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, ZSo 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 (B-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 7p12.2 (IKZF1), 9p21.3 (CDKN2A), 10p12.2 (PPIK2A), 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 identified 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 no evidence 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, Tables 1 and 2, Supplementary Table 3). rs17481869 was genotyped in UK GWAS II and German GWAS, while rs28665337 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).

 rs17481869 maps to an intergenic region at 2q22.3 with no candidate gene nearby (Fig. 2).

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.
Eighteen gene probes exceeded the threshold of 1.3 × 10⁻⁴, of which two genes passed the HEIDI test (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 Table 7). Analysis using HaploReg revealed a significant enrichment of SNPs within enhancers in primary hematopoietic stem cells (binomial test for enrichment, FDR ≤ 0.05). In whole blood-derived tissue, enrichment is highest in ALL cells (Supplementary Fig. 4, Supplementary Table 7). Analysis across all risk loci combined revealed that risk SNPs are enriched for markers of open chromatin and that the 2q22.3 risk allele (rs17481869) with decreased MYC expression (P_{SMR} = 2.09 × 10⁻⁴, b_{XY} = −0.99, and P_{SMR}, = 7.48 × 10⁻⁴, b = 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 × 10⁻⁴; Supplementary Fig. 6, Supplementary Table 10), and the 2q22.3 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 promotor regions of MYC in ES cells and GTDC1 in GM12878, respectively (Fit-Hi-C test, 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. 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

![Fig. 2](https://example.com/fig2.png) 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) show association −log₁₀P-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). −log₁₀P-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 (r² = 0), dark red (r² = 1.0). Light blue line shows recombination rates from UK10K Genomes Project. Genome coordinates are from NCBI human genome GRCh37.

---

**NATURE COMMUNICATIONS** | [DOI: 10.1038/s41467-018-03178-z](https://doi.org/10.1038/s41467-018-03178-z) | www.nature.com/naturecommunications
Impact on heritable risk. Using genome-wide complex trait analysis (GCTA)\textsuperscript{30–32} the heritability of BCP-ALL accounted for by common variants was estimated to be 0.16 (±standard error (S.E.) 0.05; frequentist test of association using SNPTEST); this association was, however, not significant after correcting for multiple testing.

**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>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><strong>All BCP-ALL</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UK GWAS I</td>
<td></td>
<td>0.15</td>
<td>0.12</td>
<td>824</td>
<td>5200</td>
<td>(1.12-1.55)</td>
</tr>
<tr>
<td>German GWAS</td>
<td></td>
<td>0.16</td>
<td>0.12</td>
<td>834</td>
<td>2024</td>
<td>(1.07-1.53)</td>
</tr>
<tr>
<td>UK GWAS II</td>
<td></td>
<td>0.15</td>
<td>0.12</td>
<td>784</td>
<td>7385</td>
<td>(1.21-1.47)</td>
</tr>
<tr>
<td>Combined</td>
<td></td>
<td>0.15</td>
<td>0.12</td>
<td>2442</td>
<td>14,609</td>
<td>(1.21-1.47)</td>
</tr>
<tr>
<td><strong>High-hyperdiploid</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UK GWAS I</td>
<td></td>
<td>0.15</td>
<td>0.12</td>
<td>289</td>
<td>5200</td>
<td>(1.11-1.88)</td>
</tr>
<tr>
<td>German GWAS</td>
<td></td>
<td>0.17</td>
<td>0.12</td>
<td>176</td>
<td>2024</td>
<td>(1.06-2.09)</td>
</tr>
<tr>
<td>UK GWAS II</td>
<td></td>
<td>0.15</td>
<td>0.12</td>
<td>251</td>
<td>7385</td>
<td>(1.05-1.81)</td>
</tr>
<tr>
<td>Combined</td>
<td></td>
<td>0.15</td>
<td>0.12</td>
<td>716</td>
<td>14,609</td>
<td>(1.21-1.87)</td>
</tr>
<tr>
<td><strong>ETV6-RUNX1-positive</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UK GWAS I</td>
<td></td>
<td>0.16</td>
<td>0.12</td>
<td>126</td>
<td>5200</td>
<td>(1.01-2.26)</td>
</tr>
<tr>
<td>German GWAS</td>
<td></td>
<td>0.09</td>
<td>0.12</td>
<td>63</td>
<td>2024</td>
<td>(0.44-1.38)</td>
</tr>
<tr>
<td>UK GWAS II</td>
<td></td>
<td>0.14</td>
<td>0.12</td>
<td>220</td>
<td>7385</td>
<td>(0.94-1.62)</td>
</tr>
<tr>
<td>Combined</td>
<td></td>
<td>0.14</td>
<td>0.12</td>
<td>409</td>
<td>14,609</td>
<td>(1.00-1.51)</td>
</tr>
</tbody>
</table>

Note: P-values for each individual study were generated using SNPVTEST 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; Phet: heterogeneity; P: index to quantify dispersion of odds ratio; CI: confidence interval.

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 are 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\textsuperscript{33} and Ph-like ALL\textsuperscript{40}, 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-RUNX1-positive BCP-ALL.

Deregulation of MYC has been reported in ALL, in some instances as a consequence of chromosomal rearrangement\textsuperscript{34}. 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\textsuperscript{35}. 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-RUNX1-positive BCP-ALL. The SNP association is intergenic positive BCP-ALL. The SNP association is intergenic with little evidence for subtype difference (0.18 ± S.E. 0.05 and 0.20 ± S.E. 0.08 for hyperdiploid and ETV6-RUNX1-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 receiver–operator characteristic (ROC) curves with the area under the curve (AUC) being 0.73 (Supplementary Fig. 10).
genes are associated with infant leukemia and intriguingly \text{GTDC1} has been identified as a 3’ MLL fusion partner in acute leukemia\cite{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\cite{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 \text{PIP4K2A} expression and 10q26.13 risk variants and \text{FAM53B} expression suggesting a mechanism for 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 UKALL1997–1999 trial by UK therapy centers and approval for UKALL2003 from the Scottish Multi-Centre Research Ethics Committee (REC02/10/032)\cite{40,41}. 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\cite{37}. 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 (Illimina, San Diego; http://www.illimina.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 Wellcome Trust Case Control Consortium 2 (http://www.wtccc.org.uk/; 51% male)\cite{42}. The German GWAS comprised 1555 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 samples post QC). Control data was generated on 2132 (50% male) healthy individuals from the Heinz Nixdorf Recall and Infanta 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,\cite{40,41} (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 consortia 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 50 ng/ul 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)\cite{37}, PLINK v1.0\cite{43}, and SNIPTEST v2.5.2 software\cite{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.\cite{37}. PLINK v1.0\cite{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 (i.e., >10%). 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 \(\lambda\) was calculated by dividing the median of the test statistics by the median expected values from a \(\chi^2\) distribution with 1 degree of freedom. Q–Q plots were generated and inflation factors estimated using R. Uncorrected and pre-inflation QQ plots of UK GWAS I, UK GWAS II, and German GWAS showed \(\lambda\) values of 1.01, 1.05, and 1.10, respectively. Prior to imputation the data sets were pre-phased by

### Table 2

<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>0.10</td>
<td>0.08</td>
<td>824</td>
<td>5200</td>
<td>1.18</td>
<td>(0.95–1.46)</td>
<td>1.37 × 10^{-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^{-2}</td>
</tr>
<tr>
<td>UK GWAS II</td>
<td>0.10</td>
<td>0.08</td>
<td>784</td>
<td>7385</td>
<td>1.52</td>
<td>(1.25–1.84)</td>
<td>2.53 × 10^{-5}</td>
</tr>
<tr>
<td>Combined</td>
<td>0.10</td>
<td>0.08</td>
<td>2442</td>
<td>14,609</td>
<td>1.32</td>
<td>(1.17–1.49)</td>
<td>5.26 × 10^{-6}</td>
</tr>
</tbody>
</table>

**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, \(p_{het}\) p heterogeneity, \(\chi^2\) index to quantify dispersion of odds ratios, CI confidence interval.
estimating haplotypes from the GWAS data sets using Segmented Haplotype ARTICLE NATURE COMMUNICATIONS | DOI: 10.1038/s41467-018-03178-z

Functional annotation

with target genes (e.g., region, and four digit genotyping data of the HLA class I and II molecules. This consisting of genotype data from 5225 individuals of European descent with

performed by testing each sentinel SNP with genes 1MB upstream and downstream

significance threshold of 5.7 × 10^-20. The association between each SNP and risk was calculated using

QC13, 50. The association between each SNP and risk was calculated using

Chromatin mark enrichment analysis

D -values were calculated as the proportion of permutations where null mapping

Hi-C analysis

based on the analysis by Schmitt et al58. Specifically we analyzed Hi-C data gen-

Sanger sequencing

Sanger sequencing

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% CIs 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 call rates/SNP >0.9, and MAFs > 0.01. Collectively the three GWAS provided

pooling the beta values and standard error for SNPs from each GWAS data sets. Estimating haplotypes from the GWAS data sets using Segmented Haplotype ARTICLE NATURE COMMUNICATIONS | DOI: 10.1038/s41467-018-03178-z

SMR analysis

SMR analysis was conducted as per Zhu et al. (at http:// ethnics.com/software/smr/index.html)59. Publicly available eQTL data was extracted from the whole blood eQTL Muther consortia, and GTEX16 v6p release portal60,61,64. 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, EGAS00001000090, EGAD0000100195, and TwinsUK, EGAS00001000090, EGAD0000100194, studies only, see http://www.uk10k.org/) as a reference. In order to account for genomic inflation post imputation in the German data set, eigenvectors were inferred using the "smartpc" component within EIGENSOFT v2.4 and adjustment was carried out by including the first two eigenvectors as covariates in SNPTEST during association analysis64,65. The inflation corrected test statistic was again calculated for all SNPs previously analyzed. All genotypes in the same LD structure at the risk-associated SNPs. After 10,000 iterations, approximate
distribution of Ca54. Association plots were generated using visP14.

HLA imputation

Classical HLA alleles were imputed, both common and rare (A, B, C, DQA1, DQB1, DRB1) and coding variants across the HLA region using SNPNLA25. 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 9891 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 described and shown high imputation quality for the HLA regions in other studies22-29. Individual GWAS studies were imputed at the 6p21 region and meta-analyzed to identify significant HLA risk alleles. A significant threshold p < 5 × 10^-8 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 Sanger sequenced using BigDye Terminator v3.1 Cycle Sequencing Kit (Life Technologies, CA, USA) and analyzed using a ABI 3700d 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 adapted26. For each risk locus, SNPs in LD were defined (i.e., r^2 > 0.8 and D’ > 0.8), and termed associated variant set (AVS). Transmission factor ChiP-seq data was downloaded from the Edinburghersive Portal for the ENCODE project for 14 ENCODE sites for H3K27ac, H3K4me1, and H3K4me3 chromatin signatures. ChiP-seq peak data for three AML and six childhood ALL cell types were obtained from the BluePrint Epigenome database (www.blueprint-epigenome.eu)55. 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 HUGHin browser57, which is based on the analysis by Schmitt 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.59 and Schmitt et al.38, respectively. Plotted topologically associating domains boundaries were obtained from the insulating score method at 40 kb bin resolution70. We searched for significant interactions (P-values generated using "Fit-Hi-C"71, between bins overlapping the currently identified ALL risk loci with target genes (e.g., "virtual 4C").

Functional annotation

SNPs in LD (r^2 > 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 data available in the Blood eQTL browser61 (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.62.

Relationship between SNP genotype and survivorship

The relationship between SNP genotype and survival was analyzed in the German AIEOP-BFM series, MRC ALL 97/99 and the UKALL2003 series. The German series consisted of 834 patients within the AIEOP-BFM 2000 trial56. Patients were treated with conventional chemotherapy (i.e., prednisone, vincristine, daunorubicin, l-asparaginase, cyclophosphamide, dexamethasone, cytarabine, 6-mercaptopurine, 6-thioguanine, methotrexate, and cytarabine), a subset of those with high-risk ALL were treated with 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 Prentice56, 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 published41,68. In ALL97, patients were stratified for the SMR test of P<0.05 < 5 × 10^-8 corresponding to a 40 kb resolution for 38 tests and for all 23 genes within 1 MB of the sentinel risk SNPs in each risk loci (38 genes probes with a top eQTL P < 5 × 10^-8). 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.

Contingent of genetic variance to familial risk

Estimation of genetic variance with each SNP was performed as per Pharoah et al.39. For an allele (i) of frequency p, relative risk R and log risk r, the risk distribution variance (V_r) is:

V_r = (1 − r)^2 p^2 + 2(1 − p)(r − 1)p^2 + r^2 p^2 − r^2 p^2

Where E is the expected value of r given by:

E = 2p(r - 1) + p^2

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

V = V_r

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 Kharaamzi et al.37. All genetic variance (V) associated with susceptibility alleles is given as V_s. The proportion of genetic risk attributable to a single allele is:

V_s/V

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 last134. Assuming a log-normal distribution LN(μ, σ^2) with mean μ and variance σ^2 the population μ was set to σ^2/2, in order that the overall mean PRS was 1.0. The

© NATURE COMMUNICATIONS | DOI: 10.1038/s41467-018-03178-z | www.nature.com/naturecommunications

NATURE COMMUNICATIONS | DOI: 10.1038/s41467-018-03178-z | www.nature.com/naturecommunications

6

8

DOI: 10.1038/s41467-018-03178-z
sibling relative risk were assumed to be 3.2. The discriminatory value of risk SNPs was examined by determining the AUC for the ROC curve.

GCTA to estimate heritability. Since artifactual 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. Typel 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 a relatedness score > 0.5. Filtering resulted in the 260,127 SNPs in the UK GWAS I and 335,899 SNPs in UK GWAS II data sets, respectively. GCTA (http://csgenomics.software/cta) was employed to estimate the fraction of the phenotypic variance attributed by SNPs given a prevalence of 0.0005 for ALL.13

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.wtcc2.org.uk/; WTCCC2-EAGAD00000000022, EAGAD00000000024). 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/20101025/) and the UK10K (n = 3781; EGAS0000100090, EGAS00001000195; EGAS0000100108; www.uk10k.org) sequenced data sets. cQTL data for various functional analyses were obtained from the MuTHER (genome-wide expression profile samples with genotype array data and methylation data; E-TABM-1140), Blood eQTL (whole-genome gene expression array data sets with RNA sequencing and genotyping data: E-TABM-1036, E-MTAB-945, E-MTAB-1708; http://www.ebi.ac.uk/giuly/journal/v45/110/abs/ng.2756.html), and ENCODE transcription factor binding data sets (transcription factor ChIP-seq data from various tissues: http://genome.ucsc.edu/ENCODE/downloads.html). ChIP-seq broad peak data for childhood AML and AML cells were obtained from the BluePrint Epigenome (dcc.blueprint-epigenome.eu) for samples S00FGCH1, S005GHH1, S00KPBH1, S017EH1, S0179DH1, S01GHH1, S0179HH1, and S0177HH1. The UK GWAS II data set can be accessed through the European Genome-Phenome Archive website (EGA, https://ega-archive.org) under the study accession EGAS0000102099. 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
22. Wellcome Trust Case Control, C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 661–678 (2007).

**Author contributions**

**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.

---

The PRACTICAL Consortium


28Department of Preventive Medicine, Keck School of Medicine, University of Southern California/Norris Comprehensive Cancer Center, Los Angeles, CA 90033, USA. 29Department of Public Health and Primary Care, Centre for Cancer Genetic Epidemiology, Strangeways Research Laboratory, University of Cambridge, Cambridge CB2 0SP, UK. 30The Institute of Cancer Research, London SM2 5NG, UK. 31Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH 44106, USA. 32Seidman Cancer Center, University Hospitals, Cleveland, OH 44106, USA. 33Department of Clinical Neurosciences, University of Cambridge, Cambridge CB2 0PY, UK. 34Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, MD 20814, USA. 35Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm 171 77, Sweden. 36Epidemiology Research Program, American Cancer Society, 250 Williams Street, Atlanta, GA 30303, USA. 37SWOG Statistical Center, Fred Hutchinson Cancer Research Center, Seattle, WA 98109-1024, USA. 38Australian Prostate Cancer Resource Centre (APCR), Australian Prostate Cancer Centre-QLD, Institute of Health and Biomedical Innovation and School of Biomedical Science, Queensland University of Technology, Brisbane 4001 QLD, Australia. 39Translational Research Institute, Brisbane 4102 QLD, Australia. 40Department of Medical Biochemistry and Genetics, Institute of Biomedicine, University of Turku, Turku FI-20014, Finland. 41Tüks Microbiology and Genetics, Department of Medical Genetics, Turku University Hospital, Turku 20521, Finland. 42BioMedTech, University of Tampere, Tampere 33014, Finland. 43Division of Nutritional Epidemiology, Institute of Environmental Medicine, Karolinska Institutet, Solna SE-171 77, Sweden. 44Institute of Cancer Sciences, University of Manchester, Manchester Academic Health Science Centre, Radiotherapy Related Research, The Christie Hospital NHS Foundation Trust, Manchester M13 9PL, UK. 45Department of Epidemiology, Harvard School of Public Health, Boston, MA 02115, USA. 46CeRePP, Pitie-Salpetriere Hospital, Paris 75020, France. 47UPMC Univ Paris 06, GRC N°S ONCOTYPE-URO, CeRePP, Tenon Hospital, Paris 75020, France. 48Department of Clinical Medicine, Aarhus University Hospital, Aarhus DK-8200, Denmark. 49Department of Medical Genetics, Aarhus University Hospital, Aarhus 8000, Denmark. 50Department of Medical Genetics, Oslo University Hospital, Oslo 0424, Norway. 51Department of Oncology, Addenbrooke’s Hospital, University of Cambridge, Cambridge CB2 0QQ, UK. 52Li Ka Shing Centre, Cancer Research UK Cambridge Research Institute, Cambridge CB2 0RE, UK. 53Cancer Epidemiology, Nuffield Department of Population Health, University of Oxford, Oxford OX3 7LF, UK. 54Department of Surgical Oncology, Princess Margaret Cancer Centre, Toronto M5G 2C4, Canada. 55Department of Radiation Oncology, Icahn School of Medicine at Mount Sinai, New York, NY 10029-5674, USA. 56Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029-5674, USA. 57Centre for Molecular Oncology, Barts Cancer Institute, John Vane Science Centre, Queen Mary University of London, London EC1M 6BQ, UK. 58Cancer Epidemiology Centre, The Cancer Council Victoria, Melbourne 3004 VIC, Australia. 59Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne 3010, Australia. 60Division of Urologic Surgery, Brigham and Women’s Hospital, Boston, MA 02115, USA. 61Fundación Pública Galega de Medicina Xenómica-SERGAS, Grupo de Medicina Xénomica, CIBERER, IDIS, Santiago de Compostela 15706, Spain. 62Centre for Research in Environmental Epidemiology (CREAL), Barcelona Institute for Global Health (ISGlobal), Barcelona 08036, Spain. 63CIBER Epidemiología y Salud Pública (CIBERESP), Madrid 28029, Spain. 64IMIM (Hospital del Mar Research Institute), Barcelona 08003, Spain. 65Universitat Pompeu Fabra (UPF), Barcelona 08005, Spain. 66Channing Division of Network Medicine, Department of Medicine, Brigham and Women’s Hospital/Harvard Medical School, Boston, MA 02115, USA. 67Department of Cancer Epidemiology, Moffitt Cancer Center, Tampa 33612, USA. 68Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA. 69Department of Epidemiology, School of Public Health, University of Washington, Seattle, WA 98195, USA. 70International Hereditary Cancer Center, Department of Genetics and Pathology, Pomeranian Medical University, Szczecin 70-204, Poland. 71Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen DK-2200, Denmark. 72Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev 2730, Denmark. 73Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg 69120, Germany. 74German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg 69120, Germany. 75Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg 69120, Germany. 76Institute for Human Genetics, University Medical Center Hamburg-Eppendorf, Hamburg 20246, Germany. 77Molecular Medicine Center, Department of Medical Chemistry and Biochemistry, Medical University, Sofia 1431, Bulgaria. 78Department of Oncology, Cross Cancer Institute, University of Alberta, Edmonton T6G 1Z2, Canada. 79Division of Radiation Oncology, Cross Cancer Institute, Edmonton T6G 2R7, Canada. 80Department of Cellular and Molecular Medicine, Molecular Endocrinology Laboratory, KU Leuven, Leuven 3000, Belgium. 81Institute of Cancer Sciences, Manchester Cancer Research Centre, University of Manchester, Manchester Academic Health Science Centre, St. Mary’s Hospital, Manchester M13 9PL, UK. 82Genomic Medicine Group, Galician Foundation of Genomic Medicine, Instituto de Investigacion Sanitaria de Santiago de Compostela (IDIS), Complejo Hospitalario Universitario de Santiago, Servicio Galego de Saúde, SERGAS, Santiago de Compostela 15706, Spain. 83Moores Cancer Center, University of California San Diego, La Jolla, CA 92039, USA. 84Department of Urology, Erasmus University Medical Center, Rotterdam 3015, The Netherlands. 85Cancer & Environment Group, Center for Research in Epidemiology and Population Health (CESP), INSERM, University Paris-Sud, University Paris-Saclay, Villejuif 94800, France.