

UCSF

UC San Francisco Previously Published Works

Title

Investigating DNA methylation as a mediator of genetic risk in childhood acute lymphoblastic leukemia.

Permalink

<https://escholarship.org/uc/item/0795j42r>

Journal

Human Molecular Genetics, 31(21)

ISSN

0964-6906

Authors

Xu, Keren

Li, Shaobo

Pandey, Priyatama

et al.

Publication Date

2022-10-28

DOI

10.1093/hmg/ddac137

Peer reviewed

1  
2  
3 1 **Investigating DNA Methylation as a Mediator of Genetic Risk in**  
4  
5  
6 2 **Childhood Acute Lymphoblastic Leukemia**  
7  
8  
9 3

10  
11 4 Keren Xu<sup>1</sup>, Shaobo Li<sup>1</sup>, Priyatama Pandey<sup>1</sup>, Alice Y. Kang<sup>2</sup>, Libby M. Morimoto<sup>2</sup>, Nicholas  
12  
13 5 Mancuso<sup>1</sup>, Xiaomei Ma<sup>3</sup>, Catherine Metayer<sup>2</sup>, Joseph L. Wiemels<sup>1</sup>, Adam J. de Smith<sup>1\*</sup>  
14  
15 6

16  
17  
18 7 <sup>1</sup> Center for Genetic Epidemiology, Department of Population and Public Health Sciences,  
19  
20 8 University of Southern California, Los Angeles, CA 90033, USA

21  
22 9 <sup>2</sup> School of Public Health, University of California, Berkeley, Berkeley, CA 94704, USA

23  
24 10 <sup>3</sup> Department of Chronic Disease Epidemiology, Yale School of Public Health, New Haven, CT  
25  
26 11 06510, USA

27  
28 12 **\*Corresponding Author:** Adam J. de Smith; Address: 1450 Biggy St., NRT-1509H, USC Norris  
29  
30 13 Comprehensive Cancer Center, Los Angeles, CA 90033; Email: [adam.desmith@med.usc.edu](mailto:adam.desmith@med.usc.edu)  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

## 25 Abstract

26 Genome-wide association studies have identified a growing number of single nucleotide  
27 polymorphisms (SNPs) associated with childhood acute lymphoblastic leukemia (ALL), yet the  
28 functional roles of most SNPs are unclear. Multiple lines of evidence suggest epigenetic  
29 mechanisms may mediate the impact of heritable genetic variation on phenotypes. Here, we  
30 investigated whether DNA methylation mediates the effect of genetic risk loci for childhood ALL.  
31 We performed an epigenome-wide association study (EWAS) including 808 childhood ALL cases  
32 and 919 controls from California-based studies using neonatal blood DNA. For differentially  
33 methylated CpG positions (DMPs), we next conducted association analysis with 23 known ALL  
34 risk SNPs followed by causal mediation analyses addressing the significant SNP-DMP pairs. DNA  
35 methylation at CpG cg01139861, in the promoter region of *IKZF1*, mediated the effects of the  
36 intronic *IKZF1* risk SNP rs78396808, with the average causal mediation effect (ACME) explaining  
37 ~30% of the total effect (ACME  $P=0.0031$ ). In analyses stratified by self-reported race/ethnicity,  
38 the mediation effect was only significant in Latinos, explaining ~41% of the total effect of  
39 rs78396808 on ALL risk (ACME  $P=0.0037$ ). Conditional analyses confirmed the presence of at  
40 least three independent genetic risk loci for childhood ALL at *IKZF1*, with rs78396808 unique to  
41 non-European populations. We also demonstrated that the most significant DMP in the EWAS,  
42 CpG cg13344587 at gene *ARID5B* ( $P=8.61 \times 10^{-10}$ ), was entirely confounded by the *ARID5B* ALL  
43 risk SNP rs7090445. Our findings provide new insights into the functional pathways of ALL risk  
44 SNPs and the DNA methylation differences associated with risk of childhood ALL.

## 50 Introduction

51 Acute lymphoblastic leukemia (ALL) is characterized by the uncontrolled proliferation of immature  
52 lymphocytes in the bone marrow and is the most common childhood cancer (1). Although current  
53 treatment protocols result in an overall survival rate that exceeds 90% in childhood ALL patients  
54 in the US (2), long-term survivors experience significant adverse effects from therapy, including  
55 subsequent neoplasms, chronic health conditions, and premature mortality (3). Understanding  
56 the causes of childhood ALL, therefore, remains essential. Genome-wide association studies  
57 (GWAS) have identified single nucleotide polymorphisms (SNPs) associated with ALL risk in  
58 genes involved in hematopoiesis and B-cell development, including *ARID5B*, *IKZF1*, *CEBPE*,  
59 *GATA3*, *IKZF3*, *ERG*, and *BMI1* (4–9); however, most of the top associated SNPs are located in  
60 non-coding regions of the genome, leaving the mechanisms through which they contribute to ALL  
61 etiology unclear.

62 Epigenetic modifications are well-recognized drivers for oncogenesis (10). As one of the  
63 components of the epigenetic machinery, DNA methylation contributes to cancer etiology and  
64 progression through various mechanisms (11); for instance, DNA hypermethylation at gene  
65 promoters can silence tumor suppressors and other cancer-related genes (12), whereas broad  
66 regions of DNA hypomethylation are associated with genomic instability (13). Furthermore, most  
67 cancers harbor genetic abnormalities that modify DNA methylation, resulting in widespread  
68 changes in gene expression (12,14–16). Several studies have reported that DNA methylation  
69 mediates the heritable genetic impact on complex diseases, such as rheumatoid arthritis, chronic  
70 obstructive pulmonary disease, and prostate cancer (17–19). However, although it has been  
71 reported that aberrant epigenetic modifications serve pivotal roles in leukemogenesis in childhood  
72 ALL (20,21), no study to our knowledge has been conducted to explore how DNA methylation  
73 modifications may function downstream of genetic risk pathways of childhood ALL.

1  
2  
3 74 Epigenome-wide association studies (EWASs) testing DNA methylation differences  
4  
5 75 between childhood ALL cases and controls at birth may also pinpoint differentially methylated  
6  
7 76 CpG positions (DMPs) involved in the development of childhood ALL, yet studies conducted so  
8  
9 77 far have mainly focused on DNA methylation changes in diagnostic leukemia samples (22–25).  
10  
11 78 Here, we performed an EWAS of ALL including 808 childhood ALL cases and 919 controls from  
12  
13 79 two ancestrally diverse independent California-based studies, the California Childhood Leukemia  
14  
15 80 Study (CCLS) and the Childhood Cancer Records Linkage Project (CCRLP), to identify ALL-  
16  
17 81 associated DMPs, and then tested the association of significant DMPs with known ALL risk SNPs  
18  
19 82 and assessed whether DNA methylation at these CpG probes may mediate the effects of the  
20  
21 83 genetic risk loci.  
22  
23  
24  
25

84

## 85 Results

26  
27  
28  
29 86 There were 850 ALL cases and 931 cancer-free controls from the CCLS and the CCRLP that had  
30  
31 87 DNA samples from neonatal dried bloodspot (DBS) assayed on either the Illumina®  
32  
33 88 HumanMethylation450 BeadChip (450K) DNA methylation arrays or Illumina® Infinium  
34  
35 89 MethylationEPIC BeadChip (EPIC) arrays. After excluding subjects with trisomy 21 (27 cases, 1  
36  
37 90 control), we included in the EWAS a total of 808 childhood ALL cases and 919 cancer-free  
38  
39 91 controls that passed DNA methylation quality control (see **Materials and Methods**). Demographic  
40  
41 92 characteristics of these subjects are summarized in **Table 1**, and the study design is illustrated in  
42  
43 93 **Figure 1**. Over half of the study participants were males, and overall 53.7% were self-reported  
44  
45 94 Latino and 31.4% were non-Latino white, with approximately equal distributions among cases and  
46  
47 95 controls across the CCLS 450K, CCLS EPIC, and CCRLP EPIC datasets. Demographic  
48  
49 96 characteristics of the subset of 683 childhood ALL cases and 804 controls included in the  
50  
51 97 methylation quantitative trait loci (mQTL) and mediation analyses were similar to the overall  
52  
53 98 dataset (**Table S1**).  
54  
55  
56  
57  
58  
59  
60

99

### 100 **Differentially methylated positions (DMPs)**

101 To identify ALL-associated DMPs on autosomal chromosomes, we first performed EWAS  
102 analyses separately in the CCLS 450K, CCLS EPIC, and CCRLP EPIC datasets. A total of  
103 363,973 CpGs were included in the overall EWAS fixed-effect meta-analysis (CCLS 450K, CCLS  
104 EPIC, and CCRLP EPIC datasets) (**Figure 1**). An additional 340,576 CpGs not included in the  
105 CCLS 450K dataset were analyzed in the EPIC array meta-analysis (CCLS EPIC and CCRLP  
106 EPIC datasets) (**Figure 1**). The number of CpGs meeting each probe filtering criterion is  
107 summarized in **Table S2**. CpG cg13344587, in an intronic region of the ALL risk gene *ARID5B*,  
108 was significantly differentially methylated by ALL case/control status ( $P=8.61 \times 10^{-10}$ ) after  
109 adjusting for multiple testing using a stringent Bonferroni correction in the overall meta-analysis  
110 ( $P$  threshold:  $0.025/363,973=6.87 \times 10^{-8}$ ) (**Figures 2A and 2B**). No additional CpGs reached  
111 Bonferroni significance in the EPIC array analysis (**Figures 2C and 2D**), nor survived the  
112 Benjamini-Hochberg false discovery rate (FDR) correction ( $FDR < 0.05$ ) in either study. Using a  
113 less stringent threshold of  $P < 1 \times 10^{-4}$  for the purposes of identifying candidate CpGs for  
114 downstream mediation analyses (26,27), we found 47 DMPs for ALL overlapping both 450K and  
115 EPIC arrays from the overall meta-analysis and 51 DMPs from the EPIC array meta-analysis  
116 (**Figures 2A and 2C; Tables 2 and S3**). The effect estimates of these CpGs in overall participants  
117 were in strong correlations with those in self-reported Latinos and non-Latino whites in stratified  
118 EWAS (**Figure S1**) and in EWAS adjusting for genetic ancestry using principal components  
119 derived from SNP array data instead of from EPISTRUCTURE, in the subset of subjects with both  
120 DNA methylation and SNP genotype data available (**Figure S1; Tables S4 and S5**).

121

### 122 **Methylation quantitative trait loci (mQTL)**

123 The 23 childhood ALL risk SNPs identified from our recent multi-ancestry GWAS meta-analysis  
124 were included in the mQTL analysis (28) (**Table S6**). These SNPs overlap 19 genomic loci, and

1  
2  
3 125 include 4 secondary associations discovered in conditional analysis adjusting for the lead SNP at  
4  
5 126 *IKZF1*, *CDKN2A*, *CEBPE*, and *IKZF3*. They were analyzed for association with the 47 DMPs  
6  
7 127 separately in the CCLS 450K, CCLS EPIC, and CCRLP EPIC datasets, and with the 51 DMPs  
8  
9 128 separately in the two EPIC datasets (**Figure 1**). Two SNP-DMP pairs *in cis* passed the Bonferroni  
10  
11 129 corrected threshold in the meta-analysis of three datasets ( $P < 2.31 \times 10^{-5}$  [0.025/(23×47)]): 1) SNP  
12  
13 130 rs7090445 and DMP cg13344587 at gene *ARID5B*, and 2) SNP rs78396808 and DMP  
14  
15 131 cg01139861 at gene *IKZF1* (**Tables 2 and 3**). No SNP-DMP associations survived the Bonferroni  
16  
17 132 corrected threshold ( $P < 2.13 \times 10^{-5}$  [0.025/(23×51)]) or the FDR correction (FDR < 0.05) in the EPIC  
18  
19 133 array meta-analysis. The two significant SNP-DMP pairs identified from the overall meta-analysis  
20  
21 134 remained significant in mQTL analysis using DNA methylation M-values (data not shown).

22  
23  
24 135 In multivariable conditional analysis to assess the effect of the SNPs on ALL risk while  
25  
26 136 adjusting for the corresponding CpG in the SNP-DMP pair, and vice-versa, the *ARID5B* SNP  
27  
28 137 rs7090445 remained significant ( $P = 8.52 \times 10^{-5}$ ) (model 2 in **Table 4**), indicating that rs7090445 had  
29  
30 138 a direct impact on ALL risk, independent of the *ARID5B* CpG cg13344587, and that cg13344587  
31  
32 139 could potentially partially mediate the effect of rs7090445 on ALL risk; however, cg13344587 was  
33  
34 140 no longer significant ( $P = 0.249$ ), indicating that DNA methylation at cg13344587 was entirely  
35  
36 141 driven by rs7090445. In contrast, the *IKZF1* SNP rs78396808 was no longer significant in the  
37  
38 142 multivariable conditional model ( $P = 0.078$ ) but the *IKZF1* CpG cg01139861 remained significant  
39  
40 143 ( $P = 9.49 \times 10^{-4}$ ), suggesting that the effect of rs78396808 on ALL risk may be mediated by  
41  
42 144 cg01139861 and that DNA methylation at cg01139861 may be affected by risk factors other than  
43  
44 145 rs78396808. We next formally tested whether there were significant mediation effects for these  
45  
46 146 two SNP-DMP pairs.

47  
48  
49 147

### 50 51 148 **Mediation analysis**

52  
53  
54 149 Causal mediation analyses were performed for the significant *ARID5B* and *IKZF1* mQTL-DMP  
55  
56 150 pairs with ALL risk separately in the CCLS 450K, CCLS EPIC, and CCRLP EPIC datasets (**Figure**

1  
2  
3 151 1). The total effect, average direct effect (ADE) and average causal mediation effect (ACME)  
4  
5 152 estimated for each mQTL-DMP pair were summarized across all three datasets using the fixed-  
6  
7 153 effect meta-analysis model.

8  
9 154 SNP rs7090445 at gene *ARID5B* had a significant total effect (estimate=0.108,  $P=5.17 \times 10^{-15}$ )  
10  
11 155 and direct effect (ADE estimate=0.093,  $P=1.06 \times 10^{-4}$ ) but a nonsignificant causal mediation  
12  
13 156 effect (ACME  $P=0.223$ ) on ALL risk through altering DNA methylation at the *ARID5B* CpG  
14  
15 157 cg13344587 (**Table 5 and Figure 3A**), demonstrating that the effect of rs7090445 on ALL risk  
16  
17 158 was independent of cg13344587.

18  
19  
20 159 In contrast, we found that *IKZF1* SNP rs78396808 had a significant total effect on ALL risk  
21  
22 160 (estimate=0.056,  $P=0.010$ ), a significant mediation effect through increasing DNA methylation at  
23  
24 161 the *IKZF1* CpG cg01139861 (ACME estimate=0.017,  $P=0.003$ ) and a nonsignificant direct effect  
25  
26 162 (estimate=0.039,  $P=0.077$ ), with the ACME explaining ~30% (0.017/0.056) of the total effect  
27  
28 163 (**Table 5 and Figure 3B**). After conditioning on the lead *IKZF1* SNP rs10230978, we observed  
29  
30 164 stronger total effect (estimate=0.089,  $P=1.24 \times 10^{-4}$ ) and direct effect (estimate=0.073,  $P=0.002$ ) of  
31  
32 165 rs78396808 on ALL risk, and a similar mediation effect through cg01139861 (estimate=0.016,  
33  
34 166  $P=0.006$ ) (**Table 5 and Figure 3B**). These indicate that the *IKZF1* SNP rs78396808 conferred  
35  
36 167 risk for ALL both through cg01139861 and independent of cg01139861. The associations  
37  
38 168 between DNA methylation at cg01139861, genotypes of rs78396808, and ALL risk in each study  
39  
40 169 set are illustrated in **Figure 4A**. The *IKZF1* SNP rs78396808 risk allele (A) increased DNA  
41  
42 170 methylation at cg01139861 (**Table 3 and Figure 4A**), and the increased DNA methylation level  
43  
44 171 at cg01139861 was associated with increased ALL risk (**Table 2 and Figure 4A**).

45  
46 172 Results from causal mediation analyses with variance estimation based on the  
47  
48 173 nonparametric bootstrap method (**Table S7**) were similar to those from the quasi-Bayesian Monte  
49  
50 174 Carlo simulation (**Table 5**).

51  
52 175

## 53 54 55 56 176 **Race/ethnicity stratified analyses**



1  
2  
3 177 Significant SNP-DMP pairs in overall participants had similar effect estimates in self-reported  
4  
5 178 Latinos, non-Latino whites, and non-Latinos (non-Latino whites plus non-Latino others) in  
6  
7 179 stratified mQTL analysis, with no significant heterogeneity by race/ethnicity in tests of moderators  
8  
9 180 in fixed-effect meta-analyses (**Table S8**). However, SNP rs78396808 at gene *IKZF1* was  
10  
11 181 associated with cg01139861 in overall participants, in Latinos, and in non-Latinos (non-Latino  
12  
13 182 whites plus non-Latino others) but was not significantly associated with cg01139861 in non-Latino  
14  
15 183 whites. SNP rs78396808 is almost monomorphic for the non-risk allele G in European populations  
16  
17 184 in the Genome Aggregation Database, whereas the risk allele A frequency is ~20% in East Asian  
18  
19 185 and Admixed American populations (29). We found the GA genotype among 11 individuals who  
20  
21 186 were self-reported non-Latino white (**Figure 4A**), likely due to admixture, which may result in the  
22  
23 187 slightly higher effect estimate and standard error for the association between rs78396808 and  
24  
25 188 cg01139861 in non-Latino whites than in Latinos (**Table S8**).

26  
27  
28 189 Next, we conducted causal mediation analyses separately in Latinos, non-Latino whites,  
29  
30 190 and non-Latinos, for those significant SNP-DMP pairs identified in overall participants. We found  
31  
32 191 a significant total effect and a significant mediation effect for rs78396808(*IKZF1*)-  
33  
34 192 cg01139861(*IKZF1*)-ALL in the overall meta-analysis in Latinos, with the total effect explained by  
35  
36 193 the ACME being ~42% (0.022/0.053) and ~28% (0.022/0.078) before and after conditioning on  
37  
38 194 the top *IKZF1* SNP rs10230978 (**Table S9**), which were both higher than that observed in overall  
39  
40 195 participants. We found no significant heterogeneity in the ACME for rs78396808(*IKZF1*)-  
41  
42 196 cg01139861(*IKZF1*)-ALL between Latinos and non-Latino whites ( $P_{\text{het}}=0.441$ ) or between Latinos  
43  
44 197 and non-Latinos ( $P_{\text{het}}=0.236$ ) (**Table S9**), likely due to lack of power given the low allele frequency  
45  
46 198 of rs78396808 in non-Latino whites.

47  
48  
49 199

#### 50 200 ***ARID5B* CpG cg13344587 is a proxy for ALL risk SNP rs7090445**

51  
52 201 To further disentangle the mQTL-DMP-ALL associations analyzed in the causal mediation  
53  
54 202 analysis, we investigated whether there was a confounding effect from the mQTL genotype on  
55  
56  
57

1  
2  
3 203 the association between DNA methylation and ALL risk by fitting three unconditional logistic  
4  
5 204 regression models in subjects with both DNA methylation and genotype data available. Results  
6  
7 205 for both rs7090445(*ARID5B*)-cg13344587(*ARID5B*)-ALL and rs78396808(*IKZF1*)-  
8  
9 206 cg01139861(*IKZF1*)-ALL are summarized in **Table 4**. We obtained the “crude effect” of every 0.1  
10  
11 207 beta-value increase in *ARID5B* cg13344587 methylation on ALL risk in model 1 ( $OR_{meta}=0.44$ ,  
12  
13 208  $P=2.87 \times 10^{-10}$ ) and the “adjusted effect” in model 2 ( $OR_{meta}=0.80$ ,  $P=0.249$ ), with a 82% reduction  
14  
15 209 after adjusting for the SNP effect (**Figure 3A**), much higher than the 10% difference for identifying  
16  
17 210 the presence of confounding (30). In addition, there was no longer a significant association  
18  
19 211 between DNA methylation at cg13344587 and ALL risk, with a greatly reduced effect estimate  
20  
21 212 ( $OR_{meta}=0.99$ ,  $P=0.970$ ) in model 3 in individuals without any copies of the rs7090445 risk allele.  
22  
23 213 Therefore, the association between decreased DNA methylation at cg13344587 and ALL risk is  
24  
25 214 consistent with confounding by SNP rs7090445. In contrast, we observed only an ~4% decrease  
26  
27 215 in the effect on ALL risk from CpG cg01139861 at *IKZF1* after adjusting for rs78396808, and the  
28  
29 216 effect remained significant in individuals without any copies of the rs78396808 risk allele  
30  
31 217 ( $OR_{meta}=1.45$ ,  $P=0.004$ ). These demonstrated no evidence of a confounding effect from  
32  
33 218 rs78396808 on the association between *IKZF1* cg01139861 and ALL risk.  
34  
35  
36  
37  
38

### 39 220 ***IKZF1* gene-specific analysis**

40  
41 221 We investigated additional *IKZF1* CpGs that might mediate the effects of SNP rs78396808 on  
42  
43 222 ALL risk. There were 36 and 42 *IKZF1* CpGs included in the overall meta-analysis and the EPIC  
44  
45 223 array meta-analysis, respectively. CpGs were tested for their association with ALL, with  
46  
47 224 cg01139861 the only one passing gene-wide significance ( $P<0.025/36$ ), and cg16499656 was an  
48  
49 225 additional CpG with  $P<0.025$  in the overall meta-analysis (**Table S10**). The CpG cg01139861  
50  
51 226 analyzed before was not included in subsequent analyses. CpG cg12431065 passed the gene-  
52  
53 227 wide significance threshold ( $0.025/42$ ), and cg10551353 was an additional CpG with  $P<0.025$  in  
54  
55 228 the EPIC array meta-analysis.  
56  
57  
58  
59  
60

229 Since rs78396808 is monomorphic in European populations, we further conducted the  
230 *IKZF1* gene-specific mQTL and causal mediation analyses in Latinos only. *IKZF1* CpGs  
231 significantly associated with ALL were tested for their associations with SNP rs78396808 with a  
232 p-value  $<0.025$  or  $<0.0125$  ( $0.025/2$ ) considered to show statistical significance in the overall  
233 meta-analysis and the EPIC array meta-analysis, respectively. We identified one significant SNP-  
234 DMP pair from the mQTL overall meta-analysis, and two significant SNP-DMP pairs from the  
235 mQTL EPIC array meta-analysis (**Table S11**). We found a significant total effect from rs78396808  
236 on ALL (estimate=0.067,  $P=0.030$ ) and a significant mediation effect from rs78396808  
237 (estimate=0.019,  $P=0.031$ ) through increasing DNA methylation at cg10551353 when  
238 conditioning on rs10230978 in Latinos, with a ~29% total effect explained by the ACME (**Table**  
239 **S12**). We also found in Latinos that DNA methylation at cg10551353 was strongly correlated with  
240 cg01139861, the original *IKZF1* CpG found to mediate the effect of rs78396808 on ALL risk  
241 (CCRLP EPIC:  $R^2=0.41$ ,  $P=3.19 \times 10^{-14}$ ; CCLS EPIC:  $R^2=0.45$ ,  $P=8.23 \times 10^{-16}$ ) (**Figure S2**). CpG  
242 cg01139861 is located in a CpG island in the promoter region of *IKZF1*, and SNP rs78396808 is  
243 in an intronic region ~116Kb downstream (**Table 2 and Figure 4B**). The additional *IKZF1* CpG  
244 cg10551353 identified here is in the 5' UTR or the TSS1500 region of several alternative  
245 transcripts, ~14Kb downstream of cg01139861 (**Table S10 and Figure 4B**).

247 ***IKZF1* SNP rs78396808 is independent of another secondary association signal at SNP**  
248 ***rs6421315***

249 In a previous GWAS of ALL in individuals of European ancestry, Vijayakrishnan *et al.* (31)  
250 identified a secondary signal at the *IKZF1* locus at SNP rs6421315 after conditioning on their lead  
251 SNP rs17133805. SNP rs17133805 is in nearly perfect linkage disequilibrium (LD) ( $R^2=0.999$ )  
252 with the lead *IKZF1* SNP rs10230978 in our multi-ancestry GWAS meta-analysis of ALL (28).  
253 Here, we explored whether the genotypes of rs6421315 and rs78396808, the secondary  
254 association signal at *IKZF1* in our multi-ancestry ALL GWAS, were independently associated with

1  
2  
3 255 ALL. The two secondary hits rs78396808 and rs6421315 had weak LD in all populations  
4  
5 256 ( $R^2=0.0188$ ,  $D'=0.4213$ ) and in Admixed Americans ( $R^2=0.072$ ,  $D'=0.5546$ ) based on LDlink (32),  
6  
7 257 and reside on different sides of a recombination peak (**Figure S3**). Additionally, SNP rs78396808  
8  
9 258 remained significantly associated with ALL in this study when conditioning on both rs10230978  
10  
11 259 and rs6421315 ( $P_{\text{meta}}=0.007$ ) (**Table S13**).

13  
14 260

15  
16 261 ***Increased DNA methylation at IKZF1 CpG cg01139861 correlated with decreased IKZF1***  
17  
18 262 ***expression***

19  
20 263 Finally, we tested the correlation for DNA methylation at cg01139861 (*IKZF1*) with gene  
21  
22 264 expression of nearby genes *IKZF1*, *FIGNL1*, and *DDC* using Spearman correlation coefficient  
23  
24 265 tests in 51 ALL tumor samples from CCLS. We excluded 11/71 samples without multiplex ligation-  
25  
26 266 dependent probe amplification (MLPA) copy-number data and 9/60 samples with one deleted  
27  
28 267 copy of *IKZF1*. Increased DNA methylation at the *IKZF1* CpG cg01139861 significantly correlated  
29  
30 268 with decreased gene expression of *IKZF1* ( $R=-0.28$ ,  $P=0.044$ ; **Figure 5**). In the 9/60 ALL tumors  
31  
32 269 with hemizygous *IKZF1* deletion, DNA methylation levels at cg01139861 were on average lower  
33  
34 270 than in the 51 cases without deletion (median = 0.226 vs. 0.410), although the difference did not  
35  
36 271 reach significance in a Wilcoxon rank sum test ( $P=0.094$ ) (**Figure S4**).

37  
38  
39 272

40  
41 273 **Discussion**

42  
43  
44 274 A role for genetic variation in the etiology of childhood ALL is well established, but little is known  
45  
46 275 regarding the association of epigenetic differences at birth and future development of ALL. We  
47  
48 276 report results from the largest neonatal DNA methylation EWAS of childhood ALL performed to  
49  
50 277 date, along with the first comprehensive mediation analysis investigating whether DNA  
51  
52 278 methylation mediates the effects of ALL genetic risk loci. We found that the *IKZF1* SNP  
53  
54 279 rs78396808 risk allele conferred risk for ALL through increasing DNA methylation at the *IKZF1*

1  
2  
3 280 promoter CpG cg01139861. In addition, the *ARID5B* CpG cg13344587, the only ALL-associated  
4  
5 281 DMP to survive Bonferroni correction, appeared to be entirely confounded by the *ARID5B* ALL  
6  
7 282 risk SNP rs7090445.  
8

9  
10 283 A limited number of epigenetic studies have been conducted previously for childhood ALL  
11  
12 284 (22–25), with DNA methylation of cases profiled using bone marrow or peripheral blood samples  
13  
14 285 collected from ALL patients at diagnosis. Few studies have investigated differential DNA  
15  
16 286 methylation associated with subsequent development of ALL. We note a paucity of ALL-  
17  
18 287 associated CpGs in our EWAS, with the *ARID5B* CpG cg13344587 being the exception that  
19  
20 288 survived Bonferroni correction. Using a more lenient threshold of  $P < 1 \times 10^{-4}$ , we identified 47 and  
21  
22 289 51 DMPs mapped to 37 and 32 genes from the overall meta-analysis and the EPIC array meta-  
23  
24 290 analysis, respectively. Altered DNA methylation at 12 of these genes (*APCDD1*, *AVPR1A*,  
25  
26 291 *CAMTA1*, *CHST8*, *EPHA10*, *EYA4*, *GP5*, *NHLRC1*, *OLFM3*, *PLOD2*, *SKAP1*, and *XKR9*) has  
27  
28 292 been observed previously in ALL tumor samples (22,23), and *ARID5B* is an established ALL  
29  
30 293 predisposition gene.  
31

32  
33 294 Intronic SNPs in *ARID5B*, which plays an important role in B-cell development, have been  
34  
35 295 associated with ALL risk in several GWAS (4,7,31). Functional analysis has shown that the ALL  
36  
37 296 risk SNP rs70904455 disrupts binding of *RUNX3* and leads to reduced *ARID5B* expression (33).  
38  
39 297 We found that the *ARID5B* SNP rs7090445-C risk allele was significantly associated with  
40  
41 298 decreased DNA methylation at the *ARID5B* CpG cg1334458, as previously reported in whole-  
42  
43 299 blood samples from cancer-free individuals (34,35). Although the rs7090445-C risk allele showed  
44  
45 300 the strongest total effect and direct effect on ALL in the causal mediation analysis, the lack of  
46  
47 301 significant mediation effect through altering DNA methylation indicates that the impact of this SNP  
48  
49 302 on ALL risk is independent of cg13344587. Further, results from our analysis of confounding  
50  
51 303 effects supported that the association between decreased DNA methylation at cg13344587 and  
52  
53 304 ALL risk appears to be entirely driven by the *ARID5B* SNP rs7090445, and hypomethylated  
54  
55 305 cg13344587 may function merely as a strong proxy of the rs7090445-C risk allele.  
56  
57  
58  
59  
60

1  
2  
3 306 In contrast, we found that DNA methylation at CpG cg01139861, located in a CpG island  
4  
5 307 in the promoter region of *IKZF1*, mediates the effects of the *IKZF1* SNP rs78396808, which we  
6  
7 308 recently reported as an independent ALL-association signal in analysis conditioned on the lead  
8  
9 309 *IKZF1* SNP rs10230978 (28). The causal mediation analysis in our overall dataset showed that  
10  
11 310 the rs78396808-A risk allele had a significant ACME on ALL risk through increasing DNA  
12  
13 311 methylation at cg01139861, explaining ~30% of the total effect on ALL. In analyses stratified by  
14  
15 312 self-reported race/ethnicity, the mediation effect for rs78396808 through increasing DNA  
16  
17 313 methylation at cg01139861 was only significant in Latinos, explaining a ~42% total effect in this  
18  
19 314 population. The SNP rs78396808 is monomorphic in European populations, although the risk  
20  
21 315 allele (A) also presents in African and South Asian populations and has a ~20% frequency in East  
22  
23 316 Asians (29); however, we did not have sufficient samples to test for mediation effects for these  
24  
25 317 population groups. In addition, the *IKZF1* SNP rs78396808 appears to be independent of another  
26  
27 318 secondary association signal at *IKZF1*, SNP rs6421315, previously identified in individuals of  
28  
29 319 European ancestry<sup>13</sup>, supporting the existence of at least three independent common genetic risk  
30  
31 320 loci for ALL across populations. In the multivariable regression adjusting for the SNP effect from  
32  
33 321 rs78396808, DNA methylation at cg01139861 remained significantly associated with ALL risk,  
34  
35 322 suggesting that other genetic or environmental risk factors may also affect this CpG site.  
36  
37  
38

39 323 Further, we found that increased DNA methylation at cg01139861 correlated with  
40  
41 324 decreased gene expression of *IKZF1* in ALL tumor samples. Gene *IKZF1* encodes the lymphoid  
42  
43 325 transcription factor IKAROS, and is essential for lymphocyte development and differentiation (36).  
44  
45 326 Somatic deletion of *IKZF1* is a common driver event in ALL, particularly in BCR-ABL1-positive  
46  
47 327 ALL (95%) (37) and in high-risk B-cell ALL (30%) (38). *IKZF1* deletions are also enriched in  
48  
49 328 patients with relapsed childhood B-cell ALL (39.3%), in whom increased promoter methylation  
50  
51 329 was also found (21). However, biallelic loss of *IKZF1* is infrequent among ALL patients at  
52  
53 330 diagnosis (39), and we found no evidence that the hypermethylated copy of cg01139861 is  
54  
55 331 selectively retained in ALL tumor samples with hemizygous *IKZF1* deletions; in fact, a lack of  
56  
57  
58  
59  
60



1  
2  
3 332 biallelic deletions along with lower DNA methylation in the retained *IKZF1* allele in patients  
4  
5 333 exhibiting a monoallelic deletion may indicate that haploinsufficiency of *IKZF1* rather than  
6  
7 334 complete abrogation is necessary for leukemogenesis. Taken together, the leukemogenic effects  
8  
9 335 of the rs78396808-A risk allele may act via downregulation of *IKZF1* gene expression partly  
10  
11 336 through increased DNA methylation at the *IKZF1* CpG cg01139861. In our targeted analysis of  
12  
13 337 CpGs across *IKZF1*, we identified one additional CpG cg10551353, in the promoter region of  
14  
15 338 several transcripts, which showed evidence of some mediation effect for rs78396808 on ALL risk.  
16  
17 339 Increased DNA methylation at cg10551353 was significantly correlated with increased DNA  
18  
19 340 methylation at cg01139861; however, cg10551353 is on the EPIC array only, so we could not  
20  
21 341 assess its association with *IKZF1* gene expression using the tumor samples assayed on 450K  
22  
23 342 arrays.

24  
25  
26 343 The current study has several strengths. First, we assayed pre-diagnostic DNA from  
27  
28 344 neonatal DBS on both genome-wide DNA methylation arrays and SNP arrays, which rules out  
29  
30 345 the possibility of reverse causality (i.e., effects of leukemia itself on DNA methylation). Second,  
31  
32 346 instead of using the traditional mediation analysis approaches relying on the restrictive and  
33  
34 347 untested assumptions, we used a more general estimation framework that provides distribution-  
35  
36 348 free estimates for causal mediation effects and accommodates nonlinearities (40,41). Moreover,  
37  
38 349 we estimated the uncertainty of the causal mediation effects through the quasi-Bayesian Monte  
39  
40 350 Carlo simulation, and we validated our results by using an alternative simulation approach based  
41  
42 351 on nonparametric bootstrap. Last, over half of the participants included in this study were self-  
43  
44 352 reported Latinos, providing an opportunity for us to perform analyses stratified by race/ethnicity,  
45  
46 353 through which we detected a stronger mediation effect for the *IKZF1* SNP rs78396808 through  
47  
48 354 cg01139861 in Latinos than in overall participants.

49  
50  
51 355 Our study does have some limitations. One potential limitation was sample size, with only  
52  
53 356 the *ARID5B* CpG cg13344587 reaching epigenome-wide significance in our EWAS for ALL. This  
54  
55 357 necessitated the use of a more lenient p-value threshold ( $P < 1 \times 10^{-4}$ ) to identify DMPs for  
56  
57  
58  
59  
60

1  
2  
3 358 subsequent mQTL and causal mediation analyses, which may have introduced false positive  
4  
5 359 results, especially in the EPIC array meta-analysis that was conducted with a smaller sample size.  
6  
7 360 We were also limited in our ability to identify DMPs for specific ALL subtypes, or CpG mediators  
8  
9 361 for ALL genetic risk loci associated with particular subtypes, analyses that would require a larger  
10  
11 362 sample size of ALL cases with well-defined tumor subtypes. Another limitation is that we were  
12  
13 363 limited by the number of CpGs on the Illumina arrays. Additional CpGs at *IKZF1* that we could not  
14  
15 364 assess with the array data may also mediate the effect of ALL risk SNP rs78396808. Bisulfite  
16  
17 365 sequencing targeting the *IKZF1* region will be required to fully capture the DNA methylation  
18  
19 366 changes that mediate the effect of rs78396808, especially at *IKZF1* regulatory regions that  
20  
21 367 correlate with gene expression patterns. Lastly, gene expression data were measured from  
22  
23 368 diagnostic leukemia samples and, although we accounted for somatic copy-number loss of *IKZF1*,  
24  
25 369 this may have affected the accuracy of the correlation results between DNA methylation at  
26  
27 370 cg01139861 and *IKZF1* expression.

30  
31 371 In conclusion, we provide evidence that increased DNA methylation at the *IKZF1* CpG  
32  
33 372 cg01139861 mediates the effects on ALL risk from SNP rs78396808, which was recently identified  
34  
35 373 as a novel independent risk locus at *IKZF1* that is specific to non-European populations. Our  
36  
37 374 findings enhance the understanding of the functional pathways of genetic risk loci for childhood  
38  
39 375 ALL and provide new insights into the DNA methylation differences associated with childhood  
40  
41 376 ALL.

42  
43 377

## 44 45 378 **Materials and Methods**

### 46 47 48 379 **Study participants**

49  
50 380 Study participants were included from two independent California-based case-control studies of  
51  
52 381 childhood leukemia, the CCLS and the CCRLP, the details of which have been described  
53  
54 382 previously (7,42). Briefly, the CCLS is a population-based case-control study conducted from  
55  
56  
57  
58  
59  
60



1  
2  
3 383 1995 to 2015 in multiple counties across California. Cases were identified within 72 hours after  
4  
5 384 diagnosis at hospitals and were eligible for participation if they met all the following criteria: (1)  
6  
7 385 age younger than 15 years, (2) without previous cancer diagnosis, (3) residence in California at  
8  
9 386 the time of diagnosis, and (4) having an English or Spanish-speaking biological parent or guardian.  
10  
11 387 Controls were randomly selected with similar eligibility criteria using birth certificates. One or two  
12  
13 388 controls were matched to each case on the date of birth, sex, and race/ethnicity. The CCRLP  
14  
15 389 linked statewide birth certificates from the California Office of Vital Records (for 1978-2009) to  
16  
17 390 statewide cancer diagnosis data from the California Cancer Registry (1988-2011). Cases were  
18  
19 391 children born in California and diagnosed with their first primary ALL at 0-15 years. Potential  
20  
21 392 controls were children born in California during the same period without prior reports of childhood  
22  
23 393 cancer. Up to four controls were randomly selected and matched to each case on the date of  
24  
25 394 birth, sex, and race/ethnicity. DBS samples were obtained from the California Biobank Program  
26  
27 395 for all participants. Cases (N=808) and controls (N=919) with available genome-wide DNA  
28  
29 396 methylation data were included in the EWAS. Analyses for identifying mQTL and mediation effects  
30  
31 397 were limited to 683 cases and 804 controls with both genome-wide DNA methylation and SNP  
32  
33 398 array data available (therefore, matching between cases and controls was broken for all analyses).  
34  
35  
36  
37  
38

### 39 **DNA methylation arrays**

40  
41 401 DNA samples were isolated from newborn DBS for 850 ALL cases and 931 cancer-free controls,  
42  
43 402 bisulfite converted, and then assayed on either the 450K DNA methylation arrays or EPIC arrays,  
44  
45 403 as previously described (43–45). EPIC arrays include >850,000 CpG probes, comprising >90%  
46  
47 404 of CpGs on the 450K array plus an additional 413,743 CpGs. ALL cases and controls were  
48  
49 405 randomized into different plates in each study set. CpG beta values were normalized to remove  
50  
51 406 batch effects according to the approach by Fortin et al. (46) Mean detection p-values were  
52  
53 407 calculated by using the “detectionP” function in the minfi (47) package through the Bioconductor  
54  
55 408 project (48,49). Functional normalization was performed with “noob” background correction (50)  
56  
57  
58  
59  
60

1  
2  
3 409 by using the “preprocessFunnorm” function in the minfi package. The beta-mixture quantile  
4  
5 410 normalization method was additionally applied (51). Samples with mean detection p-value >0.01  
6  
7 411 were considered poor quality and were removed from the analysis. CpG sites and samples that  
8  
9 412 had over 15% missing values were removed. The R package “conumee” (52) was used to  
10  
11 413 generate copy-number variation plots to detect constitutive trisomy of chromosome 21 (T21), as  
12  
13 414 previously described, and a total of 27 ALL cases and 1 control with T21 were excluded from  
14  
15 415 subsequent analyses given the profound effects of Down syndrome on DNA methylation (45).  
16  
17  
18 416

### 417 **Genome-wide SNP genotyping**

418 Genome-wide SNP array data were available from constitutive DNA samples isolated from  
419 newborn DBS for a subset of 683 cases and 804 controls. CCLS and CCRLP samples were  
420 genotyped using the Illumina Human OmniExpress V1 platform and the Affymetrix Axiom World  
421 (Latino) Array (7), respectively. Genotype data for 23 SNPs previously associated with childhood  
422 ALL were included from our recent multi-ancestry GWAS meta-analysis (28).  
423

### 424 **Identification of differentially methylated positions (DMPs)**

425 We first performed EWAS analyses separately in the CCLS 450K, CCLS EPIC, and CCRLP EPIC  
426 datasets to identify ALL-associated DMPs on autosomal chromosomes. Probes with common  
427 SNPs (minor allele frequency  $\geq 0.05$ ) in the full capture sequence or with SNPs in the targeted CpG  
428 site or its single base extension were removed (53,54). To minimize false-positive findings, we  
429 additionally removed cross-reactive probes identified previously (55–57). We fitted a logistic  
430 regression model predicting ALL case/control status as a function of DNA methylation at each  
431 remaining CpG, adjusting for sex, batch effect, cell type heterogeneity using the first ten principal  
432 components derived from ReFACTor (58), and genetic ancestry using the first ten principal  
433 components derived from EPISTRUCTURE (59). Fixed-effect meta-analysis models were used  
434 to generate summary effect estimates for the EWAS results of the CpGs overlapping both 450K

1  
2  
3 435 and EPIC arrays from three different study sets – CCLS 450K, CCLS EPIC, and CCRLP EPIC  
4  
5 436 datasets – using the R package “metafor” (60). We performed a second meta-analysis for CpGs  
6  
7 437 on EPIC arrays that were limited to the CCLS and CCRLP EPIC datasets only and not included  
8  
9 438 in the CCLS 450K dataset. The associations between CpGs and ALL were corrected for multiple  
10  
11 439 testing using a stringent Bonferroni-adjusted threshold of 0.025 (0.05/2) divided by the number of  
12  
13 440 CpGs included in each meta-analysis, and an FDR of 0.05. In addition, given that previous studies  
14  
15 441 have applied a more liberal threshold (<0.001) to identify DMPs for downstream mediation and  
16  
17 442 interaction analyses (26,27), here we applied a lenient threshold of  $1 \times 10^{-4}$  to ensure sufficient  
18  
19 443 numbers of candidate CpGs for subsequent mQTL and mediation analyses. Manhattan plots and  
20  
21 444 QQ plots were generated using the R package “CMplot” (61). For the identified DMPs, we  
22  
23 445 compared the effect estimates of these CpGs in overall participants with those in self-reported  
24  
25 446 Latinos and non-Latino whites in stratified EWAS, and in EWAS adjusting for genetic ancestry  
26  
27 447 using the first ten principal components derived from the genome-wide SNP array data with PLINK  
28  
29 448 2.0 (62) instead of from EPISTRUCTURE, in subjects with both DNA methylation and SNP  
30  
31 449 genotype data available.  
32  
33  
34  
35  
36

450

### 451 **Identification of methylation quantitative trait loci (mQTL)**

37  
38  
39 452 We carried out mQTL analyses to identify genotype-dependent DMPs associated with childhood  
40  
41 453 ALL risk using the R package “Matrix eQTL” (63). DMPs with  $P < 1 \times 10^{-4}$  from the overall meta-  
42  
43 454 analysis or the EPIC array meta-analysis were tested for association with genotypes of the 23  
44  
45 455 ALL risk SNPs. We fitted an additive linear regression model predicting methylation at each CpG  
46  
47 456 site as a function of SNP genotype (coded 0, 1, and 2), adjusting for the same covariates as for  
48  
49 457 the EWAS. The associations between SNP genotypes and DMP DNA methylation were corrected  
50  
51 458 for multiple testing using a stringent Bonferroni-adjusted threshold of  $0.025 / (\text{number of DMPs} \times$   
52  
53 459  $\text{number of SNPs})$ , and an FDR of 0.05. The mQTL analysis was first conducted separately in the  
54  
55 460 CCLS 450K, CCLS EPIC, and CCRLP EPIC datasets, and the results were subsequently meta-  
56  
57  
58  
59  
60

1  
2  
3 461 analyzed across all three datasets, and across the two EPIC datasets, in fixed-effect meta-  
4  
5 462 analysis models using “metafor”. We compared these SNP-DMP associations estimated with the  
6  
7 463 DNA methylation beta values versus those estimated when including DNA methylation M-values  
8  
9 464 in the linear regression models (64).  
10

11 465

### 12 466 **Mediation analysis**

13  
14  
15 467 We next performed model-based causal mediation analyses for the significant mQTL-DMP pairs,  
16  
17 468 using the “mediation” R package (40). First, we specified two statistical models, (1) the mediator  
18  
19 469 model for the distribution of the DMP methylation level, after conditioning on the genotype of the  
20  
21 470 mQTL and covariates including sex, ancestry, batch effect, and cell type heterogeneity, and (2)  
22  
23 471 the outcome model for the conditional distribution of ALL status, given the mQTL genotype, DMP  
24  
25 472 methylation level, and the same covariates. Models were fitted separately, and then their fitted  
26  
27 473 parameters were used as the main inputs to the mediate function, which computes the estimated  
28  
29 474 ACME, the ADE, and the total effect. Variances were estimated based on simulation. The quasi-  
30  
31 475 Bayesian Monte Carlo simulation based on normal approximation was conducted 1000 times.  
32  
33 476 Alternatively, an approach based on nonparametric bootstrap was also applied to estimate  
34  
35 477 variance for validation. Models with the *IKZF1* SNP rs78396808 were additionally adjusted for the  
36  
37 478 *IKZF1* lead SNP rs10230978, as rs78396808 was previously reported as a secondary ALL  
38  
39 479 association signal in analysis conditioned on rs10230978 (28). Results of the mediation analysis  
40  
41 480 performed separately for the CCLS 450K, CCLS EPIC, and CCRLP EPIC datasets were  
42  
43 481 summarized across all three datasets and across the two EPIC datasets using the fixed-effect  
44  
45 482 meta-analysis model.  
46  
47  
48

49 483 We repeated the mQTL analyses and the causal mediation analyses stratified by self-  
50  
51 484 reported race/ethnicity (Latinos vs. non-Latino whites, and Latinos vs. non-Latinos [i.e., non-  
52  
53 485 Latino whites plus non-Latino others]), in study participants with available genome-wide DNA  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 486 methylation and SNP array data. We included race/ethnicity as a moderator variable in fixed-  
4  
5 487 effect meta-analysis models to test for heterogeneity.  
6  
7 488

### 9 489 **Locus-specific analyses**

10  
11 490 To investigate whether there was a confounding effect from the mQTL genotype on the  
12  
13 491 association between DNA methylation and ALL risk, we fitted three unconditional logistic  
14  
15 492 regression models in subjects with both DNA methylation and genotype data available: model 1  
16  
17 493 was a logistic regression predicting ALL risk as a function of DMP DNA methylation, adjusting for  
18  
19 494 sex, batch effect, cell type heterogeneity, and genetic ancestry; model 2 was additionally adjusted  
20  
21 495 for the mQTL genotype; and model 3 was model 1 fitted in individuals without any copies of the  
22  
23 496 mQTL risk allele. The odds ratios for each 0.1 beta-value increase in DNA methylation in models  
24  
25 497 1 and 2 were considered as the “crude effect” and the “adjusted effect”, respectively. A reduction  
26  
27 498 >10% of the “adjusted effect” from the “crude effect” provides evidence of confounding (30). In  
28  
29 499 addition, a nonsignificant coefficient in model 3 indicates that the association between DNA  
30  
31 500 methylation and ALL risk is entirely confounded by the SNP effect.  
32  
33

34  
35 501 We also performed gene-specific analysis to investigate additional CpGs in the  
36  
37 502 neighboring region of the mediator CpG that could also mediate the effect of mQTL on ALL  
38  
39 503 susceptibility. First, we conducted the overall fixed-effect meta-analysis and the EPIC array fixed-  
40  
41 504 effect meta-analyses limiting to the EPIC array CpGs not in the CCLS 450K dataset for the gene-  
42  
43 505 specific DMP association testing. We retained those CpGs overlapping SNPs in the actual capture,  
44  
45 506 the single base extension, or the full capture sequence of the CpG sites to identify the potential  
46  
47 507 impact from the nearby SNPs. We applied a relaxed significance level of 0.025 in each meta-  
48  
49 508 analysis to ensure sufficient CpGs would be included in the following analysis. Next, significant  
50  
51 509 SNP-CpG associations were identified through the mQTL analysis, with a Bonferroni-adjusted  
52  
53 510 significance level of 0.025/number of CpGs. Finally, causal mediation analysis was performed to  
54  
55 511 identify mediators.  
56  
57  
58  
59  
60

1  
2  
3 5124  
5 513 **DNA methylation and gene expression analysis**

6  
7 514 DNA methylation and gene expression data were available from diagnostic leukemia (tumor)  
8  
9 515 samples of 71 ALL cases in the CCLS (25). DNA methylation data from 450K arrays were  
10  
11 516 processed as described above. Genome-wide gene expression data were generated using the  
12  
13 517 GeneChip Human Gene 1.0 ST Array (Affymetrix, Santa Clara, CA), as previously described (25).  
14  
15 518 Copy-number at 8 commonly deleted gene regions in childhood ALL was assayed in 60 out of 71  
16  
17 519 ALL tumors using MLPA, as previously described (65,66). For DMPs found to be mediators  
18  
19 520 between SNPs and ALL risk, we analyzed the associations between DNA methylation and the  
20  
21 521 expression levels of nearby genes using Spearman correlation coefficient tests, and we further  
22  
23 522 limited the correlation tests to cases without copy number deletion at the corresponding gene  
24  
25 523 regions (if available) to address potential confounding.

26  
27  
28 52429  
30  
31 525 **Funding**

32  
33 526 This work was supported by National Institutes of Health (NIH) National Cancer Institute grants  
34  
35 527 R01CA155461 and R01CA175737, National Institute for Environmental Health Sciences (NIEHS)  
36  
37 528 grants R01ES009137, P42ES004705, R24ES028524, and P50ES018172, the United States  
38  
39 529 Environmental Protection Agency grant RD83615901, and the California Tobacco-Related  
40  
41 530 Disease Research Program, Grant No. 26IR-0005A. The content of this manuscript is solely the  
42  
43 531 responsibility of the authors and does not necessarily represent the official views of the National  
44  
45 532 Institutes of Health or USEPA or TRDRP. Biospecimens and/or data used in this study were  
46  
47 533 obtained from the California Biobank Program at the California Department of Public Health  
48  
49 534 (CDPH), CBP request number 26, in accordance with Section 6555(b), 17 CCR. The CDPH is  
50  
51 535 not responsible for the results or conclusions drawn by the authors of this publication.

52  
53  
54 536  
55  
56  
57  
58  
59  
60

## 537 **Acknowledgments**

538 We thank Robin Cooley and Steve Graham (Genetic Disease Screening Program, CDPH) for  
539 their assistance and expertise in the procurement and management of DBS specimens. We also  
540 thank Hong Quach and Diana Quach at the UC Berkeley QB3 Genetic Epidemiology and  
541 Genomics Laboratory for their support in preparing and processing samples for genome-wide  
542 DNA methylation arrays. The authors additionally thank the families for their participation in the  
543 California Childhood Leukemia Study (formerly known as the Northern California Childhood  
544 Leukemia Study). For recruitment of subjects enrolled in the California Childhood Leukemia Study,  
545 the authors gratefully acknowledge the clinical investigators at the following collaborating  
546 hospitals: University of California, Davis Medical Center (Jonathan Ducore); University of  
547 California, San Francisco (Mignon Loh and Katherine Matthay); Children's Hospital of Central  
548 California (Vonda Crouse); Lucile Packard Children's Hospital (Gary Dahl); Children's Hospital  
549 Oakland (James Feusner and Carla Golden); Kaiser Permanente Roseville (formerly Sacramento)  
550 (Kent Jolly and Vincent Kiley); Kaiser Permanente Santa Clara (Carolyn Russo, Alan Wong, and  
551 Denah Taggart); Kaiser Permanente San Francisco (Kenneth Leung); Kaiser Permanente  
552 Oakland (Daniel Kronish and Stacy Month); California Pacific Medical Center (Louise Lo); Cedars-  
553 Sinai Medical Center (Fataneh Majlessipour); Children's Hospital Los Angeles (Cecilia Fu);  
554 Children's Hospital Orange County (Leonard Sender); Kaiser Permanente Los Angeles (Robert  
555 Cooper); Miller Children's Hospital Long Beach (Amanda Termuhlen); University of California,  
556 San Diego Rady Children's Hospital (William Roberts); and University of California, Los Angeles  
557 Mattel Children's Hospital (Theodore Moore).

558 The collection of cancer incidence data used in the CCRLP study was supported by the  
559 California Department of Public Health pursuant to California Health and Safety Code Section  
560 103885, Centers for Disease Control and Prevention's (CDC) National Program of Cancer  
561 Registries, under cooperative agreement 5NU58DP003862-04/DP003862, the National Cancer



1  
2  
3 562 Institute's Surveillance, Epidemiology and End Results Program under contract  
4  
5 563 HHSN261201000140C awarded to the Cancer Prevention Institute of California, contract  
6  
7 564 HHSN261201000035C awarded to the University of Southern California, and contract  
8  
9 565 HHSN261201000034C awarded to the Public Health Institute. The ideas and opinions expressed  
10  
11 566 herein are those of the author(s) and do not necessarily reflect the opinions of the State of  
12  
13 567 California, Department of Public Health, the National Institutes of Health, and the Centers for  
14  
15 568 Disease Control and Prevention or their Contractors and Subcontractors. This study used birth  
16  
17 569 data obtained from the State of California Center for Health Statistics and Informatics. The  
18  
19 570 California Department of Public Health is not responsible for the analyses, interpretations, or  
20  
21 571 conclusions drawn by the authors regarding the birth data used in this publication.  
22  
23  
24 572

### 26 573 **Conflict of Interest Statement**

28  
29 574 The authors have no conflicts of interest to declare. All co-authors have seen and agree with the  
30  
31 575 contents of the manuscript and there is no financial interest to report.  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



## References

1. Siegel, R.L., Miller, K.D., Fuchs, H.E. and Jemal, A. (2021) Cancer Statistics, 2021. *CA: A Cancer Journal for Clinicians*, **71**, 7–33.
2. Cancer Facts & Figures (2020) .
3. Essig, S., Li, Q., Chen, Y., Hitzler, J., Leisenring, W., Greenberg, M., Sklar, C., Hudson, M.M., Armstrong, G.T., Krull, K.R., *et al.* (2014) Estimating the risk for late effects of therapy in children newly diagnosed with standard risk acute lymphoblastic leukemia using an historical cohort: A report from the Childhood Cancer Survivor Study. *Lancet Oncol*, **15**, 841–851.
4. Treviño, L.R., Yang, W., French, D., Hunger, S.P., Carroll, W.L., Devidas, M., Willman, C., Neale, G., Downing, J., Raimondi, S.C., *et al.* (2009) Germline genomic variants associated with childhood acute lymphoblastic leukemia. *Nat Genet*, **41**, 1001–1005.
5. de Smith, A.J., Walsh, K.M., Francis, S.S., Zhang, C., Hansen, H.M., Smirnov, I., Morimoto, L., Whitehead, T.P., Kang, A., Shao, X., *et al.* (2018) BMI1 enhancer polymorphism underlies chromosome 10p12.31 association with childhood acute lymphoblastic leukemia. *International Journal of Cancer*, **143**, 2647–2658.
6. de Smith, A.J., Walsh, K.M., Morimoto, L.M., Francis, S.S., Hansen, H.M., Jeon, S., Gonseth, S., Chen, M., Sun, H., Luna-Fineman, S., *et al.* (2019) Heritable variation at the chromosome 21 gene ERG is associated with acute lymphoblastic leukemia risk in children with and without Down syndrome. *Leukemia*, **33**, 2746–2751.
7. Wiemels, J.L., Walsh, K.M., Smith, A.J. de, Metayer, C., Gonseth, S., Hansen, H.M., Francis, S.S., Ojha, J., Smirnov, I., Barcellos, L., *et al.* (2018) GWAS in childhood acute lymphoblastic leukemia reveals novel genetic associations at chromosomes 17q12 and 8q24.21. *Nat Commun*, **9**, 1–8.

- 1  
2  
3 600 8. Papaemmanuil, E., Hosking, F.J., Vijayakrishnan, J., Price, A., Olver, B., Sheridan, E.,  
4  
5 601 Kinsey, S.E., Lightfoot, T., Roman, E., Irving, J.A.E., *et al.* (2009) Loci on 7p12.2, 10q21.2  
6  
7 602 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. *Nat Genet*,  
8  
9 603 **41**, 1006–1010.
- 11 604 9. Perez-Andreu, V., Roberts, K.G., Harvey, R.C., Yang, W., Cheng, C., Pei, D., Xu, H.,  
12  
13 605 Gastier-Foster, J., Shuyu, E., Yew-Suang Lim, J., *et al.* (2013) Inherited GATA3 variants are  
14  
15 606 associated with Ph-like childhood acute lymphoblastic leukemia and risk of relapse. *Nat*  
16  
17 607 *Genet*, **45**, 1494–1498.
- 20 608 10. Feinberg, A.P. and Tycko, B. (2004) The history of cancer epigenetics. *Nat Rev Cancer*, **4**,  
21  
22 609 143–153.
- 24 610 11. Baylin, S.B. and Jones, P.A. (2016) Epigenetic Determinants of Cancer. *Cold Spring Harb*  
25  
26 611 *Perspect Biol*, **8**, a019505.
- 28 612 12. Jones, P.A. and Baylin, S.B. (2007) The Epigenomics of Cancer. *Cell*, **128**, 683–692.
- 30 613 13. Ehrlich, M. and Lacey, M. (2013) DNA Hypomethylation and Hemimethylation in Cancer. In  
31  
32 614 Karpf, A. R. (ed.), *Epigenetic Alterations in Oncogenesis*, Advances in Experimental  
33  
34 615 Medicine and Biology, Springer, New York, NY, pp. 31–56.
- 36 616 14. You, J.S. and Jones, P.A. (2012) Cancer Genetics and Epigenetics: Two Sides of the Same  
37  
38 617 Coin? *Cancer Cell*, **22**, 9–20.
- 41 618 15. Garraway, L.A. and Lander, E.S. (2013) Lessons from the Cancer Genome. *Cell*, **153**, 17–  
42  
43 619 37.
- 45 620 16. Shen, H. and Laird, P.W. (2013) Interplay Between the Cancer Genome and Epigenome.  
46  
47 621 *Cell*, **153**, 38–55.
- 49 622 17. Liu, Y., Aryee, M.J., Padyukov, L., Fallin, M.D., Hesselberg, E., Runarsson, A., Reinius, L.,  
50  
51 623 Acevedo, N., Taub, M., Ronninger, M., *et al.* (2013) Epigenome-wide association data  
52  
53 624 implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat*  
54  
55 625 *Biotechnol*, **31**, 142–147.

- 1  
2  
3 626 18. Dai, J.Y., Wang, X., Wang, B., Sun, W., Jordahl, K.M., Kolb, S., Nyame, Y.A., Wright, J.L.,  
4  
5 627 Ostrander, E.A., Feng, Z., *et al.* (2020) DNA methylation and cis-regulation of gene  
6  
7 628 expression by prostate cancer risk SNPs. *PLoS Genet*, **16**, e1008667.  
8  
9 629 19. Nedeljkovic, I., Lahousse, L., Carnero-Montoro, E., Faiz, A., Vonk, J.M., de Jong, K., van  
10  
11 630 der Plaats, D.A., van Diemen, C.C., van den Berge, M., Obeidat, M., *et al.* (2018) COPD  
12  
13 631 GWAS variant at 19q13.2 in relation with DNA methylation and gene expression. *Human*  
14  
15 632 *Molecular Genetics*, **27**, 396–405.  
16  
17 633 20. Hale, V., Hale, G.A., Brown, P.A. and Amankwah, E.K. (2017) A Review of DNA Methylation  
18  
19 634 and microRNA Expression in Recurrent Pediatric Acute Leukemia. *Oncology*, **92**, 61–67.  
20  
21 635 21. Hogan, L.E., Meyer, J.A., Yang, J., Wang, J., Wong, N., Yang, W., Condos, G., Hunger,  
22  
23 636 S.P., Raetz, E., Saffery, R., *et al.* (2011) Integrated genomic analysis of relapsed childhood  
24  
25 637 acute lymphoblastic leukemia reveals therapeutic strategies. *Blood*, **118**, 5218–5226.  
26  
27 638 22. Nordlund, J., Bäcklin, C.L., Wahlberg, P., Busche, S., Berglund, E.C., Eloranta, M.-L.,  
28  
29 639 Flaegstad, T., Forestier, E., Frost, B.-M., Harila-Saari, A., *et al.* (2013) Genome-wide  
30  
31 640 signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia.  
32  
33 641 *Genome Biol*, **14**, r105.  
34  
35 642 23. Chatterton, Z., Morenos, L., Mechinaud, F., Ashley, D.M., Craig, J.M., Sexton-Oates, A.,  
36  
37 643 Halemba, M.S., Parkinson-Bates, M., Ng, J., Morrison, D., *et al.* (2014) Epigenetic  
38  
39 644 deregulation in pediatric acute lymphoblastic leukemia. *Epigenetics*, **9**, 459–467.  
40  
41 645 24. Nordlund, J. and Syvänen, A.-C. (2018) Epigenetics in pediatric acute lymphoblastic  
42  
43 646 leukemia. *Seminars in Cancer Biology*, **51**, 129–138.  
44  
45 647 25. Lee, S.-T., Muench, M.O., Fomin, M.E., Xiao, J., Zhou, M., de Smith, A., Martín-Subero, J.I.,  
46  
47 648 Heath, S., Houseman, E.A., Roy, R., *et al.* (2015) Epigenetic remodeling in B-cell acute  
48  
49 649 lymphoblastic leukemia occurs in two tracks and employs embryonic stem cell-like  
50  
51 650 signatures. *Nucleic Acids Research*, **43**, 2590–2602.  
52  
53  
54  
55  
56  
57  
58  
59  
60

- 1  
2  
3 651 26. Shu, C., Justice, A.C., Zhang, X., Wang, Z., Hancock, D.B., Johnson, E.O. and Xu, K.  
4  
5 652 (2020) DNA methylation mediates the effect of cocaine use on HIV severity. *Clinical*  
6  
7 653 *Epigenetics*, **12**, 140.  
8  
9 654 27. Declerck, K., Remy, S., Wohlfahrt-Veje, C., Main, K.M., Van Camp, G., Schoeters, G.,  
10  
11 655 Vanden Berghe, W. and Andersen, H.R. (2017) Interaction between prenatal pesticide  
12  
13 656 exposure and a common polymorphism in the PON1 gene on DNA methylation in genes  
14  
15 657 associated with cardio-metabolic disease risk—an exploratory study. *Clinical Epigenetics*, **9**,  
16  
17 658 35.  
18  
19 659 28. Jeon, S., de Smith, A.J., Li, S., Chen, M., Chan, T.F., Muskens, I.S., Morimoto, L.M.,  
20  
21 660 DeWan, A.T., Mancuso, N., Metayer, C., *et al.* (2021) Genome-wide trans-ethnic meta-  
22  
23 661 analysis identifies novel susceptibility loci for childhood acute lymphoblastic leukemia.  
24  
25 662 *Leukemia*, 1–4.  
26  
27 663 29. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins,  
28  
29 664 R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., *et al.* (2019) The mutational constraint  
30  
31 665 spectrum quantified from variation in 141,456 humans. *The mutational constraint spectrum*  
32  
33 666 *quantified from variation in 141,456 humans*; preprint; Genomics, (2019) .  
34  
35 667 30. VanderWeele, T.J. (2019) Principles of confounder selection. *Eur J Epidemiol*, **34**, 211–219.  
36  
37 668 31. Vijayakrishnan, J., Qian, M., Studd, J.B., Yang, W., Kinnersley, B., Law, P.J., Broderick, P.,  
38  
39 669 Raetz, E.A., Allan, J., Pui, C.-H., *et al.* (2019) Identification of four novel associations for B-  
40  
41 670 cell acute lymphoblastic leukaemia risk. *Nat Commun*, **10**, 5348.  
42  
43 671 32. Machiela, M.J. and Chanock, S.J. (2015) LDlink: a web-based application for exploring  
44  
45 672 population-specific haplotype structure and linking correlated alleles of possible functional  
46  
47 673 variants. *Bioinformatics*, **31**, 3555–3557.  
48  
49 674 33. Studd, J.B., Vijayakrishnan, J., Yang, M., Migliorini, G., Paulsson, K. and Houlston, R.S.  
50  
51 675 (2017) Genetic and regulatory mechanism of susceptibility to high-hyperdiploid acute  
52  
53 676 lymphoblastic leukaemia at 10q21.2. *Nat Commun*, **8**.

- 1  
2  
3 677 34. Hannon, E., Weedon, M., Bray, N., O'Donovan, M. and Mill, J. (2017) Pleiotropic Effects of  
4  
5 678 Trait-Associated Genetic Variation on DNA Methylation: Utility for Refining GWAS Loci. *Am*  
6  
7 679 *J Hum Genet*, **100**, 954–959.
- 8  
9 680 35. Hannon, E., Dempster, E., Viana, J., Burrage, J., Smith, A.R., Macdonald, R., St Clair, D.,  
10  
11 681 Mustard, C., Breen, G., Therman, S., *et al.* (2016) An integrated genetic-epigenetic analysis  
12  
13 682 of schizophrenia: evidence for co-localization of genetic associations and differential DNA  
14  
15 683 methylation. *Genome Biology*, **17**, 176.
- 16  
17  
18 684 36. Harker, N., Naito, T., Cortes, M., Hostert, A., Hirschberg, S., Tolaini, M., Roderick, K.,  
19  
20 685 Georgopoulos, K. and Kioussis, D. (2002) The CD8alpha gene locus is regulated by the  
21  
22 686 Ikaros family of proteins. *Mol Cell*, **10**, 1403–1415.
- 23  
24 687 37. Mullighan, C.G., Miller, C.B., Radtke, I., Phillips, L.A., Dalton, J., Ma, J., White, D., Hughes,  
25  
26 688 T.P., Le Beau, M.M., Pui, C.-H., *et al.* (2008) BCR-ABL1 lymphoblastic leukaemia is  
27  
28 689 characterized by the deletion of Ikaros. *Nature*, **453**, 110–114.
- 29  
30 690 38. Mullighan, C.G., Su, X., Zhang, J., Radtke, I., Phillips, L.A.A., Miller, C.B., Ma, J., Liu, W.,  
31  
32 691 Cheng, C., Schulman, B.A., *et al.* (2009) Deletion of IKZF1 and Prognosis in Acute  
33  
34 692 Lymphoblastic Leukemia. *N Engl J Med*, **360**, 470–480.
- 35  
36  
37 693 39. Churchman, M.L., Low, J., Qu, C., Paietta, E.M., Kasper, L.H., Chang, Y., Payne-Turner, D.,  
38  
39 694 Althoff, M.J., Song, G., Chen, S.-C., *et al.* (2015) EFFICACY OF RETINOIDS IN IKZF1-  
40  
41 695 MUTATED BCR-ABL1 ACUTE LYMPHOBLASTIC LEUKEMIA. *Cancer Cell*, **28**, 343–356.
- 42  
43 696 40. Tingley, D., Yamamoto, T., Hirose, K., Keele, L. and Imai, K. (2014) **mediation** : R Package  
44  
45 697 for Causal Mediation Analysis. *J. Stat. Soft.*, **59**.
- 46  
47 698 41. Imai, K., Keele, L. and Tingley, D. (2010) A general approach to causal mediation analysis.  
48  
49 699 *Psychological Methods*, **15**, 309–334.
- 50  
51 700 42. Metayer, C., Zhang, L., Wiemels, J.L., Bartley, K., Schiffman, J., Ma, X., Aldrich, M.C.,  
52  
53 701 Chang, J.S., Selvin, S., Fu, C.H., *et al.* (2013) Tobacco Smoke Exposure and the Risk of

- 1  
2  
3 702 Childhood Acute Lymphoblastic and Myeloid Leukemias by Cytogenetic Subtype. *Cancer*  
4  
5 703 *Epidemiology Biomarkers & Prevention*, **22**, 1600–1611.  
6  
7 704 43. Gonseth, S., Smith, A.J. de, Roy, R., Zhou, M., Lee, S.-T., Shao, X., Ohja, J., Wrensch,  
8  
9 705 M.R., Walsh, K.M., Metayer, C., *et al.* (2016) Genetic contribution to variation in DNA  
10  
11 706 methylation at maternal smoking-sensitive loci in exposed neonates. *Epigenetics*, **11**, 664–  
12  
13 707 673.  
14  
15 708 44. Xu, K., Li, S., Whitehead, T.P., Pandey, P., Kang, A.Y., Morimoto, L.M., Kogan, S.C.,  
16  
17 709 Metayer, C., Wiemels, J.L. and de Smith, A.J. (2021) Epigenetic Biomarkers of Prenatal  
18  
19 710 Tobacco Smoke Exposure Are Associated with Gene Deletions in Childhood Acute  
20  
21 711 Lymphoblastic Leukemia. *Cancer Epidemiol Biomarkers Prev.*  
22  
23 712 45. Muskens, I.S., Li, S., Jackson, T., Elliot, N., Hansen, H.M., Myint, S.S., Pandey, P., Schraw,  
24  
25 713 J.M., Roy, R., Anguiano, J., *et al.* (2021) The genome-wide impact of trisomy 21 on DNA  
26  
27 714 methylation and its implications for hematopoiesis. *Nature Communications*, **12**, 821.  
28  
29 715 46. Fortin, J.-P., Labbe, A., Lemire, M., Zanke, B.W., Hudson, T.J., Fertig, E.J., Greenwood,  
30  
31 716 C.M. and Hansen, K.D. (2014) Functional normalization of 450k methylation array data  
32  
33 717 improves replication in large cancer studies. *Genome Biol*, **15**.  
34  
35 718 47. Aryee, M.J., Jaffe, A.E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A.P., Hansen, K.D.  
36  
37 719 and Irizarry, R.A. (2014) Minfi: a flexible and comprehensive Bioconductor package for the  
38  
39 720 analysis of Infinium DNA methylation microarrays. *Bioinformatics*, **30**, 1363–1369.  
40  
41 721 48. Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B.S., Bravo,  
42  
43 722 H.C., Davis, S., Gatto, L., Girke, T., *et al.* (2015) Orchestrating high-throughput genomic  
44  
45 723 analysis with Bioconductor. *Nat Methods*, **12**, 115–121.  
46  
47 724 49. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B.,  
48  
49 725 Gautier, L., Ge, Y., Gentry, J., *et al.* (2004) Bioconductor: open software development for  
50  
51 726 computational biology and bioinformatics. *Genome Biology*, **16**.  
52  
53  
54  
55  
56  
57  
58  
59  
60

- 1  
2  
3 727 50. Triche, T.J., Weisenberger, D.J., Van Den Berg, D., Laird, P.W. and Siegmund, K.D. (2013)  
4  
5 728 Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res*,  
6  
7 729 **41**, e90.  
8  
9 730 51. Teschendorff, A.E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D.  
10  
11 731 and Beck, S. (2013) A beta-mixture quantile normalization method for correcting probe  
12  
13 732 design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics*, **29**, 189–196.  
14  
15 733 52. Mah, C.K., Mesirov, J.P. and Chavez, L. (2018) An accessible GenePattern notebook for  
16  
17 734 the copy number variation analysis of Illumina Infinium DNA methylation arrays. *F1000Res*,  
18  
19 735 **7**, 1897.  
20  
21 736 53. Hansen, K. (2016) IlluminaHumanMethylation450kanno. ilmn12. hg19: annotation for  
22  
23 737 Illumina's 450k methylation arrays. *R package version 0.6. 0*, **10**, B9.  
24  
25 738 54. Hansen, K. (2016) IlluminaHumanMethylationEPICanno. ilm10b2. hg19: Annotation for  
26  
27 739 Illumina's EPIC methylation arrays; R package version 0.6. 0. .  
28  
29 740 55. Chen, Y., Lemire, M., Choufani, S., Butcher, D.T., Grafodatskaya, D., Zanke, B.W.,  
30  
31 741 Gallinger, S., Hudson, T.J. and Weksberg, R. (2013) Discovery of cross-reactive probes and  
32  
33 742 polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics*,  
34  
35 743 **8**, 203–209.  
36  
37 744 56. Pidsley, R., Zotenko, E., Peters, T.J., Lawrence, M.G., Risbridger, G.P., Molloy, P., Van  
38  
39 745 Djik, S., Muhlhausler, B., Stirzaker, C. and Clark, S.J. (2016) Critical evaluation of the  
40  
41 746 Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation  
42  
43 747 profiling. *Genome Biology*, **17**, 208.  
44  
45 748 57. McCartney, D.L., Walker, R.M., Morris, S.W., McIntosh, A.M., Porteous, D.J. and Evans,  
46  
47 749 K.L. (2016) Identification of polymorphic and off-target probe binding sites on the Illumina  
48  
49 750 Infinium MethylationEPIC BeadChip. *Genomics Data*, **9**, 22–24.  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



- 1  
2  
3 751 58. Rahmani, E., Zaitlen, N., Baran, Y., Eng, C., Hu, D., Galanter, J., Oh, S., Burchard, E.G.,  
4  
5 752 Eskin, E., Zou, J., *et al.* (2016) Sparse PCA corrects for cell type heterogeneity in  
6  
7 753 epigenome-wide association studies. *Nat. Methods*, **13**, 443–445.  
8  
9 754 59. Rahmani, E., Shenhav, L., Schweiger, R., Yousefi, P., Huen, K., Eskenazi, B., Eng, C.,  
10  
11 755 Huntsman, S., Hu, D., Galanter, J., *et al.* (2017) Genome-wide methylation data mirror  
12  
13 756 ancestry information. *Epigenetics & Chromatin*, **10**, 1.  
14  
15 757 60. Viechtbauer, W. (2010) Conducting meta-analyses in R with the metafor package. *Journal*  
16  
17 758 *of Statistical Software*, **36**, 1–48.  
18  
19 759 61. Yin, L., Zhang, H., Tang, Z., Xu, J., Yin, D., Zhang, Z., Yuan, X., Zhu, M., Zhao, S., Li, X., *et*  
20  
21 760 *al.* (2021) rMVP: A Memory-efficient, Visualization-enhanced, and Parallel-accelerated tool  
22  
23 761 for Genome-Wide Association Study. *Genomics, Proteomics & Bioinformatics*.  
24  
25 762 62. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J.,  
26  
27 763 Sklar, P., de Bakker, P.I.W., Daly, M.J., *et al.* (2007) PLINK: A Tool Set for Whole-Genome  
28  
29 764 Association and Population-Based Linkage Analyses. *Am J Hum Genet*, **81**, 559–575.  
30  
31 765 63. Shabalin, A.A. (2012) Matrix eQTL: ultra fast eQTL analysis via large matrix operations.  
32  
33 766 *Bioinformatics*, **28**, 1353–1358.  
34  
35 767 64. Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W.A., Hou, L. and Lin, S.M. (2010)  
36  
37 768 Comparison of Beta-value and M-value methods for quantifying methylation levels by  
38  
39 769 microarray analysis. *BMC Bioinformatics*, **11**, 587.  
40  
41 770 65. de Smith, A.J., Kaur, M., Gonseth, S., Endicott, A., Selvin, S., Zhang, L., Roy, R., Shao, X.,  
42  
43 771 Hansen, H.M., Kang, A.Y., *et al.* (2017) Correlates of Prenatal and Early-Life Tobacco  
44  
45 772 Smoke Exposure and Frequency of Common Gene Deletions in Childhood Acute  
46  
47 773 Lymphoblastic Leukemia. *Cancer Res*, **77**, 1674–1683.  
48  
49 774 66. Walsh, K.M., de Smith, A.J., Welch, T.C., Smirnov, I., Cunningham, M.J., Ma, X.,  
50  
51 775 Chokkalingam, A.P., Dahl, G.V., Roberts, W., Barcellos, L.F., *et al.* (2014) Genomic  
52  
53  
54  
55  
56  
57  
58  
59  
60



- 1  
2  
3 776 ancestry and somatic alterations correlate with age at diagnosis in Hispanic children with B-  
4  
5 777 cell ALL. *Am J Hematol*, **89**, 721–725.  
6  
7 778 67. Hahne, F. and Ivanek, R. (2016) Visualizing Genomic Data Using Gviz and Bioconductor. In  
8  
9 779 Mathé, E., Davis, S. (eds.), *Statistical Genomics*, Methods in Molecular Biology, Springer  
10  
11 780 New York, New York, NY, Vol. 1418, pp. 335–351.  
12  
13  
14 781  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

For Peer Review

## 782 **Legends to Figures**

783 **Figure 1. Study design of the EWAS, mQTL, and mediation analyses.** Flowcharts show the  
784 inclusion and exclusion criteria for CpGs included in the EWAS and SNP-DMP pairs for the mQTL  
785 and mediation analyses, separately for the overall meta-analysis and the EPIC array only meta-  
786 analysis.

787  
788 **Figure 2. Meta-analysis of the epigenome-wide association analysis.** (A) Bidirectional  
789 Manhattan plot for the overall meta-analysis, (B) QQ plot for the overall meta-analysis, (C)  
790 Bidirectional Manhattan plot for the EPIC array meta-analysis, and (D) QQ plot for the EPIC array  
791 meta-analysis. Two horizontal lines in the Manhattan plot are a Bonferroni-adjusted threshold of  
792 0.025 divided by the number of CpGs (solid line) and a lenient threshold of  $1 \times 10^{-4}$  (dash line). Y-  
793 axis for the Bidirectional Manhattan plot represents  $-\log_{10}P_{hyper}$  for hypermethylated CpGs (higher  
794 DNA methylation beta values in cases vs. in controls) and  $\log_{10}P_{hypo}$  for hypomethylated CpGs  
795 (lower DNA methylation beta values in cases vs. in controls), respectively. CpGs later included in  
796 the causal mediation analysis are labeled with gene names.

797  
798 **Figure 3. Path Diagrams showing the results of the causal mediation and confounding**  
799 **analyses.** (A) The left panel shows the causal mediation model of the *ARID5B* SNP rs7090445  
800 (independent variable), the *ARID5B* CpG cg13344587 (mediator candidate), and ALL risk  
801 (dependent variable). The right panel shows the confounding model of the *ARID5B* SNP  
802 rs7090445 (confounder), the *ARID5B* CpG cg13344587 (independent variable), and ALL risk  
803 (dependent variable). (B) The left and right panels show the causal mediation models of the *IKZF1*  
804 SNP rs78396808 (independent variable), the *IKZF1* CpG cg01139861 (mediator), and ALL risk  
805 (dependent variable), with or without conditioning on the lead *IKZF1* SNP rs10230978 identified  
806 from our recent multi-ancestry GWAS meta-analysis, respectively. The plus and minus signs

1  
2  
3 807 indicate positive correlations and negative correlations, respectively. The solid and dash lines  
4  
5 808 indicate pathways found to be statistically significant and nonsignificant, respectively. The  
6  
7 809 diagrams for the mediation models show the direct effect, the average causal mediation effect,  
8  
9 810 and the total effect estimated from the overall meta-analysis of the causal mediation analysis  
10  
11 811 results from quasi-Bayesian Monte Carlo simulation. The effect estimates correspond to the  
12  
13 812 increased probabilities of developing ALL per 1 copy increase of the SNP risk allele. The diagram  
14  
15 813 for the confounding model shows the crude effect from the logistic regression predicting ALL risk  
16  
17 814 as a function of DNA methylation at the *ARID5B* CpG cg13344587, and the adjusted effect from  
18  
19 815 the logistic regression additionally controlled for the *ARID5B* SNP rs7090445. The effect  
20  
21 816 estimates are the odds ratios for each 0.1 beta-value increase in DNA methylation from the logistic  
22  
23 817 regression models. All Models were adjusting for sex, batch effect, cell type heterogeneity, and  
24  
25 818 genetic ancestry.  
26  
27  
28  
29

30  
31 820 **Figure 4. Characteristics of the significant DMP-SNP pair at *IKZF1*.** (A) Left panels:  
32  
33 821 relationship between DNA methylation level at cg01139861 and rs78396808 genotype. Black  
34  
35 822 points represent median DNA methylation levels. A is the risk allele for rs78396808. Middle panel:  
36  
37 823 relationship between DNA methylation level at cg01139861 and ALL risk. Black points represent  
38  
39 824 median DNA methylation levels. Right panel: rs78396808 genotype frequency in cases and  
40  
41 825 controls overall, in Latinos, and in non-Latino whites. (B) Visualization of the genomic location for  
42  
43 826 DMP cg01139861, CpG cg10551353, and SNP rs78396808 at gene *IKZF1* incorporating  
44  
45 827 annotation queries to UCSC genome browser via Gviz (67). The top track shows the ideogram of  
46  
47 828 chromosome 7 with the black rectangle indicating where gene *IKZF1* is. The second track shows  
48  
49 829 the genome axis, starting from position 50342500 to position 50472798 (reference build Hg19).  
50  
51 830 The third track shows the *IKZF1* gene transcript. The fourth track shows three CpG islands located  
52  
53 831 at gene *IKZF1*. They are at chr7:50342895-50343456 (46 CpGs), chr7:50343757-50344519 (80  
54  
55 832 CpGs), and chr7: 50467566-50468400 (79 CpGs). The last track shows where DMP cg01139861,  
56  
57  
58  
59  
60

1  
2  
3 833 CpG cg10551353, and SNP rs78396808 are located. CpG cg01139861 is located in the CpG  
4  
5 834 island at chr7:50342895-50343456 in the promoter region of *IKZF1*, and SNP rs78396808 is in  
6  
7 835 an intronic region ~116Kb downstream. CpG cg10551353 is in the 5' UTR or the TSS1500 region  
8  
9 836 of several transcripts, ~14Kb downstream of cg01139861.  
10

11  
12 837

13  
14 838 **Figure 5. Scatter plot showing relationship between cg01139861 DNA methylation and**  
15  
16 839 ***IKZF1* expression levels in ALL tumor samples.** Scatter plot with linear regression line and its  
17  
18 840 95% confidence interval band showing a significantly negative correlation between DNA  
19  
20 841 methylation beta values at cg01139861 at gene *IKZF1* and gene expression log<sub>2</sub> fold changes of  
21  
22 842 *IKZF1* in 51 tumor samples from CCLS. The Spearman correlation coefficient R and its p-values  
23  
24 843 are shown in the plot.  
25

26 844

27  
28  
29 845  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 **846 Legends to Tables**  
4

5  
6 **847 Table 1. Characteristics of study participants included in the EWAS stratified by study set**  
7  
8 **848 and ALL case/control status (n = 1,727).**  
9

10  
11  
12 **850 Table 2. The 47 DMPs from the overall EWAS meta-analysis for ALL (808 cases and 919**  
13  
14 **851 controls).**  
15

16 **852**  
17  
18 **853 Table 3. Significant SNP-DMP pairs identified from the methylation quantitative trait loci**  
19  
20 **854 overall meta-analyses (683 cases and 804 controls).**  
21

22 **855**  
23  
24 **856 Table 4. Results from three logistic regression models investigating the potential**  
25  
26 **857 confounding effects.**  
27

28  
29 **858**  
30  
31 **859 Table 5. The total effect, average direct effect and average causal mediation effect**  
32  
33 **860 estimated from the causal mediation analysis (quasi-Bayesian Monte Carlo simulation, n**  
34  
35 **861 = 1000) for two mQTL-DMP pairs, summarized across the CCLS 450K, CCLS EPIC, and**  
36  
37 **862 CCRLP EPIC datasets using the fixed-effect meta-analysis model (683 cases and 804**  
38  
39 **863 controls).**  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

## Tables

**Table 1. Characteristics of study participants included in the EWAS stratified by study set and ALL case/control status (n = 1,727).**

Variables	CCLS 450K (n = 435)			CCRLP EPIC (n = 566)			CCLS EPIC (n = 726)		
	Controls (n = 225)	Cases (n = 210)	P	Controls (n = 436)	Cases (n = 130)	P	Controls (n = 258)	Cases (n = 468)	P
Gestational age (weeks), mean (SD)	39.17 (2.49)	39.33 (2.33)	0.49	39.24 (2.01)	39.16 (2.04)	0.68	39.38 (1.84)	38.99 (2.38)	0.037
Gestational age unknown (%)	6 (3.0)	7 (3.0)		23 (5.0)	3 (2.3)		4 (2.0)	192 (41.0)	
Diagnosis age (years), mean (SD)		5.21 (3.45)			3.48 (1.74)			5.70 (3.57)	
Diagnosis age unknown (%)					1 (0.8)				
Sex (%)									
Males	130 (57.8)	121 (57.6)	1.00	257 (58.9)	74 (56.9)	0.76	148 (57.4)	263 (56.2)	0.82
Females	95 (42.2)	89 (42.4)		179 (41.1)	56 (43.1)		110 (42.6)	205 (43.8)	
Race/ethnicity (%)									
Latino	107 (47.6)	109 (51.9)	0.66	251 (57.6)	69 (53.1)	0.62	136 (52.7)	176 (54.8)	0.36
Non-Latino other	38 (16.9)	33 (15.7)		62 (14.2)	22 (16.9)		32 (12.4)	49 (15.3)	
Non-Latino white	80 (35.6)	68 (32.4)		123 (28.2)	39 (30.0)		90 (34.9)	96 (29.9)	
Race/ethnicity unknown (%)								147 (31.4)	

P-values comparing the characteristics of ALL cases and controls in the CCLS 450K, CCLS EPIC, and CCRLP EPIC datasets were calculated using t-tests for the continuous variable (gestational age) and Chi-squared tests for categorical variables.

3.

**Table 2. The 47 DMPs from the overall EWAS meta-analysis for ALL (808 cases and 919 controls).**

Probe	Chr	Pos	Islands Name	Relation to Island	UCSC RefGene Name	UCSC RefGene Group	CCLS 450K		CCRLP EPIC		CCLS EPIC		Meta-analysis				
							Coef	SE	Coef	SE	Coef	SE	Coef	SE	P	P.het	i.squared
cg13344587*	chr10	63723919		OpenSea	ARID5B	Body	-6.05	2.14	-9.39	2.42	-7.08	1.79	-7.32	1.19	8.61E-10	0.58	0.00
cg19961720	chr22	17309849		OpenSea	HSFYF1	Body	23.21	6.50	28.95	12.82	18.98	9.26	22.86	4.91	3.29E-06	0.82	0.00
cg19783404	chr11	67751269		OpenSea			-10.47	3.50	-7.08	4.46	-10.96	3.44	-9.87	2.15	4.32E-06	0.77	0.00
cg20148881	chr19	34112229	chr19:34112279-34114353	N_Shore	CHST8; CHST8	TSS1500;TS S1500	-11.84	3.23	-7.08	3.63	-5.73	2.56	-7.86	1.75	7.54E-06	0.32	11.64
cg14230238	chr19	6066871		OpenSea	RFX2;RFX2	5'UTR;5'UTR	-10.09	2.99	-5.57	3.38	-5.83	2.36	-7.03	1.62	1.51E-05	0.47	0.00
cg18872749	chr17	42421993		OpenSea	GRN	TSS1500	-4.69	2.02	-6.80	2.44	-4.42	1.79	-5.06	1.18	1.64E-05	0.72	0.00
cg04036329	chr17	19771783	chr17:19771609-19771814	Island	ULK2;ULK2	TSS1500;TS S1500	-4.37	2.23	-4.83	2.47	-5.80	1.78	-5.14	1.21	2.18E-05	0.87	0.00
cg18068140	chr6	18123164	chr6:18122250-18122994	S_Shore	NHLRC1	TSS1500	-3.21	2.29	-10.80	2.59	-3.78	1.75	-5.19	1.22	2.25E-05	0.05	67.21
cg01619045	chr19	21948866	chr19:21949903-21950217	N_Shore	ZNF100	Body	18.09	5.75	38.17	22.53	19.80	8.28	19.47	4.62	2.56E-05	0.69	0.00
cg05493394	chr6	133562035	chr6:133562086-133563586	N_Shore	EYA4;EYA4;EYA4	TSS1500;TS S1500;TSS1500	-39.10	17.21	-61.49	26.24	-46.59	16.95	-46.15	10.97	2.59E-05	0.77	0.00
cg14623989	chr18	10456115	chr18:10454082-10454296	S_Shore	APCDD1	Body	-5.48	2.99	-5.19	3.41	-8.57	2.39	-6.86	1.64	2.79E-05	0.62	0.00
cg17384056	chr19	47742829	chr19:47742725-	Island			-68.50	18.49	-34.26	99.84	-95.47	47.11	-71.00	16.96	2.84E-05	0.81	0.00

			477431 51															
cg18708 233	chr 10	744515 49	chr10:7 445182 8- 744525 97	N_Sho re	CCDC1 09A	TSS15 00	-8.12	2.11	-2.64	1.96	-3.54	1.57	-4.43	1.06	2.84E-05	0.12	52.72	
cg21727 223	chr 17	667557 70		OpenS ea			4.08	2.16	5.45	2.55	5.11	1.68	4.88	1.18	3.29E-05	0.90	0.00	
cg20451 680	chr 5	542813 36		OpenS ea	ESM1;E SM1	1stExo n;1stE xon	-6.20	1.87	-1.15	2.22	-5.21	1.72	-4.56	1.10	3.52E-05	0.20	38.80	
cg23293 256	chr 10	116310 72		OpenS ea	USP6N L	Body	6.86	2.48	2.77	7.36	9.05	2.88	7.48	1.82	3.99E-05	0.68	0.00	
cg03293 350	chr 18	528910 17		OpenS ea	TCF4;T CF4	3'UTR; 3'UTR	7.18	2.93	-3.07	4.75	6.58	1.75	5.84	1.43	4.40E-05	0.14	48.74	
cg15491 120	chr 18	741568 97	chr18:7 415323 9- 741550 73	S_Sho re	ZNF516	5'UTR	23.90	7.49	9.68	9.13	16.26	6.20	17.28	4.23	4.42E-05	0.47	0.00	
cg20572 153	chr 17	992975 5		OpenS ea	GAS7; GAS7; GAS7	Body; Body; TSS20 0	-6.01	3.23	-5.07	3.54	-9.99	2.81	-7.43	1.82	4.47E-05	0.48	0.00	
cg16276 850	chr 17	384989 14	chr17:3 849752 7- 384989 63	Island	RARA; RARA; RARA; RARA; RARA	5'UTR; 1stExo n;Body ;Body; Body	7.86	3.54	12.95	4.32	6.38	3.09	8.36	2.05	4.59E-05	0.46	0.00	
cg08943 714	chr 6	139470 282		OpenS ea	HECA	Body	4.58	3.02	13.84	6.97	13.89	3.62	8.94	2.20	4.78E-05	0.11	55.12	
cg15260 921	chr 1	226308 960	chr1:22 630895 9- 226310 476	Island			-8.27	3.46	-11.46	5.15	-8.67	3.54	-9.03	2.23	5.18E-05	0.87	0.00	
cg04035 597	chr 8	197945 39	chr8:19 796843- 197980 06	N_She lf			-2.73	1.80	-4.05	2.33	-6.30	1.74	-4.46	1.10	5.18E-05	0.35	3.51	
cg27579 121	chr 6	418596 17	chr6:41 862841- 418633 35	N_She lf	USP49	5'UTR	4.02	2.15	2.28	5.26	6.79	1.81	5.42	1.34	5.19E-05	0.51	0.00	
cg01139 861*	chr 7	503432 98	chr7:50 342895- 503434 56	Island	IKZF1	TSS15 00	8.39	2.11	4.43	1.86	1.81	1.46	4.09	1.01	5.20E-05	0.04	69.80	



cg01778647	chr8	141473870	chr8:141474418-141475050	N_Shore			-7.11	3.42	-13.87	4.68	-7.32	3.34	-8.60	2.13	5.32E-05	0.45	0.00
cg10906284	chr12	63544430	chr12:63543636-63544967	Island	AVPR1A	1stExon	23.76	7.65	29.45	16.09	30.22	16.26	25.64	6.36	5.53E-05	0.91	0.00
cg06646708	chr6	18123224	chr6:1812250-18122994	S_Shore	NHLRC1	TSS1500	-2.47	2.15	-9.31	2.64	-4.85	1.91	-5.05	1.26	5.83E-05	0.13	50.71
cg02802029	chr3	145879686	chr3:145878430-145879287	S_Shore	PLOD2; PLOD2	TSS1500;TSS1500	-3.69	1.63	-3.56	1.50	-2.57	1.06	-3.07	0.77	5.86E-05	0.79	0.00
cg05029558	chr8	71581784	chr8:71581050-71581650	S_Shore	XKR9;LACTB2	5'UTR;TSS1500	-2.72	2.70	-10.41	3.01	-6.10	2.36	-6.14	1.53	6.10E-05	0.16	44.81
cg08748969	chr12	69327779	chr12:69327021-69327532	S_Shore	CPM;CPM;CPM	TSS1500;TSS1500;5'UTR	-4.97	2.17	-4.12	3.06	-6.31	2.06	-5.38	1.34	6.22E-05	0.82	0.00
cg19752094	chr11	76381800	chr11:76381449-76382295	Island	LRRC32;LRRC32	TSS1500;TSS200	-32.58	8.33	3.76	64.00	-33.90	34.94	-32.07	8.03	6.55E-05	0.85	0.00
cg03172991	chr19	13105728	chr19:13106817-13107688	N_Shore	NFIX	TSS1500	-18.38	4.87	-5.35	5.93	-8.39	4.04	-10.93	2.75	7.22E-05	0.16	44.82
cg06328724	chr19	37958752	chr19:37959852-37960615	N_Shore	ZNF570;ZNF569	TSS1500;TSS1500	-3.98	2.15	-7.16	3.00	-5.95	2.19	-5.41	1.36	7.42E-05	0.65	0.00
cg17554636	chr10	45719751	chr10:45719712-45720203	Island			-13.12	4.91	-21.92	8.85	-22.14	11.94	-15.99	4.04	7.51E-05	0.59	0.00

cg10475 928	chr 10	118901 190	chr10:1 188992 47- 118900 329	S_Shore			-4.71	1.89	-4.04	2.17	-3.88	1.57	-4.17	1.05	7.53E-05	0.94	0.00
cg25412 453	chr 1	155006 219		OpenSea	DCST1; DCST2; DCST2; DCST1	TSS200; 5'UTR; R;1stExon; TSS200	10.07	5.39	12.91	7.79	14.72	4.70	12.74	3.22	7.75E-05	0.81	0.00
cg15721 728	chr 11	662514 1	chr11:6 624552- 662507 3	S_Shore	ILK;ILK; RRP8;ILK	5'UTR; 5'UTR; TSS1500; TS S200	21.82	9.70	28.82	10.58	16.55	8.33	21.42	5.43	7.87E-05	0.66	0.00
cg04247 829	chr 1	382180 80	chr1:38 218190- 382189 77	N_Shore	EPHA10	Body	-3.25	2.89	-7.71	2.75	-5.50	1.98	-5.54	1.41	8.10E-05	0.54	0.00
cg13185 177	chr 3	194119 885	chr3:19 411760 1- 194118 988	S_Shore	GP5	5'UTR	-6.50	2.60	-2.77	2.53	-5.98	1.95	-5.23	1.33	8.24E-05	0.51	0.00
cg14362 370	chr 9	133589 654	chr9:13 358786 5- 133588 901	S_Shore	ABL1;ABL1	5'UTR; 1stExon	-3.49	3.03	-9.24	3.20	-5.92	2.17	-6.07	1.55	8.70E-05	0.42	0.00
cg19836 174	chr 5	149921 140		OpenSea	NDST1	Body	12.48	6.59	24.76	14.94	32.63	9.30	19.85	5.06	8.73E-05	0.20	38.44
cg13230 172	chr 11	120195 828	chr11:1 201958 05- 120196 372	Island	TMEM136	TSS200	60.07	17.25	71.65	42.56	33.81	33.54	56.54	14.43	8.91E-05	0.73	0.00
cg09180 564	chr 1	102308 808		OpenSea	OLFM3	Body	7.30	3.49	9.78	3.06	4.14	2.46	6.57	1.68	9.12E-05	0.35	5.45
cg23940 023	chr 14	101521 703		OpenSea	MIR485; MIR453	TSS200; TSS1500	15.66	7.45	11.71	13.37	29.80	8.55	20.25	5.18	9.21E-05	0.36	1.65
cg16728 651	chr 6	557678 65		OpenSea			3.21	2.18	6.84	2.21	3.48	1.50	4.22	1.08	9.56E-05	0.40	0.00
cg16969 852	chr 9	111775 933	chr9:11 177504 1- 111775 934	Island	CTNNA1	TSS200	57.65	16.19	79.49	31.91	11.79	22.19	47.15	12.10	9.77E-05	0.14	49.83

\*Two DMPs found to be associated with ALL risk SNPs in the subsequent mQTL analysis.

4.

1  
2  
3 All logistic regression models were adjusted for sex, batch effect, cell type heterogeneity using the first ten principal components  
4 derived from ReFACTor, and genetic ancestry using the first ten principal components derived from EPISTRUCTURE.  
5 Study heterogeneity was characterized with  $I^2$  statistics (i.squared) and their corresponding p-values (*P.het*).  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47

For Peer Review

**Table 3. Significant SNP-DMP pairs identified from the methylation quantitative trait loci overall meta-analyses (683 cases and 804 controls).**

mQTL-DMP pair	CCLS 450K		CCRLP EPIC		CCLS EPIC		Meta-analysis					
	Coef	SE	Coef	SE	Coef	SE	Coef	SE	<i>P</i>	<i>P.het</i>	i.squared	FDR
rs7090445 (chr10:63721176 at <i>ARID5B</i> ) and cg13344587 (chr10:63723919 at <i>ARID5B</i> )	-0.047	0.002	-0.052	0.002	-0.050	0.002	-0.050	0.001	<2.23E-308	0.22	34.40	<2.23E-308
rs78396808 (chr7:50459043 at <i>IKZF1</i> ) and cg01139861 (chr7:50343298 at <i>IKZF1</i> )	0.018	0.005	0.025	0.005	0.026	0.005	0.024	0.003	3.25E-17	0.51	0.00	1.75E-14

No SNP-DMP pairs passed the Bonferroni correction or FDR threshold from the mQTL EPIC array meta-analysis (479 cases and 636 controls).

All linear regression models were adjusted for sex, batch effect, cell type heterogeneity using the first ten principal components derived from ReFACTor, and genetic ancestry using the first ten principal components derived from EPISTRUCTURE. Study heterogeneity was characterized with  $I^2$  statistics (i.squared) and their corresponding p-values (*P.het*).

**Table 4. Results from three logistic regression models investigating the potential confounding effects.**

Variable	CCLS 450K	CCRLP EPIC	CCLS EPIC	Meta-analysis			
	OR (CI)	OR (CI)	OR (CI)	OR (CI)	<i>P</i>	<i>P.het</i>	i.squared
<b>rs7090445(ARID5B)-cg13344587(ARID5B)-ALL</b>							
<b>Model 1: logistic regression model predicting ALL status as a function of methylation at cg13344587, adjusting for sex, batch effect, cell type heterogeneity, and genetic ancestry (808 cases and 919 controls)</b>							
cg13344587	0.47 (0.29-0.74)	0.38 (0.23-0.61)	0.46 (0.31-0.69)	0.44 (0.34-0.57)	2.87E-10	0.77	0.00
<b>Model 2: model 1 additionally adjusted for SNP rs7090445 (808 cases and 919 controls)</b>							
cg13344587	0.64 (0.32-1.28)	0.59 (0.26-1.35)	1.05 (0.61-1.83)	0.80 (0.54-1.17)	2.49E-01	0.39	0.00
rs7090445*	1.30 (0.84-2.02)	1.42 (0.84-2.40)	2.25 (1.51-3.35)	1.67 (1.29-2.16)	8.52E-05	0.15	46.87
<b>Model 3: model 1 in subjects without any copies of the risk allele of SNP rs7090445 (154 cases and 292 controls)</b>							
cg13344587	1.26 (0.24-6.65)	0.46 (0.05-4.20)	1.05 (0.41-2.67)	0.99 (0.46-2.12)	9.70E-01	0.76	0.00
<b>(ORadj - ORcrude)/ORcrude = (0.44-0.80)/0.44 = -82%</b>							
<b>rs78396808(IKZF1)-cg01139861(IKZF1)-ALL</b>							
<b>Model 1: logistic regression model predicting ALL status as a function of methylation at cg01139861, adjusting for sex, batch effect, cell type heterogeneity, and genetic ancestry (808 cases and 919 controls)</b>							
cg01139861	2.17 (1.39-3.39)	1.51 (1.04-2.19)	1.24 (0.89-1.72)	1.51 (1.22-1.87)	1.75E-04	0.13	50.13
<b>Model 2: model 1 additionally adjusted for SNP rs78396808 (808 cases and 919 controls)</b>							
cg01139861	2.09 (1.33-3.28)	1.44 (0.99-2.11)	1.18 (0.84-1.66)	1.45 (1.16-1.81)	9.49E-04	0.14	48.88
rs78396808*	1.23 (0.79-1.91)	1.26 (0.84-1.89)	1.21 (0.83-1.76)	1.23 (0.98-1.56)	7.82E-02	0.99	0.00
<b>Model 3: model 1 in subjects without any copies of the risk allele of SNP rs78396808 (490 cases and 592 controls)</b>							
cg01139861	2.26 (1.33-3.84)	1.58 (1.02-2.45)	1.08 (0.73-1.59)	1.45 (1.13-1.88)	4.09E-03	0.08	60.29
<b>(ORadj - ORcrude)/ORcrude = (1.51-1.45)/1.51= 4%</b>							

OR: odds ratio; CI: 95% confidence interval.

ORs were calculated for every 0.1 CpG beta value increase or for every 1 copy increase of the SNP risk allele.

ORcrude is the OR from the model 1 meta-analysis and ORadj is the OR from the model 2 meta-analysis.

Study heterogeneity was characterized with  $I^2$  statistics (i.squared) and their corresponding p-values (*P.het*).

\* The effect of rs7090445 on ALL adjusted for sex, and genetic ancestry: OR=1.87 (95% CI:1.58, 2.20)  $P=2.46E-13$  (683 cases and 804 controls). The effect of rs78396808 on ALL adjusted for sex, and genetic ancestry: OR=1.31 (95% CI:1.05,1.64)  $P=0.017$  (683 cases and 804 controls).

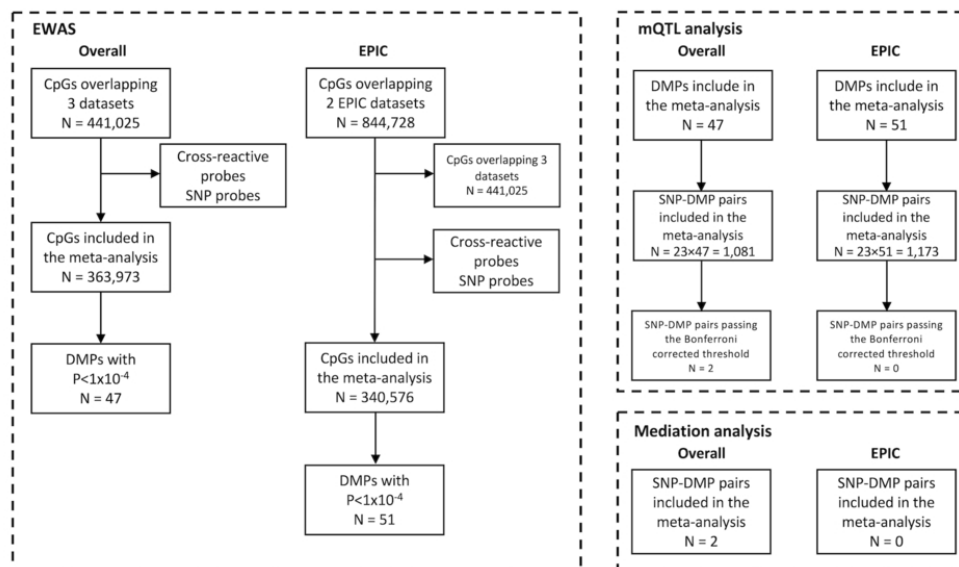
**Table 5. The total effect, average direct effect and average causal mediation effect estimated from the causal mediation analysis (quasi-Bayesian Monte Carlo simulation,  $n = 1000$ ) for two mQTL-DMP pairs, summarized across the CCLS 450K, CCLS EPIC, and CCRLP EPIC datasets using the fixed-effect meta-analysis model (683 cases and 804 controls).**

mQTL-DMP pair	Estimate	SE	Statistic	<i>P</i>	<i>P.het</i>	i.squared	Effect
rs7090445 (chr10:63721176 at <i>ARID5B</i> ) and cg13344587 (chr10:63723919 at <i>ARID5B</i> )	0.108	0.014	7.82	<b>5.17E-15</b>	0.060	64.43	total
	0.023	0.019	1.22	2.23E-01	0.455	0.00	ACME
	0.093	0.024	3.87	<b>1.06E-04</b>	0.054	65.78	ADE
rs78396808 (chr7:50459043 at <i>IKZF1</i> ) and cg01139861 (chr7:50343298 at <i>IKZF1</i> )	0.056	0.022	2.56	<b>1.05E-02</b>	0.918	0.00	total
	0.017	0.006	2.96	<b>3.09E-03</b>	0.435	0.00	ACME
	0.039	0.022	1.77	7.66E-02	0.997	0.00	ADE
rs78396808 (chr7:50459043 at <i>IKZF1</i> ) and cg01139861 (chr7:50343298 at <i>IKZF1</i> )*	0.089	0.023	3.84	<b>1.24E-04</b>	0.704	0.00	total
	0.016	0.006	2.73	<b>6.28E-03</b>	0.435	0.00	ACME
	0.073	0.024	3.08	<b>2.04E-03</b>	0.871	0.00	ADE

\*Models additionally adjusted for *IKZF1* SNP rs10230978 (top *IKZF1* SNP association in multi-ancestry ALL GWAS).

The effect estimates correspond to the increased probabilities of developing ALL per 1 copy increase of the SNP risk allele.

Study heterogeneity was characterized with  $I^2$  statistics (i.squared) and their corresponding p-values (*P.het*).



**Fig. 1**

Figure 1. Study design of the EWAS, mQTL, and mediation analyses. Flowcharts show the inclusion and exclusion criteria for CpGs included in the EWAS and SNP-DMP pairs for the mQTL and mediation analyses, separately for the overall meta-analysis and the EPIC array only meta-analysis.

81x51mm (300 x 300 DPI)



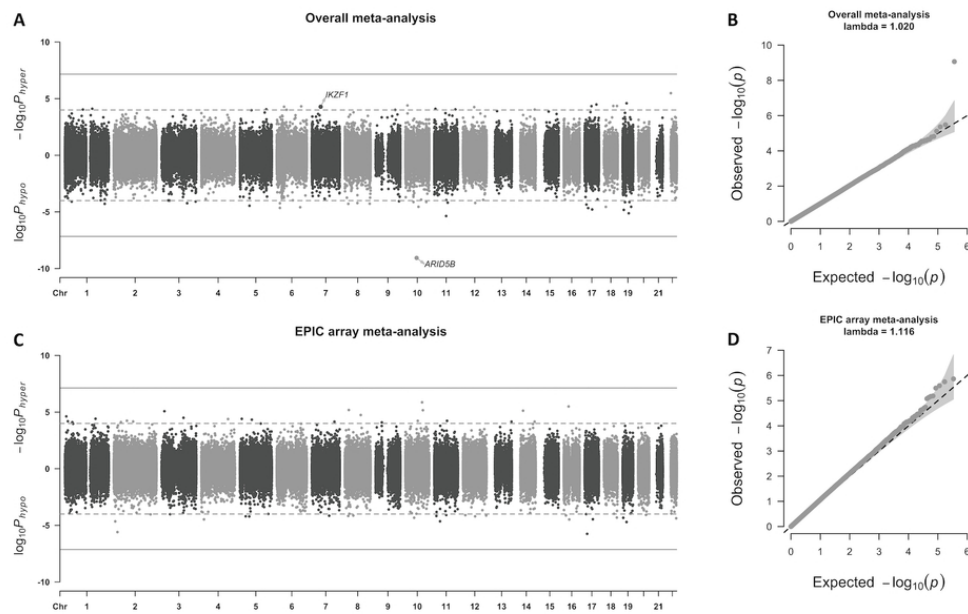
**Fig. 2**

Figure 2. Meta-analysis of the epigenome-wide association analysis. (A) Bidirectional Manhattan plot for the overall meta-analysis, (B) QQ plot for the overall meta-analysis, (C) Bidirectional Manhattan plot for the EPIC array meta-analysis, and (D) QQ plot for the EPIC array meta-analysis. Two horizontal lines in the Manhattan plot are a Bonferroni-adjusted threshold of 0.025 divided by the number of CpGs (solid line) and a lenient threshold of  $1 \times 10^{-4}$  (dash line). Y-axis for the Bidirectional Manhattan plot represents  $-\log_{10}P_{\text{hyper}}$  for hypermethylated CpGs (higher DNA methylation beta values in cases vs. in controls) and  $\log_{10}P_{\text{hypo}}$  for hypomethylated CpGs (lower DNA methylation beta values in cases vs. in controls), respectively. CpGs later included in the causal mediation analysis are labeled with gene names.

80x54mm (300 x 300 DPI)

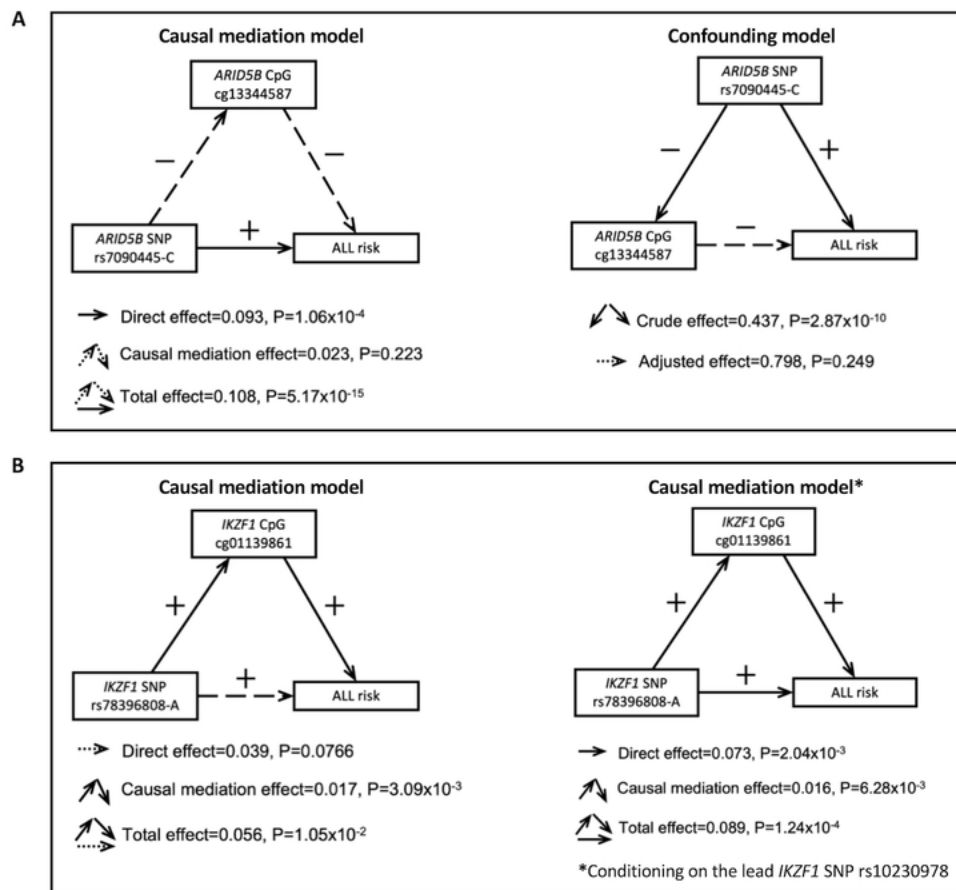


Fig. 3

Figure 3. Path Diagrams showing the results of the causal mediation and confounding analyses. (A) The left panel shows the causal mediation model of the *ARID5B* SNP rs7090445 (independent variable), the *ARID5B* CpG cg13344587 (mediator candidate), and ALL risk (dependent variable). The right panel shows the confounding model of the *ARID5B* SNP rs7090445 (confounder), the *ARID5B* CpG cg13344587 (independent variable), and ALL risk (dependent variable). (B) The left and right panels show the causal mediation models of the *IKZF1* SNP rs78396808 (independent variable), the *IKZF1* CpG cg01139861 (mediator), and ALL risk (dependent variable), with or without conditioning on the lead *IKZF1* SNP rs10230978 identified from our recent multi-ancestry GWAS meta-analysis, respectively. The plus and minus signs indicate positive correlations and negative correlations, respectively. The solid and dash lines indicate pathways found to be statistically significant and nonsignificant, respectively. The diagrams for the mediation models show the direct effect, the average causal mediation effect, and the total effect estimated from the overall meta-analysis of the causal mediation analysis results from quasi-Bayesian Monte Carlo simulation. The effect estimates correspond to the increased probabilities of developing ALL per 1 copy increase of the SNP risk allele. The diagram for the confounding model shows the crude effect from the logistic regression predicting ALL risk as a function of DNA methylation at the *ARID5B* CpG cg13344587, and the adjusted effect from the logistic regression additionally controlled for the *ARID5B* SNP rs7090445. The effect estimates are the odds ratios for each 0.1 beta-value increase in DNA methylation from the logistic regression models. All Models were adjusting for sex, batch effect, cell type heterogeneity, and genetic ancestry.

63x62mm (300 x 300 DPI)

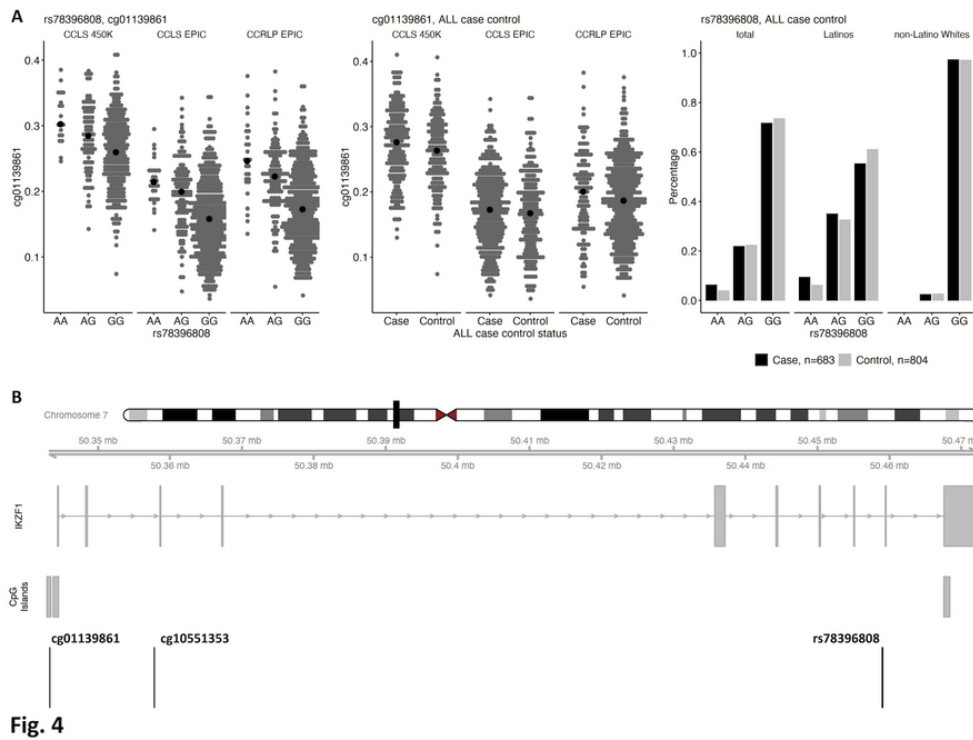


Fig. 4

Figure 4. Characteristics of the significant DMP-SNP pair at IKZF1. (A) Left panels: relationship between DNA methylation level at cg01139861 and rs78396808 genotype. Black points represent median DNA methylation levels. A is the risk allele for rs78396808. Middle panel: relationship between DNA methylation level at cg01139861 and ALL risk. Black points represent median DNA methylation levels. Right panel: rs78396808 genotype frequency in cases and controls overall, in Latinos, and in non-Latino whites. (B) Visualization of the genomic location for DMP cg01139861, CpG cg10551353, and SNP rs78396808 at gene IKZF1 incorporating annotation queries to UCSC genome browser via Gviz (67). The top track shows the ideogram of chromosome 7 with the black rectangle indicating where gene IKZF1 is. The second track shows the genome axis, starting from position 50342500 to position 50472798 (reference build Hg19). The third track shows the IKZF1 gene transcript. The fourth track shows three CpG islands located at gene IKZF1. They are at chr7:50342895-50343456 (46 CpGs), chr7:50343757-50344519 (80 CpGs), and chr7:50467566-50468400 (79 CpGs). The last track shows where DMP cg01139861, CpG cg10551353, and SNP rs78396808 are located. CpG cg01139861 is located in the CpG island at chr7:50342895-50343456 in the promoter region of IKZF1, and SNP rs78396808 is in an intronic region ~116Kb downstream. CpG cg10551353 is in the 5' UTR or the TSS1500 region of several transcripts, ~14Kb downstream of cg01139861.

76x58mm (300 x 300 DPI)

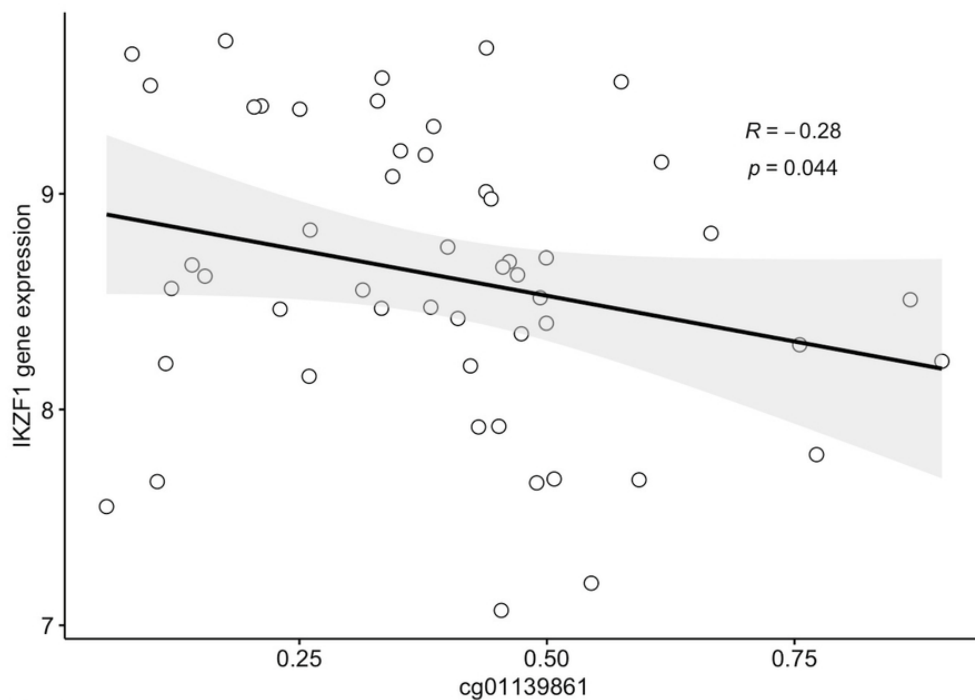


Fig. 5

Figure 5. Scatter plot showing relationship between cg01139861 DNA methylation and IKZF1 expression levels in ALL tumor samples. Scatter plot with linear regression line and its 95% confidence interval band showing a significantly negative correlation between DNA methylation beta values at cg01139861 at gene IKZF1 and gene expression log2 fold changes of IKZF1 in 51 tumor samples from CCLS. The Spearman correlation coefficient  $R$  and its  $p$ -values are shown in the plot.

76x56mm (300 x 300 DPI)