Title

Genetic Risk Factors for AIDS-Related Non-Hodgkin Lymphoma in the Multicenter AIDS Cohort Study (MACS): Candidate-Gene Study, Genome-Wide Association Study, and Pathway Analyses

Author

Keebler, Daniel

Publication Date

2016

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Genetic Risk Factors for AIDS-Related Non-Hodgkin Lymphoma in the Multicenter AIDS Cohort Study (MACS): Candidate-Gene Study, Genome-Wide Association Study, and Pathway Analyses

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Epidemiology

by

Daniel Sean Keebler

2016

ABSTRACT OF THE DISSERTATION

Genetic Risk Factors for AIDS-Related Non-Hodgkin Lymphoma in the Multicenter AIDS Cohort Study (MACS): Candidate-Gene Study, Genome-Wide Association Study, and Pathway Analyses

by

Daniel Sean Keebler

Doctor of Philosophy in Epidemiology

University of California, Los Angeles, 2016

Professor Shehnaz K. Hussain, Co-Chair

Professor Otoniel M. Martinez, Co-Chair

Background: Persons living with HIV/AIDS (PLWHA) are at elevated risk for non-Hodgkin lymphoma relative to HIV-negative individuals. AIDS-related non-Hodgkin lymphoma (AIDS-NHL) etiology is characterized by B-cell activation and chronic inflammation in the context of HIV; prior work has highlighted the importance of inflammatory biomarkers in predicting subsequent AIDS-NHL diagnosis. General-population risk factors for NHL include prior family history, suggesting a heritable genetic component to NHL risk that may encompass inflammation-related genes. Previous genome-wide association studies (GWAS) have investigated NHL in the general population, but no AIDS-NHL GWAS has yet been published. We therefore investigate single-nucleotide

polymorphisms for association with AIDS-NHL risk using a candidate-gene study in 700 HIV-positive men (178 AIDS-NHL cases), and a GWAS in 1,949 HIV-positive men (172 AIDS-NHL cases), from the Multicenter AIDS Cohort Study (MACS). Pathway analyses of GWAS results complement these two studies.

Methods: Candidate-gene study (30 SNPs; 24 genes): matched case-control design; conditional logistic regression, semi-Bayes correction and multiple imputation for missing covariate data in SAS. GWAS (n=4.86 million SNPs): imputation in MINIMAC3 for missing genotype data; logistic regression on genotype probabilities using SNPTEST v2.5.2 and adjustment for principal components from R package SNPRelate; regional plots using R package LocusExplorer. Pathway analyses: analysis of GWAS p-values using 13,094 gene sets in PASCAL and 6,212 gene sets in VEGAS2.

Results: Candidate-gene study: significant inverse association (dominant OR=0.68; 95%CI 0.47-0.99; log-additive OR=0.71, 95%CI 0.51-0.99) between NHL risk and SNP rs6815391 (3' UTR, REX1/ZPF42, 4q35.2). GWAS: genome-wide-significant signal (4q33; top SNP rs2195807; p=1.48E-08; white-only p=1.93E-07) in the vicinity of uncharacterized noncoding variant LOC100506122. Pathway analyses: repeated occurrence of gene sets capturing inflammatory processes and muscle fiber/ cytoskeletal integrity at or near the top of most scenarios.

Conclusion: Each method yielded associations with different aspects of AIDS-NHL biology. The candidate-gene study suggests involvement of the REX1/ZPF42 SNP rs6815391 in HIV replication and production of anti-inflammatory cytokines; the GWAS points toward an as-yet-uncharacterized locus (LOC100506122); pathway analyses implicate myosin genes and pathways that may be involved in early stages of B-cell activation. Further candidate-gene, *in vivo* or *in vitro* studies may clarify the biological plausibility of observed associations.

The dissertation of Daniel Sean Keebler is approved.

Roger Detels

Jian Yu Rao

Donald P. Tashkin

Shehnaz K. Hussain, Committee Co-Chair

Otoniel M. Martinez, Committee Co-Chair

University of California, Los Angeles

2016

Table of Contents

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

In Los Angeles, I am grateful for support from the UCLA Graduate Division's Dissertation Year Fellowship, the Fielding School of Public Health Department of Epidemiology, and especially the National Institutes of Health/National Cancer Institute T32 Training Program in Molecular-Genetic Epidemiology of Cancer (T32CA009142). A special debt is owed to Zuo-Feng Zhang for welcoming me into the program and providing continued support.

I am deeply grateful to the members of my committee, including Roger Detels, Donald Tashkin, Jian-Yu Rao, and especially my co-chairs, Shehnaz Hussain and Otto Martínez-Maza. The amount of time and effort they have devoted to this dissertation through weekly calls is well beyond any reasonable expectations, and this process has been much easier for it.

Zhang lab members have enriched my experience here, including Aileen Baecker, Esther Chang, Po-Yin Chang, Alan Fu, Somee Jeong, Kexin Jin, Claire Kim and Travis Meyers. Erin Peckham in particular was instrumental in helping to think through the early stages of the candidate-gene study despite an insanely hectic schedule, which I appreciate deeply. Outside the Zhang Lab, Melissa Van Dyke, Vivian Alfonso, Reena Doshi, Nicole Hoff, Jennie McKenney, Andrew Park, Solomon Makgoeng, Roch Nianogo, Ashley Kissinger, and Adam and Heather Readhead have made Los Angeles a better place to be.

In Stellenbosch and London, I thank Alex Welte and Tim Hallett for taking me on board. Thanks also to Juliet Pulliam and Jamie Lloyd-Smith for the introduction to SACEMA, to Aldi Du Toit, Martin Nieuwoudt, Reshma Kassanjee, Hilmarie Brand, Lynnemore Scheepers, Cari van Schalkwyk, Dave Matten, Gavin Hitchcock, John Hargrove, Gordon Botha, Philip Labuschagne, Paul Revill, Ellen McRobie, Jack Olney, Jeff Eaton, and especially Katrina Du Toit.

Elsewhere, my parents and extended family, Ryan Charles, Mike Tramontini, Noah Marshall-Rashid, Carrie Fawcett, Lisa Waud, Justin Andrews, Shireen and William Wetmore, Tony O’Rourke, Christine Varnado, Paul Pavwoski, Lauren Sill, Matthew Sill, and Maggie, Alex and Sadie Sill have been invaluable sources of support.

VITA

2003	University of Michigan, Ann Arbor, USA Bachelor of Arts, Biopsychology & Cognitive Science
2004	Teaching Assistant Northern Studies Program, University of Alaska Fairbanks
2005	Research Assistant Institute of Arctic Biology, University of Alaska Fairbanks
2006 – 2008	Harvard University Master of Science, Population & International Health
2008 – 2009	Health Program Manager II Alaska Division of Public Health, Anchorage, Alaska, USA
2010	Research Analyst IV Alaska Division of Public Health, Juneau, Alaska, USA
2012- 2013	Visiting Fellow & Research Consultant South African Department of Science and Technology/ National Research Foundation Centre of Excellence in Epidemiologic Modelling and Analysis (SACEMA), Stellenbosch, South Africa
2014	Senior Associate Center for Observational Research, Amgen, Inc. Thousand Oaks, CA, USA
2015	Dissertation Year Fellowship University of California, Los Angeles

PUBLICATIONS

Keebler DS, Revill P, Braithwaite RS, Phillips AN, Blaser N, Borquez A, Cambiano V, Ciaranello A, Estill J, Gray R, Hill A, Keiser O, Kessler J, Menzies NA, Nucifora KA, Salazar Vizcaya L, Walker S, Welte A, Easterbrook P, Doherty M, Hirnschall G, Hallett TB. Strategies for monitoring adults on antiretroviral treatment: A combined analysis of three mathematical models. Lancet Global Health 2014; 2(1): e35-e43 2013.

Hallett TB, Menzies NA, Revill P, **Keebler DS**, Bórquez A, McRobie E, Eaton JW. Using modelling to inform international guidelines for antiretroviral treatment. AIDS 2014; 28(S1): S1-S4.

Keebler DS, Walwyn D, Welte A. Biology as population dynamics: Heuristics for transmission risk. American Journal of Reproductive Immunology 2013; 69 (Suppl 1): 88-94.

Braithwaite RS, Nucifora KA, Toohey C, Kessler J, Uhler LM, Mentor SM, **Keebler DS**, Hallett TB. How do different eligibility guidelines for antiretroviral therapy affect the cost-effectiveness of routine viral load testing in sub-Saharan Africa? *AIDS* 2014; 28(S1): S73-S83.

Chapin, FS III, Eviner VT, Brewer C, Magness D, Talbot L, Wilcox B, **Keebler DS**. 2008. Disease effects on landscape and regional systems: A resilience framework. In Ostfeld R, Keesing F and VE Eviner (eds.), *Infectious disease ecology: effects of disease on ecosystems and of ecosystems on disease.* New York and Oxford: Princeton University Press.

Keebler DS, Hawkins H, Lien D, Massay S, Sanders L, Tofteberg C. Rural immersion as an instructional method in northern community health. *Alaska Medicine* 2005 Apr-Jun; 47(1):2-7.

CONFERENCE & PROFESSIONAL MEETING PRESENTATIONS

"How Should HIV Programmes Monitor Adults on ART? A Combined Analysis of Three Mathematical Models." International AIDS Society: 7th Conference on HIV Pathogenesis, Treatment & Prevention, Kuala Lumpur, July 2013.

"The Cost and Impact of Alternative Strategies for Monitoring Adult and Child Patients on ART." Bill & Melinda Gates Foundation: Modeling to Support Decision-Making Meeting, Seattle, WA, June 2013.

"The Cost and Impact of Alternative Strategies for Monitoring Adult and Child Patients on ART." World Health Organization: Consolidated ARV Guidelines Committee Meeting, Geneva, Switzerland, December 2012.

"Rural Immersion as an Instructional Method in Northern Community Health." American Association for the Advancement of Science: 55th Arctic Science Conference, Anchorage, Alaska, September 2004.

"Embodiment and the Translation of Inuit Emotional Worlds: Implications for Biomedical Conceptualizations of Emotional Distress." International Arctic Social Sciences Association: Fifth International Congress for Arctic Social Sciences, Fairbanks, Alaska, May 2004.

CHAPTER 1 : INTRODUCTION AND BACKGROUND

1.1. Descriptive Statistics and Epidemiology of Non-Hodgkin Lymphoma

1.1.1. Global Data

1.1.1.1. Incidence and Prevalence: Both Sexes Combined

Globocan reports that in 2012, non-Hodgkin lymphoma (NHL) was the number-ten cancer worldwide, accounting for 2.7% (n=386,000) of all new cancers and 2.4% (n=200,000) of all cancer deaths. Of these 386,000 cases, 218,000 (56%) were in males, and of these deaths, 115,000 (58%) were in males[1].

Incidence and prevalence of NHL in the general population is higher in more developed regions. In “more developed” regions—comprising all regions of Europe, North American, Australia, New Zealand and Japan—there were an estimated 188,767 incident cases of NHL in 2012, with a one year-prevalence of 132928 (12.8/100,000 persons). In “less-developed” regions—comprising Africa, Asian countries other than Japan, Latin America and the Caribbean, Melanesia, Micronesia and Polynesia—there were an estimated 195,338 incident NHL cases, with a one-year prevalence of 89,605 (2.2/100,000). This contrast is even starker in regions of “very high human development” versus regions of “low human development” (i.e. regions with a United Nations Development Programme Human Development Index (HDI) of 0.8 or greater versus those with an HDI less than 0.55). Very high human development regions had an estimated 189,112 incident cases in 2012 and a one-year prevalence of 132,018 cases (13.8/100,000); low human development regions had an estimated 36,291 incident cases in 2012 and a one-year prevalence of 14,122 cases (1.8/100,000). This higher incidence likely has to do with the population structure of higher- versus lower-resource settings: the former are characterized by a high number of older persons relative to lower-resource settings, and given

that these data are for NHL in the general population, the increased burden in higher-resource settings probably reflects the age-related increase in risk for non-AIDS-NHL (omitting pediatric cancers such as endemic Burkitt lymphoma).

Neither incidence nor prevalence reported above reflects the full impact of the HIV/AIDS epidemic in higher-resource settings, which has diminished since the 1990s: incidence data are for 2012, while the majority of individuals diagnosed with AIDS-NHL at the peak of the epidemic will have been deceased for some years and therefore not factored into one- or five-year prevalence. Using data from 10 U.S. SEER sites that ascertained HIV status in NHL cases from 1992-2009, Engels and Shiels[2] estimated that during this period, 6,784 (5.9%) of the 115,643 NHL diagnoses in the United States were among PLWHA. Of these 6,784 cases, 3,089 (45.5%) were diffuse large B-cell lymphomas (DLBCL); 568 (8.4%) were Burkitt lymphomas (BL); and 3,127 (46.1%) were either of other subtypes or of unknown subtype. Importantly, Engels and Shiels noted an increase in NHL in the United States prior to the HIV epidemic, the reasons for which have not been fully explained. This could be due to a shift in the distribution of risk factors for non-AIDS NHL, which we discuss in Section 1.1.1.2.

Greater one-year prevalence in higher-resource settings suggests that one-year survival may be poorer in low human development regions than in high human development regions. This is supported by five-year prevalence, which in more-developed regions was 518,868 prevalent cases (49.9/100,000) versus 313,975 (7.6/100,000) for less-developed regions, and 517,119 cases (54.2/100,000) for regions of “very high” human development, versus 47,738 cases (6.1/100,000) for regions of “low” human development.

1.1.1.2. Mortality: Both Sexes Combined

Explicit mortality data support the implication of prevalence data from high- versus low-resource settings: though incidence and prevalence of NHL in the general population are higher in more-developed regions, the burden of mortality is heavier in less-developed regions. Worldwide, Globocan estimates 199,670 deaths from NHL in 2012, for a crude rate of 2.8/100,000, an age-standardized weighted rate [ASR(W)] of 2.5/100,000, and a one-year cumulative risk of death (hereafter "CR") of 0.26%. Less-developed regions have a higher absolute burden of mortality [n = 124,542; crude mortality rate = 2.1/100,000], but a lower relative adjusted burden [ASR(W) = 2.3/100,000; CR = 0.24%], than more-developed regions [n=75,128; crude rate = 6.0/100,000; ASR(W) = 2.7/100,000; CR = 0.28%]. Very high human development regions had an estimated 73,445 deaths from NHL in 2012, for a crude rate of 6.4/100,000 persons, an ASR(W) of 2.8/100,000 persons, and a CR of 0.29. In contrast, low human development regions had an estimated 27,158 deaths from NHL in 2012, for a crude rate of 2.1/100,000 persons, an ASR(W) of 3.0/100,000 persons, and a CR of 0.32. Regionally, Melanesia [ASR(W) = 4.7/100,000; CR = 0.54% (n=267)] and Northern Africa [ASR(W) = 4.5/100,000; CR = 0.51% (n=7,525)] have the highest relative adjusted burdens of NHL mortality. In contrast, Eastern Asia [(n=40,860); crude rate = 2.6/100,000; ASR(W) = 1.8/100,000; CR = 0.18%] and South-Central Asia [(n=27,105); crude rate = 1.5/100,000; ASR(W) = 1.8/100,000; CR = 0.20%] have the lowest[1].

1.1.1.3. AIDS-NHL Epidemiology

Notably, data specific to AIDS-NHL, as opposed to NHL in the general population, are difficult to come by in much of the world; in the United States, nuanced data come from either cohort studies or linkage of HIV and cancer registries, but in many countries both types of registry are

either nonexistent or nonfunctional, making the calculation of reasonable estimates difficult[3,4]. A focus on data from sub-Saharan Africa, the region hardest-hit by HIV/AIDS, is informative.

In contrast to the United States, where (as discussed below) HAART has reduced the incidence of NHL, NHL incidence among adults in sub-Saharan Africa has risen in the era of HAART, while the incidence of Kaposi sarcoma has dropped precipitously. The reasons for increasing NHL incidence are unclear, but could conceivably be related to improvements in diagnostic capacity over time, increased life expectancy in both the general population and in the PLWHA population with the scaleup of ART, and chronic immune activation even in the presence of HIV.

Regardless, the worldwide burden of comorbidities in PLWHA is expected to grow in years to come, as a result of improvements in HIV-specific and general life expectancy, and also from increased "Westernization" of dietary and lifestyle habits in low- and middle-income countries. Coupled with the fact that the average global treatment gap is 66%, the intersection of HIV-specific and general-population risk factors for cancer means that malignancies in PLWHA will be a major global health issue in the coming decades[5,6].

1.1.2. United States Data

1.1.2.1. AIDS-NHL Surveillance Data

In the US, three NHL subtypes (diffuse large B-cell lymphoma, Burkitt lymphoma, and primary central nervous system lymphoma, all detailed in Section II.C) have been considered AIDS-defining cancers since 1993. Rates of NHL have plateaued in the general population since 2000[2], and the burden of AIDS-NHL has decreased since the introduction of HAART, but from 1996-2010 AIDS-NHL accounted for more cases of cancer (n=4,136) in HIV-infected individuals in

the National Cancer Institute's HIV/AIDS-Cancer Match Study (HACM) than any other, including Kaposi sarcoma (n = 2,437)[2]. PLWHA are at greatly elevated risk of NHL relative to HIV-negative persons: in the HACM, the incidence of all NHL subtypes combined in persons living with HIV or AIDS (PLWHA), over almost 1.5 million person-years of follow-up from 1996-2010, was 11 times the incidence in the general population, including HIV+ persons (i.e. the standardized incidence ratio, or SIR, was 11).

Since the introduction of HAART, the incidence of DLBCL and PCNSL, associated with low CD4 counts and poor control of EBV, has decreased, with BL and HL now contributing a greater proportion of NHL cases in PLWHA. However, SIRs are even higher for AIDS-defining NHL subtypes than for all subtypes combined (17.6, 33.7, 47.7 and 19.9 times the incidence in HIV-negative persons for DLBCL, BL, PCNSL and "NHL—not otherwise specified" respectively)[7]. Despite decreases in incidence thanks to widespread availability of HAART, this elevated risk persists even in the late era of HAART: NHL incidence among PLWHA in the HACM study was ten times that of HIV-negative persons from 2006-2010 (compare to the 1996-2010 SIR of 11, above)[8]. The fact that increased NHL risk in PLWHA persists compared to the general population, despite advances in treatment options, indicates the need for better preventive and control measures and continued investigation of NHL etiology.

1.1.2.2. AIDS-NHL Data from Cohort Studies

Cohort studies such as the MACS have also contributed to the descriptive epidemiology of cancer in PLWHA[9-12]. MACS data have shown a decline in KS and NHL incidence since the introduction of HAART[13], while the impact of HAART on the incidence of non-AIDS-defining cancers is less clear: among HIV-positive men in the MACS, the incidence rate of all cancers

combined decreased in the HAART era (965.9 cases/100,000 person-years; 95% CI = 798.8-1157.6) relative to the pre-HAART era (3601.5 cases/100,000 person-years; 95% CI = 3343.8-3871.3), but this was due largely to the HAART-induced drop in KS and AIDS-NHL incidence[14].

However, the increased risk of cancer in PLWHA appears not to be limited to AIDS-defining cancers such as AIDS-NHL: external comparison of MACS data to SEER data yields a standardized incidence ratio for all non-AIDS-defining-cancers in PLWHA vs. HIV-negative persons of 1.46 (95% CI, 1.19-1.78). This is consistent with findings from the Veterans Aging Cohort Study (VACS), in which HIV-positive veterans were at higher risk for non-AIDS defining cancers than HIV-negative veterans[15]. Plausible biological explanations for such elevated risk center on chronic immune activation and accelerated immunosenescence in PLWHA, highlighting an important etiological role for inflammation.

Epidemiological data from MACS studies have also contributed to the etiological understanding of AIDS-NHL: MACS data were used to establish that increased B-cell activation and higher levels of serum cytokines, immune activation markers, and micro-RNAs could be detected years before NHL diagnosis[16-20]. Even in the late era of HAART, the MACS continues to serve as a valuable epidemiological resource.

Pooled analyses from the InterLymph consortium have identified multiple risk factors for NHL in the general population. These were classified into ten categories: 1) familial history of hematologic malignancy (OR 1.72; 95%CI 1.54-1.93)[21]; 2) B-cell-activating autoimmune disease (OR 1.96; 95%CI 1.60-2.40); 3) hepatitis C seropositivity (OR=1.81; 95%CI 1.39-2.37); 4) atopic disease, including hay fever, eczema and allergy (OR=0.82; 95%CI 0.77-0.88); 5) blood

transfusion prior to 1990 (OR=0.76; 95%CI 0.67-0.87); 6) anthropometric factors, including height and BMI as a young adult (OR=1.95; 95%CI 1.51-2.53); 7) alcohol consumption, of more than one drink per month (OR=0.87; 95%CI 0.81-0.93); 8) duration of cigarette smoking (OR=1.06; 95%CI 0.99-1.14); 9) sun exposure (OR=0.74; 95%CI 0.66-0.83); 10) socioeconomic status (OR=0.88; 95%CI 0.83-0.93); and 11) occupational history, with teaching showing an inverse association (OR=0.86; 95%CI 0.77-0.95) with NHL risk and painting (OR=1.22; 95%CI 0.99-1.51) and general farm work (OR=1.28;95%CI 1.10-1.50) showing positive associations with NHL risk[22].

For each of these factors, the strength of association with NHL risk tends to vary according to NHL subtype. The connection of autoimmune diseases with NHL risk is especially interesting. Shiels *et al.* hypothesize that a recently-observed increase in PCNSL cases among persons aged 65+ may be due to increased use of immunosuppressive drugs for autoimmune conditions and organ transplants[23,24].

1.1.2.3. Conclusions: Epidemiology of NHL

Some have suggested that the experience of the United States with HIV-associated malignancies may foreshadow the future experience of low- and middle-income countries[25]. However, the fact that NHL incidence has not declined in sub-Saharan Africa in the era of HAART, in contrast to the United States, suggests that this experience may not be directly transferable. While improving health systems is the most obvious and urgent target for improving population health, these results also suggest a need to better understand NHL etiology in geographically and ethnically diverse populations—particularly given apparent ethnic differences in NHL

susceptibility in the United States. Again, improving our understanding of etiology is a primary motivation of this dissertation.

1.2. B-Cell Development: a Prerequisite to Understanding NHL Subtypes

NHL is an extremely heterogeneous group of malignancies defined primarily on the basis of criteria that are difficult to grasp without some knowledge of B-cell differentiation and proliferation. Thus before moving on to a discussion of NHL subtypes, we will pause briefly to consider these processes.

B-cell development begins with differentiation from hematopoietic stem cells in the bone marrow and rearrangement of immunoglobulin (Ig) gene segments. Here, the D_H and J_H segments on heavy chain μ are rearranged; the V_H segment is then joined to the DJ segment. This marks the transition of B cells from pro-B cell to pre B-cell status. During this time the pre-B cell antigen receptor complex is also formed, which carries out allelic exclusion and rearrangement of light chain genes. Next the κ and λ chains on pre B-cells are rearranged and combined with the μ chain, leading to formation and expression of an IgM molecule and marking these B-cells as immature B-cells. These cells then migrate from the bone marrow to the spleen and lymph nodes, where they further differentiate into mature B-cells and express both IgM and IgD[26]. Entering the lymph nodes from the bloodstream, some activated B-cells may migrate to sites of intense cell division and proliferation within the lymph nodes called germinal centers, characterized by light and dark zones of proliferating cells and a border region of resting B-cells called the mantle zone[27].

Importantly, it is in germinal centers that the process of somatic hypermutation occurs: here, extremely high rates of mutation at the hypervariable sites of V_H and V_L genes, which code for the receptors' antigen-binding groove, are an effective means of generating antibody receptor

diversity. However, this accelerated rate of mutation also means that deleterious oncogenic mutations can arise. The principal actor in this mechanism is the enzyme AICDA (activation-induced cytidine deaminase). After somatic hypermutation, cells with low affinity for binding antigen are deleted from the B-cell repertoire, while cells with the highest antigen-binding affinity can go on to become either memory B-cells or plasma cells. Memory B-cells divide very slowly but retain whatever genetic characteristics their rapidly-proliferating precursor cells may have had, including oncogenic mutations incurred in somatic hypermutation. Plasma cells migrate to the periphery or the bone marrow, becoming "post-germinal center" cells[27].

AICDA is active in another process that can drive lymphomagenesis: Immunoglobulin class switch recombination (CSR). CSR occurs in activated B-cells, and results in a change in the type of antibodies produced by the cell (e.g. from IgM or IgD to IgG, IgA, or IgE)[27]. During this process, portions of the Ig heavy chain's constant region are removed and reshuffled; however, this can give rise to deleterious mutations and translocations, leading to lymphomagenesis[28].

With these definitions in hand, we can now turn to the definition of NHL subtypes.

1.3. Non-Hodgkin Lymphoma: Definitions and Common HIV/AIDS-Related Subtypes

Non-Hodgkin lymphomas are cancers of B- and T-cells. The 2008 WHO classification of lymphomas distinguishes sixty different types of NHL; this list can be found in Appendix A[29,30]. WHO further distinguishes between "lymphomas also occurring in immunocompetent patients," including BL and DLBCL, and "lymphomas occurring more specifically in HIV-positive patients," including lymphomas arising in HHV-8 related multicentric Castleman's disease, and the DLBCL subtypes primary effusion lymphoma and plasmablastic lymphoma. Note that Hodgkin lymphoma, while also common in both PLWHA and immunocompetent patients, is not discussed in this dissertation.

Different subtypes tend to emerge at different CD4 counts; subtypes that emerge at very low CD4 counts (<100) also tend to be associated with Epstein-Barr virus (EBV) infection, which is poorly controlled in cases of severe immunosuppression. As reviewed in sections 1.1.1.3 and 1.1.2, the advent of HAART has changed the epidemiology of NHL subtypes in the United States, with low-CD4/EBV-associated subtypes such as PCNSL becoming rarer and subtypes such as Burkitt lymphoma (which appears at higher CD4 counts) contributing a greater proportion of AIDS-NHL cases. DLBCL is also less frequent than it used to be. Table 1.1 summarizes the epidemiology of NHL subtypes in the United States.

1.3.1. Diffuse Large B-Cell Lymphoma (DLBCL)

DLBCL is itself classified into a number of subtypes depending on cytology, gene expression patterns, and site of occurrence. With regard to cytology, DLBCL can be either immunoblastic (activated B-cell), centroblastic (activated B-cells proliferating in the dark zone of germinal centers), or anaplastic (B-cells that have lost structural differentiation). Similarly, gene expression studies have grouped DLBCL into activated B-cell and germinal center B-cell types; activated B-cell type is characterized by mutations in NFKB and B-cell receptor signaling pathways, while mutations in genes involved in histone modification (with implications for B-catenin signaling) are more common in the germinal-center DLBCL type[31]. Site of occurrence is relevant for PCNSL and primary effusion lymphoma (PEL), each discussed below.

1.3.1.1. Primary Central Nervous System Lymphoma (PCNSL)

PCNSL is defined based on the location in which it occurs, and it is not included in the 2008 WHO classification system as a distinct subtype. For instance, a DLBCL that occurs in the brain is both DLBCL and PCNSL, and most PCNSLs are DLBCLs. PCNSL has been considered an AIDS-defining diagnosis by the Centers for Disease Control and Prevention (CDC) regardless of

tumor subtype, and even in the absence of laboratory-confirmed HIV infection, since the 1980s[32]. Notably, the vast majority of PCNSL tumors are Epstein-Barr Virus (EBV) positive.

1.3.1.2. Plasmablastic Lymphoma (PL)

Plasmablastic lymphoma is highly aggressive and develops in B-cells that are not yet fully-formed plasma cells, but show immunohistochemical and cytological characteristics of plasma cells. It is associated with human herpesvirus 8 (HHV-8) infection, the same virus that causes Kaposi sarcoma (KS). PL is associated with low CD4 counts; thus its incidence has decreased in the HAART era.

1.3.1.3. Primary Effusion Lymphoma

PEL is quite rare relative to the subtypes discussed above; it emerges at very low CD4 counts (<100 or so). It tends to appear in body cavities and is also associated with HHV-8 infection; many patients presenting with PEL are also positive for KS. HIV-positive PEL cases tend also to be co-infected with EBV; in cases that are not co-infected, upregulation of the MAPK pathway has been observed. PEL is thought to emerge from a post-germinal-center B-cell, as evidence of Ig rearrangements and somatic hypermutation has been found[33]. Immunophenotypically, PEL cells tend to be CD30+, CD38+, CD71+, HLA-DR and CD138+, suggesting lymphocyte activation and plasma cell differentiation; they are CD3-, CD4-, CD8-, and CD45+, CD19-, CD20- & CD79a-.

1.3.2. Burkitt Lymphoma (BL)

Burkitt lymphoma is most famously associated with chromosomal translocation impacting c-Myc signaling. Expression data suggest that it emerges from germinal center B-cells. Additional work suggests that several genetic events, in addition to this translocation, impact pathogenesis,

including ID3, GNA13, TP53, and SMARCA[34]. These mutations impact PI3K signaling and the formation of focal adhesion complexes. The endemic form of Burkitt lymphoma, found in sub-Saharan Africa, is almost uniformly EBV-positive and malarial co-infection plays an etiologic role, while AIDS-related Burkitt lymphoma is to a lesser extent EBV-positive. Disentangling the contribution of HIV/AIDS to endemic BL epidemiology in SSA is challenging, but data from Cape Town and Kampala indicate that only a minority of BL cases in these cities present with concurrent HIV/AIDS infection.

1.4. Etiology of AIDS-NHL

AIDS-NHL lymphomagenesis is driven by two overarching processes. The first is chronic immune activation as a result of such factors as microbial translocation from the gut[35,36]; the second is the often concomitant failure to control oncoviruses, such as EBV and HHV-8 (KSHV), as a result of immunosuppression. These processes lead to increased B-cell transformation, and lymphogenic translocations and mutations. Thus B-cell lymphomagenesis is at root a matter of inflammatory response, cell differentiation and proliferation. Wnt, Notch, NFKB, and stem-cell pathways explored in Chapter 2 are involved in these processes. A more detailed exploration of their role in AIDS-NHL etiology follows in the next section.

1.4.1. Inflammation

The above, including the role of AICDA in somatic hypermutation and IgH class switch recombination, suggests that inflammation and subsequent B-cell responses should play a role in lymphomagenesis. Recent work has reinforced the connection between inflammatory processes and NHL lymphomagenesis. Elevated levels of several biomarkers, including IL-6, IL-10, CXCL13, sCD27, sCD30, neopterin, and λFLC, have been found years prior to subsequent NHL diagnosis[18,19,21]; assessment of biomarkers in prevalent cases using expression pathway analysis

has shown interactions between pre-diagnosis and prevalent-case biomarkers[37]. This finding suggests the involvement of inflammatory pathways (e.g. NFKB) in the etiology of NHL, and also suggests that expression of genes involved in these processes—such as B-cell surface receptors and cytokines—could be elevated prior to NHL diagnosis.

1.4.2. The NFKB Pathway

Nuclear factor kappa B (NF-KB) is a Rel-family transcription factor active in numerous inflammation-related signaling processes and pathways. The NF-KB pathway is well established as a positive mediator of both B- and T-cell development, proliferation and survival; proper functioning of this pathway is required to elicit immune response, but constitutive activation of the NF-KB signaling cascade is widely observed in B-cell non-Hodgkin lymphomas. This in turn results in aberrant lymphoma cell cycling and inhibition of apoptosis, a hallmark of cancer and a key part of AIDS-NHL lymphomagenesis.

The association of two genes in this pathway with HIV/AIDS-related NHL is of interest. The first of these genes, NFKBIA (nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, alpha; 14q13) codes for the alpha subunit of the IKK complex (IκBα), which is upregulated by NF-KB and inhibits the NF-KB pathway. The NFKBIA SNPs rs696 and rs8904 were assessed in a case series of Hodgkin lymphoma patients by Lake *et al.*[38], who detected these SNPs only in non-EBV HL cases, and in a case-control study by Chang *et al.*[39], who found positive associations with HL risk for homozygous variant allele carriers of rs696 and rs8904, and a negative association with heterozygous variant allele status for rs1050851. Du *et al.*[40] assessed the association of these SNPs with multiple myeloma risk among Han Chinese persons, but found no association.

The second of these genes, IKBKAP (inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase complex-associated protein), codes for the IKAP protein, which is a scaffold protein that binds NIK and IKKs into an active kinase complex. The IKBKAP SNP rs2230793 has been linked to treatment-associated neuropathy in multiple myeloma[41] but not associated with risk of onset; rs3204145 has been associated with base-of-tongue squamous cell carcinoma risk[42].

1.4.3. The Cancer Stem Cell Paradigm

The cancer stem cell (CSC) framework originally held that tumorigenesis and metastasis could arise from small populations of highly proliferative cancer stem cells of a fixed phenotype, in contrast to arising from any given cell in a tumor. This conclusion was drawn on the basis of in vitro and in vivo assays in acute myeloid leukemia cells, in which surface markers distinguished AML stem cells from other AML cells with limited potential for proliferation. More recent scholarship has suggested that there is not simply a core population of phenotypically fixed CSCs, but rather that all cancer cells retain stem cell properties that are activated or de-activated in response to cellular or microenvironmental cues[43,44]. “Stemness” is therefore a fluid phenotypic state, as differentiated cells can “de-differentiate” and re-acquire the characteristic pluripotent stem cell phenotype key to tumorigenesis and metastasis.

Controversies remain, but there is reason to think that the cancer stem-cell paradigm is relevant to lymphoma etiology as well. For instance, reactive hematopoiesis—the increased generation of blood cell precursors, and their subsequent differentiation—is a well-characterized aspect of the immune response. HSCs have been implicated directly in the pathogenesis of CLL[45], with the etiologically relevant biology defined by interactions between lymphoma cells, the tumor microenvironment, and the stem cell niche—interactions in which the Notch, Wnt, NFKB and stem-cell pathways are implicated. Additionally, the transcriptional repressor CTBP2, discussed

below, interacts with a number of stem-cell maintenance factors, and its expression appears to correlate with tumor initiation and progression via action on c-Myc signaling[46,47]. C-Myc has been identified as a key factor in stem cell homeostasis; in B-cell lymphoma cell lines, c-Myc was found to be activated by stromal signals, leading to HDAC6 upregulation and consequent increases in cell survival and lymphoma progression.

1.4.3.1. Genes Implicated in "Stemness" Phenotype

We focus on five genes active in cell "stemness" and pluripotency. OCT4/POU5F1 overexpression is sufficient to induce pluripotency, and switching off OCT4/POU5F1 expression is sufficient for loss of pluripotency and inducement of stem cell differentiation[48]. Overexpression of REX1/ZPF42, together with OCT4/POU5F1, is characteristic of embryonic stem cells and has been observed to predict transformation of follicular lymphoma to more aggressive DLBCL[49,50]. EPCAM codes for a cellular adhesion molecule that has been reported as a marker of "stemness" in hepatocellular and breast carcinoma, and that also engages Wnt signaling[51]. GLI1 plays a role in stem cell proliferation, suggesting that once CSCs have acquired the stemness phenotype, GLI1 may aid in tumorigenesis and/or metastasis; GLI1 is also a Hedgehog signaling effector that activates the Wnt pathway downstream. CTBP2 interacts with a number of stem-cell maintenance factors[46], and its expression appears to correlate with tumor initiation and progression via action on c-Myc signaling[47].

1.4.4. The Wnt/ ß-Catenin Pathway

The Wnt pathway plays a well-documented role in both embryogenesis and carcinogenesis; ß-catenin is a cadherin-associated protein responsible for cell growth and adhesion of epithelial cell layers in tissues. Wnt activation leads to greater accumulation of β-catenin in the nucleus and activation of downstream pathways, in which c-Myc and cancer-related target genes figure

prominently. Studies have shown that approximately 20% of gastric cancers contain ß-catenin mutations and that nuclear accumulation of ß-catenin is associated with a poorer prognosis[52]. Activating mutations in the Wnt pathway have been found in several cancer types including colon, prostate and ovarian cancers, lymphoblastic leukemia, and medulloblastomas[53,54]. Recently, FZD3 expression was found to be increased in esophageal cancer but not in normal esophageal tissue[55]. Wnt-β catenin signaling may also play an important role in HIV replication within astrocytes of the central nervous system[56], which conceivably could bear on the etiology of primary central nervous system lymphoma (PCNSL). Given this, and given the activity of Wnt- β catenin signaling in various cancers, and in hematopoietic and intestinal epithelial stem cells, the Wnt pathway may also impact AIDS-NHL[57].

Eleven genes active along the canonical WNT/ß-catenin signaling pathway are of particular interest: WNT2, WNT2B, WNT8A, FZD1, FZD3, CTNNB1, DVL2, AXIN1, AXIN2, PPARGC1A and TCF7L1. FZD 1 and 3 code for frizzled proteins, G-protein coupled receptors for which the Wnt proteins are ligands. This binding activates the WNT/ß-catenin pathway, which ultimately results in the transcriptional coactivator ß-catenin (encoded by CTNNB1) forming a complex with CREB-binding protein (CBP) that can induce transcription of dozens of genes downstream, including many of relevance to cancer (e.g. *c-MYC*, translocation of which is characteristic of Burkitt lymphoma). PPARGC1A interacts with CBP; DVL2 phosphorylates ß-catenin and marks it for proteosomal degradation by a destruction complex that includes the Axin protein. TCF7L1 activates the Wnt pathway in the presence of β-catenin, and acts as a repressor of the pathway in the absence of β-catenin.

1.4.5. The Notch Signaling Pathway

Notch signaling is triggered by the binding of Jagged (JAG) or Delta (DLL) ligands with Notch (NOTCH) receptors, which results in the release of the Notch intracellular domain (NICD) and activation of various target genes and downstream pathways. Notch signaling plays a prominent role in tumorigenesis, embryonic development and cell differentiation, including CD4+ and CD8+ lineage commitment. Its role in cancer appears to be cancer-and cell-type specific, with both tumor-suppressor and oncogene effects reported: for instance, >80% of T-ALL patients have Notch1 mutations, suggesting an oncogenic role in T-ALL; however, evidence suggests a tumor-suppressor role in B-ALL[58]. Seven genes coding for Notch receptors, ligands, and signaling targets are of particular interest here: Notch 3 and 4 (receptors); DLL and JAG2 (ligands); and HES2, HEY1 and HEY2, which are basic helix-loop-helix transcription factors of the Hairy family targeted by Notch signaling.

1.4.6. Conclusions: Common Ground between Inflammatory and Stem-Cell Processes

The separate categorization of stem-cell and inflammation pathways in NHL lymphomagenesis is largely artificial; it is in fact the considerable degree of overlap between them that motivates this study. For instance, OCT4 and REX1 are essential to stem-cell pluripotency, while Wnt/β-catenin and Notch-signaling pathways regulate self-renewal, proliferation, and differentiation of stem cells—e.g., via the action of β-catenin on OCT4[59]. Wnt and Notch also interact: NOTCH3 can be activated by Wnt signaling[60], and Notch proteins can regulate Wnt signaling via the interaction of β-catenin with membrane-bound Notch, leading to degradation of β-catenin and reduced downstream Wnt signaling[58]. Furthermore, a recent paper reports that co-activation of β-catenin and NFKB is needed to induce the stemness phenotype in breast cancer cells[61],

suggesting that there remains common ground to be explored between stem-cell and inflammation pathways in AIDS-NHL.

1.5. Gaps in the Literature; Rationale for the Dissertation

This dissertation comprises three aims. All three aim to investigate genetic risk factors for AIDS-related NHL; each uses one of three very different, but complementary tools for this investigation. Together, the strengths of each approach balance the shortcomings of the others; none is objectively superior to the other, and all three are integral to the dissertation.

To begin, the association between family history of NHL and risk of incident NHL suggests a genetic component to NHL risk[21]; the association between inflammatory processes and NHL suggests that variation in immune-related genes may be especially important. Therefore, we start with a targeted approach: a candidate-gene study highlighting key pathways in immune and stem-cell processes. Recognizing that our knowledge is imperfect, and that important signals of association may lie outside our set of candidate genes, we expand our perspective to cover the entire genome, using a genome-wide association study. However, this expansion comes at the cost of statistical power, and also tractability: it is challenging to make sense of millions of signals of association simultaneously. Furthermore, single-SNP approaches can be ill-suited to the study of complex biological phenomena such as inflammation and lymphomagenesis[62,63]; instead, mapping SNPs to genes, and genes to pathways, may yield results that better capture the relevant biology. To improve power and interpretability, and to balance the shortcomings of single-SNP approaches, we use a pathway analysis approach, which distills the huge number of features in GWAS output for better power and interpretability. Each is discussed in turn below.

1.5.1. Candidate-Gene Study

The biological rationale for our candidate-gene study is described in Section 1.4: Wnt, Notch, stem-cell and NFKB pathways are closely linked in stem-cell and inflammation processes, and have received some attention in NHL. However, published results have not examined these pathways using a candidate-gene approach in AIDS-NHL.

The technical rationale for choosing to perform a candidate-gene study hinges on efficiency. Candidate-gene studies employ a targeted approach to the investigation of genetic risk factors, by selecting a limited set of SNPs on important genes defined using prior knowledge or functional prediction. This targeted approach makes for efficient use of financial, biological, and technical resources.

However, successful candidate-gene studies are contingent upon the accuracy of the SNP- and gene-selection process: though targets may appear promising on the basis of prior knowledge, there is still a chance that studies will not yield meaningful associations.

1.5.2. Genome-Wide Association Study

Relative to candidate-gene studies, the major advantage of a genome-wide association study is therefore its agnostic approach: successful identification of important SNP-phenotype associations is not contingent on the set of SNPs chosen for analysis. However, this agnostic approach also gives rise to the major disadvantage of GWAS relative to candidate-gene studies: greatly reduced power. With millions of multiple comparisons in a GWAS, compared to dozens in a candidate-gene study, the penalty for multiple comparisons is much higher, and many results may be false-positives.

To improve analytic power, the National Cancer Institute has since 2012 published four meta-analyses that combine previously-published GWAS of NHL in the general population. These papers have identified loci in the human leukocyte antigen (HLA) region on chromosome 6 for

marginal zone lymphoma (MZL, a slow-growing lymphoma not often seen in PLWHA that can, however, develop into DLBCL)[64], and loci both inside and outside the HLA region for DLBCL[65,66], follicular lymphoma (FL)[67,68], and chronic lymphocytic leukemia (CLL; though called leukemia, this is a stage of small lymphocytic B-cell lymphoma)[69]. However, with the exception of DLBCL, these subtypes are not common in HIV+ individuals, and no GWAS results specific to AIDS-NHL have yet been published.

The HLA region contains multiple genes active in immune response; this region has also shown associations with host control of HIV infection[70,71]. That associations between the HLA region and NHL in the general population would be observed is sensible, given common etiologic associations with immune and B-cell activity across NHL subtypes. We would also expect these associations to be reproduced in the context of AIDS-NHL.

A GWAS specific to AIDS-NHL may uncover novel associations outside the HLA region. This would include genetic variation related to HIV protein processing; given the association of Epstein-Barr virus with Burkitt lymphoma and PCNSL (not assessed in NCI studies), genetic variation related to EBV processing (e.g. polymorphisms in IL-10[72,73]) could also be prominent. Polymorphisms in genes active in innate immune response to bacterial infection may also be especially important in the context of HIV, owing to microbial translocation from the gut and subsequent chronic inflammation[74]. As the distribution of subtypes in our sample differs from that in the NCI studies (our sample includes ~30% PCNSL cases and ~30% DLBCL cases), we may also expect to see heterogeneity in relevant polymorphisms relative to CLL, MZL, and FL.

Gene-environment interactions with risk factors for NHL in the general population (e.g. BMI, age) may also differ in the case of AIDS-NHL; however, we do not have sufficient power to investigate gene-environment interactions.

With this heterogeneity in subtypes and the small sample of our study relative to these meta-analyses, we should of course not expect to see any results approaching the level of significance found in these NCI meta-analyses. However, three additional considerations argue for the value of this GWAS.

First, GWAS can identify promising regions, or even individual SNPs, for targeted follow-up, which could occur via second-stage designs, resequencing and fine-mapping of regions of interest, or even in laboratory experiments measuring the impact of induced genetic changes on expression and epigenetic phenomena.

Second, this GWAS could serve as a component of future meta-analyses such as those cited above. With regard to meta-analyses, pooling has become the norm in GWAS, and collaboration is key. Such pooling is essential for the identification of rare variants, which occur at frequencies too low for proper analysis in individual studies. It is also essential for the investigation of rare phenotypes such as NHL, where individual studies can suffer from low power resulting from a small number of cases[75].

Third, these GWAS results provide the input for our final aim: a series of pathway analyses investigating gene- and pathway-level associations with risk of NHL, using summary statistics from single-SNP associations in the GWAS.

1.5.3. Pathway Analyses

Given that SNPs in the pathways described in aim 1 have not been investigated using a candidate-gene approach, analysis of these SNPs has not been bolstered using pathway analyses. Given that no GWAS of AIDS-NHL has yet been published, the same holds for our GWAS.

We use pathway analysis to overcome three limitations of the GWAS: low power, challenges in interpreting results for millions of SNPs, and—in common with the candidate-gene study—the challenge of using single-SNP associations to capture complex biology[62,63]. Pathway analyses have become an increasingly common method for the analysis of GWAS; by combining SNPs into genes, and genes into biological pathways, and then analyzing the association between genes and pathways with a phenotype of interest rather than a single, isolated SNP, they allow for investigation of individual SNPs within a context of shared genomic location and function[63]. Thus they can highlight entourage effects and illustrate the combined impact of multiple SNPs that in isolation might have minimal effects.

Taken together, we see that the candidate-gene study compensates for the GWAS's lack of power; the GWAS compensates for the candidate-gene study's potential to miss important SNPs when defining target sets; and the pathway analysis compensates for the GWAS's lack of power, its difficulties with interpretability, and the limitations of single-SNP approaches common to both the candidate-gene study and the GWAS. However, there is no pathway analysis without GWAS data: the three methods therefore constitute a unified approach; each informs the other, and each is valuable both intrinsically and as part of a combined analytic effort.

Table 1.1. Proportion of Selected NHL Subtypes among All NHL Cases in PLWHA and General Population, United States, 1992-2009 (SEER Data and CDC HIV Surveillance Data)[2]

Qualitative distribution	NHL subtype	Prevalence in PLWHA	Prevalence in HIV-population
Higher in PLWHA	DLBCL (all types)	45.5%	33.6%
	Burkitt Lymphoma	8.4%	1.4%
	NOS	37.2%	17.6%
Lower in PLWHA	Follicular lymphoma	2.4%	17.9%
	Marginal zone lymphoma	1.4%	7.8%
	Small lymphocytic lymphoma	0.8%	6.9%
	Peripheral T-cell lymphoma	2.8%	5.4%
	Mantle cell lymphoma	0.4%	3.3%
	Mycosis fungoides	NA	3.2%

CHAPTER 2 : CANDIDATE-GENE STUDY

2.1. Research Objectives and Methodology

We seek to identify single-nucleotide polymorphisms (SNPs) that, if associated with NHL risk, would serve either to generate novel hypotheses or confirm prior investigations regarding the role of inflammatory and stem cell-related processes in lymphomagenesis. After genotypic quality control, we use conditional logistic regression and semi-Bayes-corrected Cox proportional hazards regression (which estimates the same conditional likelihood as conditional logistic regression) to examine 30 SNPs in 24 genes along pathways including Wnt, Notch, NFKB and stem cell pluripotency-related pathways, as summarized below in Tables 2.1-2.4.

2.2 Specific Aims and Hypotheses

Using a candidate-gene approach, we aim to:

1. Genotype AIDS-NHL cases and HIV+ controls for SNPs in genes discussed in Sections 1.4.-1.5, and summarized below in Tables 2.1-2.4;
2. Determine whether the distribution of these SNPs differs between HIV+ NHL cases and controls.

The majority of the SNPs that we included in this study have not been well-characterized, making prediction of their functional consequences, and hence the direction of potential associations with NHL risk, difficult to ascertain. Nevertheless, Sections 1.4.-1.5 showed that the genes we investigate have clear biological relevance, and SNPs in these genes could impact risk of NHL. Therefore we hypothesize the following:

a) SNPs in the **NFKBIA** and **IKBKAP** genes, which repress the NFKB pathway, will be associated with risk of NHL. Constitutive activation of the NFKB pathway is an etiologic

factor in NHL lymphomagenesis; SNPs with functional impacts on NFKB repressors should therefore be associated with NHL risk.

b) SNPs in genes playing a role in degradation of β-catenin (**AXIN1, AXIN2, DVL2**) will be associated with NHL risk, owing to impact on β-catenin signaling and thus on Wnt signaling and the activation of pathways downstream from Wnt.

c) Notch signaling has shown both oncogenic and tumor-suppressor activity; we hypothesize that all SNPs involved in Notch signaling—**NOTCH3, NOTCH4, DLL1, JAG2, HES2, HEY1,** and **HEY2**—will, by impacting this signaling, be associated with AIDS-NHL risk.

d) SNPs in genes coding for Wnt-family ligands (**WNT2, WNT8A**), their receptors (**FZD1, FZD3**), associated transcription factors (**TCF7L1)** or binding sites for micro-RNA upregulators of Wnt signaling (on **WNT2B, CTNNB1** and **PPARCG1A**) will be associated with risk of NHL via disruption of β-catenin signaling and consequent activation of the Wnt pathway.

e) SNPs in **REX1/ZPF42, OCT4/POU5F1, GLI1** and **CTBP2** will be associated with risk of NHL, by virtue of impacts on c-Myc signaling, acquisition/re-acquisition of the "stemness" phenotype, and stem-cell proliferation.

2.3. Study Population

The MACS is an ongoing longitudinal study of health and behavior in men who have sex with men (MSM). It was begun in 1983 with sites in Baltimore, Los Angeles, Chicago, and Pittsburgh; since 1983, it has enrolled almost 7000 participants for twice-yearly clinic visits at which a range of detailed clinical and behavioral indicators is assessed[76]. All participants provided written informed consent; questionnaires and protocols (available at

http://statepi.jhsph.edu/macs/forms.html) are approved by site-specific institutional review boards.

2.3.1. Case and Control Selection Criteria

a. Cases: All MACS participants who developed AIDS-NHL prior to July 2010, either as their first AIDS-defining-illness or following another AIDS-defining illness, for whom biological samples were available for DNA extraction and at least one matched control could be identified (n=185). After all matching and quality-control procedures were completed, 178 of these 185 cases were retained for inclusion in analyses.

b. Controls: For each case, up to three HIV+ controls were selected randomly from risk sets based on the following matching variables: recruitment year (84-85 vs 87-91), serostatus (seroprevalent at baseline vs. seroconversion post-baseline), duration of HIV+ follow-up time (i.e. controls must be followed at least as long as the cases), race (white or non-white), and CD4+ -T cell count (categories: 0-49 mm^3, 50-99 mm^3, 100-199 mm^3, 200-349 mm^3, 350-499 mm^3, and 500 mm^3 and above). The time point at which CD4+ count was measured in cases was the last measurement before NHL diagnosis; the time point for CD4+ count in controls was the follow-up matched time-point. After all matching and quality-control procedures were completed, 522 controls were retained for inclusion in analyses.

2.3.2. Reference Date and Matching Criteria

All cases (n=178) and controls (n=522) were HIV-positive at reference date. In cases, reference date was the date of NHL diagnosis; in controls, reference date and length of follow-up time were determined as the interval in days between first HIV-positive study visit and the NHL diagnosis date of the case to which they were matched. Cases were matched on cohort (1984-1985 recruitment or 1987-1991 recruitment); length of HIV-positive cancer-negative follow-up

time; race (white/nonwhite); seroprevalent/seronegative status at baseline, and CD4+ T-cell count at reference date.

2.4. Data Collection

2.4.1. Genotypic Data

DNA was extracted from blood mononuclear cells received from the MACS specimen repository, as described in Dr. Hussain's approved MACS concept r0824, "Nucleotide Variation in DNA repair and B Cell Activation Pathways and AIDS-Associated Non-Hodgkin's B cell Lymphoma Risk." The study population is described above in Section 2.3. Genotyping was performed with a customized Fluidigm Dynamic 96.96 Array™ Assay at UCLA Genetics Laboratory. Genotyping is described in more detail in Section 2.6.

2.4.2. Covariate Data

All clinical, demographic and exposure data were collected using standard MACS protocols from 1984-2010.

2.5. Candidate Gene and SNP Selection Criteria

2.5.1. Candidate Gene Selection

Genes were originally selected as part of a multi-site cancer study in China investigating genetic and environmental risk factors for esophageal, lung, gastric, liver, and head and neck cancers. This study broadens the scope of prior investigations to include AIDS-related lymphoma. Given the canonical role of Wnt, Notch, inflammatory and stem-cell pathways in carcinogenesis, these pathways were prioritized for study; prominent genes in these pathways were identified to narrow the process of SNP selection. The biological rationale for their inclusion is explained above in Sections 1.4.1—1.4.5.

2.5.2. Candidate SNP Selection

Once important genes in Wnt, Notch, stem-cell and inflammatory pathways were identified, SNPs with a variant allele frequency >5% in the HapMap Han Chinese population lying on or near these genes were selected based on either location/genomic context (e.g. 3'UTR or missense polymorphism), prior associations with cancer in the literature, or both: the original study for which these SNPs were selected aimed more to identify novel SNPs than to replicate previously-published associations. Biological rationale for inclusion of these SNPs, and further details on the SNPs themselves, are given in tables 2.1—2.4.

2.6. Genotyping

Genotyping was performed by the UCLA Genetics Laboratory using a customized Fluidigm Dynamic 96.96 Array™ Assay[77]. Fluidigm chips use "integrated fluidic circuits," or IFCs. These consist of two layers of elastomer rubber bound together and grooved by multiple intersecting microchannels such that when pressurized liquid is fed through them, channel intersections act as valves to regulate the flow of liquid through the chip[78].

Target locations were confirmed using locus-specific primer sequences provided by Fluidigm; assays used allele-specific PCR chemistry to detect SNPs. Uniform fluorescence was ensured through use of a universal probe set in all reactions. Each plate was organized to include specimens from both cases and controls. Two or three positive laboratory controls (consisting of DNA samples purchased from the Coriell Repository) for each genotype as well as negative controls (reagent mix with no DNA) were included in each reaction plate. Replicate quality control (QC) DNA aliquots, which constitute approximately 1% of the specimens, were also distributed throughout the reaction plates. Laboratory staff were blinded to all identifiers and research information about the samples, including the identities of the QC replicates.

2.7. Laboratory and Data Quality Control Measures; Post-genotyping SNP and Sample Inclusion and Exclusion Criteria

Quality control procedures were as follows:

1) Initial measures: Ensure that blank wells on the assay plates containing no DNA were called as such;

2) Coriell sample concordance: Ensure concordance between Coriell DNA samples included on all assay plates with known genotypes, based on HapMap data;

3) Call rate determination: assess proportion of non-missing calls for each individual SNP, and drop SNP if <95%;

4) Sample coverage: assess proportion of non-missing SNP calls for each individual participant, and exclude participants with calls missing for >10% of SNPs (n ≥3 SNPs);

5) Hardy-Weinberg equilibrium: assess HWE among controls, and exclude SNPs with Bonferroni-corrected HWE p-value < 0.05/34 = 0.0015;

6) Linkage disequilibrium: assess LD using SAS PROC ALLELE; if any pair of SNPs is in high LD ($R^2 > 0.8$), drop one randomly; and

7) Duplicate sample concordance: assess concordance between original genotyping run and a duplicate run for 5% (n=53) of participants (still including any participants excluded for low sample coverage); exclude SNPs for which genotype calls differed in ≥5% (n≥3) of these 53 participants. Do not count missing calls on one plate and a successful call on another as discordant.

2.7.1. Initial Measures

Blank wells (one each per plate) were confirmed as “no calls.” Base calls and allele frequencies in Fluidigm output were compared to dbSNP to ensure that Fluidigm output was in line with expected results; see Tables 2.5—2.7 (allele frequencies).

2.7.2. Coriell Sample Concordance

Coriell sample concordance is used to ensure that the platform is actually genotyping a given sample correctly: a known sample (purchased from the Coriell repository) is used which has been genotyped as part of the HAPMAP project (which we considered the gold standard genotype), and our results are compared to those of the HAPMAP. If observed output differs from the expected output, this indicates a problem with the genotyping plate; if multiple plates fail concordance, this suggests a problem with the platform. Two or three positive Coriell controls were included in each reaction plate. A cutoff of <95% Coriell concordance (i.e. across all 12 plates, there was discordance between Coriell and genotyped results >5% of the time) was used. One hundred percent of calls for Coriell samples in our study, across all 12 plates, were concordant with the HapMap sequence.

2.7.3. Call Rate Determination

Call rate was determined for each SNP as the proportion of samples genotyped successfully among all 716 participants (e.g. if a SNP failed to be genotyped for 72 participants, its call rate would be 90%). Any SNP with <95% call rate was excluded. SNP rs9972231 (on gene JAG2, in the Notch pathway) had a call rate of 91.2% and was therefore excluded; all others passed.

2.7.4. Sample Coverage

After excluding rs9972231 due to low call rate, sample coverage was calculated for each participant as the proportion of SNPs successfully genotyped out of our remaining 33 SNPs of

interest. Participants were excluded if >10% of SNPs (i.e. more than 3 SNPs) were missing calls. Seven participants (three cases, four controls) were excluded for insufficient sample coverage (<90%).

2.7.5. Hardy-Weinberg Equilibrium

SNP rs8904 was found to be out of HWE (exact p = 0.0008, below our threshold of 0.05/34 = 0.0015) and excluded (it also failed concordance, with 13% discordance). It bears noting that two other SNPs had p >0.0015 but < 0.05: rs3815188 (NOTCH3; p = 0.0065) and rs2273368 (WNT2B; p = 0.038). SNP rs2273368 was ultimately excluded due to insufficient concordance (<95%) between original and replicate plates, described in Section 2.7.7.

2.7.6. Linkage Disequilibrium

SAS PROC ALLELE was run on seropositive controls only (n=529) and revealed two pairs of SNPs in high LD: rs696/rs8904 (r^2 = 0.846; D' = 0.9952) and rs3204145/rs1538660 (r^2 = 0.944; D' = 1). SNP rs3204145 was randomly chosen to be excluded from the rs3204145/rs1538660 pair; rs8904 was already excluded on the basis of HWE, so rs696 was retained.

2.7.7. Duplicate Sample Concordance

Five percent of samples were chosen at random for re-genotyping (n=53/1066), to ensure reproducibility and accuracy of genotypes. SNPs with >5% discordance (i.e. with different calls between the original and replicate plates among 3 or more participants, not including missing calls), were dropped, as were samples with >5% discordance (i.e. participants with different calls for three or more SNPs, again not including missing calls). Of these 53 samples, two were already dropped for insufficient sample coverage; these individuals were retained in concordance testing to provide conservative estimates. On our replicate plate, 14 of 34 SNPs of interest were

duplicated with 100% concordance; 10 of 34 had a missing call on one plate and a successful call on the other; and another 10 of 34 had discordant reads between the original and replicate plates.

One participant was dropped for discordant genotype on four SNPs, and one SNP was dropped for excessive concordance failure. Overall, three SNPs (rs9972231, rs8904, and rs2273368) failed concordance criteria, but rs9972231 was already dropped due to low call rate, and rs8904 was already dropped due to HWE. Thus only rs2273368 was dropped due to poor concordance.

2.7.8. QC Conclusions and Summary of Data for Analysis

After excluding one SNP for call rate failure (rs9972231), two SNPs for LD failure (rs3204145 and rs8904, the latter of which was also found to be out of Hardy-Weinberg equilibrium and failed replicate plate concordance), one SNP for replicate plate concordance failure (rs2273368), seven samples for coverage <90% (i.e. genotyping failure for >10% of SNPs), and one sample for concordance failure, QC procedures left us with data on 30 SNPs in 708 participants (180 cases, 528 controls), constituting 168 3:1 matched sets, ten 2:1 matched sets, two 1:1 matched sets, and two unmatched cases, for a total of 180 matched sets. Noting that two of our 180 cases had no controls, and that there were 182 unique values for r0824_set, it was discovered that sets 32 and 67 had no cases. Members of these sets, as well as the two cases with no matched controls, were therefore excluded from the analysis, leaving 178 cases and 522 controls in 168 3:1 matched sets, eight 2:1 matched sets, and two 1:1 matched sets, for a total of 178 matched sets and 700 participants.

2.8. Statistical Analysis

2.8.1. SNP and Allele Frequencies

Post quality-control allele and genotype frequencies are shown in tables 2.5—2.7, according to case-control status, for: 1) non-white participants; 2) white participants; and 3) all participants, along with dbSNP global variant allele frequency. A1 is the reference allele listed in dbSNP (after any strand flips to match the orientation used for genotyping in this study), and is not necessarily the allele with lower frequency in our population. A2 represents the variant allele. "dbSNP MAF" is the global variant allele frequency in dbSNP.

2.8.2. Description of Study Variables

2.8.2.1. Exposure Variables

Our exposures of interest are 30 SNPs on 24 genes in the Wnt, Notch, NFKB, and stem-cell pathways described above.

2.8.2.2. Outcome Variables

Our outcome of interest is non-Hodgkin lymphoma, with pathological confirmation through state-level registries and pathology reports, or, in the case of NHL diagnosed post-mortem, from autopsy reports.

2.8.2.3. Covariate Data: Substance Use and Demographics

NHL pathogenesis has been observed to begin three years prior to diagnosis[79-81]. Therefore, to avoid measurement error and spurious associations between substance use and NHL risk, only substance use occurring three or more years prior to diagnosis should be modelled, and covariates should be measured *at least* three years prior to diagnosis, i.e. in the six-month interval preceding the timepoint equal to date of diagnosis minus three years. Following Chao *et al.*[81], all substance use covariates described below were measured at this timepoint.

Six variables were ultimately used to assess substance use, out of an initial set of 18. Tobacco use was assessed as never/former/current smoker status at three-year lag; alcohol consumption was assessed as weekly/monthly/yearly use; cocaine, cannabis, and uppers/methamphetamine were assessed simply as any use (reported by participants as yes/no) in the six-month interval preceding the study visit at which responses were collected. Table 2.8 summarizes key demographic characteristics of the study population, and includes these substance use covariates.

As shown in Table 2.8, median age at reference date was 40.4 years for controls (n=522), and 41.8 years for cases (n=178). Median HIV viral load at set point, where available, was 20,730 RNA copies/mL in controls (range =300—672,810; SD=83386.23; missing for n=84 [16.1%] participants), and 31,090.50 RNA copies/mL in cases (range=400—960960; SD=127833.15; missing for n=44 [24.7%] of participants). Median CD4+ T-cell count at date of matching was 88.5 cells/mm^3 in controls (range=3.0—1361.0; SD=225.42; missing for zero participants) and 81.0 cells/mm^3 in cases (range=2.0—923.0; SD=212.43; missing for zero participants). The majority of participants were white non-Hispanic (n=472 [90.4%] controls; n=150 [84.3%] cases); were recruited during 1984-1985 (n=469 [89.8%] controls; n=159 [89.3%] cases); were seroprevalent at baseline (n=471 [90.2%] control; n=161 [90.4%] cases); and had no exposure to HAART prior to reference date (n=476 [91.2%] controls; n=161 [94.4%] cases).

Among participants with available tobacco smoking status data three years prior to reference date, n=132 (25.3%) controls and n=48 (27.0%) cases were current smokers; n=146 (28.0%) controls and n=49 (27.5%) cases were former smokers; n=148 (28.4%) controls and n=47 (26.4%) cases were never smokers. Among controls, 124 (23.8%) participants reported drinking alcohol 1-2 times per week; n=107 (20.5%) reported drinking ≤ 1-2 times per week, and 129 (24.7%) reported drinking ≥ 3-4 times per week. Data were unavailable for n=91 (17.4%)

controls. Among cases, n=42 (23.6%) reported drinking 1-2 times per week; n=49 (27.5%) reported drinking ≤ 1-2 times per week, and n=37 (20.8%) reported drinking ≥ 3-4 times per week. Data were unavailable for n=32 (18.0%) cases.

At three years prior to reference date, n=97 (18.6%) controls reported crack or other cocaine in the preceding six-month interval preceding reference visit, vs. n=18 (10.1%) cases; n=40 (7.7%) controls reported uppers/methamphetamine use, vs. n=15 (8.4%) cases; n=2 (0.4%) controls reported heroin or other opiate use, vs. n=1 case (0.6%); n=233 (44.6%) controls reported cannabis consumption, vs. n=33 (18.5%) cases.

Table 2.8 also shows that substance use data were unavailable for a high proportion of participants. This is in part a result of coding exposures using a three-year lag: for some participants, the interval between baseline visit and reference date was less than three years, so data for our timepoint of interest were unobserved. As discussed in Section 2.8.6, we performed multiple imputation to fill in any missing values for these variables at the analysis stage.

2.8.3. Covariate Selection

A priori, based on knowledge of the literature and HIV/NHL biology, the following covariates were included in our final models: HIV viral load at set point[82], median-value-imputed where missing; AIDS diagnosis prior to reference date[7]; HAART prior to reference date[7,14]; age (continuous) at reference date[83]; ever/never hepatitis C status three years prior to reference date[84-86]; and self-reported race. CD4 cell count at matching was included in initial models, but in contrast to continuous values for other matching variables, had no appreciable impact on estimates. For the sake of parsimony, and out of concern for precision given our rather small sample size (n=700), CD4 at matching was therefore not included in the model. HAART prior to reference date and any ART prior to reference date were each investigated; HAART was more

strongly associated with NHL than was "any ART" use. Assuming that "any ART" would include HAART and that using both would be redundant, HAART was chosen for its stronger association with NHL risk. Prior AIDS diagnosis, age at reference date, self-reported race, and HCV status all showed independent associations with NHL risk, and their inclusion in models led to appreciable differences in NHL-SNP associations.

Viral load at set point is an important covariate: higher viral loads at set point correlate with poorer HIV/AIDS prognosis[82,87]. When individuals seroconvert, an initial period of high viremia is followed by an approximate equilibrium, or set point[88]. Viral load at set point therefore refers to the concentration of HIV RNA viral copies per mL of blood at this equilibrium. Because MACS includes both seroconverters and participants who were seropositive at baseline (and therefore have unobserved VL at set point), the concept remains the same, but the operationalization differs, for each group. Here, viral load at set point for participants seropositive at baseline, or before visits 3-4 and recruited in 1984-1985, was taken as the value at visit 3 or 4 (to reduce any measurement variability in the start-up phase of the MACS)[87]. For those seroconverting after visit 3-4, viral load at set point is the average viral load 12 to 24.5 months following seroconversion, with the interim slope for this period approximating zero (and thus reflecting an approximate equilibrium value).

Additional covariates assessing substance use, including tobacco, alcohol, cannabis, upper/methamphetamine, and cocaine use, were explored using model selection procedures as described in the next section.

Ultimately, three broad considerations motivated the choice of covariates in this study: 1) control of confounding; 2) the impact of additional non-confounding covariates on measures of SNP-

NHL association; and 3) the impact of missing data for covariates on the precision of our estimates under a matched design.

First, under the traditional definition of a confounder as a variable associated with both exposure (any of our 30 SNPs) and outcome (NHL), only race qualifies as a confounder: there is no other variable that could simultaneously influence the probability of having a given germline polymorphism and the risk of developing NHL. Therefore, estimates adjusted only for self-reported race need no further adjustment if our goal in covariate selection is to control or confounding.

Second, whether covariates should be included in genetic association models—especially logistic models assessing binary outcomes—is not always straightforward[89]. One concern is conditioning on intermediates variables lying on the path between exposure and outcome, which can introduce bias[90]. With just 30 SNPs on genes with reasonably well-defined biological functions, this was of minimal concern.

However, only under certain conditions will the inclusion of covariates in logistic models increase precision. Such inclusion can in fact reduce power by increasing the standard error of the estimate and the width of confidence intervals, despite any increase in the magnitude of the association[89,91]. Generally, cases and controls must be drawn from the general population (or the trait must be quantitative), and the prevalence of the disease under study must exceed ~20%[,72]. With n=178 cases and n=700 total participants, NHL prevalence in our case-control sample is 25.4%, and the addition of clinical covariates to a model adjusting for race alone did in fact increase precision (see Table 2.9), though the introduction of substance-use covariates to the model had the opposite effect. Independently of questions of precision, it can still be worthwhile to investigate certain covariates, especially if these are strongly associated with the outcome.

As an example, consider the extreme case in which exposure and covariate are independent, but the covariate is a necessary cause of the outcome: a SNP carries no effect in the absence of, say, AIDS diagnosis prior to reference date. In this case, though prior AIDS diagnosis not a confounder, we would indeed want to account for it in the model, since failing to do so could erase a meaningful—albeit stratum-specific—association. Furthermore, as described further below, we apply semi-Bayes correction to shrink the width of confidence intervals in multivariate models.

Third, as shown in Table 2.8., many covariates are missing a large proportion of data. This is especially problematic because we use matched sets in conditional logistic regression, requiring that if any one member of a set is missing values for a given covariate, then members of the set are dropped. This shrinks our sample size even further, decreasing precision and potentially outweighing any benefits from including additional covariates in our models. To account for this, missing viral load data were imputed using median value imputation, and substance-use data were imputed using multiple imputation, as discussed further in section 2.8.6.

2.8.4. Model Selection

Once covariates of interest with sufficient non-missing data have been identified on the basis of prior knowledge, model selection must be carried out. Any number of methods exist; these can be grouped crudely into automated tests (forward and backward selection), change-in-estimate (point estimate or confidence intervals) methods, and the use of prior knowledge based on a DAG. All three have benefits and drawbacks: automated tests have low power to detect true confounders, can produce absurd results, and can produce artificially narrow confidence intervals, overstating the precision of one's estimates; change-in-estimate procedures can be

laborious when dealing with multiple exposures; and DAGs are conditional on one's prior knowledge being accurate and the DAG being specified correctly[92,93].

The change-in-estimate procedure, generally informed by construction of a DAG, is viewed more favorably than stepwise methods by epidemiologists[92,94]. That said, when faced with many potential confounders, a backward selection approach informed by DAGs and prior knowledge could be sensible. Here we are faced with just one confounder (i.e. race), and the question is which variables, when added to the model, have a sufficient impact on estimates to justify inclusion and the loss in precision that can come with more covariates.

To answer this question, both stepwise and change-in-estimate procedures were carried out, but model selection was greatly complicated by the sheer number of exposure-outcome models under consideration. With four models of inheritance for each of 30 SNPs yielding 150 SNP coefficients (nominal heterozygous, nominal homozygous, log-additive, dominant, and recessive), some covariates met the change-in-estimate criterion or stepwise "significance of the coefficient" criterion in certain cases, but never all. We therefore ran six different sets of models, each using a different covariate modeling strategy; we present results from each of these models for all SNPs in Section 2.9 (Table 2.9) so that readers can evaluate for themselves the impact of different modeling strategies.

These strategies include: 1) models adjusting for self-reported race only; 2) models adjusting for self-reported race, age at reference (continuous), AIDS diagnosis prior to reference date, HAART prior to reference data, log HIV viral load at set point (median-value imputed where missing), and ever/never HCV status at least three years prior to reference date; 3) model 2 with semi-Bayes correction for multiple comparisons; 4) model 2, subset to the population of white-only participants (n=622); 5) a complete-case analysis (n=389) adjusting for covariates in model

2, plus tobacco, alcohol, cannabis, upper/methamphetamine, and cocaine use as described in Section 2.8.2.3; and 6) an analysis run following multiple imputation to fill in missing values for substance use data, and adjusting for all clinical and substance-use covariates assessed in previous models.

Again, models adjusting for race alone remain valid given the traditional definition of a confounder. Models with clinical covariates are included to give a broader picture of non-confounding—but nevertheless important—factors shaping AIDS-NHL risk, and semi-Bayes correction is applied to increase precision, as described in Section 2.8.7.

2.8.5. Statistical Analyses

After covariate selection and model selection as described above, odds ratios and 95% confidence intervals were estimated using conditional logistic regression in SAS PROC LOGISTIC, with Semi-Bayes correction implemented via SAS PROC PHREG.

2.8.6. Missing Data

As shown in Table 2.8 and discussed in Section 2.8.2.3, some covariates were missing a large proportion of data, especially substance use covariates. Conditional analysis in SAS requires that participants with missing data be dropped, which complicates the interpretation of results across models. Consider two models: Model A and Model B. Model A includes no covariates, and thus retains all participants. Model B includes several covariates, and as a result, participants missing these covariate data have been dropped from the analysis. To compare odds ratios from Model A and Model B is problematic, as the two models are effectively analyzing two different populations. Given this, one cannot claim that any change in estimates from one model relative to the other are a result of any covariate's influence: rather, this may simply reflect the impact of 1) smaller sample size, and 2) analyzing a different set of participants,

There are two broad possibilities for dealing with such a situation. The first, called a "complete case analysis," is to simply exclude all participants from Model A who are missing covariate data in Model B, thus ensuring that the same set of participants is analyzed in each model. The second is to impute missing covariate data in Model B, which allows for retention of all participants and enables accurate comparison of Models A and B and the detection of covariates' impact on estimates. Desai[95], van der Heijden[96] and Donders[97] show that bias can result from complete case analysis when data are not missing completely at random, and show that imputation is superior when this is the case.

2.8.6.1. HIV Viral Load Data: Median-Value Imputation

Viral load at set point was missing for participants in cohort two, as most participants in this cohort were seroprevalent at baseline and thus had an unobservable set point. Clearly, then, viral load data are not missing completely at random, complete case analysis is inappropriate, and imputation is therefore needed. To address this missingness, Peckham applied both median value imputation and MCMC-based multiple imputation to viral load data from the same set of participants, and found no difference in the estimates of association with NHL generated under each approach[98].

We also applied both median-value and multiple imputation; models reported in Section 2.9 used median-value-imputed HIV viral load. We acknowledge that viral load data are not missing completely at random—a standard assumption when using median value imputation—but rather are limited to cohort 2, and thus membership in this cohort is related to missingness. We further acknowledge that imputation using median values, as opposed to draws from a distribution that would better capture variability in viral load measures, can overstate precision. These caveats should be borne in mind when interpreting our results, but it should also be borne in mind that

estimates of association generating using the more complex multiple imputation model run by Peckham showed no difference from median-value models, suggesting that the impact of these factors is here minimal.

2.8.6.2. Substance Use Data: Multiple Imputation

Multiple imputation using fully conditional specification (FCS, also known as chained-equations) was performed for substance-use variables and HIV viral load in SAS PROC MI and PROC MIANALYZE. FCS allows for imputation of different classes of variables by using a separate type of regression for each (e.g. logistic for binary, linear for continuous), and importantly does not assume a multivariate normal distribution (important, since we are imputing categorical data), and can impute data with an arbitrary missing pattern. This is in contrast to MCMC, which assumes a multivariate normal distribution and is appropriate for continuous variables.

In multiple imputation, multiple datasets are created, each of which represents a draw from a distribution of possible values for the variable of interest, conditional on variables used as predictors in imputation model. Imputation proceeds in three stages. First, data are imputed, generating 100 separate datasets. Second, logistic regressions are run on each of the 100 imputed datasets. Third, parameter estimates from these 100 datasets are pooled to produce summary estimates. We ensured that results were sensible by comparing the frequencies of imputed variables' values in the complete-case and imputed settings. Odds ratios reported in Table 8.9/8.10 are drawn from these summary estimates and accurately reflect the variability and uncertainty associated with imputation.

A wide range of variables should be used to inform imputation—a wider range than one would use in an analytic model, including even the outcome variable itself[99]. These variables can be

chosen using prior knowledge, or within the data themselves (e.g. an $r^2 > 0.4$ with the variable to be imputed). Variables chosen to inform the imputation therefore include not only those in our analytic model (race, age at reference date, previous HAART, previous AIDS, HCV status, log HIV RNA at set point), but also NHL case status, cohort, cd4 count at matching date, and length of follow-up (since this is related to loss to follow-up, which is in turn related to non-response for substance use questions and missingness of these variables).

The proportion of missing values in our imputed variables ranged from 18% (cannabis consumption) to 32% (uppers/methamphetamine use). Based on simulation results from Graham[100] using a scenario with 30% missingness, we ran 100 imputations. Necessary assumptions for multiple imputation, including that 1) missingness is independent of the true unobserved value of the variable being imputed; and that 2) missingness is either totally random or can be predicted on the basis of observed covariates, are discussed in Section 2.11.2.

2.8.6.3. Hepatitis C Status: Coding for Three-Year Lag

One hundred five participants were originally missing data on hepatitis C status ≥3 years prior to reference date: the interval between baseline visit and reference date for these participants was less than three years, so data for our timepoint of interest were unobserved. Rather than using imputation, this was addressed as follows. First, we examined HCV status at the visit immediately prior to, or coincident with, reference date. If this value was 0, indicating that the participant was never HCV-seropositive, then this value would also be 0 three years prior to reference date, had it been observed. These participants (n=95) were therefore coded as 0 in our ever/never three-year-lagged binary variable. Second, HCV status at all MACS visits was examined for the remaining participants originally missing three-year-lagged data (n=10; median interval between baseline and reference date = 703.5 days; range 249-1028 days; SD = 340

days). If their value was either 3 (chronically infected) or 6 (cleared) at baseline, and this value did not change between baseline and reference date, then they were coded as a 1. This was indeed the case for all ten participants (i.e. none were HCV-seroconverting, acutely infected, or discordant), resolving the missing data issue for all 105 participants.

2.8.7. Correction for Multiple Comparisons: Semi-Bayesian Approach

Corrections for multiple comparisons such as the Bonferroni are criticized as both overly conservative and lacking coherent methodological justification. Many alternatives to the Bonferroni exist; one that has received recent interest in epidemiological circles is semi-Bayes correction via shrinkage[101,102,103]. Shrinkage "pulls" coefficients toward a prior value on the basis of their variance: coefficients with high variance are effectively viewed as more suspect and "pulled" toward this prior more than coefficients with low variance, thus "penalizing" estimates with high variance.

The choice of prior reflects the investigator's *a priori* expectations regarding the associations to be observed in the data. When using a null prior reflecting an investigator's expectation of null associations, semi-Bayes correction generally has the consequence of pulling the point estimate toward the null and reducing the width of confidence intervals, which simultaneously reduces false positive findings and increases precision. The core principle is that this pull toward the null provides correction for multiple comparisons, rather than (say) dividing a p-value by the number of comparisons as in the Bonferroni approach.

Methods for semi-Bayes correction in SAS are described in detail in[98,101-103]. Briefly, implementation is via a Cox proportional-hazards model, as the likelihood estimate from a Cox model happens to equal the conditional likelihood from a conditional logistic model. The Cox

model is constructed after re-scaling and re-centering variables for interpretability; Cox results are then checked against conditional logistic results to ensure that ORs and CIs are the same.

Next a prior dataset is constructed for data augmentation. In data augmentation, we create data records including set, SNP, and covariate data, merge these into our original dataset, and estimate associations using this combined dataset. Under a null prior, the OR for the SNP-NHL association is 1.0 (95%CI 0.25-4.0). Four new matched risk-sets with a weight of 400 each are then created, in which the exposure effect in exposed cases = 0.1, a reasonable assumption. Results from Cox models run on these merged datasets, with priors as described, are presented in the next section.

2.9. Results

Table 2.9 shows odds ratios and 95%CIs obtained using six different models. The first column, "Race Only," shows results from a conditional logistic regression model adjusting for self-reported race only (n=700 participants).

The second, "Adjusted for Clinical Covariates, no SB Correction (n=700)," uses conditional logistic regression and adjusts for: a) self-reported race; b) age (continuous) at reference date; c) AIDS diagnosis prior to reference date; d) HAART use prior to reference date; e) log HIV viral load (median-value imputed where missing) at set-point; and f) ever/never hepatitis C status three years prior to reference date. Column three, "Adjusted for Clinical Covariates, no SB Correction: White Only," presents results of the same model run among white participants only (n=622). Column four shows results from a Cox proportional-hazards model adjusting for these same six covariates after semi-Bayes correction, using the full sample of 700 participants.

Column 5 presents the results of a complete-case analysis (n=389) including substance-use covariates. Column 6 presents the results of an analysis adjusting for these same covariates after

multiple imputation to fill in missing values, enabling investigation of these factors in the full sample of n=700 participants. Because the point estimates for multiply-imputed data did not change appreciably relative to models adjusting for clinical covariates, we did not perform semi-Bayes correction on multiply-imputed data.

In Sections 2.9.1-2.9.4, we highlight results from semi-Bayes-corrected models, as these yielded the most precise estimates (i.e. the narrowest confidence intervals). These models also a) adjust for multiple comparisons, and b) account for important clinical covariates, in addition to race and age. They should therefore be taken as the most reliable results. Other models are presented primarily to illustrate the impact of different modeling strategies on results, but a special note of caution is warranted for the complete-case analysis.

Complete-case analyses, despite smaller sample sizes, yielded nominally significant results for some SNPs, including IKBKAP rs2230793, WNT2 rs4730775, and FZD1 rs3750145. On the whole, these results likely reflect bias from use of the complete-case design, and differential distribution of genotypes among participants retained in the complete-case analysis (n=389) relative to that in the full sample (n=700). Given the smaller sample size, there is also greater potential for these results to be false-positives. These results should therefore be discounted.

2.9.1. REX1 (ZFP42) rs6815391: Significant Association Following Semi-Bayes Correction

REX1 (ZPF42) was the only SNP for which a significant association with NHL risk persisted after semi-Bayes correction. This was observed under dominant (OR=0.68; 95%CI 0.47-0.99) as well as log-additive (OR=0.71, 95%CI 0.51-0.99) models of inheritance. A nominally significant association was also seen in the complete-case scenario adjusting for substance use covariates (OR=0.52; 95%CI 0.29-0.93); however, this association may be due to bias resulting from use of complete-case analysis, as discussed above.

2.9.2. AXIN2 rs2240308: Suggestive Association Following Semi-Bayes Correction

After semi-Bayes correction, a suggestive association between AXIN2 SNP rs2240308 and NHL risk was seen under a dominant model of inheritance (OR=1.47; 95%CI 0.96-2.25). Without semi-Bayes correction, nominally significant positive associations were seen for heterozygous A/G genotype in models adjusting exclusively for self-reported race under a nominal model (OR=1.69, 95%CI 1.08-2.64), and also under the dominant model (OR 1.63; 95%CI 1.07-2.49).

2.9.3. WNT2 rs4730775: Suggestive Association Following Semi-Bayes Correction

A suggestive association between WNT2 rs4730775 and risk of NHL was seen under the recessive model of inheritance following semi-Bayes correction (OR=1.47, 95%CI 0.97-2.22). Without semi-Bayes correction, nominally significant positive associations under the recessive model were also seen when adjusting for clinical covariates in the white-only subset (n=622) of the full sample (OR=1.74; 95%CI 1.10-2.75), and in the complete-case analysis (n=389) adjusting for substance use and clinical covariates (T/T homozygous genotype under the nominal model OR=2.42; 95%CI 1.05-5.59; log-additive model OR=1.53; 95%CI 1.02-2.31; dominant model OR=2.12; 95%CI 1.04-4.34). As above, complete-case analysis results should be viewed with skepticism.

2.9.4. WNT8A rs4835761: Suggestive Association Following Semi-Bayes Correction

Suggestive results for WNT8A rs4835761 were observed under the dominant model after semi-Bayes correction (OR=1.43, 95%CI 0.96-2.13). No models yielded nominally significant results.

2.10. Discussion

A significant inverse association between REX1/ZPF42 SNP rs6815391 (stem-cell pathway) and risk of NHL persisted after semi-Bayes correction. Following semi-Bayes correction, significant associations did not persist for any other SNPs, but suggestive positive associations with NHL

risk were seen for AXIN2 SNP rs2240308 (Wnt/β-catenin pathway), WNT2 SNP rs4730775 (Wnt/β-catenin pathway), and WNT8A rs4835761 (Wnt/β-catenin pathway). No SNPs in the NFKB or Notch pathways had suggestive or significant results. There is biological plausibility for these associations, both with regard to hematological malignancies *per se* and with regard to HIV. We consider each in turn below.

2.10.1. REX1/ZPF42 Inhibits Expression of p38 MAPK

REX1/ZPF42 SNP rs6815391 was the only SNP for which we observed significant inverse associations with NHL risk. A search of the NHGRI-EBI Catalog[104] of published genome-wide association studies found no published associations between rs6815391 and any phenotype. However, as discussed in Section 1.4.3.1, overexpression of REX1/ZPF42, together with OCT4/POU5F1, is characteristic of embryonic stem cells and has been observed to predict transformation of follicular lymphoma to more aggressive DLBCL[49,50].

In the particular case of HIV/AIDS-associated hematological malignancies, the role of REX1/ZPF42 in p38 MAPK regulation provides further biological plausibility for this inverse association. SNP rs6815391 is found within the 3' UTR of REX1/ZPF42, suggesting that any functional impact of this polymorphism would be related to regulation of REX1/ZPF42 expression. In mesenchymal stem cells, REX1 has been found to inhibit expression of p38 MAPK via direct suppression of MKK3[105]; in primary human monocytes, activation of p38 MAPK upregulates extracellular HIV Tat-induced transcription of IL-10, an anti-inflammatory cytokine[73,106]. Notably, elevated serum levels of IL-10 have been observed in AIDS-NHL patients[80]. Activation of p38 MAPK also plays an important role in HIV replication in T-cells[107].

Because REX1/ZPF42 acts to inhibit p38 MAPK activity, and because p38 MAPK activity has deleterious effects in the context of HIV, the inverse association with NHL seen for rs6815391

further suggests that this SNP may upregulate expression of REX1/ZPF42, thereby downregulating transcription of IL-10, working to inhibit HIV replication, and reducing risk of NHL.

2.10.2. AXIN2 Degrades β-Catenin and Downregulates Wnt Signaling

The SNP rs2240308 is a missense variant 500B downstream from AXIN2, which promotes the degradation of nuclear β-catenin and is a negative regulator of Wnt signaling. β-catenin activates Wnt signaling; Wnt activation leads to greater accumulation of β-catenin in the nucleus and activation of downstream pathways, in which c-Myc and cancer-related target genes figure prominently[108]. This would suggest that rs2240308 compromises the ability of AXIN2 to degrade β-catenin effectively, thus upregulating Wnt signaling, and that the positive association between rs2240308 and NHL risk may be due to activation of these canonical cancer pathways.

On the other hand, active β-catenin represses HIV-1 replication in astrocytes (glial cells of the brain and spinal cord; the central nervous system is a known reservoir for HIV)[109]. Given this, rs2240308 would seem to be protective against NHL risk, by reducing HIV replication and thus HIV viral load.

Similarly, the literature on rs2240308 is somewhat muddled. The majority of published research on this SNP investigates its association with lung and prostate cancer in Asian populations, with some authors finding a protective effect for the G/A and A/A genotypes relative to G/G (which was our reference genotype, based on dbSNP data), others finding the opposite, and inconsistent definition of reference genotype across articles, with some using A/A as the referent and others G/G[108]. (MAF does differ somewhat between Asian and white populations, but this does not explain the inconsistency: A remains the minor allele in both populations.) Insofar as there is a

consensus, this appears to be that rs2240308 is positively associated with risk of prostate and lung cancer in Asian populations, but not in Caucasian (i.e. Turkish or Polish) populations[110].

It is therefore clear from the literature that there is some association between rs2240308 and cancers other than lymphoma, but the direction of this association varies according to study design and study population. That we have observed an association, regardless of direction, between rs2240308 and NHL, is thus consistent with the literature. However, canonical cancer biology and the biology of HIV would seem to be at cross-purposes here. Since Wnt/β-catenin signaling is a complex process, it is possible that the qualitative impact of SNP rs2240308 depends on the presence or absence of other SNPs, and on other covariates not measured in this study.

2.10.3. WNT2 and WNT8A Code for Wnt-Family Ligands

As discussed above for AXIN2, Wnt/β-catenin signaling plays an important role in oncogenic processes. WNT2 and WNT8A code for Wnt-family ligands; suggestive associations for SNPs on these genes (rs4730775 on WNT2, and rs4835761 on WNT8A) were observed following semi-Bayes correction.

SNP rs4730775 is a noncoding variant located in the 3' UTR of WNT2. It has been implicated in Peyronie disease, characterized by abnormal formation of scar tissue in the genitalia[111], and a protective association was seen for Dupuytren's disease[112], a fibromatosis involving thickening and contraction of tissue in the hand. In Chinese populations, Wallar found a weakly suggestive inverse association between this SNP and esophageal cancer (dominant model OR=0.89, 95%CI 0.75-1.07)[113]; Liu found an inverse association with liver cancer (C/T vs. C/C OR=0.71; 95%CI 0.50-0.99)[114]. SNP rs4835761 is a noncoding variant 2KB upstream of WNT8A. A nominally

significant positive association (log-additive OR=1.17; 95%CI 1.01-1.35) between this SNP and risk of bladder cancer was observed in a US population[115].

Published work has not investigated these SNPs within the context of HIV-related hematological malignancies. Given this lack of published data on these SNPs and hematological malignancies and HIV, it is reasonable to conclude that, insofar as these associations are true-positives, they may operate via canonical mechanisms of WNT/β-catenin signaling described for AXIN2. However, the possibility also exists that these are spurious associations; again, they are only suggestive.

2.10.4. The Semi-Bayes-Corrected Clinical Model Is "Best"

Model choice—i.e. models of genetic inheritance, and also covariate models—made some difference in results, though suggestive results for associations reported above were generally robust across covariate modeling scenarios. A natural question is therefore which model is "best." From a technical standpoint, the answer is that the most precise unbiased model is best, i.e. the model that, absent confounding and bias, yields the narrowest confidence intervals. From an investigative standpoint, the answer may be that the best model includes enough covariates to give a sufficient sense for factors that affect AIDS-NHL risk besides genotype, but is not overly burdened by covariates that make little or no difference to our estimates.

By either standard, the semi-Bayes corrected model adjusting for clinical covariates is the best: it allowed for investigation of prior AIDS diagnosis, prior HAART, HCV status, HIV viral load, race and age, and also yielded the most precise estimates. The average ratio of upper to lower confidence limits (RCL)[116] for semi-Bayes corrected clinical models was 2.54, compared to 3.59 for the race-only model, 3.26 for the model adjusting for clinical covariates without semi-Bayes correction, 3.55 for the model adjusting for clinical covariates without semi-Bayes correction

using a white-only (n=622) subset of the full (n=700) sample, 5.31 for the complete-case analysis, and 3.89 for the analysis of data imputed for missing substance-use data.

In models adjusting for clinical covariates, semi-Bayes correction led to a 22% decrease in the width of 95% confidence intervals (measured using the ratio of upper to lower confidence limits in corrected and uncorrected models[116]) and a 2.5 % reduction in the magnitude of odds ratios. Since semi-Bayes correction aims explicitly to shrink confidence intervals and pull point estimates toward the null, this is in line with expectations. The scale of shrinkage was generally consistent with results in Peckham[98]. On the whole, semi-Bayes correction proved to be a useful tool, offering appreciable shrinkage of confidence intervals and increased precision at little cost in terms of the magnitude of point estimates.

Substance-use covariates did not have a uniform impact on point estimates. As one would expect, the complete-case analysis using substance-use covariates had much wider confidence intervals than any other model (RCL=5.31), but multiple imputation (RCL=3.89) narrowed these intervals by 27%. It is conceivable that with more finely-grained exposure data (i.e. cumulative exposure and frequency of consumption data), these factors could be investigated with greater precision. However, many NHL cases were diagnosed shortly after baseline, which would prohibit the calculation of meaningful cumulative exposure measures, even using imputation. Furthermore, zero cells (i.e. no participants in one stratum of another covariate, such as race, reporting use within a given stratum of the substance-use covariate, such as daily use) would rapidly become an issue for frequency data, precluding successful imputation.

2.11. Strengths and Limitations

2.11.1. Strengths

Strengths of this study include the use of imputation to fill in missing values for viral load at set point and other covariates, and the use of semi-Bayes adjustment to narrow confidence intervals and account for multiple comparisons in our data. The quality and quantity of covariate data in the MACS is another strength, though we were regrettably unable to examine gene-environment interactions owing to small sample size (n=700).

2.11.2. Limitations

This study suffers from two broad sets of limitations. One set is particular to this study; it includes the use of whole-genome amplified DNA from immortalized B-cells, potential shortcomings related to multiple imputation, and small sample size. The second set is common to all candidate-gene studies: specifically, we should not expect single-SNP associations to show great magnitude, and the targeted approach of candidate-gene studies can be a weakness when a suboptimal set of SNPs/genes is chosen for analysis.

2.11.2.1. Limitations Specific to this Study

2.11.2.1.1. Use of Whole-Genome-Amplified DNA from Immortalized B-Cells

This study also used whole-genome amplified DNA from B-cells immortalized with Epstein-Barr virus[117,118]. Whole-genome amplification can suffer amplification errors such as preferential amplification and allele dropout[119], which we guard against by ensuring that allele frequencies in our sample are comparable to those listed in dbSNP. Immortalization with EBV can lead to the accrual of oncogenic mutations in B-cells, though these may be limited to p53, BCL6 and beta-globin genes[120].

2.11.2.1.2. Potential for Substance-Use Data to Be Missing-Not-At-Random (MNAR)

Multiple imputation is predicated on one of two assumptions: that data are either missing at random (MAR), i.e. that missingness can be predicted by observed variables but not by unobserved values of the variable to be imputed, or missing completely at random (MCAR), i.e. that missingness cannot be predicted by either observed variables or unobserved values of the variable to be imputed. Here "random" refers specifically to the association between the missing status and the true unobserved value of a particular variable: the key component of both MAR and MCAR is that missingness is independent of the true unobserved value of the variable to be imputed; if this is not the case, then data are missing not at random (MNAR), and standard multiple imputation approaches are inappropriate.

For example, MNAR is a possibility when investigating the use of substances that carry a high degree of stigma, such as methamphetamine. This is a matter of self-report bias: because of shame or embarrassment, persons using methamphetamine may be less likely to answer questions assessing methamphetamine use than would non-users, meaning that missingness of the methamphetamine variable is correlated with the unobserved true value of the variable. In this case, data are therefore MNAR; necessary assumptions for standard multiple imputation are not met.

2.11.2.1.3. Small Sample Size

We had just 700 participants in our sample. SNPs were selected for fairly high MAF to guard against this, but our results would still have benefitted from a larger sample. This could also have allowed for the investigation of promising lower-frequency SNPs not considered here, and for analysis of potential gene-environment interactions including substance use.

2.11.2.2. Limitations Common to All Candidate-Gene Studies

Perhaps the major limitation of this study is that it was able to investigate only a small number of SNPs (n=30), and a smaller number of genes (n=24). Though candidate genes are selected using an informed approach that takes into account biological plausibility and prior evidence, the chance always exists that the subset chosen will not yield any associations, especially given limitations on resources available for participant recruitment and data collection. We address this shortcoming in the next chapter by using an agnostic genome-wide association study to examine ~5 million SNPs across the genome, rather than the 30 examined here, for associations with NHL.

Furthermore, as discussed in the Introduction, one limitation of candidate-gene studies is imposed by biology itself: because biological networks have evolved for redundancy, genes operate in concert rather than in isolation, and we should not expect the single-SNP associations investigated using candidate-gene approaches to be of great magnitude. Our pathway analysis of GWAS data, presented in Chapter 4, aims to overcome this limitation by combining SNPs into genes, then combining genes into pathways, and analyzing associations at these levels rather than at the level of single-SNP associations.

2.12. Conclusions and Further Directions

Using conditional logistic regression and semi-Bayes adjustment for multiple comparisons in a matched-case control study of 700 HIV-positive individuals in the Multicenter AIDS Cohort, we found a significant inverse association between risk of NHL and the REX1/ZPF42 3'UTR SNP rs6815391 under dominant (OR=0.68; 95%CI 0.47-0.99) and log-additive (OR=0.71, 95%CI 0.51-0.99) models of inheritance. In addition to a possible role in cancer stem-cell processes, REX1/ZPF42 inhibits the expression of the mitogen-activated protein kinase p38 MAPK. P38 MAPK plays an important role in HIV replication in T-cells[107] and upregulates extracellular HIV

Tat-induced transcription of IL-10, elevated serum levels of which have been observed in AIDS-NHL patients[80]. This suggests that C/T or T/T variant genotype for rs6815391 may be inversely associated with serum levels of IL-10, which may help explain the inverse association between rs6815391 and NHL risk observed in this study. A natural next step would therefore be assessing the correlation between serum levels of IL-10 and rs6815391 in HIV-positive MACS participants.

The next two chapters aim to overcome the two limitations common to candidate-gene studies discussed in Section 2.11.2.2: first, that associations are contingent on having selected a fruitful set of SNPs and genes to investigate; and second, the small magnitude of single-SNP associations with phenotypes of interest. In Chapter 3, we present the results of our GWAS, examining a fuller range of variation across the genome. In Chapter 4, we move beyond single-SNP associations to examine gene- and pathway-level results for SNPs assessed in the GWAS.

Table 2.1. NF-KB Signaling Pathway SNPs Assessed in Aim 1

Role of gene/function of gene product	HUGO gene	rsID	Genomic Context
Codes for inhibitor of NF-KB signaling	NFKBIA	rs1050851	Exon - synonymous
Codes for inhibitor of NF-KB signaling	NFKBIA	rs8904	UTR 3’
Codes for inhibitor of NF-KB signaling	NFKBIA	rs696	UTR 3’
Scaffold protein: assembles active kinase complex	IKBKAP	rs2230793	Exon - missense
Scaffold protein: assembles active kinase complex	IKBKAP	rs1538660	Exon - missense
Scaffold protein: assembles active kinase complex	IKBKAP	rs3204145	Exon - missense

Table 2.2. Stem Cell-Related SNPs Assessed in Aim 1

Role of gene/function of gene product	HUGO gene	rsID	Genomic Context
Transcription factor: marker of pluripotency; needed for reacquisition/maintenance of pluripotency.	REX1 (ZFP42)	rs6815391	UTR 3’

Transcription factor: part of key transcriptional regulatory network (with SOX2 & Nanog) in embryonic stem cells. Regulates/reprograms for pluripotency.	OCT4 (POU5F1)	rs13409	UTR 3’
Transcription factor: part of key transcriptional regulatory network (with SOX2 & Nanog) in embryonic stem cells. Regulates/reprograms for pluripotency.	OCT4 (POU5F1)	rs3130932	UTR 5’
Transcription factor: regulates stem cell proliferation; Hedgehog signaling effector	GLI1	rs2228224	Exon - missense
Stem cell maintenance/transcriptional repressor	CTBP2	rs3740535	downstream variant 500B, UTR 3’
“Stemness” marker; also active in epithelial-mesenchymal transition	EPCAM	rs1126497	Exon - missense

Table 2.3. Wnt/ β-Catenin Pathway SNPs Assessed in Aim 1

Role of gene/function of gene product	**HUGO gene**	**rsID**	**Genomic Context**
Codes for ligand	WNT2	rs3729629	Intronic
Codes for ligand	WNT2	rs4730775	nc transcript variant, UTR 3’
Codes for ligand: SNP within miRNA-449 binding site (miR-449 may be positive regulator of Wnt pathway)	WNT2B	rs2273368	UTR 3’
Codes for ligand	WNT8A	rs4835761	Upstream variant 2KB
Codes for receptor	FZD1	rs3750145	UTR 3’
Codes for receptor	FZD3	rs2241802	Exon - synonymous
Codes for β-catenin: SNP within miR-589 binding site (miR-589 may be positive regulator of Wnt pathway)	CTNNB1	rs2953	UTR 3’
β-catenin degradation: phosphorylates β-catenin and marks it for proteosomal degradation.	DVL2	rs222851	Intron variant, upstream variant 2KB
β-catenin degradation: scaffold protein for GSK3β, part of β-catenin destruction complex	AXIN1	rs1981492	Intron variant
β-catenin degradation: scaffold protein for GSK3β, part of β-catenin destruction complex	AXIN2	rs2240308	downstream variant 500B, missense
Transcription factor: Wnt pathway activator in presence of β-catenin; repressor in absence of β-catenin	TCF7L1	rs6754757	Intron variant
β-catenin transcriptional complex: interacts with CBP. SNP within miRNA-200a binding site (miR-200a may be positive regulator of Wnt pathway)	PPARGC 1A	rs3774923	UTR 3’

Table 2.4. Notch Signaling Pathway SNPs Assessed in Aim 1

Role of gene/function of gene product	**HUGO gene**	**rsID**	**Genomic Context**
Codes for receptor	NOTCH3	rs3815188	Exon - synonymous

Codes for receptor	NOTCH4	rs915894	Exon - missense
Codes for receptor	NOTCH4	rs520692	Exon - missense
Codes for ligand	DLL1	rs1421	UTR 3’
Codes for ligand	DLL1	rs1033583	UTR 3’
Codes for ligand	JAG2	rs9972231	Exon - missense
Transcription factor—Notch signaling target	HES2	rs11364	UTR 3’
Transcription factor—Notch signaling target	HES2	rs8708	UTR 3’
Transcription factor—Notch signaling target	HEY1	rs1046472	UTR 3’
Transcription factor—Notch signaling target	HEY2	rs3734637	UTR 3’

Table 2.5. Allele and Genotype Frequencies in Non-White Participants (n=78).

Non-White Participants (n=78)				CASES					CONTROLS				
SNP gene & rsID	A1	A2	dbSNP MAF	N	A1 A1	A1 A2	A2 A2	MAF	N	A1 A1	A1 A2	A2 A2	MAF
DLL1 rs1033583	A	C	0.25	28	18	9	1	0.2	50	29	16	5	0.26
HES2 rs11364	G	A	0.29	28	17	10	1	0.21	50	20	23	7	0.37
OCT4 (POU5F1) rs13409	C	T	0.43	28	6	18	4	0.46	50	15	24	11	0.46
AXIN1 rs1981492	G	A	0.42	28	9	18	1	0.36	50	23	19	8	0.35
IKBKAP rs2230793	A	C	0.3	28	15	10	3	0.29	50	29	16	5	0.26
NOTCH4 rs520692	A	G	0.27	28	17	10	1	0.21	49	35	11	3	0.17
HEY1 rs1046472	C	A	0.17	28	17	10	1	0.21	50	38	12	0	0.12
FZD3 rs2241802	G	A	0.46	28	6	13	9	0.55	50	12	29	9	0.47
TCF7L1 rs6754757	T	G	0.35	28	11	12	5	0.39	50	18	24	8	0.4
DLL1 rs1421	A	G	0.1	28	21	7	0	0.13	50	39	11	0	0.11
WNT2 rs3729629	G	C	0.43	28	7	15	6	0.48	50	9	28	13	0.54
NOTCH3 rs3815188	G	A	0.22	28	17	9	2	0.23	47	26	16	5	0.28
REX1 (ZFP42) rs6815391	C	T	0.39	28	17	10	1	0.21	50	25	21	4	0.29
NOTCH4 rs915894	A	C	0.4	28	15	12	1	0.25	50	19	23	8	0.39
NFKBIA rs1050851	C	T	0.1	28	17	9	2	0.23	50	42	8	0	0.08
IKBKAP rs1538660	C	T	0.25	28	21	4	3	0.18	50	29	12	2	0.23
HEY2 rs3734637	A	C	0.45	28	7	17	4	0.45	50	17	25	8	0.41
NFKBIA rs696	G	A	0.46	28	12	10	6	0.39	50	20	21	9	0.39
AXIN2 rs2240308	G	A	0.34	28	6	16	6	0.5	50	21	22	7	0.36
WNT2 rs4730775	C	T	0.32	27	14	11	2	0.28	50	19	26	5	0.36
CTBP2 rs3740535	G	A	0.43	28	6	15	7	0.52	50	11	25	14	0.53

GLI1 rs2228224	**G**	**A**	0.36	28	10	14	4	0.39	50	17	25	8	0.41
WNT8A rs4835761	**G**	**A**	0.47	28	9	15	4	0.41	50	24	19	7	0.33
CTNNB1 rs2953	**T**	**G**	0.37	28	9	16	3	0.39	50	20	21	9	0.39
DVL2 rs222851	**A**	**G**	0.49	28	7	18	3	0.43	50	17	24	9	0.42
HES2 rs8708	**A**	**G**	0.47	28	4	14	10	0.61	50	13	24	13	0.5
FZD1 rs3750145	**A**	**G**	0.14	28	20	8	0	0.14	49	40	9	0	0.09
OCT4 (POU5F1) rs3130932	**T**	**G**	0.32	28	12	15	1	0.3	50	28	19	3	0.25
EPCAM rs1126497	**C**	**T**	0.33	28	11	12	5	0.39	50	25	18	7	0.32
PPARGC1A rs3774923	**G**	**A**	0.1	28	24	4	0	0.07	50	44	6	0	0.06

Table 2.6. Allele and Genotype Frequencies in White Participants (n=622).

White Participants Only (n=622)				**CASES**					**CONTROLS**				
SNP gene & rsID	**A1**	**A2**	**dbSNP MAF**	**N**	**A1 A1**	**A1 A2**	**A2 A2**	**MAF**	**N**	**A1 A1**	**A1 A2**	**A2 A2**	**MAF**
DLL1 rs1033583	**A**	**C**	0.25	149	67	71	11	0.31	471	214	210	47	0.32
HES2 rs11364	**G**	**A**	0.29	150	112	34	4	0.14	471	349	115	7	0.14
OCT4 (POU5F1) rs13409	**C**	**T**	0.43	150	38	82	30	0.47	472	153	235	84	0.43
AXIN1 rs1981492	**G**	**A**	0.42	149	48	71	30	0.44	464	169	214	81	0.41
IKBKAP rs2230793	**A**	**C**	0.3	150	106	38	6	0.17	470	330	126	14	0.16
NOTCH4 rs520692	**A**	**G**	0.27	150	61	68	21	0.37	470	193	229	48	0.35
HEY1 rs1046472	**C**	**A**	0.17	150	80	58	12	0.27	471	272	169	30	0.24
FZD3 rs2241802	**G**	**A**	0.46	149	49	75	25	0.42	472	162	226	84	0.42
TCF7L1 rs6754757	**T**	**G**	0.35	150	59	65	26	0.39	472	151	238	83	0.43
DLL1 rs1421	**A**	**G**	0.1	150	119	28	3	0.11	471	351	110	10	0.14
WNT2 rs3729629	**G**	**C**	0.43	150	25	87	38	0.54	471	110	244	117	0.51
NOTCH3 rs3815188	**G**	**A**	0.22	144	105	34	5	0.15	462	332	112	18	0.16
REX1 (ZFP42) rs6815391	**C**	**T**	0.39	150	102	44	4	0.17	472	296	158	18	0.21
NOTCH4 rs915894	**A**	**C**	0.4	150	49	77	24	0.42	472	175	236	61	0.38
NFKBIA rs1050851	**C**	**T**	0.1	150	89	53	8	0.23	472	285	168	19	0.22
IKBKAP rs1538660	**C**	**T**	0.25	150	107	40	3	0.15	472	329	134	9	0.16
HEY2 rs3734637	**A**	**C**	0.45	149	49	79	21	0.41	471	168	240	63	0.39
NFKBIA rs696	**G**	**A**	0.46	148	54	70	24	0.4	470	202	200	68	0.36
AXIN2 rs2240308	**G**	**A**	0.34	150	28	82	40	0.54	471	125	221	125	0.5
WNT2 rs4730775	**C**	**T**	0.32	150	39	68	43	0.51	472	137	238	97	0.46
CTBP2 rs3740535	**G**	**A**	0.43	150	85	58	7	0.24	471	267	177	27	0.25
GLI1 rs2228224	**G**	**A**	0.36	150	19	80	51	0.61	472	70	206	196	0.63

WNT8A rs4835761	**G**	**A**	0.47	150	34	79	37	0.51	472	135	233	104	0.47
CTNNB1 rs2953	**T**	**G**	0.37	149	44	71	34	0.47	472	131	234	107	0.47
DVL2 rs222851	**A**	**G**	0.49	150	24	71	55	0.61	472	59	248	165	0.61
HES2 rs8708	**A**	**G**	0.47	150	24	68	58	0.61	472	92	226	154	0.57
FZD1 rs3750145	**A**	**G**	0.14	150	103	42	5	0.17	471	343	120	8	0.14
OCT4 (POU5F1) rs3130932	**T**	**G**	0.32	150	73	63	14	0.3	471	214	200	57	0.33
EPCAM rs1126497	**C**	**T**	0.33	150	31	72	47	0.55	472	96	219	157	0.56
PPARGC1A rs3774923	**G**	**A**	0.1	149	136	12	1	0.05	471	426	44	1	0.05

Table 2.7. Allele and Genotype Frequencies in All Participants (n=700).

All Participants (n=700)				**CASES**					**CONTROLS**				
SNP gene & rsID	**A1**	**A2**	**dbSNP MAF**	**N**	**A1 A1**	**A1 A2**	**A2 A2**	**MAF**	**N**	**A1 A1**	**A1 A2**	**A2 A2**	**MAF**
DLL1 rs1033583	**A**	**C**	**0.25**	177	85	80	12	0.29	521	243	226	52	0.32
HES2 rs11364	**G**	**A**	**0.29**	178	129	44	5	0.15	521	369	138	14	0.16
OCT4 (POU5F1) rs13409	**C**	**T**	**0.43**	178	44	100	34	0.47	522	168	259	95	0.43
AXIN1 rs1981492	**G**	**A**	**0.42**	177	57	89	31	0.43	514	192	233	89	0.4
IKBKAP rs2230793	**A**	**C**	**0.3**	178	121	48	9	0.19	520	359	142	19	0.17
NOTCH4 rs520692	**A**	**G**	**0.27**	178	78	78	22	0.34	519	228	240	51	0.33
HEY1 rs1046472	**C**	**A**	**0.17**	178	97	68	13	0.26	521	310	181	30	0.23
FZD3 rs2241802	**G**	**A**	**0.46**	177	55	88	34	0.44	522	174	255	93	0.42
TCF7L1 rs6754757	**T**	**G**	**0.35**	178	70	77	31	0.39	522	169	262	91	0.43
DLL1 rs1421	**A**	**G**	**0.1**	178	140	35	3	0.12	521	390	121	10	0.14
WNT2 rs3729629	**G**	**C**	**0.43**	178	32	102	44	0.53	521	119	272	130	0.51
NOTCH3 rs3815188	**G**	**A**	**0.22**	172	122	43	7	0.17	509	358	128	23	0.17
REX1 (ZFP42) rs6815391	**C**	**T**	**0.39**	178	119	54	5	0.18	522	321	179	22	0.21
NOTCH4 rs915894	**A**	**C**	**0.4**	178	64	89	25	0.39	522	194	259	69	0.38
NFKBIA rs1050851	**C**	**T**	**0.1**	178	106	62	10	0.23	522	327	176	19	0.21
IKBKAP rs1538660	**C**	**T**	**0.25**	178	128	44	6	0.16	522	358	146	11	0.17
HEY2 rs3734637	**A**	**C**	**0.45**	177	56	96	25	0.41	521	185	265	71	0.39
NFKBIA rs696	**G**	**A**	**0.46**	176	66	80	30	0.4	520	222	221	77	0.36
AXIN2 rs2240308	**G**	**A**	**0.34**	178	34	98	46	0.53	521	146	243	132	0.49
WNT2 rs4730775	**C**	**T**	**0.32**	177	53	79	45	0.48	522	156	264	102	0.45
CTBP2 rs3740535	**G**	**A**	**0.43**	178	91	73	14	0.28	521	278	202	41	0.27
GLI1 rs2228224	**G**	**A**	**0.36**	178	29	94	55	0.57	522	87	231	204	0.61
WNT8A rs4835761	**G**	**A**	**0.47**	178	43	94	41	0.49	522	159	252	111	0.45

CTNNB1 rs2953	**T**	**G**	**0.37**	177	53	87	37	0.45	522	151	255	116	0.47
DVL2 rs222851	**A**	**G**	**0.49**	178	31	89	58	0.58	522	76	272	174	0.59
HES2 rs8708	**A**	**G**	**0.47**	178	28	82	68	0.61	522	105	250	167	0.56
FZD1 rs3750145	**A**	**G**	**0.14**	178	123	50	5	0.17	520	383	129	8	0.14
OCT4 (POU5F1) rs3130932	**T**	**G**	**0.32**	178	85	78	15	0.3	521	242	219	60	0.33
EPCAM rs1126497	**C**	**T**	**0.33**	178	42	84	52	0.53	522	121	237	164	0.54
PPARGC1A rs3774923	**G**	**A**	**0.1**	177	160	16	1	0.05	521	470	50	1	0.05

Table 2.8 Demographic Characteristics of HIV-Positive Controls (n=522) and AIDS-NHL Cases (n=178)

	HIV+, NHL- Controls (n=522)	**AIDS-NHL Cases (n=178)**
Total N	522	178
Age at reference date, median (range, SD)	40.4 (24.1-70.3, SD=7.52)	41.8 (24.9-61.3, SD = 7.75)
Age at reference date, n (%)		
24-29 years	33 (6.3%)	13 (7.3%)
30-39 years	229 (43.9%)	64 (36.0%)
40-49 years	207 (39.7%)	69 (38.8%)
>50 years	53 (10.2%)	32 (18.0%)
Race/ethnicity, n (%)		
White, non-Hispanic	472 (90.4%)	150 (84.3%)
White, Hispanic	25 (4.8%)	18 (10.1%)
Black, non-Hispanic	21 (4.0%)	10 (5.6%)
None of the above	4 (0.8%)	0 (0.0%)
Cohort, n (%)		
1984-1985 recruitment	469 (89.8%)	159 (89.3%)
1984 baseline date	401 (76.8%)	142 (79.8%)
1985 baseline date	68 (13.0%)	16 (9.0%)
1987-1991 recruitment	53 (10.2%)	19 (10.7%)
1987 baseline date	36 (6.9%)	12 (6.7%)
1988 baseline date	8 (1.5%)	2 (1.1%)
1989 baseline date	4 (0.8%)	2 (1.1%)
1990 baseline date	3 (0.6%)	1 (0.6%)
1991 baseline date	2 (0.4%)	2 (1.1%)
Reference year, n (%)		
1984-1995	442 (84.7%)	152 (85.4%)
1996-2001	64 (12.3%)	21 (11.8%)
2002-2006	16 (3.1%)	5 (2.8%)
HIV status at baseline, n (%)		
Seroprevalent	471 (90.2%)	161 (90.4%)
Seroconverter	51 (9.8%)	17 (9.6%)

	HIV+, NHL- Controls (n=522)	AIDS-NHL Cases (n=178)
HAART prior to reference date, n (%)		
Missing	0 (0.0%)	0 (0.0%)
Yes	46 (8.8%)	10 (5.6%)
No	476 (91.2%)	168 (94.4%)
ART prior to reference date, n (%)		
Missing	35 (6.7%)	58 (32.6%)
Yes	369 (70.7%)	109 (61.2%)
No	118 (22.6%)	11 (6.2%)
CD4+ T-cell slope pre-HAART, median (range, SD)	-60.69	-69.69
	(-369.99: 812.17; 70.27)	(-73.45: -283.38; 72.73)
Missing	4 (0.8%)	23 (12.9%)
CD4+ T-cell count at date of matching, median (range, SD)	88.5	81 (45.5%)
	(3.0—1361.0; 225.42)	(2.0—923.0, 212.43)
CD4+ T-cell count at date of matching, n (%)		
0-199	361 (69.2%)	124 (69.7%)
200-399	76 (14.6%)	28 (15.7%)
>=400	85 (16.3%)	26 (14.6%)
RNA set point, median (range, SD)	20730	31090.5
	(300-672810, 83386.23)	(400-960960, 127833.15)
Missing	84 (16.1%)	44 (24.7%)
AIDS diagnosis prior to reference date, n (%)		
Yes	212 (40.6%)	93 (52.2%)
No	310 (59.4%)	85 (47.8%)
Smoking status (3-year lagged), n (%)		
Missing	96 (18.4%)	34 (19.1%)
Never Smoked	148 (28.4%)	47 (26.4%)
Former Smoker	146 (28.0%)	49 (27.5%)
Current Smoker	132 (25.3%)	48 (27.0%)
Drinking: frequency since last visit (3-year lagged), n (%)		
Missing	91 (17.4%)	32 (18.0%)
>= 1 per day	29 (5.6%)	14 (7.9%)
Nearly every day	43 (8.2%)	7 (3.9%)
3 or 4 per week	57 (10.9%)	16 (9.0%)
1 or 2 per week	124 (23.8%)	42 (23.6%)
2 or 3 per month	40 (7.7%)	25 (14.0%)
1 per month	33 (6.3%)	9 (5.1%)

	HIV+, NHL- Controls (n=522)	AIDS-NHL Cases (n=178)
6-11 per year	16 (3.1%)	3 (1.7%)
1-5 per year	18 (3.4%)	12 (6.7%)
0/Refused	71 (13.6%)	18 (10.1%)
Hash/marijuana: used since last visit (yes/no, post-baseline only; 3-year lagged)?		
Missing	91 (17.4%)	33 (18.5%)
No	198 (37.9%)	76 (42.7%)
Yes	233 (44.6%)	69 (38.8%)
Crack or other cocaine: used since last visit (3-year lagged)?		
Missing	91 (17.4%)	33 (18.5%)
No	334 (64.0%)	127 (71.4%)
Yes	97 (18.6%)	18 (10.1%)
Uppers (crystal, meth, speed, ice): used since last visit (yes/no; 3-yr lagged)?		
Missing/not specified in form	168 (32.2%)	59 (33.1%)
No	314 (60.2%)	104 (58.4%)
Yes	40 (7.7%)	15 (8.4%)
Heroin or other opiates: used since last visit? (3-yr lagged)		
Missing/not specified in form	271 (51.9%)	93 (52.2%)
No	249 (47.7%)	84 (47.2%)
Yes	2 (0.4%)	1 (0.6%)

Table 2.9. Odds Ratios and 95% CIs, Uncorrected and Semi-Bayes Logistic Regression Models Adjusting for Race, Age, Prior AIDS, Prior HAART, HCV, Viral Load (Median-Value Imputed), Tobacco, Alcohol, Cannabis, Cocaine, Uppers Consumption

SNP	Self-Reported Race Only (n=700)	Adjusted for Clinical Covariates, no SB Correction (n=700)	Adjusted for Clinical Covariates, no SB Correction: White Only (n=622)	Adjusted for Clinical Covariates, SB-Corrected (n=700)	Complete Case Analysis: Substance Use and Clinical Covariates (n=389)	Imputed Analysis: Substance Use and Clinical Covariates (n=700)
DLL1 rs1033583						
AA	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)
AC	1.05 (0.74-1.49)	1.01 (0.70-1.46)	0.98 (0.66-1.46)	1.03 (0.72-1.46)	1.42 (0.80-2.53)	1.02 (0.69-1.53)
CC	0.67 (0.34-1.32)	0.62 (0.30-1.28)	0.63 (0.29-1.37)	0.69 (0.37-1.29)	0.70 (0.21-2.27)	0.53 (0.24-1.15)
Log-additive	0.92 (0.70-1.19)	0.89 (0.67-1.17)	0.87 (0.64-1.19)	0.89 (0.68-1.17)	1.08 (0.69-1.67)	0.85 (0.63-1.15)
Dominant model	0.98 (0.70-1.37)	0.94 (0.66-1.34)	0.92 (0.62-1.35)	0.95 (0.67-1.33)	1.31 (0.75-2.30)	0.92 (0.63-1.35)
Recessive model	0.66 (0.34-1.27)	0.62 (0.31-1.25)	0.64 (0.30-1.35)	0.68 (0.37-1.26)	0.56 (0.18-1.75)	0.52 (0.25-1.11)
HES2 rs11364						
GG	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)
AG	0.89 (0.60-1.33)	0.81 (0.53-1.23)	0.83 (0.52-1.31)	0.83 (0.55-1.23)	0.75 (0.38-1.50)	0.77 (0.49-1.23)
AA	0.93 (0.32-2.73)	0.84 (0.28-2.55)	1.64 (0.46-5.81)	0.90 (0.38-2.13)	1.06 (0.15-7.37)	0.86 (0.25-2.92)
Log-additive	0.91 (0.65-1.28)	0.84 (0.59-1.20)	0.95 (0.64-1.40)	0.85 (0.60-1.20)	0.83 (0.45-1.50)	0.82 (0.56-1.21)
Dominant model	0.89 (0.61-1.32)	0.81 (0.54-1.22)	0.88 (0.57-1.37)	0.82 (0.56-1.22)	0.77 (0.40-1.51)	0.78 (0.50-1.22)
Recessive model	0.97 (0.33-2.82)	0.90 (0.30-2.69)	1.69 (0.48-5.98)	0.94 (0.40-2.21)	1.18 (0.17-7.98)	0.93 (0.28-3.14)
OCT4 (POU5F1) rs13409						
CC	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)
CT	1.46 (0.98-2.18)	1.43 (0.94-2.19)	1.41 (0.89-2.22)	1.37 (0.92-2.04)	1.66 (0.85-3.24)	1.46 (0.92-2.33)
TT	1.33 (0.80-2.21)	1.42 (0.83-2.44)	1.50 (0.84-2.69)	1.34 (0.82-2.20)	1.54 (0.67-3.54)	1.46 (0.81-2.63)
Log-additive	1.18 (0.92-1.51)	1.22 (0.94-1.58)	1.24 (0.94-1.65)	1.21 (0.94-1.56)	1.26 (0.84-1.88)	1.23 (0.92-1.64)
Dominant model	1.42 (0.97-2.08)	1.43 (0.96-2.14)	1.43 (0.93-2.21)	1.39 (0.95-2.05)	1.63 (0.85-3.11)	1.46 (0.94-2.28)
Recessive model	1.05 (0.68-1.63)	1.14 (0.72-1.81)	1.21 (0.74-1.99)	1.12 (0.72-1.74)	1.09 (0.55-2.14)	1.14 (0.69-1.88)
AXIN1 rs1981492						
GG	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)
AG	1.32 (0.89-1.94)	1.18 (0.79-1.76)	1.05 (0.67-1.63)	1.16 (0.79-1.70)	1.42 (0.74-2.73)	1.07 (0.69-1.67)

SNP	Self-Reported Race Only (n=700)	Adjusted for Clinical Covariates, no SB Correction (n=700)	Adjusted for Clinical Covariates, no SB Correction: White Only (n=622)	Adjusted for Clinical Covariates, SB-Corrected (n=700)	Complete Case Analysis: Substance Use and Clinical Covariates (n=389)	Imputed Analysis: Substance Use and Clinical Covariates (n=700)
AA	1.27 (0.75-2.14)	1.12 (0.65-1.93)	1.22 (0.68-2.17)	1.09 (0.66-1.81)	1.69 (0.69-4.17)	1.08 (0.59-1.95)
Log-additive	1.15 (0.90-1.48)	1.08 (0.83-1.40)	1.09 (0.82-1.45)	1.08 (0.83-1.39)	1.32 (0.85-2.05)	1.04 (0.78-1.39)
Dominant model	1.30 (0.90-1.89)	1.17 (0.80-1.71)	1.09 (0.72-1.65)	1.15 (0.80-1.67)	1.47 (0.78-2.77)	1.07 (0.71-1.64)
Recessive model	1.07 (0.68-1.70)	1.01 (0.62-1.64)	1.18 (0.70-1.99)	1.01 (0.64-1.60)	1.33 (0.62-2.86)	1.03 (0.61-1.74)
IKBKAP rs2230793						
AA	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)
AC	0.99 (0.67-1.46)	1.01 (0.67-1.52)	0.95 (0.60-1.50)	1.01 (0.68-1.49)	0.45 (0.22-0.91)	0.92 (0.59-1.44)
CC	1.26 (0.55-2.90)	1.10 (0.46-2.61)	1.08 (0.39-2.99)	1.07 (0.51-2.23)	0.59 (0.12-2.82)	1.04 (0.41-2.64)
Log-additive	1.05 (0.77-1.43)	1.03 (0.74-1.42)	0.99 (0.69-1.42)	1.02 (0.75-1.40)	0.56 (0.32-0.99)	0.97 (0.68-1.37)
Dominant model	1.02 (0.70-1.48)	1.02 (0.69-1.51)	0.97 (0.63-1.49)	1.02 (0.70-1.49)	0.46 (0.24-0.92)	0.94 (0.61-1.43)
Recessive model	1.27 (0.56-2.87)	1.09 (0.46-2.57)	1.09 (0.40-3.01)	1.07 (0.51-2.21)	0.80 (0.18-3.58)	1.06 (0.42-2.68)
NOTCH4 rs520692						
AA	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)
AG	0.98 (0.68-1.43)	0.97 (0.66-1.43)	0.90 (0.58-1.37)	0.97 (0.67-1.40)	1.64 (0.89-3.03)	1.04 (0.68-1.59)
GG	1.35 (0.75-2.43)	1.39 (0.76-2.55)	1.38 (0.74-2.60)	1.32 (0.76-2.30)	2.13 (0.80-5.70)	1.56 (0.80-3.04)
Log-additive	1.10 (0.84-1.44)	1.11 (0.84-1.46)	1.09 (0.81-1.46)	1.10 (0.84-1.45)	1.52 (0.98-2.36)	1.17 (0.87-1.59)
Dominant model	1.04 (0.73-1.48)	1.04 (0.72-1.51)	0.98 (0.66-1.47)	1.04 (0.73-1.48)	1.72 (0.95-3.09)	1.12 (0.75-1.68)
Recessive model	1.36 (0.79-2.36)	1.41 (0.80-2.50)	1.47 (0.81-2.65)	1.34 (0.79-2.28)	1.62 (0.66-4.00)	1.53 (0.81-2.88)
HEY1 rs1046472						
CC	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)
AC	1.24 (0.87-1.78)	1.20 (0.82-1.75)	1.16 (0.77-1.76)	1.17 (0.82-1.69)	1.16 (0.65-2.08)	1.14 (0.75-1.73)
AA	1.53 (0.76-3.09)	1.43 (0.70-2.94)	1.32 (0.62-2.81)	1.32 (0.70-2.51)	0.90 (0.26-3.19)	1.38 (0.63-2.99)
Log-additive	1.24 (0.94-1.64)	1.20 (0.90-1.60)	1.15 (0.85-1.57)	1.19 (0.90-1.58)	1.06 (0.67-1.68)	1.16 (0.84-1.59)
Dominant model	1.28 (0.91-1.80)	1.23 (0.86-1.77)	1.19 (0.80-1.76)	1.21 (0.86-1.72)	1.13 (0.64-1.98)	1.17 (0.78-1.74)
Recessive model	1.39 (0.71-2.75)	1.32 (0.66-2.66)	1.24 (0.60-2.58)	1.25 (0.67-2.34)	0.85 (0.25-2.90)	1.30 (0.61-2.76)
FZD3 rs2241802						
GG	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)

SNP	Self-Reported Race Only (n=700)	Adjusted for Clinical Covariates, no SB Correction (n=700)	Adjusted for Clinical Covariates, no SB Correction: White Only (n=622)	Adjusted for Clinical Covariates, SB-Corrected (n=700)	Complete Case Analysis: Substance Use and Clinical Covariates (n=389)	Imputed Analysis: Substance Use and Clinical Covariates (n=700)
AG	1.07 (0.72-1.59)	1.18 (0.78-1.77)	1.27 (0.81-1.98)	1.15 (0.79-1.70)	1.03 (0.55-1.96)	1.20 (0.77-1.87)
AA	1.10 (0.67-1.81)	1.15 (0.69-1.92)	1.07 (0.61-1.89)	1.13 (0.70-1.81)	1.22 (0.56-2.66)	1.18 (0.67-2.08)
Log-additive	1.05 (0.83-1.34)	1.08 (0.85-1.39)	1.06 (0.81-1.39)	1.08 (0.85-1.38)	1.10 (0.75-1.61)	1.10 (0.84-1.45)
Dominant model	1.08 (0.75-1.57)	1.17 (0.80-1.72)	1.21 (0.79-1.83)	1.16 (0.80-1.67)	1.09 (0.61-1.97)	1.20 (0.79-1.82)
Recessive model	1.06 (0.69-1.62)	1.05 (0.67-1.63)	0.94 (0.57-1.55)	1.04 (0.68-1.59)	1.20 (0.60-2.41)	1.06 (0.65-1.73)
TCF7L1 rs6754757						
TT	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)
GT	0.71 (0.48-1.04)	0.73 (0.49-1.09)	0.67 (0.43-1.04)	0.76 (0.52-1.10)	0.66 (0.35-1.25)	0.71 (0.46-1.10)
GG	0.81 (0.49-1.33)	0.82 (0.49-1.39)	0.71 (0.40-1.26)	0.85 (0.52-1.39)	0.84 (0.36-1.93)	0.75 (0.42-1.32)
Log-additive	0.86 (0.67-1.11)	0.87 (0.67-1.13)	0.81 (0.61-1.09)	0.88 (0.68-1.13)	0.87 (0.58-1.33)	0.83 (0.63-1.11)
Dominant model	0.73 (0.51-1.05)	0.75 (0.52-1.10)	0.68 (0.44-1.04)	0.77 (0.54-1.10)	0.70 (0.38-1.29)	0.72 (0.48-1.09)
Recessive model	1.00 (0.64-1.56)	0.99 (0.62-1.59)	0.91 (0.55-1.51)	1.00 (0.64-1.55)	1.10 (0.54-2.27)	0.91 (0.55-1.53)
DLL1 rs1421						
AA	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)
AG	0.78 (0.51-1.20)	0.78 (0.49-1.22)	0.75 (0.46-1.22)	0.80 (0.52-1.22)	0.54 (0.25-1.16)	0.72 (0.44-1.19)
GG	0.88 (0.23-3.33)	0.99 (0.24-4.03)	0.84 (0.21-3.41)	1.00 (0.37-2.67)	0.79 (0.11-5.90)	0.89 (0.19-4.27)
Log-additive	0.82 (0.57-1.19)	0.83 (0.56-1.23)	0.80 (0.52-1.21)	0.84 (0.58-1.23)	0.64 (0.33-1.22)	0.78 (0.50-1.20)
Dominant model	0.79 (0.52-1.19)	0.79 (0.51-1.23)	0.76 (0.47-1.21)	0.81 (0.53-1.23)	0.56 (0.27-1.17)	0.73 (0.45-1.19)
Recessive model	0.90 (0.24-3.40)	1.03 (0.26-4.15)	0.88 (0.22-3.54)	1.01 (0.38-2.72)	0.94 (0.13-6.73)	0.94 (0.20-4.43)
WNT2 rs3729629						
CC	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)
CG	1.11 (0.73-1.67)	1.06 (0.69-1.63)	1.05 (0.65-1.71)	1.08 (0.72-1.61)	0.96 (0.49-1.89)	0.98 (0.61-1.58)
GG	0.78 (0.46-1.32)	0.69 (0.40-1.20)	0.65 (0.35-1.21)	0.72 (0.44-1.20)	0.58 (0.24-1.39)	0.61 (0.33-1.12)
Log-additive	0.90 (0.70-1.16)	0.85 (0.65-1.11)	0.83 (0.62-1.11)	0.85 (0.66-1.11)	0.79 (0.52-1.20)	0.79 (0.59-1.07)
Dominant model	1.00 (0.68-1.49)	0.94 (0.62-1.42)	0.92 (0.58-1.46)	0.95 (0.64-1.41)	0.83 (0.44-1.57)	0.86 (0.55-1.36)
Recessive model	0.73 (0.47-1.14)	0.66 (0.41-1.06)	0.63 (0.38-1.06)	0.69 (0.44-1.07)	0.60 (0.28-1.29)	0.61 (0.37-1.03)
NOTCH3 rs3815188						

SNP	Self-Reported Race Only (n=700)	Adjusted for Clinical Covariates, no SB Correction (n=700)	Adjusted for Clinical Covariates, no SB Correction: White Only (n=622)	Adjusted for Clinical Covariates, SB-Corrected (n=700)	Complete Case Analysis: Substance Use and Clinical Covariates (n=389)	Imputed Analysis: Substance Use and Clinical Covariates (n=700)
GG	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)
AG	0.96 (0.64-1.44)	0.83 (0.54-1.27)	0.80 (0.49-1.29)	0.85 (0.56-1.27)	0.61 (0.30-1.25)	0.87 (0.55-1.38)
AA	0.78 (0.31-2.01)	0.69 (0.26-1.88)	0.81 (0.26-2.55)	0.79 (0.36-1.74)	1.00 (0.23-4.39)	0.64 (0.22-1.87)
Log-additive	0.93 (0.68-1.28)	0.83 (0.59-1.17)	0.84 (0.57-1.23)	0.84 (0.60-1.17)	0.76 (0.43-1.34)	0.84 (0.58-1.20)
Dominant model	0.93 (0.64-1.37)	0.81 (0.54-1.21)	0.80 (0.50-1.26)	0.82 (0.56-1.21)	0.65 (0.33-1.28)	0.83 (0.54-1.28)
Recessive model	0.79 (0.31-2.02)	0.72 (0.27-1.95)	0.85 (0.27-2.68)	0.81 (0.36-1.78)	1.18 (0.28-4.97)	0.66 (0.22-1.92)
REX1 (ZFP42) rs6815391						
CC	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)
CT	0.77 (0.53-1.11)	0.68 (0.46-1.00)	0.66 (0.43-1.01)	0.71 (0.49-1.03)	0.60 (0.30-1.18)	0.72 (0.46-1.10)
TT	0.53 (0.19-1.46)	0.52 (0.18-1.57)	0.67 (0.20-2.29)	0.67 (0.29-1.55)	0.15 (0.02-1.39)	0.53 (0.17-1.70)
Log-additive	0.75 (0.55-1.04)	0.69 (0.49-0.98)	0.70 (0.48-1.02)	**0.71 (0.51-0.99)**	0.52 (0.29-0.93)	0.72 (0.50-1.05)
Dominant model	0.74 (0.51-1.07)	0.66 (0.45-0.98)	0.66 (0.43-1.00)	**0.68 (0.47-0.99)**	0.52 (0.27-1.00)	0.70 (0.46-1.07)
Recessive model	0.59 (0.21-1.61)	0.63 (0.22-1.85)	0.80 (0.24-2.67)	0.75 (0.33-1.72)	0.17 (0.02-1.56)	0.62 (0.20-1.94)
NOTCH4 rs915894						
AA	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)
AC	1.06 (0.73-1.54)	1.07 (0.72-1.58)	1.13 (0.73-1.75)	1.06 (0.73-1.54)	1.16 (0.61-2.22)	1.07 (0.70-1.65)
CC	1.05 (0.60-1.84)	1.08 (0.60-1.94)	1.44 (0.77-2.70)	1.07 (0.62-1.83)	0.70 (0.29-1.69)	1.00 (0.53-1.90)
Log-additive	1.04 (0.80-1.34)	1.05 (0.80-1.37)	1.18 (0.87-1.60)	1.04 (0.80-1.36)	0.89 (0.59-1.35)	1.02 (0.76-1.37)
Dominant model	1.06 (0.74-1.51)	1.07 (0.73-1.55)	1.18 (0.78-1.80)	1.06 (0.74-1.53)	1.01 (0.55-1.84)	1.06 (0.70-1.60)
Recessive model	1.02 (0.61-1.71)	1.04 (0.60-1.80)	1.34 (0.76-2.37)	1.04 (0.62-1.72)	0.64 (0.28-1.47)	0.96 (0.53-1.74)
NFKBIA rs1050851						
CC	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)
CT	1.09 (0.76-1.57)	1.14 (0.78-1.66)	1.05 (0.70-1.57)	1.11 (0.77-1.61)	0.86 (0.47-1.56)	1.03 (0.68-1.56)
TT	1.60 (0.70-3.64)	1.56 (0.66-3.70)	1.24 (0.49-3.14)	1.37 (0.66-2.87)	1.20 (0.36-4.02)	1.99 (0.74-5.36)
Log-additive	1.16 (0.86-1.57)	1.18 (0.87-1.62)	1.08 (0.77-1.50)	1.17 (0.86-1.59)	0.97 (0.60-1.55)	1.16 (0.83-1.64)
Dominant model	1.14 (0.80-1.62)	1.17 (0.81-1.70)	1.07 (0.72-1.58)	1.16 (0.81-1.66)	0.90 (0.50-1.60)	1.10 (0.74-1.64)
Recessive model	1.54 (0.68-3.45)	1.48 (0.64-3.45)	1.22 (0.49-3.04)	1.33 (0.64-2.75)	1.28 (0.39-4.18)	1.97 (0.74-5.25)

SNP	Self-Reported Race Only (n=700)	Adjusted for Clinical Covariates, no SB Correction (n=700)	Adjusted for Clinical Covariates, no SB Correction: White Only (n=622)	Adjusted for Clinical Covariates, SB-Corrected (n=700)	Complete Case Analysis: Substance Use and Clinical Covariates (n=389)	Imputed Analysis: Substance Use and Clinical Covariates (n=700)
IKBKAP rs1538660						
CC	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)
CT	0.81 (0.55-1.21)	0.80 (0.53-1.19)	0.84 (0.54-1.29)	0.80 (0.55-1.18)	1.08 (0.57-2.04)	0.85 (0.55-1.31)
TT	1.46 (0.52-4.12)	1.71 (0.56-5.23)	1.76 (0.45-6.93)	1.38 (0.57-3.35)	0.43 (0.03-5.59)	1.50 (0.45-5.01)
Log-additive	0.92 (0.66-1.29)	0.92 (0.65-1.31)	0.93 (0.63-1.38)	0.93 (0.66-1.30)	0.97 (0.55-1.69)	0.95 (0.65-1.38)
Dominant model	0.86 (0.59-1.25)	0.85 (0.57-1.25)	0.87 (0.57-1.33)	0.86 (0.59-1.25)	1.02 (0.55-1.89)	0.89 (0.58-1.36)
Recessive model	1.56 (0.56-4.37)	1.82 (0.60-5.57)	1.85 (0.47-7.28)	1.43 (0.59-3.48)	0.43 (0.03-5.51)	1.56 (0.47-5.22)
HEY2 rs3734637						
AA	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)
AC	1.22 (0.83-1.79)	1.20 (0.80-1.80)	1.16 (0.74-1.80)	1.18 (0.80-1.73)	1.18 (0.63-2.21)	1.19 (0.76-1.85)
CC	1.17 (0.68-2.00)	1.04 (0.59-1.84)	0.96 (0.51-1.80)	1.03 (0.61-1.74)	0.95 (0.41-2.21)	1.01 (0.54-1.87)
Log-additive	1.11 (0.86-1.43)	1.05 (0.81-1.38)	1.01 (0.75-1.36)	1.05 (0.81-1.37)	1.02 (0.69-1.51)	1.04 (0.77-1.39)
Dominant model	1.21 (0.84-1.74)	1.16 (0.79-1.71)	1.11 (0.73-1.70)	1.15 (0.79-1.67)	1.11 (0.62-1.97)	1.15 (0.75-1.75)
Recessive model	1.05 (0.64-1.70)	0.93 (0.56-1.56)	0.87 (0.50-1.53)	0.94 (0.58-1.52)	0.88 (0.40-1.94)	0.90 (0.52-1.58)
NFKBIA rs696						
GG	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)
AG	1.26 (0.87-1.82)	1.26 (0.86-1.86)	1.29 (0.85-1.98)	1.23 (0.85-1.78)	1.27 (0.67-2.40)	1.21 (0.79-1.85)
AA	1.27 (0.76-2.12)	1.32 (0.77-2.26)	1.32 (0.73-2.40)	1.26 (0.77-2.08)	1.72 (0.73-4.06)	1.32 (0.73-2.36)
Log-additive	1.15 (0.91-1.47)	1.17 (0.91-1.51)	1.18 (0.90-1.56)	1.17 (0.91-1.50)	1.30 (0.87-1.95)	1.16 (0.88-1.53)
Dominant model	1.26 (0.89-1.78)	1.28 (0.89-1.84)	1.30 (0.87-1.94)	1.26 (0.88-1.79)	1.38 (0.77-2.48)	1.24 (0.83-1.84)
Recessive model	1.13 (0.70-1.82)	1.17 (0.71-1.92)	1.16 (0.67-2.01)	1.15 (0.72-1.84)	1.54 (0.69-3.45)	1.19 (0.70-2.05)
AXIN2 rs2240308						
GG	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)
AG	1.69 (1.08-2.64)	1.55 (0.97-2.47)	1.44 (0.85-2.43)	1.44 (0.94-2.22)	0.90 (0.44-1.83)	1.45 (0.88-2.39)
AA	1.51 (0.91-2.51)	1.50 (0.88-2.55)	1.41 (0.80-2.51)	1.39 (0.86-2.25)	0.81 (0.36-1.82)	1.24 (0.70-2.21)
Log-additive	1.21 (0.95-1.53)	1.20 (0.93-1.56)	1.17 (0.89-1.55)	1.20 (0.93-1.54)	0.90 (0.60-1.35)	1.10 (0.83-1.45)
Dominant model	**1.63 (1.07-2.49)**	1.53 (0.98-2.40)	1.43 (0.87-2.34)	1.47 (0.96-2.25)	0.87 (0.44-1.71)	1.38 (0.85-2.22)

SNP	Self-Reported Race Only (n=700)	Adjusted for Clinical Covariates, no SB Correction (n=700)	Adjusted for Clinical Covariates, no SB Correction: White Only (n=622)	Adjusted for Clinical Covariates, SB-Corrected (n=700)	Complete Case Analysis: Substance Use and Clinical Covariates (n=389)	Imputed Analysis: Substance Use and Clinical Covariates (n=700)
Recessive model	1.05 (0.71-1.55)	1.10 (0.73-1.64)	1.10 (0.71-1.70)	1.09 (0.74-1.60)	0.87 (0.47-1.64)	0.96 (0.61-1.49)
WNT2 rs4730775						
CC	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)
CT	0.91 (0.60-1.38)	0.95 (0.62-1.45)	0.97 (0.60-1.55)	0.94 (0.63-1.40)	1.96 (0.92-4.20)	0.99 (0.63-1.57)
TT	1.39 (0.86-2.24)	1.48 (0.89-2.46)	1.70 (0.98-2.95)	1.42 (0.88-2.27)	2.42 (1.05-5.59)	1.60 (0.91-2.80)
Log-additive	1.16 (0.91-1.49)	1.20 (0.93-1.55)	1.30 (0.98-1.72)	1.19 (0.93-1.54)	1.53 (1.02-2.31)	1.24 (0.94-1.65)
Dominant model	1.04 (0.71-1.53)	1.09 (0.73-1.62)	1.16 (0.74-1.81)	1.08 (0.74-1.58)	2.12 (1.04-4.34)	1.14 (0.74-1.76)
Recessive model	1.47 (0.98-2.21)	1.53 (0.99-2.35)	1.74 (1.10-2.75)	1.47 (0.97-2.22)	1.56 (0.82-2.95)	1.60 (0.99-2.59)
CTBP2 rs3740535						
GG	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)
AG	1.03 (0.71-1.51)	0.96 (0.64-1.44)	0.92 (0.60-1.42)	0.97 (0.66-1.42)	0.58 (0.30-1.12)	0.88 (0.57-1.36)
AA	0.84 (0.42-1.68)	0.99 (0.49-2.00)	0.79 (0.32-1.98)	0.99 (0.53-1.86)	0.57 (0.16-1.96)	0.81 (0.37-1.77)
Log-additive	0.97 (0.73-1.28)	0.98 (0.73-1.32)	0.91 (0.65-1.27)	0.98 (0.73-1.31)	0.66 (0.40-1.11)	0.89 (0.65-1.24)
Dominant model	1.00 (0.70-1.44)	0.97 (0.66-1.42)	0.90 (0.60-1.36)	0.97 (0.67-1.40)	0.58 (0.30-1.09)	0.87 (0.58-1.32)
Recessive model	0.82 (0.42-1.62)	1.00 (0.51-2.00)	0.81 (0.33-2.02)	1.00 (0.54-1.86)	0.75 (0.23-2.44)	0.86 (0.40-1.82)
GLI1 rs2228224						
AA	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)
AG	1.42 (0.97-2.09)	1.34 (0.89-2.01)	1.39 (0.90-2.13)	1.30 (0.89-1.91)	1.41 (0.70-2.84)	1.16 (0.75-1.80)
GG	1.10 (0.65-1.87)	1.14 (0.65-1.99)	1.06 (0.57-1.99)	1.11 (0.66-1.85)	1.02 (0.41-2.57)	1.06 (0.58-1.94)
Log-additive	1.11 (0.86-1.42)	1.11 (0.86-1.45)	1.11 (0.84-1.47)	1.11 (0.86-1.44)	1.07 (0.69-1.65)	1.05 (0.79-1.40)
Dominant model	1.34 (0.92-1.94)	1.29 (0.88-1.89)	1.31 (0.87-1.96)	1.27 (0.87-1.83)	1.30 (0.67-2.50)	1.14 (0.75-1.73)
Recessive model	0.89 (0.55-1.42)	0.96 (0.58-1.58)	0.89 (0.50-1.58)	0.96 (0.60-1.54)	0.83 (0.37-1.88)	0.97 (0.56-1.67)
WNT8A rs4835761						
GG	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)
AG	1.45 (0.95-2.21)	1.49 (0.97-2.32)	1.34 (0.82-2.18)	1.42 (0.94-2.13)	1.94 (0.94-4.03)	1.45 (0.90-2.35)
AA	1.42 (0.85-2.36)	1.44 (0.85-2.45)	1.52 (0.86-2.70)	1.35 (0.83-2.20)	2.16 (0.90-5.20)	1.29 (0.72-2.30)
Log-additive	1.20 (0.94-1.54)	1.21 (0.94-1.57)	1.24 (0.93-1.64)	1.20 (0.93-1.55)	1.46 (0.95-2.24)	1.15 (0.87-1.53)

SNP	Self-Reported Race Only (n=700)	Adjusted for Clinical Covariates, no SB Correction (n=700)	Adjusted for Clinical Covariates, no SB Correction: White Only (n=622)	Adjusted for Clinical Covariates, SB-Corrected (n=700)	Complete Case Analysis: Substance Use and Clinical Covariates (n=389)	Imputed Analysis: Substance Use and Clinical Covariates (n=700)
Dominant model	1.44 (0.97-2.15)	1.48 (0.98-2.24)	1.39 (0.88-2.20)	1.43 (0.96-2.13)	2.00 (0.99-4.04)	1.40 (0.89-2.21)
Recessive model	1.11 (0.73-1.69)	1.11 (0.72-1.72)	1.26 (0.78-2.03)	1.10 (0.72-1.67)	1.33 (0.68-2.61)	1.01 (0.62-1.64)
CTNNB1 rs2953						
TT	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)
GT	1.00 (0.67-1.49)	0.99 (0.65-1.51)	0.89 (0.56-1.42)	0.99 (0.67-1.49)	0.57 (0.28-1.16)	0.95 (0.59-1.52)
GG	0.93 (0.58-1.51)	0.91 (0.55-1.51)	0.83 (0.49-1.41)	0.92 (0.58-1.47)	0.88 (0.40-1.94)	1.00 (0.58-1.71)
Log-additive	0.97 (0.76-1.23)	0.96 (0.75-1.23)	0.91 (0.70-1.19)	0.96 (0.75-1.22)	0.92 (0.61-1.39)	1.00 (0.76-1.31)
Dominant model	0.98 (0.67-1.43)	0.96 (0.65-1.43)	0.87 (0.57-1.33)	0.97 (0.66-1.42)	0.67 (0.35-1.27)	0.96 (0.62-1.49)
Recessive model	0.93 (0.62-1.40)	0.92 (0.60-1.40)	0.89 (0.57-1.40)	0.92 (0.62-1.39)	1.24 (0.64-2.42)	1.03 (0.65-1.63)
DVL2 rs222851						
GG	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)
AG	0.96 (0.65-1.41)	0.88 (0.59-1.31)	0.83 (0.54-1.26)	0.89 (0.61-1.29)	0.97 (0.53-1.77)	0.89 (0.58-1.36)
AA	1.13 (0.68-1.89)	1.16 (0.68-2.00)	1.41 (0.78-2.56)	1.14 (0.69-1.89)	0.81 (0.34-1.93)	1.28 (0.72-2.30)
Log-additive	1.05 (0.81-1.35)	1.04 (0.79-1.36)	1.09 (0.81-1.45)	1.04 (0.80-1.35)	0.92 (0.61-1.38)	1.08 (0.81-1.44)
Dominant model	1.00 (0.69-1.44)	0.94 (0.64-1.37)	0.93 (0.62-1.38)	0.94 (0.65-1.36)	0.93 (0.53-1.65)	0.97 (0.64-1.45)
Recessive model	1.16 (0.74-1.83)	1.26 (0.77-2.04)	1.57 (0.91-2.72)	1.22 (0.77-1.94)	0.83 (0.38-1.82)	1.38 (0.82-2.33)
HES2 rs8708						
GG	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)
AG	0.82 (0.56-1.20)	0.84 (0.56-1.26)	0.79 (0.51-1.22)	0.87 (0.59-1.27)	1.02 (0.54-1.93)	0.89 (0.57-1.39)
AA	0.68 (0.41-1.13)	0.68 (0.40-1.15)	0.72 (0.40-1.29)	0.72 (0.44-1.17)	0.83 (0.37-1.89)	0.73 (0.41-1.29)
Log-additive	0.83 (0.65-1.05)	0.83 (0.64-1.07)	0.84 (0.63-1.11)	0.83 (0.65-1.07)	0.93 (0.62-1.38)	0.86 (0.65-1.13)
Dominant model	0.78 (0.54-1.12)	0.79 (0.54-1.16)	0.77 (0.51-1.15)	0.81 (0.56-1.16)	0.96 (0.53-1.75)	0.84 (0.55-1.27)
Recessive model	0.76 (0.48-1.21)	0.74 (0.46-1.21)	0.82 (0.48-1.40)	0.77 (0.49-1.21)	0.82 (0.40-1.71)	0.78 (0.47-1.31)
FZD1 rs3750145						
AA	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)
AG	1.27 (0.86-1.87)	1.08 (0.71-1.64)	0.94 (0.59-1.48)	1.06 (0.72-1.58)	1.44 (0.73-2.84)	1.04 (0.66-1.66)

SNP	Self-Reported Race Only (n=700)	Adjusted for Clinical Covariates, no SB Correction (n=700)	Adjusted for Clinical Covariates, no SB Correction: White Only (n=622)	Adjusted for Clinical Covariates, SB-Corrected (n=700)	Complete Case Analysis: Substance Use and Clinical Covariates (n=389)	Imputed Analysis: Substance Use and Clinical Covariates (n=700)
GG	2.53 (0.74-8.63)	2.88 (0.76-10.91)	2.01 (0.49-8.20)	1.72 (0.64-4.59)	6.82 (1.41-32.96)	4.25 (1.02-17.76)
Log-additive	1.35 (0.96-1.89)	1.21 (0.84-1.75)	1.05 (0.71-1.56)	1.20 (0.84-1.70)	1.83 (1.05-3.20)	1.24 (0.83-1.86)
Dominant model	1.33 (0.91-1.94)	1.15 (0.77-1.73)	0.99 (0.64-1.54)	1.14 (0.78-1.68)	1.71 (0.89-3.27)	1.14 (0.73-1.78)
Recessive model	2.39 (0.70-8.13)	2.84 (0.75-10.72)	2.03 (0.50-8.27)	1.71 (0.64-4.55)	6.25 (1.29-30.26)	4.21 (1.01-17.55)
OCT4 (POU5F1) rs3130932						
TT	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)
GT	1.00 (0.69-1.45)	0.96 (0.65-1.41)	0.84 (0.54-1.29)	0.97 (0.67-1.40)	0.98 (0.52-1.85)	0.97 (0.63-1.49)
GG	0.73 (0.39-1.36)	0.78 (0.41-1.51)	0.83 (0.42-1.65)	0.82 (0.46-1.48)	0.63 (0.23-1.73)	0.87 (0.42-1.77)
Log-additive	0.91 (0.70-1.18)	0.91 (0.69-1.21)	0.88 (0.65-1.20)	0.92 (0.70-1.20)	0.85 (0.55-1.32)	0.95 (0.70-1.29)
Dominant model	0.94 (0.66-1.33)	0.92 (0.64-1.33)	0.83 (0.56-1.25)	0.93 (0.65-1.32)	0.89 (0.49-1.61)	0.95 (0.63-1.42)
Recessive model	0.73 (0.40-1.33)	0.80 (0.43-1.50)	0.90 (0.46-1.74)	0.83 (0.47-1.47)	0.63 (0.24-1.67)	0.88 (0.44-1.75)
EPCAM rs1126497						
TT	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)
CT	1.08 (0.72-1.61)	1.06 (0.70-1.61)	1.14 (0.73-1.77)	1.05 (0.71-1.56)	0.95 (0.50-1.79)	1.09 (0.69-1.72)
CC	1.07 (0.66-1.74)	1.08 (0.65-1.80)	1.11 (0.63-1.95)	1.07 (0.67-1.72)	1.08 (0.49-2.37)	1.06 (0.61-1.84)
Log-additive	1.04 (0.82-1.32)	1.04 (0.81-1.34)	1.06 (0.81-1.40)	1.04 (0.81-1.33)	1.03 (0.70-1.51)	1.04 (0.79-1.36)
Dominant model	1.08 (0.74-1.57)	1.07 (0.72-1.58)	1.13 (0.74-1.72)	1.06 (0.73-1.55)	0.99 (0.55-1.77)	1.08 (0.71-1.66)
Recessive model	1.02 (0.67-1.55)	1.04 (0.67-1.62)	1.03 (0.63-1.68)	1.04 (0.68-1.58)	1.11 (0.54-2.27)	1.01 (0.63-1.62)
PPARGC1A rs3774923						
GG	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)	(Ref)
AG	0.94 (0.52-1.70)	0.86 (0.46-1.62)	0.90 (0.44-1.82)	0.89 (0.50-1.56)	1.44 (0.54-3.86)	1.04 (0.52-2.06)
AA	2.97 (0.19-47.55)	1.12 (0.06-20.47)	0.99 (0.05-18.95)	1.02 (0.29-3.57)	>999.999	1.45 (0.07-30.64)
Log-additive	1.02 (0.59-1.77)	0.89 (0.49-1.60)	0.91 (0.48-1.75)	0.91 (0.53-1.55)	1.74 (0.73-4.18)	1.06 (0.56-2.00)
Dominant model	0.98 (0.55-1.75)	0.87 (0.47-1.62)	0.90 (0.45-1.81)	0.89 (0.51-1.57)	1.66 (0.65-4.26)	1.05 (0.54-2.06)
Recessive model	3.00 (0.19-47.96)	1.17 (0.07-20.96)	1.02 (0.05-19.18	1.03 (0.30-3.59)	>999.999	1.45 (0.07-30.33)

CHAPTER 3 : GENOME-WIDE ASSOCIATION STUDY

3.1. Research Objectives and Methods

Using an agnostic, genome-wide approach, we seek to identify single-nucleotide polymorphisms (SNPs) that, if associated with NHL risk, would serve either to generate novel hypotheses or confirm prior investigations regarding the etiology of AIDS-NHL. After quality control and imputation for missing genotype values, we use logistic regression adjusting for ancestry-indicative principal components to examine ~5 million SNPs in ~2000 participants. We then use regional LD plotting, fine-mapping, and functional characterization to identify the most likely causal SNPs from regression output, and to assess the potential biological roles of these SNPs.

3.2. Specific Aims and Hypotheses

Aim 2 is informed by the same biological background as Aim 1, but shifts our analysis from one informed by prior knowledge to one that is fully agnostic with regard to SNPs of interest. Here we intend to:

1) Use secondary GWAS data from the MACS to assess genome-wide variation and associations with NHL risk.

3.3. Study Design and Methods

3.3.1. Study Overview

An unmatched case-control design using all 1,949 seropositive MACS participants with GWAS data who passed QC steps was used. Genotyped data from seven different chips (three Affymetrix; four Illumina) was augmented using the imputation program Minimac3[121,122], run on the Michigan Imputation Server[123] using the Haplotype Reference Consortium panel of ~39 million SNPs[124]. Logistic regression on genotype probabilities generated by Minimac3 was

carried out in SNPTEST[121,125,126], adjusting for the top three ancestry-informative principal components to account for population stratification[127-129].

3.3.2. Matched Versus Unmatched Design

The candidate-gene study analyzed in Chapter 2 used a matched approach. This was sensible given the particulars of that study, but study design cannot be a one-size-fits-all approach, and there are at least two aspects of the GWAS data that recommend against use of a matched design.

First, though matching is an excellent way to increase study efficiency (and reduce costs and effort) when designing a study and prior to the collection of data, this advantage is lost when data have already been collected. In our case, subjects have already been genotyped; it is not up to us to decide whom to genotype, but rather to decide which subjects to use. We save no costs, and save no resources, by matching versus using the full GWAS sample in an unmatched design.

Second, matching on an intermediate is always to be avoided, since this introduces bias that cannot be attenuated[130]. As discussed in Chapter 1, the relatively small (n=30) number of SNPs under consideration in the candidate gene study minimized concerns surrounding matching and adjustment for intermediates. In contrast, GWAS that are not replication studies by their nature entail agnosticism with regard to exposures of interest: the point of first-stage GWAS is to look across the whole genome for signals of association that can then be followed up on with replication studies or candidate-gene studies[131]. However, agnosticism with regard to exposures entails agnosticism with regard to intermediates, and if intermediates are unknown, then we cannot be confident that we are not matching on intermediates.

3.4. Study Population

Before quality control, we included 2,027 HIV-positive participants from the Multicenter AIDS Cohort Study (MACS) with available GWAS data. Of these 2,027 participants, 172 were cases

and 1,855 were controls. Ultimately, 1,949 participants (all 172 cases and 1,777 controls) were retained in the final analysis after all QC was performed. Original data were generated by the Center for HIV/AIDS Vaccine Immunology (CHAVI) consortium[70], with subsequent genotyping of additional participants by MACS. Recruitment into CHAVI was made primarily on the basis of seroconversion status, and participants come from cohorts 1 and 2. In our final post-QC dataset, 510 of 1,949 (26%) participants were seroconverters, and 1,439 (74%) were seroprevalent at first study visit. Among cases, 85% (n=147 of 172) were seroprevalent, and 73% of controls (n=1,292 of 1,777) were seroprevalent. For comparison, as of May 2015, 48% (n=3,280) of 6,821 participants in the MACS were HIV-negative throughout the follow-up period; of the remaining 52% of participants, 87% (n=3,071) were seropositive at baseline, and 13% (n=470) were seronegative at baseline.

Tables 3.1-3.2 present additional demographic information for NHL cases and controls at first seropositive visit (baseline visit, if participants were seropositive at enrollment).

3.4.1. Case and Control Definitions and Selection Criteria

Cases: All MACS participants who developed AIDS-NHL prior to January 2016, either as their first AIDS-defining-illness, following another AIDS-defining illness, or identified only via autopsy, for whom GWAS data were available. Cases were defined by self-report, with confirmation from medical records and cancer registry match; alternatively, post-mortem diagnosis was made from death certificates for participants deceased at the time of diagnosis.

Controls: All seropositive MACS participants for whom GWAS data were available, and who did not have a diagnosis of NHL as of 1 January 2016.

3.4.2. Data Collection

All data are secondary data, collected primarily by the Center for HIV/AIDS Vaccine Immunology using platforms as described below. CHAVI data collection was motivated primarily by a desire to understand the genetics of HIV seroconversion, and thus CHAVI targeted seroconverters rather than seroprevalent participants, carrying out genotyping on Illumina platforms. Additional data collection within the MACS, in conjunction with the National Cancer Institute, targeted seroprevalent participants, and was carried out using Affymetrix platforms[132]. Because seroprevalent status is positively correlated with NHL status, this led to a differential distribution of cases and controls across Affymetrix and Illumina chips, with a higher proportion of cases found in the Affymetrix data. Differential genotyping performance was corrected for by creating a common set of SNPs using imputation.

Data were subjected to preliminary quality control and analysis at the laboratory of James Gauderman, at the University of Southern California, and then transferred to UCLA for further analyses.

3.5. Genotyping

Samples were genotyped using six different chips on two different platforms. Illumina chips included the 1Mv1, 1MDuo, and HumanHap550; Affymetrix chips included the 407, 550, and 920. Not all cases were successfully genotyped on each platform, and each platform had a different number of successfully-genotyped cases. This led to an overall missing rate of 66%. This also means that for some SNPs in particular, the effective sample size was greatly reduced. In all, samples from 2,237 HIV-positive participants passed initial QC and were transmitted to UCLA; 2,027 of these participants were HIV-positive and thus were included in further QC routines.

Position overlap between chips is shown in Table 3.3; participant overlap between chips is shown in Tables 3.4-3.6. No participant was genotyped on more than two chips, and participants typed on two chips were typed on one Affymetrix and one Illumina chip, never two Illumina or two Affymetrix chips.

3.6. Imputation of Missing Genotypes

Coverage differed widely between chips; just ~120,000 of a total of ~1.6 million positions were genotyped on all chips. Data were imputed separately for each chip using the Michigan Imputation Server[123], which runs imputation using MINIMAC3. MINIMAC3 is a fast implementation of the MACH method; both MACH and previous incarnations of MINIMAC have been used in a host of papers[122,133].

Very briefly, imputation works by making reference to a reference genome and using a regression approach to infer missing genotype based on patterns of linkage disequilibrium in haplotypes[126]. We used the Haplotype Reference Consortium (HRC) reference panel, which includes 64,976 haplotypes at 39,235,157 positions[124]. This coverage, of 39 million positions, is far greater than the ~6 million positions covered in panels such as the HapMap. The sole disadvantage of the HRC is that it currently covers only European individuals, while we had a small number of non-European individuals in our sample. Despite this, two considerations outweigh this disadvantage: first, to guard against inaccurate imputation, we use strict post-imputation QC cutoffs to ensure a high degree of certainty in these imputations. Second, even if genotypes in non-European individuals do end up being imputed with low confidence, and thus excluded from the analysis, the overall number of positions passing QC using the HRC reference panel will still be higher than the overall number that would pass QC using HapMap, given the more than sixfold greater number of positions in HRC.

It is important to ensure that reference allele frequency between study data and the reference panel used for imputation is fairly close—especially since we have some non-European participants in our study. We paid special attention to this in preparing our data for submission to the Michigan server. This correspondence can be seen graphically in a plot of one frequency against the other, and summarized using the overall R^2, the global correlation between frequencies in each panel. On no chip was this value below 0.98, suggesting extremely good overlap with the reference panel, and high quality of the subsequent imputations.

The Michigan Imputation Server produced 22 files (one for each autosomal chromosome) for each of our seven chips. Post-imputation, extensive data management and quality control steps were undertaken to ensure high data quality of our merged dataset. We describe these procedures in the following sections.

3.7. Quality Control Measures

Data underwent extensive quality control. Three sets of quality-control procedures were carried out. First, USC performed essential first steps for association analyses. Upon receipt of data from USC, we imposed stricter procedures in preparation for imputation. Finally, QC was carried out on imputation results, all following standard practice for genome-wide association studies.

3.7.1. Procedures Prior to Data Receipt: USC Methods

The laboratory of James Gauderman, at USC, conducted initial QC on the data in PLINK format. Their procedures were as follows:

1) Sex mismatch: remove 3 samples
2) Remove samples with call rate <90%
3) Remove SNPs with call rate <95%
4) Remove samples with call rate <95%
5) Remove samples with chip-chip concordance <95%

6) Remove monomorphic SNPs
7) Remove A/T C/G SNPs
    - In 1M: overlapping 1MDuo
    - In Affy: Overlapping 1M or 1MDuo
8) Update Affy to b37; flip alleles to match Illumina
9) Remove (14 total) samples discordant between platforms, related individuals
10) Remove SNPs with unaff HWE p<0.00001 in each data set
11) Remove SNPs with unaff HWE p<0.00001 in combined data sets (Illumina only)
12) Remove SNPs with <95% reproducibility using study replicates (from Illumina)
13) PCA using EIGENSOFT 4.2, PCA with HapMap3 + study samples

3.7.2. Procedures Following Data Receipt: In-House Methods

Because these data were being used for imputation, we carried out additional QC using more stringent criteria upon receipt of the data from USC, following best practices from the GWAS literature[134-136]. These were as follows:

1. Exclude SNPs with a missing rate >3% on each chip.
2. Exclude individuals with a missing rate >3% on each chip.
3. Exclude SNPs with differential missingness between cases and controls (χ^2 p-value <1*10E-06) on each chip.
4. Exclude samples with excess heterozygosity which could indicate DNA contamination, relatedness, or population stratification.
5. Remove all A>T and C>G SNPs, facilitating alignment to the HRC reference panel across multiple chips with minimal loss of information.
6. Use the Michigan Imputation Server tool to correct for any inconsistencies between Haplotype Reference Consortium data and our chips, creating 22*7=154 final processed files for submission to the server.

It was especially important to ensure that SNPs of high genotyping quality and low missingness were used. We therefore chose conservative thresholds for four procedures. Per-SNP missingness (call rate) and per-individual missingness (sample coverage) used a threshold of 3%; differential missingness between cases and controls used a χ^2 p-value <1*10E-06. Individuals with an excess degree of heterozygosity (more than two standard deviations from the mean level of heterozygosity observed on each chip) were also excluded.

A>T C>G SNPs are the greatest source of ambiguity in aligning to a reference genome, or in ensuring consistency in genotyping across chips[135]. When doing both, as we were here, the issue is doubly important. The key complication with A>T and C>G SNPs is that because A pairs with T and C with G, the actual reads cannot be determined simply from genotype when merging across chips using different orientations: we cannot tell if we are dealing with a variant allele on the forward strand and a reference allele on the reverse strand, or vice versa. MAF can be used to establish which allele is read as the variant allele, but this breaks down as we approach an MAF of 0.5.

The bulk of A>T and C>G SNPs were already removed from Affymetrix chips by USC and thus represented <0.5% of SNPs on all chips, but were a major problem in pre-imputation QC. This is shown for the 1MDuo chip on the left-hand side of Figure 3.1, distinguished by the pronounced "X-pattern". Other authors exclude these as standard practice when comparing multiple chips, and excluding these has major impact on QC results, as shown on the right-hand side of Figure 3.1. Exclusion of A>T/C>G SNPs therefore solved a major QC issue with minimal loss of information.

Following elimination of A>T and C>G SNPs, we carried out strand alignment, specified reference alleles, and updated positions to the newest version of the b37 build for consistency

with HRC data. We used the Michigan Imputation Server's recommended tool, developed by Will Rayner[137]. Using PLINK[138,139], vcftools/bcftools[140] and Unix/Linux command-line input, this tool excludes SNPs for which no position match is found in the HRC data, updates positions to newer versions of the b37 genome build as desired, and corrects for strand flips and reference allele inconsistency across Illumina and Affymetrix platforms.

3.7.3. Post-Imputation Quality Control Measures

Post-imputation quality control must also be carried out. The steps followed can be summarized as:

1) Drop one chip: Exclude SC_1Mv1 from all calculations, owing to small sample size (n=71).

2) Chip-specific QC: For each of the remaining six chips, exclude all positions with chip-specific R^2 <0.3, chip-specific MAF <0.01, and any missing calls.

3) Cross-chip QC: Identify positions that passed these criteria on every chip (i.e. the intersection of all positions passing QC across the six chips).

4) Create common datasets: For each chromosome-specific file (n=22) from each chip (n=6), drop positions that did not pass cross-chip QC in step 3, to create a common set of SNPs with no missing data for our analyses.

5) Prioritize Illumina data over Affymetrix data: Drop the 1,010 participants who were genotyped using both Affymetrix and Illumina platforms from Affymetrix datasets (i.e. limit Affymetrix data to participants genotyped exclusively on an Affymetrix chip. Illumina had slightly better overall performance in the imputations, as measured using

mean R^2 and the number of SNPs passing post-imputation QC on each chip/each platform type.

5.5) Identify all SNPs common to Affy407, 550, and 920, and all SNPs common to Illumina SP_1Mv1, HH550 and 1MDuo.

6) Merge chromosome files across chips: from the 132 chromosome-specific files (22 chromosomes*6 chips), create 22 files by merging non-overlapping participants with uniform position reads into a single dataset for each chromosome.

7) Examine QQ plots from a round of preliminary association analyses for indications that QC was insufficient; if so, revisit QC procedures as needed.

8) Impose stricter MAF cutoff: Exclude positions with MAF <0.05 in these 22 merged files.

Per recommendations by the authors of MACH and the architects of the Michigan Imputation Server, after imputation, all positions for which r^2 was < 0.3, and for which MAF was <0.01, were identified and excluded from subsequent analyses[133]. Here, the r^2 measure is the estimate of correlation between the imputed and the theoretical observed genotype in the study sample, had the latter been observed. This serves as a measure of confidence in imputation results, and is equal to 1 for typed variants.

To assess whether these QC procedures were sufficient, we ran preliminary association analyses in SNPTEST using a dataset of 1,321,413 positions genotyped on at least one chip and 5,762,520 imputed positions. Following imputation, it is standard practice to investigate departure from the expected distribution of p-values using a Q-Q plot: deviation from this distribution, seen rapidly with a simple glance, can indicate an excess of false positives[135]. In turn, this can be due to one of several reasons highlighting issues with the analysis: 1) a large number of SNPs, leading to a

proportional increase in a large number of false-positives; 2) LD between SNPs, meaning that the number of independent signals is overestimated; 3) population stratification; or 4) flaws in the imputation itself.

Examination of QQ plots[141] from the first round of association analyses in SNPTEST revealed excessive departure from the expected distribution of p-values, suggesting a high number of false positive results. Out of concern that this departure was due to poor imputation, data were subset into four groups: 1) imputed only, MAF 0.01—0.05; 2) imputed only, MAF >0.05; 3) genotyped only, MAF 0.01—0.05; and 4) genotyped only, MAF >0.05. QQ plots for each subgroup indicated that low MAF rather than poor imputation drove this departure; the departure in imputed data reflects the higher proportion of low-frequency SNPs relative to genotyped data (as one would expect when imputing with a reference panel of 39 million SNPs). Therefore we restricted our analysis to SNPs with an MAF >0.05.

Positions were excluded if they had an MAF <0.05 in the *combined* dataset, not in the chip-specific datasets. Chip-specific exclusions were made based on an MAF of <0.01. Of the positions below, 5,508,998 were common to all chips post-QC and had an MAF ≥0.05; rsIDs were assigned to 4,859,136 of these 5,508,998 SNPs. Results plotted in the following sections and used in pathway analyses (Chapter 4) use the 4.8 million SNPs with rsIDs.

3.7.4. Summary of QC Procedures and Post-QC Data

Following all QC procedures as described above, we conducted analyses using a dataset of 4,859,136 SNPs. These SNPs passed all QC criteria on each SNP with MAF >0.05 within a population of 1,949 seropositive individuals. A summary of positions excluded at each stage of the QC process is given in table 3.7.1, below.

3.8. Generation of Combined Post-Imputation Dataset

In creating a combined dataset, the primary motivation was to maximize both the quantity and the quality of information. This entailed 1) dropping the smallest (n=71 participants) dataset--the Illumina SC1Mv1--and then 2) using Illumina data (from the SP1Mv1, HH550 and 1MDuo) for participants who were typed on both Affymetrix and Illumina. Overall, imputation results are of higher quality on the Illumina chips—with the exception of SC1Mv1—as evidenced by a higher mean R^2 than the Affymetrix chips. The SC1Mv1 data had the smallest number of SNPs passing QC, such that using SC1Mv1 would reduce the intersection of all SNPs by ~1.3 million. Furthermore, SC1Mv1 captured no unique cases, and >75% of its controls were also typed on Affymetrix chips. Thus by including SC1Mv1, we actually lose information, while we gain information by excluding it. Merge procedures therefore considered not only the number of SNPs passing QC, but also the quality of information.

3.9. Covariate Selection

In contrast to the candidate-gene study, we adjust only for ancestry, for three reasons. First, no other covariate meets the definition of a confounder; second, adjustment may in fact reduce our power; and third, because we are agnostic about exposure, we must also be agnostic about intermediates, and should avoid conditioning on any factor so as to avoid introducing bias by conditioning on an intermediate.

As described in Chapter 2, whether to adjust for covariates in genetic association studies is less clear-cut than one might think. With a GWAS, circumstances are different from a candidate-gene study, and there are at least three reasons not to control for covariates other than ancestry in our analysis. First, *a priori*, and assuming no chip- or platform-specific bias in genotype ascertainment, there is no reason to think that any factor besides ancestry would meet the

traditional definition of a confounder, i.e. a factor associated with both the exposure and the outcome.

Second, our study does not meet the specific conditions that must hold in order for other (non-confounding) covariates to actually increase power or reduce the number of false-positive associations; here, including them will in fact reduce power by increasing the standard error of the estimate, despite any increase in the magnitude of the association[89,91] . Specifically, cases and controls must be drawn from the general population (or the trait must be quantitative), and the prevalence of the disease under study must exceed ~20%. Here, controls and cases were drawn from the MACS cohort, largely on the basis of seroprevalent or seroconverter status; we are investigating NHL, a binary trait; and the prevalence of NHL in our sample is ~8%. Therefore the inclusion of non-confounding covariates in our analysis is not recommended.

Third, it is unclear by what biological mechanisms SNPs of interest might operate on the risk of NHL; controlling for covariates may result in our controlling on intermediate variables in the causal pathway from SNP to NHL. For instance, if a SNP's association with NHL in fact operates indirectly, via an impact on viral load or CD4 count, then adjusting for these variables would bias the SNP-NHL association toward the null[90,142].

More generally, the goal of covariate selection should, for epidemiologists, be the control of confounding and the honest representation of uncertainty in estimates. Statistical criteria such as stepwise procedures and the change-in-estimate criterion do not necessarily have these as their goal; instead they aim for model parsimony and reducing the variance of estimates[92]. As an example, these procedures can include weak confounders, or even non-confounders, while excluding from selection strong confounders and covariates that, based on prior knowledge and the causal structure of the data, are clearly confounders.

Our primary concern regarding covariate selection in the GWAS is that, since we are agnostic about relevant exposures, and since the huge number of SNPs examined may influence NHL risk indirectly, through a direct effect on intermediate variables such as VL or CD4, adjustment for covariates could in fact introduce bias[90,142]. If a variable is potentially within the causal pathway between exposure and outcome, then it should not be matched for, and should not be adjusted for. This was less of a concern in our candidate-gene study, where a small (n=30) number of SNPs on genes with well-defined biological function afforded us sufficient prior information to adjust for multiple covariates without worry of conditioning on intermediates.

Indeed, not adjusting for a wide set of covariates is not unusual in GWAS. Adjustment for ancestry, however, is essential. This process is described in the next section.

3.10. Adjusting for Ancestry: Principal Components Analysis

Principal components analysis seeks to explain the variance of a given sample by creating a small set of novel descriptive variables from the original data. This has the advantage of reducing the dimensionality of our data: instead of creating a model that adjusts for the full set of SNPs in a study, we can reduce this set to a small handful of features called eigenvalues (usually less than ten, and even two) that capture variance effectively. The eigenvalues represent the proportion of observed variance explained by ancestry[129].

Only genotyped data were used for PCA: PCA routines cannot accommodate probabilistic genotype calls from imputed data, only "hard calls," i.e. "best-guess" imputed genotypes. As discussed in the "Statistical Analysis" section, working with imputed genotypes in the same way one works with hard calls underestimates the degree of uncertainty in imputed genotypes and is not appropriate.

Genotype data were pre-processed in PLINK to remove non-autosomal chromosomes, to identify the intersection of all SNPs with 0% missingness (n=29,747 SNPs), and to remove regions of long-range (>2Mb) LD that can distort PCA estimates in admixed populations[129]. After removal of long-range LD regions, the remaining 29,226 SNPs were pruned for LD >0.2 using a 500,000 base-pair sliding window, leaving 17,947 SNPs.

The R package SNPRelate was then used to calculate principal components[143]. Following Weale, principal components analysis used an LD-pruned dataset created using an r2 threshold of < 0.2, a sliding window size of 2Mb, and a 10% window-size increment[136].

The first three eigenvalues explained 75.1% of the observed variance (63.9%, 6.4% and 4.8%), while the first ten eigenvalues explained 87.8% of the observed variance. A cursory examination of pre- and post-PCA adjustment plots indicates that PC correction was indeed helpful, as judged by comparing Q-Q plots of the departure of observed p-values from the expected distribution before and after PC adjustment, in Figures 3.3 and 3.4 respectively. We see that this departure is much less pronounced in Figure 3.4 than in 3.3, indicating that p-values obtained after PC adjustment are much more in line with what is expected under a null distribution.

As shown in Figure 3.2, the PCA plot did indicate the presence of some outliers, which is unsurprising given 1) a small number of "non-white" participants in the MACS data, and perhaps 2) greater admixture among populations in MACS study centers (Pittsburgh, Baltimore, Los Angeles and Chicago) than in Europe. To ensure that our adjustment was sufficient to account for stratification, the R package snpStats was used to calculate the genomic inflation factor λgc in our final post-imputation dataset, using 4.86 million SNPs[144].

The genomic inflation factor is the median of the observed chi-square distribution divided by the expected chi-square distribution; like Q-Q plots, it indicates the degree of departure from the

expected distribution of test statistics and is widely used to check for biases that may inflate false positive rates[145]. A λgc value of 1 indicates no inflation, and a standard cutoff for "acceptable" inflation is 1.02—1.05. In contrast, the λgc calculated from our data was 1.0002. This indicates an exceedingly low level of population stratification; as a result, we can be confident that the results explored here are not affected by such stratification and that our adjustment for principal components was indeed sufficient.

Furthermore, for the top SNP in each region reported in Section 3.12, analyses were re-run in a sample restricting the population to members of a cluster (n=1652) identified as white European using SNPRelate. This cluster is in fact a subset of the white European population, and as such represents an extremely homogenous group, yielding more rigorous control and better insight than would analyses subsetting the population to "white" participants only. (Sub-analyses in the non-white population for these SNPs were also attempted; sample sizes were too small to attain significance, or even case-/control-specific MAF necessary to calculate an association). Results from these analyses are reported alongside results in the full sample.

Correction for the PCs plotted in Figure 3.2 had a substantial effect on the distribution of p-values. Figure 3.3 and 3.4 present Q-Q plots before and after correction, and show that correction brings the observed distribution of p-values closer to the expected distribution, indicating a decrease in the number of potential false-positive results.

Q-Q plot in Figure 3.3 shows the expected distribution of p-values on the (red) diagonal, and the observed distribution of p-values as black circular points, prior to correction for PCs. The degree of departure from the expected distribution is excessive.

In contrast, Figure 3.4 shows the impact of adjusting for principal components. We can see that the distribution of observed p-values follows the expected distribution (the diagonal red line)

much more closely than in the uncorrected plot, with departures beginning only in the upper-right quadrant. Departures in the upper-right quadrant are to be expected if there are indeed meaningful true-positive signals, and this distribution is consistent with that seen in other published GWAS. The number of meaningful signals, and thus the extent of concern with false-positive results in our final analysis, is less than suggested by this plot: plotted data are not pruned for LD, while final analyses include careful consideration of the influence of LD on results.

3.11. Statistical Analysis

3.11.1 Logistic Regression Using SNPTEST

Genotype probabilities from MINIMAC output were analyzed using SNPTEST v2.5.2 in a Linux Red Hat environment[122,125,126]. We used an additive model and the score method/EM algorithm, which deals with genotype uncertainty using likelihood maximization from missing data theory.

Addressing this uncertainty properly is essential when analyzing imputed genotypes. Any imputed genotype will necessarily have a degree of uncertainty, since these genotypes are not "hard calls" from a genotyping platform, but rather "best guesses" for an unobserved genotype made on the basis of linkage disequilibrium patterns in a reference dataset in conjunction with known genotypes from a given study sample. While it is comparatively straightforward to work with known genotypes, the uncertainty surrounding the imputation program's "best guess" greatly complicates matters. Failing to account for this uncertainty will amplify the number of false positive associations observed during analysis.

A number of approaches are used to account for this uncertainty, but they can be broken down into two general categories according to the type of data used: the analysis of dosage data and the analysis of genotype probabilities. (A third option is simply to analyze "best-guess" genotype

data as "hard calls," i.e. as though they are not imputed, which is not recommended for the reasons described above). The use of genotype probabilities, though computationally demanding, is the most rigorous option, and thus was used here. P-values from logistic regressions conducted on these genotype probabilities in SNPTEST are presented in the Manhattan plot in Figure 3.5.

This Manhattan plot is just the starting point for our analysis. We build on this plot to address three questions for each region of interest. First, are observed signals independent, or are they due to linkage disequilibrium? Second, if they are due to linkage disequilibrium, what is the pattern of LD for the region of interest? Third, what are the likely biological functions of any causal variants?

Question 1 is answered by means of conditional association plots in SNPTEST. For question 2, we create regional plots in the R/Shiny program LocusExplorer illustrating patterns of LD[146]. For question 3, we turn to both LocusExplorer and the UCSC Genome Browser[147].

3.11.2. Assessing Independence of Signals: Conditional Plots in SNPTEST

When multiple SNPs in a given region show evidence of association, it is useful to run an analysis adjusting for the top SNP[148]. This tests whether results are due to linkage disequilibrium: if no other SNPs show evidence of association after adjustment for the top SNP, then they are not independent signals. If, however, one signal disappears and another remains, then this suggests independence, and further suggests that both regions should be followed up on in analysis, rather than just one.

3.11.3. Characterizing Linkage Disequilibrium: Regional Plots in LocusExplorer

To assess and depict the impact of LD on these results, we used LocusExplorer. LocusExplorer runs a host of routines in R that link to UCSC and other databases, with a web interface via Shiny. LD files are created using the National Cancer Institute's LDLink tool, with LD statistics

calculated based on 1000 Genomes data[149]. A file for all SNPs in LD with the region's top SNP, as captured in the HapMap CEU population, is created and fed into LocusExplorer, and association results from SNPTEST are then plotted alongside LD information.

3.11.4. Identifying Potential Biological Roles: LocusExplorer and UCSC Genome Browser

Once putative causal variants have been identified, work turns to functional characterization. This includes the use of databases such as UCSC and others to determine what biological role a given variant might play.

LocusExplorer plots present two tracks from the UCSC Genome Browser and ENCODE[150] that reflect important information about gene regulation and transcription: the H3K27Ac histone mark, and deoxyribonuclease (DNase) hypersensitive regions[150]. The motivation here is to explore the potential regulatory impact of observed variation. H3K27Ac denotes a modification of histone proteins via mechanisms such as methylation, which alters chromatin accessibility and in this case may promote transcription[151]. DNase catalyzes DNA cleavage; regulatory regions, especially promoters, are known to be DNase-sensitive.

3.12. Results

3.12.1. Overall Association Results: Manhattan Plots

Without considering LD, we have identified ten SNPs with p-values lower than the p=5E-08 threshold for genome-wide significance, seven of which appear on the basis of location alone to be in high LD. This can be seen in the Manhattan plot in Figure 3.5, where a large number of SNPs on chromosome 4 are "stacked" on top of each other due to their close proximity on the chromosome. Even ten SNPs attaining genome-wide significance is a surprisingly large number, and thus it is reasonable to assume that most, if not all, of these associations are due to LD or are false-positives, especially given low MAF and large standard errors for many of these variants.

Other regions of interest, though not attaining genome-wide significance, did meet the p=5E-06 cutoff for a "suggestive" level of significance, and show the same "stacking" pattern indicating likely LD between SNPs. This can be seen on chromosomes 2, 11, 12, and also in the ~31Mb region of chromosome 4 (4p15.1). We investigate each of these regions further below.

Because of LD, a table presenting results for the top-performing SNPs (i.e. with the smallest p-values) across all 22 chromosomes contains many dozens of rows and is too unwieldy for the main text. Instead, we include chromosome-specific tables below in addition to our figures; where more than one region of interest is identified on a single chromosome (as with chromosome 4), we present tables for each region. Where typed SNPs appear in the top ten list, we contrast results with the original USC dataset to assess any divergence between associations estimated using genotyped-only data and data including both genotyped and imputed SNPs. A table with results for the top 500 SNPs across all 22 chromosomes is included as Appendix B at the end of this dissertation.

3.12.2. Chromosome 18 (18q21.32)

The top SNP in our study (p=8.84E-10; OR=0.39, 95%CI 0.29-0.53) is rs4356576, chromosome 18q21.32. This result appears to be spurious. This inverse association was not seen in the genotyped-only dataset provided by USC (n=834, OR=1.222, P=0.636); both the USC data and the imputed data used here were re-checked to ensure that no coding errors were made, but none were found.

There are at least three explanations for this result. First, the indel rs56238208 also occurs at the rs4356576 locus (18:57696044), which may have compromised the original genotyping process for rs4356576. Second, the imputation quality for rs4356576 is low relative to the rest of our SNPs: its R^2 value in MINIMAC output was just 0.3015; our cutoff is <0.3. Third, the shift in

results also could be due to a small sample size (n=836 for this SNP) and consequent lack of precision in the USC dataset. However, the qualitative shift from a positive association to an inverse association, the presence of the indel 56238208 at the same locus, the relatively poor imputation quality, and the absence of any other strongly-associated signals in the immediate region (Figure 3.5) suggests that rs4356576 is a false-positive result. We therefore regard the result for rs4356576 as an aberration rather than a meaningful signal.

3.12.3 Chromosome 4: 171-172Mb (4q33)

3.12.3.1. Logistic Regression in SNPTEST: Top Ten Results

Table 3.8 presents the top ten SNPs in the ~171-172Mb (4q33) region on chromosome 4. The top SNP in this region was rs2195807 (p=1.48E-08; p=1.93E-07 in the white-only subset). As we see, all but one of these SNPs was imputed; however, in USC's genotyped-only data, several SNPs in this region performed similarly.

3.12.3.2. Conditional Analysis in SNPTEST: Adjusting for rs2195807

Figure 3.6 shows the association of SNPs on chromosome 4 with NHL, after adjustment for rs2195807 and the top 3 PCAs. Results suggest that the 4q33 signal is independent of others on the chromosome, as results for other regions do not change. Because no signal persists in the 4q33 region after correction, the plot also suggests that rs2195807 is the sole signal in this region, and that results are shaped by LD, rather than multiple independent signals.

Thus we see that upon adjustment for rs2195807, the signal for all SNPs in the 4q33 region was erased, and that the signal for the ~50Mb region was largely unchanged. This suggests—consistent with our regional plots—strong LD in the 4q33 region, but also statistical independence between the ~50Mb and ~170Mb (4q33) regions.

It should be noted that these results do not establish rs2195807 as the causal variant. As addressed in the next section and also in the Discussion, this is a matter for fine-mapping and functional characterization, discussed below.

3.12.3.3. LD Characterization and Functional Annotation in LocusExplorer

Figure 3.7 maps all 961 unique SNPs for which we had both SNPTEST and LDlink output with LD $R^2 > 0.2$ in the vicinity of rs2198507. This plot highlights four major points. First, we can distinguish typed from imputed SNPs, on the green (second) line. Typed positions are also denoted by triangles on the plot, and imputed positions by circles. Second, that there is high LD across a wide region, as shown by the pink shading of points on the plot (darker colors indicate higher LD), and as expected by the lineup of points on the Manhattan plot (Figures 3.5 and 3.6). Third, this SNP is in high LD with SNPs falling directly on LOC100506122, which suggests that one or several of these variants may in fact be the causal variant or variants. Fourth, this SNP is in LD with SNPs across a region showing little DNase hypersensitivity or H3K27Ac histone marks (colored peaks in the yellow band below the plot). On the basis of the data on hand in the plot, this suggests that activity connected with LOC100506122 may account for observed associations.

3.12.4. Chromosome 4: ~31Mb Region (4p15.1)

3.12.4.1. Logistic Regression in SNPTEST: Top Ten Results

Table 3.9 presents the top ten SNPs falling below 50Mb on chromosome 4 (4p15.1). The top SNP in this region was rs35528558 (p=3.20E-07; white-only p=1.93E-07). Again, we see the 31Mb region (4p15.1) is most prominent, but a secondary signal may also be present at ~36Mb (4p14). We investigate this using conditional plots in the next section.

Two SNPs in Table 3.9 were typed: rs6820873 and rs12651329. In the USC data, rs6820873 showed an OR of 1.69 (vs. 1.55 here) and a p-value of 1.02e-5 (vs.4.02e-06 here); rs12651329 showed an OR of 0.60 and a p-value of 8.73e-04 in genotyped data. The shift toward the null in imputed data, coupled with smaller p-values, is encouraging: it suggests greater precision in imputed estimates, and the absence of a qualitative shift in ORs again suggests the absence of issues with reference-allele reads or strand flips in the imputation process.

3.12.4.2. Conditional Analysis in SNPTEST: Adjusting for rs35528558

Figure 3.8 shows the results of logistic regression in SNPTEST adjusting for rs35528558 in addition to the top three principal components. Interestingly, there appear to be three independent signals of interest in this region: at ~31, ~36 and ~40Mb. The top SNPs for each are rs35528558 (p=3.20E-07), rs7670868 (p=6.18E-06), and rs28802045 (p=1.35E-05), respectively.

3.12.4.3. LD Characterization and Functional Annotation in LocusExplorer

Figure 3.9 shows that rs25528558 lies directly on PCDH7, which codes for protocadherin-7, a protein active in cell adhesion. Furthermore, rs35528558 is in LD with SNPs spanning a region of DNase hypersensitivity and the H3K27Ac mark, suggesting regulatory activity in this region.

3.12.5. Chromosome 2 (2q36.1)

3.12.5.1. Logistic Regression in SNPTEST: Top Ten Results

Table 3.10 presents the top ten SNPs on chromosome 2. The top SNP in this region was rs17433868 (p=5.32E-07; white-only p=7.42E-06). As we see, all but one of these SNPs were imputed; however, in USC's genotyped-only data, several SNPs in this region (e.g. rs1897121, 2:222223235; OR=0.431, p=1.43E-05) performed similarly.

3.12.5.2. Conditional Analysis in SNPTEST: Adjusting for rs17433868 (2q36.1)

Figure 3.10 presents the results of logistic regression in SNPTEST adjusting for rs17433868 under an additive model, to determine whether the cluster of suggestive SNPs in the ~222Mb (2q36.1) region represents one signal characterized by LD, or multiple independent signals. We see no change in other signals upon correction, suggesting that this signal is indeed one signal characterized by LD, and that a more detailed investigation of this pattern in concert with functional annotation is merited.

3.12.5.3. LD Characterization and Functional Annotation in LocusExplorer

Having established in conditional association plots that our signal of interest is a single signal characterized by LD, we unpack the details of this LD pattern and provide functional annotations via the regional plot in Figure 3.11. We see that rs17433868 is in LD with SNPs lying on EPHA4, which codes for a receptor tyrosine kinase. The SNP rs17433868 is also in LD with SNPs lying within regions of histone and DNase activity, suggesting possible regulatory action.

3.12.6. Chromosome 11 (11p15.3 & 11p15.4)

3.12.6.1. Logistic Regression in SNPTEST: Top Ten Results

Table 3.11 presents the top ten SNPs on chromosome 11. The top SNP in this region was rs56289978 (p=2.31E-07; white-only p=3.20E-07). All SNPs in this table were imputed. We adjust for the top SNP Table 3.11, rs56289978, in the next section.

3.12.6.2. Conditional Analysis in SNPTEST: Adjusting for rs56289978

Figure 3.12 shows that adjustment for rs56289978 (part of the cluster of SNPs on the leftward side of the top plot, at ~11Mb), with $p < 1E-06$, does indeed do away with this signal; however, a signal to the right does persist. We address the first signal in subsequent analysis.

3.12.6.3. LD Characterization and Functional Annotation in Locus Explorer

Figure 3.13 shows limited LD for this region, and it shows also that rs56289978 falls within a desert of sorts. The sole suggestive feature in the plot below is a faint marker of DNase hypersensitivity, on which two SNPs in LD with rs56289978 fall. The nearest gene is GALNTL18, but this is over 100KB away.

3.12.7. Chromosome 12 (12q13.13 & 12p13.33)

3.12.7.1. Logistic Regression in SNPTEST: Top Ten Results

Table 3.12 presents the top ten SNPs on chromosome 12. The top SNP is rs11169939, at band q13.3, position 52271467 (p=1.02E-06; white-only p=9.63E-05). All SNPs except for rs11169945 were imputed. In USC's genotyped-only data, rs11169945 had an OR of 2.105 and a p-value of 4.76E-04. This is consistent with the OR of 2.022 observed here, while the smaller p-value observed here (8.27E-06) can be explained by the larger sample size resulting from the use of imputed data. The ~52.2Mb 12q13.13 region is represented heavily in this table and is therefore our primary region of interest; however, one SNP, rs35103713, falls at ~3Mb (12p.13.33: position 3077006).

3.12.7.2. Conditional Analysis in SNPTEST: Adjusting for rs11169939

Here we adjust for rs11169939, the top SNP in the ~52MB region (12q13.13) in the table above. Figure 3.14 shows that adjustment for this SNP does eliminate the signal in this region, suggesting that further investigation of LD patterns and functional characterization is worthwhile.

3.12.7.3. LD Characterization and Functional Annotation in Locus Explorer

Figure 3.15 shows that the LD pattern here is quite straightforward compared to that for other SNPs seen thus far, but spans ~100KB. However, the histone and DNase patterns observed here

are especially pronounced. The SNP rs11139939 lies very close to, or on, ANKRD33, and is in LD with SNPs on ACVRL1.

3.13. Discussion

We identified a genome-wide significant (p<5E-08) signal marked by pronounced LD on chromosome 4q.33. Here, the top SNP in the region (rs2195807; p=1.48E-08; white-only p=1.93E-07) is in high LD with SNPs falling directly on the uncharacterized noncoding variant LOC100506122. This variant does not fall within a region marked by significant regulatory activity (as measured by DNase hypersensitivity or H3K27Ac histone marks), and is not in close proximity to any coding variants. This could suggest that LOC100506122 has an as-yet-uncharacterized regulatory function, but mechanistic explanations for the association of rs2195807 with NHL risk remain unclear.

Using regional plots to unpack LD patterns and highlight nearby regulatory elements, we investigated suggestive associations (p<5E-06) in regions 4p15.1, 2q36.1, 11p15.3 and 12q13.13. We found that the top SNPs in these regions, with the exception of rs56289978 on 11p15.3, were either: 1) in LD with SNPs in regulatory regions (all SNPs); 2) in LD with SNPs falling on genes including the receptor tyrosine kinase gene EPHA4 (rs17433868, 2q36), the ankyrin repeat domain gene ANKRD33 and the activin A receptor-like type 1 gene ACVRL1 (rs11139939, 12q13); or 3) directly on the protocadherin-7 gene PCDH7 (rs25528558, 4p15.1).

We were unable to duplicate genome-wide significant or suggestive findings from four meta-analyses of NHL GWAS in the general population that assessed CLL[152], FL[68], MZL[64] and DLBCL[66]. Table 3.13 presents odds ratios ("Meta-OR") and p-values ("Meta-P") for the 29 SNPs attaining significance in each of these four meta-analyses alongside results from the present study ("Study OR" and "Study P"). We had results for 17 of these 29 SNPs; the other 12

failed either pre- or post-imputation QC procedures. The table shows that for these 17 SNPs, none approached the level of significance reached in the meta-analyses, and none approached a suggestive level of significance: p-values for these SNPs ranged from 2.20E-01 to 1.0.

There are at least four possible explanations for this, none of which are mutually exclusive. First, the small size of our study relative to these meta-analyses obviously plays a role. Second, with the exception of DLBCL, the distribution of subtypes in our study does not include a sufficient number of CLL, FL, or MZL cases. Relative to BL, DLBCL, and PCNSL, these subtypes are not common in PLWHA, and since their mechanisms of pathogenesis differ as well, it may be that relevant polymorphisms will also differ according to subtype. Third, the SNPs identified in this study may work to influence NHL risk primarily through HIV-related mechanisms (e.g. interaction with HIV viral proteins[153]) not prominent in general-population studies. Fourth, we did not impute the HLA region, which showed five independent associations with NHL risk in MZL, FL, and DLBCL (Table 3.13). Because of extensive LD in this region, specialized software is needed to conduct imputation[154], and this option is not available in the Michigan Imputation Server. That said, previous GWAS identified promising associations both inside and outside the HLA region, and we did not observe these associations outside the HLA region either.

We were also unable to find entries in the NHGRI-EBI Catalog of Published Genome-Wide Association Studies[104] for any of the SNPs reported here, with the exception of rs12651329 (4p14), associated with age of onset in amyotrophic lateral sclerosis (p=4.36E-06)[155]. However, the NHGRI-EBI catalog lists only SNPs with associational p-values < 1.0E-05, meaning that associations below this threshold may have been identified but not reported in the Catalog.

To better characterize suggestive signals in this study, we used regional plotting and functional analysis. In the analysis of GWAS results, one straightforward option for dealing with LD is simply to prune results, excluding all variants with an LD R^2 threshold above a certain number (usually 0.2). Alternatively, investigators may simply declare the variant with the smallest "significant" p-value lying closest to a gene of interest the causal variant. However, each of these two approaches discard a wealth of useful information, and may lead to incorrect conclusions. Proximity does not necessarily establish causation. Long-range LD (accounted for here) and regulatory influence of variants at a distance can lead to far-away variants being causal. Furthermore, the p-value is itself determined largely by the circumstances of our study, not by biology: a failure to detect meaningful signals can indicate shortcomings in study design or lack of power, rather than the absence of biological association or mechanism. Similarly, an excess of meaningful signals may be due to false-positive results also arising from study design and power issues.

In contrast, regional plotting and functional analysis make full use of LD information, and use this information to provide far more nuanced and biologically-informative results. The bulk of our fine-mapping and functional analysis here has focused on potential regulatory roles for our variants of interest, since most fall in intergenic regions. However, while providing some insight into potential mechanisms, this work also highlights the rather limited understanding of regulatory biology at present: though the vast majority of SNPs in GWAS fall in non-coding regions, the mechanisms by which they might operate to influence regulation are not always clear.

Understanding of these mechanisms may be limited further by the tissue- and phenotype-specific activity of particular variants, and a lack of sufficiently broad data across tissues and phenotypes.

In a massive survey of autoimmune disease, Farh *et al.* found that only 12% of autoimmune disease candidate variants were eQTLs[156,157], variants associated with different levels of expression as measured through RNAs. They suggest that precisely the tissue- and disease-specific activity just described accounts for this, and that the low proportion of eQTLs reflects a lack of sufficient data rather than an absence of biological mechanism—meaning that expanded collection of data is key to enriching our biological understanding.

3.14. Strengths and Limitations

3.14.1. Strengths of this Study

One strength of this study is our move beyond Manhattan plots to fine-mapping and functional annotation. Viewing LD data as a source of insight rather than as a nuisance can pay dividends in the interpretation of GWAS data. Another strength is our use of imputation to generate a common set of SNPs across multiple platforms, avoiding issues with missing data and increasing power.

3.14.2. Limitations of this Study

One potential limitation of the study is the excess number of significant findings: we observed 128 SNPs exceeding the threshold for "suggestive" significance. This is a greater number than expected, exacerbating the concern with false positives common to every GWAS. However, the bulk of these associations appear to be due to LD, which we account for and present fully in our regional association plots. In addition to LD, the large number of "significant" findings may be due to the fact that the Haplotype Reference Consortium panel covers almost seven times as many positions than other panels such as HapMap (~39 million vs. ~6 million positions), and thus we are analyzing a larger number of SNPs than many other studies. Though our proportion of false positives is similar, the absolute number is greater, and more visually striking on the

Manhattan plot in Figure 3.5 (the bulk of SNPs assessed in any study are of low significance, and thus are simply absorbed into the solid portions of the plot, meaning that no difference between results for 1 million and 5 million SNPs would be apparent). The Haplotype Reference Consortium panel was released only in late 2015, and there are no publications with which to compare our results. It will be interesting to see these once they are published.

At the same time, restricting the set of SNPs for analysis to those with MAF >0.05 means that low-frequency variants will be missed. Though this was necessary for proper quality control, and though these SNPs would likely not have shown significant associations given their low frequency (leading to more uncertain estimates), restriction of SNPs on the basis of MAF necessarily introduces an ascertainment bias.

Another potential limitation—again, common to many GWAS—is that we rely on tagSNPs for genotyping, and on imputation to fill in missing genotypes. Imputation necessarily leads to some degree of uncertainty in genotype data. We account for this in our analysis by conducting logistic regression on the posterior probabilities of each genotype for each position, rather than on "best-guess" binary data. In contrast, whole-genome sequencing, or high-density genotyping of regions of interest, would not rely on tagSNPs or imputation, and would allow for the use of "hard-call" data, which would strengthen any observed associations.

Yet another limitation is our decision not to impute the HLA region. As discussed above, this requires specialized software unavailable via the Michigan Imputation Server. Doing so could well produce interesting results, and it would no doubt allow for better comparison against existing meta-analyses of NHL GWAS, which identified a prominent role for the HLA region in MZL[64], DLBCL[66] and CLL[152].

Finally, the limitations of single-SNP associations, discussed in the context of our candidate-gene study in Chapter 2, apply to GWAS as well. Accurate, or even approximate, descriptions of biological reality may need to operate at a much greater level of complexity than the association between a single SNP and a phenotype of interest. This is the topic of our next chapter.

3.15. Further Directions

In this chapter, we have concerned ourselves with identifying one, or at most a handful, of variants associated with NHL risk, primarily via methods that exploit linkage disequilibrium patterns. The next chapter takes a different approach to identifying genetic variation associated with NHL: rather than looking at just one or a few variants, it combines evidence from all variants with sufficiently strong signals, and scales up from single-SNP associations to genes, and from genes to biological pathways. We contrast these pathway approaches with the analyses in this chapter, and situate overarching thoughts on strengths and limitations, within the context of the dissertation as a whole in Chapter 5.

Table 3.1. Characteristics of NHL Controls (n=1,777) at First Seropositive Visit

	Median/Count	SD	Min	Max
n	1865	-	-	-
age (median, at dx)	32	7.77	18	78
HIV VL at set point	20504.0	124981.0	40.0	2434693
CD4 slope pre-HAART (cells/year)	-55.56	91.50	-1545.29	392.31
HCV-positive	97 (5.5%)	-	-	-
Smoking status at visit				
Never	655 (37.5%)	-	-	-
Former	389 (22.3%)	-	-	-
Current	702 (40.2%)	-	-	-
Alcohol use since prior visit				
None	126 (7.3%)	-	-	-

1-3 drinks per week	550 (31.8%)	-	-	-
4-13 drinks per week	732 (42.3%)	-	-	-
>13 drinks per week	321 (18.6%)	-	-	-
Missing	48 (2.7%)	-	-	-
Other drug use since prior visit ("yes")				
Cannabis	1279 (73.0%)	-	-	-
Cocaine	714 (40.8%)	-	-	-
Uppers (inc. methamphetamine)	447 (25.5%)	-	-	-
Heroin/opiates	32 (1.8%)	-	-	-
Speedball	0 (0%)	-	-	-

Table 3.2. Characteristics of NHL Cases (n=172) at First Seropositive Visit

	Median/Count	SD	Min	Max
n	172	-	-	-
age (median, at dx)	34	7.30	19	56
HIV VL at set point	31133	131607.6	356	960960
CD4 slope pre-HAART (cells/year)	-64.23	71.13	-283.38	178.30
HCV-positive	16 (9.4%)	-	-	-
Smoking status at visit				
Never	69 (41.1%)	-	-	-
Former	43 (25.6%)	-	-	-
Current	56 (33.3%)	-	-	-
Alcohol use since prior visit				
None	13 (7.7%)	-	-	-
1-3 drinks per week	57 (33.8%)	-	-	-
4-13 drinks per week	70 (41.4%)	-	-	-
>13 drinks per week	29 (17.2%)	-	-	-
Missing	3			
Other drug use since prior visit ("yes")				

Cannabis	124 (72.9%)	-	-	-
Cocaine	60 (35.3%)	-	-	-
Uppers (inc. methamphetamine)	44 (26.0%)	-	-	-
Heroin/opiates	3 (1.8%)	-	-	-
Speedball	0 (0%)	-	-	-

Table 3.3. Overlap between Chips Prior to Imputation: Positions

Overlap Between Chips: Positions							
	SP_1MDuo	**SC_1Mv1**	**SP_1Mv1**	**HH550**	**Affy407**	**Affy550**	**Affy920**
SP_1MDuo	1029208	863961	879188	484840	249129	244512	247834
SC_1Mv1		893184	857161	479008	213784	208333	211121
SP_1Mv1			908821	476351	213853	210812	213828
HH550				493076	132670	127719	129694
Affy407					787240	733965	736749
Affy550						787124	759135
Affy920							788542

Table 3.4. Overlap between Chips, Cases and Controls

Overlap Between Chips: Cases and Controls							
	SP_1MDuo	**SC_1Mv1**	**SP_1Mv1**	**HH550**	**Affy407**	**Affy550**	**Affy920**
SP_1MDuo	264	**0**	**0**	**0**	**0**	**24**	**43**
SC_1Mv1		71	**0**	**0**	**46**	**3**	**7**
SP_1Mv1			578	**0**	**0**	**173**	**331**
HH550				484	**319**	**16**	**48**
Affy407					380	**0**	**0**
Affy550						545	**0**
Affy920							904

Table 3.5. Overlap between Chips, Controls Only

Overlap Between Chips: Controls Only							
	SP_1MDuo	SC_1Mv1	SP_1Mv1	HH550	Affy407	Affy550	Affy920
SP_1MDuo	261	0	0	0	0	23	42
SC_1Mv1		67	0	0	43	3	6
SP_1Mv1			501	0	0	156	281
HH550				463	306	14	47
Affy407					364	0	0
Affy550						491	0
Affy920							817

Table 3.6. Overlap between Chips, Cases Only

Overlap Between Chips: Cases Only							
	SP_1MDuo	SC_1Mv1	SP_1Mv1	HH550	Affy407	Affy550	Affy920
SP_1MDuo	3	0	0	0	0	1	1
SC_1Mv1		4	0	0	3	0	1
SP_1Mv1			77	0	0	17	50
HH550				21	13	2	1
Affy407					16	0	0
Affy550						54	0
Affy920							86

Figure 3.1. Quality-Control Plots from the Michigan Imputation Server Before and After Elimination of A.T and C>G SNPs, Illumina 1MDuo Chip

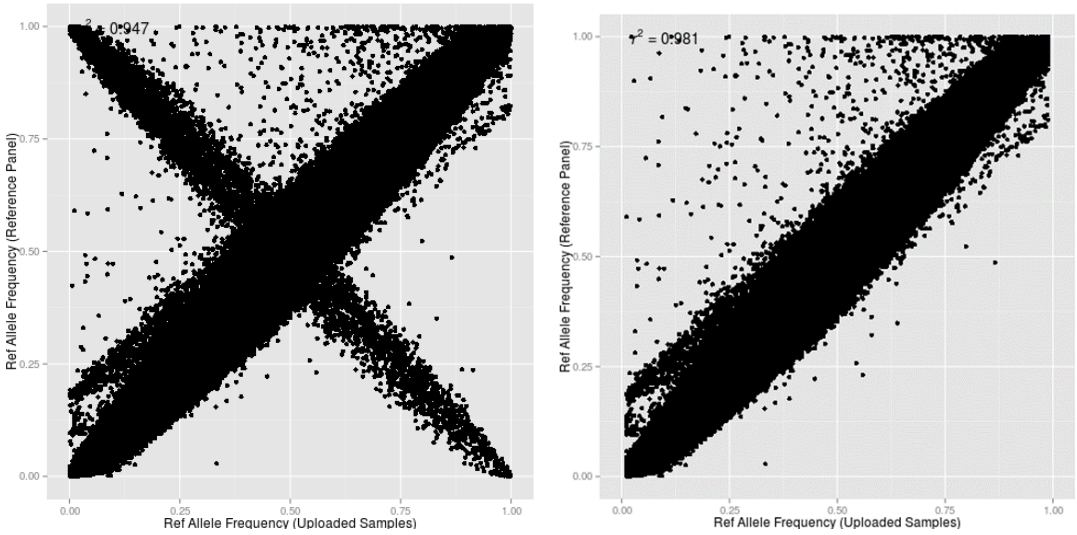


Table 3.7. Summary of Positions Before and After QC Procedures

Chip Name	Positions on Chip	Input to Imputation Server	Returned by Server	Excluded: r2 <0.3	Excluded: r2 <0.3 & MAF <0.01	Post-QC Positions
SP_1MDuo	1,029,208	942,166	39,230,259	19,790,382	29,154,934	10,075,325
SC_1Mv1	893,184	825,847	39,221,451	26,942,692	30,238,932	8,982,519
SP_1Mv1	908,821	805,073	39,221,451	18,516,348	30,862,033	8,359,418
HH550	493,076	472,251	39,210,718	19,875,791	30,804,525	8,406,193
Affy407	787,240	681,702	39,230,259	26,257,226	31,590,862	7,639,397
Affy550	787,124	656,081	39,230,259	19,834,807	30,957,965	8,272,294
Affy920	788,542	666,384	39,230,259	16,979,006	30,700,470	8,529,789

Figure 3.2. Plot of Principal Components Output, SNPRelate

Figure 3.3. Q-Q Plot Prior to Correction for Principal Components

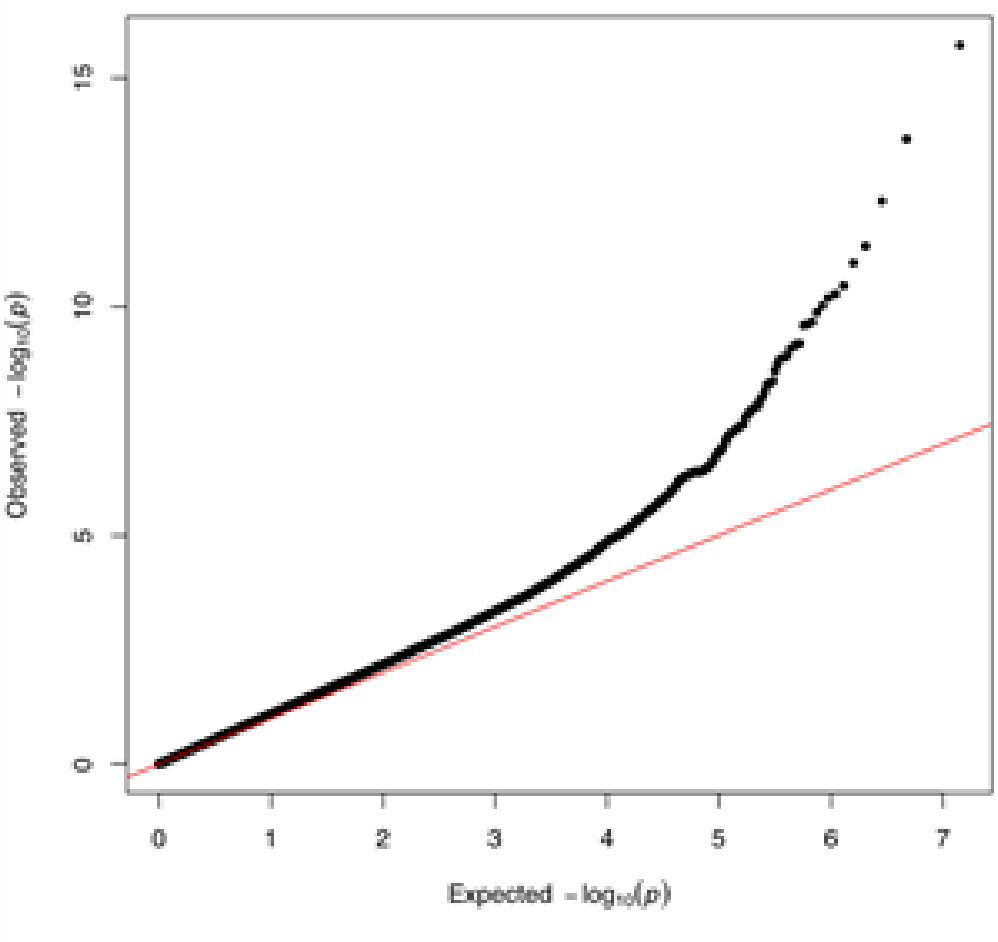


Figure 3.4. Q-Q Plot Following Correction for Top Three Principal Components

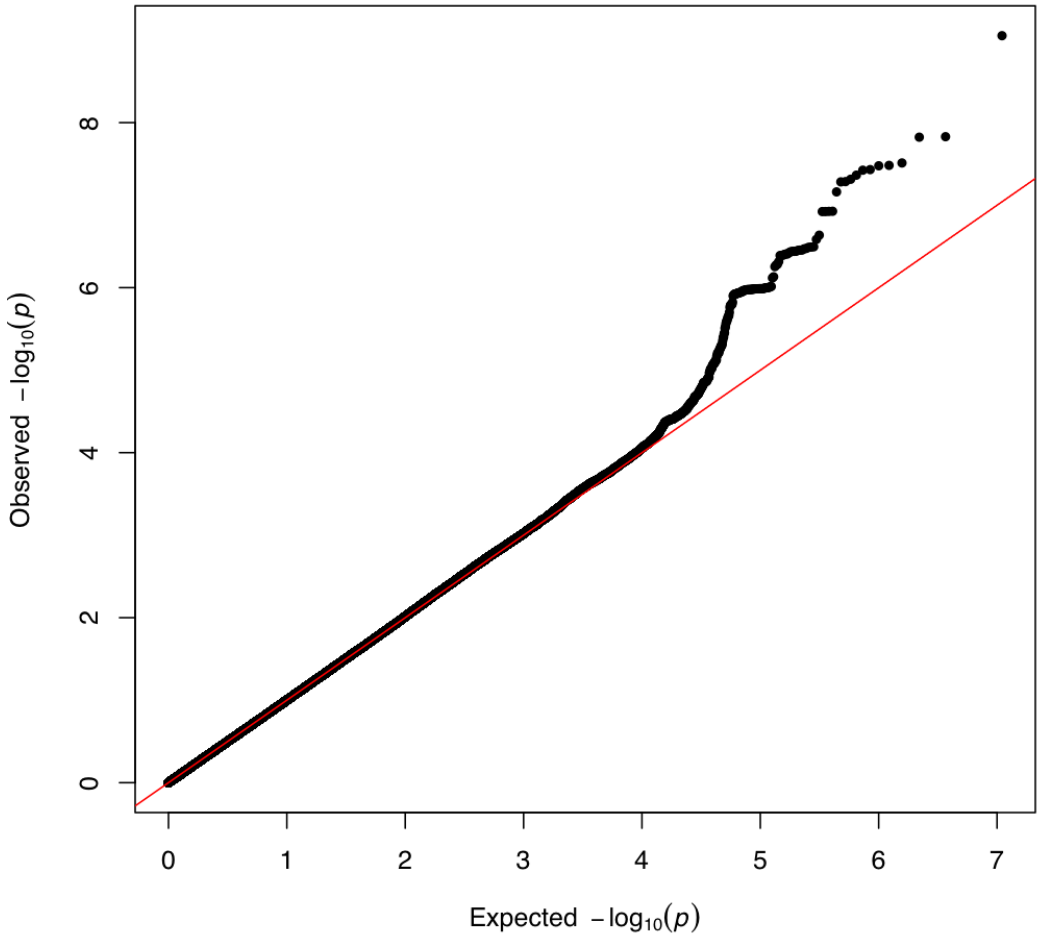


Figure 3.5. Manhattan Plot of Results, Corrected for Top Three Principal Components

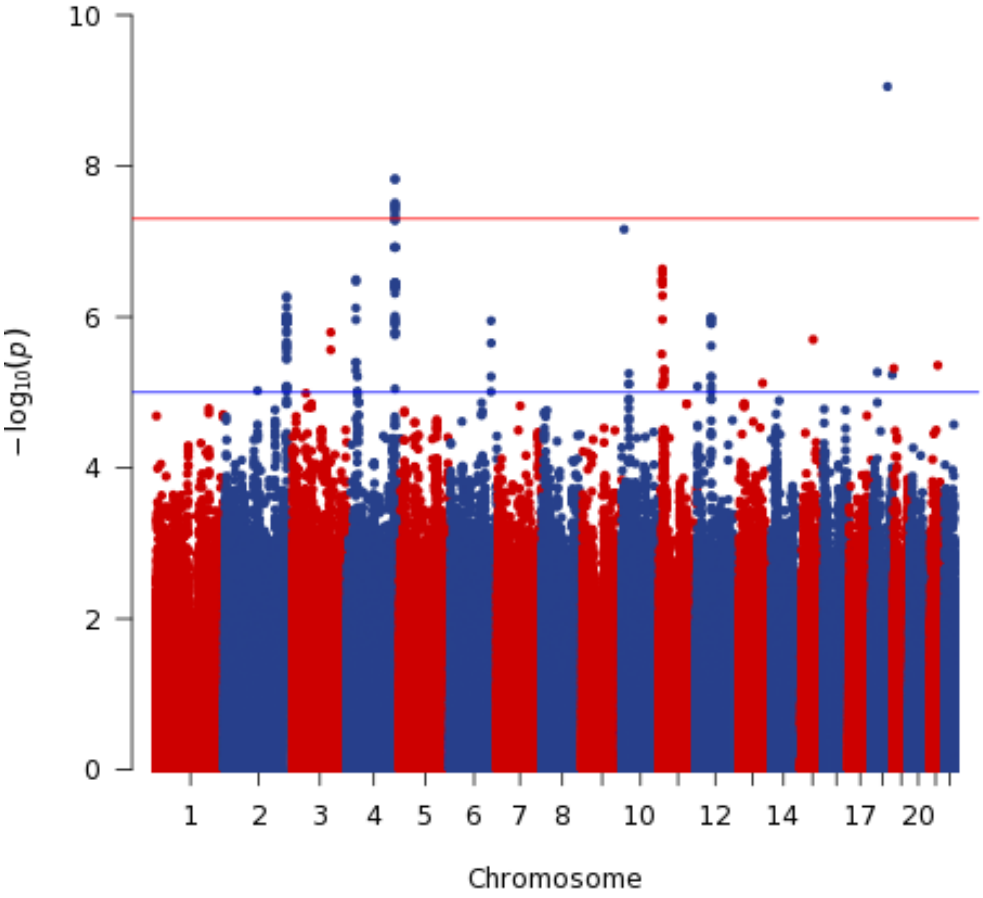


Table 3.8 Odds Ratios and P-Values for Top Ten SNPs on Chromosome 4:171-172Mb (4q33).

rsID	BP	A1	A2	MAF	OR	LL	UL	P	Typed/Imp.
rs2195807	171863210	C	T	0.054	2.33	1.60	3.39	1.48E-08	Imputed
rs80111333	171866588	T	G	0.054	2.32	1.60	3.37	1.50E-08	Imputed
rs13434452	171867485	T	A	0.057	2.20	1.52	3.20	3.08E-08	Imputed
rs10213010	171848005	T	G	0.055	2.31	1.59	3.35	3.30E-08	Imputed
rs10212953	171847608	A	G	0.055	2.31	1.59	3.35	3.34E-08	Imputed
rs28666968	171853059	A	G	0.055	2.29	1.58	3.33	3.77E-08	Typed
rs10009004	171859056	T	C	0.055	2.28	1.57	3.30	4.33E-08	Imputed
rs10049542	171878690	T	C	0.055	2.26	1.55	3.28	4.86E-08	Imputed
rs17056352	171890188	G	A	0.056	2.26	1.56	3.28	5.19E-08	Typed
rs79663997	171890355	G	A	0.056	2.26	1.55	3.28	5.22E-08	Imputed

Figure 3.6. Conditional Analysis Adjusting for rs2195807, Chromosome 4 (4q33)

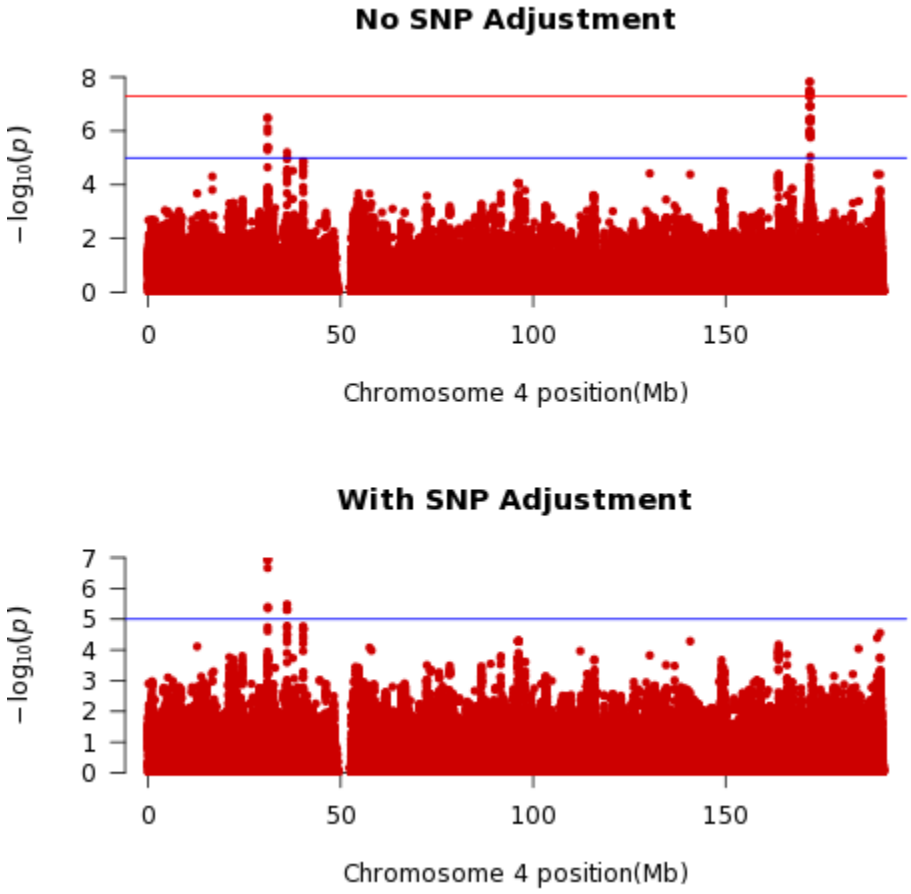


Figure 3.7 Regional Plot of SNPs in LD (r2 =0.2-1.0) with rs2195807, Chr4 (4q33)

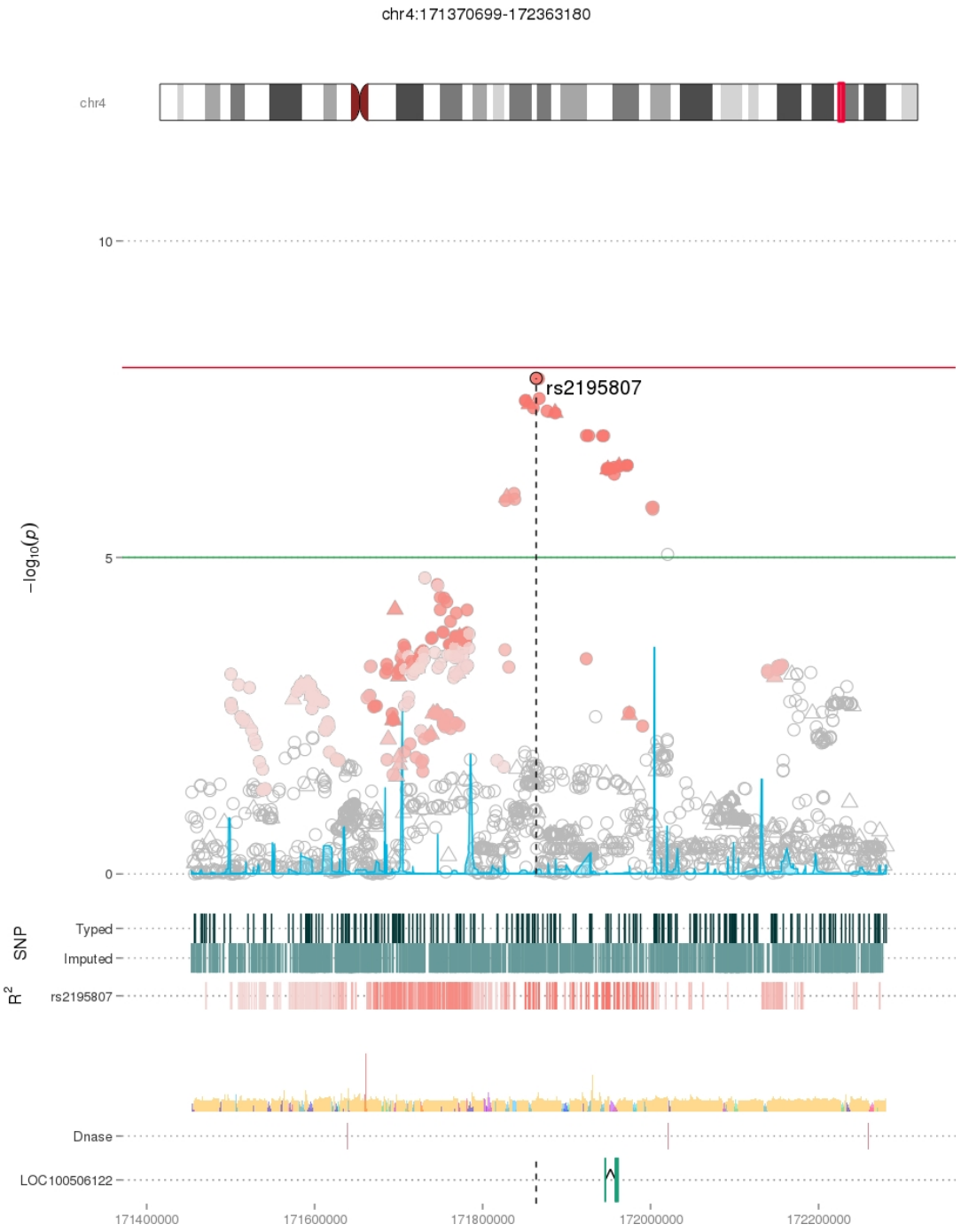


Table 3.9. Odds Ratios and P-Values for Top Ten SNPs: Chromosome 4, ~31Mb (4p15.1 & 4p14)

rsID	BP	A1	A2	MAF	OR	LL	UL	P	Typed/Imp.
rs35528558	31030698	T	C	0.210	0.47	0.33	0.66	3.20E-07	Imputed
rs35800293	31031051	A	G	0.210	0.47	0.34	0.66	3.39E-07	Imputed
rs61792945	31027150	A	G	0.209	0.49	0.35	0.68	7.62E-07	Imputed
rs35723210	31021145	C	T	0.208	0.49	0.35	0.69	1.09E-06	Imputed
rs6820873	31077545	A	C	0.470	1.55	1.24	1.94	4.02E-06	Typed
rs61792940	31009247	A	T	0.209	0.52	0.38	0.73	5.11E-06	Imputed
rs7670868	36085412	A	G	0.418	0.58	0.45	0.73	6.18E-06	Imputed
rs61797479	36082155	G	C	0.410	0.59	0.46	0.75	9.63E-06	Imputed
rs7688524	36079976	G	A	0.407	0.59	0.46	0.75	1.06E-05	Imputed
rs12651329	36079594	T	C	0.407	0.59	0.46	0.75	1.08E-05	Typed

Figure 3.8. Association Results Adjusting for rs35528558, Chromosome 4 (4p15.1)

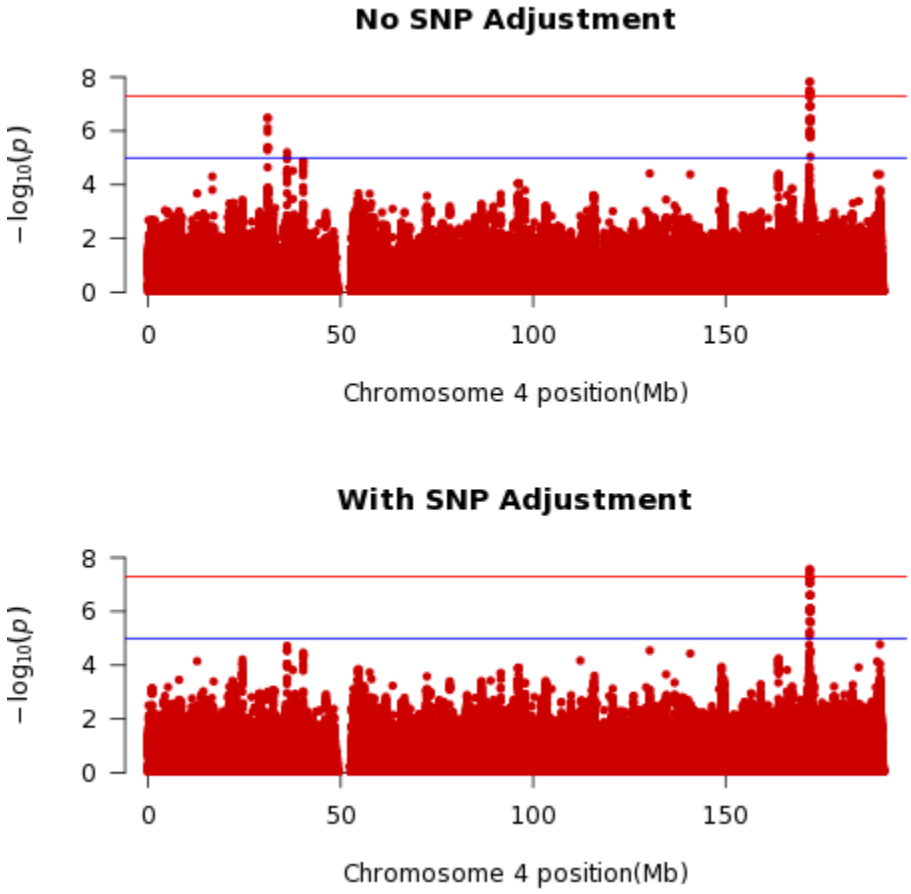


Figure 3.9. Regional Plot of SNPs in LD (r2 =0.2-1.0) with rs35528558, Chr4 (4p15.1)

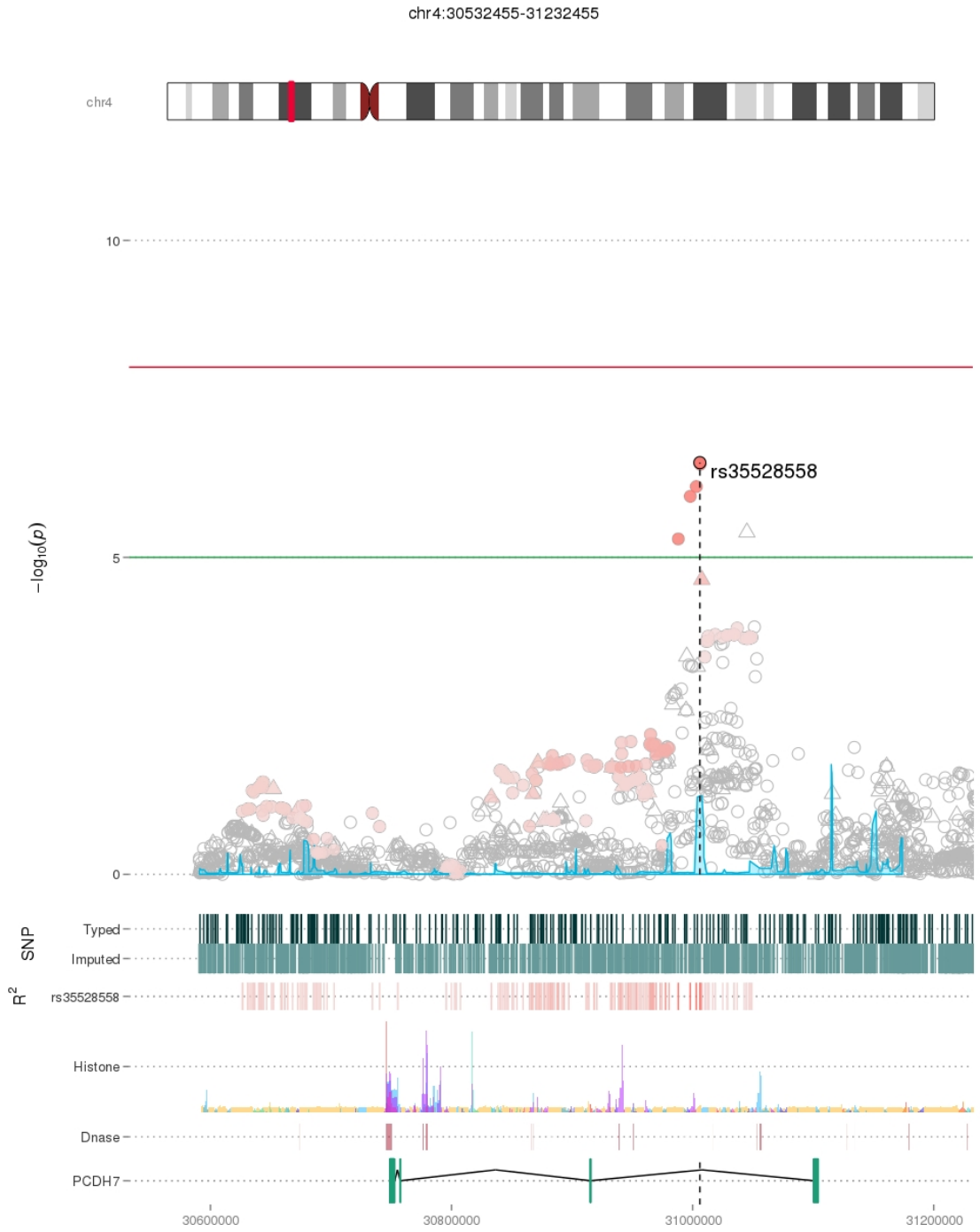


Table 3.10. Odds Ratios and P-Values for Top Ten SNPs on Chromosome 2 (2q36.1)

rsID	BP	A1	A2	MAF	OR	LL	UL	P	Typed/Imp.
rs17433868	222221224	T	C	0.253	0.510	0.377	0.689	5.32E-07	Imputed
rs11680028	222220798	A	G	0.253	0.511	0.377	0.691	5.56E-07	Imputed
rs67012780	222230197	C	T	0.348	0.514	0.395	0.668	7.39E-07	Imputed
rs11676423	222222003	T	C	0.347	0.515	0.396	0.669	9.68E-07	Imputed
rs10201690	222221003	C	T	0.346	0.518	0.398	0.674	9.95E-07	Imputed
rs17434888	222227393	T	C	0.347	0.514	0.396	0.669	1.00E-06	Imputed
rs10181647	222222356	A	C	0.347	0.515	0.396	0.670	1.00E-06	Imputed
rs17434868	222227216	C	T	0.347	0.515	0.396	0.669	1.01E-06	Imputed
rs10181883	222222534	A	G	0.347	0.515	0.396	0.670	1.01E-06	Imputed
rs10804297	222226640	T	C	0.347	0.515	0.396	0.670	1.03E-06	Typed

Figure 3.10. Association Results Adjusting for rs17433868, Chromosome 2 (2q36.1)

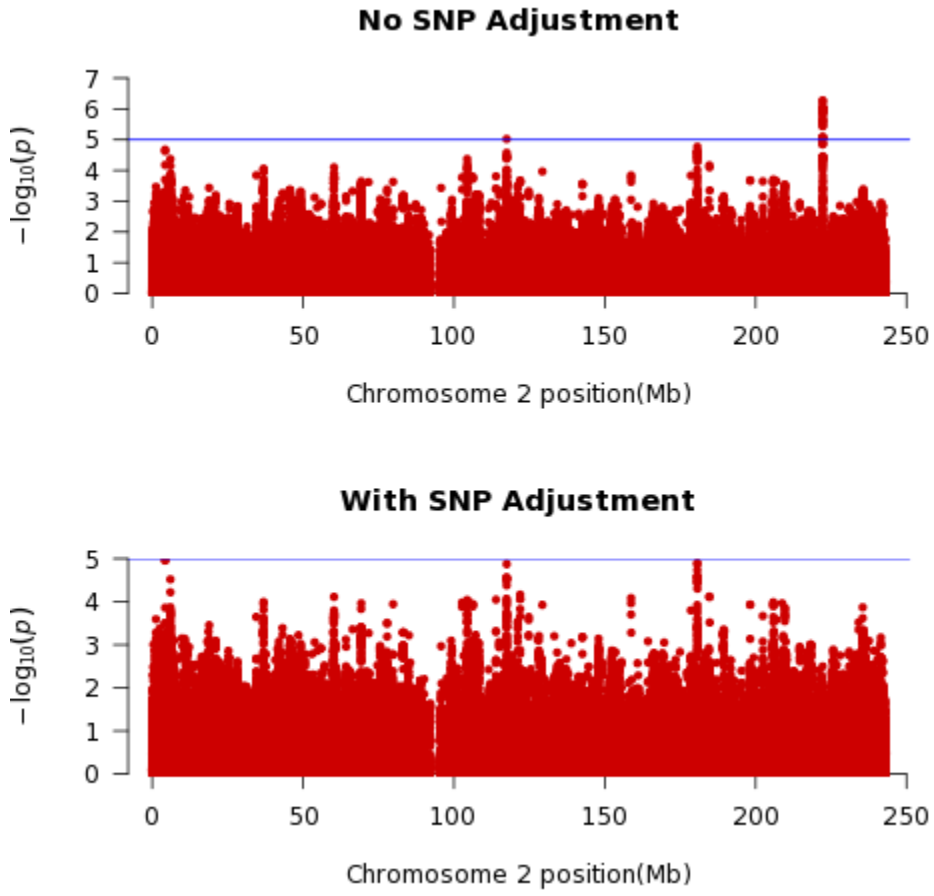


Figure 3.11. Regional Plot of SNPs in LD (r2 =0.2-1.0) with rs17433868, Chr2 (2q36.1)

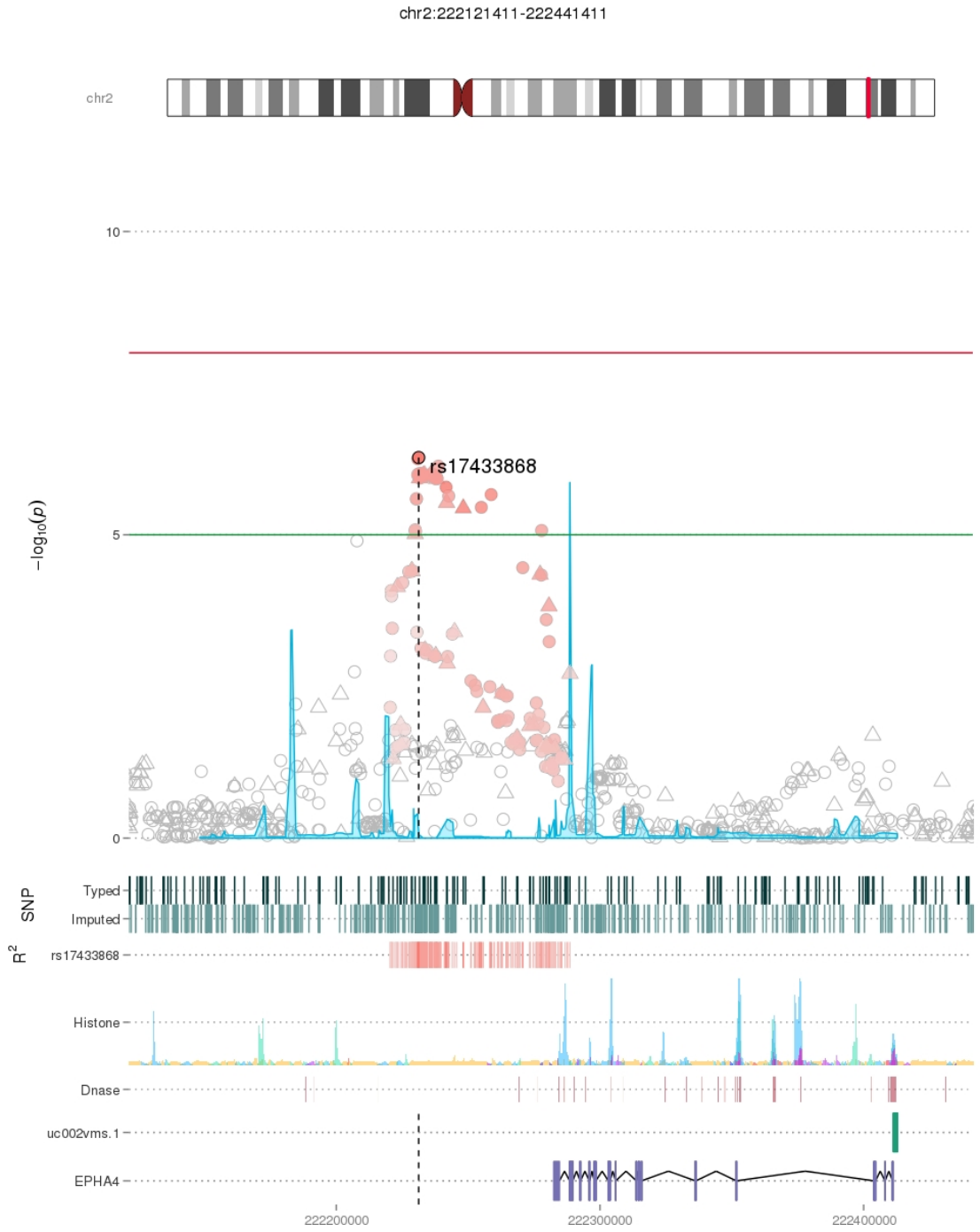


Table 3.11. Odds Ratios and P-Values for Top Ten SNPs on Chromosome 11 (11p15.3:4)

rsID	BP	A1	A2	MAF	OR	LL	UL	P	Typed/Imp.
rs56289978	11133374	G	A	0.065	2.24	1.58	3.17	2.31E-07	Imputed
rs56143914	11134564	A	G	0.065	2.23	1.57	3.16	2.60E-07	Imputed
rs77713994	11122587	G	C	0.066	2.20	1.55	3.12	3.22E-07	Imputed
rs74492376	11122577	C	A	0.066	2.20	1.55	3.12	3.24E-07	Imputed
rs78935380	11122379	A	G	0.066	2.20	1.55	3.12	3.35E-07	Imputed
rs61395681	11134584	G	A	0.066	2.19	1.54	3.10	3.63E-07	Imputed
rs58132943	11134766	G	A	0.066	2.19	1.54	3.10	3.67E-07	Imputed
rs74793062	11136115	A	G	0.066	2.18	1.54	3.10	5.22E-07	Imputed
rs80317637	11120614	C	G	0.068	2.13	1.50	3.01	1.08E-06	Imputed
rs73402352	9213436	T	C	0.119	1.82	1.36	2.43	3.11E-06	Imputed

Figure 3.12. Conditional Analysis Adjusting for rs56289978, Chromosome 11 (11p15.3)

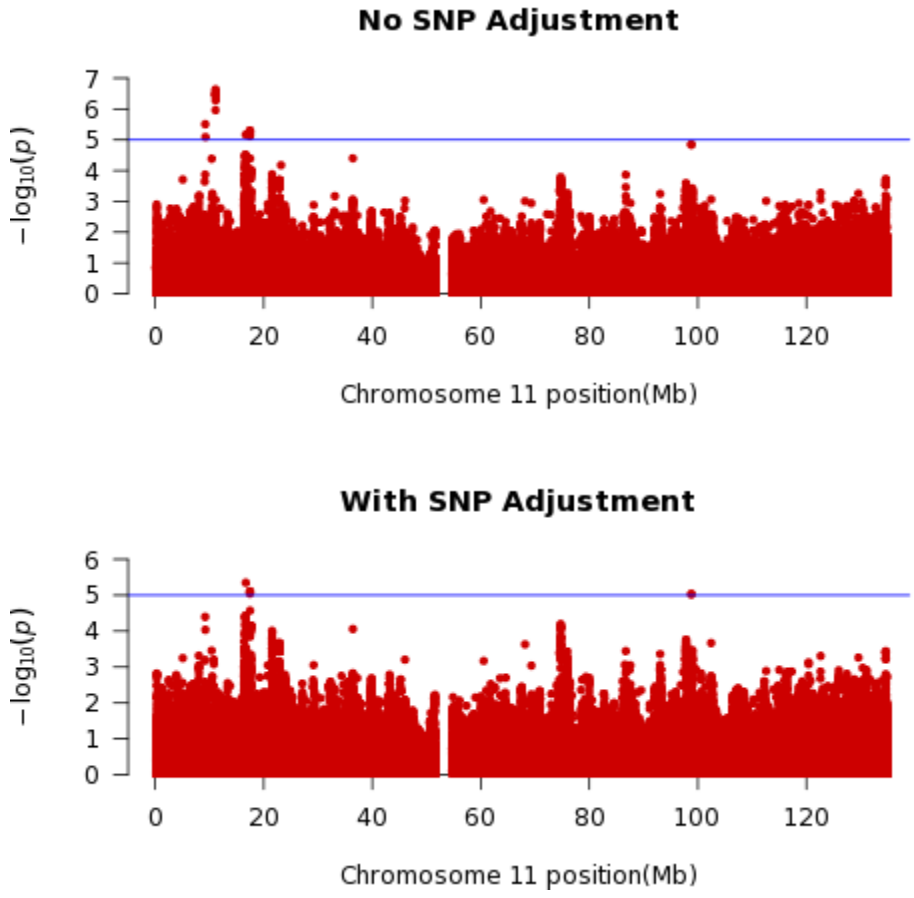


Figure 3.13. Regional Plot of SNPs in LD (r2 =0.2-1.0) with rs56289978, Chr11 (11p15.3)

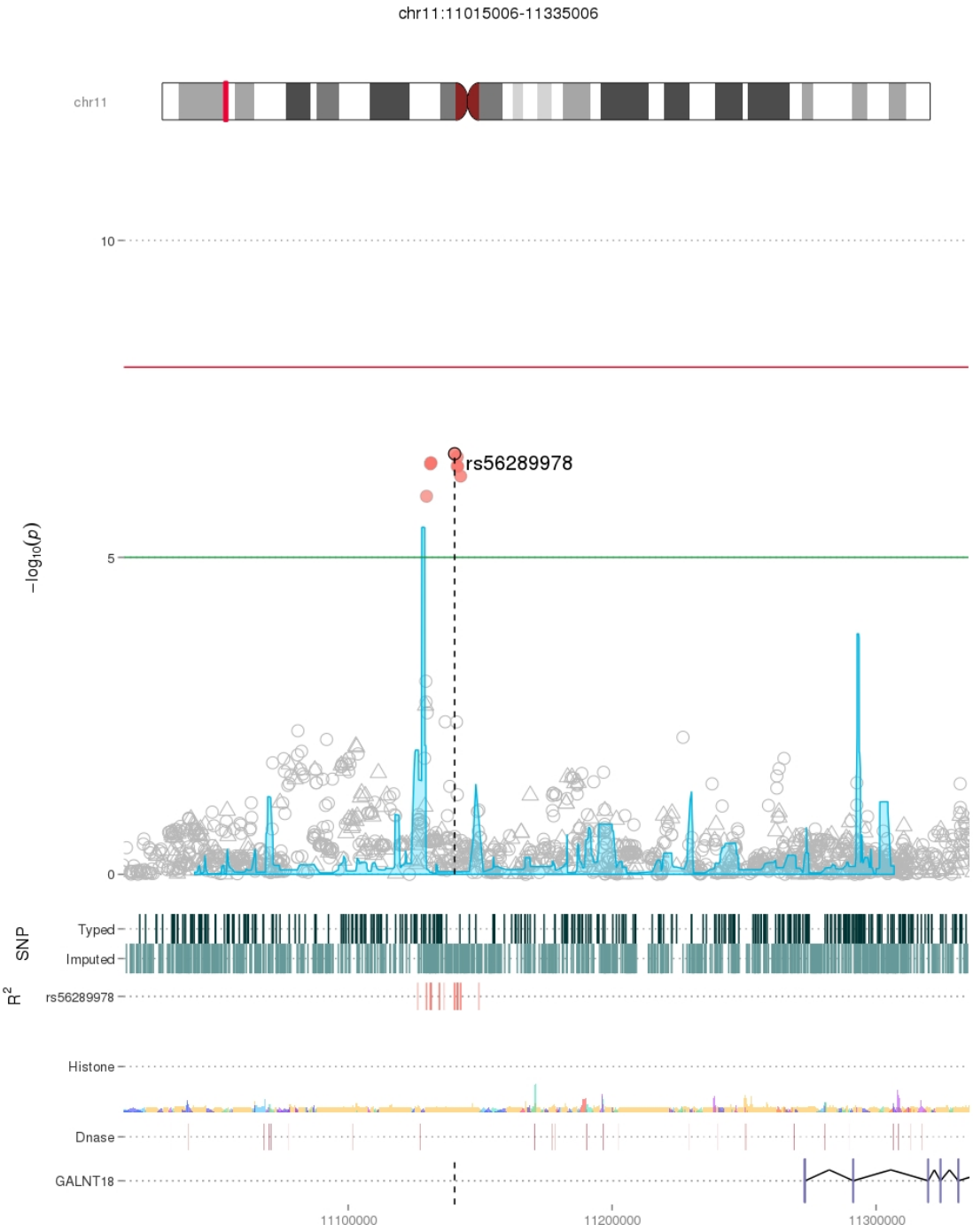


Table 3.12. Odds Ratios and P-Values for Top Ten SNPs on Chromosome 12 (12q13.13 & 12p13.33)

rsID	BP	A1	A2	MAF	OR	LL	UL	P	Typed/Imp.
rs11169939	52271467	C	T	0.103	2.002	1.485	2.701	1.02E-06	Imputed
rs76195628	52271279	T	C	0.103	2.000	1.483	2.698	1.03E-06	Imputed
rs12230130	52269931	C	T	0.103	1.989	1.474	2.685	1.08E-06	Imputed
rs12580654	52268547	G	C	0.103	1.973	1.461	2.665	1.20E-06	Imputed
rs11169944	52275199	T	C	0.142	1.696	1.286	2.237	2.42E-06	Imputed
rs11169942	52272699	G	A	0.101	1.966	1.452	2.662	6.19E-06	Imputed
rs11169943	52273248	C	T	0.101	1.968	1.454	2.665	6.20E-06	Imputed
rs11169945	52275509	T	C	0.097	2.022	1.489	2.745	8.27E-06	Typed
rs35103713	3077006	G	T	0.118	1.749	1.304	2.345	8.32E-06	Imputed
rs74626145	52271174	G	T	0.097	1.981	1.457	2.695	9.64E-06	Imputed

Figure 3.14. Conditional Analysis Adjusting for rs11169939, Chromosome 12 (12q13.13)

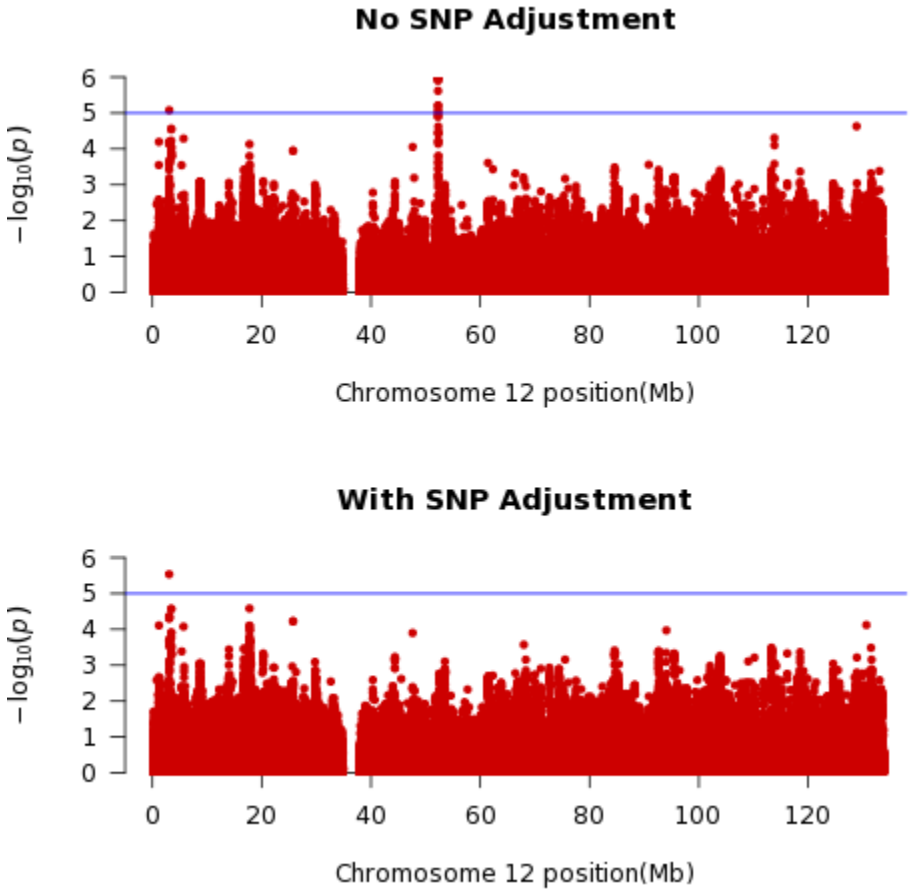


Figure 3.15. Regional Plot of SNPs in LD (r2 =0.2-1.0) with rs11169939, Chr12 (12q13.13)

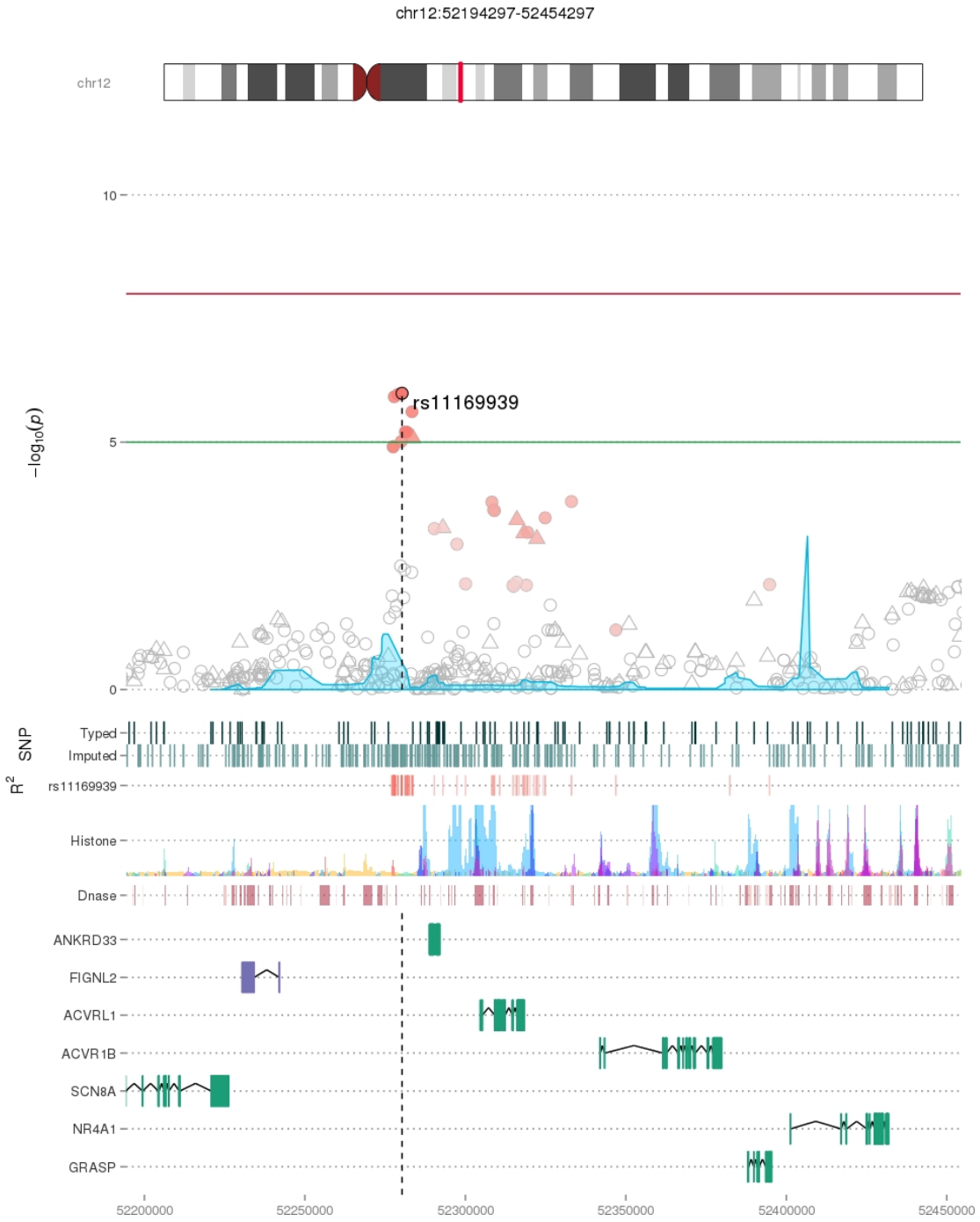


Table 3.13. Comparison of Results from Published GWAS Meta-Analyses and Current Study

	Band	Near Gene	rsID	BP	Risk	Other	Meta-OR	Meta-P	Study OR	Study P
VIJAI: MZL[64]	6p21.33	HLA-B	rs2922994	31335901	G	A	1.64 (1.39-1.92)	2.43E-09	-	-
	6p21.32	BTLN2	rs9461741	32370587	C	G	2.66 (2.08-3.39)	3.95 E-15	-	-
SKIBOLA: FL[68]	6p21.32	HLA Region	rs12195582	32444544	T	C	1.78 (1.69–1.88)	5.36E-100	-	-
	11q23.3	CXCR5	rs4938573	118741842	C	T	1.34 (1.26–1.43)	5.79 E-20	1.16 (0.87-1.55)	3.94E-01
	11q24.3	ETS1	rs4937362	128492739	T	C	1.19 (1.13–1.25)	6.76 E-11	1.05 (0.84-1.32)	9.76E-01
	3q28	LPP	rs6444305	188299902	G	A	1.21 (1.14–1.28)	1.10 E-10	1.11 (0.87-1.42)	6.33E-01
	18q21.33	BCL2	rs17749561	60783211	G	A	1.34 (1.22–1.47)	8.28 E-10	1.06 (0.71-1.57)	9.65E-01
	8q24.21	PVT1	rs13254990	129076451	T	C	1.18 (1.11–1.24)	1.06 E-08	1.00 (0.78-1.27)	6.16E-01
	17q25.3	C17orf62	rs3751913	80405552	C	T	1.23 (1.14–1.33)	2.24 E-07	-	-
	3q13.33	CD86	rs2681416	121817613	A	G	1.16 (1.09–1.22)	2.33 E-07	1.06 (0.83-1.34)	9.73E-01
	18q12.3	SCL14A2	rs11082438	42865210	G	T	1.33 (1.19–1.48)	4.01 E-07	0.73 (0.44-1.19)	2.88E-01
CERHAN: DLBCL[66]	6p25.3	EXOC2	rs116446171	484453	G	C	2.20 (1.87-2.59)	2.33E-21	-	-
	6p21.33	HLA-B	rs2523607	31322790	A	T	1.32 (1.21-1.44)	2.40E-10	-	-
	2p23.3	NCOA1	rs79480871	24694472	T	C	1.34 (1.21-1.49)	4.23E-08	1.06 (0.70-1.60)	5.18E-01
	8q24.21	PVT1	rs13255292	129076573	T	C	1.22 (1.15-1.29)	9.98E-13	0.97 (0.76-1.23)	4.38E-01
	8q24.21	PVT1	rs4733601	129269466	A	G	1.18 (1.11-1.25)	3.63E-11	0.99 (0.80-1.24)	4.36E-01
BERNDT: CLL[152]	10q23.31	ACTA2/FAS	rs4406737	90749704	G	A	1.27 (1.19-1.33)	1.22 E-14	1.15 (0.92-1.44)	4.10E-01
	18q21.33	BCL2	rs4987855	58944529	G	A	1.47 (1.32-1.61)	2.66 E-12	-	-
	18q21.33	BCL2	rs4987852	58944901	G	A	1.41 (1.27-1.56)	7.76 E-11	-	-
	11p15.5	C11orf21/TSPAN32	rs7944004	2267728	T	G	1.20 (1.13-1.27)	2.15 E-10	1.01 (0.81-1.26)	1.00E+00
	4q25	LEF1	rs898518c	109236273	A	C	1.20 (1.14-1.27)	4.24 E-10	1.02 (0.81-1.28)	8.10E-01
	2q33.1	CASP10/CASP8	rs3769825	201819625	T	C	1.19 (1.12-1.25)	2.50 E-09	-	-
	9p21.3	CDKN2B-AS1	rs1679013	22196987	C	T	1.19 (1.12-1.27)	1.27 E-08	0.90 (0.72-1.13)	2.20E-01
	18q21.32	PMAIP1	rs4368253	55773267	C	T	1.19 (1.12-1.27)	2.51 E-08	1.06 (0.83-1.34)	6.60E-01
	15q15.1	BMF	rs8024033d	38190949	C	G	1.22 (1.15-1.30)	2.71 E-10	0.89 (0.71-1.11)	3.10E-01
	2p22.2	QPCT/PRKD3	rs3770745d	37449593	T	C	1.24 (1.15-1.33)	1.68 E-08	-	-
	2q13	ACOXL/BCL2L11	rs13401811c	111332575	G	A	1.41 (1.30-1.52)	2.08 E-18	1.14 (0.87-1.51)	3.80E-01

	8q22.3	ODF1	rs2511714	103648050	G	T	1.19 (1.10-1.28)	4.72 E-06	-	-
	5p15.33	TERT	rs10069690	1332790	T	C	1.22 (1.13-1.32)	6.46 E-07	-	-

CHAPTER 4 : PATHWAY ANALYSES

4.1 Background: A Prominent Role for Pathway Analyses in Contemporary GWAS

Genome-wide association study pathway analysis (GWASPA) can be thought of as a multigenetic index[158] adapted to GWAS and informed by prior knowledge. GWASPA essentially works by 1) linking rsIDs to genes; 2) linking these genes to a pathway on the basis of shared function and scientific knowledge; and 3) testing the association between each pathway and AIDS-NHL. The approach was originally used for expression data, but has been adapted for use with SNPs in GWAS studies.

Before proceeding further, we should note that the term "pathway analysis" is generally used interchangeably to refer to one of two distinct types of analysis: a gene-set analysis, where genes are grouped into sets based on the strength of statistical associations with no necessary reference to shared function, and a pathway analysis, where genes are grouped into sets on the basis of shared function, using knowledge of the literature, laboratory data, and manual curation. Because many statistical considerations apply to both gene-set and functional pathway analyses, the term "pathway analysis" as used in this chapter will refer to both; explicit distinctions between the two will be made where relevant.

There are at least three major motivations for the use of GWASPA. The first is biological: biological networks have evolved to be redundant, and thus we find ourselves concerned primarily with complex rather than Mendelian diseases. In complex diseases, many genes with small individual effects operate in concert to generate a phenotype, rather than one or a handful of genes having large effects. These small individual effects may not be detected in isolation, but when condensed into a pathway, they may together generate a detectable effect. The second is psychological: it is difficult to make sense of the huge number of features in GWAS data;

distilling a large number of features into a smaller set of pathways makes engaging with the data more tractable. The third is statistical: with a smaller number of features, we reduce the number of simultaneous comparisons, thus reducing the stringency needed for multiple-comparisons corrections and improving power. (Also, we include SNPs regardless of their p-value—often generated by less stringent means—on the basis of background scientific knowledge.) Reproducibility may also be improved, since SNP-specific effects may vary across populations owing to genotyping, ethnic makeup, or sample size, but broader biological patterns may be less variable.

GWASPA has grown more widespread in recent years, motivated by the considerations above. During this time, a number of different pathway analysis approaches of evolving complexity have been developed, ranging from univariate analyses with Fisher's exact test to more complex multivariate and Bayesian approaches. Wojcik *et al.*[159] compared multiple pathway analysis approaches head-to-head; their results inform our own choice of methods, as described further below.

4.2 Gaps in the Literature

Given that SNPs in the pathways described in aim 1 have not been investigated using a candidate-gene approach, analysis of these SNPs has not been bolstered using pathway analyses. Given that GWAS of AIDS-NHL are underway but have yet to see publication, no pathway analysis of AIDS-NHL GWAS has been published.

4.3. Research Objectives

With pathway analyses, we shift the focus of our candidate gene study and increase the power of our GWAS, by changing the unit of analysis from single-nucleotide polymorphisms to genes and the pathways within which they operate.

4.4. Specific Aims and Hypotheses

Using a pathway-analysis approach, we aim to:

1) Use an agnostic genome-wide pathway analysis to assess variation in >13000 biological pathway and gene sets and association with AIDS-NHL.

2) Use a targeted genome-wide pathway analysis to assess variation in the candidate NHL-related pathways addressed in Aim 1.

We hypothesize the following:

- Hypothesis 1: Testing for pathway associations will increase our ability to identify meaningful associations between genetic variation and AIDS-NHL: assigning large numbers of SNPs to a smaller number of genes, and of genes to an even smaller number of pathways, reduces the multiple-comparisons penalty relative to a SNP-only approach and improves our ability to detect meaningful biological signals by reducing the chance of false negatives.
- Hypothesis 2: Even if the SNPs assessed in aim 1 are not found to be strongly associated with NHL risk, the genes on which they are found, and the pathways in which these genes are in turn found, will be associated with NHL risk.

4.5. Study Design and Methods

4.5.1. Study Overview

This aim uses summary statistics generated from the data collected in Aim 2, and uses the same 1,949 participants.

4.5.2. Software Selection

We use two programs for pathway analysis: VEGAS2[160,161] and PASCAL[162]. A tremendous range of pathway analysis programs is now available, and choosing from among these options can be challenging. Fortunately, Wojcik conducted a benchmarking analysis of several programs for both gene-level statistics and pathway-level statistics, and identified VEGAS2 as the top performer at the gene level (its pathway analysis function was not yet available), and MAGENTA as the top performer at the pathway level[159].

Motivated by this, we originally intended to use both programs. However, since the publication of Wojcik's review, the program PASCAL has become available; it uses a VEGAS2-inspired strategy to overcome computational difficulties with that program, and it also represents a major improvement over MAGENTA with regard to both performance and statistical sophistication. We therefore ran both VEGAS2 and PASCAL to calculate gene-level and pathway-level statistics; we discuss our rationale for this choice further in the next section, and the details of VEGAS2 and PASCAL methods in section 4.6 (Statistical Analysis).

4.5.2.1. Software Selection: Rationale for VEGAS2 and PASCAL

VEGAS2's pathway-level module extends the same methods used at the gene level. Though pathway-level performance has not been benchmarked, VEGAS2's rigorous statistical justification and strong performance at the gene level suggest that this performance will be duplicated at the pathway level. As a check on VEGAS2's performance, and to overcome two shortcomings with VEGAS2, we also use PACSAL.

One major shortcoming of VEGAS2 is that custom pathways cannot be specified, as it runs in a server environment with few options for customization. Since this analysis was begun, the MSigDB pathway database[163] has been updated (v5.1, January 2016), but VEGAS2 currently uses an outdated MSigDB v4.0 dataset. Rather than simply discard VEGAS2 analyses, we use

the strengths of PASCAL—flexibility, customizability, and especially the ability to specify one's own pathways—to overcome the major shortcomings of VEGAS2, in particular its use of outdated pathway data.

An additional weakness of VEGAS2 is the amount of time it takes to run. It uses Monte Carlo simulation to adjust for LD, gene size, and pathway length (described further below), which is very demanding computationally and leads to runtimes of roughly ten days for a single scenario. PASCAL's major innovation is to use one of two algorithms for LD adjustment, instead of the resampling approach described for VEGAS2[162]. (PASCAL also offers routines for more accurate estimation of p-values for SNPs at $p < 1E\text{-}15$, but this was not relevant to our case.) These algorithms cut the runtime of PASCAL up to 100-fold relative to VEGAS2; in our experience, PASCAL gene-based statistics generally ran in 5-10% the time of VEGAS2 for a dataset of ~5 million SNPs.

Importantly, both VEGAS2 and PASCAL correct for a major source of bias in pathway analysis: the length of pathways. Larger pathways have more SNPs and more genes than smaller pathways, and owing to either randomness or true biological signals are therefore more likely to be associated with a given outcome than smaller pathways[63]. This is the rationale for the standard cutoff of no fewer than 10, and no more than 200, genes in a given pathway for pathway analysis. However, this limitation can be overcome if properly accounted for. VEGAS2 and PASCAL do this by randomly sampling all genes in the study and summing them to create a pathway with length equivalent to that of the pathway of interest[160-162]. Because these genes are sampled randomly, there should be no meaningful signal, and they should also be independent, meeting a key assumption for p-value correction. The signal within the pathway of interest is thus benchmarked against the signal outside the pathway for a pathway of equivalent length,

under a null distribution. This is an example of a "competitive test", a key mark of quality in pathway analysis[63,164].

Mishra's simulation found a median Pearson correlation between gene size and gene-level p-value of -2.0E-03; a median Pearson correlation of 3.0E-03 between pathway length and pathway-level p-value, and a mean type I error rate of 0.049 across all pathways tested[161]. These very weak correlations and type I error rate below 0.05 further support the choice of VEGAS2 for pathway analysis.

4.5.2.2. Software Selection: Alternative Platforms

We considered other programs that accept summary data as input, including MAGENTA[165], Adaptive Rank Truncated Product Method (ARTP)[166-168] and PLINK[138,139]. ATPM had a high number of false-positives in Wojcik's comparison paper, and PLINK also showed poor performance[159]. Though MAGENTA performed well in Wojcik's analysis, it has three problems[165] that led us to choose alternative programs.

First is "thresholding", in which genes that fail to meet an arbitrary level of significance are dropped from the analysis before calculation of pathway scores. This can ignore important information, and the inflexible use of a fixed cutoff for dropping genes is undesirable. Second is its strategy for dealing with LD, which is as follows. Consider three genes lying in a 500MB region. First, because MAGENTA takes only the top SNP (i.e. the SNP with smallest p-value) from each gene, it calculates the top SNP for each gene. Second, for this 500MB region, it retains only the gene with the SNP of smallest p-value, discarding the other two genes from pathway-level analyses. This alleviates concerns with LD, but like thresholding, also risks discarding important functional information, and even discarding the causal gene or genes (recall the discussion of fine-mapping in Chapter 3: the strength of a statistic is not in itself an

unambiguous indication of causality). Third, the hypergeometric test used by MAGENTA can be underpowered: Lamperter[162] demonstrated better power for PASCAL vs. MAGENTA across several scenarios using binary enrichment. On both philosophical and technical grounds, we therefore opted not to use MAGENTA.

4.5.3. Pathway Selection: Broad Institute Molecular Signatures Database (MSigDB)

Pathway analysis software tests for associations between a given phenotype and a given pathway using pathways drawn from online repositories. Among others, these repositories include Gene Ontology[169], BioCarta, and the Broad Institute's Molecular Signatures Database (MSigDB)[163]. The choice of pathways has a major impact on results, and they should be chosen using three principles[63,164]. The first is that there should be a broad range of pathways from multiple evidence sources: the quality of evidence can vary by source, and it is wise to corroborate findings from one source with results in another. The second is that pathways should capture processes relevant to the phenotype of interest. The third is that the investigators should be able to evaluate the quality of evidence for a given pathway, including both the original evidence used to construct pathways and whether these pathways are updated regularly in light of any new evidence.

Based on these considerations, we chose the Molecular Signatures Database (MSigDB) Version 5.1, updated in January 2016. This repository offers more than 13,000 pathways, updated regularly, and pathways can be downloaded either as a concatenated set or as subsets constructed using different standards of evidence. The evidence source for every pathway is specified; citations and links to relevant data are provided, including in many cases the original data from experiments used in pathway construction.

The tremendous collection of pathways in MSigDB allows for investigation of an equally wide

range of scientific questions, and enables investigators to make use of the full scope of evidence available to them. This evidence can range from computational predictions based on machine learning, to gene knockout experiments in mice, to well-characterized human phenotypes supported by hundreds of citations and annotated by multiple scientific efforts. We make full use of MSigDB resources by drawing on seven different pathway collections. Three collections capture well-characterized phenomena: 1) "hallmarks" pathways[170]; 2) manually curated pathways; 3) and Gene Ontology pathways[169]. To investigate in more detail the immunological and gene regulatory aspects of NHL, we also use four specialized pathway collections: motifs (miRNA/TF binding)[171], immunologic signatures[172], computational cancer gene sets[173], and cancer signatures[163,174]. The following sections provide a brief overview of these pathway collections.

The strength of evidence for a given pathway—including both the number of sources used for annotation, the nature of the data in these sources—should always be of concern. This evidence may also be influenced in part by publication bias: that is, a wealth of evidence for a given pathway may be a result of investigators' interest and the availability of funding for research on a given set of genes, rather than the true strength of biological action for those genes relative to all others. This is a more difficult problem to correct. The best way forward is twofold: first, to present clearly the evidence base—with citations and links to original publications—for a given pathway; and second, to remind the reader that the absence of evidence for a given pathway is not "evidence of absence" of biological relevance[175].

4.5.3.1. Pathway Selection: Hallmark Pathways (50 sets)

Hallmarks (50 sets) are a special set of curated datasets for canonical pathways that are then restricted on the basis of overlap between gene sets[170]. These represent the best-characterized

biological phenomena in MSigDB and are considered to have the strongest evidence, both because the phenomena considered are well-characterized, and because of genes listed in these pathways are restricted to sets common to multiple sources in the literature, rather than just one or a few publications.

4.5.3.2. Pathway Selection: Curated Sets (4726 sets)

Curated pathways are drawn from three sources: 1) online databases such as REACTOME and BioCarta; 2) Broad Institute extraction and manipulation of gene sets from microarray articles in the literature; and 3) L2L[176] and the Myc Target Database[177].

On the basis of this, curated pathways should generally be considered to have a strong evidence base relative to databases relying primarily on computational predictions. However, at the level of individual pathways, the strength of evidence may differ according to experimental conditions and other factors. This reinforces the point that even if drawing pathways from a compendium of generally high-quality data, it is important to examine the evidence base for individual pathways of interest when interpreting associations.

4.5.3.2.1. Pathway Selection: Curated Sets: BioCarta (0/217 available sets used)

As of March 2016, BioCarta pathways were no longer being updated. It is important to revise prior conclusions in pathway structure on the basis of new evidence; use of old pathway definitions is not recommended[63]. Therefore, though BioCarta pathways are still available at MSigDB, we do not use them in PASCAL. However, VEGAS2 uses them by default, and there is no option to modify pathway choices in VEGAS2, so we report BioCarta results in VEGAS2 data.

4.5.3.2.2. Pathway Selection: Curated Sets: KEGG (186 sets)

KEGG is perhaps the best-known repository for curated pathways[178]. It is based at the University of Tokyo and hosts a number of pathways that capture processes ranging from subcellular mechanisms to specific phenotypes, including multiple types of cancer.

4.5.3.2.3. Pathway Selection: Curated Sets: Matrisome

The Matrisome is a catalog of genes active in extracellular matrix and related processes, including glycoproteins, collagens, proteoglycans, affiliated genes, ECM regulators, and secreted factors[179,180]. Associations are generated using bioinformatic prediction from both *in silico* and experimental data.

4.5.3.2.4. Pathway Selection: Curated Sets: Pathway Interaction Database (PID)

PID data capture signaling and cellular processes from the literature using manual curation and peer-review, representing a good standard of evidence with monthly updates[181]. As of February 2016, PID has been retired and moved to NDEX, the Network Data Exchange; data remain available[182]. This appears to have been done to streamline coordination between different working groups in academia and industry by hosting all data at a single repository, rather than reflecting concerns with the data themselves.

4.5.3.2.5. Pathway Selection: Curated Sets: REACTOME (674 sets)

The REACTOME Pathway Knowledgebase uses mined text and active curation of published studies to generate its list of pathways[183]. First put online in 2005, it is updated frequently and represents a useful source of curated pathways, all available in MSigDB[184].

4.5.3.2.6. Pathway Selection: Curated Sets: SigmaAldrich

SigmaAldrich is a well-known commercial provider of life sciences material, and makes available a wide set of expression and pathway data.

4.5.3.2.7. Pathway Selection: Curated Sets: UCSD Signaling Gateway

The UCSD Signaling Gateway (SG) focuses on signaling mechanisms and pathways. It is a case in point regarding our discussion of the strength of evidence for a given pathway varying widely, even within a single dataset: the SG includes a range of pathways derived from computational prediction, expression data, and manual curation, and includes comprehensive expert reviews for selected pathways of import. However, its focus is limited to specific processes.

4.5.3.2.8. Pathway Selection: Curated Sets: Signal Transduction KE

Science's Signal Transduction KE (http://stke.sciencemag.org/) makes available a range of tools, publications, and data for the study of signal transduction. It served as a source for the construction of curated pathways in MSigDB.

4.5.3.2.9. Pathway Selection: Curated Sets: SuperArray

As of March 2016, the SuperArray portal is nonfunctional, and the provenance of SuperArray data is unclear. Pathways derived from this repository will therefore not be reported in the Results section.

4.5.3.3. Pathway Selection: Gene Ontology (1454 sets)

Gene Ontology (GO) pathways are pathways of genes that share the same annotation. It should be noted that our goal in examining GO is to assess the association of genes with one of three phenomena: 1) cellular component; 2) biological process; or 3) molecular function. GO notes explicitly that investigators should not equate these three categories with "pathways" in a broad sense; rather, 2) refers specifically to phenomena such as signal transduction, and is defined by GO as "a series of events accomplished by one or more organized assemblies of molecular functions." In turn, "molecular function" denotes "activities that occur at the molecular level."

Each of these can be important for advancing etiology, but terminologically, they are not necessarily biological pathways in the sense of KEGG pathways. Importantly, these are organized hierarchically, so extensive overlap will be observed.

GO annotation uses six types of experimental evidence (EXP, IDA, IPI, IMP, IGI, IEP), ten types of computational evidence (ISS, ISO, ISA, ISM, IGC, IBA, IBD, IKR, IRD, RCA), and specification of whether references used are traceable. None of these should be taken to indicate the quality of evidence, though[185]. For instance, evidence from an experiment would generally be viewed positively, but this should also hinge on the quality of the experiment itself. GO has no such metric, and as such, one cannot say unequivocally that any broad category of evidence is better than the other in GO. Furthermore, there is no standardized metric of quality or confidence across annotation databases. The Confidence Information Ontology (CIO) is working toward this goal and is in discussions with the Gene Ontology Consortium to implement this, but this has yet to come to fruition[186]. Gaudet and Dessimoz provide a fuller discussion of biases associated with particular GO pathway databases[175].

4.5.3.4. Pathway Selection: Oncogenic (189 sets) and Immunologic (4872 sets) Signatures

The evidence base for these pathways is distinct from that for others considered so far: the majority of cancer signatures pathways were generated using data from either Gene Expression Omnibus (GEO)[187,188] or unpublished studies targeting known cancer genes shared by investigators, while only a minority were based on manual curation.

Immunologic signatures data were derived from manual curation of published literature as part of the Human Immunology Project Consortium[172]. They represent 4728/2 = 2,364 pairs of gene sets defined as follows. The phenotype of interest (which may be a molecular process) was identified. The first member of the pair represents genes upregulated in the presence of the

phenotype, and the second member represents genes downregulated in the presence of the phenotype (or vice versa). Genes chosen for this set are either differentially expressed at an FDR significance level of <0.35, or are the top 200 (in the case of upregulation) or bottom 200 (in the case of downregulation) differentially expressed genes.

4.5.3.5. Pathway Selection: Computational Cancer Datasets (858 sets)

We use two datasets generated exclusively through the analysis of cancer-related expression data. First is a set of 427 cancer gene "neighborhoods" assessing expression patterns in the neighborhood of 380 curated genes identified by Bretani *et al.*[174] to be associated with cancer. These were compiled as described in Subramanian *et al.* 2005[163] from four compendia: the Broad Institute Global Cancer Map (GCM); the Novartis Human Tissue Compendium and the Novartis Carcinoma Compendium[189]; and NCI-60 cell lines from the National Cancer Institute. GCM data include 218 tumor samples from 14 types of tumor, and 90 tissue samples from controls; Novartis data include data from 79 human and 61 mouse tissues; NCI-60 data are from the August 2010 release of the NCI Developmental Therapeutics Program's Molecular Targets Database[190,191].

Second is a set of 431 modules generated by Segal and colleagues[173]. Modules are defined by the authors as "sets of genes that act in concert to perform a specific function," and therefore represent pathways defined by common function using expression data. Segal *et al.* drew expression data from 1,975 arrays across 26 studies, fitting these to 2,849 gene sets. They identified arrays in which genes in these sets are differentially expressed, and concatenated these sets into "modules" based on gene overlap between them. They then assessed the degree of enrichment for each module in a given phenotype. Modules are then deemed "activated" or "inactivated" in a particular type of cancer, based on their patterns of expression in tissue/tumor

samples of specific type. For instance, module 1 would be defined by identifying gene sets, then identifying arrays in which members of these sets are differentially expressed, and then combining gene sets to form a module. One drawback of these modules is that they have not been updated in some time; they are taken from the original Segal *et al.* 2004 publication.

4.5.3.6. Pathway Selection: Transcription Factor and miRNA Binding Motifs (836 sets)

We also analyze 836 gene sets featuring microRNA (221 sets) and transcription factor (615 sets) target motifs. Micro-RNA motifs were identified by Xie[171] and represent regulatory elements in 3'UTRs and promoters that are conserved across human, rat, mouse and dog genomes; genes in these sets therefore share a 3'UTR miRNA binding element[192]. Transcription factor motif sets include sets that share TRANSFAC-defined TF binding sites.

4.5.3.7. Pathway Selection: Aim 1 Pathways

Finally, we sought to examine the association between AIDS-NHL and pathways containing our 24 genes of interest in aim 1. MSigDB[193] was used to identify pathways with quality evidence in which these genes were well-represented. To ensure quality of evidence, we restricted our search to hallmark pathways, curated pathways (excluding BioCarta pathways because of a lack of regular updates), and Gene Ontology pathways. We then identified pathways with 10-200 genes in which ≥5% of the genes in any pathway were also found in our list of 24 genes from aim 1, and required that this overlap be significant at a false-discovery-rate-corrected $q < 0.05$. This yielded 14 pathways, shown in Table 4.1.

4.6. Statistical Analysis

Methods for VEGAS2 and PASCAL analyses are described separately in sections 4.6.1 and 4.6.2. Each program was run using three different boundaries for assigning SNPs to a gene: a 1)

a 0kb scenario, in which only SNPs lying directly on a gene were mapped to that gene; 2) a 20kb scenario, in which SNPs lying within 20 kb of a gene's 3' or 5' UTR were also mapped to the gene; and 3) a 50kb scenario, in which SNPs lying within 20 kb of a gene's 3' or 5' UTR were also mapped to the gene. These are referred to as the "0kb", "20kb", and "50kb" scenarios throughout the text.

VEGAS2 was run on all 6,212 pathways (0kb, 20kb and 50kb scenarios) from the MSigDB v4.0 release included on the VEGAS2 server, comprising BIOCARTA, Gene Ontology, Reactome, Protein Interaction Database, and PANTHER sets.

PASCAL was run on 13,094 pathways drawn from seven collections in MSigDB v.5.1 (January 2016 update). One run was conducted on the full set of all 13,094 pathways (0kb, 20kb, and 50kb scenarios). Additional runs were conducted on: 1) hallmark pathways; 2) curated pathways; 3) oncogenic signatures; 4) immunologic signatures; 5) computational prediction data; 6) miRNA and transcription factor binding motif data; and 7) a set of 14 pathways capturing genes examined in Chapter 1 (our candidate gene study).

Q-Q plots for each program were run in 0kb, 20kb, and 50kb scenarios to examine the distribution of observed vs. expected p-values. Results from each program were analyzed and tables were generated using standard Linux command-line tools (*grep/awk/sed*) and the R/Bioconductor[194] packages *dplyr*[195], *UPSETR*[196], *qqman*[141], and *mygene*[197].

4.6.1. Statistical Analysis: VEGAS2 Analyses

VEGAS2 was run in three separate 0kb, 20kb, and 50kb scenarios, using the top 10% of SNPs on each gene, with a 2000-SNP limit on the number of SNPs assigned to a gene and a 500-kb boundary for gene clumping.

P-values for the ~5 million SNPs analyzed in chapter 3 rsIDs and p-values (here, generated by the SNPTEST analysis described in chapter 3) were input to the VEGAS2 server[198]. Running in R and PERL, VEGAS2 then converts these p-values to a chi-squared statistic (one-sided P-values are equal to the area under a chi-square distribution to the right of an observed test statistic). Using hg19 data, SNPs are assigned to genes, and these SNP-level chi-squared statistics are summed to generate a gene-level chi-squared statistic[160].

To correct for LD, Monte Carlo simulation is used: resampling from a distribution with mean 0 and variance equal to an LD matrix calculated from 1000 Genomes data is carried out, and the observed test statistics are compared with the simulated statistics. This generates an empirical p-value, which equals the proportion of resamplings in which the test statistic was as or more extreme than that originally calculated from the input data. Because small p-values translate into large chi-square statistics, the sum of these statistics will also be large; it will be rare to see values exceeding this sum, and the empirical p will therefore be small—a low proportion of simulations will yield extreme values.

Pathway analysis proceeds in three stages. First, SNPs are assigned to genes, and gene-level chi-squared statistics are calculated as described in the paragraph above. Second, genes are assigned to pathways via reference to MSigDB and pathway-level chi-square values are calculated. If genes are a) in a common pathway and b) <500kb away from one another, then these genes are clustered, and the gene-level analysis is re-run on this cluster. If not, then the gene-level analysis described in the preceding paragraph is simply run as is. Third, test statistics for genes and gene clusters in the pathway of interest are then summed, and simulation is used to calculate an empirical p-value of the pathway-level association, generally through 10^6 resamplings. As with gene-level resampling, the empirical p-value is then the proportion of resamplings in which the

test statistic was as or more extreme than that observed. This clustering may seem odd, but it is preferable to other alternatives, which include simply dropping nearby SNPs, as several other programs do (e.g. MAGENTA, described earlier). By keeping these genes and then clustering them, VEGAS2 (and as we shall see, PASCAL) maximizes available information.

Wojcik[159] reported 100% specificity, 20.41% sensitivity, and 0.16% false-positive significant results using VEGAS2 routines mapping 100% of SNPs in the input data to genes. An alternative method is to rank SNPs in increasing order of magnitude of association with NHL, and map only the top 10% of these SNPS (i.e. those with the smallest p-values) to genes; this greatly reduces the computational burden of VEGAS2, which takes up to ten days for a single run. The 10% mapping strategy performs well relative to the 100% strategy: Wojcik found 98% specificity, 28.57% sensitivity, a 0.40% false-positive rate, and a 71.43% false-negative rate. This is a worthwhile tradeoff given the computational intensity of VEGAS2; therefore all VEGAS2 runs were conducted using the top 10% of SNPs. As a sensitivity analysis, we calculated gene-level statistics with the 100% strategy, and report these alongside VEGAS2 10% and PASCAL statistics in table 4.11.

4.6.2. Statistical Analysis: PASCAL Analyses

PASCAL was run using command-line input in a Linux Red Hat environment. Three separate 0kb, 20kb, and 50kb scenarios were run, using 100% of SNPs on each gene and the sum of chi-squares strategy (SOCS), where the sum—rather than the maximum—of chi-square statistics for SNPs on a gene is used to generate the score for that gene. A 3000-SNP limit on the number of SNPs assigned to a gene and a 1MB boundary for gene clumping were used. PASCAL used either the Davies or, in the case of Davies failure, the Farebrother algorithm to generate gene-level scores. Monte Carlo simulation was used to generate empirical p-values for pathway

scores, using a minimum of 10^4 and a maximum of 10^6 simulations drawing from a distribution with mean=0 and variance equal to an LD matrix calculated from 1000 Genomes EUR data. Gene definitions were based on a b37 catalog of known genes from UCSC.

PASCAL improves on standard VEGAS2 analysis, and also on the top-performing methods in Wojcik's review. PASCAL's major innovation is to avoid the Monte Carlo simulation that VEGAS2 uses to correct for LD at the gene level, and for pathway size at the pathway level[162]. Like VEGAS, it transforms SNP-level p-values from GWAS data to chi-square scores, but then uses one of two algorithms (Davies or Farebrother) at the gene level to adjust for LD, rather than Monte Carlo simulation as performed in VEGAS2. Like VEGAS2, it also clumps together genes lying close to one another (1MB default) if they are in the same pathway to correct further for LD and functional correlation, and then calculates a gene score for this clump that is then input into pathway score calculation.

At the pathway level, PASCAL uses Monte Carlo simulation, as does VEGAS2. Empirical p-values are calculated as described for VEGAS2, and results reported here are all empirical p-values.

As a further check on PASCAL results, sensitivity analyses examining the impact of gene-fusion were run. Q-Q plots for genes, fusion genes, and pathways were created; pathway-level Q-Q plots for scenarios run using fusion genes were compared against plots for scenarios not using the fusion-gene approach.

4.6.3. Statistical Analysis: Multiple Comparisons and P-Values

There is no consensus on an appropriate threshold for "significant" p-values in pathway analysis. Many reports do not in fact apply corrections, and neither do most standard software routines for pathway analysis. Furthermore, there can be extensive overlap between genes in different

pathways, meaning that pathways do not meet the assumption of independence that is needed for such correction methods. However, Mishra accounted for gene overlap between pathways and estimated a Bonferroni-corrected p-value of 0.05/4597 = 1.09*10E-5 in simulations of the VEGAS2 pathway approach, using the same 6128 pathways that we test here in VEGAS2[161]. Both FDR and Bonferroni corrections were made to VEGAS2 and PASCAL p-values using the base R routine p.adjust, which allows for a choice of Bonferroni or various implementations of FDR; none were significant after correction[199].

We should also ask what the purpose of a p-value is in this case. Generally, the p-value is used (rightly or wrongly) to inform some sort of decision-making. The purpose of this pathway analysis is to help make sense of a huge amount of information from the GWAS by illustrating broad patterns, rather than to inform clinical decision-making, allocation of population-health resources, or study funding (including functional studies following up on promising SNPs identified via a GWAS).

Consequently the qualitative pattern of results across pathway approaches—i.e. the rankings—can be more informative than the size of the p-value attached to any particular pathway[159]. Rather than focus on the size of p-values, a focus on the similarity or dissimilarity of rankings for a specific pathway is recommended, and results are reported accordingly. We emphasize that these results will require replication: this is a small study using just 1,949 participants, of whom 172 are cases.

4.7. Results

We first present results for 6,212 pathways analyzed in VEGAS2 and 13,094 pathways analyzed in PASCAL, and compare the two sets of results. Next we present subgroup-specific analyses for pathways found only in PASCAL. As discussed, PASCAL allows for greater flexibility in both

the choice of pathways and in model parameterization. Though we recognize that there is value in using a uniform set of pathways across multiple platforms, and that a lack of comparison data from VEGAS2 may make interpretation of these subgroup-specific results more challenging, restricting our PASCAL analyses to only those pathways in VEGAS2 would have ignored important information and hampered our ability to explore gene expression signatures and miRNA and transcription-factor binding motifs of potential relevance to NHL biology. We end the results section by assessing the performance of 14 pathways containing genes targeted by our candidate-gene study in Chapter 2. We conclude with overarching thoughts on biological conclusions and the merits of PASCAL and VEGAS2 relative to one another.

4.7.1. VEGAS2 Results: 6,212 Concatenated Pathways

6,212 pathways were analyzed using VEGAS2. Curated pathways included 649 REACTOME pathways and 196 Protein Interaction Database (PID) Pathways. VEGAS2 also featured 3,748 Gene Ontology Pathways. The remaining 1,619 pathways were drawn from PANTHER[200], not found in MSigDB.

Figure 4.1 shows the distribution of expected p-values in VEGAS2 at both the gene level and the pathway level for 0, 20, and 50KB scenarios. Recall that the calculation of gene-level statistics is an important step along the way from SNPs to pathways; it is therefore important to examine both gene-level and pathway-level statistics. Q-Q plots of the former can explain potential aberrations in the latter, and pathway-level statistics can also indicate whether any problems with gene-level statistics (e.g. false-positive inflation) have been properly corrected by pathway methods.

The fit of observed p-values to the expected distribution is good across all three scenarios, at both the gene and pathway levels. At the gene level, the best fit was observed for the 0kb

definition, and we do not see the observed distribution fall short of the expected distribution, which would suggest an overly-conservative approach to defining gene boundaries. At the pathway level, we see the best fit for the 20KB definition: the 20KB definition slightly undershoots the expected distribution at smaller p-values, but the 0 and 50 KB definitions overshoot the expected distribution at larger p-values.

As we are focused on the top-performing pathways, and thus on the upper-right corner of these plots, we can be confident that our VEGAS2 results do not suffer from an excess of false positives, and if anything are fairly conservative. (Lack of departure from the expected distribution above the diagonal can also indicate the true absence of any meaningful signal, which is always a possibility).

Table 4.2 gives concatenated results for 52 VEGAS2 pathways, representing the top 25 pathways observed in the 0kb, 20kb, and 50kb scenarios. No pathways attained the conservative Bonferroni-corrected threshold of 1.09*10E-05 calculated by Mishra[161]. Seven of these pathways were in the top 25 for all three scenarios: heterotrimeric G-protein complex (Gene Ontology GO:0005834), positive regulation of translation (GO:0045727), GTP-dependent protein binding (GO:0030742), retrograde transport endosome to Golgi (GO:0042147), PID (Protein Interaction Database) ARF6 PATHWAY, GTPase activator activity (GO:0005096), and PANTHER BIOLOGICAL PROCESS Homeostasis.

As noted above, pathway performance may be driven by just one or a few genes, so it is important to examine the degree of overlap between different pathways. There is limited overlap between these pathways. GNA15 and ARBB1 are shared by the heterotrimeric G-protein complex and the PID ARF6 pathways; RGS6 and RGS9 are shared by the GTPase Activator Activity pathway and heterotrimeric g-protein complex; ARF6 and GTPase_activator_activity

share ACAP2, ARAP2, ADAP1, ACAP1, and GIT1.

Figure 4.2 assesses this overlap graphically. Since our concern is primarily with performance within the current study, we report only those genes that were included in the final pathway analyses following VEGAS2's LD-based exclusions. The difference in length, i.e. the number of genes for which statistics 1) were calculated successfully, 2) which passed LD adjustment, and 3) were therefore included in a pathway, between the final set and the initial set is not extreme for pathways of length <100; only for longer pathways do >10% of genes tend to be dropped. Full data on each pathway can be accessed by entering the pathway name at MSigDB, listed in the references.

4.7.2. PASCAL Results: 13,094 Concatenated Pathways

Using PASCAL, we examined a total of 13,094 pathways. Runs were conducted both separately by evidence type and in aggregate; results for aggregate runs are given below, and runs stratified by evidence type are given later.

Preliminary examination of Q-Q plots, shown in Figure 4.3, indicated that the distribution of observed p-values in pathway results was falling short of the expected distribution for all three scenarios, but especially for the 50KB scenario. This was not seen in Q-Q plots of gene-level statistics, where the fit was generally better. This observation suggested that PASCAL's fusion-gene strategy, which corrects for functional correlation between closely-spaced genes by collapsing genes within 1MB of one another if they are within the same pathway[162], might account for the observed distribution of pathway-level p-values. To verify this, two steps were taken. First, p-values for fusion genes were plotted (Fig. 4.3). Next, analyses were re-run on all 13,094 pathways omitting the gene fusion step (Fig. 4.4). This resulted in a marked inflation of the distribution of observed p-values, indicating that the gene-fusion step is indeed necessary for

proper analysis with PASCAL. We therefore retain our analyses using 1MB fusion, but note that they are conservative, and may reject some true positive results.

Table 4.3 shows concatenated results for 51 PASCAL pathways, representing the top 25 pathways observed in the 0kb, 20kb, and 50kb scenarios. Six pathways were in the top 25 in all three scenarios; two were immunologic signatures, two were Gene Ontology pathways, and two were from the collection of curated pathways. IGLESIAS_E2F_TARGETS_UP is a curated pathway that includes 151 cell-growth control and ductal cell and adipocyte differentiation genes upregulated in the pancreatic cells of mice with E2F1 and E2F2 (transcription factors active in cell-growth control) double-knockout[201]. The Gene Ontology pathway STRUCTURAL MOLECULE ACTIVITY was discussed in the preceding section. GSE37605_C57BL6_VS_NOD_FOXP3_FUSION_GFP_TCONV_UP is an immunologic signatures dataset comprising 163 genes in CD4+ T-cells upregulated after introduction of a FOXP3 insertion mutation in mice[202]. This mutation altered T-regulatory cell activity, blocking interaction with HIF-1a but increasing interaction with interferon regulatory factor 4 (IRF4). This mutation also increased the risk of diabetes, but decreased the risk of autoimmune disease, in these mice.

GSE25088_WT_VS_STAT6_KO_MACROPHAGE_ROSIGLITAZONE_STIM_DN, another immunological signatures dataset, includes 200 genes down-regulated in bone marrow-derived macrophages in STAT6 knockout mice[203]. To induce PPARγ transcription these cells were treated with rosiglitazone, an agonist compound used to treat type 2 diabetes by increasing sensitivity to insulin. This study concluded that STAT6 facilitates PPARγ-regulated gene expression in macrophages and dendritic cells. The Gene Ontology pathway MRNA_METABOLIC_PROCESS captures 84 genes involved in mRNA activity. REACTOME

SYNTHESIS SECRETION AND INACTIVATION OF GLP1, another curated pathway, includes 19 genes related to these activities in glucagon-like peptide-1.

Figure 4.5 examines the overlap between these six gene sets. Among overlaps >1 gene in >=2 pathways, Iglesias and SMA share five genes: ANXA1, SEPT7, LAMA4, KRT19, and FBLN2. Rosig and FOXP3 share AHI1, SYNPO, FGD6, and CYBB. Rosig and SMA share ARPC5, EPB42, RPL39, and ACTB. MRNA and FOXP3 share PRPF31 and SF3A1. Iglesias, FOXP3, and SMA share SEPT7. Iglesias and Rosig share CD53, and FOXP and GLP1 share SPCS3.

4.7.3. Comparison of PASCAL and VEGAS2 Results

We also compare results between VEGAS2 and PASCAL. VEGAS2 pathway analysis is not customizable, so VEGAS2 and PASCAL do not compare the same set of pathways, but there is overlap as described above. We highlight results from Gene Ontology and curated pathways that were common to VEGAS2 and PASCAL, but we used more recent versions of Gene Ontology and REACTOME pathways in PASCAL than did VEGAS2, meaning that even within these categories, the intersection of pathways is limited: just 734 of the 3748 v4.0 MSigDB (downloaded July 24, 2014) GO pathways in VEGAS2 overlap with those in in the MSigDB v.5.1 database used in PASCAL (which represents roughly half of the 1454 GO pathways analyzed in PASCAL).

4.7.3.1. Comparison of PASCAL and VEGAS2 Results: All Pathways

Five pathways were in the top 25 results in at least one scenario in both VEGAS2 and PASCAL: Structural Constituent of Muscle (GO:0008307), Reactome Muscle Contraction, Reactome Netrin-1 Signaling, Protein Interaction Database AVB3_Integrin pathway, and Structural Molecule Activity (GO:0005198). Table 4.4 shows p-values for these pathways in each scenario, and Figure 4.6 represents the overlap between genes in these datasets. In contrast to our earlier

results, there was extensive overlap between these pathways, largely because they capture similar phenomena relating to muscle fiber and cytoskeletal integrity.

VEGAS2 results yielded 1,515 unique genes across the top 75 pathways, 51 of which were unique. Reflecting the greater diversity of pathways examined, PASCAL results yielded 3,884 genes. 4,784 unique genes in total were examined across the two routines. 615 genes overlapped between the top 25 pathways for each scenario in both PASCAL and VEGAS2. Table 4.5 shows the most common genes within this set. Reflecting the strong representation of structural constituent pathways in our results, the most common genes included myosin-, collagen-, and nebulin-encoding genes, active in the formation of muscle fibers and cytoskeletal matrix.

Structural molecule activity (GO:0005198) is part of the Gene Ontology molecular_function ontology. It is defined as "The action of a molecule that contributes to the structural integrity of a complex or assembly within or outside a cell[204]." Structural consistent of muscle (GO:0008307) is in turn a subset of structural molecule activity. Structural constituent of muscle, with 34 genes, therefore likely represents a more refined signal.

Unsurprisingly, PASCAL detected signals in immunologic and oncogenic processes that VEGAS2 did not, because of the pathways used in each set. VEGAS2 and PASCAL may not have overlapped precisely in their analyses, but there were similarities in biological processes, e.g. transmembrane trafficking and carbohydrate activity, and guanosine triphosphate (GTP) activity.

We now consider subset analyses among evidence sets overlapping between VEGAS2 and PASCAL: Gene Ontology, and curated pathways from REACTOME and the Protein Interaction Database.

4.7.3.2. Comparison of PASCAL and VEGAS2 Results: Gene Ontology Pathways

Because of the wider range of pathways examined in PASCAL, head-to-head comparison of pathway results should focus on pathways assessed in both programs. This section presents results for both VEGAS2 and PASCAL, subset to Gene Ontology pathways or REACTOME/PID pathways.

Table 4.7 presents VEGAS2 Gene Ontology results, including the top 25 results for each gene-definition cutoff. Of these 75 pathways, 49 were distinct: i.e. they appeared in the top 25 results in more than a single scenario. The top performers in the 0, 20, and 50kb scenarios (in bold font in the table) were, respectively, GO:0005834_heterotrimeric_G-protein_complex (p=2.52E-04), GO:0008307_structural_constituent_of_muscle (p=3.12E-04), and GO:0010596_negative_regulation_of_endothelial_cell_migration (5.00E-04). Nine pathways appeared in the top 25 in all three scenarios; these pathways are primarily active in muscle fiber integrity, vesicular transport, and signal transduction by GTPases, including G-proteins. Functional coherence and robustness to gene boundary definitions increases our confidence in these signals. Alongside this overlap, it is interesting that pathways dealing with histone and protein modification and deubiquitination, along with photoreceptor cell development pathways, were in the top 25 only for the 50KB scenario. This suggests that genes operating in these pathways are especially sensitive to gene boundary definitions.

4.7.3.3. Comparison of PASCAL and VEGAS2 Results: REACTOME and Protein Interaction Database Pathways

Table 4.9 presents VEGAS2 results for REACTOME and Protein Interaction Database pathways. Here, the top performers in the 0, 20, and 50kb scenarios were, respectively, REACTOME_TRANSPORT_OF_INORGANIC_CATIONS_ANIONS_AND_AMINO_ACIDS

_OLIGOPEPTIDES (p=9.60E-04), PID_ARF6_PATHWAY (P=1.80E-03), and REACTOME_N_GLYCAN_ANTENNAE_ELONGATION (P=1.10E-03). We see much less separation between scenarios here than in the Gene Ontology pathways, with 41 distinct pathways distributed across the 75 top results and 11 pathways found in all three scenarios.

In PASCAL, 42 distinct Gene Ontology pathways appeared in the top 25 results for at least one scenario. Of these 42 pathways, 24 also appeared in the older version of GO used by VEGAS2. Across Tables 4.6 and 4.7, just three pathways were found in the top 25 results in any one scenario in both PASCAL and VEGAS2. These include Structural Constituent of Muscle (GO:0008307) and Structural Molecule Activity, for which results can be seen in Table 4.4, but also the Positive Regulation of Cellular Metabolic Process (GO:0031325) pathway (VEGAS2 0kb p-value=2.40E-03; PASCAL 20kb p-value=1.25E-02; PASCAL 50KB p-value=9.12E-03).

4.7.3.4. Comparison of PASCAL and VEGAS2 Results: Gene-Level Statistics

Despite naming differences, there is extensive overlap in function between the top pathways in VEGAS2 and PASCAL. This suggests that key genes might be in play in both settings. To investigate this, Table 4.12 compares gene-level results for PASCAL and VEGAS2. For further depth, two sets of VEGAS2 statistics are presented: one assigning only the top 10% of SNPs to a gene, which we used for our pathway analyses for the computational performance reasons listed earlier, and one assigning the top 100% of SNPs to a gene.

We see that LOC100506122, on chromosome 4, was the top performer across 0kb, 20kb, and 50kb scenarios when mapping the top 10% of SNPs to a gene, with p=1.00E-06 in each scenario. This is consistent with results from the GWAS, where we saw the strongest signal emerge from the ~171Mb-172Mb region in which LOC100506122 is found. A total of 68 SNPs were mapped to this location, for a p-value 0f 9.99*e-7. The top-performing SNP was rs28508193 (p=1.196e-

07). LOC100506122 is a non-coding RNA gene that has been assessed in lupus.

Across Vegas 100% and PASCAL scenarios, we see three genes on chromosome 16, RPUSD1, GNG13, and CHTF18, as top performers in one or more scenarios. RPUSD1 in particular was the top performer in the VEGAS2 0KB and 20KB, and PASCAL 0KB AND 50KB, scenarios, with a p-value <1.10E-04 in each case. CHTF18 was represented in the VEGAS2 50KB 100% scenario (p=8.50E-05), and GNG13 in the PASCAL 20KB scenario (5.43E-05).

Two points are important here. First, the high performance of these chromosome 16 genes is due in part to the assignment of high-impact SNPs to more than one of them, depending on gene boundaries. For instance, rs3765334 is mapped to five of the top ten genes in the VEGAS2 50KB scenario: CHTF18, RPUSD1, GNG13, MIR662, and PRR25. Second, this is taken into account in pathway analyses: PASCAL and VEGAS2 account for this at the pathway level by clumping together genes that are 1) within the same pathway, and within a certain distance of one another: 500KB in VEGAS2, and 1MB in our PASCAL iteration[160-162]. Furthermore, a search of MSigDB yielded just one pathway in which these genes overlap, namely NIKOLSKY BREAST CANCER 16P13 AMPLICON. This pathway did not appear in our top results.

We next consider PASCAL results stratified by pathway collection. This section has ALREADY given results for Gene Ontology and curated pathways in PASCAL; therefore we begin with hallmark pathways, which arguably have the strongest degree of evidence. We then move through evidence types, concluding with pathways generated by the application of machine learning techniques to microarray data and an analysis of pathways containing genes in our candidate-gene study (Chapter 2).

4.7.4. PASCAL Results, Collection-Specific

Here we expand our range of pathways using PASCAL. The purpose is twofold. First, presenting

results specific to a given collection will allow readers to place more emphasis on results for which they feel the evidence base is generally strongest. For instance, molecular biologists may be inclined to view gene expression data as especially strong evidence, and the results of gene knockout experiments in mice as providing translational evidence for human populations. Clinicians or epidemiologists may be less inclined to do so. Presenting results concatenated across all pathway collections is useful for seeing broad patterns in the data, but it can also obscure the collection-specific performance of given pathways. In contrast, presenting results according to collection type allows the reader to investigate top performers given a certain type of evidence.

Second, restricting our results to more "traditional" pathway collections such as Gene Ontology and hallmark pathways would have required us to ignore four collections of potential relevance to NHL: immunologic signatures, oncogenic signatures, computational prediction data, and miRNA and transcription-factor binding motifs. On the whole, we feel that presenting results in this way both enriches the picture emerging from our pathway analyses, and more clearly deals with issues surrounding the type and strength of evidence for particular pathways.

4.7.4.1. PASCAL Results, Collection-Specific: Hallmark Pathways

In concatenated results, the fact that no hallmark pathways appeared in the top-performing pathways would suggest that either (1) pathways with weaker evidence perform better, or that processes and entities of greatest relevance to NHL are not well-captured by hallmark processes. This section considers the performance of specific pathways within the hallmark data.

The Inflammatory Response pathway was the top performer in both 20kb and 50kb scenarios (p=8.13E-03 and 5.38E-03 respectively). Estrogen Response Early was the top performer in the 0kb scenario (p=2.86E-02), where the p-value for Inflammatory Response was 1.36E-01. In all,

15 different pathways appeared in the top ten in at least one of the three scenarios. Five appeared in all three. The fact that some pathways showed up only in the 0kb scenarios, while others showed up only in 20 and/or 50kb scenarios suggests that SNPs in the vicinity of important genes in the inflammatory pathway are not captured by the 0kb definition, but instead lie within 20 kb of their 3' or 5' UTRs. It further suggests that different genes may be contributing more heavily in different pathways.

This second point is borne out by the data. The top inflammatory response gene in the 0 and 20kb scenarios (2.84E-03 and 7.44E-03 respectively), RTP4, sits at ~171.3Mb on chromosome 3 and is described as a "probable chaperone protein which facilitates trafficking and functional cell surface expression of some G-protein coupled receptors (GPCRs)" in GeneCards. We have seen GPCR pathways pop up in other contexts. In contrast, the top gene in the 50kb scenario (p=8.40E-03) is a meta-gene comprising PTGIR (prostaglandin I2 receptor) and C5AR1 (complement component 5a receptor 1).

Among the five pathways in all three scenarios, a linear relationship between gene cutoff and p-value (larger cutoff, smaller p-value) was seen for three: Apical Surface, IL-6/JAK-STAT, and inflammatory response. This could also be because of gene overlap: perhaps a single gene or set of genes is driving this phenomenon in each scenario. Such a finding would not be unexpected, given the common biological thread between the three sets. Just five genes overlap between the inflammatory and estrogen pathways: MYC, P2RY2, RASGRP1, SLC7A2, and TPBG.

Figure 4.8 examines the degree of overlap between gene sets graphically, for the top ten pathways across scenarios. Moving from left to right, the figure shows that inflammatory response has 155 unique genes, IL-6/JAK-STAT 55 unique genes, and apical surface has 35 unique genes. It shows further that inflammatory response and IL-6/JAK-STAT share 22 genes;

apical surface and inflammatory response share four genes, while apical surface and IL-6/JAK-STAT share one gene (IL2RG). Therefore, there is no single common gene between the three that may be driving this, but it should be noted that making reference to the table in conjunction with the matrix, (200-137)/200 = 31% of inflammatory response genes are shared across other pathways. The matrix further illustrates this overlap for pathways besides apical surface and IL-6/JAK-STAT.

4.7.4.2. PASCAL Results, Collection-Specific: Immunologic Signatures

These pathways are generated from manual curation of expression data; they include both human and animal-model data. It should be borne in mind that this was a mouse pathway, and thus supported by weak evidence for association with NHL in humans relative to our curated and hallmark pathways. Observed p-values using the 20KB gene definition fell short of the expected distribution for larger p-values, but were well-aligned at the smaller p-values (<10e-3) with which we should be most concerned. The 50KB definition consistently undershot the expected distribution of p-values, again likely because of the impact on sum scores or fusion gene boundary definitions.

Seven pathways appeared in the top 25 in all three scenarios. Table 4.15 and 4.16 compare their performance. Figure 4.9 plots the intersection of these pathways. We see that these pathways tended to share just a handful of genes, with the greatest overlap seen between the flu vaccine pathway and the IGM/B-cell pathway, at seven genes.

4.7.4.3. PASCAL Results, Collection-Specific: Oncogenic Signatures

Results for oncogenic signatures data are shown in Tables 4.17 and 4.18. Twelve pathways were seen in the top 25 in each scenario, and Table 4.17 shows pathways involving KRAS expression

to be especially dominant among the top results, with KRAS.DF.V1_UP the top performer in the 0kb and 20kb scenarios (p=1.11E-03 and 1.00E-03 respectively), and KRAS.300_UP.V1_UP the top performer in the 50kb scenario (p=2.48 E-03). KRAS.DF.V1_UP captures genes induced by KRAS, in a study that found KRAS-driven cancer to require TBK1[205] of NFKB anti-apoptotic signals for survival. KRAS.300_UP.V1_UP also features genes up-regulated in epithelial cell lines over-expressing oncogenic KRAS[206]. Cytokines, growth factors, and transcription factors are heavily represented in both sets.

Table 4.18 suggests gene overlap between these pathways. ITGA2, on chromosome 5, was the top performer in one or more scenarios for five of the 12 pathways that appeared in the top 25 in each scenario. It is worth noting that ITGA2 was also one of the best-performing genes in our study, with a p-value as low as 4.42E-04 in VEGAS2 gene-based statistics for the 0kb scenario.

Figure 4.10 shows the extent of this overlap, which was especially extensive for the KRAS_LUNG_BREAST sets. KRAS.300_UP.V1_UP, the top performer in our 50kb scenario, shared 38 genes with both KRAS_LUNG_BREAST_UP_V1_UP and KRAS_600_LUNG_BREAST_UP_V1_UP; KRAS.DF.V1_UP, the top performer in our 0 and 20 kb scenarios, shared seven or more genes with these pathways.

4.7.4.4. PASCAL Results, Collection-Specific: Computational Predictions

These pathways are clusters of co-expressed genes generated as described in Subramanian *et al.* 2005 and Segal *et al.* 2004, and as discussed earlier. Results are shown in tables 4.19 and 4.20. The top-performing pathway in the 0kb scenario was Module 101 (p=1.32E-03), a ten-gene module comprising genes involved in glutathione transferase activity and glutathione conjugation reaction. Arrays in which this module was significantly induced or repressed were enriched for liver cancer, squamous-cell lung cancer and large-cell lung cancer. Module 202 was

the top performer in the 20kb (p=1.80E-03) and 50kb (p=8.69E-03) scenarios, driven in part by the gene MYL3 (20kb p=7.17E-04; 50kb p=1.05E-03). Shipp *et al.*[207] identified genes in this module as induced in diffuse large B-cell lymphoma. Notably, genes in this module are also annotated in Gene Ontology as active in muscle fiber and cytoskeletal integrity, and Module 202 contains MYL3, which has been an important gene in pathways discussed throughout the course of this chapter.

Nine pathways appeared in the top 25 in all three scenarios: Module 101 (glutathione activity; enriched for liver cancer), Module 202 (sarcomere/muscle fiber; induced in B-cell lymphoma and suppressed in prostate cancer), Module 132 (ROS metabolism/glutathione activity; enriched for liver and lung cancers), Module 310 (ROS metabolism; enriched for liver, lung and colon cancers), Module 71 (transporter activity; repressed in breast cancer), Module 122 (cell adhesion/extracellular matrix; enriched for multiple cancers), Module 183 (RNA splicing; enriched for B-cell lymphomas and prostate cancer), Module 340 (apoptosis; induced in leukemia), and GNF2_CDC27.

Investigation of gene-set overlap proved to be especially important here. We show this overlap graphically in Figure 4.11. Modules 101, 132, and 310 share ten genes (GSTP, GSTA, and GSTM-family), and Module 101 in fact had no unique genes: it shared all its genes with Modules 132 and 310. However, Module 202, our B-cell lymphoma module and the top performer in the 20 and 50kb scenarios, shared only one gene with another pathway: ANK1 with GNF2_CDC27. The Module 202 signal is therefore distinct relative to other top performers in the computational datasets, and it is encouraging that genes in this module, including MYL3, have been shown to be upregulated in our outcome of interest, B-cell lymphoma. This evidence of upregulation also gives us a clearer view of the possible impact of SNPs in muscle-fiber and

sarcomere-related pathways, including SNPs in the vicinity of MYL3, discussed throughout this chapter.

4.7.4.5. PASCAL Results, Collection-Specific: Transcription Factor & miRNA-Binding Motifs

Results are found in Tables 4.21 and 4.22. TAAWWATAG_V$RSRFC4_Q2, capturing genes with MADS-2 box type promoter regions, was a strong performer.

4.7.5. PASCAL Results: Aim 1 Pathways

As shown in Tables 4.23 and 4.24, results for the 14 pathways analyzed as an extension of our candidate-gene study were disappointing relative to those described thus far. None attained a p-value smaller than 1.36E-01. Insofar as any patterns emerged, Notch pathways performed well, and the genes NOTCH2, DLL4 and ARRB1 were strong performers in these pathways. However, no genes identified in the candidate-gene study were top performers in any of the 14 pathways assessed. The smallest p-values, at both the pathway and gene level, were seen in the 0kb scenario. Table 4.1 shows the number of candidate genes included in each pathway.

4.8. Discussion

Across 13,094 gene sets analyzed in PASCAL, and 6,212 pathways analyzed in VEGAS2, none attained significance at standard Bonferroni-corrected p-value thresholds. However, we find the use of this threshold problematic, and in comparing performance across platforms and scenarios, we focused on ranked results rather than the p-value *per se.* In doing so, clear patterns emerged: gene sets capturing inflammatory processes and muscle fiber and cytoskeletal integrity appeared at or near the top results in most every scenario.

Key inflammation-related pathways included *Iglesias E2F Targets*, comprising 151 cell-growth

control and ductal cell and adipocyte differentiation genes (0kb p=5.02E-05; 20kb p=8.00E-04; 50kb p=1.56E-03), and *GSE37605 FOXP*, capturing genes upregulated in mouse T-cells after introduction of a FOXP3 insertion mutation (0kb p=3.52E-04, 20kb p=5.10E-04, 50kb p=1.70E-03). In hallmark pathways, for which the evidence base is arguably the strongest, our PASCAL analysis identified the inflammatory response pathway as the top pathway in two of three scenarios (20kb p=8.13E-03 and 50kb p=5.38E-03). The prominent place of inflammation-related pathways in our results is consistent with prior knowledge of the biology of NHL. Analysis of expression data and machine-learning modules also highlighted key inflammatory processes, including processes previously linked to B-cell lymphoma. These gene sets included computational Module 202 (20kb p=1.80E-03, 50kb p=8.69E-03), comprising the gene MYL3 and other genes identified by Shipp et al[207] as induced in DLBCL.

Muscle-contraction-related pathways (e.g. GO "Structural Constituent of Muscle," 0kb p= 1.71E-03; 20kb p = 7.90E-04) are prominent in the top results for both PASCAL and VEGAS2. Myosin-related genes account for a high proportion of genes in these pathways; one plausible explanation for these results is that myosin has been implicated in apoptosis, inhibition of which is a classic hallmark of cancer. The prominence of muscle-related pathways could also, conceivably, be tied to enterocyte apoptosis and disruption of gap junctions in microbial translocation, an increasingly prominent process in chronic HIV-related inflammation and, in turn, potentially in the pathogenesis of NHL[74].

A more direct implication of muscle related pathway results for B-cell activation and lymphomagenesis is the link between myosins and their role in the cytoskeleton, which plays an important part in B-cell activation[208]. The cytoskeleton is involved in the aggregation of B cell receptor molecules bound to antigen, in the polarization of these complexes ("capping") and in

their internalization, which is followed by antigen processing and eventually by antigen presentation in conjunction with MHC class II molecules. Therefore, molecules involved in muscle pathways may well be linked to a central biological activity involved in B cell activation.

4.9. Strengths and Limitations

4.9.1. Strengths

This analysis addresses five major concerns in pathway analysis: pathway length, gene size, LD between SNPs, the mapping of a single SNP to multiple genes in a single pathway, and gene overlap between pathways. Both VEGAS2 and PASCAL correct for gene and pathway length using established methods based on Monte Carlo simulation, and we present both graphical and narrative summaries of gene overlap between top pathways across scenarios. Furthermore, we worked to identify key genes driving pathway results, both by describing overlap and by presenting the top gene operating in each pathway where feasible (i.e. in PASCAL). It is also possible for one SNP to be mapped to multiple genes. rs3765334, for instance, maps to five of the top ten genes in the VEGAS2 50KB scenario: CHTF18, RPUSD1, GNG13, MIR662, and PRR25. If these genes then co-occur in a single pathway, the impact of a single SNP will be double-, triple-, or even quintuple-counted by virtue of its occurrence on multiple genes, distorting results[63]. Both VEGAS2 and PASCAL account for this by using the gene-fusion approach described earlier.

We address another major issue in pathway analysis, the strength of evidence for a given pathway, by presenting subgroup analyses by evidence type and clearly identifying the source for each pathway. The breadth of biological phenomena covered by these 13,000 pathways is itself a strength, ranging from one-off expression studies in knockout mice to well-defined hallmark biological processes in humans with support from multiple studies and expert curation and

annotation. All pathways can be explored in detail at the Broad Institute's Molecular Signatures Database (MSigDB).

Our analysis also illustrates the impact of gene boundaries on SNP assignment, presenting analyses that map SNPs to genes only if they are 1) within the gene itself (0kb definition); 2) within 20kb of the 5' and 3' UTR (20kb definition); and 3) within 50kb of the 5' and 3' UTR. On the whole, top-performing pathways tended to appear in the top 25 results across more than one of these scenarios, but we did identify pathways for which this definition made a substantial qualitative difference in rankings. Our results indicate that gene boundary definition is an important consideration in pathway analysis.

Finally, our analyses were conservative. We used very stringent QC criteria in our input data, and we prioritized specificity over sensitivity in our choice of software: VEGAS2 had the highest specificity of any gene-set analysis program benchmarked by Wojcik[159]. Furthermore, pathway analysis Q-Q plots in both VEGAS2 and PASCAL show a pattern of divergence toward the null for observed p-values relative to the expected distribution, rather than away from the null.

4.9.2. Limitations

We recognize that the signals of association observed here are modest at best. This could be due to a relatively small sample size of <2,000 individuals, or it could be due to the legitimate absence of any biological signal. Given the range of phenomena assessed in our study, the former seems more likely. Regardless, full confidence in these findings would require replication in other cohorts or in meta-analyses.

A third possibility is that by prioritizing specificity over sensitivity in our choice of pathway analysis software, we have excluded true positive results that would be detected by other software programs. This could well be, but it is still preferable to err on the side of specificity

rather than sensitivity, especially given continued criticism of GWAS approaches over failures of reproducibility.

Yet another limitation involves the biological interpretation of pathway results. In making sense of a pathway with an especially strong signal, due thought should be given to biological plausibility. Given the range of biological processes covered by our 13,000 pathways, and given the large number of biological processes involved in cancer, most any result can have some degree of plausibility. However, it is encouraging that the inflammatory pathway was a top performer in hallmark data, where strength of evidence is arguably greatest; furthermore, the repeated detection of muscle-fiber related SNPs across pathway collections is indication of biological plausibility for our top results, and may provide new insights into the biological mechanisms involved in lymphomagenesis in the context of infection with HIV.

Table 4.1. Pathways Featuring Aim 1 Genes, After Application of Selection Criteria (n=14).

Gene Set Name	K	k	k/K	p	FDR q
KEGG_BASAL_CELL_CARCINOMA	55	10	0.18	2.86E-24	3.71E-20
HALLMARK_WNT_BETA_CATENIN_ SIGNALING	42	9	0.21	1.43E-22	9.31E-19
KEGG_WNT_SIGNALING_PATHWAY	151	10	0.07	1.19E-19	3.88E-16
WNT_SIGNALING	89	9	0.10	2.02E-19	5.24E-16
KEGG_MELANOGENESIS	102	7	0.07	5.12E-14	1.11E-10
KEGG_NOTCH_SIGNALING_ PATHWAY	47	5	0.11	2.98E-11	5.53E-08
FUKUSHIMA_TNFSF11_TARGETS	16	4	0.25	8.64E-11	1.12E-07
PID_BETA_CATENIN_DEG_PATHWAY	18	4	0.22	1.45E-10	1.71E-07
REACTOME_SIGNALING_BY_NOTCH	103	5	0.05	1.67E-09	1.70E-06
HALLMARK_NOTCH_SIGNALING	32	4	0.13	1.70E-09	1.70E-06
PID_PS1_PATHWAY	46	4	0.09	7.67E-09	7.11E-06
KEGG_ENDOMETRIAL_CANCER	52	4	0.08	1.27E-08	1.10E-05
REACTOME_RECEPTOR_LIGAND_BINDING_INITIATES_THE_SECOND_ PROTEOLYTIC_CLEAVAGE_OF_ NOTCH_RECEPTOR	12	3	0.25	2.40E-08	1.78E-05
KEGG_COLORECTAL_CANCER	62	4	0.06	2.61E-08	1.78E-05

Figure 4.1 Q-Q Plot of Observed Vs. Expected P-Values, Pathway-Level and Gene-Level, VEGAS 0KB, 20KB, and 50KB

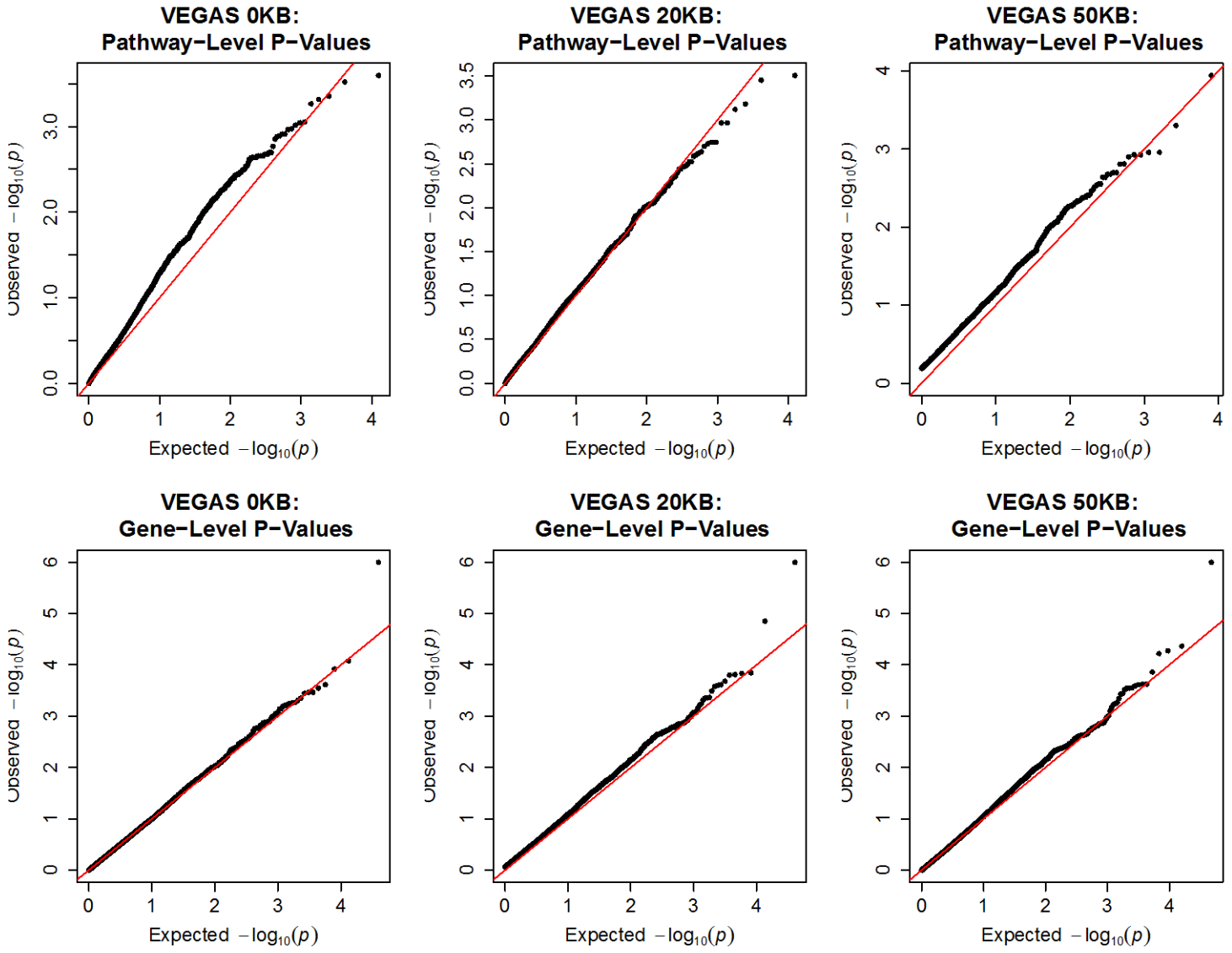


Table 4.2. VEGAS Concatenated Results: Pathway-Level Statistics, 0KB, 20KB and 50KB Scenarios

PATHWAY	PATHWAY P: 0KB	PATHWAY P: 20KB	PATHWAY P: 50KB
GO:0005834_heterotrimeric_G-protein_complex	2.52E-04	7.60E-04	1.18E-03
GO:0045727_positive_regulation_of_translation	3.00E-04	6.60E-04	2.90E-03
PC_Muscle_contraction	4.40E-04	3.00E-03	NA
GO:0008307_structural_constituent_of_muscle^{++}	4.80E-04	3.12E-04	NA
GO:0030742_GTP-dependent_protein_binding	5.40E-04	3.60E-03	1.10E-03
GO:0042147_retrograde_transport__endosome_to_Golgi	8.80E-04	1.80E-03	3.40E-03
GO:0030672_synaptic_vesicle_membrane	9.00E-04	NA	NA
REACTOME_TRANSPORT_OF_INORGANIC_CATIONS_ANIONS_AND _AMINO_ACIDS_OLIGOPEPTIDES	9.60E-04	NA	NA
GO:0031674_I_band	1.06E-03	NA	NA
BIOCARTA_STEM_PATHWAY	1.08E-03	NA	NA
GO:0005198_structural_molecule_activity^{++}	1.22E-03	NA	NA
PID_ARF6_PATHWAY	1.22E-03	1.80E-03	3.30E-03
GO:0060042_retina_morphogenesis_in_camera-type_eye	1.28E-03	1.86E-03	NA
GO:0046943_carboxylic_acid_transmembrane_transporter_activity	1.30E-03	NA	NA
GO:0005096_GTPase_activator_activity	1.40E-03	1.08E-03	1.26E-03
GO:0043236_laminin_binding	1.70E-03	NA	NA
REACTOME_MUSCLE_CONTRACTION^{++}	2.00E-03	3.00E-03	NA
PANTHER_BIOLOGICAL_PROCESS_Homeostasis	2.00E-03	3.52E-04	1.14E-04
GO:0005275_amine_transmembrane_transporter_activity	2.10E-03	NA	NA

REACTOME_STRIATED_MUSCLE_CONTRACTION	2.10E-03	2.60E-03	NA
GO:0030510_regulation_of_BMP_signaling_pathway	2.10E-03	NA	NA
GO:0005342_organic_acid_transmembrane_transporter_activity	2.20E-03	NA	NA
GO:0050906_detection_of_stimulus_involved_in_sensory_perception	2.20E-03	NA	2.80E-03
PC_Amino_acid_and_oligopeptide_SLC_transporters	2.20E-03	NA	NA
PC_Transport_of_inorganic_cations/anions_and_amino_acids/oligopeptides	2.20E-03	NA	NA
GO:0019200_carbohydrate_kinase_activity	NA	1.08E-03	NA
PC_Netrin-1_signaling	NA	2.00E-03	2.30E-03
PID_AVB3_INTEGRIN_PATHWAY^{++}	NA	2.30E-03	1.54E-03
PANTHER_BIOLOGICAL_PROCESS_Carbohydrate_metabolism	NA	2.40E-03	NA
GO:0015491_cation:cation_antiporter_activity	NA	2.50E-03	NA
GO:0002064_epithelial_cell_development	NA	3.20E-03	NA
PID_S1P_S1P3_PATHWAY	NA	3.30E-03	NA
REACTOME_NETRIN1_SIGNALING^{++}	NA	3.40E-03	2.30E-03
GO:0032410_negative_regulation_of_transporter_activity	NA	3.40E-03	NA
GO:0030695_GTPase_regulator_activity	NA	3.60E-03	NA
GO:0019321_pentose_metabolic_process	NA	3.90E-03	1.18E-03
BIOCARTA_UCALPAIN_PATHWAY	NA	4.10E-03	NA
PID_LYSOPHOSPHOLIPID_PATHWAY	NA	4.40E-03	NA
GO:0010596_negative_regulation_of_endothelial_cell_migration	NA	NA	5.00E-04
REACTOME_N_GLYCAN_ANTENNAE_ELONGATION	NA	NA	1.10E-03
GO:0010594_regulation_of_endothelial_cell_migration	NA	NA	1.56E-03

PC_Amino_acid_transport_across_the_plasma_membrane	NA	NA	2.00E-03
GO:0050909_sensory_perception_of_taste	NA	NA	2.00E-03
REACTOME_AMINO_ACID_TRANSPORT_ACROSS_THE_PLASMA_MEMBRANE	NA	NA	2.10E-03
GO:0043679_nerve_terminal	NA	NA	2.10E-03
GO:0042461_photoreceptor_cell_development	NA	NA	2.80E-03
REACTOME_G_ALPHA_S_SIGNALLING_EVENTS	NA	NA	3.00E-03
GO:0046530_photoreceptor_cell_differentiation	NA	NA	3.80E-03
REACTOME_N_GLYCAN_ANTENNAE_ELONGATION_IN_THE_MEDIAL_TRANS_GOLGI	NA	NA	3.90E-03
PID_S1P_S1P1_PATHWAY	NA	NA	3.90E-03
GO:0070646_protein_modification_by_small_protein_removal	NA	NA	4.10E-03

++ Denotes a pathway that also appears in list of top PASCAL results.

Figure 4.2. Gene Overlap between Pathways Appearing in Top 25 Results in All Three Scenarios, VEGAS Concatenated

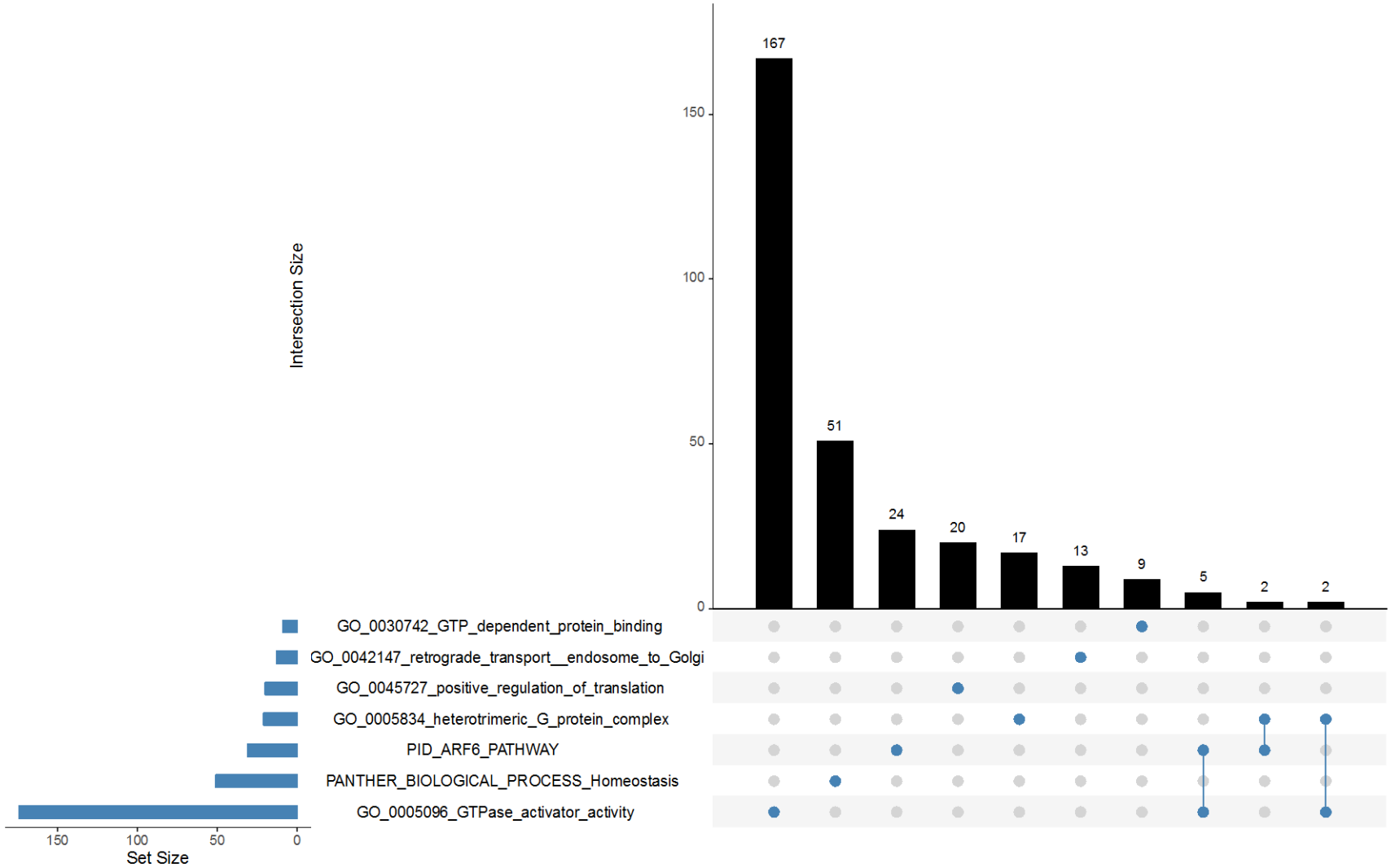


Figure 4.3. Q-Q Plot of Observed vs. Expected P-Values: Genes, Fusion Genes and Pathways, PASCAL 0, 20 & 50KB

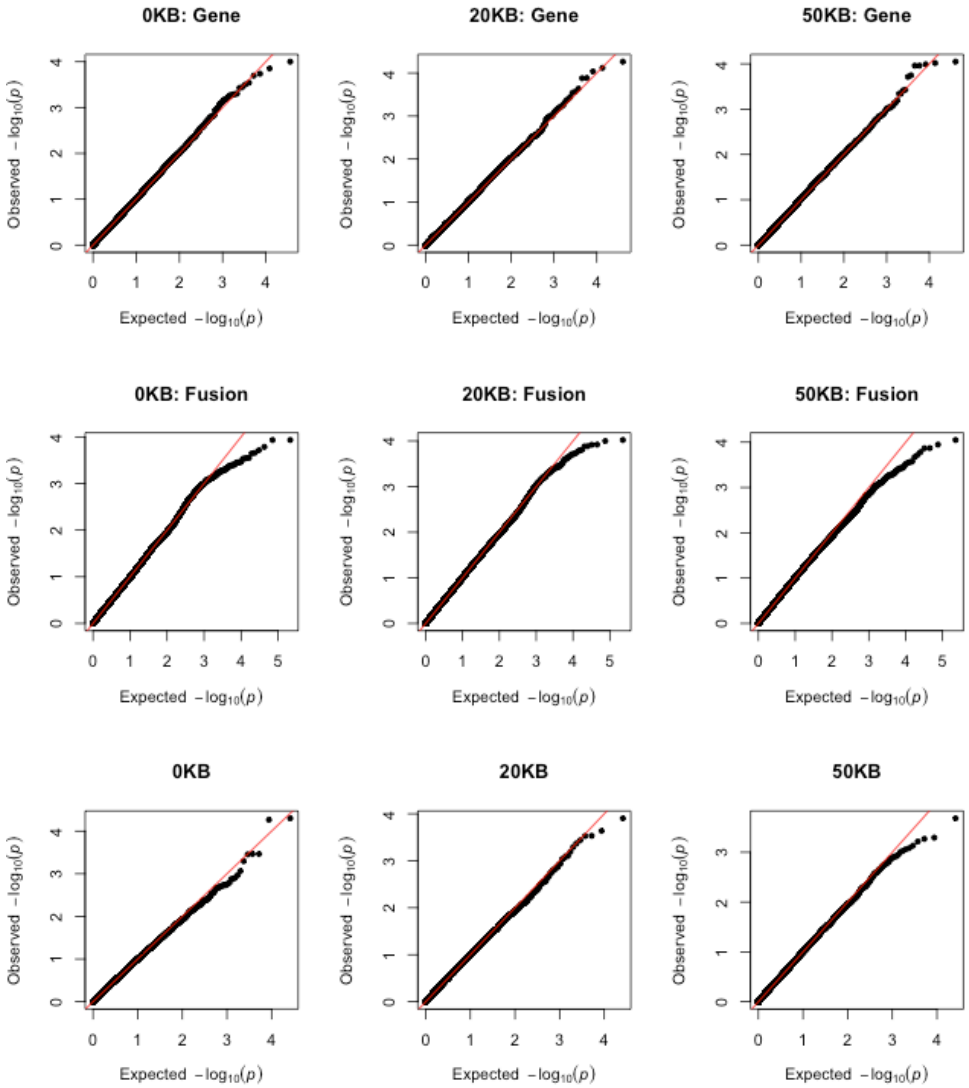


Figure 4.4. Impact of PASCAL Gene Fusion on Pathway Results

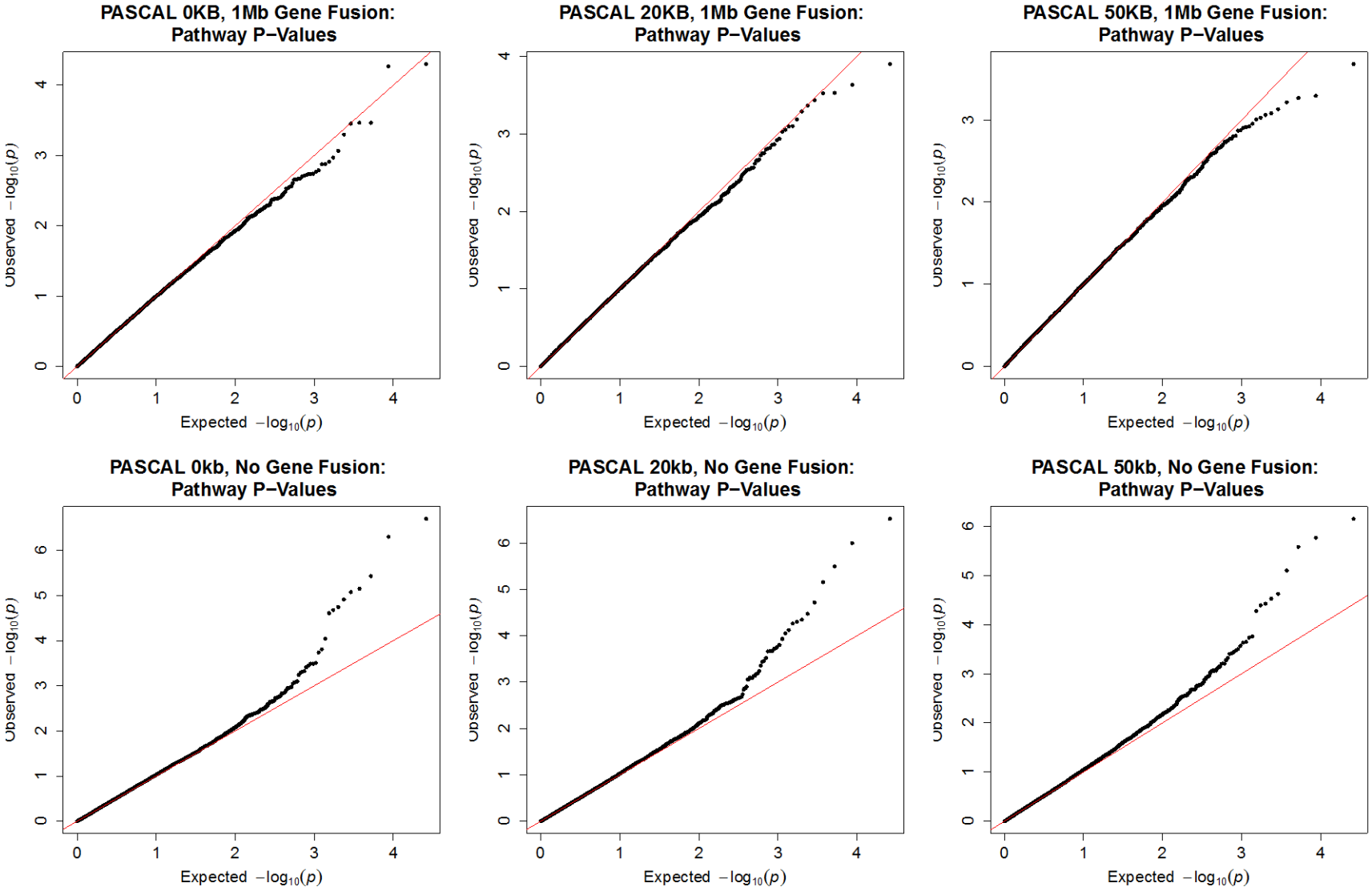


Table 4.3. Pathway-Level Statistics: Top 25 Pathways in 0KB, 20KB and 50KB Scenarios, PASCAL

PATHWAY	EVIDENCE TYPE	PATHWAY P: 0KB	PATHWAY P: 20KB	PATHWAY P: 50KB
IGLESIAS_E2F_TARGETS_UP	CURATED	5.02E-05	8.00E-04	1.56E-03
GSE29614_CTRL_VS_DAY7_TIV_FLU_VACCINE_PBMC_UP	IMM_SIG	5.42E-05	2.31E-04	.
GSE21033_CTRL_VS_POLYIC_STIM_DC_1H_DN	IMM_SIG	3.43E-04	2.94E-04	.
STRUCTURAL_MOLECULE_ACTIVITY	GO	3.43E-04	2.97E-04	5.10E-04
GSE37605_C57BL6_VS_NOD_FOXP3_FUSION_GFP_TCONV_UP	IMM_SIG	3.52E-04	5.10E-04	1.70E-03
GSE21927_SPLENIC_C26GM_TUMOROUS_VS_BONE_MARROW_MONOCYTES_UP	IMM_SIG	5.05E-04	9.30E-04	.
HOWLIN_PUBERTAL_MAMMARY_GLAND	CURATED	8.60E-04	.	.
V$TST1_01	MOTIFS	1.07E-03	.	.
KRAS_DF_V1_UP	ONCO_SIG	1.23E-03	1.20E-03	.
MODULE_101	COMPUT	1.32E-03	.	.
LIEN_BREAST_CARCINOMA_METAPLASTIC	CURATED	1.33E-03	.	.
GSE11864_UNTREATED_VS_CSF1_IN_MAC_DN	IMM_SIG	1.62E-03	.	.
STRUCTURAL_CONSTITUENT_OF_MUSCLE	GO	1.71E-03	7.90E-04	.
GSE25088_WT_VS_STAT6_KO_MACROPHAGE_ROSIGLITAZONE_STIM_DN	IMM_SIG	1.82E-03	4.28E-04	8.70E-04
XU_RESPONSE_TO_TRETINOIN_UP	CURATED	1.84E-03	1.38E-03	.
MAGRANGEAS_MULTIPLE_MYELOMA_IGLL_VS_IGLK_UP	CURATED	1.84E-03	.	.
REACTOME_MEMBRANE_TRAFFICKING	CURATED	1.90E-03	.	.
PODAR_RESPONSE_TO_ADAPHOSTIN_UP	CURATED	1.95E-03	.	6.10E-04
MRNA_METABOLIC_PROCESS	GO	1.97E-03	1.77E-03	1.34E-03
RESPONSE_TO_WOUNDING	GO	2.11E-03	.	.
FARMER_BREAST_CANCER_CLUSTER_6	CURATED	2.13E-03	.	.
REACTOME_SYNTHESIS_SECRETION_AND_INACTIVATION_OF_GLP1	CURATED	2.17E-03	1.36E-03	1.57E-03
PID_HNF3A_PATHWAY	CURATED	2.17E-03	.	.
REACTOME_GAP_JUNCTION_ASSEMBLY	CURATED	2.21E-03	.	.
REACTOME_MUSCLE_CONTRACTION	CURATED	2.48E-03	.	.
TAAWWATAG_V$RSRFC4_Q2	MOTIFS	.	1.25E-04	2.09E-04
WOO_LIVER_CANCER_RECURRENCE_UP	CURATED	.	3.65E-04	9.80E-04

LEIN_CEREBELLUM_MARKERS	CURATED	.	6.45E-04	.
GGTGAAG_MIR_412	MOTIFS	.	8.80E-04	.
COFACTOR_TRANSPORTER_ACTIVITY	GO	.	1.15E-03	2.09E-03
GTATTAT_MIR_369_3P	MOTIFS	.	1.51E-03	1.11E-03
GSE23321_CENTRAL_MEMORY_VS_NAIVE_CD8_TCELL_UP	IMM_SIG	.	1.54E-03	.
COFACTOR_TRANSPORT	GO	.	1.59E-03	.
MODULE_202	COMPUT	.	1.76E-03	.
PID_AVB3_INTEGRIN_PATHWAY	CURATED	.	1.88E-03	7.40E-04
WENG_POR_TARGETS_LIVER_DN	CURATED	.	2.12E-03	.
TIMOFEEVA_GROWTH_STRESS_VIA_STAT1_DN	CURATED	.	2.19E-03	.
GSE24142_DN2_VS_DN3_THYMOCYTE_FETAL_DN	IMM_SIG	.	2.20E-03	1.69E-03
PARK_TRETINOIN_RESPONSE_AND_PML_RARA_FUSION	CURATED	.	.	5.40E-04
REACTOME_NETRIN1_SIGNALING	CURATED	.	.	8.30E-04
SMALL_GTPASE_BINDING	GO	.	.	9.40E-04
KIM_GLIS2_TARGETS_UP	CURATED	.	.	1.20E-03
REACTOME_REGULATION_OF_HYPOXIA_INDUCIBLE_FACTOR_HIF_BY_OXYGEN	CURATED	.	.	1.22E-03
GSE17721_CTRL_VS_POLYIC_2H_BMDM_DN	IMM_SIG	.	.	1.25E-03
GTPASE_BINDING	GO	.	.	1.35E-03
CHEN_LUNG_CANCER_SURVIVAL	CURATED	.	.	1.78E-03
GSE43955_TGFB_IL6_VS_TGFB_IL6_IL23_TH17_ACT_CD4_TCELL_52H_UP	IMM_SIG	.	.	1.86E-03
MRNA_PROCESSING_GO_0006397	GO	.	.	1.86E-03
SANA_TNF_SIGNALING_DN	CURATED	.	.	2.00E-03
MARKS_ACETYLATED_NON_HISTONE_PROTEINS	CURATED	.	.	2.17E-03

Figure 4.5. Gene Overlap between Pathways Appearing in Top 25 Results in All Three Scenarios, PASCAL

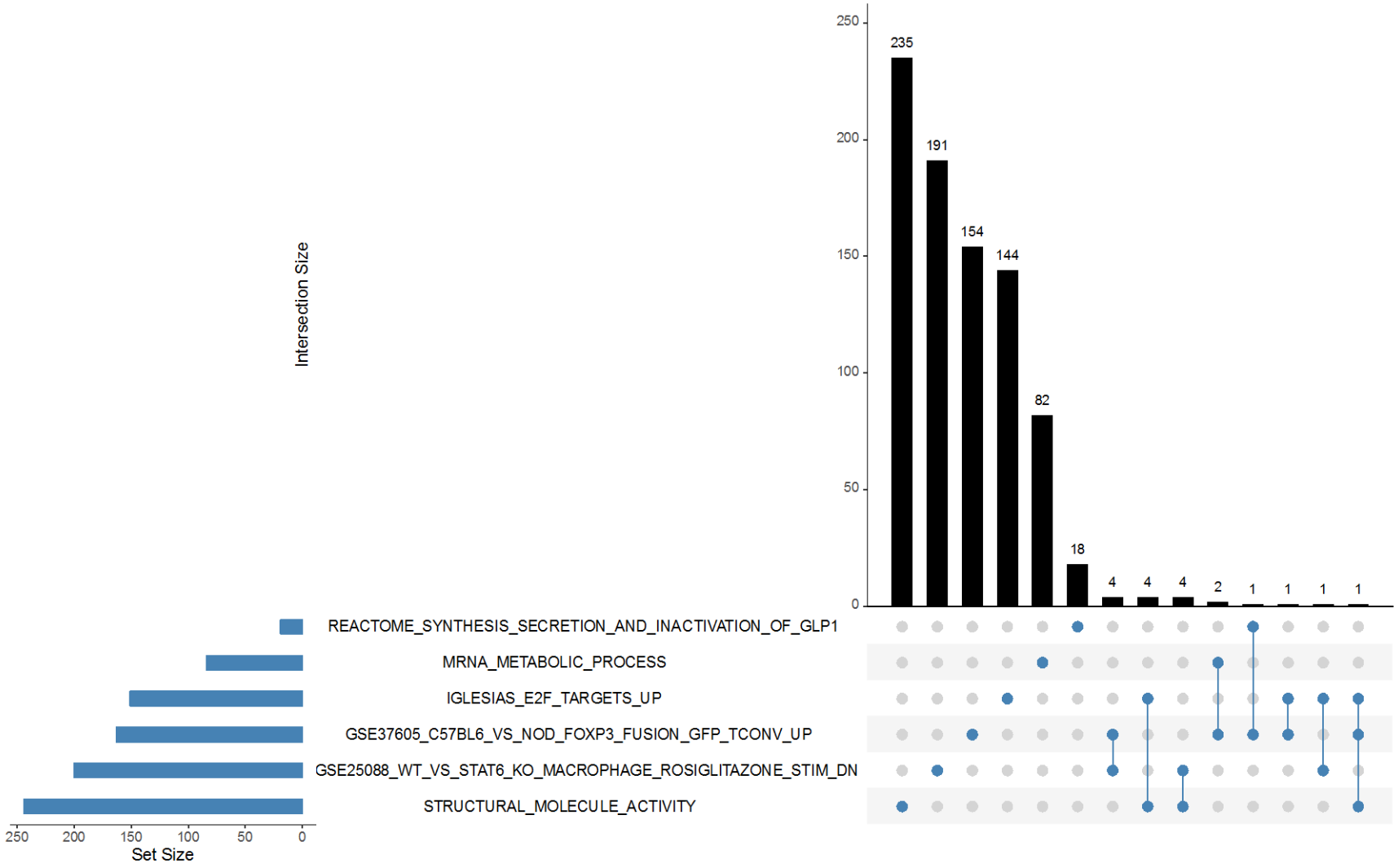


Table 4.4. Gene-Level Statistics: Top 25 Pathways in 0KB, 20KB and 50KB Scenarios, PASCAL

PATHWAY	TOPGENE0	TOPGENEP0	TOPGENE20	TOPGENEP20	TOPGENE50	TOPGENEP50
IGLESIAS_E2F_TARGETS_UP	NA	1.67E-03	POSTN	1.10E-03	POSTN	1.44E-03
GSE29614_CTRL_VS_DAY7_TIV_FLU_VACCINE_PBMC_UP	NA	2.02E-03	RIPPLY3	2.85E-03	.	.
GSE21033_CTRL_VS_POLYIC_STIM_DC_1H_DN	CSF2	3.61E-04	RIPPLY3	2.85E-03	.	.
STRUCTURAL_MOLECULE_ACTIVITY	RPS21	6.88E-04	FBLN2	8.08E-03	BFSP1	2.40E-03
GSE37605_C57BL6_VS_NOD_FOXP3_FUSION_GFP_TCONV_UP	CHTF18	5.20E-04	CHTF18	9.15E-05	CHTF18	9.58E-05
GSE21927_SPLENIC_C26GM_TUMOROUS_VS_BONE_MARROW_MONOCYTES_UP	MYL3	3.74E-04	SLC22A4	4.32E-04	.	.
HOWLIN_PUBERTAL_MAMMARY_GLAND	FOXA1	3.16E-04	.	.	.	.
V$TST1_01	LOC90768	2.24E-03	.	.	.	.
KRAS_DF_V1_UP	ETFB	1.01E-03	ETFB	6.21E-04	.	.
MODULE_101	NA	1.33E-02	.	.	.	.
LIEN_BREAST_CARCINOMA_METAPLASTIC	COL18A1	1.05E-03	.	.	.	.
GSE11864_UNTREATED_VS_CSF1_IN_MAC_DN	GPR157	5.89E-04	.	.	.	.
STRUCTURAL_CONSTITUENT_OF_MUSCLE	MYL3	3.74E-04	MYL3	7.17E-04	.	.
GSE25088_WT_VS_STAT6_KO_MACROPHAGE_ROSIGLITAZONE_STIM_DN	FGD6	1.56E-03	FGD6	2.18E-03	FGD6	2.71E-03
XU_RESPONSE_TO_TRETINOIN_UP	PTK2B	7.56E-04	PTK2B	7.41E-04	.	.
MAGRANGEAS_MULTIPLE_MYELOMA_IGLL_VS_IGLK_UP	TLX1	5.41E-04	.	.	.	.
REACTOME_MEMBRANE_TRAFFICKING	GJC1	6.17E-04	.	.	.	.

PODAR_RESPONSE_TO_ADAPHOSTIN_UP	FOXA1	3.16E-04	.	.	P4HA2	4.58E-04
MRNA_METABOLIC_PROCESS	SF3A3	3.75E-03	SF3A3	7.45E-03	KIN	1.22E-02
RESPONSE_TO_WOUNDING	ITGA2	8.23E-04	.	.	.	.
FARMER_BREAST_CANCER_CLUSTER_6	FOXA1	3.16E-04	.	.	.	.
REACTOME_SYNTHESIS_SECRETION_AND_INACTIVATION_OF_GLP1	GNG13	2.87E-04	GNG13	5.43E-05	GNG13	1.02E-04
PID_HNF3A_PATHWAY	FOXA1	3.16E-04	.	.	.	.
REACTOME_GAP_JUNCTION_ASSEMBLY	GJC1	6.17E-04	.	.	.	.
REACTOME_MUSCLE_CONTRACTION	MYL3	3.74E-04	.	.	.	.
TAAWWATAG_V$RSRFC4_Q2	.	.	MYL3	7.17E-04	FILIP1	7.29E-04
WOO_LIVER_CANCER_RECURRENCE_UP	.	.	POSTN	1.10E-03	POSTN	1.44E-03
LEIN_CEREBELLUM_MARKERS	.	.	GNG13	5.43E-05	.	.
GGTGAAG_MIR_412	.	.	SOX6	4.88E-04	.	.
COFACTOR_TRANSPORTER_ACTIVITY	.	.	NA	6.32E-04	NA	9.08E-04
GTATTAT_MIR_369_3P	.	.	PIKFYVE	1.89E-03	HAO1	1.76E-03
GSE23321_CENTRAL_MEMORY_VS_NAIVE_CD8_TCELL_UP	.	.	PCCB	6.50E-03	.	.
COFACTOR_TRANSPORT	.	.	NA	6.32E-04	.	.
MODULE_202	.	.	MYL3	7.17E-04	.	.
PID_AVB3_INTEGRIN_PATHWAY	.	.	PTK2B	7.41E-04	PTK2B	8.35E-04
WENG_POR_TARGETS_LIVER_DN	.	.	SDS	1.92E-03	.	.
TIMOFEEVA_GROWTH_STRESS_VIA_STAT1_DN	.	.	RAC1	3.51E-03	.	.

GSE24142_DN2_VS_DN3_THYMOCYTE_FETAL_DN	.	.	NA	5.24E-04	NA	1.13E-04
PARK_TRETINOIN_RESPONSE_AND_PML_RARA_FUSION	.	.	.	.	SLC2A5	1.75E-03
REACTOME_NETRIN1_SIGNALING	.	.	.	.	UNC5C	3.76E-03
SMALL_GTPASE_BINDING	.	.	.	.	FGD6	2.71E-03
KIM_GLIS2_TARGETS_UP	.	.	.	.	POSTN	1.44E-03
REACTOME_REGULATION_OF_HYPOXIA_INDUCIBLE_FACTOR_HIF_BY_OXYGEN	.	.	.	.	CUL2	6.89E-03
GSE17721_CTRL_VS_POLYIC_2H_BMDM_DN	.	.	.	.	NA	3.43E-03
GTPASE_BINDING	.	.	.	.	FGD6	2.71E-03
CHEN_LUNG_CANCER_SURVIVAL	.	.	.	.	PGAM1	1.92E-02
GSE43955_TGFB_IL6_VS_TGFB_IL6_IL23_TH17_ACT_CD4_TCELL_52H_UP	.	.	.	.	FOXA1	1.71E-03
MRNA_PROCESSING_GO_0006397	.	.	.	.	KIN	1.22E-02
SANA_TNF_SIGNALING_DN	.	.	.	.	PDLIM4	3.76E-04
MARKS_ACETYLATED_NON_HISTONE_PROTEINS	.	.	.	.	MYOD1	5.76E-03

Table 4.5. P-Values for Five Pathways in Top 25 Results for ≥1 Scenario in both VEGAS and PASCAL

PATHWAY	VEGAS P: 0KB	VEGAS P: 20KB	VEGAS P: 50KB	PASCAL P: 0KB	PASCAL P: 20KB	PASCAL P: 50KB
GO:0008307_structural_constituent_of_muscle	4.80E-04	3.12E-04	.	1.71E-03	7.90E-04	.
GO:0005198_structural_molecule_activity	1.22E-03	.	.	3.43E-04	2.97E-04	5.10E-04
REACTOME_MUSCLE_CONTRACTION	2.00E-03	3.00E-03	.	2.48E-03	.	.
PID_AVB3_INTEGRIN_PATHWAY	.	2.30E-03	1.54E-03	.	1.88E-03	7.40E-04
REACTOME_NETRIN1_SIGNALING	.	3.40E-03	2.30E-03	.	.	8.30E-04

Figure 4.6. Overlap between Five Gene Sets Occurring in Both VEGAS and PASCAL Top 25

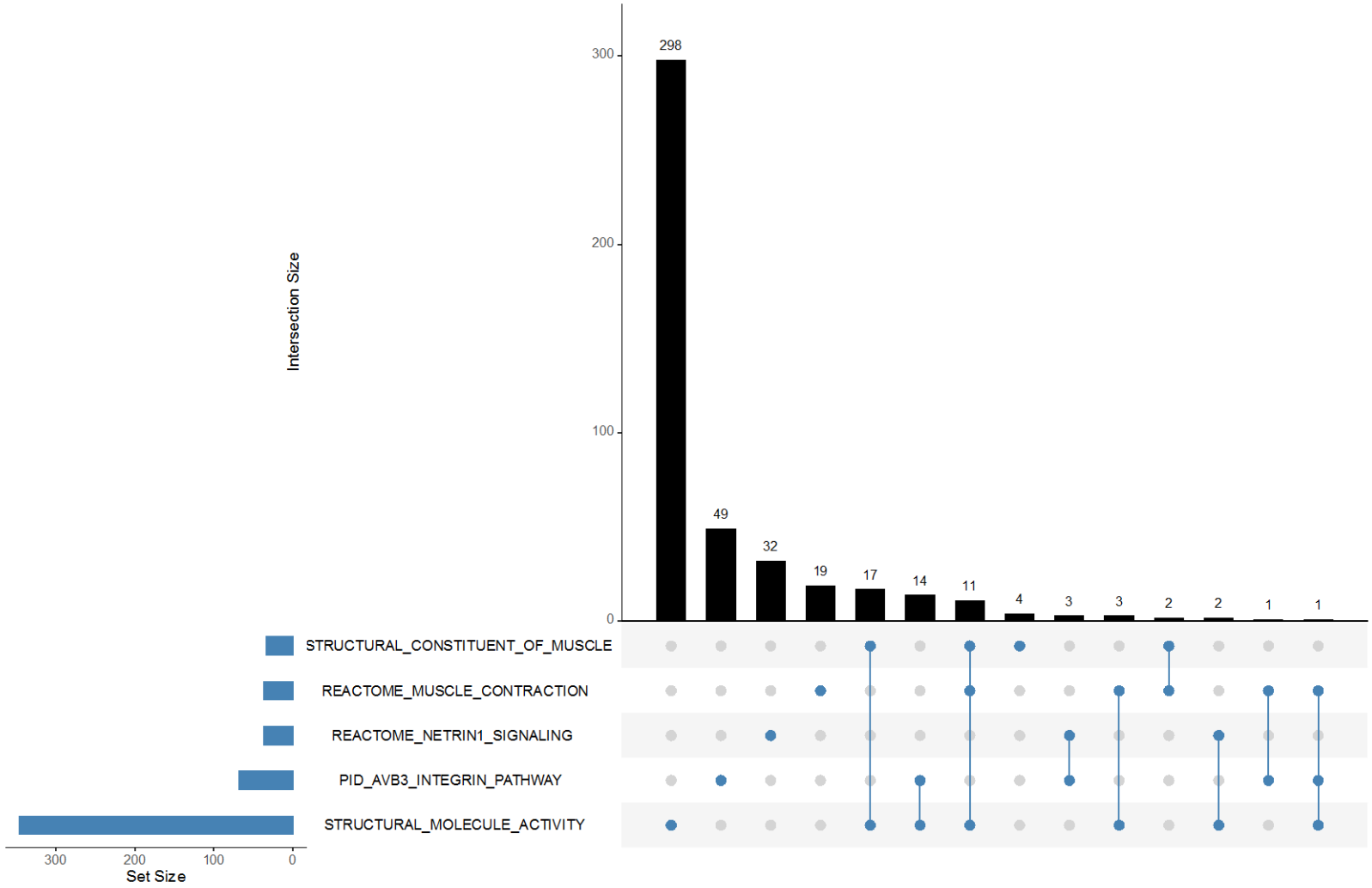


Table 4.6. Most Common Genes among Intersection of Genes in Top Pathways, VEGAS and PASCAL

GENE	PASCAL FREQ	GENE	VEGAS FREQ
MYL3 (Myosin light-chain 3, chr3)	6	RAC1 (Rho family, small GTP-binding protein Rac1, chr7)	13
COL18A1 (collagen type XVIII alpha 1, chr21)	6	MYL3 (Myosin light-chain 3, chr3)	10
ANXA1 (Annexin 1, chr9)	6	NEB (Nebulin, chr2)	10
NEB (Nebulin, chr2)	5	ITGB3 (Integrin subunit beta 3, chr17)	10
TPM4 (Tropomyosin 4, chr19)	5	SRC (SRC proto-oncogene, chr20)	10
MYL6 (Myosin light-chain 6, chr12)	5	TPM4 (Tropomyosin 4, chr19)	9
COL1A1 (Collagen type I alpha 1, chr17)	5	MYBPC1 (Myosin binding protein C, slow type, chr12)	9
FOXA1 (Forkhead box A1, chr14)	5	TLN1 (Talin 1, chr9)	9
FGD6 (FYVE, RhoGEF and PH domain containing 6, chr12)	5	SLC3A2 (Solute carrier family 3 member 2, chr11)	9
RAC1* (Rho family, small GTP-binding protein Rac1, chr7)	4	PXN (Paxillin, chr12)	9

**20 genes with frequency=4 in PASCAL, including RAC1, ITGB3, SRC, MYBPC1*

Table 4.7. Top Results for VEGAS Gene Ontology Pathways: 0, 20, and 50kb Scenarios

PATHWAY	VEGAS P: 0KB	VEGAS P: 20KB	VEGAS P: 50KB
GO:0005834_heterotrimeric_G-protein_complex	**2.52E-04**	7.60E-04	1.18E-03
GO:0045727_positive_regulation_of_translation	3.00E-04	6.60E-04	2.90E-03
GO:0008307_structural_constituent_of_muscle	4.80E-04	**3.12E-04**	4.20E-03
GO:0030742_GTP-dependent_protein_binding	5.40E-04	3.60E-03	1.10E-03
GO:0042147_retrograde_transport__endosome_to_Golgi	8.80E-04	1.80E-03	3.40E-03
GO:0030672_synaptic_vesicle_membrane	9.00E-04	6.50E-03	.
GO:0031674_I_band	1.06E-03	.	.
GO:0005198_structural_molecule_activity	1.22E-03	6.10E-03	.
GO:0060042_retina_morphogenesis_in_camera-type_eye	1.28E-03	1.86E-03	5.70E-03
GO:0046943_carboxylic_acid_transmembrane_transporter_activity	1.30E-03	8.00E-03	.
GO:0005096_GTPase_activator_activity	1.40E-03	1.08E-03	1.26E-03
GO:0043236_laminin_binding	1.70E-03	.	.
GO:0005275_amine_transmembrane_transporter_activity	2.10E-03	.	.
GO:0030510_regulation_of_BMP_signaling_pathway	2.10E-03	.	.
GO:0005342_organic_acid_transmembrane_transporter_activity	2.20E-03	.	.
GO:0050906_detection_of_stimulus_involved_in_sensory_perception	2.20E-03	5.80E-03	2.80E-03
GO:0065009_regulation_of_molecular_function	2.28E-03	.	.
GO:0043679_nerve_terminal	2.30E-03	6.50E-03	2.10E-03
GO:0031047_gene_silencing_by_RNA	2.30E-03	.	.
GO:0031328_positive_regulation_of_cellular_biosynthetic_process	2.40E-03	.	.

GO:0031325_positive_regulation_of_cellular_metabolic_process	2.40E-03	.	.
GO:0010557_positive_regulation_of_macromolecule_biosynthetic_process	2.60E-03	.	.
GO:0015491_cation:cation_antiporter_activity	2.80E-03	2.50E-03	.
GO:0045944_positive_regulation_of_transcription_from_RNA_polymerase_II_promoter	3.30E-03	.	.
GO:0030017_sarcomere	3.40E-03	.	.
GO:0019200_carbohydrate_kinase_activity	.	1.08E-03	.
GO:0002064_epithelial_cell_development	.	3.20E-03	.
GO:0032410_negative_regulation_of_transporter_activity	.	3.40E-03	.
GO:0030695_GTPase_regulator_activity	.	3.60E-03	.
GO:0019321_pentose_metabolic_process	.	3.90E-03	1.18E-03
GO:0010596_negative_regulation_of_endothelial_cell_migration	.	5.30E-03	**5.00E-04**
GO:0060589_nucleoside-triphosphatase_regulator_activity	.	5.60E-03	.
GO:0050974_detection_of_mechanical_stimulus_involved_in_sensory_perception	.	5.60E-03	.
GO:0015101_organic_cation_transmembrane_transporter_activity	.	6.60E-03	4.50E-03
GO:0030374_ligand-dependent_nuclear_receptor_transcription_coactivator_activity	.	6.80E-03	5.10E-03
GO:0016234_inclusion_body	.	7.00E-03	.
GO:0050982_detection_of_mechanical_stimulus	.	7.30E-03	.
GO:0010594_regulation_of_endothelial_cell_migration	.	.	1.56E-03
GO:0050909_sensory_perception_of_taste	.	.	2.00E-03
GO:0042461_photoreceptor_cell_development	.	.	2.80E-03
GO:0046530_photoreceptor_cell_differentiation	.	.	3.80E-03
GO:0070646_protein_modification_by_small_protein_removal	.	.	4.10E-03

GO:0016578_histone_deubiquitination	.	.	4.20E-03
GO:0016579_protein_deubiquitination	.	.	4.70E-03
GO:0042462_eye_photoreceptor_cell_development	.	.	4.90E-03
GO:0048592_eye_morphogenesis	.	.	5.30E-03
GO:0001754_eye_photoreceptor_cell_differentiation	.	.	5.30E-03
GO:0034284_response_to_monosaccharide_stimulus	.	.	5.40E-03
GO:0032925_regulation_of_activin_receptor_signaling_pathway	.	.	5.50E-03

Table 4.8. Top Results for PASCAL Gene Ontology Pathways: 0, 20, and 50kb Scenarios

PATHWAY	PASCAL P: 0KB	PASCAL P: 20KB	PASCAL P: 50KB
GO:0005198_structural_molecule_activity	**3.27E-04**	**3.24E-04**	**5.51E-04**
GO:0008307_structural_constituent_of_muscle	1.72E-03	5.80E-04	2.75E-03
GO:0016071_mrna_metabolic_process	1.89E-03	1.90E-03	1.42E-03
GO:0009611_response_to_wounding	1.94E-03	5.76E-03	.
GO:0006397_mrna_processing_go_0006397	3.80E-03	3.42E-03	2.04E-03
GO:0007043_intercellular_junction_assembly	4.42E-03	1.40E-02	.
GO:0009966_regulation_of_signal_transduction	5.03E-03	.	.
GO:0045216_intercellular_junction_assembly_and_maintenance	5.10E-03	.	.
GO:0043292_contractile_fiber	6.25E-03	.	.
GO:0007028_cytoplasm_organization_and_biogenesis	6.50E-03	.	.
GO:0008305_integrin_complex	7.62E-03	8.77E-03	8.33E-03
GO:0007589_body_fluid_secretion	7.70E-03	.	.
GO:0016491_oxidoreductase_activity_go_0016706	7.51E-03	4.65E-03	3.26E-03
GO:0006139_nucleobasenucleosidenucleotide_and_nucleic_acid_metabolic_process	7.60E-03	8.23E-03	.
GO:0005518_collagen_binding	8.77E-03	9.69E-03	.
GO:0016459_myosin_complex	9.67E-03	1.17E-02	1.23E-02
GO:0019901_protein_kinase_binding	1.01E-02	.	.
GO:0044449_contractile_fiber_part	1.05E-02	.	.
GO:0051184_cofactor_transporter_activity	1.02E-02	1.02E-03	2.16E-03
GO:0051181_cofactor_transport	1.15E-02	1.93E-03	3.64E-03

GO:0051180_vitamin_transport	1.24E-02	2.29E-03	8.43E-03
GO:0006952_defense_response	1.42E-02	.	8.42E-03
GO:0006366_transcription_from_rna_polymerase_ii_promoter	1.39E-02	.	.
GO:0031267_small_gtpase_binding	1.52E-02	1.21E-02	1.17E-03
GO:0035023_regulation_of_rho_protein_signal_transduction	1.40E-02	.	.
GO:0004364_glutathione_transferase_activity	.	4.96E-03	5.62E-03
GO:0046907_intracellular_transport	.	8.54E-03	.
GO:0051641_cellular_localization	.	9.40E-03	.
GO:0001633_secretin_like_receptor_activity	.	1.05E-02	1.42E-02
GO:0030031_cell_projection_biogenesis	.	1.16E-02	1.04E-02
GO:0016071_rna_metabolic_process	.	9.78E-03	.
GO:0031012_extracellular_matrix	.	9.99E-03	3.91E-03
GO:0005578_proteinaceous_extracellular_matrix	.	1.02E-02	4.70E-03
GO:0008375_acetylglucosaminyltransferase_activity	.	1.33E-02	.
GO:0031325_positive_regulation_of_cellular_metabolic_process	.	1.25E-02	9.12E-03
GO:0051020_gtpase_binding	.	.	1.57E-03
GO:0042169_sh2_domain_binding	.	.	6.34E-03
GO:0017016_ras_gtpase_binding	.	.	8.31E-03
GO:0043235_receptor_complex	.	.	9.09E-03
GO:0001501_skeletal_development	.	.	1.05E-02
GO:0016614_oxidoreductase_activity_acting_on_ch_oh_group_of_donors	.	.	1.09E-02
GO:0019899_enzyme_binding	.	.	1.11E-02

Table 4.9. Top Results for VEGAS REACTOME and Protein Interaction Database Pathways, 0, 20, and 50KB Scenarios

PATHWAY	VEGAS P: 0KB	VEGAS P: 20KB	VEGAS P: 50KB
REACTOME_TRANSPORT_OF_INORGANIC_CATIONS_ANIONS_AND_AMINO_ACIDS_OLIGOPEPTIDES	**9.60E-04**	8.40E-03	NA
PID_ARF6_PATHWAY	1.22E-03	**1.80E-03**	3.30E-03
REACTOME_MUSCLE_CONTRACTION	2.00E-03	3.00E-03	8.60E-03
REACTOME_STRIATED_MUSCLE_CONTRACTION	2.10E-03	2.60E-03	4.10E-03
PID_RHODOPSIN_PATHWAY	2.90E-03	.	8.50E-03
REACTOME_REGULATORY_RNA_PATHWAYS	3.10E-03	1.72E-02	.
REACTOME_AMINO_ACID_AND_OLIGOPEPTIDE_SLC_TRANSPORTERS	3.60E-03	1.29E-02	8.60E-03
PID_S1P_S1P3_PATHWAY	3.70E-03	3.30E-03	4.60E-03
PID_TCRCALCIUMPATHWAY	3.70E-03	.	.
PID_LYSOPHOSPHOLIPID_PATHWAY	4.00E-03	4.40E-03	5.80E-03
REACTOME_MICRORNA_MIRNA_BIOGENESIS	4.10E-03	.	.
PID_CONE_PATHWAY	4.20E-03	.	.
REACTOME_NETRIN1_SIGNALING	4.50E-03	3.40E-03	2.30E-03
PID_ENDOTHELINPATHWAY	6.90E-03	7.30E-03	5.40E-03
REACTOME_AMINO_ACID_TRANSPORT_ACROSS_THE_PLASMA_MEMBRANE	7.40E-03	4.70E-03	2.10E-03
PID_INTEGRIN_A4B1_PATHWAY	8.10E-03	5.20E-03	.
REACTOME_G_PROTEIN_ACTIVATION	8.20E-03	1.62E-02	.
REACTOME_GAP_JUNCTION_ASSEMBLY	8.50E-03	.	.
REACTOME_SYNTHESIS_SECRETION_AND_INACTIVATION_OF_GLP1	9.00E-03	1.08E-02	4.80E-03
REACTOME_HYALURONAN_METABOLISM	9.60E-03	.	.

PID_S1P_S1P1_PATHWAY	9.80E-03	1.22E-02	3.90E-03
REACTOME_RNA_POL_I_RNA_POL_III_AND_MITOCHONDRIAL_TRANSCRIPTION	1.00E-02	.	.
REACTOME_FORMATION_OF_TRANSCRIPTION_COUPLED_NER_TC_NER_REPAIR_COMPLEX	1.12E-02	.	.
PID_AR_TF_PATHWAY	1.17E-02	.	.
REACTOME_HYALURONAN_exprTAKE_AND_DEGRADATION	1.28E-02	.	.
PID_AVB3_INTEGRIN_PATHWAY	.	2.30E-03	1.54E-03
REACTOME_N_GLYCAN_ANTENNAE_ELONGATION	.	4.60E-03	**1.10E-03**
PID_NECTIN_PATHWAY	.	9.20E-03	.
REACTOME_DCC_MEDIATED_ATTRACTIVE_SIGNALING	.	9.80E-03	8.80E-03
REACTOME_G_ALPHA_S_SIGNALLING_EVENTS	.	1.01E-02	3.00E-03
REACTOME_N_GLYCAN_ANTENNAE_ELONGATION_IN_THE_MEDIAL_TRANS_GOLGI	.	1.08E-02	3.90E-03
REACTOME_SIGNAL_AMPLIFICATION	.	1.20E-02	9.70E-03
REACTOME_ADP_SIGNALLING_THROUGH_P2RY12	.	1.51E-02	1.14E-02
REACTOME_INCRETIN_SYNTHESIS_SECRETION_AND_INACTIVATION	.	1.76E-02	7.30E-03
REACTOME_REGULATION_OF_HYPOXIA_INDUCIBLE_FACTOR_HIF_BY_OXYGEN	.	2.00E-02	1.12E-02
REACTOME_CLASS_B_2_SECRETIN_FAMILY_RECEPTORS	.	.	7.00E-03
PID_ALK1PATHWAY	.	.	7.50E-03
REACTOME_HIGHLY_CALCIUM_PERMEABLE_POSTSYNAPTIC_NICOTINIC_ACETYLCHOLINE_RECEPTORS	.	.	8.50E-03
REACTOME_INHIBITION_OF_INSULIN_SECRETION_BY_ADRENALINE_NORADRENALINE	.	.	9.80E-03

Table 4.10. Pathway-Level Statistics: PASCAL REACTOME and Protein Interaction Database Pathways, 0, 20, and 50KB

PATHWAY	LENGTH	PATHWAY P: 0KB	PATHWAY P: 20KB	PATHWAY P: 50KB
REACTOME_MEMBRANE_TRAFFICKING	129	**1.95E-03**	8.84E-03	6.58E-03
REACTOME_GAP_JUNCTION_ASSEMBLY	18	1.96E-03	1.73E-02	.
REACTOME_SYNTHESIS_SECRETION_AND_INACTIVATION_OF_GLP1	19	2.21E-03	1.68E-03	1.36E-03
REACTOME_MUSCLE_CONTRACTION	48	2.40E-03	9.41E-03	.
PID_HNF3A_PATHWAY	44	2.54E-03	.	.
REACTOME_NETRIN1_SIGNALING	41	4.58E-03	**3.08E-03**	8.30E-04
REACTOME_HEMOSTASIS	466	5.31E-03	1.01E-02	.
REACTOME_NEURONAL_SYSTEM	279	5.33E-03	.	.
REACTOME_INCRETIN_SYNTHESIS_SECRETION_AND_INACTIVATION	22	5.52E-03	3.16E-03	2.51E-03
REACTOME_STRIATED_MUSCLE_CONTRACTION	27	6.30E-03	5.41E-03	1.07E-02
REACTOME_THROMBIN_SIGNALLING_THROUGH_PROTEINASE_ACTIVATED_ RECEPTORS_PARS	32	8.21E-03	8.82E-03	1.09E-02
REACTOME_DARPP_32_EVENTS	25	9.64E-03	.	.
REACTOME_FACTORS_INVOLVED_IN_MEGAKARYOCYTE_DEVELOPMENT_AND_ PLATELET_PRODUCTION	132	1.07E-02	.	.
REACTOME_GAP_JUNCTION_TRAFFICKING	27	1.12E-02	.	.
PID_INTEGRIN_A4B1_PATHWAY	33	1.16E-02	4.34E-03	2.60E-03
PID_AVB3_INTEGRIN_PATHWAY	75	1.20E-02	1.91E-03	**6.00E-04**
PID_EPHA2_FWD_PATHWAY	19	1.44E-02	6.57E-03	7.21E-03
REACTOME_GABA_SYNTHESIS_RELEASE_REUPTAKE_AND_DEGRADATION	17	1.44E-02	.	.

PID_ENDOTHELIN_PATHWAY	63	1.48E-02	1.09E-02	7.98E-03
REACTOME_SEMA3A_PLEXIN_REPULSION_SIGNALING_BY_INHIBITING_INTEGRIN_ ADHESION	13	1.51E-02	.	1.50E-02
REACTOME_TRANSPORT_OF_GLUCOSE_AND_OTHER_SUGARS_BILE_SALTS_AND_	89	1.55E-02	.	.
ORGANIC_ACIDS_METAL_IONS_AND_AMINE_COMPOUNDS				
REACTOME_SIGNAL_REGULATORY_PROTEIN_SIRP_FAMILY_INTERACTIONS	12	1.82E-02	1.34E-02	1.13E-02
REACTOME_CELL_CELL_COMMUNICATION	120	1.96E-02	1.45E-02	.
REACTOME_SYNTHESIS_OF_PC	18	1.96E-02	.	.
PID_AVB3_OPN_PATHWAY	31	2.01E-02	9.87E-03	9.34E-03
REACTOME_G_ALPHA_S_SIGNALLING_EVENTS	121	.	7.98E-03	4.21E-03
PID_LYSOPHOSPHOLIPID_PATHWAY	66	.	8.12E-03	1.09E-02
PID_CMYB_PATHWAY	84	.	1.03E-02	.
REACTOME_REGULATION_OF_HYPOXIA_INDUCIBLE_FACTOR_HIF_BY_OXYGEN	25	.	1.09E-02	1.32E-03
REACTOME_ADP_SIGNALLING_THROUGH_P2RY1	25	.	1.29E-02	1.08E-02
PID_ARF6_PATHWAY	35	.	1.29E-02	.
REACTOME_CLASS_B_2_SECRETIN_FAMILY_RECEPTORS	88	.	1.42E-02	4.67E-03
PID_CD40_PATHWAY	31	.	1.58E-02	.
REACTOME_DCC_MEDIATED_ATTRACTIVE_SIGNALING	13	.	1.71E-02	8.46E-03
PID_ALK1_PATHWAY	26	.	.	4.76E-03
REACTOME_OXYGEN_DEPENDENT_PROLINE_HYDROXYLATION_OF_	18	.	.	6.28E-03
HYPOXIA_INDUCIBLE_FACTOR_ALPHA				
REACTOME_N_GLYCAN_ANTENNAE_ELONGATION	14	.	.	8.74E-03
REACTOME_HIGHLY_CALCIUM_PERMEABLE_POSTSYNAPTIC_NICOTINIC_	13	.	.	1.29E-02

ACETYLCHOLINE_RECEPTORS				
PID_S1P_S1P1_PATHWAY	21	.	.	1.60E-02
PID_AR_TF_PATHWAY	53	.	.	1.70E-02

Table 4.11. Gene-Level Statistics: PASCAL REACTOME and Protein Interaction Database Pathways, 0, 20, and 50KB

PATHWAY	TOPGENE0	TOPGENEP0	TOPGENE20	TOPGENEP20	TOPGENE50	TOPGENEP50
REACTOME_MEMBRANE_TRAFFICKING	GJC1	6.17E-04	GJC1	2.30E-03	**SEC24C**	5.67E-03
REACTOME_GAP_JUNCTION_ASSEMBLY	GJC1	6.17E-04	GJC1	2.30E-03	.	.
REACTOME_SYNTHESIS_SECRETION_AND_INACTIVATION_OF_GLP1	GNG13	2.87E-04	GNG13	5.43E-05	GNG13	1.02E-04
REACTOME_MUSCLE_CONTRACTION	MYL3	3.74E-04	MYL3	7.17E-04	.	.
PID_HNF3A_PATHWAY	FOXA1	3.16E-04	.	.	.	.
REACTOME_NETRIN1_SIGNALING	RAC1	2.14E-03	UNC5C	2.91E-03	UNC5C	3.76E-03
REACTOME_HEMOSTASIS	GNG13	2.87E-04	GNG13	5.43E-05	.	.
REACTOME_NEURONAL_SYSTEM	GNG13	2.87E-04	.	.	.	.
REACTOME_INCRETIN_SYNTHESIS_SECRETION_AND_INACTIVATION	GNG13	2.87E-04	GNG13	5.43E-05	GNG13	1.02E-04
REACTOME_STRIATED_MUSCLE_CONTRACTION	MYL3	3.74E-04	MYL3	7.17E-04	MYL3	1.05E-03
REACTOME_THROMBIN_SIGNALLING_THROUGH_PROTEINASE_ACTIVATED_RECEPTORS_PARS	GNG13	2.87E-04	GNG13	5.43E-05	GNG13	1.02E-04
REACTOME_DARPP_32_EVENTS	CALM3	1.56E-03	.	.	.	.
REACTOME_FACTORS_INVOLVED_IN_MEGAKARYOCYTE_DEVELOPMENT_AND_PLATELET_PRODUCTION	RAC1	2.14E-03	.	.	.	.
REACTOME_GAP_JUNCTION_TRAFFICKING	GJC1	6.17E-04	.	.	.	.
PID_INTEGRIN_A4B1_PATHWAY	PTK2B	7.56E-04	PTK2B	7.41E-04	PTK2B	8.35E-04
PID_AVB3_INTEGRIN_PATHWAY	PTK2B	7.56E-04	PTK2B	7.41E-04	PTK2B	8.35E-04
PID_EPHA2_FWD_PATHWAY	RAC1	2.14E-03	RAC1	3.51E-03	RAC1	4.22E-03

REACTOME_GABA_SYNTHESIS_RELEASE_REUPTAKE_AND_DEGRADATION	ALDH5A1	7.70E-03	.	.	.	.
PID_ENDOTHELIN_PATHWAY	PTK2B	7.56E-04	PTK2B	7.41E-04	PTK2B	8.35E-04
REACTOME_SEMA3A_PLEXIN_REPULSION_SIGNALING_BY_INHIBITING_INTEGRIN_ADHESION	RAC1	2.14E-03	.	.	RAC1	4.22E-03
REACTOME_TRANSPORT_OF_GLUCOSE_AND_OTHER_SUGARS_BILE_SALTS_AND_ORGANIC_ACIDS_METAL_IONS_AND_AMINE_COMPOUNDS	.	8.46E-04	.	.	.	.
REACTOME_SIGNAL_REGULATORY_PROTEIN_SIRP_FAMILY_INTERACTIONS	PTK2B	7.56E-04	PTK2B	7.41E-04	PTK2B	8.35E-04
REACTOME_CELL_CELL_COMMUNICATION	PTK2B	7.56E-04	PTK2B	7.41E-04	.	.
REACTOME_SYNTHESIS_OF_PC	ACHE	9.22E-03	.	.	.	.
PID_AVB3_OPN_PATHWAY	PTK2B	7.56E-04	PTK2B	7.41E-04	PTK2B	8.35E-04
REACTOME_G_ALPHA_S_SIGNALLING_EVENTS	.	.	GNG13	5.43E-05	GNG13	1.02E-04
PID_LYSOPHOSPHOLIPID_PATHWAY	.	.	PTK2B	7.41E-04	PTK2B	8.35E-04
PID_CMYB_PATHWAY	.	.	PPP3CA	9.60E-03	.	.
REACTOME_REGULATION_OF_HYPOXIA_INDUCIBLE_FACTOR_HIF_BY_OXYGEN	.	.	CA9	1.08E-02	CUL2	6.89E-03
REACTOME_ADP_SIGNALLING_THROUGH_P2RY1	.	.	GNG13	5.43E-05	GNG13	1.02E-04
PID_ARF6_PATHWAY	.	.	ARAP2	9.59E-03	.	.
REACTOME_CLASS_B_2_SECRETIN_FAMILY_RECEPTORS	.	.	GNG13	5.43E-05	GNG13	1.02E-04
PID_CD40_PATHWAY	.	.	MYC	1.37E-02	.	.
REACTOME_DCC_MEDIATED_ATTRACTIVE_SIGNALING	.	.	RAC1	3.51E-03	RAC1	4.22E-03
PID_ALK1_PATHWAY	.	.	.	.	ACVRL1	2.30E-03

REACTOME_OXYGEN_DEPENDENT_PROLINE_HYDROXYLATION_OF_HYPOXIA_INDUCIBLE_FACTOR_ALPHA	.	.	.	.	CUL2	6.89E-03
REACTOME_N_GLYCAN_ANTENNAE_ELONGATION	.	.	.	.	MGAT4B	3.36E-03
REACTOME_HIGHLY_CALCIUM_PERMEABLE_POSTSYNAPTIC_NICOTINIC_ACETYLCHOLINE_RECEPTORS	.	.	.	.	CHRNA2	5.10E-03
PID_S1P_S1P1_PATHWAY	.	.	.	.	RAC1	4.22E-03
PID_AR_TF_PATHWAY	.	.	.	.	NR2C1	2.59E-03

Figure 4.7. Overlap between Gene Sets, PASCAL REACTOME and Protein Interaction Database (PID) Pathways

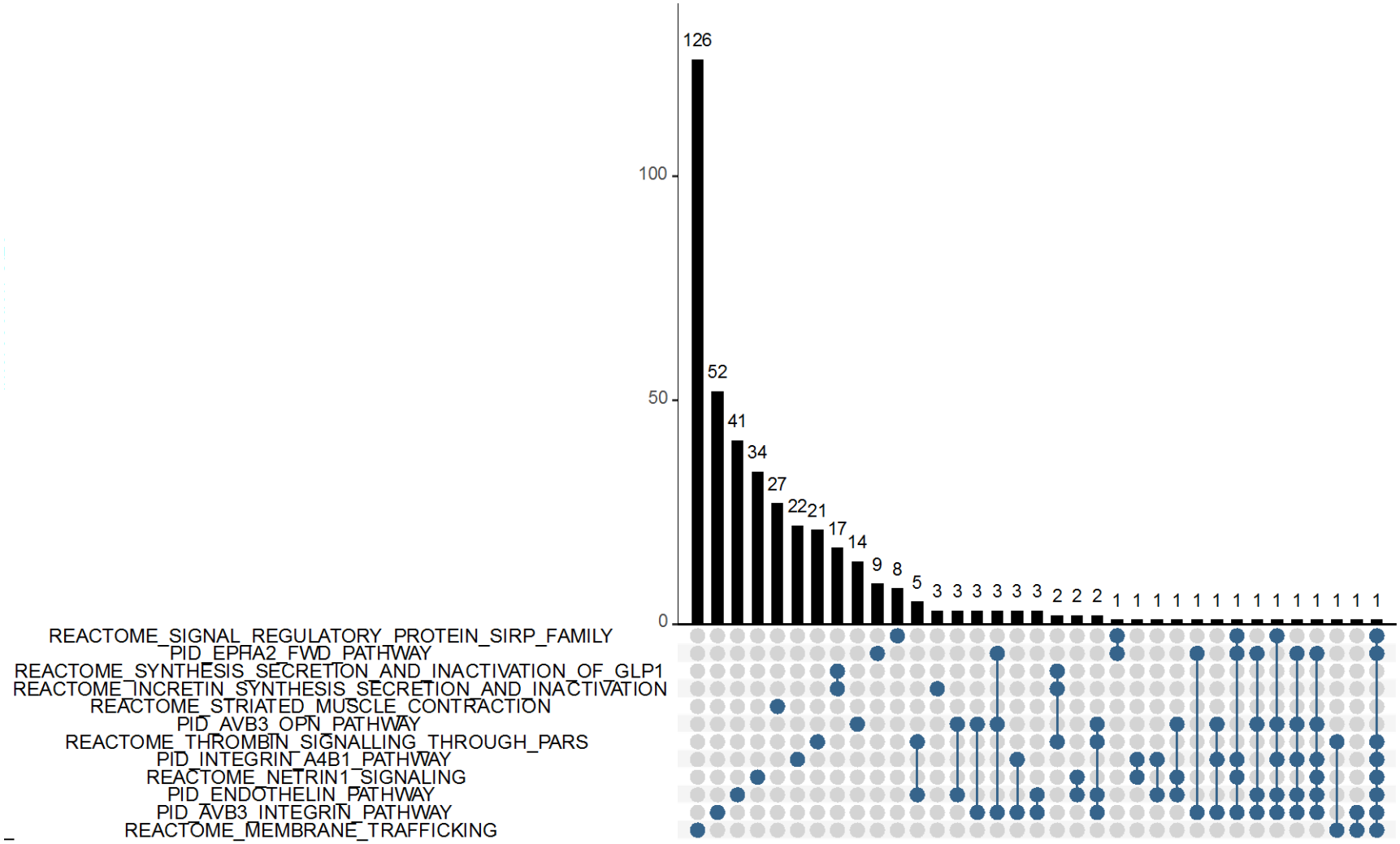


Table 4.12. Gene-Level Statistics: Top-Performing Genes across Nine Scenarios, VEGAS and PASCAL

		VEGAS: Top 10% of SNPs Mapped to Genes			VEGAS: 100% of SNPs Mapped to Genes			PASCAL: 100% of SNPs Mapped to Genes		
Gene	**Chr**	**OKB**	**20KB**	**50KB**	**0KB**	**20KB**	**50KB**	**0KB**	**20KB**	**50KB**
LOC100506122	4	**1.00E-06**	**1.00E-06**	**1.00E-06**	1.27E-04	5.03E-04	.	.	.	.
RPUSD1	16	8.30E-05	2.43E-04	2.38E-04	**1.09E-04**	1.35E-04	9.10E-05	**9.99E-05**	1.30E-04	**8.96E-05**
SHISA5	3	1.20E-04	1.58E-04	5.32E-04	.	.	.	.	.	.
SLC25A21	14	2.41E-04	2.07E-04	2.84E-04	.	.	8.83E-04	6.70E-04	6.81E-04	8.57E-04
P4HA2	5	2.84E-04	5.71E-04	6.13E-04	.	2.58E-04	4.86E-04	9.47E-04	2.87E-04	4.58E-04
PRR25	16	3.41E-04	3.22E-04	2.99E-04	1.34E-04	7.20E-05	1.17E-04	1.41E-04	7.59E-05	1.10E-04
GNG13	16	3.43E-04	2.63E-04	2.82E-04	2.89E-04	**6.40E-05**	1.04E-04	2.87E-04	**5.43E-05**	1.02E-04
CSF2	5	3.56E-04	.	.	3.55E-04	.	.	3.61E-04	.	.
ITGA2	5	4.42E-04	6.88E-04	.	.	.	.	8.23E-04	.	1.16E-03
CHTF18	16	4.90E-04	2.51E-04	2.41E-04	5.17E-04	8.30E-05	**8.50E-05**	5.20E-04	9.15E-05	9.58E-05
SDS	12	5.35E-04	4.37E-04	.	.	.	.	.	.	.
PLBD2	12	5.41E-04	.	.	.	.	.	.	.	.
P4HA2-AS1	5	5.71E-04	4.83E-04	2.36E-04	3.91E-04	4.07E-04	8.60E-05	.	.	.
LOC440311	15	5.74E-04	.	.	.	.	.	.	.	.
HYAL2	3	6.08E-04	.	.	6.78E-04	.	.	.	.	.
ERI1	8	6.13E-04	.	.	.	.	.	.	.	.
MYL3	3	6.46E-04	1.05E-03	.	3.90E-04	5.92E-04	9.02E-04	3.74E-04	7.17E-04	1.05E-03

RPL22	1	7.08E-04	.	.	.	.	.	.	.	.
CEP78	9	7.31E-04	8.91E-04	.	2.02E-04	1.80E-04	3.78E-04	1.83E-04	2.31E-04	3.87E-04
ZSCAN30	18	8.36E-04	.	.	.	.	.	.	.	.
METTL21EP	13	8.70E-04	.	.	.	.	.	.	.	.
NUDCD1	8	8.94E-04	8.29E-04	9.74E-04	.	.	.	.	.	.
ABCC8	11	9.60E-04	.	.	.	.	.	.	.	.
LINC00836	10	1.01E-03	.	.	.	.	.	.	.	.
ARAP2	4	1.03E-03	.	.	.	.	.	.	.	.
ANKRD33	12	.	1.40E-05	5.30E-05	.	.	.	.	9.89E-04	1.22E-03
MSLN	16	.	1.43E-04	3.57E-04	.	5.52E-04	1.83E-04	.	5.59E-04	1.94E-04
PFKFB4	3	.	1.47E-04	2.78E-04	.	.	.	.	.	.
MIR662	16	.	1.55E-04	3.69E-04	.	5.41E-04	1.14E-04	.	5.85E-04	1.09E-04
LOC101929221	15	.	4.31E-04	.	.	3.21E-04	.	.	.	.
MIR6830	5	.	4.41E-04	6.76E-04	.	3.94E-04	1.02E-03	.	.	.
ESYT3	3	.	6.04E-04	.	.	.	.	.	.	.
C14orf132	14	.	8.29E-04	7.80E-04	2.24E-04	3.07E-04	.	2.02E-04	2.71E-04	.
MOCS2	5	.	8.44E-04	5.78E-04	.	.	.	.	.	.
SNORA80A	21	.	1.01E-03	.	.	.	.	.	.	.
CCDC179	11	.	1.06E-03	.	.	.	.	.	.	.
ACVRL1	12	.	.	4.30E-05	.	.	.	.	.	.
FIGNL2	12	.	.	6.00E-05	.	.	.	.	.	.
MIR6823	3	.	.	1.38E-04	.	.	.	.	.	.

TREX1	3	.	.	2.56E-04	.	.	.	.	.	.
ATRIP	3	.	.	4.52E-04	.	.	.	.	.	.
LINC01105	2	.	.	5.58E-04	5.49E-04	3.54E-04	1.66E-04	.	.	.
MIR1258	2	.	.	1.00E-03	.	.	.	.	.	.
IL3	5	.	.	1.07E-03	.	.	.	.	.	.
FOXA1	14	.	.	.	3.37E-04	.	.	3.16E-04	.	.
SLC22A4	5	.	.	.	4.82E-04	4.42E-04	6.73E-04	5.11E-04	4.32E-04	7.12E-04
TEAD4	12	.	.	.	5.42E-04	.	.	6.61E-04	.	.
GJC1	17	.	.	.	5.54E-04	.	.	6.17E-04	.	.
TLX1	10	.	.	.	5.56E-04	.	.	5.41E-04	.	.
GNAI2	3	.	.	.	5.56E-04	.	.	.	.	.
LOC553103	5	.	.	.	5.90E-04	8.07E-04	6.62E-04	.	7.95E-04	6.34E-04
GPR157	1	.	.	.	6.35E-04	.	9.46E-04	5.89E-04	1.00E-03	9.51E-04
RPS21	20	.	.	.	6.84E-04	.	.	6.88E-04	.	.
PTK2B	8	.	.	.	7.88E-04	7.20E-04	8.14E-04	7.56E-04	7.41E-04	8.35E-04
ZNF397	18	.	.	.	8.14E-04	7.26E-04	.	8.39E-04	.	.
FILIP1	6	.	.	.	8.22E-04	.	8.24E-04	7.78E-04	8.58E-04	7.29E-04
LOC102724188	21	.	.	.	8.38E-04	.	.	.	.	.
VSIG10L	19	.	.	.	.	1.26E-04	.	.	1.32E-04	.
IGLON5	19	.	.	.	.	4.74E-04	.	.	4.04E-04	.
LOC400940	2	.	.	.	.	6.00E-04	1.08E-03	.	6.36E-04	9.47E-04
ETFB	19	.	.	.	.	6.02E-04	.	1.01E-03	6.21E-04	.

C1orf61	1	.	.	.	.	6.06E-04	.	.	.	.
MYL10	7	.	.	.	.	6.95E-04	.	.	.	.
PDLIM4	5	.	.	.	.	.	3.81E-04	.	.	3.76E-04
MIR6842	8	.	.	.	.	.	6.40E-04	.	.	.
ENY2	8	.	.	.	.	.	9.32E-04	.	.	9.73E-04
PRSS42	3	.	.	.	.	.	1.00E-03	.	.	1.28E-03
MIR3936	5	.	.	.	.	.	1.01E-03	.	1.00E-03	9.10E-04
C11orf58	11	.	.	.	.	.	1.13E-03	.	9.19E-04	9.24E-04
SOX6	11	.	.	.	.	.	.	4.95E-04	4.88E-04	4.56E-04
LOC150622	2	.	.	.	.	.	.	5.41E-04	3.61E-04	1.79E-04
COL18A1	21	.	.	.	.	.	.	1.05E-03	.	.

Table 4.13. Pathway-Level Statistics: All MSigDB Hallmark Pathways, 0KB, 20KB and 50KB Scenarios

PATHWAY	LENGTH	PATHWAY P: 0KB	PATHWAY P: 20KB	PATHWAY P: 50KB
HALLMARK_ESTROGEN_RESPONSE_EARLY	200	**2.86E-02**	4.15E-02	2.92E-02
HALLMARK_UNFOLDED_PROTEIN_RESPONSE	113	5.42E-02	1.31E-01	1.70E-01
HALLMARK_HYPOXIA	200	7.34E-02	1.77E-01	2.37E-01
HALLMARK_PANCREAS_BETA_CELLS	40	1.04E-01	8.87E-02	2.07E-01
HALLMARK_APICAL_SURFACE	44	1.16E-01	8.63E-02	3.02E-02
HALLMARK_MYOGENESIS	200	1.17E-01	5.02E-02	7.58E-02
HALLMARK_IL6_JAK_STAT3_SIGNALING	87	1.19E-01	1.06E-01	7.17E-02
HALLMARK_MYC_TARGETS_V1	200	1.32E-01	1.52E-01	3.94E-01
HALLMARK_MYC_TARGETS_V2	58	1.33E-01	7.73E-02	2.32E-01
HALLMARK_INFLAMMATORY_RESPONSE	200	1.36E-01	**8.13E-03**	**5.38E-03**
HALLMARK_DNA_REPAIR	150	1.40E-01	3.85E-01	3.53E-01
HALLMARK_PI3K_AKT_MTOR_SIGNALING	105	1.62E-01	4.44E-01	6.90E-01
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	200	1.66E-01	9.11E-02	3.29E-02
HALLMARK_NOTCH_SIGNALING	32	1.86E-01	4.53E-01	5.83E-01
HALLMARK_ESTROGEN_RESPONSE_LATE	200	1.94E-01	9.01E-02	7.63E-02
HALLMARK_HEDGEHOG_SIGNALING	36	2.19E-01	2.10E-01	2.70E-01
HALLMARK_KRAS_SIGNALING_UP	200	2.30E-01	2.96E-01	4.82E-01
HALLMARK_UV_RESPONSE_DN	144	2.48E-01	2.49E-01	4.05E-01
HALLMARK_ADIPOGENESIS	200	2.55E-01	1.25E-01	2.00E-01

HALLMARK_MITOTIC_SPINDLE	200	2.64E-01	2.54E-01	2.25E-01
HALLMARK_E2F_TARGETS	200	2.71E-01	1.97E-01	3.71E-01
HALLMARK_INTERFERON_ALPHA_RESPONSE	97	2.99E-01	1.76E-01	2.39E-01
HALLMARK_HEME_METABOLISM	200	3.28E-01	3.16E-01	5.68E-01
HALLMARK_COMPLEMENT	200	3.32E-01	4.88E-01	4.61E-01
HALLMARK_PROTEIN_SECRETION	96	3.47E-01	2.42E-01	4.08E-01
HALLMARK_COAGULATION	138	3.57E-01	4.03E-01	6.72E-01
HALLMARK_ANDROGEN_RESPONSE	101	3.92E-01	5.91E-01	6.84E-01
HALLMARK_TGF_BETA_SIGNALING	54	4.08E-01	2.48E-01	4.24E-01
HALLMARK_KRAS_SIGNALING_DN	200	4.25E-01	8.78E-01	9.30E-01
HALLMARK_REACTIVE_OXIGEN_SPECIES_PATHWAY	49	4.33E-01	7.92E-01	7.99E-01
HALLMARK_BILE_ACID_METABOLISM	112	4.44E-01	2.20E-01	6.86E-02
HALLMARK_G2M_CHECKPOINT	200	4.99E-01	4.45E-01	4.78E-01
HALLMARK_OXIDATIVE_PHOSPHORYLATION	200	5.34E-01	4.59E-01	5.95E-01
HALLMARK_ANGIOGENESIS	36	5.50E-01	1.69E-01	2.56E-01
HALLMARK_APICAL_JUNCTION	200	5.52E-01	4.72E-01	4.47E-01
HALLMARK_WNT_BETA_CATENIN_SIGNALING	42	5.60E-01	3.35E-01	4.61E-01
HALLMARK_GLYCOLYSIS	200	6.14E-01	1.88E-01	4.80E-02
HALLMARK_INTERFERON_GAMMA_RESPONSE	200	6.77E-01	8.55E-01	9.57E-01
HALLMARK_UV_RESPONSE_UP	158	6.81E-01	5.91E-01	6.41E-01
HALLMARK_IL2_STAT5_SIGNALING	200	6.90E-01	3.46E-01	6.96E-01
HALLMARK_PEROXISOME	104	7.03E-01	6.25E-01	5.21E-01

HALLMARK_FATTY_ACID_METABOLISM	158	7.10E-01	4.33E-01	3.12E-01
HALLMARK_MTORC1_SIGNALING	200	8.39E-01	8.13E-01	7.26E-01
HALLMARK_ALLOGRAFT_REJECTION	200	8.43E-01	4.85E-01	5.87E-01
HALLMARK_XENOBIOTIC_METABOLISM	200	8.78E-01	8.76E-01	8.96E-01
HALLMARK_APOPTOSIS	161	8.87E-01	4.39E-01	4.73E-01
HALLMARK_TNFA_SIGNALING_VIA_NFKB	200	9.24E-01	5.85E-01	5.96E-01
HALLMARK_SPERMATOGENESIS	135	9.56E-01	9.74E-01	9.46E-01
HALLMARK_P53_PATHWAY	200	9.83E-01	8.94E-01	8.93E-01
HALLMARK_CHOLESTEROL_HOMEOSTASIS	74	9.86E-01	9.38E-01	9.34E-01

Table 4.14. Gene-Level Statistics: All MSigDB Hallmark Pathways, 0KB, 20KB and 50KB Scenarios

PATHWAY	TOPGENE 0	TOPGENEP 0	TOPGENE 20	TOPGENEP 20	TOPGENE 50	TOPGENEP 50
HALLMARK_ESTROGEN_RESPONSE_EARLY	SLC22A5	2.03E-03	SLC22A5	1.05E-03	SLC22A5	1.46E-03
HALLMARK_UNFOLDED_PROTEIN_RESPONSE	SPCS3	1.21E-02	CHAC1	1.10E-02	RPS14	1.64E-02
HALLMARK_HYPOXIA	P4HA2	9.47E-04	P4HA2	2.87E-04	P4HA2	4.58E-04
HALLMARK_PANCREAS_BETA_CELLS	ABCC8	9.92E-03	PCSK1	1.24E-02	FOXO1	1.48E-02
HALLMARK_APICAL_SURFACE	GATA3	8.19E-03	GATA3	4.16E-02	AKAP7	3.18E-02
HALLMARK_MYOGENESIS	MYL3	3.74E-04	MYL3	7.17E-04	MYL3	1.05E-03
HALLMARK_IL6_JAK_STAT3_SIGNALING	IL17RB	4.24E-03	IL17RB	4.46E-03	.	5.01E-03
HALLMARK_MYC_TARGETS_V1	CCT5	3.24E-03	POLE3	4.86E-03	CANX	8.10E-03
HALLMARK_MYC_TARGETS_V2	RRP12	1.31E-02	SLC19A1	2.99E-03	SLC19A1	7.08E-03
HALLMARK_INFLAMMATORY_RESPONSE	RTP4	2.84E-03	RTP4	7.44E-03	.	8.40E-03
HALLMARK_DNA_REPAIR	SF3A3	3.75E-03	SF3A3	7.45E-03	SF3A3	1.25E-02
HALLMARK_PI3K_AKT_MTOR_SIGNALING	PIKFYVE	1.55E-03	PIKFYVE	1.89E-03	PIKFYVE	2.61E-03
HALLMARK_EPITHELIAL_MESENCHYMAL_ TRANSITION	ITGA2	8.23E-04	POSTN	1.10E-03	PDLIM4	3.76E-04
HALLMARK_NOTCH_SIGNALING	ARRB1	1.03E-02	ARRB1	1.08E-02	ARRB1	1.35E-02
HALLMARK_ESTROGEN_RESPONSE_LATE	.	2.90E-04	.	1.90E-04	SLC22A5	1.46E-03
HALLMARK_HEDGEHOG_SIGNALING	UNC5C	2.73E-03	UNC5C	2.91E-03	UNC5C	3.76E-03
HALLMARK_KRAS_SIGNALING_UP	CSF2	3.61E-04	ITGA2	1.12E-03	ITGA2	1.16E-03
HALLMARK_UV_RESPONSE_DN	.	6.16E-03	.	7.69E-03	.	1.08E-02

HALLMARK_ADIPOGENESIS	ETFB	1.01E-03	ETFB	6.21E-04	.	2.41E-03
HALLMARK_MITOTIC_SPINDLE	FGD6	1.56E-03	FGD6	2.18E-03	FGD6	2.71E-03
HALLMARK_E2F_TARGETS	STAG1	3.19E-03	STAG1	3.01E-03	STAG1	3.08E-03
HALLMARK_INTERFERON_ALPHA_RESPONSE	RTP4	2.84E-03	RTP4	7.44E-03	CCRL2	7.61E-03
HALLMARK_HEME_METABOLISM	.	5.52E-03	.	1.36E-03	.	1.09E-03
HALLMARK_COMPLEMENT	CALM3	1.56E-03	ZEB1	7.83E-03	ZEB1	6.87E-03
HALLMARK_PROTEIN_SECRETION	AP1G1	1.23E-02	AP1G1	8.45E-03	AP1G1	9.38E-03
HALLMARK_COAGULATION	ITGA2	8.23E-04	ITGA2	1.12E-03	ITGA2	1.16E-03
HALLMARK_ANDROGEN_RESPONSE	PTK2B	7.56E-04	PTK2B	7.41E-04	PTK2B	8.35E-04
HALLMARK_TGF_BETA_SIGNALING	UBE2D3	5.37E-02	BMP2	3.88E-02	TGFB1	2.41E-02
HALLMARK_KRAS_SIGNALING_DN	TLX1	5.41E-04	FGGY	2.05E-02	FGGY	1.89E-02
HALLMARK_REACTIVE_OXIGEN_SPECIES_PATHWAY	HHEX	1.09E-02	GLRX	6.08E-02	FTL	2.12E-02
HALLMARK_BILE_ACID_METABOLISM	SOAT2	1.56E-02	SOAT2	4.97E-03	HAO1	1.76E-03
HALLMARK_G2M_CHECKPOINT	STAG1	3.19E-03	STAG1	3.01E-03	STAG1	3.08E-03
HALLMARK_OXIDATIVE_PHOSPHORYLATION	ETFB	1.01E-03	ETFB	6.21E-04	MTRF1	2.27E-03
HALLMARK_ANGIOGENESIS	POSTN	1.87E-03	POSTN	1.10E-03	POSTN	1.44E-03
HALLMARK_APICAL_JUNCTION	ITGA2	8.23E-04	ITGA2	1.12E-03	ITGA2	1.16E-03
HALLMARK_WNT_BETA_CATENIN_SIGNALING	HDAC11	4.43E-03	MAML1	8.07E-03	MAML1	4.75E-03
HALLMARK_GLYCOLYSIS	P4HA2	9.47E-04	P4HA2	2.87E-04	P4HA2	4.58E-04
HALLMARK_INTERFERON_GAMMA_RESPONSE	RTP4	2.84E-03	RTP4	7.44E-03	RTP4	1.15E-02
HALLMARK_UV_RESPONSE_UP	POLE3	3.46E-03	POLE3	4.86E-03	POLE3	8.11E-03
HALLMARK_IL2_STAT5_SIGNALING	.	8.85E-03	PTH1R	2.26E-03	PTH1R	1.43E-03

HALLMARK_PEROXISOME	ABCC8	9.92E-03	IDH1	2.17E-02	IDH1	6.80E-03
HALLMARK_FATTY_ACID_METABOLISM	SLC22A5	2.03E-03	SLC22A5	1.05E-03	SLC22A5	1.46E-03
HALLMARK_MTORC1_SIGNALING	.	2.76E-03	.	3.91E-03	.	5.01E-03
HALLMARK_ALLOGRAFT_REJECTION	ACHE	9.22E-03	ACHE	1.02E-02	.	9.26E-03
HALLMARK_XENOBIOTIC_METABOLISM	MCCC2	2.05E-02	TAT	1.96E-02	IDH1	6.80E-03
HALLMARK_APOPTOSIS	SQSTM1	7.39E-03	SQSTM1	6.43E-03	.	5.03E-03
HALLMARK_TNFA_SIGNALING_VIA_NFKB	GADD45B	1.54E-02	CCRL2	7.52E-03	.	7.14E-03
HALLMARK_SPERMATOGENESIS	SLC2A5	3.56E-03	SLC2A5	4.90E-03	SLC2A5	1.75E-03
HALLMARK_P53_PATHWAY	SP1	3.21E-02	BMP2	3.88E-02	.	3.52E-02
HALLMARK_CHOLESTEROL_HOMEOSTASIS	FBXO6	1.81E-02	FBXO6	9.18E-03	.	5.27E-02

Figure 4.8. Overlap between Gene Sets, PASCAL MSigDB Hallmark Pathways

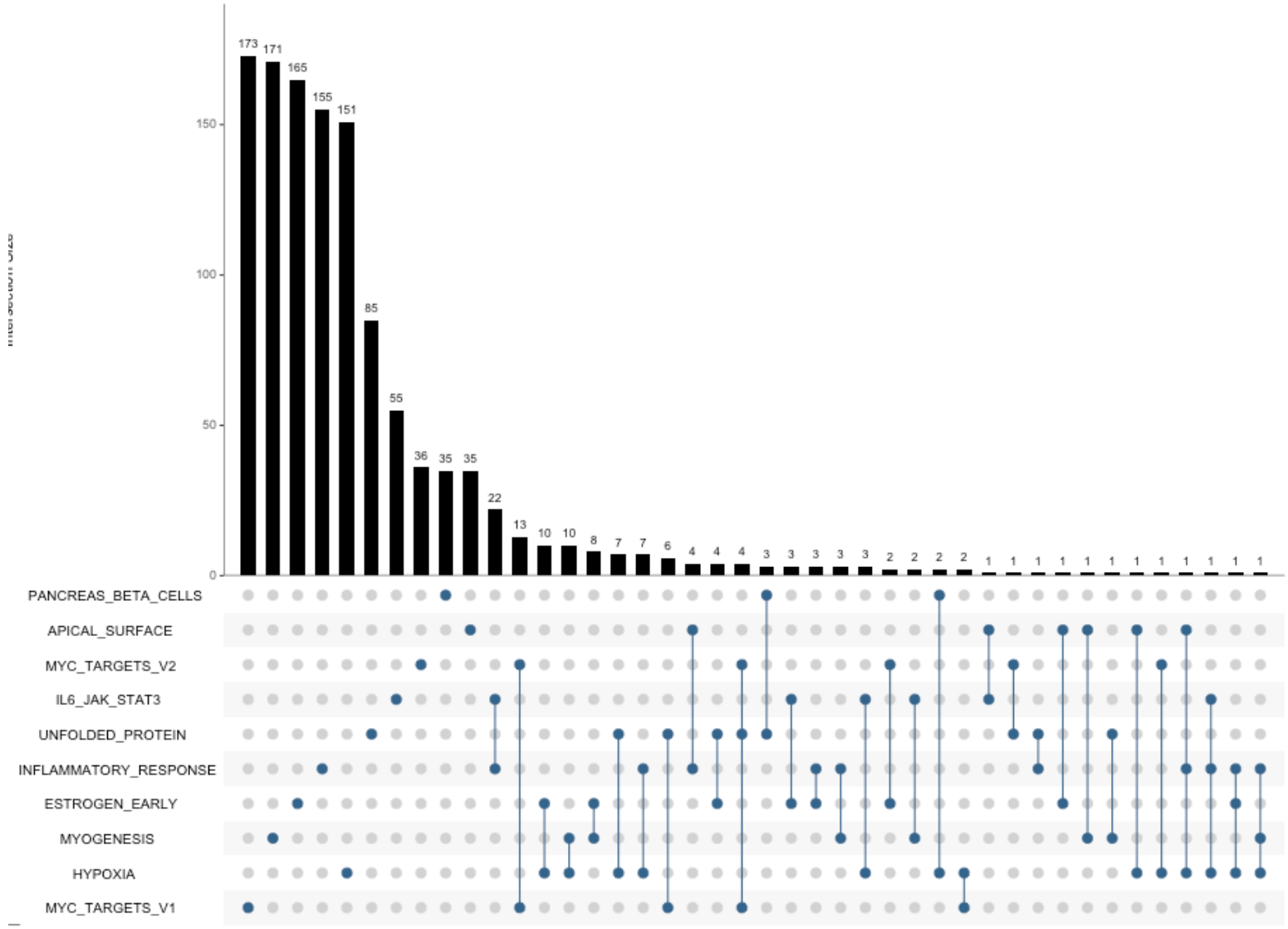


Table 4.15. Pathway-Level Statistics: Top 25 Pathways in 0KB, 20KB and 50KB Scenarios, PASCAL Immunologic Signatures

PATHWAY	LENGTH	PATH P: 0KB	PATH P: 20KB	PATH P: 50KB
GSE29614_CTRL_VS_DAY7_TIV_FLU_VACCINE_PBMC_UP	199	**4.90E-05**	**2.26E-04**	1.09E-02
GSE37605_C57BL6_VS_NOD_FOXP3_FUSION_GFP_TCONV_UP	163	3.03E-04	7.00E-04	1.85E-03
GSE21033_CTRL_VS_POLYIC_STIM_DC_1H_DN	200	3.79E-04	3.06E-04	4.02E-03
GSE21927_SPLENIC_C26GM_TUMOROUS_VS_BONE_MARROW_MONOCYTES_UP	195	5.03E-04	1.06E-03	3.64E-03
GSE25088_WT_VS_STAT6_KO_MACROPHAGE_ROSIGLITAZONE_STIM_DN	200	1.86E-03	4.62E-04	**6.80E-04**
GSE11864_UNTREATED_VS_CSF1_IN_MAC_DN	200	1.93E-03	.	.
GSE14415_NATURAL_TREG_VS_TCONV_UP	150	2.96E-03	1.05E-02	.
GSE18791_CTRL_VS_NEWCASTLE_VIRUS_DC_1H_DN	200	3.86E-03	8.03E-03	.
GSE13306_RA_VS_UNTREATED_MEM_CD4_TCELL_DN	200	3.91E-03	.	.
GSE17721_POLYIC_VS_GARDIQUIMOD_6H_BMDM_UP	200	4.19E-03	5.54E-03	.
GSE360_CTRL_VS_L_MAJOR_DC_UP	200	4.59E-03	7.92E-03	3.16E-03
GSE5542_UNTREATED_VS_IFNA_TREATED_EPITHELIAL_CELLS_24H_UP	197	5.56E-03	5.83E-03	.
GSE18281_SUBCAPSULAR_CORTICAL_REGION_VS_WHOLE_MEDULLA_THYMUS_DN	200	5.71E-03	.	.
GSE2770_UNTREATED_VS_TGFB_AND_IL12_TREATED_ACT_CD4_TCELL_48H_UP	199	5.80E-03	.	.
GSE2124_CTRL_VS_LYMPHOTOXIN_BETA_TREATED_MLN_UP	199	5.94E-03	.	7.89E-03
GSE34515_CD16_POS_MONOCYTE_VS_DC_UP	200	6.05E-03	.	.
GSE45365_HEALTHY_VS_MCMV_INFECTION_CD11B_DC_IFNAR_KO_DN	197	6.11E-03	.	.
GSE27092_WT_VS_HDAC7_PHOSPHO_DEFICIENT_CD8_TCELL_UP	200	6.71E-03	9.44E-03	.
GSE24634_NAIVE_CD4_TCELL_VS_DAY5_IL4_CONV_TREG_UP	200	7.04E-03	.	.
GSE43955_TGFB_IL6_VS_TGFB_IL6_IL23_TH17_ACT_CD4_TCELL_52H_UP	200	7.11E-03	6.92E-03	2.09E-03

GSE25846_IL10_POS_VS_NEG_CD8_TCELL_DAY7_POST_CORONAVIRUS_BRAIN_DN	200	7.15E-03	.	.
GSE17721_LPS_VS_CPG_2H_BMDM_DN	200	7.23E-03	.	.
GSE22443_NAIVE_VS_ACT_AND_IL12_TREATED_CD8_TCELL_UP	200	7.38E-03	.	.
GSE17721_PAM3CSK4_VS_CPG_0.5H_BMDM_DN	200	7.56E-03	.	.
GSE27859_DC_VS_CD11C_INT_F480_HI_MACROPHAGE_UP	200	7.79E-03	.	.
GSE23321_CENTRAL_MEMORY_VS_NAIVE_CD8_TCELL_UP	200	.	1.54E-03	5.64E-03
GSE24142_DN2_VS_DN3_THYMOCYTE_FETAL_DN	200	.	2.38E-03	1.51E-03
GSE46242_CTRL_VS_EGR2_DELETED_TH1_CD4_TCELL_UP	199	.	2.74E-03	3.28E-03
GSE25088_CTRL_VS_ROSIGLITAZONE_STIM_MACROPHAGE_DN	197	.	3.80E-03	7.90E-03
GSE12392_IFNAR_KO_VS_IFNB_KO_CD8_NEG_SPLEEN_DC_DN	200	.	4.78E-03	.
GSE3920_IFNA_VS_IFNG_TREATED_FIBROBLAST_UP	175	.	6.00E-03	.
GSE17721_CTRL_VS_POLYIC_2H_BMDM_DN	200	.	6.95E-03	1.27E-03
GSE9037_WT_VS_IRAK4_KO_LPS_4H_STIM_BMDM_UP	200	.	7.74E-03	8.88E-03
GSE14908_ATOPIC_VS_NONATOPIC_PATIENT_RESTING_CD4_TCELL_UP	200	.	7.77E-03	.
GSE13411_IGM_VS_SWITCHED_MEMORY_BCELL_UP	200	.	8.25E-03	2.79E-03
GSE43863_NAIVE_VS_LY6C_INT_CXCR5POS_CD4_EFF_TCELL_D6_LCMV_DN	191	.	8.56E-03	4.10E-03
GSE12003_MIR223_KO_VS_WT_BM_PROGENITOR_4D_CULTURE_DN	200	.	1.05E-02	.
GSE41867_NAIVE_VS_DAY15_LCMV_EFFECTOR_CD8_TCELL_DN	200	.	1.16E-02	.
GSE3337_4H_VS_16H_IFNG_IN_CD8POS_DC_DN	200	.	.	4.95E-03
GSE17301_CTRL_VS_48H_IFNA2_STIM_CD8_TCELL_UP	200	.	.	5.24E-03
GSE12392_CD8A_POS_VS_NEG_SPLEEN_IFNB_KO_DC_UP	200	.	.	6.43E-03
GSE24142_ADULT_VS_FETAL_DN3_THYMOCYTE_DN	200	.	.	6.75E-03

GSE24634_NAIVE_CD4_TCELL_VS_DAY7_IL4_CONV_TREG_UP	200	.	.	7.27E-03
GSE19888_CTRL_VS_A3R_ACT_TREATED_MAST_CELL_PRETREATED_WITH_A3R_INH_DN	200	.	.	7.51E-03
GSE13762_CTRL_VS_125_VITAMIND_DAY12_DC_UP	147	.	.	8.03E-03
GSE6269_FLU_VS_E_COLI_INF_PBMC_UP	162	.	.	9.55E-03
GSE22025_TGFB1_VS_TGFB1_AND_PROGESTERONE_TREATED_CD4_TCELL_UP	200	.	.	1.04E-02

Table 4.16. Gene-Level Statistics: Top 25 Pathways in 0KB, 20KB and 50KB Scenarios, PASCAL Immunologic Signatures

PATHWAY	TOPGENE 0	GENEP0	GENE 20	GENEP20	GENE 50	GENEP 50
GSE29614_CTRL_VS_DAY7_TIV_FLU_VACCINE_PBMC_UP	.	2.02E-03	RIPPLY3	2.85E-03	ACVRL1	2.30E-03
GSE37605_C57BL6_VS_NOD_FOXP3_FUSION_GFP_TCONV_UP	CHTF18	5.20E-04	CHTF18	9.15E-05	CHTF18	9.58E-05
GSE21033_CTRL_VS_POLYIC_STIM_DC_1H_DN	CSF2	3.61E-04	RIPPLY3	2.85E-03	RIPPLY3	4.41E-03
GSE21927_SPLENIC_C26GM_TUMOROUS_VS_BONE_MARROW_MONOCYTES_UP	MYL3	3.74E-04	SLC22A4	4.32E-04	SLC22A4	7.12E-04
GSE25088_WT_VS_STAT6_KO_MACROPHAGE_ROSIGLITAZONE_STIM_DN	FGD6	1.56E-03	FGD6	2.18E-03	FGD6	2.71E-03
GSE11864_UNTREATED_VS_CSF1_IN_MAC_DN	GPR157	5.89E-04	.	.	.	.
GSE14415_NATURAL_TREG_VS_TCONV_UP	CSF2	3.61E-04	FILIP1	8.58E-04	.	.
GSE18791_CTRL_VS_NEWCASTLE_VIRUS_DC_1H_DN	SLC22A4	5.11E-04	SLC22A4	4.32E-04	.	.
GSE13306_RA_VS_UNTREATED_MEM_CD4_TCELL_DN	.	2.22E-03	.	.	.	.
GSE17721_POLYIC_VS_GARDIQUIMOD_6H_BMDM_UP	CHTF18	5.20E-04	CHTF18	9.15E-05	.	.
GSE360_CTRL_VS_L_MAJOR_DC_UP	PCCB	5.89E-03	COL15A1	4.60E-03	COL15A1	4.16E-03
GSE5542_UNTREATED_VS_IFNA_TREATED_EPITHELIAL_CELLS_24H_UP	P4HA2	9.47E-04	P4HA2	2.87E-04	.	.
GSE18281_SUBCAPSULAR_CORTICAL_REGION_VS_WHOLE_MEDULLA_THYMUS_DN	GNG13	2.87E-04	.	.	.	.
GSE2770_UNTREATED_VS_TGFB_AND_IL12_TREA	LOC100506804	4.10E-03	.	.	.	.

TED_ACT_CD4_TCELL_48H_UP						
GSE2124_CTRL_VS_LYMPHOTOXIN_BETA_TREATED_MLN_UP	.	2.74E-03	.	.	BFSP1	2.40E-03
GSE34515_CD16_POS_MONOCYTE_VS_DC_UP	FOXA1	3.16E-04	.	.	.	.
GSE45365_HEALTHY_VS_MCMV_INFECTION_CD11B_DC_IFNAR_KO_DN	CALM3	1.56E-03	.	.	.	.
GSE27092_WT_VS_HDAC7_PHOSPHO_DEFICIENT_CD8_TCELL_UP	FOXA1	3.16E-04	FOXA1	1.60E-03	.	.
GSE24634_NAIVE_CD4_TCELL_VS_DAY5_IL4_CONV_TREG_UP	RPL11	4.35E-03	.	.	.	.
GSE43955_TGFB_IL6_VS_TGFB_IL6_IL23_TH17_ACT_CD4_TCELL_52H_UP	FOXA1	3.16E-04	FOXA1	1.60E-03	FOXA1	1.71E-03
GSE25846_IL10_POS_VS_NEG_CD8_TCELL_DAY7_POST_CORONAVIRUS_BRAIN_DN	CSF2	3.61E-04	.	.	.	.
GSE17721_LPS_VS_CPG_2H_BMDM_DN	NKX2-3	1.46E-03	.	.	.	.
GSE22443_NAIVE_VS_ACT_AND_IL12_TREATED_CD8_TCELL_UP	STAG1	3.19E-03	.	.	.	.
GSE17721_PAM3CSK4_VS_CPG_0.5H_BMDM_DN	PIKFYVE	1.55E-03	.	.	.	.
GSE27859_DC_VS_CD11C_INT_F480_HI_MACROPHAGE_UP	RTP4	2.84E-03	.	.	.	.
GSE23321_CENTRAL_MEMORY_VS_NAIVE_CD8_TCELL_UP	.	.	PCCB	6.50E-03	CHAF1B	6.16E-03
GSE24142_DN2_VS_DN3_THYMOCYTE_FETAL_DN	.	.	.	5.24E-04	.	1.13E-04
GSE46242_CTRL_VS_EGR2_DELETED_TH1_CD4_TCELL_UP	.	.	UQCR10	5.09E-03	IGFBP6	3.68E-03
GSE25088_CTRL_VS_ROSIGLITAZONE_STIM_MACROPHAGE_DN	.	.	ANKRD33	9.89E-04	ANKRD33	1.22E-03

GSE12392_IFNAR_KO_VS_IFNB_KO_CD8_NEG_SPLEEN_DC_DN	.	.	.	7.19E-04	.	.
GSE3920_IFNA_VS_IFNG_TREATED_FIBROBLAST_UP	.	.	PTK2B	7.41E-04	.	.
GSE17721_CTRL_VS_POLYIC_2H_BMDM_DN	.	.	.	1.57E-03	.	3.43E-03
GSE9037_WT_VS_IRAK4_KO_LPS_4H_STIM_BMDM_UP	.	.	ZNF397	1.00E-03	ZNF397	2.08E-03
GSE14908_ATOPIC_VS_NONATOPIC_PATIENT_RESTING_CD4_TCELL_UP	.	.	COL18A1	1.46E-03	.	.
GSE13411_IGM_VS_SWITCHED_MEMORY_BCELL_UP	.	.	C14orf132	2.71E-04	SLC22A4	7.12E-04
GSE43863_NAIVE_VS_LY6C_INT_CXCR5POS_CD4_EFF_TCELL_D6_LCMV_DN	.	.	PTH1R	2.26E-03	PTH1R	1.43E-03
GSE12003_MIR223_KO_VS_WT_BM_PROGENITOR_4D_CULTURE_DN	.	.	.	2.16E-03	.	.
GSE41867_NAIVE_VS_DAY15_LCMV_EFFECTOR_CD8_TCELL_DN	.	.	PIKFYVE	1.89E-03	.	.
GSE3337_4H_VS_16H_IFNG_IN_CD8POS_DC_DN	.	.	.	.	SOX6	4.56E-04
GSE17301_CTRL_VS_48H_IFNA2_STIM_CD8_TCELL_UP	.	.	.	.	STAG1	3.08E-03
GSE12392_CD8A_POS_VS_NEG_SPLEEN_IFNB_KO_DC_UP	.	.	.	.	NR2C1	2.59E-03
GSE24142_ADULT_VS_FETAL_DN3_THYMOCYTE_DN	.	.	.	.	SOX6	4.56E-04
GSE24634_NAIVE_CD4_TCELL_VS_DAY7_IL4_CONV_TREG_UP	.	.	.	.	ZBTB20	5.67E-03
GSE19888_CTRL_VS_A3R_ACT_TREATED_MAST_CELL_PRETREATED_WITH_A3R_INH_DN	.	.	.	.	IGFBP6	3.68E-03

GSE13762_CTRL_VS_125_VITAMIND_DAY12_DC_UP	.	.	.	.	PRR36	3.53E-03
GSE6269_FLU_VS_E_COLI_INF_PBMC_UP	.	.	.	.	PTK2B	8.35E-04
GSE22025_TGFB1_VS_TGFB1_AND_PROGESTERONE_TREATED_CD4_TCELL_UP	.	.	.	.	PTK2B	8.35E-04

Figure 4.9. Gene Set Overlap between Pathways Appearing in Top 25 in All Three Scenarios, Immunologic Signatures Data

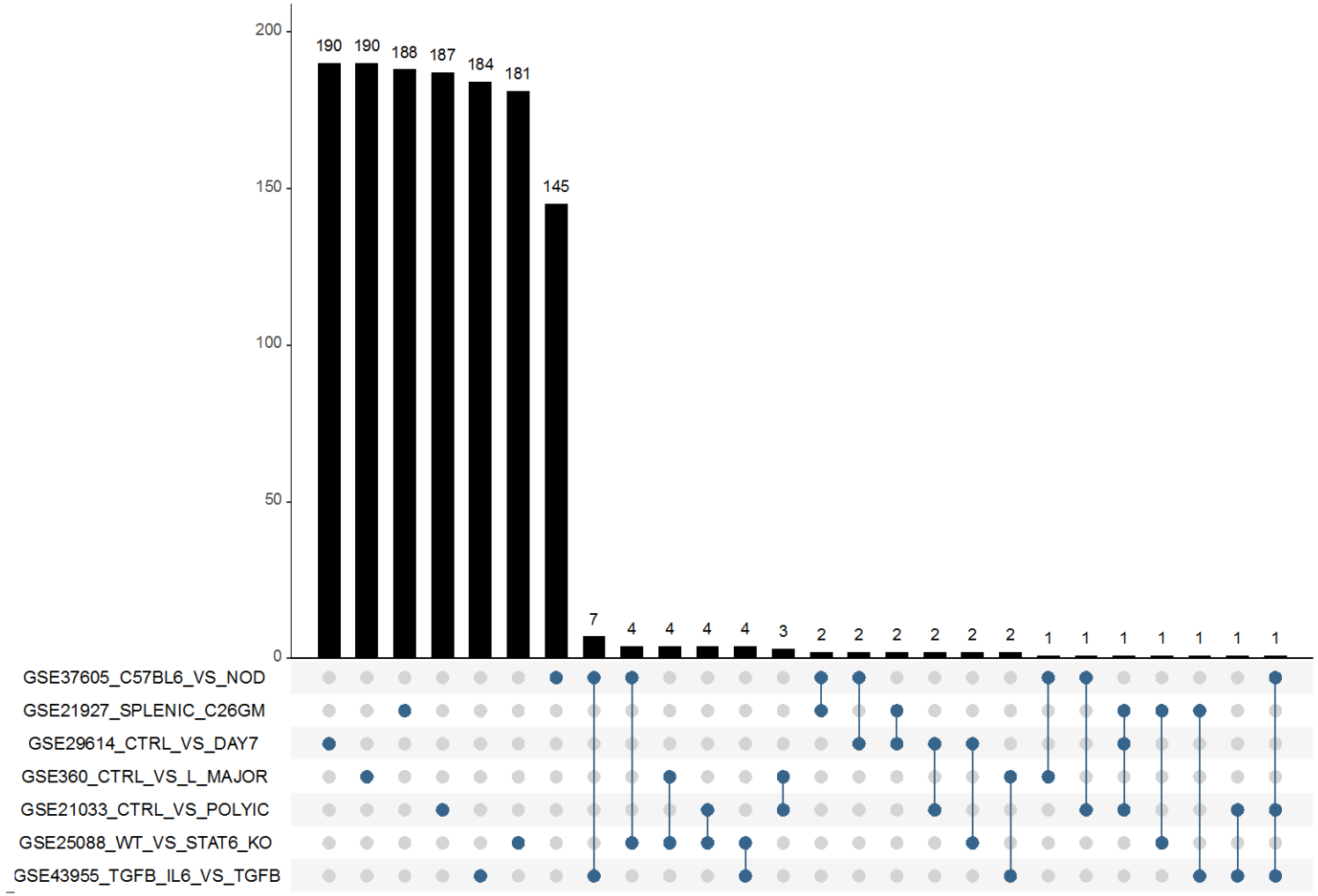


Table 4.17. Pathway-Level Statistics: Top 25 Pathways in 0KB, 20KB and 50KB Scenarios, PASCAL Oncogenic Signatures

PATHWAY	LENGTH	PATHWAY P: 0KB	PATHWAY P: 20KB	PATHWAY P: 50KB
KRAS.DF.V1_UP	193	**1.11E-03**	**1.00E-03**	1.25E-02
CRX_NRL_DN.V1_UP	140	1.12E-02	2.05E-02	7.31E-02
RELA_DN.V1_DN	141	1.75E-02	8.36E-03	2.77E-02
ESC_V6.5_UP_LATE.V1_UP	190	1.95E-02	3.27E-02	5.38E-02
VEGF_A_UP.V1_DN	193	2.62E-02	4.64E-02	4.53E-02
KRAS.600.LUNG.BREAST_UP.V1_UP	288	2.88E-02	2.95E-02	6.94E-02
KRAS.LUNG.BREAST_UP.V1_UP	145	3.41E-02	3.95E-02	7.24E-02
KRAS.300_UP.V1_UP	146	3.69E-02	4.58E-03	**2.48E-03**
HINATA_NFKB_MATRIX	10	4.40E-02	3.14E-02	5.32E-02
KRAS.BREAST_UP.V1_UP	146	4.61E-02	7.36E-02	.
ATF2_S_UP.V1_UP	193	5.51E-02	.	.
IL21_UP.V1_UP	193	5.76E-02	.	.
PKCA_DN.V1_UP	170	5.81E-02	6.76E-02	.
CSR_EARLY_UP.V1_UP	164	5.95E-02	8.34E-02	.
ALK_DN.V1_DN	148	5.99E-02	1.73E-02	2.19E-02
PIGF_UP.V1_DN	194	6.67E-02	.	.
E2F1_UP.V1_DN	193	7.18E-02	4.75E-02	.
CSR_LATE_UP.V1_UP	172	7.23E-02	.	.
GLI1_UP.V1_UP	27	7.51E-02	8.38E-03	9.74E-03
KRAS.50_UP.V1_UP	48	8.06E-02	.	4.52E-02

PRC2_SUZ12_UP.V1_DN	191	8.88E-02	5.59E-02	3.81E-02
NRL_DN.V1_UP	136	9.13E-02	.	.
PDGF_ERK_DN.V1_DN	149	9.61E-02	7.99E-02	.
PTEN_DN.V1_UP	191	1.03E-01	.	.
MTOR_UP.V1_DN	184	1.07E-01	.	.
CRX_DN.V1_UP	136	.	4.14E-02	.
KRAS.600_UP.V1_UP	287	.	5.17E-02	1.50E-02
ESC_V6.5_UP_LATE.V1_DN	186	.	5.58E-02	3.67E-02
LEF1_UP.V1_UP	195	.	7.16E-02	.
BMI1_DN_MEL18_DN.V1_UP	145	.	8.03E-02	3.85E-02
NFE2L2.V2	481	.	8.21E-02	3.49E-02
CAHOY_OLIGODENDROCUTIC	100	.	8.24E-02	3.40E-02
P53_DN.V1_DN	192	.	8.90E-02	.
PRC2_EZH2_UP.V1_DN	194	.	.	3.61E-02
ATM_DN.V1_UP	146	.	.	4.05E-02
MEL18_DN.V1_UP	141	.	.	5.39E-02
SRC_UP.V1_UP	188	.	.	7.17E-02
KRAS.BREAST_UP.V1_DN	145	.	.	7.55E-02
EIF4E_UP	100	.	.	7.89E-02
HOXA9_DN.V1_UP	194	.	.	8.17E-02

Table 4.18. Gene-Level Statistics: Top 25 Pathways in 0KB, 20KB and 50KB Scenarios, PASCAL Oncogenic Signatures

PATHWAY	P0GENE	P0GENEP	P20GENE	P20GENEP	P50GENE	P50GENEP
KRAS.DF.V1_UP	ETFB	1.01E-03	ETFB	6.21E-04	.	1.47E-03
CRX_NRL_DN.V1_UP	ACVR2B	6.23E-03	RNF207	2.58E-03	ACVR2B	6.01E-03
RELA_DN.V1_DN	IL17RB	4.24E-03	GPER1	2.83E-03	GPER1	4.15E-03
ESC_V6.5_UP_LATE.V1_UP	FOXA1	3.16E-04	P4HA2	2.87E-04	P4HA2	4.58E-04
VEGF_A_UP.V1_DN	ITGA2	8.23E-04	ITGA2	1.12E-03	ITGA2	1.16E-03
KRAS.600.LUNG.BREAST_UP.V1_UP	CSF2	3.61E-04	ITGA2	1.12E-03	ITGA2	1.16E-03
KRAS.LUNG.BREAST_UP.V1_UP	CSF2	3.61E-04	ITGA2	1.12E-03	ITGA2	1.16E-03
KRAS.300_UP.V1_UP	ITGA2	8.23E-04	ITGA2	1.12E-03	ITGA2	1.16E-03
HINATA_NFKB_MATRIX	RAC1	2.14E-03	RAC1	3.51E-03	RAC1	4.22E-03
KRAS.BREAST_UP.V1_UP	CSF2	3.61E-04	CSF2	3.00E-03	.	.
ATF2_S_UP.V1_UP	DIRAS3	4.61E-03	.	.	.	.
IL21_UP.V1_UP	DIRAS3	4.61E-03	.	.	.	.
PKCA_DN.V1_UP	MRAP	2.58E-03	MRAP	4.03E-03	.	.
CSR_EARLY_UP.V1_UP	POLE3	3.46E-03	POLE3	4.86E-03	.	.
ALK_DN.V1_DN	FOXA1	3.16E-04	PTK2B	7.41E-04	PTK2B	8.35E-04
PIGF_UP.V1_DN	COL18A1	1.05E-03	.	.	.	.
E2F1_UP.V1_DN	KMT2E	4.75E-03	KMT2E	5.01E-03	.	.
CSR_LATE_UP.V1_UP	NUDCD1	1.13E-03	.	.	.	.
GLI1_UP.V1_UP	C7orf50	5.43E-03	PTH1R	2.26E-03	PTH1R	1.43E-03

KRAS.50_UP.V1_UP	ITGA2	8.23E-04	.	.	ITGA2	1.16E-03
PRC2_SUZ12_UP.V1_DN	ITGA2	8.23E-04	ITGA2	1.12E-03	ITGA2	1.16E-03
NRL_DN.V1_UP	SCN2A	4.20E-03	.	.	.	.
PDGF_ERK_DN.V1_DN	MORC3	4.86E-03	MORC3	7.16E-03	.	.
PTEN_DN.V1_UP	RTP4	2.84E-03	.	.	.	.
MTOR_UP.V1_DN	.	2.00E-02	.	.	.	.
CRX_DN.V1_UP	.	.	ANKRD33	9.89E-04	.	.
KRAS.600_UP.V1_UP	.	.	ITGA2	1.12E-03	ITGA2	1.16E-03
ESC_V6.5_UP_LATE.V1_DN	.	.	FBLN2	8.08E-03	FBLN2	3.48E-03
LEF1_UP.V1_UP	.	.	.	1.86E-04	.	.
BMI1_DN_MEL18_DN.V1_UP	.	.	.	2.12E-04	.	4.28E-04
NFE2L2.V2	.	.	SDS	1.92E-03	HAO1	1.76E-03
CAHOY_OLIGODENDROCUTIC	.	.	P4HA2	2.87E-04	P4HA2	4.58E-04
P53_DN.V1_DN	.	.	SLC6A15	4.72E-03	.	.
PRC2_EZH2_UP.V1_DN	.	.	.	.	ITGA2	1.16E-03
ATM_DN.V1_UP	.	.	.	.	PTK2B	8.35E-04
MEL18_DN.V1_UP	.	.	.	.	.	4.28E-04
SRC_UP.V1_UP	.	.	.	.	FBLN2	3.48E-03
KRAS.BREAST_UP.V1_DN	.	.	.	.	XYLB	1.45E-02
EIF4E_UP	.	.	.	.	IDH1	6.80E-03
HOXA9_DN.V1_UP	.	.	.	.	SLC22A4	7.12E-04

Figure 4.10. Gene Set Overlap between Pathways in Top 25 in All Three Scenarios, PASCAL Oncogenic Signatures

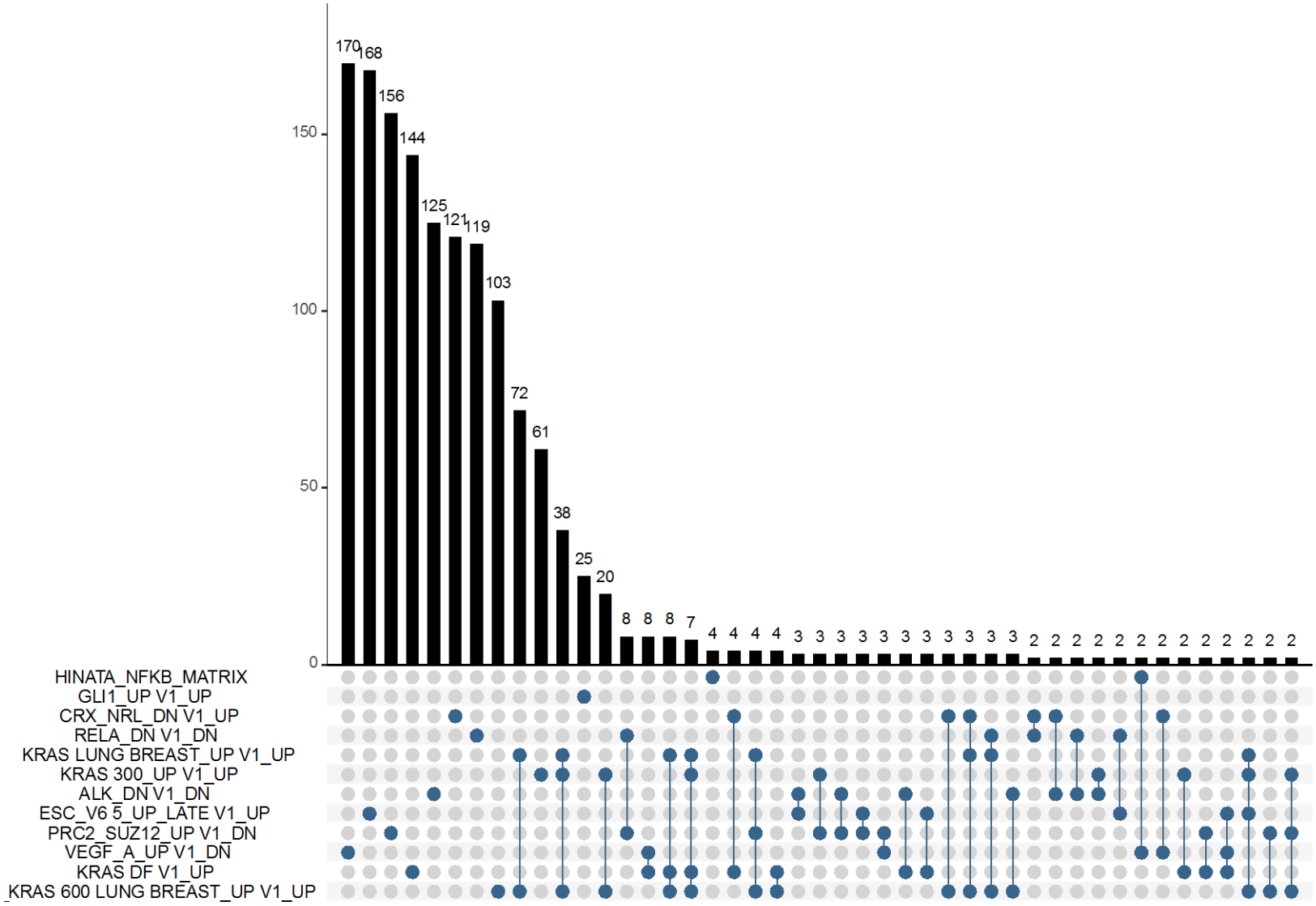


Table 4.19. Pathway-Level Statistics: Top 25 Pathways in 0KB, 20KB and 50KB Scenarios, PASCAL Computational Data

PATHWAY	LENGTH	PATHWAY P: 0KB	PATHWAY P: 20KB	PATHWAY P: 50KB
MODULE_101	10	**1.32E-03**	2.89E-03	1.32E-02
MODULE_202	28	3.41E-03	**1.80E-03**	**8.69E-03**
MODULE_132	14	4.26E-03	8.07E-03	2.18E-02
MODULE_154	75	1.24E-02	3.88E-02	.
MODULE_412	13	1.26E-02	.	.
MODULE_129	219	1.30E-02	1.80E-02	.
MODULE_357	80	1.55E-02	3.75E-02	.
MODULE_297	80	1.56E-02	.	.
MODULE_310	22	1.71E-02	7.45E-03	1.33E-02
GNF2_TTN	26	1.90E-02	.	.
MORF_SKP1A	202	2.03E-02	.	.
MODULE_543	17	2.09E-02	.	.
MODULE_275	16	2.45E-02	.	3.44E-02
MODULE_71	22	2.60E-02	2.31E-02	1.92E-02
MODULE_150	15	3.15E-02	.	.
MODULE_122	141	3.59E-02	3.19E-02	2.33E-02
GNF2_MMP1	32	3.83E-02	.	.
MODULE_183	65	3.90E-02	3.37E-02	2.22E-02
GCM_MAPK10	82	3.92E-02	.	3.48E-02
MODULE_429	14	3.99E-02	.	.

MODULE_340	26	4.10E-02	1.55E-02	1.72E-02
MORF_ANP32B	197	4.37E-02	.	.
MODULE_289	124	4.38E-02	.	.
MODULE_23	565	4.69E-02	.	.
GNF2_CDC27	60	4.92E-02	1.91E-02	2.83E-02
MODULE_131	33	.	7.93E-03	1.09E-02
GCM_RAF1	43	.	1.20E-02	1.02E-02
MODULE_33	384	.	1.50E-02	2.70E-02
MODULE_478	19	.	1.73E-02	2.90E-02
MODULE_426	88	.	2.25E-02	3.29E-03
GNF2_MAP2K3	75	.	3.22E-02	.
GCM_AIP	40	.	3.41E-02	3.56E-02
MODULE_576	110	.	3.42E-02	.
MODULE_503	112	.	3.51E-02	.
GNF2_TNFSF10	30	.	3.73E-02	2.43E-02
MODULE_356	150	.	3.78E-02	.
GNF2_CBFB	30	.	3.78E-02	.
MODULE_147	107	.	3.86E-02	.
MODULE_134	30	.	.	1.85E-02
MODULE_199	57	.	.	2.37E-02
MODULE_375	90	.	.	2.59E-02
MODULE_288	37	.	.	2.61E-02

MODULE_397	121	.	.	3.68E-02
GNF2_IL2RB	47	.	.	3.97E-02
MODULE_256	62	.	.	4.15E-02

Table 4.20. Gene-Level Statistics: Top 25 Pathways in 0KB, 20KB and 50KB Scenarios, PASCAL Computational Data

PATHWAY	TOPGENE0	TOPGENEP0	TOPGENE20	TOPGENEP20	TOPGENE50	TOPGENEP50
MODULE_101	.	1.33E-02	.	1.36E-02	.	1.85E-02
MODULE_202	MYL3	3.74E-04	MYL3	7.17E-04	MYL3	1.05E-03
MODULE_132	.	1.33E-02	.	1.36E-02	.	1.85E-02
MODULE_154	CSF2	3.61E-04	CSF2	3.00E-03	.	.
MODULE_412	ITGA2	8.23E-04	.	.	.	.
MODULE_129	C14orf132	2.02E-04	C14orf132	2.71E-04	.	.
MODULE_357	CSF2	3.61E-04	CSF2	3.00E-03	.	.
MODULE_297	CSF2	3.61E-04	.	.	.	.
MODULE_310	GSTP1	1.90E-02	GSTP1	1.79E-02	IDH1	6.80E-03
GNF2_TTN	NEB	9.90E-03	.	.	.	.
MORF_SKP1A	RAC1	2.14E-03	.	.	.	.
MODULE_543	MRC1	4.82E-03	.	.	.	.
MODULE_275	ITGA2	8.23E-04	.	.	ITGA2	1.16E-03
MODULE_71	SLC22A4	5.11E-04	SLC22A4	4.32E-04	SLC22A4	7.12E-04
MODULE_150	TCEB1	1.22E-02	.	.	.	.
MODULE_122	ITGA2	8.23E-04	MSLN	5.59E-04	MSLN	1.94E-04
GNF2_MMP1	CEMIP	2.85E-02	.	.	.	.
MODULE_183	SF3A3	3.75E-03	SF3A3	7.45E-03	SF3A3	1.25E-02
GCM_MAPK10	PHLPP2	7.99E-03	.	.	PHLPP2	9.35E-03
MODULE_429	RPS21	6.88E-04	.	.	.	.

MODULE_340	ARNTL	1.18E-02	ARNTL	1.58E-02	HAO1	1.76E-03
MORF_ANP32B	.	3.10E-03	.	.	.	.
MODULE_289	DGKI	1.66E-02	.	.	.	.
MODULE_23	ETFB	1.01E-03	.	.	.	.
GNF2_CDC27	SLC22A4	5.11E-04	SLC22A4	4.32E-04	SLC22A4	7.12E-04
MODULE_131	.	.	ITGA2	1.12E-03	ITGA2	1.16E-03
GCM_RAF1	.	.	GPER1	2.83E-03	GPER1	4.15E-03
MODULE_33	.	.	SDS	1.92E-03	PTH1R	1.43E-03
MODULE_478	.	.	P4HA2	2.87E-04	P4HA2	4.58E-04
MODULE_426	.	.	SLC19A1	2.99E-03	ACVRL1	2.30E-03
GNF2_MAP2K3	.	.	SLC22A4	4.32E-04	.	.
GCM_AIP	.	.	GPER1	2.83E-03	GPER1	4.15E-03
MODULE_576	.	.	CDC42SE2	1.13E-02	.	.
MODULE_503	.	.	CDC42SE2	1.13E-02	.	.
GNF2_TNFSF10	.	.	CCR1	2.33E-02	C5AR1	1.49E-02
MODULE_356	.	.	GPER1	2.83E-03	.	.
GNF2_CBFB	.	.	MAPRE1	4.12E-02	.	.
MODULE_147	.	.	GPER1	2.83E-03	.	.
MODULE_134	.	.	.	.	.	4.67E-03
MODULE_199	.	.	.	.	EPHA4	6.02E-03
MODULE_375	.	.	.	.	NPY1R	3.02E-03

MODULE_288	.	.	.	.	.	4.67E-03
MODULE_397	.	.	.	.	.	1.94E-04
GNF2_IL2RB	.	.	.	.	NKG7	6.48E-03
MODULE_256	.	.	.	.	PTH1R	1.43E-03

Figure 4.11. Gene Set Overlap between Pathways Appearing in Top 25 in All Three Scenarios, Computational Data

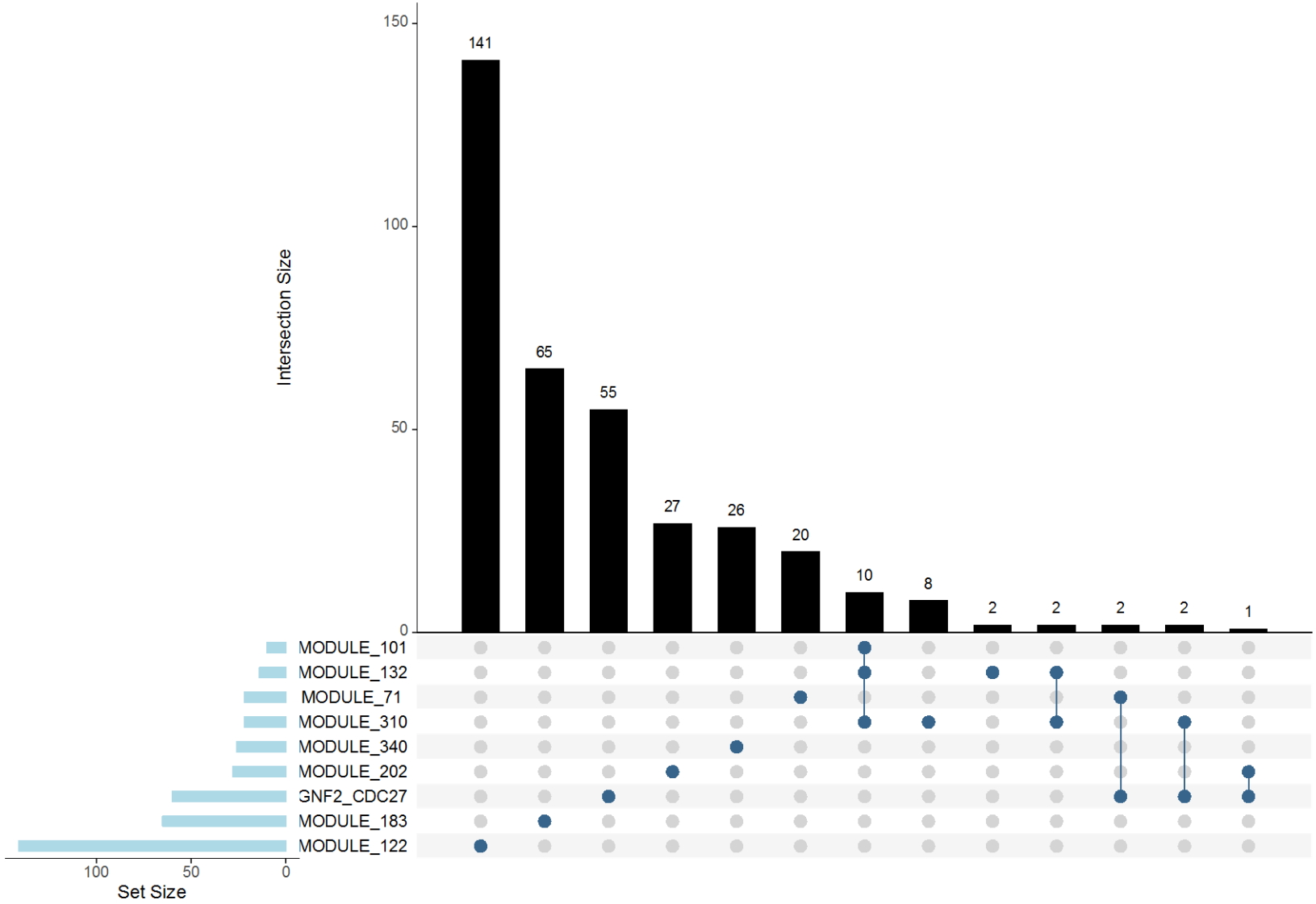


Table 4.21. Pathway-Level Statistics: Top 25 Pathways in 0KB, 20KB and 50KB Scenarios, PASCAL Transcription Factor & miRNA-Binding Motifs Data

RANK	PATHWAY	DESCRIPTION	LENGTH	PATHWAY P: 0KB	PATHWAY P: 20KB	PATHWAY P: 50KB
1	V$TST1_01+	NNKGAATTAVAVTDN promoter motif (matches POU3F1 transcription factor)[209]	264	**1.19E-03**	4.83E-03	4.78E-03
2	V$PAX2_02+	NNNAAASNN promoter motif (PAX2)[210]	260	2.83E-03	4.11E-03	1.15E-02
3	TAAWWATAG_V$RSRFC4_Q2	TAAWWATAG motif (MEF2A)[211]	167	3.33E-03	**1.09E-04**	**1.82E-04**
4	GTATTAT,MIR-369-3P+	Targets of MiRNA GTATTAT,MIR-369-3P	209	3.60E-03	1.77E-03	1.18E-03
5	TGANTCA_V$AP1_C+	TGANTCA motif (JUN)[212]	1123	6.35E-03	5.37E-03	8.45E-03
6	GTGGGTGK_UNKNOWN+	GTGGGTGK promoter motif; no TF match	295	6.46E-03	.	1.79E-02
7	RYTGCNNRGNAAC_V$MIF1_01	RYTGCNNRGNAAC promoter motif (MIF)[213]	88	8.33E-03	1.23E-02	1.42E-02
8	V$NKX22_01	TTAAGTRSTT promoter motif (NKX2-2)[214]	192	9.27E-03	9.31E-03	.
9	V$PBX1_01+	ANCAATCAW promoter motif (PBX1)[215]	254	1.02E-02	.	.
10	TATTATA,MIR-374+	Targets of MicroRNA TATTATA,MIR-374	286	1.18E-02	1.16E-02	.
11	GGTGAAG,MIR-412	Targets of MicroRNA GGTGAAG,MIR-412	63	1.24E-02	1.18E-03	5.08E-03
12	V$CDC5_01+	GATTTAACATAA promoter motif (CDC5L)[216]	258	1.39E-02	1.96E-02	6.88E-03
13	V$AP1_Q4_01+	TGAGTCAN promoter motif (JUN)[217]	263	1.41E-02	.	.
14	GGGTGGRR_V$PAX4_03+	GGGTGGRR promoter motif (PAX4)[218]	1296	1.51E-02	1.95E-02	7.03E-03
15	TGCTGAY_UNKNOWN+	TGCTGAY promoter motif (no match)	540	1.63E-02	7.92E-03	.
16	V$STAT6_02+	NNYTTCCY promoter motif (STAT6)[219]	260	1.70E-02	1.07E-02	.
17	V$TAL1BETAITF2_01+	NNNAACAGATGKTNNN promoter motif (TAL1)[220]	258	1.71E-02	.	1.45E-02

18	V$MAZR_01+	NSGGGGGGGGMCN motif; no TF match	222	1.99E-02	.	.
19	V$RFX1_02+	NNGTNRCNATRGYAACNN motif (RFX1; influences HLAII expression)[221]	280	2.08E-02	1.33E-02	.
20	V$NFKAPPAB65_01+	GGGRATTTCC motif (RELA—NFKB enhancer in B-cells)[222]	239	2.11E-02	.	.
21	V$EGR_Q6+	GTGGGSGCRRS motif (EGR1)[223]	277	2.17E-02	.	1.68E-02
22	GGCNKCCATNK_UNKNOWN	GGCNKCCATNK promoter motif (no match)	120	2.40E-02	.	.
23	V$AP2_Q6_01+	SNNNCCNCAGGCN promoter motif (GTF3A, general transcription factor IIIA)[224]	274	2.50E-02	1.52E-02	.
24	V$NRF1_Q6+	CGCATGCGCR promoter motif (NRF1)[225]	254	2.56E-02	.	.
25	V$POU1F1_Q6+	ATGAATAAWT promoter motif (POU1F1)[226]	240	2.68E-02	.	.
26	V$HLF_01+	GTTACRYAAT promoter motif (HLF, hepatic leukemia factor)[227]	254	.	4.59E-03	1.19E-02
27	ATAAGCT,MIR-21	Targets of MiRNA ATAAGCT,MIR-21	116	.	9.72E-03	.
28	TTANTCA_UNKNOWN+	TTANTCA promoter motif (no TF match)	952	.	1.31E-02	.
29	V$DR1_Q3+	RGGNCAAAGGTCA (NR2F2), chipmunk[228]	257	.	1.34E-02	1.07E-02
30	V$BACH1_01+	NNSATGAGTCATGNT (BACH1)[229]	263	.	1.37E-02	1.15E-02
31	V$E2F1_Q4+	NTTSGCGG promoter motif (E2F1)[230]	244	.	1.43E-02	.
32	V$PAX6_01	NNNNTTCACGCWTGANTKNNN promoter motif (PAX6)[231]	101	.	1.46E-02	1.63E-02
33	V$BACH2_01+	SRTGAGTCANC promoter motif (BACH2, B-cell transcription factor 2)[232]	271	.	1.53E-02	.
34	KTGGYRSGAA_UNKNOWN	KTGGYRSGAA promoter motif (no match)	76	.	1.89E-02	.
35	ATAACCT,MIR-154	Targets of miRNA ATAACCT, miR-154	62	.	1.92E-02	1.61E-02
36	V$RSRFC4_01+	RNKCTATTTWTAGMWN promoter motif	247	.	.	4.82E-03

		(MEF2A)[211]				
37	V$AP4_Q6_01+	RNCAGCTGC promoter motif (FAP4)[205]	257	.	.	4.96E-03
38	V$ATF6_01	TGACGTGG promoter motif (ATF6)[233]	125	.	.	8.07E-03
39	V$RREB1_01+	CCCCAAACMMCCCC promoter motif (RREB1)[234]	209	.	.	8.58E-03
40	V$TAL1BETAE47_01+	NNNAACAGATGKTNNN promoter motif (TAL1)[235]	250	.	.	1.16E-02
41	GGCNRNWCTTYS_UNKNOWN	GGCNRNWCTTYS promoter motif (no match)	85	.	.	1.63E-02
42	V$FXR_Q3	CARGKTSAWTRACC promoter motif (NR1H4)[236]	118	.	.	1.72E-02
43	V$HNF4_DR1_Q3+	TGAMCTTTGNCCN promoter motif (HNF4A)[228]	263	.	.	1.73E-02

Table 4.22. Gene-Level Statistics: Top 25 Pathways in 0KB, 20KB and 50KB Scenarios, PASCAL Transcription Factor & miRNA-Binding Motifs Data

PATHWAY	TOPGENE0	TOPGENEP0	TOPGENE20	TOPGENEP20	TOPGENE50	TOPGENEP50
V$TST1_01	LOC90768	2.24E-03	FOXA1	3.01E-03	STAG1	3.08E-03
V$PAX2_02	CSF2	3.61E-04	STAG1	7.41E-04	PTK2B	8.35E-04
TAAWWATAG_V$RSRFC4_Q2	MYL3	3.74E-04	SOX6	7.17E-04	FILIP1	7.29E-04
GTATTAT,MIR-369-3P	PIKFYVE	1.55E-03	PTK2B	1.89E-03	HAO1	1.76E-03
TGANTCA_V$AP1_C	FOXA1	3.16E-04	FILIP1	1.60E-03	FOXA1	1.71E-03
GTGGGTGK_UNKNOWN	ITGA2	8.23E-04	.	.	ITGA2	1.16E-03
RYTGCNNRGNAAC_V$MIF1_01	FILIP1	7.78E-04	FOXA1	8.58E-04	FILIP1	7.29E-04
V$NKX22_01	MYL3	3.74E-04	SOX6	7.17E-04	.	.
V$PBX1_01	CSF2	3.61E-04	.	.	.	.
TATTATA,MIR-374	PIKFYVE	1.55E-03	FILIP1	1.89E-03	.	.
GGTGAAG,MIR-412	SOX6	4.95E-04	PIKFYVE	4.88E-04	SOX6	4.56E-04
V$CDC5_01	FOXA1	3.16E-04	RNF11	1.60E-03	FOXA1	1.71E-03
V$AP1_Q4_01	SCAMP1	6.87E-03	.	.	.	.
GGGTGGRR_V$PAX4_03	FILIP1	7.78E-04	SQSTM1	8.58E-04	FILIP1	7.29E-04
TGCTGAY_UNKNOWN	GPR157	5.89E-04	MYL3	8.58E-04	.	.
V$STAT6_02	CSF2	3.61E-04	PIKFYVE	3.00E-03	.	.
V$TAL1BETAITF2_01	PHACTR3	1.63E-03	.	.	PDLIM4	3.76E-04
V$MAZR_01	FOXA1	3.16E-04	.	.	.	.
V$RFX1_02	GLRX5_C14orf132	2.19E-04	MGAT4B	7.41E-04	.	.

V$NFKAPPAB65_01	TLX1	5.41E-04	.	.	.	.
V$EGR_Q6	UNC5C	2.73E-03	.	.	UNC5C	3.76E-03
GGCNKCCATNK_UNKNOWN	STAG1	3.19E-03	.	.	.	.
V$AP2_Q6_01	KMT2E	4.75E-03	FILIP1	4.27E-03	.	.
V$NRF1_Q6	MAP2K7_CD320	2.37E-03	.	.	.	.
V$POU1F1_Q6	C12orf42	2.41E-03	.	.	.	.
V$HLF_01	.	.	FOXA1	1.60E-03	FOXA1	1.71E-03
ATAAGCT,MIR-21	.	.	CSF2	4.88E-04	.	.
TTANTCA_UNKNOWN	.	.	FOXA1	5.71E-03	.	.
V$DR1_Q3	.	.	CHTF18	2.71E-04	C14orf132	1.41E-03
V$BACH1_01	.	.	PTK2B	1.60E-03	FOXA1	1.71E-03
V$E2F1_Q4	.	.	FOXA1	9.15E-05	.	.
V$PAX6_01	.	.	EXOG	8.58E-04	FILIP1	7.29E-04
V$BACH2_01	.	.	C14orf132	1.60E-03	.	.
KTGGYRSGAA_UNKNOWN	.	.	FILIP1	7.40E-03	.	.
ATAACCT,MIR-154	.	.	TIMM23	1.76E-02	CUL2	6.89E-03
V$RSRFC4_01	.	.	.	.	FILIP1	7.29E-04
V$AP4_Q6_01	.	.	.	.	PTH1R	1.43E-03
V$ATF6_01	.	.	.	.	STAG1	3.08E-03
V$RREB1_01	.	.	.	.	FILIP1	7.29E-04
V$TAL1BETAE47_01	.	.	.	.	PDLIM4	3.76E-04
GGCNRNWCTTYS_UNKNOWN	.	.	.	.	CANX	8.10E-03

V$FXR_Q3	.	.	.	.	NUDCD1	1.69E-03
V$HNF4_DR1_Q3	.	.	.	.	C14orf132	1.41E-03

Figure 4.12. Gene Set Overlap between Pathways Appearing in Top 25 in All Three Scenarios, PASCAL Transcription Factor & miRNA-Binding Motifs Data

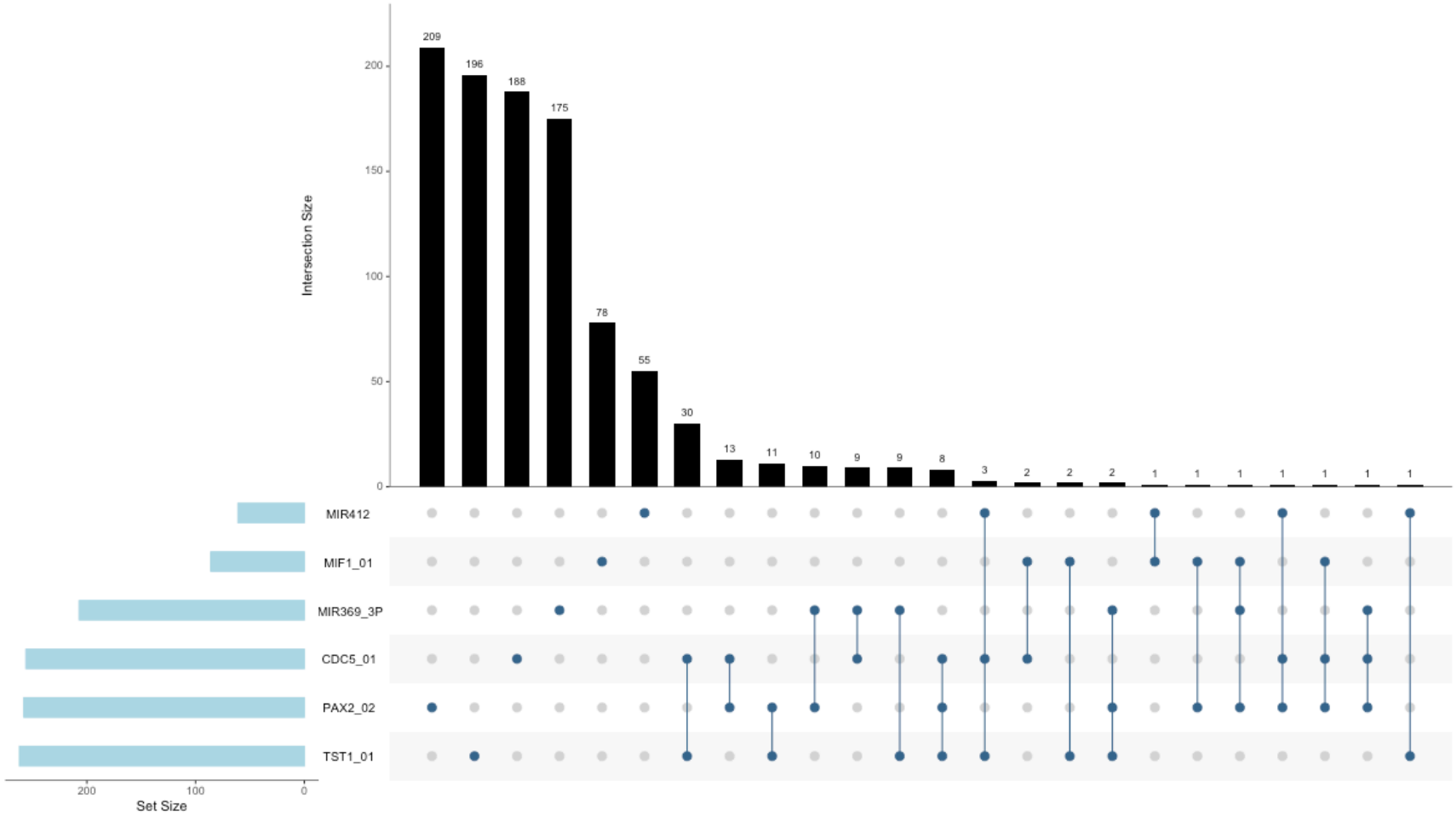


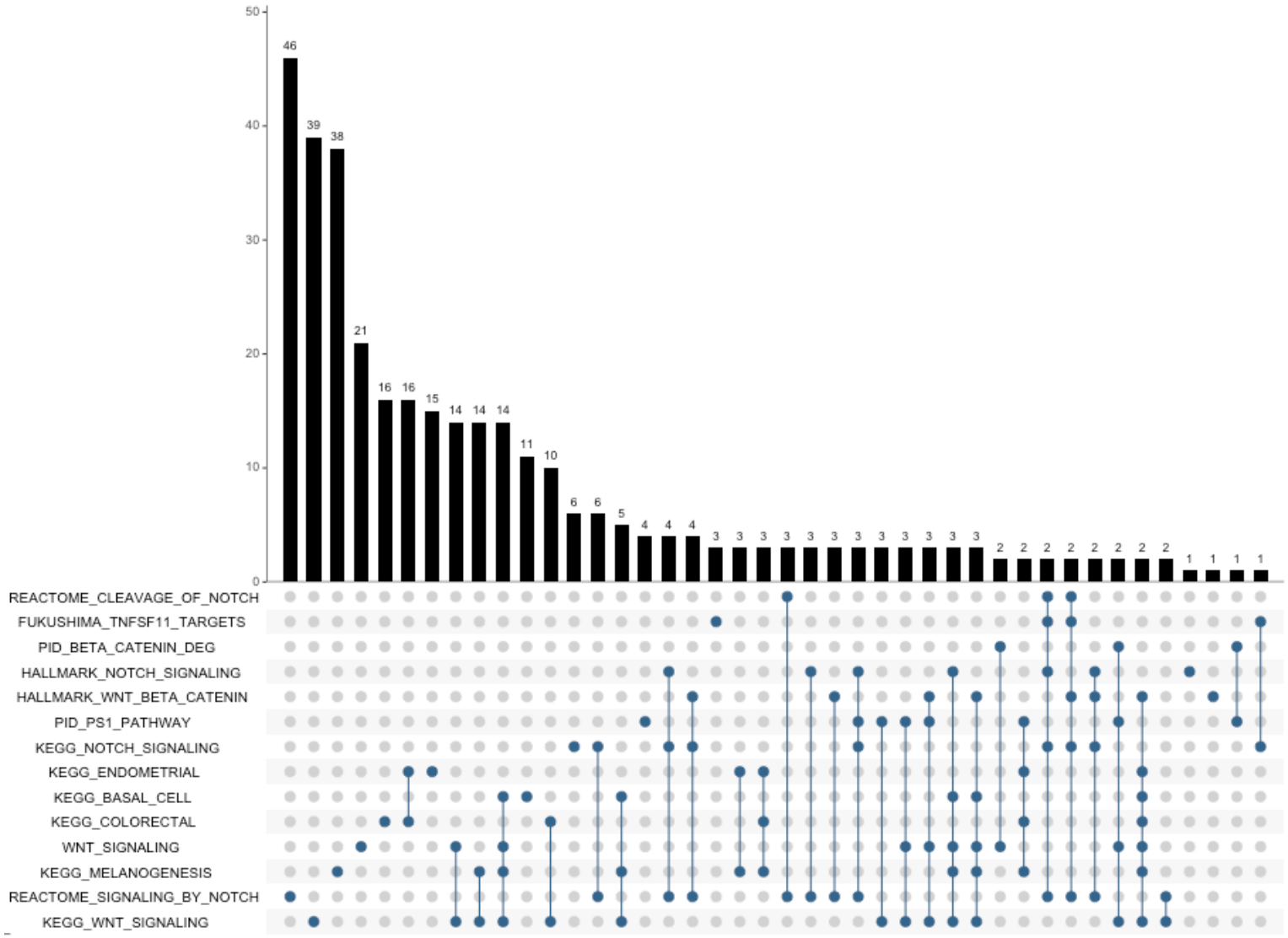
Table 4.23. Pathway-Level Results, PASCAL: Pathways Including Genes Examined in Candidate-Gene Study

PATHWAY	LENGTH	PATHWAY P: 0KB	PATHWAY P: 20KB	PATHWAY P: 50KB
KEGG_MELANOGENESIS	104	1.36E-01	5.15E-01	4.72E-01
FUKUSHIMA_TNFSF11_TARGETS	18	1.50E-01	4.03E-01	3.18E-01
HALLMARK_NOTCH_SIGNALING	34	1.87E-01	4.52E-01	5.80E-01
REACTOME_PROTEOLYTIC_CLEAVAGE_NOTCH	14	3.20E-01	2.88E-01	1.80E-01
KEGG_NOTCH_SIGNALING_PATHWAY	49	4.30E-01	5.33E-01	4.47E-01
REACTOME_SIGNALING_BY_NOTCH	105	4.74E-01	3.36E-01	1.78E-01
KEGG_WNT_SIGNALING_PATHWAY	153	5.21E-01	4.07E-01	5.05E-01
HALLMARK_WNT_BETA_CATENIN_SIGNALING	44	5.61E-01	3.38E-01	4.57E-01
PID_BETA_CATENIN_DEG_PATHWAY	20	6.40E-01	6.33E-01	5.23E-01
KEGG_BASAL_CELL_CARCINOMA	57	6.57E-01	7.57E-01	8.38E-01
WNT_SIGNALING	91	8.99E-01	8.83E-01	8.82E-01
KEGG_COLORECTAL_CANCER	64	9.25E-01	5.66E-01	4.62E-01
KEGG_ENDOMETRIAL_CANCER	54	9.35E-01	7.20E-01	8.54E-01
PID_PS1_PATHWAY	48	9.95E-01	8.57E-01	7.83E-01

Table 4.24. Gene-Level Results, PASCAL: Pathways Including Genes Examined in Candidate-Gene Study

PATHWAY	TOPGENE 0	TOPGENEP 0	TOPGENE2 0	TOPGENEP2 0	TOPGENE5 0	TOPGENEP5 0
KEGG_MELANOGENESIS	CALM3	1.56E-03	PRKACB	2.95E-02	NRAS	2.29E-02
FUKUSHIMA_TNFSF11_TARGETS	ACP1	2.62E-02	ACP1	4.05E-02	DLL4	2.58E-02
HALLMARK_NOTCH_SIGNALING	ARRB1	1.03E-02	ARRB1	1.08E-02	ARRB1	1.35E-02
REACTOME_PROTEOLYTIC_CLEAVAGE NOTCH	NOTCH2	3.40E-02	NOTCH2	4.18E-02	DLL4	2.58E-02
KEGG_NOTCH_SIGNALING_PATHWAY	MAML1	8.50E-03	MAML1	8.07E-03	MAML1	4.75E-03
REACTOME_SIGNALING_BY_NOTCH	HDAC11	4.43E-03	MAML1	8.07E-03	MAML1	4.75E-03
KEGG_WNT_SIGNALING_PATHWAY	RAC1	2.14E-03	RAC1	3.51E-03	RAC1	4.22E-03
HALLMARK_WNT_BETA_CATENIN_SIGNALIN G	HDAC11	4.43E-03	MAML1	8.07E-03	MAML1	4.75E-03
PID_BETA_CATENIN_DEG_PATHWAY	CSNK1D	6.40E-02	CSNK1D	9.42E-02	CSNK1D	9.36E-02
KEGG_BASAL_CELL_CARCINOMA	WNT7B	2.13E-02	FZD9	3.37E-02	WNT11	2.48E-02
WNT_SIGNALING	WNT7B	2.13E-02	MYC	1.37E-02	ACTB	2.41E-02
KEGG_COLORECTAL_CANCER	RAC1	2.14E-03	RAC1	3.51E-03	RAC1	4.22E-03
KEGG_ENDOMETRIAL_CANCER	NRAS	5.49E-02	MYC	1.37E-02	NRAS	2.29E-02
PID_PS1_PATHWAY	ADAM10	7.74E-02	MYC	1.37E-02	FRAT1	3.78E-02

Figure 4.13. Gene Set Overlap between Pathways Appearing in Top 25 in All Three Scenarios, PASCAL Aim 1 Pathways



CHAPTER 5 : OVERARCHING CONCLUSIONS AND PUBLIC HEALTH IMPLICATIONS

5.1. Overarching Conclusions

This study used three methods to investigate genetic risk factors for AIDS-related NHL in case-control samples of 700 and 1,949 HIV-positive individuals in the Multicenter AIDS Cohort Study: a candidate-gene study, a genome-wide association study, and pathway analyses. Motivated by a single biological concern, these methods complemented one another and balanced each other's weaknesses; yielded valuable lessons in the practical application of genomic analysis methods; and also generated suggestive if modest conclusions regarding AIDS-NHL biology and risk.

To review, our candidate-gene study identified a significant inverse association (dominant OR=0.68; 95%CI 0.47-0.99; log-additive OR=0.71, 95%CI 0.51-0.99) between risk of NHL and the SNP rs6815391, found within the 3' UTR of the stem-cell pathway gene REX1/ZPF42. REX1/ZPF42 inhibits p38 MAPK activity; in human T-cells, p38 MAPK plays an important role in HIV replication[107]; in primary human monocytes, activation of p38 MAPK upregulates extracellular HIV Tat-induced transcription of IL-10, an anti-inflammatory cytokine[73,106] elevated serum levels of which have been associated with AIDS-NHL[80]. These findings suggest that observed variation in rs6815391 may upregulate REX1/ZPF42, thus inhibiting p38 MAPK activity, HIV replication, and IL-10 production. Thus there is solid biological plausibility for an inverse association between this rs6815391 and AIDS-NHL risk.

Next, our GWAS identified a genome-wide-significant signal marked by pronounced LD on chromosome 4 (4q33; top SNP rs2195807; p=1.48E-08; white-only p=1.93E-07). Regional plots and functional annotation showed that rs2195807 is in high LD with SNPs falling directly on the

uncharacterized noncoding variant LOC100506122, but is not within a region marked by significant regulatory activity (as measured by DNase hypersensitivity or H3K27Ac histone marks): therefore, the mechanistic implications of variation in rs2195807 remain unclear. Investigation of suggestive associations (p<5E-06) in regions 4p15.1, 2q36.1, 11p15.3 and 12q13.13 indicated that with the exception of 11p15.3, the top SNPs in these regions either fell within gene regions (EPHA4, PCDH7, ANKRD33, ACVRL1) and regions of regulatory activity, or were in LD with SNPs in such regions. However, in contrast to the candidate-gene study, the GWAS did not yield a significant signal with clear biological plausibility.

Finally, we conducted an exhaustive gene-level and pathway-level analysis of our GWAS results, using 13,094 gene sets analyzed in PASCAL, and 6,212 gene sets analyzed in VEGAS2. In doing so, clear patterns emerged: gene sets capturing inflammatory processes and muscle fiber and cytoskeletal integrity appeared at or near the top results in most every scenario.

The prominent place of inflammation-related pathways in our results is consistent with prior knowledge of the biology of NHL. The performance of muscle-related pathways may reflect enrichment for myosin genes: we identified the myosin light-chain gene MYL3 as a top gene (p=3.74E-04 to 1.05E-03, depending on scenario) across pathways. Interestingly, myosins are a key component of the cytoskeleton, which plays an important part in the early stages of B-cell activation[208]. The cytoskeleton is involved in the aggregation of B cell receptor molecules bound to antigen, in the polarization of these complexes ("capping") and in their internalization, which is followed by antigen processing and eventually by antigen presentation in conjunction with MHC class II molecules. Therefore, molecules involved in muscle pathways may well be linked to a central biological activity involved in B cell activation.

The results of each aim thus highlight different aspects of AIDS-NHL biology. The candidate-

gene study suggests involvement of the REX1/ZPF42 SNP rs6815391 in HIV replication and production of anti-inflammatory cytokines; the GWAS points toward an as-yet-uncharacterized locus (LOC100506122); the pathway analyses implicate genes and pathways that may be involved in some of the earliest stages of B-cell activation.

As anticipated, the agnostic GWAS approach did indeed yield results (albeit ambiguous results) different from the candidate-gene study, while pathway analyses yielded results different from either of the two single-SNP approaches (the candidate-gene study and the GWAS). Unfortunately, extending our analysis of genes in aim 1 to the pathway level resulted in no significant or suggestive findings. Of course, this does not prove that these genes play no role in NHL; it simply establishes that given our study design and analytic choices, we were unable to detect meaningful signals.

Pathway analysis thus served as a useful complement to our single-SNP and LD-based analyses in the GWAS. It allowed us to identify key genes and key pathways rather than just one or a few SNPs that might be playing a role in the risk of AIDS-NHL, and that were missed using standard GWAS analysis alone. The repeated appearance of MYL3 and related genes in top pathways calls for revisiting a region of chromosome 3 that had a weakly suggestive regional signal in the GWAS, but that was not pronounced enough to merit follow-up on the basis of single-SNP performance alone. Better-powered GWAS, or studies using targeted resequencing, may find it promising to investigate these regions.

Again, candidate-gene studies, GWAS, and pathway analyses complement one another, and no method is really “best” (as a case in point, our candidate-gene study yielded the least ambiguous association, despite the greater complexity of the GWAS and pathway analyses). Which method is “best” depends on the question of interest, the degree of prior knowledge available to an

investigator, and the stage of investigation. As an example, consider an iterative use of these three methods. First, a GWAS identifies suggestive regions. Second, these data are fed into a pathway analysis, which suggests possible mechanisms for these genes by measuring their strength of association with biological pathways; at the same time, it suggests genes or regions that may have been missed by single-SNP analysis, but that may be promising for a candidate-gene study. Third, these results inform the design and conduct of a candidate-gene study, yielding better power at lower cost. As was the case in this dissertation, each method informs the other, and all three constitute important tools for investigation of genetic risk factors.

5.2. Public Health Implications

Didactically, the contribution of this dissertation is to highlight the value of using multiple tools to investigate genetic risk factors for NHL. It shows how each can help to overcome the shortcomings of the other, and further shows the value of examining multiple scales when investigating genetic risk, as the results obtains for standard single-SNP associations may differ from both gene-level and pathway-level statistics.

Scientifically, this dissertation makes a contribution in two ways. First, it highlights promising biological targets for follow-up, including associations between REX1/ZPF42 and IL-10 expression, the 171MB region on chromosome 4 (4q33), and the role of myosin-related polymorphisms in early B-cell activation and risk of NHL. Second, it may someday constitute one component of a subsequent meta-analysis such as those highlighted in Table 3.13. It would be immodest to claim any major significance for our results, but there may be great value in passing them along for others to build upon. The value of GWAS is in highlighting promising avenues for further investigation, whether through meta-analyses, laboratory work, or candidate-gene studies, and this study is no exception.

Programmatically, if these results were to be replicated with enough confidence, it is conceivable that they could be used to inform screening for NHL risk in HIV-positive patients at diagnosis or at time of ART initiation. Again, results obtained here are not sufficiently unambiguous to justify initiating public health efforts, but it is conceivable—broadly—that individuals with strong markers for inflammation-mediated NHL risk could be put on anti-inflammatories at time of ART initiation, thus reducing risk of subsequent NHL: we know that there is an imperfect correlation between viral load and NHL risk, that HAART alone does not erase risk of AIDS-NHL, and that chronic low-grade inflammation persists even in persons on HAART, increasing risk of NHL[10]. Furthermore, though results are not unambiguous[237], studies have found inverse associations between statins and various types of cancer. AIDS-NHL may be yet another such cancer.

CHAPTER 6 : FUTURE DIRECTIONS

We close with some thoughts on further directions for research, in light of the findings from this dissertation.

6.1. Associations between REX1/ZPF42 and Circulating IL-10

Measurement of serum IL-10 levels in HIV+ MACS participants, and assessment of any associations with variation in REX1/ZPF42 SNP rs6815391, would clarify whether the explanation given for the inverse association between rs6815391 and AIDS-NHL risk above—namely, inhibition of IL-10 expression via inhibition of p38 MAPK activity—is borne out by population-level data.

6.2. Myosin-Related SNPs: Markers of Intestinal Inflammation/Integrity & B-Cell Activation

Pathway analyses highlighted the association of myosin-related genes with NHL. As discussed, myosins play an important role in early B-cell activation. However, myosin is linked to both apoptotic regulation and the integrity of intestinal epithelium; with regard to the latter, it would be interesting to assess potential associations between 1) biomarkers of intestinal inflammation (e.g. zinc) and epithelial integrity and 2) SNPs in the pathways highlighted in VEGAS and PASCAL. At the same time, laboratory experiments to assess the impact of induced variation in MYL3 and related genes on B-cell activation could be carried out.

Doing so would also provide an interesting lesson in the use of pathway analyses to guide biological interpretation: given strong performance of a gene that was not highlighted on the basis of GWAS SNP results, and given two possible mechanisms by which it can act, experimental data would be used to clarify this. This would shed further light on both pathway analysis results and NHL biology.

6.3. Survival Analysis

A natural first step in future research is to assess the association between SNPs in our GWAS data and NHL survival, as opposed to risk of NHL. This could be easily done using PROBABEL, which takes MACH dosage files as input, and runs the R *survival* package under the hood.

6.4. Subtype Analysis

Our sample sizes are already small, but it would also be interesting to look at PCNSL in particular. Many top SNPs and top pathways included variants implicated in nervous system development, e.g. axonal guidance. This could suggest some sort of structural variation in the micro-architecture of CNS tissue that predisposes individuals to PCNSL.

6.5. Rare Variant Analysis

Common variants are those with an MAF > 5%; low-frequency variants are those with an MAF between 5% and 0.5%, and rare variants are those with an MAF < 0.5%. Some methods applicable to one class are not applicable to the others; in particular, the small sample sizes of low-frequency and rare variants requires some modification of standard statistical methods for common variants. A large number of these occurred in our GWAS data, and it could be interesting to look at them in a meta-analysis if data from other cohorts are available.

6.6. Custom Arrays for Regions of Interest

Dense genotyping of regions of interest on chromosomes 2, 4, 11, 12, and potentially 3 and 16, using custom-designed arrays, would improve power over and above that which we were able to obtain using imputation.

6.7. Improved Functional Characterization Using Omic Data

Using proteomics and expression data to better characterize basic NHL etiology, and especially the potential impact of any functional variation identified in this dissertation, is a natural next step. These data would need to be generated *de novo* in the MACS.

It is tempting to suggest laboratory confirmation of biological mechanisms, using manipulation of cell lines or animal models, to buttress the findings of genetic association studies. However, this does not necessarily imply an association with disease. Biological mechanisms identified in the laboratory—whether in cell lines or in animal models—do not necessarily translate into etiological impact in humans, at either the individual or (especially) the population level. Thus while it is tempting to suggest laboratory manipulation to establish mechanism, this in itself would not tell us that the mechanism is responsible for higher rates of NHL. This is so because of intricacies and complications at three levels: gene networks in humans, epigenetic and epigenomic networks in humans, and multiple environmental/non-heritable exposures that may individually and in concert exacerbate or alleviate any impact of the others.

If the goal is, as with most all other exposures in epidemiology, to quantify the impact of a given SNP on an outcome rather than simply point out an association with cancer, then SNPs themselves are actually a low-resolution tool. What is more relevant is under what circumstances and to what extent a given gene is expressed, and whether a given SNP has functional or regulatory effects on these patterns. This would suggest the use of expression data and a shift from analysis of binary variables (wild-type vs. variant) to continuous or at least categorical variables, i.e. mRNA levels. These take place in hugely complex networks. However, the picture grows still more complicated: mRNA levels are poor proxies for the actual level of circulating proteins; extensive post-translational modification means that mRNA levels can be poor indicators of proteins' actual biological effects. Even expression data remain a poor proxy for

actual protein levels, with some of the lowest-concentration proteins in other systems thought to exert hugely disproportionate impact. Though it will be many years before the use of proteomics data is a mainstream practice in epidemiology, and though, for now we find ourselves using blunt tools to analyze artificially simple systems, it is worth investigating the development of sharper tools and their application to more complex systems.

Appendix A. 2008 World Health Organization Classification of Non-Hodgkin Lymphomas

MATURE B-CELL NEOPLASMS, NHL-TYPE

Chronic lymphocytic leukemia/small lymphocytic lymphoma
B-cell prolymphocytic leukemia
Splenic marginal zone lymphoma
Hairy cell leukemia
Splenic lymphoma/leukemia, unclassifiable
    *Splenic diffuse red pulp small B-cell lymphoma**
    *Hairy cell leukemia-variant**
Lymphoplasmacytic lymphoma
    Waldenström macroglobulinemia
Heavy chain diseases
    Alpha heavy chain disease
    Gamma heavy chain disease
    Mu heavy chain disease
Plasma cell myeloma
Solitary plasmacytoma of bone
Extraosseous plasmacytoma
Extranodal marginal zone B-cell lymphoma of mucosa-associated lymphoid tissue (MALT lymphoma)
Nodal marginal zone B-cell lymphoma (MZL)
    Pediatric type nodal MZL
Follicular lymphoma
    Pediatric type follicular lymphoma
Primary cutaneous follicle center lymphoma
Mantle cell lymphoma
Diffuse large B-cell lymphoma (DLBCL), not otherwise specified
    T cell/histiocyte rich large B-cell lymphoma
    DLBCL associated with chronic inflammation
    Epstein-Barr virus (EBV)+ DLBCL of the elderly
Lymphomatoid granulomatosis
Primary mediastinal (thymic) large B-cell lymphoma
Intravascular large B-cell lymphoma
Primary cutaneous DLBCL, leg type
ALK+ large B-cell lymphoma
Plasmablastic lymphoma
Primary effusion lymphoma
Large B-cell lymphoma arising in HHV8-associated multicentric Castleman disease
Burkitt lymphoma
B-cell lymphoma, unclassifiable, with features intermediate between diffuse large B-cell lymphoma and Burkitt lymphoma
B-cell lymphoma, unclassifiable, with features intermediate between diffuse large B-cell lymphoma and classical Hodgkin lymphoma

NK and T-cell neoplasms, NHL-type

T-cell prolymphocytic leukemia
T-cell large granular lymphocytic leukemia
Chronic lymphoproliferative disorder of NK-cells*
Aggressive NK cell leukemia

Systemic EBV+ T-cell lymphoproliferative disease of childhood (associated with chronic active EBV infection)
Hydroa vacciniforme-like lymphoma
Adult T-cell leukemia/lymphoma
Extranodal NK/T cell lymphoma, nasal type
Enteropathy-associated T-cell lymphoma
Hepatosplenic T-cell lymphoma
Subcutaneous panniculitis-like T-cell lymphoma
Mycosis fungoides
Sézary syndrome
Primary cutaneous CD30+ T-cell lymphoproliferative disorder
    Lymphomatoid papulosis
    Primary cutaneous anaplastic large-cell lymphoma
Primary cutaneous aggressive epidermotropic CD8+ cytotoxic T-cell lymphoma*
Primary cutaneous gamma-delta T-cell lymphoma
Primary cutaneous small/medium CD4+ T-cell lymphoma*
Peripheral T-cell lymphoma, not otherwise specified
Angioimmunoblastic T-cell lymphoma
Anaplastic large cell lymphoma (ALCL), ALK+
Anaplastic large cell lymphoma (ALCL), ALK−*

Appendix B. Top 500 SNPs (MAF >0.05) from SNPTEST Logistic Regression Output, Genome-Wide Association Study

rsID	Chrom:Position	A1	A2	All MAF	Cases MAF	Controls MAF	OR (95%CI)	P-value	Typing Status
rs4356576	18:57696044	T	C	0.30	0.15	0.32	0.39 (0.29-0.53)	8.84E-10	Typed
rs2195807	4:171863210	C	T	0.05	0.11	0.05	2.33 (1.60-3.39)	1.48E-08	Imputed
rs80111333	4:171866588	T	G	0.05	0.11	0.05	2.32 (1.60-3.37)	1.50E-08	Imputed
rs13434452	4:171867485	T	A	0.06	0.11	0.05	2.20 (1.52-3.20)	3.08E-08	Imputed
rs10213010	4:171848005	T	G	0.05	0.11	0.05	2.31 (1.59-3.35)	3.30E-08	Imputed
rs10212953	4:171847608	A	G	0.05	0.11	0.05	2.31 (1.59-3.35)	3.34E-08	Imputed
rs28666968	4:171853059	A	G	0.06	0.11	0.05	2.29 (1.58-3.33)	3.77E-08	Imputed
rs10009004	4:171859056	T	C	0.06	0.11	0.05	2.28 (1.57-3.30)	4.33E-08	Typed
rs10049542	4:171878690	T	C	0.06	0.11	0.05	2.26 (1.55-3.28)	4.86E-08	Imputed
rs17056352	4:171890188	G	A	0.06	0.11	0.05	2.26 (1.56-3.28)	5.19E-08	Imputed
rs79663997	4:171890355	G	A	0.06	0.11	0.05	2.26 (1.55-3.28)	5.22E-08	Typed
rs6830294	4:171934820	G	T	0.05	0.10	0.05	2.23 (1.52-3.27)	1.19E-07	Imputed
rs78444688	4:171938263	T	C	0.05	0.10	0.05	2.23 (1.52-3.27)	1.19E-07	Imputed
rs28508193	4:171957833	A	G	0.05	0.10	0.05	2.23 (1.52-3.26)	1.20E-07	Imputed
rs28730459	4:171959720	A	G	0.05	0.10	0.05	2.23 (1.52-3.26)	1.20E-07	Imputed
rs56289978	11:11133374	G	A	0.07	0.12	0.06	2.24 (1.58-3.17)	2.31E-07	Imputed
rs56143914	11:11134564	A	G	0.07	0.12	0.06	2.23 (1.57-3.16)	2.60E-07	Imputed
rs35528558	4:31030698	T	C	0.21	0.12	0.22	0.47 (0.33-0.66)	3.20E-07	Imputed
rs77713994	11:11122587	G	C	0.07	0.12	0.06	2.20 (1.55-3.12)	3.22E-07	Imputed
rs74492376	11:11122577	C	A	0.07	0.12	0.06	2.20 (1.55-3.12)	3.24E-07	Imputed
rs78935380	11:11122379	A	G	0.07	0.12	0.06	2.20 (1.55-3.12)	3.35E-07	Imputed
rs35800293	4:31031051	A	G	0.21	0.12	0.22	0.47 (0.34-0.66)	3.39E-07	Imputed
rs7674265	4:171993094	G	A	0.05	0.10	0.05	2.15 (1.46-3.17)	3.51E-07	Imputed
rs28887169	4:171993171	G	T	0.05	0.10	0.05	2.15 (1.46-3.17)	3.51E-07	Imputed
rs7673717	4:171992835	G	A	0.05	0.10	0.05	2.16 (1.46-3.17)	3.53E-07	Imputed
rs17056430	4:171981632	A	G	0.05	0.10	0.05	2.17 (1.47-3.19)	3.63E-07	Imputed
rs17056428	4:171981534	C	A	0.05	0.10	0.05	2.17 (1.48-3.19)	3.63E-07	Imputed
rs61395681	11:11134584	G	A	0.07	0.12	0.06	2.19 (1.54-3.10)	3.63E-07	Imputed
rs58132943	11:11134766	G	A	0.07	0.12	0.06	2.19 (1.54-3.10)	3.67E-07	Typed
rs10027596	4:171974721	C	A	0.05	0.10	0.05	2.17 (1.47-3.19)	3.77E-07	Imputed
rs10012622	4:171964559	C	G	0.05	0.10	0.05	2.16 (1.47-3.17)	3.91E-07	Imputed
rs9991971	4:171968427	T	G	0.05	0.10	0.05	2.16 (1.47-3.18)	3.91E-07	Imputed
rs28530189	4:171967092	T	C	0.05	0.10	0.05	2.16 (1.47-3.18)	3.96E-07	Imputed
rs6814601	4:171965629	A	G	0.05	0.10	0.05	2.16 (1.47-3.17)	4.06E-07	Imputed
rs10034326	4:171964830	G	A	0.05	0.10	0.05	2.16 (1.47-3.17)	4.06E-07	Imputed
rs10034141	4:171964632	G	A	0.05	0.10	0.05	2.16 (1.47-3.17)	4.08E-07	Typed
rs9996746	4:171974710	T	C	0.05	0.10	0.05	2.12 (1.44-3.12)	4.84E-07	Imputed
rs74793062	11:11136115	A	G	0.07	0.12	0.06	2.18 (1.54-3.10)	5.22E-07	Imputed
rs17433868	2:222221224	T	C	0.25	0.15	0.26	0.51 (0.38-0.69)	5.32E-07	Imputed
rs11680028	2:222220798	A	G	0.25	0.15	0.26	0.51 (0.38-0.69)	5.56E-07	Imputed
rs67012780	2:222230197	C	T	0.35	0.22	0.36	0.51 (0.40-0.67)	7.39E-07	Imputed
rs61792945	4:31027150	A	G	0.21	0.12	0.22	0.49 (0.35-0.68)	7.62E-07	Imputed
rs11676423	2:222222003	T	C	0.35	0.22	0.36	0.51 (0.40-0.67)	9.68E-07	Imputed
rs28837143	4:171831514	C	T	0.06	0.11	0.06	2.03 (1.40-2.92)	9.74E-07	Imputed
rs10201690	2:222221003	C	T	0.35	0.22	0.36	0.52 (0.40-0.67)	9.95E-07	Imputed
rs17434888	2:222227393	T	C	0.35	0.22	0.36	0.51 (0.40-0.67)	1.00E-06	Imputed

rsID	Chrom:Position	A1	A2	All MAF	Cases MAF	Controls MAF	OR (95%CI)	P-value	Typing Status
rs10181647	2:222222356	A	C	0.35	0.22	0.36	0.51 (0.40-0.67)	1.00E-06	Imputed
rs17434868	2:222227216	C	T	0.35	0.22	0.36	0.51 (0.40-0.67)	1.01E-06	Imputed
rs10181883	2:222222534	A	G	0.35	0.22	0.36	0.52 (0.40-0.67)	1.01E-06	Imputed
rs11169939	12:52271467	C	T	0.10	0.18	0.10	2.00 (1.48-2.70)	1.02E-06	Imputed
rs10804297	2:222226640	T	C	0.35	0.22	0.36	0.52 (0.40-0.67)	1.03E-06	Imputed
rs10804296	2:222226628	A	C	0.35	0.22	0.36	0.52 (0.40-0.67)	1.03E-06	Imputed
rs11677083	2:222226566	C	T	0.35	0.22	0.36	0.52 (0.40-0.67)	1.03E-06	Typed
rs7570475	2:222220804	G	C	0.35	0.22	0.36	0.52 (0.40-0.67)	1.03E-06	Imputed
rs76195628	12:52271279	T	C	0.10	0.18	0.10	2.00 (1.48-2.70)	1.03E-06	Imputed
rs12694563	2:222223885	A	G	0.35	0.22	0.36	0.52 (0.40-0.67)	1.03E-06	Imputed
rs11682744	2:222226000	T	C	0.35	0.22	0.36	0.52 (0.40-0.67)	1.04E-06	Imputed
rs7582663	2:222224623	G	A	0.35	0.22	0.36	0.52 (0.40-0.67)	1.04E-06	Imputed
rs1813361	2:222222979	G	C	0.35	0.22	0.36	0.52 (0.40-0.67)	1.04E-06	Imputed
rs12467707	2:222224342	T	C	0.35	0.22	0.36	0.52 (0.40-0.67)	1.05E-06	Imputed
rs1897121	2:222223235	C	A	0.35	0.22	0.36	0.52 (0.40-0.67)	1.05E-06	Typed
rs10932910	2:222228406	C	T	0.35	0.22	0.36	0.52 (0.40-0.67)	1.05E-06	Imputed
rs16825421	2:222226345	T	C	0.35	0.22	0.36	0.52 (0.40-0.67)	1.06E-06	Imputed
rs4280448	2:222223860	T	A	0.35	0.22	0.36	0.52 (0.40-0.67)	1.06E-06	Imputed
rs10211371	2:222223711	C	T	0.35	0.22	0.36	0.52 (0.40-0.67)	1.06E-06	Imputed
rs34650322	2:222227596	C	G	0.35	0.22	0.36	0.52 (0.40-0.67)	1.07E-06	Typed
rs12230130	12:52269931	C	T	0.10	0.17	0.10	1.99 (1.47-2.68)	1.08E-06	Imputed
rs80317637	11:11120614	C	G	0.07	0.12	0.06	2.13 (1.50-3.01)	1.08E-06	Imputed
rs35723210	4:31021145	C	T	0.21	0.12	0.22	0.49 (0.35-0.69)	1.09E-06	Imputed
rs10201802	2:222221156	C	T	0.35	0.22	0.36	0.52 (0.40-0.67)	1.09E-06	Imputed
rs111399147	6:146833727	T	G	0.10	0.17	0.10	1.90 (1.41-2.58)	1.12E-06	Imputed
rs6821882	4:171821091	G	T	0.06	0.11	0.06	2.01 (1.39-2.90)	1.13E-06	Typed
rs12468036	2:222221387	A	G	0.35	0.22	0.36	0.52 (0.40-0.67)	1.14E-06	Typed
rs10167357	2:222221459	T	A	0.35	0.22	0.36	0.52 (0.40-0.67)	1.14E-06	Imputed
rs11891323	2:222229442	C	T	0.35	0.22	0.36	0.52 (0.40-0.67)	1.15E-06	Imputed
rs11692831	2:222221810	G	A	0.35	0.22	0.36	0.52 (0.40-0.67)	1.16E-06	Imputed
rs13021631	2:222221611	G	T	0.35	0.22	0.36	0.52 (0.40-0.67)	1.16E-06	Imputed
rs11901882	2:222229498	A	G	0.35	0.22	0.36	0.52 (0.40-0.67)	1.17E-06	Imputed
rs11901887	2:222229535	A	G	0.35	0.22	0.36	0.52 (0.40-0.67)	1.18E-06	Imputed
rs13022206	2:222229157	G	A	0.35	0.22	0.36	0.52 (0.40-0.67)	1.18E-06	Typed
rs13022081	2:222229301	A	G	0.35	0.22	0.36	0.52 (0.40-0.67)	1.19E-06	Imputed
rs13021697	2:222229181	A	G	0.35	0.22	0.36	0.52 (0.40-0.67)	1.19E-06	Imputed
rs28797244	4:171832521	G	T	0.06	0.12	0.06	2.06 (1.44-2.94)	1.20E-06	Imputed
rs12580654	12:52268547	G	C	0.10	0.17	0.10	1.97 (1.46-2.67)	1.20E-06	Imputed
rs7682233	4:171819122	T	C	0.06	0.11	0.06	2.00 (1.39-2.89)	1.26E-06	Typed
rs77671237	4:172030161	C	T	0.05	0.10	0.05	2.23 (1.53-3.25)	1.60E-06	Imputed
rs1679153	3:138151971	A	G	0.10	0.16	0.10	1.74 (1.27-2.37)	1.60E-06	Imputed
rs79502156	4:172028221	T	C	0.05	0.10	0.05	2.23 (1.53-3.24)	1.62E-06	Imputed
rs67167371	2:222233697	T	C	0.26	0.16	0.27	0.54 (0.40-0.72)	1.64E-06	Typed
rs77048720	4:172029772	G	A	0.06	0.10	0.05	2.21 (1.52-3.23)	1.73E-06	Imputed
rs118147441	15:63606645	C	T	0.07	0.13	0.07	2.15 (1.54-3.02)	1.99E-06	Imputed
rs34737625	2:222254215	C	T	0.27	0.17	0.28	0.52 (0.39-0.70)	2.17E-06	Imputed
rs7746723	6:146832160	A	G	0.11	0.17	0.10	1.87 (1.39-2.53)	2.22E-06	Imputed
rs10183581	2:222234905	G	A	0.34	0.22	0.35	0.53 (0.41-0.70)	2.29E-06	Imputed

rsID	Chrom:Position	A1	A2	All MAF	Cases MAF	Controls MAF	OR (95%CI)	P-value	Typing Status
rs11169944	12:52275199	T	C	0.14	0.21	0.14	1.70 (1.29-2.24)	2.42E-06	Imputed
rs12996402	2:222220208	G	C	0.35	0.23	0.36	0.55 (0.42-0.71)	2.53E-06	Imputed
rs12996064	2:222220200	C	T	0.34	0.23	0.35	0.55 (0.43-0.72)	2.58E-06	Typed
rs7425056	2:222233920	C	T	0.34	0.23	0.35	0.55 (0.42-0.71)	2.97E-06	Imputed
rs73402352	11:9213436	T	C	0.12	0.19	0.11	1.82 (1.36-2.43)	3.11E-06	Imputed
rs11675584	2:222249725	C	T	0.28	0.18	0.29	0.54 (0.41-0.72)	3.52E-06	Imputed
rs11689951	2:222241514	T	C	0.26	0.17	0.27	0.55 (0.41-0.74)	3.64E-06	Imputed
rs6820873	4:31077545	A	C	0.47	0.57	0.46	1.55 (1.24-1.94)	4.02E-06	Imputed
rs147458704	21:39039944	C	T	0.08	0.11	0.07	1.61 (1.13-2.30)	4.39E-06	Imputed
rs144317312	19:2656159	C	A	0.05	0.08	0.05	1.79 (1.18-2.72)	4.80E-06	Imputed
rs7108578	11:17458136	T	C	0.18	0.28	0.17	1.84 (1.43-2.36)	5.03E-06	Imputed
rs7119439	11:17458299	G	A	0.18	0.28	0.17	1.84 (1.43-2.36)	5.04E-06	Imputed
rs61792940	4:31009247	A	T	0.21	0.13	0.22	0.52 (0.38-0.73)	5.11E-06	Imputed
rs9958358	18:20871139	A	G	0.25	0.35	0.24	0.58 (0.45-0.73)	5.40E-06	Imputed
rs72866890	11:17466516	A	G	0.17	0.27	0.16	1.88 (1.46-2.42)	5.48E-06	Typed
rs1808069	18:74924337	C	T	0.12	0.18	0.11	1.74 (1.30-2.34)	5.80E-06	Imputed
rs7670868	4:36085412	A	G	0.42	0.30	0.43	0.58 (0.45-0.73)	6.18E-06	Imputed
rs4896870	6:146827281	G	A	0.11	0.17	0.10	1.85 (1.36-2.50)	6.18E-06	Typed
rs11169942	12:52272699	G	A	0.10	0.17	0.09	1.97 (1.45-2.66)	6.19E-06	Imputed
rs11169943	12:52273248	C	T	0.10	0.17	0.09	1.97 (1.45-2.67)	6.20E-06	Typed
rs1587570	11:16687314	C	G	0.37	0.48	0.36	1.65 (1.32-2.07)	6.81E-06	Imputed
rs7335089	13:103537748	G	A	0.06	0.10	0.06	1.79 (1.22-2.64)	7.55E-06	Imputed
rs2367887	10:26019978	T	C	0.10	0.13	0.09	0.65 (0.46-0.90)	7.68E-06	Typed
rs77492017	11:9263677	G	C	0.09	0.16	0.08	2.03 (1.48-2.77)	8.06E-06	Typed
rs10192024	2:222219639	A	T	0.34	0.23	0.35	0.56 (0.43-0.72)	8.26E-06	Imputed
rs11169945	12:52275509	T	C	0.10	0.17	0.09	2.02 (1.49-2.74)	8.27E-06	Imputed
rs35103713	12:3077006	G	T	0.12	0.18	0.11	1.75 (1.30-2.35)	8.32E-06	Imputed
rs10432451	2:222277159	C	T	0.27	0.17	0.28	0.52 (0.38-0.69)	8.48E-06	Imputed
rs10180965	2:222219687	T	C	0.34	0.23	0.35	0.56 (0.43-0.73)	8.58E-06	Imputed
rs11675679	2:222219759	G	A	0.34	0.23	0.35	0.56 (0.43-0.73)	8.88E-06	Imputed
rs112541859	4:172050979	C	T	0.10	0.15	0.09	1.77 (1.29-2.44)	8.94E-06	Typed
rs61797479	4:36082155	G	C	0.41	0.30	0.42	0.59 (0.46-0.75)	9.63E-06	Imputed
rs10168570	2:222219604	G	A	0.33	0.23	0.34	0.56 (0.43-0.73)	9.63E-06	Imputed
rs74626145	12:52271174	G	T	0.10	0.16	0.09	1.98 (1.46-2.69)	9.64E-06	Imputed
rs74878712	6:146818769	T	C	0.09	0.16	0.09	2.01 (1.47-2.75)	9.80E-06	Imputed
rs7434077	3:48541016	A	C	0.11	0.14	0.10	1.41 (1.02-1.94)	1.03E-05	Imputed
rs7688524	4:36079976	G	A	0.41	0.30	0.42	0.59 (0.46-0.75)	1.06E-05	Typed
rs12651329	4:36079594	T	C	0.41	0.30	0.42	0.59 (0.46-0.75)	1.08E-05	Imputed
rs12230886	12:52268213	T	C	0.10	0.16	0.09	1.95 (1.43-2.66)	1.22E-05	Imputed
rs2505426	10:26007840	T	C	0.10	0.14	0.10	0.71 (0.51-0.98)	1.22E-05	Imputed
rs12226940	12:52268138	A	G	0.10	0.16	0.09	1.95 (1.43-2.65)	1.24E-05	Imputed
rs35363205	2:222193130	C	G	0.34	0.25	0.35	0.63 (0.49-0.82)	1.25E-05	Typed
rs2505423	10:26007210	G	C	0.10	0.14	0.10	0.71 (0.51-0.98)	1.28E-05	Imputed
rs1445481	14:48601997	T	G	0.48	0.40	0.49	0.70 (0.56-0.88)	1.29E-05	Imputed
rs28802045	4:40265273	T	C	0.32	0.21	0.33	1.77 (1.36-2.32)	1.35E-05	Imputed
rs62303905	4:40265086	G	C	0.32	0.21	0.33	1.77 (1.36-2.32)	1.36E-05	Imputed
rs62303904	4:40265010	C	A	0.32	0.21	0.33	1.77 (1.36-2.31)	1.36E-05	Imputed
rs4800464	18:20873586	A	G	0.26	0.35	0.25	0.61 (0.48-0.77)	1.36E-05	Imputed

rsID	Chrom:Position	A1	A2	All MAF	Cases MAF	Controls MAF	OR (95%CI)	P-value	Typing Status
rs9481244	6:112580982	T	C	0.13	0.21	0.12	0.53 (0.40-0.70)	1.37E-05	Imputed
rs2104632	13:37353323	C	G	0.44	0.35	0.45	1.51 (1.20-1.91)	1.41E-05	Typed
rs10892510	11:98797719	T	C	0.45	0.36	0.46	0.65 (0.52-0.82)	1.42E-05	Imputed
rs1551778	11:98798533	A	G	0.45	0.36	0.46	0.65 (0.52-0.82)	1.42E-05	Imputed
rs147970837	7:81005372	G	A	0.06	0.10	0.05	2.16 (1.48-3.15)	1.51E-05	Imputed
rs11616636	13:37352518	A	G	0.44	0.35	0.45	1.51 (1.20-1.90)	1.52E-05	Imputed
rs9576110	13:37352778	A	G	0.44	0.35	0.45	1.51 (1.20-1.90)	1.55E-05	Typed
rs6445619	3:53997720	A	C	0.41	0.51	0.40	0.66 (0.53-0.82)	1.58E-05	Imputed
rs7617075	3:68240106	C	T	0.19	0.27	0.18	1.75 (1.36-2.25)	1.59E-05	Imputed
rs10828861	10:26005615	G	C	0.10	0.14	0.10	0.71 (0.52-0.99)	1.62E-05	Imputed
rs145775145	1:191047737	T	G	0.06	0.12	0.06	2.10 (1.47-3.00)	1.64E-05	Imputed
rs3765334	16:831496	A	G	0.31	0.20	0.32	0.56 (0.42-0.73)	1.67E-05	Imputed
rs111406727	2:180672894	G	A	0.09	0.15	0.09	1.90 (1.39-2.61)	1.70E-05	Typed
rs72800941	16:78078139	A	G	0.12	0.20	0.12	1.94 (1.46-2.57)	1.72E-05	Typed
rs6995247	8:17895014	G	A	0.06	0.08	0.05	1.45 (0.95-2.21)	1.73E-05	Imputed
rs238577	6:115356050	T	C	0.45	0.33	0.47	0.56 (0.44-0.71)	1.77E-05	Imputed
rs116463305	8:8784416	A	G	0.09	0.15	0.09	1.95 (1.42-2.67)	1.83E-05	Typed
rs12139999	1:191065967	C	T	0.07	0.12	0.06	2.01 (1.42-2.85)	1.89E-05	Imputed
rs17105883	14:37463542	G	A	0.32	0.20	0.33	0.51 (0.39-0.67)	1.93E-05	Typed
rs141150420	8:8783258	A	G	0.09	0.15	0.09	1.91 (1.39-2.62)	1.94E-05	Typed
rs12405776	1:242431557	C	T	0.07	0.09	0.06	1.43 (0.96-2.12)	1.97E-05	Imputed
rs11941771	4:40268190	G	C	0.36	0.25	0.37	0.57 (0.44-0.73)	1.98E-05	Imputed
rs11734992	4:40263136	C	T	0.39	0.27	0.40	1.81 (1.42-2.32)	2.03E-05	Imputed
rs34914741	17:64837823	C	T	0.07	0.12	0.06	2.16 (1.52-3.07)	2.04E-05	Imputed
rs4648817	1:2177017	C	T	0.12	0.18	0.11	1.82 (1.36-2.44)	2.06E-05	Imputed
rs111739640	1:242430997	T	C	0.07	0.09	0.06	1.43 (0.96-2.12)	2.08E-05	Imputed
rs360740	3:13153479	A	G	0.42	0.52	0.41	1.52 (1.22-1.89)	2.08E-05	Imputed
rs10023680	4:171704128	A	T	0.17	0.20	0.16	1.32 (1.00-1.75)	2.10E-05	Imputed
rs1105749	2:4294389	A	G	0.24	0.31	0.23	1.47 (1.16-1.88)	2.11E-05	Imputed
rs10003973	4:31032608	C	T	0.27	0.18	0.28	0.57 (0.43-0.76)	2.25E-05	Typed
rs57262748	10:26006532	T	C	0.10	0.13	0.10	0.69 (0.50-0.96)	2.26E-05	Imputed
rs13358835	5:131474170	A	G	0.09	0.15	0.08	1.91 (1.39-2.65)	2.31E-05	Imputed
rs77046703	12:128867908	A	G	0.06	0.09	0.06	1.77 (1.20-2.61)	2.34E-05	Typed
rs74634369	2:180676320	G	A	0.06	0.12	0.06	2.13 (1.49-3.04)	2.37E-05	Imputed
rs6908026	6:41345896	G	A	0.07	0.11	0.06	1.89 (1.31-2.72)	2.43E-05	Imputed
rs76480406	2:180676655	A	G	0.06	0.12	0.06	2.13 (1.49-3.04)	2.44E-05	Imputed
rs61965311	13:70268829	G	C	0.17	0.09	0.18	0.47 (0.32-0.69)	2.45E-05	Imputed
rs10065782	5:131552658	T	C	0.09	0.14	0.08	1.90 (1.38-2.63)	2.47E-05	Imputed
rs1062535	5:52351413	G	A	0.40	0.29	0.41	0.61 (0.48-0.78)	2.51E-05	Imputed
rs2897458	5:52352378	A	T	0.40	0.29	0.41	0.61 (0.48-0.78)	2.51E-05	Typed
rs7619926	3:13152176	G	A	0.40	0.50	0.39	1.54 (1.23-1.92)	2.51E-05	Typed
rs3765329	16:829955	G	C	0.30	0.20	0.31	0.56 (0.43-0.74)	2.55E-05	Imputed
rs2405266	5:131466731	G	A	0.09	0.15	0.08	1.91 (1.38-2.63)	2.56E-05	Imputed
rs59351466	2:180683031	G	A	0.06	0.12	0.06	2.12 (1.48-3.03)	2.63E-05	Imputed
rs7576618	2:117471173	C	T	0.43	0.54	0.42	1.63 (1.30-2.03)	2.67E-05	Imputed
rs133623	22:48700223	C	T	0.21	0.30	0.20	1.67 (1.30-2.13)	2.67E-05	Imputed
rs80010389	4:171721902	G	A	0.05	0.09	0.05	2.07 (1.39-3.08)	2.68E-05	Typed
rs17831624	2:180668803	C	G	0.06	0.11	0.05	2.23 (1.55-3.21)	2.71E-05	Imputed

rsID	Chrom:Position	A1	A2	All MAF	Cases MAF	Controls MAF	OR (95%CI)	P-value	Typing Status
rs58777625	4:36080604	T	C	0.40	0.30	0.41	0.61 (0.48-0.78)	2.71E-05	Imputed
rs55635320	12:3452549	A	T	0.07	0.11	0.06	1.97 (1.38-2.82)	2.73E-05	Imputed
rs79114671	2:180668809	C	A	0.06	0.11	0.05	2.23 (1.54-3.21)	2.73E-05	Imputed
rs10007434	4:171722641	T	G	0.13	0.18	0.12	1.58 (1.18-2.12)	2.76E-05	Imputed
rs55750202	12:3452572	A	C	0.07	0.11	0.06	1.97 (1.37-2.82)	2.86E-05	Imputed
rs75624804	5:131493375	A	G	0.09	0.15	0.08	1.88 (1.36-2.60)	2.88E-05	Imputed
rs9524390	13:94918702	G	A	0.39	0.48	0.38	1.51 (1.21-1.88)	2.94E-05	Imputed
rs10148776	14:37468361	C	A	0.29	0.19	0.30	0.55 (0.41-0.72)	2.95E-05	Typed
rs61730232	8:8860811	C	T	0.07	0.12	0.06	2.14 (1.51-3.03)	2.96E-05	Imputed
rs2095178	9:77452432	G	A	0.16	0.09	0.17	0.50 (0.34-0.73)	2.98E-05	Typed
rs4832921	4:37586357	G	A	0.06	0.08	0.05	1.45 (0.95-2.22)	2.98E-05	Imputed
rs11097019	4:40266986	A	C	0.31	0.21	0.32	1.75 (1.34-2.29)	2.98E-05	Typed
rs10050362	5:131473171	G	C	0.09	0.15	0.08	1.88 (1.36-2.59)	2.99E-05	Imputed
rs73257828	5:131471860	C	T	0.09	0.15	0.08	1.88 (1.36-2.59)	3.00E-05	Imputed
rs10832653	11:16638042	T	A	0.40	0.30	0.41	0.61 (0.48-0.77)	3.11E-05	Imputed
rs11926023	3:104599330	T	C	0.30	0.40	0.29	1.61 (1.28-2.02)	3.12E-05	Imputed
rs6486302	11:16637120	A	G	0.40	0.30	0.41	0.61 (0.48-0.77)	3.13E-05	Imputed
rs7117253	11:16637583	C	T	0.40	0.30	0.41	0.61 (0.48-0.77)	3.14E-05	Typed
rs73193586	3:192113314	A	T	0.12	0.19	0.12	1.78 (1.33-2.37)	3.15E-05	Imputed
rs13049974	21:32810944	T	C	0.10	0.14	0.09	1.61 (1.16-2.23)	3.16E-05	Imputed
rs111994223	7:73993926	C	A	0.15	0.20	0.14	1.53 (1.16-2.03)	3.17E-05	Imputed
rs9880353	3:104586195	G	A	0.30	0.40	0.30	1.60 (1.28-2.01)	3.22E-05	Imputed
rs9845346	3:104598873	C	A	0.30	0.40	0.30	1.60 (1.28-2.01)	3.22E-05	Typed
rs62096511	18:32847291	T	C	0.39	0.50	0.38	1.69 (1.35-2.11)	3.27E-05	Imputed
rs73973764	2:180681085	G	A	0.06	0.12	0.06	2.10 (1.47-3.00)	3.27E-05	Typed
rs73016630	3:13152457	C	T	0.07	0.13	0.06	2.08 (1.47-2.94)	3.28E-05	Imputed
rs12050275	14:37469378	T	C	0.35	0.22	0.36	0.51 (0.40-0.67)	3.30E-05	Imputed
rs13033799	2:117483063	C	T	0.44	0.55	0.43	1.62 (1.29-2.02)	3.33E-05	Imputed
rs61527852	5:131468069	C	T	0.09	0.15	0.08	1.87 (1.35-2.58)	3.34E-05	Typed
rs10787480	10:115019223	T	C	0.35	0.44	0.34	1.50 (1.20-1.88)	3.36E-05	Imputed
rs10229624	7:144964361	A	G	0.33	0.43	0.32	1.56 (1.24-1.95)	3.37E-05	Imputed
rs10741703	11:16630293	A	T	0.40	0.30	0.41	0.61 (0.48-0.77)	3.40E-05	Imputed
rs1436340	3:104597531	G	C	0.30	0.40	0.30	1.60 (1.28-2.01)	3.41E-05	Imputed
rs9990245	3:104597216	T	C	0.30	0.40	0.30	1.60 (1.28-2.01)	3.44E-05	Imputed
rs8019355	14:37470778	C	T	0.35	0.22	0.36	0.52 (0.40-0.67)	3.45E-05	Imputed
rs17520628	15:36037898	A	G	0.14	0.22	0.14	1.74 (1.33-2.29)	3.45E-05	Imputed
rs74692219	8:8878410	T	C	0.07	0.12	0.06	2.12 (1.49-3.01)	3.46E-05	Imputed
rs57094537	4:40261798	T	C	0.38	0.27	0.39	1.74 (1.36-2.23)	3.47E-05	Imputed
rs13025147	2:222268573	T	A	0.25	0.17	0.25	0.59 (0.44-0.79)	3.48E-05	Imputed
rs12510182	4:36078071	A	G	0.40	0.30	0.41	0.62 (0.49-0.79)	3.52E-05	Imputed
rs4757417	11:16615280	T	C	0.41	0.30	0.42	0.60 (0.47-0.76)	3.56E-05	Imputed
rs60250714	4:36078908	G	C	0.40	0.30	0.41	0.62 (0.49-0.79)	3.56E-05	Imputed
rs60432991	13:29121742	C	T	0.12	0.17	0.12	1.56 (1.16-2.10)	3.57E-05	Imputed
rs28373746	3:13151829	A	G	0.40	0.49	0.39	1.52 (1.22-1.90)	3.59E-05	Typed
rs10273655	7:144964544	C	T	0.33	0.43	0.32	1.55 (1.24-1.95)	3.59E-05	Imputed
rs79162715	8:8882704	T	C	0.07	0.12	0.06	2.11 (1.49-3.00)	3.60E-05	Imputed
rs10766333	11:16638484	G	C	0.41	0.30	0.42	0.60 (0.47-0.77)	3.61E-05	Imputed
rs11157835	14:52042824	G	A	0.05	0.10	0.05	2.26 (1.54-3.31)	3.61E-05	Imputed

rsID	Chrom:Position	A1	A2	All MAF	Cases MAF	Controls MAF	OR (95%CI)	P-value	Typing Status
rs58539925	4:36078621	G	C	0.40	0.30	0.41	0.62 (0.49-0.79)	3.61E-05	Typed
rs7568311	2:117468479	C	T	0.45	0.56	0.44	1.62 (1.30-2.03)	3.62E-05	Typed
rs10137636	14:37469969	C	A	0.35	0.22	0.36	0.52 (0.40-0.67)	3.63E-05	Imputed
rs11166678	8:137968725	A	G	0.06	0.09	0.06	1.58 (1.06-2.34)	3.64E-05	Imputed
rs55781930	2:180687099	C	T	0.06	0.12	0.06	2.09 (1.46-2.99)	3.66E-05	Imputed
rs3843548	8:128410346	C	T	0.16	0.08	0.16	2.20 (1.48-3.27)	3.74E-05	Imputed
rs59429726	2:180687519	C	T	0.06	0.12	0.06	2.09 (1.46-2.99)	3.75E-05	Imputed
rs10274888	7:144915784	A	G	0.33	0.43	0.32	1.55 (1.24-1.94)	3.76E-05	Typed
rs11944963	4:130302263	C	T	0.07	0.12	0.07	1.94 (1.36-2.75)	3.77E-05	Typed
rs2262994	5:65100962	A	G	0.34	0.46	0.33	1.70 (1.36-2.12)	3.80E-05	Imputed
rs11123394	2:117460883	T	G	0.45	0.57	0.44	1.63 (1.31-2.04)	3.80E-05	Imputed
rs6931756	6:167622389	A	G	0.25	0.33	0.24	1.55 (1.22-1.96)	3.80E-05	Imputed
rs11123393	2:117460746	A	C	0.45	0.57	0.44	1.63 (1.31-2.04)	3.82E-05	Imputed
rs4446661	7:144924627	A	C	0.33	0.43	0.33	1.55 (1.24-1.95)	3.84E-05	Imputed
rs73973749	2:180669764	C	T	0.06	0.11	0.05	2.18 (1.51-3.14)	3.86E-05	Typed
rs8021342	14:37465301	T	C	0.34	0.22	0.36	0.51 (0.40-0.67)	3.87E-05	Typed
rs2727951	3:2036052	A	G	0.47	0.36	0.48	1.67 (1.33-2.10)	3.88E-05	Imputed
rs74698078	5:131413721	G	A	0.09	0.15	0.08	1.98 (1.44-2.72)	3.90E-05	Imputed
rs10887183	10:85649494	C	G	0.37	0.28	0.38	0.63 (0.50-0.81)	3.91E-05	Imputed
rs1541992	3:104582019	C	A	0.33	0.42	0.32	0.63 (0.50-0.79)	3.92E-05	Imputed
rs113506281	8:8523938	G	A	0.08	0.13	0.08	1.84 (1.32-2.58)	3.93E-05	Imputed
rs7605984	2:222218395	A	C	0.31	0.21	0.32	0.57 (0.44-0.75)	3.94E-05	Imputed
rs6436255	2:222218089	C	T	0.31	0.21	0.32	0.57 (0.44-0.75)	3.94E-05	Typed
rs6436256	2:222218125	G	A	0.31	0.21	0.32	0.57 (0.44-0.75)	3.95E-05	Imputed
rs56402502	5:172041550	C	T	0.26	0.17	0.27	0.56 (0.42-0.75)	3.95E-05	Imputed
rs150370340	19:14347044	C	T	0.10	0.16	0.10	1.68 (1.23-2.29)	3.95E-05	Imputed
rs75265060	3:35104066	T	C	0.08	0.14	0.08	1.92 (1.38-2.66)	3.96E-05	Imputed
rs7340379	2:222217951	T	C	0.31	0.21	0.32	0.57 (0.44-0.75)	3.97E-05	Imputed
rs12270739	11:36432319	G	A	0.11	0.17	0.10	1.73 (1.28-2.35)	3.98E-05	Imputed
rs16991904	4:36084658	G	A	0.36	0.27	0.37	0.65 (0.51-0.83)	4.01E-05	Imputed
rs7070276	10:53983384	C	T	0.12	0.15	0.11	1.37 (1.01-1.88)	4.01E-05	Imputed
rs3212509	5:52348256	T	C	0.39	0.29	0.40	0.62 (0.49-0.79)	4.04E-05	Typed
rs3212508	5:52348124	T	G	0.39	0.29	0.40	0.62 (0.49-0.79)	4.05E-05	Imputed
rs2504188	10:26003576	A	G	0.11	0.14	0.11	0.77 (0.55-1.06)	4.05E-05	Typed
rs10432541	2:222216743	T	A	0.31	0.21	0.32	0.57 (0.44-0.75)	4.05E-05	Imputed
rs1126643	5:52347369	C	T	0.39	0.29	0.40	0.62 (0.49-0.79)	4.06E-05	Imputed
rs1902871	4:163776745	C	G	0.12	0.20	0.12	1.82 (1.37-2.42)	4.06E-05	Imputed
rs6968939	7:144928088	A	G	0.32	0.42	0.32	1.56 (1.25-1.96)	4.07E-05	Imputed
rs6816376	4:190122349	G	A	0.27	0.33	0.27	0.72 (0.57-0.91)	4.10E-05	Imputed
rs79956758	4:189385270	G	A	0.16	0.24	0.15	1.75 (1.35-2.28)	4.12E-05	Imputed
rs26325	5:115737646	C	A	0.21	0.12	0.22	2.07 (1.49-2.89)	4.12E-05	Imputed
rs11042776	11:10411218	A	G	0.28	0.38	0.27	1.66 (1.32-2.09)	4.14E-05	Imputed
rs1022309	2:117477692	T	A	0.44	0.55	0.43	1.61 (1.29-2.01)	4.15E-05	Typed
rs9856543	3:104592124	T	C	0.30	0.40	0.30	1.59 (1.27-2.00)	4.15E-05	Imputed
rs7810304	7:144945903	G	A	0.33	0.42	0.32	1.58 (1.26-1.97)	4.18E-05	Imputed
rs149833345	14:20453934	G	A	0.06	0.09	0.06	1.70 (1.15-2.52)	4.18E-05	Typed
rs4605176	16:78030251	A	G	0.11	0.18	0.10	1.96 (1.46-2.64)	4.20E-05	Imputed
rs60334489	4:36086609	T	C	0.36	0.27	0.36	0.65 (0.51-0.83)	4.21E-05	Imputed

rsID	Chrom:Position	A1	A2	All MAF	Cases MAF	Controls MAF	OR (95%CI)	P-value	Typing Status
rs12984031	19:14352149	C	T	0.09	0.14	0.08	1.75 (1.26-2.43)	4.22E-05	Imputed
rs10814700	9:38478011	T	G	0.37	0.27	0.38	0.60 (0.47-0.77)	4.22E-05	Imputed
rs76904717	8:8867523	G	A	0.07	0.12	0.06	2.11 (1.49-2.99)	4.22E-05	Imputed
rs11725198	4:163791500	T	C	0.13	0.20	0.12	1.81 (1.36-2.41)	4.23E-05	Imputed
rs2075615	5:52345220	G	A	0.39	0.29	0.40	0.62 (0.49-0.79)	4.23E-05	Imputed
rs76202981	4:171726058	G	A	0.05	0.09	0.05	2.01 (1.35-2.98)	4.26E-05	Typed
rs2564038	2:6028174	T	C	0.07	0.11	0.06	1.99 (1.39-2.85)	4.27E-05	Typed
rs4691874	4:163793257	C	T	0.13	0.20	0.13	1.73 (1.30-2.29)	4.27E-05	Imputed
rs835158	5:14873254	C	G	0.43	0.54	0.42	1.61 (1.29-2.01)	4.31E-05	Imputed
rs9826514	3:104590454	C	T	0.30	0.40	0.30	1.59 (1.27-2.00)	4.38E-05	Imputed
rs75506037	4:171731676	A	G	0.06	0.10	0.05	1.89 (1.28-2.78)	4.39E-05	Imputed
rs1585635	8:54396558	G	A	0.26	0.34	0.25	0.65 (0.52-0.83)	4.44E-05	Imputed
rs2052941	2:222276477	A	G	0.26	0.17	0.27	0.58 (0.43-0.77)	4.51E-05	Imputed
rs12643250	4:40262582	G	A	0.31	0.22	0.32	1.67 (1.28-2.18)	4.53E-05	Imputed
rs143257179	3:195845759	C	G	0.06	0.10	0.05	2.02 (1.38-2.94)	4.59E-05	Imputed
rs10432450	2:222277000	G	A	0.25	0.17	0.26	0.59 (0.44-0.79)	4.60E-05	Typed
rs16958227	15:74073127	G	C	0.06	0.10	0.06	1.78 (1.22-2.59)	4.60E-05	Imputed
rs11144107	9:77447804	T	C	0.16	0.09	0.17	0.51 (0.35-0.74)	4.62E-05	Typed
rs3778542	6:1870348	C	T	0.06	0.11	0.06	2.01 (1.40-2.88)	4.68E-05	Imputed
rs10741705	11:16670985	A	G	0.40	0.30	0.41	0.61 (0.48-0.78)	4.72E-05	Imputed
rs10919048	1:162248465	G	A	0.35	0.41	0.34	1.34 (1.07-1.68)	4.72E-05	Imputed
rs11248950	16:853728	A	C	0.38	0.28	0.40	0.59 (0.46-0.76)	4.74E-05	Typed
rs59777414	2:180685929	T	C	0.09	0.15	0.09	1.85 (1.34-2.54)	4.82E-05	Imputed
rs17134034	6:1511560	C	T	0.24	0.33	0.23	1.63 (1.29-2.07)	4.87E-05	Imputed
rs72696754	4:163790570	C	T	0.13	0.20	0.13	1.72 (1.30-2.28)	4.87E-05	Imputed
rs9996120	4:16677331	T	A	0.07	0.11	0.07	1.71 (1.20-2.44)	4.94E-05	Imputed
rs61930461	12:113821098	C	T	0.05	0.08	0.05	1.59 (1.05-2.41)	4.96E-05	Imputed
rs10131507	14:37466492	T	C	0.34	0.22	0.35	0.52 (0.40-0.67)	4.98E-05	Typed
rs80268011	4:171735386	C	T	0.06	0.09	0.05	1.88 (1.27-2.77)	5.03E-05	Typed
rs984966	5:52368922	T	A	0.39	0.30	0.40	0.63 (0.49-0.80)	5.06E-05	Imputed
rs11979826	7:144942117	A	T	0.33	0.42	0.32	1.56 (1.25-1.96)	5.07E-05	Imputed
rs10802174	1:116785162	C	T	0.10	0.14	0.10	1.43 (1.03-1.98)	5.09E-05	Imputed
rs77971395	12:5698971	G	A	0.07	0.13	0.07	1.97 (1.40-2.78)	5.15E-05	Imputed
rs1247449	10:29938005	T	C	0.05	0.10	0.05	2.24 (1.53-3.30)	5.16E-05	Imputed
rs28697231	3:104585106	G	A	0.30	0.40	0.30	1.58 (1.26-1.99)	5.16E-05	Typed
rs17831917	2:180691091	T	C	0.07	0.12	0.06	2.05 (1.43-2.93)	5.17E-05	Typed
rs10832665	11:16670394	A	C	0.41	0.31	0.42	0.61 (0.48-0.77)	5.21E-05	Typed
rs7116271	11:16657871	T	C	0.41	0.31	0.42	0.61 (0.48-0.77)	5.24E-05	Imputed
rs61831419	10:810633	T	C	0.16	0.24	0.15	1.71 (1.31-2.23)	5.35E-05	Imputed
rs2600051	3:2034177	T	C	0.47	0.36	0.48	1.65 (1.31-2.08)	5.40E-05	Typed
rs1436354	3:104584326	C	T	0.30	0.40	0.30	1.58 (1.26-1.99)	5.42E-05	Imputed
rs1223271	20:13296912	G	A	0.13	0.20	0.12	1.78 (1.34-2.36)	5.43E-05	Imputed
rs7428949	3:48545777	A	G	0.12	0.14	0.11	1.27 (0.92-1.75)	5.44E-05	Typed
rs6936429	6:167622419	T	C	0.25	0.33	0.24	1.54 (1.22-1.96)	5.61E-05	Typed
rs11732411	4:163762613	G	A	0.13	0.20	0.13	1.72 (1.30-2.28)	5.62E-05	Imputed
rs2707782	5:65089779	T	G	0.34	0.46	0.33	1.67 (1.33-2.08)	5.64E-05	Typed
rs56094842	4:163763946	G	A	0.13	0.20	0.13	1.72 (1.30-2.28)	5.65E-05	Imputed
rs10924014	1:116785299	T	C	0.11	0.14	0.10	1.44 (1.04-1.98)	5.67E-05	Imputed

rsID	Chrom:Position	A1	A2	All MAF	Cases MAF	Controls MAF	OR (95%CI)	P-value	Typing Status
rs12123812	1:190953177	C	G	0.06	0.11	0.06	2.03 (1.41-2.92)	5.73E-05	Typed
rs11723778	4:163765735	T	C	0.13	0.20	0.13	1.72 (1.30-2.28)	5.75E-05	Imputed
rs7434107	3:48541182	A	G	0.12	0.14	0.11	1.27 (0.92-1.75)	5.77E-05	Typed
rs56802651	15:74070684	T	A	0.05	0.09	0.05	2.03 (1.36-3.02)	5.82E-05	Imputed
rs117034711	16:78116370	A	C	0.12	0.20	0.11	1.90 (1.43-2.53)	5.86E-05	Imputed
rs58246662	12:3453548	A	C	0.08	0.12	0.08	1.73 (1.23-2.44)	5.87E-05	Typed
rs2729314	3:2035993	G	A	0.48	0.36	0.49	1.67 (1.32-2.10)	5.88E-05	Imputed
rs72696749	4:163769021	G	T	0.13	0.20	0.13	1.72 (1.30-2.28)	5.93E-05	Typed
rs11715364	3:119056517	T	C	0.06	0.10	0.05	0.48 (0.33-0.70)	5.96E-05	Imputed
rs13075249	3:68209602	T	C	0.18	0.27	0.18	1.73 (1.34-2.23)	5.96E-05	Imputed
rs2272896	16:836992	G	C	0.24	0.16	0.25	0.56 (0.42-0.76)	5.98E-05	Typed
rs17833070	3:13150039	T	C	0.08	0.13	0.08	1.92 (1.37-2.68)	6.01E-05	Imputed
rs7935756	11:16645234	A	G	0.41	0.30	0.42	0.61 (0.48-0.78)	6.06E-05	Imputed
rs2375531	9:1433652	A	G	0.44	0.33	0.45	0.62 (0.49-0.79)	6.06E-05	Typed
rs2727950	3:2035726	C	T	0.48	0.36	0.49	1.67 (1.32-2.10)	6.07E-05	Typed
rs10112780	8:13467552	A	T	0.43	0.55	0.42	0.59 (0.48-0.74)	6.12E-05	Imputed
rs7546515	1:116783576	A	G	0.11	0.14	0.10	1.43 (1.04-1.98)	6.12E-05	Imputed
rs2727948	3:2033717	A	G	0.47	0.36	0.48	1.65 (1.31-2.07)	6.14E-05	Imputed
rs12474090	2:222213975	C	T	0.31	0.21	0.31	0.59 (0.45-0.77)	6.16E-05	Typed
rs13062176	3:2036373	A	G	0.49	0.37	0.50	1.69 (1.34-2.12)	6.22E-05	Imputed
rs2332060	1:181133024	T	G	0.10	0.03	0.10	0.27 (0.14-0.51)	6.25E-05	Imputed
rs471511	16:5485487	C	G	0.08	0.10	0.07	1.43 (0.99-2.07)	6.26E-05	Imputed
rs10130430	14:37448518	T	C	0.34	0.22	0.35	0.54 (0.41-0.70)	6.28E-05	Imputed
rs74680894	12:3077182	T	C	0.10	0.16	0.09	1.90 (1.39-2.58)	6.28E-05	Typed
rs8009178	14:37528316	G	A	0.41	0.29	0.42	0.56 (0.44-0.71)	6.31E-05	Typed
rs2383097	9:19499373	C	G	0.25	0.33	0.24	1.61 (1.27-2.04)	6.31E-05	Imputed
rs56332741	4:163772481	G	T	0.13	0.20	0.13	1.71 (1.29-2.27)	6.33E-05	Imputed
rs12492271	3:48536456	C	T	0.12	0.14	0.12	0.79 (0.57-1.09)	6.33E-05	Imputed
rs113416252	11:16651578	A	G	0.39	0.50	0.38	1.64 (1.31-2.05)	6.37E-05	Imputed
rs76497010	12:1202630	C	T	0.05	0.09	0.05	1.96 (1.33-2.90)	6.37E-05	Imputed
rs74023533	15:74071251	A	G	0.05	0.09	0.05	2.03 (1.36-3.03)	6.40E-05	Imputed
rs6947380	7:144955190	T	C	0.33	0.42	0.32	1.56 (1.24-1.95)	6.41E-05	Imputed
rs728314	15:74070563	C	T	0.05	0.09	0.05	2.03 (1.36-3.02)	6.46E-05	Imputed
rs10832657	11:16656245	A	T	0.39	0.50	0.38	1.64 (1.31-2.05)	6.48E-05	Imputed
rs72740410	1:191115099	C	T	0.07	0.12	0.06	1.94 (1.36-2.78)	6.51E-05	Typed
rs72759417	16:858094	A	G	0.28	0.19	0.29	0.57 (0.43-0.76)	6.54E-05	Imputed
rs13103812	4:36078004	C	G	0.47	0.37	0.48	1.59 (1.26-2.00)	6.61E-05	Imputed
rs7576476	2:4296217	A	G	0.27	0.34	0.27	1.39 (1.10-1.76)	6.62E-05	Imputed
rs17662983	4:171661505	C	T	0.07	0.11	0.07	1.55 (1.07-2.23)	6.63E-05	Imputed
rs117722264	11:23189583	C	T	0.05	0.09	0.05	2.01 (1.36-2.98)	6.64E-05	Imputed
rs11850139	14:37451848	T	C	0.34	0.23	0.35	0.53 (0.41-0.69)	6.64E-05	Imputed
rs74023536	15:74072839	A	G	0.05	0.09	0.05	2.04 (1.37-3.04)	6.65E-05	Imputed
rs77308437	4:171726072	G	C	0.06	0.10	0.05	1.85 (1.25-2.72)	6.68E-05	Imputed
rs113522027	3:153020960	G	C	0.09	0.14	0.09	1.79 (1.30-2.48)	6.71E-05	Imputed
rs9930069	16:5070502	T	A	0.46	0.54	0.45	0.70 (0.56-0.88)	6.73E-05	Imputed
rs6811953	4:171764374	T	C	0.05	0.09	0.05	1.90 (1.28-2.83)	6.74E-05	Imputed
rs72734897	1:190756974	T	C	0.05	0.10	0.05	2.09 (1.42-3.08)	6.80E-05	Imputed
rs7269418	20:40550983	A	G	0.07	0.11	0.06	1.82 (1.27-2.62)	6.84E-05	Typed

rsID	Chrom:Position	A1	A2	All MAF	Cases MAF	Controls MAF	OR (95%CI)	P-value	Typing Status
rs2158724	7:101245666	C	G	0.43	0.33	0.43	1.56 (1.23-1.97)	6.85E-05	Imputed
rs34051515	12:3076417	A	G	0.10	0.16	0.09	1.88 (1.38-2.56)	6.88E-05	Imputed
rs11601816	11:16663109	A	G	0.39	0.50	0.38	1.64 (1.31-2.04)	6.90E-05	Imputed
rs35444392	12:3076401	C	A	0.10	0.16	0.09	1.87 (1.37-2.55)	6.94E-05	Imputed
rs1822233	9:32904920	C	G	0.31	0.40	0.30	0.66 (0.52-0.83)	6.94E-05	Imputed
rs10758171	9:32905899	C	G	0.31	0.40	0.30	0.66 (0.52-0.83)	6.97E-05	Typed
rs1470217	9:32903868	G	A	0.31	0.39	0.30	0.66 (0.52-0.83)	6.97E-05	Imputed
rs7251857	19:7543113	T	C	0.17	0.21	0.17	1.37 (1.05-1.80)	7.07E-05	Imputed
rs2600052	3:2035038	G	C	0.48	0.36	0.49	1.66 (1.32-2.08)	7.08E-05	Imputed
rs1946971	9:32905375	G	C	0.31	0.40	0.30	0.66 (0.52-0.83)	7.09E-05	Imputed
rs4668813	2:6040560	A	G	0.08	0.12	0.08	0.60 (0.42-0.84)	7.09E-05	Imputed
rs10119340	9:32905294	C	T	0.31	0.40	0.30	0.66 (0.52-0.83)	7.09E-05	Imputed
rs12072166	1:116780622	C	T	0.11	0.14	0.11	1.38 (1.00-1.89)	7.11E-05	Imputed
rs1348254	9:32905302	A	G	0.31	0.40	0.30	0.66 (0.52-0.83)	7.11E-05	Typed
rs12133635	1:190929221	A	G	0.06	0.11	0.06	2.02 (1.40-2.91)	7.12E-05	Imputed
rs61831428	10:824880	C	A	0.15	0.23	0.14	1.74 (1.33-2.27)	7.12E-05	Imputed
rs11706604	3:123192800	T	C	0.05	0.07	0.05	0.73 (0.47-1.12)	7.17E-05	Imputed
rs4528684	19:14351574	C	T	0.10	0.14	0.09	1.64 (1.19-2.26)	7.22E-05	Imputed
rs4674584	2:222211490	A	G	0.30	0.21	0.31	0.59 (0.45-0.77)	7.22E-05	Imputed
rs4299433	3:192095111	C	T	0.14	0.21	0.13	1.75 (1.33-2.32)	7.24E-05	Imputed
rs78150878	12:3078561	A	G	0.10	0.16	0.09	1.90 (1.39-2.59)	7.26E-05	Imputed
rs11768813	7:156619075	A	G	0.06	0.10	0.05	2.00 (1.36-2.95)	7.28E-05	Typed
rs9488527	6:115382910	A	G	0.38	0.28	0.39	0.59 (0.46-0.75)	7.32E-05	Imputed
rs11967518	6:112599997	A	T	0.38	0.48	0.37	1.62 (1.30-2.03)	7.32E-05	Imputed
rs1437084	12:17773768	T	C	0.16	0.24	0.15	1.77 (1.36-2.31)	7.37E-05	Imputed
rs152197	5:131461341	C	T	0.19	0.26	0.18	1.58 (1.22-2.04)	7.47E-05	Imputed
rs13088277	3:54073008	C	T	0.35	0.43	0.34	1.45 (1.16-1.82)	7.51E-05	Typed
rs184188275	4:171749103	A	T	0.05	0.09	0.05	1.92 (1.29-2.84)	7.53E-05	Imputed
rs9328928	16:858723	C	T	0.29	0.20	0.30	0.57 (0.43-0.75)	7.54E-05	Imputed
rs1148215	10:29955725	A	G	0.05	0.10	0.05	2.24 (1.51-3.30)	7.56E-05	Typed
rs3843549	8:128410480	G	A	0.15	0.08	0.15	2.04 (1.37-3.04)	7.57E-05	Imputed
rs1971115	18:20875301	C	T	0.25	0.34	0.24	0.60 (0.47-0.76)	7.60E-05	Imputed
rs1942136	7:13739731	T	C	0.23	0.32	0.22	1.69 (1.33-2.15)	7.62E-05	Typed
rs26321	5:115740627	G	C	0.32	0.21	0.33	1.86 (1.42-2.44)	7.70E-05	Imputed
rs154724	5:131445660	A	T	0.19	0.26	0.18	1.58 (1.23-2.04)	7.71E-05	Imputed
rs4677893	3:123191674	C	T	0.05	0.07	0.05	0.73 (0.47-1.12)	7.71E-05	Typed
rs9305896	21:20432739	C	T	0.31	0.24	0.32	1.55 (1.20-2.01)	7.73E-05	Imputed
rs71532762	7:148852452	C	T	0.06	0.10	0.06	1.88 (1.28-2.76)	7.79E-05	Imputed
rs1392996	11:16682627	A	G	0.40	0.50	0.39	0.63 (0.51-0.79)	7.79E-05	Imputed
rs448730	16:77977542	G	A	0.11	0.18	0.10	1.85 (1.38-2.49)	7.84E-05	Imputed
rs74023544	15:74075478	C	T	0.05	0.09	0.05	2.04 (1.37-3.05)	7.84E-05	Imputed
rs66495255	8:8599001	T	A	0.06	0.09	0.05	1.68 (1.12-2.51)	7.86E-05	Imputed
rs8081769	17:80431507	G	A	0.09	0.13	0.08	1.63 (1.16-2.29)	7.86E-05	Imputed
rs113017324	2:60266158	A	G	0.10	0.17	0.10	1.86 (1.37-2.52)	7.88E-05	Imputed
rs67918788	8:8599109	A	T	0.06	0.09	0.05	1.68 (1.12-2.51)	7.90E-05	Imputed
rs76401723	12:113838971	G	C	0.06	0.08	0.05	1.58 (1.05-2.37)	7.91E-05	Imputed
rs10080802	6:112598700	A	G	0.44	0.53	0.43	1.52 (1.21-1.89)	7.91E-05	Imputed
rs75995840	5:131486805	A	T	0.06	0.11	0.06	2.03 (1.41-2.92)	7.92E-05	Imputed

rsID	Chrom:Position	A1	A2	All MAF	Cases MAF	Controls MAF	OR (95%CI)	P-value	Typing Status
rs79344293	2:184704534	C	T	0.08	0.14	0.08	1.96 (1.42-2.71)	7.93E-05	Imputed
rs35215132	14:37521517	T	C	0.41	0.29	0.43	0.56 (0.44-0.71)	7.93E-05	Imputed
rs4756856	11:16641718	T	G	0.39	0.50	0.38	1.60 (1.28-2.00)	7.95E-05	Imputed
rs10033012	4:163766726	C	T	0.20	0.28	0.19	1.58 (1.23-2.03)	7.96E-05	Imputed
rs3116095	8:110323258	C	T	0.46	0.58	0.45	1.65 (1.32-2.07)	7.98E-05	Imputed
rs56222320	11:16644773	G	A	0.39	0.50	0.38	1.60 (1.28-2.00)	8.04E-05	Typed
rs55941903	11:16644680	A	C	0.39	0.50	0.38	1.60 (1.28-2.00)	8.04E-05	Typed
rs34584646	8:13473759	G	T	0.33	0.44	0.32	1.66 (1.32-2.07)	8.05E-05	Imputed
rs247285	5:131465688	C	T	0.19	0.26	0.18	1.58 (1.22-2.04)	8.07E-05	Imputed
rs2139746	14:37448831	T	G	0.34	0.23	0.35	0.54 (0.41-0.70)	8.07E-05	Imputed
rs79121161	8:8854607	G	C	0.07	0.12	0.06	2.07 (1.45-2.95)	8.11E-05	Imputed
rs111270388	5:131461720	T	A	0.19	0.26	0.18	1.58 (1.22-2.04)	8.12E-05	Imputed
rs247287	5:131461720	T	A	0.19	0.26	0.18	1.58 (1.22-2.04)	8.12E-05	Typed
rs156038	5:131486895	T	A	0.19	0.26	0.18	1.58 (1.22-2.04)	8.13E-05	Imputed
rs28627437	16:858643	T	C	0.28	0.19	0.29	0.58 (0.44-0.77)	8.17E-05	Imputed
rs17105860	14:37452512	T	C	0.34	0.23	0.35	0.54 (0.41-0.70)	8.19E-05	Imputed
rs17105858	14:37452458	C	T	0.34	0.23	0.35	0.54 (0.41-0.70)	8.21E-05	Imputed
rs72696748	4:163757259	G	A	0.13	0.19	0.12	1.78 (1.34-2.37)	8.22E-05	Imputed
rs12639284	3:123191006	T	C	0.06	0.07	0.05	0.73 (0.47-1.12)	8.23E-05	Imputed
rs128738	5:131540875	G	T	0.17	0.25	0.17	1.67 (1.29-2.16)	8.25E-05	Imputed
rs11846416	14:38041895	G	T	0.30	0.21	0.31	1.72 (1.31-2.25)	8.29E-05	Imputed
rs2063497	8:114900979	C	T	0.09	0.15	0.09	0.55 (0.40-0.75)	8.29E-05	Imputed
rs72696746	4:163756855	C	A	0.13	0.19	0.12	1.78 (1.34-2.37)	8.29E-05	Imputed
rs13004340	2:222208747	A	G	0.35	0.25	0.36	0.61 (0.47-0.78)	8.32E-05	Typed
rs2610686	2:104454596	G	T	0.29	0.37	0.28	1.57 (1.25-1.98)	8.33E-05	Typed
rs11024001	11:16657606	G	A	0.39	0.50	0.38	1.60 (1.28-2.00)	8.34E-05	Typed
rs6826501	4:36076676	C	T	0.48	0.37	0.49	1.58 (1.25-1.98)	8.36E-05	Typed
rs9819273	3:138152330	G	C	0.06	0.12	0.06	2.11 (1.48-3.02)	8.36E-05	Imputed
rs11841816	13:27559100	T	A	0.07	0.11	0.06	1.88 (1.31-2.71)	8.40E-05	Typed
rs7481887	11:16669707	A	C	0.39	0.50	0.38	1.60 (1.28-2.00)	8.48E-05	Imputed
rs2504193	10:26008795	A	G	0.11	0.13	0.11	0.77 (0.56-1.07)	8.49E-05	Imputed
rs6812119	4:96200453	C	A	0.41	0.31	0.42	1.58 (1.24-2.00)	8.53E-05	Imputed
rs13065943	3:68242218	C	T	0.18	0.27	0.18	1.70 (1.32-2.19)	8.58E-05	Imputed
rs17106133	14:37521300	G	A	0.41	0.29	0.43	0.56 (0.44-0.71)	8.69E-05	Imputed
rs11024008	11:16667834	C	T	0.39	0.50	0.38	1.60 (1.28-2.00)	8.70E-05	Typed
rs2339580	5:172042998	G	A	0.27	0.18	0.28	0.58 (0.44-0.77)	8.71E-05	Imputed
rs6571773	14:37521575	C	T	0.41	0.29	0.43	0.56 (0.44-0.71)	8.76E-05	Typed
rs214699	12:47627164	A	G	0.35	0.44	0.34	0.64 (0.51-0.80)	8.76E-05	Typed
rs4879638	9:32908420	T	C	0.24	0.33	0.23	1.58 (1.25-2.01)	8.77E-05	Imputed
rs714268	14:37520200	A	C	0.41	0.29	0.43	0.56 (0.44-0.71)	8.77E-05	Imputed
rs6532542	4:96200497	T	C	0.41	0.31	0.42	1.57 (1.24-1.99)	8.81E-05	Imputed
rs12454527	18:11685122	T	C	0.49	0.37	0.50	1.66 (1.32-2.09)	8.83E-05	Imputed
rs78841189	3:85478462	T	A	0.13	0.07	0.14	0.47 (0.31-0.72)	8.84E-05	Typed
rs2600050	3:2034114	G	C	0.48	0.37	0.49	1.65 (1.31-2.07)	8.85E-05	Imputed
rs7141528	14:37518886	T	C	0.41	0.29	0.43	0.56 (0.44-0.71)	8.86E-05	Imputed
rs6532541	4:96200493	T	A	0.41	0.31	0.42	1.57 (1.24-1.99)	8.88E-05	Imputed
rs2108061	7:144898822	T	G	0.32	0.41	0.31	1.51 (1.21-1.90)	8.90E-05	Imputed
rs12725975	1:17892396	T	C	0.41	0.52	0.40	1.60 (1.28-2.00)	8.94E-05	Imputed

rsID	Chrom:Position	A1	A2	All MAF	Cases MAF	Controls MAF	OR (95%CI)	P-value	Typing Status
rs56251761	12:3453848	C	T	0.05	0.10	0.05	2.19 (1.49-3.21)	8.94E-05	Imputed
rs6532544	4:96200635	T	C	0.41	0.31	0.42	1.57 (1.24-1.99)	9.00E-05	Imputed
rs7106053	11:17458730	A	G	0.34	0.44	0.33	1.58 (1.26-1.98)	9.01E-05	Imputed
rs7160964	14:37518425	C	A	0.41	0.29	0.43	0.56 (0.44-0.72)	9.03E-05	Imputed
rs56218015	12:3454149	C	T	0.05	0.10	0.05	2.19 (1.50-3.22)	9.03E-05	Imputed
rs35430590	2:36931321	G	A	0.36	0.28	0.37	0.65 (0.51-0.84)	9.04E-05	Typed
rs3095606	16:52584173	A	G	0.30	0.38	0.29	1.49 (1.18-1.87)	9.06E-05	Typed
rs10460526	2:36932843	C	T	0.36	0.28	0.37	0.65 (0.51-0.84)	9.08E-05	Imputed
rs73365341	10:116064528	G	A	0.05	0.06	0.05	1.34 (0.85-2.12)	9.09E-05	Typed
rs4819534	22:17274435	C	T	0.31	0.21	0.32	1.82 (1.39-2.38)	9.10E-05	Imputed
rs2061193	3:123193267	T	G	0.05	0.07	0.05	0.73 (0.47-1.13)	9.11E-05	Imputed
rs75258089	6:98556635	T	A	0.07	0.11	0.06	1.95 (1.36-2.81)	9.12E-05	Imputed
rs10010112	4:96200216	T	C	0.41	0.31	0.42	1.57 (1.24-1.99)	9.17E-05	Typed
rs7534609	1:116780258	A	T	0.10	0.14	0.10	1.43 (1.03-1.98)	9.20E-05	Imputed
rs2415378	14:37545124	A	G	0.41	0.30	0.43	0.56 (0.44-0.72)	9.24E-05	Imputed
rs74509529	15:95395832	C	T	0.11	0.17	0.10	1.80 (1.33-2.44)	9.26E-05	Imputed
rs10010041	4:96200194	T	C	0.41	0.31	0.42	1.57 (1.24-1.99)	9.26E-05	Imputed
rs78752216	15:95398436	T	C	0.11	0.17	0.10	1.80 (1.33-2.44)	9.36E-05	Imputed
rs12464898	2:117467066	C	T	0.45	0.55	0.44	1.59 (1.27-1.98)	9.39E-05	Imputed
rs17061200	13:40943522	T	C	0.40	0.49	0.39	1.54 (1.24-1.93)	9.43E-05	Imputed
rs1317983	6:43806335	T	C	0.34	0.42	0.33	0.70 (0.56-0.88)	9.47E-05	Imputed
rs6682992	1:116779561	T	C	0.10	0.14	0.10	1.43 (1.03-1.98)	9.51E-05	Imputed

REFERENCES

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, *et al.* Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *International Journal of Cancer* (2014).
2. Shiels, M.S., Engels, E. A., Linet, M. S., *et al.* The Epidemic of Non–Hodgkin Lymphoma in the United States: Disentangling the Effect of HIV, 1992–2009. *Cancer Epidemiology Biomarkers & Prevention* **22**, 1068-1078 (2013).
3. Adebamowo, C.A., Casper, C., Bhatia, K., *et al.* Challenges in the detection, prevention, and treatment of HIV-associated malignancies in low-and middle-income countries in Africa. *JAIDS Journal of Acquired Immune Deficiency Syndromes* **67**, S17-S26 (2014).
4. Parkin, D.M., Bray, F., Ferlay, J., & Jemal, A. Cancer in Africa 2012. *Cancer Epidemiology Biomarkers & Prevention* **23**, 953-966 (2014).
5. Petersen, M., Yiannoutsos, C. T., Justice, A., & Egger, M. Observational research on NCDs in HIV-positive populations: conceptual and methodological considerations. *JAIDS Journal of Acquired Immune Deficiency Syndromes* **67**, S8-S16 (2014).
6. Narayan, K.V., Miotti, P. G., Anand, N. P., Kline, L. M., Harmston, C., Gulakowski III, R., & Vermund, S. H. HIV and noncommunicable disease comorbidities in the era of antiretroviral therapy: a vital agenda for research in low-and middle-income country settings. *JAIDS Journal of Acquired Immune Deficiency Syndromes* **67**(2014).
7. Gibson, T.M., Morton, L.M., Shiels, M. S., Clarke, C. A., & Engels, E. A. Risk of non-Hodgkin lymphoma subtypes in HIV-infected people during the HAART era: a population-based study. *AIDS* **28**, 2313-2318 (2014).
8. Robbins, H.A., Shiels, M. S., Pfeiffer, R. M., & Engels, E. A. Epidemiologic contributions to recent cancer trends among HIV-infected people in the United States. *AIDS* **28**, 881-890 (2014).

9. Muñoz A, Schrager LK, Bacellar H, Speizer I, Vermund SH, Detels R, Saah AJ, Kingsley LA, Seminara D, Phair JP. Trends in the incidence of outcomes defining acquired immunodeficiency syndrome (AIDS) in the Multicenter AIDS Cohort Study: 1985-1991. *American Journal of Epidemiology* **137**, 423-428 (1993).

10. Regidor DL, Detels R, Breen EC, Widney DP, Jacobson LP, Palella F, Rinaldo CR, Bream JH, Martinez-Maza O. Effect of highly active antiretroviral therapy on biomarkers of B-lymphocyte activation and inflammation. *AIDS* **25**, 303-314 (2011).

11. Sacktor, N. The epidemiology of human immunodeficiency virus-associated neurological disease in the era of highly active antiretroviral therapy. *Journal of Neurovirology* **8**, Suppl 2115-2121 (2002).

12. International Collaboration on HIV and Cancer. Highly active antiretroviral therapy and incidence of cancer in human immunodeficiency virus-infected adults. *Journal of the National Cancer Institute* **92**, 1823-1830 (2000).

13. Jacobson LP, Yamashita T, Detels R, Margolick JB, Chmiel JS, Kingsley LA, Melnick S, Muñoz A. Impact of potent antiretroviral therapy on the incidence of Kaposi's sarcoma and non-Hodgkin's lymphomas among HIV-1-infected individuals in the Multicenter AIDS Cohort Study. *JAIDS Journal of Acquired Immune Deficiency Syndromes* **21** S34-S41 (1999).

14. Seaberg, E.C., Wiley, D., Martínez-Maza, O., *et al.* Cancer incidence in the multicenter aids cohort study before and during the HAART era. *Cancer* **116**, 5507-5516 (2010).

15. Park LS, Tate JP, Rodriguez-Barradas MC, Rimland D, Goetz MB, *et al.* Cancer Incidence in HIV-Infected Versus Uninfected Veterans: Comparison of Cancer Registry and ICD-9 Code Diagnoses. *J AIDS Clin Res* **5**(2014).

16. Thapa, D.R., Hussain, S. K., Tran, W. C *et al.* Serum MicroRNAs in HIV-Infected Individuals as Pre-Diagnosis Biomarkers for AIDS-NHL. *JAIDS Journal of Acquired Immune Deficiency Syndromes* **66**,

229-237 (2014).

17. Breen, E.C., Hussain, S. K., Magpantay, L., *et al.* B-Cell Stimulatory Cytokines and Markers of Immune Activation Are Elevated Several Years Prior to the Diagnosis of Systemic AIDS–Associated Non-Hodgkin B-Cell Lymphoma. *Cancer Epidemiology Biomarkers & Prevention* **20**, 1303-1314 (2011).

18. Hussain, S.K., Zhu, W., Chang, S. C., *et al.* Serum levels of the chemokine CXCL13, genetic variation in CXCL13 and its receptor CXCR5, and HIV-associated non-hodgkin B-cell lymphoma risk. *Cancer Epidemiology Biomarkers & Prevention* **22**, 295-307 (2013).

19. Hussain, S.K., Hessol, N. A., Levine, A. M *et al.* Serum biomarkers of immune activation and subsequent risk of non-Hodgkin B-cell lymphoma among HIV-infected women. *Cancer Epidemiology Biomarkers & Prevention* **22**, 2084-2093 (2013).

20. Vendrame, E., Hussain, S. K., Breen, E. C. *et al.* Serum levels of cytokines and biomarkers for inflammation and immune activation, and HIV-associated non-Hodgkin B-cell lymphoma risk. . *Cancer Epidemiology Biomarkers & Prevention* **23**, 343-349 (2014).

21. Cerhan, J.R. & Slager, S.L. Familial predisposition and genetic risk factors for lymphoma. *Blood* **126**, 2265-2273 (2015).

22. Morton L.M., Slager S.L., Cerhan J.R., Wang S.S., Vajdic C.M., Skibola C.F., *et al.* Etiologic heterogeneity among non-hodgkin lymphoma subtypes. *Journal of the National Cancer Institute-Monographs* **48**, 130-144 (2014).

23. Shiels MS, Pfeiffer R, Besson C, Clarke CA, Morton LM, Nogueira L, *et al.* Trends in primary central nervous system lymphoma incidence and survival in the US. *British Journal of Haematology* (2016).

24. Villano JL, Koshy M, Shaikh H, Dolecek TA, McCarthy BJ. Age, gender, and racial differences in incidence and survival in primary CNS lymphoma. *British journal of cancer* **105**, 1414-1418

(2011).

25. Rositch, A.F., & Riedel, D. J. Recent cancer trends in HIV-infected individuals in the United States: clues to global cancer trends in HIV populations. *AIDS* **28**, 925-926 (2014).

26. Pieper, K., Grimbacher, B., & Eibel, H. B-cell biology and development. *Journal of Allergy and Clinical Immunology* **131**, 959-971 (2013).

27. Murphy, K.M. *Janeway's immunobiology* (Garland Science., 2011).

28. Lenz, G. *et al.* Aberrant immunoglobulin class switch recombination and switch translocations in activated B cell–like diffuse large B cell lymphoma. *The Journal of experimental medicine* **204**, 633-643 (2007).

29. Swerdllow, S.H., Campo, E., & Harris, N. L. *WHO classification of tumours of haematopoietic and lymphoid tissues*, (IARC Press, Lyon, 2008).

30. Jaffe ES. *The 2008 WHO classification of lymphomas: implications for clinical practice and translational research.* (American Society of Hematology, 2009).

31. Alizadeh, A.A., Eisen, M. B., Davis, R. E., *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503-511 (2000).

32. Centers for Disease Control and Prevention. Revision of the CDC Surveillance Case Definition for Acquired Immune Deficiency Syndrome. *Morbidity and Mortality Weekly Report* **36**(1987).

33. Chen YB, Rahemtullah A., Hochberg E. Primary Effusion Lymphoma. *The Oncologist 2007* **12**, 569-576 (2007).

34. Greenough, A., & Dave, S. S. New clues to the molecular pathogenesis of Burkitt lymphoma revealed through next-generation sequencing. *Current opinion in hematology* **21**, 326-332 (2014).

35. Sandler, N.G., & Douek, D. C. Microbial translocation in HIV infection: causes, consequences and treatment opportunities. *Nature Reviews Microbiology* **10**, 655-666 (2012).

36. Reus, S., Portilla, J., Sánchez-Payá, J *et al.* Low-level HIV viremia is associated with microbial translocation and inflammation. *JAIDS Journal of Acquired Immune Deficiency Syndromes 2013* **62**, 129-134 (2013).

37. Nolen, B.M., Breen, E. C., Bream, J. H., Jenkins, F. J., Kingsley, L. A., Rinaldo, C. R., & Lokshin, A. E. Circulating Mediators of Inflammation and Immune Activation in AIDS-Related Non-Hodgkin Lymphoma. *PloS One* **9**, e99144.

38. Lake A, Shield L., Cordano P, *et al.* Mutations of NFKBIA, encoding IκBα, are a recurrent finding in classical Hodgkin lymphoma but are not a unifying feature of non-EBV-associated cases. *Int J Cancer* **125**, 1334-1342 (2009).

39. Chang ET, Birmann B., Kasperzyk JL, *et al.* Polymorphic variation in NFKB1 and other aspirin-related genes and risk of Hodgkin lymphoma. *Cancer Epidemiology Biomarkers & Prevention* **18**, 976-986 (2009).

40. Du J, Huo J., Shi J, *et al.* Polymorphisms of NF-κB family genes are associated with development of multiple myeloma and treatment outcome in patients undergoing bortezomib-based regimens. *Haematologica* **96**, 729-37 (2011).

41. Corthals SL, Johnson D., de Knegt Y, *et al.* Genetic associations with bortezomib mediated neuropathy in multiple myeloma. *Blood* **114**(2009).

42. Lima CSP *et al.* "Base of tongue squamous cell carcinoma susceptibility: Novel candidate genetic polymorphisms identified in genome-wide association study." in *ASCO Annual Meeting Proceeding* Vol. 30 (2012).

43. Visvader JE, Lindeman G. Cancer stem cells: current status and evolving complexities. *Cell Stem Cell* **10**, 717-728 (2012).

44. Park T, Donnenberg V, Donnenberg A, Zambidis E, Zimmerlin L. Dynamic Interactions Between Cancer Stem Cells and Their Stromal Partners. *Curr Pathobiol Rep* **2**, 41-52 (2014).

45. Kikushige Y, Ishikawa F., Miyamoto T *et al.* Self-renewing hematopoietic stem cell is the primary target in pathogenesis of human chronic lymphocytic leukemia. *Cancer cell* **20**, 246-259 (2011).

46. Tarleton HP, Lemischka I. Delayed differentiation in embryonic stem cells and mesodermal progenitors in the absence of CtBP2. *Mechanisms of development* **127**, 107-119 (2010).

47. Zhang, C., Gao, C., Xu, Y., & Zhang, Z. CtBP2 could promote prostate cancer cell proliferation through c-Myc signaling. *Gene 2014* **546**, 73-79 (2014).

48. Radzisheuskaya A, Silva JCR. Do all roads lead to Oct4? The emerging concepts of induced pluripotency. *Trends in cell biology* **24**, 275-284 (2014).

49. Gentles AJ, Alizadeh A, Lee SI, Myklebust JH, Shachaf CM, Shahbaba B *et al.* A pluripotency signature predicts histologic transformation and influences survival in follicular lymphoma patients. *Blood* **114**, 3158-3166 (2009).

50. Lee MY, Lu A, Gudas LJ. Transcriptional regulation of Rex1 (zfp42) in normal prostate epithelial cells and prostate cancer cells. *Journal of cellular physiology* **224**, 17-27 (2010).

51. Munz, M., Baeuerle, P. A., & Gires, O. The emerging role of EpCAM in cancer and stem cell signaling. *Cancer research* **69**, 5627-5629 (2009).

52. Park WS, Oh RR, Park JY, Lee SH, Shin MS, Kim YS, *et al.* Frequent Somatic Mutations of the β-Catenin Gene in Intestinal-Type Gastric Cancer. *Cancer Research* **59**, 4257-4260 (1999).

53. Mazieres J, You L, He B, Xu Z, Lee AY, Mikami I, *et al.* Inhibition of Wnt16 in human acute lymphoblastoid leukemia cells containing the t(1;19) translocation induces apoptosis. *Oncogene* **24**, 5396-5400 (2005).

54. Park J, Song JH, He T, Nam S, Lee J, Park W. Overexpression of Wnt-2 in colorectal cancers. *Neoplasma* **56**, 119-123 (2009).

55. Tanaka S, Akiyoshi T, Mori M, Wands JR, Sugimachi K. A novel frizzled gene identified in human esophageal carcinoma mediates APC/β-catenin signals. *Proceedings of the National Academy of*

Sciences **95**, 10164-9 (1998).

56. Al-Harthi, L. Interplay between Wnt/β-catenin signaling and HIV: Virologic and biologic consequences in the CNS. *J Neuroimmune Pharmacol.* **7**, 731-739 (2012).

57. Dreesen O, Brivanlou A. Signaling Pathways in Cancer and Embryonic Stem Cells. *Stem Cell Reviews.* **3**, 7-17 (2007).

58. Bertrand, F.E., Angus, C. W., Partis, W. J., & Sigounas, G. (2012). Developmental pathways in colon cancer: crosstalk between WNT, BMP, Hedgehog and Notch. *Cell Cycle* **11**, 4344-4351 (2012).

59. Kelly KF, Ng DY, Jayakumaran G, Wood GA, Koide H, Doble BW. β-catenin enhances Oct-4 activity and reinforces pluripotency through a TCF-independent mechanism. *Cell Stem Cell* **8**, 214-227 (2011).

60. Li C, Zhang S., Lu Y, Zhang Y, Wang E, *et al.* . The Roles of Notch3 on the Cell Proliferation and Apoptosis Induced by CHIR99021 in NSCLC Cell Lines: A Functional Link between Wnt and Notch Signaling Pathways. *PLoS ONE* **8**, e48659 (2013).

61. Jia, D. *et al.* β-Catenin and NF-κB co-activation triggered by TLR3 stimulation facilitates stem cell-like phenotypes in breast cancer. *Cell death and differentiation* **22**, 298-310 (2015).

62. Manolio, TA, Collins SL, Cox NJ *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747-753 (2009).

63. Mooney, M.A., Nigg, J. T., McWeeney, S. K., & Wilmot, B. (2014). Functional and genomic context in pathway analysis of GWAS data. *Trends in Genetics* **30**, 390-400 (2014).

64. Vijai, J. *et al.* A genome-wide association study of marginal zone lymphoma shows association to the HLA region. *Nature communications* **6**, 5751 (2015).

65. Slager, S.L., Jardin, F., Veron, A. S., *et al.* A Genome-Wide Association Study (GWAS) Of Event-Free Survival In Diffuse Large B-Cell Lymphoma (DLBCL) Treated With Rituximab and

Anthracycline-Based Chemotherapy: A Lysa and Iowa/Mayo Clinic SPORE Multistage Study. *Blood* **122**(2013).

66. Cerhan JR, Berndt S, Vijai J, Ghesquières H, McKay J, Wang SS, Wang Z, Yeager M, Conde L, De Bakker PI, Nieters A. Genome-wide association study identifies multiple susceptibility loci for diffuse large B cell lymphoma. *Nature genetics* **46**, 1233-8 (2014).

67. Smedby, K.E. *et al.* GWAS of follicular lymphoma reveals allelic heterogeneity at 6p21.32 and suggests shared genetic susceptibility with diffuse large B-cell lymphoma. *PLoS genetics* **7**, e1001378 (2011).

68. Skibola, C.F. *et al.* Genome-wide association study identifies five susceptibility loci for follicular lymphoma outside the HLA region. *The American Journal of Human Genetics* **95**, 462-471 (2014).

69. Speedy, H.E., Di Bernardo, M. C., Sava, G. P., Dyer, M. J., Holroyd, A., Wang, Y., *et al.* A genome-wide association study identifies multiple susceptibility loci for chronic lymphocytic leukemia. *Nature genetics* **46**, 56-60 (2014).

70. Fellay, J. *et al.* Common genetic variation and the control of HIV-1 in humans. *PLoS genetics* **5**, e1000791 (2009).

71. Fellay, J. *et al.* A whole-genome association study of major determinants for host control of HIV-1. *Science* **317**, 944-947 (2007).

72. Helminen, M., Lahdenpohja, N. & Hurme, M. Polymorphism of the interleukin-10 gene is associated with susceptibility to Epstein-Barr virus infection. *Journal of Infectious Diseases* **180**, 496-499 (1999).

73. Wong, H.-L. *et al.* Cytokine signaling pathway polymorphisms and AIDS-related non-Hodgkin lymphoma risk in the multicenter AIDS cohort study. *AIDS (London, England)* **24**, 1025 (2010).

74. Marchetti, G., Tincati, C. & Silvestri, G. Microbial translocation in the pathogenesis of HIV infection and AIDS. *Clinical microbiology reviews* **26**, 2-18 (2013).

75. Morton, L.M. *et al.* Rationale and design of the International Lymphoma Epidemiology Consortium (InterLymph) non-Hodgkin lymphoma subtypes project. *Journal of the National Cancer Institute-Monographs*, 1-14 (2014).

76. Detels R, J.L., Margolick J, Martinez-Maza O, Munoz A, Phair J, Rinaldo C, Wolinsky S. . The multicenter AIDS cohort study, 1983 to…. . *Public health* **126**, 196-8 (2012).

77. Fluidigm SNP Genotyping Analysis Version 3.1.2 Build 20111017.1807. Fluidigm: South San Francisco, CA, USA.

78. Ramakrishnan, R., Qin, J., Jones, R.C. & Weaver, L.S. Integrated fluidic circuits (IFCs) for digital PCR. *Microfluidic Diagnostics: Methods and Protocols*, 423-431 (2013).

79. Martínez-Maza, O. & Breen, E.C. B-cell activation and lymphoma in patients with HIV. *Current opinion in oncology* **14**, 528-532 (2002).

80. Breen, E.C. *et al.* Non-Hodgkin's B cell lymphoma in persons with acquired immunodeficiency syndrome is associated with increased serum levels of IL10, or the IL10 promoter− 592 C/C genotype. *Clinical Immunology* **109**, 119-129 (2003).

81. Chao, C. *et al.* Recreational amphetamine use and risk of HIV-related non-Hodgkin lymphoma. *Cancer causes & control* **20**, 509-516 (2009).

82. Mellors, J.W. *et al.* Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. *Science* **272**, 1167-1170 (1996).

83. Morton, L.M. *et al.* Lymphoma incidence patterns by WHO subtype in the United States, 1992-2001. *Blood* **107**, 265-276 (2006).

84. Terrier, B. *et al.* Characteristics of B-cell lymphomas in HIV/HCV-coinfected patients during the combined antiretroviral therapy era: an ANRS CO16 LYMPHOVIR cohort study. *JAIDS Journal of Acquired Immune Deficiency Syndromes* **63**, 249-253 (2013).

85. Chang, P.-Y. *et al.* Comment on" characteristics of B-cell lymphomas in HIV/HCV-coinfected

patients during the combined antiretroviral therapy era: an ANRS CO16 LYMPHOVIR cohort study". *Journal of acquired immune deficiency syndromes (1999)* **67**, e84-6 (2014).

86. Koff, J.L. *et al.* To each its own: linking the biology and epidemiology of NHL subtypes. *Current hematologic malignancy reports* **10**, 244-255 (2015).

87. Mellors, J.W. *et al.* Plasma viral load and CD4+ lymphocytes as prognostic markers of HIV-1 infection. *Annals of internal medicine* **126**, 946-954 (1997).

88. Schacker, T.W., Hughes, J.P., Shea, T., Coombs, R.W. & Corey, L. Biological and virologic characteristics of primary HIV infection. *Annals of internal medicine* **128**, 613-620 (1998).

89. Mefford, J. & Witte, J.S. The covariate's dilemma. *PLoS genetics* **8**(2012).

90. Hernán, M.A., Hernández-Díaz, S. & Robins, J.M. A structural approach to selection bias. *Epidemiology* **15**, 615-625 (2004).

91. Pirinen, M., Donnelly, P. & Spencer, C.C. Including known covariates can reduce power to detect genetic effects in case-control studies. *Nature genetics* **44**, 848-851 (2012).

92. Greenland, S. & Pearce, N. Statistical Foundations for Model-Based Adjustments. *Annual review of public health* **36**, 89-108 (2015).

93. Rothman, K.J., Greenland, S. & Lash, T.L. *Modern epidemiology*, (Lippincott Williams & Wilkins, 2008).

94. Weng, H.-Y., Hsueh, Y.-H., Messam, L.L.M. & Hertz-Picciotto, I. Methods of covariate selection: directed acyclic graphs and the change-in-estimate procedure. *American journal of epidemiology*, kwp035 (2009).

95. Desai, M., Kubo, J., Esserman, D. & Terry, M.B. The handling of missing data in molecular epidemiology studies. *Cancer Epidemiology Biomarkers & Prevention* **20**, 1571-1579 (2011).

96. van der Heijden, G.J., Donders, A.R.T., Stijnen, T. & Moons, K.G. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic

research: a clinical example. *Journal of clinical epidemiology* **59**, 1102-1109 (2006).

97. Donders, A.R.T., van der Heijden, G.J., Stijnen, T. & Moons, K.G. Review: a gentle introduction to imputation of missing values. *Journal of clinical epidemiology* **59**, 1087-1091 (2006).

98. Peckham, E.C. MicroRNA-related Polymorphisms and Non-Hodgkin Lymphoma Susceptibility in the Multicenter AIDS Cohort Study. Unpublished doctoral dissertation, University of California Los Angeles (2014).

99. Moons, K.G., Donders, R.A., Stijnen, T. & Harrell, F.E. Using the outcome for imputation of missing predictor values was preferred. *Journal of clinical epidemiology* **59**, 1092-1101 (2006).

100. Graham, J.W., Olchowski, A.E. & Gilreath, T.D. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science* **8**, 206-213 (2007).

101. Greenland, S. A semi-Bayes approach to the analysis of correlated multiple associations, with an application to an occupational cancer-mortality study. *Statistics in medicine* **11**, 219-30 (1992).

102. Yang, T. *et al.* Tobacco smoking, NBS1 polymorphisms, and survival in lung and upper aerodigestive tract cancers with semi-Bayes adjustment for hazard ratio variation. *Cancer causes & control : CCC* **25**, 11-23 (2014).

103. Chang, S.-C. *et al.* Single nucleotide polymorphisms of one-carbon metabolism and cancers of the esophagus, stomach, and liver in a Chinese population. *PloS one* **9**, e109235 (2014).

104. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research* **42**, D1001-D1006 (2014).

105. Bhandari, D.R. *et al.* REX-1 expression and p38 MAPK activation status can determine proliferation/differentiation fates in human mesenchymal stem cells. *PLoS One* **5**, e10493 (2010).

106. Gee, K., Angel, J.B., Mishra, S., Blahoianu, M.A. & Kumar, A. IL-10 regulation by HIV-Tat in primary human monocytic cells: involvement of calmodulin/calmodulin-dependent protein

kinase-activated p38 MAPK and Sp-1 and CREB-1 transcription factors. *The Journal of Immunology* **178**, 798-807 (2007).

107. Cohen, P.S. *et al.* The critical role of p38 MAP kinase in T cell HIV-1 replication. *Molecular Medicine* **3**, 339 (1997).

108. Ma, C. *et al.* Significant association between the Axin2 rs2240308 single nucleotide polymorphism and the incidence of prostate cancer. *Oncology letters* **8**, 789-794 (2014).

109. Narasipura, S.D. *et al.* Role of β-catenin and TCF/LEF family members in transcriptional activity of HIV in astrocytes. *Journal of virology* **86**, 1911-1921 (2012).

110. Wu, Z. *et al.* AXIN 2 rs2240308 polymorphism contributes to increased cancer risk: evidence based on a meta-analysis. *Cancer cell international* **15**, 1 (2015).

111. Dolmans, G.H. *et al.* WNT2 locus is involved in genetic susceptibility of Peyronie's disease. *The journal of sexual medicine* **9**, 1430-1434 (2012).

112. Dolmans, G.H. *et al.* Wnt signaling and Dupuytren's disease. *New England Journal of Medicine* **365**, 307-317 (2011).

113. Wallar, G. Polymorphisms in the stem cell pathway and esophageal cancer in a Chinese population. Unpublished doctoral dissertation, University of California Los Angeles (2013).

114. Liu, X. Genetic Susceptibility and Environmental Risk Factors of Liver Cancer, a Population-based Case-control Study in Jiangsu Province, China. (2015).

115. Pierzynski, J.A. *et al.* Genetic Variants in the Wnt/β-Catenin Signaling Pathway as Indicators of Bladder Cancer Risk. *The Journal of urology* **194**, 1771-1776 (2015).

116. Greenland, S. & Finkle, W.D. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American journal of epidemiology* **142**, 1255-1264 (1995).

117. Hughes, S., Arneson, N., Done, S. & Squire, J. The use of whole genome amplification in the study of human disease. *Progress in biophysics and molecular biology* **88**, 173-189 (2005).

118. Lasken, R.S. & Egholm, M. Whole genome amplification: abundant supplies of DNA from precious samples or clinical specimens. *Trends in biotechnology* **21**, 531-535 (2003).

119. Macaulay, I.C. & Voet, T. Single cell genomics: advances and future perspectives. *PLoS Genet* **10**, e1004126 (2014).

120. Epeldegui, M., Hung, Y.P., McQuay, A., Ambinder, R.F. & Martínez-Maza, O. Infection of human B cells with Epstein-Barr virus results in the expression of somatic hypermutation-inducing molecules and in the accrual of oncogene mutations. *Molecular immunology* **44**, 934-942 (2007).

121. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature genetics* **44**, 955-959 (2012).

122. Fuchsberger, C., Abecasis, G.R. & Hinds, D.A. minimac2: faster genotype imputation. *Bioinformatics*, btu704 (2014).

123. Michigan Imputation Server (https://imputationserver.sph.umich.edu/index.html#!pages/home).

124. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *bioRxiv Jan 035170* (2015).

125. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics* **39**, 906-13 (2007).

126. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nature reviews. Genetics* **11**, 499-511 (2010).

127. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* **38**, 904-909 (2006).

128. Price, A.L., Zaitlen, N.A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics* **11**, 459-463 (2010).

129. Reich, D., Price, A.L. & Patterson, N. Principal component analysis of genetic data. *Nature genetics* **40**, 491-492 (2008).

130. Mansournia, M.A., Hernán, M.A. & Greenland, S. Matched designs and causal diagrams. *International journal of epidemiology* **42**, 860-869 (2013).

131. Cantor, R.M., Lange, K. & Sinsheimer, J.S. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *The American Journal of Human Genetics* **86**, 6-22 (2010).

132. Levine, A.J. *et al.* Genome-wide association study of neurocognitive impairment and dementia in HIV-infected adults. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **159**, 669-683 (2012).

133. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic epidemiology* **34**, 816-834 (2010).

134. Anderson, C.A. *et al.* Data quality control in genetic case-control association studies. *Nature protocols* **5**, 1564-1573 (2010).

135. Turner, S. *et al.* Quality control procedures for genome-wide association studies. *Current protocols in human genetics*, 1.19. 1-1.19. 18 (2011).

136. Weale, M.E. Quality control for genome-wide association studies. *Genetic Variation: Methods and Protocols*, 341-372 (2010).

137. Rayner, W. McCarthy Group Tools. URL http://www.well.ox.ac.uk/~wrayner/tools/

138. Chang CC, C.C., Tellier LC, Vattikuti S, Purcell SM, Lee JJ. . Second-generation PLINK: rising to the challenge of larger and richer datasets. . *Gigascience* **4**(2015).

139. Purcell S, N.B., Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81**, 559-75 (2007).

140. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011).

141. Turner, S.D. qqman: an R package for visualizing GWAS results using QQ and manhattan plots. *bioRxiv*, 005165 (2014).

142. VanderWeele, T.J. Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology (Cambridge, Mass.)* **21**, 540 (2010).

143. Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326-3328 (2012).

144. Clayton, D. snpStats: SnpMatrix and XSnpMatrix classes and methods. R package version 1.16.0. (2014).

145. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997-1004 (1999).

146. Dadaev, T., Leongamornlert, D.A., Saunders, E.J., Eeles, R. & Kote-Jarai, Z. LocusExplorer: a user-friendly tool for integrated visualization of human genetic association data and biological annotations. *Bioinformatics (Oxford, England)* (2015).

147. Speir, M.L. *et al.* The UCSC Genome Browser database: 2016 update. *Nucleic acids research* **44**, D717-D725 (2016).

148. Duggal, P. *et al.* Genome-wide association study of spontaneous resolution of hepatitis C virus infection: data from multiple cohorts. *Annals of internal medicine* **158**, 235-245 (2013).

149. Machiela, M.J. & Chanock, S.J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31**, 3555-3557 (2015).

150. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).

151. Koch, C.M. *et al.* The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome research* **17**, 691-707 (2007).

152. Berndt, S.I. *et al.* Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia. *Nature genetics* **45**, 868-876 (2013).

153. Zhou, Y., Zhang, K., Yin, X., Nie, Q. & Ma, Y. HIV-1 Tat Protein Enhances Expression and Function of Breast Cancer Resistance Protein. *AIDS Research and Human Retroviruses* **32**, 1-3 (2016).

154. Zheng, X. *et al.* HIBAG–HLA genotype imputation with attribute bagging. *The pharmacogenomics journal* **14**, 192 (2014).

155. ALSGEN Consortium. Age of onset of amyotrophic lateral sclerosis is modulated by a locus on 1p34. 1. *Neurobiology of aging* **34**, 357. e7-357. e19 (2013).

156. Farh, KKH. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337-343 (2015).

157. Brown, C.C. & Wedderburn, L.R. Genetics: Mapping autoimmune disease epigenetics: what's on the horizon? *Nature Reviews Rheumatology* **11**, 131-132 (2015).

158. Mu, L.N. *et al.* Green tea drinking and multigenetic index on the risk of stomach cancer in a Chinese population. *International journal of cancer* **116**, 972-983 (2005).

159. Wojcik, G.L., Kao, W.L. & Duggal, P. Relative performance of gene-and pathway-level methods as secondary analyses for genome-wide association studies. *BMC genetics* **16**, 1 (2015).

160. Mishra, A. & Macgregor, S. VEGAS2: Software for More Flexible Gene-Based Testing. *Twin Research and Human Genetics* **18**, 86-91 (2015).

161. Mishra, A. *Methods for Genetic Epidemiology*. Unpublished Doctoral Dissertation, University of Queensland (2015).

162. Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z. & Bergmann, S. Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. *PLoS computational biology* **12**, e1004714 (2016).

163. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for

interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545-50 (2005).

164. Mooney, M.A. & Wilmot, B. Gene set analysis: A step-by-step guide. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **168**, 517-527 (2015).

165. Segrè, A.V. *et al.* Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS genetics* **6**(2010).

166. Yu, K. *et al.* Pathway analysis by adaptive combination of P-values. *Genetic epidemiology* **33**, 700-709 (2009).

167. Zhang, H. *et al.* A fast multilocus test with adaptive SNP selection for large-scale genetic-association studies. *European Journal of Human Genetics* **22**(2014).

168. Zhang, H. *et al.* A Powerful Procedure for Pathway-based Meta-Analysis Using Summary Statistics Identifies 43 Pathways Associated with Type II Diabetes in European Populations. *bioRxiv*, 041244 (2016).

169. Gene Ontology Consortium. Gene ontology consortium: going forward. *Nucleic acids research* **43**, D1049-D1056 (2015).

170. Liberzon, A. *et al.* The Molecular Signatures Database Hallmark Gene Set Collection. *Cell systems* **1**, 417-425 (2015).

171. Xie, X. *et al.* Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* **434**, 338-45 (2005).

172. Godec, J. *et al.* Compendium of Immune Signatures Identifies Conserved and Species-Specific Biology in Response to Inflammation*. Immunity **44,** 1-13* (2016).

173. Segal, E., Friedman, N., Koller, D. & Regev, A. A module map showing conditional activity of expression modules in cancer. *Nature genetics* **36**, 1090-1098 (2004).

174. Brentani, H. *et al. The generation and utilization of a cancer-oriented representation of the*

human transcriptome by using expressed sequence tags, 13418-23 (2003).

175. Gaudet, P. & Dessimoz, C. Gene Ontology: Pitfalls, Biases, Remedies. *arXiv preprint arXiv:1602.01875* (2016).

176. Newman, J.C. & Weiner, A.M. L2L: a simple tool for discovering the hidden significance in microarray expression data. *Genome biology* **6**, R81 (2005).

177. Zeller, K.I., Jegga, A.G., Aronow, B.J. & Dang, C.V. An integrated database of genes responsive to the Myc oncogenic transcription factor: identification of direct genomic targets. *Genome Biol* **4**, R69 (2003).

178. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic acids research* **44**, D457-62 (2016).

179. Naba, A. *et al.* The matrisome: in silico definition and in vivo characterization by proteomics of normal and tumor extracellular matrices. *Molecular & cellular proteomics : MCP* **11**, M111.014647 (2012).

180. Naba, A., Clauser, K.R. & Hynes, R.O. Enrichment of Extracellular Matrix Proteins from Tissues and Digestion into Peptides for Mass Spectrometry Analysis. *Journal of visualized experiments : JoVE*, e53057 (2015).

181. Schaefer, C.F. *et al.* PID: the Pathway Interaction Database. *Nucleic acids research* **37**, D674-9 (2009).

182. Pratt, D. *et al.* NDEx, the Network Data Exchange. *Cell systems* **1**, 302-305 (2015).

183. Fabregat, A. *et al.* The Reactome pathway Knowledgebase. *Nucleic acids research* **44**, D481-7 (2016).

184. Joshi-Tope, G. *et al.* Reactome: a knowledgebase of biological pathways. *Nucleic acids research* **33**, D428-32 (2005).

185. Gene Ontology Consortium. *Guide to GO Evidence Codes.* http://geneontology.org/page/guide-

go-evidence-code, (2016).

186. Bastian, F.B. *et al.* The Confidence Information Ontology: a step towards a standard for asserting confidence in annotations. *Database: the journal of biological databases and curation*, bav043 (2015).

187. Edgar, R., Domrachev, M. & Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research* **30**, 207-210 (2002).

188. Clough, E. & Barrett, T. The Gene Expression Omnibus Database. *Statistical Genomics: Methods and Protocols*, 93-110 (2016).

189. Su, A.I. *et al.* Molecular classification of human carcinomas by use of gene expression signatures. *Cancer research* **61**, 7388-93 (2001).

190. Ross, D.T. *et al.* Systematic variation in gene expression patterns in human cancer cell lines. *Nature genetics* **24**, 227-35 (2000).

191. Scherf, U. *et al.* A gene expression database for the molecular pharmacology of cancer. *Nature genetics* **24**, 236-44 (2000).

192. *http://software.broadinstitute.org/gsea/msigdb/collections.jsp*.

193. *Broad Institute of MIT and Harvard (MsigDB). "Investigate Gene Sets."*

194. Gentleman, R.C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome biology* **5**, R80 (2004).

195. Wickham, H. & Francois, R. dplyr: A grammar of data manipulation. *URL http://CRAN.R-project.org/package=dplyr. R package version 0.2* (2014).

196. Gehlenborg, N. UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets. R package version 1.2.0. http://CRAN.R-project.org/package=UpSetR. (2016).

197. Wu, C., Mark, A. & Su, A.I. MyGene. info: gene annotation query as a service. *bioRxiv*, 009332 (2014).

198. VErsatile Gene-based Association Study 2 (VEGAS2) https://vegas2.qimrberghofer.edu.au.

199. R Core Team. R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, 2012). *URL: http:// www. R-project. org* (2015).

200. Mi, H., Poudel, S., Muruganujan, A., Casagrande, J.T. & Thomas, P.D. PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic acids research* **44**, D336-D342 (2016).

201. Iglesias, A. *et al.* Diabetes and exocrine pancreatic insufficiency in E2F1/E2F2 double-mutant mice. *The Journal of clinical investigation* **113**, 1398-407 (2004).

202. Darce, J. *et al.* An N-terminal mutation of the Foxp3 transcription factor alleviates arthritis but exacerbates diabetes. *Immunity* **36**, 731-41 (2012).

203. Szanto, A. *et al.* STAT6 transcription factor is a facilitator of the nuclear receptor PPARγ-regulated gene expression in macrophages and dendritic cells. *Immunity* **33**, 699-712 (2010).

204. *AmiGo2. "Structural Molecule Activity"*, (2016).

205. Barbie, D.A. *et al.* Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462**, 108-12 (2009).

206. *http://software.broadinstitute.org/gsea/msigdb/cards/KRAS.300_UP.V1_UP.html*.

207. Shipp, M.A. *et al.* Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature medicine* **8**, 68-74 (2002).

208. Harwood, N.E. & Batista, F.D. The cytoskeleton coordinates the early events of B-cell activation. *Cold Spring Harbor perspectives in biology* **3**(2011).

209. Monuki, E.S., Kuhn, R. & Lemke, G. Cell-specific action and mutable structure of a transcription factor effector domain. *Proceedings of the National Academy of Sciences* **90**, 9978-9982 (1993).

210. Håvik, B. *et al.* A novel paired domain DNA recognition motif can mediate Pax2 repression of gene transcription. *Biochemical and biophysical research communications* **266**, 532-541 (1999).

211. Yu, Y.T. *et al.* Human myocyte-specific enhancer factor 2 comprises a group of tissue-restricted MADS box transcription factors. *Genes & development* **6**, 1783-98 (1992).

212. Mathey-Prevot, B., Andrews, N.C., Murphy, H.S., Kreissman, S.G. & Nathan, D.G. Positive and negative elements regulate human interleukin 3 expression. *Proceedings of the National Academy of Sciences* **87**, 5046-5050 (1990).

213. Reinhold, W. *et al.* The myc intron-binding polypeptide associates with RFX1 in vivo and binds to the major histocompatibility complex class II promoter region, to the hepatitis B virus enhancer, and to regulatory regions of several distinct viral genes. *Molecular and cellular biology* **15**, 3041-3048 (1995).

214. Furuta, H., Horikawa, Y., Iwasaki, N. & Hara, M. (Beta-cell) transcription factors and diabetes: Mutations in the coding region of the Beta2/NeuroD1 (NEUROD1) and Nkx2. 2 (NKX2B) genes are not associated with maturity-onset diabetes of the young in Japanese. *Diabetes* **47**, 1356 (1998).

215. Monica, K., Galili, N., Nourse, J., Saltman, D. & Cleary, M.L. PBX2 and PBX3, new homeobox genes with extensive homology to the human proto-oncogene PBX1. *Molecular and cellular biology* **11**, 6149-6157 (1991).

216. Lei, X.-H., Shen, X., Xu, X.-Q. & Bernstein, H.S. Human Cdc5, a regulator of mitotic entry, can act as a site-specific DNA binding protein. *Journal of cell science* **113**, 4523-4531 (2000).

217. Foletta, V.C. *et al.* Cloning and characterisation of the mouse fra-2 gene. *Oncogene* **9**, 3305-3311 (1994).

218. Fujitani, Y. *et al.* Identification of a portable repression domain and an E1A-responsive activation domain in Pax4: a possible role of Pax4 as a transcriptional repressor in the pancreas. *Molecular and cellular biology* **19**, 8281-8291 (1999).

219. Quelle, F.W. *et al.* Cloning of murine Stat6 and human Stat6, Stat proteins that are tyrosine

phosphorylated in responses to IL-4 and IL-3 but are not required for mitogenesis. *Molecular and cellular biology* **15**, 3336-3343 (1995).

220. Hsu, H.-L., Wadman, I., Tsan, J.T. & Baer, R. Positive and negative transcriptional control by the TAL1 helix-loop-helix protein. *Proceedings of the National Academy of Sciences* **91**, 5947-5951 (1994).

221. Katan-Khaykovich, Y. & Shaul, Y. Nuclear import and DNA-binding activity of RFX1. *European Journal of Biochemistry* **268**, 3108-3116 (2001).

222. Schmitz, M.L., dos Santos Silva, M.A. & Baeuerle, P.A. Transactivation Domain 2 (TA2) of p65 NF-κB similarity to TA1 and phorbol-ester-stimulated activity and phosphorylation in intact cells. *Journal of Biological Chemistry* **270**, 15576-15584 (1995).

223. Russo, M.W., Sevetson, B.R. & Milbrandt, J. Identification of NAB1, a repressor of NGFI-A-and Krox20-mediated transcription. *Proceedings of the National Academy of Sciences* **92**, 6873-6877 (1995).

224. McPherson, L.A. & Weigel, R.J. AP2α and AP2γ: a comparison of binding site specificity and trans-activation of the estrogen receptor promoter and single site promoter constructs. *Nucleic acids research* **27**, 4040-4049 (1999).

225. Evans, M.J. & Scarpulla, R.C. NRF-1: a trans-activator of nuclear-encoded respiratory genes in animal cells. *Genes & Development* **4**, 1023-1034 (1990).

226. Howard, P.W. & Maurer, R.A. A composite Ets/Pit-1 binding site in the prolactin gene can mediate transcriptional responses to multiple signal transduction pathways. *Journal of Biological Chemistry* **270**, 20930-20936 (1995).

227. Hunger, S.P., Brown, R. & Cleary, M.L. DNA-binding and transcriptional regulatory properties of hepatic leukemia factor (HLF) and the t (17; 19) acute lymphoblastic leukemia chimera E2A-HLF. *Molecular and cellular biology* **14**, 5986-5996 (1994).

228. Kojima, M., Takamatsu, N., Ishii, T., Kondo, N. & Shiba, T. HNF-4 plays a pivotal role in the liver-specific transcription of thechipmunk HP-25 gene. *European Journal of Biochemistry* **267**, 4635-4641 (2000).

229. Igarashi, K. *et al.* Multivalent DNA binding complex generated by small Maf and Bach1 as a possible biochemical basis for β-globin locus control region complex. *Journal of Biological Chemistry* **273**, 11783-11790 (1998).

230. Zheng, N., Fraenkel, E., Pabo, C.O. & Pavletich, N.P. Structural basis of DNA recognition by the heterodimeric cell cycle transcription factor E2F–DP. *Genes & Development* **13**, 666-674 (1999).

231. Zhang, W., Cveklovam, K., Oppermann, B., Kantorow, M. & Cvekl, A. Quantitation of PAX6 and PAX6 (5a) transcript levels in adult human lens, cornea, and monkey retina. *Molecular vision* **7**, 1 (2001).

232. Sasaki, S. *et al.* Cloning and expression of human B cell-specific transcription factor BACH2 mapped to chromosome 6q15. *Oncogene* **19**, 3739-3749 (2000).

233. Wang, Y. *et al.* Activation of ATF6 and an ATF6 DNA binding site by the endoplasmic reticulum stress response. *Journal of Biological Chemistry* **275**, 27013-27020 (2000).

234. Thiagalingam, A. *et al.* RREB-1, a novel zinc finger protein, is involved in the differentiation response to Ras in human medullary thyroid carcinomas. *Molecular and cellular biology* **16**, 5335-5345 (1996).

235. Hsu, H.-L. *et al.* Preferred sequences for DNA recognition by the TAL1 helix-loop-helix proteins. *Molecular and cellular biology* **14**, 1256-1265 (1994).

236. Urizar, N.L., Dowhan, D.H. & Moore, D.D. The farnesoid X-activated receptor mediates bile acid activation of phospholipid transfer protein gene expression. *Journal of Biological Chemistry* **275**, 39313-39317 (2000).

237. Ye, X., Mneina, A., Johnston, J.B. & Mahmud, S.M. Associations between statin use and

non-Hodgkin lymphoma (NHL) risk and survival: a meta-analysis. *Hematological oncology* (2015).