Leveraging genome and phenome-wide association studies to investigate genetic risk of acute lymphoblastic leukemia

Eleanor C. Semmes[1,2], Jayaram Vijayakrishnan[3], Chenan Zhang PhD[4], Jillian H. Hurst PhD[2], Richard S. Houlston[3], Kyle M. Walsh PhD[2,4,5,6*]

[1] Medical Scientist Training Program, Duke University, Durham, NC, United States of America

[2] Children's Health and Discovery Initiative, Department of Pediatrics, Duke University, Durham, NC, United States of America

[3] Division of Genetics and Epidemiology, The Institute of Cancer Research, Sutton, Surrey SM2 5NG, United Kingdom

[4] Department of Epidemiology and Biostatistics, University of California, San Francisco, CA, United States of America

[5] Department of Neurosurgery, Duke University, Durham, NC, United States of America

[6] Duke Cancer Institute, Duke University, Durham, NC, United States of America

Running title: Leveraging GWAS and PheWAS in childhood leukemia

Keywords: acute lymphoblastic leukemia, GWAS, PheWAS, childhood cancer, cancer epidemiology

*corresponding author
Email: kyle.walsh@duke.edu (KMW)
Mailing address: DUMC Box 3050, Durham, NC 27710
Phone: (919) 684-8732
Fax: (919) 684-5207

Conflict of interest disclosure: The authors declare no competing interests.
Word count: 250 (abstract), 4,000 (manuscript excluding abstract and references)
Number of elements: 4 tables, 2 figures

Financial support: This study was supported by R21CA242439-01 (KMW), Alex's Lemonade Stand Foundation "A" Awards (KMW), The Children's Health and Discovery Initiative of Translating Duke Health (ECS, JHH, KMW), and NIH T32CA151022-06 (CZ).

ABSTRACT

Background: Genome-wide association studies (GWAS) of childhood cancers remain limited, highlighting the need for novel analytic strategies. We describe a hybrid GWAS and phenome-wide association study (PheWAS) approach to uncover genotype-phenotype relationships and candidate risk loci, applying it to acute lymphoblastic leukemia (ALL).

Methods: PheWAS was performed for 12 ALL SNPs identified by prior GWAS and two control SNP-sets using UK Biobank data. PheWAS-traits significantly associated with ALL SNPs compared to control SNPs were assessed for association with ALL risk (959 cases, 2624 controls) using polygenic score and Mendelian randomization analyses. Trait-associated SNPs were tested for association with ALL risk in single-SNP analyses, with replication in an independent case-control dataset (1618 cases, 9409 controls).

Results: Platelet count was the trait most enriched for association with known ALL risk loci. A polygenic score for platelet count (223 SNPs) was not associated with ALL risk (P=0.82) and Mendelian randomization did not suggest a causal relationship. However, twelve platelet count-associated SNPs were nominally associated with ALL risk in COG data and 3 were replicated in UK data (rs10058074, rs210142, rs2836441).

Conclusions: In our hybrid GWAS-PheWAS approach, we identify pleiotropic genetic variation contributing to ALL risk and platelet count. Three SNPs known to influence platelet count were reproducibly associated with ALL risk, implicating genomic regions containing *IRF1*, pro-apoptotic protein *BAK1,* and *ERG* in platelet production and leukemogenesis.

Impact: Incorporating PheWAS data into association studies can leverage genetic pleiotropy to identify cancer risk loci, highlighting the utility of our novel approach.

Introduction

Genome-wide association studies (GWAS) have greatly enhanced our understanding of inherited genetic susceptibility to cancer (1), but GWAS of pediatric cancers remain limited due to lower disease incidence (2). Because of limited sample size, GWAS of childhood malignancies are often underpowered to detect variants of small-to-moderate effect size, preventing potentially important risk loci from reaching genome-wide statistical significance (*i.e.* $P<5.0x10^{-8}$) (2, 3). Novel analytic approaches are needed to investigate how germline genetic variation contributes to childhood cancer risk. The incorporation of polygenic scores (4, 5), Mendelian randomization (MR) analyses, gene-pathway analyses (6), and phenome-wide association studies (PheWAS) can augment traditional GWAS approaches to expand our understanding of the genetic etiology of pediatric malignancies and other rare diseases.

PheWAS have not been widely applied to childhood cancer etiology research, but represent a promising approach to understanding genetic risk in childhood cancer (7, 8). While GWAS examine millions of genetic loci and test for association with a single phenotype or disease, PheWAS test hundreds or thousands of phenotypes for association with a single genetic variant, essentially a reversal of the GWAS paradigm (9, 10). This methodology has recently become feasible through large collaborative efforts linking electronic health records (EHR) data with high-throughput genomic data (7). Using PheWAS to discover additional traits associated with cancer risk variants can reveal “intermediate phenotypes” (*e.g.,* height, smoking behaviors) (4) that may mediate the relationship between SNPs and cancer development. Trait-disease relationships can be further investigated using polygenic scores and MR approaches. PheWAS data can also be integrated into case-control studies to identify trait-associated genetic variants, create empirical candidate-SNP lists, and test for association with cancer case-control status. Thus, integrating PheWAS and GWAS approaches in analyses of case-control datasets may enhance our understanding of pathways driving pediatric cancer predisposition.

Acute lymphoblastic leukemia (ALL) is the most common childhood malignancy, accounting for nearly one-third of pediatric cancers (11). Its etiology is complex, but the disease is likely initiated *in utero,* with driver pre-leukemic fusion genes arising in lymphoid progenitors. ALL development is also influenced by pre/postnatal environmental exposures (*e.g.,* infections, ionizing radiation) (11-14) and by germline genetic variants. GWAS have uncovered important inherited genetic risk loci for ALL in hematopoietic transcription factors (*IKZF1, CEBPE, ARID5B, GATA3, ELK3*), cell cycle regulators (*CDKN2A/CDKN2B, SP4*), and chromatin remodeling enzymes (*BMI1*), though the precise mechanisms by which these GWAS-identified risk loci influence leukemogenesis are not completely understood (15-22).

We have developed an integrated GWAS-PheWAS approach to identify candidate traits and trait-associated variants that may modify cancer risk. We apply this methodology to ALL, uncovering novel phenotypes associated with known ALL risk variants and pleiotropic ALL risk loci, which we successfully replicate in an independent dataset. Our findings suggest that this hybrid GWAS-PheWAS methodology is a promising new approach for deciphering germline genetic risk in rare diseases, such as childhood cancers, where GWAS power remains limited.

Materials and Methods

Prior GWAS ALL risk loci. We accessed the NHGRI-EBI GWAS Catalog (https://www.ebi.ac.uk/gwas/) to compile a list of variants previously identified by GWAS as associated with B-cell precursor ALL risk in European-ancestry populations at genome-wide statistical significance (*i.e.* $P<5.0\times10^{-8}$) (access date: November 27, 2018) (23). We pruned this list of significant variants for linkage disequilibrium ($R^2\leq0.15$ in European-ancestry populations) using LDlink (24) and cross-referenced recent reviews on ALL GWAS (3), identifying 12 genome-wide significant independent ALL risk SNPs, which were included in our ALL SNP-set.

Control SNP Sets. We compiled 2 comparison SNP-sets to serve as controls for PheWAS analyses. A set of unlinked control SNPs (1000 Genomes Project) was generated using SNPsnap (Broad Institute) (25). Four control SNPs were matched to the 12 ALL risk SNPs on: minor allele frequency (±5%), surrounding gene density (±50%), distance to nearest gene (±50%) and, as a proxy for haplotype block size, the number of other SNPs in LD at $R^2 \geq 0.50$ (±50%). For several ALL risk SNPs, we could not generate more than 4 control SNPs without loosening our matching parameters, but the gain in statistical power achieved beyond a case-to-control ratio of 1:4 is minimal (26, 27).

Because ALL risk SNPs are trait-associated variants that may be more likely to associate with additional traits in PheWAS analyses, we identified a second control SNP-set by querying the GWAS catalog for chronic lymphocytic leukemia (CLL) risk SNPs. We used the same methodology as for ALL risk SNPs, yielding 31 unlinked CLL-associated variants used as another control SNP-set.

eQTL and *in silico* SNP functional analyses. We characterized ALL risk SNPs and control SNP-sets using HaploReg to annotate chromatin state and regulatory motifs surrounding each SNP (28). We examined whether variants were expression quantitative trait loci (eQTLs), protein-binding, located in DNAse hypersensitive sites, promoter or enhancer histone marks, or predicted to change transcription factor binding motifs.

UK Biobank GeneATLAS and PheWAS analyses. The UK Biobank atlas of genetic associations (http://geneatlas.roslin.ed.au.uk/) was constructed by genotyping 452,264 European-ancestry individuals for 805,426 genetic variants, performing genome-wide SNP imputation and quality-controls, and linking genetic data to EHR data (29). GeneATLAS contains data for 778 traits (118 quantitative, 660 binary) and associations with 9,113,133 genetic variants (genotyped or imputed). GeneATLAS is searchable and can be queried for

genetic (*e.g.* SNPs) or phenotypic (*e.g.*, height) data to assess genotype-phenotype associations (see Canela-Xandri *et. al.* for additional details) (29).

We queried GeneATLAS for trait associations with 12 known ALL risk SNPs and two control SNP-sets (31 CLL-associated SNPs, 48 matched SNPsnap controls). Summary statistics for traits associated with each queried variant were downloaded from GeneATLAS for downstream analyses. Significant SNP-trait associations (P<0.01) were carried forward in subsequent SNP-set analyses. Although a more stringent p-value threshold for carrying SNP-trait associations forward was considered (*e.g.* 0.05/778), this was determined to be too conservative because many of the 778 traits in GeneATLAS have high genetic correlations with each other (*e.g.* weight and hip circumference, 0.909; reticulocyte percentage and reticulocyte count, 0.952). Additionally, these individual SNP-trait associations were carried forward for SNP-set enrichment comparisons between ALL-associated SNPs and control SNP-sets, and as such the PheWAS significance threshold is somewhat arbitrary so long as it is the same threshold across all SNP-sets. PheWAS results for the 12 ALL risk SNPs and 778 traits were compared to results for the two control SNP-sets using the R Statistical Programming Environment (http://www.R-project.org/, version 3.5.2). Using Fisher's exact tests, we compared PheWAS traits associated with >1 ALL SNP between the ALL and control SNP-sets to determine if traits were enriched for association with known ALL risk variants.

ALL case-control discovery cohort. We included 959 European-ancestry ALL cases from the Children's Oncology Group (COG) in our discovery dataset (16). Genotype data were downloaded from dbGaP study accession phs000638.v1.p1, including ALL patients from COG protocols 9904 and 9905 for whom DNA was obtained from remission blood samples (30). Controls included 2624 European-ancestry subjects from the Wellcome Trust Case-Control Consortium (http://www.wtccc.org.uk/) (31). Cases and controls were genotyped on the Affymetrix 6.0 array. As described previously, genotyping quality-control (QC) measures were

implemented for cases and controls (16). We excluded samples or SNPs with genotyping call rates <98%, individuals with suggested non-European-ancestry, IBD proportion >0.20, or with discrepant sex between genotype and clinical report.

Genotype imputation. ALL case-control SNP data underwent genome-wide imputation as previously described (5). Haplotype phasing was performed with SHAPEIT (version 2.790) (32), and whole-genome imputation was performed using Minimac3 software (33) with 64,976 human haplotypes from the Haplotype Reference Consortium (2016 release) as the reference panel (34). SNPs with imputation quality (info) scores <0.60 or posterior probabilities <0.90 were excluded (16).

Platelet count polygenic score and single-SNP associations. We constructed a polygenic score for platelet count using 287 independent genetic variants associated with platelet count in a prior GWAS of blood cell trait indices (223 were included after QC filtering) (35) (**S7 Table**). The polygenic score for each individual in the ALL case-control dataset was determined based on signed, weighted beta estimates for each platelet count-associated variant, as reported in Astle *et. al.* (35) and calculated using the PLINK toolkit (36). We performed logistic regression for the platelet count polygenic score, adjusting for 10 principal components (PCs). We also tested platelet count-associated SNPs for association with ALL case-control status via single-SNP association analyses.

Mendelian randomization analyses. To assess for a causal relationship between platelet count and ALL risk, we performed formal MR analyses with the R package "MendelianRandomization"(37, 38). Using summary statistics of SNP-exposure (*i.e.,* platelet count) and SNP-outcome (*i.e.,* ALL) associations, we used the (1) inverse-variance weighted

(IVW), (2) MR-Egger, and (3) weighted median methods to test for a causal relationship between platelet count and ALL risk in our case-control dataset (39, 40).

ALL replication study. The ALL replication dataset was a meta-analysis of two prior published GWAS of B-cell precursor ALL, including German GWAS (834 cases, 2024 controls) (19) and UK GWAS II (784 cases, 7,385 controls) (20). German cases were genotyped using Illumina Human OmniExpress-12v1,0 arrays and controls were genotyped using the same platform or Illumina-HumanOmni1-Quad1_v1. UK GWAS II cases and controls were genotyped using an Illumina Infinium OncoArray-500K. Fixed-effects meta-analysis was used to estimate beta values, standard errors, and p-values for queried risk loci in this combined GWAS meta-analysis (1618 ALL cases, 9409 controls). For additional information on the GWAS meta-analysis used for replication, see Vijayakrishnan *et. al.* (21).

Results

Overview of methods. An overview of the methodology applied in our study is displayed in **Figure 1**. We used the GWAS catalog and a thorough literature review to identify known ALL risk variants from GWAS of European-ancestry populations. PheWAS analyses were then performed with the UK Biobank GeneATLAS database to test each ALL-associated variant and control variant for association with 778 traits in the UK Biobank. After determining which traits were enriched for association with the ALL SNP-set compared to control SNP-sets, we returned to the GWAS catalog to identify SNPs associated with these traits. Using polygenic score, MR, and candidate SNP approaches, we examined whether PheWAS-identified traits or trait-associated variants conferred ALL risk, and replicated single-SNP associations in an independent ALL case-control dataset.

Risk variants for PheWAS analysis. Using the GWAS catalog, we identified 12 independent (R^2≤0.15), genome-wide significant (P<5.0x10^{-8}) ALL risk SNPs (**Table 1**), which all previously replicated in independent cohorts. Two SNP-sets served as controls for our PheWAS analyses, including 31 SNPs previously associated with chronic lymphocytic leukemia (CLL) and 48 control SNPs matched to ALL risk SNPs on minor allele frequency, gene density, distance to nearest gene, and number of SNPs in LD (25). Functional annotation and *in silico* analysis of the ALL SNP-set and control SNP-sets demonstrated similar characteristics in terms of impact on chromatin structure, including promoter and enhancer histone marks, DNAse hypersensitivity, and impact on regulatory motifs; however, ALL-associated and CLL-associated SNPs were likelier to be eQTLs (**S1 Table**).

UK Biobank PheWAS analyses. We utilized the UK Biobank GeneATLAS database to conduct a PheWAS for 12 ALL-associated SNPs (**S2 Table**), 31 CLL-associated SNPs (**S3 Table**), and 48 matched control SNPs (**S4 Table**) to test for association with 778 traits. We used the same PheWAS approach and nominal significance threshold to identify SNP-trait associations (*i.e.* P<0.01) for ALL and control SNPs. The proportion of variants in each SNP-set (12 ALL-associated, 31 CLL-associated, 48 matched control) that was associated with a particular PheWAS trait was compared across groups to ascertain phenotypes enriched for association with ALL SNPs compared with control SNPs (**S5 Table)**.

We determined that 76 of the 778 traits in the database were nominally associated (P<0.01) with >1 of the 12 ALL risk SNPs. PheWAS traits significantly associated with >1 ALL risk SNP were carried forward for enrichment comparisons between ALL and control SNP-sets (**S6 Table**). All 76 PheWAS traits compared between SNP-sets are depicted in **Figure 2** showing the relative proportion of significant SNP-trait associations in each SNP-set. Platelet count was the phenotype most enriched for association with ALL risk variants. Specifically, 9 out of 12 (75%) ALL SNPs were nominally associated with platelet count, compared to 11 of 31

(35.5%) CLL SNPs (P = 0.047) and 6 of 48 (12.5%) control SNPs (P < 0.001) (**Table 2**). Notably, many of the PheWAS-identified traits were enriched for association in the ALL SNPs compared to the control SNPs, but only 5 traits were significantly enriched for association with ALL SNPs compared to both control SNPs *and* CLL SNPs, and platelet count was associated with the highest proportion of ALL SNPs (**Table 2**).

Platelet count polygenic score analyses. Given that platelet count was the trait most enriched for association with ALL SNPs in PheWAS analyses, we constructed a polygenic score for platelet count using 287 previously-published variants from a recent GWAS on blood cell indices (35) (**S7 Table**). Of these, 223 SNPs were successfully imputed (info score ≥0.60, posterior probability ≥0.90) in our ALL case-control dataset (959 cases, 2624 controls) and used in polygenic score construction. The polygenic score for platelet count was not associated with ALL case-control status in a logistic regression model adjusting for sex and 10 PCs (P=0.819).

Mendelian randomization analyses. To test for a causal relationship between platelet count and ALL risk, we used several MR analytical approaches wherein genetic variants are used as instrumental variables to assess causality in exposure/risk factor associations. Estimates from IVW (P_{IVW}=0.948), MR-Egger ($P_{MR\text{-}Egger}$=0.857, $P_{MR\text{-}intercept}$=0.912), and median-weighted ($P_{MR\text{-}median}$=0.857) MR methods were non-significant and consistent with the null polygenic score results. These MR results suggest that platelet count does not mediate ALL risk and that there is no causal relationship between these two traits.

Platelet count-associated SNPs as candidate ALL risk loci. To examine whether individual platelet count-associated variants might have pleiotropic effects on ALL risk, we performed single-SNP association analyses for 223 platelet count-associated SNPs in 959 ALL cases and 2624 controls (**S8 Table**). Twelve SNPs were nominally associated (P<0.05) with ALL case-

control status (notably, not more than expected by chance) after adjusting for sex and 10 PCs (**Table 3**). The directional effect of platelet count-associated alleles (*i.e.,* increased versus decreased platelet count) did not correlate with the direction of effect on ALL susceptibility (*i.e.,* protection versus risk).

These 12 candidate SNPs were carried forward for evaluation in an independent UK ALL case-control dataset (1618 cases, 9409 controls). Nine SNPs were successfully genotyped or imputed in this dataset, of which three associations were successfully replicated at $P<5.6x10^{-3}$ (*i.e.* 0.05/9) (**Table 4**). SNPs had similar magnitudes of effect in the UK ALL case-control and discovery data. The replicated variants map to 3 distinct genomic loci on 5q31.1, 6p21.31, and 21q22.2 (**Table 4**). To interrogate these risk loci further, we identified the genes in which these variants resided and associated genes for which these variants were expression quantitative trait loci (eQTLs) (28). We found that the 5q31.1 region was adjacent to *IRF1*, a gene encoding interferon regulatory factor 1, which regulates host immune responses, including interferon signaling. The 6p21.31 region includes *BAK1*, a pro-apoptotic protein known to be disrupted in adult-onset malignancies. Finally, the 21q22.2 region encodes the hematopoietic transcription factor *ERG*, known to be associated with ALL risk in Hispanics and children with Trisomy 21 (*i.e.* Down Syndrome) (41, 42).

Discussion

We provide a novel framework (**Figure 1**) for leveraging existing GWAS and PheWAS data to uncover traits associated with known disease risk variants and to identify trait-associated variants as possible candidate risk loci. We apply this framework to an investigation of ALL predisposition that combines rich genotype-phenotype data available from the UK Biobank with ALL case-control analyses. We first identify SNPs associated with ALL using the GWAS catalog. We then perform PheWAS on these SNPs and control SNP-sets using the UK Biobank GeneATLAS, identifying platelet count as the trait most enriched for association with ALL risk

loci. Returning to the GWAS catalog, we identify genetic determinants of platelet count. We then use a two-stage case-control design (43, 44) to examine whether SNPs associated with platelet count modify ALL risk, confirming three risk loci near *IRF1*, *BAK1,* and *ERG*.

Potential modifications to this hybrid GWAS-PheWAS approach could be implemented in future applications based on features of the cancer undergoing analysis and the datasets available. For cancers with many known GWAS hits (*e.g.* breast cancer), it may be preferable to use a more stringent p-value threshold for the PheWAS analysis to streamline subsequent SNP-set enrichment comparisons. Similarly, trait-associated SNPs could be evaluated for their association with cancer using a more stringent p-value threshold in a one-stage case-control design when sample sizes are large or when replication sets are unavailable.

We identified platelet count as significantly enriched for association with ALL risk SNPs; however, our results did not suggest a direct role for platelet count in mediating ALL risk, as the polygenic score for platelet count was not associated with ALL case-control status. Null results from MR analyses also support the conclusion that there is no causal relationship between platelet count and ALL. This indicates that platelet count and ALL may have overlapping genetic architecture due to pleiotropic loci independently influencing both traits, which appears reasonable since regulatory variants in hematopoietic transcription factors could influence each phenotype. This interpretation is supported by our single-SNP association results, identifying and replicating 3 ALL risk loci using platelet count-associated variants as candidate SNPs. Two of these ALL risk alleles were associated with higher platelet count (rs10058074 near *IRF1*, rs210142 in *BAK1*), whereas one ALL risk allele was associated with reduced platelet count (rs2836441 in *ERG*). In addition to hematopoietic transcription factor genes, pleiotropic variants in cell-cycle regulators are also candidate modifiers of both platelet count and ALL risk, as supported by our identification of a shared locus in pro-apoptotic protein *BAK1*.

The ALL risk SNP that we identify at 5q31.1 (rs10058074) is intronic, but has suggestive functional significance as a *cis*-acting eQTL for *IRF1*, a master transcriptional regulator of

immune response and oncogenesis (45, 46), as well as for *PDLIM4,* an F-actin-binding protein that influences T cell trafficking (47). This is one of the first ALL risk loci found that is related to "immune response gene elements", long posited to be important based on a hypothesized infectious etiology for ALL (11). Interferon immune responses are particularly important for controlling viral pathogens, which is notable since congenital cytomegalovirus infection was recently associated with ALL risk (48, 49). Two associated variants at 6p21.31 (rs210142, rs75080135) are also intronic, but both are *cis*-acting eQTLs for *BAK1* (BCL2 antagonist killer 1) (50), which encodes a pro-apoptotic protein that is a known CLL GWAS hit (51) and important for B cell homeostasis (52). Located 6kb apart and both associated with ALL risk in our discovery analysis, rs210142 and rs75080135 are in only weak LD in 1000 Genomes Europeans (R^2=0.11) and were both associated with ALL risk in UK replication data, although only the rs210142 association survived Bonferroni correction. The associated SNP at 21q22.2, rs2836441, is located in the 5' untranslated region of *ERG,* a transcription factor from the erythroblast transformation-specific family that is frequently deleted or alternatively spliced in the DUX4-rearranged ALL subtype (53).

Since our analyses were completed, the largest ALL GWAS to-date has been published by Vijayakrishnan and colleagues (54). This meta-analysis of four GWAS totaling 5,321 cases and 16,666 controls identified 4 novel B-ALL risk loci reaching genome-wide significance (54). Of these 4 loci, one (9q21.31) was significant for B-ALL risk overall, two (5q31.1, 6p21.31) for the high-hyperdiploid subtype, and one (17q21.32) for the ETV6-RUNX fusion subtype. Notably, their 5q31.1 and 6p21.31 risk loci overlap substantially with those identified through our hybrid GWAS-PheWAS approach. Their lead SNP at 5q31.1 (rs886285), located in C5orf56, is in weak LD (R^2=0.17) with the SNP (rs10058074) discovered through our approach, yet both variants appear to modulate expression of the master transcription factor *IRF1*. Compellingly, their lead SNP at 6p21.31 (rs210143) is only ~100 bases away (R^2=0.95) from the SNP identified through our approach (rs210142), and they too detected multiple signals in *BAK1* that implicate

decreased expression of this pro-apoptotic protein as an important hallmark of leukemogenesis. The GWAS meta-analysis from Vijayakrishnan and colleagues also confirmed *ERG* (21q22.2) as an ALL risk locus in European-ancestry populations, which we and others had previously identified as a GWAS hit for ALL in Hispanic populations (41, 42), but had been unable to replicate in European-ancestry populations.

While risk loci at 5q31.1, 6p21.31, and 21q22.2 were very recently associated with the high-hyperdiploid subtype of ALL at genome-wide significance, our results suggest that these susceptibility loci may influence ALL risk overall – not just subtype-specific risk – and may also be broadly involved in non-malignant hematopoiesis. The fact that the same risk loci identified through a largescale collaborative GWAS and a recent Hispanic ALL GWAS were uncovered through our combined GWAS-PheWAS methodology, despite our limited sample size, confirms the utility of the approach we have developed. These results provide further evidence of the importance of these loci in B-cell ALL and suggests our approach has applicability to the study of rare malignancies, including childhood cancers.

There are several limitations to our study and valid concerns of our hybrid GWAS-PheWAS approach. One limitation of using the UK Biobank for this study investigating genetic risk in a pediatric cancer is that the UK Biobank GeneATLAS PheWAS database was constructed using genetic and EHR data from adults 40-69 years of age. Thus, applying this approach to rare adult-onset diseases may be more appropriate than for pediatric diseases, as the overlap between these adult traits and pediatric phenotypes are largely unknown. For the PheWAS analyses, we used a p-value threshold of <0.01 to carry a SNP-trait association forward to enrichment analyses, rather than a Bonferroni-corrected threshold (*i.e.* 0.05/778 traits). While many of the traits in the GeneATLAS are highly correlated (*e.g.* standing height and sitting height, BMI and waist-to-hip ratio), cancers that have more than just 12 GWAS hits to evaluate via PheWAS may benefit from a more stringent threshold. Another significant limitation of our study was the limited case sample size in our discovery dataset. Because of our limited

sample size, we implemented a two-stage study design, first screening for nominally-associated SNPs in our discovery dataset (P<0.05), and then attempting replication in an independent sample. Despite these limitations, interest in our hybrid GWAS-PheWAS approach for investigating inherited genetic risk in rare diseases, where traditional approaches remain limited, appears warranted.

Although multiple GWAS in the past decade have contributed to our understanding of inherited susceptibility to ALL, there remains significant missing heritability (55). The most recent and largest ALL GWAS determined that known risk alleles accounted for 31% of the total variance in genetic risk of ALL (54); thus, there is a need for additional studies investigating ALL genetic risk loci. A recent review on the benefits and pitfalls GWAS emphasized the need for novel analytic approaches to enhance our understanding of genotype-phenotype associations in the post-GWAS era and the utility of large biorepository databases linking EHR and genotyping data, polygenic scores, and innovative study designs (8). This review also highlighted that, while increasing GWAS sample size may reveal more associations, new methods for analyzing the wealth of existing data are essential (8).

One opportunity would be to leverage LD score regression to identify traits associated with a cancer of interest. Although the sample-size limitations that apply to our study would also apply to analyses using LD score regression, replacing the PheWAS portion of our methodology with LD score regression is an intriguing approach for identifying traits with shared genetic determinants in future applications. In summary, our novel hybrid application of PheWAS represents a promising approach to investigate inherited genetic risk, especially in childhood cancers where GWAS remain underpowered and where innovative analytic strategies can help to decipher complex etiology and guide future prevention and screening strategies.

ACKNOWLEDGEMENTS

The ALL Relapse GWAS dataset was generated at St. Jude Children's Research Hospital and by the Children's Oncology Group, supported by NIH grants CA142665, CA21765, CA158568, CA156449, CA36401, CA98543, CA114766, CA140729, and U01GM92666, Jeffrey Pride Foundation, the National Childhood Cancer Foundation, and by ALSAC. Funding for the project was provided by the Wellcome Trust under award 076113 and 085475.

References

1. Sud A, Kinnersley B, Houlston RS. Genome-wide association studies of cancer: current insights and future perspectives. Nature reviews Cancer. 2017;17(11):692-704.
2. Enciso-Mora V, Hosking FJ, Sheridan E, Kinsey SE, Lightfoot T, Roman E, et al. Common genetic variation contributes significantly to the risk of childhood B-cell precursor acute lymphoblastic leukemia. Leukemia. 2012;26(10):2212-5.
3. Plon SE, Lupo PJ. Genetic Predisposition to Childhood Cancer in the Genomic Era. Annu Rev Genomics Hum Genet. 2019;20:241-63.
4. Semmes EC, Zhang C, Walsh KM. Intermediate phenotypes underlying osteosarcoma risk. Oncotarget. 2018;9(100):37345-6.
5. Zhang C, Morimoto LM, de Smith AJ, Hansen HM, Gonzalez-Maya J, Endicott AA, et al. Genetic determinants of childhood and adult height associated with osteosarcoma risk. Cancer. 2018;124(18):3742-52.
6. Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. Nature reviews Genetics. 2010;11(12):843-54.
7. Bush WS, Oetjens MT, Crawford DC. Unravelling the human genome-phenome relationship using phenome-wide association studies. Nature reviews Genetics. 2016;17(3):129-45.
8. Tam V, Patel N, Turcotte M, Bosse Y, Pare G, Meyre D. Benefits and limitations of genome-wide association studies. Nature reviews Genetics. 2019.
9. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotechnol. 2013;31(12):1102-10.
10. Denny JC, Bastarache L, Roden DM. Phenome-Wide Association Studies as a Tool to Advance Precision Medicine. Annu Rev Genomics Hum Genet. 2016;17:353-73.
11. Greaves M. A causal mechanism for childhood acute lymphoblastic leukaemia. Nature reviews Cancer. 2018;18(8):471-84.
12. Inaba H, Greaves M, Mullighan CG. Acute lymphoblastic leukaemia. Lancet (London, England). 2013;381(9881):1943-55.
13. Mori H, Colman SM, Xiao Z, Ford AM, Healy LE, Donaldson C, et al. Chromosome translocations and covert leukemic clones are generated during normal fetal development. Proceedings of the National Academy of Sciences of the United States of America. 2002;99(12):8242-7.
14. Iacobucci I, Mullighan CG. Genetic Basis of Acute Lymphoblastic Leukemia. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2017;35(9):975-83.
15. Trevino LR, Yang W, French D, Hunger SP, Carroll WL, Devidas M, et al. Germline genomic variants associated with childhood acute lymphoblastic leukemia. Nature genetics. 2009;41(9):1001-5.
16. Wiemels JL, Walsh KM, de Smith AJ, Metayer C, Gonseth S, Hansen HM, et al. GWAS in childhood acute lymphoblastic leukemia reveals novel genetic associations at chromosomes 17q12 and 8q24.21. Nature communications. 2018;9(1):286.
17. de Smith AJ, Walsh KM, Francis SS, Zhang C, Hansen HM, Smirnov I, et al. BMI1 enhancer polymorphism underlies chromosome 10p12.31 association with childhood acute lymphoblastic leukemia. International journal of cancer. 2018;143(11):2647-58.
18. 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. Nature communications. 2015;6:7553.

19. 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):3298-307.
20. 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. Nature genetics. 2009;41(9):1006-10.
21. Vijayakrishnan J, Studd J, Broderick P, Kinnersley B, Holroyd A, Law PJ, et al. Genome-wide association study identifies susceptibility loci for B-cell childhood acute lymphoblastic leukemia. Nature communications. 2018;9(1):1340.
22. Vijayakrishnan J, Kumar R, Henrion MY, Moorman AV, Rachakonda PS, Hosen I, et al. A genome-wide association study identifies risk loci for childhood acute lymphoblastic leukemia at 10q26.13 and 12q23.1. Leukemia. 2017;31(3):573-9.
23. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic acids research. 2017;45(D1):D896-d901.
24. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics (Oxford, England). 2015;31(21):3555-7.
25. Pers TH, Timshel P, Hirschhorn JN. SNPsnap: a Web-based tool for identification and annotation of matched SNPs. Bioinformatics (Oxford, England). 2015;31(3):418-20.
26. Rothman K. Modern Epidemiology. Boston, MA: Little, Brown and Company; 1986.
27. Hennessy S, Bilker WB, Berlin JA, Strom BL. Factors influencing the optimal control-to-case ratio in matched case-control studies. American journal of epidemiology. 1999;149(2):195-7.
28. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic acids research. 2012;40(Database issue):D930-4.
29. Canela-Xandri O, Rawlik K, Tenesa A. An atlas of genetic associations in UK Biobank. Nat Genet. 2018;50(11):1593-9.
30. Yang JJ, Cheng C, Devidas M, Cao X, Campana D, Yang W, et al. Genome-wide association study identifies germline polymorphisms associated with relapse of childhood acute lymphoblastic leukemia. Blood. 2012;120(20):4197-204.
31. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447(7145):661-78.
32. O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, et al. A general approach for haplotype phasing across the full spectrum of relatedness. PLoS genetics. 2014;10(4):e1004234.
33. Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nature genetics. 2016;48(10):1284-7.
34. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nature genetics. 2016;48(10):1279-83.
35. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell. 2016;167(5):1415-29.e19.
36. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics. 2007;81(3):559-75.
37. Yavorska OO, Burgess S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. Int J Epidemiol. 2017;46(6):1734-9.

38. Burgess S, Thompson SG. Use of allele scores as instrumental variables for Mendelian randomization. Int J Epidemiol. 2013;42(4):1134-44.
39. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. Genet Epidemiol. 2016;40(4):304-14.
40. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int J Epidemiol. 2015;44(2):512-25.
41. de Smith AJ, Walsh KM, Morimoto LM, Francis SS, Hansen HM, Jeon S, et al. Heritable variation at the chromosome 21 gene ERG is associated with acute lymphoblastic leukemia risk in children with and without Down syndrome. Leukemia. 2019;33(11):2746-51.
42. Qian M, Xu H, Perez-Andreu V, Roberts KG, Zhang H, Yang W, et al. Novel susceptibility variants at the ERG locus for childhood acute lymphoblastic leukemia in Hispanics. Blood. 2018.
43. Stanhope SA, Skol AD. Improved minimum cost and maximum power two stage genome-wide association study designs. PloS one. 2012;7(9):e42367.
44. Wason JM, Dudbridge F. A general framework for two-stage analysis of genome-wide association studies and its application to case-control studies. American journal of human genetics. 2012;90(5):760-73.
45. Taniguchi T, Lamphier MS, Tanaka N. IRF-1: the transcription factor linking the interferon response and oncogenesis. Biochim Biophys Acta. 1997;1333(1):M9-17.
46. Willman CL, Sever CE, Pallavicini MG, Harada H, Tanaka N, Slovak ML, et al. Deletion of IRF-1, mapping to chromosome 5q31.1, in human leukemia and preleukemic myelodysplasia. Science (New York, NY). 1993;259(5097):968-71.
47. Fu C, Li Q, Zou J, Xing C, Luo M, Yin B, et al. JMJD3 regulates CD4 T cell trafficking by targeting actin cytoskeleton regulatory gene Pdlim4. The Journal of clinical investigation. 2019;130:4745-57.
48. Wiemels JL, Talback M, Francis SS, Feychting M. Early infection with cytomegalovirus and risk of childhood hematological malignancies. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2019.
49. Francis SS, Wallace AD, Wendt GA, Li L, Liu F, Riley LW, et al. In utero cytomegalovirus infection and development of childhood acute lymphoblastic leukemia. Blood. 2017;129(12):1680-4.
50. Chittenden T, Harrington EA, O'Connor R, Flemington C, Lutz RJ, Evan GI, et al. Induction of apoptosis by the Bcl-2 homologue Bak. Nature. 1995;374(6524):733-6.
51. Slager SL, Skibola CF, Di Bernardo MC, Conde L, Broderick P, McDonnell SK, et al. Common variation at 6p21.31 (BAK1) influences the risk of chronic lymphocytic leukemia. Blood. 2012;120(4):843-6.
52. Takeuchi O, Fisher J, Suh H, Harada H, Malynn BA, Korsmeyer SJ. Essential role of BAX,BAK in B cell homeostasis and prevention of autoimmune disease. Proceedings of the National Academy of Sciences of the United States of America. 2005;102(32):11272-7.
53. Zhang J, McCastlain K, Yoshihara H, Xu B, Chang Y, Churchman ML, et al. Deregulation of DUX4 and ERG in acute lymphoblastic leukemia. Nature genetics. 2016;48(12):1481-9.
54. Vijayakrishnan J, Qian M, Studd JB, Yang W, Kinnersley B, Law PJ, et al. Identification of four novel associations for B-cell acute lymphoblastic leukaemia risk. Nature communications. 2019;10(1):5348.
55. Blanco-Gomez A, Castillo-Lluva S, Del Mar Saez-Freire M, Hontecillas-Prieto L, Mao JH, Castellanos-Martin A, et al. Missing heritability of complex diseases: Enlightenment by genetic

555 variants from intermediate phenotypes. BioEssays : news and reviews in molecular, cellular and
556 developmental biology. 2016;38(7):664-73.
557 56. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The
558 NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and
559 summary statistics 2019. Nucleic acids research. 2019;47(D1):D1005-d12.

Table 1. Summary of previously published genome-wide significant risk loci for B-cell ALL

Author (year reported) in ALL GWAS	Locus	ALL risk SNP	Gene	OR (95% CI)
Trevino LR *et al.* (2009)	7p12.2	rs11978267	*IKZF1*	1.69 (1.40-1.90)
Wiemels JL *et al.* (2018)	7p15.3	rs2390536	*SP4*	1.18 (1.11-1.24)
Wiemels JL et al. (2018)	8q24.21	rs4617118	Intergenic	1.34 (1.21-1.47)
Xu H *et al.* (2015)	9p21.3	rs3731249	*CDKN2A, CDKN2B*	1.63 (1.18-1.56)
de Smith AJ *et al.* (2018)	10p12.2	rs10741006	*PIP4K2A*	1.40 (1.40-1.53)
de Smith AJ *et al.* (2018)	10p12.31	rs12769953	*BMI1*	1.27 (1.20-1.35)
Migliorini G *et al.* (2013)	10p14	rs3824662	*GATA3*	1.31 (1.21-1.41)
Papaemmanuil E *et al.* (2009)	10q21.2	rs7089424	*ARID5B*	1.65 (1.54-1.36)
Wiemels JL *et al.* (2018)	10q26.13	rs3740540	*LHPP*	1.20 (1.15-1.28)
Vijayakrishnan J *et al.* (2017)	12q23.1	rs4762284	*ELK3*	1.19 (1.12-1.26)
Papaemmanuil E *et al.* (2009)	14q11.2	rs2239633	*CEBPE*	1.34 (1.22-1.41)
Wiemels JL *et al.* (2018)	17q21.1	rs2290400	*IKZF3*	1.18 (1.11-1.25)

Abbreviations: ALL, acute lymphoblastic leukemia; SNP, single nucleotide polymorphism; GWAS, genome-wide association study; OR, odds ratio; 95% CI, 95% confidence interval

rsIDs from GRCh37/hg19 build

Table 2. Selected PheWAS traits compared between ALL SNP-set and control SNP-sets for enrichment[a]

UK Biobank PheWAS Trait	ALL vs. CLL SNP-set	ALL vs. matched control SNP-set[b]	UK Biobank PheWAS Trait	ALL vs. CLL SNP-set	ALL vs. matched control SNP-set
	P[c]	*P*[c]		*P*[c]	*P*[c]
platelet count	**0.047**	**<0.001**	weight	0.737	**0.035**
lymphocyte count	0.191	**<0.001**	white blood cell count	0.865	**0.035**
monocyte percentage	**0.033**	**<0.001**	whole body fat mass	0.568	**0.004**
monocyte count	0.309	**0.002**	whole body fat-free mass	0.737	**0.035**
neutrophil percentage	0.531	**0.001**	whole body water mass	0.737	**0.035**
platelet crit	0.664	**<0.001**	asthma	0.815	0.393
eosinophil count	0.892	**0.002**	body fat percentage	0.815	0.080
impedance of whole body	0.103	**0.012**	body mass index	0.615	0.393
standing height	0.248	**0.039**	melanoma/malignant skin neoplasms	1.000	**0.028**
basophil count	**0.048**	**<0.001**	hematocrit percentage	0.815	0.080
comparative height (age 10)	0.169	**<0.001**	hemoglobin concentration	1.000	0.080
eosinophil percentage	0.956	**0.006**	hip circumference	1.000	0.080
neutrophil count	1.000	**0.015**	chronic rheumatic heart diseases	**0.026**	**0.028**
basal metabolic rate	0.737	**0.013**	multiple valve diseases	**0.026**	**0.028**
basophil percentage	0.073	**0.013**	mean reticulocyte volume	0.387	0.393
lymphocyte percentage	0.583	0.071	mean sphered cell volume	0.156	0.162
mean platelet volume	1.000	0.124	oily fish intake	0.418	0.080
platelet distribution width	1.000	0.071	trunk fat mass	1.000	0.080
sitting height	1.000	**0.035**	waist circumference	1.000	**0.028**
trunk fat-free mass	1.000	0.124	water intake	0.105	**0.005**
trunk predicted mass	1.000	0.124	alcohol intake frequency	0.376	0.747

Abbreviations: ALL, acute lymphoblastic leukemia; SNPs, single nucleotide polymorphisms; CLL, chronic lymphocytic leukemia
Bold values indicate nominal significance ($P < 0.05$)

[a] Individual SNP-trait associations available in S2 Table (ALL SNP-set), S3 Table (CLL SNP-set) and S4 Table (SNPsnap SNP-set)
[b] Control SNPs generated using SNPsnap matched to ALL SNPs based on minor allele frequency (±5%), surrounding gene density (±50%), distance to nearest gene (±50%), and linkage disequilibrium at $R^2 \geq 0.50$ (±50%).
[c] *P* value calculated with fisher's exact test, summary of all 76 PheWAS traits tested for enrichment in S6 Table

Table 3. Multivariate logistic regression of platelet count-associated variants and ALL risk in discovery case-control cohort[a]

Locus	SNP rsID	Effect allele[b]	EAF[c]	Gene	OR[e] (95% CI)	*P*
2q32.3	rs7585866	G	0.37	*SDPR*	0.87 (0.77-0.99)	**0.041**
4q24	rs4699154	C	0.72	near *TET2*[d]	1.22 (1.06-1.39)	**3.80×10^{-3}**
5q31.1	rs10058074	G	0.57	near *IRF1*[d]	1.15 (1.02-1.30)	**0.023**
6p21.31	rs210142	C	0.73	*BAK1*	1.17 (1.02-1.33)	**0.021**
6p21.31	rs75080135	C	0.24	*GGNBP1*	1.20 (1.02-1.40)	**0.024**
6q23.3	rs1331308	C	0.51	*HBS1L*	1.17 (1.04-1.32)	**0.011**
6q23.3	rs7776054	G	0.27	*HBS1L*	0.84 (0.73-0.97)	**0.016**
7q32.2	rs11556924	T	0.38	*ZC3HC1*	0.88 (0.77-0.99)	**0.034**
12q24.21	rs35427	T	0.60	intergenic	0.87 (0.76-0.98)	**0.026**
18q12.3	rs16977972	T	0.17	*SETBP1*	1.20 (1.01-1.41)	**0.034**
21q22.2	rs2836441	G	0.11	*ERG*	0.81 (0.67-0.96)	**0.019**
22q11.21	rs1059196	C	0.66	*SEPT5, GP1BB*	1.15 (1.00-1.32)	**0.044**

Abbreviations: SNP, single nucleotide polymorphism; EAF, effect allele frequency; OR, odds ratio; 95% CI, 95% confidence interval
Bold values indicate nominal significance ($P < 0.05$)

[a] Multivariate logistic regression adjusted for sex and top 10 ancestry-informative principal components
[b] Effect allele coded as allele previously associated with increased platelet count from Astle *et. al.* (35)
[c] Effect allele frequency in European-ancestry individuals from 1000 Genomes Project
[d] Neighboring gene located on UCSC Genome Browser
[e] Odds of ALL associated with each additional copy of the effect allele
Sample size in discovery cohort (959 Children's Oncology Group cases, 2624 controls); rsIDs from GRCh37/hg19 build

Table 4. Independent replication of ALL risk loci in combined meta-analysis of UK GWAS II and German GWAS

Locus	SNP rsID	Effect allele	Gene	OR (95% CI)[c]	*P*	$P_{heterogeneity}$
2q32.3	rs7585866[a]	G	*SDPR*	-	-	-
4q24	rs4699154	C	near *TET2*[b]	0.98 (0.89-1.06)	0.621	0.300
5q31.1	rs10058074	G	near *IRF1*[b]	1.15 (1.05-1.26)	**8.46×10^{-4}**	0.625
6p21.31	rs210142	C	*BAK1*	1.19 (1.10-1.28)	**1.20×10^{-4}**	0.780
6p21.31	rs75080135	C	*GGNBP1*	1.15 (1.05-1.25)	6.80×10^{-3}	0.038
6q23.3	rs1331308	C	*HBS1L*	0.96 (0.88-1.04)	0.264	0.008
6q23.3	rs7776054	G	*HBS1L*	1.00 (0.91-1.09)	0.941	0.165
7q32.2	rs11556924	T	*ZC3HC1*	0.99 (0.90-1.07)	0.749	0.381
12q24.21	rs35427	T	intergenic	1.03 (0.95-1.13)	0.508	0.902
18q12.3	rs16977972[a]	T	*SETBP1*	-	-	-
21q22.2	rs2836441	G	*ERG*	0.85 (0.77-0.94)	**5.13×10^{-3}**	0.733
22q11.21	rs1059196[a]	C	*SEPT5, GP1BB*	-	-	-

Abbreviations: SNP, single nucleotide polymorphism; OR, odds ratio; 95% CI, 95% confidence interval

Bold values indicate Bonferroni-corrected significance ($P < 0.05/9$) with concordant direction of effect in replication analyses

[a] Data missing since SNPs did not pass quality control filtering in the replication cohort

[b] Neighboring gene located on UCSC Genome Browser

[c] Odds of ALL associated with each additional copy of the effect allele, estimates were determined using a fixed-effects model using beta values and standard errors

Sample size in replication cohort; combined UK GWAS II and German GWAS (1618 cases, 9409 controls) (21); rsIDs from GRCh37/hg19 build

FIGURE LEGENDS:

Figure 1. Methodology for hybrid analysis of GWAS and PheWAS data. This figure illustrates our approach for investigating phenotype associations with known disease risk variants in order to identify novel candidate risk loci and/or intermediate phenotypes for subsequent analysis in case-control cohorts. Specifically, this figure depicts our application of this approach to acute lymphoblastic leukemia (ALL), which identified platelet count as a phenotype enriched for association with ALL GWAS hits and downstream analysis of the role of platelet count-associated variants in relation to ALL risk in a case-control cohort. Created with Biorender. NHGRI-EBI GWAS catalog diagram attributable to Buniello *et. al.* (56)

[a] GWAS catalog (NHGRI-EBI) - https://www.ebi.ac.uk/gwas/ (23, 56)

[b] PheWAS catalog (UK Biobank GeneATLAS) - http://geneatlas.roslin.ed.ac.uk/phewas/ (29)

[c] SNPsnap controls (Broad Institute) - https://data.broadinstitute.org/mpg/snpsnap/(25)

[d] PLINK (genome association analysis toolkit) - https://www.cog-genomics.org/plink2 (36)

Abbreviations: GWAS, genome-wide association study; PheWAS, phenome-wide association study; ALL, acute lymphoblastic leukemia; SNPs, single nucleotide polymorphisms.

Figure 2. UK Biobank PheWAS traits in ALL SNP-set versus control SNP-sets. This figure shows the percentage of ALL-associated (12 SNPs total), CLL-associated (31 SNPs total), and matched control SNPs (48 SNPs total) that were significantly associated ($P < 0.01$) with PheWAS traits in the UK Biobank Traits are depicted in descending order of percentage/proportion associated with ALL SNPs from left to right then top to bottom. (A) shows the first subset of 38 traits and (B) shows the second subset of 38 traits, since 76 traits total were significantly associated with >1 SNP in the ALL SNP-set, and thus were carried forward for statistical analysis and enrichment comparisons across SNP-sets (see **S5 Table and S6 Table** for full results of proportions and of trait enrichment comparisons between SNP sets).

Abbreviations: PheWAS, phenome-wide association study; ALL, acute lymphoblastic leukemia; CLL, chronic lymphocytic leukemia; SNP, single nucleotide polymorphism

Figure 1

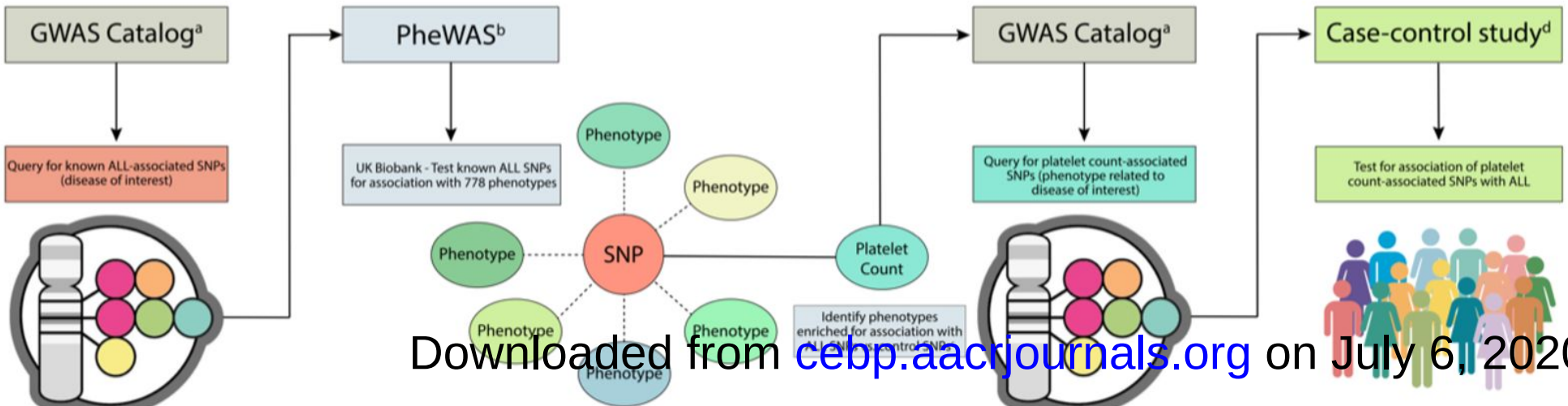


Figure 2A.

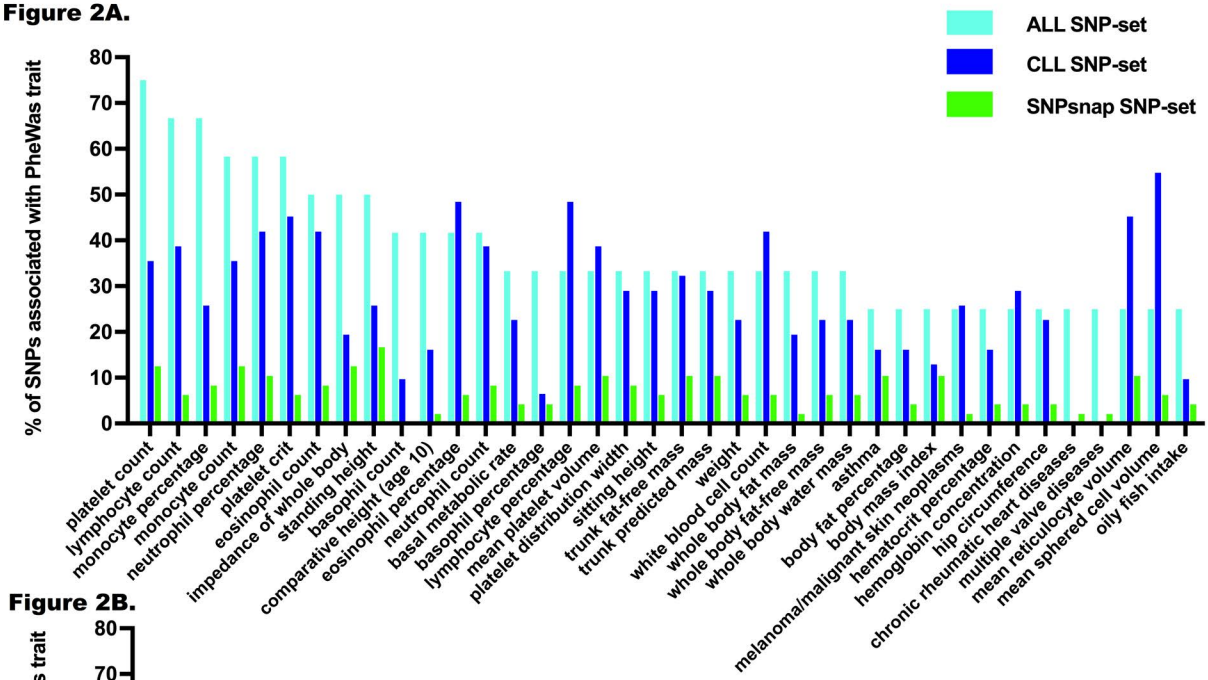
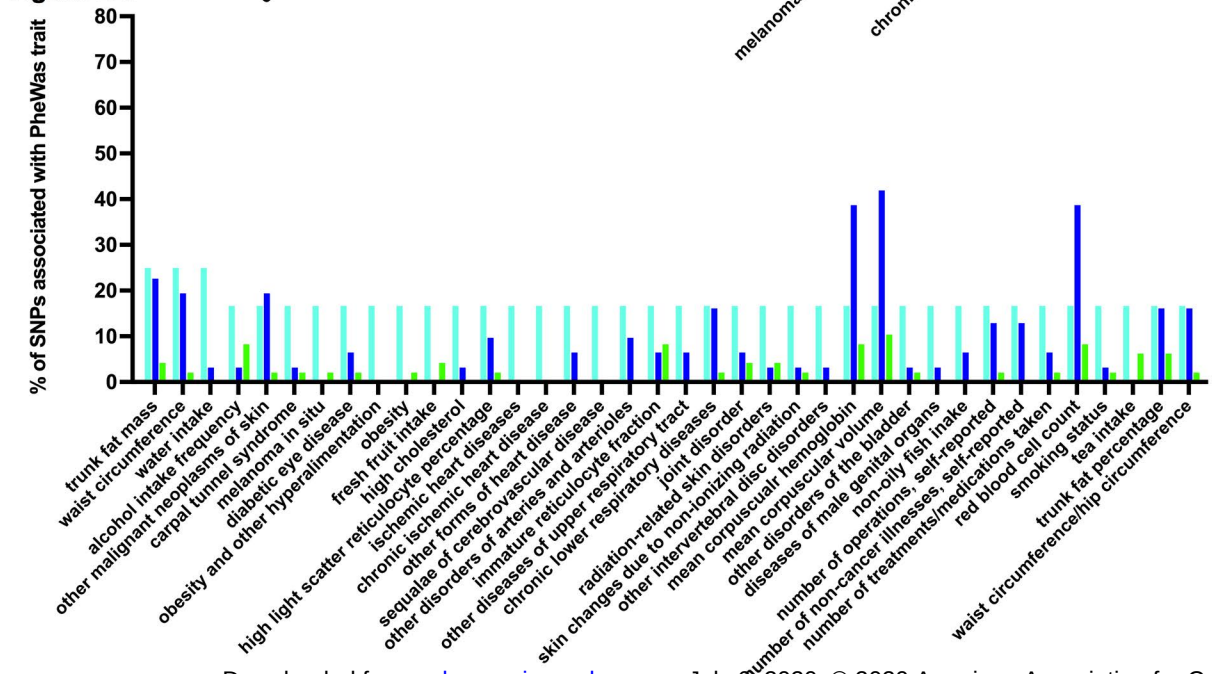


Figure 2B.



Cancer Epidemiology, Biomarkers & Prevention

AACR American Association for Cancer Research

Leveraging genome and phenome-wide association studies to investigate genetic risk of acute lymphoblastic leukemia

Eleanor C Semmes, Jayaram Vijayakrishnan, Chenan Zhang, et al.

Cancer Epidemiol Biomarkers Prev Published OnlineFirst May 28, 2020.