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. Genetic variants, particularly ones found enriched in Indigenous American (IA)-like ancestry and inherited by Hispanics/Latinos, may contribute to this disparity. In this study, we characterized the impact of IA-like ancestry on overall ALL risk and the frequency and effect size of known risk alleles in a large cohort of self-reported Hispanic/Latino individuals. We also performed genome-wide admixture mapping analysis to identify potentially novel ALL risk loci. We found that global IA ancestry was positively associated with ALL risk, but the association was not significant after adjusting for socio-economic indicators. In a series of local ancestry analyses, we uncovered that at known ALL risk loci, increasing copies of the IA-like haplotype were positively and significantly associated with ALL case-control status. Further, the IA-like haplotype had ~1.33 times the odds of harboring the risk allele compared to non-IA-like haplotypes. We found no evidence of interaction between genotype and ancestry (local or global) in relation to ALL risk. Admixture mapping identified association signals on chromosomes 2 (2q21.2), 7 (7p12.2), 10 (10q21.2), and 15 (15q22.31); however, only the variants at 7p12.2 and 10q21.2 replicated in additional cohorts. 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, which can be leveraged as targets of new precision health strategies and therapeutics.

64 **Introduction:**

65 Acute lymphoblastic leukemia (ALL) is the most common childhood cancer and accounts for
66 25% of malignancies before the age of 20[1]. The long-term survival rate of the disease in
67 children is high, occurring upwards of 85-90% [2]; however, this survival decreases to 30-40% in
68 individuals over the age of 35, to approximately 40% after relapse[3–5], and in 2024 an estimated
69 1,330 adults and children died of ALL[6–9]. Despite decades of investigation into ALL risk, the
70 etiology of the disease is not well understood and disparities in risk persist. Children of
71 Hispanic/Latino ethnicity have 1.4 and 2 times the risk of developing ALL compared to their
72 Non-Latino White and Black counterparts, respectively[10] . This disparity in risk widens in margin
73 for Hispanic/Latino adults[10] and is specific to the B-cell immunophenotype, which comprises
74 approximately 85% of cases[11] . While environmental factors are involved in this increased risk,
75 genetic factors likely contribute at least in part to this disparity[12,13]
76
77 The genetic basis of ALL can be evaluated using migration studies, ecological data, time trends,
78 heritability, and admixture and genetic studies[14] . Previous migration studies have indicated that
79 foreign- and US-born Hispanics/Latinos have comparable rates of B-cell ALL to each other, and
80 both have higher (~2.13 times) incidence rates compared to Non-Latino Whites[15,16]. The stark
81 difference in incidence between Hispanics/Latinos and Non-Latino Whites, as well as the
82 similarities in rates between US-born and foreign-born Hispanics/Latinos, may indicate a
83 substantial genetic component to the disease, rather than a primarily environmental etiology.
84 Further, ecological data has indicated that, in 2017, the countries with highest age standardized
85 incidence rates of ALL were in Latin America (*i.e.* Honduras and Mexico), while countries with the
86 lowest incidence rates were in Africa, with European countries exhibiting intermediate
87 incidence[17,18] . Most directly, previous genome-wide association studies (GWAS) have estimated
88 the heritability of ALL to be approximately 18-21% in both Hispanic/Latino and Non-Latino White
89 populations, and approximately 24% of familial relative risk has been attributed to known ALL risk

variants[19–21] . Consistent with these estimates, GWAS have identified a number of loci associated with ALL susceptibility, including risk alleles located near genes such as *IKZF1*, *ARID5B*, *GATA3*, and *ERG* that appear to be enriched in frequency or show stronger effects in Hispanic/Latino populations[13,22–24].

Notable amongst populations with differing incidence rates of ALL is the presence or dominance of a particular ancestry component. Hispanics/Latinos are an admixed population - drawing ancestral components from the American, European, and African continents[25,26] . Previous literature has alluded to associations between Indigenous American (IA)-like ancestry and ALL, suggesting that every 20% increase in IA-like ancestry was associated with a 20% increase in odds of developing B-cell ALL[27]. Because Hispanics/Latinos inherited a significant proportion of their genetic ancestries from Indigenous Americans (though there is geographical heterogeneity across the Americas), these associations with respect to global IA-like ancestries may be consistent with the migration and ecological findings above. However, associations with global ancestry can be confounded by non-genetic factors such as socio-economic status[28], for which the aforementioned analyses did not adjust. Relevant to the discussion of the role of the IA-like haplotype in ALL risk is local ancestry, or ancestry in a particular location of the genome[29]. Local ancestry analyses are less susceptible to confounding by environmental factors, and have yielded promising results in determining the association between IA-like ancestry and leukemia risk. For example, a study of the *ERG* locus found that Hispanics/Latinos who had 1 or more copies of the IA-like haplotype at the locus had 1.15-1.30 times the odds of developing ALL compared to Hispanics/Latinos with 0 copies of the haplotype[30] . Together, these data suggest that in addition to the environmental basis of disease, higher frequencies of risk alleles present in the IA-like ancestry component inherited by many Hispanics/Latinos may be contributing to disproportionate risk of ALL.

116 Given the important role of genetic factors and genetic ancestry in ALL risk, we undertook a
117 systematic admixture analysis in the largest ALL cohort of self-reported Hispanic/Latino children
118 to date. We investigated both inferred global ancestry across the genome and local ancestry at
119 known ALL loci for associations with risk, and further conducted admixture mapping to detect
120 novel risk loci that may exhibit frequency differences between ancestries and, consequently,
121 contribute to the disparities between populations. We note that given our goal to further
122 understand the disparity in ALL risk epidemiologically observed among Hispanic/Latino children,
123 we focused on inferring the genetic ancestries and characterizing their impact on ALL from self-
124 reported Hispanic/Latino individuals. We defined genetic ancestries at the continental (e.g. Africa,
125 Europe, and Americas) level, representing patterns of genetic similarities across geographical
126 space over the last ~10-50k years due to human migrations and isolation-by-distance. These
127 relatively distinct components of ancestries, though arbitrarily defined, are only recently reunited
128 in admixed populations, such as in Hispanics/Latinos. These components are inferred and defined
129 relative to reference populations available to us, and are leveraged to uncover associations with
130 traits in this research setting, but are invariably a simplified representation of the complex nature
131 of genetic ancestries. We thus append the suffix "-like" to refer to these ancestries to remind the
132 reader of this simplification and uncertainty[31].
133
134 **Methods:**
135
136 Study Subjects and Quality Control
137 We use the term "Hispanics/Latinos" as a population descriptor of our study subjects to align with
138 both the self-reported race/ethnicity categories and the definitions of the USA Office and
139 Management Budget[32] , while acknowledging that we focus on individuals of Latin American
140 origin. Our analysis utilized the following cohorts of self-reported Hispanics/Latinos: the California
141 Cancer Linkage Project (CCRLP), the Kaiser Resource for Genetic Epidemiology Research on

142 Aging Cohort (Kaiser-GERA), the California Childhood Leukemia Study (CCLS), Children's
143 Oncology Group (COG), Multi-Ethnic Study of Atherosclerosis (MESA), the Adolescent and
144 Childhood Cancer Epidemiology and Susceptibility Service for Texas and Reducing Ethnic
145 Disparities in Acute Leukemia (ACCESS/REDIAL), and the Hispanic Community Health
146 Study/Study of Latinos (HCHS/SOL). All relevant information on sample size, arrays, study
147 design, and references/dbGaP accession numbers can be found in **Table 1**. Following the
148 practice in previous studies[13,19,33], cases from CCRLP were augmented with external controls
149 from Kaiser-GERA due to being genotyped on a similar array. Cases from COG were paired with
150 external controls from MESA and ACCESS/REDIAL cases were paired with external controls from
151 HCHS/SOL. HCHS/SOL was a multi-center study, with study participants being ascertained from
152 Bronx, New York; Chicago, Illinois; Miami, Florida; and San Diego, California. The SOL San Diego
153 center was chosen as controls for ACCESS/REDIAL due to similarity in the first 2 principal
154 components and estimated ancestry distributions (**Supplementary Figures 1-2**).
155

156
157

Study	In-Text Name/ Acronym	Study Design	N	Genotype Array	Reference/ Accession Number	Usage
California Cancer Linkage Project	CCRLP	Population Based case-control study	4033	Affymetrix Axiom World Array	Wiemels et al., 2018[33]	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[34]	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[35]	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)

158
159 ***Table 1: All cohorts included in the analysis.*** *Information includes title, design, sample size, genotype array, dbGaP accession*
160 *number or reference, and how each study was used in our analysis.*

All analyses were limited to autosomes. Within each study, SNPs and individuals were removed based on the following pre-defined quality control thresholds: SNP call rate < 0.98, MAF < 0.01, $p_{HWE} < 1 \times 10^{-5}$, high relatedness ($\hat{\pi} > 0.2$), Genome Heterozygosity Rate > 6 standard deviations from the mean, or sample call rate < 0.95. Palindromic SNPs were also excluded as they could impact imputation accuracy due to unidentified strand flip or inversion between genome build[36] . The number of individuals and SNPs remaining can be found in **Supplementary Table 1**. Local ancestry was inferred using RFMIX[37]. Reference haplotypes were obtained from the genome aggregation database (gnomAD)[38] version 3, including 708 African, 671 Non-Finnish European, and 94 Indigenous American individuals previously chosen for ancestry inference in Hispanics/Latinos[39] . After restricting to the intersection of SNPs across each paired dataset, each dataset was imputed separately using the TopMed Imputation Server[40] (Kowalski et al., 2019). We observed systemic differences in imputation quality as measured by the imputation R^2 between cases and external controls, presumably due to batch effects in data generation or subtle differences in ancestry compositions between cohorts and differential imputation accuracies across populations[41]. Consequently, for all comparisons of cases to controls, we imposed filters such that only SNPs with an imputation $R^2 > 0.30$ in both cases and controls and a R^2 difference < 0.15 between cases and controls were included in the analysis.

<u>Harmonizing self-reported ethnicity and genetically inferred ancestry labels in CCRLP/Kaiser-GERA</u>

To explore our data and investigate the concordance of self-reported race/ethnicity information and genetically inferred ancestry, we plotted the first 20 Principal Components of multi-ethnic CCRLP-Kaiser/GERA study participants (10450 Hispanic/Latino, 2191 Non- Latino African, 5335 Non-Latino East Asian, 58503 Non-Latino European) and implemented a software, Harmonizing Ancestry and Race Ethnicity[42] (HARE). HARE is a two-step algorithm that identifies individuals

who may be discordant between self-identified race/ethnicity (SIRE) and genetically inferred ancestry (GIA), given the mapping of the two learned from the rest of the dataset. The first step is a training phase in which genetically inferred ancestry based on principal components are used to predict available self-reported race/ethnicity in a support vector machine. The second step is HARE strata assignment, in which the highest probability of being assigned to a stratum is contrasted with the probability that the individual belongs to their self-reported stratum using preset thresholds. We ran HARE with the default parameters and thresholds[42] . The multi-class support vector machine, with a radial kernel, was trained on the first 30 principal components calculated using PLINK[43] (v2.0) of genotype data from individuals of across all race/ethnicity groups from CCRLP/Kaiser-GERA to predict self-reported race/ethnicity. Then, the HARE algorithm compared the predicted probability of membership in the most likely strata based on genetics to the predicted probability of membership of an individual's SIRE. If the ratio of the probabilities exceeds a pre-defined threshold, an individual with existing SIRE could be assigned as missing (*i.e.* removed from analysis), while an individual without SIRE would be assigned into the top predicted strata. We first evaluated if the HARE algorithm would assign any of the study participants as missing due to discordance between SIRE and GIA. We also assessed the rate of discordance between SIRE and GIA by performing 10-fold cross validation in which we randomly masked the SIRE information of 10% of study participants for each race/ethnicity group in each trial. Percent discordance was defined as the number of individuals with self-identified race-ethnicity data classified as missing or as a stratum not corresponding to the true self-reported race/ethnicity label.

213 Global Ancestry Analyses

214 Global ancestry for each study participant was calculated by averaging local ancestry across the
215 genome. We performed logistic regression of ALL case-control status on global IA-like ancestry
216 in CCRLP and CCLS, studies in which indicators of socio-economic status (SES) variables were
217 available. Covariates included African-like ancestry, sex, and SES indicators (highest education
218 obtained by a parent and annual household income). Income was categorized in 6 strata, in
219 accordance with its treatment in CCLS[44] : <$15,000, $15,000-$29,999, $30,000-$44,999,
220 $45,000-$59,999, $60,000-$74,999, and >$75,000. In CCLS, highest educational attainment of
221 either parent was divided into 4 categories: Less than elementary school, high school, some
222 college, and a bachelor's degree or higher. In CCRLP, the strata were less than 9th grade, 9th-12th
223 grade, and some college or more. Both SES indicators were treated as ordinal categorical
224 variables, in which the lowest category was used as a reference. Linearity assumptions were
225 assessed using residual and QQ plots. Analysis was carried out using R v.4.3.1[45].

226

227 Enrichment of risk allele on IA-like ancestry haplotypes

228 We performed a series of local ancestry analyses at known ALL loci. A list of all currently known
229 ALL loci was obtained from[12] . We restricted analyses to SNPs that were independent at the local
230 ancestry level (N=20), which was determined via PLINK's clumping algorithm[46]. SNPs were
231 clumped if the local ancestry calls covering the SNPs were 90% correlated within 10 Mb of each
232 other. We excluded the T-cell ALL-specific association locus at *USP7*[47] given that the increased
233 risk of ALL in Hispanics/Latinos is limited to B-cell ALL. Across known ALL loci[12], we cross-
234 tabulated local genetic ancestry (number of copies of the IA-like, African-like, or European-like
235 haplotypes) and risk allele dosage. Then, we identified randomly in-sample 20 frequency-matched
236 alleles from across the genome and cross tabulated the occurrence of this pseudo risk allele with
237 ancestry dosage to establish a null distribution. Co-occurrence of the risk allele with a particular
238 ancestry haplotype would imply association. Differences between the null and known loci

239 distribution were evaluated via a Chi-Squared Goodness of Fit test. We also performed a logistic
240 regression of ALL case-control status on IA-like ancestry dosage at each locus, including as
241 covariates global African-like and European-like ancestry, and sex. To test for presence of
242 heterogeneous effect sizes across ancestries, we performed logistic regression of ALL case-
243 control status on risk allele genotype, IA-like ancestry dosage, sex, the first 20 principal
244 components, and an interaction term of local ancestry and risk allele dosage. Analyses were
245 carried out either using PLINK v2.0[43] or R v.4.3.1[45].
246
247 Admixture mapping, regional association testing, and replication
248 We performed a genome-wide scan for loci in which IA-like ancestry was associated with ALL
249 risk. We regressed ALL case-control status on the local IA ancestry dosage, global African-like
250 and European-like ancestry, sex, and the first 20 principal components. We performed this
251 analysis in CCRLP/Kaiser as our discovery dataset. The significance threshold for genome-wide
252 admixture mapping was 5.7 x 10^{-5}, as suggested in literature previously for Hispanic/Latino
253 populations[48,49].
254
255 After identifying potential signal through admixture mapping in CCRLP/Kaiser, we then fine-
256 mapped the admixture peak through single-variant association testing in the same dataset. The
257 admixture peak region[50] was defined as a range of positions that contains all SNPs with a -
258 log10(P) between -log10(P_{am}) and -log10(P_{am}) – 1, with P_{am} representing the p-value associated
259 with the lead SNP from admixture mapping. Genotypes of all imputed SNPs within this region
260 passing QC filtering were tested via logistic regression, with sex and the first 20 principal
261 components as covariates. To compute the significance threshold for each region, we permuted
262 case-control labels 1000 times and recorded the minimum p-value per permutation. The fifth
263 percentile of all permutation trials was recorded as the significance threshold per region. The most
264 significant variant from each peak region was assumed to be the putative causal variant, which

265 we then verified through a conditional analysis by including it as a covariate in the admixture
266 mapping regression model to determine if the admixture signal would be attenuated. We
267 attempted to replicate the signals at the lead variants from regional association testing and
268 conditional analysis in additional Hispanic/Latino cohorts: CCLS, COG/MESA, and SOL-
269 SD/REDIAL. The replication cohorts were meta-analyzed together by inverse variance weighting
270 in the software *METAL*[51].
271
272 IRB Statement:
273 This study was approved by the institutional review boards of the California Health and Human
274 Services Agency, the University of Southern California, Children's Oncology Group, Baylor
275 College of Medicine, Yale University, and University of California, Berkeley. All study participants
276 were de-identified and newborn bloodspots from the CCRLP were obtained via a consent waiver
277 from Committee for the Protection of Human Subjects of the State of California.
278
279
280 **Results:**
281
282 GWAS are often conducted stratified by race/ethnicity or ancestry in order to identify population-
283 specific associations or to control for confounding due to population stratification[52] (Peterson et
284 al., 2019). The stratifying demographic variable can appear in the form of self-reported race and
285 ethnicity, such as information from electronic health records, or inferred ancestry from genetic
286 data, such as through principal component analysis or global ancestry estimates. Usage of one
287 over the other is dependent on the goals of the study. Our study investigates genetic ancestry as
288 it relates to ALL and is largely based on the Hispanic/Latino subset (N = 4033) of the CCRLP
289 study, a population-based case-control study established in California. Moreover, we leverage
290 previous study designs, including the CCRLP and additional Hispanic/Latino cohorts, to map

291 potentially novel loci associated with ALL via admixture mapping. CCRLP, supplemented with
292 external controls from the Kaiser-GERA cohort, is the largest cohort we have access to and serves
293 as the discovery cohort (**Table 1**). Consequently, we first sought to characterize the distribution
294 of genetic ancestries in CCRLP, and to re-examine the impact, if any, of using self-reported
295 Hispanic/Latino ethnicity for our analyses.
296
297 *Harmonizing self-reported ethnicity with genetic ancestry in CCRLP/Kaiser-GERA*
298 The CCRLP/Kaiser-GERA cohort[19,33] (Wiemels et al., 2018; Jeon et al., 2022) is comprised of
299 four self-reported strata of ethnicities: Non-Latino Black (AFR; N = 2,191), Non-Latino East Asian
300 (EAS; N = 5,335), Hispanic/Latino (LAT; N = 10,450), and Non-Latino White (NLW; N = 58,503).
301 We examined the first 20 axes of the principal components analysis applied to the genetic data
302 (**Figure 1, Supplementary Figure 3**). Consistent with previous multi-ethnic studies (Wojcik et al.,
303 2019), on the first two axes, individuals fall into a cline according to their genetic ancestry, with
304 the three main components being driven by ancestries found most prevalently among self-
305 reported non-Latino African individuals, non-Latino European individuals, and non-Latino East
306 Asian individuals, representing the three vertices of the plot (**Figure 1**). In these two axes, self-
307 reported Hispanic/Latino individuals are scattered between all three ancestry components. While
308 they may appear to be genetically similar to those self-reported as East Asians in the cohort, the
309 third principal component axis shows the divergence between Hispanics/Latinos and East Asians,
310 reflecting the divergence between East Asian-like and Indigenous American (IA)-like ancestries
311 (**Supplementary Figure 3**. The relatively higher similarity between these two ancestries due to
312 their more recently shared history[53] likely contributed to the pattern on the first two PCs.

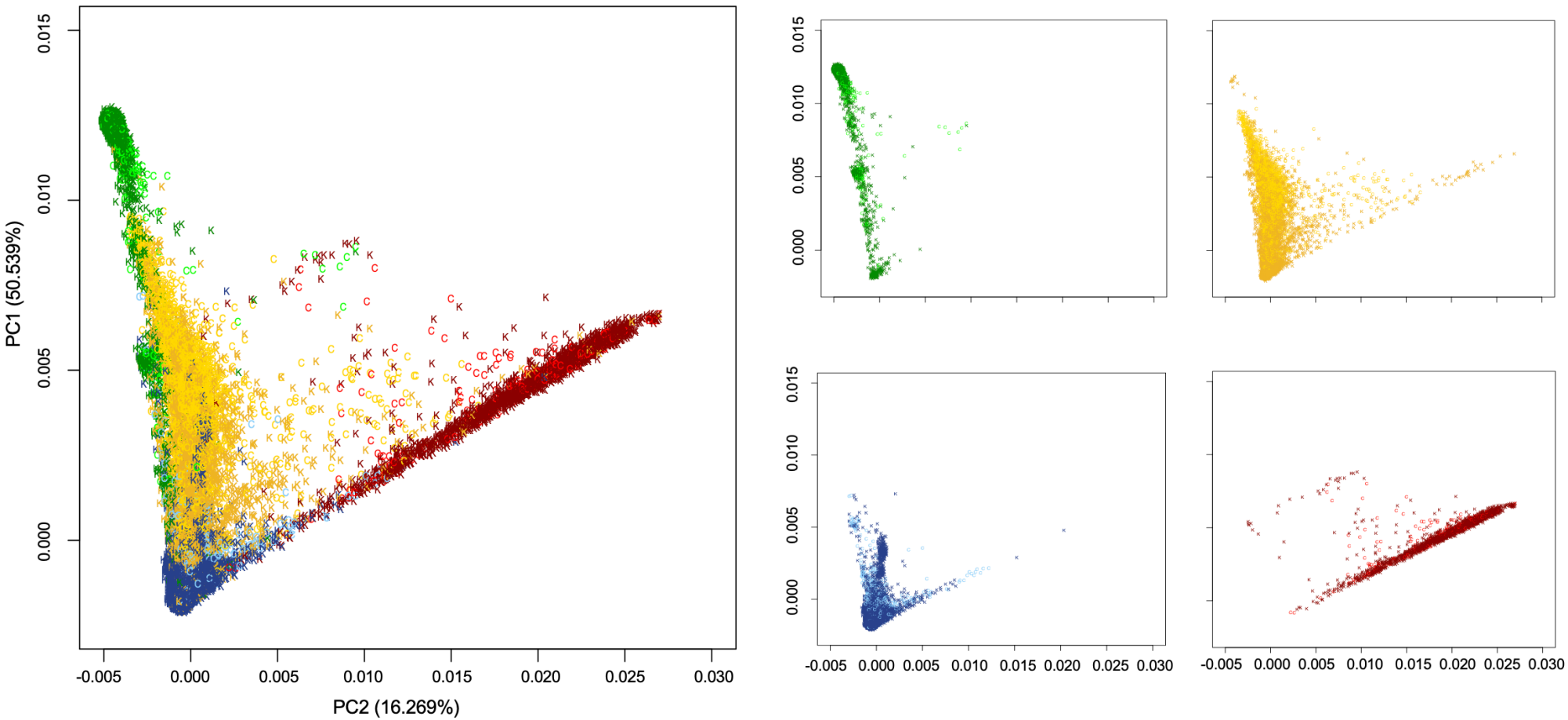


***Figure 1**: Principal component Analysis (PCA) plot of multi-ethnic study participants of CCRLP-Kaiser.* Individuals in red, yellow, green, and blue are self-reported Non-Latino Black, Hispanic/Latino, Non-Latino White, and Non-Latino East Asian individuals, respectively. Each individual from CCRLP or Kaiser/GERA is labeled with C or K, respectively.

318 We first investigated if the self-reported strata are consistent with the pattern from genetic data.
319 We applied a machine-learning based algorithm, Harmonization Ancestry and Race/Ethnicity
320 (HARE), which leverages the genetic similarities between individuals as represented in multi-
321 dimensional principal components space to identify discordances between self-labelled strata
322 with inferred genetic similarities or to impute missing self-labels[42]. Since self-reported ethnicities
323 are available for all individuals in CCRLP/Kaiser-GERA, we first evaluated the proportion of
324 individuals where the HARE algorithm identified strong discordance between the self-report and
325 the genetically inferred ancestry based on the top 30 PCs; these discordant individuals could be
326 considered for exclusion in downstream analysis. Overall, we found very low proportions of
327 samples with clear discordant labels (percent discordance = 0.47% overall, and 0.68%, 3.37%,
328 0.18%, and 0.54% for AFR, EAS, EUR, and LAT study participants, respectively (**Supplementary**
329 **Table 2**).
330
331 We also evaluated the discordance between self-reported strata and genetically inferred
332 ancestries in the HARE algorithm under its imputation mode by randomly dividing the entire
333 dataset into ten equal subsets and masking one subset at a time. In this mode, the recall
334 percentages were 82.6%, 86.9%, 77.5%, and 61.5% for AFR, EAS, NLW, and LAT, respectively.
335 In particular, among the ~38.5% of LAT individuals where the self-reported label was not recalled,
336 63.5% were considered missing because the algorithm had no clear call, while the remaining
337 were categorized into the other labels (0.80%, 0.42%, and 35.3% to AFR, EAS, and NLW,
338 respectively) and an overall distribution of estimated ancestries biased towards European-like
339 ancestry (**Supplementary Table 3; Supplementary Figure 4**). Trusting that any mis-labels (e.g.
340 transcription errors) of self-reports must be rare, and the extremely low discordance when all self-
341 reported labels are used (0.54% for LAT), we elected to proceed with our analysis using all self-
342 reported LAT individuals. Individual samples would only be removed from analysis due to
343 technical concerns identified during the quality control process, rather than excluded as, for

344 instance, an ancestry outlier based on a rather vague and undefinable criteria. Nevertheless, the
345 lower recall of LAT individuals when self-reported labels are unavailable highlights the
346 heterogeneity of LAT that exist. We thus also conducted the admixture mapping analysis using
347 the subset of ~61.5% LAT individuals based on genetically inferred ancestry to validate our
348 results; we found qualitatively concordant results (see below).
349
350 *The association between Indigenous American-like ancestries, ALL risk, and SES indicators*
351 To quantify the effects of IA-like ancestry on ALL risk, we tested the association between ALL
352 case-control status and global IA-like ancestry, while adjusting for African-like ancestry, sex, as
353 well as covariates that are proxies for socioeconomic status (SES), such as highest parental
354 education, and annual household income. We performed these assessments for all self-reported
355 LAT individuals in both CCRLP (excluding Kaiser-GERA controls; N = 4033) and CCLS (N = 605),
356 as both of these cohorts contained internal controls and are less likely to be confounded due to
357 their study design. In CCLS, IA-like ancestry was significantly associated with ALL risk, prior to
358 including SES indicators in the model, as previously reported (Walsh et al., 2013). In the base
359 model adjusting for sex and African-like ancestry, a 20% increase in IA-like ancestry was
360 associated with a 23% (OR: 1.23, 95% CI: 1.07-1.42; P = 0.003) increase in odds of developing
361 ALL. However, this association appears confounded, as inclusion of SES covariates strongly
362 attenuated the strength of association (OR: 1.12; 95% CI: 0.96-1.31) and case/control status is
363 no longer significantly associated with IA-like ancestry ($P = 0.159$; **Supplementary Table 4**). In
364 CCRLP, IA-like ancestry was not significantly associated with ALL risk in the base model adjusting
365 for only African-like ancestry and sex, and did not qualitatively change when SES covariates were
366 included (**Supplementary Table 4**).
367
368 To further support that SES indicators (education level and income) were confounders in the
369 association between ancestry and disease risk, we first examined associations between SES

370 indicators and disease risk. In CCLS, where we observed an attenuation of association after
371 adjusting for SES, higher income was associated with lower risk of disease, with those belonging
372 to higher income strata having with 0.27 to 0.76 times the odds of developing ALL (p-value range:
373 2.13×10^{-5} – 0.153; Supplementary **Table 5**). Parental education was not associated with case-
374 control status, though showing the expected trend that higher education is associated with lower
375 risk of disease (**Supplementary Tables 6**). In CCRLP, neither income nor parental education
376 were significantly associated with ALL case-control status (**Supplementary Tables 5-6**).
377
378 Next, we evaluated the relationship between ancestry and SES variables. Ancestry are often
379 correlated with SES indicators because they are correlates of self-reported race-ethnicity[28,54].
380 Self-reported race/ethnicity can act as a proxy for various socio-cultural factors, including access
381 to opportunities, healthcare, or education, which suggests that associations between ancestry
382 and variables such as income or educational attainment reflect environmental effects rather than
383 those arising from genetics[42,55,56]. Consequently, in previous studies, adjustment for SES
384 variables often attenuates the association between global ancestry and disease[28,57,58] . Consistent
385 with the literature, in both CCLS and CCRLP, we observed a positive correlation between
386 European-like ancestry and high socio-economic status (**Supplementary Tables 7-8**). Taken
387 together, because SES measures can correlate with both ancestry and disease risk, it likely
388 confounded the previously observed association[27] between IA-like ancestry and ALL risk in CCLS.
389
390 *Known ALL risk loci are enriched among Indigenous American-like ancestries*
391 We further examined the pattern of local IA-like ancestry at known ALL risk loci (**Supplementary**
392 **Table 9**). First, we tabulated the proportion of times the risk allele at known ALL loci is associated
393 with the African-like, IA-like, or European-like ancestry haplotypes, and compared these to the
394 proportions found across 1000 sets of SNPs randomly selected across the genome and matched
395 by the risk allele frequencies. Across a total of 20 SNPs from 20 loci, the risk alleles were found

396 on the African-like, IA-like, and European-like ancestry haplotypes 5.76%, 45.35%, and 48.88%
397 of the time, respectively, in CCRLP. This significantly deviates from the genome-wide expectation
398 (4.83%, 43.33%, and 51.85% for African-like, IA-like, and European-like ancestry haplotypes; p-
399 value = 1.042×10^{-36} by a chi-square goodness-of-fit test), with IA-like haplotypes showing a
400 significant enrichment in association with the risk alleles. We also replicated this observation in
401 CCLS, with IA-like haplotypes enriched with ALL risk alleles (p-value = 1.629×10^{-9}). Consistent
402 with this finding, the IA-like haplotypes have 1.32-1.33 (95% CI: 1.29 - 1.36 in CCRLP, 1.26-1.41
403 in CCLS) times the odds of carrying the risk allele compared to Non-IA-like haplotypes (p =
404 5.61×10^{-88} and 2.61×10^{-23} in CCRLP and CCLS, respectively; **Supplementary Table 10**).
405
406 If risk alleles for ALL are found disproportionately on IA-like haplotypes at known loci, we would
407 expect that IA-like ancestries show an enrichment of association in admixture mapping studies.
408 As expected, when we regressed ALL case-control status on the number of copies of IA-like
409 ancestry segments locally at each known ALL locus, while adjusting for sex and global African
410 and European ancestry proportions, we found a positive association with ALL at 14 out of 20
411 known B-cell ALL loci. At all loci in which IA-like ancestry was significantly associated with ALL at
412 the 0.05 level, IA ancestry was positively associated with risk (Binomial test, p = 0.008 **Table 2**).
413 Taken together, our results suggest that IA-like haplotypes are enriched for association with
414 increased ALL risk at known loci.
415
416
417
418
419
420
421
422
423
424
425
426

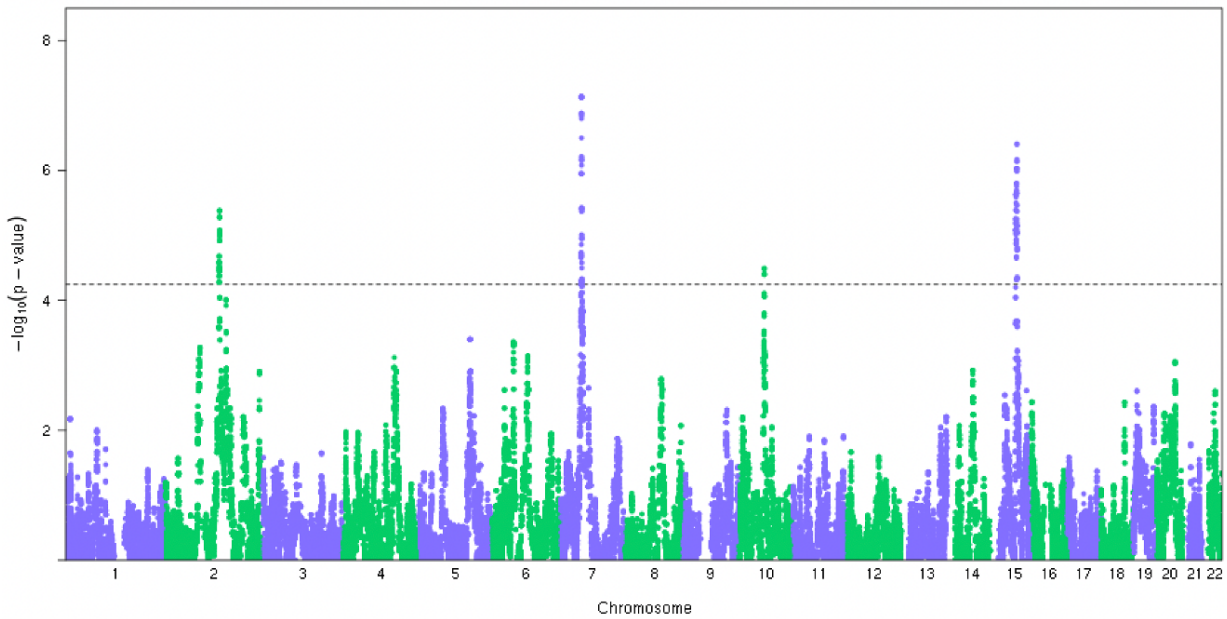
427
428
429
430

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

431
432 ***Table 2: Indigenous American-like ancestry is positively associated with risk at known ALL***
433 ***loci.*** *ALL case-control status was regressed on local Indigenous ancestry dosage, African-like*
434 *and European-like global ancestries, and sex[a]. Loci that reach nominal significance ($P < 0.05$) are*
435 *in **bold**)*
436
437
438
439

440 In addition to ALL risk alleles being found more prevalently on IA-like haplotypes, risk alleles may
441 also exert larger (or smaller) effects on ALL when they are found on the IA-like background. We
442 tested for effect size heterogeneity by including a SNPxAncestry interaction term in the standard
443 association testing model. After Bonferroni correction, no significant interactions between
444 genotype and local IA-like ancestry were observed in either CCRLP or CCLS, or when the two
445 cohorts were combined (**Supplementary Table 11**). We additionally found no evidence of
446 interaction between global IA ancestry and genotype (**Supplementary Table 12**). Therefore, we
447 found no evidence that ALL risk alleles may exert larger effects on an IA haplotype than on other
448 haplotypic backgrounds.
449
450 *Admixture mapping proposes putative novel ALL loci associated with Indigenous American-like*
451 *ancestry.*
452 Although our global ancestry analysis displayed minimal association between IA-like ancestry
453 and the risk of ALL, it is possible that this result was confounded by unascertained non-genetic
454 factors. Further, at known loci there is an enrichment of association between IA-like ancestry
455 and ALL risk, including strong associations at 7p12.2 and 10q21 that would have been genome-
456 wide significant in admixture mapping (**Table 2** above); this suggests that admixture mapping
457 may discover additional ALL risk loci. We thus performed admixture mapping using the CCRLP
458 and Kaiser/GERA Hispanic/Latino cohort as our discovery panel (1930 cases and 6417
459 controls). Our discovery admixture mapping analyses identified genome-wide significant
460 (threshold = 5.7×10^{-5} ; see **Methods**) signals at 2q21.2, 7p12.2, 10q21.2, and 15q22.31. The
461 loci on chromosomes 7 (*IKZF1*) and 10 (*ARID5B*) are known, while those on chromosomes 2
462 and 15 are potentially novel (**Figure 2**). Notably, when we restricted our analysis to a set of
463 Hispanic/Latino individuals inferred by HARE across all 10 cross-validation**,** which produced a
464 panel of 1729 cases and 4696 controls, we observed high correlation in odds ratios (Pearson's
465 correlation coefficient, r = 0. 913) and log *P*-values (r = 0.853), with a modest decrease in

466 significance of association among variants with *P*-values < 1e-3, including the four putative loci
467 (**Supplementary Table 13; Supplementary Figures 5-6**), likely reflecting the smaller sample
468 size.
469



470
471 ***Figure 2: Manhattan plot of the admixture mapping analysis.*** *The analysis identified signals*
472 *at 2q21.2, 7p12.2, 10q21.2, and 15q22.31. The significance threshold is 5.7e-5.*
473

474 We then fine-mapped each of the four loci identified by admixture mapping via single-variant
475 regional association testing (**Methods**), which uncovered the putative lead SNPs in each region
476 to be 2:133605881 (p = 4.21 e-5), 7:50395237 (p = 2.26 e-12), 10:61959980 (p = 2.25e-44),
477 and 15:66111790 (p = 3.01 e-4; **Table 3**). SNPs on chr2, 7, and 10 surpassed the regional
478 significance threshold after accounting the number of tests performed in each region via
479 permutation (**Methods**), and chr15 SNP was barely below the threshold. These SNPs exhibit
480 allele frequency differences by ancestry group (**Table 3**), and conditioning on these SNPs also
481 strongly attenuated the admixture mapping association (except for chr 15; **Table 3)**. We sought
482 replication of these associations in additional Hispanic/Latino cohorts, including cases from

483 CCLS, COG, and REDIAL (see **Methods**), totaling 1951 cases and 3,918 controls. In the meta-
484 analyzed replication cohorts, only the lead SNPs or other SNPs in LD with the lead SNPs on
485 chromosomes 7 and 10 were statistically significant (**Supplementary Table 14-15**).
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501

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}, xs1.905x10^{-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).*

519 **Discussion:**

520 Motivated by the increased risk of ALL in Hispanic/Latino children and the potential role of
521 ancestry, we carried out a systematic investigation into the impact of IA-like ancestry on childhood
522 ALL risk, both at the local and global level. Globally, we found no evidence that IA-like ancestry
523 is statistically associated with ALL risk. The lack of global ancestry association, however, did not
524 preclude an association at the local level, as IA-like ancestry was strongly associated with ALL
525 risk across known loci and risk alleles segregated with the IA haplotype more often than expected.
526 Together, these results suggest that, in addition to environmental factors, risk alleles present in
527 higher frequencies in IA populations contribute to Hispanics/Latinos' excess risk of B-cell ALL.
528 Further, this work seeks to inform best practices for performing analyses in admixed populations,
529 including considerations for stratifying by self-reported or genetically inferred ancestry and
530 estimating the effect of genetic ancestry on disease risk, in the presence of confounding by socio-
531 cultural variables.

532

533 Within the genetic epidemiology literature, careful consideration has been given to the use of self-
534 reported race/ethnicity or genetically inferred ancestry as strata in performing GWAS. For
535 instance, it is common to define "White British" based on demographic data complemented by
536 principal components of the genotype data when analyzing UK Biobank to remove concerns of
537 stratification[59–61]. However, this notion of defining a homogenous subset of the population does
538 not completely remove the effect of stratification even in a relatively homogenous cohort such as
539 the UKB[62] , and becomes even more difficult when performing analysis in multi-ancestry cohorts
540 or admixed populations. In those cases, one can pool samples across all ancestries for a "mega-
541 analysis"[63] , or perform stratified analysis using either (1) self-reported demographic labels[19,33,64]
542 or (2) some arbitrary heuristics based on genetically inferred ancestry[20,24,65,66]. The proposed
543 downside of using genetically inferred ancestry is the prescription of arbitrary thresholds to define
544 ancestry strata. Across studies, the inclusion criteria vary, with individuals of self-reported White,

545 Black and Asian race/ethnicity being defined as study participants with over 75-95% European-
546 like, 50-90% African-like, and 75-90% Asian-like ancestries, respectively[20,24,65–68] .
547
548 Previous genetic studies involving Hispanic/Latino participants have defined their strata as
549 individuals whose IA ancestry was greater or equal to 10% *and* greater than the proportion of
550 African ancestry[66,69,70]. Given the heterogeneity of Hispanics/Latinos as a population, this method
551 of categorization may prove untenable. The admixture proportions of Hispanic/Latino individuals
552 vary by country of origin with Caribbean and mainland Hispanics/Latinos exhibiting the highest
553 levels of African-like and IA-like ancestry, respectively[71,72]. Moreover, the estimated admixture
554 proportions can vary depending on the reference samples available, as well as the geographic
555 and temporal definitions of an ancestry component one wishes to estimate. Given these nuances,
556 we have tended to favor relying on self-reported demographic labels as strata when performing
557 multi-ancestry GWAS[33,73] or when limiting our analyses to Hispanics/Latinos[13]. Indeed, even if
558 overt categorizations impact little of the statistical power of the study design and may provide a
559 bit more robustness to the study, the underlying message conveyed by explicit categorization of
560 people based on estimated genetic ancestry may bring more harm than benefits.
561
562 In the present study, we evaluated the impact of our choice of relying on self-reported
563 demographic labels. When considering the multi-ancestry cohorts of CCRLP/Kaiser-GERA
564 together, the HARE analysis revealed low misclassification of individuals (0.18-3.37% across four
565 ethnic groups), suggesting that the algorithm rarely has enough evidence to suggest that a
566 person's self-reported label is inconsistent with their genetically inferred ancestry. The
567 discordance between HARE-imputed labels and the self-reported label is greater when we
568 masked the self-reported labels in a subset of the individuals. The most frequent discordance was
569 found among Hispanics/Latinos, reflecting their genetic heterogeneity across principal component
570 spaces. This is consistent with Hispanics/Latinos having the most complicated admixture profile

571 and history in our cohort, and with the notion that it is challenging to genetically define and clearly
572 delineate the Hispanic/Latino population based on genetics[63] . When we conducted our admixture
573 mapping study using this genetically inferred subset of individuals (N = 6425), we found highly
574 concordant results (r = 0.913 for effect sizes; **Supplementary Figures 5-6**). This analysis did not
575 identify additional signals; instead, the analysis was underpowered, with only the loci on
576 chromosomes 7 and 15 surpassing the genome-wide significance threshold. Given these results,
577 we elected to continue using self-reported demographic labels for inclusion of our studies
578 involving Hispanics/Latinos. We continue to use rigorous statistical approaches to control for
579 possibility of population stratification in association studies, as well as detailed characterization
580 and scrutiny of any potential findings.
581
582 Our in-depth investigation into inferred genetic ancestry also adds to and clarifies previous
583 findings. A previous study using CCLS data found a positive, statistically significant association
584 between global IA-like ancestry and ALL risk[27] . This result may have been affected by selection
585 bias, due to the lower participation of CCLS controls than cases in lower SES areas[74]. In our re-
586 analysis of the CCLS dataset, we indeed found that the association of ALL risk with global IA-like
587 ancestry was attenuated and no longer statistically significant (though trending in the positive
588 direction) after adjusting for income and education level. In the CCRLP, in which controls were
589 ascertained through statewide birth records[33] and thus should not be impacted by selection bias,
590 we found no significant association between genetic ancestry and ALL risk. Associations between
591 global ancestries and disease risks are often reported, but these associations can be confounded
592 by unaccounted environmental factors. In Hispanics/Latinos, African-like or IA-like ancestries
593 have been shown to be inversely correlated with education levels and overall socioeconomic
594 status [28,57,75,76] and accounting for these abolished the previously statistically significant
595 association between ancestry and type-2 diabetes risk[28]. In other studies of African Americans,
596 the associations between global ancestry and diseases, such as hypertension and lung cancer,

597 also dissolved upon inclusion of socio-economic status in the model[57,77]. Therefore, a more robust
598 approach to investigate the relationship between global ancestry and disease risk would be to
599 include non-genetic factors or employ a family-based design[58,78].
600
601 Local ancestry is unlikely to be correlated with environmental factors, allowing for robust
602 estimation of ancestry-disease associations. We found that averaged across all known ALL loci
603 the risk alleles segregated with the IA-like haplotype more often than expected by chance (**Table**
604 **2**, **Supplemental Table 10**). On the other hand, we found no interaction between IA-like ancestry
605 and the risk genotype at either the local or global levels (**Supplemental Tables 11-12**),
606 suggesting that there is little evidence the risk alleles exert different effects in different ancestry
607 context. Our admixture mapping identified associations on chromosomes 2, 7, 10, and 15, but
608 SNPs at the putatively novel loci on chromosomes 2 and 15 did not replicate in additional
609 Hispanic/Latino cohorts. The chromosome 2 locus (2q21.2) is a relative gene desert but overlaps
610 *MGAT5*, which has been reported to be involved in cancer metastasis and tumor suppression in
611 mice[79] (Granovsky et al., 2000). Further, transcriptional regulation of the gene is driven by the
612 Ras-Raf-Ets pathway, which is oncogenic[80–82]. On chromosome 15, the gene closest to the
613 putative risk SNP is *RAB11A*, a member of the RAS oncogene family, and has been previously
614 associated with risk of lung squamous cell carcinoma[83] and ovarian cancer[84] . Further, this region
615 contains multiple SNP associations with blood cell phenotypes[85,86] , of relevance for the etiology
616 of childhood ALL[21] . Adjusting for the lead SNP did not strongly attenuate the chromosome 15
617 association identified by admixture mapping, suggesting that more complex genomic variation
618 such as structural variants, which may be in low LD with imputed SNPs[87], could underlie this
619 association and would require further validation. While there may be biological plausibility for
620 these potentially novel loci, the lack of replication could also be due to lack of statistical power,
621 heterogeneity of Hispanic/Latino populations across the United States, or that our initial findings
622 were statistical artefacts.

Finally, the results of our study have implications on future investigations of the genetic architecture of ALL. As reflected in our findings, the elevated ALL risk in Hispanic/Latino children is to some extent driven by the higher frequencies of known ALL risk alleles on IA-like haplotypes, which may underlie the shift towards higher mean polygenic risk scores for ALL in Hispanics/Latinos than in non-Latino Whites[73,88]. Moreover, in line with risk alleles being found consistently enriched on IA-like haplotypes, IA-like ancestry at known loci is also more likely to be associated with increasing risk. These consistent directional associations with risk could be indicative of a signature of past adaptation. There is some evidence that ALL risk loci, such as the *IKZF1* locus that we found to be genome-wide significant in admixture mapping, may be under positive selection in Hispanic/Latino populations[13]. Given the overall enrichment of risk alleles on IA haplotypes, the selection may be more polygenic. Careful scrutiny is needed to rule out other potential confounding[59,89] though polygenic adaptation has been demonstrated in other traits[90,91]. Understanding this architecture, as well as continued discovery of ALL risk alleles that may segregate with the Indigenous American haplotype or particular ancestries broadly, can aid in precision medicine efforts and improve diagnostics, cancer screening, risk assessment, and awareness, with the ultimate end of reducing disparities[92].

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/).

674 **Data Availability Statement**

675 Genetic data from CCRLP and CCLS were obtained from the California Biobank. Because
676 samples and data are property of the State of California, researchers who would like to request
677 to use the data should establish a protocol with the California Department of Public Health
678 Committee for the Protection of Human Subjects and other relevant agencies to understand if
679 approval to share data are acceptable. The State has provided the following guidance on data
680 sharing, as noted in the following statement: "California has determined that researchers
681 requesting the use of California Biobank biospecimens for their studies will need to seek an
682 exemption from NIH or other granting or funder requirements regarding the uploading of study
683 results into an external bank or repository (including into the NIH dbGaP or other bank or
684 repository). This applies to any uploading of genomic data and/or sharing of these biospecimens
685 or individual data derived from these biospecimens. Such activities have been determined to
686 violate the statutory scheme at California Health and Safety Code Section 124980 (j), 124991
687 (b), (g), (h) and 103850 (a) and (d), which protect the confidential nature of biospecimens and
688 individual data derived from biospecimens. Investigators may agree to share aggregate data on
689 SNP frequency and their associated p values with other investigators and may upload such
690 frequencies into repositories including the NIH dbGaP repository providing: a) the denominator
691 from which the data is derived includes no fewer than 20,000 individuals; b) no cell count is for
692 <5 individuals; and c) no correlations or linkage probabilities between SNPs are provided." Our
693 data, from CCRLP and CCLS, are each comprised of less than 20,000 subjects and therefore
694 cannot be uploaded to a repository such as dbGaP. Data from Children's Oncology Group and
695 REDIAL were obtained by special request (Yang et al., 2011; Harris et al., 2023). Data from
696 MESA (phs000209.v13.p3) and HCHS/SOL (phs000810.v2.p2.c1, phs000810.v2.p2.c2) were
697 obtained from dbGaP.

698

References

1. Bhojwani, D., Yang, J. J. & Pui, C.-H. Biology of Childhood Acute Lymphoblastic Leukemia. *Pediatr. Clin. North Am.* **62**, 47–60 (2015).
2. Rheingold, S. R. *et al.* Determinants of survival after first relapse of acute lymphoblastic leukemia: a Children's Oncology Group study. *Leukemia* **38**, 2382–2394 (2024).
3. Einsiedel, H. G. *et al.* Long-term outcome in children with relapsed ALL by risk-stratified salvage therapy: results of trial acute lymphoblastic leukemia-relapse study of the Berlin-Frankfurt-Münster Group 87. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **23**, 7942–7950 (2005).
4. Nguyen, K. *et al.* Factors Influencing Survival After Relapse From Acute Lymphoblastic Leukemia: A Children's Oncology Group Study. *Leuk. Off. J. Leuk. Soc. Am. Leuk. Res. Fund UK* **22**, 2142–2150 (2008).
5. Hunger, S. P. & Raetz, E. A. How I treat relapsed acute lymphoblastic leukemia in the pediatric population. *Blood* **136**, 1803–1812 (2020).
6. Devilli, L. *et al.* Long-term and quality of survival in patients treated for acute lymphoblastic leukemia during the pediatric age. *Hematol. Rep.* **13**, 8847 (2021).
7. Pui, C.-H., Robison, L. L. & Look, A. T. Acute lymphoblastic leukaemia. *The Lancet* **371**, 1030–1043 (2008).
8. Sasaki, K. *et al.* Acute lymphoblastic leukemia: A population-based study of outcome in the United States based on the surveillance, epidemiology, and end results (SEER) database, 1980–2017. *Am. J. Hematol.* **96**, 650–658 (2021).
9. American Cancer Society. Key Statistics for Acute Lymphocytic Leukemia (ALL). *American Cancer Society* https://www.cancer.org/cancer/types/acute-lymphocytic-leukemia/about/key-statistics.html (2024).

725 10. Feng, Q. *et al.* Trends in Acute Lymphoblastic Leukemia Incidence in the United States
726 by Race/Ethnicity From 2000 to 2016. *Am. J. Epidemiol.* **190**, 519–527 (2021).
727 11. Quiroz, E. *et al.* Immunophenotype of acute lymphoblastic leukemia in minorities-
728 analysis from the SEER database. *Hematol. Oncol.* **40**, 105–110 (2022).
729 12. de Smith, A. J., Jiménez-Morales, S. & Mejía-Aranguré, J. M. The genetic risk of acute
730 lymphoblastic leukemia and its implications for children of Latin American origin. *Front.*
731 *Oncol.* **13**, 1299355 (2024).
732 13. de Smith, A. J. *et al.* A noncoding regulatory variant in IKZF1 increases acute
733 lymphoblastic leukemia risk in Hispanic/Latino children. *Cell Genomics* **4**, (2024).
734 14. Thomas, D. C. *Statistical Methods in Genetic Epidemiology*. (Oxford University Press,
735 Oxford ; New York, 2004).
736 15. Swords, R. *et al.* Acute leukemia in adult Hispanic Americans: a large-population study.
737 *Blood Cancer J.* **6**, e484 (2016).
738 16. Pollyea, D. A. *et al.* Acute Leukemia in Hispanic Americans: Incidence and Incidence
739 Rate Differences by Nativity. *Blood* **120**, 3160 (2012).
740 17. Dong, Y. *et al.* Leukemia incidence trends at the global, regional, and national level
741 between 1990 and 2017. *Exp. Hematol. Oncol.* **9**, 14 (2020).
742 18. Yi, M., Zhou, L., Li, A., Luo, S. & Wu, K. Global burden and trend of acute lymphoblastic
743 leukemia from 1990 to 2017. *Aging* **12**, 22869–22891 (2020).
744 19. Jeon, S., Smith, A. J. & Li, S. Genome-wide trans-ethnic meta-analysis identifies novel
745 susceptibility loci for childhood acute lymphoblastic leukemia. *Leukemia* **36**, 865–868 (2022).
746 20. Vijayakrishnan, J. *et al.* Identification of four novel associations for B-cell acute
747 lymphoblastic leukaemia risk. *Nat. Commun.* **10**, 5348 (2019).
748 21. Kachuri, L. *et al.* Genetic determinants of blood-cell traits influence susceptibility to
749 childhood acute lymphoblastic leukemia. *Am. J. Hum. Genet.* **108**, 1823–1835 (2021).

750 22. Qian, M. *et al.* Novel susceptibility variants at the ERG locus for childhood acute
751 lymphoblastic leukemia in Hispanics. *Blood* **133**, 724–729 (2019).
752 23. Xu, H. *et al.* ARID5B genetic polymorphisms contribute to racial disparities in the
753 incidence and treatment outcome of childhood acute lymphoblastic leukemia. *J. Clin. Oncol.*
754 *Off. J. Am. Soc. Clin. Oncol.* **30**, 751–757 (2012).
755 24. Perez-Andreu, V. *et al.* Inherited GATA3 variants are associated with Ph-like childhood
756 acute lymphoblastic leukemia and risk of relapse. *Nat. Genet.* **45**, 1494–1498 (2013).
757 25. Bryc, K., Durand, E. Y., Macpherson, J. M., Reich, D. & Mountain, J. L. The genetic
758 ancestry of african americans, latinos, and european Americans across the United States.
759 *Am. J. Hum. Genet.* **96**, 37–53 (2015).
760 26. Homburger, J. R. *et al.* Genomic Insights into the Ancestry and Demographic History of
761 South America. *PLOS Genet.* **11**, e1005602 (2015).
762 27. Walsh, K. *et al.* Associations between genome-wide Native American ancestry, known
763 risk alleles and B-cell ALL risk in Hispanic children. *Leukemia* **27**, 10.1038/leu.2013.130
764 (2013).
765 28. Florez, J. C. *et al.* Strong association of socioeconomic status with genetic ancestry in
766 Latinos: implications for admixture studies of type 2 diabetes. *Diabetologia* **52**, 1528–1536
767 (2009).
768 29. Thornton, T. A. & Bermejo, J. L. Local and Global Ancestry Inference, and Applications
769 to Genetic Association Analysis for Admixed Populations. *Genet. Epidemiol.* **38**, S5–S12
770 (2014).
771 30. de Smith, A. J. *et al.* Heritable variation at the chromosome 21 gene ERG is associated
772 with acute lymphoblastic leukemia risk in children with and without Down syndrome.
773 *Leukemia* **33**, 2746–2751 (2019).
774 31. Committee on the Use of Race, Ethnicity, and Ancestry as Population Descriptors in
775 Genomics Research *et al. Using Population Descriptors in Genetics and Genomics*

776 *Research: A New Framework for an Evolving Field*. 26902 (National Academies Press,
777 Washington, D.C., 2023). doi:10.17226/26902.
778 32. Zavala, V. A. *et al.* Cancer health disparities in racial/ethnic minorities in the United
779 States. *Br. J. Cancer* **124**, 315–332 (2021).
780 33. Wiemels, J. L. *et al.* GWAS in childhood acute lymphoblastic leukemia reveals novel
781 genetic associations at chromosomes 17q12 and 8q24.21. *Nat. Commun.* **9**, 286 (2018).
782 34. Ma, X., Buffler, P. A., Layefsky, M., Does, M. B. & Reynolds, P. Control Selection
783 Strategies in Case-Control Studies of Childhood Diseases. *Am. J. Epidemiol.* **159**, 915–921
784 (2004).
785 35. Harris, R. D. *et al.* Ethnic-Specific Predictors of Neurotoxicity Among Pediatric Acute
786 Lymphoblastic Leukemia Patients Following High-Dose Methotrexate. *Cancer* **129**, 1287–
787 1294 (2023).
788 36. Sheng, X. *et al.* Inverted genomic regions between reference genome builds in humans
789 impact imputation accuracy and decrease the power of association testing. *Hum. Genet.*
790 *Genomics Adv.* **4**, (2023).
791 37. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative
792 modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**,
793 278–288 (2013).
794 38. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in
795 141,456 humans. *Nature* **581**, 434–443 (2020).
796 39. Li, S. *et al.* Localized variation in ancestral admixture identifies pilocytic astrocytoma risk
797 loci among Latino children. *PLoS Genet.* **18**, e1010388 (2022).
798 40. Kowalski, M. H. *et al.* Use of 100,000 NHLBI Trans-Omics for Precision Medicine
799 (TOPMed) Consortium whole genome sequences improves imputation quality and detection
800 of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS Genet.*
801 **15**, 1008500 (2019).

802 41. Cahoon, J. L. *et al.* Imputation accuracy across global human populations. *Am. J. Hum.*
803 *Genet.* **111**, 979–989 (2024).
804 42. Fang, H. *et al.* Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in
805 Genome-wide Association Studies. *Am. J. Hum. Genet.* **105**, 763–772 (2019).
806 43. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer
807 datasets. *GigaScience* **4**, 7 (2015).
808 44. Metayer, C. *et al.* Tobacco Smoke Exposure and the Risk of Childhood Acute
809 Lymphoblastic and Myeloid Leukemias by Cytogenetic Subtype. *Cancer Epidemiol.*
810 *Biomarkers Prev.* **22**, 1600–1611 (2013).
811 45. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation
812 for Statistical Computing (2023).
813 46. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-
814 Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
815 47. Qian, M. *et al.* Genome-Wide Association Study of Susceptibility Loci for T-Cell Acute
816 Lymphoblastic Leukemia in Children. *J. Natl. Cancer Inst.* **111**, 1350–1357 (2019).
817 48. Brown, L. A. *et al.* Admixture Mapping Identifies an Amerindian Ancestry Locus
818 Associated with Albuminuria in Hispanics in the United States. *J. Am. Soc. Nephrol. JASN*
819 **28**, 2211–2220 (2017).
820 49. Schick, U. M. *et al.* Genome-wide Association Study of Platelet Count Identifies
821 Ancestry-Specific Loci in Hispanic/Latino Americans. *Am. J. Hum. Genet.* **98**, 229–242
822 (2016).
823 50. Qin, H. & Zhu, X. Power Comparison of Admixture Mapping and Direct Association
824 Analysis in Genome-Wide Association Studies. *Genet. Epidemiol.* **36**, 235–243 (2012).
825 51. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of
826 genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).

827 52. Peterson, R. E. *et al.* Genome-wide association studies in ancestrally diverse
828 populations: opportunities, methods, pitfalls, and recommendations. *Cell* **179**, 589–603
829 (2019).
830 53. Reich, D. *et al.* Reconstructing Native American population history. *Nature* **488**, 370–374
831 (2012).
832 54. Iyer, H. S., Gomez, S. L., Cheng, I. & Rebbeck, T. R. Relative impact of genetic ancestry
833 and neighborhood socioeconomic status on all-cause mortality in self-identified African
834 Americans. *PLOS ONE* **17**, e0273735 (2022).
835 55. Institute of Medicine (US) Committee on Understanding and Eliminating Racial and
836 Ethnic Disparities in Health Care. *Unequal Treatment: Confronting Racial and Ethnic*
837 *Disparities in Health Care*. (National Academies Press (US), Washington (DC), 2003).
838 56. Burchard, E. G. *et al.* The Importance of Race and Ethnic Background in Biomedical
839 Research and Clinical Practice. *N. Engl. J. Med.* **348**, 1170–1175 (2003).
840 57. Aldrich, M. C. *et al.* Socioeconomic Status and Lung Cancer: Unraveling the Contribution
841 of Genetic Admixture. *Am. J. Public Health* **103**, e73–e80 (2013).
842 58. Sun, H. *et al.* The impact of global and local Polynesian genetic ancestry on complex
843 traits in Native Hawaiians. *PLOS Genet.* **17**, e1009273 (2021).
844 59. Sohail, M. *et al.* Polygenic adaptation on height is overestimated due to uncorrected
845 stratification in genome-wide association studies. *eLife* **8**, e39702 (2019).
846 60. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data.
847 *Nature* **562**, 203–209 (2018).
848 61. Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to
849 Common Complex Disease. *Cell* **167**, 1415-1429.e19 (2016).
850 62. Chen, M. *et al.* Evidence of Polygenic Adaptation in Sardinia at Height-Associated Loci
851 Ascertained from the Biobank Japan. *Am. J. Hum. Genet.* **107**, 60–71 (2020).

852 63. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery for
853 complex traits. *Nature* **570**, 514–518 (2019).
854 64. Kachuri, L. *et al.* Genetically adjusted PSA levels for prostate cancer screening. *Nat.*
855 *Med.* **29**, 1412–1423 (2023).
856 65. Klarin, D. *et al.* Genetics of blood lipids among ~300,000 multi-ethnic participants of the
857 Million Veteran Program. *Nat. Genet.* **50**, 1514–1523 (2018).
858 66. Xu, H. *et al.* Novel susceptibility variants at 10p12.31-12.2 for childhood acute
859 lymphoblastic leukemia in ethnically diverse populations. *J. Natl. Cancer Inst.* **105**, 733–742
860 (2013).
861 67. Gautam, Y. *et al.* Genome-wide admixture and association analysis identifies African
862 ancestry specific risk loci of eosinophilic esophagitis in African American. *J. Allergy Clin.*
863 *Immunol.* **151**, 1337–1350 (2023).
864 68. Shrine, N. *et al.* Multi-ancestry genome-wide association analyses improve resolution of
865 genes and pathways influencing lung function and chronic obstructive pulmonary disease
866 risk. *Nat. Genet.* **55**, 410–422 (2023).
867 69. Archer, N. P. *et al.* Family-based exome-wide association study of childhood acute
868 lymphoblastic leukemia among Hispanics confirms role of ARID5B in susceptibility. *PLOS*
869 *ONE* **12**, e0180488 (2017).
870 70. Karol, S. E. *et al.* Genetics of ancestry-specific risk for relapse in acute lymphoblastic
871 leukemia. *Leukemia* **31**, 1325–1332 (2017).
872 71. Conomos, M. P. *et al.* Genetic diversity and association studies in US Hispanic/Latino
873 populations: applications in the Hispanic Community Health Study/Study of Latinos. *Am. J.*
874 *Hum. Genet.* **98**, 165–184 (2016).
875 72. Bryc, K. *et al.* Genome-wide patterns of population structure and admixture among
876 Hispanic/Latino populations. *Proc. Natl. Acad. Sci.* **107**, 8954–8961 (2010).

877 73. Jeon, S. *et al.* Genome-wide trans-ethnic meta-analysis identifies novel susceptibility loci
878 for childhood acute lymphoblastic leukemia. *Leukemia* **36**, 865–868 (2022).
879 74. Wheeler, D. C., Boyle, J., Carli, M., Ward, M. H. & Metayer, C. Neighborhood
880 Deprivation, Indoor Chemical Concentrations, and Spatial Risk for Childhood Leukemia. *Int.*
881 *J. Environ. Res. Public. Health* **20**, 3582 (2023).
882 75. Zou, J. Y. *et al.* Genetic and socioeconomic study of mate choice in Latinos reveals
883 novel assortment patterns. *Proc. Natl. Acad. Sci.* **112**, 13621–13626 (2015).
884 76. da Silva, T. M. *et al.* Biogeographical ancestry is associated with socioenvironmental
885 conditions and infections in a Latin American urban population. *SSM - Popul. Health* **4**, 301–
886 306 (2018).
887 77. Non, A. L., Gravlee, C. C. & Mulligan, C. J. Education, Genetic Ancestry, and Blood
888 Pressure in African Americans and Whites. *Am. J. Public Health* **102**, 1559–1565 (2012).
889 78. Asgari, S. *et al.* A positively selected FBN1 missense variant reduces height in
890 Peruvians. *Nature* **582**, 234–239 (2020).
891 79. Granovsky, M. *et al.* Suppression of tumor growth and metastasis in Mgat5-deficient
892 mice. *Nat. Med.* **6**, 306–312 (2000).
893 80. Hirata, A. *et al.* Increased risk of metastasis in patients with incidental use of renin-
894 angiotensin system inhibitors: a retrospective analysis for multiple types of cancer based on
895 electronic medical records. *Hypertens. Res.* **45**, 1869–1881 (2022).
896 81. Kang, R. *et al.* Transcriptional Regulation of the N-Acetylglucosaminyltransferase V
897 Gene in Human Bile Duct Carcinoma Cells (HuCC-T1) Is Mediated by Ets-1 *. *J. Biol. Chem.*
898 **271**, 26706–26712 (1996).
899 82. Buckhaults, P., Chen, L., Fregien, N. & Pierce, M. Transcriptional Regulation ofN-
900 Acetylglucosaminyltransferase V by the srcOncogene *. *J. Biol. Chem.* **272**, 19575–19581
901 (1997).

902 83. Gombodorj, N. *et al.* RAB11A Expression Is Associated With Cancer Aggressiveness
903 Through Regulation of FGFR-Signaling in Lung Squamous Cell Carcinoma. *Ann. Surg.*
904 *Oncol.* **29**, 7149–7162 (2022).
905 84. Wang, Y., Ren, Y., Li, N., Zhao, J. & Zhao, S. Rab11a promotes the malignant
906 progression of ovarian cancer by inducing autophagy. *Genes Genomics* **44**, 1375–1384
907 (2022).
908 85. Vuckovic, D. *et al.* The Polygenic and Monogenic Basis of Blood Traits and Diseases.
909 *Cell* **182**, 1214-1231.e11 (2020).
910 86. Chen, M.-H. *et al.* Trans-ethnic and ancestry-specific blood-cell genetics in 746,667
911 individuals from 5 global populations. *Cell* **182**, 1198-1213.e14 (2020).
912 87. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes.
913 *Nature* **526**, 75–81 (2015).
914 88. Jeon, S. *et al.* Evaluating Genomic Polygenic Risk Scores for Childhood Acute
915 Lymphoblastic Leukemia in Latinos. *MedRxiv Prepr. Serv. Health Sci.* 2023.06.08.23291167
916 (2023) doi:10.1101/2023.06.08.23291167.
917 89. Berg, J. J. *et al.* Reduced signal for polygenic adaptation of height in UK Biobank. *eLife*
918 **8**, e39725 (2019).
919 90. Le, M. K. *et al.* 1,000 ancient genomes uncover 10,000 years of natural selection in
920 Europe. 2022.08.24.505188 Preprint at https://doi.org/10.1101/2022.08.24.505188 (2022).
921 91. Chen, M. *et al.* Evidence of Polygenic Adaptation in Sardinia at Height-Associated Loci
922 Ascertained from the Biobank Japan. *Am. J. Hum. Genet.* **107**, 60–71 (2020).
923 92. Jorde, L. B. & Bamshad, M. J. Genetic Ancestry Testing What Is It and Why Is It
924 Important? *JAMA* **323**, 1089–1090 (2020).
925