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

Jayaram Vijayakrishnan, Maoxiang Qian, James B. Studd, Wenjian Yang, Ben Kinnersley, Philip J. Law, Peter Broderick, Elizabeth A. Raetz, James Allan, Ching-Hon Pui, Ajay Vora, William E. Evans, Anthony Moorman, Allen Yeoh, Wentao Yang, Chunliang Li, Claus R. Bartram, Charles G. Mullighan, Martin Zimmermann, Stephen P. Hunger, Martin Schrappe, Mary V. Relling, Martin Stanulla, Mignon L. Loh, Richard S. Houlston, Jun J. Yang

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-RUNXI 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-RUNXI 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.
Acute lymphoblastic leukaemia (ALL) is the most common paediatric cancer with B-cell precursor ALL (B-ALL) accounting for ~85% of the cases\(^1\). 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\(^1,3\). Both the absence of specific environmental risk factors and early onset suggest a strong inherited genetic basis for susceptibility\(^4-6\). 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 (PIPK4K2A), 10q26.13 (LHPP), 12q23.1 (ELK3), 10p14 (GATA3), 10q21.2 (ARID5B), 14q11.2 (CEBPE), 16p13.3 (USP7) and 21q22.2 (ERG)\(^7-16\). 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\(^17\), 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\(^7,9,12,18,19\).

To gain a more comprehensive insight into susceptibility to ALL, we performed a meta-analysis of four GWAS from the North America\(^3,18,20\) and Europe\(^7,15\), 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\(^7,9,12,13,18,20\), 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\(^31,22\). Quantile–quantile (Q–Q) plots for SNPs did not show evidence of substantive over dispersion ($\lambda_{CV}$ 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\(^9,12,18\), 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, BAKI), 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\(^13\), 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 (PIPK4K2A) (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\(^23\), and the GM12878 lymphoblastoid cell line from ENCODE\(^24,25\) (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 $r^2 > 0.8$, $P < \text{max} \times 50$; 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 (Hi-C) data from naive B cells\(^26\) (Supplementary Table 9) as well as Hi-C and H3K27Ac ChIP data in human ALL cells\(^27\) (Supplementary Fig. 8). We also sought to identify target genes by performing quantitative trait locus (QTL) analysis of mRNA expression (eQTL) data from GTEx\(^28\), Blood eQTL\(^29\), MuTHER\(^30\) and CAGE\(^31\) databases, and DNA methylation (mQTL) (Supplementary Table 10). We annotated risk loci with variants mapping to haematopoietic transcription factor (TF)-binding sites (Fig. 2) and Supplementary Figures 3, 4, 5, 6, 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...
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...
performed a TWAS integrating genomic and expression data. 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-Sal-lari et al. 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 (Benjamin–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_SI GWAS cohorts. A failure to demonstrate additional relationships may, however, be reflective of limited statistical power.

**Contribution of risk SNPs to heritability.** Using LD-adjusted kinship (LDAK), 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 FQFT. 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. Our recently discovered T-ALL risk locus USP7 was also not significant in this GWAS because of its lineage-specific effect on ALL susceptibility. BAK1 is essential for B-cell homoeostasis and its knockout mice accumulate immature and mature follicular B cells. BAK1 induces apoptosis by binding to and antagonising anti-apoptotic proteins, including BCL2. Reduced BAK1 expression relieves repression of BCL2, inhibiting apoptosis and conferring a pro-survival advantage. 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 leukemogenesis. Notably, the 6p21 region is also pleiotropic, influencing chronic lymphocytic leukaemia (CLL) and testicular cancer risk. Moreover, the strongest association for both CLL and testicular cancer risk. Finally, rs210143-T, to the BAK1 promoter and a strong negative association with expression suggests decreased BAK1 promotes ALL leukemogenesis. Notably, the 6p21 region is also pleiotropic, influencing chronic lymphocytic leukaemia (CLL) and testicular cancer risk. Moreover, the strongest association for both CLL and testicular cancer risk.

| CHR | SNP (Subtype) | Locus (gene) | Position (BP) | Risk allele | RAF | OR (95% CI) | P-value |
|-----|---------------|--------------|---------------|-------------|-----|------------|---------|
| 2   | rs17481869(ETV6-RUNXI) | 2q22.3       | 146124454     | A           | 0.08| 1.74       | 2.37 × 10⁻⁹ |
| 5   | rs8626825 (High-Hyperdiploidy) | 5q31.1 (C5orf56) | 31765206     | T           | 0.34| 1.29       | 1.56 × 10⁻⁸ |
| 6   | rs210143 (High-Hyperdiploidy) | 6p21.31 (B AKI) | 33546930     | C           | 0.73| 1.30       | 2.21 × 10⁻⁸ |
| 7   | rs17133805 | 7p12.2 (IKZF1) | 50477514     | G           | 0.32| 1.65       | 5.28 × 10⁻⁷ |
| 8   | rs75777619 | 8q24.21      | 130185176     | A           | 0.12| 1.26       | 2.30 × 10⁻⁹ |
| 9   | rs76925697 | 9q21.31      | 83747371      | A           | 0.96| 1.52       | 2.11 × 10⁻⁸ |
| 10  | rs113650570 | 10p12.2 (GATA3) | 21976402 | A | 0.02| 2.32       | 8.06 × 10⁻⁸ |
| 12  | rs2296624 | 12p12.2 (PIPIK2A) | 22856946 | C | 0.67| 1.25       | 2.79 × 10⁻⁷ |
| 12  | rs1779301 | 12q26.13 (LLHP) | 126292655    | C | 0.66| 1.22       | 5.72 × 10⁻³ |
| 14  | rs4762284 | 14q23.1 (ELK3) | 96612762      | C | 0.32| 1.15       | 3.75 × 10⁻⁷ |
| 14  | rs2239630 | 14q11.2 (CEBP) | 23589349     | A | 0.45| 1.28       | 1.72 × 10⁻¹ |
| 17  | rs10053104 (ETV6-RUNXI) | 17q21.32 (IGF2BP1) | 47092076 | T | 0.47| 1.33       | 1.82 × 10⁻⁸ |
| 21  | rs9976326 (High-Hyperdiploidy) | 21q22.2 (ERG) | 39776485 | T | 0.25| 1.33       | 4.79 × 10⁻⁷ |

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

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. 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 NR3CI whose alterations are associated with poor outcome and high risk in ALL patients. 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, FLII, ERG, MYB and CEBPa/β, and this may account for its enrichment at risk loci. Several groups have shown activity of BET inhibitors in AML cells lines. NR3CI 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 and induces cell death in ALL cells by lowering the expression of B-cell survival factors. Further validation will be required to establish a role of disrupted NR3CI 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.
GWAS data. The four GWAS datasets have been the subject of previous publications: (i) UK GWAS I—284 cases, 2699 controls from the 1958 British Birth Cohort (UKBB); (ii) 2501 controls from the UK Blood Service controls; (iii) 15 cases and 2501 controls from the UK Blood Service controls; (iv) UK GWAS—1155 Berlin–Frankfurt–Munster (BFM) trial (1993–2004 cases), 2132 Heinz Nixdorf Recall study controls; (v) UK GWAS II—1021 cases from Medical Research Council UK ALL 2003 and ALL-97/99 trials, 2976 PRACTICAL Consortium and 4464 Breast Cancer Association Consortium controls; (vi) COG_SJ GWAS, with the sum of these four ancestries being 100% for any given subject (P9904/P9905/P9906). St. Jude Therapy Trial XIIIB/XV and 2057 non-ALL controls of European ancestry from the Multi-Ethnic Study of Atherosclerosis (MESA) study (dbGAP phs000209.v9). The replication study included 2237 cases and 3461 non-ALL controls of 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 XIIIB/XV and COG_P9906).

Statistical analysis of GWAS data. Analyses were undertaken using R v3.2.3,68, PLINK v1.09,69 SNPTEST v2.5.222 and IMPUTE v2.360 software. Standard quality-control measures were applied to each GWAS.61 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-analyses. SNPs with call rate < 95% were excluded or shown deviation from Hardy–Weinberg equilibrium ($P < 10^{-5}$). Appropriateness case–control matching was evaluated using Q–Q plots inflation test statistics. The inflation factor λ was calculated to indicate the degree of genomic inflation, by dividing the median of the test value distribution with 1 degree of freedom (Supplementary Fig. 2). Prediction of the untyped genotypes was carried out using 1000 Genomes Project (Phase I) and UK10K as reference.62,63 To account for genomic inflation post imputation, top Eigenvectors from the principal component analysis were used as covariates in the final association analysis.64 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 v1.23. 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 I² 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 using ADMIXTURE (version 1.30).46 We calculated Chorom’s Q score to test for heterogeneity and the I² statistic to test for heterogeneity.

Chromatin-state annotation. 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 XIIIB/XV and COG_P9906).

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 v1.23. Association meta-analyses only included SNPs with info score > 0.8, imputed call rates > 0.9 and MAFs > 0.01. We calculated Chorom’s Q score to test for heterogeneity and the I² 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 using ADMIXTURE (version 1.30).46 We calculated Chorom’s Q score to test for heterogeneity and the I² statistic to test for heterogeneity.

Statistical analysis of GWAS data. Analyses were undertaken using R v3.2.3,68, PLINK v1.09,69 SNPTEST v2.5.222 and IMPUTE v2.360 software. Standard quality-control measures were applied to each GWAS.61 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-analyses. SNPs with call rate < 95% were excluded or shown deviation from Hardy–Weinberg equilibrium ($P < 10^{-5}$). Appropriateness case–control matching was evaluated using Q–Q plots inflation test statistics. The inflation factor λ was calculated to indicate the degree of genomic inflation, by dividing the median of the test value distribution with 1 degree of freedom (Supplementary Fig. 2). Prediction of the untyped genotypes was carried out using 1000 Genomes Project (Phase I) and UK10K as reference.62,63 To account for genomic inflation post imputation, top Eigenvectors from the principal component analysis were used as covariates in the final association analysis.64 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 v1.23. Association meta-analyses only included SNPs with info score > 0.8, imputed call rates > 0.9 and MAFs > 0.01. We calculated Chorom’s Q score to test for heterogeneity and the I² 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 using ADMIXTURE (version 1.30).46 We calculated Chorom’s Q score to test for heterogeneity and the I² statistic to test for heterogeneity.

Chromatin-state annotation. 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 XIIIB/XV and COG_P9906).
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 LDuk version 4.935 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 × 10−7), 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.39. For an allele (i) of frequency p, relative risk R and log risk r, the risk distribution variance (V) is:

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

where E is the expected value of r given by:

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

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

\[ V = \sum V_i \]

Eighteen risk loci were included in the calculation of thePRS 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.63–66,91 PRS were generated as per Pharoah et al.79 assuming a log-normal distribution LN(μ,σ²) with mean μ and variance σ². The population μ was set to Φ2/(2π), 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 survival 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.20,34 No significant association was observed for these novel risk loci (i.e., P > 0.05).20,34

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

Data availability

UK10K datasets were obtained from the Wellcome Trust Case Control Consortium 2 (http://www.wtcc.org.uk/uk/). Phase 3 of the 1000 Genomes Project (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20150223/) and the UK10K (n = 3781; EGAS0000100079; https://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, EGAS0000100097; EGAS0000100289 and EGAS0000100396, respectively). The SJ_COG GWAS data for ALL cases are deposited in the NHG 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/ChromHMM/Segments/ChmmModels/coreMarks/

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 leukaemia, 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 across childhood acute lymphoblastic leukaemia? A review of variation in risk by subtype. Cancer Epidemiol. Biomarkers Prev. 28, cebp0801 (2019).
7. Papaemmanuil, E. et al. Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukaemia. Nat. Genet. 41, 1005–1010 (2009).
8. Shrober, A. L. et al. Variation in CDKN2A at 9p21.3 influences childhood acute lymphoblastic leukaemia 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 leukaemia and phenotype. Blood 122, 3299–3307 (2013).
10. Vijayakrishnan, 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. Vijayakrishnan, J. et al. A genome-wide association study identifies risk loci for childhood acute lymphoblastic leukaemia at 10q26.13 and 12q23.1. Leukemia 31, 573–579 (2017).
12. Vijayakrishnan, J. et al. Genome-wide association study identifies susceptibility loci for B-cell childhood acute lymphoblastic leukaemia. Nat. Commun. 9, 1340 (2018).
13. Trevino, L. R. et al. Germline genomic variants associated with childhood acute lymphoblastic leukaemia. Nat. Genet. 41, 1001–1005 (2009).
14. Qian, M. et al. Novel susceptibility variants at the ERG locus for childhood acute lymphoblastic leukaemia in Hispanics. Blood 133, 724–729 (2019).
15. Qian, M. et al. Genome-wide association study of susceptibility loci for T-cell acute lymphoblastic leukaemia in children. J. Natl Cancer Inst. djw043 (2019).
16. Xu, H. et al. Novel susceptibility variants at 10p12.31-12.2 for childhood acute lymphoblastic leukaemia 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 leukaemia 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 leukaemia 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 leukaemia. Blood 120, 4197–4202 (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 tissues in twins. Nat. Genet. 44, 1191–1198 (2012).
32. Karol, S. E. et al. Genetics of ancestry-specific risk for relapse in acute lymphoblastic leukemia. Leukemia 31, 1325–1332 (2017).
33. Speed, D. et al. Reevaluation of SNP heritability in complex human traits. Nat. Genet. 49, 986–992 (2017).
34. Prieto, A. et al. Identification of polygenic risk associated with inherited colorectal cancer screening. Ann. Oncol. 27, 429–434 (2016).
35. Takeuchi, O. et al. Essential role of BAK, BAX in B cell homeostasis and prevention of autoimmune disease. Proc. Natl. Acad. Sci. USA 102, 11272–11277 (2005).
36. Chittenden, T. et al. Induction of apoptosis by the Bcl-2 homologue BAX. Nature 377, 736–739 (1995).
37. Leu, J. I. & George, D. L. Hepatic IGFBP1 is a prosurvival factor that binds to positive acute lymphoblastic leukemia. Nat. Genet. 48, 1279–1382 (2016).
38. Irving, J. A. et al. Integration of genetic and clinical risk factors improves association detection. PLoS Genet. 15, e1007889 (2019).
39. Qian, M. et al. Novel susceptibility variants at the ERG locus for childhood acute lymphoblastic leukemia: a report from Children’s Oncology Group Study AALL0232. J. Clin. Oncol. 34, 2380–2388 (2016).
40. Chen, J. et al. miR-125b inhibitor enhance the chemosensitivity of their correlations to expression patterns in normal hematopoietic subpopulations. Proc. Natl. Acad. Sci. USA 102, 19969–19007 (2005).
41. Anders, A. et al. Moleculardesign of childhood acute leukemia and its curation to expression patterns in normal hematopoietic subpopulations. Proc. Natl. Acad. Sci. USA 102, 19969–19007 (2005).
42. Dey, S. et al. A linear complexity phasing method for whole-genome haplotyping. Nat. Genet. 34, 1285–1288 (2016).
43. Huang, X. et al. Haplotype blocks (1046) 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. Blood 133, 724–729 (2018).
45. Jiang, J. et al. Detection of DUX4-IGH translocations in acute lymphoblastic leukemia. Nat. Genet. 48, 1481–1489 (2016).
46. Willman, C. L. et al. Deletion of IGF1–mapping to chromosome 5q11, in human leukemia and preleukemic myelodysplasia. Science 259, 968–971 (1993).
47. Boulton, I. J. et al. Allelic loss of IGF1 in myelodysplasia and acute myeloid leukemia: retention of IGF1 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 leukemia. Nature 478, 529–533 (2011).
51. Zuber, J. et al. RNA screen identifies Brd4 as a therapeutic target in acute myeloid leukaemia. Nature 478, 524–526 (2011).
52. Gottin, M. A. & Chapman, K. E. The anti-inflammatory and immunosuppressive effects of glucocorticoids, recent developments and mechanistic insights. Mol. Cell Endocrinol. 335, 2–13 (2011).
53. Krum, K. et al. Suppression of B-cell development genes is key to glucocorticoid efficacy in treatment of acute childhood lymphoblastic leukemia. Blood 120, 3000–3008 (2012).
54. Larsen, E. C. et al. Dexamethasone and high-dose methotrexate improve outcome for children and young adults with high-risk B-cell lymphoblastic leukemia: a report from Children’s Oncology Group Study AALL0232. J. Clin. Oncol. 34, 2380–2388 (2016).
55. Mertens, J. A. 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).
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.I.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/.