

1 **Genetic variation near CXCL12 is associated with susceptibility** 2 **to HIV-related non-Hodgkin lymphoma**

3 Christian W. Thorball^{1,2}, Tiphaine Oudot-Mellakh³, Christian Hammer^{4,5}, Federico A.
4 Santoni⁶, Jonathan Niay³, Dominique Costagliola⁷, Cécile Goujard^{8,9}, Laurence Meyer¹⁰,
5 Sophia S. Wang¹¹, Shehnaz K. Hussain¹², Ioannis Theodorou³, Matthias Cavassini¹³, Andri
6 Rauch¹⁴, Manuel Battegay¹⁵, Matthias Hoffmann¹⁶, Patrick Schmid¹⁷, Enos Bernasconi¹⁸,
7 Huldrych F. Günthard^{19,20}, Paul J. McLaren^{21,22}, Charles S. Rabkin²³, Caroline Besson²³⁻
8 ^{25,*}, Jacques Fellay^{1,2,26,*}

9

10 ¹School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne,
11 Switzerland; ²Swiss Institute of Bioinformatics, Lausanne, Switzerland; ³Centre de
12 génétique moléculaire et chromosomique, GH La Pitié Salpêtrière, Paris, France;
13 ⁴Department of Cancer Immunology, Genentech, South San Francisco, CA, USA;
14 ⁵Department of Human Genetics, Genentech, South San Francisco, CA, USA; ⁶Service of
15 Endocrinology, Diabetology and Metabolism, Lausanne University Hospital, Lausanne,
16 Switzerland; ⁷Sorbonne Universités, INSERM, UPMC Université Paris 06, Institut Pierre
17 Louis d'épidémiologie et de Santé Publique (IPLESP UMRS 1136), Paris, France; ⁸Inserm,
18 CESP, U1018, Paris-Sud University, Le Kremlin-Bicêtre, France; ⁹Department of Internal
19 Medicine, Bicêtre Hospital, AP-HP, Le Kremlin-Bicêtre, France; ¹⁰INSERM U1018,
20 Centre de recherche en Épidémiologie et Santé des Populations, Paris-Sud University, Paris-
21 Saclay University, Le Kremlin-Bicêtre, France; ¹¹Division of Health Analytics, City of

22 Hope Beckman Research Institute and City of Hope Comprehensive Cancer Center;
23 ¹²Department of Medicine, Cedars-Sinai Medical Center, Los Angeles, CA, USA;
24 ¹³Service of Infectious Diseases, Lausanne University Hospital and University of
25 Lausanne, 1015, Lausanne, Switzerland; ¹⁴Department of Infectious Diseases, Bern
26 University Hospital, University of Bern, Switzerland; ¹⁵Department of Infectious Diseases
27 and Hospital Epidemiology, University Hospital Basel, University of Basel, 4031, Basel,
28 Switzerland; ¹⁶Division of Infectious Diseases and Hospital Epidemiology, Kantonsspital
29 Olten, Switzerland; ¹⁷Division of Infectious Diseases, Cantonal Hospital of St. Gallen,
30 9007, St. Gallen, Switzerland; ¹⁸Division of Infectious Diseases, Regional Hospital, 6900,
31 Lugano, Switzerland; ¹⁹Department of Infectious Diseases and Hospital Epidemiology,
32 University Hospital Zurich, 8091, Zurich, Switzerland; ²⁰Institute of Medical Virology,
33 University of Zurich, 8057, Zurich, Switzerland; ²¹JC Wilt Infectious Diseases Research
34 Centre, National Microbiology Laboratory, Public Health Agency of Canada, Winnipeg,
35 Canada; ²²Department of Medical Microbiology and Infectious Diseases, University of
36 Manitoba, Winnipeg, Canada; ²³Infections and Immunoepidemiology Branch, Division of
37 Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA;
38 ²⁴CESP, UVSQ, INSERM, Université Paris-Saclay, 94805, Villejuif, France; ²⁵Department
39 of Hematology and Oncology, Hospital of Versailles, 78150 Le Chesnay, France;
40 ²⁶Precision Medicine Unit, Lausanne University Hospital (CHUV) and University of
41 Lausanne, Lausanne, Switzerland.

42

43 * J.F. and C.B. jointly directed this work.

44 **Corresponding author:**

45 Jacques Fellay, School of Life Sciences, École Polytechnique Fédérale de Lausanne, 1015,

46 Lausanne, Switzerland. E-mail: jacques.fellay@epfl.ch. Phone number: +41216931849

47

48

49

50

51

52

53

54

55

56

57

58

59

60

61 **Abstract**

62 Human immunodeficiency virus (HIV) infection is associated with a substantially
63 increased risk of non-Hodgkin lymphoma (NHL). High plasma viral load, low CD4+ T cell
64 counts and absence of antiretroviral treatment (ART) are known predictive factors for
65 NHL. Even in the era of suppressive ART, HIV-infected individuals remain at increased
66 risk of developing NHL compared to the general population. To search for human genetic
67 determinants of HIV-associated NHL, we performed case-control genome-wide
68 association studies (GWAS) in three cohorts of HIV+ patients of European ancestry and
69 meta-analyzed the results. In total, 278 cases and 1924 matched controls were included.
70 We observed a significant association with NHL susceptibility in the C-X-C motif
71 chemokine ligand 12 (*CXCL12*) region on chromosome 10. A fine mapping analysis
72 identified rs7919208 as the most likely causal variant ($P = 4.77e-11$). The G>A
73 polymorphism creates a new transcription factor binding site for BATF and JUND.
74 Analyses of topologically associating domains and promoter capture Hi-C data revealed
75 significant interactions between the rs7919208 region and the promoter of *CXCL12*, also
76 known as stromal-derived factor 1 (*SDF-1*). These results suggest a modulatory role of
77 *CXCL12* regulation in the increased susceptibility to NHL observed in the HIV-infected
78 population.

79

80

81

82

83 **Introduction**

84 Human immunodeficiency virus (HIV) infection is associated with a markedly increased
85 risk of several types of cancer compared to the general population.¹⁻³ This elevated cancer
86 risk can be attributed partly to viral-induced immunodeficiency, frequent co-infections
87 with oncogenic viruses (e.g., Epstein-Barr virus (EBV), hepatitis B and hepatitis C viruses,
88 human herpesvirus 8 (HHV-8) and papillomavirus), and increased prevalence of traditional
89 risk factors such as smoking.^{4,5} However, all of these risk factors may not entirely explain
90 the excess cancer burden seen in the HIV+ population.⁶

91 A previous study performed in the Swiss HIV Cohort Study (SHCS) identified two AIDS-
92 defining cancers, Kaposi sarcoma and non-Hodgkin lymphoma (NHL) as the main types
93 of cancer found among HIV positive patients (NHL representing 34% of all identified
94 cancers).⁴ The relative risk of developing NHL in HIV patients was highly elevated
95 compared to the general population (period-standardized incidence ratio (SIR) = 76.4).⁴
96 High HIV plasma viral load, absence of antiretroviral therapy (ART) as well as low CD4+
97 T cell counts are known predictive factors for NHL.^{7,8} The introduction of ART into
98 clinical practice has led to improved overall survival and restoration of immunity by
99 decreasing viral load and increasing CD4+ T cell counts, and has led to a decreased risk of
100 developing NHL. However, the risk remains substantially elevated compared to the general
101 population (SIR = 9.1 (8.3–10.1))⁹ and NHL still represents 20% of all cancers in people
102 living with HIV in the ART era.¹⁰ Non-Hodgkin's lymphomas associated with HIV are
103 predominantly aggressive B-cell lymphomas. Although they are heterogeneous, they share
104 several pathogenic mechanisms involving chronic antigen stimulation, impaired immune

105 response, cytokine deregulation and reactivation of the oncogenic viruses EBV and HHV-
106 8.¹¹

107 The emergence of genome-wide approaches in human genomics has led to the discovery
108 of many associations between common genetic polymorphisms and susceptibility to
109 several diseases including HIV infection and multiple types of cancer.^{12,13} Recent genome-
110 wide association studies (GWAS) of NHL have identified multiple susceptibility loci in
111 the European population.¹⁴⁻²² These variants are located in the genes *LPXN*²¹, *BTNL2*²³,
112 *EXOC2*, *NCOA1*¹⁴, *PVT1*^{14,22}, *CXCR5*, *ETSI*, *LPP*, and *BCL2*²² for various subtypes of
113 NHL, as well as *BCL6* in the Chinese population.²⁴ Strong associations with variation in
114 human leukocyte antigen (HLA) genes have also been reported.^{15,18,22} However, in the
115 setting of HIV infection, no genome-wide analysis has been reported concerning the
116 occurrence of NHL and the specific mechanisms driving their development remain largely
117 unknown.

118 Here we report the results of the first genome-wide analysis of NHL susceptibility in
119 individuals chronically infected with HIV. We combined three HIV cohort studies from
120 France, Switzerland and the USA and searched for associations between >6 million single
121 nucleotide polymorphisms (SNPs) and a diagnosis of NHL. We identified a novel genetic
122 locus near *CXCL12* as associated with the development of NHL among HIV+ individuals.

123

124

125

126 **Materials and methods**

127 *Ethics statement*

128 The Swiss HIV Cohort Study (SHCS), the Primo ANRS and ANRS CO16 Lymphovir
129 cohorts (ANRS) and the Multicenter AIDS Cohort Study (MACS) cohorts have been
130 approved by the competent ethics committees / institutional review boards of all
131 participating institutions. A written informed consent, including consent for human genetic
132 testing, was obtained from all study participants.

133

134 *Study participants and contributing centers*

135 **Swiss HIV Cohort Study (SHCS)**

136 The SHCS is a large, ongoing, multicenter cohort study of HIV-positive individuals that
137 includes >70% of adult living with HIV in Switzerland. At follow-up visits every 6 months,
138 demographic, clinical, laboratory, and ART information has been prospectively recorded
139 since 1988.²⁵ Cancer diagnoses are verified thoroughly using checking charts including
140 information on biopsies and imaging. To minimize potential treatment bias and population
141 stratification, we only considered as cases patients diagnosed with NHL between 2000 and
142 2017 and of European ancestry, as determined by principal component analysis (PCA)
143 (supplemental Figure 1A). Controls were matched based on age, ancestry, CD4+ T cell
144 counts and viral load results. To be eligible as controls, they also had to be diagnosed with
145 HIV prior to 2005 and have no registered cancer diagnosis of any type as of 2017. Patients
146 were genotyped using Illumina HumanOmniExpress-24 Beadchips, or genotypes were

147 obtained in the context of a previous GWAS in the SHCS on various platforms including
148 Illumina HumanCore-12, HumanHap550, Human610 and Human1M Beadchips.

149

150 **French Primo ANRS and ANRS CO16 Lymphovir cohorts (ANRS)**

151 The French ANRS CO16 lymphovir cohort of HIV related lymphomas enrolled adult
152 patients at diagnosis of lymphoma in 32 centers between 2008 and 2015.²⁶ Pathological
153 materials were centralized, and diagnoses of NHL were based on World Health
154 Organization criteria. Patients were genotyped using Illumina Human Omni5 Exome 4v
155 beadchips. Additional cases and controls were included from the ANRS PRIMO Cohort,
156 which has been enrolling patients during primary HIV-1 infection in 95 French Hospitals
157 since 1996.²⁷ Patients were genotyped using Illumina Sentrix Human Hap300 Beadchips.
158 Only patients of European ancestry, as determined by PCA, were included in the study
159 (supplemental Figure 1B).

160

161 **The Multicenter AIDS Cohort Study (MACS)**

162 The MACS has enrolled gay and bisexual HIV infected men in 4 US cities since 1984. The
163 NHL cases were predominately diagnosed prior to the year 2000. Data collected include
164 demographic variables (age, race, ethnicity and HIV transmission category), CD4+ T cell
165 count, HIV viral load and tumor histology. Eligible cases had a diagnosis of HIV-related
166 NHL, available genotyping data and at least one CD4+ T cell count obtained within 2 years
167 of the NHL diagnosis. Controls were matched on MACS study site, age at NHL diagnosis
168 (+/- 2 years) and CD4+ T cell count at NHL diagnosis (within the following groups 0-99 /

169 100 -199 / 200-499 / >499 cells/ μ L). Patients were genotyped using Illumina
170 HumanHap550 and Human1M Beadchips.²⁸ As in the other cohorts, only individuals of
171 European ancestry were included, as determined by PCA (supplemental Figure 1C).

172

173 ***Quality control and imputation of genotyping data***

174 The genotyping data from each cohort was filtered and imputed in a similar way, with each
175 genotyping array processed separately to minimize potential batch effects. All variants
176 were first flipped to the correct strand orientation with BCFTOOLS (v1.8) using the human
177 genome build GRCh37 as reference. Variants were removed if they had a larger than 20%
178 minor allele frequency (MAF) deviation from the 1000 genomes phase 3 EUR reference
179 panel or if they showed a larger than 10% MAF deviation between genotyping chips in the
180 same cohort.

181 The QC filtered genotypes were phased with EAGLE2²⁹ and missing genotypes were
182 imputed using PBWT³⁰ with the Sanger Imputation Service³¹, taking the 1000 Genomes
183 Project Phase 3 panel as reference. Only high-quality variants with an imputation score
184 (INFO > 0.8) were retained for further analyses.

185

186 ***Genome-wide association testing and meta-analysis***

187 To search for associations between human genomic variation and the development of HIV-
188 related NHL, we first performed separate GWAS within each cohort (SHCS, ANRS and
189 MACS) prior to combining the results in a meta-analysis.

190 For each cohort separately, the imputed variants were filtered out using PLINK
191 (v2.00a2LM)³² based on missingness (> 0.1), minor allele frequency (< 0.02) and deviation
192 from Hardy-Weinberg Equilibrium ($P_{HWE} < 1e-6$). Determination of population structure
193 and calculation of principal components was done using EIGENSTRAT (v6.1.4)³³ and the
194 HapMap3 reference panel³⁴. All individuals not clustering with the European HapMap3
195 samples were excluded from further analyses. The samples were screened using KING
196 (v2.1.3)³⁵ to ensure no duplicate or cryptic related samples were included. Single-marker
197 case-control association analyses were performed using linear mixed models, with genetic
198 relationship matrices calculated between pairs of individuals according to the leave-one-
199 chromosome-out principle, as implemented in GCTA mlma-loco (v1.91.4beta).^{36,37} Sex
200 was included as a covariate, except in the MACS cohort, which only includes men.

201 The results of the three GWAS were combined across cohorts using a weighted Z-score-
202 based meta-analysis in PLINK (v1.90b5.4), after exclusion of the variants that were not
203 present in all three cohorts.

204

205 *Fine mapping of associated regions*

206 Fine mapping of the *CXCL12* locus was performed using PAINTOR (v3.1)³⁸ to identify
207 the most likely causal variant(s). All variants within 200kb of the top associated SNP and
208 with a p-value below 0.005 were included in the model. The linkage disequilibrium (LD)
209 matrix was created using PLINK and genotype data from the SHCS cohort. PAINTOR was
210 first run against all genomic annotation databases provided with the software, including the
211 FANTOM5, ENCODE and the Roadmap Epigenomics Project. For the final model, the top

212 5 annotations based on improvement to model fit and cell type relevance were selected to
213 obtain the posterior probabilities and the 99% credible set of the variants most likely to be
214 causal based on the association from Bayes' factors.

215

216 *Predictive effect of potentially causal variants*

217 The potential functional impact of the predicted causal variants was assessed using
218 DeepSEA³⁹, a deep learning-based sequence model trained on available chromatin and
219 transcription factor data from ENCODE and Roadmap Epigenomics. DeepSEA provides a
220 functional significance score for each variant, which is a measure of the evolutionary
221 conservation and the significance of the magnitude of the predicted chromatin effects. For
222 the variants with a functional significance score of less than 0.01, we analyzed the predicted
223 changes in specific chromatin modifications or transcription factor (TF) binding
224 probabilities. Chromatin or TF binding changes with E-values below 0.001 and normalized
225 probabilities of observing a binding event above 0.2 were considered relevant. The TF
226 position weight matrices (PWMs) for TFs with a high probability of binding (normalized
227 probability $\geq 50\%$) were obtained from the JASPAR CORE 5.0 database.⁴⁰

228

229 *Long-range chromatin interactions*

230 Predicted topological associating domains (TADs) near the genome-wide significant locus
231 in GM12878 lymphoblastoid cells were obtained from publicly available data⁴¹ and
232 visualized using the 3D Genome Browser.⁴²

233 Potential interactions between the genome-wide significant locus and promoters of nearby
234 genes were analyzed using publicly available promoter capture Hi-C data in GM12878
235 lymphoblastoid cells. The Hi-C data was processed through the CHiCAGO pipeline and
236 visualized with CHiCP.^{43,44} Interaction scores ≥ 5 were considered significant, as described
237 previously.⁴⁵

238

239 *Expression quantitative trait loci (eQTL) analyses*

240 The role of rs7919208 as an eQTL was examined in GEUVADIS⁴⁶ and in response to
241 various pathogens, although not including HIV, in the Milieu Intérieur Consortium
242 cohort.⁴⁷ Furthermore, eQTL information was also obtained from the GTEx (v7)⁴⁸ Portal
243 on 03/22/2019.

244 Bulk RNA Barcoding and sequencing (BRB-seq)⁴⁹ was performed on RNA from
245 peripheral blood mononuclear cells (PBMCs) of 452 individuals from the SHCS with
246 available genotyping data.

247 *Comparison to GWAS hits in the general population*

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

253 The effect of rs7919208 in the general population cohorts was assessed directly using the
254 NIH database for Genotypes and Phenotypes (dbGaP) accession # phs000801 cohorts for
255 chronic lymphocytic leukemia (CLL), DLBCL (Diffuse large B-cell lymphoma), FL
256 (Follicular lymphoma) and MZL (Marginal zone lymphoma) and corresponding
257 controls.^{14,22,23,50} The genotype data was imputed, processed and analyzed using the same
258 pipeline and methods as described above for the HIV cohorts, with duplicate samples
259 identified and removed using KING and including age and sex as covariates.

260

261 *Statistical analyses*

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

264

265 *Data sharing statement*

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

269

270

271

272

273

274 **Results**

275 *Study participants and association testing*

276 To identify human genetic determinants of HIV-associated NHL, we performed case-
277 control genome-wide association studies (GWAS) in three groups of HIV+ patients of
278 European ancestry (SHCS, ANRS and MACS). The characteristics of the study participants
279 are presented in Table 1. In total, genotyping data were obtained for 278 cases
280 (NHL+/HIV+) and 1924 matched controls (NHL-/HIV+). With this sample size, we had
281 80% power to detect a common genetic variant (10% minor allele frequency) with a relative
282 risk of 2.5, assuming an additive genetic model and using Bonferroni correction for
283 multiple testing ($P_{\text{threshold}} = 5e-8$).⁵¹

284 After genome-wide imputation and quality control, 6.2 million common variants were
285 tested for association with the development of NHL using linear mixed models including
286 sex as a covariate. Results were combined across cohorts using a weighted Z-score-based
287 meta-analysis (Figure 1A). The genomic inflation factor (λ) was in all cases very
288 close to 1 [1.00–1.01], indicating an absence of systematic inflation of the association
289 results (Figure 1B; supplemental Figure 2).

290

291 *Association results*

292 We observed significant associations with the development of HIV-related NHL at a single
293 locus on chromosome 10, downstream of *CXCL12* (Figure 1C). A total of 7 SNPs in this
294 locus had p-values lower than the genome-wide significance threshold ($P < 5e-8$), with

295 rs7919208 displaying the strongest association (Table 2). This association was only
296 detected in the SHCS and ANRS cohorts and not among MACS study participants
297 (supplemental Table 1).

298

299 *Fine mapping of the CXCL12 locus*

300 To identify the causal variant(s) among associated SNPs and determine their potential
301 functional effects, we used a multi-level fine mapping approach, combining the statistical
302 fine mapping tool PAINTOR to obtain a 99% credible set and the deep learning framework
303 DeepSEA to predict any effects on chromatin marks and transcription factor binding these
304 variants may have.

305 Using PAINTOR, we identified a single variant, rs7919208, having a high posterior
306 probability (= 100%) of being causal among the 99% credible set based on the integration
307 of the association results, LD structure and enrichment of genomic features in this locus
308 (Figure 2).

309 Consistent with the PAINTOR result, DeepSEA also identified rs7919208 as the sole
310 variant, among the 99% credible set, predicted to have a functional impact by significantly
311 increasing the probability of binding by the B cell transcription factors BATF (log₂ fold-
312 change = 3.27) and JUND (log₂ fold-change = 2.91) (supplemental Table 2). Further
313 analysis of the genomic sequence surrounding rs7919208 and the JASPAR transcription
314 factor binding site (TFBS) motifs for BATF and JUND revealed that rs7919208 G->A
315 polymorphism creates the TFBS motif required for the binding of these transcription
316 factors (Figure 3A).

317 ***Long-range chromatin interactions***

318 To assess the potential functional links between the TFBS created in the presence of the
319 minor allele of rs7919208 and the nearby genes, we performed an analysis of promoter
320 capture Hi-C data and topologically associating domains (TADs). We used the well-
321 characterized GM12878 lymphoblastoid cell line produced by EBV transformation of B
322 lymphocytes collected from a female European donor as model.

323 First, to examine the interaction potential of the rs7919208 region with nearby promoters,
324 we analyzed available promoter capture Hi-C data obtained from the GM12878 cell line.
325 This analysis revealed a significant interaction between the rs7919208 region and the
326 *CXCL12* promoter, suggesting a possible modulating impact of rs7919208 on the
327 transcription of that gene (Figure 3B). Second, to further validate this observed genomic
328 interaction, we analyzed available TAD calls from GM12878 cells⁴¹, using the 3D Genome
329 Browser for visualization⁴² (Figure 3C). We observed that rs7919208 is located within a
330 large TAD together with *CXCL12*, signifying the interaction potential of the new TFBS at
331 rs7919208 and *CXCL12*.

332

333 ***Transcriptomic effects of rs7919208***

334 We did not observe any association between rs7919208 and mRNA expression levels of
335 *CXCL12* in peripheral blood or PBMCs from multiple publicly available datasets,
336 including GTEx (v7)⁴⁸, GEUVADIS⁴⁶ and the Milieu Intérieur Consortium⁴⁷
337 (supplemental Figure 3). Of note, *CXCL12* expression levels were very low in all datasets.

338 HIV infection causes many profound transcriptomic changes.⁵² Thus, in order to examine
339 the effect of rs7919208 on *CXCL12* in the context of HIV infection, we extracted RNA
340 from PBMCs of 452 individuals in the SHCS with available genotyping data and sequenced
341 them using the Bulk RNA Barcoding and sequencing (BRB-seq) approach.⁴⁹ However, the
342 expression levels of *CXCL12* were below the limit of detection for most individuals,
343 preventing an eQTL analysis.

344

345 *No replication of susceptibility loci found in the general population*

346 To assess whether the genetic contribution to the risk of developing NHL is similar or
347 distinct in the HIV+ population compared to the general population, we extracted the p-
348 values of all variants found to be genome-wide significant in previous GWAS performed
349 in the general population^{14,21–24,53} and compared them to our results. We did not replicate
350 any of the previously published genome-wide associated variants, even at nominal
351 significance level ($P < 0.05$), despite sufficient statistical power for many of the variants,
352 thus indicating that the genetic susceptibility of NHL is distinct between the HIV+ and the
353 general population (supplemental Table 3). To further examine this possibility, we tested
354 whether the NHL/HIV+ associated variant rs7919208 is associated with an increased risk
355 of NHL in the general population. We performed a series of case/control GWAS of four
356 NHL subtypes (CLL, DLBCL, FL and MZL) as well as a combined GWAS with all NHL
357 subtypes (supplemental Table 4; supplemental Figure 4) and assessed the association
358 evidence at rs7919208. We found no association between rs7919208 and any of the
359 subtypes in the general population, even at nominal significance.

360 **Discussion**

361 In this genome-wide analysis, including a total of 278 NHL HIV+ cases and 1924 HIV+
362 controls from three independent cohorts, we identified a novel NHL susceptibility locus on
363 chromosome 10 near the *CXCL12* gene. The strong signal observed in the meta-analysis
364 was driven by the associations detected in the SHCS and ANRS cohorts and there was no
365 evidence of association in the MACS cohort. Notably, most NHL cases in the MACS
366 cohort date back to the pre-ART era, while only NHL cases diagnosed after the year 2000
367 were included in the SHCS and ANRS analyses. Conceivably, NHL occurring in the early
368 years of the HIV pandemic may have been primarily driven by severe immunosuppression,
369 which could have obscured any influence of human genetic variation among the cases in
370 the MACS sample. Precise phenotype definition is crucial in designing large-scale genetic
371 studies since any environmental noise tends to decrease the likelihood of identifying
372 potential genetic influences.

373 NHL is a relatively rare cancer even among HIV infected individuals, making it difficult
374 to collect the large numbers of cases that would typically be included in contemporary
375 genome-wide genetic studies. Indeed, a recent study from the Data Collection on Adverse
376 events of Anti-HIV Drugs (D:A:D) group showed an NHL incidence rate of 1.17/1000
377 person-years of follow-up over the past 15 years (392 new cases in >40,000 HIV-infected
378 individuals).⁸ Still, we were able to obtain clinical and genetic data from a total of 278
379 patients with confirmed NHL diagnosis. By matching them with a larger number of
380 controls from the same cohorts, we had enough power to identify associated variants of
381 relatively large effects in the *CXCL12* region.

382 Several groups have already suggested a potential role for *CXCL12* variation in HIV-
383 related NHLs. A prospective study correlated increased *CXCL12* expression with
384 subsequent NHL development in HIV-infected children but not in uninfected children.⁵⁴
385 The number of A alleles at the *CXCL12*-3' variant (rs1801157) has also previously been
386 associated with an increased risk of developing HIV-related NHL during an 11.7 year
387 follow-up period.⁵⁵ Thus, our data further support the role of *CXCL12* as a critical
388 modulator of the individual risk of developing NHL in the HIV population.

389 The role of *CXCL12* and its receptor chemokine receptor 4 (*CXCR4*) in cancer in the
390 general population is well established, with the levels of *CXCL12* and *CXCR4* found to be
391 increased in multiple types of cancer and to be associated with tumor progression.^{56,57}
392 Furthermore, *in vivo* inhibition of either *CXCR4* or *CXCL12* signaling is capable of
393 disrupting early lymphoma development in severe combined immunodeficient (SCID)
394 mice transfused with EBV+ PBMCs.⁵⁸ These results and others have already led to the
395 development and testing of several small molecules targeting either *CXCL12* or *CXCR4*
396 to inhibit tumor progression.⁵⁶

397 We could not identify any significant relationship between rs7919208 and the expression
398 levels of *CXCL12* in PBMCs or EBV transformed lymphocytes. This can be due to multiple
399 factors such as the low expression levels of *CXCL12* in most tissues, aside from stromal
400 cells, or that rs7919208 through creation of the *BATF* and *JUND* binding site represent an
401 induced or dynamic eQTL. These types of eQTLs are often found in regions deprived of
402 regulatory annotations, since these have been examined in static cell types.⁵⁹ HIV-induced
403 overexpression of *BATF*⁶⁰ could also explain why rs7919208 is only a risk factor in the
404 HIV population and not in the general population.

405 Previous analyses in the general population have discovered both shared and distinct
406 associations for NHL subtypes.^{14,21–24,53} However, similar analyses were not possible in
407 our sample since NHL subtype information was not available for many of our cases.
408 Furthermore, information on serostatus for relevant co-infections with EBV or other
409 oncogenic viruses was not available and could therefore not be assessed. In particular, EBV
410 has been largely associated with the development of NHL and other lymphomas and is
411 considered a driver of a subset of NHLs in the general population.⁶¹ Variants in the HLA
412 region have consistently been associated with all NHL subtypes in HIV uninfected
413 populations regardless of EBV serostatus. We did not find any evidence of HLA
414 associations in our analyses of HIV-related NHL. This lack of replication of HLA variants
415 and of all other previously identified risk variants from the general population suggests that
416 distinct genes or pathways influence susceptibility to NHL in the HIV+ population
417 compared to the general population.⁶²

418 In summary, we have identified variants significantly associated with the development of
419 NHL in the HIV population. Fine mapping of the associated locus and subsequent analyses
420 of TADs, promoter capture Hi-C data as well as deep-learning models of mutational effects
421 on transcription factor binding, points to a causative model involving the gain of a BATF
422 and JUND transcription binding site downstream of *CXCL12* capable of physically
423 interacting with the *CXCL12* promoter. These results suggest an important modulating role
424 of *CXCL12* in the development of HIV-related NHL.

425

426

427 **Acknowledgments**

428 This study has been financed within the framework of the Swiss HIV Cohort Study,
429 supported by the Swiss National Science Foundation (grant #177499), by SHCS project
430 #789 and by the SHCS research foundation. The data are gathered by the Five Swiss
431 University Hospitals, two Cantonal Hospitals, 15 affiliated hospitals and 36 private
432 physicians (listed in <http://www.shcs.ch/180-health-care-providers>).

433

434 Members of the Swiss HIV Cohort Study:

435 Anagnostopoulos A, Battegay M, Bernasconi E, Böni J, Braun DL, Bucher HC, Calmy A,
436 Cavassini M, Ciuffi A, Dollenmaier G, Egger M, Elzi L, Fehr J, Fellay J, Furrer H, Fux
437 CA, Günthard HF (President of the SHCS), Haerry D (deputy of "Positive Council"), Hasse
438 B, Hirsch HH, Hoffmann M, Hösli I, Huber M, Kahlert CR (Chairman of the Mother &
439 Child Substudy), Kaiser L, Keiser O, Klimkait T, Kouyos RD, Kovari H, Ledergerber B,
440 Martinetti G, Martinez de Tejada B, Marzolini C, Metzner KJ, Müller N, Nicca D, Paioni
441 P, Pantaleo G, Perreau M, Rauch A (Chairman of the Scientific Board), Rudin C, Scherrer
442 AU (Head of Data Centre), Schmid P, Speck R, Stöckle M (Chairman of the Clinical and
443 Laboratory Committee), Tarr P, Trkola A, Vernazza P, Wandeler G, Weber R, Yerly S.

444

445 This work further benefited from the ANRS funding of both the Primo and Lymphovir
446 cohorts.

447 Foundation Monahan and Fulbright funded the stay of CB at the National Cancer Institute
448 (NCI).

449 The Genome-Wide Association Study (GWAS) of Non-Hodgkin Lymphoma (NHL)
450 project was supported by the intramural program of the Division of Cancer Epidemiology
451 and Genetics (DCEG), National Cancer Institute (NCI), National Institutes of Health
452 (NIH). The datasets have been accessed through the NIH database for Genotypes and
453 Phenotypes (dbGaP) under accession # phs000801. A full list of acknowledgements can be
454 found in the supplementary note (Berndt SI et al., Nature Genet., 2013, PMID: 23770605).

455

456 **Authorship**

457 C.W.T., J.F., P.J.M., C.S.R., C.B., C.H. and T.O.M. contributed to the conception and
458 design of the study. C.W.T., J.F., P.J.M., F.A.S., D.C., L.M., C.G., I.T., S.K.H., M.C., A.R.,
459 M.B., M.H., P.S., E.B., H.F.G., C.S.R. and C.B. contributed to the acquisition of data.
460 C.W.T., T.O.M., C.H., F.A.S., C.B., C.S.R. and J.F. contributed to the analysis and
461 interpretation of data. C.W.T., J.F., C.S.R., C.B. and S.W. contributed to the drafting the
462 article and revising it critically for important intellectual content.

463 All authors critically reviewed and approved the final manuscript.

464 Conflict of Interest Disclosure: Christian Hammer is a full-time employee of F. Hoffmann–
465 La Roche/Genentech. The remaining authors declare no competing financial interests.

466 Correspondence: Jacques Fellay, School of Life Sciences, École Polytechnique Fédérale
467 de Lausanne, Lausanne, Switzerland; e-mail: jacques.fellay@epfl.ch.

468 References

- 469 1. Patel P, Hanson DL, Sullivan PS, et al. Incidence of Types of Cancer among HIV-
470 Infected Persons Compared with the General Population in the United States, 1992–
471 2003. *Ann Intern Med.* 2008;148(10):728–736.
- 472 2. Vogel M, Friedrich O, Lüchters G, et al. Cancer risk in HIV-infected individuals on
473 HAART is largely attributed to oncogenic infections and state of
474 immunocompetence. *European Journal of Medical Research.* 2011;16(3):101.
- 475 3. Robbins HA, Pfeiffer RM, Shiels MS, et al. Excess Cancers Among HIV-Infected
476 People in the United States. *J. Natl. Cancer Inst.* 2015;107(4):.
- 477 4. Clifford GM, Polesel J, Rickenbach M, et al. Cancer Risk in the Swiss HIV Cohort
478 Study: Associations With Immunodeficiency, Smoking, and Highly Active
479 Antiretroviral Therapy. *JNCI Journal of the National Cancer Institute.*
480 2005;97(6):425–432.
- 481 5. Engels EA. Non-AIDS-defining malignancies in HIV-infected persons: etiologic
482 puzzles, epidemiologic perils, prevention opportunities. *AIDS.* 2009;23(8):875–885.
- 483 6. Borges ÁH, Dubrow R, Silverberg MJ. Factors contributing to risk for cancer among
484 HIV-infected individuals, and evidence that earlier combination antiretroviral therapy
485 will alter this risk: *Current Opinion in HIV and AIDS.* 2014;9(1):34–40.
- 486 7. Guiguet M, Boué F, Cadranet J, et al. Effect of immunodeficiency, HIV viral load,
487 and antiretroviral therapy on the risk of individual malignancies (FHDH-ANRS
488 CO4): a prospective cohort study. *The Lancet Oncology.* 2009;10(12):1152–1159.
- 489 8. Shepherd L, Ryom L, Law M, et al. Differences in Virological and Immunological
490 Risk Factors for Non-Hodgkin and Hodgkin Lymphoma. *J Natl Cancer Inst.*
491 2018;110(6):598–607.
- 492 9. Hleyhel M, Belot A, Bouvier AM, et al. Risk of AIDS-Defining Cancers Among
493 HIV-1–Infected Patients in France Between 1992 and 2009: Results From the
494 FHDH-ANRS CO4 Cohort. *Clinical Infectious Diseases.* 2013;57(11):1638–1647.
- 495 10. Robbins HA, Pfeiffer RM, Shiels MS, et al. Excess Cancers Among HIV-Infected
496 People in the United States. *JNCI: Journal of the National Cancer Institute.*
497 2015;107(4):.
- 498 11. Swerdlow SH. WHO classification of tumours of haematopoietic and lymphoid
499 tissues. International Agency for Research on Cancer; 2017.
- 500 12. McLaren PJ, Carrington M. The impact of host genetic variation on infection with
501 HIV-1. *Nat Immunol.* 2015;16(6):577–583.
- 502 13. Sud A, Kinnersley B, Houlston RS. Genome-wide association studies of cancer:
503 current insights and future perspectives. *Nature Reviews Cancer.* 2017;17(11):692–
504 704.
- 505 14. Cerhan JR, Berndt SI, Vijai J, et al. Genome-wide association study identifies
506 multiple susceptibility loci for diffuse large B cell lymphoma. *Nature Genetics.*
507 2014;46(11):1233–1238.
- 508 15. Conde L, Halperin E, Akers NK, et al. Genome-wide association study of follicular
509 lymphoma identifies a risk locus at 6p21.32. *Nat Genet.* 2010;42(8):661–664.
- 510 16. Frampton M, da Silva Filho MI, Broderick P, et al. Variation at 3p24.1 and 6q23.3
511 influences the risk of Hodgkin’s lymphoma. *Nature Communications.* 2013;4:.

- 512 17. Kumar V, Matsuo K, Takahashi A, et al. Common variants on 14q32 and 13q12 are
513 associated with DLBCL susceptibility. *J Hum Genet.* 2011;56(6):436–439.
- 514 18. Moutsianas L, Enciso-Mora V, Ma YP, et al. Multiple Hodgkin lymphoma–
515 associated loci within the HLA region at chromosome 6p21.3. *Blood.*
516 2011;118(3):670–674.
- 517 19. Skibola CF, Bracci PM, Halperin E, et al. Genetic variants at 6p21.33 are associated
518 with susceptibility to follicular lymphoma. *Nat Genet.* 2009;41(8):873–875.
- 519 20. Urayama KY, Jarrett RF, Hjalgrim H, et al. Genome-Wide Association Study of
520 Classical Hodgkin Lymphoma and Epstein–Barr Virus Status–Defined Subgroups.
521 *JNCI J Natl Cancer Inst.* 2012;104(3):240–253.
- 522 21. Vijai J, Kirchhoff T, Schrader KA, et al. Susceptibility Loci Associated with Specific
523 and Shared Subtypes of Lymphoid Malignancies. *PLoS Genetics.*
524 2013;9(1):e1003220.
- 525 22. Skibola CF, Berndt SI, Vijai J, et al. Genome-wide Association Study Identifies Five
526 Susceptibility Loci for Follicular Lymphoma outside the HLA Region. *The American*
527 *Journal of Human Genetics.* 2014;95(4):462–471.
- 528 23. Vijai J, Wang Z, Berndt SI, et al. A genome-wide association study of marginal zone
529 lymphoma shows association to the HLA region. *Nat Commun.* 2015;6:.
- 530 24. Tan DEK, Foo JN, Bei J-X, et al. Genome-wide association study of B cell non-
531 Hodgkin lymphoma identifies 3q27 as a susceptibility locus in the Chinese
532 population. *Nature Genetics.* 2013;45(7):804–807.
- 533 25. Schoeni-Affolter F, Ledergerber B, Rickenbach M, et al. Cohort Profile: The Swiss
534 HIV Cohort Study. *Int J Epidemiol.* 2010;39(5):1179–1189.
- 535 26. Besson C, Lancar R, Prevot S, et al. Outcomes for HIV-associated diffuse large B-
536 cell lymphoma in the modern combined antiretroviral therapy era. *AIDS.*
537 2017;31(18):2493.
- 538 27. Dalmaso C, Carpentier W, Meyer L, et al. Distinct Genetic Loci Control Plasma
539 HIV-RNA and Cellular HIV-DNA Levels in HIV-1 Infection: The ANRS Genome
540 Wide Association 01 Study. *PLoS ONE.* 2008;3(12):e3907.
- 541 28. Fellay J, Ge D, Shianna KV, et al. Common Genetic Variation and the Control of
542 HIV-1 in Humans. *PLoS Genetics.* 2009;5(12):e1000791.
- 543 29. Loh P-R, Danecek P, Palamara PF, et al. Reference-based phasing using the
544 Haplotype Reference Consortium panel. *Nature Genetics.* 2016;48(11):1443–1448.
- 545 30. Durbin R. Efficient haplotype matching and storage using the positional Burrows–
546 Wheeler transform (PBWT). *Bioinformatics.* 2014;30(9):1266–1272.
- 547 31. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for
548 genotype imputation. *Nat Genet.* 2016;advance online publication:
- 549 32. Chang CC, Chow CC, Tellier LC, et al. Second-generation PLINK: rising to the
550 challenge of larger and richer datasets. *Gigascience.* 2015;4(1):1–16.
- 551 33. Price AL, Patterson NJ, Plenge RM, et al. Principal components analysis corrects for
552 stratification in genome-wide association studies. *Nat. Genet.* 2006;38(8):904–909.
- 553 34. The International HapMap 3 Consortium. Integrating common and rare genetic
554 variation in diverse human populations. *Nature.* 2010;467(7311):52–58.
- 555 35. Manichaikul A, Mychaleckyj JC, Rich SS, et al. Robust relationship inference in
556 genome-wide association studies. *Bioinformatics.* 2010;26(22):2867–2873.

- 557 36. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A Tool for Genome-wide
558 Complex Trait Analysis. *The American Journal of Human Genetics*. 2011;88(1):76–
559 82.
- 560 37. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in
561 the application of mixed-model association methods. *Nat Genet*. 2014;46(2):100–
562 106.
- 563 38. Kichaev G, Roytman M, Johnson R, et al. Improved methods for multi-trait fine
564 mapping of pleiotropic risk loci. *Bioinformatics*. 2017;33(2):248–255.
- 565 39. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep
566 learning–based sequence model. *Nature Methods*. 2015;12(10):931–934.
- 567 40. Khan A, Fornes O, Stigliani A, et al. JASPAR 2018: update of the open-access
568 database of transcription factor binding profiles and its web framework. *Nucleic
569 Acids Res*. 2018;46(D1):D260–D266.
- 570 41. Rao SSP, Huntley MH, Durand NC, et al. A 3D Map of the Human Genome at
571 Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*.
572 2014;159(7):1665–1680.
- 573 42. Wang Y, Song F, Zhang B, et al. The 3D Genome Browser: a web-based browser for
574 visualizing 3D genome organization and long-range chromatin interactions. *Genome
575 Biology*. 2018;19(1):151.
- 576 43. Schofield EC, Carver T, Achuthan P, et al. CHiCP: a web-based tool for the
577 integrative and interactive visualization of promoter capture Hi-C datasets.
578 *Bioinformatics*. 2016;32(16):2511–2513.
- 579 44. Mifsud B, Tavares-Cadete F, Young AN, et al. Mapping long-range promoter
580 contacts in human cells with high-resolution capture Hi-C. *Nature Genetics*.
581 2015;47(6):598–606.
- 582 45. Cairns J, Freire-Pritchett P, Wingett SW, et al. CHiCAGO: robust detection of DNA
583 looping interactions in Capture Hi-C data. *Genome Biol*. 2016;17(1):127.
- 584 46. Lappalainen T, Sammeth M, Friedländer MR, et al. Transcriptome and genome
585 sequencing uncovers functional variation in humans. *Nature*. 2013;501(7468):506–
586 511.
- 587 47. Piasecka B, Duffy D, Urrutia A, et al. Distinctive roles of age, sex, and genetics in
588 shaping transcriptional variation of human immune responses to microbial
589 challenges. *PNAS*. 2018;115(3):E488–E497.
- 590 48. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature*.
591 2017;550(7675):204–213.
- 592 49. Alpern D, Gardeux V, Russeil J, et al. BRB-seq: ultra-affordable high-throughput
593 transcriptomics enabled by bulk RNA barcoding and sequencing. *Genome Biology*.
594 2019;20(1):71.
- 595 50. Berndt SI, Skibola CF, Joseph V, et al. Genome-wide association study identifies
596 multiple risk loci for chronic lymphocytic leukemia. *Nature Genetics*.
597 2013;45(8):868–876.
- 598 51. Johnson JL, Abecasis GR. GAS Power Calculator: web-based power calculator for
599 genetic association studies. *bioRxiv*. 2017;164343.
- 600 52. Mohammadi P, Desfarges S, Bartha I, et al. 24 Hours in the Life of HIV-1 in a T Cell
601 Line. *PLOS Pathogens*. 2013;9(1):e1003161.

- 602 53. Lim U, Kocarnik JM, Bush WS, et al. Pleiotropy of Cancer Susceptibility Variants on
603 the Risk of Non-Hodgkin Lymphoma: The PAGE Consortium. *PLoS ONE*.
604 2014;9(3):e89791.
- 605 54. Sei S, O'Neill DP, Stewart SK, et al. Increased Level of Stromal Cell-Derived Factor-
606 1 mRNA in Peripheral Blood Mononuclear Cells from Children with AIDS-related
607 Lymphoma. *Cancer Res.* 2001;61(13):5028–5037.
- 608 55. Rabkin CS, Yang Q, Goedert JJ, et al. Chemokine and Chemokine Receptor Gene
609 Variants and Risk of Non-Hodgkin's Lymphoma in Human Immunodeficiency
610 Virus-1–Infected Individuals. *Blood*. 1999;93(6):1838–1842.
- 611 56. Meng W, Xue S, Chen Y. The role of CXCL12 in tumor microenvironment. *Gene*.
612 2018;641:105–110.
- 613 57. Peled A, Klein S, Beider K, Burger JA, Abraham M. Role of CXCL12 and CXCR4
614 in the pathogenesis of hematological malignancies. *Cytokine*. 2018;109:11–16.
- 615 58. Piovan E, Tosello V, Indraccolo S, et al. Chemokine receptor expression in EBV-
616 associated lymphoproliferation in hu/SCID mice: implications for CXCL12/CXCR4
617 axis in lymphoma generation. *Blood*. 2005;105(3):931–939.
- 618 59. Strober BJ, Elorbany R, Rhodes K, et al. Dynamic genetic regulation of gene
619 expression during cellular differentiation. *Science*. 2019;364(6447):1287–1290.
- 620 60. Quigley M, Pereyra F, Nilsson B, et al. Transcriptional analysis of HIV-specific
621 CD8+ T cells shows that PD-1 inhibits T cell function by upregulating BATF. *Nat*.
622 *Med*. 2010;16(10):1147–1151.
- 623 61. Gasser O, Bihl FK, Wolbers M, et al. HIV Patients Developing Primary CNS
624 Lymphoma Lack EBV-Specific CD4⁺ T Cell Function Irrespective of Absolute
625 CD4⁺ T Cell Counts. *PLoS Medicine*. 2007;4(3):6.
- 626 62. SH S, E C, NL H, et al. WHO Classification of Tumours of Haematopoietic and
627 Lymphoid Tissues.

628

Tables

Table 1. Summary of included samples and studies

Cohort	Cases	Controls	Lambda	Genotyping chips	Years of NHL diagnosis	Control inclusion criteria
SHCS	145	1090	1.00	Illumina	2000 - 2017	HIV < 2005, no cancer diagnosis as of 2017 & matched with age
<i>Age (median)</i>	<i>61</i>	<i>58</i>		HumanOmniExpress-24, Human1M, Human610, HumanHap550, HumanCore-12		
<i>Sex (%Male)</i>	<i>91%</i>	<i>80%</i>				
ANRS	61	562	1.00	Illumina Human	2008 - 2015	No cancer diagnosis
<i>Age (median)</i>	<i>50</i>	<i>34</i>		Omni5 Exome 4v 1-2, Illumina 300		
<i>Sex (%Male)</i>	<i>89%</i>	<i>87%</i>				
MACS	72	272	1.01	Illumina 1MV1, Human1M-Duo, HumanHap550	1985 - 2013	Matched to cases in terms of age, treatment & time of infection
<i>Age (median)</i>	<i>69</i>	<i>68</i>				
<i>Sex (%Male)</i>	<i>100%</i>	<i>100%</i>				

Cohort and patient characteristics for the SHCS, ANRS and MACS cohorts. Lambda indicates the genomic inflation factor from the individual cohort GWAS.

Table 2. Significant association with HIV-related NHL

Chr	Pos	SNP	Ref	Alt	P	OR
10	44673557	rs7919208	A	G	4.77e-11	1.23
10	44677967	rs149399290	T	C	3.09e-08	1.20
10	44678218	rs17155463	T	A	3.09e-08	1.20
10	44678262	rs17155474	C	T	3.09e-08	1.20
10	44678454	rs17155478	T	C	3.09e-08	1.20
10	44678898	rs12249837	G	A	3.09e-08	1.20
10	44680902	rs10608969	T	TAAAGA	3.09e-08	1.20

Variants significantly associated with HIV-related NHL in a weighted Z-score-based meta-analysis of all individuals included in the SHCS, ANRS and MACS cohorts. Odds ratios (OR) were transformed from betas using the formula $OR = \exp(\text{beta})$.

Figure Legends

Figure 1. Genome-wide association analysis. (A) Schematic of analysis pipeline. (B) Quantile-quantile plot of the observed $-\log_{10}(\text{p-value})$ (black dots, y-axis) versus expected $-\log_{10}(\text{p-values})$ under the null hypothesis (red line) to check for any genomic inflation of the observed p-values. No genomic inflation is observed, with the genomic inflation factor $\lambda = 0.99$. (C) Manhattan plot of all obtained p-values for each variant included in the meta-analysis. The genome-wide threshold ($P = 5e-8$) for significance is marked by a dotted line. Only variants at the CXCL12 locus were found to be significant.

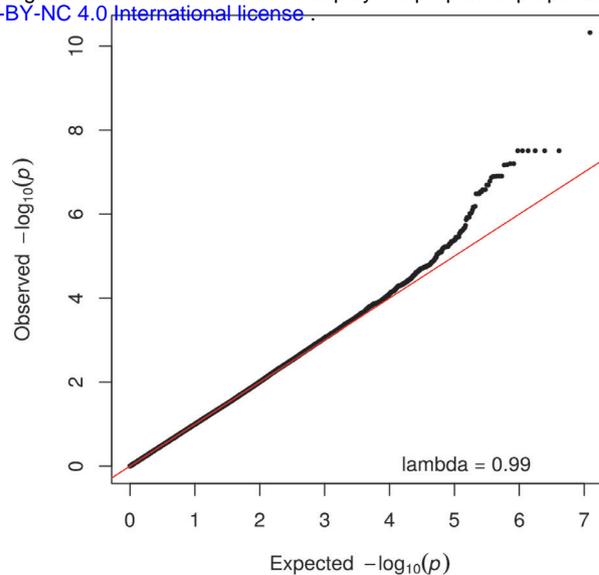
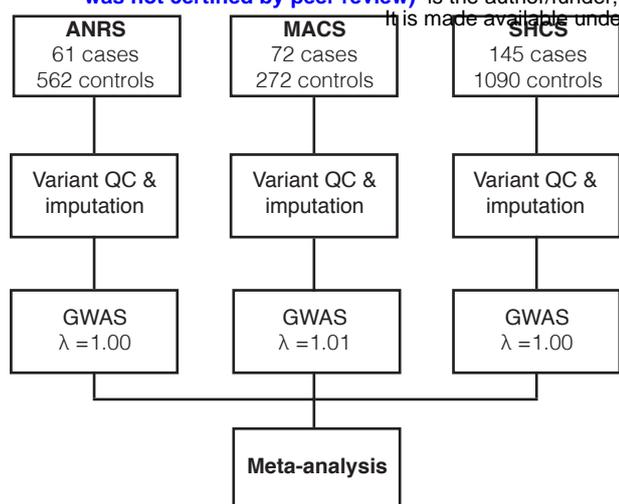
Figure 2. Fine mapping of genome-wide significant hits with PAINTOR. (A) The 99% credible set and posterior probabilities of being the causal variant. The genomic positions are listed on the x-axis. Bottom tracks represent DNAase and chromatin marks obtained from GM12878 cells as well as TFBS from the Roadmap Epigenomics Project and ENCODE in the region. (B) Locus plot of the associated variants, highlighting the LD relationship, based on the SHCS cohort. The top variant rs7919208 is marked by a black diamond.

Figure 3. Novel transcription factor binding site and long-range interactions. (A) Canonical motifs of BATF and JUND with the underlying genomic reference sequence and the nucleotide change caused by rs7919208. (B) Promoter capture Hi-C analysis in the GM12878 cell line of the region with the predicted causal variant and CXCL12. Variants and their level of association in the meta-analysis are marked in the inner grey circle.

Genome-wide significant variants are colored green. Purple lines indicate significant interactions between promoter and other genomic regions. (C) TADs in the GM12878 cell line in the region of CXCL12. The yellow and blue boxes indicate the called TADs from the Hi-C contact map above. The plot is centered on rs7919208.

Figure 1

A medRxiv preprint doi: <https://doi.org/10.1101/19011999>; this version posted November 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY-NC 4.0 International license](https://creativecommons.org/licenses/by-nc/4.0/).



C

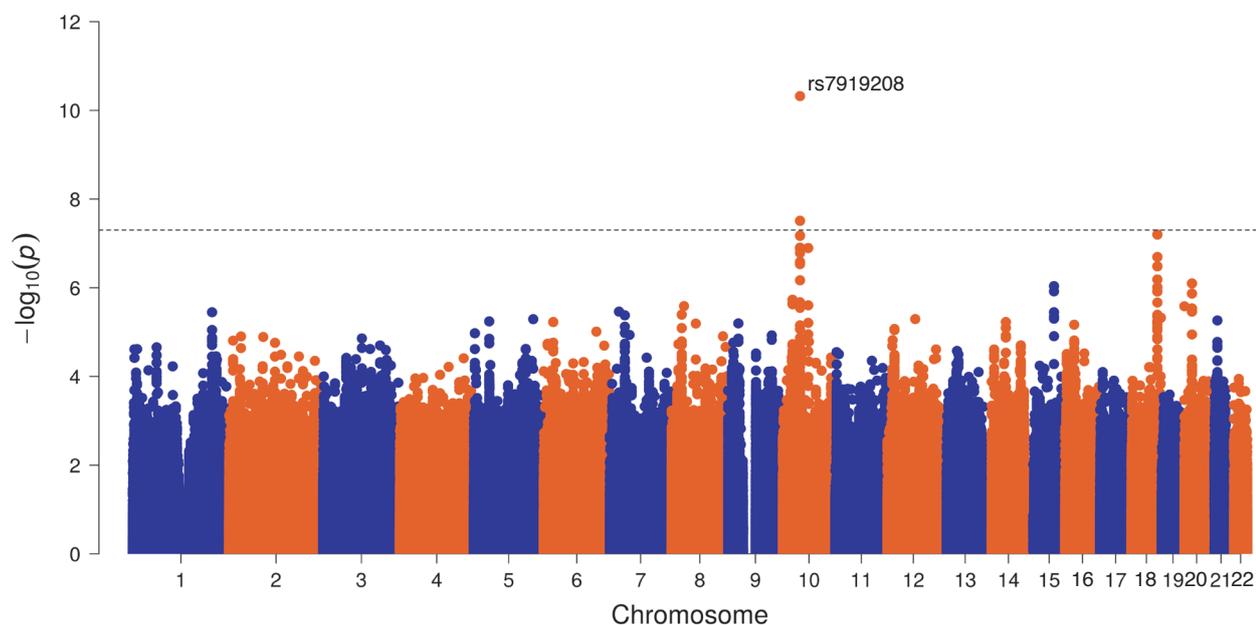


Figure 2

A medRxiv preprint doi: <https://doi.org/10.1101/19011999>; this version posted November 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY-NC 4.0 International license](https://creativecommons.org/licenses/by-nc/4.0/).

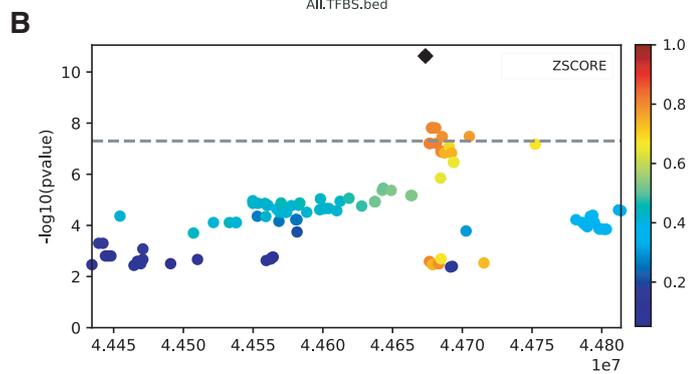
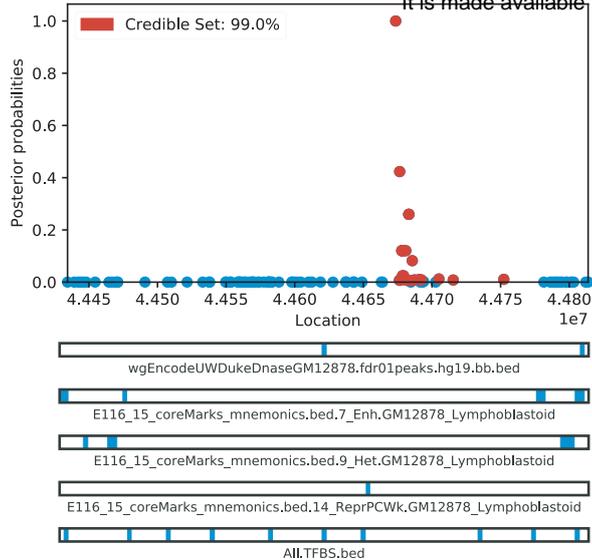


Figure 3

A medRxiv preprint doi: <https://doi.org/10.1101/19011999>; this version posted November 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY-NC 4.0 International license](#).

