A genome-wide association study of marginal zone lymphoma shows association to the HLA region

Joseph Vijai1,*, Zhaomiao Wang1, Sonja I. Berndt1,*, Christine F. Skibola1,5, Susan L. Slater2, Silvia de Sanjose3,8, Mads Melbye9,10, Bengt Glimelius11,12, Paige M. Bracci13, Lucia Conde4,5, Brenda M. Birmann14, Sophia S. Wang15, Angela R. Brooks-Wilson16,17, Qing Lan3, Paul I.W. de Bakker18,19, Roel C.H. Vermeulen19,20, Carol Portlock1, Stephen M. Ansell21, Brian K. Link22, Jacques Ribo5, Kari E. North23,24, Jian Gu25, Henrik Hjalgrim9, Wendy Cozen26,27, Nikolaus Becker28, Lauren R. Teras29, John S. Spinei30,31, Jennyere23,32,33,35,36,37,38,39, Maria Grazia Ennas40, Alain Monnereau12,42,43, Kimberly A. Bertrand14,44, Demetrius Albanese3, Tracy Lightfoot45, Meredith Yeager46, Charles C. Chung46, Laurie Burdett47, Amy Hutchinson48, Charles Lawrence49, Rebecca Montalvan50, Liming Liang44,47, Jinyan Huang44, Baoshan Ma44,48, Danylo J. Villan49, Anna Maria Corniness,50 Tinu Thomas1, Anne J. Novak51, Ahmet Dogan52, Mark Liebow25,48,49, Carrie A. Thompson53, Thomas E. Witzig54, Thomas M. Hamburgen55, George J. Weiner56, Martyn T. Smith57, Elizabeth A. Holly58, Rebecca D. Jackson59, Lesley F. Tinker59, Yuanqing Ye25, Hans-Olov Adam44,52, Karin E. Smedby53, Annecia De Roog51,54, Patricia Hartge3, Lindsay M. Morton3, Richard K. Severson55, Yolanda Benavente7,8, Paolo Boffetta56, Paul Brennan57, Lenka Foretova58, Marc Maynadie59, James McKay60, Anthony Staines61, W. Ryan Diver29, Claire M. Vajdic62, Bruce K. Armstrong63, Anne Kricker63, Tongzhang Zheng34, Theodore R. Holford64, Gianluca Severi35,36,65, Paolo Vineis37,65, Meredith Yeager2, Charles C. Chung46, Laurie Burdett47, Amy Hutchinson48, Charles Lawrence49, Rebecca Montalvan50, Liming Liang44,47, Jinyan Huang44, Baoshan Ma44,48, Danylo J. Villan49, Anna Maria Corniness,50 Tinu Thomas1, Anne J. Novak51, Ahmet Dogan52, Mark Liebow25,48,49, Carrie A. Thompson53, Thomas E. Witzig54, Thomas M. Hamburgen55, George J. Weiner56, Martyn T. Smith57, Elizabeth A. Holly58, Rebecca D. Jackson59, Lesley F. Tinker59, Yuanqing Ye25, Hans-Olov Adam44,52, Karin E. Smedby53, Annecia De Roog51,54, Patricia Hartge3, Lindsay M. Morton3, Richard K. Severson55, Yolanda Benavente7,8, Paolo Boffetta56, Paul Brennan57, Lenka Foretova58, Marc Maynadie59, James McKay60, Anthony Staines61, W. Ryan Diver29, Claire M. Vajdic62, Bruce K. Armstrong63, Anne Kricker63, Tongzhang Zheng34, Theodore R. Holford64, Gianluca Severi35,36,65, Paolo Vineis37,65, Giovann M. Ferr66, Rosalio Riccio67, Lucia Milig68, Jacqueline Clavel61,42, Edward Giovannucci14,44,46,9, Peter Kraft44,47, Jarno Virtamo70, Alex Smith45, Eleanor Kane45, Eve Roman46, Brian C.H. Chiu1, Joseph F. Fraumeni3, Xifeng Wu25,71, James R. Cerhan6,71, Kenneth Offit1,1, Stephen J. Chanock3,4, Nathaniel Rothman1,7 & Alexandra Nieters6,7,8

Marginal zone lymphoma (MZL) is the third most common subtype of B-cell non-Hodgkin lymphoma. Here we perform a two-stage GWAS of 1,281 MZL cases and 7,127 controls of European ancestry and identify two independent loci near BTLN2 (rs9461741, P = 3.95 × 10−15) and HLA-A (rs2922994, P = 2.43 × 10−9) in the HLA region significantly associated with MZL risk. This is the first evidence that genetic variation in the major histocompatibility complex influences MZL susceptibility.

Received 16 May 2014 | Accepted 4 Nov 2014 | Published 8 Jan 2015

* These authors contributed equally to this work. Correspondence and requests for materials should be addressed to J.V. (email: josephv@mskcc.org).
Marginal zone lymphoma (MZL) encompasses a group of lymphomas that originate from marginal zone B cells present in extranodal tissue and lymph nodes. Three subtypes of MZL have been defined, extranodal MZL of mucosa-associated lymphoid tissue (MALT), splenic MZL and nodal MZL, which together account for 7–12% of all non-Hodgkin lymphoma (NHL) cases. Geographic differences in incidence have been observed, and inflammation, immune system dysregulation and infectious agents, such as Helicobacter pylori, have been implicated particularly for the gastric MALT subtype\(^2\), but little else is known of MZL aetiology.

Here we perform the first two-stage, subtype-specific genome-wide association study (GWAS) of MZL and identify two independent single-nucleotide polymorphisms (SNPs) within the HLA region associated with MZL risk. Together with recent studies on other common subtypes of NHL, these results point to shared susceptibility loci for lymphoma in the HLA region.

Results

Stage 1 MZL GWAS. As part of a larger NHL GWAS, 890 MZL cases and 2,854 controls from 22 studies in the United States and Europe (Supplementary Table 1) were genotyped using the Illumina OmniExpress array. Genotype data from the Illumina Omni2.5 was also available for 3,536 controls from three of the 22 studies\(^3\). After applying rigorous quality control filters (Supplementary Table 2, Methods), data for 611,856 SNPs with minor allele frequency (MAF) >1% in 825 cases and 6,221 controls of European ancestry (Supplementary Fig. 1) remained for the stage 1 analysis (Supplementary Table 3). To discover variants associated with risk, logistic regression analysis was performed on these SNPs adjusting for age, gender and three significant eigenvectors computed using principal components analysis (Supplementary Fig. 2, Methods). Examination of the quantile–quantile (Q–Q) plot (Supplementary Fig. 3) showed minimal detectable evidence for population substructure (\(\lambda = 1.01\)) with some excess of small \(P\) values. A Manhattan plot revealed association signals at the HLA region (Supplementary Fig. 4; rs62133:31,061,211–32,620,572) on chromosome 6 reaching genome-wide significance. Removal of all SNPs in the HLA region resulted in an attenuation of the excess of small \(P\) values observed in the Q–Q plot, although some excess still remained. To further explore associations within the HLA region and identify other regions potentially associated with risk, common SNPs available in the 1000 Genomes project data release 3 were imputed (Methods).

Stage 2 genotyping. Ten SNPs in promising loci with \(P < 7.5 \times 10^{-6}\) in the stage 1 discovery were selected for replication (stage 2) in an additional 456 cases and 906 controls of European ancestry (Supplementary Tables 1 and 3). Of the SNPs selected for replication, two SNPs were directly genotyped on the OmniExpress, while the remaining eight were imputed with high accuracy (median info score = 0.99) in stage 1 (Supplementary Table 4). Replication was carried out using Taqman genotyping. In the combined meta-analysis of 1,281 cases and 7,127 controls, we identified two distinct loci (Table 1, Fig. 1, Supplementary Table 4) at chromosomes 6p21.32 and 6p21.33 that reached the threshold of genome-wide statistical significance (\(P < 5 \times 10^{-8}\)). These are rs9461741 in the butyrophilin-like 2 (MHC class II associated) (BTNL2) gene at 6p21.32 in HLA class II (\(P = 3.95 \times 10^{-15}\), odds ratio (OR) = 2.66, confidence interval (CI) = 2.08–3.39) and rs2922994 at 6p21.33 in HLA class I (\(P = 2.43 \times 10^{-9}\), OR = 1.64, CI = 1.39–1.92). These two SNPs were weakly correlated (\(r^2 = 0.008\) in 1000 Genomes CEU population), and when both were included in the same statistical model, both SNPs remained strongly associated with MZL risk (rs9461741, \(P = 2.09 \times 10^{-13}\), rs2922994, \(P = 6.03 \times 10^{-13}\)), suggesting that the two SNPs are independent. Both SNPs were weakly correlated with other SNPs in the HLA region previously reported to be associated with other NHL subtypes or Hodgkin lymphoma (\(r^2 < 0.14\) for all pairwise comparisons). None of the previously reported SNPs were significantly associated with MZL risk after adjustment for multiple testing (\(P < 0.0025\) in our study, suggesting the associations are subtype-specific (Supplementary Table 5). Another SNP rs7750641 (\(P = 3.34 \times 10^{-8}\), Supplementary Table 4) in strong linkage disequilibrium (LD) with rs2922994 (\(r^2 = 0.85\)) also showed promising association with MZL risk. rs7750641 is a missense variant in transcription factor 19 (TCF19), which encodes a DNA-binding protein implicated in the transcription of genes during the G1–S transition in the cell cycle\(^4\). The non-HLA SNPs genotyped in stage 2 were not associated with MZL risk (Supplementary Table 4).

HLA alleles. To obtain additional insight into plausible functional variants, we imputed the classical HLA alleles and amino acid residues using SNP2HLA\(^5\) (Methods). No imputed HLA alleles or amino acid positions reached genome-wide significance (Supplementary Table 6). However, for HLA class I, the most promising associations were observed with HLA-B*08 (\(7.94 \times 10^{-8}\), HLA-B*08:01 (\(7.79 \times 10^{-8}\) and the HLA-B allele encoding an aspartic acid residue at position 9 (Asp9) (\(7.94 \times 10^{-8}\), located in the peptide binding groove of the protein. HLA-B*08:01 and Asp9 are highly correlated (\(r^2 = 0.99\)), and thus their effect sizes were identical (OR = 1.67, 95% CI: 1.38–2.01). They are both also in strong LD with rs2922994 (\(r^2 = 0.97\)). Due to the fact that they are collinear, the effects of the SNPs and alleles were indistinguishable from one another in conditional modelling. For HLA class II, a suggestive association was observed with HLA-DRB1*01:02 (OR = 2.24, 95% CI: 1.64–3.07, \(P = 5.08 \times 10^{-7}\); Supplementary Table 6), which is moderately correlated with rs9461741 (\(r^2 = 0.69\)). Conditional analysis revealed that the effects of rs9461741 (the intragenic SNP in BTNL2) and HLA-DRB1*01:02 were indistinguishable statistically (stage 1: \(r^2 = 0.69\), \(P_{\text{adjusted}} = 0.064\) and HLA-DRB1*01:02, \(P_{\text{adjusted}} = 0.29\). A model containing both HLA-B*08:01 and HLA-DRB1*01:02 showed that the two alleles were independent (HLA-B*08:01: \(P_{\text{adjusted}} = 4.65 \times 10^{-8}\) and HLA-DRB1*01:02: \(P_{\text{adjusted}} = 2.97 \times 10^{-7}\)), further supporting independent associations in HLA class I and II loci.

MALT versus non-MALT. Heterogeneity between the largest subtype of MZL, namely MALT and other subtypes grouped as non-MALT, was evaluated for the MZL associated SNPs (Supplementary Table 7). The effects were slightly stronger for MALT, but no evidence for substantial heterogeneity was observed (\(P_{\text{heterogeneity}} = 0.05\)). Studies have suggested that H. pylori infection is a risk factor for gastric MZL\(^2\). An examination of SNPs previously suggested to be associated with H. pylori infection in independent studies\(^6\) did not reveal any significant association with the combined MZL or the MALT subtype in this study (Supplementary Table 8). Toll-like receptors (TLR) are considered strong candidates in mediating inflammatory immune response to pathogenic insults. A previous study reported\(^7\) a nominally significant association with TLR10–TLR1–TLR6 region with MZL risk. After excluding the cases and controls in the previous report\(^7\), we found little additional support for this locus (MZL: \(P = 0.006\), OR = 1.18 and MALT: \(P = 0.38\), OR = 1.08).

© 2015 Macmillan Publishers Limited. All rights reserved.
Secondary functional analyses. To gain additional insight into potential biological mechanisms, expression quantitative trait loci (eQTL) analyses were performed in two datasets consisting of lymphoblastoid cell lines (Methods). Significant associations were seen for rs2922994 and rs7750641 with HLA-B and HLA-C (Supplementary Table 9) while suggestive associations (false discovery rate, FDR < 0.05) for correlated SNPs of rs2922994 \( (r^2 = 0.8) \) in HLA class I and RNF5 (Supplementary Table 10) were observed. No significant eQTL association was observed for rs9461741 or other correlated HLA class II SNPs. Chromatin state analysis (Methods) using ENCODE data revealed correlated SNPs of rs2922994 showed a chromatin state consistent with the prediction for an active promoter (rs3094005) or satisfied the state of a weak promoter (rs2844577) in the lymphoblastoid cell line.
GM12878 (Supplementary Fig. 5). GM12878 is the only lymphoblastoid cell line from which high-quality whole-genome annotation data for chromatin state is readily available. Analyses were also performed with HaploReg (Supplementary Table 11) and RegulomeDB (Supplementary Table 12) that showed overlap of the SNPs with functional motifs, suggesting plausible roles in gene regulatory processes.

Discussion
The most statistically significant SNP associated with MZL, rs9461741, is located in HLA class II in the intron between exons 3 and 4 of the BTNL2 gene. BTNL2 is highly expressed in lymphoid tissues8 and has close homology to the B7 costimulatory molecules, which initiate lymphocyte activation as part of antigen presentation. Evidence is consistent with BTNL2 acting as a negative regulator of T-cell proliferation and cytokine production9,10 and attenuating T-cell-mediated responses in the gut10. We were unable to statistically differentiate the effects of rs9461741 from HLA-DRB1*01:02 and, thus, our observed association could be due to HLA-DRB1. HLA-DRB1 has been shown to be associated with other autoimmune diseases, including rheumatoid arthritis11 and selective IgA deficiency12. Similarly, rs2922994 is located 11 kb upstream of HLA-B, which is known to play a critical role in the immune system by presenting peptides derived from the endoplasmic reticulum lumen. rs7750641, a missense variant in TCF19, was previously associated with pleiotropic effects on blood-based phenotypes13 and is highly expressed in germinal center cells and up-regulated in human pro-B and pre-B cells14. Autoimmune diseases, such as Sjögren’s syndrome and systemic lupus erythematosus, are established risk factors for developing MZL, with the strongest associations seen between Sjögren’s syndrome and the MALT subtype15. Of note, SNPs in HLA-B and the classical alleles HLA-DRB1*01:02 are strongly associated with Sjögren’s syndrome16,17, while HLA-DRB1*03 has been associated with rheumatoid arthritis17. The multiple independent associations in the HLA region and their localization to known functional autoimmune and B-cell genes suggest possible shared genetic effects that span both lymphoid cancers and autoimmune diseases. Chronic autoimmune stimulation leading to overactivity and defective apoptosis of B cells, and secondary inflammation events triggered by genetic and environmental factors are biological mechanisms that may contribute to the pathogenesis of MZL.

We have performed the largest GWAS of MZL to date and identified two independent SNPs within the HLA region that are robustly associated with the risk of MZL. In addition to the known diversity in etiology and pathology, there is mounting evidence of genetic heterogeneity across the NHL subtypes of lymphoma. However, the HLA region appears to be commonly associated with multiple major subtypes, such as MZL, CLL,16,19, DLBCL,19 and FL.20,23. Further studies are needed to identify biological mechanisms underlying these relationships and advance our knowledge regarding their interactions with associated environmental factors that may modulate disease risks.

Methods
Stage 1 MZL GWAS study subjects and ethics. As part of a larger NHL GWAS initiative, we conducted a GWAS of MZL using 890 cases and 2,854 controls of European descent from 22 studies of NHL (Supplementary Table 1 and Supplementary Table 2), including nine prospective cohort studies, eight population-based case–control studies, and five clinic or hospital-based case–control studies. All studies were approved by the respective Institutional Review Boards as listed. These are ATBC(NCI Special Studies Institutional Review Board), BCCA: UBC BC Cancer Agency Research Ethics Board, CPS-II: American Cancer Society, LCSS: Northern and Yorkshire Research Ethics Committee, ENGELA: IRB000003888—Comité d’Évaluation Ethique de l’Inserm IRB # 1, EPIC: Imperial College London, Epilymph: International Agency for Research on Cancer, HPFS: Harvard School of Public Health (HSPH) Institutional Review Board, Iowa-Mayo SPORE: University of Iowa Institutional Review Board, Italian GALE: Comité Etico Azienda Ospedaliero Universitaria di Cagliari, Mayo Clinic Case–Control: Mayo Clinic Institutional Review Board, MCCS: Cancer Council Victoria’s Human Research Ethics Committee, MD Anderson: University of Texas MD Anderson Cancer Center Institutional Review Board, MSKCC: Memorial Sloan Kettering Cancer Center Institutional Review Board, NCI-SEER (NCI Special Study Institutional Review Board), NIHs: Partners Human Research Committee, Brigham and Women’s Hospital, NS: NSW Cancer Council Ethics Committee, NYU-WHS: New York University School of Medicine Institutional Review Board, PLCO: (NCI Special Studies Institutional Review Board), SCALE: Scientific Ethics Committee for the Capital Region of Denmark, SCALE: Regional Ethical Review Board in Stockholm (Section 4) IRB#: UCSF: University of California San Francisco Committee on Human Research, WHI: Fred Hutchinson Cancer Research Center, Yale: Human Investigation Committee, Yale University School of Medicine. Informed consent was obtained from all participants.

SNPs were ascertained from cancer registries or hospitals or through self-report verified by medical and pathology reports. To determine the NHL subtype, phenotype data for all NHL cases were reviewed centrally at the International Lymphoma Epidemiology Consortium (InterLymph) Data Coordinating Center and harmonized using the hierarchical classification proposed by the InterLymph Pathology Working Group24,25 based on the World Health Organization (WHO) classification26.

Genotyping and quality control. All MZL cases with sufficient DNA (n = 890) and a subset of controls (n = 2,854) frequency matched by age, sex and study to the entire group of NHL cases, along with 4% quality control duplicates, were genotyped on the Illumina OmniExpress at the NCI Core Genotyping Resource (CGR). Genotypes were called using Illumina GenomeStudio software, and quality control (QC) was conducted on the genotype data previously generated on the Illumina Omni2.5 from an additional 3,536 controls from three of the 22 studies (ATBC, CPS-II and PLCO) and were also included23, resulting in a total of 825 cases and 6,221 controls for the stage 1 analysis (Supplementary Table 3). Of these additional 3,536 controls, 703 (~235 from each study) were selected to be representative of their cohort and cancer free23, while the remainder were cancer-free controls from an unpublished study of prostate cancer in the PLCO. SNPs with call rate <95%, with Hardy–Weinberg equilibrium P < 1 × 10−10, or with a MAF <1% were excluded from analysis, leaving 611,856 SNPs for analysis. To evaluate population substructure, a principal components analysis was conducted using EIGENSTRAT28 -http://genepath.med.harvard.edu/eigens, mean heterozygosity <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. Furthermore, unexpected duplications (>99.9% concordance) and first-degree relatives based on identity by descent sharing with Pi-hat >0.40 were excluded. Ancestry was assessed using the Genotyping Library and Utilities (GLU-http://code.google.com/p/glui-genetics/), struct.admix module based on the method by Pritchard et al.27 and participants with <80% European ancestry were excluded (Supplementary Fig. 1). After exclusions, 825 MZL cases and 6,221 controls remained. The SNPs were tested for association with NHL in a single stage analysis using PLIF (http://www.international-lymphoma.org/plif) version 3 stage 1 after imputation with the 1000 Genomes Project (1KG-http://www.international-lymphoma.org/) version 3 data28. The MAF <1% and controls with a quality score <info> <0.3 were excluded from analysis, leaving 8,478,065 SNPs for association testing. Association testing was conducted using the imputed data was conducted using SNPTEST32,33-https://mathgen.stats.ox.ac.uk/genetics_software/snpset/snpset.html version 2, assuming dosages for the genotypes and adjusting for age, sex and three significant principle components. In a null model for MZL risk, the first three eigenvectors EV1, EV2, EV3 were taken forward for de novo mutation analysis.

Stage 2 replication of SNPs from the GWAS. After ranking the SNPs by P value and LD filtering (r² >0.05), 10 SNPs from the most promising loci identified from stage 1 were selected for replication in an additional 456 cases and 906 controls (Supplementary Tables 1 and 4). Wherever possible, we selected either the best directly genotyped SNP or
the most significant imputed SNP for the locus. Only imputed SNPs with an information score >0.8 were considered for replication. Only SNPs with MAF >1% in at least one of the controls were taken forward for replication in regions where they appeared as singletons or obvious artifacts. For the HLA region, we selected one additional SNP (rs7750641) that was highly correlated with rs2992994 for additional confirmation. Genotyping was conducted using custom TaqMan genotyping assays (Applied Biosystems) validated at the NCI Core Genotyping Resource. Genotyping was done at four centres. HapMap control samples genotyped across two centres yielded 100% concordance as did blind duplicates (~5% of total samples). Due to the small number of samples, the MD Anderson, Mayo and NCI replication studies were pooled together for association testing; however, MSKCC samples were analysed separately to account for the available information on Ashkenazi ancestry. Association results were adjusted for age and gender and study site in the pooled analysis. The results from the stage 1 and stage 2 studies were then combined using a fixed effect meta-analysis method (Supplementary Table 9).

HapMap imputation and analysis. To determine if specific coding variants within HLA genes contributed to the diverse association signals, we imputed the classical HLA alleles (A, B, C, DQA1, DQB1, DRB1) and coding variants across the HLA region (chr6:20–40 Mb) using SNP2HLA14 http://www.broadinstitute.org/mpg/snp2hla/). The HapMap control data set was based on a reference panel from the Type 1 Diabetes Genetics Consortium (T1DGC) consisting of genotype data from 5,225 individuals. The concordance rates obtained were 97.32, 98.5, 98.14 and 97.49% for HLA-A, B, C and DRB1, respectively, in the NCI GWAS suggesting robust performance of the imputation method. Due to the limited number of SNPs (7,253) in the T1DGC reference set, imputation of HLA SNPs was conducted with IMPUTE2 and the 1KGP reference set as described above. A total of 68,488 SNPs, 201 classical HLA alleles (two- and four-digit resolution) and 1,038 AA markers including 103 AA positions that were ‘multi-allelic’ with three to six different residues present at each position, were successfully imputed (info score >0.3 for SNPs or r2>0.3 for alleles and AAs) and available for downstream analysis. Multi-allelic markers were analysed as binary markers (for example, allele present or absent) and a meta-analysis was conducted where we tested SNPs, HLA alleles and AAs across the HLA region for association with MZL using PLINK33 or SNPTEST35 as described above.

eQTL analysis. We conducted an eQTL analysis using two independent datasets: childhood asthma14 and HapMap15. As described previously14 for the childhood asthma data set15, peripheral blood lymphocytes were transformed into lymphoblastoid cell lines for 830 parents and offspring from 206 families of European ancestry. Data from 405 controls were used for the analysis as follows: using extracted RNA, gene expression was assessed with the Affymetrix HG-U133 Plus 2.0 chip. Genotyping was conducted using the Illumina Human-1 Beadchip and Illumina HumanHap300K Beadchip, and imputation performed using data from 1KGP. All SNPs selected for replication were tested for eQTL associations (defined as gene transcripts within 1 Mb), assuming an additive genetic model, adjusting for non-genetic effects in the gene expression value. Association testing was conducted using a variance component-based score test67 in MERLIN77, which accounts for the correlation between siblings. To gain insight into the relative importance of associations with our SNPs compared with other SNPs in the region, we also conducted additional analyses, in which both the MZL SNP and the most significant SNP for the particular gene transcript (that is, peak SNP) were included in the analysis. In conditional analyses, in which both the MZL SNP and the most strongly associated SNP were included in the analysis, the following results were observed.

Bioinformatics ENCODE and chromatin state dynamics. To assess chromatin state dynamics, we used Chromos, which has precomputed data from ENCODE on nine cell types using Chip-Seq experiments76. These consist of B-lymphoblastoid cells (GM12878), hepatocellular carcinoma cells (HepG2), embryonic stem cells, erythroleukemia cell lines (K562), umbilical vein endothelial cells, skeletal muscle myoblasts, normal lung fibroblasts, normal epidermal keratinocytes and mammary epithelial cells. These precomputed data have genome-segmentation performed using a multivariate hidden Markov model to reduce the combinatorial space to a set of interpretable chromatin states. The output from Chromos lists data into 15 chromatin states corresponding to repressed, poised and active promoters, strong and weak enhancers, putative insulators, transcribed regions and large-scale repressed and inactive domains (Supplementary Fig. 5).

References
1. Doglion, C., Wortheroson, A. C., Moschini, A., de Boni, M. & Isaacs, P. G. High incidence of primary gastric lymphoma in northeastern Italy. Lancet 339, 834–835 (1992).
2. Parsonnet, J. et al. Helicobacter pylori infection and gastric lymphoma. New Engl. J. Med. 330, 1267–1271 (1994).
3. Wang, Z. et al. Improved imputation of common and uncommon SNPs with a new reference set. Nat. Genet. 44, 6–7 (2012).
4. Ku, D. H. et al. A new growth-regulated complementary DNA with the sequence of a putative trans-activating factor. Cell Growth Differ. 2, 179–186 (1991).
5. Jia, X. et al. Imputing amino acid polymorphisms in human leukocyte antigens. PLoS ONE 8, e64683 (2013).
6. Mayerle, J. et al. Identification of genetic loci associated with Helicobacter pylori serologic status. JAMA 309, 1912–1920 (2013).
7. Purdue, M. P. et al. A pooled investigation of Toll-like receptor gene variants and risk of non-Hodgkin lymphoma. Carcinogenesis 30, 275–281 (2009).
8. Arnett, H. A. et al. BTN2L1, a butyrophilin-B7-like molecule, is a negative costimulatory molecule modulated in intestinal inflammation. J. Immunol. 178, 1523–1533 (2007).
9. Nguyen, T., Liu, X. K., Zhang, Y. & Dong, C. BTN2L1, a butyrophilin-like molecule that functions to inhibit T cell activation. J. Immunol. 176, 7354–7360 (2006).
10. Swansson, R. M. et al. Butyrophilin-like 2 modulates B7 costimulation to induce Foxp3 expression and regulatory T cell development in mature T cells. J. Immunol. 190, 2027–2035 (2013).
11. Kallberg, H. et al. Gene-gene and gene-environment interactions involving HLA-DRB1, HTF22 and smoking in two subsets of rheumatoid arthritis. Arthritis Rheum. 60, 867–875 (2009).
12. Ferreira, R. C. et al. High-density SNP mapping of the HLA region identifies multiple independent susceptibility loci associated with selective IgA deficiency. PLoS Genet. 8, e1002476 (2012).
13. Ferreira, M. A. et al. Sequence variants in three loci influence monocyte counts and erythrocyte volume. Am. J. Hum. Genet. 85, 745–749 (2009).
14. Hystad, M. E. et al. Characterization of early stages of human B cell development by gene expression profiling. J. Immunol. 179, 3662–3671 (2007).
15. Dias, C. & Isenberg, D. A. Susceptibility of patients with rheumatic diseases to B-cell non-Hodgkin lymphoma. Nat. Rev. Rheumatol. 7, 360–368 (2011).
16. Lessard, C. et al. Variants at multiple loci implicated in both innate and adaptive immune responses are associated with Sjögren’s syndrome. Nat. Genet. 45, 1284–1292 (2013).
17. Raychaudhuri, S. et al. Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. Nat. Genet. 44, 291–296 (2012).
18. Bernd, S. I. et al. Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia. Nat. Genet. 45, 868–876 (2013).
19. Cerhan, J. R. et al. Genome-wide association study identifies multiple susceptibility loci for diffuse large B cell lymphoma. Nat. Genet. 46, 1233–1238 (2014).
20. Smalley, K. E. et al. GWAS of follicular lymphoma reveals allelic heterogeneity within the 3p21.12 locus and suggests shared genetic susceptibility with diffuse large B-cell lymphoma. PLoS Genet. 7, e1001378 (2011).
21. Conde, L. et al. Genome-wide association study of follicular lymphoma identifies a risk locus at 6p21.32. Nat. Genet. 42, 661–664 (2010).
22. Vijai, J. et al. Susceptibility loci associated with specific and shared subtypes of lymphoid malignancies. PLoS Genet. 9, e1003200 (2013).
23. Sklar, P. et al. Genome-wide association study identifies five susceptibility loci for follicular lymphoma outside the HLA region. Am. J. Hum. Genet. 95, 462–471 (2014).
24. 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).
25. Turner, J. I. 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).
26. Swerdlow, S., Campo, E. & Harris, N. World Health Organization Classification of Tumours of Haematopoietic and Lymphoid Tissues (IARC Press, 2008).
27. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. Genetics 155, 945–959 (2000).
28. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38, 904–909 (2006).
29. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 5, e1000529 (2009).
30. Abecasis, G. R. et al. A map of human genome variation from population-scale sequencing. Nature 467, 1061–1073 (2010).
31. Ferreira, T. & Marchini, J. Modeling interactions with known risk loci—a Bayesian model averaging approach. Ann. Hum. Genet. 75, 1–9 (2011).
32. Wang, S. et al. Absence of human leukocyte antigen class I and II alleles in non-Hodgkin lymphoma etiology. Blood 115, 4820–4823 (2010).
33. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575 (2007).
34. Dixon, A. L. et al. A genome-wide association study of global gene expression. Nat. Genet. 39, 1202–1207 (2007).
35. Cheung, V. G. et al. Polygenic cis- and trans-regulation of human gene expression. PLoS Biol. 8, 9 (2010).
36. Chen, W. M. & Abecasis, G. R. Family-based association tests for genomewide association scans. Am. J. Hum. Genet. 81, 913–926 (2007).
37. Abecasis, G. R. & Wigginton, J. E. Handling marker-marker linkage disequilibrium in genome-wide studies with clustered markers. Am. J. Hum. Genet. 77, 754–767 (2005).
38. Barendsoom, M. & Manke, T. ChroMoS: an integrated web tool for SNP classification, prioritization and functional interpretation. Bioinformatics 29, 2197–2198 (2013).
39. Ernst, J. et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473, 43–49 (2011).

Acknowledgements
Support for individual studies: ATBC—Intramural Research Program of the National Institutes of Health, National Institute on Aging, Division of Cancer Epidemiology and Genetics, U.S. Public Health Service contracts (N01-CN-45165, N01-RC-45035, N01-RC-37004); BCCA (74419). Robert and Kate Niehaus Clinical Cancer Genetics Research Initiative (57470). U01 HG007033; NCI-SEIR—Intramural Research Program of the National Cancer Institute, National Institutes of Health, and Public Health Service (N01-PC-56546,N01-PC-56709, N01-PC-57015, N01-PC-57105). The NIH was supported in part by National Institutes of Health grants CA87969, CA94499, CA149445, CA098812, CA098566 (K.A.B.), and K07 CA115687 (B.M.B.). We would like to thank the participants and staff of the Nurses’ Health Study for their valuable contributions as well as the following state cancer registries for their help: AL, CA, CT, DE, FL, GA, MA, ME, MI, MN, NY, NC, ND, OH, OK, OR, PA, RI, SC, TN, TX, VA, WA, WV. In addition, this study was approved by the Connecticut Department of Public Health (DPH) Human Investigations Committee. Certain data used in this publication were obtained from the DPH. The authors assume full responsibility for analyses and interpretation of these data; Iowa-Mayo SPORE (G.J.W., J.R.C., T.E.W.—National Institutes of Health (CA97274). Specialized Programs of Research Excellence (SPORE) in Human Cancer (P50 CA75774). Molecular Epidemiology of Non-Hodgkin Lymphoma Survival (R01 CA92153). National Center for Advancing Translational Science (UL1 TR000135). MCCS (G.G.G., G.S.)—The Melbourne Collaborative Cohort Study research fund 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; MD Anderson (X.W.)—Institutional support to the Center for Translational and Public Health Genomics; MSKCC (K.O.)—Geoffrey Beene Cancer Research Center, Lymphoma Foundation (LFS541). Barbara K. Lipman Lymphoma Research Fund (74419), Robert and Kate Niehaus Clinical Cancer Genetics Research Initiative (57470), U01 HG007033; NCI-SEIR—Intramural Research Program of the National Cancer Institute, National Institutes of Health, and Public Health Service (N01-PC-56546,N01-PC-56709, N01-PC-57015, N01-PC-57105). The NIH was supported in part by National Institutes of Health grants CA87969, CA94499, CA149445, CA098812, CA098566 (K.A.B.), and K07 CA115687 (B.M.B.). We would like to thank the participants and staff of the Nurses’ Health Study for their valuable contributions as well as the following state cancer registries for their help: AL, CA, CT, DE, FL, GA, MA, ME, MI, MN, NY, NC, ND, OH, OK, OR, PA, RI, SC, TN, TX, VA, WA, WV. In addition, this study was approved by the Connecticut Department of Public Health (DPH) Human Investigations Committee. Certain data used in this publication were obtained from the DPH. The authors assume full responsibility for analyses and interpretation of these data; NS (C.M.Vaide). The NIH was also supported by the Australian National Health and Medical Research Council (DP990602) and the Sardinian Department of Health Services as part of the statewide 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 HHSN261201000140C awarded to the Cancer Prevention Institute of California, contract HHSN261201000035C awarded to the University of Southern California, contract HHSN261201000346C awarded to the Public Health Institute; and the Centers for Disease Control and Prevention’s National Program of Cancer Registries, under agreement #1U8 DP000807-01 awarded to the Public Health Institute. The ideas and opinions expressed herein are those of the authors, and endorsement by the State of California, the California Department of Health Services, the National Cancer Institute, or the Centers for Disease Control and Prevention or their contractors and subcontractors is not intended nor should be inferred. WHI—WHI investigators are: Program Office—(National Heart, Lung, and Blood Institute, Bethesda, Maryland) Jacques Rossouw, Shari Ludlam, Dale Burwen, Joan McGowan, Leslie Ford, and Nancy Geller; Clinical Coordinating Center—(Fred Hutchinson Cancer Research Center, Seattle, WA) Garnet Anderson, Ross Prentice, Andrea LaCroix, and Charles Kooperberg; Investigators and Academic Centers—(Brigham and Women’s Hospital, Boston, MA) JoAnn E. Eary and Elizabeth Ganz; (Cancer Prevention Research Institute of Texas, Houston, TX) Linda G. Cohn and Robert C. Vinik; (University of North Carolina, Chapel Hill) Cheryl Fry et al.; (University of Utah, Salt Lake City, UT) Erin E.10.1038/ncomms6751 | DOI: 10.1038/ncomms6751 | www.nature.com/naturecommunications
Author contributions

J.Vijai, S.I.B., C.F.S., L.C., B.M.B., S.S.W., A.R.B.-W., Q.L., H.H., W.C., L.R.T., J.J.S., Y.Z., M.P.P., L.V., G.D., B.G., P.M.B., L.C., B.M.B., S.S.W., A.R.B.-W., Q.L., R.C.H.V., C.P., S.M.A., B.K.L., J.R., K.E.N., J.G., H.H., W.C., N.B., L.R.T., J.J.S., J.T., Y.Z., M.P.P., G.G.G., R.S.K., A.Z.-J., M.G.E., A.M. Monnereau, K.A.B., D.A., T.I., D.J.V., A.Maria, M.C., T.T., A.J.N., A.D., M.I., C.A.T., T.E.W., T.M.H., G.J.W., M.T.S., E.A.H., R.D.J., L.F.T., Y.Y., H.-O.A., K.E.S., A.J.D.R., P.H., L.M.M., R.K.S., Y.B., P. Boffetta, P. Brennan, L.F., M. Maynadie, J.M., A. Staines, W.R.D., C.M.V., B.K.A., A.K., T.Z., T.R.H., G.S., P.Y., G.M.F., R.R., L.M., J.C., E.G., P.K., J. Virtamo, A. Smith, E.K., E.R., B.C.H.C., J.F.F., X.W., J.R.C., K.O., N.R. and A.N. organized and designed the study. J.Vijai, S.I.B., A.H., X.W., J.R.C., K.O., S.J.C. and N.R. conducted and supervised the genotyping of samples. J.Vijai, Z.W., S.I.B., C.F.S., S.d.S., L.C., P.I.W.d.B., J.G., M.Y., C.C.C., L.L., J.H., B.M., S.J.C. and N.R. contributed to the design and execution of statistical analysis. J.Vijai, Z.W., S.I.B., C.F.S., J.R.C., K.O., S.J.C., N.R. and A.N. wrote the first draft of the manuscript. J.Vijai, C.F.S., S.L.S., S.d.S., M. Melbye, R.G., P.M.B., L.C., B.M.B., S.S.W., A.R.B.-W., Q.L., R.C.H.V., C.P., S.M.A., B.K.L., J.R., K.E.N., J.G., H.H., W.C., N.B., L.R.T., J.J.S., J.T., Y.Z., M.P.P., G.G.G., R.S.K., A.Z.-J., M.G.E., A. Monnereau, K.A.B., D.A., T.I., D.J.V., A.M. Maria, M.C., T.T., A.J.N., A.D., M.I., C.A.T., T.E.W., T.M.H., G.J.W., M.T.S., E.A.H., R.D.J., L.F.T., Y.Y., H.-O.A., K.E.S., A.J.D.R., P.H., L.M.M., R.K.S., Y.B., P. Boffetta, P. Brennan, L.F., M. Maynadie, J.M., A. Staines, W.R.D., C.M.V., B.K.A., A.K., T.Z., T.R.H., G.S., P.Y., G.M.F., R.R., L.M., J.C., E.G., P.K., J. Virtamo, A. Smith, E.K., E.R., B.C.H.C., J.F.F., X.W., J.R.C., K.O., N.R. and A.N. conducted the epidemiological studies and contributed samples to the GWAS and/or follow-up genotyping. All authors contributed to the writing of the manuscript.

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://npg.nature.com/reprintsandpermissions/

How to cite this article: Vijai, J. et al. A genome-wide association study of marginal zone lymphoma shows association to the HLA region. Nat. Commun. 6:5751 doi: 10.1038/ncomms6751 (2015).

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/