This is a repository copy of Genome-wide association study identifies susceptibility loci for B-cell childhood acute lymphoblastic leukemia.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/129865/

Version: Published Version

Article:
Vijayakrishnan, Jayaram, Studd, James, Broderick, Peter et al. (92 more authors) (2018) Genome-wide association study identifies susceptibility loci for B-cell childhood acute lymphoblastic leukemia. Nature Communications. 1340. ISSN 2041-1723

https://doi.org/10.1038/s41467-018-03178-z

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.
Genome-wide association study identifies susceptibility loci for B-cell childhood acute lymphoblastic leukemia


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.

1 Division of Genetics and Epidemiology, The Institute of Cancer Research, Sutton, Surrey SM2 5NG, UK. 2 Division of Molecular Genetic Epidemiology, German Cancer Research Centre, 69120 Heidelberg, Germany. 3 Northern Institute for Cancer Research, Newcastle University, Newcastle upon Tyne NE2 4HH, UK. 4 Wolfson Childhood Cancer Research Centre, Northern Institute for Cancer Research, Newcastle University, Newcastle upon Tyne NE1 7RU, UK. 5 Department of Haematology, Great Ormond Street Hospital, London WC1N 3JH, UK. 6 Department of Health Sciences, University of York, York YO10 5DD, UK. 7 Department of Paediatric and Adolescent Haematology and Oncology, Leeds General Infirmary, Leeds LS1 3EX, UK. 8 Medical Genetics Research Group, Leeds Institute of Molecular Medicine, University of Leeds, Leeds LS9 7TF, UK. 9 Paediatric and Familial Cancer Research Group, Institute of Cancer Sciences, St. Mary’s Hospital, Manchester M13 9WL, UK. 10 Department of Human Genetics, Institute of Human Genetics, University of Heidelberg, 69120 Heidelberg, Germany. 11 Department of Genomics, Institute of Human Genetics, Life & Brain Centre, University of Bonn, D-53012 Bonn, Germany. 12 Department of Biomedicine, Human Genomics Research Group, University Hospital and University of Basel, 4031 Basel, Switzerland. 13 Institute for Medical Informatics, Biometry and Epidemiology, University Hospital Essen, University of Duisburg-Essen, 45147 Essen, Germany. 14 Department of Oncology, Centre for Cancer Genetic Epidemiology, Cambridge, Cambridge CB1 8RN, UK. 15 Department of Public Health and Primary Care, Centre for Cancer Genetic Epidemiology, University of Cambridge, Cambridge CB1 8RN, UK. 16 Department of Oncology, Centre for Cancer Genetic Epidemiology, University of Cambridge, Strangeways Laboratory, Cambridge CB1 8RN, UK. 17 Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK. 18 Genomic Epidemiology Group, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany. 19 Division of Breast Cancer Research, The Institute of Cancer Research, London SW7 3RP, UK. 20 Royal Marsden NHS Foundation Trust, London SW3 6JJ, UK. 21 Institute of Population Health, University of Manchester, Manchester M13 9PL, UK. 22 Warwick Medical School, University of Warwick, Coventry CV4 7AL, UK. 23 Department of Applied Health Research, University College London, London WC1E 7HB, UK. 24 Centre for Evolution and Cancer, Institute of Cancer Research, Sutton, Surrey SM2 5NG, UK. 25 Department of Paediatric Haematology and Oncology, Hannover Medical School, 30625 Hannover, Germany. 26 General Paediatrics, University Hospital Schleswig-Holstein, 24105 Kiel, Germany. 27 Center for Primary Health Care Research, Lund University, 221 00 Lund, Sweden. Correspondence and requests for materials should be addressed to R.S.H. (email: richard.houlston@icr.ac.uk). #Full list of consortium members appears at the end of the paper.
Acutelymphoblastic 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 7p12.2 (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 totaling 2442 cases and 14,609 controls. We report two previously unidentifed 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-positve 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...
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 \( P_{\text{SMR}} > 0.8 \) threshold of \( 1.3 \times 10^{-4} \), of which two genes passed the HEIDI 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 PIPK2A (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 2q22.3 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 in GM12878 lymphoblastoid and H1 human embryonic stem (ES) cells using Hi-C data.

**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...
analysis. Specifically, we investigated a possible relationship between BCP-ALL risk and HLA alleles by imputing the 6p21 region using the Type I Diabetes Genetics Consortium (T1DGC) as reference.27–29 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)30–32 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-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).

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 10q22 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 Ph-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-RUNX1-positive BCP-ALL.

Deregulation of MYC has been reported in ALL in some instances as a consequence of chromosomal rearrangement.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 lncRNAs in mediating risk.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 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.36 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 Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<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>
<tr>
<td>High-hyperdiploid</td>
<td></td>
<td></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>
<tr>
<td>ETV6-RUNX1-positive</td>
<td></td>
<td></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.

---

**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>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>Controls</td>
</tr>
<tr>
<td>UK GWAS I</td>
<td>0.15</td>
</tr>
<tr>
<td>German GWAS</td>
<td>0.16</td>
</tr>
<tr>
<td>UK GWAS II</td>
<td>0.15</td>
</tr>
<tr>
<td>Combined</td>
<td>0.15</td>
</tr>
<tr>
<td>P_{het} = 0.77</td>
<td>$\mu = 0.9%$</td>
</tr>
</tbody>
</table>

- **UK GWAS I**
  - 0.15, 0.12, 289, 5200, 1.45, (1.11-1.88), 6.3 x 10^{-3}, 2.29 x 10^{-2}, 2.19 x 10^{-2}, 2.55 x 10^{-5}, $P_{\text{het}} = 0.94$, $\mu = 0%$, $\mu = 0%$
- **German GWAS**
  - 0.17, 0.12, 176, 2024, 1.49, (1.06-2.09), 2.29 x 10^{-2}, 2.19 x 10^{-2}, 2.55 x 10^{-5}, $P_{\text{het}} = 0.94$, $\mu = 0%$, $\mu = 0%$
- **UK GWAS II**
  - 0.15, 0.12, 251, 7385, 1.38, (1.05-1.81), 2.19 x 10^{-2}, 2.55 x 10^{-5}, $P_{\text{het}} = 0.94$, $\mu = 0%$, $\mu = 0%$
- **Combined**
  - 716, 14,609, 1.49, (1.21-1.87), 2.55 x 10^{-5}, $P_{\text{het}} = 0.94$, $\mu = 0%$, $\mu = 0%$

**Impact on heritable risk.** Using genome-wide complex trait analysis (GCTA)30–32 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-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 GTDC1 has been identified as a 3’ MLL fusion partner in acute leukemia37.

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 expression33,38,39. 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 PIP4K2A expression and 10q26.13 risk variants and 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 of ALL. Hence, a large number of risk SNPs may as yet be present at low frequency. By inference, these types of loci are within regions of B-cell active chromatin is consistent with such a model of disease susceptibility in ALL. It is therefore noteworthy.

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 UK ALL2003 (683 cases; 307 females, mean age: 5.9 years) and ALL97/99 trials (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/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.

Statistic 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/); PLINK v1.943, and SNPTEST v2.5.2 software44. GenoStudio 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 al45. PLINK v1.94 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, IJPT/CHB, and YRI populations as a reference) were excluded using the smartpca package, part of EigenSOFT v4.246,47. 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 < 10−6). 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 χ² distribution with 1 degree of freedom. Q–Q plots were generated and inflation factors estimated using R Uncorrected and pre-imputation 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 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>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>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>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>2442</td>
<td>14,609</td>
<td>1.32</td>
<td>(1.17-1.49)</td>
<td>2.96 × 10⁻⁶</td>
</tr>
<tr>
<td>High-hyperdiploid</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UK GWAS I</td>
<td>0.06</td>
<td>289</td>
<td>5200</td>
<td>0.86</td>
<td>(0.61-1.22)</td>
<td>4.03 × 10⁻¹</td>
</tr>
<tr>
<td>German GWAS</td>
<td>0.08</td>
<td>176</td>
<td>2024</td>
<td>0.98</td>
<td>(0.64-1.48)</td>
<td>9.11 × 10⁻¹</td>
</tr>
<tr>
<td>UK GWAS II</td>
<td>0.10</td>
<td>251</td>
<td>7385</td>
<td>1.48</td>
<td>(1.06-2.08)</td>
<td>2.13 × 10⁻²</td>
</tr>
<tr>
<td>Combined</td>
<td>0.10</td>
<td>716</td>
<td>14,609</td>
<td>1.10</td>
<td>(0.89-1.35)</td>
<td>0.38</td>
</tr>
<tr>
<td>ETV6-RUNX1-positive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UK GWAS I</td>
<td>0.11</td>
<td>126</td>
<td>5200</td>
<td>2.01</td>
<td>(1.20-3.39)</td>
<td>8.52 × 10⁻³</td>
</tr>
<tr>
<td>German GWAS</td>
<td>0.12</td>
<td>63</td>
<td>2024</td>
<td>1.72</td>
<td>(0.88-3.38)</td>
<td>1.14 × 10⁻¹</td>
</tr>
<tr>
<td>UK GWAS II</td>
<td>0.13</td>
<td>220</td>
<td>7385</td>
<td>2.34</td>
<td>(1.64-3.35)</td>
<td>2.90 × 10⁻⁶</td>
</tr>
<tr>
<td>Combined</td>
<td>0.13</td>
<td>409</td>
<td>14,609</td>
<td>2.14</td>
<td>(1.64-2.80)</td>
<td>3.20 × 10⁻⁸</td>
</tr>
</tbody>
</table>

Note: P-values for each individual study were generated using SNPlTEST 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, I² index to quantify dispersion of odds ratios, CI confidence interval.

### 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 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>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>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>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>2442</td>
<td>14,609</td>
<td>1.32</td>
<td>(1.17-1.49)</td>
<td>2.96 × 10⁻⁶</td>
</tr>
<tr>
<td>High-hyperdiploid</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UK GWAS I</td>
<td>0.06</td>
<td>289</td>
<td>5200</td>
<td>0.86</td>
<td>(0.61-1.22)</td>
<td>4.03 × 10⁻¹</td>
</tr>
<tr>
<td>German GWAS</td>
<td>0.08</td>
<td>176</td>
<td>2024</td>
<td>0.98</td>
<td>(0.64-1.48)</td>
<td>9.11 × 10⁻¹</td>
</tr>
<tr>
<td>UK GWAS II</td>
<td>0.10</td>
<td>251</td>
<td>7385</td>
<td>1.48</td>
<td>(1.06-2.08)</td>
<td>2.13 × 10⁻²</td>
</tr>
<tr>
<td>Combined</td>
<td>0.10</td>
<td>716</td>
<td>14,609</td>
<td>1.10</td>
<td>(0.89-1.35)</td>
<td>0.38</td>
</tr>
<tr>
<td>ETV6-RUNX1-positive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UK GWAS I</td>
<td>0.11</td>
<td>126</td>
<td>5200</td>
<td>2.01</td>
<td>(1.20-3.39)</td>
<td>8.52 × 10⁻³</td>
</tr>
<tr>
<td>German GWAS</td>
<td>0.12</td>
<td>63</td>
<td>2024</td>
<td>1.72</td>
<td>(0.88-3.38)</td>
<td>1.14 × 10⁻¹</td>
</tr>
<tr>
<td>UK GWAS II</td>
<td>0.13</td>
<td>220</td>
<td>7385</td>
<td>2.34</td>
<td>(1.64-3.35)</td>
<td>2.90 × 10⁻⁶</td>
</tr>
<tr>
<td>Combined</td>
<td>0.13</td>
<td>409</td>
<td>14,609</td>
<td>2.14</td>
<td>(1.64-2.80)</td>
<td>3.20 × 10⁻⁸</td>
</tr>
</tbody>
</table>

Note: P-values for each individual study were generated using SNPlTEST 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, I² 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.\textsuperscript{41,49} Prediction of the untyped SNPs was carried out using IMPUTE v2.3.0 based on the data from the 1000 Genomes Project (Phase I integrated variant set, v3.20110121, http://www.1000genomes.org, 9 December 2013) and UK10K (ALSPAC, EGAS00001000000/EGAD00001000195, and TwinsUK, EGAS000010001008/EGAD00001000194, studies only, http://www.uk10k.org/) as reference. In order to correct for genomic inflation post imputation in the German data set, eigenvectors were inferred using the “smartpca” component within EIGENSOFT v2.4 and adjustment was carried out by including the first two eigenvectors as covariates in SNPTEST during association analysis.\textsuperscript{46,47} The inflation factor \(\lambda_{\text{1000G}}\) and \(\lambda_{\text{1,000UK}}\) were again calculated for all SNPs, respectively, for QC.\textsuperscript{13,19} The association between each SNP and risk was calculated using SNPTEST assuming an additive model using a \textsuperscript{virtual 4C} for open chromatin the variant set enrichment method of Cowper-Sal Lari \textsuperscript{60} from the SNPTEST output. Meta-analyses were performed using META v1.7 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/\(\text{SNP} \geq 0.9\), and MAFs \(\leq 0.01\). Collectively the three GWAS provided genotype data on 2442 cases (mean age at diagnosis 5.6 years; 54% male) and 14,609 controls (43% male) with data for 6,755,215 SNPs.\textsuperscript{28,29} We calculated Cochran’s \(Q\) statistic to test for heterogeneity and the \(P\) statistic to quantify the proportion of the total variation that was caused by heterogeneity.\textsuperscript{52} LD metrics were calculated in PLINK\textsuperscript{53} and vcftools\textsuperscript{52} using UK10K genomic data. LD blocks were defined on the basis of hapMap recombination rate, as defined by the Oxford recombination hotspots, and on the basis of the distribution of \(\text{CIs}\).\textsuperscript{54,55} Association plots were generated using visP\textsuperscript{14}.

**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 SN2PHLA.\textsuperscript{29} The imputation was based on a reference panel from the 1000 Genomes Project, consisting of the HLA-region 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 used previously and showed high imputation quality for SNPTEST assuming an additive model using a \textsuperscript{EIGENSOFT v2.4} and adjustment was carried out by including the Oxford recombination hotspots, and on the basis of genomic in

**Sanger sequencing.** To assess the accuracy of imputed genotypes, a random series of samples was Sanger sequenced using BigDye\textsuperscript{®} 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 Table S12.

**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 adapted.\textsuperscript{56} For each risk locus, SNPs in LD were defined (i.e., \(r^2 \geq 0.8\) and \(D^2 > 0.8\)), and termed associated variant set (AVS). Transcription factor ChIP-Seq data for three AML and six childhood ALL cell types were obtained from the Blue-Print\textsuperscript{60} and the normal with variance:

\[
\text{variance} = \mu + \sigma^2
\]

\(\sigma^2\) was set after Bonferroni correction as the variance threshold of 5.7 \times 10^{-5} which shows the presence of significant HLA risk alleles. A significant SNPs included the top SNP from the current meta-analysis from each previously published locus in addition to the two risk loci discovered in this study. The 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 interactions between risk loci. PRS were generated as per Malas et al. assuming a log-normal distribution \(\text{LN}(\mu, \sigma^2)\) with mean \(\mu\), and variance \(\sigma^2\). The population \(\mu\) was set to \(\sigma^2/2\), in order that the overall mean PRS was 1.0.
sibling relative risk were assumed to be 3.2.\(^4\) 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.\(^5\) Typed SNPs were excluded if they had a MAF <0.1 or a HWE p-value <0.01. SNPs were also excluded if a differential missingness rate between cases and controls was P <0.05. In addition, individuals were excluded if having a relatedness score >0.05. 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/gcta/) was employed to estimate the fraction of the phenotypic variance attributed by SNPs given a prevalence of 0.0005 for ALL.\(^6\)

**Data availability.** The UK GWAS I control set comprised 2699 individuals in the 1958 British Birth Cohort (HapI1.2M-Duo Custom arrays 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-EGAD000000000022, EGAD000000000024). The reference panels used in the imputation can be obtained from the 1090 genomes phased haplotypes (n = 1092) from the Phase I integrated variant set release (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101021/) and the UK10k (n = 3781; EGAS000000000909, EGAD000000000195, EGAS000000001018; www.uk10k.org) sequenced data sets. eQTL data for various functional analyses were obtained from the MuTHER studies (genome-wide expression profiled 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.nature.com/njgr/journal/v45/n10/abs/ncg4276.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 ALL and AML cells were obtained from the BluePrint Epigenome (dcc.blueprint-epigenome.eu) for samples S00FGCH1, S005GFF1, S00KBPH1, S017EHH1, S0179DH1, S01GQHH1, S0176H1, 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 EGAD00001002809. 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**

15. Wellcome Trust Case Control, C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 477, 661–678 (2007).
44. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new
multinot possible method for genome-wide association studies by imputation
47. Price, A. L. et al. Principal components analysis corrects for stratification in
49. Delaneau, O., Zagury, J. F. & Marchini, J. Improved whole-chromosome phasing
for disease and population genetic studies. Nat. Methods 10, 5–6
(2013).
50. de Bakker, P. I. et al. Practical aspects of imputation-driven meta-analysis of
51. Liu, J. Z. et al. Meta-analysis and imputation refines the association of 15q25 with
52. Higgins, J. P. & Thompson, S. G. Quantifying heterogeneity in a meta-
53. Danecek, P. et al. The variant call format and VCFtools. Bioinformatics 27,
54. Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. A fine-scale
map of recombination rates and hotspots across the human genome. Science
56. Westra, H.-J. et al. Systematic identi
57. Martin, J. S. et al. HUGIn: Hi-C unifying genomic interrogator.
58. Schmitt, A. D., Hu, M. & Ren, B. Genome-wide mapping and analysis of
60. Westra, H.-J. et al. Systematic identification of trans eQTLs as putative drivers
62. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies
63. Nica, A. C. et al. The architecture of gene regulatory variation across multiple
(2016).
64. Dixon, J. R. et al. Topological domains in mammalian genomes identified by
66. Nica, A. C. et al. The architecture of gene regulatory variation across multiple
(2016).
(2008).
68. Vora, A. et al. Toxicity and ef
70. R: A language and environment for statistical computing (R Foundation
72. R: A language and environment for statistical computing (R Foundation
(2008).
75. Vora, A. et al. Toxicity and ef
76. Pharoah, P. D., Antoniou, A. C., Easton, D. F. & Ponder, B. A. Polygenes, risk
77. R: A language and environment for statistical computing (R Foundation
(2008).
80. Vora, A. et al. Toxicity and ef
82. R: A language and environment for statistical computing (R Foundation
(2008).
85. Vora, A. et al. Toxicity and ef
86. Pharoah, P. D., Antoniou, A. C., Easton, D. F. & Ponder, B. A. Polygenes, risk
87. R: A language and environment for statistical computing (R Foundation