medRxiv preprint doi: https://doi.org/10.1101/2021.05.07.21256849; this version posted June 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Genome-wide trans-ethnic meta-analysis identifies novel susceptibility loci for childhood acute lymphoblastic leukemia

Transethnic GWAS on acute lymphoblastic leukemia

Soyoung Jeon[1,2], Adam J. de Smith[1], Shaobo Li[1,2], Minhui Chen[1], Tsz Fung Chan[1], Ivo S. Muskens[1], Libby M. Morimoto[3], Andrew T. DeWan[4,5], Nicholas Mancuso[1,6,7], Catherine Metayer[3], Xiaomei Ma[5], Joseph L. Wiemels[1], Charleston W.K. Chiang[1,6]

Affiliations:

1. Center for Genetic Epidemiology, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA
2. Cancer Biology and Genomics Graduate Program, Program in Biological and Biomedical Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA
3. Division of Epidemiology & Biostatistics, School of Public Health, University of California, Berkeley, CA
4. Center for Perinatal, Pediatric and Environmental Epidemiology, Yale School of Public Health, New Haven, CT
5. Department of Chronic Disease Epidemiology, Yale School of Public Health, New Haven, CT
6. Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA
7. Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, CA

Correspondence: Joseph Leo Wiemels, Center for Genetic Epidemiology, 1450 Biggy St, Los Angeles, California; email: wiemels@usc.edu, phone: (323) 442-7865, fax: (323) 442-7749; Charleston W.K. Chiang, 1450 Biggy St, Los Angeles, California; email: charleston.chiang@med.usc.edu, phone: (323)442-8052, fax: (323) 442-7749

Text word count: 4029

Abstract word count: 228

Numbers of figure and tables: 4 figures, 2 tables

Number of references: 57

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.05.07.21256849; this version posted June 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Key Points

- **We identified three novel loci on childhood acute lymphoblastic leukemia near genes previously associated with multiple blood cell traits**
- **Polygenic risk scores using known and novel risk variants showed similar efficacy in Latino and non-Latino white Americans**

Abstract

The risk of childhood acute lymphoblastic leukemia (ALL) differs across ethnic groups and there exist gaps in our understanding of the genetic risk of ALL as most studies took place in populations of predominantly European ancestries. In an effort to address these limitations, we performed a genetic meta-analysis of ALL in 76,317 participants across four ethnic groups, including 17,814 non-European individuals and 3,482 total cases. We replicated 15 out of 16 previously identified loci associated with ALL in our trans-ethnic analysis. We further identified five novel associations at genome-wide significance, including three novel loci and two secondary associations at previously known loci (17q12 and near *CEBPE*). The three putatively novel loci (rs9376090 near *MYB/HBS1L*, rs10998283 near *TET1*, and rs9415680 near *JMJD1C/NRBF2*) were previously shown to be associated with multiple blood cell traits and other hematopoietic cancers. Polygenic risk scores constructed from our trans-ethnic meta-analysis showed similar efficacy in independent Latino (LAT) and non-Latino white (NLW) ALL cohorts (AUC ~ 0.67-0.68) and could partly explain the increased risk of ALL in LAT compared to NLW. Cross-population analysis also showed high but significantly less than 100% genetic correlation between LAT and NLW, suggesting potential differences in the underlying genetic architecture between ethnic groups. In summary, our findings enhance the understanding of genetic contribution to ALL risk across diverse populations and highlight the importance to include multiple ethnic groups in GWAS.

medRxiv preprint doi: https://doi.org/10.1101/2021.05.07.21256849; this version posted June 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Introduction

Acute lymphoblastic leukemia (ALL) is the most common type of childhood cancer worldwide, with substantial racial and ethnic differences in incidence and treatment outcome[1,2]. Previous genome-wide association studies (GWAS) have confirmed the genetic basis of ALL susceptibility by identifying a number of risk loci for childhood ALL[3,4,5(p12),6–8] and estimating the heritability to be 21% (ref.[9]). However, the known risk loci together account for a relatively small portion of the total variance in genetic risk of ALL[9], suggesting that additional susceptibility alleles may be discovered in larger studies. Furthermore, these studies were generally performed in cohorts with a predominantly European ancestry. Latino children have the highest risk of ALL in the United States, with an incidence rate ~15-40% higher than in non-Latino whites[10–12] and an increased chance of relapse and poorer overall survival[13,14]. Yet, we have a limited understanding of the genetic architecture of ALL in non-European populations and the generalizability of findings from existing GWAS to non-European populations (but see recent efforts for studying the genetic etiology of childhood ALL in Latinos[15–17]). While environmental or social factors likely underlie some if not the majority of the differences in risk between ethnic groups, there may also be a difference in the genetic risk architecture that modulates risk across ethnic groups and would argue for the greater inclusion of other ethnic groups in genetic studies of ALL.

Given this context, we performed a trans-ethnic GWAS of childhood ALL in a discovery panel consisting of 76,317 individuals from an assembled multi-ethnic cohort. We note the complexity of discussing race, ethnicity and ancestry in a genetic study. As a convention, we used the following terms and abbreviations to refer to each ethnic group in our study: African American (AFR), East Asian (EAS), Latino American (LAT), and non-Latino white (NLW). These population labels are largely based on self-reported ethnic identity and we confirmed that they largely correlate with genetic ancestry as defined by the reference populations in 1000 Genomes[18] (Methods). Our cohort consisted of 3,482 cases and 72,835 controls for an effective sample size of 13,292, which is, to our knowledge, the largest trans-ethnic GWAS for ALL to date. We identified three novel ALL risk loci and tested the novel findings from our discovery panel in two additional independent cohorts. We further compared the efficacy of polygenic risk scores (PRS) to stratify individuals based on their risk of ALL in the two largest subgroups of our data, LAT and NLW. PRS models are known to be poorly transferred to non-European populations[19], but multi-ethnic designs may be more effective in identifying alleles with shared effects across population without explicit fine-mapping and produce more comparable PRS models between populations[20,21]. Finally, we leveraged our genome-wide summary statistics to contrast the genetic architecture of ALL between LAT and NLW populations.

medRxiv preprint doi: https://doi.org/10.1101/2021.05.07.21256849; this version posted June 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Materials and methods

Study Cohorts

The California Childhood Cancer Record Linkage Project (CCRLP) includes all children born in California during 1982-2009 and diagnosed with ALL at the age of 0-14 years per California Cancer Registry records. Children who were born in California during the same period and not reported to California Cancer Registry as having any childhood cancer were considered potential controls. Detailed information on sample matching, preparation and genotyping has been previously described.[4] Because ALL is a rare childhood cancer, for the purpose of a genetic study we followed previous practice[4] and incorporated additional controls using adult individuals from the Kaiser Resource for Genetic Epidemiology Research on Aging Cohort (GERA; dbGaP accession: phs000788.v1.p2). The GERA cohort was chosen because a very similar genotyping platform had been used: Affymetrix Axiom World arrays. For replications we included two independent ALL cohorts: (1) individuals of predominantly European ancestry from the Children's Oncology Group (COG; dbGAP accession: phs000638.v1.p1) as cases and from Wellcome Trust Case–Control Consortium[22] (WTCCC) as controls; and (2) individuals of European and Latino ancestry from the California Childhood Leukemia Study (CCLS), a non-overlapping California case-control study (1995-2008).[23] The quality control and imputation for both the discovery and replication cohorts were conducted in ethnic strata and generally followed previous pipelines of ALL GWAS, but with additional attention paid to incorporate the entire GERA cohort and ensuring data quality post-imputation. See Supplemental Methods for details. This study was approved by Institutional Review Boards at the California Health and Human Services Agency, University of Southern California, Yale University, and the University of California San Francisco.

Association Testing

We used SNPTEST[24] (v2.5.2) to test the association between imputed genotype dosage and case-control status in logistic regression, after adjusting for the top 20 principal components (PCs). Sex was not included as a covariate, and we found sex was not correlated with genotype dosage of any of the putatively associated SNPs (data not shown). Results from the four ethnic-stratified analyses were combined via the fixed-effect meta-analysis with variance weighting using METAL[25]. Only variants passing QC in at least three of the four ethnic groups were meta-analyzed. A genome-wide threshold of 5×10^{-8} was used for significance in the discovery stage. A Bonferroni-corrected significance of 0.00312 (=0.05/16) was used for replication of previously reported susceptibility variants[3–8,26–28]. Cochran's *Q*-test for heterogeneity was performed using METAL[25]. To perform conditional analysis in identifying secondary associations within a locus, the lead SNP was additionally included in the regression model, again using 5×10^{-8} as threshold for significance.

Polygenic Risk Score Analysis

Polygenic risk scores (PRS) for ALL were constructed using PLINK (v2.0) by summing the genotype dosages of risk alleles, each weighted by its effect size from our discovery GWAS meta-analysis. PRS were constructed based on: (1) lead SNPs in the 16 known loci (N = 18

medRxiv preprint doi: https://doi.org/10.1101/2021.05.07.21256849; this version posted June 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

SNPs, including variants from the two secondary signals in *IKZF1* and *CDKN2A/B* that were previously reported; for which we used the corresponding effect sizes from conditional analysis), and (2) by additionally including the novel hits (N = 23 SNPs, including the additional 3 novel loci and 2 novel conditional associations). Associations between PRS and case-control status for ALL were tested in each group adjusting for 20 PCs using R. To evaluate the predictive power of PRS, Area Under the receiver operating characteristic Curve (AUC) were calculated using pROC package[29] in R.

Genetic architecture of ALL within and between populations

To investigate the genetic architecture of ALL and contrasting this architecture between NLW and LAT populations, we estimated the percentage of familial relative risk (FRR) explained by associated variants individually or in aggregate, the heritability ascribable to all post-QC imputed SNPs with MAF $\geq$ 0.05, the genetic correlation between NLW and LAT, and the genome-wide proportion of causal variants that are population-specific or population-shared. See the Supplemental Methods for details.

medRxiv preprint doi: https://doi.org/10.1101/2021.05.07.21256849; this version posted June 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Results

Trans-ethnic Genetic Associations with ALL

We performed a trans-ethnic meta-analysis GWAS for childhood ALL. After quality control filtering, our dataset consisted of 3,482 cases and 72,835 controls (Supplementary Table S1; Methods) in total. In contrast to the previous trans-ethnic analysis[4], we included additional controls for NLW and added the EAS cohort. Furthermore, we tested the association at 7,628,894 imputed SNPs, including low frequency (MAF between 1-5%) variants that were not previously systematically tested. We aggregated the summary statistics across the four ethnic groups in a fixed-effect meta-analysis. The genomic control inflation factor was 1.022 after excluding 16 previously known ALL-associated loci (Table 1), suggesting our meta-analysis was reasonably robust to any confounding due to population stratification (Figure 1). In total, twelve loci reached genome-wide significance (i.e, $P < 5.0 \times 10^{-8}$) in our analysis.

We found that for the 16 previously published risk loci for ALL[3–9,17,27,28], all were associated with ALL at the nominal level ($P < 0.05$) or have a SNP nearby with strong association (Table 1). Nearly all of the published risk SNPs show consistent direction of effects across ethnic groups (13/16 SNPs with heterogeneity P-value > 0.05; P = 0.0384, 0.006, 0, 0.000259 for AFR, EAS, LAT, NLW respectively for consistent direction of effect by the sign test, Supplemental Table S2). In some cases, the published SNP is not the SNP with the most significant association in our dataset, though usually our top SNP in the locus is in strong LD with the reported SNP (Table 1). Given the larger sample size and trans-ethnic analysis, the best associated variants in our analysis may reflect the more likely causal / shared association across populations. Two loci at *C5orf56* and *TLE1* are noted. At the *C5orf56* locus on 5q31, the variant previously reported in an independent European-ancestry cohort (rs886285) to be associated with a particular subtype of ALL (HD-ALL)[9] was not nominally associated with ALL overall (P = 0.63) in our dataset. A weakly linked SNP (rs11741255; r2 = 0.35 in NLW, 0.19 in LAT) in the same locus approximately 20kb away was significantly associated with ALL in our data ($P = 1.69 \times 10^{-4}$) but may reflect a chance association. At the *TLE1* locus on 5q21, neither the published variant nor our top variant in the locus would be considered significantly associated after Bonferroni correction (minimum $P = 1.06 \times 10^{-2}$ for rs62579826), possibly due to heterogeneity driven by EAS in which both the published variant and our top variant are monomorphic[30].

More importantly, we discovered three putatively novel susceptibility loci: one at 6q23 and two at 10q21 (Figure 1). The strongest association signal in 6q23 is at rs9376090 ($P = 8.23 \times 10^{-9}$, OR=1.27) in the intergenic region between *MYB* and *HBS1L* (Figure 2A). This association is mainly driven by NLW presumably due to its large sample size (Supplementary Table S1). In 10q21, there were two independent signals that showed genome-wide significance. One locus was identified with the lead SNP rs9415680 ($P=7.27 \times 10^{-8}$, OR=1.20), within a broad association peak, with apparently long-range LD with SNPs covering *NRBF2, JMJD1C*, and parts of *REEP3* (Figure 2B). The second locus in 10q21 was identified 5Mb away, with lead SNP rs10998283 ($P=3.92 \times 10^{-8}$, OR=1.15) in an intronic region in *TET1* (Figure 2C). The association signals for both loci in 10q21 were largely driven by LAT. We used the convention of the nearest genes to refer to these loci for the remainder of the manuscript, acknowledging that they may not be the causal genes.

medRxiv preprint doi: https://doi.org/10.1101/2021.05.07.21256849; this version posted June 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

To replicate our findings in independent datasets, we tested the associations of the three novel variants and their LD proxies (with $P < 5 \times 10^{-7}$;n=141) in independent samples from the COG/WTCCC and CCLS cohorts (Methods). For the *MYB/HBS1L* locus, which was driven by NLW in the discovery cohort, we replicated the signal in COG/WTCCC cohort (rs9376090, $P_{COG} = 4.87\times10^{-3}$, $P_{COG+discovery\ analysis} = 1.23\times10^{-10}$; Supplementary Table S3), but did not replicate in CCLS likely owing to the small sample size of NLW. For the *TET1* locus, in which the original association was driven by LAT in the discovery, three of the four SNPs with $P < 5\times10^{-7}$ in the discovery cohort nominally replicated in CCLS. The lead SNP after meta-analyzing the discovery cohort and the replication cohort of CCLS was rs79226025 ($P_{CCLS} = 3.04 \times10^{-2}$, $P_{CCLS+discovery} = 6.81 \times10^{-9}$; Supplementary Table S3). For the *NRBF2* / *JMJD1C* locus, we did not observe an association in the replication cohorts.

We also performed conditional analyses adjusting for the lead SNP at each locus and identified a secondary signal in four out of the 16 previously known loci (Table 2, Figure 3). In all cases, the LD between the secondary hit and the top hit in the locus are low (Table 2). The additional second associations in *CDKN2A* and *IZKF1* loci were previously noted[9(p1)]. In *CEBPE* (rs60820638, $P=5.38 \times10^{-8}$) and 17q12 (rs12944882, $P=7.71 \times10^{-10}$), these secondary signals represent novel associations. In particular, at the *CEBPE* locus, previous reports suggest multiple correlated variants with functional evidence[31,32] . Our analysis is consistent with the two previous variants (rs2239635 and rs2239630) being or tagging the same underlying signal, while the new association we identified (rs60820638) is an independent association (Supplementary Table S4).

Polygenic Risk Score

To assess the combined effect of all identified risk alleles for ALL, we constructed a PRS model in our discovery cohort, using either the 18 SNPs from 16 previously known loci or the 23 known plus novel SNPs and their associated effect sizes from the trans-ethnic meta-analysis. We then computed and tested the PRS for NLW and LAT individuals in the independent CCLS and COG/WTCCC cohorts. The scores generated with the known risk loci were significantly associated with case-control status in all groups ($P_{CCLS\ NLW}=2.22\times10^{-17}$, $P_{CCLS\ LAT}=4.78\times10^{-23}$, $P_{COG/WTCCC} =2.99\times10^{-62}$, Supplementary Table S5). Adding the three novel loci identified in this study and the two novel secondary signals further strengthened the evidence of the association in COG/WTCCC ($P =6.93\times10^{-63}$) and CCLS LAT ($P = 5.75 \times10^{-24}$), while the evidence of association stayed about the same in CCLS NLW ($P =2.03 \times10^{-17}$). The predictive accuracy as measured by AUC are similar between NLW and LAT, at around 67-68%, consistent with the hypothesis that trans-ethnic meta-analysis will enable PRS to be more transferrable between populations.

We also examined the distribution of PRS in CCRLP individuals (Figure 4). We found that while the shape of the PRS distribution is consistent with a normal distribution (Kolmogorov-Smirnov P = 0.918 and 0.303 for LAT and NLW, respectively) and appears similar between LAT and NLW (standard deviation of 0.728 and 0.735 respectively; F-Test P = 0.633), the scores in LAT are shifted to the right compared to the scores in NLW (mean of 5.101 and 4.641 respectively, Welch t-test $P = 1.3\times10^{-122}$). The observed pattern was consistent when the scores were stratified by case-control status (mean of 5.324 and 4.881 in LAT and NLW cases, respectively, P=3.956

medRxiv preprint doi: https://doi.org/10.1101/2021.05.07.21256849; this version posted June 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

$x10^{-58}$; mean of 4.895 and 4.414 in LAT and NLW controls, respectively, with P= 1.493 $x10^{-78}$). This observation was also replicated in CCLS with mean of 5.119 in LAT and 4.607 in NLW (P=4.596 $x10^{-51}$). Therefore, results from our PRS analyses are consistent with the notion that differences in allele frequency of ALL risk loci between populations may complement other non-genetic factors for ALL risk, and partly explain the increased ALL risk in LAT relative to NLW children and LAT.

Genetic architecture of ALL in Latinos and non-Latino whites

We estimated the relative contributions of each variant to ALL risk by computing the familial relative risk (Supplementary Table 6). In CCLS, where effect size estimates are expected to be less biased by winner's curse, the known risk variants accounted for 22.7% and 23.2% of familial relative risk in LAT and NLW, and the addition of novel variants increased these estimates to 24.3% and 24.8%, respectively (Supplemental Table S6).

The heritability of ALL attributable to all common SNPs (MAF $\geq$ 0.05) was estimated to be 20.3 ± 3.2% in NLW and 4.1 ± 2.0% in LAT using the GCTA-LDMS framework[33], and 20.2 ± 4.7%% in NLW and 11.1 ± 3.6% in LAT using the phenotype-correlation-genotype-correlation (PCGC) regression framework (Supplemental Table S7A). The heritability estimates in NLW are consistent in both approaches and with that previous reported[9(p1)]. Because the imputation quality using HRC reference panel is expected to be high for variants with MAF between 1-5% in NLW, our dataset also provides the opportunity to estimate the frequency-stratified contribution to the heritability of ALL in NLW. The inclusion of low frequency variants increased the estimated heritability in NLW to 29.8 ± 4.3% using REML (divided ~16.2% due to common variants, 13.5% due to low frequency variants; Supplemental Table S7B). Taking advantage of the admixed nature of LAT, whereby ancestry segments could capture effects beyond that directly attributable to assayed SNPs (such as the estimate from GCTA-LDMS), we also adopted an approach described in Zaitlen et al[34] to estimate the total narrow-sense heritability for ALL in LAT to be 37.3 ± 6.9%. Taken together, multiple lines of evidence suggest that increasing sample sizes will identify additional low frequency associations to ALL in the future.

Furthermore, we estimated the genetic correlation of ALL between NLW and LAT to be high (rG = 0.714 ± standard error 0.130) but significantly different from 1 (P = 0.014, Supplementary Table 8). This indicates the genetic architectures of NLW and LAT may be similar as expected from correlated effect sizes (Supplementary Figure S1) but not perfectly concordant. We complemented this analysis further by estimating the number of population-specific and shared causal alleles using the program PESCA[35]. The PESCA framework defines the set of causal variants as all variants tested to have a non-zero effect, even if the effect is indirect and only statistical rather than biological in nature. Using this framework, we estimated that approximately 32.5% of SNPs inferred to be causal are shared between NLW and LAT (1.71% of all common SNPs were inferred to have nonzero effects in both NLW and LAT; 1.69% and 1.87% were inferred to have population-specific nonzero effects in NLW and LAT, respectively). Together, these results suggests that there may be ethnic-specific genetic risk profiles or differential interactions with the environment that contributes to differences in disease risk between NLW and LAT. However, it should be noted that these analyses adopted the REML

medRxiv preprint doi: https://doi.org/10.1101/2021.05.07.21256849; this version posted June 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

framework or used the GCTA-LDMS estimates as hyperparameters, which could be biased in the context of LAT population here (see Discussion).

Discussion

By incorporating data across four ethnic groups, we have performed the largest trans-ethnic meta-analysis GWAS of childhood ALL to date. We identified three putatively novel susceptibility loci and two additional independent risk associations at previously reported loci. Our analysis suggests that the known and novel ALL risk alleles together explained about 25% of the familial relative risk in both NLW and LAT populations, and that the trans-ethnic PRS we constructed, although relatively simple and utilizing only the genome-wide associated variants, performed similarly in both NLW and LAT in predicting ALL (AUC ~ 67-68%).

In support of their potential role in ALL etiology, each of the three novel loci harbors genes and/or variants with a role in hematopoiesis and leukemogenesis as annotated by HaploReg (version 4.1) [36] and GTEx portal[37]. The associated variants in 6q23 are located between *HBS1L* and *MYB*, a myeloblastosis oncogene that encodes a critical regulator protein of lymphocyte differentiation and hematopoiesis[38] . This locus is already well known for associations with multiple blood cell measurements, severity of major hemoglobin disorders, and β-thalassemia[39,40]. The associated SNPs in our study fall within *HBS1L-MYB* intergenic region known to harbor multiple variants that reduce transcription factor binding, affect long-range interaction with *MYB*, and impact *MYB* expression[39,41]. The lead SNP rs9376090 is in a predicted enhancer region in K562 leukemia cells and GM12878 lymphoblastoid cells, and is a known GWAS hit for platelet count[38] and hemoglobin concentration[42,43]. Also, it is an eQTL in lymphocytes and whole blood[37] for *ALDH8A1*, which encodes aldehyde dehydrogenases, a cancer stem cell marker and a regulator self-renewal, expansion, and differentiation.

One of the associated loci in 10q21 has a distinct haplotype structure, with 130 highly correlated SNPs ($r^2 > 0.8$) associated with ALL (Figure 2B). This haplotype structure is observed in LAT and EAS, and the associations are driven by alleles with higher frequency in LAT and EAS than NLW or AFR (Supplementary Table S1, Supplementary Figure S2). This 400kb region is rich with genetic variants associated with blood cell traits such as platelet count, myeloid white cell count, and neutrophil percentage of white cells[44,45]. It is also associated with IL-10 levels[46] which was shown to be in deficit in ALL cases[47]. The signal region is contained within the intron of *JMJD1C*, a histone demethylase that a recent study has found to regulate abnormal metabolic processes in AML[48]. Previous studies have found that it acts as a coactivator for key transcription factors to ensure survival of AML cells[49] and self-renewal of mouse embryonic stem cells[50].

The second locus in 10q21 contains intronic variants in the *TET1* gene, which is well known for its oncogenicity in several malignancies including AML[51]. A recent study showed the epigenetic regulator *TET1* is highly expressed in T-cell ALL and is crucial for human T-ALL cell growth in vivo[52]. We found the associations at this locus to be slightly stronger for T-ALL than for B-ALL in a small subset of individuals with ALL subtype information, though the difference is not statistically significant (Supplemental Table S9). Of the four significant variants in this locus, SNP rs58627364 lies in the promoter region of *TET1* while the remaining three variants did not appear to overlap functional elements (Supplementary Figure S3). However, none of these SNPs

medRxiv preprint doi: https://doi.org/10.1101/2021.05.07.21256849; this version posted June 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

were observed as eQTL for *TET1* in whole blood or lymphoblastoid cells[53]; future studies may want to investigate whether these SNPs affect *TET1* expression in hematopoietic stem or progenitor cells.

In addition to identifying putative novel ALL risk loci, we capitalized on the large numbers of Latinos and non-Latino whites included in our study to explore the genetic architecture of ALL in these two populations. In the NLW population, we estimated that ~ 29% of the heritability of ALL was attributed to a combination of common and low-frequency (MAF between 1-5%) imputed variants using both GCTA-LDMS and PCGC regression (Supplementary Table S7B). This estimate is higher than the previous estimate of ~20% (ref. [9]), suggesting that there are additional low frequency variants associated with ALL that may be discovered in larger scaled studies. The picture is less clear among LAT, where the estimated heritability was perhaps unrealistically low using GCTA-LDMS (4.1% in univariate analysis; Supplementary Table S7A, 8). This estimate contrasted strongly against other lines of evidence that showed similar estimated effect sizes (r2 = 0.819; Supplementary Figure S1) and familial relative risks explained by GWAS loci (Supplementary Table S6) between NLW and LAT. Previous studies have noted the downward bias in REML heritability estimate in case-control studies, which is exacerbated when the covariates in the model (*i.e.* PCs and ancestry) are correlated with the disease status[54]. We thus also followed previous suggestions and used the PCGC regression to obtain variance component estimates[54,55], resulting in a higher heritability estimate (11.1%). Because of the underrepresentation of Native American or other non-European haplotypes in HRC panel, *a priori* we did not estimate the heritability including low-frequency imputed variants in LAT. When we did attempt to estimate heritability in this setting, we obtained a strongly negative REML-based heritability estimates (Supplementary Table S7B), suggesting potential model instability or misspecification attributed to the admixed nature of LAT[56]. Consequently, we also recommend caution when interpreting the estimated genetic correlation between LAT and NLW.

Nevertheless, we used bivariate version of the REML analysis to compute genetic correlation of ALL between populations, as had been done previously for prostate cancer with individual level data[57]. Our estimated genetic correlation (r_G = 0.71) is significantly less than 1, apparently suggesting a significant population-specific components of the disease architecture between LAT and NLW. This would be consistent with the findings of the *ERG* locus[16,17], a Latino-specific association with ALL, and suggest that future ethnic-specific GWAS across different ethnic groups for ALL will be insightful. This is also consistent with our observation in the PESCA analysis, where we found that only 32.5% of the estimated causal alleles are shared between LAT and NLW. These insights should still be treated with caution because the sample size for ALL, a rare disease, is still relatively small compared to complex traits examined using PESCA[35], and because the REML-based heritability estimates for LAT used as hyperparameter by PESCA may be biased. Therefore, more focused efforts to investigate the genetic architecture for ALL, particularly in admixed populations like the Latinos, is needed.

Future studies aimed to uncover the genetic risk factors for ALL could focus on multiple avenues. First, there will be a need to further increase the sample size of the study cohort, which would provide additional venues to replicate the putative novel findings here and identify more associated alleles at lower frequency. Second, there should be a focus on ethnic-specific GWAS for ALL, as ethnic-specific associations could be missed in a trans-ethnic GWAS. An example is

medRxiv preprint doi: https://doi.org/10.1101/2021.05.07.21256849; this version posted June 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

the *ERG* locus, which is not genome-wide significant in our meta-analysis. Finally, while not explored extensively in this particular study, there should be a focus on disentangling the different subtypes of ALL, and to study other aspects of the disease pathogenesis such as disease progression or risk of relapse, though these data are less available and may require more focused ascertainment and cohort creation.

Acknowledgement

This work was supported by research grants from the National Institutes of Health (R01CA155461, R01CA175737, R01ES009137, P42ES004705, P01ES018172, P42ES0470518 and R24ES028524) and the Environmental Protection Agency (RD83451101), United States. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health and the EPA. The collection of cancer incidence data used in this study was supported by the California Department of Public Health as part of the statewide cancer reporting program mandated by California Health and Safety Code Section 103885; the National Cancer Institute's Surveillance, Epidemiology and End Results Program under contract HHSN261201000140C awarded to the Cancer Prevention Institute of California, contract HHSN261201000035C awarded to the University of Southern California, and contract HHSN261201000034C awarded to the Public Health Institute; and the Centers for Disease Control and Prevention's National Program of Cancer Registries, under agreement U58DP003862-01 awarded to the California Department of Public Health. The biospecimens and/or data used in this study were obtained from the California Biobank Program, (SIS request #26), Section 6555(b), 17 CCR. The California Department of Public Health is not responsible for the results or conclusions drawn by the authors of this publication. We thank Hong Quach and Diana Quach for DNA isolation support. We thank Martin Kharrazi, Robin Cooley, and Steve Graham of the California Department of Public Health for advice and logistical support. We thank Eunice Wan, Simon Wong, and Pui Yan Kwok at the UCSF Institute of Human Genetics Core for genotyping support. This study makes use of data generated by the Wellcome Trust Case–Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113 and 085475. Genotype data for COG ALL cases are available for download from dbGaP (Study Accession: phs000638.v1.p1). Data came from a grant, the Resource for Genetic Epidemiology Research in Adult Health and Aging (RC2 AG033067; Schaefer and Risch, PIs) awarded to the Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH) and the UCSF Institute for Human Genetics. The RPGEH was supported by grants from the Robert Wood Johnson Foundation, the Wayne and Gladys Valley Foundation, the Ellison Medical Foundation, Kaiser Permanente Northern California, and the Kaiser Permanente National and Northern California Community Benefit Programs. The RPGEH and the Resource for Genetic Epidemiology Research in Adult Health and Aging are described here: https://divisionofresearch.kaiserpermanente.org/genetics/rpgeh/rpgehhome. For recruitment of subjects enrolled in the CCLS replication set, the authors gratefully acknowledge the clinical investigators at the following collaborating hospitals: University of California Davis Medical Center (Dr. Jonathan Ducore), University of California San Francisco (Drs. Mignon Loh and Katherine Matthay), Children's Hospital of Central California (Dr. Vonda Crouse), Lucile Packard Children's Hospital (Dr. Gary Dahl), Children's Hospital Oakland (Dr. James Feusner), Kaiser Permanente Roseville (formerly Sacramento) (Drs. Kent Jolly and Vincent Kiley), Kaiser

medRxiv preprint doi: https://doi.org/10.1101/2021.05.07.21256849; this version posted June 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Permanente Santa Clara (Drs. Carolyn Russo, Alan Wong, and Denah Taggart), Kaiser Permanente San Francisco (Dr. Kenneth Leung), and Kaiser Permanente Oakland (Drs. Daniel Kronish and Stacy Month). The authors additionally thank the families for their participation in the California Childhood Leukemia Study (formerly known as the Northern California Childhood Leukemia Study). Finally, the authors acknowledge the Center for Advanced Research Computing (CARC; https://carc.usc.edu) at the University of Southern California for providing computing resources that have contributed to the research results reported within this publication.

Authorship

Contribution: J.L.W., C.W.K.C., A.J.D. conceived and supervised this project; S.J., S.L., M.C. performed data analysis; N.M., I.S.M., L.M.M., A.T.D., C.M., X.M. provided resources; S.J., A.J.D., C.W.K.C., J.L.W. wrote the manuscript with input from all the coauthors.
Conflict-of-interest disclosure: The authors declare no competing interests.
Correspondence: Joseph Leo Wiemels, Center for Genetic Epidemiology, 1450 Biggy St, Los Angeles, California, email: wiemels@usc.edu; Charleston Chiang, 1450 Biggy St, Los Angeles, California, email: charleston.chiang@med.usc.edu

REFERENCES

1. Giddings BM, Whitehead TP, Metayer C, Miller MD. Childhood leukemia incidence in California: High and rising in the Hispanic population: Hispanic Childhood Leukemia Incidence. *Cancer*. 2016;122(18):2867-2875. doi:10.1002/cncr.30129

2. Lim JY-S, Bhatia S, Robison LL, Yang JJ. Genomics of racial and ethnic disparities in childhood acute lymphoblastic leukemia. *Cancer*. 2014;120(7):955-962. doi:10.1002/cncr.28531

3. Vijayakrishnan J, Kumar R, Henrion MYR, 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):573-579. doi:10.1038/leu.2016.271

4. Wiemels JL, Walsh KM, de Smith AJ, et al. GWAS in childhood acute lymphoblastic leukemia reveals novel genetic associations at chromosomes 17q12 and 8q24.21. *Nat Commun*. 2018;9(1):286. doi:10.1038/s41467-017-02596-9

5. Papaemmanuil E, Hosking FJ, Vijayakrishnan J, et al. Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. *Nat Genet*. 2009;41(9):1006-1010. doi:10.1038/ng.430

6. Perez-Andreu V, Roberts KG, Harvey RC, et al. Inherited GATA3 variants are associated with Ph-like childhood acute lymphoblastic leukemia and risk of relapse. *Nat Genet*. 2013;45(12):1494-1498. doi:10.1038/ng.2803

7. Treviño LR, Yang W, French D, et al. Germline genomic variants associated with childhood acute lymphoblastic leukemia. *Nat Genet*. 2009;41(9):1001-1005. doi:10.1038/ng.432

medRxiv preprint doi: https://doi.org/10.1101/2021.05.07.21256849; this version posted June 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

8. Xu H, Yang W, Perez-Andreu V, 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):733-742. doi:10.1093/jnci/djt042

9. Vijayakrishnan J, Qian M, Studd JB, et al. Identification of four novel associations for B-cell acute lymphoblastic leukaemia risk. *Nat Commun*. 2019;10(1):5348. doi:10.1038/s41467-019-13069-6

10. Feng Q, de Smith AJ, Vergara-Lluri M, et al. Trends in Acute Lymphoblastic Leukemia Incidence in the United States by Race/Ethnicity From 2000 to 2016. *Am J Epidemiol*. Published online October 9, 2020:kwaa215. doi:10.1093/aje/kwaa215

11. Linabery AM, Ross JA. Trends in childhood cancer incidence in the U.S. (1992-2004). *Cancer*. 2008;112(2):416-432. doi:10.1002/cncr.23169

12. Barrington-Trimis JL, Cockburn M, Metayer C, Gauderman WJ, Wiemels J, McKean-Cowdin R. Trends in childhood leukemia incidence over two decades from 1992 to 2013. *Int J Cancer*. 2017;140(5):1000-1008. doi:10.1002/ijc.30487

13. Bhatia S, Sather HN, Heerema NA, Trigg ME, Gaynon PS, Robison LL. Racial and ethnic differences in survival of children with acute lymphoblastic leukemia. *Blood*. 2002;100(6):1957-1964. doi:10.1182/blood-2002-02-0395

14. Kadan-Lottick NS. Survival Variability by Race and Ethnicity in Childhood Acute Lymphoblastic Leukemia. *JAMA*. 2003;290(15):2008. doi:10.1001/jama.290.15.2008

15. Popejoy AB, Fullerton SM. Genomics is failing on diversity. *Nature*. 2016;538(7624):161-164. doi:10.1038/538161a

16. Qian M, Xu H, Perez-Andreu V, et al. Novel susceptibility variants at the ERG locus for childhood acute lymphoblastic leukemia in Hispanics. *Blood*. 2019;133(7):724-729. doi:10.1182/blood-2018-07-862946

17. de Smith AJ, Walsh KM, Morimoto LM, 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):2746-2751. doi:10.1038/s41375-019-0514-9

18. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393

19. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet*. 2019;51(4):584-591. doi:10.1038/s41588-019-0379-x

20. Wojcik GL, Graff M, Nishimura KK, et al. Genetic analyses of diverse populations improves discovery for complex traits. *Nature*. 2019;570(7762):514-518. doi:10.1038/s41586-019-1310-4

21. Mahajan A, Spracklen CN, Zhang W, et al. *Trans-Ancestry Genetic Study of Type 2 Diabetes Highlights the Power of Diverse Populations for Discovery and Translation*. Genetic and Genomic Medicine; 2020. doi:10.1101/2020.09.22.20198937

medRxiv preprint doi: https://doi.org/10.1101/2021.05.07.21256849; this version posted June 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

22. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007;447(7145):661-678. doi:10.1038/nature05911

23. Metayer C, Zhang L, Wiemels JL, et al. Tobacco smoke exposure and the risk of childhood acute lymphoblastic and myeloid leukemias by cytogenetic subtype. *Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol*. 2013;22(9):1600-1611. doi:10.1158/1055-9965.EPI-13-0350

24. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*. 2007;39(7):906-913. doi:10.1038/ng2088

25. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*. 2010;26(17):2190-2191. doi:10.1093/bioinformatics/btq340

26. Walsh KM, de Smith AJ, Hansen HM, et al. A Heritable Missense Polymorphism in *CDKN2A* Confers Strong Risk of Childhood Acute Lymphoblastic Leukemia and Is Preferentially Selected during Clonal Evolution. *Cancer Res*. 2015;75(22):4884-4894. doi:10.1158/0008-5472.CAN-15-1105

27. Vijayakrishnan J, Henrion M, Moorman AV, et al. The 9p21.3 risk of childhood acute lymphoblastic leukaemia is explained by a rare high-impact variant in CDKN2A. *Sci Rep*. 2015;5(1):15065. doi:10.1038/srep15065

28. de Smith AJ, Walsh KM, Francis SS, et al. BMI1enhancer polymorphism underlies chromosome 10p12.31 association with childhood acute lymphoblastic leukemia: *BMI* 1 enhancer polymorphism in ALL. *Int J Cancer*. 2018;143(11):2647-2658. doi:10.1002/ijc.31622

29. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12(1):77. doi:10.1186/1471-2105-12-77

30. Genome Aggregation Database Consortium, Karczewski KJ, Francioli LC, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581(7809):434-443. doi:10.1038/s41586-020-2308-7

31. Wiemels JL, de Smith AJ, Xiao J, et al. A functional polymorphism in the CEBPE gene promoter influences acute lymphoblastic leukemia risk through interaction with the hematopoietic transcription factor Ikaros. *Leukemia*. 2016;30(5):1194-1197. doi:10.1038/leu.2015.251

32. Studd JB, Yang M, Li Z, et al. Genetic predisposition to B-cell acute lymphoblastic leukemia at 14q11.2 is mediated by a CEBPE promoter polymorphism. *Leukemia*. 2019;33(1):1-14. doi:10.1038/s41375-018-0184-z

33. The LifeLines Cohort Study, Yang J, Bakshi A, et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet*. 2015;47(10):1114-1120. doi:10.1038/ng.3390

medRxiv preprint doi: https://doi.org/10.1101/2021.05.07.21256849; this version posted June 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

34. Zaitlen N, Pasaniuc B, Sankararaman S, et al. Leveraging population admixture to characterize the heritability of complex traits. *Nat Genet*. 2014;46(12):1356-1362. doi:10.1038/ng.3139

35. Shi H, Burch KS, Johnson R, et al. Localizing Components of Shared Transethnic Genetic Architecture of Complex Traits from GWAS Summary Data. *Am J Hum Genet*. 2020;106(6):805-817. doi:10.1016/j.ajhg.2020.04.012

36. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res*. 2012;40(Database issue):D930-934. doi:10.1093/nar/gkr917

37. Lonsdale J, Thomas J, Salvatore M, et al. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013;45(6):580-585. doi:10.1038/ng.2653

38. Lin BD, Carnero-Montoro E, Bell JT, et al. 2SNP heritability and effects of genetic variants for neutrophil-to-lymphocyte and platelet-to-lymphocyte ratio. *J Hum Genet*. 2017;62(11):979-988. doi:10.1038/jhg.2017.76

39. Stadhouders R, Aktuna S, Thongjuea S, et al. HBS1L-MYB intergenic variants modulate fetal hemoglobin via long-range MYB enhancers. *J Clin Invest*. 2014;124(4):1699-1710. doi:10.1172/JCI71520

40. Guo MH, Nandakumar SK, Ulirsch JC, et al. Comprehensive population-based genome sequencing provides insight into hematopoietic regulatory mechanisms. *Proc Natl Acad Sci*. 2017;114(3):E327-E336. doi:10.1073/pnas.1619052114

41. Li M, Jiang P, Cheng K, et al. Regulation of MYB by distal enhancer elements in human myeloid leukemia. *Cell Death Dis*. 2021;12(2):223. doi:10.1038/s41419-021-03515-z

42. Astle WJ, Elding H, Jiang T, et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell*. 2016;167(5):1415-1429.e19. doi:10.1016/j.cell.2016.10.042

43. van Rooij FJA, Qayyum R, Smith AV, et al. Genome-wide Trans-ethnic Meta-analysis Identifies Seven Genetic Loci Influencing Erythrocyte Traits and a Role for RBPMS in Erythropoiesis. *Am J Hum Genet*. 2017;100(1):51-63. doi:10.1016/j.ajhg.2016.11.016

44. Tajuddin SM, Schick UM, Eicher JD, et al. Large-Scale Exome-wide Association Analysis Identifies Loci for White Blood Cell Traits and Pleiotropy with Immune-Mediated Diseases. *Am J Hum Genet*. 2016;99(1):22-39. doi:10.1016/j.ajhg.2016.05.003

45. Buniello A, MacArthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. 2019;47(D1):D1005-D1012. doi:10.1093/nar/gky1120

46. Ahola-Olli AV, Würtz P, Havulinna AS, et al. Genome-wide Association Study Identifies 27 Loci Influencing Concentrations of Circulating Cytokines and Growth Factors. *Am J Hum Genet*. 2017;100(1):40-50. doi:10.1016/j.ajhg.2016.11.007

medRxiv preprint doi: https://doi.org/10.1101/2021.05.07.21256849; this version posted June 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

47. Chang JS, Zhou M, Buffler PA, Chokkalingam AP, Metayer C, Wiemels JL. Profound deficit of IL10 at birth in children who develop childhood acute lymphoblastic leukemia. *Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol*. 2011;20(8):1736-1740. doi:10.1158/1055-9965.EPI-11-0162

48. Lynch JR, Salik B, Connerty P, et al. JMJD1C-mediated metabolic dysregulation contributes to HOXA9-dependent leukemogenesis. *Leukemia*. 2019;33(6):1400-1410. doi:10.1038/s41375-018-0354-z

49. Chen M, Zhu N, Liu X, et al. JMJD1C is required for the survival of acute myeloid leukemia by functioning as a coactivator for key transcription factors. *Genes Dev*. 2015;29(20):2123-2139. doi:10.1101/gad.267278.115

50. Xiao F, Liao B, Hu J, et al. JMJD1C Ensures Mouse Embryonic Stem Cell Self-Renewal and Somatic Cell Reprogramming through Controlling MicroRNA Expression. *Stem Cell Rep*. 2017;9(3):927-942. doi:10.1016/j.stemcr.2017.07.013

51. Cimmino L, Dawlaty MM, Ndiaye-Lobry D, et al. TET1 is a tumor suppressor of hematopoietic malignancy. *Nat Immunol*. 2015;16(6):653-662. doi:10.1038/ni.3148

52. Bamezai S, Demir D, Pulikkottil AJ, et al. TET1 promotes growth of T-cell acute lymphoblastic leukemia and can be antagonized via PARP inhibition. *Leukemia*. Published online May 15, 2020. doi:10.1038/s41375-020-0864-3

53. Lappalainen T, Sammeth M, Friedländer MR, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*. 2013;501(7468):506-511. doi:10.1038/nature12531

54. Weissbrod O, Flint J, Rosset S. Estimating SNP-Based Heritability and Genetic Correlation in Case-Control Studies Directly and with Summary Statistics. *Am J Hum Genet*. 2018;103(1):89-99. doi:10.1016/j.ajhg.2018.06.002

55. Golan D, Lander ES, Rosset S. Measuring missing heritability: Inferring the contribution of common variants. *Proc Natl Acad Sci*. 2014;111(49):E5272-E5281. doi:10.1073/pnas.1419064111

56. Steinsaltz D, Dahl A, Wachter KW. On Negative Heritability and Negative Estimates of Heritability. *Genetics*. 2020;215(2):343-357. doi:10.1534/genetics.120.303161

57. the PRACTICAL consortium, Mancuso N, Rohland N, et al. The contribution of rare variation to prostate cancer heritability. *Nat Genet*. 2016;48(1):30-35. doi:10.1038/ng.3446

medRxiv preprint doi: https://doi.org/10.1101/2021.05.07.21256849; this version posted June 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Table 1. Summary statistics for the reported variants, the top variant in the loci from our meta-analysis, and the linkage disequilibrium between the two variants in NLW and LAT.

	Reported SNP				Top SNP in this study				r^2	
Gene	Chr	Pos	rsID (reference)	P-value	Chr	Pos	rsID	P-value	NLW	LAT
C5orf56	5	131765206	rs886285 (ref [9])	0.63	5	131811182	rs11741255	1.69×10^{-4}	0.35	0.19
BAK1	6	33546930	rs210143 (ref [9])	4.49×10^{-8}	6	33546837	rs210142	4.27×10^{-8}	1	1
IKZF1	7	50470604	rs4132601 (ref [5])	1.13×10^{-33}	7	50477144	rs10230978	3.92×10^{-34}	0.98	0.97
8q24	8	130156143	rs4617118 (ref [4])	1.04×10^{-12}	Same					
CDKN2A	9	21970916	rs3731249 (ref [27])	1.29×10^{-18}	9	21975319	rs36228834	1.90×10^{-18}	0.99	1
TLE1	9	83747371	rs76925697 (ref [9])	5.37×10^{-2}	9	83728588	rs62579826	1.06×10^{-2}	0.81	0.98
GATA3	10	8104208	rs3824662 (ref [6])	4.24×10^{-9}	Same					
PIP4K2A	10	22852948	rs7088318 (ref [8])	6.50×10^{-19}	10	22853102	rs7075634	2.42×10^{-19}	0.96	0.97
BMI1	10	22423302	rs11591377 (ref [28])	8.21×10^{-10}	10	22374489	rs1926697	5.24×10^{-10}	0.84	0.88
ARID5B	10	63723577	rs10821936 (ref [7])	4.78×10^{-67}	10	63721176	rs7090445	7.36×10^{-70}	0.98	0.99
LHPP	10	126293309	rs35837782 (ref [3])	6.90×10^{-4}	Same					
ELK3	12	96612762	rs4762284 (ref [3])	2.42×10^{-3}	12	96645605	rs78405390	4.68×10^{-5}	0.13	0.22
CEBPE	14	23589057	rs2239633 (ref [5])	3.0×10^{-14}	14	23589349	rs2239630	2.12×10^{-21}	0.74	0.78
IKZF3	17	38066240	rs2290400 (ref [4])	$2.09\ 10^{-6}$	17	37957235	rs17607816	1.42×10^{-7}	0.02	0.22
IGF2BP1	17	47092076	rs10853104 (ref [9])	2.93×10^{-2}	17	47217004	rs6504598	4.87×10^{-4}	0.02	0.02
ERG	21	39789606	rs8131436 (ref [17])	6.97×10^{-5}	21	39784752	rs55681902	9.36×10^{-6}	0.62	0.65

We focused on the variants within 1Mb of the previously reported susceptibility variants[3–8,26–28] and reported the association results of the published lead SNP as well as the top SNP at each locus from our meta-analysis. Note that out of the 16 loci, three (8q24.21, *IKZF3*, and *BMI1*) were initially identified and five (*IKZF1*, *PIP4K2A*, *ARID5B*, *CDKN2A*, *CEBPE*) were previously shown to be replicated using a smaller but largely overlapping subset of this dataset[4,28]. For these loci, our findings here would not necessarily constitute an independent replication. Gene names (gene) are given based on the nearest gene unless the variant is in gene desert. Chromosome (Chr) and position (Pos) are given in hg19 coordinates. r2 denotes the squared correlation of the reported SNP and our top SNP in NLW from discovery cohort; NLW and LAT denote the non-Latino white and Latino cohorts, respectively.

medRxiv preprint doi: https://doi.org/10.1101/2021.05.07.21256849; this version posted June 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Table 2. Summary of conditional analysis to identify secondary associations at known loci.

Gene	Chr	Pos	rsID	Risk allele	OR	$P_{conditional}$	$P_{discovery}$	r2
IKZF1	7	50459043	rs78396808	A	1.632	$3.46x10^{-26}$	$2.7x10^{-16}$	*0.06
CDKN2A/B	9	21993964	rs2811711	T	1.355	7.2x10-^{10}	$1.85x10^{-11}$	0.01
CEBPE	14	23592617	rs60820638	A	1.193	$5.38x10^{-8}$	0.102	0.16
IZKF3	17	37983492	rs12944882	T	1.204	$7.71x10^{-10}$	$2.81x10^{-7}$	0.02

For each of the four significant association after conditional analysis, we show the genomic coordinates in hg19, effect size (OR), the P-values with or without conditioning on the lead SNP from the discovery meta-analysis in the locus, and the r2 between the lead SNP and secondary association.

Chr., chromosome; Pos;. Position in hg19; OR: Effect size; $P_{conditional}$: p-value from the conditional analysis; $P_{discovery}$: p –value from meta-analysis without conditioning on any SNP; r2: squared correlation of the conditioned SNP and the most significantly associated SNP from conditional analysis.

*calculated in Latino population as the variant was filtered out for low MAF in NLW cohort.

Figure1. Summary result of the trans-ethnic meta-analysis on ALL.
Results of the meta-analysis is represented by the Manhattan plot. The novel loci from this study are marked with dotted lines and labeled with the nearest genes. Significance threshold at genome-wide significance level ($5x10^{-8}$) is marked with a horizontal dashed grey line in the Manhattan plot. The y-axis is truncated at $-\log10(1x10^{-50})$ to improve readability. The insert shows the Quantile-Quantile plot. Deviation from the expected p-value distribution is evident only in the tail. There is little evidence of inflation of the test statistics in general as the genomic inflation factor is 1.024.

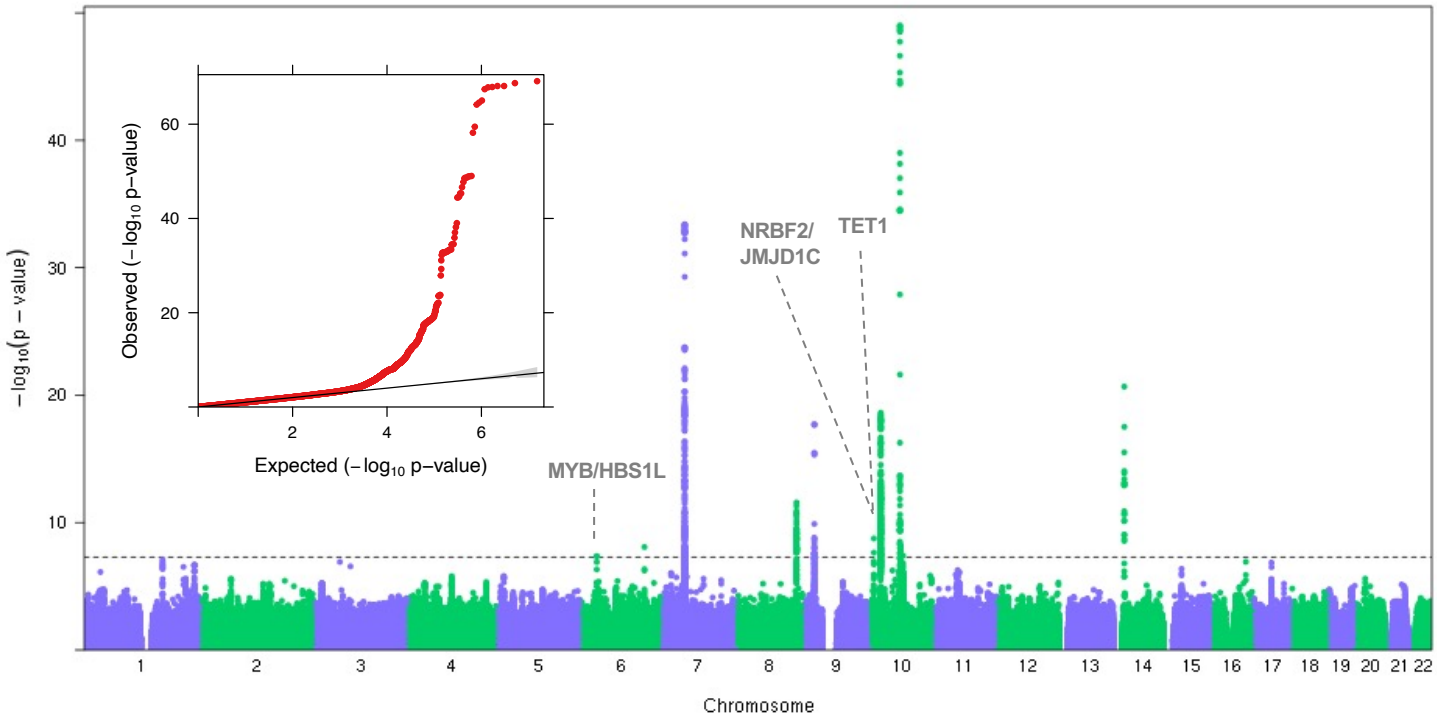


Figure2. Novel loci associated with childhood ALL in trans-ethnic meta-analysis. LocusZoom plots showing 1 Mb region around the identified loci near (A) MYB/HBS1L on chr6, (B) NRBF2/JMJD1C on chr10, and (C) TET1 on chr10 are shown. Diamond symbol indicates the lead SNP in each locus. Color of remaining SNPs is based on linkage disequilibrium (LD) as measured by r2 with the lead SNP in non-Latino white. All coordinates in x-axis are in hg19.

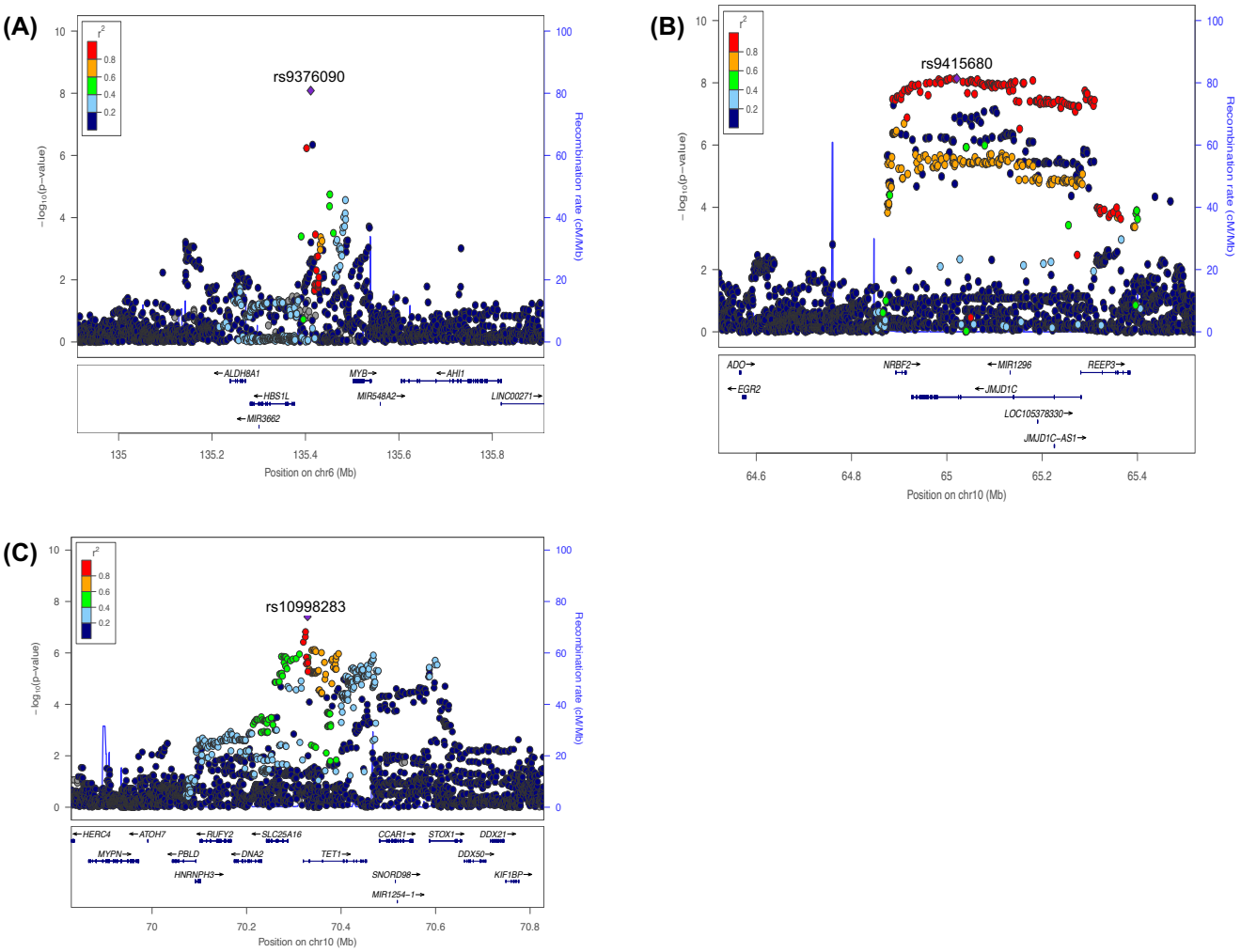
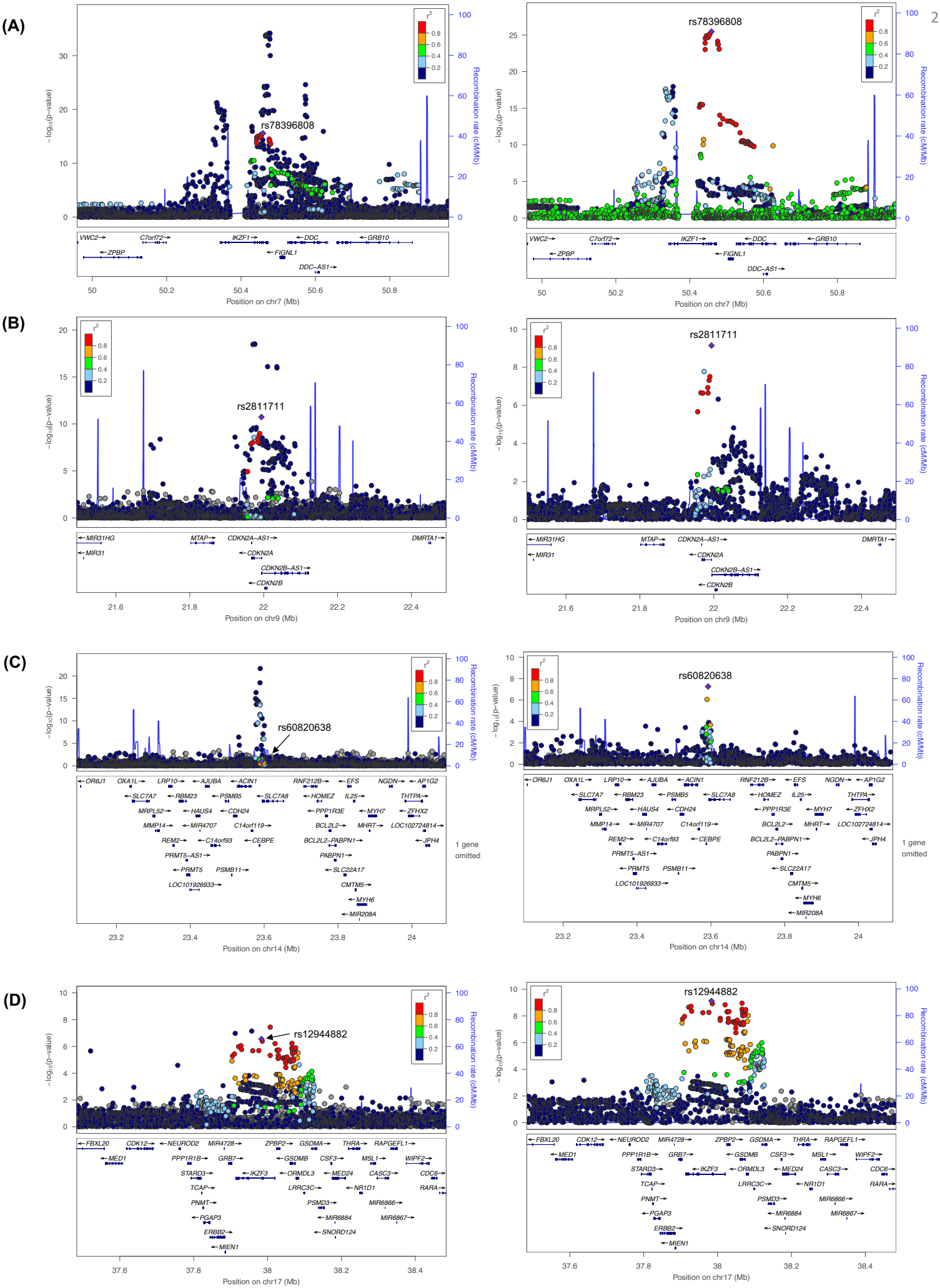


Figure 3. Secondary association signal ($p < 5x10^{-8}$) with ALL found in previously known loci through conditional analysis. LocusZoom plot displaying the 1 Mb region found to harbor a second novel variant associated with ALL through conditional analysis: (A) IKZF1 (B) CDKN2A (C) CEBPE (D) IKZF3. For each locus, we display the pattern of association before(left) and after(right) conditioning on the top associated variant in the locus. In both cases, diamond indicates the lead SNP in the conditional analysis. Color of the remaining SNPs is based on linkage disequilibrium (LD) with the lead variant in the conditional analysis in non-Latino white. Genomic coordinates on x-axis are in hg19.


(A)
rs78396808
rs78396808
(B)
rs2811711
rs2811711
(C)
rs60820638
rs60820638
1 gene omitted
1 gene omitted
(D)
rs12944882
rs12944882
−log10(p−value)
Recombination rate (cM/Mb)
r2
Position on chr7 (Mb)
Position on chr9 (Mb)
Position on chr14 (Mb)
Position on chr17 (Mb)
VWC2
C7orf72
IKZF1
DDC
GRB10
ZPBP
FIGNL1
DDC−AS1
MIR31HG
MTAP
CDKN2A−AS1
DMRTA1
MIR31
CDKN2A
CDKN2B−AS1
CDKN2B
OR6J1
OXA1L
LRP10
AJUBA
ACIN1
RNF212B
EFS
NGDN
AP1G2
SLC7A7
RBM23
PSMB5
SLC7A8
HOMEZ
IL25
THTPA
MRPL52
HAUS4
CDH24
PPP1R3E
MYH7
ZFHX2
MMP14
MIR4707
C14orf119
BCL2L2
MHRT
LOC102724814
REM2
C14orf93
CEBPE
BCL2L2−PABPN1
JPH4
PRMT5−AS1
PABPN1
PRMT5
PSMB11
SLC22A17
LOC101926933
CMTM5
MYH6
MIR208A
FBXL20
CDK12
NEUROD2
MIR4728
ZPBP2
GSDMA
THRA
RAPGEFL1
MED1
PPP1R1B
GRB7
GSDMB
CSF3
MSL1
WIPF2
STARD3
IKZF3
ORMDL3
MED24
CASC3
CDC6
TCAP
LRRC3C
NR1D1
RARA
PNMT
PSMD3
MIR6866
PGAP3
MIR6884
MIR6867
ERBB2
SNORD124
MIEN1

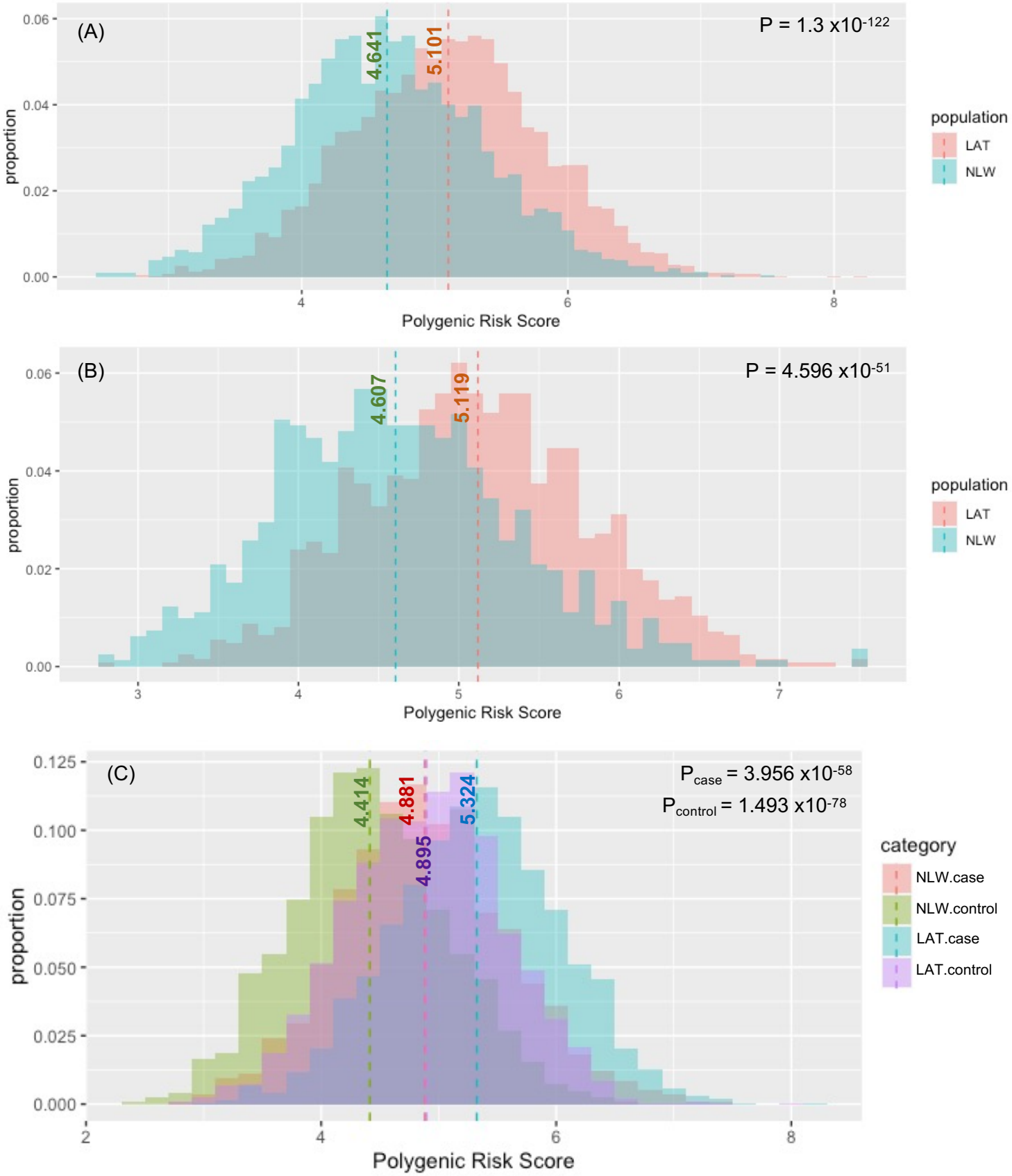


Figure 4. Polygenic Risk Score (PRS) distribution based on GWAS loci for ALL.
We compared the PRS distribution between LAT and NLW cohorts in (A) CCRLP and (B) CCLS cohorts. PRS were constructed by summing up imputed dosage weighted by effect size for each Latino(red) and non-Latino white individual(green). In (C) We further stratified the PRS in CCRLP cohort by case/control status. The population mean is indicated with vertical dash lines with the mean score shown. P-values on the right upper corner of each graph is from one-sided t-test comparing the difference in PRS between LAT and NLW overall or within cases and controls.

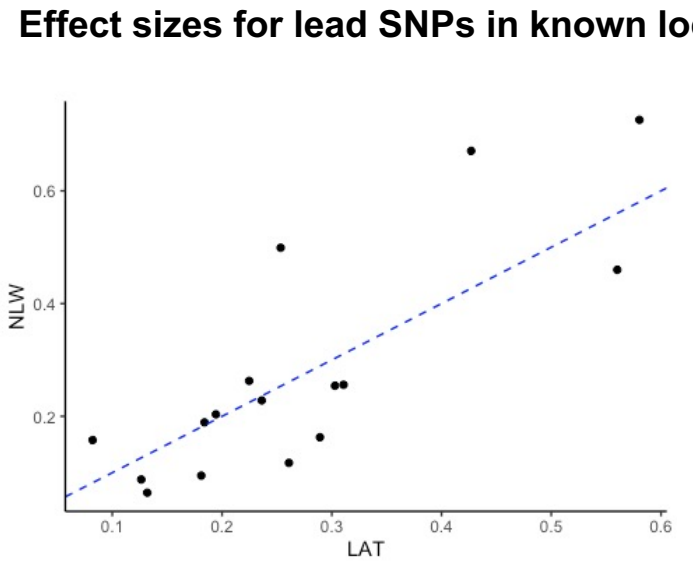


Figure S1. **Effect sizes for lead SNPs in each of the 16 known loci**. The effect size estimates(Beta) from GWAS in only the LAT or NLW subset from our discovery cohort are shown. The correlation coefficient (r) is 0.819. The dashed blue line is y=x.

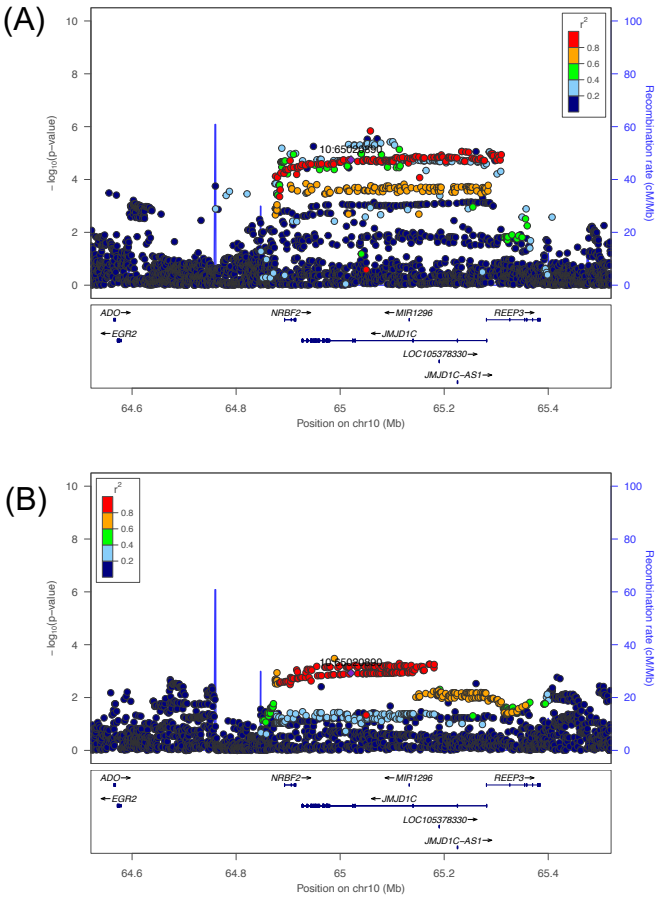


Figure S2. Association signal around NRBF2/JMJD1C locus on chr10 in LAT and EAS cohorts.

LocusZoom plots show distinct haplotypes showing association with ALL in (A) LAT and (B) EAS cohorts in our study. Diamond symbol indicates the lead SNP in each cohort. Color of remaining SNPs is based on linkage disequilibrium (LD) as measured by r2 with the lead SNP in the respective cohort. All coordinates in x-axis are in hg19.

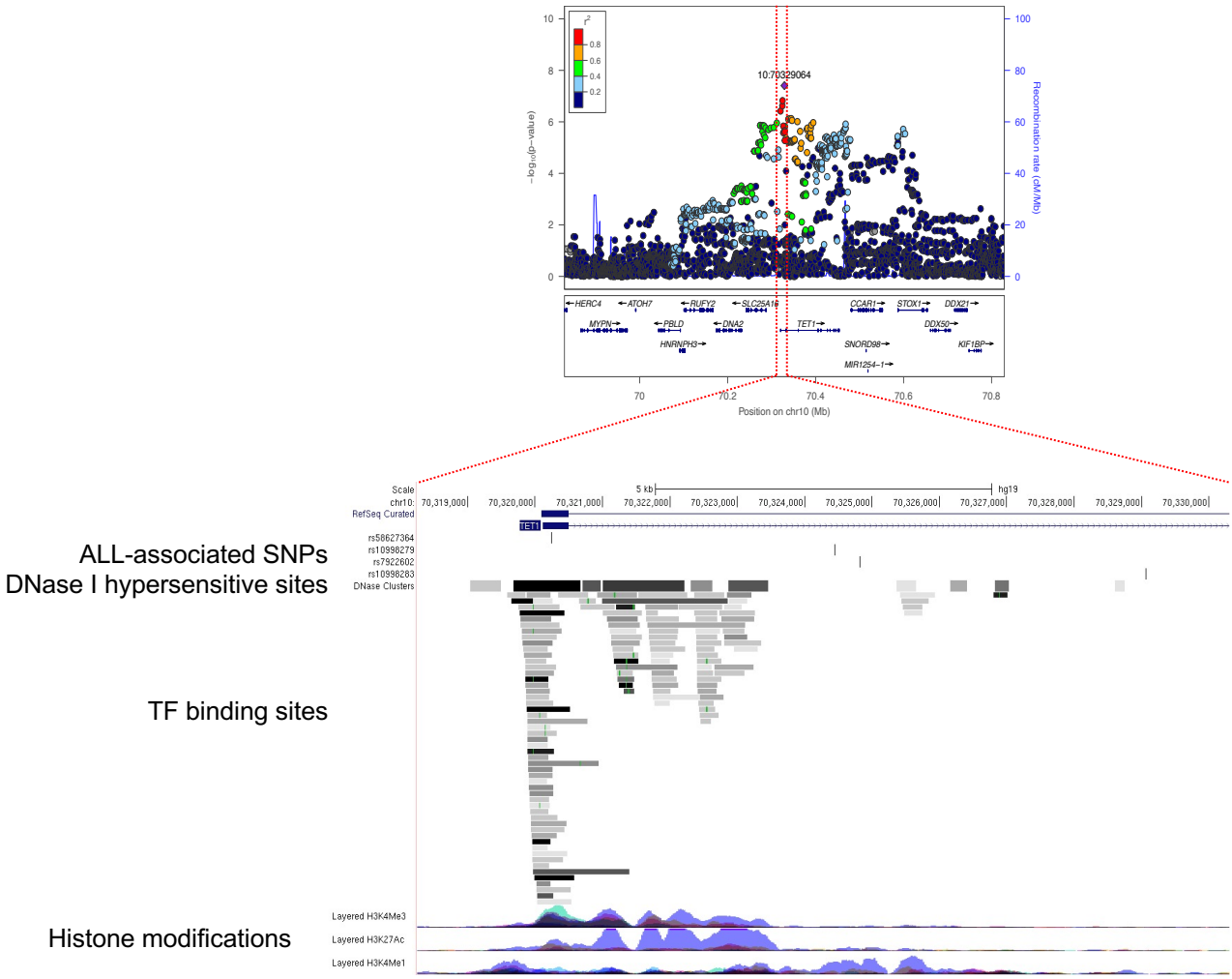


Figure S3. Functional annotation of the TET1 locus.
For the immediately nearby location around the top associated SNPs in our meta-analysis(blue vertical lines), we extracted the functionally annotated genomic/epigenomic features from multiple cell types in ENCODE data. Functional data were retrieved from UCSC genome browser.