












ARTICLE

<https://doi.org/10.1038/s41467-019-13069-6>

OPEN

# Identification of four novel associations for B-cell acute lymphoblastic leukaemia risk

Jayaram Vijayakrishnan <sup>1,19</sup>, Maoxiang Qian<sup>2,3,19</sup>, James B. Studd <sup>1</sup>, Wenjian Yang<sup>2</sup>, Ben Kinnersley <sup>1</sup>, Philip J. Law <sup>1</sup>, Peter Broderick <sup>1</sup>, Elizabeth A. Raetz<sup>4</sup>, James Allan<sup>5</sup>, Ching-Hon Pui <sup>6,7</sup>, Ajay Vora<sup>8</sup>, William E. Evans <sup>2,7</sup>, Anthony Moorman<sup>9</sup>, Allen Yeoh<sup>10,11</sup>, Wentao Yang<sup>2</sup>, Chunliang Li <sup>12</sup>, Claus R. Bartram<sup>13</sup>, Charles G. Mullighan <sup>6,7,14</sup>, Martin Zimmerman<sup>15</sup>, Stephen P. Hunger<sup>16</sup>, Martin Schrappe<sup>17</sup>, Mary V. Relling<sup>2,7</sup>, Martin Stanulla<sup>15</sup>, Mignon L. Loh<sup>18</sup>, Richard S. Houlston <sup>1\*</sup> & Jun J. Yang <sup>2,6,7\*</sup>

There is increasing evidence for a strong inherited genetic basis of susceptibility to acute lymphoblastic leukaemia (ALL) in children. To identify new risk variants for B-cell ALL (B-ALL) we conducted a meta-analysis with four GWAS (genome-wide association studies), totalling 5321 cases and 16,666 controls of European descent. We herein describe novel risk loci for B-ALL at 9q21.31 (rs76925697,  $P = 2.11 \times 10^{-8}$ ), for high-hyperdiploid ALL at 5q31.1 (rs886285,  $P = 1.56 \times 10^{-8}$ ) and 6p21.31 (rs210143 in *BAK1*,  $P = 2.21 \times 10^{-8}$ ), and *ETV6-RUNX1* ALL at 17q21.32 (rs10853104 in *IGF2BP1*,  $P = 1.82 \times 10^{-8}$ ). Particularly notable are the pleiotropic effects of the *BAK1* variant on multiple haematological malignancies and specific effects of *IGF2BP1* on *ETV6-RUNX1* ALL evidenced by both germline and somatic genomic analyses. Integration of GWAS signals with transcriptomic/epigenomic profiling and 3D chromatin interaction data for these leukaemia risk loci suggests deregulation of B-cell development and the cell cycle as central mechanisms governing genetic susceptibility to ALL.

<sup>1</sup> Division of Genetics and Epidemiology, The Institute of Cancer Research, 15 Cotswold Road, Sutton, Surrey SM2 5NG, UK. <sup>2</sup> Department of Pharmaceutical Sciences, St. Jude Children's Research Hospital, Memphis, Tennessee, USA. <sup>3</sup> Children's Hospital and Institutes of Biomedical Sciences, Fudan University, Shanghai, China. <sup>4</sup> Division of Pediatric Hematology and Oncology, New York University Langone Health, New York, New York, USA. <sup>5</sup> Northern Institute for Cancer Research, Newcastle University, Newcastle upon Tyne NE2 4HH, UK. <sup>6</sup> Department of Oncology, St. Jude Children's Research Hospital, Memphis, Tennessee, USA. <sup>7</sup> Hematological Malignancies Program, St. Jude Children's Research Hospital, Memphis, Tennessee, USA. <sup>8</sup> Great Ormond Hospital, London, UK. <sup>9</sup> Wolfson Childhood Cancer Research Centre, Northern Institute for Cancer Research, Newcastle University, Newcastle upon Tyne NE1 7RU, UK. <sup>10</sup> Centre for Translational Research in Acute Leukaemia, Department of Paediatrics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. <sup>11</sup> VIVA-University Children's Cancer Centre, Khoo Teck Puat-National University Children's Medical Institute, National University Hospital, National University Health System, Singapore, Singapore. <sup>12</sup> Department of Tumor Cell Biology, St. Jude Children's Research Hospital, Memphis, Tennessee, USA. <sup>13</sup> Institute of Human Genetics, University Hospital, Heidelberg, Germany. <sup>14</sup> Department of Pathology, St. Jude Children's Research Hospital, Memphis, Tennessee, USA. <sup>15</sup> Department of Paediatric Haematology and Oncology, Hannover Medical School, 30625 Hannover, Germany. <sup>16</sup> Department of Paediatrics and Centre for Childhood Cancer Research, Children's Hospital of Philadelphia and the Perelman School of Medicine at The University of Pennsylvania, Philadelphia, Pennsylvania, USA. <sup>17</sup> Department of Paediatrics, University Medical Centre Schleswig-Holstein, Kiel, Germany. <sup>18</sup> Department of Pediatrics, Benioff Children's Hospital and the Helen Diller Family Comprehensive Cancer Center, University of California San Francisco, San Francisco, California, USA. <sup>19</sup> These authors contributed equally: Jayaram Vijayakrishnan, Maoxiang Qian. \*email: [richard.houlston@icr.ac.uk](mailto:richard.houlston@icr.ac.uk); [Jun.Yang@stjude.org](mailto:Jun.Yang@stjude.org)

Acute lymphoblastic leukaemia (ALL) is the most common paediatric cancer with B-cell precursor ALL (B-ALL) accounting for ~85% of the cases<sup>1</sup>. Although the peak age of diagnosis of ALL is between ages 2 and 5 years, some initiating somatic genomic abnormalities (e.g., chromosomal translocations) can be detectable at birth<sup>2,3</sup>. Both the absence of specific environmental risk factors and early onset suggest a strong inherited genetic basis for susceptibility<sup>4–6</sup>. Our understanding of ALL susceptibility has been informed by genome-wide association studies (GWAS) identifying 11 regions harbouring risk variants: 7p12.2 (*IKZF1*), 8q24.21, 9p21.3 (*CDKN2A/B*), 10p12.2 (*PIP4K2A*), 10q26.13 (*LHPP*), 12q23.1 (*ELK3*), 10p14 (*GATA3*), 10q21.2 (*ARID5B*), 14q11.2 (*CEBPE*), 16p13.3 (*USP7*) and 21q22.2 (*ERG*)<sup>7–16</sup>. ALL is a biologically heterogeneous disease with subtypes defined by recurrent initiating genetic abnormalities. After initiation, however, leukaemia cells acquire a constellation of secondary lesions. The two most common subtypes of B-ALL are *ETV6-RUNX1* fusion positive and high-hyperdiploid (HD) ALL<sup>17</sup>, each accounting for 20–25% of cases. HD ALL is characterised by a chromosome number > 51 due to the non-random gain of specific chromosomes. Subtype-specific GWAS associations have so far been identified at 10q21.2 (*ARID5B*) associated with HD ALL, 10p14 (*GATA3*) for Philadelphia chromosome-like ALL, and 2q22.3 associated with *ETV6-RUNX1*-positive ALL<sup>7,9,12,18,19</sup>.

To gain a more comprehensive insight into susceptibility to ALL, we performed a meta-analysis of four GWAS from the North America<sup>13,18,20</sup> and Europe<sup>7,9,12</sup>, with additional replication. We report both the discovery of four new susceptibility regions for ALL and refined risk estimates for the previously reported loci. In addition, we have investigated the gene regulatory mechanisms underlying the genetic associations observed at these risk loci by integrating genome-wide chromosome conformation capture (Hi-C) data and chromatin immunoprecipitation-sequencing (ChIP-seq), epigenomic and transcriptomic profiling to pinpoint target genes.

## Results

**GWAS meta-analysis and replication.** We conducted a meta-analysis of four GWAS B-ALL datasets: UK GWAS I, German GWAS, UK GWAS II and the COG\_SJ GWAS<sup>7,9,12,13,18,20</sup>, totalling 5321 cases and 16,666 controls of European descent. Following established quality-control measures for each GWAS dataset (Supplementary Fig. 1), the genotypes of ~10 million single-nucleotide polymorphisms (SNPs) in each study were imputed. After filtering out SNPs on the basis of minor allele frequency (MAF) and imputation quality, we assessed associations between ALL status and SNP genotype in each study using logistic regression. Risk estimates were combined through an inverse-variance-weighted fixed-effects meta-analysis<sup>21,22</sup>. Quantile–quantile (Q–Q) plots for SNPs did not show evidence of substantive over dispersion ( $\lambda_{GC}$  values 1.02–1.08; Supplementary Fig. 2). Given the biological heterogeneity of ALL, as evidenced by subtype-specific associations at a number of previously published regions<sup>9,12,18</sup>, we analysed the association between genotype and all B-ALL cases, and the common subtypes of HD and *ETV6-RUNX1*-positive ALL. Risk loci that were genome-wide significant only with a particular ALL subtype were defined as subtype-specific associations.

Meta-analysis identified 16 risk loci above genome-wide significance ( $P < 5 \times 10^{-8}$ , by inverse-variance method based on a fixed-effects model), of which 10 are previously reported B-ALL risk loci (Fig. 1 and Supplementary Table 1). Of the six new genome-wide significant candidate risk loci, one was generic to all B-ALL, three were specific for high-HD ALL and two were

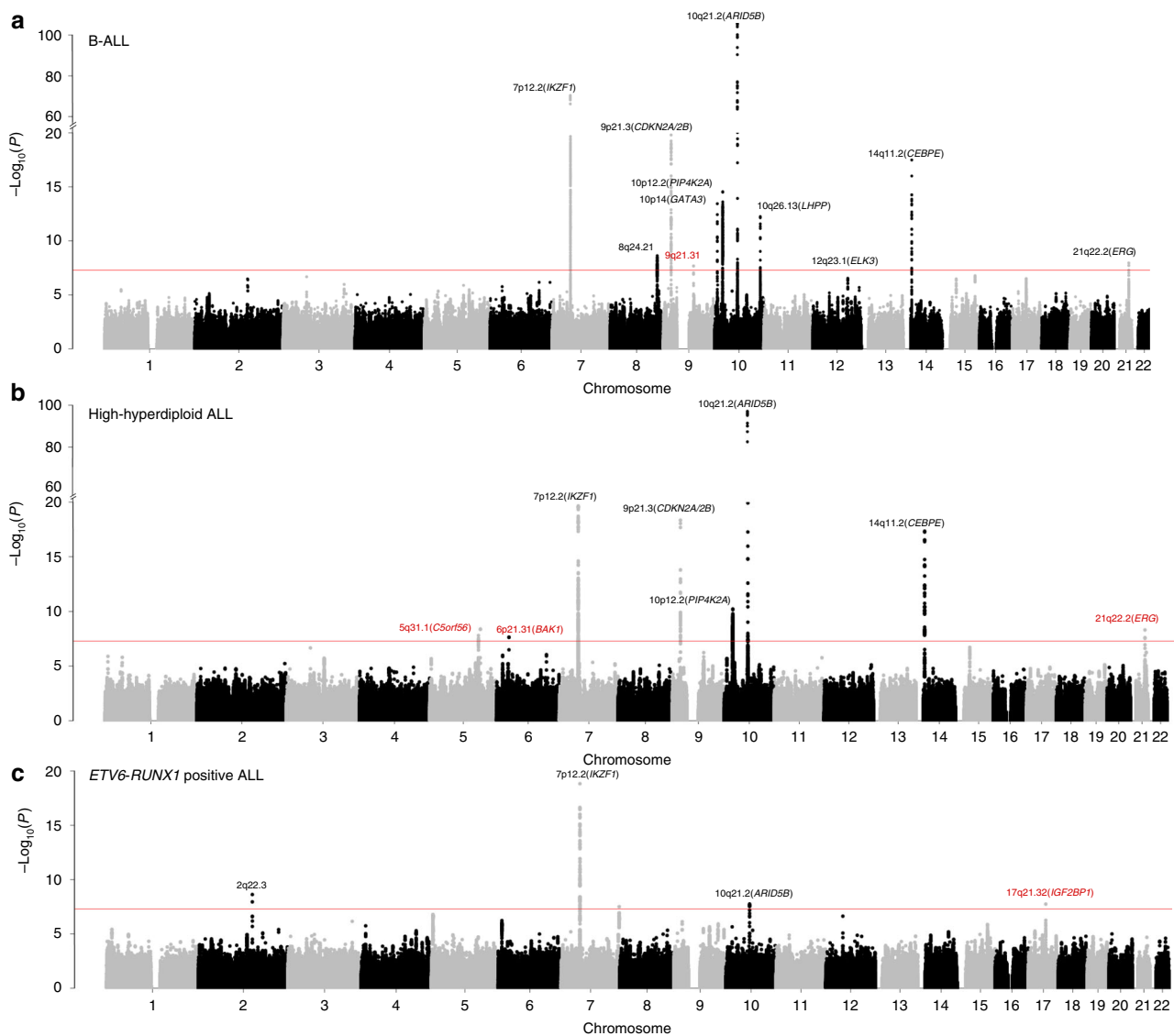
specific for *ETV6-RUNX1*-positive ALL (Supplementary Table 1). These six SNP associations were interrogated in an independent series of 2237 cases and 3461 controls (COG\_SJ GWAS non-European American (EA); Supplementary Tables 2 and 3). Four of the six SNPs were validated in the replication series ( $P < 0.05$ , by additive logistic regression test): for all B-ALL at 9q21.3 (rs76925697, nearest gene *TLE1*), for HD ALL at 5q31.1 (rs886285, *C5orf56*) and 6p21.31 (rs210143, *BAK1*), and for *ETV6-RUNX1*-positive ALL at 17q21.32 (rs10853104, *IGF2BP1*) (Table 1 and Supplementary Tables 2, 4 and 5). In addition to providing further evidence for the 21q22.2 association for all B-ALL<sup>14</sup>, we also identified a subtype-specific association for HD with rs9976326 (Table 1 and Supplementary Tables 1, 4 and 5).

Next, we performed a conditional analysis on the sentinel risk SNP at each locus to search for further independent signals at new and previously reported risk regions. We confirmed the presence of previously reported dual association signals at 9p21.3 (*CDKN2A/B*) and 10p12.2 (*PIP4K2A*) (Supplementary Table 6). In addition, independent risk variants were identified at 21q22.2 (*ERG*) and 7p12.2 (*IKZF1*) (Supplementary Table 7 and Supplementary Figs. 3, 4, 5, 6 and 7).

**Functional annotation of new risk loci.** To gain insight into the biological basis of association signals, we examined the epigenetic landscape of risk regions in B cells. For each of the new risk regions, we evaluated chromatin profiles using ChromHMM, ATAC-seq data in primary B cells from the Roadmap Epigenomics consortia<sup>23</sup>, and the GM12878 lymphoblastoid cell line from ENCODE<sup>24,25</sup> (Fig. 2 and Supplementary Figs. 3, 4, 5, 6, 7). As the strongest associated GWAS SNP may not represent the causal variant, we examined variants in linkage disequilibrium (LD) with the top risk SNP in each region (defined by  $r^2 > 0.8$ ,  $P < P_{\min} \times 50$ ; Supplementary Table 8). Genomic spatial proximity and chromatin looping between non-coding DNA and target genes are key to gene regulation; we therefore interrogated promoter capture Hi-C (CHiC) data from naive B cells<sup>26</sup> (Supplementary Table 9) as well as Hi-C and H3K27Ac ChIP data in human ALL cells<sup>27</sup> (Supplementary Fig. 8). We also sought to identify target genes by performing quantitative trait locus (QTL) analysis of mRNA expression (eQTL) data from GTEx<sup>28</sup>, Blood eQTL<sup>29</sup>, MuTHER<sup>30</sup> and CAGE<sup>31</sup> databases, and DNA methylation (mQTL) (Supplementary Table 10). We annotated risk loci with variants mapping to haematopoietic transcription factor (TF)-binding sites (Fig. 2, Supplementary Figs. 3, 4, 5, 6 and 7, and Supplementary Table 11). Using Summary data-based Mendelian Randomisation (SMR) analysis, we examined for pleiotropy between GWAS signal and *cis*-eQTL for genes within 1 Mb of the sentinel SNP to identify a possible causal relationship between gene expression and disease (Supplementary Tables 12 and 13).

Lead SNPs at 6p21 are located within an intron 1 kb downstream of the *BAK1* transcription start site and possess histone marks characteristic of active promoter activity and open chromatin accessibility (Fig. 2a). The top SNP, rs210143, falls within a TF-binding cluster and the C-risk allele is associated with reduced *BAK1* expression ( $P_{\text{Blood}} = 3.3 \times 10^{-310}$ , by linear regression test). SMR analysis confirmed a significant association with *BAK1* expression and ALL consistent with a likely causal relationship (Supplementary Table 12). The 6p21 association was confined to HD ALL only, whereas risk variants did not reach genome-wide significance for either *ETV6-RUNX1* or all B-ALL. *BAK1* was not differentially expressed in leukaemic blasts from any ALL subtype (Supplementary Fig. 9).

The HD ALL-specific association at 5q31 (*C5orf56*) localises to genomic regions featuring ChIP-seq marks indicative of



**Fig. 1** Manhattan plots of association for **a** B-ALL, **b** high-hyperdiploid ALL and **c** *ETV6-RUNX1*-positive ALL. y-axis shows genome-wide  $P$ -values (two-sided, calculated using SNPTTEST v2.5.2 assuming an additive model) of > 6 million successfully imputed autosomal SNPs in 5321 cases and 16,666 controls. The x-axis shows the chromosome number. The red horizontal line represents the genome-wide significance threshold of  $P = 5.0 \times 10^{-8}$ . New associations are labelled in red. Other risk loci were reported in previous GWAS using subsets of ALL cohorts included herein.

regulatory elements. Although SNP rs2522044 is eQTL for *SLC22A4* and *C5orf56* ( $P_{\text{Blood}} = 8.8 \times 10^{-51}$  and  $7.3 \times 10^{-12}$ , respectively, by linear regression test), a looping interaction between the top SNP rs886285 and the immune regulatory gene *IRF1* was observed (Fig. 2b and Supplementary Table 9). SMR analysis did not reveal any association with *C5ORF56*, *SLC22A4* or *IRF1* expression, nor did these genes show subtype-specific expression in ALL blasts (Supplementary Table 12).

Risk SNPs at 17q21 localising to the second intron of *IGF2BP1* lack evidence of cis-regulatory activity. However, the strongest associated SNP, rs10853104, maps to a TF-binding cluster and is predicted to disrupt a conserved CTCF-binding motif (Fig. 2c), suggesting an influence on topological-associated domain structure. As the 17q21 association was unique to *ETV6-RUNX1*-positive ALL, we investigated the relationship between ALL subtype and expression of genes within 1 Mb of rs10853104. *ETV6-RUNX1*-positive ALL cells showed significant overexpression of *IGF2BP1* compared with other ALL subtypes

(Supplementary Figs. 10 and 11;  $P = 3.68 \times 10^{-23}$ , by two-sided Wilcoxon's rank-sum test).

The lead SNP rs76925697 at a new B-ALL risk locus in 9q21 resides 500 kb centromeric to *TLE1* within a genomic region devoid of chromatin marks indicative of regulatory function (Fig. 2d). We also did not observe any evidence for eQTLs or TF binding. However, in ALL cells, the region containing the risk SNP showed strong looping with a distal enhancer within *TLE1* (Supplementary Fig. 8). Finally, we identified an HD ALL-specific association at the previously reported 21q22 locus within intron 3 of *ERG*. Notably, the T-risk allele of the lead SNP rs9976326 is predicted to disrupt binding of the haematological TF AML1/RUNX1 and is associated with reduced gene methylation.

Transcriptome-wide association studies (TWASs) investigating the association of genetically predicted gene expression with disease can identify new susceptibility genes by aggregating evidence across variants, thereby increasing study power. We

**Table 1 Summary of results for genome-wide significant childhood ALL risk loci.**

| CHR | SNP (Subtype)                     | Locus (gene)                | Position (BP) | Risk allele | RAF  | OR (95% CI)      | P-value                 |
|-----|-----------------------------------|-----------------------------|---------------|-------------|------|------------------|-------------------------|
| 2   | rs17481869( <i>ETV6-RUNX1</i> )   | 2q22.3                      | 146124454     | A           | 0.08 | 1.74 (1.45–2.09) | $2.37 \times 10^{-09}$  |
| 5   | *rs886285 (High-Hyperdiploidy)    | 5q31.1 ( <i>C5orf56</i> )   | 131765206     | T           | 0.34 | 1.29 (1.18–1.41) | $1.56 \times 10^{-08}$  |
| 6   | *rs210143 (High-Hyperdiploidy)    | 6p21.31 ( <i>BAK1</i> )     | 33546930      | C           | 0.73 | 1.30 (1.19–1.43) | $2.21 \times 10^{-08}$  |
| 7   | rs17133805                        | 7p12.2 ( <i>IKZF1</i> )     | 50477514      | G           | 0.32 | 1.65 (1.56–1.74) | $5.28 \times 10^{-71}$  |
| 8   | rs75777619                        | 8q24.21                     | 130185176     | G           | 0.12 | 1.26 (1.17–1.36) | $2.30 \times 10^{-09}$  |
| 9   | *rs76925697                       | 9q21.31                     | 83747371      | A           | 0.96 | 1.52 (1.31–1.76) | $2.11 \times 10^{-08}$  |
| 9   | rs113650570                       | 9p21.3 ( <i>CDKN2A</i> )    | 21976402      | A           | 0.02 | 2.32 (2.03–2.65) | $8.06 \times 10^{-35}$  |
| 10  | rs10821936                        | 10q21.2 ( <i>ARID5B</i> )   | 63723577      | C           | 0.33 | 1.80 (1.71–1.89) | $1.19 \times 10^{-106}$ |
| 10  | rs3824662                         | 10p14 ( <i>GATA3</i> )      | 8104208       | A           | 0.19 | 1.29 (1.21–1.38) | $3.57 \times 10^{-14}$  |
| 10  | rs2296624                         | 10p12.2 ( <i>PIP4K2A</i> )  | 22856946      | C           | 0.67 | 1.25 (1.18–1.32) | $2.79 \times 10^{-15}$  |
| 10  | rs12779301                        | 10q26.13 ( <i>LHPP</i> )    | 126292655     | C           | 0.66 | 1.22 (1.15–1.29) | $5.72 \times 10^{-13}$  |
| 12  | rs4762284                         | 12q23.1 ( <i>ELK3</i> )     | 96612762      | T           | 0.32 | 1.15 (1.12–1.19) | $3.75 \times 10^{-07}$  |
| 14  | rs2239630                         | 14q11.2 ( <i>CEBPE</i> )    | 23589349      | A           | 0.45 | 1.28 (1.22–1.35) | $1.72 \times 10^{-21}$  |
| 17  | *rs10853104 ( <i>ETV6-RUNX1</i> ) | 17q21.32 ( <i>IGF2BP1</i> ) | 47092076      | T           | 0.47 | 1.33 (1.21–1.47) | $1.82 \times 10^{-08}$  |
| 21  | rs9976326 (High-Hyperdiploidy)    | 21q22.2 ( <i>ERG</i> )      | 39776485      | T           | 0.25 | 1.33 (1.21–1.46) | $4.79 \times 10^{-09}$  |

BP base pair, CHR chromosome, CI confidence intervals, OR odds ratio, RAF risk allele frequency. OR and CI are derived from current meta-analysis. \*New loci discovered in current meta-analyses. Other risk loci were reported in previous GWAS using subsets of ALL cohorts included herein.

performed a TWAS integrating genomic and expression data<sup>32</sup>. This analysis confirmed the risk loci described above but did not identify any additional associations independent of GWAS signals, which were statistically significant (Supplementary Figs. 12 and 13).

To implicate recurrent disruption of TF-binding sites at ALL risk loci genome wide, we performed TF-binding enrichment analysis as per Cowper-Salari et al.<sup>33</sup>. This analysis identified over-representation of TF binding at risk SNPs compared with a random SNPs subset. A number of TFs somatically mutated in B-ALL, including *PBX1* (Benjamini–Hochberg corrected  $P$ -value [ $P_{BH}$ ] = 0.007), *TCF3* ( $P_{BH}$  = 0.007), *ETS1* ( $P_{BH}$  = 0.009), *RUNX1* ( $P_{BH}$  = 0.012) and *ERG* ( $P_{BH}$  = 0.030) (Supplementary Fig. 14) were enriched at risk loci providing evidence that germline variation and somatic alterations may impact on the same pathways. In addition, we identified *BRD4* ( $P_{BH}$  = 0.007) and *NR3C1* ( $P_{BH}$  = 0.009) binding sites as significantly enriched at risk loci, suggesting their disruption contributes to leukaemogenesis.

**Relationship between new risk alleles and clinical features.** We did not find an association between sex or age at diagnosis of ALL with the new risk SNPs using case-only analysis. We also found no statistically significant relationship between SNP genotype and patient outcome using data from German and COG\_SJ GWAS cohorts<sup>20,34</sup>. A failure to demonstrate additional relationships may, however, be reflective of limited statistical power.

**Contribution of risk SNPs to heritability.** Using LD-adjusted kinships (LDAK)<sup>35</sup>, the heritability of ALL ascribable to all common variation was identified as 21% (SD  $\pm$  0.065) (Supplementary Table 14). Together, the risk loci identified so far accounted for 31% of the total variance in genetic risk of ALL (Supplementary Table 15). To assess the collective impact of all identified risk SNPs we constructed polygenic risk scores (PRS) considering the combined effect of all risk SNPs modelled under a log-normal relative risk distribution after correcting the Z-scores for Winner's curse using FIQT<sup>36</sup>. Based on their PRS score, an individual in the top 1% of genetic risk would have a 4.7-fold increased risk of ALL when compared with an individual with median genetic risk (Supplementary Fig. 15).

## Discussion

Our analysis provides evidence of four new associations with the risk of developing ALL. Besides providing additional evidence for

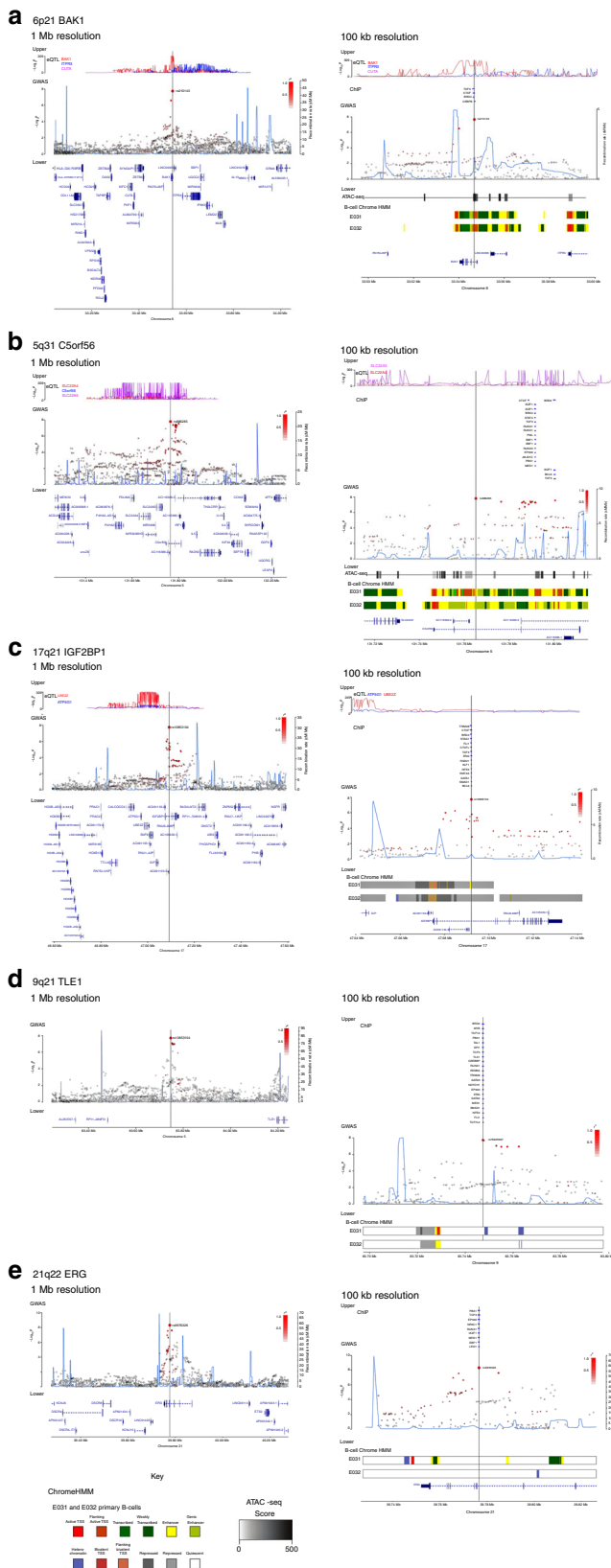
genetic susceptibility to ALL, these new risk loci provide further insights into the biological basis of ALL development. Integrating information from Hi-C data with chromatin profiling and eQTL/mQTL data implicates a number of genes with strong a priori evidence as the functional basis of associations, e.g., at 6p21.31 the pro-apoptotic protein *BAK1*, at 21q22.2 the haematological ETS TF *ERG* and at 17q21.32 proliferation factor *IGF2BP1*. Conditional analysis revealed two novel secondary associations at 7p12.2 (*IKZF1*) and 21q22 (*ERG*), in addition to the previously identified signals at 10p12.2 (*PIP4K2A*) and 9p21.3 (*CDKN2A/2B*). Two of the genome-wide significant associations from our discovery meta-analysis were not replicated. This may be the consequence of a different population allelic structure between cohorts of different ancestry (Europeans in the discovery and non-Europeans in the replication) or population-specific associations<sup>14</sup>. Our recently discovered T-ALL risk locus *USP7* was also not significant in this GWAS because of its lineage-specific effect on ALL susceptibility<sup>15</sup>.

*BAK1* is essential for B-cell homeostasis and its knockout mice accumulate immature and mature follicular B cells. *BAK1* induces apoptosis by binding to and antagonising anti-apoptotic proteins, including *BCL2*<sup>37–39</sup>. Reduced *BAK1* expression relieves repression of *BCL2*, inhibiting apoptosis and conferring a pro-survival advantage<sup>40</sup>. Proximity of the lead 6p21 risk variant, rs210143-T, to the *BAK1* promoter and a strong negative association with expression suggests decreased *BAK1* promotes ALL leukaemogenesis. Notably, the 6p21 region is also pleiotropic, influencing chronic lymphocytic leukaemia (CLL) and testicular cancer risk. Moreover, the strongest association for both CLL and ALL is rs210143, suggesting a similar mechanistic basis.

The 12q21 association at *IGF2BP1* is specific for *ETV6-RUNX1*-positive ALL and this subtype also significantly over-expresses *IGF2BP1*<sup>41</sup>. We did not observe a significant association between *IGF2BP1* genotype and its expression in *ETV6-RUNX1* ALL, plausibly because the subtle effects of this germline risk variant on *IGF2BP1* transcription were masked by the drastic upregulation as a result of *ETV6-RUNX1* fusion. The subtype-specific nature of the association may be explained by the observation that in *ETV6-RUNX1* positive ALL *IGF2BP1* binds to the *ETV6-RUNX1* transcript increasing its stability and expression<sup>42</sup>. *IGF2BP1* has been implicated in promoting proliferation and cell survival via the post-transcriptional regulation of a number of genes including *KRAS*, *MYC* and *PTEN*<sup>43</sup>.

Our analysis confirms the ALL association at 21q22 (*ERG*) recently reported in Hispanics<sup>44</sup>. In addition, we report a new





**Fig. 2** Regional plots of association results and recombination rates for the newly identified risk loci. **a** 6p21 (rs210143), **b** 5q31 (rs886285), **c** 17q21, **d** 9q21.3 (rs76925697), **e** 21q22 (rs9976326). Loci are shown at both 1 Mb (left) and 100 kb (right) resolutions. Upper panes show FDR corrected eQTL  $P$ -values extracted from the Blood database; ChIP transcription factor binding sites shown as blue bars. GWAS plots show association  $-\log_{10}P$ -values (left y-axis) of SNPs shown according to their chromosomal positions (x-axis). Light blue line shows recombination rates in (cM/Mb) from UK10K Genomes Project (right y-axis). Lead SNPs are denoted by large circles labelled by rsID. Colour intensity of each symbol reflects LD, white ( $r^2 = 0$ ), dark red ( $r^2 = 1.0$ ). Genome coordinates are from NCBI human genome GRCh37. Lower pane shows chromatin-state segmentation tracks (ChromHMM) from primary B cells and gene positions from Gencode v27 comprehensive gene annotation. Where no significant results were obtained upper and lower panes are omitted.

domain-containing TF important for normal hematopoietic development. Somatic alteration of *ERG* is recurrent in ALL and rs9976326 is in close proximity to hotspot deletions<sup>45</sup>.

Although risk SNPs at 5q31 reside in *C5orf56*, which has no established role in B-cell biology, Hi-C interactions implicate the TF IRF1, which is required for normal T-cell development and is deleted in 50% of acute myelogenous leukaemia<sup>46,47</sup>. The intergenic region at 9q21.3 (near *TLE1*) has no clear candidate and the biological basis of the association is unclear.

TF-enrichment analysis revealed gene *BRD4* with no previous indication from germline or somatic studies in ALL and the gene *NR3C1* whose alterations are associated with poor outcome and high risk in ALL patients<sup>48</sup>. *BRD4*, a member of the BET protein family, is a transcriptional co-activator that binds acetylated histones recruiting TFs to DNA. *BRD4* has been found to co-localise with the lymphoid TFs SPI1, FLI1, *ERG*, MYB and CEBP $\alpha/\beta$ <sup>49</sup>, and this may account for its enrichment at risk loci. Several groups have shown activity of BET inhibitors in AML cells lines<sup>50,51</sup>. *NR3C1* is the glucocorticoid receptor, the target of the immunomodulatory hormones glucocorticoids, including cortisol, and drugs including dexamethasone and prednisone. The effect of these compounds is potent immune-suppression and reduced inflammation. Glucocorticoid treatment reduces circulating B-cell numbers<sup>52</sup> and induces cell death in ALL cells by lowering the expression of B-cell survival factors<sup>53</sup>. Further validation will be required to establish a role of disrupted *NR3C1* signalling in the genesis of ALL.

Deciphering the functional consequences of risk loci is inherently challenging, as analyses are complicated by background haplotype structure. We have relied in part on integration of GWAS signals with in silico and publicly accessible epigenetic data; hence, these predictions require experimental verification through functional assays in the future.

In summary, our study provides further evidence for inherited susceptibility to ALL and support for subtype specificity at risk loci. The different subtypes of B-ALL presumably reflect the different aetiology and evolutionary trajectories of progenitor cells influenced by inherited variation. Our findings further support a model of ALL susceptibility based on transcriptional dysregulation consistent with altered B-cell differentiation, where dysregulation of apoptosis and cell cycle signalling features as recurrently modulated pathways. Genes elucidated from GWAS functional annotation may represent promising therapeutic targets for drug discovery. Finally, although our GWAS meta-analysis is the largest of its kind, greater sample sizes are likely to uncover additional associations underscoring the need for collaborative analyses.

HD ALL association with rs9976326. The SNP rs9976326 and the top SNP reported in Hispanics (rs2836371) are separated by 3 kb and correlated (pairwise LD values  $r^2 = 0.52$ ,  $D' = 0.85$  and  $r^2 = 0.60$ ,  $D' = 0.87$  in European and admixed Americans 1000 genomes populations, respectively). *ERG* encodes an *ETS*

## Methods

**Ethics.** Collection of samples and clinical information was undertaken with informed consent and ethical review board approval. Specifically, Medical Research Council UKALL97/99 trial by UK therapy centres and approval for UKALL2003 from the Scottish Multi-Centre Research Ethics Committee (REC:02/10/052), the UK Bloodwise Childhood Leukaemia Cell Bank, the United Kingdom Childhood Cancer Study, and University of Heidelberg; AALL0232 (clinicaltrials.gov NCT00075725)<sup>54</sup> and P9904/P9905/P9906 (NCT00005585/NCT00005596/NCT00005603)<sup>55</sup> from the Children's Oncology Group (COG); and Total Therapy XIII/XV (NCI-T93-0101D/NCT00137111)<sup>56,57</sup> from the St. Jude Children's Research Hospital. The diagnosis of ALL was established in accordance with World Health Organization guidelines.

**GWAS data.** The four GWAS datasets have been the subject of previous publications: (i) UK GWAS I—824 cases, 2699 controls from the 1958 British Birth Cohort and 2501 controls from the UK Blood Service controls<sup>7</sup>; (ii) German GWAS—1155 Berlin–Frankfurt–Münster (BFM) trial (1993–2004) cases, 2132 Heinz Nixdorf Recall study controls<sup>9</sup>; (iii) UK GWAS II—1021 cases from Medical Research Council UK ALL-2003 and ALL-97/99 trials, 2976 PRACTICAL Consortium and 4446 Breast Cancer Association Consortium controls<sup>12</sup>; (iv) COG\_SJ GWAS—2,879 cases of European ancestry from the COG AALL0232, COG P9904/P9905/P9906, St. Jude Total Therapy XIII/XV and 2057 non-ALL controls of European ancestry from the Multi-Ethnic Study of Atherosclerosis (MESA) study (dbGAP phs000209.v9)<sup>13,18,20</sup>.

The replication study included 2237 cases and 3461 non-ALL controls of non-European ancestry from the same cohort as COG\_SJ GWAS.

The UK GWAS I, UK GWAS II and German GWAS series were genotyped using Illumina Human 317K Human OmniExpress-12v1.0 or Infinium OncoArray-500K arrays. The COG\_SJ GWAS and replication series were genotyped using Affymetrix Human SNP 6.0 (St. Jude Total XVI, COG P9904/9905, MESA) and Affymetrix GeneChip Human 500k Mapping arrays (St. Jude Total XIII/XV and COG P9906).

**Statistical analysis of GWAS data.** Analyses were undertaken using R v3.2.3<sup>58</sup>, PLINK v1.9<sup>59</sup>, SNPTEST v2.5.2<sup>22</sup> and IMPUTE v2.3<sup>60</sup> software. Standard quality-control measures were applied to each GWAS<sup>61</sup>. 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 (and Native American in COG\_SJ dataset) as a reference) were excluded for discovery GWAS and meta-analysis. SNPs with call rate < 95% were excluded or showed deviation from Hardy–Weinberg equilibrium ( $P < 10^{-5}$ ). Appropriateness case–control matching was evaluated using Q–Q plots inflation test statistics. The inflation factor  $\lambda$  was calculated to indicate the degree of genomic inflation, by dividing the median of the test statistics by the median expected values from a  $\chi^2$  distribution with 1 degree of freedom (Supplementary Fig. 2). Prediction of the untyped genotypes was carried out using 1000 Genomes Project (Phase 1) and UK10K as reference<sup>62,63</sup>. To account for genomic inflation post imputation, top Eigenvectors from the principal component analysis were used as covariates in the final association analysis<sup>64</sup>; the top two and five Eigenvectors for the UK\_German GWAS and the COG\_SJ GWAS, respectively. No further adjustments for  $P$ -values were applied. The association between each SNP and risk was calculated assuming an additive model and meta-analyses were performed using META v1.7<sup>21,22</sup>. Association meta-analyses only included SNPs with info score > 0.8, imputed call rates > 0.9 and MAFs > 0.01. We calculated Cochran's  $Q$  statistic to test for heterogeneity and the  $I^2$  statistic to quantify the proportion of the total variation that was caused by heterogeneity.

In COG\_SJ dataset, genetic ancestry (European [CEU], African [YRI], East Asian [JPT/CHB] and Native American) was determined by using ADMIXTURE (version 1.3.0)<sup>65</sup>, with the sum of these four ancestries being 100% for any given subject. EA, African American and Asian were defined as having > 95% European genetic ancestry, > 70% African ancestry and > 90% Asian ancestry, respectively. Hispanics were individuals for whom Native American ancestry was > 10% and greater than African ancestry, as previously described<sup>18</sup>. Using a large reference panel of human haplotypes from the Haplotype Reference Consortium (HRC r1.1 2016)<sup>66</sup> in Michigan Imputation Server<sup>66,67</sup> with ShapeIT (v2.r790)<sup>68</sup> as the phasing tool, we imputed untyped SNPs genome-wide. SNPs were excluded if (1) imputation quality metric  $R^2 < 0.3$  (indicating inadequate accuracy of the imputed genotype); (2) minor allele frequency in cases and controls < 0.01; (3) HWE  $P < 1 \times 10^{-5}$  in cases and controls classified as European American. Using Q–Q plots inflation test statistics, we estimated an inflation factor  $\lambda$  of 1.09 in the replication series.

The discovery GWAS  $P$ -value was thresholded at  $5 \times 10^{-8}$  for genome-wide significance and replication  $P$ -value was thresholded at 0.05 for validation. For all four variants validated in the replication analysis, we estimated a false discovery rate < 5% with nominal  $P$ -value < 0.05, using Benjamini–Hochberg procedure.

We performed the same statistical analyses for all the datasets unless specifically stated.

**Summary Mendelian randomisation analysis.** SMR analysis was conducted as per Zhu et al.<sup>69</sup>. The most significant eQTL or mQTL for each gene was used as an

instrumental variable to test for an association between expression levels of the gene and B-ALL using summary statistics from the meta-analysis GWAS dataset. The expression levels of the gene identified should be significantly associated with the disease as a result of true pleiotropy as opposed to correlation due to linkage between the GWAS variants and functional eQTL variants; accordingly, the heterogeneity in dependent instruments (HEIDI) analysis was performed as per Zhu et al.<sup>69</sup> Publicly available eQTL data were extracted from the CAGE eQTL dataset (peripheral blood,  $n = 2765$ )<sup>31</sup>, GTEx eQTL v7, whole blood ( $n = 369$ ) and Epstein–Barr Virus-transformed lymphocytes ( $n = 117$ ), and blood eQTL datasets<sup>29,70,71</sup>. To investigate regulatory elements associated with B-ALL, we utilised the methylation QTL datasets Aberdeen (Blood,  $n = 639$ ) and UCL (Blood,  $n = 665$ )<sup>72</sup>. All eQTL or mQTL summary datasets were pruned to only those probes with  $P_{eQTL/mQTL} < 5 \times 10^{-8}$ . GWAS summary statistics files were generated from the meta-analysis of UK GWAS I, UK GWAS II, German GWAS and COG\_SJ datasets. Reference files were generated by merging 1000 genomes phase 3 and UK10K (ALSPAC and TwinsUK) data. Summary eQTL files for the GTEx samples were generated from downloaded v7 'all\_SNPgene\_pairs' files. Only probes with eQTL  $P < 5.0 \times 10^{-8}$  were considered in the SMR analysis. HEIDI test  $P$ -values < 0.05 were taken to indicate significant heterogeneity.

**Association test of predicted gene expression with ALL risk.** Associations between predicted gene expression and ALL risk were examined using MetaXcan, accounting for LD<sup>32</sup>. SNP weights and their respective covariance for all GTEx tissues were obtained from predict.db (<http://predictdb.org/>), which is based on GTEx version 7 eQTL data. To combine S-PrediXcan data across the different tissues taking into account tissue–tissue correlations, we used S-MultiXcan. To determine whether associations between genetically predicted gene expression and ALL risk were influenced by variants previously identified by GWAS, we performed conditional analyses adjusting for GWAS risk SNPs (Supplementary Table 16) predicted by GCTA-COJO stepwise logistic regression analysis<sup>73,74</sup>. Adjusted output files were provided as the input GWAS summary statistics for S-PrediXcan analyses as above.

**Functional-epigenetic annotation.** Promoter ChIC, chromatin-state annotation and TF analyses were performed on lead SNPs, defined as any SNP with a  $P$ -value <  $P(\min) \times 50$  and  $R^2 > 0.8$  from the lead SNP at a locus.

**eQTL data.** SNP gene expression associations were extracted from the Blood<sup>29</sup>, CAGE<sup>31</sup>, and MuTHER<sup>70</sup> eQTL datasets. Only associations from the Blood dataset are shown in Fig. 2.

**ATAC-seq.** Chromatin accessibility in the lymphoblastoid cell line GM12878 was extracted from GSE47753<sup>75</sup>.

**Chromatin-state annotation.** Chromatin-state segregation data, analysed by ChromHMM, as shown for the primary B-cell lines E031 and E032, and the lymphoblastoid cell line GM12878 from the roadmap<sup>76</sup> and Encode<sup>24</sup> projects, respectively.

**Promoter capture Hi-C.** Promoter-looping interactions were downloaded and filtered for a  $-\log(\text{weighted } P) \geq 5$  in naive B cells only<sup>26</sup>. Interactions were called using CHICAGO<sup>77</sup>. Interactions overlapping lead SNPs in each locus are reported.

**Hi-C and histone mark ChIP-seq in ALL cells.** Hi-C and H3K27Ac ChIP-seq were performed in human ALL cell line Nalm6 at St. Jude<sup>27</sup>. For Hi-C, the Nalm6 cell line was cultured under recommended conditions to about 80% confluence. Five million cells were crosslinked with 1% formaldehyde for 10 min at room temperature, then digested with 125 units of MboI and labelled by biotinylated nucleotides and were proximity ligated. After reverse crosslinking, ligated DNA was purified and sheared to 300–500 bp, then ligation junctions were pulled down with streptavidin beads and prepared as a general Illumina library. The Hi-C sample was sequenced paired-end 76 cycles on Illumina HiSeq 4000. For the H3K27Ac ChIP-seq, a frozen cell pellet containing 10 million cells was sent to Active Motif for ChIP and library preparation. The sample was divided into an aliquot for ChIP using an antibody to H3K27ac (Active Motif) and an input control. Single-end sequencing was performed using an Illumina NextSeq 500 generating 76 cycles for each sequencing read. Histone acetylation mark and chromatin looping signals were directly downloaded from the NCBI GEO GSE115494 dataset. Loop interactions were called using HiCCUPS<sup>78</sup> from Juicer tools v1.12.01 under default parameters at a resolution of 5 kb and 10 kb. Enriched interaction was reported with a false discovery rate < 0.1.

**TF-enrichment analysis.** TF-binding enrichment analysis was performed according to the method of Cowper-Salari et al.<sup>33</sup> examining SNPs in LD with the sentinel SNP (i.e.,  $r^2 > 0.8$  and  $D' > 0.8$ ). Publicly available TF ChIP-seq data were obtained from ChIP-Atlas (<http://chip-atlas.org/>). TF-binding sites were filtered for those with a MACS peak  $Q$ -value > 100 and from cells lines with a 'blood'

annotation. Overlapping binding sites from the same ChIP target were merged. For each mark, the overlap of the SNPs and the binding sites was assessed to generate a mapping tally. A null distribution was produced by performing 10,000 permutations, randomly selecting LD blocks with the same number of SNPs as the test set, and calculating the null mapping tally. *P*-values were calculated by normalising the tallies to the median of the null distribution.

**Heritability analysis.** We used LDAK version 4.9<sup>35</sup> to estimate the polygenic variance (i.e., heritability) ascribable to all genotyped and imputed GWAS SNPs. Heritability ascribed to all the genotyped and imputed SNPs was calculated from summary data after filtering; information score filtering (>0.99), allele frequency (>0.01) and Hardy-Weinberg deviation ( $P < 1 \times 10^{-5}$ ), resulting in 1,553,634 SNPs for analyses. SNP-specific weightings were calculated reflecting correlations across SNPs (predictors) using UK10K and 1000 genomes data, after adjusting for LD, MAF and genotype certainty.

**Contribution of genetic variance to familial risk.** Estimation of risk variance associated with each SNP was performed as per Pharoah et al.<sup>79</sup> For an allele (*i*) of frequency *p*, relative risk *R* and log risk *r*, the risk distribution variance (*V<sub>i</sub>*) is:

$$V_i = (1 - p)^2 E^2 + 2p(1 - p)(r - E)^2 + p^2(2r - E)^2 \quad (1)$$

where *E* is the expected value of *r* given by:

$$E = 2p(1 - p)r + 2p^2r \quad (2)$$

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

$$V = \sum V_i \quad (3)$$

The percentage of total variance was calculated assuming a familial risk of childhood ALL of 3.2 (95% confidence interval (CI) 1.5–5.9) as per Kharazmi et al.<sup>80</sup>. All genetic variance (*V*) associated with susceptibility alleles is given as  $\sqrt{3.2}$ <sup>80</sup>. The proportion of genetic risk attributable to a single allele is:

$$V_i V^{-1}$$

Eighteen 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 11 variants are thought to act independently, as previous studies have shown no interaction between risk loci<sup>7,9–11</sup>. PRS were generated as per Pharoah et al.<sup>79</sup> 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.

**Relationship between SNP genotype and ALL clinical features.** The relationship between SNP genotype and survival was analysed in the German AIEOP-BFM series, which consisted of 834 patients within the AIEOP-BFM 2000 trial. Patients were treated with conventional chemotherapy (i.e., prednisone, vincristine, daunorubicin, l-asparaginase, cyclophosphamide, ifosfamide, cytarabine, 6-mercaptopurine, 6-thioguanine and methotrexate), a subset of those with high-risk ALL were treated with cranial irradiation and/or stem cell transplantation. Events, for event-free survival, 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 Prentice, and compared using Gray's test. Cox regression analysis was used to estimate hazard ratios and 95% CIs adjusting for clinically relevant covariates. Similar analyses of SNP genotype with treatment response and outcome measures were performed in the COG\_SJ series as reported previously<sup>20,34</sup>. No significant association was observed for these novel risk SNPs (i.e.,  $P > 0.05$ )<sup>20,34</sup>.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

UK controls were obtained from the Wellcome Trust Case Control Consortium 2 (<http://www.wtccc.org.uk/>; 50.7% male.<sup>81</sup> WTCCC2:EGAD00000000022 and EGAD00000000024). Imputation reference panels are available from 1000 G phase I (<ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/>) and the UK10K ( $n = 3781$ ; EGAS00001000090, EGAD00001000195 and EGAS00001000108; [www.uk10k.org](http://www.uk10k.org)). The UK GWAS I, UK GWAS II and German GWAS data for ALL cases are available through the European Genome-Phenome Archive website (EGA, <https://ega-archive.org>, EGAS00001003937, EGAS00001002809 and EGAS00001003936, respectively). The SJ\_COG GWAS data for ALL cases are deposited in the NIH dbGAP (<https://www.ncbi.nlm.nih.gov/gap/>) under phs000638.v1.p1 and phs000637.v1.p1. ATAC-seq dataset GSE47753\_GM12878\_ATACseq\_50k\_AllReps\_ZINBA\_pp08.bed.gz was downloaded from Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE47753>). ChromHMM data for primary B-cell are available at <http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/coreMarks/>

[jointModel/final/](http://jointModel/final/) and ChromHMM annotation for GM12878 is available from <http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHmm/wgEncodeBroadHmmGm12878HMM.bed.gz>. Promoter ChIC data are available at <https://osf.io/u8tzp/>. eQTL and mQTL data for SMR analysis were downloaded from <https://cns.genomics.com>. GTEx version 7 data are available at <https://gtexportal.org/home/datasets>. Requests for other data should be directed to the authors.

Received: 22 April 2019; Accepted: 17 October 2019;

Published online: 25 November 2019

## References

1. Stiller, C. *Childhood Cancer in Britain: Incidence, Survival, Mortality*. (Oxford Univ. Press, Oxford, 2007).
2. Greaves, M. F. & Wiemels, J. Origins of chromosome translocations in childhood leukaemia. *Nat. Rev. Cancer* **3**, 639–649 (2003).
3. Gruhn, B. et al. Prenatal origin of childhood acute lymphoblastic leukemia, association with birth weight and hyperdiploidy. *Leukemia* **22**, 1692–1697 (2008).
4. Perera, F. P. Environment and cancer: who are susceptible? *Science* **278**, 1068–1073 (1997).
5. Stiller, C. A. & Parkin, D. M. Geographic and ethnic variations in the incidence of childhood cancer. *Br. Med. Bull.* **52**, 682–703 (1996).
6. Williams, L. A., Yang, J. J., Hirsch, B. A., Marcotte, E. L. & Spector, L. G. Is there etiologic heterogeneity between subtypes of childhood acute lymphoblastic leukemia? A review of variation in risk by subtype. *Cancer Epidemiol. Biomarkers Prev.* **28**, cebp.0801.2018 (2019).
7. Papaemmanuil, E. et al. Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. *Nat. Genet.* **41**, 1006–1010 (2009).
8. Sherborne, A. L. et al. Variation in CDKN2A at 9p21.3 influences childhood acute lymphoblastic leukemia risk. *Nat. Genet.* **42**, 492–494 (2010).
9. Migliorini, G. et al. Variation at 10p12.2 and 10p14 influences risk of childhood B-cell acute lymphoblastic leukemia and phenotype. *Blood* **122**, 3298–3307 (2013).
10. Vijaykrishnan, J. et al. The 9p21.3 risk of childhood acute lymphoblastic leukaemia is explained by a rare high-impact variant in CDKN2A. *Sci. Rep.* **5**, 15065 (2015).
11. Vijaykrishnan, J. et al. A genome-wide association study identifies risk loci for childhood acute lymphoblastic leukemia at 10q26.13 and 12q23.1. *Leukemia* **31**, 573–579 (2017).
12. Vijaykrishnan, J. et al. Genome-wide association study identifies susceptibility loci for B-cell childhood acute lymphoblastic leukemia. *Nat. Commun.* **9**, 1340 (2018).
13. Trevino, L. R. et al. Germline genomic variants associated with childhood acute lymphoblastic leukemia. *Nat. Genet.* **41**, 1001–1005 (2009).
14. Qian, M. et al. Novel susceptibility variants at the ERG locus for childhood acute lymphoblastic leukemia in Hispanics. *Blood* **133**, 724–729 (2019).
15. Qian, M. et al. Genome-wide association study of susceptibility loci for T-cell acute lymphoblastic leukemia in children. *J. Natl Cancer Inst.* djz043 (2019).
16. Xu, H. et al. Novel susceptibility variants at 10p12.31–12.2 for childhood acute lymphoblastic leukemia in ethnically diverse populations. *J. Natl Cancer Inst.* **105**, 733–742 (2013).
17. Mullighan, C. G. et al. Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* **446**, 758–764 (2007).
18. Perez-Andreu, V. et al. Inherited GATA3 variants are associated with Ph-like childhood acute lymphoblastic leukemia and risk of relapse. *Nat. Genet.* **45**, 1494–1498 (2013).
19. Perez-Andreu, V. et al. A genome-wide association study of susceptibility to acute lymphoblastic leukemia in adolescents and young adults. *Blood* **125**, 680–686 (2015).
20. Yang, J. J. et al. Genome-wide association study identifies germline polymorphisms associated with relapse of childhood acute lymphoblastic leukemia. *Blood* **120**, 4197–4204 (2012).
21. Liu, J. Z. et al. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat. Genet.* **42**, 436–440 (2010).
22. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).
23. Bernstein, B. E. et al. The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.* **28**, 1045–1048 (2010).
24. Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
25. Consortium, E. P. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* **9**, e1001046 (2011).



26. Javierre, B. M. et al. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell* **167**, 1369–1384 e19 (2016).
27. Tian, L. et al. Long-read sequencing unveils IGH-DUX4 translocation into the silenced IGH allele in B-cell acute lymphoblastic leukemia. *Nat. Commun.* **10**, 2789 (2019).
28. Consortium, G. T. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
29. Westra, H. J. et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
30. Grundberg, E. et al. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat. Genet.* **44**, 1084–1089 (2012).
31. Lloyd-Jones, L. R. et al. The genetic architecture of gene expression in peripheral blood. *Am. J. Hum. Genet.* **100**, 228–237 (2017).
32. Barbeira, A. N. et al. Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS Genet.* **15**, e1007889 (2019).
33. Cowper-Salari, R. et al. Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat. Genet.* **44**, 1191–1198 (2012).
34. Karol, S. E. et al. Genetics of ancestry-specific risk for relapse in acute lymphoblastic leukemia. *Leukemia* **31**, 1325–1332 (2017).
35. Speed, D. et al. Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* **49**, 986–992 (2017).
36. Frampton, M. J. et al. Implications of polygenic risk for personalised colorectal cancer screening. *Ann. Oncol.* **27**, 429–434 (2016).
37. Takeuchi, O. et al. Essential role of BAX, BAK in B cell homeostasis and prevention of autoimmune disease. *Proc. Natl Acad. Sci. USA* **102**, 11272–11277 (2005).
38. Chittenden, T. et al. Induction of apoptosis by the Bcl-2 homologue Bak. *Nature* **374**, 733–736 (1995).
39. Leu, J. I. & George, D. L. Hepatic IGFBP1 is a prosurvival factor that binds to BAK, protects the liver from apoptosis, and antagonizes the proapoptotic actions of p53 at mitochondria. *Genes Dev.* **21**, 3095–3109 (2007).
40. Chen, J. et al. miR-125b inhibitor enhance the chemosensitivity of glioblastoma stem cells to temozolomide by targeting Bak1. *Tumour Biol.* **35**, 6293–6302 (2014).
41. Andersson, A. et al. Molecular signatures in childhood acute leukemia and their correlations to expression patterns in normal hematopoietic subpopulations. *Proc. Natl Acad. Sci. USA* **102**, 19069–19074 (2005).
42. Stokus, M., Vaitkeviciene, G., Eidukaite, A. & Griskevicius, L. ETV6/RUNX1 transcript is a target of RNA-binding protein IGF2BP1 in t(12;21)(p13;q22)-positive acute lymphoblastic leukemia. *Blood Cells Mol. Dis.* **57**, 30–34 (2016).
43. Huang, X. et al. Insulin-like growth factor 2 mRNA-binding protein 1 (IGF2BP1) in cancer. *J. Hematol. Oncol.* **11**, 88 (2018).
44. Qian, M. et al. Novel susceptibility variants at the ERG locus for childhood acute lymphoblastic leukemia in Hispanics. *Blood* **133**, 724–729 (2018).
45. Zhang, J. et al. Deregulation of DUX4 and ERG in acute lymphoblastic leukemia. *Nat. Genet.* **48**, 1481–1489 (2016).
46. Willman, C. L. et al. Deletion of IRF-1, mapping to chromosome 5q31.1, in human leukemia and preleukemic myelodysplasia. *Science* **259**, 968–971 (1993).
47. Boulton, J. et al. Allelic loss of IRF1 in myelodysplasia and acute myeloid leukemia: retention of IRF1 on the 5q- chromosome in some patients with the 5q- syndrome. *Blood* **82**, 2611–2616 (1993).
48. Irving, J. A. et al. Integration of genetic and clinical risk factors improves prognostication in relapsed childhood B-cell precursor acute lymphoblastic leukemia. *Blood* **128**, 911–922 (2016).
49. Roe, J. S., Mercan, F., Rivera, K., Pappin, D. J. & Vakoc, C. R. BET bromodomain inhibition suppresses the function of hematopoietic transcription factors in acute myeloid leukemia. *Mol. Cell* **58**, 1028–1039 (2015).
50. Dawson, M. A. et al. Inhibition of BET recruitment to chromatin as an effective treatment for MLL-fusion leukaemia. *Nature* **478**, 529–533 (2011).
51. Zuber, J. et al. RNAi screen identifies Brd4 as a therapeutic target in acute myeloid leukaemia. *Nature* **478**, 524–528 (2011).
52. Coutinho, A. E. & Chapman, K. E. The anti-inflammatory and immunosuppressive effects of glucocorticoids, recent developments and mechanistic insights. *Mol. Cell Endocrinol.* **335**, 2–13 (2011).
53. Kruth, K. A. et al. Suppression of B-cell development genes is key to glucocorticoid efficacy in treatment of acute lymphoblastic leukemia. *Blood* **129**, 3000–3008 (2017).
54. Larsen, E. C. et al. Dexamethasone and high-dose methotrexate improve outcome for children and young adults with high-risk B-acute lymphoblastic leukemia: a report from Children's Oncology Group Study AALL0232. *J. Clin. Oncol.* **34**, 2380–2388 (2016).
55. Borowitz, M. J. et al. Clinical significance of minimal residual disease in childhood acute lymphoblastic leukemia and its relationship to other prognostic factors: a Children's Oncology Group study. *Blood* **111**, 5477–5485 (2008).
56. Pui, C. H. et al. Improved outcome for children with acute lymphoblastic leukemia: results of Total Therapy Study XIII B at St Jude Children's Research Hospital. *Blood* **104**, 2690–2696 (2004).
57. Pui, C. H. et al. Treating childhood acute lymphoblastic leukemia without cranial irradiation. *N. Engl. J. Med.* **360**, 2730–2741 (2009).
58. Team, R. D. C. R. *A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, Vienna, Austria, 2008).
59. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
60. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).
61. Anderson, C. A. et al. Data quality control in genetic case-control association studies. *Nat. Protoc.* **5**, 1564–1573 (2010).
62. Consortium, U. K. et al. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
63. Genomes Project, C. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
64. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
65. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
66. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
67. Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
68. Delaneau, O., Marchini, J. & Zagury, J. F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2011).
69. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
70. Nica, A. C. et al. The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet.* **7**, e1002003 (2011).
71. Consortium, G. T. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
72. Hannon, E., Weedon, M., Bray, N., O'Donovan, M. & Mill, J. Pleiotropic effects of trait-associated genetic variation on DNA methylation: utility for refining GWAS loci. *Am. J. Hum. Genet.* **100**, 954–959 (2017).
73. Zhu, Z. et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat. Commun.* **9**, 224 (2018).
74. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. Genome-wide complex trait analysis (GCTA): methods, data analyses, and interpretations. *Methods Mol. Biol.* **1019**, 215–236 (2013).
75. Buenostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
76. Roadmap Epigenomics, C. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
77. Cairns, J. et al. CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biol.* **17**, 127 (2016).
78. Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
79. Pharoah, P. D., Antoniou, A. C., Easton, D. F. & Ponder, B. A. Polygenes, risk prediction, and targeted prevention of breast cancer. *N. Engl. J. Med.* **358**, 2796–2803 (2008).
80. Kharazmi, E. et al. Familial risks for childhood acute lymphocytic leukaemia in Sweden and Finland: far exceeding the effects of known germline variants. *Br. J. Haematol.* **159**, 585–588 (2012).
81. 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).

## Acknowledgements

In the UK, funding was provided by Bloodwise and Cancer Research UK (C1298/A8362). In the United States, this work was partly supported by National Institutes of Health Grant Numbers CA21765, CA98543, CA114766, CA98413, CA180886, CA180899, GM92666, GM115279, and GM097119, and the American Lebanese Syrian Associated Charities. We thank the patients and parents who participated in the Children's Oncology Group (COG) protocols included in this study, the clinicians and research staff at St. Jude Children's Research Hospital and COG institutions, Jeanette Pullen (University of Mississippi, Jackson, MS) for assistance in the classification of patients with ALL and Mark Shriver (Pennsylvania State University, University Park, PA) for sharing single-nucleotide polymorphism genotype data of the Native American references. M.Q. is supported by the Initial Funding for New PI of Fudan University, the National Natural Science Foundation of China (81973997) and the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning. S.P.H. is the Jeffrey E. Perelman Distinguished Chair in Pediatrics at The Children's Hospital of



Philadelphia. M.L.L. is the University of California, San Francisco Benioff Chair of Children's Health and the Deborah and Arthur Ablin Chair of Pediatric Molecular Oncology.

### Author contributions

J.J.Y. and R.S.H. designed the overall study. Association analysis and statistical data analysis were performed by J.V. and M.Q. Functional analysis was undertaken by J.B.S., J.V., W.Y. and M.Q. W.Y., B.K. and P.J.L. provided bioinformatics support. P.B. supervised the data production of UK GWAS II. J.A., A.V. and A.M. provided samples recruited on the ALL-97/99 and ALL-2003 trials. C.R.B., M. Stanulla, M. Schrappe and M.Z. provided samples through the Berlin–Frankfurt–Münster (BFM) trial (1993–2004); M. Stanulla and M.Z. conducted outcome analysis on BFM samples. E.A.R., C.-H.P., W.E.E., C.G.M., S.P.H., M.V.R. and M.L.L. supervised the sample collection and data production of the COG\_SJ cohort. C.-H.P., W.E.E., A.Y., C.L., S.P.H., M.V.R., M.L.L., R.S.H. and J.J.Y. interpreted the data and the research findings. The manuscript was drafted by J.J.Y., R.S.H., J.V., M.Q. and J.B.S., and was reviewed by all of the co-authors.

### Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41467-019-13069-6>.

**Correspondence** and requests for materials should be addressed to R.S.H. or J.J.Y.

**Peer review information** *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

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



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019