The role of genetic and environmental immune dysregulation in the etiology of childhood acute lymphoblastic leukemia and other complex diseases

By

Amelia Dale Wallace

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Epidemiology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Lisa F. Barcellos, Chair
Professor Sarah Stanley
Professor Catherine Metayer
Professor Steve Selvin

Fall 2017

Abstract

The role of genetic and environmental immune dysregulation in the etiology of childhood acute lymphoblastic leukemia and other complex diseases

Amelia Dale Wallace

Doctor of Philosophy in Epidemiology

University of California, Berkeley

Professor Lisa Barcellos, Chair

Childhood acute lymphoblastic leukemia (ALL) is the most common childhood cancer. Epidemiologic studies suggest that of genetic and environmental factors influencing immune dysregulation play an important role in the etiology of the disease. In the following dissertation, we explore three hypotheses, one established and two nascent, wherein potential immunomodulatory risk factors are tested for association with ALL.

Allergic disease has long been suspected to play a role in the development of childhood ALL. Studies conducted over the last several decades have yielded mixed results. In the first chapter, we examine the association between allergy, a common immune-mediated disorder, and ALL in the California Childhood Leukemia Study (CCLS), a case-control study of 977 children diagnosed with ALL and 1037 matched controls (1995-2015). History of allergies in the first year of life was obtained from interviews, mainly reported by mothers. Logistic regression analyses were conducted to estimate odds ratios (ORs) and 95% confidence intervals (CIs), controlling for birth order, day care attendance, and mode of delivery. In addition, we conducted meta-analyses with data from the CCLS and 12 published studies and employed a new method to estimate between-study heterogeneity (R_b).

Overall, no associations were observed between childhood ALL risk and specific allergy phenotypes or any allergy, as a group. However, having any allergy was associated with an increased risk of ALL among the youngest study participants. In the meta-analysis random-effect models, reduced odds of ALL was associated with hay fever (metaOR=0.65, 95% CI: 0.47, 0.90); however, restricting the analysis to studies that used medical records for assessment of allergy or recently published studies led to null or attenuated results. Overall, our findings do not support a clear association between allergy and childhood ALL. We conclude that the degree to which epidemiological studies can inform the relationship between allergies and risk of childhood ALL is limited by between-study heterogeneity.

In the second chapter, we explore a putative germline genetic risk factor that could contribute to the increased incidence of ALL in Hispanics, who have the highest rate of

disease (4.3 cases per 100,000 per year). Our current understanding of the complex etiology of childhood ALL has failed to explain this disproportionate burden. Recently, one predominant somatic point mutation signature (TpC>T point mutations) related to the innate immune enzyme APOBEC3B was identified in ALL tumor genomes. A common deletion polymorphism in this gene is prevalent in Hispanics and has been associated with an increased risk of other cancers that have the APOBEC3B-related point mutation signature.

We genotyped and tested for association this ~29Kb deletion polymorphism and risk of childhood ALL in 518 cases and 608 controls in the CCLS, where Hispanics account for ~45% of the study population. We found no evidence for germline risk of disease among carriers of the deletion polymorphism, or any SNP in the APOBEC3 gene megalocus in a larger set of 1,083 cases and 1,137 controls. To ensure that ethnic heterogeneity did not mask true associations, local genetic ancestry was inferred with RFMix. Adjustment for local genetic ancestry did not meaningfully alter the observed associations. Further, no specific local genetic ancestry in the APOBEC3 megalocus was independently associated with disease. While somatic mutation induced by the APOBEC3B enzyme may influence tumor progression, there is no evidence that APOBEC3 polymorphisms are associated with disease etiology.

Another leading etiologic hypothesis for ALL related to immune dysregulation states that a specific infectious agent – or mis-timing of common infection – early in life causes the disease. In an untargeted viral discovery study previously carried out in the CCLS, we observed high expression of human endogenous retroviruses (HERVs) in diagnostic bone marrow samples. Approximately 8% of the human genome is comprised of HERVs originating from historic retroviral integration into germ cells. The function of HERVs as regulators of gene expression is well established. Less well studied are insertional polymorphisms of HERVs and their contribution to the heritability of complex phenotypes like cancer. The most recent integration of HERV, HERV-K, is expressed in a range of complex human conditions from cancer to neurologic diseases. In an exploratory study to better understand potential links between HERV-K and ALL, we undertook a phenome-wide association study of HERV-K polymorphisms and present the results in the third and final chapter. Using an in-house computational pipeline and whole-genome sequencing data from the diverse 1000 Genomes Phase 3 population (n=2,504), we identified 48 polymorphic HERV-K insertions that are tagged by adjacent SNPs. To test the potential role of polymorphic HERV-K in the heritability of complex diseases, existing databases were queried for enrichment of established relationships between the HERV-K insertion-associated SNPs (hiSNPs), and tissue specific gene expression and disease phenotypes.

Overall, hiSNPs for the 48 polymorphic HERV-K sites were statistically enriched ($p<1.0E^{-16}$) for eQTLs across 44 human tissues. Fifteen of the 48 HERV-K insertions had hiSNPs annotated in the EMBL-EBI GWAS Catalog and cumulatively associated with >100 phenotypes. Experimental factor ontology enrichment analysis suggests that polymorphic HERV-K specifically contribute to neurologic and immunologic disease phenotypes, including traits related to intracranial volume (FDR 2.00E-09), Parkinson’s disease (FDR 1.80E-09), and autoimmune diseases (FDR 1.80E-09). These results provide strong candidates for context-specific study of polymorphic HERV-K insertions in disease-related traits and observed enrichment in immune-related traits aw well as virally induced cancers warrant further study in ALL, despite identifying no existing associations specifically with the disease.

Overall, these studies make an important contribution to our understanding of the immune dysregulation that precedes childhood-onset ALL. Specifically, this work has lain to rest existing hypotheses while at the same time generating new ones. Future research building off this work will bring us closer to effective prevention strategies for this devastating disease.

To my parents, Donald and Deborah Wallace, for always encouraging me to follow my passion, wherever it might lead – I am so proud that this is where has taken me.

Table of Contents

I. Preface

I.i Childhood Acute Lymphoblastic Leukemia, A Cancer Born of Immune Dysregulation

Cancer is the second leading cause of death among children in the United States. Leukemia comprises the majority of childhood cancer, with more than 50,000 children diagnosed worldwide per year. Acute lymphoblastic leukemia (ALL) accounts for 80% of these cases[1]. Although the five-year survival rate for childhood ALL now exceeds 85%, survivors often suffer long-term medical comorbidities and neuropsychological complications as a result of treatments[2]. Thus, prevention remains an important goal.

ALL is a cancer of lymphoid lineage progenitor hematopoietic cells with the vast majority occurring in pre-B cells, whereas pre-T cell leukemia is less frequent (~13% of all childhood ALL cases)[3]. While most other childhood cancers demonstrate a low and stable incidence rate from infancy to age 18, ALL produces a steep epidemic curve that spikes from ages 2-6 and then drops down to very low rates lasting until elder adulthood[4]. The incidence of ALL is slightly higher in boys than girls (1.27 times the rate) and demonstrates a differential incidence by race and ethnicity, where Hispanics experience the highest risk (4.29 cases/100,000) in the US and African Americans experience the lowest (1.87 cases/100,000)[4,5].

The etiology of ALL is unknown, except in specific cases where ionizing radiation or genetic syndromes can cause the disease[6,7]. Early life immune dysregulation is thought to play a predominant role. The current paradigm for immune function posits that the immune system is activated in response to infection, and relaxed during periods of homeostasis[8]. Immune dysregulation can be triggered by endogenous and exogenous mechanisms. Disruption of this balance is known to cause a myriad of diseases[9-12], and we hypothesize a role for this mechanism in childhood leukemia.

I.ii The Human Immune System, Homeostasis, and Disease

The human immune system is a powerful mediator of both health and disease. The immune response is one of the most, if not *the* most important evolutionary adaptation in humans. It is critical for securing the fitness of any organism through survival to reproductive age. It is a continually evolving mechanism to protect against exogenous infection. Both the innate and adaptive arms of the immune system wield powerful, destructive tools to eliminate the vast genera of human parasites, from viruses to helminths[13]. Collateral damage to human tissues is prevented by tight regulation of these mechanisms. Maintaining an equilibrium, wherein the immune response is strong enough to fight off infection, but not so strong or non-specific as to cause damage to host tissues, is critical for maintaining health.

Dysregulation resulting in insufficient immunity is most often associated with infection. For example immune suppression can create susceptibility to opportunistic infection by microorganisms that normally colonize healthy individuals (e.g. yeast infections) or reactivation of latent, persistent infection, for example Varicella-Zoster, the chicken pox virus, causing shingles in the elderly[14]. More recently, it has been suggested that insufficient immune response could result in cancer, indeed most so called cancer causing genes (p53, Rb, etc.) are immune related genes[15]. The tumor surveillance hypothesis suggests that the immune system serves as a first-line of defense against early tumor cells, eliminating them before uncontrolled replication yields cancer[16]. Evidence of tumor surveillance exists in both mouse models and human studies. Several knockout experiments of key immune genes in mice suggest that NK cells, T cells, interferons, and other immune molecules are key in eliminating precancerous cells[17].

Conversely, hyperactive immunity can result in a range of complex diseases from allergy and asthma, to autoimmune diseases, to chronic inflammatory conditions, such as liver cirrhosis. Some of the most critical effector mechanisms of the adaptive immune system require the ability to differentiate self from non-self; and amongst non-self antigens, the benign from the pathogenic. A common precursor to immune-mediated complex disease is the failure to make these important distinctions. For example, autoimmune diseases are thought to arise from an inappropriate immune response to self-antigen, resulting in the destruction of healthy tissues. Allergies or Type I hypersensitivities, result from mounting an antibody-mediated immune response to benign, non-self stimuli, such as pollen or dust. A hyperactive immune response is also hypothesized to be causally related to some cancers via chronic inflammation, for example *Helicobacter pylori* associated gastric cancers and hepatitis-associated hepatocellular carcinomas[18,19].

Unobstructed and delicately balanced signaling feedback loops maintain immunologic homeostasis in healthy individuals. For example, typically, immunosuppressive pathways are activated following successful elimination of infections and self-reactive lymphocytes are destroyed via central and peripheral tolerance mechanisms. Immune dysregulation occurs when these pathways are interrupted and the causes of immune dysregulation can be endogenous or exogenous.

I.iii Endogenous Immune Dysregulation

Inherited and spontaneous genetic mutations can be an important endogenous source of immune dysregulation. For example, individuals born with loss of function mutations in the *IL2RG* gene are unable to activate lymphocytes, resulting in X-linked severe combined immunodeficiency (XSCID). Without a functioning immune system, individuals with XSCID succumb to infection in infancy and, left untreated, die by the age of two[20]. More nuanced examples of genetic immune dysregulation can be found in autoimmune diseases, most of which have a strong genetic component. The strongest genetic signals for the heritability of this class of diseases are in variants that encode specific alleles of human leukocyte antigen (HLA) genes, the most complex region of the human genome. HLA molecules are responsible for presenting antigen to the adaptive immune system, thus activating it in the presence of infection. The mechanism by which HLA alleles increase risk of autoimmunity is currently unknown, however it is thought that these variants facilitate the presentation of- and adaptive response to self-antigen[9].

I.iv Exogenous Immune Dysregulation

Immune dysregulation can also result from exogenous environmental exposures. Occupational cohort studies have shown that exposure to common herbicides used in agriculture (organo-phosphates and carbamates) decrease effector immune cell responses and increase risk of atopic and autoimmune diseases[21]. Cigarette smoke is a pervasive toxic exposure whose immunomodulatory effects are often overlooked. Studies in mouse models show that dendritic cells and T cells exert suppressed responses to *ex vivo* bacterial and viral antigen stimulation following acute exposure to cigarette smoke[22,23]. Viruses are another fascinating example of exogenous entities with highly evolved immunomodulatory mechanisms. Cytomegalovirus (CMV) has the largest genome of any human virus and even encodes a viral version of the human immunosuppressive cytokine, IL-10. By interrupting antigen presentation pathways and expressing *cmv*IL-10, lytic CMV is able to establish lifelong infection in most individuals[24].

I.v Immune Dysregulation in Childhood ALL

Acute lymphoblastic leukemia (ALL) is a cancer of a central component of the immune system– lymphocytes. Epidemiologic evidence strongly suggests that immune dysregulation precedes disease, however, the endogenous and/or exogenous triggers of dysregulation are poorly understood. Immune dysregulation among children who go on to develop ALL may, for some, begin *in utero.* A study published within our group in 2011 shows lower levels of five cytokines measured in neonatal (i.e. pre-diagnostic) blood samples from ALL cases compared to healthy controls. After adjustment in multivariable models, the strongest deficit was observed in IL-10, an immunosuppressive cytokine[25].

Further, the infection histories of children who develop ALL differ significantly from healthy controls. *In utero* infection with cytomegalovirus (CMV), a common herpes virus that is the leading infectious cause of birth defects, is nearly 4 times as likely to occur in children who will develop ALL than in healthy children[26]. Two large registry-based studies in Taiwan and the United Kingdom showed that after birth, children that develop ALL have more medically diagnosed infections than healthy children. Specifically, having any medically diagnosed infection occurring at least one year before ALL onset increases risk of disease ~4 fold[27]. This association hints at an abnormal immune system that either allows more infections, more fulminant infections, or both in children that go on to develop ALL. Finally, in a study that we conducted using diagnostic bone marrow samples, we observed a disease-specific profile of expressed viral transcripts at the time of hospital admittance compared to children with acute myeloid leukemia, which is thought to have an etiology distinct from ALL[26].

ALL tumor cells themselves bare hallmarks of immune dysregulation. Next generation sequencing studies of tumors from the most common ALL subtype, *ETV6-RUNX1* ALL [also called t(12:21) or *TEL-AML1* fusion], have shed light on somatic mutational patterns that allude to immune-mediated molecular mechanisms of mutagenesis. Specifically, the tumor cells display two major patterns of somatic mutation. The first is a pattern of genome-wide deletions that harbor recognition signal sequences at their breakpoints characteristic of Recombination Activating Gene (RAG1/2)-mediated non-homologous end joining[28]. RAG1/2 is expressed in developing lymphocytes and is necessary for V(D)J recombination, which creates variation in the B- and T-cell receptor repertoires. The second is a point mutation signature, specifically a clustered pattern of TpC>T point mutations, unique to a family of cytidine deaminases whose members include Activation Induced Deaminase (AID) and APOBECs (Apolipoprotein B mRNA-editing enzyme, catalytic polypeptide-like), important in adaptive and innate (respectively) defense against infection[28]. These signatures are the primary somatic mutational signatures in ALL, suggesting a critical role of RAG1/2 and APOBEC in the development of ALL. However, a study of 51 high hyperdiploid subtype ALL tumors did not identify these signatures, suggesting that they are not ubiquitous across subtypes[29].

I.vi Endogenous Sources of Immune Dysregulation in Childhood ALL

Approximately 25% of childhood ALL cases are the aforementioned *ETV6-RUNX1* cytogenetic subtype, which is hallmarked by a chromosomal translocation at t(12:21) that results in a fusion of two genes (*ETV6* and *RUNX1)* and the expression of a chimeric transcription factor[30]. Surveys in population-based samples of cord blood collected at birth have shown that the t(12:21) translocation occurs prenatally, and is not fully penetrant, occurring ~100 times more frequently than ALL itself. This chromosomal anomaly has been shown to disrupt normal hematopoiesis[28]. By binding sets of promoters distinct from either of its parent transcription factors, cells containing the translocation have distinctly different expression profiles, which have been shown to stall development of pre-B cells and alter activity of TGFβ, an important regulator of immunity[30,31].

Common heritable genetic variation has also been associated with ALL in genome-wide association studies, yet there is no clear link to immune dysregulation. To date, six established germline single nucleotide polymorphisms (SNPs) are associated with disease and lie in or near the following genes: *IKZF1, ARID5B, GATA3, CDKN2A, CEBPE,* and *PIP4K2A* [32]. Together these SNPs account for ~8% of the estimated variability in risk attributable to germline genetics[33]. Global Native American genetic ancestry has also been associated with an increased risk of ALL, which may in part explain the increased incidence of disease among Hispanics[33]. Interestingly, the functional effects of these variants on ALL susceptibility are not well understood. One exception is an ALL risk SNP on chromosome 7 lies in region of transcriptional regulation for the *IKZF1* gene, which codes for an important regulator of hematopoiesis and immune function[34]. CEBPE and GATA3 may also contribute to lymphocyte development[35,36] but functional studies will be required to determine the specific effects of these genetic variants on immune dysregulation and ALL. Candidate studies of variants in specific immune genes further suggest that there are endogenous sources of immune dysregulation, including in the KIR locus, which encodes activating and inhibitory NK and T cell receptors[37]. Finally, Down syndrome, a congenital condition resulting from trisomy of the 21st chromosome, is associated with ~20-fold increased risk of ALL[38], and children with Down syndrome have marked immune dysregulation including abnormal white blood cells counts and increased susceptibility to autoimmune disease and infection[39].

I.vii Exogenous Sources of Immune Dysregulation in Childhood ALL

Many environmental factors that modify immune function have been associated with risk of ALL. There is strong evidence suggesting factors that moderate ‘normal’ exposure to infections confer protection from ALL including vaginal birth, having older siblings, breastfeeding >6 months, and attending daycare[40-42]. This evidence, which collectively suggests that mis-timing of exposure to common infection increases risk of ALL has precipitated two leading etiologic hypotheses for the disease: Mel Greaves’ Delayed Infection hypothesis and Leo Kinlin’s Population Mixing hypothesis[43,44]. Both of these hypotheses posit that improper immune priming via either delayed exposure to infection akin to the hygiene hypothesis for allergy and autoimmune disease (Delayed Infection), or introduction of a novel agent by migration (Population Mixing), results in a dysregulated immune reaction and leukemia in susceptible individuals.

Risk factors for ALL in which immunotoxicity is a plausible mechanism include paternal cigarette smoking (Odds Ratio 1.25) and prenatal occupational pesticide exposure (Odd Ratio 2.09)[45,46]. Beyond direct accumulation of mutations in germ cells, *in utero* cigarette smoke exposure is hypothesized to trigger developmental immunotoxicity. Studies using mouse models demonstrate that prenatal smoke exposure results in reduced cytoxic T-lymphocyte lytic activity in offspring[47]. Independent functional studies of common agricultural chemicals have shown that several classes have immunotoxic effects[21]. Whether these are the chemicals are specifically driving the association with ALL is yet to be determined.

I.viii The Role of Reverse Causality, Heterogeneity, and Bias

Several factors complicate our ability to delineate the sources of immune dysregulation that contribute to ALL etiology. First, for several of the aforementioned factors associated with disease, it cannot be determined whether they are a cause of immune dysregulation, a product of it, or both. In daycare attendance, for example, a healthy, but ALL-susceptible child who does not attend daycare at a young age may not be exposed to infection via normal child-to-child contact, potentially resulting in immune dysregulation and disease. On the other hand, a child that does not attend daycare may already have dysregulated immunity and be too sick to be placed in daycare. In this case, daycare attendance may prevent immune dysregulation, or it may simply act as a marker for healthier children. Similarly, congenital CMV infection and early childhood infections may be markers of immune dysregulation, causes, or both. The directed acyclic graph below (Figure 1) highlights the complexity of the possible causal relationships between the immunomodulatory risk factors for ALL.

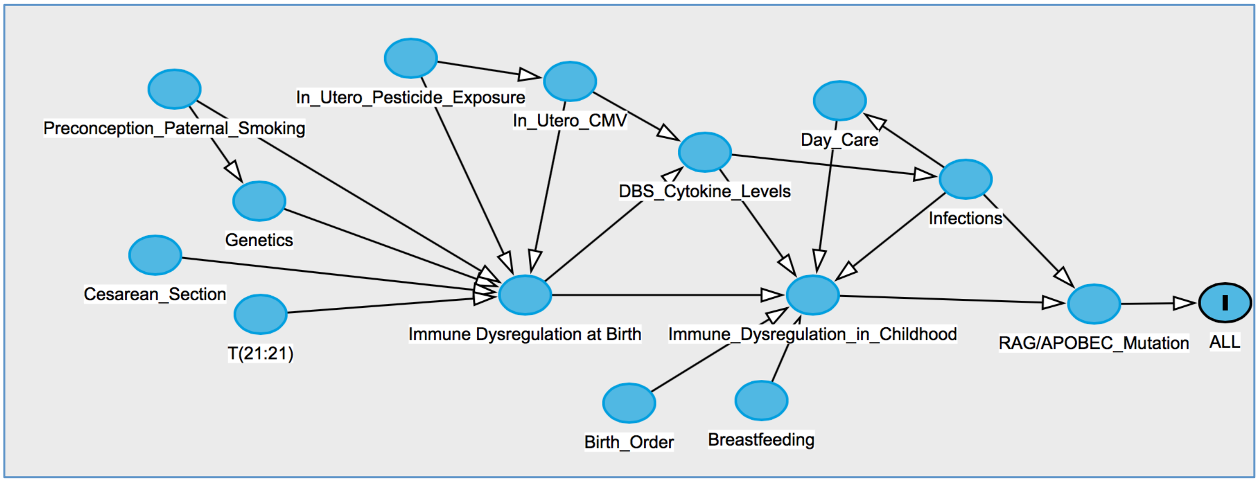


Figure I.viii.i: Directed acyclic graph showing the hypothesized mediation of many established risk factors for ALL by immune dysregulation at birth and in childhood.

Establishing a causal mechanism for diseases is further complicated by the fact that ALL is heterogeneous and can be divided into distinct subtypes. As mentioned above, only the ETV6-RUNX1 ALL subtype tumors are known to have somatic mutations providing direct evidence of immune dysregulation, though studies in other subtypes are currently underway. ALL is already an extremely rare disease, and thus most molecular and epidemiologic studies including those identifying the aforementioned disease risk factors are often underpowered to explore heterogeneity by cytogenetic subtypes. Rarity also limits the design of epidemiologic studies of ALL to case-control studies, where exposure to potential causal agents must be recalled by parents, sometimes years later. Thus it is possible that bias and/or confounding produces observed associations in some studies, which may explain a lack of consistency in direction and magnitude of exposure effects across studies. As such, we often must rely on systematic reviews and meta-analyses to draw clear inference from the body of available evidence.

In the first two chapters of this dissertation, two mechanisms of immune dysregulation are investigated as potential risk factors for childhood ALL within the California Childhood Leukemia Study (CCLS), a diverse study population capturing the majority of childhood leukemia cases occurring in California over the past 20 years. In the first chapter, we examine an exogenous immunomodulatory exposure, allergy, which has been of etiologic interest in ALL for decades, as measured by self-report in both the CCLS population and then via a comprehensive meta-analysis spanning 13 studies. In the second chapter, we evaluate an endogenous genetic variant (*APOBEC3B* deletion) that produces a modified innate immune effector molecule, which has been previously associated with the somatic mutational signature that is predominate in some ALL tumor genomes and other cancer types. Finally, in Chapter 3, we present a study of a previously unrecognized class of unique polymorphic genetic features, human endogenous retroviruses, that have the ability to modify immune function through transcriptional regulation and interaction with exogenous viruses, for associations multiple complex diseases including childhood ALL.

I.ix References

1 Group, U. S. C. S. W. *United States Cancer Statistics: 1999–2010 Incidence and Mortality Web-based Report*, < http://www.cdc.gov/uscs> (2013).

2 Cancer incidence in five continents. Volume IX. *IARC Sci. Publ.*, 1-837 (2008).

3 Pui, C. H., Behm, F. G. & Crist, W. M. Clinical and biologic relevance of immunologic marker studies in childhood acute lymphoblastic leukemia. *Blood* **82**, 343-362 (1993).

4 Ries LAG, S. M., Gurney JG, Linet M, Tamra T, Young JL, Bunin GR (eds). Cancer Incidence and Survival among Children and Adolescents: United States SEER Program 1975-1995, National Cancer Institute, SEER Program. *NIH Pub* **No. 99-4649** (1999).

5 Siegel, D. A. *et al.* Rates and Trends of Pediatric Acute Lymphoblastic Leukemia - United States, 2001-2014. *MMWR Morb. Mortal. Wkly. Rep.* **66**, 950-954, doi:10.15585/mmwr.mm6636a3 (2017).

6 Yokota, T. & Kanakura, Y. Genetic abnormalities associated with acute lymphoblastic leukemia. *Cancer Sci.* **107**, 721-725, doi:10.1111/cas.12927 (2016).

7 Schmitz-Feuerhake, I. *et al.* Leukemia in the proximity of a German boiling-water nuclear reactor: evidence of population exposure by chromosome studies and environmental radioactivity. *Environ. Health Perspect.* **105 Suppl 6**, 1499-1504 (1997).

8 Marques, R. E., Marques, P. E., Guabiraba, R. & Teixeira, M. M. Exploring the Homeostatic and Sensory Roles of the Immune System. *Front. Immunol.* **7**, 125, doi:10.3389/fimmu.2016.00125 (2016).

9 Kuchroo, V. K., Ohashi, P. S., Sartor, R. B. & Vinuesa, C. G. Dysregulation of immune homeostasis in autoimmune diseases. *Nat. Med.* **18**, 42-47, doi:10.1038/nm.2621 (2012).

10 Chen, L. *et al.* The biomarkers of immune dysregulation and inflammation response in Parkinson disease. *Transl Neurodegener* **5**, 16, doi:10.1186/s40035-016-0063-3 (2016).

11 Deleidi, M., Jaggle, M. & Rubino, G. Immune aging, dysmetabolism, and inflammation in neurological diseases. *Front. Neurosci.* **9**, 172, doi:10.3389/fnins.2015.00172 (2015).

12 Castle, S. C. Clinical relevance of age-related immune dysfunction. *Clin. Infect. Dis.* **31**, 578-585, doi:10.1086/313947 (2000).

13 Simon, A. K., Hollander, G. A. & McMichael, A. Evolution of the immune system in humans from infancy to old age. *Proc. Biol. Sci.* **282**, 20143085, doi:10.1098/rspb.2014.3085 (2015).

14 Chen, S. Y. *et al.* Incidence of herpes zoster in patients with altered immune function. *Infection* **42**, 325-334, doi:10.1007/s15010-013-0550-8 (2014).

15 Munoz-Fontela, C., Mandinova, A., Aaronson, S. A. & Lee, S. W. Emerging roles of p53 and other tumour-suppressor genes in immune regulation. *Nat. Rev. Immunol.* **16**, 741-750, doi:10.1038/nri.2016.99 (2016).

16 Ribatti, D. The concept of immune surveillance against tumors. The first theories. *Oncotarget* **8**, 7175-7180, doi:10.18632/oncotarget.12739 (2017).

17 Swann, J. B. & Smyth, M. J. Immune surveillance of tumors. *J. Clin. Invest.* **117**, 1137-1146, doi:10.1172/JCI31405 (2007).

18 Parsonnet, J. *et al.* Helicobacter pylori infection and the risk of gastric carcinoma. *N. Engl. J. Med.* **325**, 1127-1131, doi:10.1056/NEJM199110173251603 (1991).

19 But, D. Y., Lai, C. L. & Yuen, M. F. Natural history of hepatitis-related hepatocellular carcinoma. *World J. Gastroenterol.* **14**, 1652-1656 (2008).

20 Allenspach, E., Rawlings, D. J. & Scharenberg, A. M. in *GeneReviews(R)* (eds M. P. Adam *et al.*) (1993).

21 Corsini, E., Sokooti, M., Galli, C. L., Moretto, A. & Colosio, C. Pesticide induced immunotoxicity in humans: a comprehensive review of the existing evidence. *Toxicology* **307**, 123-135, doi:10.1016/j.tox.2012.10.009 (2013).

22 Dietert, R. R. Developmental immunotoxicity (DIT), postnatal immune dysfunction and childhood leukemia. *Blood Cells Mol. Dis.* **42**, 108-112, doi:10.1016/j.bcmd.2008.10.005 (2009).

23 Selgrade, M. K. Immunotoxicity: the risk is real. *Toxicol. Sci.* **100**, 328-332, doi:10.1093/toxsci/kfm244 (2007).

24 Ressing, M. E. *et al.* Epstein-Barr virus evasion of CD8(+) and CD4(+) T cell immunity via concerted actions of multiple gene products. *Semin. Cancer Biol.* **18**, 397-408, doi:10.1016/j.semcancer.2008.10.008 (2008).

25 Wiemels, J. L. *et al.* Prenatal origin of acute lymphoblastic leukaemia in children. *Lancet* **354**, 1499-1503 (1999).

26 Francis, S. S. *et al.* In utero cytomegalovirus infection and development of childhood acute lymphoblastic leukemia. *Blood* **129**, 1680-1684, doi:10.1182/blood-2016-07-723148 (2017).

27 Chang, J. S., Tsai, C. R., Tsai, Y. W. & Wiemels, J. L. Medically diagnosed infections and risk of childhood leukaemia: a population-based case-control study. *Int. J. Epidemiol.* **41**, 1050-1059, doi:10.1093/ije/dys113 (2012).

28 Papaemmanuil, E. *et al.* RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia. *Nat. Genet.* **46**, 116-125, doi:10.1038/ng.2874 (2014).

29 Paulsson, K. *et al.* The genomic landscape of high hyperdiploid childhood acute lymphoblastic leukemia. *Nat. Genet.* **47**, 672-676, doi:10.1038/ng.3301 (2015).

30 Linka, Y. *et al.* The impact of TEL-AML1 (ETV6-RUNX1) expression in precursor B cells and implications for leukaemia using three different genome-wide screening methods. *Blood Cancer J* **3**, e151, doi:10.1038/bcj.2013.48 (2013).

31 Torrano, V., Procter, J., Cardus, P., Greaves, M. & Ford, A. M. ETV6-RUNX1 promotes survival of early B lineage progenitor cells via a dysregulated erythropoietin receptor. *Blood* **118**, 4910-4918, doi:10.1182/blood-2011-05-354266 (2011).

32 Moriyama, T., Relling, M. V. & Yang, J. J. Inherited genetic variation in childhood acute lymphoblastic leukemia. *Blood* **125**, 3988-3995, doi:10.1182/blood-2014-12-580001 (2015).

33 Walsh, K. M. *et al.* Associations between genome-wide Native American ancestry, known risk alleles and B-cell ALL risk in Hispanic children. *Leukemia* **27**, 2416-2419, doi:10.1038/leu.2013.130 (2013).

34 Payne, K. J. & Dovat, S. Ikaros and tumor suppression in acute lymphoblastic leukemia. *Crit. Rev. Oncog.* **16**, 3-12 (2011).

35 Wiemels, J. L. *et al.* A functional polymorphism in the CEBPE gene promoter influences acute lymphoblastic leukemia risk through interaction with the hematopoietic transcription factor Ikaros. *Leukemia* **30**, 1194-1197, doi:10.1038/leu.2015.251 (2016).

36 Zhu, J., Yamane, H. & Paul, W. E. Differentiation of effector CD4 T cell populations (*). *Annu. Rev. Immunol.* **28**, 445-489, doi:10.1146/annurev-immunol-030409-101212 (2010).

37 de Smith, A. J. *et al.* The role of KIR genes and their cognate HLA class I ligands in childhood acute lymphoblastic leukemia. *Blood* **123**, 2497-2503, doi:10.1182/blood-2013-11-540625 (2014).

38 Hitzler, J. K. & Zipursky, A. Origins of leukaemia in children with Down syndrome. *Nat. Rev. Cancer* **5**, 11-20, doi:10.1038/nrc1525 (2005).

39 Kusters, M. A., Verstegen, R. H., Gemen, E. F. & de Vries, E. Intrinsic defect of the immune system in children with Down syndrome: a review. *Clin. Exp. Immunol.* **156**, 189-193, doi:10.1111/j.1365-2249.2009.03890.x (2009).

40 Ma, X. *et al.* Daycare attendance and risk of childhood acute lymphoblastic leukaemia. *Br. J. Cancer* **86**, 1419-1424, doi:10.1038/sj.bjc.6600274 (2002).

41 Francis, S. S. *et al.* Mode of delivery and risk of childhood leukemia. *Cancer Epidemiol. Biomarkers Prev.* **23**, 876-881, doi:10.1158/1055-9965.EPI-13-1098 (2014).

42 Westergaard, T. A., P.; Pedersen, J.; Olsen, J.; Frisch, M.; Sorensen, H.; Wohlfahrt, J.; Melbye, M.;. Birth Characteristics, Sibling Patterns, and Acute Leukemia Risk in Childhood: a Population-Based Cohort Study. *J. Natl. Cancer Inst.* **89**, 939-947 (1997).

43 Greaves, M. Infection, immune responses and the aetiology of childhood leukaemia. *Nat. Rev. Cancer* **6**, 193-203, doi:10.1038/nrc1816 (2006).

44 Kinlen, L. J. & Stiller, C. Population mixing and excess of childhood leukemia. *BMJ* **306**, 930 (1993).

45 Liu, R., Zhang, L., McHale, C. M. & Hammond, S. K. Paternal smoking and risk of childhood acute lymphoblastic leukemia: systematic review and meta-analysis. *J. Oncol.* **2011**, 854584, doi:10.1155/2011/854584 (2011).

46 Wigle, D. T., Turner, M. C. & Krewski, D. A systematic review and meta-analysis of childhood leukemia and parental occupational pesticide exposure. *Environ. Health Perspect.* **117**, 1505-1513, doi:10.1289/ehp.0900582 (2009).

47 Ng, S. P., Silverstone, A. E., Lai, Z. W. & Zelikoff, J. T. Effects of prenatal exposure to cigarette smoke on offspring tumor susceptibility and associated immune mechanisms. *Toxicol. Sci.* **89**, 135-144, doi:10.1093/toxsci/kfj006 (2006).

II. Acknowledgements

First I would like to thank my mentors, Dr. Lisa Barcellos, Dr. Catherine Metayer, and Dr. Joseph Wiemels for their guidance and support throughout this entire process and specifically Dr. Barcellos for her unwavering enthusiasm and for pushing me as a scientist; Dr. Metayer for providing data, a workspace, financial support, and for lighting the occasional fire under me exactly when it was needed; and Dr. Wiemels for providing scientific mentorship and a unique perspective on science and academia.

Second, I would like to thank all of the co-authors of this work for their feedback and contributions. In particular I would like to thank Dr. Stephen Francis and Jake Wendt for affording me the opportunity to work as a member of their team.

Next I would like to acknowledge all faculty and staff in Epidemiology at the School of Public Health for their hard work in providing students an incredible education; and my classmates – the support that they provide for each other makes this program a great community to be a part of. I want to specifically thank soon-to-be Dr. Jennifer Ames for her feedback and support throughout this process.

Thank you to the University of California Cancer Research Coordinating Committee for proving financial support during the 2016-2017 academic year; and to the Epidemiology department and executor of the Patricia A Buffler scholarship award for additional financial support.

This work could not have been possible without the support and resources from all members of the Integrative Cancer Research Group (specifically Alice Kang, Libby Morimoto, Todd Whitehead, Amanda Singer, John Nides, Steve Selvin, and Pagan Morris); the Genetic Epidemiology and Genomics Laboratory (specifically Milena Gianfrancesco, Giovanna Cruz, Brooke Rhead, Xiaorong Shao, Calvin Chi, Cameron Adams, Mary Horton, Hong Quach, Diana Quach, Gary Artim, and Indro Fedrigo); and the laboratory of Dr. Lee Riley (specifically Dr. Riley for providing lab space, support, and encouragement to pursue epidemiology, Robbie Snyder, Melaine Delcroix, and Sheila Adams-Sapper).

Finally, I would like to thank my dissertation committee members: Dr. Lisa Barcellos, Dr. Catherine Metayer, Dr. Steve Selvin, and Dr. Sarah Stanley for their guidance.

1. Allergy and Risk of Childhood Acute Lymphoblastic Leukemia

1.1 Allergy as a manifestation or cause of immune dysregulation

Having a greater number of medically diagnosed early childhood infections is a hallmark of immune dysregulation in children who go on to develop acute lymphoblastic leukemia (ALL)[1]. Atopic allergic disease could similarly be considered a marker of dysregulated immunity; one that is common among children in the United States and with increasing incidence[2]. Atopic allergies, also called Type-I hypersensitivities, represent a propensity for inappropriate B-cell activation and IgE antibody production against innocuous antigens. The epidemiology of allergy and ALL have striking similarities wherein exposures that modify timing of infections in early childhood act as risk factors in both diseases (e.g. birth by cesarean section, lower birth order, and absence from day care)[3,4]. In fact, the hygiene hypothesis (described in detail in the section 1.2) for the etiology of allergic disease in part motivates Greaves' Delayed-Infection hypothesis for the etiology of childhood ALL. The similarity in risk factors and the engagement of B-lineage lymphocytes in both diseases has led to extensive study of the relationship between them.

Allergy and ALL both appear rooted in immune dysregulation, but is there a causal relationship between the two diseases? The series of directed acyclic graphs below explore possible causal relationships. Like daycare and medically diagnosed infection histories, allergy could be a product of immune dysregulation in children who go on to develop ALL, which is causally unrelated to ALL but serves as proxy evidence for underlying causal mechanisms shared by both diseases (Figure 1.1.1A). Alternatively, allergy may contribute to immune dysregulation through pathways of chronic immune activation or other mechanisms, thus playing a causal role in disease (Figure 1.1.1B).

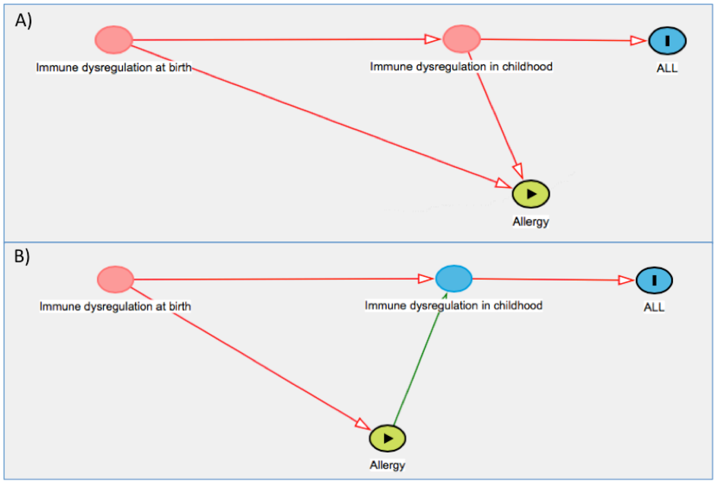

A)
Immune dysregulation at birth
Immune dysregulation in childhood
ALL
Allergy
B)
Immune dysregulation at birth
Immune dysregulation in childhood
ALL
Allergy

Figure 1.1.1: Directed acyclic graph representing hypothesized causal relationships between allergy and ALL. A) Allergy is an independent product of immune dysregulation and occurrence

of allergy and ALL may correlate if mechanisms are shared B) Allergy contributes to ALL through exacerbation of immune dysregulation.

Unfortunately, despite the fact that several epidemiologic studies have been conducted to nail down the magnitude and direction of the allergy-ALL relationship, results are mixed. Elucidating the causal relationship between the two diseases first requires a clear understanding of the association between them. If occurrence of allergy in early childhood is inversely associated with ALL, that would suggest that the immunology underlying allergy, or having allergy itself, is protective against childhood ALL. On the other hand, if the two diseases are correlated, that might suggest a shared underlying mechanism of pathogenesis.

The studies that have strived to tease apart the nature of the relationship between these two diseases generate heterogeneous results with the strength and direction of association possibly influenced by study design, exposure assessment, and source of controls. Especially in case-control studies where allergy exposure in early life is recalled sometimes years after the leukemia diagnosis (or reference date for controls), estimates of the association are subject to bias and uncontrolled confounding. In the California Childhood Leukemia Study (CCLS), one of the largest samples in the US with high Hispanic representation (46%), we have collected self-reported allergy to various allergens in ~2600 cases and controls. In the following article, we incorporate the results of an analysis of self-reported allergy and ALL within the CCLS into a larger meta-analysis in an attempt to clarify this association. Identifying a clear association could motivate functional studies focused on mechanistic connections to elucidate causal relationship between allergy and ALL, if one exists.

1.2 ARTICLE: Allergies and childhood acute lymphoblastic leukemia: A case-control study and meta-analysis

INTRODUCTION

Childhood acute lymphoblastic leukemia (ALL) is a cancer of lymphoblasts, which are precursor immune cells. ALL is thought to arise through immune dysregulation, beginning *in utero* and extending throughout early life [5]. Because allergies are also believed to result from aberrant immune function (i.e. an immune response to innocuous antigen), they have long been suspected as potential causes of childhood ALL [4]. In the early 1950s, a case report suggested allergy could cause acute bone marrow injury, resulting in childhood leukemia [6]. This report precipitated decades of research on the subject, ultimately yielding inconsistent results.

Epidemiologic evidence of shared risk factors between allergic disease and ALL as well as coinciding upward trends in their incidence suggest a relationship between the two conditions [2,7]. Surrogates of early life exposure to normal antigenic challenge, including a higher birth order [8,9], daycare attendance [10,11], and vaginal birth [12,13], are protective against both ALL and allergy. In fact, these observations form the basis of the leading etiologic hypothesis for both diseases, the so-called “hygiene hypothesis”.

The hygiene hypothesis postulates that a lack of early-life exposure to normal immune challenges leads to aberrant immune system responses later in life. This hypothesis was initially employed to suggest that improvements to sanitation in the developed world during the 20th century led to the emergence of allergic diseases [3]. Subsequently, Greaves *et al.* adapted this hypothesis by proposing that "delayed infection" early in childhood, followed by a hyper-responsiveness to infectious challenge later in childhood, generates the genetic mutations necessary to develop childhood ALL [14]. The hypothesis implies that the two diseases – allergy and ALL – have related biological mechanisms.

Two separate but overlapping meta-analyses by Dahl *et al.* in 2009 [15] and Linabery *et al.* in 2010 [16] have been conducted on allergy and the risk of childhood ALL. Though the quantitative results of the two meta-analyses were nearly identical, disparate interpretations of the results by the two author teams (suggestion of causation and bias, respectively) leave the existence of a relationship between allergy and ALL in question. Here, we examine the association between allergy in the first year of life and the risk of ALL in children enrolled in the California Childhood Leukemia Study (CCLS). In addition, we update previous meta-analyses with four new epidemiologic studies of allergy and childhood leukemia published since 2010 [17-20] including results from the present study, increasing the previous sample from 30,763 to 40,123 children and including the largest medical record-based study conducted to date on this topic. Further, we provide a statistical evaluation of factors contributing to between-study heterogeneity using meta-regression. To estimate the percent variance due to between-study heterogeneity, we implement a new method with an alternative estimator that does not violate the assumption of constant within study variance, which is required for the I^2% measure used in previous meta-analyses and is almost never met [21]. Our methods and the addition of important new studies allow us to comprehensively evaluate the association of this challenging exposure with ALL risk.

MATERIALS AND METHODS

CCLS analysis

Study population. The CCLS is a case-control study conducted in California between 1995 and 2015 [22]. Case ascertainment was rapid, usually within 72 hours from diagnosis, and ~84% of incident childhood leukemia cases were consented from participating hospitals. Eligibility criteria for cases and controls were as follows: 1) living in the study area at the time of recruitment; 2) having an English or Spanish-speaking parent; 3) being less than 15 years old at diagnosis or reference date for controls; and 4) having no previous diagnosis of cancer. From 1995 to 2008, one or two controls were selected using birth certificate information from the Office of Vital Records, and individually matched to cases on date of birth, sex, child's Hispanic ethnicity, and maternal race. No controls were enrolled from 2009 to 2015, and the overall participation rate was ~68%. A summary of participation rates in the CCLS by study phase is summarized in Supplemental Table S1. The University of California, Berkeley's Institutional Review Board, California Department of Health Services, and all participating clinical institutions approved the study. Informed consent was obtained from all study participants.

Sample size. A total of 1477 leukemia cases and 1226 controls were eligible to participate and interviewed. Participants with non-ALL leukemia (n=40), missing exposure or covariate data (n=252) or Down syndrome (n=57) were excluded. To ensure that all allergies occurred prior to the leukemia prodrome, participants whose date of diagnosis or reference date was less than 18 months after birth were excluded (n=153), leaving a minimum time interval of at least 6 months between potential exposure (allergy in the first year of life) and disease diagnosis. The final sample size was 977 ALL cases and 1073 controls (of which 653 were matched ALL/control pairs.

Exposure classification. We examined allergic diseases occurring in the first year of life as reported through interview with the biological parents (~98% mothers), wherein the parent was asked *if the child had allergies, what the child was allergic to, and whether or not a physician was consulted*. See Appendix I for structure of allergy-related questions by recruitment phase from the CCLS questionnaire. Allergens were categorized as hay fever (defined here as allergy to plants/grass/dust/mold), food, drugs and medications, soap/cosmetics, animal/insect, and other. Variables were constructed for any allergy (binary, defined as at least one of the above allergies) and total count of reported allergens. A total of 41 allergic participants were excluded from specific allergy phenotype analyses because the specific allergen(s) was unknown. A variable combining allergy to food and drugs was constructed for use in the subsequent meta-analysis. Of the resulting allergy phenotype categories, those with less than 10 exposed controls were not tested for association with disease (including animal/insect allergy and soap/cosmetic allergy).

Statistical Analysis Matched and unmatched data analyzed with logistic regression analyses were conducted to estimate the odds ratios (OR) and 95% confidence intervals (CI). Individually matched and unmatched results were similar, therefore only those from the unmatched analyses using the entire set of available cases and controls are presented for increased statistical power. All models were adjusted for child's age (centered on median age=4.61), sex, Hispanic ethnicity, maternal race, and for possible confounders determined *a priori* as being associated with both ALL and allergic disease through a literature search [i.e. mode of delivery[13,23], birth order[8,24], daycare attendance in the first year of life[10,11], and household income]. Independent models included an interaction term to test for potential effect modification by age at diagnosis, sex, child's Hispanic ethnicity, mode of delivery, and day care attendance. Models including interaction terms where the p-value for interaction was < 0.2 were then compared to the base model (no interaction term) for goodness-of-fit using the likelihood ratio test. For models in which the likelihood ratio test produced a p-value < 0.05, the marginal effects of allergy phenotype on ALL risk, derived from the model including the interaction term, are presented. To address potential control-selection bias potentially introduced by not recruiting controls to the study from 2009-2015, analyses were repeated excluding cases recruited during that time period; the results did not change (unpublished observation).

Meta-analysis

Search criteria. Using PubMed, we identified all epidemiologic studies published between November 1, 2008, the date of the first published meta-analysis[15], and June 1, 2016 using the following search terms: MeSH terms = “leukemia” OR "leukemia, lymphoid” AND "allergy and immunology” OR "hypersensitivity”; All Fields = “atopy” OR “atopic” OR "hypersensitivity” OR "allergy” OR "allergy and immunology” OR “allergic” AND “leukemia”. Results were filtered by age to include only children (birth to 18 years).

In addition, all eligible and ineligible studies reported in the 2009 and 2010 meta-analyses (including unpublished data) were retrieved for the present study.[15,16]

Study eligibility. Studies eligible for the present meta-analysis included original reports, written in English, investigating the relationship of childhood ALL (age 18 and under) and allergy, asthma, or atopic diseases. Only studies presenting odds ratios and confidence intervals/standard error were included. If multiple studies utilized overlapping study populations, the study with the largest sample size was chosen for increased power. Reasons for exclusion included incompatible methods of exposure measurement, such as serum level IgE[25], restriction of studies to specific subgroups (e.g., Down syndrome[26], inclusion of individuals age ≥18[27-29]), and overlap with included studies (Figure 1).

Data extraction. Information on authors; publication year; study design; assessment method for exposure, outcome, and covariates; and measures of association and standard error was extracted from eligible publications using a standardized form. We extracted six binary exposure variables including any allergy, eczema, asthma, hay fever, food/drug allergy, and hives in addition to the binary outcome of ALL. For studies presenting results using multiple latency periods between exposure and disease [1] or multiple methods of exposure assessment [30], reported results with the highest precision were included to improve statistical power.

Statistical analyses. Using the Metafor and hetmeta packages [21,31] in R [32], summary associations between ALL risk and allergy phenotypes were estimated using data from 4 to 13 studies.

An inverse-variance-weighted (DerSimonian-Laird) random-effects model was used to estimate summary odds ratios for each allergy phenotype. A new estimator published in Crippa *et al.* (2016) was used to estimate the percent of variation attributable to between-study heterogeneity (R_b%), rather than the typical I^2%, to reduce bias in the estimate. The *p*-value for heterogeneity was also estimated using Cochran's *Q* test. Study-level characteristics, referred to as moderators, thought to potentially contribute to heterogeneity were selected and abstracted prior to analyses. These include medical record vs. self-reported exposure assessment, nested case-control vs. other study design, ≥80% vs. <80% control response rate, hospitalized vs. healthy controls, specified vs. not specified latency period, and publication year. Potential moderators were tested for correlation using an *a priori* determined Pearson correlation coefficient r^2 threshold of 0.3 [33]. Among correlated moderators, the one that could be used to characterize the most studies was selected for inclusion in the mixed-effects models. Mixed-effects models were constructed when significant heterogeneity was detected for models of ALL risk with any of the allergy phenotypes (R_b>0). Separate models were built for each potential moderator in turn and a full model including all uncorrelated moderators was constructed. Stratified measures of effect and associated *p*-values, $P_{Qm,}$ were estimated from the omnibus test of model coefficients. Because the CCLS study only measured allergy exposure during the first year of life, analyses were also conducted excluding the CCLS study to assess potential sources of bias.

Publication bias was formally assessed for analysis of any allergy via a regression test of funnel plot asymmetry.

RESULTS

CCLS analysis

In general, controls were comparable to cases except for the fact that controls tended to be of higher income than cases (Supplemental Table S2). The odds ratios for any allergy and most allergy phenotypes ranged from 0.75 (hay fever) to 1.29 (any allergy), but there was no strong statistical evidence of association overall (Table 1). However, multiplicative interaction was detected between any allergy and age at diagnosis/reference date. Stratified analyses show that risk of ALL was associated with occurrence of any allergy during the first year of life only amongst the youngest study subjects (Figure 2). Food allergy and allergy count were also suggestive of interaction with age, though the goodness of fit test for including the interaction term did not meet the significance threshold *P*<0.05.

Meta analysis

Characteristics of the 13 studies included in the meta-analysis are listed in Supplemental Table S3 [17,19,20,24,30,34-40]. Thirteen studies reported results specific to ALL and sample sizes ranged from less than 200 to several thousands (Figure 3).

Results from the random-effects models are shown in Figure 3 and Table 3 for any allergy and specific allergy phenotypes, respectively. Child's history of hay fever was the sole allergy to be associated with a reduced risk of ALL (OR 0.65, 95% CI: 0.47, 0.90). No statistical evidence of association was seen for other types of allergies.

About 22% of the variability in effect estimates for any allergy is due to heterogeneity (R_b CI: 0,79%), while no heterogeneity was reported within allergy types (Table 2). To formally assess potential study-level characteristics contributing to heterogeneity of any allergy and ALL, a meta-regression was performed using *a priori* selected, uncorrelated potential moderators, including medical record vs. self-reported exposure assessment. Medical record-based studies (Figure 3) were associated with an attenuated relationship between any allergy and ALL risk in the mixed-effects model (P_{Qm} <0.05, Model 1, Table 3). Model 1 was then applied to four allergy phenotypes and the same trend was observed; *i.e.*, the odds ratio for ALL risk associated with each of the allergy phenotypes was larger (and consistently null) in medical record-based studies compared to non-record based studies (Supplementary Figure 1). However, with fewer studies analyzed by subcategory, none of the formal tests of moderators (P_{Qm}) were statistically significant (unpublished observation).

We furthered explored whether control selection method added heterogeneity to the meta-analysis. Within studies where exposure assessment was self-reported, the odds ratio for ALL risk associated with any allergy was larger (null) in studies with population-based controls compared to studies with hospital-based controls; however the mixed-effects model comparing studies with hospital-based controls (3 studies) to those with population-based controls (6 studies) produced a test of moderators statistic that was not statistically significant (P_{Qm} =0.12)(Figure 3).

Finally, we hypothesized that publication year could contribute to between-study heterogeneity because the incidence rates of allergy and ALL are both changing over time, study quality is improving, and public perception of allergy is changing [41,42]. In a univariate mixed-effects model, publication year was associated with an increase in the odds ratio relating any allergy to ALL (P_{Qm} <0.05, Model 2, Table 3), indicating that early studies tended to report an inverse relationship but more recent studies have reported null or positive associations (Supplementary Figure 2). In the full model containing both the medical record indicator and publication year (Model 3, Table 3) the independent effects of medical-record exposure assessment and publication year remain compelling, though no longer reach statistical significance. All three models applied reduced the residual between-study heterogeneity estimate (R_b%) to 0.

The observed asymmetry in the funnel plot estimated using the random-effects model suggested evidence of publication bias for studies examining the relationship between any allergy and ALL, (Supplementary Figure 3, *P*=0.06). There were not enough studies that included specific allergy phenotypes to formally assess publication bias by subcategory.

DISCUSSION

Based on the CCLS study and the largest meta-analysis to date -- combining data for over 8000 ALL cases and 25,000 controls – our findings suggest there is no clear association between allergy and the risk of ALL. However, several observations have emerged from our analyses, raising methodological concerns and biological issues discussed below.

In the CCLS, we observed an increased risk of ALL given any allergy occurring in the first year of life. However, age-stratified analyses showed that the youngest study participants drove this association. It was more common for allergy to precede ALL when diagnoses occurred at 1.5 -3 years of age than if disease arose later in life. Allergy beyond the first year of life was not assessed in the CCLS and so it is unclear whether allergy occurs more frequently in cases proximal to diagnosis at any age, or if this phenomenon is limited to the youngest cases. Results from a large, record-based study in Taiwan support the hypothesis that any allergy is more common proximal to diagnosis across the life course, where compared to allergy in the first year of life and >1 year before diagnosis, allergy <1 year before diagnosis conferred the highest risk of childhood ALL [19]. The association could also be explained by differential recall, where parents of younger children at the time of interview could more accurately recall allergy occurring in the first year of life than parents of older children.

Early-onset allergy to drugs or medications conferred a larger increase in the risk of ALL than any other specific allergy phenotype evaluated in the CCLS population, although the association was not statistically significant. While potentially a chance finding, there are several factors that could be contributing to this observation. First, true drug hypersensitivities have the potential to influence the allergy - ALL association. However, many adverse drug reactions are not immune-mediated and further, those that are are heterogeneous in type and can be IgE-, IgG-, or T cell-mediated [43] and may have differential effects on ALL etiology. If a causal relationship exists between allergies to drugs and ALL further functional studies are needed for full characterization. The relatively high odds ratio for ALL associated with an allergy to drugs or medications could also be due to confounding. We could not control for the number or type of medications taken in the first year of life, which was likely higher in cases than controls, yielding more opportunity for cases to develop a drug allergy that is causally-independent of ALL. This line of reasoning is supported by the observation that cases experience more frequent infection-related medical office visits prior to leukemia onset than controls, in both self-report [18] and medical record-based studies [1,44] of ALL; and are therefore more likely to be exposed to a broader spectrum of medications. There is little epidemiologic evidence from previous studies to support an association between ALL and allergy to drugs *per se*, as previous studies have typically examined a combined food/drug allergy phenotype [36,37,40].

The CCLS population is large and unique. However, the study also presents some limitations. Selection bias is of concern in the CCLS, wherein participating controls are of higher socioeconomic status than cases and non-participating controls [45]. Higher socioeconomic status has been previously associated with increased odds of allergic disease in children [46], potentially biasing our results in the direction of an inverse association. Parental interview assessed only allergy occurring in the first year of life. Allergic disease is relatively rare in infants [2] and while it may be a useful measure of early-life immune dysregulation, low exposure prevalence in our study limited the power to detect statistically significant associations with ALL. Further, certain allergies, like hay fever, can be difficult to diagnose in very young children. This restriction on exposure assessment may also limit the generalizability of our results to the childhood allergy – ALL relationship overall. However, Chang *et al.* (2012) also reported null associations between allergy phenotypes occurring before age one and ALL, with the exception of an increased risk for atopic dermatitis, an exposure that was not included in the CCLS questionnaire. A limitation of our study is that we do not have medical record data for comparison and the contribution of uncontrolled confounding to our result cannot be ruled out. However, we believe that several analytic advantages may have contributed to the discrepancy between previous studies and ours. First, we incorporated a latency period into our study, wherein we excluded cases and controls for which allergy occurred within 6 months of leukemia diagnosis or reference date, respectively. Thus, we reduced the potential that our observed associations were due to reverse causality. Second, our study controlled for shared risk factors for allergy and ALL, including birth order, day care attendance, and mode of delivery, which potentially confounded the results from previous studies.

In this updated meta-analysis, most associations with allergy were attenuated, compared to previous meta-analyses[15,16], and only hay fever maintained a statistically significant inverse association with ALL. With the addition of recent studies including a large, medical-record based analysis, it is clear that between-study heterogeneity limits the interpretation of any summary measure of effect. Hay fever is the most common allergy under study [2] but is also the most commonly misdiagnosed [47], as its symptoms are similar to and easily mistaken for upper respiratory tract infections and vice versa. Using any allergy as an exposure variable allowed for inclusion of the greatest number of studies for meta-analysis and thus allowed us to statistically assess study-level sources of heterogeneity for the first time. Medical-record exposure assessment or correlated features (record-based or nested design, incorporation of a latency period, hospital vs. healthy controls, and control response rate) were important sources of between-study heterogeneity. Our meta-regression (Model 1) showed that, in medical record-based studies, no association was observed between allergy of any type and ALL; and within self-report-based studies, any allergy and hay fever resulted in statistically significant inverse relationships. A parallel pattern has emerged with the association of early-life infection and ALL, wherein questionnaire-based studies have found an inverse association between infection and ALL and medical-record based studies have found infection to be a risk factor [5]. This difference in association is possibly the result of reporting bias, which may also be occurring in the allergy-ALL relationship. Linabery *et al.* argued that misclassification of the allergy phenotype is likely occurring in both the case of parental report and medical record; however, the record-based studies are more likely to be non-differential with respect to case status, and thus less subject to bias. Further, a recent systematic review showed a large and significant difference in prevalence of childhood atopic disorders comparing those diagnosed by a physician to parental-report [48]. For example, the annual prevalence of allergic rhinitis according to physician report ranged from 0.4-4.1% whereas the self-reported annual prevalence ranged from 15.1-37.8% (41). The authors speculate that consistent under-reporting among general medical practitioners and over-reporting by parents create this difference. There is thus potential for self-report studies of allergy and childhood leukemia to produce study results that have both higher power and higher susceptibility to bias, potentially explaining the strong inverse associations between the two diseases in self-reported studies, whereas medical record-based studies present more attenuated relationships.

Within studies using self-reported exposure assessment, the test for heterogeneity between studies using population-based controls and hospital-based controls was not significant, supporting our decision to include both types of study. However, utilizing hospitalized controls does exaggerate the inverse relationship between allergy and ALL among self-reported studies, possibly due to control selection bias, as potential controls receiving treatment for allergy-related symptoms were included in some of these studies and not in others.

Another significant source of heterogeneity is publication year. Over time, the strong inverse association between allergy and ALL is shifting towards no association, with some of the most recent studies finding a positive association between the two diseases. Several factors could contribute to this observation, including improvements in study quality over time, changing prevalence of both allergy and ALL, and potentially changing opinions or knowledge regarding allergy and hygiene among the general public [49,50].

Numerous sources of bias complicate interpretation of these results. Results from the CCLS and another recent study [19] suggest that the timing of allergy relative to ALL diagnosis may modify observed associations. Across the 13 studies included in the meta-analysis, timing of exposure assessment relative to diagnosis was largely consistent. Thus, tests of between-study heterogeneity related to timing of allergy could not be assessed, though it may be contributing to the underlying variation in effect estimates. Publication bias may be limiting the interpretability of our meta-analysis. Within questionnaire-based studies, questionnaire structure can result in differential parental response between studies. For example, a questionnaire may ask whether allergy was ever diagnosed by a physician, or have only asked about a specific subset of allergens, or as in the case of the CCLS, restricted to a particular time window. These factors may make the interpretation of the effect of combined allergy categories difficult. Medical records are likely more easily standardized across studies, but combining those studies with self-reported studies poses similar problems. Recall bias is also of concern in questionnaire-based studies, but not in record-based studies; wherein parents of children with cancer may not remember a comparably insignificant allergic event. On the other hand, medical-record based studies are likely capture only a subset of more severe allergy, an exposure that may not be representative of allergy overall. Similarly, selection bias is on concern particularly in studies where participant contact is required (i.e. to conduct a questionnaire). Our results show a clear exaggeration of associations when controls are selected in clinics, and in the CCLS study, like many modern case-control studies, controls are not exchangeable with cases in terms of socio-economic status, opening the door to residual confounding.

CONCLUDING REMARKS

Due to the complexities outlined above, understanding the causal relationship between allergy and childhood leukemia cannot be achieved with additional status quo epidemiologic studies, rather it will require a combination of large, prospective, perhaps biomarker-based cohort studies and basic biological and mechanistic studies. Functional studies can demonstrate biologically meaningful relationships between the two diseases [51]. For example, the tumor surveillance hypothesis has been well demonstrated in biological studies [51], wherein individuals who are prone to allergic response are considered to have a hyper-vigilant antigen recognition and response thus producing more successful cancer cell surveillance and elimination. The vast heterogeneity of the effect of allergy observed within and between cancer sites suggests that the mechanisms by which these two families of disease intersect are not mutually exclusive and speak to the potential pleiotropic effects of allergy across individuals and tissue sites [51].

In the absence of complex and expensive studies, germline genetic predictors of allergy could serve as an informative proxy for unbiased exposure assessment of allergy in the context of cancer risk. From the current analysis, the once apparently strong inverse relationship between allergy and ALL has waned to an absent or ambiguous association. As the tide changes on the story of allergy and ALL, so must our tactics for studying the intersection of these complex diseases.

ACKNOWLEGEMENTS

The authors thank the families that participated in the CCLS. Without their time and effort none of our studies would be possible. The authors also thank our clinical collaborators throughout California for their continued support of our research and commitment to their patients: University of California Davis Medical Center (Dr. Jonathan Ducore), University of California San Francisco (Drs. Mignon Loh and Katherine Matthay), Children's Hospital of Central California (Dr. Vonda Crouse), Lucile Packard Children's Hospital (Dr. Gary Dahl), Children's Hospital Oakland (Dr. James Feusner), Kaiser Permanente Roseville (former Sacramento; Drs. Kent Jolly and Vincent Kiley), Kaiser Permanente Santa Clara (Drs. Carolyn Russo, Alan Wong, and Denah Taggar), Kaiser Permanente San Francisco (Dr. Kenneth Leung), and Kaiser Permanente Oakland (Drs. Daniel Kronish and Stacy Month). Finally, the authors thank the entire California Childhood Leukemia Study staff and the former UCB Survey Research Center for their effort and dedication.

1.3 Tables 1.1-1.3 Allergies and childhood acute lymphoblastic leukemia

Table 1.1 Odds ratios and 95% confidence intervals for ALL cases vs. controls who reported having allergy to specific allergens in the first year of life in the CCLS

Allergen		**Cases (n=977)**	**Controls (n=1073)**	**Odds Ratio[a]**	**95% Confidence Interval**	***P*-Value**
Any Allergy	No	827	942			
	Yes	128	119	**1.29[b]**	0.97,1.72	0.08
Hay Fever	No	927	1020			
	Yes	22	30	**0.75**	0.42,1.32	0.32
Food	No	906	995			
	Yes	40	45	**1.09**	0.69,1.71	0.71
Drugs and Medications	No	909	1011			
	Yes	38	26	**1.63**	0.96,2.75	0.07
Food/Drug	No	873	970			
	Yes	74	70	**1.24**	0.87,1.76	0.24
Allergy Count	No	851	954			
1 Allergen vs. None	Yes	88	91	**1.08**	0.79,1.49	
2+ Allergens vs. None	Yes	16	16	**1.24**	0.60,2.56	0.56[c]

[a] Adjusted for age, sex, maternal race, child's Hispanic status, income, mode of delivery, daycare attendance, and birth order

[b] Marginal effect where model additionally includes interaction term median-centered age*any allergy. See figure 2 for results from age strata

[c] Test for linear trend

Table 1.2. Results of Random Effects Meta-Analyses for Various Allergic Exposures and ALL Risk from Current and Prior Meta-Analyses

		Any Allergy				Eczema			
Author	Year	N	Summary OR	95% CI	I^2% R_b %	N	Summary OR	95% CI	I^2% R_b %
Dahl et al	2009	8	0.67	0.54,0.82	NR	5	0.68	0.56,0.83	29
Linabery et al	2010	6	0.69	0.54,0.89	80	5	0.74	0.58,0.96	62
Present study		10	0.76	0.58,1.01	22	7	0.86	0.64,1.14	0

		Hay Fever				Food/Drug			
Author	Year	N	Summary OR	95% CI	I^2% R_b %	N	Summary OR	95% CI	I^2% R_b %
Dahl et al	2009	5	0.53	0.43,0.65	28				
Linabery et al	2010	3	0.55	0.46,0.66	3				
Present study		8	0.65	0.47,0.90	0	6	0.73	0.52,1.03	0

		Asthma				Hives			
Author	Year	N	Summary OR	95% CI	I^2% or R_b %	N	Summary OR	95% CI	I^2% or R_b %
Dahl et al	2009	6	0.82	0.63,1.10	43				
Linabery et al	2010	7	0.79	0.61,1.02	44	2	0.93	0.73,1.19	0
Present study		9	0.85	0.63,1.16	0	4	1.23	0.80,1.87	0

CI, 95% confidence interval; I^2%, percent of total variability due to heterogeneity; OR, odds ratio; N, number of included studies; R_b%, percent of total variability due to heterogeneity

Table 1.3. Results from 3 Random Effects Meta-Regression Models in the Association Between Any Allergy and ALL

Model		Summary Beta	95% CI	*P* value	R_b%	P_{Qm}
Model 1	Intercept	-0.46	-0.76, -0.16	0.003	0	0.026
	Medical Record	0.55	0.07, 1.04	0.026		
Model 2	Intercept	-0.98	-1.63,-0.34	0.003	0	0.017
	Publication year	0.04	0.01,0.07	0.017		
Model 3	Intercept	-0.98	-1.63, -0.34	0.003	0	0.020
	Medical Record	0.41	-0.10, 0.92	0.114		
	Publication year	0.03	-0.003,0.07	0.073		

CI, confidence interval; I^2%, percent of total variability due to heterogeneity; P_{Qm}, *P*-value for Cochran's test of Moderators

1.4 Figures 1.1-1.3 Allergies and childhood acute lymphoblastic leukemia

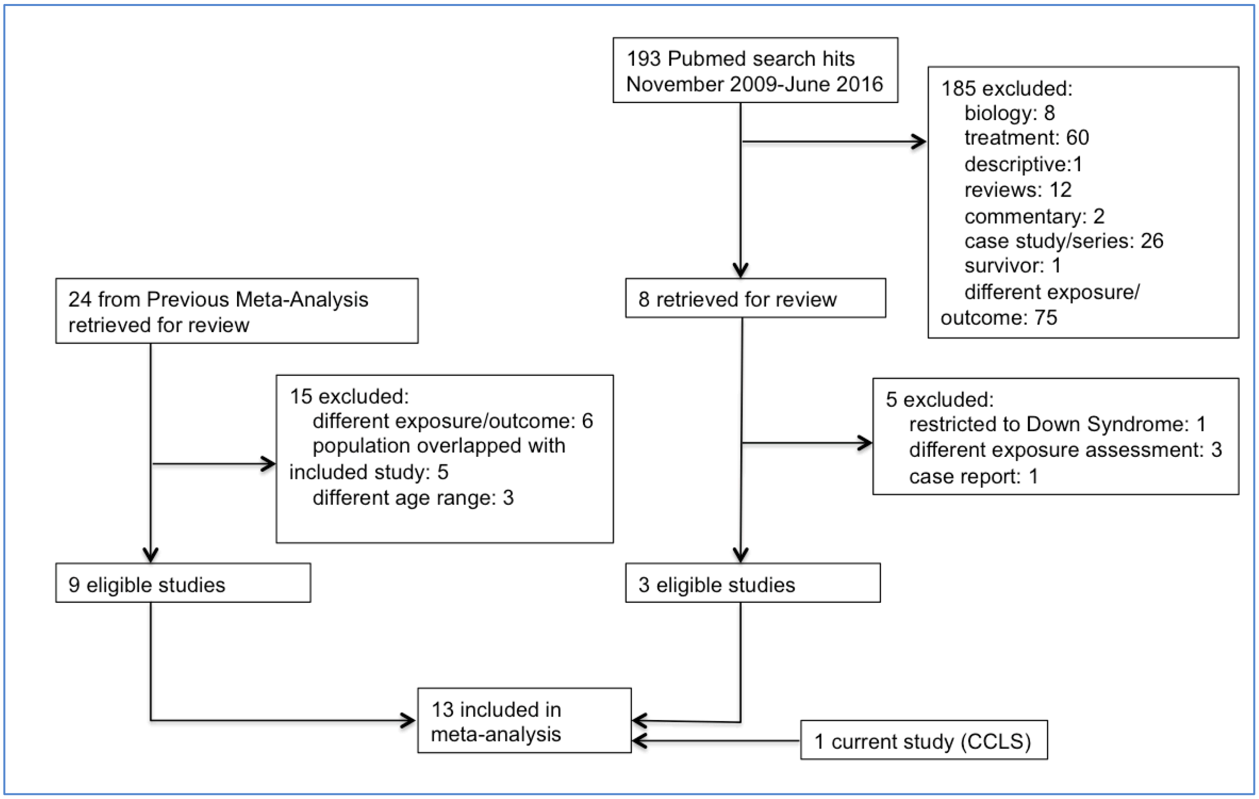


Figure 1.1: Flow chart - Meta-analysis study search and selection process

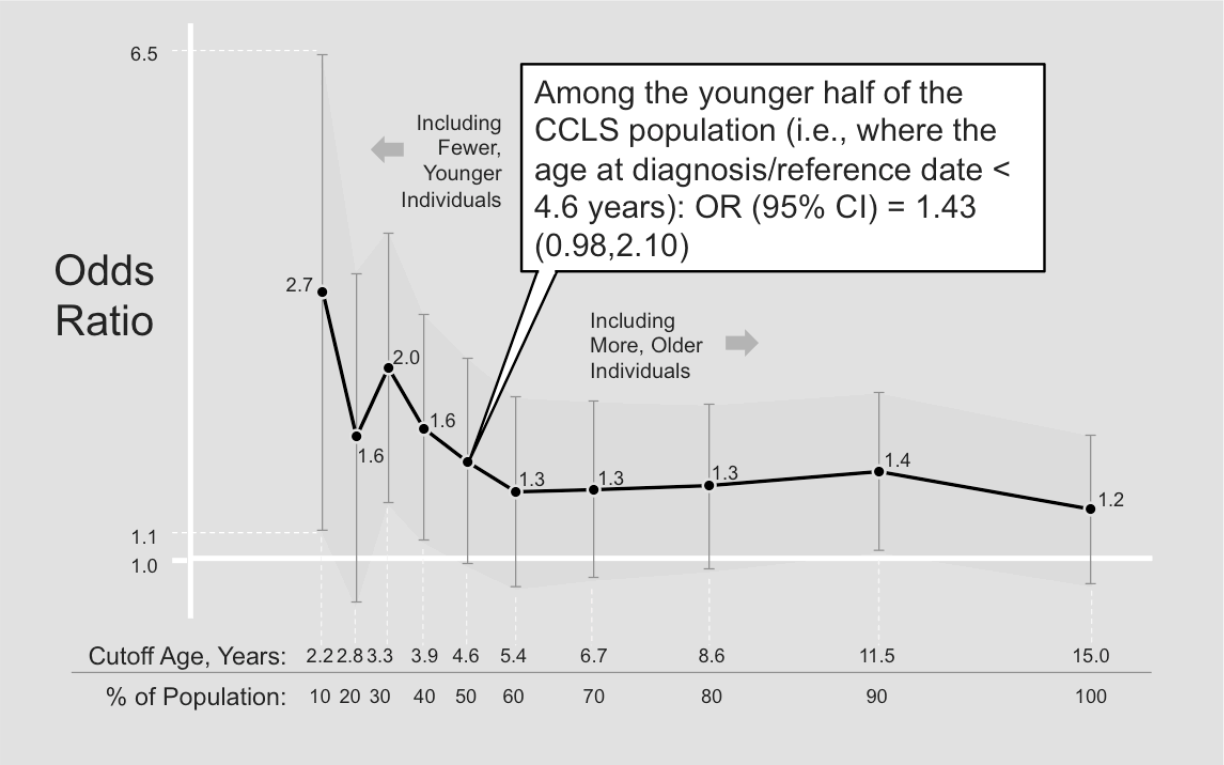
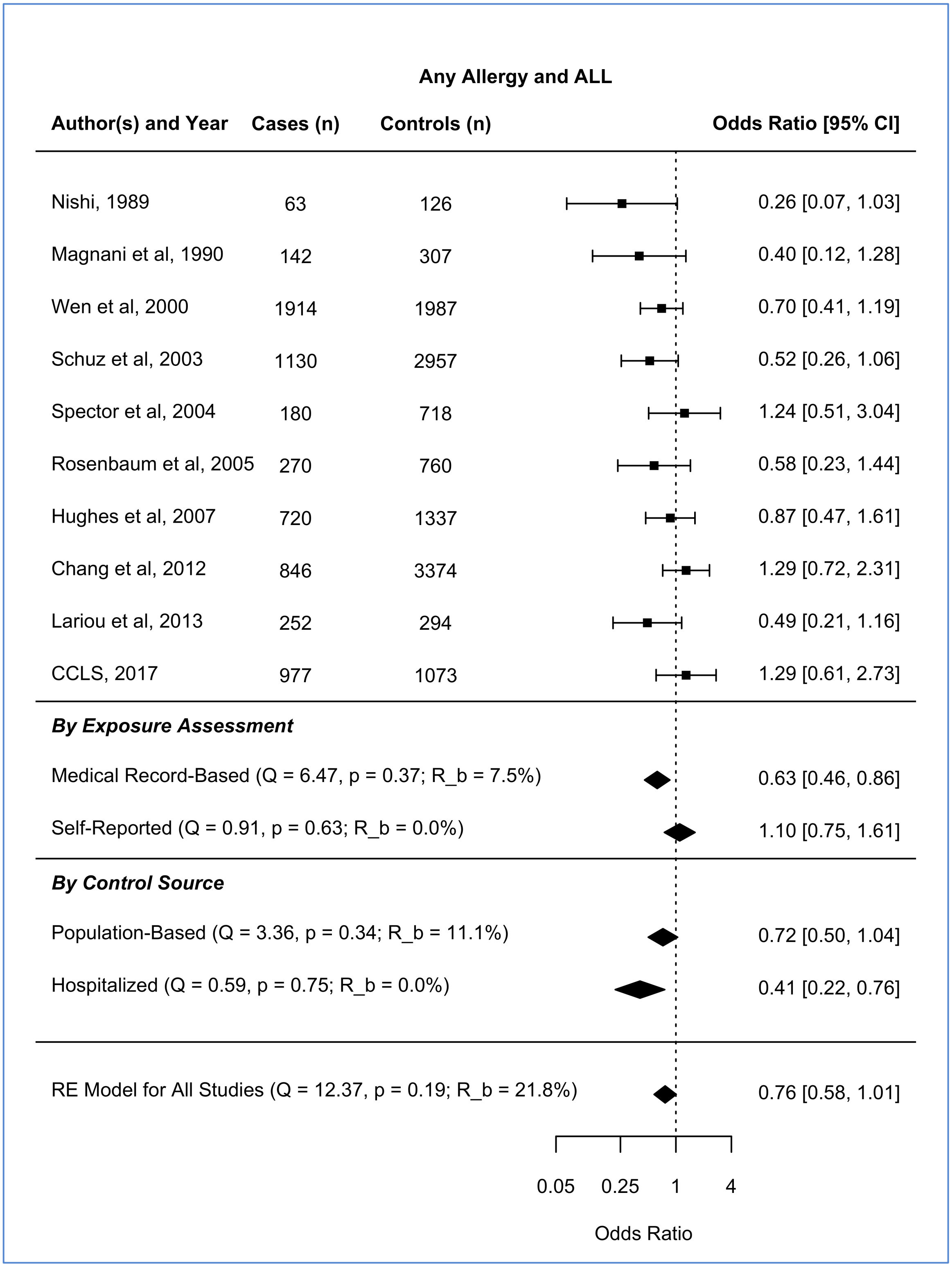


Figure 1.2: Sensitivity analysis testing the association of any allergy and ALL risk among the youngest individuals in the CCLS by decile. Models were adjusted for sex, maternal race, child's Hispanic status, income, mode of delivery, daycare attendance, and birth order. Text annotation box provides example interpretation for stratification at the 50th percentile.



Author(s) and Year	Cases (n)	Controls (n)	Odds Ratio [95% CI]
Nishi, 1989	63	126	0.26 [0.07, 1.03]
Magnani et al, 1990	142	307	0.40 [0.12, 1.28]
Wen et al, 2000	1914	1987	0.70 [0.41, 1.19]
Schuz et al, 2003	1130	2957	0.52 [0.26, 1.06]
Spector et al, 2004	180	718	1.24 [0.51, 3.04]
Rosenbaum et al, 2005	270	760	0.58 [0.23, 1.44]
Hughes et al, 2007	720	1337	0.87 [0.47, 1.61]
Chang et al, 2012	846	3374	1.29 [0.72, 2.31]
Lariou et al, 2013	252	294	0.49 [0.21, 1.16]
CCLS, 2017	977	1073	1.29 [0.61, 2.73]
By Exposure Assessment			
Medical Record-Based (Q = 6.47, p = 0.37; R_b = 7.5%)			0.63 [0.46, 0.86]
Self-Reported (Q = 0.91, p = 0.63; R_b = 0.0%)			1.10 [0.75, 1.61]
By Control Source			
Population-Based (Q = 3.36, p = 0.34; R_b = 11.1%)			0.72 [0.50, 1.04]
Hospitalized (Q = 0.59, p = 0.75; R_b = 0.0%)			0.41 [0.22, 0.76]
RE Model for All Studies (Q = 12.37, p = 0.19; R_b = 21.8%)			0.76 [0.58, 1.01]

Figure 1.3: Forest plots - Any Allergy and Acute Lymphoblastic Leukemia random-effects model, meta-regression medical record vs. self report, and meta-regression hospitalized vs. population-based controls

1.5 Supplementary Materials: Allergies and childhood acute lymphoblastic leukemia

1.5.1 Supplementary Tables 1.1-1.3

Table S1.1 CCLS Enrollment and Participation Summary ALL cases and controls

Study Phase	Year(s)		Number Eligible	Number Consented	Number Interviewed	Participation Rate (Interviewd/Eligible)
1	1995-1999	Cases	273	226	214	78.39%
		Controls	366	279	205	56.01%
2	1999-2002	Cases	324	284	273	84.26%
		Controls	508	361	360	70.87%
3	2002-2009	Cases	704	624	515	73.15%
		Controls	826	662	661	80.20%
5[b]	2011-2015	Cases	1005	771	475	47.26%
		Total Cases	2306	1905	1477	64.05
		Total Controls	1700	1302	1226	72.12

[b] No Cases or Controls recruited during study phase 4,no controls recruited during phase 5

Table S1.2. Baseline Characteristics of CCLS ALL Cases and Controls

		ALL (n=977)		Controls (n=1073)	
		N	**%**	**N**	**%**
Age	Median (Range)	4.56(1.5-15.0)		4.63(1.5-15.0)	
Sex					
	Male	**509**	52.6	**612**	57.0
	Female	**468**	48.3	**461**	43.0
Maternal Race					
	White	**802**	82.9	**923**	86.0
	Black	**31**	3.2	**27**	2.5
	Other	**144**	14.9	**123**	11.5
Child's Hispanic Status					
	Hispanic	**517**	53.4	**474**	44.2
	Non-Hispanic	**460**	47.5	**599**	55.8
Income					
	<$15,000	**161**	16.6	**100**	9.3
	$15,000-$29,999	**195**	20.1	**127**	11.8
	$30,000-$44,999	**137**	14.2	**139**	13.0
	$45,000-$59,999	**132**	13.6	**142**	13.2
	$60,000-$74,999	**60**	6.2	**117**	10.9
	$75,000+	**292**	30.2	**448**	41.8
Mode of Delivery					
	Vaginal	**732**	75.6	**856**	79.8
	Cesarean	**245**	25.3	**217**	20.2
Daycare in 1st year of life					
	Yes	**331**	34.2	**277**	25.8
	No	**646**	66.7	**796**	74.2
Birth Order	Mean (Range)	1.0(0.0-8.0)		1.1(0.0-9.0)	

ALL, acute lymphoblastic leukemia

Table S1.3. Characteristics of Studies Included in the Meta-Analysis of Allergy and ALL

Author Name (Pub Year)	**Study Population**	**Ages at Diag-nosis**	**Source of Cases**	**Source of Controls**	**Number Cases/ Controls**	**Source of Exposure Data**
Nish*i et al. (1989)*	Hokkaido Prefecture, Japan	< 15 (1981-1987)	Receiving treatment for non-T ALL at 9 Hokkaido Hospitals	Hospital: excluded non-routine examinations	63/126	Questionnaire
Magnani *et al. (1990)*	Hospital in Turin, Italy	< 18 (1974-1984)	Hospital	Hospital: excluded cancer, down syndrome, beta thalassemia, infectious mononucleosis, splenic enlargement	142/307	Questionnaire
Wen *et al. (2000)*	Children Cancer Group (CCG), USA	< 15 (1989-1993)	CCG Member Institutions	Random digit dialing	1914/1987	Questionnaire
Schuz *et al. (2003)*	Germany	< 15 (1992-1994)	German Childhood Cancer Registry	Population registry	1294/2957	Questionnaire
Jourdan-Da Silva *et al. (2004)*	France	< 15 (1990-1998)	National Cancer Registry of Childhood Leukemia and Lymphoma	Random digit dialing	473/567	Questionnaire
Spector et al. (2004)	Kaiser and Group Health HMOs (CA, OR, WA), USA	< 7 (1985-1999)	HMOs	HMOs	180/718 ALL 147/586 B-ALL	Medical record
Rosenbaum *et al. (2005)*	Population-based western and central New York, USA	< 15 (1980-1991)	records/registries form 4 medical centers	Birth registry	270/760	Questionnaire

Soderberg *et al. (2006)*	Sweden	< 19 (1987-1999)	Swedish Cancer Registry	Population registry	875 ALL/14865	Hospital Discharge Registry
Hughes *et al. (2007)*	Great Britain	< 15 (1991-1996)	Great Britain	Primary care population registers	821/1337	Medical Records
Rudant *et al. (2010)*	France	1-14 (2003-2004)	French National Registry of childhood blood malignancies	Random digit dialing	765/1681	Questionnaire
Chang *et al. (2012)*	National Health Insurance Research Database (NHIRD), Taiwan	1-10 (2000-2008)	Catastrophic illness dataset	Health insurance database	846/3374	NHIRD
Lariou *et al. (2013)*	Six Nationwide Registry for Childhood Heamatological Malignancies (NARCHEM) hospitals, Greece	< 15 (1999-2003)	NARCHEM	Hospital: excluded atopic admisson diagnoses	252/294	Questionnaire, Serum IgE
Present Study	California Childhood Leukemia Study, USA	< 15 (1995-2008)	CCLS California	Birth registry	1070/1073	Questionnaire

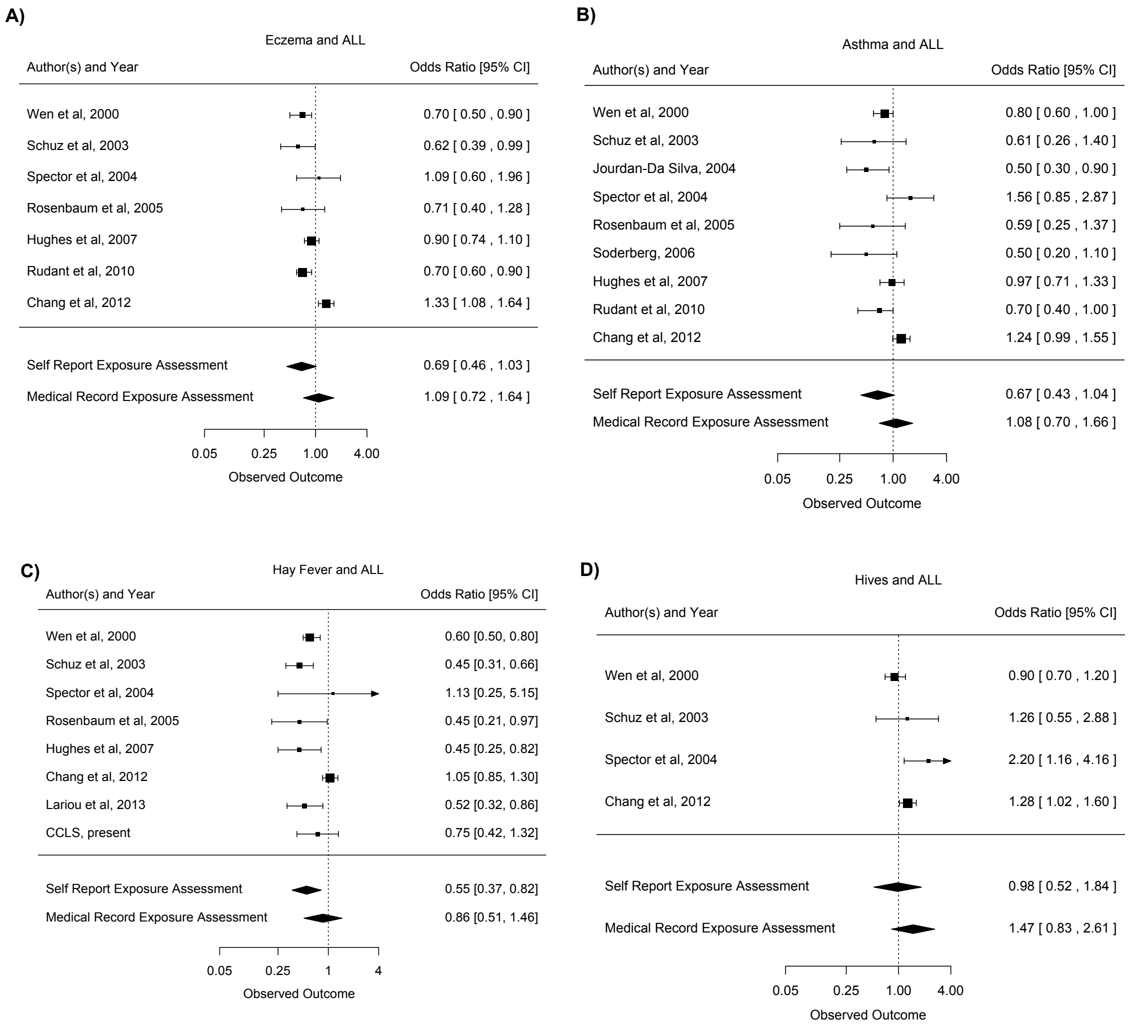
Table S1.3. Characteristics of Studies Included in the Meta-Analysis of Allergy and ALL (Continued)

Author Name (Publication Year)	**Matching**	**Matching/ Adjusting Variables**	**Exposure Variables (w/latency)**	**Outcome Variables**	**Citation**
Nish*i et al. (1989)*	Yes	Age, sex, neighborhood	Atopic diathesis (asthma or atopic dermatitis)	ALL	27
Magnani *et al. (1990)*	No	SES	Allergy (asthma, rhinitis as indicator)	ALL	26
Wen *et al. (2000)*	Yes	Age, race, area code/months breastfeeding, maternal education, race, family income	Asthma, hay fever, food/drug, eczema, hives, any, count	T ALL, early B ALL, Pre-B ALL, B ALL not specified, unknown	33
Schuz *et al. (2003)*	Yes	Sex, age, neighborhood/ age, gender, year of birth, urbanization, SES	Parental report of physician diagnosed atopic disease, hay fever, neurodermatitis, asthma, contact eczema, hives, food/drug, other	ALL	30
Jourdan-Da Silva *et al. (2004)*	Yes	Frequency by age, sex, geography	Asthma	ALL, AML	24
Spector et al. (2004)	Yes	HMO, gender, DOB, /breastfeeding maternal age, birth weight, sibship, race	Asthma, atopy/hives, eczema, food/drug/bee, pollen/dust/dander, any,	ALL, B-ALL	32

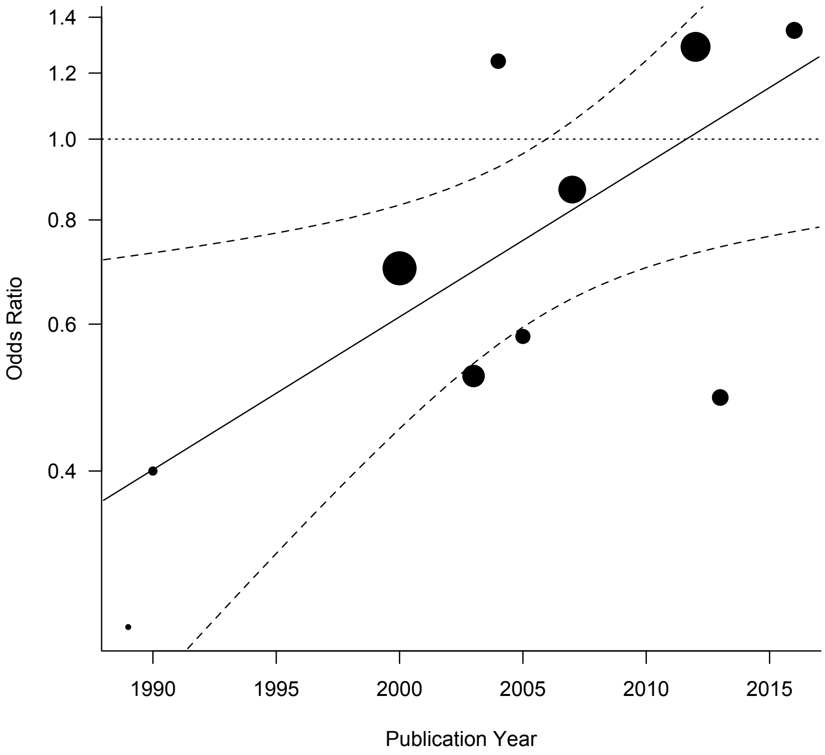
Rosenbaum *et al. (2005)*	Yes	Frequency by sex, race, birth year/maternal smoking, education, breastfeeding	Asthma, eczema, food/drug/bee, dust/pollen, cat/dog	ALL	28
Soderberg *et al. (2006)*	Yes	Sex, 5-year age group, year of diagnosis/SES	Asthma	ALL, AML	31
Hughes *et al. (2007)*	Yes	Sex, age, region	Any allergy, eczema, asthma, hay fever	ALL, Common ALL, AML	20
Rudant *et al. (2010)*	No	Age, sex, parent profession, urban/rural status, maternal age	Asthma, bronchitis with wheezing, eczema, asthma with eczema	ALL, AML	29
Chang *et al. (2012)*	Yes	Date of birth, sex, reference date/income, urbanization	Before 1 year of age or before 1 year pre-dx: any allergy, count, rhinitis, asthma, urticaria, atopic dermatitis	ALL	19
Lariou *et al. (2013)*	Yes	Sex, age/maternal education, maternal age at birth, breastfeeding, maternal smoking during pregnancy, birthweight, birth order	Any allergy, food, respiratory, other	ALL	25
Present Study	Yes	Age, sex, child hispanic status, maternal race/income, mode of delivery, daycare attendence, birth order	Before 1 year of age: any allergy, hay fever, food, drug	ALL, AML	*NA*

1.5.2 Supplementary Figures 1.1-1.3

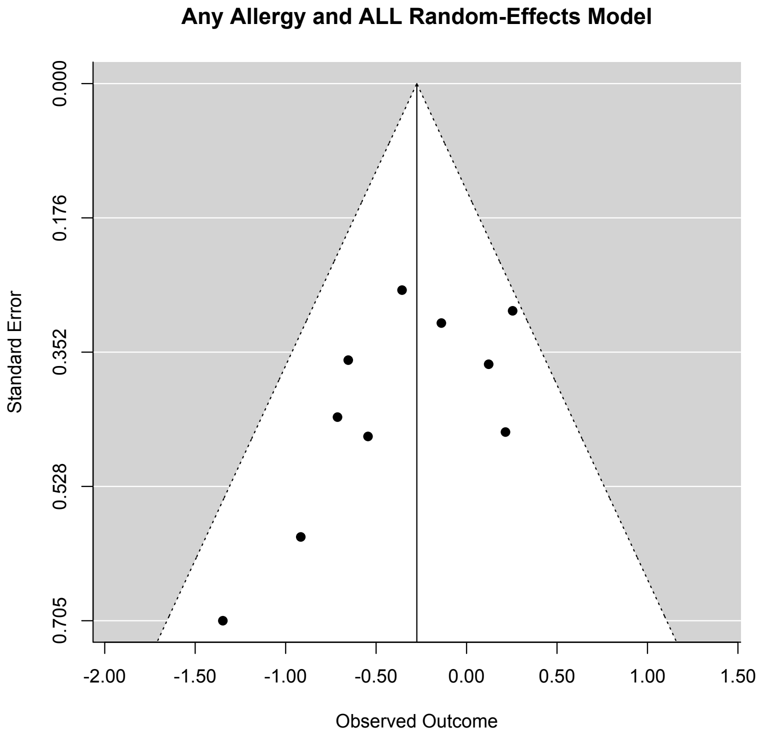
Supplementary Figure 1. Mixed effects model results for allergy and risk of acute lymphoblastic leukemia stratified by exposure assessment. A) eczema B) asthma C) hay fever D) hives.



Supplementary Figure 2. Individual study results of any allergy and risk of acute lymphoblastic leukemia from 1989-2016. Size of dot proportional to weight in mixed effect models



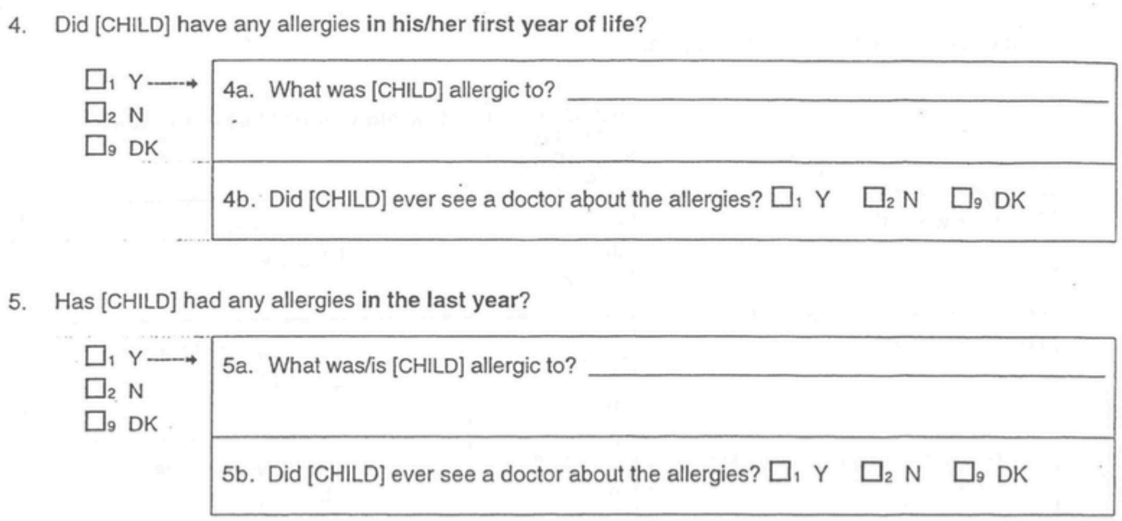
Supplementary Figure 3. Funnel plot for any allergy and ALL based on random effects model



1.5.3 Appendix I

CCLS Questionnaire: Allergy related questions by phase

Phase 1:



4. Did [CHILD] have any allergies **in his/her first year of life?**

□1 Y →
□2 N
□9 DK

4a. What was [CHILD] allergic to? ____________________

4b. Did [CHILD] ever see a doctor about the allergies? □1 Y □2 N □9 DK

5. Has [CHILD] had any allergies **in the last year?**

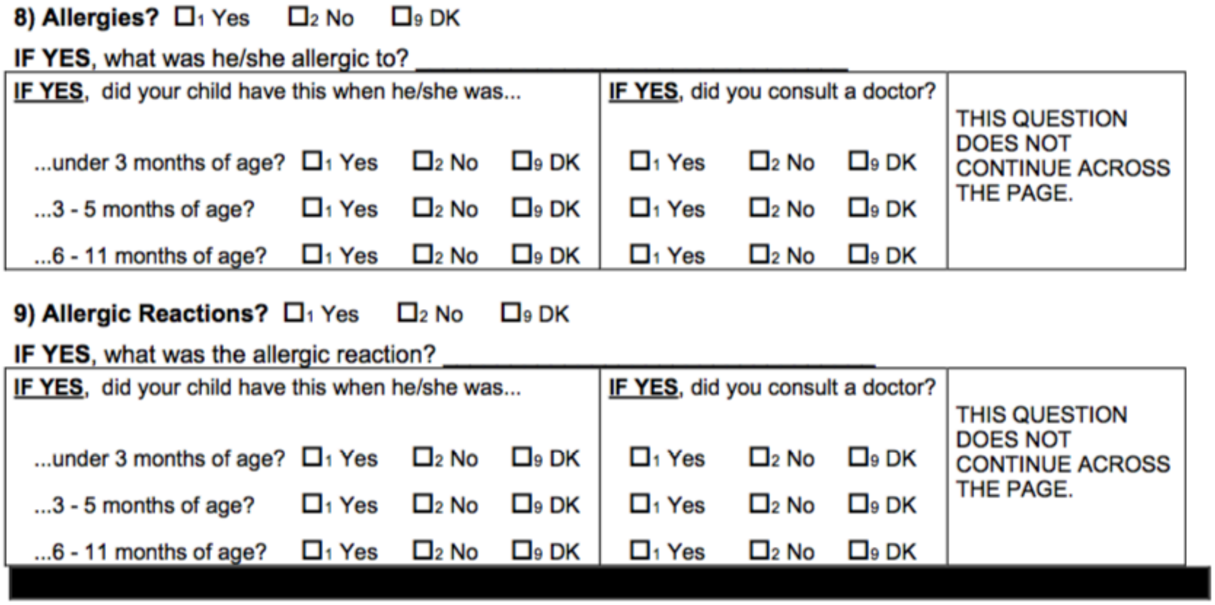
□1 Y →
□2 N
□9 DK

5a. What was/is [CHILD] allergic to? ____________________

5b. Did [CHILD] ever see a doctor about the allergies? □1 Y □2 N □9 DK

Phase 2:

The next questions are about any illnesses and infections your child may have had in **his/her first year of life**. USE SHOW CARD #5. **Did [CHILD] ever have.......**



8) Allergies? □1 Yes □2 No □9 DK

IF YES, what was he/she allergic to? ______________________________

IF YES, did your child have this when he/she was...				**IF YES**, did you consult a doctor?			THIS QUESTION DOES NOT CONTINUE ACROSS THE PAGE.
...under 3 months of age?	□1 Yes	□2 No	□9 DK	□1 Yes	□2 No	□9 DK	
...3 - 5 months of age?	□1 Yes	□2 No	□9 DK	□1 Yes	□2 No	□9 DK	
...6 - 11 months of age?	□1 Yes	□2 No	□9 DK	□1 Yes	□2 No	□9 DK	

9) Allergic Reactions? □1 Yes □2 No □9 DK

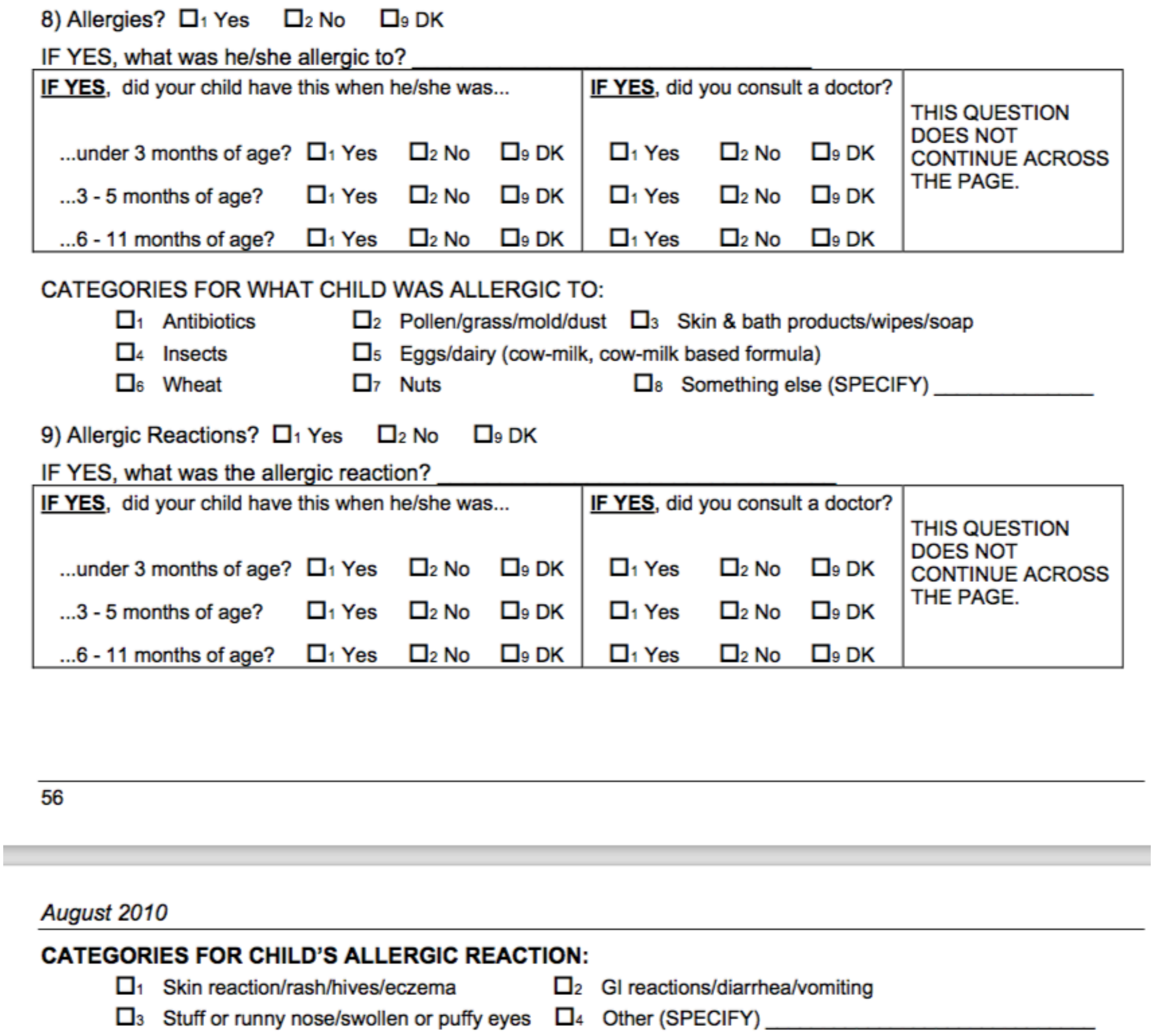
IF YES, what was the allergic reaction? ______________________________

IF YES, did your child have this when he/she was...				**IF YES**, did you consult a doctor?			THIS QUESTION DOES NOT CONTINUE ACROSS THE PAGE.
...under 3 months of age?	□1 Yes	□2 No	□9 DK	□1 Yes	□2 No	□9 DK	
...3 - 5 months of age?	□1 Yes	□2 No	□9 DK	□1 Yes	□2 No	□9 DK	
...6 - 11 months of age?	□1 Yes	□2 No	□9 DK	□1 Yes	□2 No	□9 DK	

Phase 3:

The next questions are about any illnesses and infections your child may have had in [fill CHILD]'s first year of life or reference date.

Did [*fill CHILD*] ever have.......



8) Allergies? □1 Yes □2 No □9 DK

IF YES, what was he/she allergic to? ______________________________

IF YES, did your child have this when he/she was...				**IF YES**, did you consult a doctor?			THIS QUESTION DOES NOT CONTINUE ACROSS THE PAGE.
...under 3 months of age?	□1 Yes	□2 No	□9 DK	□1 Yes	□2 No	□9 DK	
...3 - 5 months of age?	□1 Yes	□2 No	□9 DK	□1 Yes	□2 No	□9 DK	
...6 - 11 months of age?	□1 Yes	□2 No	□9 DK	□1 Yes	□2 No	□9 DK	

CATEGORIES FOR WHAT CHILD WAS ALLERGIC TO:

□1 Antibiotics □2 Pollen/grass/mold/dust □3 Skin & bath products/wipes/soap
□4 Insects □5 Eggs/dairy (cow-milk, cow-milk based formula)
□6 Wheat □7 Nuts □8 Something else (SPECIFY) _____________

9) Allergic Reactions? □1 Yes □2 No □9 DK

IF YES, what was the allergic reaction? ______________________________

IF YES, did your child have this when he/she was...				**IF YES**, did you consult a doctor?			THIS QUESTION DOES NOT CONTINUE ACROSS THE PAGE.
...under 3 months of age?	□1 Yes	□2 No	□9 DK	□1 Yes	□2 No	□9 DK	
...3 - 5 months of age?	□1 Yes	□2 No	□9 DK	□1 Yes	□2 No	□9 DK	
...6 - 11 months of age?	□1 Yes	□2 No	□9 DK	□1 Yes	□2 No	□9 DK	

56

August 2010

CATEGORIES FOR CHILD'S ALLERGIC REACTION:

□1 Skin reaction/rash/hives/eczema □2 GI reactions/diarrhea/vomiting
□3 Stuff or runny nose/swollen or puffy eyes □4 Other (SPECIFY) ________________________

Phase 5:

Now I am going to ask about allergic reactions to a variety of things.

40) **From birth to *[DIAG/REF DATE]*, did *[CHILD]* have allergic reactions to *[CATI: Fill in allergen]*?**		**What was *[CHILD's]* age when he/she had his/her first allergic reaction to *[CATI: Fill in allergen]*?**	**Overall, would you rate the severity of [*CHILD*]'s reaction to [*CATI: Fill in allergen*] as mild, moderate or severe?** 1 = Mild, 2 = Moderate 3 = Severe 9 = DK 00 = Refusal					**Did a doctor or a health professional ever tell you that [*CHILD*] was/is allergic to [*CATI: Fill in allergen*]?**
Allergen	**Yes, No, DK**	**Age**	**Mild**	**Moderate**	**Severe**	**DK**	**REF**	**Yes, No, DK**
40a) Foods	□1 Y □2 N □9 DK □00 Refusal	____ Years □2 Tested positive □9 DK □00 Refusal	1	2	3	9	00	□1 Y □2 N □9 DK □00 Refusal
40b) Animals or insects	□1 Y □2 N □9 DK □00 Refusal	____ Years □2 Tested positive □9 DK □00 Refusal	1	2	3	9	00	□1 Y □2 N □9 DK □00 Refusal
40c) House dust or molds	□1 Y □2 N □9 DK □00 Refusal	____ Years □2 Tested positive □9 DK □00 Refusal	1	2	3	9	00	□1 Y □2 N □9 DK □00 Refusal
40d) Plants, pollen, hay fever	□1 Y □2 N □9 DK □00 Refusal	____ Years □2 Tested positive □9 DK □00 Refusal	1	2	3	9	00	□1 Y □2 N □9 DK □00 Refusal
40e) Drugs or medications	□1 Y □2 N □9 DK □00 Refusal	____ Years □2 Tested positive □9 DK □00 Refusal	1	2	3	9	00	□1 Y □2 N □9 DK □00 Refusal
40f) Soap, cosmetics, detergents	□1 Y □2 N □9 DK □00 Refusal	____ Years □2 Tested positive □9 DK □00 Refusal	1	2	3	9	00	□1 Y □2 N □9 DK □00 Refusal
40g) Other	□1 Y □2 N □9 DK □00 Refusal	____ Years □2 Tested positive □9 DK □00 Refusal	1	2	3	9	00	□1 Y □2 N □9 DK □00 Refusal

1.6 References

1 Chang, J. S., Tsai, C. R., Tsai, Y. W. & Wiemels, J. L. Medically diagnosed infections and risk of childhood leukaemia: a population-based case-control study. *Int. J. Epidemiol.* **41**, 1050-1059, doi:10.1093/ije/dys113 (2012).

2 Jackson, K. D., Howie, L. D. & Akinbami, L. J. Trends in allergic conditions among children: United States, 1997-2011. *NCHS Data Brief*, 1-8 (2013).

3 Yazdanbakhsh, M., Kremsner, P. G. & van Ree, R. Allergy, parasites, and the hygiene hypothesis. *Science* **296**, 490-494, doi:10.1126/science.296.5567.490 (2002).

4 Chang, J. S., Wiemels, J. L. & Buffler, P. A. Allergies and childhood leukemia. *Blood Cells Mol. Dis.* **42**, 99-104, doi:10.1016/j.bcmd.2008.10.003 (2009).

5 Wiemels, J. Perspectives on the causes of childhood leukemia. *Chem. Biol. Interact.* **196**, 59-67, doi:10.1016/j.cbi.2012.01.007 (2012).

6 Black, A. B. & Meynell, M. J. Aleukaemic myeloid leukaemia presenting as aplastic anaemia. *Br. Med. J.* **1**, 1430-1431 (1951).

7 Siegel, D. A. *et al.* Cancer incidence rates and trends among children and adolescents in the United States, 2001-2009. *Pediatrics* **134**, e945-955, doi:10.1542/peds.2013-3926 (2014).

8 Bernsen, R. M., de Jongste, J. C. & van der Wouden, J. C. Birth order and sibship size as independent risk factors for asthma, allergy, and eczema. *Pediatr. Allergy Immunol.* **14**, 464-469 (2003).

9 Westergaard, T. A., P.; Pedersen, J.; Olsen, J.; Frisch, M.; Sorensen, H.; Wohlfahrt, J.; Melbye, M.;. Birth Characteristics, Sibling Patterns, and Acute Leukemia Risk in Childhood: a Population-Based Cohort Study. *J. Natl. Cancer Inst.* **89**, 939-947 (1997).

10 Hagerhed-Engman, L., Bornehag, C. G., Sundell, J. & Aberg, N. Day-care attendance and increased risk for respiratory and allergic symptoms in preschool age. *Allergy* **61**, 447-453, doi:10.1111/j.1398-9995.2006.01031.x (2006).

11 Ma, X. *et al.* Daycare attendance and risk of childhood acute lymphoblastic leukaemia. *Br. J. Cancer* **86**, 1419-1424, doi:10.1038/sj.bjc.6600274 (2002).

12 Thavagnanam, S. F., J.; Bromley, A.; Shields, M. D.; Cardwell, C. R.;. A meta-analysis of the association between Caesarian section and childhood asthma. *Clin. Exp. Allergy*, 629-633, doi:10.1111/j.1365-2222.2007.02780.x Clinical (2007).

13 Francis, S. S. *et al.* Mode of delivery and risk of childhood leukemia. *Cancer Epidemiol. Biomarkers Prev.* **23**, 876-881, doi:10.1158/1055-9965.EPI-13-1098 (2014).

14 Greaves, M. Infection, immune responses and the aetiology of childhood leukaemia. *Nat. Rev. Cancer* **6**, 193-203, doi:10.1038/nrc1816 (2006).

15 Dahl, S., Schmidt, L. S., Vestergaard, T., Schuz, J. & Schmiegelow, K. Allergy and the risk of childhood leukemia: a meta-analysis. *Leukemia* **23**, 2300-2304, doi:10.1038/leu.2009.162 (2009).

16 Linabery, A. M., Jurek, A. M., Duval, S. & Ross, J. A. The association between atopy and childhood/adolescent leukemia: a meta-analysis. *Am. J. Epidemiol.* **171**, 749-764, doi:10.1093/aje/kwq004 (2010).

17 Rudant, J. *et al.* Childhood acute leukemia, early common infections, and allergy: The ESCALE Study. *Am. J. Epidemiol.* **172**, 1015-1027, doi:10.1093/aje/kwq233 (2010).

18 Urayama, K. Y. *et al.* Early life exposure to infections and risk of childhood acute lymphoblastic leukemia. *Int. J. Cancer* **128**, 1632-1643, doi:10.1002/ijc.25752 (2011).

19 Chang, J. S., Tsai, Y. W., Tsai, C. R. & Wiemels, J. L. Allergy and risk of childhood acute lymphoblastic leukemia: a population-based and record-based study. *Am. J. Epidemiol.* **176**, 970-978, doi:10.1093/aje/kws263 (2012).

20 Lariou, M. S. *et al.* Allergy and risk of acute lymphoblastic leukemia among children: a nationwide case control study in Greece. *Cancer Epidemiol.* **37**, 146-151, doi:10.1016/j.canep.2012.10.012 (2013).

21 Crippa, A., Khudyakov, P., Wang, M., Orsini, N. & Spiegelman, D. A new measure of between-studies heterogeneity in meta-analysis. *Stat. Med.* **35**, 3661-3675, doi:10.1002/sim.6980 (2016).

22 Ma, X., Buffler, P. A., Layefsky, M., Does, M. B. & Reynolds, P. Control selection strategies in case-control studies of childhood diseases. *Am. J. Epidemiol.* **159**, 915-921 (2004).

23 Renz-Polster, H. *et al.* Caesarean section delivery and the risk of allergic disorders in childhood. *Clin. Exp. Allergy* **35**, 1466-1472, doi:10.1111/j.1365-2222.2005.02356.x (2005).

24 Jourdan-Da Silva, N. *et al.* Infectious diseases in the first year of life, perinatal characteristics and childhood acute leukaemia. *Br. J. Cancer* **90**, 139-145, doi:10.1038/sj.bjc.6601384 (2004).

25 Chang, J. S. *et al.* Maternal immunoglobulin E and childhood leukemia. *Cancer Epidemiol. Biomarkers Prev.* **18**, 2221-2227, doi:10.1158/1055-9965.EPI-09-0212 (2009).

26 Nunez-Enriquez, J. C. *et al.* Allergy and acute leukaemia in children with Down syndrome: a population study. Report from the Mexican inter-institutional group for the identification of the causes of childhood leukaemia. *Br. J. Cancer* **108**, 2334-2338, doi:10.1038/bjc.2013.237 (2013).

27 Gibson, R. *et al.* Epidemiology of diseases in adult males with leukemia. *J. Natl. Cancer Inst.* **56**, 891-898 (1976).

28 Viadana, E. & Bross, I. D. Use of the medical history to predict the future occurrence of leukemias in adults. *Prev. Med.* **3**, 165-170 (1974).

29 Zheng, W. *et al.* Prior medical conditions and the risk of adult leukemia in Shanghai, People's Republic of China. *Cancer Causes Control* **4**, 361-368 (1993).

30 Hughes, A. M. *et al.* Allergy and risk of childhood leukaemia: results from the UKCCS. *Int. J. Cancer* **121**, 819-824, doi:10.1002/ijc.22702 (2007).

31 Viechtbauer, W. Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software* **36**, 1-48 (2010).

32 R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria, 2015).

33 Cohen, J. *Statistical power analysis for the behavioral sciences*. 2nd edn, (L. Erlbaum Associates, 1988).

34 Magnani, C., Pastore, G., Luzzatto, L. & Terracini, B. Parental occupation and other environmental factors in the etiology of leukemias and non-Hodgkin's lymphomas in childhood: a case-control study. *Tumori* **76**, 413-419 (1990).

35 Nishi, M. & Miyake, H. A case-control study of non-T cell acute lymphoblastic leukaemia of children in Hokkaido, Japan. *J. Epidemiol. Community Health* **43**, 352-355 (1989).

36 Rosenbaum, P. F., Buck, G. M. & Brecher, M. L. Allergy and infectious disease histories and the risk of childhood acute lymphoblastic leukaemia. *Paediatr. Perinat. Epidemiol.* **19**, 152-164, doi:10.1111/j.1365-3016.2005.00634.x (2005).

37 Schuz, J., Morgan, G., Bohler, E., Kaatsch, P. & Michaelis, J. Atopic disease and childhood acute lymphoblastic leukemia. *Int. J. Cancer* **105**, 255-260, doi:10.1002/ijc.11054 (2003).

38 Soderberg, K. C., Jonsson, F., Winqvist, O., Hagmar, L. & Feychting, M. Autoimmune diseases, asthma and risk of haematological malignancies: a nationwide case-control study in Sweden. *Eur. J. Cancer* **42**, 3028-3033, doi:10.1016/j.ejca.2006.04.021 (2006).

39 Spector, L. *et al.* Medically recorded allergies and the risk of childhood acute lymphoblastic leukaemia. *Eur. J. Cancer* **40**, 579-584, doi:10.1016/j.ejca.2003.08.024 (2004).

40 Wen, W. *et al.* Allergic disorders and the risk of childhood acute lymphoblastic leukemia (United States). *Cancer Causes Control* **11**, 303-307 (2000).

41 Howard, L. M. & Wessely, S. The psychology of multiple allergy. *BMJ* **307**, 747-748 (1993).

42 Altman, D. R. & Chiaramonte, L. T. Public perception of food allergy. *J. Allergy Clin. Immunol.* **97**, 1247-1251 (1996).

43 Roujeau, J. C. Clinical heterogeneity of drug hypersensitivity. *Toxicology* **209**, 123-129, doi:10.1016/j.tox.2004.12.022 (2005).

44 Roman, E. *et al.* Childhood acute lymphoblastic leukemia and infections in the first year of life: a report from the United Kingdom Childhood Cancer Study. *Am. J. Epidemiol.* **165**, 496-504, doi:10.1093/aje/kwk039 (2007).

45 Slusky, D. A. *et al.* Potential role of selection bias in the association between childhood leukemia and residential magnetic fields exposure: A population-based assessment. *Cancer Epidemiol.* **38**, 307-313, doi:10.1016/j.canep.2014.02.010 (2014).

46 Galobardes, B. *et al.* Childhood wheezing, asthma, allergy, atopy, and lung function: different socioeconomic patterns for different phenotypes. *Am. J. Epidemiol.* **182**, 763-774, doi:10.1093/aje/kwv045 (2015).

47 Greiner, A. N., Hellings, P. W., Rotiroti, G. & Scadding, G. K. Allergic rhinitis. *Lancet* **378**, 2112-2122, doi:10.1016/S0140-6736(11)60130-X (2011).

48 Pols, D. H. *et al.* Atopic dermatitis, asthma and allergic rhinitis in general practice and the open population: a systematic review. *Scand. J. Prim. Health Care* **34**, 143-150, doi:10.3109/02813432.2016.1160629 (2016).

49 Hadley, C. Food allergies on the rise? Determining the prevalence of food allergies, and how quickly it is increasing, is the first step in tackling the problem. *EMBO Rep* **7**, 1080-1083, doi:10.1038/sj.embor.7400846 (2006).

50 Bloomfield, S. F., Stanwell-Smith, R., Crevel, R. W. & Pickup, J. Too clean, or not too clean: the hygiene hypothesis and home hygiene. *Clin. Exp. Allergy* **36**, 402-425, doi:10.1111/j.1365-2222.2006.02463.x (2006).

51 Hoste, E., Cipolat, S. & Watt, F. M. Understanding allergy and cancer risk: what are the barriers? *Nat. Rev. Cancer* **15**, 131-132 (2015).

2. A Germline Deletion in the *APOBEC3B* Gene and Risk of Childhood Acute Lymphoblastic Leukemia

2.1 The *APOBEC3B* polymorphism as an endogenous manifestation of immune dysregulation

The etiology of childhood ALL is influenced by germline genetic risk, as evidenced by twin and other genetic studies[1,2]. Disease concordance among monozygotic twin pairs is estimated to be around 10% whereas concordance among dizygotic twin pairs is exceedingly rare, suggesting that genetics contribute to disease pathogenesis independent of shared environment[3]. Genome-wide association studies have identified six genetic loci that confer risk of ALL. Candidate gene studies and studies of genetic ancestry further suggest that immune-related loci, such as the KIR complex, and Native American genetic ancestry may also confer some risk. As discussed in Section III.VI, a direct link between ALL heritability and immune dysregulation preceding disease remains ambiguous. The remaining hidden heritability, i.e. the ~90% of variance in disease occurrence attributable to genetics that remains unexplained, and the suggestive links to immune dysregulation indicate that there is more to discover.

One potential variant that could contribute to hidden heritability is a common germline deletion (~30Kb) in the *APOBEC3B* gene, which, due to scant linkage disequilibrium, cannot be accurately identified via array-based genome-wide association studies. This deletion is a rare example of a heritable genetic variant that can directly affect somatic mutation in cancer. The APOBEC3B enzyme is among the family of cytidine deaminases posited as driving somatic point mutation in t(12:21) ALL among other cancers that bear the TpC>T point mutation signature[4,5]. This enzyme family normally halts infection through TpC>T hypermutation of foreign nucleic acids, thus playing an integral part in innate immune response to DNA viruses, retroviruses, and retrotransposons[6]. Due to the APOBEC3B enzyme's nuclear localization, its high expression in certain cancers, and its ability to cause mutation to the human genome, it has become of primary interest in the etiology of cancers bearing its TpC>T point mutation signature, as has its germline deletion variant[6-8]. In addition to ALL, this somatic mutation signature is predominant in known virally induced cancer genomes (i.e. cervical, head and neck cancers)[4]. The common germline deletion in the *APOBEC3B* gene and resulting fusion transcript is associated with increased susceptibility to infections[9,10], an increased risk of breast and liver cancers[11,12], and the overrepresentation of the TpC>T somatic point mutation signature in tumor genomes of those harboring the deletion polymorphism (referred to as 'hypermutators')[13]. Among individuals with Native American genetic ancestry, the prevalence of the deletion is ~60%, and much lower in Europeans (~7%)[14].

The elevated allele frequency of the *APOBEC3B* deletion variant among Hispanics, an admixed population of Native American, African, and European[15] genetic ancestries, make it a compelling candidate to account for 1) the unexplained increase in genetic risk of ALL associated with Native American genetic ancestry and 2) the overall increased risk of ALL among Hispanics. APOBEC3B and the product of its deletion variant are primarily active during viral infection and endogenous retrotransposition[6]. In accordance with the leading etiologic hypothesis of childhood ALL, it is plausible that the abnormal exposure to infection experienced by children who go on to develop disease activates an immune response that includes APOBEC3 enzymatic activity, which in turn results in collateral damage to the genome. While it is clear that APOBEC3B enzymatic activity does not act alone to drive carcinogenesis, it significantly contributes to the landscape of somatic mutation present in the major subtype of ALL tumors and it's germline deletion polymorphism may explain the hidden heritability contributing to ALL risk.

The following article describes a study wherein we genotyped the *APOBEC3B* deletion polymorphism and examined its association with ALL in the CCLS. The polymorphism was evaluated using a PCR assay as evaluation of this variant could not be achieved using SNP array data. While an accurate tagging SNP of this variant was identified using the 1000 Genomes Phase 3 reference population (European r^2=0.90), neither the tagging SNP, nor the deletion variant could be reliably imputed in our data. In the California Childhood Leukemia Study dataset, the imputed tagging SNP was missing in nearly 60% of genotyped individuals and the imputation score for those successfully called was moderate at ~0.6. Further, CNV analysis using two SNP probes on the Illumina platform that exist within the deletion region also could not produce reliable genotypes. Heterozygous deletion carriers could not be differentiated from homozygous carriers, perhaps due to quenching of the probes. The technical challenges that restrict us to direct genotyping or whole genome sequencing are likely not limited to the *APOBEC3B* deletion, but extend to many structural variants genome-wide. Thus, it comes as no surprise that so much of the heritability of ALL and many complex diseases remains unexplained. Future research will require development of techniques geared toward identifying these types of variants.

2.2 LETTER TO THE EDITOR: A germline deletion of *APOBEC3B* does not contribute to subtype-specific childhood acute lymphoblastic leukemia etiology

Approximately 4/100,000 children are diagnosed with acute lymphoblastic leukemia (ALL) in the United States annually. Early life exposures related to immune priming (i.e. vaginal birth, daycare attendance, and high birth order) [16] and having fewer infections requiring medical treatment [17] have been inversely associated with disease, suggesting an etiologic role of infectious agents, perhaps via dysregulation of the immune system.

Patterns of somatic mutation in ALL tumors give further insight into disease etiology. An innate immune enzyme, APOBEC3B (Apolipoprotein B mRNA- editing enzyme, catalytic polypeptide-like 3B)*,* inhibits viral infection by inducing TpC>T nucleotide changes in foreign nucleic acid. This mutation signature has been identified in tumor genomes of several cancer types with known [cervical, head and neck, stomach [18]] or hypothesized [breast [19]] infectious etiologies, and is attributed to aberrant enzymatic activity of APOBEC3B. The APOBEC3B point mutation signature is also predominant in ALL [4], but with subtype specificity. The signature is present in *ETV6-RUNX1* fusion ALL [5] but absent in high hyperdiploid ALL [20]. The high expression of

APOBEC3B in lymphoblasts further justifies examination of functional APOBEC3B polymorphisms in ALL etiology.

A ~30kb germline deletion polymorphism at the *APOBEC3B* locus has been associated with increased risk of several cancers that bear the APOBEC3B mutation signature [21] and studies have shown that the deletion transcript yields an enzyme with a higher *in vitro* propensity for collateral genomic DNA damage than its wild type counterpart [13]. In fact, the signature TpC>T point mutation burden is higher in the tumors of ALL and breast cancer patients carrying the germline *APOBEC3B* deletion compared with those without [13]. The deletion is common in populations of Native American ancestry (~60%), and relatively rare in Europeans and Africans (6% and 0.9%, respectively) [22]. Hispanic children, whose genetic ancestry is typically comprised of a mixture of Native American, European, and African ancestries, are at greatest risk for developing ALL in the United States [23]. While it has been suggested that the *APOBEC3B* deletion polymorphism contributes to the patterns of somatic mutations observed in ALL tumors [13], it is not known whether the variant contributes to disease risk. Here, we report results from the first association study of germline *APOBEC3B* variants in childhood ALL risk.

The *APOBEC3B* deletion genotype was assessed in California Childhood Leukemia Study (CCLS) case and control subjects (see Supplementary Methods for enrollment details) with a PCR-based assay (n=1,126). The deletion was tested for association with childhood ALL status overall and within *ETV6-RUNX1* fusion and high hyperdiploid ALL subtypes. Overall, controls tended to be wealthier than cases with a higher proportion self-reporting as white and non-Hispanic (Table 1).

APOBEC3B deletion copy-number was detected using a validated polymerase chain reaction (PCR) method described previously [22]. A total of 518 ALL cases and 608 controls were genotyped using this PCR assay, with copy number (homozygote wild-type, heterozygote, and homozygote deletion) determined from agarose gel electrophoresis results (Figure S1). A chi-square test for Hardy-Weinberg equilibrium was performed; the null hypothesis of deletion equilibrium was accepted among controls after stratifying by Hispanic vs. non-Hispanic ethnicity (*p*-value=0.45 and 0.45, respectively). Ethnic heterogeneity in the CCLS population is supported by the distribution of multidimensional scaling (MDS) components compared to reference populations (Figure S2).

After adjusting for global genetic ancestry (first 3 MDS components), no association was observed between the *APOBEC3B* deletion and overall ALL risk for the additive, dominant, or recessive models, nor after stratification by cytogenetic subtype (Table 2). When study subjects were stratified by self-reported Hispanic status, results did not change (Tables S1).

Previous studies have identified SNPs within the *APOBEC3* region that are associated with cancer risk independent of the *APOBEC3B* deletion [21,24]. Thus, we tested all SNPs across the *APOBEC3* gene region (chr22:39,200,000-39,650,000) that passed quality filtering (8,275 SNPs) for association with ALL in 1,083 cases and 1,137 controls. After controlling for genetic ancestry, no variant reached statistical significance after correcting for multiple testing (Figure S3). A SNP ~20Kb upstream of *APOBEC3A* and previously associated with bladder cancer [21], rs1014971, was not associated with

ALL risk (OR 0.91, *p*-value=0.33). The top association was seen for rs73424730 (OR 1.35, *p*-value=0.004), a SNP ~100Kb downstream of the *APOBEC3H* gene.

To determine whether the observed absence of association between the *APOBEC3B* deletion and ALL was due to confounding by genetic ancestry, local ancestry at the *APOBEC3* megalocus was inferred using a discriminative modeling approach, RFMix [25]. The *APOBEC3B* deletion polymorphism is a highly population-stratified genetic variant [22]. Thus, in the admixed CCLS population, it is possible that adjustment for global genetic ancestry was insufficient, resulting in residual confounding or a washing-out of the true effect of interest. After adjusting for regional ancestral proportions for four continental ancestries, there remained no association between the *APOBEC3B* deletion variant and childhood ALL overall (Table S2). To ensure that effect heterogeneity was not washing-out true associations, RFmix assigned genetic ancestries (African, Native American, European, and East Asian) at each SNP in the *APOBEC3* locus were tested for independent association with ALL to determine whether local genetic ancestry was associated with disease. African ancestry at the *APOBEC3* megalocus was nominally associated with ALL risk, but did not reach statistical significance after correcting for multiple tests (Figure S4).

Despite the previously observed presence of the APOBEC3B-mediated point mutation signature in the tumor genomes of a subset of ALL patients, and a higher burden of point mutations among *APOBEC3B* deletion carriers, we found no apparent relationship between the germline *APOBEC3B* deletion and risk of developing the disease. Moreover, we found no evidence of association between any SNPs across the *APOBEC3* gene region and childhood ALL, including at a locus previously associated with the APOBEC mutation signature in bladder cancer [26].

Evidence from studies of other cancers suggests a complex relationship between germline variation at the *APOBEC3* gene region and tumorigenesis, for instance the deletion is associated with an increased risk of breast cancer [13] but appears protective in bladder cancer [21]. Further, the precise contribution of germline *APOBEC* polymorphisms to the presence of the APOBEC mutation signature in an individual tumor has yet to be determined [27]. There is evidence that the *APOBEC3B* deletion polymorphism changes the mutagenic behavior of the enzyme [13]; however, the presence of the mutational signature in a tumor genome does not imply that the deletion is present, as the signature could arise by some other means (i.e. increased enzyme expression by other mechanisms).

Childhood ALL is a rare and heterogeneous disease, and more prevalent in admixed populations, making the study of any potential germline genetic risk factors challenging. While it is unlikely that the *APOBEC3B* deletion polymorphism is a strong independent risk factor for disease overall, our sample size was limited in statistical power to assess small effects, especially following stratification by cytogenetic subtype. Associations observed in other cancers suggest the deletion alters disease risk by 20-30% [21]. These cancers have mixed somatic mutation signatures, suggesting mutagenesis occurs from multiple sources [4]. Conversely, the APOBEC mutational signature is the only one observed in ALL other than common, spontaneous cytosine deamination[4]; therefore, we hypothesized a moderate association with the *APOBEC3B* deletion would be present in ALL. Further, the CCLS study population is representative of cases occurring in the state of California, and reflects the substantial racial and ethnic diversity therein. Thus, residual confounding by genetic admixture remains a challenge in investigating this variant, which differs significantly in frequency across populations based on ancestral origin. However, stratifying by self-reported Hispanic status in this study yielded no apparent association in either group. To further account for potential confounding effects of admixture, local ancestry at the *APOBEC3* megalocus was estimated. Adjusting for local ancestral proportions did not change the observed null association of the *APOBEC3B* deletion polymorphism with ALL risk, nor were any of the four observed regional genetic ancestries associated independently with ALL risk.

The apparent lack of association between the germline deletion of *APOBEC3B* and risk of ALL does not provide support that this mutagen, active in tumor cells, is a driver of tumorigenesis. In a previous study, our group showed expression of double-stranded (ds)DNA viruses in primary, treatment naïve, childhood ALL tumors [28]. dsDNA viruses are a primary target of the APOBEC3B enzyme. Transient expression of these viruses in leukemic cells could thus produce a passenger-type APOBEC3B point mutation signature in ALL. A virus-mediated induction of APOBEC3B expression may be dominant over the impact on APOBEC3B expression by the polymorphism studied here. Alternatively, the APOBEC mutational signature may reflect an infectious etiology of ALL in a subset of cases with heritable predisposition occurring in unrelated genes. Experimental studies in mice have shown that PAX5 mutation, a predisposing risk factor in some childhood ALL cases, can result in development of ALL following exposure to common infection [29]. Though there is no apparent relationship between the germline *APOBEC3B* deletion polymorphism and ALL risk, delineating a potential role for the polymorphism in tumor progression may have treatment implications, warranting further study.

2.3 Tables 2.1-2.2: A germline deletion of *APOBEC3B*

Table 2.1. CCLS Study Participant Characteristics

	Cases (n=1083)		Controls (n=1137)	
	N	%	N	%
Tagging SNP Genotyped	1080	99.7	901	79.2
PCR Genotyped	518	47.8	608	53.5
Interviewed	975	90.0	1137	100.0
Mean Age at Dx (Range)	5.54(0.0-14.96)		5.49(0.0-14.93)	
Missing	*124*	*11.4*	*0*	*0.0*
Sex				
Male	621	57.3	652	57.3
Female	462	42.7	485	42.7
Missing	*0*	*0.0*	*0*	*0.0*
Ethnicity				
Hispanic	498	46.0	515	45.3
Non-Hispanic	461	42.6	622	54.7
Missing	*124*	*11.4*	*0*	*0.0*
Maternal Race				
White/Caucasian	787	72.7	983	86.5
African American	34	3.1	34	3.0
Mixed/Other	138	12.7	120	10.6
Missing	*124*	*11.4*	*0*	*0.0*
Household Income				
<15,000	154	14.2	109	9.6
15,000-29,999	187	17.3	150	13.2
30,000-44,999	143	13.2	141	12.4
45,000-59,999	133	12.3	162	14.2
60,000-74,999	59	5.4	125	11.0
≥75,000	283	26.1	450	39.6
Missing	*124*	*11.4*	*0*	*0.0*

Table 2.2: Odds ratios for the association between the APOBEC3B deletion polymorphism and risk of childhood ALL overall and stratified by cytogenetic subtype

	Cases (n)	Controls (n)	Model	OR*	95% CI	*P*-Value
Overall	518	608	Addititve	0.96	0.77-1.22	0.77
wt/wt	360	441	-	ref	-	-
wt/del	143	144	Dominant	1.03	0.79-1.36	0.81
del/del	15	23	Recessive	0.62	0.31-1.21	0.16
Common ALL	146	608				
wt/wt	100	441	-	ref	-	-
wt/del;del/del	46	167	Dominant	1.08	0.71-1.62	0.73
Hyperdiploid	117	608				
wt/wt	80	441	-	ref	-	-
wt/del;del/del	37	167	Dominant	0.84	0.54-1.34	0.47
t1221	64	608				
wt/wt	47	441	-	ref	-	-
wt/del;del/del	17	167	Dominant	1.10	0.60-2.07	0.77

OR odds ratio; CI confidence interval; *wt* wildtype; *del APOBEC3B* deletion; ref reference
*Adjusted for global genetic ancestry and sex

2.4 Supplementary Materials: A germline deletion of *APOBEC3B*

2.4.1 Supplementary Methods

Individuals were enrolled in the CCLS between 1995 and 2015 with rapid, comprehensive case ascertainment from 80% of California hospitals, allowing capture of ~76% of all cases, usually within 72 hours of diagnosis. Controls were individually matched to cases on age, sex, child's Hispanic ethnicity, and maternal race [30]. For consented individuals, saliva or buccal swab specimens were collected at the time of interview. Among cases, cytogenetic features of primary tumors were abstracted from medical records. The CCLS was approved by the University of California Berkeley Institutional Review Board and by all participating institutions. Informed consent was obtained from all participating subjects. A subset of subjects with biospecimens and genetic data available were selected for the present genetic study and were largely representative of the CCLS study population (Table S1). Subjects with Down syndrome (n=52) were excluded.

PCR-Based APOBEC3B Deletion Genotyping: DNA was extracted from buccal and saliva samples and resuspended in Tris EDTA buffer. Copy-number of the deletion was assessed using a validated polymerase chain reaction (PCR) method described in Kidd *et al* [14]. In brief, PCR amplification was carried out using the AmpliTaq Gold™ DNA Polymerase with Buffer II and $MgCl_2$ (ThermoFisher Scientific) following manufacturer's protocol. Two PCR reactions were run for each subject, one using a primer pair spanning the deletion breakpoints and resulting in a 700bp product (undeleted product is too large to be amplified), and another internal to the deletion breakpoints and producing a 490bp product. PCR products were combined and run on a 2% agarose gel by electrophoresis and visualized to determine genotype: subjects displaying no product for the "deletion" primer pair but showing the 490bp product for the "internal" primers were classed as homozygous undeleted; subjects with the 700bp deletion product plus the 490bp internal product were classed as heterozygous deleted; and subjects with the 700bp deletion product but no internal product were homozygous deleted (Figure S1). Each individual was genotyped once, and those individuals that appeared homozygous for the deletion were genotyped again using an independent set of primers [14]. A positive result from the second reaction implied a heterozygous genotype (primer pair sequences available on request).

SNP Genotyping. In the CCLS, genome-wide genotype data were previously produced from DNA extracted from dried blood spots, saliva, or buccal cells genotyped using the Illumina HumanOmniExpress 12v1-1, HumanOmniExpressExome 8v1-2, and InfiniumOmniExpress 8v1-4 platforms, containing >700,000 SNP markers. SNPs with a call-rate less than 90% were excluded, as were individuals genotyped at less than 90% of markers.

SNP Imputation. Probes lying within the APOBEC3B deletion region (chr22: 39,357,694-39,388,574) were removed. Imputation was then carried out with IMPUTE2 [31] using the 1000 Genomes Phase 3 reference haplotypes.

Genetic Ancestry Estimation. Multidimensional scaling (MDS) components were derived using PLINK1.9 [32]. Imputed SNPs were pruned for independence and the singular value decomposition-based algorithm was performed on an inter-sample distance matrix. The first 3 components were included in adjusted models to account for global genetic ancestry.

Local Genetic Ancestry Inference. The *APOBEC3* megalocus on chromosome 22 contains seven ABOBEC3 gene-family members (APOBEC3A-D, F-H). SNP-wise local genetic ancestry was estimated for 2,899 SNPs at the APOBEC3 megalocus (chr22:39,200,000-39,650,000) after removal of multi-allelic SNPs, SNPs with missing reference or alternate alleles, and SNPs within the *APOBEC3B* deletion region. First, the 2,899 selected SNPs were phased using BEAGLE 4.0 [33] independently in the CCLS and in five reference ancestral populations included in Phase 3 of the 1000 Genomes. Following phasing, local ancestry was inferred in the admixed CCLS data based on the five reference populations using RFMix, wherein ancestry was estimated for 65 0.1cM windows over three EM-iterations. Ancestral proportions across the APOBEC3 megalocus were calculated for each individual as the proportion of haplotypes assigned to each of the 5 reference ancestries. Dummy variables for the additive and dominant contribution of each reference ancestry were also assigned SNP-wise for each individual after pruning SNPs for local ancestry-based independence.

Statistical Analyses. Logistic regression was used to determine the association of the ~30kb deletion of *APOBEC3B* (assessed by PCR and tagging SNP, respectively) with ALL under three different models of inheritance: additive, autosomal dominant, and recessive, adjusted for genetic ancestry and sex. Stratified analyses by Hispanic status and cytogenetic subtype were also conducted. All PCR-based analyses were carried out using R[34] and all tagging SNP-based analyses were carried out using SNPTest v2.5.2 to account for imputation uncertainty [35]. Association analysis of *APOBEC3* gene region SNPs was carried out in PLINK 1.9 using logistic regression tests under the additive model, adjusting for genetic ancestry [32].

2.4.2 Supplementary Tables 2.1-2.2

Table S2.1: Odds ratios for the association between the *APOBEC3B* deletion polymorphism and risk of childhood ALL stratified by self-reported Hispanic ethnicity

Hispanic	Cases (n)	Controls (n)	Model	OR*	95% CI	*P*-Value
Total	270	288	Additive	1.18	0.77-1.60	0.28
wt/wt	164	195	-	ref		-
wt/del	94	81	Dominant	1.28	0.90-1.83	0.18
del/del	12	12	Recessive	0.94	0.40-2.17	0.88
Non-Hispanic White	Cases (n)	Controls (n)	Model	OR*	95% CI	*P*-Value
Total	236	320	Additive	0.67	0.38-1.16	0.16
wt/wt	187	246	-	ref		-
wt/del	46	63	Dominant	0.64	0.35-1.15	0.14
del/del	3	11	Recessive	0.81	0.37-8.70	0.86

OR odds ratio; CI confidence interval; *wt* wildtype; *del APOBEC3B* deletion; ref reference

*Adjusted for genetic ancestry and sex

Table S2.2: Odds ratios for the association between the *APOBEC3B* deletion polymorphism adjusted for continental ancestral proportions in the *APOBEC3* gene region

Regional ancestral proportion adjustment	Cases (n)	Controls (n)	Model	OR*	95% CI	*P*-Value
African	518	608	Additive	0.99	0.79-1.23	0.89
wt/wt	360	441	-	ref	-	-
wt/del	143	144	Dominant	1.05	0.80-1.37	0.73
del/del	15	23	Recessive	0.66	0.34-1.27	0.23
Native American	518	608	Additive	0.95	0.76-1.20	0.68
wt/wt	360	441	-	ref	-	-
wt/del	143	144	Dominant	1.01	0.77-1.33	0.92
del/del	15	23	Recessive	0.62	0.31-1.19	0.16
European	518	608	Additive	0.99	0.79-1.24	0.92
wt/wt	360	441	-	ref	-	-
wt/del	143	144	Dominant	1.05	0.80-1.38	0.7
del/del	15	23	Recessive	0.66	0.34-1.27	0.22
East Asian	518	608	Additive	1.02	0.82-1.28	0.86
wt/wt	360	441	-	ref	-	-
wt/del	143	144	Dominant	1.09	0.84-1.42	0.52
del/del	15	23	Recessive	0.7	0.36-1.34	0.28

OR odds ratio; CI confidence interval; *wt* wildtype; *del APOBEC3B* deletion; ref reference

*Adjusted for local genetic ancestry

2.4.2 Supplementary Figures 2.1-2.4

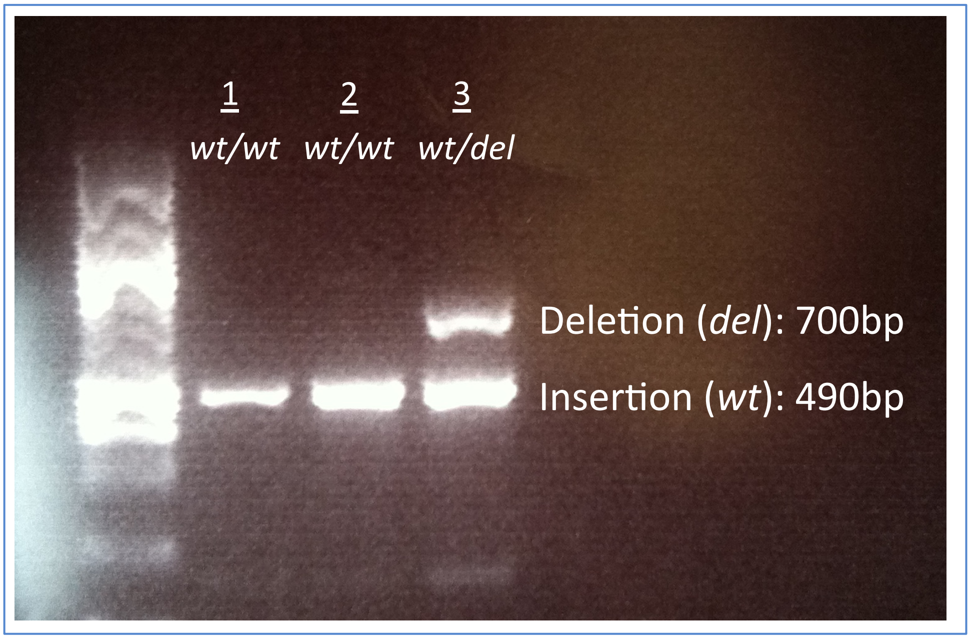


Figure S2.1 APOBEC3B PCR-based genotyping assay. Individuals 1, 2 homozygous non-carries of the deletion, individual 3 is heterozygous, carrying one copy of the ~30Kb deletion polymorphism.

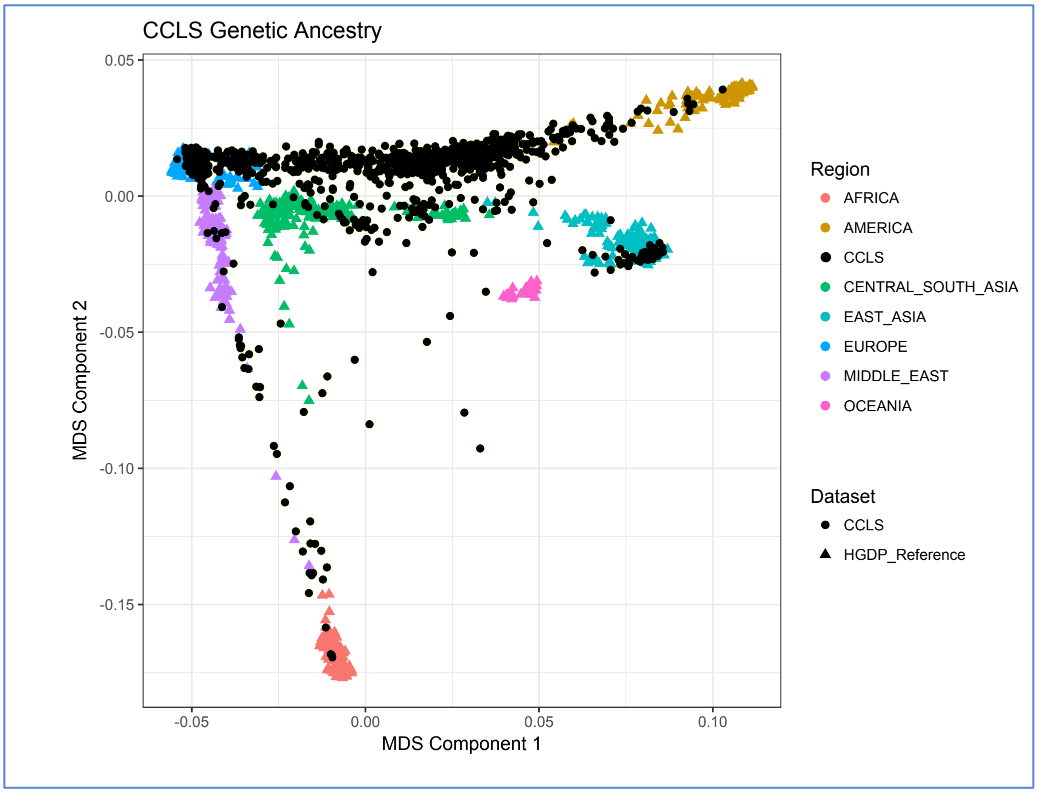


Figure S2.2 MDS plot comparing genetic ancestry of *A3B* genotyped CCLS subjects to the Human Genome Diversity Project reference

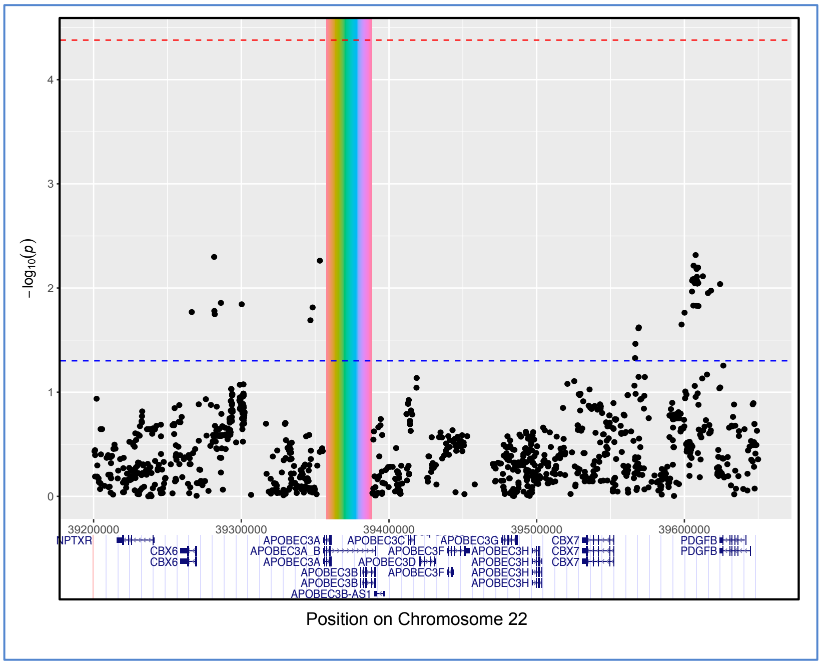


Figure S2.3 CCLS *APOBEC3* Region GWAS. Blue dashed line significance threshold *p*=0.05; red dashed line Bonferroni adjusted significance threshold *p*=0.00004; rainbow box *APOBEC3B* deletion position

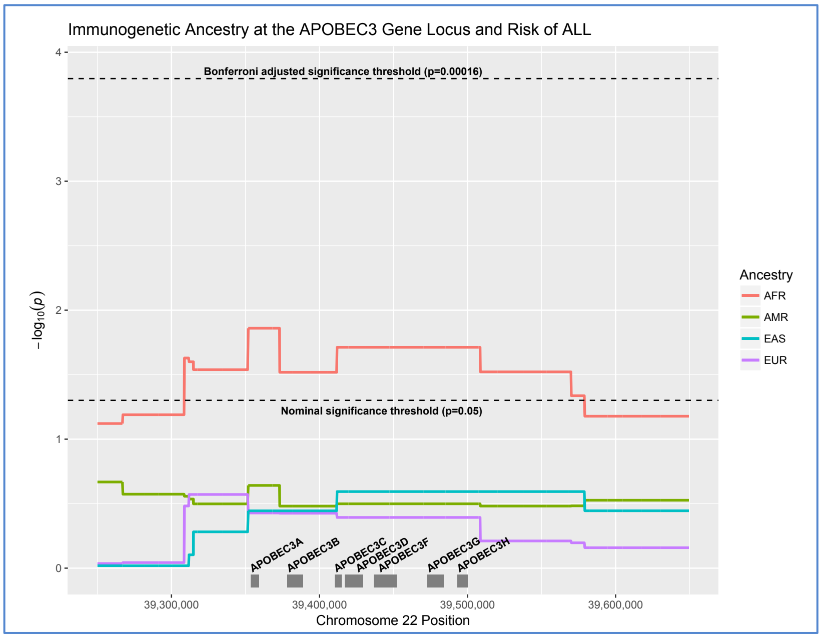


Figure S2.4 Association of immunogenetic ancestry at the APOBEC3 megalocus and ALL risk

2.5 References

1 Sherborne, A. L. & Houlston, R. S. Risk of Childhood Acute Lymphoblastic Leukemia: Identification of Inherited Susceptibility. **4**, 105-111, doi:10.1007/978-94-007-6591-7_11 (2013).

2 Levine, R. L. Inherited susceptibility to pediatric acute lymphoblastic leukemia. *Nat. Genet.* **41**, 957-958, doi:10.1038/ng0909-957 (2009).

3 Greaves, M. F., Maia, A. T., Wiemels, J. L. & Ford, A. M. Leukemia in twins: lessons in natural history. *Blood* **102**, 2321-2333, doi:10.1182/blood-2002-12-3817 (2003).

4 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421, doi:10.1038/nature12477 (2013).

5 Papaemmanuil, E. *et al.* RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia. *Nat. Genet.* **46**, 116-125, doi:10.1038/ng.2874 (2014).

6 Vieira, V. C. & Soares, M. A. The role of cytidine deaminases on innate immune responses against human viral infections. *Biomed Res Int* **2013**, 683095, doi:10.1155/2013/683095 (2013).

7 Cescon, D. W., Haibe-Kains, B. & Mak, T. W. APOBEC3B expression in breast cancer reflects cellular proliferation, while a deletion polymorphism is associated with immune activation. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 2841-2846, doi:10.1073/pnas.1424869112 (2015).

8 Burns, M. B. *et al.* APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature* **494**, 366-370, doi:10.1038/nature11881 (2013).

9 An, P. *et al.* APOBEC3B deletion and risk of HIV-1 acquisition. *J. Infect. Dis.* **200**, 1054-1058, doi:10.1086/605644 (2009).

10 Jha, P. *et al.* Deletion of the APOBEC3B gene strongly impacts susceptibility to falciparum malaria. *Infect. Genet. Evol.* **12**, 142-148, doi:10.1016/j.meegid.2011.11.001 (2012).

11 Xuan, D. *et al.* APOBEC3 deletion polymorphism is associated with breast cancer risk among women of European ancestry. *Carcinogenesis* **34**, 2240-2243, doi:10.1093/carcin/bgt185 (2013).

12 Zhang, T. *et al.* Evidence of associations of APOBEC3B gene deletion with susceptibility to persistent HBV infection and hepatocellular carcinoma. *Hum. Mol. Genet.* **22**, 1262-1269, doi:10.1093/hmg/dds513 (2013).

13 Nik-Zainal, S. *et al.* Association of a germline copy number polymorphism of APOBEC3A and APOBEC3B with burden of putative APOBEC-dependent mutations in breast cancer. *Nat. Genet.* **46**, 487-491, doi:10.1038/ng.2955 (2014).

14 Kidd, J. N. T. T., E.; Kaul, R.; Eichler, E.; . Population Stratification of a Common APOBEC Gene Deletion Polymorphism. *PLoS Genetics* **3**, doi:10.1371 (2007).

15 Walsh, K. M. *et al.* Associations between genome-wide Native American ancestry, known risk alleles and B-cell ALL risk in Hispanic children. *Leukemia* **27**, 2416-2419, doi:10.1038/leu.2013.130 (2013).

16 Buffler, P. A. S., S.; Matthay, KK.; Wiencke, JK.; Wiemels, JL.; Reynolds, P.;. Daycare attendance and risk of childhood acute lymphoblastic leukaemia. *Br. J.*

Cancer **86**, 1419-1424, doi:10.1038/sj/bjc/6600274 http://www.bjcancer.com (2002).

17 Chang, J. S., Tsai, C. R., Tsai, Y. W. & Wiemels, J. L. Medically diagnosed infections and risk of childhood leukaemia: a population-based case-control study. *Int. J. Epidemiol.* **41**, 1050-1059, doi:10.1093/ije/dys113 (2012).

18 de Martel, C. *et al.* Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. *Lancet Oncol.* **13**, 607-615, doi:10.1016/S1470-2045(12)70137-7 (2012).

19 Lawson, J. S. & Heng, B. Viruses and breast cancer. *Cancers (Basel)* **2**, 752-772, doi:10.3390/cancers2020752 (2010).

20 Paulsson, K. *et al.* The genomic landscape of high hyperdiploid childhood acute lymphoblastic leukemia. *Nat. Genet.* **47**, 672-676, doi:10.1038/ng.3301 (2015).

21 Middlebrooks, C. D. *et al.* Association of germline variants in the APOBEC3 region with cancer risk and enrichment with APOBEC-signature mutations in tumors. *Nat. Genet.* **48**, 1330-1338, doi:10.1038/ng.3670 (2016).

22 Kidd, J. M., Newman, T. L., Tuzun, E., Kaul, R. & Eichler, E. E. Population stratification of a common APOBEC gene deletion polymorphism. *PLoS Genet* **3**, e63, doi:10.1371/journal.pgen.0030063 (2007).

23 Barrington-Trimis, J. L. *et al.* Rising rates of acute lymphoblastic leukemia in Hispanic children: trends in incidence from 1992 to 2011. *Blood* **125**, 3033-3034, doi:10.1182/blood-2015-03-634006 (2015).

24 Gohler, S. *et al.* Impact of functional germline variants and a deletion polymorphism in APOBEC3A and APOBEC3B on breast cancer risk and survival in a Swedish study population. *J. Cancer Res. Clin. Oncol.* **142**, 273-276, doi:10.1007/s00432-015-2038-7 (2016).

25 Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278-288, doi:10.1016/j.ajhg.2013.06.020 (2013).

26 Rothman, N. *et al.* A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. *Nat. Genet.* **42**, 978-984, doi:10.1038/ng.687 (2010).

27 Chan, K. *et al.* An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat. Genet.* **47**, 1067-1072, doi:10.1038/ng.3378 (2015).

28 Francis, S. S. *et al.* In utero cytomegalovirus infection and development of childhood acute lymphoblastic leukemia. *Blood* **129**, 1680-1684, doi:10.1182/blood-2016-07-723148 (2017).

29 Martin-Lorenzo, A. *et al.* Infection Exposure is a Causal Factor in B-cell Precursor Acute Lymphoblastic Leukemia as a Result of Pax5-Inherited Susceptibility. *Cancer Discov.* **5**, 1328-1343, doi:10.1158/2159-8290.CD-15-0892 (2015).

30 Ma, X., Buffler, P. A., Layefsky, M., Does, M. B. & Reynolds, P. Control selection strategies in case-control studies of childhood diseases. *Am. J. Epidemiol.* **159**, 915-921 (2004).

31 Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529, doi:10.1371/journal.pgen.1000529 (2009).

32 Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7, doi:10.1186/s13742-015-0047-8 (2015).

33 Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084-1097, doi:10.1086/521987 (2007).

34 R: A language and environment for statistical computing (R Foundation for Statistical Computing, 2012).

35 Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906-913, doi:10.1038/ng2088 (2007).

3. The Role of Polymorphic Human Endogenous Retrovirus-K Insertions in Multiple Diseases

3.1 The propensity for immunomodulation by polymorphic human endogenous retrovirus-K in ALL and other diseases

The mounting epidemiologic evidence suggesting that infection plays a role in the etiology of childhood ALL led our group to pose the following questions: Is there a single infectious agent that causes childhood ALL? And if so, can it be identified in diagnostic bone marrow samples from ALL cases? To answer these questions we conducted a study of untargeted viral and bacterial metagenomics in children with ALL, and as a comparison group, in children with acute myeloid leukemia – a disease that does not have an infectious etiology. Two interesting differences were observed among viral transcripts derived from ALL vs. AML bone marrows. First was the presence of cytomegalovirus (CMV), a common infection from a family of viruses, the herpesviridae, with known cancer causing members, and the second was elevated expression of human endogenous retroviruses (HERVs), which are known to be aberrantly expressed in tumor cells and also to activate in the presence of herpesvirus infection (Figure 3.1.1).

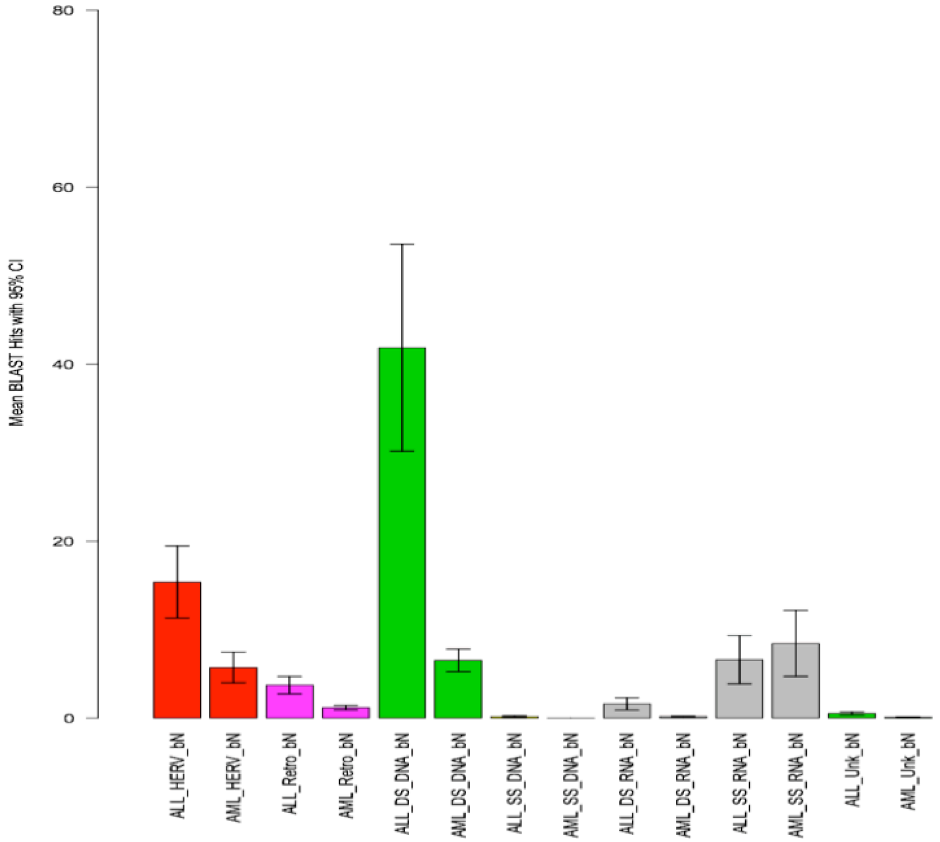


Figure 3.1.1 RNA extracted from diagnostic bone marrow samples from 60 ALL cases and 31 AML cases were pooled into 3 pools each, sequenced, and queried for the presence of viral transcripts. The plot represents the average number of sequencing reads per pool that hit to

different classes of virus using blastn in acute ALL and AML cases, respectively. HERV human endogenous; Retro retrovirus; DS_DNA double stranded DNA virus; SS_DNA single stranded DNA virus; DS_RNA double stranded RNA virus; SS_RNA single stranded RNA virus; Unk unknown virus;

HERVs have long been suspected as a cause of cancer[1]. In fact, HERV transcript expression has been observed in tumor cells from multiple cancer types and several other factors contribute to a compelling mechanistic hypothesis by which they could cause disease (Figure 3.1.2).

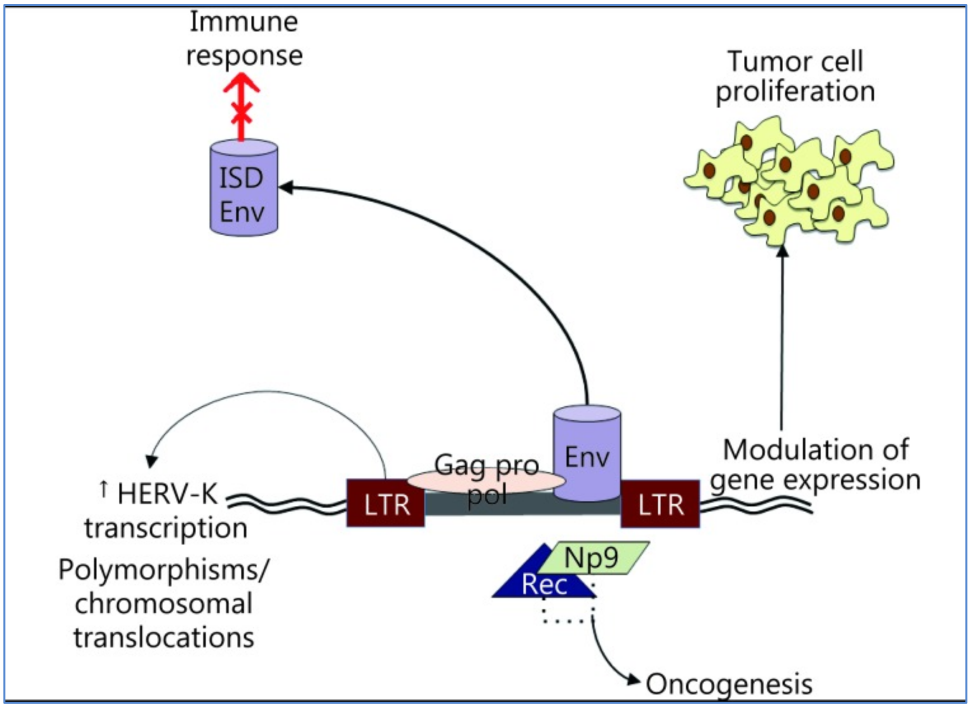


Figure 3.1.2 Mechanisms of oncogenesis. Long terminal repeats (LTRs) recruit transcription factors from the infected cell for retroviral gene transcription. These LTRs can also enhance transcription of the host cell genes, leading to uncontrolled tumor cell proliferation. Some HERVs encode potentially oncogenic proteins like Np9 and Rec that interact with transcription factors and activate immunosuppressor pathways, promoting oncogenesis. HERVs can also induce chromosomal translocations in somatic cells that could lead to tumor proliferation. HERVs also can promote an immunosuppressive response that may lead to cancer formation and spreading, because the Env protein has an immunosuppressive domain (ISD). Source: *Cancer Biol Med. 2016 Dec; 13(4): 483–488*

Following the identification of HERV transcripts in ALL bone marrows, we gained a strong interest in the potential mechanisms by which HERVs might contribute to disease risk. Perhaps the most interesting hypothesis considers the immunomodulatory properties of HERVs, which can be either suppressive or activating. For example, immune suppression is evident during pregnancy where HERVs play an apparent role in regulating immune tolerance at the maternal-fetal interface[2]. Immune activating activities, in contrast, can occur in the presence of exogenous viral infections, wherein specific viruses, including CMV, have been shown to activate HERV superantigen expression, resulting in constitutive, non-specific activation of T cells[3]. In fact, HERVs have been implicated in the etiology of HIV and Epstein Barr virus-associated lymphomas where each of these viruses are known to activate HERV-encoded superantigen production and a ~7-fold increase in HERV transcript expression is detected in these lymphomas compared to healthy cells[4]. Given that we recently identified a 4-fold increased risk of ALL with *in utero* CMV infection, we hypothesized that a similar biological mechanism could underlie the interaction between CMV and HERVs in ALL. Further, as described in Section 2.1, APOBEC enzymes are activated in response to HERV transcription and exert strong restriction of these elements via hypermutation *in vitro*[5]. This activity may also contribute to somatic mutations that trigger onset of leukemia in children.

Until recently, it was thought that the ancient retroviral integrations of HERVs were fixed across humans, thus limiting their ability to differentially affect disease risk. As such, HERVs have been largely overlooked in studies of disease etiology. However, recent studies have shown that HERV-K, the youngest endogenous retroviral family in humans, is indeed polymorphic in some locations in the genome. The most comprehensive characterization of polymorphic HERV-K insertions in human genomes was published in 2016[6] and so they are not present in human reference panels of genetic variation. As such, none have been tested for association with human disease in genome-wide association studies. As a first step in our ultimate goal to investigate the role of HERVs in ALL, we developed a computational pipeline to identify HERV-K insertional polymorphisms to 1) identify additional novel polymorphic HERV-K loci to the ~40 already known and 2) test all insertions for association with human diseases that have established genetic associations in the literature. The article below describes our computational pipeline and the results of this study, which lends credence to future applications in the context of ALL as discussed in the conclusion in Section 4.

3.2 ARTICLE: To ERV is human: A phenotype-wide scan linking polymorphic human endogenous retrovirus-K insertions to complex phenotypes

INTRODUCTION

Retroviruses are a class of RNA virus that undergoes reverse transcription to DNA during the infectious cycle inside a host cell. At the proviral stage, retroviral DNA integrates into the host DNA to produce viral proteins. Integration into germ cells can result in endogenization, wherein the virus can be vertically transmitted via standard Mendelian inheritance mechanisms. Endogenous retroviruses (ERV) are ancient examples of proviruses that integrated and endogenized into the human genome >40mya [7]. In modern humans, ERVs account for approximately 8% of the genome [7]. Their relative stability, as well as the conservation of orthologs in other primate genomes suggests that they induce genome plasticity and can enhance evolutionary fitness [8,9]. Retroviruses are reliant on the fitness of their host for survival and the long-standing evolutionary cooperation between ERVs and humans may represent a symbiotic relationship [10]. The positive selection of persistent ERVs in the genome may have resulted from increasing the probability of survival to reproductive age [via adaptive effects on placentation [11]; and immune [12] and brain development [13]]. The phenotypic effects of ERVs on the post-reproductive adult, however, remain unclear and are of growing interest [14] [15] [16].

Previous studies have described the potential mechanisms by which ERVs influence human phenotypes. ERV insertions introduce viral genes and, due to their inter-individual homology can generate copy-number variants via non-allelic homologous recombination [17]. They modify transcription by adding enhancers [18] and promoters [19], disrupting intron structure, causing RNA interference [20] adding poly-A tails [21], and altering DNA methylation [22]. ERV expression, typically restricted in healthy tissues except during placental development, is detected in diseases including cancers and autoimmune disorders [reviewed in [23,24]].

While the functional effects of ERVs are well established, the vast majority of ERV insertions are fixed across individuals, and so their potential contribution to phenotypic variation has been largely overlooked. The HERV-K (HML-2) subfamily, however, contains human-specific, unfixed insertions, ranging from fully intact provirus to solo long terminal repeat (LTR) sequences [6]. HERV-K represents the most recent ERV integration into the human genome and numerous insertions have neither been eliminated from the genome (via negative selection or drift), nor fixed (via positive selection or drift). The polymorphic nature of these insertions suggests a potential contribution to causal variation in the heritability of complex phenotypes. Targeted studies have identified specific HERV-K integrations that affect disease risk, for example a polymorphic HERV-K inserted within the complement component 4 (C4A/B) gene confers strong genetic risk of schizophrenia [14].

Technical limitations have proved a major obstacle in the untargeted identification of polymorphic HERV-K insertions for application to clinical and epidemiologic studies. With the emergence of next-generation sequencing technologies, methods are being developed for the untargeted identification of ERVs among other mobile genetic elements in human genomes[6,25,26]. Here, we examine phenotypic effects of all polymorphic HERV-K insertions identifiable from a large, publically available whole genome sequencing (WGS) dataset. With our computational pipeline, we identified HERV-K insertion locations using data from the diverse 1000 Genomes Phase 3 population (n=2,504). By identifying a subset of polymorphic HERV-K insertions with strong associations to adjacent ‘tagging’ SNPs, we have leveraged several comprehensive SNP annotation databases to test for enrichment of established relationships between HERV-K insertion-associated SNPs (hiSNPs), tissue-specific gene expression, and diverse disease phenotypes.

METHODS

To elucidate the broad phenotypic effects of polymorphic HERV-K insertions, we developed a computational pipeline to identify the presence/absence of known and novel HERV-K insertions in individual WGS data (Fig S1, https://github.com/unreno/chimera). All HERV-K insertions included in this study were nominated with HERVnGoSeq or comprehensive literature review and validated via sequencing in previous studies or by confirming presence in the GRCh37 human reference.

Most studies of common genetic disease-risk variants published to-date rely on SNPs, which are easy and cheap to measure compared to other structural genetic variants. These SNP studies rely on linkage disequilibrium (LD) wherein the disease-associated SNP is not necessarily the causal variant but instead tags the causal variant (outlined in Fig S2). To test our underlying hypothesis that disease-associated SNPs are, in some cases, tagging polymorphic HERV-K insertions, which are the true causal variants, we next identified SNPs associated with each HERV-K insertion and queried existing SNP:disease databases for phenotypic associations.

HERVnGoSeq Computational Pipeline

We developed a custom bioinformatics pipeline, HERVnGoSeq, to map the genomic locations of HERV-K insertions using whole genome sequencing (WGS) data. Quality-filtered raw WGS were aligned to HERV-K113. Reads that partially aligned to HERV-K113 - chimeric reads - were trimmed and the non-HERV portions of the reads were extracted. The trimmed chimeric reads were then aligned to the human genome (GRCh37/Hg19). The base-pair position of the trimmed end of the read where HERV-K sequence was removed was called as the insertion point. Insertion points were collected for both the forward and reverse complement alignments separately. Insertion points within 1,000bp of each other were grouped to represent a single insertion point. The presence of putative HERV-K insertions were assigned to each individual if they had at least one chimeric read that aligned to that insertion point. Absence of an insertion was inferred for individuals when they lacked any chimeric reads representing the specific insertion. The complete pipeline and description is available at https://github.com/unreno/chimera.

HERV-K Identification/Validation

Putative polymorphic HERV-K elements were nominated via HERVnGoSeq. Sequence similarity between the reference index, HERV-K113 LTR, and the HERV-K10 LTR, which was ancestrally co-opted to form another mobile element class, Sine-VNTR-Alu composite elements [SVA, [27]], resulted in nomination of insertion sites of SVA-A, B, and C in addition to HERV-K when the LTR portion of the SVA element was sufficiently conserved. Thus, true HERV-K insertion sites nominated by HERVnGoSeq were identified as follows: *HERV-K present in reference* Using the dataset of mobile genetic elements present in the GRCh37 derived from RepBase [28](UCSC RepeatMasker track), HERVnGoSeq nominated sites were confirmed as HERV-K if mapped within a known HERV-K ± 100bp. *HERV-K absent from reference* The remaining polymorphic insertion sites were determined to be HERV-K only if the insertion ± 100bp was previously reported and confirmed with sequencing by a previous study[6,29-42]. Otherwise, they could not be distinguished from SVAs *in silico*. Additional HERV-K insertions missed by HERVnGoSeq but identified by previous studies and genotyped in the 1KG population were also included (n=5)[6,42]. Prevalence of each insertion site was estimated based on either the presence/absence calls from HERVnGoSeq or from genotypes of HERV-K insertions from previous studies.

Identification of SNPs associated with polymorphic HERV-K insertions

All HERV-K insertion sites were tested for SNP associations. For each of the 2,504 individuals in 1000 Genomes Phase 3, a binary indicator of presence/absence of the HERV-K insertion was generated via HERVnGoSeq or by recoding genotypes generated from previous publications [6,42]. Variant files for 1000 Genomes Phase 3 were downloaded from the FTP site (NCBI FTP site: ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp). After stratifying by continental population and removing related individuals[43], all biallelic SNPs present in the 1000 Genomes Phase 3 variant files were tested for association with the presence of each of the polymorphic HERV-K insertions using logistic regression adjusted for population stratification by including the first 6 multidimensional scaling (MDS) vectors in all models. MDS components were generated from all 1000 Genomes variants following pruning for common SNPs (minor allele frequency MAF>0.05) and for independence followed by random thinning to 10% of variants. All logistic regression modeling and MDS estimation were conducted in Plink 1.9 [44].

Manhattan plots were generated to visualize associations between genome-wide SNPs and polymorphic HERV-K insertions. For each of the "*taggable"* HERV-K insertion sites (i.e. those that showed a single, strong association peak in the Manhattan plot), HERV-K insertion-associated SNPs (hiSNPs) were defined as all SNPs within a 1Mb window of the insertion with p-value for association less than or equal to the Bonferroni adjusted p-value threshold for significance (0.05/Total SNPs in 1Mb window).

Sensitivity of hiSNP set generation

To confirm that the binary HERV-K insertion presence/absence calls made by HERVnGoSeq generated hiSNP sets similar to hiSNPs called using a different pipeline, we identified 12 HERV-K insertions detected by HERVnGoSeq and the 1000 Genomes by Sudmant *et al* [42]. For the 12 HERV-K insertion sites detected by HERVnGoSeq and genotyped by Sudmant *et al.*, logistic regression-based SNP associations were estimated from the binary HERVnGoSeq calls and the dichotomized 1000 Genomes genotype calls within Europeans. hiSNPs generated by HERVnGoSeq and 1000 Genomes calls using the method described above were compared. To ensure that hiSNPs associated by logistic regression were representative of SNPs that are in LD with polymorphic HERV-K insertions, complete genotype data for the 12 overlapping HERVnGoSeq/1000 Genomes sites were used to identify tagging SNPs via the r^2 measure of LD, defined as having an $r^2>0.2$.

Expression quantitative trait loci (eQTL)

Polymorphic HERV-K insertion hiSNPs across all HERV-K insertion sites were pooled and tested for eQTL enrichment against all common SNPs included in the tissue-specific Genotype-Tissue Expression (GTEx) Project (n= 11,555,102) [45] using a Fisher's exact test. Enrichment of hiSNPs annotated as GTEx eQTLs were also calculated separately by HERV-K insertion site and tissue type using Fisher's exact test.

Genome-wide association and Experimental Factor Ontology enrichment

To investigate whether the hiSNPs for the polymorphic HERV-K insertion sites have established phenotypic associations, the EMBL-EBI GWAS Catalogue [46] was queried for the presence of hiSNPs. To test for broader phenotypic enrichment across the HERV-K insertion sites, experimental factor ontology enrichment analyses were conducted for all pooled hiSNPs using the XGR online tool (http://galahad.well.ox.ac.uk:3020) with significant enrichments having a false discovery rate < 0.05.

SNP density

To determine whether the absence of hiSNPs for some HERV-K insertions sites was due to the absence of any proximal SNPs, SNP density was calculated for all insertion sites. All SNPs in the 1000 Genomes Phase 3 dataset were counted within a 1MB, 500Kb, and 100Kb window centred on each polymorphic HERV-K insertion site. The mean SNP densities for HERV-K insertions with hiSNPs was compared to HERV-K insertions with prevalence estimates from 0.2-0.8 and no hiSNPs using a Student's t-Test. Two HERV-K sites located in unlocalized contigs (chr1_gl00192_random) were excluded.

Hotspot distance

To examine whether some HERV-K insertions lacked hiSNPs due to their proximity to recombination hotspots, we selected two recombination hotspot maps and calculated the distance between the HERV-K insertion sites and recombination hotspots. We identified two datasets mapping the genomic locations of recombination hotspots genome-wide – one using a population-average LD-based mapping method [4,697 hotspots, [47]] and the other using ChIP-seq to identify PRDM9 binding sites among five individuals [62,110 hotspots, [48]]. Repeated random sampling with replacement was used to estimate the distribution of mean distances between randomly selected genomic locations and the nearest recombination hotspot. In order to compare this distribution to the distribution of distances of polymorphic HERV-Ks, a pool of random genomic locations was created wherein locations were matched to the HERV-K sites by chromosome and GC content of flanking 2Kb region. HERV-K insertion sites on chromosome Y and unmapped contigs were excluded (n=11). First, the percentage GC content of the 2kb flanking each HERV-K was calculated using data from the UCSC gc5Base Track [49] in Hg19. Next, all chromosomes were divided into 2Kb segments and percentage GC content was calculated for each. GC content for all segments was calculated to the tenth of a percent. For each HERV-K insertion, 2Kb genomic segments were added to the sampling pool if they had identical GC content and were on the same chromosome. This resulted in an average of 444 random locations from which to sample for each HERV-K site (~187,000 total). From this pool, 172 random genomic locations (defined as the center of the 2Kb fragment), matched 1:1 to respective HERV-K insertions, were sampled and the mean distance to nearest recombination hotspots were calculated. This sampling procedure was repeated 1000 times. The entire process was carried out independently with each of the recombination hotspot maps.

RESULTS

Polymorphic HERV-K identification

Our computational pipeline, HERVnGoSeq, nominated 1,381 putative HERV-K insertion sites among 2,504 human whole genomes from the 1000 Genomes Phase 3 population where sequencing reads partially aligned to the HERV-K113 LTR. HERV-

K113 represents the most recent HERV-K integration and thus is most likely to be polymorphic and to have preserved function [38]. Of the 1,381 sites, 403 HERV-K insertions mapped to reference HERV-K breakpoint sequences in GRCh37/Hg19. A total of 783 putative sites mapped to reference SINE-alu-VNTR (SVA) insertions and were discarded. Of the remaining 195 non-reference putative insertions, 23 had been previously annotated and confirmed as HERV-K via sequencing in recent studies [6,42] and the remainder, many of which are likely SVA, will require future targeted sequencing to confirm. With the rapidly increasing rate of discovery of HERV-K insertions in human genomes, we were able to collate an additional 5 HERV-K insertion sites discovered in a parallel study [6], which were also genotyped in the 1000 Genomes population. In total, the 431 (403 reference and 28 non-reference) HERV-K insertions were tested for SNP associations (S1 Table). The dichotomized presence/absence, rather than genotypes, of the nominated 431 HERV-K insertion sites were tested for SNP associations to reduce potential misclassification induced by poor sensitivity of calls due to low sequencing depth (~4x). An additional 16 HERV-K insertion sites identified in independent populations of diseased individuals [The Cancer Genome Atlas [39]; dbRIP [50]] were not detected by HERVnGoSeq, nor any parallel study utilizing the 1000 Genomes Project, suggesting that the diversity of HERV-K insertion sites expands beyond what is represented in the 1000 Genomes population.

hiSNP identification

All HERV-K insertions detected in more than one individual (n=431) were tested for SNP associations. HERVnGoSeq did not identify any HERV-K insertions occurring in 100% of 2,504 individuals; however, low coverage sequencing data likely resulted in an underestimation of the prevalence of some insertion sites. Thus it is likely that some sites with high prevalence called by HERVnGoSeq are actually fixed across populations. Genome-wide univariate logistic regression stratified by continental population and adjusted for genetic ancestry revealed 48 polymorphic HERV-K insertion sites with significant SNP associations (hiSNPs) in at least one continental population after correction for multiple testing (Table 1, S2 Table, S3-S50 Figs). Of these, 13 were not previously known to be polymorphic. The majority of the remaining 385 HERV-K insertion sites with no identifiable hiSNPs were rare (prevalence < 0.2, n=48), private (n=14), or common and potentially fixed (prevalence > 0.8, n=190), which may explain the lack of association with neighbouring SNPs. However, 129 HERV-K sites that appear to be common and unfixed (prevalence restricted to 0.2 - 0.8) yet have no hiSNPs and thus could not be evaluated for phenotype enrichment in this study (Fig 1). Among the 48 HERV-K insertion sites with strong SNP associations, hiSNP sets were selected within a 1-megabase window to ensure that no strongly associated SNPs were excluded arbitrarily due to genomic distance. The median number of hiSNPs associated with each of these 48 HERV-K insertions was 279 (Table 1).

To ensure that the selected hiSNPs represented SNPs in true linkage disequilibrium with the HERV-K insertions, a subset of twelve sites that were identified by HERVnGoSeq and also genotyped in an independent study of structural variation in the 1000 Genomes Project [42] were selected for sensitivity analyses (Supplementary Methods). hiSNP sets selected by HERVnGoSeq were compared to hiSNP sets derived from genotyped insertions via logistic regression on dichotomized genotypes, and LD via r^2 values based on maximum likelihood phasing. HERVnGoSeq logistic regression-based hiSNP sets consistently nominated the greatest number of hiSNPs across the twelve sites and, in ten out of twelve sites, >75% of HERVnGoSeq hiSNPs overlapped with hiSNPs derived from the genotyped insertions via logistic regression or r^2 (S2 Table, S51-63 Figs).

Tissue-specific differential gene expression and disease enrichment

The Genotype-Tissue expression (GTEx) project provides expression quantitative trait loci (eQTL) analysis results from genotype and gene expression data derived from 449 individuals across 44 human tissues [45]. We tested whether the HERVnGoSeq derived hiSNPs for each of the HERV-K insertion sites were enriched for eQTLs based on these data. Because the GTEx individuals are ~85% white, we restricted these analyses to the hiSNP sets identified among the 30 polymorphic HERV-K insertion sites in the European continental population (S1 Table) to reduce confounding by population stratification. We observed enrichment ($p<0.05$) of hiSNPs for eQTLs in at least one tissue type for 21 of the 30 sites by Fisher's exact test (Fig 2). HERV-K insertion sites contributed the most eQTL associations to subcutaneous adipose tissue and thyroid tissue with 16 HERV-K sites each. Thirteen of the sites with hiSNP sets enriched for eQTLs also include SNPs associated with disease by GWAS (Fig 2). The number of genes for which individual HERV-K insertion hiSNPs served as eQTLs ranged from 1 to 75 (S3 Table). Often in the instances where a large number of genes were affected, the HERV-K insertion and genes occurred on blocks of extended LD (i.e. the major histocompatibility complex).

We next examined the 30 hiSNP sets identified in the European continental population for annotation in the NHGRI-EBI GWAS Catalog. Half of the HERV-K insertion sites had at least one hiSNP with a genome-wide significant association with a disease phenotype (Fig 3). In total, European polymorphic HERV-K insertions are associated with 80 human phenotypes (Table 2).

Experimental factor ontology enrichment analysis suggests that polymorphic HERV-K insertions broadly associate with neurologic and immunologic disease phenotypes, including traits related to intracranial volume (FDR 4.40E-08), Parkinson's disease (FDR 1.80E-09), and autoimmune diseases (FDR 1.80E-09) (S4 Table).

Analyses of 'untaggable' HERV-K insertion sites

The majority of polymorphic HERV-K insertions identified via HERVnGoSeq were not associated with any nearby SNPs (n=129) and could not be evaluated for existing phenotypic associations in this study. Fifty-three of these HERV-K insertions (estimated mean prevalence: 45.6%, range: <1% - 79.7%), occur within genes (S5 Table).
The distribution of polymorphic HERV-K in specific chromosomal regions (ex: telomeres, centromeres) did not explain the lack of strong hiSNP associations in the 129 identified polymorphic sites (Fig 1A). However, we suspected that they might differ from sites with strong SNP associations in two respects – proximal SNP density and distance to nearest recombination hotspots, which both effect neighbouring patterns of LD [51,52]. We found significantly lower SNP densities in the areas around HERV-K insertion sites without hiSNPs (mean 2813.2) than the areas around HERV-K insertions with hiSNPs (mean 3392.7, difference in means: 579.5 SNPs, p-value 0.0007).

However, the presence of SNPs flanking the 129 HERV-K insertions without hiSNPs suggests that SNP density is not a sufficient determining factor.

ERVs can be involved in homologous and non-homologous recombination events [17]. Some are enriched for PRDM9 binding motifs [17] and elimination through recombination is a major mechanism by which ERV sequences are removed from the human genome [53]. We investigated whether recombination at HERV-K insertion sites explained the lack of hiSNP associations with these 129 HERV-K insertions by measuring their proximity to known recombination hotspots.

HERV-K insertion sites without hiSNP associations were farther on average from mapped recombination hotspots than their hiSNP-associated counterparts (222.2kb difference in distance to LD hotspots, *p*-value 0.04, and 40.1kb difference in distance to ChIP-seq hotspots, *p*-value 0.008). To determine whether the distance of polymorphic HERV-K insertions from recombination hotspots was greater than expected by chance, we compared the mean distance of these 172 polymorphic HERV-K insertions (48 with hiSNPs and 129 without) to the distances from hotspots of repeated random samples of 172 genomic locations. To mitigate potential confounding, random genomic locations were matched to HERV-K insertions sites on chromosome and flanking 2kb GC content. Polymorphic HERV-K without hiSNPs were farther from recombination hotspots than randomly selected genomic locations (p-value <0.005), whereas insertions with hiSNPs were not farther (or closer) to recombination hotspots than expected by chance (Fig 4).

DISCUSSION

This study shows that polymorphic HERV-K insertions occur in regions of the genome enriched for phenotypic function and, furthermore, that these insertion variants co-occur with established disease-risk variants, providing previously-untested candidates for the functional elements underlying the heritability of numerous complex diseases. Using our computational pipeline, HERVnGoSeq, and the diverse 1000 Genomes population, we confirmed the presence of 33 known polymorphic HERV-K insertions and identified an additional 13 confirmed sites via strong SNP associations not previously recognized as polymorphic. Of the total 48 HERV-K insertions under investigation, 18 have hiSNP sets enriched for eQTLs and 15 contained disease-associated SNPs identified in prior GWAS.

The collective evidence put forth by annotated hiSNPs supports a role for HERV-K insertions in inducing phenotypic effects. There is compelling evidence that polymorphic HERV-K insertions affect brain function. Previous studies have established links between HERV-K and amyotrophic lateral sclerosis [54], HIV-associated dementia [55], and Schizophrenia [14]. Our results further support these links and provide specific candidate polymorphisms that may explain these observations. We found two polymorphic HERV-K insertions (chr5:64388440 and chr6:32648026) whose hiSNP sets include a GWAS hit for schizophrenia. The association of the hiSNP at the insertion chr6:32648026 has already been attributed to the presence of a polymorphic HERV-K at the complement component 4 (C4) locus, resulting in altered expression at *C4A* and *C4B*, which we also observed in our eQTL enrichment results (Table S3)[14]. The hiSNP set at the second site at chr5:64388440 contains a SNP that is associated with schizophrenia symptoms relating to hallucination, delusion, and paranoia [56] and both the SNP and the HERV-K are located directly upstream of and serve as eQTLs for

ADAMTS6, a gene among a family that experimentally induces neurite growth in cultured neurons [57]. One of the most strongly associated hiSNPs for another HERV-K at chr4:120263688 was previously identified in a GWAS examining the genetics of cognitive performance using proxy-phenotypes [58]. Experimental factor ontology enrichment analysis of hiSNPs also suggested a largely neurological phenotypic effect of polymorphic HERV-K wherein Parkinson's disease, intracranial volume, and temporal arteritis were the second, third, and fourth most significantly enriched terms, respectively. Enrichment analysis also suggests that HERV-K insertion sites may have a functional role in autoimmune diseases. The role of polymorphic HERV-K in immunity, particularly insertions within the HLA, is difficult to delineate. While no specific associations have been established between HERV-K and autoimmunity, strong evidence suggests a link between multiple sclerosis and expression of HERV-W [59]; and the role of ERVs in autoimmunity has been long suspected, but its study has been hindered by technological limitations. With the increasing availability of next-generation sequencing data and computational methods like HERVnGoSeq, the time is ripe for a thorough investigation of polymorphic HERVs in autoimmune disease.

HERV-K expression has frequently been noted in human cancers and has also been of interest as an etiologic factor. We found hiSNP associations with Hepatocellular carcinoma (HCC) tagging two polymorphic HERV-K insertions. Recent studies identified an increase in HERV-K expression in HCC vs. normal tissue [60] and also discovered that HCC tumor mutations are frequently caused by APOBEC enzymes, a component of the human innate immune system primarily active against ERVs [61]. As such, the role of polymorphic HERV-K in interaction with hepatitis viruses and HCC appears warranted.

The HERV-K LTR is known to contain enhancer elements and thus the degree to which hiSNPs were enriched for eQTLs was not unexpected. An advantage of using the GTEx database is the ability to determine tissue-specific eQTL activity, which can help discern the phenotypic effects of polymorphic HERV-K. For example, we observe that the hiSNPs for a polymorphic HERV-K at chr19:22414379 are enriched for eQTLs in adipose tissue, suggesting that the insertion could affect fat storage. Indeed, the hiSNP set for this insertion also contains a GWAS SNP associated with changes in body mass index over time [62].

It is possible that the HERV-K with hiSNP eQTL enrichment is not itself the variant altering *cis*-gene expression, particularly in cases where the HERV is inserted on a large LD block, for example within the HLA region (chr6:32505702, chr6:3264803, chr6:32746812). Often these regions are not fully explored in functional follow-up studies due to their complexity. Since HERV-K LTRs contain functional elements, they serve as strong candidates for eQTLs regardless of regional complexity, warranting additional studies.

We could not leverage SNP annotations to illuminate the function of the majority of polymorphic HERV-K insertions because they did not associate with any neighbouring SNPs. In addition to the 48 HERV-K insertions with hiSNPs, we also nominated 129 reference HERV-K insertions that are likely polymorphic and that may have yet-undetected phenotypic associations. It is possible that some of the lower prevalence sites without hiSNPs are fixed and were called polymorphic only because of poor detection sensitivity of HERVnGoSeq. However, this seems unlikely, as our prevalence estimates for previously recognized non-reference polymorphic HERV-K were usually within ~10% of previous studies' estimates (Fig S65). Pairwise correlations of SNPs directly adjacent to these HERV-K insertions suggest that there is LD in these regions (data not shown), but that the HERV-K insertions are 'dark variants' that are not correlated with proximal SNPs. One potential explanation why the majority of polymorphic HERV-K insertions fail to have strong SNP associations is the greater than expected distance to the nearest recombination hotspot. Patterns of LD are known to strengthen the closer variants are to a hotspot, with complete loss of LD within the hotspot itself. The observed decay of HERV:SNP LD farther from hotspots requires further study. It is also possible that there is hotspot activity near or within these HERVs that were not identified and included in the two hot spot maps used for this study. Breakdown of LD patterns surrounding these HERV-K insertions may also be explained by other mechanisms that could not be investigated in the present study, including frequent sporadic non-allelic homologous recombination events, evolutionarily recent integration, off-target mutagenic activity of HERV-K repressors such as APOBEC enzymes, or hypermethylation resulting in sporadic deamination of methylated cytosines.

In our survey of phenotypic associations with polymorphic HERV-K insertions, the greatest limitation was the poor sensitivity of detection of HERV-K due to the low coverage of the sequencing data available for the 1000 Genomes population. Consequently, we were not able to call genotypes or estimate prevalence with high precision. Similar pipelines that have used these data to genotype mobile genetic elements often include an imputation step. Our observation, that a significant number of HERV-K insertions lack SNP associations, likely impedes the ability and reliability of imputation-based methods for genotyping these polymorphisms. This may also explain why so few members of the HERV-K family have been recognized as polymorphic. We anticipate that the accuracy of genotyping HERV-K insertions will greatly increase with higher coverage sequencing data.

Polymorphic HERV-K elements are associated with the germline risk of myriad phenotypes. While this study provides a starting point for further investigation, disease-specific epidemiologic and functional studies are needed to elucidate the role of specific polymorphic HERV-K insertions in complex diseases.

3.3 Tables 3.1-3.5: To ERV is human

Table 3.1: Polymorphic HERV-K insertions identified by HERVnGoSeq

Polymorphic Reference HERV-K Insertions			Polymorphic Non-Reference HERV-K Insertions		
Coordinate GRCh37/hg19	**Prevalence[$]**	**Average hiSNP Count[$]**	**Coordinate GRCh37/hg19**	**Prevalence[$]**	**Average hiSNP Count[$]**
chr3:14132679	0.96	453	chr1:106015875	0.04	436
chr3:125609298*	0.54	194	chr1:111802592	0.59	234
chr3:129776131*	0.47	203	chr1:223578304	0.01	206
chr3:195654395*	0.96	205	chr4:9603240	0.67	969
chr4:120263688*	0.68	1416	chr4:9981605	0.02	650
chr5:8937848	0.83	345	chr5:4537604	0.01	176
chr6:32505702*, chr6_cox_hap2:3953713, chr6_qbl_hap6:3740216	0.13	1817	chr5:64388440	0.07	226
chr6:32746812*, chr6_ssto_hap7:4177515	0.08	222	chr5:80442266	0.05	49
chr7:16237347*	0.81	259	chr6:32648036, chr6_mann_hap4:4099133, chr6_ssto_hap7:4073239	0.35	4965
chr7:158029477	0.28	239	chr6:161270899	0.84	576
chr8:7355392	0.14	108	chr7:158773385	0.01	102
chr8:18651453	0.52	199	chr11:60449890	0.07	292
chr8:37050885	0.32	125	chr12:44313657	0.27	593
chr10:135355522	0.18	155	chr12:124066477	0.13	444
chr11:71478951*	0.82	379	chr13:90743183	0.12	292
chr11:71875417	0.88	117	chr15:63374594	0.68	238
chr12:55727210	0.76	385	chr19:21841536	0.20	613
chr14:20552746*	0.30	144	chr19:22414379	0.43	993
chr17_ctg5_hap1:1389475*	0.17	2300	chr19:22457244	0.01	907
chr17_ctg5_hap1:504028*	0.16	2443	chr19:29855781	0.55	536
chr19:386675*	0.10	19	chr19:57996939	0.02	191
chr19:52924209*	0.39	37	chr20:12402387	0.03	271
chr20:25215439*	0.84	93	chrX:93606603	0.02	186
chr21:15654234*	0.65	19	chr8:7671216*	0.07	60

*Not previously recognized as polymorphic

[$] Averaged across 5 super populations

Table 3.2: HERV-K insertions with hiSNPs Annotated in NHGRI-EBI GWAS Catalog and associated traits

Non-MHC HERV-K insertion sites		
HERV-K Insertion Site	**Disease/Trait**	**SNP ID**
chr1:111802591	Interferon alpha levels in systemic lupus erythematosus	rs7411387***
chr4:120263688	Corneal astigmatism	rs11098499**
	Educational attainment	rs10028773***
chr5:64388440	Schizophrenia	rs17206232*
chr5:8937853	Obesity-related traits	rs11134338**
chr6:161270898	Lipoprotein (a) - cholesterol levels	rs1620921***
	Lipoprotein (a) levels	rs9355814, rs783147
	Protein quantitative trait loci	rs7770628
chr10:135355522	Obesity-related traits	rs2249694
chr12:124066477	Pubertal anthropometrics	rs786425*
chr12:55727213	Contrast sensitivity	rs12230513*
chr13:90743183	Longevity	rs2882281
chr15:63374594	Blood metabolite levels	rs1472631
	Mean platelet volume	rs11071720
	Metabolic traits	rs2652822
	Platelet count	rs3809566
	Social communication problems	rs17828380*
chr17:44361947	Bone mineral density	rs1864325
	Corticobasal degeneration	rs12185268
	Epithelial ovarian cancer	rs183211
	Idiopathic pulmonary fibrosis	rs17690703
	Interstitial lung disease	rs1981997
	Intracranial volume	rs9303525
	Male-pattern baldness	rs12373124
	Ovarian cancer in BRCA1 mutation carriers	rs183211
	Parkinson's disease	rs12185268, rs17577094, rs17649553, rs183211, rs199515, rs199533, rs415430, rs8070723
	Progressive supranuclear palsy	rs8070723
	Subcortical brain region volumes	rs17689882, rs8072451
chr19:22414379	Body mass index (change over time)	rs8105895**
	Chagas cardiomyopathy in Tripanosoma cruzi seropositivity	rs2262909
	Dental caries	rs10404998**, rs1865075**, rs931608*
	Response to statin therapy (LDL-C)	rs931608*
	Telomere length	rs1975174***, rs412658***
MHC HERV-K insertion sites		
HERV-K Insertion Site	**Disease/Trait**	**SNP ID**

chr6:32505702	Cervical cancer	rs9272143
	Hepatitis B vaccine response	rs3135363
	Hepatitis C induced liver cirrhosis	rs3135363
	Hepatocellular carcinoma	rs9272105
	Leishmaniasis (visceral)	rs9271858
	Response to interferon beta therapy	rs9272105
	Rheumatoid arthritis	rs2157337*
	Systemic sclerosis	rs3129763
chr6:32648036	Alzheimer's disease (late onset)	rs9271192
	Antinuclear antibody levels	rs2395185**
	Arthritis (juvenile idiopathic)	rs2395148
	Asthma	rs3117098, rs7775228***, rs9268516*, rs9272346*
	Asthma and hay fever	rs9273373*
	Atopic dermatitis	rs9469099
	Chronic lymphocytic leukemia	rs674313
	Circulating myeloperoxidase levels (serum)	rs3134931
	Cystic fibrosis severity	rs9268905**
	Dementia and core Alzheimer's disease neuropathologic changes	rs7453498*
	Epstein-Barr virus immune response (EBNA-1)	rs477515**
	Follicular lymphoma	rs12195582, rs2647012*
	Hepatitis B vaccine response	rs477515**
	Hepatitis C induced liver cirrhosis	rs3817963*
	Hepatocellular carcinoma (hepatitis B virus related)	rs9275319**
	Hodgkin's lymphoma	rs2395185**, rs6903608*
	Hypothyroidism	rs3129720
	IgA nephropathy	rs2856717*, rs660895*, rs7763262, rs9275596*
	IgE grass sensitization	rs7775228***
	Inflammatory bowel disease	rs477515**
	Leprosy	rs9271100
	Lung adenocarcinoma	rs3817963*
	Lung cancer	rs2395185**
	Lupus nephritis in systemic lupus erythematosus	rs2647012*
	Lymphoma	rs2647045***, rs2647046*, rs9268853**
	Multiple sclerosis (OCB status)	rs3129720, rs3817963*, rs9275563
	Narcolepsy (age of onset)	rs7744020*
	Nasopharyngeal carcinoma	rs28421666*
	Nephropathy	rs9275596*
	Neurofibrillary tangles	rs34075049**

	Parkinson's disease	rs2395163, rs9275326
	Peanut allergy	rs9275596*
	Primary biliary cirrhosis	rs7774434***
	Rheumatoid arthritis	rs12194148*, rs12525220, rs660895*, rs7748270, rs9268839*, rs9275406
	Sarcoidosis	rs2076530
	Schizophrenia	rs9274623
	Sjogren's syndrome	rs9271588
	Systemic lupus erythematosus	rs2647012*, rs9271100
	Systemic sclerosis	rs9275390
	Type 1 diabetes	rs9272346*
	Ulcerative colitis	rs1063355*, rs2395185**, rs6927022*, rs9268480*, rs9268853**, rs9268877**, rs9268923**
	Vitiligo	rs3806156**
	Waist-hip ratio	rs2076529
	Waist-to-hip ratio adjusted for body mass index	rs7759742*
chr6:32746812	Kawasaki disease	rs2857151

HERV-K::SNP association *p-value < 1.0e-10, **p-value < 1.0e-15, ***p-value < 1.0e-20

3.4 Figures 3.1-3.4: To ERV is human

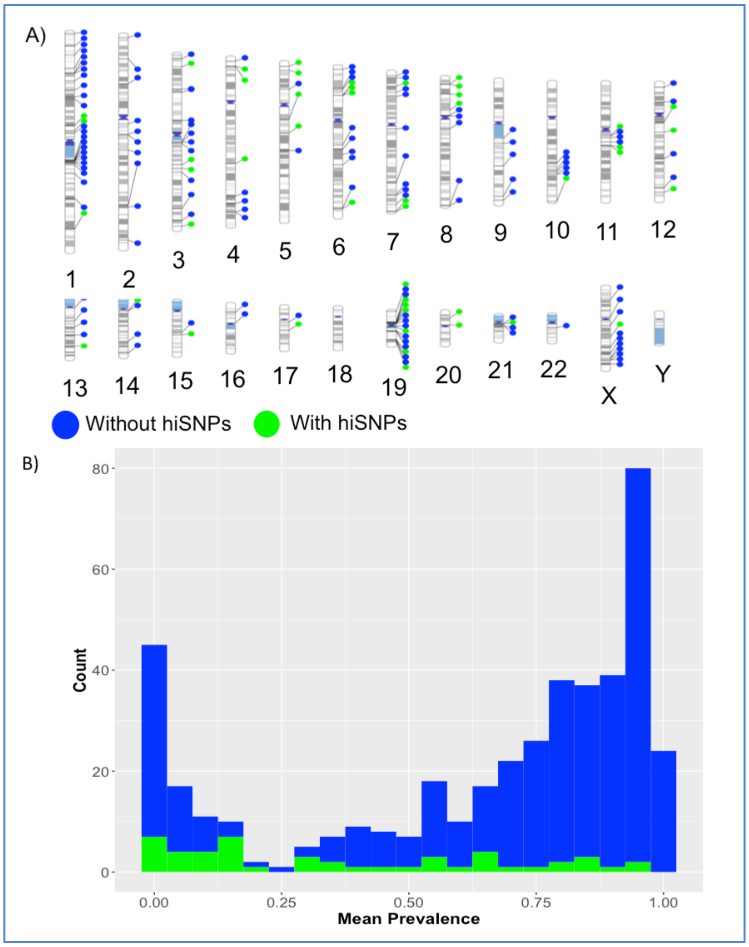


Fig 3.1. A) Ideogram. Relative genomic locations of HERV-K insertion locations with (green) and without (blue) hiSNPs identified via HERVnGoSeq. **B) Histogram**. Frequency distribution of identified HERV-K insertion prevalences for insertions with (green) and without (blue) hiSNPs

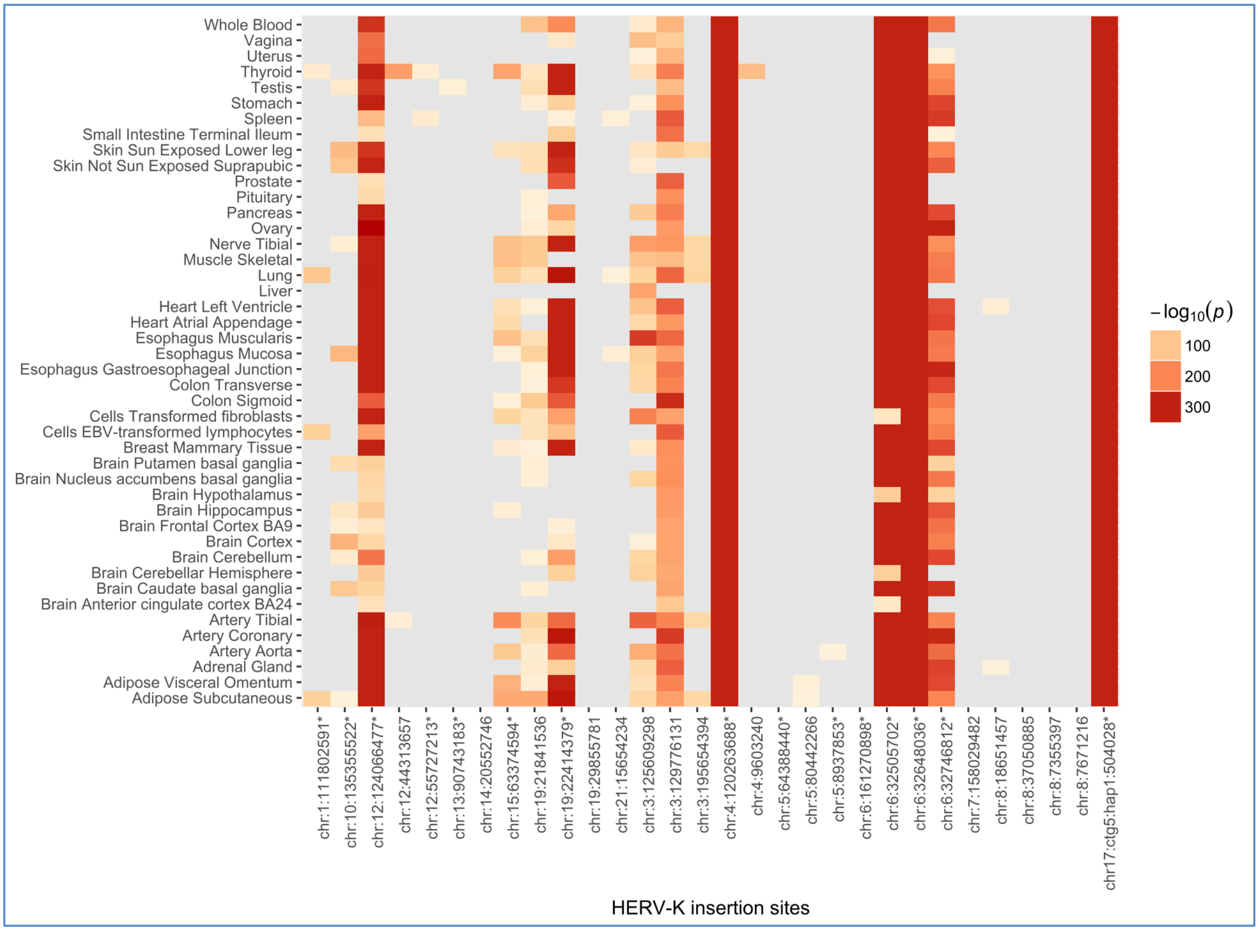


Fig 3.2. Heat map of log-transformed *p*-values for enrichment of hiSNPs that are eQTLs. Includes the 30 polymorphic HERV-K insertion sites with strong SNP associations in the European continental population. Results are stratified across 44 human tissue types.

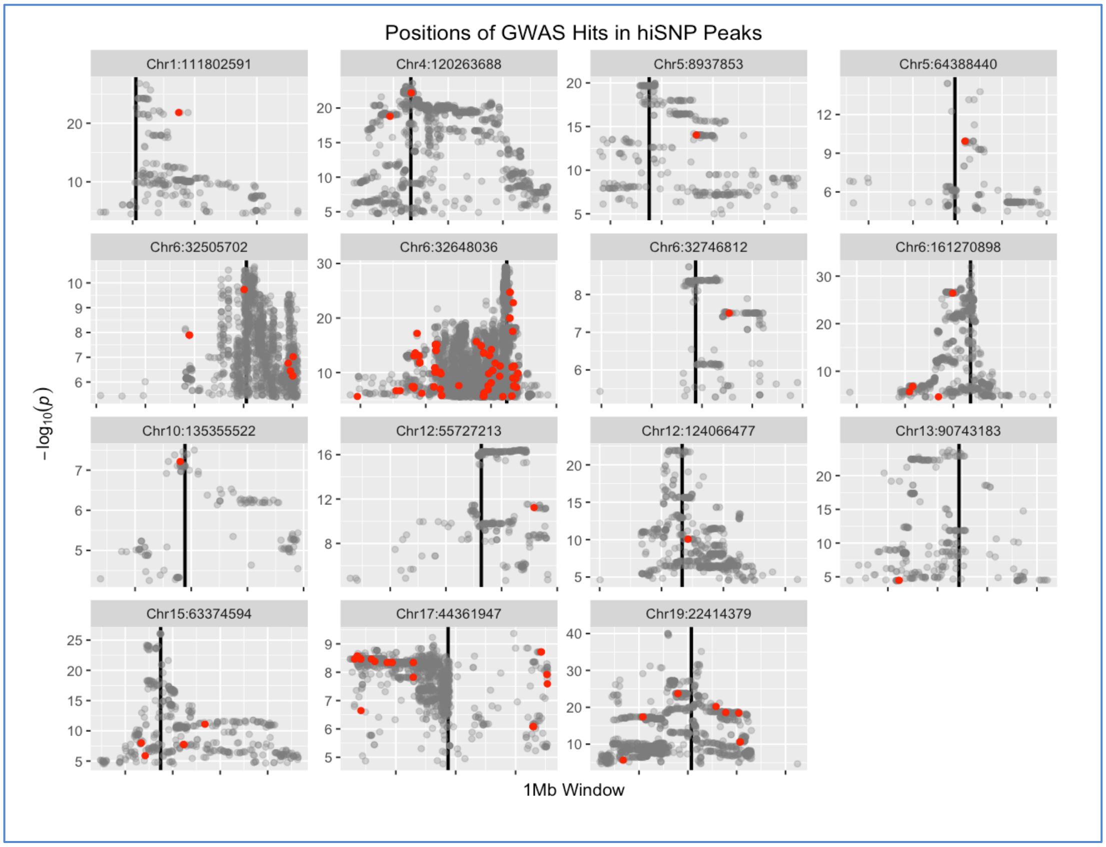


Fig 3.3. Manhattan plots. Plots of hiSNP sets for 15 HERV-K insertion sites among the 30 polymorphic HERV-K insertion sites with strong SNP associations in the European continental population. Vertical black lines denote HERV-K insertion locations. hiSNPs that are annotated in the NHGRI-EBI GWAS Catalog are represented by red points.

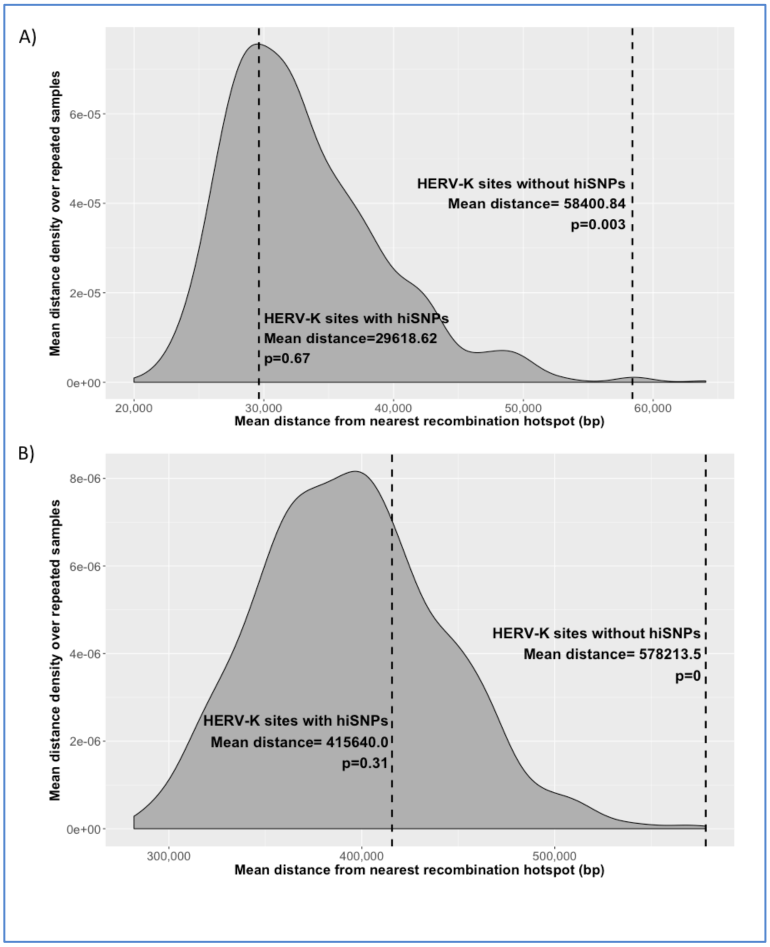


Fig 3.4. Mean distance to nearest recombination hotspot. Distances indicated for polymorphic HERV-K insertions with and without hiSNPs (dashed lines) and the distribution of mean distances of random genomic locations matched to HERV-K insertions on proximal GC content and chromosome. Distributions were derived from 1000 repeated random samples with replacement. **A)** Distances from nearest ChIP-seq-based recombination hotspot, **B)** Distances from nearest LD-based recombination hotspot.

3.5 Supplementary Materials: To ERV is human

3.5.1 Supplementary Tables 3.1-3.5

Table S3.1: Polymorphic HERV-K insertions identified with HERVnGoSeq, prevalence, and hiSNP set size, stratified by continental population

Polymorphic Reference HERV-K Insertions	**African**		**Amerindian**		**European**		**East Asian**		**South Asian**	
Coordinate GRCh37/hg19	**Prev**	**hiSNP Count**	**Prev**	**hiSNP Count**	**Prev**	**hiSNP Count**	**Prev**	**hiSNP Count**	**Prev**	**hiSNP Count**
chr3:14132679	0.87	453	0.97	-	0.97	-	0.98	-	1.00	-
chr3:125609298*	0.55	307	0.49	36	0.52	268	0.54	164	0.61	-
chr3:129776131*	0.62	15	0.39	186	0.36	273	0.50	283	0.49	258
chr3:195654395*	0.99	-	0.93	236	0.96	105	0.95	275	0.98	-
chr4:120263688*	0.66	1593	0.63	1253	0.67	1431	0.62	1465	0.79	1337
chr5:8937848	0.82	306	0.81	337	0.73	389	0.91	-	0.88	348
chr6:32505702*, chr6_cox_hap2:3953713, chr6_qbl_hap6:3740216	0.21	2407	0.10	1265	0.11	1862	0.12	1953	0.11	1600
chr6:32746812*, chr6_ssto_hap7:4177515	0.09	-	0.09	-	0.09	222	0.08	-	0.07	-
chr7:16237347*	0.84	130	0.87	-	0.83	-	0.65	387	0.84	-
chr7:158029477	0.19	251	0.29	193	0.33	446	0.33	146	0.28	159
chr8:7355392	0.32	44	0.11	125	0.08	204	0.04	-	0.12	57
chr8:18651453	0.65	203	0.42	116	0.26	238	0.76	263	0.51	175
chr8:37050885	0.48	182	0.22	110	0.18	150	0.33	110	0.40	75
chr10:135355522	0.06	88	0.23	107	0.22	132	0.17	292	0.22	-
chr11:71478951*	0.88	-	0.81	-	0.78	-	0.77	379	0.87	-
chr11:71875417	0.90	93	0.88	-	0.88	-	0.82	187	0.91	72
chr12:55727210	0.68	279	0.84	317	0.83	365	0.60	618	0.83	348
chr14:20552746*	0.49	72	0.24	131	0.18	205	0.29	218	0.30	96
chr17_ctg5_hap1:1389475*	0.11	-	0.19	-	0.23	2300	0.11	-	0.21	-
chr17_ctg5_hap1:504028*	0.11	-	0.17	-	0.25	2443	0.09	-	0.21	-
chr19:386675*	0.18	18	0.05	18	0.05	-	0.15	22	0.09	-
chr19:52924209*	0.50	33	0.36	-	0.32	-	0.34	-	0.42	41
chr20:25215439*	0.85	93	0.82	-	0.82	-	0.83	-	0.88	-
chr21:15654234*	0.70	-	0.63	-	0.62	19	0.60	-	0.69	-
Polymorphic Non-Reference HERV-K Insertions										
chr1:106015875	0.19	436	0.01	-	0.00	-	0.00	-	0.00	-
chr1:111802592	0.63	193	0.64	173	0.52	268	0.58	331	0.58	206
chr1:223578304	0.06	206	0.01	-	0.00	-	0.00	-	0.00	-
chr4:9603240	0.63	901	0.66	634	0.83	422	0.57	1306	0.66	1580
chr4:9981605	0.08	650	0.00	-	0.00	-	0.00	-	0.00	-
chr5:4537604	0.04	176	0.00	-	0.00	-	0.00	-	0.00	-
chr5:64388440	0.17	400	0.07	94	0.06	184	0.02	-	0.02	-
chr5:80442266	0.03	24	0.08	60	0.08	88	0.00	-	0.05	24

chr6:32648036, chr6_mann_hap4:4099133, chr6_ssto_hap7:4073239	0.26	5287	0.42	3282	0.27	6113	0.33	3938	0.45	6206
chr6:161270899	0.81	446	0.83	472	0.67	874	0.98	-	0.89	513
chr7:158773385	0.04	102	0.00	-	0.00	-	0.00	-	0.00	-
chr11:60449890	0.03	8	0.22	541	0.00	-	0.09	328	0.01	-
chr12:44313657	0.48	540	0.18	507	0.11	451	0.36	887	0.23	580
chr12:124066477	0.32	571	0.09	382	0.12	665	0.03	17	0.07	583
chr13:90743183	0.34	591	0.10	155	0.11	301	0.01	-	0.04	120
chr15:63374594	0.64	224	0.56	156	0.63	481	0.86	-	0.71	91
chr19:21841536	0.32	717	0.08	337	0.08	248	0.24	1342	0.29	423
chr19:22414379	0.64	626	0.35	1016	0.35	1306	0.35	1011	0.44	1007
chr19:22457244	0.04	907	0.00	-	0.00	-	0.00	-	0.00	-
chr19:29855781	0.62	604	0.48	578	0.51	547	0.72	575	0.43	377
chr19:57996939	0.07	191	0.01	-	0.00	-	0.00	-	0.00	-
chr20:12402387	0.12	271	0.01	-	0.00	-	0.00	-	0.00	-
chrX:93606603	0.06	186	0.02	-	0.00	-	0.00	-	0.00	-
Novel Sites Found in HNGS Not Previously Annotated										
chr8:7671216*	0.19	49	0.04	-	0.04	112	0.02	-	0.06	19

*Not previously recognized as polymorphic

Prev, Prevalence

Table S3.2: Number of hiSNPs identified via HERVnGoSeq and 1000 Genomes Structural Variant Group; European Super Population Only

Insertion Site Coordinate GRCh37/hg19	1000 Genomes Variant ID	Number HERVnGoSeq hiSNPs	Number 1KG SV hiSNPs	Number SNPS LD > 0.2
chr8_37050885_R_POST	DEL_pindel_25540	96	36	64
chr4_9603240_R_PRE	SVA_umary_SVA_205	321	370	0
chr5_64388440_F_POST	SVA_umary_SVA_245	33	31	42
chr5_80442266_R_PRE	SVA_umary_SVA_252	75	28	8
chr6_161270898_F_POST	SVA_umary_SVA_320	814	736	258
chr1_111802591_R_PRE	SVA_umary_SVA_45	243	240	34
chr12_44313657_F_POST	SVA_umary_SVA_540	377	289	121
chr12_124066477_R_PRE	SVA_umary_SVA_556	394	399	103
chr13_90743183_R_PRE	SVA_umary_SVA_583	221	236	145
chr15_63374594_R_PRE	SVA_umary_SVA_637	431	343	40
chr19_21841536_R_PRE	SVA_umary_SVA_745	193	176	108
chr19_29855781_R_PRE	SVA_umary_SVA_748	468	457	295

Table S3.3: List of HERV-K insertion sites whose hiSNP sets contain eQTL and corresponding list of differentially expressed genes across all tissue types

HERV-K Location	eQTL Genes
chr1:111802591	CHI3L2, CHIA, CHIAP2, OVGP1, RP11-552M11.8 DDX20, DRAM2, ATP5F1, CD53, UBE2FP3, HIGD1AP12, PGCP1, PIFO, RP11-284N8.3
chr10:135355522	SYCE1 CYP2E1, PAOX
chr12:124066477	ATP6V0A2, DDX55, DNAH10, DNAH10OS, EIF2B1, RP11-486O12.2, TCTN2, RP11-282O18.7, TMED2, C12orf65, MPHOSPH9, HCAR2, GTF2H3, RP11-282O18.3, SNRNP35, PITPNM2, ZNF664, RILPL2, VPS37B, RILPL1, ABCB9, RPL27P12, SETD8
chr12:44313657	PUS7L RP11-350F4.2, TWF1, TMEM117, RP11-210N13.1
chr12:55727213	DNAJC14, TESPA1, RAB5B
chr13:90743183	LINC00559
chr14:20552746	PNP
chr15:63374594	APH1B, LACTB, RP11-244F12.3, RPS27L, TPM1 RP11-1069G10.1, CA12, RP11-321G12.1, USP3
chr17:44361947	ARL17A, ARL17B, CRHR1, CRHR1-IT1, DND1P1, KANSL1, KANSL1-AS1, LRRC37A, LRRC37A2, LRRC37A4P, NSFP1, PLEKHM1, RP11-259G18.2, RP11-259G18.3, RP11-707O23.5, RPS26P8, WNT3, MAPT, RP11-259G18.1, RP11-669E14.6, RP11-798G7.8, ARHGAP27, CTD-2020K17.1, FMNL1, LRRC37A17P, RP11-798G7.5, SPPL2C, AC091132.1, FAM215B, MAP3K14, MAPT-AS1, RP11-798G7.6, NSF, NMT1, CTB-39G8.3, RP11-669E14.4, RP11-293E1.1, RPRML, RP11-995C19.2
chr19:21841536	RP11-420K14.3, RP11-420K14.6, RP11-678G14.2, RP11-678G14.3, VN1R84P, ZNF100, ZNF429, ZNF626, ZNF708 ZNF43, CTD-2561J22.5, ZNF66, ZNF85, RP11-420K14.1, RP11-157B13.7, RP11-420K14.2
chr19:22414379	RP11-678G14.3, ZNF208, ZNF257, ZNF429, ZNF98, RP11-678G14.2, ZNF738 CTB-176F20.3, CTD-2561J22.5, RP11-420K14.1, RP11-420K14.2, AC025811.3, CTB-135N1.2, ZNF100, LINC00664, ZNF676, AC003973.6, CTD-2561J22.2, CTC-457E21.3, RP11-420K14.6, ZNF431, AC011516.1, CTD-2291D10.2, RP11-157B13.6, RP11-157B13.7, ZNF209P, ZNF729, AC004004.2, CTD-2561J22.6, VN1R85P, ZNF492, RP11-420K14.3, ZNF708
chr19:29855781	CTC-525D6.5
chr21:15654234	ABCC13, RBM11
chr3:125609298	FAM86JP, RP11-379B18.5, SLC41A3, ALDH1L1-AS1, RP11-379B18.1, RP11-666A20.4 ENPP7P4, ALG1L, RP11-666A20.3, ROPN1B, GS1-388B5.1, OR7E29P, OR7E53P, RPS3AP14
chr3:129776131	COL6A4P2, FAM86HP, RP11-77P16.4, ALG1L2, COL6A6 RP11-93K22.13, TRH, TMCC1-AS1
chr3:195654394	AC024937.6, RP11-480A16.1 SDHAP2, SDHAP1
chr4:120263688	C4orf3, RP11-21I10.2, RP11-33B1.1, RP11-384K6.6, RP11-548H18.2, FABP2, RP11-33B1.4, METTL14, KLHL2P1, RP11-33B1.3, RP11-455G16.1, USP53, MYOZ2, PDE5A, RP11-33B1.2
chr4:9603240	ENPP7P10, RP11-1396O13.1, ENPP7P11, FAM86KP, OR7E83P, OR7E84P, OR7E85P, OR7E86P, RP11-1396O13.13, RP11-1396O13.22, RP11-747H12.6, UNC93B7
chr5:64388440	ADAMTS6 PPWD1
chr5:80442266	RASGRF2 CKMT2-AS1
chr5:8937853	SEMA5A
chr6:161270898	SLC22A3, RP3-428L16.1, RP11-235G24.3

chr6:32505702	C4A, C4B, CYP21A1P, CYP21A2, HCG23, HLA-DPB2, HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DQB1-AS1, HLA-DQB2, HLA-DRA, HLA-DRB1, HLA-DRB5, HLA-DRB6, MICB, NOTCH4, RPL32P1, STK19P, XXbac-BPG254F23.6, PSMB9, RNF5, STK19, TNXA, TNXB, ATF6B, CFB, HLA-DOB, PRRT1, SKIV2L, HLA-DMA, PPT2, EGFL8, HLA-DPB1, AGER, DDAH2, HLA-DMB, C6orf25, CSNK2B, GPANK1, LY6G6C, TAP2, LY6G5B, SLC44A4, VWA7, APOM, AGPAT1, DXO, HLA-DPA1, BTNL2, WASF5P, HLA-DRB9, MSH5, TAP1, XXbac-BPG254F23.7, PBX2, C2
chr6:32648036	ATP6V1G2, BAG6, C2, C4A, C4B, CFB, CYP21A1P, CYP21A2, HCG23, HLA-DOB, HLA-DPB1, HLA-DPB2, HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DQB1-AS1, HLA-DQB2, HLA-DRA, HLA-DRB1, HLA-DRB5, HLA-DRB6, LSM2, LY6G5B, NOTCH4, PRRT1, RPL32P1, SKIV2L, STK19P, TAP2, XXbac-BPG254F23.6, TNXA, AGER, RNF5, STK19, TNXB, MICB, ATF6B, HLA-DPA1, PPT2, EGFL8, HLA-DMA, HSPA1A, HSPA1B, LY6G6C, XXbac-BPG154L12.4, HLA-DMB, PBX2, TAPBP, MICA, RPS18, C6orf25, CSNK2B, GPANK1, LY6G6E, SLC44A4, VWA7, AGPAT1, DXO, LY6G5C, MSH5, APOM, FKBPL, BTNL2, PSMB9, LYPLA2P1, XXbac-BPG181B23.7, HLA-DRB9, LY6G6D, C6orf47, NELFE, C6orf10, PSMB8, TAP1, XXbac-BPG254F23.7, SYNGAP1
chr6:32746812	HLA-DOB, HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DQB1-AS1, HLA-DQB2, HLA-DRB1, HLA-DRB5, HLA-DRB6, NOTCH4, PRRT1, TAP2, XXbac-BPG254F23.6, C4A, TAPBP, RNF5, HSD17B8, ATF6B, CYP21A2, TAP1, XXbac-BPG254F23.7, NONE
chr7:158029482	NONE
chr8:18651457	RPL35P6, ASAH1
chr8:37050885	NONE
chr8:7355397	NONE
chr8:7671216	NONE

Table S3.4: Experimental Factor Ontology enrichment results for all hiSNPs in the European continental population

Term ID	Term Name	# of SNPs annotated	Z-score	P-value	FDR	# of SNPs overlapped	SNPs
EFO:0005140	autoimmune disease	1733	8.16	2.80E-11	1.80E-09	30	rs9268923, rs2395185, rs477515, rs3129720, rs9275563, rs3817963, rs1816854, rs9271100, rs1063355, rs9268877, rs9268853, rs6927022, rs9271060, rs9271588, rs9268480, rs9272346, rs2647012, rs3806156, rs7411387, rs7748270, rs12194148, rs9268839, rs2070676, rs2070677, rs12525220, rs660895, rs9275406, rs2395148, rs2157337, rs111945767
EFO:0002508	Parkinson's disease	136	12	3.30E-11	1.80E-09	10	rs8070723, rs12185268, rs183211, rs9275326, rs415430, rs2395163, rs199515, rs199533, rs17577094, rs17649553
EFO:0004886	intra cranial volume	5	19.8	2.00E-09	4.40E-08	3	rs9303525, rs8072451, rs17689882
EFO:1001209	temporal arteritis	5	19.8	2.00E-09	4.40E-08	3	rs9268905, rs4252134, rs9275592
EFO:0006803	vasculitis	5	19.8	2.00E-09	4.40E-08	3	rs9268905, rs4252134, rs9275592
EFO:0000574	lymphoma	42	11	3.60E-08	6.50E-07	5	rs2395185, rs9268853, rs2647045, rs2647046, rs6903608
EFO:0005206	oligoclonal band measurement	9	14.7	5.10E-08	7.90E-07	3	rs3129720, rs9275563, rs3817963
EFO:0004194	IGA glomerulonephritis	32	10.1	3.30E-07	0.0000045	4	rs9275596, rs660895, rs2856717, rs7763262
EFO:0005129	hepatitis C induced liver cirrhosis	5	13.1	9.10E-07	0.000011	2	rs3817963, rs3135363
EFO:0004614	apolipoprotein A 1 measurement	7	11.1	0.0000032	0.000031	2	rs783147, rs7770628
EFO:0005856	arthritis	409	6.02	0.0000028	0.000031	10	rs7748270, rs12194148, rs9268839, rs2070676, rs2070677, rs12525220, rs660895, rs9275406, rs2395148, rs2157337
EFO:0001642	lymphoid neoplasm	266	6.22	0.0000037	0.000034	8	rs2395185, rs9268853, rs2647045, rs2647046, rs2647012, rs674313, rs6903608, rs12195582
EFO:0002461	skeletal system disease	918	5.4	0.0000043	0.000036	15	rs9271100, rs9275390, rs3129763, rs2647012, rs7411387, rs7748270, rs12194148, rs9268839, rs2070676, rs2070677, rs12525220, rs660895, rs9275406, rs2395148, rs2157337
EFO:0007790	Epstein Barr virus nuclear antigen 1 IgG measurement	8	10.3	0.0000051	0.000037	2	rs9268923, rs2516049
EFO:0000729	ulcerative colitis	274	6.09	0.0000048	0.000037	8	rs9268923, rs2395185, rs9271100, rs1063355, rs9268877, rs9268853, rs6927022, rs9268480
EFO:0000405	digestive system disease	1291	5.14	0.0000059	0.00004	18	rs9268923, rs2395185, rs477515, rs3817963, rs7774434, rs3828805, rs1816854, rs9271100, rs3135363, rs1063355, rs9268877, rs9268853, rs6927022, rs9271060, rs9271588, rs9275319, rs9272105, rs9268480
EFO:0005755	rheumatic disease	738	5.35	0.0000007	0.000045	13	rs9271100, rs2647012, rs7411387, rs7748270, rs12194148, rs9268839, rs2070676, rs2070677, rs12525220, rs660895, rs9275406, rs2395148, rs2157337
EFO:0005803	hematological system disease	297	5.76	0.0000091	0.000054	8	rs2395185, rs9268853, rs2647045, rs2647046, rs2647012, rs674313,

							rs6903608, rs12195582
Orphanet:101435	Rare genetic eye disease	62	7.02	0.0000095	0.000054	4	rs9268838, rs8070723, rs9275390, rs3129763
EFO:0000685	rheumatoid arthritis	317	5.5	0.000015	0.000083	8	rs7748270, rs12194148, rs9268839, rs12525220, rs660895, rs9275406, rs2395148, rs2157337
EFO:0000508	genetic disorder	328	5.37	0.00002	0.000099	8	rs9268905, rs9268838, rs8070723, rs12185268, rs12373124, rs9275390, rs9268947, rs3129763
EFO:0000182	hepatocellular carcinoma	12	8.34	0.00002	0.000099	2	rs9275319, rs9272105
EFO:0001379	endocrine system disease	333	5.31	0.000023	0.00011	8	rs3129720, rs3817963, rs9268905, rs7774434, rs3135363, rs9275319, rs9272105, rs9268947
EFO:0000783	myositis	15	7.41	0.00004	0.00018	2	rs9275338, rs9275330
Orphanet:586	Cystic fibrosis	17	6.93	0.00006	0.00019	2	rs9268905, rs9268947
Orphanet:306708	Frontotemporal neurodegeneration with movement disorder	16	7.16	0.000049	0.00019	2	rs8070723, rs12185268
Orphanet:165661	Genetic pancreatic disease	17	6.93	0.00006	0.00019	2	rs9268905, rs9268947
EFO:0000183	Hodgkins lymphoma	16	7.16	0.000049	0.00019	2	rs2395185, rs6903608
EFO:0003767	inflammatory bowel disease	658	4.69	0.000048	0.00019	11	rs9268923, rs2395185, rs477515, rs1816854, rs9271100, rs1063355, rs9268877, rs9268853, rs6927022, rs9271060, rs9268480
EFO:0001421	liver disease	143	5.42	0.00005	0.00019	5	rs3817963, rs7774434, rs3135363, rs9275319, rs9272105
Orphanet:399998	Male infertility due to obstructive azoospermia of genetic origin	17	6.93	0.00006	0.00019	2	rs9268905, rs9268947
Orphanet:307058	Miscellaneous movement disorder due to genetic neurodegenerative disease	16	7.16	0.000049	0.00019	2	rs8070723, rs12185268
Orphanet:400003	Rare genetic disorder with obstructive azoospermia	17	6.93	0.00006	0.00019	2	rs9268905, rs9268947
Orphanet:156610	Rare genetic respiratory disease	17	6.93	0.00006	0.00019	2	rs9268905, rs9268947
Orphanet:183521	Rare genetic movement disorder	19	6.52	0.000084	0.00026	2	rs8070723, rs12185268
EFO:0004280	movement disorder	20	6.34	0.000099	0.0003	2	rs8070723, rs12185268
EFO:0006925	lipoprotein A measurement	29	5.16	0.00031	0.00091	2	rs1620921, rs9355814
EFO:1001986	connective tissue disease	31	4.96	0.00038	0.00095	2	rs9275390, rs3129763
Orphanet:98702	Connective tissue disease with eye involvement	31	4.96	0.00038	0.00095	2	rs9275390, rs3129763
Orphanet:165652	Rare genetic gastroenterological disease	30	5.06	0.00034	0.00095	2	rs9268905, rs9268947
EFO:1001993	scleroderma	31	4.96	0.00038	0.00095	2	rs9275390, rs3129763
EFO:0000717	systemic scleroderma	31	4.96	0.00038	0.00095	2	rs9275390, rs3129763
EFO:0004505	telomere length	30	5.06	0.00034	0.00095	2	rs412658, rs1975174
EFO:0003956	seasonal allergic	37	4.48	0.00064	0.0016	2	rs7775228, rs9273373

	rhinitis						
EFO:0000616	neoplasm	1719	3.4	0.00078	0.0019	17	rs2395185, rs3817963, rs7774434, rs3828805, rs9268853, rs9275319, rs9272105, rs9272143, rs183211, rs28421666, rs2647045, rs2647046, rs2647012, rs674313, rs6903608, rs9275642, rs12195582
EFO:0004554	genomic measurement	41	4.21	0.00086	0.002	2	rs412658, rs1975174
Orphanet:275742	Genetic infertility	46	3.93	0.0012	0.0027	2	rs9268905, rs9268947
Orphanet:399980	Rare genetic male infertility	46	3.93	0.0012	0.0027	2	rs9268905, rs9268947
EFO:1001513	liver neoplasm	105	3.66	0.0014	0.003	3	rs7774434, rs9275319, rs9272105
EFO:0006930	brain volume measurement	107	3.62	0.0015	0.0032	3	rs9303525, rs8072451, rs17689882
EFO:0005772	neurodegenerative disease	967	3.21	0.0015	0.0033	11	rs8070723, rs12185268, rs183211, rs9275326, rs415430, rs2395163, rs199515, rs199533, rs9271192, rs17577094, rs17649553
EFO:0000311	cancer	1541	3.11	0.0017	0.0035	15	rs2395185, rs3817963, rs3828805, rs9268853, rs9275319, rs9272105, rs9272143, rs183211, rs2647045, rs2647046, rs2647012, rs674313, rs6903608, rs9275642, rs12195582
Orphanet:71859	Rare genetic neurological disorder	53	3.59	0.0018	0.0038	2	rs8070723, rs12185268
EFO:0003819	dental caries	122	3.3	0.0024	0.0048	3	rs931608, rs1865075, rs10404998
EFO:0003086	kidney disease	301	3.13	0.0026	0.0051	5	rs9275596, rs2647012, rs660895, rs2856717, rs7763262
EFO:0003966	eye disease	303	3.11	0.0027	0.0052	5	rs9271588, rs9268838, rs8070723, rs9275390, rs3129763
EFO:0004343	waist-hip ratio	129	3.16	0.0029	0.0055	3	rs7759742, rs5020946, rs2076529
EFO:0003785	allergy	217	3.05	0.0032	0.006	4	rs7775228, rs9273373, rs9469099, rs4713555
EFO:0005854	allergic rhinitis	66	3.11	0.0034	0.0063	2	rs7775228, rs9273373
EFO:0004274	gout	69	3.02	0.0039	0.007	2	rs2070676, rs2070677
EFO:0004732	lipoprotein measurement	681	2.83	0.004	0.0071	8	rs931608, rs783147, rs7770628, rs2858310, rs9378212, rs1620921, rs9275052, rs9355814
EFO:0000096	neoplasm of mature B-cells	146	2.88	0.0045	0.0079	3	rs2647012, rs674313, rs12195582
EFO:0003853	respiratory system neoplasm	154	2.76	0.0054	0.0094	3	rs2395185, rs3817963, rs3828805
EFO:0004872	inflammatory biomarker measurement	262	2.59	0.007	0.012	4	rs783147, rs7770628, rs2858310, rs9378212
EFO:0000524	head disease	385	2.48	0.0085	0.014	5	rs931608, rs3828805, rs9271588, rs1865075, rs10404998
EFO:0007861	body ratio measurement	178	2.45	0.009	0.015	3	rs7759742, rs5020946, rs2076529
EFO:0003769	endocrine neoplasm	187	2.34	0.011	0.017	3	rs7774434, rs9275319, rs9272105
EFO:0000684	respiratory system disease	1556	2.3	0.011	0.018	13	rs2395185, rs3817963, rs9268905, rs3828805, rs9268947, rs9272346, rs2076530, rs7775228, rs17690703, rs3117098, rs9268516, rs1981997, rs9273373
EFO:0000274	atopic eczema	110	2.13	0.014	0.022	2	rs9469099, rs4713555
EFO:0000701	skin disease	561	2.19	0.014	0.022	6	rs12373124, rs9275390, rs3129763, rs9469099, rs3806156, rs4713555
EFO:0000270	asthma	449	2.09	0.017	0.025	5	rs9272346, rs7775228, rs3117098, rs9268516, rs9273373

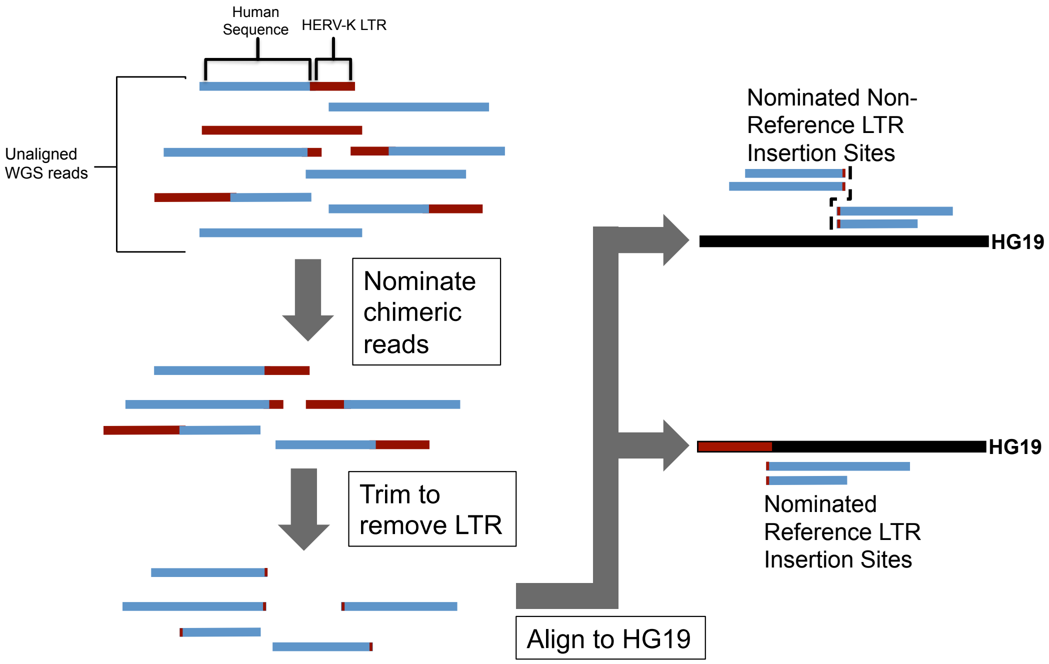
EFO:0006846	autoimmune disease biomarker	117	2.02	0.016	0.025	2	rs11071720, rs2395185
EFO:0005036	platelet measurement	117	2.02	0.016	0.025	2	rs11071720, rs3809566
EFO:0004264	vascular disease	215	2.06	0.017	0.025	3	rs9268905, rs4252134, rs9275592
EFO:0003885	multiple sclerosis	221	2	0.019	0.027	3	rs3129720, rs9275563, rs3817963
EFO:0004464	brain measurement	347	1.94	0.021	0.03	4	rs9303525, rs34075049, rs8072451, rs17689882
EFO:0001071	lung carcinoma	129	1.85	0.021	0.03	2	rs2395185, rs3817963
EFO:0004458	C-reactive protein measurement	139	1.73	0.026	0.036	2	rs2858310, rs9378212
EFO:0004645	response to vaccine	154	1.56	0.033	0.046	2	rs477515, rs3135363
EFO:0005105	lipid or lipoprotein measurement	1016	1.6	0.043	0.058	8	rs931608, rs783147, rs7770628, rs2858310, rs9378212, rs1620921, rs9275052, rs9355814

Table S3.5: List of HERV-K insertions without hiSNPs that occur within genes

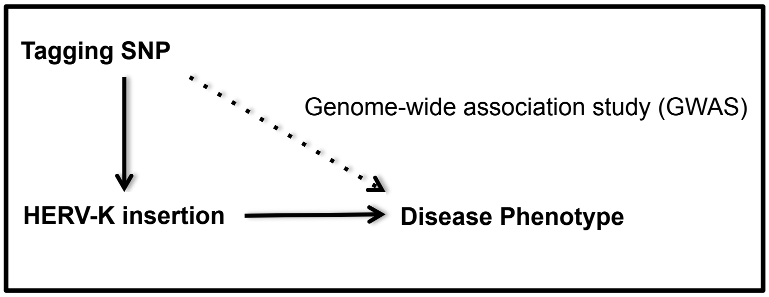
HERV-K	Chromosome	Gene Start	Gene End	Gene Name
chr1_144603408	chr1	144146810	144830407	NBPF9
chr1_144603408	chr1	144146810	144830407	NBPF8
chr1_144603408	chr1	144596310	144604297	BC073801
chr1_145502680, chr1_146069249, chr1_146119262, chr1_146343450, chr1_146383502	chr1	145293370	146467744	NBPF10
chr1_146069249	chr1	146032541	146082431	NBPF24
chr1_146343450, chr1_146383502	chr1	146334189	146460493	NBPF12
chr1_148565785	chr1	148558187	148596267	NBPF15
chr1_15462793	chr1	15438310	15478960	TMEM51-AS1
chr1_160660573	chr1	160648535	160681641	CD48
chr1_246245987	chr1	245912641	246580714	SMYD3
chr1_45979181	chr1	45976706	45987610	PRDX1
chr1_75842769	chr1	75667815	76076799	SLC44A5
chr2_112721387	chr2	112656190	112786945	MERTK
chr2_3109196	chr2	2898819	3129798	AK095310
chr2_98747295	chr2	98703594	98833427	VWA3B
chr21_19306004	chr21	19273579	19639687	CHODL
chr3_100056136	chr3	100053561	100068855	NIT2
chr3_100384984	chr3	100328432	100414323	GPR128
chr3_130166558	chr3	130064358	130203690	COL6A5
chr3_134235732	chr3	134204574	134283870	CEP63
chr3_188969159	chr3	188665002	189041271	TPRG1
chr3_9889344	chr3	9834231	9896822	ARPC4-TTLL3
chr3_99991338	chr3	99979660	100044096	TBC1D23
chr5_180694828	chr5	180688212	180699308	LOC100507602
chr5_55452824	chr5	55395506	55529186	ANKRD55
chr6_160215763	chr6	160211021	160219461	MRPL18
chr6_74514859	chr6	74405507	74538041	CD109
chr7_139152150	chr7	139138087	139168457	KLRG2
chr7_140252784	chr7	140218219	140302342	DENND2A
chr7_27781273	chr7	27778991	27869386	TAX1BP1
chr8_113784400	chr8	113235158	114389382	CSMD3
chr8_91696260	chr8	91634222	91803859	TMEM64
chr10_101580568	chr10	101542462	101611662	ABCC2
chr10_104215093	chr10	104209573	104216050	LOC100505761
chr10_104215093	chr10	104213596	104216049	AX746750
chr10_19799694	chr10	19778022	19896829	C10orf112
chr11_60481953	chr11	60467046	60483285	MS4A8
chr11_62625990	chr11	62623483	62656355	SLC3A2
chr12_21797876	chr12	21788274	21810728	LDHB

chr12_4831367	chr12	4829751	4881892	GALNT8
chr13_43615646	chr13	43597361	43683306	DNAJC15
chr15_59126496	chr15	59063392	59149734	FAM63B
chr16_14729998	chr16	14726667	14763093	BFAR
chr18_66610471	chr18	66382490	66722426	CCDC102B
chr19_12214752	chr19	12203077	12225494	ZNF788
chr19_36725235	chr19	36705503	36729675	ZNF146
chr19_38021670	chr19	37997840	38034239	ZNF793
chr19_39122732	chr19	39109721	39126125	EIF3K
chr19_55462618	chr19	55434876	55477611	NLRP7
chrX_122813497	chrX	122734411	122866904	THOC2
chrX_134437628	chrX	134382887	134477957	DKFZp451F083
chrX_134437628	chrX	134382887	134478012	BC029787
chrX_134540021	chrX	134540020	134540794	AB062081
chrX_151298090	chrX	151282520	151307050	MAGEA10-MAGEA5
chrX_154138254	chrX	154064063	154250998	F8

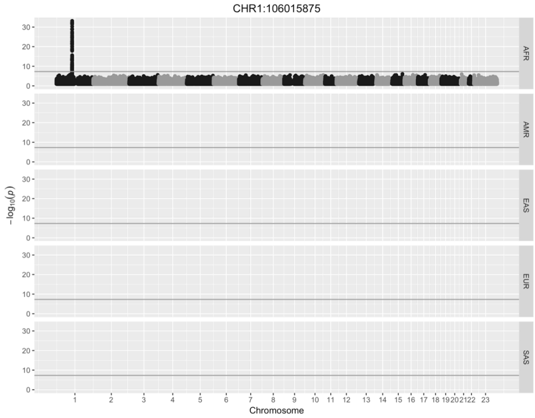
3.5.2 Supplementary Figures 3.1-3.65

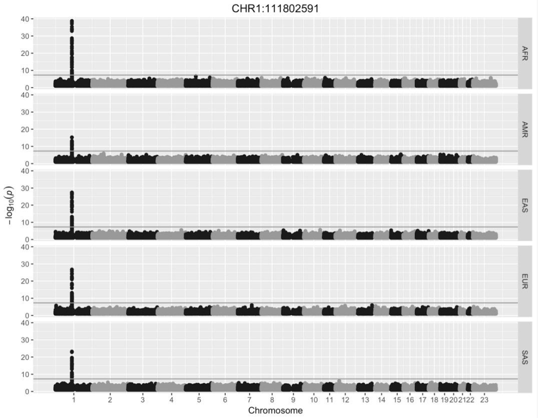


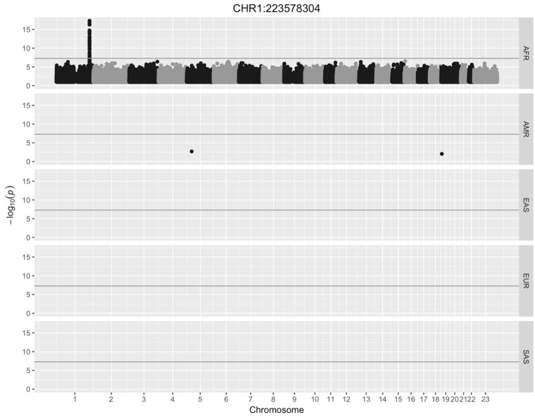
S3.1 Fig. Schematic representation of the HERVnGoSeq computational pipeline.

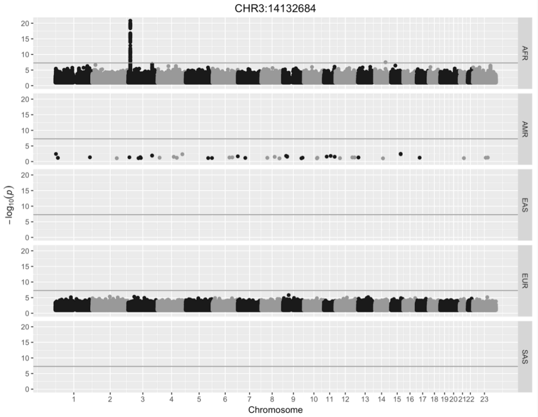


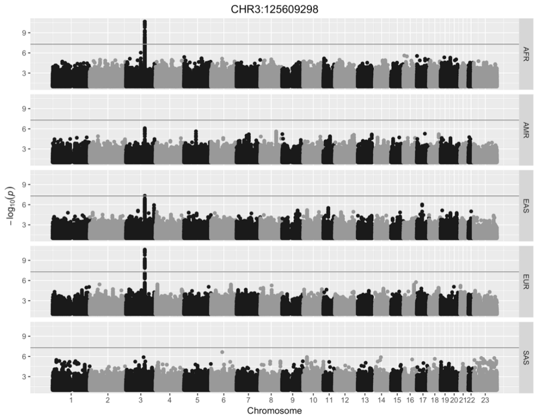
S3.2 Fig. Causal diagram of the hypothesized relationship between tagging SNPs, HERV-K insertion polymorphisms, and human disease phenotypes. SNPs associated with disease phenotypes via genome-wide association studies are tagging (i.e. in strong LD with) the true causal variant, a polymorphic HERV-K insertion.

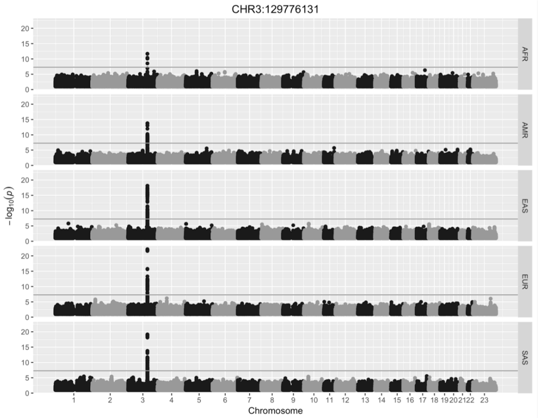

CHR1:106015875

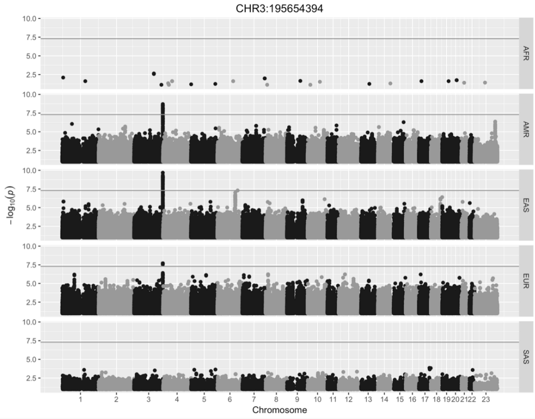

CHR1:111802591

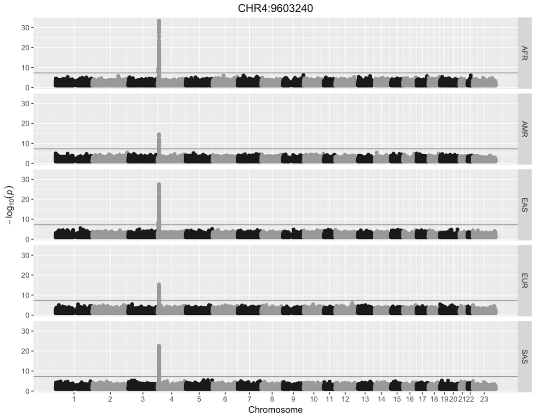

CHR1:223578304

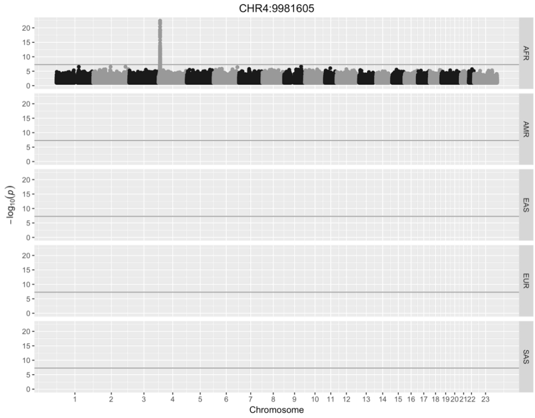

CHR3:14132684

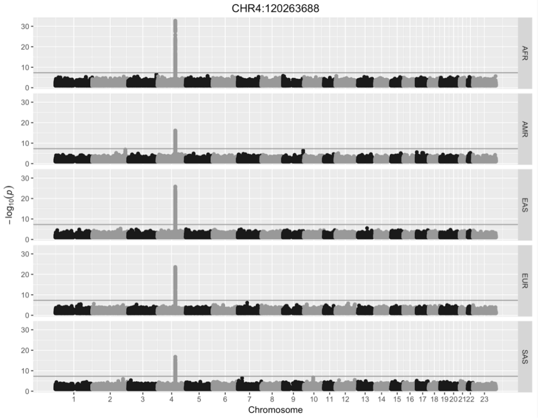

CHR3:125609298

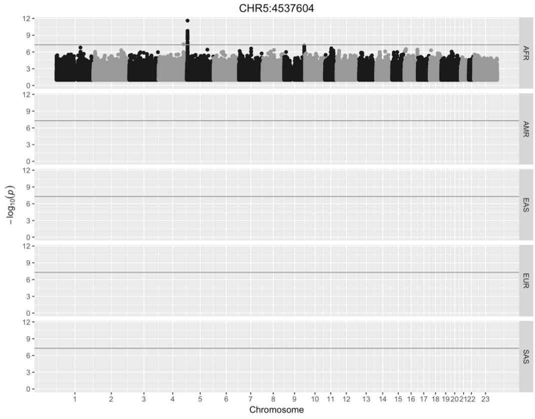

CHR3:129776131

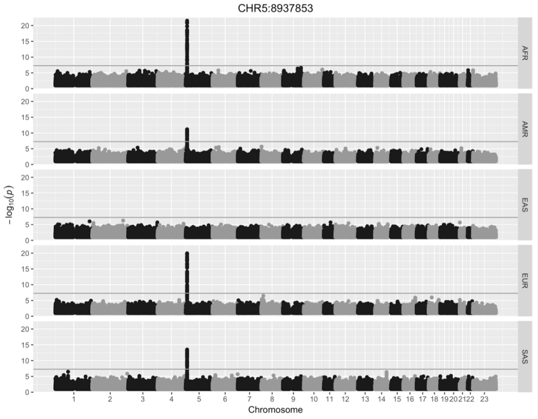

CHR3:195654394

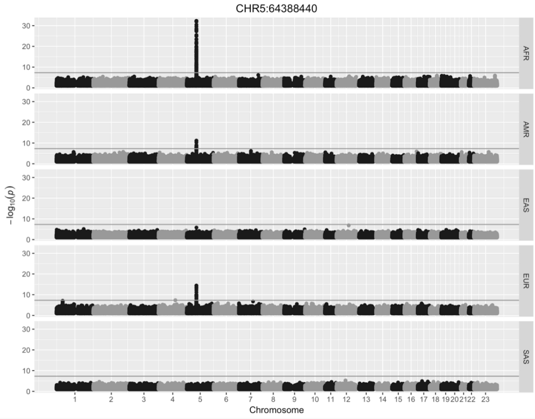

CHR4:9603240

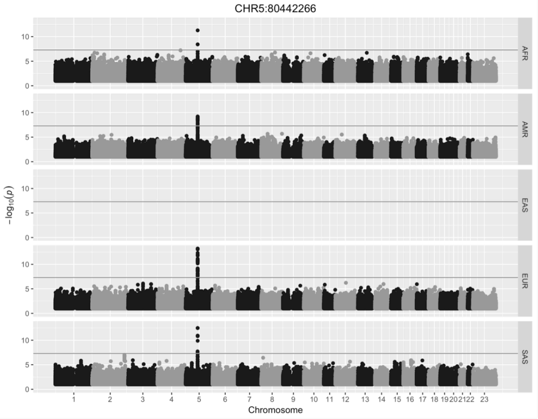

CHR4:9981605

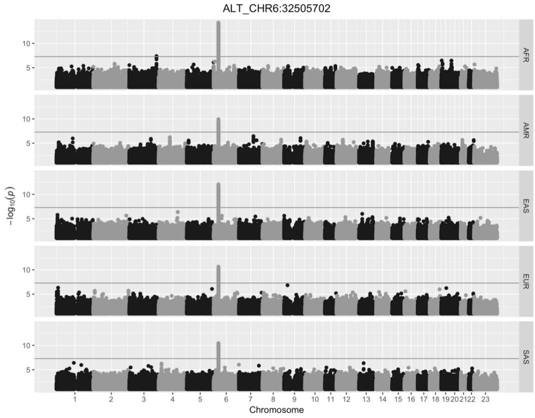

CHR4:120263688

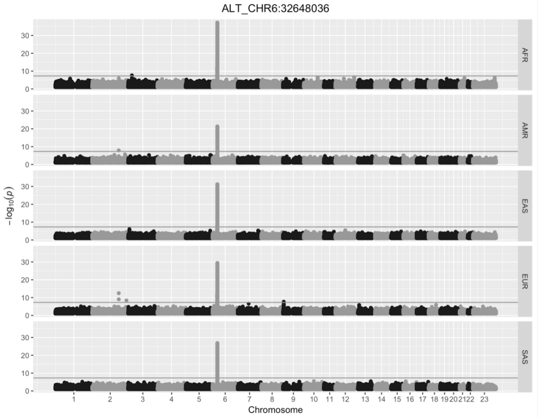

CHR5:4537604

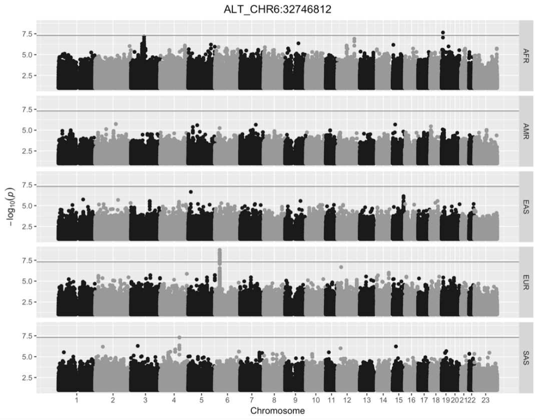

CHR5:8937853

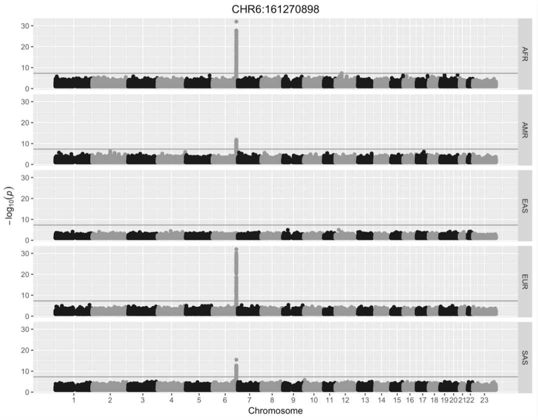

CHR5:64388440
$-\log_{10}(p)$
0
10
20
30
AFR
AMR
EAS
EUR
SAS
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Chromosome

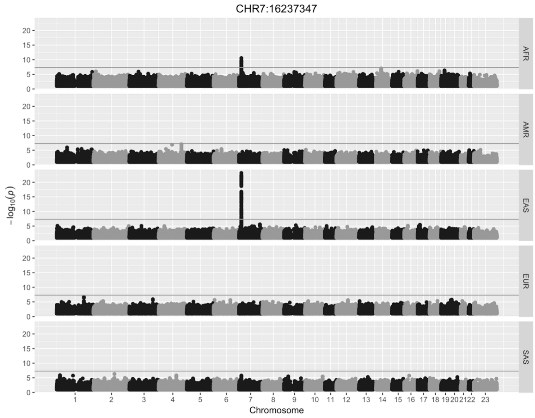

CHR5:80442266
$-\log_{10}(p)$
0
5
10
AFR
AMR
EAS
EUR
SAS
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Chromosome

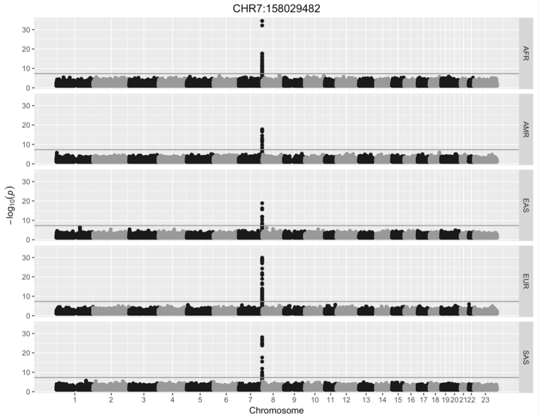

ALT_CHR6:32505702
$-\log_{10}(p)$
5
10
AFR
AMR
EAS
EUR
SAS
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Chromosome

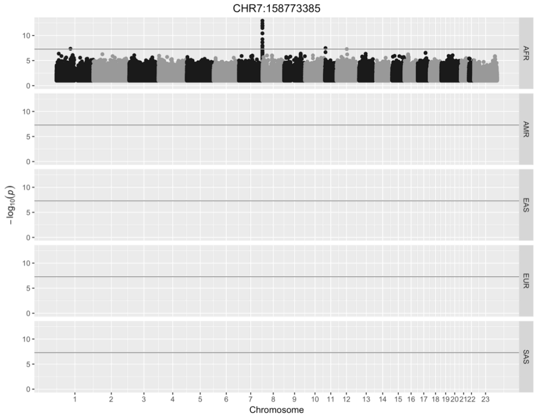

ALT_CHR6:32648036
$-\log_{10}(p)$
0
10
20
30
AFR
AMR
EAS
EUR
SAS
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Chromosome

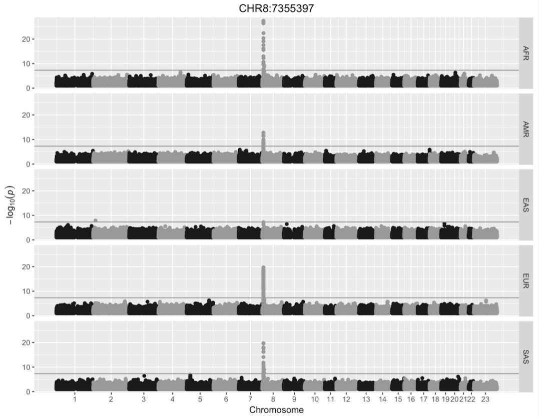

ALT_CHR6:32746812
$-\log_{10}(p)$
2.5
5.0
7.5
AFR
AMR
EAS
EUR
SAS
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Chromosome

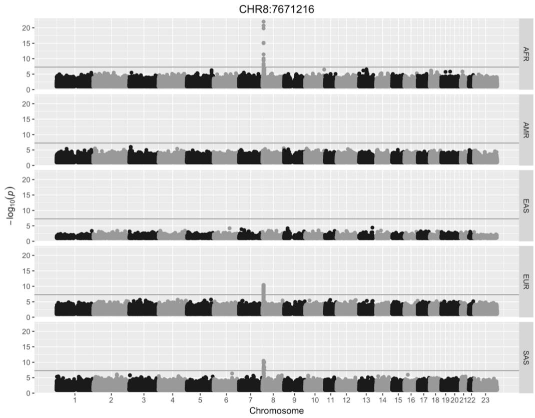

CHR6:161270898
$-\log_{10}(p)$
0
10
20
30
AFR
AMR
EAS
EUR
SAS
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Chromosome

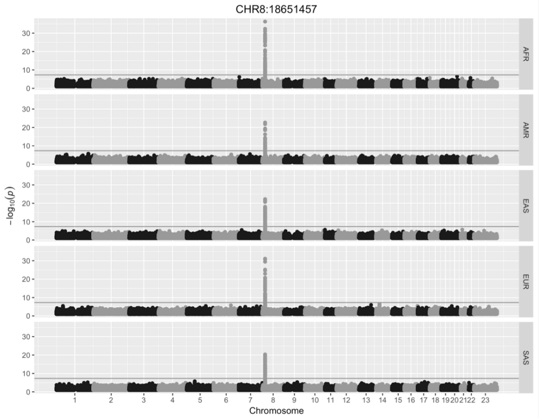

CHR7:16237347
AFR
AMR
EAS
EUR
SAS
$-\log_{10}(p)$
0
5
10
15
20
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Chromosome

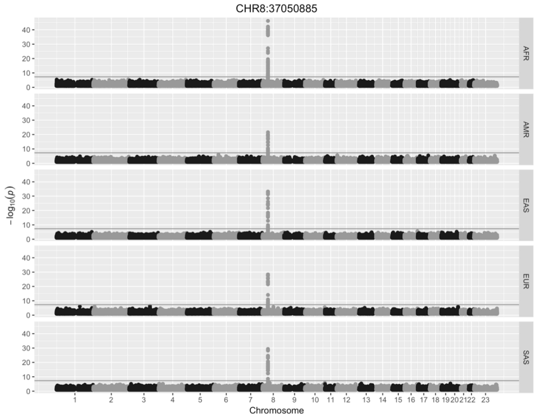

CHR7:158029482
AFR
AMR
EAS
EUR
SAS
$-\log_{10}(p)$
0
10
20
30
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Chromosome

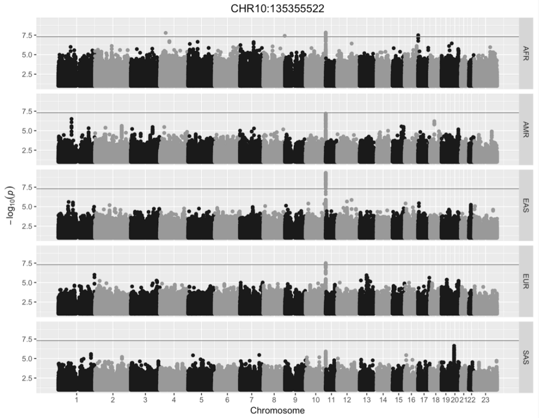

CHR7:158773385
AFR
AMR
EAS
EUR
SAS
$-\log_{10}(p)$
0
5
10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Chromosome

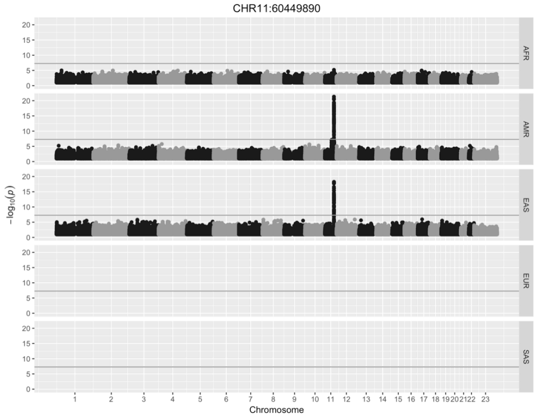

CHR8:7355397
AFR
AMR
EAS
EUR
SAS
$-\log_{10}(p)$
0
10
20
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Chromosome

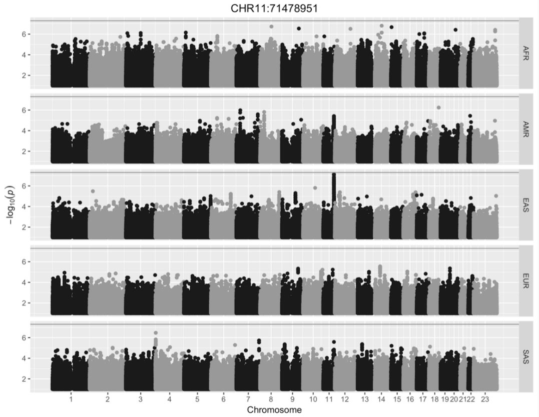

CHR8:7671216
AFR
AMR
EAS
EUR
SAS
$-\log_{10}(p)$
0
5
10
15
20
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Chromosome

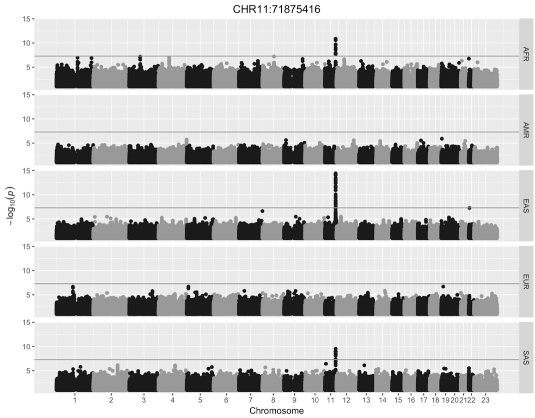

CHR8:18651457
AFR
AMR
EAS
EUR
SAS
$-\log_{10}(p)$
0
10
20
30
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Chromosome

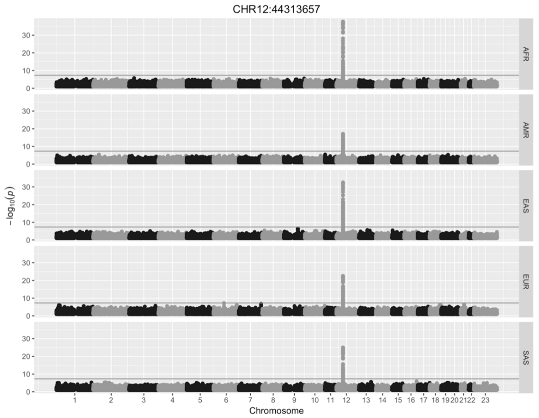

CHR8:37050885

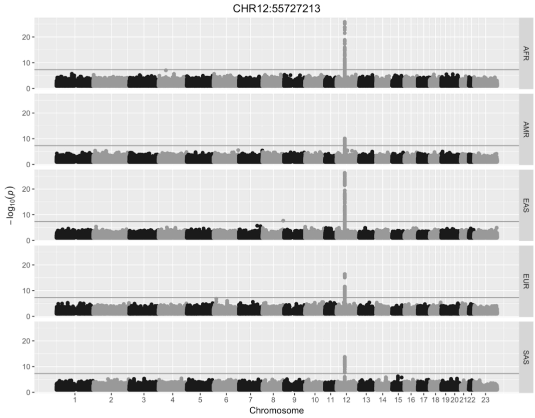

CHR10:135355522

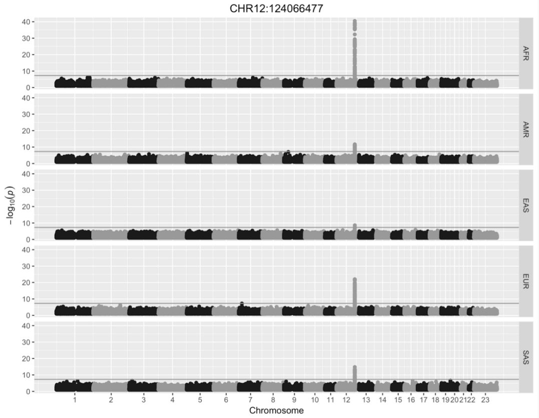

CHR11:60449890

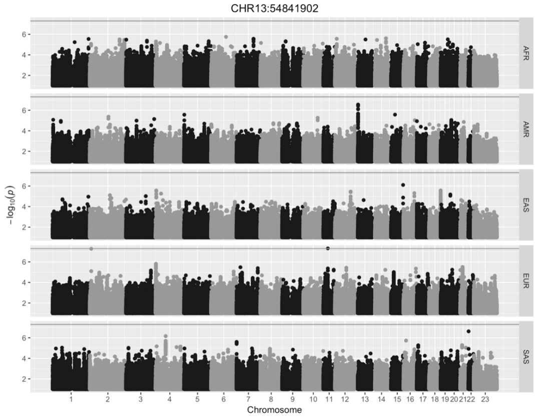

CHR11:71478951

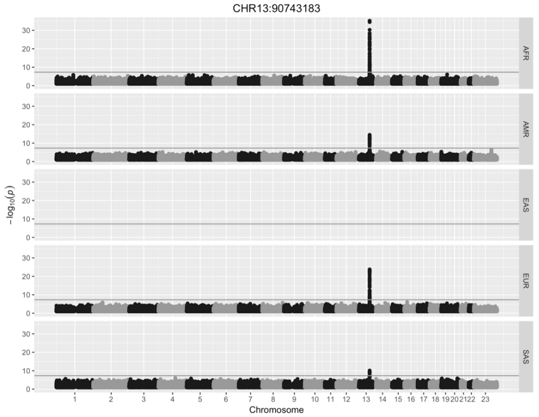

CHR11:71875416

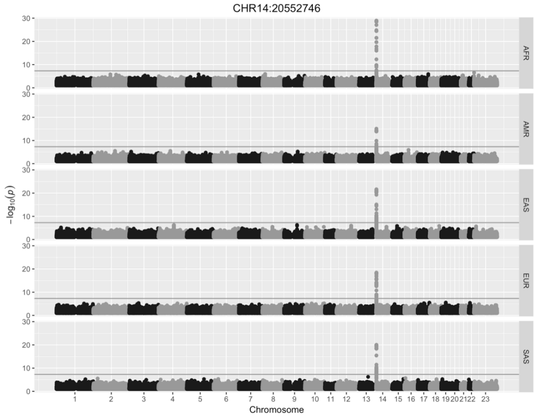

CHR12:44313657

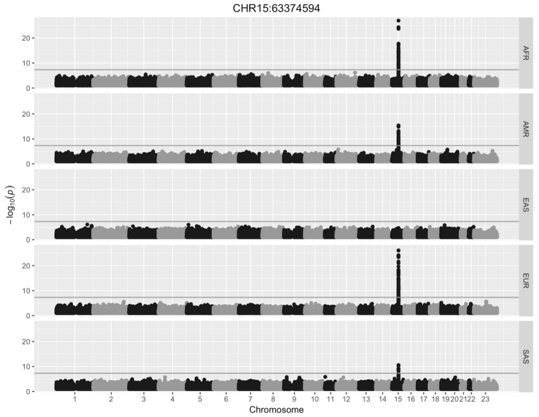

CHR12:55727213

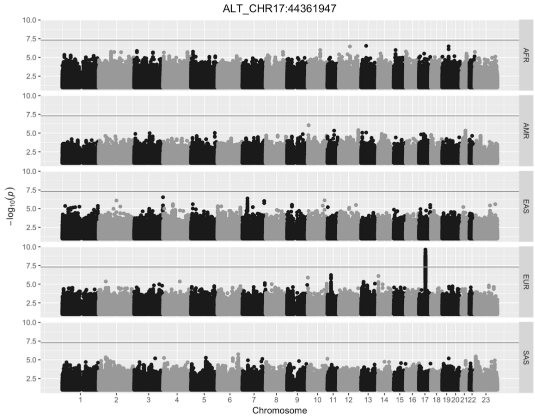

CHR12:124066477

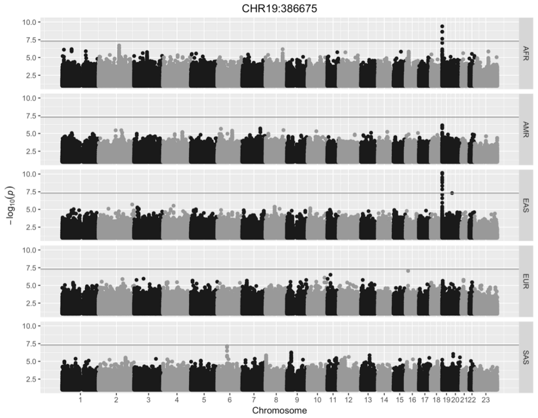

CHR13:54841902

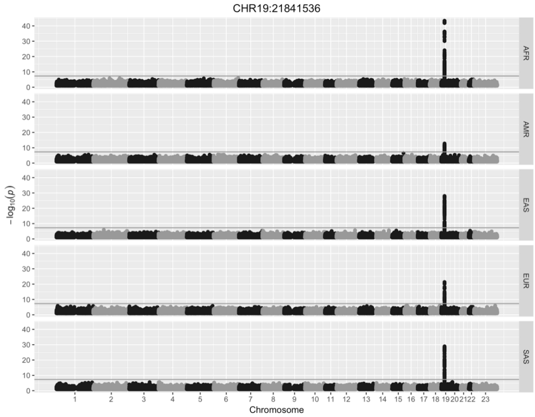

CHR13:90743183

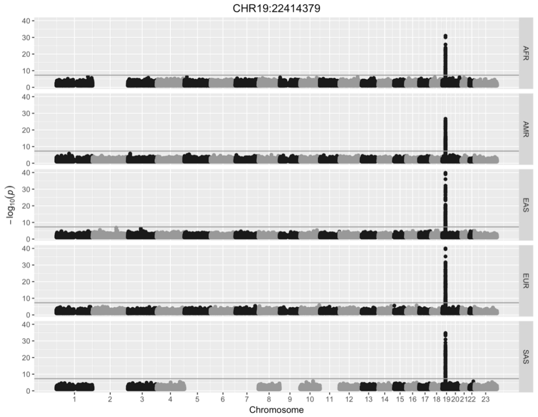

CHR14:20552746

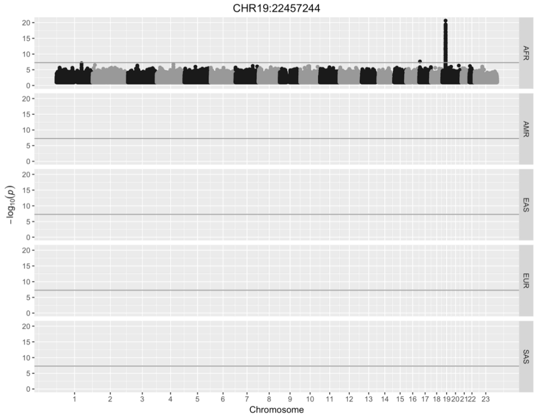

CHR15:63374594

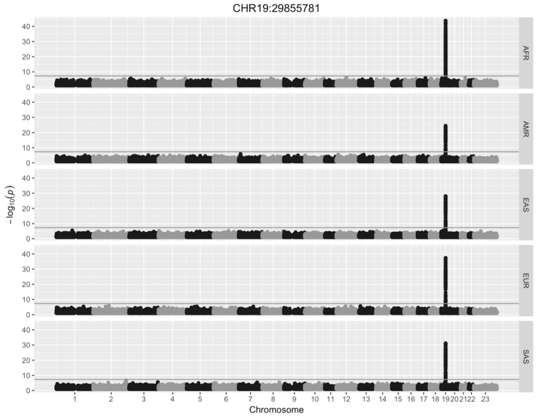

ALT_CHR17:44361947

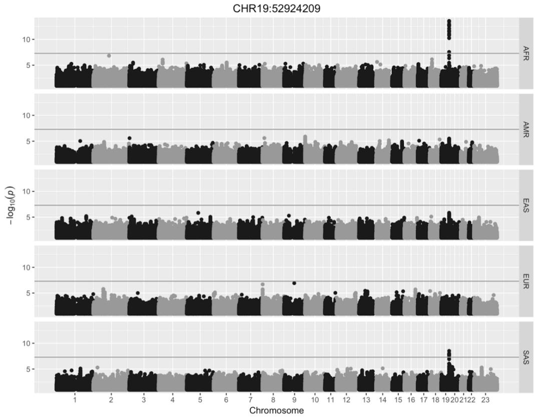

CHR19:386675

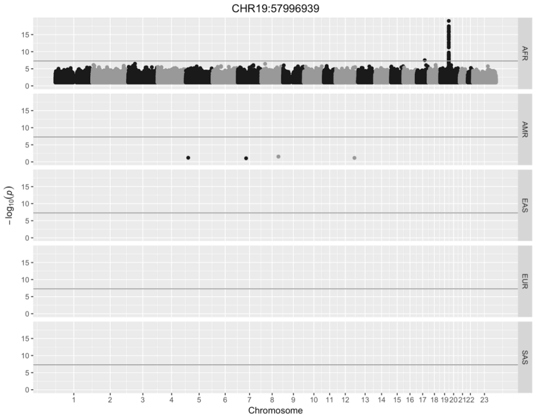

CHR19:21841536

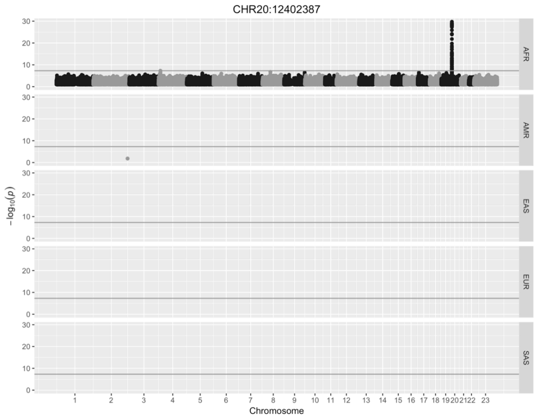

CHR19:22414379

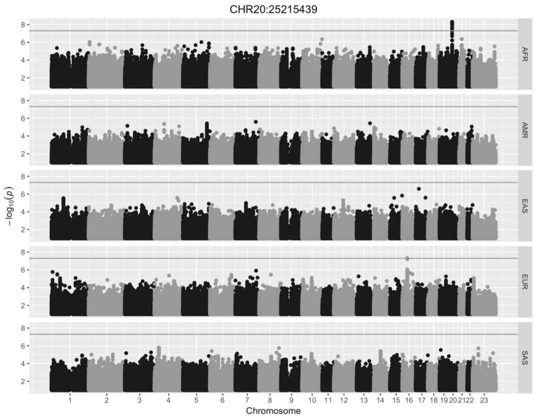

CHR19:22457244

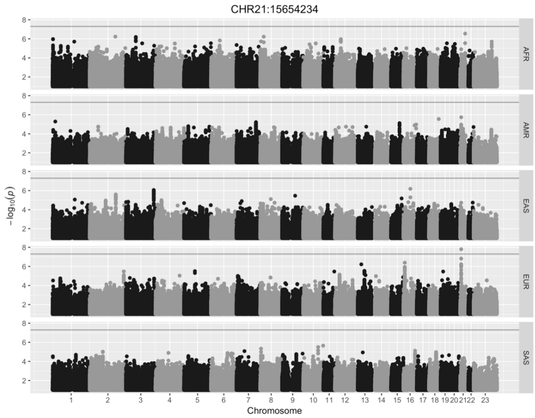

CHR19:29855781

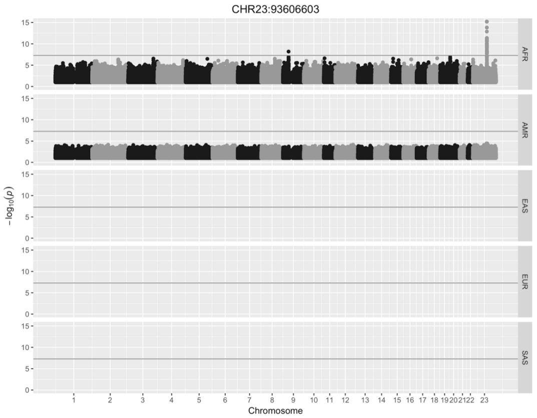

CHR19:52924209


CHR19:57996939

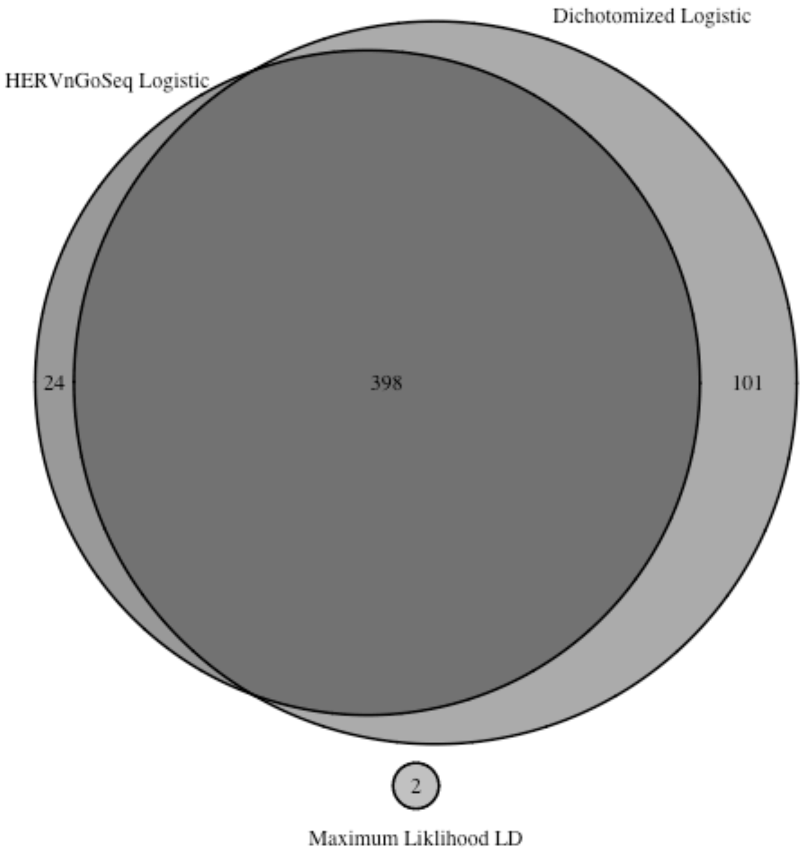

CHR20:12402387

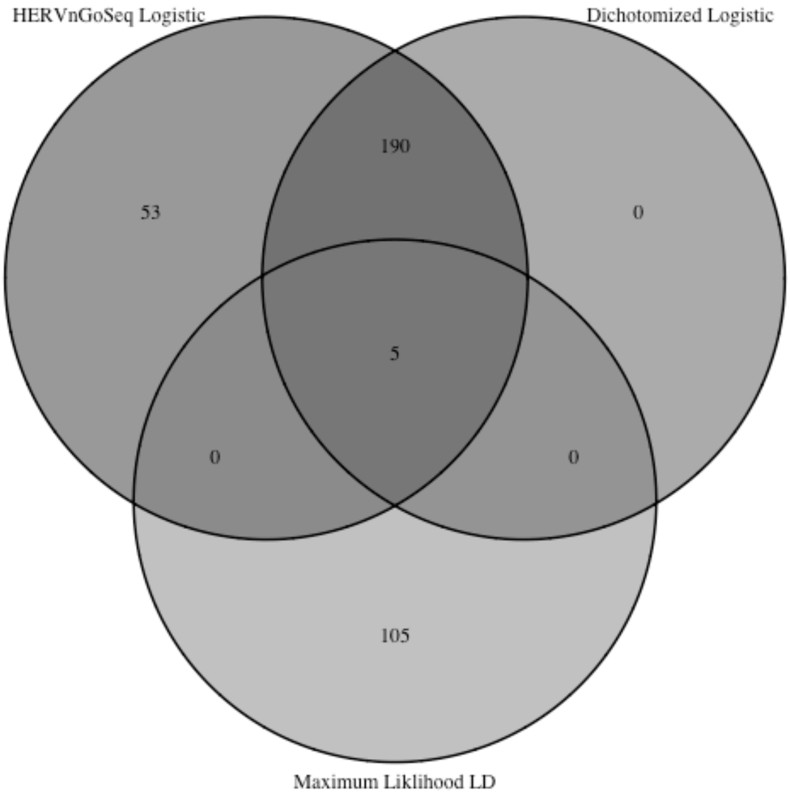

CHR20:25215439

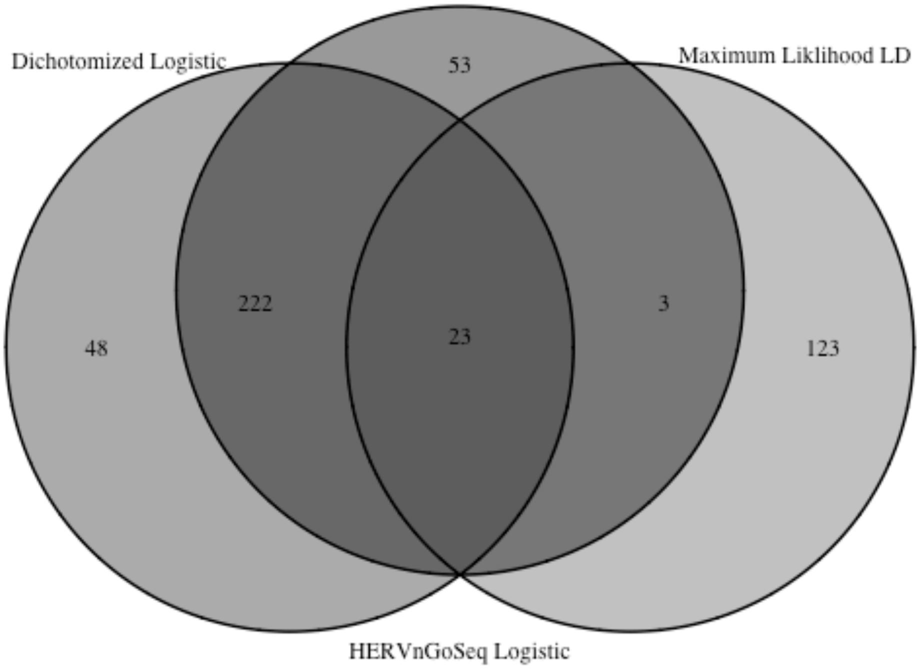

CHR21:15654234

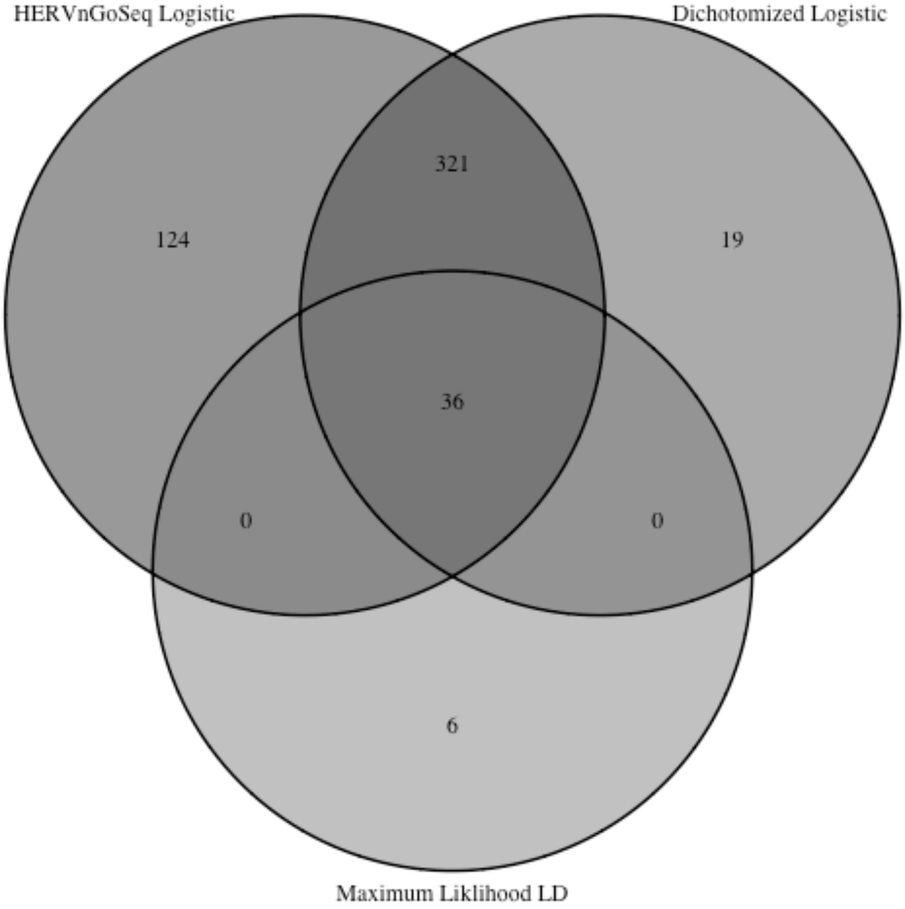

CHR23:93606603

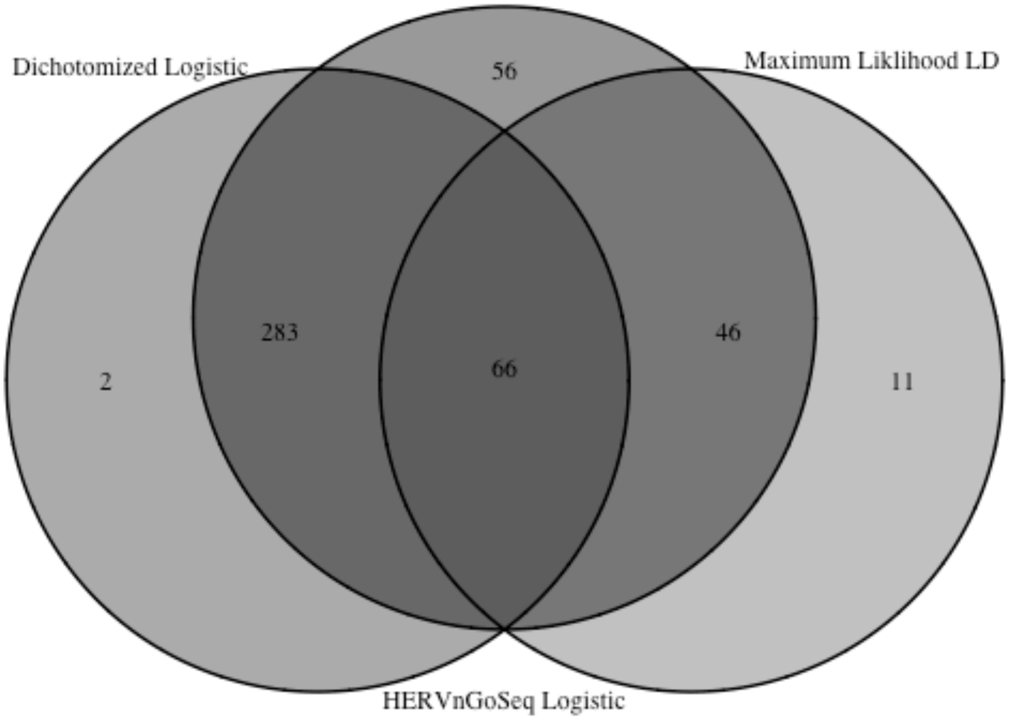
S3.3-S3.50 Fig. Manhattan plots for each of the 48 HERV-K insertion sites that are 'taggable' by SNPs. Logistic regression p-values stratified by 5 super populations. AFR African, AMR Native American, EUR European, EAS East Asian, SAS South Asian.

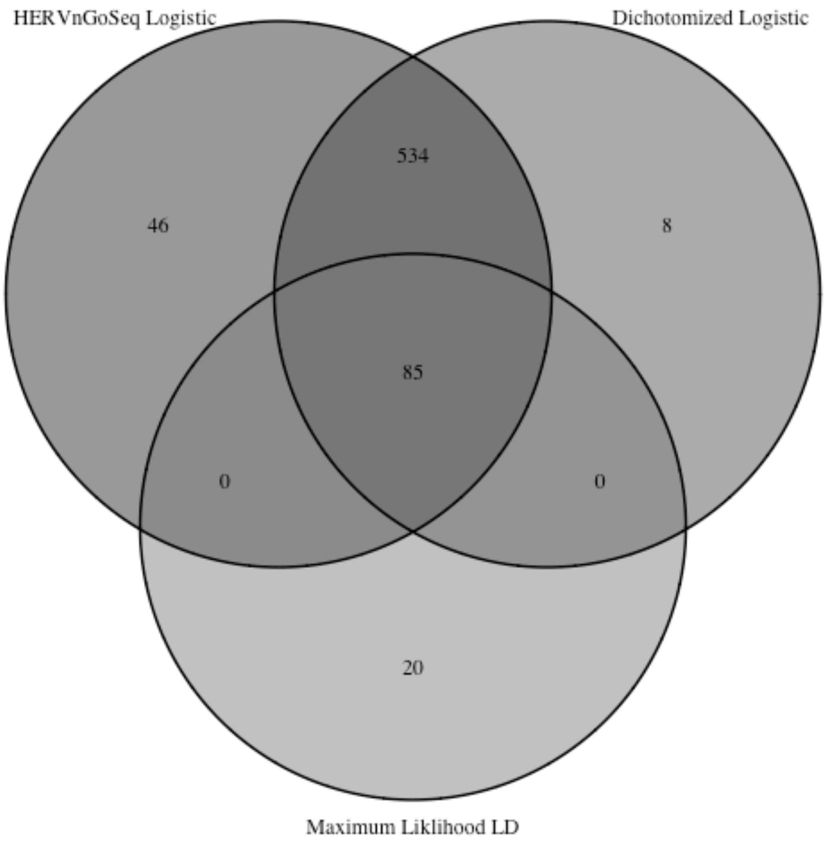

Dichotomized Logistic
HERVnGoSeq Logistic
24
398
101
2
Maximum Liklihood LD

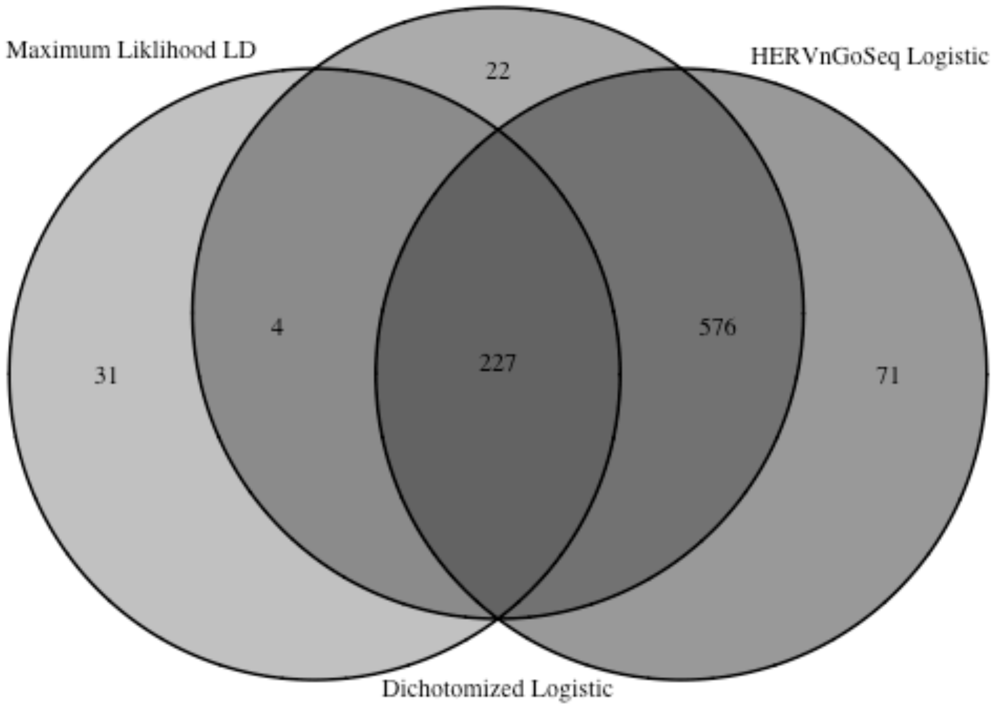

HERVnGoSeq Logistic
Dichotomized Logistic
190
53
0
5
0
0
105
Maximum Liklihood LD

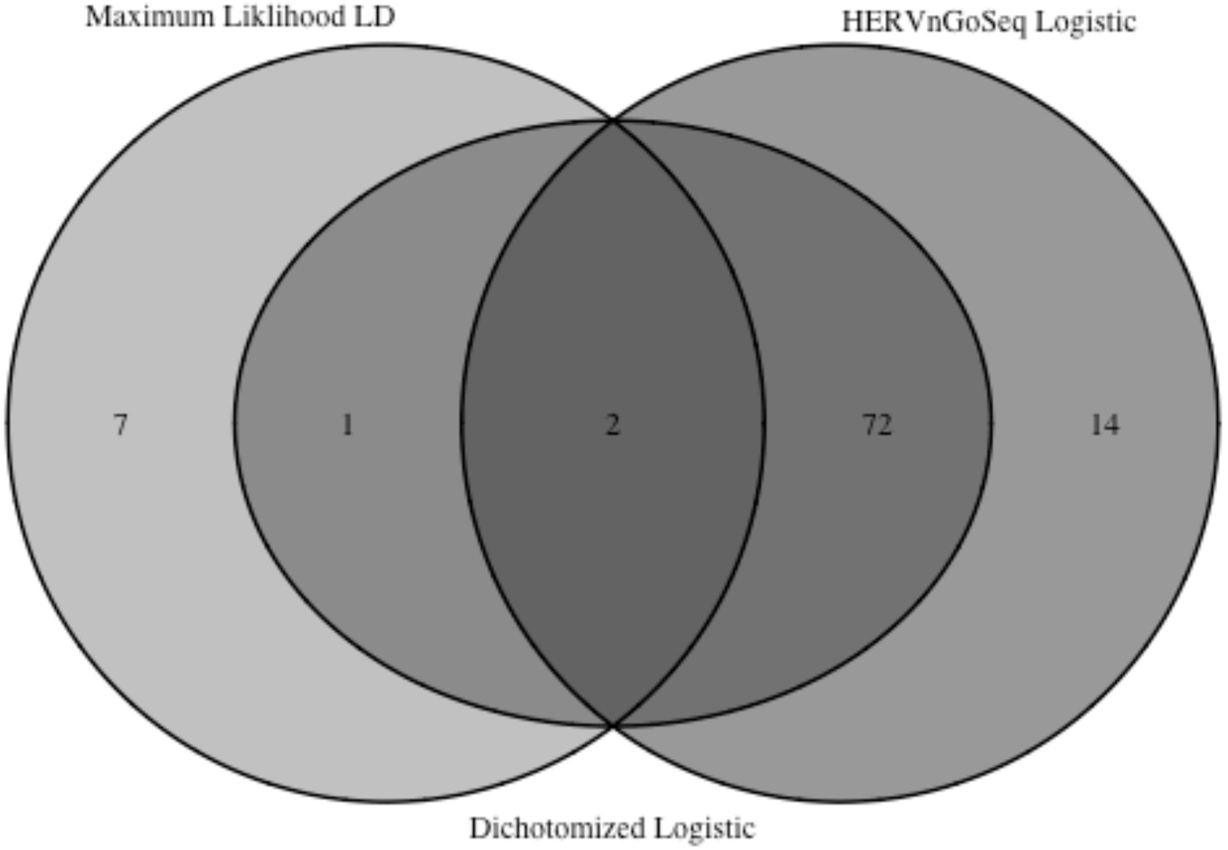

Dichotomized Logistic
53
Maximum Liklihood LD
222
23
3
48
123
HERVnGoSeq Logistic

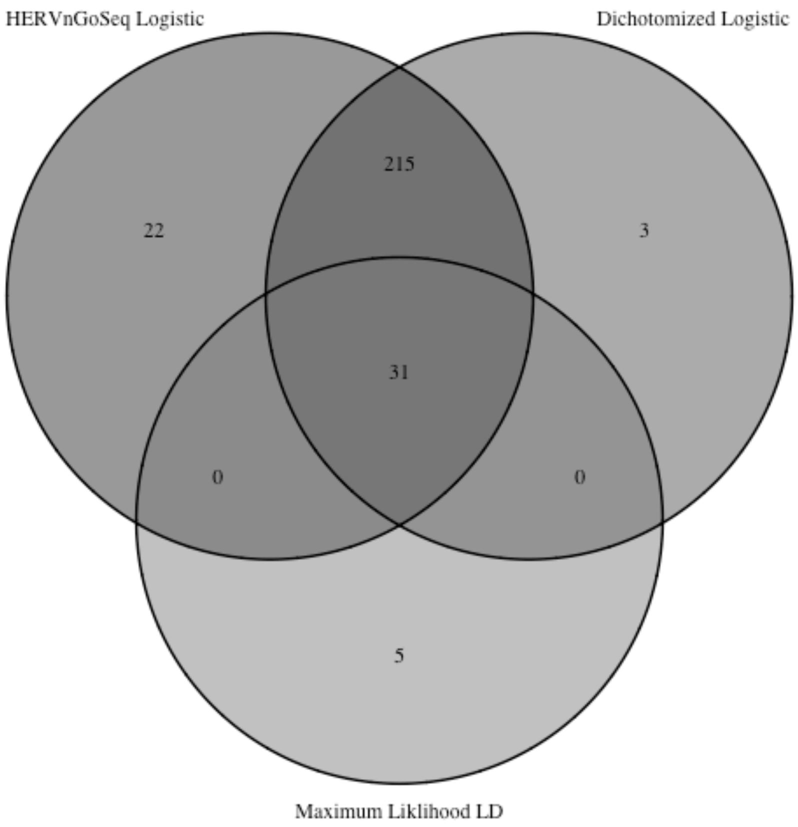

HERVnGoSeq Logistic
Dichotomized Logistic
321
124
19
36
0
0
6
Maximum Liklihood LD

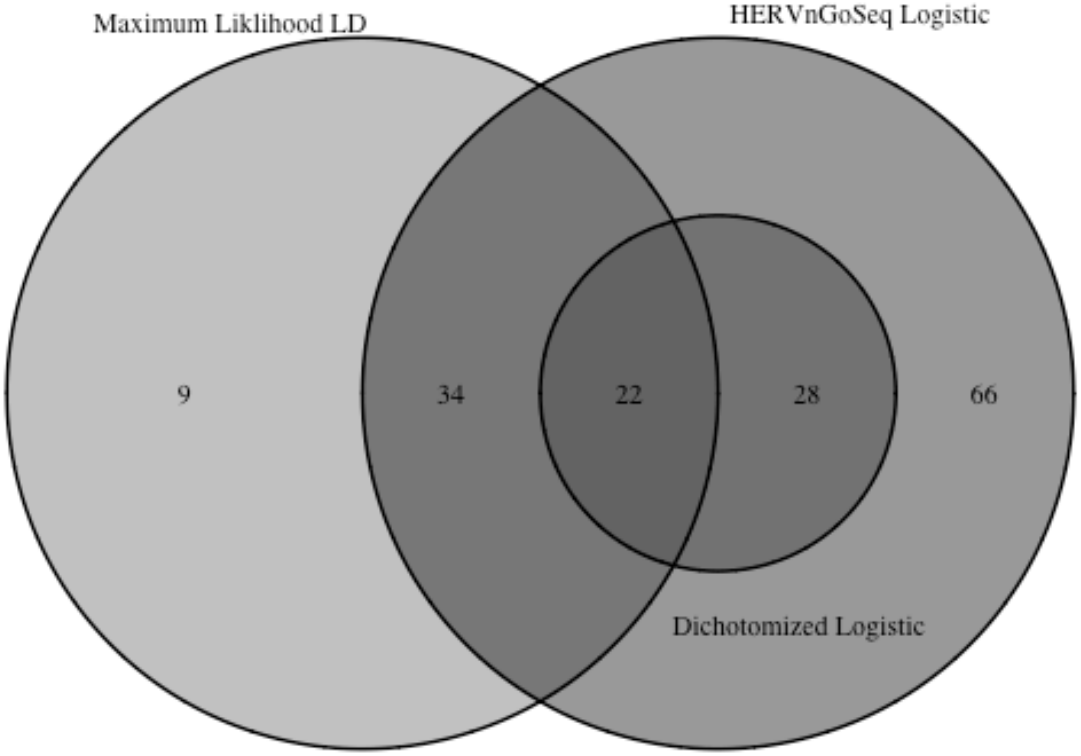

Dichotomized Logistic
Maximum Liklihood LD
HERVnGoSeq Logistic
56
283
66
46
2
11

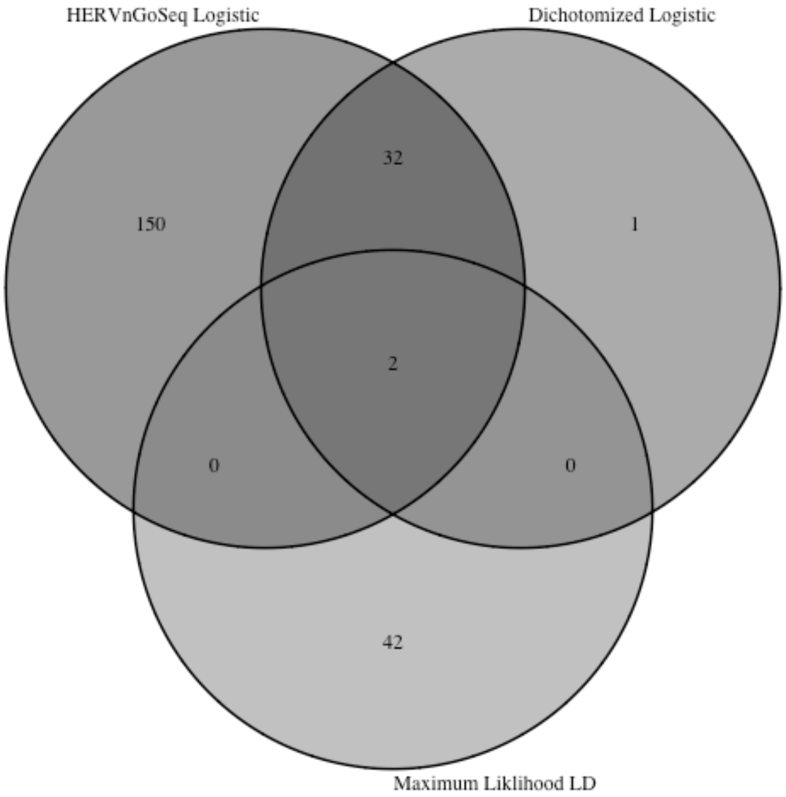

HERVnGoSeq Logistic
Dichotomized Logistic
Maximum Liklihood LD
534
46
8
85
0
0
20

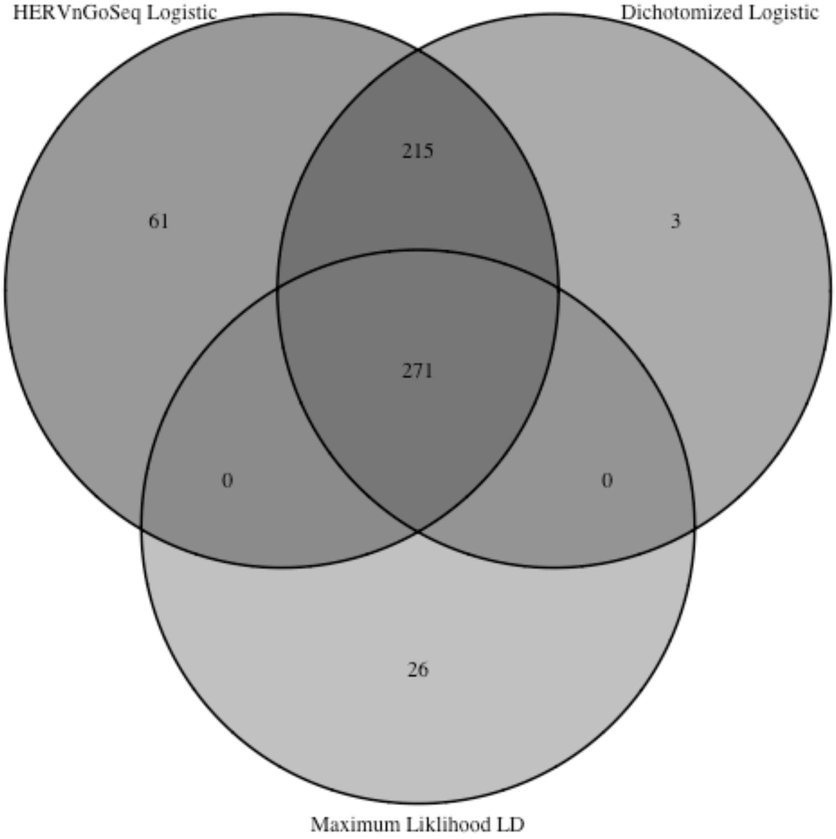

Maximum Liklihood LD
22
HERVnGoSeq Logistic
4
576
31
227
71
Dichotomized Logistic


Maximum Liklihood LD
HERVnGoSeq Logistic
7
1
2
72
14
Dichotomized Logistic

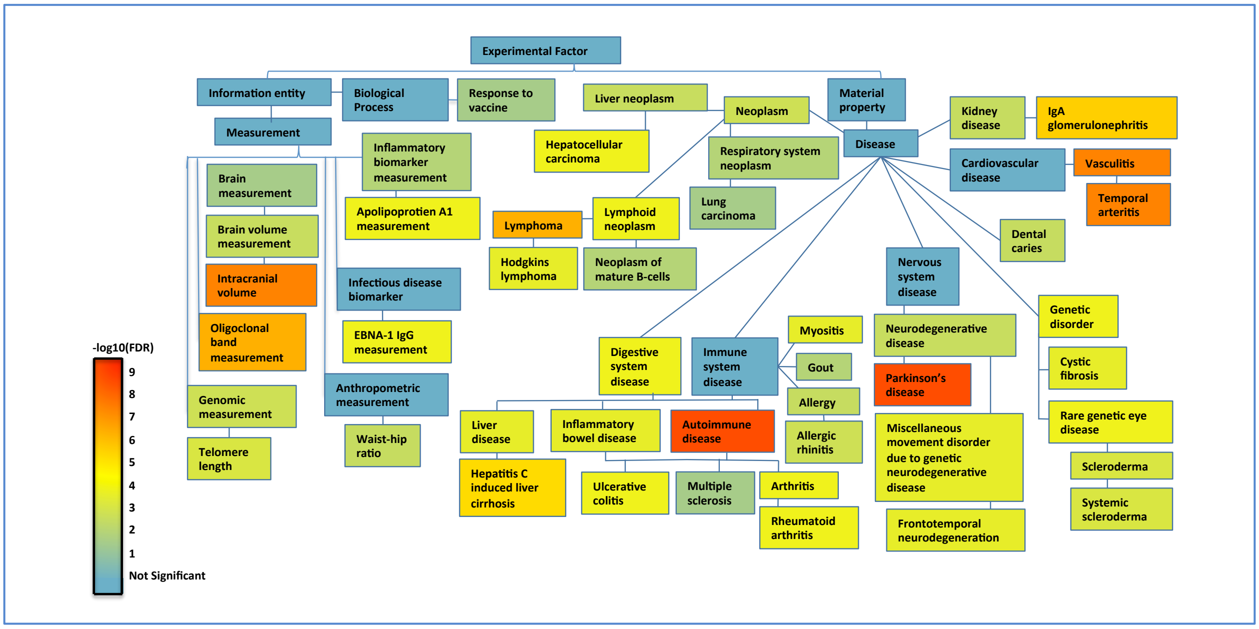

HERVnGoSeq Logistic
Dichotomized Logistic
215
22
3
31
0
0
5
Maximum Liklihood LD


Maximum Liklihood LD
HERVnGoSeq Logistic
9
34
22
28
66
Dichotomized Logistic

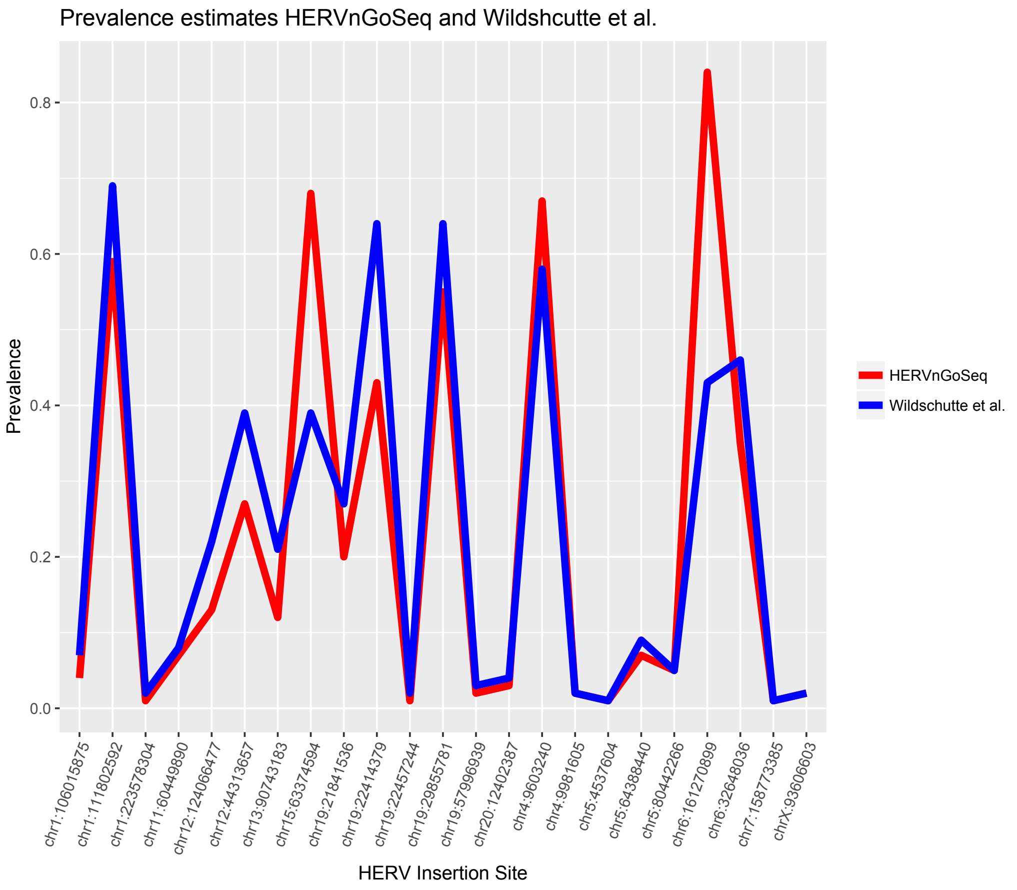

HERVnGoSeq Logistic
Dichotomized Logistic
32
150
1
2
0
0
42
Maximum Liklihood LD


HERVnGoSeq Logistic
Dichotomized Logistic
215
61
3
271
0
0
26
Maximum Liklihood LD

S3.51-S3.63 Fig. Venn diagrams. Diagrams show the overlap between hiSNP sets derived by logistic regression of HERVnGoSeq presence/absence calls, logistic regression of dichotomized genotypes called by the 1000 Genomes Structural Variant Group, and the r^2 measure of linkage disequilibrium



S3.64 Fig. Hierarchical chart of experimental factor ontology enrichment terms. Hue intensity is inversely proportional to the magnitude of the p-value for significance.



S3.65 Fig. Estimated prevalence of 23 non-reference HERV-K insertion sites detected in the 1000 Genomes population by both HERVnGoSeq (red) and a previous publication (blue).

3.6 References

1 Gonzalez-Cao, M. *et al.* Human endogenous retroviruses and cancer. *Cancer Biol Med* **13**, 483-488, doi:10.20892/j.issn.2095-3941.2016.0080 (2016).

2 Mangeney, M. *et al.* Placental syncytins: Genetic disjunction between the fusogenic and immunosuppressive activity of retroviral envelope proteins. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 20534-20539, doi:10.1073/pnas.0707873105 (2007).

3 Huber, B. T., Hsu, P. N. & Sutkowski, N. Virus-encoded superantigens. *Microbiol. Rev.* **60**, 473-482 (1996).

4 zur Hausen, H. The search for infectious causes of human cancers: where and why (Nobel lecture). *Angew. Chem. Int. Ed. Engl.* **48**, 5798-5808, doi:10.1002/anie.200901917 (2009).

5 Vieira, V. C. & Soares, M. A. The role of cytidine deaminases on innate immune responses against human viral infections. *Biomed Res Int* **2013**, 683095, doi:10.1155/2013/683095 (2013).

6 Wildschutte, J. H. *et al.* Discovery of unfixed endogenous retrovirus insertions in diverse human populations. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E2326-2334, doi:10.1073/pnas.1602336113 (2016).

7 Bannert, N. & Kurth, R. The evolutionary dynamics of human endogenous retroviral families. *Annu Rev Genomics Hum Genet* **7**, 149-173, doi:10.1146/annurev.genom.7.080505.115700 (2006).

8 Feschotte, C. & Gilbert, C. Endogenous viruses: insights into viral evolution and impact on host biology. *Nat Rev Genet* **13**, 283-296, doi:10.1038/nrg3199 (2012).

9 Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10**, 691-703, doi:10.1038/nrg2640 (2009).

10 Jern, P. & Coffin, J. M. Effects of retroviruses on host genome function. *Annu Rev Genet* **42**, 709-732, doi:10.1146/annurev.genet.42.110807.091501 (2008).

11 Simpson, G. R. *et al.* Endogenous D-type (HERV-K) related sequences are packaged into retroviral particles in the placenta and possess open reading frames for reverse transcriptase. *Virology* **222**, 451-456, doi:10.1006/viro.1996.0443 (1996).

12 Hurst, T. P. & Magiorkinis, G. Activation of the innate immune response by endogenous retroviruses. *J Gen Virol* **96**, 1207-1218, doi:0.1099/jgv.0.000017
10.1099/jgv.0.000017 (2015).

13 Mortelmans, K., Wang-Johanning, F. & Johanning, G. L. The role of human endogenous retroviruses in brain development and function. *APMIS* **124**, 105-115, doi:10.1111/apm.12495 (2016).

14 Sekar, A. *et al.* Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177-183, doi:10.1038/nature16549 (2016).

15 Li, W. *et al.* Human endogenous retrovirus-K contributes to motor neuron disease. *Science translational medicine* **7**, doi:ARTN 307ra153
10.1126/scitranslmed.aac8201 (2015).

16 Bowen, L. N. *et al.* HIV-associated motor neuron disease: HERV-K activation and response to antiretroviral therapy. *Neurology* **87**, 1756-1762, doi:10.1212/WNL.0000000000003258 (2016).

17 Campbell, I. M. *et al.* Human endogenous retroviral elements promote genome instability via non-allelic homologous recombination. *BMC Biol.* **12**, 74, doi:10.1186/s12915-014-0074-4 (2014).

18 Chuong, E. B., Rumi, M. A., Soares, M. J. & Baker, J. C. Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nat Genet* **45**, 325-329, doi:10.1038/ng.2553 (2013).

19 Fuchs, N. V. *et al.* Human endogenous retrovirus K (HML-2) RNA and protein expression is a marker for human embryonic and induced pluripotent stem cells. *Retrovirology* **10**, 115, doi:10.1186/1742-4690-10-115 (2013).

20 Ling, J. *et al.* The ERV-9 LTR enhancer is not blocked by the HS5 insulator and synthesizes through the HS5 site non-coding, long RNAs that regulate LTR enhancer function. *Nucleic Acids Res* **31**, 4582-4596 (2003).

21 Kim, H. S. Genomic impact, chromosomal distribution and transcriptional regulation of HERV elements. *Mol Cells* **33**, 539-544, doi:10.1007/s10059-012-0037-y (2012).

22 Kreimer, U., Schulz, W. A., Koch, A., Niegisch, G. & Goering, W. HERV-K and LINE-1 DNA Methylation and Reexpression in Urothelial Carcinoma. *Front Oncol* **3**, 255, doi:10.3389/fonc.2013.00255 (2013).

23 Cegolon, L. *et al.* Human endogenous retroviruses and cancer prevention: evidence and prospects. *BMC Cancer* **13**, 4, doi:10.1186/1471-2407-13-4 (2013).

24 Nexo, B. A. *et al.* Are human endogenous retroviruses triggers of autoimmune diseases? Unveiling associations of three diseases and viral loci. *Immunol Res* **64**, 55-63, doi:10.1007/s12026-015-8671-z (2016).

25 Ray, D. A. & Batzer, M. A. Reading TE leaves: new approaches to the identification of transposable element insertions. *Genome Res.* **21**, 813-820, doi:10.1101/gr.110528.110 (2011).

26 Witherspoon, D. J. *et al.* Mobile element scanning (ME-Scan) by targeted high-throughput sequencing. *BMC Genomics* **11**, 410, doi:10.1186/1471-2164-11-410 (2010).

27 Hancks, D. C. & Kazazian, H. H., Jr. SVA retrotransposons: Evolution and genetic instability. *Semin. Cancer Biol.* **20**, 234-245, doi:10.1016/j.semcancer.2010.04.001 (2010).

28 Bao, W. D., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**, doi:UNSP 11
10.1186/s13100-015-0041-9 (2015).

29 Barbulescu, M. *et al.* Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. *Curr. Biol.* **9**, 861-868, doi:Doi 10.1016/S0960-9822(99)80390-X (1999).

30 Belshaw, R. *et al.* Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K(HML2): Implications for present-day activity. *J. Virol.* **79**, 12507-12514, doi:10.1128/Jvi.79.19.12507-12507-12514.2005 (2005).

31 Bennettt, E. A., Coleman, L. E., Tsui, C., Pittard, W. S. & Devine, S. E. Natural genetic variation caused by transposable elements in humans. *Genetics* **168**, 933-951, doi:10.1534/genetics.104.031757 (2004).

32 Dangel, A. W. *et al.* The Dichotomous Size Variation of Human-Complement C4 Genes Is Mediated by a Novel Family of Endogenous Retroviruses, Which Also Establishes Species-Specific Genomic Patterns among Old-World Primates. *Immunogenetics* **40**, 425-436 (1994).

33 Hughes, J. F. & Coffin, J. M. Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: Implications for human and viral evolution. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 1668-1672, doi:10.1073/pnas.0307885100 (2004).

34 Macfarlane, C. & Simmonds, P. Allelic variation of HERV-K(HML-2) endogenous retroviral elements in human populations. *J. Mol. Evol.* **59**, 642-656, doi:10.1007/s00239-004-2656-1 (2004).

35 Mamedov, I., Lebedev, Y., Hunsmann, G., Khusnutdinova, E. & Sverdlov, E. A rare event of insertion polymorphism of a HERV-K LTR in the human genome. *Genomics* **84**, 596-599, doi:10.1016/j.ygeno.2004.04.010 (2004).

36 Mayer, J. *et al.* An almost-intact human endogenous retrovirus K on human chromosome 7. *Nat. Genet.* **21**, 257-258 (1999).

37 Moyes, D. L. *et al.* The distribution of the endogenous retroviruses HERV-K113 and HERV-K115 in health and disease. *Genomics* **86**, 337-341, doi:DOI 10.1016/j.ygeno.2005.06.004 (2005).

38 Turner, G. *et al.* Insertional polymorphisms of full-length endogenous retroviruses in humans. *Curr. Biol.* **11**, 1531-1535, doi:Doi 10.1016/S0960-9822(01)00455-9 (2001).

39 Marchi, E., Kanapin, A., Magiorkinis, G. & Belshaw, R. Unfixed Endogenous Retroviral Insertions in the Human Population. *J. Virol.* **88**, 9529-9537, doi:10.1128/Jvi.00919-14 (2014).

40 Lee, E. *et al.* Landscape of Somatic Retrotransposition in Human Cancers. *Science* **337**, 967-971, doi:10.1126/science.1222077 (2012).

41 Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56-64, doi:10.1038/nature06862 (2008).

42 Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75-+, doi:10.1038/nature15394 (2015).

43 Gazal, S., Sahbatou, M., Babron, M. C., Genin, E. & Leutenegger, A. L. High level of inbreeding in final phase of 1000 Genomes Project. *Sci. Rep.* **5**, 17453, doi:10.1038/srep17453 (2015).

44 Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7, doi:10.1186/s13742-015-0047-8 (2015).

45 Carithers, L. J. & Moore, H. M. The Genotype-Tissue Expression (GTEx) Project. *Biopreserv Biobank* **13**, 307-308, doi:10.1089/bio.2015.29031.hmm (2015).

46 MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896-D901, doi:10.1093/nar/gkw1133 (2017).

47 Genomes Project, C. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:10.1038/nature09534 (2010).

48 Pratto, F. *et al.* Recombination initiation maps of individual human genomes. *Science* **346**, 826-+, doi:UNSP 1256442
10.1126/science.1256442 (2014).

49 Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**, D493-D496, doi:10.1093/nar/gkh103 (2004).

50 Wang, J. *et al.* dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. *Hum. Mutat.* **27**, 323-329, doi:10.1002/humu.20307 (2006).

51 Ke, X. *et al.* The impact of SNP density on fine-scale patterns of linkage disequilibrium. *Hum. Mol. Genet.* **13**, 577-588, doi:10.1093/hmg/ddh060 (2004).

52 Ardlie, K. G., Kruglyak, L. & Seielstad, M. Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet* **3**, 299-309, doi:10.1038/nrg777 (2002).

53 Katzourakis, A., Pereira, V. & Tristem, M. Effects of recombination rate on human endogenous retrovirus fixation and persistence. *J. Virol.* **81**, 10712-10717, doi:10.1128/JVI.00410-07 (2007).

54 Alfahad, T. & Nath, A. Retroviruses and amyotrophic lateral sclerosis. *Antiviral Res.* **99**, 180-187, doi:10.1016/j.antiviral.2013.05.006 (2013).

55 Garrison, K. E. *et al.* T cell responses to human endogenous retroviruses in HIV-1 infection. *PLoS Pathog.* **3**, e165, doi:10.1371/journal.ppat.0030165 (2007).

56 Fanous, A. H. *et al.* Genome-wide association study of clinical dimensions of schizophrenia: polygenic effect on disorganized symptoms. *Am. J. Psychiatry* **169**, 1309-1317, doi:10.1176/appi.ajp.2012.12020218 (2012).

57 Hamel, M. G. *et al.* Multimodal signaling by the ADAMTSs (a disintegrin and metalloproteinase with thrombospondin motifs) promotes neurite extension. *Exp. Neurol.* **210**, 428-440, doi:10.1016/j.expneurol.2007.11.014 (2008).

58 Rietveld, C. A. *et al.* Common genetic variants associated with cognitive performance identified using the proxy-phenotype method. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 13790-13794, doi:10.1073/pnas.1404623111 (2014).

59 Schmitt, K. *et al.* Comprehensive analysis of human endogenous retrovirus group HERV-W locus transcription in multiple sclerosis brain lesions by high-throughput amplicon sequencing. *J. Virol.* **87**, 13837-13852, doi:10.1128/JVI.02388-13 (2013).

60 Ma, W. J. *et al.* Human Endogenous Retroviruses-K (HML-2) Expression Is Correlated with Prognosis and Progress of Hepatocellular Carcinoma. *Biomed Research International*, doi:Artn 8201642
10.1155/2016/8201642 (2016).

61 Chiu, Y. L. & Greene, W. C. The APOBEC3 cytidine deaminases: an innate defensive network opposing exogenous retroviruses and endogenous

retroelements. *Annu. Rev. Immunol.* **26**, 317-353, doi:10.1146/annurev.immunol.26.021607.090350 (2008).

62 McQueen, M. B. *et al.* The National Longitudinal Study of Adolescent to Adult Health (Add Health) sibling pairs genome-wide data. *Behav. Genet.* **45**, 12-23, doi:10.1007/s10519-014-9692-4 (2015).

4. Conclusion

4.1 Concluding Remarks and Future Directions

With the ultimate goal of preventing the most common childhood cancer, epidemiologists have amassed evidence of exogenous (e.g. environmental) and endogenous (e.g. genetic) sources of immune dysregulation preceding the onset of childhood acute lymphoblastic leukemia (ALL). Still, important questions remain to be answered regarding disease etiology. The epidemiologic studies presented here contribute clarity to both long-held and nascent theories and raise new questions regarding the causation of ALL by exploring three important hypotheses. In the first chapter, we addressed a decades-old hypothesis that early life allergic disease is associated with ALL. Next, in light of compelling new evidence that APOBEC3 enzymes of the innate immune system cause somatic mutation in ALL, we tested for association of a germline deletion of *APOBEC3B* with ALL in the second chapter. Finally, we took the first steps to explore a hypothesis born of earlier work within our group suggesting that human endogenous retrovirus-K (HERV-K) is active in ALL tumors. In the final chapter we sought to identify a role of polymorphic HERV-K insertions in the heritability diseases including ALL via established SNP associations. More broadly, this work taken together highlights the limitations and strengths of epidemiology in determining the causes of childhood ALL and the challenges and opportunities moving forward.

Epidemiologic studies of the association between allergic disease and childhood leukemia have been conducted since the 1950s. A number of these studies have yielded evidence of an inverse relationship between the two diseases, supporting a hypothesis that allergy, or the immune functions that generate allergy, are protective against ALL. More recently, studies using medical records to assess allergy exposure have exhibited a positive association with ALL, supporting the hypothesis that allergy is a risk factor of the disease, or that they share a common biological mechanism. In the CCLS overall, no associations were observed between childhood ALL risk and specific allergy phenotypes or any allergy. However, having any allergy was associated with an increased risk of ALL among the youngest study participants. Amidst this landscape of conflicting associations, we were able to incorporate results from the CCLS to evaluate sources of between-study heterogeneity using meta-regression methods. In the meta-analysis random-effect models, a reduced odds of ALL was associated with hay fever (metaOR=0.65, 95% CI: 0.47, 0.90); however, restricting the analysis to studies that used medical records for assessment of allergy or recently published studies, led to null or attenuated results. In the conclusion of the chapter, we state, "*understanding the causal relationship between allergy and childhood leukemia cannot be achieved with additional status quo epidemiologic studies*" and propose new strategies for future work. The differences due to study related factors, for example, the differences in effect observed between medical diagnosis of allergy vs. self report, may be due to the fact that they are representing two distinct exposures. It is plausible that medically diagnosed allergy represents a more severe underlying biology, or it may simply reflect confounding by socio-economic status as proxied by healthcare utilization. The inconsistency in exposure measurement across studies is greatly limiting in summarizing a body of evidence and one of several challenges that must be resolved in future investigations.

The primary motivation behind establishing an association (positive or negative) between allergy and ALL is to better understand the underlying immune dysregulation preceding ALL. If the results of the meta-analysis had been strongly suggestive of a consistent relationship in either direction, functional studies may have been warranted to tease out a biological mechanism. Modern causal inference methods, such as Mendelian randomization, have been developed to address issues related to exposure misclassification and uncontrolled confounding. This method could be used in the future to evaluate the causal effect of allergy on ALL by leveraging a germline genetic proxy for allergy exposure; nevertheless, this method also has its limitations related to the predictive power of germline genetics for the allergy phenotype and pleiotropic effects of those variants that are related to allergy.

Unlike allergy where the relationship remains unclear, the results from our study of the germline deletion of *APOBEC3B* and ALL risk are convincingly null. The lack of association of the deletion with ALL suggests that source of dysregulation for APOBEC3 enzymatic activity that incurs somatic mutation is not related to the germline genetic variant. It is not known under what circumstances APOBEC-mediated collateral damage to the human genome occurs during ALL pathogenesis. Infection is a likely possibility because wild type APOBEC3B enzyme is primarily active against double stranded DNA viruses and retrotransposons localized in the nucleus. Cytomegalovirus (CMV) is of key interest in exploring the relationship between APOBEC3B, infection, and ALL for two important reasons. First, our group has identified CMV as the only exogenous viral agent unique to diagnostic bone marrow samples of ALL as compared to bone marrow of similarly immune suppressed individuals (children with AML) and further demonstrated that children who go on to develop ALL have a prevalence-odds of *in utero* CMV infection 3.51 times higher than healthy controls. Secondly, CMV has a tropism for B-cells in bone marrow[1] and thus may potentially activate APOBEC-related immunity specifically in the cells that go on to become cancerous. Given this, we have an alternative hypothesis as to how the TpC>T point mutation signature arises in ALL: that that they arise in ALL cases that were congenitally infected with CMV. Specifically, our group has received funding through the UCSF Cancer Center to conduct whole genome sequencing on diagnostic bone marrow mononuclear cells from 15 ALL cases born with *in utero* CMV infection and 15 without. With these data we should be able to quantify the burden of APOBEC and RAG somatic mutation signatures in each group. Positive results would provide strong evidence that CMV initiates specific immune dysregulation that in turn initiates leukemogenesis in susceptible children.

In the final chapter of this dissertation we implemented a new method for tagging polymorphic HERV-K insertions and report no observation of existing genetic associations between HERV-K insertions with ALL. Yet through the associations with other classes of disease and biological processes we provide evidence that polymorphic HERV-K insertions may prove interesting in ALL afterall. One of the most important limitations to our study design, wherein we relied on the taggability of polymorphic HERV-K insertions to identify existing locus-phenotype associations, was the observation that the vast majority of probable polymorphic HERV-K insertions were not associated with any neighboring SNPs. Without directly genotyping these HERVs in disease contexts, the influence that they may or may not have on ALL or any disease cannot be determined. Further, our study relied on published SNP:phenotype associations, which are likely not comprehensive in capturing all genetic variation associated with diseases like ALL. Thus, it is possible that untaggable HERV-K insertions or missing data on genetic loci important to ALL contributed to our lack of findings specific to this disease.

The non-ALL phenotype associations that we did find with the 48 taggable polymorphic HERV-K insertions provide evidence that these elements could play a role in immune dysregulation and cancer. Specifically, using experimental factor ontology enrichment analysis, we found overall strong associations of polymorphic HERV-K with various responses to infection including IgG antibody response to a herpesvirus, Epstein Barr virus (FDR 0.000037). Further, HERV-K polymorphisms were associated with autoimmune diseases (FDR 1.8E-09) and virally induced neoplasms including lymphoma (FDR 6.5E-07) and hepatocellular carcinoma (FDR 0.000099). Given these observations, one could posit that having a higher burden or specific polymorphic HERV-K insertions could interact with exogenous viral infection to dysregulate the immune system and that this dysregulation can lead to autoimmune disease or cancer. Observational and experimental evidence supports this hypothesis. *In vitro* experiments provide evidence that HERV-K expression increases upon infection with herpesviruses (EBV and CMV) as well as with exogenous retroviruses like HIV[2]. As discussed in Section 3.1, HERVs have a demonstrated ability to modulate immunity through both suppressive and activating pathways. Further, HERV expression can be detected in autoimmune tissues, including cerebral spinal fluid and brain tissue from patients with multiple sclerosis, as well as in tumors from cases of lymphoma, breast cancer, prostate and ovarian cancers[3]. Thus, there is much to be done in order to understand the role of HERVs, and specifically polymorphic HERV-K insertions, in ALL and other cancers with suggestive immune dysregulation and/or infectious etiologies.

The puzzle remains for future research to disentangle the relationship between immune dysregulation and childhood ALL. Epidemiologic studies serve as an important step in understanding the causes of this disease and here we have presented three examples of epidemiologic studies wherein results have generated a set of strong hypotheses to test in the context of ALL and other diseases. With advancements in next generation molecular and computational methods as well as methods in causal inference, future observational and experimental studies will be able to definitively answer lingering questions. What is going wrong in the immune system of children who go on to develop ALL? Are the sources of immune dysregulation effecting children with leukemia exogenous, endogenous, or both? Is immune dysregulation causal of ALL, or is it simply prodrome of disease with a years-long latency? There remains little doubt that immune dysregulation precedes the onset of ALL; it is our hope as public health practitioners that the answers to these remaining questions will precipitate effective prevention strategies for this leading cause of morbidity and death in children.

4.2 References

1 Maciejewski, J. P. & St Jeor, S. C. Human cytomegalovirus infection of human hematopoietic progenitor cells. *Leuk. Lymphoma* **33**, 1-13, doi:10.3109/10428199909093720 (1999).

2 zur Hausen, H. The search for infectious causes of human cancers: where and why (Nobel lecture). *Angew. Chem. Int. Ed. Engl.* **48**, 5798-5808, doi:10.1002/anie.200901917 (2009).

3 Ryan, F. P. Human endogenous retroviruses in health and disease: a symbiotic perspective. *J. R. Soc. Med.* **97**, 560-565, doi:10.1258/jrsm.97.12.560 (2004).