Novel Susceptibility Variants at the *ERG* Locus for Childhood Acute Lymphoblastic Leukemia in Hispanics

Supplementary Text, Tables and Figures for Maoxiang Qian *et al.*

Supplementary Text

Genetic ancestry

Genetic ancestry was determined for all the case and control subjects (i.e., individuals from COG AALL0232 [clinicaltrials.gov NCT00075725], COG P9904/P9905/P9906 [NCT00005585/NCT00005596/NCT00005603], St. Jude Total Therapy XIIIB/XV [NCI-T93-0101D/NCT00137111], MESA [dbGAP phs000209.v9], GALA, and Guatemala) by using Admixture (version 1.3.0),[1] on the basis of genotypes at 30,000 SNPs randomly selected from the Affymetrix SNP arrays. HapMap samples from descendants of Northern Europeans (CEU, N = 60), West Africans (YRI, N = 60), East Asians (CHB/JPT, N = 90) and Native American references[2] (N = 105) were used to represent European, African, East Asian, and Native American ancestries, respectively. We assumed that these 4 ancestries summed to 100% in each genotyped individuals. Hispanics are generally considered as an admixed population: two-way admixture between Native American and European populations or three-way admixture amongst Native American, European, and African ancestries.[2] Therefore, European American (EA), African American (AA), and Asians were defined as having >95% European genetic ancestry, >70% African ancestry, and >90% Asian ancestry, respectively. Hispanics were individuals for whom Native American ancestry was >10% and greater than African ancestry, as previously described.[3] Cutoffs of genetic ancestry used for each race/ethnic group were based on % ancestry distribution in ALL cohorts as reported previously.[4]

Genotyping and quality control

SNP genotyping was performed in germline DNA using the Affymetrix Human SNP Array 6.0 (COG AALL0232, COG P9904/P9905, MESA, GALA, and Guatemala) or the Affymetrix GeneChip Human Mapping 500K Array (COG P9906 and St. Jude Total Therapy XIIIB/XV). Genotype calls were determined as described previously.[3] The sample quality control procedures were performed on the basis of SNP call rate and minor allele frequency

(**Supplemental Fig. 2**). Individuals with one of the following features were excluded: 1) discordant sex; 2) genotype failure rate >= 0.05; 3) heterozygosity rate >= 5 s.d. from the mean 4) heterozygosity rate >= 3 s.d. from the mean and genotype failure rate >= 0.03; 5) identity-by-descent score > 0.185 and lower call rate to other individuals. The SNPs used in the GWAS were filtered on the basis of allele frequency, call rate, and deviation from Hardy-Weinberg equilibrium, as described previously.[3]

Genome-wide association analyses

In the discovery GWAS, the association between each SNP and ALL susceptibility was tested by comparing the genotype frequencies between ALL cases and non-ALL controls with an additive logistic regression model in PLINK (version 1.90),[5] including the top six principal components inferred using EIGENSTRAT[6] (continuous variables) as covariates to control for population stratification. Quantile-quantile plots indicated only minimal inflation at the tail of the distribution (λ = 1.09, **Supplemental Fig. 9**). Genotype dosage effect was examined for rs2836365, with OR = 1.55 [1.14-2.13] and 2.42 [1.75-3.35] for heterozygotes and homozygotes, respectively. The association of the rs2836365 genotype was also evaluated in genetically-defined EAs (N = 2,317 ALL and 2,050 controls) and AAs (N = 227 ALL and 1,380 controls), with cases from the same COG cohorts and MESA controls. The association of *ERG* genotype with Native American genetic ancestry was again evaluated in 167 Guatemalan children with ALL enrolled at the Unidad Nacional de Oncología Pediátrica, Guatemala City, Guatemala.

Using a large reference panel of human haplotypes from the Haplotype Reference Consortium (HRC r1.1 2016)[7] in Michigan Imputation Server[7,8] with ShapeIT (v2.r790)[9] as the phasing tool, we imputed additional SNPs within a 1 Mb region centered on the novel *ERG* variant (rs2836365). Regional plots were illustrated with LocusZoom.[10] In the replication studies, we tested the top hits at *ERG* locus with the additive logistic regression model conditioning on genetic structure.

Logistic regression model was used to evaluate SNP genotype associations with subtypes (e.g., *ETV6-RUNX1* ALL vs other ALL), sex, age at diagnosis (≥ or < 10 years of age), presenting whole blood cell count (≥ or < 50 × 10^9 cell/L), and minimal residual disease (MRD) at the end of remission induction therapy, adjusting for genetic ancestry. R (version 3.3.3) statistical software was used for all analyses unless indicated otherwise. All statistical tests were two-sided.

Functional annotation of ALL risk variants in *ERG*

To explore possible functional effects of *ERG* variants, we aligned the position of the ALL risk variants at this locus with GWAS signals for other blood cell-related traits at this locus (NHGRI GWAS Catalog), somatic *ERG* deletions in ALL (*Nat Genet* 48, 1481-1489, 2016), ATAC-seq signals in different types of hematopoietic cells (*Nat Genet* 48, 1193-1203, 2016), and placental mammal base-wise conservation scores by phyloP (UCSC Genome Browser). We also explored eQTL effects of the *ERG* risk variant using the GTEx (https://gtexportal.org/home/) and observed marginal association with *ERG* expression. The exact function of the *ERG* risk variant remains unclear and should be examined in future studies.

Association of germline and somatic *ERG* variations in ALL

ERG subtype and somatic *ERG* deletion in leukemia blasts were ascertained and defined as described previously.[11] Briefly, the *ERG* subtype was determined on the basis of gene expression profiling, and *ERG* deletion was assessed as copy number loss using Affymetrix SNP arrays. All cases in the *ER*G subtype had overexpression of *DUX4*, e.g., due to the *IGH-DUX4* rearrangement. We speculate that germline and somatic *ERG* variation have influence ALL leukemogenesis via distinct mechanisms. Compared to *DUX4*-mediated *ERG* deregulation in leukemia cells, inherited risk allele in *ERG* would have only modestly influence canonical *ERG* activity. In particular, because germline *ERG* allele was not associated with the risk of *IGH-DUX4* rearrangement, it is unlikely that this *ERG* variant would be related to expression of the alternative *ERG* transcript, which was directly oncogenic in lymphoid cells. In the same ALL cohort, we also tested association of *ERG* germline variant with somatic aberrations in *IKZF1* and *CRLF2* (**Supplementary Fig. 7B and 7C**).

Ancestry-dependent effects of the ALL risk allele in *ERG*

There are a number of possible explanations for the ancestry-dependent effects of *ERG* allele on ALL risk: 1) there is a common causal variant at this locus associated with ALL risk in all race groups, but in Africans it is in weak linkage disequilibrium (LD) with and thus poorly tagged by SNPs on the Affymetrix array (rs2836365). This variation in LD pattern by race directly contributed to the lack of association signal in Africans; 2) alternatively, *ERG* allele confers ALL risk in conjunction with another yet unknown genetic variants and the epistasis of these two contributes to leukemia pathogenesis. If the yet unknown interacting variant is absent in Africans (but common in Hispanics), the *ERG* allele alone would not show any association with ALL risk in Africans.

Reference

1. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 2009;19(9):1655-1664.
2. Mao X, Bigham AW, Mei R, et al. A genomewide admixture mapping panel for Hispanic/Latino populations. *Am J Hum Genet*. 2007;80(6):1171-1178.
3. Xu H, Yang W, Perez-Andreu V, et al. Novel susceptibility variants at 10p12.31-12.2 for childhood acute lymphoblastic leukemia in ethnically diverse populations. *J Natl Cancer Inst*. 2013;105(10):733-742.
4. Yang JJ, Cheng C, Devidas M, et al. Ancestry and pharmacogenomics of relapse in acute lymphoblastic leukemia. *Nat Genet*. 2011;43(3):237-241.
5. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559-575.
6. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet*. 2006;2(12):e190.
7. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet*. 2016;48(10):1279-1283.
8. Das S, Forer L, Schonherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet*. 2016;48(10):1284-1287.
9. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. *Nat Methods*. 2011;9(2):179-181.
10. Pruim RJ, Welch RP, Sanna S, et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*. 2010;26(18):2336-2337.
11. Zhang J, McCastlain K, Yoshihara H, et al. Deregulation of DUX4 and ERG in acute lymphoblastic leukemia. *Nat Genet*. 2016;48(12):1481-1489.

Supplementary Table 1. Clinical characteristics and molecular subtypes of the COG Hispanic ALL cases

	COG P9904/9905/9906 (N = 362)	COG AALL0232 (N = 624)
Age at diagnosis		
≥10	70 (19.3%)	412 (66.0%)
<10	292 (80.7%)	212 (34.0%)
Gender		
Male	185 (80.7%)	358 (57.4%)
Female	177 (19.3%)	266 (42.6%)
Leucocyte count at diagnosis (10^9 cell/L)		
⩾50	68 (19.3%)	298 (47.8%)
<50	294 (80.7%)	326 (52.2%)
DNA Index		
<1.16	252 (69.6%)	531 (85.1%)
≥1.16	95 (26.2%)	78 (12.5%)
Unkown	15 (4.1%)	15 (2.4%)
MRD		
Positive	67 (18.5%)	123 (19.7%)
Negative	265 (73.2%)	492 (78.8%)
Unkown	30 (8.3%)	9 (1.4%)
Tumor subtype		
ETV6-RUNX1	72 (19.9%)	72 (11.5%)
BCR-ABL1	0 (0.0%)	15 (2.4%)
MLL -rearranged	6 (1.7%)	0 (0.0%)
TCF3-PBX1	23 (6.4%)	NA
Hyperdiploid	94 (26.0%)	73 (11.7%)
B-other	167 (46.1%)	292 (46.8%)
Unkown	0 (0%)	172 (27.6%)

Abbreviations: COG, Children's Oncology Group; NA, not available.

Supplementary Table 2. Genome-wide significant variants identified in the discovery GWAS

SNP	CHR	Position	REF	ALT	Gene	Genotype ascertainment	Risk allele frequency		*P* value	OR (95% CI)
							Case (N = 940)	Control (N = 681)		
rs12719019	7	50476139	C*	T	*IKZF1*	*Genotyped*	0.76	0.65	4.35×10^{-8}	1.59 (1.35-1.89)
rs2390536	7	21485397	G	A*	*SP4*	*Imputed*	0.22	0.20	4.7×10^{-4}	1.39 (1.16-1.67)
rs46170118	8	130156143	A	G*	*8q24.1 (MYC)*	*Imputed*	0.15	0.14	0.2	1.15 (0.94-1.41)
rs28665337	8	130194104	C	A*	*8q24.21*	*Imputed*	0.13	0.11	0.1	1.21 (0.96-1.52)
rs3731217	9	21984661	A*	C	*CDKN2A*	*Imputed*	0.93	0.90	0.02	1.37 (1.04-1.79)
rs17756311	9	22053895	G	A*	*CDKN2A/B*	*Genotyped*	0.07	0.06	0.1	1.25 (0.93-1.69)
rs3824662	10	8104208	C	A*	*GATA3*	*Genotyped*	0.46	0.32	3.86×10^{-8}	1.56 (1.33-1.83)
rs4748793	10	22483011	A*	G	*COMMD3/BMI1/EBLN1*	*Genotyped*	0.81	0.78	0.05	1.19 (1.00-1.43)
rs7075634	10	22853102	T	C*	*PIP4K2A*	*Genotyped*	0.81	0.74	0.004	1.30 (1.09-1.55)
rs10821936	10	63723577	C*	T	*ARID5B*	*Genotyped*	0.60	0.44	1.83×10^{-12}	1.72 (1.47-2.00)
rs35837782	10	126293309	A	G*	*LHPP*	*Imputed*	0.53	0.50	0.004	1.24 (1.07-1.44)
rs4762284	12	96612762	A	T*	*ELK3*	*Imputed*	0.51	0.46	0.3	1.09 (0.94-1.26)
rs4982731	14	23585333	C*	T	*CEBPE*	*Genotyped*	0.44	0.39	0.04	1.16 (1.00-1.35)
rs2290400	17	38066240	T*	C	*IKZF3*	*Imputed*	0.62	0.58	0.3	1.09 (0.93-1.27)

Abbreviations: CHR, chromosome; REF, reference allele (GRCh37/hg19); ALT, alternative allele; OR, odds ratio; CI, confidence interval. *P* values and odds ratios were estimated by the additive logistic regression test adjusting genetic structure; *asterisk indicates risk allele for ALL susceptibility.

Supplementary Table 3. Association signals at the novel ALL risk locus *ERG* in Hispanics

SNP	CHR	Position[a]	Alleles[b]	Cohort[c]	Risk allele frequency (total number of subjects)		*P* value[d]	OR (95% CI)[e]
					Case	Control		
rs2836365	21	39768274	A/G*	Discovery	0.44 (940)	0.34 (681)	3.76×10^{-8}	1.56 (1.33-1.83)
				Replication	0.45 (144)	0.37 (441)	0.01	1.43 (1.07-1.89)
rs2836371	21	39773528	T/C*	Discovery	0.43 (940)	0.32 (681)	1.42×10^{-9}	1.64 (1.40-1.93)
				Replication	0.44 (144)	0.36 (441)	0.01	1.45 (1.09-1.93)

CHR, chromosome; OR, odds ratio; CI, confidence interval. Association of SNP genotype with ALL was evaluated by comparing allele frequencies in ALL cases and non-ALL controls, after adjusting for genetic structure.

[a]Chromosomal locations are based on GRCh37/hg19; [b]The second allele marked by asterisk is the risk allele for ALL. [c]Discovery cohort, COG AALL0232 and P9904/5; replication cohort, COG P9906 and SJ Total 13B/15. [d]*P* values were estimated by the additive logistic regression test adjusting for genetic structure. [e]OR values represent the increase in risk of developing ALL for each copy of the risk allele compared with subjects who do not carry the risk allele.

Supplementary Table 4. Association results for all examined SNPs at the *ERG* locus

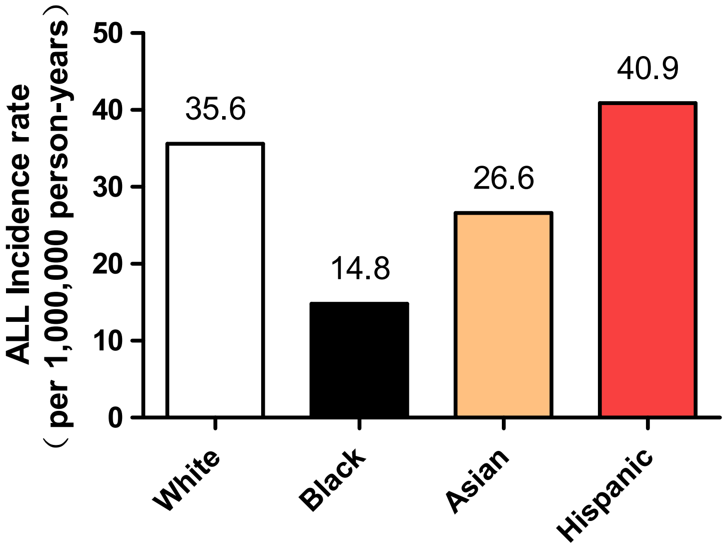
SNP	CHR	Position	REF	ALT*	Genotype ascertainment	Discovery				Replication			
						Risk allele frequency		*P* value	OR (95% CI)	Risk allele frequency		*P* value	OR (95% CI)
						Case (N = 940)	Control (N = 681)			Case (N = 144)	Control (N = 441)		
rs2836371	21	39773528	T	C	Imputed	0.43	0.32	1.42×10^{-9}	1.64 (1.40-1.93)	0.44	0.36	0.010	1.45 (1.09-1.93)
rs12053730	21	39769249	T	C	Imputed	0.43	0.32	5.96×10^{-9}	1.61 (1.37-1.89)	0.44	0.37	0.017	1.41 (1.07-1.88)
rs2836367	21	39769032	T	C	Imputed	0.43	0.32	7.76×10^{-9}	1.6 (1.37-1.88)	0.44	0.37	0.017	1.41 (1.07-1.88)
rs2836368	21	39769113	T	C	Imputed	0.43	0.32	7.76×10^{-9}	1.6 (1.37-1.88)	0.44	0.37	0.017	1.41 (1.07-1.88)
rs2836366	21	39768750	T	C	Imputed	0.43	0.32	9.08×10^{-9}	1.6 (1.36-1.88)	0.44	0.37	0.017	1.41 (1.07-1.88)
rs2836359	21	39764211	T	A	Imputed	0.43	0.32	1.06×10^{-8}	1.6 (1.36-1.87)	0.43	0.36	0.018	1.41 (1.06-1.87)
rs9983914	21	39766421	C	T	Imputed	0.42	0.32	2.88×10^{-8}	1.57 (1.34-1.84)	0.43	0.36	0.014	1.42 (1.07-1.89)
rs8131436	21	39789606	G	C	Imputed	0.44	0.34	3.90×10^{-8}	1.56 (1.33-1.82)	0.44	0.37	0.013	1.44 (1.08-1.93)
rs2836365	21	39768274	A	G	Genotyped	0.44	0.34	3.76×10^{-8}	1.56 (1.33-1.83)	0.44	0.37	0.018	1.4 (1.06-1.86)
rs2836361	21	39764588	G	A	Imputed	0.43	0.33	4.03×10^{-8}	1.56 (1.33-1.83)	0.44	0.37	0.015	1.42 (1.07-1.88)
rs2836362	21	39764784	T	C	Imputed	0.43	0.33	4.03×10^{-8}	1.56 (1.33-1.83)	0.44	0.37	0.015	1.42 (1.07-1.88)
rs3787889	21	39770856	A	T	Imputed	0.43	0.33	4.78×10^{-8}	1.56 (1.33-1.83)	0.43	0.37	0.031	1.37 (1.03-1.81)
rs2836360	21	39764437	T	C	Imputed	0.42	0.32	8.41×10^{-8}	1.55 (1.32-1.81)	0.42	0.35	0.014	1.42 (1.08-1.89)
rs9984638	21	39766569	T	C	Imputed	0.43	0.33	1.05×10^{-7}	1.54 (1.31-1.8)	0.44	0.37	0.017	1.41 (1.06-1.87)
rs2836363	21	39766031	C	T	Imputed	0.43	0.33	2.21×10^{-7}	1.52 (1.30-1.78)	0.43	0.35	0.015	1.42 (1.07-1.87)
rs9983942	21	39766497	C	A	Imputed	0.43	0.33	2.57×10^{-7}	1.52 (1.29-1.78)	0.43	0.36	0.020	1.4 (1.06-1.85)
rs2836357	21	39763354	C	A	Genotyped	0.43	0.33	2.97×10^{-7}	1.51 (1.29-1.77)	0.43	0.35	0.015	1.42 (1.07-1.87)
rs2836369	21	39769817	G	C	Imputed	0.38	0.29	3.37×10^{-7}	1.52 (1.30-1.79)	0.41	0.34	0.024	1.39 (1.04-1.85)
rs2246468	21	39762883	T	C	Imputed	0.28	0.18	3.57×10^{-7}	1.62 (1.35-1.95)	0.25	0.20	0.045	1.41 (1.01-1.96)
rs2836364	21	39767874	C	T	Imputed	0.37	0.28	4.07×10^{-7}	1.52 (1.29-1.79)	0.41	0.34	0.023	1.39 (1.05-1.85)
rs8127488	21	39789438	C	T	Imputed	0.39	0.30	4.74×10^{-7}	1.51 (1.29-1.78)	0.40	0.33	0.016	1.42 (1.07-1.89)
rs743298	21	39765510	C	T	Imputed	0.37	0.28	5.14×10^{-7}	1.52 (1.29-1.78)	0.41	0.33	0.009	1.46 (1.1-1.94)
rs55681902	21	39784752	T	C	Imputed	0.38	0.29	9.28×10^{-7}	1.49 (1.27-1.75)	0.39	0.33	0.031	1.37 (1.03-1.82)
rs9976326	21	39776485	A	T	Imputed	0.37	0.28	4.53×10^{-6}	1.46 (1.24-1.71)	0.39	0.31	0.012	1.44 (1.08-1.92)

Abbreviations: CHR, chromosome; REF, reference allele (GRCh37/hg19); ALT, alternative allele; OR, odds ratio; CI, confidence interval.
P values and odds ratios were estimated by the additive logistic regression test adjusting genetic structure; *asterisk indicates risk allele for ALL susceptibility.

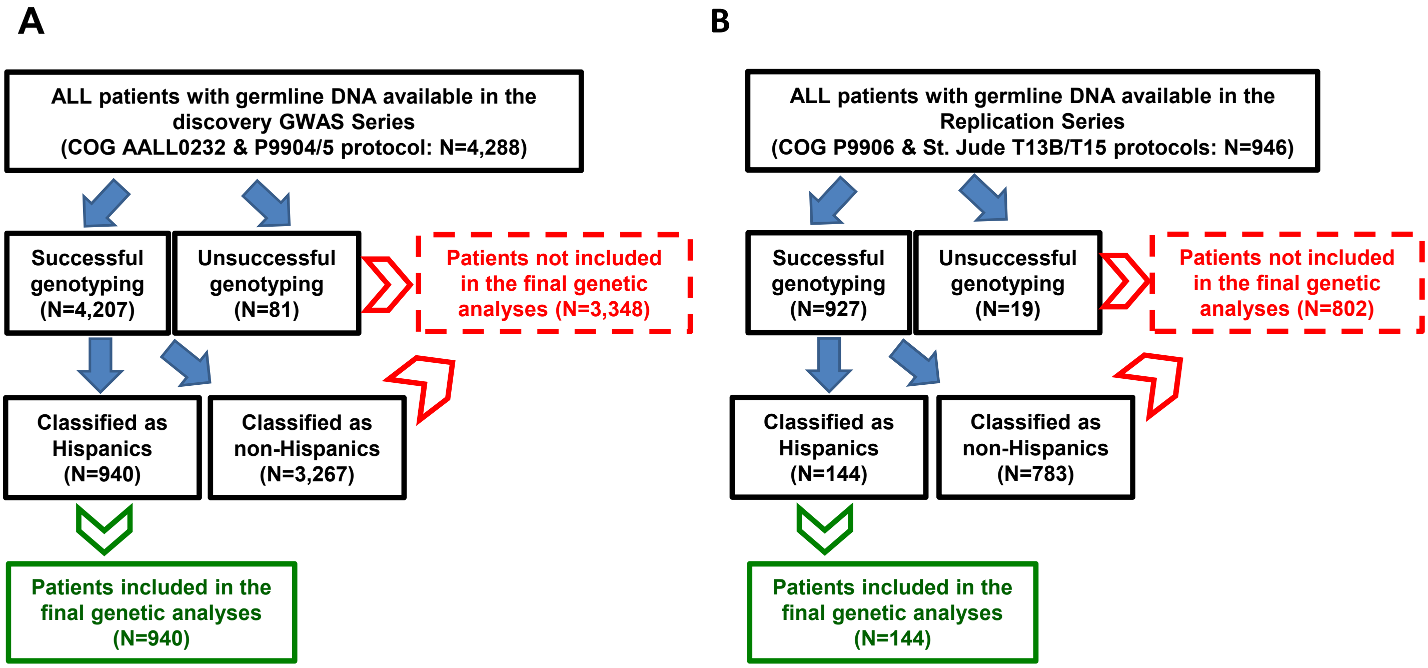
Supplementary Table 5. Association results for the top *ERG* SNPs in Hispanic, European, and African Americans

					Hispanic American				European American				African American			
SNP	CHR	Position	REF	ALT*	Risk allele frequency		*P* value	OR (95% CI)	Risk allele frequency		*P* value	OR (95% CI)	Risk allele frequency		*P* value	OR (95% CI)
					Case (N = 940)	Control (N = 681)			Case (N = 2,317)	Control (N = 2,050)			Case (N = 227)	Control (N = 1,380)		
rs2836365	21	39768274	A	G	0.44	0.34	3.76×10^{-8}	1.56 (1.33-1.83)	0.35	0.32	0.02	1.12 (1.02-1.22)	0.18	0.19	0.75	0.96 (0.74-1.24)
rs2836371	21	39773528	T	C	0.43	0.32	1.42×10^{-9}	1.64 (1.40-1.93)	0.35	0.32	0.008	1.14 (1.04-1.25)	0.10	0.09	0.50	1.13 (0.80-1.59)

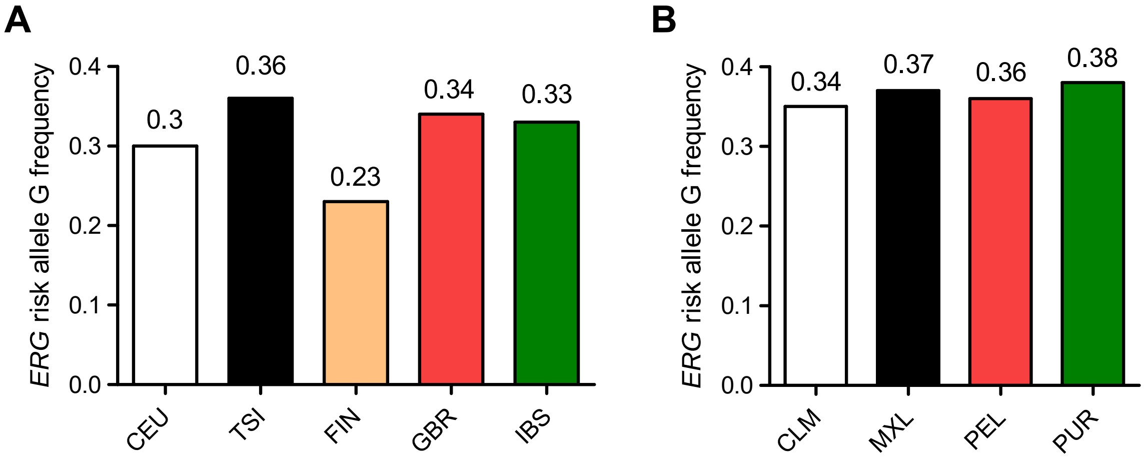
Abbreviations: CHR, chromosome; REF, reference allele (GRCh37/hg19); ALT, alternative allele; OR, odds ratio; CI, confidence interval.*P* values and odds ratios were estimated by the additive logistic regression test adjusting genetic structure; *asterisk indicates risk allele for ALL susceptibility.



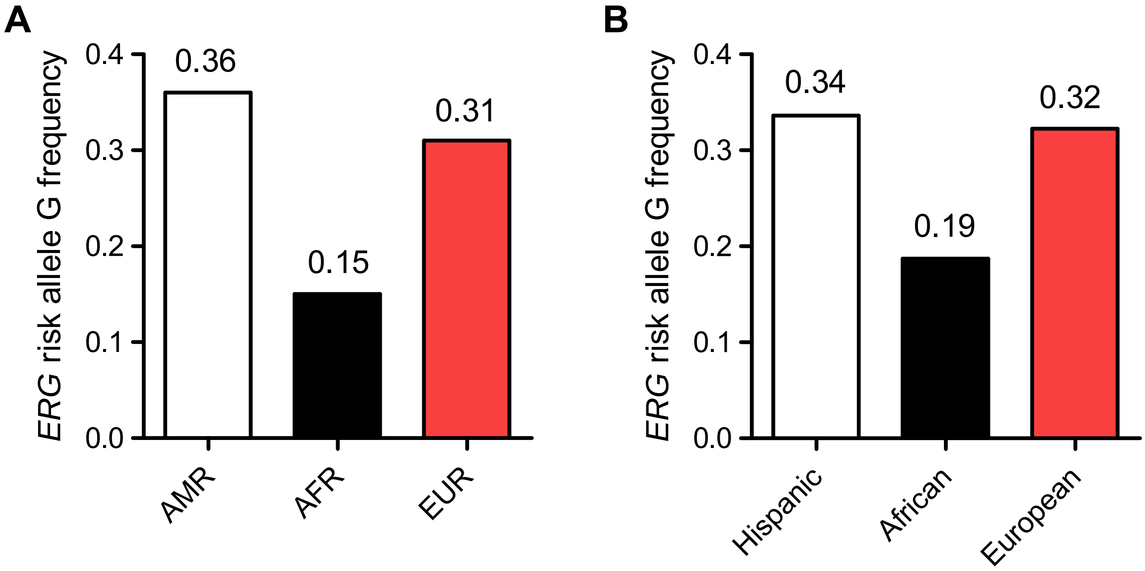
Supplementary Figure 1. ALL incidence rate in different races (i.e., White, Black, Asian, and Hispanic) obtained by 13 registries of the National Cancer Institute's Surveillance, Epidemiology, and End Results program diagnosed between 1992 and 2004 among children aged birth to 19 years in the U.S (adopted from *Cancer* 2008; 112:416). ALL, acute lymphoblastic leukemia.


A
ALL patients with germline DNA available in the discovery GWAS Series (COG AALL0232 & P9904/5 protocol: N=4,288)
Successful genotyping (N=4,207)
Unsuccessful genotyping (N=81)
Patients not included in the final genetic analyses (N=3,348)
Classified as Hispanics (N=940)
Classified as non-Hispanics (N=3,267)
Patients included in the final genetic analyses (N=940)
B
ALL patients with germline DNA available in the Replication Series (COG P9906 & St. Jude T13B/T15 protocols: N=946)
Successful genotyping (N=927)
Unsuccessful genotyping (N=19)
Patients not included in the final genetic analyses (N=802)
Classified as Hispanics (N=144)
Classified as non-Hispanics (N=783)
Patients included in the final genetic analyses (N=144)

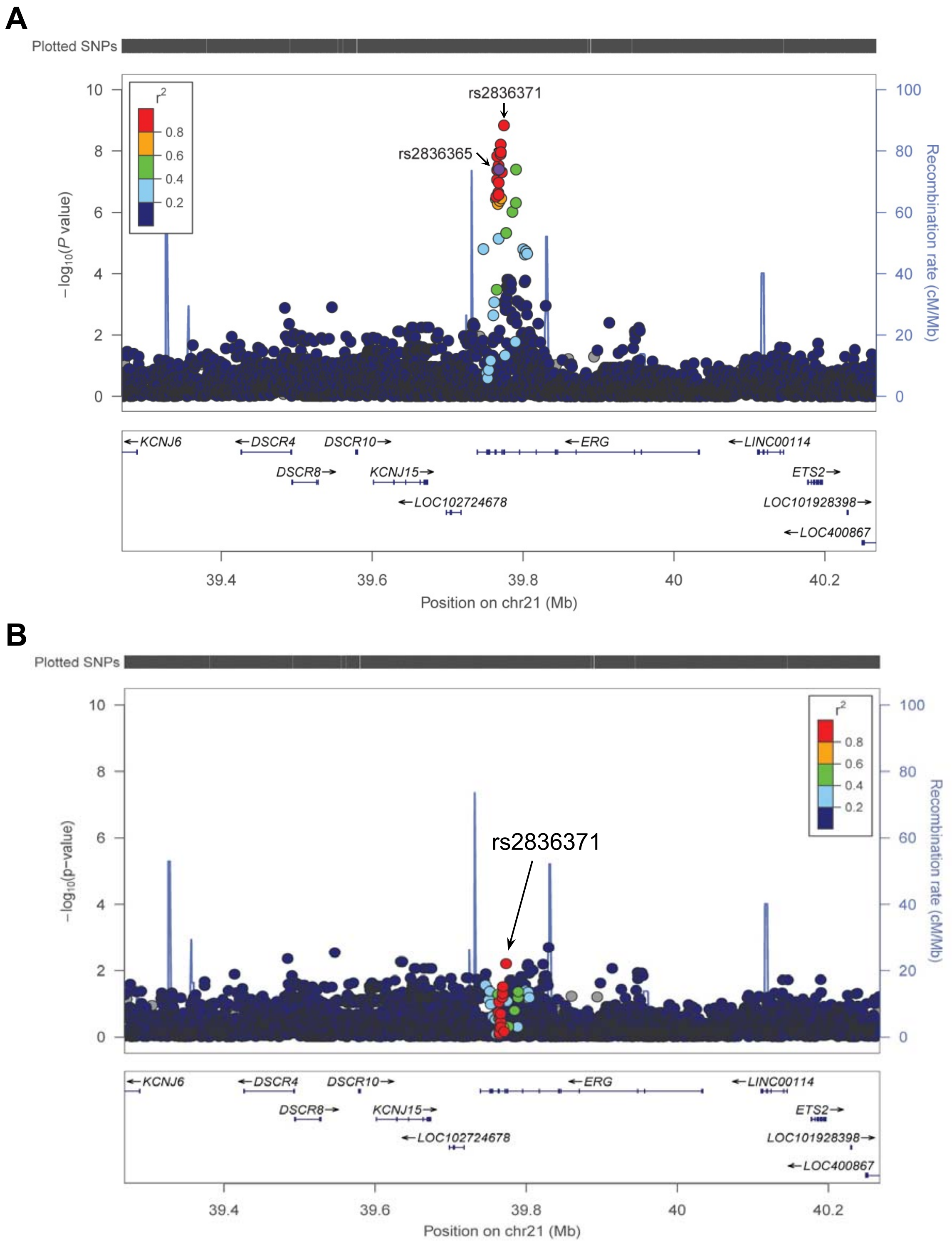
Supplementary Figure 2. Flow chart of patients inclusion/exclusion for the genetic analyses. Individuals with one of the following features were excluded: 1) discordant sex (between clinical data and genotype-inferred sex); 2) genotype failure rate > 5%; 3) heterozygosity rate >= 5 s.d. from the mean; 4) heterozygosity rate >= 3 s.d. from the mean if genotype failure rate >= 0.03; 5) identity-by-descent score > 0.185 and lower call rate to other individuals.



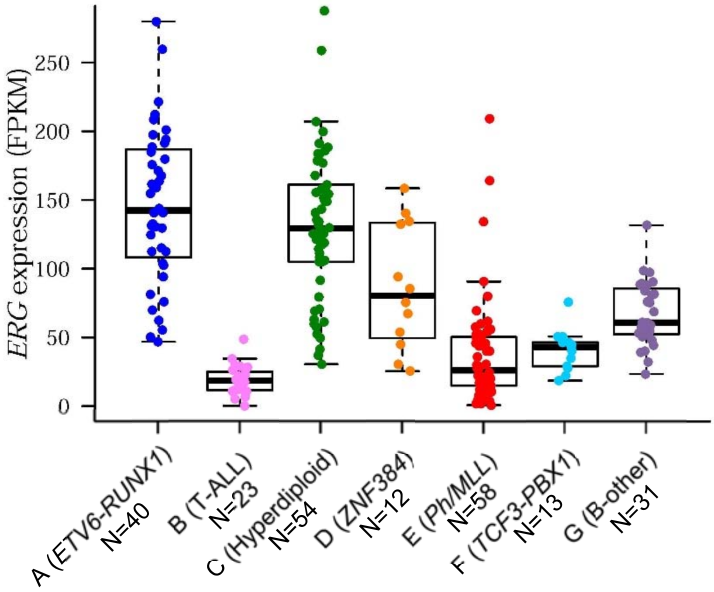
Supplementary Figure 3. *ERG* risk allele G frequency in different sub-populations with European ancestry (A) or in different American sub-populations (B) from the data generated by the 1000 Genomes Project. CEU, Utah Residents (CEPH) with Northern and Western European Ancestry; TSI, Toscani in Italia; FIN, Finnish in Finland; GBR, British in England and Scotland; IBS, Iberian Population in Spain; CLM, Colombian in Medellin, Colombia; MXL, Mexican Ancestry in Los Angeles, California; PEL, Peruvian in Lima, Peru; PUR, Puerto Rican in Puerto Rico.



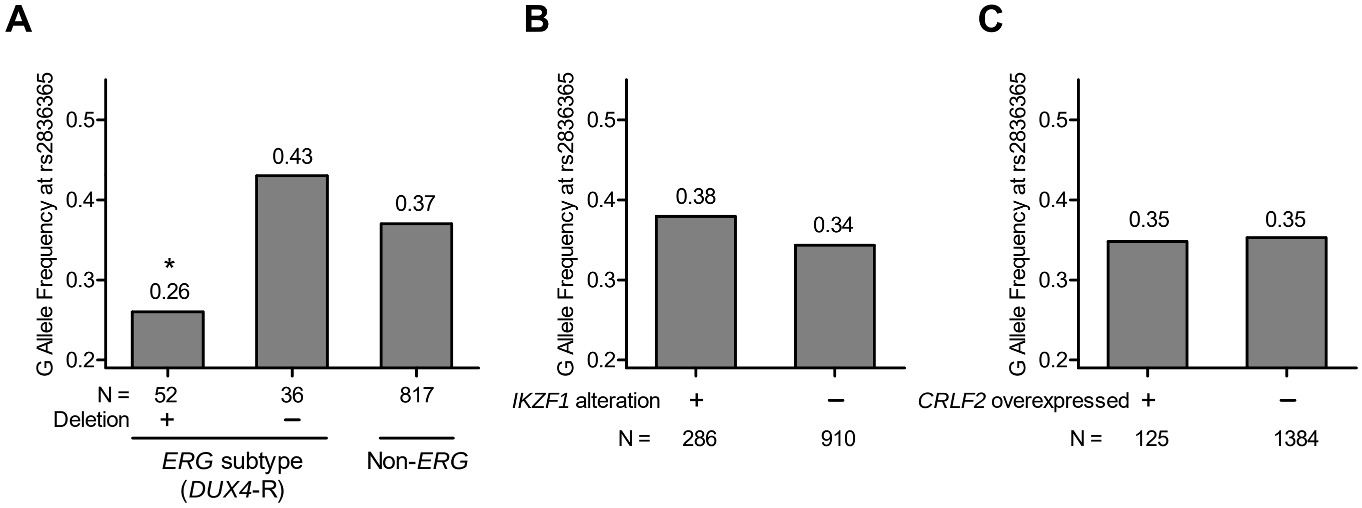
Supplementary Figure 4. *ERG* risk allele G frequency in different populations from the data generated by the 1000 Genomes Project (A) or by MESA (B). Across all populations we examined (Hispanics, Europeans and Africans), *ERG* allele frequency derived from MESA is highly consistent with 1000 Genomes. AMR, (Native) American; AFR, African; EUR, European.



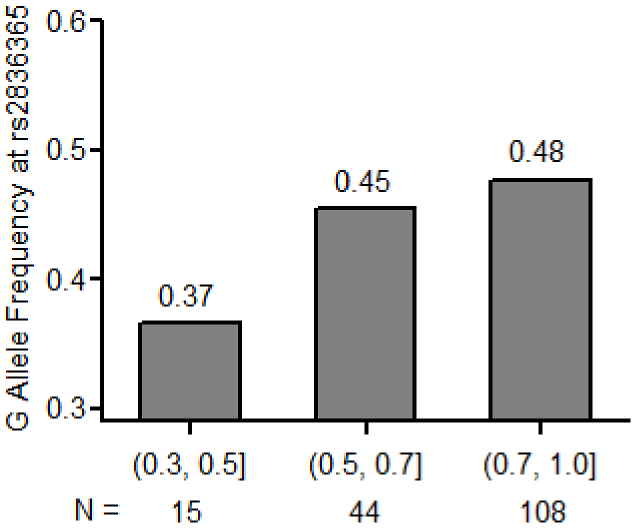
Supplementary Figure 5. Univariate and conditional analysis of genotype association at the *ERG* locus. Genotype imputation was performed using the Michigan Imputation Server, as described in **Methods**. Both directly-genotyped and imputed SNPs are included in the association analyses. The lead SNP from the original discovery GWAS (rs2836365, OR = 1.56, 95% CI: 1.33-1.83, $P = 3.8 \times 10^{-8}$) is indicated by its rsID and purple diamond shape, and other SNPs are displayed by color showing their extent of linkage disequilibrium with this lead SNP. Recombination rate, chromosomal position (hg19), and nearby genes (RefSeq) are indicated. The top SNP from the imputation analysis is rs2836371 (OR = 1.64, 95% CI: 1.40-1.93, $P = 1.4 \times 10^{-9}$). P values were calculated by additive logistic regression test in univariate analysis (**A**), or after adjusting for genotype at rs2836365 (conditional analysis, **B**).



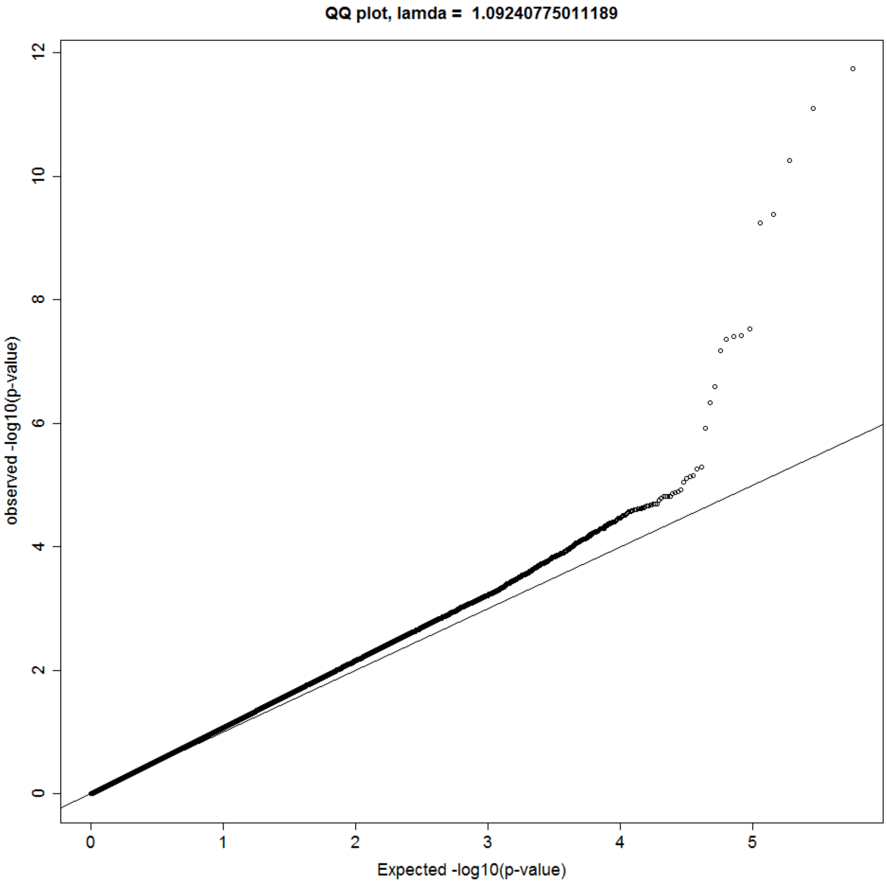
Supplemental Figure 6. Expression of the *ERG* gene in seven ALL subgroups identified from hierarchical clustering. The data were from whole-transcriptome sequencing of 231 children with newly diagnosed ALL enrolled in the Ma-Spore frontline ALL trials (*Genome Res.* 2017 Feb;27(2):185-195). Each sample is represented by a dot and is color-coded according to the subgroups it belongs to. Box plots depict the interquartile range, and whiskers indicate the maximal and minimal observations that are within 1.5 times the length of the box.



Supplementary Figure 7. The frequency of the *ERG* risk allele is negatively correlated with the incidence of somatic *ERG* deletion in ALL, but not with *DUX4*-R, *IKZF1* alteration or *CRLF2* overexpression. Risk allele frequency of rs2836365 were estimated for ALL cases of *ERG* subtype (*DUX4*-R) with or without *ERG* deletion and ALL cases of non-*ERG* subtype **(A),** for ALL cases with or without *IKZF1* alteration **(B)**, and for ALL cases with or without *CRLF2* overexpression **(C)**. Somatic genomics data (*DUX4* rearrangement, *ERG* deletion, *IKZF1* deletion, and *CRLF2* rearrangements) were reported previously (*Nat Genet* 48, 1481-1489, 2016). ALL, acute lymphoblastic leukemia. Logistic regression test with rs2836365 genotype adjusting for genetic ancestry, **P* < 0.05.



Supplementary Figure 8. The frequency of the *ERG* risk allele is correlated with Native American genetic ancestry in Guatemalan children. Risk allele frequency of rs2836365 were estimated for Guatemalan ALL children with increasing levels of Native American genetic ancestry (30-50%, 50%-70%, and 70-100%). ALL, acute lymphoblastic leukemia.



Supplementary Figure 9. Quantile-quantile (Q-Q) plot of logistic regression test for GWAS. The negative logarithm of the observed (*y* axis) and the expected (*x* axis) *P* value is plotted for each SNP (dot), and the black line indicates the null hypothesis of no true association. Deviation from the expected *P* value distribution is evident only in the tail area (λ = 1.09), suggesting that population stratification was adequately controlled by adjusting for genetic structure.