Inherited Susceptibility in Childhood Leukemia among a California Hispanic Population

By

Ling-I Hsu

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 Patricia A. Buffler, Chair
Professor Lisa Barcellos
Professor George Sensabaugh

Fall 2013

Inherited Susceptibility in Childhood Leukemia among a California Hispanic Population

ABSTRACT

Inherited Susceptibility in Childhood Leukemia among a California Hispanic Population

By

Ling-I Hsu

Doctor of Philosophy in Epidemiology

University of California, Berkeley

Professor Patricia A. Buffler, Chair

The incidence of acute lymphoblastic leukemia (ALL) has been found to be nearly 20% higher among Hispanics than non-Hispanic Whites in California. Ethnic differences in ALL incidences may be attributed to the differences in the frequency of genetic factors or increased Native American ancestry. In addition to biological factors, suggestive evidence exists for other factors including agricultural pesticide usage, socioeconomic status, and timing of early exposure to infectious agents or other environmental exposures that differ among the Hispanic population. Since ALL is the most common childhood malignancy, characterizing the genetic variation and unraveling the complex interplay between genetic and environmental factors are crucial for understanding the disease etiology. Recent genome-wide association studies (GWAS) in non-Hispanic White populations have indicated that inherited genetic variations in key regulators for lymphoid differentiation contribute to childhood ALL susceptibility (*IKZF1*, *ARID5B*, and *CEBPE*). However, few studies have been explored these loci in the Hispanic population, and fewer have assessed the interplay of environment factors. This dissertation is focused on identifying and characterizing genetic components, gene and environment interactions and biological pathways among the high risk population of childhood ALL.

In Chapter 2, the relationship between eight selected single nucleotide polymorphisms (SNPs) identified in the previous GWAS: 10q21.2 (rs7089424, rs10821936, rs7073837, rs10740055, rs10994982, *ARID5B*), 14q11.2 (rs2239633, *CEBPE*), and 7p12.2 (rs4132601, rs11978267, *IKZF1*) and the risk for childhood ALL was investigated in both non-Hispanic White (NHW) and Hispanic populations of the California Childhood Leukemia Study (CCLS). Logistic regression assuming a log-additive genetic model was used to estimate odds ratios (OR) associated with each SNP within *IKZF1*, *CEBPE*, and *ARID5B* among 594 NHW children (225 cases and 369 controls) and 706 Hispanic children (300 cases and 406 controls). We found significant associations for five *ARID5B* variants in both Hispanics (P values of 1.0×10^{-9} to 0.004) and NHWs (P values of 2.2×10^{-6} to 0.018). Risk estimates were in the same direction in both groups and strengthened when restricted to B-cell hyperdiploid ALL. Similar results were observed for the *CEBPE* variant. *IKZF1* variants showed some varieties in susceptibility loci. Evidence of interaction was not observed for these eight variants and surrogates for early life

exposure to infections, such as daycare attendance, birth order and history of infections. The findings provide additional support for the role of inherited genetic susceptibility in childhood ALL and insights into ALL pathogenesis in diverse populations.

In Chapter 3, the relationship between variation within three candidate lymphoid cell development genes (*IKZF1*, *CEBPE*, and *ARID5B)* and the risk of childhood ALL was extensively examined in the Hispanic population. Genotypic data for 323 Hispanic ALL cases and 454 controls from the CCLS were generated using Illumina OmniExpress v1 platform. Statistically significant associations between genotypes at 7p12.2 (*IKZF1*), 10q21.2 (*ARID5B*), and 14q11.2 (*CEBPE*) and ALL risk are found; odds ratio (OR) =0.50, 95% confidence interval (CI): 0.35-0.71 (P value =0.004), OR=2.12, 95% CI: 1.70-2.65 (P value $=1.16 \times 10^{-9}$), OR=1.69, 95% CI: 1.37-2.08 (P value $=2.35 \times 10^{-6}$), respectively. The rs11980379 and rs4132601 risk alleles within *IKZF1* were associated with *IKZF1* expression. As shown by present study findings and previous published studies, inherited predisposition seems to be subtype-specific, suggesting different etiologies for different ALL subtypes. Potential interactions between the genetic variation and surrogates for early life exposure to infections, such as daycare attendance and birth order, on the ALL risk were not observed on a multiplicative scale. The results further identify more susceptibility loci and underscore the importance of lymphoid cell development genes on ALL pathogenesis.

Finally, in Chapter 4, pathway-based analyses were employed in Hispanic GWAS data of the CCLS to examine if different biological pathways were overrepresented in ALL and major ALL disease subtypes, including B-cell ALL, hyperdiploid B-ALL, and *TEL-AML1* ALL. For pathway analyses, genes that had at least one significantly associated SNP (P value <0.001) were selected, while adjusted for age, gender, and genetic ancestry. The top five overrepresented KEGG pathways in ALL include axon guidance ($P_{FDR}=5.1\times10^{-06}$), protein digestion and absorption ($P_{FDR}=7.2\times10^{-04}$), melanogenesis ($P_{FDR}=0.001$), leukocyte transendothelial migration ($P_{FDR}=0.002$), and focal adhesion ($P_{FDR}=0.002$). Between different disease subtypes, pathway analyses results indicate that hyperdiploid B-ALL and *TEL-AML1* ALL involve distinct biological mechanisms compared to ALL, while focal adhesion is a shared mechanism between different ALL disease subtypes. Furthermore, targeted maximum likelihood estimation (TMLE) method incorporating with least absolute shrinkage and selection operator (LASSO) were used for data reduction and to select a list of candidate genes for directing future studies, while accounting for correlation between SNPs. Several genes including *COL6A6*, *COL5A1*, *DVL1*, *TCF7L1*, *MAP2K2*, *VAV3*, *CTNNA2*, *CDK6*, *RRAS2*, and *CAMK2D* warrant future investigations. The findings suggest that pathway analyses and novel causal methods can provide additional insights into selecting regions for targeted sequencing and these enriched biological pathways can be explored as new therapeutic targets for childhood ALL.

To my wonderful Ph.D. adviser
Patricia A. Buffler

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1

Chapter 2

Chapter 3

Chapter 4

LIST OF FIGURES

Chapter 1

Chapter 3

ACKNOWLEGMENT

First and the foremost I am indebted to my outstanding faculty adviser, Professor Patricia Buffler. Without her constant support, guidance, and encouragement, the dissertation would not be possible. I am extremely honored to be able to work with her on childhood leukemia research projects. She encouraged me to not only grow as an independent thinker but also as a confident researcher. Her elegance, generosity and genuine interests in sciences forever change me. I could not have image having a better advisor and mentor for my Ph.D study. Her sudden death a few months ago devastated us. I would like to dedicate my dissertation to honor her and hope the dissertation can reflect a small part of her. I am forever in her debt.

I would like to sincerely thank Dr. Lisa Barcellos, whose introductory course in genetic epidemiology inspired me to the field. Her constant passion in genetic epidemiology and her insights into my research projects make the research possible. I would also want to thank Professor George Sensabaugh for his feedback throughout the dissertation writing process and Professor Art Reingold for providing his comments and support during this tough period.

I am thankful to Anand Chokkalingam who is always my advocate during my Ph.D study. He constantly gives me guidance both in research ideas and professional development. I also want to thank Catherine Metayer, Professor Joe Wiemels, and Professor Steve Selvin for their assistance with research concepts and data analyses.

I am grateful to my fellow students and colleagues in the California Childhood Leukemia Study and the Barcellos Genetic Epidemiology and Genomic Laboratory for their friendship and constant support- Karen Bartley, Giovanna Cruz, Jeffrey Chang, Melinda Aldrich, Kevin Urayama, Yang Wang, Julia Selezneva, Alice Kang, Hong Quach, Gary Artim, Michaela George, and Milena Gianfrancesco. I especially want to thank Farren Briggs for his continual insights and encouragement to keep me on track, and XiaoRong Shao for her statistical expertise and confidence in me. I also want to thank my fellow cohorts, whose talents always inspire me, particularly Hope Biswas, Raymond Lo, and Aracely Tamayo. Without them, I would not have completed the process.

I am thankful to all of my friends who are always by my side and teach me never give up - Yi-Wen Lin, Ya-Ting Chuang, John Liu, Yu-Hsuan Su, Joyce Kong, Laura Sue, Summer Han, Tiffany Chen, Janet Pan, I-Husan Chen, Hongyou Lu, Hung-Chia Yang, Sara Bodach, Qijuan Li, Tzu-Yi Chuang, Ellen Hsu, Ting-Huei Chen, and Angela Hsieh. The everyday challenges of graduate school would not be tolerable without their love and support.

Finally, I want to thank my parents, Chiang-Pao Hsu and Yi-Huei Lu, and my sister Chin-I Hsu for their unconditional love and unwavering support. Everything would not be possible without their faith in me. Without them, I would definitely not be who I am nor where I am today.

Chapter One
Introduction to Childhood Leukemia and Hypothesis on Childhood Leukemia Etiology

CHILDHOOD LEUKEMIA

An overview of childhood leukemia

Leukemia is the most common childhood malignancy, accounting for 31% of all cancers diagnosed in children younger than 15 years of age (1). The annual incidence rate of childhood leukemia in the United States is approximately 45 cases per million, which equates to approximately 3,250 cases of childhood leukemia diagnosed each year (1). Within the major subtypes of childhood leukemia, 79% of these cases are acute lymphoblastic leukemia (ALL), followed by 17% of acute myeloid leukemia (AML), 4% of chronic myeloid leukemia (CML), and other types (2). The etiology of the major leukemia subtypes (ALL and AML) is suspected to be different based on both cell lineage and epidemiological studies of incidence and risk factors (3).

Substantial geographic variation exists in childhood leukemia incidence rates. Incidence rates for ALL are highest in European, American (North, Central and South), and Oceania (Australia and New Zealand) countries, followed by intermediate rates in Asian countries, and have the lowest rates in African countries (4). In California, Hispanics have the highest age-adjusted childhood leukemia rates (56.2 per million), followed by non-Hispanic Whites (44.6 per million), Asian/Pacific Islanders (40.0 per million), and African Americans (29.1 per million) (5).

Epidemiologic studies of childhood leukemia have examined a number of possible risk factors, including environmental and genetic factors, to determine the etiology of the disease. However, there is still little known about the causes of and risk factors for childhood leukemia (6). From a global perspective, established evidence for an increased risk of ALL includes gender (30% higher incidence in males compared to females), age (peak incidence is between the ages of 2 and 5), race (there is an approximate1.6 times higher risk in non-Hispanic White children compared to African American children), prenatal exposure to x-rays, therapeutic radiation, high birth weight, and specific genetic syndromes (1, 7). However, these factors explain less than 10% of the cases, leaving the remaining 90% unresolved (8). Other putative risk factors include non-ionizing electromagnetic fields, parental occupation, parental smoking, exposure to pesticides or solvents, immunological factors, and genetic susceptibility (9).

Disease-free survival of childhood ALL has improved over the last several decades, reaching 80% in developed countries (10). However, treatment outcomes also differ by ethnicity: African Americans and Hispanics have lower survival rate compared to non-Hispanic Whites and Asian Americans (11). The contribution of genetic, environmental, or sociocultural factors to these observed differences in disease incidences and treatment responses remains unclear. This dissertation is focused on characterizing the genetic components and relevant environmental exposures and deciphering how these factors contribute to the multifactorial pathogenesis of childhood ALL in the Hispanic population.

Natural History and Biology

Acute lymphoblastic leukemia (ALL) is a heterogeneous disease characterized by the predominance of immature hematopoietic cells (7). Though the etiology of childhood leukemia is unclear, recent research has indicated that both genetic susceptibility and environmental exposures are likely to be involved. It is possible that both prenatal and postnatal environmental exposures play a crucial role in triggering the onset of leukemia. The process leading to the onset of childhood leukemia is likely to involve at least two genetic events, or "hits" (12). The first event probably occurs *in utero* and often causes a chromosome translocation and formation of a fusion gene. The second event occurs during the postnatal period, causing a proliferation of the leukemia clones (13).

Immunophenotyping by flow cytometry is essential in diagnosing ALL and distinguishing subtypes with therapeutic implications, including B-cell ALL and T-cell ALL. Approximately 75% of childhood B-cell ALL cases have a recurring chromosomal alteration which is detectable by karyotyping, fluorescence *in situ* hybridization (FISH), or other molecular techniques. Many leukemia patients have chromosomal alterations, including t(12;21) *TEL-AML1*, 11q23 *MLL-AF4*, t(8;21) *AML1-ETO*, t(15;17) *PML-RARA*, t(1;19) *E2A/PBX1*, inv(16) *CBFB-MYH11*, and hyperdiploidy (14-19). While these appear to be important initiating events in leukemogenesis, the alterations are insufficient to cause leukemia (20). For example, screening of neonatal cord-blood samples has revealed a putative leukemic clone with the *TEL-AML1* fusion gene in 1% of newborn babies, which is a frequency 100 times higher than the prevalence of ALL (21).

Individual subtypes of ALL may have distinct etiologies. Molecular abnormalities associated with particular subtypes may be linked to specific causal mechanisms and have been used for risk stratification and treatment specification (22-24). For example, infant ALL is usually associated with *MLL* rearrangement and has a high concordance rate in monozygotic twins (approaching 100% for those with a single placenta) (25). In contrast, B-cell ALL peaks between two and five years of age and has a relatively much lower concordance rate of 5-25% (26). Approximately one-third of B-cell ALL patients show an increase in the chromosome number (i.e., hyperdiploidy), which make up a unique biologic subset associated with increased *in vitro* apoptosis in a variety of chemotherapeutic agents (27). A favorable prognosis is observed in hyperdiploid B-ALL patients, which can be attributed to the impact of trisomies of chromosomes 4, 10, and 17 (28). Another most prevalent translocations (~25%) is the t(12;21) chromosomal translocation, which results in the chimeric fusion gene *TEL-AML1* (*ETV6-RUNX1*) (29). *TEL-AML1* patients are thought to have an excellent prognosis and are associated with good risk features such as female gender, low white cell count, and CD10+immunophenotype (29).

Up to 15% of childhood B-cell ALL patients lack a previously identified chromosomal rearrangement, but do exhibit a gene-expression profile similar to that of t (9; 22) *BCR-ABL1* positive ALL and often have deletion or mutation of *IKZF1* (30). The incidence of t (9; 22) subtype increases with age, from 2% in children to 20% in adults (20–39 years of age) (29). The presence of the t (9; 22) translocation has been associated with a poor prognosis. Due to the advent of tyrosine kinase inhibitors, the treatment outcome has improved for *BCR-ABL1* positive patients (29).

The growing body of literature indicates that cytogenetic characteristics may vary between different ethnic groups and might be associated with ALL treatment outcomes. Aldrich et al. found a similar distribution of cytogenetic characteristics among non-Hispanic Whites and Hispanics in the California Childhood Leukemia Study (CCLS), with the exception of t(12;21)

TEL-AML, which is more common in non-Hispanic Whites (31). Similarly, a study from St. Jude Children's Research Hospital observed that among African American children, there is a higher prevalence of the t(1;19) *E2A-PBX1* translocation, which is associated with incomplete remission and poor treatment outcome (32).

Heterogeneity of Hispanics in the United States

According to the United States Census Bureau, Hispanics are individuals who self-report as being Spanish, Latino or Hispanic, and can be of any race. The reported incidence rate of childhood ALL has been found to be approximately 20% higher among Hispanic children than the rate among non-Hispanic White children (5). This higher risk is possibly due to an increased prevalence of ALL risk alleles in populations with Native American ancestry, as well as ethnic differences in exposure to environmental risk factors (33-35). Hispanics are a recent admixed population, meaning the proportions of European, African, and Native American genetic ancestry can vary considerably (36). In the CCLS, the Hispanic population showed a similar mean of European, Native American, and African ancestry (~52%, ~40% and ~8%, respectively for both cases and controls) between cases and controls using ancestry informative markers (AIMs), probably due to careful individual matching among cases and controls (37).

Since Hispanics are an admixed population, they are more susceptible to population stratification, a type of confounding resulting from allele frequency differences in cases and controls, due to systematic differences in ancestry rather than association with diseases (38). Linkage disequilibrium (LD) decays quickly in randomly mating populations; however, in populations with recent admixture, LD may produce spurious associations with markers that are unlinked to disease loci. If a disease has a higher incidence in one ancestral subpopulation, then this subgroup will be overrepresented among the cases. Wacholder et al. define population stratification as "the distortion of the relationship between a genotype of interest and disease due to the effect of a true risk factor that is related to the genotype" (39). These spurious associations can also occur when cases and controls are sampled from the same admixed populations in which the proportions of ancestry vary between individuals, as is the case for the Hispanic population (40).

One way to account for population structure is to use principal components analysis (PCA) (41). PCA is a statistical method for exploring datasets with a large number of measurements by reducing the dimensions to a few principal components (PCs) that describe the pattern. The first PC is the linear combination of measurements that accounts for the largest amount of variability in the data; the second PC is the second most variable summary among all possible linear combinations (42). Price et al. propose using PCA to correct for population stratification in genetic studies (41, 43). First, PCA is applied to the genotype data to infer continuous axes of genetic variation. Afterwards, adjusted genotypes and phenotypes are calculated to compute the association statistics (The adjusted genotypes are the residuals from the regression of the original genotypes against the continuous axis of variation. The adjusted phenotypes are similarly determined) (41, 43). This method has been widely implemented in genetic studies in admixed populations (44).

Infection-related hypotheses in childhood leukemia

Current hypotheses suggest that immune function and responses to infection are likely to play a key role in ALL etiology (20). Two infection-related hypotheses have been proposed in childhood ALL. Kinlen postulates that the mixing of immunologically isolated populations with new residences may lead to an increased risk of ALL, due to changes in the population dynamics of infectious diseases (45). Thus viewed, leukemia may be a rare response to a common infection (46). Greaves hypothesizes that the absence in early childhood of an immune challenge and priming, combined with "delayed" exposures to infection, might subsequently cause adverse immune responses to common infectious agents, and thus increase the risk of childhood ALL (34). Greaves's theory parallels the hygiene hypothesis that was developed to explain an increased prevalence of allergies in developed countries: in early life, fewer infections and less exposure to bacteria and endotoxin may lead to a less well-modulated immune system and over-reactive T-helper 2 cells (47). Kinlen's hypothesis predicts a lack of immunity to common viral or other infective agents while Greaves's hypothesis postulates an abnormal and delayed response to infections. Even though these two hypotheses differ in their details, they are compatible with one another.

The literature on early infections and ALL is heterogeneous, with various definitions of common infections and various quantifying indices (48). Several epidemiologic studies have evaluated the roles of immune function and responses to infection in ALL etiology using proxies of exposure to infections, including daycare attendance, birth order, both the child's and mother's history of infections, breastfeeding, urban versus rural locations, play-group attendance, and parental social contact in the work place (49-51). The transmission of infectious agents is believed to be promoted through different types of social settings because of the immaturity of children's immune systems together with a general lack of hygienic behaviors. In developed countries, most exposures to common childhood infections result from contacts with other children (52). For example, it has been well documented that daycare attendance increases the risk for infections and the risk of infections increases with the number of socializing children (53). In 2004, McNally and Eden reviewed 10 studies that examined the association between daycare attendance and ALL, and the work of Urayama et al. (2010) update and extend the review to 14 studies (48, 54). Neither study show adverse effects of daycare attendance with the risk of ALL; moreover, the majority of studies demonstrate a protective effect. The results from the recent meta-analysis reveal a reduced risk of ALL for non-Hispanic White children who attend daycare facilities (OR= 0.76, 95% CI: 0.67–0.87) (54).

Birth order is another marker of infectious exposure, with later-born children presumed to be more often exposed to infectious agents by older siblings, as well as exposure at earlier ages. Analyses among non-Hispanic White children in the CCLS show evidence of a reduced risk of childhood ALL associated with having older siblings (55). Most of the cohort studies and the case-control studies based on birth registries report null or negative associations between birth order and ALL, whereas the results are less consistent in other studies (56). Any relationship between birth order and ALL may be diluted if the birth interval is large or if a child acquires one or more infections from other sources, such as via parental social contact.

Histories of common childhood infections can also be used as a measure of probable exposure to infections during early childhood. Among various common infections assessed in the CCLS, having ear infections during the first year of life was associated with a protective effect towards ALL (55).The results published by CCLS are consistent with two additional studies that present evidence of a reduced ALL risk in conjunction with ear infections (57, 58). Several other studies provide evidence of a reduced ALL risk associated with other types of common infections, including the common cold (58), gastrointestinal infections (59-61), and neonatal infections (58, 62). On the other hand, certain other studies report no association (63, 64), or the reverse, an increased risk of childhood ALL (65, 66), especially those studies utilizing general medical records (65-67). It is possible that these conflicting results reflect one of many mechanisms involved in the etiology of childhood ALL. Clinically diagnosed infectious illnesses may not capture the same infection experiences as self-reported infection histories (66).

Taken as a whole, the most comprehensive and consistent results support that exposure to infectious agents early in life protects against ALL are from daycare attendance literature. Nevertheless, in the absence of direct serologic evidence, the biological mechanisms behind immune responses remain unclear (65).

Genetic Susceptibility to Childhood Leukemia

Studies of leukemia among identical twins indicate a substantially higher risk for the twin of a patient to also be diagnosed with ALL (68). The estimated concordance rate for ALL in identical twins is between 5% and 25%. During the peak incidence of ALL at the age of between two and five years, the concordance rate is showed to be 15% (8). Children with affected siblings have a two- to four-fold greater risk of developing leukemia (69). Direct evidence for inherited genetic syndromes to ALL is provided by the high risk associated with Bloom's syndrome, neurofibromatosis, ataxia telangiectasia, and constitutional trisomy 21 (Down syndrome); however, these explain only a small proportion of ALL diagnoses (70). It is likely that the inheritance of multiple low-risk variants or the complex interactions between genetic variants and environmental exposures, contribute to the ALL disease risk.

Candidate-gene approaches implicate inherited polymorphisms of several genes that could play a role in leukemogenesis, but these findings are inconclusive (71). Genetic susceptibility studies of genes that encode enzymes with critical roles in xenobiotic metabolism such as *GSTM1*, *GSTT1* and *CYP1A1*, show associations with an increased risk of childhood leukemia (72). Polymorphisms in the folate metabolism and oxidative stress response genes as well as in cell cycle checkpoint genes and DNA repair genes may also increase susceptibility to ALL (73-76). Genes involved in the immune system, such as human lymphocyte antigen (*HLA*), have been associated with childhood ALL (77, 78). Additionally, genes involved in encoding proteins with key roles in lymphoid development (*PAX5*, *IKZF1*, *EBF1*, and *LMO2*), cell-cycle regulation and tumor suppression (*CDKN2A/CDKN2B*, *PTEN*, and *RB1*), lymphoid signaling (*BTLA*, *CD200*, and *TOX*), and transcriptional regulation and coactivation (*TBL1XR1*, *ETV6*, and *ERG*) are all suspected to be involved with ALL pathogenesis (30).

To date, most research on childhood ALL focuses on candidate-gene approaches and examines ALL in the non-Hispanic White population. On the other hand, genome-wide association studies (GWAS) use an agnostic approach ("hypothesis-free"), comparing single nucleotide polymorphisms (SNPs) across the whole genome in a large sample of cases and controls. These studies usually contain hundreds of thousands of markers and a genome-wide significance threshold (nominal P-value $< 5\times10^{-8}$) is required, ideally accompanied by independent replication datasets. The first GWAS in childhood ALL was published in 2009 by two different research groups, with a focus on the non-Hispanic White population (79, 80). The two studies provide the first evidence that inherited genetic variations are associated with childhood ALL, including *IKZF1* (encoding the early lymphoid transcription factor IKAROS), *ARID5B* (encoding the AT-rich interactive domain 5B transcription factor), and *CEBPE* (encoding the transcription factor CCAAT/enhancer binding protein, epsilon) (**Table 1**). Notably, these risk variants annotate genes involved in transcriptional regulation and differentiation of B-cell progenitors (30, 81). Subsequently, follow-up studies show consistent associations with the risk of childhood ALL in different populations, including a large German replication study (82), a Thai population (83), a Polish population (84), a French-Canadian cohort (85), a multiethnic population (86), and African American children from St. Jude Children's Research Hospital (87). Currently, GWAS of childhood ALL have been successfully identified as common genetic variants in *IKZF1*, *ARID5B*, *CEBPE*, *CDKN2A/2B*, *BMI1-PIP4K2*, and *GATA3* (79, 80, 85, 88, 89, 90). However, fewer studies explore these loci in Hispanic populations using a genome-wide approach (89).

One of the current hypotheses suggests that childhood leukemia results from chromosomal alterations and mutations that disrupt the normal differentiation process of lymphoid or myeloid progenitor cells (91). GWAS results indicate that genes involved in B-lymphocyte development, such as *IKZF1*, *ARID5B*, and *CEBPE*, might play a crucial role in ALL predispositions (33, 79, 80, 86, 92, 93). Animal models in knockout mice also support the biological plausibility of these genes, which are essential for B-cell development (82, 83, 85, 94, 95). Altogether, the findings from GWAS and animal studies further support the current hypothesis, indicating that inherited genetic variation in these B-cell regulatory genes might contribute to the pathogenesis of ALL and may also have important implications for ALL treatment.

Previously Identified B-lymphocyte Development Genes from GWAS

Hematopoiesis is controlled by a number of regulatory networks, including a series of specific transcription factors. Hematopoietic stem cells (HSCs) produce multi potent progenitors (MPPs) that give rise to lymphoid-primed multi potent progenitors (LMPPs) and common lymphoid progenitors (CLPs) (96). B lymphocytes are then generated from hematopoietic progenitors, which are carefully regulated by lineage-specific transcription factors (96, 97). The differentiation and proliferation of B cells require the proper regulation of transcription factors that activate specific genes and restrict the differentiation of HSCs.

One of the critical transcription factors for B-cell lineage specification is Ikaros (encoded by *IKZF1)*, which acts as a key regulator of hematopoietic differentiation (81). Ikaros is required for lymphocyte development since its deletion halts the development of B lymphocytes (98). In mice models, B cells remain absent throughout the lifetime of Ikaros-deficient mice (99). The growing body of literature has established that *IKZF1* is one of the most clinically relevant tumor suppressors in ALL (81). *IKZF1* alterations are presented in more than 70% of *BCR-ABL1* lymphoid leukemia patients and are associated with poor response to treatment (100, 101). Studies from the Children's Oncology Group (COG) leukemia study group demonstrate that patients with deletion or mutations of *IKZF1* have nearly three times the risk of treatment failure (101).

Less is known about *CEBPE* and *ARID5B*. *CEBPE*, together with other CEBP family members, have been shown to play a pivotal role in hematopoietic proliferation and differentiation (102). *CEBPE* is involved in functional maturation and terminal differentiation of myeloid cells as well as a target of chromosomal translocation in ALL (102). Deregulated *CEBPE* expression may lead to malignant transformation (102). *ARID5B* is a member of the AT-rich interaction domain family of transcription factors which are essential for embryogenesis and growth regulation (103). *ARID5B* knockout mice exhibit defects in the B lymphoid progenitors (94). Thus, it is plausible that germline variation at the *ARID5B* locus affects susceptibility to ALL by altering *ARID5B* function in B-lineage development. As demonstrated in previous study findings, the associations between *ARID5B* genotypes and ALL risk are most prominent in hyperdiploid ALL subtypes, implying that ALL predispositions might be subtype specific and different subtypes involve different etiologies (80, 104).

Pathway-based Analysis

The rationale behind the GWAS design is that common variation in the human genome, as exhibited by SNPs with frequency greater than 5%, is responsible for the risk of complex disorders (105). The study design of GWAS is based on being able to statistically detect the association of a SNP that is in linkage disequilibrium with a predisposing gene variant. Even though several common genetic variants in *IKZF1*, *ARID5B*, *CEBPE*, *CDKN2A/2B*, *BMI1-PIP4K2*, and *GATA3* have been identified through GWAS, these variants only have odds ratios of around 1.2 to 1.5 and do little to explain leukemia risk (86, 106). It has been argued that genetic variants contributing to disease susceptibility are not being captured by current GWAS paradigms (107). These undetected associations are important because they may be effective in elucidating the biologic basis of diseases and leading to possible treatments. Recently, complementary approaches for GWAS analysis have been developed, including the use of imputation for association tests, and pathway-based association approaches (108). Genotype imputation predicts untyped markers in target samples using a densely typed reference set (i.e., 1000 Genome Project). Imputation allows meta-analysis of studies genotyped on different genotyping platforms and allows association testing of variants that are not well captured by the chips. Pathway analysis is the use of prior biological knowledge underlying complex diseases to integrate into association analyses (109). In particular, analyzing genomic data through gene sets that are defined by functional pathways offer greater potential for discovery and to find the connections of disease mechanisms.

Several recently published studies have clearly demonstrated the use and importance of pathway-based approaches, which complement standard single-marker analysis in extracting more biological information from existing GWAS datasets (110). Complex diseases may result from the accumulation of the effects of genetic variants within pathways. Pathway analyses provide a way of integrating the results of the GWAS and the genes into a known molecular pathway to test whether the pathway is overrepresented in the disease. The analyses address two important elements of post-GWAS prioritization: the selected pathway provides a biological foundation for statistically combining the GWAS association results and offers one biological interpretation as well as possible clinical relevance. Several pathway classification methods are available, including the BioCarta pathway database (111), the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (112), and the Gene Ontology (GO) database (113). Both KEGG and BioCarta contain manually curated pathways in different biological processes, whereas the GO database contains mostly electronic annotations for human genes, and provides an ontology of defined terms representing gene properties. The strength of the pathway-based approach has been illustrated in several GWAS focused on common diseases, including breast cancer, Parkinson's disease, Crohn's disease, and bipolar disorder (114-117).

Gene and Environmental Interactions

Many complex traits are believed to be the result of the combined effects of genes, environmental factors, and their interactions (118). Understanding the relationships between genetic variants and environmental exposures can help identify high-risk subgroups in the population and provide better insights into mechanisms of the disease. Fetal life and early childhood appear to be critical developmental stages that are more sensitive to environmental stressors; unexplained disease risk in ALL may be partly due to gene-environment interaction.

In childhood ALL, the fact that there is a greater risk in children than in adults has been linked to the developmental immaturity of a child's immune system and differential exposure to environmental toxins. Variants in the genes that are responsible for important biological functions, such as xenobiotic, folate or immune pathways, may affect a child's response to environmental exposures, and thus influence the risk of ALL (71). In CCLS, several studies have been published to examine the complex interactions between genetic variants and proposed environmental factors in the risk of childhood ALL, including *MDR1* variants and indoor insecticide exposure, DNA repair genes and X-ray exposure, and adaptive immunity genes and early life infections (119-121).

THE CALIFORNIA CHILDHOOD LEUKEMIA STUDY (CCLS)

The CCLS is an ongoing population-based case control study which began in 1995. Incident cases of newly diagnosed childhood leukemia (age 0–14 years) were rapidly gathered from major clinical centers in the study area, usually within 72 hours of diagnosis. Cases were initially identified in four hospitals and later expanded to nine hospitals in the San Francisco Bay Area and Central Valley. In 2010, there were a total of 35 counties in Northern and Central California participating in the study (**Figure 1**). Out of 1,172 eligible patients, 997 (85%) gave their consent to participate and have completed interviews. For each case, one or two healthy controls were randomly selected from the birth certificates maintained by the Center for Health Statistics of the California Department of Public Health, matching on child's age, sex, Hispanic status (a child was considered Hispanic if either parent self-reported as Hispanic), and maternal race (White, Black, Asian/Pacific Islander, Native American, and Other/Mixed). The detailed control selection process was described elsewhere (122). Among the control subjects who were contacted and considered eligible, 86% participated. Thurs, there were a total of 1,226 controls selected at random by their California birth certificates, making the CCLS one of the largest case-control studies on childhood leukemia.

Comparisons with population-based surveillance data from the California Cancer Registry (1997-2003) revealed that the 703 cases enrolled in the study between 1997 to 2003, which represented approximately 75% of diagnosed childhood leukemia cases among residents of the study counties (*n*=948). Ma et al. (2004) compared birth certificate control subjects with 'ideal' control subjects (population-based controls from the general birth certificate pool who would have been obtained under optimal circumstances), and found little difference in demographic characteristics between the birth certificate control sample and their "ideal" control sample, suggesting that the CCLS is an approximately population-based study (122).

Cases and controls were eligible to enter the study if they were under 15 years of age, lived in the study area at the time of diagnosis, had at least one parent who spoke either English or Spanish, and had no prior malignancy history. Due to the unique demographic composition of the study area, approximately 42% of study population is Hispanic, 41% is non-Hispanic White, and 17% are in another race or ethnic groups. This provides a unique opportunity to explore the etiology of ALL in the Hispanic population.

Participants were classified as Hispanic if parents responded "yes" to at least one of the following questions during an in-person, home interview: 1) "Is child Spanish/Hispanic/Latino?" and 2) "Is child Mexican, Mexican-American or Chicano?" The race of the child was categorized as: 1) White, 2) Black or African American, 3) Native American, 4) Asian or Pacific Islander, or 5) Other. A child was considered non-Hispanic White if both parents reported as being non-Hispanic ethnicity and White race. Children of parents reporting different races were placed in the category of Mixed/Other. Any child with a parent self-reporting Hispanic ethnicity was considered Hispanic.

Data Collection

Questionnaires were designed to obtain epidemiologic data that related to ALL risk and captured the ethnically diverse California population, particularly focusing on the timing and types of exposures of interest. Data were primarily collected using a personal interview conducted at the home of the respondent in either English or Spanish. Detailed information on childhood infections, child's vaccination history, breastfeeding history, chemical exposures, maternal infection history, parental smoking history and maternal dietary history were collected.

Immunophenotype was determined for ALL cases by using flow cytometry profiles. Those that were positive for CD19 or CD10 (≥20%) were classified as B-lineage and those expressing CD2, CD3, CD4, CD5, CD7, or CD8 (≥20%) were classified as T-lineage (123). When extra copies of chromosomes 21 and X were identified by FISH assays, assignment of hyperdiploid status (51-67 chromosomes) was made (31). *TEL-AML1* translocations were identified by the fusion of the *TEL* and *AML1* loci.

Biological Specimen Collection

The CCLS has obtained buccal cytobrush specimens from 98% of interviewed cases, controls and their biological mothers. Other biospecimens collected from CCLS study participants include: 1) archived newborn blood (ANB) specimens obtained from the California Department of Public Health's Genetic Disease Labortary (~90% of subjects) and 2) diagnostic pre-treatment bone marrow and peripheral blood from enrolled cases. Buccal cytobrush DNA specimens served as the primary DNA source for both cases and controls in the genetic studies. If buccal cell DNA was unavailable, ANB specimens served as a secondary DNA source, and DNA collected from other sources served as additional backups.

RESEARCH HYPOTHESES ADDRESSED IN THIS DISSERTATION

Hypothesis 1 (Chapter 2): Selected single nucleotide polymorphisms (SNPs) identified in the previous GWAS has the same effect on childhood ALL risk in the Hispanic population of the CCLS. Furthermore, the effect of genetic susceptibility on childhood ALL risk is modified by proxies of early life exposure to infections.

Hypothesis 2 (Chapter 3): Common genetic variations in normal lymphoid development genes (*IKZF1*, *CEBPE*, and *ARID5B*) are associated with increased risk of childhood ALL in the Hispanic population and the risk of childhood ALL associated with specific alleles is modified by proxies of early life exposure to infections.

Hypothesis 3 (Chapter 4): Particular biological pathways, which are essential for cell growth or cell differentiation, are overrepresented in childhood ALL. Different biological pathways are overrepresented in major disease subtypes, including B-cell ALL, hyperdiploid B-ALL and *TEL-AML1* ALL.

KNOWLEDGE TO BE GAINED

Current GWAS for childhood ALL have been limited to the non-Hispanic White population, whereas genetic variants may be specific to certain ethnic groups and allele frequencies can differ among ethnic groups (124). Understanding the complete spectrum of genetic susceptibility to childhood ALL require studies conducted on diverse populations, especially for Hispanics who appear to have the highest risk of childhood ALL.

ALL is a heterogeneous disease and is thought to have originated from accumulation of various critical genetic lesions in progenitor cells that are committed to differentiate in the B-cell pathway (2). There is a need for a comprehensive assessment of the genes that play important roles in regulating lymphocyte differentiation and which have been linked to ALL susceptibility in previous GWAS: *IKZF1* (encoding the early lymphoid transcription factor IKAROS), *CEBPE* (encoding the transcription factor CCAAT/enhancer-binding protein, epsilon), and *ARID5B* (encoding AT rich interactive domain 5B) in the Hispanic population. These studies help elucidate similarities and differences in genetic structures among different populations as well as identify specific genome regions for disease-predisposing variants in the high-risk population. Furthermore, examination between genetic variants within the three candidate genes, together with surrogates for infectious exposures, can provide additional information to show underlying immune-related hypotheses of childhood leukemia. Lastly, we further hypothesize that genetic predisposition to ALL might be mediated by multiple genes responsible for same biological pathway. Pathway-based analyses of Hispanic GWAS data can prioritize biological pathways, and possibly identify molecular targeted therapies for ALL treatment, in addition to providing complementary information of GWAS results. Taken together, these studies will increase our understanding of ALL pathogenesis and hopefully will contribute to improve future disease prognoses and treatment outcomes.

REFERENCES

1. Ries LAG SM, Gurney JG. Cancer Incidence and Survival among Children and Adolescents: United States SEER Program 1975-1995. Bethesda, MD: National Cancer Institute, SEER Program. NIH Pub. 1999:No. 99-4649.

2. Pui CH. Childhood Leukemia: Cambridge University Press; 2006.

3. Puumala SE, Ross JA, Aplenc R, Spector LG. Epidemiology of childhood acute myeloid leukemia. Pediatr Blood Cancer 2013;60(5):728-33.

4. Parkin DM KE, Draper GJ. International incidence of Childhood Cancer, Vol II. Lyon, France: International Agency for Research on Cancer. Nat Rev Genet 1998.

5. Campleman SL WW. Childhood cancer in California 1988 to 1999 Volume I: birth to age 14. Sacramento, CA: California Department of Health Services, Cancer Surveillance Section. 2004:16-17.

6. Belson M, Kingsley B, Holmes A. Risk factors for acute leukemia in children: a review. Environ Health Perspect 2007;115(1):138-45.

7. Wiemels J. Perspectives on the causes of childhood leukemia. Chem Biol Interact 2012;196(3):59-67.

8. Eden T. Aetiology of childhood leukaemia. Cancer Treat Rev 2010;36(4):286-97.

9. Buffler PA, Kwan ML, Reynolds P, Urayama KY. Environmental and genetic risk factors for childhood leukemia: appraising the evidence. Cancer Invest 2005;23(1):60-75.

10. Pui CH, Sandlund JT, Pei D, Rivera GK, Howard SC, Ribeiro RC, et al. Results of therapy for acute lymphoblastic leukemia in black and white children. JAMA 2003;290(15):2001-7.

11. Bhatia S, Sather HN, Heerema NA, Trigg ME, Gaynon PS, Robison LL. Racial and ethnic differences in survival of children with acute lymphoblastic leukemia. Blood 2002;100(6):1957-64.

12. Greaves MF. Speculations on the cause of childhood acute lymphoblastic leukemia. Leukemia 1988;2(2):120-5.

13. Greaves M. Molecular genetics, natural history and the demise of childhood leukaemia. Eur J Cancer 1999;35(14):1941-53.

14. Gale KB, Ford AM, Repp R, Borkhardt A, Keller C, Eden OB, et al. Backtracking leukemia to birth: Identification of clonotypic gene fusion sequences in neonatal blood spots. Proceedings of the National Academy of Sciences 1997;94:13950-13954.

15. Greaves MF, Wiemels J. Origins of chromosome translocations in childhood leukaemia. Nat Rev Cancer 2003;3(9):639-49.

16. McHale CM, Wiemels JL, Zhang L, Ma X, Buffler PA, Feusner J, et al. Prenatal origin of childhood acute myeloid leukemias harboring chromosomal rearrangements t(15;17) and inv(16). Blood 2003;101(11):4640-1.

17. McHale CM, Wiemels JL, Zhang L, Ma X, Buffler PA, Guo W, et al. Prenatal origin of TEL-AML1-positive acute lymphoblastic leukemia in children born in California. Genes Chromosomes Cancer 2003;37(1):36-43.

18. Wiemels JL, Cazzaniga G, Daniotti M, Eden OB, Addison GM, Masera G, et al. Prenatal Origin of Acute Lymphoblastic Leukaemia in Children. Lancet 1999;354:1499-1503.

19. Wiemels JL, Xiao Z, Buffler PA, Maia AT, Ma X, Dicks BM, et al. In utero origin of t(8;21) AML1-ETO translocations in childhood acute myeloid leukemia. Blood 2002;99(10):3801-5.

20. Greaves M. Molecular Genetics, Natural History and the Demise of Childhood Leukaemia. European Journal of Cancer 1999;35(2):173-185.
21. Mori H, Colman SM, Xiao Z, Ford AM, Healy LE, Donaldson C, et al. Chromosome translocations and covert leukemic clones are generated during normal fetal development. Proc Natl Acad Sci U S A 2002;99(12):8242-7.
22. Maia AT, van der Velden VH, Harrison CJ, Szczepanski T, Williams MD, Griffiths MJ, et al. Prenatal origin of hyperdiploid acute lymphoblastic leukemia in identical twins. Leukemia 2003;17(11):2202-6.
23. Panzer-Grumayer ER, Fasching K, Panzer S, Hettinger K, Schmitt K, Stockler-Ipsiroglu S, et al. Nondisjunction of chromosomes leading to hyperdiploid childhood B-cell precursor acute lymphoblastic leukemia is an early event during leukemogenesis. Blood 2002;100(1):347-9.
24. Taub JW, Konrad MA, Ge Y, Naber JM, Scott JS, Matherly LH, et al. High frequency of leukemic clones in newborn screening blood samples of children with B-precursor acute lymphoblastic leukemia. Blood 2002;99(8):2992-6.
25. Greaves MF. Aetiology of acute leukaemia. Lancet 1997;349(9048):344-9.
26. Inaba H, Greaves M, Mullighan CG. Acute lymphoblastic leukaemia. Lancet 2013;381(9881):1943-55.
27. Trueworthy R, Shuster J, Look T, Crist W, Borowitz M, Carroll A, et al. Ploidy of lymphoblasts is the strongest predictor of treatment outcome in B-progenitor cell acute lymphoblastic leukemia of childhood: a Pediatric Oncology Group study. J Clin Oncol 1992;10(4):606-13.
28. Carroll WL, Bhojwani D, Min DJ, Raetz E, Relling M, Davies S, et al. Pediatric acute lymphoblastic leukemia. Hematology Am Soc Hematol Educ Program 2003:102-31.
29. Moorman AV. The clinical relevance of chromosomal and genomic abnormalities in B-cell precursor acute lymphoblastic leukaemia. Blood Rev 2012;26(3):123-35.
30. Mullighan CG. The molecular genetic makeup of acute lymphoblastic leukemia. Hematology Am Soc Hematol Educ Program 2012;2012:389-96.
31. Aldrich MC, Zhang L, Wiemels JL, Ma X, Loh ML, Metayer C, et al. Cytogenetics of Hispanic and White children with acute lymphoblastic leukemia in California. Cancer Epidemiol Biomarkers Prev 2006;15(3):578-81.
32. Pollock BH, DeBaun MR, Camitta BM, Shuster JJ, Ravindranath Y, Pullen DJ, et al. Racial differences in the survival of childhood B-precursor acute lymphoblastic leukemia: a Pediatric Oncology Group Study. J Clin Oncol 2000;18(4):813-23.
33. Walsh KM, Chokkalingam AP, Hsu LI, Metayer C, de Smith AJ, Jacobs DI, et al. Associations between genome-wide Native American ancestry, known risk alleles and B-cell ALL risk in Hispanic children. Leukemia 2013.
34. Greaves M. Infection, immune responses and the aetiology of childhood leukaemia. Nat Rev Cancer 2006;6(3):193-203.
35. Yang JJ, Cheng C, Devidas M, Cao X, Fan Y, Campana D, et al. Ancestry and pharmacogenomics of relapse in acute lymphoblastic leukemia. Nat Genet 2011;43(3):237-41.
36. Price AL, Patterson N, Yu F, Cox DR, Waliszewska A, McDonald GJ, et al. A genomewide admixture map for Latino populations. Am J Hum Genet 2007;80(6):1024-36.
37. Chokkalingam AP AM, Bartley K, Hsu LI, Metayer C, et al. Matching on Race and Ethnicity in Case-Control Studies as a Means of Control for Population Stratification. Epidemiol 2011;1:101.

38. Choudhry S, Coyle NE, Tang H, Salari K, Lind D, Clark SL, et al. Population stratification confounds genetic association studies among Latinos. Hum Genet 2006;118(5):652-64.
39. Wacholder S, Rothman N, Caporaso N. Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. Cancer Epidemiol Biomarkers Prev 2002;11(6):513-20.
40. Tian C, Gregersen PK, Seldin MF. Accounting for ancestry: population substructure and genome-wide association studies. Hum Mol Genet 2008;17(R2):R143-50.
41. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38(8):904-9.
42. Reich D, Price AL, Patterson N. Principal component analysis of genetic data. Nat Genet 2008;40(5):491-2.
43. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet 2006;2(12):e190.
44. Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W, Mahoney MW, et al. PCA-correlated SNPs for structure identification in worldwide human populations. PLoS Genet 2007;3(9):1672-86.
45. Kinlen L. Evidence for an infective cause of childhood leukaemia: comparison of a Scottish new town with nuclear reprocessing sites in Britain. Lancet 1988;2(8624):1323-7.
46. Kinlen LJ. Infection and childhood leukaemia near nuclear sites. Lancet 1997;349(9066):1702.
47. Strachan DP. Family size, infection and atopy: the first decade of the "hygiene hypothesis". Thorax 2000;55 Suppl 1:S2-10.
48. McNally RJ, Eden TO. An infectious aetiology for childhood acute leukaemia: a review of the evidence. Br J Haematol 2004;127(3):243-63.
49. Urayama KY, Ma X, Buffler PA. Exposure to infections through day-care attendance and risk of childhood leukaemia. Radiat Prot Dosimetry 2008;132(2):259-66.
50. Chang JS, Metayer C, Fear NT, Reinier K, Yin X, Urayama K, et al. Parental social contact in the work place and the risk of childhood acute lymphoblastic leukaemia. Br J Cancer 2007;97(9):1315-21.
51. Roman E, Simpson J, Ansell P, Lightfoot T, Smith A. Infectious proxies and childhood leukaemia: findings from the United Kingdom Childhood Cancer Study (UKCCS). Blood Cells Mol Dis 2009;42(2):126-8.
52. Rosenbaum PF, Buck GM, Brecher ML. Early child-care and preschool experiences and the risk of childhood acute lymphoblastic leukemia. Am J Epidemiol 2000;152(12):1136-44.
53. Nystad W, Skrondal A, Magnus P. Day care attendance, recurrent respiratory tract infections and asthma. Int J Epidemiol 1999;28(5):882-7.
54. Urayama KY, Buffler PA, Gallagher ER, Ayoob JM, Ma X. A meta-analysis of the association between day-care attendance and childhood acute lymphoblastic leukaemia. Int J Epidemiol 2010;39(3):718-32.
55. Urayama KY, Ma X, Selvin S, Metayer C, Chokkalingam AP, Wiemels JL, et al. Early life exposure to infections and risk of childhood acute lymphoblastic leukemia. Int J Cancer 2011;128(7):1632-43.

56. Rudant J, Orsi L, Menegaux F, Petit A, Baruchel A, Bertrand Y, et al. Childhood acute leukemia, early common infections, and allergy: The ESCALE Study. Am J Epidemiol 2010;172(9):1015-27.
57. Neglia JP, Linet MS, Shu XO, Severson RK, Potter JD, Mertens AC, et al. Patterns of infection and day care utilization and risk of childhood acute lymphoblastic leukaemia. Br J Cancer 2000;82(1):234-40.
58. van Steensel-Moll HA, Valkenburg HA, van Zanen GE. Childhood leukemia and infectious diseases in the first year of life: a register-based case-control study. Am J Epidemiol 1986;124(4):590-4.
59. Dockerty JD, Skegg DC, Elwood JM, Herbison GP, Becroft DM, Lewis ME. Infections, vaccinations, and the risk of childhood leukaemia. Br J Cancer 1999;80(9):1483-9.
60. Jourdan-Da Silva N, Perel Y, Mechinaud F, Plouvier E, Gandemer V, Lutz P, et al. Infectious diseases in the first year of life, perinatal characteristics and childhood acute leukaemia. Br J Cancer 2004;90(1):139-45.
61. Rosenbaum PF, Buck GM, Brecher ML. Allergy and infectious disease histories and the risk of childhood acute lymphoblastic leukaemia. Paediatr Perinat Epidemiol 2005;19(2):152-64.
62. Perrillat F, Clavel J, Auclerc MF, Baruchel A, Leverger G, Nelken B, et al. Day-care, early common infections and childhood acute leukaemia: a multicentre French case-control study. Br J Cancer 2002;86(7):1064-9.
63. MacArthur AC, McBride ML, Spinelli JJ, Tamaro S, Gallagher RP, Theriault GP. Risk of childhood leukemia associated with vaccination, infection, and medication use in childhood: the Cross-Canada Childhood Leukemia Study. Am J Epidemiol 2008;167(5):598-606.
64. Schuz J KU, Meinert R, Kaatsch P, Michaelis J. Association of childhood leukaemia with factors related to the immune system. British Journal of Cancer 1999;80(3-4):585-590.
65. Roman E, Simpson J, Ansell P, Kinsey S, Mitchell CD, McKinney PA, 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 2007;165(5):496-504.
66. Chang JS, Tsai CR, Tsai YW, Wiemels JL. Medically diagnosed infections and risk of childhood leukaemia: a population-based case-control study. Int J Epidemiol 2012;41(4):1050-9.
67. Cardwell CR, McKinney PA, Patterson CC, Murray LJ. Infections in early life and childhood leukaemia risk: a UK case-control study of general practitioner records. Br J Cancer 2008;99(9):1529-33.
68. Greaves MF, Maia AT, Wiemels JL, Ford AM. Leukemia in twins: lessons in natural history. Blood 2003;102(7):2321-33.
69. Hemminki K, Jiang Y. Risks among siblings and twins for childhood acute lymphoid leukaemia: results from the Swedish Family-Cancer Database. Leukemia 2002;16(2):297-8.
70. Hodgson SM, E, editor. A Practical Guide to Human Cancer Genetics. Cambridge: Cambridge University Press; 2007.
71. Urayama KY, Chokkalingam AP, Manabe A, Mizutani S. Current evidence for an inherited genetic basis of childhood acute lymphoblastic leukemia. Int J Hematol 2013;97(1):3-19.
72. Stanulla M, Schrappe M, Brechlin AM, Zimmermann M, Welte K. Polymorphisms within glutathione S-transferase genes (GSTM1, GSTT1, GSTP1) and risk of relapse in childhood B-cell precursor acute lymphoblastic leukemia: a case-control study. Blood 2000;95(4):1222-8.

73. Schnakenberg E, Mehles A, Cario G, Rehe K, Seidemann K, Schlegelberger B, et al. Polymorphisms of methylenetetrahydrofolate reductase (MTHFR) and susceptibility to pediatric acute lymphoblastic leukemia in a German study population. BMC Med Genet 2005;6:23.
74. Healy J, Belanger H, Beaulieu P, Lariviere M, Labuda D, Sinnett D. Promoter SNPs in G1/S checkpoint regulators and their impact on the susceptibility to childhood leukemia. Blood 2007;109(2):683-92.
75. Wiemels JL, Smith RN, Taylor GM, Eden OB, Alexander FE, Greaves MF. Methylenetetrahydrofolate reductase (MTHFR) polymorphisms and risk of molecularly defined subtypes of childhood acute leukemia. Proc Natl Acad Sci U S A 2001;98(7):4004-9.
76. Guha N, Chang JS, Chokkalingam AP, Wiemels JL, Smith MT, Buffler PA. NQO1 polymorphisms and de novo childhood leukemia: a HuGE review and meta-analysis. Am J Epidemiol 2008;168(11):1221-32.
77. Taylor GM, Dearden S, Ravetto P, Ayres M, Watson P, Hussain A, et al. Genetic susceptibility to childhood common acute lymphoblastic leukaemia is associated with polymorphic peptide-binding pocket profiles in HLA-DPB1*0201. Hum Mol Genet 2002;11(14):1585-97.
78. Urayama KY, Chokkalingam AP, Metayer C, Ma X, Selvin S, Barcellos LF, et al. HLA-DP genetic variation, proxies for early life immune modulation and childhood acute lymphoblastic leukemia risk. Blood 2012;120(15):3039-47.
79. Papaemmanuil E, Hosking FJ, Vijayakrishnan J, Price A, Olver B, Sheridan E, et al. Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. Nat Genet 2009;41(9):1006-10.
80. Trevino LR, Yang W, French D, Hunger SP, Carroll WL, Devidas M, et al. Germline genomic variants associated with childhood acute lymphoblastic leukemia. Nat Genet 2009;41(9):1001-5.
81. Payne KJ, Dovat S. Ikaros and tumor suppression in acute lymphoblastic leukemia. Crit Rev Oncog 2011;16(1-2):3-12.
82. Prasad RB, Hosking FJ, Vijayakrishnan J, Papaemmanuil E, Koehler R, Greaves M, et al. Verification of the susceptibility loci on 7p12.2, 10q21.2, and 14q11.2 in precursor B-cell acute lymphoblastic leukemia of childhood. Blood 2010;115(9):1765-7.
83. Vijayakrishnan J, Sherborne AL, Sawangpanich R, Hongeng S, Houlston RS, Pakakasama S. Variation at 7p12.2 and 10q21.2 influences childhood acute lymphoblastic leukemia risk in the Thai population and may contribute to racial differences in leukemia incidence. Leuk Lymphoma 2010;51(10):1870-4.
84. Pastorczak A, Gorniak P, Sherborne A, Hosking F, Trelinska J, Lejman M, et al. Role of 657del5 NBN mutation and 7p12.2 (IKZF1), 9p21 (CDKN2A), 10q21.2 (ARID5B) and 14q11.2 (CEBPE) variation and risk of childhood ALL in the Polish population. Leuk Res 2011;35(11):1534-6.
85. Healy J, Richer C, Bourgey M, Kritikou EA, Sinnett D. Replication analysis confirms the association of ARID5B with childhood B-cell acute lymphoblastic leukemia. Haematologica 2010;95(9):1608-11.
86. Xu H, Yang W, Perez-Andreu V, Devidas M, Fan Y, Cheng C, et al. Novel susceptibility variants at 10p12.31-12.2 for childhood acute lymphoblastic leukemia in ethnically diverse populations. J Natl Cancer Inst 2013;105(10):733-42.

87. Yang W, Trevino LR, Yang JJ, Scheet P, Pui CH, Evans WE, et al. ARID5B SNP rs10821936 is associated with risk of childhood acute lymphoblastic leukemia in blacks and contributes to racial differences in leukemia incidence. Leukemia 2010;24(4):894-6.

88. Migliorini G, Fiege B, Hosking FJ, Ma Y, Kumar R, Sherborne AL, et al. Variation at 10p12.2 and 10p14 influences risk of childhood B-cell acute lymphoblastic leukemia and phenotype. Blood 2013.

89. Xu H, Yang W, Perez-Andreu V, Devidas M, Fan Y, Cheng C, et al. Novel Susceptibility Variants at 10p12.31-12.2 for Childhood Acute Lymphoblastic Leukemia in Ethnically Diverse Populations. J Natl Cancer Inst 2013.

90. Walsh KM, de Smith AJ, Chokkalingam AP, Metayer C, Dahl GV, Hsu LI, et al. Novel childhood ALL susceptibility locus BMI1-PIP4K2A is specifically associated with the hyperdiploid subtype. Blood 2013;121(23):4808-9.

91. Pui CH, Robison LL, Look AT. Acute lymphoblastic leukaemia. Lancet 2008;371(9617):1030-43.

92. Walsh KM, Cooper MA, Holle R, Rakov VA, Roeder WP, Ryan M. National Athletic Trainers' Association position statement: lightning safety for athletics and recreation. J Athl Train 2013;48(2):258-70.

93. Sherborne AL, Hosking FJ, Prasad RB, Kumar R, Koehler R, Vijayakrishnan J, et al. Variation in CDKN2A at 9p21.3 influences childhood acute lymphoblastic leukemia risk. Nat Genet 2010;42(6):492-4.

94. Wilsker D, Patsialou A, Dallas PB, Moran E. ARID proteins: a diverse family of DNA binding proteins implicated in the control of cell growth, differentiation, and development. Cell Growth Differ 2002;13(3):95-106.

95. Cortes M, Wong E, Koipally J, Georgopoulos K. Control of lymphocyte development by the Ikaros gene family. Curr Opin Immunol 1999;11(2):167-71.

96. Nutt SL, Kee BL. The transcriptional regulation of B cell lineage commitment. Immunity 2007;26(6):715-25.

97. Ramirez J, Lukin K, Hagman J. From hematopoietic progenitors to B cells: mechanisms of lineage restriction and commitment. Curr Opin Immunol 2010;22(2):177-84.

98. John LB, Ward AC. The Ikaros gene family: transcriptional regulators of hematopoiesis and immunity. Mol Immunol 2011;48(9-10):1272-8.

99. Wang JH, Nichogiannopoulou A, Wu L, Sun L, Sharpe AH, Bigby M, et al. Selective defects in the development of the fetal and adult lymphoid system in mice with an Ikaros null mutation. Immunity 1996;5(6):537-49.

100. Mullighan CG, Miller CB, Radtke I, Phillips LA, Dalton J, Ma J, et al. BCR-ABL1 lymphoblastic leukaemia is characterized by the deletion of Ikaros. Nature 2008;453(7191):110-4.

101. Mullighan CG, Su X, Zhang J, Radtke I, Phillips LA, Miller CB, et al. Deletion of IKZF1 and prognosis in acute lymphoblastic leukemia. N Engl J Med 2009;360(5):470-80.

102. Akasaka T, Balasas T, Russell LJ, Sugimoto KJ, Majid A, Walewska R, et al. Five members of the CEBP transcription factor family are targeted by recurrent IGH translocations in B-cell precursor acute lymphoblastic leukemia (BCP-ALL). Blood 2007;109(8):3451-61.

103. Patsialou A, Wilsker D, Moran E. DNA-binding properties of ARID family proteins. Nucleic Acids Res 2005;33(1):66-80.

104. Xu H, Cheng C, Devidas M, Pei D, Fan Y, Yang W, et al. ARID5B genetic polymorphisms contribute to racial disparities in the incidence and treatment outcome of childhood acute lymphoblastic leukemia. J Clin Oncol 2012;30(7):751-7.
105. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 2008;9(5):356-69.
106. Pui CH, Carroll WL, Meshinchi S, Arceci RJ. Biology, risk stratification, and therapy of pediatric acute leukemias: an update. J Clin Oncol 2011;29(5):551-65.
107. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature 2009;461(7265):747-53.
108. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet 2010;11(7):499-511.
109. Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 2005;37(4):413-7.
110. Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. Nat Rev Genet 2010;11(12):843-54.
111. Nishimura D. BioCarta. Biotech Software & Internet Report 2001;2(3):117-120.
112. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28(1):27-30.
113. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, et al. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 2004;32(Database issue):D258-61.
114. Menashe I, Maeder D, Garcia-Closas M, Figueroa JD, Bhattacharjee S, Rotunno M, et al. Pathway analysis of breast cancer genome-wide association study highlights three pathways and one canonical signaling cascade. Cancer Res 2010;70(11):4453-9.
115. Torkamani A, Topol EJ, Schork NJ. Pathway analysis of seven common diseases assessed by genome-wide association. Genomics 2008;92(5):265-72.
116. Holmans P, Moskvina V, Jones L, Sharma M, Vedernikov A, Buchel F, et al. A pathway-based analysis provides additional support for an immune-related genetic susceptibility to Parkinson's disease. Hum Mol Genet 2013;22(5):1039-49.
117. Li D, Duell EJ, Yu K, Risch HA, Olson SH, Kooperberg C, et al. Pathway analysis of genome-wide association study data highlights pancreatic development genes as susceptibility factors for pancreatic cancer. Carcinogenesis 2012;33(7):1384-90.
118. Thomas D. Gene--environment-wide association studies: emerging approaches. Nat Rev Genet 2010;11(4):259-72.
119. Chokkalingam AP, Bartley K, Wiemels JL, Metayer C, Barcellos LF, Hansen HM, et al. Haplotypes of DNA repair and cell cycle control genes, X-ray exposure, and risk of childhood acute lymphoblastic leukemia. Cancer Causes Control 2011;22(12):1721-30.
120. Chang JS, Wiemels JL, Chokkalingam AP, Metayer C, Barcellos LF, Hansen HM, et al. Genetic polymorphisms in adaptive immunity genes and childhood acute lymphoblastic leukemia. Cancer Epidemiol Biomarkers Prev 2010;19(9):2152-63.
121. Urayama KY, Wiencke JK, Buffler PA, Chokkalingam AP, Metayer C, Wiemels JL. MDR1 gene variants, indoor insecticide exposure, and the risk of childhood acute lymphoblastic leukemia. Cancer Epidemiol Biomarkers Prev 2007;16(6):1172-7.
122. Ma X, Buffler PA, Layefsky M, Does MB, Reynolds P. Control selection strategies in case-control studies of childhood diseases. Am J Epidemiol 2004;159(10):915-21.

123. Hrusak O, Trka J, Zuna J, Polouckova A, Kalina T, Stary J. Acute lymphoblastic leukemia incidence during socioeconomic transition: selective increase in children from 1 to 4 years. Leukemia 2002;16(4):720-5.

124. Rosenberg NA, Huang L, Jewett EM, Szpiech ZA, Jankovic I, Boehnke M. Genome-wide association studies in diverse populations. Nat Rev Genet 2010;11(5):356-66.

125. Pui CH, Relling MV, Downing JR. Acute lymphoblastic leukemia. N Engl J Med 2004;350(15):1535-48.

126. Georgopoulos K, Bigby M, Wang JH, Molnar A, Wu P, Winandy S, et al. The Ikaros gene is required for the development of all lymphoid lineages. Cell 1994;79(1):143-56.

127. Lahoud MH, Ristevski S, Venter DJ, Jermiin LS, Bertoncello I, Zavarsek S, et al. Gene targeting of Desrt, a novel ARID class DNA-binding protein, causes growth retardation and abnormal development of reproductive organs. Genome Res 2001;11(8):1327-34.

128. Ma X, Buffler PA, Wiemels JL, Selvin S, Metayer C, Loh M, et al. Ethnic difference in daycare attendance, early infections, and risk of childhood acute lymphoblastic leukemia. Cancer Epidemiol Biomarkers Prev 2005;14(8):1928-34.

129. Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. Stat Med 1990;9(7):811-8.

130. Westergaard T, Andersen PK, Pedersen JB, Olsen JH, Frisch M, Sorensen HT, et al. Birth characteristics, sibling patterns, and acute leukemia risk in childhood: a population-based cohort study. J Natl Cancer Inst 1997;89(13):939-47.

131. Dockerty JD, Draper G, Vincent T, Rowan SD, Bunch KJ. Case-control study of parental age, parity and socioeconomic level in relation to childhood cancers. Int J Epidemiol 2001;30(6):1428-37.

132. Fear NT, Simpson J, Roman E. Childhood cancer and social contact: the role of paternal occupation (United Kingdom). Cancer Causes Control 2005;16(9):1091-7.

133. Baye TM, Wilke RA. Mapping genes that predict treatment outcome in admixed populations. Pharmacogenomics J 2010;10(6):465-77.

134. Ma X, Buffler PA, Selvin S, Matthay KK, Wiencke JK, Wiemels JL, et al. Daycare attendance and risk of childhood acute lymphoblastic leukaemia. Br J Cancer 2002;86(9):1419-24.

135. Bartley K, Metayer C, Selvin S, Ducore J, Buffler P. Diagnostic X-rays and risk of childhood leukaemia. Int J Epidemiol 2010;39(6):1628-37.

136. Hansen HM, Wiemels JL, Wrensch M, Wiencke JK. DNA quantification of whole genome amplified samples for genotyping on a multiplexed bead array platform. Cancer Epidemiol Biomarkers Prev 2007;16(8):1686-90.

137. Jewell N. Statistics for Epidemiology. Boca Raton, Florida: Chapman&Hall/CRC; 2004.

138. Hunter DJ. Gene-environment interactions in human diseases. Nat Rev Genet 2005;6(4):287-98.

139. Sherborne AL, Houlston RS. What are genome-wide association studies telling us about B-cell tumor development? Oncotarget 2010;1(5):367-72.

140. Jensen K, Schaffer L, Olstad OK, Bechensteen AG, Hellebostad M, Tjonnfjord GE, et al. Striking decrease in the total precursor B-cell compartment during early childhood as evidenced by flow cytometry and gene expression changes. Pediatr Hematol Oncol 2010;27(1):31-45.

141. Pui CH. Childhood leukemias. N Engl J Med 1995;332(24):1618-30.

142. Kinlen LJ. Epidemiological evidence for an infective basis in childhood leukaemia. Br J Cancer 1995;71(1):1-5.

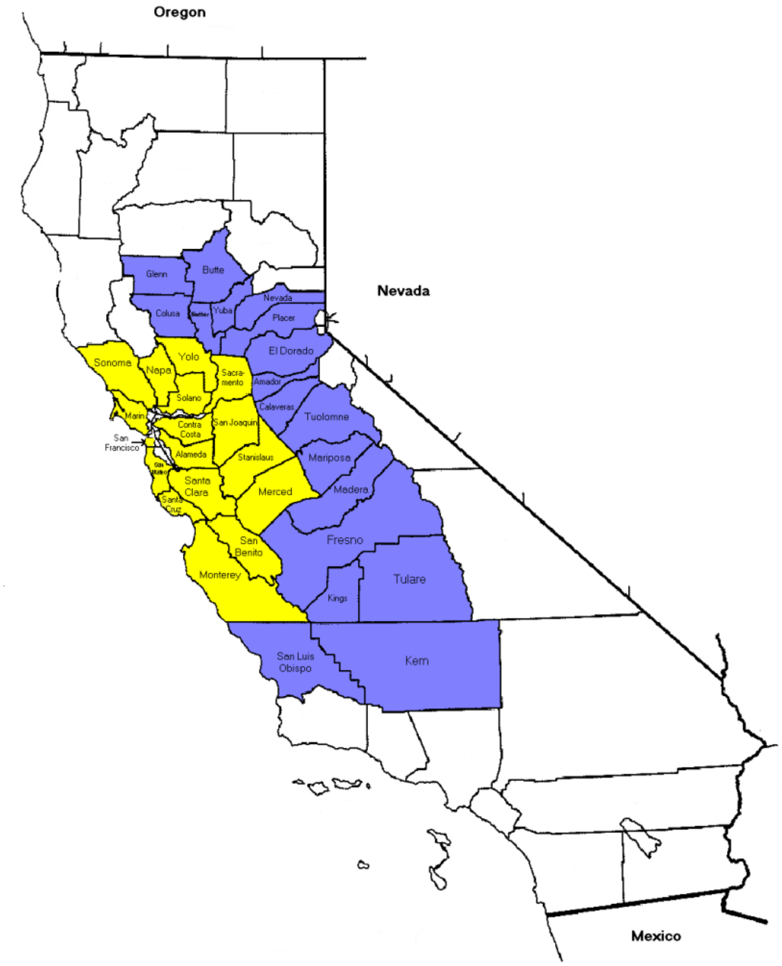
143. Lu N, Samuels ME, Shi L, Baker SL, Glover SH, Sanders JM. Child day care risks of common infectious diseases revisited. Child Care Health Dev 2004;30(4):361-8.
144. Osterholm MT. Infectious disease in child day care: an overview. Pediatrics 1994;94(6 Pt 2):987-90.
145. Gilham C, Peto J, Simpson J, Roman E, Eden TO, Greaves MF, et al. Day care in infancy and risk of childhood acute lymphoblastic leukaemia: findings from UK case-control study. BMJ 2005;330(7503):1294.

Table 1: Effect sizes of associated SNPs identified from previous GWAS

Variant (Gene)	Observed log additive OR (79, 80)
rs4132601 (*IKZF1*)	1.69 (1.40-1.90)
rs7089424 (*ARID5B*)	1.65 (1.54-1.76)
rs10821936(*ARID5B*)	1.91 (1.60-2.20)
rs7073837 (*ARID5B*)	1.65 (1.54-1.76)
rs10740055 (*ARID5B*)	1.53 (1.35-1.89)
rs10994982 (*ARID5B*)	1.61(1.30-1.90)
rs2239633 (*CEBPE*)	1.34 (1.22-1.45)

Figure 1: Map of 35 participating counties of the CCLS

The areas colored by yellow are the original 17 counties participating in the CCLS; the areas colored by purple are additional 18 counties participating in the CCLS later.



CHAPTER 2
Replications of the Susceptibility Loci on 7p12.2, 10q21.2, and 14q11.2 in Acute Lymphoblastic Leukemia in the CCLS and their Interactions with Infection-Related Exposures

ABSTRACT

The etiology of childhood acute lymphoblastic leukemia (ALL) is largely unknown. Epidemiologic studies have provided evidence for a relationship between early life exposure to infections and the risk of childhood ALL. Genome-wide association studies have identified common genetic variants at 10q21.2 (rs7089424, rs10821936, rs7073837, rs10740055, rs10994982, *ARID5B*), 14q11.2 (rs2239633, *CEBPE*), and 7p12.2 (rs4132601, rs11978267, *IKZF1*) as determinants of childhood ALL risk. The current analyses examined the association between these previously identified SNPs and the risk of childhood ALL among 594 non-Hispanic White children (225 cases and 369 controls) and 706 Hispanic children (300 cases and 406 controls) recruited from the California Childhood Leukemia Study (CCLS). In addition, we examined whether there are genetically susceptible subgroups in which a delay in exposure to infections, indicated by daycare attendance by six months old and presence of older siblings, may have differential influences on risk of childhood ALL. All risk estimates were evaluated separately for non-Hispanic White and Hispanic populations. Single SNP analyses were performed using a log-additive model while adjusting for age, sex, and for Hispanics only, child's race. Gene-environment interaction was evaluated using a logistic regression model containing an interaction term while adjusting for child's age, gender, income, and for Hispanics only, child's race.

Five SNPs in *ARID5B* conferred the most significant and consistent results in both non-Hispanic White and Hispanic populations. Two SNPs in *IKZF1* and one SNP in *CEBPE* were confirmed to be associated with increased ALL risk in our non-Hispanic White study population. Suggestive multiplicative interaction was found between SNP rs7073837 of *ARID5B* and daycare attendance by six months old in the risk of childhood ALL in both populations ($OR_{interaction}$=0.57, p=0.09 for non-Hispanic Whites; OR interaction=0.43, p=0.04 for Hispanics). However, the results did not persist after correction for multiple comparisons. Evidence of interaction was not observed for other genotypes and measures of exposure to infections examined.

The findings provide further support for the potential roles of these three genes in the pathogenesis of ALL. Further investigations are needed to fine-map susceptibility loci around these genes and identify additional environmental factors that may modulate the effects of these loci.

INTRODUCTION

Leukemia is the most common childhood malignancy, accounting for 31% of all cancers diagnosed in children younger than 15 years of age (1). Within the major subtypes of childhood leukemia defined by cell lineage, 79% of these cases are acute lymphoblastic or lymphocytic leukemia (ALL), followed by 17% of acute myeloid leukemia (AML) (2). ALL is a heterogeneous disease characterized by the predominance of immature hematopoietic cells. The disease comprises a group of disorders characterized by chromosomal alterations, including aneuploidy (hyper- and hypodiploid) and chromosomal rearrangements (3). Established evidence for an increased risk of ALL includes sex, age, race, prenatal exposure to x-rays, therapeutic radiation, high birth weight and specific genetic syndromes (1). However, these factors explain less than 10% of the cases, leaving the remaining 90% unexplained.

Two recent genome-wide association studies (GWAS) in Caucasian populations have independently shown that germline polymorphisms of the *IKZF1*, *ARID5B*, and *CEBPE* genes are associated with the development of childhood ALL (4, 5). In these studies, associations were found between childhood ALL risk and variants at 10q21.2 (rs7089424, rs10821936, rs7073837, rs10740055, rs10994982, *ARID5B*), 14q11.2 (rs2239633, *CEBPE*), and 7p12.2 (rs4132601, rs11978267, *IKZF1*). In both studies, the loci in *ARID5B* were found to have particularly strong effects for the hyperdiploid subtype of childhood ALL (4, 5). Associations with all of these loci have been confirmed in subsequent replication studies among non-Hispanic White populations (6, 7). Most importantly, all the genes encode proteins with known or postulated roles in normal lymphoid development and/or lymphoid leukemogenesis (8-10).

Given the accumulating support for a role of immunologic factors in the etiology of childhood ALL, infection-related exposures and immune-related processes have emerged as strong candidate risk factors involved in the etiology (11). Greaves hypothesizes that the absence of an early immune challenge and priming during early childhood, combined with ‘delayed’ exposures to infection, might result in subsequent adverse immune responses to common infectious agents, thereby increasing the risk of childhood ALL (12). This suggests that exposure to infections early in life may provide a protective effect against childhood ALL. Prior studies have found associations between ALL and proxy measures of exposure to infections, including daycare attendance (13-16), birth order (17, 18), child’s history of infections (19, 20), play group attendance (21), and parental social contacts in the work place (22, 23), while negative findings have also been reported (24-26).

The reported incidence rate of childhood ALL among California Hispanics is higher than the rate in any other racial/ethnic subgroup in California, including non-Hispanic Whites (27). No biologic factors have been identified which account for this difference in risk, but it may be related to a combination of environmental and genetic risk factors. Hispanics are a recent genetically admixed group and may not share the same susceptibility loci as non-Hispanic Whites (28). In the present study, we attempted to validate these prior identified genetic variants in the non-Hispanic White study population and extend the findings to the Hispanic population in the California Childhood Leukemia Study (CCLS). To test the hypothesis that early life exposure to infections modifies the effect of genetic susceptibility on childhood ALL, we also performed analyses to assess the interactions between the eight candidate SNPs and two proxy measures of early childhood infection, namely daycare attendance by six months old and having an older sibling. These two proxy measures have shown the most compelling results in previous CCLS publications on the risk of childhood ALL (15, 16, 29).

MATERIALS AND METHODS

Study populations

The California Childhood Leukemia Study (CCLS) is an ongoing population-based case control study that began in 1995. Incident cases of newly diagnosed childhood leukemia (age 0–14 years) were rapidly ascertained from major clinical centers in the study area, usually within 72 hours of diagnosis. Cases were initially identified from four and later nine hospitals in the San Francisco Bay Area and Central Valley. Comparisons with population-based surveillance data from the California Cancer Registry (1997-2003) indicated that the 703 cases enrolled in the study between 1997 to 2003 represented around 75% of diagnosed childhood leukemia cases among residents of the study counties (*n*=948), making the study approximately population based. For each case, one or two healthy controls were randomly selected from the state birth registry maintained by the Center for Health Statistics of the California Department of Public Health (CDPH), matching on child's age, sex, Hispanic status, and maternal race (White, Black, Asian/Pacific Islander, Native American, and Other/Mixed). A detailed description of control selection in the CCLS was reported elsewhere (30). A total of 86% of case subjects determined eligible consented to participate and 86% of controls subjects contacted and considered eligible participated (31). Cases and controls were eligible to enter the study if they were under 15 years of age, resided in the study area at the time of diagnosis, had at least one parent who speaks either English or Spanish, and had no prior history of malignancy.

The current analysis included 706 Hispanic (300 cases and 406 controls) and 594 non-Hispanic White (225 cases and 369 controls) subjects recruited between 1995 and 2008 who had available DNA specimens. The child's own race/ethnicity was defined according to that of both parents. For example, a child was considered non-Hispanic White if both parents reported being non-Hispanic ethnicity and White race. Children of parents reporting different races were considered to be of Mixed/Other race. Any child with a parent reporting Hispanic ethnicity was considered Hispanic. Detailed demographic characteristics such as age, gender, and household income, as well as environmental exposures such as pesticide and paints/solvent use, infection history, and smoking history were collected during an in-person home interview.

This study was reviewed and approved by the institutional review committees at the University of California Berkeley, the CDPH, and the participating hospitals. Written informed consent was obtained from all parent respondents.

Biospecimen collection and DNA processing

DNA specimens from buccal cytobrushes were obtained from case and control children either at the hospital or during the in-home personal interview. DNA specimens were extracted and processed within 48 hours of collection by heating in the presence of 0.5N NaOH. DNA was re-purified later either manually using Gentra Puregene reagents (QIAGEN, USA, Valencia, CA) or an automated organic DNA extraction protocol (AutoGen, Holliston, MA). WGA products were cleaned with a Montage PCR9 filter plate (Millipore, Billerica, MA). When buccal cytobrush DNA was inadequate or not available (26.6% of subjects), DNA was isolated from dried bloodspots collected at birth and archived by the Genetic Diseases Screening Program of the CDPH. After extraction using the QIAamp DNA Mini Kit (QIAGEN, USA, Valencia, CA), these DNA samples were whole-genome amplified using REPLI-g reagents (QIAGEN, USA, Valencia, CA).We previously genotyped DNA specimens from both buccal cells and dried blood spot (DBS) for nine subjects; genotype concordance between paired samples was 98.9% (32). Regardless of source, DNA specimens were quantified using human-specific Alu-PCR to confirm a minimum level of amplifiable human DNA, and randomized prior to genotype (33).

We performed Sequenom iPlex genotyping of 7 of 8 previously identified SNPs (4, 5): *ARID5B* (rs10994982, rs10740055, rs7073837, rs7089424, rs10821936), *IKZF1* (rs11978267), and *CEBPE* (rs2239633). The average SNP call rate was 96.7%. Genotypes for duplicate DNA specimens (N=154 per SNP) showed 100% concordance. The one remaining SNP in *IKZF1* (rs4132601) was typed using TaqMan assays. The average SNP call rate was 98.2%. Genotypes for duplicate DNA specimens (N=146 per SNP) showed 100% concordance. All SNPs were tested for deviation from Hardy-Weinberg equilibrium (SAS version 9.2; SAS Institute Inc., Cary, NC). All SNPs included in this analysis (N=8) had a call rate > 90%, had minor allele frequencies > 5% in both Hispanics and non-Hispanics, and did not fail Hardy-Weinberg equilibrium (p< 0.01) in both Hispanic and non-Hispanic controls.

<u>Proxy measures of early childhood exposure to infections</u>

Information on daycare attendance, birth order, and history of infectious illness was collected through in-person interviews with the biological mothers. Detailed information on collection of early childhood exposure to infections was presented previously (13). Briefly, the child's birth order was determined based on a detailed pregnancy history obtained for the biological mother. Information on the child's social contacts outside the home was obtained through a history of daycare attendance before the date of diagnosis for cases and reference dates for controls or before age six, whichever occurred first. For each daycare the child attended, information on age attended, duration of time attended, hours per week, and numbers of other children were obtained. These data were used to calculate "total child-hours of exposure" for each child (13, 29). Child-hours at each daycare facility was calculated as follows: (number of months attending the day care) x (mean hours per week at this day care) x (number of other children at this day care) x (4.35 weeks per month). The measure for "total child-hours of exposure" for each child was calculated by summing the child-hours at each daycare attended. For children who never attended daycare, an age of 72 months was assigned as the age when first started daycare, and a value of zero was assigned for duration of stay, mean hours per week, mean number of children, total number of children, and total child-hours. In this study, we made daycare attendance variable into a dichotomous variable (ever/never) because the distribution of child-hours attendance was binomial for participants included in the study .

Respondents were also asked for a history of common infectious illnesses the child had during the first year of life, including severe diarrhea/vomiting, ear infection, persistent cough, mouth and eye infection, influenza, and unspecified "other infection.

Statistical analysis

Single SNP analyses were conducted assuming a log-additive model (0, 1, or 2 copies of the variant allele), using unconditional logistic regression. The odds ratio (OR) and 95% confidence intervals (CI) were calculated to estimate the risk of ALL associated with each SNP while adjusting for age, sex and for Hispanics only child's race. The results from sensitivity analyses which compared the risk estimates from conditional logistic regression and those from unconditional logistic regression adjusted for the matching variables showed very similar estimates. For each SNP, the referent allele was set to match the allele reported previously (4, 5), even in instances where this allele was less common in our study populations. All risk analyses were performed separately for non-Hispanic White and Hispanic children. Correction for multiple testing was performed using the Benjamini-Hochberg false discovery rate (FDR) method with a type I error rate of 5%; nominal P-values are shown (14). Similarly, logistic regression adjusting for age, sex, income and for Hispanics only child's race was used to estimate the odds ratio (OR) and 95% confidence interval (CI) associated with the two surrogate measures of early life infection exposure, daycare attendance by age six months and having an older sibling.

To test for heterogeneity (interaction), we focused on the effects of the two social contact measures, birth order and daycare attendance on ALL risk because they offered the most complete data. The data for other infection variables were less complete compared to the social contact measures and offered less statistical power to detect interactions. The joint effects of the two infectious exposure variables and the eight SNPs were evaluated using a multiplicative logistic regression model that included a product term for presence or absence of the risk variant and the two proxy measures of early exposure to common infections (daycare attendance by age six months and having an older sibling), while adjusting for age, sex, income and for Hispanics only child's race. There were 515 ALL cases and 748 controls in the gene-environment interaction analyses, after excluding subjects under the age one year to allow for sufficient exposure to infectious factors before the development of childhood leukemia. The product term was the interaction odds ratio ($OR_{Interaction}$), defined as the ratio of the joint effect of the two infection exposure variables and the eight genetic variants and the product of the individual effects ($OR_{Interaction} = OR_{GE} / (OR_G \text{ x } OR_E)$). It is expected that OR_{GE} associated with the infection exposure variables and genetic variants is equal to the product term of OR_G (independent effect of the genetic variants in the absence of the infection exposure variables) and OR_E (independent effect of the infection exposure variables in the absence of the genetic variants) when there is no joint effect. If there is a multiplicative interaction between the infection exposure variables and the genetic variants, then $OR_{GE} \neq OR_G \times OR_E$ and will differ significantly from one. A p-value of 0.2 or less for interaction was considered statistically significant given the available sample size, i.e. <300 observations (34). Interactions between the SNPs and the infection exposure variables were evaluated separately for Hispanics and non-Hispanic Whites because the patterns of daycare attendance and family structure differed significantly by ethnicity (13).

RESULTS

Study characteristics of 706 Hispanics (300 cases and 406 controls) and 594 non-Hispanic Whites (225 cases and 369 controls) are described in **Table 1**. Cases and controls were comparable in the distribution of sex and age. Cases generally had lower annual household income compared to controls ($p<0.001$). Because Hispanics are a recently admixed group (28), a large proportion (59%-68%) of our Hispanic population reported “Mixed or Other” race. We observed a similar frequency of B-cell precursor (BCP) ALL and BCP high-hyperdiploid ALL among non-Hispanic White and Hispanic cases. The frequency of BCP ALL was 90 % among non-Hispanic Whites and 93 % among Hispanics. The frequency of BCP high-hyperdiploid ALL (>50 chromosomes) was similar for non-Hispanic Whites and Hispanics; i.e., 31%.

The associations of the eight candidate SNPs with the risk of ALL in non-Hispanic White children are shown in **Table 2.** Consistent with findings in European populations, five SNPs from the region that annotate the *ARID5B* gene were strongly associated with ALL risk [rs10994982, OR $_{\text{per-allele}}$ = 1.37 (p=0.018); rs10740055, OR $_{\text{per-allele}}$ = 1.50 (p= 0.003); rs7073837, OR $_{\text{per-allele}}$ = 1.43(p= 0.009); rs7089424, OR $_{\text{per-allele}}$ = 1.84 ($p=2.2\times10^{-6}$); rs10821936, OR $_{\text{per-allele}}$ = 1.79 ($p=4.8\times10^{-6}$)]. Odds ratio estimates were in the same direction and were of similar strength as those previously reported (4-7). These associations for these five SNP withstood multiple testing corrections and remained significant after controlling for a false-discovery rate of 5 %. Analyses were further performed among ALL subgroups. Although the power decreased with smaller sample size in the subgroup analyses, five *ARID5B* SNPs remained statistically significant in the BCP ALL subgroup. The associations also remained significant when the analysis was confined to BCP high-hyperdioploid ALL ($p<0.05$). The variant rs2239633 within *CEBPE* was also significantly associated with ALL risk among non-Hispanic Whites (p=0.005). When restricted to BCP and BCP high-hyperdioploid ALL, the associations persisted. The two variants (rs11978267 and rs4132601) within *IKZF1* also showed significant associations with ALL ($p=7.8\times10^{-6}$ and 8.4×10^{-6}) and the effects were similar across subgroups.

The associations of the eight candidate SNPs with the risk of ALL in Hispanic children are shown in **Table 3.** Similar effects were observed in *ARID5B* SNPs for the Hispanic population. All five *ARID5B* SNPs showed strong evidence of an increased risk associated with ALL [rs10994982, OR $_{\text{per-allele}}$ = 1.53 (p=0.004); rs10740055, OR $_{\text{per-allele}}$ = 1.61 ($p= 1\times10^{-5}$); rs7073837, OR $_{\text{per-allele}}$ = 1.76 ($p= 4.3\times10^{-7}$); rs7089424, OR $_{\text{per-allele}}$ = 1.98 ($p=1\times10^{-9}$); rs10821936, OR $_{\text{per-allele}}$ = 1.99 ($p=1.2\times10^{-9}$)]. The risk estimates associated with SNPs were similar among ALL and BCP ALL cases. Interestingly, the risk estimates were stronger and remained significant when the analysis was restricted to ALL cases with BCP high-hyperdiploid ALL [rs10994982, OR $_{\text{per-allele}}$ =2.21(95% CI,1.52-3.21); rs10740055, OR $_{\text{per-allele}}$ = 2.43 (95% CI ,1.67-3.55); rs7073837, OR $_{\text{per-allele}}$ = 2.66 (95% CI, 1.75-4.02); rs7089424, OR $_{\text{per-allele}}$ = 3.22 (95% CI, 2.18-4.73); rs10821936, OR $_{\text{per-allele}}$ = 3.08 (95% CI, 2.11-4.48)]. The variant (rs2239633) within *CEBPE* did not reveal a significant association with ALL risk, but showed a suggestive effect (OR $_{\text{per-allele}}$ = 1.24, p=0.06). When confined to BCP high-hyperdiploid ALL, the variant (rs2239633) was significantly associated with ALL subtype (OR $_{\text{per-allele}}$ = 1.81, p=0.003). In contrast to the observations among non-Hispanic Whites, two SNPs (rs11978267 and rs4132601) in *IKZF1* were not significantly associated with risk of ALL among Hispanics (p=0.08 and 0.14, respectively).

The association of daycare attendance and birth order with risk of ALL was examined (**Table 4**). Among non-Hispanic Whites, an inverse association between having an older sibling and ALL risk (OR=0.65; 95% CI, 0.46-0.93) was observed. A suggestive association was also observed for daycare attendance by age six months, which showed a reduced risk of ALL among children who attended daycare (OR=0.66, 95% CI, 0.43-1.01). These associations were not observed among Hispanics.

Potential interactions between each of the eight candidate SNPs and the early life infection variables (daycare attendance and birth order) were examined within each ethnic group. The results for the joint effect of the eight SNPs and daycare attendance before age six months with the risk of childhood ALL indicated that the strongest interaction was observed between rs7073837 within *ARID5B* and daycare attendance in Hispanics ($OR_{interaction}$=0.43, p=0.04) (**Table 5**). The potential interaction suggested that in the presence of the two factors, the risk is lower than what would be expected based on a multiplicative scale. Among non-Hispanic Whites, the same variant (rs7073837) also showed a suggestive interaction between daycare attendance and the variant ($OR_{interaction}$=0.57; p=0.09) (**Table 5**). However, none of the interaction p-values remained statistically significant after correction for multiple comparisons. The interaction ORs for the remaining genetic polymorphisms were all non-significant, with p-values >0.20, indicating no joint influence between daycare attendance and those genetic polymorphisms in the risk of ALL using a multiplicative model of interaction.

Further stratified analyses among non-Hispanic White children with no daycare attendance before age of six months indicated that the variant rs7073837 within *ARID5B* was associated with an increased risk of childhood ALL ($OR_{per-allele}$ =1.74, 95% CI=1.26, 2.41). The same analysis of the variant among children who attended daycare showed a risk estimate in the opposite direction ($OR_{per-allele}$ =0.99, 95% CI=0.56, 1.76). A similar effect was observed among Hispanic children. The variant rs7073837 within *ARID5B* was associated with increased ALL risk among Hispanic children who did not a attend daycare facility before six months of age ($OR_{per-allele}$ =2.16, 95% CI=1.62, 2.89). The risk estimate for the variant rs7073837 and risk of ALL were in the opposite direction among children who attended daycare ($OR_{per-allele}$ =0.94, 95% CI=0.41, 2.18).

The results for the joint effect of the eight candidate SNPs and birth order with the risk of childhood ALL are shown in **Table 6**. Unlike daycare attendance, birth order showed less strong evidence of multiplicative interaction with candidate SNPs in the risk of childhood ALL. Joint effects for the two *ARID5B* SNPs, rs7089424 and rs10821936, and having an older sibling showed a suggestive effect of interaction ($OR_{interaction}$=1.48 (p=0.17) and $OR_{interaction}$=1.43 (p=0.17)) among non-Hispanic Whites. These two variants did not show the same association among Hispanics. Additionally, there was no evidence of interactions between other genetic variants in *ARID5B*, *CEBPE*, and *IKZF1* and having an older sibling in the both the non-Hispanic White and Hispanic populations, with a p-value < 0.2. Furthermore, interactions between other infectious illness and the eight candidate SNPs were not observed (data not shown).

DISCUSSION

In this population-based case-control study, we aimed to validate results for SNPs identified in previous GWAS in non-Hispanic White populations and extend the findings to a Hispanic population. The ethnic diversity of our California study population offers us the opportunity to perform analyses of non-Hispanic White and Hispanic children separately. In addition, since the etiology of childhood ALL is likely to result from a combination of environmental exposures and susceptibility factors influenced by individual genetics (35), we examined potential interactions between indicators of early exposure to infections and these inherited ALL polymorphisms among non-Hispanic White and Hispanic children. Consistent with previous findings (4-7), we confirmed that previous GWAS-identified SNPs in B-cell development genes *(ARID5B*, *CEBPE*, and *IKZF1)* are associated with childhood ALL risk in the CCLS non-Hispanic White population. Intriguingly, the *ARID5B* and *CEBPE* SNPs are also significantly associated with B-cell ALL risk in the CCLS Hispanic population.

Current evidence suggests that childhood leukemia may result from chromosomal alterations and mutations that disrupt the normal differentiation process of lymphoid or myeloid progenitor cells (36). Inherited genetic variation in these regulatory genes may contribute to the pathogenesis of ALL and may have important implications for ALL treatment (37). Our analysis suggests that *ARID5B* is associated with increased ALL risk in both the non-Hispanic White and Hispanic populations. It is notable that these five SNPs are validated in different ethnic groups with relatively common minor allele frequencies (30%-48%). In previous replication studies, five SNPs within *ARID5B* showed consistent positive associations with the risk of ALL (6, 7). Similar associations have been found in SNP rs10821936 of *ARID5B* with ALL risk in African American, Thai and Hispanic populations (37-39). *ARID5B* SNPs rs7089424 and rs10821936 showed the strongest signal in our study and are in strong linkage disequilibrium (r^2=0.95) with each other. The *ARID5B* SNP rs7073937 which is in moderate LD with the pair rs7089424 and rs10821936 (r^2=0.65), yielded the second-strongest signal associated with increased ALL risk. All five variants are mapped to intron regions of *ARID5B*, which encodes AT- rich interactive domain 5B. *ARID5B* is a member of the AT-rich interaction domain family of transcription factors that is essential for embryogenesis and growth regulation (40). In an animal model, *ARIDB5* homozygous null mice showed immune abnormalities, including defects in the B lymphoid progenitors (41). Expression of *ARID5B* is also correlated with *RAG1* expression in bone marrow (42), which indicates that altering *ARID5B* function is associated with early B-cell development and may contribute to leukemogenesis (43).

We were able to replicate the reported associations of *IKZF1* with ALL in the non-Hispanic White population, but not in the Hispanic population. In a previous German case-control replication study, the results showed convincing support for an association between rs4132601 and ALL risk (6). However, the Canadian study failed to replicate the association between SNPs rs4132601 and rs11978267 and ALL risk (7). In addition, the p-value for rs4132601 among Hispanics was of only borderline significance ($p = 0.08$). The discrepancy in the findings might due to the fact that the *IKZF1* associations observed among the non-Hispanic Whites in our study and among the European Caucasian populations in previous studies are indirect, i.e. due to linkage disequilibrium with other causal variants. The minor allele frequencies of rs4132601 and rs11978267 are the same in our two populations (27% in non-Hispanic Whites and 27% in Hispanics) and are in complete LD with each other (r^2 =1). These two SNPs are annotated to chromosome region 7p12.2, the Ikaros family zinc finger 1 (*IKZF1*). This gene *IKZF1* encodes Ikaros protein, which is responsible for regulating the development and function of the immune system and acts as a master regulator of hematopoietic differentiation (44). *IKZF1* is known to be associated with survival in childhood ALL (45). Studies from the Children Oncology Group (COG) leukemia study group have shown that patients with deletion or mutations of *IKZF1* have nearly three times the risk of treatment failure (45).

We also observed a similar effect of the *CEBPE* SNP rs2239633 in the non-Hispanic White population, as reported previously, and for the Hispanic population in B-cell ALL and hyperdiploid ALL subtypes. The SNP rs2239633 maps to the gene *CEBPE*, encoding CCAAT/enhancer-building protein epsilon. *CEBPE* is involved in functional maturation and terminal differentiation of myeloid cells, as well as being a target of chromosomal translocation in ALL (8). Deregulated *CEBPE* expression may lead to malignant transformation (8). The minor allele frequency is 49% in the non-Hispanic White population and 39% in the Hispanic population. Therefore, the lack of association in the Hispanic population is likely due to the potential diversity in underlying genetic structure.

The greater risk of ALL in children than in adults has been linked to developmental immaturity of the immune system and differential exposures to environmental toxins (21). By contributing to normal maturation of the immune system, early common infections would protect the child against leukemia, while a lack of such early exposures would make the child more vulnerable to overreact to later infections (46). Greaves's "delayed infection" hypothesis suggests that the lack of early immunomodulatory exposures may lead to deficiencies in the child's immune system and later contribute to the potential for the development of leukemia (12, 47). The occurrence of common infections during early childhood has been associated with a child's social contacts with other children within the home, and also exposures outside the home. Birth order is considered to be a surrogate marker of infection-related exposures within the home, with later born children presumed to be exposed to older siblings more often and exposed to infectious agents at an earlier age (48, 49). Daycare attendance represents an opportunity for child's social contact outside the home and has been used as a surrogate of early exposure to infection. In developed countries, infections spread through the fecal-oral and respiratory routes occur frequently in daycare facility (49). In accordance with Greaves's hypothesis, previous studies have found associations between a reduced risk of childhood ALL and daycare attendance and high birth order (13, 16, 18, 29, 50). More compelling evidence with respect to supporting Greaves's delayed hypothesis has been observed for daycare attendance. A recent meta-analysis of 14 case-control studies indicated that daycare attendance either before age one or age two was associated with a reduced risk of ALL (OR=0.76, 95% CI (0.67- 0.87))(15). Children who have attended daycare are assumed to have been exposed to infections at an earlier age compared to those who have not attended day care. Several epidemiologic studies, including CCLS have shown that the strongest reduction in ALL risk occurs when daycare attendance is starts before the six months of age (29, 50-52).

The gene-environment analysis presented here suggests that lack of immune priming early in life may be associated with an adverse immune response to infection. Variations in the genes responsible for development may affect children's immune responses and possibly their risk of ALL. Our data suggest that the variant rs7073837 within *ARID5B* gene may modify the effect of daycare attendance on the risk of childhood ALL. Specifically, an increased ALL risk associated with the variant A allele of *ARID5B* rs7073837 was observed among those children who did not attend daycare facility before age of six months in both non-Hispanic White and Hispanic children in our study, even though 8.6% of Hispanic control children and 26.6% of non-Hispanic White control children attended day care before age of six months. As stated previously, the SNP rs7073837 is located in intron 2 of *ARID5B* gene, and this gene plays a vital role in the regulation of embryonic development and cell growth. Even though this analysis suggested potential interactions between daycare attendance and a SNP within *ARID5B*, the lack of power and the potential for random variations due to the small numbers of study subjects make definitive inferences impossible. More power to address such hypothesis could come from adding more study subjects through continuing recruitment of subjects and through collaboration between study groups.

We observed suggestive gene-environment interactions between SNPs rs7089424 and rs10821936 within the *ARID5B* gene and having an older sibling among non-Hispanic Whites (p=0.17 for both populations). The increased ALL risk associated with the G allele of rs7089424 was stronger among non-first born children. Similarly, an increased ALL risk associated with the C allele of rs10821936 was stronger among non-first born children. This interaction was not observed in the Hispanic population. The observed differences between the two ethnicity groups may result from the fact that Hispanic children tend to have more siblings and live in larger families. Therefore, this variable may not be a good proxy for early life infection within the home in the Hispanic population.

The results of our analyses need to be interpreted in the context of several limitations. One of the limitations and challenges of assessing gene-environment interaction is the lack of statistical power. As noted previously, we observed differences in the distribution of daycare attendance by age six months and having an older sibling in the Hispanic and the non-Hispanic White populations. As a result, all risk analyses were conducted in the non-Hispanic White and Hispanic population separately. Given the reduced sample size, i.e. less than 300 observations per group, the statistical power to detect the multiplicative interaction is weak to moderate. The study had 80% power to detect an interaction odds ratio of 2.0, assuming minor allele frequency of 30 %, an exposure prevalence of 20%, an estimated OR of 1.3 for the genotype and the exposure, and a two-sided significance level of 0.05. Even though the CCLS is one of the largest case-control studies of childhood ALL in the United States, gene-environment interactions with moderate to small effect sizes may not have been detected. In addition, daycare attendance, birth order and infection histories were intended to serve as surrogate measures of infectious exposure. The self-reported infection variables are subject to exposure misclassification, since information was not always corroborated with child’s medical records or recalled correctly. There are other possible sources of exposure to infectious agents that we did not consider in the study, such as parental occupation contacts, and/or total number of individuals living in the household (53). Other factors such as breastfeeding, maternal medication use, immunization, and mother’s infection history, may also affect the early development of immune system.

Another consideration is population stratification because one of our study populations is Hispanic, a recently genetically mixed group. As an admixed group, Hispanics have non-Hispanic White, Amerindian, and African ancestry. Asian populations are closely related to Amerindian populations and our own analysis of genetic ancestry, using ancestry informative markers (AIMs), has shown that California Hispanics have little African ancestry (data not shown). Nonetheless, population stratification is likely to be minimal in the CCLS due to careful individual-matching of cases and controls on child’s Hispanic status and maternal race. A previous analysis accounting for AIMs-based genetic ancestry in the CCLS found no major effects of population stratification (54).

To our knowledge, this is the first study to investigate the SNPs identified through GWAS in the Hispanic population and evaluate their interactions with proxies for early life exposure to infections on the risk of ALL. We sought to validate previously identified GWAS SNPs in non-Hispanic Whites and Hispanics in the CCLS. We also examined the putative joint effects of early life exposure to infection and the eight candidate SNPs on the risk of ALL. Our non-Hispanic White population exhibited similar risk estimates for SNPs on *ARID5B*, *CEBPE* and *IKZF1* compared to the previous GWAS. We found that genetic variations in the *ARID5B* and *CEBPE* genes are associated with the development of ALL in two diverse populations, pointing to the importance of the regions in the etiology of ALL.

Even though the study results suggest potential interactions between early exposures to infection and *ARID5B* genes, larger studies of homogeneous phenotypes are required to replicate the findings and address whether these gene-environment interactions are subtype specific. Further confirmation is needed to determine functional variants around the significant genomic regions within these three genes. More effort is also required to discover whether the genetic determinants of childhood ALL are population-specific or overlap between these populations.

REFERENCES

1. Ries LAG SM, Gurney JG. Cancer Incidence and Survival among Children and Adolescents: United States SEER Program 1975-1995. Bethesda, MD: National Cancer Institute, SEER Program. NIH Pub. 1999:No. 99-4649.

2. Pui CH, editor. Childhood Leukemia: Cambridge University Press; 2006.

3. Pui CH, Relling MV, Downing JR. Acute lymphoblastic leukemia. N Engl J Med 2004;350(15):1535-48.

4. Papaemmanuil E, Hosking FJ, Vijayakrishnan J, Price A, Olver B, Sheridan E, et al. Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. Nat Genet 2009;41(9):1006-10.

5. Trevino LR, Yang W, French D, Hunger SP, Carroll WL, Devidas M, et al. Germline genomic variants associated with childhood acute lymphoblastic leukemia. Nat Genet 2009;41(9):1001-5.

6. Prasad RB, Hosking FJ, Vijayakrishnan J, Papaemmanuil E, Koehler R, Greaves M, et al. Verification of the susceptibility loci on 7p12.2, 10q21.2, and 14q11.2 in precursor B-cell acute lymphoblastic leukemia of childhood. Blood 2010;115(9):1765-7.

7. Healy J, Richer C, Bourgey M, Kritikou EA, Sinnett D. Replication analysis confirms the association of ARID5B with childhood B-cell acute lymphoblastic leukemia. Haematologica 2010;95(9):1608-11.

8. Akasaka T, Balasas T, Russell LJ, Sugimoto KJ, Majid A, Walewska R, et al. Five members of the CEBP transcription factor family are targeted by recurrent IGH translocations in B-cell precursor acute lymphoblastic leukemia (BCP-ALL). Blood 2007;109(8):3451-61.

9. Georgopoulos K, Bigby M, Wang JH, Molnar A, Wu P, Winandy S, et al. The Ikaros gene is required for the development of all lymphoid lineages. Cell 1994;79(1):143-56.

10. Lahoud MH, Ristevski S, Venter DJ, Jermiin LS, Bertoncello I, Zavarsek S, et al. Gene targeting of Desrt, a novel ARID class DNA-binding protein, causes growth retardation and abnormal development of reproductive organs. Genome Res 2001;11(8):1327-34.

11. Greaves M. Molecular Genetics, Natural History and the Demise of Childhood Leukaemia. European Journal of Cancer 1999;35(2):173-185.

12. Greaves M. Infection, immune responses and the aetiology of childhood leukaemia. Nat Rev Cancer 2006;6(3):193-203.

13. Ma X, Buffler PA, Wiemels JL, Selvin S, Metayer C, Loh M, et al. Ethnic difference in daycare attendance, early infections, and risk of childhood acute lymphoblastic leukemia. Cancer Epidemiol Biomarkers Prev 2005;14(8):1928-34.

14. Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. Stat Med 1990;9(7):811-8.

15. Urayama KY, Buffler PA, Gallagher ER, Ayoob JM, Ma X. A meta-analysis of the association between day-care attendance and childhood acute lymphoblastic leukaemia. Int J Epidemiol 2010;39(3):718-32.

16. Urayama KY, Ma X, Selvin S, Metayer C, Chokkalingam AP, Wiemels JL, et al. Early life exposure to infections and risk of childhood acute lymphoblastic leukemia. Int J Cancer 2011;128(7):1632-43.

17. Westergaard T, Andersen PK, Pedersen JB, Olsen JH, Frisch M, Sorensen HT, et al. Birth characteristics, sibling patterns, and acute leukemia risk in childhood: a population-based cohort study. J Natl Cancer Inst 1997;89(13):939-47.

18. Dockerty JD, Draper G, Vincent T, Rowan SD, Bunch KJ. Case-control study of parental

age, parity and socioeconomic level in relation to childhood cancers. Int J Epidemiol 2001;30(6):1428-37.

19. Roman E, Simpson J, Ansell P, Kinsey S, Mitchell CD, McKinney PA, 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 2007;165(5):496-504.

20. Rudant J, Orsi L, Menegaux F, Petit A, Baruchel A, Bertrand Y, et al. Childhood acute leukemia, early common infections, and allergy: The ESCALE Study. Am J Epidemiol 2010;172(9):1015-27.

21. Eden T. Aetiology of childhood leukaemia. Cancer Treat Rev 2010;36(4):286-97.

22. Chang JS, Metayer C, Fear NT, Reinier K, Yin X, Urayama K, et al. Parental social contact in the work place and the risk of childhood acute lymphoblastic leukaemia. Br J Cancer 2007;97(9):1315-21.

23. Fear NT, Simpson J, Roman E. Childhood cancer and social contact: the role of paternal occupation (United Kingdom). Cancer Causes Control 2005;16(9):1091-7.

24. Rosenbaum PF, Buck GM, Brecher ML. Early child-care and preschool experiences and the risk of childhood acute lymphoblastic leukemia. Am J Epidemiol 2000;152(12):1136-44.

25. Neglia JP, Linet MS, Shu XO, Severson RK, Potter JD, Mertens AC, et al. Patterns of infection and day care utilization and risk of childhood acute lymphoblastic leukaemia. Br J Cancer 2000;82(1):234-40.

26. MacArthur AC, McBride ML, Spinelli JJ, Tamaro S, Gallagher RP, Theriault GP. Risk of childhood leukemia associated with vaccination, infection, and medication use in childhood: the Cross-Canada Childhood Leukemia Study. Am J Epidemiol 2008;167(5):598-606.

27. Campleman SL WW. Childhood cancer in California 1988 to 1999 Volume I: birth to age 14. Sacramento, CA: California Department of Health Services, Cancer Surveillance Section. 2004:16-17.

28. Baye TM, Wilke RA. Mapping genes that predict treatment outcome in admixed populations. Pharmacogenomics J 2010;10(6):465-77.

29. Ma X, Buffler PA, Selvin S, Matthay KK, Wiencke JK, Wiemels JL, et al. Daycare attendance and risk of childhood acute lymphoblastic leukaemia. Br J Cancer 2002;86(9):1419-24.

30. Ma X, Buffler PA, Layefsky M, Does MB, Reynolds P. Control selection strategies in case-control studies of childhood diseases. Am J Epidemiol 2004;159(10):915-21.

31. Bartley K, Metayer C, Selvin S, Ducore J, Buffler P. Diagnostic X-rays and risk of childhood leukaemia. Int J Epidemiol 2010;39(6):1628-37.

32. Chokkalingam AP, Bartley K, Wiemels JL, Metayer C, Barcellos LF, Hansen HM, et al. Haplotypes of DNA repair and cell cycle control genes, X-ray exposure, and risk of childhood acute lymphoblastic leukemia. Cancer Causes Control 2011;22(12):1721-30.

33. Hansen HM, Wiemels JL, Wrensch M, Wiencke JK. DNA quantification of whole genome amplified samples for genotyping on a multiplexed bead array platform. Cancer Epidemiol Biomarkers Prev 2007;16(8):1686-90.

34. Jewell N. Statistics for Epidemiology. Boca Raton, Florida: Chapman&Hall/CRC; 2004.

35. Hunter DJ. Gene-environment interactions in human diseases. Nat Rev Genet 2005;6(4):287-98.

36. Pui CH, Robison LL, Look AT. Acute lymphoblastic leukaemia. Lancet 2008;371(9617):1030-43.

37. Xu H, Cheng C, Devidas M, Pei D, Fan Y, Yang W, et al. ARID5B genetic

polymorphisms contribute to racial disparities in the incidence and treatment outcome of childhood acute lymphoblastic leukemia. J Clin Oncol 2012;30(7):751-7.

38. Yang W, Trevino LR, Yang JJ, Scheet P, Pui CH, Evans WE, et al. ARID5B SNP rs10821936 is associated with risk of childhood acute lymphoblastic leukemia in blacks and contributes to racial differences in leukemia incidence. Leukemia 2010;24(4):894-6.

39. Vijayakrishnan J, Sherborne AL, Sawangpanich R, Hongeng S, Houlston RS, Pakakasama S. Variation at 7p12.2 and 10q21.2 influences childhood acute lymphoblastic leukemia risk in the Thai population and may contribute to racial differences in leukemia incidence. Leuk Lymphoma 2010;51(10):1870-4.

40. Patsialou A, Wilsker D, Moran E. DNA-binding properties of ARID family proteins. Nucleic Acids Res 2005;33(1):66-80.

41. Sherborne AL, Houlston RS. What are genome-wide association studies telling us about B-cell tumor development? Oncotarget 2010;1(5):367-72.

42. Jensen K, Schaffer L, Olstad OK, Bechensteen AG, Hellebostad M, Tjonnfjord GE, et al. Striking decrease in the total precursor B-cell compartment during early childhood as evidenced by flow cytometry and gene expression changes. Pediatr Hematol Oncol 2010;27(1):31-45.

43. Wilsker D, Patsialou A, Dallas PB, Moran E. ARID proteins: a diverse family of DNA binding proteins implicated in the control of cell growth, differentiation, and development. Cell Growth Differ 2002;13(3):95-106.

44. John LB, Ward AC. The Ikaros gene family: transcriptional regulators of hematopoiesis and immunity. Mol Immunol 2011;48(9-10):1272-8.

45. Mullighan CG, Su X, Zhang J, Radtke I, Phillips LA, Miller CB, et al. Deletion of IKZF1 and prognosis in acute lymphoblastic leukemia. N Engl J Med 2009;360(5):470-80.

46. Pui CH. Childhood leukemias. N Engl J Med 1995;332(24):1618-30.

47. Kinlen LJ. Epidemiological evidence for an infective basis in childhood leukaemia. Br J Cancer 1995;71(1):1-5.

48. Lu N, Samuels ME, Shi L, Baker SL, Glover SH, Sanders JM. Child day care risks of common infectious diseases revisited. Child Care Health Dev 2004;30(4):361-8.

49. Osterholm MT. Infectious disease in child day care: an overview. Pediatrics 1994;94(6 Pt 2):987-90.

50. Perrillat F, Clavel J, Auclerc MF, Baruchel A, Leverger G, Nelken B, et al. Day-care, early common infections and childhood acute leukaemia: a multicentre French case-control study. Br J Cancer 2002;86(7):1064-9.

51. Gilham C, Peto J, Simpson J, Roman E, Eden TO, Greaves MF, et al. Day care in infancy and risk of childhood acute lymphoblastic leukaemia: findings from UK case-control study. BMJ 2005;330(7503):1294.

52. Jourdan-Da Silva N, Perel Y, Mechinaud F, Plouvier E, Gandemer V, Lutz P, et al. Infectious diseases in the first year of life, perinatal characteristics and childhood acute leukaemia. Br J Cancer 2004;90(1):139-45.

53. McNally RJ, Eden TO. An infectious aetiology for childhood acute leukaemia: a review of the evidence. Br J Haematol 2004;127(3):243-63.

54. Chokkalingam AP AM, Bartley K, Hsu LI, Metayer C, et al. Matching on Race and Ethnicity in Case-Control Studies as a Means of Control for Population Stratification. Epidemiol 2011;1:101.

Table 1. Characteristics of study participants by Hispanic status, CCLS, 1995-2008

	Non-Hispanic Whites		***Hispanics***	
	Cases, n (%)	***Control, n (%)***	***Cases, n (%)***	***Control, n (%)***
Study subjects	225 (37.8)	369 (62.2)	300 (42.5)	406 (57.5)
Sex				
Male	125 (55.6)	221 (59.9)	156 (52.0)	214 (52.8)
Female	100 (44.4)	148 (40.1)	144(48.0)	192 (47.2)
Age				
Mean age, y (SE)	5.6 (3.4)	5.7 (3.8)	5.6 (3.6)	5.5 (3.6)
Income				
<$15,000	14 (6.2)	14 (3.8)	73 (24.4)	65 (16.1)
$15,000-$29,999	20 (8.9)	16 (4.3)	78 (26.1)	89 (21.9)
$30,000-$44,999	26 (11.6)	29 (7.8)	61 (20.4)	81 (20.0)
$45,000-$59,999	37 (16.5)	47 (12.8)	40 (13.4)	62 (15.3)
$60,000-$74,999	26 (11.6)	41 (11.1)	16 (5.3)	38 (9.4)
≥$75,000	101 (45.1)	221 (60.1)	31 (10.4)	71 (17.5)
Race				
White/Caucasian	225 (100)	369 (100)	114 (38.0)	116 (28.5)
African American	-	-	2 (0.7)	0 (0)
Native American	-	-	6 (2.0)	11 (2.7)

Asian or Pacific Islander	-	-	0 (0)	3 (0.7)
Mixed or Other			178(59.3)	276 (67.9)
Disease subtypes (case only)				
B-cell	202 (90.2)	-	279 (93.3)	-
T-cell	21 (9.4)	-	20 (6.7)	-
Both lineage	1 (0.4)	-	0 (0.0)	-
Cytogenetic characteristics (case-only)				
B-cell hyperdiploid (51-67 chromosomes)	64 (30.9)	-	85 (31.1)	-

Table 2. Association of eight candidate SNPs and their characteristics with risk of ALL in Non-Hispanic White children by type of disease and compared to results from previous GWAS (4, 5), CCLS, 1995-2008

Chr.	SNP	Gene	Allele†	Risk allele frequency	Log additive OR‡ (CI)	P-value	Log additive OR‡ (CI) B-cell ALL	Log additive OR‡ (CI) B-cell hyperdiploid	Previous GWAS OR (CI) (4, 5)
10	rs10994982[a]	*ARID5B*	**A**/G	0.48	1.37(1.08-1.73)	0.018	1.41(1.10-1.80)	1.67(1.14-2.47)	1.61(1.30-1.90)
10	rs10740055[b]	*ARID5B*	A/**C**	0.48	1.50(1.18-1.90)	0.003	1.57(1.23-2.02)	1.76(1.19-2.61)	1.53(1.41-1.64)
10	rs7073837[b]	*ARID5B*	**A**/C	0.45	1.43(1.11-1.85)	0.009	1.59(1.19-2.11)	2.30(1.48-3.59)	1.58(1.35-1.89)
10	rs7089424[b]	*ARID5B*	**G**/T	0.29	1.84(1.43-2.37)	2.2x10^{-6}	1.98(1.50-2.61)	2.71(1.79-4.12)	1.65(1.54-1.76)
10	rs10821936[a]	*ARID5B*	**C**/T	0.31	1.79(1.40-2.28)	4.8x10^{-6}	1.83(1.21-2.36)	2.43(1.64-3.61)	1.91(1.60-2.20)
14	rs2239633[b]	*CEBPE*	**C**/T	0.51	1.44(1.13-1.82)	0.005	1.83(1.42-2.36)	1.88(1.25-2.82)	1.34(1.22-1.45)
7	rs11978267[a]	*IKZF1*	A/**G**	0.27	1.81(1.41-2.32)	7.8x10^{-6}	1.77(1.37-2.30)	1.76(1.18-2.62)	1.69(1.40-1.90)
7	rs413260[b]	*IKZF1*	**G**/T	0.27	1.80(1.41-2.31)	8.4x10^{-6}	1.78(1.37-2.30)	1.75(1.18-2.61)	1.69(1.58-1.81)

† Bolded letter indicates risk allele from previous GWAS; ‡ Odds ratios (OR) and 95% confidence intervals (CI) were calculated using log additive models adjusting for age and sex.

a. SNPs identified by Trevino, L. R. et al.
b. SNPs identified by Papaemmanuil, E et al.

Table 3. Association of eight candidate SNPs and their characteristics with risk of ALL in Hispanic children by type of disease, CCLS, 1995-2008

Chr.	SNP	Gene	Function	Allele †	Risk allele frequency	Log additive OR‡ (CI)	P-value	Log additive OR‡ (CI) B-cell ALL	Log additive OR‡ (CI) B-cell hyperdiploid
10	rs10994982	*ARID5B*	INTRON	**A**/G	0.54	1.53(1.23-1.90)	0.0004	1.65(1.31-2.07)	2.21(1.52-3.21)
10	rs10740055	*ARID5B*	INTRON	A/**C**	0.54	1.61(1.29-2.00)	1×10^{-5}	1.72(1.37-2.17)	2.43(1.67-3.55)
10	rs7073837	*ARID5B*	INTRON	**A**/C	0.52	1.76(1.39-2.23)	4.3×10^{-7}	2.22(1.68-2.94)	2.66(1.75-4.02)
10	rs7089424	*ARID5B*	INTRON	**G**/T	0.39	1.98(1.59-2.48)	1×10^{-9}	2.20(1.72-2.81)	3.22(2.18-4.73)
10	rs10821936	*ARID5B*	INTRON	**C**/T	0.41	1.99(1.61-2.47)	1.2×10^{-9}	2.15(1.69-2.71)	3.08(2.11-4.48)
14	rs2239633	*CEBPE*	near 5’	**C**/T	0.61	1.24(0.99-1.55)	0.06	1.26(0.99-1.58)	1.81(1.24-2.66)
7	rs11978267	*IKZF1*	INTRON	A/**G**	0.27	1.20(0.95-1.51)	0.14	1.18(0.93-1.50)	1.21(0.84-1.73)
7	rs4132601	*IKZF1*	3 UTR	**G**/T	0.27	1.22(0.97-1.54)	0.08	1.21(0.95-1.54)	1.27(0.88-1.82)

† Bolded letter indicates risk allele from previous GWAS; ‡ Odds ratios (OR) and 95% confidence intervals (CI) were calculated using log additive models adjusting for age, sex, and race.

Table 4. Odds Ratio (95%CI) for association of daycare attendance and birth order with risk of childhood ALL in non-Hispanic White and Hispanic children, CCLS, 1995-2008

	Non-Hispanic Whites					Hispanics				
	Cases		Controls		OR (95% CI)*	Cases		Controls		OR (95% CI) †
	n	%	n	%		n	%	n	%	
No. of subjects	222	38.6	353	61.4		293	42.6	395	57.4	
Censored at age 6 months										
Never	180	81.1	256	72.5	1.00 (Ref)	262	89.4	360	91.1	1.00 (Ref)
Ever	41	18.5	94	26.6	0.66 (0.43-1.01)	29	9.9	34	8.6	1.53 (0.87-2.69)
Older siblings										
No	102	45.9	127	35.9	1.00 (Ref)	106	36.2	138	34.9	1.00 (Ref)
Yes	115	51.8	216	61.2	0.65 (0.46-0.93)	182	62.1	251	63.5	0.94 (0.67-1.30)

*Odds ratios (OR) and 95% confidence intervals (CI) were calculated using unconditional logistic regression adjusting for child's age, sex and annual household income.

† Odds ratios (OR) and 95% confidence intervals (CI) were calculated using unconditional logistic regression adjusting for child's age, sex, race, and annual household income.

Table 5. Interaction between eight candidate SNPs and daycare attendance with the risk of childhood ALL by Hispanic status, CCLS, 1995-2008

			Non-Hispanic Whites		**Hispanics**	
Gene	**SNP**	**Day Care attendance** **At age 6 months**	**OR (CI)**[a]	**$P_{interaction}$**	**OR (CI)**[b]	**$P_{interaction}$**
ARID5B	rs10994982	No	1.48 (1.12-1.95)	0.49	1.60(1.26-2.04)	0.33
		Yes	1.18 (0.68-2.04)		1.08(0.51-2.31)	
	rs10740055	No	1.58(1.19-2.09)	0.83	1.64(1.29-2.09)	0.67
		Yes	1.67(0.95-2.95)		1.37(0.63-3.02)	
	rs7073837	No	1.74 (1.26-2.41)	0.09	2.16 (1.62-2.89)	0.04
		Yes	0.99 (0.56-1.76)		0.94 (0.41-2.18)	
	rs7089424	No	2.09 (1.53-2.88)	0.87	2.12(1.64-2.75)	0.88
		Yes	1.95(1.07-3.55)		2.01(0.85-4.75)	
	rs10821936	No	1.97 (1.46-2.66)	0.61	2.14 (1.67-2.74)	0.49
		Yes	1.65 (0.98-2.76)		1.69 (0.73-3.88)	
CEBPE	rs2239633	No	1.60 (1.19-2.15)	0.51	1.16 (0.92-1.48)	0.22

		Yes	1.32 (0.76-2.28)		1.83 (0.75-4.48)	
IKZF1	rs11978267	No	1.67 (1.24-2.24)	0.37	1.14 (0.88-1.46)	0.56
		Yes	2.25 (1.25-3.96)		2.48 (0.93-6.63)	
	rs4132601	No	1.65 (1.23-2.22)	0.30	1.16 (0.90-1.49)	0.30
		Yes	2.32 (1.31-4.13)		2.58 0.97-6.90)	

* Children under the age of 1 were excluded.

a. Odds ratios (OR) and 95% confidence intervals (CI) calculated using log additive model adjusting for child’s age, gender and annual household income

b. Odds ratios (OR) and 95% confidence intervals (CI) calculated using log additive model adjusting for child’s race, age, gender and annual household income

Table 6. Interaction between eight candidate SNPs and having an older sibling with the risk of childhood ALL by Hispanic status, CCLS, 1995-2008

			Non-Hispanic Whites		Hispanics	
Gene	**SNP**	**Having an older sibling**	**OR (CI)**[a]	$\mathbf{P_{interaction}}$	**OR (CI)**[b]	$\mathbf{P_{interaction}}$
ARID5B	rs10994982	No	1.29(0.88-1.91)	0.76	**1.52 (1.04-2.23)**	0.88
		Yes	1.45 (1.04-2.00)		1.52 (1.14-2.03)	
	rs10740055	No	1.51 (1.01-2.26)	0.95	1.53 (1.05-2.23)	0.77
		Yes	1.59 (1.15-2.21)		1.65 (1.23-2.20)	
	rs7073837	No	1.35 (0.86-2.08)	0.57	1.89 (1.18-3.01)	0.92
		Yes	1.59 (1.09-2.31)		1.95 (1.39-2.73)	
	rs7089424	No	1.66 (1.09-2.52)	0.17	1.82 (1.21-2.75)	0.47
		Yes	2.53 (1.71-3.75)		2.26 (1.65-3.10)	
	rs10821936	No	1.55 (1.05-2.29)	0.17	1.77 (1.22-2.59)	0.34
		Yes	2.26 (1.58-3.23)		2.27 (1.68-3.07)	
CEBPE	rs2239633	No	1.49 (0.99-2.25)	0.92	1.21 (0.81-1.89)	0.93
		Yes	1.47 (1.05-2.07)		1.22 (0.91-1.64)	

IKZF1	rs11978267	No	1.71 (1.14-2.57)	0.95	1.23 (0.82-1.83)	0.79
		Yes	1.77 (1.24-2.51)		1.13 (0.83-1.54)	
	rs4132601	No	1.68 (1.12-2.53)	0.84	1.26 (0.84-1.89)	0.78
		Yes	1.80 (1.26-2.57)		1.15 (0.84-1.57)	

*Children under the age of 1 were excluded.

a. Odds ratios (OR) and 95% confidence intervals (CI) calculated using log additive model adjusting for child’s age, gender and annual household income

b. Odds ratios (OR) and 95% confidence intervals (CI) calculated using log additive model adjusting for child’s race, age, gender and annual household income

Chapter 3
Association of Genetic Variation in *IKZF1, ARID5B,* and *CEBPE* and Surrogates for Early Life Infections with the Risk of Acute Lymphoblastic Leukemia in Hispanic Children

ABSTRACT

Background: Genome-wide association studies focusing on European-ancestry populations have identified ALL risk loci on: 7p12.2 (*IKZF1*), 10q21.2 (*ARID5B*), and 14q11.2 (*CEBPE*). This chapter extended the analyses in previous chapter (Chapter 2) and comprehensively assessed variation within these three genes to capture their impacts on ALL risk in the California Hispanic population. We also further assessed the joint effects between the genetic variation and surrogates for early life infections, including presence of older siblings, daycare attendance, and ear infections in infancy.

Methods: Genotypic data for 323 Hispanic ALL cases and 454 controls from the California Childhood Leukemia Study (CCLS) were generated using Illumina OmniExpress v1 platform. Logistic regression assuming a log-additive model estimated odds ratios (OR) associated for each SNP, adjusted for age, sex, and the first five principal components that captured population substructure. In addition, we examined potential interactions between six inherited ALL risk alleles and surrogates for early life infections using logistic regression models that included an interaction term.

Results: Significant associations between genotypes at 7p12.2 (*IKZF1*), 10q21.2 (*ARID5B*), and 14q11.2 (*CEBPE*) and ALL risk were identified; rs7780012, OR=0.50, 95% confidence interval (CI): 0.35-0.71 (P=0.004); rs7089424, OR=2.12, 95% CI: 1.70-2.65 ($P=1.16 \times 10^{-9}$); rs4982731, OR=1.69, 95% CI: 1.37-2.08 ($P=2.35 \times 10^{-6}$), respectively. Evidence for multiplicative interactions between genetic variants and surrogates for early life infections with ALL risk was not observed.

Conclusion: This is the first study to comprehensively assess the genetic variants within the three previously identified genes and evaluate their potential interactions with surrogates for early life infections in the Hispanic population. Significant main effects were detected between ALL risk and variants in *IKZF1*, *ARID5B*, and *CEBPE*. However, we did not identify significant interactions between these SNPs and surrogates for early life infections among Hispanic children.

INTRODUCTION

Leukemia is the most common malignancy under the age of 15, accounting for 31% of all cancer cases in this age group (1). Several risk factors for acute lymphoblastic leukemia (ALL) have been established, including sex, age, race, prenatal exposure to x-rays, therapeutic radiation, and specific genetic syndromes (2). Direct evidence for inherited genetic susceptibility is demonstrated by the high risk of ALL associated with Bloom's syndrome, neurofibromatosis, ataxia telangiectasia and constitutional trisomy 21(3). These predisposing disorders account for only <5% of all diagnosed cases. Most genetic association studies of ALL have focused on candidate genes, primarily those implicated in the metabolism of carcinogens, folate metabolism, immune function, and cell-cycle regulation (4, 5). Recent genome-wide association studies (GWAS) have identified common genetic variation near *IKZF1* (7p12.2), *ARID5B* (10q21.2), and *CEBPE* (14q11.2) that influences ALL risk in non-Hispanic White populations (6-9). However, only one study has explored these loci in Hispanic populations using a genome-wide approach (9). Given the observation of a high incidence of childhood ALL in California Hispanics, understanding the influence of variants within these genes with comprehensive adjustment for genetic ancestry is important for characterizing the relationship between candidate loci and ALL, and furthering our understanding of disease pathogenesis.

Epidemiologic studies have provided indirect evidence for an infectious etiology of ALL, though a specific infectious agent has yet to be identified (10). Greaves hypothesized that the absence of an early immune challenge and priming during early childhood, combined with 'delayed' exposures to infection, might subsequently result in adverse immune responses to common infectious agents, thereby increasing the risk of childhood ALL (10). This suggests that exposure to infections early in life may have a protective effect against childhood ALL. Associations of ALL have been observed with proxy measures of exposure to infection, including daycare attendance (11-14), birth order (15, 16), and child's history of infections (17, 18). The most compelling evidence of a relationship between early life infection and development of childhood ALL is from studies of daycare attendance, which is considered a surrogate for exposure to multiple microbial agents (19). The results from a recent meta-analysis of published studies conducted by Urayama *et al*. showed a significantly reduced risk of ALL among non-Hispanic Whites children who attend daycare facilities (13).

Ethnic differences in the risk of ALL are well-recognized; the incidence of ALL is nearly 20% higher among Hispanics than among non-Hispanic Whites in California (20). This higher risk is possibly due to an increased prevalence of ALL risk alleles in populations with Native American ancestry, as well as ethnic differences in exposure to environmental risk factors (10, 21). In the current study, we examine associations between ALL risk in Hispanic children and genetic variation in candidate B-cell development genes previously linked to ALL risk in GWAS of non-Hispanic White populations (*IKZF1*, *ARID5B*, and *CEBPE*). We further investigate potential interactions between these genetic variants and surrogates for early life exposure to infections, including presence of older siblings, daycare attendance, and ear infections during infancy, in California Hispanics.

MATERIALS AND METHODS

Study populations

The California Childhood Leukemia Study (CCLS) is a population-based case-control study which began in 1995. Incident cases of newly diagnosed childhood leukemia (age 0–14 years) were rapidly ascertained from major clinical centers in the study area, usually within 72 hours of diagnosis. Cases were initially identified from four hospitals (later expanded to nine) in the San Francisco Bay Area and Central Valley. For each case, one or two healthy controls matched on child's age, sex, Hispanic status (a child was considered Hispanic if either parent self-reported Hispanic) and maternal race (White, Black, Asian/Pacific Islander, Native American, and Other/Mixed) were randomly selected from the state birth registry maintained by the Center for Health Statistics of the California Department of Public Health (CDPH). A detailed description of control selection in the CCLS has been previously reported (22). A total of 86% of case subjects determined to be eligible consented to participate and 86% of controls subjects among those contacted and considered eligible participated (23).

Cases and controls were eligible to enter the study if they were under 15 years of age, resided in the study area at the time of diagnosis, had at least one parent who speaks either English or Spanish, and had no prior history of malignancy. The current analysis included 777 Hispanics (323 cases and 454 controls) in the CCLS who consented to participate and were interviewed between 1995 and 2008, and for whom archived newborn blood (ANB) spots were available. Immunophenotype was determined for ALL using flow cytometry profiles (CD10 and CD19 for B-cell lineage) (24). When fluorescence *in situ* hybridization (FISH) assays conducted at University of California Berkeley identified extra copies of chromosomes 21 and X, assignment of high hyperdiploid status (51-67 chromosomes) was made (24).

This study was reviewed and approved by the institutional review committees at the University of California Berkeley, the CDPH, and the participating hospitals. Written informed consent was obtained from all parent respondents.

Genotyping and quality control

Samples were genotyped at the Genetic Epidemiology and Genomics Laboratory, School of Public Health, University of California, Berkeley, using the Illumina OmniExpress v1 platform, which contains 730,525 markers. Quality control filtering removed SNPs that were not on autosomal chromosomes, were missing in >2% of samples, had minor allele frequency (MAF) of <2%, or showed significant deviation from Hardy-Weinberg equilibrium in controls ($P < 1 \times 10^{-5}$). The resulting data set of 634,037 SNPs was then subjected to additional quality control filtering in all samples. We excluded samples for which < 98% of loci were successfully genotyped, samples with discordant sex profiles (birth certificate vs. genetically determined sex) and samples displaying cryptic relatedness (based on identity-by-descent calculations with pi-hat cutoff of 0.15). Ten pairs of duplicate samples were included to assess assay reproducibility, with average concordance >99.99%. In the current study, we included all SNPs that passed quality control and were located within ±10 kb of selected candidate genes. 119 SNPs were included in these analyses.

Proxy measures of early childhood exposure to infections

Information on daycare attendance, birth order and history of infections was collected through in-person interview with the biological mothers. Detailed information on collection of early childhood exposure to infections was presented previously (11). Briefly, the child's birth order was determined based on a detailed pregnancy history obtained for the biological mother. Information on the child's social contacts outside the home was obtained through a history of daycare attendance before the date of diagnosis for cases and a reference date for controls, or before age six, whichever occurred first. To examine the influence of daycare attendance during a specific time window of exposure, daycare attendance was censored at six months and one year of age. For each daycare the child attended, information on age attended, duration of time attended, hours per week, and numbers of other children were obtained (11, 25). These data were used to calculated "total child-hours of exposure" for each child. Child-hours at each daycare facility was calculated as follows: (number of months attending the day care) x (mean hours per week at this day care) x (number of other children at this day care) x (4.35 weeks per month) (11, 25). In this study, we made daycare attendance into a dichotomous variable (ever/never) because the distribution of child-hour attendance was binomial for the Hispanic population included in the study.

Respondents were also asked for a history of common infectious illnesses the child had during the first year of life, including severe diarrhea/vomiting, ear infection, persistent cough, mouth and eye infection, influenza, and unspecified "other infection".

Statistical analysis

Single SNP analyses were conducted assuming a log-additive model (0, 1, or 2 copies of the minor allele), using unconditional logistic regression in PLINK v1.07 (26). To adjust for potential population stratification in study samples, a principal component analysis approach was implemented in EIGENSTRAT (27). The odds ratio (OR) and 95% confidence intervals (CI) for each SNP were calculated to estimate the risk of ALL associated with each additional copy of the minor allele, adjusted for age, sex, and the first five genetic principal components. Significance criteria based on the Benjamini and Hochberg (BH) procedure for controlling the false discovery rate (FDR) were determined by using PLINK v1.07 with a type I error rate of 5% (12). Significant regions were plotted using the online tool SNAP (28). Logistic regression adjusting for age, sex, and income was used to estimate the OR and 95% CI associated with proxies for early life infections (presence of older siblings, daycare attendance, and ear infections) and ALL risk.

Gene expression analysis

The association of genotypes and mRNA expression in *IKZF1*, *ARID5B*, and *CEBPE* was examined using mRNA data from lymphoblastoid cell lines derived from 45 MEX (Mexican ancestry in Los Angeles, California) HapMap individuals available from the database of the Gene Expression Variation (GENEVAR) project (29). Spearman's rank correlation coefficient was used to estimate the strength of the relationship between genotypes and the intensity of gene expression (30). Non-parametric permutation *P*-values were also provided to further evaluate the significance of nominal *P*-values as implemented in GENEVAR (29).

Gene-environment interaction analysis

To determine which SNPs in the three candidate regions contribute independently to disease susceptibility for subsequent geneXenvironment interaction analysis, we performed conditional haplotype analysis on all significant SNPs within the three gene regions, as implemented in PLINK v1.07. To evaluate whether the signals from these three candidate genes were independent from each other, the association between childhood ALL and each significant locus was tested using logistic regression, conditioning on all other SNPs in the region. To test for heterogeneity (interaction), we focused on the association between daycare attendance by critical development periods (age six months and one year of age), presence of older siblings, and ear infections in infancy and ALL. The joint effects of proxies for early life infections and the six SNPs selected from conditional haplotype analysis were evaluated using logistic regression, while adjusting for age, sex, income and the first five genetic principal components (PCs). The three infection variables were chosen for evaluation in the joint effect analysis based on previous CCLS publications (14). A dominant genetic model was assumed given small sample size. A *P*-value of 0.2 or less for interaction was considered statistically significant, given the available sample size, i.e. <350 observations (31). The study has 80% power to detect an interaction odds ratio of 2.0, assuming a minor allele frequency of 30 %, an exposure prevalence of 20%, an estimated OR of 1.3 for the genotype and the exposure, and a two-sided significance level of 0.05.

RESULTS

Quality control filtering yielded 777 Hispanic individuals (323 ALL cases and 454 controls) and 119 SNPs were relevant to this analysis. Study characteristics for the Hispanic participants are described in **Table 1**. The distributions of child's sex, age, and race/ethnicity were similar between cases and controls. Cases generally had lower annual household income compared to controls. Because Hispanics are a recently admixed group (32), a proportion (34%-37%) of our Hispanic population reported "Mixed or Other" race and 49% -51% of them reported "White and Caucasian" race. In our data, the frequency of the B-cell precursor (BCP) ALL was 91.9% among Hispanics (N=297) and the frequency of BCP high-hyperdiploid ALL (>50 chromosomes) for Hispanics was 30% (N=97).

Single SNP analyses

A summary of *IKZF1* gene information and childhood ALL association results for SNPs with $P<$ 0.05 (based on correction for FDR using BH procedure) is provided in **Table 2**. Ten of thirty-four SNPs examined in the *IKZF1* gene showed evidence of associations among the Hispanic population (P_{FDR} <0.05). The most significant single SNP result was rs7780012 (OR=0.50, 95% CI 0.35-0.71, $P_{\mathrm{FDR}} = 0.004$), which maps to intron 2 of the *IKZF1* gene **(Supplementary Fig. S1)**. The association remained significant when analyses were restricted to BCP ALL (OR=0.52, 95% CI 0.36-0.74) or BCP high-hyperdiploid ALL (OR=0.56, 95% CI 0.40-0.77) (**Table 5**).

A summary of *ARID5B* gene information and childhood ALL association results for SNPs with P < 0.05 (based on correction for FDR using BH procedure) is provided in **Table 3**. Eleven of fifty-seven SNPs examined in *ARID5B* showed evidence of association among the Hispanic population (P_{FDR} <0.05). The most significant single SNP association, which maps to intron 3 of the gene *ARID5B*, was rs7089424 (OR= 2.12, 95% CI 1.70-2.65, $P_{\mathrm{FDR}} = 1.16 \times 10^{-9}$) (**Supplementary Fig. S2**). The association signal remained significant when analyses were restricted to BCP ALL (OR=2.30, 95% CI 1.83-2.94) or BCP high-hyperdiploid ALL (OR=3.05, 95% CI 2.13-4.36) (**Table 5**). Two additional SNPs in *ARID5B* (rs7090445 and rs4506592) were associated with ALL at a significance level of 10^{-8} after FDR adjustment and are in linkage disequilibrium (LD) (r^2 =0.77). Interestingly, the positive associations between ALL subtypes and *ARID5B* SNPs were stronger and remained significant when analyses were restricted to ALL cases with BCP high-hyperdiploid ALL (rs7090445, OR= 3.07, 95% CI 2.14-4.39; rs4506592, OR=0.35, 95% CI 0.25-0.50) **(Table 5)**.

A summary of *CEBPE* gene information and childhood ALL association results for SNPs with $P < 0.05$ based on correction for FDR using the BH procedure is provided in **Table 4**. Eleven of twenty-eight SNPs showed evidence of association among Hispanic population (P_{FDR} <0.05). The most significant single SNP association, which maps to the 3' region of *CEBPE*, was rs4982731 (OR= 1.69, 95% CI 1.37-2.08, $P_{\mathrm{FDR}} = 2.35 \times 10^{-5}$) (**Supplementary Fig. S3**). SNP rs10143875 in *CEBPE* was also associated with ALL risk (OR= 1.67, 95% CI 1.36-2.07, P_{FDR} = 2.35×10^{-5}) and is in complete LD with rs4982731 (r^2=1). When restricted to BCP high-hyperdiploid ALL, results for variants rs4982731 and rs10143875 were more strongly associated with this ALL subtype (OR= 2.47, 95% CI 1.76-3.48; OR=2.45, 95% CI 1.75-3.45) **(Table 5)**.

Expression quantitative trait loci (eQTLs) analysis

To explore whether the observed SNP associations might influence gene expression, we investigated the correlation between the most significant SNP within three candidate genes and publicly available mRNA expression data **(Figure 1)**. Associations between both rs11980379 and rs4132601 risk genotypes and reduced *IKZF1* expression were observed (*P*=0.028 and *P* =0.029 respectively; **Figure 1**), with lower expression being associated with the risk alleles. Associated variants within *ARID5B* and *CEBPE* did not appear to influence gene expression (data not shown).

Gene-environment interaction

To minimize the numbers of statistical tests, six independent SNPs from the three genes were selected using conditional haplotype analysis for further gene-environment interaction analysis, including three SNPs from *IKZF1*, one SNP from *ARID5B*, and two SNPs from *CEBPE*. Tests of association for the six SNPs stratified by daycare attendance (ever/never) were performed using a multiplicative model of interaction to estimate the joint effects between daycare attendance for different time periods and the six genetic variants (**Supplementary Table S1 and Supplementary Table S2**). Daycare attendance censored at age six months and rs4982731in *CEBPE* showed suggestive evidence for interaction on a multiplicative scale ($P_{\text{interaction}}$=0.07; **Supplementary Table S1**); however, after controlling for multiple comparisons, the results were not significant. We did not find any evidence of interaction between the other SNPs and daycare attendance censored at age one year after FDR correction ($P_{\text{interaction}}$ >0.20). Similarly, the presence of older siblings and ear infections in infancy showed no evidence of multiplicative interactions with the genetic variants in the risk of childhood ALL (data not shown).

DISCUSSION

Our study is among the first to comprehensively assess genetic variation within the previously identified genes (*IKZF1*, *ARID5B*, and *CEBPE*) and further examine the joint effects between the genetic variation and proxies for early life infections. There was no effect modification of genetic associations by proxies for early life infections in the Hispanic population we studied. In addition, our study examined the functional relevance on gene expression levels in the Hispanic population. Among the genetic variants tested, rs11980379 and rs4132601 risk alleles within *IKZF1* were correlated with reduced *IKZF1* mRNA expression. Our results align well with previous observations of frequent *IKZF1* somatic deletion in leukemic cells, indicating that a reduction in *IKZF1* levels is pro-leukemic (6).

To date, most research on childhood ALL has focused on non-Hispanic White populations. GWAS have established that inherited genetic variants are associated with childhood ALL, including *IKZF1* (encoding the early lymphoid transcription factor IKAROS), *ARID5B* (encoding the AT-rich interactive domain 5B transcription factor), and *CEBPE* (encoding the transcription factor CCAAT/enhancer-binding protein, epsilon) (6-9, 33). Subsequently, follow-up studies have shown consistent genetic associations with the risk of childhood ALL in different populations, including a large German replication study (34), a Thai population (35), a Polish population (36), a French-Canadian cohort (37), and African American children from St. Jude Children’s Research Hospital (38).

In a previous publication, we replicated the results concerning previously identified GWAS SNPs from non-Hispanic White population in the CCLS Hispanic population (39). Our current study took a more comprehensive approach and used GWAS data to look at these three gene regions. The results are consistent with these previous published studies, indicating that causal SNPs are likely to be within or nearby these B-cell development-related genes. In particular, family members of the Ikaros gene are crucial for regulation of cell-fate decisions during hematopoiesis (40, 41). In our California Hispanic population, we were able to confirm the association between *IKZF1* and childhood ALL risk. *IKZF1*, which encodes Ikaros, is the founding member of a family of zinc finger transcription factors required for the development of all lymphoid lineages (40). *IKZF1* alterations are present in more than 70% of *BCR-ABL1* lymphoid leukemias and have been associated with poor prognoses in *BCR-ABL* ALL (42, 43). Evidence from homozygous mutant mice has shown that deletion of *IKZF1* leads to a rapid development of leukemia (44). Interestingly, the observation of a correlation between the rs4132601 genotype and *IKZF1* mRNA expression level in Epstein-Barr virus transformed lymphocytes in our study is consistent with previous findings, and with the hypothesis that the variant may influence ALL risk by affecting on early B-cell differentiation (6, 45).

We observed an association at 10q21.2 encoding the AT-rich interactive domain 5B (*ARID5B*) with childhood ALL in the Hispanic population. Given the biological heterogeneity of ALL, risk variants are likely to have differential effects on ALL risk, depending on cell lineage and phenotype (46). Subtype analysis of B-cell precursor ALL provides strong evidence that variants at 10q21.2-*ARID5B* are highly associated with the risk of developing high hyperdiploid childhood ALL in the Hispanic population, consistent with prior findings (6, 47). The strongest association with ALL is with rs7089424, which has been previously reported in the non-Hispanic White population (6, 8). Although *ARID5B* has not been studied extensively, it is highly conserved and plays a key role in embryonic development. In addition, *ARID5B* homozygous (Arid5b-/-) mice show immune abnormalities, including a reduction in the B-cell progenitor (48). In the present study, we also found an association between SNPs in *CEPBE* and childhood ALL risk. Prior studies have suggested a role for *CEBPE* (CCAAT/enhancer-binding protein, epsilon) in the development of childhood ALL. *CEBPE* is a suppressor of myeloid leukemogenesis (49, 50). *CEBPE*, along with other *CEBP* family members, is targeted by recurrent immunoglobulin heavy (IGH) translocation in B-cell precursor ALL, supporting a role in susceptibility to ALL (51).

While the risk of developing ALL may be determined by complex interactions between genetic and environmental factors (10), epidemiologic studies have thus far provided only indirect evidence that ALL has an infectious etiology, and no specific agent has been implicated (17). In the current analysis, we investigated the joint effect of the genetic variants and surrogates of exposure to early life infections on ALL risk. To minimize the number of statistical tests, six SNPs within the three genes (*IKZF1*, *ARID5B* and *CEBPE*) were selected using conditional haplotype analysis. As in previous CCLS studies, we used daycare attendance, presence of older siblings, and ear infections in infancy as proxies for early life exposures to infection (14). No evidence of interaction was observed between these proxies for early life exposure to common infections with six genetic variants based on a multiplicative scale. Compared with the non-Hispanic White population in the CCLS, Hispanic children have fewer hours of daycare attendance, have more children living in the same household, and have lower family income and parental education (data not shown). All of these factors might contribute to different patterns of childhood exposures to infection, as well as response to infections. Other measures of early life infections among Hispanics, such as the total number of people living in the household at the time of child's birth and/or parental or other child's social contacts, may be better proxies. However, such measures are not available in our study.

In an admixed population such as Hispanics, population stratification is a potential problem. However, its effect is likely to be minimal in the CCLS, due to the careful matching of race and ethnicity among the subjects (52). Further, we used a principal components analysis (PCA) to successfully reduce the effects of potential population stratification (genomic control factor λ = 1.02) (**Supplementary Fig. S4**) (27). Another major strength of the present study is the detailed assessment of early life infections exposures, such as daycare attendance. We calculated a composite variable "child-hours" to measure how long each individual attended a daycare facility and the number of other children at the facility (11). However, the distribution of child-hour attendance is binomial for the Hispanic population included in the study; therefore, we used a dichotomous daycare attendance variable (ever/never) in the analysis.

One limitation of this study for assessing gene-environment interaction is the sample size. Even though the CCLS is one of the largest case-control studies of ALL in the United States with relevant biospecimens and environmental data, gene-environment interactions with moderate to small effect sizes may not be detected. Another potential limitation of the gene-environment analyses in our study is the influence of uncontrolled or residual confounding on risk estimates. The consideration of other surrogate indicators, such as the total number of children/adults living in the household or serologic evidence of prior infection, may be more suitable to address the gene-environment interaction in the Hispanic population. Furthermore, while performing gene-environment analyses, we focused on only six SNPs selected from conditional haplotype analysis; however, it is possible that other SNPs within the genes or the combination of these SNPs might have a joint effect with daycare attendance on ALL risk.

Our study showed that variants within *IKZF1*, *ARID5B*, and *CEBPE* were associated with increased ALL risk, and the effects for *ARID5B* and *CEBPE* were most prominent in the high-hyperdiploid ALL subtype in the California Hispanic population. Even though we did not observe significant gene-environment interactions in the study, identification of interactions between genetic variants and environmental risk factors may require much larger datasets. Replication studies with larger sample sizes in other Hispanic populations will be desirable to extend our results. Additional functional studies and re-sequencing of the relevant genetic regions are needed to better understand the role of these genes in early hematopoiesis.

Grant support

This study was supported by grants from the National Institute of Environmental Health Sciences (PS42 ES04705 and R01 ES09137), the National Cancer Institute (R25CA112355), and Children with Cancer, United Kingdom.

Acknowledgements

Participating hospitals and clinical collaborators included University of California Davis Medical Center (Dr Jonathan Ducore), University of California San Francisco (Dr Mignon Loh and Dr 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 Sacramento (Dr Vincent Kiley), Kaiser Permanente Santa Clara (Dr Carolyn Russo and Dr Alan Wong), Kaiser Permanente San Francisco (Dr Kenneth Leung), Children's Hospital of Los Angeles (Dr Cecilia Fu) and Kaiser Permanente Oakland (Dr Stacy Month). We also acknowledge our collaborators at the Northern California Cancer Center and the entire California Childhood Leukemia Study staff for their effort and dedication.

REFERENCES

1. Kaatsch P. Epidemiology of childhood cancer. Cancer Treat Rev 2010;36(4):277-85.
2. Pui CH, Relling MV, Downing JR. Acute lymphoblastic leukemia. N Engl J Med 2004;350(15):1535-48.
3. Eden T. Aetiology of childhood leukaemia. Cancer Treat Rev 2010;36(4):286-97.
4. Urayama KY, Chokkalingam AP, Manabe A, Mizutani S. Current evidence for an inherited genetic basis of childhood acute lymphoblastic leukemia. Int J Hematol 2013;97(1):3-19.
5. Vijayakrishnan J, Houlston RS. Candidate gene association studies and risk of childhood acute lymphoblastic leukemia: a systematic review and meta-analysis. Haematologica 2010;95(8):1405-14.
6. Papaemmanuil E, Hosking FJ, Vijayakrishnan J, Price A, Olver B, Sheridan E, et al. Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. Nat Genet 2009;41(9):1006-10.
7. Orsi L, Rudant J, Bonaventure A, Goujon-Bellec S, Corda E, Evans TJ, et al. Genetic polymorphisms and childhood acute lymphoblastic leukemia: GWAS of the ESCALE study (SFCE). Leukemia 2012;26(12):2561-4.
8. Trevino LR, Yang W, French D, Hunger SP, Carroll WL, Devidas M, et al. Germline genomic variants associated with childhood acute lymphoblastic leukemia. Nat Genet 2009;41(9):1001-5.
9. Xu H, Yang W, Perez-Andreu V, Devidas M, Fan Y, Cheng C, et al. Novel Susceptibility Variants at 10p12.31-12.2 for Childhood Acute Lymphoblastic Leukemia in Ethnically Diverse Populations. J Natl Cancer Inst 2013.
10. Greaves M. Infection, immune responses and the aetiology of childhood leukaemia. Nat Rev Cancer 2006;6(3):193-203.
11. Ma X, Buffler PA, Wiemels JL, Selvin S, Metayer C, Loh M, et al. Ethnic difference in daycare attendance, early infections, and risk of childhood acute lymphoblastic leukemia. Cancer Epidemiol Biomarkers Prev 2005;14(8):1928-34.
12. Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. Stat Med 1990;9(7):811-8.
13. Urayama KY, Buffler PA, Gallagher ER, Ayoob JM, Ma X. A meta-analysis of the association between day-care attendance and childhood acute lymphoblastic leukaemia. Int J Epidemiol 2010;39(3):718-32.
14. Urayama KY, Ma X, Selvin S, Metayer C, Chokkalingam AP, Wiemels JL, et al. Early life exposure to infections and risk of childhood acute lymphoblastic leukemia. Int J Cancer 2011;128(7):1632-43.
15. Westergaard T, Andersen PK, Pedersen JB, Olsen JH, Frisch M, Sorensen HT, et al. Birth characteristics, sibling patterns, and acute leukemia risk in childhood: a population-based cohort study. J Natl Cancer Inst 1997;89(13):939-47.
16. Dockerty JD, Draper G, Vincent T, Rowan SD, Bunch KJ. Case-control study of parental age, parity and socioeconomic level in relation to childhood cancers. Int J Epidemiol 2001;30(6):1428-37.
17. Roman E, Simpson J, Ansell P, Kinsey S, Mitchell CD, McKinney PA, 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 2007;165(5):496-504.

18. Rudant J, Orsi L, Menegaux F, Petit A, Baruchel A, Bertrand Y, et al. Childhood acute leukemia, early common infections, and allergy: The ESCALE Study. Am J Epidemiol 2010;172(9):1015-27.
19. O'Connor SM, Boneva RS. Infectious etiologies of childhood leukemia: plausibility and challenges to proof. Environ Health Perspect 2007;115(1):146-50.
20. Campleman SL WW. Childhood cancer in California 1988 to 1999 Volume I: birth to age 14. Sacramento, CA: California Department of Health Services, Cancer Surveillance Section. 2004:16-17.
21. Walsh KM, Chokkalingam AP, Hsu LI, Metayer C, de Smith AJ, Jacobs DI, et al. Associations between genome-wide Native American ancestry, known risk alleles and B-cell ALL risk in Hispanic children. Leukemia 2013.
22. Ma X, Buffler PA, Layefsky M, Does MB, Reynolds P. Control selection strategies in case-control studies of childhood diseases. Am J Epidemiol 2004;159(10):915-21.
23. Bartley K, Metayer C, Selvin S, Ducore J, Buffler P. Diagnostic X-rays and risk of childhood leukaemia. Int J Epidemiol 2010;39(6):1628-37.
24. Aldrich MC, Zhang L, Wiemels JL, Ma X, Loh ML, Metayer C, et al. Cytogenetics of Hispanic and White children with acute lymphoblastic leukemia in California. Cancer Epidemiol Biomarkers Prev 2006;15(3):578-81.
25. Ma X, Buffler PA, Selvin S, Matthay KK, Wiencke JK, Wiemels JL, et al. Daycare attendance and risk of childhood acute lymphoblastic leukaemia. Br J Cancer 2002;86(9):1419-24.
26. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81(3):559-75.
27. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38(8):904-9.
28. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PI. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics 2008;24(24):2938-9.
29. Yang TP, Beazley C, Montgomery SB, Dimas AS, Gutierrez-Arcelus M, Stranger BE, et al. Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. Bioinformatics 2010;26(19):2474-6.
30. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, et al. Population genomics of human gene expression. Nat Genet 2007;39(10):1217-24.
31. Jewell N. Statistics for Epidemiology. Boca Raton, Florida: Chapman&Hall/CRC; 2004.
32. Baye TM, Wilke RA. Mapping genes that predict treatment outcome in admixed populations. Pharmacogenomics J 2010;10(6):465-77.
33. Han S, Lee KM, Park SK, Lee JE, Ahn HS, Shin HY, et al. Genome-wide association study of childhood acute lymphoblastic leukemia in Korea. Leuk Res 2010;34(10):1271-4.
34. Prasad RB, Hosking FJ, Vijayakrishnan J, Papaemmanuil E, Koehler R, Greaves M, et al. Verification of the susceptibility loci on 7p12.2, 10q21.2, and 14q11.2 in precursor B-cell acute lymphoblastic leukemia of childhood. Blood 2010;115(9):1765-7.
35. Vijayakrishnan J, Sherborne AL, Sawangpanich R, Hongeng S, Houlston RS, Pakakasama S. Variation at 7p12.2 and 10q21.2 influences childhood acute lymphoblastic

leukemia risk in the Thai population and may contribute to racial differences in leukemia incidence. Leuk Lymphoma 2010;51(10):1870-4.

36. Pastorczak A, Gorniak P, Sherborne A, Hosking F, Trelinska J, Lejman M, et al. Role of 657del5 NBN mutation and 7p12.2 (IKZF1), 9p21 (CDKN2A), 10q21.2 (ARID5B) and 14q11.2 (CEBPE) variation and risk of childhood ALL in the Polish population. Leuk Res 2011;35(11):1534-6.

37. Healy J, Richer C, Bourgey M, Kritikou EA, Sinnett D. Replication analysis confirms the association of ARID5B with childhood B-cell acute lymphoblastic leukemia. Haematologica 2010;95(9):1608-11.

38. Yang W, Trevino LR, Yang JJ, Scheet P, Pui CH, Evans WE, et al. ARID5B SNP rs10821936 is associated with risk of childhood acute lymphoblastic leukemia in blacks and contributes to racial differences in leukemia incidence. Leukemia 2010;24(4):894-6.

39. Chokkalingam AP, Hsu LI, Metayer C, Hansen HM, Month SR, Barcellos LF, et al. Genetic variants in ARID5B and CEBPE are childhood ALL susceptibility loci in Hispanics. Cancer Causes Control 2013;24(10):1789-95.

40. John LB, Ward AC. The Ikaros gene family: transcriptional regulators of hematopoiesis and immunity. Mol Immunol 2011;48(9-10):1272-8.

41. Schmitt C, Tonnelle C, Dalloul A, Chabannon C, Debre P, Rebollo A. Aiolos and Ikaros: regulators of lymphocyte development, homeostasis and lymphoproliferation. Apoptosis 2002;7(3):277-84.

42. Kuiper RP, Waanders E, van der Velden VH, van Reijmersdal SV, Venkatachalam R, Scheijen B, et al. IKZF1 deletions predict relapse in uniformly treated pediatric precursor B-ALL. Leukemia 2010;24(7):1258-64.

43. Mullighan CG, Su X, Zhang J, Radtke I, Phillips LA, Miller CB, et al. Deletion of IKZF1 and prognosis in acute lymphoblastic leukemia. N Engl J Med 2009;360(5):470-80.

44. Virely C, Moulin S, Cobaleda C, Lasgi C, Alberdi A, Soulier J, et al. Haploinsufficiency of the IKZF1 (IKAROS) tumor suppressor gene cooperates with BCR-ABL in a transgenic model of acute lymphoblastic leukemia. Leukemia 2010;24(6):1200-4.

45. Greaves MF, Wiemels J. Origins of chromosome translocations in childhood leukaemia. Nat Rev Cancer 2003;3(9):639-49.

46. Walsh KM, de Smith AJ, Chokkalingam AP, Metayer C, Dahl GV, Hsu LI, et al. Novel childhood ALL susceptibility locus BMI1-PIP4K2A is specifically associated with the hyperdiploid subtype. Blood 2013;121(23):4808-9.

47. Paulsson K, Forestier E, Lilljebjorn H, Heldrup J, Behrendtz M, Young BD, et al. Genetic landscape of high hyperdiploid childhood acute lymphoblastic leukemia. Proc Natl Acad Sci U S A 2010;107(50):21719-24.

48. Sherborne AL, Houlston RS. What are genome-wide association studies telling us about B-cell tumor development? Oncotarget 2010;1(5):367-72.

49. Akagi T, Thoennissen NH, George A, Crooks G, Song JH, Okamoto R, et al. In vivo deficiency of both C/EBPbeta and C/EBPepsilon results in highly defective myeloid differentiation and lack of cytokine response. PLoS One 2010;5(11):e15419.

50. Bedi R, Du J, Sharma AK, Gomes I, Ackerman SJ. Human C/EBP-epsilon activator and repressor isoforms differentially reprogram myeloid lineage commitment and differentiation. Blood 2009;113(2):317-27.

51. Akasaka T, Balasas T, Russell LJ, Sugimoto KJ, Majid A, Walewska R, et al. Five members of the CEBP transcription factor family are targeted by recurrent IGH translocations in B-cell precursor acute lymphoblastic leukemia (BCP-ALL). Blood 2007;109(8):3451-61.

52. Chokkalingam AP AM, Bartley K, Hsu LI, Metayer C, et al. Matching on Race and Ethnicity in Case-Control Studies as a Means of Control for Population Stratification. Epidemiol 2011;1:101.

Table 1. Characteristics of Hispanic case-control study subjects, CCLS, 1995-2008

	Cases, n (%)	Controls, n (%)
Study subjects	323 (41.6)	454 (58.4)
Sex		
Male	173 (53.6)	240 (52.9)
Female	150 (46.4)	214 (47.1)
Age		
Mean age, y(SE)	5.3(3.4)	5.3 (3.4)
Income		
<$15,000	79 (24.5)	74 (16.3)
$15,000-$29,999	88 (27.2)	106 (23.3)
$30,000-$44,999	64 (19.8)	87 (19.2)
$45,000-$59,999	41 (12.7)	63 (13.9)
$60,000-$74,999	17 (5.3)	42 (9.3)
≥$75,000	34 (10.5)	82 (18.0)
Race		
White/Caucasian	161 (49.8)	237 (52.2)
African American	14 (4.3)	15 (3.3)
Native American	0 (0)	4 (0.9)
Asian or Pacific Islander	28 (8.7)	40 (8.8)
Mixed or others	120 (37.2)	158 (34.8)
Cytogenetics (case-only)		
B-cell precursor (BCP) ALL	297 (91.9)	-
BCP high-hyperdiploid ALL (>50 chromosome)	97 (30.0)	-
Daycare attendance (Mean and SE)		
Thousand child hours before 6 mo of age (SE)	0.24 (1.15)	0.18 (0.85)
Thousand child hours before 1 yr of age (SE)	0.94 (3.83)	0.65 (2.46)

Table 2. Odds ratio (95%CI) and p-values for significant[*] *IKZF1* SNPs associated with childhood ALL risk in the Hispanic population, CCLS, 1995-2008

Chr.	SNP	Minor Allele (frequency)	Base-pair location	OR[a]	95% CI	P_{values}[b]	P_{values}[c]
7	rs7780012	A (0.12)	50438720	0.50	0.35, 0.71	1.31×10^{-04}	0.004
7	rs11980379	G (0.28)	50469981	1.47	1.17, 1.85	9.31×10^{-04}	0.011
7	rs4132601	C (0.28)	50470604	1.47	1.17, 1.85	9.31×10^{-04}	0.011
7	rs716719	A (0.46)	50325717	1.38	1.12, 1.69	0.002	0.018
7	rs6952409	A (0.23)	50462935	1.41	1.11, 1.78	0.004	0.029
7	rs6964823	A (0.31)	50460096	0.73	0.59, 0.91	0.006	0.033
7	rs7781977	G (0.49)	50346134	0.75	0.61, 0.93	0.007	0.034
7	rs9886239	A (0.48)	50336551	0.76	0.62, 0.93	0.008	0.035
7	rs4917017	G (0.49)	50335232	0.76	0.62, 0.94	0.009	0.035
7	rs12719019	A (0.28)	50476139	0.75	0.60, 0.95	0.014	0.049

a. Odds ratio (OR) and 95% confidence interval (CI) calculated using log additive models adjusting for age, sex and first 5 genetic principal components.

b. P-values calculated using log additive models adjusting for age, sex and first 5 genetic principal components (PCs) without adjustment for multiple comparisons.

c. P-values based on correction for False Discovery Rate (FDR) using Benjamini and Hochberg (BH) procedure.

* Ten of thirty-four SNPs had significant P-values adjusted for FDR.

Table3. Odds ratio (95% CI) and p-values for significant* *ARID5B* SNPs associated with childhood ALL risk in the Hispanic population, CCLS, 1995-2008

Chr.	SNP	Minor Allele (frequency)	Base-pair location	OR[a]	95% CI	P_{values}[b]	P_{values}[c]
10	rs7089424	C (0.49)	63752159	2.12	1.70, 2.65	2.04×10^{-11}	1.16×10^{-09}
10	rs7090445	G (0.49)	63721176	2.08	1.67, 2.60	6.24×10^{-11}	1.57×10^{-09}
10	rs4506592	G (0.49)	63727187	0.48	0.38, 0.60	8.24×10^{-11}	1.57×10^{-09}
10	rs7073837	C (0.48)	63699895	0.55	0.45, 0.69	6.84×10^{-08}	9.75×10^{-07}
10	rs10821938	C (0.45)	63724773	0.56	0.45, 0.69	1.24×10^{-07}	1.42×10^{-06}
10	rs10994981	A (0.41)	63708007	0.65	0.52, 0.80	7.94×10^{-05}	7.54×10^{-04}
10	rs6479778	A (0.29)	63689077	1.52	1.21, 1.90	3.26×10^{-04}	0.002
10	rs2893881	G (0.30)	63688672	1.47	1.18, 1.84	7.59×10^{-04}	0.005
10	rs4948491	G (0.46)	63696889	1.43	1.16, 1.76	9.41×10^{-04}	0.006
10	rs4948488	G (0.40)	63685154	1.44	1.16, 1.79	9.67×10^{-04}	0.006
10	rs12249208	A (0.03)	63730012	0.36	0.18, 0.72	0.004	0.019

a. Odds ratio (OR) and 95% confidence interval (CI) calculated using log additive models adjusting for age, sex and the first 5 genetic principal components.

b. P-values calculated using log additive models adjusting for age, sex and first 5 genetic principal components (PCs) without adjustment for multiple comparisons.

c. P-values based on correction for False Discovery Rate (FDR) using Benjamini and Hochberg (BH) procedure.

* Eleven of fifty-seven SNPs had significant P-values adjusted for FDR.

Table 4. Odds ratio (95% CI) and p-values for significant[*] *CEBPE* SNPs associated with childhood ALL risk in the Hispanic population, CCLS, 1995-2008

Chr.	SNP	Minor Allele (frequency)	Base-pair location	OR[a]	95% CI	P_{values}[b]	P_{values}[c]
14	rs4982731	G (0.41)	23585333	1.69	1.37, 2.08	1.15×10^{-06}	2.35×10^{-05}
14	rs10143875	A (0.41)	23584265	1.67	1.36, 2.07	1.68×10^{-06}	2.35×10^{-05}
14	rs6572981	A (0.18)	23599252	0.53	0.39, 0.70	1.64×10^{-05}	1.23×10^{-04}
14	rs2144827	A (0.14)	23587231	0.50	0.36, 0.68	1.76×10^{-05}	1.23×10^{-04}
14	rs17794251	A (0.41)	23593442	1.53	1.24, 1.88	6.50×10^{-05}	3.64×10^{-04}
14	rs7155790	C (0.27)	23589586	0.62	0.49, 0.79	9.06×10^{-05}	4.23×10^{-04}
14	rs2236135	G (0.11)	23595721	0.56	0.39, 0.80	0.001	0.005
14	rs2239629	C (0.11)	23598128	0.57	0.40, 0.81	0.002	0.005
14	rs17198995	G (0.05)	23595423	0.44	0.26, 0.73	0.002	0.005
14	rs2180395	A (0.10)	23596621	0.56	0.39, 0.80	0.002	0.005
14	rs2073305	A (0.09)	23601073	0.59	0.41, 0.85	0.005	0.012

a. Odds ratio (OR) and 95% confidence interval (CI) calculated using log additive models adjusting for age, sex and the first 5 genetic principal components.

b. P-values calculated using log additive models adjusting for age, sex and first 5 genetic principal components (PCs) without adjustment for multiple comparisons.

c. P-values based on correction for False Discovery Rate (FDR) using Benjamini and Hochberg (BH) procedure.

* Eleven of twenty-eight SNPs had significant P-values adjusted for FDR.

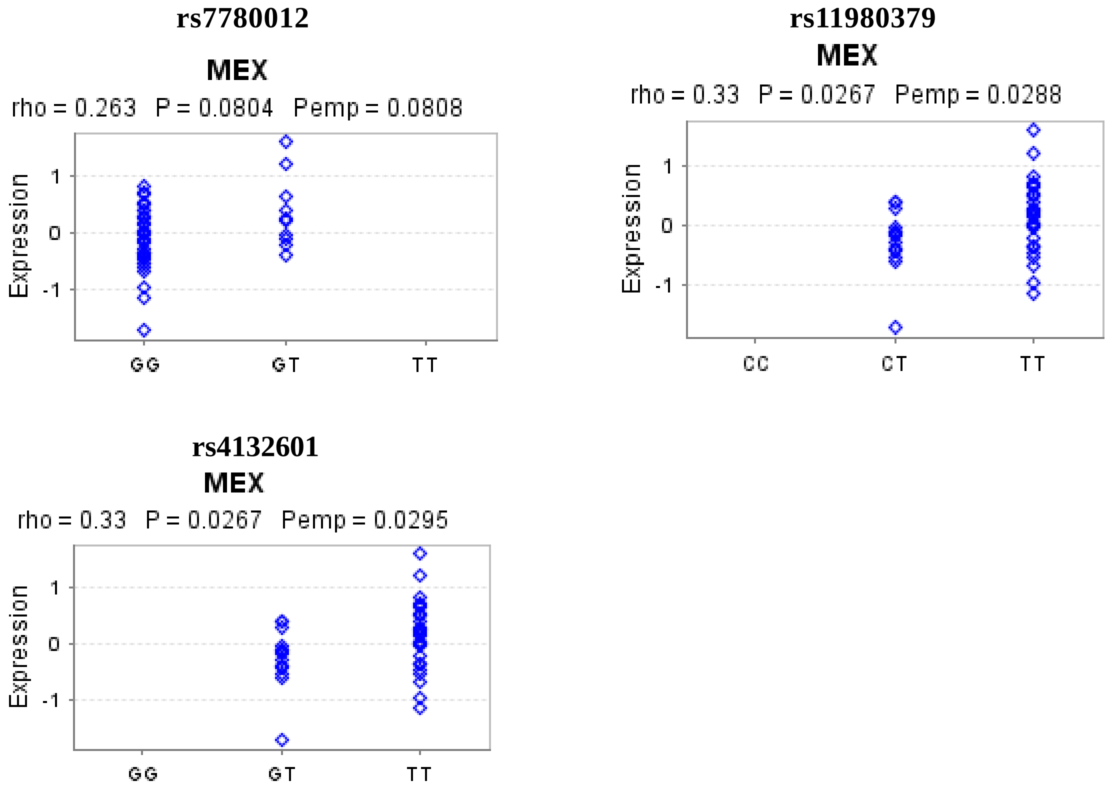
Table 5. Odds ratio (95% CI) for association with ALL by subsets of significant[*] SNPs in *ARID5B*, *CEBPE* and *IKZF1* and immunological subtypes of B-cell precursor (BCP) ALL and BCP high-hyperdiploid ALL in the Hispanic population, CCLS, 1995-2008

Gene	ALL (n=323)	B-cell precursor (BCP) ALL (n=297)	BCP high-hyperdiploid ALL (n=97)
	OR[a] (95% CI)	OR[a] (95% CI)	OR[a] (95% CI)
IKZF1			
rs7780012	0.50 (0.35, 0.71)	0.52 (0.36,0.74)	0.56 (0.40,0.77)
rs11980379	1.47 (1.17, 1.85)	1.47 (1.17,1.86)	1.50 (1.06,2.12)
rs4132601	1.47 (1.17, 1.85)	1.47 (1.17,1.86)	1.50 (1.06,2.11)
rs716719	1.38 (1.12, 1.69)	1.37 (1.11,1.69)	1.72(1.25,2.38)
rs6952409	1.41 (1.11, 1.78)	1.40 (1.10,1.78)	1.42 (0.99,2.02)
rs694823	0.73 (0.59, 0.91)	0.73 (0.58-0.91)	0.67 (0.47,0.95)
ARID5B			
rs7089424	2.12 (1.70, 2.65)	2.30(1.83, 2.94)	3.05 (2.13,4.36)
rs7090445	2.08 (1.67, 2.60)	2.24 (1.78,2.81)	3.07 (2.14,4.39)
rs4506592	0.48 (0.38, 0.60)	0.45 (0.36,0.56)	0.35 (0.25,0.50)
rs7073837	0.55 (0.45, 0.69)	0.51 (0.41,0.64)	0.42 (0.30,0.60)
rs10821938	0.56 (0.45, 0.69)	0.53 (0.42,0.66)	0.42 (0.30-0.60)
rs10994981	0.65 (0.52, 0.80)	0.61 (0.49,0.76)	0.49 (0.35-0.70)
rs6479778	1.52 (1.21,1.90)	1.55 (1.23,1.95)	1.77 (1.26-2.46)
rs2893881	1.47(1.18,1.84)	1.51 (1.20,1.89)	1.71 (1.20-2.42)
CEBPE			
rs4982731	1.69 (1.37,2.08)	1.77 (1.43,2.20)	2.47 (1.76,3.48)
rs10143875	1.67 (1.36,2.07)	1.75 (1.41,2.18)	2.45(1.75,3.45)
rs6572981	0.53 (0.39,0.70)	0.53 (0.39,0.71)	0.40 (0.24,0.67)
rs2144827	0.50 (0.36,0.68)	0.49 (0.35,0.68)	0.32 (0.17,0.60)
rs17794251	1.53 (1.24,1.88)	1.54 (1.24,1.89)	2.00 (1.44,2.78)

a. Odds ratio (OR) and 95% confidence interval (CI) calculated using log additive models adjusting for age, sex and top 5 genetic principal components.

* SNPs had significant p-values adjusted for FDR with ALL risk and immunological subtypes of BCP ALL and BCP high-hyperdiploid ALL.

Figure 1: Correlation between rs7780012, rs11980379, and rs4132601 genotypes and *IKZF1* expression in HapMap MEX dataset*


rs7780012

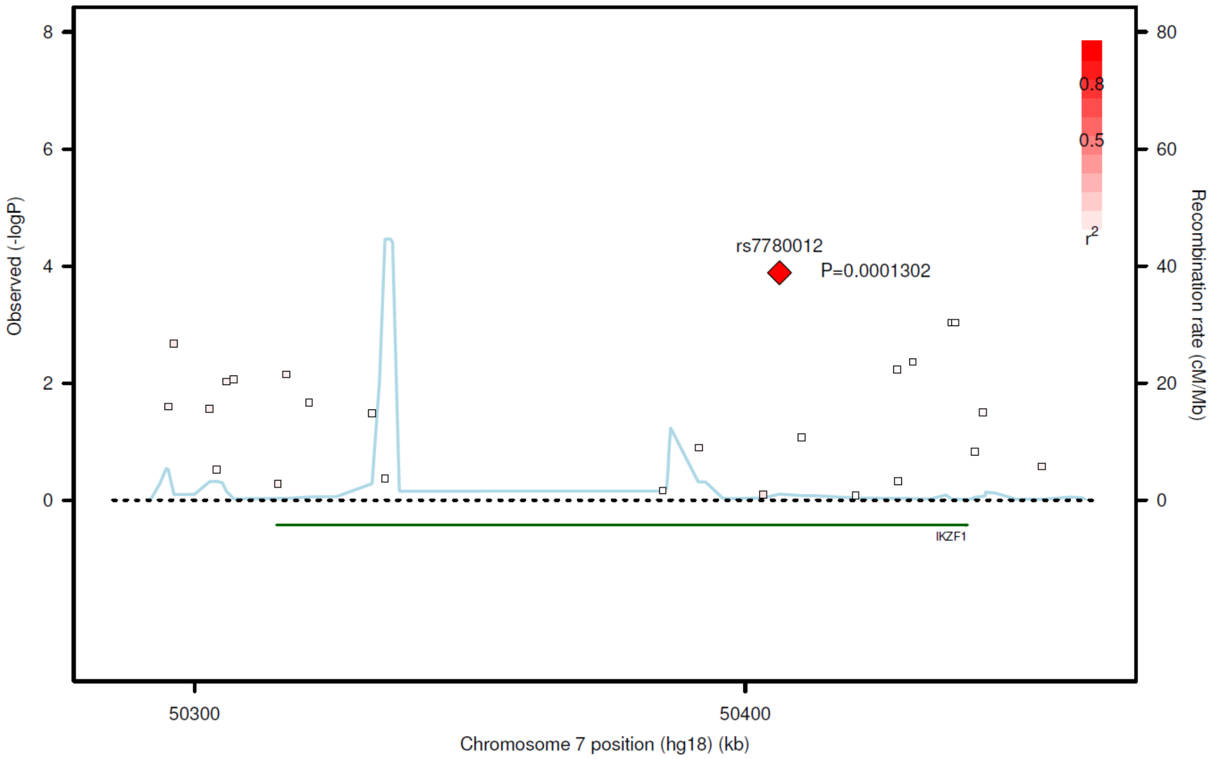
rs11980379

rs4132601

The correlation between normalized *IKZF1* gene expression and alleles of rs7780012, rs11980379 and rs4132601 were examined using Spearman rank correlation.

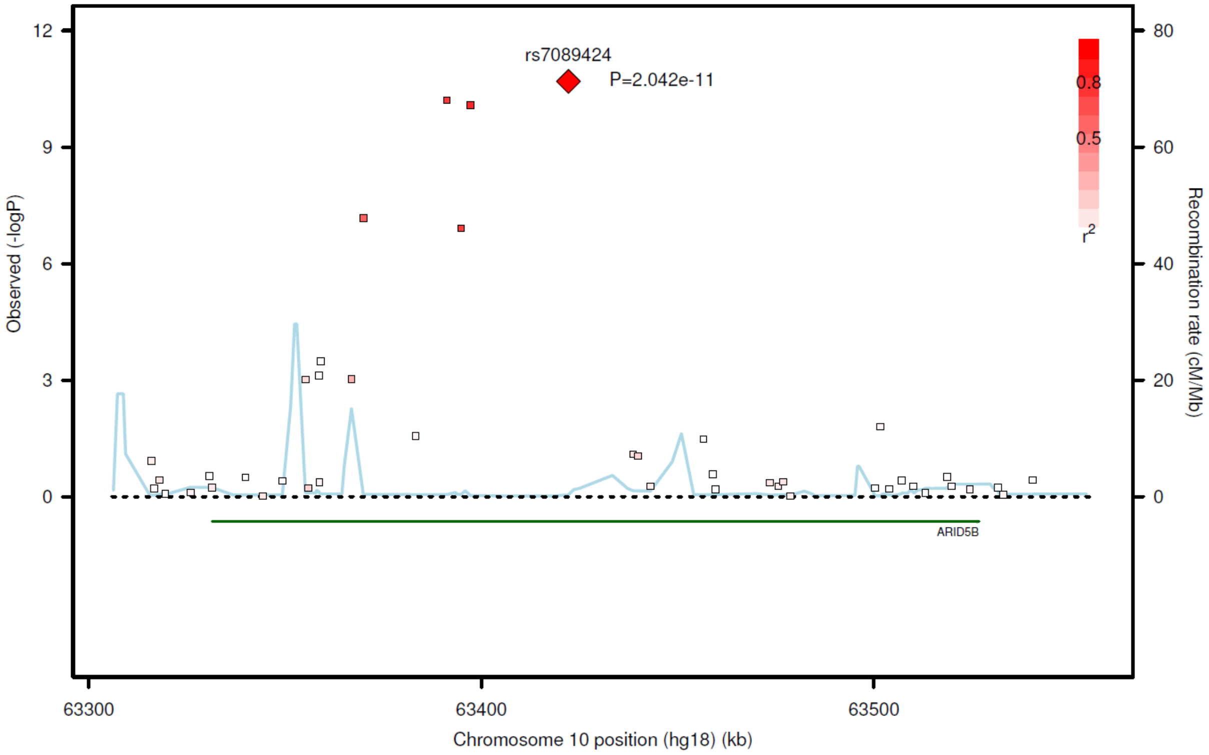
* Publicly available Sentrix Human-6 Expression BeadChips (Illumina) data on 45 lymphoblastoid cell lines derived from healthy Mexican ancestry (MEX) in Los Angeles, California as part of the HapMap 3 project as of June, 2009 (GENEVAR project).
Pemp=Empirical p-value derived from 10,000 permutations

Supplementary Fig. S1: Association and recombination plots for *IKZF1* (7p12.2 loci) associated with childhood ALL risk in the Hispanic population, CCLS, 1995-2008



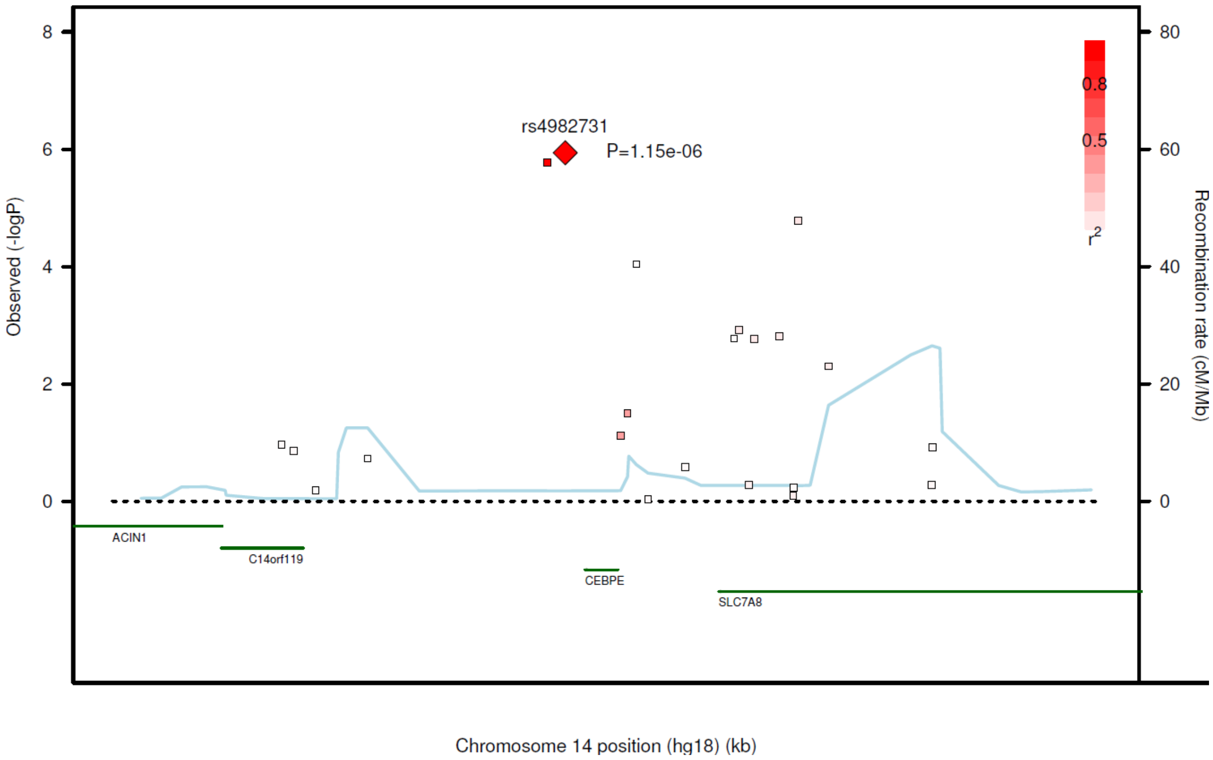
Genotyped SNPs are represented as squares and diamonds and a larger diamond indicated the top-hit association in each region. The strength of linkage disequilibrium between each SNPs and the top hit is indicated by the color of the symbol. Recombination rates, plotted in light blue, are based on HapMap3 MEX samples, and genomic coordinates are based on National Center for Biotechnology Information (NCBI) Build 36 of the human genome (released in March 2006).

Supplementary Fig. S2: Association and recombination plots for *ARID5B* (10q21.2 loci) associated with childhood ALL risk in the Hispanic population, CCLS, 1995-2008



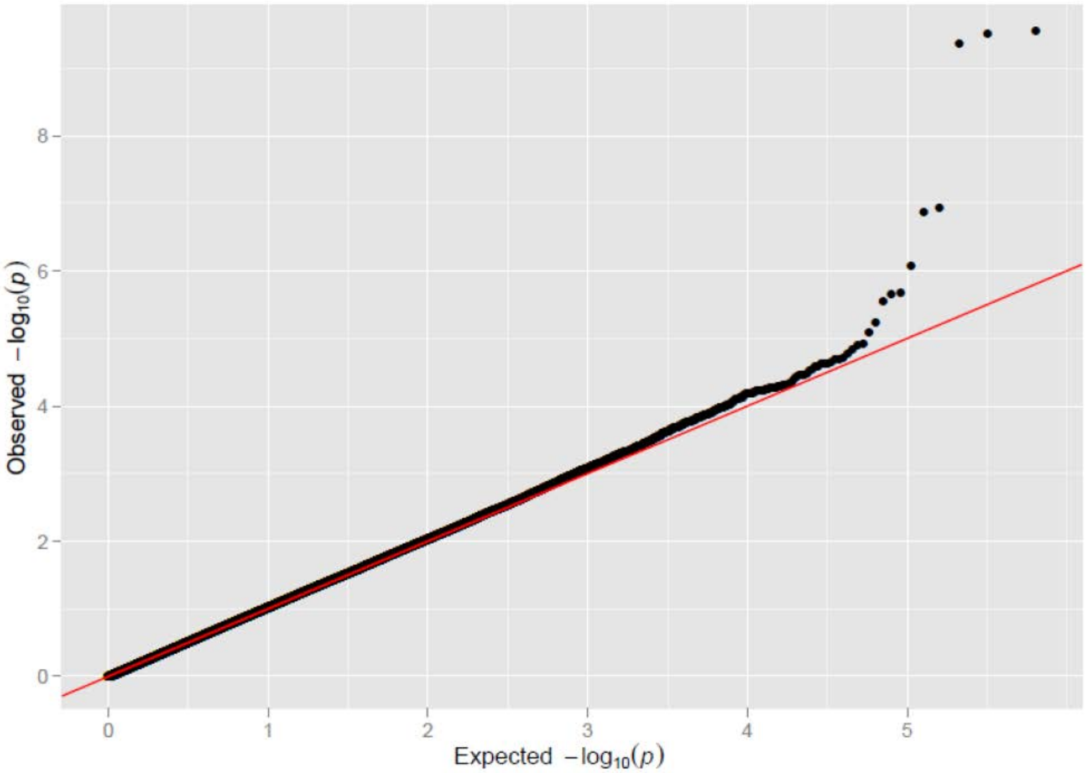
Genotyped SNPs are represented as squares and diamonds and a larger diamond indicated the top-hit association in each region. The strength of linkage disequilibrium between each SNPs and the top hit is indicated by the color of the symbol. Recombination rates, plotted in light blue, are based on HapMap3 MEX samples, and genomic coordinates are based on National Center for Biotechnology Information (NCBI) Build 36 of the human genome (released in March 2006).

Supplementary Fig. S3: Association and recombination plots for *CEBPE* (14q11.2 loci) associated with childhood ALL risk in the Hispanic population, CCLS, 1995-2008



Genotyped SNPs are represented as squares and diamonds and a larger diamond indicated the top-hit association in each region. The strength of linkage disequilibrium between each SNPs and the top hit is indicated by the color of the symbol. Recombination rates, plotted in light blue, are based on HapMap3 MEX samples, and genomic coordinates are based on National Center for Biotechnology Information (NCBI) Build 36 of the human genome (released in March 2006).

Supplementary Fig. S4. QQ-plot for Hispanic genome wide data



Quantile-quantile (Q-Q) plot comparing the distributing of the observed versus the expected –log10 P values (log-additive model, adjusted for age and sex) of the 634,037 SNPs. The black line indicates distribution expected under the null hypothesis of no association and the red line indicates the inflation of the test statistics (λ=1.02)

Supplementary Table S1: P-values for gene by environment interaction of six candidate SNPs[†] and daycare attendance by age of 6 months with the risk of ALL in Hispanic Children, CCLS, 1995-2008

Gene	SNP	Censored at Age 6 months	Cases N‡	Controls N‡	OR (95% CI)[a]	$P_{interaction}$
CEBPE	rs4982731	No	81/198	162/237	1.64 (1.17-2.29)	0.07
		Yes	4/30	16/22	6.42 (1.56-26.48)	
CEBPE	rs17794251	No	85/194	167/231	1.65 (1.17-2.31)	0.99
		Yes	9/25	14/24	2.45 (0.64-9.42)	
ARID5B	rs4506592	No	101/175	66/333	0.35 (0.24-0.51)	0.70
		Yes	13/21	8/30	0.35 (0.09-1.31)	
IKZF1	rs4132601	No	128/151	226/173	1.55 (1.14-2.13)	0.71
		Yes	15/19	22/16	2.36 (0.69-8.03)	
IKZF1	rs6964823	No	165/104	185/214	0.59 (0.43-0.82)	0.61
		Yes	14/20	15/23	0.59 (0.17-2.09)	
IKZF1	rs4917017	No	93/186	101/298	0.69 (0.49-0.97)	0.99
		Yes	6/28	4/34	0.93 (0.18-4.73)	

a. Odds ratios (OR) and 95% confidence intervals (CI) calculated using logistic regression model adjusting for child's age, gender, annual household income and the first five genetic principal components.

† Six candidate SNPs were selected from all significant SNPs using a conditional haplotype analysis.

‡ Dominant models used to provide counts of cases with wild genotype versus having any copy of minor alleles and counts of controls with wild genotype versus having any copy of minor alleles by daycare attendance by 6 months (yes/no).

Supplementary Table S2. P-values for gene by environment interaction of six candidate SNPs[†] and daycare attendance by age one with the risk of ALL in Hispanic Children, CCLS, 1995-2008

Gene	SNP	Censored at Age one	Cases N‡	Controls N‡	OR (95% CI)[a]	$P_{interaction}$
CEBPE	rs4982731	No	78/192	153/223	1.65 (1.17-2.33)	0.15
		Yes	7/36	25/36	3.08 (1.07-8.85)	
	rs17794251	No	81/89	155/220	1.62 (1.15-2.29)	0.78
		Yes	13/30	26/35	2.12 (0.78-5.76)	
ARID5B	rs4506592	No	98/172	65/311	0.37 (0.25-0.53)	0.70
		Yes	12/27	9/52	0.26 (0.08-0.82)	
IKZF1	rs4132601	No	123/147	216/160	1.62 (1.17-2.23)	0.74
		Yes	20/23	32/29	1.32 (0.52-3.33)	
	rs6964823	No	162/108	175/200	0.58 (0.42-0.81)	0.35
		Yes	17/26	24/37	1.13 (0.42-3.02)	
	rs4917017	No	91/179	94/282	0.67 (0.47-0.96)	0.65
		Yes	8/35	11/50	1.29 0.42-4.02)	

a. Odds ratios (OR) and 95% confidence intervals (CI) calculated using logistic regression model adjusting for child's age, gender, annual household income and the first five genetic principal components.

† Six candidate SNPs were selected from all significant SNPs using a conditional haplotype analysis.

‡ Dominant models used to provide counts of cases with wild genotype versus having any copy of minor alleles and counts of controls with wild genotype versus having any copy of minor alleles by daycare attendance by one year old (yes/no).

Chapter 4
Pathway Analysis of Hispanic Genome-Wide Association Study in Childhood Leukemia

ABSTRACT

The incidence of acute lymphoblastic leukemia (ALL) has been found to be nearly 20% higher among Hispanics than non-Hispanic Whites. Genome-wide association studies (GWAS) have showed evidence for association in *IKZF1*, *ARID5B*, *CEBPE*, *CDKN2A*, *GATA3*, and *BM1-PIP4K2A*. However, these loci only account for <10% of leukemia genetic risks, indicating additional susceptibility loci are yet to be discovered. We applied pathway-based analyses in Hispanic GWAS data of the California Childhood Leukemia Study (CCLS) to examine if different biological pathways were overrepresented in ALL and major ALL disease subtypes, including B-cell ALL, hyperdiploid B-ALL, and *TEL-AML1* ALL. Furthermore, we applied causal inference and data reduction methods to prioritize a list of candidate genes within each identified overrepresented pathway in ALL for future investigation, while accounting for the correlation between SNPs. The study population is comprised of 323 Hispanic ALL cases and 454 controls from the CCLS, genotyped using Illumina OmniExpress v1 platform. For pathway analyses, we selected genes that had at least one associated significantly SNP (P<0.001), when adjusted for age, gender, and genetic ancestry. Between different disease subtypes, pathway analyses results indicate that hyperdiploid B-ALL and *TEL-AML1* ALL involve distinct biological mechanisms compared to ALL. Focal adhesion is a shared mechanism between different ALL disease subtypes. For childhood ALL, the top five overrepresented KEGG pathways include axon guidance (P_{FDR}=5.1×10^{-06}), protein digestion and absorption (P_{FDR}=7.2×10^{-04}), melanogenesis (P_{FDR}=0.001), leukocyte transendothelial migration (P_{FDR}=0.002), and focal adhesion (P_{FDR}=0.002). Interestingly, 90% of the identified pathways are associated with cancer progression, such as downstream *Wnt* signaling and *MAPK* pathway. Several candidate genes have identified through targeted maximum likelihood estimation (TMLE) method in ALL, including *COL6A6*, *COL5A1*, *DVL1*, *TCF7L1*, *MAP2K2*, *VAV3*, *CTNNA2*, *CDK6*, *RRAS2*, and *CAMK2D* warrant future investigations.

This is the first study to show distinct biological pathways are overrepresented in different leukemia subtypes using pathway-based approaches and to identify a list of candidate genes using newly developed causal method, TMLE. The findings demonstrate that newly developed bioinformatics tools and causal inference methods can provide insights to further understand leukemia pathogenesis.

INTRODUCTION

Leukemia is characterized by the uncontrolled proliferation of hematopoetic cells in the bone marrow (1). Acute lymphoblastic leukemia (ALL) is the most common subtype of childhood leukemia, comprising nearly 80% of diagnosis cases (2). Established evidence for increased risk of ALL, including sex, age, race, prenatal exposure to x-rays, therapeutic radiation, and specific genetic syndromes (3). Direct evidence for inherited genetic susceptibility is demonstrated by the high risk of ALL associated with Bloom's syndrome, neurofibromatosis, ataxia telangiectasia, and constitutional trisomy 21(3). However, these predisposing disorders only account for <5% of leukemia cases (4).

Current evidence suggests that leukemia might result from chromosomal alterations and genetic variations that disrupt the normal process of lymphoid progenitor cells differentiation (1). Around 75% of childhood ALL cases have chromosomal aberrations that can be detected by karyotyping, fluorescent *in situ* hybridization (FISH) or other molecular techniques (5). In B-cell ALL, these aberrations include hyperdiploid (>50 chromosomes), hypodiploid (<44 chromosomes), and chromosomal translocations such as 11q23 *MLL-AF4*, t (12; 21) *TEL-AML1*, t (1; 19) *E2A-PBX1*, and t (9; 22) *BCR-ABL1* (5, 6). Hyperdiploid and *TEL-AML1* rearranged childhood ALL account for approximately 25% and 22% of the entire childhood ALL populations, respectively (7). It is known that different cytogenetic subtypes have different disease prognosis and are suspected to have distinct underlying biological mechanisms (8).

The first genome-wide association study (GWAS) in childhood ALL was published in 2009 with a focus on non-Hispanic White populations (9, 10), and subsequent GWA studies in diverse populations have been confirmed the previous identified associations and further found new susceptibility loci (7, 11-15). All these studies have provided evidence on inherited genetic variations are associated with childhood ALL, including *IKZF1* (7p12.2), *ARID5B* (10q21.2), *CEBPE* (14q11.2), *CDKN2A* (9p21.3), *GATA3* (10p14), and *BM1-PIP4K2A* (10p12.31-12.2). However, these loci only account for <10% of leukemia genetic risk, indicating additional susceptibility loci are yet to be discovered and advanced bioinformatics tools may further guide future directions (16).

GWAS have typically focused on the analysis of single markers, which may lack statistical power to uncover the relatively small effect sizes (odds ratio <2.0) conferred by genetic variants. Some genes may be associated with disease status but may not reach a stringent genome-wide significance threshold ($P < 10^{-8}$). Given the limitations of conventional single-marker association analysis, complementary approaches for GWAS analysis have been developed, including pathway-based approaches. In pathway analysis, a group of related genes in the same biological functional pathway are jointly tested with a disease of interest (17, 18). The method not only can help prioritize the biological pathways, which are most likely to be involved in the disease etiology, but identify novel loci that are unable to detect through traditional GWAS approaches. Several published studies have demonstrated that multiple related genes in the same functional pathway may work together to confer disease susceptibility such as in breast cancer, Parkinson's disease, and Crohn's disease (19)

In this study, we applied pathway-based analyses in Hispanic GWAS data of the California Childhood Leukemia Study (CCLS), testing for biological functions that are significantly enriched in ALL. Furthermore, we compared whether different biological pathways were overrepresented in major leukemia subtypes, including B-cell ALL, hyperdiploid B-ALL and *TEL-AML1* ALL. Lastly, we explored targeted maximum likelihood estimation (TMLE) method incorporating with least absolute shrinkage and selection operator (LASSO) to select a list of candidate genes within each identified pathway in ALL while accounting for the complex correlation between SNPs.

MATERIALS AND METHODS

Study populations

The CCLS is an ongoing population-based case control study which began in 1995. Incident cases of newly diagnosed childhood leukemia (age 0–14 years) were rapidly ascertained from major clinical centers in the study area, usually within 72 hours of diagnosis. Cases were initially identified from four, and later expanded to nine, hospitals in the San Francisco Bay Area and Central Valley. For each case, one or two healthy controls were randomly selected from the state birth registry maintained by the Center for Health Statistics of the California Department of Public Health (CDPH), matching on child's age, sex, Hispanic status (a child was considered Hispanic if either parent self-reported as Hispanic) and maternal race (White, Black, Asian/Pacific Islander, Native American, and Other/Mixed). A detailed description of control selection in the CCLS has been previously reported (20). A total of 86% of case subjects determined eligible consented to participate, and 86% of controls subjects participated among those contacted and considered eligible (21).

Cases and controls were eligible to enter the study if they were under 15 years of age, resided in the study area at the time of diagnosis, had at least one parent who speaks either English or Spanish, and had no prior history of malignancy. The current analysis included 777 Hispanics (323 ALL cases and 454 controls) in the CCLS who enrolled and were interviewed subjects between 1995 and 2008, and for whom had available archived newborn blood (ANB) spot specimens. Detailed cytogenetic classification was described previously (22). Immunophentypic and cytogenetic classification were abstracted from children's medical records, and reviewed by a consulting clinical oncologist. Immunophenotype was determined for ALL cases using flow cytometry profiles and those who were positive for CD19 or CD10 (≥20%) were classified as B-cell ALL (23). Cytogenetic classification was determined by pretreated bone marrow specimens at the time of diagnosis using conventional G-banding or FISH. When extra copies of chromosomes 21 and X were identified by FISH assays, an assignment of high hyperdiploid status (51-67 chromosomes) was made (24). *TEL-AML1* translocations were also identified by FISH assays.

This study was reviewed and approved by the institutional review committees at the University of California Berkeley, the CDPH, and the participating hospitals. Written informed consent was obtained from all parent respondents.

Genotyping and quality control

Samples were genotyped at the Genetic Epidemiology and Genomics Laboratory, School of Public Health, University of California, Berkeley, using the Illumina OmniExpress v1 platform which contains 730,525 markers. Quality control filtering removed SNPs that were not on autosomal chromosomes, were missing in >2% of samples, had minor allele frequency (MAF) of <2% or showed significant deviation from Hardy-Weinberg equilibrium in controls ($P < 1 \times 10^{-5}$). The resulting data set of 634,037 SNPs was then subjected to additional quality control filtering in all samples. We excluded samples for which < 98% of loci were successfully genotyped, samples with discordant sex profiles (birth certificate vs. genetically determined sex) and samples displaying cryptic relatedness (based on identity-by-descent calculations with pi-hat cutoff of 0.15). Ten pairs of duplicate samples were included to assess assay reproducibility, with average concordance >99.99%. The above quality control filtering yielded 777 Hispanic individuals (323 ALL cases and 454 controls) and 634,037 SNPs. To adjust for potential population stratification in study samples, a principal component analysis approach was implemented in EIGENSTRAT (25). For pathway analyses, we selected SNPs that showed marginal associations (P<0.001) with ALL and different ALL disease subtypes (B-cell ALL, hyperdiploid B-ALL, and *TEL-AML1* ALL), while adjusted for age, gender and the first five principal components. At the significance threshold of 0.001, we were able to detect an odds ratio of 1.7 at a minor allele frequency of 15% for childhood ALL and B-cell ALL; an odds ratio of 2.2 at a minor allele frequency of 15% for hyperdiploid B-ALL; and an odds ratio of 2.7 at a minor allele frequency of 20% for *TEL-AML1* ALL (data not shown). Therefore, all SNPs were further filtered based on these criteria and subsequently mapped to genes if they were located within a genomic region based on National Center for Biotechnology Information's dbSNP browser (build 137).

Pathway analyses

Pathway analyses were performed by WEB-based GEne SeT AnaLysis Toolkit (WebGestalt) (26) and Database for Annotation, Visualization and Integrated Discovery, DAVID V6.7 (27). We used three pathway resources for this investigation: the BioCarta pathway database (28), the Kyoto Encyclopedia of Genes and Genomes (KEGG) (29), and the Gene Ontology (GO) database (30). Further, we explored two bioinformatics tools to investigate whether different classification methods generated consistent results. We compared two pathway tools that classified genes based on KEGG pathways: Webgestalt ''KEGG'' (26) and DAVID "KEGG" (27). Second, we investigated two additional pathway classification resources in DAVID: GO and BioCarta databases (27).

To identify functional categories with significantly enrichment in a gene set, we compared the gene set of interest to all human genes. Pathway tools tested whether the number of genes from each pathway in our list of predicted candidate genes is higher than expected given the number of genes selected from the total number of genes. WebGestalt uses the hypergeometric test to evaluate the significance of enrichment (26). In DAVID, a modified Fisher's exact test is adopted to measure the significance of the gene-enrichment in annotated pathways (27). Benjamini and Hochberg procedure controlling for the false discovery rate (FDR) was performed for each pathway analysis tool to adjust for multiple comparisons (31).

After identifying the enriched pathways in ALL, SNPs with the most significant *P*-value for each gene was selected to construct the unweighted genetic risk score for each pathway. We used an additive genetic model and assigned a numerical value for each genotype based on the number of risk alleles for each SNP. The cumulative effect of risk alleles was determined by counting the total number of risk alleles for each individual. A logistic regression model was then used to estimate the cumulative effects of multiple risk alleles within the same functional pathway on the risk of ALL.

Targeted maximum likelihood estimation (TMLE)

To further generate a list of candidate genes within each identified biological pathway for future investigations , while accounting for the confounding effects of other genes, R package, targeted maximum likelihood estimation (TMLE) incorporating with LASSO was applied (32, 33). LASSO can handle massive, highly correlated data and select variables of importance while making predictions. TMLE is based on the general maximum likelihood estimate (MLE) framework and combines it with robust estimation using the efficient influence curve, which measures the influence of one observation on the estimator (32, 33). TMLE helps reduce the bias for the targeted parameter, and provides formal statistical inference. Detailed TMLE methodology is explained elsewhere (32, 33). By incorporating LASSO into TMLE, the approach not only helps with data reduction and candidate gene selection, but also produces robust statistical inference. The method was applied to estimate the effects for SNPs of interest (defined as *P* <0.001) on disease risk, while accounting for the effects of all other SNPs (defined as *P* <0.05) within the same biological pathway. Based on the *P*-values estimated from TMLE method, we generated a list of candidate genes for each identified pathway (*P* <0.05). Only individuals with no missing genetic data were included. All SNPs were pre-screened: those with correlations less than 0.1 or greater than 0.9 with SNPs of interest were excluded in the model since SNPs that have independent effects or are highly correlated each other could not provide additional information to the model (32, 33).

RESULTS

Study characteristics of 777 Hispanics (323 cases and 454 controls) are described in **Table 1**. The distributions of child's sex, age, and race/ethnicity were similar between cases and controls. Since Hispanics are a recently admixed group (34), a proportion (34-37%) of our Hispanic population reported "Mixed or Other" race and 49-51% of them reported "White and Caucasian" race. The frequency of hyperdiploid ALL (>50 chromosomes) is 30%, and the frequency of *TEL-AML1* ALL is 8%. The number of genes reaching significance levels of 0.05, 0.01, 0.005 and 0.001 in the CCLS GWAS dataset are shown in **Supplementary Table S1**. The number of genes increased as less stringent criteria were used in defining significant SNP associations. In the analysis, we used a stringent significant *P* of 0.001 for pathway analysis. Guided by the *P* cutoff and power calculation on different disease subtypes, we were able to map 625 SNPs to 187 genes for childhood ALL, 638 SNPs to 183 genes for B-cell ALL, 404 SNPs to 96 genes for hyperdiploid B-ALL, and 486 SNPs to 110 genes for *TEL-AML1* ALL, respectively.

Table 2 presents the comparisons between the top 10 KEGG pathways in different ALL disease subtypes. The common pathway that consists across different ALL disease subtypes is focal adhesion, which physically connects the extracellular matrix to the cytoskeleton and has long been speculated to mediate cell migration (35). The overrepresented biological pathways for childhood ALL and B-cell ALL are similar (i.e. axon guidance, leukocyte transendothelial migration, and focal adhesion). On the other hand, hyperdiploid B-ALL and *TEL-AML1* ALL show common and distinct biological pathways compared to ALL. 60% of the most significantly overrepresented pathways in childhood hyperdiploid B-ALL are different compared to ALL, including bacterial invasion of epithelial cells, and metabolic pathways. The significant overrepresented pathways in childhood *TEL-AML1* also show some similarity with childhood ALL, including tight junction, axon guidance, and focal adhesion, and some differences, including cell adhesion molecules (CAMs), soluble N-ethylmaleimide-sensitive fusion protein receptor (SNARE) interactions in vesicular transport, and sulfur metabolism. Complete information on overrepresented KEGG pathways associated with different ALL disease subtypes (B-ALL, hyperdiploid B-ALL and *TEL-AML1* ALL) are shown in **Supplementary Table S2-S4**.

Table 3 presents the top 10 ranking KEGG pathways enriched in childhood ALL, which clearly implicates cancer-related pathways (i.e. pathways related to cell proliferation, cell differentiation, and cell signaling). All ten pathways remained significant after imposing a FDR multiple testing correction (P<0.05). The top five KEGG pathways include axon guidance ($P_{\text{FDR}}=5.1\times10^{-06}$), protein digestion and absorption ($P_{\text{FDR}}=7.2\times10^{-04}$), melanogenesis ($P_{\text{FDR}}=0.001$), leukocyte transendothelial migration ($P_{\text{FDR}}=0.002$), and focal adhesion ($P_{\text{FDR}}=0.002$). Among these pathways, leukocyte transendothelial migration pathway is essential for immune response and inflammatory reaction, which may be associated with leukemia pathogenesis. Notably, all these pathways are connected to downstream *PI3K*, *MAPK*, or *Wnt* signaling pathway, and these pathways have been linked to multiple human malignancies (36-38).

To examine the cumulative effects of the most significant SNPs for each gene, we calculated unweighted genetic scores by summing the number of risk alleles carried by each individual for each pathway. The risk of ALL increased as the number of risk alleles increased within each biological pathway (P for trend <0.05) **(Table 3)**. For example, the odds of developing childhood ALL significantly increased with each additional copy of risk alleles for genes in the focal adhesion pathway (OR=1.96; 95% confidence interval (CI): 1.60-2.41).

After identifying the enriched pathways in ALL, we further prioritized a gene list by applying data reduction and causal inference methods. Based on LASSO and TMLE results, we generated a list of candidate genes for each biological pathway, while accounting for confounding effects of other genes within the same pathway**. Table 4** presents an example of TMLE results for focal adhesion pathway. SNPs within *VAV3*, *COL6A6*, and *COL5A1* genes have much more significant P-values (P<0.05) while other SNPs are no long significant, suggesting that the effect of the pathway may be driven by these three genes. By applying the same criteria, important genes for each identified pathways were selected as shown in **Table 3** and **Supplementary Table S5**. For example, in axon guidance pathway, *UNC5*, *EPHB1*, and *PLXNC1* may play a more central role in ALL disease development than other genes **(Table 3)**. Similarly, in leukocyte transendothelial migration pathway, *VAV3*, and *CTNNA2* may be worth of future investigations **(Table 3)**.

Furthermore, to compare the outcomes of the different pathway databases, genes were classified into pathways using the Gene Ontology and BioCarta databases for childhood ALL. The only significant BioCarta pathway associated with childhood ALL is integrin signaling pathway, which is triggered when integrins in the cell membrane bind to extracellular matrix components (**Supplementary Table S6**). When Gene Ontology term was investigated, there are significant enrichment associated with cell morphogenesis involved in neuron differentiation, cellular component morphogeneisis, and cell motion (**Supplementary Table S7**). The results for different pathway tools Webgestalt and DAVID are similar (**Supplementary Table S8**). Overall, we observed consistency between different pathway analysis tools when analyzing the same dataset.

DISCUSSION

This is the first study to show distinct biological pathways are overrepresented in different leukemia disease subtypes using pathway analyses approaches. We further apply TMLE incorporating with LASSO to select variables of importance and provide formal statistical inferences for each overrepresented pathway in ALL. The results demonstrate that newly developed bioinformatics tools and causal inference method may illuminate new and biologically relevant pathways and genes to improve our current understanding of pathogenesis in childhood leukemia.

Realizing the limitations of conventional single-marker association analysis, complementary approaches for GWAS analysis have been developed in recent years (18). Pathway-based analyses provide a complementary approach to combine effects of many loci, allowing for small contributions to overall disease susceptibility by individual SNPs that are otherwise missed by conventional single-SNP GWAS analysis (17). By taking into account prior biological knowledge about genes and pathways, we may have a better chance to identify novel genes and biological mechanisms that are involved in disease pathogenesis (19). Additionally, as the most associated gene in a pathway might not be the best candidate for therapeutic intervention, targeting susceptibility pathways might also have clinical implications for finding additional drug targets. Several novel molecular targeted agents are under investigations for ALL treatment such as tyrosine kinase inhibitors, Fms-like tyrosine kinase, NOTCH1 inhibitors, and mTOR inhibitors (38). The enrichment pathways that identified in our study may further guide sophisticated targeted treatment strategies for ALL.

Subtypes of childhood ALL exhibit specific molecular characteristics are known to be important in risk stratification and treatment specification at diagnosis (39). However, little is known about the underlying mechanisms leading to different ALL disease subtypes with specific chromosome abnormalities. These molecular characteristics may have distinctive biological mechanisms. Our pathway analysis results clearly support that hyperdiploidy B-ALL and *TEL-AML* ALL might have different disease pathogenesis compared with childhood ALL. Compared with childhood ALL, pathways involved with hyperdiploidy B-All are more related to signal transduction and metabolism pathways. Compared with childhood ALL, pathways involved with *TEL-AML* are more related to tissue and organ morphogenesis and the maintenance of cell and tissue structure and function. These interactions between transmembrane molecules lead to a direct or indirect control of cellular activities such as adhesion, proliferation, and apoptosis. Our study provides a foundation to examine each of the biological pathways that is specific to ALL disease subtypes to further understand the disease etiology of ALL subtypes. The shared mechanism across different ALL disease subtypes is focal adhesion, which consists of large protein complexes organized at the basal surface of cells. Cells are usually surrounded by the extracellular matrix (ECM), and adhesion of cells to the ECM is the key to the regulation of cellular morphology, migration, proliferation, survival, and differentiation (40). These functions are indispensable during development, for maintenance of tissue architecture, and the induction of tissue repair, which have been indicated to involve with tumor formation and progression (35).

It is well known that childhood ALL is caused by interactions between multiple genetic factors and environmental factors, and that complex molecular networks and cellular pathways play key roles in development of childhood leukemia. Our pathway analyses of Hispanic GWAS identify overrepresentation of association signals in several pathways (axon guidance, protein digestion and absorption, melanogenesis, leukocyte transendothelial migration, focal adhesion, endometrial cancer, glioma, pathways in cancer, tight junction, and regulation of actin cytoskeleton) and suggest involvement in at least three biological processes (anatomical structure morphogenesis, organ morphogenesis, and cellular component movement).

The identified pathways that are overrepresented with childhood ALL in this study are mainly those that are associated with other malignances such as endometrial cancer and glioma or those that are associated cell communication and cell motility such as focal adhesion, tight junction, and regulation of actin cytoskeleton. All these identified pathways are involved in different cellular processes that mediate signal transduction cascades leading to cell proliferation, cell migration, and cell adhesion. For example, regulation of actin cytoskeleton is related to cell migration, which is required for many biological processes, such as embryonic morphogenesis, immune surveillance, tissue repair and regeneration (41). Aberrant regulation of cell migration drives progression of many diseases, including cancer invasion and metastasis (41). Tight junction pathway is also associated with cell proliferation, transformation, and metastasis and it is essential for the development of the body during embryogenesis (42). Moreover, this pathway regulates integrin signaling that is required for changes in cell shape and motility (42).

Other identified cancer-related pathways, including downstream *MAPK*, *PI3K-AKT*, *Jak-STAT*, and *Wnt* signaling pathway, have been showed to be closely related to cancer progression (38, 43). An important extension to the pathway analysis is highlighted the *RAS*/*RAF*/*MAPK* canonical signaling cascade as the common downstream pathway associated with childhood ALL in this analysis. This cascade plays an essential role in transmitting extracellular signals from growth factors to promote the growth, proliferation, differentiation, and survival of cells, and modification in its activity has been linked to multiple human malignancies (36). As *MAPK* activation plays critical roles in the regulation of proliferation and differentiation, aberrant activation of this pathway could be considered as a likely cause of hematopoietic disease (44). The importance of *MAPK* signaling in the regulation of hematopoiesis is underscored by activation of the downstream *MAPK* genes in myeloid leukemia (44).

Another interesting pathway that identified through pathway analyses is leukocyte transendothelial migration. There are multiple lines of evidence for a potential role of leukocyte migration in childhood ALL development. Leukocyte migration from the blood into tissues is essential for immune surveillance and inflammation responses (45). It is regulated by a reaction cascade involving sequential interactions of adhesion receptors and chemokines (46). In addition, chemokines regulate the biological processes of hematopoietic cells to cellular activation, and differentiation (47). Gene within the pathway such as *VAV3* is warranted for further investigation given its relation to regulating B-cell receptor signaling pathway and aberration of the gene may lead to B-cell malignancies (48).

In addition to identifying enriched pathways, the study further selected a list of candidate genes that can be used for future targeted sequencing and functional studies to assess the genetic effects on ALL susceptibility. The data reduction algorithm, LASSO, together with causal inference method, TMLE, produce a target list of candidate genes while accounting for the correlation between SNPs. Several genes have been identified through the approach, including *COL6A6*, *COL5A1*, *DVL1*, *TCF7L1*, *MAP2K2*, *VAV3*, *CTNNA2*, *CDK6*, *RRAS2*, and *CAMK2D*. *VAV3* gene shows up as top-rank gene in several pathways. The gene is recruited and activated on epidermal growth factor and insulin-like growth factor, which has been linked to prostate cancer development (49). *RRAS2*, a member of the *RAS* superfamily of small GTP-binding proteins, encodes protein that associates with the plasma membrane and may function as a signal transducer. The results in *RRAS* knockouts indicate that this family gene may be associated with cell development and during antigen-induced responses in T and B cells (50).

Other genes that contributed to the association with pathways in cancer and ALL are *DVL3*, *CDK6*, *TCF7L1*, *MAP2K2*, and *CTNNA2*. Among these genes, *TCF7L1* and *DVL1* are members of the *Wnt* pathway (51). Aberrant activation of *Wnt* signaling pathway has been documented in various human cancers including myeloid leukemia (52). This signaling pathway ultimately activates other genes involved in B cell proliferation and differentiation and regulates the identity and function of epidermal and embryonic stem cells (51). Extensive work also has established in *MAP2K2* and *CDK6* regulation role in cell growth and cell death (36, 53). *MAP2K2* belongs to *RAS*/*RAF*/*MAPK* transduction signaling that play an essential role in connecting cell-surface receptors to transcription factors (36). Enhanced *CDK6* expression has been documented in lymphoma and leukemia. Several reports have shown chromosomal translocations in patients suffering from B-lymphoid malignancies involving with *CDK6* (54).

To our knowledge, this is the first report using pathway based analyses and newly developed causal method in Hispanic GWAS data of childhood ALL. These identified pathways are presumed to play a role in disease pathogenesis through variations in specific genes that have not yet been identified. The results strongly suggest that development of ALL are modulated by several critical cellular processes, including cell growth, differentiation, survival, and migration. Other strength of our study is the detailed information on cytogenetic subtypes, which enable us to show distinct biological mechanisms are involved with different disease subtypes. Furthermore, we employed TMLE method to prioritize genes that may serve integral functions for tumor development. All these genes have been linked to human malignances and established biological relevance.

Our results should be interpreted in the context of the limitations. A limitation of the pathway analysis method is the requirement for specification of a *P* cutoff in defining the list of significantly associated SNPs. Clearly, the choice of this threshold could be arbitrary. We chose a relatively stringent cutoff $P < 0.001$. Another important limitation of the pathway-based approach is the incomplete biological annotation of the human genome. At present, the function of many human genes is unknown; therefore, they cannot be assigned to known pathways. Moreover, susceptibility loci in intergenic regions are also not included in this study. As a result, when using this approach, only a small portion of the human genome variation can be studied. In particular, the results may favor pathways with more complete gene information and large genes containing many SNPs which are more likely to contain significant SNPs by chance alone. Additionally, there is no gold standard on pathway definition, and different databases have different guidelines for their pathway construction and curation. Consequently, the gene content of pathways representing the same biological process may vary between different databases, and this may have some impact on the analyses. We aimed at minimizing this effect by selecting pathways from three commonly used resources.

In conclusion, pathway analysis findings are uniquely and naturally connected to the functional biology underlying childhood leukemia. The identifications of cancer-related and inflammatory-related pathways support the power of the method to highlight pathways with established relevance to childhood leukemia etiology. In addition, the results elucidate numerous candidate genes for further explorations. Future studies are needed to confirm and sequence the identified genes in a larger childhood leukemia dataset.

Grant support

This study was supported by grants from the National Institute of Environmental Health Sciences (PS42 ES04705 and R01 ES09137), the National Cancer Institute (R25CA112355), and Children with Cancer, United Kingdom.

Acknowledgements

Participating hospitals and clinical collaborators included University of California Davis Medical Center (Dr Jonathan Ducore), University of California San Francisco (Dr Mignon Loh and Dr 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 Sacramento (Dr Vincent Kiley), Kaiser Permanente Santa Clara (Dr Carolyn Russo and Dr Alan Wong), Kaiser Permanente San Francisco (Dr Kenneth Leung), Children's Hospital of Los Angeles (Dr Cecilia Fu) and Kaiser Permanente Oakland (Dr Stacy Month). We also acknowledge our collaborators at the Northern California Cancer Center and the entire California Childhood Leukemia Study staff for their effort and dedication.

REFERENCES

1. Pui CH, Relling MV, Downing JR. Acute lymphoblastic leukemia. N Engl J Med 2004;350(15):1535-48.

2. Stiller CA, Parkin DM. Geographic and ethnic variations in the incidence of childhood cancer. Br Med Bull 1996;52(4):682-703.

3. Eden T. Aetiology of childhood leukaemia. Cancer Treat Rev 2010;36(4):286-97.

4. Belson M, Kingsley B, Holmes A. Risk factors for acute leukemia in children: a review. Environ Health Perspect 2007;115(1):138-45.

5. Mullighan CG. The molecular genetic makeup of acute lymphoblastic leukemia. Hematology Am Soc Hematol Educ Program 2012;2012:389-96.

6. Buffler PA, Wood SM, Suarez L, Kilian DJ. Mortality follow-up of workers exposed to 1,4-dioxane. J Occup Med 1978;20(4):255-9.

7. Ellinghaus E, Stanulla M, Richter G, Ellinghaus D, te Kronnie G, Cario G, et al. Identification of germline susceptibility loci in ETV6-RUNX1-rearranged childhood acute lymphoblastic leukemia. Leukemia 2012;26(5):902-9.

8. Williams DL, Tsiatis A, Brodeur GM, Look AT, Melvin SL, Bowman WP, et al. Prognostic importance of chromosome number in 136 untreated children with acute lymphoblastic leukemia. Blood 1982;60(4):864-71.

9. Papaemmanuil E, Hosking FJ, Vijayakrishnan J, Price A, Olver B, Sheridan E, et al. Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. Nat Genet 2009;41(9):1006-10.

10. Trevino LR, Yang W, French D, Hunger SP, Carroll WL, Devidas M, et al. Germline genomic variants associated with childhood acute lymphoblastic leukemia. Nat Genet 2009;41(9):1001-5.

11. Han S, Lee KM, Park SK, Lee JE, Ahn HS, Shin HY, et al. Genome-wide association study of childhood acute lymphoblastic leukemia in Korea. Leuk Res 2010;34(10):1271-4.

12. Orsi L, Rudant J, Bonaventure A, Goujon-Bellec S, Corda E, Evans TJ, et al. Genetic polymorphisms and childhood acute lymphoblastic leukemia: GWAS of the ESCALE study (SFCE). Leukemia 2012.

13. Xu H, Yang W, Perez-Andreu V, Devidas M, Fan Y, Cheng C, et al. Novel Susceptibility Variants at 10p12.31-12.2 for Childhood Acute Lymphoblastic Leukemia in Ethnically Diverse Populations. J Natl Cancer Inst 2013.

14. Walsh KM, Chokkalingam AP, Hsu LI, Metayer C, de Smith AJ, Jacobs DI, et al. Associations between genome-wide Native American ancestry, known risk alleles and B-cell ALL risk in Hispanic children. Leukemia 2013.

15. Sherborne AL, Hosking FJ, Prasad RB, Kumar R, Koehler R, Vijayakrishnan J, et al. Variation in CDKN2A at 9p21.3 influences childhood acute lymphoblastic leukemia risk. Nat Genet 2010;42(6):492-4.

16. Enciso-Mora V, Hosking FJ, Sheridan E, Kinsey SE, Lightfoot T, Roman E, et al. Common genetic variation contributes significantly to the risk of childhood B-cell precursor acute lymphoblastic leukemia. Leukemia 2012;26(10):2212-5.

17. Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS results: A review of statistical methods and recommendations for their application. Am J Hum Genet 2010;86(1):6-22.

18. Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. Nat Rev Genet 2010;11(12):843-54.

19. Torkamani A, Topol EJ, Schork NJ. Pathway analysis of seven common diseases assessed by genome-wide association. Genomics 2008;92(5):265-72.
20. Ma X, Buffler PA, Layefsky M, Does MB, Reynolds P. Control selection strategies in case-control studies of childhood diseases. Am J Epidemiol 2004;159(10):915-21.
21. Bartley K, Metayer C, Selvin S, Ducore J, Buffler P. Diagnostic X-rays and risk of childhood leukaemia. Int J Epidemiol 2010;39(6):1628-37.
22. Metayer C, Zhang L, Wiemels JL, Bartley K, Schiffman J, Ma X, et al. Tobacco smoke exposure and the risk of childhood acute lymphoblastic and myeloid leukemias by cytogenetic subtype. Cancer Epidemiol Biomarkers Prev 2013;22(9):1600-11.
23. Hrusak O, Porwit-MacDonald A. Antigen expression patterns reflecting genotype of acute leukemias. Leukemia 2002;16(7):1233-58.
24. Aldrich MC, Zhang L, Wiemels JL, Ma X, Loh ML, Metayer C, et al. Cytogenetics of Hispanic and White children with acute lymphoblastic leukemia in California. Cancer Epidemiol Biomarkers Prev 2006;15(3):578-81.
25. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38(8):904-9.
26. Zhang B, Kirov S, Snoddy J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res 2005;33(Web Server issue):W741-8.
27. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 2009;4(1):44-57.
28. Nishimura D. BioCarta. Biotech Software & Internet Report 2001;2(3):117-120.
29. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28(1):27-30.
30. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, et al. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 2004;32(Database issue):D258-61.
31. Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. Stat Med 1990;9(7):811-8.
32. Gruber SaVdL, M. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. International Journal of Biostatistics 2010;6.
33. van der Laan M, and D. Rubin. Targeted maximum likelihood learning. The International Journal of Biostatistics 2006;2.
34. Baye TM, Wilke RA. Mapping genes that predict treatment outcome in admixed populations. Pharmacogenomics J 2010;10(6):465-77.
35. McLean GW, Carragher NO, Avizienyte E, Evans J, Brunton VG, Frame MC. The role of focal-adhesion kinase in cancer - a new therapeutic opportunity. Nat Rev Cancer 2005;5(7):505-15.
36. Geest CR, Coffer PJ. MAPK signaling pathways in the regulation of hematopoiesis. J Leukoc Biol 2009;86(2):237-50.
37. Downward J. Targeting RAS signalling pathways in cancer therapy. Nat Rev Cancer 2003;3(1):11-22.
38. Courtney KD, Corcoran RB, Engelman JA. The PI3K pathway as drug target in human cancer. J Clin Oncol 2010;28(6):1075-83.
39. Inaba H, Greaves M, Mullighan CG. Acute lymphoblastic leukaemia. Lancet 2013;381(9881):1943-55.

40. Gumbiner BM. Cell adhesion: the molecular basis of tissue architecture and morphogenesis. Cell 1996;84(3):345-57.

41. Zhao X, Guan JL. Focal adhesion kinase and its signaling pathways in cell migration and angiogenesis. Adv Drug Deliv Rev 2011;63(8):610-5.

42. Sawada N. Tight junction-related human diseases. Pathol Int 2013;63(1):1-12.

43. Dreesen O, Brivanlou AH. Signaling pathways in cancer and embryonic stem cells. Stem Cell Rev 2007;3(1):7-17.

44. Johnson DE. Src family kinases and the MEK/ERK pathway in the regulation of myeloid differentiation and myeloid leukemogenesis. Adv Enzyme Regul 2008;48:98-112.

45. Infante E, Ridley AJ. Roles of Rho GTPases in leucocyte and leukaemia cell transendothelial migration. Philos Trans R Soc Lond B Biol Sci 2013;368(1629):20130013.

46. Muller WA. Mechanisms of leukocyte transendothelial migration. Annu Rev Pathol 2011;6:323-44.

47. Schroeder MA, DiPersio JF. Mobilization of hematopoietic stem and leukemia cells. J Leukoc Biol 2012;91(1):47-57.

48. Inabe K, Ishiai M, Scharenberg AM, Freshney N, Downward J, Kurosaki T. Vav3 modulates B cell receptor responses by regulating phosphoinositide 3-kinase activation. J Exp Med 2002;195(2):189-200.

49. Lyons LS, Rao S, Balkan W, Faysal J, Maiorino CA, Burnstein KL. Ligand-independent activation of androgen receptors by Rho GTPase signaling in prostate cancer. Mol Endocrinol 2008;22(3):597-608.

50. Delgado P, Cubelos B, Calleja E, Martinez-Martin N, Cipres A, Merida I, et al. Essential function for the GTPase TC21 in homeostatic antigen receptor signaling. Nat Immunol 2009;10(8):880-8.

51. Anastas JN, Moon RT. WNT signalling pathways as therapeutic targets in cancer. Nat Rev Cancer 2013;13(1):11-26.

52. Mikesch JH, Steffen B, Berdel WE, Serve H, Muller-Tidow C. The emerging role of Wnt signaling in the pathogenesis of acute myeloid leukemia. Leukemia 2007;21(8):1638-47.

53. Kollmann K, Heller G, Schneckenleithner C, Warsch W, Scheicher R, Ott RG, et al. A Kinase-Independent Function of CDK6 Links the Cell Cycle to Tumor Angiogenesis. Cancer Cell 2013;24(2):167-81.

54. Parker EP, Siebert R, Oo TH, Schneider D, Hayette S, Wang C. Sequencing of t(2;7) translocations reveals a consistent breakpoint linking CDK6 to the IGK locus in indolent B-cell neoplasia. J Mol Diagn 2013;15(1):101-9.

Table 1. Characteristics of Hispanic case-control study subjects, CCLS, 1995-2008

	Cases, n (%)	**Controls, n (%)**
Study subjects	*323 (41.6)*	454 (58.4)
Sex		
Male	173 (53.6)	240 (52.9)
Female	150 (46.4)	214 (47.1)
Age		
Mean age, y(SE)	5.3 (3.4)	5.3 (3.4)
Race		
White/Caucasian	161 (49.8)	235 (51.8)
African American	14 (4.3)	15 (3.3)
Native American	0 (0)	4 (0.9)
Asian or Pacific Islander	26 (8.1)	40 (8.8)
Mixed or others	120 (37.2)	156 (34.4)
Cytogenetics (case-only)		
B-cell ALL	297 (91.9)	-
Hyperdiploid B-cell ALL (>50 chromosome)	97 (30.0)	-
TEL-AML ALL	40 (8.1)	

Table 2. Comparisons between different ALL disease subtypes and associated biological pathways [a]

Pathway	ALL	B-ALL	Hyperdiploid B-ALL	TEL-AML ALL
Axon guidance	√	√	√	
Protein digestion and absorption	√	√	√	
Melanogenesis	√	√		
Leukocyte transendothelial migration	√	√		
Focal adhesion	√	√	√	√
Endometrial cancer	√			
Glioma	√			
Pathways in cancer	√	√	√	√
Tight junction	√			√
Regulation of actin cytoskeleton	√	√		
Gap junction		√		
Histidine metabolism		√		
Pancreatic secretion		√		
Bacterial invasion of epithelial cells			√	
Metabolic pathways			√	
Small cell lung cancer			√	√
Amoebiasis			√	
Valine, leucine and isoleucine degradation			√	
Purine metabolism			√	
Sulfur metabolism				√
Cell adhesion molecules (CAMs)				√
ABC transporters				√
SNARE interactions in vesicular transport				√
Fat digestion and absorption				√
Non-small cell lung cancer				√

√ top-10 ranking KEGG pathways associated with disease status

√ adjusted *P* value based on correction for False Discovery Rate (FDR) using Benjamini and Hochberg (BH) is smaller than 0.05.

a) SNPs which showed association with each childhood ALL disease subtypes (P < 0.001) and filtered by power calculation were included in this study. The analysis was limited to KEGG pathways where at least two genes were present in the submitted list and used a hypergeometric test to compare the submitted list to a reference of all human genes using WebGestalt v.2 (http://bioinfo.vanderbilt.edu/webgestalt/).

Table 3. Overrepresented KEGG pathways among the top results of CCLS Hispanic GWAS on childhood ALL and identification of important genes using causal inference approach

KEGG pathway	Genes	obs	exp	ratio	*P* value	P_{adjust}	OR* (95% CI)	Important genes**
Axon guidance	*LRRC4C PLXNC1 SLIT3 EPHB1 NTN1 UNC5B GNAI1 NGEF*	8	0.55	14.46	9.64×10^{-08}	5.1×10^{-06}	1.59 (1.35-1.88)	*UNC5B* *EPHB1* *PLXNC1*
Protein digestion and absorption	*CPA2 COL4A2 SLC7A8 COL5A1 COL6A6*	5	0.35	14.39	2.71×10^{-05}	7.1×10^{-04}	2.03 (1.60-2.58)	*CPA2* *COL6A6* *COL5A1* *SLC7A8*
Melanogenesis	*TCF7L1 CAMK2D DVL3 MAP2K2 GNAI1*	5	0.43	11.54	7.81×10^{-05}	0.0014	1.58 (1.35-1.84)	*DVL3* *TCF7L1* *CAMK2D* *MAP2K2*
Leukocyte transendothelial migration	*ITGAL VAV3 MYL2 CTNNA2 GNAI1*	5	0.50	10.05	2.1×10^{-04}	0.0021	1.64 (1.35-1.98)	*VAV3* *CTNNA2*
Focal adhesion	*VAV3 MYL2 COL4A2 TLN1 COL5A1 COL6A6*	6	0.86	6.99	2.1×10^{-04}	0.0021	1.96 (1.60-2.41)	*VAV3* *COL6A6* *COL5A1*

Endometrial cancer	*TCF7L1 MAP2K2 CTNNA2*	3	0.22	13.45	0.001	0.013	1.53 (1.25-1.87)	*TCF7L1* *MAP2K2* *CTNNA2*
Glioma	*CAMK2D MAP2K2 CDK6*	3	0.28	10.8	0.002	0.014	1.59 (1.33-1.91)	*CDK6* *CAMK2D* *MAP2K2*
Pathways in cancer	*TCF7L1 DVL3 COL4A2 MAP2K2 CDK6 CTNNA2*	6	1.40	4.29	0.003	0.014	1.63 (1.41-1.90)	*DVL3* *CDK6* *TCF7L1* *MAP2K2* *CTNNA2*
Tight junction	*MYL2 RRAS2 CTNNA2 GNAI1*	4	0.57	7.06	0.002	0.014	1.52 (1.24-1.87)	*TCF7L1* *RRAS2* *CTNNA2*
Regulation of actin cytoskeleton	*ITGAL VAV3 MYL2 RRAS2 MAP2K2*	5	0.91	5.47	0.002	0.014	1.69 (1.44-2.00)	*VAV3* *MAP2K2* *RRAS2*

Abbreviations: Exp, expected; KEGG, Kyoto Encyclopedia of Genes and Genomes; Obs, observed; OR, odds ratio.
SNPs which showed association with childhood ALL (*P* < 0.001) were included in this study. Of the 187 genes submitted for analysis, 185 were incorporated for analysis using a hypergeometric test to compare the submitted list to a reference of all human genes using WebGestalt v.2 (http://bioinfo.vanderbilt.edu/webgestalt/). The analysis was limited to KEGG pathways where at least two genes were present in the submitted list. The top-10

ranking KEGG pathways are shown. Adjusted P-values were based on correlation for False Discovery Rate (FDR) using Benjamini and Hochberg (BH) procedure.

*Odds Ratios and 95% CI for cumulative effects of SNPs within each pathway on the risk of childhood ALL was calculated using a logistic regression model.

**Genes were selected based on the *P* values from targeted maximum likelihood estimation (TMLE) (*P* <0.05).

Table 4. TMLE analysis suggests *VAV3*, *COL6A6*, and *COL5A1* as important genes within the focal adhesion pathway

Genes	SNPs	Marginal P-value	TMLE *P*-value
VAV3	rs17485868	4.22×10^{-5}	1.31×10^{-18}
VAV3	rs12126655	6.66×10^{-4}	0.542
VAV3	rs10494081	7.75×10^{-5}	0.121
COL6A6	rs16830219	3.66×10^{-4}	8.63×10^{-17}
TLN1	rs2295795	3.71×10^{-4}	0.017
COL5A1	rs12554098	8.73×10^{-4}	5.96×10^{-16}
COL4A2	rs9555707	8.64×10^{-5}	0.089

Supplementary Table S1. Assessment of numbers of genes meeting various significance thresholds from CCLS Hispanic GWAS

	Number of genes (*P*<0.05)	**Number of genes (*P* <0.01)**	**Number of genes (*P* <0.005)**	**Number of genes (*P* <0.001)**
ALL	4322	1290	791	187
B-ALL	4354	1295	789	183
Hyperdiploid B-ALL	4187	1201	717	96
TEL-AML	4014	1165	704	110

Supplementary Table S2. Overrepresented KEGG pathways among the top results of CCLS Hispanic GWAS on childhood B-ALL

KEGG pathway	Genes	obs	exp	Enrichment ratio	*P* value	*P* $_{\text{adjust}}$
Axon guidance	*SLIT3 UNC5B EPHB1 NGEF GNAI1*	5	0.54	9.29	2.1×10^{-04}	0.004
Leukocyte transendothelial migration	*ITGAL VAV3 MYL2 CTNNA2 GNAI1*	5	0.48	10.33	1.1×10^{-04}	0.004
Focal adhesion	*VAV3 PDGFC MYL2 COL4A2 TLN1*	5	0.83	5.99	0.001	0.029
Protein digestion and absorption	*CPA2 COL4A2 SLC7A8*	3	0.34	8.87	0.005	0.046
Gap junction	*PDGFC GNAI1 TUBB1*	3	0.38	7.99	0.006	0.046
Histidine metabolism	*ASPA ALDH1A3*	2	0.12	16.52	0.006	0.046
Pancreatic secretion	*CPA2 SLC4A4 PNLIP*	3	0.42	7.12	0.009	0.047
Melanogenesis	*TCF7L1 CAMK2D GNAI1*	3	0.42	7.12	0.009	0.047
Pathways in cancer	*TCF7L1 COL4A2 CDK6 CTNNA2 CBLB*	5	1.36	3.67	0.012	0.053
Regulation of actin cytoskeleton	*ITGAL VAV3 PDGFC MYL2*	4	0.89	4.50	0.012	0.053

Abbreviations: Exp, expected; KEGG, Kyoto Encyclopedia of Genes and Genomes; Obs, observed; OR, odds ratio. SNPs which showed association with childhood B-ALL ($P < 0.001$) were included in this study. Of the 183 genes submitted for analysis, 180 were incorporated for analysis using a hypergeometric test to compare the submitted list to a reference of all human genes using WebGestalt v.2 (http://bioinfo.vanderbilt.edu/webgestalt/). The analysis was limited to KEGG pathways where at least two genes were present in the submitted list. The top-10 ranking KEGG pathways are shown. Adjusted P-values were based on correlation for False Discovery Rate (FDR) using Benjamini and Hochberg (BH) procedure.

Supplementary Table S3. Overrepresented KEGG pathways among the top results of CCLS Hispanic GWAS on childhood hyperdiploid B-ALL

KEGG pathway	Genes	obs	exp	Enrichment ratio	*P* value	P_{adjust}
Small cell lung cancer	*PTK2 FHIT COL4A2*	3	0.18	16.37	8.1×10^{-04}	0.009
Metabolic pathways	*BCKDHB GALNT5 GMDS CES5A HPSE2 TBXAS1 MECR AOX1*	8	2.44	3.28	0.003	0.016
Valine, leucine and isoleucine degradation	*BCKDHB AOX1*	2	0.09	21.08	0.004	0.016
Bacterial invasion of epithelial cells	*PTK2 CAV3*	2	0.15	13.25	0.010	0.024
Focal adhesion	*PTK2 CAV3 COL4A2*	3	0.43	6.96	0.009	0.024
Protein digestion and absorption	*COL4A2 SLC7A8*	2	0.17	11.45	0.013	0.026
Amoebiasis	*PTK2 COL4A2*	2	0.23	8.75	0.021	0.037
Pathways in cancer	*PTK2 COL4A2 APPL1*	3	0.70	4.27	0.037	0.049
Axon guidance	*SLIT3 PTK2*	2	0.28	7.19	0.031	0.049
Purine metabolism	*FHIT PDE2A*	2	0.35	5.73	0.048	0.057

Abbreviations: Exp, expected; KEGG, Kyoto Encyclopedia of Genes and Genomes; Obs, observed; OR, odds ratio. SNPs which showed association with childhood hyperdiploidy B-ALL ($P < 0.001$) were included in this study. Of the 96 genes submitted for analysis, 93 were incorporated for analysis using a hypergeometric test to compare the submitted list to a reference of all human genes using WebGestalt v.2 (http://bioinfo.vanderbilt.edu/webgestalt/). The analysis was limited to KEGG pathways where at least two genes were present in the submitted list. The top-10 ranking KEGG pathways are shown. Adjusted P-values were based on correlation for False Discovery Rate (FDR) using Benjamini and Hochberg (BH) procedure.

Supplementary Table S4. Overrepresented KEGG pathways among the top results of CCLS Hispanic GWAS on childhood *TEL-AML1* ALL

KEGG pathway	**Genes**	**obs**	**exp**	**Enrichment ratio**	***P* value**	***P* $_{\text{adjust}}$**
Pathways in cancer	*RARB COL4A1 AKT1 CTNNA3 DCC*	5	0.82	6.12	0.001	0.011
Sulfur metabolism	*PAPSS2 CHST12*	2	0.03	61.43	0.001	0.011
Small cell lung cancer	*RARB COL4A1 AKT1*	3	0.21	14.09	0.001	0.012
Focal adhesion	*PDGFD COL4A1 AKT1 ITGA9*	4	0.05	7.99	0.001	0.012
Tight junction	*MAGI1 AKT1 CTNNA3*	3	0.33	9.08	0.004	0.018
Cell adhesion molecules	*ESAM ITGA9 PVRL3*	3	0.33	9.01	0.004	0.018
SNARE interactions in vesicular transport	*STX16 VTI1A*	2	0.09	22.18	0.004	0.018
ABC transporters	*ABCG5 ABCG8*	2	0.11	18.15	0.005	0.018
Fat digestion and absorption	*ABCG5 ABCG8*	2	0.12	17.36	0.006	0.018
Non-small cell lung cancer	*RARB AKT1*	2	0.41	14.79	0.008	0.021

Abbreviations: Exp, expected; KEGG, Kyoto Encyclopedia of Genes and Genomes; Obs, observed; OR, odds ratio. SNPs which showed association with childhood TEL-AML ALL (*P* < 0.001) were included in this study. Of the 110 genes submitted for analysis, 108 were incorporated for analysis using a hypergeometric test to compare the submitted list to a reference of all human genes using WebGestalt v.2 (http://bioinfo.vanderbilt.edu/webgestalt/). The analysis was limited to KEGG pathways where at least two genes were present in the submitted list. The top-10 ranking KEGG pathways are shown. Adjusted P-values were based on correlation for False Discovery Rate (FDR) using Benjamini and Hochberg (BH) procedure.

Supplementary Table S5. Complete list of TMLE *P*-values for each important gene that are selected for each biological pathway

Pathway	GENE	SNPs	Marginal *P* value	TMLE *P* value
Axon Guidance	*PLXNC1*	rs10777585	8.3×10^{-4}	1.8×10^{-7}
Axon Guidance	*EPHB1*	rs7636504	9.6×10^{-4}	$<2\times10^{-16}$
Axon Guidance	*UNC5B*	rs2244140	2.0×10^{-4}	$<2\times10^{-16}$
protein digestion and absorption	*COL6A6*	rs16830219	3.0×10^{-4}	2.8×10^{-15}
protein digestion and absorption	*CPA2*	rs2171493	3.2×10^{-5}	4.8×10^{-20}
protein digestion and absorption	*COL5A1*	rs12554098	8.7×10^{-4}	5.7×10^{-13}
protein digestion and absorption	*SLC7A8*	rs2144827	2.2×10^{-5}	4.6×10^{-10}
melanogenesis	*TCF7L1*	rs2568222	1.6×10^{-4}	1.5×10^{-11}
melanogenesis	*DVL3*	rs2175525	8.8×10^{-4}	5.5×10^{-39}
melanogenesis	*CAMK2D*	rs6842886	7.1×10^{-4}	8.3×10^{-9}
melanogenesis	*CAMK2D*	rs10488958	4.1×10^{-4}	4.3×10^{-9}
melanogenesis	*CAMK2D*	rs12512765	9.7×10^{-4}	3.1×10^{-8}
melanogenesis	*MAP2K2*	rs350916	5.9×10^{-4}	9.4×10^{-6}
leukocyte transendothelial migration	*VAV3*	rs17485868	4.2×10^{-5}	1.8×10^{-18}
leukocyte transendothelial migration	*CTNNA2*	rs11687661	1.1×10^{-4}	1.8×10^{-4}
focal adhesion	*VAV3*	rs17485868	4.2×10^{-5}	1.3×10^{-18}
focal adhesion	*COL6A6*	rs16830219	3.0×10^{-4}	8.6×10^{-17}
focal adhesion	*COL5A1*	rs12554098	8.7×10^{-4}	5.9×10^{-16}
endometrial cancer	*CTNNA2*	rs11687661	1.1×10^{-4}	9.6×10^{-5}
endometrial cancer	*TCF7L1*	rs2568222	1.6×10^{-4}	5.7×10^{-11}
endometrial cancer	*MAP2K2*	rs350916	5.9×10^{-4}	4.9×10^{-6}
glioma	*CAMK2D*	rs6842886	7.0×10^{-4}	1.1×10^{-5}
glioma	*CAMK2D*	rs10488958	4.0×10^{-4}	4.7×10^{-9}
glioma	*CAMK2D*	rs12512765	9.7×10^{-4}	1.9×10^{-7}
glioma	*CDK6*	rs2282979	2.6×10^{-4}	1.1×10^{-27}
glioma	*MAP2K2*	rs350916	5.9×10^{-4}	8.5×10^{-6}
pathways in cancer	*CTNNA2*	rs11687661	1.1×10^{-4}	8.7×10^{-5}
pathways in cancer	*TCF7L1*	rs2568222	1.6×10^{-4}	1.2×10^{-11}
pathways in cancer	*DVL3*	rs2175525	8.8×10^{-4}	3.1×10^{-34}
pathways in cancer	*CDK6*	rs2282979	2.6×10^{-4}	1.2×10^{-27}
pathways in cancer	*MAP2K2*	rs350916	5.9×10^{-4}	4.3×10^{-6}
tight junction	*CTNNA2*	rs11687661	1.1×10^{-4}	1.7×10^{-4}
tight junction	*RRAS2*	rs2970332	7.6×10^{-4}	1.1×10^{-5}
regulation of actin cytoskeleton	*VAV3*	rs17485868	4.2×10^{-5}	3.9×10^{-20}

regulation of actin cytoskeleton	*RRAS2*	rs2970332	7.6×10^{-4}	2.2×10^{-5}
regulation of actin cytoskeleton	*MAP2K2*	rs350916	5.9×10^{-4}	1.4×10^{-6}

Marginal *P* value was calculated by logistic regression models while adjusted for age, gender and the first five principal components.
TMLE *P* value was estimated using R package (TMLE) for each SNP while accounted for other SNPs within the same biological pathway.

Supplementary Table S6. Overrepresented GO categories among the top results of the CCLS Hispanic GWAS on childhood ALL

GO category	Function	*P* value	*P* $_{FDR}$
GOTERM_BP_FAT	cell morphogenesis involved in neuron differentiation	2.1×10^{-07}	3.3×10^{-04}
GOTERM_BP_FAT	cell morphogenesis involved in differentiation	1.1×10^{-06}	0.001
GOTERM_BP_FAT	cellular component morphogenesis	1.3×10^{-06}	0.002
GOTERM_BP_FAT	cell morphogenesis	1.9×10^{-06}	0.003
GOTERM_BP_FAT	cell motion	2.5×10^{-06}	0.004
GOTERM_BP_FAT	cell projection organization	2.8×10^{-06}	0.004
GOTERM_BP_FAT	neuron projection morphogenesis	1.3×10^{-05}	0.022
GOTERM_BP_FAT	neuron development	3.1×10^{-05}	0.051
GOTERM_BP_FAT	axon guidance	3.5×10^{-05}	0.056
GOTERM_BP_FAT	post-embryonic development	3.5×10^{-05}	0.057
GOTERM_BP_FAT	axonogenesis	3.9×10^{-05}	0.063
GOTERM_BP_FAT	cell projection morphogenesis	4.5×10^{-05}	0.072
GOTERM_BP_FAT	cell part morphogenesis	6.4×10^{-05}	0.101
GOTERM_BP_FAT	neuron projection development	6.4×10^{-05}	0.115
GOTERM_BP_FAT	neuron differentiation	8.7×10^{-05}	0.144
GOTERM_BP_FAT	sensory organ development	1.4×10^{-04}	0.236
GOTERM_CC_FAT	cell leading edge	2.4×10^{-04}	0.317

Abbreviations: BP, biological process; CC, cellular component
List of 17 most significantly overrepresented GO categories for childhood ALL (cutoff for significant SNPs: *P*<0.001).

Supplementary Table S7. Overrepresented BioCarta pathways among the top results of the CCLS Hispanic GWAS on childhood ALL

BioCarta pathway from DAVID	Genes	*P* value	*P* $_{\text{adjust}}$
Integrin Signaling Pathway	*MAP2K2 TLN1 TNS1*	0.036	0.871

SNPs which showed association with childhood ALL (*P*<0.001) were included in this study and were mapped backed to regions on the genome, and the predicted candidate genes were used for analysis. The number of observed genes was compared to the number of expected genes per pathway. The top ranking BioCarta pathway is shown.

Supplementary Table S8. Overrepresented KEGG pathways among the top results of the CCLS Hispanic GWAS on childhood ALL using Webgestalt and DAVID software

KEGG pathway from Webgestalt	*P* value	KEGG pathway from DAVID	*P* value
Axon guidance	9.64×10^{-08}	Axon guidance	2.3×10^{-04}
Protein digestion and absorption	2.71×10^{-05}	Melanogenesis	0.016
Melanogenesis	7.82×10^{-05}	Leukocyte transendothelial migration	0.028
Leukocyte transendothelial migration	2.12×10^{-04}	Focal adhesion	0.046
Focal adhesion	2.82×10^{-04}	Endometrial cancer	0.092

SNPs which showed association with childhood ALL (*P*<0.001) were included in this study and were mapped backed to regions on the genome, and the predicted candidate genes were used for analysis. The number of observed genes was compared to the number of expected genes for each KEGG pathway. The top-5 ranking KEGG pathways per pathway classification tool are shown.

Chapter 5
CONCLUSIONS and FUTURE DIRECTIONS

CONCLUSIONS

Acute lymphoblastic leukemia (ALL) is the most common type of childhood malignancy and represents a significant public health burden (1). Ethnic differences in the risk of ALL are well-recognized as the incidence of ALL is nearly 20% higher among Hispanics than non-Hispanic Whites in California (2). This higher risk is possibly due to an increased prevalence of ALL risk alleles in populations with Native American ancestry, as well as ethnic differences in exposure to environmental risk factors (3, 4). However, most studies conducted thus far focus on individuals with European ancestry. Extending the research to Hispanics is crucial in order to understand ethnic-specific patterns of disease by identifying causal variants and their potential interactions with the environment (5). Understanding ethnic-specific similarities and differences can help us better understand the etiology of childhood ALL with the ultimate goal of translating the knowledge into disease prevention and possibly targeted treatment.

Childhood ALL is a cancer of immature lymphoid progenitor cells that have encountered a series of alterations within key cellular pathways. A growing body of literature from genome wide association studies (GWAS) of non-Hispanic Whites (NHW) supports the hypothesis that genetic variation in key regulatory genes that direct B-lymphocyte differentiation, including *IKZF1*, *CEBPE*, and *ARID5B*, play a role in the etiology of ALL (6). Genetic susceptibility loci explain only a small portion of cases and evidence that environmental exposures contribute to overt development of childhood ALL. For example, early life exposure to infection is hypothesized to play an important role in ALL. The aim of the current work was to investigate prior reported associations in a Hispanic population and to extend analyses to investigate broader associations with biological pathways that may be involved in the development of leukemia. Additionally, the dissertation sought to identify genetically susceptible subgroups in which a delay in exposure to infections may have differential influences on the risk of developing childhood ALL. The specific aims of this dissertation are to:

(i) Assess previously reported SNP associations within the non-Hispanic White and Hispanic populations of the California Childhood Leukemia Study (CCLS).

(ii) Comprehensively assess the genetic variants within the three previously identified genes (*IKZF1*, *CEBPE*, and *ARID5B*) and evaluate their potential interactions with surrogates for early life infections in the Hispanic population.

(iii) Apply newly developed bioinformatics tools and causal inference methods to prioritize biological pathways and identify a list of candidate gens for future investigation.

These projects highlight the possibility of delineating the properties of variants previously found in the non-Hispanic White population in the Hispanic population and localization of new variants in which true functionality may reside, with an attempt to understand childhood ALL pathogenesis.

Chapter 2 aimed to validate selected single nucleotide polymorphisms (SNPs) identified in the previous GWAS (7, 8): five SNPs in *ARID5B* (rs7089424, rs10821936, rs7073837, rs10740055, and rs10994982), one SNP in *CEBPE* (rs2239633), and two SNPs in *IKZF1* (rs4132601, and rs11978267) in both CCLS non-Hispanic White and Hispanic populations. We found genetic variants in the *ARID5B* and *CEBPE* genes were associated with the development of ALL in both Hispanic and non-Hispanic White populations. Risk estimates were in the same direction in both groups and strengthened when restricted to the B-cell hyperdiploid ALL subtype, as with the previously published GWAS (7, 8). In contrast, *IKZF1* variants displayed varying susceptibility loci between two populations. Furthermore, we pursued gene-environment analyses with the *a priori* hypothesis that genetic variants on these B-cell development genes were associated with ALL risk, modified by early life infection experiences such as daycare attendance and birth order. However, no significant multiplicative interaction was observed for these eight SNPs and surrogates for early life exposure to infections, after controlling for multiple comparisons. Further investigations are needed to fine-map susceptibility loci around these three genes and identify additional environmental factors that may modulate the effects of these loci. More effort is also required to discover whether the genetic determinants of childhood ALL are population-specific or overlap between these populations. In summary, this chapter provides additional evidence for inherited genetic childhood ALL susceptibility variants in these three candidate genes.

Chapter 3 comprehensively assessed three B-cell development genes (*ARID5B*, *CEBPE*, and *IKZF1*) and the joint effect of the genetic variants within these genes and surrogates for early life infections on the risk of ALL, using Hispanic GWAS data. Significant associations between genotypes at 7p12.2 (*IKZF1*), 10q21.2 (*ARID5B*), and 14q11.2 (*CEBPE*) and ALL risk were identified and the effects for *ARID5B* and *CEBPE* were most prominent in the hyperdiploid ALL subtype. Among these genetic variants, rs4132601 and rs11980379 genotype were correlated with *IKZF1* mRNA expression level, suggesting the SNPs might be in linkage disequilibrium with a casual variant. Conditional haplotype analysis was used to select SNPs that had independent effect on ALL risk for further gene and environment analyses. Evidence for multiplicative interactions between the selected genetic variants and surrogates for early life infections with ALL risk was not observed. The findings underscore the importance of B-cell development genes (*IKZF1*, *ARIDB5*, and *CEBPE*) and that these variants collectively play a major role in the development of childhood ALL. Further investigations are needed to sequence these gene regions and to investigate variants in the context of ALL treatment. A more refined measure of early life infections among Hispanics should be considered, such as the total number of people living in the household and/or parental or other child's social contacts. Future studies utilizing gene expression profiles and animal studies will greatly enhance our understanding of the underlying biological mechanisms for ALL.

Chapter 4 successfully identified several overrepresented biological pathways in childhood ALL among Hispanics, including axon guidance, protein digestion and absorption, melanogenesis, leukocyte transendothelial migration, focal adhesion, endometrial cancer, glioma, pathways in cancer, tight junction, and regulation of actin cytoskeleton pathways. Among the different ALL disease subtypes, pathway analyses results suggested that hyperdiploid B-ALL and *TEL-AML1* ALL involve distinct biological mechanisms compared to ALL and focal adhesion is a shared mechanism between different ALL disease subtypes. The findings demonstrate the successful application of bioinformatics tools to indentify biologic pathways that may serve as future potential therapeutic targets. Additionally, enriched biological pathways seem to be subtype-specific for ALL, compatible with the different etiologies hypotheses. By incorporating data reduction and causal inference methods (least absolute shrinkage and selection operator (LASSO) and targeted maximum likelihood estimation (TMLE)), a list of candidate genes has been identified in ALL while accounting for the correlation between SNPs, including *COL6A6*, *COL5A1*, *DVL1*, *TCF7L1*, *MAP2K2*, *VAV3*, *CTNNA2*, *CDK6*, *RRAS2*, and *CAMK2D* warrant future investigations. The findings show that newly developed bioinformatics tools and causal inference methods may shed light on relevant biological processes of childhood ALL and illuminate new candidate genes for future functional studies. Further research is needed to apply different pathway identification methods and to integrate gene expression and transcriptome data to further understand the pathogenesis of childhood ALL. Since the majority of genes in the genome are relatively unknown and their biological function still needs to be established, further studies should also consider using computational predictions of cellular processes from genomic and molecular information to better characterize gene function.

FUTURE DIRECTIONS

The present work confirmed associations with genetic variants previously identified in studies of non-Hispanic Whites among Hispanics helping to elucidate similarities and differences in genetic structure between two populations. The susceptibility loci for ALL are concentrated to genes directly related to hematopoietic differentiation and development (*IKZF1*, *ARIDB5*, and *CEBPE*) and these associations may vary by genetic ancestry background. We further hypothesize that genetic predisposition to ALL might be mediated by multiple genes involved in the same biological pathways. Results from incorporating pathway-based analyses and novel causal inference methods provide a new research approach to efficiently integrating biological information from multiple SNPs with weaker effects and accounting for confounding effects simultaneously to identify a set of targeted genes that merit further exploration. The work described in the dissertation provides a few insights into future research directions, including (1) different approaches for detecting gene and environment interactions, (2) investigation of rare variants, and (3) the use of newly developed bioinformatics tools and the Encyclopedia of DNA elements (ENCODE) project (9).

Understanding the relationships between genetic polymorphisms and environmental exposures can help identify high-risk subgroups in the population and have important implications for personalized medicine. A conventional gene-environment analysis restricts the search for interactions to cases where one or both factors show a marginal association (10). This approach can be more powerful than exhaustive pair-wise scans but risks missing potentially relevant interactions between variants with weak or non-significant marginal effects. In some circumstances, even though the marginal effects are not detectable, the genetic association may only occur at a specific subtype depending on environmental exposures or conversely, the environmental factor may only act as a risk factor in the presence of a susceptible genotype (11). Another aspect for detecting gene-environment interaction is any interaction is scale dependent, and it is therefore essential to state whether the presence of an interaction is a departure from an additive or multiplicative model on a scale of absolute risk or odds ratio. Sample size is another concern for gene-environment analyses. The key determinants of sample size requirement are the prevalence of the exposures, the allele frequency, the mode of inheritance, and the interaction OR. Even though the work presented in this dissertation only has limited power to make definitive inference on the gene-environment interactions reported, the CCLS is one of the largest case-control studies for childhood leukemia in the United States and has successfully collected comprehensive data on genetics and environmental exposures. Continued recruitment and international collaboration by means of large consortia such as the Childhood Leukemia International Consortium (CLIC) will be needed to attain sufficient power to confirm the reported interactions. Further efforts involving data harmonization between studies will be required in order to make a definitive conclusion (12).

Due to the fact that GWAS have only identified a small fraction of the common susceptibility loci of low penetrance, increasing attention has been given to the possibility that complex networks between genes and rare variants might account for some of the missing heritability (13). For pathway analyses, there are two types of approaches applicable to GWAS research (14). In this dissertation, the first type of pathway-based approaches is used, '*p*-value enrichment approach', which tests for overrepresentation of SNPs associated with disease within a pathway. The other approach is proposed by Wang et al., which adapted the gene set enrichment analysis (GSEA) method to GWAS, examining the distribution of association between outcome and genes within a pathway versus that between outcome and other genes using a modified Kolmogorov-Smirnov test (15). This approach uses individual-level SNP genotypes information to derive gene-level and pathway-level test statistics and usually requires phenotype permutations. It will be worthwhile to compare the findings using two different types of pathway analyses approaches.

The rationale behind the rare variants hypothesis is that genetic contributions to complex diseases arise from the interactions of many uncommon variants (16). Rare variants, which have minor allele frequencies (MAFs) between 0.1 and 1%, are often evolved from more recent mutations and less subjected to natural selection. Thus, association studies of rare variants hold the potential for identifying genetic components that are functionally relevant and explain a larger proportion of inherited susceptibility (17). Several association tests have been proposed for rare variants. These tests are often involved with pooling or collapsing multiple rare variants, such as combined multivariate and collapsing (CMC) test. Another more flexible method is developed by Wu et al, which used the sequence kernel association test (SKAT) to test the association between rare and common variants and disease status (18). With the success of whole genome sequencing, rare and private variants with moderate to large effects on many complex traits might be discovered.

One of the challenges in the “post-GWAS” era is to characterize biological functions of the identified risk loci and their relevance to disease etiology. These biological insights can be translated to clinical benefits, including biomarker screening and treatment specification (19). The availability of next-generation sequencing and the coverage of ENCODE annotations have enhanced our understanding of genetic variants across the entire genome and also the ability to link anonymous associations to a function element. SNPs identified by GWAS are showed to be enriched within non-coding functional elements, with a majority residing in or near protein-coding genes (9). Combining ENCODE information and the profiling of epigenetic and transcriptomic data with allele-specific information derived from deep sequencing of the target regions will provide specific insights on the impact of putative causal variants, with the potential to reveal alleles with more impact on the molecular and clinical level.

Little is known about the causes of childhood leukemia. The use of contemporary technologies to identify genetic alterations in ALL has been tremendously informative, but much work remains to be done. Studies on childhood leukemia are subject to small sample sizes due to the rarity and heterogeneity of the disease. Collaborating with international consortia can further confirm the study findings and identify new genetic susceptibility loci of ALL with greater power, ultimately improving our understanding of leukemogenesis and facilitating disease prevention and identification of treatment targets for ALL.

REFERENCES

1. Pui CH, Evans WE. Treatment of acute lymphoblastic leukemia. N Engl J Med 2006;354(2):166-78.
2. Campleman SL WW. Childhood cancer in California 1988 to 1999 Volume I: birth to age 14. Sacramento, CA: California Department of Health Services, Cancer Surveillance Section. 2004:16-17.
3. Walsh KM, Chokkalingam AP, Hsu LI, Metayer C, de Smith AJ, Jacobs DI, et al. Associations between genome-wide Native American ancestry, known risk alleles and B-cell ALL risk in Hispanic children. Leukemia 2013.
4. Greaves M. Infection, immune responses and the aetiology of childhood leukaemia. Nat Rev Cancer 2006;6(3):193-203.
5. Rosenberg NA, Huang L, Jewett EM, Szpiech ZA, Jankovic I, Boehnke M. Genome-wide association studies in diverse populations. Nat Rev Genet 2010;11(5):356-66.
6. Inaba H, Greaves M, Mullighan CG. Acute lymphoblastic leukaemia. Lancet 2013;381(9881):1943-55.
7. Papaemmanuil E, Hosking FJ, Vijayakrishnan J, Price A, Olver B, Sheridan E, et al. Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. Nat Genet 2009;41(9):1006-10.
8. Trevino LR, Yang W, French D, Hunger SP, Carroll WL, Devidas M, et al. Germline genomic variants associated with childhood acute lymphoblastic leukemia. Nat Genet 2009;41(9):1001-5.
9. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489(7414):57-74.
10. Thomas D. Gene--environment-wide association studies: emerging approaches. Nat Rev Genet 2010;11(4):259-72.
11. Murcray CE, Lewinger JP, Gauderman WJ. Gene-environment interaction in genome-wide association studies. Am J Epidemiol 2009;169(2):219-26.
12. Metayer C, Milne E, Clavel J, Infante-Rivard C, Petridou E, Taylor M, et al. The Childhood Leukemia International Consortium. Cancer Epidemiol 2013;37(3):336-47.
13. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature 2009;461(7265):747-53.
14. Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. Nat Rev Genet 2010;11(12):843-54.
15. Wang L, Zhang B, Wolfinger RD, Chen X. An integrated approach for the analysis of biological pathways using mixed models. PLoS Genet 2008;4(7):e1000115.
16. Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs. rare allele hypotheses for complex diseases. Curr Opin Genet Dev 2009;19(3):212-9.
17. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet 2008;40(6):695-701.
18. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 2011;89(1):82-93.
19. Freedman ML, Monteiro AN, Gayther SA, Coetzee GA, Risch A, Plass C, et al. Principles for the post-GWAS functional characterization of cancer risk loci. Nat Genet 2011;43(6):513-8.