Genome-wide association study identifies susceptibility loci for B-cell childhood acute lymphoblastic leukemia

Jayaram Vijayakrishnan1, James Studd1, Peter Broderick1, Ben Kinnersley1, Amy Holroyd1, Philip J. Law1, Rajiv Kumar2, James M. Allan3, Christine J. Harrison4, Anthony V. Moorman4, Ajay Vora5, Eve Roman6, Sivaramakrishna Rachakonda2, Sally E. Kinsey7, Eamonn Sheridan8, Pamela D. Thompson9, Julie A. Irving3, Rolf Koehler10, Per Hoffmann11,12, Markus M. Nöthen11, Stefanie Heilmann-Heimbach11, Karl-Heinz Jöckel13, Douglas F. Easton14,15, Paul D.P. Pharaoh14,15, Alison M. Dunning16, Julian Peto17, Frederico Canzian18, Anthony Swerdlow1,19, Rosalind A. Eeles1,20, ZSoa Kote-Jarai1, Kenneth Muir21,22, Nora Pashayan15,23, The PRACTICAL consortium, Mel Greaves24, Martin Zimmerman25, Claus R. Bartram10, Martin Schrappe26, Martin Stanulla25, Kari Hemminki2,27 & Richard S. Houlston1

Genome-wide association studies (GWAS) have advanced our understanding of susceptibility to B-cell precursor acute lymphoblastic leukemia (BCP-ALL); however, much of the heritable risk remains unidentified. Here, we perform a GWAS and conduct a meta-analysis with two existing GWAS, totaling 2442 cases and 14,609 controls. We identify risk loci for BCP-ALL at 8q24.21 (rs28665337, \( P = 3.86 \times 10^{-9} \), odds ratio (OR) = 1.34) and for ETV6-RUNX1 fusion-positive BCP-ALL at 2q22.3 (rs17481869, \( P = 3.20 \times 10^{-8} \), OR = 2.14). Our findings provide further insights into genetic susceptibility to ALL and its biology.
Acute lymphoblastic leukemia (ALL) is the most common pediatric cancer in western countries, of which B-cell precursor acute lymphoblastic leukemia (BCP-ALL) accounts for approximately 80% of cases. The etiology of ALL is poorly understood and no specific environmental risk factor has so far been identified aside from indirect evidence for an infective origin. Independent of concordance disease in monozygotic twins, which has an in utero origin evidence, albeit indirect, for inherited predisposition to ALL is provided by the elevated risk seen in siblings of ALL cases. Previous genome-wide association studies (GWAS) have suggested susceptibility to ALL is polygenic, identifying single-nucleotide polymorphisms (SNPs) in eight loci influencing ALL risk at 7p11.22 (IKZF1), 9p21.3 (CDKN2A), 10p12.2 (PIPK2A), 10q26.13 (LHPP), 12q23.1 (ELK3), 10p14 (GATA3), 10q21.2 (ARID5B), and 14q11.2 (CEBPE). ALL is biologically heterogeneous and subtype associations have been identified for 10q21.2 (ARID5B) associated with high-hyperdiploid BCP-ALL (i.e., >50 chromosomes) and 10p14 (GATA3) associated with Ph-like BCP-ALL.

Statistical modeling of GWAS data indicates that much of the heritable risk of ALL ascribable to common genetic variation remains to be discovered. To gain a more comprehensive insight into predisposition to ALL we performed a meta-analysis of two previously published GWAS and a new GWAS together totaling 2442 cases and 14,609 controls. We report two previously unidentified risk loci, providing further insights into the genetic and biological basis of this disease.

**Results**

**Association analysis.** We analyzed data from three studies of European ancestry: a new GWAS from the United Kingdom–UK GWAS II, and two previously reported GWAS–UK GWAS I and a German GWAS (Supplementary Figs. 1, 2 and Supplementary Table 1). After imposing pre-determined (see “Methods”) quality metrics to each of the three GWAS, the studies provided genotype data on 2442 cases and 14,609 controls. To increase genomic resolution, we imputed >10 million SNPs using whole-genome reference genotype data from 1000 Genomes Project (n = 1092) and UK10K (n = 3781). Quantile-quantile plots of SNPs (minor allele frequency (MAF) > 0.01) post-imputation showed evi dence of substantive over-dispersion introduced by imputation (genomic inflation λ for UK GWAS I, UK GWAS II, and German GWAS were 1.02, 1.05, and 1.01, respectively; Supplementary Fig. 3).

Pooling data from the three GWAS, we derived joint odds ratios (ORs), 95% confidence intervals (CIs), and associated per allele P-values under a fixed-effects model for each SNP with MAF > 0.01. Given the biological heterogeneity of BCP-ALL, overall and subtype-specific ORs were derived for BCP-ALL, high-hyperdiploid ALL (i.e., >50 chromosomes), and ETV6–RUNX1 fusion-positive BCP-ALL. This combined meta-analysis further substantiated previously published risk SNPs (Fig. 1, Supplementary Table 2). In addition to previously reported loci we identified three risk loci for BCP-ALL at 8q24.21 (rs28665337, hg19 chr8:g.130194104) and 5q21.3 (rs7449087, hg19 chr5: g.107928071), and for ETV6–RUNX1–positive ALL at 2q22.3 (rs17481869, hg19 chr2:g.146124454) (Fig. 2). The ETV6–RUNX1–positive ALL at 2q22.3 was imputed (info score > 0.97) in all three data sets, imputation fidelity was confirmed through Sanger sequencing in a subset of samples (r² = 0.98, Supplementary Table 4). The fidelity of imputation of SNP rs7449087 was poor (r² = 0.81) with no correlated directly typed SNP with P-value < 1 × 10⁻⁶; hence we did not consider this represented a bona fide association (Supplementary Table 4). Conditional analysis did not provide evidence for multiple independent signals at either 8q24.21 or 2q22.3.

The 8q24.21 variant rs28665337 maps 35 kb 3’ of the long intergenic non-coding RNA 977 (LINC00977, Fig. 2). The 8q24.21 region harbors variants associated with multiple cancers, including colorectal, prostate, bladder cancer also B-cell malignancies such as diffuse large B-cell lymphoma, Hodgkin lymphoma, and chronic lymphocytic leukemia (Supplementary Table 5). The linkage disequilibrium (LD) blocks delineating these cancer risk loci are distinct from the 8q24.21 BCP-ALL association signal suggesting this risk locus is unique to BCP-ALL (pairwise LD metrics r² < 0.2; Supplementary Table 5).

The region harbors variants associated with multiple cancers, including colorectal, prostate, bladder cancer also B-cell malignancies such as diffuse large B-cell lymphoma, Hodgkin lymphoma, and chronic lymphocytic leukemia (Supplementary Table 5). The linkage disequilibrium (LD) blocks delineating these cancer risk loci are distinct from the 8q24.21 BCP-ALL association signal suggesting this risk locus is unique to BCP-ALL (pairwise LD metrics r² < 0.2; Supplementary Table 5).

**Relationship between SNP genotype and patient outcome.** We examined the relationship between SNP genotype and patient outcome using data from UK GWAS II and German GWAS. Neither rs28665337 or rs17481869 showed a consistent association with either event-free survival (EFS) or risk of relapse, even when stratified by ETV6–RUNX1 status (Supplementary Table 6).

**Functional annotation of risk loci.** To gain insight into the biological basis underlying the association signals at these as well as previously identified risk loci, we examined the epigenetic landscape of BCP-ALL risk loci genome wide. For each risk locus we evaluated profiles of three histone marks of active chromatin

---

**Fig. 1** Manhattan plot of association. y-axis shows genome-wide P-values (two-sided, calculated using SNPTEST v2.5.2 assuming an additive model) of >6 million successfully imputed autosomal SNPs in 2442 cases and 14,609 controls. The x-axis shows the chromosome number. The red horizontal line represents the genome-wide significance threshold of P = 5.0 × 10⁻⁸.
Figure 2: Regional plots of association results and recombination rates for the identified risk loci. a) 8q24.21 (rs28665337), b) 2q22.3 (rs17481869). Plots (generated using visPIG\(^{14}\)) show association \(^{-}\log_{10}\)P-values (left y-axis) of genotyped (triangles) and imputed (circles) SNPs in the GWAS samples (2442 cases and 14,609 controls) and recombination rates (right y-axis). \(^{-}\log_{10}\)P-values were calculated assuming an additive model in SNPTEST v2.5.2 and are shown according to their chromosomal positions (x-axis). Lead SNPs are denoted by large circles or triangles labeled by rsID. Color intensity of each symbol reflects LD, white (\(r^2 = 0\)), dark red (\(r^2 = 1.0\)). Light blue line shows recombination rates from UK10K Genomes Project. Genome coordinates are from NCBI human genome GRCh37.

(H3K27ac, H3K4me1, and H3k4me3) using ChIP-seq data of 14 cell types from ENCODE, including lymphoblastoid cell line (GM12878), and multiple ALL and acute myeloid leukemia (AML) samples from the Blue-Print Epigenome database (Supplementary Fig. 4, Supplementary Table 7). Since the strongest associated GWAS SNP may not represent the causal variant, we examined signals across an interval spanning all variants in LD with the most strongly associated SNP at each risk loci (\(r^2 > 0.8\) and D' > 0.8 based on the 1000 Genomes EUR reference panel). The analysis across all risk loci combined revealed that risk SNPs are enriched for markers of open chromatin and that enrichment is highest in ALL cells (Supplementary Fig. 4, Supplementary Table 7). Analysis using HaploReg\(^{17}\) revealed a significant enrichment of SNPs within enhancers in primary hematopoietic stem cells (binomial test for enrichment, \(P = 0.0034\); Supplementary Data 1). Collectively these data support a model of disease etiology where risk loci influence BCP-ALL risk through cis regulatory effects on transcription.

We used summary-level Mendelian randomization (SMR) analysis to test for concordance between GWAS and cis-eQTL-associated SNPs with all correlated SNPs (\(r^2 > 0.8\)) within 1 Mb of the lead SNP at each locus (Supplementary Tables 8 and 9) deriving \(b_{XY}\) statistics, which estimate the effect of gene expression on childhood ALL risk. This analysis showed variation in the expression of CDKN2B, FAM53B, FIGNL1, and PIP5K2A were associated with risk loci (Supplementary Fig. 5, Supplementary Tables 8 and 9). Eight gene probes exceeded the \(P_{\text{SMR}}\) threshold of 1.3 \times 10^{-4}, of which two genes passed the HEIDTI test for heterogeneity (\(P_{\text{HEIDI}} > 0.05\)). In whole blood-derived tissue, the 10q26.13 locus was associated with FAM53B expression and the 10p12.2 locus was associated with PIP4K2A (alias PIP5K2A) expression (\(P_{\text{SMR}} = 2.09 \times 10^{-4}\), \(b_{XY} = -0.99\), and \(P_{\text{SMR}} = 7.48 \times 10^{-4}\), \(b_{XY} = 0.32\), respectively; Supplementary Fig. 5, Supplementary Table 9). Following from SMR analysis we also investigated whether the most strongly associated SNP at each risk locus, individually, was associated with the expression of genes within a 2 MB window to ensure capture of long range interactions. This provided evidence for a relationship between the 8q24.21 risk allele (rs28665337) and increased expression of MYC (t-test, \(P = 7.20 \times 10^{-4}\); Supplementary Fig. 6, Supplementary Table 10), and the 2q22 risk allele (rs17481869) with decreased GTDC1 expression (t-test, \(P = 0.037\); Supplementary Fig. 6, Supplementary Table 10). Since chromatin looping interactions are fundamental for regulation of gene expression, we interrogated physical interactions at respective genomic regions defined by rs28665337 and rs17481869 in GM12878 lymphoblastoid and H1 human embryonic stem (ES) cells using Hi-C data. Acknowledging limitations that these cell types may not fully reflect ALL biology, the regions containing rs28665337 and rs17481869 show significant chromatin looping interactions with the promoter regions of MYC in ES cells and GTDC1 in GM12878, respectively (Fit-Hi-C test\(^{18}\), Supplementary Figs. 7, 8).

**HLA alleles and risk.** A relationship between variation within the major histocompatibility complex (MHC) region and risk of ALL has long been speculated\(^{19-26}\). However, most studies have failed to address the complex LD patterns within the MHC or issues relating to population stratification. In view of the inconsistencies and limitations of published studies we conducted a more rigorous
analysis. Specifically, we investigated a possible relationship between BCP-ALL risk and HLA alleles by computing the 6p21 region using the Type I Diabetes Genetics Consortium (T1DGC) as reference. The strongest association from a combined analysis of all three GWAS was provided by SNP rs9469021, which maps 167 Kb centromeric to HLA-B (combined \( P = 3.5 \times 10^{-3} \); frequentist test of association using SNPTEST); this association was, however, not significant after correcting for multiple testing.

Impact on heritable risk. Using genome-wide complex trait analysis (GCTA) the heritability of BCP-ALL accounted for by common variants was estimated to be 0.16 (standard error (S. E.) 0.03, REML analysis \( P_\text{meta} = 4.25 \times 10^{-8} \)) with little evidence for subtype difference (0.18 ± S.E. 0.05 and 0.20 ± S.E. 0.08 for hyperdiploid and \( ETV6\)-positive BCP-ALL, respectively). The 11 known susceptibility variants account for 34% of the familial risk (Supplementary Table 11). The impact of BCP-ALL SNPs are among the strongest GWAS associations of any malignancy, raising the possibility of clinical utility for risk prediction. To examine this, we generated polygenic risk scores (PRS) based on the composite effect of all risk SNPs assuming a log-normal relative risk distribution. Using this approach for all risk SNPs, individuals in the top 1% of genetic risk had a 7.5-fold relative risk of BCP-ALL (Supplementary Fig. 9). The individual risk discrimination provided by the variants is shown in the log-normal relative risk distribution. Using this approach for all risk SNPs assuming a polygenic risk score (PRS) based on the composite effect of all risk SNPs assuming a log-normal relative risk distribution. Using this approach for all risk SNPs, individuals in the top 1% of genetic risk had a 7.5-fold relative risk of BCP-ALL (Supplementary Fig. 9).

Discussion

The evidence for the two risk loci we report has been based on a meta-analysis of three independent GWAS data sets. While the combined association \( P \)-values for each risk locus is genome-wide significant with each series providing support for association we acknowledge that we did not provide additional replication. For rare cancers such as childhood ALL, ascertaining case series which are appropriately ethnically matched and are sufficiently powered to provide independent replication is inherently problematic. Moreover as exemplified by the 10q21 and 10p14 risk loci, associations can be highly subtype-specific which adds to the difficulty in obtaining appropriate replication series. Accepting such caveats our analysis provides evidence for the existence of two additional risk loci for childhood BCP-ALL at 2q22.3 and 8q24.21.

We did not observe an association between risk SNPs at either 2q22.3 and 8q24.21 with patient survival. This is consistent with the impact of risk variants operating at an early stage of ALL evolution rather than disease progression per se. We acknowledge this analysis only has power to demonstrate a 10% difference in patient outcome. To robustly determine the relationship between genotype and outcome requires larger patient cohorts.

Given the existence of different subtypes of BCP-ALL, presumably reflecting the different etiology and evolutionary trajectories, it is perhaps not surprising that some SNPs display subtype-specific effects. Notable in this respect are the 10q21.2 and 10p14 variants that specifically influence high-hyperdiploid BCP-ALL and \( ETV6\)-like ALL, respectively. As with 7p12.2, 9p21.3, 10p12.2, 14q11.2, and the currently identified 8q24.21 locus has generic effects on the risk of BCP-ALL. In contrast the 2q22.3 association was highly specific for \( ETV6\)-positive BCP-ALL.

Deregulation of \( MYC \) has been reported in ALL, in some instances as a consequence of chromosomal rearrangement. Studies in other cancers have shown that disease-specific risk loci at 8q24.21 lie within tissue-specific enhancers interacting with \( MYC \) or \( PVT1 \) promoters. Furthermore, recent Hi-C analysis of this region has demonstrated a complicated 3D structure implicating various lincRNAs in mediating risk. Hence, it is plausible that the susceptibility to ALL has a similar mechanistic basis, brought about through involvement of the lincRNA 00977.

Risk conferred by rs17481869 (2q22.3) was specific to \( ETV6\)-positive BCP-ALL. The SNP association is intergenic with no obvious candidate gene in the vicinity, presently hindering the suggestion of testable hypotheses regarding its functional basis. eQTL data does, however, provide evidence implicating \( GTDC1 \). \( GTDC1 \) encodes a glucosyltransferase whose expression is relatively high in peripheral blood leukocytes. Chromosomal rearrangements of \( MLL \) (mixed lineage leukemia)

| Table 1 rs28665337 (8q24.21) genotypes and risk associated with BCP-ALL, high-hyperdiploid, and \( ETV6\)-RUNX1-positive childhood BCP-ALL subtypes |

<table>
<thead>
<tr>
<th>All BCP-ALL</th>
<th>RAF</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cases</td>
<td>Controls</td>
</tr>
<tr>
<td>UK GWAS I</td>
<td>0.15</td>
<td>0.12</td>
</tr>
<tr>
<td>German GWAS</td>
<td>0.16</td>
<td>0.12</td>
</tr>
<tr>
<td>UK GWAS II</td>
<td>0.15</td>
<td>0.12</td>
</tr>
<tr>
<td>Combined</td>
<td>0.15</td>
<td>0.12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>High-hyperdiploid</th>
<th>RAF</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cases</td>
<td>Controls</td>
</tr>
<tr>
<td>UK GWAS I</td>
<td>0.15</td>
<td>0.12</td>
</tr>
<tr>
<td>German GWAS</td>
<td>0.17</td>
<td>0.12</td>
</tr>
<tr>
<td>UK GWAS II</td>
<td>0.15</td>
<td>0.12</td>
</tr>
<tr>
<td>Combined</td>
<td>0.15</td>
<td>0.12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( ETV6)-RUNX1-positive</th>
<th>RAF</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cases</td>
<td>Controls</td>
</tr>
<tr>
<td>UK GWAS I</td>
<td>0.16</td>
<td>0.12</td>
</tr>
<tr>
<td>German GWAS</td>
<td>0.09</td>
<td>0.12</td>
</tr>
<tr>
<td>UK GWAS II</td>
<td>0.14</td>
<td>0.12</td>
</tr>
<tr>
<td>Combined</td>
<td>0.14</td>
<td>0.12</td>
</tr>
</tbody>
</table>

Note: \( P \)-values for each individual study were generated using SNPTEST v2.5.2 software. Combined \( P \)-values and estimates were obtained using a fixed-effects model using beta values and standard errors. RAF risk allele frequency, OR odds ratio, \( P_{\text{het}} \) heterogeneity, \( P \) index to quantify dispersion of odds ratios, CI confidence interval.
genes are associated with infant leukemia and intriguingly GTDC1 has been identified as a 3’ MLL fusion partner in acute leukemia\textsuperscript{37}.

Most cancer GWAS risk loci map to non-coding regions of the genome and in so-far as they have been deciphered their functional basis has been attributed to changes in regulatory regions influencing gene expression\textsuperscript{33,38,39}. The finding that the current and previously identified risk SNPs show a propensity to map within regions of B-cell active chromatin is consistent with such a model of disease susceptibility in ALL. It is therefore noteworthy that SMR analysis revealed significant relationships between 10p12.2 risk variants and these associations.

Our analysis sheds further light on inherited predisposition to childhood ALL. Functional characterization of risk loci identified should provide additional insight into the biological and etiological basis of this malignancy. While the power of our meta-analysis to identify common variants loci (MAF ≥ 0.2) associated with relative risks ≥ 1.2 was around 80%, we acknowledge that we had low power to detect alleles conferring more moderate effects or were present at low frequency. By inference, these types of variant may be responsible for a larger proportion of the heritable risk of ALL. Hence, a large number of risk SNPs may as yet be unidentified. Finally, as we have demonstrated, considering ALL subtypes individually should reveal additional specific risk variants.

**Methods**

**Ethics.** The ascertainment patient samples and associated clinical information was conducted with informed consent according to ethical board approval. Specifically, ethical committee approval was obtained for Medical Research Council UKALL37/99 trial by UK therapy centers and approval for UKALL2003 from the Scottish Multi-Centre Research Ethics Committee (REC02/10/052)\textsuperscript{30,41}. Additionally ethical approval was granted by the Childhood Leukemia Cell Bank, the United Kingdom Childhood Cancer Study, and University of Heidelberg.

**Published GWAS samples.** The United Kingdom (UK) GWAS I and German GWAS have been previously published\textsuperscript{32}. In summary, UK GWAS I comprised (numbers post quality control (QC)) 824 BCP-ALL cases (360 female, average age at diagnosis 5.5 years) genotyped using Human 317K arrays (Illumina, San Diego; http://www.illumina.com); control genotypes were obtained from 2699 individuals from the 1958 British Birth Cohort (Hap1.2M-Duo Custom array data) and 2501 from the UK Blood Service produced by the Welcome Trust Case Control Consortium 2 (http://www.wtccc.org.uk/; 51% male)\textsuperscript{42}. The German GWAS comprised 1155 cases (620 male; mean age at diagnosis 6 years) from the Berlin–Frankfurt–Münster (BFM) trials (1993–2004) genotyped using Illumina Human OmniExpress-12v1.0 arrays (834 sample posts QC). Control data was generated on 2132 (50% male) healthy individuals from the Heinz Nixdorf Recall and infants 784 individuals genotyped using Illumina HumanOmniQuad v1 and 1428 individuals genotyped on Illumina-HumanOmniExpress-12v1.0 platform. In total 2024 controls remained post QC in the German cohort.

**New GWAS samples.** UK GWAS II consisted of 1021 BCP-ALL cases recruited to Medical Research Council UK ALL-2003 (2003–2011) (683 cases; 307 females, mean age: 5.9 years) and ALL-97/99 trials\textsuperscript{5,43} (338 cases, 160 females, mean age: 4.9 years) obtained from the Bloodwise Childhood Leukemia Cell Bank (www.cellbank.org). DNA was extracted from cell pellets by standard ethanol precipitation methods. Samples were then genotyped on an Infinium OncoArray-500K BeadChip from Illumina comprising a 250K SNP genome-wide backbone and a 250K custom content selected across multiple consoritum within COGS (Collaborative Oncological Gene-Environmental Study). OncoArray genotyping was carried out in accordance with the manufacturer’s recommendations by the High-Throughput Genomics Group, Oxford Genomics Center. Prior to genotyping DNA samples were quantified by Quant-IT PicoGreen (Thermo Fisher Scientific, MA, USA), normalized and 30 ng/μl aliquots plated in 96 deep-well plates. Post QC we obtained genotype data for 784 cases (365 female; mean age at diagnosis 5.3 years). Controls consisted of: (1) 2976 cancer-free, men ascertained by the PRACTICAL Consortium; (2) 4446 cancer-free women from the UK through the Breast Cancer Association Consortium. All controls were genotyped on Infinium OncoArray-500K BeadChip arrays.

**Statistical and bioinformatics analysis of GWAS data sets.** Analyses and/or data management were undertaken using R V3.2.3 (R Core Team 2013; http://www.R-project.org)\textsuperscript{37}, PLINK v1.0\textsuperscript{45}, and SNIPTEST v2.5.2 software\textsuperscript{44}. GenomeStudio software (Illumina, San Diego; Available at: http://www.illumina.com) was used to extract genotypes from raw data. QC of all GWAS data sets was performed as suggested by Anderson et al\textsuperscript{46}. PLINK v1.0\textsuperscript{45} was used for conducting the sample and SNP QC steps. Specifically, individuals with low call rate (<95%) as well as all individuals with non-European ancestry (using the HapMap version 2 CEU, JPT/CHB, and YRI populations as a reference) were excluded using the smartpca package, part of EIGENSOFT v4.2\textsuperscript{46}. SNPs with a call rate <95% were excluded as were those with a MAF <0.01 or displaying significant deviation from Hardy-Weinberg equilibrium (i.e. P < 10\textsuperscript{-8}). The adequacy of case-control matching and possibility of differential genotyping of cases and controls were formally evaluated using QQ plots of test statistics. The inflation factor λ was calculated by dividing the median of the test statistics by the median expected values from a χ\textsuperscript{2} distribution with 1 degree of freedom. Q-Q plots were generated and inflation factors estimated using R. Uncorrected and pre-quantification QQ plots of UK GWAS I, UK GWAS II, and German GWAS showed λ values of 1.01, 1.05, and 1.10, respectively. Prior to imputation the data sets were pre-phased by

---

**Table 2** rs17481869 (2q22.3) genotypes and risk associated with BCP-ALL, high-hyperdiploid, and ETV6-RUNX1 childhood BCP-ALL subtypes

<table>
<thead>
<tr>
<th>All BCP-ALL</th>
<th>RAF</th>
<th>Number</th>
<th>Cases</th>
<th>Controls</th>
<th>OR</th>
<th>CI</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>UK GWAS I</td>
<td>0.08</td>
<td>0.07</td>
<td>824</td>
<td>5200</td>
<td>1.18</td>
<td>(0.95–1.46)</td>
<td>1.37 × 10\textsuperscript{-1}</td>
</tr>
<tr>
<td>German GWAS</td>
<td>0.10</td>
<td>0.08</td>
<td>834</td>
<td>2024</td>
<td>1.25</td>
<td>(1.01–1.56)</td>
<td>4.33 × 10\textsuperscript{-2}</td>
</tr>
<tr>
<td>UK GWAS II</td>
<td>0.10</td>
<td>0.07</td>
<td>784</td>
<td>7385</td>
<td>1.52</td>
<td>(1.25–1.84)</td>
<td>2.53 × 10\textsuperscript{-5}</td>
</tr>
<tr>
<td>Combined</td>
<td>0.10</td>
<td>0.07</td>
<td>2442</td>
<td>14,609</td>
<td>1.32</td>
<td>(1.17–1.49)</td>
<td>5.26 × 10\textsuperscript{-6}</td>
</tr>
</tbody>
</table>

**High-hyperdiploid**

| UK GWAS I   | 0.06 | 0.07   | 289   | 5200     | 0.86 | (0.61–1.22) | 4.03 × 10\textsuperscript{-1} |
| German GWAS | 0.08 | 0.07   | 176   | 2024     | 0.98 | (0.64–1.48) | 9.11 × 10\textsuperscript{-1} |
| UK GWAS II  | 0.10 | 0.07   | 251   | 7385     | 1.48 | (1.06–2.08) | 2.13 × 10\textsuperscript{-2} |
| Combined    | 0.10 | 0.07   | 716   | 14,609   | 1.10 | (0.89–1.35) | 0.38 |

**ETV6-RUNX1-positive**

| UK GWAS I   | 0.11 | 0.07   | 126   | 5200     | 2.01 | (1.20–3.39) | 8.52 × 10\textsuperscript{-3} |
| German GWAS | 0.12 | 0.08   | 63    | 2024     | 1.72 | (0.88–3.38) | 1.14 × 10\textsuperscript{-1} |
| UK GWAS II  | 0.13 | 0.07   | 220   | 7385     | 2.34 | (1.64–3.35) | 2.90 × 10\textsuperscript{-6} |
| Combined    | 0.13 | 0.07   | 409   | 14,609   | 2.14 | (1.64–2.80) | 3.20 × 10\textsuperscript{-8} |

Note: P-values for each individual study were generated using SNIPTEST v2.5.2 software. Combined P-values and estimates were obtained using a fixed-effects model using beta values and standard errors. RAF risk allele frequency, OR odds ratio, P\textsubscript{het} P-heterogeneity, P index to quantify dispersion of odds ratios, CI confidence interval.
estimating haplotypes from the GWAS data sets using Segmented HAPlotype Estimation and Imputation Tool to make imputation less computationally intensive. Prediction of the unmapped SNPs was carried out using IMPUTE v2.3.0 based on the data from 1000 Genomes Project (Phase 1 integrated variant set, v3.20101123, http://www.1000genomes.org, 9 December 2013) and UK10K (ALSPAC, EGAS00000100090/EGAD00001000195, and TwinsUK, EGAS00000100010/EGAD00001000194, studies only, see www.uk10k.org) as references. In order to account for genomic inflation post imputation in the German data set, eigenvectors were inferred using the “smartpcas” component within EIGENSOFT v2.4 and adjustment was carried out by including the first two eigenvectors as covariates in SNPTEST during association analysis. The inflation correction factor λ was again calculated for all SNPs per population (i.e., OC13,20). The association between each SNP and risk was calculated using SNPTEST assuming an additive model using a “-frequentist” test and applying a default genotype calling probability threshold of 0.9. Where applicable the first two eigenvectors were used as covariates in the association analyses for that data set. ORs and 95% CI were obtained from the beta values and standard errors obtained from the SNPTEST output. Meta-analyses were performed using META v1.23 pooling the beta values and standard error for SNPs from each GWAS data sets. Association meta-analyses only included markers with info scores >0.8, imputed pooling the beta values and standard error for SNPs from each GWAS data sets. LD metrics were calculated in PLINK43 and vcftools53 using UK10K genomic data. LD blocks were defined on the basis of HapMap recombination rate, as defined by using the Oxford recombination hotspots, and on the basis of the distribution of Ce65. Association plots were generated using visPic14.

HLA imputation. Classical HLA alleles were imputed, both common and rare (A, B, C, DQA1, DBQ1, DRB1) and coding variants across the HLA region using SNP2HLA. The imputation was based on a reference panel from the T1DGC consisting of genotype data from 5225 individuals of European descent with genotyping data of 8961 common SNPs and indel polymorphisms across the HLA region, and four digit genotyping data of the HLA class I and II molecules. This reference panel has been previously and showed high imputation quality for the HLA regions in other studies20-29. Individual GWAS studies were imputed at the 6p21 region and meta-analyzed to identify significant HLA risk alleles. A significance threshold of p<1×10^-16 was set after Bonferroni correction as the number of SNP tests was 8654.

Sanger sequencing. To assess the accuracy of imputed genotypes, a random series of samples was sequenced using BigDye® Terminator v3.1 Cycle Sequencing Kit (Life Technologies, CA, USA) and analyzed using a ABI 3700td sequencer (Applied Biosystems, CA, USA). Oligonucleotide primer sequences are provided in Supplementary Table 12.

Chromatin mark enrichment analysis. To assess for an over-representation of markers for open chromatin the variant set enrichment method of Cowper-Sal Lari et al. was adapted50. For each risk locus, SNPs in LD were defined (i.e., r2 > 0.8 and D’ > 0.8), and termed associated variant set (AVS). Transcription factor ChIP-seq data were obtained from the ENCODE project for the hg19 genome assembly (i.e., H3K27ac, H3k4me1, and H3K4me3 chromatin signatures. ChIP-seq broad peak data for three AML and six childhood ALL cell types were obtained from the Blue-Print Epigenome database (www.blueprint-epigenome.eu)15. For each mark, overlap of SNPs in the AVS and the ChIP peak were derived, generating a mapping score. The null hypothesis was tested by scoring randomly chosen SNPs with the same LD structure at the risk-associated SNPs. After 10,000 iterations, approximate P-values were calculated as the proportion of permutations where null mapping score was at least equal to the AVS mapping score. Enrichment was calculated normalizing scores to the median of the null model.

Hi-C analysis. Hi-C analysis was conducted using the HUGIn browser57, which is based on the analysis by Smithit et al. Specifically we analyzed Hi-C data generated on the H1 ES cells and GM12878 lymphoblastoid cell lines originally described in Dixon et al. and Smithit et al. respectively. Plotted topologically associating domains boundaries were obtained from the insulating score method at 40 kb bin resolution58. We searched for significant interactions (P-values generated using “Tt-Hi-C”58) between bins mapping the currently identified ALL risk loci with target genes (e.g., “virtual 4C”).

Functional annotation. SNPs in LD (r2 > 0.8) with the top SNPs from each risk loci were assessed for histone marks in relevant tissue, proteins bound and location were annotated using HaploReg57 (Supplementary Data 1). eQTL analysis was performed by testing each sentinel SNP with genes 1MB upstream and downstream using the whole blood eQTL data available in the Blue qTL browser64 (Supplementary Data 1). Methylation quantitative trait loci (mQTL) for all known BCP-ALL risk loci where assessed using the mQTL Database (www.mqtldb.org), which shows the presence of significant methylated CpG sites at various stages of life as described by Gaunt et al.59.

SMR analysis. SMR analysis was conducted as per Zhu et al. (http://cancer.sanger.ac.uk/software/smr/index.html)61. Publicly available eQTL data was extracted from the whole blood eQTL Muther consortia, and GTEX6 v6 release portal61,64. GWAS summary statistics were generated from the meta-analysis of UK GWAS I, UK GWAS II, and German GWAS data sets. Reference files were generated by merging 1000 genomes phase 3 and UK10K (ALSPAC and TwinsUK) vcs. Summary eQTL files for the GTEa samples were generated from downloaded v6p “all SNP” eQTL files. REND files were generated from downloaded SNP-gene eQTL data, which were converted into a query file format as mentioned in the SMR online guide (http://cnsgenomics.com/software/smr) and then using the “make-besd” command to make binary versions of the files. Only probes with eQTL P < 5×10^-8 were considered in the SMR analysis. A PCA test of the SMR test at P<5×10^-8 was considered significant for both the original and meta-analysis with specificity set at 80% and sensitivity set at 90% for all loci in UK10K. HEIDI test P-values <0.05 were taken to indicate significant heterogeneity as suggested by Zhu et al. For the two genes passing the thresholds, plots of eQTL and GWAS associations as well as plots of GWAS and eQTL effect sizes were constructed.

Relationship between SNP genotype and survival. The relationship between SNP genotype and survival analysis was performed in the German AEIOF-BFM series, MRC ALL 97/99 and the UKALL2003 series. The German series consisted of 834 patients within the AEIOF-BFM 2000 trial65. Patients were treated with conventional chemotherapy (i.e., prednisone, vincristine, daunorubicin, l-asparaginase, cyclophosphamide, dexamethasone, cytarabine, 6-mercaptopurine, 6-thioguanine) using a protocol with high doses of cranial irradiation and/or stem cell transplantation. Events, for EFS, were defined as resistance to therapy, relapse, secondary cancer, or death. Kaplan–Meier methodology was used to estimate survival rates, with differences between groups tested using the log-rank method (two-sided P-values). Cumulative incidences of competing events were calculated using the methodology of Kalbfleisch and Prentice66, and compared using Gray’s test67. Cox regression analysis was used to estimate hazard ratios and 95% CIs adjusting for clinically relevant covariates. The full details regarding the recruitment, classification, and treatment of patients on MRC ALL97/99 (1997–2002) or UKALL2003 (2003–2011) have been published already41,67. In ALL97/99, patients were classified as standard or high risk based on the Oxford score. In ALL99 and UKALL2003, patients were initially assigned to regimen A or B based on whether they were NCI standard or high risk. Regimen A comprised a three drug induction following by consolidation, CNS-directed therapy, interim maintenance, delayed intensification, and continuing therapy. Regimen B patients additionally received a four drug induction and BFM consolidation. Treatment response and cytogenetics were used to re-assign high-risk patients to regimen C to receive augmented BFM consolidation and Capizzi maintenance. In ALL99 and ALL2003, early treatment response was measured by marrow morphology at day 8/15 for regimen B A-treated patients. In addition, ALL2003 patients were randomized to regimen C if their induction of minimal residual disease levels—evaluated by real-time quantitative PCR analysis of immunoglobulin and T-cell receptor gene rearrangements—were >0.01%. Survival analysis considered two endpoints: EFS defined as time to relapse, second tumor or death, censoring at last contact; and relapse rate defined as time to relapse for those achieving complete remission and surviving at death in remission or last contact. Survival rates were calculated and compared using Kaplan–Meier methods and log-rank tests. All analyses were performed using Intercooled Stata 13.0 (Stata Corporation, USA).

Contribution of genetic variance to familial risk. Estimation of risk variance associated with each SNP was performed as per Pharoah et al. For an allele (i) of frequency p, relative risk R and log risk r, the risk distribution variance (Vr) is: 

\[V_r = (1 - p)^2 \cdot \log(1 + e) - 2 \cdot p \cdot (1 - p) \cdot \log(1 + e)\]

where E is the expected value of r given by: 

\[E = 2 \cdot p \cdot (1 - p) + 2 \cdot p^2\]

For multiple alleles the distribution of risk in the population tends toward the normal with variance: 

\[V_{\text{total}} = \sum_{i=1}^{n} V_r(i)\]

The percentage of total variance was calculated assuming a familial risk of childhood ALL of 3.2 (95% CI 1.5–5.9) as per Kharaizmi et al. All genetic variance (V) associated with susceptibility alleles is given as V32. The proportion of genetic risk attributable to a single allele is: 

\[V_i / V_{\text{total}}\]

Eleven risk loci were included in the calculation of the PRS for childhood ALL by selecting the top SNP from the current meta-analysis from each previously published loci in addition to the two risk loci discovered in this study. The eleven variants are thought to act independently as previous studies have shown no interaction between risk loci. PRS were generated as per ALL with et al. assuming a log-normal distribution LN(μ, σ2) with mean μ and variance σ2. The population μ was set as σ2/2, in order that the overall mean PRS was 1.0. The

DATABASE

**ARTICLE NATURE COMMUNICATIONS | DOI: 10.1038/s41467-018-03178-z**

NATURE COMMUNICATIONS | DOI: 10.1038/s41467-018-03178-z www.nature.com/naturecommunications
sibling relative risk were assumed to be 3.24. The discriminatory value of risk SNPs was examined by determining the AUC for the ROC curve.

**GCTA to estimate heritability.** Since artefactual differences in allele frequencies between cases and controls have the potential to bias estimation genetic variation, additional QC measures were imposed on the GWAS data sets which have been advocated by Lee et al.13. Typed SNPs were excluded if they had a MAF < 0.01 or a HWE p-value < 0.05. SNPs were also excluded if a differential missingness test between cases and controls was P < 0.05. In addition, individuals were excluded if having a relatedness score of >0.5. Filtering resulted in the 260,127 SNPs in the UK GWAS I and 355,899 SNPs in UK GWAS II data sets, respectively. GCTA (http://csgenomics.software/gcta/) was employed to estimate the fraction of the phenotypic variance attributed by SNPs given a prevalence of 0.0005 for ALL.20

**Data availability.** The UK GWAS I control set comprised 2699 individuals in the 1958 British Birth Cohort (Hapi2M-Duo Custom array data) and 2501 individuals from the UK Blood Service obtained from the publicly accessible data generated by the Wellcome Trust Case Control Consortium 2 (http://www.wtccc.org.uk/; WTCCC2:EGAD00000000022, EGAD00000000024). The reference panels used in the imputation can be obtained from the 1000 genomes phased haplotypes (n = 1092) from the Phase I integrated variant set release (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/) and the UK10k (n = 3781; EGAS00001000909, EGAS00001000195, EGAS00001001088; www.uk10k.org) sequenced data sets. cQTL data for various functional analyses were obtained from the MuTHER (genotype-expression profiled samples with genotype array data and methylation data; E-TABM-1144), Blood cQTL (whole-genome gene expression array data sets with RNA sequencing and genotyping data; E-TABM-1036, E-MTAB-945, E-MTAB-1708; http://www.nature.com/gnl/journal/v45/n10/abs/11073414.html), and ENCODE transcription factor binding data sets (transcription factor Chip-seq data from various tissues; http://encode.umd.edu/ENCODE/download.html). ChIP-seq broad peak data for childhood ALL and AML cells were obtained from the BluePrint Epigenome (dcc.blueprint-epigenome.eu) for samples S00FGCH1, S005GFIH, S00KPH1, S017HE1, S0179HD1, S01GFRH1, S01GHQH1, S017HE1H1, and S0177HH1H. The UK GWAS II data set can be accessed through the European Genome-Phenome Archive (EGA, https://ega-archive.org) under the study accessions EGAS0000102529. All other relevant data are available on request to the authors.

Received: 20 July 2017 Accepted: 25 January 2018

**Published online:** 09 April 2018

**References**


Additional information Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-03178-z.

Competing interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.