Title

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

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

Investigating DNA Methylation as a Mediator of Genetic Risk in Childhood Acute Lymphoblastic Leukemia

Keren Xu[1], Shaobo Li[1], Priyatama Pandey[1], Alice Y. Kang[2], Libby M. Morimoto[2], Nicholas Mancuso[1], Xiaomei Ma[3], Catherine Metayer[2], Joseph L. Wiemels[1], Adam J. de Smith[1*]

[1] Center for Genetic Epidemiology, Department of Population and Public Health Sciences, University of Southern California, Los Angeles, CA 90033, USA

[2] School of Public Health, University of California, Berkeley, Berkeley, CA 94704, USA

[3] Department of Chronic Disease Epidemiology, Yale School of Public Health, New Haven, CT 06510, USA

***Corresponding Author:** Adam J. de Smith; Address: 1450 Biggy St., NRT-1509H, USC Norris Comprehensive Cancer Center, Los Angeles, CA 90033; Email: adam.desmith@med.usc.edu

Abstract

Genome-wide association studies have identified a growing number of single nucleotide polymorphisms (SNPs) associated with childhood acute lymphoblastic leukemia (ALL), yet the functional roles of most SNPs are unclear. Multiple lines of evidence suggest epigenetic mechanisms may mediate the impact of heritable genetic variation on phenotypes. Here, we investigated whether DNA methylation mediates the effect of genetic risk loci for childhood ALL. We performed an epigenome-wide association study (EWAS) including 808 childhood ALL cases and 919 controls from California-based studies using neonatal blood DNA. For differentially methylated CpG positions (DMPs), we next conducted association analysis with 23 known ALL risk SNPs followed by causal mediation analyses addressing the significant SNP-DMP pairs. DNA methylation at CpG cg01139861, in the promoter region of *IKZF1*, mediated the effects of the intronic *IKZF1* risk SNP rs78396808, with the average causal mediation effect (ACME) explaining ~30% of the total effect (ACME P=0.0031). In analyses stratified by self-reported race/ethnicity, the mediation effect was only significant in Latinos, explaining ~41% of the total effect of rs78396808 on ALL risk (ACME P=0.0037). Conditional analyses confirmed the presence of at least three independent genetic risk loci for childhood ALL at *IKZF1*, with rs78396808 unique to non-European populations. We also demonstrated that the most significant DMP in the EWAS, CpG cg13344587 at gene *ARID5B* (P=8.61x10^{-10}), was entirely confounded by the *ARID5B* ALL risk SNP rs7090445. Our findings provide new insights into the functional pathways of ALL risk SNPs and the DNA methylation differences associated with risk of childhood ALL.

Introduction

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

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

Epigenome-wide association studies (EWASs) testing DNA methylation differences between childhood ALL cases and controls at birth may also pinpoint differentially methylated CpG positions (DMPs) involved in the development of childhood ALL, yet studies conducted so far have mainly focused on DNA methylation changes in diagnostic leukemia samples (22–25). Here, we performed an EWAS of ALL including 808 childhood ALL cases and 919 controls from two ancestrally diverse independent California-based studies, the California Childhood Leukemia Study (CCLS) and the Childhood Cancer Records Linkage Project (CCRLP), to identify ALL-associated DMPs, and then tested the association of significant DMPs with known ALL risk SNPs and assessed whether DNA methylation at these CpG probes may mediate the effects of the genetic risk loci.

Results

There were 850 ALL cases and 931 cancer-free controls from the CCLS and the CCRLP that had DNA samples from neonatal dried bloodspot (DBS) assayed on either the Illumina® HumanMethylation450 BeadChip (450K) DNA methylation arrays or Illumina® Infinium MethylationEPIC BeadChip (EPIC) arrays. After excluding subjects with trisomy 21 (27 cases, 1 control), we included in the EWAS a total of 808 childhood ALL cases and 919 cancer-free controls that passed DNA methylation quality control (see **Materials and Methods**). Demographic characteristics of these subjects are summarized in **Table 1**, and the study design is illustrated in **Figure 1**. Over half of the study participants were males, and overall 53.7% were self-reported Latino and 31.4% were non-Latino white, with approximately equal distributions among cases and controls across the CCLS 450K, CCLS EPIC, and CCRLP EPIC datasets. Demographic characteristics of the subset of 683 childhood ALL cases and 804 controls included in the methylation quantitative trait loci (mQTL) and mediation analyses were similar to the overall dataset (**Table S1**).

Differentially methylated positions (DMPs)

To identify ALL-associated DMPs on autosomal chromosomes, we first performed EWAS analyses separately in the CCLS 450K, CCLS EPIC, and CCRLP EPIC datasets. A total of 363,973 CpGs were included in the overall EWAS fixed-effect meta-analysis (CCLS 450K, CCLS EPIC, and CCRLP EPIC datasets) (**Figure 1**). An additional 340,576 CpGs not included in the CCLS 450K dataset were analyzed in the EPIC array meta-analysis (CCLS EPIC and CCRLP EPIC datasets) (**Figure 1**). The number of CpGs meeting each probe filtering criterion is summarized in **Table S2**. CpG cg13344587, in an intronic region of the ALL risk gene *ARID5B*, was significantly differentially methylated by ALL case/control status (P=8.61x10^{-10}) after adjusting for multiple testing using a stringent Bonferroni correction in the overall meta-analysis (P threshold: 0.025/363,973=6.87x10^{-8}) (**Figures 2A and 2B**). No additional CpGs reached Bonferroni significance in the EPIC array analysis (**Figures 2C and 2D**), nor survived the Benjamini-Hochberg false discovery rate (FDR) correction (FDR<0.05) in either study. Using a less stringent threshold of P<1x10^{-4} for the purposes of identifying candidate CpGs for downstream mediation analyses (26,27), we found 47 DMPs for ALL overlapping both 450K and EPIC arrays from the overall meta-analysis and 51 DMPs from the EPIC array meta-analysis (**Figures 2A and 2C; Tables 2 and S3**). The effect estimates of these CpGs in overall participants were in strong correlations with those in self-reported Latinos and non-Latino whites in stratified EWAS (**Figure S1**) and in EWAS adjusting for genetic ancestry using principal components derived from SNP array data instead of from EPISTRUCTURE, in the subset of subjects with both DNA methylation and SNP genotype data available (**Figure S1; Tables S4 and S5**).

Methylation quantitative trait loci (mQTL)

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

include 4 secondary associations discovered in conditional analysis adjusting for the lead SNP at *IKZF1*, *CDKN2A*, *CEBPE*, and *IKZF3*. They were analyzed for association with the 47 DMPs separately in the CCLS 450K, CCLS EPIC, and CCRLP EPIC datasets, and with the 51 DMPs separately in the two EPIC datasets (**Figure 1**). Two SNP-DMP pairs *in cis* passed the Bonferroni corrected threshold in the meta-analysis of three datasets ($P<2.31\times10^{-5}$ [0.025/(23×47)]): 1) SNP rs7090445 and DMP cg13344587 at gene *ARID5B*, and 2) SNP rs78396808 and DMP cg01139861 at gene *IKZF1* (**Tables 2 and 3**). No SNP-DMP associations survived the Bonferroni corrected threshold ($P<2.13\times10^{-5}$ [0.025/(23×51)]) or the FDR correction (FDR<0.05) in the EPIC array meta-analysis. The two significant SNP-DMP pairs identified from the overall meta-analysis remained significant in mQTL analysis using DNA methylation M-values (data not shown).

In multivariable conditional analysis to assess the effect of the SNPs on ALL risk while adjusting for the corresponding CpG in the SNP-DMP pair, and vice-versa, the *ARID5B* SNP rs7090445 remained significant ($P=8.52\times10^{-5}$) (model 2 in **Table 4**), indicating that rs7090445 had a direct impact on ALL risk, independent of the *ARID5B CpG* cg13344587, and that cg13344587 could potentially partially mediate the effect of rs7090445 on ALL risk; however, cg13344587 was no longer significant ($P=0.249$), indicating that DNA methylation at cg13344587 was entirely driven by rs7090445. In contrast, the *IKZF1* SNP rs78396808 was no longer significant in the multivariable conditional model ($P=0.078$) but the *IKZF1* CpG cg01139861 remained significant ($P=9.49\times10^{-4}$), suggesting that the effect of rs78396808 on ALL risk may be mediated by cg01139861 and that DNA methylation at cg01139861 may be affected by risk factors other than rs78396808. We next formally tested whether there were significant mediation effects for these two SNP-DMP pairs.

Mediation analysis

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

1). The total effect, average direct effect (ADE) and average causal mediation effect (ACME) estimated for each mQTL-DMP pair were summarized across all three datasets using the fixed-effect meta-analysis model.

SNP rs7090445 at gene *ARID5B* had a significant total effect (estimate=0.108, P=5.17x10^{-15}) and direct effect (ADE estimate=0.093, P=1.06x10^{-4}) but a nonsignificant causal mediation effect (ACME P=0.223) on ALL risk through altering DNA methylation at the *ARID5B* CpG cg13344587 (**Table 5 and Figure 3A**), demonstrating that the effect of rs7090445 on ALL risk was independent of cg13344587.

In contrast, we found that *IKZF1* SNP rs78396808 had a significant total effect on ALL risk (estimate=0.056, P=0.010), a significant mediation effect through increasing DNA methylation at the *IKZF1* CpG cg01139861 (ACME estimate=0.017, P=0.003) and a nonsignificant direct effect (estimate=0.039, P=0.077), with the ACME explaining ~30% (0.017/0.056) of the total effect (**Table 5 and Figure 3B**). After conditioning on the lead *IKZF1* SNP rs10230978, we observed stronger total effect (estimate=0.089, P=1.24x10^{-4}) and direct effect (estimate=0.073, P=0.002) of rs78396808 on ALL risk, and a similar mediation effect through cg01139861 (estimate=0.016, P=0.006) (**Table 5 and Figure 3B**). These indicate that the *IKZF1* SNP rs78396808 conferred risk for ALL both through cg01139861 and independent of cg01139861. The associations between DNA methylation at cg01139861, genotypes of rs78396808, and ALL risk in each study set are illustrated in **Figure 4A**. The *IKZF1* SNP rs78396808 risk allele (A) increased DNA methylation at cg01139861 (**Table 3 and Figure 4A**), and the increased DNA methylation level at cg01139861 was associated with increased ALL risk (**Table 2 and Figure 4A**).

Results from causal mediation analyses with variance estimation based on the nonparametric bootstrap method (**Table S7**) were similar to those from the quasi-Bayesian Monte Carlo simulation (**Table 5**).

Race/ethnicity stratified analyses

Significant SNP-DMP pairs in overall participants had similar effect estimates in self-reported Latinos, non-Latino whites, and non-Latinos (non-Latino whites plus non-Latino others) in stratified mQTL analysis, with no significant heterogeneity by race/ethnicity in tests of moderators in fixed-effect meta-analyses (**Table S8**). However, SNP rs78396808 at gene *IKZF1* was associated with cg01139861 in overall participants, in Latinos, and in non-Latinos (non-Latino whites plus non-Latino others) but was not significantly associated with cg01139861 in non-Latino whites. SNP rs78396808 is almost monomorphic for the non-risk allele G in European populations in the Genome Aggregation Database, whereas the risk allele A frequency is ~20% in East Asian and Admixed American populations (29). We found the GA genotype among 11 individuals who were self-reported non-Latino white (**Figure 4A**), likely due to admixture, which may result in the slightly higher effect estimate and standard error for the association between rs78396808 and cg01139861 in non-Latino whites than in Latinos (**Table S8**).

Next, we conducted causal mediation analyses separately in Latinos, non-Latino whites, and non-Latinos, for those significant SNP-DMP pairs identified in overall participants. We found a significant total effect and a significant mediation effect for rs78396808(*IKZF1*)-cg01139861(*IKZF1*)-ALL in the overall meta-analysis in Latinos, with the total effect explained by the ACME being ~42% (0.022/0.053) and ~28% (0.022/0.078) before and after conditioning on the top *IKZF1* SNP rs10230978 (**Table S9**), which were both higher than that observed in overall participants. We found no significant heterogeneity in the ACME for rs78396808(*IKZF1*)-cg01139861(*IKZF1*)-ALL between Latinos and non-Latino whites (P_{het}=0.441) or between Latinos and non-Latinos (P_{het}=0.236) (**Table S9**), likely due to lack of power given the low allele frequency of rs78396808 in non-Latino whites.

ARID5B CpG cg13344587 is a proxy for ALL risk SNP rs7090445

To further disentangle the mQTL-DMP-ALL associations analyzed in the causal mediation analysis, we investigated whether there was a confounding effect from the mQTL genotype on

the association between DNA methylation and ALL risk by fitting three unconditional logistic regression models in subjects with both DNA methylation and genotype data available. Results for both rs7090445(*ARID5B*)-cg13344587(*ARID5B*)-ALL and rs78396808(*IKZF1*)-cg01139861(*IKZF1*)-ALL are summarized in **Table 4**. We obtained the "crude effect" of every 0.1 beta-value increase in *ARID5B* cg13344587 methylation on ALL risk in model 1 (OR_{meta}=0.44, P=2.87x10^{-10}) and the "adjusted effect" in model 2 (OR_{meta}=0.80, P=0.249), with a 82% reduction after adjusting for the SNP effect (**Figure 3A**), much higher than the 10% difference for identifying the presence of confounding (30). In addition, there was no longer a significant association between DNA methylation at cg13344587 and ALL risk, with a greatly reduced effect estimate (OR_{meta}=0.99, P=0.970) in model 3 in individuals without any copies of the rs7090445 risk allele. Therefore, the association between decreased DNA methylation at cg13344587 and ALL risk is consistent with confounding by SNP rs7090445. In contrast, we observed only an ~4% decrease in the effect on ALL risk from CpG cg01139861 at *IKZF1* after adjusting for rs78396808, and the effect remained significant in individuals without any copies of the rs78396808 risk allele (OR_{meta}=1.45, P=0.004). These demonstrated no evidence of a confounding effect from rs78396808 on the association between *IKZF1* cg01139861 and ALL risk.

IKZF1 gene-specific analysis

We investigated additional *IKZF1* CpGs that might mediate the effects of SNP rs78396808 on ALL risk. There were 36 and 42 *IKZF1* CpGs included in the overall meta-analysis and the EPIC array meta-analysis, respectively. CpGs were tested for their association with ALL, with cg01139861 the only one passing gene-wide significance (P<0.025/36), and cg16499656 was an additional CpG with P<0.025 in the overall meta-analysis (**Table S10**). The CpG cg01139861 analyzed before was not included in subsequent analyses. CpG cg12431065 passed the gene-wide significance threshold (0.025/42), and cg10551353 was an additional CpG with P<0.025 in the EPIC array meta-analysis.

Since rs78396808 is monomorphic in European populations, we further conducted the *IKZF1* gene-specific mQTL and causal mediation analyses in Latinos only. *IKZF1* CpGs significantly associated with ALL were tested for their associations with SNP rs78396808 with a p-value <0.025 or <0.0125 (0.025/2) considered to show statistical significance in the overall meta-analysis and the EPIC array meta-analysis, respectively. We identified one significant SNP-DMP pair from the mQTL overall meta-analysis, and two significant SNP-DMP pairs from the mQTL EPIC array meta-analysis (**Table S11**). We found a significant total effect from rs78396808 on ALL (estimate=0.067, *P*=0.030) and a significant mediation effect from rs78396808 (estimate=0.019, *P*=0.031) through increasing DNA methylation at cg10551353 when conditioning on rs10230978 in Latinos, with a ~29% total effect explained by the ACME (**Table S12**). We also found in Latinos that DNA methylation at cg10551353 was strongly correlated with cg01139861, the original *IKZF1* CpG found to mediate the effect of rs78396808 on ALL risk (CCRLP EPIC: R^2=0.41, *P*=3.19x10^{-14}; CCLS EPIC: R^2=0.45, *P*=8.23x10^{-16}) (**Figure S2**). CpG cg01139861 is located in a CpG island in the promoter region of *IKZF1*, and SNP rs78396808 is in an intronic region ~116Kb downstream (**Table 2 and Figure 4B**). The additional *IKZF1* CpG cg10551353 identified here is in the 5′ UTR or the TSS1500 region of several alternative transcripts, ~14Kb downstream of cg01139861 **(Table S10 and Figure 4B)**.

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

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

ALL. The two secondary hits rs78396808 and rs6421315 had weak LD in all populations (R^2=0.0188, D'=0.4213) and in Admixed Americans (R^2=0.072, D'=0.5546) based on LDlink (32), and reside on different sides of a recombination peak (**Figure S3**). Additionally, SNP rs78396808 remained significantly associated with ALL in this study when conditioning on both rs10230978 and rs6421315 (P_{meta}=0.007) (**Table S13**).

Increased DNA methylation at IKZF1 CpG cg01139861 correlated with decreased IKZF1 expression

Finally, we tested the correlation for DNA methylation at cg01139861 (*IKZF1*) with gene expression of nearby genes *IKZF1, FIGNL1, and DDC* using Spearman correlation coefficient tests in 51 ALL tumor samples from CCLS. We excluded 11/71 samples without multiplex ligation-dependent probe amplification (MLPA) copy-number data and 9/60 samples with one deleted copy of *IKZF1*. Increased DNA methylation at the *IKZF1* CpG cg01139861 significantly correlated with decreased gene expression of *IKZF1* (R=-0.28, *P*=0.044; **Figure 5**). In the 9/60 ALL tumors with hemizygous *IKZF1* deletion, DNA methylation levels at cg01139861 were on average lower than in the 51 cases without deletion (median = 0.226 vs. 0.410), although the difference did not reach significance in a Wilcoxon rank sum test (P=0.094) (**Figure S4**).

Discussion

A role for genetic variation in the etiology of childhood ALL is well established, but little is known regarding the association of epigenetic differences at birth and future development of ALL. We report results from the largest neonatal DNA methylation EWAS of childhood ALL performed to date, along with the first comprehensive mediation analysis investigating whether DNA methylation mediates the effects of ALL genetic risk loci. We found that the *IKZF1* SNP rs78396808 risk allele conferred risk for ALL through increasing DNA methylation at the *IKZF1*

promoter CpG cg01139861. In addition, the *ARID5B* CpG cg13344587, the only ALL-associated DMP to survive Bonferroni correction, appeared to be entirely confounded by the *ARID5B* ALL risk SNP rs7090445.

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

Intronic SNPs in *ARID5B*, which plays an important role in B-cell development, have been associated with ALL risk in several GWAS (4,7,31). Functional analysis has shown that the ALL risk SNP rs70904455 disrupts binding of *RUNX3* and leads to reduced *ARID5B* expression (33). We found that the *ARID5B* SNP rs7090445-C risk allele was significantly associated with decreased DNA methylation at the *ARID5B* CpG cg1334458, as previously reported in whole-blood samples from cancer-free individuals (34,35). Although the rs7090445-C risk allele showed the strongest total effect and direct effect on ALL in the causal mediation analysis, the lack of significant mediation effect through altering DNA methylation indicates that the impact of this SNP on ALL risk is independent of cg13344587. Further, results from our analysis of confounding effects supported that the association between decreased DNA methylation at cg13344587 and ALL risk appears to be entirely driven by the *ARID5B* SNP rs7090445, and hypomethylated cg13344587 may function merely as a strong proxy of the rs7090445-C risk allele.

In contrast, we found that DNA methylation at CpG cg01139861, located in a CpG island in the promoter region of *IKZF1*, mediates the effects of the *IKZF1* SNP rs78396808, which we recently reported as an independent ALL-association signal in analysis conditioned on the lead *IKZF1* SNP rs10230978 (28). The causal mediation analysis in our overall dataset showed that the rs78396808-A risk allele had a significant ACME on ALL risk through increasing DNA methylation at cg01139861, explaining ~30% of the total effect on ALL. In analyses stratified by self-reported race/ethnicity, the mediation effect for rs78396808 through increasing DNA methylation at cg01139861 was only significant in Latinos, explaining a ~42% total effect in this population. The SNP rs78396808 is monomorphic in European populations, although the risk allele (A) also presents in African and South Asian populations and has a ~20% frequency in East Asians (29); however, we did not have sufficient samples to test for mediation effects for these population groups. In addition, the *IKZF1* SNP rs78396808 appears to be independent of another secondary association signal at *IKZF1*, SNP rs6421315, previously identified in individuals of European ancestry [13], supporting the existence of at least three independent common genetic risk loci for ALL across populations. In the multivariable regression adjusting for the SNP effect from rs78396808, DNA methylation at cg01139861 remained significantly associated with ALL risk, suggesting that other genetic or environmental risk factors may also affect this CpG site.

Further, we found that increased DNA methylation at cg01139861 correlated with decreased gene expression of *IKZF1* in ALL tumor samples. Gene *IKZF1* encodes the lymphoid transcription factor IKAROS, and is essential for lymphocyte development and differentiation (36). Somatic deletion of *IKZF1* is a common driver event in ALL, particularly in BCR-ABL1-positive ALL (95%) (37) and in high-risk B-cell ALL (30%) (38). *IKZF1* deletions are also enriched in patients with relapsed childhood B-cell ALL (39.3%), in whom increased promoter methylation was also found (21). However, biallelic loss of *IKZF1* is infrequent among ALL patients at diagnosis (39), and we found no evidence that the hypermethylated copy of cg01139861 is selectively retained in ALL tumor samples with hemizygous *IKZF1* deletions; in fact, a lack of

biallelic deletions along with lower DNA methylation in the retained *IKZF1* allele in patients exhibiting a monoallelic deletion may indicate that haploinsufficiency of *IKZF1* rather than complete abrogation is necessary for leukemogenesis. Taken together, the leukemogenic effects of the rs78396808-A risk allele may act via downregulation of *IKZF1* gene expression partly through increased DNA methylation at the *IKZF1* CpG cg01139861. In our targeted analysis of CpGs across *IKZF1*, we identified one additional CpG cg10551353, in the promoter region of several transcripts, which showed evidence of some mediation effect for rs78396808 on ALL risk. Increased DNA methylation at cg10551353 was significantly correlated with increased DNA methylation at cg01139861; however, cg10551353 is on the EPIC array only, so we could not assess its association with *IKZF1* gene expression using the tumor samples assayed on 450K arrays.

The current study has several strengths. First, we assayed pre-diagnostic DNA from neonatal DBS on both genome-wide DNA methylation arrays and SNP arrays, which rules out the possibility of reverse causality (i.e., effects of leukemia itself on DNA methylation). Second, instead of using the traditional mediation analysis approaches relying on the restrictive and untested assumptions, we used a more general estimation framework that provides distribution-free estimates for causal mediation effects and accommodates nonlinearities (40,41). Moreover, we estimated the uncertainty of the causal mediation effects through the quasi-Bayesian Monte Carlo simulation, and we validated our results by using an alternative simulation approach based on nonparametric bootstrap. Last, over half of the participants included in this study were self-reported Latinos, providing an opportunity for us to perform analyses stratified by race/ethnicity, through which we detected a stronger mediation effect for the *IKZF1* SNP rs78396808 through cg01139861 in Latinos than in overall participants.

Our study does have some limitations. One potential limitation was sample size, with only the *ARID5B* CpG cg13344587 reaching epigenome-wide significance in our EWAS for ALL. This necessitated the use of a more lenient p-value threshold ($P<1 \times 10^{-4}$) to identify DMPs for

subsequent mQTL and causal mediation analyses, which may have introduced false positive results, especially in the EPIC array meta-analysis that was conducted with a smaller sample size. We were also limited in our ability to identify DMPs for specific ALL subtypes, or CpG mediators for ALL genetic risk loci associated with particular subtypes, analyses that would require a larger sample size of ALL cases with well-defined tumor subtypes. Another limitation is that we were limited by the number of CpGs on the Illumina arrays. Additional CpGs at *IKZF1* that we could not assess with the array data may also mediate the effect of ALL risk SNP rs78396808. Bisulfite sequencing targeting the *IKZF1* region will be required to fully capture the DNA methylation changes that mediate the effect of rs78396808, especially at *IKZF1* regulatory regions that correlate with gene expression patterns. Lastly, gene expression data were measured from diagnostic leukemia samples and, although we accounted for somatic copy-number loss of *IKZF1*, this may have affected the accuracy of the correlation results between DNA methylation at cg01139861 and *IKZF1* expression.

In conclusion, we provide evidence that increased DNA methylation at the *IKZF1* CpG cg01139861 mediates the effects on ALL risk from SNP rs78396808, which was recently identified as a novel independent risk locus at *IKZF1* that is specific to non-European populations. Our findings enhance the understanding of the functional pathways of genetic risk loci for childhood ALL and provide new insights into the DNA methylation differences associated with childhood ALL.

Materials and Methods

Study participants

Study participants were included from two independent California-based case-control studies of childhood leukemia, the CCLS and the CCRLP, the details of which have been described previously (7,42). Briefly, the CCLS is a population-based case-control study conducted from

1995 to 2015 in multiple counties across California. Cases were identified within 72 hours after diagnosis at hospitals and were eligible for participation if they met all the following criteria: (1) age younger than 15 years, (2) without previous cancer diagnosis, (3) residence in California at the time of diagnosis, and (4) having an English or Spanish-speaking biological parent or guardian. Controls were randomly selected with similar eligibility criteria using birth certificates. One or two controls were matched to each case on the date of birth, sex, and race/ethnicity. The CCRLP linked statewide birth certificates from the California Office of Vital Records (for 1978-2009) to statewide cancer diagnosis data from the California Cancer Registry (1988-2011). Cases were children born in California and diagnosed with their first primary ALL at 0-15 years. Potential controls were children born in California during the same period without prior reports of childhood cancer. Up to four controls were randomly selected and matched to each case on the date of birth, sex, and race/ethnicity. DBS samples were obtained from the California Biobank Program for all participants. Cases (N=808) and controls (N=919) with available genome-wide DNA methylation data were included in the EWAS. Analyses for identifying mQTL and mediation effects were limited to 683 cases and 804 controls with both genome-wide DNA methylation and SNP array data available (therefore, matching between cases and controls was broken for all analyses).

DNA methylation arrays

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

by using the "preprocessFunnorm" function in the minfi package. The beta-mixture quantile normalization method was additionally applied (51). Samples with mean detection p-value >0.01 were considered poor quality and were removed from the analysis. CpG sites and samples that had over 15% missing values were removed. The R package "conumee" (52) was used to generate copy-number variation plots to detect constitutive trisomy of chromosome 21 (T21), as previously described, and a total of 27 ALL cases and 1 control with T21 were excluded from subsequent analyses given the profound effects of Down syndrome on DNA methylation (45).

Genome-wide SNP genotyping

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

Identification of differentially methylated positions (DMPs)

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

and EPIC arrays from three different study sets – CCLS 450K, CCLS EPIC, and CCRLP EPIC datasets – using the R package “metafor" (60). We performed a second meta-analysis for CpGs on EPIC arrays that were limited to the CCLS and CCRLP EPIC datasets only and not included in the CCLS 450K dataset. The associations between CpGs and ALL were corrected for multiple testing using a stringent Bonferroni-adjusted threshold of 0.025 (0.05/2) divided by the number of CpGs included in each meta-analysis, and an FDR of 0.05. In addition, given that previous studies have applied a more liberal threshold (<0.001) to identify DMPs for downstream mediation and interaction analyses (26,27), here we applied a lenient threshold of 1×10^{-4} to ensure sufficient numbers of candidate CpGs for subsequent mQTL and mediation analyses. Manhattan plots and QQ plots were generated using the R package “CMplot” (61). For the identified DMPs, we compared the effect estimates of these CpGs in overall participants with those in self-reported Latinos and non-Latino whites in stratified EWAS, and in EWAS adjusting for genetic ancestry using the first ten principal components derived from the genome-wide SNP array data with PLINK 2.0 (62) instead of from EPISTRUCTURE, in subjects with both DNA methylation and SNP genotype data available.

Identification of methylation quantitative trait loci (mQTL)

We carried out mQTL analyses to identify genotype-dependent DMPs associated with childhood ALL risk using the R package “Matrix eQTL” (63). DMPs with $P<1 \times 10^{-4}$ from the overall meta-analysis or the EPIC array meta-analysis were tested for association with genotypes of the 23 ALL risk SNPs. We fitted an additive linear regression model predicting methylation at each CpG site as a function of SNP genotype (coded 0, 1, and 2), adjusting for the same covariates as for the EWAS. The associations between SNP genotypes and DMP DNA methylation were corrected for multiple testing using a stringent Bonferroni-adjusted threshold of 0.025/(number of DMPs × number of SNPs), and an FDR of 0.05. The mQTL analysis was first conducted separately in the CCLS 450K, CCLS EPIC, and CCRLP EPIC datasets, and the results were subsequently meta-

analyzed across all three datasets, and across the two EPIC datasets, in fixed-effect meta-analysis models using "metafor". We compared these SNP-DMP associations estimated with the DNA methylation beta values versus those estimated when including DNA methylation M-values in the linear regression models (64).

Mediation analysis

We next performed model-based causal mediation analyses for the significant mQTL-DMP pairs, using the "mediation" R package (40). First, we specified two statistical models, (1) the mediator model for the distribution of the DMP methylation level, after conditioning on the genotype of the mQTL and covariates including sex, ancestry, batch effect, and cell type heterogeneity, and (2) the outcome model for the conditional distribution of ALL status, given the mQTL genotype, DMP methylation level, and the same covariates. Models were fitted separately, and then their fitted parameters were used as the main inputs to the mediate function, which computes the estimated ACME, the ADE, and the total effect. Variances were estimated based on simulation. The quasi-Bayesian Monte Carlo simulation based on normal approximation was conducted 1000 times. Alternatively, an approach based on nonparametric bootstrap was also applied to estimate variance for validation. Models with the *IKZF1* SNP rs78396808 were additionally adjusted for the *IKZF1* lead SNP rs10230978, as rs78396808 was previously reported as a secondary ALL association signal in analysis conditioned on rs10230978 (28). Results of the mediation analysis performed separately for the CCLS 450K, CCLS EPIC, and CCRLP EPIC datasets were summarized across all three datasets and across the two EPIC datasets using the fixed-effect meta-analysis model.

We repeated the mQTL analyses and the causal mediation analyses stratified by self-reported race/ethnicity (Latinos vs. non-Latino whites, and Latinos vs. non-Latinos [i.e., non-Latino whites plus non-Latino others]), in study participants with available genome-wide DNA

methylation and SNP array data. We included race/ethnicity as a moderator variable in fixed-effect meta-analysis models to test for heterogeneity.

Locus-specific analyses

To investigated whether there was a confounding effect from the mQTL genotype on the association between DNA methylation and ALL risk, we fitted three unconditional logistic regression models in subjects with both DNA methylation and genotype data available: model 1 was a logistic regression predicting ALL risk as a function of DMP DNA methylation, adjusting for sex, batch effect, cell type heterogeneity, and genetic ancestry; model 2 was additionally adjusted for the mQTL genotype; and model 3 was model 1 fitted in individuals without any copies of the mQTL risk allele. The odds ratios for each 0.1 beta-value increase in DNA methylation in models 1 and 2 were considered as the “crude effect” and the “adjusted effect”, respectively. A reduction >10% of the “adjusted effect” from the “crude effect” provides evidence of confounding (30). In addition, a nonsignificant coefficient in model 3 indicates that the association between DNA methylation and ALL risk is entirely confounded by the SNP effect.

We also performed gene-specific analysis to investigate additional CpGs in the neighboring region of the mediator CpG that could also mediate the effect of mQTL on ALL susceptibility. First, we conducted the overall fixed-effect meta-analysis and the EPIC array fixed-effect meta-analyses limiting to the EPIC array CpGs not in the CCLS 450K dataset for the gene-specific DMP association testing. We retained those CpGs overlapping SNPs in the actual capture, the single base extension, or the full capture sequence of the CpG sites to identify the potential impact from the nearby SNPs. We applied a relaxed significance level of 0.025 in each meta-analysis to ensure sufficient CpGs would be included in the following analysis. Next, significant SNP-CpG associations were identified through the mQTL analysis, with a Bonferroni-adjusted significance level of 0.025/number of CpGs. Finally, causal mediation analysis was performed to identify mediators.

DNA methylation and gene expression analysis

DNA methylation and gene expression data were available from diagnostic leukemia (tumor) samples of 71 ALL cases in the CCLS (25). DNA methylation data from 450K arrays were processed as described above. Genome-wide gene expression data were generated using the GeneChip Human Gene 1.0 ST Array (Affymetrix, Santa Clara, CA), as previously described (25). Copy-number at 8 commonly deleted gene regions in childhood ALL was assayed in 60 out of 71 ALL tumors using MLPA, as previously described (65,66). For DMPs found to be mediators between SNPs and ALL risk, we analyzed the associations between DNA methylation and the expression levels of nearby genes using Spearman correlation coefficient tests, and we further limited the correlation tests to cases without copy number deletion at the corresponding gene regions (if available) to address potential confounding.

Funding

This work was supported by National Institutes of Health (NIH) National Cancer Institute grants R01CA155461 and R01CA175737, National Institute for Environmental Health Sciences (NIEHS) grants R01ES009137, P42ES004705, R24ES028524, and P50ES018172, the United States Environmental Protection Agency grant RD83615901, and the California Tobacco-Related Disease Research Program, Grant No. 26IR-0005A. The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or USEPA or TRDRP. Biospecimens and/or data used in this study were obtained from the California Biobank Program at the California Department of Public Health (CDPH), CBP request number 26, in accordance with Section 6555(b), 17 CCR. The CDPH is not responsible for the results or conclusions drawn by the authors of this publication.

Acknowledgments

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

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

Institute's Surveillance, Epidemiology and End Results Program under contract HHSN261201000140C awarded to the Cancer Prevention Institute of California, contract HHSN261201000035C awarded to the University of Southern California, and contract HHSN261201000034C awarded to the Public Health Institute. The ideas and opinions expressed herein are those of the author(s) and do not necessarily reflect the opinions of the State of California, Department of Public Health, the National Institutes of Health, and the Centers for Disease Control and Prevention or their Contractors and Subcontractors. This study used birth data obtained from the State of California Center for Health Statistics and Informatics. The California Department of Public Health is not responsible for the analyses, interpretations, or conclusions drawn by the authors regarding the birth data used in this publication.

Conflict of Interest Statement

The authors have no conflicts of interest to declare. All co-authors have seen and agree with the contents of the manuscript and there is no financial interest to report.

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.

8. Papaemmanuil, E., Hosking, F.J., Vijayakrishnan, J., Price, A., Olver, B., Sheridan, E., Kinsey, S.E., Lightfoot, T., Roman, E., Irving, J.A.E., *et al.* (2009) Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. *Nat Genet*, **41**, 1006–1010.
9. Perez-Andreu, V., Roberts, K.G., Harvey, R.C., Yang, W., Cheng, C., Pei, D., Xu, H., Gastier-Foster, J., Shuyu, E., Yew-Suang Lim, J., *et al.* (2013) Inherited GATA3 variants are associated with Ph-like childhood acute lymphoblastic leukemia and risk of relapse. *Nat Genet*, **45**, 1494–1498.
10. Feinberg, A.P. and Tycko, B. (2004) The history of cancer epigenetics. *Nat Rev Cancer*, **4**, 143–153.
11. Baylin, S.B. and Jones, P.A. (2016) Epigenetic Determinants of Cancer. *Cold Spring Harb Perspect Biol*, **8**, a019505.
12. Jones, P.A. and Baylin, S.B. (2007) The Epigenomics of Cancer. *Cell*, **128**, 683–692.
13. Ehrlich, M. and Lacey, M. (2013) DNA Hypomethylation and Hemimethylation in Cancer. In Karpf, A. R. (ed.), *Epigenetic Alterations in Oncogenesis*, Advances in Experimental Medicine and Biology, Springer, New York, NY, pp. 31–56.
14. You, J.S. and Jones, P.A. (2012) Cancer Genetics and Epigenetics: Two Sides of the Same Coin? *Cancer Cell*, **22**, 9–20.
15. Garraway, L.A. and Lander, E.S. (2013) Lessons from the Cancer Genome. *Cell*, **153**, 17–37.
16. Shen, H. and Laird, P.W. (2013) Interplay Between the Cancer Genome and Epigenome. *Cell*, **153**, 38–55.
17. Liu, Y., Aryee, M.J., Padyukov, L., Fallin, M.D., Hesselberg, E., Runarsson, A., Reinius, L., Acevedo, N., Taub, M., Ronninger, M., *et al.* (2013) Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat Biotechnol*, **31**, 142–147.

18. Dai, J.Y., Wang, X., Wang, B., Sun, W., Jordahl, K.M., Kolb, S., Nyame, Y.A., Wright, J.L., Ostrander, E.A., Feng, Z., *et al.* (2020) DNA methylation and cis-regulation of gene expression by prostate cancer risk SNPs. *PLoS Genet*, **16**, e1008667.
19. Nedeljkovic, I., Lahousse, L., Carnero-Montoro, E., Faiz, A., Vonk, J.M., de Jong, K., van der Plaat, D.A., van Diemen, C.C., van den Berge, M., Obeidat, M., *et al.* (2018) COPD GWAS variant at 19q13.2 in relation with DNA methylation and gene expression. *Human Molecular Genetics*, **27**, 396–405.
20. Hale, V., Hale, G.A., Brown, P.A. and Amankwah, E.K. (2017) A Review of DNA Methylation and microRNA Expression in Recurrent Pediatric Acute Leukemia. *Oncology*, **92**, 61–67.
21. Hogan, L.E., Meyer, J.A., Yang, J., Wang, J., Wong, N., Yang, W., Condos, G., Hunger, S.P., Raetz, E., Saffery, R., *et al.* (2011) Integrated genomic analysis of relapsed childhood acute lymphoblastic leukemia reveals therapeutic strategies. *Blood*, **118**, 5218–5226.
22. Nordlund, J., Bäcklin, C.L., Wahlberg, P., Busche, S., Berglund, E.C., Eloranta, M.-L., Flaegstad, T., Forestier, E., Frost, B.-M., Harila-Saari, A., *et al.* (2013) Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia. *Genome Biol*, **14**, r105.
23. Chatterton, Z., Morenos, L., Mechinaud, F., Ashley, D.M., Craig, J.M., Sexton-Oates, A., Halemba, M.S., Parkinson-Bates, M., Ng, J., Morrison, D., *et al.* (2014) Epigenetic deregulation in pediatric acute lymphoblastic leukemia. *Epigenetics*, **9**, 459–467.
24. Nordlund, J. and Syvänen, A.-C. (2018) Epigenetics in pediatric acute lymphoblastic leukemia. *Seminars in Cancer Biology*, **51**, 129–138.
25. Lee, S.-T., Muench, M.O., Fomin, M.E., Xiao, J., Zhou, M., de Smith, A., Martín-Subero, J.I., Heath, S., Houseman, E.A., Roy, R., *et al.* (2015) Epigenetic remodeling in B-cell acute lymphoblastic leukemia occurs in two tracks and employs embryonic stem cell-like signatures. *Nucleic Acids Research*, **43**, 2590–2602.

26. Shu, C., Justice, A.C., Zhang, X., Wang, Z., Hancock, D.B., Johnson, E.O. and Xu, K. (2020) DNA methylation mediates the effect of cocaine use on HIV severity. *Clinical Epigenetics*, **12**, 140.

27. Declerck, K., Remy, S., Wohlfahrt-Veje, C., Main, K.M., Van Camp, G., Schoeters, G., Vanden Berghe, W. and Andersen, H.R. (2017) Interaction between prenatal pesticide exposure and a common polymorphism in the PON1 gene on DNA methylation in genes associated with cardio-metabolic disease risk—an exploratory study. *Clinical Epigenetics*, **9**, 35.

28. Jeon, S., de Smith, A.J., Li, S., Chen, M., Chan, T.F., Muskens, I.S., Morimoto, L.M., DeWan, A.T., Mancuso, N., Metayer, C., *et al.* (2021) Genome-wide trans-ethnic meta-analysis identifies novel susceptibility loci for childhood acute lymphoblastic leukemia. *Leukemia*, 1–4.

29. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., *et al.* (2019) The mutational constraint spectrum quantified from variation in 141,456 humans. *The mutational constraint spectrum quantified from variation in 141,456 humans*; preprint; Genomics, (2019) .

30. VanderWeele, T.J. (2019) Principles of confounder selection. *Eur J Epidemiol*, **34**, 211–219.

31. Vijayakrishnan, J., Qian, M., Studd, J.B., Yang, W., Kinnersley, B., Law, P.J., Broderick, P., Raetz, E.A., Allan, J., Pui, C.-H., *et al.* (2019) Identification of four novel associations for B-cell acute lymphoblastic leukaemia risk. *Nat Commun*, **10**, 5348.

32. Machiela, M.J. and Chanock, S.J. (2015) LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics*, **31**, 3555–3557.

33. Studd, J.B., Vijayakrishnan, J., Yang, M., Migliorini, G., Paulsson, K. and Houlston, R.S. (2017) Genetic and regulatory mechanism of susceptibility to high-hyperdiploid acute lymphoblastic leukaemia at 10q21.2. *Nat Commun*, **8**.

34. Hannon, E., Weedon, M., Bray, N., O'Donovan, M. and Mill, J. (2017) Pleiotropic Effects of Trait-Associated Genetic Variation on DNA Methylation: Utility for Refining GWAS Loci. *Am J Hum Genet*, **100**, 954–959.
35. Hannon, E., Dempster, E., Viana, J., Burrage, J., Smith, A.R., Macdonald, R., St Clair, D., Mustard, C., Breen, G., Therman, S., *et al.* (2016) An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential DNA methylation. *Genome Biology*, **17**, 176.
36. Harker, N., Naito, T., Cortes, M., Hostert, A., Hirschberg, S., Tolaini, M., Roderick, K., Georgopoulos, K. and Kioussis, D. (2002) The CD8alpha gene locus is regulated by the Ikaros family of proteins. *Mol Cell*, **10**, 1403–1415.
37. Mullighan, C.G., Miller, C.B., Radtke, I., Phillips, L.A., Dalton, J., Ma, J., White, D., Hughes, T.P., Le Beau, M.M., Pui, C.-H., *et al.* (2008) BCR-ABL1 lymphoblastic leukaemia is characterized by the deletion of Ikaros. *Nature*, **453**, 110–114.
38. Mullighan, C.G., Su, X., Zhang, J., Radtke, I., Phillips, L.A.A., Miller, C.B., Ma, J., Liu, W., Cheng, C., Schulman, B.A., *et al.* (2009) Deletion of IKZF1 and Prognosis in Acute Lymphoblastic Leukemia. *N Engl J Med*, **360**, 470–480.
39. Churchman, M.L., Low, J., Qu, C., Paietta, E.M., Kasper, L.H., Chang, Y., Payne-Turner, D., Althoff, M.J., Song, G., Chen, S.-C., *et al.* (2015) EFFICACY OF RETINOIDS IN IKZF1-MUTATED BCR-ABL1 ACUTE LYMPHOBLASTIC LEUKEMIA. *Cancer Cell*, **28**, 343–356.
40. Tingley, D., Yamamoto, T., Hirose, K., Keele, L. and Imai, K. (2014) **mediation** : *R* Package for Causal Mediation Analysis. *J. Stat. Soft.*, **59**.
41. Imai, K., Keele, L. and Tingley, D. (2010) A general approach to causal mediation analysis. *Psychological Methods*, **15**, 309–334.
42. Metayer, C., Zhang, L., Wiemels, J.L., Bartley, K., Schiffman, J., Ma, X., Aldrich, M.C., Chang, J.S., Selvin, S., Fu, C.H., *et al.* (2013) Tobacco Smoke Exposure and the Risk of

Childhood Acute Lymphoblastic and Myeloid Leukemias by Cytogenetic Subtype. *Cancer Epidemiology Biomarkers & Prevention*, **22**, 1600–1611.

43. Gonseth, S., Smith, A.J. de, Roy, R., Zhou, M., Lee, S.-T., Shao, X., Ohja, J., Wrensch, M.R., Walsh, K.M., Metayer, C., *et al.* (2016) Genetic contribution to variation in DNA methylation at maternal smoking-sensitive loci in exposed neonates. *Epigenetics*, **11**, 664–673.
44. Xu, K., Li, S., Whitehead, T.P., Pandey, P., Kang, A.Y., Morimoto, L.M., Kogan, S.C., Metayer, C., Wiemels, J.L. and de Smith, A.J. (2021) Epigenetic Biomarkers of Prenatal Tobacco Smoke Exposure Are Associated with Gene Deletions in Childhood Acute Lymphoblastic Leukemia. *Cancer Epidemiol Biomarkers Prev*.
45. Muskens, I.S., Li, S., Jackson, T., Elliot, N., Hansen, H.M., Myint, S.S., Pandey, P., Schraw, J.M., Roy, R., Anguiano, J., *et al.* (2021) The genome-wide impact of trisomy 21 on DNA methylation and its implications for hematopoiesis. *Nature Communications*, **12**, 821.
46. Fortin, J.-P., Labbe, A., Lemire, M., Zanke, B.W., Hudson, T.J., Fertig, E.J., Greenwood, C.M. and Hansen, K.D. (2014) Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol*, **15**.
47. Aryee, M.J., Jaffe, A.E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A.P., Hansen, K.D. and Irizarry, R.A. (2014) Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, **30**, 1363–1369.
48. Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B.S., Bravo, H.C., Davis, S., Gatto, L., Girke, T., *et al.* (2015) Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods*, **12**, 115–121.
49. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 16.

50. Triche, T.J., Weisenberger, D.J., Van Den Berg, D., Laird, P.W. and Siegmund, K.D. (2013) Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res*, **41**, e90.
51. Teschendorff, A.E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D. and Beck, S. (2013) A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics*, **29**, 189–196.
52. Mah, C.K., Mesirov, J.P. and Chavez, L. (2018) An accessible GenePattern notebook for the copy number variation analysis of Illumina Infinium DNA methylation arrays. *F1000Res*, **7**, 1897.
53. Hansen, K. (2016) IlluminaHumanMethylation450kanno. ilmn12. hg19: annotation for Illumina's 450k methylation arrays. *R package version 0.6. 0*, **10**, B9.
54. Hansen, K. (2016) IlluminaHumanMethylationEPICanno. ilm10b2. hg19: Annotation for Illumina's EPIC methylation arrays; R package version 0.6. 0. .
55. Chen, Y., Lemire, M., Choufani, S., Butcher, D.T., Grafodatskaya, D., Zanke, B.W., Gallinger, S., Hudson, T.J. and Weksberg, R. (2013) Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics*, **8**, 203–209.
56. Pidsley, R., Zotenko, E., Peters, T.J., Lawrence, M.G., Risbridger, G.P., Molloy, P., Van Djik, S., Muhlhausler, B., Stirzaker, C. and Clark, S.J. (2016) Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biology*, **17**, 208.
57. McCartney, D.L., Walker, R.M., Morris, S.W., McIntosh, A.M., Porteous, D.J. and Evans, K.L. (2016) Identification of polymorphic and off-target probe binding sites on the Illumina Infinium MethylationEPIC BeadChip. *Genomics Data*, **9**, 22–24.

58. Rahmani, E., Zaitlen, N., Baran, Y., Eng, C., Hu, D., Galanter, J., Oh, S., Burchard, E.G., Eskin, E., Zou, J., *et al.* (2016) Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nat. Methods*, **13**, 443–445.

59. Rahmani, E., Shenhav, L., Schweiger, R., Yousefi, P., Huen, K., Eskenazi, B., Eng, C., Huntsman, S., Hu, D., Galanter, J., *et al.* (2017) Genome-wide methylation data mirror ancestry information. *Epigenetics & Chromatin*, **10**, 1.

60. Viechtbauer, W. (2010) Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, **36**, 1–48.

61. Yin, L., Zhang, H., Tang, Z., Xu, J., Yin, D., Zhang, Z., Yuan, X., Zhu, M., Zhao, S., Li, X., *et al.* (2021) rMVP: A Memory-efficient, Visualization-enhanced, and Parallel-accelerated tool for Genome-Wide Association Study. *Genomics, Proteomics & Bioinformatics*.

62. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., *et al.* (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet*, **81**, 559–575.

63. Shabalin, A.A. (2012) Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, **28**, 1353–1358.

64. Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W.A., Hou, L. and Lin, S.M. (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, **11**, 587.

65. de Smith, A.J., Kaur, M., Gonseth, S., Endicott, A., Selvin, S., Zhang, L., Roy, R., Shao, X., Hansen, H.M., Kang, A.Y., *et al.* (2017) Correlates of Prenatal and Early-Life Tobacco Smoke Exposure and Frequency of Common Gene Deletions in Childhood Acute Lymphoblastic Leukemia. *Cancer Res*, **77**, 1674–1683.

66. Walsh, K.M., de Smith, A.J., Welch, T.C., Smirnov, I., Cunningham, M.J., Ma, X., Chokkalingam, A.P., Dahl, G.V., Roberts, W., Barcellos, L.F., *et al.* (2014) Genomic

776 ancestry and somatic alterations correlate with age at diagnosis in Hispanic children with B-
777 cell ALL. *Am J Hematol*, **89**, 721–725.
778 67. Hahne, F. and Ivanek, R. (2016) Visualizing Genomic Data Using Gviz and Bioconductor. In
779 Mathé, E., Davis, S. (eds.), *Statistical Genomics*, Methods in Molecular Biology, Springer
780 New York, New York, NY, Vol. 1418, pp. 335–351.

781

Legends to Figures

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.

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 1×10^{-4} (dash line). Y-axis for the Bidirectional Manhattan plot represents $-\log_{10}P_{hyper}$ for hypermethylated CpGs (higher DNA methylation beta values in cases vs. in controls) and $\log_{10}P_{hypo}$ 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.

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.

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.

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 $\log_2$ fold changes of *IKZF1* in 51 tumor samples from CCLS. The Spearman correlation coefficient R and its p-values are shown in the plot.

Legends to Tables

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

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

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

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

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).

Tables

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

	CCLS 450K (n = 435)			CCRLP EPIC (n = 566)			CCLS EPIC (n = 726)		
Variables	**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.

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

							CCLS 450K		CCRLP EPIC		CCLS EPIC		Meta-analysis				
Probe	**Chr**	**Pos**	**Islands Name**	**Relation to Island**	**UCSC RefGene Name**	**UCSC RefGene Group**	**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	HSFYP1	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;TSS1500	-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;TSS1500	-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;TSS1500;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

			47743151														
cg18708233	chr10	74451549	chr10:74451828-74452597	N_Shore	CCDC109A	TSS1500	-8.12	2.11	-2.64	1.96	-3.54	1.57	-4.43	1.06	2.84E-05	0.12	52.72
cg21727223	chr17	66755770		OpenSea			4.08	2.16	5.45	2.55	5.11	1.68	4.88	1.18	3.29E-05	0.90	0.00
cg20451680	chr5	54281336		OpenSea	ESM1;ESM1	1stExon;1stExon	-6.20	1.87	-1.15	2.22	-5.21	1.72	-4.56	1.10	3.52E-05	0.20	38.80
cg23293256	chr10	11631072		OpenSea	USP6NL	Body	6.86	2.48	2.77	7.36	9.05	2.88	7.48	1.82	3.99E-05	0.68	0.00
cg03293350	chr18	52891017		OpenSea	TCF4;TCF4	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
cg15491120	chr18	74156897	chr18:74153239-74155073	S_Shore	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
cg20572153	chr17	9929755		OpenSea	GAS7;GAS7;GAS7	Body;Body;TSS200	-6.01	3.23	-5.07	3.54	-9.99	2.81	-7.43	1.82	4.47E-05	0.48	0.00
cg16276850	chr17	38498914	chr17:38497527-38498963	Island	RARA;RARA;RARA;RARA;RARA	5'UTR;1stExon;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
cg08943714	chr6	139470282		OpenSea	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
cg15260921	chr1	226308960	chr1:226308959-226310476	Island			-8.27	3.46	-11.46	5.15	-8.67	3.54	-9.03	2.23	5.18E-05	0.87	0.00
cg04035597	chr8	19794539	chr8:19796843-19798006	N_Shelf			-2.73	1.80	-4.05	2.33	-6.30	1.74	-4.46	1.10	5.18E-05	0.35	3.51
cg27579121	chr6	41859617	chr6:41862841-41863335	N_Shelf	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
cg01139861*	chr7	50343298	chr7:50342895-50343456	Island	IKZF1	TSS1500	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	chr6:18122250-18122994	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

cg10475928	chr10	118901190	chr10:118899247-118900329	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
cg25412453	chr1	155006219		OpenSea	DCST1;DCST2;DCST2;DCST1	TSS200;5'UTR;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
cg15721728	chr11	66251141	chr11:66245526625073	S_Shore	ILK;ILK;RRP8;ILK	5'UTR;5'UTR;TSS1500;TSS200	21.82	9.70	28.82	10.58	16.55	8.33	21.42	5.43	7.87E-05	0.66	0.00
cg04247829	chr1	38218080	chr1:38218190-38218977	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
cg13185177	chr3	194119885	chr3:194117601-194118988	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
cg14362370	chr9	133589654	chr9:133587865-133588901	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
cg19836174	chr5	149921140		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
cg13230172	chr11	120195828	chr11:120195805-120196372	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
cg09180564	chr1	102308808		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
cg23940023	chr14	101521703		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
cg16728651	chr6	55767865		OpenSea			3.21	2.18	6.84	2.21	3.48	1.50	4.22	1.08	9.56E-05	0.40	0.00
cg16969852	chr9	111775933	chr9:111775041-111775934	Island	CTNNAL1	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.

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

	CCLS 450K		CCRLP EPIC		CCLS EPIC		Meta-analysis					
mQTL-DMP pair	**Coef**	**SE**	**Coef**	**SE**	**Coef**	**SE**	**Coef**	**SE**	***P***	***P*.het**	**i.squared**	**FDR**
rs7090445 (chr10:63721176 at *ARID5B*) and cg13344587 (chr10:63723919 at *ARID5B*)	-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 *IKZF1*) and cg01139861 (chr7:50343298 at *IKZF1*)	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.

	CCLS 450K	CCRLP EPIC	CCLS EPIC	Meta-analysis			
Variable	**OR (CI)**	**OR (CI)**	**OR (CI)**	**OR (CI)**	***P***	***P.het***	**i.squared**
rs7090445(ARID5B)-cg13344587(ARID5B)-ALL							
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)							
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
Model 2: model 1 additionally adjusted for SNP rs7090445 (808 cases and 919 controls)							
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
Model 3: model 1 in subjects without any copies of the risk allele of SNP rs7090445 (154 cases and 292 controls)							
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
(ORadj - ORcrude)/ORcrude = (0.44-0.80)/0.44 = -82%							
rs78396808(IKZF1)-cg01139861(IKZF1)-ALL							
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)							
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
Model 2: model 1 additionally adjusted for SNP rs78396808 (808 cases and 919 controls)							
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
Model 3: model 1 in subjects without any copies of the risk allele of SNP rs78396808 (490 cases and 592 controls)							
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
(ORadj - ORcrude)/ORcrude = (1.51-1.45)/1.51= 4%							

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	*P*	*P.het*	i.squared	Effect
rs7090445 (chr10:63721176 at *ARID5B*) and cg13344587 (chr10:63723919 at ARID5B)	0.108	0.014	7.82	**5.17E-15**	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	**1.06E-04**	0.054	65.78	ADE
rs78396808 (chr7:50459043 at IKZF1) and cg01139861 (chr7:50343298 at IKZF1)	0.056	0.022	2.56	**1.05E-02**	0.918	0.00	total
	0.017	0.006	2.96	**3.09E-03**	0.435	0.00	ACME
	0.039	0.022	1.77	7.66E-02	0.997	0.00	ADE
rs78396808 (chr7:50459043 at IKZF1) and cg01139861 (chr7:50343298 at IKZF1)*	0.089	0.023	3.84	**1.24E-04**	0.704	0.00	total
	0.016	0.006	2.73	**6.28E-03**	0.435	0.00	ACME
	0.073	0.024	3.08	**2.04E-03**	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*).

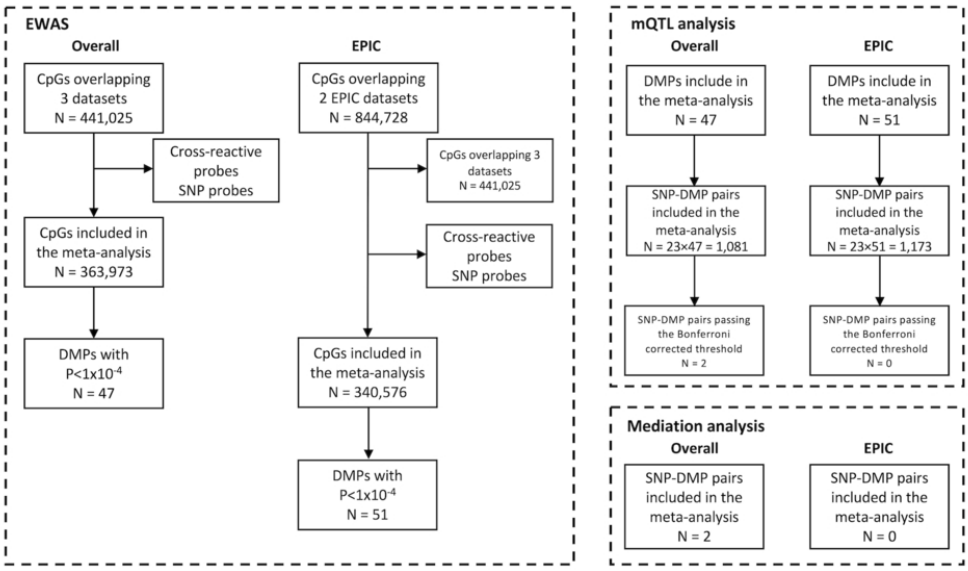

Fig. 1

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)

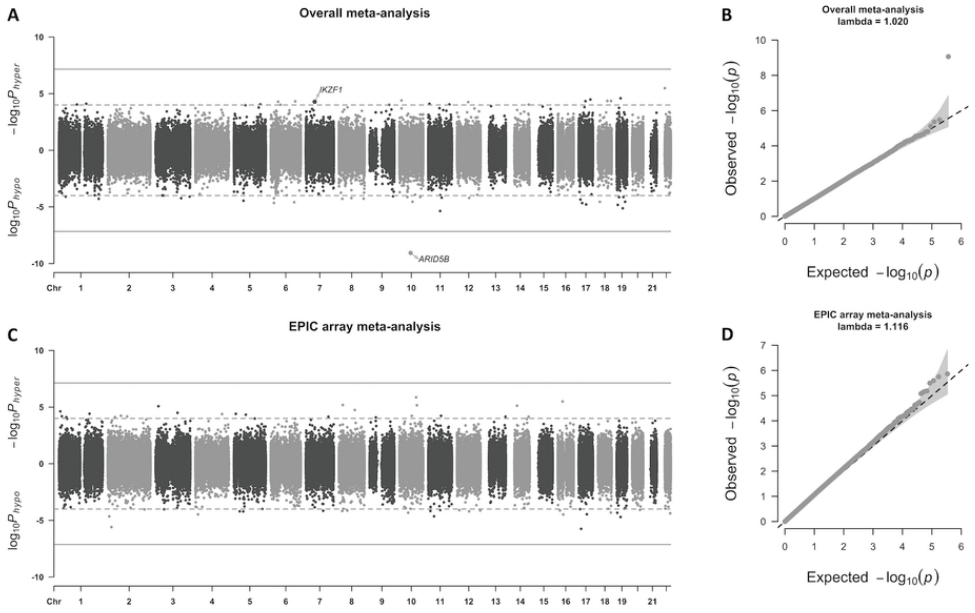

Fig. 2

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 x 10-4 (dash line). Y-axis for the Bidirectional Manhattan plot represents -log10Phyper for hypermethylated CpGs (higher DNA methylation beta values in cases vs. in controls) and log10Phypo 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)

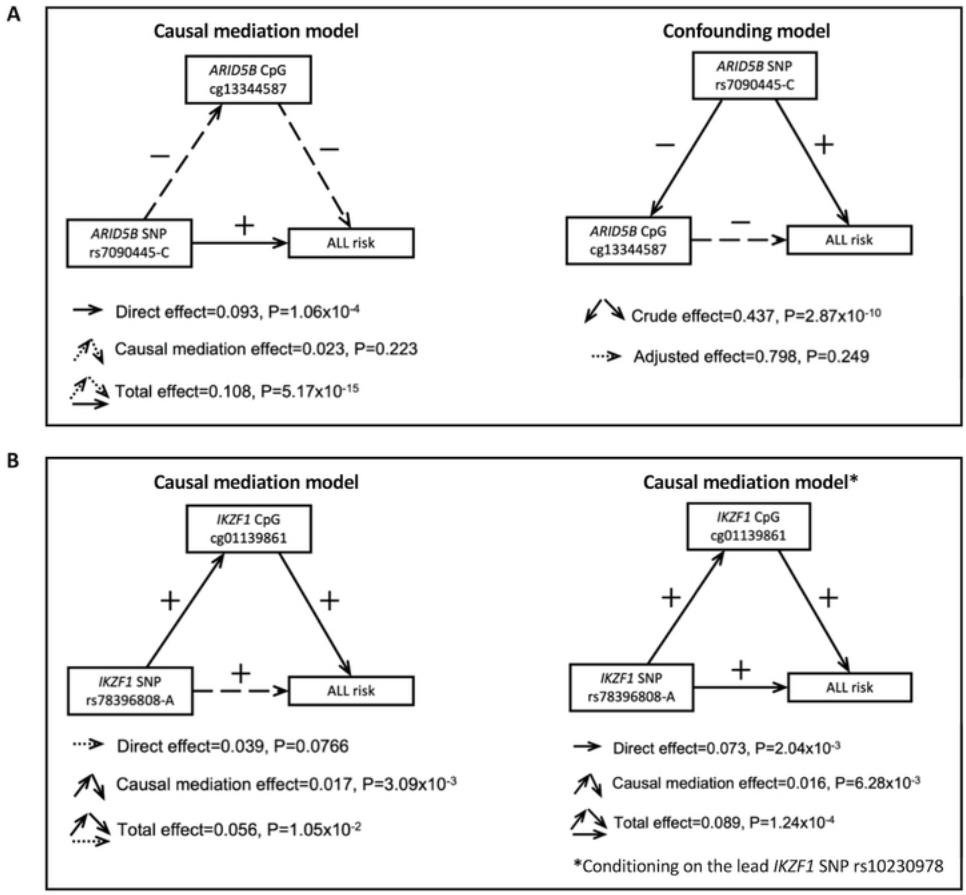

Fig. 3

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)

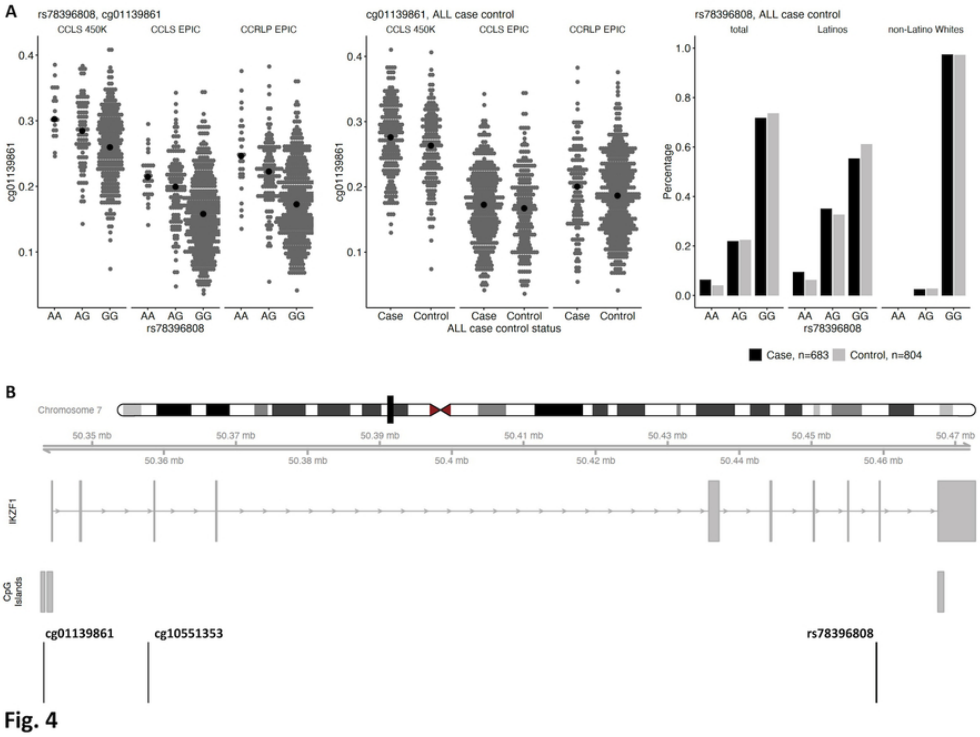

Fig. 4

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)

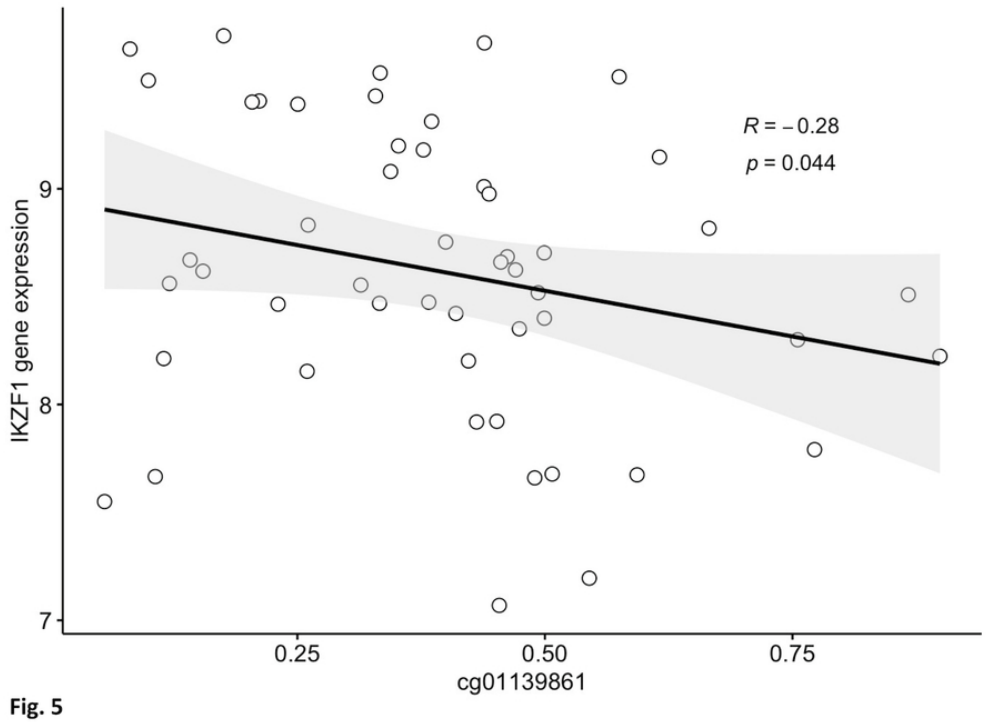

Fig. 5

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)