Published in final edited form as:
Leukemia. 2017 March ; 31(3): 573–579. doi:10.1038/leu.2016.271.

A genome-wide association study identifies risk loci for childhood acute lymphoblastic leukemia at 10q26.13 and 12q23.1

Jayaram Vijayakrishnan[1], **Rajiv Kumar**[2], **Marc Y.R. Henrion**[1], **Anthony V Moorman**[4], **P Sivaramakrishna Rachakonda**[2], **Ismail Hosen**[2], **Miguel Inacio da Silva Filho**[2], **Amy Holroyd**[1], **Sara E. Dobbins**[1], **Rolf Koehler**[3], **Hauke Thomsen**[2], **Julie A. Irving**[4], **James M. Allan**[4], **Tracy Lightfoot**[5], **Eve Roman**[5], **Sally E. Kinsey**[6,7], **Eamonn Sheridan**[7], **Pamela D Thompson**[8], **Per Hoffmann**[9,16], **Markus M. Nöthen**[9], **Stefanie Heilmann-Heimbach**[9], **Jöckel Karl-Heinz**[11], **Mel Greaves**[14], **Christine J. Harrison**[4], **Claus R. Bartram**[3], **Martin Schrappe**[13], **Martin Stanulla**[13], **Kari Hemminki**[2,15], and **Richard S. Houlston**[1]

[1]Division of Genetics and Epidemiology, The Institute of Cancer Research, Sutton, Surrey, United Kingdom [2]Division of Molecular Genetic Epidemiology, German Cancer Research Centre, Heidelberg, Germany [3]Institute of Human Genetics, University of Heidelberg, Heidelberg, Germany [4]Northern Institute for Cancer Research, Newcastle University, Newcastle upon Tyne, United Kingdom [5]Epidemiology and Cancer Statistics Group, Department of Health Sciences, University of York, York, United Kingdom [6]Department of Paediatric and Adolescent Haematology and Oncology, Leeds General Infirmary, Leeds, United Kingdom [7]Leeds Institute of Molecular Medicine, University of Leeds, Leeds, United Kingdom [8]Paediatric and Familial Cancer, Institute of Cancer Sciences, Manchester, United Kingdom [9]Institute of Human Genetics, University of Bonn, Bonn, Germany [10]German Centre for Neurodegenerative Diseases, Bonn, Germany [11]Institute for Medical Informatics, Biometry and Epidemiology, University Hospital Essen, University of Duisburg–Essen, Essen, Germany [12]Department of Paediatric Haematology and Oncology, Hannover Medical School, Hannover, Germany [13]General Paediatrics, University Hospital Schleswig-Holstein, Kiel, Germany [14]Haemato-Oncology Research Unit, Division of Molecular Pathology, Institute of Cancer Research, Sutton, Surrey, United Kingdom [15]Center for Primary Health Care Research, Lund University, Malmö, Sweden [16]Human Genomics Research Group, Department of Biomedicine, University Hospital Basel, Basel, Switzerland

Abstract

Genome-wide association studies (GWASs) have shown that common genetic variation contributes to the heritable risk of childhood acute lymphoblastic leukemia (ALL). To identify new susceptibility loci for the largest subtype of ALL, B-cell precursor ALL (BCP-ALL), we conducted a meta-analysis of two GWAS with imputation using 1000 Genomes and UK10K

Corresponding author: Richard S Houlston; richard.houlston@icr.ac.uk; Tel: +44 (0) 208 722 4175; Fax: +44 (0) 208 722 4365.

Disclosure of Conflicts of Interest
The authors declare no conflicts of interest.

Europe PMC Funders Author Manuscripts

Project data as reference (totaling 1,658 cases and 7,224 controls). After genotyping an additional 2,525 cases and 3,575 controls we identify new susceptibility loci for BCP-ALL mapping to 10q26.13 (rs35837782, *LHPP*, $P = 1.38x10^{-11}$) and 12q23.1 (rs4762284, *ELK3*, $P = 8.41\times10^{-9}$). We also provide confirmatory evidence for the existence of independent risk loci at 9p21.3, but show that the association marked by rs77728904 can be accounted for by linkage disequilibrium with the rare high-impact *CDKN2A* p.Ala148Thr variant rs3731249. Our data provide further insights into genetic susceptibility to ALL and its biology.

Keywords

10q26.13; 12q23.1; risk; B-cell; acute lymphoblastic leukemia

Introduction

Acute lymphoblastic leukemia (ALL) is the major pediatric cancer in western countries, with B-cell precursor (BCP) ALL accounting for approximately 80% of ALL cases1. Despite this, the etiology of ALL is poorly understood and although there is indirect evidence for an infective origin, no specific environmental risk factor has been identified2, 3. Evidence for inherited predisposition to ALL is provided by the increased risk shown in siblings of cases independent of the concordance in monozygotic twins, which has an *in utero* etiology4. Support for polygenic susceptibility to ALL has come from genome-wide association studies (GWAS)5–9. While these studies have so far identified single-nucleotide polymorphisms (SNPs) at seven loci influencing BCP-ALL at 7p12.2 (*IKZF1*), 9p21.3 (*CDKN2A*, two risk loci), 10p12.2 (near *PIP4K2A*), 10p14 (*GATA3*), 10q21.2 (near *ARID5B*) and 14q11.2 (near *CEBPE*), statistical modelling using Genome-Wide Complex Trait Analysis predicts that additional risk loci conferring modest effects should be identifiable by further GWAS10.

Recovery of untyped genotypes through imputation provides a mechanism of exploiting GWAS datasets to identify new risk alleles11. Additionally it enables fine mapping and refinement of association signals, for example, in identification of the *CDKN2A* p.Ala148Thr variant rs3731249 (hg19 chr9:g.21970916 G>A) as contributing to the 9p21.3 association signal8. Recently, the use of the 1000 Genomes Project and the UK10K projects as a combined reference panel has been shown to improve imputation accuracy compared with using the 1000 Genomes Project data alone12, 13.

Here we report imputation using the 1000 Genomes and the UK10K Project data as reference and meta-analysis of two GWAS to identify new susceptibility alleles for BCP-ALL. After replication genotyping in three additional case-control series we have identified new risk loci for BCP-ALL at 10q26.13 and 12q23.1. Our findings provide further insights into the genetic and biological basis of this hematological malignancy.

Europe PMC Funders Author Manuscripts

Methods

Ethics

Collection of samples and clinico-pathological information from subjects was undertaken with informed consent in accordance with the Declaration of Helsinki and ethical board approval. Ethical committee approval was obtained for Medical Research Council UKALL97/99 trial by individual UK treatment centers and approval for UKALL2003 was obtained from the Scottish Multi-Centre Research Ethics Committee (REC:02/10/052)14, 15. Additional ethical approval was obtained under the auspices of the Childhood Leukaemia Cell Bank, the United Kingdom Childhood Cancer Study and University of Heidelberg.

GWAS data

The United Kingdom (UK)-GWAS and German-GWAS data sets have been previously reported6, 7. Briefly, the UK-GWAS was based on constitutional DNA (*i.e.* remission samples) of 459 white BCP-ALL cases from the United Kingdom Childhood Cancer Study (UKCCS; http://www.ukccs.org/; 258 male; mean age at diagnosis 5.3 years); 342 cases from the UK Medical Research Council (MRC) ALL 97/99 (1997-2002) trial (190 male; mean age of diagnosis 5.7 years) and 23 cases from Northern Institute for Cancer Research (16 male). Genotyping was performed using Illumina Human 317K arrays (Illumina, San Diego; Available at: http://www.illumina.com). For controls we used publicly accessible data generated by the Wellcome Trust Case Control Consortium 2 (http://www.wtccc.org.uk/) from 2,699 individuals in the 1958 British Birth Cohort (Hap1.2M-Duo Custom array data) and 2,501 individuals from the UK Blood Service. The German GWAS was comprised of 1,155 cases (620 male; mean age at diagnosis 6.0 years) ascertained through the Berlin-Frankfurt-Münster (BFM) trials (1993-2004) genotyped using Illumina Human OmniExpress-12v1.0 arrays. For controls we used genotype data from 2,132 healthy individuals from the Heinz Nixdorf Recall (HNR) study; consisting of 704 individuals genotyped using Illumina-HumanOmni1-Quad_v1 and 1,428 individuals genotyped on Illumina-HumanOmniExpress-12v1.0 platform. In total we obtained 1,658 BCP ALL cases and 7224 matched controls from the two GWAS series combined.

Quality control of GWAS samples

The quality control steps of UK and GERMAN GWAS study samples have been have been described in previous studies6, 7. After quality control steps we obtained 824 cases and 5,200 controls for the UKGWAS data set and 834 cases and 2,024 controls from the German data sets that were then used for further genotyping and imputation analysis.

Replication series and genotyping

The UK replication series comprised 1,150 patients (504 male; mean age at diagnosis 6.2 years) ascertained through the UK ALL-2003 (2003-2011) and ALL-97/99 trials14, 15. Immunophenotyping of diagnostic samples was undertaken using standard methods. The 2,100 controls (702 male) were ethnically-matched healthy individuals with no personal history of cancer recruited to the National Study of Colorectal Cancer Genetics (NSCCG)16

and the Genetic Lung Cancer Predisposition Study (GELCAPS)17. Genotyping of cases and controls was performed using competitive allele-specific polymerase chain reaction KASPAR chemistry (LCG Biosciences Ltd., Hertfordshire, UK). The German replication series consisted of 1,501 patients ascertained (794 males; mean age at diagnosis, 6.2 years ascertained through the BFM trials (1993-2004)18. The controls comprised 1,516 individuals (762 males; mean age, 58.2 years) ethnically-matched healthy individuals of German origin recruited at the Institute of Transfusion Medicine in Manheim, Germany, 2004. Samples having SNP call rates of <90% were excluded from the analysis. To ensure quality of genotyping in all assays, at least 2 negative controls and 1% to 2% duplicates (concordance >99.9%) were genotyped. All primers and probes used are detailed in Table S8. Combing both replication series we had access to 2,651 B-cell ALL cases and 3,616 matched controls for the current study.

Sanger sequencing

To confirm the fidelity of imputation a random subset of samples were sequenced using BigDye® Terminator v3.1 Cycle Sequencing Kit (Life Technologies) in conjunction with ABI 3700xl semi-automated sequencers (Applied Biosystems). Primer sequences are detailed in Table S10.

Statistical and bioinformatics

Main data analysis were undertaken using R version 2.15.2 (R Core Team, 2013; http://www.R-project.org/), PLINK v1.919 and SNPTEST v2.4.1 software20. The two GWAS data sets were imputed for over 10 million variants using IMPUTE2 v2.3.0 software 21, 22 and data from the 1000 Genomes Project (Phase 1 integrated variant set, v3.20101123, http://www.1000genomes.org, 9 December 2013) and UK10K (ALSPAC, EGAS00001000090 / EGAD00001000195, and TwinsUK, EGAS00001000108 / EGAD00001000194, studies only; http://www.uk10k.org/) as reference. Data sets were first phased using SHAPEIT v2.12 prior to imputation to accurately estimate haplotypes23. The adequacy of case-control matching and possibility of differential genotyping between cases and controls was evaluated using quantile-quantile plots of test statistics to compute λ_{100}. Test of association between imputed SNPs and childhood ALL was performed using a missing data likelihood score test under a frequentist additive model in software SNPTEST. Eigenvectors for the German data set were inferred using *smartpca* component within EIGENSOFT v2.424, 25 and Eigenstrat adjustment was carried out by including the first 2 eigenvectors as covariates in SNPTEST during association analysis. Post imputation and SNPTEST, only markers with info scores >0.4, imputed call rates/SNP >0.9, MAFs >0.005 and a posterior imputation quality threshold of 0.5 or higher were included in further analysis. SNPs that deviated from Hardy Weinberg equilibrium (HWE) at *P*-values $<10^{-5}$ were also excluded from further analysis. Meta-analysis of post QC GWAS datasets was conducted in META 1.3.120, 21, 26, under a fixed-effects model using the inverse variance approach. We calculated Cochran's Q statistic to test for heterogeneity and the I^2 statistic to quantify the proportion of the total variation attributable to heterogeneity27. The presence of secondary association signals due to allelic heterogeneity in risk loci were carried out using a conditional analysis in SNPTEST by adjusting for the sentinel SNP using the '–condition-on' option. Logistic regression association analysis and meta-analysis of the replication data sets under fixed-

effects were carried out using STATA v.10 software (Stata Corporation, College Station, TX, USA).

Linkage disequilibrium (LD) metrics were calculated using vcftools v0.1.12b26 (http://vcftools.sourceforge.net) using UK10K data. HapMap recombination rate (cM/Mb) were defined by Oxford recombination hotspots28, 29.

Chromatin state dynamics

To explore the epigenetic profile of association signals, we used 15-state chromatin segmentation data learned by computationally integrating chIP-seq data for GM12878 lymphoblastoid cells inferred from ENCODE Histone Modification data (H4K20me1, H3K9ac, H3K4me3, H3K4me2, H3K4me1, H3K36me3, H3K27me3, H3K27ac, and CTCF) and binarized using a multivariate Hidden Markov Model (http://genome.ucsc.edu/ENCODE/)30. Risk SNPs and their proxies (*i.e.,* $r^2 > 0.8$ in the 1000 Genomes EUR reference panel) were annotated for putative functional effect using HaploReg v331 and RegulomeDB32 and SeattleSeq33 Annotation. These servers make use of data from ENCODE30, genomic evolutionary rate profiling (GERP)34 conservation metrics, combined annotation dependent depletion (CADD) scores35 and PolyPhen scores36. Similarly we searched for overlap with “super-enhancer” regions as defined by Hnisz *et al*, 37 restricting analysis to GM12878 cells.

Expression quantitative trait locus analysis

Expression quantitative trait locus (eQTL) analysis was performed for all genes in 1Mb regions spanning rs4762284 and rs35837782 by querying mRNA expression data from MuTHER38 and Blood eQTL browser39.

Chromosome karyotyping and 9p21.3 deletion status

Conventional cytogenetic studies on diagnostic ALL tumor cells were conducted using standard karyotyping methodologies, and standard criteria for the definition of a clone were applied. Genomic copy number at 9p21.3 was assayed using FISH and MLPA as previously described40, 41

Relationship between SNP genotype and survivorship

To investigate if genotype is associated with clinical phenotype or outcome, we analysed data on patients recruited to AIEOP-BFM 200018. Briefly, patients received standard chemotherapy (*i.e.*, prednisone, vincristine, daunorubicin, l-asparaginase, cyclophosphamide, ifosfamide, cytarabine, 6-mercaptopurine, 6-thioguanine, and methotrexate) with a subset of high-risk patients treated with cranial irradiation and/or stem cell transplantation. Event-free survival (EFS) was defined as the time from diagnosis to the date of last follow-up in complete remission or to the first event. Events were resistance to therapy (nonresponse), relapse, secondary neoplasm, or death from any cause. Failure to achieve remission due to early death or nonresponse was considered as an event at time zero and patients lost to follow-up were censored at the time of their withdrawal. Patients were stratified into 3 categories: standard, intermediate, and high risk. Although minimal residual disease (MRD) analysis was the main stratification criterion, high risk was also defined by

prednisone poor-response or ≥5% leukemic blasts in bone marrow on day 33, or t (9;22)/ t(4;11) positivity or their molecular equivalents (BCR-ABL/MLL-AF4-fusion) independent of MRD status. Standard patients were MRD-negative on treatment day 33 (TP1) and 78 (TP2) and had no high-risk criteria. High-risk patients were defined as having residual disease ($\geq 10^{-3}$ cells) at TP2. Intermediate patients had positive-MRD detection at either TP1 or TP2, but had a cell count of $<10^{-3}$ at TP2. The Kaplan–Meier method was used to estimate survival rates, differences were compared with the two-sided log-rank test42, 43. Cumulative incidence functions for competing events were constructed by the method of Kalbfleisch and Prentice and were compared employing the Gray's test44, 45. Computations were performed using SASv9.1 (SAS, Cary, NC, USA).

Heritability analysis

We used Genome-wide Complex Trait Analysis (GCTA) to estimate the polygenic variance (*i.e.* heritability) ascribable to all GWAS SNPs 46. SNPs were excluded based on minor allele frequency (<0.01), missing genotype rate (0.05) and deviation from HWE (P<0.05). Individuals were excluded for exhibiting an excess of missing genotype (>0.02) and where two individuals were closely related (genetic relatedness score > 0.05). A genetic relationship matrix (GRM) of pairs of samples was used as input for the restricted maximum likelihood analysis to estimate the heritability explained by the selected set of SNPs. Regions of high LD in the genome were excluded from the analysis. Imposing a prevalence of 0.00052 for childhood ALL we estimated the heritability explained by risk SNPs identified by GWAS as located within autosomal regions associated with ALL. For each risk SNP the heritability was estimated for all chromosomes simultaneously using the risk SNP genotype as a covariate. In chromosomes bearing multiple independent risk loci, all the risk SNPs in that chromosome were used as covariates to get the combined contribution of risk SNPs towards heritability. The heritability associated with the risk SNPs was taken to be the difference between the heritability of the chromosome on which it is found as calculated with and without covariate adjustment for the SNP.

Calculation of polygenic risk scores

In addition to the two new risk loci described here seven previously reported risk loci were included in the calculation of the Polygenic Risk Scores (PRS) for childhood ALL (rs10828317, 10p12.2; rs3824662, 10p14; rs7089424, 10q21.2; rs2239633, 14q11.2; rs4132601, 7p12.2; rs3731249, 9p21.3; rs3731217, 9p21.3; rs35837782, 10q26.13; rs4762284, 12q23.1). The eight variants are thought to act independently as previous studies have shown no interaction between risk loci5–7. PRS were constructed using methods established by Pharoah *et al,* based on log-normal distribution $LN(\mu,\sigma^2)$ of mean μ, and variance σ^2 (i.e. relative risk is normally distributed on a logarithmic scale)47. Standardized incidence ratios for familial risk in singleton siblings and twins for childhood ALL were assumed to be 3.24. Familial risk was calculated by dividing polygenic variation over the square root of familial risk.

Results

Association analysis

To identify new susceptibility loci for BCP-ALL we conducted a pooled meta-analysis of two GWAS in populations of European ancestry, the UK-GWAS and the German-GWAS (see Methods). After filtering, the studies provided genotype data on 1,658 cases and 7,224 controls. To achieve consistent and dense genome-wide coverage, we imputed unobserved genotypes at >10 million SNPs using a combined reference panel comprising 1,092 individuals from the 1000 Genomes Project and 3,781 individuals from the UK10K project. Quantile-quantile plots of SNPs (minor allele frequency (MAF) >0.5%) post-imputation did not show evidence of substantive over-dispersion introduced by imputation (genomic inflation λ_{100} for UK and German GWAS was 1.016 and 1.009 respectively; Figure S1).

Pooling data from both GWAS we derived joint odds ratios (ORs) and 95% confidence intervals (CIs) under a fixed-effects model for each SNP with MAF >0.5% and associated per allele *P* values. From this analysis we identified the top ranked SNPs in 20 distinct regions and not previously implicated in the risk of developing BCP-ALL (Table S1). After confirming the fidelity of imputation by Sanger sequencing (Table S2) we successfully designed and optimized allele-specific PCR (KASPAR) assays for 14 SNPs. We sought validation of associations by genotyping additional UK and German case-control series totaling 2,525 cases and 3,575 controls (Table S3).

In the combined analysis of data from these replication series, rs35837782 (10q26.13, hg19 chr10:g.126293309) and rs4762284 (12q23.1, hg19 chr12:g.96612762) showed significant support for an association with BCP-ALL, with *P*-values and ORs of 3.66×10^{-6}, 1.20 and 3.88×10^{-4} , 1.16 respectively (Table 1; Table S4; Figure S3). In a meta-analysis of the discovery GWAS and replication series, these associations attained genome-wide significance (rs35837782, $P = 1.38 \times 10^{-11}$) and (rs4762284, $P = 8.41 \times 10^{-9}$) (Table 1; Table S4; Figure S3).

Conditional association analyses

To explore the possibility of multiple risk loci at 10q26.13 and 12q23.1 and previously identified GWAS risk loci we performed conditional analyses. At 10q26.13 and 12q23.1 we found no evidence for signals independent of SNPs rs35837782 and rs4762284. Similarly at 7p12.2, 10p12.2, 10p14, 10q21.2 and 14q11.2 we found no support for the existence of multiple risk loci.

We and others have recently sought to decipher the GWAS signal at the 9p21.3 locus[8, 48–50]. Our conditional analysis supports the assertion of an additional locus in this region, independent of the original GWAS SNP rs3731217, which is best captured by the rare coding SNP rs3731249 (MAF=0.03 in CEU, r2=0.005, D'=1.00 with rs3731217; Table S5). rs3731249, encoding *CDKN2A* p.Ala148Thr, has been shown to reduce tumor suppressor function of p16INK4A, increase susceptibility to leukemic transformation of hematopoietic progenitor cells and to be preferentially retained in ALL cells[49]. The more common variant rs662463 correlated with rs77728904 has concurrently been suggested as a plausible causative variant underlying this new association signal[48] (MAF=0.07, r^2=0.16, D'=1.00

Europe PMC Funders Author Manuscripts

with rs3731249). Despite some evidence that rs77728904 variant is a cis-eQTL for *CDKN2B*48, this association signal is entirely captured by rs3731249 (*P*-values before and after conditioning: 6.26 x $10^{-7}$and 0.10 respectively; Table S5). Here our analysis has been constrained to the identification of variants which can be imputed with high fidelity, hence it does not exclude the possibility of rarer variants with higher impact, especially indels potentially impacting on ALL risk. This exemplifies the difficultly in elucidating the genetic basis of such functionally rich genomic regions. Once correcting for these two signals, no additional statistically significant association was detected in this region.

Relationship between the new ALL risk SNPs and tumor profile

Given the biological heterogeneity of BCP-All, we analyzed the association between rs35837782 and rs4762284 genotypes and the major subtypes of BCP-ALL, hyperdiploidy (*i.e.* >50 chromosomes), *ETV6-RUNX1* and others (Table S4; Figure S3). Analysis of these data provided no consistent evidence that the risk of rs35837782 and rs4762284 was confined to hyperdiploid, *ETV6-RUNX1* or non-hyperdiploid/non-*ETV6-RUNX1* subtypes of B-ALL. Similarly, we found no evidence for a relationship between rs35837782 and rs4762284 genotypes and other chromosomally defined forms of BCP-ALL defined by t(9;22)(q34;q11), t(1;19)(q23;p13), and t(4;11)(q21;q23) karyotype or *CDKN2A* deletion status after adjustment for multiple testing (Table S6). Finally, we found no evidence that rs35837782 and rs4762284 genotype was associated with age at diagnosis or sex or influenced patient outcome as defined by event-free survival by analyzing data on 810 patients from the AIEOP-BFM 2002 trial (Figure S4, Table S7, and Table S8).

Impact on the heritable risk

By fitting all SNPs from GWAS simultaneously, the estimated heritability of ALL attributable to all common variation is 12.1% (± 3.8%). This estimate represents the additive variance and therefore, does not include the potential impact of gene-gene interactions or dominance effects or gene-environment interactions impacting on ALL risk. Moreover, given the evidence, albeit indirect, of a role for infectious exposure in relation to ALL risk, it is possible that substantive gene-environment effects operate. While the currently identified risk SNPs (newly discovered and previously identified) only account for 19% of the additive heritable risk the OR effect sizes of the ALL risk SNPs are among the highest reported in GWAS of any cancer type and in combination they impact significantly on disease risk with those in the top 1% of genetic risk having a 6.2-fold relative of ALL (Figure S5). The power of our GWAS to identify common alleles conferring relative risks of 1.5 or greater (such as the 7p12.2 variant) is high (~80%). Hence, there are unlikely to be many additional SNPs with similar effects for alleles with frequencies greater than 0.3 in populations of European ancestry. In contrast, our analysis had limited power to detect alleles with smaller effects and/or MAF<0.1.

Biological inference

At 10q26.13, rs35837782 localizes to intron 6 on the gene encoding phospholysine phosphohistidine inorganic pyrophosphate phosphatase (LHPP; Fig. 1) with genes *FAM53B* and *METTL10* mapping nearby. The SNP rs4762284 at 12q23.1 maps to intron 1 of the

gene encoding the ETS-domain protein (ELK3) , with nearby genes including *CDK17* (Fig. 1).

To explore the epigenetic profile of association signals at each of the two new risk loci we used HaploReg and RegulomeDB to examine whether the sentinel SNPs and those in high LD (*i.e.,* $r^2 > 0.8$ in the 1000 Genomes EUR reference panel) annotate putative transcription factor binding or enhancer elements (Table S9). The SNP rs4762284 resides within a region of open chromatin, common across multiple cell lines, consistent with a regulatory element such as an enhancer or a promoter. To gain further insight into the functional basis of rs35837782 and rs4762284 associations we examined for an association between SNP genotype and expression of genes mapping within 1Mb of sentinel SNPs. We made use of publicly available expression data on blood cells, lymphoblastoid cell lines from HapMap3, Geneva, and the Multiple Tissue Human Expression Resource pilot data. In blood rs4762284 genotype was associated with *ELK3* expression at $P = 6.85 \times 10^{-4}$) with the risk allele correlated with reduced expression (Table S10)39.

Discussion

In this analysis of BCP-ALL, we have identified common variants at 10q26.13 and 12q23.1. It has recently been proposed that many GWAS signals are a consequence of ‘synthetic associations’, resulting from the combined effect of one or more rare causal variants rather than simply LD with a common risk variant51, 52. Support for such a model in ALL is provided by the rare high-impact variant rs3731249 in *CDKN2A*8 which is in LD with rs77728904. Since imputation using UK10K as reference can accurately recover genotypes for variants with MAFs of 0.5%12 the possibility that either 10q26.13 or 12q23.1 associations have a similar genetic basis is highly unlikely.

Given the existence of immuno-genetic subtypes of BCP-ALL, it is perhaps not surprising there is variability in the genetic effects on ALL risk by subtype, with 10q21.2 variants influencing hyperdiploid ALL and 10p14 variants influencing non-hyperdiploid/non-*ETV6-RUNX1* disease6, 7. In contrast to the 7p12.2 and 10p12.2 risk variants6, 7 the 10q26.13 and 12q23.1 loci have generic effects on the development of ALL.

Because rs35837782 and rs4762284 localize to *LHPP* and *ELK3*, respectively, it is plausible that the functional basis of these associations are mediated through these genes. ELK3, an ETS-domain transcription factor is an attractive candidate for defining ALL susceptibility *a priori* since it plays a role in both B-cell development and IgH gene regulation53. ELK3, which is a member of ETS family of transcription factors, interacting with TCF3 transcription factor 3 (E2A immunoglobulin enhancer-binding factors E12/E47) which is involved in several ALL specific gene fusions including *TCF3-PBX1*/t(1;19)(q23;p13) and *TCF3-HLF*/t(17;19)(q23;p13) ALL54. ELK3 is highly expressed primarily at the early stages of B-lymphocyte development with expression declining drastically upon B-cell maturation, correlating with the activity of the enhancer of the immunoglobulin heavy chain53. Hence genetically determined reduced expression is compatible with B-cell developmental arrest, a hallmark of ALL. In contrast to ELK3 evidence for a role for *LHPP*, which encodes a diphosphatase, in B-cell development or B-cell malignancy is yet to be

established55. While the identified risk SNPs map within regions of active chromatin within B-cells and thus have a role in the B-cell cis-regulatory network *a priori*, additional laboratory follow-up is required to decipher their functional basis.

In summary, our findings represent a further important step in defining the contribution of inherited genetic variants to the risk of developing ALL. Our current and previous findings are notable because we have defined associations of several regions with susceptibility to ALL, and these regions harbor plausible candidate genes for further investigation. Moreover they emphasize the role of genetically determined expression of B-cell developmental genes being key players in ALL. Given that there remains significant missing heritability for ALL, future GWAS-based studies in concert with functional analyses are likely to lead to further insights into ALL biology.

Supplementary Files

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

In the UK Bloodwise provided principal funding for the study (13027, 13044). Support from Cancer Research UK (C1298/A8362, supported by the Bobby Moore Fund) is also acknowledged. Support from José Carreras Stiftung under project DJCLS R 13/19 is also acknowledged. The study made use of genotyping data from the Wellcome Trust Case Control Study on the 1958 Birth Cohort. We thank S. Richards and J. Burrett (Clinical Trials Service Unit, Oxford), J. Simpson (Univ. York), P. Thomson and A. Hussain (Cancer Immunogenetics, School of Cancer Sciences, Univ. Manchester) for assistance with data harmonisation. We thank the Childhood Leukaemia Cell Bank for access to samples. We are grateful to the UK Cancer Cytogenetics Group (UKCCG) for data collection and provision of samples. P. Thompson (Paediatric and Familial Cancer, Institute of Cancer Sciences, The University of Manchester) is funded by Children with Cancer. Genotyping of German cases and controls was partly covered by funding from Tumorzentrum Heidelberg-Mannheim and Deutsche Krebshilfe. We are grateful to all the study subjects for their participation and their clinicians who contributed to the blood sample and data collection. The Heinz Nixdorf Foundation supported the population based Heinz Nixdorf Recall (HNR) study; the genotyping of the Illumina HumanOmni-1 Quad BeadChips of the HNR subjects was financed by the German Centre for Neurodegenerative Disorders (DZNE, Bonn). Additionally, the study is funded by the German Ministry of Education and Science and the German Research Council (DFG; Project SI 236/8-1, SI236/9-1, ER 155/6-1). We are extremely grateful to all investigators who contributed to the generation of this dataset and to all subjects for their participation.

References

1. Stiller CA, Parkin DM. Geographic and ethnic variations in the incidence of childhood cancer. British medical bulletin. 1996 Oct; 52(4):682–703. [PubMed: 9039726]
2. Greaves M. Infection, immune responses and the aetiology of childhood leukaemia. Nature reviews Cancer. 2006 Mar; 6(3):193–203. [PubMed: 16467884]
3. Crouch S, Lightfoot T, Simpson J, Smith A, Ansell P, Roman E. Infectious illness in children subsequently diagnosed with acute lymphoblastic leukemia: modeling the trends from birth to diagnosis. American journal of epidemiology. 2012 Sep 1; 176(5):402–408. [PubMed: 22899827]
4. Kharazmi E, da Silva Filho MI, Pukkala E, Sundquist K, Thomsen H, Hemminki K. Familial risks for childhood acute lymphocytic leukaemia in Sweden and Finland: far exceeding the effects of known germline variants. British journal of haematology. 2012 Dec; 159(5):585–588. [PubMed: 23025517]
5. Sherborne AL, Hosking FJ, Prasad RB, Kumar R, Koehler R, Vijayakrishnan J, et al. Variation in CDKN2A at 9p21.3 influences childhood acute lymphoblastic leukemia risk. Nature genetics. 2010 Jun; 42(6):492–494. [PubMed: 20453839]

6. Migliorini G, Fiege B, Hosking FJ, Ma Y, Kumar R, Sherborne AL, et al. Variation at 10p12.2 and 10p14 influences risk of childhood B-cell acute lymphoblastic leukemia and phenotype. Blood. 2013 Nov 7; 122(19):3298–3307. [PubMed: 23996088]
7. Papaemmanuil E, Hosking FJ, Vijayakrishnan J, Price A, Olver B, Sheridan E, et al. Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. Nature genetics. 2009 Sep; 41(9):1006–1010. [PubMed: 19684604]
8. Vijayakrishnan J, Henrion M, Moorman AV, Fiege B, Kumar R, da Silva Filho MI, et al. The 9p21.3 risk of childhood acute lymphoblastic leukaemia is explained by a rare high-impact variant in CDKN2A. Scientific reports. 2015; 5:15065. [PubMed: 26463672]
9. Sherborne AL, Hemminki K, Kumar R, Bartram CR, Stanulla M, Schrappe M, et al. Rationale for an international consortium to study inherited genetic susceptibility to childhood acute lymphoblastic leukemia. Haematologica. 2011 Jul; 96(7):1049–1054. [PubMed: 21459794]
10. Enciso-Mora V, Hosking FJ, Sheridan E, Kinsey SE, Lightfoot T, Roman E, et al. Common genetic variation contributes significantly to the risk of childhood B-cell precursor acute lymphoblastic leukemia. Leukemia. 2012 Oct; 26(10):2212–2215. [PubMed: 22456626]
11. Servin B, Stephens M. Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS genetics. 2007 Jul.3(7):e114. [PubMed: 17676998]
12. Huang J, Howie B, McCarthy S, Memari Y, Walter K, Min JL, et al. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. Nature communications. 2015; 6:8111.
13. Walter K, Min JL, Huang J, Crooks L, Memari Y, et al. Consortium UK. The UK10K project identifies rare variants in health and disease. Nature. 2015 Oct 1; 526(7571):82–90. [PubMed: 26367797]
14. Hann I, Vora A, Richards S, Hill F, Gibson B, Lilleyman J, et al. Benefit of intensified treatment for all children with acute lymphoblastic leukaemia: results from MRC UKALL XI and MRC ALL97 randomised trials. UK Medical Research Council's Working Party on Childhood Leukaemia. Leukemia. 2000 Mar; 14(3):356–363. [PubMed: 10720126]
15. Vora A, Goulden N, Wade R, Mitchell C, Hancock J, Hough R, et al. Treatment reduction for children and young adults with low-risk acute lymphoblastic leukaemia defined by minimal residual disease (UKALL 2003): a randomised controlled trial. The Lancet Oncology. 2013 Mar; 14(3):199–209. [PubMed: 23395119]
16. Penegar S, Wood W, Lubbe S, Chandler I, Broderick P, Papaemmanuil E, et al. National study of colorectal cancer genetics. British journal of cancer. 2007 Nov 5; 97(9):1305–1309. [PubMed: 17895893]
17. Wang Y, Broderick P, Webb E, Wu X, Vijayakrishnan J, Matakidou A, et al. Common 5p15.33 and 6p21.33 variants influence lung cancer risk. Nature genetics. 2008 Dec; 40(12):1407–1409. [PubMed: 18978787]
18. Conter V, Bartram CR, Valsecchi MG, Schrauder A, Panzer-Grumayer R, Moricke A, et al. Molecular response to treatment redefines all prognostic factors in children and adolescents with B-cell precursor acute lymphoblastic leukemia: results in 3184 patients of the AIEOP-BFM ALL 2000 study. Blood. 2010 Apr 22; 115(16):3206–3214. [PubMed: 20154213]
19. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics. 2007 Sep; 81(3):559–575. [PubMed: 17701901]
20. Liu JZ, Tozzi F, Waterworth DM, Pillai SG, Muglia P, Middleton L, et al. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. Nature genetics. 2010 May; 42(5):436–440. [PubMed: 20418889]
21. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nature genetics. 2007 Jul; 39(7):906–913. [PubMed: 17572673]
22. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS genetics. 2009 Jun.5(6):e1000529. [PubMed: 19543373]

Europe PMC Funders Author Manuscripts

23. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nature methods. 2012 Feb; 9(2):179–181.
24. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics. 2006 Aug; 38(8):904–909. [PubMed: 16862161]
25. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS genetics. 2006 Dec. 2(12):e190. [PubMed: 17194218]
26. de Bakker PI, Ferreira MA, Jia X, Neale BM, Raychaudhuri S, Voight BF. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Human molecular genetics. 2008 Oct 15; 17(R2):R122–128. [PubMed: 18852200]
27. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Statistics in medicine. 2002 Jun 15; 21(11):1539–1558. [PubMed: 12111919]
28. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. A fine-scale map of recombination rates and hotspots across the human genome. Science. 2005 Oct 14; 310(5746):321–324. [PubMed: 16224025]
29. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al. The structure of haplotype blocks in the human genome. Science. 2002 Jun 21; 296(5576):2225–2229. [PubMed: 12029063]
30. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6; 489(7414):57–74. [PubMed: 22955616]
31. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic acids research. 2012 Jan; 40(Database issue):D930–934. [PubMed: 22064851]
32. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome research. 2012 Sep; 22(9): 1790–1797. [PubMed: 22955989]
33. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009 Sep 10; 461(7261):272–276. [PubMed: 19684571]
34. Cooper GM, Stone EA, Asimenos G, Program NCS. Green ED, Batzoglou S, et al. Distribution and intensity of constraint in mammalian genomic sequence. Genome research. 2005 Jul; 15(7): 901–913. [PubMed: 15965027]
35. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014 Mar; 46(3):310–315. [PubMed: 24487276]
36. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nature methods. 2010 Apr; 7(4):248–249. [PubMed: 20354512]
37. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-Andre V, Sigova AA, et al. Super-enhancers in the control of cell identity and disease. Cell. 2013 Nov 7; 155(4):934–947. [PubMed: 24119843]
38. Nica AC, Parts L, Glass D, Nisbet J, Barrett A, Sekowska M, et al. The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. PLoS genetics. 2011; 7(2):e1002003. [PubMed: 21304890]
39. Westra H-J, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nature genetics. 2013; 45(10):1238–1243. 10//print. [PubMed: 24013639]
40. Sulong S, Moorman AV, Irving JA, Strefford JC, Konn ZJ, Case MC, et al. A comprehensive analysis of the CDKN2A gene in childhood acute lymphoblastic leukemia reveals genomic deletion, copy number neutral loss of heterozygosity, and association with specific cytogenetic subgroups. Blood. 2009 Jan 1; 113(1):100–107. [PubMed: 18838613]
41. Schwab CJ, Chilton L, Morrison H, Jones L, Al-Shehhi H, Erhorn A, et al. Genes commonly deleted in childhood B-cell precursor acute lymphoblastic leukemia: association with cytogenetics and clinical features. Haematologica. 2013 Jul; 98(7):1081–1088. [PubMed: 23508010]

42. Kaplan EL, M P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958; (53):457–481.
43. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer chemotherapy reports. 1966 Mar; 50(3):163–170. [PubMed: 5910392]
44. John, D., Kalbfleisch, RLP. The Statistical Analysis of Failure Time Data. Second Edition. Wiley Series in Probability and Statistics; 2002.
45. Gray RJ. A class of K-sample tests for comparing the cumulative incidence of a competing risk. Ann Stat. 1988; (19):1141–1154.
46. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. American journal of human genetics. 2011 Jan 7; 88(1):76–82. [PubMed: 21167468]
47. Pharoah PD, Antoniou A, Bobrow M, Zimmern RL, Easton DF, Ponder BA. Polygenic susceptibility to breast cancer and implications for prevention. Nature genetics. 2002 May; 31(1): 33–36. [PubMed: 11984562]
48. Hungate EA, Vora SR, Gamazon ER, Moriyama T, Best T, Hulur I, et al. A variant at 9p21.3 functionally implicates CDKN2B in paediatric B-cell precursor acute lymphoblastic leukaemia aetiology. Nature communications. 2016; 7:10635.
49. Xu H, Zhang H, Yang W, Yadav R, Morrison AC, Qian M, et al. Inherited coding variants at the CDKN2A locus influence susceptibility to acute lymphoblastic leukaemia in children. Nature communications. 2015; 6:7553.
50. Walsh KM, de Smith AJ, Hansen HM, Smirnov IV, Gonseth S, Endicott AA, et al. A Heritable Missense Polymorphism in CDKN2A Confers Strong Risk of Childhood Acute Lymphoblastic Leukemia and Is Preferentially Selected during Clonal Evolution. Cancer research. 2015 Nov 15; 75(22):4884–4894. [PubMed: 26527286]
51. Anderson CA, Soranzo N, Zeggini E, Barrett JC. Synthetic associations are unlikely to account for many common disease genome-wide association signals. PLoS biology. 2011; 9(1):e1000580. [PubMed: 21267062]
52. Wray NR, Purcell SM, Visscher PM. Synthetic associations created by rare variants do not explain most GWAS results. PLoS biology. 2011; 9(1):e1000579. [PubMed: 21267061]
53. Lopez M, Oettgen P, Akbarali Y, Dendorfer U, Libermann TA. ERP, a new member of the ets transcription factor/oncoprotein family: cloning, characterization, and differential expression during B-lymphocyte development. Molecular and cellular biology. 1994 May; 14(5):3292–3309. [PubMed: 7909357]
54. Moorman AV. The clinical relevance of chromosomal and genomic abnormalities in B-cell precursor acute lymphoblastic leukaemia. Blood reviews. 2012 May; 26(3):123–135. [PubMed: 22436535]
55. Yokoi F, Hiraishi H, Izuhara K. Molecular cloning of a cDNA for the human phospholysine phosphohistidine inorganic pyrophosphate phosphatase. Journal of biochemistry. 2003 May; 133(5):607–614. [PubMed: 12801912]

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

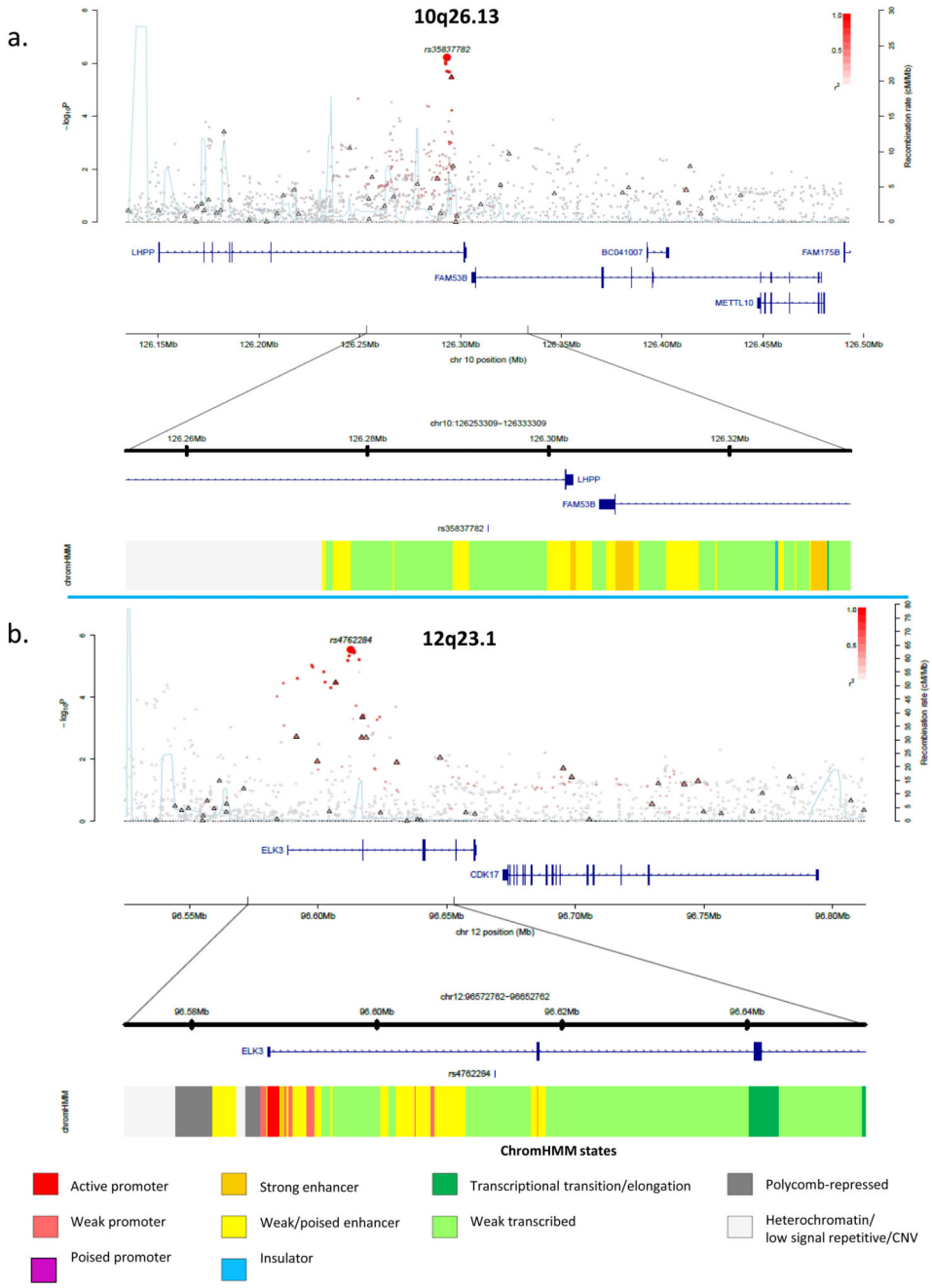

Figure 1. Regional plots of association results and recombination rates for the newly identified risk loci for BCP-ALL (a-b).

Results for 10q26.13 (rs35837782, a) and 12q23.1 (rs4762284, b). Plots (using visPig) show association results of both genotyped (triangles) and imputed (circles) SNPs in the GWAS samples and recombination rates. –log10 P values (y axes) of the SNPs are shown according to their chromosomal positions (x axes). The sentinel SNP in each combined analysis is shown as a large circle or triangle and is labelled by its rsID. The color intensity of each symbol reflects the extent of LD with the top genotyped SNP, white ($r^2 = 0$) through to dark red ($r^2 = 1.0$). Genetic recombination rates, estimated using UK10K Genomes Project

samples, are shown with a light blue line. Physical positions are based on NCBI build 37 of the human genome. Also shown are the relative positions of genes and transcripts mapping to the region of association. Genes have been redrawn to show their relative positions; therefore, maps are not to physical scale. The lower panel is the chromatin-state segmentation track (ChromHMM) for lymphoblastoid cells using data from the HapMap ENCODE Project.

Europe PMC Funders Author Manuscripts

Table 1

Risk to childhood acute lymphoblastic leukemia at loci 10q26.13 and 12q23.1.

RAF: Risk Allele Frequency for rs35837782 is G and risk allele for rs4762284 is T, a. OR: Odds Ratio, b. CI: Confidence Interval.

	Case genotypes				Control genotypes						
rs35837782 (10q26.13, *LHPP*)	**RAF**	**AA**	**AG**	**GG**	**RAF**	**AA**	**AG**	**GG**	**OR[a]**	**95% CI[b]**	***P*-value**
UK GWAS	0.67	93	358	373	0.62	745	2,510	1,945	1.27	(1.14-1.41)	$2.04X10^{-5}$
German GWAS	0.67	93	365	376	0.63	268	952	804	1.19	(1.05-1.36)	0.0072
Meta									1.24	(1.14-1.34)	$6.04 X 10^{-7}$
UK replication 1	0.67	73	211	251	0.62	150	475	405	1.19	(1.03-1.39)	0.022
UK replication 2	0.67	56	265	253	0.61	167	497	391	1.33	(1.14-1.55)	0.0002
German replication	0.65	184	607	621	0.63	204	701	574	1.13	(1.02-1.26)	0.0203
Meta									1.20	(1.11-1.29)	$3.66 X 10^{-6}$
Combined meta									**1.21**	**(1.15-1.28)**	**$1.38X10^{-11}$** (P_{het}= 0.48, I^2= 0%)
rs4762284 (12q23.1, *ELK3*)	**RAF**	**AA**	**AT**	**TT**	**RAF**	**AA**	**AT**	**TT**	**OR[a]**	**95% CI[b]**	***P*-value**
UK GWAS	0.33	373	358	94	0.30	2,578	2,160	462	1.18	(1.05-1.32)	0.0046
German GWAS	0.33	372	366	95	0.28	1,072	773	179	1.30	(1.14-1.49)	$1.01X10^{-4}$
Meta									1.23	(1.13-1.34)	$2.87X10^{-6}$
UK replication 1	0.31	258	227	52	0.30	497	408	98	1.03	(0.88-1.21)	0.6870
UK replication 2	0.34	232	287	52	0.29	534	434	85	1.30	(1.11-1.53)	0.0008
German replication	0.31	650	625	124	0.28	746	619	109	1.15	(1.02-1.29)	0.0178
Meta									1.16	(1.07-1.25)	$3.85X10^{-4}$
Combined meta									**1.19**	**(1.12-1.26)**	**$8.41X10^{-9}$** (P_{het}= 0.17, I^2= 38%)