Supplementary Information

Potential role in hematopoiesis, leukemogenesis, and ALL etiology for putatively novel loci. We annotated each of the three putatively associated novel loci to ALL using HaploReg (version 4.1(1) and GTEx portal(2). Chromatin interactions in the genomic neighborhood of the new top hits were evaluated using the program Juicebox(3) and high resolution Hi-C data from the B-lymphoblastoid cell line GM12878(4). We summarize our findings below. 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(5) . This locus is already well known for associations with multiple blood cell measurements, severity of major hemoglobin disorders, and β-thalassemia(6,7). 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(6,8). 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(5) and hemoglobin concentration(9,10). Also, it is an eQTL in lymphocytes and whole blood(2) 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 1B). 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 (Table S4, Supplementary Figure A).

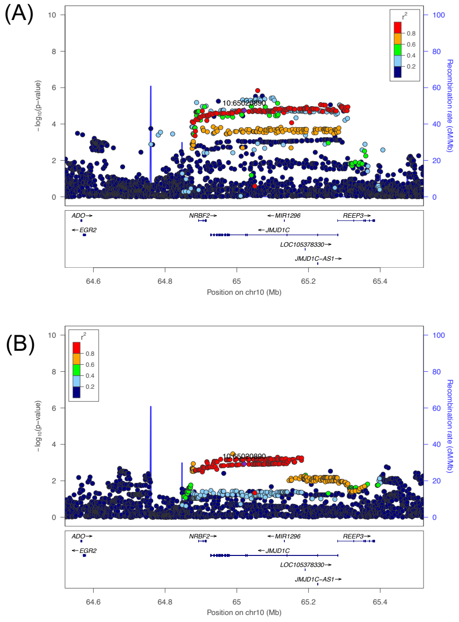

Figure A. **Association signal around NRBF2/JMJD1C locus on chr10 in LAT and EAS cohorts.**
LocusZoom plots show distinct haplotypes showing association with ALL in (A) LAT and (B) EAS cohorts in our study. Diamond symbol indicates the lead SNP in each cohort. Color of remaining SNPs is based on linkage disequilibrium (LD) as measured by r2 with the lead SNP in the respective cohort. All coordinates in x-axis are in hg19.

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(11,12). It is also associated with IL-10 levels(13) which was shown to be in deficit in ALL cases(14). 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(15). Previous studies have found that it acts as a coactivator for key transcription factors to ensure survival of AML cells(16) and self-renewal of mouse embryonic stem cells(17). We note that the association at this locus did not replicate in COG and CCLS replication cohort and should be assessed in additional datasets.

The second locus in 10q21 contains intronic variants in the *TET1* gene, which is well known for its oncogenicity in several malignancies including AML(18). 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(19). 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 (Table S5). 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 B).

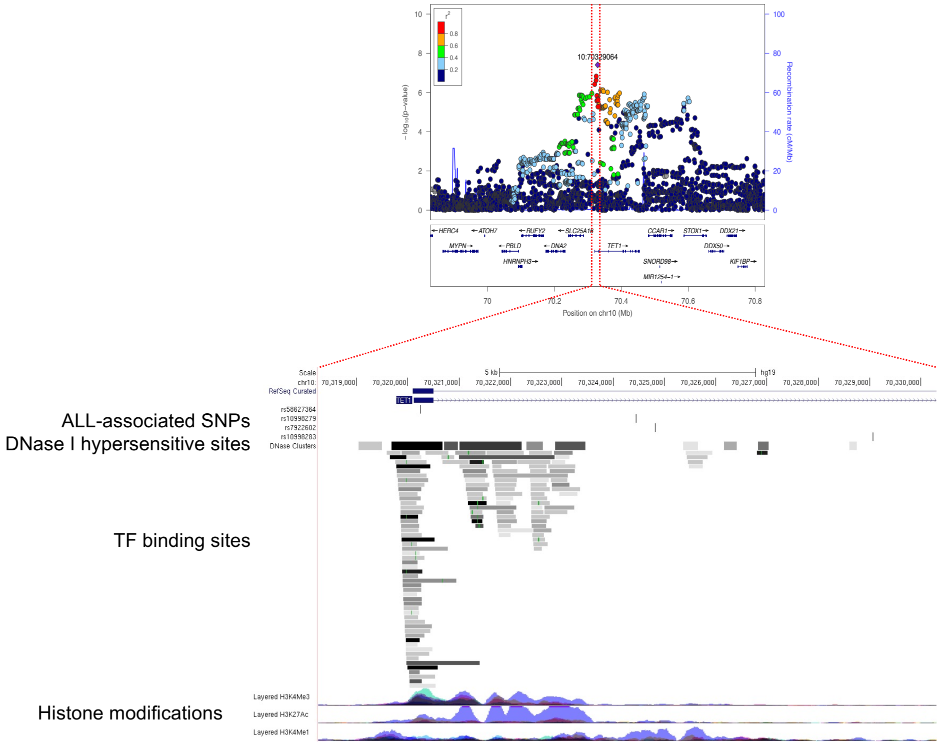

Figure B. Functional annotation of the TET1 locus.
For the immediately nearby location around the top associated SNPs in our meta-analysis(blue vertical lines), we extracted the functionally annotated genomic/epigenomic features from multiple cell types in ENCODE data. Functional data were retrieved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

However, none of these SNPs were observed as eQTL for *TET1* in whole blood or lymphoblastoid cells(20); future studies may want to investigate whether these SNPs affect *TET1* expression in hematopoietic stem or progenitor cells.

Supplementary Methods

Study Cohorts

In our efforts to be sensitive to the genomic history of human populations, 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 self-reported ethnic group in our study: African American (AFR), East Asian (EAS), Latino American (LAT), and non-Latino white (NLW). As a baseline for the current analysis, we supplemented the previous trans-ethnic analysis by including additional controls for NLW and added the EAS cohort(21). 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. Our age cutoff of 15 was designed to capture most childhood ALL cases, whose incidence peaks between ages 2-7, and to limit dilution of genetic effects of different subtypes that affect older teenagers. Also, we note that specific cytogenetic subtyping is not available for our entire discovery cohort since our diagnostic catchment extends back to 1988. 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(21). Because ALL is a rare childhood cancer, for the purpose of a genetic study we followed previous practice(21) 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: (i) 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 (ii) 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.

Data Processing and Quality Control

The quality control (QC) on single nucleotide polymorphism (SNP) array genotypes and samples were carried out in each population and dataset in parallel, performed in two stages: pre-imputation and post-imputation. In pre-imputation QC, the sex chromosomes were excluded, and SNPs were filtered out on the basis of call rate (<98%), minor allele frequency (MAF<0.01), genome-wide relatedness (PI_HAT>0.02), genome heterozygosity rate (mean heterozygosity± 6Std), and deviation from Hardy-Weinberg equilibrium in controls ($P<10^{-5}$). Samples with call rate < 95% were also removed. In general, individuals were included in each of the four ethnic groups based on their self-reported ethnicity. We did not attempt to reassign individuals to different ethnic groups based on estimated genetic ancestry. We performed principal components analysis for each population including our study subjects along with 1000 Genomes Project reference data (24) to identify extreme outlier individuals that clustered apart from other individuals in their self-reported race/ethnicity groups: this identified a small subset of individuals (29 cases and 51 controls from

CCRLP, 31 individuals from GERA) among self-reported Asians (total n=722) that clustered with South Asian reference individuals, as well as 5 self-reported African American individuals out of n=3572 from the GERA cohort that clustered with East Asian reference individuals. These appear to result from a lack of finer-scale ethnic labels for self-report, or a mis-labeling of individual records, and these individuals were dropped from our analysis.

To control for potential batch effect and systematic bias between array types, we further performed two separate GWASs for quality control purposes. First, stratified by ethnicity and restricted to SNPs passing the filters described above in both CCRLP and GERA, we compared CCRLP controls and GERA individuals. Second, using the GERA NLW cohort we compared individuals that were genotyped on the Axiom type "A" to those genotyped on the type "O" reagent kit (NLW was the only cohort in GERA that was genotyped using both reagent kits). Twenty principal components (PCs) were included as covariates of the logistic regression. In both comparisons we observed inflation of the test statistics suggesting a subset of SNPs exhibited evidence of batch effect, thus we removed variants with $P < 0.01$ in any of the comparisons from all populations.

We then performed genome-wide imputation with the overlapping set of remaining SNPs (N = 431,543 in AFR, 259,468 in EAS, 547,575 in LAT and 362,977 in NLW) in each dataset using Haplotype Reference Consortium (HRC v r1.1 2016) as a reference in the Michigan Imputation Server(25). The different number of SNPs passing QC and used in imputation reflects the fact that each ethnic group in CCRLP was genotyped using Affymetrix World arrays optimized for the Latino population (i.e., Axiom LAT array). In post-imputation QC, we filtered variants in each ethnic group by imputation quality ($R^2 < 0.3$), MAF (< 0.01), and allele frequency difference between non-Finnish Europeans in the Genome Aggregation Database (gnomAD)(26) and CCRLP NLW controls (> 0.1). We next performed another GWAS between CCRLP controls and GERA individuals to conservatively protect against between-cohort batch effects after imputation, and removed variants with $P < 1x10^{-5}$. In principal components analysis using imputed data, we identified and removed 31 individuals in GERA LAT that were extreme outliers after imputation in PCs 1 to 20. Stratified by ethnicity, the CCRLP and GERA datasets were then merged to perform GWAS of ALL. In total, 124, 318, 1878, 1162 cases and 2067, 5017, 8410, 57341 controls, in AFR, EAS, LAT and NLW, respectively were used in GWAS for ALL. A total of 7,628,894 SNPs that remained in at least three ethnic groups were tested in our GWAS discovery analysis.

For replication cohorts, we generally followed the same quality control pipeline. For COG and WTCCC, because self-identified ethnicity was not available to us, we performed global ancestry estimations using ADMIXTURE and the 1000 Genomes populations as reference and removed individuals with < 90% estimated European ancestry from the analysis. This resulted in a total of 1504 and 2931 NLW cases and controls, respectively, from COG/WTCCC, and 472 NLW cases, 340 NLW controls, 750 LAT cases and 504 LAT controls, from CCLS.

Association Testing

We used SNPTEST(27) (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(28). Only variants passing QC in at least three of the four ethnic groups were meta-analyzed. A genome-wide threshold of 5 x 10^{-8} was used for significance in the discovery stage. As a convention, we referred to the locus by the names of the closest genes from the index variant; we acknowledge that these genes may not be the causal gene. A Bonferroni-corrected significance of 0.00312 (=0.05/16) was used for replication of previously reported susceptibility variants(21,29–36). Cochran's Q-test for heterogeneity was performed using METAL(28). To perform conditional analysis in identifying secondary associations within a locus, the lead SNP was additionally included in the regression model, again using 5 x 10^{-8} as threshold for significance.

Familial risk per variant

The percentage of familial relative risk (FRR) explained by each genetic variant was calculated as per Schumacher et al(37)4/14/22 9:47:00 AM . The familial relative risk due to locus k (λ_k) is given by

$$\lambda_k = \frac{p_k r_k^2 + q_k}{(p_k r_k + q_k)^2}$$

where p_k is the frequency of the risk allele for locus k in each population, $q_k = 1 - p_k$, and r_k is the estimated per-allele odds ratio from meta-analysis. The percentage of familial relative risk is calculated as $\sum_k \log \lambda_k / \log \lambda_0$ where λ_0 is the observed familial risk to first-degree relatives of ALL cases, assumed to be 3.2 as per Kharazmi et al. (38).

Heritability Estimates

We estimated heritability ascribable to all post-QC imputed SNPs with MAF ≥ 0.05 in our GWAS data using the genome-wide complex trait analysis software (GCTA)(39). We followed the GCTA-LDMS approach to estimate heritability from imputed data(40), which recommended stratifying

SNPs into bins based on their LD scores and/or minor allele frequency. Using GCTA, we computed the genetic relationship matrix (GRM) of pairs of samples using SNPs in each bin, and used the multiple GRMs as input to obtain a restricted maximum likelihood (REML) estimate of heritability. All individuals in discovery analysis were used for LAT (n=10,288). For computational efficiency and for maintaining a close balance in sample size to the LAT data, we randomly sampled 10,000 NLW GERA controls to be included with all of CCRLP NLW cases and controls (total N = 12,391). We used a prevalence of 4.41x10^{-4} and 4.09x10^{-4} for childhood ALL in LAT and NLW respectively based on data from the Surveillance Research Program, (National Cancer Institute SEER*Stat software version 8.3.8; https://seer.cancer.gov/seerstat) to convert the estimated heritability to the liability scale. Because the NLW are expected to be much better imputed using HRC than LAT, particularly at rare variants, our genome-wide imputed data potentially could be used to partition the contribution of low frequency ($0.01 \leq MAF < 0.05$) and common ($MAF \geq 0.05$) variants in NLW population. In this case, we performed GCTA-LDMS analysis in 8 strata: two MAF strata (low frequency and common) by four quartiles of LD score strata. We also used these same GRMs to estimate heritability using phenotype-correlation-genotype-correlation (PCGC) regression as implemented in LDAKv5.1(41,42).

We further applied an approach to estimate heritability in LAT population using local ancestry(43). In brief, we first estimated the local ancestry in LAT using RFMix(44), using the combined 1000 Genomes Project and Human Genome Diversity Project(26) as ancestry references. Specifically, we used AFR (excluding ACB and ASW individuals; n=716), self-reported Non-Finnish European (NFE; n=617), and subjects having > 85% global AMR ancestry(based on ADMIXTURE(45); n=94) as the reference for African, European, and Native American ancestries. We used the local ancestry to estimate the genetic similarity and the heritability explained by local ancestry $h^2{}_\gamma$, calculated the genetic distance F_{STC} between the ancestral populations, and the mean admixture proportion θ. Following Zaitlen et al(43), the heritability is then calculated using formula $h^2{}_\gamma = 2\,\theta(1-\theta)\,h^2F_{STC}$. Because the original approach is only applicable to two-way admixed populations, we assigned the ancestry call as missing if the most likely local ancestry call for a locus is AFR. This sets approximately 5% of the genome as missing. The number of copies of local ancestry are standardized to have zero mean and unit variance to compute the genetic similarity matrix. The heritability explained by local ancestry *h2* was estimated in GCTA, with the global AMR ancestry normalized by non-African ancestry as a quantitative covariate. The genetic distance F_{STC} between the ancestral populations was computed based on the allele frequencies in the reference population as $\frac{(f_{AMR}-f_{NFE})^2}{2f\,(1-f)}$ where f_{AMR} and f_{NFE} are allele frequencies in AMR and NFE reference panel and the expected frequency in the admixed population, f, is the average of ancestral frequencies weighted by the average normalized global ancestries. The genome wide F_{STC} is the average value across 417,635 sites where the minor allele frequencies are greater than 0.05 in both ancestral populations.

To measure genetic correlation between LAT and NLW, we used SNPs with MAF ≥ 0.05 in both populations to generate GRM using R as per Mancuso et al (46). The individuals used in univariate REML for each ethnicity were used for the bivariate analysis (n=22,679). We used imputed dosage data to estimate GRM for each unique pair of ancestry groups as

$$A = \frac{1}{m}\begin{bmatrix} Z_1Z_1^t & Z_1Z_2^t \\ Z_2Z_1^t & Z_2Z_2^t \end{bmatrix}$$

where m is number of SNPs and Z_1 and Z_2 are the standardized genotype matrices for LAT and NLW, respectively. We estimated genetic correlation using bivariate GREML in GCTA(39).

Investigation of Genetic Architecture

To quantify the extent to which latent causal variants for ALL are shared or population-specific between LAT and NLW, we analyzed our GWAS summary data using the tool PESCA(47). Briefly, PESCA analyzes GWAS summary data from multiple populations jointly to infer the genome-wide proportion of causal variants that are population-specific or population-shared. For computational efficiency, PESCA requires first defining LD blocks that are approximately independent in both populations and assumes that a SNP in a given block is independent from all SNPs in all other blocks. We computed pairwise LD matrix in both NLW and LAT using ~329K directly genotyped SNPs shared in both populations. Then, following Shi et al.(47), we generated the trans-ethnic LD matrix by using the larger r^2 value of the NLW or LAT-specific pairwise LD, and used LDetect (48) to define LD blocks within the transethnic LD matrix. By setting mean LD block size to 200 SNPs and using default parameters, we obtained 1,653 blocks that are approximately independent, which is approximately similar to previous reports in East Asians and Europeans(47). We then followed Shi et al. to estimate the numbers of population-specific and shared causal SNPs using PESCA(47). We restricted our analysis to 1.3M SNPs with MAF > 0.05, $r^2 < 0.95$, and with summary association statistics available in both NLW and LAT. We first estimated the genome-wide proportion of population-specific and shared causal variants with the heritability estimated above (0.2033 and 0.0413 in NLW and LAT, respectively) using default parameters in PESCA, parallelizing the analysis in groups of 10 LD blocks at a time. Using the estimated genome-wide proportions of population-specific and shared causal variants as prior probabilities, we then estimated the posterior probability of each SNP to be causal in a single population (population-specific) or both populations (shared), and inferred the posterior expected numbers of population-specific/shared causal SNPs in each LD block by summing the per-SNP posterior probabilities of being causal in a single or both populations. Critically, while PESCA is an analysis based on summary statistics and not designed for admixed populations, it can be applied to admixed population such as LAT if in-sample LD is used(47). Through the PESCA analysis, we found 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. In other words, approximately 1.71% / (1.71%+1.69%+1.87%) = 32.5% of SNPs inferred to be causal are shared between NLW and LAT.

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 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(49) in R.

REFERENCES

1. 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 Jan;40(Database issue):D930-934.

2. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013 Jun;45(6):580–5.

3. Robinson JT, Turner D, Durand NC, Thorvaldsdóttir H, Mesirov JP, Aiden EL. Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data. Cell Syst. 2018 Feb 28;6(2):256-258.e1.

4. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell. 2014 Dec;159(7):1665–80.

5. Lin BD, Carnero-Montoro E, Bell JT, Boomsma DI, de Geus EJ, Jansen R, et al. 2SNP heritability and effects of genetic variants for neutrophil-to-lymphocyte and platelet-to-lymphocyte ratio. J Hum Genet. 2017 Nov;62(11):979–88.

6. Stadhouders R, Aktuna S, Thongjuea S, Aghajanirefah A, Pourfarzad F, van IJcken W, et al. HBS1L-MYB intergenic variants modulate fetal hemoglobin via long-range MYB enhancers. J Clin Invest. 2014 Apr 1;124(4):1699–710.

7. Guo MH, Nandakumar SK, Ulirsch JC, Zekavat SM, Buenrostro JD, Natarajan P, et al. Comprehensive population-based genome sequencing provides insight into hematopoietic regulatory mechanisms. Proc Natl Acad Sci. 2017 Jan 17;114(3):E327–36.

8. Li M, Jiang P, Cheng K, Zhang Z, Lan S, Li X, et al. Regulation of MYB by distal enhancer elements in human myeloid leukemia. Cell Death Dis. 2021 Feb;12(2):223.

9. 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 Nov;167(5):1415-1429.e19.

10. van Rooij FJA, Qayyum R, Smith AV, Zhou Y, Trompet S, Tanaka T, 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 Jan;100(1):51–63.

11. Tajuddin SM, Schick UM, Eicher JD, Chami N, Giri A, Brody JA, 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 Jul;99(1):22–39.

12. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019 Jan 8;47(D1):D1005–12.

13. Ahola-Olli AV, Würtz P, Havulinna AS, Aalto K, Pitkänen N, Lehtimäki T, et al. Genome-wide Association Study Identifies 27 Loci Influencing Concentrations of Circulating Cytokines and Growth Factors. Am J Hum Genet. 2017 Jan;100(1):40–50.

14. 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 Aug;20(8):1736–40.

15. Lynch JR, Salik B, Connerty P, Vick B, Leung H, Pijning A, et al. JMJD1C-mediated metabolic dysregulation contributes to HOXA9-dependent leukemogenesis. Leukemia. 2019 Jun;33(6):1400–10.

16. Chen M, Zhu N, Liu X, Laurent B, Tang Z, Eng R, et al. JMJD1C is required for the survival of acute myeloid leukemia by functioning as a coactivator for key transcription factors. Genes Dev. 2015 Oct 15;29(20):2123–39.

17. Xiao F, Liao B, Hu J, Li S, Zhao H, Sun M, et al. JMJD1C Ensures Mouse Embryonic Stem Cell Self-Renewal and Somatic Cell Reprogramming through Controlling MicroRNA Expression. Stem Cell Rep. 2017 Sep 12;9(3):927–42.

18. Cimmino L, Dawlaty MM, Ndiaye-Lobry D, Yap YS, Bakogianni S, Yu Y, et al. TET1 is a tumor suppressor of hematopoietic malignancy. Nat Immunol. 2015 Jun;16(6):653–62.

19. Bamezai S, Demir D, Pulikkottil AJ, Ciccarone F, Fischbein E, Sinha A, et al. TET1 promotes growth of T-cell acute lymphoblastic leukemia and can be antagonized via PARP inhibition. Leukemia [Internet]. 2020 May 15 [cited 2021 Jan 31]; Available from: http://www.nature.com/articles/s41375-020-0864-3

20. Lappalainen T, Sammeth M, Friedländer MR, ‘t Hoen PAC, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013 Sep 1;501(7468):506–11.

21. 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. Nat Commun. 2018 Dec;9(1):286.

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 Jun 7;447(7145):661–78.

23. Metayer C, Zhang L, Wiemels JL, Bartley K, Schiffman J, Ma X, et al. Tobacco smoke exposure and the risk of childhood acute lymphoblastic and myeloid leukemias by cytogenetic subtype. Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol. 2013 Sep;22(9):1600–11.

24. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015 Oct 1;526(7571):68–74.

25. the Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016 Oct;48(10):1279–83.

26. Genome Aggregation Database Consortium, Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020 May;581(7809):434–43.

27. 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 Jul;39(7):906–13.

28. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010 Sep 1;26(17):2190–1.

29. Vijayakrishnan J, Kumar R, Henrion MYR, 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 Mar;31(3):573–9.

30. Papaemmanuil E, Hosking FJ, Vijayakrishnan J, Price A, Olver B, Sheridan E, et al. Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. Nat Genet. 2009 Sep;41(9):1006–10.

31. Perez-Andreu V, Roberts KG, Harvey RC, Yang W, Cheng C, Pei D, et al. Inherited GATA3 variants are associated with Ph-like childhood acute lymphoblastic leukemia and risk of relapse. Nat Genet. 2013 Dec;45(12):1494–8.

32. Treviño LR, Yang W, French D, Hunger SP, Carroll WL, Devidas M, et al. Germline genomic variants associated with childhood acute lymphoblastic leukemia. Nat Genet. 2009 Sep;41(9):1001–5.

33. Xu H, Yang W, Perez-Andreu V, Devidas M, Fan Y, Cheng C, et al. Novel susceptibility variants at 10p12.31-12.2 for childhood acute lymphoblastic leukemia in ethnically diverse populations. J Natl Cancer Inst. 2013 May 15;105(10):733–42.

34. Walsh KM, de Smith AJ, Hansen HM, Smirnov IV, Gonseth S, Endicott AA, 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 Nov 15;75(22):4884–94.

35. Vijayakrishnan J, Henrion M, Moorman AV, Fiege B, Kumar R, Inacio da Silva Filho M, et al. The 9p21.3 risk of childhood acute lymphoblastic leukaemia is explained by a rare high-impact variant in CDKN2A. Sci Rep. 2015 Dec;5(1):15065.

36. de Smith AJ, Walsh KM, Francis SS, Zhang C, Hansen HM, Smirnov I, et al. BMI1enhancer polymorphism underlies chromosome 10p12.31 association with childhood acute lymphoblastic leukemia: *BMI* 1 enhancer polymorphism in ALL. Int J Cancer. 2018 Dec 1;143(11):2647–58.

37. Schumacher FR, Al Olama AA, Berndt SI, Benlloch S, Ahmed M, Saunders EJ, et al. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. Nat Genet. 2018 Jul;50(7):928–36.

38. Kharazmi E, da Silva Filho MI, Pukkala E, Sundquist K, Thomsen H, Hemminki K. Familial risks for childhood acute lymphocytic leukaemia in Sweden and Finland: far exceeding the effects of known germline variants. Br J Haematol. 2012 Sep;n/a-n/a.

39. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A Tool for Genome-wide Complex Trait Analysis. Am J Hum Genet. 2011 Jan;88(1):76–82.

40. The LifeLines Cohort Study, Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AAE, et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat Genet. 2015 Oct;47(10):1114–20.

41. 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 Jul;103(1):89–99.

42. Speed D, Hemani G, Johnson MR, Balding DJ. Improved Heritability Estimation from Genome-wide SNPs. Am J Hum Genet. 2012 Dec;91(6):1011–21.

43. Zaitlen N, Pasaniuc B, Sankararaman S, Bhatia G, Zhang J, Gusev A, et al. Leveraging population admixture to characterize the heritability of complex traits. Nat Genet. 2014 Dec;46(12):1356–62.

44. Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am J Hum Genet. 2013 Aug 8;93(2):278–88.

45. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009 Sep 1;19(9):1655–64.

46. the PRACTICAL consortium, Mancuso N, Rohland N, Rand KA, Tandon A, Allen A, et al. The contribution of rare variation to prostate cancer heritability. Nat Genet. 2016 Jan;48(1):30–5.

47. Shi H, Burch KS, Johnson R, Freund MK, Kichaev G, Mancuso N, et al. Localizing Components of Shared Transethnic Genetic Architecture of Complex Traits from GWAS Summary Data. Am J Hum Genet. 2020 Jun;106(6):805–17.

48. Berisa T, Pickrell JK. Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics. 2015 Sep 22;btv546.

49. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011 Dec;12(1):77.