Research Article

Cancer Epidemiology, Biomarkers & Prevention

Pathway Analysis of Genome-wide Association Study in Childhood Leukemia among Hispanics

Ling-I Hsu[1], Farren Briggs[2], Xiaorong Shao[1], Catherine Metayer[1], Joseph L. Wiemels[3], Anand P. Chokkalingam[1], and Lisa F. Barcellos[1]

Abstract

Background: The incidence of acute lymphoblastic leukemia (ALL) is nearly 20% higher among Hispanics than non-Hispanic Whites. Previous studies have shown evidence for association between risk of ALL and variation within *IKZF1*, *ARID5B*, *CEBPE*, *CDKN2A*, *GATA3*, and *BM1-PIP4K2A* genes. However, variants identified only account for <10% of the genetic risk of ALL.

Methods: We applied pathway-based analyses to genome-wide association study (GWAS) data from the California Childhood Leukemia Study to determine whether different biologic pathways were overrepresented in childhood ALL and major ALL subtypes. Furthermore, we applied causal inference and data reduction methods to prioritize candidate genes within each identified overrepresented pathway, while accounting for correlation among SNPs.

Results: Pathway analysis results indicate that different ALL subtypes may involve distinct biologic mechanisms. Focal adhesion is a shared mechanism across the different disease subtypes. For ALL, the top five overrepresented Kyoto Encyclopedia of Genes and Genomes pathways include axon guidance, protein digestion and absorption, melanogenesis, leukocyte transendothelial migration, and focal adhesion ($P_{FDR} < 0.05$). Notably, these pathways are connected to downstream *MAPK* or *Wnt* signaling pathways which have been linked to B-cell malignancies. Several candidate genes for ALL, such as *COL6A6* and *COL5A1*, were identified through targeted maximum likelihood estimation.

Conclusions: This is the first study to show distinct biologic pathways are overrepresented in different ALL subtypes using pathway-based approaches, and identified potential gene candidates using causal inference methods.

Impact: The findings demonstrate that newly developed bioinformatics tools and causal inference methods can provide insights to furthering our understanding of the pathogenesis of leukemia. *Cancer Epidemiol Biomarkers Prev; 25(5); 815–22. ©2016 AACR.*

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 diagnoses (2). Strong evidence for increased risk of ALL due to sex, age, race/ethnicity, prenatal exposure to X-rays, therapeutic radiation, and specific genetic syndromes has been established (3). Direct evidence for inherited genetic susceptibility is demonstrated by the high risk of ALL associated with Bloom syndrome, neurofibromatosis, ataxia telangiectasia, and constitutional trisomy 21 (3). 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 (4). 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 (5–7).

Current evidence suggests that leukemia results from chromosomal alterations and genetic variations that disrupt the normal process of lymphoid progenitor cell differentiation (1). Around 75% of childhood ALL cases have chromosomal aberrations that can be detected by karyotyping, FISH, or other molecular techniques (8). In B-cell precursor 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* (8, 9). Hyperdiploid and *TEL-AML1* rearranged childhood ALL account for approximately 25% and 22% of the entire childhood ALL populations, respectively (10). It is known that different cytogenetic subtypes have different disease prognoses, and are suspected to have distinct underlying biologic mechanisms (11).

The first genome-wide association study (GWAS) in childhood ALL was published in 2009 with a focus on Caucasian populations (12, 13), and subsequent GWA studies in diverse populations have confirmed previous genetic associations and identified new susceptibility loci (5, 14–17). These studies confirmed genetic contributions to childhood ALL susceptibility, and include variation within *IKZF1* (7p12.2), *ARID5B* (10q21.2), *CEBPE* (14q11.2), *CDKN2A* (9p21.3), *GATA3* (10p14), and *BM1-PIP4K2A* (10p12.31-12.2). However, variants within these loci account for <10% of the overall estimated genetic risk of leukemia (18).

GWAS are focused on the analysis of single markers, and depending on sample size, may lack statistical power to uncover small effects (OR <2.0) conferred by most common genetic

[1]School of Public Health, University of California, Berkeley, Berkeley, California. [2]Department of Epidemiology and Biostatistics, School of Medicine, Case Western Reserve University, Cleveland, Ohio. [3]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, California.

Note: Supplementary data for this article are available at Cancer Epidemiology, Biomarkers & Prevention Online (http://cebp.aacrjournals.org/).

Corresponding Author: Ling-I Hsu, University of California, Berkeley, 1995 University Avenue, Suite 460, Berkeley, CA 94704-7394. Phone: 847-418-7362; Fax: 510-642-9319; E-mail: lingi.hsu@berkeley.edu

doi: 10.1158/1055-9965.EPI-15-0528

variants identified, to date, for complex diseases. Some variants may be associated with disease status, but may not reach a stringent genome-wide significance threshold ($P = 5 \times 10^{-8}$). Complementary approaches to traditional GWAS analysis have been developed, including pathway-based analysis. In pathway analysis, a group of related genes in the same biologic functional pathway are jointly tested for association with a disease of interest (19, 20). The method is used to help prioritize biologic pathways most likely to be involved in the disease etiology, and to identify new loci not previously detected through GWAS. Several published studies have demonstrated that multiple related genes in the same functional pathway may confer disease susceptibility to breast cancer, Parkinson disease, and Crohn disease (21).

In the current study, we applied pathway-based analyses to GWAS data from the California Childhood Leukemia Study (CCLS) among Hispanics, and tested for evidence of biologic functions that are significantly enriched in ALL. Furthermore, we compared whether different biologic pathways were overrepresented in major leukemia subtypes, including B-cell ALL, hyperdiploid B-ALL, and *TEL-AML1* ALL. Finally, we utilized the targeted maximum likelihood estimation (TMLE) method with the least absolute shrinkage and selection operator (LASSO) to identify a list of candidate genes within each significantly enriched pathway in ALL, while accounting in models, for the complex correlation between SNPs.

Materials and Methods

Study populations

The CCLS is population-based case–control study. 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 4 (1995–1999), and later expanded to 9 (2000–2008), 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 ethnicity (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 reported previously (22). A total of 86% of case subjects determined eligible consented to participate, and 86% of controls subjects participated among those contacted and considered eligible (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 ALL cases and 454 controls) in the CCLS who enrolled and were interviewed subjects between 1995 and 2008, and for whom archived newborn blood spot specimens were available. Detailed cytogenetic classification was described previously (24). Immunophenotypic 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 ($\geq$20%) were classified as B-cell ALL (25). 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 (26). *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 dataset 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, principal component analysis was used as implemented in EIGENSTRAT (27). For pathway analyses, we selected SNPs that showed marginal associations ($P < 0.001$) with ALL and each ALL disease subtypes (B-cell ALL, hyperdiploid B-ALL, and *TEL-AML1* ALL), in models adjusted for age, gender, and the first five principal components, using PLINK v1.07. At the significance threshold of 0.001, given our sample size, power was present to detect an OR of 1.7 at a MAF of 15% for childhood ALL and B-cell ALL; an OR of 2.2 at a MAF of 15% for hyperdiploid B-ALL; and an OR of 2.7 at a MAF of 20% for *TEL-AML1* ALL. Therefore, all post-quality control SNPs for analysis were first filtered on the basis of 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; ref.28) and Database for Annotation, Visualization and Integrated Discovery, DAVID V6.7 (29). We used three pathway resources for this investigation: the BioCarta pathway database (30), the Kyoto Encyclopedia of Genes and Genomes (KEGG; ref.31), and the Gene Ontology (GO) database (32). Furthermore, 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" (28) and DAVID "KEGG" (29). Second, we investigated two additional pathway classification resources in DAVID: GO and BioCarta databases (30).

To identify functional categories with significant evidence for 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 determine significance (28). In DAVID, evidence for gene enrichment in annotated pathways is evaluated using a modified Fisher exact test to test for significance (29). The Benjamini and Hochberg (BH) procedure controlling for the false discovery rate (FDR) was performed for each pathway analysis tool to adjust for multiple comparisons (33).

After identifying the enriched pathways in ALL, SNPs with the most significant *P* value for each gene were selected to construct an 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, using SAS 9.2.

TMLE

To further generate a list of candidate genes within each identified biologic pathway, TMLE incorporating LASSO was applied, as implemented in the R package (34, 35). 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 (34, 35). TMLE helps reduce the bias for the targeted parameter, and provides formal statistical inference. Details of the TMLE method are explained elsewhere (34, 35). 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 within genes 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 genes in the same biologic pathway. On the basis of the *P* values estimated from TMLE method, we generated a list of candidate genes for each identified pathway ($P < 0.05$). Individuals with no missing genetic data were included ($n = 764$). All SNPs were prescreened: those with correlations less than 0.1 or greater than 0.9 with SNPs of interest were excluded in the model as SNPs that have independent effects or are highly correlated each other could not provide additional information to the model (34, 35).

Results

Study characteristics of 777 Hispanics (323 cases and 454 controls) are described in Table 1. The distributions of sex, age, and race/ethnicity were similar between cases and controls. As Hispanics are a recently admixed group (36), a proportion (34%–37%) of our Hispanic population reported "Mixed or Other" race and 49% to 51% of them reported "White and Caucasian" race. The frequency of hyperdiploid ALL (>50 chromosomes) was 30%, and the frequency of *TEL-AML1* ALL was 8%. The number of genes with SNPs reaching significance levels of 0.05, 0.01, 0.005, and 0.001 in the CCLS GWAS dataset are shown in Supplementary Table S1. A stringent significant *P* of 0.001 for pathway analysis was used for analysis. Guided by the *P* cutoff and power calculations for different disease subtypes, we mapped 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 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, year (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 presents the comparisons between the top 10 KEGG pathways in different ALL disease subtypes. The focal adhesion pathway was common across all ALL disease subtypes. This biologic pathway physically connects the extracellular matrix to the cytoskeleton and has long been speculated to mediate cell migration (37). The overrepresented biologic 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 biologic pathways compared with ALL. A total of 60% of the overrepresented pathways in childhood hyperdiploid B-ALL demonstrating significance are different compared with ALL, including bacterial invasion of epithelial cells and metabolic pathways. Overrepresented pathways in childhood *TEL-AML1* demonstrating significance also show similarity with childhood ALL, including tight junction, axon guidance, and focal adhesion, and some differences, including cell adhesion molecules (CAM), 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 Tables S2–S4.

Table 3 presents the top 10 ranked KEGG pathways enriched in childhood ALL which include cancer-related pathways (i.e., pathways related to cell proliferation, cell differentiation, and cell signaling). All 10 pathways remained significant after correction for multiple tests ($P < 0.05$). The top five KEGG pathways include axon guidance ($P_{FDR} = 5.1 \times 10^{-6}$), protein digestion and absorption ($P_{FDR} = 7.2 \times 10^{-4}$), melanogenesis ($P_{FDR} = 0.001$), leukocyte transendothelial migration ($P_{FDR} = 0.002$), and focal adhesion ($P_{FDR} = 0.002$). Among these pathways, leukocyte transendothelial migration pathway is essential for immune response and inflammatory reaction, which may be associated with leukemia pathogenesis. Notably, these pathways are connected to downstream *PI3K*, *MAPK*, or *Wnt* signaling pathway, and have been linked to multiple human malignancies (38–40).

Table 2. Comparisons between different ALL disease subtypes and associated biologic 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				√*
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 FDR using BH is smaller than 0.05.

[a]SNPs which showed association with each childhood ALL disease subtype ($P < 0.001$) 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/).

To examine the cumulative effects of the most significant SNPs for each pathway-associated 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 biologic pathway ($P_{trend} < 0.05$; Table 3). For example, the odds of developing childhood ALL significantly increased with each additional risk allele 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. On the basis of LASSO and TMLE results, we generated a list of candidate genes for each biologic pathway, while accounting for confounding effects of other genes within the same pathway. Table 4 presents an example of TMLE results for the focal adhesion pathway. SNPs within *VAV3*, *COL6A6*, and *COL5A1* genes have much more significant *P* values ($P < 0.05$) whereas other SNPs are no longer 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 the axon guidance pathway, variation within *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, variation within *VAV3*, and *CTNNA2* showed the strongest evidence of association (Table 3).

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

Discussion

This is the first study to show distinct biologic pathways are overrepresented in different leukemia disease subtypes using a pathway analysis approach. We further applied TMLE, incorporating LASSO, to select variants of importance and to provide formal statistical inferences for each significantly enriched pathway in ALL. The results demonstrate that newly developed bioinformatics tools and causal inference methods may illuminate new and biologically relevant pathways and genes to improve current understanding of pathogenesis in childhood leukemia. Our pathway-based association analysis reveals

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

KEGG pathway	Genes	Obs	Exp	Ratio	*P*	P_{adjust}	OR[a] (95% CI)	Important genes[b]
Axon guidance	*LRRC4C, PLXNC1, SLIT3, EPHB1, NTN1, UNC5B, GNAI1, NGEF*	8	0.55	14.46	9.64×10^{-8}	5.1×10^{-6}	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^{-5}	7.1×10^{-4}	2.03 (1.60–2.58)	*CPA2* *COL6A6* *COL5A1* *SLC7A8*
Melanogenesis	*TCF7L1, CAMK2D, DVL3, MAP2K2, GNAI1*	5	0.43	11.54	7.81×10^{-5}	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^{-4}	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^{-4}	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* *TCF7L1*
Tight junction	*MYL2, RRAS2, CTNNA2, GNAI1*	4	0.57	7.06	0.002	0.014	1.52 (1.24–1.87)	*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*

NOTE: 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 FDR using BH procedure.

Abbreviations: Exp, expected; KEGG, Kyoto Encyclopedia of Genes and Genomes; Obs, observed.

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

[b]Genes were selected on the basis of the *P* values from TMLE ($P < 0.05$).

Table 4. TMLE results suggest *VAV3*, *COL6A6*, and *COL5A1* are important genes within the focal adhesion pathway

Genes	SNPs	Marginal *P*	TMLE *P*
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

strong connections between leukemia development, immune regulation, and cancer-related pathways.

Pathway-based analyses provide a complementary approach to combine the effects of many loci; small contributions to overall disease susceptibility conferred by genes with weakly associated SNPs are otherwise missed by conventional GWAS analysis (19). By taking into account prior biologic knowledge about genes and pathways, we may have a better chance to identify novel genes and biologic mechanisms involved in disease pathogenesis (21). In addition, 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 investigation for ALL treatment such as tyrosine kinase inhibitors, Fms-like tyrosine kinase, *NOTCH1* inhibitors, and mTOR inhibitors (40). The enriched pathways identified in this study may further guide sophisticated targeted treatment strategies for ALL.

Subtypes of childhood ALL exhibiting specific molecular characteristics are known to be important in risk stratification and treatment specification at diagnosis (41). However, little is known about the underlying mechanisms leading to different ALL disease subtypes with specific chromosome abnormalities. Our results suggest that hyperdiploidy B-ALL and *TEL-AML* ALL are associated with different biologic pathways and perhaps different mechanisms for disease pathogenesis compared with childhood ALL. Pathways uniquely enriched in hyperdiploidy B-ALL include signal transduction and metabolism, whereas pathways involved with tissue and organ morphogenesis and the maintenance of cell and tissue structure and function were enriched in *TEL-AML* ALL. Interactions between transmembrane molecules lead to a direct or indirect control of cellular activities such as adhesion, proliferation, and apoptosis (42, 43). Results from the current study underscore the need to consider specific biologic pathways for ALL disease subtypes to further understand the disease etiology. The shared mechanism across different ALL disease subtypes, in this analysis, is focal adhesion, which consists of large protein complexes organized at the basal surface of cells. These proteins 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 (43, 44). Although a direct relationship between focal adhesion and leukemia has not been observed, elevated level of focal adhesion kinase has shown in various cancers, including thyroid, prostate, cervix, and colon cancer (37, 45).

The use of high-resolution genomic profiling to characterize the genetic basis of leukemogensis has indicated that high frequency of recurrent somatic alternations in key signaling pathways, including B-cell development differentiation, the *TP53*/*RB* tumor suppressor pathway and *Ras* signaling (46). Many of the genes encoded proteins with key functions in lymphoid development, lymphoid signaling, transcriptional regulations, or immune responses (46). The identified pathways that are overrepresented with childhood ALL in this study are consistent with current literature, mainly those associated with other malignances or cell communication and cell motility such as focal adhesion, tight junction, and regulation of actin cytoskeleton. All identified pathways are involved in different cellular processes that mediate signal transduction cascades leading to cell proliferation, cell migration, and cell adhesion. It has been demonstrated these pathways may contribute to the regulation of hematopoietic progenitor cells and are essential mediators for both immune and inflammatory responses (47, 48). For example, regulation of actin cytoskeleton and tight junction pathway are related to cell migration, which are required for many biologic processes, such as embryonic morphogenesis, immune surveillance, tissue repair and regeneration (44). Aberrant regulation of cell migration drives cancer progression and metastasis (49, 49). 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 (45, 50). 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. 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 (48, 50).

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*. The *VAV3* gene shows up as the top-ranked gene in several pathways. The gene is recruited and activated on EGF and insulin-like growth factor given its relation to regulating B-cell receptor signaling pathway and aberration of the gene may lead to B-cell malignancies (51, 52). *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 (53). *TCF7L1* and *DVL1* are members of the *Wnt* pathway (54). Aberrant activation of *Wnt* signaling pathway has been documented in various human cancers including myeloid leukemia (55). 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 (54). Enhanced *CDK6* expression has also been documented in lymphoma and leukemia (56, 57).

To our knowledge, this is the first report using pathway-based analyses and newly developed causal approaches to analysis of 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 is modulated by several critical cellular processes, including cell growth, differentiation,

survival, and migration. Another strength of our study is the detailed information on cytogenetic subtypes, which enables us to show distinct biologic mechanisms are involved with different disease subtypes. Furthermore, we employed the TMLE method to prioritize genes that may serve integral functions for tumor development. All prioritized genes have been previously linked to human malignances but not ALL.

Our results should be interpreted in the context of several 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 $P < 0.001$ to enable us to refine the gene sets and focus on SNPs most likely to represent a nonspurious association. Another important limitation of the pathway-based approach is the incomplete biologic annotation of the human genome, and the complete functional characterization of many human genes is unknown. Moreover, susceptibility loci in intergenic regions were 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 are more likely to contain significant SNPs by chance alone. In addition, 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 biologic 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 identification of cancer-related and inflammatory-related pathways supports the power of this methodologic framework to highlight pathways with established relevance to childhood leukemia etiology. In addition, the results highlight several strong candidate genes for further investigations. Future studies are needed to confirm and sequence the identified genes in a larger Hispanic childhood leukemia dataset and in other ethnic patient populations.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: L. Hsu, F. Briggs, L.F. Barcellos
Development of methodology: L. Hsu, F. Briggs, J.L. Wiemels, L.F. Barcellos
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): L. Hsu, C. Metayer, A.P. Chokkalingam, L.F. Barcellos
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): L. Hsu, F. Briggs, X. Shao, C. Metayer, L.F. Barcellos
Writing, review, and/or revision of the manuscript: L. Hsu, F. Briggs, X. Shao, C. Metayer, J.L. Wiemels, A.P. Chokkalingam, L.F. Barcellos
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): L. Hsu
Study supervision: C. Metayer, J.L. Wiemels, L.F. Barcellos

Acknowledgments

Participating hospitals and clinical collaborators included University of California Davis Medical Center (Dr. Jonathan Ducore), University of California San Francisco (Drs. Mignon Loh and Katherine Matthay), Children's Hospital of Central California (Dr. Vonda Crouse), Lucile Packard Children's Hospital (Dr. Gary Dahl), Children's Hospital Oakland (Dr. James Feusner), Kaiser Permanente Roseville (former Sacramento; Drs. Kent Jolly and Vincent Kiley), Kaiser Permanente Santa Clara (Drs. Carolyn Russo, Alan Wong, and Denah Taggar), Kaiser Permanente San Francisco (Dr. Kenneth Leung), and Kaiser Permanente Oakland (Drs. Daniel Kronish and Stacy Month). The authors also acknowledge the entire CCLS staff for their effort and dedication.

Grant Support

L. Hsu, X. Shao, C. Metayer, J.L. Wiemles, A.P. Chokkalingam, and L.F. Barcellos received grants from the National Institute of Environmental Health Sciences (PS42 ES04705 and R01 ES09137), the NCI (R25CA112355), and Children with Cancer, United Kingdom.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received July 20, 2015; revised February 5, 2016; accepted February 17, 2016; published OnlineFirst March 3, 2016.

References

1. Pui CH, Relling MV, Downing JR. Acute lymphoblastic leukemia. N Engl J Med 2004;350:1535–48.
2. Stiller CA, Parkin DM. Geographic and ethnic variations in the incidence of childhood cancer. Br Med Bull 1996;52:682–703.
3. Eden T. Aetiology of childhood leukaemia. Cancer Treat Rev 2010;36: 286–97.
4. Campleman SL, Wright WE. Childhood cancer in California 1988 to 1999 volume I: birth to age 14. Sacramento, CA: California Department of Health Services, Cancer Surveillance Section; 2004. p. 16–7.
5. Walsh KM, Chokkalingam AP, Hsu LI, Metayer C, de Smith AJ, Jacobs DI, et al. Associations between genome-wide NativeAmerican ancestry, known risk alleles and B-cell ALL risk in Hispanic children. Leukemia 2013;27: 2416–9.
6. Hsu LI, Chokkalingam AP, Briggs FB, Walsh K, Crouse V, Fu C, et al. 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. Cancer Causes Control 2015;26:609–19
7. Belson M, Kingsley B, Holmes A. Risk factors for acute leukemia in children: a review. Environ Health Perspect 2007;115:138–45.
8. Mullighan CG. The molecular genetic makeup of acute lymphoblastic leukemia. Hematology Am Soc Hematol Educ Program 2012;2012: 389–96.
9. Buffler PA, Wood SM, Suarez L, Kilian DJ. Mortality follow-up of workers exposed to 1,4-dioxane. J Occup Med 1978;20:255–9.
10. 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;26:902–9.
11. 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;60:864–71.
12. 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;41:1006–10.
13. 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;41:1001–5.
14. 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;34:1271–4.
15. 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;26:2561–4.
16. 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; 105:733–42.
17. 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;42:492–4.
18. 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;26:2212–5.
19. Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. Am J Hum Genet 2010;86:6–22.
20. Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. Nat Rev Genet 2010;11:843–54.
21. Torkamani A, Topol EJ, Schork NJ. Pathway analysis of seven common diseases assessed by genome-wide association. Genomics 2008; 92:265–72.
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;159:915–21.
23. Bartley K, Metayer C, Selvin S, Ducore J, Buffler P. Diagnostic X-rays and risk of childhood leukaemia. Int J Epidemiol 2010;39:1628–37.
24. 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;22:1600–11.
25. Hrusak O, Porwit-MacDonald A. Antigen expression patterns reflecting genotype of acute leukemias. Leukemia 2002;16:1233–58.
26. 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: 578–81.
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;38:904–9.
28. Zhang B, Kirov S, Snoddy J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res 2005;33:W741–8.
29. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 2009;4:44–57.
30. Nishimura D. Biotech Software & Internet Report. July 2004, 2(3):117–20.
31. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28:27–30.
32. 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:D258–61.
33. Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. Stat Med 1990;9:811–8.
34. Gruber S, van der Laan MJ. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. Int J Biostat 2010;6: Article 26.
35. van der Laan MJ, Rubin D. Targeted maximum likelihood learning. Int J Biostat 2006;2.
36. Baye TM, Wilke RA. Mapping genes that predict treatment outcome in admixed populations. Pharmacogenomics J 2010;10:465–77.
37. 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;5:505–15.
38. Geest CR, Coffer PJ. MAPK signaling pathways in the regulation of hematopoiesis. J Leukoc Biol 2009;86:237–50.
39. Downward J. Targeting RAS signalling pathways in cancer therapy. Nat Rev Cancer 2003;3:11–22.
40. Courtney KD, Corcoran RB, Engelman JA. The PI3K pathway as drug target in human cancer. J Clin Oncol 2010;28:1075–83.
41. Inaba H, Greaves M, Mullighan CG. Acute lymphoblastic leukaemia. Lancet 2013;381:1943–55.
42. Gumbiner BM. Cell adhesion: the molecular basis of tissue architecture and morphogenesis. Cell 1996;84:345–57.
43. Zhao X, Guan JL. Focal adhesion kinase and its signaling pathways in cell migration and angiogenesis. Adv Drug Deliv Rev 2011;63:610–5.
44. Sawada N. Tight junction-related human diseases. Pathol Int 2013;63: 1–12.
45. Dreesen O, Brivanlou AH. Signaling pathways in cancer and embryonic stem cells. Stem Cell Rev 2007;3:7–17.
46. Zhang J, Mullighan CG, Harvey RC, Wu G, Chen X, Edmonson M, et al. Key pathways are frequently mutated in high risk childhood acute lymphoblastic leukemia: a report from the Children's Oncology Group. Blood 2011;118:3080–7
47. Golias C, Tsoutsi E, Matziridis A, Makridis P, Batistatou A, Charalabopoulos K. Review. Leukocyte and endothelial cell adhesion molecules in inflammation focusing on inflammatory heart disease. In Vivo 2007; 21:757–69.
48. Infante E, Ridley AJ. Roles of Rho GTPases in leucocyte and leukaemia cell transendothelial migration. Philos Trans R Soc Lond B Biol Sci 2013; 368:20130013.
49. Muller WA. Mechanisms of leukocyte transendothelial migration. Annu Rev Pathol 2011;6:323–44.
50. 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.
51. 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;195:189–200.
52. 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;22:597–608.
53. 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;10:880–8.
54. Anastas JN, Moon RT. WNT signalling pathways as therapeutic targets in cancer. Nat Rev Cancer 2013;13:11–26.
55. 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;21:1638–47.
56. 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;24:167–81.
57. 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; 15:101–9.