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

Soyoung Jeon[1,2], Adam J. de Smith[1], Shaobo Li[1,2], Minhui Chen[1], Ivo S. Muskens[1], Libby M. Morimoto[3], Andrew T. DeWan[4,5], Nicholas Mancuso[1,6,7], Catherine Metayer[3], Xiaomei Ma[5,8], 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
8. Yale Cancer Outcomes, Public Policy, and Effectiveness Research (COPPER) Center, Yale University, New Haven, CT

Correspondence: J.L.W. (wiemels@usc.edu), C.W.K.C. (charleston.chiang@med.usc.edu)

ABSTRACT

The incidence patterns of childhood acute lymphoblastic leukemia (ALL) differ across ethnic groups but have been studied mostly in populations of predominantly European ancestries. Risk variants identified from previous genome-wide association studies (GWAS) do not fully explain heritable risk. In an effort to address these limitations, we performed a meta-analysis of ALL in 76,317 participants across four ethnic groups, including 17,814 non-European individuals and 3,482 total cases. We generally replicated previously identified loci associated with ALL (15 out of 16 loci replicated after Bonferroni corrections). 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* on chr6q23.3, rs10998283 near *TET1* on chr10q21.3, and rs9415680 near *JMJD1C/NRBF2* on chr10q21.3) were previously shown to be associated with multiple blood cell traits and other hematopoietic cancers. When trans-ethnic information is used, polygenic risk scores constructed from GWAS loci in 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.

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 that confer risk of childhood ALL including in *ARID5B, IKZF1, CEBPE, CDKN2A, PIP4K2A, LHPP* and *ELK3*, among others[3–8]. However, these studies were generally composed of participants with a predominantly European ancestry, and efforts to replicate these findings in other ethnic groups have been limited, leading to a gap in our understanding of the genetic architecture of ALL in non-European populations. For example, among major ethnic groups in the United States, Latino children have the highest risk of ALL, with an incidence rate ~15-40% higher than in non-Latino whites[9–11], and an increased chance of relapse and poorer overall survival[12,13]. Yet, individuals outside of the non-Latino white group (Latinos, African Americans and Asians) are generally underrepresented in GWAS, with the exception of recent efforts for childhood ALL in Latinos[14–16]. 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 of the disease across ethnic groups.

The previously identified risk loci for ALL are generally common (except for a low frequency missense variant in *CDKN2A*[17]) in European ancestry populations. The heritability of ALL was estimated to be 21% in the European ancestry population, but the known risk loci together account for a relatively small portion of the total variance in genetic risk of ALL[18]. This suggests that additional susceptibility alleles may yet to be discovered in larger genetic studies of ALL. Moreover, polygenic risk scores (PRS) hold great promise for preventive risk assessment and stratification in a population, but are known to be poorly transferred to non-European populations[19]. It has been shown that PRS derived from multi-ethnic GWAS summary statistics would be more transferrable between populations, since multi-ethnic designs are more effective to identify alleles with shared effects across population without explicit fine-mapping[20,21].

Given this context, we performed a trans-ethnic GWAS of childhood ALL in a discovery panel consisting of 76,317 individuals from a multi-ethnic cohort derived from the California birth population and Kaiser's Cohort for Genetic Epidemiology Research on Aging. Our cohort consisted of 3,482 cases and 72,835 controls for an effective sample size of 13,292(Methods). We note the complexity of discussing race, ethnicity and ancestry in a genetic study. As a convention, we used the following term 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[22] (Methods). To our knowledge, this is 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 independent cohorts with cases from the California Childhood Leukemia Study and the Children's Oncology Group cohort. We further compared the efficacy of PRS to stratify individuals based on their risk of ALL in the two largest subgroups of our data, LAT and NLW, and contrasted the genetic architecture of ALL between them.

SUBJECTS AND METHODS

Study Subjects in Discovery Cohort

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. For each case, one control subject was randomly selected from a pool of potential controls and matched to case on year and month of birth, sex (56% male, 44% female), and ethnicity. A detailed information on sample preparation and genotyping has been previously described.[4] Furthermore, because ALL is a rare childhood cancer, for the purpose of a genetic study we incorporated additional controls using adult individuals from the Kaiser Resource for Genetic Epidemiology Research on Aging Cohort (GERA). The GERA cohort was chosen because a very similar genotyping platform had been used: Affymetrix Axiom World arrays. The GERA genotype information was downloaded from dbGaP (Study Accession: phs000788.v1.p2).

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 did perform principal components analysis for each population including our study subjects along with 1000 Genomes Project reference data[22] 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 performed two separate GWASs. First, stratified by ethnicity and restricted to post-QC SNPs 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[23]. 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)[24] and CCRLP NLW controls (> 0.1). We next performed another GWAS between CCRLP controls and GERA individuals, 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. An effective population was calculated using this equation:

$$N_{eff} = \frac{4}{\frac{1}{N_{cases}} + \frac{1}{N_{controls}}}$$

where N_{eff} is the effective sample size, N_{cases} is the number of cases, and $N_{controls}$ is the number of control subjects. A total of 7,628,894 SNPs that remained in at least three ethnic groups were tested in our GWAS discovery analysis.

Association Testing in Discovery Cohort

In each ethnic strata, we used SNPTEST[26] (v2.5.2) to test the association between imputed genotype dosage at each SNP and case-control status in logistic regression, after adjusting for the top 20 PCs. The result 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. We also tested to ensure sex is not correlated with the genotype dosage of SNPs with $P < 5 \times 10^{-7}$ in our meta-analysis, thus our results are not explained by sex as a confounder.

For replication of previously identified ALL risk loci, we focused on the variants within 1Mb of the previously reported susceptibility variants[3–8,17,27,28]. We report the association results of the published lead SNP as well as the top SNP at each locus from our meta-analysis. For replication of the known loci, a significance of 0.00312 (=0.05/16) was used.

To identify secondary signals at each of the three putatively novel and 16 previously known risk loci, we performed multivariate conditional analysis adjusting for the lead SNP and 20 PCs. A second round of conditional analyses conditioning on the top and the second top SNP resulted in no additional SNPs reaching genome-wide significance at any of the loci examined. A genome-wide threshold of 5×10^{-8} was used for significance.

176 **Association Testing in Replication Cohorts**
177
178 We included two independent ALL case-control replication studies: (1) individuals of
179 predominantly European ancestry from the Children's Oncology Group (COG, [dbGAP
180 accession number phs000638.v1.p1]) cohort as cases and from Wellcome Trust Case–Control
181 Consortium[29] (WTCCC) as controls, genotyped on either the Affymetrix Human SNP Array 6.0
182 (WTCCC, COG trials AALL0232 and P9904/9905) or the Affymetrix GeneChip Human
183 Mapping 500K Array (COG P9906 and St. Jude Total Therapy XIIIB/XV[18]); and (2) individuals
184 of European and Latino ancestry from the California Childhood Leukemia Study (CCLS), a non-
185 overlapping California case-control study (1995-2008).[30] The quality control procedures and
186 imputation were performed in accordance with the discovery cohort. For COG and WTCCC,
187 because self-identified ethnicity was not available to us, we performed global ancestry
188 estimations using ADMIXTURE and the 1000 Genomes populations as reference and removed
189 individuals with < 90% estimated European ancestry from the analysis. This resulted in 1504
190 cases and 2931 controls from COG/WTCCC, and in the CCLS, 426 NLW cases, 278 NLW
191 controls, 758 LAT cases and 549 LAT controls for our replication analysis. In each dataset, we
192 tested the imputed SNP dosages in the three putatively novel loci with $P < 5 \times 10^{-7}$ (number of
193 SNPs=141) in logistic regression with top 20 PCs as covariates using PLINK[31] (v2.3 alpha).
194

195 **Polygenic Risk Score Analysis**

196 Polygenic risk scores (PRS) for ALL were constructed for each individual in CCLS LAT, NLW,
197 and COG/WTCCC by summing the number of risk alleles, each weighted by its effect size from
198 our discovery GWAS meta-analysis. PRS were constructed in two different ways: (1) a PRS with
199 lead SNPs in the 16 known loci (N = 18 SNPs, including variants from the two secondary signals
200 in *IKZF1* and *CDKN2A/B* that were previously reported; for which we used the corresponding
201 effect sizes from conditional analysis), and (2) a PRS including the novel hits (A total of 19 loci
202 and 23 SNPs, including the additional 3 novel loci and 2 novel conditional associations). The
203 lead SNPs in known loci included in both models were from this study. PRS was computed using
204 PLINK (v2.0) where the effect size of each risk allele was multiplied by the genotype dosage and
205 summed to generate a score for each sample. Associations between PRS and case-control status
206 for ALL were tested in each group adjusting for 20 principal components using R. To evaluate
207 the predictive power of PRS, Area Under the receiver operating characteristic Curve (AUC) were
208 calculated using pROC package[32] in R.
209

210 **Heritability Estimates**

211 We estimated heritability ascribable to all post-QC imputed SNPs with MAF ≥ 0.05 in our
212 GWAS data using the genome-wide complex trait analysis software (GCTA)[33]. We followed the
213 GCTA-LDMS approach to estimate heritability from imputed data[34], which recommended
214 stratifying SNPs into bins based on their LD scores and/or minor allele frequency. Using GCTA,
215 we computed the genetic relationship matrix (GRM) of pairs of samples using SNPs in each bin,
216 and used the multiple GRMs as input to obtain a restricted maximum likelihood (REML)
217 estimate of heritability. All individuals in discovery analysis were used for LAT (n=10,288). For
218 computational efficiency and for maintaining a close balance in sample size to the LAT data, we
219 randomly sampled 10,000 NLW GERA controls to be included with all of CCRLP NLW cases

220 and controls (total N = 12,391). We used a prevalence of 4.41×10^{-4} and 4.09×10^{-4} for childhood
221 ALL in LAT and NLW respectively based on data from the Surveillance Research Program,
222 (National Cancer Institute SEER*Stat software version 8.3.8; https://seer.cancer.gov/seerstat) to
223 convert the estimated heritability to the liability scale. Because the NLW are expected to be
224 much better imputed using HRC than LAT, particularly at rare variants, our genome-wide
225 imputed data potentially could be used to partition the contribution of low frequency ($0.01 \leq$
226 $MAF < 0.05$) and common ($MAF \geq 0.05$) variants in NLW population. In this case, we
227 performed GCTA-LDMS analysis in 8 strata: two MAF strata (low frequency and common) by
228 four quartiles of LD score strata.

229 To measure genetic correlation between LAT and NLW, we used SNPs with $MAF \geq 0.05$ in
230 both populations to generate GRM using R as per Mancuso et al.[35] The individuals used in
231 univariate REML for each ethnicity were used for the bivariate analysis (n=22,679). We used
232 imputed dosage data to estimate GRM for each unique pair of ancestry groups as

233 $$A = \frac{1}{m}\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \cdot \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}^t$$

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

236

237 **Investigation of Genetic Architecture**

238 To quantify the extent to which latent causal variants for ALL are shared or population-specific
239 between LAT and NLW, we analyzed our GWAS summary data using the tool PESCA[36].
240 Briefly, PESCA analyzes GWAS summary data from multiple populations jointly to infer the
241 genome-wide proportion of causal variants that are population-specific or population-shared. For
242 computational efficiency, PESCA requires first defining LD blocks that are approximately
243 independent in both populations and assumes that a SNP in a given block is independent from all
244 SNPs in all other blocks. We computed pairwise LD matrix in both NLW and LAT using ~329K
245 directly genotyped SNPs shared in both populations. Then, following Shi et al.[36], we generated
246 the trans-ethnic LD matrix by using the larger r^2 value of the NLW or LAT-specific pairwise LD,
247 and used LDetect [37] to define LD blocks within the transethnic LD matrix. By setting mean LD
248 block size to 200 SNPs and using default parameters, we obtained 1,653 blocks that are
249 approximately independent, which is approximately similar to previous reports in East Asians
250 and Europeans.[36] We then followed Shi et al. to estimate the numbers of population-specific and
251 shared causal SNPs using PESCA[36]. We restricted our analysis to 1.3M SNPs with $MAF > 0.05$,
252 $r^2 < 0.95$, and with summary association statistics available in both NLW and LAT. We first
253 estimated the genome-wide proportion of population-specific and shared causal variants with the
254 heritability estimated above (0.2033 and 0.0413 in NLW and LAT, respectively) using default
255 parameters in PESCA, parallelizing the analysis in groups of 10 LD blocks at a time. Using the
256 estimated genome-wide proportions of population-specific and shared causal variants as prior
257 probabilities, we then estimated the posterior probability of each SNP to be causal in a single
258 population (population-specific) or both populations (shared), and inferred the posterior expected
259 numbers of population-specific/shared causal SNPs in each LD block by summing the per-SNP
260 posterior probabilities of being causal in a single or both populations. Critically, while PESCA is

261 an analysis based on summary statistics and not designed for admixed populations, it can be
262 applied to admixed population such as LAT if in-sample LD is used.[36]

263 **Familial risk per variant**

264 The percentage of familial relative risk (FRR) explained by each genetic variant was calculated
265 as per Schumacher et al[38]5/7/2021 10:41:00 PM . The familial relative risk due to locus k (λ_k) is
266 given by

267 $$\lambda_k = \frac{p_k r_k^2 + q_k}{(p_k r_k + q_k)^2}$$

268 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
269 the estimated per-allele odds ratio from meta-analysis. The percentage of familial relative risk is
270 calculated as $\sum_k \log \lambda_k / \log \lambda_0$ where λ_0 is the observed familial risk to first-degree relatives of
271 ALL cases, assumed to be 3.2 as per Kharazmi et al.

272

273 **RESULTS**

274 **Trans-ethnic Genetic Associations with ALL**

275 We performed a trans-ethnic meta-analysis GWAS for childhood ALL. After quality control
276 filtering, our dataset consisted of 3,482 cases and 72,835 controls (124 cases/ 2,067 controls in
277 AFR, 318 cases/5,017 controls in EAS, 1,878 cases/ 8,410 controls in LAT, 1,162 cases/ 57,341
278 controls in NLW; see Materials and Methods) in total. Compared to the previous trans-ethnic
279 analysis, we included additional controls for NLW, added data from EAS population. In
280 addition, we also tested the association at a total of 7,628,894 SNPs after imputation using the
281 Haplotype Reference Consortium reference panel (HRC v r1.1.2016). We aggregated the
282 summary statistics across the four ethnic groups in a fixed-effect meta-analysis. The genomic
283 control inflation factor was 1.023 (1.022 after removing 16 previously known ALL-associated
284 loci, Table 1), suggesting our meta-analysis was not impacted by uncontrolled confounding due
285 to population stratification (Figure 1). Twelve loci reached genome-wide significance (i.e, $P <$
286 5.0×10^{-8}).
287
288 We first examined the evidence of association at 16 previously published risk loci for ALL[3–
289 8,16,18,27,28] . Across the 16 loci, all replicated at the nominal level ($P < 0.05$) or have a SNP nearby
290 with strong association (Table 1), including nine that reached genome-wide significance. Note
291 that out of the 16 loci, three(8q24.21, *IKZF3*, and *BMI1*) were initially identified and five
292 (*IKZF1, PIP4K2A, ARID5B, CDKN2A, CEBPE*) were previously shown to be replicated using a
293 smaller but largely overlapping subset of this dataset[4,28]. For these loci, our findings here would
294 not necessarily constitute an independent replication. In some cases, the published SNP is not the
295 SNP with the most significant association in our dataset, though usually our top SNP in the locus
296 is in strong LD with the reported SNP (Table 1). Interestingly, at the *C5orf56* locus on 5q31
297 previously identified in an independent cohort of predominantly European ancestry, the reported
298 variant (rs886285) was not nominally significant ($P = 0.63$) in our dataset, presumably due to
299 this SNP being associated with a particular subtype of ALL (HD-ALL)[18]. Another SNP in the
300 same locus approximately 20kb away (rs11741255) but in low LD (r2 = 0.35 in NLW, 0.19 in

301 LAT) was significantly associated with ALL in our data (P = 1.69×10^{-4}) but may reflect a chance
302 association. At another locus (*TLE1* on 5q21), neither the published variant nor our top variant in
303 the locus would be considered significantly associated after correcting for multiple testing of 16
304 SNPs in this analysis (minimum P = 1.06×10^{-2} for rs62579826), possibly due to heterogeneity
305 driven by EAS in which both the published variant and our top variant are monomorphic[24].
306 Therefore, further characterization at these two loci is needed to determine whether there is a
307 reproducible association signal to ALL and if so, which are the more likely causal variants.
308
309 More importantly, we discovered three putatively novel susceptibility loci: one at 6q23 and two
310 at 10q21 (Figure 1). The strongest association signal in 6q23 is at rs9376090 (P= 8.23×10^{-9},
311 OR=1.27) in the intergenic region between *MYB* and *HBS1L* (Figure 2A). This association is
312 mainly driven by NLW presumably due to its large sample size (Supplementary Table S1). In
313 10q21, there were two independent signals that showed genome-wide significance. One locus
314 was identified with the lead SNP rs9415680 (P=7.27×10^{-8}, OR=1.20), within a broad
315 association peak, with apparently long-range LD with SNPs covering *NRBF2, JMJD1C*, and
316 parts of *REEP3* (Figure 2B). The second locus in 10q21 was identified 5Mb away, with lead
317 SNP rs10998283 (P=3.92×10^{-8}, OR=1.15) in an intronic region in *TET1* (Figure 2C). The
318 association signals for both loci in 10q21 were largely driven by LAT. We used the convention
319 of the nearest genes to refer to these loci for the remainder of the manuscript, acknowledging that
320 they may not be the causal genes.
321
322 To replicate the novel associations in independent datasets, we tested the associations of the
323 three novel variants and their LD proxies (with P < 5×10^{-7};n=141) in independent samples from
324 COG/WTCCC NLW, and CCLS LAT and NLW. For the *MYB/HBS1L* locus, which was driven
325 by NLW in the discovery cohort, we replicated the signal in COG/WTCCC cohort (rs9376090,
326 $P_{COG} = 4.87 \times 10^{-3}$, $P_{COG+discovery\ analysis} = 1.23 \times 10^{-10}$; Supplementary Table S2), but did not replicate
327 in CCLS likely owing to the small sample size of NLW. For the *TET1* locus, in which the
328 original association was driven by LAT in the discovery, three of the four SNPs with P < 5×10^{-7}
329 in the discovery cohort nominally replicated in CCLS. The lead SNP after meta-analyzing the
330 discovery cohort and the replication cohort of CCLS was rs79226025 ($P_{CCLS} = 3.04 \times 10^{-2}$,
331 $P_{CCLS+discovery} = 6.81 \times 10^{-9}$; Supplementary Table S2). For the *NRBF2* / *JMJD1C* locus, we did
332 not observe an association in the replication cohorts.
333
334 **Novel secondary signals at known loci**

335 We performed conditional analyses adjusting for the lead SNP at each of the known and novel
336 loci to test for independent secondary associations contributing to ALL risk. Using a genome-
337 wide significance threshold of P<5×10^{-8}, we identified a secondary signal in four out of the 16
338 previously known loci (Table 2, Figure 3). In all cases, the LD between the secondary hit and the
339 top hit in the locus are low (Table 2). The additional second associations in *CDKN2A* and *IZKF1*
340 loci were previously noted[18]. In *CEBPE* (rs60820638, P=5.38×10^{-8}) and 17q12 (rs12944882,
341 P=7.71×10^{-10}), these secondary signals represent novel associations. In particular, at the *CEBPE*
342 locus, previous reports suggest multiple correlated variants with functional evidence[39,40] . Our
343 analysis is consistent with the two previous variants (rs2239635 and rs2239630) being or tagging
344 the same underlying signal, while the secondary association we identified (rs60820638) is an
345 independent association (Supplementary Table S3). In the case of 17q12 the lead SNP in our

346 discovery GWAS is the novel association, as the conditional analysis revealed the previously
347 known SNP at the locus. The LD between the two associated SNPs is low (r2 = 0.02). None of
348 the loci tested showed residual association surpassing genome-wide significance threshold after
349 conditioning on the first or second associated variants, and none of the three putatively novel loci
350 showed additional association in conditional analysis.

351

352 **Polygenic Risk Score**

353

354 To assess the combined effect of all identified risk alleles for ALL, we constructed an ALL PRS
355 model in our discovery cohort, using either the 18 SNPs from 16 previously known loci or the 23
356 known plus novel SNPs (including the novel conditional hits) and their associated effect sizes
357 from the trans-ethnic meta-analysis in CCRLP/GERA. We then computed the PRS for NLW and
358 LAT individuals in CCLS and COG, as they are the largest ethnic groups available across CCLS,
359 COG/WTCCC, and our discovery cohort. The scores generated with the known risk loci were
360 significantly associated with case-control status in all groups ($P_{CCLS\ NLW}=2.22x10^{-17}$, P_{CCLS}
361 $_{LAT}=4.78x10^{-23}$, $P_{COG/WTCCC}=2.99x10^{-62}$, Supplementary Table S4). Adding the three novel loci
362 identified in this study and the two novel secondary signals further strengthened the evidence of
363 the association in COG/WTCCC ($P=6.93x10^{-63}$) and CCLS LAT ($P= 5.75\ x10^{-24}$), while the
364 evidence of association stayed about the same in CCLS NLW ($P=2.03\ x10^{-17}$). The predictive
365 accuracy as measured by AUC are similar between NLW and LAT, at around 67-68%, consistent
366 with the hypothesis that trans-ethnic meta-analysis will enable PRS to be more transferrable
367 between populations.

368

369 Also, we explored the distribution of PRS in CCRLP individuals (Supplementary Figure S1). We
370 found that while the shape of the PRS distribution is consistent with a normal distribution
371 (Kolmogorov-Smirnov P = 0.918 and 0.303 for LAT and NLW, respectively) and appears
372 similar between LAT and NLW (standard deviation of 0.728 and 0.735 respectively; F-Test P =
373 0.633), the scores in LAT are shifted to the right compared to the scores in NLW (mean of 5.101
374 and 4.641 respectively, Welch t-test $P = 1.3x10^{-122}$). The observed pattern was consistent when
375 the scores were stratified by case-control status (mean of 5.324 and 4.881 in LAT and NLW
376 cases, respectively, $P=3.956\ x10^{-58}$; mean of 4.895 and 4.414 in LAT and NLW controls,
377 respectively, with $P= 1.493\ x10^{-78}$). This observation was also replicated in CCLS with mean of
378 5.119 in LAT and 4.607 in NLW ($P=4.596\ x10^{-51}$). Therefore, results from our PRS analyses are
379 consistent with the notion that differences in allele frequency of ALL risk loci between
380 populations may complement other non-genetic factors for ALL risk, and partly explain the
381 increased ALL risk in LAT relative to NLW children and LAT.

382

383 **Genetic architecture of ALL in Latinos and non-Latino whites**

384

385 We estimated the relative contributions of each variant to ALL risk by computing the familial
386 relative risk (Supplementary Table 5). In our discovery cohort (CCRLP/GERA), the known risk
387 variants account for 22.9% and 21.7% familial relative risk in LAT and NLW respectively, while
388 the addition of the novel variants increased this estimate to 25.9% and 24.6%, respectively. In
389 CCLS, where effect size estimates are expected to be less biased by winner's curse, the pattern is
390 similar: 22.7% and 23.2% explained in LAT and NLW using known risk variants, and 24.3%

391 and 24.8% explained in LAT and NLW using both known and novel variants (Supplemental
392 Table S5).
393
394 We also estimated the heritability of ALL attributable to all common SNPs (MAF ≥ 0.05) using
395 the GCTA-LDMS framework.[34] We estimated the heritability to be 20.3% in NLW and 4.1% in
396 LAT (Supplemental Table S6A). The low estimate in LAT remained with using only the CCRLP
397 cohort (Supplemental Table S6C). The low estimates could be unreliable because of the admixed
398 nature of LAT (see Discussion). The estimated heritability in NLW is consistent with that
399 previous reported[18]. Because the imputation quality using HRC reference panel is expected to be
400 high for low frequency variants between 1-5% MAF in NLW, our dataset also provides the
401 opportunity to estimate the stratified contribution by allele frequency to the heritability of ALL
402 in NLW. When low frequency variants are included, the estimated heritability in NLW increased
403 to 29.8% (16.2% due to common variants, 13.5% due to low frequency variants; Supplemental
404 Table S6B). Furthermore, we sought to measure the genetic correlation for ALL between NLW
405 and LAT, using bivariate GREML analysis implemented in GCTA. The genetic correlation was
406 high (rG = 0.714 ± standard error 0.130) but significantly different from 1 (P = 0.014,
407 Supplementary Table 7), indicating the genetic architectures may be similar as expected from
408 correlated effect sizes (Supplementary Figure S2) but not perfectly concordant. Taken at face
409 value, this may suggest that there may be ethnic-specific genetic risk profiles or differential
410 interactions with the environment that contributes to differences in disease risk between NLW
411 and LAT, although it may be subject to similar concerns as the low univariate estimate due to
412 admixture.
413
414 Finally, we further characterized the genetic architecture of ALL in NLW and LAT by
415 estimating the number of population-specific and shared causal alleles using the program
416 PESCA[36]. The PESCA framework defines the set of causal variants as all variants tested to have
417 a non-zero effect, even if the effect is indirect and only statistical rather than biological in nature.
418 Using this framework, we estimated that approximately 1.71% of common SNPs have nonzero
419 effects in both NLW and LAT, and that 1.69% and 1.87% had population-specific nonzero
420 effects in NLW and LAT, respectively. By dividing the genome into 1,653 approximately
421 independent LD regions (Methods), the mean expected numbers of population-specific/shared
422 causal SNPs were 13.86 (SD 7.23) NLW-specific, 15.32 (SD 8.00) LAT-specific, and 14.02 (SD
423 7.32) shared by the two populations (Supplementary Figure S3).
424
425
426 **DISCUSSION**
427
428 In summary, we have performed the largest trans-ethnic meta-analysis GWAS of childhood ALL
429 to date. We incorporated data across four ethnic groups (2,191 AFR, 5,335 EAS, 10,288 LAT,
430 and 58,503 NLW individuals) and identified three putatively novel susceptibility loci. In
431 addition, conditional analyses identified novel secondary associations with ALL in two
432 previously reported loci. Our analysis suggests that the known and novel ALL risk alleles
433 together explained about 25% of the familial relative risk in both NLW and LAT populations,
434 and that the trans-ethnic PRS we constructed, although relatively simple and utilizing only the
435 genome-wide associated variants, performed similarly in both NLW and LAT in predicting ALL
436 (AUC ~ 67-68%).

While the three putatively novel risk loci were not consistently replicated, possibly due to small sample sizes in our replication cohorts and population differences in SNP risk effects, each of the three loci harbors genes and/or variants with reported functions in hematopoiesis and leukemogenesis that support a potential role in ALL etiology. Among the set of significantly associated SNPs ($P_{meta} < 5x10^{-7}$) in the three putatively novel loci, annotations from HaploReg (version 4.1) [41] and GTEx portal[42] provided insights of the biological basis of these novel associations. 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[43] . This locus is already well known for associations with multiple blood cell measurements, severity of major hemoglobin disorders, and β-thalassemia, with multiple putative functional variants reported[44,45]. The *HBS1L-MYB* intergenic variants are also known to reduce transcription factor binding, affect long-range interaction with *MYB*, and impact *MYB* expression. 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[43] and hemoglobin concentration[46,47]. Also, it is an eQTL for *ALDH8A1in* lymphocytes and whole blood[42]. *ALDH8A1* 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. 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 S4). 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[48,49]. It is also associated with IL-10 levels[50] which was shown to be in deficit in ALL cases[51]. 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[52]. Previous studies have found that it acts as a coactivator for key transcription factors to ensure survival of AML cells[53] and self-renewal of mouse embryonic stem cells[54]. Moreover, several significantly associated variants in this region found in our study alter binding motif of some important transcription factors such as Pax5 and TCF12(ref.[41]).

The second locus in 10q21 contains intronic variants in the *TET1* gene, which is well known for its oncogenicity in several malignancies including AML[55]. 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[56]. We investigated whether there is a stronger association of this locus with T-ALL with subtype-stratified analysis using a subset of cases in CCRLP with subtype information. The locus appears to be associated in both T-ALL and B-ALL and we observed a trend of slightly higher association for T-ALL, but the difference is not statistically significant, likely due to only a small number of T-ALL cases available (Supplemental Table S8). Although the associated locus overlaps *TET1*, the most associated SNPs in our meta-analysis were observed as eQTLs of *DNA2, SLC25A16, STOX1* in whole blood or lymphoblastoid cells[57]; the role of these genes in leukemia has not been explored, and we also cannot verify whether the SNPs may affect *TET1* expression in hematopoietic stem or progenitor cells. 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 S5). Further, rs58627364 is

483 reported to bind to POL2, CCNT2, ETS1, HEY1, and HMGN3 in K562 cells in ENCODE data,
484 and also significantly alters BCL, BDP1 and NF-kappa B binding motifs. The second locus in
485 10q21 contains intronic variants in the *TET1* gene, which is well known for its oncogenicity in
486 several malignancies including AML[55]. A recent study showed the epigenetic regulator *TET1* is
487 highly expressed in T-cell ALL and is crucial for human T-ALL cell growth in vivo[56]. We
488 investigated whether there is a stronger association of this locus with T-ALL with subtype
489 stratified analysis using a small number of cases in CCRLP with subtype information available
490 (Supplemental Table S8). The locus appears to be associated with both T-ALL and B-ALL and
491 we observed a trend of slightly higher association for T-ALL, but the difference is not
492 statistically significant, possibly due to a small number of T-ALL cases.
493
494 In addition to the discovery of putative novel ALL risk loci, we also capitalized on the large
495 numbers of Latinos and non-Latino whites included in our study to explore the genetic
496 architecture of ALL in these two populations, by estimating the heritability explained and the
497 proportion of population-specific and shared causal alleles. In the NLW population, we estimated
498 that 20.3% of the heritability of ALL was attributed to common imputed variants using GCTA-
499 LDMS (Supplementary Table S6A). The estimated heritability increased to 29.8%
500 (Supplementary Table S6B) if low frequency variants down to MAF of 1% were included,
501 suggesting that there are additional low frequency variants associated with ALL that may be
502 discovered in larger scaled studies. However, it is puzzling that the estimated heritability was so
503 low in LAT (4.1% in univariate analysis, and 7.3% in bivariate analysis; Supplementary Table
504 S6A, 7). Because the accuracy of imputation at low frequency variants is expected to be
505 significantly worse in LAT (due to the underrepresentation of Native American or other non-
506 European haplotypes in HRC panel), a priori we did not estimate the heritability including low-
507 frequency variants in LAT. However, when we did attempt to estimate heritability in this setting,
508 we obtain a strongly negative heritability estimates, suggesting blatant model instability or
509 misspecification, likely attributed to the admixed nature of LAT[58]. This thus brings in question
510 whether the 4-7% estimates from common variant analysis are robust and accurate. Because the
511 estimated familial relative risk summed over GWAS loci between NLW and LAT are similar
512 (Supplementary Table S5), and the estimated effect sizes at known associated loci are highly
513 correlated (r2 = 0.819; Supplementary Figure S2), we suspect the heritability in Latino
514 population may be underestimated with the standard REML approach. Our observation suggests
515 that further efforts to develop a framework to robustly estimate heritability in complex admixed
516 populations are needed, such as extending model proposed in Zaitlin et al.[59] to 3-way admixed
517 scenarios.
518
519 As a consequence of the potential bias in heritability estimates in LAT, our estimated genetic
520 correlation between LAT and NLW should be interpreted with caution. On the surface, our
521 estimated genetic correlation ($r_G = 0.71$) is significantly less than 1, suggesting that there are
522 significant population-specific components of the disease architecture between LAT and NLW.
523 This would be consistent with the findings of the *ERG* locus[15,16], a Latino-specific locus
524 associated with ALL, and suggest that future ethnic-specific GWAS across different ethnic
525 groups for ALL will be insightful. This is also consistent with our observation in the PESCA
526 analysis, where we estimated the population-specific and shared causal alleles. We found that
527 only 32.5% of the estimated causal alleles are shared between LAT and NLW. This contrasts
528 starkly with the previous investigation of nine complex traits using PESCA between East Asian

529 and European populations, where most of the causal alleles are shared between populations
530 across all 9 traits. The result of the PESCA analysis could be partly explained by drastic
531 difference in power between LAT and NLW currently in our study, and that generally the sample
532 sizes for ALL are still significantly lower than the complex traits examined in Shi et al.[36]
533 Another factor could be that because the heritability estimates are used to inform PESCA
534 analysis, a bias in heritability estimates could impact the estimated number of population-
535 specific and shared causal alleles. Therefore, the genetic architecture for ALL, particularly in
536 admixed populations like the Latinos, remain unclear and will require more focused effort to
537 investigate.
538
539 Future studies aimed to uncover the genetic risk factors for ALL could focus on multiple
540 avenues. First, there will be a need to further increase the sample size of the study cohort, which
541 would provide additional venues to replicate the putative novel findings here and identify more
542 associated loci. Second, there should be a focus on ethnic-specific GWAS for ALL, as ethnic-
543 specific associations could be missed in a trans-ethnic GWAS. An example is the *ERG* locus,
544 which is not genome-wide significant in our meta-analysis. Finally, while not explored
545 extensively in this particular study, there should be a focus on disentangling the different
546 subtypes of ALL, and to study other aspects of the disease pathogenesis such as disease
547 progression or risk of relapse, though these data are less available and may require more focused
548 ascertainment and cohort creation.
549
550
551 **DESCRIPTION OF SUPPLEMENTAL DATA**
552 Supplemental Data include five figures and eight tables
553
554 **DATA AVAILABILITY**
555 Any uploading of genomic data and/or sharing of these biospecimens or individual data derived
556 from these biospecimens has been determined to violate the statutory scheme of the California
557 Health and Safety Code Sections 124980(j), 124991(b), (g), (h), and 103850 (a) and (d), which
558 protect the confidential nature of biospecimens and individual data derived from biospecimens.
559 This study was approved by Institutional Review Boards at the California Health and Human
560 Services Agency, University of Southern California, Yale University, and the University of
561 California San Francisco. The de-identified newborn dried blood spots for the CCRLP
562 (California Biobank Program SIS request # 26) were obtained with a waiver of consent from the
563 Committee for the Protection of Human Subjects of the State of California.
564
565 **ACKNOWLDGEMENT**
566 This work was supported by research grants from the National Institutes of Health
567 (R01CA155461, R01CA175737, R01ES009137, P42ES004705, P01ES018172, P42ES0470518
568 and R24ES028524) and the Environmental Protection Agency (RD83451101), United States.
569 The content is solely the responsibility of the authors and does not necessarily represent the
570 official views of the National Institutes of Health and the EPA. The collection of cancer
571 incidence data used in this study was supported by the California Department of Public Health as
572 part of the statewide cancer reporting program mandated by California Health and Safety Code
573 Section 103885; the National Cancer Institute's Surveillance, Epidemiology and End Results
574 Program under contract HHSN261201000140C awarded to the Cancer Prevention Institute of

575 California, contract HHSN261201000035C awarded to the University of Southern California,
576 and contract HHSN261201000034C awarded to the Public Health Institute; and the Centers for
577 Disease Control and Prevention's National Program of Cancer Registries, under agreement
578 U58DP003862-01 awarded to the California Department of Public Health. The biospecimens
579 and/or data used in this study were obtained from the California Biobank Program, (SIS request
580 #26), Section 6555(b), 17 CCR. The California Department of Public Health is not responsible
581 for the results or conclusions drawn by the authors of this publication. We thank Hong Quach
582 and Diana Quach for DNA isolation support. We thank Martin Kharrazi, Robin Cooley, and
583 Steve Graham of the California Department of Public Health for advice and logistical support.
584 We thank Eunice Wan, Simon Wong, and Pui Yan Kwok at the UCSF Institute of Human
585 Genetics Core for genotyping support. This study makes use of data generated by the Wellcome
586 Trust Case–Control Consortium. A full list of the investigators who contributed to the generation
587 of the data is available from www.wtccc.org.uk. Funding for the project was provided by the
588 Wellcome Trust under award 076113 and 085475. Genotype data for COG ALL cases are
589 available for download from dbGaP (Study Accession: phs000638.v1.p1). Data came from a
590 grant, the Resource for Genetic Epidemiology Research in Adult Health and Aging (RC2
591 AG033067; Schaefer and Risch, PIs) awarded to the Kaiser Permanente Research Program on
592 Genes, Environment, and Health (RPGEH) and the UCSF Institute for Human Genetics. The
593 RPGEH was supported by grants from the Robert Wood Johnson Foundation, the Wayne and
594 Gladys Valley Foundation, the Ellison Medical Foundation, Kaiser Permanente Northern
595 California, and the Kaiser Permanente National and Northern California Community Benefit
596 Programs. The RPGEH and the Resource for Genetic Epidemiology Research in Adult Health
597 and Aging are described here:
598 https://divisionofresearch.kaiserpermanente.org/genetics/rpgeh/rpgehhome. For recruitment of
599 subjects enrolled in the CCLS replication set, the authors gratefully acknowledge the clinical
600 investigators at the following collaborating hospitals: University of California Davis Medical
601 Center (Dr. Jonathan Ducore), University of California San Francisco (Drs. Mignon Loh and
602 Katherine Matthay), Children's Hospital of Central California (Dr. Vonda Crouse), Lucile
603 Packard Children's Hospital (Dr. Gary Dahl), Children's Hospital Oakland (Dr. James Feusner),
604 Kaiser Permanente Roseville (formerly Sacramento) (Drs. Kent Jolly and Vincent Kiley), Kaiser
605 Permanente Santa Clara (Drs. Carolyn Russo, Alan Wong, and Denah Taggart), Kaiser
606 Permanente San Francisco (Dr. Kenneth Leung), and Kaiser Permanente Oakland (Drs. Daniel
607 Kronish and Stacy Month). The authors additionally thank the families for their participation in
608 the California Childhood Leukemia Study (formerly known as the Northern California
609 Childhood Leukemia Study). Finally, the authors acknowledge the Center for Advanced
610 Research Computing (CARC; https://carc.usc.edu) at the University of Southern California for
611 providing computing resources that have contributed to the research results reported within this
612 publication.
613

614 **DECLARATION OF INTEREST**

615 The authors declare no competing interests.

616 **Table 1**. Summary statistics for the reported variants, the top variant in the loci from our meta-
617 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 [18])	0.63	5	131811182	rs11741255	1.69×10^{-4}	0.35	0.19
BAK1	6	3354693	rs1048728818 (ref [18])	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 [18])	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 [18])	2.93×10^{-2}	17	47217004	rs6504598	4.87×10^{-4}	0.02	0.02
ERG	21	39789606	rs8131436 (ref [16])	6.97×10^{-5}	21	39784752	rs55681902	9.36×10^{-6}	0.62	0.65

618 Definition of terms: Reported SNP; variant associated with ALL in previous papers. Out top
619 SNP: variant with lowest P-value at 1Mb around the reported SNP in the transethnic meta-
620 analysis; Gene: nearest gene unless the variant is in gene desert. Chr., chromosome; Pos.,
621 Position in hg19; P-value, logistic regression test value, r2: squared correlation of the reported
622 SNP and our top SNP; NLW: non-Latino white cohort; LAT: Latino cohort

623 **Table 2.** Summary of conditional analysis to identify secondary associations at known loci.
624

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

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

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

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

634 **REFERENCES**

635 1. Giddings, B. M., Whitehead, T. P., Metayer, C. & Miller, M. D. Childhood leukemia incidence in
636 California: High and rising in the Hispanic population: Hispanic Childhood Leukemia Incidence.
637 *Cancer* **122**, 2867–2875 (2016).

638 2. Lim, J. Y.-S., Bhatia, S., Robison, L. L. & Yang, J. J. Genomics of racial and ethnic disparities in
639 childhood acute lymphoblastic leukemia. *Cancer* **120**, 955–962 (2014).

640 3. Vijayakrishnan, J. *et al.* A genome-wide association study identifies risk loci for childhood acute
641 lymphoblastic leukemia at 10q26.13 and 12q23.1. *Leukemia* **31**, 573–579 (2017).

642 4. Wiemels, J. L. *et al.* GWAS in childhood acute lymphoblastic leukemia reveals novel genetic
643 associations at chromosomes 17q12 and 8q24.21. *Nat. Commun.* **9**, 286 (2018).

644 5. Papaemmanuil, E. *et al.* Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of
645 childhood acute lymphoblastic leukemia. *Nat. Genet.* **41**, 1006–1010 (2009).

646 6. Perez-Andreu, V. *et al.* Inherited GATA3 variants are associated with Ph-like childhood acute
647 lymphoblastic leukemia and risk of relapse. *Nat. Genet.* **45**, 1494–1498 (2013).

648 7. Treviño, L. R. *et al.* Germline genomic variants associated with childhood acute lymphoblastic
649 leukemia. *Nat. Genet.* **41**, 1001–1005 (2009).

650 8. Xu, H. *et al.* Novel susceptibility variants at 10p12.31-12.2 for childhood acute lymphoblastic
651 leukemia in ethnically diverse populations. *J. Natl. Cancer Inst.* **105**, 733–742 (2013).

652 9. Feng, Q. *et al.* Trends in Acute Lymphoblastic Leukemia Incidence in the United States by
653 Race/Ethnicity From 2000 to 2016. *Am. J. Epidemiol.* kwaa215 (2020)
654 doi:10.1093/aje/kwaa215.

655 10. Linabery, A. M. & Ross, J. A. Trends in childhood cancer incidence in the U.S. (1992-
656 2004). *Cancer* **112**, 416–432 (2008).

657 11. Barrington-Trimis, J. L. *et al.* Trends in childhood leukemia incidence over two decades
658 from 1992 to 2013. *Int. J. Cancer* **140**, 1000–1008 (2017).

659 12. Bhatia, S. *et al.* Racial and ethnic differences in survival of children with acute
660 lymphoblastic leukemia. *Blood* **100**, 1957–1964 (2002).

661 13. Kadan-Lottick, N. S. Survival Variability by Race and Ethnicity in Childhood Acute
662 Lymphoblastic Leukemia. *JAMA* **290**, 2008 (2003).

663 14. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164
664 (2016).

665 15. Qian, M. *et al.* Novel susceptibility variants at the ERG locus for childhood acute
666 lymphoblastic leukemia in Hispanics. *Blood* **133**, 724–729 (2019).

667 16. de Smith, A. J. *et al.* Heritable variation at the chromosome 21 gene ERG is associated
668 with acute lymphoblastic leukemia risk in children with and without Down syndrome. *Leukemia*
669 **33**, 2746–2751 (2019).

670 17. Walsh, K. M. *et al.* A Heritable Missense Polymorphism in *CDKN2A* Confers Strong Risk
671 of Childhood Acute Lymphoblastic Leukemia and Is Preferentially Selected during Clonal
672 Evolution. *Cancer Res.* **75**, 4884–4894 (2015).

673 18. Vijayakrishnan, J. *et al.* Identification of four novel associations for B-cell acute
674 lymphoblastic leukaemia risk. *Nat. Commun.* **10**, 5348 (2019).

675 19. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health
676 disparities. *Nat. Genet.* **51**, 584–591 (2019).

677 20. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery for
678 complex traits. *Nature* **570**, 514–518 (2019).

679 21. Mahajan, A. *et al. Trans-ancestry genetic study of type 2 diabetes highlights the power*
680 *of diverse populations for discovery and translation*.
681 http://medrxiv.org/lookup/doi/10.1101/2020.09.22.20198937 (2020)
682 doi:10.1101/2020.09.22.20198937.

683 22. The 1000 Genomes Project Consortium. A global reference for human genetic variation.
684 *Nature* **526**, 68–74 (2015).

685 23. the Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for
686 genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).

687 24. Genome Aggregation Database Consortium *et al.* The mutational constraint spectrum
688 quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

689 25. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of
690 genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).

691 26. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method
692 for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913
693 (2007).

694 27. Vijayakrishnan, J. *et al.* The 9p21.3 risk of childhood acute lymphoblastic leukaemia is
695 explained by a rare high-impact variant in CDKN2A. *Sci. Rep.* **5**, 15065 (2015).

696 28. de Smith, A. J. *et al.* BMI1enhancer polymorphism underlies chromosome 10p12.31
697 association with childhood acute lymphoblastic leukemia: *BMI* 1 enhancer polymorphism in ALL.
698 *Int. J. Cancer* **143**, 2647–2658 (2018).

699 29. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000
700 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).

701 30. Metayer, C. *et al.* Tobacco smoke exposure and the risk of childhood acute
702 lymphoblastic and myeloid leukemias by cytogenetic subtype. *Cancer Epidemiol. Biomark. Prev.*
703 *Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol.* **22**, 1600–1611 (2013).

704 31. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer
705 datasets. *GigaScience* **4**, 7 (2015).

706 32. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare
707 ROC curves. *BMC Bioinformatics* **12**, 77 (2011).

708 33. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A Tool for Genome-wide
709 Complex Trait Analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

710 34. The LifeLines Cohort Study *et al.* Genetic variance estimation with imputed variants finds
711 negligible missing heritability for human height and body mass index. *Nat. Genet.* **47**, 1114–
712 1120 (2015).

713 35. the PRACTICAL consortium *et al.* The contribution of rare variation to prostate cancer
714 heritability. *Nat. Genet.* **48**, 30–35 (2016).

715 36. Shi, H. *et al.* Localizing Components of Shared Transethnic Genetic Architecture of
716 Complex Traits from GWAS Summary Data. *Am. J. Hum. Genet.* **106**, 805–817 (2020).

717 37. Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in
718 human populations. *Bioinformatics* btv546 (2015) doi:10.1093/bioinformatics/btv546.

719 38. Schumacher, F. R. *et al.* Association analyses of more than 140,000 men identify 63
720 new prostate cancer susceptibility loci. *Nat. Genet.* **50**, 928–936 (2018).

721 39. Wiemels, J. L. *et al.* A functional polymorphism in the CEBPE gene promoter influences
722 acute lymphoblastic leukemia risk through interaction with the hematopoietic transcription factor
723 Ikaros. *Leukemia* **30**, 1194–1197 (2016).

724 40. Studd, J. B. *et al.* Genetic predisposition to B-cell acute lymphoblastic leukemia at
725 14q11.2 is mediated by a CEBPE promoter polymorphism. *Leukemia* **33**, 1–14 (2019).

726 41. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states,
727 conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic*
728 *Acids Res.* **40**, D930-934 (2012).

729 42. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**,
730 580–585 (2013).

731 43. Lin, B. D. *et al.* 2SNP heritability and effects of genetic variants for neutrophil-to-
732 lymphocyte and platelet-to-lymphocyte ratio. *J. Hum. Genet.* **62**, 979–988 (2017).

733 44. Stadhouders, R. *et al.* HBS1L-MYB intergenic variants modulate fetal hemoglobin via
734 long-range MYB enhancers. *J. Clin. Invest.* **124**, 1699–1710 (2014).

735 45. Guo, M. H. *et al.* Comprehensive population-based genome sequencing provides insight
736 into hematopoietic regulatory mechanisms. *Proc. Natl. Acad. Sci.* **114**, E327–E336 (2017).

737 46. Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to
738 Common Complex Disease. *Cell* **167**, 1415-1429.e19 (2016).

739 47. van Rooij, F. J. A. *et al.* Genome-wide Trans-ethnic Meta-analysis Identifies Seven
740 Genetic Loci Influencing Erythrocyte Traits and a Role for RBPMS in Erythropoiesis. *Am. J.*
741 *Hum. Genet.* **100**, 51–63 (2017).

742 48. Tajuddin, S. M. *et al.* Large-Scale Exome-wide Association Analysis Identifies Loci for
743 White Blood Cell Traits and Pleiotropy with Immune-Mediated Diseases. *Am. J. Hum. Genet.*
744 **99**, 22–39 (2016).

745 49. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association
746 studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012
747 (2019).

748 50. Ahola-Olli, A. V. *et al.* Genome-wide Association Study Identifies 27 Loci Influencing
749 Concentrations of Circulating Cytokines and Growth Factors. *Am. J. Hum. Genet.* **100**, 40–50
750 (2017).

751 51. Chang, J. S. *et al.* Profound deficit of IL10 at birth in children who develop childhood
752 acute lymphoblastic leukemia. *Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res.*
753 *Cosponsored Am. Soc. Prev. Oncol.* **20**, 1736–1740 (2011).

754 52. Lynch, J. R. *et al.* JMJD1C-mediated metabolic dysregulation contributes to HOXA9-
755 dependent leukemogenesis. *Leukemia* **33**, 1400–1410 (2019).

756 53. Chen, M. *et al.* JMJD1C is required for the survival of acute myeloid leukemia by
757 functioning as a coactivator for key transcription factors. *Genes Dev.* **29**, 2123–2139 (2015).

758 54. Xiao, F. *et al.* JMJD1C Ensures Mouse Embryonic Stem Cell Self-Renewal and Somatic
759 Cell Reprogramming through Controlling MicroRNA Expression. *Stem Cell Rep.* **9**, 927–942
760 (2017).

761 55. Cimmino, L. *et al.* TET1 is a tumor suppressor of hematopoietic malignancy. *Nat.*
762 *Immunol.* **16**, 653–662 (2015).

763 56. Bamezai, S. *et al.* TET1 promotes growth of T-cell acute lymphoblastic leukemia and
764 can be antagonized via PARP inhibition. *Leukemia* (2020) doi:10.1038/s41375-020-0864-3.

765 57. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional
766 variation in humans. *Nature* **501**, 506–511 (2013).

767 58. Steinsaltz, D., Dahl, A. & Wachter, K. W. On Negative Heritability and Negative
768 Estimates of Heritability. *Genetics* **215**, 343–357 (2020).

769 59. Zaitlen, N. *et al.* Leveraging population admixture to characterize the heritability of
770 complex traits. *Nat. Genet.* **46**, 1356–1362 (2014).

771

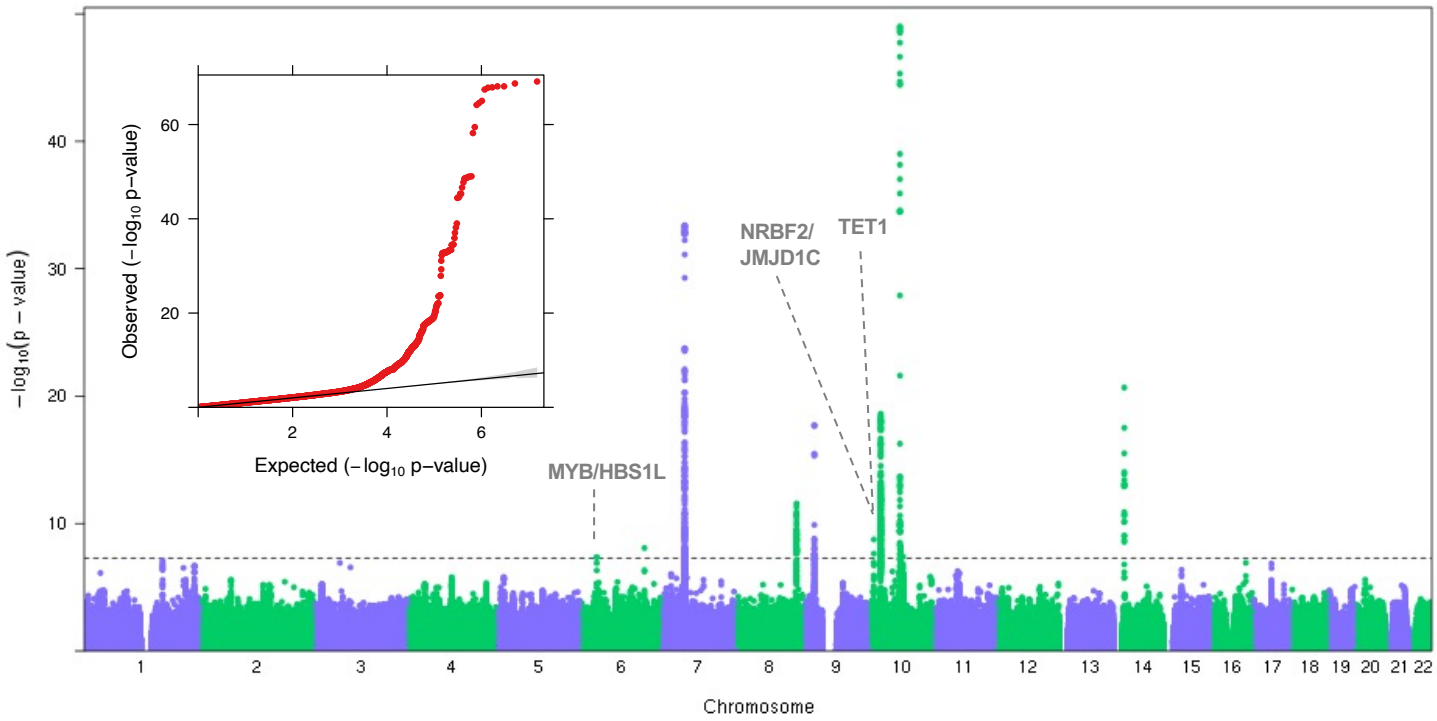

Figure1. Summary result of the trans-ethnic meta-analysis on ALL.
Results of the meta-analysis is represented by the Manhattan plot. The novel loci from this study are marked with dotted lines and labeled with the nearest genes. Significance threshold at genome-wide significance level ($5x10^{-8}$) is 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.

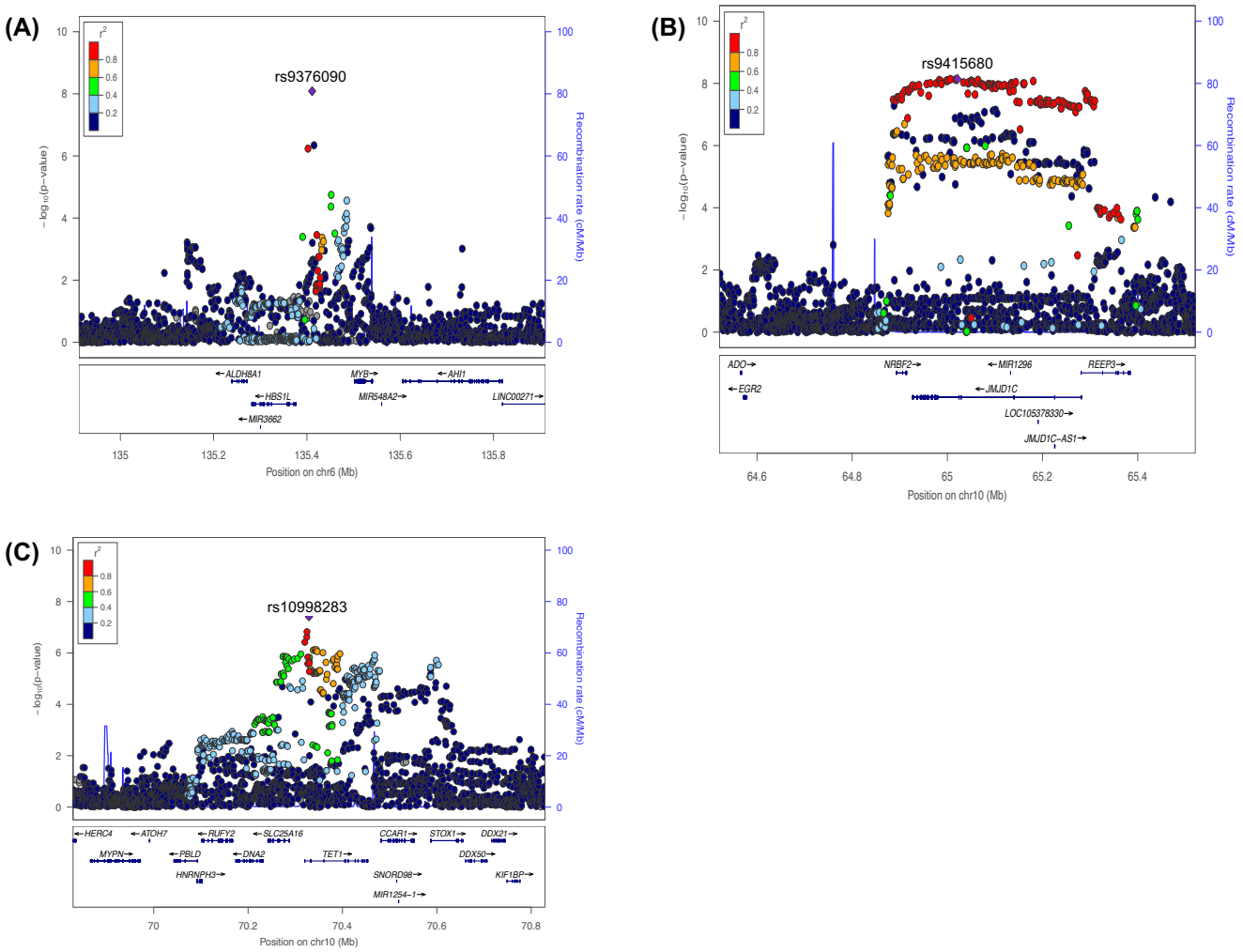

Figure2. Novel loci associated with childhood ALL in trans-ethnic meta-analysis. LocusZoom plots showing 1 Mb region around the identified loci near (A) MYB/HBS1L on chr6, (B) NRBF2/JMJD1C on chr10, and (C) TET1 on chr10 are shown. Diamond symbol indicates the lead SNP in each locus. Color of remaining 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.

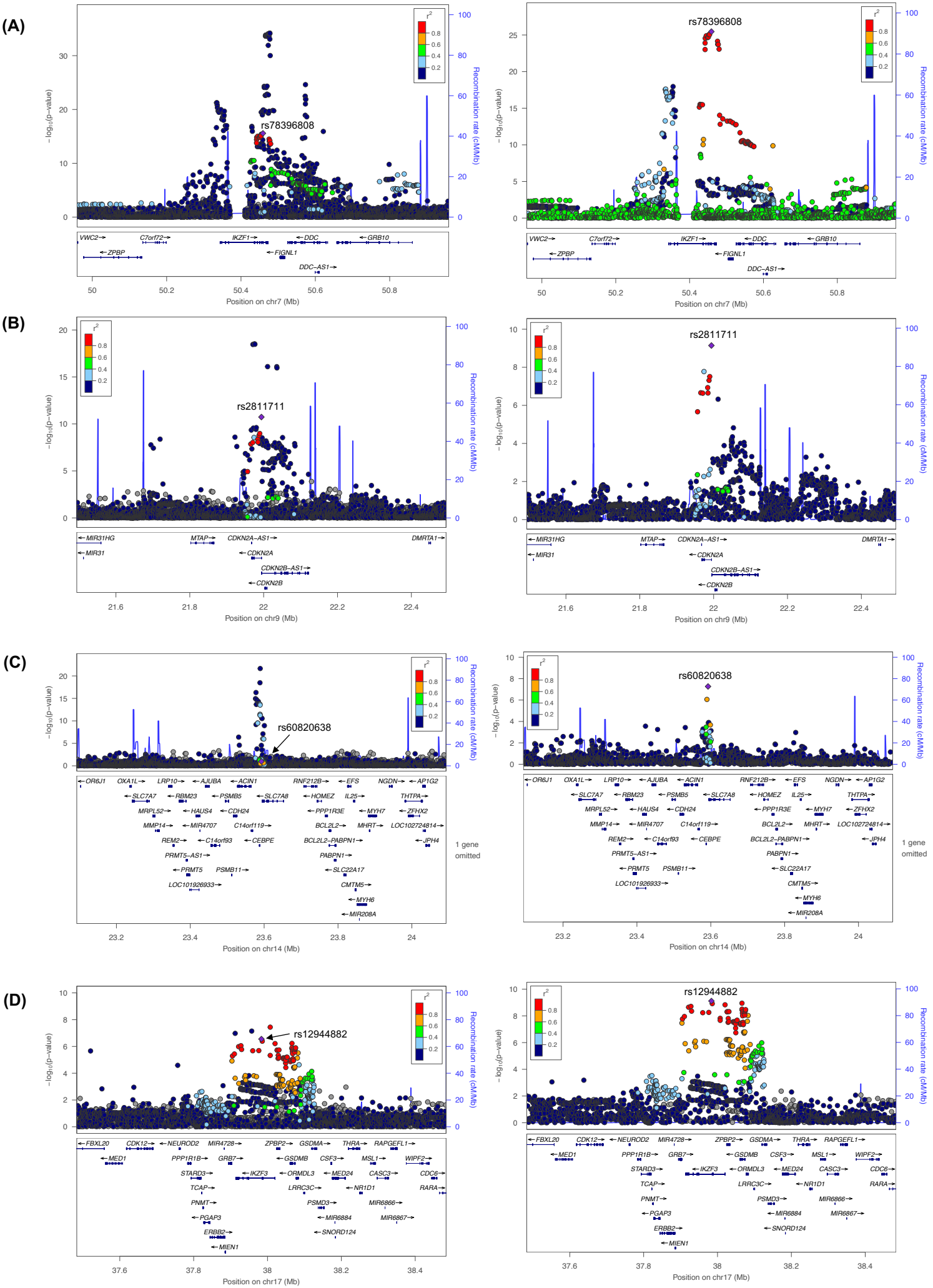
(A)
rs78396808
rs78396808
(B)
rs2811711
rs2811711
(C)
rs60820638
rs60820638
1 gene omitted
1 gene omitted
(D)
rs12944882
rs12944882
$-\log_{10}$(p-value)
Recombination rate (cM/Mb)
r^2
0.8
0.6
0.4
0.2
VWC2
C7orf72
IKZF1
DDC
GRB10
ZPBP
FIGNL1
DDC-AS1
Position on chr7 (Mb)
MIR31HG
MTAP
CDKN2A-AS1
DMRTA1
MIR31
CDKN2A
CDKN2B-AS1
CDKN2B
Position on chr9 (Mb)
Position on chr14 (Mb)
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
Position on chr17 (Mb)

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.