Novel Gene and Network Associations Found for Acute Lymphoblastic Leukemia Using Case-Control and Family-Based Studies in Multiethnic Populations

Priyanka Nakka1,2, Natalie P. Archer3, Heng Xu4, Philip J. Lupo5, Benjamin J. Raphael6, Jun J. Yang7, and Sohini Ramachandran1,2

Abstract

Background: Acute lymphoblastic leukemia (ALL) is the most common childhood cancer, suggesting that germline variants influence ALL risk. Although multiple genome-wide association (GWA) studies have identified variants predisposing children to ALL, it remains unclear whether genetic heterogeneity affects ALL susceptibility and how interactions within and among genes containing ALL-associated variants influence ALL risk.

Methods: Here, we jointly analyzed two published datasets of case–control GWA summary statistics along with germline data from ALL case–parent trios. We used the gene-level association method PEGASUS to identify genes with multiple variants associated with ALL. We then used PEGASUS gene scores as input to the network analysis algorithm HotNet2 to characterize the genomic architecture of ALL.

Results: Using PEGASUS, we confirmed associations previously observed at genes such as ARID5B, IKZF1, CDKN2A/2B, and PIP4K2A, and we identified novel candidate gene associations. Using HotNet2, we uncovered significant gene subnetworks that may underlie inherited ALL risk: a subnetwork involved in B-cell differentiation containing the ALL-associated gene CEBPE, and a subnetwork of homeobox genes, including MEIS1.

Conclusions: Gene and network analysis uncovered loci associated with ALL that are missed by GWA studies, such as MEIS1. Furthermore, ALL-associated loci do not appear to interact directly with each other to influence ALL risk, and instead appear to influence leukemogenesis through multiple, complex pathways.

Impact: We present a new pipeline for post hoc analysis of association studies that yields new insight into the etiology of ALL and can be applied in future studies to shed light on the genomic underpinnings of cancer.

Introduction

Acute lymphoblastic leukemia (ALL) is the most common childhood cancer in Western countries, with a peak incidence range of 2 to 5 years of age (1–3). The early age of onset suggests that the etiology of ALL begins very early in development, possibly prenatally (4, 5). The risk of ALL also increases significantly in patients with certain congenital syndromes, such as Down syndrome and ataxia-telangiectasia (6, 7). In addition, there is a significantly higher risk of ALL in siblings of affected cases compared with those without ALL siblings, which is particularly evident in concordant cases of ALL in monozygotic twins (1, 2). Taken together, these observations suggest that germline genetic variation may contribute to ALL susceptibility; however, the genetic mechanisms that generate predisposition to ALL are not completely understood.

Development of childhood ALL is thought to be caused by (i) chromosomal translocations (such as TEL-AML1 fusions) or hyperdiploidy, which can happen in utero, followed by (ii) secondary somatic gene deletions or mutations that ultimately lead to disease (1, 2). Different underlying causes for the second, crucial step in the natural history of ALL (gene deletions or mutations that cause ALL) have been postulated, including aberrant reactions to infections in infancy and genetic variation in immune-response pathways (1). Germline variation can influence either step in this process, and case–control genome-wide association (GWA) studies have successfully identified ALL-associated SNPs in genes including ARID5B, IKZF1, CEBPE, PIP4K2A, CDKN2A/2B, and GATA3 (8–15). Supplementary Table S1 shows genes containing variants that have been associated with ALL in at least two GWA studies at a genome-wide significant level (P < 10−8). In spite of these findings, it remains unknown to what extent these genes interact to affect ALL risk.

Recent work from our group and others has yielded novel methods for exploring gene networks in the context of a GWA
To identify genes and gene subnetworks influencing genetic predisposition for ALL, we analyzed two datasets of P values from previously published case–control ALL GWA studies (14, 15) and P values from an ALL case–parent trio study (25, 26). We used the gene-level association method PEGASUS (20) to identify novel candidate genes associated with ALL. We then applied the gene set enrichment method DAVID (17, 18) to identify enriched functional categories in PEGASUS-identified gene-level associations with ALL. The PEGASUS gene-level P values or “gene scores” were then used as input to the HotNet2 algorithm (27) to uncover multiple novel gene interaction subnetworks that are significantly associated with ALL, shedding light on the underlying biological mechanisms that may cause genetic predisposition to ALL.

Materials and Methods

Subjects/datasets

For the “discovery stage,” we analyzed data from a case–control study of ALL with 1,773 affected children (see ref. 15 for details on the original study). Briefly, genome-wide SNP-level P values for 247,505 variants across the exome were obtained on 1,773 children of European descent with B-ALL and 10,448 non-ALL controls of European descent (15). All individuals were genotyped using the Illumina Infinium HumanExome array.

For the “replication stage,” we analyzed data from an independent multiethnic case–control study of ALL (14). Briefly, genome-wide SNP-level P values for 709,059 variants across the genome were obtained on 1,605 ALL case subjects of multiple ethnicities and 6,661 controls (see ref. 14 for further details). Xu and colleagues (14) inferred ancestry components using STRUCTURE (28) on these data to assign individuals genome-wide proportions of European, Native American, Asian, and African ancestry; European Americans were defined as individuals with greater than 95% European ancestry and Hispanic Americans as individuals with Native American ancestry that is greater than 10% and Native American ancestry that is greater than African ancestry. Following these guidelines, we classified 963 cases and 1,381 controls as European Americans and 305 cases and 1,088 controls as Hispanic American. We also classified 88 cases and 1,363 controls who had greater than 70% African ancestry as African Americans; however, we did not analyze this sample further due to its small sample size. All individuals were genotyped using the Affymetrix Human SNP Array 6.0.

To test for additional replication of gene-level association signals, we analyzed a dataset of 368 ALL case–parent trios. This population has been described previously (25, 26). Briefly, all individuals were genotyped using the Illumina Infinium HumanExome Bead Chip. SNP-level P values for 237,436 exonic variants throughout the genome were obtained through multinomial modeling, which was conducted using EMIM software (29).

All patient data described here have been previously analyzed and published (14, 15, 25, 26). In this study, we analyze summary statistics from these previous publications, with one exception: For the replication dataset (14), we obtained genotype data from dbGAP (project ID 6249, principal investigator: S. Ramachandran) to calculate empirical LD for PEGASUS. Please see the works by Xu and colleagues (15), Xu and colleagues (14), and Archer and colleagues (25) for details about institutional review of these studies and written informed consent.

Gene-level association testing

To identify genes associated with ALL in each of the two case–control datasets and the trio analysis dataset, we performed gene-level tests of association using PEGASUS (20). Briefly, individual SNP statistics were drawn from a $\chi^2$ distribution correlated by empirical LD, and the distribution of the sum of correlated $\chi^2$ statistics within a gene is the null distribution for gene-level statistics; this distribution was then numerically integrated to calculate a gene-level $P$ value with machine precision (20). We calculated gene-level $P$ values or “scores” for 19,000 genes using gene boundaries of 50,000 bp upstream and downstream of the genes to account for regulatory regions; gene start and end positions were downloaded from the UCSC Genome Browser. For PEGASUS testing on GWA results from Xu and colleagues (15), we used genotype data from the 1000 Genomes EUR population as proxies to calculate LD, as the case–control study included only individuals of European descent (30). We used genotype data from the multiethnic GWA study (14) to calculate empirical LD for PEGASUS analysis. We used LD information empirically calculated from the trios for the trio analysis-based gene-level test (25, 26).

In addition to calculating gene scores for the multiethnic GWA dataset (the “replication stage” dataset), we also calculate gene scores separately for inferred European American cases and Hispanic American cases from this dataset to compare gene-level association results between these two ancestries.

Pathway analysis

We performed pathway analysis with HotNet2 (27), a topology-based method for finding significantly associated subnetworks within protein–protein interaction (PPI) networks. HotNet2 uses directed heat diffusion along interaction networks where every gene, or a “node” in the network graph, has a “heat score” based on its gene score. Although originally developed for analyzing somatic mutation data from cancer datasets, HotNet2 has been used to uncover gene subnetworks significantly associated with common traits and diseases using P values from GWA studies on common variants (20). We used negative log-transformed gene scores generated by PEGASUS as heat scores in HotNet2.

We used HotNet2 to find gene interaction subnetworks containing genes that PEGASUS identifies as strongly associated with ALL. As described by Nakka and colleagues (20), HotNet2 does not perform well when too many genes are assigned similar heat scores, so we use a gene score threshold determined by local FDR (IFDR) for the GWA-based PEGASUS gene scores (15). We calculated IFDR for PEGASUS gene scores using the twilight R package.
package (31) and determined a cutoff for gene scores at the first "elbow" or inflection point in the graph of 1 – IFDR against gene scores (Supplementary Fig. S1).

In addition to gene scores, HotNet2 also requires a PPI network (32–37) as input. A primary concern for selecting a PPI network for our analysis here was the connectivity of previously associated ALL genes in the networks; without immediate neighbors in the network graph, a gene known to be associated with ALL will not be included in the resulting gene subnetworks output by HotNet2. We chose to combine the iRefIndex network (36) and KEGG pathway database (34, 35) and use the resulting combined network as input to HotNet2 because it contains at least 9 interactions for each ALL-associated gene (Supplementary Table S2).

As there is no straightforward way to test for replication of entire significant subnetworks identified using the discovery stage gene scores as input, we instead attempted to replicate the individual genes contained in these subnetworks in the replication gene score dataset to a nominal significance threshold (gene $P < 0.05$).

## Results

### PEGASUS results based on case–control GWA $P$-values

We performed discovery-stage PEGASUS analysis using case–control GWA $P$-values (15). ARID5B, IKZF1, CDKN2A/2B, and PIP4K2A all contain SNPs associated with ALL at a genome-wide significant level in previous GWAS and expression studies (8–13). We then tested for replication of the 42 resulting gene hits ($P < 10^{-5}$, Bonferroni-corrected for the number of haplotype blocks in the genome; ref. 38) using a replication dataset of gene-level $P$ values calculated from a second dataset of case–control GWA P values (Table 1; Supplementary Table S3; ref. 14). We find that eight genes, ARID5B, IKZF1, FIGNL1, CDKN2A, CDKN2B, DDC, PIP4K2A, and HLA-DQB1, are replicated (replication $P$ values $\leq 0.05$). We also find that only ARID5B is replicated in the trio-based gene-level PEGASUS analysis (replication $P = 1.61 \times 10^{-5}$), and we ascribe this to the small sample size of the trio data (Supplementary Table S4).

### HotNet2 results using GWA-based PEGASUS scores as input

The subnetwork in Fig. 1A shows multiple genes involved in hematopoiesis and leukemogenesis (39–42). MEIS1, PKNOXI, HOXA2, HOXA5, HOXA7, HOXA11, HOXA13, and HOXB4 (shown in the subnetwork) are homeobox genes, which encode the HOX transcription factors. HOX transcription factors bind to DNA and regulate genes involved in the differentiation of the embryo and also the differentiation, self-renewal, and proliferation of hematopoietic stem cells (39, 40). In leukemogenesis, a chromosomal translocation, such as t(12;21), creates the TEL-AML1 fusion gene, which retains binding domains necessary for the homing of hematopoietic progenitor cells to the bone marrow and the DNA-binding component of a transcription factor called core-binding factor (40). The fusion gene then initiates an abnormal transcriptional cascade that affects the HOX genes downstream (40). The altered transcriptional cascade affects the differentiation and self-renewal capacity of hematopoietic stem cells (39, 40). Leukemogenesis can also be triggered via the HOX-regulatory pathway through translocations involving the MLL gene (41, 42). MLL fusion proteins have enhanced transcriptional activity, which disrupts the normal pattern of HOX gene expression and leads to changes in self-renewal and growth of hematopoietic stem cells that eventually result in leukemia (40–42). We note that SNPs within these genes were only marginally significant (SNP $P$ values: $0.007 < P < 0.2$) in the GWA-level analysis and so would have been missed by standard approaches to interpreting GWA results. However, by using network analysis following PEGASUS analysis of GWA $P$ values, we were able to uncover significant gene networks containing these homeobox genes.

SNPs located in the gene CEBPE (Fig. 1B) have been previously associated with ALL in GWAS studies at genome-wide significant levels ($P = 4 \times 10^{-10}$ and $P = 5.6 \times 10^{-6}$; refs. 9, 43). CEBPE encodes CCAAT/enhancer binding protein epsilon, which suppresses myeloid leukemogenesis and is mutated in a subset of cases (9). Genes in the C/EBP family, such as CEBPG (also shown in the subnetwork), are involved in hematopoietic cell development, especially granulopoiesis (hematopoiesis of granulocytes), and are sometimes targeted by recurrent immunoglobulin heavy chain translocations in B-cell precursor ALL (9, 44). ATF5 (activating transcription factor 5) is a transcription factor that activates the transactivation activity of C/EBP family members upon stimulus by IL1B, a proinflammatory cytokine (45). Polymorphisms in the ATF5 gene were associated with outcome after treatment of ALL with asparaginase (46). MLLT6, which encodes myeloid/lymphoid or mixed-lineage leukemia; translocated to $6$, protein, is a gene that is commonly translocated in ALL to create an MLL fusion gene, which encodes a chimeric protein that ultimately leads to leukemia (47, 48). MLLT6 is part of a family of nuclear
transcription factors (48). JAM2 (junctional adhesion molecule 2; Fig. 1B) belongs to the immunoglobulin superfamily and is expressed by vascular endothelium and B lymphocytes. The level of JAM2 expression defines B-cell differentiation stages, and the encoded protein plays a role in the homing of B cells to lymphoid organs (such as the spleen, bone marrow, and lymph nodes); disruption of its normal activity leads to tumorigenesis (49). This subnetwork shows genes such as CEPTB, which contains genome-wide significant SNPs, interacting with other genes involved in hematopoiesis. These networks have not been identified in previous pathway analyses on ALL (24).

Additional significant subnetworks from our analyses are shown in Supplementary Figs. S2 and S3 and are annotated in Supplementary Tables S3–S8 (9, 12, 50–70).

Ethnicity-specific HotNet2 results

ALL is known to have higher incidence and a worse prognosis in patients with high levels of Native American ancestry (14, 71). We tested for germline signatures of this phenomenon by performing gene and network analysis of European American cases and controls and Hispanic American cases and controls in the multiethnic GWA dataset (14) separately (Fig. 2). We find that although there are gene hits (PEGASUS P < 10−6) shared between the two cohorts, such as ARID5B, there are also 18 and 3 genes that achieved significance in only the European American and Hispanic American cohorts, respectively (Fig. 2A). We also find that a significant subnetwork centered on MEIS1 is only identified in network analysis of the Hispanic American cohort-derived PEGASUS gene scores and is missed in network analysis of the European American cohort (Fig. 2B). These candidate genes were not identified in a previous analysis of pathways in Hispanic individuals and can represent useful targets for future functional validation (24).

Gene set enrichment analysis with DAVID

We used significant genes (gene P < 10−5) resulting from PEGASUS analysis on GWA P values as input to the gene set enrichment analysis method DAVID (17, 18) to test for genome annotation enrichment of GO terms and KEGG pathways. Five annotation clusters achieved a significant enrichment score (greater than 1.3). The first category contains genes that are enriched for GO terms involving regulation of monocyte, myeloid cell, and leukocyte differentiation, such as IKZF1, JUN, CSF1, ACIN1, and HIST4H4 (Table 2). Another category includes genes with annotations related to hematopoiesis and immune system development, such as DNASE2, CEBPA, KL65, CEBPE, IKZF1, CSF1, ACIN1, KL65, and FLVR1 (Table 2). These results broadly confirm our results from HotNet2 analysis in which we identified multiple gene subnetworks involved in B-cell differentiation and hematopoiesis.

Discussion

Here, we present a gene-level and network analysis of published case–control and family-based association studies that yield new insight into the genomic underpinnings of ALL. Using the gene-level association method PEGASUS, we confirmed and replicated associations at multiple genes previously associated with ALL in GWA studies, such as ARID5B, IKZF1, CDKN2A/B, and PIP4K2A, and we identified novel gene associations (Table 1; Supplementary Table S3). We also found that the gene ARID5B is replicated in gene-level analysis of a multiethnic family-based association study (Supplementary Table S4; refs. 25, 26). Our findings suggest that gene and network analysis can be used to draw on multiple data types (case–control and trio-based studies) and genotyping platforms (exome-wide or genome-wide SNP chips) to yield new insight into complex diseases like ALL; our approach can also be used post hoc on published GWA studies, increasing the return on investment of the GWA approach. Furthermore, we note that although the SNP panels used in the three datasets analyzed here vary drastically in both SNP density and content, using our method, we are able to generate datasets of gene scores of similar size that can be compared directly (Table 1; Supplementary Tables S3 and S4). We also note that the trio dataset had a small sample size of trios and several monomorphic loci, which are uninformative for trio-based analysis (154, 092 monomorphic SNPs of...
Gene and Network Analysis of Acute Lymphoblastic Leukemia

Figure 2.
Ethnicity-specific PEGASUS and HotNet2 results. Using PEGASUS, we calculated gene-level P values using GWA SNP P values from association studies on European American cases and controls and Hispanic American cases and controls (14). The Venn diagram in A, 18 significant gene hits (PEGASUS P values < 10^{-5}) in the European American cohort only (red), 3 significant gene hits in the Hispanic American cohort only (blue), and 5 significant gene hits in both cohorts (purple). B, A gene subnetwork found using PEGASUS gene scores derived from the Hispanic American cohort as input. Bold, italicized, and underlined genes are genes that are replicated in the European American cohort (PEGASUS P < 0.05).

237,436 total SNPs, or 64.9%), which may account for the small number of gene hits from the discovery dataset that were replicated in this dataset (Supplementary Table S4). In the future, PEGASUS can be used to jointly and quantitatively explore differences between candidate genes derived from both case–control and family-based association studies.

The goal of our study is similar to that of Hsu and colleagues (24), to identify genes and gene sets associated with ALL risk, but our approaches are quite distinct. First, PEGASUS (20) reports a gene score for each gene in the genome that is sensitive to genes containing multiple variants of moderate association with a trait of interest while controlling for LD, which may vary with ancestral background. The PEGASUS gene score allows for testing for significant associations at the gene level, for gene set enrichment analysis using known canonical pathways (Table 2), and for detection of novel gene subnetworks associated with a phenotype when used in conjunction with HotNet2 (Figs. 1 and 2). PEGASUS gene scores are not limited by preexisting annotations and allow for the calculation of FDRs. Hsu and colleagues (24) do not calculate gene scores, but instead use a GWA SNP-level threshold (P < 0.001) to identify candidate genes for gene set enrichment analysis. We also note that the pathways that Hsu and colleagues (24) identify do not contain a number of known ALL-associated genes, whereas our network and pathway results (Figs. 1 and 2; Table 2) contain genes such as CEPBE and IKZF1, which have been identified previously in GWA studies of ALL. The candidate genes identified by both these studies yield new insight into the pathogenesis of ALL; further studies may integrate both approaches and test whether different molecular subtypes of ALL are characterized by differing genetic architecture (see Table 2 in the work by Hsu and colleagues; ref. 24).

After network analysis with HotNet2 (27) using our PEGASUS results as input, we found multiple significant gene interaction networks containing genes previously associated with ALL and leukemogenesis, such as CEBPE and MEIS1 (Fig. 1). A subnetwork centering on CEBPE contains genes in the C/EBP family and other interacting genes, which are transcription factors involved in hematopoiesis and are thought to suppress leukemogenesis and, thus, may influence the development of ALL. In addition, we note that although MEIS1 and other HOX genes have been suspected to influence leukemogenesis (39–42), germline variants in MEIS1 have failed to achieve genome-wide significance in any GWA study performed to date on ALL (8–15). However, we do identify MEIS1 and other interacting HOX genes as significantly mutated in cases by using PEGASUS gene scores as input to network analysis with HotNet2. Thus, PEGASUS, along with gene set enrichment analyses or HotNet2, can be applied after case–control association studies to gain additional insight into associated loci that are missed by the GWA framework, thus yielding new insight into disease from previously published GWA studies.

We also uncover multiple novel gene interaction subnetworks that may influence ALL risk. For example, we identify a network centered around the TNKS pathway that is involved in miRNA-mediated transcriptional regulation that may be involved in ALL risk (58–63), and we uncover a gene subnetwork containing UNCG3B1 and other genes that plays an important role in innate and adaptive immunity (Supplementary Fig. S2; Supplementary Tables S6 and S7; refs. 64–66). An open question in the literature is whether genes associated with ALL in GWA studies (such as CEBPE, IKZF1, ARID5B, etc.) work in concert to influence the phenotype or through separate pathways. In our network analysis, we find genes such as CEBPE, MEIS1, and DDC are contained in distinct subnetworks. Thus, we conclude that ALL cases may contain heterogeneous sets of mutations that influence leukemogenesis via multiple subnetworks; however, further experiments are needed to test this result. Taken together, these network results provide new...
Table 2. Genome annotation enrichment clusters

<table>
<thead>
<tr>
<th>Annotation cluster 1</th>
<th>Enrichment Score: 2.09893050347324</th>
<th>Term</th>
<th>P</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0045657: positive regulation of monocyte differentiation</td>
<td>4.62E-04</td>
<td>JUN, CSF1, ACIN1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0045655: regulation of monocyte differentiation</td>
<td>9.16E-04</td>
<td>JUN, CSF1, ACIN1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0002763: positive regulation of myeloid leukocyte differentiation</td>
<td>0.001361426</td>
<td>IKZF1, JUN, CSF1, ACIN1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0045639: positive regulation of myeloid cell differentiation</td>
<td>0.007938727</td>
<td>IKZF1, JUN, CSF1, ACIN1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0045637: regulation of myeloid cell differentiation</td>
<td>0.01372659</td>
<td>HIST4H4, IKZF1, JUN, CSF1, ACIN1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0002761: regulation of myeloid leukocyte differentiation</td>
<td>0.014424008</td>
<td>IKZF1, JUN, CSF1, ACIN1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Annotation cluster 2</th>
<th>Enrichment score: 1.60058768735878</th>
<th>Term</th>
<th>P</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topological domain: Lumenal</td>
<td>0.002999392</td>
<td>TCIRG1, GCNT4, TPST2, ST6GAL2, GOLT1B, IGFB2R, LMAN2L, CSF1, ASPH2D, GALNT4, EXT1, AB0, PPAP2B, MOXD1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Golgi apparatus</td>
<td>0.01871437</td>
<td>TPST2, GCNT4, ST6GAL2, AP6G2, GOLT1B, LMAN2L, GALNT4, TMRF, SGM1, TNKS, PTPRFN1, EXT1, AB0, PPAP2B, GOLGA4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Annotation cluster 3</th>
<th>Enrichment score: 1.541258731532563</th>
<th>Term</th>
<th>P</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0045099: myeloid cell differentiation</td>
<td>0.007081276</td>
<td>DNAE2, CEBPA, CEBPE, CSF1, ACIN1, KLFI, FLVCR1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0050225: macrophage differentiation</td>
<td>0.004154563</td>
<td>CEBPA4, CEBPE, CSF1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0050097: hemopoiesis</td>
<td>0.009664139</td>
<td>DNAE2, CEBPA, KLFI6, CEBPE, IKZF1, FCN1, ACIN1, KLFI1, FLVCR1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0050218: erythrocyte differentiation</td>
<td>0.016394444</td>
<td>DNAE2, ACIN1, KLFI1, FLVCR1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0045634: hemopoietic or lymphoid organ development</td>
<td>0.016484528</td>
<td>DNAE2, CEBPA, KLFI6, CEBPE, IKZF1, FCN1, ACIN1, KLFI1, FLVCR1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0002520: immune system development</td>
<td>0.022674988</td>
<td>DNAE2, CEBPA, KLFI6, CEBPE, IKZF1, FCN1, ACIN1, KLFI1, FLVCR1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0054101: erythrocyte homeostasis</td>
<td>0.02593847</td>
<td>DNAE2, ACIN1, KLFI1, FLVCR1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0048872: homeostasis of number of cells</td>
<td>0.036577024</td>
<td>DNAE2, CSF1, ACIN1, KLFI1, FLVCR1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Annotation cluster 4</th>
<th>Enrichment score: 1.4825623926872973</th>
<th>Term</th>
<th>P</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0046983: protein dimerization activity</td>
<td>0.018422727</td>
<td>CEBPA, CHKA, IKZF1, CEBPE, CSF1, HP54, EEA1, RRAGC, ABCG8, CDA13, ABCG5, JUN, UBAL2, GYS2, EXT1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0042802: identical protein binding</td>
<td>0.04149504</td>
<td>CEBPA, CHKA, CEP72, CEBPE, CSF1, HP54, FBPI, FHL2, EEA1, GUCY2C, CDH13, DOK2, GYS2, EXT1, PHLD3A</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Annotation cluster 5</th>
<th>Enrichment score: 1.3866597881487874</th>
<th>Term</th>
<th>P</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0054657: cellular carbohydrate biosynthetic process</td>
<td>0.000691521</td>
<td>GLT2SD1, FBPI, GYS2, FBPD2, EXT1, PPARGC1A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0016051: carbohydrate biosynthetic process</td>
<td>0.01096541</td>
<td>GLT2SD1, FBPI, GYS2, FBPD2, EXT1, PPARGC1A</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NOTE: Table 2 shows genome annotation enrichment clusters for ALL-associated genes. We performed GO and KEGG pathway annotation enrichment analysis using DAVID (17, 18) for genescores generated from GWA results. We find that five annotation clusters achieved a significant enrichment score (greater than 1.3). Significant annotations (P < 0.05) within these clusters include hematopoiesis, immune system development, erythrocyte differentiation, and regulation of monocyte differentiation. Gene names that are bold represent genes that appear in significant HotNet2 gene subnetworks in this study (Fig. 1; Supplementary Figs. 52 and 53).

hypotheses regarding the etiology and mechanism of ALL onset that can be investigated further in functional studies. Finally, we performed gene and network analysis of European American cases and controls and Hispanic American cases and controls in the multiethnic GWA dataset (14) separately (Fig. 2). We found gene hits (PEGASUS P < 10^{-4}) that were shared between the two cohorts, but also 18 and 3 genes that achieved significance in only the European American and Hispanic American cohorts, respectively (Fig. 2A). A significant subnetwork centered on MEIS1 was also identified in network analysis of the Hispanic American cohort-derived PEGASUS gene scores, but not in network analysis of the European American cohort (Fig. 2B). This result reafirms the need for multiethnic association studies of complex diseases to fully determine how mutations interact to produce complex traits and how treatments can best target complex diseases across ethnicities.
Our study uses novel methodology to quantitatively combine SNP-level GWA analyses of ALL, and we characterize candidate genes and gene subnetworks that may influence ALL risk using multiple large, case-control datasets and a trio dataset with affected children. One limitation of this study is that we rely exclusively on genotype data to replicate our results, as opposed to functional experiments. Still, gene set enrichment analysis and previous studies of hematopoiesis and leukemogenesis confirm that genes identified as ALL-associated by our quantitative framework are biologically relevant to ALL onset and progression (Table 2). Another caveat of our analysis is that we do not have GWA SNP-level values for different molecular subtypes of ALL for the datasets analyzed here, and so we analyze all subtypes of B-cell ALL cases together. When larger datasets of ALL patients become available, molecular subtypes of ALL could be analyzed separately using our approach to test whether genetic heterogeneity underlies risk for different subtypes of ALL. We also lacked sufficient sample size to analyze African American cases and controls separately; when large enough datasets become available, we can carry out further ethnicity-specific analyses using our method. In addition, one challenge that future gene-level association studies should address is producing effect size estimates for gene association scores. Finally, our network analysis using HotNet2 is dependent on publicly available PPI network databases for information about gene interactions, which may be incomplete and may contain inaccuracies.

Our study is the first systematic gene and network analysis of multiple ALL datasets, including exome- and genome-wide case-control studies and a case-parent association study, resulting in novel candidate loci and gene interactions that may lend new insight into the genomic underpinnings of ALL. In particular, we find multiple significant gene subnetworks containing previously identified ALL susceptibility loci that appear to instigate leukemogenesis through multiple different pathways and, thus, may be independent risk loci. PEGASUS, when combined with network analysis, offers a new, powerful approach for identifying shared and unique signals of gene-level associations in complex traits across multiple GWA datasets and can be similarly extended to integrate analysis of multiple data types (e.g., gene expression data, somatic data, and germline mutations).

References


Studies in Multiethnic Populations

Priyanka Nakka, Natalie P. Archer, Heng Xu, et al.


Updated version Access the most recent version of this article at: doi:10.1158/1055-9965.EPI-17-0360

Supplementary Material Access the most recent supplemental material at: http://cebp.aacrjournals.org/content/suppl/2017/07/28/1055-9965.EPI-17-0360.DC1

Cited articles This article cites 70 articles, 14 of which you can access for free at: http://cebp.aacrjournals.org/content/26/10/1531.full#ref-list-1

E-mail alerts Sign up to receive free email-alerts related to this article or journal.
Reprints and Subscriptions To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.
Permissions To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.