Genetic variation near *CXCL12* is associated with susceptibility to HIV-related non-Hodgkin lymphoma

Christian W. Thorball,[1,2] Tiphaine Oudot-Mellakh,[3] Nava Ehsan,[4,5] Christian Hammer,[6,7] Federico A. Santoni,[8] Jonathan Niay,[3] Dominique Costagliola,[9] Cécile Goujard,[10,11] Laurence Meyer,[12] Sophia S. Wang,[13] Shehnaz K. Hussain,[14] Ioannis Theodorou,[3] Matthias Cavassini,[15] Andri Rauch,[16] Manuel Battegay,[17] Matthias Hoffmann,[18] Patrick Schmid,[19] Enos Bernasconi,[20] Huldrych F. Günthard,[21,22] Pejman Mohammadi,[4,5] Paul J. McLaren,[23,24] Charles S. Rabkin,[25] Caroline Besson[25-27] and Jacques Fellay[1,2,28]

[1]School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland; [2]Swiss Institute of Bioinformatics, Lausanne, Switzerland; [3]Centre de Génétique Moléculaire et Chromosomique, GH La Pitié Salpêtrière, Paris, France; [4]The Scripps Research Translational Institute, La Jolla, CA, USA; [5]Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA, USA; [6]Department of Cancer Immunology, Genentech, South San Francisco, CA, USA; [7]Department of Human Genetics, Genentech, South San Francisco, CA, USA; [8]Service of Endocrinology, Diabetology and Metabolism, Lausanne University Hospital, Lausanne, Switzerland; [9]Sorbonne Universités, INSERM, UPMC Université Paris 06, Institut Pierre Louis d'Épidémiologie et de Santé Publique (IPLESP UMRS 1136), Paris, France; [10]INSERM, CESP, U1018, Paris-Sud University, Le Kremlin-Bicêtre, France; [11]Department of Internal Medicine, Bicêtre Hospital, AP-HP, Le Kremlin-Bicêtre, France; [12]INSERM U1018, Centre de Recherche en Épidémiologie et Santé des Population, Paris-Sud University, Paris-Saclay University, Le Kremlin-Bicêtre, France; [13]Division of Health Analytics, City of Hope Beckman Research Institute and City of Hope Comprehensive Cancer Center, Duarte, CA, USA; [14]Department of Medicine, Cedars-Sinai Medical Center, Los Angeles, CA, USA; [15]Service of Infectious Diseases, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland; [16]Department of Infectious Diseases, Bern University Hospital, University of Bern, Switzerland; [17]Department of Infectious Diseases and Hospital Epidemiology, University Hospital Basel, University of Basel, Basel, Switzerland; [18]Division of Infectious Diseases and Hospital Epidemiology, Kantonsspital Olten, Olten, Switzerland; [19]Division of Infectious Diseases, Cantonal Hospital of St. Gallen, St. Gallen, Switzerland; [20]Division of Infectious Diseases, Regional Hospital of Lugano, Lugano, Switzerland; [21]Department of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, Zurich, Switzerland; [22]Institute of Medical Virology, University of Zurich, Zurich, Switzerland; [23]JC Wilt Infectious Diseases Research Center, National Microbiology Laboratory, Public Health Agency of Canada, Winnipeg, Manitoba, Canada; [24]Department of Medical Microbiology and Infectious Diseases, University of Manitoba, Winnipeg, Manitoba, Canada; [25]Infections and Immunoepidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA; [26]CESP, UVSQ, INSERM, Université Paris-Saclay, Villejuif, France; [27]Department of Hematology and Oncology, Hospital of Versailles, Le Chesnay, France and [28]Precision Medicine Unit, Lausanne University Hospital (CHUV) and University of Lausanne, Lausanne, Switzerland.

 doi:10.3324/haematol.2020.247023

Received: January 9, 2020.

Accepted: July 14, 2020.

Pre-published: July 16, 2020.

Correspondence: *JACQUES FELLAY* - jacques.fellay@epfl.ch

Supplementary Methods

Study participants and contributing centers

Swiss HIV Cohort Study (SHCS)

The SHCS is a large, ongoing, multicenter cohort study of HIV-positive individuals that includes >70% of adult living with HIV in Switzerland. At follow-up visits every 6 months, demographic, clinical, laboratory, and ART information has been prospectively recorded since 1988.[1] Cancer diagnoses are verified thoroughly using checking charts including information on biopsies and imaging. To minimize potential treatment bias and population stratification, we only considered as cases patients diagnosed with NHL between 2000 and 2017 and of European ancestry, as determined by principal component analysis (PCA) (supplemental Figure 1A). Controls were matched based on age, ancestry, CD4+ T cell counts and viral load results. To be eligible as controls, they also had to be diagnosed with HIV prior to 2005 and have no registered cancer diagnosis of any type as of 2017. Patients were genotyped using Illumina HumanOmniExpress-24 Beadchips, or genotypes were obtained in the context of a previous GWAS in the SHCS on various platforms including Illumina HumanCore-12, HumanHap550, Human610 and Human1M Beadchips.

French Primo ANRS and ANRS CO16 Lymphovir cohorts (ANRS)

The French ANRS CO16 lymphovir cohort of HIV related lymphomas enrolled adult patients at diagnosis of lymphoma in 32 centers between 2008 and 2015.[2] Pathological materials were centralized, and diagnoses of NHL were based on World Health Organization criteria. Patients were genotyped using Illumina Human Omni5 Exome 4v beadchips. Additional cases and controls (matched on self-reported ancestry) were included from the ANRS PRIMO Cohort,

which has been enrolling patients during primary HIV-1 infection in 95 French Hospitals since 1996.[3] Patients were genotyped using Illumina Sentrix Human Hap300 Beadchips. Only patients of European ancestry, as determined by PCA, were included in the study (supplemental Figure 1B).

The Multicenter AIDS Cohort Study (MACS)

The MACS has enrolled gay and bisexual HIV infected men in 4 US cities since 1984. The NHL cases were diagnosed between 1985 and 2013 (median year of diagnosis: 1992). Data collected include demographic variables (age, race, ethnicity and HIV transmission category), CD4+ T cell count, HIV viral load and tumor histology. Eligible cases had a diagnosis of HIV-related NHL, available genotyping data and at least one CD4+ T cell count obtained within 2 years of the NHL diagnosis. Controls were matched on MACS study site, age at NHL diagnosis (+/- 2 years) and CD4+ T cell count at NHL diagnosis (within the following groups 0-99 / 100 -199 / 200-499 / >499 cells/µL). Patients were genotyped using Illumina HumanHap550 and Human1M Beadchips.[4] As in the other cohorts, only individuals of European ancestry were included, as determined by PCA (supplemental Figure 1C).

Fine mapping of associated regions

Fine mapping of the *CXCL12* locus was performed using PAINTOR (v3.1) to identify the most likely causal variant(s). All variants within 200kb of the top associated SNP and with a p-value below 0.005 were included in the model. The linkage disequilibrium (LD) matrix was created using PLINK and genotype data from the SHCS cohort. PAINTOR was first run against all genomic annotation databases provided with the software, including the FANTOM5, ENCODE and the Roadmap Epigenomics Project. For the final model, the top 5 annotations

based on improvement to model fit and cell type relevance were selected to obtain the posterior probabilities and the 99% credible set of the variants most likely to be causal based on the association from Bayes' factors.

Predictive effect of potentially causal variants

The potential functional impact of the predicted causal variants was assessed using DeepSEA[6], a deep learning-based sequence model trained on available chromatin and transcription factor data from ENCODE and Roadmap Epigenomics. DeepSEA provides a functional significance score for each variant, which is a measure of the evolutionary conservation and the significance of the magnitude of the predicted chromatin effects. For the variants with a functional significance score of less than 0.01, we analyzed the predicted changes in specific chromatin modifications or transcription factor (TF) binding probabilities. Chromatin or TF binding changes with E-values below 0.001 and normalized probabilities of observing a binding event above 0.2 were considered relevant. The TF position weight matrices (PWMs) for TFs with a high probability of binding (normalized probability $\geq$ 50%) were obtained from the JASPAR CORE 5.0 database.[7]

Long-range chromatin interactions

Predicted topological associating domains (TADs) near the genome-wide significant locus in GM12878 lymphoblastoid cells were obtained from publicly available data[8] and visualized using the 3D Genome Browser.[9]

Potential interactions between the genome-wide significant locus and promoters of nearby genes were analyzed using publicly available promoter capture Hi-C data in GM12878

lymphoblastoid cells. The Hi-C data was processed through the CHiCAGO pipeline and visualized with CHiCP.[10,11] Interaction scores ≥ 5 were considered significant, as described previously.[12]

Expression quantitative trait loci (eQTL) analyses

The role of rs7919208 as an eQTL was examined in GEUVADIS[13] and in response to various pathogens, although not including HIV, in the Milieu Intérieur Consortium cohort.[14] Furthermore, eQTL information was also obtained from the GTEx (v7)[15] Portal on 03/22/2019.

Bulk RNA Barcoding and sequencing (BRB-seq)[16] was performed on RNA from peripheral blood mononuclear cells (PBMCs) of 452 individuals from the SHCS with available genotyping data.

Allele-specific effects of rs7919208

To assess allelic effect of rs7919208 SNP on *CXCL12* expression, we used haplotype-level ASE data from the GTEx v8, containing 15,253 samples spanning 49 human tissues and 838 individuals[17,18]. Briefly, this haplotype level data was generated using phASER v1.0.1, which incorporates RNA-seq and DNA-seq data with population phasing to allow phasing over longer distances and aggregating ASE signal from all available SNPs within a gene[19]. We compared allelic imbalance in *CXCL12* expression between the individuals homozygous reference and heterozygous for rs7919208. Allelic imbalance was quantified as the log ratio between the two allelic counts or log allelic fold change (log aFC)[20]. We use the absolute value of log aFC in a one-sided ranksum test to ensure robustness to rare variant effects and phasing errors. All tissue

with at least 20 total ASE counts were included. P-values were corrected for multiple hypothesis testing using the Benjamini-Hochberg procedure.

Transcript usage effect associated with rs7919208

The CXCL12 transcript isoform expression and the effect of the rs7919208 SNP on this transcript isoform expression was assessed across the top six highest expressed tissues for CXCL 12 in GTEx v8.[18] The analysis of the effect on transcript usage was restricted to the two transcripts ENST00000343575.10 and ENST00000374429.6, as they account for 94% of the expressed isoforms. Associations were analyzed using linear regressions and significant associations confirmed with rank based inverse normal transformed values to eliminate potential outlier-driven effects.

Comparison to GWAS hits in the general population

An attempt at replicating variants previously associated with NHL in the general population was performed by extraction of the p-values of the SNPs reported to be associated in previous NHL GWAS. A variant was considered replicated if it had a nominally significant association p-value ($P < 0.05$) plus similar effect direction in the meta-analysis.

The effect of rs7919208 in the general population cohorts was assessed directly using the NIH database for Genotypes and Phenotypes (dbGaP) accession # phs000801 cohorts for chronic lymphocytic leukemia (CLL), DLBCL (Diffuse large B-cell lymphoma), FL (Follicular lymphoma) and MZL (Marginal zone lymphoma) and corresponding controls.[21–24] The genotype data was imputed, processed and analyzed using the same pipeline and methods as

described above for the HIV cohorts, with duplicate samples identified and removed using KING and including age and sex as covariates.

Statistical analyses

All statistical analyses were performed using the R statistical software (v3.3.3), unless otherwise specified.

Data sharing statement

Full summary statistics will be made available in the GWAS catalog (https://www.ebi.ac.uk/gwas) upon publication. The raw genotype data can be obtained through the respective cohorts.

References

1. Schoeni-Affolter F, Ledergerber B, Rickenbach M, et al. Cohort Profile: The Swiss HIV Cohort Study. Int J Epidemiol 2010;39(5):1179–1189.
2. Besson C, Lancar R, Prevot S, et al. Outcomes for HIV-associated diffuse large B-cell lymphoma in the modern combined antiretroviral therapy era. AIDS 2017;31(18):2493.
3. Dalmasso C, Carpentier W, Meyer L, et al. Distinct Genetic Loci Control Plasma HIV-RNA and Cellular HIV-DNA Levels in HIV-1 Infection: The ANRS Genome Wide Association 01 Study. PLoS ONE 2008;3(12):e3907.
4. Fellay J, Ge D, Shianna KV, et al. Common Genetic Variation and the Control of HIV-1 in Humans. PLOS Genetics 2009;5(12):e1000791.
5. Kichaev G, Roytman M, Johnson R, et al. Improved methods for multi-trait fine mapping of pleiotropic risk loci. Bioinformatics 2017;33(2):248–255.
6. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. Nature Methods 2015;12(10):931–934.

7. Khan A, Fornes O, Stigliani A, et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res 2018;46(D1):D260–D266.

8. Rao SSP, Huntley MH, Durand NC, et al. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell 2014;159(7):1665–1680.

9. Wang Y, Song F, Zhang B, et al. The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. Genome Biology 2018;19(1):151.

10. Schofield EC, Carver T, Achuthan P, et al. CHiCP: a web-based tool for the integrative and interactive visualization of promoter capture Hi-C datasets. Bioinformatics 2016;32(16):2511–2513.

11. Mifsud B, Tavares-Cadete F, Young AN, et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nature Genetics 2015;47(6):598–606.

12. Cairns J, Freire-Pritchett P, Wingett SW, et al. CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. Genome Biol 2016;17(1):127.

13. Lappalainen T, Sammeth M, Friedländer MR, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature 2013;501(7468):506–511.

14. Piasecka B, Duffy D, Urrutia A, et al. Distinctive roles of age, sex, and genetics in shaping transcriptional variation of human immune responses to microbial challenges. PNAS 2018;115(3):E488–E497.

15. GTEx Consortium. Genetic effects on gene expression across human tissues. Nature 2017;550(7675):204–213.

16. Alpern D, Gardeux V, Russeil J, et al. BRB-seq: ultra-affordable high-throughput transcriptomics enabled by bulk RNA barcoding and sequencing. Genome Biology 2019;20(1):71.

17. Castel SE, Aguet F, Mohammadi P, Consortium Gte, Ardlie KG, Lappalainen T. A vast resource of allelic expression data spanning human tissues. bioRxiv 2019;792911.

18. Aguet F, Barbeira AN, Bonazzola R, et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. bioRxiv 2019;787903.

19. Castel SE, Mohammadi P, Chung WK, Shen Y, Lappalainen T. Rare variant phasing and haplotypic expression from RNA sequencing with phASER. Nature Communications 2016;7(1):1–6.

20. Mohammadi P, Castel SE, Brown AA, Lappalainen T. Quantifying the regulatory effect size of cis-acting genetic variation using allelic fold change. Genome Res [Epub ahead of print].

21. Berndt SI, Skibola CF, Joseph V, et al. Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia. Nature Genetics 2013;45(8):868–876.

22. Cerhan JR, Berndt SI, Vijai J, et al. Genome-wide association study identifies multiple susceptibility loci for diffuse large B cell lymphoma. Nature Genetics 2014;46(11):1233–1238.

23. Skibola CF, Berndt SI, Vijai J, et al. Genome-wide Association Study Identifies Five Susceptibility Loci for Follicular Lymphoma outside the HLA Region. The American Journal of Human Genetics 2014;95(4):462–471.

24. Vijai J, Wang Z, Berndt SI, et al. A genome-wide association study of marginal zone lymphoma shows association to the HLA region. Nat Commun;6.

Supplemental Tables

Supplemental Table 1. Cohort level association statistics for genome-wide significant variants in the meta-analysis

SNP	ANRS BETA	ANRS P	SHCS BETA	SHCS P	MACS BETA	MACS P
rs7919208	0.32	2.78E-10	0.12	2.66E-03	-0.06	0.57
rs149399290	0.28	7.93E-07	0.13	2.30E-03	-0.03	0.79
rs17155463	0.28	7.93E-07	0.13	2.30E-03	-0.03	0.79
rs17155474	0.28	7.93E-07	0.13	2.30E-03	-0.03	0.79
rs17155478	0.28	7.93E-07	0.13	2.30E-03	-0.03	0.79
rs12249837	0.28	7.93E-07	0.13	2.30E-03	-0.03	0.79
rs10608969	0.28	7.93E-07	0.13	2.30E-03	-0.03	0.79

The analyses were performed using linear mixed models with GCTA within each cohort. No associations were seen in the MACS cohort. P-values and beta values are presented to show the level and direction of the association.

Supplemental Table 2. Top predicted changes associated with rs7919208 allelic variation in GM12878

Transcription factor	Effect (Log2fold change)	E-value	Normalized Prob. (Reference)	Normalized Prob. (Alternative)
BATF	3.27	0.00004	0.13	0.60
JUND	2.91	0.00009	0.12	0.50
MEF2A	2.06	0.00026	0.09	0.28
MEF2C	1.98	0.00022	0.08	0.26
BCL11A	2.19	0.00040	0.07	0.25
P300	1.74	0.00055	0.08	0.22
IRF4	2.12	0.00033	0.06	0.21

Significant changes induced by rs7919208 as predicted by DeepSEA for transcription factors with a normalized probability (Prob.) above 0.20 in the GM12878 lymphoblastoid cell line. The E-value is the expected proportion of variants with a larger predicted effect between the reference and alternative allele for a certain chromatin feature based on predicted effects calculated for variants in The 1000 Genomes Project.

Supplemental Table 3. Comparisons with genome-wide significant variants identified in GWAS of NHL in the general population

SNP	Gene	Publication	Subtype	META P	SHCS P	ANRS P	MACS P	OR	Mean MAF	POWER (P<0.05)	POP
rs116446171	EXOC2	Cerhan et al.[14]	DLBCL	NA	NA	NA	0.08	2.20	0.01	58%	EUR
rs12195582	HLA region	Skibola et al.[22]	FL	0.33	0.47	0.51	NA	1.78	0.41	100%	EUR
rs12289961	LPXN	Vijai et al. (2013)[21]	DLBCL+FL	0.88	0.40	0.14	0.46	1.29	0.22	74%	EUR
rs13254990	PVT1	Skibola et al.[22]	FL	0.75	NA	0.58	0.82	1.18	0.32	48%	EUR
rs13255292	PVT1	Cerhan et al.[14]	DLBCL	0.83	NA	0.72	0.90	1.22	0.32	62%	EUR
rs17203612	HLA class II	Skibola et al.[22]	FL	NA	0.82	NA	NA	1.44	0.38	98%	EUR
rs17749561	BCL2	Skibola et al.[22]	FL	0.74	NA	0.72	0.98	1.34	0.10	61%	EUR
rs2523607	HLA-B	Cerhan et al.[14]	DLBCL	0.50	0.30	0.77	NA	1.32	0.08	49%	EUR
rs2922994	HLA-B	Vijai et al. (2015)[23]	MZL	0.55	0.30	0.66	NA	1.64	0.08	93%	EUR
rs3130437	HLA class I	Skibola et al.[22]	FL	0.33	0.47	NA	0.46	1.23	0.38	68%	EUR
rs4733601	PVT1	Cerhan et al.[14]	DLBCL	0.29	NA	0.38	0.57	1.18	0.49	51%	EUR
rs4937362	ETS1	Skibola et al.[22]	FL	0.62	0.61	0.90	NA	1.17	0.45	47%	EUR
rs4938573	CXCR5	Skibola et al.[22]	FL	0.60	0.25	0.72	0.18	1.34	0.19	81%	EUR
rs6444305	LPP	Skibola et al.[22]	FL	NA	NA	NA	NA	1.21	NA	NA	EUR
rs6457327	HLA	Lim et al.[53]	DLBCL+FL	0.66	0.80	0.85	0.18	1.30	0.35	85%	EUR
rs6773854	BCL6	Tan et al.[22]	DLBCL	0.30	0.51	0.73	0.07	1.44	0.21	95%	CHN
rs79480871	NCOA1	Cerhan et al.[14]	DLBCL	NA	NA	NA	NA	1.34	0.08	54%	EUR
rs9461741	BTNL2	Vijai et al. (2015)[23]	MZL	0.97	0.72	0.67	NA	2.66	0.03	99%	EUR

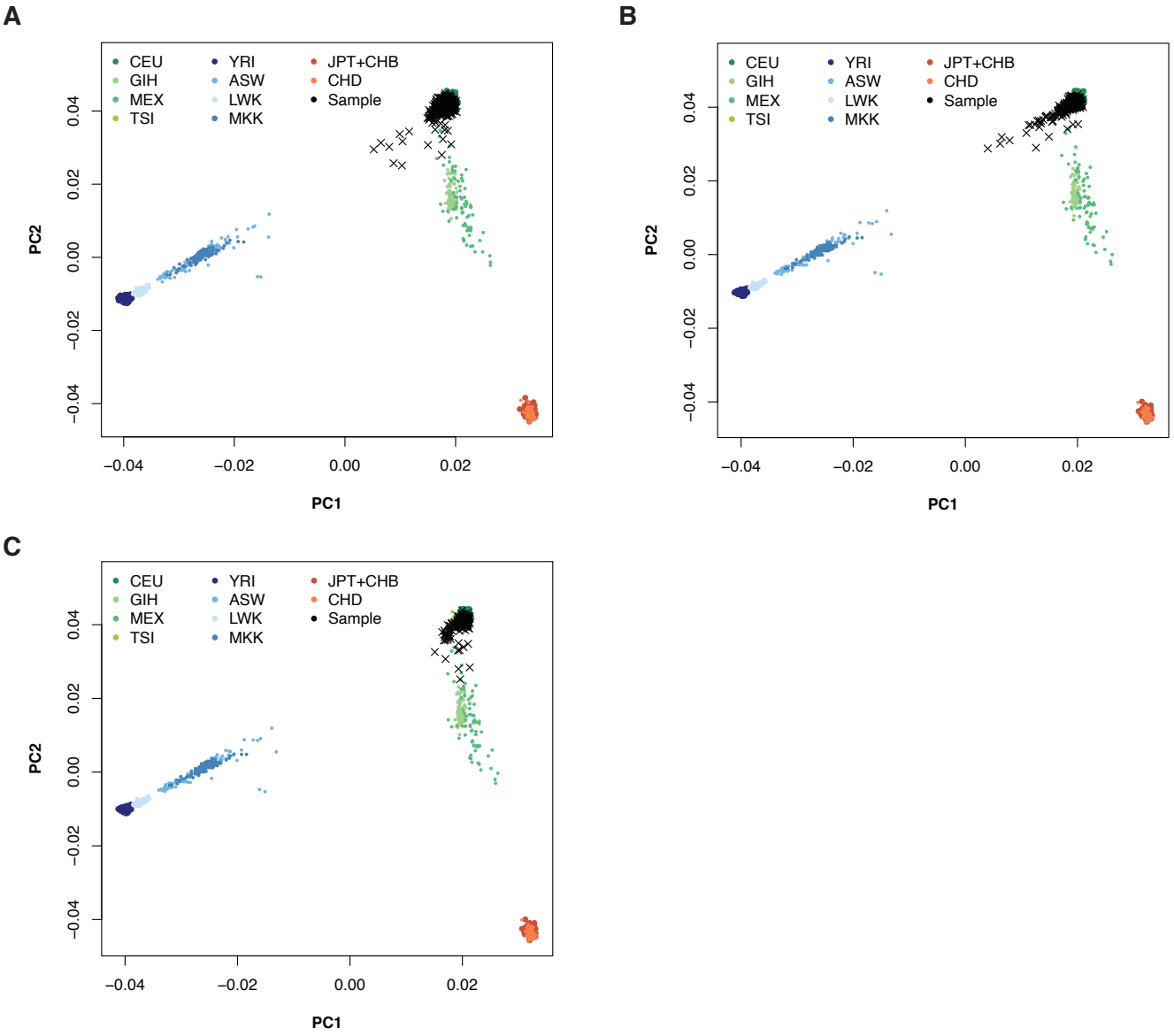
Comparisons with genome-wide significant variants identified in published GWAS of NHL in the general population. The NHL subtypes includes DLBCL (Diffuse large B-cell lymphoma), FL (Follicular lymphoma) and MZL (Marginal zone lymphoma). P-values for the HIV meta-analysis and the individual cohort GWAS are shown per variant. The statistical power to replicate the published variants under an additive model at $P < 0.05$, given their published odds ratios (OR) and the mean observed minor allele frequencies (MAF) in the HIV cohorts is also listed. Most of the published GWAS was on European (EUR) patients and Chinese (CHN).

Supplemental Table 4. The effect of rs7919208 in GWAS in the general population

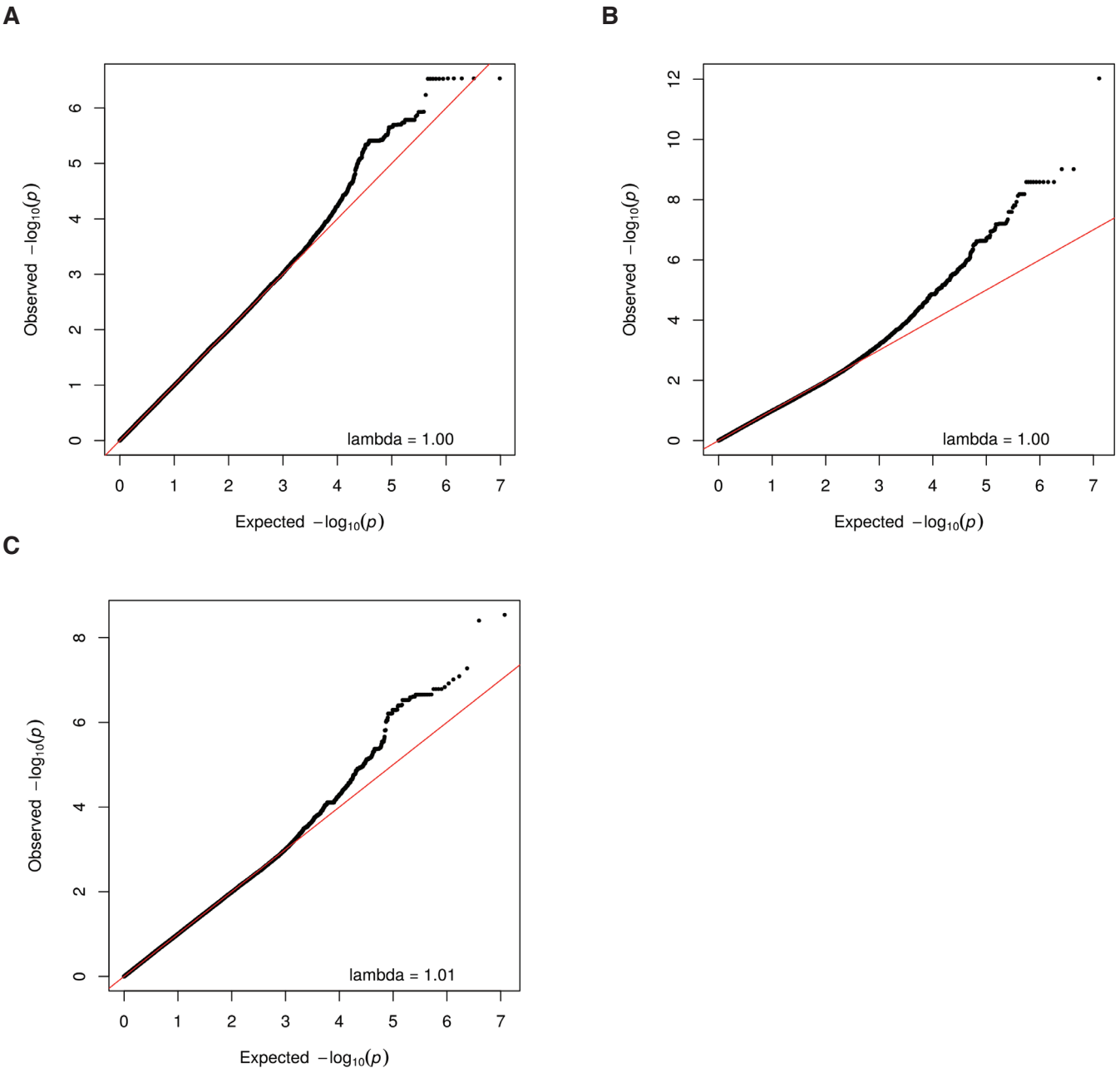
Subtype	Cases	Controls	Lambda	P (rs7919208)	OR (rs7919208)
CLL	1033	2635	1.04	0.29	0.97
DLBCL	2173	2635	1.04	0.65	1.01
FL	1753	2635	1.03	0.36	0.97
MZL	617	2635	1.01	0.56	0.98
Combined	5556	2635	1.04	0.60	0.99

The association of rs7919208 in the general (non-HIV) population across NHL subtypes. Lambda indicates the genome-wide inflation factor for the GWAS performed for each subtype to ensure the test-statistics observed are valid. The calculated p-values and odds ratios for rs7919208 are listed for each GWAS.

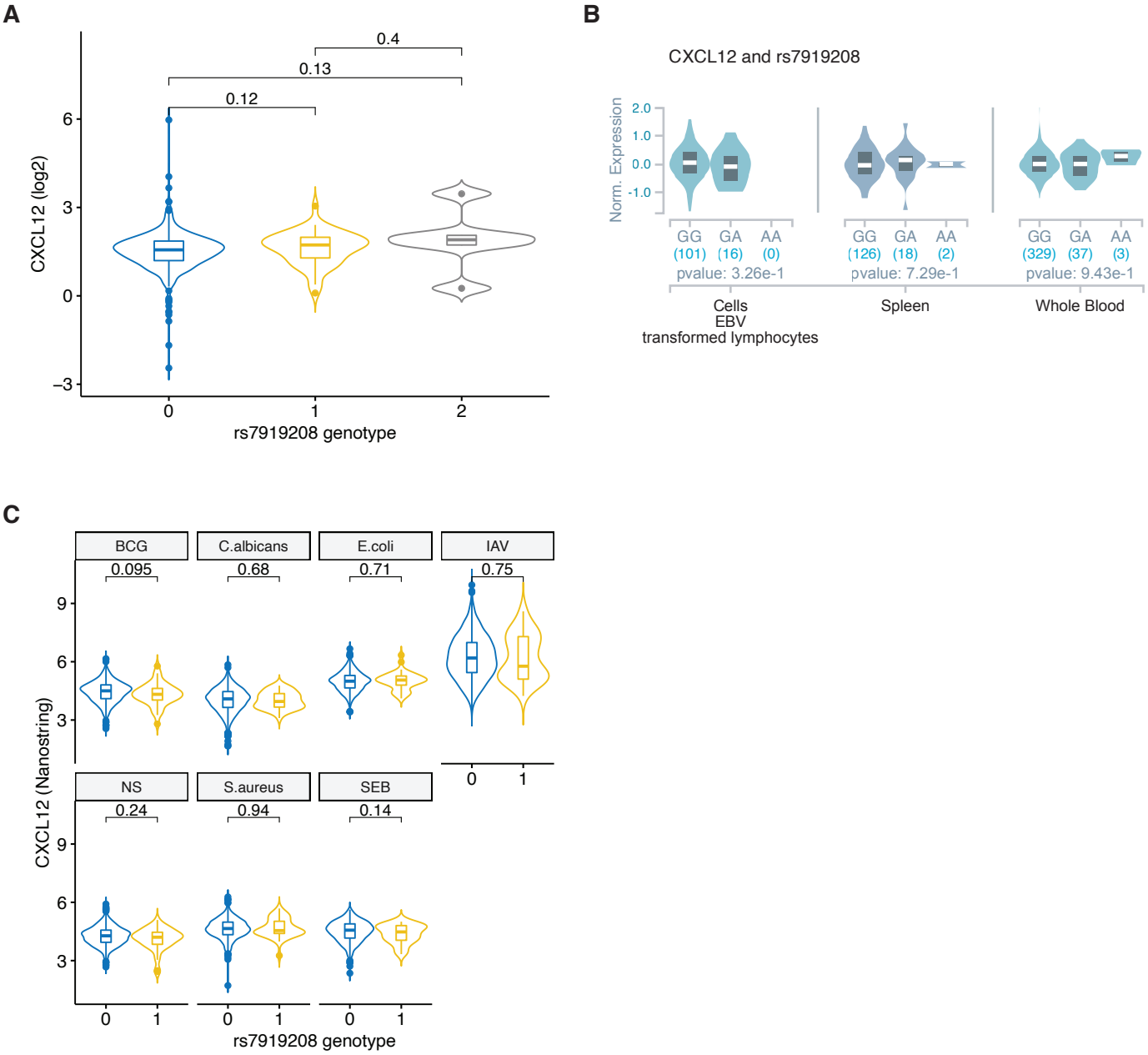
Supplemental Figures



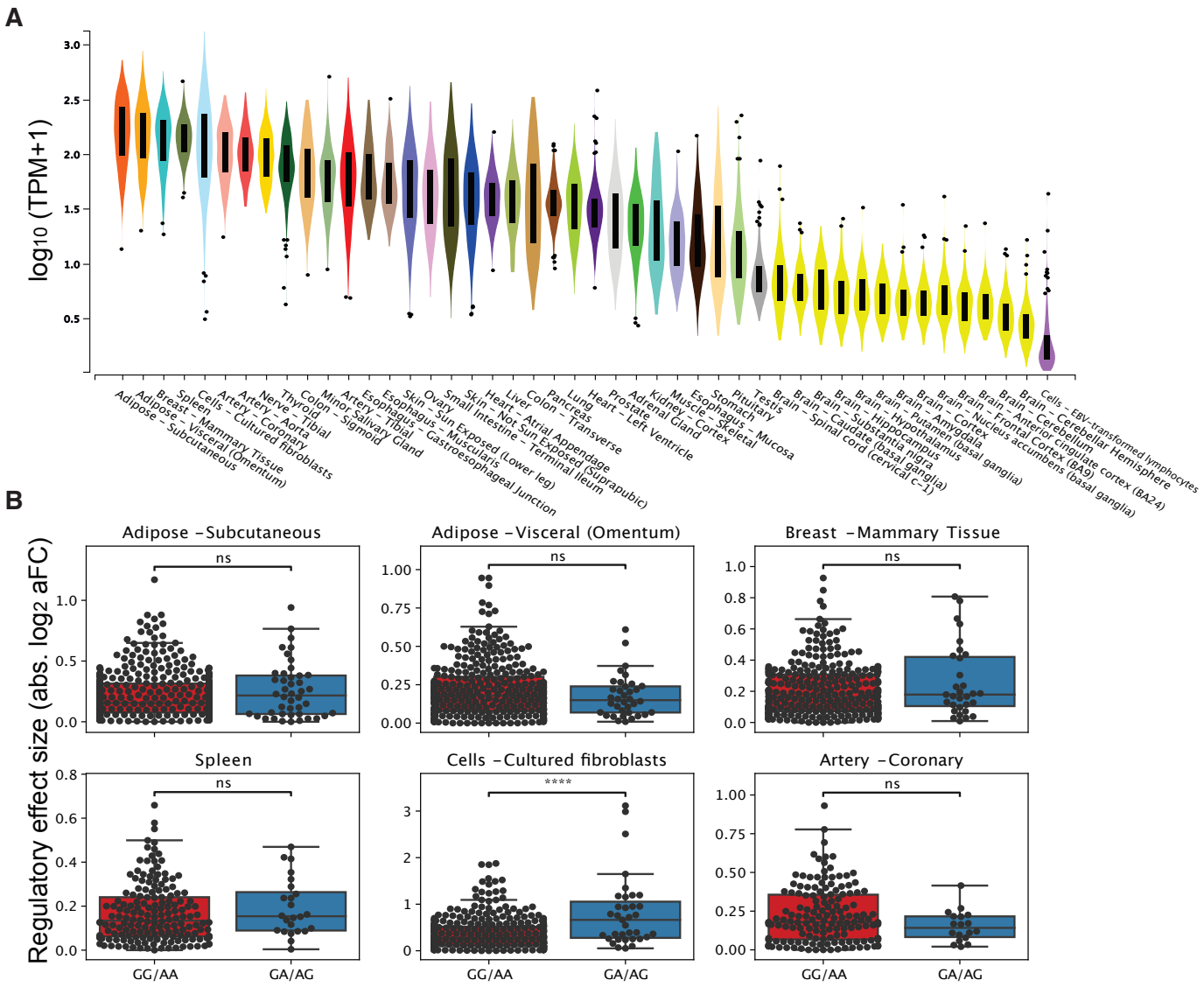
Supplemental Figure 1. Principal component analyses (PCA) with the HapMap project. The black crosses represent individuals genotyped and included in this study. Individuals of European ancestry colocalizes with the HapMap reference samples from CEU (Northern Europeans from Utah) and TSI (Tuscans from Italy). (A) The SHCS cohort. (B) The ANRS cohort. (C) The MACS cohort.



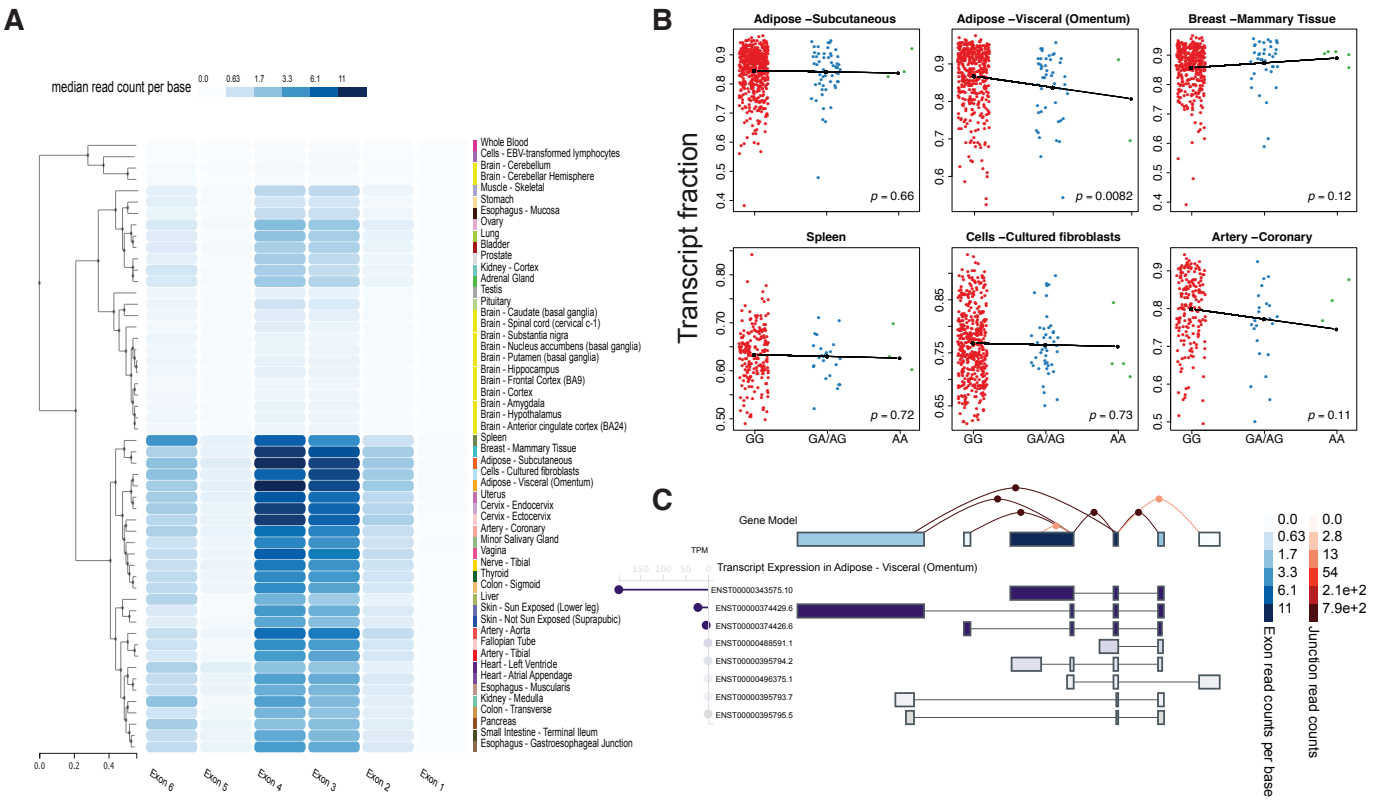
Supplemental Figure 2. Quantile-quantile plots for the initial cohort level GWAS. Lambda indicates the genome-wide inflation factor. Values ~1 denotes the lack of genomic inflation due to confounding factors. (A) Plot for the GWAS in the SHCS cohort. (B) Plot or the GWAS in the ANRS cohort. (C) Plot for the GWAS in the ANRS cohort.



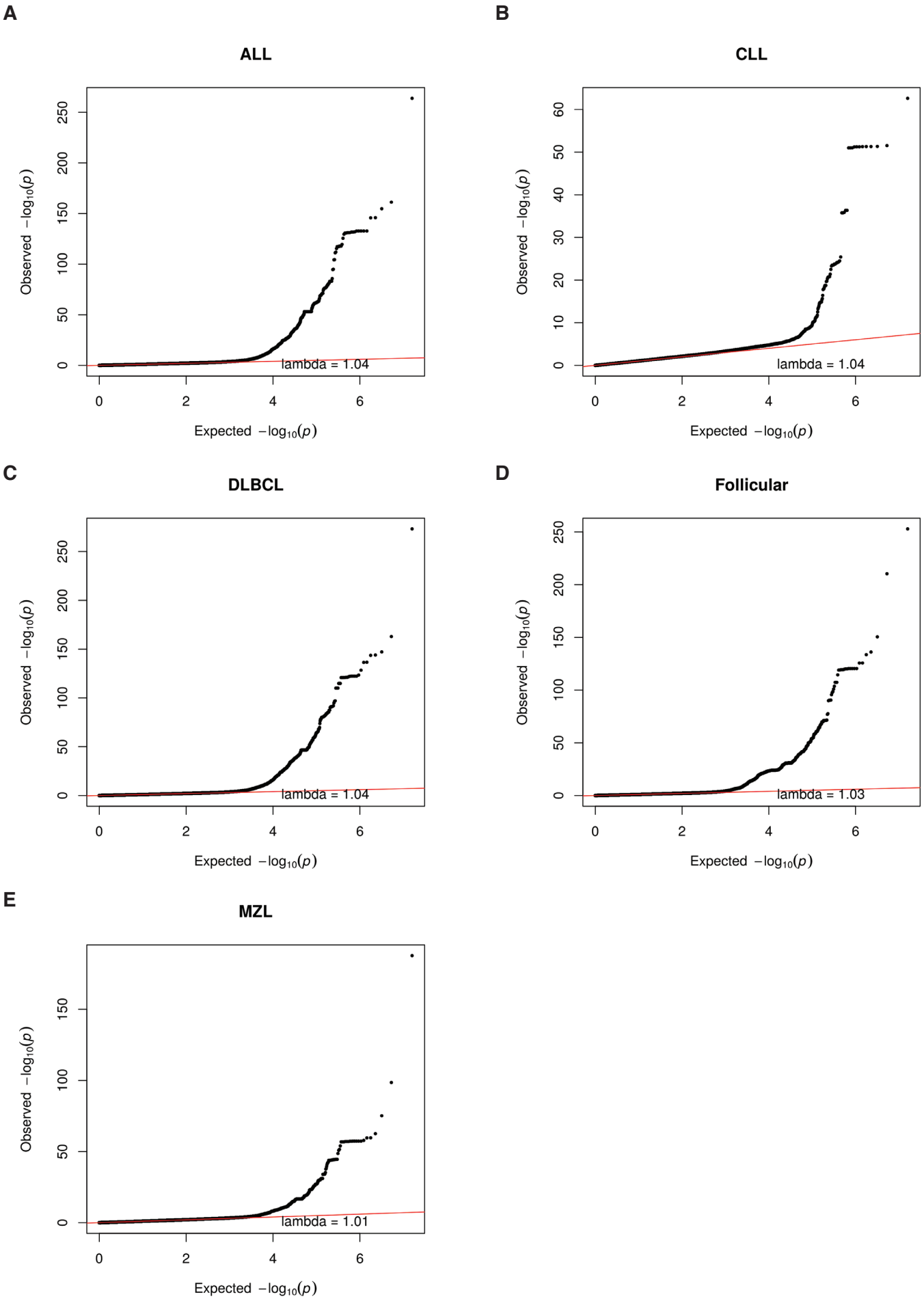
Supplemental Figure 3. eQTL information on rs7919208. (A) Expression levels of CXCL12 in EBV transformed lymphocytes from the GEUVADIS consortium according to the rs7919208 genotype. Differences between genotype groups were tested using Wilcoxon rank-sum tests with the obtained p-values shown on the figures. (B) Relationship between CXCL12 and the rs7919208 genotype in GTEx across EBV transformed lymphocytes, Spleen and Whole Blood. (C) Relationship between CXCL12 expression and rs1919208 using Nanostring in the Milieu Interieur Consortium for stimulated and non-stimulated (NS) PBMCs. Stimulants used were Mycobacterium bovis (BCG), Candida albicans, Escherichia coli, Influenza A virus (IAV), Staphylococcus aureus and Staphylococcal enterotoxin B (SEB).



Supplemental Figure 4. **Allele-specific effect in *CXCL12* associated with rs7919208.** A) Gene expression for *CXCL12* across tissues available in GTEx project data v8 release (GTEx Consortium, 2019). Gene expression values are shown in TPM (Transcript Per Million), calculated from a gene model with isoforms collapsed to a single gene (https://gtexportal.org/home/gene/CXCL12). B) Allelic effect of rs7919208 SNP on *CXCL12* expression in top six tissues with highest gene expression. Individuals heterozygous for rs7919208 show increased allelic imbalance *CXCL12* in fibroblasts (FDR adjusted p-value 0.0006, one-sided ranksum test). Each dot is an individual in GTEx v8 data. The increased allelic imbalance was not observed in any other GTEx tissues (indicated as 'ns'). The regulatory effect size (y-axis) is the absolute log allelic Fold Change (aFC) that is the ratio between the allelic expression of the two haplotypes (Mohammadi et al. 2017).



Supplemental Figure 5. Transcript isoform expression for CXCL12 and transcript usage effect associated with rs7919208. A) Exon expression of CXCL12 across tissues in GTEx v8 release data (GTEx Consortium, 2019), (https://gtexportal.org/home/gene/CXCL12). B) Effect of rs7919208 SNP on transcript isoform expression of CXCL12 in top six tissues that CXCL12 has highest expression. The expression is restricted to ENST00000343575.10 and ENST00000374429.6 only, and the y-axis shows the transcription fraction from ENST00000343575.10. The variant shows significant effect on usage of ENST00000343575.10 and ENST00000374429.6 transcript isoforms in adipose visceral tissue (linear regression p-value, 0.0082 and FDR adjusted p-value, 0.0492). We found similar results repeating this analysis with rank based inverse normal transform values that eliminate potential outlier-driven effects. (Adipose visceral linear regression p-value, 0.0063 and FDR adjusted p-value, 0.0378). Each dot is an individual in GTEx v8 data. The black line is the linear fit between transcript usage of ENST00000343575.10 (y-axis) and individual genotype groups (x-axis). The linear regression p-values are shown for each tissue. C) Representation of transcript expression model for adipose visceral tissue with eight transcript isoforms (GTEx Consortium, 2019), (https://gtexportal.org/home/gene/CXCL12). On average about 94% of expressed transcripts are derived from transcript isoforms ENST00000343575.10 and ENST00000374429.6 in all tissues. Transcript structure and overall isoform usage patterns are similar across all 6 tested tissues.



Supplemental Figure 6. Quantile-quantile plots for the general population NHL GWAS. Lambda indicates the genome-wide inflation factor. Values ~1 denotes the lack of genomic inflation due to confounding factors. (A) Plot for the GWAS of all NHL subtypes combined. (B) Plot for GWAS of chronic lymphocytic leukemia (CLL). (C) Plot for Diffuse large B-cell lymphoma (DLBCL). (D) Plot for Follicular lymphoma. (E) Plot for Marginal zone lymphoma (MZL).