A childhood acute lymphoblastic leukemia genome-wide association study identifies novel sex-specific risk variants

Sandeep K. Singh, PhDab, Philip J. Lupo, PhDb, Michael E. Scheurer, PhDb, Anshul Saxena, MPHa, Amy E. Kennedy, PhD, Boubakari Ibrahimou, PhD, Manuel Alejandro Barbieri, PhDb, Ken I. Mills, PhD, Jacob L. McCauley, PhD, Mehmet Fatih Okcu, MD, MPH, Mehmet Tsviik Dorak, MD, MPH,∗

Abstract

Childhood acute lymphoblastic leukemia (ALL) occurs more frequently in males. Reasons behind sex differences in childhood ALL risk are unknown. In the present genome-wide association study (GWAS), we explored the genetic basis of sex differences by comparing genotype frequencies between male and female cases in a case-only study to assess effect-modification by sex.

The case-only design included 236 incident cases of childhood ALL consecutively recruited at the Texas Children’s Cancer Center in Houston, Texas from 2007 to 2012. All cases were non-Hispanic whites, aged 1 to 10 years, and diagnosed with confirmed B-cell precursor ALL. Genotyping was performed using the Illumina HumanCoreExome BeadChip on the Illumina Infinium platform. Besides the top 100 statistically most significant results, results were also analyzed by the top 100 highest effect size with a nominal statistical significance (P < 0.05).

The statistically most significant sex-specific association (P = 4 × 10−9) was with the single nucleotide polymorphism (SNP) rs4813720 (RASSF2), an expression quantitative trait locus (eQTL) for RASSF2 in peripheral blood. rs4813720 is also a strong methylation QTL (meQTL) for a CpG site (cg22485289) within RASSF2 in pregnancy, at birth, childhood, and adolescence. cg22485289 is one of the hypermethylated CpG sites in ALL compared with pre-B cells. Two missense SNPs, rs12722042 and rs12722039, in the HLA-DQA1 gene yielded the highest effect sizes (odds ratio [OR] = 14; P < 0.01) for sex-specific results. The HLA-DQA1 SNPs belong to DQA1*01 and confirmed the previously reported male-specific association with DQA1*01. This finding supports the proposed infection-related etiology in childhood ALL risk for males. Further analyses revealed that most SNPs (either direct effect or through linkage disequilibrium) were within active enhancers or active promoter regions and had regulatory effects on gene expression levels.

Cumulative data suggested that RASSF2 rs4813720, which correlates with increased RASSF2 expression, may counteract the suppressor effect of estrogen-regulated miR-17-92 on RASSF2 resulting in protection in males. Given the amount of sex hormone-related mechanisms suggested by our findings, future studies should examine prenatal or early postnatal programming by sex hormones when hormone levels show a large variation.

Abbreviations: ADAM28 = ADAM metalloproteinase domain 28, ALL = acute lymphoblastic leukemia, ARID5B = AT rich interactive domain 5b, BCM = Baylor College of Medicine, CI = confidence interval, eQTL = expression quantitative trait locus, ER = estrogen receptor, ERCC1 = excision repair cross-complementation group 1, FDR = false discovery rate, FIU = Florida International University, GSEA = gene set enrichment analysis, GWAS = genome-wide association study, H3K27ac = histone-3 lysine-27 acetylation, H3K4me1 = histone-3 lysine-4 monomethylation, H3K4me3 = histone-3 lysine-4 trimethylation, H3K9ac = histone-3 lysine-9 acetylation, HLA-DQA1 = major histocompatibility complex, class II, DQA1*01 = Hardy–Weinberg equilibrium, IFNG = interferon gamma, IFI1 = interferon regulatory factor 1, KAT7 = Klysine acetyltransferase 7, LD = linkage disequilibrium, meQTL = methylation quantitative trait locus, MILE study = The Microarray Innovations in Leukemia study, NF = nuclear factor, OR = odds ratio, QC = quality control, QQ = quantile-quantile, RAP1GAP2 = RAP1 GTPase activating protein 2, RASSF2 = Ras association
1. Introduction

Acute lymphoblastic leukemia (ALL) is the most common pediatric malignancy, with an annual incidence rate of 42 cases per million children under age 15 in the United States.\[^1\] While overall 5-year survival rate has improved to 80% on average, the US annual incidence rate increased by 0.8% per year from 1975 to 2012 in children aged 14 years and younger.\[^1,2\] Like many other diseases and cancers, childhood ALL risk shows consistent sex differences.\[^1,3\] Males show a higher incidence rate in comparison to females (45 cases per million children versus 39 cases per million children per year, respectively) in the United States.\[^3\] Relapse and secondary malignancies are also more common in males.\[^6\] Reasons underlying sex differences in childhood ALL risk are still unknown, as it has not been extensively evaluated.\[^4\]

Sex differences in susceptibility to complex disorders is an active research area, and some progress has been made in recent years.\[^7,8\] Studies that examined sex differential for etiological clues identified several sex-specific genetic markers also for childhood ALL risk.\[^9-14\] For example, a study that examined previously identified childhood ALL single nucleotide polymorphisms (SNPs) in genome-wide association studies (GWAS)\[^15,16\] reported sex-specific effects of 2 statistically correlated ARID5B (AT-rich interactive domain 5b) SNPs (rs10994982 \[P = 0.01\] and rs10740055 \[P = 0.03\]) with ALL risk in males (odds ratio \(OR = 3.79\) and \(OR = 4.35\), respectively; ORs for females \(OR = 1.03\) and \(1.37\), respectively).\[^12\] Candidate gene studies have also reported sex differences in genetic associations in childhood ALL. Two SNPs in ERCCI (excision repair cross-complementation group 1) are associated with childhood ALL risk among males, but not with females.\[^13\] Likewise, several multiple sclerosis risk markers, such as HLA-DRA (rs3135388), HLA-C (rs9264942), HSPA1B (rs1061581), and IFNG (rs2069727), also yielded sex-specific associations with childhood ALL risk.\[^10\] Associations with an intronic SNP (rs12203592) in IRF4 and an intergenic SNP (rs2395185) near HLA-DRA further suggest the existence of sex-specific genetic risk variants.\[^9,11\]

The robust and replicated sex-specific association with an IFNG (interferon gamma) polymorphism (rs2069727; \(P_{interaction} = 0.008\)) in asthma suggests the existence of genetic contribution to sex differences in nonmalignant childhood disorders.\[^17\] The same SNP also shows a sex-specific association with childhood ALL risk.\[^10\] A recent genome-wide meta-analysis of asthma revealed sex-specific candidate risk markers in interferon regulatory factor 1 (IRF1) and RAP1 GTPase activating protein 2 (RAP1GAP2) with regulatory effects in gene regulation as expression quantitative trait loci (eQTL).\[^18\] Yao et al\[^19\] examined 11,672 disease-associated polymorphisms from the National Human Genome Research Institute GWAS catalog and dbGaP database in relation to gene expression levels in the whole blood derived RNA and identified 14 sex-specific eQTLs. A recent survey of correlations between SNPs and gene expression levels by sex showed that up to 15% of autosomal SNPs have a gender bias in their correlations.\[^20\] Similarly, another study reported that 582 autosomal genes have sex-specific differences in their expression levels.\[^21\] Such sex-specific signals are likely to be diluted or cancelled out when both genders are analyzed simultaneously. Likewise, reanalysis of GWAS datasets revealed that coronary artery diseases, Crohn disease, rheumatoid arthritis, and type 1 diabetes show sex-specific associations with increased risk in only 1 gender.\[^22,23\] In the present GWAS, we explored the genetic basis of sex differences by comparing genotype frequencies between male and female cases with childhood ALL using a case-only study design.

2. Materials and methods

2.1. Study design

We used a case-only design because of its greater statistical power in detecting effect modification, in this case, sex and genotype interactions, in comparison to other traditional epidemiological designs.\[^24\] An assumption of the case-only design requires genotype and sex to be independent in the healthy population. There is no reason to believe that healthy male and female controls have different genotype frequencies in autosomal chromosomes. Recruitment of cases in a rare disease like childhood ALL is a challenge; hence, preferring a statistically more powerful research design was more practical.

2.2. Subject recruitment

Institutional review board approval was granted by Florida International University (FIU), Office of Research Integrity, and Baylor College of Medicine (BCM) prior to the study. Samples for the current study were obtained from a parent case-control study of childhood ALL and has been explained elsewhere in detail.\[^11\] Briefly, incident cases of childhood ALL were consecutively recruited at the Texas Children’s Cancer Center (TXCCC), BCM, in Houston, Texas from 2007 to 2012. The present study used 236 cases from the parent study for the case-only analysis. They were all non-Hispanic whites, aged 1 to 10 years, and diagnosed with confirmed B-cell precursor ALL. Genomic DNA was extracted from blood samples at TXCCC, BCM. Established childhood ALL risk associations identified in previous GWAS\[^14-16\] have been replicated to validate the parent case-control study sample using TaqMan allelic discrimination assays at FIU.\[^11\]

2.3. Genotyping

Genotype data were generated using the Illumina Human-CoreExome BeadChip (Illumina, San Diego, CA) on the Illumina Infinium platform at the John P. Hussman Institute of Human Genomics, University of Miami, Florida.

2.4. Data analysis

Quality control (QC) steps were performed in a hierarchical fashion using PLINK and R packages.\[^25\] Sample QC was done prior to SNP QC. The Hardy–Weinberg equilibrium (HWE) test was used retrospectively in the whole group of subjects for
sex-specific analysis. Unconditional logistic regression was used to calculate allelic ORs and 95% confidence intervals (CIs) after coding males as cases and females as controls. The coding of male and female cases for statistical analysis meant that an OR greater than 1.0 indicates greater risk for males compared with females.

Power analysis and sample software was used to calculate statistical power for various risk genotype frequencies. The study was statistically powered to detect qualitative interactions, that is sex-specific associations differing in direction (risk or protection); rather than associations in the same direction and differing in magnitude (quantitative interactions). Permutation testing (10,000 permutations) was performed for all sex-specific results to rule out chance findings using a threshold of $P<0.05$. Results were adjusted for potential population stratification using the genomic control method.[26] A quantile-quantile (Q-Q) plot was used to rule out systematic errors (Supplementary Figure 1, http://links.lww.com/MD/B374). Besides the top 100 statistically most significant results (Supplementary Table 1, http://links.lww.com/MD/B375), results were also analyzed by the OR with a nominal statistical significance ($P<0.05$) (Supplementary Table 2, http://links.lww.com/MD/B376).

### 2.5. Bioinformatic analysis

To functionally annotate the SNPs that have shown sex-specific associations, a number of bioinformatic tools were used. Variants were annotated for their effects on the gene expression, protein, and the predicted function using several bioinformatic and empirical tools. ANNOVAR was used to obtain functionality scores (DANN, FATHMM, GWAVA, SIFT, PolyPhen2, Mutation Taster, Mutation Assessor, LRT, FATHMM, and MetaLR), as well as conservation scores.[27] ANNOVAR and SNPnexus[28] were used to predict the effect of variants on transcription factor binding sites (TFBS), microRNA binding sites and identification of variants that disrupt enhancers, repressors, and promoters. Complementary annotations were performed using CADD,[29] RegulomeDB,[30] HaploReg v4,[31] SnIPA,[32] and rVarBase.[33] GWAS3D was used to predict genetic variants or variants in linkage disequilibrium (LD) affecting regulatory pathways and essential disease/trait associations by integrating functional genomics, chromatin state, sequence motif, and conservation information.[34] It also provides visualization tools to complement the results. To assess correlations of SNPs with gene expression levels in peripheral blood cells, we used SNP and Copy Number Annotation (SCAN) database[35] and Blood eQTL[36] databases. For each SNP, the Blood eQTL database and SCAN provided information from experimental data on gene expression regulation in peripheral blood cells and lymphoblastoid cell lines, respectively. Likewise, we screened mQTLdb for methylation-QTLs (meQTL) to examine effects of sex-specific variants on CpG islands[37,38] at birth, during childhood and adolescence as well as during pregnancy. Statistically similar (proxy) SNP sets for each SNP were obtained from HaploReg v4,[31] and were submitted to the tools listed above for a number of analyses. For direct observation of TFBSs nearby SNPs, we examined the Swiss Regulon browser.[38] Gene set enrichment analysis (GSEA) was performed using DAVID v6.7.[39] The interactions of candidate genes with other genes and noncoding RNAs were investigated using NPinter v3.0,[40] which provides experimental data. Computational miRNA targets on protein-coding genes were explored using TargetScan.[41] To map the HLA-DQA1 SNPs to HLA-DQA1 types, we examined complete sequences of HLA-DQA1 alleles in the IPD-IMGT/HLA Database (http://www.ebi.ac.uk/ipd/imgt/hla).

### 2.6. The microarray innovations in leukemia (MILE) study

The MILE study generated microarray-based gene expression profiles from 2096 patients with standard subtypes of acute and chronic leukemia (and myelodysplastic syndromes) in 11 laboratories on 3 continents.[42] The aim of the study was to examine gene expression profiles for diagnosis and subclassification. The individual level data from the 2096 patients were available to compare expression levels of different genes of interest in different leukemia subtypes.

### 3. Results

After stringent QC, 209 subjects and 271,069 SNPs were included in the statistical analysis. In the final dataset, genotype call rates were >99.9%. All QC steps and their results are reported in Table 1. The final sample included 116 males (55.5%) and 93 females (44.5%) with no statistically significant difference in age distribution ($P=0.43$). The mean age was 4.43 ($\pm$2.58) years for males and 4.73 ($\pm$2.32) years for females. The genomic inflation factor ($\lambda$), an indicator of population stratification, was 1.0, which suggested no population stratification, as confirmed by the Q-Q plot (Supplementary Figure 1, http://links.lww.com/MD/B374). The design was ideal to detect sex-specific associations as the reflection of allele frequencies higher in one sex than the other. Naturally, the allele frequencies in the pooled sample would be in between sex-specific frequencies. We checked this assumption by using the frequencies for the European sample in the HapMap project (Tables 2 and 3). At least for the higher

### Table 1

<table>
<thead>
<tr>
<th>QC steps</th>
<th>Criteria for QC</th>
<th>Numbers filtered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Sex discordant</td>
<td>We used genetically identified sex information based on sex chromosomes instead of reported sex information</td>
<td>5</td>
</tr>
<tr>
<td>Sample call rate</td>
<td>Call rate &lt;99%</td>
<td>14</td>
</tr>
<tr>
<td>Heterozygosity rate</td>
<td>Autosomal heterozygosity rate outside of ±3 standard deviations from the mean</td>
<td>8</td>
</tr>
<tr>
<td>Duplicate</td>
<td>Samples excluded as duplicate (PI_HAT &gt; 90%)</td>
<td>0</td>
</tr>
<tr>
<td>SNP Genotype rate</td>
<td>Call rate &lt;99%</td>
<td>6261</td>
</tr>
<tr>
<td>MAF</td>
<td>MAF &lt;0.03</td>
<td>260,641</td>
</tr>
<tr>
<td>Mismissings between males and females</td>
<td>Differential missingsness between males and females with $P&lt;1.0 \times 10^{-3}$</td>
<td>478</td>
</tr>
<tr>
<td>HWE</td>
<td>HWE ($P&lt;10^{-3}$) was used in the whole group</td>
<td>0</td>
</tr>
</tbody>
</table>

HWE = Hardy-Weinberg equilibrium, MAF = minor allele frequency, QC = quality control, SNP = single nucleotide polymorphism.

* These SNP's had very low allele frequencies and would be noninformative as they show little variation in the sample set being analyzed. Statistical power to detect associations with them was insufficient.
Table 2

SNPs associated with sex-specific childhood ALL risk at the significance level $P<10^{-4}$.

<table>
<thead>
<tr>
<th>SNP</th>
<th>Gene/region</th>
<th>Chr</th>
<th>Minor allele</th>
<th>Location</th>
<th>MAF (males)</th>
<th>HapMap CEU MAF</th>
<th>MAF (females)</th>
<th>ORallele (95% CI)</th>
<th>P value $^1$</th>
<th>P value (permutation)$^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs4813720</td>
<td>RASSF2</td>
<td>10</td>
<td>A</td>
<td>Intron</td>
<td>0.23</td>
<td>0.37</td>
<td>0.44</td>
<td>0.30 (0.18–0.50)</td>
<td>3.85 $\times$ 10^{-6}</td>
<td>1.0 $\times$ 10^{-4}</td>
</tr>
<tr>
<td>rs251237</td>
<td>HSPA6/PROSER3</td>
<td>17</td>
<td>A</td>
<td>Intron</td>
<td>0.42</td>
<td>0.39</td>
<td>0.21</td>
<td>3.02 (1.87–4.89)</td>
<td>6.79 $\times$ 10^{-6}</td>
<td>1.0 $\times$ 10^{-4}</td>
</tr>
<tr>
<td>rs615369</td>
<td>10q24.31</td>
<td>10</td>
<td>G</td>
<td>Intron</td>
<td>0.17</td>
<td>0.32</td>
<td>0.36</td>
<td>0.30 (0.18–0.51)</td>
<td>7.42 $\times$ 10^{-6}</td>
<td>1.0 $\times$ 10^{-4}</td>
</tr>
</tbody>
</table>

$^1$ Interaction odds ratio per allele ($OR_{allele}$) for the additive model.

$^2$ P value adjusted for genonic control.

$^3$ P value permutation = point-wise P value from maxT permutation analysis after 10,000 permutation.

Table 3

SNPs with the highest effect sizes (OR $>10$; P $<0.05$) showing sex-specific associations with childhood ALL risk.

<table>
<thead>
<tr>
<th>SNP</th>
<th>Gene/region</th>
<th>Chr</th>
<th>Minor allele</th>
<th>Location</th>
<th>MAF (males)</th>
<th>HapMap CEU MAF</th>
<th>MAF (females)</th>
<th>ORallele (95% CI)</th>
<th>P value $^1$</th>
<th>P value (permutation)$^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs12722042</td>
<td>HLA-DQA1</td>
<td>6</td>
<td>G</td>
<td>Exon</td>
<td>0.07</td>
<td>0.04</td>
<td>0.05</td>
<td>14.72 (1.91–113.2)</td>
<td>0.008</td>
<td>6.0 $\times$ 10^{-6}</td>
</tr>
<tr>
<td>rs11905348</td>
<td>CD44</td>
<td>10</td>
<td>A</td>
<td>Intron</td>
<td>0.07</td>
<td>0.04</td>
<td>0.05</td>
<td>14.72 (1.91–113.9)</td>
<td>0.01</td>
<td>0.001</td>
</tr>
<tr>
<td>rs12722009</td>
<td>HLA-DQA1</td>
<td>6</td>
<td>A</td>
<td>Exon</td>
<td>0.07</td>
<td>0.04</td>
<td>0.05</td>
<td>14.08 (1.84–107.9)</td>
<td>0.009</td>
<td>5.0 $\times$ 10^{-4}</td>
</tr>
<tr>
<td>rs80040022</td>
<td>UMODL1</td>
<td>21</td>
<td>A</td>
<td>Exon</td>
<td>0.06</td>
<td>0.04</td>
<td>0.05</td>
<td>13.66 (1.77–105.5)</td>
<td>0.01</td>
<td>0.001</td>
</tr>
<tr>
<td>rs61753605</td>
<td>PRM2</td>
<td>6</td>
<td>C</td>
<td>Exon</td>
<td>0.06</td>
<td>0.02</td>
<td>0.03</td>
<td>13.37 (1.73–103.2)</td>
<td>0.01</td>
<td>0.002</td>
</tr>
<tr>
<td>rs35666085</td>
<td>OSGR5</td>
<td>22</td>
<td>G</td>
<td>Exon</td>
<td>0.06</td>
<td>0.05</td>
<td>0.06</td>
<td>12.63 (1.63–97.9)</td>
<td>0.02</td>
<td>0.003</td>
</tr>
<tr>
<td>rs143021649</td>
<td>CTNNB1</td>
<td>3</td>
<td>T</td>
<td>Intron</td>
<td>0.06</td>
<td>0.04</td>
<td>0.05</td>
<td>11.61 (1.49–90.5)</td>
<td>0.02</td>
<td>0.003</td>
</tr>
<tr>
<td>rs6799524</td>
<td>PRSS1</td>
<td>3</td>
<td>G</td>
<td>Intron</td>
<td>0.05</td>
<td>0.006</td>
<td>0.03</td>
<td>10.62 (1.35–83.2)</td>
<td>0.02</td>
<td>0.008</td>
</tr>
<tr>
<td>rs10003468</td>
<td>CDS2</td>
<td>4</td>
<td>G</td>
<td>Intron</td>
<td>0.05</td>
<td>0.02</td>
<td>0.05</td>
<td>10.62 (1.35–83.2)</td>
<td>0.02</td>
<td>0.006</td>
</tr>
</tbody>
</table>

$^1$ Interaction odds ratio per allele ($OR_{allele}$) for the additive model.

$^2$ P value adjusted for genonic control.

$^3$ P value permutation = point-wise P value from maxT permutation analysis after 10,000 permutation.

All = acute lymphoblastic leukemia, CEU = European sample in HapMap project, Chr = chromosome, MAF = minor allele frequency, SNP = single nucleotide polymorphism.

*point-wise*
microRNA binding site in 3’ UTR. To gain further insight into RASSF2 function, we obtained the list of experimentally confirmed target gene list for mir-19a/19b (n=298) from NPinter database, and subjected this list to GSEA on DAVID. The highest fold enrichment (12.4) was observed for RAS association category, which was also the statistically most significant (false discovery rate [FDR]=0.004). Apoptosis was also one of the most highly significant categories (fold enrichment=3.5; FDR=0.007). Thus, the only microRNA that had an experimentally confirmed physical interaction with RASSF2 had targets cumulatively involved in RAS signaling and apoptosis (mir-19a/19b belongs to the mir-17-92 cluster).

The second ranking SNP rs23127 (OR\textsubscript{Interaction}=3.0), located in the genes PROSER3/HSPB6, yielded a high risk for males compared with females for childhood ALL (Table 2).

---

**Table 4**

<table>
<thead>
<tr>
<th>SNP</th>
<th>Target CpG ID (Gene)</th>
<th>meQTL effect size (regression beta value)</th>
<th>ALL mean methylation</th>
<th>pre-B cell mean methylation</th>
<th>DNA methylation status in ALL</th>
<th>Genetic association effect size (odds ratio)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs4137207†</td>
<td>cg22485289 (RASSF2)</td>
<td>8.3 x 10\textsuperscript{-2}</td>
<td>0.432</td>
<td>0.279</td>
<td>0.764</td>
<td>Hypomethylation</td>
</tr>
<tr>
<td>rs810669†</td>
<td>cg14120049 (HSPB6)</td>
<td>6.6 x 10\textsuperscript{-2}</td>
<td>0.867</td>
<td>0.651</td>
<td>0.344</td>
<td>Hypermethylation</td>
</tr>
<tr>
<td>rs1706505†</td>
<td>cg24363200 (SNQ1)</td>
<td>1.4 x 10\textsuperscript{-2}</td>
<td>0.591</td>
<td>0.505</td>
<td>0.873</td>
<td>Hypomethylation</td>
</tr>
<tr>
<td>rs7402818†</td>
<td>cg11957475 (RASSF1)</td>
<td>8.4 x 10\textsuperscript{-2}</td>
<td>0.420</td>
<td>0.863</td>
<td>0.600</td>
<td>Hypermethylation</td>
</tr>
<tr>
<td>rs1101225†</td>
<td>cg25427504</td>
<td>5.1 x 10\textsuperscript{-2}</td>
<td>-0.453</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>rs1101225†</td>
<td>cg02535924</td>
<td>1.6 x 10\textsuperscript{-2}</td>
<td>0.472</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
</tr>
</tbody>
</table>

**Table 4**

SNPs with sex-specific associations, their meQTL status (for childhood period) and target CpG sites,\textsuperscript{[47]} and the status of DNA methylation in ALL and healthy pre-B cells.\textsuperscript{[60]}
Table 3 and Supplementary Table 2 (http://links.lww.com/MD/B376) show the results based on the highest effect sizes. Two statistically correlated SNPs ($r^2 = 1$), rs12722042 (CADD = 16.02 and DANN = 1) and rs12722039 (CADD = 5.02 and DANN = 0.49), from the HLA-DQA1 (major histocompatibility complex, class II, DQ alpha 1) gene and a SNP, rs11992342, from the ADAM28 (ADAM metallopeptidase domain 28) region yielded highest effect sizes (OR_HLM/controls = 14; P ≤ 0.01) for sex-specific results. Both HLA-DQA1 SNPs were missense variants and marked by several histone modification marks, such as Histone-3 lysine-4 monomethylation (H3K4me1), Histone-3 lysine-4 trimethylation (H3K4me3), Histone-3 lysine-9 acetylation (H3K9ac), and Histone-3 lysine-27 acetylation (H3K27ac) indicating the location within an active promoter site. When we examined whether there are crucial TFBSs in the vicinity of these SNPs located at chromosome 6 positions 32,605,284 and 32,605,309 (hg19 coordinates), Swiss Regulon browser indicated that there was an androgen receptor binding site within 100 bp of these SNPs (32,605,225 to 32,605,2546). ADAM28 rs11992342 maps to an active enhancer region, and affects binding affinity of several TFs of STAT family. Top ranking 100 results based on OR included associations with more missense variants than the results based on P values (34 vs 2) (Supplementary Tables 1 and 2, http://links.lww.com/MD/B375, http://links.lww.com/MD/B376). Our results predicted several of these missense variants may have deleterious effects on final protein structure (Supplementary Table 4, http://links.lww.com/MD/B378). Most of the identified sex-specific SNPs were located in cis-regulatory elements, such as promoters and enhancers or may have eQTL effects (Supplementary Tables 1 and 2, http://links.lww.com/MD/B375, http://links.lww.com/MD/B376).

Our mQTLdb screening results showed that several variants were meQTLs at birth, during childhood and adolescence, as well as during pregnancy (Supplementary Table 5, http://links.lww.com/MD/B376). We examined the involvement of the CpG sites linked to our GWAS associations in B-cell differentiation and leukemogenesis by using the data provided by Kulis et al.[30] A total of 24 variants were reported to be pivotal for B-lymphocyte differentiation and their target CpG sites found to be either hypermethylated or hypomethylated in ALL (Table 4). Most notably, the top ranking SNP rs4813720 is an meQTL for the CpG island cg22485289 within RASSF2. Both HLA-DQA1 SNPs (rs12722042 and rs12722039) are meQTLs for the CpG island cg24593918 within HLA-DQB1.

GWA3D analysis for the top 100 ranking sets of SNPs unraveled a number of long-range interactions between the regions of SNP locations and other genes, some of which being on different chromosomes (Supplementary Figures 2 and 3, http://links.lww.com/MD/B374). GWA3D reported 84 variants with TF binding affinity changes or regulatory signals based on the GM12878 cell line and HapMap CEU population for 100 statistically most significant results. A total of 66 variants have been detected affecting long-range interactions (Supplementary Figure 2, http://links.lww.com/MD/B374). Cumulatively, GWA3D results suggested a potential role for STAT and MYC TFs in sex-differences in childhood ALL risk. Among the top 100 results ranked by the effect size, GWA3D analysis showed that 68 of the variants were associated with regulatory signals and TF binding affinity changes on the GM12878 cell line and HapMap CEU population. Long-range interactions were noted for 60 variants (Supplementary Figure 3, http://links.lww.com/MD/B374).

As shown in Table 5, in all ALL subtypes examined in the MILE study, RASSF2 expression was lower than controls, and

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>MIR17 cluster</th>
<th>MIR17HG cluster</th>
<th>RASSF2</th>
<th>MYC</th>
</tr>
</thead>
<tbody>
<tr>
<td>P value</td>
<td>P value</td>
<td>Ratio</td>
<td>P value</td>
<td>Ratio</td>
</tr>
<tr>
<td>ALL with all subtypes</td>
<td>1.78</td>
<td>0.001</td>
<td>1.36</td>
<td>0.08</td>
</tr>
<tr>
<td>T-ALL</td>
<td>2.5 x 10^-6</td>
<td>0.02</td>
<td>1.35</td>
<td>0.12</td>
</tr>
<tr>
<td>c-ALL/Pre-B-ALL with 1 (B/L)</td>
<td>3.1 x 10^-6</td>
<td>0.002</td>
<td>1.15</td>
<td>0.03</td>
</tr>
<tr>
<td>MATURE B-ALL with 1 (B/L)</td>
<td>3.1 x 10^-6</td>
<td>0.002</td>
<td>1.15</td>
<td>0.03</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>MIR17 cluster</th>
<th>MIR17HG cluster</th>
<th>RASSF2</th>
<th>MYC</th>
</tr>
</thead>
<tbody>
<tr>
<td>P value</td>
<td>P value</td>
<td>Ratio</td>
<td>P value</td>
<td>Ratio</td>
</tr>
<tr>
<td>ALL with all subtypes</td>
<td>1.78</td>
<td>0.001</td>
<td>1.36</td>
<td>0.08</td>
</tr>
<tr>
<td>T-ALL</td>
<td>2.5 x 10^-6</td>
<td>0.02</td>
<td>1.35</td>
<td>0.12</td>
</tr>
<tr>
<td>c-ALL/Pre-B-ALL with 1 (B/L)</td>
<td>3.1 x 10^-6</td>
<td>0.002</td>
<td>1.15</td>
<td>0.03</td>
</tr>
<tr>
<td>MATURE B-ALL with 1 (B/L)</td>
<td>3.1 x 10^-6</td>
<td>0.002</td>
<td>1.15</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Expression differences in mir-17-92, RASSF2, and MYC in acute lymphoblastic leukemia subtypes in the MILE study.

Table 6

[Table 6](http://links.lww.com/MD/B374)
mir-17-92 levels were higher with statistical significance varying from $P = 0.02$ to $6.8 \times 10^{-0.80}$. The statistically most significant change in RASSF2 was in T-ALL with the expression ratio to controls being 0.23 (mir-17-92 showed an expression ratio of 1.71 in T-ALL). The statistically most significant change in mir-17-92 was in ALL with t(12;21) with the expression ratio to controls being 3.17 (RASSF2 showed an expression ratio of 0.50 in the same ALL subtype). In all subtypes, RASSF2 and mir-17-92 expression levels showed an inverse correlation (Table 5). The expression levels of MYC did not appear to correlate with mir-17-92 levels except in pro-B and mature B-ALL subtypes (Table 5). While the strong inverse correlation between RASSF2 and mir-17-92 confirmed the expected relationship between them, the lack of a correlation between MYC and mir-17-92 suggested that mir-17-92 increase in ALL is not exclusively caused by MYC, but may be due to genomic amplifications as observed in other malignancies.[41]

4. Discussion

This is the first GWAS designed to identify sex-specific childhood ALL risk markers. As a second exploratory approach, we also analyzed results based on OR, to examine clinically meaningful results that may not have reached the strict statistical significance due to their lower frequencies. These results are free from multiple comparison issues, as each permutation on each SNP assesses the role of chance.

The statistically most significant association was with an intronic SNP in RASSF2, which is a novel tumor suppressor gene and a member of the RAS family which regulates a wide range of biological processes, including KRAS signaling.[46,47] It induces apoptosis and cell cycle arrest, shows frequent methylation in several cancers, and rs4813720 correlates with RASSF2 expression levels.[48] RASSF2 ablation downregulates genes involved in the immune response, hematological development, as well as genes activating nuclear factor (NF)-κB signaling.[46]

RASSF2 has emerged as a candidate gene involved in sex-specific modification of risk for childhood ALL in the present study for the first time. This result is biologically plausible. RASSF2 is a tumor suppressor gene via its inhibitory effect on the proto-oncogene KRAS.[49] KRAS is one of the most frequently mutated proto-oncogenes in childhood ALL.[44,49] The top ranking sex-specific risk marker for childhood ALL, rs4813720, correlates with RASSF2 expression level[48] as well as the methylation levels in a key CpG site cg22485289 within the promoter of RASSF2 at birth and childhood.[17] This particular CpG site is one of the key sites found to be hypomethylated in childhood ALL compared with pre-B cells.[50] Thus, rs4813720 appears to be involved in the prevention of childhood ALL in males via maintaining the expression levels of tumor suppressive RASSF2 in pre-B cells presumably via a methylation-related mechanism. Linkage of DNA methylation quantitative trait loci to human cancer risk is already well documented,[48] and our results represent yet another example.

The sex effect in the prevention of childhood ALL by RASSF2 expression may also be biologically plausible. The key to the sex effect may be the events upstream of RASSF2. The mir-17-92 cluster, which is in an amplified genomic region in human malignant B-cell lymphomas, has been shown to be an oncomiRNA also in c-Myc-induced lymphomagenesis in mice.[51] mir-17-92 also plays a major role in normal B-cell development.[52,53] In an experimental study, RASSF2 has been found to be a target gene for the mir-92 polycistron.[54] Our bioinformatics work also confirmed a physical interaction with one of the micro RNAs that derive from mir-17-92 polycistron, mir-19a/19b, and revealed a predicted binding site for it (see Results). Crucially, estrogen administration increases mir-17-92 levels via MYC overexpression[55,56] and downregulates RASSF2 expression.[57] Thus, in addition to the already known involvement of sex hormones in the gender effect in cancer,[58] we propose a novel mechanism for childhood ALL: estrogen exposure increases the oncogenic microRNA species mir-19a/19b levels, which in turn downregulates RASSF2 (as well as another tumor suppressor PTEN[59]). Downregulation of RASSF2 abolishes KRAS inhibition, and the KRAS oncogene contributes to childhood ALL development. Males are already less susceptible to this mechanism due to lesser exposure to estrogen, as well as due to the presence of binding sites to Y chromosome-linked transcription factor SRY,[47] and are further protected if they possess the minor allele of rs4813720, which increases RASSF2 expression levels. If experimentally confirmed, this would be the first demonstration of the mechanism of a sexually antagonistic association of a SNP with cancer susceptibility. Although sex hormone levels are very low during childhood in both males and females, programming of autosomal gene expression by sex hormones during prenatal development or during the temporary androgen surge in early infancy (called mini-puberty) is a possibility. Mini-puberty is particularly interesting as inter individual variability in sex hormone levels, which may rise to the puberty levels but temporarily, may be very remarkable.[60,61] Unfortunately, there are no studies on the long-term effects of these sex hormone exposures in early infancy on biological systems although behavioral studies continue to show the relevance of this exposure.[60,62]

Having a marker that is associated with prevention for childhood ALL in males has clinical importance. Males develop childhood ALL more frequently, have a higher relapse rate, and worse prognosis.[41] Due to having clinically more aggressive form of the disease, they also receive more intensive treatment resulting in higher toxicity. The marker is within RASSF2 and acts by increasing its expression, which in turn, binds to and inactivates KRAS. KRAS is frequently mutated in childhood ALL[44,49] and mutant KRAS worsens the prognosis.[49] Thus, any pharmacological intervention to increase the expression of RASSF2 should be beneficial for male patients with childhood ALL, especially for those with KRAS mutations, to counter the deleterious effects of mutant KRAS.

Examination of associations with the highest odds ratios identified a pair of HLA region SNPs. The HLA-DQA1 SNPs (rs12722042 and rs12722039) belong to DQA1*01, confirming the previously reported male-specific association with DQA1*01, and supporting the proposed infection-related etiology in childhood ALL risk for males.[63] Both HLA-DQA1 SNPs are located in an active promoter region and showed eQTL effects on the HLA-DQB1 gene. These SNPs have regulatory effects on gene expression rather than on protein structure. Intriguingly, the HLA-DQA1 SNPs are in close proximity (<100 bp) of an androgen receptor binding site. This observation may be relevant in their risk associations with childhood ALL in males. The SNP rs33597309, correlated statistically with rs12722042 ($r^2 = 0.79$), has shown a risk association with esophageal squamous cell carcinoma in a Chinese population[64] (esophageal carcinoma has a high male-to-female ratio).

The validity of the case-only design depends on the independence of exposure and genotype.[24] Departures from this
independence may affect the results. Such an association would be invalid due to the violation of the major assumption of case-only approach. In the case of gender, it seems unlikely to have different genotype frequencies in autosomal chromosomes between males and females in the healthy population. Conforming to the assumptions of the case-control design still does not make the results valid until replicated by a second independent study. The lack of a previously observed association of RASSF2 SNPs with childhood ALL does not argue against the validity of our result. This is most likely due to cancelling of associations in opposite directions in males and females when the analysis is performed on the overall sample. None of the previous childhood ALL GWAS performed sex-specific analyses. Despite the biological plausibility of our results, given the limited sample size of our study and lack of replication, it is important to reexamine these results in an independent study before proceeding to functional studies. These results cannot be generalized in any other ethnic population, as all of these samples were non-Hispanic whites.

In conclusion, our results suggest that novel sex-specific risk variants for childhood ALL exist. Functional analyses suggest that most variants (either direct effect or through LD) have regulatory effects, which increases the likelihood of causal associations. For instance, the statistically most significant association rs4813720 correlates with RASSF2 expression and may play a significant role in childhood ALL etiology. Our results also confirm the previously observed male-specific association with DQA1*01. While ranking results by OR is not a traditional approach for GWAS, this exploratory approach yielded several missense variants with effects on protein function. The statistically most significant association was with a SNP within RASSF2, which interacts with KRAS. Given the difficulties of developing drugs to inactivate KRAS directly, the demonstration of the effect of RASSF2 in prevention from childhood ALL may even offer a new target for drug development efforts. Feasibility of countering the downstream effects of mir-17 may even offer a new target for drug development efforts.

While ranking results by OR is not a traditional approach for GWAS, this exploratory approach yielded several missense variants with effects on protein function. The statistically most significant association was with a SNP within RASSF2, which interacts with KRAS. Given the difficulties of developing drugs to inactivate KRAS directly, the demonstration of the effect of RASSF2 in prevention from childhood ALL may even offer a new target for drug development efforts. Feasibility of countering the downstream effects of mir-17 may even offer a new target for drug development efforts.

References


