Genome-wide association study identifies multiple susceptibility loci for diffuse large B cell lymphoma

Diffuse large B cell lymphoma (DLBCL) is the most common lymphoma subtype and is clinically aggressive. To identify genetic susceptibility loci for DLBCL, we conducted a meta-analysis of 3 new genome-wide association studies (GWAS) and 1 previous scan, totaling 3,857 cases and 7,666 controls of European ancestry, with additional genotyping of 9 promising SNPs in 1,359 cases and 4,557 controls. In our multi-stage analysis, five independent SNPs in four loci achieved genome-wide significance marked by rs116446171 at 6p25.3 (EXOC2; $P = 2.33 \times 10^{-21}$), rs2523607 at 6p21.33 (HLA-B; $P = 2.40 \times 10^{-10}$), rs79480871 at 2p23.3 (NCOA1; $P = 4.23 \times 10^{-8}$) and two independent SNPs, rs13255292 and rs4733601, at 8q24.21 (PVT1; $P = 9.98 \times 10^{-13}$ and 3.63 $\times 10^{-11}$, respectively). These data provide substantial new evidence for genetic susceptibility to this B cell malignancy and point to pathways involved in immune recognition and immune function in the pathogenesis of DLBCL.

In stage 1, with adjustment for sex, age and four eigenvectors (Online Methods), we observed an enrichment of SNPs with smaller $P$ values in comparison to the null distribution of no association in the quantile-quantile plot, with a $\lambda$ value (genomic inflation factor) of 1.016 (Supplementary Fig. 2). Two SNPs exceeded the threshold for genome-wide significance ($P < 5 \times 10^{-8}$), and 20 SNPs showed highly suggestive associations ($P < 5 \times 10^{-7}$) (Supplementary Fig. 3). All but one SNP (rs13255292) mapped to the human leukocyte antigen (HLA) region of chromosome 6 (29.5–33.2 Mb on human genome version 19 (hg19) coordinates).

In stage 2, we included data from 2 unpublished GWAS (GELA/EPIC and Mayo) plus 1 published GWAS (UCSF7), totaling 1,196 DLCLB cases and 1,445 controls (Online Methods and Supplementary Tables 1 and 3). Because different genotyping platforms were used, we imputed common SNPs for each study on the basis of 1,000 Genomes Project release version 3 (ref. 12) using IMPUTE2 (ref. 13) (Supplementary Table 4). In meta-analysis of all genotyped and high-quality imputed SNPs from stages 1 and 2 ($n = 8,363,971$), we identified 19 SNPs at genome-wide significance ($P < 5 \times 10^{-8}$) (Supplementary Table 5) and 134 SNPs at a suggestive level of significance ($P < 5 \times 10^{-7}$) (Supplementary Table 6); 123 of the 153 total SNPs mapped to the HLA region on chromosome 6. On the basis of these results, we selected and successfully designed TaqMan primers for 8 promising SNPs ($P < 5 \times 10^{-9}$) outside the HLA region and 1 SNP from the HLA region for stage 3 de novo genotyping in an additional 1,359 DLBCL cases and 4,557 controls (Online Methods and Supplementary Tables 1 and 3).

In a meta-analysis of all three stages (Supplementary Table 7), we identified four non-HLA SNPs in three newly associated loci at 6p25.3 (rs116446171, $P = 2.33 \times 10^{-21}$) near EXOC2, 8q24.21 (rs13255292, $P = 9.98 \times 10^{-13}$; rs4733601, $P = 3.63 \times 10^{-11}$) near PVTI and MYC, and 2p23.3 (rs79480871, $P = 4.23 \times 10^{-8}$) near NCOA1 (Fig. 1a–c and Table 1). The two 8q24.21 SNPs displayed minimal linkage disequilibrium (LD; $r^2 = 0.03$ in the 1,000 Genomes Project CEU population of Northern and Western European ancestry). Furthermore, in conditional analysis, both rs13255292 (conditional OR = 1.22; $P = 1.39 \times 10^{-12}$) and rs4733601 (conditional OR = 1.18; $P = 2.84 \times 10^{-10}$) remained genome-wide significant; together, these data support the presence of two independent SNPs associated with DLBCL at 8q24.21. We also observed two suggestive SNPs ($P < 5 \times 10^{-7}$) (Supplementary Table 8), one at 5q31.3 (rs79464052, $P = 5.57 \times 10^{-8}$) in ARAP3 (Supplementary Fig. 4) and one at 3q13.33 (rs2681416), although the latter SNP did not replicate in stage 2 or 3.

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

Received 26 June; accepted 4 September; published online 28 September 2014; doi:10.1038/ng.3105

© 2014 Nature America, Inc. All rights reserved.
Within the HLA region, rs2523607 ($P = 3.35 \times 10^{-9}$) was carried forward for replication in stage 3. This SNP, located at 6p21 in HLA-B, reached combined $P = 2.40 \times 10^{-10}$ in the meta-analysis of all three stages (Fig. 1d and Table 1). To further evaluate the association of HLA variants with DLBCL risk, we imputed classical HLA alleles at six loci (HLA-A, HLA-B, HLA-C, HLA-DRB1, HLA-DQA1 and HLA-DQB1) in the four GWAS data sets from stages 1 and 2 and conducted a meta-analysis (Online Methods). The imputation accuracy for HLA types was high (>95.2%) when imputed types were compared to the results from HLA sequencing (four-digit resolution) previously performed on a subset of the National Cancer Institute (NCI) samples scanned as part of this study in stage 1 (Online Methods). Of all the SNPs and classical HLA alleles tested across the major histocompatibility complex (MHC), only the rs2523607 SNP (OR = 1.34; $P = 3.3 \times 10^{-9}$ in stages 1 and 2) and the HLA-B*08:01 classical allele (OR = 1.30; $P = 3.16 \times 10^{-8}$ in stages 1 and 2) reached genome-wide significance (Supplementary Table 9). These markers were in very high LD ($r^2 = 0.91$), and, after adjusting for the
The EXOC2 protein functions at the interface between host defense and epithelial cell polarity, cell motility and cytokinesis and in protection and the intercellular transfer of viral proteins and virions. The EXOC2 protein functions at the interface between host defense and epithelial cell polarity, cell motility and cytokinesis and in protection and the intercellular transfer of viral proteins and virions.

The susceptibility locus at 6p25.3 (rs116446171) maps near a plausible DLBCL candidate gene, EXOC2 (encoding exocyst complex component 2), whose gene product forms part of a large multiprotein complex responsible for vesicle trafficking and maintenance and the intercellular transfer of viral proteins and virions.

The susceptibility locus at 6p25.3 (rs116446171) maps near a plausible DLBCL candidate gene, EXOC2 (encoding exocyst complex component 2), whose gene product forms part of a large multiprotein complex responsible for vesicle trafficking and maintenance and the intercellular transfer of viral proteins and virions.

The susceptibility locus at 6p25.3 (rs116446171) maps near a plausible DLBCL candidate gene, EXOC2 (encoding exocyst complex component 2), whose gene product forms part of a large multiprotein complex responsible for vesicle trafficking and maintenance and the intercellular transfer of viral proteins and virions.

The susceptibility locus at 6p25.3 (rs116446171) maps near a plausible DLBCL candidate gene, EXOC2 (encoding exocyst complex component 2), whose gene product forms part of a large multiprotein complex responsible for vesicle trafficking and maintenance and the intercellular transfer of viral proteins and virions.

The susceptibility locus at 6p25.3 (rs116446171) maps near a plausible DLBCL candidate gene, EXOC2 (encoding exocyst complex component 2), whose gene product forms part of a large multiprotein complex responsible for vesicle trafficking and maintenance and the intercellular transfer of viral proteins and virions.

The susceptibility locus at 6p25.3 (rs116446171) maps near a plausible DLBCL candidate gene, EXOC2 (encoding exocyst complex component 2), whose gene product forms part of a large multiprotein complex responsible for vesicle trafficking and maintenance and the intercellular transfer of viral proteins and virions.

The susceptibility locus at 6p25.3 (rs116446171) maps near a plausible DLBCL candidate gene, EXOC2 (encoding exocyst complex component 2), whose gene product forms part of a large multiprotein complex responsible for vesicle trafficking and maintenance and the intercellular transfer of viral proteins and virions.

The susceptibility locus at 6p25.3 (rs116446171) maps near a plausible DLBCL candidate gene, EXOC2 (encoding exocyst complex component 2), whose gene product forms part of a large multiprotein complex responsible for vesicle trafficking and maintenance and the intercellular transfer of viral proteins and virions.

The susceptibility locus at 6p25.3 (rs116446171) maps near a plausible DLBCL candidate gene, EXOC2 (encoding exocyst complex component 2), whose gene product forms part of a large multiprotein complex responsible for vesicle trafficking and maintenance and the intercellular transfer of viral proteins and virions.

The susceptibility locus at 6p25.3 (rs116446171) maps near a plausible DLBCL candidate gene, EXOC2 (encoding exocyst complex component 2), whose gene product forms part of a large multiprotein complex responsible for vesicle trafficking and maintenance and the intercellular transfer of viral proteins and virions.

The susceptibility locus at 6p25.3 (rs116446171) maps near a plausible DLBCL candidate gene, EXOC2 (encoding exocyst complex component 2), whose gene product forms part of a large multiprotein complex responsible for vesicle trafficking and maintenance and the intercellular transfer of viral proteins and virions.

The susceptibility locus at 6p25.3 (rs116446171) maps near a plausible DLBCL candidate gene, EXOC2 (encoding exocyst complex component 2), whose gene product forms part of a large multiprotein complex responsible for vesicle trafficking and maintenance and the intercellular transfer of viral proteins and virions.

The susceptibility locus at 6p25.3 (rs116446171) maps near a plausible DLBCL candidate gene, EXOC2 (encoding exocyst complex component 2), whose gene product forms part of a large multiprotein complex responsible for vesicle trafficking and maintenance and the intercellular transfer of viral proteins and virions.

The susceptibility locus at 6p25.3 (rs116446171) maps near a plausible DLBCL candidate gene, EXOC2 (encoding exocyst complex component 2), whose gene product forms part of a large multiprotein complex responsible for vesicle trafficking and maintenance and the intercellular transfer of viral proteins and virions.

The susceptibility locus at 6p25.3 (rs116446171) maps near a plausible DLBCL candidate gene, EXOC2 (encoding exocyst complex component 2), whose gene product forms part of a large multiprotein complex responsible for vesicle trafficking and maintenance and the intercellular transfer of viral proteins and virions.
the p160/steroid receptor coactivator (SRC) family, and the latter gene encodes a protein that is a member of a family of proteins involved in clathrin-mediated endocytosis and might also augment the induction of T cell receptor endocytosis. However, our bioinformatic analysis did not identify a clear link between our SNP and the genes in this region, supporting the need to refine this signal in future work.

Through imputation with SNP2HLA46, our strongest associations in the HLA region were with the HLA-B SNP rs2523607 and the HLA-B*08:01 allele, which are in very high LD; on the basis of our available sample size, we cannot definitely rule out an orthogonal effect of rs2523607 in favor of HLA-B*08:01. HLA-B encodes an HLA class 1 heavy chain, which heterodimerizes with a light chain (β2 microglobulin) molecule, whereby it has a central role in presenting intracellularly processed self or foreign antigens to CD8+ cytotoxic T lymphocytes. Class I molecules have been linked to a variety of immune-mediated diseases and cancers, including Hodgkin lymphoma, follicular lymphoma, DLBCL,5,14,37,38 and, more recently, marginal zone lymphoma (J. Vijai, Z.W., S.I.B., C.F.S. and S.L.S. et al., unpublished data). Our results strongly suggest HLA-B*08:01 as the primary MHC association with DLBCL risk. This classical allele is carried by the so-called ancestral 8.1 haplotype associated with complex diseases (for example, type 1 diabetes)39. Classical alleles for other HLA loci might also be involved (including those on the 8.1 haplotype), but larger sample sizes will be required to evaluate this possibility.

Our study represents the largest DLBCL GWAS in individuals of European descent. We did not observe a notable signal for a locus previously reported for DLBCL at 3q27 in East Asians40, rs6773854 (reported with OR = 1.47 and \(P = 1.14 \times 10^{-11}\)), identified with a discovery set of 253 B cell NHL cases (148 DLBCLs). Although our current study had a similar minor allele frequency (MAP) for this SNP of 0.22 among controls, we observed an OR of 1.06 and a \(P\) value of 0.81 for association (Supplementary Table 13), suggesting that the reported marker might not be correlated with the functional susceptibility allele in individuals of European ancestry. Of the two suggestive loci \((P < 5 \times 10^{-7})\) reported in the literature40,41, we did not observe an association for rs751837 with DLBCL (OR = 0.97; \(P = 0.46\)), identified in a small Chinese GWAS (OR = 3.51; \(P = 3.3 \times 10^{-7}\))40, but we did observe a consistent albeit attenuated association for rs10484561 (OR = 1.18; \(P = 1.5 \times 10^{-4}\)), which was initially reported for a subset of the studies in stage 1 (OR = 1.36; \(P = 1.46 \times 10^{-7}\)42). Previously, an InterLymph study of ~1,800 DLBCLs and ~6,500 controls reported a strong signal for a dinucleotide haplotype in the LTA-TNF locus (LTA 252A>G (rs909253)/TNF 308G>A (rs1800629)) at 6p21.3 (OR = 1.31; \(P = 2.9 \times 10^{-8}\))43. Although nearly all of the cases from the previous publication were included in our current GWAS, the signal we observed overall was weaker (OR = 1.15; \(P = 8.5 \times 10^{-4}\)). The attenuation was not explained by study design (case-control, cohort) or adjustment for population substructure (data not shown) but could be due to population sampling differences, heterogeneity or chance.

To explore the heritability of DLBCL, we estimated the contribution of all common SNPs to the variance explained by fitting all genotyped autosomal SNPs simultaneously using the method proposed by Yang et al.44 in the stage 1 data set. We estimated that common SNPs, including but not limited to the loci discovered in this study, explained approximately 16% of the variance in DLBCL risk overall.

In summary, our findings represent an important step in defining the contribution of common genetic variants to risk for DLBCL. Our findings are notable because we have newly defined associations of several regions with susceptibility to DLBCL, and these regions harbor plausible candidate genes for further investigation. Further studies are required to discover additional common susceptibility loci as well as functional analyses that can explain the biological underpinnings of these new susceptibility loci.

URLs: HapMap, http://www.hapmap.org/

METHODS

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

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank C. Allmer, E. Angelucci, A. Bigelow, S. Buehler, K. Butterbach, A. Chabrier, J.M. Conners, M. Corines, M. Cornelis, K. Corsano, H. Dykes, L. Ershler, A. Gabbas, R.P. Gallagher, R.D. Gascoyne, P. Hui, I. Irish, L. Jacobus, K. Laireszeg, A.S. Lai, J. Lunde, M. McAdams, R. Montalvan, L. Padyukov, M. Rais, T. Rattle, L. Rigacci, K. Snyder, G. Specchia, M. Stagner, G. Thomas, C. Tornow, G. Wood and M. Yang. The overall GWAS project was supported by the Intramural Program of the US National Institutes of Health/National Cancer Institute. A list of support provided to individual studies appears in the Supplementary Note.

AUTHORS CONTRIBUTIONS

J.R.C., S.I.B., S.S.W., A.N., A.R.B.-W., Q.L., G. Severi, M. Melbye, L.R.T., M.P.P., C.L., R.B.M., S.L.S., S.d.S., K.E.S., C.F.S., N.R. and S.J.C. organized and designed the study. J.R.C., L.C., L.B., A.H., P.M.B., E.A.H., L.S., G. Salles, C.F.S., N.R. and S.J.C. conducted and supervised the genotyping of samples. J.R.C., S.I.B., J. Vijai, Z.W., M.Y., L.C., P.E.W.d.B., D.C., J.G., D. Zhi, Y.W.A., J.H., B.M., L.I., J.-H.P., C.C.C., N.C., S.d.S., K.E.S., C.F.S., N.R. and S.J.C. contributed to the design and execution of statistical analyses. J.R.C., S.I.B., J. Vijai, H.G., J.M., S.S.W., Z.W., M.Y., L.C., A.N., D.C., A.R.M., C.F.S., A.I., J.D.R., L.G., P.B.R., F.S., K.E.S., N.R. and S.J.C. wrote the first draft of the manuscript. J.R.C., J. Vijai, H.G., J.M., S.S.W., Z.W., M.Y., L.C., A.N., D.C., A.R.M., C.F.S., A.I., J.D.R., L.G., P.B.R., F.S., K.E.S., N.R. and S.J.C. wrote the first draft of the manuscript. J.R.C., J. Vijai, H.G., J.M., S.S.W., Z.W., M.Y., L.C., A.N., D.C., A.R.M., C.F.S., A.I., J.D.R., L.G., P.B.R., F.S., K.E.S., N.R. and S.J.C. wrote the first draft of the manuscript. J.R.C., J. Vijai, H.G., J.M., S.S.W., Z.W., M.Y., L.C., A.N., D.C., A.R.M., C.F.S., A.I., J.D.R., L.G., P.B.R., F.S., K.E.S., N.R. and S.J.C. wrote the first draft of the manuscript. J.R.C., J. Vijai, H.G., J.M., S.S.W., Z.W., M.Y., L.C., A.N., D.C., A.R.M., C.F.S., A.I., J.D.R., L.G., P.B.R., F.S., K.E.S., N.R. and S.J.C. wrote the first draft of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Siegel, R., Naishadham, D. & Jemal, A. Cancer statistics, 2013. CA Cancer J. Clin. 63, 11–30 (2013).
2. Flowers, C.R., Sinha, R. & Vose, J.M. Improving outcomes for patients with diffuse large B-cell lymphoma. CA Cancer J. Clin. 60, 393–408 (2010).
3. Wang, S.S. et al. Family history of hematopoietic malignancies and risk of non-Hodgkin lymphoma (HNL): a pooled analysis of 10,211 cases and 11,905 controls from the International Lymphoma Epidemiology Consortium (InterLymph). Blood 109, 3479–3488 (2007).
4. Goldin, L.R., Bjorkholm, M., Kristinsson, S.Y., Turesson, I. & Landgren, O. Highly increased familial risks for specific lymphoma subtypes. Br. J. Haematol. 146, 91–94 (2009).
5. Skibola, C.F. et al. Tumor necrosis factor (TNF) and lymphotoxin-α (LTA) polymorphisms and risk of non-Hodgkin lymphoma in the InterLymph Consortium. Am. J. Epidemiol. 171, 267–276 (2010).
6. Skibola, C.F. et al. Genetic variants at 6p21.33 are associated with susceptibility to follicular lymphoma. Nat. Genet. 41, 873–875 (2009).
7. Conde, L. et al. Genome-wide association study of follicular lymphoma identifies a risk locus at 6p21.32. Nat. Genet. 42, 661–664 (2010).
8. Smedby, K.E. et al. GWAS of follicular lymphoma reveals allelic heterogeneity at 6p21.32 and suggests shared genetic susceptibility with diffuse large B-cell lymphoma. PLoS Genet. 7, e1001378 (2011).
9. Vijai, J. et al. Susceptibility loci associated with specific and shared subtypes of lymphoid malignancies. PLoS Genet. 9, e1003220 (2013).
10. Tan, D.E. et al. Genome-wide association study of B cell non-Hodgkin lymphoma identifies 3q27 as a susceptibility locus in the Chinese population. Nat. Genet. 45, 804–807 (2013).
11. Wang, Z. et al. Improved imputation of common and uncommon SNPs with a new reference set. Nat. Genet. 44, 6–7 (2012).

12. Abecasis, G.R. et al. A map of human genome variation from population-scale sequencing. Nature 467, 1061–1073 (2010).

13. 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).

14. Wang, S.S. et al. Human leukocyte antigen class I and II alleles in non-Hodgkin lymphoma etiology. Blood 115, 4820–4823 (2010).

15. Ward, L.D. & Kellis, M. HaploReg: a resource for exploring chromatin states, prioritization and functional interpretation. Bioinformatics 29, 2197–2198 (2013).

16. Ernst, J. et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473, 43–49 (2011).

17. Mukerji, J., Olivieri, K.C., Misra, V., Agopian, K.A. & Gabuzda, D. Proteomic analysis of HIV-1 nef cellular binding partners reveals a role for eoxcytosome components in mediating enhancement of intercellular nanotube formation. Retrovirology 9, 33 (2012).

18. Mantovani, A. & Balkwill, F. RalGTPases and cancer: linchpin support of the Ras-mediated tumorigenesis. Nat. Rev. Cancer 8, 133–140 (2008).

19. Nandi, S. et al. Gene expression signatures in diffuse large B-cell lymphoma. Nat. Genet. 43, 830–837 (2011).

20. Bodemann, B.O. & White, M.A. Ral GTPases and cancer: linchpin support of the Ras-mediated tumorigenesis. Nat. Rev. Cancer 8, 223–231 (2010).

21. Kassies, D. Ral GTPases in tumorigenesis: emerging from the shadows. Exp. Cell Res. 319, 2337–2342 (2013).

22. Di Bernardo, M.C. et al. A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia. Nat. Genet. 40, 1204–1210 (2008).

23. Berndt, S.I. et al. Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia. Nat. Genet. 45, 868–876 (2013).

24. Wang, S.S. et al. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42, 1271–1276 (2010).

25. Wang, S.S. et al. Common gene variants in the tumor necrosis factor (TNF) and TNF receptor superfamilies and NF-kB transcription factors and non-Hodgkin lymphoma risk. PLoS ONE 4, e5360 (2009).

26. Wacholder, S., Yeager, M. & Liao, L.M. Invited commentary: more surprises from a gene desert. Am. J. Epidemiol. 175, 488–491 (2012).

27. Crowther-Swanepoel, D. et al. Common variants at 2q37.3, 8q24.21, 15q21.3 and 19q13.4 influence chronic lymphocytic leukemia risk. Nat. Genet. 42, 132–136 (2010).

28. Enciso-Mora, V. et al. A genome-wide association study of Hodgkin’s lymphoma identifies new susceptibility loci at 2p16.1 (REL), 8q24.21 and 10p14 (GATA3). Nat. Genet. 42, 1126–1130 (2010).

29. Graham, M. & Adams, J.M. Chromosome 8 breakpoint far 3’ of the c-myc oncogene in a Burkitt’s lymphoma 2:8 variant translocation is equivalent to the murine pvt-1 locus. EMBO J 5, 2845–2851 (1986).

30. Love, C. et al. The genetic landscape of mutations in Burkitt lymphoma. Nat. Genet. 44, 1321–1325 (2012).

31. Savage, K.J. et al. MYC gene rearrangements are associated with a poor prognosis in diffuse large B-cell lymphoma patients treated with R-CHOP chemotherapy. Blood 114, 3533–3537 (2009).

32. Pasqualucci, L. et al. Analysis of the coding genome of diffuse large B-cell lymphoma. Nat. Genet. 43, 830–837 (2011).

33. Ollate, S.A., Tsai, S.Y., Tsai, M.J. & O’Malley, B.W. Sequence and characterization of a coactivator for the steroid hormone receptor superfamily. Science 270, 1354–1357 (1995).

34. Novokhatksa, O. et al. Adaptor proteins intersect 1 and 2 bind similar proline-rich ligands but are differentially recognized by SH2 domain-containing proteins. PLoS ONE 8, e075046 (2013).

35. McGaw, M.K. et al. The intersec 2 adaptor links Wiskott Aldrich Syndrome protein (WASp)-mediated actin polymerization to T cell antigen receptor endocytosis. J. Exp. Med. 194, 1777–1787 (2001).

36. Jia, X. et al. Imputing amino acid polymorphisms in human leukocyte antigens. Proc. Natl Acad. Sci. USA 102, 92–97 (2005).

37. Howell, W.M. HLA and disease: guilt by association. Int. J. Immunogenet. 41, 1–12 (2014).

38. Klitz, W., Aldrich, C.L., Fildes, N., Horning, S.J. & Begovich, A.B. Localization of predisposition to Hodgkin disease in the HLA class II region. Am. J. Hum. Genet. 54, 497–505 (1994).

39. Price, P. et al. The genetic basis for the association of the B1 ancestral haplotype (B1, BS, DR3) with multiple immunopathological diseases. Immunol. Rev. 167, 257–274 (1999).

40. Kumar, V. et al. Common variants on 14q32 and 13q12 are associated with DLBCL susceptibility. J. Hum. Genet. 56, 436–439 (2011).

41. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42, 565–569 (2010).

James R Cerhan1,94, Sonja I Berndt2,94, Joseph Vijai3,94, Hervé Ghèsquieres4,5,94, James McKay6,94, Sophia S Wang7,94, Zhaoming Wang8,94, Meredith Yeager8, Lucia Conde9,10, Paul I W de Bakker11,12, Alexandra Nieters13, David Cox14, Laurie Burdett8, Alain Monnereau15–17, Christopher R Flowers18, Anneclaire J De Roos19,20, Angela R Brooks-Wilson21,22, Qing Lan2, Gianluca Severi23–25, Mads Melbye26,27, Jian Gu28, Rebecca D Jackson29, Eleanor Kane30, Lauren R Teras31, Mark P Purdie3, Claire M Vajdic32, John J Spinelli33,34, Graham G Giles24,25, Demetrius Albanes2, Rachel S Kelly35,36, Mariagrazia Zucca37, Kimberly A Bertrand38,39, Anne Zeleniuch-Jacquotte38,39,40, Charles Lawrence41, Amy Hutchinson42, Degui Zhi42, Thomas M Habermann43, Brian K Link44, Anne J Novak45, Ahmet Dogan46, Yan W Asmann46, Mark Liebow43, Carrie A Thompson43, Stephen A Ansell43, Thomas E Witzig43, George J Weiner44, Amelie S Veron44, Diana Zelenika47, Hervé Tilly48, Corinne Haïoun49, Thierry Jo Molina50, Henrik Hjalgrim26, Bengt Gillemot51,52, Hans-Olov Adami35,53, Paige M Bracci54, Jacques Riby9,10, Martyn T Smith10, Elizabeth A Holly55, Wendy Cozen55,56, Patricia Hartge5, Lindsay M Morton2, Richard K Severson57, Lesley F Tinker59, Kari E North58,59, Nikolaus Becker60, Yolanda Benavente61,62, Paolo Bofetta63, Paul Brennan64, Lenka Foretova65, Marc Maynadie66, Anthony Staines67, Tracy Lightfoot68, Simon Crouch30, Alex Smith30, Eve Roman30, W Ryan Diver31, Kenneth Offit3, Andrew Zelenetz3, Andrew Zelenetz3, Robert J Klein3, Danylo J Villano3, Tongzhang Zheng69, Yawei Zhang69, Theodore R Holford70,99, Anne Kricker70, Jenny Turner71,72, Melissa C Southey73, Jacqueline Clavel15,16, Jarmo Virtamo74, Stephanie Weinstein2, Elio Riboli75, Paolo Vineis23,36, Rudolph Kaaks60, Dimitrios Trichopoulos35,76,77, Roel C Vermeulen12,78, Heiner Boeing79, Anne Tjonneland80, Emanuele Angelucci81, Simonetta Di Lollo82, Marco Rais83, Brenda M Birnbaum38, Francine Laden35,38,84, Edward Giovannucci35,38,85, Peter Kraft35,86, Jinyan Huang35, Baoshan Ma35,87, Yuanying Ye28, Brian C H Chiu88, Joshua Sampson2, Liming Liang35,86, Ju-Hyun Park89, Charles C Chung2, Dennis D Weisenburger90, Nilanjan Chatterjee2, Joseph F Fraumeni Jr2, Susan L Slager1, Xifeng Wu28,95, Silvia de Sanjose61,62,95, Karin E Smedby91,95, Gilles Salles59,92,93,95, Christine F Skibola9,10,95, Nathaniel Rothman29,25 & Stephen J Chanock2,95
ONLINE METHODS
Stage 1: DLBCL GWAS. As part of a larger initiative, we conducted a GWAS of DLBCL 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 were approved by their respective institutional review boards, and informed consent was obtained for all participants. Cases were ascertained from cancer registries, clinics or hospitals, or through self-report verified by medical and pathology reports. To determine NHL subtype, phenotype data for all NHL cases were harmonized to the hierarchical classification proposed by the InterLymph Pathology Working Group based on World Health Organization (WHO) classification.

All DLBCL cases with sufficient DNA (n = 2,878) 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 BeadChip at the NCI Cancer Genomic Research Laboratory (CGR). Genotypes were called using Illumina GenomStudio software, and quality control duplicates showed >99% concordance. Monomorphic SNPs and SNPs with a call rate <95% were excluded. Samples with a call rate of 59.3%, a mean heterozygosity of <0.25 or >0.33 on the basis of the autosomal SNPs or sex discordance (>5% heterozygosity on the X chromosome for males and <20% heterozygosity on the X chromosome for females) were excluded. Furthermore, unexpected duplicates (>99.9% concordance) and first-degree relatives, defined as having identity-by-descent (IBD) sharing of Pi-hat > 0.40, were excluded. Ancestry was assessed using the Genotyping Library and Utilities (GLU) struct.admix module based on the method by Pritchard et al. and participants with <80% European ancestry were excluded (Supplementary Fig. 6). After exclusions, 2,661 (92.5%) cases and 2,685 (94.1%) controls remained (Supplementary Table 2). Genotype data previously generated on the Illumina Omni2.5 BeadChip from an additional 3,536 controls from 3 of the studies (ATBC, CPSII and PLCO) were also included, resulting in a total of 2,661 cases and 6,221 controls for the stage 1 analysis. Of these additional 3,536 controls, 703 (2.3% from each study) were selected to be representative of their cohort and cancer-free individuals, whereas the remainder were cancer-free controls from an unpublished study of prostate cancer in PLCO. SNPs with a call rate of <95%, with a Hardy-Weinberg equilibrium P value of <1 × 10^{-6} or with a MAF of <1% were excluded from analysis, leaving 611,844 SNPs for analysis (Supplementary Table 4). To evaluate population substructure, a principal-components analysis (PCA) was performed using the GLU version 1.0, struct.pca module, which is similar to EIGENSTRAT. Plots of the first five principal components are shown in Supplementary Figure 7. Association testing was conducted assuming a log-additive genetic model, adjusting for age, sex and four significant principal components. All data analysis and management was conducted using GLU.

Stage 2: in silico analysis of three independent DLBCL GWAS. Three independent DLBCL GWAS provided genotyping data for a meta-analysis (Supplementary Table 1), which included data generated with the following commercial SNP microarrays: Illumina HumanHap660W for Mayo (393 DLBCL cases and 172 controls), HumanCNV370-Duo for UCSI (254 DLBCL cases and 748 controls) and HumanHap610K for GELA (549 cases). In all studies, subjects with a genotyping call rate of <95%, duplicates, related individuals and SNPs with a call rate of <95% were removed before imputation (Supplementary Table 4). The GELA study was conducted on cases only; controls were drawn from a pool of 928 individuals from the French component of the EPIC cohort, who were previously scanned on Illumina HumanHap660W or IlluminaHap610K. We subsequently chose a subset of 525 individuals with matched ancestry as determined by PCA. In total, there were 1,196 cases and 1,445 controls in stage 2.

Imputation was conducted separately for each study in stages 1 and 2 using IMPUTE2 (ref. 13) and the 1000 Genomes Project version 3 (ref. 12). The imputation analysis was restricted to common SNPs (MAF cutoff of >0.01 with an imputation accuracy INFO score of >0.3).

Association testing was conducted for each study using SNPTEST version 2, adjusting for age, sex and any significant principal components. We evaluated the top ten eigenvectors for the GELA, Mayo and UCSF studies in each baseline risk model, adjusting for both age and sex. On the basis of the significance level (P < 0.05) of the regression coefficient for eigenvectors, we chose to adjust for three eigenvectors (EV1, EV7 and EV8) for GELA in the final association model, whereas no eigenvectors met the criteria for adjustment for either the Mayo or UCSI study.

All meta-analyses were performed using the fixed-effects inverse variance method based on the P estimates and standard errors from each study.

Stage 3: replication studies and technical validation. In stage 3, 8 SNPs in the most promising loci outside of the HLA region and one SNP from the HLA region (Supplementary Table 7) were taken forward for de novo replication in an additional 1,359 cases and 4,557 controls from 4 studies (Supplementary Table 1), except for rs2523607, which was not genotyped in any of the studies (NCI replication). Genotyping was conducted using custom TaqMan genotyping assays (Applied Biosystems) at the NCI CGR. Each assay was optimized and validated with 270 HapMap samples and additional CEPH samples (SNP500Cancer), and these samples were used as genotyping controls for clustering and reproducibility analysis. All validated assays had 99% or higher concordance with HapMap data, and completion with control DNA was >97%. Duplicates from stage 3 samples (64 pairs; 3%) yielded 100% concordance.

In technical validation, we observed a high correlation of genotyping calls from the OmniExpress microarray with confirmatory TaqMan assays in 455 stage 1 duplicate samples for 2 genotyped (rs13255292, r2 = 1.00; rs4733601, r2 = 1.00) and 4 imputed (rs11644617, r2 = 0.92; rs2523607, r2 = 0.99; rs2681416, r2 = 1.00; rs79480871, r2 = 0.94) SNPs. We also observed a high correlation of genotyping calls from the Illumina HumanHap660W microarray with confirmatory TaqMan assays in stage 2 duplicate samples from the Mayo study (n = 165) for 2 genotyped (rs13255292, r2 = 1.00; rs4733601, r2 = 1.00) and 4 imputed (rs11644617, r2 = 1.00; rs2523607, r2 = 1.00; rs79480871, r2 = 0.85; rs79464052, r2 = 0.95) SNPs.

HLA induction and analysis. We imputed dense SNPs as well as classical HLA alleles (HLA-A, HLA-B, HLA-C, HLA-DRB1, HLA-DQA1 and HLA-DQB1) and coding variants across the HLA region (chr. 6: 29.5–33.2 Mb, hg19) in the stage 1 (NCI) and stage 2 (Mayo, UCSF and GELA/EPIC) studies using SNP2HLA. Imputation was based on a reference panel from the Type 1 Diabetes Genetics Consortium (T1DGC) and consisted of genotypes from 5,223 individuals of European descent who were typed for HLA-A, HLA-B, HLA-C, HLA-DQA1, HLA-DQB1, HLA-DRB1, DP1 and DRB1 4-digit alleles. To assess imputation accuracy, we compared the imputed HLA alleles to DNA sequencing data (to four digits) available for a subset of samples from the NCI GWAS and found high concordance rates for HLA-A (97.3%), HLA-B (98.5%), HLA-C (98.1%) and HLA-DRB1 (97.5%). Owing to the limited number of SNPs (n = 7,253) in the T1DGC reference set, imputation of HLA SNPs was conducted with IMPUTE2 and the 1000 Genomes Project reference set as described above. A total of 68,488 SNPs, 201 classical HLA alleles (at 2- and 4-digit resolution) and 1,038 AA markers, including 103 AA positions that were ‘multiallelic’ with 3 to 6 different residues present at each position, were successfully imputed (INFO score >0.3 for SNPs or r2 >0.3 for alleles and AA markers) and available for analysis. Multiallelic markers were analyzed as binary markers (for example, having the allele present or absent) and using a global test, and a meta-analysis was conducted where we tested SNPs, HLA alleles and AA markers across the HLA region for association with DLBCL using PLINK or SNPTEST as described above.

Expression quantitative trait locus analysis. To evaluate the effect of our top loci (and SNPs found to be in LD on the basis of r2 > 0.8) in the HapMap CEU population, release 28 on gene expression, we conducted an eQTL analysis on lymphoblastoid cell lines using 2 independent data sets: childhood asthma and HapMap. For the childhood asthma data set, peripheral blood lymphocytes were transformed into lymphoblastoid cell lines for 830 parents and offspring from 206 families of European ancestry. Using extracted RNA, gene expression was assessed with the Affymetrix HG-U133 Plus 2.0 chip. Genotyping was conducted using the Illumina HumanIM BeadChip and Illumina HumanHap300K BeadChip, and imputation was performed using data from the 1000 Genomes Project. All SNPs selected for replication were tested for cis associations (defined as an effect on gene transcripts for genes
within 1 Mb of the SNP), assuming an additive genetic model adjusted for non-genetic effects in the gene expression value. To gain insight into the relative importance of associations with our SNPs in comparison to other SNPs in the region, we also conducted conditional analyses in which both the DLBCL SNP and the most significant SNP for the particular gene transcript (peak SNP) were included in the same model. Only cis associations that reached \( P \leq 6.8 \times 10^{-3} \), which corresponds to an FDR of 1%, are reported (Supplementary Table 10).

The HapMap data set consisted of a publicly available RNA sequencing data set for transformed lymphoblastoid cell lines from 41 CEPH CEU individuals (HapMap), with samples available from the Gene Expression Omnibus (GEO) repository under accession GSE16921. Genotyping data for the same HapMap CEU individuals were directly downloaded from HapMap. Because rs2523607, rs79480871 and rs116446171 were not genotyped in HapMap, we selected rs3130923, rs6746301 and rs7762424 as respective proxies, as they were the most strongly linked SNPs available in HapMap (\( r^2 = 0.94, 0.69 \) and 0.54 in 1000 Genomes Project CEU data, respectively). Correlation between expression and genotype for each SNP-probe pair was tested using the Spearman’s rank correlation test with \( t \)-distribution approximation and was estimated with respect to the minor allele in the HapMap CEU population. \( P \) values were adjusted using the Benjamini-Hochberg FDR correction, and eQTLs were considered significant at FDR \( \leq 0.05 \) (Supplementary Table 11).

Bioinformatics: ENCODE and chromatin state dynamics. Using 1000 Genomes Project data, we identified SNPs with \( r^2 \geq 0.8 \) with our sentinel SNPs that were reported to be nonsynonymous or nonsense variants. We used HaploReg v2 (ref. 15), which is a tool for exploring noncoding functional annotation using ENCODE data, to evaluate the genome surrounding our SNPs (Supplementary Table 12). To assess chromatin state dynamics, we used ChroMoS16, which has precomputed data from ENCODE on nine cell types based on ChIP-seq analyses17. These precomputed data have genome segmentation performed using the multivariate hidden Markov model to reduce the combinatorial space to a set of interpretable chromatin states. The output from ChroMoS separates 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).

Heritability analyses. To estimate the contribution of all common SNPs to the variance explained, we used the method proposed by Yang et al.44, which was extended to dichotomous traits55 and implemented in Genome-Wide Complex Trait Analysis (GCTA) software55. The genetic similarity matrix was estimated from our stage 1 data using all genotyped autosomal SNPs with a MAF of >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 DLBCL was 0.0074 (ref. 54).

Estimate of recombination hotspots. To identify recombination hotspots in a region, we used SequenceLDhot45, a program that uses the approximate marginal likelihood method and calculates likelihood ratio statistics at a set of possible hotspots. We tested 5 unique sets of 100 control samples. The PHASE v2.1 program was used to calculate background recombination rates57,58, and LD heat maps were visualized in \( r^2 \) using the snp.plotter program59.