The impact of Indigenous American-like ancestry on risk of acute lymphoblastic leukemia in Hispanic/Latino children

Jalen Langie[1,2], Tsz Fung Chan[1,2], Wenjian Yang[3], Alice Y. Kang[4], Libby Morimoto[4], Daniel O. Stram[5], Nicholas Mancuso[1,2], Xiaomei Ma[6], Catherine Metayer[4], Philip J. Lupo[7,8], Karen R. Rabin[7,8], Michael E. Scheurer[7,8], Joseph L. Wiemels[1,2], Jun J. Yang[3], Adam J. de Smith[1,2,10], Charleston W.K. Chiang[1,2,9,10]

[1] Center for Genetic Epidemiology, Department of Population and Public Health Sciences, University of Southern California Keck School of Medicine, Los Angeles, CA, USA.

[2] USC Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA, USA.

[3] Department of Pharmacy and Pharmaceutical Sciences, St. Jude Children's Research Hospital, Memphis, Tennessee, USA.

[4] School of Public Health, University of California Berkeley, Berkeley, CA, USA.

[5] Department of Population and Public Health Sciences, University of Southern California Keck School of Medicine, Los Angeles, CA, USA.

[6] Department of Chronic Disease Epidemiology, Yale School of Public Health, New Haven, CT, USA.

[7] Division of Hematology-Oncology, Department of Pediatrics, Baylor College of Medicine, Houston, TX, USA.

[8] Texas Children's Cancer and Hematology Centers, Texas Children's Hospital, Houston, TX, USA.

[9] Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA, USA.

[10] Co-senior authors

Abstract:

Acute lymphoblastic leukemia (ALL) is the most common childhood cancer, with Hispanic/Latino children having a higher incidence of ALL than other racial/ethnic groups. Among genetic variants previously implicated in ALL risk, a number of them were found to be enriched in Indigenous American (IA)-like ancestries and inherited by many Hispanic/Latino individuals. However, due to potential confounding from environmental factors, the association between IA-like ancestry and risk for ALL has remained unclear. In this study, we characterized the impact of IA-like ancestry on overall ALL risk and on the frequency and effect size of known risk alleles, while accounting for non-genetic correlates of ancestry. Contrary to previous findings, we found that global IA-like ancestry was not significantly associated with ALL risk after adjusting for socio-economic indicators. On the other hand, locally at known ALL risk regions, we uncovered that increasing copies of the IA-like haplotype were positively and significantly associated with ALL risk (*e.g.* the IA-like haplotype had ~1.33 times the odds of harboring the risk allele compared to non-IA-like haplotypes), but we found no evidence of interaction between genotype and ancestry in relation to ALL. Admixture mapping identified replicable association signals at chr7p12.2 and chr10q21.2, consistent with the benefit of leveraging genetic ancestry in identifying genetic risk loci. Taken together, our results suggest that increased risk of ALL in Hispanic/Latino children may be conferred by higher frequency of risk alleles within IA-like ancestry, and that local ancestry-based analyses are robust stratagems to elucidate genetic etiology of disease.

Introduction:

Acute lymphoblastic leukemia (ALL) is the most common childhood cancer and accounts for 25% of malignancies before the age of 20 (ref. [1]). The long-term survival rate of the disease in

60 children is upwards of 85-90% [2], but survival rate decreases to approximately 40% after
61 relapse[3–6] and 30-40% in individuals over the age of 35. Despite decades of investigation into
62 ALL risk, the etiology of the disease is not well understood and disparities in risk persist.
63 Children of Hispanic/Latino ethnicity have 1.4 and 2 times the risk of developing ALL compared
64 to their Non-Latino White and Black counterparts, respectively[7] . This disparity in risk widens in
65 margin for Hispanic/Latino adults[7] and is specific to the B-cell immunophenotype – which
66 comprises approximately 85% of cases[8] . While environmental factors are involved in this
67 increased risk, genetic factors, which have been historically understudied in this population[9,10],
68 likely contribute in part to this disparity[11,12]
69
70 Genomic, migrational, and ecological data strongly suggest a complex genetic etiology of ALL,
71 both broadly across humans and within diverse Latin American populations. Most directly,
72 previous genome-wide association studies (GWAS) have estimated the heritability of ALL to be
73 approximately 18-21% in both Hispanic/Latino and Non-Latino White populations, and
74 approximately 24% of familial relative risk has been attributed to known ALL risk variants[13–15] .
75 Consistent with these estimates, GWAS have identified a number of loci associated with ALL
76 susceptibility, including risk alleles located near genes such as *IKZF1*, *ARID5B*, *GATA3*, and *ERG*
77 that appear to have higher frequency or stronger effects in Hispanic/Latino populations[12,16–18].
78 Previous migration studies have indicated that non-native- and US-born Hispanic/Latino
79 individuals have comparable rates of B-cell ALL to each other, and both have higher (~2.13 times)
80 incidence rates compared to their Non-Latino White counterparts[19,20]. Furthermore, ecological
81 data has indicated that countries with the highest age standardized incidence rates of ALL were
82 in Latin America (*i.e.* Honduras and Mexico), while countries with the lowest incidence rates were
83 in Africa, with European countries exhibiting intermediate incidence[21,22] .
84

85 Notable amongst populations with differing incidence rates of ALL is the presence or dominance
86 of a particular ancestry component. Due to the unique demographic history of Latin America,
87 Hispanics/Latino populations are admixed - drawing ancestral components from the American,
88 European, and African continents[23,24] . Previous literature has alluded to associations between
89 Indigenous American (IA)-like ancestry and ALL, suggesting that every 20% increase in IA-like
90 ancestry was associated with a 20% increase in odds of developing B-cell ALL[25]. Because
91 Hispanics/Latinos inherited a significant proportion of their genetic ancestries from Indigenous
92 Americans (though there is geographical heterogeneity across the Americas), this association
93 was presumed to explain the migration and ecological findings above. However, global ancestry
94 can be confounded by non-genetic factors such as socio-economic status[26], which was not
95 considered in previous analyses[25] . Further exploration of the association between IA-like
96 ancestry and ALL risk would need to account for this potential confounding.
97
98 Given the important role of genetic factors and genetic ancestry in ALL risk, we undertook a
99 systematic admixture analysis in the largest ALL cohort of self-reported Hispanic/Latino children
100 to date. Here, we sought to clarify the association between IA-like ancestry and ALL risk through
101 a study design that would better account for non-genetic confounders and leveraged local
102 ancestry analyses, which are less susceptible to confounding by environmental factors and have
103 previously yielded promising results in examining ancestry-ALL relationships[30]. We investigated
104 both inferred global ancestry across the genome and local ancestry at known ALL loci for
105 associations with risk. Further, we employed admixture mapping[27], which is best powered to
106 detect loci that exhibit frequency differences between ancestries and may underlie the risk
107 differences between populations[28,29].
108
109 We note that, given our goal to further understand the disparity in ALL risk epidemiologically
110 observed among Hispanic/Latino children, we inferred the genetic ancestries and characterized

111 their associations with ALL risk in self-reported Hispanic/Latino individuals. We defined genetic
112 ancestries at the continental (e.g. Africa, Europe, and Americas) level, which represents patterns
113 of genetic similarities across geographical space over the last ~10-50k years due to human
114 migrations and isolation-by-distance. These relatively distinct components of ancestries, though
115 arbitrarily defined, are only recently reunited in admixed populations, such as in individuals of
116 Hispanic/Latino origin. These components are inferred and defined relative to reference
117 populations available to us and are leveraged to uncover associations with traits in this research
118 setting but are invariably a simplified representation of the complex nature of genetic ancestries.
119 We thus append the suffix "-like" to refer to these ancestries to remind the reader of this
120 simplification and uncertainty[30].
121
122 **Methods:**
123
124 Study Subjects and Quality Control
125 We use the term "Hispanics/Latinos" as a population descriptor of our study subjects to align with
126 both the self-reported race/ethnicity categories and the definitions of the USA Office and
127 Management Budget[31], while acknowledging that we focus on individuals of Latin American origin.
128 Our analysis utilized the following cohorts of self-reported Hispanic/Latino study participants: the
129 California Cancer Linkage Project (CCRLP), the Kaiser Resource for Genetic Epidemiology
130 Research on Aging Cohort (Kaiser-GERA), the California Childhood Leukemia Study (CCLS),
131 Children's Oncology Group (COG), Multi-Ethnic Study of Atherosclerosis (MESA), the Adolescent
132 and Childhood Cancer Epidemiology and Susceptibility Service for Texas and Reducing Ethnic
133 Disparities in Acute Leukemia (ACCESS/REDIAL), and the Hispanic Community Health
134 Study/Study of Latinos (HCHS/SOL). All relevant information on sample size, arrays, study
135 design, and references/dbGaP accession numbers can be found in **Table 1**. Following the
136 practice in previous studies[12,13,32], cases from CCRLP were augmented with external controls

137 from Kaiser-GERA genotyped on a similar array. Cases from COG were paired with external
138 controls from MESA and ACCESS/REDIAL cases were paired with external controls from
139 HCHS/SOL. HCHS/SOL was a multi-center study, with study participants being ascertained from
140 Bronx, New York; Chicago, Illinois; Miami, Florida; and San Diego, California. The SOL San Diego
141 center was chosen as controls for ACCESS/REDIAL due to similarity in the first 2 principal
142 components and estimated ancestry distributions (**Supplementary Figures 1-2**).
143

144
145

Study	In-Text Name/ Acronym	Study Design	N	Genotype Array	Reference or Accession Number	Usage
California Cancer Linkage Project	CCRLP	Population Based case-control study	4033	Affymetrix Axiom World Array	Wiemels et al., 2018 (ref [32])	Global and local ancestry analyses; Discovery Admixture Mapping Cohort (combined with Kaiser-GERA)
Kaiser Resource for Genetic Epidemiology Research on Aging Cohort	Kaiser-GERA	Cohort study (controls only)	6417	Affymetrix Axiom World Array	phs000788.v1.p2	Discovery Admixture Mapping Cohort (combined with CCRLP)
California Childhood Leukemia Study	CCLS	Population Based case-control study (cases and controls)	1120	Illumina OmniExpress v.1	Ma et al., 2004 (ref [33])	Global and local ancestry analysis; Replication cohort
Children's Oncology Group (AALL0232, COG9904/9905)	COG	Clinical Trial (cases only)	836	Genome-Wide Human SNP Array 6.0	phs001350.v1.p1	Replication Cohort (combined with MESA)
Multi-Ethnic Study of Atherosclerosis	MESA	Cohort Study (controls only)	1118	Genome-Wide Human SNP Array 6.0	phs000209.v13.p3	Replication Cohort (combined with COG)
Adolescent and Childhood Cancer Epidemiology and Susceptibility Service for Texas and Reducing Ethnic Disparities in Acute Leukemia	ACCESS/ REDIAL	Hospital Cases (cases only)	510	Illumina Global Screening Array	Harris et al., 2023 (ref [34])	Replication Cohort (combined with HCHS/SOL)
Hispanic Community Health Study/Study of Latinos	HCHS/SOL	Cohort Study (controls only)	2285	Illumina Omni2.5M array	phs000810.v2.p2.c2; phs000810.v2.p2.c1	Replication cohort (combined with cases from Texas)

146
147 ***Table 1: All cohorts included in the analysis.*** *Information includes title, design, sample size, genotype array, dbGaP accession*
148 *number or reference, and how each study was used in our analysis.*

149
150 All analyses were limited to autosomes. Within each study, SNPs and individuals were removed
151 based on the following pre-defined quality control thresholds: SNP call rate < 0.98, MAF < 0.01,
152 $p_{HWE} < 1 \times 10^{-5}$, high relatedness ($\hat{\pi} > 0.2$), Genome Heterozygosity Rate > 6 standard
153 deviations from the mean, or sample call rate < 0.95. Palindromic SNPs were also excluded as
154 they could impact imputation accuracy due to unidentified strand flip or inversion between genome
155 build[35] . The number of individuals and SNPs remaining can be found in **Supplementary Table**
156 **1**. Local ancestry was inferred using RFMIX[36]. Reference haplotypes were obtained from the
157 genome aggregation database (gnomAD)[37] version 3, including 708 African, 671 Non-Finnish
158 European, and 94 Indigenous American individuals previously chosen for ancestry inference in
159 Hispanics/Latinos[38] . After restricting to the intersection of SNPs across each paired dataset, each
160 dataset was imputed separately using the TopMed Imputation Server[39]. We observed systemic
161 differences in imputation quality as measured by the imputation R^2 between cases and external
162 controls, presumably due to batch effects in data generation or subtle differences in ancestry
163 compositions between cohorts and differential imputation accuracies across populations[40].
164 Consequently, for all comparisons of cases to controls, we imposed filters such that only SNPs
165 with an imputation $R^2 > 0.30$ in both cases and controls and a R^2 difference < 0.15 between cases
166 and controls were included in the analysis.
167
168
169 <u>Global Ancestry Analyses</u>
170 Global ancestry for each study participant was calculated by averaging local ancestry estimates
171 across the genome. We performed logistic regression of ALL case-control status on global IA-like
172 ancestry in CCRLP and CCLS, studies in which indicators of socio-economic status (SES)
173 variables were available. Covariates included African-like ancestry, sex, and SES indicators
174 (highest education obtained by a parent and annual household income). Income was categorized

175 in 6 strata, in accordance with its treatment in CCLS[41] : <$15,000, $15,000-$29,999, $30,000-
176 $44,999, $45,000-$59,999, $60,000-$74,999, and >$75,000. In CCLS, highest educational
177 attainment of either parent was divided into 4 categories: Less than elementary school, high
178 school, some college, and a bachelor's degree or higher. In CCRLP, the strata were less than 9th
179 grade, 9th-12th grade, and some college or more. Both SES indicators were treated as ordinal
180 categorical variables, in which the lowest category was used as a reference. Linearity
181 assumptions were assessed using residual and QQ plots. Analysis was carried out using R v.4.3.1
182 (ref. [42]).
183
184 Enrichment of risk allele on IA-like ancestry haplotypes
185 We performed local ancestry analyses at known ALL loci, based on the list provided in ref. [11]. We
186 excluded the T-cell ALL-specific association locus at *USP7*[43] given that the increased risk of ALL
187 in Hispanics/Latinos is limited to B-cell ALL. We restricted analyses to remaining SNPs that were
188 independent at the local ancestry level (*i.e.* choosing SNPs that were not in admixture linkage
189 disequilibrium; N=20) using PLINK's clumping algorithm[44]. SNPs were clumped if the diploid local
190 IA-like ancestry assignments covering the SNPs were 90% correlated and within 10 Mb of each
191 other.
192
193 A total of three analyses were performed using local ancestry estimates. First, across known ALL
194 loci[11], we cross-tabulated local genetic ancestry (number of copies of the IA-like, African-like, or
195 European-like haplotypes) and risk allele dosage to identify association of risk alleles with
196 ancestry. To establish a null distribution to which we would compare our observed ancestry-risk
197 allele association, we randomly selected 1000 sets of 20 frequency-matched alleles from across
198 the genome, without replacement. Differences between the null and known loci distributions were
199 evaluated via a Chi-Squared Goodness of Fit test. We restricted this analysis to CCRLP and
200 CCLS controls, to approximate the frequency of the risk allele in the general population. Second,

201 we performed a logistic regression of ALL case-control status on IA-like ancestry dosage at each
202 locus, including as covariates global African-like and European-like ancestry, and sex, to evaluate
203 how likely is the IA-like ancestry associated with increasing ALL risk. Third, to test for presence
204 of heterogeneous effect sizes across ancestries, we performed logistic regression of ALL case-
205 control status on risk allele genotype, IA-like ancestry dosage, sex, the first 20 principal
206 components, and an interaction term of local ancestry and risk allele dosage. Analyses were
207 carried out either using PLINK v2.0 (ref. [45]) or R v.4.3.1 (ref. [42]).
208
209 Admixture mapping, regional association testing, and replication
210 We performed a genome-wide scan for loci in which IA-like ancestry was associated with ALL
211 risk. We regressed ALL case-control status on the local IA ancestry dosage, global African-like
212 and European-like ancestry, sex, and the first 20 principal components. We performed this
213 analysis in CCRLP/Kaiser as our discovery dataset. The significance threshold for genome-wide
214 admixture mapping was 5.7 x 10^{-5}, as suggested in literature previously for Hispanic/Latino
215 populations[46,47].
216
217 We fine-mapped putative admixture association region through single-variant association testing
218 in the same dataset. The admixture peak region[48] was defined as a range of positions that
219 contains all SNPs with a -log10(P) between -log10(P_{am}) and -log10(P_{am}) – 1, with P_{am} representing
220 the p-value associated with the lead SNP from admixture mapping. Genotypes of all imputed
221 SNPs within this region passing QC filtering were tested via logistic regression, with sex and the
222 first 20 principal components as covariates. Significance threshold for each region were computed
223 using 1000 rounds of permutation of case-control labels, choosing the threshold that controls for
224 the 5% false discovery rate. We attempted to replicate the association at the lead variants from
225 regional association testing and conditional analysis in additional Hispanic/Latino cohorts: CCLS,

226 COG/MESA, and SOL-SD/REDIAL. The replication cohorts were meta-analyzed together by
227 inverse variance weighting in the software *METAL*[49].
228
229 Evaluating the impact of using self-reported race/ethnicity labels versus genetic inferred ancestry
230 We opted to perform admixture mapping using all self-reported Hispanic/Latino individuals to be
231 inclusive and avoid forming study exclusion criteria based on estimated genetic ancestry.
232 However, alternative approaches could choose to include participants based on inferred genetic
233 ancestries. To examine, in our context, the impact of informing study inclusion based on
234 genetically inferred ancestry, we implemented a software, Harmonizing Ancestry and Race
235 Ethnicity[50] (HARE; Fang et. al, 2019) on genetic data from CCRLP/Kaiser-GERA
236 (**Supplementary Figure 3** . HARE is a two-step machine learning-based algorithm that identifies
237 individuals who may have discordant self-identified race/ethnicity (SIRE) and genetically inferred
238 ancestry (GIA) information, given the mapping of the two learned from the rest of the dataset. The
239 first step is a training phase in which genetically inferred ancestry based on principal components
240 are used to predict available self-reported race/ethnicity in a support vector machine (SVM). The
241 second step is HARE strata assignment, in which the highest probability of being assigned to a
242 stratum is contrasted with the probability that the individual belongs to their self-reported stratum
243 using preset thresholds. If the ratio of the probabilities exceeds a pre-defined threshold, an
244 individual with existing SIRE could be assigned as missing (*i.e.* removed from analysis), while an
245 individual without SIRE would be assigned into the top predicted strata.
246
247 We applied HARE with the default parameters and thresholds[50], training a multi-class SVM with
248 a radial kernel on the first 30 principal components (calculated using PLINK[45] v2.0) of genotype
249 data from individuals across all self-reported race/ethnicity groups from CCRLP/Kaiser-GERA
250 (10,450 Hispanic/Latino, 2,191 Non- Latino African, 5,335 Non-Latino East Asian, 58,503 Non-
251 Latino European individuals). We first evaluated if the HARE algorithm would exclude any study

252 participants due to discordance between SIRE and GIA. We also assessed the rate of
253 discordance between SIRE and GIA by performing 10-fold cross validation in which we randomly
254 masked the SIRE information of 10% of study participants for each race/ethnicity group in each
255 trial. This assessment simulates a scenario where one may define study exclusion criteria based
256 on comparing the GIA to a known reference panel. Percent discordance was defined as the
257 number of individuals classified as missing or as a stratum not corresponding to the masked SIRE
258 label. We then investigated the impact of this study exclusion criteria on our admixture mapping
259 result.
260
261
262 IRB Statement:
263 This study was approved by the institutional review boards of the California Health and Human
264 Services Agency, the University of Southern California, Children's Oncology Group, Baylor
265 College of Medicine, Yale University, and University of California, Berkeley. All study participants
266 were de-identified and newborn bloodspots from the CCRLP were obtained via a consent waiver
267 from Committee for the Protection of Human Subjects of the State of California.
268
269
270 **Results:**
271
272 *The association between global Indigenous American-like ancestries, ALL risk, and SES*
273 *indicators*
274 Previous studies have identified a significant and positive association between global IA-like
275 ancestries and ALL risk[25] (Walsh et al., 2013) but did not adjust for socioeconomic status (SES)
276 variables that often confound ancestry-disease associations. To quantify the effects of IA-like
277 ancestry on ALL risk, we tested the association between ALL case-control status and global IA-

like ancestry, while adjusting for African-like ancestry, sex, as well as covariates that are proxies for SES, such as highest parental education and annual household income. We performed these assessments for all self-reported LAT individuals in both CCRLP (excluding Kaiser-GERA controls; N = 4033) and CCLS (N = 605), as both of these cohorts have available SES variables, contained internal controls and are less likely to be confounded due to their study design. In CCLS, IA-like ancestry was significantly associated with ALL risk, prior to including SES indicators in the model, as previously reported (Walsh et al., 2013). In the base model adjusting for sex and African-like ancestry, a 20% increase in IA-like ancestry was associated with a 23% (OR: 1.23, 95% CI: 1.07-1.42; P = 0.003) increase in odds of developing ALL. However, this association appears confounded, as inclusion of SES covariates strongly attenuated the strength of association (OR: 1.12; 95% CI: 0.96-1.31) and case/control status is no longer significantly associated with IA-like ancestry ($P = 0.159$; **Supplementary Table 2**). In CCRLP, IA-like ancestry was not significantly associated with ALL risk in the base model adjusting for only African-like ancestry and sex, and did not qualitatively change when SES covariates were included (**Supplementary Table 2**). Regression analysis between the SES variables and ALL case-control status and ancestry, respectively, further supported the suggestion that socio-cultural factors can act as confounders in ancestry-trait relations. For instance, we observed higher income to be associated with lower risk of disease (*e.g.* 0.27 – 0.76 times the odds of developing ALL; **Supplementary Tables 3-4**), and higher European-like ancestry is associated with higher socio-economic status (**Supplementary Tables 5-6**). Together, because SES measures can correlate with both ancestry and disease risk, the previously observed association between IA-like ancestry and ALL risk (Walsh et al., 2013) may have been confounded.

304 *Known ALL risk loci are enriched among Indigenous American-like ancestral haplotypes*

305 While global ancestry association analyses can be confounded by socioeconomic indicators, use
306 of local ancestry to identify associations at risk loci may be more robust. Further, the absence of
307 an association between IA-like ancestry and ALL risk globally would not necessarily preclude an
308 association locally at known risk loci. Thus, we performed three analyses leveraging the pattern
309 of local ancestry at known B-cell ALL risk loci[11] (**Supplementary Table 7**) to examine if the risk
310 alleles are found more frequently or show larger effects on IA-like ancestry haplotypes. First, we
311 tabulated the proportion of times the risk allele at known ALL loci is associated with the African-
312 like, IA-like, or European-like ancestry haplotypes, and compared these to the proportions found
313 across 1000 sets of SNPs randomly selected across the genome and matched by the risk allele
314 frequencies. Across a total of 20 quasi-independent ALL SNPs (**Methods**)**,** the risk alleles were
315 found on the African-like, IA-like, and European-like ancestry haplotypes 5.76%, 45.35%, and
316 48.88% of the time, respectively, in CCRLP. This significantly deviates from the genome-wide
317 expectation (4.83%, 43.33%, and 51.85% for African-like, IA-like, and European-like ancestry
318 haplotypes; p-value = 1.04×10^{-36} by a chi-square goodness-of-fit test), with IA-like haplotypes
319 showing a significant enrichment in association with the risk alleles. We also replicated this
320 observation in CCLS, with IA-like haplotypes enriched with ALL risk alleles (p-value = $1.629 \times 10^{-}$
321 ^{9}). Consistent with this finding, the IA-like haplotypes have 1.32-1.33 (95% CI: 1.29 - 1.36 in
322 CCRLP, 1.26-1.41 in CCLS) times the odds of carrying the risk allele compared to Non-IA-like
323 haplotypes (p = 5.61×10^{-88} and 2.61×10^{-23} in CCRLP and CCLS, respectively; **Supplementary**
324 **Table 8**).

325

326 In the second local ancestry analysis, we regressed ALL case-control status on the number of
327 copies of IA-like ancestry segments locally at each known ALL locus, while adjusting for sex and
328 global African and European ancestry proportions. Because we anticipate that risk alleles for ALL
329 are found disproportionately on IA-like haplotypes across these loci, we would expect that IA-like

330 ancestries show an enrichment of association increasing the risk of ALL in the setting of admixture
331 mapping studies. Indeed, we found a positive association of IA-like ancestry with ALL at 14 out of
332 20 known B-cell ALL loci. At all loci in which IA-like ancestry was nominally significantly ($p < 0.05$)
333 associated with ALL, IA ancestry was positively associated with risk (Binomial test, $p = 0.008$
334 **Table 2**). To confirm that the known risk alleles were in fact driving the ancestry association, we
335 included risk genotype as an additional covariate in each model. At each locus, upon adjustment
336 for risk genotype, the association between IA-ancestry and ALL risk reduced in significance
337 (**Supplementary Table 9**). Taken together, our results suggest that IA-like haplotypes are
338 enriched for association with increased ALL risk at known loci, driven by increased frequency of
339 risk allele.

340
341

Gene	Chr	Location	OR	L95	U95	P-value[a]
BCL11A	2	2p16.1	0.94	0.86	1.02	0.138
RPL6P5	2	2q22.3	1.01	0.93	1.10	0.751
C5orf56	5	5q31.1	1.07	0.98	1.17	0.113
BAK1	6	6p21.31	0.946	0.87	1.03	0.208
MYB/HBS1L	6	6q23	0.99	0.91	1.07	0.749
IKZF1	**7**	**7p12.2**	**1.32**	**1.21**	**1.43**	**2.990×10^{-10}**
CCDC26	8	8q24.21	0.93	0.85	1.01	0.084
CDKN2A/2b	9	9p21.3	0.93	0.86	1.02	0.115
TLE1	9	9q21.31	1.04	0.96	1.13	0.358
GATA3	**10**	**10p14**	**1.14**	**1.04**	**1.24**	**0.003**
BMI1/PIP4K2A	**10**	**10p12.2/10p12.31**	**1.16**	**1.07**	**1.27**	**5.422×10^{-4}**
ARID5B	**10**	**10q21.2**	**1.29**	**1.19**	**1.41**	**4.924×10^{-9}**
JMJD1C	**10**	**10q21.3**	**1.27**	**1.16**	**1.38**	**5.383×10^{-8}**
TET1	**10**	**10q21.3**	**1.15**	**1.06**	**1.25**	**0.001**
LHPP	10	10q26.13	0.97	0.89	1.06	0.508
ELK3	**12**	**12q23.1**	**1.10**	**1.01**	**1.19**	**0.038**
CEBPE	**14**	**14q11.2**	**1.10**	**1.01**	**1.20**	**0.022**
IKZF3	17	17q21.1	1.03	0.95	1.11	0.515
IGF2BP1	17	17q21.32	1.02	0.93	1.11	0.728

ERG	21	21q22.2	1.02	0.93	1.11	0.711

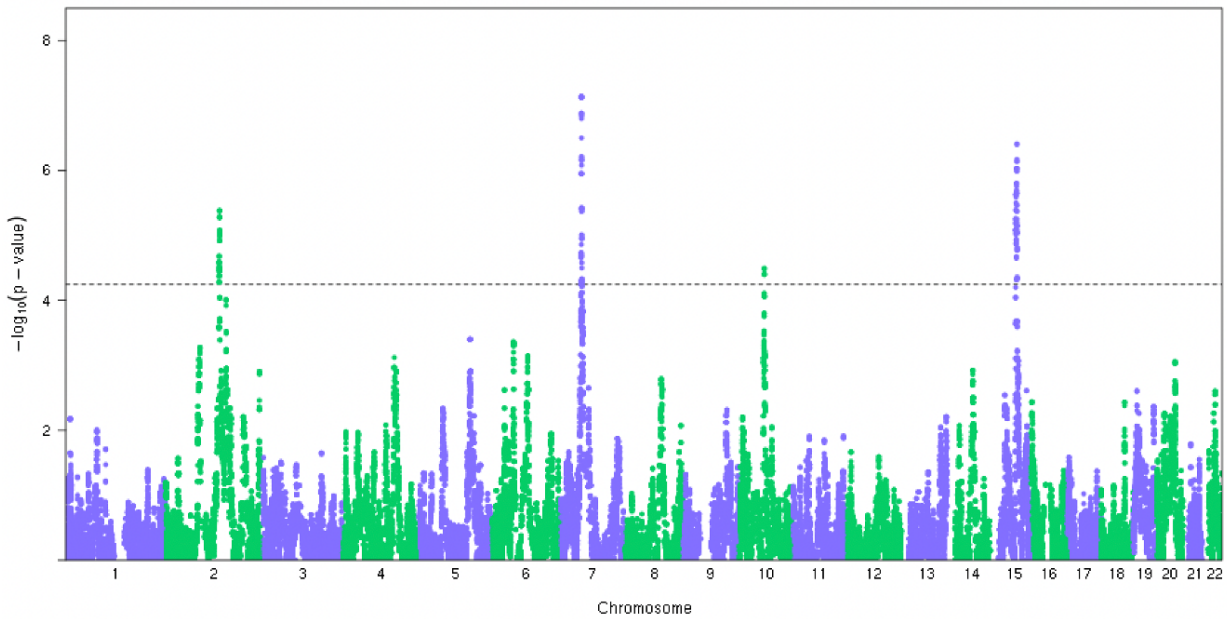
Table 2: Indigenous American-like ancestry is positively associated with risk at known ALL loci. *ALL case-control status was regressed on local Indigenous ancestry dosage, African-like and European-like global ancestries, and sex[a]. Loci that reach nominal significance (P < 0.05) are in* ***bold****)*

In addition to ALL risk alleles being found more prevalently on IA-like haplotypes, risk alleles may also exert larger (or smaller) effects on ALL when they are found on the IA-like background. Therefore, for the third local ancestry analysis, we tested for effect size heterogeneity by including a SNPxAncestry interaction term in the standard association testing model. After Bonferroni correction, no significant interactions between genotype and local IA-like ancestry were observed in either CCRLP or CCLS, or when the two cohorts were combined (**Supplementary Table 10**). We additionally found no evidence of interaction between global IA ancestry and genotype (**Supplementary Table 11**). Therefore, we found no evidence that ALL risk alleles may exert larger effects on an IA haplotype than on other haplotypic backgrounds.

Admixture mapping proposes putative novel ALL loci associated with Indigenous American-like ancestry.

Although results from our global ancestry analysis showed minimal association between IA-like ancestry and ALL risk after adjusting for SES indicators, it does not preclude the possibility that some risk alleles show drastic differences in frequency between ancestries, and thus are amenable to being discovered through admixture mapping, particularly since we observed that IA-like ancestries are enriched for association with risk alleles for ALL (**Table 2** above). We thus performed admixture mapping using the CCRLP and Kaiser/GERA Hispanic/Latino cohort as our discovery panel (1930 cases and 6417 controls). Our discovery admixture mapping

367 analyses identified genome-wide significant (threshold = 5.7 x 10^{-5} ; see **Methods**) signals at
368 2q21.2, 7p12.2, 10q21.2, and 15q22.31. The loci on chromosomes 7 (*IKZF1*) and 10 (*ARID5B*)
369 are known, while those on chromosomes 2 and 15 are potentially novel (**Figure 2**).
370



371
372 ***Figure 2: Manhattan plot of the admixture mapping analysis.*** *The analysis identified signals*
373 *at 2q21.2, 7p12.2, 10q21.2, and 15q22.31. The significance threshold is 5.7e-5.*
374

375 We then fine-mapped each of the four admixture loci via single-variant regional association
376 testing (**Methods**), which uncovered the putative lead SNPs in each region to be 2:133605881
377 (p = 4.209 x 10^{-5}), 7:50395237 (p = 2.262 x 10^{-12}), 10:61959980 (p = 2.245 x 10^{-44}), and
378 15:66111790 (p = 3.011 x 10^{-4} ; **Table 3**). SNPs on chr2, 7, and 10 surpassed the regional
379 significance threshold after accounting the number of tests performed in each region via
380 permutation (**Methods**), and chr15 SNP was slightly below the threshold. These SNPs exhibit
381 allele frequency differences by ancestry group (**Table 3**), and conditioning on these SNPs also
382 generally attenuated the admixture mapping association (**Table 3)**. We sought replication of
383 these associations in additional Hispanic/Latino cohorts, including cases from CCLS, COG, and

384 REDIAL (see **Methods**), totaling 1,951 cases and 3,918 controls. In the meta-analyzed
385 replication cohorts, only the lead SNPs or other SNPs in LD with the lead SNPs on
386 chromosomes 7 and 10 were statistically significant (**Supplementary Table 12-13**). Through
387 power calculations[51] , we found that, given observed odds ratios in the discovery cohort and
388 combined allele frequencies in the replication datasets, our replication analysis had only 1.2%
389 and 0.5% power to detect the associations on chromosomes 2 and 15, respectively; in contrast,
390 we had 17.8% power to identify the signal on chromosome 7 and 100% power to detect the
391 locus on chromosome 10.
392
393
394 While local ancestry is not likely to be associated with socio-cultural variables, such analyses
395 can still be confounded by population stratification (Mersha 2015). In our study, we opted to
396 include all self-reported Hispanic/Latino individuals in our admixture mapping analysis and
397 adjusted for potential confounding by including global ancestries as covariates in our admixture
398 mapping model. Nevertheless, use of genetically inferred ancestry is pervasive and often
399 heralded as a convenient, albeit arbitrary way of identifying racial/ethnic strata to better control
400 for confounding by population structure (Fang et al., 2019) To evaluate whether our use of self-
401 reported data may lead to starkly different results from using genetically inferred ancestry, we
402 employed the Harmonizing Ancestry Race Ethnicity algorithm (HARE; Fang et al., 2019) to
403 identify individuals whose self-reported race ethnicity labelled may not be consistent with
404 genetically inferred ancestry compared to a reference. Consistent with the goals of our study
405 and best practices in using population labels in human genetic studies[30], HARE does not
406 attempt to re-assign individuals to strata; rather, individuals with discordant self-report and
407 genetically inferred data are indicated and thus can be excluded from analysis.
408

409 Using all available self-reported race/ethnicity labels in the multi-ethnic CCRLP + Kaiser cohort,
410 we observed very low rates of discordance between HARE-predicted labels and the self-
411 reported labels (percent discordance = 0.47% overall, and 0.68%, 3.37%, 0.18%, and 0.54% for
412 AFR, EAS, EUR, and LAT study participants, respectively. **Supplementary Table 14**),
413 suggesting that in general there is not enough genetic evidence to exclude participants from
414 analysis (nor do we advocate for removing individuals from analysis on the basis of genetic
415 make-up alone). In a scenario where individuals might be missing self-reported race/ethnicity
416 labels, HARE can be used to predict the label based on a reference. We simulated such a
417 scenario using 10-fold cross validations, randomly masking in turn 10% of the cohort while using
418 the remaining 90% as the reference for the HARE algorithm to learn the relationship between
419 self-reported labels and genetically inferred ancestry (**Methods**). In this case, we observed the
420 algorithm to be less confident, recalling 61.5% of LAT individuals, compared to 77.5%, 82.6%,
421 and 86.9% of NLW, AFR, and EAS individuals, respectively. Of those not recalled back as
422 Latino, 63.5% were assigned as missing, 35.3% were assigned as Non-Latino White, 0.4%
423 were assigned as Non-Latino East Asian, and 0.8% were assigned as Non-Latino Black
424 (**Supplementary Table 15**). As a sensitivity analysis, we restricted our admixture mapping
425 analysis to the set of Hispanic/Latino individuals inferred by HARE across 10 cross-validation
426 trials**,** which produced a panel of 1,729 cases and 4,696 controls (**Supplementary Table 16**).
427 Overall, we observed high correlation in odds ratios (Pearson's correlation coefficient, r = 0.
428 913) and log *P*-values (r = 0.853), with a modest decrease in significance among variants with
429 *P*-values < 1 x 10^{-3}, including the four putative loci (**Supplementary Table 12; Supplementary**
430 **Figures 5-6**), likely reflecting the smaller sample size.
431
432
433

Chr	Pos (Hg38)	rsID	Locus	Ref	Risk	$OR_{Top\ SNP}$ $(95\%\ CI)$	$P_{Top\ SNP}$	RAF^{*}_{EUR}	RAF_{AFR}	RAF_{IA}	Gene	P_{crude}	P_{adj}	Peak Region
2	134116620	rs117345013	2q21.2	T	C	1.22 (1.11, 1.72)	4.209×10^{-5}	0.010	0.000	0.390	Unknown	4.156×10^{-6}	0.005	133605881, 134236147
7	50390453	rs6969942	7p12.2	C	T	1.35 (1.25, 1.47)	2.262×10^{-12}	0.507	0.773	0.870	IKZFI	7.315×10^{-8}	0.002	50352094, 50457722
10	61959980	rs4948492	10q21.2	T	C	1.85 (1.59, 2.13)	2.245×10^{-44}	0.295	0.162	0.441	ARID5B	3.210×10^{-5}	0.202	61928110, 62306779
15	65877726	rs12438510	15q22.31	G	A	1.23 (1.10,1.39)	3.011×10^{-4}	0.000	0.000	0.340	RAB11A	3.916×10^{-5}	6.718×10^{-5}	65540837, 66146213

***Table 3**: Odds Ratios, 95% CI's, and p-values of top SNPs from regional association testing, under the admixture peak, as well as the crude and adjusted p-values of the admixture mapping tag SNP.* *The regional significance thresholds, determined via permutation to account for the regional multiple testing burden, are 7.906x10^{-5}, 6.256x10^{-4}, 1.905x10^{-3}, and 1.270x10^{-4} for 2q21.2, 7p12.2, 10q21.2, and 15q22.31, respectively.*Risk allele frequencies (RAF) per ancestry group were obtained from European (Orcadian, Russian, Tuscan, French, Basque, Bergamo, Sardinian, Adygei), African (Bantu, Mbuti, Biaka, San, Yoruba, Mandenka), and Indigenous American (Surui, Karitiana, Maya, Pima, Colombian) superpopulation data from the Human Genome Diversity Project (accessed from the Genome Aggregation Database[38] v4).*

451 **Discussion:**

452 Motivated by the increased risk of ALL in Hispanic/Latino children, we carried out a systematic
453 investigation into the impact of IA-like ancestry on childhood ALL risk, both at the local and global
454 level. We found that, socio-economic status indicators confound the relationship between global
455 IA-like ancestry and ALL risk, and adjustment for these variables attenuates the association. The
456 lack of global ancestry association, however, did not preclude an association at the local level, as
457 IA-like ancestry was significantly associated with ALL risk across known loci and risk alleles
458 segregated with the IA haplotype more often than expected. On the other hand, we found no
459 interaction between IA-like ancestry and the risk genotype at either the local or global levels,
460 suggesting that there is little evidence the risk alleles exert different effects in different ancestry
461 contexts. Together, these results suggest that, in addition to environmental factors, risk alleles
462 present in higher frequencies in IA populations may contribute to Hispanic/Latino individuals'
463 excess risk of B-cell ALL.

464

465 Further, our in-depth investigation into inferred genetic ancestry also adds to and clarifies previous
466 findings. A previous study found a positive, statistically significant association between global IA-
467 like ancestry and ALL risk[25] . This result may have been affected by selection bias, due to the
468 lower participation of controls than cases in lower SES areas[52]. Associations between global
469 ancestries and disease risks are often reported, but these associations are often attenuated once
470 non-genetic factors are included in the model or when a family-based design is used[53,54]. Indeed,
471 observed associations between ancestry and type-2 diabetes risk in Hispanic/Latinos[26] and
472 between ancestry and hypertension and lung cancer in African Americans[55,56] were attenuated
473 once socio-economic covariates were accounted for. In our analysis of the CCLS dataset, we
474 found that the association of ALL risk with global IA-like ancestry was attenuated and no longer
475 statistically significant (though trending in the positive direction) after adjusting for income and
476 education level. We acknowledge that disentangling the association between global genetic

477 ancestry and ALL risk is complex, especially given the observed correlations between SES
478 indicators, ALL risk, and genetic ancestry reported in this study and elsewhere[57,58]. In the CCRLP,
479 however, in which controls were ascertained through statewide birth records[32] and thus should
480 not be impacted by selection bias, we found no significant association between genetic ancestry
481 and ALL risk.
482
483 Our admixture mapping identified associations on chromosomes 2, 7, 10, and 15, but SNPs at
484 the putatively novel loci on chromosomes 2 and 15 did not replicate in additional Hispanic/Latino
485 cohorts. The chromosome 2 locus (2q21.2) is a relative gene desert but overlaps *MGAT5*, which
486 has been reported to be involved in cancer metastasis and tumor suppression in mice[59] . Further,
487 transcriptional regulation of the gene is driven by the Ras-Raf-Ets pathway, which is oncogenic[60–
488 62]. On chromosome 15, the gene closest to the putative risk SNP is *RAB11A*, a member of the
489 RAS oncogene family, and has been previously associated with risk of lung squamous cell
490 carcinoma[63] and ovarian cancer[64] . Further, this region contains multiple SNP associations with
491 blood cell phenotypes[65,66], suggesting a possible connection to the etiology of childhood ALL[15] .
492 Adjusting for the lead SNP did not strongly attenuate the chromosome 15 association identified
493 by admixture mapping, suggesting that more complex genomic variation such as structural
494 variants, which may be in low LD with imputed SNPs[67], could underlie this association and would
495 require further validation. While there may be biological plausibility for these potentially novel loci,
496 the lack of replication could also be due to lack of statistical power, heterogeneity of
497 Hispanic/Latino populations across the United States, or that our initial findings were statistical
498 artefacts.
499
500 Within the genetic epidemiology literature, careful consideration has been given to the use of self-
501 reported race/ethnicity or genetically inferred ancestry as strata in performing GWAS. For
502 instance, it is common to define “White British” based on demographic data complemented by

503 principal components of the genotype data when analyzing UK Biobank to remove concerns of
504 stratification[68–70]. However, analysis in a relatively homogenous cohort such as the UKB is not
505 completely free from the effect of stratification[71], and the notion of defining a homogenous subset
506 of the cohort becomes even more difficult when performing analysis in multi-ancestry cohorts or
507 admixed populations. In those cases, one can pool samples across all ancestries for a "mega-
508 analysis"[72] , or perform stratified analysis based on either (1) self-reported demographic
509 labels[13,32,73] or (2) some arbitrary heuristics using genetically inferred ancestry[14,18,74,75]. The
510 downside of using genetically inferred ancestry is the prescription of arbitrary thresholds to define
511 ancestry strata. Across studies, the inclusion criteria vary, with individuals of self-reported White,
512 Black and Asian race/ethnicity being defined as study participants with over 75-95% European-
513 like, 50-90% African-like, and 75-90% Asian-like ancestries, respectively[14,18,74–77].
514
515 Previous genetic studies involving Hispanic/Latino participants have defined their strata as
516 individuals whose IA ancestry was greater or equal to 10% *and* greater than the proportion of
517 African ancestry[75,78,79]. Given the heterogeneity of Hispanics/Latinos as a population, this method
518 of categorization may prove untenable. The admixture proportions of Hispanic/Latino individuals
519 vary by country of origin with Caribbean and mainland Hispanics/Latinos exhibiting the highest
520 levels of African-like and IA-like ancestry, respectively[80,81]. Moreover, the estimated admixture
521 proportions can vary depending on the reference samples available, as well as the geographic
522 and temporal definitions of an ancestry component one wishes to estimate. Given these nuances,
523 we have tended to favor relying on self-reported demographic labels as strata when performing
524 multi-ancestry GWAS[32,82] or when limiting our analyses to Hispanics/Latinos[12]. Further,
525 the results of our HARE analysis showed that the algorithm rarely has enough evidence to exclude
526 an individual from analysis when self-reported labels are present, and imputing missing labels
527 have low recall rate for a heterogenous population like the Latinos. In cases where self-reported
528 labels are missing at high proportions, it may be preferable to conduct a mega-analysis instead

529 of stratified analysis. Indeed, even if overt categorizations impact little of the statistical power of
530 the study design and may provide a bit more robustness to the study, the underlying message
531 conveyed by explicit categorization of people based on estimated genetic ancestry may bring
532 more harm than benefits.
533
534 Finally, the results of our study have implications on future investigations of the genetic
535 architecture of ALL and designs of ancestry-based analyses. As reflected in our findings, the
536 elevated ALL risk in Hispanic/Latino children is to some extent driven by the higher frequencies
537 of known ALL risk alleles on IA-like haplotypes, which may underlie the shift towards higher mean
538 polygenic risk scores for ALL in Hispanic/Latino individuals compared to in non-Latino White
539 individuals [82,83]. Moreover, these consistent directional associations with risk could be indicative
540 of a signature of past adaptation. There is some evidence that ALL risk loci, such as the *IKZF1*
541 locus that we found to be genome-wide significant in admixture mapping, may be under positive
542 selection in Hispanic/Latino populations[12]. Given the overall enrichment of risk alleles on IA
543 haplotypes, the selection may be more oligo-genic or polygenic, though careful scrutiny is needed
544 to rule out other potential confounding[68,84]. Irrespective of whether ALL risk loci are under transient
545 directional selection, the increased prevalence of risk alleles on IA-like haplotypes suggest that
546 continued discovery in larger cohorts with IA-like ancestries and/or utilizing mapping approaches
547 that incorporate signals from both ancestry and allelic associations may be fruitful in discovering
548 more ALL risk loci and in understanding the genetic contributions to risk disparities.
549
550
551
552
553
554

Acknowledgements

We would like to thank the families and patients who participated in each study for ALL, and Soyoung Jeon for helpful conversations contributing to this research. Data used in this study were generated with supports from multiple grants from the National Institutes of Health (R01CA155461, R01CA175737, R01ES009137, P42ES004705, P50ES018172, P42ES0470518, R24ES028524, P20CA262733, U24ES028524, and R01CA262263), the Environmental Protection Agency (RD83615901), the Cancer Prevention and Research Institute of Texas (RP160771, RP210064), and St. Baldrick's Foundation (consortium grant 522277 with support from the Micaela's Army Foundation). Analysis performed in this study was supported by F31CA278359 to J.L., R01CA262263 to A.J.D., and R35GM142783 to C.W.K.C. A.J.D. is a Scholar of the Leukemia & Lymphoma Society. The content and conclusions from this work do not reflect the official views of the sponsors. The collection of cancer incidence data used in this study was supported by the California Department of Public Health pursuant to California Health and Safety Code Section 103885; Centers for Disease Control and Prevention's (CDC) National Program of Cancer Registries, under cooperative agreement 1NU58DP007156; the National Cancer Institute's Surveillance, Epidemiology and End Results Program under contract HHSN261201800032I awarded to the University of California, San Francisco, contract HHSN261201800015I awarded to the University of Southern California, and contract HHSN261201800009I awarded to the Public Health Institute. The ideas and opinions expressed herein are those of the authors and do not reflect the opinions of the State of California, Department of Public Health, the National Cancer Institute, and the Centers for Disease Control and Prevention or their Contractors and Subcontractors. The biospecimens and/or data used in this study were obtained from the California Biobank Program (CBP request #26 and #1380) Section 6555(b), 17 CCR). The California Department of Public Health is not responsible for the results or conclusions drawn by the authors of this publication. Computation of this work is supported by USC's Center for Advanced Research Computing (https://www.carc.usc.edu/).

Data Availability Statement

Genetic data from CCRLP and CCLS were obtained from the California Biobank. Because samples and data are property of the State of California, researchers who would like to request to use the data should establish a protocol with the California Department of Public Health Committee for the Protection of Human Subjects and other relevant agencies to understand if approval to share data are acceptable. The State has provided the following guidance on data sharing, as noted in the following statement: "California has determined that researchers requesting the use of California Biobank biospecimens for their studies will need to seek an exemption from NIH or other granting or funder requirements regarding the uploading of study results into an external bank or repository (including into the NIH dbGaP or other bank or repository). This applies to any uploading of genomic data and/or sharing of these biospecimens or individual data derived from these biospecimens. Such activities have been determined to violate the statutory scheme at California Health and Safety Code Section 124980 (j), 124991 (b), (g), (h) and 103850 (a) and (d), which protect the confidential nature of biospecimens and individual data derived from biospecimens. Investigators may agree to share aggregate data on SNP frequency and their associated p values with other investigators and may upload such frequencies into repositories including the NIH dbGaP repository providing: a) the denominator from which the data is derived includes no fewer than 20,000 individuals; b) no cell count is for <5 individuals; and c) no correlations or linkage probabilities between SNPs are provided." Our data, from CCRLP and CCLS, are each comprised of less than 20,000 subjects and therefore cannot be uploaded to a repository such as dbGaP. Data from Children's Oncology Group and REDIAL were obtained by special request (Yang et al., 2011; Harris et al., 2023). Data from MESA (phs000209.v13.p3) and HCHS/SOL (phs000810.v2.p2.c1, phs000810.v2.p2.c2) were obtained from dbGaP.

607 **References**
608

609 1. Bhojwani, D., Yang, J. J. & Pui, C.-H. Biology of Childhood Acute Lymphoblastic Leukemia.
610 *Pediatr. Clin. North Am.* **62**, 47–60 (2015).
611 2. Rheingold, S. R. *et al.* Determinants of survival after first relapse of acute lymphoblastic
612 leukemia: a Children's Oncology Group study. *Leukemia* **38**, 2382–2394 (2024).
613 3. Devilli, L. *et al.* Long-term and quality of survival in patients treated for acute lymphoblastic
614 leukemia during the pediatric age. *Hematol. Rep.* **13**, 8847 (2021).
615 4. Pui, C.-H., Robison, L. L. & Look, A. T. Acute lymphoblastic leukaemia. *The Lancet* **371**,
616 1030–1043 (2008).
617 5. Sasaki, K. *et al.* Acute lymphoblastic leukemia: A population-based study of outcome in the
618 United States based on the surveillance, epidemiology, and end results (SEER) database,
619 1980–2017. *Am. J. Hematol.* **96**, 650–658 (2021).
620 6. American Cancer Society. Key Statistics for Acute Lymphocytic Leukemia (ALL). *American*
621 *Cancer Society* https://www.cancer.org/cancer/types/acute-lymphocytic-leukemia/about/key-
622 statistics.html (2024).
623 7. Feng, Q. *et al.* Trends in Acute Lymphoblastic Leukemia Incidence in the United States by
624 Race/Ethnicity From 2000 to 2016. *Am. J. Epidemiol.* **190**, 519–527 (2021).
625 8. Quiroz, E. *et al.* Immunophenotype of acute lymphoblastic leukemia in minorities- analysis
626 from the SEER database. *Hematol. Oncol.* **40**, 105–110 (2022).
627 9. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164
628 (2016).
629 10. Mills, M. C. & Rahal, C. The GWAS Diversity Monitor tracks diversity by disease in real
630 time. *Nat. Genet.* **52**, 242–243 (2020).

631 11. de Smith, A. J., Jiménez-Morales, S. & Mejía-Aranguré, J. M. The genetic risk of acute
632 lymphoblastic leukemia and its implications for children of Latin American origin. *Front.*
633 *Oncol.* **13**, 1299355 (2024).
634 12. de Smith, A. J. *et al.* A noncoding regulatory variant in IKZF1 increases acute
635 lymphoblastic leukemia risk in Hispanic/Latino children. *Cell Genomics* **4**, (2024).
636 13. Jeon, S., Smith, A. J. & Li, S. Genome-wide trans-ethnic meta-analysis identifies novel
637 susceptibility loci for childhood acute lymphoblastic leukemia. *Leukemia* **36**, 865–868 (2022).
638 14. Vijayakrishnan, J. *et al.* Identification of four novel associations for B-cell acute
639 lymphoblastic leukaemia risk. *Nat. Commun.* **10**, 5348 (2019).
640 15. Kachuri, L. *et al.* Genetic determinants of blood-cell traits influence susceptibility to
641 childhood acute lymphoblastic leukemia. *Am. J. Hum. Genet.* **108**, 1823–1835 (2021).
642 16. Qian, M. *et al.* Novel susceptibility variants at the ERG locus for childhood acute
643 lymphoblastic leukemia in Hispanics. *Blood* **133**, 724–729 (2019).
644 17. Xu, H. *et al.* ARID5B genetic polymorphisms contribute to racial disparities in the
645 incidence and treatment outcome of childhood acute lymphoblastic leukemia. *J. Clin. Oncol.*
646 *Off. J. Am. Soc. Clin. Oncol.* **30**, 751–757 (2012).
647 18. Perez-Andreu, V. *et al.* Inherited GATA3 variants are associated with Ph-like childhood
648 acute lymphoblastic leukemia and risk of relapse. *Nat. Genet.* **45**, 1494–1498 (2013).
649 19. Swords, R. *et al.* Acute leukemia in adult Hispanic Americans: a large-population study.
650 *Blood Cancer J.* **6**, e484 (2016).
651 20. Pollyea, D. A. *et al.* Acute Leukemia in Hispanic Americans: Incidence and Incidence
652 Rate Differences by Nativity. *Blood* **120**, 3160 (2012).

653 21. Dong, Y. *et al.* Leukemia incidence trends at the global, regional, and national level
654 between 1990 and 2017. *Exp. Hematol. Oncol.* **9**, 14 (2020).
655 22. Yi, M., Zhou, L., Li, A., Luo, S. & Wu, K. Global burden and trend of acute
656 lymphoblastic leukemia from 1990 to 2017. *Aging* **12**, 22869–22891 (2020).
657 23. Bryc, K., Durand, E. Y., Macpherson, J. M., Reich, D. & Mountain, J. L. The genetic
658 ancestry of african americans, latinos, and european Americans across the United States. *Am.*
659 *J. Hum. Genet.* **96**, 37–53 (2015).
660 24. Homburger, J. R. *et al.* Genomic Insights into the Ancestry and Demographic History of
661 South America. *PLOS Genet.* **11**, e1005602 (2015).
662 25. Walsh, K. *et al.* Associations between genome-wide Native American ancestry, known
663 risk alleles and B-cell ALL risk in Hispanic children. *Leukemia* **27**, 10.1038/leu.2013.130
664 (2013).
665 26. Florez, J. C. *et al.* Strong association of socioeconomic status with genetic ancestry in
666 Latinos: implications for admixture studies of type 2 diabetes. *Diabetologia* **52**, 1528–1536
667 (2009).
668 27. Shriver, M. D. *et al.* Skin pigmentation, biogeographical ancestry and admixture
669 mapping. *Hum. Genet.* **112**, 387–399 (2003).
670 28. Qin, H. & Zhu, X. Power comparison of admixture mapping and direct association
671 analysis in genome-wide association studies. *Genet. Epidemiol.* **36**, 235–243 (2012).
672 29. McKeigue, P. M. Prospects for Admixture Mapping of Complex Traits. *Am. J. Hum.*
673 *Genet.* **76**, 1–7 (2005).
674 30. Committee on the Use of Race, Ethnicity, and Ancestry as Population Descriptors in
675 Genomics Research *et al. Using Population Descriptors in Genetics and Genomics Research:*

676 *A New Framework for an Evolving Field*. 26902 (National Academies Press, Washington,
677 D.C., 2023). doi:10.17226/26902.
678 31. Zavala, V. A. *et al.* Cancer health disparities in racial/ethnic minorities in the United
679 States. *Br. J. Cancer* **124**, 315–332 (2021).
680 32. Wiemels, J. L. *et al.* GWAS in childhood acute lymphoblastic leukemia reveals novel
681 genetic associations at chromosomes 17q12 and 8q24.21. *Nat. Commun.* **9**, 286 (2018).
682 33. Ma, X., Buffler, P. A., Layefsky, M., Does, M. B. & Reynolds, P. Control Selection
683 Strategies in Case-Control Studies of Childhood Diseases. *Am. J. Epidemiol.* **159**, 915–921
684 (2004).
685 34. Harris, R. D. *et al.* Ethnic-Specific Predictors of Neurotoxicity Among Pediatric Acute
686 Lymphoblastic Leukemia Patients Following High-Dose Methotrexate. *Cancer* **129**, 1287–
687 1294 (2023).
688 35. Sheng, X. *et al.* Inverted genomic regions between reference genome builds in humans
689 impact imputation accuracy and decrease the power of association testing. *Hum. Genet.*
690 *Genomics Adv.* **4**, (2023).
691 36. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative
692 modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–
693 288 (2013).
694 37. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in
695 141,456 humans. *Nature* **581**, 434–443 (2020).
696 38. Li, S. *et al.* Localized variation in ancestral admixture identifies pilocytic astrocytoma
697 risk loci among Latino children. *PLoS Genet.* **18**, e1010388 (2022).

698 39. Kowalski, M. H. *et al.* Use of 100,000 NHLBI Trans-Omics for Precision Medicine
699 (TOPMed) Consortium whole genome sequences improves imputation quality and detection
700 of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS Genet.*
701 **15**, 1008500 (2019).
702 40. Cahoon, J. L. *et al.* Imputation accuracy across global human populations. *Am. J. Hum.*
703 *Genet.* **111**, 979–989 (2024).
704 41. Metayer, C. *et al.* Tobacco Smoke Exposure and the Risk of Childhood Acute
705 Lymphoblastic and Myeloid Leukemias by Cytogenetic Subtype. *Cancer Epidemiol.*
706 *Biomarkers Prev.* **22**, 1600–1611 (2013).
707 42. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation
708 for Statistical Computing (2023).
709 43. Qian, M. *et al.* Genome-Wide Association Study of Susceptibility Loci for T-Cell Acute
710 Lymphoblastic Leukemia in Children. *J. Natl. Cancer Inst.* **111**, 1350–1357 (2019).
711 44. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-
712 Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
713 45. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer
714 datasets. *GigaScience* **4**, 7 (2015).
715 46. Brown, L. A. *et al.* Admixture Mapping Identifies an Amerindian Ancestry Locus
716 Associated with Albuminuria in Hispanics in the United States. *J. Am. Soc. Nephrol. JASN* **28**,
717 2211–2220 (2017).
718 47. Schick, U. M. *et al.* Genome-wide Association Study of Platelet Count Identifies
719 Ancestry-Specific Loci in Hispanic/Latino Americans. *Am. J. Hum. Genet.* **98**, 229–242
720 (2016).

721 48. Qin, H. & Zhu, X. Power Comparison of Admixture Mapping and Direct Association
722 Analysis in Genome-Wide Association Studies. *Genet. Epidemiol.* **36**, 235–243 (2012).
723 49. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of
724 genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
725 50. Fang, H. *et al.* Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in
726 Genome-wide Association Studies. *Am. J. Hum. Genet.* **105**, 763–772 (2019).
727 51. Johnson, J. L. & Abecasis, G. R. GAS Power Calculator: web-based power calculator for
728 genetic association studies. 164343 Preprint at https://doi.org/10.1101/164343 (2017).
729 52. Wheeler, D. C., Boyle, J., Carli, M., Ward, M. H. & Metayer, C. Neighborhood
730 Deprivation, Indoor Chemical Concentrations, and Spatial Risk for Childhood Leukemia. *Int.*
731 *J. Environ. Res. Public. Health* **20**, 3582 (2023).
732 53. Sun, H. *et al.* The impact of global and local Polynesian genetic ancestry on complex
733 traits in Native Hawaiians. *PLOS Genet.* **17**, e1009273 (2021).
734 54. Asgari, S. *et al.* A positively selected FBN1 missense variant reduces height in Peruvians.
735 *Nature* **582**, 234–239 (2020).
736 55. Aldrich, M. C. *et al.* Socioeconomic Status and Lung Cancer: Unraveling the
737 Contribution of Genetic Admixture. *Am. J. Public Health* **103**, e73–e80 (2013).
738 56. Non, A. L., Gravlee, C. C. & Mulligan, C. J. Education, Genetic Ancestry, and Blood
739 Pressure in African Americans and Whites. *Am. J. Public Health* **102**, 1559–1565 (2012).
740 57. Wang, L., Gomez, S. L. & Yasui, Y. Racial and Ethnic Differences in Socioeconomic
741 Position and Risk of Childhood Acute Lymphoblastic Leukemia. *Am. J. Epidemiol.* **185**,
742 1263–1271 (2017).

743 58. Spector, L. G. *et al.* RE: 'RACIAL AND ETHNIC DIFFERENCES IN
744 SOCIOECONOMIC POSITION AND RISK OF CHILDHOOD ACUTE
745 LYMPHOBLASTIC LEUKEMIA'. *Am. J. Epidemiol.* **188**, 1192–1193 (2019).
746 59. Granovsky, M. *et al.* Suppression of tumor growth and metastasis in Mgat5-deficient
747 mice. *Nat. Med.* **6**, 306–312 (2000).
748 60. Hirata, A. *et al.* Increased risk of metastasis in patients with incidental use of renin-
749 angiotensin system inhibitors: a retrospective analysis for multiple types of cancer based on
750 electronic medical records. *Hypertens. Res.* **45**, 1869–1881 (2022).
751 61. Kang, R. *et al.* Transcriptional Regulation of the N-Acetylglucosaminyltransferase V
752 Gene in Human Bile Duct Carcinoma Cells (HuCC-T1) Is Mediated by Ets-1 *. *J. Biol. Chem.*
753 **271**, 26706–26712 (1996).
754 62. Buckhaults, P., Chen, L., Fregien, N. & Pierce, M. Transcriptional Regulation ofN-
755 Acetylglucosaminyltransferase V by the srcOncogene *. *J. Biol. Chem.* **272**, 19575–19581
756 (1997).
757 63. Gombodorj, N. *et al.* RAB11A Expression Is Associated With Cancer Aggressiveness
758 Through Regulation of FGFR-Signaling in Lung Squamous Cell Carcinoma. *Ann. Surg.*
759 *Oncol.* **29**, 7149–7162 (2022).
760 64. Wang, Y., Ren, Y., Li, N., Zhao, J. & Zhao, S. Rab11a promotes the malignant
761 progression of ovarian cancer by inducing autophagy. *Genes Genomics* **44**, 1375–1384
762 (2022).
763 65. Vuckovic, D. *et al.* The Polygenic and Monogenic Basis of Blood Traits and Diseases.
764 *Cell* **182**, 1214-1231.e11 (2020).

765 66. Chen, M.-H. *et al.* Trans-ethnic and ancestry-specific blood-cell genetics in 746,667
766 individuals from 5 global populations. *Cell* **182**, 1198-1213.e14 (2020).
767 67. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes.
768 *Nature* **526**, 75–81 (2015).
769 68. Sohail, M. *et al.* Polygenic adaptation on height is overestimated due to uncorrected
770 stratification in genome-wide association studies. *eLife* **8**, e39702 (2019).
771 69. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data.
772 *Nature* **562**, 203–209 (2018).
773 70. Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links
774 to Common Complex Disease. *Cell* **167**, 1415-1429.e19 (2016).
775 71. Chen, M. *et al.* Evidence of Polygenic Adaptation in Sardinia at Height-Associated Loci
776 Ascertained from the Biobank Japan. *Am. J. Hum. Genet.* **107**, 60–71 (2020).
777 72. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery for
778 complex traits. *Nature* **570**, 514–518 (2019).
779 73. Kachuri, L. *et al.* Genetically adjusted PSA levels for prostate cancer screening. *Nat.*
780 *Med.* **29**, 1412–1423 (2023).
781 74. Klarin, D. *et al.* Genetics of blood lipids among ~300,000 multi-ethnic participants of the
782 Million Veteran Program. *Nat. Genet.* **50**, 1514–1523 (2018).
783 75. Xu, H. *et al.* Novel susceptibility variants at 10p12.31-12.2 for childhood acute
784 lymphoblastic leukemia in ethnically diverse populations. *J. Natl. Cancer Inst.* **105**, 733–742
785 (2013).

786 76. Gautam, Y. *et al.* Genome-wide admixture and association analysis identifies African
787 ancestry specific risk loci of eosinophilic esophagitis in African American. *J. Allergy Clin.*
788 *Immunol.* **151**, 1337–1350 (2023).
789 77. Shrine, N. *et al.* Multi-ancestry genome-wide association analyses improve resolution of
790 genes and pathways influencing lung function and chronic obstructive pulmonary disease risk.
791 *Nat. Genet.* **55**, 410–422 (2023).
792 78. Archer, N. P. *et al.* Family-based exome-wide association study of childhood acute
793 lymphoblastic leukemia among Hispanics confirms role of ARID5B in susceptibility. *PLOS*
794 *ONE* **12**, e0180488 (2017).
795 79. Karol, S. E. *et al.* Genetics of ancestry-specific risk for relapse in acute lymphoblastic
796 leukemia. *Leukemia* **31**, 1325–1332 (2017).
797 80. Conomos, M. P. *et al.* Genetic diversity and association studies in US Hispanic/Latino
798 populations: applications in the Hispanic Community Health Study/Study of Latinos. *Am. J.*
799 *Hum. Genet.* **98**, 165–184 (2016).
800 81. Bryc, K. *et al.* Genome-wide patterns of population structure and admixture among
801 Hispanic/Latino populations. *Proc. Natl. Acad. Sci.* **107**, 8954–8961 (2010).
802 82. Jeon, S. *et al.* Genome-wide trans-ethnic meta-analysis identifies novel susceptibility loci
803 for childhood acute lymphoblastic leukemia. *Leukemia* **36**, 865–868 (2022).
804 83. Jeon, S. *et al.* Evaluating Genomic Polygenic Risk Scores for Childhood Acute
805 Lymphoblastic Leukemia in Latinos. *MedRxiv Prepr. Serv. Health Sci.* 2023.06.08.23291167
806 (2023) doi:10.1101/2023.06.08.23291167.
807 84. Berg, J. J. *et al.* Reduced signal for polygenic adaptation of height in UK Biobank. *eLife*
808 **8**, e39725 (2019).
809