Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia

Genome-wide association studies (GWAS) have previously identified 13 loci associated with risk of chronic lymphocytic leukemia or small lymphocytic lymphoma (CLL). To identify additional CLL susceptibility loci, we conducted the largest meta-analysis for CLL thus far, including four GWAS with a total of 3,100 individuals with CLL (cases) and 7,667 controls.

In the meta-analysis, we identified ten independent associated SNPs in nine new loci at 10q23.31 (ACTA2 or FAS (ACTA2/FAS), \( P = 1.22 \times 10^{-14} \)), 18q21.33 (BCL2, \( P = 7.76 \times 10^{-11} \)), 11p15.5 (C11orf21, \( P = 2.15 \times 10^{-10} \)), 4q25 (LEFT1, \( P = 4.24 \times 10^{-10} \)), 2q33.1 (CASP10 or CASP8 (CASP10/CASP8), \( P = 2.50 \times 10^{-9} \)), 9p21.3 (CDKN2B-AS1, \( P = 1.27 \times 10^{-8} \)), 18q21.32 (PMAIP1, \( P = 2.51 \times 10^{-8} \)), 15q15.1 (BMF, \( P = 2.71 \times 10^{-10} \)) and 2p22.2 (QPCT, \( P = 1.68 \times 10^{-8} \)), as well as an independent signal at an established locus (2q13, ACOXL, \( P = 2.08 \times 10^{-10} \)). We also found evidence for two additional promising loci below genome-wide significance at 8q22.3 (ODF1, \( P = 5.40 \times 10^{-8} \)) and 5p15.33 (TERT, \( P = 1.92 \times 10^{-7} \)). Although further studies are required, the proximity of several of these loci to genes involved in apoptosis suggests a plausible underlying biological mechanism.

RESULTS

As part of a larger initiative in non-Hodgkin’s lymphoma (NHL) (called the NHL GWAS), we genotyped 2,343 cases and 2,854 controls of European descent from 22 studies using the Illumina OmniExpress Beadchip (Online Methods and Supplementary Table 1). Of those 5,197 subjects, 94% passed rigorous quality-control criteria (Online Methods and Supplementary Table 2), and 549,934 SNPs successfully passed quality-control criteria with a median call rate >98%. We also used genotype data previously generated on the Illumina Omni2.5 from an additional 3,536 controls and 1 case from three studies, giving a total of 2,179 cases and 6,221 controls for the analysis of the NHL GWAS data (Supplementary Table 3).

In the NHL GWAS (stage 1) analysis, we observed an enrichment of SNPs with small \( P \) values compared to the null distribution with a \( \lambda \) value of 1.026 in the quantile-quantile plot (Supplementary Fig. 1). After exclusion of previously established loci, an excess of small \( P \) values still remained, suggesting there are additional susceptibility loci. In our stage 1 analysis, we observed SNPs from ten unique loci (defined as separated by at least 50 kb and with linkage disequilibrium (LD), \( r^2 < 0.05 \)) that reached genome-wide significance (\( P < 5 \times 10^{-8} \)), including eight established loci and two new loci (Supplementary Fig. 2).

We then performed a meta-analysis of the NHL GWAS data with three other independent CLL GWAS\(^5,9\) that had a combined total of 921 cases and 1,446 controls (stage 2; Supplementary Tables 1 and 3). Because these other CLL GWAS studies were conducted on different commercial SNP microarrays, we imputed common SNPs from the 1000 Genomes Project\(^{10}\) using IMPUTE2 (ref. 11) (Online Methods and Supplementary Table 4). In the meta-analysis of the data from stages 1 and 2, associations for all 13 established loci showed a consistent direction of effect with previously reported studies, and 10 loci achieved \( P < 5 \times 10^{-8} \) (Supplementary Table 5). However, two previously established loci, 15q25.2 and 19q13.3, were only nominally significant in the meta-analysis (\( P = 0.03 \) and \( P = 0.008 \), respectively), and we found no significant association in stage 1 for the 15q25.2 locus (\( P = 0.10 \)). A suggestive locus on 18q21.1 that did not meet genome-wide significance in previous studies\(^{12}\) was also nominally significant here (\( P = 5.06 \times 10^{-4} \)). From the meta-analysis of stages 1 and 2, we identified ten promising SNPs in the eight new loci and one promising SNP in an established locus that we carried forward for a \textit{de novo} replication in stage 3; this stage included an additional 392 cases and 4561 controls and \textit{in silico} replication in an independent CLL GWAS with 396 cases and 311 controls (Online Methods and Supplementary Tables 1.3 and 4).

Significant new loci

Seven of the ten SNPs in new loci reached genome-wide significance in the meta-analysis of all three stages: 10q23.31 (ACTA2/FAS, \( P = 1.22 \times 10^{-14} \)), 18q21.33 (BCL2, \( P = 2.66 \times 10^{-12} \)), 11p15.5 (C11orf21, \( P = 2.15 \times 10^{-10} \)), 4q25 (LEFT1, \( P = 4.24 \times 10^{-10} \)), 2q33.1 (CASP10/CASP8 (CASP10/CASP8), \( P = 2.50 \times 10^{-9} \)), 9p21.3 (CDKN2B-AS1, \( P = 1.27 \times 10^{-8} \)), and 18q21.32 (PMAIP1, \( P = 2.51 \times 10^{-8} \)) (Table 1 and Fig. 1). Further, within the 18q21.33 locus, a second SNP (rs4987855) in low LD (\( r^2 = 0.01 \)) with rs4987855 and located only 372 bp away also reached genome-wide significance (Table 1; \( P = 7.76 \times 10^{-11} \)); we determined this SNP to be independent in conditional analyses (\( P_{\text{conditional}} = 3.87 \times 10^{-7} \); Table 2).

A full list of authors and affiliations appears at the end of the paper.

Received 16 November 2012; accepted 2 May 2013; published online 16 June 2013; doi:10.1038/ng.2652
To explore these regions in greater detail and identify additional loci that we may have missed using just the genotyped SNPs in stage 1, we imputed the stage 1 data of our NHL GWAS using 1000 Genomes Project data (February 2012 release) and performed a meta-analysis of the results from stages 1 and 2. The most significant SNPs at three of our new loci, 10q23.31 (rs2147420) 18q21.33 (rs4987856) and 4q25 (rs2003869), were highly correlated ($r^2 \geq 0.95$) with our strongest genotyped SNPs, rs4406737, rs4987858 and rs898518, respectively (Supplementary Table 6). We found only modest correlations ($r^2$ range 0.18–0.58) for the most significant imputed SNPs at 11p15.5 (rs2521269), 2q33.1 (rs11688943) and 9p21.3 (rs1359742) with our strongest genotyped SNPs in each of the respective regions. The most significant of the imputed SNPs at 18q21.32 (rs35741867) seemed to be independent of our strongest genotyped SNP (rs4368253, $r^2 = 0.003$, $P_{conditional} < 7.89 \times 10^{-7}$ for both SNPs), suggesting a possible second, independent signal (Table 2).

Meta-analysis of our imputed scan data identified two new loci, 15q15.1 (BMF, $P = 2.71 \times 10^{-10}$) and 2p22.2 (QPCT, $P = 1.68 \times 10^{-8}$) (Table 1 and Fig. 1). In addition, although our genotyped SNP at 5p15.33 (TERT, rs10069690, $P = 1.92 \times 10^{-7}$) (Supplementary Table 7)

### Table 1: Association results for new susceptibility loci

| Chr. | Nearest gene(s) | SNP | Position | Risk allelea | Other allele | RAF | Stage | OR (95% CI) | $P$  |
|------|----------------|-----|----------|--------------|--------------|-----|-------|------------|------|
| Newly identified loci | | | | | | | | | |
| 10q23.31 | ACTA2/FAS | rs4406737 | 90,749,704 | G | A | 0.57 | Stage 1 | 1.30 (1.21–1.40) | $3.30 \times 10^{-12}$ |
| | | | | | | | Stage 2 | 1.17 (1.03–1.32) | 0.01 |
| | | | | | | | Stage 3 | 1.27 (1.06–1.52) | 0.007 |
| | | | | | | | Combinedb | 1.27 (1.19–1.33) | $1.22 \times 10^{-14}$ |
| 18q21.33 | BCL2 | rs4987855c | 58,944,529 | G | A | 0.91 | Stage 1 | 1.47 (1.28–1.69) | $5.51 \times 10^{-8}$ |
| | | | | | | | Stage 2 | 1.47 (1.18–1.85) | 0.0007 |
| | | | | | | | Stage 3 | 1.43 (1.12–1.82) | 0.004 |
| | | | | | | | Combinedb | 1.47 (1.32–1.61) | $2.66 \times 10^{-12}$ |
| 11p15.5 | C11orf21, TSPAN32 | rs7944004 | 2,267,728 | T | G | 0.49 | Stage 1 | 1.43 (1.26–1.63) | $2.67 \times 10^{-8}$ |
| | | | | | | | Stage 2 | 1.24 (0.98–1.56) | 0.07 |
| | | | | | | | Stage 3 | 1.52 (1.17–1.97) | 0.002 |
| | | | | | | | Combinedb | 1.41 (1.27–1.56) | $7.76 \times 10^{-11}$ |
| 4q25 | LEF1 | rs898518c | 109,236,273 | A | C | 0.59 | Stage 1 | 1.15 (1.02–1.32) | 0.03 |
| | | | | | | | Stage 2 | 1.27 (1.11–1.45) | 0.0006 |
| | | | | | | | Combinedb | 1.20 (1.13–1.27) | $2.15 \times 10^{-10}$ |
| 2q33.1 | CASP10/CASP8 | rs3769825 | 201,819,625 | T | C | 0.45 | Stage 1 | 1.15 (1.03–1.32) | 0.01 |
| | | | | | | | Stage 2 | 1.22 (1.07–1.40) | 0.004 |
| | | | | | | | Combinedb | 1.19 (1.12–1.25) | $2.50 \times 10^{-9}$ |
| 9p21.3 | CDKN2B-AS1 | rs1679013 | 22,196,987 | C | T | 0.52 | Stage 1 | 1.15 (1.09–1.27) | $3.65 \times 10^{-5}$ |
| | | | | | | | Stage 2 | 1.24 (1.08–1.41) | 0.002 |
| | | | | | | | Stage 3 | 1.18 (1.02–1.37) | 0.03 |
| | | | | | | | Combinedb | 1.19 (1.12–1.27) | $1.27 \times 10^{-8}$ |
| 18q21.32 | PMAIP1 | rs4368253 | 55,773,267 | C | T | 0.69 | Stage 1 | 1.22 (1.14–1.32) | $2.72 \times 10^{-8}$ |
| | | | | | | | Stage 2 | 1.22 (1.08–1.39) | 0.003 |
| | | | | | | | Combinedb | 1.22 (1.15–1.30) | $2.71 \times 10^{-10}$ |
| 15q15.1 | BMF | rs8024033d | 38,190,949 | C | G | 0.51 | Stage 1 | 1.29 (1.18–1.40) | $8.23 \times 10^{-9}$ |
| | | | | | | | Stage 2 | 1.10 (0.95–1.28) | 0.21 |
| | | | | | | | Combinedb | 1.24 (1.15–1.33) | $1.68 \times 10^{-8}$ |
| 2p22.2 | QPCT, PRKD3 | rs3770745d | 37,449,593 | T | C | 0.22 | Stage 1 | 1.43 (1.28–1.56) | $9.76 \times 10^{-13}$ |
| | | | | | | | Stage 2 | 1.45 (1.23–1.72) | $9.39 \times 10^{-6}$ |
| | | | | | | | Stage 3 | 1.32 (1.08–1.59) | 0.007 |
| | | | | | | | Combinedb | 1.41 (1.30–1.52) | $2.08 \times 10^{-18}$ |

*a The risk allele is the allele corresponding to the estimated OR. b Number of cases and controls in the joint analysis of stages 1–3: rs4406737 (3,481/12,170), rs4987858 (3,885/12,446), rs4987852 (3,880/12,497), rs7944004 (3,869/12,476), rs898518 (3,879/12,441), rs3769825 (3,885/12,471), rs1679013 (3,482/12,148), rs4368253 (3,885/12,473), rs8024033 (3,097/7,665), rs3770745 (3,097/7,663) and rs13401811 (3,839/12,264). c For the International Cancer Genome Consortium (ICGC) study in stage 3, results for proxy SNPs were provided (rs4987856, proxy for rs4987858, $r^2 = 1.0$; rs7698317, proxy for rs898518, $r^2 = 1.0$; rs1554005, proxy for rs13401811, $r^2 = 1.0$). d Identified from the 1000 Genomes meta-analysis of stages 1 and 2 with imputation information $>0.9$ in the NHL GWAS. e New independently associated SNP in a previously reported susceptibility locus at 2q13. f Chr., chromosome; RAF, risk allele frequency in controls; OR, per-allele OR adjusted for age, sex and significant principal components. Bolded text indicates the combined meta-analysis results.
did not reach genome-wide significance, we did observe an imputed SNP in this region that reached genome-wide significance (rs7705526, \( P = 3.75 \times 10^{-8} \)). We found another promising locus at 8q22.3 (ODF1, \( P = 5.40 \times 10^{-8} \)) (Supplementary Table 7). Additional studies are needed to confirm these findings, particularly the signal on 5p15.33, which is already known to harbor risk variants for multiple cancers\(^{13-20}\).

**Figure 1** Association results, recombination hotspots and LD plots for the regions newly associated with CLL. (a–i) Top, association results of GWAS data from the stage 1 NHL GWAS (gray diamonds), the stage 2 combined data (blue diamond), the stage 3 combined data (purple diamond) and the combined data from stages 1–3 (red diamond) plotted against the –log\(_{10}\) \( P \) values (left y axis). Overlaid are the likelihood ratio statistics (right y axis) estimating putative recombination hotspots across the region on the basis of five unique sets of 100 randomly selected control samples. Bottom, LD heat map based on \( r^2 \) values from the total control populations for all SNPs included in the GWAS. Shown are results for the 10q23.31 (a), 18q21.33 (b), 11p15.5 (c), 4q25 (d), 2q33.1 (e), 9p21.3 (f), 18q21.32 (g), 15q15.1 (h) and 2p22.2 (i) regions.
An examination of established loci revealed a new SNP in 2q13 (BCL2L11, rs13401811, \( P = 6.09 \times 10^{-17} \); Table 1 and Fig. 2) that was independent of the previously reported SNP. After conditioning on the established 2q13 SNP (rs17483466, \( r^2 = 0.02 \)), the new SNP rs13401811 remained strongly associated with CLL risk (\( P_{\text{conditional}} = 1.60 \times 10^{-12} \); Table 2). We found a putative second signal at the established 2q37.3 locus (Supplementary Table 5; rs7578199, \( P = 5.39 \times 10^{-7} \)) that was in low LD with (\( r^2 = 0.01 \)), and independent of, the previously reported SNP rs757978 (\( P_{\text{conditional}} = 6.10 \times 10^{-6} \); Table 2), although rs7578199 was not genome-wide significant. We observed another possible second signal on 6p21.32 (Supplementary Table 5; HLA, rs9273363, \( P = 2.24 \times 10^{-10} \)). rs9273363 showed some evidence
of conditional independence with the originally reported SNPs ($r^2 \leq 0.25$, $P_{\text{conditional}} \leq 3.50 \times 10^{-5}$; Table 2). However, rs9273363 may be part of a shared HLA haplotype, and thus accurate HLA typing is needed to further clarify its level of independence. In addition, we observed a SNP at 15q21.3 (Supplementary Table 5; rs11636802, $P = 1.68 \times 10^{-11}$) that had stronger statistical significance than the previously reported SNP, rs7169431 ($P = 1.72 \times 10^{-05}$). Although only modestly correlated ($r^2 = 0.16$), rs11636802 explained all of the risk associated with rs7169431 in a conditional analysis (Table 2), suggesting that this SNP may be a better marker for the locus.

Explanations of variance

Heritability analysis indicated that the ten independent SNPs in our new loci together with the new independent SNP at 2q13 (Table 1) explain approximately 5% more familial risk in addition to the ~12% explained by the established loci. When we explored the contribution of all common variants to the genetic heritability of CLL (using a method that estimates the variance explained by fitting all genotyped autosomal SNPs simultaneously21,22; Online Methods)21,22, we estimated that common SNPs have the potential to explain up to ~46% of the familial risk, suggesting that more common loci, probably of small effects, still remain to be discovered. However, the analysis also implied that common SNPs probably do not explain all of the familial risk, and other factors, such as uncommon SNPs with modest effects or rare highly penetrant variants, probably also have a role.

Biological inferences

Five of the new loci (10q23.31, 18q21.33, 2q33.1, 18q21.32 and 15q15.1) identified in this study, as well as the new SNP at the established 2q13 locus, are located in or near genes involved in apoptosis. rs4406737 is located on 10q23.31 between the first and second exons of FAS, a member of the tumor necrosis factor receptor superfamily that has a crucial role in the initiation of the signaling cascade of the caspase family in apoptosis. Mutations in FAS leading to defective FAS-mediated apoptosis have been documented in inherited lymphoproliferative disorders associated with autoimmunity23,24, and families with germline FAS mutations have a substantially increased risk of other lymphoma subtypes25.

The two newly identified SNPs at 18q21.33 (rs4987855 and rs4987852) map to the 3’ UTR of CBL2 (B-cell CLL/lymphoma 2), which encodes an essential outer mitochondrial membrane protein that blocks lymphocyte apoptosis. Constitutive expression of BCL2 through (t(14;18) and other translocations is common in follicular lymphomas, but the (t(14;18) translocation is also found in CLL, albeit rarely26. Both SNPs are located within a narrow region of BCL2 in which the majority of (t(14;18) translocation breakpoints occur27. rs4987855 is in LD with a SNP (rs4987856, $r^2 = 1.0$) that is located within 200 bp of a putative microRNA binding site for miR-195 (ref. 28) and has been found to be nominally correlated with BCL2 expression (Supplementary Table 8; $P = 0.02$)29. Forced overexpression of BCL2 in mice leads to an increased incidence of B-cell lymphomas30.

The new SNPs at 18q21.32 and 15q15.1, as well as the new SNP at the established 2q13 locus, are located near BCL2 family member genes. rs4368253 is located approximately 51 kb downstream from PMAIP1 (phorbol-12-myristat-13-acetate-induced protein 1), which encodes the proapoptotic BCL2 protein NOXA. Regulation of apoptosis through NOXA is crucial for B-cell expansion after antigen triggering31. Downregulation of NOXA contributes to the persistence of CLL B cells in the lymph node environment32, rs8024033 is located approximately 5.4 kb upstream of BCL2 modifying factor, which encodes an apoptotic activator that binds

**Figure 1** Continued
to BCL2 proteins. BMF has been implicated in the survival of chronic lymphocytic leukemia cells, and loss of Bmf in mice leads to B-cell hyperplasia and an accelerated development of radiation-induced thymic lymphomas. The new SNP (rs13401811) at 2q13, a locus previously implicated in risk of CLL, and more generally in B-cell NHL, is located approximately 262 kb upstream of BCL2L11 (BCL2-like 11). BCL2L11 encodes a proapoptotic member of the BCL2 family, BIM, which has a key role in the regulation of apoptosis in T- and B-cell homeostasis. Loss of Bim accelerates Myc-induced leukemia in mice, and this SNP has been previously reported to be nominally associated with CLL in a small candidate gene study.

The new 2q33.1 SNP (rs3769825) resides in intron 2 of CASP8 (caspase-8) and is in LD with a missense SNP (rs13006529, $r^2 = 0.71$) in the nearby gene CASP10 (caspase-10) (Supplementary Table 9), both of which have a central role in cell apoptosis. SNPs within this region have been associated with breast cancer, esophageal cancer, and melanoma susceptibility. SNPs in CASP8/CASP10, including one in moderate LD with the SNP we found here (rs11674246, $r^2 = 0.66$), were previously nominally associated with CLL risk in smaller case-control studies.

The remaining four new loci (11p15.5, 4q25, 9p21.3 and 2p22.2) map to other biologically interesting genes. The 4q25 SNP, rs898518, is located between the fourth and fifth exons of LEF1 (lymphoid enhancer-binding factor 1), which encodes a transcription factor involved in the WNT signaling pathway, an essential component in the normal homeostasis of hematopoietic stem cells. Aberrant protein expression of LEF1 has been observed in CLL cells and monoclonal B-cell lymphocytosis, suggesting that LEF1 has an early role in CLL leukemogenesis. rs1679013 maps to an intergenic region on 9p21.3 roughly 200 kb upstream from CDKN2B-AS1, an antisense noncoding RNA implicated in the risk of acute lymphocytic leukemia. The 2p22.2 SNP (rs3770745) is located approximately 52 kb upstream of PRKD3 (protein kinase D3), which encodes a protein that interacts with transcriptional repressor, B-cell lymphoma 6 (BCL6). The 11p15.5 region contains many imprinted genes and has been implicated in Beckwith-Wiedemann syndrome, a disorder characterized by excessive growth and a high incidence of childhood tumors.

**DISCUSSION**

Our large GWAS of CLL identified ten associated SNPs in nine new loci and one new independently associated SNP in a previously reported locus. Together with the previously established loci, the cumulative set of SNPs correspond to an area under the curve of 0.73. Although further studies are required to fine map the regions, the proximity of several of these loci to genes involved in apoptosis suggests a possible underlying mechanism of biological relevance. Our results further support a substantial contribution of common gene variants in the pathogenesis of CLL.

**METHODS**

Methods and any associated references are available in the online version of the paper.

**ACKNOWLEDGMENTS**

We thank C. Allmer, E. Angelucci, A. Bigelow, I. Brock, K. Butterbach, A. Chabrier, D. Chan-Lam, J.M. Conners, D. Connelly, M. Cornelis, K. Corsano, C. Dalley, D. Cox, H. Cramp, R. Cutting, H. Dykes, L. Ershler, A. Gabbas, R.P. Gallagher, R.D. Gascoyne, P. Hui, L. Irish, L. Jacobus, S. Kaul, J. Lunde, M. McAdams, R. Montalvan, R. Rais, T. Rattle, L. Rigacci, K. Snyder, G. Specchia, M. Stagner, P. Taylor, G. Thomas, C. Tornow, G. Wood, M. Yang and M. Zucca for their assistance. The overall GWAS project was supported by the intramural program of the Division of Cancer Epidemiology and Genetics, NCI, US National Institutes of Health. A full list of acknowledgments is provided in the Supplementary Note.

**AUTHOR CONTRIBUTIONS**

S.I.B., C.E.S., N.J.C., A.N., W.C., S.S.W., L.R.T., A.R.B.-W., P.H., M.P.P., R.M.B., B.K.A., P.C., Y.Z., G.S., A.Z.-J., C.L., K.E.S., J.M., P.V., J.J.S., A.K., S.d.S., H.H.,...
Sonja I Berndt, Christine F Skibola, Vijai Joseph, Nicole J Camp, Alexandra Nieters, Zhaoming Wang, Wendy Cozen, Alain Monnereau, Sophia S Wang, Rachel S Kelly, Lauren R Teras, Nilanjan Chatterjee, Charles C Chung, Meredith Yeager, Angela R Brooks-Wilson, Patricia Hartge, Mark P Purdue, Brenda M Birmann, Bruce K Armstrong, Pierluigi Cocchi.
and Welfare, Helsinki, Finland. 70Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut, USA. 71Program in Cancer Biology and Genetics, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. 72Cancer Control Research, BC Cancer Agency, Vancouver, British Columbia, Canada. 73Department of Environmental Health, Harvard School of Public Health, Boston, Massachusetts, USA. 74Department of Nutrition, Harvard School of Public Health, Boston, Massachusetts, USA. 75Department of Pathology, Macquarie University, Sydney, New South Wales, Australia. 76Department of Histopathology, Douglass Hanly Moir Pathology, Macquarie Park, New South Wales, Australia. 77Princess of Wales Clinical School, University of New South Wales, Sydney, New South Wales, Australia. 78Department of Biomedical Science, University of Cagliari, Monreale, Cagliari, Italy. 79Interdisciplinary Department of Medicine, University of Bari, Bari, Italy. 80Environmental and Occupational Epidemiology Unit, ISPO, Florence, Italy. 81Epidemiology and Genetics Unit, Department of Health Sciences, University of York, York, UK. 82Dongguk University-Seoul, Seoul, South Korea. 83Department of Epidemiology, University of North Carolina, Chapel Hill, North Carolina, USA. 84Department of Oncology, University of Sheffield, Sheffield, UK. 85Department of Haematology, Sheffield Teaching Hospitals National Health Service (NHS) Foundation Trust, Sheffield, UK. 86Genomic Medicine Group Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), University of Santiago de Compostela, Santiago de Compostela, Spain. 87Department of Biochemistry and Molecular Biology, Institute of Oncology, University of Oviedo, Oviedo, Spain. 88August Pi I Sunyer Biomedical Research Institute (IDIBAPS), Hospital Clinic of Barcelona, University of Barcelona, Barcelona, Spain. 89These authors contributed equally to this work. 90These authors jointly directed this work. Correspondence should be addressed to S.L.S. (slager@mayo.edu).
ONLINE METHODS

Stage 1: NHL GWAS. As part of a larger initiative, we conducted a GWAS of CLL using cases and controls of European descent from 22 studies of NHL (Supplementary Table 1), including 9 prospective cohort studies, 8 population-based case-control studies and 5 clinic- or hospital-based case-control studies. All studies obtained informed consent from the participants and approval from the respective Institutional Review Boards for this study. As described in Supplementary Table 1, cases were ascertained from cancer registries, clinics or hospitals or through self report verified by medical and pathology reports. The phenotype information for all cases with NHL was reviewed centrally at the International Lymphoma Epidemiology Consortium (InterLymph) Data Coordinating Center and harmonized according to the hierarchical classification proposed by the InterLymph Pathology Working Group based on the World Health Organization (WHO) classification (2008)50,51.

All cases with CLL with sufficient DNA (n = 2,343) and a subset of available controls frequency matched by age and sex to cases (n = 2,854), including 4% quality-control duplicates, were genotyped on the Illumina OmniExpress at the NCI Cancer Genomic Research Laboratory (CGR). Genotypes were called using Illumina GenomStudio software, and quality-control duplicates showed >99% concordance. Extensive quality-control metrics were applied to the data. Monomorphic SNPs and SNPs with call rate <93% were excluded. Samples with call rate ≤93%, mean heterozygosity ≤0.25 or >0.33 based on the autosomal SNPs or gender discordance (≥5% heterozygosity on the X chromosome for males and <20% heterozygosity on the X chromosome for females) were excluded. Unexpected duplicates (>99.9% concordance) and first-degree relatives on the basis of identity-by-descent sharing with Pi-hat > 0.40 were removed. Ancestry was assessed using the Genotyping Library and Utilities (GLU) struct.admix module based on the method proposed by Pritchard et al.52 and participants with <80% European ancestry were excluded (Supplementary Fig. 3). After exclusions, 2,178 (93%) cases and 2,685 (94%) controls remained (Supplementary Table 2). Genotype data previously generated on the Illumina Omni12.5 from an additional 3,535 controls and 1 case from three of the studies (Alpha-Tocopherol, Beta-Carotene Lung Cancer Prevention Study (ATBC), American Cancer Society Cancer Prevention Study–II Cohort (CPSII) and Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (PLCCO)) were also included53, resulting in a total of 2,179 cases and 6,221 controls for the stage 1 analysis. Of these additional controls, 703 (~235 from each study) were selected to be representative of their cohort and were cancer free54. The remaining 2,823 controls were cancer-free controls from an unpublished study of prostate cancer in PLCCO. SNPs with call rate <99%, Hardy-Weinberg equilibrium P < 1 × 10^-8 or minor allele frequency <1% were excluded from analysis, leaving 549,934 SNPs for analysis. To evaluate population structure, a principal components analysis was performed using the GLU, version 1.0, struct.pca module, which is similar to EIGENSTRAT55. Plots of the first ten principal components are shown in Supplementary Figure 4. Association testing was conducted assuming a log-additive genetic model adjusting for age, sex and significant principal components. All data analyses and management were conducted using GLU.

Stage 2: three Independent CLL GWAS. Three independent CLL GWAS provided genotype data for a meta-analysis (Supplementary Table 1). In all three studies, subjects with genotyping call rate <95%, duplicates, related individuals and SNPs with call rate <95% were removed before imputation (Supplementary Table 4). Imputation was conducted separately for each study using IMPUTE2 (ref. 11) and a hybrid of the 1000 Genomes Project version 2 (February 2012 release) and Division of Cancer Epidemiology and Genetics (DCEG) European reference panels12,13. SNPs were imputed for a total of 921 cases and 1,446 controls. Association testing was conducted for each study using SNPTEST version 2 adjusting for age, sex and significant principal components obtained informed consent from the participants and approval from the NCI Core Genotyping Resource, and genotyping for the Utah/Sheffield study was conducted at the Core Research Facilities at the University of Utah. Blind duplicates (~5%) yielded 100% concordance. The ICGC study provided results for eight SNPs (or proxies) that were genotyped on the Affymetrix 6.0 SNP microarray (Supplementary Table 4). Association results for the NCI Rep and Utah/Sheffield studies were adjusted for age and sex, and results from ICGC were adjusted for age, sex and significant principal components. A comparison of the genotyping calls from the OmniExpress microarray and confirmatory TaqMan assays (n = 384) yielded 99.9% concordance.

Meta-analysis. Meta-analyses were performed using the fixed-effects inverse-variance method based on the β estimates and standard errors from each study. For all SNPs listed in Tables 1 and 2, no substantial heterogeneity was observed among the studies in stage 1 or among the studies in stages 1–3 combined after Bonferroni correction (P heterogeneous ≥ 0.02 for all SNPs).

Further follow-up analyses. Using 1000 Genomes data, we identified SNPs with P2 > 0.7 with our lead SNP that were reported to be non synonymous or nonsynonymous variants. We used HaploReg56, which is a tool for exploring non-coding functional annotation using ENCODE data, to evaluate the genome surrounding our SNPs (Supplementary Table 9). In addition, we evaluated cis associations between all new and promising SNPs discovered in this study and the expression of nearby genes in lymphoblastoid cell lines from subjects of European descent from three publicly available data sets29,55,56 (Supplementary Table 8).

Heritability analyses. To evaluate the familial risk explained by the new loci identified in this study, we estimated the contribution of each SNP to heritability using the equation $\rho_{SNP}^2 = \beta^2 \times 2(1 - f)$, where β is the log-odds ratio per copy of the risk allele and f is the allele frequency, and then summed the contributions of all newly associated SNPs. Using the equation derived by Pharoah et al.57 to estimate the total heritability from the sibling relative risk (RR = 8.5 from Goldin et al.58), we then calculated the proportion of familial risk explained by dividing the summed contributions of the newly associated SNPs by the total heritability.

To estimate the contribution of all common SNPs to familial risk, we used the method proposed by Yang et al.59, which was extended to dichotomous traits22 and implemented in the Genome-wide Complex Trait Analysis (GCTA) software58. The genetic similarity matrix was estimated from our discovery scan using all genotyped autosomal SNPs with a minor allele frequency >0.01. We used restricted maximum likelihood (REML), the default option for GCTA, to fit the appropriate variance components model that included the top ten eigenvectors as covariates. The final estimate of heritability on the underlying liability scale assumed that the lifetime risk of CLL was 0.005. From this estimate, we calculated the proportion of familial risk explained based on a familial relative risk of 8.5. Details of fitting the variance components model and transforming from the observed to the liability scale have been previously documented22.

Estimate of recombination hotspots. To identify recombination hotspots in the region, we used SequenceDhоО60, a program that uses the approximate marginal likelihood method61 and calculates likelihood ratio statistics at a set of possible hotspots. We tested five unique sets of 100 control samples. The PHASE v2.1 program was used to calculate background recombination rates62,63 and LD heat maps were visualized using the snp.plotter program64.

50. Morton, L.M. et al. Proposed classification of lymphoid neoplasms for epidemiologic research from the Pathology Working Group of the International Lymphoma Epidemiology Consortium (InterLymph). Blood 110, 695–708 (2007).
51. Turner, J.J. et al. InterLymph hierarchical classification of lymphoid neoplasms for epidemiologic research based on the WHO classification (2008): update and future directions. Blood 116, e90–e98 (2010).
52. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. Genetics 155, 945–959 (2000).
53. Price, A.L. et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38, 904–909 (2006).
54. 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–D934 (2012).
55. Dixon, A.L. et al. A genome-wide association study of global gene expression. Nat. Genet. 39, 1202–1207 (2007).
56. Stranger, B.E. et al. Population genomics of human gene expression. Nat. Genet. 39, 1217–1224 (2007).
57. Pharoah, P.D. et al. Polygenic susceptibility to breast cancer and implications for prevention. Nat. Genet. 31, 33–36 (2002).
58. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 88, 76–82 (2011).
59. Fearnhead, P. SequenceLDhot: detecting recombination hotspots. Bioinformatics 22, 3061–3066 (2006).
60. Fearnhead, P., Harding, R.M., Schneider, J.A., Myers, S. & Donnelly, P. Application of coalescent methods to reveal fine-scale rate variation and recombination hotspots. Genetics 167, 2067–2081 (2004).
61. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165, 2213–2233 (2003).
62. Crawford, D.C. et al. Evidence for substantial fine-scale variation in recombination rates across the human genome. Nat. Genet. 36, 700–706 (2004).
63. Luna, A. & Nicodemus, K.K. SNP.plotter: an R-based SNP/haplotype association and linkage disequilibrium plotting package. Bioinformatics 23, 774–776 (2007).