641–51.

Chapter 7: Low HDL-C: lessons learned from the Turkish Heart Study



International Congress Series 1262 (2004) 193–199



www.ics-elsevier.com

Low HDL-C: lessons learned from the Turkish Heart Study

U. Hodoğlugil^a, D. Williamson^a, R.W. Mahley^{a,b,*}

^aGladstone Institute of Cardiovascular Disease, University of California, San Francisco, P.O. Box 419100, San Francisco, CA 94141-9100, USA

^bDepartments of Pathology and Medicine, University of California, San Francisco, P.O. Box 419100, San Francisco, CA 94141-9100, USA

Abstract. Low levels of high-density lipoprotein cholesterol (HDL-C) are highly prevalent in Turks. Analysis of approximately 10,000 Turkish men and women revealed the lowest HDL-C levels of any previously characterized population (~10–15 mg/dl lower than in western Europeans or Americans). These low HDL-C levels are primarily of genetic origin. A survey of the activities of candidate enzymes and transfer proteins showed that Turks have uniquely elevated hepatic lipase mass and activity, and their hepatic lipase levels are 25–30% higher than in non-Turkish American controls. As in other populations, Turkish newborns of both sexes have low total cholesterol and HDL-C levels (~30 mg/dl), and 8–9-year-old boys and girls have virtually identical HDL-C levels (50–60 mg/dl) similar to those in other populations. After puberty, however, HDL-C levels decrease markedly to typical adult levels, 36–37 mg/dl in males and 40–43 mg/dl in females. The mechanism for these striking reductions is unknown. Recent studies have shown that polymorphisms in the ATP-binding cassette A1 protein (ABCA1) and cholesterol ester transfer protein (CETP) are associated with altered HDL-C in Turks. Furthermore, a polymorphism in acyl CoA:diacylglycerol acyltransferase (DGAT) is associated with altered body mass index (BMI), HDL-C, and blood pressure in Turks. © 2004 Elsevier B.V. All rights reserved.

Keywords: HDL; Hepatic lipase; ATP-binding cassette A1; Cholesterol ester transfer protein; Acyl CoA:diacylglycerol acyltransferase

1. Introduction

Analysis of risk factors for coronary heart disease conducted as part of the Turkish Heart Study revealed that low levels of high-density lipoprotein cholesterol (HDL-C) are highly prevalent in Turks [1]. Lipid studies of approximately 10,000 Turkish adult men and women living in six different regions of Turkey demonstrated that the Turks have

* Corresponding author. Gladstone Institute of Cardiovascular Disease, University of California, San Francisco, P.O. Box 419100, San Francisco, CA 94141-9100, USA. Tel.: +1-415-826-7500; fax: +1-415-285-5632.

E-mail address: rmahley@gladstone.ucsf.edu (R.W. Mahley).

some of the lowest HDL-C levels of any population in the world (~ 53% of men and 26% of women had HDL-C levels <35 mg/dl). Turks living in Germany and the United States also had low HDL-C levels, suggesting that the condition is at least partly of genetic origin [2,3]. Low HDL-C was associated with an increase in hepatic lipase (levels 25–30% higher than in non-Turkish American controls) [2]. The prevalence of low HDL-C levels in Turks is not explained by hypertriglyceridemia or the metabolic syndrome [1–7]. The observations related to the unique lipid profile of Turks have been confirmed and extended by Onat et al. [4,5].

2. Discussion

2.1. Characterization of HDL subclasses

To define more precisely the factors associated with low HDL in Turks, we have characterized the HDL subclasses [6]. Turks have low levels of HDL₂, LpAI, and pre β -1 HDL and increased levels of LpAI/AII particles (potentially an atherogenic lipid profile). The frequency distributions of HDL-C and LpAI levels were skewed toward bimodality in Turkish women but were unimodal in Turkish men. The apolipoprotein (apo) E genotype affected HDL-C and LpAI levels in women only. In women, but not men, the ϵ 2 allele was strikingly more prevalent in those with the highest levels of HDL-C and LpAI than in those with the lowest levels [6]. The higher prevalence of the ϵ 2 allele in these subgroups of women was not explained by plasma triglyceride or total cholesterol level, age, or body mass index (BMI). The modulating effects of apoE isoforms on lipolytic hydrolysis of HDL by hepatic lipase (apoE2 prevents efficient hydrolysis) or on lipoprotein receptor binding (apoE2 interacts poorly with the LDL receptor) may account for differences in HDL-C levels in Turkish women (the ϵ 2 allele is associated with higher HDL levels). In Turkish men, who have substantially higher levels of hepatic lipase activity than women, the modulating effect of apoE may be overwhelmed. The gender-specific effect of the apoE genotype on HDL-C and LpAI levels in association with elevated levels of hepatic lipase provides new insights into the metabolism of HDL.

2.2. Effect of puberty on plasma lipids

Our most recent studies were designed to determine if HDL-C levels are unusually low throughout life in Turks and to assess the effect of puberty on plasma lipids in Turks [7]. The plasma lipids of cord blood of healthy newborns ($n=105$) and 8–10-year-old school children ($n=225$) have been analyzed in ethnic Turks [7]. Lipid values in typical western European populations and in Turks are shown in Table 1. Typically, newborns have very low total cholesterol and HDL-C levels that are virtually identical in males and females. Both total cholesterol and HDL-C levels rise with age, and there are no gender differences before puberty. The HDL-C levels of 8–9-year-old western European children were 50–60 mg/dl and were virtually identical in both males and females. Turkish newborns and prepubescent children have somewhat lower total cholesterol levels, but have similar HDL-C levels (mean for the prepubescent group, ~ 50 mg/dl) (Table 1).

After puberty, however, HDL-C levels in western European males decrease to typical adult levels (~ 47 mg/dl) and in females decrease much less to typical adult levels (~ 55–

Table 1
Plasma lipid levels change with age

	Western Europeans		Turks	
	Total cholesterol (mg/dl)	HDL-C (mg/dl)	Total cholesterol (mg/dl)	HDL-C (mg/dl)
Newborns	<100	~ 30	~ 60	~ 30
Male/female differences	none	none	none	none
Children	135–165	55–60	~ 140	50–60
Male/female differences	none	none	none	none
Young adults	>160	47–57	>160	37–43
Male/female differences	M>F	M: 66–60→47 F: 55–60→55	M>F	M: 58→37* F: 55→43

* Values shown here are for Turks in the upper socioeconomic groups whose diets are similar to those of western European populations.

57 mg/dl) (Table 1) (for reviews, see Refs. [7–9]). On the other hand, the HDL-C levels in Turkish boys drop from a mean of ~ 58 to 37 mg/dl and then the levels remain at 36–37 mg/dl during adulthood [7]. The HDL-C levels in Turkish girls decrease from ~ 55 to 43 mg/dl and remain at an average of 40–43 mg/dl in adulthood [7]. The changes in HDL-C levels in Turks are illustrated in Fig. 1.

The mechanism for the striking reduction in HDL-C levels after puberty in Turks is unknown. We hypothesize that androgen production plays a major role in modulating HDL-C levels at puberty. Consistent with this hypothesis is the observation from Hergenç et al. [10], suggesting that Turks have lower levels of sex hormone-binding globulin, which should result in increased levels of free bioactive testosterone in both males and females. Hepatic lipase production is regulated by androgens, and high levels of androgens are associated with increased levels and activity of hepatic lipase. Thus, high levels of free

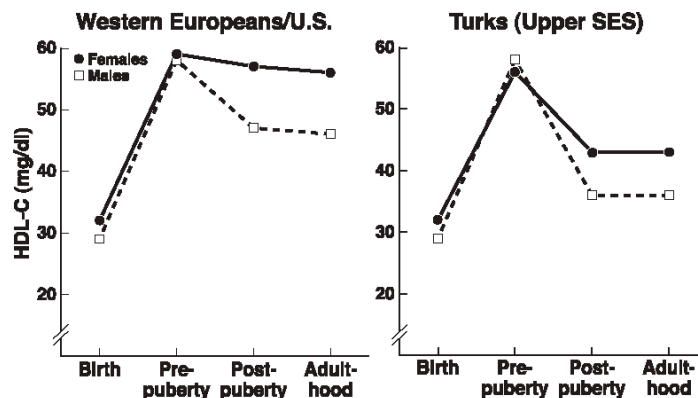


Fig. 1. Changes in HDL-C levels with age. There is a dramatic difference in the magnitude of the decrease in HDL-C levels post puberty in the western European versus Turkish males and females. Values for the Turkish children are those from upper socioeconomic status (SES) groups since the diet of this group closely resembles the European diets with respect to carbohydrate and fat.

testosterone may explain the high levels of hepatic lipase activity and protein mass that are characteristic of Turkish males and females.

We have begun to explore single nucleotide polymorphic sites that may be associated with lipid abnormalities and coronary artery disease in the Turkish population. To date, we have examined polymorphic sites in acyl CoA:diacylglycerol acyltransferase (DGAT), cholesterol ester transfer protein (CETP), and the ATP-binding cassette A1 protein (ABCA1).

2.3. DGAT polymorphisms

Studies characterizing DGAT promoter polymorphisms were undertaken using genomic DNA from Turkish Heart Study participants [11]. Plasma lipid and lipoprotein profiles in Turks have demonstrated lower plasma HDL-C levels and increased triglyceride levels. Turks also have slightly higher systolic and diastolic blood pressures than other populations. In addition, as in other populations, obesity has become a major health issue for Turkish people. About 50% of Turkish men and women were overweight (BMI > 25 kg/m²), and 11% of men and 22% of women were obese (BMI > 30 kg/m²). A more recent study has shown an increase in the prevalence of obesity in Turks, with 38% of Turkish women being obese [12]. These observations suggested that studies of the enzymes involved in triglyceride metabolism and adipose tissue biology might be revealing.

DGAT catalyzes the synthesis of triglyceride from fatty acyl CoA and diacylglycerol, and DGAT deficiency in mice is associated with leanness and resistance to diet-induced obesity [13–15]. In a recent study, we identified five polymorphisms within the human DGAT promoter and 5' noncoding sequence [11]. One common variant, a C to T transition 79 bases 3' of the transcriptional start site was associated with lower BMI, higher HDL-C, and lower systolic and diastolic blood pressures ($p=0.003, 0.058, 0.020$,

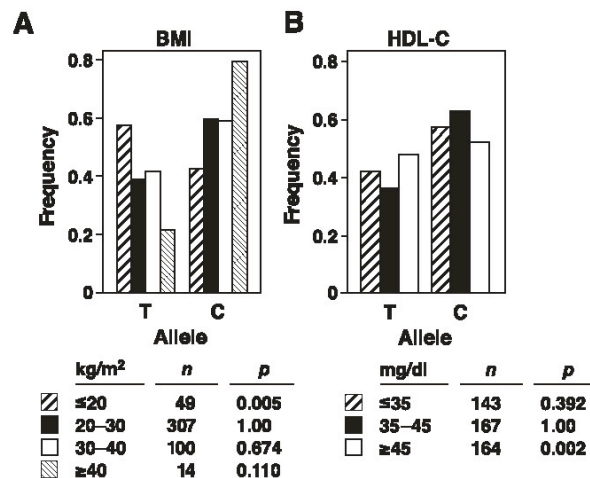


Fig. 2. Distributions and frequencies of the C79T alleles according to BMI and HDL-C in women. All p values were determined by chi-square test for independence. (A) Each group was compared with the control group (BMI > 20 to < 30 kg/m²). (B) The low and high HDL-C groups were compared with the moderate HDL group.

and 0.015, respectively) in randomly selected Turkish women ($n=476$). In a subsequent case-control study, the 79TT genotype and 79T allele were significantly associated with lower BMI ($p=0.0003$ and 0.0001 , respectively), higher HDL-C levels ($p=0.0039$ and 0.0016), lower diastolic ($p=0.0057$ and 0.0034), and lower systolic blood pressure ($p=0.0144$ and 0.0077) in Turkish women (Fig. 2). Functional analysis of the C79T polymorphism in the DGAT promoter by transient transfection experiments in cultured adipocytes, hepatocytes, and intestinal cells revealed a 20–33% decrease in promoter activity for the 79T allele compared with the 79C allele. Our data indicate that, in Turkish women, the DGAT C79T polymorphism is associated with alterations in three parameters that are features of the metabolic syndrome—body weight, HDL-C level, and blood pressure—and suggest that this polymorphism may contribute to these effects.

2.4. Cholesterol ester transfer protein

CETP facilitates the transfer of cholesterol esters and neutral lipids between HDL and apoB-containing lipoproteins (for a review, see Ref. [16]). High levels of CETP activity result in low HDL-C levels. Several groups have shown that the TaqIB polymorphism in CETP affects HDL-C levels [17–19]. Genotyping of 1219 Turkish males and 792 Turkish females revealed that the B2 polymorphism was associated with significantly higher HDL-

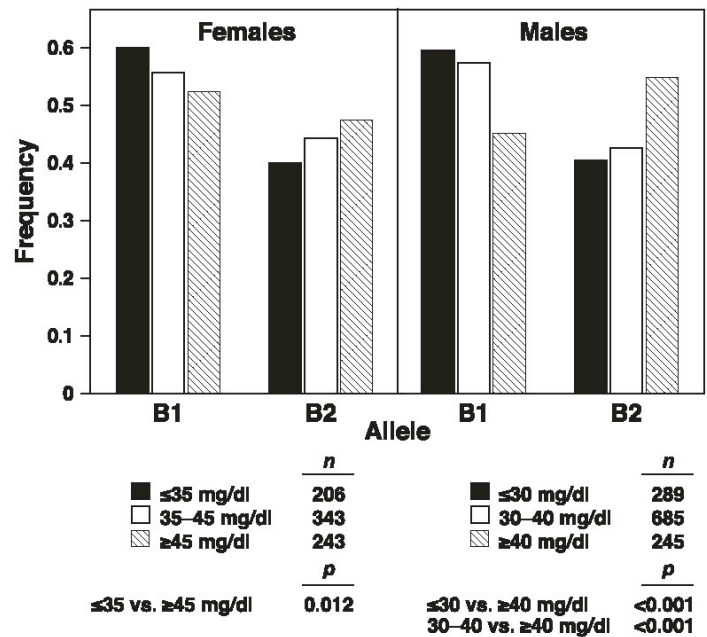


Fig. 3. Frequencies of the CETP TaqIB variation in a random male and female Turkish population. Alleles are defined as B1 (common) and B2 (rare). Allele frequencies are stratified by HDL-C level. Statistical significance was determined by chi-square analysis. n, number of subjects for each group.

C levels (Fig. 3). Individuals with the CETP B2B2 genotype had lower plasma CETP activity and higher HDL-C levels than those with the B1B1 genotype.

2.5. ATP-binding cassette A1 protein

ABCA1, which transports free cholesterol and phospholipid across plasma membranes to form HDL-C in the plasma (for a review, see Ref. [20]), represents an ideal target to study in this Turkish population with low HDL-C. Thus far, two polymorphic sites, one in the promoter region and the other in the coding sequence, have proven to be of interest.

In a random Turkish population of ~2700 men and women, the C–14T promoter polymorphism was associated with variable HDL-C levels. In men, the less frequent –14T (~38% in the Turkish population) allele was associated with significantly higher HDL-C levels (~5.5% increase) than the –14C allele. There was no association between the C–14T polymorphism and HDL-C levels in Turkish females. In cultured Cos-7, Chinese hamster ovary, and HepG2 cells expressing a luciferase construct encoding the –14T promoter polymorphism, luciferase activity increased by 25–75%.

The V771M variant in exon 16 was also associated with significantly higher HDL-C levels in Turkish males but not in females. We have begun to analyze haplotypes to demonstrate the effect of other ABCA1 polymorphic sites on V771M, especially in Turkish females. These studies are likely to reveal the importance of ABCA1 polymorphisms on HDL-C levels in Turks.

3. Summary

Continued studies of the Turkish population will shed light on the multiple genetic factors that modulate HDL-C levels. Undoubtedly, the interaction between genes and environment plays a key role in determining the lipid profile characteristic of Turks.

Acknowledgements

We are indebted to our associates at the American Hospital, Istanbul, especially Guy Pépin and Sibel Tanir in the Gladstone Institute (Istanbul), and Dr. K. Erhan Palaoğlu at the American Hospital (Istanbul). We thank Sylvia Richmond and Theresa Hashim for manuscript preparation and Stephen Ordway and Gary Howard for editorial assistance. We acknowledge the generous support of the American Hospital, especially Mr. George Rountree, and the J. David Gladstone Institutes. This work was supported in part by R01 grant HL71027 from the National Institutes of Health.

References

- [1] R.W. Mahley, et al., Turkish Heart Study: lipids, lipoproteins, and apolipoproteins, *J. Lipid Res.* 36 (1995) 839–859.
- [2] T.P. Bersot, et al., Elevated hepatic lipase activity and low levels of high density lipoprotein in a normotriglyceridemic, nonobese Turkish population, *J. Lipid Res.* 40 (1999) 432–438.
- [3] S. Lüttmann, et al., Electrophoretic screening for genetic variation in apolipoprotein C-III: identification of a novel apoC-III variant, apoC-III(Asp45→Asn), in a Turkish patient, *J. Lipid Res.* 35 (1994) 1431–1440.
- [4] A. Onat, et al., Plasma lipoproteins and apolipoproteins in Turkish adults: overall levels, associations with

- other risk parameters and HDL's role as a marker of coronary risk in women (in Turkish), Arch. Turk. Soc. Cardiol. 27 (1999) 72–79.
- [5] A. Onat, Risk factors and cardiovascular disease in Turkey, Atherosclerosis 156 (2001) 1–10.
- [6] R.W. Mahley, et al., Low levels of high density lipoproteins in Turks, a population with elevated hepatic lipase: high density lipoprotein characterization and gender-specific effects of apolipoprotein E genotype, J. Lipid Res. 41 (2000) 1290–1301.
- [7] R.W. Mahley, et al., Plasma lipids in Turkish children: impact of puberty, socioeconomic status, and nutrition on plasma cholesterol and HDL, J. Lipid Res. 42 (2001) 1996–2006.
- [8] C.M. Loughrey, et al., Race and gender differences in cord blood lipoproteins, Atherosclerosis 148 (2000) 57–65.
- [9] L.S. Webber, et al., Tracking of serum lipids and lipoproteins from childhood to adulthood. The Bogalusa Heart Study, Am. J. Epidemiol. 133 (1991) 884–899.
- [10] G. Hergeng, et al., Associations of obesity markers, insulin, and sex hormones with HDL-cholesterol levels in Turkish and German individuals, Atherosclerosis 145 (1999) 147–156.
- [11] E.H. Ludwig, et al., DGAT1 promoter polymorphism associated with alterations in body mass index, high density lipoprotein levels and blood pressure in Turkish women, Clin. Genet. 62 (2002) 68–73.
- [12] A. Onat, et al., Indices of obesity and central obesity in Turkish adults: distinct rise in obesity in 1990–98 more pronounced among men, (in Turkish). Arch. Turk. Soc. Cardiol. 27 (1999) 209–217.
- [13] S.J. Smith, et al., Obesity resistance and multiple mechanisms of triglyceride synthesis in mice lacking DGAT, Nat. Genet. 25 (2000) 87–90.
- [14] S. Cases, et al., Identification of a gene encoding an acyl CoA: diacylglycerol acyltransferase, a key enzyme in triacylglycerol synthesis, Proc. Natl. Acad. Sci. U. S. A. 95 (1998) 13018–13023.
- [15] R.V. Farese Jr., S. Cases, S.J. Smith, Triglyceride synthesis: insights from the cloning of diacylglycerol acyltransferase, Curr. Opin. Lipidol. 11 (2000) 229–234.
- [16] A. Ritsch, J.R. Patsch, Cholesteryl ester transfer protein: gathering momentum as a genetic marker and as drug target, Curr. Opin. Lipidol. 14 (2003) 173–178.
- [17] M. Corbex, et al., Extensive association analysis between the CETP gene and coronary heart disease phenotypes reveals several putative functional polymorphisms and gene–environment interaction, Genet. Epidemiol. 19 (2000) 64–80.
- [18] J.M. Ordovas, et al., Association of cholesteryl ester transfer protein–*TaqIB* polymorphism with variations in lipoprotein subclasses and coronary heart disease risk. The Framingham Study, Arterioscler. Thromb. Vasc. Biol. 20 (2000) 1323–1329.
- [19] M.E. Brousseau, et al., Cholesteryl ester transfer protein *TaqI B2B2* genotype is associated with higher HDL cholesterol levels and lower risk of coronary heart disease end points in men with HDL deficiency. Veterans Affairs HDL Cholesterol Intervention Trial, Arterioscler. Thromb. Vasc. Biol. 22 (2002) 1148–1154.
- [20] A.D. Attie, J.P. Kastelein, M.R. Hayden, Pivotal role of ABCA1 in reverse cholesterol transport influencing HDL levels and susceptibility to atherosclerosis, J. Lipid Res. 42 (2001) 1717–1726.

Thesis Discussion

Subject to the recommendation of my Thesis Committee, I believe that AIMS 1 and 2 have been completed.

1.23.5 Future Technology

I'd like to share some thoughts about what I've learned during my graduate program. It is evident the future of statistical genetics and genomics will continue to favor high throughput methods that can survey hundreds of thousands, to millions of markers simultaneously. Within the next three years we expect to have more than a 1,000 human genomes resequenced, and genotyping individual genomes will become a commonplace, pay-for-service. So obviously, as the number of data sets continues to grow, our ability to interpret these data will require novel computational methods. We need to ask ourselves the question:

What tools will predict the biological mechanism for the tens of thousands of associations that will be found?

I argue Delta-MATCH is a good example of a tool that can do this, at least for transcription factor binding. Interestingly, Delta-MATCH may be more relevant today than it was when I started the project. I think unlike many tools, the strength of Delta-MATCH is that it is extensible, and might endure the test of time because it is relatively easy to integrate additional data to its core.

Perhaps one of the most important things that I've learned during my graduate studies is how much time it takes to coordinate the collection of a large biological cohort, and how

quickly technologies become outdated. For example in just the course of the three years it took to collect these HIV samples, our lab progressed through using three separate genotyping platforms, HPLC, RFLP and TaqMan. Using these platforms I was only able to genotype dozens of SNPs, in a handful of candidate genes.

1.23.6 Controlling for Ethnicity

It should be emphasized a proper study design must require enough samples to have enough power to make a statistical conclusion. The study design of AIM 2 was initially flawed because although it appeared in the preliminary survey of TLR9 genotyping that the rs5743836 T>C minor (C) allele was enriched in the elite controllers (CVL-1), when compared to the noncontrollers (CVL-4), it was found that the elite controller group was enriched with a disproportionate amount of African-Americans. So an observation that appeared as a correlation to an HIV viremia phenotype was intrinsically confounded by an ethnicity-specific genotype frequency.

This design flaw was hard to avoid because I had little control over how many of individuals of each ethnicity were recruited into each CVL classification group. Part of the problem was that I wasn't provided, or didn't demand the gender and ethnicity of many of the blood samples before they were genotyped. This had the benefit of keeping the samples in a randomized order and blinded me from having a handling bias, but was harmful in that it resulted in me genotyping a much larger number of samples than were statistically compared in the final analyses. This ultimately wasted a lot of time, effort and money. For example, although over one hundred HIV-infected individuals from Brazil were genotyped for TLR9 and CCR5, the samples could not be properly analyzed because of the inability to control for the high frequency of ethnic admixture that is

intrinsic to the Brazilian population. Moreover, an entire subset of *hislat* individuals were probably genotyped without sufficient power to detect a significant association because their total numbers were low.

Why were there so many African-American controllers in this cohort? It could be because *afam* individuals have a genetic predisposition that protects them against viral replication similar to how the CCR5 del32 allele protects *whites* from HIV infection. Or, it may be that there were recruitment biases that targeted the *afam* population. What can be said is that when studying a genetic locus that is known to have strong ethnicity-specific polymorphism frequencies, care should be taken to control for ethnicity as best as possible during sample collection. Perhaps in the future, as we better define sets of ancestry informative markers, it will be possible to control for ethnicity not just globally across the genome, but locally with a resolution at the level of the gene haplotype. It appears there are groups achieving this computational goal, and it will soon be possible to study the genetic differences in a case and control study design between groups of admixed populations. This would create the benefit of allowing statisticians to use all of the samples in a cohort regardless of ethnicity, ultimately providing the clinicians more power to detect an association per recruitment effort.

1.23.7 Studying Rare Phenotypes

Another major contribution to why this project was inherently difficult to conclude was because the phenotype of the case group was exceedingly rare. It is estimated that less than 1 % of the HIV-infected population can be categorized as true elite controllers. This means that in order to collect 47 *white* elite controllers in group 1, there may have been close to 4,700 HIV-infected people screened. Even if this rareness is underestimated, it

follows that studying any rare phenotype will require a massive clinical recruitment effort that may require the contribution of collaborators, which always requires a lot of time.

If I were to restart the genotyping project again today with the DNA samples in hand, I would probably favor conducting a genome wide association study using an Affymetrix, or Illumina SNP chip. As is expected, this approach is being pursued by one of my collaborators (Bruce Walker), with many of the same biological samples that have been contributed to my cohort.

1.23.8 Transitioning

The database and freezer stock of biological samples from the HIV-1 investigation have been passed back to members of the McCune lab, and of my genotyping notebooks will be provided. The data for the genetic investigation of [CCR5](#), [TLR9](#), and [IRF5](#), are being consolidated for publication, but the data for the survey of [APOE](#) will not be published.

A manuscript announcing the Delta-MATCH database will soon be submitted. Once published, the Delta-MATCH website will be opened to the public, and its source code will be open sourced. Katie Pollard and Francois Guillemot are investigating the NEUROG2/[PAX6](#) interaction in HAR152 in the United Kingdom. I hope in the future, the Delta-MATCH query tool will continue to be optimized so its derived list of candidate SNPs may be validated through collaborations using high-throughput genotyping technologies ([Affymetrix](#), [Illumina](#), etc.).

I conclude, it will be the challenge of the future, to build hypothesis-generating tools like Delta-MATCH that integrate useful orthogonal data sets and can predict the biological

mechanism of human diseases, so that these predictions may be validated by the molecular biologists.

Bibliography

1. Kel, A.E., et al., *MATCH: A tool for searching transcription factor binding sites in DNA sequences*. Nucleic Acids Res, 2003. **31**(13): p. 3576-9.
2. Ng, P.C. and S. Henikoff, *SIFT: Predicting amino acid changes that affect protein function*. Nucleic Acids Res, 2003. **31**(13): p. 3812-4.
3. Wingender, E., et al., *TRANSFAC: a database on transcription factors and their DNA binding sites*. Nucleic Acids Res, 1996. **24**(1): p. 238-41.
4. Loots, G.G. and I. Ovcharenko, *rVISTA 2.0: evolutionary analysis of transcription factor binding sites*. Nucleic Acids Res, 2004. **32**(Web Server issue): p. W217-21.
5. Guo, Y. and D.C. Jamison, *The distribution of SNPs in human gene regulatory regions*. BMC Genomics, 2005. **6**: p. 140.
6. Riva, A. and I.S. Kohane, *SNPper: retrieval and analysis of human SNPs*. Bioinformatics, 2002. **18**(12): p. 1681-5.
7. Shen, J., P.L. Deininger, and H. Zhao, *Applications of computational algorithm tools to identify functional SNPs in cytokine genes*. Cytokine, 2006. **35**(1-2): p. 62-6.
8. Wang, P., et al., *SNP Function Portal: a web database for exploring the function implication of SNP alleles*. Bioinformatics, 2006. **22**(14): p. e523-9.
9. Kasprzyk, A., et al., *Ensmart: a generic system for fast and flexible access to biological data*. Genome Res, 2004. **14**(1): p. 160-9.
10. Shah, N., et al., *SNP-VISTA: an interactive SNP visualization tool*. BMC Bioinformatics, 2005. **6**: p. 292.
11. Kashuk, C., et al., *ViewGene: a graphical tool for polymorphism visualization and characterization*. Genome Res, 2002. **12**(2): p. 333-8.
12. Mahley, R.W. and S.C. Rall, Jr., *Apolipoprotein E: far more than a lipid transport protein*. Annu Rev Genomics Hum Genet, 2000. **1**: p. 507-37.
13. Mahley, R.W., K.H. Weisgraber, and Y. Huang, *Apolipoprotein E4: a causative factor and therapeutic target in neuropathology, including Alzheimer's disease*. Proc Natl Acad Sci U S A, 2006. **103**(15): p. 5644-51.
14. Corder, E.H., et al., *HIV-infected subjects with the E4 allele for APOE have excess dementia and peripheral neuropathy*. Nat Med, 1998. **4**(10): p. 1182-4.
15. Buttini, M., et al., *Dominant negative effects of apolipoprotein E4 revealed in transgenic models of neurodegenerative disease*. Neuroscience, 2000. **97**(2): p. 207-10.
16. Michelsen, K.S., et al., *Lack of Toll-like receptor 4 or myeloid differentiation factor 88 reduces atherosclerosis and alters plaque phenotype in mice deficient in apolipoprotein E*. Proc Natl Acad Sci U S A, 2004. **101**(29): p. 10679-84.
17. Graham, R.R., et al., *A common haplotype of interferon regulatory factor 5 (IRF5) regulates splicing and expression and is associated with increased risk of systemic lupus erythematosus*. Nat Genet, 2006. **38**(5): p. 550-5.
18. Graham, D.S., et al., *Association of IRF5 in UK SLE Families Identifies a Variant Involved in Polyadenylation*. Hum Mol Genet, 2006.

19. Graham, R.R., Kyogoku, D., et al., *Three functional variants of interferon regulatory factor 5 (IRF5) define risk and protective haplotypes for human SLE*. prepublication, 2007.
20. Bochud, P.Y., et al., *Polymorphisms in Toll-like receptor 9 influence the clinical course of HIV-1 infection*. *Aids*, 2007. **21**(4): p. 441-446.
21. Huang, Y., et al., *The role of a mutant CCR5 allele in HIV-1 transmission and disease progression*. *Nat Med*, 1996. **2**(11): p. 1240-3.
22. Algood, H.M. and J.L. Flynn, *CCR5-deficient mice control Mycobacterium tuberculosis infection despite increased pulmonary lymphocytic infiltration*. *J Immunol*, 2004. **173**(5): p. 3287-96.
23. *A haplotype map of the human genome*. *Nature*, 2005. **437**(7063): p. 1299-320.
24. Blanchette, M., et al., *Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression*. *Genome Res*, 2006. **16**(5): p. 656-68.
25. Ferretti, V., et al., *PReMod: a database of genome-wide mammalian cis-regulatory module predictions*. *Nucleic Acids Res*, 2007. **35**(Database issue): p. D122-6.
26. Suarez, B.K., et al., *An analysis of identical single-nucleotide polymorphisms genotyped by two different platforms*. *BMC Genet*, 2005. **6 Suppl 1**: p. S152.
27. Matys, V., et al., *TRANSFAC: transcriptional regulation, from patterns to profiles*. *Nucleic Acids Res*, 2003. **31**(1): p. 374-8.
28. Matys, V., et al., *TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes*. *Nucleic Acids Res*, 2006. **34**(Database issue): p. D108-10.
29. Inoue, M., et al., *A promoter polymorphism of the alpha2-HS glycoprotein gene is associated with its transcriptional activity*. *Diabetes Res Clin Pract*, 2007.
30. Siddiq, A., et al., *A synonymous coding polymorphism in the alpha2-Heremans-schmid glycoprotein gene is associated with type 2 diabetes in French Caucasians*. *Diabetes*, 2005. **54**(8): p. 2477-81.
31. Saxena, R., et al., *Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels*. *Science*, 2007. **316**(5829): p. 1331-6.
32. Lewontin, R.C., *On measures of gametic disequilibrium*. *Genetics*, 1988. **120**(3): p. 849-52.
33. Pritchard, J.K. and M. Przeworski, *Linkage disequilibrium in humans: models and data*. *Am J Hum Genet*, 2001. **69**(1): p. 1-14.
34. Fellay, J., et al., *A whole-genome association study of major determinants for host control of HIV-1*. *Science*, 2007. **317**(5840): p. 944-7.
35. Iafrate, A.J., et al., *Detection of large-scale variation in the human genome*. *Nat Genet*, 2004. **36**(9): p. 949-51.
36. Komura, D., et al., *Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays*. *Genome Res*, 2006. **16**(12): p. 1575-84.
37. Kent, W.J., et al., *The human genome browser at UCSC*. *Genome Res*, 2002. **12**(6): p. 996-1006.
38. Karolchik, D., et al., *The UCSC Genome Browser Database*. *Nucleic Acids Res*, 2003. **31**(1): p. 51-4.

39. Hirata, Y., et al., *Uncoupling store-operated Ca²⁺ entry and altered Ca²⁺ release from sarcoplasmic reticulum through silencing of junctophilin genes*. *Biophys J*, 2006. **90**(12): p. 4418-27.
40. Takeshima, H., et al., *Junctophilins: a novel family of junctional membrane complex proteins*. *Mol Cell*, 2000. **6**(1): p. 11-22.
41. Minamisawa, S., et al., *Junctophilin type 2 is associated with caveolin-3 and is down-regulated in the hypertrophic and dilated cardiomyopathies*. *Biochem Biophys Res Commun*, 2004. **325**(3): p. 852-6.
42. Guidetti, P. and R. Schwarcz, *3-Hydroxykynurenine and quinolinate: pathogenic synergism in early grade Huntington's disease?* *Adv Exp Med Biol*, 2003. **527**: p. 137-45.
43. Giorgini, F., et al., *A genomic screen in yeast implicates kynurenine 3-monooxygenase as a therapeutic target for Huntington disease*. *Nat Genet*, 2005. **37**(5): p. 526-31.
44. Pollard, K.S., et al., *Forces shaping the fastest evolving regions in the human genome*. *PLoS Genet*, 2006. **2**(10): p. e168.
45. Pollard, K.S., et al., *An RNA gene expressed during cortical development evolved rapidly in humans*. *Nature*, 2006. **443**(7108): p. 167-72.
46. Scardigli, R., et al., *Crossregulation between Neurogenin2 and pathways specifying neuronal identity in the spinal cord*. *Neuron*, 2001. **31**(2): p. 203-17.
47. Kawaguchi, A., et al., *Differential expression of Pax6 and Ngn2 between pair-generated cortical neurons*. *J Neurosci Res*, 2004. **78**(6): p. 784-95.
48. Scardigli, R., et al., *Direct and concentration-dependent regulation of the proneural gene Neurogenin2 by Pax6*. *Development*, 2003. **130**(14): p. 3269-81.
49. O'Brien, T.R., et al., *Serum HIV-1 RNA levels and time to development of AIDS in the Multicenter Hemophilia Cohort Study*. *JAMA*, 1996. **276**(2): p. 105-10.
50. Lambotte, O., et al., *HIV controllers: a homogeneous group of HIV-1-infected patients with spontaneous control of viral replication*. *Clin Infect Dis*, 2005. **41**(7): p. 1053-6.
51. Bafica, A., et al., *The induction of Toll-like receptor tolerance enhances rather than suppresses HIV-1 gene expression in transgenic mice*. *J Leukoc Biol*, 2004. **75**(3): p. 460-6.
52. Schoenemeyer, A., et al., *The interferon regulatory factor, IRF5, is a central mediator of toll-like receptor 7 signaling*. *J Biol Chem*, 2005. **280**(17): p. 17005-12.
53. Bachelierie, F., et al., *HIV enhancer activity perpetuated by NF-kappa B induction on infection of monocytes*. *Nature*, 1991. **350**(6320): p. 709-12.
54. O'Neill, L.A., *How Toll-like receptors signal: what we know and what we don't know*. *Curr Opin Immunol*, 2006. **18**(1): p. 3-9.
55. Lazarus, R., et al., *Single-nucleotide polymorphisms in the Toll-like receptor 9 gene (TLR9): frequencies, pairwise linkage disequilibrium, and haplotypes in three U.S. ethnic groups and exploratory case-control disease association studies*. *Genomics*, 2003. **81**(1): p. 85-91.
56. Barrett, J.C., et al., *Haploview: analysis and visualization of LD and haplotype maps*. *Bioinformatics*, 2005. **21**(2): p. 263-5.

57. Hunt, P.W., et al., *Prevalence of CXCR4 tropism among antiretroviral-treated HIV-1-infected patients with detectable viremia*. J Infect Dis, 2006. **194**(7): p. 926-30.
58. Latz, E., et al., *TLR9 signals after translocating from the ER to CpG DNA in the lysosome*. Nat Immunol, 2004. **5**(2): p. 190-8.
59. Li, Y., et al., *IL-18 gene therapy develops Th1-type immune responses in Leishmania major-infected BALB/c mice: is the effect mediated by the CpG signaling TLR9?* Gene Ther, 2004. **11**(11): p. 941-8.
60. Hemmi, H., et al., *A Toll-like receptor recognizes bacterial DNA*. Nature, 2000. **408**(6813): p. 740-5.
61. Lachheb, J., et al., *Toll-like receptors and CD14 genes polymorphisms and susceptibility to asthma in Tunisian children*. Tissue Antigens, 2008.
62. Carvalho, A., et al., *Polymorphisms in Toll-Like Receptor Genes and Susceptibility to Pulmonary Aspergillosis*. J Infect Dis, 2008. **197**(4): p. 618-621.
63. Novak, N., et al., *Putative association of a TLR9 promoter polymorphism with atopic eczema*. Allergy, 2007. **62**(7): p. 766-72.
64. Carvalho, A., et al., *T-1237C polymorphism of TLR9 gene is not associated with multiple sclerosis in the Portuguese population*. Mult Scler, 2008.
65. Mockenhaupt, F.P., et al., *Toll-like receptor (TLR) polymorphisms in African children: Common TLR-4 variants predispose to severe malaria*. Proc Natl Acad Sci U S A, 2006. **103**(1): p. 177-82.
66. Hong, J., et al., *TLR2, TLR4 and TLR9 polymorphisms and Crohn's disease in a New Zealand Caucasian cohort*. J Gastroenterol Hepatol, 2007. **22**(11): p. 1760-6.
67. De Jager, P.L., et al., *Genetic variation in toll-like receptor 9 and susceptibility to systemic lupus erythematosus*. Arthritis Rheum, 2006. **54**(4): p. 1279-82.
68. Demirci, F.Y., et al., *Association study of Toll-like receptor 5 (TLR5) and Toll-like receptor 9 (TLR9) polymorphisms in systemic lupus erythematosus*. J Rheumatol, 2007. **34**(8): p. 1708-11.
69. Ito, A., et al., *Lack of association of Toll-like receptor 9 gene polymorphism with Behcet's disease in Japanese patients*. Tissue Antigens, 2007. **70**(5): p. 423-6.
70. Malissen, B. and J.J. Ewbank, *'TaiLoRing' the response of dendritic cells to pathogens*. Nat Immunol, 2005. **6**(8): p. 749-50.
71. Cheung, V.G., et al., *Mapping determinants of human gene expression by regional and genome-wide association*. Nature, 2005. **437**(7063): p. 1365-9.
72. Mancl, M.E., et al., *Two discrete promoters regulate the alternatively spliced human interferon regulatory factor-5 isoforms. Multiple isoforms with distinct cell type-specific expression, localization, regulation, and function*. J Biol Chem, 2005. **280**(22): p. 21078-90.
73. Shumway, S.D., M. Maki, and S. Miyamoto, *The PEST domain of IkappaBalpha is necessary and sufficient for in vitro degradation by mu-calpain*. J Biol Chem, 1999. **274**(43): p. 30874-81.
74. Weisgraber, K.H., S.C. Rall, Jr., and R.W. Mahley, *Human E apoprotein heterogeneity. Cysteine-arginine interchanges in the amino acid sequence of the apo-E isoforms*. J Biol Chem, 1981. **256**(17): p. 9077-83.

75. Utermann, G., N. Pruin, and A. Steinmetz, *Polymorphism of apolipoprotein E. III. Effect of a single polymorphic gene locus on plasma lipid levels in man.* Clin Genet, 1979. **15**(1): p. 63-72.
76. Tarr, P.E., et al., *Modeling the influence of APOC3, APOE, and TNF polymorphisms on the risk of antiretroviral therapy-associated lipid disorders.* J Infect Dis, 2005. **191**(9): p. 1419-26.
77. Valcour, V., et al., *Age, apolipoprotein E4, and the risk of HIV dementia: the Hawaii Aging with HIV Cohort.* J Neuroimmunol, 2004. **157**(1-2): p. 197-202.

Appendices

1.24 AIM 1 Extras (Delta-MATCH)

Figure 201 The BIOBASE MATCH Program Version 10.2

MATCH™ 10.2
Matrix Search for Transcription Factor Binding Sites
your login name: d

BIOBASE
Biological Databases /
Biologische Datenbanken GmbH

Sequence Selection:

Select one of your stored sequences:

Our example: (EMBL-Seq; RNTAFEL standard; DNA; ROD; 11973 BP)

Upload a file: no file selected

Enter a [new sequence](#):

(Allowed formats are: [RAW](#), [FASTA](#), [TRANSFAC](#), [EMBL](#), [GenBank](#), [DDBJ](#), [IG](#))

Matrix or profile selection: [Get help](#)

[Profiles](#) (group of matrices):

Use only [high quality](#) matrices

Include matrices created by the user

Cut-off selection for matrix group or our profile:

[minimize false positives](#)

[minimize false negatives](#)

[minimize the sum of both error rates](#)

[mat. sim.](#)

[core sim.](#)

Archive:

Select a previous search result:

Delete a previously stored sequence:

Mail to info@biobase.de

Figure 202 How to Calculate a MATCH Score [1]

The matrix similarity score mSS (as well as the core similarity score) for a subsequence x of the length L is calculated in the following way:

$$mSS = \frac{Current - Min}{Max - Min} \quad 1$$

$$Current: \sum_{i=1}^L I(i)f_{i,b_i}$$

$f_{i,B}$, frequency of nucleotide B to occur at the position i of the matrix ($B \in \{A, T, G, C\}$)

$$Min: \sum_{i=1}^L I(i)f_i^{\min}$$

f_i^{\min} , frequency of the nucleotide which is rarest in position i in the matrix

$$Max: \sum_{i=1}^L I(i)f_i^{\max}$$

f_i^{\max} , highest frequency in position i .

The information vector

$$I(i) = \sum_{B \in \{A, T, G, C\}} f_{i,B} \ln(4f_{i,B}), \quad i = 1, 2, \dots, L \quad 2$$

describes the conservation of the positions i in a matrix (5). Multiplication of the frequencies with the information vector leads to a higher acceptance of mismatches in less conserved regions, whereas mismatches in highly conserved regions are very much discouraged. This leads to a better performance in recognition of TF binding sites if compared with methods that do not use the information vector (6).

Figure 203 MATCH Score Calculation [1]

How have the matrices been constructed?

For matrix generation, orthologous factors and their appropriate binding sites are selected from those sites which have been collected in TRANSFAC® and TRANSFAC® site sequences are extended, 10bp at both ends, by using the EMBL links in the site entries. The extended sequences are aligned by using the Gibbs site sampling algorithm (Lawrence et al. (1993) Science 262: 208-214). In the initial step of the sampling, a short matrix is constructed either by using the average of the annotated "sites core" versus lower case in the site sequences) or, where no "sites core" is highlighted, by using a window size of $w = 6$. Using the Match™ algorithm, this initial matrix is then used to build an ungapped alignment of the site set. On the basis of this alignment, a weight matrix $M(i, j)$ is constructed for each window (i, j) , where $(1 \leq i, j \leq \text{length}(\text{alignment}))$ and $(w = j - i + 1 > 6 \text{ bp})$. For each of the constructed matrices, rates of false positives (FP) and false negatives (FN) are estimated: for the estimation of the FP rates, binding sites from the set (Jack-knife test) as well as computationally generated oligonucleotides are used. In the Jack-knife test the set is temporally removed and the matrix is constructed on the basis of the remaining sites. The removed site is then searched by the matrix. This is done for each site to calculate the FN rate of the matrix. On the basis of the FP and FN rates, the best matrix window (i, j) is selected. Finally, the different cut-offs (minFN, minFP, ...) are selected as described in the respective section of the documentation for Match™. Please note, not all matrices have been constructed according to the above scheme. Individual entries and the linked references for information, how the respective matrices were constructed.

Table 39 Delta-MATCH Tissue Types.

	Tissue
1	Adipocyte Specific
2	Liver Specific
3	Immune Cell Specific
4	Lung Specific
5	Muscle Specific
6	Nerve Cell Specific
7	Pituitary Specific
8	Pancreatic Beta Cell Specific
9	Cell Cycle Specific
10	Glioma

Table 40 NF-kB TFBS Matrixes Used by Delta-MATCH

	factor	mat_id	matrix_length	FP
1	NF-kappaB	V\$NFKAPPAB50_01	10	1.000
2	NF-kappaB	V\$NFKAPPAB65_01	10	0.991
3	NF-kappaB	V\$NFKAPPAB_01	10	0.984
4	NF-kappaB	V\$NFKB_C	12	0.988
5	NF-kappaB	V\$NFKB_Q6	14	0.955
6	NF-kappaB	V\$NFKB_Q6_01	16	0.876

Table 41 Distribution of Potential Scores (dif_z) for NF-kB TFBS Matrixes

TFBS matrix		0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
V\$NFKAPPAB50_01	4547844	0	0	0	0	0	0	0	0	0	0	0
V\$NFKAPPAB65_01	4547677	167	167	167	167	167	167	167	167	167	167	167
V\$NFKB_C	4547761	63	48	48	48	48	48	48	46	33	13	13
V\$NFKAPPAB_01	4547339	505	505	361	361	220	212	212	153	153	153	153
V\$NFKB_Q6_01	4523684	9753	8889	6115	4746	1843	776	597	352	323	0	0
V\$NFKB_Q6	4546894	676	671	409	333	179	72	71	61	61	7	7

Table 42 List of 351 Transcription Factors

1	ACAAAT	76	C_EBP	151	HIF-1	226	NGFI-C	301	Staf
2	AFFP1	77	D-1type	152	HLF	227	Nkx2-2	302	STAT
3	Ahr	78	DBP	153	HMG	228	Nkx2-5	303	STAT1
4	Ahr:Arnt	79	DEAF1	154	HMG1Y	229	NKX25	304	STAT3
5	AHRHIF	80	DEC	155	Hmx3	230	NKX3A	305	STAT5A
6	AIRE	81	deltaEF1	156	HNF-1	231	NKX6-1	306	STAT5B
7	alpha-CP1	82	E12	157	HNF-3	232	Nkx6-2	307	STATx
8	Aix-4	83	E2	158	HNF-3alpha	233	Nrf-1	308	Stra13
9	aMEF-2	84	E2A	159	HNF-3beta	234	NRF-2	309	TALI
10	AML	85	E2F	160	HNF-4	235	Nrf2	310	TATA
11	AML-1a	86	E2F-1	161	HNF-4alpha	236	NRSF	311	Tax/CREB
12	AML1	87	E2F-1:DP-1	162	HNF-4alpha1	237	Oct-1	312	TBP
13	AP-1	88	E2F-1:DP-2	163	HNF-6	238	OCT-x	313	TBX5
14	AP-2	89	E2F-4:DP-1	164	HNF1	239	Octamer	314	TCF-4
15	AP-2alpha	90	E2F-4:DP-2	165	HNF3	240	Olf-1	315	TCF11
16	AP-2alphaA	91	E47	166	HNF4	241	Osf2	316	TCF11:MatG
17	AP-2gamma	92	E4BP4	167	Hox-1.3	242	p300	317	TEF
18	AP-2rep	93	E4F1	168	HOUA3	243	p53	318	TEF-1
19	AP-3	94	EBF	169	HOUA4	244	Pax	319	Tel-2
20	AP-4	95	Ebox	170	HSF	245	Pax-1	320	TFE
21	APOLYA	96	EGR	171	HSF1	246	Pax-2	321	TFII-1
22	AR	97	Egr-1	172	HSF2	247	Pax-3	322	TFIIA
23	AREB6	98	Egr-2	173	HTF	248	Pax-4	323	TGIF
24	Arnt	99	Egr-3	174	ICSBP	249	Pax-5	324	Tst-1
25	ARP-1	100	ELF-1	175	Ik-1	250	Pax-6	325	TTF-1
26	ATATA	101	Elk-1	176	Ik-2	251	Pax-8	326	TTF1
27	ATF	102	En-1	177	Ik-3	252	Pax-9	327	UFH3BETA
28	ATF-1	103	ER	178	IRF1	253	PAX6	328	USF
29	ATF3	104	ETF	179	IRF	254	PBX	329	USF2
30	ATF4	105	ETS	180	IRF-1	255	Pbx-1	330	v-ErbA
31	ATF6	106	Evi-1	181	IRF-2	256	PBX1	331	v-Jun
32	Bach1	107	FAC1	182	IRF-7	257	Pbx1b	332	v-Maf
33	Bach2	108	FOX	183	IRF1	258	PEA3	333	v-Myb
34	Barbie	109	FOXO3	184	ISRE	259	PEBP	334	VBP
35	Bel-1	110	FOXJ2	185	KROX	260	Pit-1	335	VDR
36	BLIMP1	111	FOXM1	186	LBP-1	261	PITX2	336	VDR
37	Brachyury	112	FOXO1	187	LEF1	262	PLZF	337	Wim
38	BRCA1	113	FOXO3	188	LEFTCF1	263	Poly	338	XBP-1
39	Brm-2	114	FOXO4	189	Lentiviral	264	POU1F1	339	XFD-1
40	c-Ets-1	115	FOXP1	190	LF-A1	265	POU3F2	340	XFD-2
41	c-Ets-1(p54)	116	FOXP3	191	Lhx3	266	POU6F1	341	XFD-3
42	c-Ets-2	117	Freac-2	192	Lmo2	267	PPAR	342	XPF-1
43	c-Maf	118	Freac-3	193	LRF	268	PPAR,	343	Xvent-1
44	c-Myb	119	Freac-4	194	LUN-1	269	PPARalpha:RXR-alpha	344	YY1
45	c-Myc:Max	120	Freac-7	195	LXR	270	PPARG	345	Zec
46	c-R4	121	FXR	196	LXR	271	PR	346	ZF5
47	C/EBP	122	FXR/RXR-alpha	197	Lyt-1	272	PTF1-beta	347	Zic1
48	C/EBPalpha	123	GABP	198	MAF	273	PUL1	348	Zic2
49	C/EBPbeta	124	GATA	199	Max	274	R	349	Zic3
50	C/EBPdelta	125	GATA-1	200	MAZ	275	Rb-E2F-1:DP-1	350	ZID
51	C/EBPgamma	126	GATA-2	201	MAZR	276	Retroviral	351	Zta
52	cap	127	GATA-3	202	MEF-2	277	REF		
53	Cart-1	128	GATA-4	203	MEF-3	278	REF1		
54	CBF	129	GATA-6	204	MEIS1	279	Koaz		
55	CCAAT	130	GATA-X	205	MIF-1	280	RORalpha1		
56	Cdx5	131	GC	206	MOVO-B	281	RORalpha2		
57	CDP	132	GCM	207	MRF-2	282	RP58		
58	CDX	133	GCMF	208	Mx-1	283	RREB-1		
59	Cdx-2	134	GEN_INI	209	MTF-1	284	RSRFC4		
60	CdxA	135	Gfi-1	210	Muscle	285	S8		
61	CHOP/C/EBPalph	136	GFI1	211	MYB	286	SEF-1		
62	Churchill	137	GFI1B	212	Myc	287	SF-1		
63	CHX10	138	GLI	213	MvxD	288	SMAD		
64	CI2	139	GR	214	myogenin	289	SMAD-3		
65	Clox	140	GZF1	215	MZF1	290	SMAD-4		
66	COMP1	141	Hand1:E47	216	N-Myc	291	SOX		
67	COUP	142	HEB	217	Ncx	292	Sox-5		
68	COUP-TF	143	Helios	218	NERF1a	293	SOX-9		
69	COUPTF	144	HEN1	219	neural-restr.-silencer-element	294	Sp1		
70	CP2	145	HES1	220	NF-1	295	Sp3		
71	CP2/LBP-1c/LSF	146	HFH-1	221	NF-AT	296	Spz1		
72	CRE-BP1	147	HFH-3	222	NF-E2	297	SREBP		
73	CREB	148	HFH-4	223	NF-kappaB	298	SREBP-1		
74	CREBATF	149	HFH-8	224	NF-muE1	299	SRE		
75	Crx	150	HIC1	225	NF-Y	300	SRY		

Table 43 List of 584 Matrix Names

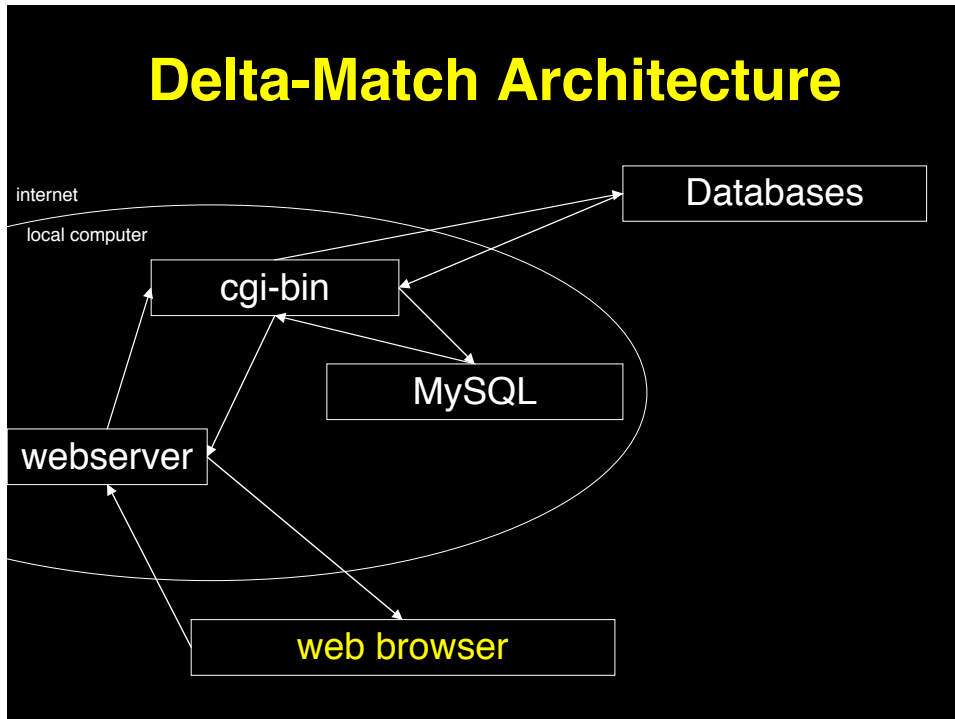
(see file "550_matrixes.txt" at the **Delta-MATCH > Downloads** web page)

http://dingo.ucsf.edu/~dwilliamson/scripts/delta_match/acc_files/550_matrixes.txt

Table 44 Distribution of Polymorphisms in the human genome (hg18.snp126)

Polmorphism Position	Count
total in hg18.snp126	11647909
10kb_up	647311
10kb_down	648916
5'UTR	16376
3'UTR	84503
exons	212764
introns	3415853
conserved	397802
cpgislands	88432
regpotential	5356000
insertion/deletions	2204226
simplerepeats	571302
repeatmasker	5280806
microsatellite	46995
nonbinary	72334
mapped to 2 or more positions	305668

Figure 204 Architectural Diagram for the Delta-MATCH Query Tool (DMQT)



1.24.1 Delta-MATCH MYSQL Databases and Tables

If you want to see the details of the embedded Delta-MATCH XML tags, download the following file:

DM_MYSQL_databases.pdf

1.24.2 The Delta-MATCH XML DTD

If you want to see the details of the embedded Delta-MATCH XML tags, download the following file:

http://dingo.ucsf.edu/~dwilliamson/scripts/delta_match/dm_result.dtd

1.24.3 List of Delta-MATCH Errors

The Delta-MATCH Query Tool will return a number of error messages to the browser.

Errors are returned in the order they are found.

- Error 1 - no matrixes passed your selected criteria
- Error 2 - more than 1,500 rsnumbers passed your selected criteria
- Error 3 - no rsnumbers were found that passed your selected criteria
- Error 4 - no rsnumbers were found in the select gene names
- Error 5 - could not connect to database
- Error 6 - more than 5 gene names were submitted
- Error 7 - no gene names were found
- Error 8 - rsnumber file was not uploaded properly
- Error 9 - no premod modules were found

Figure 205 Delta-MATCH Error 1 - no matrixes passed your selected criteria

This error states “no matrixes passed your combined selected criteria”. In example 4, there was an internal conflict between the matrix name selected in STEP 1 (V\$NFKB_Q6) and the “Matrix Quality” type (*qual* = “low”) because V\$NFKB_Q6 is actually a “high quality” matrix. The quality of this matrix name (mat_id) can be verified by viewing the “550_matrixes.txt” file and noting the number “1” in the column “quality” next to the V\$NFKB_Q6 mat_id. The warning suggests trying to adjust a number of parameters that will eliminate the conflict including changing or unchecking the matrix quality box, or changing the matrix length sub-selected under “Show Matrix Details” box.

ERROR 1- no matrixes passed your selected criteria

Return to the input page and try changing the following parameters

- primary matrix selection (1-4)
- minimum potential score (potential)
- maximum returned rnumbers value
- matrix quality (qual)
- matrix length (mat_len)

You may want to start by unchecking the 'Quality' checkbox

[Go Back](#) to the input page

[e-mail](#) Delta-MATCH your questions and comments

Figure 206 Delta-MATCH Error 2- more than 1,500 rnumbers passed your selected criteria

A maximum of 1,500 rnumbers will be returned per query. If you receive have reached the 1,500 limit, you will receive a warning and may want to resubmit your query using a more stringent set of criteria. The matrix names (mat_id) are searched in alphabetical order. Therefore if you have receive a warning after submitting a search that included more than one matrix, it is likely your results do not include all of the important

Delta-MATCH predictions for every matrix submitted. You may also consider breaking your job down into smaller tasks.

ERROR 2 - more than 1500 rsnnumbers passed your selected criteria

Warning!!! Your returned results HAVE BEEN TRUNCATED

The combined parameters you have selected may not be stringent enough

Only the first 1500 results have been returned in alphabetical order of the list of matrix names that passed your specified criteria

You may have overlooked some important results in this query

Please consider going back to the input page and resubmit your query with a more stringent set of parameters. You may try to:

- Selecting fewer matrixes, during the primary matrix selection (STEP 1)
- Reduce the number of Gene Names being searched
- Reduce the number of rsnnumbers in you upload file
- Increase the minimum potential score cutoff a couple of decimal points (keep Min Potential Score >= 0.8)
- Check the Matrix Quality box with the quality set to high
- Decrease the number of results returned per matrix (Max Returned rsnnumbers)

Figure 207 Delta-MATCH Error 3 - no rsnnumbers were found that passed your selected criteria

If no rsnnumbers passed the cumulative selected criteria you will receive Error. Consider decreasing the stringency of the query and try again. Error 3 is returned in Example 6 because there were no biologically relevant polymorphisms identified for the specified TFBS matrix [V\$ACAAT_B (0)].

ERROR 3 - no rsnnumbers were found that passed your selected criteria

The combined parameters you have selected may be too stringent, or there were no significant Delta-MATCH predictions for this set of matrixes

Please consider the following to return more rsnnumbers

- Verify the selected matrixes have results in there database tables by viewing the number next of the matrix name in drop down menu of STEP 1 -1 [e.g. V\$ACAAT_B (0)]
- Decrease the minimum potential score cutoff a couple of decimal points (keep Min Potential Score >= 0.3)
- If you have resubmitted this search using a 'log' file, be sure the original search didn't include any uploaded text files.
- Uncheck the Matrix Quality box (Quality = wither high or low)
- Increase the number of results returned per matrix (Max Returned rsnnumbers)

If the above steps fail you may then try

- Selecting more matrixes, during the primary matrix selection (STEP 1)

[Go Back](#) to the input page

[e-mail](#) Delta-MATCH your questions and comments

Figure 208 Delta-MATCH Error 4- no rsnumbers were found in the select gene names

Error 4 may be returned when no rsnumbers are found because either the gene names were not found, or if there were truly no polymorphisms found in the associated gene windows prior to testing any of the other input criteria.

ERROR 4 - no rsnumbers were found in the select gene names

The reason no rsnumbers were found include:

- There may actually be no rsnumbers located at the loci for the specified genes (Are the specified genes short in length?)
- There may have been a field name mismatch for the specified database table
- You may have mistyped the appropriate gene names
- The gene names do not exist or exist under another accession name

You may consider downloading the help file that is in the 'Search By Gene Name' section to see some example database table name / field name combinations that work

Return to the input page and try verify the name of the genes you are choosing by selecting the 'Search for Gene Without Returning Results' checkbox

[Go Back](#) to the input page

[e-mail](#) Delta-MATCH your questions and comments

Figure 209 Delta-MATCH Error 5 - could not connect to database

You may receive Error 5 if the webserver hosting the Delta-MATCH Query Tool is unable to connect to the computer hosting the Delta-MATCH MySQL database. This error might occur during a system update or power failure. Please be patient and try again. If the problem persists, please contact the author.

Error: Could not connect to database. dm2_5_million
Error: Could not connect to database. hg18
Error: Could not connect to database. hg17
Error: Could not connect to database. go
Error: Could not connect to database. dm2_acc
Error: Could not connect to database. 500k
Error: Could not connect to database. hapmap
Error: Could not connect to database. premod
Error: Could not connect to database. affy_chips
Error: Could not connect to database. illumina
Error: Could not connect to database. chavi
Error: Could not connect to database. dgv

ERROR 5- could not connect to database

The computer holding the Delta-MATCH MySQL database is temporarily unavailable

- verify the host database computer (boxer) is powered on
- verify mysql is running on the host database computer
- verify the network is available
- verify the hard drives on the host database computer are good

Please be patient and try again

If this error persist for more that a day, please contact the Delta-MATCH administrator

[Go Back](#) to the input page

[e-mail](#) Delta-MATCH your questions and comments

Figure 210 Delta-MATCH Error 6 - more than 5 gene names were submitted

Error 6 will be returned to the browser more than the maximum allowable number of gene names have been submitted. This warning is not critical.

ERROR 6 - more than 5 gene names were submitted

Warning!!! The maximum number of gene names allowed to be submitted is 5

Only the results from the first 5 submitted gene names will be returned

[Go Back](#) to the input page

Figure 211 Delta-MATCH Error 7 - no gene names were found

There were no GENE NAMES matching the 'UCSC hg18 Table Name' 'Field Name' pair found. You may consider downloading the help file that is in the 'Search By Gene Name' section to see some example database table name / field name combinations that work.

ERROR 7 - no GENE NAMES were found

There were no GENE NAMES matching the 'UCSC hg18 Table Name' 'Field Name' pair found

lease go back and augment the parameters in the 'Search by Gene Names'

You may consider downloading the help file that is in the 'Search By Gene Name' section to see some example database table name / field name combinations that work

Please direct question and comments to deltamatch@commandcreate.org

Figure 212 Delta-MATCH Error 8 - rsnumber file was not uploaded properly

ERROR 8 - rsnumber file was not uploaded properly

The rsnumber file was not uploaded properly
name of file uploaded: 'CV_David_W_Williamson.doc'
type of file uploaded: 'application/msword'

The uploaded rsnumber file wasn't a 'text/plain' file, or wasn't formatted appropriately
([download example file](#))

The text file should contain only the list of rsnumbers, one per line
Create your file in unix with the 'vi' editor, or be sure to save file as the type 'MS-DOS Text' if using MS-Word

Please go back and try uploading another file

[Go Back](#) to the input page

[e-mail](#) Delta-MATCH your questions and comments

Figure 213 Delta-MATCH Error 9 - no premod modules were found

ERROR 9 - no premod modules were found

You submitted these PReMod terms: 'M00769,M00701,wefasaefa'

There were no PReMod modules that matching this set

Verify that submitted terms match a 'FACTOR' or 'MODULE_MATRIX' of the 'key' file
(matches must be exact and are case sensitive)

View the [PReMod key file](#)

Please go back and try a different PReMod selection

[Go Back](#) to the input page

[e-mail](#) Delta-MATCH your questions and comments

Figure 214 Delta-MATCH Graphic Motif



Figure 215 Delta-MATCH Resources (Graphics)

Bioinformatics
at the University of California San Francisco

php **MySQL** **python™**

The Apache Software Foundation
<http://www.apache.org/>

BIOBASE **TRANSFAC** **fedora^f**
BIOLOGICAL DATABASES

BIOCONDUCTOR **R** **X**
open source software for bioinformatics

THE J. DAVID GLADSTONE INSTITUTES
GLADSTONE INSTITUTE OF CARDIOVASCULAR DISEASE
AT THE UNIVERSITY OF CALIFORNIA SAN FRANCISCO

AIM 2 Extras (A Genetic Survey)

Figure 216 Haploview Linkage Disequilibrium Legend

(http://www.broad.mit.edu/mpg/haploview/haploview_doc.pdf excerpts, page 3)

Table 1.1. Standard Color Scheme

	$D' < 1$	$D' = 1$
LOD < 2	white	blue
LOD # 2	shades of pink/red	bright red

Table 1.3. r^2 Color Scheme

$r^2 = 0$	white
$0 < r^2 < 1$	shades of grey
$r^2 = 1$	black

Figure 217 The DNA Degenerate Alphabet

The DNA degenerate alphabet							
A	Adenosine	R = A or G	puRine	B = C, G or T	not A		
C	Cytidine	Y = C or T	pyrimidine	D = A, G or T	not C		
G	Guanosine			H = A, C or T	not G		
T	Thymidine	N = A, C, G or T	aNy	V = A, C or G	not T		
		K = G or T	Keto (in large groove)	S = G or C	Strong (3 H bonds)		
		M = A or C	aMino (in large groove)	W = A or T	Weak (2 H bonds)		
complement of :		A C G T	R Y	K M	S W	B D H V	N
is :		T G C A	Y R	M K	S W	V H D B	N

1.25 Other Software by David W. Williamson

1.25.1 What Color Eyes Would Your Children Have? (flash)

Flash version (hosted at TheTech)

<http://museum.thetech.org/ugenetics/eyeCalc/eyecalculator.html>

1.25.2 What Color Eyes Would Your Children Have? (html)

Simple html Version

http://127.0.0.1/~david/scripts/Eye_Calculator_Radio/Eye_Calculator_Radio.html

1.25.3 SNP Enzyme Finder

http://127.0.0.1/~david/scripts/SNP_Enzyme_Finder/SNP_Enzyme_Finder.html

1.25.4 Haplotype Mapper

http://127.0.0.1/~david/scripts/Haplotype_Mapper/Haplotype_Mapper.html

Figure 218 David W. Williamson's Contact and Business Card

THE J. DAVID GLADSTONE INSTITUTES
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO



**GLADSTONE INSTITUTE OF
CARDIOVASCULAR DISEASE**

David W. Williamson
Research Associate
Bioinformaticist
Doctoral Candidate
dwilliamson@gladstone.ucsf.edu
www.gladstone.ucsf.edu

1650 Owens Street
San Francisco, CA 94158

Telephone (415) 734-4945
Facsimile (415) 355-0855

1.26 Ph.D. Thesis Defense (February 06, 2008)

1.26.1 Seminar Announcement

seminar

Ph.D. Thesis Seminar

**Genetic and Bioinformatic Approaches
to Identify Polymorphic Modulators
of Transcription Factor Binding and
Disease Phenotypes Including
HIV-1 Viremia**

David Williamson
UCSF Ph.D. Graduate Program in
Biological and Medical Informatics
Mahley Lab

DATE: Wednesday, February 6, 2008
TIME: 3:00 – 4:00 p.m.
LOCATION: Robert W. Mahley Auditorium
Gladstone Institutes
1650 Owens Street, 1st Floor
HOST/CONTACT: Robert W. Mahley, M.D., Ph.D.
415-734-2062


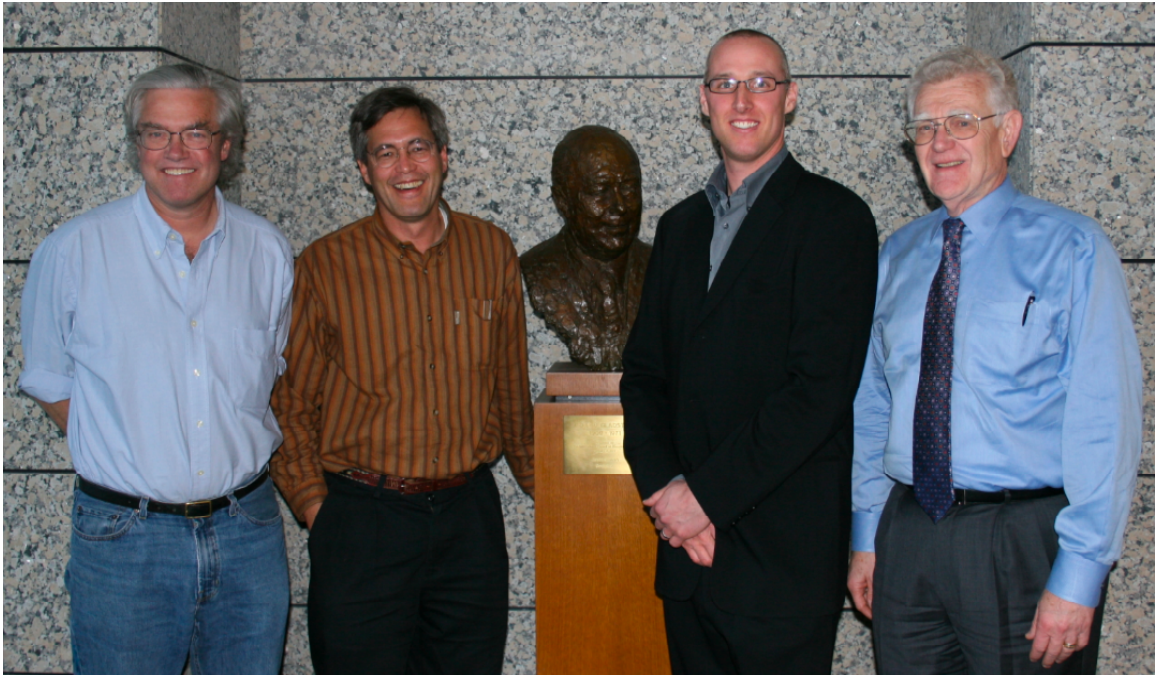
 **GLADSTONE INSTITUTE OF
CARDIOVASCULAR DISEASE**

Figure 219 Joseph “Mike” McCune, Bruce Conklin, David Williamson, Robert Mahley



February 06, 2008

19 UCSF Library Release

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses and dissertations. Copies of all UCSF theses and dissertations will be routed to the library via the Graduate Division. The library will make all theses and dissertations accessible to the public and will preserve these to the best of their abilities, in perpetuity.

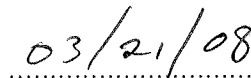
Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis or dissertation to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

.....

Author Signature

David Wayne Williamson

.....

Date