Common 2q37.3, 8q24.21, 15q21.3, and 16q24.1 variants influence chronic lymphocytic leukemia risk

Dalemari Crowther-Swanepoel[1*], Peter Broderick[1*], Maria Chiara Di Bernardo[1*], Sara E Dobbins[1], María Torres[2], Mahmoud Mansouri[3], Clara Ruiz-Ponte[2], Anna Enjuanes[4], Richard Rosenquist[3], Angel Carracedo[2], Jesper Jurlander[5], Elias Campo[4], Gunnar Juliusson[6], Emilio Montserrat[7], Karin E Smedby[8], Martin JS Dyer[9], Estella Matutes[10], Claire Dearden[10], Nicola J Sunter[11], Andrew G Hall[11], Tryfonia Mainou-Fowler[12], Graham H Jackson[13], Geoffrey Summerfield[14], Robert J Harris[15], Andrew R Pettitt[15], David J Allsup[16], James R Bailey[17], Guy Pratt[18], Chris Pepper[19], Chris Fegan[20], Anton Parker[21], David Oscier[21], James M Allan[11], Daniel Catovsky[10], Richard S Houlston[1¥]

1. Section of Cancer Genetics, Institute of Cancer Research, Sutton, Surrey. UK
2. Genomic Medicine Group, University of Santiago de Compostela and Galician Foundation of Genomic Medicine, CIBERER, Santiago de Compostela, Spain
3. Department of Genetics and Pathology, Uppsala University, Uppsala, Sweden
4. Hematopathology Unit, Center for Biomedical Diagnosis Hospital Clinic, University of Barcelona, Spain
5. Department of Hematology, Leukemia Laboratory, Rigshospitalet, Copenhagen, Denmark
6. Lund Strategic Research Center for Stem Cell Biology and Cell Therapy, Hematology and Transplantation, Lund University, Lund, Sweden
7. Department of Hematology, Institut d'Investigacions Biomediques August Pi i Sunyer, Hospital Clinic, University of Barcelona, Spain
8. Unit of Clinical Epidemiology, Dept of Medicine, Karolinska Institutet, Stockholm, Sweden
9. MRC Toxicology Unit, Leicester University, Leicester. UK
10. Section of Haemato-oncology, Institute of Cancer Research, Sutton, Surrey. UK
11. Northern Institute for Cancer Research, Paul O'Gorman Building, Newcastle University, Newcastle-upon-Tyne. UK
12. Haematological sciences, Leech Building, The Medical School, Newcastle University, Newcastle-upon-Tyne. UK
13. Department of Haematology, Royal Victoria Infirmary, Newcastle-upon-Tyne. UK
14. Department of Haematology, Queen Elizabeth Hospital, Gateshead, Newcastle-upon-Tyne. UK
15. Division of Haematology, University of Liverpool School of Cancer Studies, Liverpool, UK.
16. Department of Haematology, Hull Royal Infirmary, Hull. UK
17. Hull York Medical School and University of Hull, Hull. UK
18. Department of Haematology, Birmingham Heartlands Hospital, Birmingham. UK

19. Department of Haematology, School of Medicine, Cardiff University. Cardiff. UK
20. Cardiff and Vale NHS Trust, Heath Park, Cardiff. UK
21. Royal Bournemouth Hospital, Bournemouth, UK.

*Joint authors at this position

¥Corresponding author
Richard Houlston
Institute of Cancer Research
15 Cotswold Rd, Sutton, Surrey SM2 5NG, UK
Tel: +44-(0)-208-722-4175; Fax: +44-(0)-208-722-4359; e-mail: richard.houlston@icr.ac.uk

To identify novel risk variants for chronic lymphocytic leukemia (CLL) we conducted a genome-wide association study of 299,983 tagging SNPs, with validation in four additional series totaling 2,503 cases and 5,789 controls. We identified four risk loci for CLL at 2q37.3 (rs757978, *FARP2*; odds ratio [OR] = 1.39; *P* = 2.11 x 10^{-9}), 8q24.21 (rs2456449; OR = 1.26; *P* = 7.84 x 10^{-10}), 15q21.3 (rs7169431; OR = 1.36; *P* = 4.74 x 10^{-7}) and 16q24.1 (rs305061; OR = 1.22; *P* = 3.60 x 10^{-7}). There was also evidence for risk loci at 15q25.2 (rs783540, *CPEB1*; OR = 1.18; *P* = 3.67 x 10^{-6}) and 18q21.1 (rs1036935; OR = 1.22; *P* = 2.28 x 10^{-6}). These data provide further evidence for genetic susceptibility to this B-cell hematological malignancy.

B-cell chronic lymphocytic leukemia (CLL; MIM 151400) is the most common lymphoid malignancy in Western countries[1]. While the seven-fold increased risk of CLL in first-degree relatives of CLL patients provides strong evidence for inherited susceptibility[2], the genetic basis of predisposition is largely unknown.

We have previously reported the results of a genome-wide association (GWA) study based on the analysis of 299,983 tagging single nucleotide polymorphisms (SNPs) in 505 CLL cases and 1,438 control individuals (henceforth referred to as Stage 1)[3] and fast tracking analysis of the smallest *P*-values in this GWA study. Through this we identified SNPs mapping to 2q13 (rs17483466), 2q37.1 (*SP140*; rs13397985), 6p25.3 (*IRF4*; rs872071), 11q24.1 (rs735665), 15q23 (rs7176508), and 19q13.32 (*PRKD2*; rs11083846) that confer a modest increase in risk[3]. Given the success of this GWA study we have conducted a further follow-up study and report four newly identified susceptibility loci for CLL.

In Stage 2, we attempted to genotype 180 SNPs in 540 UK CLL cases - UK-replication series 1. The SNPs were chosen by a hypothesis-free (agnostic) strategy on the basis of *P*-values from the Armitage trend test, excluding those correlated with an $r^2>0.8$ with previously identified association signals. After imposing stringent quality control metrics 162 SNP genotypes were recovered for 519 of the cases. We used publicly accessible data on 2,695 UK blood donor controls to compare genotype frequencies. In Stage 3, we genotyped the 19 SNPs that showed the strongest association from combined analysis of Stages 1 and 2 in two case-control series: UK-replication series 2 (660 cases, 809 controls), Spanish-replication series (424 cases, 450 controls). In Stage 4 we genotyped the 10 SNPs which displayed the strongest association from a combined analysis of Stages 1-3 in the Swedish-replication series (395 cases, 397 controls).

The combined joint analysis of these data provided conclusive evidence of an association between seven SNPs and CLL on the basis of the conventionally accepted threshold for genome-wide significance (i.e. $P < 5.0 \times 10^{-7}$; Supplementary Table 1). These SNPs map to five independent genomic regions. One of the SNPs, rs11668878, maps to 19q13.32, a region we have previously reported to be a risk locus for CLL (defined by rs11083846)[3]. Linkage disequilibrium (LD) exists between rs11668878 and rs11083846 ($r^2=0.27$, $D'=1.0$) and conditional analysis did not provide evidence for a second disease locus ($P > 0.05$).

Under a fixed effects model the strongest statistical evidence for a novel association was attained with rs2456449 which maps to 8q24.21 at 128,262,163bp (OR = 1.26; 95% confidence interval [CI]: 1.17-1.35; $P = 7.84 \times 10^{-10}$; Figure 1, Supplementary Table 1). The association between rs2456449 and CLL risk was consistent between case-control series (Figure 1;

P_{het}=0.95, I^2=0%). rs2466024, localizing to 128,257,201bp, also provided statistically significant evidence for the 8q24.21 association (P = 4.61 x 10^{-7}; Supplementary Table 1). rs2456449 and rs2466024 are in strong LD (r^2=0.75, D'=1.0) and map to a 40kb region of LD (128,241,868-128,282,415bp defined by recombination hotspots and r^2 metrics; Figure 2, Supplementary Figure 1). Conditional analysis provided no evidence for an independent role of rs2466024, with the rs2456449 genotype sufficient to capture the locus variation.

Genome-wide association studies of cancer have shown that the 128-130Mb genomic interval at 8q24.21 harbors multiple independent loci with different tumor specificities: prostate (rs16901979; 128,194,098bp)[4], breast (rs13281615; 128,424,800bp)[5], colorectal-prostate (rs6983267; 128,482,487bp)[6,7], prostate (rs1447295; 128,554,220bp)[8] and bladder (rs9642880; 128,787,250bp)[9] cancer. The LD blocks defining these specific loci however appear distinct from the 8q24.21 CLL association signal. This is reflected in the LD metrics between rs2456449, and rs6983267 (r^2=0.00, D'=0.13), rs13281615 (r^2=0.00, D'=0.01), rs16901979 (r^2=0.01, D'=1.0), rs1447295 (r^2=0.04, D'=1.0), rs9642880 (r^2= 0.02, D'=0.19), which provide little evidence for correlation between loci. The 8q24.21 region to which the cancer associations map is bereft of genes and predicted transcripts, hence the causal basis of associations is likely to be indirect. rs6983267, defining the colorectal-prostate cancer locus, has recently been shown to affect TCF4 binding to an enhancer for the *MYC* (avian myelocytomatosis viral oncogene homolog; MIM 190080) promoter, providing a mechanistic basis for this 8q24.21 association[10,11]. It is possible that the biological basis of the other 8q24.21 cancer risk loci is via *MYC* through similar long range cis-acting mechanisms. We have previously shown that variation in *IRF4* influences the risk of CLL[3]. If the 8q24.21 locus influences CLL risk through differential *MYC* expression then the association is especially intriguing as *MYC* is a direct target of *IRF4* in activated B-cells[12].

The second strongest statistical evidence for an association signal was provided by rs757978, which maps to exon 9 of *FARP2* (alias FERM, RhoGEF and pleckstrin domain protein 2) at 2q37.3 (242,019,774bp). The overall estimate of effect associated with rs757978 was an OR of 1.39 (95% CI: 1.25–1.56; P = 2.11 x 10^{-9}; P_{het} =0.13, I^2=43%; Figure 1, Supplementary Figure 1, Supplementary Table 1). rs11681497, which maps to intron 4 of *FARP2* (241,993,006bp) and is in LD with rs757978 (r^2=1, D'=1), also provided strong evidence for the 2q37.3 association (P = 4.53 x 10^{-9}; Figure 2), however conditional analysis provided no evidence for an independent role. FARP2 is a member of Cdc42-GEF family of proteins which are involved in signaling downstream of G protein-coupled receptors[13]. We examined if there was a relationship between genotype and *FARP2* expression in Epstein Barr Virus (EBV) transformed lymphocytes which might account for the 2q37.3 association. No data was available for rs757978 but no association between rs11681497 genotype and mRNA expression level was shown (Supplementary Figure

2). rs757978 leads to the substitution of threonine for isoleucine at amino acid 260 (T260I) in the expressed protein. As *in silico* analysis using both SIFT and PolyPhen programs predicts T260I to be functionally deleterious it is possible that the association signal at 2q37.3 is a direct consequence of T260I.

The third strongest association was provided by rs305061 (OR = 1.22; 95% CI: 1.12-1.32; P = 3.60 x 10^{-7}; P_{het} =0.38, I^2=5%; Figure 1, Supplementary Table 1), which maps within a 30kb region of LD at 16q24.1 (84,533,160bps; Figure 2). rs305061 localizes 19kb telomeric to the *IRF8* (Interferon regulatory factor 8; MIM 601565) gene. Variation in *IRF8* represents a strong candidate for the basis of the observed association as IRF8 is one of several transcription factors which regulates response to α and β interferons by binding the interferon stimulated response element. Moreover, *IRF8* is involved in B-cell lineage specification, immunoglobulin light chain gene rearrangement, the distribution of mature B-cells into the marginal zone and follicular B-cell compartments, as well as regulation of germinal center reaction[14]. Variation at 16q24.1, defined by rs17445836 which maps 61kb telomeric to *IRF8* (84,575,164bps), has recently been shown to be a risk locus for multiple sclerosis (MS)[15]. MS has been reported to be associated with CLL risk in some studies[16,17]. While rs17445836 and rs305061 are not strongly correlated (r^2=0.13, D'=0.82), it is possible that a common genetic basis to both diseases is mediated through the same causal variant at 16q24.1.

The fourth strongest association was identified with rs7169431 which maps to 15q21.3 (54,128,188bps; OR = 1.36; 95% CI: 1.21-1.53; P = 4.74 x 10^{-7}; P_{het} = 0.51, I^2 = 0%; Figure 1, Supplementary Table 1). rs7169431 localizes to a 25kb region of LD flanked by *NEDD4* (neural precursor cell expressed, developmentally downregulated 4; MIM 602278), an E3 (c-Cbl) ubiquitin ligase, and *RFX7* (regulatory factor X, 7; MIM 612660) which is a member of the regulatory factor X family of transcription factors (Figure 2, Supplementary Figure 1). While there is currently no evidence for a direct role of *NEDD4* in CLL it represents a credible candidate gene because of its role in regulating viral latency and pathogenesis of EBV. Specifically NEDD4 regulates the latent membrane protein 2A (LMP2A) of EBV which mimics signaling induced by the B-cell receptor (BCR) altering B-cell development[18].

In addition to these four loci there was evidence for two additional disease loci at 15q25.2 (rs783540) and 18q21.1 (rs1036935), although the associations did not attain genome-wide significance (OR = 1.18; 95% CI: 1.10-1.27; P = 3.67 x 10^{-6} ; P_{het} =0.07, I^2=53% and OR = 1.22; 95% CI: 1.12-1.32; P = 2.28 x 10^{-6}, P_{het} = 0.39, I^2= 3%, respectively; Supplementary Table 1, Supplementary Figure 1). rs783540 localizes to intron 2 of the gene encoding CPEB1 (cytoplasmic polyadenylation element binding protein 1; MIM 607342) which has an established role in the regulation of cyclin B1 during embryonic cell division-differentiation. Two genes mapping centromeric to rs1036935, *CXXC1* (protein containing CXXC domain; MIM 609150) and *MBD1* (methyl-CpG-binding domain protein 1; MIM 156535) are involved in gene regulation. *MBD1* expression in EBV transformed lymphocytes was shown to correlate with risk genotype (Supplementary Figure 2). Although *MBD1* has no documented role in CLL, it has potential to affect CLL development through translational control of *MYC* via MDBP binding[19].

CLL shows male predominance and can be classified on the basis of the presence or absence of somatic hypermutations of the immunoglobulin heavy-chain variable (IGVH) genes[20,21], with mutated-CLL having a better prognosis. We assessed the relationship between age, gender, family history of CLL or another related B-cell malignancy, mutation status, and SNP genotypes by case-only logistic regression (Supplementary Table 3). None of the SNPs displayed evidence of a relationship with age or gender (based on cases from all phases). Having a family history of B-cell malignancy (based on Stages 1 and 2) was associated with a higher frequency of *FARP2* rs11681497 and rs757978 risk genotypes (P = 0.031 and 0.037 respectively), compatible with familial cases being enriched for genetic susceptibility. IGVH mutational status (based on Stages 1, 2, and 4) was associated with rs305061, with risk genotype correlating with unmutated CLL (P = 0.0002). Since rs305061 maps to *IRF8* this relationship is compatible with dysfunctional B-cell activation signaling being associated with possession of risk genotype.

To gain insight into the allelic architecture of predisposition to CLL we examined for interactive effects between 2q37.3, 8q24.21, 15q21.3, and 16q24.1 variants and the six previously identified loci at 2q13 (rs17483466), 2q37.1 (rs13397985), 6p25.3 (rs872071), 11q24.1 (rs735665), 15q23 (rs7176508) and 19q13.32 (rs11083846). The only evidence for an interaction between loci was provided by rs305061 and rs7176508 (P = 0.045), albeit non-significant after adjusting for multiple testing (Supplementary Table 2). The proportion of case and control subjects grouped according to the number of risk alleles that they carry is detailed in Figure 3. The distribution of risk alleles follows a normal distribution in both cases and controls, but with a shift toward a higher number of risk alleles in the cases. While the risks of CLL associated with the 10 variants so far identified are individually modest, the carrier frequencies of risk alleles are high in the European population and hence the loci make a major contribution to the development of CLL with the population attributable fraction ascribable to the 10 loci is approximately 87%. Moreover, the risk of CLL increases with increasing numbers of variant alleles for the 10 loci ($OR_{per\text{-}allele}$= 1.39, 95% CI: 1.35-1.44; $P = 2.73 \times 10^{-88}$; Figure 3, Supplementary Table 4). Individuals with 13+ risk alleles have a >7-fold increase in CLL risk compared to those with a median number of risk alleles. These ORs may be underestimates because the additive model on which analyses are based assumes equal weighting across the SNPs. We estimate that the 10 loci we have identified account for approximately 10% of the excess familial risk of CLL, assuming a polygenic model. It is however acknowledged that the present data only provide estimates of the effect on susceptibility attributable to variation at the loci and the effect of the actual causal variant responsible for the association is likely to be greater.

In conclusion these results provide evidence for low risk variants predisposing to B-cell CLL and insight into the development of this hematological malignancy. Furthermore, the reciprocal familial risks between CLL and other B-cell malignancies[2] raises the possibility that these variants may also influence the risk of related B-cell tumors.

URLs

Online Inheritance in Man: http://www.ncbi.nlm.nih.gov/sites/entrez
The R suite can be found at http://www.r-project.org/
Detailed information on the tag SNP panel can be found at http://www.illumina.com/
dbSNP: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=snp
HAPMAP: http://www.hapmap.org/
KBioscience: http://kbioscience.co.uk/
WGAViewer: http://www.genome.duke.edu/centers/pg2/downloads/wgaviewer.php
International immunogenetics information system: http://imgt.cines.fr
IWCLL:http://www.icr.ac.uk/research/research_sections/cancer_genetics/cancer_genetics_teams/molecular_and_population_genetics/fcll/index.shtml
The European Genome-phenome Archive (Wellcome Trust Case-Control Consortium [WTCCC]): http://www.ebi.ac.uk/ega/page.php
POLYPHEN: http://www.bork.embl-heidelberg.de/PolyPhen/
SIFT: http://blocks.fhcrc.org/sift//SIFT.html
SNAP (SNP Annotation and Proxy Search): http://www.broadinstitute.org/mpg/snap/

ACKNOWLEDGEMENTS

Leukaemia Research Fund (LRF) provided principal funding for the study (LRF05001 and 06002). Additional funding was provided by CLL Global Research Foundation, Cancer Research UK (C1298/A8362 supported by the Bobby Moore Fund), and the Arbib Fund. Dalemari Crowther-Swanepoel (DC-S) was in receipt of a PhD studentship from the Institute of Cancer Research. This study makes use of data generated by the Wellcome Trust Case-Control Consortium 1 and 2. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113 and 085475. In the UK NHS funding for the Royal Marsden Biomedical Research Centre is acknowledged. EC and AE acknowledge CICYT funding (SAF 08/3630) and the RTICC from the Insituto de Salud Carlos III. AC and MT acknowledge Xunta de Galicia funding for Spanish control population (PGIDIT 08CSA005208PR) and the support of the Spanish National Genotyping Center (USC Node). We would also like to thank Dr Leonid Padyukov, Karolinska Institutet, Stockholm, Sweden, and the EIRA group for providing control samples from the Swedish population for the Swedish replication study. We are grateful to all clinicians for their involvement in patient ascertainment. Finally we are grateful to all the patients and other individuals for their participation.

AUTHOR CONTRIBUTIONS

RSH designed, obtained funding, directed the study, and oversaw analyses. RSH drafted the manuscript, with help from DC-S, PB, MCDB and SED. MCDB performed statistical analyses. DC-S, MCDB, SED performed bioinformatics analyses. RSH and DC established the parent study. RSH, DC and DC-S developed patient recruitment, sample acquisition and performed sample collection of cases. For the GWA and UK-replication series 1, DC-S and PB supervised laboratory management and oversaw genotyping of cases. DC-S conducted sequencing. JMA and DJA conceived of the Newcastle-based CLL study. JMA established the study, supervised laboratory management and oversaw genotyping of cases and controls. NJS performed sample management of cases and controls. AGH developed the Newcastle Haematology Biobank, incorporating the Newcastle-based CLL study. TM-F, GHJ, GS, RJH, ARP, DO, DJA, JRB, GP, CP, and CF developed patient recruitment, sample acquisition and performed sample collection of cases. Coordinated of the Spain replication series was conducted by EC and AE. EC and AE provided CLL samples and AC controls and compiled detailed phenotypic information from cases and controls respectively. Genotyping was performed by MT and AC and MT supervised laboratory management and quality control. For the Swedish case-control study MM performed sample collection and prepared DNA. RR performed sample collection for all cases, while JJ, GJ and KES performed sample collection of cases in the SCALE study. PB and DC-S performed genotyping of cases and controls.

All authors contributed to the final paper.

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

Note: Supplementary information is available on the Nature Genetics website

METHODS

Initial genotyping for genome-wide scan

This study describes the follow-up genotyping of a previously reported genome-wide scan of CLL[3]. Briefly 505 cases of CLL were genotyped using HumanCNV370-Duo BeadChips (Illumina, San Diego, USA). For controls we made use of publicly accessible HumanHap550K BeadChip genotype data generated on 1,438 individuals from the WTCCC British 1958 Birth Cohort. Quality control metrics included removal of samples with call rates under 90% and SNP assays with call rates under 95%. Subjects with evidence of cryptic relatedness and non-European background were excluded from the analysis. We considered only the 299,983 autosomal SNPs with minor allele frequencies (MAF) exceeding 1% in both cases and controls, and with no extreme evidence of departure from Hardy-Weinberg equilibrium (HWE; $P > 10^{-5}$) in cases or controls. Comparison of observed and expected distributions showed little evidence for an inflation of test statistics (inflation factor λ=1.053). Full details on the scan were reported previously[3].

Replication samples

In Stage 2 (**UK-replication series 1**) we genotyped SNPs in 540 CLL cases (353 males, 187 females; mean age at diagnosis 60.3 years; SD ± 12.2) ascertained through the Royal Marsden NHS Hospitals Trust between 1998 and 2006. All cases were UK residents and were self reported to be of European ancestry. Genotyping was conducted using the Sequenom MassARRAY system (http://www.sequenom.com/). For controls we made use of publicly accessible genotype data generated on 2,736 UK blood donor controls. We excluded SNPs on the basis of deviation from HWE using a threshold of $P < 1.0 \times 10^{-5}$ in either the cases or controls. We also removed SNPs with MAF <0.05. To identify and exclude individuals with non-Western European ancestry, case and control data was merged with individuals of different ethnicities from HapMap, genome-wide IBS distances for markers shared between HapMap and our SNP panel were determined, and dissimilarity measures used to perform principal component analysis. After imposing these stringent quality control measures 162 SNP genotypes were available on 519 cases and 2,695 controls and this dataset formed the basis of our analysis.

In Stage 3 we genotyped 19 SNPs showing the strongest evidence of an association from joint analysis of Stages 1 and 2 in two independent case-control studies. **UK-replication series 2**: Cases were ascertained through the Newcastle CLL Consortium - 660 United Kingdom Caucasian patients with CLL originally diagnosed between 1970 and 2006 who attended seven hematology units in the United Kingdom (Newcastle Royal Victoria Infirmary, Gateshead Queen Elizabeth Hospital, Birmingham Heartlands Hospital, Hull Royal Infirmary, Liverpool Royal

University Hospital, Royal Bournemouth Hospital and The University of Wales College of Medicine Hospital). Peripheral blood for DNA extraction was taken between 1998 and 2006 for all cases (Hull 1993-1996; Birmingham and Cardiff 1998-2006; Newcastle and Gateshead 1998-2005; Liverpool 2000-2006; Bournemouth 1995-2007). Controls comprised 809 UK Caucasian healthy individuals aged between 16 and 69 (mean age 48.2 years, SD ± 14.3, 509 males, 443 females) recruited to a study of acute leukemia conducted between April 1991 and December 1996, previously described[22]. **Spanish-replication series**: The Spanish replication series involved 445 cases and 450 controls. All cases were Spanish residents and were self reported to be of European ancestry. Controls were obtained from the MedXen Control population cohort (Xunta de Galicia) and were selected to match cases according to geographic origin, age, and sex.

In Stage 4 we genotyped 10 SNPs, showing the strongest evidence of an association from joint analysis of Stages 1-3 in the **Swedish-replication series**: 304 samples from the Swedish part of a population-based case-control study, called SCALE (Scandinavian Lymphoma Etiology)[23] and 91 samples from the biobank at the Department of Pathology, Uppsala University Hospital, Uppsala, Sweden, were included for analysis. For the SCALE samples, peripheral blood was collected from 275 cases during 1999-2001, within a median of 4 months from diagnosis (range, 0-29 months) while in 29 cases, follow-up samples ascertained in 2007-2008 were used. The samples received from the biobank at Uppsala University Hospital were collected between 1982-2005. All samples were diagnosed according to the recently updated criteria[24] and displayed the characteristic CLL immunophenotype. The Swedish replication series totaled 148 females and 247 males and had an average age at diagnosis of 62.6 years. Swedish controls were collected at Karolinska Institutet, Stockholm, Sweden and included 397 healthy individuals aged between 33 and 71 (mean age 61.3, SD± 6.8, 251 males, 146 females).

In all studies, for both incident and prevalent cases, the diagnosis of CLL has been confirmed in accordance with the current WHO classification guidelines[25].

Ethics

Collection of blood samples and clinico-pathologiocal information from patients and controls was undertaken with informed consent and relevant institutional ethical review board approval in accordance with the tenets of the Declaration of Helsinki.

Replication series genotyping

DNA was extracted from EDTA venous blood samples using standard methodologies and Picogreen quantified (Invitrogen, Paisley, United Kingdom). To ensure quality control of genotyping, a series of duplicate samples were genotyped in the same batches. For all SNP assays >99% concordant results were obtained. Genotyping in UK-replication 1 was performed by Sequenom MassARRAY system (http://www.sequenom.com/). We attempted to type 180 SNPs, although it was not possible to design functional assays for 18 SNPs with this technology. Details of the methodology are available on request. To exclude technical artifacts of genotyping, we included a random series of 24 samples previously genotyped using Illumina 550K HapMap arrays (Illumina, San Diego, USA). We successfully genotyped 519 unique subjects passing quality control metrics, excluding validation samples, and study duplicates. Genotyping of UK-replication 2 and the Swedish-replication series were conducted by competitive allele-specific PCR KASPar chemistry (KBiosciences Ltd, Hertfordshire, UK; http://www.kbioscience.co.uk/); primer sequences and conditions available on request. Excluding validation samples and study duplicates, passing quality control metrics we genotyped 1,469 unique subjects in UK-replication series 2 and 792 unique subjects in the Swedish replication series. Genotyping of Spanish samples was conducted using the Sequenom MassARRAY system (http://www.sequenom.com/) in two i-Plex assays. Excluding validation samples and study duplicates, passing quality control metrics we genotyped 874 unique subjects. Mouthwash or buccal DNA samples were available from 89 of the cases from Stage 1 and Stage 2. In these samples, SNP genotypes were 99% concordant with genotypes obtained from typing of blood DNA samples. Together with the fact that the associations identified do not map to any of the regions of the genome commonly associated with copy number variation in CLL[26], these results argue against bias from differential genotyping as a consequence of allelic imbalance influencing study findings.

IGVH mutational status

IGVH mutational status was determined according to BIOMED-2 protocols as described previously[27,28] or commercial reagents (*InVivo*Scribe Technologies, San Diego, USA). Clonality was assessed by size discrimination of PCR products using semi-automated ABI3730xl/ABI3700 sequencers in conjunction with Genescan software (Applied Biosystems, Foster City, USA). Sequences obtained were submitted to online database IMGT/V-QUEST[29]. In accordance with published criteria, sequences with a germline identity of ≥ 98% were classified as unmutated, and those displaying identity <98% as mutated.

Statistical analysis

Statistical analyses were undertaken using R (v2.3.1) and STATA (v10.0) Software. Deviation of the genotype frequencies in the controls from those expected under HWE was assessed by the χ^2 test (1 degree of freedom). The association between each SNP and risk of CLL was assessed by the Cochran-Armitage trend test. Odds ratios (ORs) and associated 95% confidence intervals (CIs) were calculated by unconditional logistic regression. Relationships between multiple SNPs showing association with CLL risk in the same region were investigated using logistic regression analysis and the impact of additional SNPs from the same region was assessed by a likelihood-ratio test.

Associations by gender, age and mutation status were examined by logistic regression in case-only analyses. The combined effect of pairs of loci identified with risk was investigated by logistic regression modeling; evidence for interactive effects between SNPs assessed by a likelihood ratio test. The OR and trend test for increasing numbers of deleterious alleles was estimated by counting two for a homozygote and one for a heterozygote assuming equal weights.

Meta-analysis was conducted using standard methods[30]. Cochran's Q statistic to test for heterogeneity[30] and the I^2 statistic[31] to quantify the proportion of the total variation due to heterogeneity were calculated. Large heterogeneity is typically defined as I^2 ≥75%. The population attributable fraction was estimated from $1-\prod_i 1-(x_i-1)/x_i$ where $x_i = (1-p)^2 + 2p(1-p)OR_1 + p^2OR_2$, p is the population allele frequency, and OR_1 and OR_2 are the ORs associated with hetero- and homozygosity respectively. The sibling relative risk attributable to a given SNP was calculated using the formula[32]:

$$\lambda^* = \frac{p(pr_2 + qr_1)^2 + q(pr_1 + q)^2}{[p^2r_2 + 2pqr_1 + q^2]^2}$$

where p is the population frequency of the minor allele, q=1-p, and r_1 and r_2 are the relative risks (estimated as OR) for heterozygotes and rare homozygotes, relative to common homozygotes. Assuming a multiplicative interaction the proportion of the familial risk attributable to a SNP was calculated as $\log(\lambda^*)/\log(\lambda_0)$, where λ_0 is the overall familial relative risk estimated from epidemiological studies, assumed to be 7.25[2].

Bioinformatics

We used Haploview software (v3.2) to infer the LD structure of the genome in the regions containing loci associated with disease risk. We applied two *in silico* algorithms, PolyPhen and SIFT, to predict the impact of non-synonymous SNPs on protein function.

Relationship between SNP genotype and expression levels

To examine for a relationship between SNP genotype and expression levels of *FARP2, NEDD4, MBD1,* and *CPEB1* in lymphocytes we made use of publicly available expression data generated from analysis of 90 Caucasian derived Epstein-Barr virus–transformed lymphoblastoid cell lines using Sentrix Human-6 Expression BeadChips (Illumina, San Diego, USA)[33,34]. Online recovery of data was performed using WGAViewer Version 1.25 Software. Differences in the distribution of levels of mRNA expression between SNP genotypes were compared using a Wilcoxon-type test for trend[35].

FIGURE LEGENDS

Figure 1: Forest plots of effect size and direction for the SNPs associated with CLL risk. (a) 2q37.3 (rs757978), (b) 8q24.21 (rs2456449), (c) 15q21.3 (rs7169431), (d) 16q24.1 (rs305061). Boxes denote OR point estimates, their areas being proportional to the inverse variance weight of the estimate. Horizontal lines represent 95% confidence intervals. The diamond (and broken line) represents the summary OR computed under a fixed effects model, with 95% confidence interval given by the width of the diamond. The unbroken vertical line is at the null value (OR = 1.0).

Figure 2: Four previously unidentified loci, 2q37.3, 8q24.21, 15q21.3, and 16q24.1 showing genome-wide level of evidence of association to CLL.

(a) Illustration of the **2q37.3** locus, with the local recombination rate plotted in light blue over this 600-kb chromosomal segment centered on **rs757978**. Each square represents a SNP found in this locus and the most associated SNP in the combined analysis, **rs757978**, is marked by a blue diamond. The color intensity of each square reflects the extent of LD with rs757978 - red ($r^2 > 0.8$) through to white ($r^2 < 0.3$). Physical positions are based on build 36 of the human genome. **rs757978** is located in exon 9 of *FARP2*.

(**b**) Illustration of the **8q24.21** locus, with the most associated SNP in this locus, **rs2456449**, highlighted by a blue diamond. Here, we also present all SNPs found within a 600-kb window centered on **rs2456449** and define SNP colors based on LD with **rs2456449**.

(**c**) Illustration of the **15q21.3** locus, with the most associated SNP in this locus, **rs7169431**, highlighted by a blue diamond. Here, we also present all SNPs found within a 600-kb window centered on **rs7169431** and define SNP colors based on LD with **rs7169431**. *NEDD4* and *RFXDC2* map centromeric and telomeric to **rs7169431**.

(d) Illustration of the **16q24.1** locus, with the most associated SNP in this locus, **rs305061**, highlighted by a blue diamond. Here, we also present all SNPs found within a 600-kb window centered on **rs305061** and define SNP colors based on LD with **rs305061**. In this case, *IRF8* is the only gene found in the vicinity to the association signal.

Linkage disequilibrium maps are presented for all four loci in Supplementary Figure 1a–d online.

Figure 3: (a) Distribution of risk alleles in controls (blue bars) and CLL cases (red bars) for the 10 loci (rs757978, rs2456449, rs7169431 and rs305061 and the six previously identified loci[3] - rs17483466, rs13397985, rs872071, rs735665, rs7176508, and rs11083846)**; (b) Plot of the increasing ORs for CLL with increasing number of risk alleles.** The ORs are relative to the median number of 7 risk alleles; Vertical bars correspond to 95% confidence intervals. The distribution of risk alleles follows a normal distribution in both case and controls, with a shift

towards a higher number of risk alleles in cases. Analysis is based on data from Stages 1, 2 and UK-replication series 2. Horizontal line denotes the null value (OR=1.0).

SUPPLEMENTARY MATERIAL

Supplementary Table 1: Genotypes for the 19 SNPs in cases and controls in each of the stages. Also shown are Odds ratios and associated 95% confidence intervals, for each of the SNPs genotyped in each Stage.

Supplementary Table 2: Pairwise analysis of rs757978, rs2456449, rs7169431, rs305061 and six previously identified loci[3]. For each row-column combination, numbers show the *P*-value and number of samples (N) the result is based on, for inclusion of an interaction term between the two SNPs. The interactions are based on data from all stages for the four new loci and on data from Stage 1, 2, and UK-replication series 2 for the six previously identified loci.

Supplementary Table 3: Clinico-pathological association testing. Age and gender are based on data from all stages, mutational status is based on data from Stages 1, 2 and 4, and family history is based on Stage 1 and 2 only.

Supplementary Table 4: Odds ratios corresponding to increasing number of risk alleles in rs757978, rs2456449, rs7169431 and rs305061 and the six previously identified loci. The analysis is based on data from Stages 1, 2, and UK-replication series 2.

Supplementary Figure 1: Regional plots of the four confirmed associations at 2q37.3 (a), 8q24.21 (b), 15q21.3 (c), 16q24.1 (d), and the associations at 15q25.2 (e) and 18q21.1 (f). The upper panel shows single marker association statistics (as $-\log_{10}$ values) as a function of genomic position (NCBI build 36.1), with black dots corresponding to *P*-values from Stage 1 and red dots corresponding to combined *P*-values. Also shown are the relative position of genes mapping to each region of association. In the lower panel are the estimated statistics of the square of the correlation coefficient (r^2) for HapMap SNPs, generated using Haploview software (v3.2). The values indicate the LD relationship between each pair of SNPs; the darker the shading, the greater extent of LD. Exons of genes have been redrawn to show the relative positions in the gene, therefore maps are not to physical scale.

Supplementary Figure 2: Relationship between lymphocyte mRNA expression levels of (a) *FARP2* and rs11681497 genotype, (b) *NEDD4* and rs7169431 genotype*,* (c) *MBD1* and rs1036935 genotype and (d) *CPEB1* and rs783540 genotype*.* Data on rs757978 was not available and the relationship between variation at 2q23.3 and *FARP2* expression was assessed using rs11681497 genotype. Expression of genes is based on data from analysis of 90 Epstein-Barr virus–transformed lymphoblastoid cell lines using Sentrix Human-6 Expression BeadChip

(Illumina, San Diego, USA)[33,34]. Data was recovered using WGAViewer Version 1.25. Differences in the distribution of expression by SNP genotype were compared using a Wilcoxon-type test for trend[35].

REFERENCES

1. Stevenson, F. & Caligaris-Cappio, F. Chronic lymphocytic leukemia: revelations from the B-cell receptor. *Blood*, 4389–95 (2004).
2. Goldin, L.R., Pfeiffer, R.M., Li, X. & Hemminki, K. Familial risk of lymphoproliferative tumors in families of patients with chronic lymphocytic leukemia: results from the Swedish Family-Cancer Database. *Blood* **104**, 1850-4 (2004).
3. Di Bernardo, M.C. et al. A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia. *Nat Genet* **40**, 1204-10 (2008).
4. Gudmundsson, J. et al. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat Genet* **39**, 631-7 (2007).
5. Easton, D.F. et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087-93 (2007).
6. Tomlinson, I. et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat Genet* **39**, 984-8 (2007).
7. Yeager, M. et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet* **39**, 645-649 (2007).
8. Amundadottir, L.T. et al. A common variant associated with prostate cancer in European and African populations. *Nat Genet* **38**, 652-8 (2006).
9. Kiemeney, L.A. et al. Sequence variant on 8q24 confers susceptibility to urinary bladder cancer. *Nat Genet* **40**, 1307-12 (2008).
10. Pomerantz, M.M. et al. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat Genet* **41**, 882-4 (2009).
11. Tuupanen, S. et al. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat Genet* **41**, 885-90 (2009).
12. Shaffer, A.L. et al. IRF4 addiction in multiple myeloma. *Nature* **454**, 226-31 (2008).
13. Miyamoto, Y., Yamauchi, J. & Itoh, H. Src kinase regulates the activation of a novel FGD-1-related Cdc42 guanine nucleotide exchange factor in the signaling pathway from the endothelin A receptor to JNK. *J Biol Chem* **278**, 29890-900 (2003).
14. Wang, H. & Morse, H.C., 3rd. IRF8 regulates myeloid and B lymphoid lineage diversification. *Immunol Res* **43**, 109-17 (2009).
15. De Jager, P.L. et al. Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. *Nat Genet* **41**, 776-82 (2009).
16. Soderberg, K.C., Jonsson, F., Winqvist, O., Hagmar, L. & Feychting, M. Autoimmune diseases, asthma and risk of haematological malignancies: a nationwide case-control study in Sweden. *Eur J Cancer* **42**, 3028-33 (2006).

17. Cartwright, R.A. et al. Chronic lymphocytic leukaemia: case control epidemiological study in Yorkshire. *Br J Cancer* **56**, 79-82 (1987).

18. Ikeda, M. & Longnecker, R. The c-Cbl proto-oncoprotein downregulates EBV LMP2A signaling. *Virology* **385**, 183-91 (2009).

19. Zhang, X.Y., Supakar, P.C., Wu, K.Z., Ehrlich, K.C. & Ehrlich, M. An MDBP site in the first intron of the human c-myc gene. *Cancer Res* **50**, 6865-9 (1990).

20. Hamblin, T.J., Davis, Z., Gardiner, A., Oscier, D.G. & Stevenson, F.K. Unmutated Ig V(H) genes are associated with a more aggressive form of chronic lymphocytic leukemia. *Blood* **94**, 1848-54 (1999).

21. Damle, R.N. et al. Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. *Blood* **94**, 1840-7 (1999).

22. Allan, J.M. et al. Polymorphism in glutathione S-transferase P1 is associated with susceptibility to chemotherapy-induced leukemia. *Proc Natl Acad Sci U S A* **98**, 11592-7 (2001).

23. Smedby, K.E. et al. Ultraviolet radiation exposure and risk of malignant lymphomas. *J Natl Cancer Inst* **97**, 199-209 (2005).

24. Hallek, M. et al. Guidelines for the diagnosis and treatment of chronic lymphocytic leukemia: a report from the International Workshop on Chronic Lymphocytic Leukemia updating the National Cancer Institute-Working Group 1996 guidelines. *Blood* **111**, 5446-56 (2008).

25. Müller-Hermelink H, M.E., Catovsky D, Harris N. . Chronic lymphocytic leukaemia /small lymphocytic lymphoma. In: Jaffe ES, Harris NL, Stein H, Vardiman JW, eds. World Health Organization Classification of Tumours: Pathology and Genetics of Tumours of Haematopoietic and Lymphoid Tissues. *Lyon, France: IARC Press*, 127–130. (2001).

26. Pfeifer, D. et al. Genome-wide analysis of DNA copy number changes and LOH in CLL using high-density SNP arrays. *Blood* **109**, 1202-10 (2007).

27. van Dongen, J.J. et al. Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. *Leukemia* **17**, 2257-317 (2003).

28. van Krieken, J.H. et al. Improved reliability of lymphoma diagnostics via PCR-based clonality testing: report of the BIOMED-2 Concerted Action BHM4-CT98-3936. *Leukemia* **21**, 201-6 (2007).

29. Brochet, X., Lefranc, M.P. & Giudicelli, V. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res* **36**, W503-8 (2008).

30. Petitti, D. Meta-analysis Decision Analysis and Cost-Effectiveness Analysis. *Oxford, New York, Oxford*.(1994).

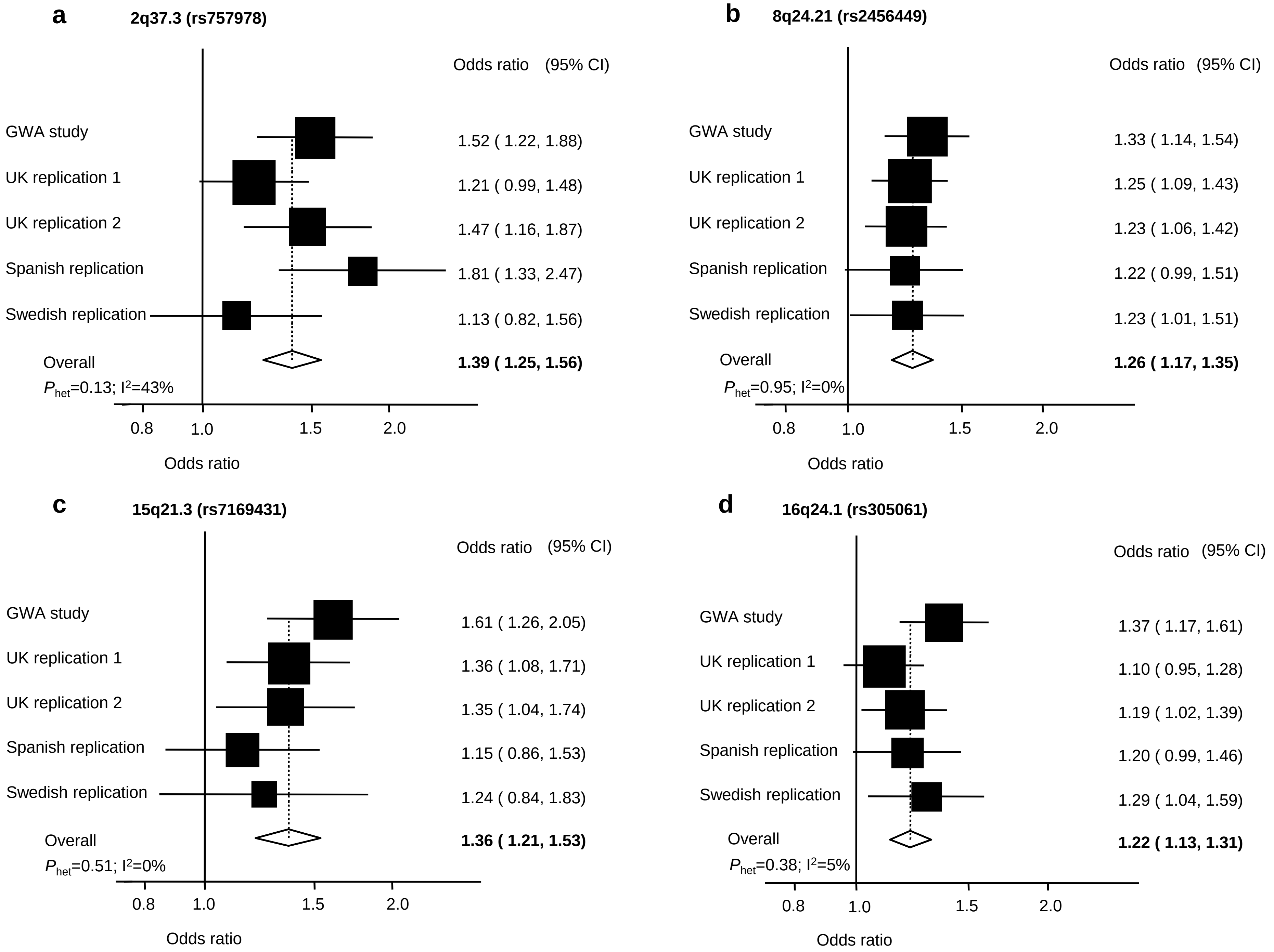
31. Higgins, J.P. & Thompson, S.G. Quantifying heterogeneity in a meta-analysis. *Stat Med* **21**, 1539-1558 (2002).

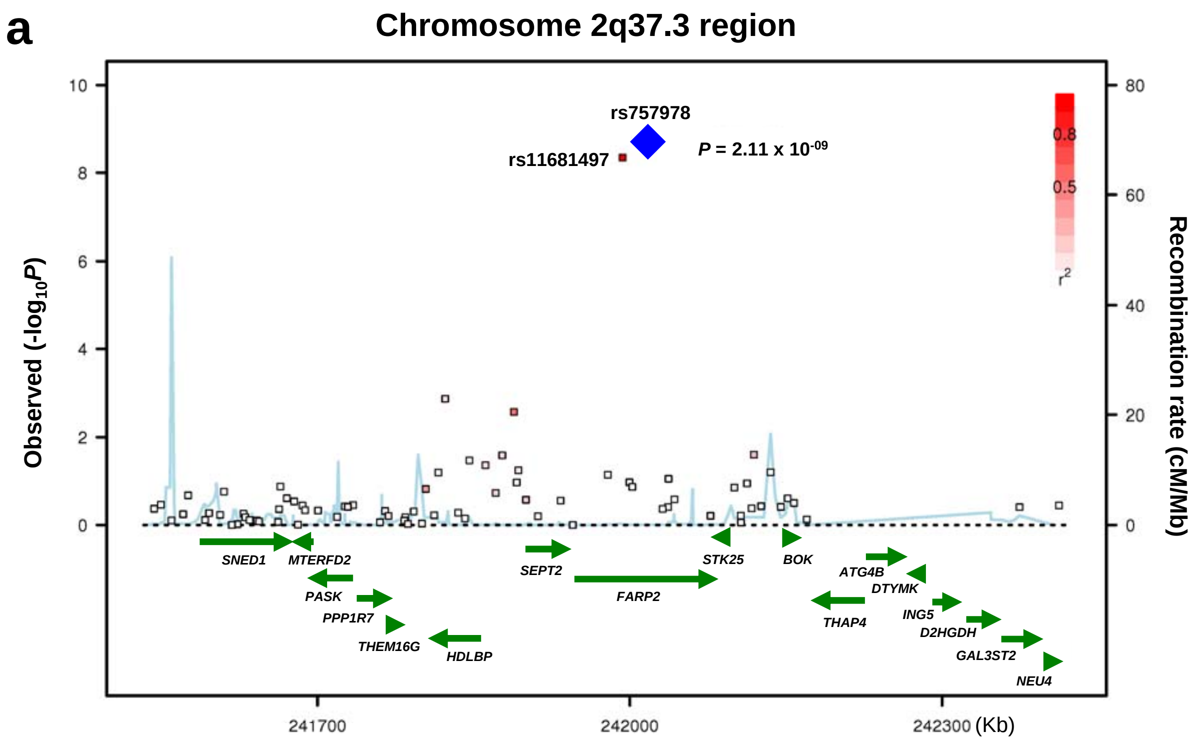
32. Houlston, R.S. & Ford, D. Genetics of coeliac disease. *QJM* **89**, 737-43 (1996).

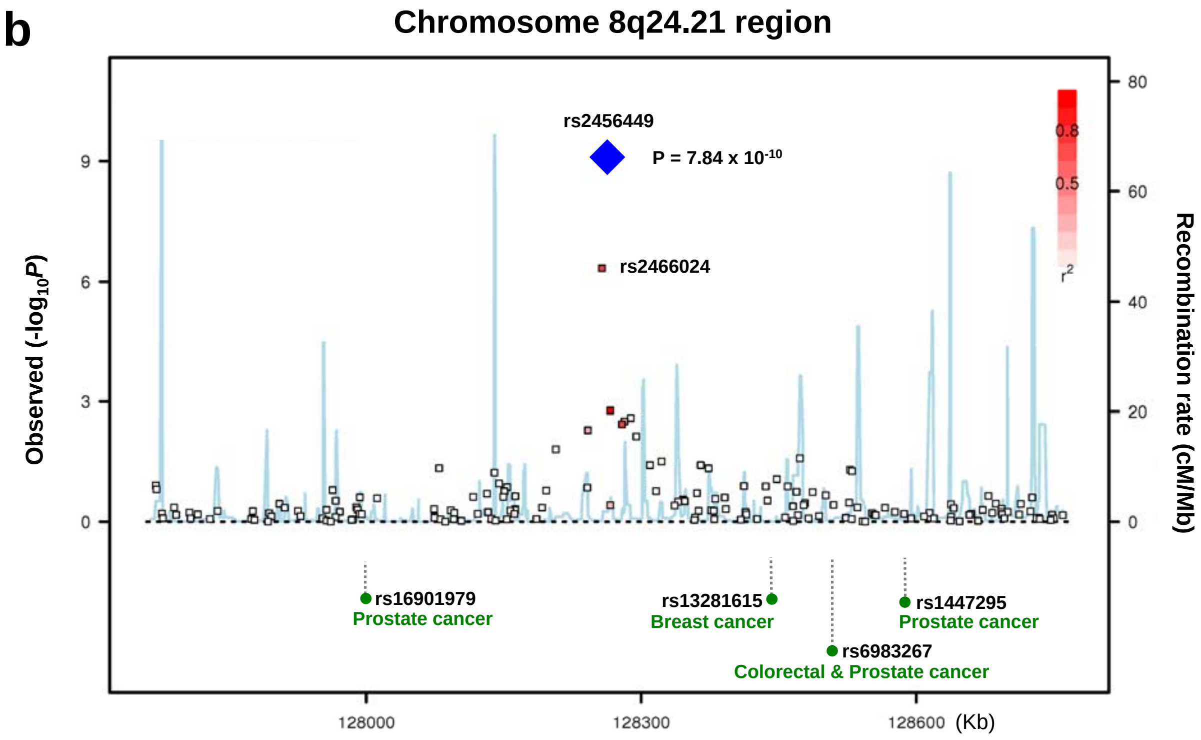
33. Stranger, B.E. et al. Genome-wide associations of gene expression variation in humans. *PLoS Genet* **1**, e78 (2005).

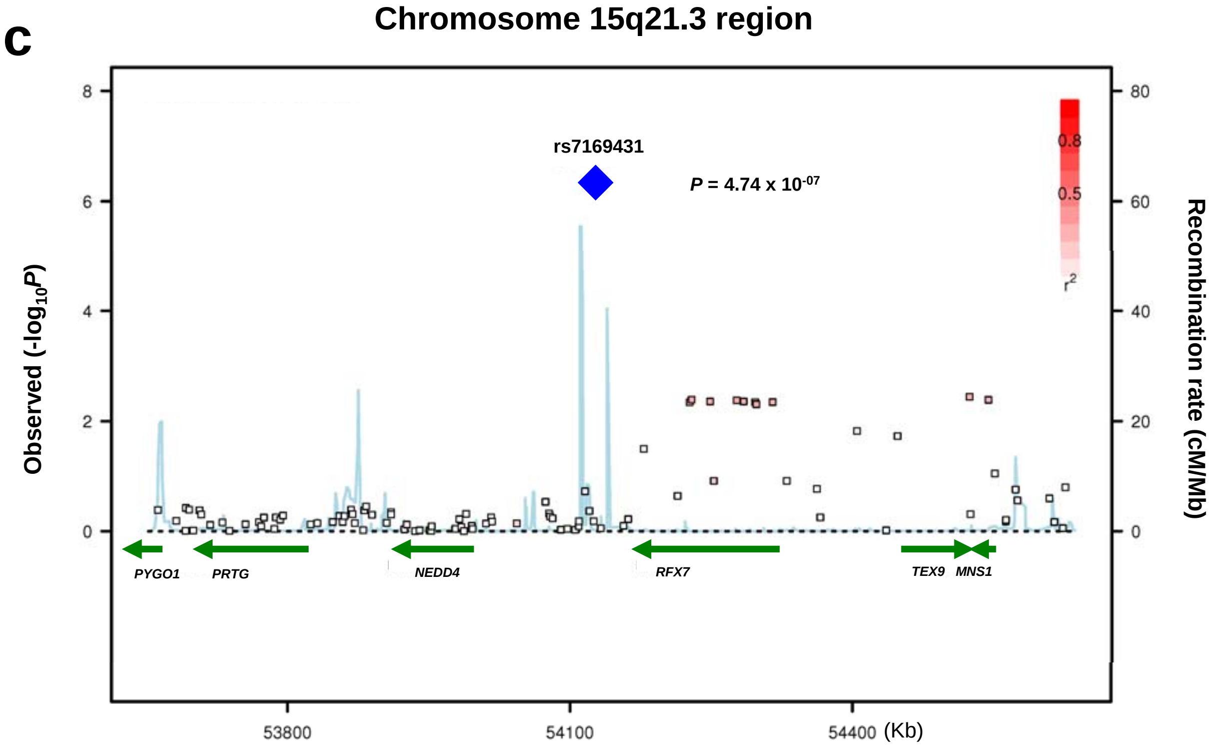
34. Stranger, B.E. et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848-53 (2007).

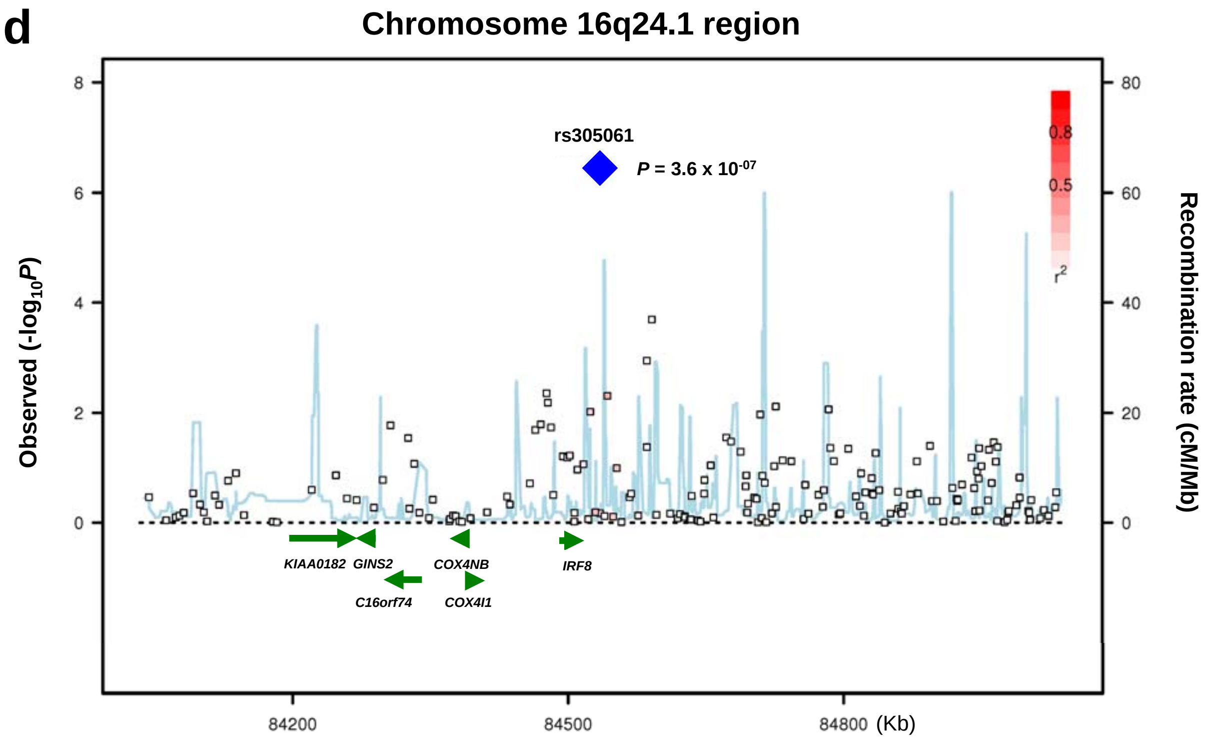
35. Cuzick, J. A Wilcoxon-type test for trend. *Stat Med.* **4**, 87-90. (1985).

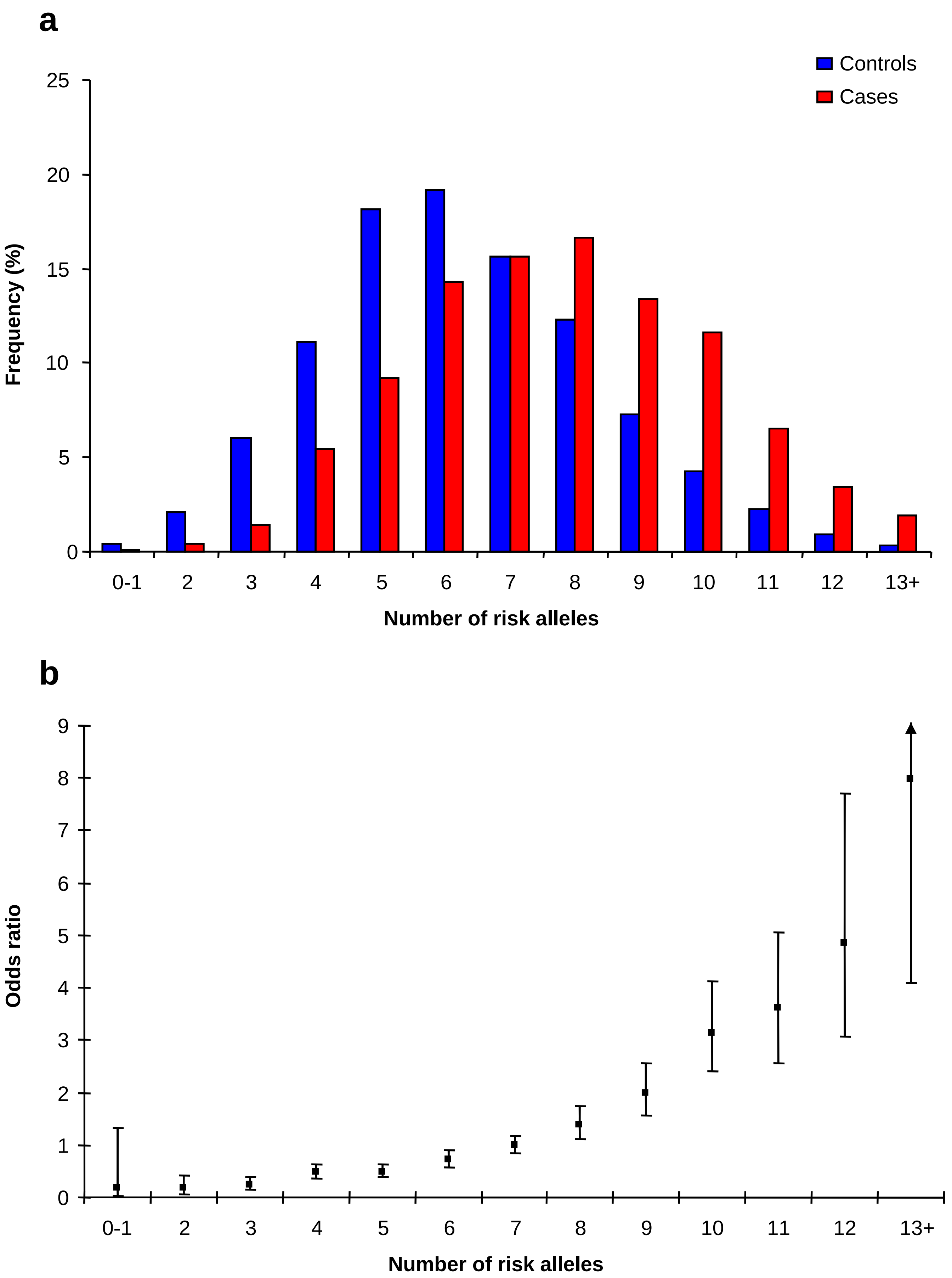

a
2q37.3 (rs757978)
Odds ratio (95% CI)
GWA study
1.52 (1.22, 1.88)
UK replication 1
1.21 (0.99, 1.48)
UK replication 2
1.47 (1.16, 1.87)
Spanish replication
1.81 (1.33, 2.47)
Swedish replication
1.13 (0.82, 1.56)
Overall
1.39 (1.25, 1.56)
P_{het}=0.13; I^2=43%
0.8
1.0
1.5
2.0
Odds ratio
b
8q24.21 (rs2456449)
Odds ratio (95% CI)
GWA study
1.33 (1.14, 1.54)
UK replication 1
1.25 (1.09, 1.43)
UK replication 2
1.23 (1.06, 1.42)
Spanish replication
1.22 (0.99, 1.51)
Swedish replication
1.23 (1.01, 1.51)
Overall
1.26 (1.17, 1.35)
P_{het}=0.95; I^2=0%
0.8
1.0
1.5
2.0
Odds ratio
c
15q21.3 (rs7169431)
Odds ratio (95% CI)
GWA study
1.61 (1.26, 2.05)
UK replication 1
1.36 (1.08, 1.71)
UK replication 2
1.35 (1.04, 1.74)
Spanish replication
1.15 (0.86, 1.53)
Swedish replication
1.24 (0.84, 1.83)
Overall
1.36 (1.21, 1.53)
P_{het}=0.51; I^2=0%
0.8
1.0
1.5
2.0
Odds ratio
d
16q24.1 (rs305061)
Odds ratio (95% CI)
GWA study
1.37 (1.17, 1.61)
UK replication 1
1.10 (0.95, 1.28)
UK replication 2
1.19 (1.02, 1.39)
Spanish replication
1.20 (0.99, 1.46)
Swedish replication
1.29 (1.04, 1.59)
Overall
1.22 (1.13, 1.31)
P_{het}=0.38; I^2=5%
0.8
1.0
1.5
2.0
Odds ratio


a
Chromosome 2q37.3 region
Observed (-log$_{10}P$)
Recombination rate (cM/Mb)
10
8
6
4
2
0
80
60
40
20
0
rs757978
rs11681497
P = 2.11 x 10^{-09}
0.8
0.5
r^2
SNED1
MTERFD2
PASK
PPP1R7
THEM16G
HDLBP
SEPT2
FARP2
STK25
BOK
ATG4B
DTYMK
THAP4
ING5
D2HGDH
GAL3ST2
NEU4
241700
242000
242300 (Kb)


b
Chromosome 8q24.21 region
rs2456449
P = 7.84 x 10^{-10}
rs2466024
0.8
0.5
r^2
Observed (-log$_{10}$$P$)
Recombination rate (cM/Mb)
9
6
3
0
80
60
40
20
0
rs16901979
Prostate cancer
rs13281615
Breast cancer
rs1447295
Prostate cancer
rs6983267
Colorectal & Prostate cancer
128000
128300
128600
(Kb)


c
Chromosome 15q21.3 region
rs7169431
P = 4.74 x 10-07
0.8
0.5
r2
Observed (-log10P)
Recombination rate (cM/Mb)
8
6
4
2
0
80
60
40
20
0
PYGO1
PRTG
NEDD4
RFX7
TEX9
MNS1
53800
54100
54400
(Kb)


d
Chromosome 16q24.1 region
rs305061
$P = 3.6 \times 10^{-07}$
0.8
0.5
r^2
Observed ($-\log_{10}P$)
Recombination rate (cM/Mb)
0
2
4
6
8
0
20
40
60
80
KIAA0182
GINS2
COX4NB
IRF8
C16orf74
COX4I1
84200
84500
84800
(Kb)


a
Controls
Cases
25
20
15
10
5
0
Frequency (%)
0-1
2
3
4
5
6
7
8
9
10
11
12
13+
Number of risk alleles
b
9
8
7
6
5
4
3
2
1
0
Odds ratio
0-1
2
3
4
5
6
7
8
9
10
11
12
13+
Number of risk alleles