Title

Characterization of Complex Genetic Component Contributing to the Susceptibility for Multiple Sclerosis and Rheumatoid Arthritis

Author

Briggs, Farren Basil Shaw

Publication Date

2010

Peer reviewed|Thesis/dissertation

Characterization of Complex Genetic Component Contributing to the Susceptibility for Multiple Sclerosis and Rheumatoid Arthritis

By

Farren Basil Shaw Briggs

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Epidemiology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Lisa F. Barcellos, Chair
Professor Alan E. Hubbard
Professor Martyn T. Smith
Professor Glenys Thomson

Spring 2010

Abstract

Characterization of Complex Genetic Component Contributing to the Susceptibility for Multiple Sclerosis and Rheumatoid Arthritis

By

Farren Basil Shaw Briggs

Doctor of Philosophy in Epidemiology

University of California, Berkeley

Professor Lisa Barcellos, Chair

Autoimmune diseases (ADs) are a major public health concern, as the third most common category of disease in the US, following cancer and heart disease. As a result, autoimmunity has become one of the most active genetic and epidemiologic research areas. However, unraveling the etiological mechanisms in ADs has proven difficult. There is strong evidence suggesting a complex genetic component contributes to all ADs. For most ADs, the prominent genetic risk locus is within the major histocompatibility complex (MHC) on chromosome 6p21.3. Unfortunately, identifying non-MHC susceptibility loci has proven difficult in studies of ADs, for which multigenic patterns of inheritance are observed. Recently, through concerted international efforts, several genome-wide association (GWA) studies and subsequent replication analyses have confirmed many new susceptibility loci for ADs with very modest effects. The remainder of genetic variants contributing to AD susceptibility are unknown. It is clear that current approaches will be limited to completely characterize the complex genetic component in ADs. This dissertation is focused on the use of strong epidemiological approaches and novel analytical strategies to identify additional non-MHC genetic risk factors for two complex ADs: multiple sclerosis (MS) and rheumatoid arthritis (RA).

In Chapter 1, the relationship between variation in DNA repair pathways genes and risk for MS was investigated. Univariate association testing, epistatic tests of interactions, logistic regression modeling and non-parametric Random Forests analyses were performed using genotypes from 1,343 MS cases and 1,379 healthy controls of European ancestry. A total of 485 single nucleotide polymorphisms (SNPs) within 72 genes related to DNA repair pathways, including base excision repair, nucleotide excision repair, and double strand breaks repair, were investigated. A SNP variant within *GTF2H4* on 6p21.33 was significantly associated with MS (odds ratio=0.7, P=3.5 x 10^{-5}) after accounting for multiple testing, and was not due to linkage disequilibrium with the well-established *HLA-DRB1*1501* allele within the MHC. Despite clear evidence for an association between *GTF2H4* and MS, collectively, these results, derived from a well-powered study, do not support a strong role for common variation within DNA repair pathway genes in MS.

In Chapter 2, the relationship between variation within 8 candidate hypothalamic-pituitary-adrenal (HPA) axis genes and susceptibility to MS was comprehensively investigated. A total of 326 SNPs were investigated in 1,343 MS cases and 1,379 healthy controls of European ancestry using a multi-analytical strategy. Random Forests, a supervised machine learning algorithm, identified 8 SNPs within the corticotropin releasing hormone receptor 1 or *CRHR1* locus on 17q21.31 as important predictors of MS. Based on univariate analyses, five *CRHR1* variants were associated with decreased risk for disease following a conservative correction for multiple tests. Independent replication was observed in a large meta-analysis comprised of 2,624 MS cases and 7,220 healthy controls of European ancestry. Collectively, results provide strong evidence for the involvement of *CRHR1* in MS.

In Chapter 3, epistatic interactions with a well-established genetic risk factor (*PTPN22* 1858T) in RA were investigated. The analysis consisted of four stages: Stage I (data reduction) – identifying candidate chromosomal regions in 292 affected RA sibling pairs, by predicting *PTPN22* concordance using multipoint identity-by-descent probabilities and Random Forests; Stage II (extension analysis) – testing detailed genetic data within candidate chromosomal regions for epistasis with *PTPN22* 1858T in 677 RA cases and 750 controls using logistic regression; Stage III (replication analysis) – confirmation of epistatic interactions in 947 RA cases and 1,756 controls; Stage IV (combined analysis) – a pooled analysis including all 1,624 RA cases and 2,506 control subjects for final estimates of effect size. A total of 7 epistatic interactions identified in Stage II were replicated in Stage III. A SNP variant (rs7200573) within *CDH13* demonstrated the strongest evidence for interaction (p=1.5 x 10^{-4}) with *PTPN22*. There was also evidence for epistasis between *PTPN22* and SNP variants within *MYO3A*, *CEP72* and near *WFDC1*.

The research conducted in Chapters 1 through 3 describes the use of analytical approaches based on strong hypotheses, multi-stage analyses, and the use of robust non-parametric methods in tandem with conventional association testing. Results described in these chapters are scientifically important, as they contribute to our understanding of the underlying genetic architecture in two very debilitating ADs (MS and RA), and also provide methodological frameworks for investigating other chronic diseases with a complex genetic component.

To my wonderful family:
My beautiful and caring mother, Aloma;
My supportive father, John;
And my sister and equal, Venetia.

TABLE OF CONTENTS

INTRODUCTION

Autoimmune diseases (ADs) are a major public health concern and one of the most active genetic and epidemiologic research areas. ADs are the third most common disease category in the United States; approximately 14.7-23.5 million Americans (5–8% of the U.S. population) are afflicted with at least one of eighty recognized ADs, and prevalence is increasing (1). Furthermore, women are disproportionately affected by ADs, there are no curative therapies and severe lifetime economic burdens on familial and societal structures result for those with disease (2, 3). In general, ADs originate from a breakdown in one or more of the several basic mechanisms involved in immune tolerance, which allows immune cells to distinguish between self and non-self structures. This failure results in the individual's immune system becoming auto-reactive and capable of damaging self cells, tissues and/or organs. Despite increasing efforts, unraveling the etiological mechanisms of ADs has proven difficult. A role for both genes and environment is strongly supported by multiple areas of research. This dissertation is focused on characterizing the genetic component in two complex ADs: multiple sclerosis and rheumatoid arthritis.

Multiple sclerosis (MS [MIM 126200]) is a complex demyelinating autoimmune disease of the central nervous system, with two distinct but overlapping neuro-pathological phases, inflammatory and neurodegenerative, resulting in the accrual of various neurological deficits (4). MS is primarily prevalent in temperate regions, with an approximate prevalence of 0.1% in the U.S. The peak age of onset is age 20 to 40 years; however onset can occur at any age. Women are twice as likely to develop MS as men; however men are more likely to have a more severe disease progression. The clinical progression of MS is heterogeneous, unpredictable and complex, and includes symptomatic disturbances in visual, motor and sensory systems, affecting cognition, coordination, balance and bowel/bladder/sexual functions (5). The varied disease manifestations result from the development of plaques/lesions, primarily in the white matter of the brain and spinal cord, and/or loss of myelin sheath on oligodendrocytes, which interrupts signal transduction in the CNS. There are, however, distinct disease courses: relapsing remitting (RR) (including secondary progressive (SP)), primary progressive (PP), and rare subtypes such as progressive relapsing MS. At onset, 85% of patients have RRMS, characterized by episodic attacks (relapses) followed by periods of partial or total recovery (remissions); and 40% of these patients will develop SPMS, characterized by progressive disability. Nearly 15% of patients present with PPMS at onset, having progressive disability without remission, and are normally older with equal risk in males and females. Full knowledge of specific pathological mechanisms that distinguish phenotypic variation in MS remains unknown. To date, only a few factors that contribute to the susceptibility and phenotypic heterogeneity of MS have been identified, including a strong genetic component (4), as well as several promising environmental risk factors (6-8), including tobacco smoke (9-14).

Rheumatoid arthritis (RA [MIM 180300]) is a common chronic multi-system autoimmune disease, resulting from persistent inflammatory synovitis and subsequent erosion of the joint architecture (polyarthritis) (15). The prevalence of RA varies across populations, however an estimated prevalence of 1% is reported in Caucasians of North

America and Europe (15). Women are two to three times more likely to develop RA than men, with age of onset peaking in the fifth decade of life (16). The clinical manifestation of RA is heterogeneous, where some affected individuals have relatively benign presentation versus others with severe physical disability, due to progressive bone erosion and comorbidity with coronary artery disease, infection and lymphoma (17). A substantial portion of RA patients (~60%) have autoantibody responses, primarily rheumatoid factor and the highly specific cyclic citrullinated peptide antibodies (18). The relationship between autoantibody presence and RA has not been completely resolved. Several factors confer susceptibility for RA, including a prominent genetic component (15), and several environmental risk factors (16). The most recognized and preventable environmental risk factor is cigarette smoking, with an attributable risk of approximately 20% (19).

There is strong evidence suggesting a complex genetic component contributes to both MS and RA. Twin studies from different populations have consistently demonstrated a higher disease concordance in monozygotic twins compared to dizygotic twins in both MS (approximately 30% and 5%, respectively) and RA (approximately 12% and 3%, respectively), with an estimated genetic heritability of >60% for both diseases in Northern European populations (20-22). As with most ADs, the prominent genetic risk locus for MS and RA is within the major histocompatibility complex (MHC) on chromosome 6p21.3 (23). *HLA-DRB1*, a human leukocyte antigen (HLA) class II gene within the MHC, is the primary susceptibility locus conferring approximately 50% and 37% of the genetic risk in MS and RA, respectively (4, 24). Susceptibility to MS has been primarily associated with the *HLA-DR2* or *HLA-DRB1*15* haplotype (*DQB1*0602*, *DQA1*0102*, *DRB1*1501*, *DRB5*0101*) (4). Haplotype analyses in admixed African Americans have demonstrated that the *HLA-DRB1*15* allele is the primary susceptibility locus for MS (25, 26). Recent efforts have further demonstrated the association with *HLA-DRB1* is complex and other independent HLA and non-HLA susceptibility loci within the MHC exist (27, 28). In RA several *HLA-DRB1* alleles confer susceptibility. These alleles share a common sequence of amino acid residues in the antigen-recognition portion of the HLA molecule (29). These alleles have been termed the shared-epitope (SE; *HLA-DRB1* alleles: 0101, 0102, 0401, 0404, 0405, 0408, 0413, 1001, and 1402), and may contribute to anti-citrulline antibody production (30, 31). As with MS, the association with the MHC in RA is complex, and there is evidence supporting the role of multiple MHC susceptibility loci (31, 32).

The identification of non-MHC susceptibility loci has proven difficult in these complex ADs. A multigenic pattern of inheritance in MS and RA was first demonstrated by the limited success of family-based linkage studies. A genome-wide linkage analysis is a hypothesis free scan of the genome, aimed at identifying chromosomal loci that cosegregate with disease in related individuals from large pedigrees (model-based), or exhibit evidence of excess sharing in affected sibpairs or other relative pairs (model-free). Many genome-wide linkage screens have been conducted in both MS and RA, and for both diseases, *HLA-DRB1* was consistently identified as the key susceptibility locus (4, 15). Similarly, in both ADs, there was limited overlap in linkage signals observed at other chromosomal regions, and the possible susceptibility loci in these diverse regions have

not yet been identified (33-35). These initial genome-wide studies underscore a complex polygenic pattern of inheritance contributing to disease susceptibility of MS and RA. Recently, through concerted international efforts, several large scale genome-wide association (GWA) studies (described below) and subsequent replication analyses have now confirmed several novel non-MHC susceptibility loci, in both MS and RA; however only a very modest proportion of the hereditary risk has been explained by these findings (23).

GWA studies attempt to identify relatively common single nucleotide polymorphisms (SNPs; minor allele frequency [MAF] >1-5%) in linkage disequilibrium (LD) with causal susceptibility loci in a complex disease. GWA studies assume the common-disease-common-variant (CDCV) hypothesis; that common disease susceptibility is a result of the joint action of several common variants with relatively small to moderate effects, and that a significant proportion of disease alleles is shared among unrelated affected individuals (36). GWA studies are an attractive approach for investigating the genetic basis of complex diseases. Like the previous linkage studies, they are hypothesis free and unconstrained by *a priori* assumptions. The success of GWA studies, however, is dependent on several factors (37, 38). These include the availability of sufficiently large study samples from clearly defined study populations capable of contributing relevant genetic information regarding the research question. In addition, informative SNPs that capture extensive genetic variation across the whole genome and that can be inexpensively and efficiently genotyped must be used. Currently up to 1,000,000 common SNPs are available on various commercial platforms. Finally, the success of GWA studies relied heavily on the application of robust statistical methods that are capable of identifying causal genetic associations despite the dimensionality conflict of investigating excessively large number of variables. The conventional test statistic has been a single-point, one degree of freedom test of association, where significance is defined as $p<5 \times 10^{-8}$. Thus far, GWA studies have begun to unravel the underlying biological mechanisms involved in MS and RA, through identification of common variants with detectable odds ratios [ORs] of 1.1 to 1.5. Furthermore, GWA studies have provided substantive information regarding the extent of the genetic contributions of common variants to susceptibility, overlap across particular diseases, and that current approaches will be limited in ability to identify the entire genetic contribution for most complex diseases (39).

Few have started to explore the missing heritability of complex diseases unexplained by initial GWA findings. Amongst the many explanations are: (1) a much larger number of disease variants with smaller effects are yet to be found, due to low power of recent GWA studies; (2) rare and other structural (i.e. copy number variants, copy neutral variation, and epigenetic variation) variants poorly detected by current platforms are involved; (3) limited power has been present to investigate *gene* x *gene* (epistasis) and higher order multigenic interactions; and (4) a lack of accounting for genetic heterogeneity in the context of environmental exposures (40). GWA studies will continue to be a substantial tool in uncovering additional genetic risk factors, particularly with the development of more comprehensive genotyping platforms and the acquisition of substantially larger study populations. For example, approximately 60,000 individuals are

required to provide sufficient power to identify the majority of variants with OR=1.1 (39). Nevertheless, it is apparent that current approaches, primarily a marginal test for association, for investigating genetic susceptibility are not able to identify a substantial fraction of the genetic burden. Currently there is no consensus regarding appropriate methods for evaluating the genetic component of complex diseases, however, there are strong epidemiologic approaches and alternative analytical strategies that complement GWA studies and can be used to identify susceptibility loci.

One such alternative explored in this dissertation is the application of a judicious hypothesis-driven candidate gene association (CGA) study (41). CGA studies are a powerful and cost-efficient approach that allows for targeted investigation of selected alleles within candidate genes. These studies provide inferential advantages relative to untargeted screening strategies in GWA studies and allow for detecting associations that would otherwise not meet genome-level criteria ($p<5 \times 10^{-8}$). Through rapid advances in molecular research, candidate genes can be more confidently assigned to functional pathways. Consequently, formulating hypotheses at the levels of pathways, including all relevant genes as candidates, may be a more efficient epidemiologic approach allowing that allows for global conclusions about the relationship between a biological pathway and disease (41). However, a successful CGA study requires: 1) adequately powered data sets; 2) strong hypotheses supported from prior research; 3) dense coverage of genetic variation with candidate genes; and 4) comprehensive investigation of a candidate pathway, including analytical approaches that can consider variation within several related genes simultaneously.

In GWA studies, hundreds of thousands of SNPs are tested for association, but only those reaching conservative genome-wide significance ($p<5 \times 10^{-8}$) are generally investigated in validation analyses. Attempts to disentangle true associations ($p>5 \times 10^{-8}$) from statistical noise have posed a tremendous challenge, but have led to the implementation of multi-stage studies and diverse methodological approaches, beyond single-locus association tests, to extract additional information from GWA studies. Both of these themes are incorporated in this dissertation as objective constructs allowing for robust conclusions. Multi-stage analyses in genetic epidemiology are conventionally a two-stage approach where results from the first stage are validated in the second. Furthermore, multi-stage analyses can incorporate additional analytical layers, allowing for hypothesis generation, data reduction (prioritization), and model specification; and have been used in this dissertation. The use for various methodologies can be similarly rationalized, and incorporated in multi-stage analyses. A primary analytical concern in genetic epidemiological research, is underestimating the genetic contribution of a given locus to disease, since the conventional univariate approach is not able to detect complex inheritance patterns (42). As a result, various approaches have been developed to consider a single locus in the context of all other loci (i.e. allowing for the presence of interactions), and may be more appropriately suited for investigating complex multigenic diseases (43). One such methodological alternative explored in this dissertation, is a non-parametric supervised machine learning algorithm, Random Forests (44) which has been shown to be considerably more efficient than univariate methods (45). In this dissertation, multi-stage analyses using both conventional and non-parametric methods are explored.

As previously mentioned, genetic epidemiologists are constrained by the available tools for investigating complex genetic diseases. Using non-parametric methods and investigating hypotheses at the level of a biological pathway may be more efficient approaches to identifying genetic factors contributing to complex diseases. There is clear evidence supporting a multigenic etiology in MS and RA, and therefore it is likely that risk variants act in concert to influence susceptibility. There is no consensus regarding appropriate approaches for investigating *gene* x *gene* and higher-order interactions, however combining several methods may be optimal (46, 47). Furthermore, comprehensive investigation at the level of a functional pathway is also necessary, as pathway-based analyses may amplify the effects of individual polymorphisms (each with a modest effect). Multigenic approaches addressing these limitations have been prudently explored in this dissertation, as it is likely that statistical modeling of interactions may not correspond to a true biological interaction (48).

In Chapter 1, a multi-analytic CGA investigation of common genetic variation in DNA repair pathways and susceptibility for MS was pursued. Considering the limitations of current methodologies, as discussed, a comprehensive parallel analysis was conducted to investigate the detailed genetic information in 72 DNA repair related genes using: 1) conventional association testing, and 2) non-parametric methods. In addition to marginal tests for association, tests for *gene* x *gene* interactions at the level of specific DNA repair pathways, and a multigenic analysis approach were used to investigate the impact of functional variants on MS risk. A non-parametric approach utilized Random Forests and Classification and Regression Tree algorithms to investigate the relationship between genetic variation and prediction of MS. Results obtained using both approaches are compared and reported.

Chapter 2 describes a multi-stage investigation of common genetic variation in genes involved in the hypothalamus-pituitary-adrenal (HPA) axis and susceptibility for MS. In this chapter, a CGA study was conducted; however, Random Forests was used as a means of data reduction, by identifying genetic variants that contributed most to prediction of MS status. These variants were subsequently investigated using logistic regression in the CGA study. A replication analysis was then conducted to confirm initial findings using a large independent data set. Additional analyses were conducted, including comparisons of extended haplotypes in MS cases and controls, to further refine the association signal observed.

Lastly, in Chapter 3, *gene* x *gene* investigation (epistasis) was explicitly investigated using a multi-stage and multi-analytical approach in three independent RA data sets. Specifically, epistatic relationships with *PTPN22*, the primary non-MHC risk locus in RA, were explored. First, Random Forests was used to identify candidate genomic regions. These regions were then investigated using detailed genetic information from a GWA study; data from a second GWA study were used to conduct replication studies.

In conclusion, given the current limitations of GWA studies for resolving the genetic etiology in complex diseases, Chapters 1 through 3 describe analytical approaches based on strong hypotheses, consisting of multi-stage analyses and application of robust non-

parametric methods in tandem with conventional association testing methods. Results in these chapters are scientifically important, as they contribute to our understanding of the underlying genetic architecture in two very debilitating ADs (MS and RA), and also provide strong methodological frameworks for investigating the complex genetic component in other diseases.

REFERENCES

1. National Institutes of Health Autoimmune Disease Coordinating Committee Report. Bethesda (MD): The Institutes, 2002.
2. Fairweather D, Rose NR. Women and autoimmune diseases. Emerg Infect Dis 2004;10:2005-11.
3. Jacobson DL, Gange SJ, Rose NR, Graham NM. Epidemiology and estimated population burden of selected autoimmune diseases in the United States. Clin Immunol Immunopathol 1997;84:223-43.
4. Oksenberg JR, Barcellos LF. Multiple sclerosis genetics: leaving no stone unturned. Genes Immun 2005;6:375-87.
5. Zuvich RL, McCauley JL, Pericak-Vance MA, Haines JL. Genetics and pathogenesis of multiple sclerosis. Semin Immunol 2009;21:328-33.
6. Marrie RA. Environmental risk factors in multiple sclerosis aetiology. Lancet Neurol 2004;3:709-18.
7. Ascherio A, Munger KL. Environmental risk factors for multiple sclerosis. Part II: Noninfectious factors. Ann Neurol 2007;61:504-13.
8. Ascherio A, Munger KL. Environmental risk factors for multiple sclerosis. Part I: the role of infection. Ann Neurol 2007;61:288-99.
9. Antonovsky A, Leibowitz U, Smith HA, et al. Epidemiologic Study of Multiple Sclerosis in Israel. I. an Overall Review of Methods and Findings. Arch Neurol 1965;13:183-93.
10. Villard-Mackintosh L, Vessey MP. Oral contraceptives and reproductive factors in multiple sclerosis incidence. Contraception 1993;47:161-8.
11. Thorogood M, Hannaford PC. The influence of oral contraceptives on the risk of multiple sclerosis. Br J Obstet Gynaecol 1998;105:1296-9.
12. Hernan MA, Olek MJ, Ascherio A. Cigarette smoking and incidence of multiple sclerosis. Am J Epidemiol 2001;154:69-74.
13. Riise T, Nortvedt MW, Ascherio A. Smoking is a risk factor for multiple sclerosis. Neurology 2003;61:1122-4.
14. Nortvedt MW, Riise T, Maeland JG. Multiple sclerosis and lifestyle factors: the Hordaland Health Study. Neurol Sci 2005;26:334-9.
15. Silman AJ, Pearson JE. Epidemiology and genetics of rheumatoid arthritis. Arthritis Res 2002;4 Suppl 3:S265-72.
16. Alamanos Y, Drosos AA. Epidemiology of adult rheumatoid arthritis. Autoimmun Rev 2005;4:130-6.
17. Callahan LF, Pincus T. Mortality in the rheumatic diseases. Arthritis Care Res 1995;8:229-41.
18. Schellekens GA, Visser H, de Jong BA, et al. The diagnostic properties of rheumatoid arthritis antibodies recognizing a cyclic citrullinated peptide. Arthritis Rheum 2000;43:155-63.
19. Criswell LA, Merlino LA, Cerhan JR, et al. Cigarette smoking and the risk of rheumatoid arthritis among postmenopausal women: results from the Iowa Women's Health Study. Am J Med 2002;112:465-71.
20. Compston A, Coles A. Multiple sclerosis. Lancet 2002;359:1221-31.

21. MacGregor AJ, Snieder H, Rigby AS, et al. Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. Arthritis Rheum 2000;43:30-7.
22. Hawkes CH, Macgregor AJ. Twin studies and the heritability of MS: a conclusion. Mult Scler 2009;15:661-7.
23. Invernizzi P, Gershwin ME. The genetics of human autoimmune disease. J Autoimmun 2009;33:290-9.
24. Deighton CM, Walker DJ, Griffiths ID, Roberts DF. The contribution of HLA to rheumatoid arthritis. Clin Genet 1989;36:178-82.
25. Oksenberg JR, Barcellos LF, Cree BA, et al. Mapping multiple sclerosis susceptibility to the HLA-DR locus in African Americans. Am J Hum Genet 2004;74:160-7.
26. Caillier SJ, Briggs F, Cree BA, et al. Uncoupling the roles of HLA-DRB1 and HLA-DRB5 genes in multiple sclerosis. J Immunol 2008;181:5473-80.
27. Barcellos LF, Sawcer S, Ramsay PP, et al. Heterogeneity at the HLA-DRB1 locus and risk for multiple sclerosis. Hum Mol Genet 2006;15:2813-24.
28. Yeo TW, De Jager PL, Gregory SG, et al. A second major histocompatibility complex susceptibility locus for multiple sclerosis. Ann Neurol 2007;61:228-36.
29. Gregersen PK, Silver J, Winchester RJ. The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. Arthritis Rheum 1987;30:1205-13.
30. van Gaalen FA, van Aken J, Huizinga TW, et al. Association between HLA class II genes and autoantibodies to cyclic citrullinated peptides (CCPs) influences the severity of rheumatoid arthritis. Arthritis Rheum 2004;50:2113-21.
31. Ding B, Padyukov L, Lundstrom E, et al. Different patterns of associations with anti-citrullinated protein antibody-positive and anti-citrullinated protein antibody-negative rheumatoid arthritis in the extended major histocompatibility complex region. Arthritis Rheum 2009;60:30-8.
32. Lee HS, Lee AT, Criswell LA, et al. Several regions in the major histocompatibility complex confer risk for anti-CCP-antibody positive rheumatoid arthritis, independent of the DRB1 locus. Mol Med 2008;14:293-300.
33. A meta-analysis of genomic screens in multiple sclerosis. The Transatlantic Multiple Sclerosis Genetics Cooperative. Mult Scler 2001;7:3-11.
34. Jawaheer D, Seldin MF, Amos CI, et al. A genomewide screen in multiplex rheumatoid arthritis families suggests genetic overlap with other autoimmune diseases. Am J Hum Genet 2001;68:927-36.
35. Amos CI, Chen WV, Lee A, et al. High-density SNP analysis of 642 Caucasian families with rheumatoid arthritis identifies two new linkage regions on 11p12 and 2q33. Genes Immun 2006;7:277-86.
36. Reich DE, Lander ES. On the allelic spectrum of human disease. Trends Genet 2001;17:502-10.
37. McCarthy MI, Abecasis GR, Cardon LR, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 2008;9:356-69.

38. Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS results: A review of statistical methods and recommendations for their application. Am J Hum Genet;86:6-22.
39. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. Science 2008;322:881-8.
40. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. Nature 2009;461:747-53.
41. Jorgensen TJ, Ruczinski I, Kessing B, Smith MW, Shugart YY, Alberg AJ. Hypothesis-driven candidate gene association studies: practical design and analytical considerations. Am J Epidemiol 2009;170:986-93.
42. Moore JH. A global view of epistasis. Nat Genet 2005;37:13-4.
43. Heidema AG, Boer JM, Nagelkerke N, Mariman EC, van der AD, Feskens EJ. The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases. BMC Genet 2006;7:23.
44. Breiman L. Random Forests. Machine Learning 2001;45:5-32.
45. Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P. Screening large-scale association study data: exploiting interactions using random forests. BMC Genet 2004;5:32.
46. Musani SK, Shriner D, Liu N, et al. Detection of gene x gene interactions in genome-wide association studies of human population data. Hum Hered 2007;63:67-84.
47. Cordell HJ. Genome-wide association studies: Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet 2009.
48. Siemiatycki J, Thomas DC. Biological models and statistical interactions: an example from multistage carcinogenesis. Int J Epidemiol 1981;10:383-7.

ACKNOWLEDGEMENTS

My academic career would not have been possible without the exceptional academic support and mentorship of my advisor, Lisa Barcellos. I have appreciated her unwaivering dedication to my academic and professional advancement, and I truly look forward to working closely with her as my career develops.

I would like to thank my dissertation committee members, Alan Hubbard, Martyn Smith, and Glenys Thomson, for guiding me over the last few years.

My Berkeleyan life would not have been possible without the friendships of fellow Cal colleagues: Aubree Gordon, Darlene Francis, Miranda Ritterman, Sara Tartof, and the members of the SPH Genetics and Genomics Laboratory. I would also like to acknowledge my friendship with John Herkenhoff and his support during my graduate life, but most importantly for bringing Joci (my goldendoodle girl) into my life.

I appreciate all those individuals who, and moments that, contributed to my journey from Belmopan, to Belize City, to Worcester (particularly my Environmental History class which ignited my passion for Public Health), to Boston (especially Paola Arlotta who believed), to Providence (especially Jeffrey Blume and Rachel Morello-Frosch who encouraged), and finally to Berkeley.

I have been blest with amazing friendships with individuals who are strong, confident, successful and inspiring, and I am so humbly grateful: Kellee Garbutt, who has kept me grounded and honest since the age of twelve; Kirsten Demoga and Euriphile Joseph, who have been most loyal and present since college; and Lora Kim and Noelle Stout, who have shared their lives completely and truly, and fostered my personal growth with their generosity.

Lastly and most importantly, I am indebted to my family, and I hope I have made them proud. I am unbelievably grateful for being raised and never knowing that the *impossible* existed, that all things were *possible*, that my future was a destination and the journey was my choice. I am blest to have an awe-inspiring sister and fantastic best friend, Venetia Briggs, who blazed ahead and set an extremely high standard for my journey. I am honored to have a steadfast father, John Briggs, who adventurously defined his own path. I am utterly humbled to have a giving mother, Aloma Briggs, who loved me beyond words; she is the most important person in my life and this dissertation would have been impossible without her love, support, and words of wisdom.

Chapter 1

Variation Within DNA Repair Pathway Genes and Risk for Multiple Sclerosis.

ABSTRACT

Multiple sclerosis (MS) is a complex autoimmune disease of the central nervous system with a prominent genetic component. The primary genetic risk factor is the *HLA-DRB1*1501* allele; however, much of the remaining genetic contribution to MS has not been elucidated. Here, the relationship between variation in DNA repair pathway genes and risk for MS was investigated. Single-locus association testing, epistatic tests of interactions, logistic regression modeling and non-parametric Random Forests analyses were performed using genotypes from 1,343 MS cases and 1,379 healthy controls of European ancestry. A total of 485 single nucleotide polymorphisms (SNPs) within 72 genes related to DNA repair pathways, including base excision repair, nucleotide excision repair, and double strand breaks repair, were investigated. A SNP variant within *GTF2H4* on 6p21.33 was significantly associated with MS (odds ratio=0.7, P=3.5 x 10^{-5}) after accounting for multiple testing, and was not due to linkage disequilibrium with *HLA-DRB1*1501*. While other candidate genes examined here warrant further follow-up studies, collectively, these results, derived from a well-powered study, do not support a strong role for common variation within DNA repair pathway genes in MS.

INTRODUCTION

Multiple sclerosis (MS) is a chronic inflammatory and demyelinating autoimmune disease of the central nervous system with a substantial genetic component (1). Susceptibility to MS has been primarily associated with human leukocyte antigen (*HLA*) class II genes within the major histocompatibility complex (MHC) on chromosome 6p21; specifically the *HLA-DR2* or *HLA-DRB1*15* haplotype (*DQB1*0602*, *DQA1*0102*, *DRB1*1501*, *DRB5*0101*) (1), with recent haplotype analyses in admixed African Americans confirming *HLA-DRB1*15* as the primary susceptibility allele (2, 3).

The identification of all non-MHC genetic risk factors has proven difficult. Recent genome wide association (GWA) and replication studies have begun to identify and confirm additional genetic risk factors for MS, including *IL7RA*, *IL2RA*, *CLEC16A*, *CD58*, *TNFRSFA1*, *IRF8*, *KIF21B*, and *TMEM39A*, although their effects on risk are modest at best (4-10). A substantial component of the genetic susceptibility to MS remains unknown, underscoring a complex polygenic pattern of inheritance contributing to disease susceptibility. While GWA studies are attractive for many reasons, including their 'hypothesis free' nature, it is clear experiments using current technology will be limited in their ability to identify the entire genetic contribution for most complex diseases, including MS (11). Candidate gene studies have historically failed to identify susceptibility loci with conclusive evidence. However, revisiting candidate gene studies with: (1) adequately powered data sets, (2) hypotheses supported from prior research, and (3) analytical approaches that can consider variation within biological pathways and thus, variation within several related genes simultaneously, remains an important strategy for disease gene identification (12, 13). There are several candidate biological pathways thought to contribute to MS susceptibility. The current study focused on genetic variation within pathways related to DNA repair.

A potential role for DNA repair genes in MS is suggested by a significant increase in total DNA (nuclear [nDNA] and mitochondrial [mtDNA] DNA) damage in active MS lesions compared to normal appearing white matter of MS brains (14). Further investigation suggests damage predominantly affects mtDNA (15). Gene expression studies have shown multiple genes involved in DNA repair are differentially expressed in MS lymphocytes and lesions when compared to control tissues, including genes involved in base excision repair (BER; i.e. *PARP1*, *OGG1*, *UNG*), nucleotide excision repair (NER; i.e. *RPA1*, *ERCC5*), non-homologous end joining repair (NHEJ; i.e. *PRKDC*) and several other DNA repair mechanisms (i.e. direct reversal and mismatch repair) (16-18). There is also strong evidence supporting involvement of genotoxic agents in MS, including tobacco smoke and nitric oxide (19-21). These lines of evidence, strongly suggest a plausible biological role for DNA repair in the etiology of MS. We investigated a total of 72 genes related to (or within) four distinct DNA repair pathways: BER, NER, and repair of double-strand breaks (DSB: homologous recombination [HR] and NHEJ).

MATERIALS AND METHODS

Study population

Through the collaborative efforts of the International Multiple Sclerosis Genetics Consortium (IMSGC), the dataset was comprised of 2,961 (1,488 MS cases and 1,512 controls) participants recruited from three clinical centers (University of California, San Francisco; Harvard/MIT

Board Institute; and Cambridge University) (10). All MS cases met well-established disease criteria (22, 23). Unrelated controls were obtained from these same US sites and from the British 1958 Birth Cohort Study. These controls were selected to provide nearly equivalent gender and age distributions (10). Informed consent was obtained from all study participants and approvals from local institutional review boards were secured at each recruitment site prior to enrollment. All participants self-reported as non-Hispanic whites.

Genotyping and quality control

DNA repair genes were identified from a detailed gene inventory (Table 1) (24). Genic single nucleotide polymorphisms (SNPs) were selected for genotyping based on their function as a tagging SNP or as a SNP which may introduce a deleterious amino acid substitution affecting protein function. Tagging SNPs were identified by the Integrated Haplotype Analysis Pipeline (iHAP; an automated pipeline based on the algorithms and logical rules developed for HapBlock), with user-defined parameters (25). iHAP retrieved CEU (Utah Residents with Northern and Western European Ancestry) Haplotype Map (HapMap) data (NCBI Build 35; Release #21, July 2006) (26), contructed haplotype blocks defined as a set of SNPs (one or more) which capture 80% or more of all observed haplotypes (alpha=0.80), and identified tagging SNPs as the minimum set of SNPs which can distinguish all common haplotypes within a block (common haplotype threshold [beta]= frequency of 0.05). Deleterious SNPs were identified using the Sorting Intolerant From Tolerant (SIFT) program, which predicts how protein function may be affected by amino acid substitution by comparing sequence homology and physical properties of amino acids (27). All non-synonymous SNPs (retrieved from dbSNP July, 2007) for genes of interest, identified as having a potential damaging effect on protein function were selected for genotyping, in addition to tagging SNPs.

All individuals were genotyped for 564 SNPs, as a subset of 48,767 custom SNPs using the Illumina Infinium 60K BeadChip assay (28). A rigorous quality control protocol was utilized. Briefly, Whole-genome Association Study Pipeline (WASP) assessed sample and SNP genotyping efficiency (<95%), allele frequencies, gender errors, and Hardy-Weinberg equilibrium (HWE; <0.0001) (4). This process was performed recursively. Samples were excluded if the probability of Caucasian European descent was <0.90. Additional quality control processes were performed, including the assessment of population outliers, as previously described (10). The final quality control analysis yielded 2,722 individuals and 46,874 SNPs, of which 501 SNPs were relevant to this analysis. We excluded 16 SNP variants with a minor allele frequency <0.01. Therefore, a total of 485 SNPs (including 25 non-synonymous variants) within 72 genes related to DNA repair pathways (22 BER genes, 26 NER genes, 15 HR genes, and 9 NHEJ genes) were investigated in 1,343 MS cases and 1,379 healthy controls of European ancestry (for a complete gene list see Supplementary Table 1). The rs3135388 (A/G) SNP was used to determine the presence of the *HLA-DRB1*1501* allele as previously described (29).

Statistical analysis

We investigated the power to detect marginal and epistatic genetic associations, assuming a two-sided type 1 error of 5% (α=0.05). Results indicated our investigation was well powered (80%) to detect allelic odds ratios (ORs) $\leq$0.8 and $\geq$1.2 and epistatic ratio of odds ratios for interaction $\leq$0.6 and $\geq$1.4 for almost all models considered (*data not shown*). Therefore, all analyses were

performed and results interpreted in accordance with these established criteria. All subsequent statistical tests were two-sided.

Allele frequencies between MS cases and controls were compared, and ORs, 95% confidence intervals (95% CI), and significance criteria based on Benjamini and Hochberg procedure for controlling the False Discovery Rate (FDR-BH) were determined, using PLINK v1.06 (30). The rs3135388 SNP which is highly correlated with the *HLA-DRB1*1501* allele (4), was also included; the presence or absence of the A variant was used to classify MS cases and controls for *HLA-DRB1*1501* stratified analyses (**1501* carriers: 702 cases, 345 controls; **1501* non-carriers: 638 cases, 1,030 controls). Haplotype blocks were constructed using D' confidence intervals (31), and frequencies were compared using Haploview v4.1 (32). The conditional haplotype method (CHM) was used to distinguish primary from secondary MHC associations. The CHM tests for homogeneity of relative allele frequencies in cases and controls at a test locus on haplotypes identical for alleles at another locus (33).

We tested all two-way interactions between SNP variants within a single pathway using unconditional logistic regression as implemented in PLINK (SNPs were coded as 0, 1, and 2; where 0 represents homozygous genotype for the major allele); the test for epistasis was based on the coefficient of the interaction term (where *P*-value of the interaction term reflects the difference in the likelihood between the full model and a reduced model containing only main effects).

Furthermore, the potential biological involvement of DNA repair pathways was explored by specifically investigating the combined effect of non-synonymous SNPs (N=25) present in 18 genes (Table 6), across all pathways. We hypothesized the presence of multiple missense variants may contribute, additively, to risk for MS. The combined effect of missense variants was determined by counting the total number of missense alleles for each individual. An appropriate reference group was determined by identifying the number of missense variants that reflected the lowest quartile (≤4 missense alleles [25.5% of the study population]; similar to previous approaches (34, 35)). Wilcoxon-type test for trend (36) and unconditional logistic regression analyses, adjusted for gender and *HLA-DRB1*1501* (rs3135388) carrier status, were conducted using STATA v9.2 (College Station, TX). The combined effect score was treated as both a categorical and continuous variable to determine the risk per strata and per allele.

Non-parametric methods were also used in parallel to univariate tests of association. Random Forests, a supervised machine learning algorithm that grows recursively partitioned trees without pruning (37), and Classification And Regression Trees (CART), a decision tree that presents a hierarchical arrangement of investigated variables were utilized. Random Forests is essentially an extension of CART, however, it produces a collection of trees independently grown using bootstrap aggregating and random selection of predictors to determine classification at each node. In Random Forests, all predictors are considered simultaneously and the classification accuracy of the forest is assessed. Each predictor is randomly permuted across all trees and used to generate a variable importance (VI) score. The VI scores rank predictors by their importance in classifying the outcome in the context of all predictors without model specification, and are robust to uninformative predictors and outliers. Additionally, the VI score potentially includes the effect of multiplex interactions between the predictors, as each variable selected at a node is

essentially important conditional on the variable selected at the prior node. To assist with the interpretation of important predictors identified by Random Forests, CART was used to explore and illustrate the conditional relationship(s) amongst these predictors.

For the Random Forests analyses, all genetic variants were coded as genotypes and missing genotypes were imputed using Beagle v2.1.3 (38). All genetic variants were used to predict MS disease status using Random Forests v6.40.179 (http://salford-systems.com/) with *mtry*=sqrt(p) and *ntree*=5,000. The analysis was repeated twice: 1) excluding *HLA-DRB1*1501* (rs3135388), and 2) excluding all chromosome 6p21 variants (rs3135388 and general transcription factor IIH polypeptide 4 [*GTF2H4*] variants). *Important* (top-ranking) predictors from the Random Forests analysis were determined based on the distribution of the VI scores (see Supplementary Figures 1-3), and were chosen for further investigation by CART.

RESULTS

A total of 485 SNPs in 72 genes related to DNA repair pathways were investigated in 1,343 MS cases and 1,379 healthy controls of European ancestry. The rs3135388 A variant, which is highly correlated with *HLA-DRB1*1501* (4) on chromosome 6p21.32, was significantly associated with MS risk, as expected (OR=2.7, 95% CI: 2.4, 3.1, $P_{unadjusted}<7.8 \times 10^{-48}$).

Five SNPs demonstrated some evidence for association with MS susceptibility, however, after adjusting for multiple testing, one significant association with MS susceptibility was observed (Table 2). A variant (rs1264307) within *GTF2H4*, a NER gene on chromosome 6p21.33, was significantly associated with MS risk (OR=0.73, $P_{FDR\text{-}BH}<3.5 \times 10^{-5}$). Interestingly, the *GTF2H4* SNP (rs1264307) was not in linkage disequilibrium (LD; $r^2<0.01$) with the *HLA-DRB1*1501* tagging SNP (rs3135388). Similarly, *HLA-DRB1*1501* stratified analyses did not reveal significant associations (Table 3 and 4), with the exception of *GTF2H4*, which was associated with decreased MS risk in **1501* negative individuals (OR=0.74, $P_{FDR\text{-}BH}=0.03$; Table 4). Haplotype analyses including two other *GTF2H4* SNPs did not reveal additional information, demonstrating that the association was primarily due to rs1264307 (*results not shown*). CHM analysis of the *HLA-DRB1*1501* tagging SNP (rs3135388) and the *GTF2H4* variant (rs1264307) demonstrated the *HLA-DRB1*1501* allele conferred increased risk independent of *GTH2H4* (Table 5).

A total of 31,666 epistatic interactions were investigated between SNPs within each specific pathway (6,780 BER interactions, 15,911 NER interactions, 5,659 HR interactions, and 3,316 NHEJ interactions were formally tested); no interactions were significant after adjusting for multiple testing (*results not shown*). The multigenic investigation (combined effect) of the 25 non-synonymous variants in 18 DNA repair genes (Table 6), did not demonstrate any association with MS susceptibility ($P>0.05$; Table 7).

Chromosome 6p21 variants (*HLA-DRB1*1501* and *GTF2H4* [rs1264307]) were very *important* predictors of MS susceptibility from Random Forests analysis (see Figures 1 and 2). As strong effects may mask the importance of other predictors, we removed, sequentially, *HLA-DRB1*1501* data, followed by all chromosome 6p21 genotype data (*HLA-DRB1*1501* and three *GTF2H4* SNP variants), and performed two additional Random Forests analyses. A total of nine

important predictors were additionally observed (see Figure 3; Table 8). Seven of these demonstrated a $P_{unadjusted}<0.05$ when analyzed using single locus association testing (Table 9), and included variants within *BRCA2*, *DDB2*, *ERCC3*, *RAD23A*, *RPA3*, *XAB2* and *XRCC4*. Using CART, a classification tree including five of the nine *important* predictors was constructed. The relationship between these genetic variants is illustrated in Figure 4. Interestingly, four variants were located within NER genes. The top-ranking Random Forests SNP and the variant used in the first node of the classification tree was an intronic *XAB2* variant (rs4134860; OR=1.26, $P_{unadjusted}=9.7 \times 10^{-4}$ from single-locus testing).

DISCUSSION

DNA damage from both endogenous and exogenous sources continuously challenges the genomic integrity of both nDNA and mtDNA. Fortunately, DNA repair processes help maintain genetic stability, and play an important role in maintaining a healthy immune and nervous system (39, 40). Over 150 genes are involved in the known nDNA and mtDNA repair pathways. The current study focused on common genetic variation within 72 genes derived from four principal DNA repair pathways: BER, NER, HR and NHEJ. We observed significant evidence for association between a SNP in *GTF2H4*, a NER gene, and risk for MS. Significant results for other genetic variants based on marginal, epistatic, and multigenic tests of association were not observed after stringent correction for multiple testing.

Using a non-parametric approach comprised of the Random Forests and CART algorithms, evidence was observed for a predictive relationship for MS based on 9 variants in NER, HR and NHEJ genes. Specifically, variants within NER genes were most prominent among predictors of MS; that is, four of five variants incorporated into the classification tree were in *XAB2*, *RPA3*, *ERCC3* and *DDB2*. In general, NER is responsible for removing nDNA lesions distorting the DNA helix (i.e. bulky DNA adducts) through several steps, which include: lesion detection, incision of the damaged strand, removal of the lesion-containing oligonucleotide, replacement of the removed oligonucleotide, and DNA ligation (41, 42).

The *GTF2H4* rs1264307 SNP, located in intron 11, was significantly associated with MS susceptibility, and the association was predominantly in **1501* non-carriers. *GTF2H4* encodes an integral subunit (p52) of the important transcription factor IIH (TFIIH) protein complex, the helicase responsible for unwinding DNA structure, allowing repair (41). The rs1264307 variant is <150bp from exon 12 in *GTF2H4* and is in strong LD ($r^2=1$) with several SNPs beyond exons 12 and 13 among CEU subjects (exon 12 and 13 variants were not available in phased CEU data from the HapMap project (26)). Additionally, SNP rs1264307 resides within a haplotype block spanning 38kb and three other genes (*VARS2*, *SFTA2* and *DPCR1*). *GTF2H4* is located on chromosome 6p21.33, among several MHC class I genes; however, it was not in LD ($r^2<0.01$) with *HLA-DRB1*1501* (rs3135388) on chromosome 6p21.32. Conditional haplotype analysis confirmed *HLA-DRB1*1501* conferred increased risk of MS, independent of *GTF2H4* (Supplementary Table 4). A SNP (rs1264303) within the 5' UTR region of *VARS2* and immediately centromeric to *GTF2H4* was also in strong LD ($r^2=1$) with our variant among CEU subjects, and shows evidence for association with MS risk (OR=0.83, $P=1.3 \times 10^{-3}$) in an independent dataset (931 cases, 2,431 controls, *data not shown*) (4). However, the potential involvement of *HLA*-class I genes (43), and the well-established allelic heterogeneity at the *HLA*-

DRB1 locus in MS (44, 45) underscore the need to fully examine *GTF2H4* and nearby variation, in conjunction with full MHC data and classical HLA loci. While *HLA-DRB1*1501* status was established for all MS cases and controls in the current study, complete *HLA-DRB1* typing was not available.

It was important to simultaneously consider genetic variation in these pathways, as multigene variants have been associated with diminished DNA repair capacity and persistent DNA damage (46, 47), and by extension could influence apoptosis (48) and cellular senescence (49). In the current study, we specifically investigated 25 non-synonymous coding SNPs in 18 DNA repair genes. However, a combined effect of missense variants based on carrier status for number of alleles at each locus showed no evidence of association with MS.

We cannot exclude the potential role of other, yet unidentified, environmental factors interacting with DNA repair genes to influence risk for MS. Interactions between DNA repair and exposure to tobacco smoke have been implicated in some cancers (50, 51). Whether gene–environment relationships are important in MS, given strong evidence for smoking as a risk factor (20, 21, 52-54), needs further investigation. Moreover, although common genetic variants within DNA repair genes were genotyped and/or captured and subsequently excluded as major risk factors for MS in the present study, risk due to rare variants, perhaps in particular subsets of MS cases, cannot be excluded.

The primary strengths of this pathway-driven candidate gene investigation in MS are: 1) the availability of dense genotyping across defined candidate genes, allowing for a comprehensive investigation of four DNA repair pathways; 2) the use of a homogenous study population; 3) the use of a large study population that was statistically well powered to identify modest genetic effects; 4) the implementation of stringent significance criteria, allowing for clearer interpretation of results; 5) the application of a multigenic approach to investigate biologically functional variants from these pathways; and 6) the parallel application of non-parametric methods to complement marginal tests of association. Despite the strengths of this investigation, these results must be interpreted cautiously as a replication analysis was not conducted to confirm these reported associations, and no environmental exposures were available. Determining the biological relevance of DNA repair pathways, particularly NER, in MS will require more extensive genetic and molecular research.

REFERENCES

1. Oksenberg JR, Barcellos LF. Multiple sclerosis genetics: leaving no stone unturned. Genes Immun 2005;6:375-87.
2. Oksenberg JR, Barcellos LF, Cree BA, et al. Mapping multiple sclerosis susceptibility to the HLA-DR locus in African Americans. Am J Hum Genet 2004;74:160-7.
3. Caillier SJ, Briggs F, Cree BA, et al. Uncoupling the roles of HLA-DRB1 and HLA-DRB5 genes in multiple sclerosis. J Immunol 2008;181:5473-80.
4. Hafler DA, Compston A, Sawcer S, et al. Risk alleles for multiple sclerosis identified by a genomewide study. N Engl J Med 2007;357:851-62.
5. The International Multiple Sclerosis Genetics Consortium (IMSGC). Refining genetic associations in multiple sclerosis. Lancet Neurol 2008;7:567-9.
6. Rubio JP, Stankovich J, Field J, et al. Replication of KIAA0350, IL2RA, RPL5 and CD58 as multiple sclerosis susceptibility genes in Australians. Genes Immun 2008;9:624-30.
7. Perera D, Stankovich J, Butzkueven H, et al. Fine mapping of multiple sclerosis susceptibility genes provides evidence of allelic heterogeneity at the IL2RA locus. J Neuroimmunol 2009;211:105-9.
8. De Jager PL, Baecher-Allan C, Maier LM, et al. The role of the CD58 locus in multiple sclerosis. Proc Natl Acad Sci U S A 2009;106:5264-9.
9. De Jager PL, Jia X, Wang J, et al. Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. Nat Genet 2009;41:776-82.
10. The International Multiple Sclerosis Genetics Consortium (IMSGC). Comprehensive follow-up of the first genome-wide association study of multiple sclerosis identifies KIF21B and TMEM39A as susceptibility loci. Hum Mol Genet 2010;19:953-962.
11. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. Science 2008;322:881-8.
12. Thomas DC. The need for a systematic approach to complex pathways in molecular epidemiology. Cancer Epidemiol Biomarkers Prev 2005;14:557-9.
13. Baranzini SE, Galwey NW, Wang J, et al. Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. Hum Mol Genet 2009;18:2078-90.
14. Vladimirova O, O'Connor J, Cahill A, Alder H, Butunoi C, Kalman B. Oxidative damage to DNA in plaques of MS brains. Mult Scler 1998;4:413-8.
15. Lu F, Selak M, O'Connor J, et al. Oxidative damage to mitochondrial DNA and activity of mitochondrial enzymes in chronic active lesions of multiple sclerosis. J Neurol Sci 2000;177:95-103.
16. Tajouri L, Mellick AS, Ashton KJ, et al. Quantitative and qualitative changes in gene expression patterns characterize the activity of plaques in multiple sclerosis. Brain Res Mol Brain Res 2003;119:170-83.
17. Mandel M, Gurevich M, Pauzner R, Kaminski N, Achiron A. Autoimmunity gene expression portrait: specific signature that intersects or differentiates between multiple sclerosis and systemic lupus erythematosus. Clin Exp Immunol 2004;138:164-70.
18. Satoh J, Nakanishi M, Koike F, et al. Microarray analysis identifies an aberrant expression of apoptosis and DNA damage-regulatory genes in multiple sclerosis. Neurobiol Dis 2005;18:537-50.

19. Giovannoni G, Heales SJ, Land JM, Thompson EJ. The potential role of nitric oxide in multiple sclerosis. Mult Scler 1998;4:212-6.
20. Hernan MA, Olek MJ, Ascherio A. Cigarette smoking and incidence of multiple sclerosis. Am J Epidemiol 2001;154:69-74.
21. Riise T, Nortvedt MW, Ascherio A. Smoking is a risk factor for multiple sclerosis. Neurology 2003;61:1122-4.
22. McDonald WI, Compston A, Edan G, et al. Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis. Ann Neurol 2001;50:121-7.
23. Thompson AJ, Montalban X, Barkhof F, et al. Diagnostic criteria for primary progressive multiple sclerosis: a position paper. Ann Neurol 2000;47:831-5.
24. Wood RD, Mitchell M, Lindahl T. Human DNA repair genes, 2005. Mutat Res 2005;577:275-83.
25. Song CM, Yeo BH, Tantoso E, et al. iHAP--integrated haplotype analysis pipeline for characterizing the haplotype structure of genes. BMC Bioinformatics 2006;7:525.
26. The International HapMap Project. Nature 2003;426:789-96.
27. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 2003;31:3812-4.
28. Steemers FJ, Chang W, Lee G, Barker DL, Shen R, Gunderson KL. Whole-genome genotyping with the single-base extension assay. Nat Methods 2006;3:31-3.
29. de Bakker PI, McVean G, Sabeti PC, et al. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat Genet 2006;38:1166-72.
30. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559-75.
31. Gabriel SB, Schaffner SF, Nguyen H, et al. The structure of haplotype blocks in the human genome. Science 2002;296:2225-9.
32. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2005;21:263-5.
33. Valdes AM, Thomson G. Detecting disease-predisposing variants: the haplotype method. Am J Hum Genet 1997;60:703-16.
34. Wu X, Gu J, Grossman HB, et al. Bladder cancer predisposition: a multigenic approach to DNA-repair and cell-cycle-control genes. Am J Hum Genet 2006;78:464-79.
35. Beuten J, Gelfond JA, Franke JL, et al. Single and multigenic analysis of the association between variants in 12 steroid hormone metabolism genes and risk of prostate cancer. Cancer Epidemiol Biomarkers Prev 2009;18:1869-80.
36. Cuzick J. A Wilcoxon-type test for trend. Stat Med 1985;4:87-90.
37. Breiman L. Random Forests. Machine Learning 2001;45:5-32.
38. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 2007;81:1084-97.
39. Rivera-Munoz P, Malivert L, Derdouch S, et al. DNA repair and the immune system: From V(D)J recombination to aging lymphocytes. Eur J Immunol 2007;37 Suppl 1:S71-82.
40. Fishel ML, Vasko MR, Kelley MR. DNA repair in neurons: so if they don't divide what's to repair? Mutat Res 2007;614:24-36.

41. Nouspikel T. DNA repair in mammalian cells : Nucleotide excision repair: variations on versatility. Cell Mol Life Sci 2009;66:994-1009.
42. Larsen NB, Rasmussen M, Rasmussen LJ. Nuclear and mitochondrial DNA repair: similar pathways? Mitochondrion 2005;5:89-108.
43. Yeo TW, De Jager PL, Gregory SG, et al. A second major histocompatibility complex susceptibility locus for multiple sclerosis. Ann Neurol 2007;61:228-36.
44. Barcellos LF, Sawcer S, Ramsay PP, et al. Heterogeneity at the HLA-DRB1 locus and risk for multiple sclerosis. Hum Mol Genet 2006;15:2813-24.
45. Dyment DA, Herrera BM, Cader MZ, et al. Complex interactions among MHC haplotypes in multiple sclerosis: susceptibility and resistance. Hum Mol Genet 2005;14:2019-26.
46. Jones IM, Thomas CB, Xi T, Mohrenweiser HW, Nelson DO. Exploration of methods to identify polymorphisms associated with variation in DNA repair capacity phenotypes. Mutat Res 2007;616:213-20.
47. Shin A, Lee KM, Ahn B, et al. Genotype-phenotype relationship between DNA repair gene genetic polymorphisms and DNA repair capacity. Asian Pac J Cancer Prev 2008;9:501-5.
48. Norbury CJ, Zhivotovsky B. DNA damage-induced apoptosis. Oncogene 2004;23:2797-808.
49. d'Adda di Fagagna F. Living on a break: cellular senescence as a DNA-damage response. Nat Rev Cancer 2008;8:512-22.
50. Horikawa Y, Gu J, Wu X. Genetic susceptibility to bladder cancer with an emphasis on gene-gene and gene-environmental interactions. Curr Opin Urol 2008;18:493-8.
51. Taioli E. Gene-environment interaction in tobacco-related cancers. Carcinogenesis 2008;29:1467-74.
52. Nortvedt MW, Riise T, Maeland JG. Multiple sclerosis and lifestyle factors: the Hordaland Health Study. Neurol Sci 2005;26:334-9.
53. Sundstrom P, Nystrom L, Hallmans G. Smoke exposure increases the risk for multiple sclerosis. Eur J Neurol 2008;15:579-83.
54. Di Pauli F, Reindl M, Ehling R, et al. Smoking is a risk factor for early conversion to clinically definite multiple sclerosis. Mult Scler 2008;14:1026-30.

Table 1: DNA Repair Genes Investigated

Gene	Biological Pathway
APEX1	BER
APEX2	BER
LIG3	BER
MBD4	BER
MPG	BER
MUTYH	BER
NEIL1	BER
NEIL2	BER
NTHL1	BER
OGG1	BER
PARP1	BER
PARP2	BER
PNKP	BER
POLB	BER
POLG	BER
RFC1	BER
RFC2	BER
RFC4	BER
RFC5	BER
SMUG1	BER
TDG	BER
XRCC1	BER
BRCA1	HR
BRCA2	HR
EME1	HR
MRE11A	HR
MUS81	HR
NBN	HR
RAD51	HR
RAD51C	HR
RAD51L1	HR
RAD51L3	HR
RAD52	HR
RAD54B	HR
RAD54L	HR
XRCC2	HR
XRCC3	HR
CCNH	NER
CDK7	NER
DDB2	NER
ERCC1	NER
ERCC2	NER

ERCC3	NER
ERCC4	NER
ERCC5	NER
ERCC6	NER
ERCC8	NER
GTF2H1	NER
GTF2H4	NER
GTF2H5	NER
MMS19L	NER
MNAT1	NER
PCNA	NER
POLD1	NER
POLE	NER
RAD23A	NER
RAD23B	NER
RPA1	NER
RPA2	NER
RPA3	NER
XAB2	NER
XPA	NER
XPC	NER
DCLRE1C	NHEJ
FEN1	NHEJ
LIG4	NHEJ
NHEJ1	NHEJ
PRKDC	NHEJ
WRN	NHEJ
XRCC4	NHEJ
XRCC5	NHEJ
XRCC6	NHEJ

Table 2: Marginal Association Results ($P_{\text{unadjusted}}<0.01$) for DNA Repair Variants and MS Susceptibility

Chr	SNP	Base-pair Location	$P_{\text{unadjusted}}$	$P_{\text{FDR-BH}}$	OR	95% CI	Gene	DNA Repair pathway
6	rs1264307	30,988,736	7.1×10^{-8}	3.5×10^{-5}	0.73	0.66, 0.82	*GTF2H4*	NER
19	rs4134860	7,592,407	9.7×10^{-4}	0.23	1.26	1.10, 1.44	*XAB2*	NER
5	rs9293329	82,432,343	0.0055	0.59	0.77	0.64, 0.93	*XRCC4*	NHEJ
2	rs4150454	127,755,014	0.0069	0.59	1.16	1.04, 1.30	*ERCC3*	NER
7	rs2110554	7,722,116	0.0072	0.59	1.32	1.08, 1.61	*RPA3*	NER

Table 3: Marginal Association Results ($P_{unadjusted}$<0.01) for DNA Repair Variants and MS Susceptibility in *HLA-DRB1*1501* Carriers[1].

Chr	SNP	Base-pair Location	$P_{unadjusted}$	P_{FDR-BH}	OR	95% CI	Gene	DNA Repair pathway
19	rs4134860	7,592,407	0.0010	0.26	1.50	1.18, 1.91	*XAB2*	NER
13	rs11571789	31,857,240	0.0011	0.26	0.63	0.48, 0.83	*BRCA2*	HR
13	rs11571686	31,820,331	0.0018	0.28	0.65	0.49, 0.85	*BRCA2*	HR
17	rs1131636	1,747,939	0.0034	0.31	0.76	0.63, 0.91	*RPA1*	NER
19	rs794078	7,591,843	0.0037	0.31	0.73	0.59, 0.90	*XAB2*	NER
1	rs2255403	224,642,893	0.0040	0.31	0.70	0.55, 0.89	*PARP1*	BER
1	rs752307	224,618,152	0.0045	0.31	0.70	0.55, 0.90	*PARP1*	BER
13	rs206079	31,818,618	0.0053	0.32	1.30	1.08, 1.56	*BRCA2*	HR
2	rs4150454	127,755,014	0.0099	0.53	1.28	1.06, 1.55	*ERCC3*	NER

[1] Using 702 MS cases and 345 controls who carried the rs3135388 A risk variant.

Table 4: Marginal Association Results ($P_{unadjusted}$<0.01) for DNA Repair Variants and MS Susceptibility in *HLA-DRB1*1501* Non-carriers[1].

Chr	SNP	Base-pair Location	$P_{unadjusted}$	P_{FDR-BH}	OR	95% CI	Gene	DNA Repair pathway
6	rs1264307	30,988,736	5.7×10^{-5}	0.028	0.74	0.64, 0.86	*GTF2H4*	NER
7	rs3094406	151,993,798	0.0021	0.33	1.42	1.14, 1.78	*XRCC2*	HR
19	rs3213266	48,767,476	0.0025	0.33	0.67	0.51, 0.87	*XRCC1*	BER
12	rs5744761	131,762,012	0.0033	0.33	1.74	1.20, 2.52	*POLE*	NER
11	rs16920467	93,827,574	0.0037	0.33	1.61	1.16, 2.23	*MRE11A*	HR
6	rs1264308	30,987,966	0.0040	0.33	1.33	1.09, 1.61	*GTF2H4*	NER
10	rs11593133	15,035,155	0.0049	0.34	0.81	0.69, 0.94	*DCLRE1C*	NHEJ
5	rs9293329	82,432,343	0.0056	0.34	0.70	0.55, 0.90	*XRCC4*	NHEJ
1	rs12410307	46,502,985	0.0085	0.46	0.72	0.56, 0.92	*RAD54L*	HR

[1] Using 638 MS cases and 1,030 controls who did not carry the rs3135388 A risk variant.

Table 5: Conditional Haplotype Method of *HLA-DRB1*1501* (rs3135388**A**) and *GTF2H4* (rs1264307).

***GTF2H4-HLA-DRB1*1501*- haplotype**	**Case/Control counts**	**OR (95% CI)**	***P*-value**
CC	1265/1452	Ref.	
C**A**	600/272	2.53 (2.15, 2.99)	>0.00001
TC	621/932	Ref.	
T**A**	202/104	2.92 (2.24, 3.81)	>0.00001

Table 6: Marginal Assosciation Results for 18 DNA Repair Missense SNPs Variants and MS Susceptibility

Chr	SNP	$P_{unadjusted}$	OR	**1501* carriers		**1501* non-carriers		Gene	DNA Repair pathway
				$P_{unadjusted}$	OR	$P_{unadjusted}$	OR		
1	rs3219484	0.46	1.08	0.87	0.97	0.58	1.08	*MUTYH*	BER
3	rs2227999	0.23	1.15	0.77	1.06	0.24	1.19	*XPC*	NER
3	rs2307298	0.27	1.32	0.60	1.27	0.24	1.46	*MBD4*	BER
5	rs2266690	0.50	1.05	0.37	1.11	0.57	1.05	*CCNH*	NER
7	rs3218536	0.68	0.96	0.30	0.84	0.96	0.99	*XRCC2*	HR
8	rs8178017	0.15	1.20	0.11	1.49	0.27	1.20	*PRKDC*	NHEJ
10	rs3740526	0.21	1.07	0.91	1.01	0.09	1.13	*MMS19L*	NER
12	rs5745066	0.21	1.29	0.55	1.23	0.32	1.31	*POLE*	NER
13	rs766173	0.32	0.86	**0.029**	0.57	0.37	1.18	*BRCA2*	HR
13	rs2227869	0.48	0.91	0.40	0.81	0.85	1.03	*ERCC5*	NER
13	rs17655	0.57	0.96	0.97	1.01	0.78	0.98	*ERCC5*	NER
13	rs1805388	0.33	0.93	0.92	0.99	0.27	0.90	*LIG4*	NHEJ
13	rs1805389	0.36	0.90	0.39	0.83	0.94	0.99	*LIG4*	NHEJ
14	rs3093921	0.29	0.81	0.78	1.10	0.08	0.60	*PARP2*	BER
14	rs3093926	**0.047**	1.24	0.06	1.44	0.31	1.16	*PARP2*	BER
14	rs861539	0.93	1.01	0.53	1.06	0.65	1.03	*XRCC3*	HR
15	rs2307441	0.72	0.95	0.55	0.88	0.61	0.91	*POLG*	BER
16	rs1800067	0.32	1.11	0.08	1.39	0.70	1.05	*ERCC4*	NER
17	rs5030755	0.98	1.00	0.94	0.99	0.50	0.92	*RPA1*	NER
17	rs1799967	0.24	1.32	0.17	1.98	0.43	1.26	*BRCA1*	HR
17	rs16941	0.27	0.94	**0.038**	0.81	0.56	1.05	*BRCA1*	HR
17	rs799917	0.43	0.96	**0.044**	0.82	0.33	1.08	*BRCA1*	HR
17	rs4986850	0.08	0.84	0.09	0.75	0.60	0.93	*BRCA1*	HR
17	rs1799950	0.39	1.10	0.53	1.13	0.53	1.09	*BRCA1*	HR
19	rs13181	0.31	0.94	0.51	1.07	0.06	0.87	*ERCC2*	NER

Table 7: Combined effects of missense variants

No. of missense alleles	Case/Control counts	Adjusted OR [1] (95% CI)	*P*-value
≤4	341/353	Ref. [c]	
5	199/189	1.09 (0.84, 0.99)	0.50
6	206/228	0.91 (0.71, 1.42)	0.48
7	213/208	1.11 (0.86, 1.17)	0.44
8	162/173	0.96 (0.73, 1.42)	0.78
9	102/95	1.21 (0.87, 1.26)	0.26
10	51/63	0.83 (0.55, 1.68)	0.39
11	36/35	1.12 (0.67, 1.26)	0.66
12	19/24	0.90 (0.47, 1.87)	0.74
13	11/6	2.12 (0.76, 1.71)	0.15
14	3/2	0.73 (0.10, 5.94)	0.75
15	1/2	0.77 (0.07, 8.59)	0.83
Per allele		1.00 (0.98, 1.03)	0.74
P for trend			0.82

[1]Odds Ratios and 95% CI from unconditional logistic regression, adjusted for gender and *HLA-DRB1*1501* (rs3135388).

Table 8: *Important* Predictors From Random Forest Results Across Runs

VI Rank	All Variants		No *HLA-DRB1*		No Chromosome 6p21 variants	
	SNP	**Gene**	**SNP**	**Gene**	**SNP**	**Gene**
1	rs3135388	*HLA-DRB1*	rs1264307	*GTF2H4*	rs4134860	*XAB2*
2	rs1264307	*GTF2H4*	rs4134860	*XAB2*	rs2974754	*RAD23A*
3			rs4150454	*ERCC3*	rs7783714	*RPA3*
4			rs1231201	*PRKDC*	rs9293329	*XRCC4*
5			rs9293329	*XRCC4*	rs4134813	*XAB2*
6			rs2974754	*RAD23A*	rs1231201	*PRKDC*
7			rs9562605	*BRCA2*	rs4150454	*ERCC3*
8					rs9562605	*BRCA2*
9					rs2957873	*DDB2*

Table 9: Important Predictors Identified by Random Forests, When Excluding Chromosome 6p21 Variants.

Chr	SNP	Base-pair Location	VI rank [1]	$P_{unadjusted}$	OR	95% CI	Gene	DNA Repair pathway
19	rs4134860	7,592,407	1	9.7×10^{-4}	1.26	1.10, 1.44	*XAB2*	NER
19	rs2974754	12,922,983	2	0.042	0.89	0.80, 1.00	*RAD23A*	NER
7	rs7783714	7,658,402	3	0.022	0.87	0.78, 0.98	*RPA3*	NER
5	rs9293329	82,432,343	4	0.0055	0.77	0.64, 0.93	*XRCC4*	NHEJ
19	rs4134813	7,600,021	5	0.10	1.10	0.98, 1.23	*XAB2*	NER
8	rs1231201	49,008,716	6	0.53	0.97	0.87, 1.08	*PRKDC*	NHEJ
2	rs4150454	127,755,014	7	0.0069	1.16	1.04, 1.30	*ERCC3*	NER
13	rs9562605	31,788,026	8	0.045	0.88	0.78, 1.00	*BRCA2*	HR
11	rs2957873	47,205,870	9	0.031	1.16	1.01, 1.32	*DDB2*	NER

[1] Rank of variable importance score for *important* predictors.

Figure 1: Variable Importance Scores for the Top 50 Predictors for the Random Forests Analysis Where All Genetic Variants Were Used to Predict MS.

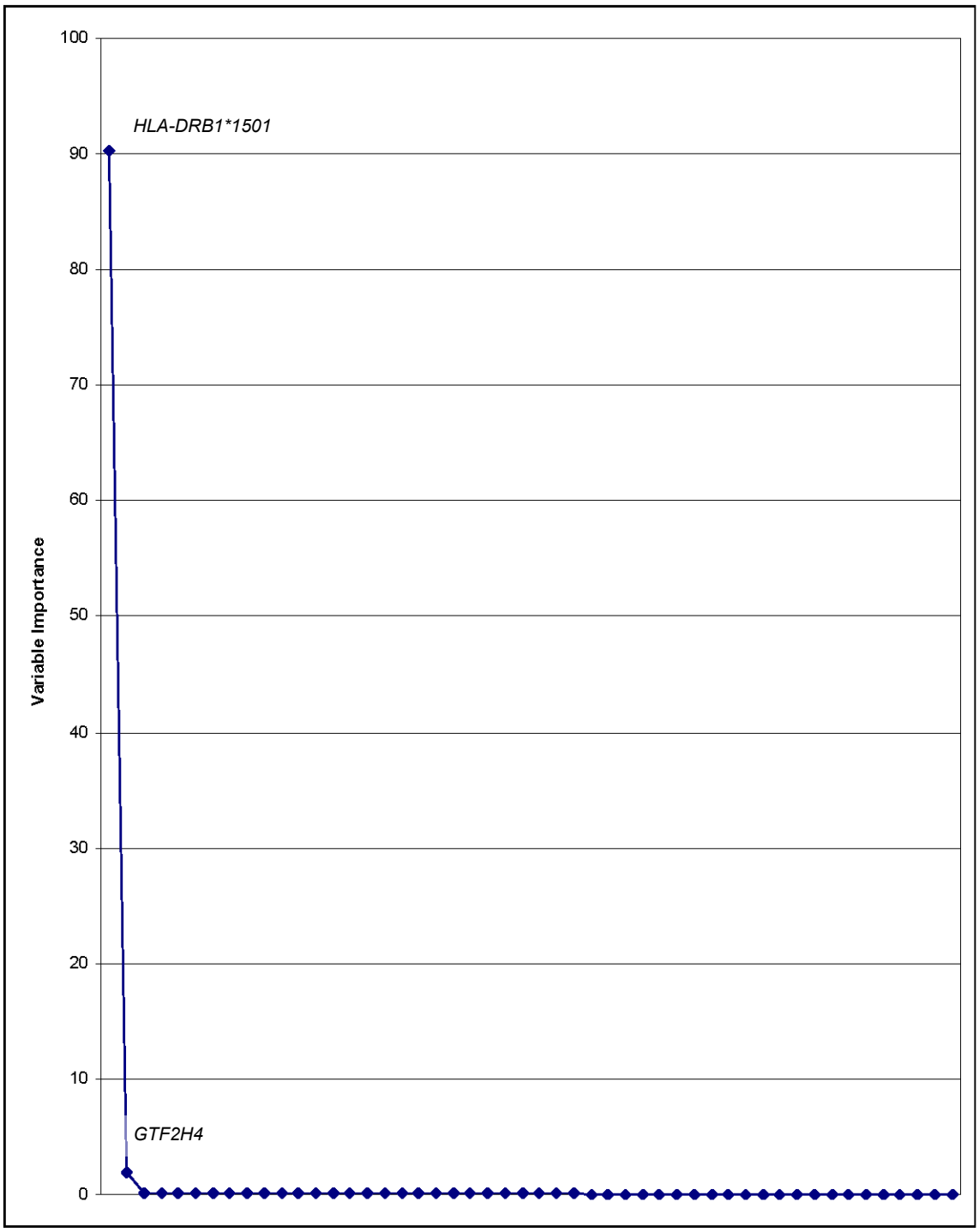


Figure 2: Variable Importance Scores for the Top 50 Predictors for the Random Forests Analysis Where All Genetic Variants, Except *HLA-DRB1*1501*, Were Used to Predict MS.

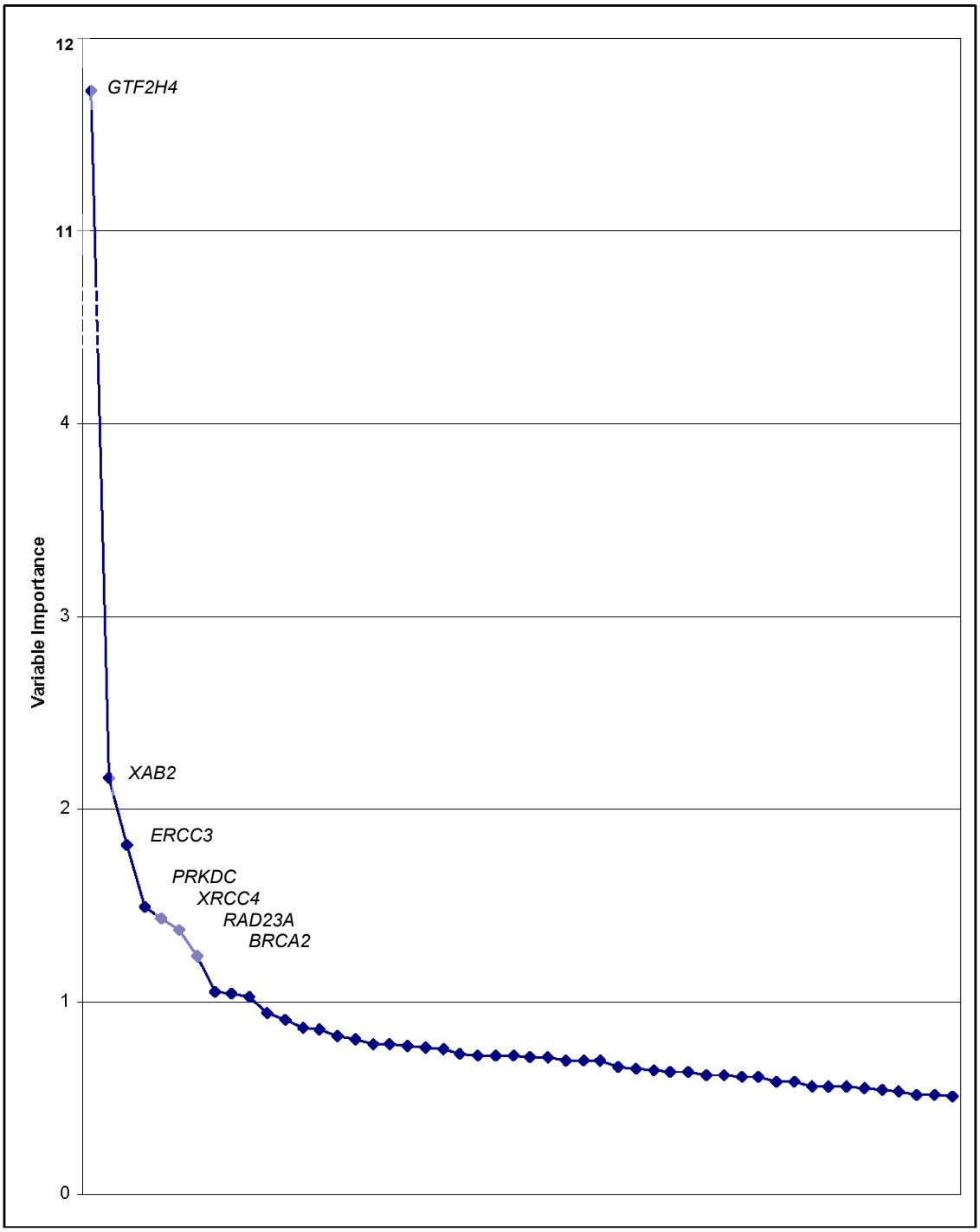


Figure 3: Variable Importance Scores for the Top 50 Predictors for the Random Forests Analysis Where All Genetic Variants, Except those on Chromosome 6p21 (*HLA-DRB1*1501* and *GTF2H4* variants), Were Used to Predict MS.

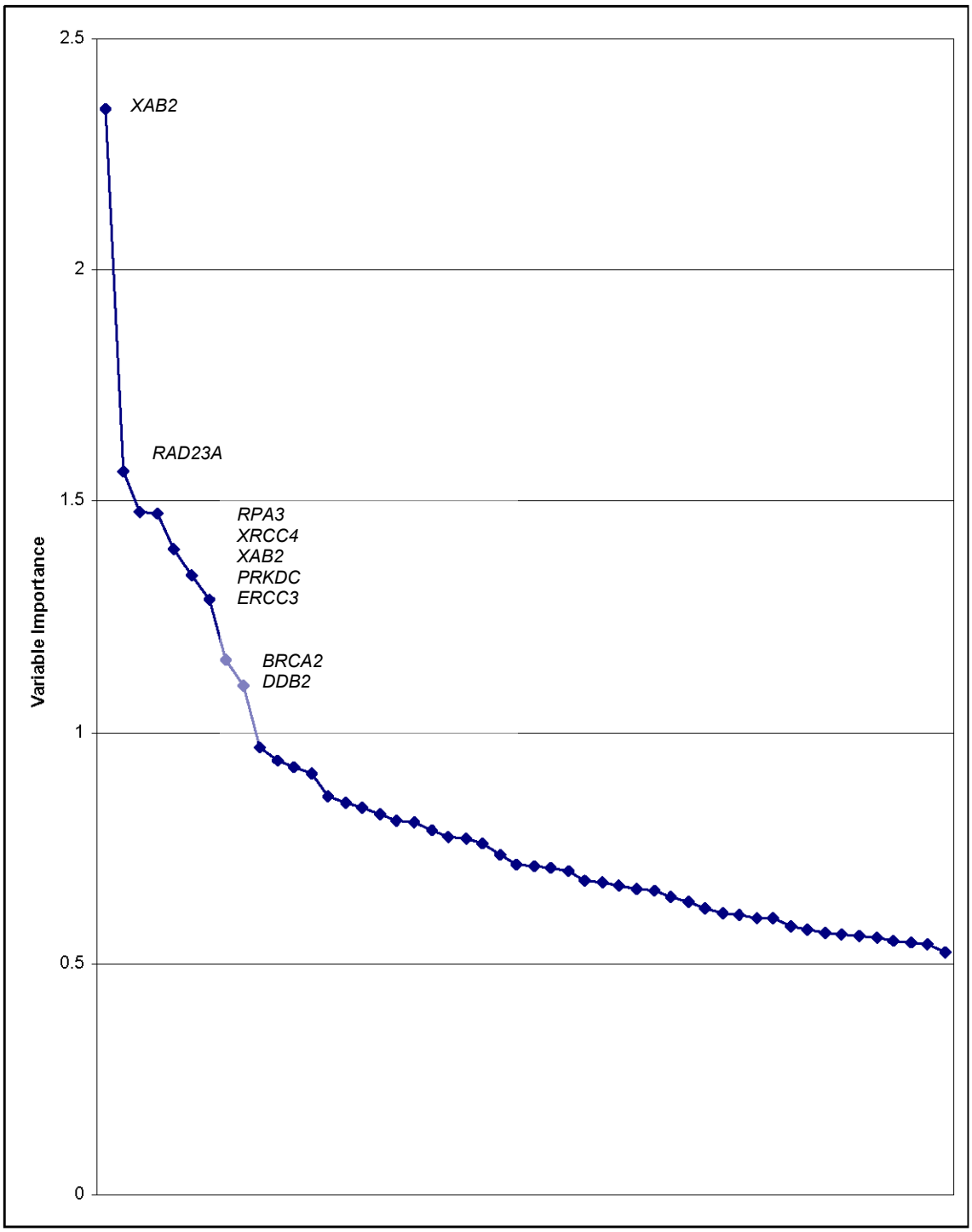
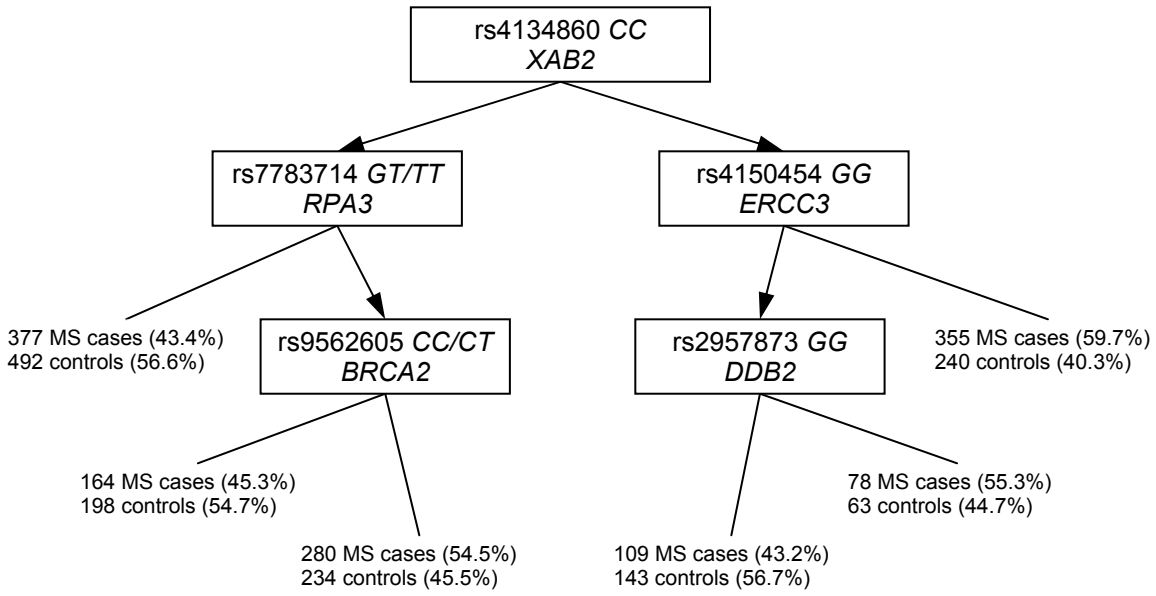


Figure 4: Classification and regression tree (CART) analysis of *important* predictors (not on chromosome 6p21) identified by Random Forests (Table 4).



Chapter 2

Corticotropin Releasing Hormone Receptor 1 (*CRHR1*) is a Novel Multiple Sclerosis Susceptibility Locus.

ABSTRACT

The primary genetic risk factor in multiple sclerosis (MS) is the *HLA-DRB1*1501* allele; however, much of the remaining genetic contribution to MS has yet to be elucidated. Several lines of evidence support a role for neuroendocrine system involvement in autoimmunity which may, in part, be genetically determined. Here, we comprehensively investigated variation within 8 candidate hypothalamic-pituitary-adrenal (HPA) axis genes and susceptibility to MS. A total of 326 SNPs were investigated in 1,343 MS cases and 1,379 healthy controls of European ancestry using a multi-analytical strategy. Random Forests, a supervised machine learning algorithm, identified 8 intronic SNPs within the corticotropin releasing hormone receptor 1 or *CRHR1* locus on 17q21.31 as important predictors of MS. Based on univariate analyses: six *CRHR1* variants were associated with decreased risk for disease following a conservative correction for multiple tests. Independent replication was observed for *CRHR1* in a large meta-analysis comprised of 2,624 MS cases and 7,220 healthy controls of European ancestry. Results from a combined meta-analysis of all 3,967 MS cases and 8,599 controls provide strong evidence for the involvement of *CRHR1* in MS (rs242936: $p=9.7 \times 10^{-5}$). Further investigation of mechanisms involved in HPA axis regulation and response to stress in MS pathogenesis is warranted.

INTRODUCTION

Multiple sclerosis (MS; MIM #126200) is a clinically heterogeneous, autoimmune disease of the central nervous system (CNS) including two distinct, though intersecting, neuro-pathological phases: inflammation and neurodegeneration (1). The pathogenesis of MS involves a substantial genetic component, with *HLA* class II genes within the major histocompatibility complex (MHC) on chromosome 6p21 conferring approximately 50% of the genetic risk (2). Admixture analyses have shown the primary susceptibility locus for MS within the MHC is *HLA-DRB1,* and more specifically, the HLA-*DRB1*15* allele (3, 4). The complexities surrounding the *HLA-DRB1* contribution are the focus of multiple ongoing research efforts within the field (5-7). The identification of non-MHC susceptibility loci, while progressing, is far from complete. Recent genome wide association (GWA) and replication studies have begun to unravel the polygenic etiology of MS. Modest associations for over a dozen variants, including SNPs within *IL7RA*, *IL2RA*, *CLEC16A*, *CD58*, *TNFRSFA1*, *IRF8*, *KIF21B*, and *TMEM39A*, have been described (8-14). To date, GWA studies using currently available technology have had limited success refining the genetic etiology for most complex diseases, including MS (15, 16). Candidate gene approaches based on strong hypotheses and well-powered datasets still remain an important strategy for detection of disease-associated variants (17, 18).

MS pathogenesis is thought to involve multiple biological pathways that contribute to inflammatory and neurodegenerative components of the disease. Genetic variation within these pathways is therefore likely to be associated with disease predisposition. One biological pathway involved in MS is the hypothalamic-pituitary-adrenal (HPA) axis, a principal component of the neuroendocrine system that regulates individual response to physical and emotional stress, and maintains homeostasis with strong neuroimmune modulating properties (19, 20). Various measures of HPA axis activity have been investigated in small studies of clinically heterogeneous MS cases, where impaired HPA axis activity in MS cases appears to be dependent on clinical phenotypes and disease duration (21). A recent comprehensive analysis demonstrated strong evidence for HPA hyperactivity in 173 well-defined MS patients and 60 healthy controls (22).

The HPA axis is further implicated in pathology of MS by evidence supporting psychological stress as a risk factor for MS onset and exacerbations in several studies, despite methodological differences (23). MS cases are more likely to have experienced stressful life events prior to symptom onset than matched healthy controls for the same prodromal time period (24, 25) or controls with other neurologic or rheumatic diseases (26). Several studies suggest that stressful life events often precede disease relapse in MS cases (27-33); and interestingly, MS cases who experienced family or work conflicts have an increased risk of developing a new gadolinium enhancing (Gd+) MRI lesion (27). Psychological stress has also been associated with both susceptibility and disease progression for other autoimmune diseases (34-39). In addition, a recent retrospective cohort study examined hospital discharge records and reported significant associations between individuals who experience traumatic childhood stress and increased rates of first hospitalizations for any of 21 autoimmune diseases, including MS (40).

Although evidence supports the involvement of the HPA axis in the pathology of MS, the exact nature of the relationship and the underlying biological mechanism(s) remain unclear. In this study, we investigated variation within 8 candidate genes involved in mediating and/or

regulating the neuroendocrine activity of the HPA axis using a supervised machine learning algorithm (Random Forests). A total of 326 informative SNPS within *BDNF*, *CRHBP*, *CRHR1*, *CRHR2*, *HCRTR1*, *HCRTR2*, *OPRD1* and *OPRK1* were genotyped in 1,343 MS cases and 1,379 healthy controls of European ancestry. Random Forests was first used to rank important SNP predictors of case status. Univariate association tests, haplotype and logistic regression analyses were then used to further characterize relationships between important variants and risk for MS. We confirmed our significant associations in a large meta-analysis of 2,624 MS and 7,220 control subjects of European ancestry (13). We report for the first time, *CRHR1*as a novel susceptibility locus for MS, in a multi-stage analysis (Figure 1).

RESULTS

Stage I – Discovery Analysis

An exploratory dataset was comprised of 1,343 MS cases and 1,379 healthy controls of European ancestry (14). The rs3135388 (A/G) SNP was used to determine the presence of the *HLA-DRB1*1501* allele in MS cases and controls as previously described (8). As expected, the T variant was significantly associated with MS risk (odds ratio [OR]=2.7, 95% confidence interval [CI]: 2.4-3.1, p=7.8 x 10^{-48}) and demonstrates that our case-control sample is representative of those described in other studies. This SNP had a minor allele frequency (MAF) of 29.8% in cases and 13.6% in controls.

Using this dataset, Random Forests identified 8 intronic SNPs within *CRHR1* on chromosome 17q21.31 as important predictors of MS (Table 1; Figure 2). Six of the 8 important *CRHR1* SNPs were significantly associated with decreased risk of MS in logistic regressions models (p<0.05; Table 1), adjusted for gender and *HLA-DRB1*1501*. All 8 important SNPs demonstrated similar effect sizes of decreased risk of MS.

Stage II – Replication Analysis

Four of the 8 important *CRHR1* SNPs identified in Stage I were examined in an independent group of 2,624 MS and 7,220 control subjects of European ancestry to determine whether Stage I results could be replicated (Table 1). Data from MS cases and controls were comprised of three GWA scans generated on different SNP arrays, and subsequently each scan was imputed using a single panel of 2.56 million SNPs for meta-analysis, as previously described (13). Meta-analysis for *CRHR1* SNPs was conducted using a fixed-effects logistic model based on the observed and expected allele dosage, and adjusted for cohort of origin, gender, and *HLA-DRB1*1501* (13). Interestingly, the four available SNPs (of 8) within *CRHR1* were also significantly associated with decreased risk of MS in this independent dataset (Table 1).

Stage III – Combined Analysis

The data from Stages I and II were combined to generate final effect estimates for the four of the 8 important *CRHR1* SNPs. The meta-analysis from Stage II was rerun, including the Stage I data as an additional cohort. The four *CRHR1* variants were significantly (p<5 x 10^{-4}) more associated with decreased risk of MS in the combined meta-analysis of 3,967 MS cases and 8,599 controls subjects of European ancestry (Table 2).

Stage IV – Extension Analysis

CRHR1 is located within a polymorphic inversion (~900 kb) on 17q21. There are two 17q21 haplotypes, H1 and H2, comprised of SNPs which are in complete disequilibrium, as H2 is inverted relative to H1 (41). The H2 inversion is most frequent in Southern Europe and Southwest Asia, decreasing outward in all directions (42). To further investigate the *CRHR1* association, we examined linkage disequilibrium patterns for four of the 8 SNPs genotyped in CEU (Utah Residents with Northern and Western European Ancestry) Haplotype Map (Hapmap) data (version 3; Release #R2). Results show that the four SNPs tag a total of 8 *CRHR1* variants (r^2>0.5; *data not shown*).

Finally, combining all available genetic data for *CRHR1* from Stages I and II, we tested for association with MS using adjusted logistic regression as previously described (Figure 3). Results indicate there is very strong evidence for association between *CRHR1* and MS. A H2 inversion tagging SNP (rs1396862) was present in both data sets (Stage I and Stage II), and all logistic regression models were then rerun adjusting for presence of the inversion. Similar results were observed (Figure 4), and show the replicated *CRHR1* association is exerting an effect independent of the H2 inversion. In addition, results did not vary greatly when individuals carrying the H2 haplotype were excluded from analyses (*data not shown*).

Haplotypic analyses were conducted to further resolve the association between *CRHR1* SNPs and MS using dense genotypic data (N=89 SNPs) available from Stage I. Using D' confidence intervals (43) as implemented in Haploview v4.2 (44), a single haplotype block spanning nearly the entire *CRHR1* gene (46 kb; Figure 5) and including 81 SNPs was observed. H2 was identified by the presence of the rs1396862T variant; with a prevalence of 22.9%, comparable to frequencies observed in European Americans and most Western, Central and Southeastern Europeans (42). A total of 13 sub-haplotypes with a frequency $\geq$1% were observed, including 11 H1 and two H2 sub-haplotypes (Figure 6). The H1.5 sub-haplotype frequency significantly differed in MS cases and healthy controls (Figure 6; Table 3); it was significant associated with decreased risk of MS when compared to all other H1 sub-haplotypes (OR=0.64, 95% CI: 0.50-0.83, $p= 4.1 \times 10^{-4}$). Results did not differ when compared to all other H1 and H2 haplotypes (*data not shown*). Interestingly, H1.5 was the only sub-haplotype with all 8 SNPs identified in Stage I, and thus, concurred with results obtained from Random Forests analyses (Figure 6).

DISCUSSION

Results from the current study of variants within 8 candidate genes involved in mediating and/or regulating the neuroendocrine activity of the HPA axis demonstrate strong evidence for association between *CRHR1* (GeneID 1394) and MS susceptibility. A large, well-powered study design including both discovery and replication stages were utilized. European ancestry estimates based on genetic markers were used to remove population outliers for case-control datasets in both stages; therefore the potential impact of population stratification was minimized. The 17q21 region has been a key locus of interest in MS: it was identified as a possible MS locus in a meta-analysis of several genome-wide linkage screens (45). However, subsequent GWA scans have not successfully identified *CRHR1* or any other susceptibility locus in this region (46-48). Further, results for *CRHR1* in the current study would not meet criteria for genome-wide

significance (p>5 x 10^{-8}). Our study underscores the importance of considering biological function in genetic studies of MS, as well as other autoimmune disorders.

Close to 100% of the common genetic variation in *CRHR1* was captured in Stage I analyses, based on $r^2 \geq 0.95$ in CEU Hapmap data (Release #21). A total of 89 *CRHR1* SNP variants were investigated using Random Forests and the 8 SNPs identified as important were located in introns. Therefore, an explicit functional role is not evident. However, *in silico* analysis (http://fastsnp.ibms.sinica.edu.tw/) suggests four of the 8 important SNPs (rs171442, rs242938, rs173365, and rs17689966) are potential intronic enhancers that may alter transcription factor binding. Interestingly, the 8 important SNPs identified by Random Forest reside on a single haplotype (H1.5).

A summary of all *CRHR1* data considered in the current study shows evidence for association with MS was present for SNPs within intron 1 and across exons 3 through 14 (Figure 3 and Figure 4). The strongest signals are located within exons 3 and 4. Five missense polymorphisms have since been identified (dbSNP; accessed April 2010) in *CRHR1* exons 3 (rs16940655), 4 (rs41280114), 6 (rs75638861), and 14 (rs61732578 [may lead to altered splice regulation], rs75738089). The rs16940655 (*CRHR1* V60A) variant encodes a conservative amino-acid exchange (http://fastsnp.ibms.sinica.edu.tw/), and was genotyped in Stage I of the current study; however it was excluded from analyses due to low MAF (<1% in controls). It was present in 1.12% of MS cases and 0.87% of controls (p=0.39; logistic regression model adjusted for *HLA-DRB1*1501* and gender). Owing to low MAF, rs16940655 does not appear to be in 'strong' LD with any *CRHR1* variant investigated in Stage I (or in CEU Hapmap data); however, it is located near our replicated SNP (<1.7 kb from rs242939).

CRHR1 is located within a large region (~900 kb) of high linkage disequilibrium on chromosome 17q21.31, resulting from a local chromosomal inversion. The global distribution of two 17q21.31 haplotypes, H1 and H2 (the inversion), varies. A decrease in H2 frequency that follows a south to north gradient across Europe has been demonstrated with population data (42). We repeated all analyses of *CRHR1* with adjustment for the presence of the inversion, and also excluded individuals with the inversion from analyses, to address concerns related to population stratification; results did not vary. Similar to findings reported herein, a previous investigation of H2 haplotype tagging SNP rs9468 in 937 UK trio families showed no evidence for association (49). It is important to note that some *CRHR1* variants identified by Random Forests in Stage 1 also tagged ($r^2 \geq 0.5$) SNP variants within the nearby *IMP5* locus, based on assessment of linkage disequilibrium in CEU Hapmap data. Complete genotype data for variants within this locus were not available for MS cases and controls utilized in the current study. Therefore, detailed characterization of genetic variation within the *CRHR1-IMP5* region is needed. However, our findings based on association with replication strongly support a role for *CRHR1* variation in MS susceptibility.

The neuroendocrine, autonomic and behavioral stress response is centralized to a common corticotrophin-releasing hormone (CRH)-mediated mechanism (50). CRH is released into hypophyseal portal system by neurons in the paraventicular nucleus of the hypothalamus, and delivered to the anterior pituitary. There are two G protein-coupled CRH receptors that exert the effect of CRH: *CRHR1* and *CRHR2*, however only *CRHR1* is expressed within the pituitary and

demonstrates a strong affinity for CRH. The interaction of CRH and CRHR1 at the pituitary is exclusively responsible for activating the biosynthesis and release of the adrenocorticotropic hormone (ACTH) into peripheral circulation, which culminates in steriodogenesis and the release of glucocorticoids (GCs; i.e. cortisol) from cortex of the adrenal gland via ACTH receptors (51). GC receptors (GR; expressed by all nucleated cells) transduce the stress response signal in end organs, directly affecting transcription of GC-sensitive genes, with activational and/or inhibitory actions in various systems, including negative feedback at the pituitary to stop the response to stressful stimuli.

CRH is also found peripherally, and regulates inflammation through direct activation of *CRHR1* receptors on mast cells and monocytes/macrophages and via *in vivo* secretion of TNF, IL-1 and IL-6 (52-55). Interestingly, *CRHR1* is expressed on resting human B and T lymphocytes and other several other immune related cells and tissues (56-62). Furthermore, urocortin, a *CRHR1* ligand structurally similar to CRH, promotes microvessel permeability, and *CRHR1* is significantly overexpressed in early development at the blood-brain-barrier than in adulthood in animal models (63, 64).

A total of eight *CRHR1* isoforms: α (*no exon 6*), β (*contains all 14 exons*), c (*no exons 3 and 6*), d (*no exons 6 and 13*), e (*no exons 3 and 4, resulting in a frame shift and early stop codon in exon 8*), f (*no exon 12, resulting in a frame shift*), g (*no exon 11 and portions of exons 10 and 12 are missing*), and h (*contains a cryptic exon between exons 4 and 5, resulting in a frame shift and early stop codon in exon 5*) have been described, and have varied tissue distribution (65, 66). While *CRHR1*α appears to be the primary functional CRHR1 receptor (66), the diversity of *CRHR1* isoforms underscores its possible pleiotropic functions in various biological mechanisms, as well as complex responsive to CRHR1 ligands. For example, *CRHR1*α signaling is attenuated or amplified by other soluble isoforms (*CRHR1*e and *CRHR1*h, respectively) (67). Multiple *CRHR1* isoforms (α, β, c, e, and f) have been detected in mast cells (54).

In conclusion, this pathway-driven investigation adds to prior evidence that the HPA axis and related stress response mechanisms contribute to MS susceptibility. Notwithstanding the successes of GWA studies, they have demonstrated difficulty in fully refining the genetic etiology of complex diseases (16). Discussions on future directions for complex disease studies emphasize the need for incorporating biological knowledge into genetic investigations (17). The strengths of this investigation are: 1) genotyping of a large homogenous study population that was well powered to identify modest genetic effects; 2) the application of a robust non-parametric algorithm to guide association testing; 3) the replication of findings in a large, independent study population; and 4) extensive genotypic coverage of investigated loci. A key limitation of this investigation is that genetic data for other key HPA axis genes were not available; therefore it is necessary to investigate these unexplored relationships in MS, and in the context of the variants investigated here.

CRHR1 is a critical component of the HPA axis. An impaired HPA axis has been suspected to contribute to autoimmunity, including MS (20, 68-70), and genetic variation within related genes has been associated with the development of affective disorders and other stress-related clinical conditions, both marginally and through *gene* x *environment* interactions (71-74). Thus, further

investigations of these clinical conditions, exposure to stressful life events, and genetic variation with the HPA axis in MS is necessary. Results from this investigation provide evidence for a non-MHC mediated mechanism in the pathogenesis of MS.

MATERIALS and METHODS

Stage I – Discovery Analysis

Ten genes involved in mediating and/or regulating the neuroendocrine activity of the HPA axis were selected for investigation. Genic SNPs were selected for genotyping primarily based on their function as a tagging SNP. Using Tagger, as implemented in Haploview v3.3 (44), tagging SNPs were defined as a set of SNPs which captured 100% of the genetic variation in CEU Hapmap data (NCBI Build 35; Release #21, July 2006), with pairwise $r^2 \geq 0.95$. Additional SNPs not available in CEU Hapmap data were identified and selected from dbSNP (retrieved July, 2007) based on proximity to exons and to expand coverage where there was none (e.g. there were no *CRH* SNPs available in CEU Hapmap data (Release #21), therefore 4 SNPs were selected from dbSNP for genotyping).

A total of 486 SNPs in 10 genes were genotyped as a subset of 48,767 custom SNPs using the Illumina Infinium 60K BeadChip assay (75), in 2,961 (1,488 MS cases and 1,512 controls) participants recruited from three clinical centers (University of California, San Francisco; Harvard/MIT Board Institute; and Cambridge University) (14). Unrelated controls were obtained from these same US sites and from the British 1958 Birth Cohort Study. These controls were selected to provide nearly equivalent gender and age distributions (14). All participants self-reported as non-Hispanic whites. All MS cases met well-established disease criteria (76, 77). Informed consent was obtained from all study participants and approvals from local institutional review boards were secured at each recruitment site prior to enrollment.

A rigorous quality control protocol was utilized; Whole-genome Association Study Pipeline assessed sample and SNP genotyping efficiency (<95%), allele frequencies, gender errors, and Hardy-Weinberg equilibrium ($p<0.0001$), recursively. Samples were excluded if the probability of Caucasian European descent was <0.90. Additional quality control processes were performed, including the assessment of population outliers (14). The final quality control analysis yielded 2,722 individuals and 46,874 SNPs, of which 380 SNPs were relevant to this analysis. We further excluded 54 SNPs with a MAF<0.01. Missing genotypes were imputed in cases and controls using Beagle v2.1.3 (78). Therefore, a total of 326 SNPs within 8 genes related to the HPA axis were investigated in 1,343 MS cases and 1,379 healthy controls of European ancestry. The rs3135388 (A/G) SNP was used to determine the presence of the *HLA-DRB1*1501* allele as previously described (79).

We investigated the power to detect log-additive genetic associations. ORs ranging from 0.1 to 3.0 were examined, assuming a two-sided type 1 error of 5% ($\alpha=0.05$). Results indicated that our discovery analysis was sufficiently powered (>75%) to detect allelic OR ≤ 0.8 and ≥ 1.2 for almost all models considered.

All SNP genotypes in cases and controls were coded as 0, 1 or 2 copies of the minor allele, and investigated using Random Forests v6.40.179 (http://salford-systems.com/) with *mtry*=sqrt(p)

and *ntree*=5,000. Briefly, Random Forests independently grows a collection of recursively partitioned trees without pruning using bootstrap aggregating and random selection of a subset of all predictors to determine classification at each node (80). Approximately a third of the observations is not included in each bootstrap sample, but is used to assess the classification accuracy of each tree in the forest, by comparing the predicted versus the actual outcome. This process is repeated with each predictor, in a tree, randomly permuted. Finally, a single importance score for each variable is determined by any increase change in misclassification and used to generate a variable importance (VI) score. The VI scores rank predictors by their importance in classifying the outcome in the context of all predictors without model specification, and are robust to uninformative predictors and outliers. Additionally, the VI score potentially includes the effect of multiplex interactions between the predictors, as each variable selected at a node is essentially important conditional on the variable selected at the prior node; therefore containing additional information when compared to univariate tests.

Important (top-ranking) predictors from the Random Forests analysis were determined based on the distribution of the VI scores. Allele frequencies were compared between MS cases and controls, and ORs, 95% confidence intervals (95% CI) were determined using unconditional logistic regression models adjusted for gender and *HLA-DRB1*1501* as implemented in PLINK v1.06 (81).

Stage II – Replication Analysis

Genotypic data available from a recently published independent meta-analysis (three GWA scans) of 2,624 MS and 7,220 control subjects of European ancestry was used to confirm Stage I results (13). Briefly, all individuals were imputed on a single panel of 2.56 million SNPs, and a fixed-effects logistic model based on the observed and expected allele dosage, adjusted for cohort of origin (N=6), gender, and *HLA-DRB1*1501* using *xtlogit* implemented in STATA v9.2 (StataCorp LP, College Station, TX) (see De Jager et al) (13).

Stage III – Combined Analysis

A meta-analysis combining the genotypic data from Stage I and Stage II (3,967 MS and 8,599 control subjects) determined final effect size estimates using a fixed-effects logistic model based on the observed and expected allele dosage, adjusted for cohort of origin (N=7), gender, and *HLA-DRB1*1501* using STATA v9.2.

Stage IV – Extension Analysis

A LD analysis was conducted in CEU Hapmap data (version 3; Release #R2) to identify chromosomal regions tagged by the important predictors using Haploview v4.2 (44). Genetic variation within the tagged chromosomal regions were also investigated in Stage I, Stage II, and Stage III data sets using the appropriate logistic regression models as previously described. Logistic regression models were rerun to include the 17q21.31 inversion tagging variant (rs1396862) as an independent variable.

Using Stage I data, haplotype blocks were constructed using D' confidence intervals (43), and frequencies determined using Haploview v4.2. Haplotype frequencies were rounded to the nearest whole number and compared using Fisher's exact test (*cci*; STATA v9.2).

REFERENCES

1. Hauser SL, Oksenberg JR. The neurobiology of multiple sclerosis: genes, inflammation, and neurodegeneration. Neuron 2006;52:61-76.
2. Oksenberg JR, Barcellos LF. Multiple sclerosis genetics: leaving no stone unturned. Genes Immun 2005;6:375-87.
3. Oksenberg JR, Barcellos LF, Cree BA, et al. Mapping multiple sclerosis susceptibility to the HLA-DR locus in African Americans. Am J Hum Genet 2004;74:160-7.
4. Caillier SJ, Briggs F, Cree BA, et al. Uncoupling the roles of HLA-DRB1 and HLA-DRB5 genes in multiple sclerosis. J Immunol 2008;181:5473-80.
5. Dyment DA, Herrera BM, Cader MZ, et al. Complex interactions among MHC haplotypes in multiple sclerosis: susceptibility and resistance. Hum Mol Genet 2005;14:2019-26.
6. Barcellos LF, Sawcer S, Ramsay PP, et al. Heterogeneity at the HLA-DRB1 locus and risk for multiple sclerosis. Hum Mol Genet 2006;15:2813-24.
7. Lincoln MR, Ramagopalan SV, Chao MJ, et al. Epistasis among HLA-DRB1, HLA-DQA1, and HLA-DQB1 loci determines multiple sclerosis susceptibility. Proc Natl Acad Sci U S A 2009;106:7542-7.
8. Hafler DA, Compston A, Sawcer S, et al. Risk alleles for multiple sclerosis identified by a genomewide study. N Engl J Med 2007;357:851-62.
9. The International Multiple Sclerosis Genetics Consortium (IMSGC). Refining genetic associations in multiple sclerosis. Lancet Neurol 2008;7:567-9.
10. Rubio JP, Stankovich J, Field J, et al. Replication of KIAA0350, IL2RA, RPL5 and CD58 as multiple sclerosis susceptibility genes in Australians. Genes Immun 2008;9:624-30.
11. Perera D, Stankovich J, Butzkueven H, et al. Fine mapping of multiple sclerosis susceptibility genes provides evidence of allelic heterogeneity at the IL2RA locus. J Neuroimmunol 2009;211:105-9.
12. De Jager PL, Baecher-Allan C, Maier LM, et al. The role of the CD58 locus in multiple sclerosis. Proc Natl Acad Sci U S A 2009;106:5264-9.
13. De Jager PL, Jia X, Wang J, et al. Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. Nat Genet 2009;41:776-82.
14. The International Multiple Sclerosis Genetics Consortium (IMSGC). Comprehensive follow-up of the first genome-wide association study of multiple sclerosis identifies KIF21B and TMEM39A as susceptibility loci. Hum Mol Genet 2010;19:953-962.
15. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. Science 2008;322:881-8.
16. Bush WS, Sawcer SJ, de Jager PL, et al. Evidence for polygenic susceptibility to multiple sclerosis--the shape of things to come. Am J Hum Genet 2010;86:621-5.
17. Thomas DC. The need for a systematic approach to complex pathways in molecular epidemiology. Cancer Epidemiol Biomarkers Prev 2005;14:557-9.
18. Baranzini SE, Galwey NW, Wang J, et al. Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. Hum Mol Genet 2009;18:2078-90.
19. Miller DB, O'Callaghan JP. Neuroendocrine aspects of the response to stress. Metabolism 2002;51:5-10.

20. Webster Marketon JI, Glaser R. Stress hormones and immune function. Cell Immunol 2008;252:16-26.
21. Heesen C, Gold SM, Huitinga I, Reul JM. Stress and hypothalamic-pituitary-adrenal axis function in experimental autoimmune encephalomyelitis and multiple sclerosis - a review. Psychoneuroendocrinology 2007;32:604-18.
22. Ysrraelit MC, Gaitan MI, Lopez AS, Correale J. Impaired hypothalamic-pituitary-adrenal axis activity in patients with multiple sclerosis. Neurology 2008;71:1948-54.
23. Goodin DS, Ebers GC, Johnson KP, Rodriguez M, Sibley WA, Wolinsky JS. The relationship of MS to physical trauma and psychological stress: report of the Therapeutics and Technology Assessment Subcommittee of the American Academy of Neurology. Neurology 1999;52:1737-45.
24. Grant I, Brown GW, Harris T, McDonald WI, Patterson T, Trimble MR. Severely threatening events and marked life difficulties preceding onset or exacerbation of multiple sclerosis. J Neurol Neurosurg Psychiatry 1989;52:8-13.
25. Liu XJ, Ye HX, Li WP, Dai R, Chen D, Jin M. Relationship between psychosocial factors and onset of multiple sclerosis. Eur Neurol 2009;62:130-6.
26. Warren S, Greenhill S, Warren KG. Emotional stress and the development of multiple sclerosis: case-control evidence of a relationship. J Chronic Dis 1982;35:821-31.
27. Mohr DC, Goodkin DE, Bacchetti P, et al. Psychological stress and the subsequent appearance of new brain MRI lesions in MS. Neurology 2000;55:55-61.
28. Ackerman KD, Heyman R, Rabin BS, et al. Stressful life events precede exacerbations of multiple sclerosis. Psychosom Med 2002;64:916-20.
29. Buljevac D, Hop WC, Reedeker W, et al. Self reported stressful life events and exacerbations in multiple sclerosis: prospective study. Bmj 2003;327:646.
30. Brown RF, Tennant CC, Sharrock M, Hodgkinson S, Dunn SM, Pollard JD. Relationship between stress and relapse in multiple sclerosis: Part II. Direct and indirect relationships. Mult Scler 2006;12:465-75.
31. Golan D, Somer E, Dishon S, Cuzin-Disegni L, Miller A. Impact of exposure to war stress on exacerbations of multiple sclerosis. Ann Neurol 2008;64:143-8.
32. Mitsonis CI, Zervas IM, Mitropoulos PA, et al. The impact of stressful life events on risk of relapse in women with multiple sclerosis: a prospective study. Eur Psychiatry 2008;23:497-504.
33. Potagas C, Mitsonis C, Watier L, et al. Influence of anxiety and reported stressful life events on relapses in multiple sclerosis: a prospective study. Mult Scler 2008;14:1262-8.
34. Cutolo M, Straub RH. Stress as a risk factor in the pathogenesis of rheumatoid arthritis. Neuroimmunomodulation 2006;13:277-82.
35. Maunder RG, Levenstein S. The role of stress in the development and clinical course of inflammatory bowel disease: epidemiological evidence. Curr Mol Med 2008;8:247-52.
36. Stojanovich L, Marisavljevich D. Stress as a trigger of autoimmune disease. Autoimmun Rev 2008;7:209-13.
37. Karaiskos D, Mavragani CP, Makaroni S, et al. Stress, coping strategies and social support in patients with primary Sjogren's syndrome prior to disease onset: a retrospective case-control study. Ann Rheum Dis 2009;68:40-6.
38. Tomer Y, Huber A. The etiology of autoimmune thyroid disease: a story of genes and environment. J Autoimmun 2009;32:231-9.

39. Kim JE, Cho DH, Kim HS, et al. Expression of the corticotropin-releasing hormone-proopiomelanocortin axis in the various clinical types of psoriasis. Exp Dermatol 2007;16:104-9.
40. Dube SR, Fairweather D, Pearson WS, Felitti VJ, Anda RF, Croft JB. Cumulative childhood stress and autoimmune diseases in adults. Psychosom Med 2009;71:243-50.
41. Stefansson H, Helgason A, Thorleifsson G, et al. A common inversion under selection in Europeans. Nat Genet 2005;37:129-37.
42. Donnelly MP, Paschou P, Grigorenko E, et al. The Distribution and Most Recent Common Ancestor of the 17q21 Inversion in Humans. Am J Hum Genet.
43. Gabriel SB, Schaffner SF, Nguyen H, et al. The structure of haplotype blocks in the human genome. Science 2002;296:2225-9.
44. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2005;21:263-5.
45. GAMES and the Transatlantic Multiple Sclerosis Genetics Cooperative. A meta-analysis of whole genome linkage screens in multiple sclerosis. J Neuroimmunol 2003;143:39-46.
46. Hafler DA, Compston A, Sawcer S, et al. Risk alleles for multiple sclerosis identified by a genomewide study. New England Journal of Medicine 2007;357:851-62.
47. Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene). Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. Nat Genet 2009;41:824-8.
48. Baranzini SE, Wang J, Gibson RA, et al. Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. Hum. Mol. Genet. %R 10.1093/hmg/ddn388 2009;18:767-778.
49. Goris A, Maranian M, Walton A, et al. No evidence for association of a European-specific chromosome 17 inversion with multiple sclerosis. Eur J Hum Genet 2006;14:1064.
50. Borrelli E. A chilled-out knockout. Nat Genet 1998;19:108-9.
51. Papadimitriou A, Priftis KN. Regulation of the hypothalamic-pituitary-adrenal axis. Neuroimmunomodulation 2009;16:265-71.
52. Theoharides TC, Singh LK, Boucher W, et al. Corticotropin-releasing hormone induces skin mast cell degranulation and increased vascular permeability, a possible explanation for its proinflammatory effects. Endocrinology 1998;139:403-13.
53. Theoharides TC, Donelan JM, Papadopoulou N, Cao J, Kempuraj D, Conti P. Mast cells as targets of corticotropin-releasing factor and related peptides. Trends Pharmacol Sci 2004;25:563-8.
54. Cao J, Papadopoulou N, Kempuraj D, et al. Human mast cells express corticotropin-releasing hormone (CRH) receptors and CRH leads to selective secretion of vascular endothelial growth factor. J Immunol 2005;174:7665-75.
55. Agelaki S, Tsatsanis C, Gravanis A, Margioris AN. Corticotropin-releasing hormone augments proinflammatory cytokine production from macrophages in vitro and in lipopolysaccharide-induced endotoxin shock in mice. Infect Immun 2002;70:6068-74.
56. Baker C, Richards LJ, Dayan CM, Jessop DS. Corticotropin-releasing hormone immunoreactivity in human T and B cells and macrophages: colocalization with arginine vasopressin. J Neuroendocrinol 2003;15:1070-4.
57. Baigent SM. Peripheral corticotropin-releasing hormone and urocortin in the control of the immune response. Peptides 2001;22:809-20.

58. Goetzl EJ, Chan RC, Yadav M. Diverse mechanisms and consequences of immunoadoption of neuromediator systems. Ann N Y Acad Sci 2008;1144:56-60.
59. Singh VK, Fudenberg HH. Binding of [125I]corticotropin releasing factor to blood immunocytes and its reduction in Alzheimer's disease. Immunol Lett 1988;18:5-8.
60. Audhya T, Jain R, Hollander CS. Receptor-mediated immunomodulation by corticotropin-releasing factor. Cell Immunol 1991;134:77-84.
61. Mousa SA, Bopaiah CP, Stein C, Schafer M. Involvement of corticotropin-releasing hormone receptor subtypes 1 and 2 in peripheral opioid-mediated inhibition of inflammatory pain. Pain 2003;106:297-307.
62. McEvoy AN, Bresnihan B, FitzGerald O, Murphy EP. Corticotropin-releasing hormone signaling in synovial tissue from patients with early inflammatory arthritis is mediated by the type 1 alpha corticotropin-releasing hormone receptor. Arthritis Rheum 2001;44:1761-7.
63. Cureton EL, Ereso AQ, Victorino GP, et al. Local secretion of urocortin 1 promotes microvascular permeability during lipopolysaccharide-induced inflammation. Endocrinology 2009;150:5428-37.
64. Hsuchou H, Kastin AJ, Wu X, Tu H, Pan W. Corticotropin-releasing hormone receptor-1 in cerebral microvessels changes during development and influences urocortin transport across the blood-brain barrier. Endocrinology;151:1221-7.
65. Pisarchik A, Slominski AT. Alternative splicing of CRH-R1 receptors in human and mouse skin: identification of new variants and their differential expression. Faseb J 2001;15:2754-6.
66. Hillhouse EW, Grammatopoulos DK. The molecular mechanisms underlying the regulation of the biological activity of corticotropin-releasing hormone receptors: implications for physiology and pathophysiology. Endocr Rev 2006;27:260-86.
67. Pisarchik A, Slominski A. Molecular and functional characterization of novel CRFR1 isoforms from the skin. Eur J Biochem 2004;271:2821-30.
68. Calcagni E, Elenkov I. Stress system activity, innate and T helper cytokines, and susceptibility to immune-related diseases. Ann N Y Acad Sci 2006;1069:62-76.
69. Kern S, Ziemssen T. Brain-immune communication psychoneuroimmunology of multiple sclerosis. Mult Scler 2008;14:6-21.
70. Kemeny ME, Schedlowski M. Understanding the interaction between psychosocial stress and immune-related diseases: a stepwise progression. Brain Behav Immun 2007;21:1009-18.
71. Muller MB, Wurst W. Getting closer to affective disorders: the role of CRH receptor systems. Trends Mol Med 2004;10:409-15.
72. Liu Z, Zhu F, Wang G, et al. Association of corticotropin-releasing hormone receptor1 gene SNP and haplotype with major depression. Neurosci Lett 2006;404:358-62.
73. Wasserman D, Sokolowski M, Rozanov V, Wasserman J. The CRHR1 gene: a marker for suicidality in depressed males exposed to low stress. Genes Brain Behav 2008;7:14-9.
74. Gillespie CF, Phifer J, Bradley B, Ressler KJ. Risk and resilience: genetic and environmental influences on development of the stress response. Depress Anxiety 2009;26:984-92.
75. Steemers FJ, Chang W, Lee G, Barker DL, Shen R, Gunderson KL. Whole-genome genotyping with the single-base extension assay. Nat Methods 2006;3:31-3.

76. McDonald WI, Compston A, Edan G, et al. Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis. Ann Neurol 2001;50:121-7.
77. Thompson AJ, Montalban X, Barkhof F, et al. Diagnostic criteria for primary progressive multiple sclerosis: a position paper. Ann Neurol 2000;47:831-5.
78. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 2007;81:1084-97.
79. de Bakker PI, McVean G, Sabeti PC, et al. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat Genet 2006;38:1166-72.
80. Breiman L. Random Forests. Machine Learning 2001;45:5-32.
81. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559-75.

Table 1: *CRHR1* SNP variants on chromosome 17 identified as important predictors of MS by Random Forests in Stage I analysis

VI rank[1]	SNP	Base-pair Location	Function	Minor allele	Stage I				Stage II	
					MAF cases	MAF controls	*p*-value [2]	OR [3] (95% CI)	*p*-value	OR [4] (95% CI)
8	rs12940065	41,221,636	intron 1	A	0.16	0.17	0.42	0.94 (0.81-1.09)	--	--
3	rs34174655	41,231,188	intron 1	A	0.08	0.10	**0.0029**	0.74 (0.61-0.90)	--	--
2	rs171442	41,247,686	intron 2	A	0.06	0.08	**0.0087**	0.74 (0.59-0.93)	--	--
6	rs242939	41,251,360	intron 3	C	0.06	0.08	**0.016**	0.76 (0.61-0.95)	**0.0078**	0.82 (0.71-0.95)
1	rs242938	41,251,717	intron 3	A	0.06	0.08	**0.0076**	0.74 (0.60-0.92)	--	--
7	rs242936	41,254,990	intron 4	A	0.10	0.12	**0.022**	0.81 (0.68-0.97)	**0.0024**	0.82 (0.73-0.93)
4	rs173365	41,256,855	intron 4	A	0.43	0.46	**0.032**	0.89 (0.79-0.99)	**0.0029**	0.89 (0.83-0.96)
5	rs17689966	41,266,236	intron 8	G	0.43	0.46	0.055	0.90 (0.80-1.00)	**0.0030**	0.89 (0.83-0.96)

[1] Important (top-ranking) predictors from the Random Forests analysis were determined based on the distribution of the VI scores (Figure 2).
[2] Logistic regression model were the variant is coded to reflect genotypes (0, 1, and 2, where 0 represents homozygous dominant genotype) and *HLA-DRB1*1501* and gender were included in the model (PLINK v1.06).
[3] Adjusted proportional ORs from logistic regression models with *HLA-DRB1*1501* and gender each model (PLINK v1.06).
[4] Fixed-effects meta-analysis of three GWA scans for MS susceptibility based on the observed (imputed) and expected alleles dosage of each SNP, taking into account the empirically observed variance of the allele dosage, was conducted using logistic regression models, adjusted for location, *HLA-DRB1*1501* and gender (STATA v9.0).

Table 2: Replicated SNP variants (Table 1) in the combined data set.

SNP	Chr	Base-pair Location	Gene	Function	Minor allele	Stage III	
						***p*-value**	**OR [1] (95% CI)**
rs242939	17	41,251,360	*CRHR1*	intron 3	C	4.1×10^{-4}	0.81 (0.72-0.91)
rs242936	17	41,254,990	*CRHR1*	intron 4	A	9.7×10^{-5}	0.82 (0.74-0.90)
rs173365	17	41,256,855	*CRHR1*	intron 4	A	1.9×10^{-4}	0.89 (0.84-0.95)
rs17689966	17	41,266,236	*CRHR1*	intron 8	G	3.0×10^{-4}	0.89 (0.84-0.95)

[1] Fixed-effects meta-analysis of three GWA scans for MS susceptibility based on the observed (imputed) and expected alleles dosage of each SNP, taking into account the empirically observed variance of the allele dosage, was conducted using logistic regression models, adjusted for location, *HLA-DRB1*1501* and gender (STATA v9.0).

Table 3: Association of *CRHR1* H1 haplotypes with MS [1].

Haplotypes	Case Counts	Control Counts	*p*-value [2]	OR [3] (95% CI)
H1.1	410.7	380.4	0.190	1.11 (0.95-1.30)
H1.2	361.5	359.7	0.869	1.02 (0.86-1.20)
H1.3	330.0	304.0	0.226	1.11 (0.94-1.32)
H1.4	235.5	262.0	0.251	0.90 (0.74-1.08)
H1.5	115.8	175.6	**4.1×10^{-4}**	0.64 (0.50-0.83)
H1.6	125.3	123.4	0.896	1.03 (0.79-1.34)
H1.7	95.0	94.1	0.941	1.02 (0.75-1.38)
H1.8	84.8	78.2	0.576	1.10 (0.80-1.53)
H1.9	71.8	83.4	0.413	0.87 (0.62-1.21)
H1.10	52.9	42.6	0.303	1.25 (0.81-1.92)
H1.11	31.0	24.2	0.345	1.31 (0.74-2.34)
Total H1	1993.4	2008.7		

[1] Haplotype counts were determined by summing the fractional likelihoods of each individual haplotype (Haploview v4.2). H1 haplotypes were determined by the absence of rs1396862T (Figure 5); sub-haplotypes with a frequency $\geq$1% are shown. The nomenclature of sub-haplotypes was determined by the frequency in the total population, e.g. the most frequent H1 sub-haplotype was named H1.1.

[2] Significance was determined using Fisher's exact test (Stata v9.2).

[3] Reference is all other H1 haplotypes.

Figure 1: Summary of analytical approach to investigate HPA axis related genes in MS.

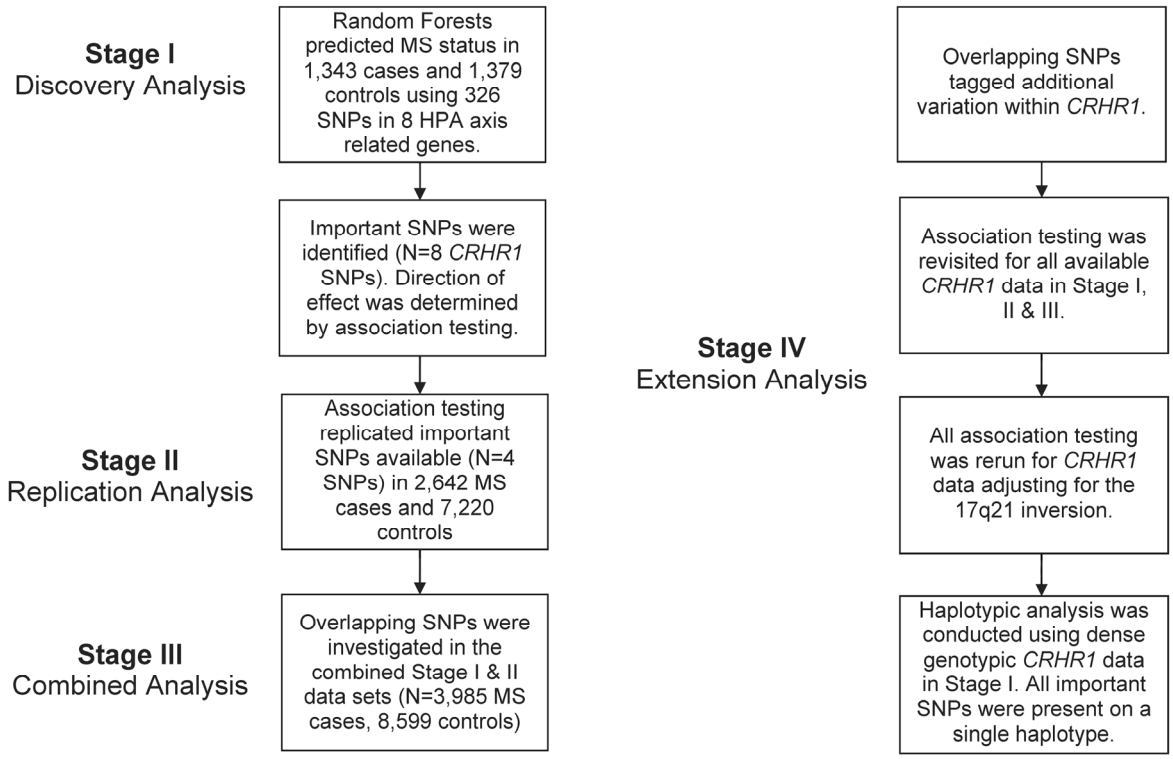
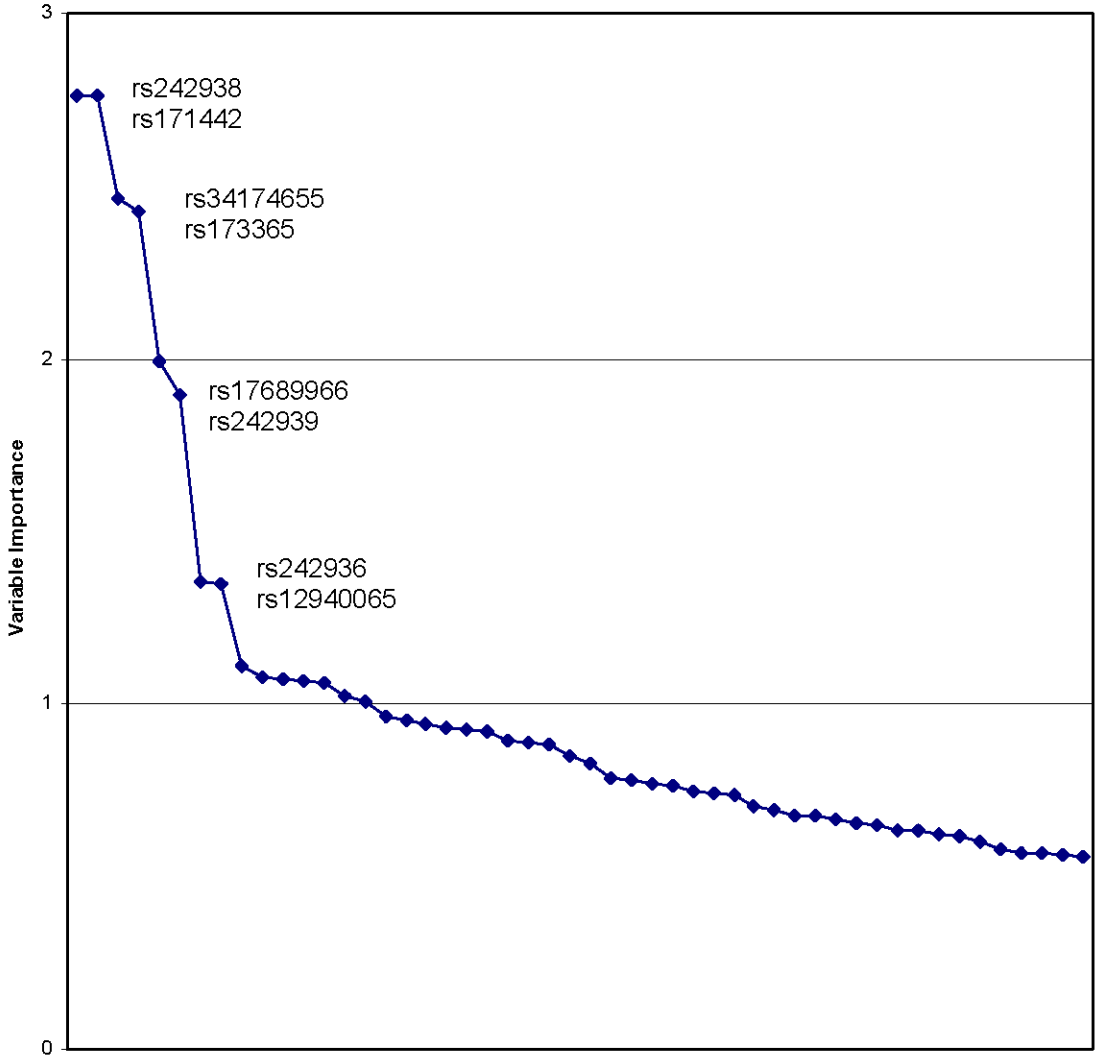
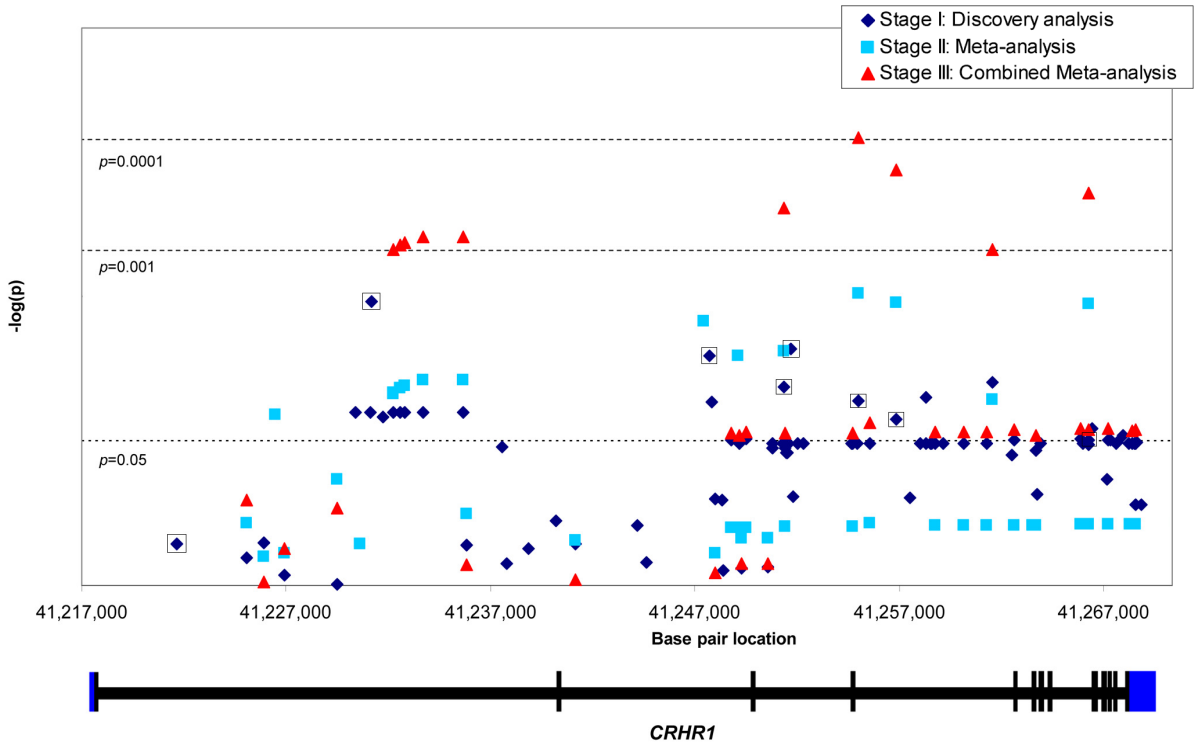


Figure 2: Variable Importance scores for the top 50 predictors for the Random Forests analysis (Stage I) where all genetic variants were used to predict MS in 1,343 cases and 1,379 healthy controls [1].



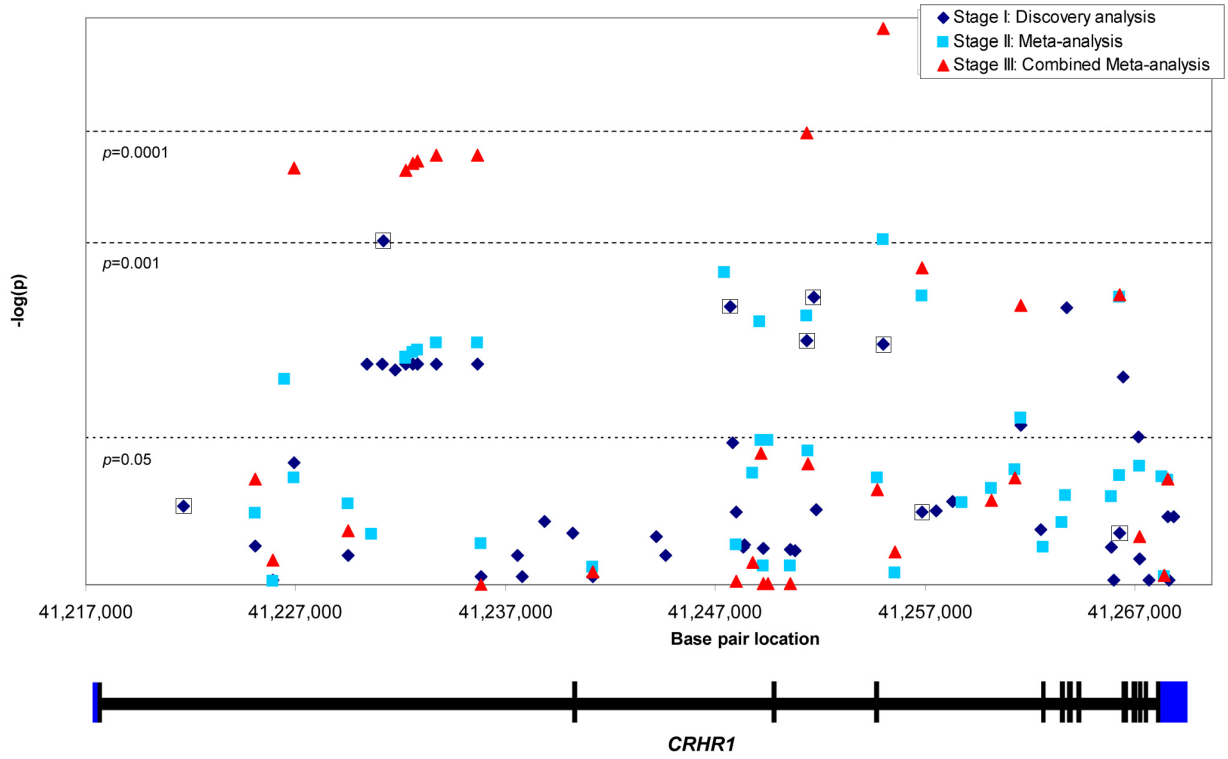
[1] Random Forests v6.40.179 (http://salford-systems.com/) with *mtry*=sqrt(p) and *ntree*=5,000 was used. All SNP genotypes in cases and controls were coded as 0, 1 or 2 copies of the minor allele.

Figure 3: Visualization of significance values (-log(*p*)) for all *CRHR1* variants available in both study populations [1].



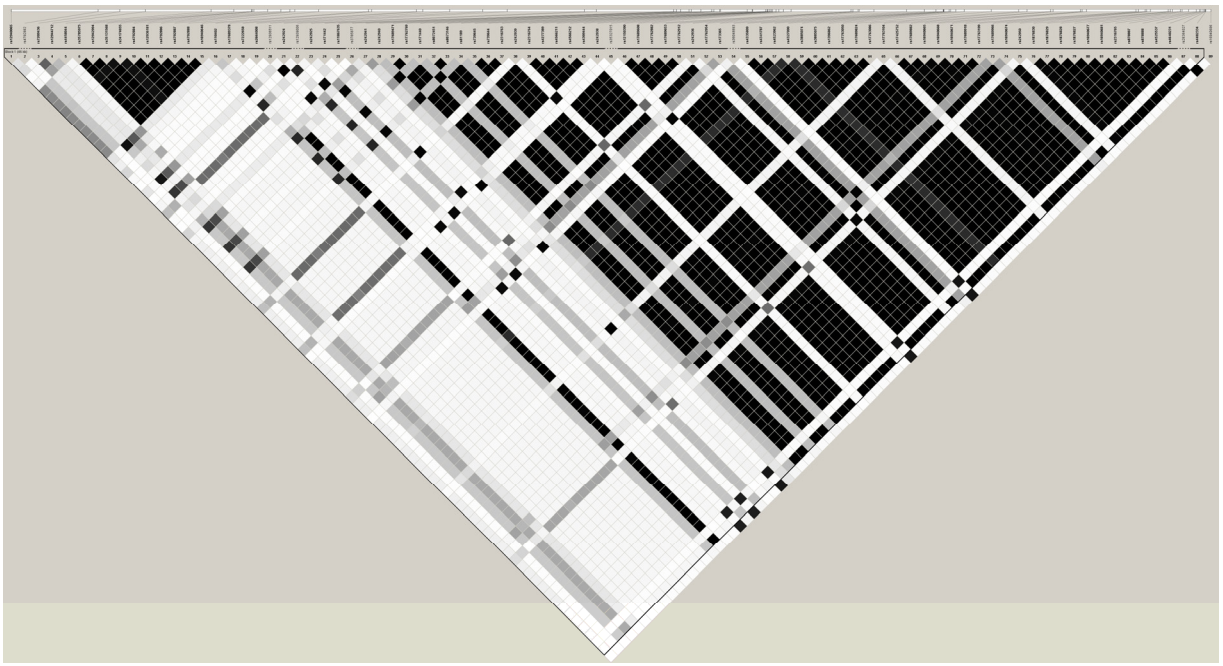
[1] A total of 89 *CRHR1* SNP variants were available in the discovery dataset (Stage I: diamonds); 43 *CRHR1* SNP variants in the independent meta-analysis (Stage II: squares); and 35 *CRHR1* SNPs variants in the combined meta-analysis (Stage III: triangles). Uncorrected significance values are based on logistic regressions models respective to each analysis (see Materials and Methods). SNP variants identified as important by Random Forests in Stage I are outlined by squares. Blue blocks represent untranslated regions and black blocks are exons (13 blocks representing 14 exons).

Figure 4: Visualization of significance values (-log(*p*)) for all *CRHR1* SNP variants available in both study populations, adjusted for the presence of the 17.q21.31 inversion (tagging SNP rs1396862) [1].



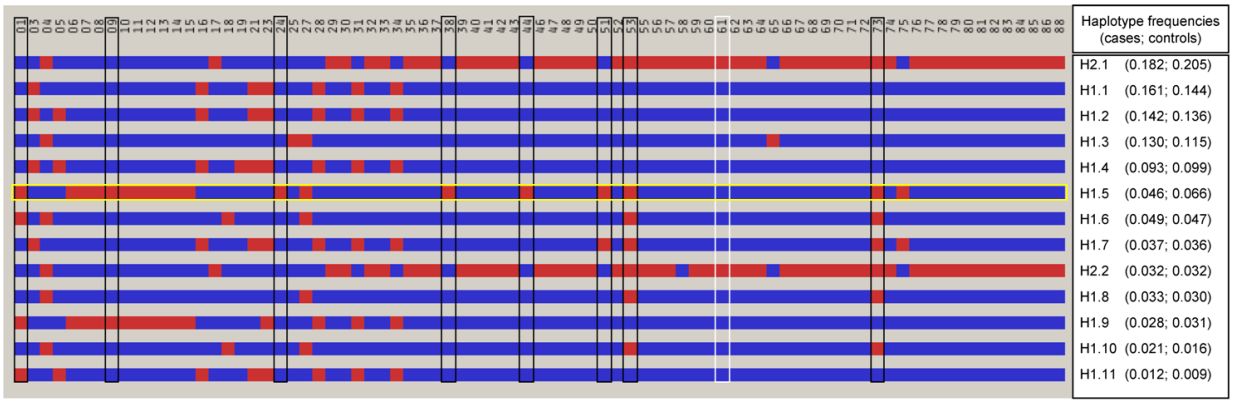
[1] Adjusted logistic regression models were performed for 51 SNP variants in the discovery dataset (Stage I: diamonds); 43 SNP variants in the independent meta-analysis (Stage II: squares); and 30 SNPs variants in the combined meta-analysis (Stage III: triangles). Uncorrected significance values are based on logistic regressions models respective to each analysis, with the inclusion of the 17q21.31 inversion-tagging SNP variant (rs1396862; see Materials and Methods). SNP variants identified as important by Random Forests in Stage I are outlined by squares. Blue blocks represent untranslated regions and black blocks are exons (13 blocks representing 14 exons).

Figure 5. Linkage disequilibrium pattern of *CRHR1* SNPs in 1,343 MS cases and 1,379 controls [1].



[1] Haplotype block was determined using D' confidence intervals in Haploview v4.2. Pairwise r^2 are color coded to illustrate LD on a monochromatic scale, where $r^2=1$ is shaded black.

Figure 6: Visualization of *CRHR1* haplotypes with ≥1% frequency in 1,343 MS cases and 1,379 controls of European ancestry [1].



[1] Major alleles are colored blue; minor alleles are colored red. H1 and H2 haplotypes were determined by the absence or presence of rs1396862 minor allele (SNP #61 outlined in white). Sub-haplotype nomenclature was determined by frequency, with most frequent sub-haplotypes named H1.1 and H2.1. The 8 SNPs identified by Random Forests are vertically outlined in black; note that the minor alleles for all 8 SNPs are on one sub-haplotype: H1.5 (horizontally outlined in yellow).

Chapter 3

Supervised Machine Learning and Logistic Regression Identifies Novel Epistatic Risk Factors with *PTPN22* for Rheumatoid Arthritis.

Published: *Genes and Immunity* (21 January 2010) doi:10.1038/gene.2009.110

ABSTRACT
Investigating genetic interactions (epistasis) has proven difficult despite the recent advances of both laboratory methods and statistical developments. With no 'best' statistical approach available, combining several analytical methods may be optimal for detecting epistatic interactions. Using a multi-stage analysis that incorporated supervised machine learning and methods of association testing, we investigated epistatic interactions with a well-established genetic factor (*PTPN22* 1858T) in a complex autoimmune disease (rheumatoid arthritis [RA]). Our analysis consisted of four principal stages: Stage I (data reduction) – identifying candidate chromosomal regions in 292 affected sibling pairs, by predicting *PTPN22* concordance using multipoint identity-by-descent probabilities and a supervised machine learning algorithm (Random Forests); Stage II (extension analysis) – testing detailed genetic data within candidate chromosomal regions for epistasis with *PTPN22* 1858T in 677 cases and 750 controls using logistic regression; Stage III (replication analysis) – confirmation of epistatic interactions in 947 cases and 1,756 controls; Stage IV (combined analysis) – a pooled analysis including all 1,624 RA cases and 2,506 control subjects for final estimates of effect size. A total of 7 replicating epistatic interactions were identified. SNP variants within *CDH13*, *MYO3A*, *CEP72* and near *WFDC1* demonstrated significant evidence for interaction with *PTPN22*, affecting susceptibility to RA.

INTRODUCTION

Genome-wide association (GWA) studies, which provide the ability to simultaneously investigate hundreds of thousands of genetic markers in large numbers of individuals, have successfully led to the discovery of genetic risk factors with modest effects in several complex diseases, including autoimmune diseases (1, 2). Nevertheless, it is apparent that current approaches to genetic analysis, which include almost exclusively, marginal associations using a univariate approach, are not able to identify a substantial fraction of the genetic burden. This may reflect the involvement of rare variants, copy number variation, *gene* x *gene* interactions, *gene* x *environment* interactions and/or epigenetic mechanisms. Currently there is no consensus regarding appropriate approaches for evaluating these components of complex diseases.

Investigating genetic or *gene* x *gene* interactions (also known as 'epistasis', where the action of one gene is modified by one or several other genes) has proven difficult, despite recent advances of both laboratory methods and statistical developments. For example, the 15th biennial Genetic Analysis Workshop (GAW15) investigated genetic interactions in rheumatoid arthritis (RA [MIM 180300]) using several data sets. A variety of statistical approaches were used to investigate epistasis in RA in both family and population-based data sets with varying genetic marker density; results varied greatly, demonstrating that robust and comprehensive approaches not restricted by a small sample size or sparse data are necessary (3-8). With no 'best' statistical approach available, combining several analytical methods may be optimal for detecting epistatic interactions (9, 10). Here, we performed a comprehensive multi-stage genetic investigation with a replication analysis to reveal epistatic relationships, involving the well-established *PTPN22* (protein tyrosine phosphatase non-receptor 22; GeneID 26191) risk variant, that alter susceptibility to RA.

RA is a chronic multi-system autoimmune disease, resulting from persistent inflammatory synovitis and subsequent erosion of the joint architecture. It is considered a complex disease with a multi-factorial etiology influenced by both genetic and environmental risk factors. Genetic predisposition to RA is suspected to involve multiple genes with incomplete penetrance, interacting in a variety of biological pathways (11). A large number of previous studies have investigated the role of genetic factors in RA. The *HLA-DRB1* (GeneID 3123) shared epitope (SE; *HLA-DRB1* alleles: 0101, 0102, 0401, 0404, 0405, 0408, 0413, 1001, and 1402) and *PTPN22* 1858T alleles have been consistently associated with greater risk (12-14). Similar to *HLA-DRB1*, *PTPN22* is a strong biological candidate as it regulates T-cell receptor (TCR) signaling and confers risk for multiple autoimmune diseases, though there is limited knowledge regarding the exact etiological mechanism(s) (15, 16).

Previous genome-wide linkage scans in RA identified evidence for additional loci, however uncovering the susceptibility genes within these regions has been difficult (17, 18). Recent candidate gene and GWA studies have reported strong evidence for several genes with modest contributions to RA susceptibility, including *CTLA4*, *PADI4*, *REL*, *STAT4*, and *TRAF1-C5*, as well as loci within chromosomal regions 6q23, 10p15, 12q13 and 22q13 (19-26). Multiplex interactions between genetic and environmental risk factors (*gene* x *gene*, *gene* x *environment*, *gene(n)* x *environment*) have been implicated in RA, however the scope of previous studies has been restricted primarily to candidate risk factors (14, 27-30). In order to understand the etiologic mechanisms involved in RA, a clear understanding of the intricate genetic architecture

underlying susceptibility, including the complete identification of epistatic relationships between genes, particularly those that do not demonstrate independent, or marginal, disease associations, is necessary. Using a multi-stage strategy framed within a genome-wide perspective, we identified evidence for novel epistatic risk factors interacting with *PTPN22* to influence RA susceptibility.

RESULTS

We investigated epistatic interactions with the *PTPN22* 1858T risk variant (rs2476601) in RA cases (available through the North American Rheumatoid Arthritis Consortium [NARAC] and who were predominantly positive for the presence of antibodies to cyclic citrullinated peptide [anti-CCP]; Table 1) and healthy controls of European origin using four principal stages: Stage I (data reduction) – identification of relevant chromosomal regions in 292 affected sibling pairs (ASPs) using Random Forests, a supervised machine learning algorithm; Stage II (extension analysis) – testing for epistasis with *PTPN22* 1858T in 677 cases and 750 controls; Stage III (replication analysis) – confirmation of epistatic interactions in 947 cases and 1,756 controls; Stage IV (combined analysis) – a combined analysis including all 1,624 RA cases and 2,506 control subjects to determine final estimates of effect size (Figure 1).

Stage I (data reduction)

A non-parametric approach using a supervised machine learning algorithm (Random Forests) and multipoint identity-by-descent (mIBD) probabilities identified promising regions harboring epistatic candidates for *PTPN22* in 292 ASPs from 512 multiplex RA families. This case-only analysis investigated mIBD probabilities for 379 microsatellite markers (MSMs) from across the genome (~10 cM coverage), as well as gender and *HLA-SE*, were used to predict *PTPN22* 1858T concordance in ASPs using the Random Forests algorithm (see Materials and methods). Random Forests assigns each predictor a variable importance (VI) score, with more *important* predictors having greater values. Based on the distribution of the VI scores, the five top-ranking predictors were considered *important* predictors of *PTPN22* concordance in the ASPs (see Supplementary Figure 1), as there appeared to be a clear distinction between the VI scores for these predictors compared to the remaining predictors. The five *important* predictors were mIBD probabilities for MSMs in the following chromosomal regions: 4q34, 5p15, 10p11, 14q23 and 16q23 (Table 2).

Stage II (extension analysis)

We investigated *important* chromosomal regions through association tests in individuals with at least 90% Northern European ancestry (677 cases; 750 healthy controls) from the collective NARAC I data set (908 cases; 1,260 healthy controls) for which dense genome-wide SNP data were available (Table 1). We evaluated our power to detect *gene* x *gene* interactions (see Materials and methods) and were sufficiently powered to investigate interactions under a dominant mode of inheritance. The presence of the *PTPN22* 1858T risk variant was significantly associated with RA in Caucasians when using the full NARAC I data set (odds ratio [OR]=2.09, 95% CI: 1.69-2.60), as previously described (22). Similar results were observed when the analysis was restricted to individuals with at least 90% Northern European ancestry (OR=1.89, 95% CI: 1.45-2.46). A total of 10,589 SNPs within 5Mb of an *important* locus satisfied quality control criteria (see Materials and methods; Table 2) and were investigated using the Breslow-Day (BD) test for homogeneity. A total of 665 SNPs significantly modified the effect of *PTPN22* on susceptibility to RA (p-value$_{BD}$<0.05). Under a dominant model, 449 SNPs tested for epistasis

(logistic regression) with *PTPN22* demonstrated significant (p-value$_I$<0.05; see Materials and methods) evidence for interaction in the NARAC I data set.

Stage III (replication analysis)

The replication analysis was performed in the NARAC II data set (947 cases and 1756 controls of European ancestry). The presence of the *PTPN22* 1858T risk variant was also significantly associated with RA in the replication data set as expected (OR=1.81, 95% CI: 1.49-2.20); results for the NARAC II cases have been previously reported (26, 31). Four hundred forty-seven of the 449 SNPs that demonstrated evidence for epistasis (Stage II) met quality control criteria and were formally tested for epistasis with *PTPN22* in the NARAC II data set, under a dominant model. A total of 7 epistatic relationships were replicated (p-value$_I$<0.05; Table 3). Individuals were stratified as carriers or non-carriers for the minor allele of each replicating SNP to characterize the relationship of *PTPN22* 1858T with RA susceptibility; consistent associations were observed across both Stage II (NARAC I) and Stage III (NARAC II) populations (Table 3).

Stage IV (combined analysis)

To determine final estimates of epistasis, replicated findings (N=7) were investigated in the combined analysis of the data sets from Stage II (NARAC I) and III (NARAC II), and included 1,624 RA cases and 2,506 healthy controls; minor allele frequencies for these SNPs did not differ between data sets demonstrating no evidence for heterogeneity (data not shown) (Table 4). The effect of *PTPN22* was significantly increased amongst individuals who were carriers of the minor allele for four SNP variants versus non-carriers, including: an intronic *CDH13* variant (rs1895535 AA/AG: OR=3.95, 95% CI: 2.39-6.55, p=1.0 x 10^{-7}; rs1895535 GG: OR=1.69, 95% CI: 1.44-1.98, p<1 x 10^{-8}); an intronic *MYO3A* variant (rs12573019 AA/AG: OR=2.78, 95% CI: 2.05-3.76, p<1 x 10^{-8}; rs12573019 GG: OR =1.60, 95% CI: 1.34-1.90; p=1.7 x 10^{-7}); and two SNPs with no known function (1. rs4695888 CT/TT: OR=2.10, 95% CI: 1.76-2.51, p<1 x 10^{-8}; rs4695888 CC: OR=1.22, 95% CI: 0.91-1.65, p=0.18; and 2. rs1168587 CT/TT: OR=2.13, 95% CI: 1.78-2.56, p<1 x 10^{-8}; rs1168587 CC: OR=1.29, 95% CI: 0.98-1.70, p=0.07). The effect of *PTPN22* was reduced in carriers of three SNP variants versus non-carriers, including an intronic *CDH13* variant (rs7200573 AA/AC: OR=1.37, 95% CI: 1.11-1.70, p=0.0034; rs7200573 CC: OR=2.47, 95% CI: 1.99-3.06, p<1 x 10^{-8}); a variant 7kb upstream *WFDC1* (rs11865624 CC/CT: OR=1.01, 95% CI: 0.66-1.55, p=0.95; rs11865624 TT: OR=2.01, 95% CI: 1.71-2.37, p=1 x 10^{-8}); and an intronic *CEP72* variant (rs7726839 AG/GG: OR=1.46, 95% CI: 1.16-1.82, p=0.0011; rs7726839 GG: OR=2.22, 95% CI: 1.81-2.73, p< 1 x 10^{-8}) (Table 4).

DISCUSSION

We investigated a role for epistasis in RA, a complex autoimmune disease, between *PTPN22* and other variants across the genome using large, well-characterized study populations with detailed clinical and genetic information. A multi-stage analysis, with extension and replication, that combined robust non-parametric and parametric methods was utilized. We report evidence for novel epistatic interactions with *PTPN22* in RA for variants within *CDH13* (GeneID 1012), *MYO3A* (GeneID 53904), *CEP72* (GeneID 55722) and near *WFDC1* (GeneID 58189). Interestingly, when RA cases were compared to controls, none of the final SNP variants demonstrated significant associations with RA (p<0.05; data not shown); that is, marginal effects in the absence of *PTPN22* were not present for any SNP.

PTPN22 is the second strongest known genetic risk factor for RA, and confers risk in several other autoimmune diseases (15, 16). Located on chromosome 1p13.3-13.1, *PTPN22* encodes a non-receptor classical class I tyrosine protein, lymphoid tyrosine phosphatase (LYP), which negatively regulates the TCR signaling by dephosphorylating several molecules (i.e. Src family kinases) immediately downstream of the TCR (32, 33). The 1858T risk variant replaces the amino acid at position 620 from an arginine to a tryptophan, resulting in a gain of function mutation that increases the capacity of LYP to negatively regulate TCR signaling (34). There is also evidence suggesting that the 1858T risk variant impairs B-cell receptor signaling and subsequent proliferation (35-37). Given the importance of *PTPN22* in RA and autoimmunity, the current evidence for statistical interactions which appear to modify RA risk conferred by variation within *PTPN22* sheds new light on potential biological mechanisms for future genetic and molecular investigations. Indeed, epistatic relationships observed in the current study may also be relevant to other autoimmune diseases.

We identified and replicated 7 epistatic interactions with *PTPN22* under a dominant genetic model. In this analysis *CDH13*, on chromosome 16, demonstrated the strongest epistatic relationship with *PTPN22* in RA. Two *CDH13* SNP variants approximately 350 kb apart (rs7200573 and rs1895535) showed significant evidence for interaction with *PTPN22*. For rs7200573, approximately 14 kb from exon 8, *PTPN22* conferred increased risk of RA in both carriers and non-carriers of rs7200573A; however *PTPN22* risk was significantly (p-value$_I$=0.00015) less in carriers (OR=1.37) relative to non-carriers (OR=2.47). Interestingly, rs1895535, approximately 22 kb from exon 5, demonstrated the opposite association for *PTPN22* and RA risk; *PTPN22* risk was significantly (p-value$_I$=0.0016) greater in rs1895535A carriers (OR=3.95) relative to non-carriers (OR=1.69). This variation in association suggests that these risk variants exist on separate haplotypes; r^2 between the two SNPs was <0.1. Unfortunately, SNP variants within exons 5 and 8 of *CDH13* have not been identified, and in available CEPH HapMap data (http://hapmap.org), the linkage disequilibrium (LD) between the *CDH13* epistatic risk variants and other SNP variants near the respective exons is also low (r^2<0.1). There was also evidence supporting epistasis for an intronic *MYO3A* SNP variant (rs12573019), an intronic *CEP72* SNP variant (rs7726839), and a SNP variant 7 kb upstream of *WFDC1* (rs11865624), on chromosomes 10, 5 and 16, respectively. *PTPN22* associated risk for RA was significantly increased in carriers of the *MYO3A* minor allele, while *PTPN22* risk was significantly increased in non-carriers of the *CEP72* and *WFDC1* minor alleles. An investigation of the underlying haplotype block structure for the epistatic risk variants in CEPH, suggests that the identified variants may be tagging functional variants within their respective genes. For example, the *MYO3A* variant, between exons 9 and 10 (~2.2 kb range), occurs within a large 150 kb haplotype block spanning 22 exons; the *CEP72* risk variant, located in the first intron, exists within a 38 kb haplotype block that extends through the entire gene; and the risk variant near *WFDC1* has modest LD (r^2=0.2-0.3) with several SNP variants within the 5' untranslated region and first intron of *WFDC1*.

Detailed functional data are not available for these epistatic candidates; however there is sufficient evidence to suggest a plausible biological relationship between these loci and RA. For example, *CDH13* (cadherin 13; T-cadherin (truncated); H-cadherin (heart)) encodes a unique cadherin lacking transmembrane and cytosolic domains necessary for homophilic adhesive activity of classical cadherins (38). T-cadherin is likely involved in signal transduction and not

cell-cell adhesion, as it concentrates in lipid raft domains of the plasma membrane, affects cellular migration, angiogenesis, survival under oxidative stress, and contributes to the invasive potential of various cancers (39-44). Interestingly, *CDH13* has also been associated with attention-deficit/hyperactivity disorder, blood pressure, and adult height (45-47).

Additionally, *MYO3A* (myosin IIIA) encodes a unique myosin motor protein that contains an N-terminal kinase domain and is primarily expressed within the retina and inner ear of vertebrates (48). Mutations within the motor domain of *MYO3A* results in nonsyndromic hearing loss DFNB30 (49). Both sensorineural and conductive hearing loss have increased prevalence in RA patients (50-52). Furthermore, *MYO3A* influences sterocillia shape and length, and is capable of inducing filopodial actin protrusions in culture cells, and thus may have an unknown function in immune cell locomotion similar to *MYO2A* (53, 54).

CEP72 (centrosomal protein 72kb) encodes a centromere protein that is critical for chromosomal alignment and proper tension generation between sister chromatids during mitosis (55). Interestingly, anti-centromere antibodies are primarily observed in patients with CREST syndrome, but have also been observed in other autoimmune diseases (56-59).

Another candidate identified here, *WFDC1* (whey acidic protein four-disulphide core domain 1), encodes ps20 and has a highly conserved core domain (60). WFDC1/ps20 is a multifunctional protein that facilitates endothelial cell motility and angiogenesis, inhibits cell proliferation, and promotes cellular senescence (60-62). CD4 T-cells normally express ps20 after restimulation and IL2 expansion (63). ps20 also increases CD4 T-cell permissiveness to HIV spread via CD54 integrin expression, and identifies a subset of CD4 memory T-cells at an early differentiation stage (CD45RO+/CD28+/CD57-) (63). Despite the lack of detailed experimental data, there is plausible statistical and biological evidence to support further investigation of these candidates. Our results underscore important lessons derived from recent GWA studies, including findings that most replicated associations do not involve previous candidate genes, suggesting new biological hypotheses, and that many have implicated non–protein coding regions (64).

In this analysis, we investigated epistatic interactions with *PTPN22* in RA using a multi-stage approach that combined robust non-parametric and parametric methods. There are several clear advantages to our methodology: (1) Random Forests is a model-free approach, which attaches a measure of importance to each predictor. The VI reported by Random Forests incorporates additional information compared to a univariate test for a predictor, as it reflects both the individual effect and the possible effect through multiplex interactions. In this analysis, we applied Random Forests to an ASP data set that was sufficiently powered to detect marginal genetic associations with OR$\leq$0.5 and $\geq$2.0 for a dominant model (data not shown). (2) We utilized conventional logistic regression models to test for epistasis assuming multiplicative interaction, which are readily interpretable. (3) We included a replication analysis to confirm epistatic interactions. In addition, we utilized well-defined study populations with detailed clinical information. The principal limitation in this analysis is our interpretation of the Random Forests results in Stage I. First, there is no clear standard for identifying an *important* variable; we based our selection on the empirical distribution of the VI scores and subjectively determined that the there was a clear distinction in the VI for the five top-ranking predictors versus all others (see Supplementary Figure 1). Second, the VI potentially includes the effect of multiplex

interactions between the predictors, as each variable selected at a node is essentially *important* conditional on the variable selected at the prior node in a single tree; however, given that VI scores are based on the forest, it is unknown how heirachial relationships are ranked. Multiplex genetic interactions were not explored in this analysis (i.e. three-way or other higher order interactions), which may explain why epistatic interactions were detected in four of the five *important* regions. Furthermore, the imposition of a dominant genetic model for epistatic candidates might not have been the appropriate assumption to follow-up the Random Forests findings.

The *HLA-DRB1* shared epitope (SE) is a critical genetic component of RA in Northern European Caucasians, conferring approximately 30-50% of the genetic risk, which suggests an important role for antigen presentation and subsequent T-cell activation in RA pathogenesis; recent experimental data suggests that the SE triggers pro-oxidant signaling and an innate immune response (65-68). Given the significance of *HLA-DRB1* in RA, we specifically included *HLA-DRB1* and a dense set of MSMs across chromosome 6p21, which provided extensive coverage of the major histocompatibility complex region, as predictors in Stage I of the analysis. Interestingly, Random Forests did not identify any chromosome 6p21 marker, including the *HLA-DRB1* locus, as an *important* predictor of *PTPN22* carrier status in the ASP investigation in Stage I.z Results did not change when SE status, specifically, was considered in ASPs (data not shown); therefore we did not further investigate the relationship between *HLA-DRB1* and *PTPN22*. Our results are in strong agreement with several recent studies (69-71), but in contrast with others (14, 72). We do acknowledge that a large proportion of the ASPs were positive for HLA-SE. Additionally, the RA cases in Stage II and III were anti-CCP positive, the phenotypic subgroup for which the classical *HLA-DRB1* association has been exclusively established (73).

In summary, results from the current study show the genetic contribution to RA risk is more complex than originally considered. Here, we identified novel candidate genes (*CDH13*, *MYO3A*, *CEP72*, *WFDC1*) that modify the effect of *PTPN22* associated risk for RA, and provide an important framework for future studies of *gene* x *gene* interactions. Genetic studies in complex disease must include application of multi-analytical strategies. Efforts to explore higher-order interactions are needed and will require very large sample sizes and clearly defined phenotypes.

MATERIALS AND METHODS

Informed consent was obtained from all participants and approvals from local institutional review boards were secured at each recruitment site prior to enrollment. All RA cases satisfied the American College of Rheumatology (ACR) 1987 classification criteria for RA (74).

Study Population

Stage I. A total of 292 ASPs from 512 multiplex RA families recruited by the NARAC were used as previously described (Table 1) (12). Briefly, families were eligible if: ≥ two siblings satisfied the ACR 1987 criteria for RA (74), ≥ one sibling had erosions on hand radiographs, and ≥ one sibling had disease onset between the ages of 18 and 60 years. Families were excluded from participation if other diseases associated with similar articular symptoms were present.

Greater than 90% of RA cases were positive for the presence of anti-CCP antibody, an antibody that is highly specific to RA (75); 82.1% were positive for rheumatoid factor (data not shown).

Stage II. A collective data set of 908 RA cases was used in the extension analysis, which included 464 unrelated probands from the NARAC ASP families (described above; of which 127 probands overlapped with Stage I, results did not vary when overlapping probands were excluded), 168 RA cases from the National Data Bank for Rheumatic Diseases, 162 RA cases from the National Inception Cohort of Rheumatoid Arthritis and 114 RA cases from the Study of New Onset Rheumatoid Arthritis, and 1,260 healthy control subjects from the New York Health Project, as previously described (Table 1) (22, 76-79). All RA cases were positive for the presence of anti-CCP antibody; 85.9% were positive for rheumatoid factor (data not shown). The study was then restricted to subjects with at least 90% Northern European ancestry (see Laboratory Procedures) to avoid bias due to population stratification (N=682 RA cases, 752 controls). Similarly, all RA cases in this group were positive for anti-CCP; 85.0% were positive for rheumatoid factor (data not shown). Subjects were also excluded if genotype data for SNP rs2476601 (*PTPN22* 1858T) was missing. The final data set for extension analysis was comprised of 677 RA cases and 750 healthy controls (Table 1).

Stage III. The NARAC II data set used in the replication analysis consists of 952 RA cases, which included 175 RA probands from the NARAC family studies, 332 RA cases from the Veterans Affairs Rheumatoid Arthritis Registry, 160 RA cases from the Studies of the Etiologies of Rheumatoid Arthritis cohort, 105 RA probands as members of the Multiple Autoimmune Disease Genetics Consortium, 86 RA patients from the UCSF Rheumatoid Arthritis Genetics Project and 94 RA patients from the Early Rheumatoid Arthritis Treatment Evaluation Registry and 1,760 control subjects, as previously described (Table 1) (12, 22, 26, 31, 80-82). Patients for whom anti-CCP data were available were all anti-CCP positive (data was not available for 31 Veterans Affairs Rheumatoid Arthritis Registry subjects). Control data were taken from publicly available control data sets in the Illumina iControl database (Illumina, San Diego, CA; http://www.illumina.com/iControlDB) and the neurodevelopmental control group obtained from the NIH Laboratory of Neurogenetics (http://neurogenetics.nia.nih.gov/paperdata/public/). The control genotypes were selected from the entire set of European American genotypes available in these resources based on the following data filters: 1) >90% complete genotyping data; 2) >90% European continental ancestry. The European continental ancestry was determined using ancestry informative markers previously described (83). Subjects were also excluded if genotype data for SNP rs2476601 (*PTPN22* 1858T) was missing. The final data set for the replication analysis was comprised of 947 RA cases and 1,756 healthy controls (Table 1). There was no overlap between individuals in the replication dataset, and those used in both Stage I and Stage II analyses.

Laboratory Procedures

Stage I. RA ASPs were genotyped for *PTPN22* 1858T (rs2476601), the *HLA-DRB1* locus and 379 MSMs from the Marshfield Set 8A Combo List (http://research.marshfieldclinic.org) with additional MSMs in specific chromosomal regions (i.e. the HLA complex) as previously described (12, 69). The MSMs provided approximately 10 cM genome-wide coverage on average.

Stage II. The NARAC I subjects were genotyped for 545,080 SNPs at the Feinstein Institute for Medical Research. The NARAC I RA case samples and 601 control samples were genotyped with the Infinium HumanHap550 v1.0 (Illumina), 411 controls on HumanHap550 v3.0 and 248 controls on Infinium HumanHap300 and HumanHap240S arrays (22). Genotypes were collected for samples across the three Illumina platforms and plate membership was assessed by the top 10 principal components (EIGENSTRAT); no systematic differences were observed (22). Subjects were excluded if more than 5% of genotypes were missing, had non-European ancestry, had evidence of relatedness, or had evidence of possible DNA contamination (22). Subjects were evaluated for Northern European ancestry by applying the software program STRUCTURE to an ancestry informative set of 704 SNPs (84, 85).

Stage III. The NARAC II subjects were genotyped using 373,400 SNPs on the Illumina HapMap370 BeadChip at the Feinstein Institute for Medical Research, as previously described (26). Subjects with more than 5% missing genotype data or showing evidence of non-European ancestry were excluded. In addition, samples showing evidence of relatedness with other samples in the study population, or possible DNA contamination were also excluded. Of the SNP variants identified in Stage II (see Statistical Analysis), a total of 219 SNP variants were imputed in the NARAC II data set using maximum likelihood imputation and applying the greedy algorithm as implemented in MACH v1.0.16 (86). Five Markov Chain iterations were set to obtain map estimates which were used as conditions for finding the most likely genotype. The NARAC I healthy controls were used as the reference population.

Statistical Analysis

Stage I. A graphical summary of the analytical approach for this study, including all steps, statistical methods and significance criteria is provided in Figure 1. The first stage of analysis utilized Random Forests, a supervised machine learning algorithm that grows recursively partitioned trees without pruning (87). Each tree is independently grown on a bootstrapped sample of observations, and at each node in the tree, the predictor that best discriminates the outcome is selected from a random subset of predictors. Classification accuracy of the forest is assessed for observations not included in the bootstrapped sample by comparing the predicted versus the actual outcome across all trees. Finally, the value of each predictor is randomly permuted across all trees, and a single VI score for each predictor is determined by the change in classification accuracy (misclassification). The VI score reflects the relationship between the increase in misclassification of the outcome when the predictor is permuted and classification accuracy is evaluated across all trees in the forests. This score can be ordered to suggest *important* predictors that most contribute to outcome classification without model specification, however the Type I error rate remains unknown.

Our Random Forests analysis explored the specific hypothesis that ASPs may share other genetic regions relevant to the biological mechanism(s) through which their shared *PTPN22* status confers risk for RA; therefore we used mIBD probabilities for 379 MSMs to predict *PTPN22* concordance in ASPs using Random Forests v5.1 (http://www.stat.berkeley.edu/~breiman/RandomForests/cc_software.htm). ASPs were categorized as positively or negatively concordant for *PTPN22* 1858T carrier status (both siblings having at least one variant (N=83 pairs), or neither sibling having a variant (N=209 pairs)). All MSMs were in Hardy-Weinberg equilibrium (HWE) in unrelated individuals and had

inheritance consistency within families (12), and mIBD probabilities for each MSM were generated using GENEHUNTER v2.1 (88). Based on the distribution of the VI scores (see Supplementary Figure 1), five *important* (top-ranking) predictors from the Random Forests analysis were chosen for further study (Table 2).

Stage II. All SNPs within 5Mb of an *important* locus were selected for investigation in the NARAC I case-control data set (N=11,207 SNPs). SNPs were excluded if they did not meet set criteria for HWE ($p>1x10^{-4}$), genotype call rates (>90% completeness), or minor allele frequency (MAF; >0.01). A final subset of SNPs used for this stage of analysis (N=10,589 SNPs) had an average call rate of 99.2% (Table 2).

Power to detect *gene* x *gene* interactions with *PTPN22* in Stage II was investigated, using dominant and recessive inheritance modes. We assumed a two-sided type 1 error of 5% ($\alpha=0.05$), and a frequency and effect estimate for *PTPN22* 1858T of 9.0% and OR=1.8 based on published estimates. Power to detect ROR_I ranging from 0.1 to 3.0 was examined. Our analyses revealed that there was sufficient power (~70-99%) to detect *gene* x *gene* interactions with $ROR_I \leq 0.5$ and ≥ 2.0 for almost all dominant models considered, and for recessive models where MAF>35% in Stage II (data not shown). Therefore, we restricted our investigation to dominant genetic models, and all analyses were performed and results interpreted in accordance with these established criteria.

Identified SNPs were evaluated for effect modification of *PTPN22* risk in NARAC I cases and controls using the Breslow-Day test for homogeneity implemented in PLINK v1.06 (89). Significant SNPs ($p\text{-value}_{BD}<0.05$) were selected as candidates for epistasis with *PTPN22*, and were formally tested for epistasis using logistic regression models in Stata v9.2 (StataCorp LP, College Station, TX). The test for epistasis was based on the coefficient of the interaction term (where *p*-value of the interaction term [ratio of odds ratio; ROR_I] reflects the difference in the likelihood between the full model and a reduced model containing only main effects); interactions demonstrating a significance level $p\text{-value}_I<0.05$ were considered significant.

Stage III. The SNPs that provided evidence for epistatic interaction with *PTPN22* in Stage II were further investigated. Power to detect *gene* x *gene* interactions between *PTPN22* and each final candidate SNP was investigated for the replication (NARAC II) data set using previously defined parameters. Our analyses revealed that power was sufficient (~70-99%) in Stage III to detect *gene* x *gene* interactions with $ROR_I \leq 0.6$ and ≥ 1.6 for almost all dominant models considered. Candidate SNPs that satisfied quality control criteria previously mentioned, were formally tested for epistasis using logistic regression models in Stata v9.2. Interactions demonstrating a significance level $p\text{-value}_I<0.05$ were considered significant.

Stage IV. A combined analysis including all 1,624 RA cases and 2,506 control subjects gave final estimates of effect size for replicating interactions. Allelic frequencies for all SNPs pursued in the fourth stage were first tested for heterogeneity (Stata v9.2) in each case and control group separately before combining. Only those demonstrating homogeneity across groups were included. SNP stratified analyses (minor allele carrier versus non-carrier subjects [dominant genetic model]) were used to further characterize the epistatic relationships for *PTPN22* using logistic regression (Stata v9.2).

REFERENCES

1. McCarthy MI, Abecasis GR, Cardon LR, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 2008;9:356-69.
2. Gregersen PK, Olsson L. Recent Advances in the Genetics of Autoimmune Disease. Annu Rev Immunol 2009;27:363-91.
3. Bell JT. A two-dimensional genome scan for rheumatoid arthritis susceptibility loci. BMC Proc 2007;1 Suppl 1:S63.
4. Mei L, Li X, Yang K, et al. Evaluating gene x gene and gene x smoking interaction in rheumatoid arthritis using candidate genes in GAW15. BMC Proc 2007;1 Suppl 1:S17.
5. Glaser B, Nikolov I, Chubb D, et al. Analyses of single marker and pairwise effects of candidate loci for rheumatoid arthritis using logistic regression and random forests. BMC Proc 2007;1 Suppl 1:S54.
6. Ma L, Dvorkin D, Garbe JR, Da Y. Genome-wide analysis of single-locus and epistasis single-nucleotide polymorphism effects on anti-cyclic citrullinated peptide as a measure of rheumatoid arthritis. BMC Proc 2007;1 Suppl 1:S127.
7. Ritchie MD, Bartlett J, Bush WS, Edwards TL, Motsinger AA, Torstenson ES. Exploring epistasis in candidate genes for rheumatoid arthritis. BMC Proc 2007;1 Suppl 1:S70.
8. Ding Y, Cong L, Ionita-Laza I, Lo SH, Zheng T. Constructing gene association networks for rheumatoid arthritis using the backward genotype-trait association (BGTA) algorithm. BMC Proc 2007;1 Suppl 1:S13.
9. Musani SK, Shriner D, Liu N, et al. Detection of gene x gene interactions in genome-wide association studies of human population data. Hum Hered 2007;63:67-84.
10. Cordell HJ. Genome-wide association studies: Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet 2009.
11. Gregersen PK, Behrens TW. Genetics of autoimmune diseases--disorders of immune homeostasis. Nat Rev Genet 2006;7:917-28.
12. Jawaheer D, Seldin MF, Amos CI, et al. Screening the genome for rheumatoid arthritis susceptibility genes: a replication study and combined analysis of 512 multicase families. Arthritis Rheum 2003;48:906-16.
13. Begovich AB, Carlton VE, Honigberg LA, et al. A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. Am J Hum Genet 2004;75:330-7.
14. Kallberg H, Padyukov L, Plenge RM, et al. Gene-gene and gene-environment interactions involving HLA-DRB1, PTPN22, and smoking in two subsets of rheumatoid arthritis. Am J Hum Genet 2007;80:867-75.
15. Gregersen PK, Lee HS, Batliwalla F, Begovich AB. PTPN22: setting thresholds for autoimmunity. Semin Immunol 2006;18:214-23.
16. Vang T, Miletic AV, Arimura Y, Tautz L, Rickert RC, Mustelin T. Protein tyrosine phosphatases in autoimmunity. Annu Rev Immunol 2008;26:29-55.
17. Jawaheer D, Seldin MF, Amos CI, et al. A genomewide screen in multiplex rheumatoid arthritis families suggests genetic overlap with other autoimmune diseases. Am J Hum Genet 2001;68:927-36.
18. Amos CI, Chen WV, Lee A, et al. High-density SNP analysis of 642 Caucasian families with rheumatoid arthritis identifies two new linkage regions on 11p12 and 2q33. Genes Immun 2006;7:277-86.

19. Rodriguez MR, Nunez-Roldan A, Aguilar F, Valenzuela A, Garcia A, Gonzalez-Escribano MF. Association of the CTLA4 3' untranslated region polymorphism with the susceptibility to rheumatoid arthritis. Hum Immunol 2002;63:76-81.
20. Suzuki A, Yamada R, Chang X, et al. Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. Nat Genet 2003;34:395-402.
21. Remmers EF, Plenge RM, Lee AT, et al. STAT4 and the risk of rheumatoid arthritis and systemic lupus erythematosus. N Engl J Med 2007;357:977-86.
22. Plenge RM, Seielstad M, Padyukov L, et al. TRAF1-C5 as a risk locus for rheumatoid arthritis--a genomewide study. N Engl J Med 2007;357:1199-209.
23. Thomson W, Barton A, Ke X, et al. Rheumatoid arthritis association at 6q23. Nat Genet 2007;39:1431-3.
24. Barton A, Thomson W, Ke X, et al. Rheumatoid arthritis susceptibility loci at chromosomes 10p15, 12q13 and 22q13. Nat Genet 2008;40:1156-9.
25. Barton A, Thomson W, Ke X, et al. Re-evaluation of putative rheumatoid arthritis susceptibility genes in the post-genome wide association study era and hypothesis of a key pathway underlying susceptibility. Hum Mol Genet 2008;17:2274-9.
26. Gregersen PK, Amos CI, Lee AT, et al. REL, encoding a member of the NF-kappaB family of transcription factors, is a newly defined risk locus for rheumatoid arthritis. Nat Genet 2009.
27. John S, Amos C, Shephard N, et al. Linkage analysis of rheumatoid arthritis in US and UK families reveals interactions between HLA-DRB1 and loci on chromosomes 6q and 16p. Arthritis Rheum 2006;54:1482-90.
28. Newman WG, Zhang Q, Liu X, et al. Rheumatoid arthritis association with the FCRL3 -169C polymorphism is restricted to PTPN22 1858T-homozygous individuals in a Canadian population. Arthritis Rheum 2006;54:3820-7.
29. Julia A, Moore J, Miquel L, et al. Identification of a two-loci epistatic interaction associated with susceptibility to rheumatoid arthritis through reverse engineering and multifactor dimensionality reduction. Genomics 2007;90:6-13.
30. Klareskog L, Stolt P, Lundberg K, et al. A new model for an etiology of rheumatoid arthritis: smoking may trigger HLA-DR (shared epitope)-restricted immune reactions to autoantigens modified by citrullination. Arthritis Rheum 2006;54:38-46.
31. Criswell LA, Pfeiffer KA, Lum RF, et al. Analysis of families in the multiple autoimmune disease genetics consortium (MADGC) collection: the PTPN22 620W allele associated with multiple autoimmune phenotypes. Am J Hum Genet 2005;76:561-71.
32. Hill RJ, Zozulya S, Lu YL, Ward K, Gishizky M, Jallal B. The lymphoid protein tyrosine phosphatase Lyp interacts with the adaptor molecule Grb2 and functions as a negative regulator of T-cell activation. Exp Hematol 2002;30:237-44.
33. Wu J, Katrekar A, Honigberg LA, et al. Identification of substrates of human protein-tyrosine phosphatase PTPN22. J Biol Chem 2006;281:11002-10.
34. Vang T, Congia M, Macis MD, et al. Autoimmune-associated lymphoid tyrosine phosphatase is a gain-of-function variant. Nat Genet 2005;37:1317-9.
35. Rieck M, Arechiga A, Onengut-Gumuscu S, Greenbaum C, Concannon P, Buckner JH. Genetic variation in PTPN22 corresponds to altered function of T and B lymphocytes. J Immunol 2007;179:4704-10.

36. Arechiga AF, Habib T, He Y, et al. Cutting edge: the PTPN22 allelic variant associated with autoimmunity impairs B cell signaling. J Immunol 2009;182:3343-7.
37. Zikherman J, Hermiston M, Steiner D, Hasegawa K, Chan A, Weiss A. PTPN22 deficiency cooperates with the CD45 E613R allele to break tolerance on a non-autoimmune background. J Immunol 2009;182:4093-106.
38. Dames SA, Bang E, Haussinger D, Ahrens T, Engel J, Grzesiek S. Insights into the low adhesive capacity of human T-cadherin from the NMR structure of Its N-terminal extracellular domain. J Biol Chem 2008;283:23485-95.
39. Philippova MP, Bochkov VN, Stambolsky DV, Tkachuk VA, Resink TJ. T-cadherin and signal-transducing molecules co-localize in caveolin-rich membrane domains of vascular smooth muscle cells. FEBS Lett 1998;429:207-10.
40. Philippova M, Ivanov D, Allenspach R, Takuwa Y, Erne P, Resink T. RhoA and Rac mediate endothelial cell polarization and detachment induced by T-cadherin. Faseb J 2005;19:588-90.
41. Joshi MB, Philippova M, Ivanov D, Allenspach R, Erne P, Resink TJ. T-cadherin protects endothelial cells from oxidative stress-induced apoptosis. Faseb J 2005;19:1737-9.
42. Philippova M, Banfi A, Ivanov D, et al. Atypical GPI-anchored T-cadherin stimulates angiogenesis in vitro and in vivo. Arterioscler Thromb Vasc Biol 2006;26:2222-30.
43. Kuphal S, Martyn AC, Pedley J, et al. H-cadherin expression reduces invasion of malignant melanoma. Pigment Cell Melanoma Res 2009;22:296-306.
44. Adachi Y, Takeuchi T, Nagayama T, Ohtsuki Y, Furihata M. Zeb1-mediated T-cadherin repression increases the invasive potential of gallbladder cancer. FEBS Lett 2009;583:430-6.
45. Franke B, Neale BM, Faraone SV. Genome-wide association studies in ADHD. Hum Genet 2009.
46. Org E, Eyheramendy S, Juhanson P, et al. Genome-wide scan identifies CDH13 as a novel susceptibility locus contributing to blood pressure determination in two European populations. Hum Mol Genet 2009;18:2288-96.
47. Axenovich TI, Zorkoltseva IV, Belonogova NM, et al. Linkage analysis of adult height in a large pedigree from a Dutch genetically isolated population. Hum Genet 2009.
48. Dose AC, Hillman DW, Wong C, Sohlberg L, Lin-Jones J, Burnside B. Myo3A, one of two class III myosin genes expressed in vertebrate retina, is localized to the calycal processes of rod and cone photoreceptors and is expressed in the sacculus. Mol Biol Cell 2003;14:1058-73.
49. Walsh T, Walsh V, Vreugde S, et al. From flies' eyes to our ears: mutations in a human class III myosin cause progressive nonsyndromic hearing loss DFNB30. Proc Natl Acad Sci U S A 2002;99:7518-23.
50. Elwany S, el Garf A, Kamel T. Hearing and middle ear function in rheumatoid arthritis. J Rheumatol 1986;13:878-81.
51. Ozcan M, Karakus MF, Gunduz OH, Tuncel U, Sahin H. Hearing loss and middle ear involvement in rheumatoid arthritis. Rheumatol Int 2002;22:16-9.
52. Murdin L, Patel S, Walmsley J, Yeoh LH. Hearing difficulties are common in patients with rheumatoid arthritis. Clin Rheumatol 2008;27:637-40.
53. Schneider ME, Dose AC, Salles FT, et al. A new compartment at stereocilia tips defined by spatial and temporal patterns of myosin IIIa expression. J Neurosci 2006;26:10243-52.

54. Jacobelli J, Bennett FC, Pandurangi P, Tooley AJ, Krummel MF. Myosin-IIA and ICAM-1 regulate the interchange between two distinct modes of T cell migration. J Immunol 2009;182:2041-50.
55. Oshimori N, Li X, Ohsugi M, Yamamoto T. Cep72 regulates the localization of key centrosomal proteins and proper bipolar spindle formation. Embo J 2009.
56. Miyawaki S, Asanuma H, Nishiyama S, Yoshinaga Y. Clinical and serological heterogeneity in patients with anticentromere antibodies. J Rheumatol 2005;32:1488-94.
57. Walker JG, Fritzler MJ. Update on autoantibodies in systemic sclerosis. Curr Opin Rheumatol 2007;19:580-91.
58. Wu R, Shovman O, Zhang Y, Gilburd B, Zandman-Goddard G, Shoenfeld Y. Increased prevalence of anti-third generation cyclic citrullinated peptide antibodies in patients with rheumatoid arthritis and CREST syndrome. Clin Rev Allergy Immunol 2007;32:47-56.
59. Salliot C, Gottenberg JE, Bengoufa D, Desmoulins F, Miceli-Richard C, Mariette X. Anticentromere antibodies identify patients with Sjogren's syndrome and autoimmune overlap syndrome. J Rheumatol 2007;34:2253-8.
60. Larsen M, Ressler SJ, Gerdes MJ, et al. The WFDC1 gene encoding ps20 localizes to 16q24, a region of LOH in multiple cancers. Mamm Genome 2000;11:767-73.
61. McAlhany SJ, Ressler SJ, Larsen M, et al. Promotion of angiogenesis by ps20 in the differential reactive stroma prostate cancer xenograft model. Cancer Res 2003;63:5859-65.
62. Madar S, Brosh R, Buganim Y, et al. Modulated expression of WFDC1 during carcinogenesis and cellular senescence. Carcinogenesis 2009;30:20-7.
63. Alvarez R, Reading J, King DF, et al. WFDC1/ps20 is a novel innate immunomodulatory signature protein of human immunodeficiency virus (HIV)-permissive CD4+ CD45RO+ memory T cells that promotes infection by upregulating CD54 integrin expression and is elevated in HIV type 1 infection. J Virol 2008;82:471-86.
64. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. Science 2008;322:881-8.
65. MacGregor AJ, Snieder H, Rigby AS, et al. Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. Arthritis Rheum 2000;43:30-7.
66. Jawaheer D, Li W, Graham RR, et al. Dissecting the genetic complexity of the association between human leukocyte antigens and rheumatoid arthritis. Am J Hum Genet 2002;71:585-94.
67. Ling S, Li Z, Borschukova O, Xiao L, Pumpens P, Holoshitz J. The rheumatoid arthritis shared epitope increases cellular susceptibility to oxidative stress by antagonizing an adenosine-mediated anti-oxidative pathway. Arthritis Res Ther 2007;9:R5.
68. Ling S, Pi X, Holoshitz J. The rheumatoid arthritis shared epitope triggers innate immune signaling via cell surface calreticulin. J Immunol 2007;179:6359-67.
69. Lee AT, Li W, Liew A, et al. The PTPN22 R620W polymorphism associates with RF positive rheumatoid arthritis in a dose-dependent manner but not with HLA-SE status. Genes Immun 2005;6:129-33.
70. Harrison P, Pointon JJ, Farrar C, Brown MA, Wordsworth BP. Effects of PTPN22 C1858T polymorphism on susceptibility and clinical characteristics of British Caucasian rheumatoid arthritis patients. Rheumatology (Oxford) 2006;45:1009-11.
71. Costenbader KH, Chang SC, De Vivo I, Plenge R, Karlson EW. Genetic polymorphisms in PTPN22, PADI-4, and CTLA-4 and risk for rheumatoid arthritis in two longitudinal

cohort studies: evidence of gene-environment interactions with heavy cigarette smoking. Arthritis Res Ther 2008;10:R52.
72. Kokkonen H, Johansson M, Innala L, Jidell E, Rantapaa-Dahlqvist S. The PTPN22 1858C/T polymorphism is associated with anti-cyclic citrullinated peptide antibody-positive early rheumatoid arthritis in northern Sweden. Arthritis Res Ther 2007;9:R56.
73. Ding B, Padyukov L, Lundstrom E, et al. Different patterns of associations with anti-citrullinated protein antibody-positive and anti-citrullinated protein antibody-negative rheumatoid arthritis in the extended major histocompatibility complex region. Arthritis Rheum 2009;60:30-8.
74. Arnett FC, Edworthy SM, Bloch DA, et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 1988;31:315-24.
75. Schellekens GA, Visser H, de Jong BA, et al. The diagnostic properties of rheumatoid arthritis antibodies recognizing a cyclic citrullinated peptide. Arthritis Rheum 2000;43:155-63.
76. Wolfe F, Michaud K, Gefeller O, Choi HK. Predicting mortality in patients with rheumatoid arthritis. Arthritis Rheum 2003;48:1530-42.
77. Fries JF, Wolfe F, Apple R, et al. HLA-DRB1 genotype associations in 793 white patients from a rheumatoid arthritis inception cohort: frequency, severity, and treatment bias. Arthritis Rheum 2002;46:2320-9.
78. Weisman B, Bombardier C, Massarotti E. Analysis at one year of an inception cohort of early rheumatoid arthritis: the SONORA study. Arthritis Rheum 2003;48.
79. Mitchell MK, Gregersen PK, Johnson S, Parsons R, Vlahov D. The New York Cancer Project: rationale, organization, design, and baseline characteristics. J Urban Health 2004;81:301-10.
80. Sokka T, Pincus T. An Early Rheumatoid Arthritis Treatment Evaluation Registry (ERATER) in the United States. Clin Exp Rheumatol 2005;23:S178-81.
81. Mikuls TR, Kazi S, Cipher D, et al. The association of race and ethnicity with disease expression in male US veterans with rheumatoid arthritis. J Rheumatol 2007;34:1480-4.
82. Kolfenbach JR, Deane KD, Derber LA, et al. A prospective approach to investigating the natural history of pre-clinical rheumatoid arthritis (RA) using first-degree relatives of probands with RA. Arthritis Care Res 2009;(in press).
83. Kosoy R, Nassir R, Tian C, et al. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. Hum Mutat 2008;30:69-78.
84. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. Am J Hum Genet 2000;67:170-81.
85. Seldin MF, Shigeta R, Villoslada P, et al. European population substructure: clustering of northern and southern populations. PLoS Genet 2006;2:e143.
86. Li Y, Abecasis G. Mach 1.0: Rapid Haplotype Reconstruction and Missing Genotype Inference. Am J Hum Genet 2006;S79.
87. Breiman L. Random Forests. Machine Learning 2001;45:5-32.
88. Markianos K, Daly MJ, Kruglyak L. Efficient multipoint linkage analysis through reduction of inheritance space. Am J Hum Genet 2001;68:963-77.
89. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559-75.

Table 1: Clinical and demographic features of the RA family and case-control datasets

Study Population		N	Female: Male Ratio	Age of onset (years)	*PTPN22* 1858T carrier (%) [1]	Shared Epitope carrier (%) [2]	Anti-CCP status (%)
Stage I	RA cases	530	3.3:1	39.2	29.1% (n=154)	83.8% (n=444)	90.8% (n=481)
	ASPs	292	--	--	83	227	259
Stage II (NARAC I)	All RA cases	908	2.8:1	45.5 (n=884)	27.8% (n=898)	97.9% (n=865)	100% (n=907)
	All controls	1,260	2.5:1	NA	15.6% (n=1,253)	44.6% (n=1,159)	NA
	NE [3] RA cases	682	2.8:1	45.7 (n=670)	27.6% (n=677)	97.7% (n=641)	100% (n=681)
	NE [3] controls	752	2.3:1	NA	16.8% (n=750)	46.7% (n=752)	NA
Stage III (NARAC II)	All RA cases	952	1.2:1 [4]	47.2 (n=917)	27.4% (n=947)	NA	100% (n=921)
	All controls	1,760	NA	NA	17.2% (n=1,756)	NA	NA

[1] Individuals were *PTPN22* 1858T carriers if they had one or more 1858T alleles. In Stage II, the *PTPN22* 1858T allele was associated with RA using all individuals (OR=2.09, 95% CI: 1.69-2.60) and when restricted by Northern European ancestry (OR=1.89, 95% CI: 1.45-2.46). In Stage III, the *PTPN22* 1858T allele was associated with RA (OR=1.81, 95% CI: 1.49-2.20).
[2] Individuals were Shared Epitope carriers if they had one or more of the following *HLA-DRB1* alleles: 0101, 0102, 0401, 0404, 0405, 0408, 0413, 1001, and 1402. However, four-digit typing was not available for all subjects, thus cases with *HLA-DRB1**04/01/10 were assumed to be carriers for Shared Epitope.
[3] Subjects with at least 90% Northern European ancestry.
[4] Lower Female : Male ratio reflects overrepresentation of male RA cases from the Veterans Affairs Rheumatoid Arthritis Registry.

Table 2: Top genomic regions identified by the Random Forests analysis in Stage I (292 ASPs) and number of SNPs to be investigated in Stage II (NARAC I: 677 cases, 750 controls)

Chr	MSM	Band	No. SNPs [1]
4	*D4S2431*	4q34.1	1,580
5	*D5S2488*	5p15.33	1,167
10	*D10S1426*	10p11.23	2,016
14	*D14S592*	14q23.1	2,006
16	*D16S402*	16q23.3	3,820

[1] SNP variants were selected if present in the NARAC I GWA data set, satisfied quality control criteria, and were within 5 Mb of an *important* MSM identified by Random Forests (see Supplementary Figure 1).

Table 3: Replicating results for tests of interaction between *PTPN22* and candidate SNPs in two populations of European origins (Stage II and III), and the effect of *PTPN22* when stratified by SNP of interest.

Chr	SNP	p_{BD} [1]	Stage II [2]				Stage III [3]			
			p_I	ROR_I [4] (95% CI)	SNP carriers [5]	SNP non-carriers [6]	p_I	ROR_I [4] (95% CI)	SNP carriers [5]	SNP non-carriers [6]
					OR_{PTPN22} (95% CI)	OR_{PTPN22} (95% CI)			OR_{PTPN22} (95% CI)	OR_{PTPN22} (95% CI)
4	rs4695888T	0.041	0.016	2.02 (1.14-3.60)	2.26 (1.66-3.06)	1.11 (0.68-1.81)	0.038	1.59 (1.03-2.45)	2.04 (1.64-2.54)	1.29 (0.88-1.87)
5	rs7726839G	0.025	0.021	0.55 (0.33-0.91)	1.39 (0.97-2.00)	2.55 (1.77-3.66)	0.046	0.68 (0.46-0.99)	1.45 (1.09-1.94)	2.15 (1.67-2.76)
10	rs12573019A	0.029	0.022	1.99 (1.11-3.57)	3.13 (1.89-5.19)	1.58 (1.17-2.12)	0.048	1.56 (1.00-2.42)	2.53 (1.73-3.71)	1.63 (1.31-2.03)
10	rs1168587T	0.019	0.018	1.96 (1.12-3.44)	2.30 (1.69-3.13)	1.17 (0.73-1.87)	0.037	1.55 (1.03-2.35)	2.07 (1.65-2.60)	1.33 (0.94-1.89)
16	rs1895535A	0.018	0.038	2.94 (1.06-8.12)	5.03 (1.89-13.43)	1.71 (1.31-2.24)	0.0076	2.42 (1.26-4.62)	4.02 (2.17-7.45)	1.67 (1.36-2.03)
16	rs7200573A	0.026	0.021	0.54 (0.32-0.91)	1.42 (1.00-2.02)	2.61 (1.79-3.81)	0.0015	0.54 (0.37-0.79)	1.33 (1.01-1.74)	2.46 (1.89-3.21)
16	rs11865624C	0.028	0.050	0.48 (0.23-1.00)	1.01 (0.51-2.01)	2.12 (1.60-2.79)	0.020	0.51 (0.28-0.90)	1.00 (0.59-1.71)	1.98 (1.62-2.42)

[1] Breslow-Day Test was used as a test of effect modification, evaluating the homogeneity of association (OR) between a SNP variant and the risk for RA, across each strata of *PTPN22* 1858T carrier status in NARAC I data (677 RA cases and 750 controls). Only those SNPs with $p\text{-value}_{BD}<0.05$ were tested for epistasis.
[2] Using NARAC I data: 677 RA cases and 750 controls.
[3] Using NARAC II data: 947 RA cases and 1,756 controls.
[4] The test of epistasis is based on the interaction term (reported ratio of odds ratios [ROR_I] and 95% CI) between the dominant genetic model of *PTPN22* 1858T (CT/TT vs. CC) and the dominant genetic model for the SNP variant.
[5] Using RA cases and controls who carried the minor allele of the SNP variant of interest.
[6] Using RA cases and controls who did not carry the minor allele of the SNP variant of interest.

Table 4: Results for tests for interaction between *PTPN22* and replicating SNPs in the combined data set, and results for the effect of *PTPN22* 1858T when stratified by SNP variants of interest under a dominant model (Stage IV) [1].

Chr	SNP	MAF [2]	Stage IV		SNP carrier [4]		SNP non-carrier [5]		Gene
			ROR_I [3] (95% CI)	p_I	OR (95% CI)	p	OR (95% CI)	p	
4	rs4695888T	0.50	1.71 (1.22-2.43)	**0.0021**	2.10 (1.76-2.51)	**<1 x 10^{-8}**	1.22 (0.91-1.65)	0.184	
5	rs7726839G	0.26	0.65 (0.48-0.89)	**0.0061**	1.46 (1.16-1.82)	**0.0011**	2.22 (1.81-2.73)	**<1 x 10^{-8}**	*CEP72*
10	rs12573019A	0.13	1.74 (1.23-2.47)	**0.0019**	2.78 (2.05-3.76)	**<1 x 10^{-8}**	1.60 (1.34-1.90)	**1.7 x 10^{-7}**	*MYO3A*
10	rs1168587T	0.43	1.66 (1.19-2.31)	**0.0028**	2.13 (1.78-2.56)	**<1 x 10^{-8}**	1.29 (0.98-1.70)	0.073	
16	rs1895535A	0.06	2.34 (1.38-3.97)	**0.0016**	3.95 (2.39-6.55)	**1.0 x 10^{-7}**	1.69 (1.44-1.98)	**<1 x 10^{-8}**	*CDH13*
16	rs7200573A	0.29	0.56 (0.41-0.75)	**0.00015**	1.37 (1.11-1.70)	**0.0034**	2.47 (1.99-3.06)	**<1 x 10^{-8}**	*CDH13*
16	rs11865624C	0.07	0.50 (0.32-0.79)	**0.0031**	1.01 (0.66-1.55)	0.946	2.01 (1.71-2.37)	**<1 x 10^{-8}**	5' *WFDC1*

[1] Using the combined data from NARAC I and NARAC II: 1,647 RA cases and 2,506 controls.
[2] Minor allele frequencies (MAF) were determined using data from controls.
[3] The test of epistasis is based on the interaction term (reported ratio of odds ratios [ROR_I] and 95% CI) between the dominant genetic model of *PTPN22* 1858T (CT/TT vs. CC) and the dominant genetic model for the SNP variant.
[4] Using RA cases and controls who carried the minor allele of the SNP variant of interest.
[5] Using RA cases and controls who did not carry the minor allele of the SNP variant of interest.

Figure 1. Summary of Analytical Approach to Identify Evidence for Epistatic Relationships with *PTPN22* in RA.

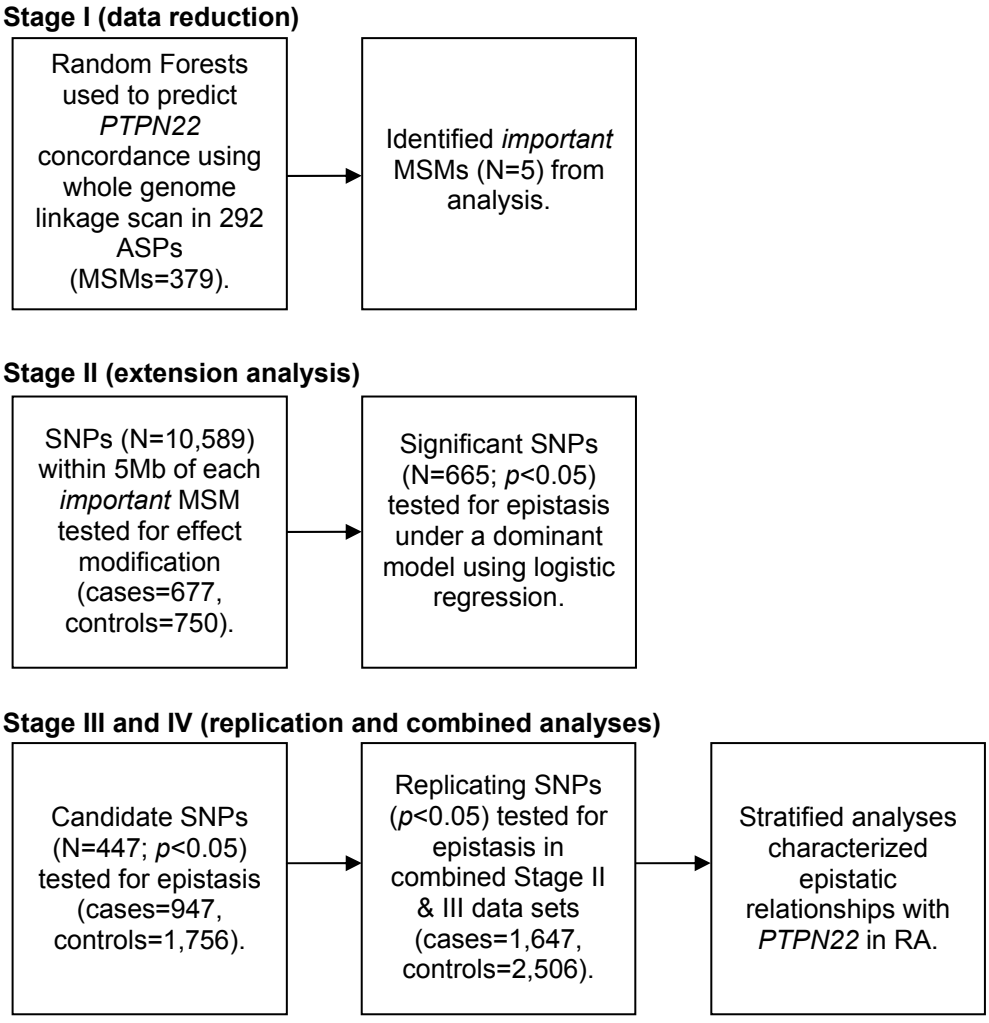
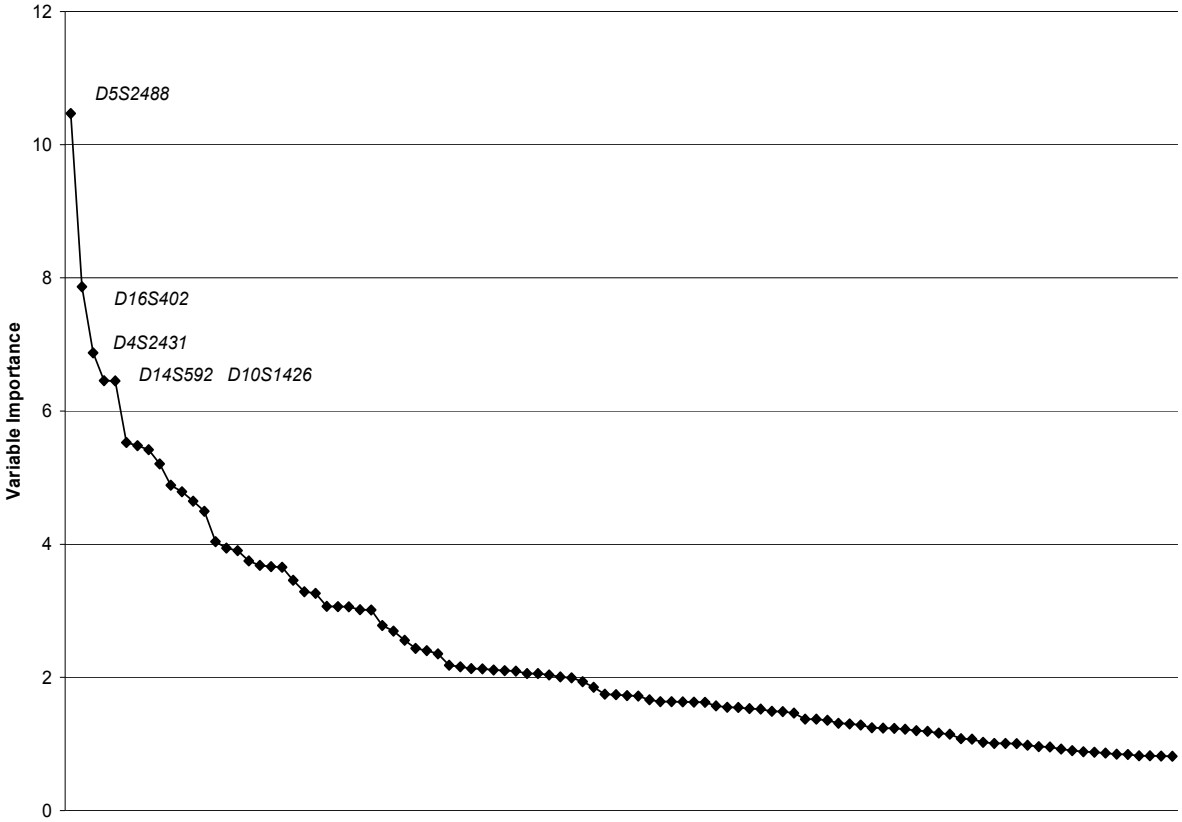


Figure 2: Variable Importance (VI) Scores for the Top 100 Predictors for the Random Forests Analysis. Gender, *HLA-SE* status, and Multipoint Identity-By-Descent Probabilities for 379 Microsatellite Markers, were used to Predict *PTPN22* 1858T Concordance in 292 Affected Sibling Pairs. There is a Clear Distinction Between the VI Scores for the Five Top-Ranking Predictors Compared to the Remaining Predictors.



CONCLUSIONS

Autoimmune diseases (ADs) are a major global public health concern. Despite, intensive epidemiological research over the last two decades, our understanding of the etiological mechanisms contributing to ADs is limited. However, we do know there is a prominent and complex genetic component. Genome-wide association (GWA) studies have laid the foundation for unraveling the complex genetic component of ADs and other diseases; however it is clear that other analytical frameworks are necessary to further refine the genetic etiology of complex diseases. This includes the identification of additional susceptibility loci that are common variants, as well as other genetic and epigenetic variants. The work described in this dissertation demonstrates the utility of hypothesis-driven research and the application of robust and comprehensive analytical strategies in characterizing the complex genetic component in two ADs: multiple sclerosis (MS) and rheumatoid arthritis (RA).

Chapter 1 comprehensively investigated common variation within DNA repair pathway genes and susceptibility of MS. Two analytical approaches were performed in parallel. Significant evidence for association between a SNP in *GTF2H4*, a nucleotide excision repair (NER) gene, and risk for MS was observed. Significant results for other genetic variants based on marginal, epistatic, and multigenic tests of association were not observed after stringent correction for multiple testing. Random Forests and CART analyses largely overlapped with uncorrected marginal tests of association; however additional NER genes were identified as candidate genes for future research. Further investigation of DNA repair pathways in the context of environmental exposures is necessary, as is the investigation of rare variants, perhaps in particular subsets of MS cases. Further research is necessary to definitively investigate the biological relationship with MS pathogenesis and other DNA repair pathways, including direct reversal, mismatch repair, and translesion synthesis, with particular emphasis on research efforts that can incorporate both genetic and environmental risk factors under a unified framework. In summary, this chapter provides limited evidence for the involvement of common genetic variation in biologically-relevant DNA repair pathways in the multifactorial pathogenesis of MS.

Chapter 2 successfully identified a novel MS susceptibility locus, *CRHR1*, using a well-powered hypothesis-driven study design including both exploratory and replication stages. *CRHR1* is located within 17q21 chromosomal region, and has been a key locus of interest in MS, however this is the first susceptibility locus identified in this region. This finding underscores the importance of considering biological function in genetic studies of MS, as well as other ADs, as this replicated association would not have been identified by genome-wide association (GWA) studies ($p<5 \times 10^{-8}$). *CRHR1* is a critical component of the hypothalamus-pituitary-adrenal (HPA) axis, and is present on various immune cells. An impaired HPA axis has been suspected to contribute to autoimmunity, including MS and has been associated with the development of affective disorders and other stress-related clinical conditions, both marginally and through *gene* x *environment* interactions. Thus, further investigations of these conditions, experiences of stressful life events, and genetic variation within HPA axis genes in MS is necessary. This investigation provides evidence for a non-major histocompatibility complex (MHC) mediated mechanism in the pathogenesis of MS.

Chapter 3 successfully investigated epistasis (*gene* x *gene* interactions) with the prominent non-MHC susceptibility locus for RA, *PTPN22*, across the genome. Strong evidence for novel

epistatic interactions with *PTPN22* and variants within *CDH13*, *MYO3A*, *CEP72*, and near *WFDC1*, was revealed. A clear research hypothesis was pursued using large, well-characterized study populations with detailed clinical and genetic information in a multi-stage analysis that included extension and replication stages, and combined robust non-parametric and parametric methods. These results underscore the importance of exploring analytical strategies capable of detecting complex genetic risks in the absence of marginal effects, as none of the final SNP variants demonstrated significant associations with RA ($p<0.05$) when cases were compared to controls. However, this investigation only explored two-way interactions and not higher-order relationships which are likely to contribute to RA and other ADs. Furthermore, a dominant genetic model assumption used in this study for epistatic candidates may not have been appropriate. Future studies should explore additional genetic models. The replicated epistatic interactions observed in this study were of modest significance ($p<1 \times 10^{-5}$). An exploration of higher-order interactions will require much larger sample sizes and clearly defined phenotypes.

Recently, genetic epidemiologists have acknowledged that GWA studies and the common-variant-common-disease hypothesis are seemingly limited in characterizing the complete genetic component of complex ADs, including MS and RA. However, work described in this dissertation demonstrates the use of strong epidemiologic study design, clearly defined hypotheses, and robust analytical strategies, may further our understanding of the contribution of common variation to the genetic architecture of complex diseases. There are several explicit research initiatives that are needed to further understand the statistical associations reported in this dissertation. These include, but are not restricted to, the exploration of: (1) rare variants, (2) other biological-related genes/pathways, (3) new methods for investigating biological pathways, (4) approaches to detect associations, both marginal and epistatic, that do not meet criteria for genome-wide significance, and (5) considerations of context, or 'the environment'.

Rare Variants

The marginal and epistatic variants identified in this dissertation have largely unknown function and, therefore, focused DNA sequencing will be necessary to elucidate and describe true causal variants, their impact on protein function, and subsequent disease pathogenesis. Given the limited success of the common variant hypothesis, many new investigations are attempting to pursue large-scale sequencing studies to characterize the disease contribution of rare variants, which occur in <1% of the population. Most promising for improving our understanding of human genetic variation is the 1,000 Genomes Project, which is cataloguing both allelic and structural variation in at least 1,000 human genomes from ethnically diverse populations. However, building on the rare variant hypothesis, is the notion of genetic heterogeneity; where multiple rare variants contribute to disease, both marginally and through genetic interactions. Similar to common variants in GWA studies, rare variant analyses will present a large number of statistical challenges that are likely to lead to the development of new and powerful methods. One possible method that can aid a DNA sequencing investigation of candidate genes is to first prioritize SNPs based on *in silico* analyses. These analyses can help predict the impact of candidate SNPs on the final protein function, and can help distinguish conservative amino-acid changes from deleterious changes, and so forth, minimizing the bias introduced by multiple testing. Tools currently used for investigating GWA studies will be useful, but additional approaches to explore genetic heterogeneity will be necessary, such as tests capable of simultaneously considering all variation within a gene.

Pathway analysis

One of the greatest limitations of epidemiologic studies is the assumption that biological significance can be captured through statistical means. As discussed previously, GWA studies are hypothesis-free, and do not make assumptions regarding underlying disease etiology. This approach has successfully identified dozens of MS and RA susceptibility loci, to date. Many of these loci have immune-related function; however, some findings are disparate and further investigation of etiological mechanisms affected by the presence of these variants is necessary. Gene products may have pleiotropic effects; nonetheless, we know that these functional roles occur within the constructs of biological pathways. It is naïve to assume that a biological system cannot compensate for functional inadequacies in one component protein or does not have built in redundancies to ensure mechanistic success. By restricting genetic investigation to the identification of a single variant within a single gene, we lose the inferential advantage of having experimental biological knowledge. Pathway based analyses have not been fully developed or explored, but there are a few approaches currently available that may help identify additional susceptibility loci from association studies and creatively model biological pathways. The first approach is utilizing Gene Ontology (GO) categories, which consists of three vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. By performing hypergeometric tests comparing a gene set of liberally significant ($p<0.05$) GWA results compared to a reference gene set of all GWA genes tested, we may be able to identify GO categories that are significantly enriched. This procedure can inform the functional profile of associated genes within GO hierarchies and improve interpretation of GWA findings. As a result, a more focused approach based on biological mechanisms can guide future investigations.

Another pathway-driven approach that I hope to explore is the application of machine-learning algorithms to investigate genetic variation within biological pathways and susceptibility for MS. Data-adaptive techniques can simultaneously consider all variants, thus this approach provides information not available from marginal association tests. I hope to explore various approaches, including conventional methodologies, to develop an operational framework for testing biological pathways. One of the key algorithms, in addition to Random Forests, that will be explored is the Deletion/Substitution/Addition (D/S/A) algorithm. It is a data-adaptive technique that tests a series of models and allows the search to move among statistical models that are not nested. D/S/A can detect complex interactions and correlation patterns in addition to main effects, for example, D/S/A may determine that the best model predicting the outcome includes main effects, interactions (e.g. WY), polynomials (e.g. W^2Y^3), or any combination thereof (e.g. $Z + W^2Y$); therefore it is a seemingly appropriate statistical approach for uncovering complex biological relationships.

Context

I have only begun to discuss the importance of exploring context, in the form of biological pathways, but there are many additional layers that must be considered. As human beings, the biology of our lives is nested: genetic variation is nested within genes and chromosomes, which is within nuclei of cells, within tissues of organs, within biological pathways/mechanisms/systems, within a physical body, within a physical environment, and finally within a society. Variation at any of these levels impacts our physiology, thus by considering only the most specific level we lose sight of the larger picture. To date, genetic

studies have not routinely or comprehensively incorporated environmental data, largely in part to the complexities and costs of collecting and investigating such data in the context of genetic predisposition. A large impediment is appropriate conceptualization and detection of *gene* x *environment* interaction in the presence and/or absence of main effects. Several candidate *gene* x *environment* studies have been conducted, but have largely been underpowered to detect associations. Nonetheless many research initiatives are moving forward and exploring uncharted territory as the role for environment in MS and RA pathology is strongly implicated by modest disease concordance in monozygotic (MZ) twins.

Context matters. This is further evident in epigenetic investigations of MZ twins, which have demonstrated divergent DNA methylation patterns over time, resulting in differences in the potential expressivity of identical genes. How does this affect the penetrance of a genetic variant? Is this a surrogate for environmental exposure(s)? And which exposures? Variation in diet, smoking, exercise, experiences of discrimination, and other stressful life events can alter the epigenome. Knowing that both environmental and social exposures can biologically impact gene expressivity, is it sufficient to only explore genetic variation? The answer is no. Understanding genetic variation that contributes to disease is critical, and I hope to continue characterizing these relationships for the length of my academic career. However, my future research will focus on questions related to health outcomes that consider both the role of genetics and the greater and very complex context in which we live.