Cell Genomics, Volume *4*

Supplemental information

A noncoding regulatory variant in *IKZF1* increases acute lymphoblastic leukemia risk in Hispanic/Latino children

Adam J. de Smith, Lara Wahlster, Soyoung Jeon, Linda Kachuri, Susan Black, Jalen Langie, Liam D. Cato, Nathan Nakatsuka, Tsz-Fung Chan, Guangze Xia, Soumyaa Mazumder, Wenjian Yang, Steven Gazal, Celeste Eng, Donglei Hu, Esteban González Burchard, Elad Ziv, Catherine Metayer, Nicholas Mancuso, Jun J. Yang, Xiaomei Ma, Joseph L. Wiemels, Fulong Yu, Charleston W.K. Chiang, and Vijay G. Sankaran

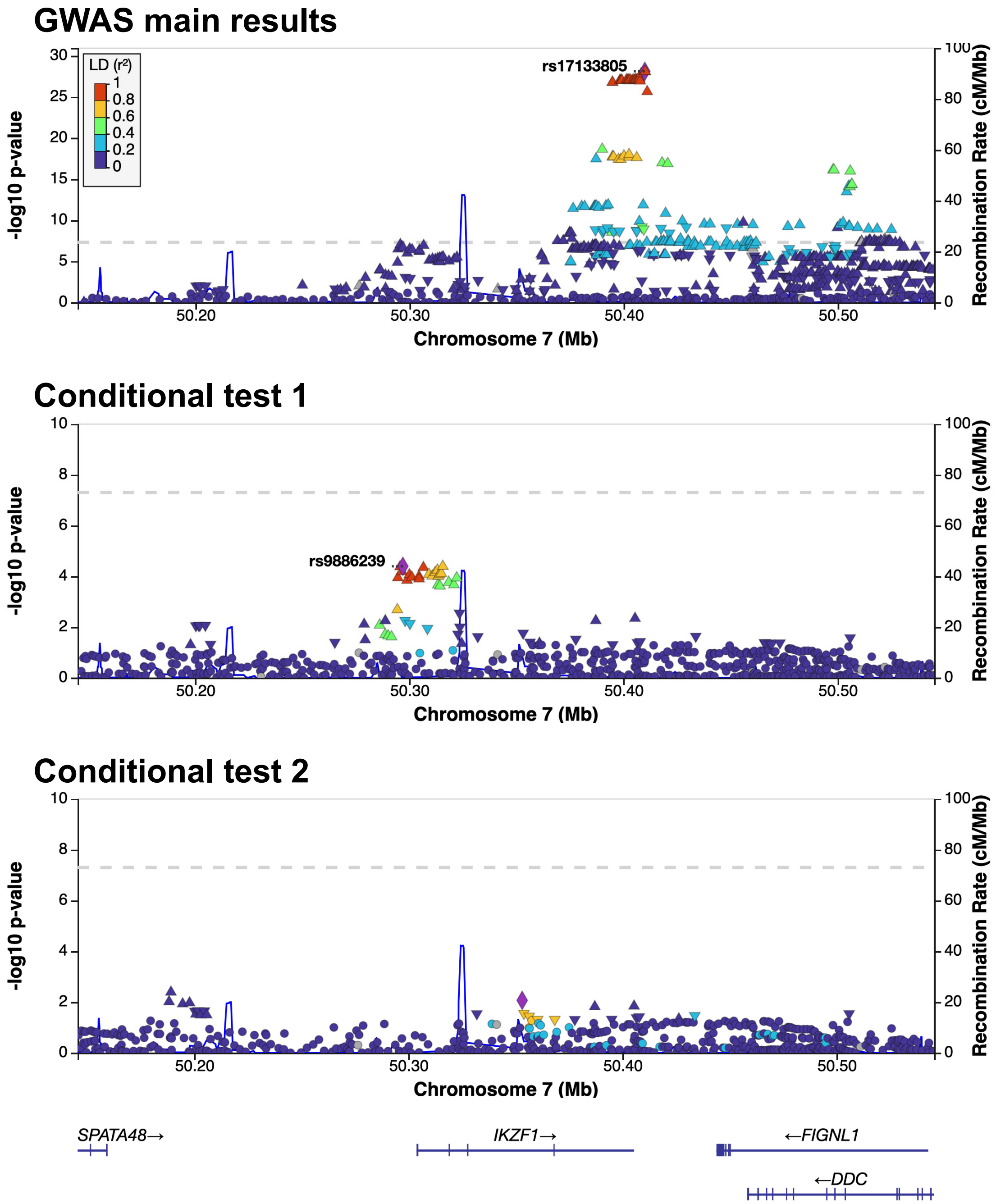
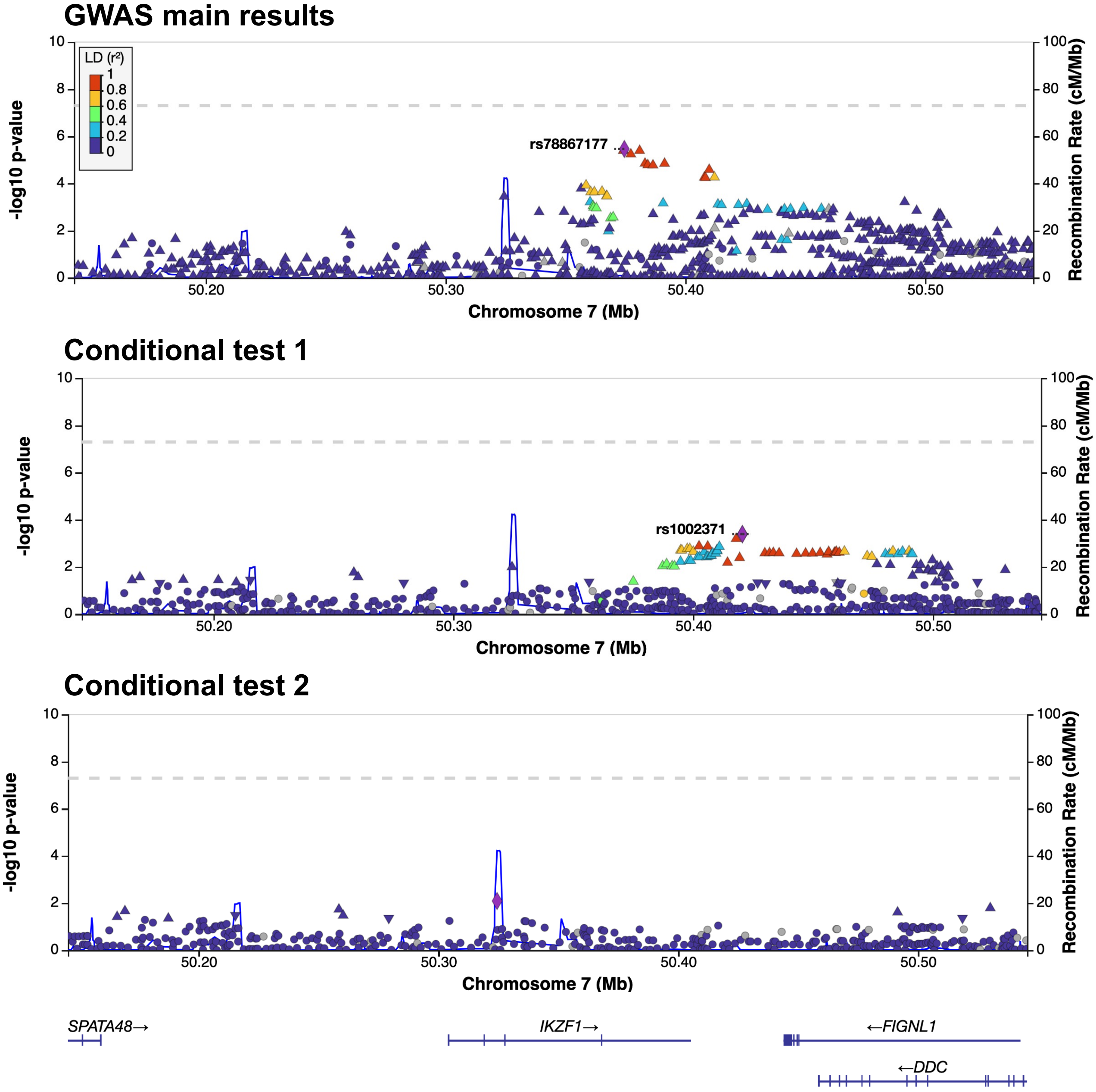


Figure S1 - Two independent childhood ALL association signals identified in non-Hispanic Whites, related to Figure 1. LocusZoom plots showing an approximately 600 Mb region at chromosome 7p12.2 centered on the *IKZF1* gene region (+/-250 kb), from (A) GWAS main results, the unconditional GWAS of childhood ALL in non-Hispanic Whites, and (B) Conditional test 1, results conditioned on the lead SNP (rs17133805) in the main GWAS. GWAS conditioned on the lead SNPs in the main GWAS and Conditional test 1 revealed no further ALL association peaks (C). Diamond symbols indicate the lead SNP in each locus. Color of remaining SNPs is based on linkage disequilibrium (LDS) as measured by r^2 with the lead SNP in each signal. All coordinates are in genome build hg38.


GWAS main results

Conditional test 1

Conditional test 2

Figure S2 - Childhood ALL association signals in East Asians, related to Figure 1. LocusZoom plots showing an approximately 600 Mb region at chromosome 7p12.2 centered on the *IKZF1* gene region (+/- 250 kb), from (A) GWAS main results, the unconditional GWAS of childhood ALL in East Asians, and (B) Conditional test 1, results conditioned on the lead SNP (rs78867177) in the main GWAS. GWAS conditioned on the lead SNPs in the main GWAS and Conditional test 1 revealed no further ALL association peaks (C). Diamond symbols indicate the lead SNP in each locus. Color of remaining SNPs is based on linkage disequilibrium (LDS) as measured by r^2 with the lead SNP in each signal. All coordinates are in genome build hg38.

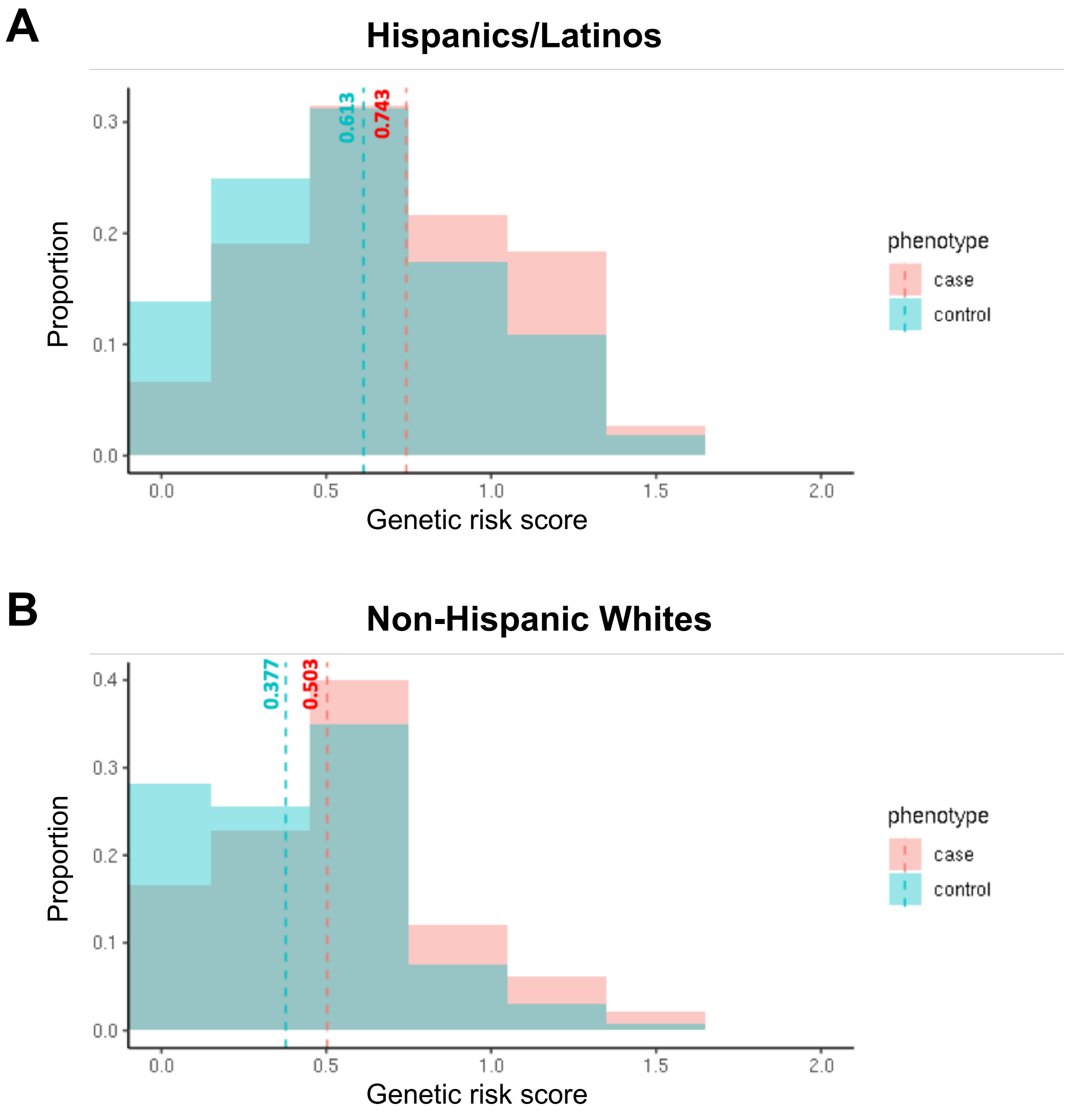


Figure S3 - *IKZF1* genetic risk score distribution in Hispanic/Latino and non-Hispanic White children in the CCRLP, related to STAR Methods. We compared the genetic risk score (GRS) distribution between Hispanics/Latinos (A) and non-Hispanic Whites (B) in the CCRLP. GRS were calculated using the three independent lead SNPs in Hispanics/Latinos and the two independent lead SNPs in non-Hispanic Whites, weighted by their corresponding marginal or conditional effect estimates. Subjects were further stratified by case/control status. The population mean is indicated with vertical dash lines with the mean score shown.

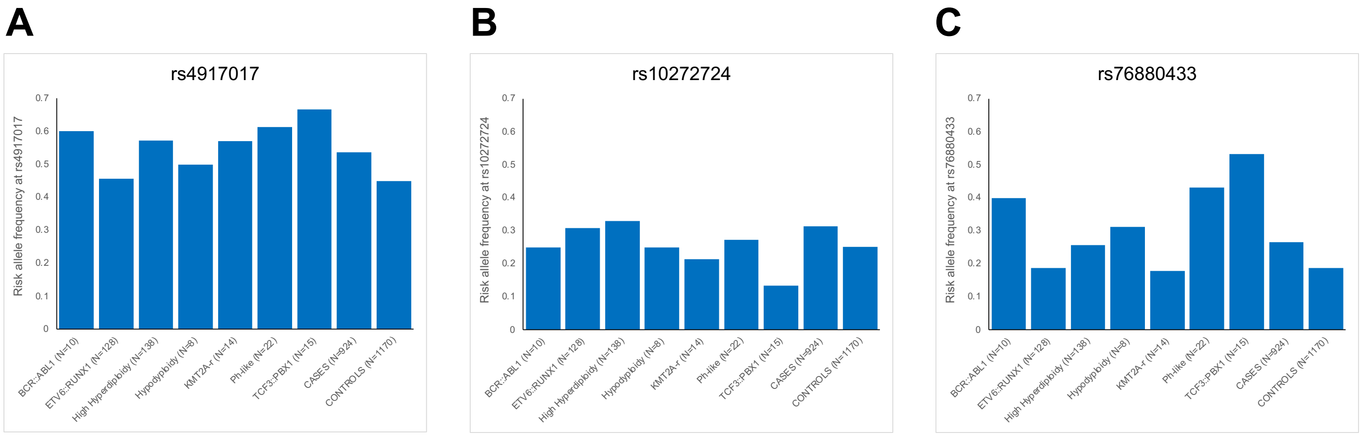


Figure S4 - *IKZF1* SNP risk allele frequencies across childhood ALL molecular subtypes, related to STAR Methods. Risk allele frequencies of lead SNPs at the three independent ALL association loci in Hispanics/Latinos - signal 1 (rs4917017), signal 2 (rs10272724), and signal 3 (rs76880433) - in childhood ALL cases from the Children’s Oncology Group (COG)/St. Jude Children’s Hospital cohorts.

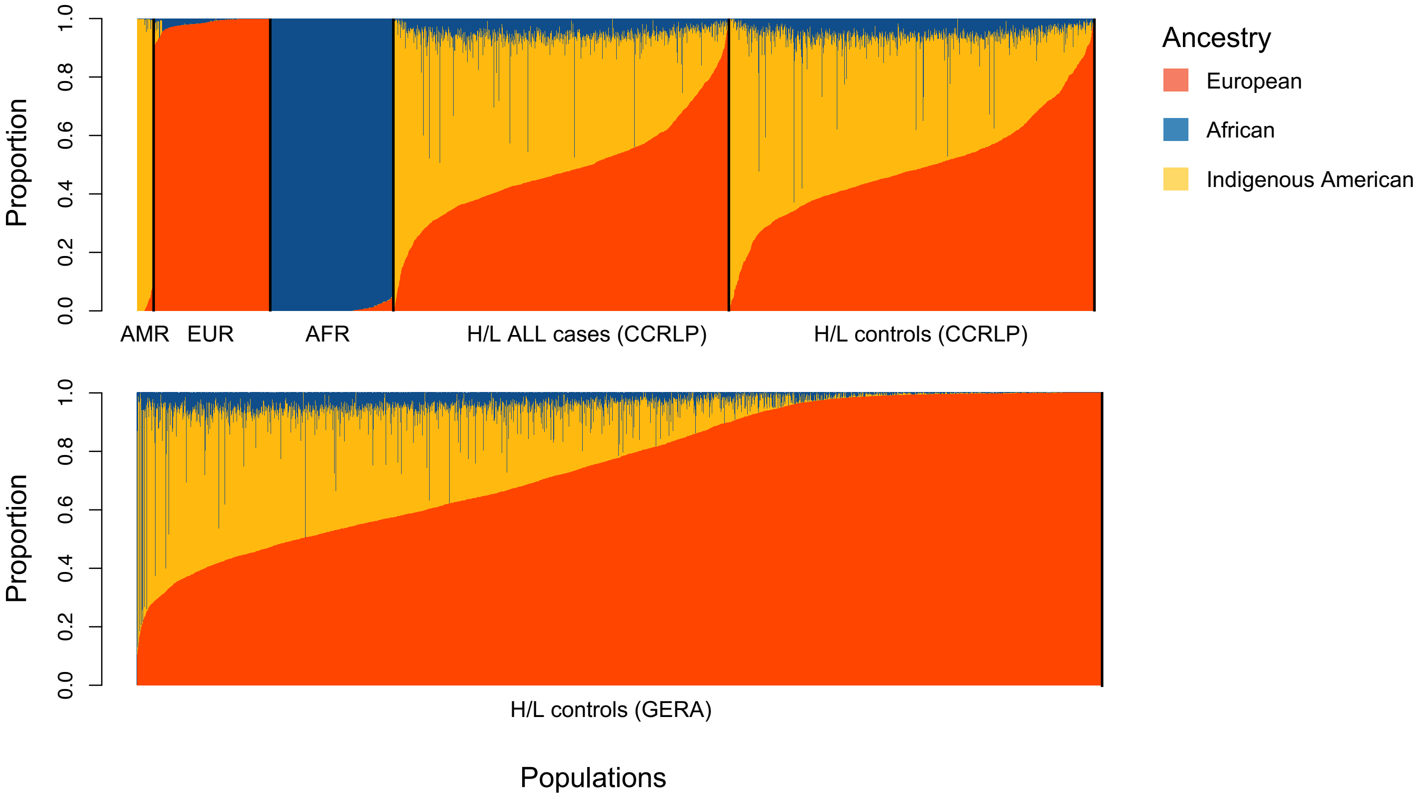


Figure S5 - Global ancestry proportions estimated in Hispanic/Latino study subjects, related to Figure 1. For each individual, local ancestry was inferred using RFMix, using a reference panel consisting of 671 non-Finnish European individuals (EUR) for European ancestry, 708 African individuals (AFR) for African ancestry, and 94 selected Latino/Admixed American (AMR) individuals for Indigenous American ancestry from gnomAD. Proportions of global Indigenous American, European, and African ancestry were calculated by summing local ancestry estimates across the genome. In the combined set of Hispanic/Latino (H/L) subjects in our study (CCRLP plus GERA), the average global ancestry proportions were estimated to be 28.0% Indigenous American, 67.6% European, and 4.4% African ancestry. Stratifying by study/cohort and case/control status, CCRLP cases had on average 42.5% Indigenous American, 52.0% European, and 5.5% African ancestry, CCRLP controls had 41.7% Indigenous American, 52.7% European, and 5.6% African ancestry, and GERA controls had 19.2% Indigenous American, 77.1% European, and 3.7% African ancestry.

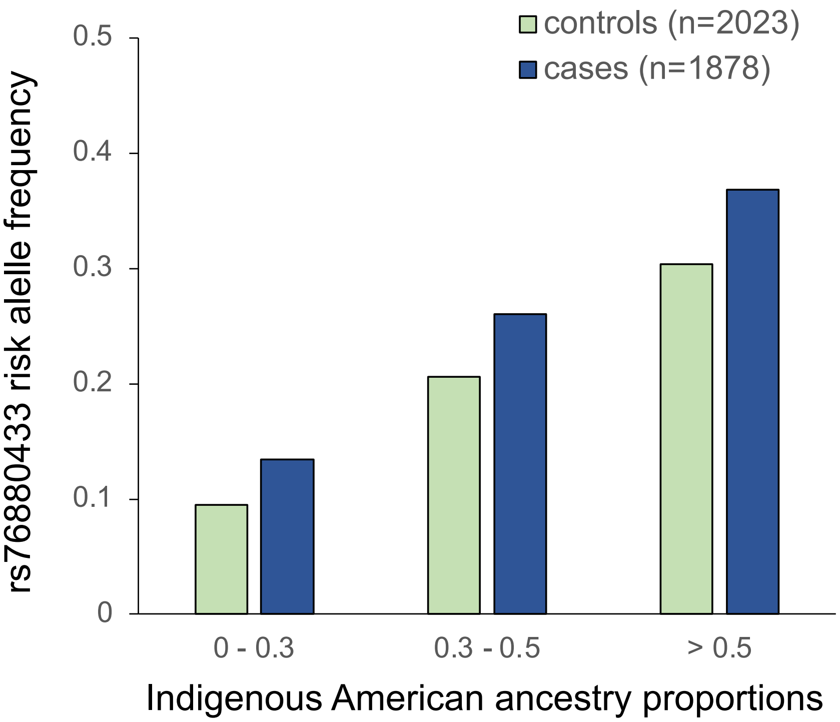


Figure S6 - *IKZF1* signal 3 lead SNP rs76880433 risk allele frequency in Hispanics/Latinos by Indigenous American ancestry proportions, related to Figure 1. Risk allele frequency for rs76880433 is highest in cases and controls with >50% Indigenous American ancestry.

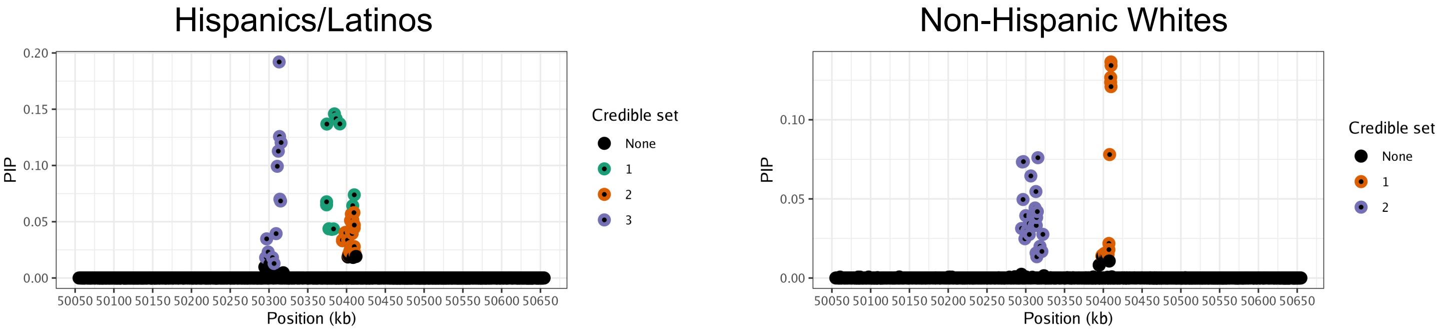


Figure S7 - SuSiE credible set variants in Hispanics/Latinos and non-Hispanic Whites in the CCRLP, related to Figure 1 and Figure S1. Three credible sets were reported by the SUm of SIngle Effects (SuSiE) model in Hispanics/Latinos, and two credible sets were reported in non-Hispanic Whites. Posterior inclusion probability (PIP) values are shown on the Y-axes.

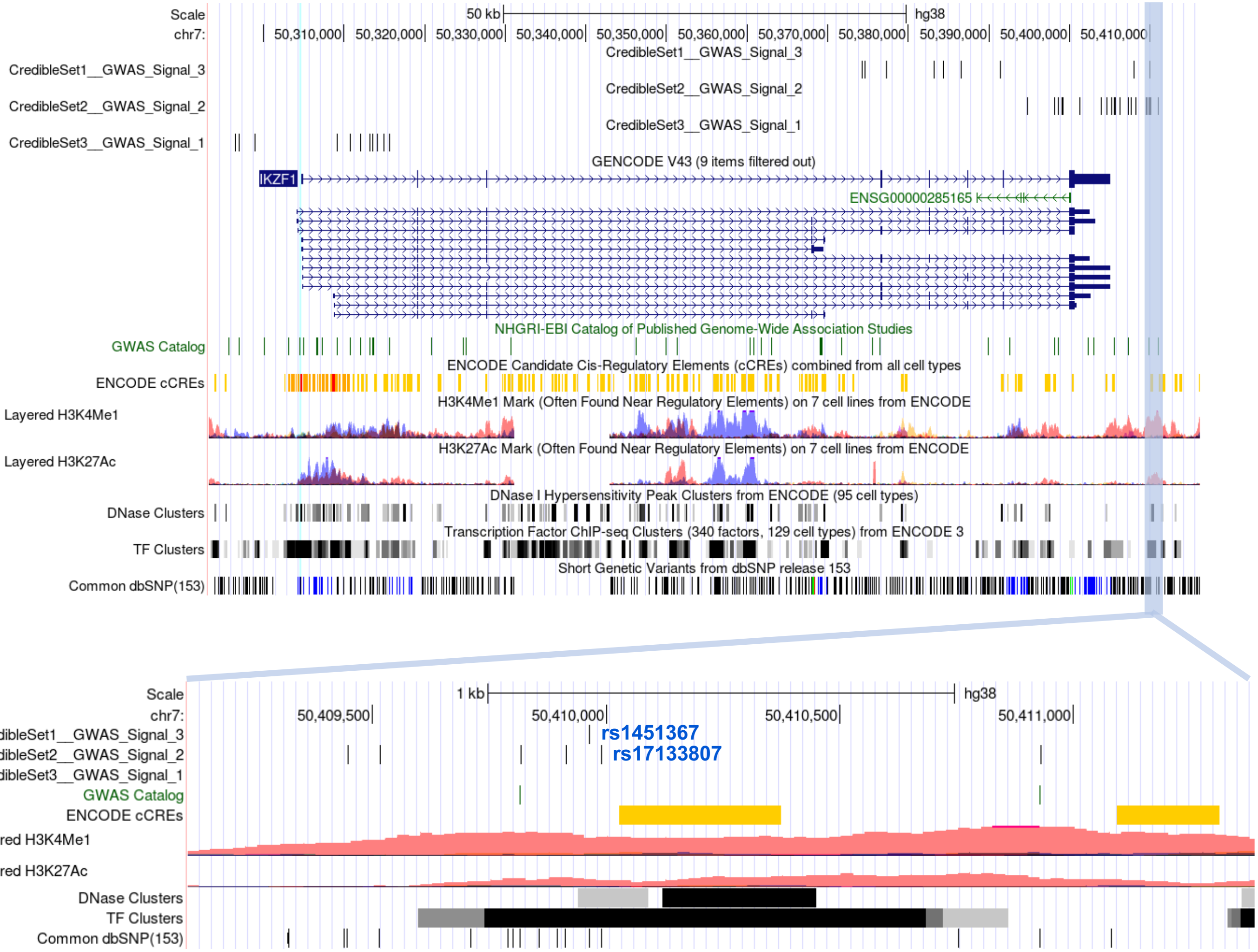

Scale
50 kb
hg38
chr7:
50,310,000
50,320,000
50,330,000
50,340,000
50,350,000
50,360,000
50,370,000
50,380,000
50,390,000
50,400,000
50,410,000
CredibleSet1__GWAS_Signal_3
CredibleSet2__GWAS_Signal_2
CredibleSet3__GWAS_Signal_1
GENCODE V43 (9 items filtered out)
IKZF1
ENSG00000285165
NHGRI-EBI Catalog of Published Genome-Wide Association Studies
GWAS Catalog
ENCODE Candidate Cis-Regulatory Elements (cCREs) combined from all cell types
ENCODE cCREs
H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K4Me1
H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
DNase I Hypersensitivity Peak Clusters from ENCODE (95 cell types)
DNase Clusters
Transcription Factor ChIP-seq Clusters (340 factors, 129 cell types) from ENCODE 3
TF Clusters
Short Genetic Variants from dbSNP release 153
Common dbSNP(153)
Scale
1 kb
hg38
chr7:
50,409,500
50,410,000
50,410,500
50,411,000
CredibleSet1__GWAS_Signal_3
CredibleSet2__GWAS_Signal_2
CredibleSet3__GWAS_Signal_1
GWAS Catalog
ENCODE cCREs
Layered H3K4Me1
Layered H3K27Ac
DNase Clusters
TF Clusters
Common dbSNP(153)
rs1451367
rs17133807

Figure S8 - Position of Hispanic/Latino credible set variants in relation to *IKZF1* gene and regulatory elements, related to Figure 2. Top: three independent credible sets of variants associated with childhood ALL in Hispanics/Latinos, corresponding to GWAS signals 1 (credible set 3), 2 (credible set 2), and 3 (credible set 1), and their chromosome 7 position (hg38) in relation to regulatory elements, in the UCSC Genome Browser. Only one variant in credible set 1/GWAS signal 3 - rs1451367 - overlaps with predicted regulatory elements based on peaks for H3K4Me1, H3K27Ac, and DNase I hypersensitivity.

Bottom: The putative causal variant rs1451367 in credible set 1/GWAS signal 3 lies only 26bp upstream of the credible set 2/GWAS signal 2 variant rs17133807, both of which reside in the same predicted enhancer element.

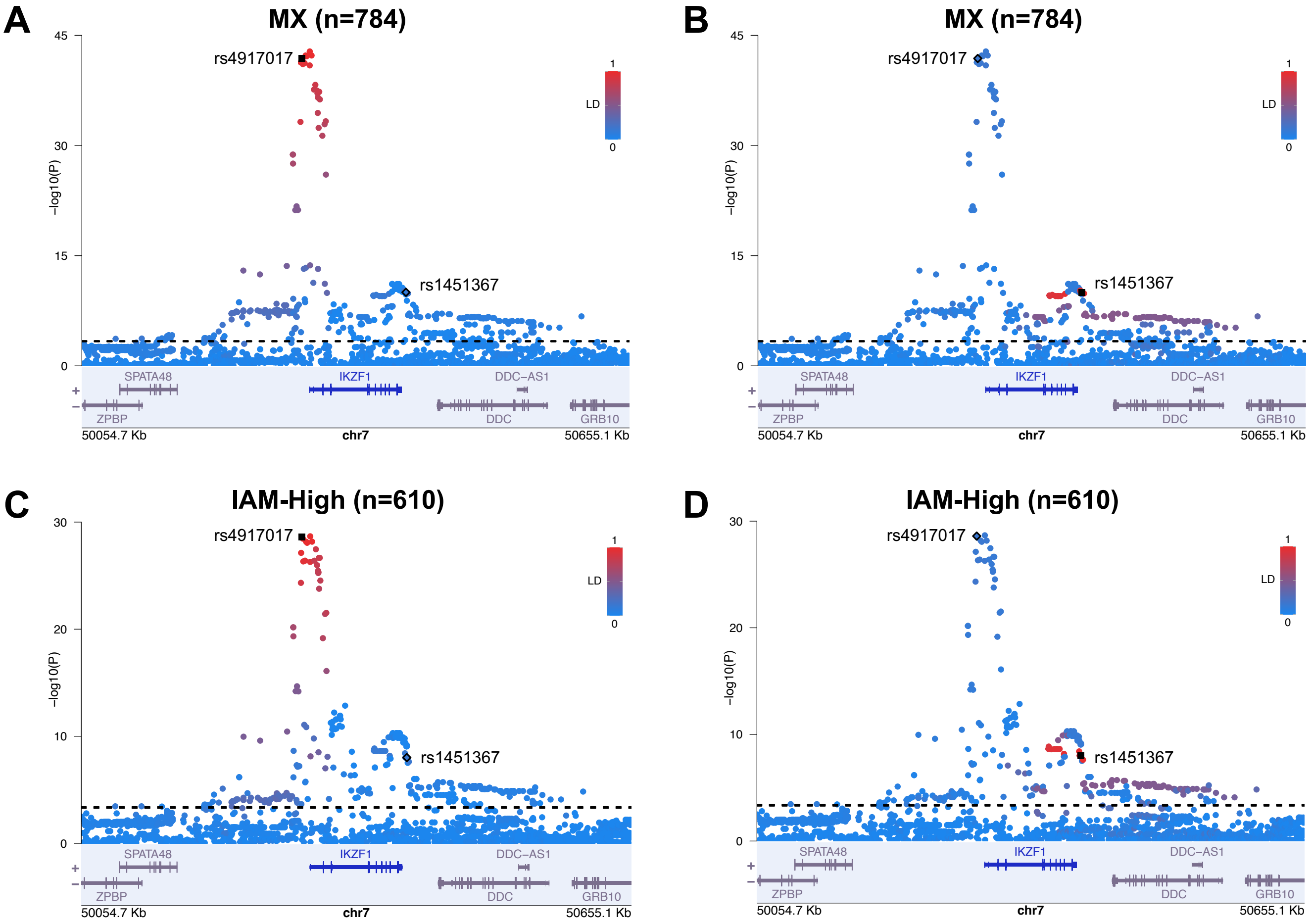


Figure S9 - Expression quantitative trait locus (eQTL) results for *IKZF1* SNPs in Mexican American children, related to STAR Methods. Regional plots depicting associations with whole blood *IKZF1* expression in children from GALA II. Analyses were conducted separately in Mexican Americans (MX), based on self-identified race/ethnicity, and in participants with >50% global Indigenous American (IAM_{High}) genetically inferred ancestry. Linkage disequilibrium (LD) r^2 is visualized with respect to rs4917017 in A) and C) and with respect to rs1451367 in B) and D).

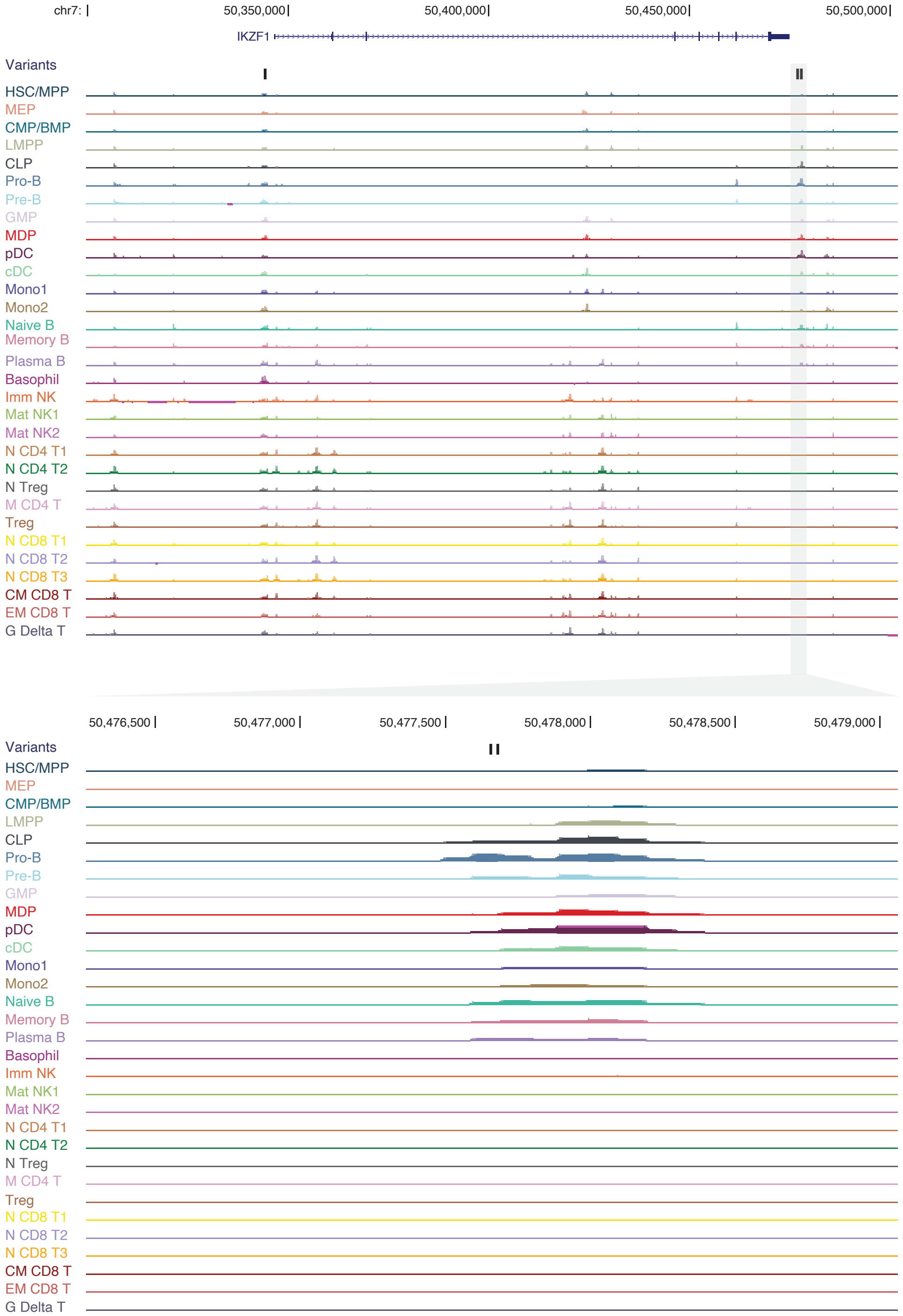


Figure S10 - Putative causal variants in three independent *IKZF1* association loci in Hispanics/Latinos in relation to chromatin accessibility across hematopoietic cell types, related to Figure 2. Top: Chromatin accessibility across the *IKZF1* region was assessed using single-cell ATAC-sequencing from human hematopoiesis, as we have previously described.[27] Bottom: Putative causal variants rs17133807 and rs1451367 in Hispanic/Latino ALL GWAS signals 2 and 3 overlapped the same regulatory element, with strong accessibility in B-cells and their precursors, and the greatest accessibility at the pro-B stage, but minimal or no accessibility in T-cells or myeloid cells.

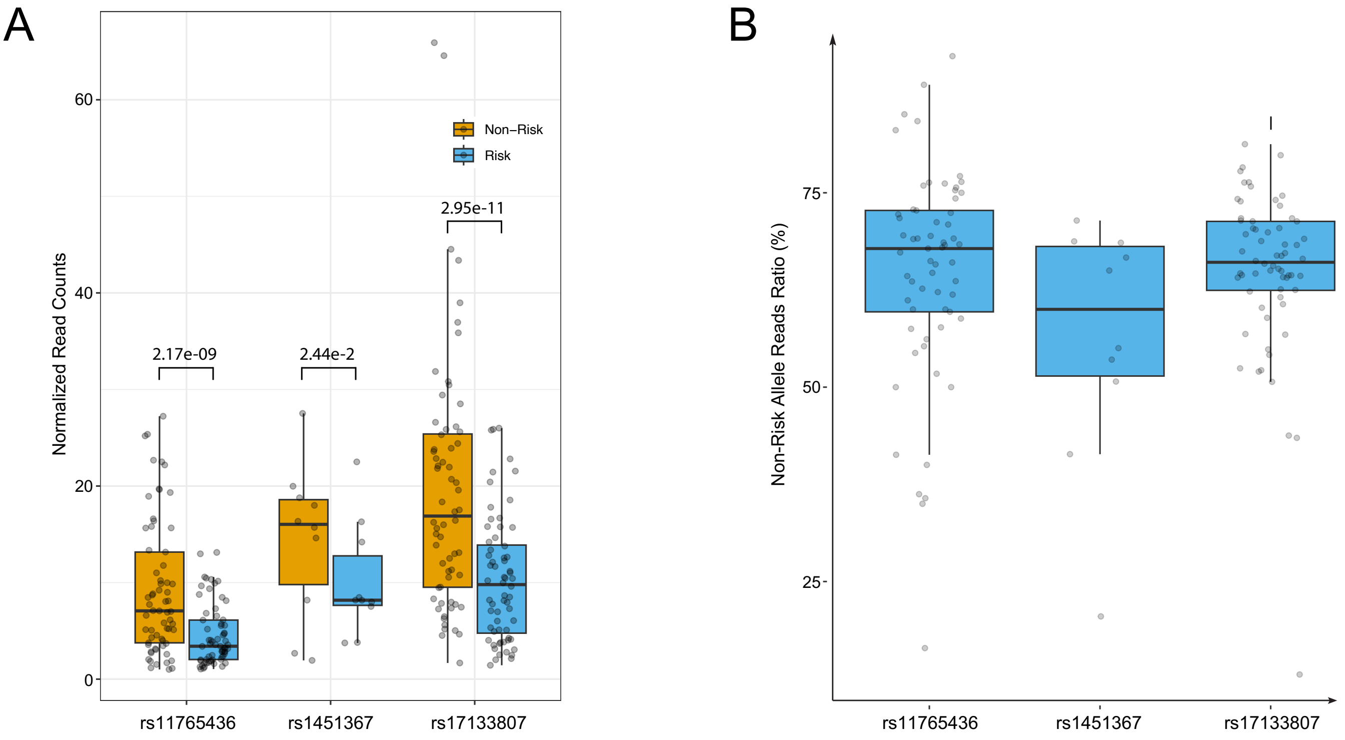


Figure S11 - Allele-specific chromatin accessibility at putative causal *IKZF1* variants in B-cell ALL patients, related to Figure 2. Among 156 B-cell ALL patients, we assessed ATAC-sequencing read counts for the risk and non-risk alleles for putative causal variants underlying the Hispanic/Latino childhood ALL GWAS signals 1-3. (A) Boxplots displaying the normalized read counts for the non-risk versus risk alleles in B-cell ALL patients heterozygous for SNPs rs11765436 (non-risk/risk:T/A, n=65), rs1451367 (non-risk/risk:C/T, n=10), and rs17133807 (non-risk/risk:G/A, n=64). For each SNP, ATAC-sequencing reads were significantly increased for the non-risk allele compared with the risk allele in paired Wilcoxon signed rank tests (1-tailed), supporting a bias towards chromatin accessibility for the non-risk alleles. (B) Boxplots displaying the ratio of normalized read counts for non-risk versus risk alleles in SNP heterozygotes (calculated by non-risk/[non-risk + risk] allele read counts). ATAC-sequencing read counts were normalized by reads per million (see Table S11 for details).

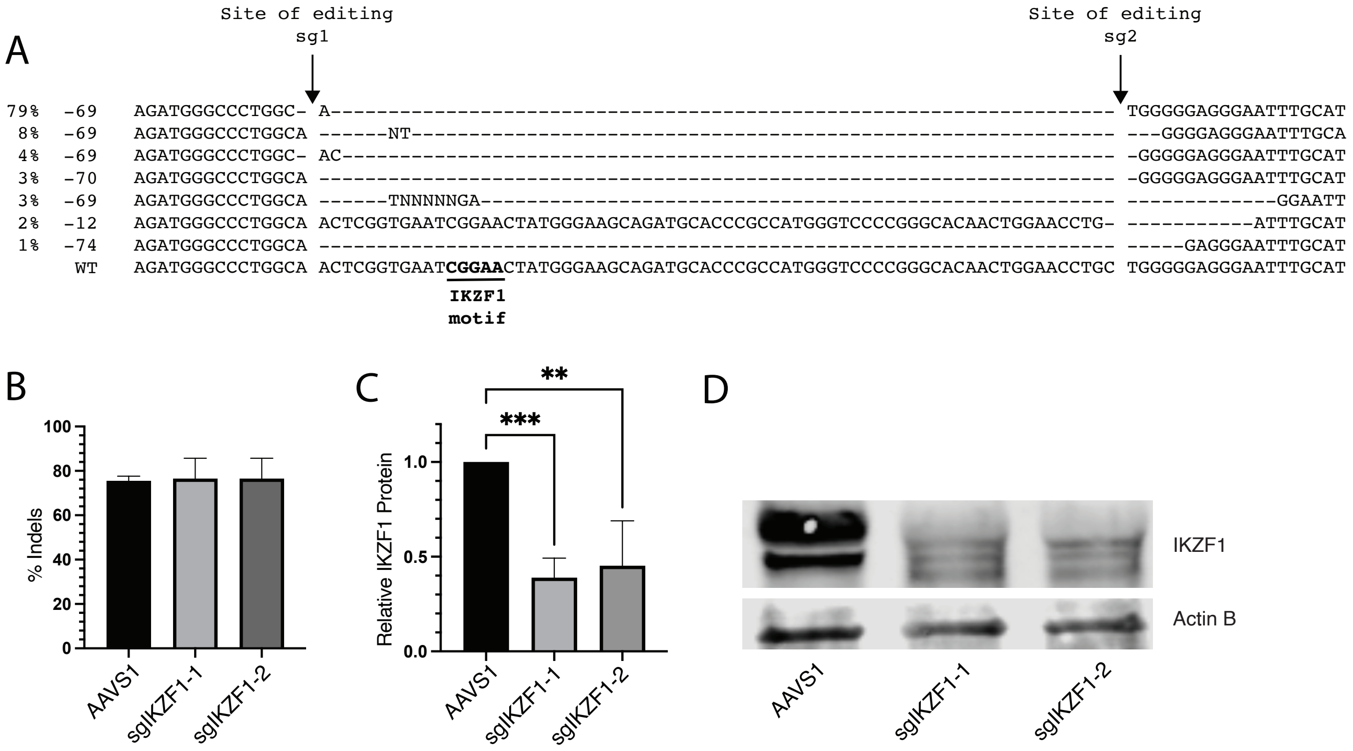


Figure S12 - CRISPR/Cas9-genome editing targeting the *IKZF1* binding motif within the enhancer sequence or *IKZF1* coding sequence in HSPCs, related to STAR Methods.

(A) Genome editing targeting the *IKZF1* binding motif within the enhancer sequence in HSPCs. Editing efficiency after 3 weeks of B-cell culture from HSPCs following RNP delivery of Cas9 and dual sgRNAs for introduction of microdeletions spanning the IKZF1 TF binding motif within the enhancer region. Prediction of editing outcomes by sanger sequencing shows major edited sequences result in successful introduction of microdeletions and disruption of the IKZF1 TF binding motif. (B) Genome editing targeting the *IKZF1* coding sequence in HSPCs. Editing efficiency was detected at 72 hours after RNP delivery of Cas9 and sgRNA (sgAAVS1, sgIKZF1-1 or sgIKZF1-2) by nucleofection. (C, D) Western blot showing decreased IKZF1 protein expression 96 hours after genome editing in HSPCs. Bar graphs demonstrating quantitative expression change (C) and representative images of the western blots are shown (D). Experiments were performed in triplicate.

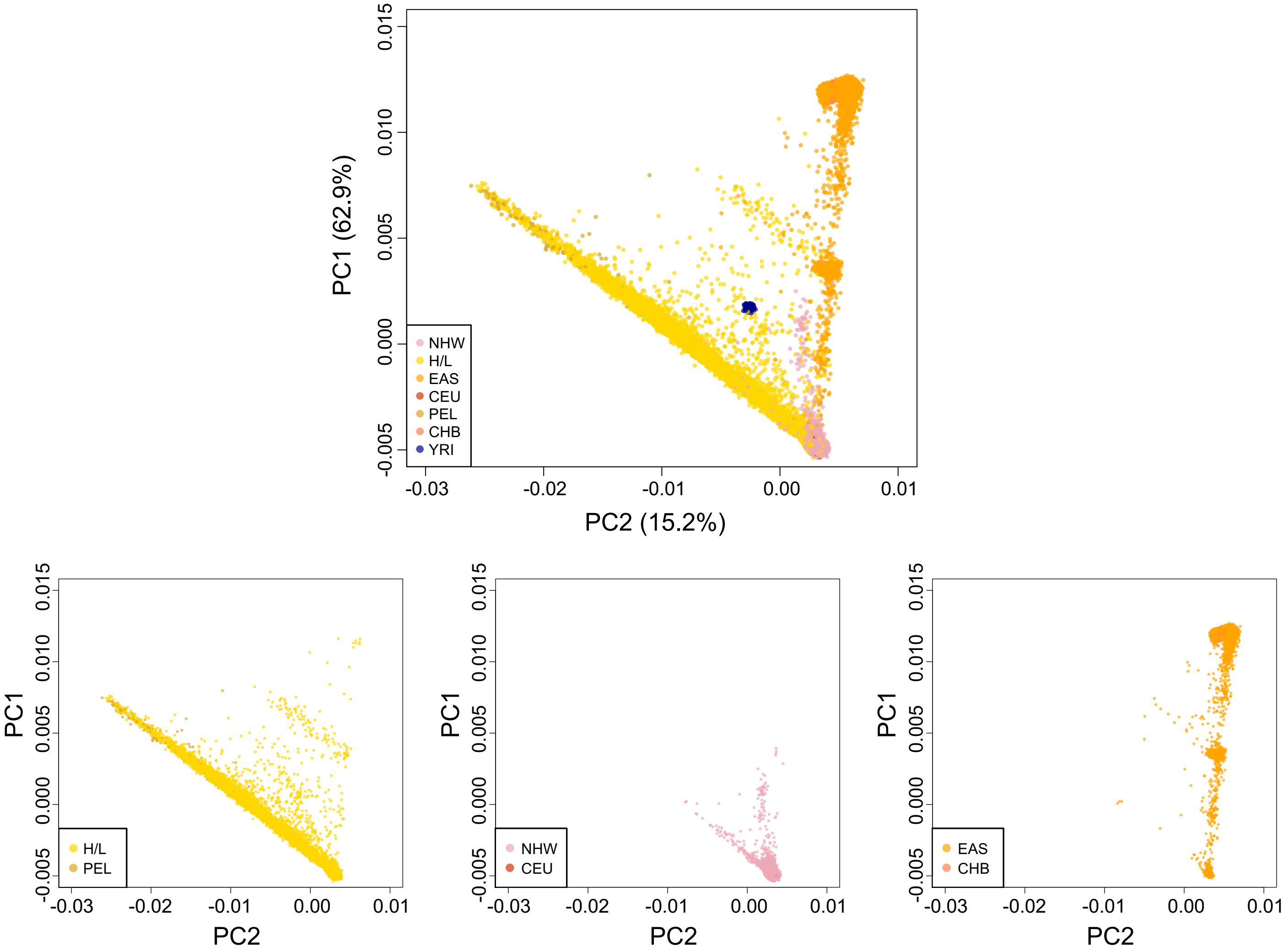


Figure S13 - PCA analysis of study subjects together with 1000 Genome project reference populations, related to STAR Methods.

Principal components analysis (PCA) plots were generated for Hispanic/Latino (H/L), non-Hispanic White (NHW), and East Asian (EAS) study subjects in CCRLP/GERA along with individuals from 1000 Genomes Project populations CEU (Northern Europeans from Utah), PEL (Peruvian in Lima, Peru), CHB (Han Chinese), and YRI (Yoruba) (top). In general, NHW subjects cluster with CEU towards the bottom right, EAS subjects cluster with CHB towards the top right, and H/L subjects cluster from the left with PEL towards the bottom right. The plots below display each of the CCRLP/GERA populations plotted individually along with their closest corresponding 1KG population. PCs were calculated using PLINK (version 1.90) and plots were generated using *R* (version 4.3.1).