A genome-wide association study identifies multiple susceptibility loci for chronic lymphocytic leukemia

A genome-wide association study identifies multiple susceptibility loci for chronic lymphocytic leukemia

Helen E Speedy[1*], Maria Chiara Di Bernardo[1*], Georgina P Sava[1], Martin J S Dyer[2], Amy Holroyd[1], Yufei Wang[1], Nicola J Sunter[3], Larry Mansouri[4], Gunnar Juliusson[5], Karin E Smedby[6], Göran Roos[7], Sandrine Jayne[8], Aneela Majid[8], Claire Dearden[9], Andrew G Hall[3], Tryfonia Mainou-Fowler[10], Graham H Jackson[11], Geoffrey Summerfield[12], Robert J Harris[13], Andrew R Pettitt[13], David J Allsup[14], James R Bailey[15], Guy Pratt[16], Chris Pepper[17], Chris Fegan[18], Richard Rosenquist[4], Daniel Catovsky[9], James M Allan[3], Richard S Houlston[1+]

1. Division of Genetics and Epidemiology, Institute of Cancer Research, Sutton, Surrey, UK
2. Department of Cancer Studies and Molecular Medicine, University of Leicester, Leicester, UK
3. Northern Institute for Cancer Research, Newcastle University, Newcastle upon Tyne, UK
4. Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden.
5. Lund Strategic Research Center for Stem Cell Biology and Cell Therapy, Hematology and Transplantation, Lund University, Lund, Sweden
6. Unit of Clinical Epidemiology, Department of Medicine, Karolinska Institutet, Stockholm, Sweden
7. Department of Medical Biosciences, Pathology, Umeå University, Umeå, Sweden
8. Medical Research Council Toxicology Unit, Leicester University, Leicester, UK
9. Haemato-Oncology, Division of Pathology, Institute of Cancer Research, Sutton, Surrey, UK
10. Haematological Sciences, Medical School, Newcastle University, Newcastle-upon-Tyne, UK
11. Department of Haematology, Royal Victoria Infirmary, Newcastle-upon-Tyne, UK
12. Department of Haematology, Queen Elizabeth Hospital, Gateshead, Newcastle-upon-Tyne, UK
13. Department of Molecular and Clinical Cancer Medicine, University of Liverpool, Liverpool, UK.
14. Department of Haematology, Hull Royal Infirmary, Hull, UK
15. Hull York Medical School and University of Hull, Hull, UK
16. Department of Haematology, Birmingham Heartlands Hospital, Birmingham, UK
17. Department of Haematology, School of Medicine, Cardiff University, Cardiff, UK
18. Cardiff and Vale National Health Service Trust, Heath Park, Cardiff, UK

*Equal authorship at this position.

[+]Corresponding author:

E-mail: Richard.houlston@icr.ac.uk; Tel: +44 (0) 208-722-4175; Fax: +44 (0) 208-722-4365

ABSTRACT

Genome-wide association studies (GWASs) of chronic lymphocytic leukemia (CLL) have shown that common genetic variation contributes to the heritable risk of CLL. To identify additional CLL susceptibility loci, we conducted a GWAS and performed a meta-analysis with a published GWAS totaling 1,739 cases and 5,199 controls with validation in an additional 1,144 cases and 3,151 controls. A combined analysis identified novel susceptibility loci mapping to 3q26.2 (rs10936599, *P*=1.74x10^{-9}), 4q26 (rs6858698, *P*=3.07x10^{-9}), 6q25.2 (*IPCEF1*, rs2236256, *P*=1.50x10^{-10}) and 7q31.33 (*POT1*, rs17246404, *P*=3.40x10^{-8}). Additionally we identified a promising association at 5p15.33 (*CLPTM1L*, rs31490, *P*=1.72x10^{-7}) and validated recently reported putative associations at 5p15.33 (*TERT*, rs10069690, *P*=1.12x10^{-10}) and 8q22.3 (rs2511714, *P*=2.90x10^{-9}). These findings provide further insights into the genetic and biological basis of inherited genetic susceptibility to CLL.

Chronic lymphocytic leukemia (CLL) is the most common hematological malignancy in Western countries[1] and is characterized by a 8-fold increased risk in first-degree relatives[2]. Genome-wide association studies (GWASs) have so far identified common variants at 24 loci that contribute to the heritable risk of CLL [3-6]. Current projections for the number of independent regions harboring common variants associated with CLL suggest that additional risk loci conferring modest effects should be identified by expansion of discovery GWAS datasets.

To identify additional novel susceptibility loci for CLL, we conducted an independent primary scan of CLL and performed a genome-wide meta-analysis with a previously published GWAS followed by analysis of the top single nucleotide polymorphisms (SNPs) in two separate case-control series.

In the primary scan (UK-CLL-2), 1,271 CLL cases were genotyped using the Illumina Omni Express BeadChip. The newly scanned cases comprised 1,111 cases collected through the Institute of Cancer Research and Royal Marsden NHS Hospitals Trust and 160 from the Newcastle CLL Consortium. For controls, we made use of publicly accessible Hap1.2M-Duo Custom array data generated on 2,501 individuals from the UK Blood Service Control Group.

Our previous GWAS of CLL (UK-CLL-1) was based on a scan of 517 CLL cases (155 potentially enriched for genetic susceptibility by virtue of family history) using Illumina HumanCNV370-Duo BeadChips[3]. For controls, we made use of Hap1.2M-Duo Custom array data generated on 2,698 individuals from the Wellcome Trust Case Control Consortium 2 (WTCCC2) 1958 birth cohort (also known as the National Child Development Study).

Quantile-quantile plots of the genome-wide *P*-values showed there was minimal inflation of the test statistics rendering substantial cryptic population substructure or differential genotype calling between cases and controls unlikely in each of the GWAS (genomic control inflation factor, λ=1.04 and 1.05 respectively; Supplementary Figure 1). To harmonize the two GWAS datasets, we imputed UK-CLL-1 to recover untyped SNPs directly genotyped in UK-CLL-2, using data from the 1000 Genomes Project as reference. Using data on all cases and controls from each GWAS, we derived joint odds ratios (ORs) and confidence intervals (CIs) under a fixed effects model for each SNP and associated *P*-values, restricting analysis to SNPs with MAF >1%. After filtering on the basis of pre-specified quality-control measures (Online Methods), the two studies provided genotype data on 450K autosomal SNPs for 1,739 CLL cases and 5,199 controls (Supplementary Figure 2). In the meta-analysis, associations for all 22 established loci

(24 variants) showed a consistent direction of effect with previously reported studies with 9 loci having a *P*-value of <5.0x10^{-8} (Supplementary Figure 3, Supplementary Table 1).

In the combined analysis, we identified two SNPs (rs2236256, rs6062501) showing genome-wide significant evidence of association (*i.e.* $P \leq 5.0\times10^{-8}$) and mapping to distinct loci not previously associated with CLL risk (Supplementary Table 2). We also identified promising association signals (*i.e.* $P<1.0\times10^{-5}$) at 11 additional loci (Supplementary Table 2). We applied 1000 Genomes imputation to UK-CLL-1 and UK-CLL-2 at these loci to investigate if a statistically significant stronger SNP association could be identified, recovering an additional SNP which was significant at the genome-wide threshold (rs6858698; Supplementary Table 2). We performed replication genotyping of six SNPs selected on the basis of statistical significance (rs2236256, rs6062501, rs6858698) and gene centricity coupled with considerations of potential candidacy (*i.e.* role in B-cell or cancer biology; rs10936599, rs31490, rs17246404) in case-control series from the UK (803 cases, 2,780 controls) and Sweden (341 cases, 371 controls) (Online Methods; Supplementary Figure 2, Supplementary Table 3). In a meta-analysis of all four datasets, four SNPs showed evidence for an association with CLL risk which was genome-wide significant (Figure 1, Online Methods, Supplementary Table 3).

The strongest evidence of association was attained with rs2236256 at 6q25.2 (Odds ratio [OR]=1.23, 95% confidence interval [C.I.] 1.15-1.30; P=1.50x10^{-10}; P_{het}=0.21, I^2=34%; Figure 1, Supplementary Table 3). The SNP localises to the 3' UTR of the interaction protein for cytohesin exchange factors 1 gene (*IPCEF1*, Figure 2). *IPCEF1* encodes the C-terminal half of CNK3 one of the CNK scaffolds involved in signal transduction downstream of Ras. To explore the epigenetic profile of the 6q25.2 association signal, we used chromatin state segmentation, based upon lymphoblastoid cell line (LCL) data from the ENCODE project[7]. rs2236256 is predicted to reside within a strong enhancer element, therefore raising the possibility of functionality *per se* (Figure 2, Supplementary Table 4).

The second strongest signal was shown by rs10936599 at 3q26.2 (OR=1.26, 95% C.I. 1.17-1.35; P=1.74x10^{-9}; P_{het}=0.99, I^2=0%; Figure 1, Supplementary Table 3) which is responsible for the H717H polymorphism in the myoneurin gene (*MYNN*; MIM 606042). Carrier status for the rs10936599-C allele has previously been shown to influence the risk of both colorectal cancer and multiple myeloma[8,9]. While *MYNN* encodes a zinc finger protein of unknown function expressed principally in muscle, rs10936599 (169,492,101bps) maps within a 250Kb region of LD which also encompasses the telomerase RNA component gene (*TERC*; MIM 602322, Figure 2). The imputed SNP rs2293607, which is correlated with rs10936599 (r^2=1.0, D'=1.0) and maps 63bps 5' to *TERC,* has recently been shown to be associated with *TERC* mRNA expression in a colorectal cancer cell line[10]. rs2293607 is predicted to lie within an active promoter region (Figure 2) and is a transcription factor binding site for multiple proteins, including OCT2, ELF1, NFKB and CMYC, which are important in the regulation of B-cell gene expression (Supplementary Table 4). While we found no evidence for a relationship between rs10936599, and telomere length in 246 CLL patients (Supplementary Table 5), carrier status for the rs10936599-C risk allele is previously been associated with significantly longer telomeres in leukocytes[10,11].

The third significant association was at rs6858698 on 4q26 (OR=1.31, 95% C.I. 1.20-1.44; P=3.07x10^{-9}; P_{het}=0.18, I^2=39%; Figure 1, Supplementary Table 3) which maps 760bp upstream of the calcium/calmodulin-dependent protein kinase II-delta (*CAMK2D*; MIM 607708, Figure 2) gene. While *CAMK2D* has as yet no documented role in CLL, calcium/calmodulin signalling has a role in the regulation of gene expression following B-cell activation, through inhibition of transcription factor E2A[12]. Since rs6858698 resides within a conserved region (Genomic Evolutionary Rate Profiling [GERP] score=2.40) and maps to a predicted promoter with binding sites for multiple transcription factors, it is possible that the SNP is responsible for the 4q26 association (Figure 2; Supplementary Table 4).

The fourth significant SNP association was shown by rs17246404 at 7q31.33 (OR=1.22, 95% C.I. 1.14-1.31; P=3.40x10^{-8}; P_{het}=0.77, I^2=0%; Figure 1, Supplementary Table 3) which maps to the 3' UTR of protection of telomeres 1 gene (*POT1*; MIM 606478, Figure 2). POT1 forms part of the shelterin complex which functions to protect telomeres and maintain chromosomal stability. Recently, sequence analysis has demonstrated somatic mutations of *POT1* in 3.5% of all CLL and 9% of immunoglobulin heavy chain variable (*IGHV*) unmutated CLL[13]. Since *POT1* mutated cells have numerous telomeric and chromosomal abnormalities it has been suggested that *POT1* mutation facilitates the acquisition of the malignant features of CLL cells. Telomeric deprotection has been postulated to be an early step in the evolution of CLL, not merely a consequence of telomere shortening but also of shelterin alteration. Hence it is plausible that the functional basis of the 7q31.33 association is through aberrant *POT1* expression.

In addition to these four risk loci, we identified a promising association at 5p15.33 defined by rs31490 (OR=1.18, 95% C.I. 1.11-1.26, P=1.72x10^{-7}, Figure 1, Supplementary Table 3), which resides within intron 2 of cleft lip and palate transmembrane protein 1-like gene (*CLPTM1L*; alias cisplatin resistance-related protein 9, *CRR9*; MIM 612585, Figure 2) which encodes a transcript whose over expression has been linked to *cis*-platinum resistance by enhancing apoptosis. A recent GWAS of CLL has reported promising associations at 5p15.33 defined by rs10069690 and at 8q22.33 defined by rs2511714[6]. Combining the *P*-values for rs10069690 and rs2511714 obtained in our meta-analysis (*P*=1.0x10^{-4} and 1.0x10^{-3} respectively) with published data[6] provides robust evidence for both associations (combined *P*-values 1.10x10^{-10} and 2.90x10^{-9} respectively; Supplementary Figure 4). rs10069690 maps to intron 4 of *TERT* (telomerase reverse transcriptase; MIM 187270), at 5p15.33 and is independent of the rs31490 association. rs2511714 maps 5.6kb telomeric to the gene encoding homo sapiens outer dense fiber of sperm tails 1 (*ODF1*; MIM 182878) within an ~10 kb region of LD.

CLL cells are characterised by short telomeres despite a low proliferative index hence both *CLPTM1L* and *TERT* which map to 5p15.33 represent attractive candidates for CLL susceptibility *a priori* assuming that the causal variant exerts an influence through a *cis* effect. Intriguingly, the 5p15.33 *TERT-CLPTM1L* region is characterised by the multiple distinct risk loci with different tumour specificities. rs31490 is highly correlated with rs401681 which also maps within intron 2 of *CLPTM1L* (r^2=0.93, D'=0.98). The T allele of rs401681 which increases CLL risk also increases the risk of pancreatic cancer[14] and melanoma[15,16] but is associated with reduced lung[17] and bladder cancer risk[18]. Carrier status for the CLL risk allele, rs10069690-T has previously been shown to increase breast and ovarian cancer[19,20] risk but reduce risk of testicular cancer[21].

CLL is characterized by male predominance and can be classified on the basis of the presence or absence of somatic hypermutations of the *IGHV* genes, with *IGHV*-mutated CLL typically having a more benign prognosis[22]. None of the seven SNPs, rs10936599, rs6858698, rs2236256, rs17246404, rs31490, rs10069690 and rs2511714, showed a significant relationship with these traits, after adjustment for multiple testing (Supplementary Table 5). While individually none of seven SNP or previously identified risk SNPs showed a relationship with age at diagnosis, cases diagnosed young tended to carry a higher number of risk alleles (*P*=0.01), consistent with early-onset CLL being enriched for genetic susceptibility.

To gain understanding of the functional basis of the new CLL loci, we examined publically available eQTL data obtained from the analysis of LCLs[23]. Only one of the seven SNPs, rs10936599, was associated (*P*<0.05) with transcript levels of a *cis*-gene (Supplementary Table 6). Irrespective of the functional basis of the associations we have identified, the potential impact of common alleles on gene expression will be modest and could occur at any time before diagnosis of CLL. Moreover, expression differences may only be relevant to a subpopulation of cells that provide 'targets' for leukemogenic mutations.

To gain insight into the allelic architecture of predisposition to CLL, we examined for interactive effects between 3q26.2, 4q26, 5p15.33, 6q25.2 and 7q31.33 SNPs and the previously identified CLL risk SNPs (Supplementary Table 7). There was no evidence of significant interaction (*i.e. P*>0.05 after adjustment for multiple testing) compatible with each locus having an independent effect on CLL risk.

Our GWAS of CLL has identified four novel loci and validated two recently identified promising associations, bringing the total number of common risk variants for CLL to 30. While additional studies are required to decipher the functional basis of the risk loci the proximity of several of the loci to genes having a role in telomeres suggests a plausible mechanism of biological relevance.

The contribution of all common variation to the heritability of CLL is around 46%-59%[6,24]. Collectively the 30 risk variants thus far identified show an area under the curve of 0.74 and account for 19% of the familial risk of CLL. It is therefore likely that a significant number of additional common variants remain to be discovered by further GWAS-based initiatives. While the power of our study to detect the common loci conferring risks of ≥1.3 was high we had low power to detect alleles with smaller effects and/or minor allele frequencies (MAFs) <0.1. By implication, variants with such profiles are likely to represent a much larger class of susceptibility loci for CLL, because of truly small effect sizes or submaximal LD with tagging SNPs. Further efforts to expand the scale of GWAS, in terms of both sample size and SNP coverage, and to increase the number of SNPs taken forward to large-scale replication should therefore identify additional risk variants for CLL.

URLs

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

HapMap: http://www.hapmap.org;

1000genomes: http://www.1000genomes.org

LGC Genomics: http://www.lgcgenomics.com

SNAP http://www.broadinstitute.org/mpg/snap

IMPUTE: https://mathgen.stats.ox.ac.uk

Wellcome Trust Case Control Consortium: www.wtccc.org.uk

Mendelian Inheritance In Man: http://www.ncbi.nlm.nih.gov/omim

1958 Birth Cohort: http://www.cls.ioe.ac.uk/studies.asp?section=000100020003

UCSC genome browser: http://genome.ucsc.edu

HaploReg: http://www.broadinstitute.org/mammals/haploreg/haploreg.php

RegulomeDB: http://regulome.stanford.edu/

International Immunogenetics Information System, http://imgt.cines.fr/

ACKNOWLEDGEMENTS

Leukaemia Lymphoma Research provided principal funding for the study (LRF05001 and 06002). Support from Cancer Research UK (C1298/A8362 supported by the Bobby Moore Fund) and the Arbib Fund is acknowledged. The study made use of genotyping data on the 1958 Birth Cohort; a full list of the investigators who contributed to the generation of these data is available from www.wtccc.org.uk. We thank L. Padyukov (Karolinska Institutet) 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. 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 http://www.wtccc.org.uk/. We are grateful to all investigators and to all the patients and individuals for their participation and finally we would also like to thank the clinicians, other hospital staff and study staff that contributed to the blood sample and data collection for this study.

AUTHOR CONTRIBUTIONS

R.S.H. obtained financial support, designed and provided overall project management. R.S.H. drafted the manuscript. H.E.S. performed project management and supervised genotyping, M.C.D.B. performed bioinformatic and statistical analyses; G.P.S. and A.H. performed genotyping; Y.W. and M.C.D.B. performed imputation analysis; D.C. and R.S.H. established the ICLLLC; C.D. and D.C performed recruitment of samples. In Sweden: L.M. Performed sample collection and prepared DNA; R.R. performed collection of all cases while G.J and K.E.S performed sample collection in the SCALE study; G.R. performed telomere analysis. In Newcastle: J.M.A. and D.J.A. conceived of the Newcastle CLL Consortium (NCLLC). J.M.A. obtained financial support, supervised laboratory management and oversaw genotyping of NCLLC cases. N.J.S. performed sample management of cases. A.G.H. developed the Newcastle Haematology Biobank, incorporating NCLLC. T.M.-F., G.H.J., G.S., R.J.H., A.R.P., D.J.A., J.R.B., G.P., C.P. and C.F. developed protocols for recruitment of individuals with CLL and sample acquisition and performed sample collection of cases. In Leicester: M.J.S.D. performed overall management, collection and processing of samples; S.J. and A.M. performed DNA extractions and IGVH mutation assays. All authors contributed to the final paper.

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interest.

TABLE AND FIGURE LEGENDS

FIGURES

Figure 1: Plot of the odds ratios of chronic lymphocytic leukemia associated with (a) rs10936599, (b) rs6858698, (c) rs31490, (d) rs2236256 and (e) rs17246404. Studies were weighted according to the inverse of the variance of the log of the odds ratio (OR) calculated by unconditional logistic regression. Horizontal lines: 95% confidence intervals (95% CI). Box: OR point estimate; its area is proportional to the weight of the study. Diamond (and broken line): summary OR computed under a fixed effects model, with 95% CI given by its width. Unbroken vertical line: null value (OR=1.0).

Figure 2: Regional plots of association results, recombination rates and chromatin state segmentation track for (a) 3q26.2 (rs10936599), (b) 4q26 (rs6858698), (c) 5p13.32 (rs31490), (d) 6q25.2 (rs2236256) and (e) 7q31.33 (rs17246404) risk loci. Association results of both genotyped (triangles) and imputed (circles) SNPs in the GWAS samples and recombination rates for rates. $-\log_{10}$ *P*-values (*y* axis) of the SNPs are shown according to their chromosomal positions (*x* axis). The top genotyped SNP in each combined analysis is shown as a large diamond and is labeled by its rsID. 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 HapMap Utah residents of Western and Northern European ancestry (CEU) 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 the relative positions; therefore, maps are not to physical scale. The lower panel shows the exons and introns of the genes of interest; observed SNP and the chromatin state segmentation track (ChromHMM) for lymphoblastoid cell line data derived from the ENCODE project.

ONLINE METHODS

Ethics

Collection of samples and clinico-pathological information from subjects was undertaken with informed consent and relevant ethical review board approval in accordance with the tenets of the Declaration of Helsinki.

Subjects

For both incident and prevalent cases, the diagnosis of CLL has been confirmed in accordance with WHO classification guidelines[25].

GWAS datasets

UK-CLL-2 comprised 1,271 CLL cases collected from two ongoing initiatives: (i). 1,111 through a UK national study of CLL genetics coordinated by the ICR (712 male, mean age at diagnosis 62.6 years; s.d. 11.1); (ii) 160 were collected through the Newcastle CLL Consortium (NCLLC), (96 male, mean age at diagnosis 65.3 years; s.d. 10.6) from patients attending six haematology units in the UK (Newcastle Royal Victoria Infirmary, Gateshead Queen Elizabeth Hospital, Birmingham Heartlands Hospital, Hull Royal Infirmary, Liverpool Royal University Hospital and The University of Wales College of Medicine Hospital). The proportion of incident and prevalent cases varied between clinical centres. All cases were UK residents and had self-reported European ancestry. Cases were genotyped using the Illumina Omni Express BeadChip according to manufacturer's recommendations. For controls, we made use of publicly accessible Hap1.2M-Duo Custom array data generated by the WTCCC on 2,501 individuals from the UK Blood Service Control Group.

UK-CLL-1: Details of this study have been previously reported[3,4]. Briefly, peripheral blood samples were collected from 517 individuals with CLL (364 male, mean age at diagnosis 61.7 years; s.d. 9.9): 155 with at least one relative affected with CLL or a B-cell lymphoproliferative disorder (LPD) (95 male, mean age at diagnosis 58.9 years, s.d. 10.8), ascertained between 1997 and 2007 through the International CLL linkage consortium (ICLLLC), an ongoing initiative to ascertain and collect CLL cases and families segregating CLL and B-cell LPD through clinical centers and 362 cases (269 male, mean age at diagnosis 62.9 years, s.d. 9.3) ascertained between 1999 and 2004 through Leukaemia Research CLL-4, a randomized phase III trial[26]. Blood samples from the cases were collected at registration. All cases were British residents and had self-reported European ancestry. Genotyping of cases was performed using

Illumina Human 317K arrays according to the manufacturer's protocols (Illumina, San Diego, USA). For controls, we used publicly accessible data generated by the Wellcome Trust Case Control Consortium on 2,698 individuals from the 1958 Birth Cohort (58C; also known as the National Child Development Study)[27].

Quality control of GWAS datasets

DNA samples with GenCall scores <0.25 at any locus were considered "no calls". A SNP was deemed to have failed if <95% of DNA samples generated a genotype at the locus. Cluster plots were manually inspected for all SNPs considered for replication. The same quality control metrics on the primary GWAS data were applied as in the previous UK-CLL-1[3]. We removed from the analysis individuals showing sex discrepancy (0 samples) and samples for whom <95% of SNPs were successfully genotyped (13 samples; Supplementary Figure 2). We computed identity-by-state (IBS) probabilities for all pairs (cases and controls) to search for duplicates and closely related individuals amongst samples (defined as IBS ≥0.80, thereby excluding first-degree relatives). For all related pairs the sample having the highest call rate was retained, thereby eliminating 13 samples. To identify individuals who might have non-Western European ancestry, we merged our case and control data with phase II HapMap samples (60 western European [CEU], 60 Nigerian [YRI], 90 Japanese [JPT] and Han Chinese [CHB]). For each pair of individuals we calculated genome-wide IBS distances on markers shared between HapMap and our SNP panel, and used these as dissimilarity measures upon which to perform principal component analysis. The first two principal components for each individual were plotted and 9 samples showing marked separation from the CEU cluster were excluded from analyses (Supplementary Figure 5). We filtered out SNPs having a minor allele frequency (MAF) <1%, and a call rate <95% in cases or controls. We also excluded SNPs showing departure from Hardy-Weinberg equilibrium (HWE) at $P<10^{-5}$. For replication and validation analysis call rates were >95% per 384-well plate for each SNP; cluster plots were visually examined by two researchers.

Replication series and genotyping

The UK-replication series comprised 340 CLL cases collected through the NCLLC and 463 cases collected through the Leicester Haematology Tissue Bank of consecutive cases presenting to CLL clinic in Leicester. Controls comprised 2,780 healthy individuals ascertained through the National Study of Colorectal Cancer (1999–2006)[28] (n=676; 225 male); the Genetic Lung Cancer Predisposition Study[29] (n=693; 162 male); the Colorectal Adenoma Gene-Environment Interactions Study (n=686; 382 male); the Study of the Genetic Epidemiology of Colorectal Cancer (n=174; 88 male); and the RMHNHST family history DNA database (n=551; 217 male). These controls were the spouses or unrelated friends of individuals with malignancies. None had a personal history of malignancy at time of ascertainment. Both cases and controls were British residents and had self reported to be of European ancestry.

The Swedish replication series comprised 271 cases from SCALE (Scandinavian Lymphoma Etiology) study[30] and 70 from the biobank at Uppsala University Hospital, Uppsala, Sweden (220 male, mean age at diagnosis of 62.5 years, s.d. 9.0). For the SCALE samples, peripheral blood was collected from 231 cases during 1999-2001, within a median of 4 months from diagnosis (range, 0-29 months) while in 40 cases, follow-up samples were taken in 2007-2008. Samples from the biobank at Uppsala University Hospital were collected between 1982 and 2005. Controls were ascertained through the Karolinska Institutet, Stockholm, Sweden and included 371 healthy individuals (239 male) aged between 33 and 71 years (mean age 61.4 years, s.d. 6.8).

Genotyping was performed using competitive allele-specific PCR KASPar chemistry (LGC Genomics Ltd, Hertfordshire, UK). All primers and probes used are available on request. Samples having SNP call rates of <90% were excluded from the analysis. To ensure quality of genotyping in all assays, at least two negative controls and 1-2% duplicates (showing a concordance >99.99%) were genotyped.

Mutational status

IGHV gene mutation status was determined according to BIOMED-2 protocols as described previously[31]. Sequence analysis was conducted using Chromas software version 2.23 (Applied Biosystems) and the international immunogenetics information system database. In accordance with published criteria, we classified sequences with a germline identity of ≥98% as unmutated and those with identity of <98% as mutated[32].

Telomere length

Telomere length of genomic DNA extracted from mononuclear blood cells was determined using real-time PCR[33,34]. Mean inter-assay CV for relative telomere length was 4-8%[35]. Telomere/single copy gene values were calculated by $2^{-\Delta Ct}$ methodology.

Statistical and bioinformatic analysis

Main analyses were undertaken using R (v2.6), Stata v.10 (State College, Texas, US) and PLINK (v1.06)[36] software. The association between each SNP and risk was assessed by the Cochran-Armitage trend test.

The adequacy of the case-control matching and possibility of differential genotyping of cases and controls were formally evaluated using quantile-quantile (Q-Q) plots of *P*-values derived from test statistics. The inflation factor λ was based on the 90% least significant SNPs[37]. Odds ratios (ORs) and associated 95% confidence intervals (CIs) were calculated by unconditional logistic regression. Meta-analysis was performed using the fixed-effects inverse-variance method based on the β and standard errors[38]. Cochran's Q statistic to test for heterogeneity and the I^2 statistic to quantify the proportion of the total variation due to heterogeneity were calculated[39]. I^2 values ≥75% are considered characteristic of large heterogeneity[39], and in such cases the random-effects model was applied.

Associations by sex, age and *IGHV* mutation status were examined by logistic regression in case-only analyses. Telomere length across genotypes was compared by the Kruskal-Wallis test. The combined effect of pairs of loci identified with risk was investigated by logistic regression modelling. Evidence for interactive effects between SNPs was assessed by a likelihood ratio test comparing the fit of the null model and a model including the interaction term; the difference in log-likelihood ratio statistics was assessed by χ^2 test. The sibling relative risk attributable to each SNP was calculated using the formula:

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

where *p* is the population frequency of the minor allele, *q*=1-*p*, and r_1 and r_2 are ORs for heterozygotes and rare homozygotes, relative to common homozygotes. Assuming an overall sibling relative risk (λ_0) of 8.5 for CLL[2] and a multiplicative interaction the proportion of the familial risk attributable to SNP genotypes was calculated as $\log(\lambda^*)/\log(\lambda_0)$.

Prediction of the untyped SNPs was carried out using IMPUTEv2, based on the 1000 Genomes phase 1 integrated variant set (b37) from March 2012. Imputed data were analysed using SNPTEST v2 to account for uncertainties in SNP prediction. Association meta-analyses only included markers with info scores >0.4, HWE $P>10^{-5}$ and missingness rate <0.05. Meta-analyses were carried out in R using the genotype probabilities from IMPUTEv2, where a SNP was not directly typed.

LD metrics were calculated in PLINK using 1000 Genomes data and plotted using SNAP. LD blocks were defined on the basis of HapMap recombination rate (cM/Mb) as defined using the Oxford recombination hotspots[40] and on the basis of distribution of confidence intervals defined by Gabriel *et al.*[41]

To explore the epigenetic profile of association signals we made use of chromatin state segmentation in lymphoblastoid cell lines data generated by the ENCODE Project[7]. The states were inferred from ENCODE Histone Modification data (H4K20me1, H3K9ac, H3K4me3, H3K4me2, H3K4me1, H3K36me3, H3K27me3, H3K27ac and CTCF) binarized using a multivariate Hidden Markov Model. We made use of HaploReg[42] and RegulomeDB[43] to examine whether any of the SNPs or their proxies (*i.e.* r^2>0.8 in 1000 Genomes EUR reference panel) annotate putative transcription factor binding/enhancer elements. We assessed sequence conservation using Genomic Evolutionary Rate Profiling (GERP). GERP scores (−12 to 6, with 6 being indicative of complete conservation) reflect the proportion of substitutions at that site rejected by selection compared with observed substitutions expected under a neutral evolutionary model, based on sequence alignment of 34 mammalian species[44].

Relationship between SNP genotype and mRNA expression

To look for a relationship between SNP genotype and expression levels in lymphocytes we made use of publicly available expression data generated on lymphoblastoid cell lines from the MuTHER resource, using a combination of Illumina HumanHap300, HumanHap610Q, 1M-Duo and 1.2MDuo 1M chips (Illumina, San Diego, USA). Data were examined using Genevar[45].

REFERENCES

1. Stevenson, F.K. & Caligaris-Cappio, F. Chronic lymphocytic leukemia: revelations from the B-cell receptor. *Blood* **103**, 4389-95 (2004).
2. Goldin, L.R., Bjorkholm, M., Kristinsson, S.Y., Turesson, I. & Landgren, O. Elevated risk of chronic lymphocytic leukemia and other indolent non-Hodgkin's lymphomas among relatives of patients with chronic lymphocytic leukemia. *Haematologica* **94**, 647-53 (2009).
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. Crowther-Swanepoel, D. *et al.* Common variants at 2q37.3, 8q24.21, 15q21.3 and 16q24.1 influence chronic lymphocytic leukemia risk. *Nat Genet* **42**, 132-6 (2010).
5. Slager, S.L. *et al.* Common variation at 6p21.31 (BAK1) influences the risk of chronic lymphocytic leukemia. *Blood* **120**, 843-6 (2012).
6. Berndt, S.I. *et al.* Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia. *Nat Genet* **45**, 868-76 (2013).
7. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28**, 817-25 (2010).
8. Houlston, R.S. *et al.* Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. *Nat Genet* **42**, 973-7 (2010).
9. Chubb, D. *et al.* Common variation at 3q26.2, 6p21.33, 17p11.2 and 22q13.1 influences multiple myeloma risk. *Nat Genet* **45**, 1221-5 (2013).
10. Jones, A.M. *et al.* TERC polymorphisms are associated both with susceptibility to colorectal cancer and with longer telomeres. *Gut* **61**, 248-54 (2012).
11. Codd, V. *et al.* Identification of seven loci affecting mean telomere length and their association with disease. *Nat Genet* **45**, 422-7, 427e1-2 (2013).
12. Verma-Gaur, J., Hauser, J. & Grundstrom, T. Negative feedback regulation of antigen receptors through calmodulin inhibition of E2A. *J Immunol* **188**, 6175-83 (2012).
13. Ramsay, A.J. *et al.* POT1 mutations cause telomere dysfunction in chronic lymphocytic leukemia. *Nat Genet* **45**, 526-30 (2013).
14. Petersen, G.M. *et al.* A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. *Nat Genet* **42**, 224-8 (2010).
15. Barrett, J.H. *et al.* Genome-wide association study identifies three new melanoma susceptibility loci. *Nat Genet* **43**, 1108-13 (2011).
16. Stacey, S.N. *et al.* New common variants affecting susceptibility to basal cell carcinoma. *Nat Genet* **41**, 909-14 (2009).
17. Wang, Y. *et al.* Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat Genet* **40**, 1407-9 (2008).
18. Rothman, N. *et al.* A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. *Nat Genet* **42**, 978-84 (2010).
19. Bojesen, S.E. *et al.* Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. *Nat Genet* **45**, 371-84, 384e1-2 (2013).
20. Haiman, C.A. *et al.* A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer. *Nat Genet* **43**, 1210-4 (2011).
21. Karlsson, R. *et al.* Investigation of six testicular germ cell tumor susceptibility genes suggests a parent-of-origin effect in SPRY4. *Hum Mol Genet* **22**, 3373-80 (2013).
22. 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).
23. Grundberg, E. *et al.* Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet* **44**, 1084-9 (2012).
24. Di Bernardo, M.C., Broderick, P., Catovsky, D. & Houlston, R.S. Common genetic variation contributes significantly to the risk of developing chronic lymphocytic leukemia. *Haematologica* **98**, e23-4 (2013).
25. Swerdlow SH, Campo E & N, H. World Health Organization Classification of Tumours of Haematopoietic and Lymphoid Tissues. Lyon, France: IARC Press. (2008).

26. Catovsky, D. *et al.* Assessment of fludarabine plus cyclophosphamide for patients with chronic lymphocytic leukaemia (the LRF CLL4 Trial): a randomised controlled trial. *Lancet* **370**, 230-9 (2007).
27. Power, C. & Elliott, J. Cohort profile: 1958 British birth cohort (National Child Development Study). *Int J Epidemiol* **35**, 34-41 (2006).
28. Penegar, S. *et al.* National study of colorectal cancer genetics. *Br J Cancer* **97**, 1305-9 (2007).
29. Eisen, T., Matakidou, A. & Houlston, R. Identification of low penetrance alleles for lung cancer: the GEnetic Lung CAncer Predisposition Study (GELCAPS). *BMC Cancer* **8**, 244 (2008).
30. Smedby, K.E. *et al.* Ultraviolet radiation exposure and risk of malignant lymphomas. *J Natl Cancer Inst* **97**, 199-209 (2005).
31. 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).
32. 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).
33. Cawthon, R.M. Telomere measurement by quantitative PCR. *Nucleic Acids Res* **30**, e47 (2002).
34. Nordfjall, K., Osterman, P., Melander, O., Nilsson, P. & Roos, G. hTERT (-1327)T/C polymorphism is not associated with age-related telomere attrition in peripheral blood. *Biochem Biophys Res Commun* **358**, 215-8 (2007).
35. Mansouri, L. *et al.* Short telomere length is associated with NOTCH1/SF3B1/TP53 aberrations and poor outcome in newly diagnosed chronic lymphocytic leukemia patients. *Am J Hematol* **88**, 647-51 (2012).
36. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
37. Clayton, D.G. *et al.* Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat Genet* **37**, 1243-6 (2005).
38. Petitti, D. Meta-analysis decision analysis and cost-effectiveness analysis. *Oxford University Press, Oxord, New York* (1994).
39. Higgins, J.P. & Thompson, S.G. Quantifying heterogeneity in a meta-analysis. *Stat Med* **21**, 1539-58 (2002).
40. Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**, 321-4 (2005).
41. Gabriel, S.B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225-9 (2002).
42. Ward, L.D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* **40**, D930-4 (2012).
43. Boyle, A.P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* **22**, 1790-7 (2012).
44. Cooper, G.M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**, 901-13 (2005).
45. Yang, T.P. *et al.* Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. *Bioinformatics* **26**, 2474-6 (2010).

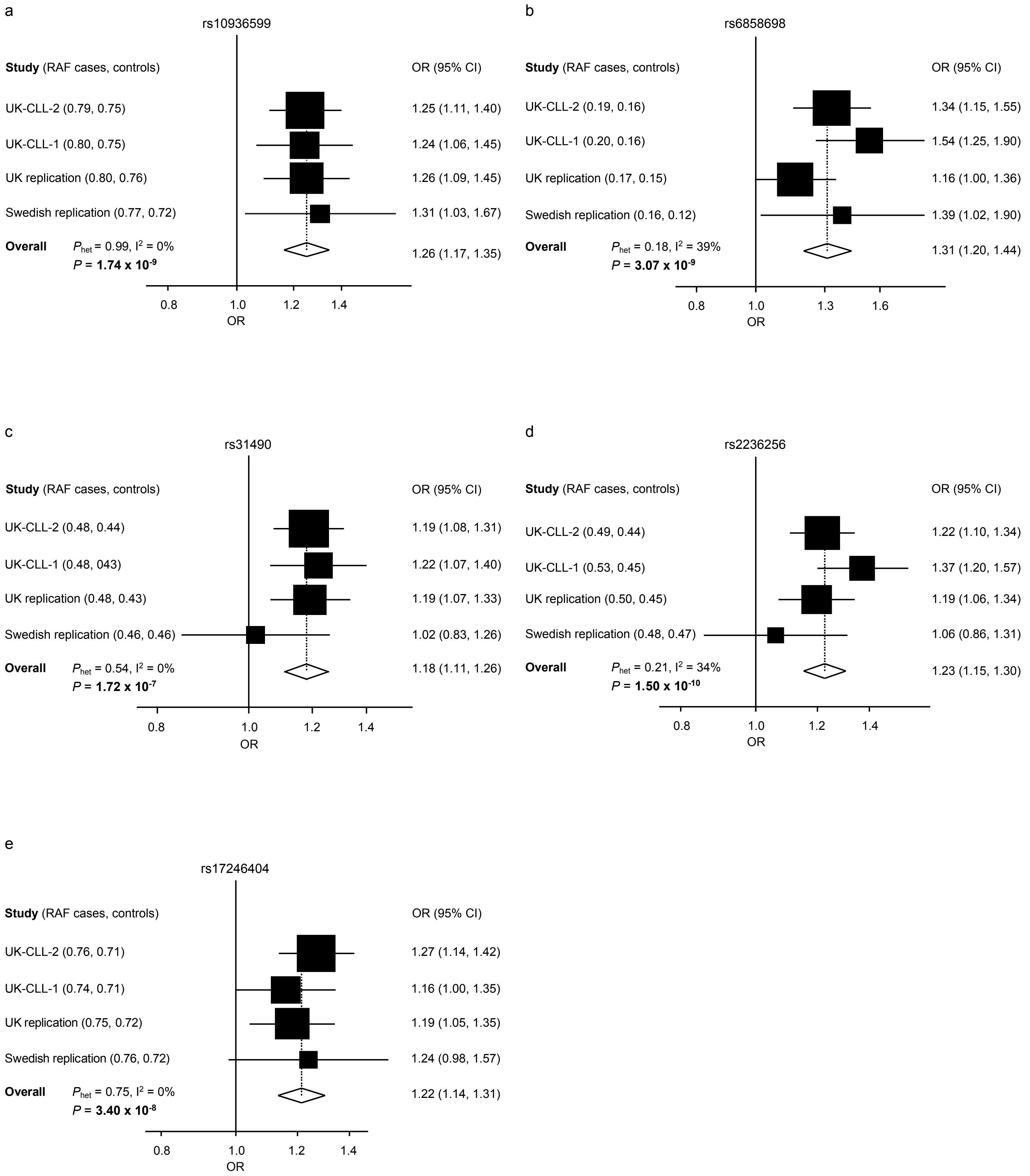
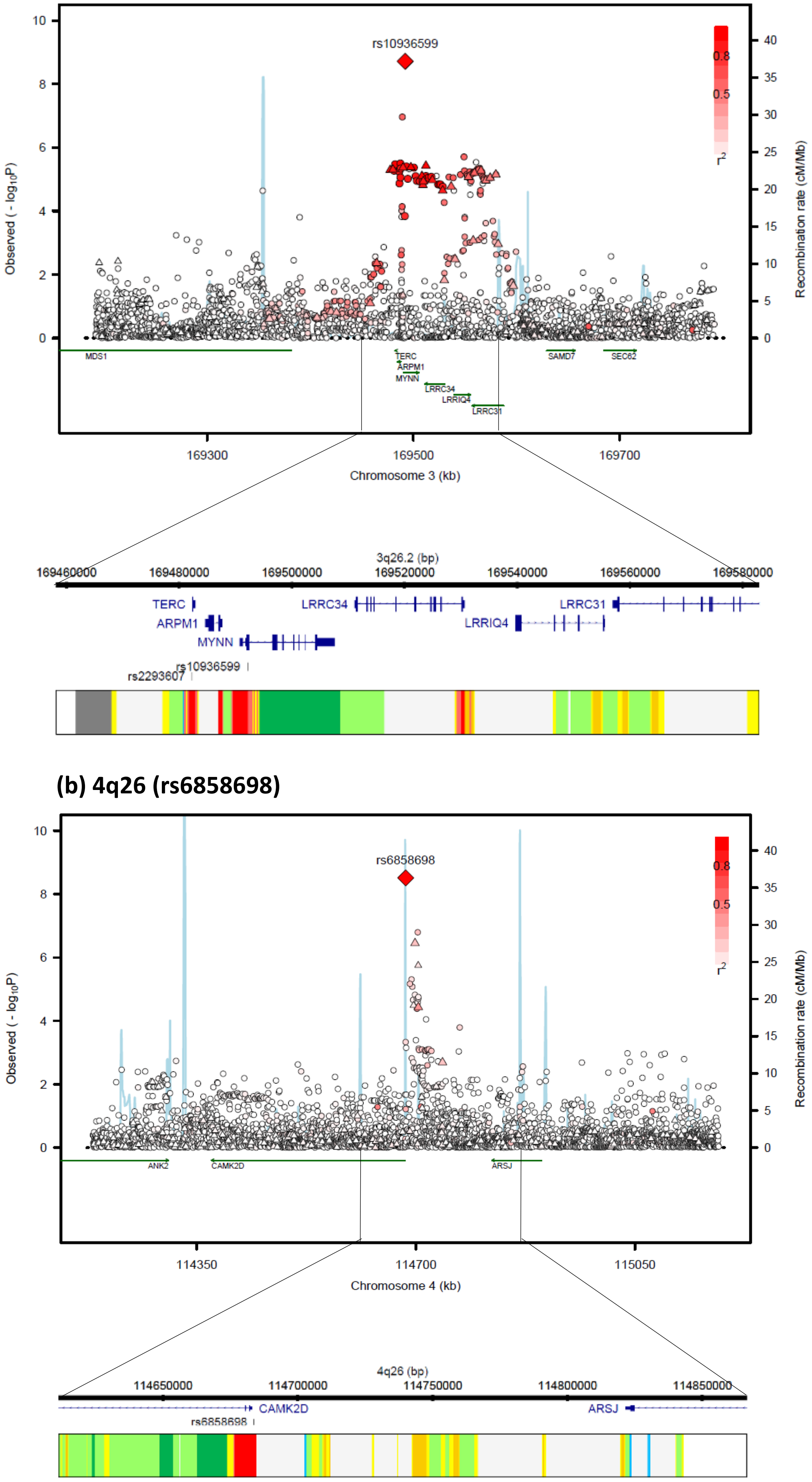

a
rs10936599
Study (RAF cases, controls)
OR (95% CI)
UK-CLL-2 (0.79, 0.75)
1.25 (1.11, 1.40)
UK-CLL-1 (0.80, 0.75)
1.24 (1.06, 1.45)
UK replication (0.80, 0.76)
1.26 (1.09, 1.45)
Swedish replication (0.77, 0.72)
1.31 (1.03, 1.67)
Overall
P_{het} = 0.99, I^2 = 0%
P = 1.74 x 10^{-9}
1.26 (1.17, 1.35)
0.8
1.0
1.2
1.4
OR
b
rs6858698
Study (RAF cases, controls)
OR (95% CI)
UK-CLL-2 (0.19, 0.16)
1.34 (1.15, 1.55)
UK-CLL-1 (0.20, 0.16)
1.54 (1.25, 1.90)
UK replication (0.17, 0.15)
1.16 (1.00, 1.36)
Swedish replication (0.16, 0.12)
1.39 (1.02, 1.90)
Overall
P_{het} = 0.18, I^2 = 39%
P = 3.07 x 10^{-9}
1.31 (1.20, 1.44)
0.8
1.0
1.3
1.6
OR
c
rs31490
Study (RAF cases, controls)
OR (95% CI)
UK-CLL-2 (0.48, 0.44)
1.19 (1.08, 1.31)
UK-CLL-1 (0.48, 043)
1.22 (1.07, 1.40)
UK replication (0.48, 0.43)
1.19 (1.07, 1.33)
Swedish replication (0.46, 0.46)
1.02 (0.83, 1.26)
Overall
P_{het} = 0.54, I^2 = 0%
P = 1.72 x 10^{-7}
1.18 (1.11, 1.26)
0.8
1.0
1.2
1.4
OR
d
rs2236256
Study (RAF cases, controls)
OR (95% CI)
UK-CLL-2 (0.49, 0.44)
1.22 (1.10, 1.34)
UK-CLL-1 (0.53, 0.45)
1.37 (1.20, 1.57)
UK replication (0.50, 0.45)
1.19 (1.06, 1.34)
Swedish replication (0.48, 0.47)
1.06 (0.86, 1.31)
Overall
P_{het} = 0.21, I^2 = 34%
P = 1.50 x 10^{-10}
1.23 (1.15, 1.30)
0.8
1.0
1.2
1.4
OR
e
rs17246404
Study (RAF cases, controls)
OR (95% CI)
UK-CLL-2 (0.76, 0.71)
1.27 (1.14, 1.42)
UK-CLL-1 (0.74, 0.71)
1.16 (1.00, 1.35)
UK replication (0.75, 0.72)
1.19 (1.05, 1.35)
Swedish replication (0.76, 0.72)
1.24 (0.98, 1.57)
Overall
P_{het} = 0.75, I^2 = 0%
P = 3.40 x 10^{-8}
1.22 (1.14, 1.31)
0.8
1.0
1.2
1.4
OR

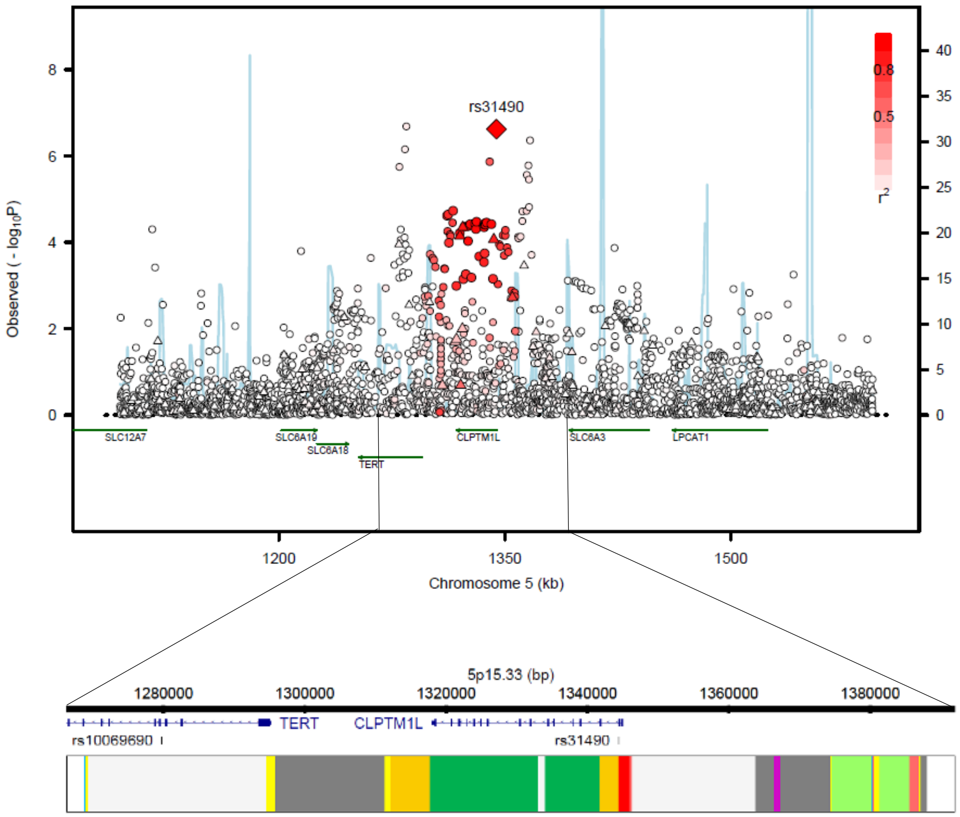
Figure 1



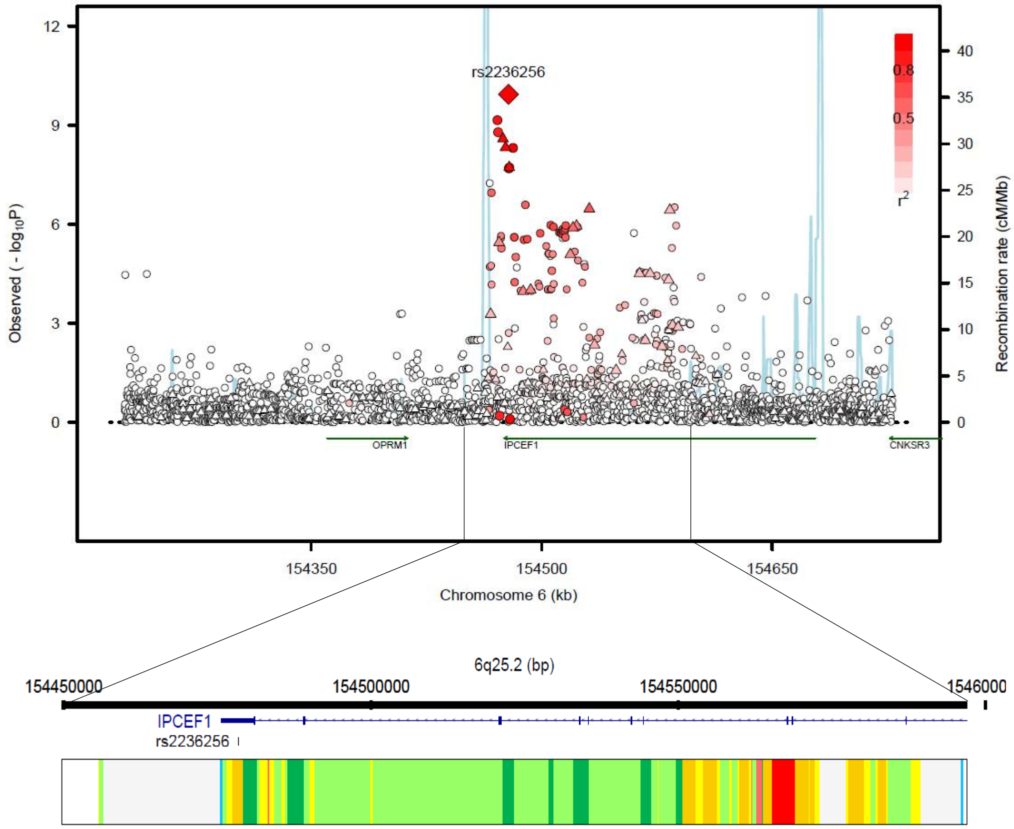
(a) 3q26.2 (rs10936599)

(b) 4q26 (rs6858698)

(c) 5p15.33 (rs31490)



(d) 6q25.2 (rs2236256)



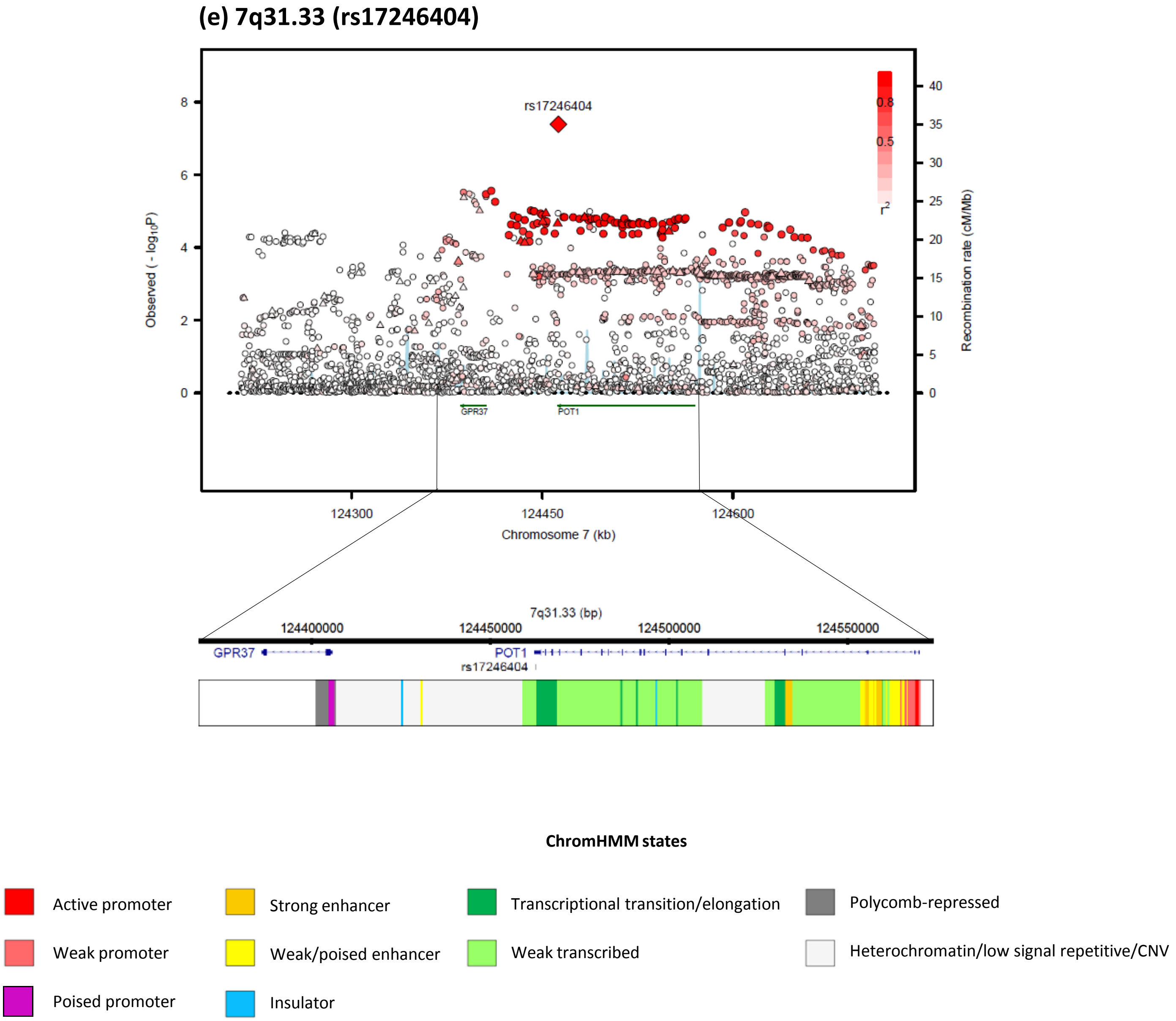

(e) 7q31.33 (rs17246404)
rs17246404
Observed (- $\log_{10}$P)
8
6
4
2
0
0.8
0.5
r^2
40
35
30
25
20
15
10
5
0
Recombination rate (cM/Mb)
GPR37
POT1
124300
124450
124600
Chromosome 7 (kb)
7q31.33 (bp)
124400000
124450000
124500000
124550000
GPR37
POT1
rs17246404
ChromHMM states
Active promoter
Weak promoter
Poised promoter
Strong enhancer
Weak/poised enhancer
Insulator
Transcriptional transition/elongation
Weak transcribed
Polycomb-repressed
Heterochromatin/low signal repetitive/CNV

Figure 2