Genome-wide association analysis implicates dysregulation of immunity genes in chronic lymphocytic leukaemia

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

Citation

Published Version
doi:10.1038/ncomms14175

Citable link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:31731882

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA
Genome-wide association analysis implicates dysregulation of immunity genes in chronic lymphocytic leukaemia

Phillip J. Law et al.#

Several chronic lymphocytic leukaemia (CLL) susceptibility loci have been reported; however, much of the heritable risk remains unidentified. Here we perform a meta-analysis of six genome-wide association studies, imputed using a merged reference panel of 1,000 Genomes and UK10K data, totalling 6,200 cases and 17,598 controls after replication. We identify nine risk loci at 1p36.11 (rs34676223, \( P = 5.04 \times 10^{-13} \)), 1q42.13 (rs41271473, \( P = 1.06 \times 10^{-10} \)), 1q42.13 (rs41271473, \( P = 1.06 \times 10^{-10} \)), 4q24 (rs71597109, \( P = 1.37 \times 10^{-10} \)), 4q35.1 (rs57214277, \( P = 3.69 \times 10^{-8} \)), 6p21.31 (rs3800461, \( P = 1.97 \times 10^{-8} \)), 11q23.2 (rs61904987, \( P = 2.64 \times 10^{-11} \)), 18q21.1 (rs1036935, \( P = 3.27 \times 10^{-8} \)), 19p13.3 (rs7254272, \( P = 4.67 \times 10^{-8} \)) and 22q13.33 (rs140522, \( P = 2.70 \times 10^{-9} \)). These new and established risk loci map to areas of active chromatin and show an over-representation of transcription factor binding for the key determinants of B-cell development and immune response.

Correspondence and requests for materials should be addressed to R.H. (email: richard.houlston@icr.ac.uk) or to S.L.S. (email: slager.susan@mayo.edu).

#A full list of authors and their affiliations appears at the end of the paper.
Chronic lymphocytic leukaemia (CLL) is an indolent B-cell malignancy that has a strong genetic component, as evidenced by the eightfold increased risk seen in relatives of CLL patients. Our understanding of CLL genetics has been transformed by genome-wide association studies (GWAS) that have identified risk alleles for CLL. So far, common genetic variation at 33 loci has been shown to influence CLL risk. Although projections indicate that additional risk variants for CLL can be discovered by GWAS, the statistical power of the individual existing studies is limited.

To gain a more comprehensive insight into CLL predisposition, we analysed genome-wide association data from populations of European ancestry from Europe, North America and Australia, identifying nine new risk loci. Our findings provide additional insights into the genetic and biological basis of CLL risk.

Results

Association analysis. After quality control, the six GWAS provided single-nucleotide polymorphism (SNP) genotypes on 4,478 cases and 13,213 controls (Supplementary Tables 1 and 2). To increase genomic resolution, we imputed >10 million SNPs using the 1000 Genomes Project combined with UK10K as reference. Quantile–Quantile (Q–Q) plots for SNPs with minor allele frequency (MAF) >0.5% show no evidence of substantive overdispersion (λ between 1.00 and 1.10 across the studies; Supplementary Fig. 1). Meta-analysing the association test results from the six series, we derived joint odds ratios per-allele and 95% confidence intervals under a fixed-effects model. Conditional analysis of GWAS data showed no evidence for additional independent signals at these nine loci. In analyses limited to the exomes of 141 CLL cases from 66 families, we found no evidence to suggest that any of the association signals might be a consequence of linkage disequilibrium (LD) with a rare disruptive coding variant.

Several of the newly identified risk SNPs map in or near to genes with established roles in B-cell biology, hence representing credible candidates for susceptibility to CLL. The 4q24 association marked by rs71597109 (Fig. 2) maps to intron 1 of the gene encoding BANK1 (B-cell scaffold protein with ankyrin repeats 1), a B-cell-specific scaffold protein. SNPs at this locus have been associated with systemic lupus erythematosus risk. BANK1 expression is only seen in functional B-cell antigen receptor (BCR)-expressing B cells, mediating effects through LYN-mediated tyrosine phosphorylation of inositol triphosphate receptors. BANK1-deficient mice display higher levels of mature B cells and spontaneous germinal centre B cells, while studies in humans found lower BANK1 transcript levels in CLL versus normal B cells. The 19p13.3 association marked by rs7254272 (Fig. 2) maps 2.5 kb 5′ to ZBTB7A (zinc finger and BTB domain-containing protein 7a, alias LRF, leukaemia/lymphoma-related factor, pokemnon). ZBTB7A is a master regulator of B versus T lymphoid fate. Loss of ZBTB7A results in aberrant activation of the NOTCH pathway in lymphoid progenitors. NOTCH is constitutively activated in CLL and is a determinant of resistance to apoptosis in CLL cells.

Figure 1 | Manhattan plot of association P values. Shown are the genome-wide P values (two-sided) of >10 million successfully imputed autosomal SNPs in 4,478 cases and 13,213 controls from the discovery phase. Text labelled in red are previously identified risk loci, and text labelled in blue are newly identified risk loci. The red horizontal line represents the genome-wide significance threshold of $P = 5.0 \times 10^{-8}$.
upstream of MDS2 (Fig. 2), which is the fusion partner of ETV6 in t(1;12)(p36;p13) myelodysplasia. Based on RNA sequencing (RNA-seq) data from patients, MDS2 is overexpressed in CLL versus normal cells and also differentially expressed between two experimentally determined CLL subgroups. The SNP rs57214277 maps to q43.5.1 and resides 140 kb centromeric to IRF2 (interferon regulatory factor 2, Fig. 2). Interferon (IFN)-β, a family of antiviral immune genes, induces IRF2 that inhibits the reactivation of murine gamma herpesvirus. Furthermore, SNPs in strong LD with rs57214277 are associated with increased expression of IRF2 as well as trans-regulation of a network of genes in lipopolysaccharide and IFNγ-treated monocytes. rs140522 maps to 22q13.33 (Fig. 2), which has previously been associated with multiple sclerosis risk. This region of LD contains four genes, of which only NCAHF2 (non-SMC condensin II complex subunit H2) shows differential expression between CLL and normal B cells (~2.5-fold lower levels in CLL), and plays an essential role in mitotic chromosome assembly and segregation. rs41271473, rs3800461, rs61904987 and rs1036935 mark genes that have roles in WNT signalling (ROHU, autophagy (C6orf106), transcriptional activation (CXXC1), kinetochore association (SKAI, ZW10) and protein degradation (USP28, TMRPRSS5; Fig. 3).

**Table 1 | Summary results for SNPs associated with CLL risk.**

<table>
<thead>
<tr>
<th>SNP</th>
<th>Locus</th>
<th>Position (bp, hg19)</th>
<th>Risk allele</th>
<th>Data set</th>
<th>RAF (case; control)</th>
<th>OR</th>
<th>95% CI</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs34676223</td>
<td>1p36.11</td>
<td>23943735</td>
<td>C</td>
<td>Discovery</td>
<td>(0.74; 0.71)</td>
<td>1.16</td>
<td>(1.09; 1.22)</td>
<td>2.69 x 10^-7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replication</td>
<td>(0.74; 0.69)</td>
<td>1.29</td>
<td>(1.18; 1.42)</td>
<td>4.69 x 10^-8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Combined</td>
<td></td>
<td>1.19</td>
<td>(1.14; 1.25)</td>
<td>5.04 x 10^-13</td>
</tr>
<tr>
<td>rs41271473</td>
<td>1q42.13</td>
<td>228880296</td>
<td>G</td>
<td>Discovery</td>
<td>(0.81; 0.79)</td>
<td>1.19</td>
<td>(1.12; 1.26)</td>
<td>4.69 x 10^-8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replication</td>
<td>(0.82; 0.79)</td>
<td>1.20</td>
<td>(1.08; 1.34)</td>
<td>5.59 x 10^-4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Combined</td>
<td></td>
<td>1.19</td>
<td>(1.13; 1.26)</td>
<td>1.06 x 10^-10</td>
</tr>
<tr>
<td>rs71597109</td>
<td>4q24</td>
<td>102741002</td>
<td>C</td>
<td>Discovery</td>
<td>(0.72; 0.69)</td>
<td>1.17</td>
<td>(1.11; 1.24)</td>
<td>1.02 x 10^-8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replication</td>
<td>(0.73; 0.71)</td>
<td>1.15</td>
<td>(1.05; 1.26)</td>
<td>3.46 x 10^-3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Combined</td>
<td></td>
<td>1.17</td>
<td>(1.11; 1.22)</td>
<td>1.37 x 10^-10</td>
</tr>
<tr>
<td>rs57214277</td>
<td>4q35.1</td>
<td>185254772</td>
<td>T</td>
<td>Discovery</td>
<td>(0.44; 0.41)</td>
<td>1.14</td>
<td>(1.08; 1.19)</td>
<td>9.56 x 10^-7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replication</td>
<td>(0.43; 0.39)</td>
<td>1.12</td>
<td>(1.03; 1.21)</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Combined</td>
<td></td>
<td>1.13</td>
<td>(1.08; 1.18)</td>
<td>3.69 x 10^-8</td>
</tr>
<tr>
<td>rs3800461</td>
<td>6p21.31</td>
<td>34616322</td>
<td>C</td>
<td>Discovery</td>
<td>(0.13; 0.11)</td>
<td>1.21</td>
<td>(1.12; 1.31)</td>
<td>4.20 x 10^-7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replication</td>
<td>(0.12; 0.11)</td>
<td>1.17</td>
<td>(1.03; 1.34)</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Combined</td>
<td></td>
<td>1.20</td>
<td>(1.13; 1.23)</td>
<td>1.97 x 10^-8</td>
</tr>
<tr>
<td>rs61904987</td>
<td>1q23.2</td>
<td>113517203</td>
<td>T</td>
<td>Discovery</td>
<td>(0.14; 0.12)</td>
<td>1.23</td>
<td>(1.14; 1.32)</td>
<td>4.44 x 10^-8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replication</td>
<td>(0.13; 0.12)</td>
<td>1.26</td>
<td>(1.12; 1.42)</td>
<td>1.20 x 10^-4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Combined</td>
<td></td>
<td>1.24</td>
<td>(1.16; 1.30)</td>
<td>2.46 x 10^-11</td>
</tr>
<tr>
<td>rs1036935</td>
<td>18q21.1</td>
<td>47843534</td>
<td>A</td>
<td>Discovery</td>
<td>(0.25; 0.22)</td>
<td>1.17</td>
<td>(1.10; 1.24)</td>
<td>2.81 x 10^-7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replication</td>
<td>(0.24; 0.22)</td>
<td>1.11</td>
<td>(1.01; 1.23)</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Combined</td>
<td></td>
<td>1.15</td>
<td>(1.10; 1.21)</td>
<td>3.27 x 10^-8</td>
</tr>
<tr>
<td>rs7254272</td>
<td>19p13.3</td>
<td>4069119</td>
<td>A</td>
<td>Discovery</td>
<td>(0.20; 0.18)</td>
<td>1.18</td>
<td>(1.11; 1.26)</td>
<td>4.61 x 10^-7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replication</td>
<td>(0.19; 0.18)</td>
<td>1.13</td>
<td>(1.01; 1.26)</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Combined</td>
<td></td>
<td>1.17</td>
<td>(1.10; 1.23)</td>
<td>4.67 x 10^-8</td>
</tr>
<tr>
<td>rs140522</td>
<td>22q13.33</td>
<td>50971266</td>
<td>T</td>
<td>Discovery</td>
<td>(0.35; 0.32)</td>
<td>1.16</td>
<td>(1.10; 1.22)</td>
<td>2.20 x 10^-8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replication</td>
<td>(0.35; 0.33)</td>
<td>1.10</td>
<td>(1.01; 1.2)</td>
<td>0.025</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Combined</td>
<td></td>
<td>1.15</td>
<td>(1.10; 1.20)</td>
<td>2.70 x 10^-9</td>
</tr>
</tbody>
</table>

**Note:** base pair position; CLL, chronic lymphocytic leukaemia; I², proportion of the total variation due to heterogeneity; OR, odds ratio; RAF, risk allele frequency; SNP, single-nucleotide polymorphism; 95% CI, 95% confidence interval.

RAF is risk allele frequency across all of the discovery and replication data sets, respectively. ORs are derived with respect to the risk allele. Text in bold highlight the P-value in the combined data.
these nine risk variants have generic effects on CLL development rather than tumour progression per se.

Functional annotation of new risk loci. To gain insight into the biological basis underlying the novel association signals, we first evaluated profiles for three histone marks (H3K4me1, H3K27ac marking active chromatin and the repressive mark H3K27me3) at each locus, in GM12878 lymphoblastoid cell line (LCL; ref. 18) as well as primary CLL samples19 (Supplementary Fig. 2). We also examined ATAC-seq profiles from CLL samples and primary B cells as a marker of chromatin accessibility19,20. Since the strongest associated GWAS SNP may not represent the causal variant, we examined signals across an interval spanning all variants in LD \( r^2 > 0.2 \) with the sentinel SNP (based on the 1000 Genomes EUR reference panel). These data revealed regions of active chromatin state at all nine risk loci, in at least one of the cell types. Furthermore, based on the analyses of Hnisz et al.21, five of the loci fall within regions designated as ‘super enhancers’ in either LCLs or CD19 B cells (Supplementary Fig. 2). Overall,
these findings suggest that the risk loci annotate regulatory regions and may, therefore, have an impact on CLL risk through modulation of enhancer or promoter activity.

Given the possibility that SNPs might influence enhancer or promoter activity by causing changes in transcription factor (TF) binding, we next evaluated the SNPs at each GWAS locus based on their overlap with TF-binding sites. In the absence of comprehensive TF chromatin immunoprecipitation sequencing (ChIP-Seq) data from CLL samples, we used regions of chromatin accessibility defined by ATAC-seq data as a surrogate marker for TF binding, identifying 47 SNPs in LD with the sentinel SNPs that also overlapped ATAC-seq peaks. Using motifbreakR to predict whether these SNPs might disrupt TF-binding motifs, we found 478 potentially disrupted motifs, corresponding to 349 TF-binding sites (Supplementary Table 7). Moreover, at 10 of the SNPs, the altered motif matched the TFs bound in ChIP-seq data from the ENCODE project (Supplementary Table 8 and Supplementary Fig. 3).

In particular, we noted that rs13149699 at 4q35 (r² = 0.83 with lead SNP rs57214277) was predicted to disrupt SPI1 binding. In addition, rs13149699 showed evidence of evolutionary constraint, and in LCL ChIP-seq data, the SNP was bound by SPI1 as well as other TFs with roles in B-cell function including IRF4, PAX5, POU2F2 (alias OCT2) and RELA (Supplementary Table 8).

We explored whether there was any association between the genotypes of the nine new risk SNPs and the transcript levels of genes within 1 Mb of each respective variant by performing expression quantitative trait loci (eQTL) analysis using gene expression profiles of 468 CLL cases. In addition, we interrogated publicly accessible expression data on whole blood and LCLs (Supplementary Data 2). There were significant (false discovery rate (FDR) < 0.05) and consistent eQTLs between rs13149699 and rs61904987 and SKAI, rs140522 and ODF3B, and rs41271473 and TYMP.

Biological inference of all CLL risk loci. Given our observation that the nine novel risk loci annotate putative regulatory regions, we sought to examine the epigenetic landscape of CLL risk loci on a broader scale, evaluating the enrichment of both histone
and $D > 0.8$) and determined the overlap between these variants and ENCODE ChIP-seq data. Imposing a $P$ value threshold of $5.37 \times 10^{-4}$ (that is, 0.05/93, based on permutation), we identified a significant enrichment of histone marks associated with active enhancer and promoter elements (HK4Me1, H3K27ac and H3K9ac) as well as actively transcribed regions (H3K36me3). We also identified an over-representation of TF binding for POLR2A, IRF4, RUNX3, NFATC1, STAT5A, PML and WRNIP1 (Fig. 4). In addition, although not statistically significant, POU2F2 showed evidence for enriched binding ($P = 7.78 \times 10^{-4}$). Several of these TFs have established roles in B-cell function. OCT2, IRF4 and RUNX3 have been shown to be targeted for hypomethylation in B cells$^{24}$. MYC is a direct target of IRF4 in activated B cells, with IRF4 being itself a direct target of MYC transactivation. It is noteworthy that variations at IRF4 and $8q24$-MYC are recognized risk factors for CLL$^{2,3}$. Collectively, these findings are consistent with CLL GWAS SNPs mapping within regions of active chromatin state that exert effects on B-cell cis-regulatory networks.

We investigated the genetic pathways between the gene products in proximity to the GWAS SNPs using the LENS pathway tool$^{25}$. These gene products were primarily involved in immune response, BCR-mediated signalling, apoptosis and maintenance of chromosome integrity, as well as interconnectivity between the gene products (Fig. 5). Pathways that were enriched included those related to interferon signalling and apoptosis (Supplementary Data 3).

### Impact of risk SNPs on heritability of CLL

By fitting all SNPs from GWAS simultaneously using Genome-wide Complex Trait Analysis, the estimated heritability of CLL attributable to all common variation is 34% (± 5%), thus having potential to explain 57% of the overall familial risk. This estimate represents the additive variance and, therefore, does not include the potential impact of interactions or dominance effects or gene–environment interactions, having an impact on CLL risk. The currently identified risk SNPs (newly discovered and previously identified) only account for 25% of the additive heritable risk.

### Discussion

Besides providing additional evidence for genetic susceptibility to CLL, the new and established risk loci identified further insights into the biological basis of CLL development. These loci annotate genes that participate in interconnecting cellular pathways, which are central to B-cell development. In particular, we note the involvement of BCR-mediated signalling with immune responses and apoptosis. Importantly, gene discovery initiatives can have an impact on the successful development of new therapeutic agents$^{26}$. In this respect it is notable that Ibrutinib$^{27}$ (a BTK inhibitor) and Idelalisib$^{28}$ (a PI3KCD inhibitor) mediate their effects through interference of BCR signalling, and Venetoclax$^{29}$ targets the anti-apoptotic behaviour of BCL-2. The power of our GWAS to identify common alleles conferring relative risks of 1.2 or greater (such as the rs35923643 variant) is high (~80%). Hence, there are unlikely to be many additional SNPs with similar effects for alleles with frequencies greater than 0.2 in populations of European ancestry. In contrast, our analysis had limited power to detect alleles with smaller effects and/or MAF <0.1. Hence, further GWAS studies in concert with functional analyses should lead to additional insights into CLL biology and afford the prospect of development of novel therapies.

### Methods

#### Ethics

Figure 5 | Hive Plot of common protein–protein interactions in CLL. Each arm represents a functional annotation term, each arc represents an interaction between two proteins and the distance from the centre of the plot corresponds to a greater number of protein–protein interactions (higher degree of the node). The left arm represents proteins annotated as being involved in BCR signalling; the top arm represents proteins annotated as immune response; the right arm represents proteins involved in apoptosis; and the bottom arm represents proteins involved in DNA damage and chromosomal integrity. Selected proteins known to be involved in CLL risk are shown.

**Genome-wide association studies.** The meta-analysis was based on six GWAS.10–17 (Supplementary Tables 1 and 2). Briefly, the six GWAS comprised—UK-CLL1: 517 CLL cases and 2,698 controls; UK-CLL2: 1,403 CLL cases, 2,501 controls; Genetic Epidemiology of CLL (GEC) Consortium: 396 CLL cases and 296 controls; NHL GWAS Consortium: 1,851 CLL cases and 6,649 controls; UCSF: 214 CLL cases, 751 controls; Utah: 331 CLL cases, 420 controls.

**Quality control of GWAS.** Standard quality-control measures were applied to the GWAS.10 Specifically, individuals with low call rate (<95%) as well as all individuals evaluated to be of non-European ancestry (using the HapMap version 2 CEU, JPT/CHB and YRI populations as a reference) were excluded. For apparent first-degree relative pairs, we removed the control from a case–control pair; otherwise, we excluded the individual with the lower call rate. SNPs with a call rate <95% were excluded as were those with a MAF <0.01 or displaying significant deviation from Hardy–Weinberg equilibrium (that is, \(P < 10^{-6}\)). GWAS data were imputed to >10 million SNPs with the IMPUTE2 v2.3 software using a merged reference panel consisting of data from 1000 Genomes Project (phase 1 integrated release 3 March 2012)10 and UK10K (ref. 11). Genotypes were aligned to the positive strand in both imputation and genotyping. Imputation was conducted separately for each study, and in each the data were pruned to a common set of SNPs between cases and controls before imputation. We set thresholds for imputation quality to retain potential risk variants with MAF >0.005 for validation. Poorly imputed SNPs defined by an information measure <0.80 were excluded. Tests of association between imputed SNPs and CLL was performed using logistic regression under an additive genetic model in SNPTESTv2.5 (ref. 33). The adequacy of the case–control matching and possibility of differential genotyping of cases and controls were formally evaluated using Q–Q plots of test statistics (Supplementary Fig. 1). The inflation factor \(λ\) was based on the 90% least-significant SNPs.10 Where appropriate, principal components, generated using common SNPs, were included in the analysis to limit the effects of cryptic population stratification that otherwise might cause inflation of test statistics.

Eigenvectors for the GWAS data sets were inferred using smartpca (part of EIGENSOFT15) by merging cases and controls with Phase II HapMap samples.

**Replication studies and technical validation.** The 16 SNPs in the most promising loci were taken forward for de novo replication (Supplementary Table 2). The UK replication series comprised 645 cases collected through the NCLLC and Leiceste Haematology Tissue Bank and 2,341 controls comprised 2,780 healthy individuals ascertained through the National Study of Colorectal Cancer (1999–2006; ref. 36). These controls were the spouses or unrelated friends of individuals with malignancies. None had a personal history of malignancy at the time of ascertainment. Both cases and controls were British residents and had self-reported European ancestry. The Mayo replication series comprised 407 newly diagnosed cases and 1,207 clinic-based controls from the Mayo Clinic CLL case-control study37. The eligibility criteria of the cases were age 20 years and older, consented within 9 months of their initial diagnosis at presentation to Mayo Clinic and no history of HIV. The eligibility criteria for the controls were age 20 years and older, a resident of Minnesota, Iowa or Wisconsin at the time of appointment at Mayo Clinic, no history of lymphoma or leukaemia and no history of HIV infection. Controls were frequency matched to the regional case distribution on 5-year age group, sex and geographic area. In silico replication was performed in 444 cases and 609 controls from International Cancer Genome Consortium (ICGC), and 226 cases and 228 controls from the WHI study38,39.

The fidelity of imputation as assessed by the concordance between imputed and directly genotyped SNPs was examined in a subset of samples (Supplementary Table 5). Replication genotyping of UK samples was performed using competitive allele-specific PCR KASPar chemistry (LGC, Hertfordshire, UK); replication genotyping of Mayo samples was performed using Sequenom MassARRAY (Sequenom Inc., San Diego, CA, USA). Primers are listed in Supplementary Table 9. Call rates for SNP genotypes were >95% in each of the replication series. To ensure the quality of genotyping in all assays, at least two negative controls and duplicate samples (showing a concordance of >99%) were genotyped at each centre. To exclude technical artefacts in genotyping, we performed cross-platform validation of 96 samples and sequenced a set of 96 randomly selected samples from each case and control series to confirm genotyping accuracy. Assays were found to be performing robustly; concordance was >99%.

**Meta-analysis.** Meta-analyses were performed using the fixed-effects inverse-variance method based on the \(β\) estimates and s.e.'s from each study using META v1.6 (ref. 40). 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. Using the meta-analysis summary statistics and LD correlations from a reference panel of the 1000 Genomes Project combined with UK10k we used Genome-wide Complex Trait Analysis to estimate the polygenic variance (that is, heritability) ascribable to all genotyped and imputed GWAS SNPs. SNPs were excluded based on low MAF (MAF <0.01), poor imputation (info score <0.4) and evidence of departure from Hardy Weinberg Equilibrium (HWE) (P <0.05). Individuals were excluded for poor imputation and where two individuals were closely related. A genetic relationship matrix of pairs of samples was used as input for the restricted maximum likelihood analysis to estimate the heritability explained by the selected set of SNPs. To transform the estimated heritability to the liability scale, we used the lifetime risk\textsuperscript{5,56} for CLL, which is estimated to be 0.006 by SEER (http://seer.cancer.gov/statfacts/html/ly.html). The variance of the risk distribution due to the identified risk loci was calculated as described by Pharoah et al.\textsuperscript{7}, assuming that the relative risk when a first-degree relative has CLL is 8.5 (ref. 1).

Pathway analysis. To investigate the interaction between the gene products of the GWAS SNPs, we performed a pathway analysis. We selected the closest coding genes for the lead-associated SNPs and then performed pathway analysis using the LENS tool\textsuperscript{22}, which identifies gene product and protein–protein interactions from HPRD\textsuperscript{58} and BioGRID\textsuperscript{59}. Enrichment of pathways was assessed using Fisher’s exact test, comparing the overlap of the genes in the network with the genes in the pathway. Pathway data were obtained from REACTOME\textsuperscript{30}. Cytoscape was used to perform network analyses\textsuperscript{61} and the Hive Plot was drawn using HiveR (academic.depauw.edu/~hanson/HiveR/HiveR.html).

Data availability. Genotype data that support the findings of this study have been deposited in the database of Genotypes and Phenotypes (dbGaP) under accession code phs000802.v2.p1 and in the European Genome-phenome Archive (EGA) under accession codesEGAS00000000909, EGAD00000001915, EGAD000000001008, EGAD00000000022 and EGAD00000000024.

Transcriptional profiling data from the MuTHER consortium that support the findings of this work have been deposited in the European Bioinformatics Institute (Part of the European Molecular Biology Laboratory, EMBL-EBI) under accession code E-TABM-1140. Data from Blood eQTL have been deposited in the EBI-EMBL under accession codes E-TABM-1036, E-MTAB-945 and E-MTAB-1708. GTEx data are deposited in dbGaP under accession code phs000424.v6.p1. The remaining data are contained within the paper and its Supplementary files or are available from the authors upon reasonable request.

References


supported by a Leukemia and Lymphoma Society Career Development Award, the Bernstein Family Fund for Leukemia and Lymphoma Research and the National Institutes of Health (K08CA134919), National Center for Advancing Translational Science (UL1 TR001135). HPFS—The HPFS was supported in part by National Institutes of Health grants CA167552, CA149445 and CA098122. We would like to thank the participants and staff of the Health Professionals Follow-up Study for their valuable contributions as well as the following state cancer registries for their help: A.I., AL, AK, AR, CA, CO, CT, DC, DE, FL, GA, ID, IL, IN, IA, KY, LA, ME, MD, MA, MI, MN, MS, MO, MT, NE, NV, NH, NJ, NM, NY, NC, ND, OH, OK, OR, PA, RI, SC, TN, TX, UT, VA, WA, WY. We assume full responsibility for analyses and interpretation of these data. Iowa-Mayo SPOR—NCI Specialized Programs of Research Excellence (SPORE) in Human Cancer (5P01CA97274); National Cancer Institute (P30 CA086682, P30 CA15083); Henry J. Predomin Foundation. Italian G—Fondazione per lo studio dei canceri (FPC) (grant number 2012); Fondazione Banco di Sardegna 2010-2012 and Regione Autonoma della Sardegna (LR7 CRP-59812/2012; MGE). Mayo Clinic Case—Control—National Institutes of Health (RO1 CA92153); National Cancer Institute (P30 CA15083). MCCS—the Melbourne Collaborative Cohort Study recruitment was funded by VicHealth and Cancer Council Victoria. The MCCS was further supported by Australian NHMRC grants 209057, 251553 and 504711 and by infrastructure provided by Cancer Council Victoria. Cases and their vital status were ascertained through the Victorian Cancer Registry (VCR). MD Anderson—Institutional support to the Center for Translational and Public Health Genomics. MSKCC—Geoffrey Beene Cancer Research Grant, Lymphoma Foundation (LE5541); Barbara K. Lipman Lymphoma Research Fund (74419); Robert and Kate Niehaus Clinical Cancer Genetics Research Initiative (57476); U10 HG007033. ENCOCER—U10 CA178800. T2D-GENES—Intramural Research Program of the National Cancer Institute, National Institutes of Health and Public Health Service (N01-PC-65064, N01-PC-67008, N01-PC-67009, N01-PC-67610, N01-PC-71105). NIH—The NIH was supported in part by National Institutes of Health grants CA186107, CA87969, CA49449, CA149445 and CA99812. We would like to thank the participating Health Professionals’ Study Health Councils for their valuable contributions as well as the following state cancer registries for their help: A.L., A.Z., A.R., C.A., C.O., C.T., D.E., F.L., G.A., I.D., I.L., I.N., I.A., K.Y., L.A., L.E., M.D., M.A., M.I., N.E., N.H., N.J., N.Y., N.C., N.D., O.H., O.K., O.R., P.A., R.I., S.C., T.N., T.X., V.A., W.A. and W.Y. The authors assume full responsibility for analyses and interpretation of these data. NSW—was supported by grants from the Australian National Health and Medical Research Council (NHMRC) and the University of Sydney Faculty of Medicine. NYU-WHS—National Cancer Institute (RO1 CA988661, P30 CA016807); National Institute of Environmental Health Sciences (ES000260). PLCO—This research was supported by the Intramural Research Program of the National Cancer Institute and by contracts from the Division of Cancer Prevention, National Cancer Institute, NIH, DHEW. SCALE—Swedish Cancer Society (2009-659). Stockholm Cancer Registry (20140209) and the Strategic Research Program in Epidemiology at Karolinska Institute. Swedish Cancer Society grant (02 6661). National Institutes of Health (SR01 CA09669-02); Plan Denmark. UCSF2—the UCSF studies were supported by the NCI, National Institutes of Health, CA014682, CA154435, CA45614, CA89745, CA7014. The collection of cancer incidence data used in this study was supported by the California Department of Health Services as part of the state's cancer reporting program mandated by California Health and Safety Code Section 103885; the National Cancer Institute’s Surveillance, Epidemiology and End Results Program under contract HHSN26120100140C awarded to the Cancer Prevention Institute of California, contract HHSN261201000035C awarded to the University of Southern California, and contract HHSN261201000034C awarded to the Public Health Institute for the Centers for Disease Control and Prevention’s National Program of Cancer Registries, under agreement #1U58 DP000807-01 awarded to the Public Health Institute. The ideas and opinions expressed herein are those of the investigators who contributed to the generation of the data available in the www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113. We are grateful to all investigators and all the patients and individuals for their participation. We also thank the clinicians, other hospital staff and study staff that contributed to the blood sample and data collection for this study.

Author contributions

Additional information
Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications
Competing financial interests: The authors declare no competing financial interests. Reprints and permission information is available online at http://npq.nature.com/reprintsandpermissions/
How to cite this article: Law, P. J. et al. Genome-wide association analysis implicates dysregulation of immunity genes in chronic lymphocytic leukaemia. Nat. Commun. 8, 14175 doi: 10.1038/ncomms14175 (2017).
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

1 Division of Genetics and Epidemiology, The Institute of Cancer Research, London SW7 3RP, UK. 2 Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland 20892, USA. 3 Department of Internal Medicine, Huntsman Cancer Institute, University of Utah School of Medicine, Salt Lake City, Utah 84112, USA. 4 Department of Epidemiology, School of Public Health and Comprehensive Cancer Center, University of Alabama at Birmingham, Birmingham, Alabama 35233, USA. 5 Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA. 6 Northern Institute for Cancer Research, Newcastle University, Newcastle upon Tyne NE2 4HH, UK. 7 Center for Chronic Immunodeficiency, University Medical Center Freiburg, Freiburg, Baden-Württemberg 78108, Germany. 8 Institut d’Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Hospital Clinic, Barcelona 08036, Spain. 9 Registre des hémopathies malignes de la Gironde, Institut Bergonié, Inserm U1219 EPICENE, 33076 Bordeaux, France. 10 Epidemiology of Childhood and Adolescent Cancers Group, Inserm, Center of Research in Epidemiology and Statistics Sorbonne Paris Cité (CRESS), Paris F-94807, France. 11 Université Paris Descartes, Paris 75270, France. 12 Epidemiology Research Program, American Cancer Society, Atlanta, Georgia 30303, USA. 13 Grupo de Medicina Xenomica, Universidade de Santiago de Compostela, Centro Nacional de Genotipado (CeGen-PRB2-ISCI1I), CIBERER, 15782 Santiago de Compostela, Spain. 14 Channing Division of Network Medicine, Department of Medicine, Brigham and Women’s Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA. 15 Ernest and Helen Scott Haematological Research Institute, University of Leicester, Leicester LE2 7LX, UK. 16 Department of Preventive Medicine, USCK School of Medicine, University of Southern California, Los Angeles, California 90033, USA. 17 Norris Comprehensive Cancer Center, USCK School of Medicine, University of Southern California, Los Angeles, California 90033, USA. 18 Unit of Clinical Epidemiology, Department of Medicine Solna, Karolinska Institutet, Hematology Center, Karolinska University Hospital, Stockholm 17176, Sweden. 19 The Royal Marsden NHS Foundation Trust, London SM2 8PT, UK. 20 Genome Sciences Centre, BC Cancer Agency, Vancouver British Columbia, Canada V5Y2L3. 21 Department of Biomedical Physiology and Kinesiology, Simon Fraser University, Burnaby, British Columbia, Canada V5A1S6. 22 Haematological Sciences, Medical School, Newcastle University, Newcastle-upon-Tyne NE2 4HH, UK. 23 Centre for Big Data Research in Health, University of New South Wales, Sydney, New South Wales 2052, Australia. 24 Department of Haematology, Royal Viscount Inflammatory, Newcastle upon Tyne NE1 4LP, UK. 25 Department of Public Health, Clinical and Molecular Medicine, University of Cagliari, Monserrato, Cagliari 09042, Italy. 26 Department of Environmental Health Sciences, Yale School of Public Health, New Haven, Connecticut 06520, USA. 27 Cancer Epidemiology Centre, Cancer Council Victoria, Melbourne, Victoria 3004, Australia. 28 Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, University of Melbourne, Melbourne, Victoria 3010, Australia. 29 Westat, Rockville, Maryland 20850, USA. 30 Division of Hematology, Mayo Clinic, Rochester, Minnesota 55905, USA. 31 Department of Medicine, Mayo Clinic, Rochester, Minnesota 55905, USA. 32 Department of Epidemiology Research, Division of Health Surveillance and Research, Statens Serum Institut, 2300 Copenhagen, Denmark. 33 Department of Medicine, Stanford University School of Medicine, Stanford, California 94305, USA. 34 Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, 75105 Uppsala, Sweden. 35 Department of Epidemiology Research Program, American Cancer Society, Atlanta, Georgia 30303, USA. 36 Department of Internal Medicine, Carver College of Medicine, University of Iowa, Iowa City, Iowa 52242, USA. 37 Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California 94118, USA. 38 Division of Endocrinology, Diabetes and Metabolism, Ohio State University, Columbus, Ohio 43210, USA. 39 Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA. 40 Cancer Epidemiology Research Programme, Catalan Institute of Oncology-IDIBELL, L’Hospitalet de Llobregat, Barcelona 08908, Spain. 41 CIBER de Epidemiología y Salud Pública (CIBERESP), Barcelona 08036, Spain. 42 The Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA. 43 International Agency for Research on Cancer, Lyon 69372, France. 44 Registre des Hémopathies Malignes de Côte d’Or, University of Burgundy and Dijon University Hospital, Dijon 21070, France. 45 Department of Computational Biology, St Jude Children’s Research Hospital, Memphis, Tennessee 38105, USA. 46 Department of Family Medicine and Public Health Sciences, Wayne State University, Detroit, Michigan 48201, USA. 47 School of Public Health, Imperial College London, London W2 IPG, UK. 48 MRC-PHE Centre for Environment and Health, School of Public Health, Imperial College London, London W2 IPG, UK. 49 Human Genetics Foundation, 10126 Turin, Italy. 50 Institute for Risk Assessment Sciences, Utrecht University, Utrecht 3508 TD, The Netherlands. 51 Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht 3584 CX, The Netherlands. 52 Genetic Epidemiology Laboratory, Department of Pathology, University of Melbourne, Melbourne, Victoria 3010, Australia. 53 Epidemiology of Childhood and Adolescent Cancers Group, Inserm, Center of Research in Epidemiology,
and Statistics Sorbonne Paris Cité (CRESS), Paris F-94807, France. 54 Université Paris Descartes, 75270 Paris, France. 55 Cancer Control Research, BC Cancer Agency, Vancouver, British Columbia, Canada V5Z1L3. 56 School of Population and Public Health, University of British Columbia, Vancouver, British Columbia, Canada V6T1Z3. 57 Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts 02115, USA. 58 Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts 02115, USA. 59 Department of Biomedical Science, University of Cagliari, Monseerrato, Cagliari 09042, Italy. 60 Department of Haematology, Queen Elizabeth Hospital, Gateshead NE9 6SX, UK. 61 Interdisciplinary Department of Medicine, University of Bari, Bari 70124, Italy. 62 Department of Molecular and Clinical Cancer Medicine, University of Liverpool, Liverpool L69 3BX, UK. 63 Environmental and Occupational Epidemiology Unit, Cancer Prevention and Research Institute (ISPO), Florence 50139, Italy. 64 Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. 65 Carolina Center for Genome Sciences, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. 66 Queens Centre for Haematology and Oncology, Castle Hill Hospital, Hull and East Yorkshire NHS Trust, Cottingham HU16 5JQ, UK. 67 Department of Haematology, Birmingham Heartlands Hospital, Birmingham B9 5SS, UK. 68 Division of Cancer and Genetics, School of Medicine, Cardiff University, Cardiff CF14 4XN, UK. 69 Cardiff and Vale National Health Service Trust, Heath Park, Cardiff CF14 4XW, UK. 70 Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Kingdom of Saudi Arabia. 71 Division of Molecular Pathology, The Institute of Cancer Research, London SW7 3RP, UK. 72 Unidad de Hematología, Hospital Clinic, IDIBAPS, Universitat de Barcelona, Barcelona 08036, Spain. 73 Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota 55905, USA. * These authors contributed equally to this work. ** These authors jointly supervised the work.