Title: Novel gene and network associations found for acute lymphoblastic leukemia using case-control and family-based studies in multi-ethnic populations

Authors: Priyanka Nakka[1,2], Natalie P. Archer[3], Heng Xu[4], Philip J. Lupo[5], Benjamin J. Raphael[6], Jun J. Yang[7], Sohini Ramachandran[1,2,*]

Author Affiliations:
[1] Department of Ecology and Evolutionary Biology, Brown University, Providence, Rhode Island, 02912, USA
[2] Center for Computational Molecular Biology, Brown University, Providence, Rhode Island, 02912, USA
[3] Maternal and Child Health Epidemiology Unit, Texas Department of State Health Services, Austin, Texas, 78756, USA
[4] National Key Laboratory of Biotherapy, Sichuan University, Chengdu, 610041, China
[5] Department of Pediatrics, Baylor College of Medicine, Houston, Texas, 77030, USA
[6] Department of Computer Science, Princeton University, Princeton, New Jersey, 08540, USA
[7] Pharmaceutical Sciences Department, St. Jude Children's Research Hospital, Memphis, TN 38105, USA

*__Corresponding author__: Sohini Ramachandran; Email: sramachandran@brown.edu; Mailing address: 164 Angell St, Box G-W, Brown University, Providence, RI 02912.

Running title: Gene and network analysis of acute lymphoblastic leukemia

Keywords: Leukemias and lymphomas; Pediatric cancers; New software for data analysis; Genotype/phenotype correlations

Grant Support:
S. Ramachandran received grants from the National Science Foundation (NSF) (CAREER Award DBI-1452622), and the US National Institutes of Health (NIH) (R01GM118652 and COBRE award P20GM109035). S. Ramachandran is a Pew Scholar in the Biomedical Sciences, funded by the Pew Charitable Trust, and is an Alfred P. Sloan Research Fellow. B.J. Raphael is supported by a Career Award at the Scientific Interface from the Burroughs Wellcome Fund, an Alfred P. Sloan Research Fellowship, US National Science Foundation (NSF) grant IIS-1016648, an NSF CAREER Award (CCF-1053753), and US National Institutes of Health (NIH) grants R01HG007069 and R01CA180776. P.J. Lupo is supported by Cancer Prevention Research Institute of Texas grant RP140258 and Alex's Lemonade Stand Foundation Epidemiology Grant. J.J. Yang is supported by NIH grant CA176063 and American Lebanese Syrian Associated Charities at St. Jude Children's Research Hospital.

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 analyze two published datasets of case-control GWA summary statistics along with germline data from ALL case-parent trios. We use the gene-level association method PEGASUS to identify genes with multiple variants associated with ALL. We then use PEGASUS gene scores as input to the network analysis algorithm HotNet2 to characterize the genomic architecture of ALL.

Results:

Using PEGASUS, we confirm associations previously observed at genes such as *ARID5B, IKZF1, CDKN2A/2B,* and *PIP4K2A,* and we identify novel candidate gene associations. Using HotNet2, we uncover 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 uncovers loci associated with ALL that are missed by GWA studies such as *MEIS1*. Further, 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 2-5 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 to 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 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 ($p < 5 \times 10^{-8}$). 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 framework (16–23), including an analysis of gene sets associated with ALL by Hsu et al (24), which is discussed in detail later. In the gene-level method, PEGASUS (20), association *p*-values are combined within genes while correcting for linkage disequilibrium (LD); this approach allows us to account for genetic heterogeneity -- when different causal mutations of small effect in the same gene or pathway may be present across cases -- and to test for genes and pathways significantly associated with ALL. Because several ALL susceptibility genes are known to play integral roles in lymphoid development, cell type differentiation and leukemogenesis, it is important to determine if these genes act in concert or separately in ALL patients. Using PEGASUS gene scores as input into gene-set or pathway analysis allows us to test for gene interaction subnetworks that are significantly enriched for genes associated with ALL.

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 1773 affected children (see (15) for details on the original study). Briefly, genome-wide SNP-level *p*-values for 247,505 variants across the exome were obtained on 1773 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 1605 ALL case subjects of multiple ethnicities and 6661 controls (see (14) for further details). Xu et al. (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 classify 963 cases and 1381 controls as European American and 305 cases and 1008 controls as Hispanic American. We also classified 88 cases and 1363 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 has 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, PI Ramachandran) to calculate empirical LD for PEGASUS. Please see Xu et al (15), Xu et al (14) and Archer et al (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 are drawn from a chi-square distribution correlated by empirical LD, and the distribution of the sum of correlated chi-square statistics within a gene is the null distribution for gene-level statistics; this distribution is then numerically integrated to calculate a gene-level *p*-value with machine precision (20). We calculate 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 are downloaded from the UCSC Genome Browser. For PEGASUS testing on GWA results from Xu et al. (15), we use genotype data from the 1000 Genomes EUR population as proxies to calculate LD since the case-control study included only individuals of European descent (30). We use genotype data from the multi-ethnic GWA study (14) to calculate empirical LD for PEGASUS analysis. We use 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 multi-ethnic GWA dataset (the "replication stage" dataset), we also calculate gene scores separately for inferred European American cases and Hispanic Americans 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 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. Though 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 use 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 in Nakka et al. (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 false discovery rate (lFDR) for the GWA-based PEGASUS gene scores (15). We calculate lFDR for PEGASUS gene scores using the twilight R package (31) and determine a cut-off for gene scores at the first "elbow" or inflection point in the graph of 1 – lFDR against gene scores (Supplementary Figure S1).

In addition to gene scores, HotNet2 also requires a protein-protein interaction (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).

Since there is no straightforward way to test for replication of entire significant subnetworks identified using the discovery stage gene scores as input, we instead attempt to replicate the individual genes contained in these subnetworks in the replication gene score dataset to a nominal significance threshold (gene *p*-value < 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 GWA and expression studies (8–13). We then tested for replication of the 42 resulting gene hits ($p < 10^{-3}$, Bonferroni-corrected for the number of haplotype blocks in the genome (38)) using a replication dataset of gene-level *p*-values calculated from a second dataset of case-control GWA *p*-values (14) (Table 1 and Supplemental Table S3). We find that eight genes, *ARID5B, IKZF1*, *FIGNL1*, *CDKN2A, CDKN2B, DDC, PIP4K2A,* and *HLA-DQB1*, are replicated (replication *p*-values ≤ 0.005309704). We also find that only *ARID5B* is replicated in the trio-based gene-level PEGASUS analysis (replication *p*-value: 1.61×10^{-6}), 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 Figure 1A shows multiple genes involved in hematopoiesis and leukemogenesis (39–42). *MEIS1*, *PKNOX1*, *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 results 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* (Figure 1B) have been previously associated with ALL in GWA studies at genome-wide significant levels ($p = 4\text{x}10^{-10}$ and $p = 5.6\text{x}10^{-8}$) (9,43). *CEBPE* encodes CCAAT/enchancer 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 (Interleukin-1β), 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) (Figure 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 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 *CEBPE*, 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 Figures S2 and S3 and are annotated in Supplementary Tables S5-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 (Figure 2). We find that while there are gene hits (PEGASUS *p*-value < 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 (Figure 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 (Figure 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 use significant genes (gene *p*-value < 10^{-3}) 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, KLF6, CEBPE, IKZF1, CSF1, ACIN1, KLF1*, and *FLVCR1* (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 and Supplementary Table S3). We also found that the gene *ARID5B* is replicated in gene-level analysis of a multi-ethnic family-based association study (25,26) (Supplementary Table S4). 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. Further, we note that though 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 and 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 analyses (154,092 monomorphic SNPs of 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 et al (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 (Figures 1 and 2). PEGASUS gene scores are not limited by pre-existing annotations and allow for the calculation of false discovery rates. Hsu et al (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 et al (24) identify do not contain a number of known ALL-associated genes, whereas our network and pathway results (Figures 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 Hsu et al (24)).

After network analysis with HotNet2 (27) using our PEGASUS results as input, we found multiple significant gene interaction network containing genes previously associated with ALL and leukemogenesis, such as *CEBPE* and *MEIS1* (Figure 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 though *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 micro-RNA mediated transcriptional regulation that may be involved in ALL risk (58–63), and we uncover a gene subnetwork containing *UNC93B1* and other genes that plays an important role in innate and adaptive immunity (64–66) (Supplementary Figure S2 and Supplementary Tables S6-S7). 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 hypotheses regarding the etiology and mechanism of ALL onset that can be investigated further in functional studies.

Lastly, 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 (Figure 2). We found gene hits (PEGASUS p-value $< 10^{-6}$) 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 (Figure 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 (Figure 2B). This result re-affirms the need for multi-ethnic 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 p-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. Lastly, 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).

Acknowledgements:

We thank the patients and their parents who participated in the case-control and case-parent studies analyzed in this study.

References:

1. Greaves M. Infection, immune responses and the aetiology of childhood leukaemia. Nat Rev Cancer [Internet]. Nature Publishing Group; 2006 Mar 9 [cited 2016 Nov 16];6(3):193–203. Available from: http://www.nature.com/doifinder/10.1038/nrc1816
2. Greaves MF, Maia AT, Wiemels JL, Ford AM. Leukemia in twins: lessons in natural history. Blood. 2003;102(7).
3. Hjalgrim LL, Rostgaard K, Schmiegelow K, Söderhäll S, Kolmannskog S, Vettenranta K, et al. Age- and sex-specific incidence of childhood leukemia by immunophenotype in the Nordic countries. J Natl Cancer Inst [Internet]. Oxford University Press; 2003 Oct 15 [cited 2016 Nov 17];95(20):1539–44. Available from: http://www.ncbi.nlm.nih.gov/pubmed/14559876
4. Gruhn B, Taub JW, Ge Y, Beck JF, Zell R, Häfer R, et al. Prenatal origin of childhood acute lymphoblastic leukemia, association with birth weight and hyperdiploidy. Leukemia [Internet]. Nature Publishing Group; 2008 Sep 12 [cited 2017 May 31];22(9):1692–7. Available from: http://www.nature.com/doifinder/10.1038/leu.2008.152
5. Marshall GM, Carter DR, Cheung BB, Liu T, Mateos MK, Meyerowitz JG, et al. The prenatal origins of cancer. Nat Rev Cancer [Internet]. NIH Public Access; 2014 [cited 2017 May 31];14(4):277–89. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24599217
6. Hasle H, Clemmensen IH, Mikkelsen M. Risks of leukaemia and solid tumours in individuals with Down's syndrome. Lancet [Internet]. 2000 Jan 15 [cited 2017 May 31];355(9199):165–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10675114
7. Morrell D, Cromartie E, Swift M. Mortality and cancer incidence in 263 patients with ataxia-telangiectasia. J Natl Cancer Inst [Internet]. 1986 Jul [cited 2017 May 31];77(1):89–92. Available from: http://www.ncbi.nlm.nih.gov/pubmed/3459930
8. 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;122(19).
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 [Internet]. Nature Publishing Group; 2009 Sep 16 [cited 2016 Nov 21];41(9):1006–10. Available from: http://www.nature.com/doifinder/10.1038/ng.430
10. Perez-Andreu V, Roberts KG, Harvey RC, Yang W, Cheng C, Pei D, et al. Inherited GATA3 variants are associated with Ph-like childhood acute lymphoblastic leukemia and risk of relapse. Nat Genet [Internet]. Nature Research; 2013 Oct 20 [cited 2016 Nov 21];45(12):1494–8. Available from: http://www.nature.com/doifinder/10.1038/ng.2803
11. 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 [Internet]. Nature Research; 2010 Jun 9 [cited 2016 Nov 22];42(6):492–4.

Available from: http://www.nature.com/doifinder/10.1038/ng.585
12. Treviño LR, Yang W, French D, Hunger SP, Carroll WL, Devidas M, et al. Germline genomic variants associated with childhood acute lymphoblastic leukemia. Nat Genet [Internet]. Nature Publishing Group; 2009 Sep 16 [cited 2016 Nov 17];41(9):1001–5. Available from: http://www.nature.com/doifinder/10.1038/ng.432
13. Walsh KM, De Smith AJ, Hansen HM, Smirnov I V, Gonseth S, Endicott AA, et al. Prevention and Epidemiology A Heritable Missense Polymorphism in CDKN2A Confers Strong Risk of Childhood Acute Lymphoblastic Leukemia and Is Preferentially Selected during Clonal Evolution.
14. 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.
15. Xu H, Zhang H, Yang W, Yadav R, Morrison AC, Qian M, et al. Inherited coding variants at the CDKN2A locus influence susceptibility to acute lymphoblastic leukaemia in children. Nat Commun [Internet]. Nature Publishing Group; 2015 Jan 24 [cited 2015 Nov 4];6:7553. Available from: http://www.nature.com/ncomms/2015/150624/ncomms8553/full/ncomms8553.html
16. Holden M, Deng S, Wojnowski L, Kulle B. GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. Bioinformatics [Internet]. 2008 Dec 1 [cited 2015 Aug 25];24(23):2784–5. Available from: http://bioinformatics.oxfordjournals.org/content/24/23/2784.full
17. Huang DW, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, et al. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol [Internet]. 2007 Jan [cited 2015 Aug 25];8(9):R183. Available from: http://genomebiology.com/2007/8/9/R183
18. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc [Internet]. Nature Publishing Group; 2008 Dec [cited 2016 Dec 5];4(1):44–57. Available from: http://www.nature.com/doifinder/10.1038/nprot.2008.211
19. Jia P, Zheng S, Long J, Zheng W, Zhao Z. dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks. Bioinformatics [Internet]. 2011 Jan 1 [cited 2015 Aug 13];27(1):95–102. Available from: http://bioinformatics.oxfordjournals.org/content/early/2010/11/02/bioinformatics.btq615
20. Nakka P, Raphael BJ, Ramachandran S. Gene and Network Analysis of Common Variants Reveals Novel Associations in Multiple Complex Diseases. Genetics. 2016;
21. Raychaudhuri S, Plenge RM, Rossin EJ, Ng ACY, Purcell SM, Sklar P, et al. Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. PLoS Genet [Internet]. Public Library of Science; 2009 Jun 26 [cited 2015 Apr 5];5(6):e1000534. Available from: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000534
22. Rossin EJ, Lage K, Raychaudhuri S, Xavier RJ, Tatar D, Benita Y, et al. Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. PLoS Genet [Internet]. Public Library of Science; 2011 Jan

13 [cited 2015 Aug 7];7(1):e1001273. Available from: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1001273#pgen.1001273.s013

23. Segrè A V, Groop L, Mootha VK, Daly MJ, Altshuler D. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. PLoS Genet [Internet]. Public Library of Science; 2010 Aug 12 [cited 2015 Aug 19];6(8):e1001058. Available from: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1001058
24. Hsu L-I, Briggs F, Shao X, Metayer C, Wiemels JL, Chokkalingam AP, et al. Pathway Analysis of Genome-wide Association Study in Childhood Leukemia among Hispanics. [cited 2017 May 24]; Available from: http://cebp.aacrjournals.org/content/cebp/early/2016/04/14/1055-9965.EPI-15-0528.full.pdf
25. Archer NP, Perez-Andreu V, Scheurer ME, Rabin KR, Peckham-Gregory EC, Plon SE, et al. Family-based exome-wide assessment of maternal genetic effects on susceptibility to childhood B-cell acute lymphoblastic leukemia in hispanics. Cancer [Internet]. 2016 Dec 1 [cited 2017 Mar 20];122(23):3697–704. Available from: http://doi.wiley.com/10.1002/cncr.30241
26. Archer, Natalie P., Lupo PJ. Family-based exome-wide association study of childhood acute lymphoblastic leukemia among Hispanics confirms role of ARID5B in susceptibility. PLoS One. 2017;In Review.
27. Leiserson MDM, Vandin F, Wu H-T, Dobson JR, Eldridge J V, Thomas JL, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat Genet [Internet]. 2015 Dec 15 [cited 2014 Dec 15];47(2):106–14. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25501392
28. Pritchard JK, Stephens M, Donnelly P. Inference of Population Structure Using Multilocus Genotype Data. Genetics [Internet]. 2000 [cited 2017 Mar 31];155(2). Available from: http://www.genetics.org/content/155/2/945.short
29. Howey R, Cordell HJ, Weinberg C, Umbach D, Shi M, Umbach D, et al. PREMIM and EMIM: tools for estimation of maternal, imprinting and interaction effects using multinomial modelling. BMC Bioinformatics [Internet]. BioMed Central; 2012 [cited 2016 Nov 24];13(1):149. Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-149
30. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. Nature [Internet]. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2015 Sep 30 [cited 2015 Sep 30];526(7571):68–74. Available from: http://dx.doi.org/10.1038/nature15393
31. Scheid S, Spang R. twilight; a Bioconductor package for estimating the local false discovery rate. Bioinformatics [Internet]. 2005 Jun 15 [cited 2015 Jul 22];21(12):2921–2. Available from: http://bioinformatics.oxfordjournals.org/content/21/12/2921.long
32. Das J, Yu H. HINT: High-quality protein interactomes and their applications in understanding human disease. BMC Syst Biol [Internet]. 2012 Jan [cited 2015 Apr 26];6(1):92. Available from: http://www.biomedcentral.com/1752-0509/6/92

33. Rolland T, Taşan M, Charloteaux B, Pevzner SJ, Zhong Q, Sahni N, et al. A proteome-scale map of the human interactome network. Cell [Internet]. Elsevier; 2014 Nov 20 [cited 2016 Nov 25];159(5):1212–26. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25416956
34. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res [Internet]. 2012 Jan [cited 2014 Jul 9];40(Database issue):D109-14. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3245020&tool=pmcentrez&rendertype=abstract
35. Kanehisa M. A database for post-genome analysis. Trends Genet [Internet]. 1997 Sep [cited 2015 May 18];13(9):375–6. Available from: http://www.sciencedirect.com/science/article/pii/S0168952597012237
36. Razick S, Magklaras G, Donaldson IM. iRefIndex: A consolidated protein interaction database with provenance. BMC Bioinformatics [Internet]. 2008 [cited 2016 Nov 25];9(1):405. Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-405
37. Khurana E, Fu Y, Chen J, Gerstein M. Interpretation of Genomic Variants Using a Unified Biological Network Approach. Rzhetsky A, editor. PLoS Comput Biol [Internet]. 2013 Mar 7 [cited 2016 Nov 25];9(3):e1002886. Available from: http://dx.plos.org/10.1371/journal.pcbi.1002886
38. Wojcik GL, Kao WHL, Duggal P. Relative performance of gene- and pathway-level methods as secondary analyses for genome-wide association studies. BMC Genet [Internet]. 2015 Jan [cited 2015 May 17];16(1):34. Available from: http://www.biomedcentral.com/1471-2156/16/34
39. Armstrong SA. Molecular Genetics of Acute Lymphoblastic Leukemia. J Clin Oncol [Internet]. 2005 Sep 10 [cited 2016 Nov 28];23(26):6306–15. Available from: http://www.jco.org/cgi/doi/10.1200/JCO.2005.05.047
40. Pui C-H, Relling M V., Downing JR. Acute Lymphoblastic Leukemia. N Engl J Med [Internet]. Massachusetts Medical Society ; 2004 Apr 8 [cited 2016 Nov 28];350(15):1535–48. Available from: http://www.nejm.org/doi/abs/10.1056/NEJMra023001
41. Mullighan CG, Hermans F, Kaspers G, Hall A, Behm F, Williams D. Molecular genetics of B-precursor acute lymphoblastic leukemia. J Clin Invest [Internet]. American Society for Clinical Investigation; 2012 Oct [cited 2017 Mar 20];122(10):3407–15. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23023711
42. Rozovskaia T, Feinstein E, Mor O, Foa R, Blechman J, Nakamura T, et al. Upregulation of Meis1 and HoxA9 in acute lymphocytic leukemias with the t(4 : 11) abnormality. Oncogene [Internet]. Nature Publishing Group; 2001 Feb 15 [cited 2017 Mar 20];20(7):874–8. Available from: http://www.nature.com/doifinder/10.1038/sj.onc.1204174
43. 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 [Internet]. 2012 May 11 [cited 2016 Nov

28];26(5):902–9. Available from: http://www.nature.com/doifinder/10.1038/leu.2011.302
44. Sun J, Zheng J, Tang L, Healy J, Sinnett D, Dai Y. Association between CEBPE Variant and Childhood Acute Leukemia Risk: Evidence from a Meta-Analysis of 22 Studies. Mills K, editor. PLoS One [Internet]. 2015 May 4 [cited 2017 Apr 24];10(5):e0125657. Available from: http://dx.plos.org/10.1371/journal.pone.0125657
45. Zhang Z-Y, Li S-Z, Zhang H-H, Wu Q-R, Gong J, Liang T, et al. Stabilization of ATF5 by TAK1–Nemo-Like Kinase Critically Regulates the Interleukin-1β-Stimulated C/EBP Signaling Pathway. Mol Cell Biol [Internet]. 2015 Mar 1 [cited 2016 Nov 28];35(5):778–88. Available from: http://mcb.asm.org/lookup/doi/10.1128/MCB.01228-14
46. Rousseau J, Gagné V, Labuda M, Beaubois C, Sinnett D, Laverdière C, et al. ATF5 polymorphisms influence ATF function and response to treatment in children with childhood acute lymphoblastic leukemia. Blood. 2011;118(22).
47. Meyer C, Hofmann J, Burmeister T, Gröger D, Park TS, Emerenciano M, et al. The MLL recombinome of acute leukemias in 2013. Leukemia [Internet]. Nature Publishing Group; 2013 Nov 30 [cited 2016 Nov 28];27(11):2165–76. Available from: http://www.nature.com/doifinder/10.1038/leu.2013.135
48. Nebral K, Denk D, Attarbaschi A, König M, Mann G, Haas OA, et al. Incidence and diversity of PAX5 fusion genes in childhood acute lymphoblastic leukemia. Leukemia [Internet]. Nature Publishing Group; 2009 Jan 20 [cited 2016 Nov 28];23(1):134–43. Available from: http://www.nature.com/doifinder/10.1038/leu.2008.306
49. Doñate C, Ody C, McKee T, Ruault-Jungblut S, Fischer N, Ropraz P, et al. Homing of Human B Cells to Lymphoid Organs and B-Cell Lymphoma Engraftment Are Controlled by Cell Adhesion Molecule JAM-C. Cancer Res. 2013;73(2).
50. Gass JN, Gunn KE, Sriburi R, Brewer JW. Stressed-out B cells? Plasma-cell differentiation and the unfolded protein response. Trends Immunol. 2004;25(1):17–24.
51. Kharabi Masouleh B, Geng H, Hurtz C, Chan LN, Logan AC, Chang MS, et al. Mechanistic rationale for targeting the unfolded protein response in pre-B acute lymphoblastic leukemia. Proc Natl Acad Sci U S A [Internet]. National Academy of Sciences; 2014 May 27 [cited 2016 Nov 27];111(21):E2219-28. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24821775
52. Morishima N, Nakanishi K, Nakano A. Activating transcription factor-6 (ATF6) mediates apoptosis with reduction of myeloid cell leukemia sequence 1 (Mcl-1) protein via induction of WW domain binding protein 1. J Biol Chem [Internet]. American Society for Biochemistry and Molecular Biology; 2011 Oct 7 [cited 2016 Nov 27];286(40):35227–35. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21841196
53. Gilbert J, Haber M, Bordow SB, Marshall GM, Norris MD. Use of Tumor-Specific Gene Expression for the Differential Diagnosis of Neuroblastoma from Other Pediatric Small Round-Cell Malignancies. The American Journal of Pathology. 1999.
54. Wang Y, Vera L, Fischer WH, Montminy M. The CREB coactivator CRTC2 links hepatic ER stress and fasting gluconeogenesis. Nature [Internet]. Nature Publishing Group; 2009 Jun 21 [cited 2016 Nov 27];460(7254):534. Available from: http://www.nature.com/doifinder/10.1038/nature08111
55. Casero RA, Marton LJ. Targeting polyamine metabolism and function in cancer and other

hyperproliferative diseases. Nat Rev Drug Discov [Internet]. Nature Publishing Group; 2007 May [cited 2016 Nov 27];6(5):373–90. Available from: http://www.nature.com/doifinder/10.1038/nrd2243

56. Murray-Stewart TR, Woster PM, Casero RA. Targeting polyamine metabolism for cancer therapy and prevention. Biochem J. 2016;473(19).
57. Thomas T, Thomas TJ. Polyamine metabolism and cancer. J Cell Mol Med [Internet]. Blackwell Publishing Ltd; 2003 Apr [cited 2016 Nov 27];7(2):113–26. Available from: http://doi.wiley.com/10.1111/j.1582-4934.2003.tb00210.x
58. De Keersmaecker K, Atak ZK, Li N, Vicente C, Patchett S, Girardi T, et al. Exome sequencing identifies mutation in CNOT3 and ribosomal genes RPL5 and RPL10 in T-cell acute lymphoblastic leukemia. Nat Genet [Internet]. Nature Research; 2012 Dec 23 [cited 2016 Aug 25];45(2):186–90. Available from: http://www.nature.com/doifinder/10.1038/ng.2508
59. Gutierrez-Camino A, Lopez-Lopez E, Martin-Guerrero I, Piñan MA, Garcia-Miguel P, Sanchez-Toledo J, et al. Noncoding RNA–related polymorphisms in pediatric acute lymphoblastic leukemia susceptibility. Pediatr Res [Internet]. Springer Nature; 2014 Jun 11 [cited 2016 Nov 28];75(6):767–73. Available from: http://www.nature.com/doifinder/10.1038/pr.2014.43
60. Shirai Y-T, Suzuki T, Morita M, Takahashi A, Yamamoto T. Multifunctional roles of the mammalian CCR4-NOT complex in physiological phenomena. Front Genet [Internet]. Frontiers Media SA; 2014 [cited 2016 Nov 28];5:286. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25191340
61. Verghese ET, Drury R, Green CA, Holliday DL, Lu X, Nash C, et al. MiR-26b is down-regulated in carcinoma-associated fibroblasts from ER-positive breast cancers leading to enhanced cell migration and invasion. J Pathol [Internet]. John Wiley & Sons, Ltd; 2013 Nov [cited 2016 Nov 28];231(3):388–99. Available from: http://doi.wiley.com/10.1002/path.4248
62. McCubrey JA, Steelman LS, Bertrand FE, Davis NM, Abrams SL, Montalto G, et al. Multifaceted roles of GSK-3 and Wnt/β-catenin in hematopoiesis and leukemogenesis: opportunities for therapeutic intervention. Leukemia [Internet]. Nature Publishing Group; 2014 Jan 19 [cited 2016 Nov 28];28(1):15–33. Available from: http://www.nature.com/doifinder/10.1038/leu.2013.184
63. Youns M, Fu Y-J, Zu Y-G, Kramer A, Konkimalla VB, Radlwimmer B, et al. Sensitivity and resistance towards isoliquiritigenin, doxorubicin and methotrexate in T cell acute lymphoblastic leukaemia cell lines by pharmacogenomics. Naunyn Schmiedebergs Arch Pharmacol [Internet]. Springer-Verlag; 2010 Sep 29 [cited 2016 Nov 28];382(3):221–34. Available from: http://link.springer.com/10.1007/s00210-010-0541-6
64. Monlish DA, Bhatt ST, Schuettpelz LG. The Role of Toll-Like Receptors in Hematopoietic Malignancies. Front Immunol [Internet]. Frontiers Media SA; 2016 [cited 2016 Nov 29];7:390. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27733853
65. Takeda J, Miyata T, Kawagoe K, Iida Y, Endo Y, Fujita T, et al. Deficiency of the GPI anchor caused by a somatic mutation of the PIG-A gene in paroxysmal nocturnal hemoglobinuria. Cell. Cell Press; 1993;73(4):703–11.

66. Lindqvist CM, Nordlund J, Ekman D, Johansson A, Moghadam BT, Raine A, et al. The Mutational Landscape in Pediatric Acute Lymphoblastic Leukemia Deciphered by Whole Genome Sequencing. Hum Mutat [Internet]. 2015 Jan [cited 2016 Nov 29];36(1):118–28. Available from: http://doi.wiley.com/10.1002/humu.22719
67. San Jose-Eneriz E, Agirre X, Roman-Gomez J, Cordeu L, Garate L, Jimenez-Velasco A, et al. Downregulation of DBC1 expression in acute lymphoblastic leukaemia is mediated by aberrant methylation of its promoter. Br J Haematol [Internet]. Blackwell Publishing Ltd; 2006 Jul [cited 2016 Dec 3];134(2):137–44. Available from: http://doi.wiley.com/10.1111/j.1365-2141.2006.06131.x
68. Rachakonda SP, Penack O, Dietrich S, Blau O, Blau IW, Radujkovic A, et al. Single-Nucleotide Polymorphisms Within the Thrombomodulin Gene (THBD) Predict Mortality in Patients With Graft-Versus-Host Disease. J Clin Oncol [Internet]. 2014 Oct 20 [cited 2016 Dec 5];32(30):3421–7. Available from: http://jco.ascopubs.org/cgi/doi/10.1200/JCO.2013.54.4056
69. Borgoño CA, Diamandis EP. The emerging roles of human tissue kallikreins in cancer. Nat Rev Cancer [Internet]. Nature Publishing Group; 2004 Nov [cited 2016 Dec 5];4(11):876–90. Available from: http://www.nature.com/doifinder/10.1038/nrc1474
70. Chen X, Zheng J, Zou Y, Song C, Hu X, Zhang C, et al. IGF binding protein 2 is a cell-autonomous factor supporting survival and migration of acute leukemia cells. J Hematol Oncol [Internet]. BioMed Central; 2013 [cited 2016 Dec 5];6(1):72. Available from: http://jhoonline.biomedcentral.com/articles/10.1186/1756-8722-6-72
71. 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.

Tables:

Table 1. Case-control GWA-based PEGASUS Gene Hits.

Gene ID	*Chromosome*	*Start position (hg19)*	*End position (hg19)*	*Discovery stage PEGASUS p-values: GWA p-values* (15)	*Replication stage PEGASUS p-values: GWA p-values* (14)
ARID5B*	10	63661012	63856707	2.22E-16	2.22E-16*
IKZF1	7	50343678	50472798	2.22E-16	2.22E-16
FIGNL1	7	50511826	50518088	2.22E-16	2.22E-16
CDKN2A	9	21967750	21994490	1.97E-07	0.000928457
DDC	7	50526133	50633154	1.14E-05	4.82E-12
PIP4K2A	10	22823765	23003503	2.36E-05	4.45E-07
CDKN2B	9	22002901	22009312	0.000140848	1.66E-05
HLA-DQB1	6	32627240	32634466	0.000870177	0.005309704

Table 1 shows case-control GWA-based PEGASUS gene hits. In the discovery-stage analysis, we apply PEGASUS to case-control GWA *p*-values (15) using the 1000 Genomes Project EUR population (30) as a reference for LD. We then replicated eight of the 42 resulting gene hits ($p < 10^{-3}$; Bonferroni-corrected for the number of haplotype blocks in the genome (38)), shown in bold above, by applying PEGASUS to a second dataset of case-control GWA *p*-values

(replication *p*-value < 0.05) (14). The full list of 42 gene hits is shown in Supplementary Table S3. The gene marked with an asterisk (*), *ARID5B*, was also replicated using the trio analysis-based PEGASUS replication test (replication *p*-value = 1.61×10^{-6}) (Supplementary Table S4). *ARID5B, IKZF1*, *CDKN2A/2B* and *PIP4K2A* all contain SNPs previously associated with ALL in GWA studies (8–13).

Author Manuscript Published OnlineFirst on July 27, 2017; DOI: 10.1158/1055-9965.EPI-17-0360
Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

Table 2. Genome Annotation Enrichment Clusters.

Annotation Cluster 1	**Enrichment Score: 2.098930503473424**		
	Term	***p*-value**	**Genes**
	GO:0045657: positive regulation of monocyte differentiation	4.62E-04	*JUN, CSF1, ACIN1*
	GO:0045655: regulation of monocyte differentiation	9.16E-04	*JUN, CSF1, ACIN1*
	GO:0002763: positive regulation of myeloid leukocyte differentiation	0.001361426	***IKZF1***, *JUN, CSF1, ACIN1*
	GO:0045639: positive regulation of myeloid cell differentiation	0.007937827	***IKZF1***, *JUN, CSF1, ACIN1*
	GO:0045637: regulation of myeloid cell differentiation	0.011372639	*HIST4H4*, ***IKZF1***, *JUN, CSF1, ACIN1*
	GO:0002761: regulation of myeloid leukocyte differentiation	0.014424008	***IKZF1***, *JUN, CSF1, ACIN1*
Annotation Cluster 2	**Enrichment Score: 1.600658678135878**		
	Term	***p*-value**	**Genes**
	topological domain:Lumenal	0.002991392	*TCIRG1, GCNT4,* ***TPST2****, ST6GAL2,* ***GOLT1B****, IGF2R, LMAN2L, CSF1, ASPHD2, GALNT4,* ***EXT1****,* ***ABO****,* ***PPAP2B****, MOXD1*
	golgi apparatus	0.011871437	***TPST2****, GCNT4, ST6GAL2, AP1G2,* ***GOLT1B****, LMAN2L, GALNT4, TMF1, SGSM1,* ***TNKS****, PTGFRN,* ***EXT1****,* ***ABO****,* ***PPAP2B****,* ***GOLGA4***
Annotation Cluster 3	**Enrichment Score: 1.5412837811532563**		
	Term	***p*-value**	**Genes**
	GO:0030099: myeloid cell differentiation	0.001081276	*DNASE2, CEBPA,* ***CEBPE****, CSF1, ACIN1, KLF1, FLVCR1*
	GO:0030225: macrophage differentiation	0.004134616	*CEBPA,* ***CEBPE****, CSF1*
	GO:0030097: hemopoiesis	0.009664139	*DNASE2, CEBPA,* ***KLF6****,* ***CEBPE****,* ***IKZF1****, CSF1, ACIN1, KLF1, FLVCR1*
	GO:0030218: erythrocyte differentiation	0.016399444	*DNASE2, ACIN1, KLF1, FLVCR1*
	GO:0048534: hemopoietic or lymphoid organ development	0.016484328	*DNASE2, CEBPA, KLF6,* ***CEBPE****,* ***IKZF1****, CSF1, ACIN1, KLF1, FLVCR1*
	GO:0002520: immune system development	0.022674988	*DNASE2, CEBPA, KLF6,* ***CEBPE****,* ***IKZF1****, CSF1, ACIN1, KLF1, FLVCR1*
	GO:0034101: erythrocyte homeostasis	0.023193847	*DNASE2, ACIN1, KLF1, FLVCR1*
	GO:0048872: homeostasis of number of cells	0.036577024	*DNASE2, CSF1, ACIN1, KLF1, FLVCR1*
Annotation Cluster 4	**Enrichment Score: 1.4825623926872973**		
	Term	***p*-value**	**Genes**
	GO:0046983: protein dimerization activity	0.011842272	*CEBPA,* ***CHKA****,* ***IKZF1****,* ***CEBPE****, CSF1,* ***HPS4****,* ***EEA1****,* ***RRAGC****,* ***ABCG8****, CDH13,* ***ABCG5****, JUN, UBA3,* ***GYS2****,* ***EXT1***
	GO:0042802: dentical protein binding	0.041495104	*CEBPA,* ***CHKA****,* ***CEP72****,* ***CEBPE****, CSF1,* ***HPS4****,* ***FBP1****, FHL2,* ***EEA1****, GUCY2D, CDH13, DOK2,* ***GYS2****,* ***EXT1****, PHLDA3*
Annotation Cluster 5	**Enrichment Score: 1.3866597881487874**		
	Term	***p*-value**	**Genes**
	GO:0034637: cellular carbohydrate biosynthetic process	0.001691251	*GLT25D1,* ***FBP1****,* ***GYS2****,* ***FBP2****,* ***EXT1****, PPARGC1A*
	GO:0016051: carbohydrate biosynthetic process	0.01096541	*GLT25D1,* ***FBP1****,* ***GYS2****,* ***FBP2****,* ***EXT1****, PPARGC1A*
	GO:0033692: cellular polysaccharide biosynthetic process	0.041446472	*GLT25D1,* ***GYS2****,* ***EXT1***

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 gene scores 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 (Figure 1 and Supplementary Figures S2 and S3).

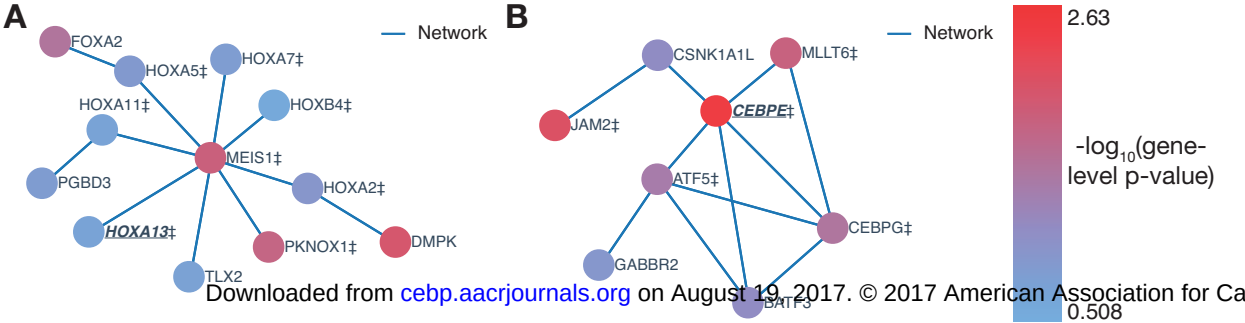
Figure Legends

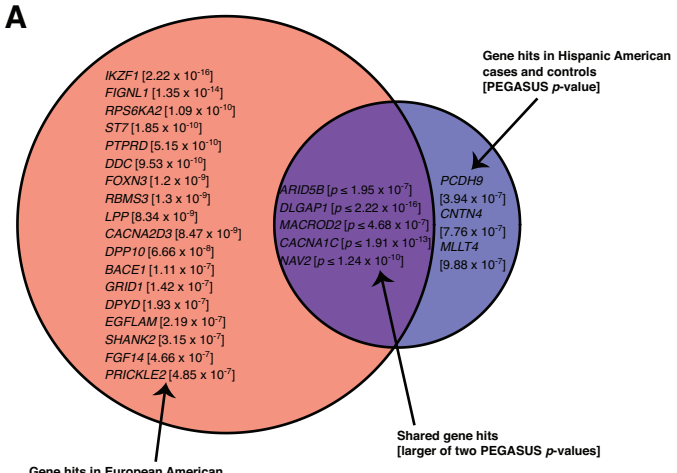
Figure 1. HotNet2 Results using PEGASUS Gene Scores as Input.

Figure 1 shows two subnetworks for ALL containing known ALL-associated loci from significant runs of HotNet2 (27) ($p \leq 0.05$ for multiple subnetwork sizes), using PEGASUS gene scores based on GWA SNP *p*-values as input. Panel A shows multiple *HOX* genes involved in hematopoiesis and leukemogenesis, and panel B shows genes such as *CEBPE*, which contains genome-wide significant SNPs, interacting with other genes involved in hematopoiesis. Circles represent genes in each subnetwork and are colored by heat score (negative log-transformed PEGASUS gene scores); the color bar indicates the lowest heat score (blue or "cold" genes) and the highest heat score (red or "hot" genes) in each subnetwork. Lines between genes indicate a direct gene-gene interaction from the iRefIndex (36) and KEGG (34,35) databases. Genes that are bolded, italicized and underlined represent genes with nominally significant PEGASUS scores (*HOXA13*: replication *p*-value = 0.04; *CEBPE*: replication *p*-value = 9.20×10^{-10}) in the replication GWA dataset (14). Genes marked with the double dagger symbol (‡) are genes that been associated with ALL or a related phenotype in previous GWA studies not analyzed here or functional studies (9,39–49).

Figure 2. Ethnicity-Specific PEGASUS and HotNet2 Results.

Figure 2 shows ethnicity-specific PEGASUS and HotNet2 results. Using PEGASUS, we calculate 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 Panel A shows 18 significant gene hits (PEGASUS *p*-values $< 10^{-6}$) 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). Panel B shows 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*-value < 0.05).





Cancer Epidemiology, Biomarkers & Prevention

AACR American Association for Cancer Research

Novel gene and network associations found for lymphoblastic leukemia using case-control and family-based studies in multi-ethnic populations

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

Cancer Epidemiol Biomarkers Prev Published OnlineFirst July 27, 2017.