Title
The Genetics of Kidney Transplantation Outcomes and Autoimmune Disease

Author
Musone, Stacy Lynn

Publication Date
2010

Peer reviewed|Thesis/dissertation

The Genetics of Kidney Transplantation Outcomes and Autoimmune Disease

by

Stacy Lynn Musone

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Pharmaceutical Sciences and Pharmacogenomics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

ACKNOWLEDGEMENTS

The work presented in this dissertation was a collaborative effort contributed to by many individuals whose efforts often go unrecognized. There have been many times when I did not feel worthy enough to continue in this field, and it was the encouragement received from the people in this acknowledgement that drove me to persevere. I don't want to forget those who influenced me prior to coming to UCSF. Alex Parker, Alessandro Foti and Alyce Chen took a chance on me when I was an undergraduate and did not know a thing about human genetics. They taught me the basics of laboratory technique and set me up for success. Joel Hirschhorn taught me about high throughput genetics and complex traits and his excitement about the field was infectious.

I would like to thank both Christian Vaisse and Bruce Conklin for allowing me to rotate with them during my first year of graduate school. It was while in Christian's lab that I learned about cell culture experiments for the first time and I thank Christian and his lab for their patience with me. Bruce encouraged me to think big and creatively and I learned a lot about stem cell research and GPCR's while in his lab.

Every person, past and present, in the Kwok laboratory has helped me to get through graduate school. I would like to thank each of them – you know who you are. A few people deserve to be singled out, especially Ludmila Pawlikowska, Stephanie Hesselson, Wilson Liao, Jennifer Pons and Ting-Fung Chan for offering so much guidance to me on both a professional and personal level. Bani Tamraz and Ernest Lam, my co-graduate students, should be thanked for lending a listening ear when I needed to vent. Brad Dispensa generously, and often at inconvenient times, offered the best technical support ever dispensed and I would not have been able to analyze my data without him. Excellent technicians make our lab what it is: in particular, Justin Chen, Connie Ha, Matthew Akana, Theodora Wingert, Eunice Wan, Catherine Chu and Annie Poon have contributed to my projects tremendously and I thank them enormously.

I would also like to acknowledge Ludmila Pawlikowska and rotation student Joseph Lee for helping to design the Barcoding panel featured in Chapter 2. Andrew Sung and Jasmin Eshragh performed the barcode genotyping. Summer interns Lauren Lee and Yang Cao and rotation student Brett Johnson contributed to the sequencing projects in Chapter 4.

The lab wouldn't exist without Pui-Yan Kwok, my mentor. Thanks for giving me the chance to tackle such a large main project and for encouraging me to continue to work on the A20 side project. I might not have always been adequate at meeting the projects' goals, but I gave it my best. Thank you for your patience and calm leadership, something that is really rare to find.

Daniel Salomon and Steve Horvath were co-principal investigators on the kidney transplantation project. Daniel gave me free reign with the data as well as personal attention any time I had naïve questions about transplantation. He and his lab, especially Tony Mondala and Daniel Campbell, deserve all the credit for making the kidney project possible.

Lindsey Criswell and Averil Ma were co-mentors for the A20/TNFAIP3 project. I thank them for putting up with me, offering their expertise and guiding the direction of the project. Wilson Liao and Annie Poon contributed the control sequencing data used in Chapter 5. Nataliya Shiffrin and Timothy Lu from the Ma lab provided the NFκB response assay data in Chapter 6. I thank Kimberly Taylor for valuable input on Chapter 5 and she deserves all the credit for the TNFAIP3-SLE association study featured in Chapter 6 on which she was co-first author. This work was originally published in Nature Genetics:

Musone SL*, Taylor KE*, Lu T, Nititham J, Ferreira RC, Ortmann W, Shiffrin N, Petri MA, M. Kamboh I, Manzi S, Seldin MF, Gregersen PK, Behrens TW, Ma A, Kwok PY, Criswell LA. Multiple polymorphisms in the TNFAIP3 region are independently associated with systemic lupus erythematosus. *Nat Genet* **40**, 1062-4 (2008).

The bubble I have lived in over the past few years included an inner circle of classmates who made it tolerable; in fact, they made it quite enjoyable. Jennifer Yokoyama, Howard Horng, Jason Fernandes, Somayeh Ahmadiantehrani, Melissa Calton, Michael Hicks. These are lifelong friends who I thank for their

witty humor, generosity and acceptance of my desire to drive a bus. Connie Ha and Justin Chen, who I became friends with in the lab and who have since both moved on, have remained close friends who I can count on. While in San Francisco, I met my fiancé, David Dilworth, who has given sage advice and made me laugh way too often. It's a good thing we weren't in graduate school at the same time or the Earth might have imploded. Thanks for being somewhat patient, mostly helpful and almost willing to have chickens someday. Seriously, I look forward to a bright future together with you.

My family has been a quiet support system. I moved 3,000 miles away from them to attend graduate school and although it has been difficult being so far away, I couldn't thank them enough. I have a large extended family that inspires me to do my best. My two brothers, Jason and Christopher, are amazing people who have always encouraged me to do what I want. In the absence of our late father, they have selflessly helped to protect me and ensure that I pursue education. Our mother, Patricia, is the sweetest woman, who lends her ear to my terrible story-telling without judgment. Thank you for your unconditional support and love.

The Genetics of Kidney Transplantation Outcomes and Autoimmune Disease

Stacy Lynn Musone

Abstract

Kidney transplantation recipients face rejection despite anti-rejection drugs and matching efforts. Biopsy confirmed acute rejection (AR) and chronic allograft nephropathy (CAN) are 2 rejection phenotypes of interest. We conducted a genome-wide association study (GWAS) in European-derived kidney transplant donors and recipients. Well-functioning transplant donors (TX; N=261) were compared to AR (N=90) or CAN (N=105) participants. The same comparisons were conducted in recipients (TX N=226; AR N=71; CAN N=105). Analyses were adjusted for multiple comparisons and additionally for population substructure by including the first 2 multi-dimensional scaling dimension values as covariates in logistic regression. The most significant findings were identified in the TX vs. CAN tests (lowest unadjusted P=6.51E-08), which also displayed the largest odds ratios, and the least significant findings were identified for the TX vs. AR tests. Results need to be validated in an independent collection.

Proportion of identity by state, pi-hat, was calculated between donor-recipient pairs and compared between different donor-types and also by outcome (TX, AR, CAN). No significant difference within donor-type matched pairs was observed between outcome phenotypes in European-derived samples.

A set of primers was developed to sequence 112 candidate genes for AR and CAN. A custom resequencing tiling array was designed and tested. Technological development in the field called for testing the same panel on next-generation sequencing technology. We improved quality control metrics by trimming the reads and successfully called single nucleotide polymorphisms (SNPs) at a rate of 1/1000 bases sequenced.

Finally, TNFAIP3, a candidate gene for autoimmune disease (AID) was sequenced in samples multiply affected with AID (N=123) and in controls (N=397). One novel intronic insertion/deletion polymorphism was significantly associated with multiple AID diagnoses (Fisher's Exact P-value=0.0090; OR (95% CI)

7.053(1.67-29.79). Coding polymorphism rs2230926 was tested for association in a panel of individuals from families with multiple AIDs. Significant association was observed with all affected individuals (P=0.0336) as well as psoriasis, Crohn's disease and rheumatoid arthritis, with marginal association for Sjogren's and Graves disease. Additionally, we conducted an association study of the entire gene locus in lupus and identified 3 independent signals of association, including coding SNP rs2230926.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Appendix

LIST OF FIGURES

Chapter 2

Chapter 3

Chapter 4

Chapter 6

CHAPTER 1

INTRODUCTION TO KIDNEY TRANSPLANTATION AND GENETIC STUDIES

1.1. Kidney Transplantation

1.1.1. Transplantation Success and Failure

End-stage renal disease (ESRD), where the kidneys' ability to filter the blood has decreased to a level requiring mechanical filtration or transplantation, has a diverse etiology ranging from inherited disorders such as polycystic kidney disease to bacterial infections to hypertension. Over half a million people in the U.S. receive treatment for ESRD [1]. The most common cause of ESRD is diabetes, the incidence of which has dramatically increased over the last two decades [2]. Although patients could receive hemodialysis, kidney transplantation is the treatment of choice for patients suffering from ESRD. More than 15,000 kidney transplantations were performed last year in the U.S. while more than 70,000 patients are currently waiting for a suitable donor [1]. Acute rejection and chronic allograft nephropathy remain obstacles to post-transplantation health despite donor-recipient matching and immunosuppressive therapy.

The first successful kidney transplant, which was encouraged by the success of skin grafts between identical twins in 1937 [3], was performed in 1954 between monozygotic twins [4, 5] after a series of failed attempts between genetically dissimilar individuals. Therefore, a genetic component for the rejection response in transplantation has been observed since transplantation's beginnings.

Kidney donors can be either deceased or living. Although kidney transplantations with organs from deceased donors in the United States outnumber those from living donors, transplantations with grafts from living donors are more likely to be successful. The deceased-donor organ survival rate at one year after transplantation is 90%, whereas living donor organs have a 96% survival rate [1]. Startlingly, at 10 years post transplant, only 39% of deceased-donor organs and 57% of living-donor organs survive [1]. Patients with failed transplants can be re-transplanted or

put on dialysis. Transplant outcomes depend on several factors, including matching of donors to recipients, anti-rejection immunosuppressive therapy and genetic predisposition.

Several criteria are used to match donors and recipients. ABO blood type is the first simple screen, followed by human leukocyte antigen (HLA) matching. The classical approach to HLA testing is through crossmatching, a technique that tests the donor's blood antigens against the recipient's serum antibodies. Absence of reactivity between the two is an indicator that the recipient will not have an immediate and severe immune reaction against the transplanted organ. In addition, 3 HLA genes encoding Class I and Class II molecules are genotyped or sequenced in donors and recipients and checked for concordance. HLA matching improves kidney transplantation results by reducing the number of grafts lost by forty percent [6]. Centers across the U.S. implement slightly different combinations of these techniques to test for matching organs [6, 7].

Once transplanted, recipients receive medications that suppress their immune system, preventing it from fully rejecting the foreign allograft. Standard triple therapy consists of a calcineurin inhibitor (CNI, either Cyclosporin or Tacrolimus (FK506)), mycophenolate mofetil (a B and T cell proliferation inhibitor), and low dose prednisone [8]. A newer treatment option is Sirolimus, a mammalian target of rapamycin (mTor) inhibitor that can be used instead of a CNI, but it is not yet part of first-line standard therapy in renal transplantation [9]. This therapy regimen must be maintained for the remainder of the patient's life, except in the rare case of allograft acceptance. It is important to note that CNIs have nephrotoxic effects, thus slowly poisoning the very organ they are meant to protect [10].

Despite these matching techniques and drug treatments, many transplant recipients experience acute organ rejection (AR), which is mediated by T cells responding to donor organ antigens and can be treated well with a pulse dose of corticosteroids. AR typically occurs within the first three months post-transplantation. The rejection that occurs more slowly over time is referred to as

chronic allograft nephropathy (CAN), a complex phenotype that can be quantified with a Banff score on biopsy histology [11]. It is characterized histologically as interstitial fibrosis and tubular atrophy. This phenotype is thought to represent a compound effect of both the anti-rejection immunosuppressive medication and the body's immune response to the allograft.

1.2. Complex Human Genetics

1.2.1. Introduction to Genetic Association Studies of Complex Disease

The human genome is highly polymorphic and it is this variant nature that gives us human diversity. These polymorphic loci contribute not just to our physical appearance, but to our susceptibility to disease. Studies in families allow geneticists to link a genetic polymorphism to a particular trait or disease segregating in family members. Genetic association studies were proposed by Risch and Merikangas [12] in 1996 in order to move genetic analysis from family linkage studies towards population studies, whereby increased power would help identify contributors of modest effect in unrelated individuals. Single nucleotide polymorphisms (SNPs) are genotyped in cases and controls and tested for allele or genotype frequency differences between the two groups.

The common disease - common variant hypothesis began to be tested in large case-control collections [13]. Studies were conducted on candidate genes picked by geneticists from the literature based on a protein's known function. Many genes were tested and it became clear that selecting a gene based on prior knowledge had strengths and weaknesses. One could successfully choose a gene encoding a protein that would be significantly associated to the disease or trait of interest. However, genes would only be tested for which functional studies had been conducted. Other weaknesses of the candidate gene approach became apparent as either false positives or lack of measurable association bore out. Lack of replication became an increasing problem due to "the winner's curse," the phenomenon whereby the first reported association between a gene and disease overestimated the risk attributed by the identified variant. It became clear that large population studies were necessary to achieve the power to

avoid false negatives and replication in independent sample collections was required to ensure avoidance of false positives.

A nice example of the need for replication in independent collections and that large sample sizes are needed in candidate gene association studies is the case of PPARG in type II diabetes. An initial report of association of this variant had been followed up by 5 reports. Only 1 of these follow-up reports observed significant association with disease. A family-based study with replication in 3 independent collections of 16 previously-associated loci for type II diabetes failed to confirm association for all but the PPARG Pro12Ala SNP. Additionally, the combination of all previous reports revealed the modest effect of the variant and significant association with diabetes [14].

During this time, the technology for genotyping genetic variants advanced at a rapid rate, following Moore's law. Genotyping a dense set of markers distributed across the genome became affordable and the genome-wide association study (GWAS) was born. Over the past few years, GWAS have successfully identified polymorphic loci contributing to many common diseases including type II diabetes [15-19], breast cancer [20] and prostate cancer [21, 22]. The success of these studies lends confidence to the whole genome approach taken in this dissertation.

However, problems do exist with GWAS. Namely, samples sizes must typically be large in order to have enough power to identify truly associated variants with small effects. Also, false positive association due to population stratification must be dealt with as small frequency differences due to membership in subpopulations between cases and controls could be significant [23]. There are many ways to deal with this, including assigning individuals to a population based on clustering of genotypes to conduct a stratified or structured association [24], adding covariates from principal component analysis (PCA) or multi-dimensional scaling (MDS) to logistic regression [25], or testing for homogeneity of the case-control population based upon PCA or MDS values.

The many comparisons made in GWAS to test each SNP for association also needs to be taken into account. Genomic control correction, which uses the median test statistic across the study to adjust for both heterogeneity in the samples and multiple comparisons, is one method to address this concern [26]. A conservative approach would be to adjust the p-value for the number of tests conducted, Bonferroni correction, but this does not take into account the correlation amongst many SNPs and thus, the lack of independence between tests. Other methods include Sidak stepdown p-value correction, which due to its step-wise correction is less conservative than Bonferroni correction. Permutation testing allows scientists to calculate an empirical p-value for each SNP by swapping case-control status of individuals and calculating the number of tests rejecting the null hypothesis over many iterations. Finally, false discovery rate correction adjusts p-values based upon the proportion of tests expected to reject the null hypothesis of no association simply by chance [27].

In order to follow up findings from GWAS, results should first be validated in an independent collection. Other methods for further investigation of the associated locus include fine-mapping through additional genotyping. As SNPs are inherited together in blocks, or haplotypes, it is possible that the associated variant is in linkage disequilibrium (LD) with a causative or increased-risk bearing variant not directly measured in the panel of SNPs genotyped for the GWAS. This method may also identify multiple independent effects. Another method for follow-up is deep sequencing in cases and controls to identify rare variants in addition to common ones. Deep sequencing will identify variants and association testing reveals loci contributing to the measured trait. Rare variants for a common disease have successfully been identified in obesity [28] and type I diabetes [29].

1.2.2. Genetic Studies of Outcomes in Kidney Transplantation

Candidate gene approaches have reported associations of SNPs and microsatellite markers to both AR [30-69] and CAN [31, 33, 47, 57, 70-77] in donors and recipients, but many of the studies suffer from small sample sizes, varying phenotype definitions and lack of replication. Genes that

have been repeatedly studied include those encoding cytokines, such as chemokines, chemokine receptors and interleukins. At least one study has assessed the dynamic between donor and recipient polymorphisms outside of the HLA on transplant outcome [78, 79]. Ethnic differences in long-term graft survival have been noted. African Americans have poorer long term outcomes with kidney transplantation than all other ethnic groups [80-83]. One study has shown that it is not just access to care that contributes to this phenomenon [84] and another suggests that HLA mismatching is also not to blame [85]. Asians have a better long-term graft survival rate than all other ethnicities, with Caucasian and Hispanic patients having intermediate outcomes [82, 83].

1.3. Statement of Purpose

Here I have introduced the general concepts that will be featured in this dissertation, kidney transplantation and complex human genetics. Kidney transplantation outcomes do not solely depend upon HLA matching and differences in transplant outcome by ancestry and previously published association reports in individual genes hint at a genetic component to AR and CAN. We hypothesize that these traits are complex genetic traits that can be studied through population based case-control genetic approaches. It would be useful to take an unbiased genome-wide approach to help disentangle the genetic roots of these complex phenotypes in kidney donors and recipients towards gaining a better understanding of the underlying biology of rejection, identify potential new drug targets for anti-rejection treatment or predict those who are at risk of experiencing AR or CAN. Additionally, by taking advantage of the paired nature of transplantation and our genetic data, we could potentially impact future matching techniques.

We will conduct GWAS comparing TX versus AR or CAN or the combination of the two traits in donors and recipients separately, for a total of six GWAS in order to test our hypothesis that these traits are genetic in nature. Additionally, we will compare proportion of the genome shared between donor-recipient pairs by outcome to test the hypothesis that pairs with good outcomes share more of the genome than those with poor outcomes (AR or CAN). The rising incidence of ESRD and as a result, kidney transplantation, increases the urgency for a deeper biological

understanding of transplant outcomes and the genetic risk or protection contributed by both donors and recipients.

1.4. Summary of Chapters

In Chapter 2, I will highlight some of the technical issues one encounters when conducting a GWAS for any trait. We have developed a SNP barcoding panel tool to ensure that the samples applied to the genotyping arrays are the same as those in the original plate in which they arrived. We have also implemented tools, such as identity by state calculations to measure genetic relatedness amongst the samples that may be inappropriate, suggesting sample swaps. Finally, this chapter explores ancestry assignment tools in order to assign individuals to populations, thus avoiding spurious association due to underlying population stratification.

Chapter 3 encompasses all of the details of the GWAS for the two rejection traits in kidney transplantation, from DNA collection to association testing of SNPs, while accounting for the quality control metrics introduced in Chapter 2. Our search for variants contributing to AR and CAN in transplant donors and also in the recipients is, to our knowledge, the first GWAS conducted to date on these phenotypes. Chapter 4 delves into resequencing efforts to analyze particular genes in greater detail. In our study, these genes were selected from expression and proteomic studies conducted by collaborators, but this would also be the natural next step to analyze variants and closest gene neighbors identified in our GWAS.

Finally, Chapters 5 and 6 will take us to a different project altogether. This project focuses on a gene of interest in autoimmune disease, TNFAIP3, which encodes the protein A20. Chapter 5 details the sequencing of the gene in individuals each affected with multiple autoimmune diseases as well as controls. Specific genotyping of a non-synonymous variant is performed in multiple autoimmune disease collections to test for association. In Chapter 6, we perform association tests on the same candidate gene, TNFAIP3, by analyzing of genotype data surrounding the gene locus generated as part of a GWAS of systemic lupus erythematosus.

1.5. References

1. System, U.S.R.D., *USRDS 2009 Annual Data Report: Atlas of Chronic Kidney Disease and End-Stage Renal Disease in the United States*, N.I.o.D.a.D.a.K.D. National Institutes of Health, Editor. 2009: Bethesda, MD.

2. CDC's Division of Diabetes Translation, N.C.f.C.D.P.a.H.P., *National Diabetes Surveillance System: Incidence of Diabetes in the Population Aged 18-79 Years*. 2007, Centers for Disease Control and Prevention (CDC), National Center for Health Statistics, Division of Health Interview Statistics.

3. Brown, J., *Homografting of skin: with report of success in identical twins.* Surgery, 1937. **1**(558).

4. Guild, W.R., et al., *Successful homotransplantation of the kidney in an identical twin.* Trans Am Clin Climatol Assoc, 1955. **67**: p. 167-73.

5. Harrison, J.H., J.P. Merrill, and J.E. Murray, *Renal homotransplantation in identical twins.* Surg Forum, 1956. **6**: p. 432-6.

6. Takemoto, S., et al., *HLA matching for kidney transplantation.* Hum Immunol, 2004. **65**(12): p. 1489-505.

7. Goes, N. and A. Chandraker, *Human leukocyte antigen matching in renal transplantation: an update.* Curr Opin Nephrol Hypertens, 2000. **9**(6): p. 683-7.

8. Marcen, R., *Immunosuppressive drugs in kidney transplantation: impact on patient survival, and incidence of cardiovascular disease, malignancy and infection.* Drugs, 2009. **69**(16): p. 2227-43.

9. Flechner, S.M., et al., *De novo kidney transplantation without use of calcineurin inhibitors preserves renal structure and function at two years.* Am J Transplant, 2004. **4**(11): p. 1776-85.

10. Gaston, R.S., *Chronic calcineurin inhibitor nephrotoxicity: reflections on an evolving paradigm.* Clin J Am Soc Nephrol, 2009. **4**(12): p. 2029-34.

11. Solez, K., et al., *Banff 07 classification of renal allograft pathology: updates and future directions.* Am J Transplant, 2008. **8**(4): p. 753-60.

12. Risch, N. and K. Merikangas, *The future of genetic studies of complex human diseases.* Science, 1996. **273**(5281): p. 1516-7.

13. Lohmueller, K.E., et al., *Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease.* Nat Genet, 2003. **33**(2): p. 177-82.

14. Altshuler, D., et al., *The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes.* Nat Genet, 2000. **26**(1): p. 76-80.

15. Diabetes Genetics Initiative of Broad Institute of Harvard and, M.I.T., et al., *Genome-Wide Association Analysis Identifies Loci for Type 2 Diabetes and Triglyceride Levels.* Science, 2007: p. 1142358.

16. Scott, L.J., et al., *A Genome-Wide Association Study of Type 2 Diabetes in Finns Detects Multiple Susceptibility Variants.* Science, 2007: p. 1142382.

17. Sladek, R., et al., *A genome-wide association study identifies novel risk loci for type 2 diabetes.* Nature, 2007. **445**(7130): p. 881-885.

18. Zeggini, E., et al., *Replication of Genome-Wide Association Signals in U.K. Samples Reveals Risk Loci for Type 2 Diabetes.* Science, 2007: p. 1142364.

19. Consortium, T.W.T.C.C., *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature, 2007. **447**(7145): p. 661-678.

20. Easton, D.F., et al., *Genome-wide association study identifies novel breast cancer susceptibility loci.* Nature, 2007. **447**(7148): p. 1087-93.

21. Gudmundsson, J., et al., *Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24.* Nat Genet, 2007. **39**(5): p. 631-7.

22. Yeager, M., et al., *Genome-wide association study of prostate cancer identifies a second risk locus at 8q24.* Nat Genet, 2007. **39**(5): p. 645-9.

23. Campbell, C.D., et al., *Demonstrating stratification in a European American population.* Nat Genet, 2005. **37**(8): p. 868-72.

24. Pritchard, J.K., M. Stephens, and P. Donnelly, *Inference of population structure using multilocus genotype data.* Genetics, 2000. **155**(2): p. 945-59.

25. Patterson, N., A.L. Price, and D. Reich, *Population structure and eigenanalysis.* PLoS Genet, 2006. **2**(12): p. e190.

26. Devlin, B. and K. Roeder, *Genomic control for association studies.* Biometrics, 1999. **55**(4): p. 997-1004.

27. Storey, J.D. and R. Tibshirani, *Statistical significance for genomewide studies.* Proc Natl Acad Sci U S A, 2003. **100**(16): p. 9440-5.

28. Vaisse, C., et al., *Melanocortin-4 receptor mutations are a frequent and heterogeneous cause of morbid obesity.* J Clin Invest, 2000. **106**(2): p. 253-62.

29. Nejentsev, S., et al., *Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes.* Science, 2009. **324**(5925): p. 387-9.

30. Dutkiewicz, G., et al., *Lack of Association of Polymorphisms 239+34A/C in the SOD1 Gene and 47C/T in the SOD2 Gene With Delayed Graft Function and Acute and Chronic Rejection of Kidney Allografts.* Transplant Proc, 2009. **41**(9): p. 3701-3.

31. Brabcova, I., et al., *Genetic variability of major inflammatory mediators has no impact on the outcome of kidney transplantation.* Transplantation, 2007. **84**(8): p. 1037-44.

32. Alakulppi, N.S., et al., *Lack of association between thrombosis-associated and cytokine candidate gene polymorphisms and acute rejection or vascular complications after kidney transplantation.* Nephrol Dial Transplant, 2007.

33. Asderakis, A., et al., *Association of polymorphisms in the human interferon-gamma and interleukin-10 gene with acute and chronic kidney transplant outcome: the cytokine effect on transplantation.* Transplantation, 2001. **71**(5): p. 674-7.

34. Breulmann, B., et al., *Influence of cytokine genes polymorphisms on long-term outcome in renal transplantation.* Clin Transplant, 2007. **21**(5): p. 615-21.

35. Cartwright, N.H., et al., *A study of cytokine gene polymorphisms and protein secretion in renal transplantation.* Transpl Immunol, 2001. **8**(4): p. 237-44.

36. Canossi, A., et al., *Renal allograft immune response is influenced by patient and donor cytokine genotypes.* Transplant Proc, 2007. **39**(6): p. 1805-12.

37. de Reuver, P., et al., *Recipient ctla-4 +49 G/G genotype is associated with reduced incidence of acute rejection after liver transplantation.* Am J Transplant, 2003. **3**(12): p. 1587-94.

38. Dmitrienko, S., et al., *Immune response gene polymorphisms in renal transplant recipients.* Transplantation, 2005. **80**(12): p. 1773-82.

39. Hahn, A.B., et al., *TNF-alpha, IL-6, IFN-gamma, and IL-10 gene expression polymorphisms and the IL-4 receptor alpha-chain variant Q576R: effects on renal allograft outcome.* Transplantation, 2001. **72**(4): p. 660-5.

40. Hutchings, A., et al., *Association of cytokine single nucleotide polymorphisms with B7 costimulatory molecules in kidney allograft recipients.* Pediatr Transplant, 2002. **6**(1): p. 69-77.

41. Loucaidou, M., et al., *Cytokine polymorphisms do not influence acute rejection in renal transplantation under tacrolimus-based immunosuppression.* Transplant Proc, 2005. **37**(4): p. 1760-1.

42. Marder, B.A., et al., *The impact of costimulatory molecule gene polymorphisms on clinical outcomes in liver transplantation.* Am J Transplant, 2003. **3**(4): p. 424-31.

43. Marshall, S.E., et al., *The impact of recipient cytokine genotype on acute rejection after renal transplantation.* Transplantation, 2000. **70**(10): p. 1485-91.

44. Marshall, S.E., et al., *Donor cytokine genotype influences the development of acute rejection after renal transplantation.* Transplantation, 2001. **71**(3): p. 469-76.

45. Muller-Steinhardt, M., et al., *The interleukin-6 -174promoter polymorphism is associated with long-term kidney allograft survival.* Kidney Int, 2002. **62**(5): p. 1824-7.

46. Muller-Steinhardt, M., et al., *Cooperative influence of the interleukin-6 promoter polymorphisms -597, -572 and -174 on long-term kidney allograft survival.* Am J Transplant, 2004. **4**(3): p. 402-6.

47. Pelletier, R., et al., *Evidence for a genetic predisposition towards acute rejection after kidney and simultaneous kidney-pancreas transplantation.* Transplantation, 2000. **70**(4): p. 674-80.

48. Pawlik, A., et al., *IL-2 and TNF-alpha promoter polymorphisms in patients with acute kidney graft rejection.* Transplant Proc, 2005. **37**(5): p. 2041-3.

49. Poli, F., et al., *Tumour necrosis factor-alpha gene polymorphism: implications in kidney transplantation.* Cytokine, 2000. **12**(12): p. 1778-83.

50. Poli, F., et al., *TNF-alpha IFN-gamma IL-6, IL-10, and TGF-beta1 gene polymorphisms in renal allografts.* Transplant Proc, 2001. **33**(1-2): p. 348-9.

51. Sankaran, D., et al., *Cytokine gene polymorphisms predict acute graft rejection following renal transplantation.* Kidney Int, 1999. **56**(1): p. 281-8.

52. Slavcheva, E., et al., *Cytotoxic T-lymphocyte antigen 4 gene polymorphisms and susceptibility to acute allograft rejection.* Transplantation, 2001. **72**(5): p. 935-40.

53. Tinckam, K., et al., *The relative importance of cytokine gene polymorphisms in the development of early and late acute rejection and six-month renal allograft pathology.* Transplantation, 2005. **79**(7): p. 836-41.

54. Wang, J., et al., *IMPDH1 Gene Polymorphisms and Association With Acute Rejection in Renal Transplant Patients.* Clin Pharmacol Ther, 2007.

55. Wisniewski, A., et al., *Possible association of cytotoxic T-lymphocyte antigen 4 gene promoter single nucleotide polymorphism with acute rejection of allogeneic kidney transplant.* Transplant Proc, 2006. **38**(1): p. 56-8.

56. Wramner, L.G., et al., *Impaired kidney graft survival is associated with the TNF-alpha genotype.* Transplantation, 2004. **78**(1): p. 117-21.

57. Hoffmann, S., et al., *Donor genomics influence graft events: the effect of donor polymorphisms on acute rejection and chronic allograft nephropathy.* Kidney Int, 2004. **66**(4): p. 1686-93.

58. Azarpira, N., et al., *Influence of recipient and donor IL-10, TNFA and INFG genotypes on the incidence of acute renal allograft rejection.* Mol Biol Rep, 2008.

59. Azarpira, N., et al., *Vitamin D receptor genotypes and kidney allograft rejection.* Mol Biol Rep, 2009.

60. Gorgi, Y., et al., *Ctla-4 exon 1 (+49) and promoter (-318) gene polymorphisms in kidney transplantation.* Transplant Proc, 2006. **38**(7): p. 2303-5.

61. Gorgi, Y., et al., *Mannose binding lectin (+54) exon 1 gene polymorphism in tunisian kidney transplant patients.* Transplant Proc, 2009. **41**(2): p. 660-2.

62. Gorgi, Y., et al., *Human platelet antigens: HPA-1, -2, -3, -4, and -5 polymorphisms in kidney transplantation.* Transplant Proc, 2007. **39**(8): p. 2568-70.

63. Grinyo, J., et al., *Association of four DNA polymorphisms with acute rejection after kidney transplantation.* Transpl Int, 2008.

64. Hoffmann, T.W., et al., *Impact of a Polymorphism in the IL-12p40 Gene on the Outcome of Kidney Transplantation.* Transplant Proc, 2009. **41**(2): p. 654-6.

65. Nogueira, E., et al., *Incidence of donor and recipient toll-like receptor-4 polymorphisms in kidney transplantation.* Transplant Proc, 2007. **39**(2): p. 412-4.

66. Palmer, S.M., et al., *Donor polymorphisms in Toll-like receptor-4 influence the development of rejection after renal transplantation.* Clin Transplant, 2006. **20**(1): p. 30-6.

67. Sfar, I., et al., *The PTPN22 C1858T (R620W) Functional Polymorphism in Kidney Transplantation.* Transplant Proc, 2009. **41**(2): p. 657-9.

68. Lee, H., et al., *Influence of recipient and donor IL-1alpha, IL-4, and TNFalpha genotypes on the incidence of acute renal allograft rejection.* J Clin Pathol, 2004. **57**(1): p. 101-3.

69. Vamvakopoulos, J.E., et al., *Interleukin 1 and chronic rejection: possible genetic links in human heart allografts.* Am J Transplant, 2002. **2**(1): p. 76-83.

70. Brown, K.M., et al., *Influence of donor C3 allotype on late renal-transplantation outcome.* N Engl J Med, 2006. **354**(19): p. 2014-23.

71. Fekete, A., et al., *Association between heat shock protein 70s and toll-like receptor polymorphisms with long-term renal allograft survival.* Transpl Int, 2006. **19**(3): p. 190-6.

72. McLaren, A.J., et al., *Adhesion molecule polymorphisms in chronic renal allograft failure.* Kidney Int, 1999. **55**(5): p. 1977-82.

73. Pawlik, A., et al., *The FcgammaRIIa polymorphism in patients with chronic kidney graft rejection.* Transplant Proc, 2004. **36**(5): p. 1311-3.

74. Pawlik, A., et al., *The cytokine gene polymorphisms in patients with chronic kidney graft rejection.* Transpl Immunol, 2005. **14**(1): p. 49-52.

75. Ayed, K., et al., *Polymorphism of the renin-angiotensin-aldosterone system in patients with chronic allograft dysfunction.* Transpl Immunol, 2006. **15**(4): p. 303-9.

76. Azarpira, N., et al., *Angiotensinogen, angiotensine converting enzyme and plasminogen activator inhibitor-1 gene polymorphism in chronic allograft dysfunction.* Mol Biol Rep, 2008.

77. Ozaki, K.S., et al., *Improved renal function after kidney transplantation is associated with heme oxygenase-1 polymorphism.* Clin Transplant, 2008.

78. Freedman, B.I., et al., *Potential donor-recipient MYH9 genotype interactions in posttransplant nephrotic syndrome after pediatric kidney transplantation.* Am J Transplant, 2009. **9**(10): p. 2435-40.

79. Lacha, J., et al., *Effect of cytokines and chemokines (TGF-beta, TNF-alpha, IL-6, IL-10, MCP-1, RANTES) gene polymorphisms in kidney recipients on posttransplantation outcome: influence of donor-recipient match.* Transplant Proc, 2005. **37**(2): p. 764-6.

80. Eckhoff, D.E., et al., *Racial disparities in renal allograft survival: a public health issue?* J Am Coll Surg, 2007. **204**(5): p. 894-902; discussion 902-3.

81. Rudge, C., et al., *Renal transplantation in the United Kingdom for patients from ethnic minorities.* Transplantation, 2007. **83**(9): p. 1169-73.

82. Katznelson, S. and J.M. Cecka, *The great success of Asian kidney transplant recipients.* Transplantation, 1997. **64**(12): p. 1850-2.

83. Katznelson, S., D.W. Gjertson, and J.M. Cecka, *The effect of race and ethnicity on kidney allograft outcome.* Clin Transpl, 1995: p. 379-94.

84. Chakkera, H.A., et al., *Influence of race on kidney transplant outcomes within and outside the Department of Veterans Affairs.* J Am Soc Nephrol, 2005. **16**(1): p. 269-77.

85. Chertow, G.M. and E.L. Milford, *Poorer graft survival in African-American transplant recipients cannot be explained by HLA mismatching.* Adv Ren Replace Ther, 1997. **4**(1): p. 40-5.

CHAPTER 2

DATA CLEANING TOOLS FOR GENOME-WIDE ASSOCIATION STUDIES

2.1. Development of a DNA Barcode SNP Genotyping Panel

2.1.1. Introduction

DNA samples included in studies involving high-throughput SNP genotyping technologies could be mixed up at a number of steps in the sample preparation process. In our laboratory, we employ SNP genotyping microarrays that require a single DNA sample to be pulled from a 96-well plate and added to a microarray. As a precaution against mixing samples up, especially as it could swap our cases and controls, we have developed a DNA barcoding SNP genotyping panel to ensure DNA samples on the array are the same individual as in our starting 96-well plates. A 48-plex SNP genotyping technology was selected to allow for a moderately priced assay with enough resolution to identify individuals. It is expected that single DNA sample swaps will not occur within the 48-plex genotyping assay as it is performed in 384-well plates and all samples are handled with liquid handling robotics or 12-tip pipettes. Inversions made on a full plate or row are theoretically more easily identified by sex mismatches than an individual sample.

2.1.2. Materials and Methods

2.1.2.1. SNP Selection

We picked 2 SNPs from each chromosome that, according to HapMap, had at least a 30% minor allele frequency (MAF) in each of the 4 populations genotyped in the first phase of the International HapMap Project (European, Chinese, Japanese and Yoruban) [1]. Two chromosome Y SNPs were exceptions to this rule. The 48 chosen SNPs can be viewed in Table 2.1. For each chromosome, one SNP was amplified off of Nsp I digested DNA and the other off the Sty I fraction of the genome for the Affymetrix 500K Mapping Array genotyping assay used in our laboratory for genome-wide association studies. This set of SNPs is also present on future iterations of Affymetrix genotyping technology, including the Genome-Wide Human SNP Array 6.0, which we currently employ in the laboratory. In order to ensure the highest number of SNPs

passing quality control for barcode genotyping on the 48-plex SNPstream genotyping platform, we only selected A/G SNPs.

2.1.2.2. Genotyping

Genotyping was performed on the SNPstream (Beckman Coulter) instrument according to manufacturer's instructions with 5ng of DNA. Genotype clusters for all samples and SNPs were manually checked and adjusted by 384-well quadrant, or 96 sample batches. Genotype data was cleaned by first removing individuals with a genotyping rate less than 90%. This removed 367 of 1233 individuals, or 30% of all samples, leaving 866 for comparison to the Affymetrix genotypes. SNPs were then removed for having less than 90% genotyping which removed 3 SNPs out of 48.

Affymetrix genotypes were extracted for all 48 SNPs for individuals passing initial array QC (contrast QC >0.4). Genotypes were called with the birdseed-v2 algorithm in 2 large batches from Genome-Wide Human 6.0 intensity files for 15 plates of samples. Data was then cleaned using the 90% individual and SNP genotyping thresholds as above. This left 1157 individuals remaining out of 1160 and all 48 SNPs for comparison to the barcode panel genotypes.

2.1.2.3. Analysis

Affymetrix pedigree files were generated with custom scripts and dataset cleaning and merging was conducted with Plink v1.06 [2]. Affymetrix sample ID's were randomly reconfigured to assess specificity of barcode identification and merged in Plink to calculate non-missing mismatches between the two sets of genotypes.

2.1.3. Results

Thirteen duplicate samples were checked for concordance within the SNPstream genotyping assay and only one sample presented a problem (18/48 discordant genotypes). This could have been due to a second aliquot that was sent to our lab having actually represented a different DNA sample or, since this individual did not pass initial array QC for Affymetrix genotyping, it may

simply represent a poor quality DNA aliquot. As this sample failed initial QC, Affymetrix genotypes were never generated for this sample and concordance between technologies cannot be checked.

Nine duplicate samples were checked for concordance within the Affymetrix genotyping platform and, again, only one sample presented a problem by having 28/48 discordant genotypes. This sample initially had 27 discordant genotypes between the two technologies and was subsequently re-genotyped on the Affymetrix array (see details below). Eight hundred twenty-four samples overlapped between technologies; 42 were in the SNPstream dataset and not in Affymetrix while 291 were in Affymetrix and not SNPstream. The concordance rate across the 2 platforms was 99.5%. When randomly reassigning DNA sample ID's to the Affymetrix genotypes and then merging with the SNPstream barcode genotypes to assess specificity, concordance was 38.7% and discordant genotypes occurred for 823/824 individuals with a mean of 26.5 discordant genotypes per comparison ranging from 14-35.

Sixty-six samples had one or more discordant genotypes, 14 individuals had 2 or more discordant genotypes and 6 had greater than 2 discordant genotypes. Four samples had greater than 20 discordant genotypes and all would have been removed from GWAS analysis for sex mismatches (3; mismatch between genotype results and clinical data) or inappropriately high identity by state (IBS) with another sample (1 individual). One of the others had an Affymetrix call rate of 93%, which would have excluded the sample from association analysis for this study. The final sample was not re-genotyped and will be removed from association analysis. One sample was re-genotyped due to high discordance between the technologies; the first comparison yielded 27/45 discordant genotypes (40% concordant). The re-genotyped Affymetrix sample and SNPstream genotypes were 100% concordant across the 45 SNPs compared.

2.1.4. Discussion

Through the comparison of over 800 samples between our SNP Barcode Panel and Affymetrix genotypes, it was revealed that 14 individuals had 2 or more discordant genotypes. This is less than 2% of the sample and represents a similar rate to sex mismatches, which are theoretically able to resolve up to 50% of sample swaps due to the binary nature of the trait. These 14 samples probably got swapped with another sample at a step in the microarray genotyping process when DNA is individually pipetted from a 96-well plate onto a chip. We were able to correct the genotyping of 1 sample and exclude 1 other due to discordant genotypes between the sample barcode and microarray assays. However, all other samples with high discordance were identified as problematic due to sex discordance or inappropriately high IBS with another non-related sample in the dataset. Sex concordance and IBS checking amongst samples require no extra genotyping or cost expenditure when already performing genome-wide SNP genotyping. Therefore, it is a tool that is not much more useful at identifying sample swaps than the combination of sex concordance and IBS checking in our partially related dataset. It may represent an important quality control metric in fully unrelated sets of samples.

							HapMap MAF by Population			
rs Number	**Probe Set ID**	**Chr**	**BP**	**Strand**	**Enzyme**	**Flank**	Japanese	Han Chinese	CEPH	Yoruba
rs4649343	SNP_A-2069004	1	231842102	-	NSP	acaatgccacgttcac[A/G]agatgaccaattgcct	0.4889	0.4333	0.4500	0.3100
rs10915413	SNP_A-1805954	1	5271439	+	STY	gtccgtgagaaagaaa[C/T]ggtagagtcagcgtag	0.3556	0.3889	0.4600	0.4674
rs7597996	SNP_A-1922789	2	226212802	+	NSP	ttagagtgtgtttcaa[C/T]gcatcctaataacctg	0.3667	0.3889	0.3300	0.3900
rs896222	SNP_A-1961806	2	2686842	+	STY	aaatgtgctctgaaaa[A/G]tcgtgttggcaattgt	0.3778	0.3556	0.4400	0.3600
rs9858096	SNP_A-2026694	3	195401596	-	NSP	aaacaatcaaataagc[A/G]acagttctgatcccaa	0.4667	0.3667	0.4100	0.5000
rs11128506	SNP_A-2130000	3	1041688	+	STY	ataccaagctacaata[C/T]acatctgaagacttac	0.4444	0.4667	0.4600	0.4694
rs2118922	SNP_A-1957474	4	178404068	-	NSP	ttttataaatacgaaa[C/T]gttggcactagctaat	0.4556	0.4762	0.4681	0.4667
rs3775816	SNP_A-2122575	4	20144815	+	STY	agtgcttatatggatc[C/T]gactaccctgtcctgc	0.4889	0.4000	0.4000	0.3400
rs566750	SNP_A-2304659	5	134513184	+	NSP	gctctgagggccgcac[A/G]atgttctggtgaattg	0.4268	0.3095	0.3021	0.4896
rs200107	SNP_A-2196325	5	8708690	+	STY	catgttactgattcca[A/G]tcttggttcacaaatc	0.3333	0.4091	0.4400	0.3900
rs2981956	SNP_A-1929767	6	167619694	+	NSP	ttcatcatatcgagta[C/T]ctttccagtctgtctg	0.3111	0.3444	0.3900	0.4700
rs4145201	SNP_A-2011607	6	17007422	-	STY	tgtcagagcggacgac[C/T]gggtttgaactgaaag	0.3889	0.3889	0.3061	0.3400
rs4298423	SNP_A-1866651	7	151274842	+	NSP	tcccctaattccttca[C/T]gtggttattctagctt	0.4333	0.4444	0.3200	0.3900
rs10279220	SNP_A-2070008	7	7923172	+	STY	acacagtggtttctga[A/G]gatccaaccacatatt	0.4778	0.4111	0.4000	0.3500
rs1383474	SNP_A-2162252	8	136203829	+	NSP	attttaggaatttgat[A/G]agattcagtagcacat	0.4205	0.3068	0.4694	0.3367
rs895695	SNP_A-2149411	8	3232222	-	STY	atgtactagatagaac[A/G]actgtccttcaacttt	0.3000	0.3111	0.4700	0.4800
rs1335259	SNP_A-1860556	9	120464167	-	NSP	tcagttgatcttaaca[C/T]ggaaagagagacaggt	0.4667	0.4000	0.3000	0.3100
rs10963302	SNP_A-1995548	9	1783472	-	STY	agatccaaatggaaag[A/G]ccaggtaggaagctag	0.4432	0.3409	0.4583	0.3980
rs2997238	SNP_A-2143559	10	122484709	-	NSP	atgactgattttgaca[C/T]tccttgggtctcttct	0.3333	0.3889	0.3878	0.3500
rs10159718	SNP_A-2040264	10	1129183	-	STY	aaccaccaaatcgcta[C/T]agtttaggaccccagg	0.4000	0.4111	0.3800	0.4800
rs623823	SNP_A-2284126	11	133492804	+	NSP	gatatgacagtggagt[A/G]aaacaaagcaaacgaa	0.4778	0.4333	0.3700	0.3000
rs11030008	SNP_A-2237097	11	3815192	+	STY	ctctcaagctctaaca[A/G]tccttgtgcttgggga	0.4111	0.3333	0.3700	0.4400
rs12812747	SNP_A-2167421	12	113859041	+	NSP	caacagttttgctatc[A/G]catgaccttagatgac	0.4556	0.3444	0.4082	0.3700
rs10773982	SNP_A-4273463	12	1698410	+	STY	aactgctgtttgattc[A/G]aaatgattctttgtgt	0.4444	0.4333	0.3061	0.4184
rs9527109	SNP_A-1942111	13	53060483	+	NSP	ttggattagaacttca[A/G]gaattgttcactattg	0.3111	0.3444	0.4300	0.3000

							HapMap MAF by Population			
rs Number	**Probe Set ID**	**Chr**	**BP**	**Strand**	**Enzyme**	**Flank**	Japanese	Han Chinese	CEPH	Yoruba
rs7317204	SNP_A-1901244	13	23603705	+	STY	tggatgctgaaatgtc[A/G]agatttacaagacatt	0.3718	0.4375	0.5000	0.3776
rs876561	SNP_A-2165440	14	91014664	+	NSP	caaagcacattacagg[A/G]aagcaaagatttaaca	0.4444	0.4889	0.3900	0.3700
rs11623278	SNP_A-1855505	14	31857898	-	STY	tcaagtgttgatctta[A/G]agcaggtaaatactaa	0.3750	0.4889	0.4700	0.3100
rs12900029	SNP_A-2162848	15	93754271	+	NSP	attttgctgtgttaaa[A/G]ctgtgtatcctgaact	0.3778	0.4111	0.3900	0.3100
rs792419	SNP_A-1800827	15	32840541	-	STY	gtagactgagatgaaa[C/T]gagatgggtgtttgca	0.4535	0.3182	0.3061	0.3333
rs11150186	SNP_A-2072939	16	78250169	+	NSP	acaggaaggcatttca[A/G]taatcattcagcgagg	0.4556	0.4444	0.3100	0.4700
rs17680913	SNP_A-2101078	16	9192461	-	STY	agatcaccttcacttg[A/G]agaaatgttacttcac	0.4048	0.4222	0.4896	0.3400
rs4789786	SNP_A-2174296	17	78097641	-	NSP	cacacttacaaggcga[A/G]ctgcaccgagaggtga	0.4000	0.4333	0.4900	0.4900
rs7217233	SNP_A-1782901	17	8556959	+	STY	ggcagagtagttgtta[C/T]gtaatctcaggagagt	0.3444	0.4222	0.4600	0.4000
rs7504842	SNP_A-1805015	18	74150981	-	NSP	gtcatccttacgaaac[A/G]ctcacattagctaagg	0.4302	0.4878	0.5000	0.4688
rs751355	SNP_A-1898123	18	9418420	+	STY	tgctgtgtgtttactc[C/T]agaacgccctgctcac	0.3667	0.4333	0.4900	0.3400
rs2075415	SNP_A-1954760	19	37173086	+	NSP	attggttttcctgcca[C/T]ctatttcgctaggaga	0.4000	0.4268	0.3085	0.4900
rs1055919	SNP_A-4280892	19	4803137	+	STY	agcgggagcccatgac[C/T]gattggacgccgccgg	0.4186	0.4865	0.5000	0.4239
rs6015552	SNP_A-1936767	20	57706242	+	NSP	ttctttatggcaatcc[A/G]agaatgaactaatgca	0.4778	0.3556	0.4000	0.3200
rs879012	SNP_A-2300641	20	957788	-	STY	gcctgccacagggaac[A/G]actatcctgtccccat	0.4091	0.4222	0.4400	0.4898
rs2070435	SNP_A-2020783	21	46786139	-	NSP	ttagaccgttcatcga[C/T]aggactgcttttcttg	0.5000	0.4659	0.3800	0.3700
rs2822859	SNP_A-2015285	21	15005394	-	STY	aatggttaagcacaaa[C/T]agttgatctctgaatt	0.4333	0.3667	0.3600	0.3400
rs2272789	SNP_A-1848186	22	34009951	+	NSP	tatcttctatgcggca[C/T]tagccaaaaaagaaga	0.3333	0.3556	0.3100	0.3000
rs361594	SNP_A-1935007	22	16957338	-	STY	ttctgaggaccataca[A/G]atcggttttccaccta	0.4444	0.4889	0.3000	0.5000
rs4829909	SNP_A-2093866	X	136811935	+	NSP	agaagatttgaaagag[A/G]atggggacctatgatt	0.3778	0.4186	0.3469	0.3469
rs2694710	SNP_A-1922765	X	3163474	-	STY	ttagaggtgattatca[A/G]tatagtctatcagttt	0.4205	0.3721	0.4898	0.4490
rs17250121	SNP_A-8332143	Y	19296941	+	NSP	gttaagtgtttataga[C/T]taggtctgtttctccc	0.4348	0.1364	0.3000	0.0000
rs1276034	SNP_A-8390603	Y	22393444	+	STY	tcaaaggtatatgtca[A/G]aagacccaaaatcctc	0.0000	0.0000	0.3333	0.0000

Table 2.1: Barcode Panel SNP Information. 50 SNPs included in barcode genotyping assay. Rs Number: reference sequence number; ProbeSet ID - Affymetrix SNP identification number, Chr - chromosome; BP - base pair position; Enzyme - Indicates from which restriction digestion a SNP is amplified for Affymetrix assay; MAF - minor allele frequency.

2.2. Comparing Identity by State Values to Detect Sample Errors

2.2.1. Introduction

Identity by state (IBS) values tell one how alike any two samples are to one another. It is equivalent to adding together the proportion of SNP genotypes with 2 alleles shared and those sharing 1 allele across the genome to gain the proportion of the total genome shared between the pair. In our study, we have many donor-recipient pairs who are not related and some that are related. We can use the proportion of the genome shared IBS, termed pi-hat, to detect outlier values that indicate sample mix-ups, which could occur as DNA is individually aliquoted into a DNA plate for the first time or when we aliquot a prepared sample onto a genotyping microarray.

2.2.2. Materials and Methods

Samples were genotyped with one of two Affymetrix genotyping microarrays – Human Mapping 500K Array Set or Genome-Wide Human 6.0. Genotypes were called with BRLMM (500K) or Birdseed-v2 (6.0) and subjected to a 95% sample and SNP call rate. Linkage disequilibrium (LD), or correlation amongst SNP genotypes, was calculated in 50 SNP bins in windows sliding 2 SNPs forward after each bin with a variance inflation factor (VIF) threshold of 2. VIF = $1 / (1-R^2)$, where R^2 is the correlation between SNPs. A set of 426,476 SNPs was pruned to 147,201 for 1277 individuals. IBS was determined with the calculation of pi hat between all samples as follows: Pi Hat (Proportion IBD) = P (IBD=2) + 0.5*P (IBD=1), where P = proportion and IBD = identity by descent. P (IBD=2) refers to the proportion of genotypes with 2 shared alleles between the 2 individuals being compared as P (IBD=1) refers to 1 shared allele between the 2 individuals. LD and IBS calculations were made in Plink.

2.2.3. Results

Mean pi-hat between any two individuals is 0.0367, but the median pi-hat was 0 indicating that the mean was skewed higher by a set of samples with high IBS. Figure 2.1 is a histogram of pi-hat values for comparisons between all 1277 individuals (814,726 comparisons) displayed for those comparisons with pi-hat ≤ 0.15 (N=814,238). The mean pi-hat rose to 0.199 when

narrowing the comparisons to donor-recipient pairs (N=585 pairs). Living related donor pairs shared 0.484 while unrelated pairs shared .080 of their genome. Cadaverous donor pairs shared 0.037; similar to the value for any two randomly compared individuals. Twenty-four samples were marked for removal from the study for having inappropriately high IBS with another sample. Pi-hat of 0.3 or greater was considered inappropriately high except for related donor-recipient pairs, for whom a pi-hat greater than 0.9 was considered inappropriate. Two pairs of samples with inappropriate IBS were re-genotyped on the Affymetrix 6.0 array and pi-hat was re-calculated. One of the samples had been swapped during the first experiment, because upon re-genotyping the pi-hat with its living donor was 0.5 and the inappropriate IBS with the other pair was removed. IBS analysis was also able to identify and confirm double-donor or double-recipient samples that were aliquoted twice, once for each donor-recipient pair. Double donors with double aliquots had a mean pi-hat of 0.997 (N=15).

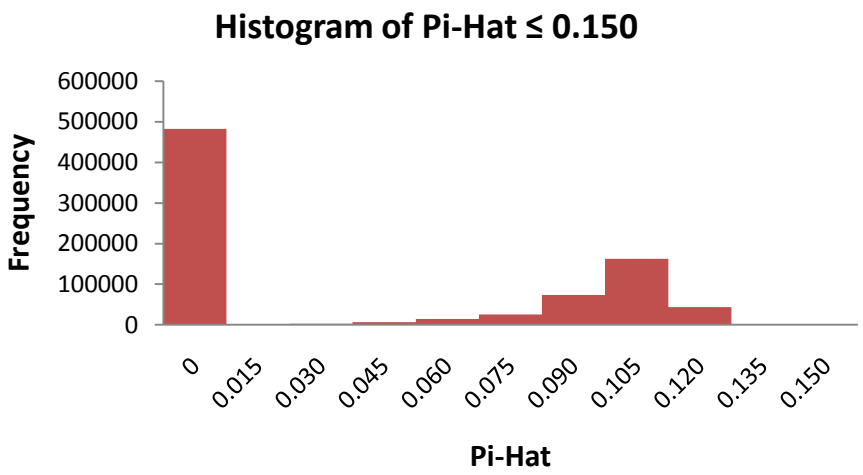


Figure 2.1: Histogram of Pi-hat for Combinations ≤ 0.150. N=1277; Displaying pi-hat for 814,238 of 814,726 comparisons made. Mean pi-hat is 0.0367; median is 0. All individuals were compared against all other individuals in the collection.

2.2.4. Discussion

IBS comparisons are able to identify DNA sample swaps in our study by taking advantage of the relatedness between many living donor pairs. It would not be able to identify swaps of samples from unrelated pairs. Further, we are able to confirm the relatedness or unrelatedness of living

donor-recipient pairs as occasionally pairs are misclassified. Although many studies do not knowingly include related individuals, this analysis would be useful in identifying any duplicate samples and unknowingly related samples.

2.3. Ancestry Analysis

2.3.1. Introduction

Small differences in allele frequencies due to population stratification can cause one to falsely identify a variant as associated with a trait of interest when the difference is simply due to the cases and controls being members of different populations or subpopulations [3, 4]. In order to avoid and correct for this, we implemented a method of assigning individuals to a population based on ancestry informative markers (AIMs). This will assign individuals to a population based on an assumption of the number of underlying clusters or populations. Additionally we will use multidimensional scaling (MDS) to help control for smaller, usually intercontinental, differences when conducting association testing in one ancestral group.

2.3.2. Materials and Methods

2.3.2.1. STRUCTURE

Samples were genotyped and cleaned for sample and SNP call rate as in section 2.2.2, but without LD pruning. Analysis included 1277 individuals. 2,230 unique AIMs were selected from three journal articles for European Americans [5], African Americans [6] and Latino populations [7]. Nine hundred seventy-two of the SNPs were genotyped on our microarray and 631 passed a call rate cutoff of 95%. Of the 631 SNPs used in this analysis, 33 are from the African American panel, 460 from the Latino panel and 139 from the European panel of markers. One SNP overlaps the European and Latino panels. Data were formatted in Plink. STRUCTURE [8] was used to estimate membership in a population, assuming 4 clusters (K) exist. These 4 clusters correspond to continents – Africa, the Americas, Asia, and Europe. Initially we ran 20 iterations each for K=1 to K=10 using a burnin rate of 10,000 with 10,000 reps. Other parameters were set

to default such that an admixture model was used with correlation between SNPs. K=4 fit the data best.

2.3.2.2. Multi-Dimensional Scaling

MDS is a method used to measure the distance between objects, in this case the objects are individual people. A matrix of similarity is calculated from pairwise distances (IBS). Various levels of the matrix, or dimensions, explain different things such as population substructure. We implemented MDS calculation for 1277 individuals in Plink with samples genotyped, cleaned and LD pruned as in Section 2.2.2.

2.3.3. Results

2.3.3.1. STRUCTURE

The ancestral groups assigned through this analysis fit the clinical ethnicity information well, but not perfectly. The membership coefficients for each of the 4 clusters, separated by reported ethnicity information, can be found in Table 2.2. Figure 2.2 is a colored bar plot of the 4 clusters also organized by clinical ethnicity. Hispanics were mostly categorized as a mixture of European and Native American ancestry while African Americans were a mix of African and European ancestry. The majority of samples are of Caucasian, or European, ancestry. Mean membership proportions were used to assign an ancestral group to all individuals such that assignments for GWAS analysis are based upon empirical data and to include samples of unknown ethnicity in populations for inclusion in analysis.

	Ancestral Group & Corresponding K			
	Africa	**Asia**	**Europe**	**Americas**
Clinical Information (N)	**1**	**2**	**3**	**4**
Asian (49)	0.0060	**0.8943**	0.0824	0.0172
African American (142)	**0.8749**	0.0086	**0.1114**	0.0053
Hispanic (184)	0.0570	0.0363	**0.4461**	**0.4606**
Native American (7)	0.0013	0.0117	0.0381	**0.9486**
Caucasian (799)	0.0191	0.0112	**0.9545**	0.0153

Table 2.2: Reported Ethnicity and STRUCTURE Assigned Ancestral Group Membership. Columns 1-4 represent a continental ancestral group and each row displays a particular clinically

assigned ethnicity group with corresponding mean membership coefficient for K 1-4. Cells with >10% membership are bolded. Data not shown for samples assigned as other ethnicity or unknown in clinical database (N=96). Total N=1277.

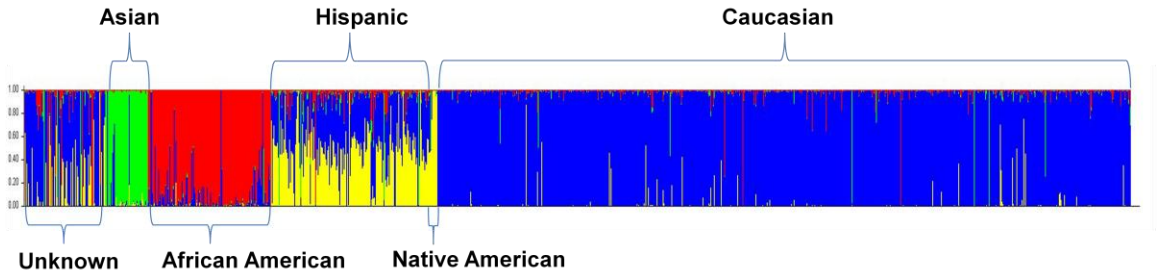


Figure 2.2: Bar Plot of Ancestral Group Membership Proportions. Each color represents a different cluster and vertical bars represent individual samples colored by the proportion of membership in one of the 4 clusters (K) assigned during STRUCTURE analysis. Labels are clinical ethnicities assigned upon enrollment. N=1277.

2.3.3.2. Multidimensional Scaling

The variance explained by each dimension in our LD pruned genotype dataset is shown in Figure 2.3, where the fraction explained levels off after the 3rd dimension. This means that the majority of variance will be explained by dimensions 1-3. MDS separates 3 ancestral populations - Africa, Europe and Asia - very well with just the first 2 dimensions, where Native Americans cluster very close to Asians (Figure 2.4). Even though it is well established that African Americans are admixed with European ancestry, these individuals form their own cluster due to their high proportion of African ancestry. Hispanic individuals, another admixed population, can be a mixture of these 3 ancestral populations in various proportions, as demonstrated in Figure 2.5 where Hispanic are overlaid onto the same plot from Figure 2.4. The 3rd dimension separates Native Americans from Asians (Figure 2.6) while the 4th dimension explains intra-European stratification (Figure 2.7). Population membership for each individual was determined with STRUCTURE in the previous section of this chapter.

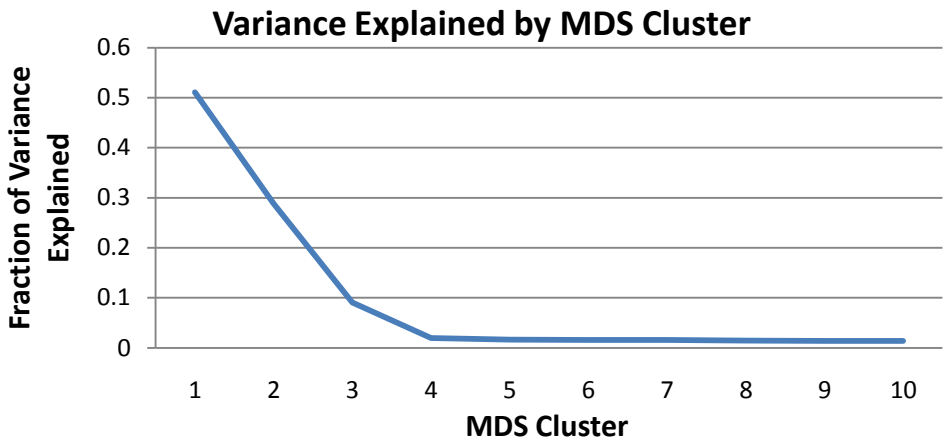


Figure 2.3: Variance Explained by MDS Clusters 1-10. MDS dimensions calculated on linkage disequilibrium pruned SNP dataset for 1277 individuals. MDS – multidimensional scaling.

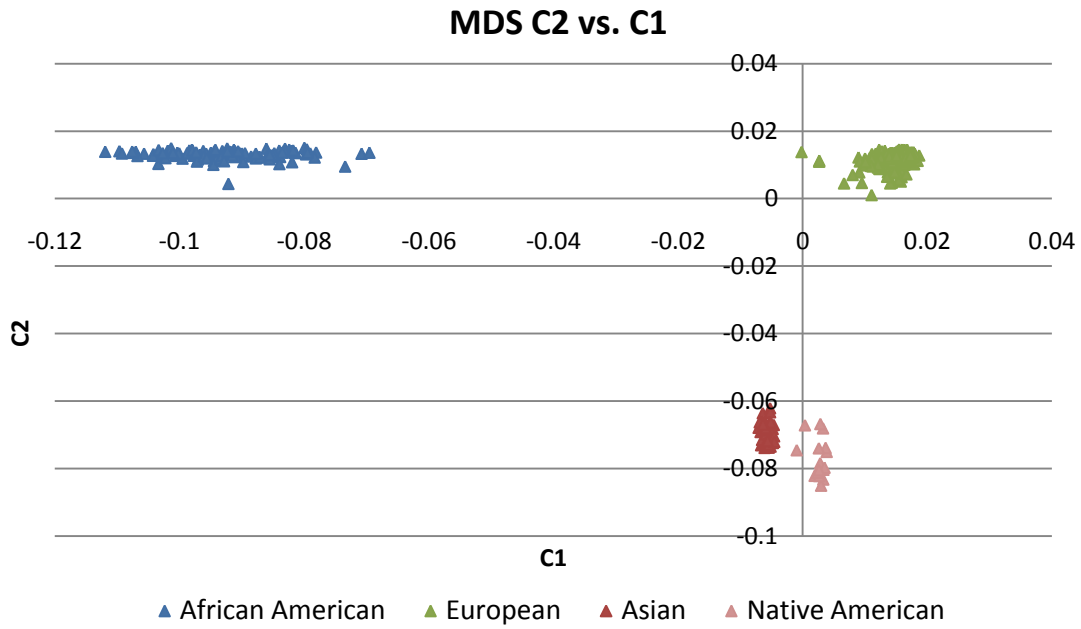


Figure 2.4: MDS C2 vs. C1 for 4 Structure Assigned Populations. Position on the first 2 MDS dimensions plotted for each individual to show separation of 4 populations listed above. Population assignment determined with STRUCTURE analysis. MDS - Multidimensional scaling; C - Cluster or dimension. N=968.

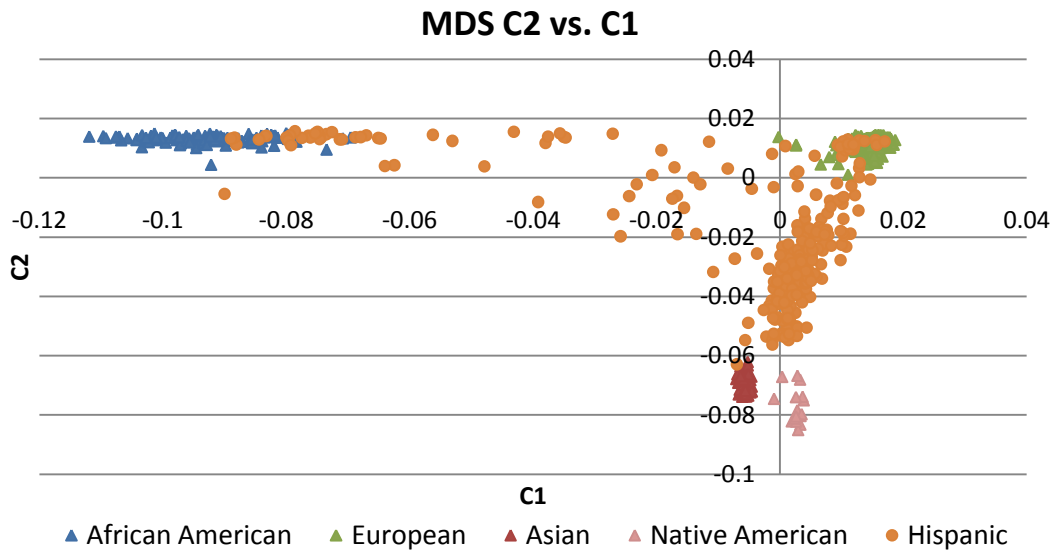


Figure 2.5: MDS C2 vs. C1 for Structure Assigned Populations with Hispanics. Position on the first 2 MDS dimensions plotted for each individual to show separation of 5 populations listed above. Population assignment determined with STRUCTURE analysis. MDS - Multidimensional scaling; C – Cluster or dimension. N=1277.

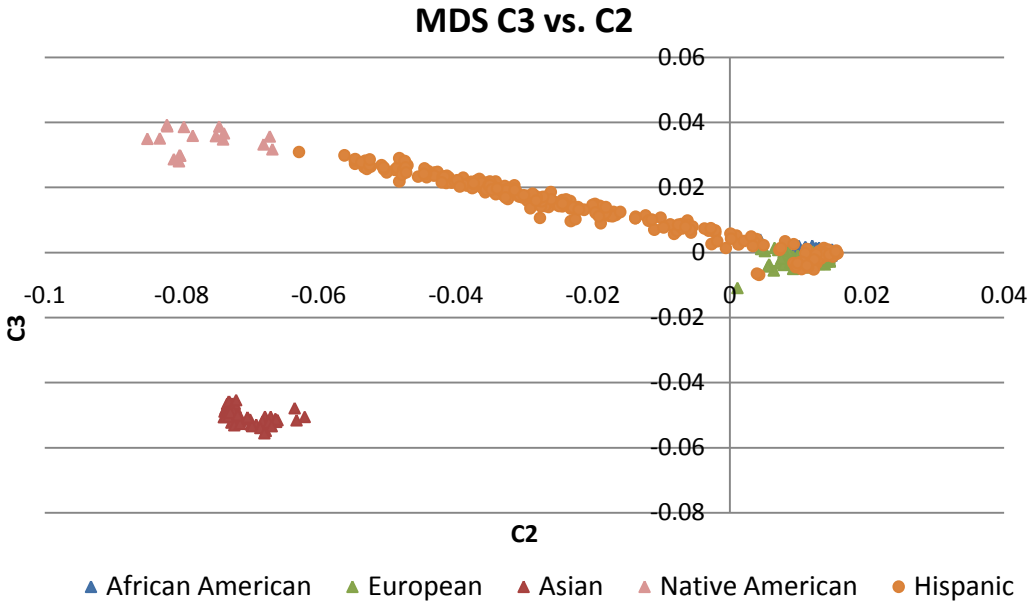


Figure 2.6: MDS C3 vs. C2 for Structure Assigned Populations with Hispanics. Position on the 2nd & 3rd MDS dimensions plotted for each individual to show separation of 5 populations listed above. Population assignment determined with STRUCTURE analysis. MDS - Multidimensional scaling; C – Cluster or dimension. N=1277.

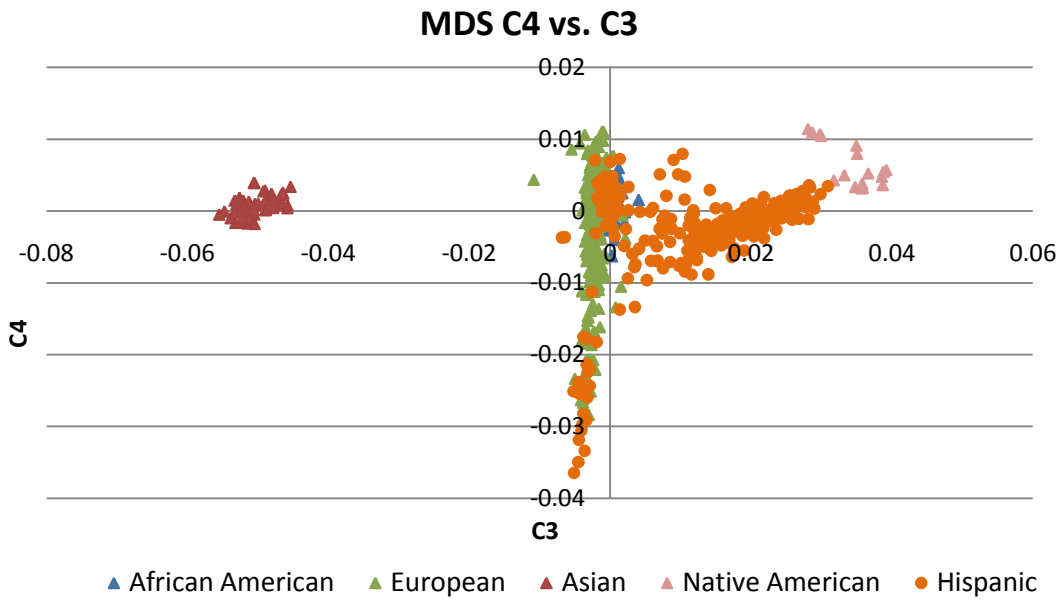


Figure 2.7: MDS C4 vs. C3 for Structure Assigned Populations with Hispanics. Position on the 3rd & 4th MDS dimensions plotted for each individual to show separation of 5 populations listed above. Population assignment determined with STRUCTURE analysis. MDS - Multidimensional scaling; C – Cluster or dimension. N=1277.

2.3.4. Discussion

This section clearly demonstrates the genetic differences between populations. It also demonstrates that in genetic studies it is important not to assume membership in a population based on clinical assignments as some individuals will cluster much better with a different population. This is best illustrated in Figure 2.2 where some bars, or individuals, are 100% different color, or population, from the other individuals in their clinically assigned ethnicity group. Also, one can resolve membership in a population even further than a singular assignment with the implementation of MDS. If the dimensions explaining much of the variance are used as covariates in logistic regression, easily implemented in Plink, MDS should help correct for subtle population substructure within an ancestral population such as Europeans.

2.4. References

1. The International HapMap, C., *A haplotype map of the human genome.* Nature, 2005. **437**(7063): p. 1299-1320.

2. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.

3. Campbell, C.D., et al., *Demonstrating stratification in a European American population.* Nat Genet, 2005. **37**(8): p. 868-72.

4. Clayton, D.G., et al., *Population structure, differential bias and genomic control in a large-scale, case-control association study.* Nat Genet, 2005. **37**(11): p. 1243-6.

5. Price, A.L., et al., *Discerning the ancestry of European Americans in genetic association studies.* PLoS Genet, 2008. **4**(1): p. e236.

6. Smith, M.W., et al., *A high-density admixture map for disease gene discovery in african americans.* Am J Hum Genet, 2004. **74**(5): p. 1001-13.

7. Price, A.L., et al., *A genomewide admixture map for Latino populations.* Am J Hum Genet, 2007. **80**(6): p. 1024-36.

8. Pritchard, J.K., M. Stephens, and P. Donnelly, *Inference of population structure using multilocus genotype data.* Genetics, 2000. **155**(2): p. 945-59.

CHAPTER 3

GENOME-WIDE ASSOCIATION STUDY OF ACUTE REJECTION AND CHRONIC ALLOGRAFT NEPHROPATHY

3.1. Introduction

Over 15,000 kidney transplants are performed in the United States each year and, despite donor-recipient matching techniques, acute rejection (AR) and chronic allograft nephropathy (CAN) remain obstacles to post transplant health [1]. We have conducted a genome-wide association study (GWAS) in European-derived donors and recipients to identify polymorphisms associated with these outcomes when compared to healthy transplants (TX). We have also compared donor-recipient pi-hat, or proportion of the genome shared identically by state (IBS), between outcome groups by donor type. The 3 donor types in this study are living related (LRD), living unrelated (LURD) and cadaverous (CAD).

Like linkage studies in families, GWAS allows one to enter the realm of human genetic analysis without any prior knowledge or assumptions about which areas of the genome might be related to predisposition towards a particular trait. As family studies of rejection phenotypes are not feasible, GWAS is the best choice for an agnostic genetic study for this project. In GWAS, the whole genome is probed, leading to the possible identification of SNPs in or near genes and pathways never previously implicated in the phenotype of interest and potentially revealing novel areas of investigation that could inform basic research or guide future treatment protocols or drug discovery efforts.

We have rigorously defined our phenotypes and required all participants to undergo kidney biopsy to confirm their phenotype with histology. This ensures that other groups can attempt to replicate our findings on their own sample collections while utilizing the same phenotype definitions. We applied several methods of multiple comparisons correction and controlled for population stratification.

3.2. Materials and Methods

3.2.1. DNA Collection

Study participants of various ethnic backgrounds (Caucasian, African American, Hispanic, Asian, Native American) were enrolled from eleven centers throughout the U.S. and recipients have been followed-up between 12 and 24 months post transplant for clinical assessment of outcome and protocol biopsies. Perfectly matched HLA living donor pairs were excluded. Donor and recipient blood was collected and sent to a centralized location at Scripps Research Institute in La Jolla, California for DNA extraction with Qiagen's QIAamp DNA Blood Midi Kit per manufacturer's instructions. Kidney transplants with the proper consent, anti-rejection regimen including a calcineurin inhibitor, and without active immune-related disorders, type I or type II diabetes, chronic active hepatitis, human immunodeficiency virus, cytomegalovirus, BK nephritis or bacterial pyelonephritis were elected for inclusion. Samples categorized as acute dysfunction, no rejection (ADNR) were excluded. AR and CAN were confirmed through biopsy and histology read by a single pathologist following Banff criteria [2].

Additional criteria for each of the three phenotypes are as follows:

Acute Rejection: Recipients within the first year of transplantation with a serum creatinine at least 25% above established baseline with biopsy proven tubulointerstitial cellular rejection with or without vascular rejection. Additional exclusion criteria are anatomical obstruction, vascular compromise, hemolytic uremic syndrome, and drug intensification within two weeks prior to biopsy. The symptoms shall also not be due to dehydration or drug effects or toxicity.

Chronic Allograft Nephropathy: Patients at least one year post-transplant with a serum creatinine at least 25% above established baseline as determined by a minimum of 3 measurements over at least 2 months and with a greater than 15% decrease in creatinine clearance from baseline. Additional exclusion criteria are a serum creatinine greater than 3.5mg/dl, poorly controlled hypertension (>130/80), anatomical obstruction, vascular compromise, and recurrent or de novo glomerulonephritis or focal segmental glomerulosclerosis. The symptoms shall also not be due to dehydration or drug effects or toxicity.

Normal Functioning Graft Without Rejection (TX): Patients at least one year post-transplant with at least 3 serum creatinine readings over a 3 month period that change less than 20% and lack a pattern of increasing levels. Women must have a serum creatinine level less than or equal to 1.5mg/dl; men must have a level less than or equal to 1.6mg/dl. Subjects must have a creatinine clearance of at least 45ml/min. Additional exclusion criteria are AR, CAN or nephropathy by biopsy, a history of rejection, acute dysfunction and poorly controlled hypertension (>130/80).

3.2.2. Genotyping

Subjects were genotyped on the Affymetrix Human Mapping 500K Array Set or Genome-Wide Human SNP Array 6.0 according to the manufacturer's instructions. In brief, this involved enzymatic digestion, ligation of an adapter, single primer amplification of ligated DNA segments, product clean up, random fragmentation and labeling the DNA before hybridization onto the array. After sample hybridization, arrays were washed and stained before being scanned with a laser to record intensity values. Mapping 500K Arrays had to pass DM call rate threshold of 92% and 6.0 arrays had to have Quality Control (QC) contrast >=0.4 to be elected for genotype calling. Genotypes were determined with BRLMM (500K) or Birdseed-v2 (6.0) in the Affymetrix Power Tools Suite. The 500K genotyping was performed in 1 large batch for Nsp arrays and 1 large batch for Sty arrays and then merged. The average sample call rate was 97.9%.

Genotyping of 6.0 samples was performed in small batches. Initially we genotyped these samples in 2 large batches and also genotyped them in smaller 1 or 2 plate batches with Birdseed-v2. The mean call rate for the large batch genotyping was 98.50% with a standard deviation of 1.50. Small batch genotyping produced a mean genotyping rate of 99.13% with a standard deviation of 0.99. Approximately 21,000 more SNPs passed a 95% call rate threshold with small batch genotyping versus large batches (881,843 vs. 860,691 SNPs). Small batch genotyping was chosen for its increased sample and SNP call rates.

3.2.3. Data Quality Control

Individuals and SNPs were subjected to a 95% call rate threshold. Duplicates passing this call rate were merged for 500K and the sample with a higher call rate was kept for 6.0. Samples with discordant sex between clinical and genetic data were removed from the study. Samples with > 2 discordant genotypes between 50 overlapping SNPs on Nsp and Sty 500K arrays were removed and re-genotyped. A custom barcode genotyping panel was performed for samples run on 6.0 arrays and those with >2 discordant genotypes were re-genotyped on an Affymetrix array or removed from the study. Unrelated samples with pairwise identity by state (IBS) pi-hat > 0.3 and related donor-recipient pairs with IBS pi-hat > 0.9 were removed from the study.

We scanned a total of 2,495 microarrays (554 Nsp / 548 Sty 500K; 1,393 6.0). Four hundred fifty five 500K samples advanced to genotyping with the BRLMM algorithm. After merging duplicates and removing samples failing the 95% call rate threshold, 434 samples remained. For 6.0, 1,283 samples advanced to genotyping and 1,279 passed the 95% call rate threshold. Between the two array types, 40 samples were removed for sex mismatch, 25 for inappropriately high IBS, 8 for sex mismatch and high IBS, 1 for inappropriately high IBS and Barcode mismatches, 10 duplicates, 15 double-donor double-aliquots, 1 for sex mismatch and not being part of the study, and 35 for not being part of the study. This included reference samples that were genotyped on each plate as a positive control for the microarray process and samples removed after initial enrollment. This left 1,578 samples remaining. Before accounting for ancestry, the number of samples for each array type and outcome can be seen in Table 3.1.

Contrast QC values were correlated with genotyping call rate for 6.0 samples (Figure 3.1). Duplicate samples had a mean concordance rate of 99.75% for 500K and 99.31% for 6.0 genotyping; 99.53% when both samples surpassed the 95% call rate threshold. Duplicate sample concordance was correlated with call rate of the pair (Figure 3.2). This demonstrates the importance of having a call rate threshold as genotypes are less accurate for samples with lower

call rate. This was probably due to a lack of ability to resolve intensity differences between genotypes and also explains the correlation between contrast QC values and genotype call rates.

Another QC parameter to consider is heterozygosity, as a raised level may indicate sample contamination. It is important to take population membership into consideration for setting heterozygosity thresholds as it differs between populations (Figure 3.3). Calculating heterozygosity means within populations, as determined by Structure, no samples fell outside a 95% confidence interval, thus none were removed.

For our analysis, a merged set of overlapping SNPs from the 500K and 6.0 Affymetrix genotyping microarrays was used. The 500K array began with a set of 500,618 SNPs that shrunk to 453,647 after implementing a 95% call rate threshold. The 6.0 array contained 909,622 SNPs to begin with and 881,843 after call rate QC. These two sets of markers were merged and resulted in a set of 431,326 SNPs genotyped on all 1,578 DNA samples. The average SNP call rate was 0.988 with a mean minor allele frequency of 0.21 (Histograms for both in Figure 3.4 and Figure 3.5, respectively).

To check for systematic differences between the two array types, an association study was conducted between them. All 500K genotyped samples were compared to all 6.0 genotyped samples. An excess of SNPs was observed in the tail of the p-value distribution visualized on a Q-Q plot (Figure 3.6). Using a Bonferroni adjusted p-value cutoff of $P<1.16x10^{-07}$ for 431,326 SNPs, 109 SNPs were marked for examination during association testing of our traits of interest.

Phenotype	500K	6.0	Total
CAN	88	256	344
AR	80	207	287
TX	224	658	882
AR/CAN	1	1	2
No Outcome	8	55	63
TOTAL	401	1177	1578

Table 3.1: Outcome Phenotypes for Each Genotyping Array Type. Samples with No Outcomes did not have an outcome assigned at the time of analysis. CAN – chronic allograft nephropathy; TX – good outcome; AR – acute rejection. 500K and 6.0 refer to Affymetrix Genotyping Microarray products used in this study.

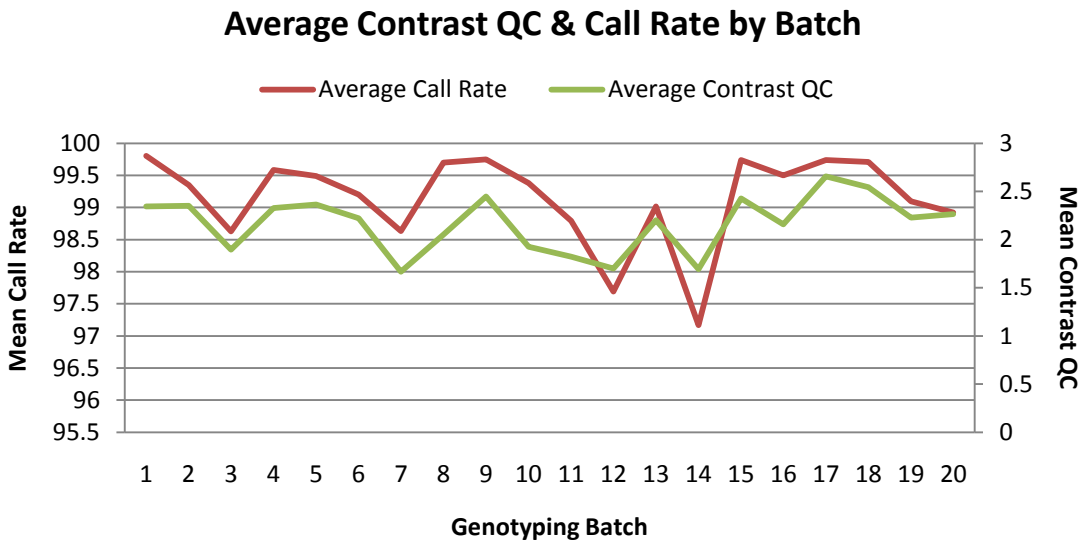


Figure 3.1: Average Contrast QC & Call Rate by Batch. Mean genotyping call rate is displayed in red and labeled on the left axis. Mean contrast QC is displayed in green and the axis is labeled on the right of the figure.

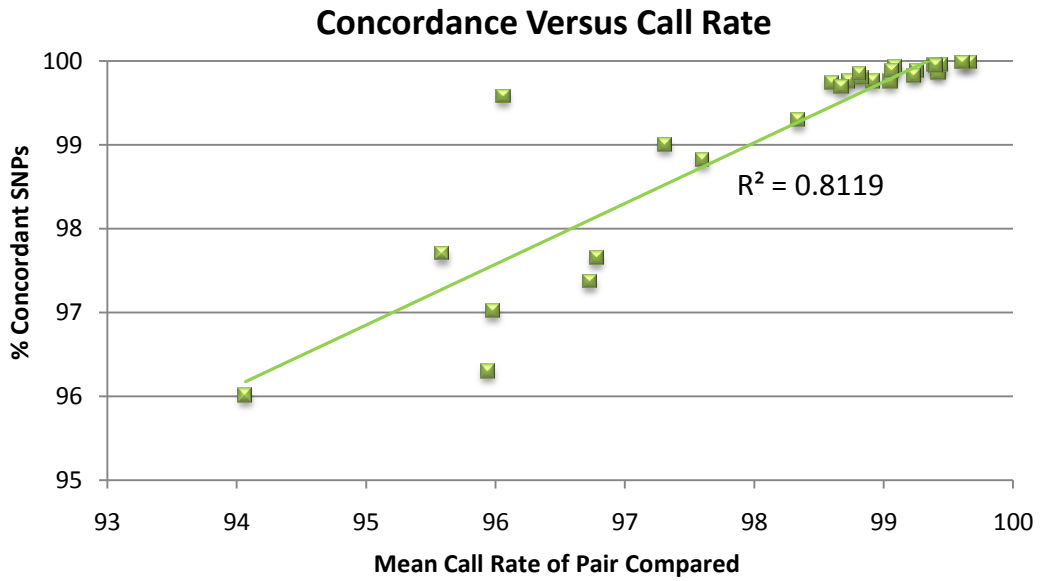


Figure 3.2: Duplicate Sample Concordance versus Call Rate. Data is shown for 27 6.0 pairs. The call rate displayed is the mean call rate between the two samples being checked for concordance.

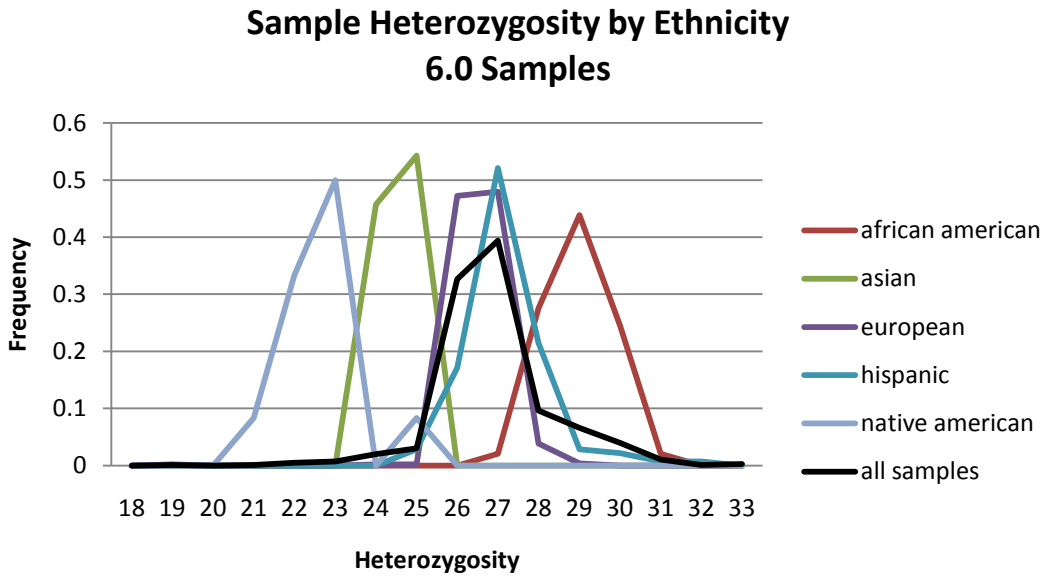


Figure 3.3: Heterozygosity by Population. Data shown for 6.0 samples, but a similar trend is observed for 500K samples. Population membership was determined by STRUCTURE.

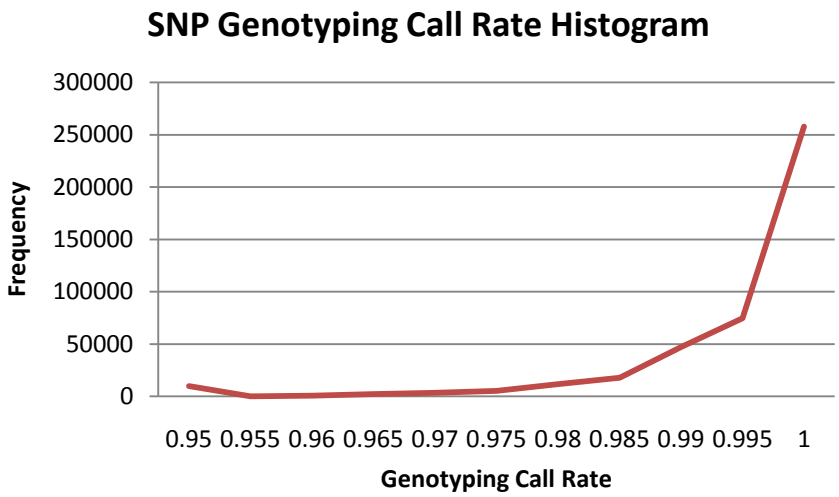


Figure 3.4: SNP Genotyping Call Rate Histogram. Data shown is for 95% European subjects only. N = 431,326 SNPs and 883 individuals. Mean call rate is 0.988.

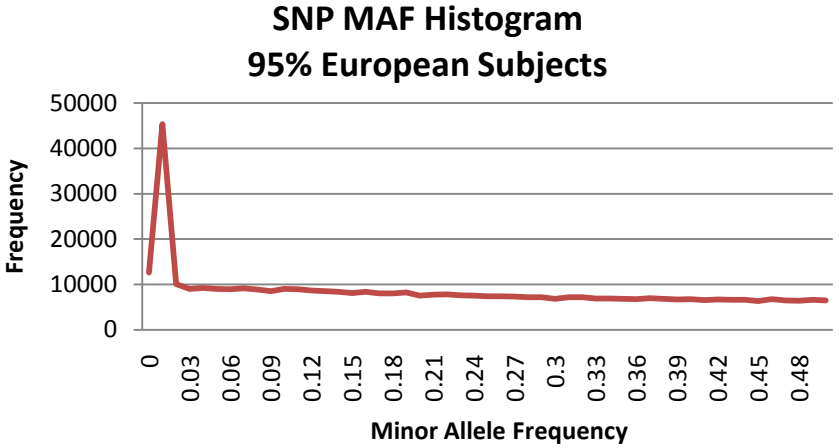


Figure 3.5: SNP Minor Allele Frequency Histogram. Mean MAF = 0.21. N = 431,326 SNPs and 883 individuals. MAF – minor allele frequency.

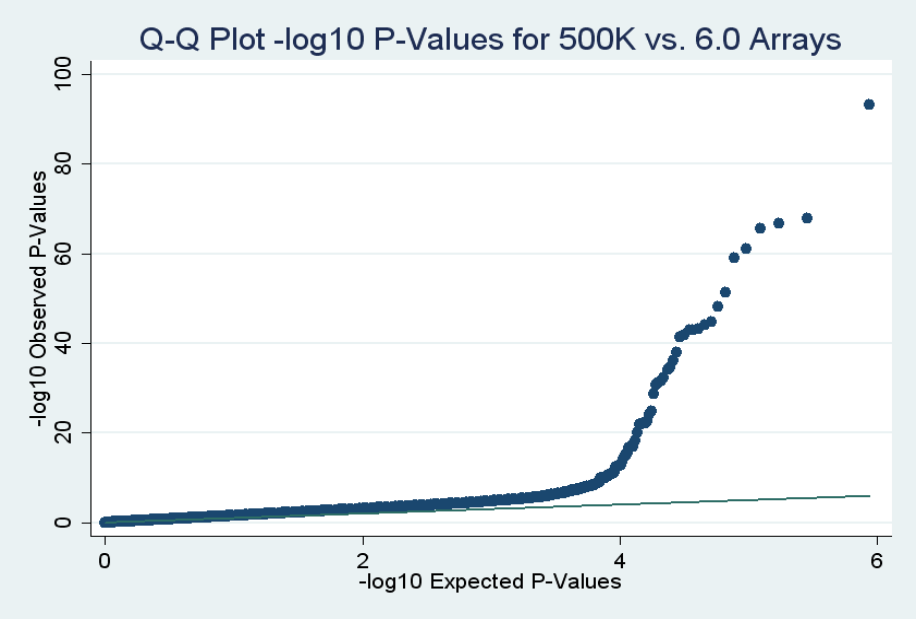


Figure 3.6: Q-Q Plot of –log10 P-Values for 500K vs. 6.0 Genotype Association Test. 431,326 SNPs compared. Observed –log10 P-Values are plotted on the Y axis and expected P-values are on the X axis. Each dot represents a SNP. This test compared all samples genotyped on a 500K array versus all 6.0 genotyped samples to identify between platform differences.

3.2.4. Ancestry Testing

Membership in an ancestral population was determined through the method discussed in Section 2.3.3.1 whereby a subset of SNPs was analyzed with Structure, a tool used to group people together based on genotype calls (Figure 3.7) [3]. Individuals with at least 95% membership in the European population were selected for this primary analysis as it is the largest population in our study (N=903), making up 57% of the total sample. After sample QC, the number was reduced to 883.

Multidimensional scaling (MDS), analogous to principal components analysis, was also used to calculate dimensions explaining population strata. MDS was calculated on an LD pruned set of SNPs as in Section 2.3.2.2 for the full set of individuals and for the subset of European individuals. The number of SNPs after pruning for LD was 252,009. The first 2 dimensions for the full set of individuals are plotted in Figure 3.8 where individuals are colored by their membership in the 95% European cluster or not. One notes that the 95% European group clusters tightly

together. The variance explained by each MDS Cluster within the Europeans is displayed in a scree plot in Figure 3.9. Most of the variation can be explained by the first 2 dimensions, after which the “elbow” bends and only a small fraction of additional variation is explained by the remaining 10 clusters displayed in the figure. Compared to variance explained per dimension in the full sample collection, the variance explained per dimension for Europeans is quite small. In Chapter 2, Figure 2.3, one can see that the first dimension explains greater than 50% of the variance, whereas in Europeans the first dimension explains approximately 0.13 of the variance. The first 3 European MDS dimensions are plotted and color coded by transplant outcome in Figure 3.10 to ensure that the phenotypes are distributed evenly along the MDS axes. One TX individual appears to be an outlier in the plot of C1 and C2 and two TX individuals are outliers when plotted for C2 and C3, where C refers to dimensions, or clusters.

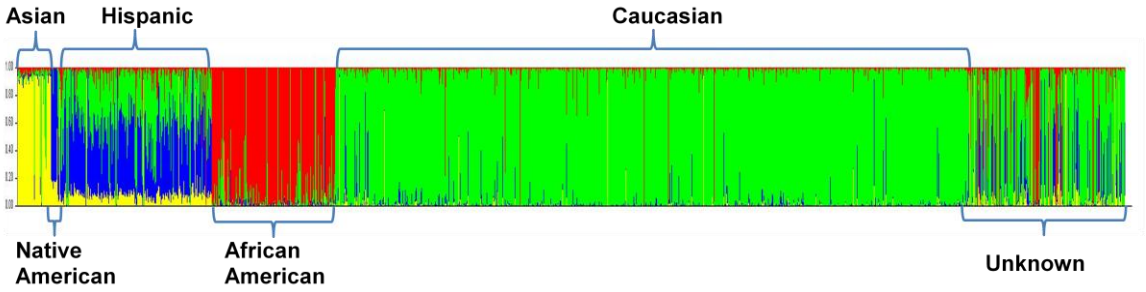


Figure 3.7: Bar Plot of Ancestral Group Membership Proportions. Each color represents a different cluster and vertical bars represent individual samples colored by the proportion of membership in one of the 4 clusters (K) assigned during STRUCTURE analysis. Labels are clinical ethnicities assigned upon enrollment. N=1,697.

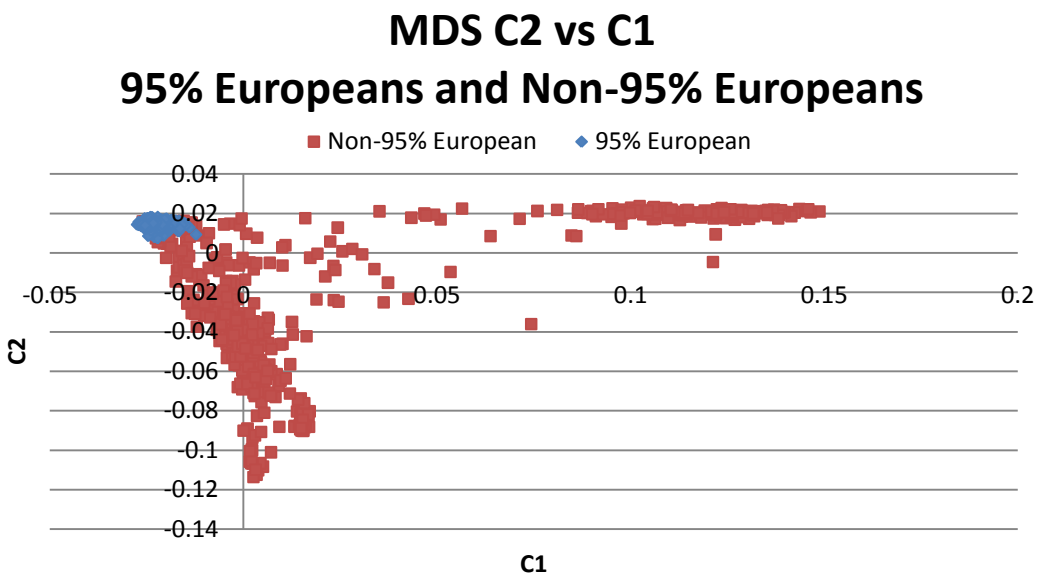


Figure 3.8: MDS C2 vs. C1 for 95% Europeans and Non-95% Europeans. Displays tight group of 95% European samples in blue versus all other samples colored in red. N=1,578. 95% European membership determined through Structure analysis. MDS – multi-dimensional scaling; C – Cluster or dimension.

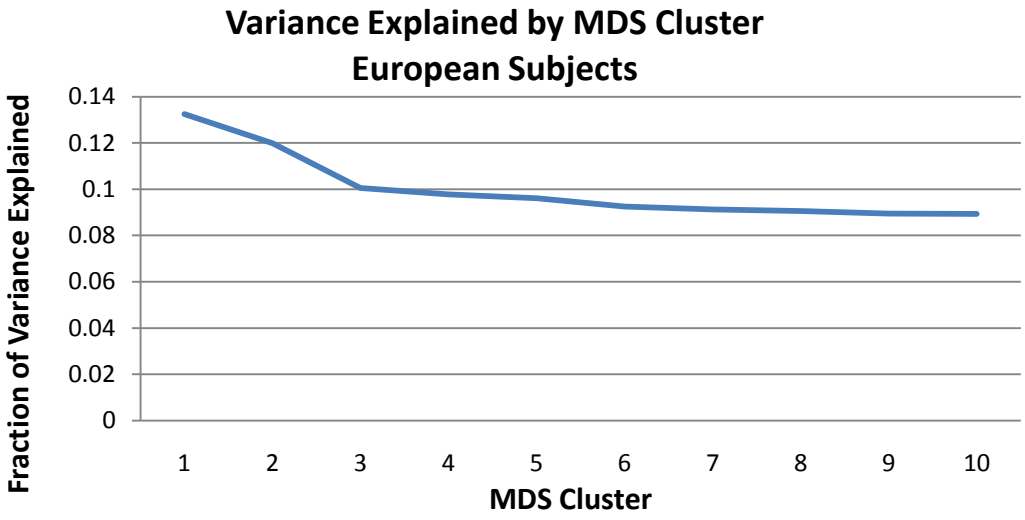


Figure 3.9: Variance Explained by MDS Clusters 1-10 within 95% European Subjects. 95% European membership determined through Structure analysis. MDS dimensions calculated on linkage disequilibrium pruned SNP dataset for 883 individuals. MDS – multidimensional scaling.

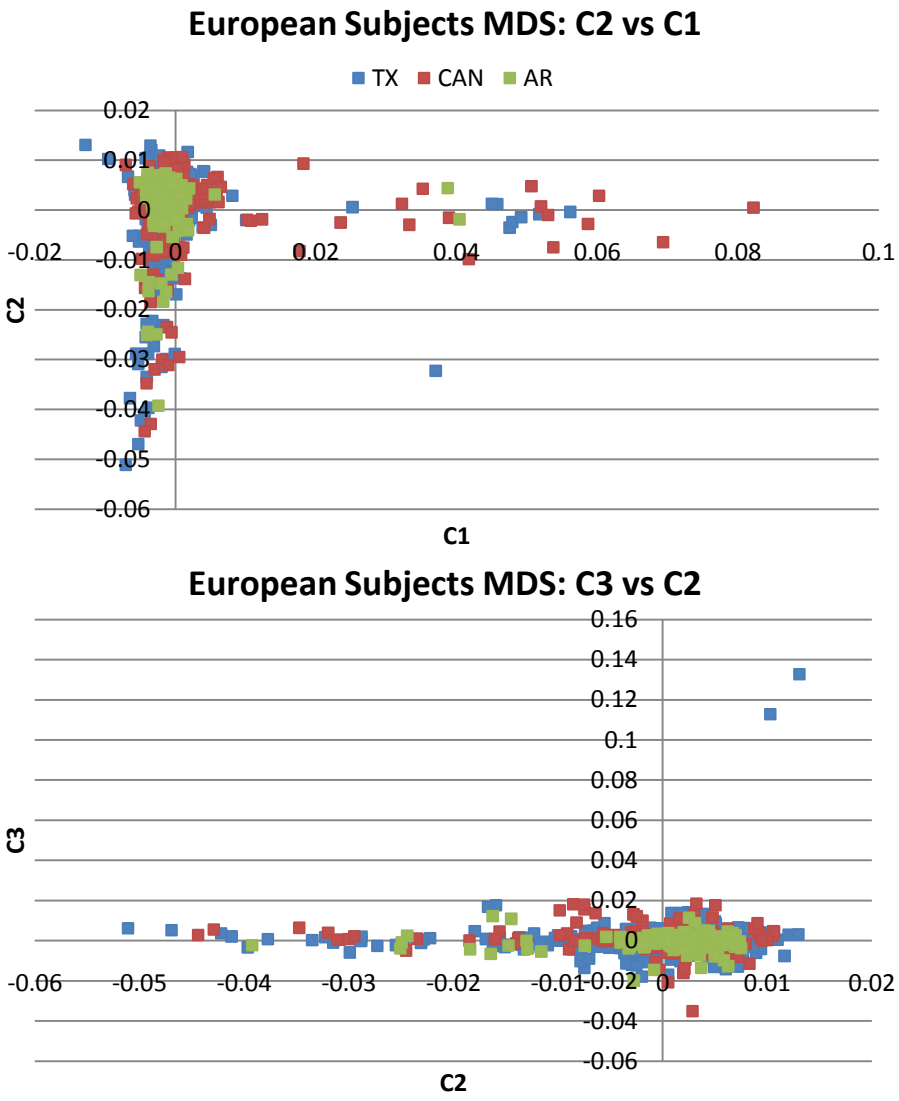


Figure 3.10: First 3 MDS Dimensions within Europeans by Outcome Phenotype. N = 883. CAN – chronic allograft nephropathy; TX – good outcome; AR – acute rejection.

3.2.5. Tests for Association

Chi-square test of association between cases (either AR or CAN or a combination of the 2) and controls (TX) were conducted in donors and recipients separately, as in Table 3.2 in Plink [4]. Out of the 883 samples passing QC metrics, including ancestry analysis, 857 had known outcome phenotypes. SNPs with a minor allele frequency less than 1% were removed before association testing (57,967 removed). Hardy-Weinberg Equilibrium p-values were calculated

separately for each comparison and in cases and controls and SNPs with a p-value < 10^{-05} were removed. This resulted in the removal of 51 SNPs for all donor tests and 55 for all recipient tests. The number of SNPs tested was 373,308 for donor tests and 373,304 for recipient tests with an individual genotyping rate of 99.4%.

It is important to implement tools to avoid false positives due to population substructure and multiple comparisons [5, 6]. We applied a simple correction for stratification with genomic control, a global correction method based on median chi square for all tests in a comparison [7]. We also adjusted p-values with several methods for multiple comparison corrections, including Sidak step-down adjustment and false discovery rate, both implemented in Plink [8]. Finally, we performed logistic regression with the first 2 MDS dimensions as covariates in order to adjust for subtle population stratification. We consider our primary analyses to be allelic chi-squared tests of association with unadjusted p-values.

Donors		Recipients	
Controls (TX) N	**Cases (N)**	**Controls (TX) N**	**Cases (N)**
261	AR (90)	226	AR (71)
	CAN (105)		CAN (105)
	AR+CAN (194)		AR+CAN (176)

Table 3.2: Number of Donors and Recipients for Each of 3 Outcomes. Displays the 6 GWAS comparisons, 3 for donors and 3 for recipients. N - number of samples; TX – well functioning transplant (controls); AR – acute rejection (cases); CAN – chronic allograft nephropathy (cases).

3.2.6. IBS Calculation between Pairs by Outcome

Ancestry was determined in Structure with a set of AIMs and samples with greater than 95% European ancestry were included in this analysis (N=903). LD was calculated in bins of 50 SNPs shifting 2 SNPs forward after each bin; SNPs with greater than 0.8 r^2 were removed, pruning the SNP set from 431,326 to 252,914. IBS proportions, termed pi-hat, were calculated between every sample pairing as in Chapter 2, Section 2.2.3. Random pairs of individuals with pi-hat >0.3 were determined to share an inappropriate level of identity and were both removed from the study. Pairs sharing more than 0.9 IBS were also removed from the study in addition to individuals

whose genetically determined sex did not match the clinical data entry. Pi-hat was compared between outcome groups (TX, AR, CAN) within donor-type classes (LRD, LURD and CAD) in Stata with analysis of variance (ANOVA).

3.3. Results

3.3.1. Association Testing of SNPs

We conducted 6 GWAS with our dataset, comparing well functioning transplants (TX) to AR, CAN or the combination of AR and CAN for both donors and recipients of kidney transplants. The genomic inflation factor for all comparisons was very close to 1, ranging from 1 to 1.016. Quantile-quantile plots for $-\log_{10}$(P-values) were generated for all comparisons (Figured 3.11). An excess of observations, or points above the line of symmetry, in the tail end of the spectrum indicates that more SNPs were significant than expected simply by chance given the number of tests performed. This trend is not evident for either donor or recipient TX vs. AR tests indicating that our study may be underpowered for this trait.

Manhattan plots in Figures 3.12 and 3.13 display unadjusted chi-square $-\log_{10}$ (P-values) for SNPs in order along each chromosome for donor and recipient tests, respectively. Lists of the most significant associations for each comparison can be found in Tables 3.3 to 3.8. The most significant findings were observed for the TX vs. CAN comparisons for both donors and recipients. The least significant findings were revealed for TX vs. AR comparisons, with TX vs. AR + CAN comparisons in the middle. Generally, the top TX vs. AR comparisons were for common SNPs (MAF > 5%), whereas top TX vs. CAN results were for rare variants (MAF < 5%). No AR comparisons remained significant after multiple comparisons corrections were implemented. Logistic regression p-values were quite similar to unadjusted chi-square p-values, with the exception of when the MAF for a case or control group was zero. In these cases, the logistic regression p-value was much higher than the chi-square p-value.

Using the International HapMap's calculation as a threshold (5×10^{-08}) [9], genome-wide significance was achieved or was very close to being achieved in all but the donor and recipient TX vs. AR comparisons. The top 2 hits for donor and recipient CAN and AR + CAN tests were also found to be highly significant when performing association by array type testing. After taking this into consideration, only CAN tests for donors and recipients achieved (recipients, lowest P 1.43×10^{-09}) or nearly achieved (donors, lowest P 6.51×10^{-08}) genome-wide significance. Odds ratios for risk alleles ranged from 1.706 to 76.22 and protective ORs ranged from 0.045 to 0.589 in the most highly significant SNPs ($P < 10^{-04}$) for all 6 comparisons. The TX vs. CAN comparisons for both donors and recipients displayed the highest OR's in the study.

The most significant finding in the study, identified in the recipient TX vs. CAN comparison, was SNP_A-2207560 (rs17578850) on chromosome 4 between TBC1 domain family, member 1 (TBC1D1) and phosphoglucomutase 2 (PGM2) (unadjusted p = 1.43×10^{-09}; OR (95% CI) – 17.02 (5.016 - 57.77)). The SNP did not meet genome-wide significance after adjustment for the false discovery rate (FDR p = 0.0024) or in logistic regression analysis with MDS covariates Logistic p = 4.66×10^{-05}). TBC1D1 is known to regulate cell growth and differentiation [10] while PGM2 functions as both a phosphoglucomutase and a phosphopentomutase and might play a role in congenital immunideficienies [11].

General themes for the function of genes represented in the list of top results for all comparisons are a role in kidney function (KCNH8, KCNMA1, CACNA2D1, SLC5A11), immune function (CD5L, CD83 IL1B & IL1A, ALCAM, MAPK13, MBIP, IL13RA1), structure and movement (MYO3B, MAPRE1) and cancer (BCAR3, MYCN, MCC, TUSC1, RSU1, CRK, RIT2). Genes listed were the closest gene to an associated SNP, and does not necessarily mean that the variant is located within the coding region. However, it is possible that the association is due to variants affecting the encoded transcript due to LD patterns in the region.

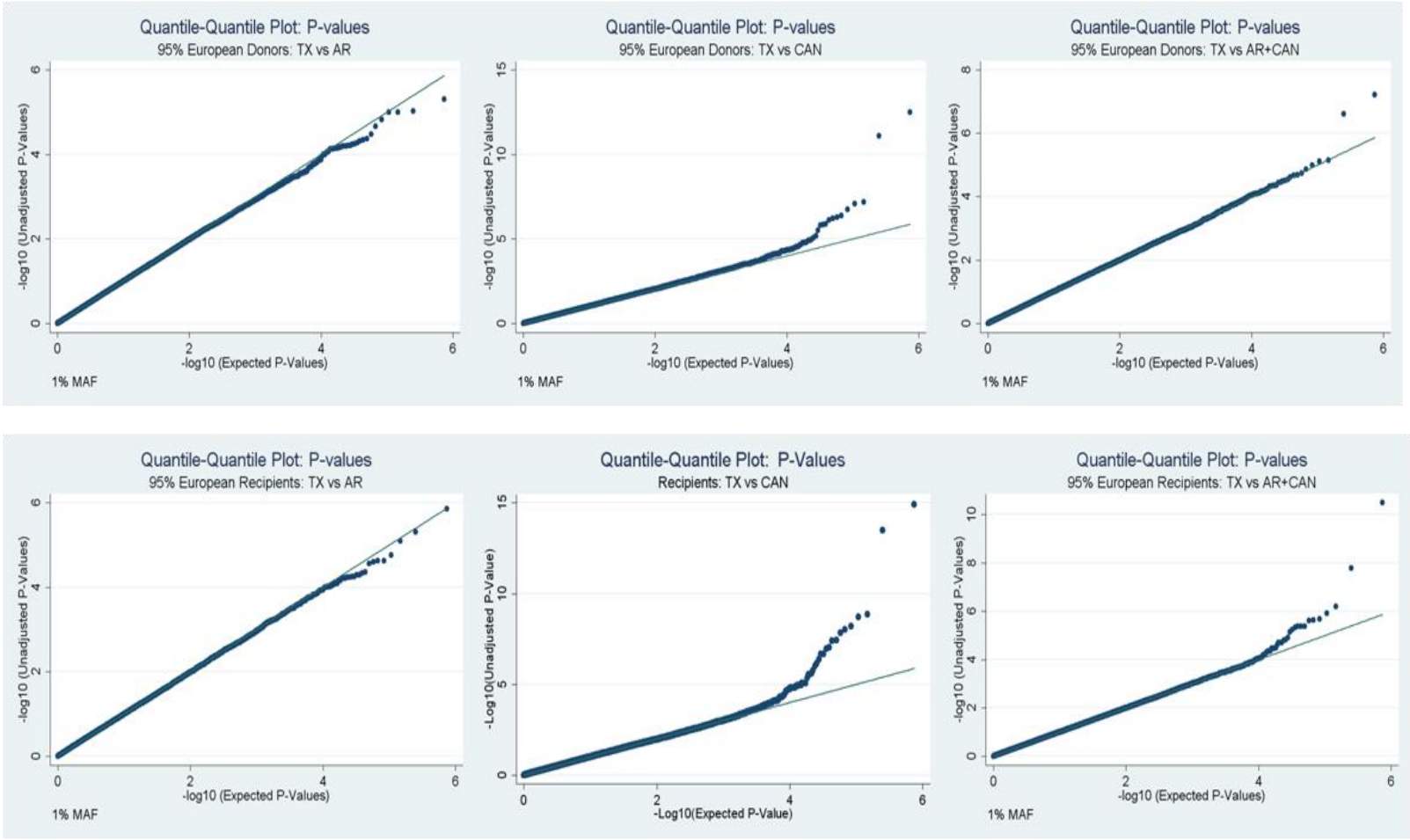


Figure 3.11: Q-Q Plots of P-Values for All 6 Genome-Wide Association Comparisons. Each plot represents observed vs. expected –log10 (P-Value) for all comparisons made in this study where each point represents 1 SNP. Top row displays donor tests while bottom row displays recipients. From left to right, the plots show the comparisons TX vs. AR, TX vs. CAN and TX vs. AR+CAN. Q-Q – quantile-quantile; TX – good outcome; CAN – chronic allograft nephropathy; AR – acute rejection.

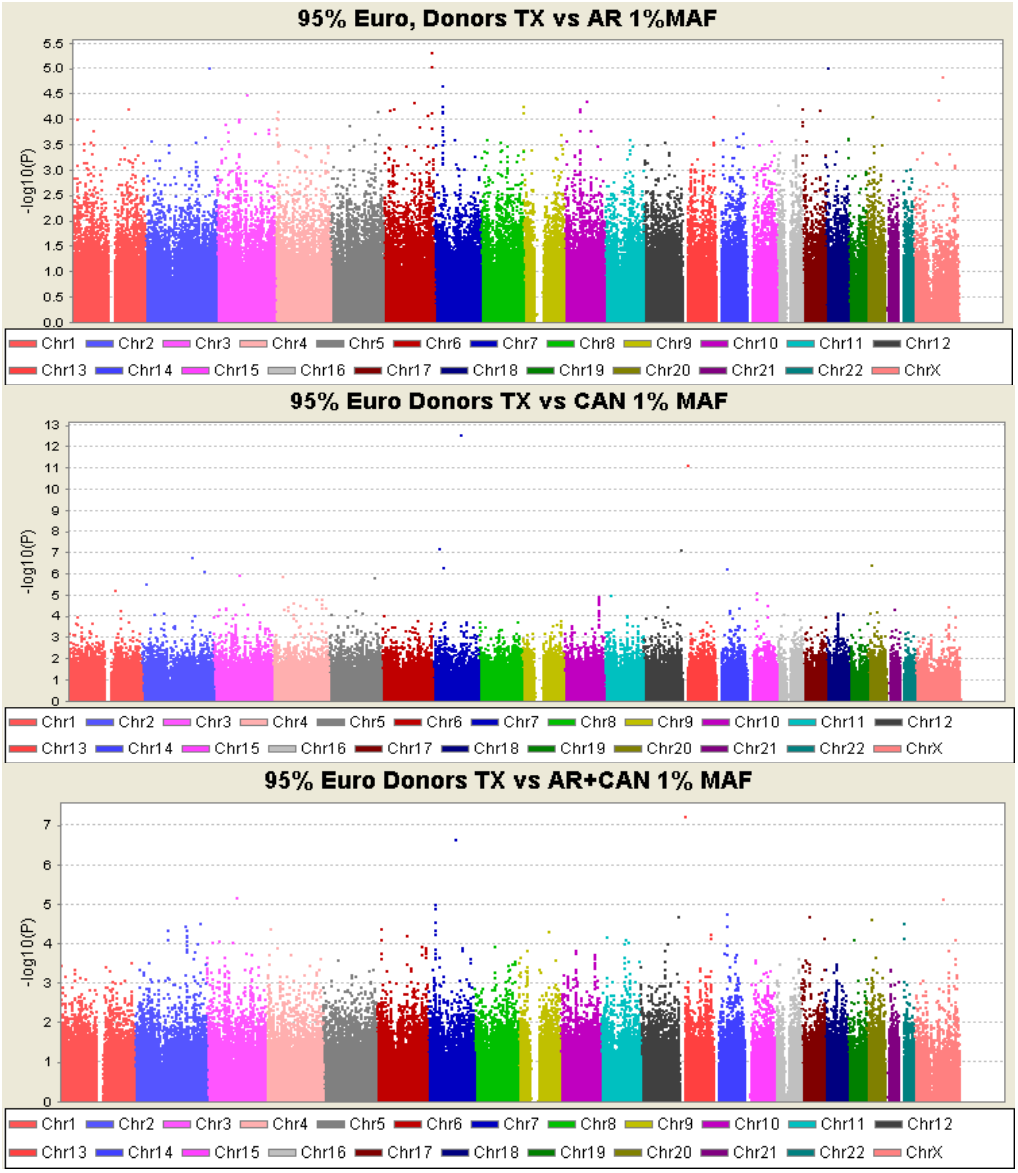


Figure 3.12: Manhattan Plots for 3 Genome-wide Comparisons in Donors. Each color represents a different chromosome and each point represents 1 SNP. The Y axis is –log10 (P-value), such that higher peaks indicate more significant results. Each image represents tests for markers with a study-wide minor allele frequency (MAF) of >1%.

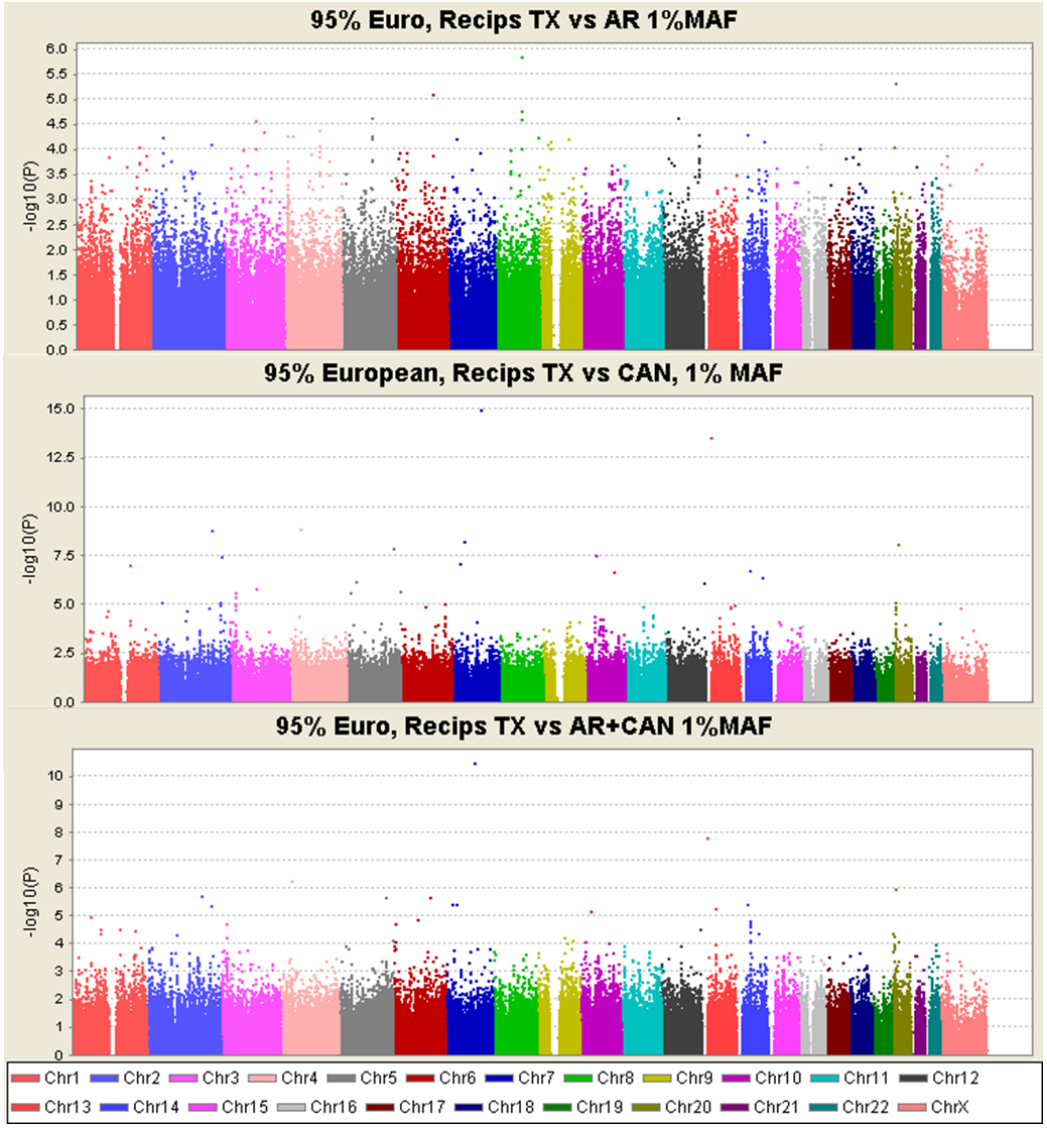


Figure 3.13: Manhattan Plots for 3 Genome-wide Comparisons in Recipients. Each color represents a different chromosome and each point represents 1 SNP. The Y axis is –log10 (P-value), such that higher peaks indicate more significant results. Each image represents tests for markers with a study-wide minor allele frequency (MAF) of >1%.

CHR	SNP	BP	A1	F_A	F_U	A2	P	OR	L95	U95	SIDAK	Logistic P
6	SNP_A-2057496	165510054	B	0.3989	0.2231	A	4.85E-06	2.311	1.606	3.326	0.8365	3.61E-06
6	SNP_A-2147974	165512851	B	0.3889	0.2203	A	9.55E-06	2.252	1.565	3.242	0.9717	6.48E-06
18	SNP_A-4280782	10458453	A	0.1685	0.3442	B	9.99E-06	0.3862	0.2507	0.5949	0.976	3.92E-05
2	SNP_A-1966977	220479819	A	0.1517	0.3238	B	1.01E-05	0.3735	0.2385	0.585	0.9769	2.32E-05
23	SNP_A-2131760	99565504	B	0.6444	0.4289	A	1.53E-05	2.413	1.61	3.618	0.9967	5.10E-05
7	SNP_A-1819917	27621458	B	0.3722	0.2126	A	2.20E-05	2.195	1.52	3.171	0.9997	3.79E-05
3	SNP_A-1849438	104112736	A	0.07778	0.01533	B	3.37E-05	5.419	2.234	13.14	1	3.95E-04
23	SNP_A-2027280	85975189	B	0.1418	0.0403	A	4.24E-05	3.934	1.96	7.899	1	9.32E-05
10	SNP_A-2149130	78416914	A	0.1207	0.03696	B	4.52E-05	3.576	1.873	6.826	1	1.97E-03
6	SNP_A-1841538	103082619	B	0.1222	0.03861	A	4.83E-05	3.467	1.844	6.519	1	1.18E-04

Table 3.3: Top Association Results for Donors TX vs. AR. Genomic inflation factor is 1. All FDR P-values = 1. Logistic regression was conducted with first 2 MDS dimensions as covariates. CHR – chromosome; SNP – single nucleotide polymorphism; BP – base pair; A1 – allele 1; F_A – frequency in cases; F_U – frequency in controls; A2 – allele 2; P – unadjusted P-value for chi-square test; OR – odds ratio; L95 – lower bound of 95% confidence interval; U95 – upper bound of 95% confidence interval; SIDAK – Sidak step down p-value corrected for multiple comparisons; FDR – false discovery rate adjusted P-value.

CHR	SNP	BP	A1	F_A	F_U	A2	P	OR	L95	U95	SIDAK	FDR	Logistic P	AA
7	SNP_A-2198171	96289860	B	0.1087	0.0019	A	2.82E-13	63.05	8.397	473.4	1.05E-07	1.41E-06	5.61E-03	Y
13	SNP_A-2111564	19871368	A	0.0969	0.0019	B	7.98E-12	55.07	7.319	414.3	2.98E-06	2.00E-05	8.13E-03	Y
7	SNP_A-2047727	25742745	A	0.0631	0.0019	B	6.51E-08	34.96	4.543	269	0.024	0.1028	7.49E-04	N
12	SNP_A-4273558	127628558	B	0.0545	0.0000	A	8.21E-08	NA	NA	NA	0.0302	0.1028	9.98E-01	N
2	SNP_A-2124574	171057232	B	0.0784	0.0077	A	1.75E-07	10.98	3.624	33.26	0.06311	0.1748	8.51E-04	N
20	SNP_A-1865116	15703244	A	0.0490	0.0000	B	4.09E-07	NA	NA	NA	0.1417	0.3414	9.98E-01	N
7	SNP_A-2091720	40106997	A	0.0481	0.0000	B	5.29E-07	NA	NA	NA	0.1792	0.3782	9.98E-01	N
14	SNP_A-2109300	35604109	B	0.0693	0.0059	A	6.17E-07	12.54	3.562	44.11	0.2056	0.3858	7.37E-05	N
2	SNP_A-2040275	214317752	B	0.0545	0.0019	A	7.57E-07	29.77	3.818	232.2	0.2462	0.421	1.20E-03	N
3	SNP_A-4242945	88660873	A	0.0446	0.0000	B	1.28E-06	NA	NA	NA	0.3787	0.6381	9.98E-01	N

Table 3.4: Top Association Results for Donors TX vs. CAN. Genomic inflation factor is 1.00912. Logistic regression was conducted with first 2 MDS dimensions as covariates. Abbreviations are the same as Table 3.3. AA – Assoc by Array Type.

CHR	SNP	BP	A1	F_A	F_U	A2	P	OR	L95	U95	SIDAK	FDR	Logistic P	AA
13	SNP_A-2111564	19871368	A	0.0618	0.0019	B	6.00E-08	33.81	4.545	251.5	0.0221	0.3002	1.74E-02	Y
7	SNP_A-2198171	96289860	B	0.0565	0.0019	A	2.45E-07	30.96	4.135	231.8	0.0874	0.6130	1.92E-02	Y
3	SNP_A-1849438	104112736	A	0.0747	0.0153	B	7.18E-06	5.19	2.346	11.48	0.9315	1.00	6.36E-05	N
23	SNP_A-2131760	99565504	B	0.6021	0.4289	A	7.78E-06	2.014	1.479	2.743	0.9452	1.00	1.25E-05	N
7	SNP_A-4266247	27527347	A	0.5052	0.3591	B	1.05E-05	1.822	1.394	2.382	0.9804	1.00	1.99E-05	N
7	SNP_A-2047727	25742745	A	0.0417	0.0019	B	1.37E-05	22.57	2.979	170.9	0.9940	1.00	2.86E-03	N
14	SNP_A-2221474	47472645	A	0.3229	0.1981	B	1.88E-05	1.931	1.425	2.616	0.9991	1.00	2.15E-05	N
12	SNP_A-4273558	127628558	B	0.0344	0.0000	A	2.05E-05	NA	NA	NA	0.9995	1.00	9.97E-01	N
17	SNP_A-4197954	25955997	B	0.0652	0.1600	A	2.14E-05	0.3663	0.2271	0.5907	0.9997	1.00	1.42E-04	N
20	SNP_A-1865116	15703244	A	0.0340	0.0000	B	2.43E-05	NA	NA	NA	0.9999	1.00	9.97E-01	N

Table 3.5: Top Association Results for Donors TX vs. AR + CAN. Genomic inflation factor is 1.00484. Logistic regression was conducted with first 2 MDS dimensions as covariates. Abbreviations are the same as Table 3.3. AA – Assoc by Array Type.

CHR	SNP	BP	A1	F_A	F_U	A2	P	OR	L95	U95	SIDAK	Logistic P
8	SNP_A-1894798	85085311	A	0.4507	0.2400	B	1.42E-06	2.598	1.75	3.857	0.4123	5.05E-06
20	SNP_A-2261086	12307879	B	0.0441	0.2140	A	4.88E-06	0.1696	0.07252	0.3964	0.8383	3.13E-05
6	SNP_A-4265910	123915176	A	0.3662	0.1858	B	8.12E-06	2.531	1.67	3.836	0.9517	3.12E-05
8	SNP_A-1994083	85087422	A	0.3239	0.1593	B	1.79E-05	2.529	1.641	3.897	0.9987	4.42E-05
5	SNP_A-1840780	100211413	A	0.0786	0.0114	B	2.35E-05	7.419	2.531	21.74	0.9998	0.000713
12	SNP_A-2027976	50060401	A	0.6643	0.4600	B	2.42E-05	2.323	1.562	3.454	0.9999	4.14E-05
8	SNP_A-1917509	85089160	B	0.3380	0.1726	A	2.59E-05	2.448	1.601	3.744	0.9999	5.26E-05
3	SNP_A-1973791	106948094	B	0.2113	0.0830	A	2.78E-05	2.961	1.752	5.005	1	0.000159
4	SNP_A-2298432	116979120	B	0.2887	0.1394	A	4.41E-05	2.507	1.598	3.931	1	0.00011
3	SNP_A-1974342	132430125	A	0.4500	0.2677	B	4.65E-05	2.238	1.511	3.315	1	6.4E-05

Table 3.6: Top Association Results for Recipients TX vs. AR. Genomic inflation factor is 1.0164. All FDR P-values = 1. Logistic regression was conducted with first 2 MDS dimensions as covariates. Abbreviations are the same as Table 3.3.

CHR	SNP	BP	A1	F_A	F_U	A2	P	OR	L95	U95	SIDAK	FDR	Logistic P	AA
7	SNP_A-2198171	96289860	B	0.1485	0.0023	A	1.16E-15	76.22	10.31	563.3	4.15E-10	5.81E-09	4.29E-03	Y
13	SNP_A-2111564	19871368	A	0.1421	0.0045	B	3.47E-14	36.44	8.569	155	1.29E-08	8.68E-08	8.87E-04	Y
4	SNP_A-2207560	37565122	A	0.1029	0.0067	B	1.43E-09	17.02	5.016	57.77	0.000535	0.002393	4.66E-05	N
2	SNP_A-4249372	1.81E+08	B	0.0784	0.0000	A	1.95E-09	NA	NA	NA	0.000728	0.002442	0.9971	N
7	SNP_A-2091720	40106997	A	0.0735	0.0000	B	6.38E-09	NA	NA	NA	0.002377	0.006381	0.9972	N
20	SNP_A-1865116	15703244	A	0.0728	0.0000	B	9.47E-09	NA	NA	NA	0.003529	0.0079	0.9972	N
5	SNP_A-4265328	1.54E+08	A	0.0777	0.0022	B	1.53E-08	37.98	5.001	288.4	0.0057	0.01095	0.00067	N
10	SNP_A-2195203	35954523	A	0.1020	0.0111	B	3.62E-08	10.11	3.738	27.37	0.01341	0.02263	0.000118	N
2	SNP_A-2040275	2.14E+08	B	0.0743	0.0022	A	4.14E-08	35.86	4.702	273.4	0.01534	0.02303	0.00048	N
7	SNP_A-2047727	25742745	A	0.0625	0.0000	B	9.06E-08	NA	NA	NA	0.03324	0.04532	0.9974	N

Table 3.7: Top Association Results for Recipients TX vs. CAN. Genomic inflation factor is 1.00067. Logistic regression was conducted with first 2 MDS dimensions as covariates. Abbreviations are the same as Table 3.3. AA – Assoc by Array Type.

CHR	SNP	BP	A1	F_A	F_U	A2	P	OR	L95	U95	SIDAK	FDR	Logistic P	AA
7	SNP_A-2198171	96289860	B	0.1029	0.0023	A	3.38E-11	50.15	6.833	368	1.26E-05	0.000169	0.00934	Y
13	SNP_A-2111564	19871368	A	0.0833	0.0045	B	1.64E-08	20	4.72	84.74	0.006113	0.04111	0.004561	Y
4	SNP_A-2207560	37565122	A	0.0727	0.0067	B	6.24E-07	11.62	3.48	38.83	0.2077	1	0.000144	N
20	SNP_A-1865116	15703244	A	0.0520	0.0000	B	1.23E-06	NA	NA	NA	0.3679	1	0.997	N
2	SNP_A-4249372	1.81E+08	B	0.0491	0.0000	A	2.11E-06	NA	NA	NA	0.5449	1	0.9971	N
5	SNP_A-4265328	1.54E+08	A	0.0549	0.0022	B	2.36E-06	26.2	3.49	196.7	0.5858	1	0.002057	N
6	SNP_A-2300587	1.24E+08	B	0.1314	0.2679	A	2.46E-06	0.4136	0.2845	0.6012	0.6005	1	4.94E-06	N
14	SNP_A-2109300	35604109	B	0.0482	0.0000	A	4.21E-06	NA	NA	NA	0.7922	1	0.9972	N
7	SNP_A-2047727	25742745	A	0.0462	0.0000	B	4.26E-06	NA	NA	NA	0.7963	1	0.9972	N
7	SNP_A-2091720	40106997	A	0.0462	0.0000	B	4.26E-06	NA	NA	NA	0.7963	1	0.9972	N

Table 3.8: Top Association Results for Recipients TX vs. AR + CAN. Genomic inflation factor is 1. Logistic regression was conducted with first 2 MDS dimensions as covariates. Abbreviations are the same as Table 3.3. AA – Assoc by Array Type.

3.3.2. IBS Differences between Pairs by Outcome

Average pi-hat between any two European-derived samples is 0.01. Mean pi-hat between all donor-recipient pairs of known donor type and outcome is 0.215 (N=258 pairs). Living related donor pairs (N=114) had a mean pi-hat of 0.459 while living unrelated donor pairs (N=84) had pi-hat of 0.035. Cadaverous donor pairs (N=60) had a mean pi-hat of 0.006 (Figure 3.12).

Pi-hat by outcome across all donor types is 0.208, 0.186 and 0.229 for AR, CAN and TX, respectively, revealing a non-significant trend towards increased pi-hat for those pairs with a good outcome. In comparing IBS between outcomes within LRD, LURD and CAD groups, only a small trend is observed for LRD, where TX pairs have the highest mean pi-hat and AR pairs have the lowest mean pi-hat (Figure 3.13). All ANOVA tests for differences in IBS for 3 outcomes (TX, AR, CAN) within the 3 donor-type classes were non-significant (Table 3.9). Two LURD pairs had pi-hat > 0.4 and after removal of these 2 samples the same trend is observed as for LRD where TX pairs have a slightly higher mean pi-hat. The new values become 0.029 TX, 0.017 CAN and 0.015 AR, but ANOVA remains non-significant.

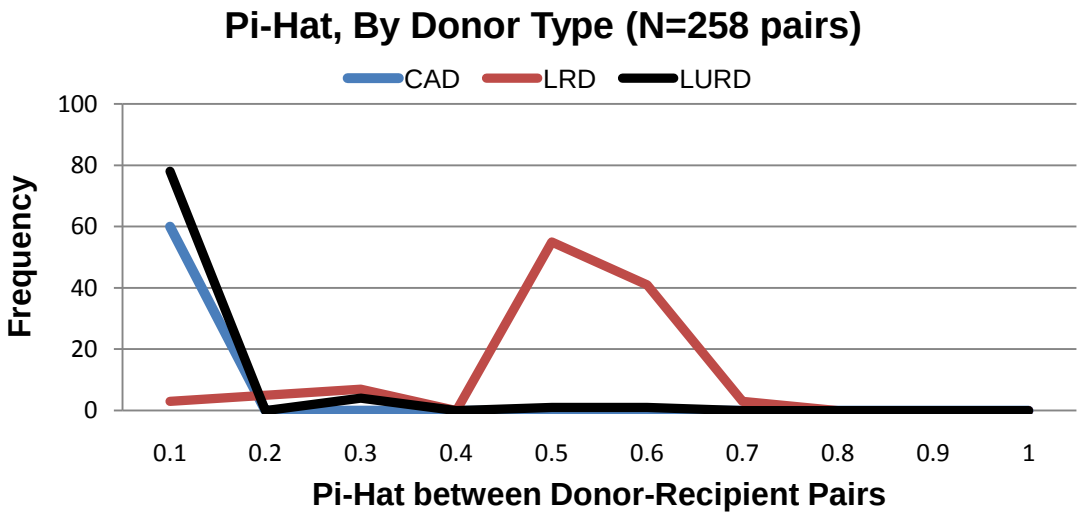
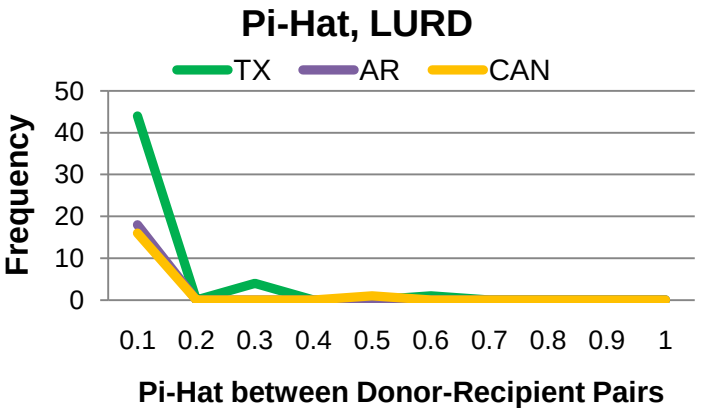
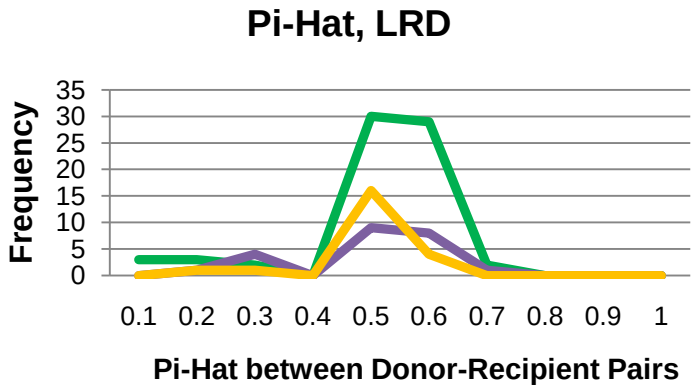


Figure 3.14: Pi-Hat between Donor-Recipient Pairs by Donor Type. CAD – cadaverous donor (N=60). LRD – living related donor (N=114). LURD – living unrelated donor (N=84). Mean pi-hat for all donor-recipient pairs is 0.215.





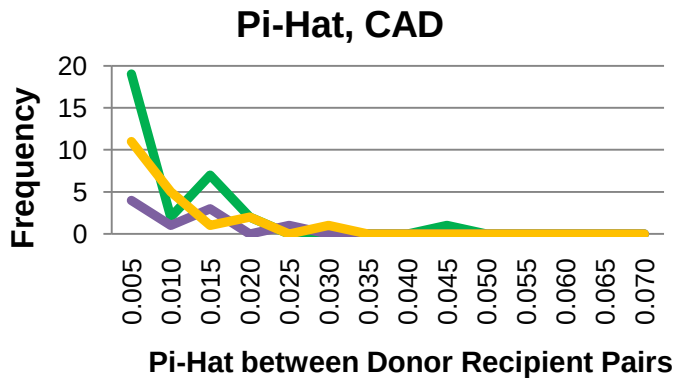


Figure 3.15: Pi-Hat between Donor-Recipient Pairs by Outcome. LURD – living unrelated donor; LRD – living related donor (N=114); (N=84); CAD – cadaverous donor (N=60).

	CAD			LRD			LURD		
Outcome	**Mean**	**SD**	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**	**N**
TX	0.006	0.009	31	0.462	0.131	70	0.039	0.095	49
CAN	0.006	0.008	20	0.460	0.100	22	0.044	0.113	17
AR	0.007	0.008	9	0.447	0.129	22	0.015	0.020	18

Table 3.9: Summary and Association of Pi-Hat Values for 3 Donor Types by Outcomes. All ANOVA tests non-significant. CAD – cadaverous donor (N=60); LRD – living related donor (N=114); LURD – living unrelated donor (N=84); SD – standard deviation; TX – good outcome; AR – acute rejection; CAN – chronic allograft nephropathy.

3.4. Discussion

We have conducted the first GWAS of rejection phenotypes in kidney transplantation donors and recipients. We have taken advantage of the paired data through global IBS analysis and investigated differences in the mean proportions shared between outcome phenotypes. Our GWAS results do not identify the known chromosome 6 HLA locus' association with rejection. However, as perfectly-matched HLA donor-recipient pairs were not included in this study and anti-rejection drugs help control against the severe rejection contributed to by this locus, we did not expect to identify this region. However, we do identify loci involved with immune function (CD5L, CD83 IL1B & IL1A, ALCAM, MAPK13, MBIP, IL13RA1), which was expected. IL1B haplotypes have been previously associated with multiple acute rejection episodes in heart transplant recipients [12]. An IL1A promoter polymorphism has been tested for association with acute rejection in renal transplant donors (63 cases vs. 63 controls) and recipients (74 cases vs. 70 controls), but no significant difference was identified in either group (p=.685 & p = 0.634 in donors and recipients, respectively) [13].

A weakness of our study is the low power we have to identify SNPs associated to AR, visualized in the Q-Q plots in Figure 3.9. This lack of power would be aided by the addition of more samples to the analysis, an ongoing goal of the project. Further validation of the GWAS findings, especially the CAN results, through testing in other ethnic groups for which we have data and also in an

independent transplant collection will be necessary to assure we have avoided false positive findings. Future analyses on this same dataset could also include evaluating copy number variation differences between rejection phenotypes.

The identity by state findings demonstrate that there are not global differences in the genome contributing to outcomes in kidney transplantation. However, as we already know the importance of the HLA locus and may identify more regions through GWAS analysis (after adding more samples and confirming our results in additional collections), we may wish to study IBS patterns in more detail in specific areas of the genome. This would allow us to take advantage of the paired nature of our dataset, which is wholly ignored in the GWAS analyses.

3.5. References

1. System, U.S.R.D., *USRDS 2009 Annual Data Report: Atlas of Chronic Kidney Disease and End-Stage Renal Disease in the United States*, N.I.o.D.a.D.a.K.D. National Institutes of Health, Editor. 2009: Bethesda, MD.

2. Solez, K., et al., *Banff 07 classification of renal allograft pathology: updates and future directions.* Am J Transplant, 2008. **8**(4): p. 753-60.

3. Pritchard, J.K., M. Stephens, and P. Donnelly, *Inference of population structure using multilocus genotype data.* Genetics, 2000. **155**(2): p. 945-59.

4. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.

5. Clayton, D.G., et al., *Population structure, differential bias and genomic control in a large-scale, case-control association study.* Nat Genet, 2005. **37**(11): p. 1243-6.

6. Rice, T.K., N.J. Schork, and D.C. Rao, *Methods for handling multiple testing.* Adv Genet, 2008. **60**: p. 293-308.

7. Devlin, B. and K. Roeder, *Genomic control for association studies.* Biometrics, 1999. **55**(4): p. 997-1004.

8. Benjamini, Y.a.Y.H., *Controlling the false discovery rate—a practical and powerful approach to multiple testing.* J. R. Stat. Soc. Ser, 1995. **B**(57): p. 289-300.

9. The International HapMap, C., *A haplotype map of the human genome.* Nature, 2005. **437**(7063): p. 1299-1320.

10. White, R.A., et al., *The gene encoding TBC1D1 with homology to the tre-2/USP6 oncogene, BUB2, and cdc16 maps to mouse chromosome 5 and human chromosome 4.* Cytogenet Cell Genet, 2000. **89**(3-4): p. 272-5.

11. Maliekal, P., et al., *Molecular identification of mammalian phosphopentomutase and glucose-1,6-bisphosphate synthase, two members of the alpha-D-phosphohexomutase family.* J Biol Chem, 2007. **282**(44): p. 31844-51.

12. Vamvakopoulos, J.E., et al., *Interleukin 1 and chronic rejection: possible genetic links in human heart allografts.* Am J Transplant, 2002. **2**(1): p. 76-83.

13. Lee, H., et al., *Influence of recipient and donor IL-1alpha, IL-4, and TNFalpha genotypes on the incidence of acute renal allograft rejection.* J Clin Pathol, 2004. **57**(1): p. 101-3.

CHAPTER 4

DESIGN OF A RESEQUENCING PANEL FOR INVESTIGATION OF RARE VARIANTS IN GENE TARGETS

4.1. Introduction

Resequencing genes by Sanger sequencing has long been a method of investigation for rare genetic conditions, but was typically low throughput and time intensive [1]. Since 2004, a wave of new technology has made sequencing more high throughput and cost effective and has also lent itself to variant detection in common diseases which are typically heterogeneous and multigenic and for which individual polymorphisms typically contribute a small fraction of the total genetic risk [2, 3]. The large number of sequences that can be attained, on the order of mega and gigabases, in a single experiment mean that large numbers of cases and controls can be pooled, making association to a phenotype of interest feasible.

For this project, over 100 candidate genes were selected by our collaborators from gene expression and proteomic studies for deep resequencing in our kidney transplant donor-recipient collection. This effort was intended to complement our GWAS conducted in Chapter 3. Individuals were to be resequenced on a custom-designed and manufactured tiling array. Large-format arrays can sequence up to 300,000 bases in both directions with 99.95% accuracy, according to the manufacturer. The technology in the sequencing field has rapidly advanced since the beginning of this project and it is now feasible to do high throughput sequencing on platforms offered by other companies. Here, I will present the original chip design and testing along with a modified approach to sequencing through a next-generation technology.

4.2. Materials and Methods

4.2.1. Sequence Selection and LR-PCR for Sequencing on Microarrays

A custom resequencing array was designed that contains sequencing probes for 118 candidate genes with a mean of 2,298 base pairs to be sequenced per gene (Table 4.1 & Figure 4.1). Sequence to be tiled onto the array was formatted by downloading gene sequence from the

Ensembl [4] database, removing repetitive sequences identified with RepeatMasker [5] and removing homologous sequences identified with Miropeats [6]. Repetitive and homologous sequences are not worth tiling on a microarray as they cannot be distinguished from one another and accurate sequence information cannot be attained. For each gene, coding sequence, exon-intron boundaries and promoter regions were included. A schematic of the process displays how much sequence was lost at each design stage (Figure 4.2) and displays the sequence lost per gene (Figure 4.3). After accounting for PCR design failure, 232,993 bases from 112 genes will be the maximum sequencing output of our custom designed sequencing microarray.

DNA samples were amplified using long-range PCR primers designed by Perlegen, Inc. or with Primer3 [7, 8]. Each primer pair was tested on 3 DNA samples and had to amplify at least 2 with a single band to be included in the passing primer panel. A full list of the 369 primer pairs can be found in the Appendix. Long range PCR (LR-PCR) conditions were a modified version of the Affymetrix protocol using a 12µL reaction. At least 9.5µg of DNA was needed for each individual to be resequenced and the assay was compatible with whole-genome amplified DNA. Two µM each primer, forward and reverse, were mixed with 2.5mM dNTPs, Takara LA Taq, LA PCR buffer II and water. Thermal cycling consisted of denaturing DNA at 94˚C for 2 minutes followed by 35 cycles of 94˚C for 15 seconds and 64˚C for 12 minutes, with final elongation of 64˚C for 17 minutes. All reactions were screened by gel electrophoresis and quantified before being pooled in equimolar quantities with a liquid handling robot.

We used Quant-iT Pico Green dsDNA Assay Kit for quantifying the samples. A standard curve was generated by serial dilution of a lambda DNA standard from 5.120ng/µl to 0.03ng/µl. DNA samples for quantifiication were diluted 1:250 with 1x TE buffer, vortexed and then 5µL of the dilution was added to a plate containing 10µL 2x Pico Green Reagent. The whole protocol was carried out with a liquid handling robot. Readings were taken on an Envision spectrophotometer plate reader and concentrations were determined by fitting to the standard curve. Pooled PCR reactions were cleaned with a Clontech filter plate per manufacturer's instructions and subjected

to the Affymetrix post-PCR protocol. After scanning, samples were subjected to base analysis and SNP calling with the Affymetrix resequencing array software, GeneChip Sequence Analysis Software (GSEQ).

Genes	Fragments	Bases	Primer Pairs
118	1,318	271,183	387

Table 4.1: Resequencing Array Summary

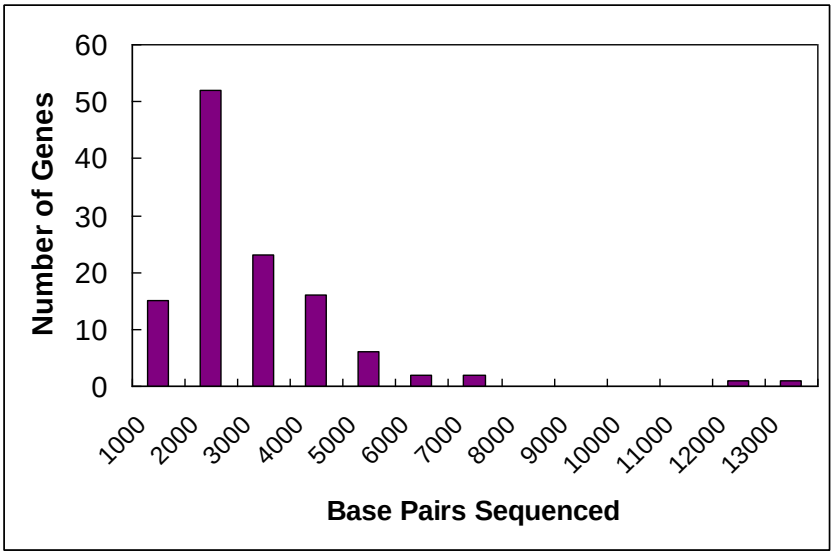


Figure 4.1: Histogram of Base Pairs Sequenced Per Gene. Mean base pairs sequenced per gene is 2,298.

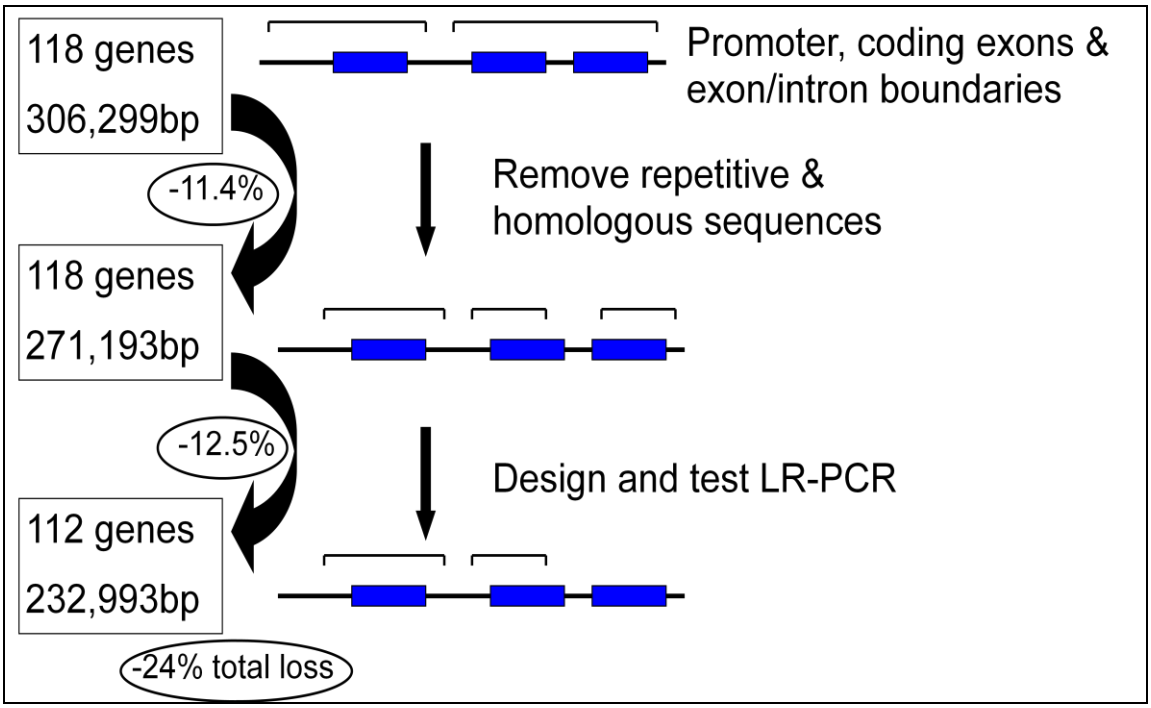


Figure 4.2: Schematic of Sequencing Microarray Sequence Selection Process. The process began with 118 genes and ended with 112 genes tiled onto the array and successfully amplified with LR-PCR primers. Percentage represents fraction of sequence lost at each design step. LR-PCR – Long range – polymerase chain reaction.

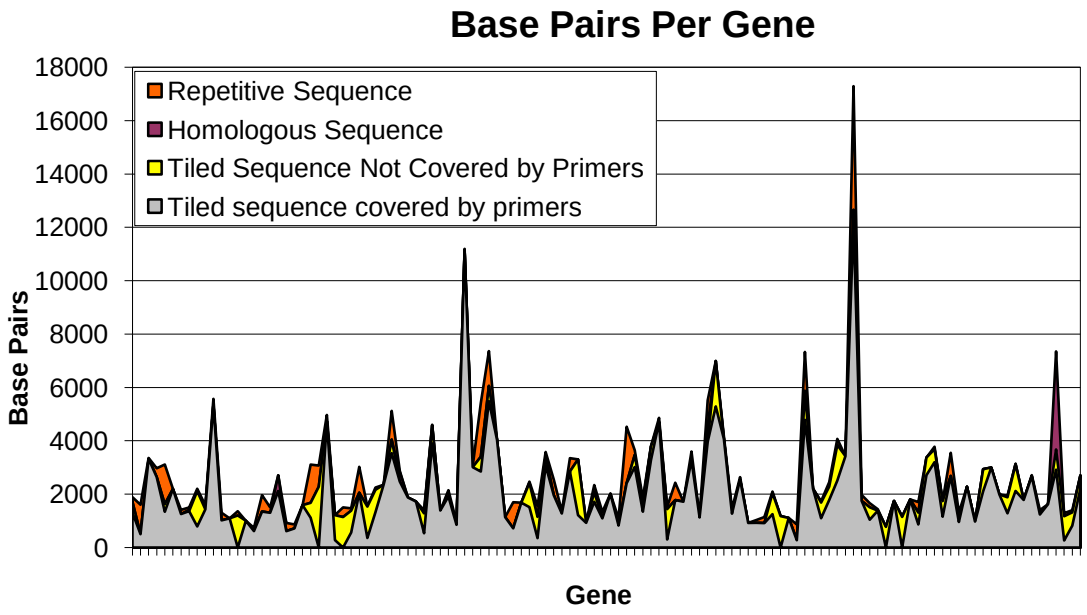


Figure 4.3: Sequence Selection per Gene in Base Pairs. Grey portion shows sequence remaining, while yellow, orange and purple sections display sequence removed for reason listed in legend for 118 genes.

4.2.2. Alternative Strategy Testing – Next Generation Sequencing

Next generation, high throughput sequencing technology has rapidly developed over the past five years. We decided to use our panel of PCR primers to prepare DNA libraries to be analyzed on a new technology, a 4-base sequencing by synthesis assay. The advantage this technology has over the Affymetrix technology is that all amplified bases can be sequenced, as opposed to just those bases that are able to be probed through microarray hybridization. This increases the potential amount of sequenced bases from 233Kb to 3.1Mb and means that in addition to coding, intron/exon boundaries and promoters, intronic regions and more 5' and 3' regions of the genes can be covered, depending on the placement of PCR primers.

We prepared the samples for Illumina Solexa sequencing by following the PCR method and pooling strategy in section 4.2.1. Library preparation was performed according to the manufacturer's protocol. Briefly, this involved shredding the amplified DNA samples through nebulization to less than 800bp fragments, polishing ends to be blunt, adding a 3'-dA overhang, ligating adaptors, and amplifying adapted DNA fragments with PCR. We quantified the library with Power SYBR Green quantitative real-time PCR using a well-performing previously sequenced library as a standard and Solexa PCR primers complementary to the adaptors. The DNA libraries were first quantified on a spectrophotometer and diluted to 10nM. Ten microliter reactions contained 1µl of DNA and 5uM each primer. Triplicate reactions were cycled and read on an Applied Biosystems 7900HT machine using a standard 40 cycle Absolute Quantification protocol.

Solid phase sequencing was achieved on the Genome Analyzer by flowing 5pM DNA into channels previously populated with a dense layer of primers complementary to the adaptors attached to each end of the DNA fragments in the library preparation process. Attached DNA strands were then copied through repeated bridge amplification and the resulting clusters of identical sequence were subjected to 4 base reversible terminator sequencing chemistry. Bases were read after each of 36 cycles by laser excitation and image capture. Intensities were

converted to text outputs with the Genome Analyzer Pipeline software. Firecrest captured intensity values, Bustard called bases and Gerald, or more specifically Eland, aligned the reads to the human genome. Polymorphisms were called with Casava.

We wanted to test the power of this sequencing technology on pooled DNA sequences with the intention of future resequencing pools of cases and controls. We made four libraries which included 1 CEPH DNA sample, 5 samples and 10 samples as well as 1 whole genome amplified (WGA) sample from a kidney transplantation study to test the robustness of this assay on this type of DNA.

4.3. Results

4.3.1. Hybridization Tests for Custom Designed Resequencing Array

Preliminary testing of the array was conducted. The first test was to pool, in equimolar quantities, 59 PCR products from a test DNA sample for hybridization onto the array. Ninety-three percent of bases covered by the amplified regions were called with an average fragment call rate of 96% and 8 SNPs in 6 genes were identified, all of which were in dbSNP. These data are preliminary and only to be interpreted as proof of our ability to successfully hybridize onto our custom array as Affymetrix recommends that their base-calling algorithm, GSEQ, be run with a minimum of 15 samples at once.

The first full run saw the pooling of 353 PCR reactions (96% of all PCR attempted) which included 210,874 bases out of a possible 232,993 (90.5% of all bases). The base call rate achieved was 90.43% of tiled bases. If fragments with less than 60% of bases called were excluded, the remaining bases displayed a 93.21% call rate. Of the fragments pooled, 96% passed the call rate threshold of 60% and included 201,941 bases. SNPs were called, totaling 13,855 in total. Of these, 7,121 were identified in non-pooled fragments. The remaining 6,734 were in pooled fragments. However, 1205 were in failed fragments (due to low call rate) and 5,466 were surrounded by low quality sequence. This left 63 remaining heterozygous polymorphisms which

equates to 1 polymorphism per approximately 3000 bases. The number of polymorphisms is about 3 times fewer than expected and could be due to a weakness in calling just 1 sample at a time.

4.3.2. Preliminary Testing of Next Generation Sequencing Technology

We dedicated 2 lanes to each of the 4 libraries prepared for sequencing, filling up the 8 available lanes on the flow cell. We expected a yield of 125Mb per lane, but only achieved approximately one half to two-thirds of this amount, obtaining between 59.9Mb and 80Mb of raw sequence per lane. We used a control lane of PhiX sequence from an earlier experiment run on the same machine to control for base pair composition, which is important if a library has an unequal amount of A's, C's, G's and T's. For a summary of yield per lane, clusters passing filters (PF), aligning to the genome and error rates, see Table 4.2. Accounting for the percent clusters PF, percent aligned and the number of bases pooled after LR-PCR for each library, the CEPH 1-plex sample had mean coverage of 28.3x per base when utilizing the full 36-base read. The 5-plex had 29x coverage per base, which represents 5.8x coverage per base per individual. The 10-plex library had mean 23.3x coverage per base, or 2.3x coverage per base per individual. The WGA DNA sample had the lowest mean coverage per base of 14.4x even though it had the highest output of raw bases. This was due to the low percent of aligned bases, around 32% for each lane and might have to do with the sample being whole genome amplified before the sequence selection through LR-PCR. We visually observed alignments in Illumina's Genome Studio software and observed overrepresentation of reads at LR-PCR primer sequence loci used in our DNA selection method.

Eighty-five to ninety percent of the clusters passed basic quality metrics calculated in the Pipeline for all lanes. Illumina recommends that greater than 80% of sequence aligns in order to advance to SNP calling with their Casava software. The percent of sequence aligned to the human genome was low for 2 of the libraries; both lanes of the 10-plex pool and the WGA DNA sample from the kidney transplantation study (sample B10018). Additionally, the error rate for all 8

libraries surpassed the threshold suggested by Illumina, which advises < 1.2% +/- 0.3% for a 35-base read. However, after trimming the read length to the first 30 bases from the original full data set of 36 bases, only the WGA B10018 sample failed the percent aligned and error rate filters. See Table 4.3 for a lane summary with the shorter read length. This demonstrates that the error rate increases with the number of cycles. It also highlights that, although we cannot prove it here, WGA DNA might be problematic for next-generation sequencing.

We called polymorphisms for the CEPH 1-plex sample with Casava and 3,250 SNPs were identified, which is approximately 1 SNP for every 1,000 bases sequenced. Mean bases used for each SNP call was 239 with a standard deviation of 1,003 (min = 3; max = 17,514). A Venn diagram of SNP types displays that 57% of polymorphisms were heterozygotes, some were homozygotic changes from the reference genome and very few contained 2 bases different from the reference (Figure 4.4). Based upon PCR failure, we pooled 3.076Mb of DNA for this sample and we used 32-base length reads for SNP calling. This was the longest length that passed QC metrics for Casava SNP calling for the 2 lanes of data. Totaling the 2 lanes of passing filter bases and aligned bases, 77.4Mb of sequence was attained for the sample giving us a mean base coverage of 25.2x.

		36 base Gerald Alignment to hg18 Reference Genome			
L	**Contents**	**Yield (Kb)**	**% PF Clusters**	**% Align (PF)**	**% Error Rate (PF)**
1	CEPH 1-plex	63590	88.71 +/- 0.50	79.68 +/- 1.59	3.17 +/- 0.07
2	CEPH 1-plex	59904	88.37 +/- 0.69	79.51 +/- 1.04	3.16 +/- 0.07
3	CEPH 5-plex	64137	87.50 +/- 0.78	71.99 +/- 3.12	2.88 +/- 0.07
4	CEPH 5-plex	64045	86.96 +/- 2.49	72.17 +/- 4.11	2.96 +/- 0.09
5	CEPH 10-plex	67940	86.36 +/- 0.74	59.68 +/- 2.24	6.82 +/- 0.05
6	CEPH 10-plex	68065	86.84 +/- 1.03	60.20 +/- 1.46	6.85 +/- 0.08
7	B10018	79551	84.57 +/- 1.07	31.81 +/- 0.49	6.95 +/- 0.09
8	B10018	80161	84.38 +/- 1.68	31.98 +/- 0.38	7.01 +/- 0.12

Table 4.2: Lane Yields and Error Rates for 4 Libraries Sequenced by Next-Generation Technology. Data analyzed with Pipeline 1.0, with exception of Gerald, from Pipeline 1.1 (beta) with PhiX for intensity control from an earlier experiment on the same machine. Text in red indicates that metric is below the quality threshold set by Illumina for variant calling. B10018 is a male transplant recipient and from whole genome amplified DNA; L – Lane; Kb – kilobases; PF – pass filter.

		30 base Gerald Alignment to hg18 Reference Genome			
L	**Contents**	**Yield (Kb)**	**% PF Clusters**	**% Align (PF)**	**% Error Rate (PF)**
1	CEPH 1-plex	52991	88.71 +/- 0.50	89.37 +/- 1.80	1.07 +/- 0.02
2	CEPH 1-plex	49920	88.37 +/- 0.69	89.11 +/- 1.19	1.08 +/- 0.03
3	CEPH 5-plex	53447	87.50 +/- 0.78	77.26 +/- 3.37	0.96 +/- 0.02
4	CEPH 5-plex	53370	86.96 +/- 2.49	77.53 +/- 4.46	0.99 +/- 0.05
5	CEPH 10-plex	56617	86.36 +/- 0.74	74.02 +/- 2.87	1.60 +/- 0.02
6	CEPH 10-plex	56721	86.84 +/- 1.03	74.77 +/- 1.81	1.62 +/- 0.04
7	B10018	66293	84.57 +/- 1.07	43.00 +/- 0.68	2.33 +/- 0.04
8	B10018	66800	84.38 +/- 1.68	43.43 +/- 0.45	2.36 +/- 0.07

Table 4.3: Lane Yields and Error Rates for 4 Libraries Sequenced by Next-Generation Technology after Trimming Results to 30-base Reads. Data analyzed with Pipeline 1.0, with exception of Gerald, from Pipeline 1.1 (beta) with PhiX for intensity control from an earlier experiment on the same machine. Text in red indicates that metric is below the quality threshold set by Illumina for variant calling. B10018 is a male transplant recipient and from whole genome amplified DNA; L – Lane; Kb – kilobases; PF – pass filter.

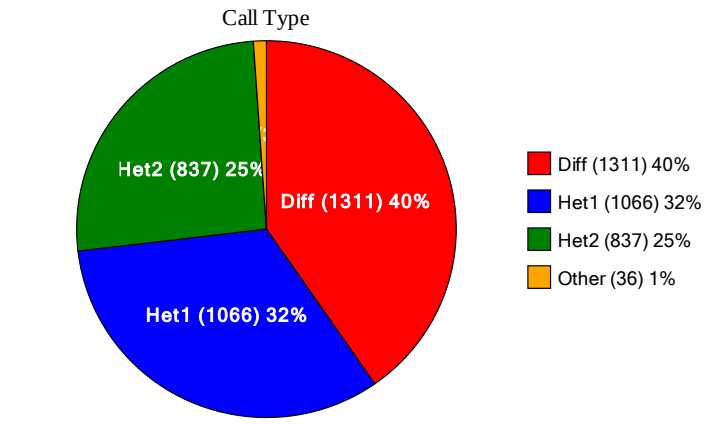


Figure 4.4: Venn Diagram of SNP Call Types for Next-Gen Sequenced Single-Plex Sample. Het1 and Het2 refer to which allele is listed first in genotype call, Het1 first allele matches references and Het2 2nd allele matches the reference. Other type is when 2 bases are called which both differ from the reference genome. Diff – homozygotic call different from reference human genome; Het – heterozygote.

4.4. Discussion

A large panel of LR-PCR primers has been prepared in order to amplify DNA from 112 genes for sequencing. We began the project with the intention to sequence cases and controls on microarrays. We successfully hybridized samples to the microarray, but did not choose to use

them for sequencing of cases and controls for kidney transplant outcomes due to the development of next generation technologies. Microarray sequencing experiments are limited in their value now that such next-generation sequencing technologies are available. These newer sequencing methods have fewer limitations when choosing regions for sequencing. Additionally, in the future, whole genome sequencing without selection of regions of interest may become a more ordinary occurrence, rendering the microarray sequencing method obsolete. We took advantage of our panel of primers to test a next generation sequencing by synthesis assay, which could be used to sequence our kidney transplantation samples.

Future sequencing experiments using this panel of LR-PCR primers could include modifications to the oligos, specifically a 5' block, to avoid the stacking of reads at PCR primer positions evident when sequencing by synthesis on flow cells [9]. Also, we could use a different method for enrichment of DNA regions of interest. One method would be to use long oligos for DNA selection in solution [10]. Another is to select DNA regions of interest through hybridization onto a microarray before sequencing with a next-generation sequencing technology. However, this technology encounters the same weakness as sequencing directly on a microarray [11]. Mainly, DNA must observe microarray hybridization kinetics and limitations to probe design. A third method would be to perform PCR in microdroplets, which would eliminate the labor intensive quantification and normalization step traditional PCR methods, like ours, require [12].

4.5. References

1. Schuster, S.C., *Next-generation sequencing transforms today's biology.* Nat Methods, 2008. **5**(1): p. 16-8.

2. Mardis, E.R., *Next-generation DNA sequencing methods.* Annu Rev Genomics Hum Genet, 2008. **9**: p. 387-402.

3. Nejentsev, S., et al., *Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes.* Science, 2009. **324**(5925): p. 387-9.

4. Hubbard, T.J., et al., *Ensembl 2007.* Nucleic Acids Res, 2007. **35**(Database issue): p. D610-7.

5. Smit, A.F.A., R. Hubley, and P. Green. [cited; Available from: http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker.

6. Parsons, J.D., *Miropeats: graphical DNA sequence comparisons.* Comput Appl Biosci, 1995. **11**(6): p. 615-9.

7. Hinds, D.A., et al., *Whole-genome patterns of common DNA variation in three human populations.* Science, 2005. **307**(5712): p. 1072-9.

8. Rozen, S. and H. Skaletsky, *Primer3 on the WWW for general users and for biologist programmers.* Methods Mol Biol, 2000. **132**: p. 365-86.

9. Harismendy, O. and K. Frazer, *Method for improving sequence coverage uniformity of targeted genomic intervals amplified by LR-PCR using Illumina GA sequencing-by-synthesis technology.* Biotechniques, 2009. **46**(3): p. 229-31.

10. Gnirke, A., et al., *Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing.* Nat Biotechnol, 2009. **27**(2): p. 182-9.

11. Albert, T.J., et al., *Direct selection of human genomic loci by microarray hybridization.* Nat Methods, 2007. **4**(11): p. 903-5.

12. Tewhey, R., et al., *Microdroplet-based PCR enrichment for large-scale targeted sequencing.* Nat Biotechnol, 2009. **27**(11): p. 1025-31.

CHAPTER 5

SEQUENCING OF TNFAIP3 AND ASSOCIATION OF VARIANTS WITH MULTIPLE AUTOIMMUNE DISEASES

5.1. Abstract

The TNFAIP3 locus has been associated with multiple autoimmune diseases. Here, we sequence the coding portions of the gene to identify polymorphisms that could explain some of the disease associations. A collection of 123 individuals with multiple autoimmune diseases (AIDs; mean=2.2 confirmed diagnoses) and 397 controls is used for initial sequencing with additional genotyping of the most common coding polymorphism, rs2230926, in a large sample of Caucasian individuals from families with multiple AIDs (n=1,099) and unrelated controls (n=743). Thirty-two polymorphisms were identified in the sequencing collection, including 17 novel and 11 coding variants. One novel insertion/deletion polymorphism was significantly associated with multiple autoimmune disease diagnoses (p-value 0.0047; OR (95% CI) – 7.053 (1.67 – 29.79). Further, significant association between rs2230926 alleles and disease is observed for Sjögren's syndrome, psoriasis, Crohn's, rheumatoid arthritis and all autoimmune disease affected individuals in a collection of families with multiple autoimmune diseases. Single disease collections in multiple ethnicities were also genotyped (systemic lupus erythematosus, multiple sclerosis, psoriasis) and association was observed for MS in Caucasians.

5.2. Introduction

Autoimmune diseases (AIDs) are characterized by the misidentification of self as foreign with a resultant immune response that attack's one's own cells and organs. Inheritance patterns have been studied for many of these disorders and they are generally accepted as having a genetic component to susceptibility. Genetic predisposition is multifactorial and disease incidence varies from rare (e. g. idiopathic thrombocytopenic purpura has a population prevalence of .08% in U. S. adults) to common (rheumatoid arthritis (RA) has a Caucasian population prevalence of 5%). Although AIDs affect different systems or organs, it has been well noted that these diseases can

cluster in families and even within individuals. One example is a study of AID clustering in families with multiple sclerosis (MS) described by Barcellos et al. [1].

With such overlapping disease prevalence, it is not surprising that several genetic loci have been associated with more than 1 AID. The hallmark locus is the human leukocyte antigen, which plays a large role in autoimmunity. Another locus is the TNFAIP3 gene and surrounding genomic region which has to date been associated with RA [2-4], systemic lupus erythematosus (SLE) [5-11], psoriasis [12], coeliac disease [13, 14], type 1 diabetes [15], ulcerative colitis [16], Crohn's disease [17] and juvenile idiopathic arthritis [18]. This gene encodes A20, a protein involved in inhibiting signals from the tumor necrosis factor, toll-like receptor and nucleotide-binding oligomerization domain pathways [19-21]. Dysregulation of these pathways results in inflammation and programmed cell death.

With the exception of missense polymorphism rs2230926 (F127C) in SLE, associations to date have been identified outside of coding regions of the gene. One explanation for such associations is that the polymorphisms are in linkage disequilibrium with putatively causal polymorphisms that were not genotyped directly. We sought to identify such mutations for these autoimmune disease associations by sequencing the coding portions of the gene in individuals from the Multiple Autoimmune Disease Genetics Consortium (MADGC) collection [22] who are each affected with multiple autoimmune diseases. This collection includes families affected by more than one autoimmune disease, and here we perform sequencing in individuals who are themselves affected by more than one disease. This sample set provides an opportunity to search for mutations that may be relevant to more than one autoimmune disease, as shown for the PTPN22 AID association using this same collection [22].

TNFAIP3 (NM_006290), at 6q23, is composed of 9 exons with a non-coding exon 1 and partially coding exon 9. The 790 amino acids include an N-terminal cysteine protease OTU domain (Cys103) and 7 C-terminal zinc finger motifs that perform its deubiquitination and ubiquitination

functions [23], respectively. In this study, we sequence 123 subjects each affected with multiple autoimmune diseases and 397 controls. We also perform additional genotyping of the most common coding polymorphism, rs2230926, in the remainder of the MADGC collection in addition to individual disease collections.

5.3. Materials and Methods

5.3.1. DNA Collections

We selected 123 affected subjects from the MADGC collection with at least two of nine core autoimmune diseases each; see Table 5.1 for list and counts of all confirmed disease combinations observed. The four most common combinations were the 2-disease combinations of Hashimoto's thyroiditis with 1 of the following: RA (n=19), SLE (n=11), MS (n=15), or type I diabetes (TID; n=12). The mean disease count was 2.2 with a maximum of 6. Numbers of affected individuals per disease are listed in Table 5.2. Most subjects were Caucasian (N=108), 11 were Caucasian/Native American, 1 was Caucasian/Asian, and 3 were Hispanic. Eighteen families had multiple members sequenced (38 individuals) while the remaining 85 individuals had no relatives sequenced in this study. For association testing, 1 member of each family was randomly selected and a panel of 91 unrelated Caucasian cases was formed. Healthy Caucasian controls (n=397) were enrolled at the University of California San Francisco and includes some individuals from the SOPHIE (Study Of PHarmacogenetics in Ethnically diverse populations) collection.

Genotyping of SNP rs2230926 was conducted in the remainder of the MADGC collection (Caucasians; n=1,099) and in DNA samples from SLE, MS and plaque psoriasis collections. African American, Asian, Caucasian and Hispanic controls were from the SOPHIE collection and additional African American and Caucasian controls were from healthy, normal controls used previously by the MS consortium. Association testing for differences between control groups, including the MADGC family controls, revealed no significant difference so controls have been

combined within ethnic groups for this study to increase our statistical power (data not shown). All subjects gave written informed consent in accordance with the IRB at their respective institution.

Disease Combination	N	Freq
RA, Hashimoto's	19	0.154
Hashimoto's, MS	15	0.122
TID, Hashimoto's	12	0.098
SLE, Hashimoto's	11	0.089
RA, Graves'	4	0.033
SLE, Sjögren's Syndrome	3	0.024
RA, Sjögren's Syndrome	3	0.024
RA, JIA	3	0.024
Hashimoto's, Psoriasis	3	0.024
UC, MS	2	0.016
SLE, MS	2	0.016
RA, SLE	2	0.016
RA, Psoriasis	2	0.016
RA, MS	2	0.016
RA, TID	2	0.016
TID, MS	2	0.016
Graves', Sjögren's Syndrome	2	0.016
Graves', MS	2	0.016
CD, Psoriasis	2	0.016
SLE, UC	1	0.008
SLE, Sjögren's Syndrome, Crest Syndrome	1	0.008
SLE, Psoriasis, Myasthenia Gravis	1	0.008
SLE, Hashimoto's, Vitiligo	1	0.008
SLE, Hashimoto's, Scleroderma, Polymyositis & Dermatomyositis	1	0.008
SLE, Hashimoto's, Myasthenia Gravis	1	0.008
SLE, Hashimoto's, Idiopathic Thrombocytopenia Purpura (ITP)	1	0.008
SLE, Graves', Myasthenia Gravis	1	0.008
SLE, Graves', JIA	1	0.008
SLE, Graves'	1	0.008
RA, SLE, Psoriasis	1	0.008
RA, SLE, Hashimoto's	1	0.008
RA, Hashimoto's, Pernicious or Hemolytic Anemia	1	0.008
RA, Hashimoto's, MS	1	0.008
RA, Graves', Hashimoto's	1	0.008
RA, CD	1	0.008
Psoriasis, MS	1	0.008

Psoriasis, JIA	1	0.008
JIA, Sjögren's Syndrome	1	0.008
TID, SLE, Hashimoto's, Vitiligo, Myasthenia Gravis, IgA deficiency	1	0.008
TID, SLE, Hashimoto's, Vitiligo	1	0.008
TID, JIA	1	0.008
TID, Graves'	1	0.008
Hashimoto's, UC, Pernicious or Hemolytic Anemia	1	0.008
Hashimoto's, Sjögren's Syndrome, Autoimmune Hepatitis	1	0.008
Hashimoto's, Psoriasis, MS	1	0.008
Hashimoto's, CD	1	0.008
Graves', UC	1	0.008
Graves', Psoriasis	1	0.008
Graves', CD	1	0.008

Table 5.1: Disease Combinations among 123 Sequenced MADGC Participants. Confirmed disease combinations for all 123 affected individuals listed in descending order by frequency. Freq – frequency of combination.

Affectation Status	Hash	RA	SLE	MS	TID	Graves'	Psoriasis	Other AIDs	Sjög	IBD	JIA
Affected	74	43	32	28	20	16	13	12	11	10	7
Unaffected	47	80	89	95	103	106	110	104	105	112	116
Reported, Unconfirmed	2	0	2	0	0	1	0	7	7	1	0
Fraction Affected	0.612	0.350	0.264	0.228	0.163	0.131	0.106	0.103	0.095	0.082	0.057

Table 5.2: Autoimmune Disease Distribution in 123 Sequenced MADGC Participants. Autoimmune diseases present in participants listed in order of decreasing frequency from left to right. Fraction affected only includes confirmed cases. Hash - Hashimoto's; RA - rheumatoid arthritis; SLE - systemic lupus erythematosus; MS - multiple sclerosis; TID – type I diabetes; Other AIDs - other autoimmune diseases which include autoimmune Addison's Disease, autoimmune hepatitis, CREST syndrome (limited scleroderma), idiopathic thromobocytopenic purpura, mixed or undifferentiated connective tissue disease, myasthenia gravis, pernicious or hemolytic anemia, polymyositis and dermatomyositis, scleroderma and vitiligo; Sjög - Sjögren's Syndrome; IBD - inflammatory bowel disease; JIA - juvenile idiopathic arthritis.

5.3.2. Sequencing

To sequence all protein coding bases, eight sequencing reactions were performed for each DNA sample. Four sets of PCR primers were from SeattleSNPs (http://pga.gs.washington.edu/) while the other 4 were designed using Primer3 [24]. Detailed primer information can be found in Table 5.3. Primer sets were checked through ePCR on the UCSC genome browser to ensure one unique genomic hit and were also inspected for a lack of known SNPs according to dbSNP.

PCR was performed with 8ng DNA, 0.4µM each forward and reverse primer, 1x buffer, 4mM dNTPs, and 0.3U Qiagen HotStar Taq in a 10µL reaction. PCR was cleaned up by incubation with 1x SAP (PCR Clean-Up Reagent, PerkinElmer Life Sciences, Inc.) at 37°C for one hour. Sequencing reactions contained 2.5µL of clean PCR product, 0.375µM primer and 8.3% Applied Biosystems (ABI) BigDye Terminator v3.1 in a 12µL reaction. Excess dye terminator removal was performed with genCLEAN plates following manufacturer's instructions before sequencing on an ABI 3730xL DNA Analyzer. Sequencing was performed in one direction, except for regions with insertion-deletion polymorphisms and novel polymorphisms which were confirmed by sequencing the other strand.

Exons	Amp. Coordinates	Forward Primer, 5' to 3'	Reverse Primer, 5' to 3'	PD
2	138233445-138234423	GGGGCTAAAGAGGAAACACC	CTTCATGAATGGGGATCCAG	Primer3
3	138237232-138238024	CCCTGTGTGCTCCTCCTTAG	CCACTGGAGGTTTCTGGTGT	Primer3
4 & 5	138238303-138239063	TCCCCAACTTTTGAGTTTGC	AAGCAAAAAGGAAAACCCTGA	Primer3
6	138239258-138240237	CAAGTAAACGCCTGTCAGGTTAG	ACCATGCACAAGACTCTGAATTT	SeattleSNPs
7	138240758-138241872	CGTCTTAGTTACTCATGGCTGCT	TAAATGTCCTGGTAACATCCTGG	SeattleSNPs
7	138241682-138242615	GTTCAGTGAGACCACTGCCAT	TGAGAGATTTCCAAACCACATCT	SeattleSNPs
8	138242382-138243465	GCAGCTCCTAATATCACATTCCA	TCTGTCTGTTCGCTCCTTATGAT	SeattleSNPs
9	138243737-138244509	CCTTGCTCAGGCAGGTAAAG	AGCCAAGACGATGAAGCAGT	Primer3

Table 5.3: Sequencing Primer Pairs. Amp. Coordinates – amplicon positions on chromosome 6; PD – primer design software/source.

5.3.3. Genotyping

Genotyping was performed with a predesigned ABI TaqMan assay for SNP rs2230926 following the manufacturer's protocol. We used 2x PCR Universal Master Mix and 4.5ng DNA in a 5µL reaction. Duplicates and no template controls were checked for quality control purposes.

5.3.4. Analysis

Sequencing traces were analyzed with Sequencher (Gene Codes). Hardy-Weinberg equilibrium (HWE) p-values were calculated in Haploview [25] to assess sequencing quality and a p-value of 0.001 was used as the significance threshold for exclusion. Individual polymorphism tests for association between sequenced cases and controls were conducted in Plink [26]. We used Fisher's exact test and also conducted adaptive permutation tests by swapping case-control status to calculate empirical p-values for each variant. In order to mitigate the potential for false positives due to population stratification, we restricted the analysis to Caucasian samples, and we also trimmed the panel to unrelated individuals at the same time, which reduced the number of cases from 123 to 91.

A single haplotype block was defined using the spine of LD definition in Haploview. Haplotype tests for association were conducted in Plink for 24-variant combinations for all frequencies and also restricted to those with a frequency greater than 1 percent. The 24 variants were polymorphic in cases or controls when analysis was restricted to unrelated Caucasian cases.

Weighted sums analysis was performed to test for association with disease for a group of variants, a powerful method especially for rare variants where each polymorphism contributes only a small amount of risk. We used a custom script according to the method of Madsen et al. [27]. We checked for differences in common (>2% MAF), rare, exonic, intronic, non-synonymous, synonymous, and untranslated region (UTR) variants between cases and controls. Association testing for rs2230926 was performed in Plink and HWE p-values were checked in Haploview with criteria as above.

5.4. Results

5.4.1. Sequencing of TNFAIP3 in Cases and Controls

We identified 33 polymorphisms through the sequencing of 246 case and 794 control chromosomes (Table 5.4). One was dropped from analysis for being out of HWE (rs3214646) and probably does not represent a true polymorphic locus. Eleven were in protein coding regions; 8 of these were non-synonymous and 3 were synonymous. The synonymous SNP, Leu725Leu, is located in zinc-finger motif 6. Seventeen were novel, or not in the public database dbSNP, including 9 of the coding variants. Seven variants were missing from the control sequencing and 1 from the case sequencing data and were not included in comparisons between cases and controls. For the 2 variants detected in cases only, rs5029964 was in 1 of 2 family members sequenced and novel_2 was in an individual with no other family members sequenced.

SNP ID	SNP Coordinate	Seq In Ctrls	Seq In Cases	SNP property	Alleles	MAF
rs5029933	138233755	N	Y	Intron 1	A/G	0.049
novel_1	138233963	N	Y	Intron 1	G/C	0.008
rs3214646*	138234018	N	Y	Intron 1	T/-	0.500
novel_8	138234044	Y	Y	Exon 2, 5' UTR	T/G	0.003
novel_2	138234294	Y	Y	Exon 2, Ser79Arg	C/G	0.001
rs5029938	138237326	N	Y	Intron 2	C/T	0.049
rs643177	138237386	N	Y	Intron 2	C/T	0.248
rs5029939	138237416	N	Y	Intron 2	C/G	0.041
novel_3	138237419	N	Y	Intron 2	A/C	0.004
rs5029940	138237657-9	Y	Y	Intron 2 (-15 to -18 from Ex. 3)	-/CCT	0.352
novel_9	138237684	Y	Y	Exon 3, Asn102Ser	A/G	0.001
rs2230926	138237759	Y	Y	Exon 3, Phe127Cys	T/G	0.029
novel_10	138237849	Y	Y	Exon 3, Leu157Pro	T/C	0.001
rs5029947	138238510	Y	Y	Intron 3 (-8bp from Ex. 4)	C/G	0.004
rs5029948	138239022	Y	Y	Intron 5	C/T	0.052
rs661561	138239024	Y	Y	Intron 5	C/A	0.342
rs5029964	138239034	Y	Y	Intron 5	A/G	0.001
rs582757	138239517	Y	Y	Intron 5	T/C	0.268
novel_4	138239582	Y	Y	Intron 5	C/-	0.01
novel_11	138241009	Y	N	Intron 6	A/G	0.001
rs610604	138241110	Y	Y	Intron 6	T/G	0.323
novel_12	138241591	Y	Y	Exon 7, Arg439Gln	G/A	0.001
novel_13	138241913	Y	Y	Exon 7, Glu546Glu	G/A	0.001
rs5029953	138242453	Y	Y	Intron 7	G/A	0.009
rs5029965	138242545	Y	Y	Intron 7	G/A	0.011
novel_5	138242933	Y	Y	Exon 8, Thr647Pro	A/C	0.004
novel_14	138243823	Y	Y	Intron 8	G/A	0.006
novel_15	138243916	Y	Y	Exon 9, Pro714Ser	C/T	0.001
novel_6**	138243951	Y	Y	Exon 9, Leu725Leu	G/A	0.004
novel_16	138244007	Y	Y	Exon 9, Gly744Asp	G/A	0.001
rs5029956	138244071	Y	Y	Exon 9, Pro765Pro	C/T	0.003
novel_7	138244250	Y	Y	Exon 9, 3' UTR	G/T	0.013
novel_17	138244323	Y	Y	Exon 9, 3' UTR	G/A	0.001

Table 5.4: Polymorphism Discovery Summary for Cases and Controls. Coordinates obtained from hg18. Flanking sequences are on the positive strand of the genome and SNP alleles are shown as Major/Minor. Seq - Sequenced, data is not available for groups that were not sequenced at that base. Ctrls - Controls; MAF- Minor Allele Frequency. *rs3214646 removed for violation of Hardy-Weinberg Equilibrium (P=3.7009E-36). **Novel SNP 6 is located within zinc-finger motif 6.

5.4.2. Association Testing of Sequenced Variants

Fisher's exact tests for association of identified variants in cases versus controls were performed for 24 SNPs and insertion/deletion polymorphisms. Comparing 91 unrelated, Caucasian multiply affected individuals to 397 Caucasian controls revealed significant association for one intronic insertion/deletion polymorphism with multiple AID diagnoses (Novel_4; Fisher P = 0.0090; OR = 7.053, 95% CI 1.67 - 29.79; Table 5.5). It also remains significant after permutation testing (permuted P = 0.0047). One SNP was not polymorphic in this restricted dataset (rs5029964), so it was not tested for association.

An omnibus test for association of 24-marker haplotypes with a frequency at least 1% was highly significant, with a p-value of 2.94×10^{-05}. Keeping the haplotype frequency threshold of 1% or greater revealed 3 significant haplotypes, none of which contained the risk allele for Novel_4 (data not shown). When we included all frequency haplotypes, 8 reached significance given an alpha of 0.05 and one was borderline significant (Table 5.6 contains results for these 9 haplotypes). Additionally, we tested for differences in polymorphisms between cases and controls with weighted sums analysis and found cases to be enriched for 5' and 3' UTR variants (one-side p-value 0.04).

SNP	A1	F_A	F_U	A2	OR	L95	U95	Fisher P	EMP1	NP
novel_8	G	0.0000	0.0040	T	0	0	NA	1	1	6
novel_2	C	0.0055	0.0000	G	NA	NA	NA	0.1944	0.1818	98
rs5029940	C	0.3407	0.3539	A	0.9433	0.6705	1.327	0.7952	1	6
novel_9	G	0.0000	0.0013	A	0	0	NA	1	0.8571	6
rs2230926	G	0.0275	0.0241	T	1.142	0.4185	3.119	0.7908	0.7778	8
novel_10	C	0.0000	0.0013	T	0	0	NA	1	0.8571	6
rs5029947	G	0.0000	0.0013	C	0	0	NA	1	0.3261	45
rs5029948	T	0.0550	0.0536	C	1.026	0.5031	2.093	1	1	6
rs661561	A	0.3407	0.3445	C	0.9831	0.6985	1.384	1	1	6
rs5029964	0	0.0000	0.0000	A	---	---	---	---	---	---
rs582757	C	0.2582	0.2739	T	0.9228	0.6383	1.334	0.7106	1	6
novel_4	A	0.0275	0.0040	C	7.053	1.67	29.79	0.0090	0.0047	4700
rs610604	G	0.3132	0.3240	T	0.9514	0.6714	1.348	0.8596	1	6
novel_12	A	0.0000	0.0013	G	0	0	NA	1	0.3261	45
novel_13	A	0.0000	0.0013	G	0	0	NA	1	0.2464	68
rs5029953	A	0.0000	0.0077	G	0	0	NA	0.6013	0.55	19
rs5029965	A	0.0110	0.0119	G	0.9222	0.1976	4.305	1	0.8571	6
novel_5	C	0.0055	0.0040	A	1.387	0.1434	13.41	0.5787	0.625	15
novel_14	A	0.0000	0.0080	G	0	0	NA	0.6031	0.7273	10
novel_15	T	0.0000	0.0013	C	0	0	NA	1	0.1818	98
novel_6	A	0.0110	0.0013	G	8.367	0.7545	92.78	0.09841	0.1418	133
novel_16	A	0.0000	0.0013	G	0	0	NA	1	0.2143	83
rs5029956	T	0.0000	0.0040	C	0	0	NA	1	0.7778	8
novel_7	T	0.0000	0.0133	G	0	0	NA	0.2239	0.2623	60
novel_17	A	0.0000	0.0013	G	0	0	NA	1	0.2118	84

Table 5.5: Association Testing of Sequenced Variants. 91 unrelated, multiply affected, Caucasian MADGC cases vs. 397 controls. OR's for variants with F_A or F_U of 0 cannot be calculated. SNPs are ordered by genomic position. rs5029964 was not polymorphic when restricted to these samples. A1 – Allele 1; F_A – Frequency in cases; F_U – Frequency in controls; A2 – Allele 2; OR – Odds ratio; L95 – Lower 95% confidence interval; U95 – Upper 95% confidence interval; EMP1 – Empirical P-value; NP – Number of permutations; NA – not available.

HAPLOTYPE	Hap Freq	F_A	F_U	P	Novel_4 Risk
TGAATTCCACCTGGGGAGCGGCGG	0.0464	0.0039	0.0566	0.0024	N
TGCATTCCCTCGGGGGAGCGGCGG	0.0578	0.0050	0.0704	0.0007	N
TGAATTCCCTCTGGGGAGCGGCGG	0.5312	0.6128	0.5117	0.0141	N
TGCATTCCACCGGGGGAGCGGCGG	0.1938	0.2459	0.1813	0.0478	N
TGAATTCCCTCTGGGGAGCAGCGG	0.0032	0.0110	0.0013	0.0372	N
TGAATTCCCTATGGGGAGCGGCGG	0.0038	0.0142	0.0014	0.0124	Y
TGCATTCCACAGGGGGAGCGGCGG	0.0016	0.0084	0.0000	0.0111	Y
TCAATTCCCTCTGGGGAGCGGCGG	0.0011	0.0055	0.0000	0.0406	N
TGAATTCCATCTGGGGAGCGGCGG	0.0163	0.0000	0.0202	0.0530	N

Table 5.6: Haplotype Testing Results between Sequenced Cases and Controls. Significant or borderline significant haplotypes listed for 24 polymorphisms sequenced in cases and controls listed in order as in Table 5.4 with rs5029940 coded as C/A and novel_4 coded as A/C. 91 unrelated, multiply affected, Caucasian MADGC cases vs. 397 controls. Hap Freq – haplotype frequency; F_A – frequency in cases; F_U – frequency in controls; P – unadjusted p-value. Novel-4 Risk refers to risk allele for SNP significantly associated with disease in single marker testing.

5.4.3. Association Testing of rs2230926

As the coding SNP, rs2230926, was previously associated with SLE, we genotyped it in the entire MADGC collection which included 1099 affected Caucasian participants and 815 unaffected Caucasian family controls. We also genotyped the SNP in 743 unrelated healthy Caucasian controls from the MS consortium and the SOPHIE collection (Table 5.7). Controls were combined as no significant difference was observed between the groups. The strongest OR observed was for Sjögren's syndrome (p = 0.0523; OR = 3.092), followed by psoriasis (p = 0.0030; OR = 2.489), Crohn's disease (p = 0.0378; OR **=** 2.267), ulcerative colitis (p = 0.4759; OR = 0.4929) and RA (p = 0.0178; OR = 1.883) which, except for ulcerative colitis, reached at least borderline significance given an alpha of 0.05. Significant association was also achieved for the comparison of all affected individuals versus controls (p = 0.0336; OR (95% CI) = 1.385 (1.024 - 1.872). Borderline significant association was observed for Graves' disease and for multiply affected individuals,

where samples with at least 2 confirmed AIDs were tested against controls (p = 0.0744 & 0.0617, respectively).

We performed additional genotyping of rs2230926 in multi-ethnic disease-specific cohorts for MS, SLE, and psoriasis (Table 5.8). The plaque psoriasis samples were tested separately as subtypes of psoriasis are thought to represent genetically distinct diseases and this subtype represented the majority of cases in our collection. Significant association was observed for Caucasian MS samples (p= 0.0116; OR 1.787, 95% CI 1.132-2.821), which also had the strongest OR out of individual diseases tested. The next highest OR's were observed in Asian American SLE (1.373) and Hispanic SLE (1.331) tests, neither of which reached significance. The OR's closest to 1 were 1.066 for plaque psoriasis in a Caucasian sample and 1.085 in African American SLE.

	Cases			Controls	
Disease	**N**	**G count**	**MAF**	**OR (95%CI)**	**P**
Sjogren's	18	3	0.083	3.092 (0.9308 - 10.27)	0.0523
Psoriasis	88	12	0.068	**2.489 (1.335 - 4.64)**	**0.0030**
Crohn's Disease**	56	7	0.063	**2.267 (1.025 - 5.014)**	**0.0378**
Ulcerative Colitis**	35	1	0.014	0.4929 (0.06769 - 3.589)	0.4759
RA	162	17	0.052	**1.883 (1.106 - 3.206)**	**0.0178**
Graves' Disease	86	9	0.052	1.878 (0.9293 - 3.795)	0.0744
Multiply affected	158	15	0.047	1.695 (0.9684 - 2.966)	0.0617
P or H anemia*	22	2	0.045	1.62 (0.386 - 6.796)	0.5058
SLE	131	11	0.042	1.491 (0.7863 - 2.825)	0.2183
IBD (IC, UC, CD)	97	8	0.041	1.463 (0.699 - 3.061)	0.3098
All Affected	1099	86	0.039	**1.385 (1.024 - 1.872)**	**0.0336**
MS	209	10	0.024	0.8336 (0.4301 - 1.616)	0.5894
ITP*	15	1	0.033	1.173 (0.158 - 8.706)	0.8760
Hashimoto's	266	16	0.030	1.055 (0.6144 - 1.81)	0.8470
TID	84	5	0.030	1.043 (0.4181 - 2.604)	0.9276
JIA	32	0	0.000	--	0.1703
Vitiligo*	12	0	0.000	--	0.4009

Table 5.7: MADGC Collection Genotyping and Allelic Association of rs2230926. Results sorted in descending order by OR strength. Controls included 1558 individuals with MAF of 0.0286. P-values < 0.05 and corresponding OR's are in bold. MAF - minor allele frequency; HWE - Hardy-Weinberg equilibrium, OR - odds ratio; CI - confidence interval; SLE - systemic lupus erythematosus; RA - rheumatoid arthritis; MS - multiple sclerosis; TID - type I diabetes; IBD - inflammatory bowel disease; IC - idiopathic colitis; UC - ulcerative colitis; CD - Crohn's disease; JIA - juvenile idiopathic arthritis; P or H anemia - pernicious or hemolytic anemia; ITP - idiopathic

thromobocytopenic purpura. Diseases marked with * were not part of the 9 core diseases in the study. Diseases marked ** are subtypes of IBD.

		Cases		Controls				
Sample	**Ethnicity**	**N**	**MAF**	**N**	**MAF**	**HWE P**	**P-value**	**OR (95% CI)**
MS	Cauc	373	0.048	743	0.028	1	0.0116	1.787 (1.132-2.821)
SLE	AsAm	201	0.050	177	0.037	0.055	0.3817	1.373 (0.6729-2.803)
SLE	His	185	0.070	214	0.054	1	0.332	1.331 (0.7458-2.375)
Psoriasis-All	Cauc	701	0.031	743	0.028	0.560	0.5467	1.142 (0.7415-1.759)
MS	AfAm	773	0.340	656	0.319	0.341	0.2344	1.1 (0.9402-1.286)
SLE	AfAm	150	0.337	656	0.319	0.526	0.5456	1.085 (0.8319-1.416)
Plaq Psor	Cauc	664	0.029	743	0.028	0.630	0.7771	1.066 (0.6835-1.664)

Table 5.8: Allelic Tests for Association of rs2230926 with Psoriasis, MS and SLE. Results sorted in descending order by OR. Collections are separated by ethnicity. Psoriasis-All refers to all subtypes; Plaq Psor – plaque psoriasis; Cauc – Caucasian; AsAm – Asian American; His – Hispanic; AfAm – African American; N – number of individuals; MAF – minor allele frequency; HWE P – Hardy-Weinberg Equilibrium p-value; OR – odds ratio; CI – confidence interval.

5.5. Discussion

This study represents, to our knowledge, the first comprehensive screening of coding exons of the gene encoding A20. We have screened a population affected by multiple autoimmune diseases given recent association with several autoimmune phenotypes. This gave us the opportunity to test for association of variants with multiple AID diagnoses. Additionally, we performed more extensive genotyping and association testing of the previously associated coding SNP rs2230926.

We identified 32 polymorphisms, 17 novel and 11 coding, in cases and controls. One intronic insertion/deletion polymorphism was significantly associated with multiple AID diagnoses after correcting for multiple comparisons. We also identified 9 haplotypes significantly or marginally associated with multiple autoimmune disease diagnoses, 2 containing the intronic polymorphism associated with disease in single marker tests. Cases were found to be enriched for 5' and 3' UTR variants compared to controls. As we chose to focus on coding exons in our study, we may miss important polymorphisms present in regulatory regions, but have captured protein coding SNPs and insertion/deletion polymorphisms.

We did not observe association with SLE for rs2230926 within the full MADGC collection and this could be due to lack of power; we only had 23% power given our observed OR and sample size at an alpha of 0.05. This could also be due to the MADGC collection families having different predisposition towards disease. We did observe significant association with this variant and an increased risk to the combination of all AIDs, Sjögren's syndrome, psoriasis, Crohn's disease, and RA in the context of families affected by multiple autoimmune diseases. The differences in OR's observed between UC (0.4929) and Crohn's (2.267) are quite striking as they are two types of inflammatory bowel disease. Also striking are the differences in OR's for the thyroid diseases Hashimoto's (1.055) and Graves' disease (1.878). Our data support a role for this variant as a general autoimmune disease susceptibility risk factor.

Finally, we observed significant association with MS in Caucasians in a disease specific DNA panel. We also tested for association in an African American MS collection; Asian American, Hispanic and African American SLE participants; and Caucasian psoriasis samples. No significant association was observed in these tests, but we lacked power given the sample size, minor allele frequency and OR for each comparison. The most well powered comparison was in the Caucasian MS sample, with a power of 55-66% given a relative risk of 1.7-1.8.

In conclusion, we have identified many polymorphisms in this gene and have identified one new insertion/deletion variant associated with multiple autoimmune disease diagnoses. The known coding variant, rs2230926, was identified as a general autoimmune disease risk factor and was also associated with Sjögren's, psoriasis, Crohn's, and RA in a large Caucasian collection of families with multiple autoimmune diseases. Finally, in a set of multi-ethnic single-disease cohorts, the same coding variant was associated with MS in a Caucasian sample.

5.6. References

1. Barcellos, L.F., et al., *Clustering of autoimmune diseases in families with a high-risk for multiple sclerosis: a descriptive study.* Lancet Neurol, 2006. **5**(11): p. 924-31.

2. Plenge, R.M., et al., *Two independent alleles at 6q23 associated with risk of rheumatoid arthritis.* Nat Genet, 2007. **39**(12): p. 1477-82.

3. Dieguez-Gonzalez, R., et al., *Analysis of TNFAIP3, a feedback inhibitor of nuclear factor-kappaB and the neighbor intergenic 6q23 region in rheumatoid arthritis susceptibility.* Arthritis Res Ther, 2009. **11**(2): p. R42.

4. Thomson, W., et al., *Rheumatoid arthritis association at 6q23.* Nat Genet, 2007. **39**(12): p. 1431-3.

5. Cai, L.Q., et al., *A single-nucleotide polymorphism of the TNFAIP3 gene is associated with systemic lupus erythematosus in Chinese Han population.* Mol Biol Rep, 2009.

6. Han, J.W., et al., *Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus.* Nat Genet, 2009. **41**(11): p. 1234-7.

7. Bates, J.S., et al., *Meta-analysis and imputation identifies a 109 kb risk haplotype spanning TNFAIP3 associated with lupus nephritis and hematologic manifestations.* Genes Immun, 2009. **10**(5): p. 470-7.

8. Cai, L.Q., et al., *A single-nucleotide polymorphism of the TNFAIP3 gene is associated with systemic lupus erythematosus in Chinese Han population.* Mol Biol Rep. **37**(1): p. 389-94.

9. Shimane, K., et al., *The association of a nonsynonymous single-nucleotide polymorphism in TNFAIP3 with systemic lupus erythematosus and rheumatoid arthritis in the Japanese population.* Arthritis Rheum. **62**(2): p. 574-9.

10. Graham, R.R., et al., *Genetic variants near TNFAIP3 on 6q23 are associated with systemic lupus erythematosus.* Nat Genet, 2008. **40**(9): p. 1059-61.

11. Musone, S.L., et al., *Multiple polymorphisms in the TNFAIP3 region are independently associated with systemic lupus erythematosus.* Nat Genet, 2008. **40**(9): p. 1062-4.

12. Nair, R.P., et al., *Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways.* Nat Genet, 2009. **41**(2): p. 199-204.

13. Trynka, G., et al., *Coeliac disease associated risk variants in TNFAIP3 and REL implicate altered NF-{kappa}B signalling.* Gut, 2009.

14. Coenen, M.J., et al., *Common and different genetic background for rheumatoid arthritis and coeliac disease.* Hum Mol Genet, 2009. **18**(21): p. 4195-203.

15. Fung, E.Y., et al., *Analysis of 17 autoimmune disease-associated variants in type 1 diabetes identifies 6q23/TNFAIP3 as a susceptibility locus.* Genes Immun, 2009. **10**(2): p. 188-91.

16. Wang, K., et al., *Comparative genetic analysis of inflammatory bowel disease and type 1 diabetes implicates multiple loci with opposite effects.* Hum Mol Genet.

17. Consortium, T.W.T.C.C., *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature, 2007. **447**(7145): p. 661-678.

18. Prahalad, S., et al., *Variants in TNFAIP3, STAT4, and C12orf30 loci associated with multiple autoimmune diseases are also associated with juvenile idiopathic arthritis.* Arthritis Rheum, 2009. **60**(7): p. 2124-2130.

19. Dixit, V.M., et al., *Tumor necrosis factor-alpha induction of novel gene products in human endothelial cells including a macrophage-specific chemotaxin.* J Biol Chem, 1990. **265**(5): p. 2973-8.

20. Boone, D.L., et al., *The ubiquitin-modifying enzyme A20 is required for termination of Toll-like receptor responses.* Nat Immunol, 2004. **5**(10): p. 1052-60.

21. Hitotsumatsu, O., et al., *The ubiquitin editing enzyme A20 restricts NOD2 triggered signals.* Immunity, 2008. **28**: p. 381-390.

22. Criswell, L.A., et al., *Analysis of families in the multiple autoimmune disease genetics consortium (MADGC) collection: the PTPN22 620W allele associates with multiple autoimmune phenotypes.* Am J Hum Genet, 2005. **76**(4): p. 561-71.

23. Wertz, I.E., et al., *De-ubiquitination and ubiquitin ligase domains of A20 downregulate NF-kappaB signalling.* Nature, 2004. **430**(7000): p. 694-9.

24. Rozen, S. and H. Skaletsky, *Primer3 on the WWW for general users and for biologist programmers.* Methods Mol Biol, 2000. **132**: p. 365-86.

25. Barrett, J.C., et al., *Haploview: analysis and visualization of LD and haplotype maps.* Bioinformatics, 2005. **21**(2): p. 263-5.

26. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.

27. Madsen, B.E. and S.R. Browning, *A groupwise association test for rare mutations using a weighted sum statistic.* PLoS Genet, 2009. **5**(2): p. e1000384.

CHAPTER 6

MULTIPLE POLYMORPHISMS IN THE TNFAIP3 REGION ARE INDEPENDENTLY ASSOCIATED WITH SYSTEMIC LUPUS ERYTHEMATOSUS[1]

6.1. Abstract

The tumor necrosis factor alpha-induced protein 3 (TNFAIP3) gene encodes a ubiquitin editing enzyme, A20, that restricts NFκB dependent signaling and prevents inflammation. We show that 3 independent SNPs in the TNFAIP3 region are associated with systemic lupus erythematosus (SLE) among individuals of European ancestry. Further, an A20 protein bearing the risk allele of a non-synonymous SNP, rs2230926, displays a decreased ability to restrict TNF-induced NFκB activity in vitro. These findings provide critical links between A20 and the etiology of SLE.

6.2. Introduction

Autoimmune diseases are characterized by persistent or recurrent inflammation in the absence of explicit microbial infection. SLE is the prototypic systemic autoimmune disease. Although the disease is genetically complex, substantial work over the past decade has led to the identification of several reproducible genetic risk factors for SLE [1].

A20, the product of the TNFAIP3 gene, is an NFκB inducible protein expressed in multiple cell types and required for preventing spontaneous inflammation [2]. The elimination of A20 from mice leads to severe spontaneous inflammation, cachexia and premature death [2]. A20 regulates the ubiquitylation of key signaling proteins and restricts the duration of both tumor necrosis factor and Toll-like receptor induced NFκB signals [3-5]. Thus, A20 is a potent endogenous anti-inflammatory molecule. As the TNFAIP3 gene is well conserved between humans and mice, and given recent evidence supporting association of this gene with rheumatoid

[1] This work was previously published and is reproduced with permission.

Musone SL*, Taylor KE*, Lu T, Nititham J, Ferreira RC, Ortmann W, Shiffrin N, Petri MA, M. Kamboh I, Manzi S, Seldin MF, Gregersen PK, Behrens TW, Ma A, Kwok PY, Criswell LA. Multiple polymorphisms in the TNFAIP3 region are independently associated with systemic lupus erythematosus. *Nat Genet* **40**, 1062-4 (2008).

*S.L.M. and K.E.T. contributed equally to this work.

arthritis (RA) [6, 7], we hypothesized that hypomorphic mutations of the TNFAIP3 gene might also be associated with SLE. Previous GWAS analysis had not identified this gene locus as a strong candidate region for association with SLE.

6.3. Materials and Methods

6.3.1. Subjects

SLE cases were obtained from four sources. Patients from the University of California, San Francisco (UCSF) were participants in the UCSF Lupus Genetics Project and were recruited from UCSF Arthritis Clinics and private rheumatology practices in northern California, as well as by nationwide outreach [8]. SLE patients contributed by the Autoimmune Biomarkers Collaborative Network (ABCoN) [9] were recruited from the Hopkins Lupus cohort [10]. A third case series was part of the Multiple Autoimmune Disease Genetics Consortium (MADGC) collection [11]. Finally, a fourth set of cases recruited from the Pittsburgh Lupus Registry were obtained from the University of Pittsburgh [12]. Unrelated healthy controls were from the New York Health Project (NYHP) [13] (http://www.amdec.org/amdec_initiatives/nycp.html). All cases were confirmed for SLE diagnosis by documentation of at least four American College of Rheumatology (ACR) criteria [14] in medical record reviews (95%) or by written confirmation from a treating rheumatologist. Cases were typical of SLE case series of European descent, being 93% female and having an average age of onset of 35 years (SD ± 13 years). Twenty-eight percent of subjects meet ACR criteria for renal disease and 79% meet ACR criteria for arthritis, as has been reported previously [15, 16]. The Institutional Review Boards of all investigative institutions approved these studies, and all cases and controls gave written informed consent.

6.3.2. Genotyping and SNP Selection

All cases and controls were genotyped using the Illumina HumanHap550 array, as reported previously [15]. ABCoN and MADGC cases and a subset of NYHP controls (n = 869) were genotyped on the version 1 Illumina 550K panel. All other subjects were genotyped on the version 3 Illumina 550K panel. Additional genotyping for rs2230926 in ABCoN and MADGC

cases was performed using a pre-validated TaqMan (Applied Biosystems) assay according to manufacturer's instructions. SNPs were removed from analysis that had a minor allele frequency less than 5% (with the exception of the non-synonymous SNP, rs2230926), greater than 10% missing genotypes, or Hardy-Weinberg equilibrium $p < 0.001$ in controls. Of the 158 SNPs in the extended TNFAIP3 region, 143 passed quality control filters; in the initial 500-kb region, 115 passed quality control filters.

6.3.3. Statistical Analysis

Subjects were first removed for whom there was evidence of duplication or relatedness in the Illumina 550K data, using IBS estimation in PLINK [17] (http://pngu.mgh.harvard.edu/purcell/plink), and who had < 90% of genotypes called. While all subjects were of self-reported European ancestry, in order to guarantee genetic homogeneity we performed ancestry analysis using STRUCTURE [18] and a set of 235 ancestry-informative markers (AIMs) contained in the Illumina 550K panel. Subjects were removed who had < 90% estimated European ancestry.

We conducted allelic tests of cases and controls using Haploview [19]. Conditional analyses to determine independent effects were performed in Whap [20] (http://pngu.mgh.harvard.edu/purcell/whap), which uses log-ratio testing of alternative models. Stata 9.2 (http://www.stata.com/) was used for multivariate logistic regression of the three independent SNPs. Tagger [21] was used to measure r^2 between SNPs in the HapMap CEU population to determine proxies for SNPs not genotyped in our samples.

We performed stratified analyses designed to determine whether population substructure within our European subjects explained the associations of TNFAIP3 region SNPs with SLE. We first used a set of 1409 EUROSTRUCTURE AIMS [22] to estimate percent northern versus southern European ancestry. We also used the first 4 principal components determined by EIGENSTRAT [23] using whole-genome Illumina 550K data, as in Taylor et al., [16] to determine a subset of

genetically homogeneous subjects and therefore account for more subtle substructure than simply north-south. Greater than or equal to 90% membership in the northern population and membership in the homogeneous subset were each then used as stratifiers in allelic analyses of the top 3 SNPs. Strata were analyzed separately and then combined using the Mantel-Haenszel method; tests of heterogeneity and combined ORs were performed with Stata 9.2.

6.3.4. NFκB Response Assay

Human A20 cDNAs corresponding to the major and minor alleles at rs2230926 were generated by RT-PCR and Quik-change mutagenesis (Stratagene). These cDNAs were verified by sequencing and transiently transfected into 293T cells along with NFκB-luciferase and CMV-renilla reporter constructs, stimulated with 10 ng/ml TNF for 6 hours and then lysed for renilla and luciferase assays using a dual luciferase reporter assay (Promega). A20 and actin protein expression levels were determined by immunoblotting of whole cell lysates and densitometric quantification. Relative A20 expression levels between samples were determined after quantitating and normalizing A20 expression to actin expression for each sample. All assays were performed at least three times and p-values were determined by unpaired Student's T test.

6.4. Results

To examine the potential role of TNFAIP3 in SLE, we utilized data from a recently published genome-wide association study [15]. Table 6.1 shows the number of cases and controls before and after quality control filters. In total, 1,239 SLE cases and 1629 controls were included in this analysis. We initially selected 129 contiguous SNPs from the TNFAIP3 region on chromosome 6, extended with flanking regions approximately 250kb on either side of the gene coding sequence (138,000 kb to 138,500 kb). This region also captures the PERP gene, an apoptosis effector. Since we observed significant SNPs in LD blocks at the boundaries of the initial region, we later extended our analysis to 158 SNPs in the region from 137,975 kb to 138,550 kb.

	Illumina 550K genotyped*	Post-QC**
Cohort 1 (ABCoN and MADGC) cases	446	394
Cohort 2 (U. C. San Francisco) cases	611	564
Cohort 3 (U. Pittsburgh) cases	319	281
Total cases	**1376**	**1239**
NYHP controls	**1762**	**1629**
*After removal of duplicate samples and first-degree relatives. **After removal of subjects with < 90% genotyping or < 90% European ancestry by STRUCTURE[18] analysis.		

Table 6.1: Summary of Genotypes by Source Before and After Quality-Control Filters.

Additional genotyping was performed for the TNFAIP3 non-synonymous coding SNP rs2230926 in 393 of the SLE cases, as they were typed on the version 1 Illumina 550K panel which did not include this SNP. In the controls, 869 were typed on the version 1 array and therefore did not have data available for rs2230926. A subgroup analysis of the testing below using only cases (n=1,239) and controls (n=760) that were typed for rs2230926 revealed essentially the same results (data not shown).

A total of 21 SNPs in the region had allelic $p \leq 0.005$ (Table 6.2). At this screening stage we used a liberal cutoff, given at least 10 independent haplotype blocks in the region. SNP rs13192841 had the smallest p-value, 5.4×10^{-8} (OR 1.4, 95% CI 1.2 – 1.6), while SNP rs2230926 had the highest OR, 2.0 (95% CI 1.4 – 3.0, $p=3.0 \times 10^{-4}$). All of these top 21 were in the initial 500kb region covered by 129 SNPs (Figure 6.1); the extension to a 575kb region with 29 additional SNPs did not yield new candidates. Based on data from the HapMap CEU population ($r^2=1$), SNPs rs6933404 and rs2327832 are perfect proxies for RA-associated SNP rs6920220 while SNPs rs13192841 and rs12527282 are perfect proxies for another RA-associated SNP, rs10499194 [6, 7].

SNP*	Name	A	Case,Ctrl Ratio Counts	Case,Ctrl Freq	OR (95% CI)	P value
2	rs6933404	G	588:1882, 614:2640	0.24, 0.19	1.3 (1.2 - 1.5)	5.6E-06
4	rs600469	G	1229:1245, 1480:1746	0.50, 0.46	1.2 (1.05 - 1.3)	0.0044
5	rs13192841	G	1753:543, 2208:960	0.76, 0.70	1.4 (1.2 - 1.6)	5.4E-08
6	rs12527282	G	1833:629, 2239:993	0.75, 0.69	1.3 (1.15 - 1.5)	1.8E-05
8	rs2327832	G	578:1872, 587:2613	0.24, 0.18	1.4 (1.2 - 1.6)	1.4E-06
10	rs686851	G	1229:1247, 1486:1760	0.50, 0.46	1.2 (1.1 - 1.3)	0.0038
11	rs1002658	G	2051:415, 2583:639	0.83, 0.80	1.2 (1.1 - 1.4)	0.0039
12	rs525977	G	1229:1247, 1493:1763	0.50, 0.46	1.2 (1.05 - 1.3)	0.0045
13	rs6904167	G	1202:1214, 1466:1728	0.50, 0.46	1.2 (1.05 - 1.3)	0.0042
17	rs636393	A	1059:625, 855:663	0.63, 0.56	1.3 (1.1 - 1.5)	2.0E-04
18	rs602414	A	1557:917, 1882:1374	0.63, 0.58	1.2 (1.1 - 1.4)	8.5E-05
58	rs2230926	C	114:2342, 36:1484	0.05, 0.02	2.0 (1.4 - 3.0)	3.0E-04
105	rs2484066	C	1403:1045, 1729:1507	0.57, 0.53	1.2 (1.1 - 1.3)	0.0036
106	rs9494941	A	1556:908, 1908:1346	0.63, 0.59	1.2 (1.1 - 1.3)	5.0E-04
108	rs1931867	A	1534:888, 1858:1296	0.63, 0.59	1.2 (1.1 - 1.3)	8.0E-04
110	rs6922466	A	1905:531, 2378:844	0.78, 0.74	1.3 (1.1 - 1.4)	1.0E-04
111	rs12660547	A	1853:625, 2296:962	0.75, 0.71	1.2 (1.1 - 1.4)	3.0E-04
112	rs12661926	A	1852:626, 2293:963	0.75, 0.70	1.2 (1.1 - 1.4)	3.0E-04
113	rs7773257	A	2149:329, 2734:520	0.87, 0.84	1.2 (1.07 - 1.4)	0.0043
114	rs6920846	A	1696:782, 2071:1183	0.68, 0.64	1.2 (1.1 - 1.4)	2.0E-04
115	rs4896318	G	1153:523, 948:560	0.69, 0.63	1.3 (1.1 - 1.5)	4.0E-04

Table 6.2: SNPs with Allelic P-Value < 0.005 from Haploview [19]. *SNP number refers to order in Figure 6.1, containing 115 SNPs passing QC in the initial 500-kb region. A – Allele; Ctrl – Control. Freq – Frequencies; OR – Odds ratio; CI – Confidence interval.

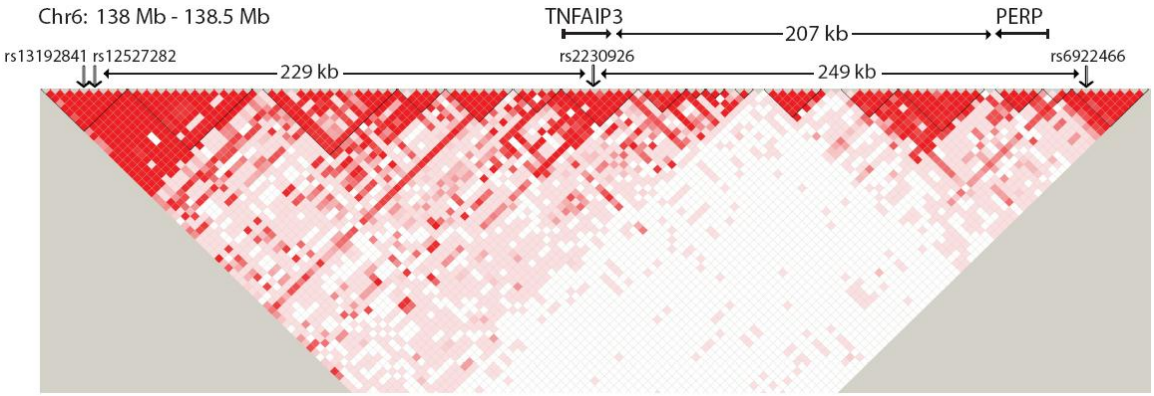


Figure 6.1: TNFAIP3 Region Showing D' for Genotypes of All Study Subjects and Location of Independently Associated SNPs. SNPs shown are those passing QC in the initial 500-kb region. Independent SNPs based on conditional analysis (Table 6.3) are indicated by RS number. SNPs rs13192841 and rs12527282 are collinear; one allele determines the other in > 99% of subjects. D' plot, generated in Haploview [19], indicates D' between pairs of SNPs; deeper red indicates higher D'.

Figure 6.1 also shows the LD pattern among the study genotypes. With multiple blocks of high LD in the region, it is clear that the 21 signals are not all independent. Therefore, we performed conditional analysis, starting with the top SNP, to test the additional candidate SNPs for independence, i.e. for significance when conditioning on the values of the previously-confirmed top SNPs. We first confirmed the independence of rs2230926, with p=0.0014 conditional on rs13192841. Then we tested all other candidate SNPs (with allelic $p < 0.005$) conditional on rs13192841 and rs2230926 (Table 6.3). The most significant SNP was rs6922466, p=0.00037. Next we tested all candidate SNPs conditioning on all 3 independent SNPs and there was not strong evidence for additional independent signals (all $p \geq 0.027$ in 17 tests). SNP rs12527282 was collinear with rs13192841 in conditional analysis; one allele determined the other in > 99% of estimated haplotypes. Finally, we conditioned on rs2230926 within its LD block using all SNPs passing QC, with no additional significant SNPs (all p>0.15, data not shown). As seen in Figure 6.1, the final 3 SNPs are in different LD blocks; each pairwise r^2 is < 0.01.

We further confirmed the three independent signals with multivariate logistic regression using an additive model (Table 6.4). This shows protective effects of the rs13192841 minor allele with an

OR of 0.72 (95% CI 0.62-0.83, $p=7.9x10^{-6}$) and the rs6922466 minor allele with an OR of 0.76 (95% CI 0.65 – 0.88, p=0.00039). In contrast, the minor allele of rs2230926 was associated with an increased risk of SLE with an OR of 1.88 (95% CI 1.27 – 2.79, p=0.0016).

Lastly, we performed stratified analyses of allelic tests to ensure that the associations were not explained by substructure within the European population. We stratified by a) whether or not subjects had ≥ 90% Northern European ancestry, and b) whether or not subjects were in a genetically homogeneous subset determined by principal components analysis (PCA). Overall, results of these stratified analyses (Table 6.5) were consistent with the results summarized above. For rs13192841 and rs6922466, the largest magnitudes of effect (lowest OR for protective SNPs), 0.67 (95% CI 0.55-0.81, $p=2.9x10^{-5}$) and 0.72 (95% CI 0.60-0.88, p=0.0008) respectively, were in the homogeneous subset of subjects. For the infrequent exonic SNP, rs2230926, the homogeneous subset association was OR=1.53 (95% CI 0.84-2.96, p=0.15); given the number of subjects in this subset, we had only 65% power to detect an OR of 1.53. Combining the homogeneous and non-homogeneous strata using the Mantel-Haenszel method produced OR=1.87 (95% CI 1.26-2.77, p=0.0013) with p=0.34 for the heterogeneity of the stratum-specific associations. We conclude that, while some signal from rs2230926 may be due to intra-European population substructure, there is strong evidence for a signal remaining after controlling for this.

SNP*	SNP	Location	p-value conditional on rs2230926 and rs13192841	p-value conditional on rs2230926, rs13192841, and rs6922466
2	rs6933404	138000928	0.025	0.086
4	rs600469	138003365	0.74	0.53
5	**rs13192841**	138008907	N/A	N/A
6	**rs12527282**	138008945	(collinear)	(collinear)
8	rs2327832	138014761	0.013	0.054
10	rs686851	138021664	0.79	0.47
11	rs1002658	138023277	0.31	0.59
12	rs525977	138027345	0.79	0.47
13	rs6904167	138029601	0.72	0.027
17	rs636393	138049223	0.37	0.77
18	rs602414	138053358	0.56	0.94
58	**rs2230926**	138237759	N/A	N/A
105	rs2484066	138317462	0.048	0.77
106	rs9494941	138473046	0.0019	0.15
108	rs1931867	138482531	0.0035	0.16
110	**rs6922466**	138486623	**0.00037**	N/A
111	rs12660547	138489755	0.00079	0.13
112	rs12661926	138489803	0.00078	0.13
113	rs7773257	138491248	0.022	0.72
114	rs6920846	138491762	0.0026	0.33
115	rs4896318	138492967	0.0089	0.20

Conditioned p-values obtained from Whap [20]. SNPs rs13192841 and rs12527282 are collinear, i.e. one allele determines the other in > 99% of haplotypes. *SNP number refers to order in Figure 6.1, containing 115 SNPs passing QC in the initial 500-kb region.

Table 6.3: Conditional Tests for All SNPs with Single-Marker Allelic P < 0.005.

		Minor allele		**Risk allele**	
	p-value	**OR**	**95% CI**	**OR**	**95% CI**
rs13192841	7.9e-6	0.72	0.62 – 0.83	1.39	1.20 – 1.61
rs2230926	0.0016	1.88	1.27 – 2.79	1.88	1.27 - 2.79
rs6922466	0.00039	0.76	0.65 – 0.88	1.32	1.13 – 1.54

Table 6.4: Multivariate Logistic Regression for rs13192841, rs2230926, and rs6922466 Using Additive Model. Interaction terms were insignificant by log ratio testing (not shown). OR – Odds ratio; CI – Confidence interval.

	rs13192841		
Subgroup (n=called genotypes)	**p-value**	**OR**	**heterogeneity p-value†**
All combined raw (n=2731)	6.10E-08	0.71 (0.63 - 0.81)	-
North European* > 90% (n=1456)	0.00069	0.75 (0.63 - 0.89)	0.46
North European* < 90% (n=1273)	3.2E-05	0.68 (0.57 - 0.82)	
Strata MH† combined	1.10E-07	0.72 (0.63 - 0.81)	
Homogeneous** subset (n=1191)	2.90E-05	0.67 (0.55 - 0.81)	0.067
Not in homogeneous** subset (n=1540)	0.065	0.85 (0.71 - 1.01)	
Strata MH† combined	0.000032	0.76 (0.67 - 0.87)	
	rs2230926		
Subgroup	**p-value**	**OR**	**heterogeneity p-value†**
All combined raw (n=1987)	0.00025	2.01 (1.36 - 3.03)	-
North European* > 90% (n=923)	0.025	2.07 (1.07 - 4.37)	0.88
North European* < 90% (n=1063)	0.0073	1.94 (1.16 - 3.30)	
Strata MH† combined	0.00048	1.99 (1.34 - 2.94)	
Homogeneous subset** (n=959)	0.15	1.53 (0.84 - 2.96)	0.34
Not in homogeneous** subset (n=1028)	0.0021	2.23 (1.29 - 3.94)	
Strata MH† combined	0.0013	1.87 (1.26 - 2.77)	
	rs6922466		
Subgroup	**p-value**	**OR**	**heterogeneity p-value†**
All combined raw (n=2828)	0.00012	0.78 (0.69 - 0.89)	-
North European* > 90% (n=1502)	0.0016	0.76 (0.64 - 0.90)	0.52
North European* < 90% (n=1324)	0.035	0.82 (0.68 - 0.99)	
Strata MH† combined	0.00018	0.79 (0.70 - 0.89)	
Homogeneous subset** (n=1233)	0.0008	0.72 (0.60 - 0.88)	0.15
Not in homogeneous** subset (n=1595)	0.13	0.87 (0.73 - 1.05)	
Strata MH† combined	0.00076	0.80 (0.70 - 0.91)	

Table 6.5: Associations between TNFAIP3 SNPs and SLE by Ancestry Strata and Combined Using Allelic Model.

*based on STRUCTURE[18] analysis and 1,409 EUROSTRUCTURE AIMs [22]

**based on 4 principal components from EIGENSTRAT [23] analysis with 550K data [16]

†Mantel-Haenszel combined odds ratios (OR), p-values, and test of heterogeneity of the stratum-specific associations.

The three independently associated SNPs include one coding and two non-coding polymorphisms. The coding SNP, rs2230926, is a non-synonymous variant resulting in a phenylalanine-to-cysteine change at residue 127 of the A20 protein. To begin to test the biological impact of this SNP, we compared the ability of human A20 proteins encoded by the major (127F) and minor (127C, risk) allele cDNAs to inhibit TNF induced NFκB signaling. These experiments revealed that the minor 127C protein is comparably stable to the 127F protein. Importantly, the 127C A20 protein is modestly, but consistently, less effective at inhibiting TNF induced NFκB activity when similar amounts of the two proteins are expressed (Figure 2). This reduced anti-inflammatory activity of A20 may allow excessive cellular responses to TNF. In addition, as A20 is essential for restricting cellular responses triggered by Toll-like receptors (TLRs), NOD2, and potentially other pro-inflammatory stimuli, it is likely that a hypomorphic A20 protein may contribute to multiple facets of excessive inflammation and autoimmunity in humans bearing this polymorphism [4, 5].

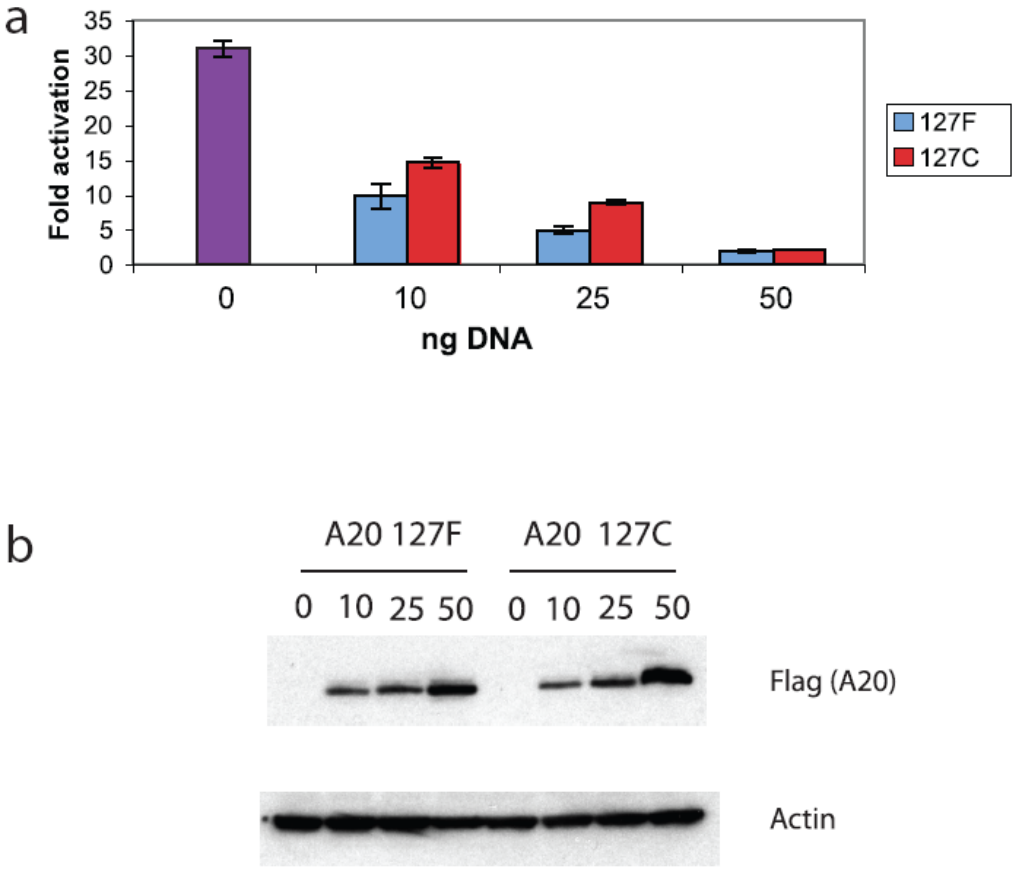


Figure 6.2: Decreased NFκB Inhibition by rs2230926, Phe127Cys. Cells were transfected with varying amounts of TNFAIP3 constructs bearing either 127F or 127C alleles. NFκB activity was measured after stimulation with TNF. (a) Cells bearing the minor Cys allele had approximately 5-fold less inhibition of NFκB levels. Error bars represent the standard deviation; n=3. As shown in (b), similar levels of protein were present for the two constructs.

6.5. Discussion

Our findings show that three independent SNPs in the TNFAIP3 region are associated with SLE. These polymorphisms may cause reduced expression or activity of A20's anti-inflammatory activity, predisposing patients to develop SLE. Considered together with recent studies correlating TNFAIP3 SNPs with RA [6, 7], it is apparent that TNFAIP3 is a potent regulator of susceptibility to autoimmunity in humans. In addition, Graham et al. also observe association to TNFAIP3 in their genome wide association study of SLE [24].We identify independent SNPs in the same two LD blocks as Graham et al., plus a third LD block. Future work could attempt to

clarify the precise location of the effects seen in these LD blocks through fine mapping and additional functional experiments as well as investigating association with specific subphenotypes. Since we have limited our study to people of European descent, future studies including other ethnic groups are necessary, especially since SLE affects people of non-European ancestry at an increased frequency compared to Caucasians. It is also important to note that our region of interest covers not only TNFAIP3, but also the PERP gene. A missense SNP in the PERP gene, rs648802, was not genotyped in our panel, but a near perfect proxy based on the HapMap CEU data, rs563495 (r^2=0.966), was genotyped in our cohort and did not have an allelic p-value meeting our criteria for significance. Given the recently demonstrated association of human TNFAIP3 SNPs with RA and prior functional studies of Tnfaip3 deficient mice, our current genetic and functional experiments support the notion that TNFAIP3 is a causative gene associated with SLE as well as RA. Hence, TNFAIP3 may be an important determinant for multiple autoimmune diseases.

6.6. References

1. Crow, M.K., *Collaboration, genetic associations, and lupus erythematosus.* N Engl J Med, 2008. **358**(9): p. 956-61.

2. Lee, E.G., et al., *Failure to regulate TNF-induced NF-kappaB and cell death responses in A20-deficient mice.* Science, 2000. **289**(5488): p. 2350-4.

3. Wertz, I.E., et al., *De-ubiquitination and ubiquitin ligase domains of A20 downregulate NF-kappaB signalling.* Nature, 2004. **430**(7000): p. 694-9.

4. Boone, D.L., et al., *The ubiquitin-modifying enzyme A20 is required for termination of Toll-like receptor responses.* Nat Immunol, 2004. **5**(10): p. 1052-60.

5. Hitotsumatsu, O., et al., *The ubiquitin editing enzyme A20 restricts NOD2 triggered signals.* Immunity, 2008. **28**: p. 381-390.

6. Plenge, R.M., et al., *Two independent alleles at 6q23 associated with risk of rheumatoid arthritis.* Nat Genet, 2007. **39**(12): p. 1477-82.

7. Thomson, W., et al., *Rheumatoid arthritis association at 6q23.* Nat Genet, 2007. **39**(12): p. 1431-3.

8. Thorburn, C.M., et al., *Association of PDCD1 genetic variation with risk and clinical manifestations of systemic lupus erythematosus in a multiethnic cohort.* Genes Immun, 2007. **8**(4): p. 279-287.

9. Bauer, J.W., et al., *Elevated Serum Levels of Interferon-Regulated Chemokines Are Biomarkers for Active Human Systemic Lupus Erythematosus.* PLoS Med, 2006. **3**(12): p. e491.

10. Petri, M., *Hopkins Lupus Cohort. 1999 update.* Rheum Dis Clin North Am, 2000. **26**(2): p. 199-213, v.

11. Criswell, L.A., et al., *Analysis of families in the multiple autoimmune disease genetics consortium (MADGC) collection: the PTPN22 620W allele associates with multiple autoimmune phenotypes.* Am J Hum Genet, 2005. **76**(4): p. 561-71.

12. Demirci, F.Y.K., et al., *Association of a common interferon regulatory factor 5 (IRF5) variant with increased risk of systemic lupus erythematosus (SLE).* Ann Hum Genet 2006. **71**: p. 308-311.

13. Mitchell, M.K., et al., *The New York Cancer Project: rationale, organization, design, and baseline characteristics.* J Urban Health, 2004. **81**(2): p. 301-10.

14. Hochberg, M.C., *Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus.* Arthritis Rheum, 1997. **40**(9): p. 1725.

15. Hom, G., et al., *Association of Systemic Lupus Erythematosus with C8orf13-BLK and ITGAM-ITGAX.* N Engl J Med, 2008.

16. Taylor, K.E., et al., *Specificity of the STAT4 genetic association for severe disease manifestations of systemic lupus erythematosus.* PLoS Genet, 2008. **4**(5): p. e1000084.

17. Purcell S, Neale B, and T.L. Todd-Brown K, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC, *PLINK: a toolset for whole-genome association and population-based linkage analysis.* American Journal of Human Genetics, 2007. **81**.

18. Pritchard, J.K., M. Stephens, and P. Donnelly, *Inference of population structure using multilocus genotype data.* Genetics, 2000. **155**(2): p. 945-59.

19. Barrett, J.C., et al., *Haploview: analysis and visualization of LD and haplotype maps.* Bioinformatics, 2005. **21**(2): p. 263-5.

20. Purcell, S., M. Daly, and P. Sham, *WHAP: haplotype-based association analysis.* Bioinformatics, 2007. **23**(2): p. 255-256.

21. de Bakker, P.I., et al., *Efficiency and power in genetic association studies.* Nat Genet, 2005. **37**(11): p. 1217-23.

22. Seldin, M.F., et al., *European population substructure: clustering of northern and southern populations.* PLoS Genet, 2006. **2**(9): p. e143.

23. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nat Genet, 2006. **38**(8): p. 904-9.

24. Graham, R.R., et al., *Genetic variants near TNFAIP3 on 6q23 are associated with systemic lupus erythematosus.* Nat Genet, 2008. **40**(9): p. 1059-61.

APPENDIX

TABLE OF PRIMERS FOR KIDNEY TRANSPLANTATION RESEQUENCING PROJECT

Gene	Primer Pair	Chromosome:Positions	Primer Sequence	Primer Sequence
CD79B	362324	17:59355607-59363052	AGTGCCCAAGTTCCCAGGAATGTTCTATACGG	GCATGAGGAGCAAACCAGAGTTAGCAACCA
CD79B	362325	17:59362545-59370342	CATGCCCTCTATGTTCCCGGAGCTACTTC	GTGTGAACGCATGTGGTAAGGTGTGTACTCAG
SELL	63348	1:167918249-167927006	CCTATTTGCACTGTAGCCTCGCCGTCTGT	AGTTCCTGCCAACACTACGTCTGTAGGGTCAC
SELL	63347	1:167926980-167932656	AGGATGAGATGGCAGATTGGGAAGGGATAAAC	AGACGGCGAGGCTACAGTGCAAATAGGACT
SELL	63346	1:167932633-167942227	GGGCCATGACTCCTTCCTAATACTTTACCCTC	CTTCCCAATCTGCCATCTCATCCTACCT
SELL	63345	1:167942206-167954009	CGTAGGCTAGGGAAGGGTTGGGTAGTGGTC	TATTAGGAAGGAGTCATGGCCCAGGGATCAAG
ANXA1	321733	9:74947265-74958921	TAAACACAACACTGATAACGGCGTTTC	AGCAATTAAGACTACACAATGGCATGACTTAC
ANXA1	321734	9:74958889-74965563	TGTAAGTCATGCCATTGTGTAGTCTTAATTGC	GAGTATCTTCATCAGTTCCAAGGCCCTAAAT
ANXA1	321735	9:74965534-74976907	TTTAGGGCCTTGGAACTGATGAAGATACTCTA	GTGACTATAACACATGATACGCGGGACA
VCAM1	393219	1:100953517-100963153	TTTCTGTTACATAAAGGCAAATGGAGTCATAG	GCAATCAGATAGGTCATCAGTGAAACTAAGG
VCAM1	393220	1:100963128-100972864	GTTTCACTGATGACCTATCTGATTGCCATATC	TAGATTCACATTCATATACTCCCGCATCCTT
VCAM1	393221	1:100971403-100977515	AATTATTAGTTCGCAGGTTGAAGCCATACGGA	TGTACACAGGTGTTCATTTGTTGCCCATTAAC
BHMT	349347	5:78444708-78451780	GGTTACAGATTCTAAATGCAGCATCGT	GTTGAACAGTTTGCCATAACAATAACACTAAT
BHMT	349348	5:78451758-78461210	ATTGTTATGGCAAACTGTTCAACTTGGTAAAG	CTCACATGGCGATTACAACTTGCTAACTT
BHMT	349349	5:78461163-78465844	AGGGCAAGGTTAGATTAGCAAGTTAGCAAG	TATCAAGGAGGCTATTTGGGAGAACAGTCTAT
SLA	294894	8:134116496-134125008	TACCTTCTTCCCGTTTCCTTGGTTATCTATTG	GGTGAATATGGGCTCCTAGCAGGTGATTAT
SLA	294895	8:134124634-134135420	TGGTCAAGCAGCGGCAGTTAGATTTCGGTATT	CCCAGTGACTTTGTCCCGTTTGGCTAATGGT
SLA	294902	8:134135382-134145881	TTATCATCACCATTAGCCAAACGGGACA	GAGTATCTGCCCTTGGGAATGCTTGTAAGATT
STAT1	379829	2:191535091-191545652	CTTCTGGCAACTCTTCATTTCTTCTTATCCTT	GGTTGAATCAGTCTCTAAATTGGGTTGACTTC
STAT1	379830	2:191545603-191556406	AGGACAGCAAACCATTGTGAAGTCAACC	GTAAGCTAGGGCAGTCAGTGATTCCATACCAC
STAT1	379831	2:191556375-191567026	GTGGTATGGAATCACTGACTGCCCTAGCTTAC	ACAGATGTTATCTCCCTTCGTGGCCTACAG
STAT1	379832	2:191566996-191574970	ACTGTAGGCCACGAAGGGAGATAACATC	CTGTATAGGGCACTGTGTAAGAAAGGTTGACA
STAT1	379833	2:191574941-191586475	TCAACCTTTCTTACACAGTGCCCTATACAGC	TCAGATGAATGAGTACATGACTTCGGAATCTT
MYCBP	387368	1:39100795-39105197	TGAGATAGGCAAGATGTGAAGTCGATATT	CCGAAATGAAAGAGAAGTATGAAGCTATTGTA
MYCBP	387369	1:39105177-39110718	CATACTTCTCTTTCATTTCGGCCAGTTCTAGG	CTCGAGGGCATATCAAGGTGGTTACATT
MYCBP	387370	1:39107822-39113439	TGTAATCAGACAAGAACTCTATACCCAATCAT	CAGCCAAATCCTGACAATCATAGTGTA
SLC25A24	393958	1:108478825-108482513	CATGATTCAAACTGGTGCCAAACATCTAAG	TAGTACACCTTGGGCATCTACAGCTCTATTCC

Gene	Primer Pair	Chromosome:Positions	Primer Sequence	Primer Sequence
KIRREL	396592	1:156324039-156334647	CACAGATGACCCGATAGCCAAAGGATAGC	AGCCGATACACTGCAACCTCAAATTCCTAGAT
DDX17	5703	22:37208348-37214953	AGCCGGCCATTAAGAGATGATACCAAG	GCCCAATTCCCAGCAGTAGAGAGAATACTAAA
DDX17	5704	22:37214933-37219987	CTCTACTGCTGGGAATTGGGCTGAATAACCTC	GGATGCGCAGAGATGGGTGAGTATAGCAC
SIGLEC10	190170	19:56604023-56612217	TTCCCACGGCTACTCTTGAACTCTTTATGGAC	TTTCGGGTGGAGAGAGGAAGCTATGTGA
CD4	96858	12:6771791-6775567	TCTCAGTCTCTCTTTGCCTCACTTTGGATCTA	GAGAGAGGACTTGGTATGGCTTTGGCTAACTT
CD4	96859	12:6788488-6794936	CTTCCCAACTTCCTGATTTCTAACAATAACAG	GCACTAAACTGGCACACTACCTTATGTCTACC
IFNAR1	1879	21:33625589-33637513	CCAGGGCTGAGCTGATGTAGTATCCTATG	CCAGGAGAGATGTGTATCACTATTGCCTTATC
IFNAR1	1880	21:33637471-33646191	TAGAAGCTGAAGATAAGGCAATAGTGATACAC	AGCTATCGCCAGAATAATGTCTAACAAG
IFNAR1	1881	21:33646179-33655433	TCTGGCGATAGCTGACTGATACAGATAGCAG	AAGGCGAAGGTGGTAGTGGTAGAGGACTAGAT
CD3D	244806	11:117712783-117718584	CAAGGTATTATGAATGTGTGGAACGCGAGAA	GGCTGATAGTTCGGTGACCTGGCTTTATCTAC
KLF9	88657	9:72186264-72198143	TGGGCGCTGAAGTTCTCTGATATTTGTCAATC	GGCAAGTACCCGGTCTCTGTCTTCTCTATGGA
KLF9	88658	9:72202949-72207466	TGATACAACAGAGACAGTGTAGGGCTTAGGTG	TTCAACCCGAAAGATAGGCCATAAGTGA
IGFBP4	360453	17:35856607-35867448	GCCCTCCCACACCTCCTAGTCATCATAGAAG	CATCATGGCCAACTGGTAGGTTCCTAAGTCTC
IGFBP4	360454	17:35867274-35872084	CCATGAAGTCACCGGGATGAACCTATC	CCTAGAGGATGGCATCAGGTAAGGTAGCAAAC
IL10RA	76672	11:117365680-117369191	GTTCACACTGCCAACTGTCAGAGTCACTAAAT	AACAGGGCAGGACTTTGGAGAATGTTAGATA
IL10RA	76671	11:117369170-117378347	CAGGTCTGAAATGGCGTGATATTAGTCAAGTT	TCTGACAGTTGGCAGTGTGAACCTAGAGA
ACVR1B	17_ACVR1B	12:50653543-50663643	TGAGTTTCAGAAGTAAATGTAAAAGCAGAAAGA	CTGGATCTACCACCACCAAACAGGAAAG
ACVR1B	18_ACVR1B	12:50663541-50674612	AACATTACAGATGAGACAGGTGCTGTGTTT	GTGAGAGGGGCTGCACTTCTGGTTCTT
ADD3	10_ADD3	10:111844922-111855071	GCAAGGTGGTCAGTAGCTTCTCAGGTT	TTGAGAAAAGGGGGTAAATTAGAACTTAGG
ADD3	11_ADD3	10:111855004-111865193	TGCAGTAGAGAAAATCCACAAGTTCTCTTT	TAGTTCTGATCAGGTCCTCCTCCCCATTTT
ADD3	12_ADD3	10:111865329-111875324	AAAGGGAACAGATGAACTTGTTAAAGGAGA	CTTGTGTCAAGTGAAAATATGATGCTGAAA
ADD3	13_ADD3	10:111875009-111885087	TGGGGTAAATGAGAAGAATTATGTAAGCTC	AAAGTTATGGAAAAGCATAGTTCACAGGTT
AIF1	AIF1	6:31689974-31693566	ACTCCCAGCTACAGAAAAGGAAATATGTTG	TCTTTGCTTTATTTAATAGCACCAGTTCTTTC
AQP1	2_AQP1	7:30926981-30932650	GAGGGGATGTGAGTCATTCCCTGTAAC	ATACCCAATTCCCATAGCTTGGCTGGCACT
ARNT	4_ARNT	1:149076323-149086594	GGGTTGTTGTGCCTGAGAAAAGCACTG	CCCTCAGTTTGACCCACCTCATCTCTC
ARNT	5_ARNT	1:149066215-149076415	TGAAAGGTAAGAGGATAAAGCCAGTGAATA	GCTGGCTTCTAAAACATTTTTCTTTCTCAG
ARNT	6_ARNT	1:149056084-149066084	GACCAGGCCAGTTAGATGACCATAAAAT	AGTGGAAAAACACTGGAAGATTACTGACTG
ARNT	7_ARNT	1:149047816-149056623	TTTAATTTGTTTTAAGGTTCTTTCTCCAGATTC	TGCGGTAATGAACATAAAAGGCTATAAAAGA
C1QB	C1QB	1:22851255-22861249	TGTGACATTTATCACAGAAGCAATGAAG	AAGTTCATGGGCTGGGACTGTATCTGGT
C3	3_C3	19:6641194-6652194	AGCTAGGACCTGATAAGTCCCCTTCAG	GGAAATAGTCTTAGCTCTGTATCCCAAGTTTTC
CA12	6_CA12	15:61402059-61412054	CCTCAGTGTCTTCTCTGCAGAATGATTTAC	TCTAGCAGAGTACAGACCATCAGAAATGTT
CASP1	7_CASP1	11:104401574-104412178	TTAGCTAAATAATCTGCAAATAATCATCCTTGA	GGAAAGAATTTGGTTAAAGACATGCAAGTTA

Gene	Primer Pair	Chromosome:Positions	Primer Sequence	Primer Sequence
CCL4	CCL4	17:31454321-31457965	TGATTCTTACTGGTTAATTTGTCCTGTTCTTT	GGATCAGGCTTTATTGTGCATGTAATTCTA
CCL5	CCL5	17:31222112-31232086	AAAGCAAGAAATTCCCACAAGAGGACT	GTTTACATAAAGGGACAAGCTTGGGAAAGT
CCL19	CCL19	9:34678553-34682249	TGATTAATGAATGGAAAGAAGCTAAGTCAGA	CTGTGCTTCCCTGGCTACTCCTTCTTC
CCL21	CCL21	9:34698059-34700454	CAATTTAATAATCAGATTTGCTGGAAGGAG	GATCTGTTCCCTTCTTTCTCACATTGTCT
CCR2	CCR2	3:46370161-46378414	AATCCATGATACCACTGAAACCAGCAC	GGATCATCATCTTCCACCTGCTTAACTC
CCR3	1_CCR3	3:46226320-46232835	GTTACCACAATGAGGAACAGCTTTTGTAGT	AAGCTCACCCAATGTGTGCAGTGTGTA
CD2	1_CD2	1:117097941-117108054	TCTCTATGCTTCTTTGTTTTGGAATAAGTTTT	ACGGCCAGCACCCTCTCACAGAGTTTC
CD2	2_CD2	1:117107537-117114223	CCAAGTTCCAGTTTGTTCTTCAGATGAC	CAATTGTTACCTAGCTGGATTTTGTGTTTG
CD3D	CD3D	11:117714329-117719495	CAATAGGTGGAACAATTCCGAGGACAG	CTGTCTCAAACTCCTGCTTTCCAGGTAGAT
CD4	1_CD4	12:6768466-6780036	CCTGGCCAGAGACGCCTAGAGGAACAG	GGATAAGAATGCCACTGAGTTTCTGTCAAC
CD4	3_CD4	12:6788137-6801137	GGCCATTATCATTTTGTCTAACATTGTATCC	GTCTCTGCCCCTCTCACTCTCCAGGAC
CD14	CD14	5:139990511-139994059	ATTCAACATAGTCCCATCCCTTCCTTCCTA	TGAGTCTGGTTCTGGTAATGTCTGTAACTC
CD36	1_CD36	7:80068898-80079410	CATTCTACAACCCTTCAAAAAGATTCAAAA	ATTTCAGGTGTGCACGTTTAATTACTGAAG
CD36	5_CD36	7:80108854-80118853	AATGTCATTTAGGTCTGGCAAGTGTCTATG	TGATTTCAATTTATTTTGTCTGAAGGCATTT
CD36	6_CD36	7:80118823-80128990	AAATGCCTTCAGACAAAATAAATTGAAATC	CCCATTTTGCAGCAACAAATTGCAAAG
CD36	7_CD36	7:80130102-80142224	AGTTTTGGCAGGATCTGGCAGTAATTTT	AATGGATGCTGATGAATCCAGGCTATT
CD47	2_CD47	3:109272954-109283390	TCAAAGAGGGCCCATATCATTACATTAAAA	ATTTTTCTTATGTTTCACGGGAGGAGGT
CD47	3_CD47	3:109263106-109273266	CCAGCCACTCAACACATGAATACATACTTA	CATACATGTTATTGTTTCTGCTGCTAGCTC
CD47	4_CD47	3:109253106-109263172	CCATTCTTTCAGAGAAGGTATGAGATACTTGA	CATTTTTAAGCCTGAAATTGTCATTCTGTG
CD47	5_CD47	3:109243665-109253636	ACTCTTTCATTAACATCTCTGGCACCTTTA	GCCATTAAATGCTTTCTGAACTCCATTAG
CD163	1_CD163	12:7538036-7548095	GGTGTACTTCATCCCATTAGAAAAGAAATG	AATTGGATAGTAGTGGGGGAAGAGTTTACA
CD163	2_CD163	12:7521773-7532094	CAGAAAAGCCAAGACAAACAAGTGTAGG	TTCATGGTTCTCAGTTTACAACACTAGAATTT
CD44	4_CD44	11:35154623-35165454	TGTCCTAAACTGAACTTATTACTGTCTCCAAAT	CCAACTGTCCTTACTTAGTCTTAGGGAAATGA
CD44	6_CD44	11:35166419-35176537	ATGGCTACGACATGAGATGTGCTGTCTC	AACATTCTATTTCTTCTTCCATAATCTCATCAA
CD44	7_CD44	11:35176449-35186508	TGAGAAGTAGTTAATGTGAAAAATGGGTGA	AAAATCAGGTTGAGAGACCTCCACAAG
CD44	8_CD44	11:35186447-35196777	CCCAGAGTGTGAAACTGTCTTCATTGT	GCATCATAGGAGCCAACATAAACACTATAA
CD44	9_CD44	11:35197324-35208386	ATAAATGGCTTCTCAGTGATTCAGAATGTG	GACAGGATGGAAAACCTTTGGACAGTG
CD48	1_CD48	1:158940436-158948792	GACTACAAGGGTGGGAAATCCATTCTTT	AAAAATGCTTTCCTGTAAAAGTGGCTCAT
CD48	3_CD48	1:158914186-158922899	GCTTTTAATCACCCTTTGGCTTACCCTAGT	TGTGGAAAAGACAAAACCCTACCTTTCTTA
CD52	CD52	1:26515986-26520613	TGCTACTTTCTCTACCAAAATCACAAAATTC	ATCCATGAGAAGGGAGGAAAGAAGTAGAG
CD53	1_CD53	1:111216295-111225356	TTAGCTAGTGATACATTTGGGTTGTCACAT	GATTACAGAAGGCCACAACAATCTGTG
CD53	2_CD53	1:111232971-111243968	CTGTAGCACAGCTCTAGGGTACAGTGAATC	TCAGCTAAGTGACTGGTTGAAATCAGGATA

Gene	Primer Pair	Chromosome:Positions	Primer Sequence	Primer Sequence
CD86	1_CD86	3:123256368-123266466	CGCTTTCATCAATGTAATCCTTGGCTAT	CTGATTCATCTTCATCTTTCCAACATTAAA
CD86	6_CD86	3:123305911-123314327	AAACCCAACAAACATAACCCACAAACTTAC	TTCAGATGTCTTCAGATTAACACAAAATGAAT
CD86	7_CD86	3:123318839-123323483	CACATACACACATAAAAGAGAACTTCCAGTGA	ATTCTGCCTAATGACCTCTTTCAACCTTCT
CDH2	2_CDH2	18:23991646-24001787	TGCTACCTGATATGATATTTTAGGGAACTCA	CCTGCTGCCATAATAAACTTTCTCCTAAA
CDH2	17_CDH2	18:23836336-23847992	TGTTTCATGCTACTGTATTTTTATCCTTTAACC	GCCTTCAACATTGCATACATTTTCAACTAA
CDH2	19_CDH2	18:23816345-23827666	CAGAACAACTGTTATGTCTTCTTGAGGAATTAG	CAAGCCAGGAACTATGACAGACACTATAAA
CDH2	22_CDH2	18:23791669-23801743	TCGCCCTTAAGACATTTTAAATCATAGACC	TTTCCTGCTCAGCATCCTCTTCATATTTA
CFB	CFB	6:32020515-32028852	AAAAACCATGTTCCCCAACTTGACAGAT	GACACACACTTTGAGAAAGAGGGAGTAGC
CLEC2B	1_CLEC2B	12:9905538-9914310	CTTGACTGTTTTCCTTTTTGGAACAAAACT	CAGTGAATGCACTATACTAAACATCCGCTAA
CLEC2B	2_CLEC2B	12:9896037-9902771	TGATATATCTTCCTGAAGGCACCAGAATTA	CTTTGGAACCCAGCTTGACACCAAGTAG
COL4A1	7_COL4A1	13:109687926-109698048	AATGGAATTCACTGTAATCTTGAGAAGAGC	GGCTGGTTGGTTTCTTCCACTGTATAATAG
COL4A1	10_COL4A1	13:109657981-109668065	AGTAGTCATCTCCGGCTTTGGCGTATGAT	TCTGCAGAAAATCAAATTTCAATAGGAAGA
COL4A1	11_COL4A1	13:109647932-109658048	TGAAATTCTCCTCCTGGGTTTAGATACTTG	ATCTGTTCTTTCTCCTACCCAAATTCATGT
COL4A1	12_COL4A1	13:109637990-109648055	GAGTTCTCATGGATGAGGCCACTGTTG	AACTGCTCAACATTCATAGGAACACACA
COL4A1	13_COL4A1	13:109627985-109638057	ACAATGAACCACACATCAATCAGAACCTT	TGGCCAAGCACAGTTTCAACCCCATGT
COL4A1	14_COL4A1	13:109617981-109628044	CCTTCTTCCAGTAGGATTCCTGGTGCT	CCTCAAGTGCAATCAGTTTTCTTTAGAATTA
COL4A1	15_COL4A1	13:109607983-109618088	AAGGAAAATACTCGTTCCTTAGCATTTCTT	GGAAGACATAAAAATCTGGAAATGTAAAGTCCT
COL4A1	16_COL4A1	13:109598419-109608418	GAAGGAAAAGTAACAACTTTTGTGGGTTTC	TTTGAATGTTGGGATCCACAGTGTTATATT
COL4A2	6_COL4A2	13:109807087-109817194	ACTTTGCTTGGCATCGGGCAGTTTATAG	CCTAAGGACATTTTGAAATGACTGTTTCAT
COL4A2	12_COL4A2	13:109875015-109887080	AACCGTAACTGATCATGAGTATGTATTGTGA	TTTGCTGAACACATTTTGCTAGACAGTTC
COL4A2	14_COL4A2	13:109888287-109901348	TGATTTTTACCCATTACCATCCTCAAATTA	TATTTTTGCAGATAATGCAGCCATCTCTT
COL4A2	17_COL4A2	13:109917080-109927156	ATTAACTCGGTTTGCACCCAGGCTACT	CACTAAGGGCATCTGTTTTCTTCTAGTTTTT
COL4A2	18_COL4A2	13:109927086-109937373	GGAAATGCCTGAATATAAGGAGGTCCA	CCCTGGCCTGCTACTAAGGGTTCACAC
COL4A2	19_COL4A2	13:109936921-109947147	AAGGTCTGGGCCCCTTGATGTGTTTCAG	AGGCTGTGCTGATGGACTGACCCTAGAT
COL4A2	20_COL4A2	13:109951938-109961936	ATGGTGTGGAGGGAAAATAGTAGATTTGAA	TGCCTAATCATCCATTCAAAGAAAAATACC
CSF1R	1_CSF1R	5:149463576-149473672	TGCTGTTGAGTTTTTCTATCTGTGTACTTTTT	AAGGGGGTGAGGTCCAAGTACTAAGAGT
CSF1R	3_CSF1R	5:149436044-149446276	AGTGTCCCGATCCCTGCGGAGCTAGTA	CATTTCCTGTGTCAGACCCTGTTCTAAGTA
CSPG2	2_CSPG2	5:82815023-82826153	GTGATTTATGACCCACTTTAAACCTGAATG	CAAAGGAGAAATGACAGTGCTGAAATC
CSPG2	5_CSPG2	5:82842389-82852860	GCCTTTTCTTGTTAATCAATCAGCAAATTA	TTCTGATGGAACAGAAGGTACTAACACTCC
CSPG2	7_CSPG2	5:82862788-82872865	TGAAAAGTTGATATACGGTGTCACTGTGGT	GCTGAAAAGAAGGACCAGCATGTTTTT
CSPG2	8_CSPG2	5:82877038-82887108	GCGTTTTATGCTAAAATACACGCAAACATT	TGTAATTAGTTGGCAGAAATTCACAAGATCA
CSPG2	11_CSPG2	5:82902784-82913281	GGACTCAAGACTAGGAAACTGGTCTCTTATTTA	TGCCCACACTATTTCTATATATTCGCTCAT

Gene	Primer Pair	Chromosome:Positions	Primer Sequence	Primer Sequence
CTNNAL1	2_CTNNAL1	9:110795900-110805896	GAAATCCATGATGTTGGCATCTATTACAAA	ATTTTCTTACAAAAGCCCTCCTGTCAATTC
CTNNAL1	3_CTNNAL1	9:110786071-110796401	TGGCTCTAAGTCTTTCCTTTCCATTCTATTT	TGCTGACTTCTTACTCCACTGTATTTAACAGA
CTNNAL1	4_CTNNAL1	9:110776057-110786173	GCCTGCAGGGTTTTTCCAGTCTCATTG	CCACATAAGAACTTAAAAAGATTTTGTTTCACC
CTNNAL1	5_CTNNAL1	9:110766116-110776368	TGTGATAAACTGAGTGGCTTTCCACAG	AGCCACTTCTCTTCATTTCCCAACAGT
CTNNAL1	6_CTNNAL1	9:110756057-110766272	TGGCAAATCATTTTGATGCTTTGTATTTAT	AACATTTAAGAAAGTAAGGAACACCCTCAAG
CTNNAL1	7_CTNNAL1	9:110744330-110755988	GCCAGAGAGATCATTGGTCATATTTGTTTA	CCTTGCTGTACAAAATGCTTATTAGGGAAC
CTSB	2_CTSB	8:11739378-11748442	ACCCACATAACAGAGAGGTGTCCTGAT	ACTCCTGACCACTTGGTTTCCTTTTGAG
CXCL10	CXCL10	4:77161074-77164681	CTGTGAAATTAAGTTTTGCCACGATTCAT	GATTTGGAGATTAGGCCAAGCTCTGTTAT
CXCL13	CXCL13	4:78745000-78753014	CAGGAAGCCCAAACCTCTGATTTCTTAG	CGAATTAAAGAATGAAGTGATTCCTGATCTCTA
CXCL9	CXCL9	4:77140515-77148282	CTTAGCCACTGTTAATTTGGGTCTCTTACA	AAGGAAGATGCAGCAGAGGTAAGTGGTTAG
CXCR4	CXCR4	2:136587900-136590834	GTTAACTGGATCAGTGGCGGGGTAATG	GGTTATCTACACTGAGGATACTGGATGAGGA
EGF	1_EGF	4:111052953-111063017	GTATCTCTTCATTTGGCCTCAGGGATTT	TTGAAAGACAGTCACTTATCCTCTTGTAGGT
EGF	4_EGF	4:111082922-111093015	AACAATTTACCTGGGTGAGCTAATTTATGA	ACAGGATTAAATTCAGCCATCTTTTTGTG
EGF	5_EGF	4:111099853-111109913	AAGATTTAGCAGTGTCCTCTGTCAAAACAC	ACACAGATGCAATCTTGATGAAAGGAG
EGF	7_EGF	4:111112507-111122328	TTGAAAATTCCAATGGTAATGCTGTTATTT	TCATTGCATAACTTTTGTTTATTTTTCATCA
EGF	8_EGF	4:111123497-111132928	TCTTACAGAATCCATTTCTCCCCTGTTCT	GCCCTATACATTTTTAGGTTTAAGGGGACT
EGF	9_EGF	4:111132891-111143198	TTCATGAAAGTCCCCTTAAACCTAAAAATG	TCCTTCCAAGAAATCACATGTTTTAAACTAAT
EGF	10_EGF	4:111142780-111153419	TGTCTTCTGTGTCTCTCAAAGGATTAATTG	TGTAGTAACAATATCTTGGCTGCAAGAAAAA
EGFR	13_EGFR	7:55173658-55183752	TTTTTCTTTCTTCTTTCTTTTCTGGTTTGA	GGCTTGAAATGTCAGTTCTTTATAGACTTTTGA
EGFR	14_EGFR	7:55186423-55199196	GCGTCATCAGTTTCTCATCATTTCACT	CTAATGTGTGTGCTAATGTCACCGACACC
EGFR	16_EGFR	7:55203622-55213885	TGCCAAATATAGAAAGAGGGGATTTAGTCA	TGAAAATATCCCTTTAACTTGGATAAGTGCT
EGFR	18_EGFR	7:55223665-55233835	ATCCCCTTTTAACCTCAGTGCTGTATTTC	TCATCATTACTGGTTACTGTTTCTTGATGTTT
EGFR	19_EGFR	7:55233541-55243530	ACTGAGTGTGTATCCTGTCTGGAGCATAAT	ATGGAGTCCAAGCTTTGAGTACTGACTGA
EIF5A	EIF5A	17:7150032-7157495	GAGAGAGGGAGGGCAAAGAAAATAAAGTT	GCAGTGCAGTACCTATCTCAGCCACAG
EPS15	4_EPS15	1:51709581-51719727	ACGCTCCCCAAGAAGAAACAAAATCTAAC	CACATGTGATTAACCAAGAGTTTCCTTTCA
EPS15	6_EPS15	1:51698617-51708584	CCAATACTTAATGAATTTCCATGTGACAAA	AGATAAAATGTCAAAACTTTTCTGCCTTCA
EPS15	8_EPS15	1:51677865-51688208	CTTATGCCTTTTCCACACCCTGGAATC	GGAACACGGAAATTAAGGAAGTACCAATAA
EPS15	10_EPS15	1:51658019-51668241	CACCATTTCCTGTGTATTGAAAGTCTGTCT	CAGTCTACATCTCAGTACCAATGGGTAAATC
EPS15	13_EPS15	1:51630657-51638579	GAGCATAATGATCTATACTGCTTCAGGGTTTA	ACAATTAATGATAAATCTTTGCACCCTCCT
EPS15	16_EPS15	1:51594243-51604669	TGAATGTGATTAACCAAAAGAAAACATGAA	GCAAACACACTGAGAGGTATCTGTTGATCT
ERBB3	2_ERBB3	12:54772975-54782538	AGCCAGGAGAACCCAAGAAAGAAGAAG	TCATAAACCAAAATTACACTTTCCTCTGCTA
ERBB4	42_ERBB4	2:212691813-212702049	TTCATCCTCATTGGAATTTGAGATAATTAAA	CATGCATTTGTTATTCATTATCTGAGAGAAGTT

Gene	Primer Pair	Chromosome:Positions	Primer Sequence	Primer Sequence
ERBB4	76_ERBB4	2:212351953-212362122	TGGAAGGTGCTATGTGTATTTGTAGCTAATTT	CTTCGGGATAATCTCCCACATGGTAAC
ERBB4	79_ERBB4	2:212321632-212331632	TCAAAAAGCTGACACATCTCAGTATATTCTTT	CTTTTGCAGAAGCCTCTTTGACCCCTCTGT
ERBB4	84_ERBB4	2:212271961-212282056	CCACTTTAGACTAACCCAGACTTCTAAGGAGA	GCTCTAGCCACTGTTACCCTTTATAGATGA
ERBB4	87_ERBB4	2:212241874-212252059	GGAAGGATCTGCATAGAGTCTTGTAACCTC	AGAACATTTCCCAGATTTCCAGTGTACTTT
ERBB4	91_ERBB4	2:212203101-212212511	AGGCGAAGAAAACTTTTTAAATATGGGAAC	TCAATGCAGATATTTTCTAGATCTGCTGCTA
ERBB4	92_ERBB4	2:212191699-212201074	AACAAAACTAACAGACAAGGAATGCATGTT	CCACCACCACATAGTCTTTCTAACACTCTT
ERBB4	98_ERBB4	2:212131820-212142038	AGGGAGTGGTAAGGCTACTGAGCAATG	TGAATGGCTAACTACATGTAAGATCCAATTTT
ERBB4	112_ERBB4	2:211991979-212002210	GGGTGAAAAGGGATAACTGAAAGTATAACA	GGTAGATGAGTAAAAGGTGTAGAGCAGATTTGA
ERBB4	116_ERBB4	2:211955584-211962496	ATAACCTCACTATAGGGGTTCAGACCATGT	GCCACCCCTGAAAGTATCAGTAGTTTTC
FGF9	2_FGF9	13:21152251-21162250	AGCAAGGGGAAAGAAGGAAGTAAAATAAGA	TATTAACTCAGGGAACCCAGAGAGACTACG
FGF9	4_FGF9	13:21172189-21177646	TATTTTTGTTGTTACTGCCCATGAGTTTTG	CCAATATATAAGCTGGAATATGGCTGGAGA
GAP43	1_GAP43	3:116824262-116834435	AACATTGTTTATTATTTCTGAGCTCAAGTGC	GGGATGTAGCTAGAGTGAAAGTTTAAAGGTG
GAP43	6_GAP43	3:116874321-116884419	ATTTCCCTGCGTCTGTTTCTTTACTGC	CCTTCATCTTGTCTTTTGTCTGCTTACTTT
GAP43	10_GAP43	3:116913849-116923766	TGTTTCCTCATCAGAGTCAGACTTTACAGAA	CCTTAGCAAAACCTAGAATAACTTTTCACTGG
GBP1	1_GBP1	1:89294111-89304316	CATAAAACACAAACAGCTCTTAGAAAACAATC	TGATGAGCACCTAGGACATATCTGGTATAA
GBP1	2_GBP1	1:89289742-89294612	TCAGTTTCTCCAGCATCATTGACTTCTATT	CCAAACTGGGGAGGAAGTAACTTAATTTTG
GNA14	12_GNA14	9:79333168-79341667	AAAAATGCACATATGGCAAATAGGAGAAAT	CTCTTCCTCTCTCTTTCCCGCCTCATTATT
GNA14	22_GNA14	9:79228222-79239316	GCCAATGGGTTGGCTTCCTTTCCTTTG	CAGCATGAGGTGAGAAACACAATTAAAAAT
GZMA	GZMA	5:54433218-54442849	GGCTCCCTTCTAAGGTCACTTGATTTCTAA	AATGAACTACCATGGTTGAAAACGAATTT
HCK	2_HCK	20:30123061-30131619	ATCAGAAGACTTCCCCGCATGAGGCTCT	AGCTTCAGCTGGATGTTTCTTTGTTTT
HCK	4_HCK	20:30133128-30143501	CAGTGAATGCCTGGGCTTTGTCTCTTC	GTTAGAATGGAGACCCATGCTTTGCTTTTT
HCK	5_HCK	20:30143140-30153362	TACTGAAAAATTTGGGTCTCTGGGTCTC	GCAAATAGATATTCAGGAATTGGAAGGACA
HCLS1	1_HCLS1	3:122854239-122863139	CAAGTCTTGACACGCAGTAAAACAGGATAC	AAACCTTAAATGAAAGAGACTTTGGGCACT
HCLS1	3_HCLS1	3:122831925-122839976	CCCCATTCTTAAATGGAAATTAGAATCACA	AAGAAAGACCAGAGCACAAGCCTGAGAC
HHIP	2_HHIP	4:145796079-145806165	CATCTTCTCAGCAGTAGGTTATTGTAGTGTG	AAATAATTTTTCTTTGTGGGCATTTATACATTT
HHIP	9_HHIP	4:145869029-145879247	AGAACATTAGCTTTATCCATCCTAAACTCAAA	CAATAAAAGCAGACCTTTTAGGAAGATTGA
HLA-F	HLA-F	6:29798320-29805133	GTTCCTGATAATTCAGGGGTTACCAAGATT	CACTTGTAAGATGCAGTGAGCACTGATAAA
IFNAR1	2_IFNAR1	21:33628791-33638926	AGGGGACAGTGCAGATTCAGGCAGGTG	TGCAATAGATGGCTAAGATCAAATGAAGAA
IFNAR1	3_IFNAR1	21:33639348-33650230	GCCTTTATCTTCTTGCCAGTTATCTCACTT	AATGTAGGAAACGTTTGTCTCAACTCTTTT
IGFBP4	2_IGFBP4	17:35862190-35868515	CACTAACTCCCACTTCTCTCCCACTCAG	AACCCTCAACACCACTTTGTTCTTTATCTC
IL10RA	2_IL10RA	11:117371339-117378415	GTAGGGATTCGCAGAAACCTAGACACATCT	GGAGAAAGTGACCAACACTACTTTTCTTTC
IL16	5_IL16	15:79301585-79311665	TATACCAGAATCTGACACTGGAGGATGAAG	TTGGATATGCTGAAAAATAAACAACAACAA

Gene	Primer Pair	Chromosome:Positions	Primer Sequence	Primer Sequence
IL16	8_IL16	15:79338992-79349183	GTCTATGAGTGAAAGGATGATGGTGATTCT	CTAGCCACTCTTGTGTTTGTCTCTTCTGAT
IL16	11_IL16	15:79361933-79371933	CTGTAATAACCAGGGGGCCATGTCTGT	AGAATCTTCACAGCTGAGCTTCAGAAAAC
IL16	12_IL16	15:79371187-79381662	CACCCTCAACTGGAAAGAAATAGTGACTTC	TGGGCTCATCAGACTCACCTACACTCT
IL16	13_IL16	15:79381149-79389497	TAAATGAGAAAGGTAAGGCTCAATCAAGGT	CTGGTTGTTGCTGATGCAAAGTCACAT
IL2RG	IL2RG	X:70242982-70249117	AGTTTCTGTAAGTGGCTTCTCCAATCACCT	GACCAGAAACTGCTGATTATCTGAGAAGAG
IL6R	7_IL6R	1:152703566-152709562	TTTTTAAGTAAATGGATTTGGGACTGATGA	TATTCAGAAAGGTTTGGTTGCCCTGTG
IL6ST	2_IL6ST	5:55298393-55308219	GGATACATATTATGGTGAGTTTTGCCATGT	GGGAAACTTAGTAAATAAAGGAATAAGGCTGAC
IL6ST	4_IL6ST	5:55282869-55295992	TCTAGGTCTCATCAAACTCTTTGCAAGTAA	AAATACCAAAAAGGTGAGGAATAACACTTTC
IL7R	1_IL7R	5:35892134-35902710	TGAATTTTTGTCTTCAGATTCTTTTAAAGTGG	TGATGGGTAACGTAATGGATTTTTATAGGT
IL7R	2_IL7R	5:35902682-35912601	CCTATAAAAATCCATTACGTTACCCATCAA	AAATAGCTGAATCATTGGGTCACCTTAAAC
ITGA6	4_ITGA6	2:173038429-173050293	CCCACTATCTCCTAGTTCTGACTGATTTAACTG	AGTGATGAAAGCGCTGGTTAAAGGGAGA
ITGA6	6_ITGA6	2:173052527-173064611	GCAGCTAAGGATGCTCTCTAGTATGTGAAT	GTAATGGGGCTATGAGGGAAATGAGCAC
ITGA6	8_ITGA6	2:173070043-173080042	AAGAAAAAGGTAAAATGGTGTCCATCAGTC	TATGTGAACTGAGGCTAAAAAGGTTCACAG
KLF9	3_KLF9	9:72189066-72198349	TTGTATTTTTGACTTTCCTGTCTCCATTAAA	AAGCCTCTTAAAGTGATTTATCTCCCAAAA
LAPTM5	1_LAPTM5	1:30993729-31003904	GTGAGGAAGTGGGAGAGAACACCAGCTC	GTCTTCAACCCTTCCCAACAGACAATG
LRP2	5_LRP2	2:169878599-169888266	GAAGTAAGGGGCACCTGCAATTGTTATTAG	CGCTAGGATTTGAATGCCACAATACATA
LRP2	9_LRP2	2:169837738-169847811	TGCTTAATAAATGCTGATAACCTTGACTACAA	GAAACCCAAACCATCATTATCTGTTCTGT
LRP2	10_LRP2	2:169829039-169838163	AAAAAGCAGTGAAGTGTGGGTTTTCTTTT	GTAGGTTGCTCAGCATACGAACTTGAA
LRP2	11_LRP2	2:169819440-169826366	CCTACCTTTTAAAATATTCCTCATAGTCAGTGG	TTATTGCTATTTTGGGAATGTTGTGTTGTC
LRP2	12_LRP2	2:169807286-169813364	TTTGTGAATTTCTTTTGGAAAGTTCTTTTG	TGAATTCATCAGAGTCAATACTAAACACAGC
LRP2	13_LRP2	2:169797734-169807801	AGTGTGCCAGTGGGGATAAATGTATTG	CACTTTAGGACAACTCGTGCTGGCTAC
LRP2	14_LRP2	2:169788052-169798043	AACCACCCCACTATATACTTTCCCTAGTCC	CTGGCCACTACTGCATTGTACCTAACACTA
LRP2	15_LRP2	2:169777713-169788000	ACCACATGAACCAAAAGTTTCTAATTATGC	AATTAATTTCTATTGCATGTGTTTCTTTCACA
LRP2	16_LRP2	2:169767745-169777823	GTCTTCAGCCAAAGATGATCAAACAATTT	AAGATGGATGCAGCCACCGTTCACATT
LRP2	17_LRP2	2:169756179-169766735	GCCTTGGAAGGTAAATTCAAAGTGTCTCT	CTATCCATTCATCTCCCTGGCCCTCTGT
LRP2	19_LRP2	2:169737742-169747806	CCCCTTTAGAAAGTACACAGATCACCAAAG	AACAAAGAAAAGATGACATGGGAAATACAC
LRP2	20_LRP2	2:169727649-169737801	CAAGTGCCATTTATTAGATGAGAGATGAGA	CAAGATGCAGGAGGGCTGTGAAACCAG
LRP2	21_LRP2	2:169710256-169722317	ACCACTTGACTGTTATTTGTGGCTGATCT	GGGCTAATATTTTTACTTACCTGCTGATGGT
LRP2	23_LRP2	2:169697730-169707801	GATAAGTGATGCCAATAGAAAATGTGAGTG	ATACCTTCTTACTCCCACTGTTGACCAAAT
LRP2	24_LRP2	2:169692106-169698262	TGTTTAAATCTTGAAAGCATTTTCCCTGAT	AAAGTGTATATTCCTGCCATGGTTCCTGTT
LST1	LST1	6:31661202-31665676	GGGGAGGAAGTAGAAGGTTCTTGAATTTG	CGTTTCCTCCATGACCACCAGGCTGAG
LY96	1_LY96	8:75065129-75070556	ACCGTGCCCAGCTGAGAATTTGTTTTAT	ATTCCTTGGATTTTCTAACCATCTGTAGCA

Gene	Primer Pair	Chromosome:Positions	Primer Sequence	Primer Sequence
LY96	2_LY96	8:75079439-75085940	TCTGGGGAGATATTTTCAATCATCATCAAT	TTCTGAGGAAGGAGATCCTACCATTCAACT
LY96	4_LY96	8:75099072-75103839	TGTGATAAATACTCTCAAAGAAACAAGCAAGT	TTTGAATTAGGTTGGTGTAGGATGACAAAC
MMP7	1_MMP7	11:101896403-101907230	ATCTGCCTCCTGCCATCTTTCCCCTGTAT	TCATTGTATAGTATGTCTTCACAGAAAGAGTGC
NDFIP1	3_NDFIP1	5:141491497-141504554	TGGGTTCTTATCTTTCTAATTGGCTTTATG	AGGCAGGGACTGTGCCTTCTTTACTTT
NNMT	4_NNMT	11:113667556-113677820	GTGGCATTCCTTGGGTTCCTCAGGTGT	AGAGACAGAATACCCAATAAAAAGCTGTCA
NNMT	6_NNMT	11:113682768-113688962	GTGGCAGTTCTCTGAGGTAAATGCTCTCT	TTTTTGGATTGCTGGGATGAACAGAAATG
NRP1	7_NRP1	10:33592103-33602260	ACTTGTGCGCAGTACACCCTCCTGAATA	GCATAACGGATGCATAGTTTAGCCTTAAAA
NRP1	9_NRP1	10:33575641-33585751	CTGTGTATACAGATTTGCCCTGCCAGT	CCCAGTGAGCTATTTTTGTCAGTTCTCTAC
NRP1	12_NRP1	10:33545680-33556019	GAACCAACTTAGGGCCATGCTGTCTTT	AATGCTGATGAGCACATTTAGTTCTTTG
NRP1	13_NRP1	10:33535682-33545891	AGGAAAGGGCAATTGAGAAAATGACAG	CCAAATCTAAGGCAGATGGGGATTCTA
NRP1	14_NRP1	10:33525642-33535752	AATGTTTTGAGTGGGTTATTGTACCGTTTA	TAATCAGGTCAACAGTGACAGCGTCCTTTA
OPTN	3_OPTN	10:13200443-13210073	CCTTGACTTAAGCTGTGATGGTCTCTGTTA	ACTGAGCACTTTCCAAATGTTTCAATTTTA
OPTN	4_OPTN	10:13211643-13221285	ATCCATTTGAATGGTTGGATCATGAGTTAT	GGTAAATGAGTAGCAAAGGCCTGGTTTC
PBX1	4_PBX1	1:162824938-162835011	ACTCCTCACCCTTTGTCCTCATCACTC	AGAGAATGAAAAGAGGGGCAAAATCATC
PBX1	5_PBX1	1:162834951-162845011	CAAAGAAGAGTGTTTCTGCCAGTAGCTTTA	AAGAACAAAGAGGCCTCCTGAAAAGTCT
PBX1	6_PBX1	1:162847791-162857848	GCCGGTGTTTTGATAGTGTCAGGTTAC	AAAAGCTATGCAACCAGGGTTCATTTAC
PBX1	9_PBX1	1:162874812-162885011	ATGAGTAGGTACTGCCTGCCCGTTTTT	ATGGCGGTGGGAGAGGACCTAACTAGA
PDGFC	13_PDGFC	4:157982452-157993005	GACAAAAGGAAGAAAGAAGAAACATGTAGG	GCTAACTATCTGTGGTCATTTGATAAAAGAGTC
PDGFC	17_PDGFC	4:157942053-157952364	ATCCAAACCTTCCAAACCAAAAGTTGACAT	ACCAACCTCTACATATGCCTTTCTTGCATA
PDGFC	20_PDGFC	4:157903261-157913320	GCAAGGCTTTTGTTTTTGGAAGAAAAT	TTTTAGGCCCAGGACATGGAGCTTTAG
PIK3C3	1_PIK3C3	18:37788652-37798821	CGCAAGTATTATGTCCAAAGTAGCATGAAT	CTAGATTTCATTCATGAGTTTAGAAGCCTTTC
PIK3C3	2_PIK3C3	18:37798543-37808713	TCCTGGGCTATTCTGTCAGTTAGGATAGTA	GGAACAATGGGTAAATTTGCTCAAGTAAG
PIK3C3	4_PIK3C3	18:37818195-37828437	AAATTGTTATGCCACCTGTGTCTATTGTTC	CTGTATATCATCTCAGTCCCACCCAAAGTT
PIK3C3	5_PIK3C3	18:37828406-37838717	TAAACTTTGGGTGGGACTGAGATGATATAC	GGAAGTAAATTTTCATTTCACAGAGGAGTC
PIK3C3	6_PIK3C3	18:37847053-37857220	AAAGCATCTTGAATCCTTCCCACTTTTT	TTCAATCAAACACAAAGAATATGTAACTTGAAA
PIK3C3	8_PIK3C3	18:37858372-37868716	TTGTTTGAATAGATGTTTGGCTTGAATGA	AAAACATAAAGCACCATGTCTCACAAAGAT
PIK3C3	9_PIK3C3	18:37871504-37884737	CCCTAATGTATCTCATTGAAAACCCAACTT	CCAATAAAGGGTTCAGGTTTCATTCAA
PIK3C3	11_PIK3C3	18:37891709-37901814	CGAAGAATTTAATAGGCTTCATCGTGAACT	ATCATTCACCAAAACCCAACAAGTGTC
PIK3C3	13_PIK3C3	18:37908204-37916428	CATTAGCAAAGCTAGTTCTGTTTTAGGACCA	TCATTTTTACCATTTTTCTAATCAGCAGGT
PLG	1_PLG	6:161042707-161052800	TGTCATTTGGTGTAGGATGTAGAGATATTAACG	CCAAACATAGTCATCTTTGATCTTTTCTCA
PLG	2_PLG	6:161052700-161063235	TGAAGACCTAGAACATAGAAGAATGCCTAGTT	AAAACCTGGGTGTGAAAGAACAGATAGAGT
PLG	3_PLG	6:161063206-161073199	ACTCTATCTGTTCTTTCACACCCAGGTTTT	CAGTCCCCAATAATGGAAACTTTAAAGAAA

Gene	Primer Pair	Chromosome:Positions	Primer Sequence	Primer Sequence
PLG	4_PLG	6:161072667-161082666	TGATTCTGTCATCCTAGAGAAACCTGACAT	TCTGAAAACAGAGAGAGAGAGAGAGACTCA
PLSCR1	2_PLSCR1	3:147725569-147735661	AGGGTAGATCTTTACTTCAGACTTGCAAAA	CCTGCAGGCTAGTTGGTCAGCATCCTC
PLSCR1	3_PLSCR1	3:147715786-147725748	GCACAGAGCCACATTCCAAAGTAAGTTAT	GGAAGGATAGAGTGTAAGAGCTAAACATTTCA
PPP1CC	2_PPP1CC	12:109643077-109654167	TTTTGCCCCTTTATCTAAATCTCGTATTGT	TGGAGTGAAGAGTCTTTCATTTGCTGTTAT
PPP2CB	2_PPP2CB	8:30770039-30780759	GCATTTTTGTATAGTCTGATTGTCCATCTTT	GCAGGGATTATGTTGTTTATGTTCCCTTAG
PPP2CB	3_PPP2CB	8:30761911-30770872	CTAGGTGAAAAGAGGAAAAACTTGCTGGT	AATTACCTACTAAAATTTTTCGACTGGCTTTT
PRKCA	20_PRKCA	17:61918844-61929157	GTAAAGCTTTTTGCTGCCTCTGCAAATTGT	GCTTTCTCACTTAAGACACACAGAGAAAATG
PRKCA	39_PRKCA	17:62108849-62118914	GGTTTTTCTCAGTGCCGGGCTTGACAG	CTGCCCAGTCTCTTGCAGTTGTGTTTTT
PRKCA	44_PRKCA	17:62158832-62168963	GTGGGTACAAGTTTTCCTTTATGCAGATTA	GTGATTCTCAATCTTGCCTATGTCTGTGA
PRKCA	48_PRKCA	17:62198837-62208911	AAAGAAATTAGAGACTCCTTGTAGGTTCCTATG	TGTTTCCTCTGTTGCAAAGATAGAGACTATT
PRKCA	49_PRKCA	17:62208886-62218502	TCTCTATCTTTGCAACAGAGGAAACAGACT	CACTTTTCAGTTAGCTCTGGATCAGGAAAC
PRKCA	51_PRKCA	17:62228439-62238243	ATACTCACCATTCATTGTGGTGTGAAATCT	AGGTGAGGAAATCAGCAAAGATTGACTTTA
PRKCB1	20_PRKCB1	16:23944074-23954349	CTGTCACAACAACCCATTGACCACTGT	CTGTCGATGTGGGCCTGGATGTAGATG
PRKCB1	26_PRKCB1	16:24003811-24013294	TGTTGAAGAGACAGGAAGGTAAGGACATTT	TGTTTTTAGTACAAGGACCAAACAAAACATATC
PRKCB1	29_PRKCB1	16:24034283-24044582	CCAAGGACTTGAGCATTTATCCAAAGAG	AGAATCCACGCATTCAATCTGGGCACT
PRKCB1	34_PRKCB1	16:24093273-24104842	CTGAGCATTGCCAAGCATATGGTGTCCT	CTTAACCTCAACACGGCGTGGTATTTTT
PRKCB1	39_PRKCB1	16:24133811-24140370	GAGATGGGATTATGCAGATGGCCTATGG	TTTGCTCTTTTCTGCTACCCACTAGCCTTA
PTPRC	1_PTPRC	1:196873780-196883962	AAATATAAGAACCCTACAATAATGCTTCCAAAC	GGACTATTGTCTAGTACCTCTTCACATCTACCA
PTPRC	6_PTPRC	1:196923776-196933943	ACAGTGATGCTGAAGTCTTGGAATTTTCT	TTTAAAAATGTGGGAATGAAAAGAATGTTG
PTPRC	7_PTPRC	1:196933866-196943981	CCTTTTGAAGGTTCTGTATTTCACGTCAC	TCACAAGTAAAGGTTTCAATATTTTTCCATTTT
PTPRC	8_PTPRC	1:196943882-196953953	CAGGGGTTGAAAAGTTTCAGTTACATGATT	AATGGTAACGTTCATGGGGGCCATTAC
PTPRC	9_PTPRC	1:196953862-196964044	GAACATGACTGTCTCCATGACATCAGATAA	GCTAGACTGGTTGAATAATCACCTCCAAAG
PTPRC	10_PTPRC	1:196963436-196971490	CTGCATTTTCACTACATAAGAAAAGGTGAA	AACTATTATCTAAGGCCACAATGACCTCCT
PTPRC	11_PTPRC	1:196976586-196984397	ACAAGTCTTTACATGGGGGTTAAGATTTTC	GTATTTTCCCCTTACTTTCAACCTCCCAGT
RSU1	1_RSU1	10:16889696-16900102	CGCTTTAGTGGTGTTTTCAAATGACTTTTA	GCCCAATAACACTATTTTATTTGCTTAACACAG
RSU1	4_RSU1	10:16859991-16870090	CAGACTATACATCTCAGGGATAGAGCAATG	GGCCATCATTTATTGGACATAAGGTCATA
RSU1	6_RSU1	10:16840017-16850094	TCAAAGATGTAAGAAATATACTGATGTGCACTG	CTCCACCTCCTGGTGGCAAGAATCTACT
RSU1	7_RSU1	10:16831116-16839876	AAGGAATCAGCAGTTTTCAAAAAGAGATAAC	ATGAATGGGTACACGAGACTCAACTTACAT
RSU1	13_RSU1	10:16769685-16780077	GCTCTGTGCCCAAATGTAGCTGACTTC	TACAGGAACAGTTGAGCAATTAAACAAAAA
SAMHD1	2_SAMHD1	20:34994797-35004436	CATGTCAATTGTAAGAAGCATTAAATTTCAGA	TACTTAAATGCACTTCCAGCACTTCCATCT
SAMHD1	3_SAMHD1	20:34985316-34992808	AAGTTGCCCTAAAGGTATGTTAGAATGTTTAAG	TGCCAACTGTGAATATACTAGCAAACAGAA
SAMHD1	4_SAMHD1	20:34973612-34982240	CGGGGAAAGAATAGTTATAGGGCAAGTT	GGAATGAAGATATACTTCCTCAACACAAATG

Gene	Primer Pair	Chromosome:Positions	Primer Sequence	Primer Sequence
SAMHD1	5_SAMHD1	20:34965668-34974573	AAAAATGAAAGTTGCAAGCCAATAAATG	CCACAACTTTTTCCTCTGTGCTTGTATG
SAMHD1	6_SAMHD1	20:34952750-34961504	CTGTTTCCTCTGCCTGCAATTCTTTTTCTT	GAATCTCAGAGTAAACCACAATGATTTAAAGGT
SELP	1_SELP	1:167854810-167866485	TTACCAATAAAATTTCCCAAGCGCCAGAAA	CTTTCCTATGCCCTTTCTACCAATATGATG
SELP	2_SELP	1:167838274-167849672	CTTGTCTGGCTAGCTCAGTCCTATCTGA	GACTCTGTAAGGAAGTAAAACTTTCAGTGACAT
SELP	4_SELP	1:167826500-167836568	AATTTTCACATGGTGTTTTGCAGAGTTCTA	ACAATAGAGTTCATATAAGGATTTCCATGTCC
SELP	5_SELP	1:167823660-167826667	AGCAATAGTTTTCAAAAAGGGACAGTATGC	ATGGCTCCAGGTTTCTGTCTTATGCTTTAT
SFRP1	5_SFRP1	8:41237647-41247151	CTGTCTGTCCCTCTTTTTCTTCTTTTCTCT	TATGGACATTTTCCTTGGCTGGACTAGTAT
TLR2	TLR2	4:154842331-154847285	AGAAAATCCAGAATAAATATGCATGGTATGA	TTCACATGGTTCCCAGCTTAAGAAAGTTAC
TNFAIP8	TNFAIP8	5:118755851-118757553	CCTTTTGATTTGCTTAATCTGCCATTTTAC	CCACAGTACTGATTTCAGATAAGCCATTTTT
TNFSF10	1_TNFSF10	3:173714213-173724693	AATCTGTAAAAGGATAGTGACAGCGAGACA	CCCAGTTAACGTCTTTAAACAATGGTGGT
TNFSF10	2_TNFSF10	3:173706056-173714963	TTTTATATTCTTCCCACATTGCTGATGTCT	AAAGATTAGAAGTCTTTTCCCCCATTTTTAG
ZC3HAV1	3_ZC3HAV1	7:138415488-138425571	GGTGTAAGCCAATGATATGAAAAATTACTGT	TTCTAGCCCTATATGCCATGTTTCTACGAT
ZC3HAV1	4_ZC3HAV1	7:138405482-138415570	TGTTCATTATTTTGTTTCATTTTGGCTTTT	CCAACAAACCTGTGTCACATTTAATGAATA
ZC3HAV1	5_ZC3HAV1	7:138388581-138400377	GTTCACATCTCATTGCTAAAACGAAATCAT	GCATTTGATATTACCGTGAGCTTACTGTCTC
ANXA1	201_ANXA1	9:74955726-74956919	AATGATCAAATTTTTGGCATTACCTTTGTT	ACCAAGAGTACAATGAGCCAGTATTAAGCA
ANXA1	202_ANXA1	9:74965371-74975126	TTGTTTTGTTTTAGGGCAATGTAATAGAGC	TCTTGTGACGTCATTTTATTTTCAGCTACA
AQP1	201_AQP1	7:30917203-30918952	GAGGAAAGTCCTAAACTGTCCCTATCTTCA	TTGATTCCTAGAGGTGGTTTATTTGGAAAC
AQP1	202_AQP1	7:30927433-30930720	ATGTTTCTAAAAGTGGCCCAGGAGTAAGT	GAGCAAGATAATGCAGTGATAGATGGAAAG
ARNT	201_ARNT	1:149091212-149097699	GATATGAGCATTTGGGATTTTTAGCAAACT	TGTCAGAGGGTATGACAGAGTCACAGATTA
BHMT	201_BHMT	5:78442817-78443924	ATGTTTGGGTATAGGGGTAGAAGGTCATTT	TAGGAGCAGTATCTCTAGGGAGTTTTCTGG
C3	204_C3	19:6627981-6638064	TATCTGGGAAATTCTCAAAAATGGACAAAT	ACAAAACAACAACAACAAAAACCCATACAT
CD36	201_CD36	7:80130051-80133870	CACTGAGGCAAAGAAATGTAATCATCTAGG	ATTTTGTTGTGGGGATATAAAGGCAAGTAA
CD36	202_CD36	7:80136362-80141657	CCAAATGAACTTCACTGGAAGAAAAGTG	GCACAAGTGCTTTATTTGTGCTATTGTTAC
CD4	201_CD4	12:6793332-6798954	GAGTTGGTGCTCTCCAAATAAGGATATGAT	ATTAAGCCTCTGGAAACTAGAGAGCAACAC
CD44	201_CD44	11:35116038-35117728	AAAAGGCTTCCCCTGAAGAATATTACAAAC	GCCCAACAAACTTTCTTCTCTTTCTCTTTT
COL4A2	204_COL4A2	13:109939619-109946026	CTAAGCAAACCGCCTATGATACACACTAAA	GTGCCTCACTAACACTGATGGGATAAAATA
CSF1R	201_CSF1R	5:149472158-149474136	TGGACAGTGAGGACAGTTATGTCTTGTAAA	TTTTATCAACCTTTGTCTTAGTGTGGCATC
CSF1R	203_CSF1R	5:149413504-149421778	CTCTCAGTGCTATGTTCATGAGACGACCT	ATTTCCATGAAGATAAGGGGATTAGGAAAG
CSPG2	201_CSPG2	5:82802420-82803748	TTGATACTTCCAAGAAGTCCCTGTTACTCA	AAGTGCCTACATTTTCTTAACTTCGCCTTA
DDX17	202_DDX17	22:37219808-37225616	GTAGGAGCCTTGCAGAGTTAAAGTGATACA	TTTGTAAGGTTTGTCAGGCTTCATGAATAG
ERBB3	201_ERBB3	12:54764333-54769963	TGACACAGTCTACTCCCTACTCCCAAATAG	AAAAACAGGTGGTCATAGTGGGTTTTT
ERBB4	202_ERBB4	2:212245753-212252476	TGGCTTGAGAGAAGTGTGGTTTATTCTTTA	CAATTTCCATCCCAAAGACAGAATTAGTTT

Gene	Primer Pair	Chromosome:Positions	Primer Sequence	Primer Sequence
ERBB4	203_ERBB4	2:212230551-212231928	AAAAGGATCTTGATAATGTTCTGGAGCTGT	AGTAAGAAAGTTGGCTTGAGAAGGAAAGTG
ERBB4	204_ERBB4	2:212133648-212135476	TCAGCTCACCTTACTCTACCTGCAACTAAC	TTCGGAAGTGCTATATTTCAAACTTCATTG
GAP43	201_GAP43	3:116824073-116825471	CACTTTAATCTCTCATGCCCAAATTGTAAA	CTACCCTCACCCTAATTAGTAGCAGCATTC
GAP43	202_GAP43	3:116877345-116878717	GTCCGAAAGACATTACTTTTCCCATACAAT	ACATATACACACACTACTCACCAACCGATG
GAP43	203_GAP43	3:116922036-116923069	TTGATTGTAAGGCTGAGAGATTAAAGACCA	TCCAACCAACTAATGAGAAGGTAGAACAGA
HCK	202_HCK	20:30134923-30141088	TATGGTCTCTTGGCTTTTTGTAGGGTAGAT	AGAAAATTAGGTTTTCAGCAGTGTTTTTGC
HHIP	201_HHIP	4:145786730-145793701	AAAAGAAGAAGAAAAGAGGGAACGAAACAT	GGATTGAGAATCCAACGTGTGTAAAGATTA
HMOX1	202_HMOX1	22:34115140-34120007	TTGGAACATCTTATCTCTTAAGGTGGATGA	AAAACACAAGACACAACATCTTTTCAGGTT
IFNGR1	202_IFNGR1	6:137566116-137570176	GGAATTTATTGCTTCATGCTGTATTTGTGT	TCACATGGCTTTCCAAATTAGTTGTTCTAT
IL10RA	202_IL10RA	11:117368983-117371914	ATGTATTTTTAATGTGCTCCCCAAGAAGTC	TAGTAGATCATATCAAAATGGCTCGTGGTC
IL16	201_IL16	15:79348673-79352959	CTTCCTTGATAAACCTTAGTGAGCCTTTTG	AACTGAGACAGGATCTTGGATGGAAAG
IL16	202_IL16	15:79357759-79359358	GCCTAAAGATTAAGAAGGGATAGGAGGAGA	TAGGTTCGAATGGTATGTCTTGAGCTTCTA
IL6R	201_IL6R	1:152667371-152675933	TTTTAACCTTCCAGTTTGGTCCCTAGAGTA	CCCCAATGGCAATTATTACTACCTTTAACA
ITGA6	202_ITGA6	2:173037782-173042568	ATGAGAGAGGACGAACTCTTGTAATGTGAC	AAGTACTGGTAGAGGGGAAAAGAATGTCAA
ITGA6	203_ITGA6	2:173043674-173049625	AAACCTTCCATTTTTCATCTAGTTTCATGG	TCAAAAGAAAAAGAAATAGCTTCACAATTCA
MAT2B	202_MAT2B	5:162871440-162878294	CTTAACTTTAGAATTGGCTTGCAGATATGG	AATCCTGAAAATATGGAGGTTACGAGAATG
NNMT	201_NNMT	11:113632788-113633895	ATAGAAGCATCGCCCTATCTTTAGCTGTAA	GATCATATCTGGGTCAAAGAGCATGTAAAG
NRP1	201_NRP1	10:33658542-33660344	ATTTATGGGCAGATAAAACAACAACAACAA	GACCCACTTTAAAACTTGTGAAAACCTCTC
OPTN	202_OPTN	10:13190975-13198494	CAAAAATGTCCAAAATGTAACTGGAGAGAA	ATTCAAACCAAATATTCTGGAAAGATCCTG
PIK3C3	201_PIK3C3	18:37829628-37838642	AATGACGTGTGTATTATGGGACTAGTGGAG	ATCCTCTTCCCTTCTAAAAATCTCTCTTGC
PLG	201_PLG	6:161092881-161094250	AAGCAAAAAGTAAAGAAACAACAACCAACC	AGAATCCAGCAGTCAGAAAATAACACAAAA
PLG	202_PLG	6:161053730-161059722	CTTTTGAGGCCTTTATATTTCTCCTGACTG	CCAAATTTCTGAAAAAGAAGTGATTGTGAG
PLSCR1	201_PLSCR1	3:147744586-147750575	TGAAAGCTACTGGATTTCTACTGTCTCCTG	TACCTGTAAAGCAGTGTAGCTAAGGGGAAG
PRKCA	202_PRKCA	17:62199274-62200890	TAGAAATTCAAATCCATGTTACTCCCTTCC	ACACTGTCTACACTCCAGGTTGTAGCTTTC
PRKCA	203_PRKCA	17:62215099-62216468	TACATGTATAGCATGGTGTCTCCAAAACAG	AGGCTGACTTCTTAAAGGCTCTCAGTTTAC
RSU1	201_RSU1	10:16776423-16778063	TTTAGGAAGTTCTTTCACACTTTGCCTTTT	TATGTTCTGATCTGAATGGTGAATTTCCTT
RSU1	202_RSU1	10:16674338-16676036	TTATAGGGCACTTTGTCAGCATTTAATCAG	AAACCTCAGCTTGCATTTATTTTGATTTTC
SLA	201_SLA	8:134183559-134185192	GTTAATGCTTGAATGACTCCTGACTTCACT	TCCAAAGCTGTTATCCCTTCTTCAGATAAT
TAP1	202_TAP1	6:3999236-4003216	ACATGAATGAAAGCCTTTTGTGAAGAGTAA	GTTTGTACTCCAGGAAGTCTGCATTATCAC
TIMP1	201_TIMP1	X:47325794-47331245	TAGTTTTCTACTGACCCACTCACTTGCTTC	TAAAATATAAAAACCCAACATTTGGCATCC
TNFSF10	202_TNFSF10	3:173714934-173724775	AGCTTATGACATCTGATAGTGGGGAGATTT	AGACATCAGCAATGTGGGAAGAATATAAAA
VEGF	202_VEGF	6:43853257-43860569	GTACCCTGATGAGATCGAGTACATCTTCAA	TCGGTGATTTAGCAGCAAGAAAAATAAAAT

Gene	Primer Pair	Chromosome:Positions	Primer Sequence	Primer Sequence
ADD3	301_ADD3	10:111753610-111755856	TAGAGAATGGAGTCAGTTGTTTTGGACAAT	TACGTGTCTGTGTCTTCTGTTCTCTAGGTG
C3	301_C3	19:6667570-6672438	GACAGGTACAAAAGCTCTAGAAATGAGGAC	ATGATAATTCTATGAGAACACCCTCCTTCC
C3	303_C3	19:6652886-6660033	AATCAGAGGGGAAATGGAGATAAGATTTG	CCAGAGCTGTTTCTTCCTTCAATAAACTCT
C3	304_C3	19:6641261-6648952	ATTGACAGCGTTTAGTTCACAGGCTTC	CATGTAGCACTGATGTAGAAAGCACTTTTG
CSF1R	301_CSF1R	5:149427343-149433653	CGCATTGACTAATTTATGACCAGAAGAAAG	AATTGTGGCTTTGGCTAATAGGACAGTAAC
ITGB2	301_ITGB2	21:45147371-45155839	AGCTTTCTCTTCCTGGCTATGTTTCTGC	GCCGTCTATATGTGTTGTGGTCTTTAATGT
ITGB2	302_ITGB2	21:45138467-45146547	TACATAAACACACATGCCCACATATGTACC	CATCCTCTGTGTAAGGACAGAAACACCTC
PLSCR1	301_PLSCR1	3:147736270-147738404	TGCTTAAAGTTGGCAATAATCAAAACAAAA	GAAAAGCAAGGAGTCTAGTCCTGGAGATTA
RSU1	301_RSU1	10:16862926-16864739	AATCTTACTGAGCAGATTAAACCACCATGA	ATATACGCTTTGGCTAACAACGACTAATCC
ZC3HAV1	301_ZC3HAV1	7:138443898-138446151	GCTTCAGTAGGAGAGTTTGGAAGTTTTGAT	GGTAGTAGGGAGGGAAAGACTCAAGATACC
ZC3HAV1	302_ZC3HAV1	7:138396119-138400735	ATGGTCTTTCTTTACTTCCCTTCACGACTA	TCATTTTCCGATCTAGTATCCTTTCAGTCA
ZC3HAV1	303_ZC3HAV1	7:138382698-138390896	TTGCTTAATGCTAACACATTAGGACCTTTG	ATCAGAATTTGTTTAAAACCTCCCAGATGA
CD86	422106	3:123286816-123297440	GATTCCCAGTAACCTTCTTATTTTGTAGTTCA	CCCAAACAACCCTATACACATCCTATTC
CD87	422107	3:123297432-123309211	TTGTTTGGGAGCAAGAGTGAAATTGGTATGGA	ATCACCGCTGGAGAAGGGGTCAAGGTA
GNA14	401_GNA14	9:79452132-79456335	GAGTTGCTTCCTCAGAAGAGATGTAATTGA	ATAAGGTACAGCGGTCAAGAGGTTAAGTGT
HCLS1	401_HCLS1	3:122841802-122849830	ACTTTCTCTTAGGGATAGAGGTGTCCCTTC	TGTCCGTTACTATTAAGACCTCTAGAGTTGAGC
HHIP	401_HHIP	4:145845892-145854595	ACATATACTATTGTGTGGGGGACAAAAACA	GTTGTTGAATTTGCAAGACATACCAAGTTT
HHIP	402_HHIP	4:145854566-145861774	AAACTTGGTATGTCTTGCAAATTCAACAAC	GCCACAAATCAACTGATGTTTGAAAGTTAT
IL6ST	401_IL6ST	5:55270124-55279832	AATGCAGATGAGGATTTGTGTGGTATTTAG	CTAGTTTTGTCCAGACAAGGTTTTCTGATG
LRP2	401_LRP2	2:169853046-169860488	AACATTTCTCAAGCACAGAGAGTAATGTCC	GTGAGATATAAAGCTGGCATCAAAATCAAA
PTPRC	401_PTPRC	1:196984581-196992431	CAGTGCATGGTTGACCTAGTTAATTTTCAT	TATGAGTAGAACAAGGAGGACATCTTGAGG
SIGLEC10	401_SIGLEC10	19:56606154-56613894	AGAGTTTACTCCATTCATTGCAAATTACCC	CTAAGAGACCCTCATTGGAACTTGACTTCT

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Stacy L. Moore
Author Signature

04/02/2010
Date