nature

https://doi.org/10.1038/s41586-020-2277-x

Supplementary information

Complement genes contribute sex-biased vulnerability in diverse disorders

In the format provided by the authors and unedited

Nolan Kamitaki✉, Aswin Sekar, Robert E. Handsaker, Heather de Rivera, Katherine Tooley, David L. Morris, Kimberly E. Taylor, Christopher W. Whelan, Philip Tombleson, Loes M. Olde Loohuis, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Michael Boehnke, Robert P. Kimberly, Kenneth M. Kaufman, John B. Harley, Carl D. Langefeld, Christine E. Seidman, Michele T. Pato, Carlos N. Pato, Roel A. Ophoff, Robert R. Graham, Lindsey A. Criswell, Timothy J. Vyse✉ & Steven A. McCarroll✉

Supplementary Material for

Complement genes contribute sex-biased vulnerability in diverse illnesses

Nolan Kamitaki, Aswin Sekar, Robert E. Handsaker, Heather de Rivera,
Katherine Tooley, David L. Morris, Kimberly E. Taylor, Christopher W. Whelan,
Philip Tombleson, Loes M. Olde Loohuis, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Michael Boehnke, Robert P. Kimberly, Kenneth M. Kaufman, John B. Harley,
Carl D. Langefeld, Christine E. Seidman, Michele T. Pato, Carlos N. Pato, Roel A. Ophoff,
Robert R. Graham, Lindsey A. Criswell, Timothy J. Vyse, Steven A. McCarroll

Contents:

Supplementary Note 1 - Fine mapping of an independent association signal in the MHC class II region 1

Supplementary Note 2 - Sex bias of *C3* and *C4* gene expression across human tissues 4

Supplementary References 5

Supplementary Note 1 - Fine mapping of an independent association signal in the MHC class II region

Linkage disequilibrium of *C4* variation to other variants in the MHC genomic region differs by ancestry

The linkage-disequilibrium (LD) relationships of *C4* variation to other genetic variation in the MHC region differ greatly between European-ancestry and African American cohorts , and are in general far stronger among Europeans. For example, **Extended Data Fig. 4d, e** shows the LD-correlation (r^2) of SNPs across the MHC region to an estimate of *C4* composite risk (the composite estimate of *C4*-derived SLE risk derived from the genotype-group risk measurements in **Fig. 2a)**. (Other *C4* features, such as total *C4* gene copy number, also exhibit strikingly different correlations with genetic markers between the two populations.). Most notably, LD in European ancestry is widespread across the extended MHC genomic region (**Extended Data Fig. 4d**) – and particularly strong in the nearby MHC class II region (32-33 Mb) – while strong LD in African Americans is localized primarily to a much-smaller region immediately flanking the *C4* genes (**Extended Data Fig. 4e**).

A direct comparison of the two population-specific LD patterns confirms that nearly all variants with LD to *C4* variation have greater linkage in the European-ancestry cohort than in the African American cohort (**Extended Data Fig. 4f**).

Initial (*C4*-naïve) association analysis produces divergent association results in the MHC genomic region for European-ancestry and African American cohorts

In unconditional (*C4*-naïve) association analysis of SLE for each variant in the MHC region, there is little correlation between results from European-ancestry and African American cohorts[3,5] (**Extended Data Fig. 4g, h**). In particular, a SNP's level of association in one population offers very little predictive information about its level of association to SLE in the other population (**Extended Data Fig. 4g**). This could in principle result from multiple population-specific variants or even population-specific biology – but more-parsimonious explanations are both strongly preferred, and more strongly constrained and testable by available data.

Of course, *C4* alleles have both allele-frequency and LD differences between these populations (**Extended Data Table 1**) and therefore could be a potential contributor to the population differences visible in **Extended Data Fig. 4g, h**. To evaluate this possibility, the variants in **Extended Data Fig. 4g, h** are colored orange in proportion to their European-ancestry LD (r^2) to *C4* composite risk, revealing that SNPs with European-specific associations tend to have strong LD to *C4* in Europeans. This highlights the strong effect that *C4* alleles would be likely to have in shaping the relative association strengths of genetic markers throughout the MHC genomic region.

Controlling for *C4* composite risk aligns the residual association signal in the MHC genomic region across ancestries

If, beginning with the European-ancestry cohort, we now consider SNPs, not in a naïve association analysis, but in a joint association analysis together with *C4* (i.e. with *C4* composite risk as a covariate), then the association statistics for variants in the two cohorts begin to align with each other more strongly (**Extended Data Fig. 4i**, which should be compared to **Extended Data Fig. 4g**). Further adjusting the association statistics for the African American cohort analysis to account for *C4* changes the overall pattern more modestly (**Extended Data Fig. 4j, k**); this likely reflects reduced LD to *C4* alleles among African Americans. The genetic signal for SLE in both populations (in this C4-adjusted analysis) now converges onto two variants (rs2105898 and rs9271513, separated by 975 bp) which are in strong (r^2>0.9) LD across both populations and can be considered as a haplotype. rs2105898 is used as an index SNP in further analyses in the following sections and as such is highlighted in **Extended Data Fig. 4j, k**, with variants colored purple in proportion to

their European-ancestry LD (r^2) with rs2105898. Population-specific *HLA* alleles (*DRB1**15:01 and *DRB1**15:03) have been proposed as potential explanations for the apparently divergent association signals across European-ancestry and African American populations[3,5]; in **Extended Data Fig. 4j**, these alleles are shown with grey triangles.

Much of the remaining trans-ancestry differences in association pattern in the MHC region appear to be explained by differences in LD patterns between populations; in **Extended Data Fig. 4l**, blueness of variants represents greater LD to rs2105898 in the European-ancestry cohort (relative to LD among African Americans) and redness represents greater LD in the African American cohort (relative to LD among European-ancestry individuals). Notably, the many variants with relatively stronger association signals among Europeans (including *DRB1**1501) exhibit stronger LD to rs2105898 among European-ancestry individuals, while select variants with relatively stronger association signals among African Americans (including *DRB1**1503) exhibit stronger LD to rs2105898 among African Americans. The strong LD this risk haplotype has with *DRB1**15:01 in Europeans and *DRB1**15:03 in African Americans may explain earlier findings of population-specific associations with *DRB1**15:01 in Europeans and *DRB1**15:03 in African Americans.

This analysis also indicates that while much, if not all, of the European ancestry-specific association after controlling for *C4* composite risk can be accounted for by European ancestry-specific LD to rs2105898, this is not true for African Americans, who may harbor at least one additional, independent genetic effect not explained by the above analysis.

The *C4*-independent association signal comprising rs2105898 and another linked variant defines strong pan-tissue expression QTLs for *HLA* class II genes

Data from the GTEx Consortium[52] (v7) included 227 instances (gene/tissue pairs) in which this haplotype of two variants (rs2105898 and rs9271513) associated with elevated (*HLA-DRB1*, *-DRB5*, *-DQA1*, and *-DQB1*) or reduced (*HLA-DRB6*, *-DQA2*, and *-DQB2*) expression of an *HLA* class II gene with at least nominal ($p < 10^{-4}$) significance. Some of the strongest associations at each gene ($p < 10^{-8}$ to 10^{-76}) were in whole blood, but expression QTLs elsewhere can also reflect the presence of blood and immune cells within those tissues[53]. Although eQTL analyses of *HLA* genes may be affected by read-alignment artifacts in these genes' hyperpolymorphic domains, most such observed signals are robust after adjusting for individual *HLA* alleles[54].

The haplotype with elevated expression of *HLA-DRB1*, *-DRB5*, *-DQA1*, and *-DQB1* (allele frequency 0.20 among Europeans, 0.22 among African Americans) associated with increased SLE risk (odds ratio) of 1.52 (95% CI: 1.44-1.61; $p < 10^{-48}$) in Europeans and 1.49 (95% CI: 1.35-1.63; $p < 10^{-16}$) in African Americans in analyses adjusting for *C4* effects. The risk haplotype tagged by rs2105898 tended to be on low-risk *C4* haplotypes in Europeans, a relationship that may have made both genetic influences (*C4* and the rs2105898 haplotype) harder to recognize in earlier work; controlling for either rs2105898 or *C4* (**Extended Data Fig. 3b**) greatly increased the association of SLE with the other genetic influence (**Extended Data Table 3**). Controlling for the simpler (2.3)*C4A+C4B* model in SNP associations with SjS (as precision of estimates of individual alleles were low due to sample size) also pointed strongly to the same haplotype, with the same allele of rs2105898 associating in the same direction but larger effect (OR: 1.96; 95% CI: 1.64-2.34; $p < 10^{-12}$) as compared to SLE (**Extended Data Fig. 3d**).

The rs2105898 haplotype affects the XL9 hotspot of active chromatin and transcription factor binding

The two variants defining this short haplotype reside within the XL9 regulatory region[55,56], a well-studied region of open chromatin that contains abundant chromatin marks characteristic of active enhancers and transcription factor binding sites. As characterized by HaploReg v4.1[57], rs2105898 in particular lies within multiple histone marks that are associated with active enhancers (6 tissues), in the XL9 region of open

chromatin (15 tissues), and under ChIP-seq binding peaks for 19 transcription factors (**Extended Data Fig. 5a**, data from the ENCODE project[58,59] and Roadmap Epigenomics Consortium[60]).

rs2105898 disrupts a binding site for the ZNF143 transcription factor

We identified transcription factors whose binding motif is significantly affected by rs2105898. The strongest hit (ZNF143) is also among the transcription factors that have been determined by ChIP-seq analysis (from the ENCODE project) to bind to DNA sequence at rs2105898 (**Extended Data Fig. 5b**). ZNF143 is a widely expressed zinc-finger transcription factor that has been found to anchor chromatin interactions that connect distal regulatory elements with gene promoters[61].

Two databases (HaploReg, CIS-BP TF[62]) evaluate ZNF143 as having low or no binding to the minor (reference) allele of rs2105898 and very high affinity to the major (alternate) allele of rs2105898:

CIS-BP (log score)
Reference (T) allele: 4.459
Alternate (G) allele: 13.273

HaploReg (log score)
Reference (T) allele: -0.4
Alternate (G) allele: 11.5

ZNF143 is a recently identified component of complexes that maintain topologically associated domains (TADs) in concert with CTCF and cohesin (SMC1, SMC3, RAD21, STAG1/2), both of which also have numerous ChIP-seq peaks overlapping rs2105898. Specifically, ZNF143 has been found to directly bind and regulate promoter interaction with distal enhancers, congruous with the observation of numerous RNA polymerase ChIP-seq peaks at rs2105898 but with nearest promoter being 14.5kb away (*HLA-DQA1*, downstream). Furthermore, as this region lies in the genomic neighborhood of many genes for which rs2105898 is a multi-tissue eQTL (*HLA-DRB1*, *-DRB5*, *-DRB6* upstream and *-DQA1*, *-DQA2*, *-DQB1*, and *-DQB2* downstream) it seems plausible that by regulating ZNF143 binding, rs2105898 alters the interaction between this enhancer region and the promoters of the numerous proximal *HLA* class II genes. We also note that there have been other findings on the *DRB1**15:01 haplotype in European-ancestry individuals (among whom *DRB1**15:01 is in high LD with rs2105898) including hypomethylation[63] that could in principle reflect the effect of this haplotype.

rs2105898 is in strong LD with index SNPs for other autoimmune disorders

rs2105898 also has high LD to the most strongly associated SNPs for other autoimmune phenotypes in the NHGRI-EBI GWAS catalog[64]. Of these associations, the strongest is to the peak SNP for multiple sclerosis oligoclonal band status (r^2=0.88,D'=0.98). Also in high LD to rs2105898 is a shared peak SNP for associations to broad multiple sclerosis, immunoglobulin A production, ulcerative colitis, and Crohn's disease (all r^2=0.49,D'=0.98).

Supplementary Note 2 - Sex bias of *C4* and *C3* gene expression across human tissues

We analyzed expression levels of C4 and C3 across 46 human tissues for which the GTEx project35 has analyzed (by RNA-seq) tissues from both men and women in the most recent release at time of writing (v8). Specifically, for each tissue we applied a non-parametric unsigned Mann-Whitney rank–sum test to compare transcripts per million (TPM) measurements between men and women and calculated Bonferroni-corrected p-values. For C4, the TPM values for C4A and C4B were summed (per sample).

At a cutoff of $\alpha = 0.01$, results for brain, blood, liver and lymphoblastoid cells were insignificant; the only two tissues with significant sex differences for C4 expression were "Adipose - Subcutaneous" (higher in men) and "Breast - Mammary Tissue" (higher in women). For C3, three tissues showed higher expression in women: "Breast - Mammary Tissue", "Skin - Not Sun Exposed (Suprapubic)", and "Skin - Sun Exposed (Lower leg)". These differences may well reflect sexual dimorphism in cell-type composition; for example, breast tissue undergoes monthly remodeling with infiltrating immune cells in women[65]. More broadly, breast tissue and subcutaneous fat distribution[66] are highly dimorphic between men and women. Though we cannot formally exclude the possibility that these tissues contribute to the sex difference protein concentrations we observed in CSF (and earlier studies have also observed in plasma), we note all cases except for "Adipose - Subcutaneous" involve higher detected expression levels in women and thus have the opposite sign of the observed difference in CSF/plasma protein concentration.

Supplementary References

52 GTEx Consortium *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213, doi:10.1038/nature24277 (2017).

53 Raj, P. *et al.* Regulatory polymorphisms modulate the expression of HLA class II molecules and promote autoimmunity. *Elife* **5**, doi:10.7554/eLife.12089 (2016).

54 Aguiar, V. R. C., Cesar, J., Delaneau, O., Dermitzakis, E. T. & Meyer, D. Expression estimation and eQTL mapping for HLA genes with a personalized pipeline. *PLoS Genet* **15**, e1008091, doi:10.1371/journal.pgen.1008091 (2019).

55 Majumder, P., Gomez, J. A. & Boss, J. M. The human major histocompatibility complex class II HLA-DRB1 and HLA-DQA1 genes are separated by a CTCF-binding enhancer-blocking element. *J Biol Chem* **281**, 18435-18443, doi:10.1074/jbc.M601298200 (2006).

56 Majumder, P., Gomez, J. A., Chadwick, B. P. & Boss, J. M. The insulator factor CTCF controls MHC class II gene expression and is required for the formation of long-distance chromatin interactions. *J Exp Med* **205**, 785-798, doi:10.1084/jem.20071843 (2008).

57 Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* **40**, D930-934, doi:10.1093/nar/gkr917 (2012).

58 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).

59 Rosenbloom, K. R. *et al.* ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res* **41**, D56-63, doi:10.1093/nar/gks1172 (2013).

60 Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).

61 Bailey, S. D. *et al.* ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. *Nat Commun* **2**, 6186, doi:10.1038/ncomms7186 (2015).

62 Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431-1443, doi:10.1016/j.cell.2014.08.009 (2014).

63 Kular, L. *et al.* DNA methylation as a mediator of HLA-DRB1*15:01 and a protective variant in multiple sclerosis. *Nat Commun* **9**, 2397, doi:10.1038/s41467-018-04732-5 (2018).

64 Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005-D1012, doi:10.1093/nar/gky1120 (2019).

65 Ramakrishnan, R., Khan, S. A. & Badve, S. Morphological changes in breast tissue with menstrual cycle. *Mod Pathol* **15**, 1348-1356, doi:10.1097/01.MP.0000039566.20817.46 (2002).

66 Shi, H. & Clegg, D. J. Sex differences in the regulation of body weight. *Physiol Behav* **97**, 199-204, doi:10.1016/j.physbeh.2009.02.017 (2009).