Article

Complement genes contribute sex-biased vulnerability in diverse disorders

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

Received: 14 October 2019

Accepted: 28 February 2020

Published online: 11 May 2020

Check for updates

Nolan Kamitaki[1,2]✉, Aswin Sekar[1,2], Robert E. Handsaker[1,2], Heather de Rivera[1,2], Katherine Tooley[1,2], David L. Morris[3], Kimberly E. Taylor[4], Christopher W. Whelan[1,2], Philip Tombleson[3], Loes M. Olde Loohuis[5,6], Schizophrenia Working Group of the Psychiatric Genomics Consortium*, Michael Boehnke[7], Robert P. Kimberly[8], Kenneth M. Kaufman[9], John B. Harley[9], Carl D. Langefeld[10], Christine E. Seidman[1,11,12], Michele T. Pato[13], Carlos N. Pato[13], Roel A. Ophoff[5,6], Robert R. Graham[14], Lindsey A. Criswell[4], Timothy J. Vyse[3]✉ & Steven A. McCarroll[1,2]✉

Many common illnesses, for reasons that have not been identified, differentially affect men and women. For instance, the autoimmune diseases systemic lupus erythematosus (SLE) and Sjögren's syndrome affect nine times more women than men[1], whereas schizophrenia affects men with greater frequency and severity relative to women[2]. All three illnesses have their strongest common genetic associations in the major histocompatibility complex (MHC) locus, an association that in SLE and Sjögren's syndrome has long been thought to arise from alleles of the human leukocyte antigen (HLA) genes at that locus[3–6]. Here we show that variation of the complement component 4 (C4) genes *C4A* and *C4B*, which are also at the MHC locus and have been linked to increased risk for schizophrenia[7], generates 7-fold variation in risk for SLE and 16-fold variation in risk for Sjögren's syndrome among individuals with common C4 genotypes, with *C4A* protecting more strongly than *C4B* in both illnesses. The same alleles that increase risk for schizophrenia greatly reduce risk for SLE and Sjögren's syndrome. In all three illnesses, C4 alleles act more strongly in men than in women: common combinations of *C4A* and *C4B* generated 14-fold variation in risk for SLE, 31-fold variation in risk for Sjögren's syndrome, and 1.7-fold variation in schizophrenia risk among men (versus 6-fold, 15-fold and 1.26-fold variation in risk among women, respectively). At a protein level, both C4 and its effector C3 were present at higher levels in cerebrospinal fluid and plasma[8,9] in men than in women among adults aged between 20 and 50 years, corresponding to the ages of differential disease vulnerability. Sex differences in complement protein levels may help to explain the more potent effects of C4 alleles in men, women's greater risk of SLE and Sjögren's syndrome and men's greater vulnerability to schizophrenia. These results implicate the complement system as a source of sexual dimorphism in vulnerability to diverse illnesses.

SLE (commonly referred to as lupus) is a systemic autoimmune disease of unknown cause. Risk of SLE is largely (66%) heritable[10], although it may have environmental triggers, as onset often follows events that damage cells, such as infection and severe sunburn[11]. Most patients with SLE produce autoantibodies against nucleic acid complexes, including ribonucleoproteins and DNA[12].

In genetic studies, SLE is most strongly associated with variation across the MHC locus, which contains the HLA genes[3]. However, conclusive attribution of this association to specific genes and alleles has been difficult; the identities of the most likely genetic sources have been frequently revised as genetic studies have grown in size[4,5]. In several other autoimmune diseases, including type 1 diabetes, coeliac disease

[1]Department of Genetics, Harvard Medical School, Boston, MA, USA. [2]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [3]Department of Medical and Molecular Genetics, King's College London, London, UK. [4]Rosalind Russell/Ephraim P. Engleman Rheumatology Research Center, Division of Rheumatology, UCSF School of Medicine, San Francisco, CA, USA. [5]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA, USA. [6]Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, CA, USA. [7]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA. [8]Division of Clinical Immunology and Rheumatology, University of Alabama at Birmingham, Birmingham, AL, USA. [9]Center for Autoimmune Genomics and Etiology (CAGE), Department of Pediatrics, Cincinnati Children's Medical Center & University of Cincinnati and the US Department of Veterans Affairs Medical Center, Cincinnati, OH, USA. [10]Department of Biostatistical Sciences, Wake Forest School of Medicine, Winston–Salem, NC, USA. [11]Howard Hughes Medical Institute, Chevy Chase, MD, USA. [12]Cardiovascular Division, Brigham and Women's Hospital, Boston, MA, USA. [13]SUNY Downstate Medical Center, Brooklyn, NY, USA. [14]Human Genetics, Genentech, South San Francisco, CA, USA. *A list of participants and their affiliations appears in the online version of the paper. ✉e-mail: nolan_kamitaki@hms.harvard.edu; timothy.vyse@kcl.ac.uk; mccarroll@hms.harvard.edu

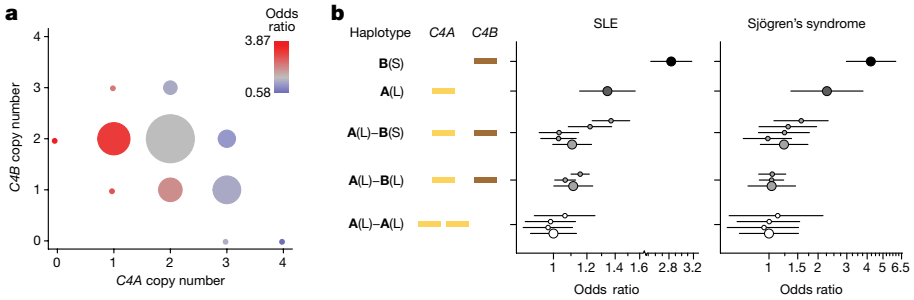


Fig. 1 | Association of SLE and Sjögren's syndrome with C4 alleles. a, Levels of SLE risk associated with 11 common combinations of *C4A* and *C4B* gene copy number. The colour of each circle reflects the level of SLE risk (odds ratio) associated with a specific combination of *C4A* and *C4B* gene copy numbers relative to the most common combination (two copies of *C4A* and two copies of *C4B*) in grey. The area of each circle is proportional to the number of individuals with that number of *C4A* and *C4B* genes. Paths from left to right on the plot reflect the effect of increasing *C4A* gene copy number (greatly reduced risk); paths from bottom to top reflect the effect of increasing *C4B* gene copy number (modestly reduced risk); and diagonal paths from upper left to lower right reflect the effect of exchanging *C4B* for *C4A* copies (modestly reduced risk). Data are from analysis of 6,748 patients with SLE and 11,516 unaffected controls of European ancestry. The odds ratios are reported with confidence intervals in Extended Data Fig. 2c. **b**, Risk of SLE and Sjögren's syndrome associated with common combinations of *C4A* and *C4B* gene copy number and flanking SNP haplotype. For each C4 locus structure, separate odds ratios are reported for each SNP haplotype background on which the C4 locus structure segregates. Data are from analyses of 6,748 patients with SLE and 11,516 controls (left) and 673 patients with Sjögren's syndrome and 1,153 controls (right). Error bars represent 95% confidence intervals around the effect-size estimate for each allele.

and rheumatoid arthritis, strong effects of the MHC locus arise from HLA alleles that cause the peptide-binding groove of HLA proteins to present a disease-critical autoantigen[13,14]. By contrast, in SLE, genetic variants in the MHC locus—including single nucleotide polymorphisms (SNPs) and HLA alleles—are broadly associated with the presence of diverse autoantibodies[15].

The *C4A* and *C4B* genes are also present in the MHC genomic region, between the class I and class II HLA genes. Classical complement proteins help eliminate debris from dead and damaged cells, attenuating the visibility of diverse intracellular proteins to the adaptive immune system. *C4A* and *C4B* commonly vary in genomic copy number[16] and encode complement proteins with distinct affinities for molecular targets[17,18]. SLE frequently presents with hypocomplementaemia that worsens during flares, possibly reflecting increased active consumption of complement[19]. Rare cases of severe, early-onset SLE can involve complete deficiency of a complement component (C4, C2 or C1Q)[20,21], and one of the strongest common-variant associations in SLE maps to *ITGAM*, which encodes a receptor for C3, the effector of C4 (ref. [22]). Although total C4 gene copy number is associated with SLE risk[23,24], this association is thought to arise from linkage disequilibrium (LD) with alleles of nearby HLA genes[25], which have been the focus of fine-mapping analyses[3,4].

The complex genetic variation of *C4A* and *C4B*—which consists of many alleles with different numbers of *C4A* and *C4B* genes—has been challenging to analyse in large cohorts. A recently feasible approach to this problem is based on imputation: people share long haplotypes with the same combinations of SNP and C4 alleles, such that *C4A* and *C4B* gene copy numbers can be imputed from SNP data[7]. To analyse *C4A* and *C4B* in large cohorts, we developed a way to identify C4 alleles from whole-genome sequence (WGS) data (Extended Data Fig. 1a, b), and then analysed WGS data from 1,265 individuals (from the Genomic Psychiatry Cohort[26,27]) to create a large multi-ancestry panel of 2,530 reference haplotypes of MHC-region SNPs, *C4A* alleles and *C4B* alleles (Extended Data Fig. 1c)—ten times as large as in earlier work[7]. We then analysed SNP data from the largest SLE genetic-association study[3] (ImmunoChip; 6,748 patients with SLE and 11,516 control subjects of European ancestry) (Extended Data Fig. 2a, b), imputing C4 alleles to estimate the SLE risk associated with common combinations of *C4A* and *C4B* gene copy numbers (Fig. 1a).

Groups of research participants with the eleven most common combinations of *C4A* and *C4B* gene copy number exhibited sevenfold variation in their relative risk of SLE (95% confidence interval (CI), [5.88, 8.61]; $P < 10^{-117}$ in total, Fig. 1a, Extended Data Fig. 2c). The relationship between SLE risk and C4 gene copy number exhibited consistent, logical patterns across the 11 genotype groups. For each *C4B* copy number, higher *C4A* copy number was associated with reduced SLE risk (Fig. 1a, Extended Data Fig. 2c). Conversely, for each *C4A* copy number, higher *C4B* copy number was associated with more modestly reduced SLE risk (Fig. 1a). Logistic-regression analysis estimated that the protection afforded by each copy of *C4A* (odds ratio 0.54; 95% confidence interval (CI): [0.51, 0.57]) was equivalent to that of 2.3 copies of *C4B* (odds ratio 0.77; 95% CI: [0.71, 0.82]). We calculated an initial C4 risk score as 2.3 times the number of *C4A* genes plus the number of *C4B* genes in an individual's genome. Despite clear limitations of this risk score—it is imperfectly imputed from flanking SNP haplotypes ($r^2 = 0.77$, Extended Data Table 1) and only approximates C4-derived risk by using a simple, linear model (to avoid overfitting the genetic data)—SNPs across the MHC genomic region tended to be associated with SLE in proportion to their level of LD with this risk score (Extended Data Fig. 3a).

Combinations of many different C4 alleles generate the observed variation in *C4A* and *C4B* gene copy number; particular *C4A* and *C4B* gene copy numbers have also arisen recurrently on multiple SNP haplotypes[7] (Extended Data Fig. 1c). Analysis of SLE risk in relation to each of these C4 alleles and SNP haplotypes reinforced the conclusion that *C4A* contributes strong protection, and *C4B* contributes more modest protection, from SLE, and that C4 genes (rather than nearby variants) are the principal drivers of this variation in risk levels (Fig. 1b).

These results prompted us to consider whether other autoimmune disorders with similar patterns of genetic association at the MHC genomic region might also be driven in part by variation of *C4A* and *C4B*. Primary Sjögren's syndrome is a heritable (54%)[28] systemic autoimmune disorder of exocrine glands, characterized primarily by dry eyes and mouth with other systemic effects. At a protein level, Sjögren's syndrome is (like SLE) characterized by diverse autoantibodies, including antinuclear antibodies targeting ribonucleoproteins[29], and hypocomplementaemia[30]. The largest source of common genetic risk for Sjögren's syndrome lies in the MHC genomic locus[31], with associations to the same haplotype(s) as in SLE[6] and with heterogeneous HLA associations in different ancestries[32]. We imputed C4 alleles into existing SNP data from a European-ancestry Sjögren's syndrome case–control cohort (673 cases and 1,153 controls). As in SLE, logistic-regression analyses found both *C4A* copy number (odds ratio 0.41; 95% CI: [0.34, 0.49]) and *C4B* copy number (OR: 0.67; 95% CI:

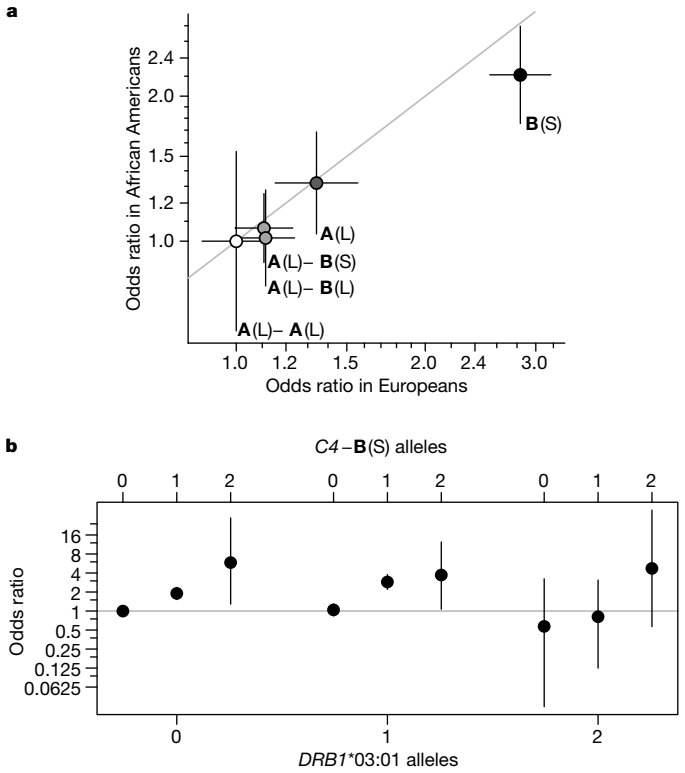


Fig. 2 | C4 and trans-ancestral analysis of the MHC-association signal in SLE. **a**, Common C4 alleles exhibit similar strengths of association to SLE (odds ratios) in European-ancestry and African American (1,494 SLE cases; 5,908 controls) cohorts. Error bars represent 95% confidence intervals around the effect size estimate for each sex. **b**, Analysis of SLE risk across combinations of C4-B(S) and *DRB1**03:01 genotypes in an African American SLE case–control cohort, in which the two alleles exhibit very little LD ($r^2 = 0.10$). On each *DRB1**03:01 genotype background, additional C4-B(S) alleles increase risk (that is, within each grouping). Whereas on each C4-B(S) background, *DRB1**03:01 alleles have no appreciable relationship with risk (this can be seen by comparing, for example, the first of the three points from each group). Error bars represent 95% confidence intervals around the effect-size estimate for each combination of C4-B(S) and *DRB1**03:01.

[0.53, 0.86]) to be protective against Sjögren's syndrome, generating a 16-fold variation in risk for Sjögren's syndrome (95% CI, [8.59, 30.89]; $P < 10^{-23}$ in total) among individuals with common C4 genotypes. The risk-equivalent ratio of *C4B* to *C4A* gene copies was similar in Sjögren's syndrome and SLE (about 2.3 to 1); furthermore, as with SLE, nearby SNPs associated with Sjögren's syndrome in proportion to their LD with a C4-derived risk score ((2.3)C4A + C4B) (Extended Data Fig. 3b), where C4A and C4B are the respective gene copy numbers. The distribution of Sjögren's syndrome risk across the individual *C4A* and *C4B* alleles and haplotypes revealed a pattern that, as in SLE, supported a greater protective effect from *C4A* than *C4B*, and little effect of flanking SNP haplotypes (Fig. 1b).

The association of SLE and Sjögren's syndrome with C4 gene copy number has long been attributed to the HLA-*DRB1*03:01* allele. In European populations, *DRB1*03:01* is in strong LD ($r^2 = 0.71$) with the common C4-B(S) allele, which lacks any *C4A* gene and is the highest-risk C4 allele in our analysis (Fig. 1b); many MHC-region SNPs associated with SLE and Sjögren's syndrome in proportion to their linkage-disequilibrium correlations with both C4 gene variation and *DRB1**03:01 (Extended Data Fig. 4a, b). Cohorts with other ancestries can have recombinant haplotypes that disambiguate the contributions of alleles that are in LD in Europeans. Among African Americans, we found that common C4 alleles exhibited far less LD with HLA alleles; in particular, the LD between C4-B(S) and *DRB1**03:01 was low ($r^2 = 0.10$) (Extended Data Table 2). Thus, genetic data from an African-American SLE cohort (1,494 cases and 5,908 controls) made it possible to distinguish between these potential genetic effects. Joint-association analysis of *C4A*, *C4B* and *DRB1*0301* implicated *C4A* ($P < 10^{-14}$) and *C4B* ($P < 10^{-5}$) but not *DRB1*0301* ($P = 0.29$) (Extended Data Table 3). Each C4 allele was associated with effect sizes of similar magnitude on SLE risk in Europeans and African Americans (Fig. 2a). An analysis specifically of combinations of C4-B(S) and *DRB1**03:01 allele dosages in African Americans showed that C4-B(S) alleles consistently increased SLE risk regardless of *DRB1**03:01 status, whereas *DRB1**03:01 had no consistent effect when controlling for C4-B(S) (Fig. 2b). Although C4 alleles had less LD with nearby variants on African American than on European haplotypes, SNPs across the genomic region associated with SLE in proportion to linkage-disequilibrium correlations with C4 variation in African Americans (Extended Data Fig. 4c).

Accounting for C4 alleles in jointly analysing the SLE-association data from African American and European ancestry cohorts also enabled mapping of an additional, more-modest genetic effect independent of *C4A* and *C4B*. This effect (tagged by rs2105898 and rs9271513) appeared to involve noncoding variation in the HLA class II XL9 region that is associated most strongly with expression levels (rather than the coding sequence) of many HLA class II genes (Extended Data Figs. 3c, d, 4d–l, 5 and Supplementary Note 1).

Alleles at C4 that increase dosage of *C4A* (and to a more modest extent *C4B*) appear to protect strongly against SLE and Sjögren's syndrome (Fig. 1a, b). By contrast, alleles that increase expression of *C4A* in the brain are more common among research participants with schizophrenia[6]. These same illnesses exhibit marked, and opposite, sex differences: SLE and Sjögren's syndrome are nine times more common among women of childbearing age than among men of a similar age[1], whereas in schizophrenia, women exhibit less severe symptoms, more frequent remission of symptoms, lower relapse rates and lower overall incidence[2]. Although the vast majority of genetic associations in complex diseases are shared between men and women[33], the SNPs most strongly associated with SLE risk within the MHC region are associated with larger potential effect sizes in men[34]. Thus, we sought to evaluate the possibility that the effects of C4 alleles on risk in SLE, Sjögren's syndrome and schizophrenia might differ between men and women.

Analysis indicated that the effects of C4 alleles were stronger in men. When a sex-by-C4 interaction term was included in association analyses, this term was significant for both SLE ($P = 0.002$) and schizophrenia ($P = 0.0024$), with larger C4 effects in men for both disorders. (Analysis of Sjögren's syndrome had limited power owing to the small number of men affected by Sjögren's syndrome). For both SLE and schizophrenia, the individual *C4A* and *C4B* alleles were consistently associated with stronger effects in men than women (Fig. 3a, b). SNPs across the MHC genomic region exhibited sex-biased association with SLE, Sjögren's syndrome and schizophrenia to the extent of their LD with C4 gene variation (Extended Data Fig. 6a–c).

The stronger effects of C4 alleles on male relative to female risk could arise from sex differences in C4 RNA expression, C4 protein levels or downstream responses to C4. Analysis of RNA expression in human tissues, using data from GTEx[35], identified no sex differences in C4 RNA expression in brain, blood, liver or lymphoblastoid cells (a more detailed description of this analysis can be found in Supplementary Note 2). We then analysed C4 protein in cerebrospinal fluid (CSF) from two panels of adult research participants ($n = 589$ total) in whom we had also measured C4 gene copy number (by direct genotyping or imputation). CSF C4 protein levels correlated strongly with C4 gene copy number ($P < 10^{-10}$, Extended Data Fig. 7a), so we normalized C4 protein measurements to the number of C4 gene copies. CSF from adult men contained on average 27% more C4 protein per C4 gene copy than CSF from women (meta-analysis $P = 9.9 \times 10^{-6}$, Fig. 3c). C4 acts by activating the complement component 3 (C3) protein, promoting C3 deposition

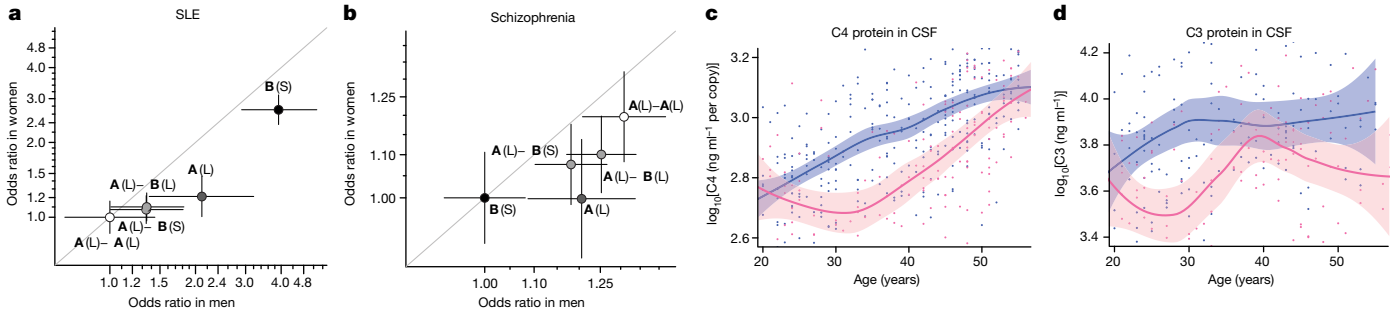


Fig. 3 | Sex differences in the magnitude of C4 genetic effects and complement protein concentrations. a, SLE risk (odds ratios) associated with the four most common C4 alleles in men (*x* axis) and women (*y* axis) among 6,748 affected and 11,516 unaffected individuals of European ancestry. For each sex, the lowest-risk allele (C4-A(L)-A(L)) is used as a reference (odds ratio of 1.0). Shading of each point reflects the relative level of SLE risk (darker indicates greater risk) conferred by *C4A* and *C4B* copy numbers as in Fig. 2b. Error bars represent 95% confidence intervals around the effect size estimate for each sex. **b**, Schizophrenia risk (odds ratios) associated with the four most common C4 alleles in men (*x* axis) and women (*y* axis) among 28,799 affected and 35,986 unaffected individuals of European ancestry, aggregated by the Psychiatric Genomics Consortium[43]. For each sex, the lowest-risk allele (C4-B(S)) is used as a reference (odds ratio of 1.0). For visual comparison with **a**, shading of each allele reflects the relative level of SLE risk. Error bars represent 95% confidence intervals around the effect size estimate for each sex. **c**, Concentrations of C4 protein in CSF sampled from 340 adult men (blue) and 167 adult women (pink) as a function of age with locally estimated scatterplot smoothing (LOESS). Concentrations are normalized to the number of C4 gene copies in an individual's genome (a strong independent source of variance, Extended Data Fig. 7a) and shown on a $\log_{10}$ scale as a LOESS curve. Shaded regions represent 95% confidence intervals derived during LOESS. **d**, Levels of C3 protein in CSF from 179 adult men and 125 adult women as a function of age. Concentrations are shown on a $\log_{10}$ scale as a LOESS curve. Shaded regions represent 95% confidence intervals derived during LOESS.

onto targets in tissues. CSF levels of C3 protein were also on average 42% higher among men than women (meta-analysis $P = 7.5 \times 10^{-7}$, Fig. 3d).

The elevated concentrations of C3 and C4 proteins in CSF of men parallel earlier findings showing that, in plasma, C3 and C4 are also present at higher levels in men than women[8,9]. The large sample size ($n > 50{,}000$) of the plasma studies enables sex differences to be further analysed as a function of age. Both men and women undergo age-dependent elevation of C4 and C3 levels in plasma, but this occurs early in adulthood (20–30 years of age) in men and closer to menopause (40–50 years of age) in women, with the result that male–female differences in complement protein levels are observed primarily during the reproductive years (20–50 years of age)[8,9]. We replicated these findings using measurements of C3 and gene copy number-corrected C4 protein in plasma from adults, finding (as in the earlier plasma studies[8,9] and in CSF; Fig. 3c, d) that these differences are most pronounced during the reproductively active years of adulthood (20–50 years of age) (Extended Data Fig. 7b–d). We also observed that patients with Sjögren's syndrome have lower C4 serum levels than unaffected individuals ($P < 1\text{x}10^{-20}$, Extended Data Fig. 7e) even after correcting for C4 gene copy number ($P < 1\text{x}10^{-8}$, Extended Data Fig. 7f), suggesting that hypocomplementaemia in Sjögren's syndrome is not simply due to C4 genetics but also reflects disease effects on background complement levels, for example, owing to complement consumption. The ages of pronounced sex difference in complement levels correspond with the ages at which men and women differ in disease incidence: in schizophrenia, men outnumber women among cases incident in early adulthood, but not among cases incident after 40 years of age[2]; in SLE, women greatly outnumber men among cases incident during the child-bearing years, but not among cases incident after 50 years of age or during childhood[36]; in Sjögren's syndrome, the high relative vulnerability of women declines in magnitude after 50 years of age[37].

Our results indicate that the MHC genomic region shapes vulnerability in lupus and Sjögren's syndrome—two of the three most common rheumatic autoimmune diseases—in a very different way than in type I diabetes, rheumatoid arthritis and coeliac disease. In those diseases, precise interactions between HLA protein variants and specific autoantigens determine risk[13,14]. In SLE and Sjögren's syndrome, however, the genetic variation implicated here points instead to the continuous, chronic interaction of the immune system with a large number of potential autoantigens. Because complement facilitates the rapid clearance of debris from dead and injured cells, increased levels of C4 protein probably attenuate interactions between the adaptive immune system and ribonuclear self-antigens at sites of cell injury, pre-empting the development of autoimmunity. The additional C4-independent genetic risk effect described here (associated with rs2105898) may also affect autoimmunity broadly, rather than in an antigen-specific manner, by regulating expression of many HLA class II genes (including *DRB1*, *DQA1* and *DQB1*). Mouse models of SLE indicate that once tolerance is broken for one self-antigen, autoreactive germinal centres generate B cells targeting other self-antigens[38]; such 'epitope spreading' could lead to autoreactivity against many related autoantigens, regardless of which antigen(s) are involved in the earliest interactions with immune cells. Further supporting such a model, higher copy number of C4 is associated with lower risk of AQP4-IgG-seropositive neuromyelitis optica[39], in which seropositive patients have increased incidence of other non-organ-specific autoantibodies such as those seen in SLE and Sjögren's syndrome[40]. B cells also express the complement receptors CR1 and CR2[41], providing an additional candidate mechanism for regulation by C4 and C3.

We note that the role of complement proteins in preventing the emergence of autoimmunity may be very different than their (potentially disease-exacerbating) role once autoimmunity has been established. Also, our genetic findings address the development of SLE and Sjögren's syndrome rather than complications that arise in any specific organ. A few per cent of patients with SLE develop neurological complications that can include psychosis[42]; although psychosis is also a symptom of schizophrenia, neurological complications of SLE do not resemble schizophrenia more broadly, and probably have a different aetiology.

The same C4 alleles that increase vulnerability to schizophrenia appeared to protect strongly against SLE and Sjögren's syndrome. This pleiotropy will need to be considered in efforts to engage the complement system therapeutically. The complement system contributed to these pleiotropic effects more strongly in men than in women. Moreover, though the natural allelic series at C4 enabled human-genetic analysis to establish dose–risk relationships for C4 in men and women, sexual dimorphism in the levels of complement protein also included complement component 3 (C3). Why and how this sexual dimorphism in the complement system has evolved in

humans poses interesting questions for immune and evolutionary biology.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2277-x.

1. Ngo, S. T., Steyn, F. J. & McCombe, P. A. Gender differences in autoimmune disease. *Front. Neuroendocrinol.* **35**, 347–369 (2014).
2. Abel, K. M., Drake, R. & Goldstein, J. M. Sex differences in schizophrenia. *Int. Rev. Psychiatry* **22**, 417–428 (2010).
3. Langefeld, C. D. et al. Transancestral mapping and genetic load in systemic lupus erythematosus. *Nat. Commun.* **8**, 16021 (2017).
4. Rioux, J. D. et al. Mapping of multiple susceptibility variants within the MHC region for 7 immune-mediated diseases. *Proc. Natl Acad. Sci. USA* **106**, 18680–18685 (2009).
5. Hanscombe, K. B. et al. Genetic fine mapping of systemic lupus erythematosus MHC associations in Europeans and African Americans. *Hum. Mol. Genet.* **27**, 3813–3824 (2018).
6. Cruz-Tapias, P., Rojas-Villarraga, A., Maier-Moore, S. & Anaya, J. M. HLA and Sjögren's syndrome susceptibility. a meta-analysis of worldwide studies. *Autoimmun. Rev.* **11**, 281–287 (2012).
7. Sekar, A. et al. Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177–183 (2016).
8. Gaya da Costa, M. et al. Age and sex-associated changes of complement activity and complement levels in a healthy Caucasian population. *Front. Immunol.* **9**, 2664 (2018).
9. Ritchie, R. F. et al. Reference distributions for complement proteins C3 and C4: a practical, simple and clinically relevant approach in a large cohort. *J. Clin. Lab. Anal.* **18**, 1–8 (2004).
10. Lawrence, J. S., Martins, C. L. & Drake, G. L. A family survey of lupus erythematosus. 1. Heritability. *J. Rheumatol.* **14**, 913–921 (1987).
11. Lipsky, P. E. Systemic lupus erythematosus: an autoimmune disease of B cell hyperactivity. *Nat. Immunol.* **2**, 764–766 (2001).
12. Ippolito, A. et al. Autoantibodies in systemic lupus erythematosus: comparison of historical and current assessment of seropositivity. *Lupus* **20**, 250–255 (2011).
13. Lee, K. H., Wucherpfennig, K. W. & Wiley, D. C. Structure of a human insulin peptide–HLA–DQ8 complex and susceptibility to type 1 diabetes. *Nat. Immunol.* **2**, 501–507 (2001).
14. Raychaudhuri, S. et al. Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat. Genet.* **44**, 291–296 (2012).
15. Morris, D. L. et al. MHC associations with clinical and autoantibody manifestations in European SLE. *Genes Immun.* **15**, 210–217 (2014).
16. Bánlaki, Z., Doleschall, M., Rajczy, K., Fust, G. & Szilágyi, A. Fine-tuned characterization of RCCX copy number variants and their relationship with extended MHC haplotypes. *Genes Immun.* **13**, 530–535 (2012).
17. Isenman, D. E. & Young, J. R. The molecular basis for the difference in immune hemolysis activity of the Chido and Rodgers isotypes of human complement component C4. *J. Immunol.* **132**, 3019–3027 (1984).
18. Law, S. K., Dodds, A. W. & Porter, R. R. A comparison of the properties of two classes, C4A and C4B, of the human complement component C4. *EMBO J.* **3**, 1819–1823 (1984).
19. Birmingham, D. J. et al. The complex nature of serum C3 and C4 as biomarkers of lupus renal flare. *Lupus* **19**, 1272–1280 (2010).
20. Ross, S. C. & Densen, P. Complement deficiency states and infection: epidemiology, pathogenesis and consequences of neisserial and other infections in an immune deficiency. *Medicine* **63**, 243–273 (1984).
21. Wu, Y. L., Hauptmann, G., Viguier, M. & Yu, C. Y. Molecular basis of complete complement C4 deficiency in two North-African families with systemic lupus erythematosus. *Genes Immun.* **10**, 433–445 (2009).
22. International Consortium for Systemic Lupus Erythematosus. Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in *ITGAM, PXK, KIAA1542* and other loci. *Nat. Genet.* **40**, 204–210 (2008).
23. Yang, Y. et al. Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. *Am. J. Hum. Genet.* **80**, 1037–1054 (2007).
24. Jüptner, M. et al. Low copy numbers of complement C4 and homozygous deficiency of C4A may predispose to severe disease and earlier disease onset in patients with systemic lupus erythematosus. *Lupus* **27**, 600–609 (2018).
25. Boteva, L. et al. Genetically determined partial complement C4 deficiency states are not independent risk factors for SLE in UK and Spanish populations. *Am. J. Hum. Genet.* **90**, 445–456 (2012).
26. Pato, M. T. et al. The genomic psychiatry cohort: partners in discovery. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **162**, 306–312 (2013).
27. Sanders, S. J. et al. Whole genome sequencing in psychiatric disorders: the WGSPD consortium. *Nat. Neurosci.* **20**, 1661–1668 (2017).
28. Kuo, C. F. et al. Familial risk of Sjögren's syndrome and co-aggregation of autoimmune diseases in affected families: a nationwide population study. *Arthritis Rheumatol.* **67**, 1904–1912 (2015).
29. Fayyaz, A., Kurien, B. T. & Scofield, R. H. Autoantibodies in Sjögren's Syndrome. *Rheum. Dis. Clin. North Am.* **42**, 419–434 (2016).
30. Ramos-Casals, M. et al. Hypocomplementaemia as an immunological marker of morbidity and mortality in patients with primary Sjögren's syndrome. *Rheumatology* **44**, 89–94 (2005).
31. Chused, T. M., Kassan, S. S., Opelz, G., Moutsopoulos, H. M. & Terasaki, P. I. Sjögren's syndrome association with HLA-Dw3. *N. Engl. J. Med.* **296**, 895–897 (1977).
32. Taylor, K. E. et al. Genome-wide association analysis reveals genetic heterogeneity of Sjögren's syndrome according to ancestry. *Arthritis Rheumatol.* **69**, 1294–1305 (2017).
33. Khramtsova, E. A., Davis, L. K. & Stranger, B. E. The role of sex in the genomics of human complex traits. *Nat. Rev. Genet.* **20**, 173–190 (2019).
34. Hughes, T. et al. Analysis of autosomal genes reveals gene–sex interactions and higher total genetic risk in men with systemic lupus erythematosus. *Ann. Rheum. Dis.* **71**, 694–699 (2012).
35. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
36. Brinks, R. et al. Age-specific and sex-specific incidence of systemic lupus erythematosus: an estimate from cross-sectional claims data of 2.3 million people in the German statutory health insurance 2002. *Lupus Sci. Med.* **3**, e000181 (2016).
37. Kim, H. J. et al. Incidence, mortality, and causes of death in physician-diagnosed primary Sjögren's syndrome in Korea: A nationwide, population-based study. *Semin. Arthritis Rheum.* **47**, 222–227 (2017).
38. Degn, S. E. et al. Clonal evolution of autoreactive germinal centers. *Cell* **170**, 913–926 (2017).
39. Estrada, K. et al. A whole-genome sequence study identifies genetic risk factors for neuromyelitis optica. *Nat. Commun.* **9**, 1929 (2018).
40. Pittock, S. J. et al. Neuromyelitis optica and non organ-specific autoimmunity. *Arch. Neurol.* **65**, 78–83 (2008).
41. Erdei, A. et al. Expression and role of CR1 and CR2 on B and T lymphocytes under physiological and autoimmune conditions. *Mol. Immunol.* **46**, 2767–2773 (2009).
42. Unterman, A. et al. Neuropsychiatric syndromes in systemic lupus erythematosus: a meta-analysis. *Semin. Arthritis Rheum.* **41**, 1–11 (2011).
43. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).

Schizophrenia Working Group of the Psychiatric Genomics Consortium

Stephan Ripke[15,16], Benjamin M. Neale[15,16,17,18], Aiden Corvin[19], James T. R. Walters[20], Kai-How Farh[15], Peter A. Holmans[20,21], Phil Lee[15,16,18], Brendan Bulik-Sullivan[15,16], David A. Collier[22,23], Hailiang Huang[15,17], Tune H. Pers[17,24,25], Ingrid Agartz[26,27,28], Esben Agerbo[29,30,31], Margot Albus[32], Madeline Alexander[33], Farooq Amin[34,35], Silviu A. Bacanu[36], Martin Begemann[37], Richard A. Belliveau Jr[16], Judit Bene[38,39], Sarah E. Bergen[16,40], Elizabeth Bevilacqua[16], Tim B. Bigdeli[36], Donald W. Black[41], Richard Bruggeman[42], Nancy G. Buccola[43], Randy L. Buckner[44,45,46], William Byerley[47], Wiepke Cahn[48], Guiqing Cai[49,50], Murray J. Cairns[51,52,53], Dominique Campion[54], Rita M. Cantor[55], Vaughan J. Carr[51,56], Noa Carrera[20], Stanley V. Catts[51,57], Kimberly D. Chambert[16], Raymond C. K. Chan[58], Ronald Y. L. Chen[59], Eric Y. H. Chen[59,60], Wei Cheng[61], Eric F. C. Cheung[62], Siow Ann Chong[63], C. Robert Cloninger[64], David Cohen[65], Nadine Cohen[66], Paul Cormican[19], Nick Craddock[20,21], Benedicto Crespo-Facorro[67], James J. Crowley[68], David Curtis[69,70], Michael Davidson[71], Kenneth L. Davis[50], Franziska Degenhardt[72,73], Jurgen Del Favero[74], Lynn E. DeLisi[75,76], Ditte Demontis[31,77,78], Dimitris Dikeos[79], Timothy Dinan[80], Srdjan Djurovic[28,81], Gary Donohoe[19,82], Elodie Drapeau[50], Jubao Duan[83,84], Frank Dudbridge[85], Naser Durmishi[86], Peter Eichhammer[87], Johan Eriksson[88,89,90], Valentina Escott-Price[20], Laurent Essioux[91], Ayman H. Fanous[92,93,94,95], Martilias S. Farrell[68], Josef Frank[96], Lude Franke[97], Robert Freedman[98], Nelson B. Freimer[99], Marion Friedl[100], Joseph I. Friedman[50], Menachem Fromer[15,16,18,101], Giulio Genovese[16], Lyudmila Georgieva[20], Elliot S. Gershon[102], Ina Giegling[100,103], Paola Giusti-Rodríguez[68], Stephanie Godard[104], Jacqueline I. Goldstein[15,17], Vera Golimbet[105], Srihari Gopal[106], Jacob Gratten[107], Lieuwe de Haan[108], Marina Mitjans[37], Marian L. Hamshere[20], Mark Hansen[109], Thomas Hansen[31,110], Vahram Haroutunian[50,111,112], Annette M. Hartmann[100], Frans A. Henskens[51,113,114], Stefan Herms[72,73,115], Joel N. Hirschhorn[17,25,116], Per Hoffmann[72,73,115], Andrea Hofman[72,73], Mads V. Hollegaard[117], David M. Hougaard[117], Masashi Ikeda[118], Inge Joa[119], Antonio Julià[120], René S. Kahn[48], Luba Kalaydjieva[121], Sena Karachanak-Yankova[122], Juha Karjalainen[97], David Kavanagh[20], Matthew C. Keller[123], Brian J. Kelly[52], James L. Kennedy[124,125,126], Andrey Khrunin[127], Yunjung Kim[68], Janis Klovins[128], James A. Knowles[129], Bettina Konte[100], Vaidutis Kucinskas[130], Zita Ausrele Kucinskiene[130], Hana Kuzelova-Ptackova[131], Anna K. Kähler[40], Claudine Laurent[33,132], Jimmy Lee Chee Keong[63,133], S. Hong Lee[107], Sophie E. Legge[20], Bernard Lerer[134], Miaoxin Li[59,60,135], Tao Li[136], Kung-Yee Liang[137], Jeffrey Lieberman[138], Svetlana Limborska[127], Carmel M. Loughland[51,52], Jan Lubinski[139], Jouko Lönnqvist[140], Milan Macek Jr[131], Patrik K. E. Magnusson[40], Brion S. Maher[141], Wolfgang Maier[142], Jacques Mallet[143], Sara Marsal[120], Manuel Mattheisen[31,77,78,144], Morten Mattingsdal[28,145], Robert W. McCarley[75,76], Colm McDonald[146], Andrew M. McIntosh[147,148], Sandra Meier[96], Carin J. Meijer[108], Bela Melegh[38,39], Ingrid Melle[28,149], Raquelle I. Mesholam-Gately[75,150], Andres Metspalu[151], Patricia T. Michie[51,152], Lili Milani[151], Vihra Milanova[153], Younes Mokrab[22], Derek W. Morris[19,82], Ole Mors[31,77,154], Kieran C. Murphy[155], Robin M. Murray[156], Inez Myin-Germeys[157], Bertram Müller-Myhsok[158,159,160], Mari Nelis[151], Igor Nenadic[161], Deborah A. Nertney[162], Gerald Nestadt[163], Kristin K. Nicodemus[164], Liene Nikitina-Zake[128], Laura Nisenbaum[165], Annelie Nordin[166], Eadbhard O'Callaghan[167], Colm O'Dushlaine[16], F. Anthony O'Neill[168], Sang-Yun Oh[169], Ann Olincy[98], Line Olsen[31,110], Jim Van Os[157,170], Psychosis Endophenotypes International Consortium*, Christos Pantelis[51,171], George N. Papadimitriou[79], Agnes A. Steixner[37], Elena Parkhomenko[50], Michele T. Pato[129], Tiina Paunio[172,173], Milica Pejovic-Milovancevic[174], Diana O. Perkins[175], Olli Pietiläinen[173,176], Jonathan Pimm[70], Andrew J. Pocklington[20], John Powell[156], Alkes Price[17,177], Ann E. Pulver[163], Shaun M. Purcell[101], Digby Quested[178], Henrik B. Rasmussen[31,110], Abraham Reichenberg[50], Mark A. Reimers[179], Alexander L. Richards[20], Joshua L. Roffman[44,46], Panos Roussos[101,180], Douglas M. Ruderfer[20,101], Veikko Salomaa[90], Alan R. Sanders[83,84], Ulrich Schall[51,52], Christian R. Schubert[181], Thomas G. Schulze[96,182], Sibylle G. Schwab[183], Edward M. Scolnick[16], Rodney J. Scott[51,53,184], Larry J. Seidman[75,150], Jianxin Shi[185], Engilbert Sigurdsson[186], Teimuraz Silagadze[187], Jeremy M. Silverman[50,188], Kang Sim[63], Petr Slominsky[127], Jordan W. Smoller[16,18], Hon-Cheong So[59], Chris C. A. Spencer[189], Eli A. Stahl[17,101], Hreinn Stefansson[190], Stacy Steinberg[190], Elisabeth Stogmann[191], Richard E. Straub[192], Eric Strengman[48,193], Jana Strohmaier[96], T. Scott Stroup[138], Mythily Subramaniam[63], Jaana Suvisaari[140], Dragan M. Svrakic[64], Jin P. Szatkiewicz[68], Erik Söderman[26], Srinivas Thirumalai[194], Draga Toncheva[122], Paul A. Tooney[51,52,53], Sarah Tosato[195], Juha Veijola[196,197], John Waddington[198], Dermot Walsh[199], Dai Wang[106], Qiang Wang[136], Bradley T. Webb[36], Mark Weiser[71], Dieter B. Wildenauer[200], Nigel M. Williams[20], Stephanie Williams[68], Stephanie H. Witt[96], Aaron R. Wolen[179], Emily H. M. Wong[59], Brandon K. Wormley[36], Jing Qin Wu[51,53], Hualin Simon Xi[201], Clement C. Zai[124,125], Xuebin Zheng[202], Fritz Zimprich[191], Naomi R. Wray[107], Kari Stefansson[190], Peter M. Visscher[107], Wellcome Trust Case–Control Consortium 2*, Rolf Adolfsson[166], Ole A. Andreassen[28,149], Douglas H. R. Blackwood[148], Elvira Bramon[203], Joseph D. Buxbaum[49,50,111,204], Anders D. Børglum[31,77,78,154], Sven Cichon[72,73,115,205], Ariel Darvasi[206], Enrico Domenici[207], Hannelore Ehrenreich[37], Tõnu Esko[17,25,116,151], Pablo V. Gejman[83,84], Michael Gill[19], Hugh Gurling[70], Christina M. Hultman[40], Nakao Iwata[118], Assen V. Jablensky[51,200,208,209], Erik G. Jönsson[26,28], Kenneth S. Kendler[210], George Kirov[20], Jo Knight[124,125,126], Todd Lencz[211,212,213], Douglas F. Levinson[33], Qingqin S. Li[106], Jianjun Liu[202,214], Anil K. Malhotra[211,212,213], Steven A. McCarroll[16,116], Andrew McQuillin[70], Jennifer L. Moran[16], Preben B. Mortensen[29,30,31], Bryan J. Mowry[107,215], Markus M. Nöthen[72,73], Roel A. Ophoff[48,55,99], Michael J. Owen[20,21], Aarno Palotie[16,18,176], Carlos N. Pato[129], Tracey L. Petryshen[16,75,216], Danielle Posthuma[217,218,219], Marcella Rietschel[96], Brien P. Riley[210], Dan Rujescu[100,103], Pak C. Sham[59,60,135], Pamela Sklar[101,111,180], David St Clair[220], Daniel R. Weinberger[192,221], Jens R. Wendland[181], Thomas Werge[31,110,222], Mark J. Daly[15,16,17], Patrick F. Sullivan[40,68,175] & Michael C. O'Donovan[20,21]

Psychosis Endophenotype International Consortium

Maria J. Arranz[170,223], Steven Bakker[48], Stephan Bender[224,225], Elvira Bramon[170,226,227], David A. Collier[22,23], Benedicto Crespo-Facorro[228,229], Jeremy Hall[148], Conrad Iyegbe[170], Assen V. Jablensky[230], René S. Kahn[48], Luba Kalaydjieva[121,231], Stephen Lawrie[148], Cathryn M. Lewis[170], Kuang Lin[170], Don H. Linszen[232], Ignacio Mata[228,229], Andrew M. McIntosh[148], Robin M. Murray[156], Roel A. Ophoff[99], Jim Van Os[157,170], John Powell[170], Dan Rujescu[100,103], Muriel Walshe[170], Matthias Weisbrod[225] & Durk Wiersma[233]

Wellcome Trust Case–Control Consortium 2

Peter Donnelly[192,234], Ines Barroso[235], Jenefer M. Blackwell[236,237], Elvira Bramon[205], Matthew A. Brown[238], Juan P. Casas[239,240], Aiden Corvin[19], Panos Deloukas[235], Audrey Duncanson[241], Janusz Jankowski[242], Hugh S. Markus[243], Christopher G. Mathew[244], Colin N. A. Palmer[245], Robert Plomin[23], Anna Rautanen[192], Stephen J. Sawcer[246], Richard C. Trembath[244], Ananth C. Viswanathan[247,248], Nicholas W. Wood[249], Chris C. A. Spencer[192], Gavin Band[192], Céline Bellenguez[192], Peter Donnelly[192,234], Colin Freeman[192], Eleni Giannoulatou[192], Garrett Hellenthal[192], Richard Pearson[192], Matti Pirinen[192], Amy Strange[192], Zhan Su[192], Damjan Vukcevic[192], Cordelia Langford[235], Ines Barroso[235], Hannah Blackburn[235], Suzannah J. Bumpstead[235], Panos Deloukas[235], Serge Dronov[235], Sarah Edkins[235], Matthew Gillman[235], Emma Gray[235], Rhian Gwilliam[235], Naomi Hammond[235], Sarah E. Hunt[235], Alagurevathi Jayakumar[235], Jennifer Liddle[235], Owen T. McCann[235], Simon C. Potter[235], Radhi Ravindrarajah[235], Michelle Ricketts[235], Avazeh Tashakkori-Ghanbaria[235], Matthew Waller[235], Paul Weston[235], Pamela Whittaker[235], Sara Widaa[235], Christopher G. Mathew[244], Jenefer M. Blackwell[236,237], Matthew A. Brown[238], Aiden Corvin[19], Mark I. McCarthy[250] & Chris C. A. Spencer[192]

[15]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. [16]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [17]Medical and Population Genetics Program, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [18]Psychiatric and Neurodevelopmental Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. [19]Neuropsychiatric Genetics Research Group, Department of Psychiatry, Trinity College Dublin, Dublin, Ireland. [20]MRC Centre for Neuropsychiatric Genetics and Genomics, Institute of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, UK. [21]National Centre for Mental Health, Cardiff University, Cardiff, UK. [22]Eli Lilly, Windlesham, UK. [23]Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, King's College London, London, UK. [24]Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark. [25]Division of Endocrinology and Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, MA, USA. [26]Department of Clinical Neuroscience, Psychiatry Section, Karolinska Institutet, Stockholm, Sweden. [27]Department of Psychiatry, Diakonhjemmet Hospital, Oslo, Norway. [28]NORMENT, KG Jebsen Centre for Psychosis Research, Institute of Clinical Medicine, University of Oslo, Oslo, Norway. [29]Centre for Integrative Register-based Research, CIRRAU, Aarhus University, Aarhus, Denmark. [30]National Centre for Register-based Research, Aarhus University, Aarhus, Denmark. [31]The Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH, Aarhus, Denmark. [32]State Mental Hospital, Haar, Germany. [33]Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, USA. [34]Department of Psychiatry and Behavioral Sciences, Atlanta Veterans Affairs Medical Center, Atlanta, GA, USA. [35]Department of Psychiatry and Behavioral Sciences, Emory University, Atlanta, GA, USA. [36]Virginia Institute for Psychiatric and Behavioral Genetics, Department of Psychiatry, Virginia Commonwealth University, Richmond, VA, USA. [37]Clinical Neuroscience, Max Planck Institute of Experimental Medicine, Göttingen, Germany. [38]Department of Medical Genetics, University of Pécs, Pécs, Hungary. [39]Szentagothai Research Center, University of Pécs, Pécs, Hungary. [40]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [41]Department of Psychiatry, University of Iowa Carver College of Medicine, Iowa City, IA, USA. [42]Department of Psychiatry, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands. [43]School of Nursing, Louisiana State University Health Sciences Center, New Orleans, LA, USA. [44]Athinoula A. Martinos Center, Massachusetts General Hospital, Boston, MA, USA. [45]Center for Brain Science, Harvard University, Cambridge, MA, USA. [46]Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA. [47]Department of Psychiatry, University of California at San Francisco, San Francisco, CA, USA. [48]Department of Psychiatry, Rudolf Magnus Institute of Neuroscience, University Medical Center Utrecht, Utrecht, The Netherlands. [49]Department of Human Genetics, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [50]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [51]Schizophrenia Research Institute, Sydney, New South Wales, Australia. [52]Priority Centre for Translational Neuroscience and Mental Health, University of Newcastle, Newcastle, New South Wales, Australia. [53]School of Biomedical Sciences and Pharmacy, University of Newcastle, Callaghan, New South Wales, Australia. [54]Centre Hospitalier du Rouvray and INSERM U1079 Faculty of Medicine, Rouen, France. [55]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA, USA. [56]School of Psychiatry, University of New South Wales, Sydney, New South Wales, Australia. [57]Royal Brisbane and Women's Hospital, University of Queensland, Brisbane, Australia. [58]Institute of Psychology, Chinese Academy of Science, Beijing, China. [59]Department of Psychiatry, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China. [60]State Key Laboratory for Brain and Cognitive Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China. [61]Department of

Computer Science, University of North Carolina, Chapel Hill, NC, USA. [62]Castle Peak Hospital, Hong Kong, China. [63]Institute of Mental Health, Singapore, Singapore. [64]Department of Psychiatry, Washington University, St Louis, MO, USA. [65]Department of Child and Adolescent Psychiatry, Assistance Publique Hopitaux de Paris, Pierre and Marie Curie Faculty of Medicine and Institute for Intelligent Systems and Robotics, Paris, France. [66]Blue Note Biosciences, Princeton, NJ, USA. [67]University Hospital Marqués de Valdecilla, Instituto de Formación e Investigación Marqués de Valdecilla, University of Cantabria, Santander, Spain. [68]Department of Genetics, University of North Carolina, Chapel Hill, NC, USA. [69]Department of Psychological Medicine, Queen Mary University of London, London, UK. [70]Molecular Psychiatry Laboratory, Division of Psychiatry, University College London, London, UK. [71]Sheba Medical Center, Tel Hashomer, Israel. [72]Department of Genomics, Life and Brain Center, Bonn, Germany. [73]Institute of Human Genetics, University of Bonn, Bonn, Germany. [74]Applied Molecular Genomics Unit, VIB Department of Molecular Genetics, University of Antwerp, Antwerp, Belgium. [75]Department of Psychiatry, Harvard Medical School, Boston, MA, USA. [76]VA Boston Health Care System, Brockton, MA, USA. [77]Centre for Integrative Sequencing, iSEQ, Aarhus University, Aarhus, Denmark. [78]Department of Biomedicine, Aarhus University, Aarhus, Denmark. [79]First Department of Psychiatry, University of Athens Medical School, Athens, Greece. [80]Department of Psychiatry, University College Cork, Cork, Ireland. [81]Department of Medical Genetics, Oslo University Hospital, Oslo, Norway. [82]Cognitive Genetics and Therapy Group, School of Psychology and Discipline of Biochemistry, National University of Ireland Galway, Galway, Ireland. [83]Department of Psychiatry and Behavioral Neuroscience, University of Chicago, Chicago, IL, USA. [84]Department of Psychiatry and Behavioral Sciences, NorthShore University HealthSystem, Evanston, IL, USA. [85]Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK. [86]Department of Child and Adolescent Psychiatry, University Clinic of Psychiatry, Skopje, Republic of Macedonia. [87]Department of Psychiatry, University of Regensburg, Regensburg, Germany. [88]Department of General Practice, Helsinki University Central Hospital, University of Helsinki, Helsinki, Finland. [89]Folkhälsan Research Center, Helsinki, Finland, Biomedicum Helsinki 1, Helsinki, Finland. [90]National Institute for Health and Welfare, Helsinki, Finland. [91]Translational Technologies and Bioinformatics, Pharma Research and Early Development, F. Hoffman-La Roche, Switzerland. [92]Department of Psychiatry, Georgetown University School of Medicine, Washington, DC, USA. [93]Department of Psychiatry, Keck School of Medicine of the University of Southern California, Los Angeles, CA, USA. [94]Department of Psychiatry, Virginia Commonwealth University School of Medicine, Richmond, VA, USA. [95]Mental Health Service Line, Washington VA Medical Center, Washington, DC, USA. [96]Department of Genetic Epidemiology in Psychiatry, Central Institute of Mental Health, Medical Faculty Mannheim, University of Heidelberg, Heidelberg, Germany. [97]Department of Genetics, University Medical Centre Groningen, University of Groningen, Groningen, The Netherlands. [98]Department of Psychiatry, University of Colorado Denver, Aurora, CO, USA. [99]Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, CA, USA. [100]Department of Psychiatry, University of Halle, Halle, Germany. [101]Division of Psychiatric Genomics, Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [102]Departments of Psychiatry and Human Genetics, University of Chicago, Chicago, IL, USA. [103]Department of Psychiatry, University of Munich, Munich, Germany. [104]Departments of Psychiatry and Human and Molecular Genetics, INSERM, Institut de Myologie, Hôpital de la Pitiè-Salpêtrière, Paris, France. [105]Mental Health Research Centre, Russian Academy of Medical Sciences, Moscow, Russia. [106]Neuroscience Therapeutic Area, Janssen Research and Development, Raritan, NJ, USA. [107]Queensland Brain Institute, The University of Queensland, Brisbane, Queensland, Australia. [108]Department of Psychiatry, Academic Medical Centre University of Amsterdam, Amsterdam, The Netherlands. [109]Illumina, La Jolla, CA, USA. [110]Institute of Biological Psychiatry, Mental Health Centre Sct. Hans, Mental Health Services Copenhagen, Copenhagen, Denmark. [111]Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [112]J. J. Peters VA Medical Center, Bronx, New York, NY, USA. [113]Priority Research Centre for Health Behaviour, University of Newcastle, Newcastle, New South Wales, Australia. [114]School of Electrical Engineering and Computer Science, University of Newcastle, Newcastle, New South Wales, Australia. [115]Division of Medical Genetics, Department of Biomedicine, University of Basel, Basel, Switzerland. [116]Department of Genetics, Harvard Medical School, Boston, MA, USA. [117]Section of Neonatal Screening and Hormones, Department of Clinical Biochemistry, Immunology and Genetics, Statens Serum Institut, Copenhagen, Denmark. [118]Department of Psychiatry, Fujita Health University School of Medicine, Toyoake, Japan. [119]Regional Centre for Clinical Research in Psychosis, Department of Psychiatry, Stavanger University Hospital, Stavanger, Norway. [120]Rheumatology Research Group, Vall d'Hebron Research Institute, Barcelona, Spain. [121]Centre for Medical Research, The University of Western Australia, Perth, Western Australia, Australia. [122]Department of Medical Genetics, Medical University, Sofia, Bulgaria. [123]Department of Psychology, University of Colorado Boulder, Boulder, CO, USA. [124]Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, Ontario, Canada. [125]Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada. [126]Institute of Medical Science, University of Toronto, Toronto, Ontario, Canada. [127]Institute of Molecular Genetics, Russian Academy of Sciences, Moscow, Russia. [128]Latvian Biomedical Research and Study Centre, Riga, Latvia. [129]Department of Psychiatry and Zilkha Neurogenetics Institute, Keck School of Medicine at University of Southern California, Los Angeles, CA, USA. [130]Faculty of Medicine, Vilnius University, Vilnius, Lithuania. [131]Department of Biology and Medical Genetics, 2nd Faculty of Medicine and University Hospital Motol, Prague, Czech Republic. [132]Department of Child and Adolescent Psychiatry, Pierre and Marie Curie Faculty of Medicine, Paris, France. [133]Duke-NUS Graduate Medical School, Singapore, Singapore. [134]Department of Psychiatry, Hadassah-Hebrew University Medical Center, Jerusalem, Israel. [135]Centre for Genomic Sciences, The University of Hong Kong, Hong Kong, China. [136]Mental Health Centre and Psychiatric Laboratory, West China Hospital, Sichuan University, Chengdu, China. [137]Department of Biostatistics, Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD, USA. [138]Department of Psychiatry, Columbia University, New York, NY, USA. [139]Department of Genetics and Pathology, International Hereditary Cancer Center, Pomeranian Medical University in Szczecin, Szczecin, Poland. [140]Department of Mental Health and Substance Abuse Services, National Institute for Health and Welfare, Helsinki, Finland. [141]Department of Mental Health, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA. [142]Department of Psychiatry, University of Bonn, Bonn, Germany. [143]Centre National de la Recherche Scientifique, Laboratoire de Génétique Moléculaire de la Neurotransmission et des Processus Neurodégénératifs, Hôpital de la Pitié Salpêtrière, Paris, France. [144]Department of Genomics Mathematics, University of Bonn, Bonn, Germany. [145]Research Unit, Sørlandet Hospital, Kristiansand, Norway. [146]Department of Psychiatry, National University of Ireland Galway, Galway, Ireland. [147]Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, UK. [148]Division of Psychiatry, University of Edinburgh, Edinburgh, UK. [149]Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway. [150]Massachusetts Mental Health Center Public Psychiatry Division of the Beth Israel Deaconess Medical Center, Boston, MA, USA. [151]Estonian Genome Center, University of Tartu, Tartu, Estonia. [152]School of Psychology, University of Newcastle, Newcastle, New South Wales, Australia. [153]First Psychiatric Clinic, Medical University, Sofia, Bulgaria. [154]Department P, Aarhus University Hospital, Risskov, Denmark. [155]Department of Psychiatry, Royal College of Surgeons in Ireland, Dublin, Ireland. [156]King's College London, London, UK. [157]South Limburg Mental Health Research and Teaching Network, EURON, Maastricht University Medical Centre, Maastricht, The Netherlands. [158]Institute of Translational Medicine, University of Liverpool, Liverpool, UK. [159]Max Planck Institute of Psychiatry, Munich, Germany. [160]Munich Cluster for Systems Neurology (SyNergy), Munich, Germany. [161]Department of Psychiatry and Psychotherapy, Jena University Hospital, Jena, Germany. [162]Department of Psychiatry, Queensland Brain Institute and Queensland Centre for Mental Health Research, University of Queensland, Brisbane, Queensland, Australia. [163]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [164]Department of Psychiatry, Trinity College Dublin, Dublin, Ireland. [165]Eli Lilly, Lilly Corporate Center, Indianapolis, IN, USA. [166]Department of Clinical Sciences, Psychiatry, Umeå University, Umeå, Sweden. [167]DETECT Early Intervention Service for Psychosis, Blackrock, Ireland. [168]Centre for Public Health, Institute of Clinical Sciences, Queen's University Belfast, Belfast, UK. [169]Lawrence Berkeley National Laboratory, University of California at Berkeley, Berkeley, CA, USA. [170]Institute of Psychiatry, King's College London, London, UK. [171]Melbourne Neuropsychiatry Centre, University of Melbourne and Melbourne Health, Melbourne, Victoria, Australia. [172]Department of Psychiatry, University of Helsinki, Helsinki, Finland. [173]Public Health Genomics Unit, National Institute for Health and Welfare, Helsinki, Finland. [174]Medical Faculty, University of Belgrade, Belgrade, Serbia. [175]Department of Psychiatry, University of North Carolina, Chapel Hill, NC, USA. [176]Institute for Molecular Medicine Finland, FIMM, University of Helsinki, Helsinki, Finland. [177]Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA. [178]Department of Psychiatry, University of Oxford, Oxford, UK. [179]Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA, USA. [180]Institute for Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [181]PharmaTherapeutics Clinical Research, Pfizer Worldwide Research and Development, Cambridge, MA, USA. [182]Department of Psychiatry and Psychotherapy, University of Gottingen, Göttingen, Germany. [183]Psychiatry and Psychotherapy Clinic, University of Erlangen, Erlangen, Germany. [184]Hunter New England Health Service, Newcastle, New South Wales, Australia. [185]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA. [186]University of Iceland, Landspitali, National University Hospital, Reykjavik, Iceland. [187]Department of Psychiatry and Drug Addiction, Tbilisi State Medical University (TSMU), Tbilisi, Georgia. [188]Research and Development, Bronx Veterans Affairs Medical Center, New York, NY, USA. [189]Wellcome Trust Centre for Human Genetics, Oxford, UK. [190]deCODE Genetics, Reykjavik, Iceland. [191]Department of Clinical Neurology, Medical University of Vienna, Vienna, Austria. [192]Lieber Institute for Brain Development, Baltimore, MD, USA. [193]Department of Medical Genetics, University Medical Centre Utrecht, Utrecht, The Netherlands. [194]Berkshire Healthcare NHS Foundation Trust, Bracknell, UK. [195]Section of Psychiatry, University of Verona, Verona, Italy. [196]Department of Psychiatry, University of Oulu, Oulu, Finland. [197]University Hospital of Oulu, Oulu, Finland. [198]Molecular and Cellular Therapeutics, Royal College of Surgeons in Ireland, Dublin, Ireland. [199]Health Research Board, Dublin, Ireland. [200]School of Psychiatry and Clinical Neurosciences, The University of Western Australia, Perth, Western Australia, Australia. [201]Computational Sciences CoE, Pfizer Worldwide Research and Development, Cambridge, MA, USA. [202]Human Genetics, Genome Institute of Singapore, A*STAR, Singapore, Singapore. [203]University College London, London, UK. [204]Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [205]Institute of Neuroscience and Medicine (INM-1), Research Center Juelich, Juelich, Germany. [206]Department of Genetics, The Hebrew University of Jerusalem, Jerusalem, Israel. [207]Neuroscience Discovery and Translational Area, Pharma Research and Early Development, F. Hoffman-La Roche, Basel, Switzerland. [208]The Perkins Institute for Medical Research, The University of Western Australia, Perth, Western Australia, Australia. [209]Centre for Clinical Research in Neuropsychiatry, School of Psychiatry and Clinical Neurosciences, The University of Western Australia, Perth, Western Australia, Australia. [210]Virginia Institute for Psychiatric and Behavioral Genetics, Departments of Psychiatry and Human and Molecular Genetics, Virginia Commonwealth

University, Richmond, VA, USA. [211]The Feinstein Institute for Medical Research, Manhasset, NY, USA. [212]The Hofstra NS-LIJ School of Medicine, Hempstead, NY, USA. [213]The Zucker Hillside Hospital, Glen Oaks, NY, USA. [214]Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore. [215]Queensland Centre for Mental Health Research, University of Queensland, Brisbane, Queensland, Australia. [216]Center for Human Genetic Research and Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA. [217]Department of Child and Adolescent Psychiatry, Erasmus University Medical Centre, Rotterdam, The Netherlands. [218]Department of Complex Trait Genetics, Neuroscience Campus Amsterdam, VU University Medical Center Amsterdam, Amsterdam, The Netherlands. [219]Department of Functional Genomics, Center for Neurogenomics and Cognitive Research, Neuroscience Campus Amsterdam, VU University, Amsterdam, The Netherlands. [220]University of Aberdeen, Institute of Medical Sciences, Aberdeen, UK. [221]Departments of Psychiatry, Neurology, Neuroscience and Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MD, USA. [222]Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark. [223]Fundació de Docència i Recerca Mútua de Terrassa, Universitat de Barcelona, Barcelona, Spain. [224]Child and Adolescent Psychiatry, University of Technology Dresden, Dresden, Germany. [225]Section for Experimental Psychopathology, General Psychiatry, Heidelberg, Germany. [226]Institute of Cognitive Neuroscience, University College London, London, UK. [227]Mental Health Sciences Unit, University College London, London, UK. [228]Centro Investigación Biomédica en Red Salud Mental, Madrid, Spain. [229]University Hospital Marqués de Valdecilla, Instituto de Formación e Investigación Marqués de Valdecilla, University of Cantabria, Santander, Spain. [230]Centre for Clinical Research in Neuropsychiatry, The University of Western Australia, Perth, Western Australia, Australia. [231]Western Australian Institute for Medical Research, The University of Western Australia, Perth, Western Australia, Australia. [232]Department of Psychiatry, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands. [233]Department of Psychiatry, University Medical Center Groningen, University of Groningen, The Netherlands. [234]Department of Statistics, University of Oxford, Oxford, UK. [235]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, UK. [236]Cambridge Institute for Medical Research, University of Cambridge School of Clinical Medicine, Cambridge, UK. [237]Telethon Institute for Child Health Research, Centre for Child Health Research, University of Western Australia, Subiaco, Western Australia, Australia. [238]Diamantina Institute of Cancer, Immunology and Metabolic Medicine, Princess Alexandra Hospital, University of Queensland, Brisbane, Queensland, Australia. [239]Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK. [240]Department of Epidemiology and Public Health, University College London, London, UK. [241]Molecular and Physiological Sciences, The Wellcome Trust, London, UK. [242]Peninsula School of Medicine and Dentistry, Plymouth University, Plymouth, UK. [243]Clinical Neurosciences, St George's University of London, London, UK. [244]Department of Medical and Molecular Genetics, School of Medicine, Guy's Hospital, King's College London, London, UK. [245]Biomedical Research Centre, Ninewells Hospital and Medical School, Dundee, UK. [246]Department of Clinical Neurosciences, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK. [247]Institute of Ophthalmology, University College London, London, UK. [248]National Institute for Health Research, Biomedical Research Centre at Moorfields Eye Hospital, National Health Service Foundation Trust, London, UK. [249]Department of Molecular Neuroscience, Institute of Neurology, London, UK. [250]Oxford Centre for Diabetes, Endocrinology and Metabolism, Churchill Hospital, Oxford, UK.

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Creation of a C4 reference panel from WGS data

We constructed a reference panel for imputation of C4 structural haplotypes using WGS data for 1,265 individuals from the Genomic Psychiatry Cohort[26]. The reference panel included individuals of diverse ancestry, including 765 Europeans, 250 African Americans and 250 people of reported Latino ancestry.

We estimated the diploid C4 copy number, and estimated separately the diploid copy number of the contained human endogenous retrovirus (HERV) sequence, using Genome STRiP[44]. In brief, Genome STRiP carefully calibrates measurements of read depth across specific genomic segments of interest by estimating and normalizing away sample-specific technical effects such as the effect of GC content on read depth (estimated from the genome-wide data). To measure total C4 gene copy number, we analysed the segments 6:31948358–31981050 and 6:31981096–32013904 (hg19), masking the intronic HERV segments that distinguish short (S) from long (L) C4 gene isotypes. To measure copy number of the HERV sequence, we analysed segments 6:31952461–31958829 and 6:31985199–31991567 (hg19). Across the 1,265 individuals, the resultant locus-specific copy-number estimates exhibited a strongly multi-modal distribution (Extended Data Fig. 1a) from which individuals' total C4 copy numbers could be readily inferred.

We then estimated the numbers of *C4A* and *C4B* genes in each individual genome. To do this, we extracted reads mapping to the paralogous sequence variants that distinguish *C4A* from *C4B* (hg19 coordinates 6:31963859–31963876 and 6:31996597–31996614) in each individual, combining reads across the two sites. We included only reads that aligned to one of these segments in its entirety. We then counted the number of reads matching the canonical active site sequences for *C4A* (CCC TGT CCA GTG TTA GAC) and *C4B* (CTC TCT CCA GTG ATA CAT). We combined these counts with the likelihood estimates of diploid C4 copy number (from Genome STRiP) to determine the maximum likelihood combination of *C4A* and *C4B* in each individual (Extended Data Fig. 1b). We estimated the genotype quality of the *C4A* and *C4B* estimate from the likelihood ratio between the most likely and second most likely combinations.

To phase the C4 copy number measurements into haplotypes, we first used the GenerateHaploidCNVGenotypes utility in Genome STRiP to estimate haplotype-specific copy-number likelihoods for C4 (total C4 gene copy number), *C4A*, *C4B* and HERV using the diploid likelihoods from the prior step as input. Default parameters for GenerateHaploidCNVGenotypes were used, plus -genotypeLikelihoodThreshold 0.0001. The output was then processed by the GenerateCNVHaplotypes utility in Genome STRiP to combine the multiple estimates into likelihood estimates for a set of unified structural alleles. GenerateCNVHaplotypes was run with default parameters, plus -defaultLogLikelihood -50, -unknownHaplotypeLikelihood -50, and -sampleHaplotypePriorLikelihood 2.0. The resultant VCF output was phased using Beagle 4.1 (beagle_4.1_27Jul16.86a) in two steps: first, performing genotype refinement from the genotype likelihoods using the Beagle gtgl = and maxlr = 1000000 parameters, and then running Beagle again on the output file using gt = to complete the phasing.

Our previous work suggested that several C4 structures segregate on multiple haplotypes, and probably arose by recurrent mutation on different haplotype backgrounds[7]. The GenerateCNVHaplotypes utility requires as input an enumerated set of structural alleles to assign to the samples in the reference cohort, including any structurally equivalent alleles, with distinct labels to mark them as independent, plus a list of samples to assign (with high likelihood) to specific labelled input alleles to disambiguate among these recurrent alleles. The selection of the set of structural alleles to be modelled, along with the labelling strategy, is important to our methodology and the performance of the reference panel. In the reference panel, each input allele represents a specific copy number structure and optionally includes a label that differentiates the allele from other independent alleles with equivalent structure. We use the notation <H_n_n_n_n_L> to identify each allele, where the four integers following the H are, respectively, the (redundant) haploid count of the total number of C4 copies, *C4A* copies, *C4B* copies and HERV copies on the haplotype. For example, <H_2_1_1_1> was used to represent the 'AL-BS' haplotype. The optional final label L is used to distinguish potentially recurrent haplotypes with otherwise equivalent structures (under the model) that should be treated as independent alleles for phasing and imputation.

To build the reference panel, we experimentally evaluated a large number of potential sets of structural alleles and methods for assigning labels to potentially recurrent alleles. For each evaluation, we built a reference panel using the 1,265 reference samples, and then evaluated the performance of the panel via cross-validation, leaving out 10 different samples in each trial (5 samples in the last trial) and imputing the missing samples from the remaining samples in the panel. The imputed results for all 1,265 samples were then compared to the original diploid copy number estimates to evaluate the performance of each candidate reference panel (Extended Data Table 1).

Using this procedure, we selected a final panel for downstream analysis that used a set of 29 structural alleles representing 16 distinct allelic structures (as listed in the reference panel VCF file). Each allele contained from one to three copies of C4. Three allelic structures (AL-BS, AL-BL and AL-AL) were represented as a set of independently labelled alleles with 9, 3 and 4 labels, respectively.

To identify the number of labels to use on the different alleles and the samples to 'seed' the alleles, we generated spider plots of the C4 locus based on initial phasing experiments run without labelled alleles, and then clustered the resulting haplotypes in two dimensions based on the *y*-coordinate distance between the haplotypes on the left and right sides of the spider plot. Clustering was based on visualizing the clusters (Extended Data Fig. 1c) and then manually choosing both the number of clusters (labels) to assign and a set of confidently assigned haplotypes to use to seed the clusters in GenerateCNVHaplotypes. This procedure was iterated multiple times using cross-validation, as described above, to evaluate the imputation performance of each candidate labelling strategy.

Within the dataset used to build the reference panel, there is evidence for individuals carrying seven or more diploid copies of C4, which implies the existence of (rare) alleles with four or more copies of C4. In our experiments, attempting to add additional haplotypes to model these rare four-copy alleles reduced overall imputation performance. Consequently, we conducted all downstream analyses using a reference panel that models only alleles with up to three copies of C4. In the future, larger reference panels might benefit from modelling these rare four-copy alleles.

The reference panel will be available in dbGaP (accession number pending) with broad permission for research use.

Genetic data for SLE

For analysis of SLE, collection and genotyping of the European-ancestry cohort (6,748 cases, 11,516 controls, genotyped by ImmunoChip) as previously described[3]. Collection and genotyping of the African American cohort (1,494 cases, 5,908 controls, genotyped by OmniExpress) as previously described[5].

Genetic data for Sjögren's syndrome

For analysis of Sjögren's syndrome, collection and genotyping of the European-ancestry cohort (673 cases, 1,153 controls, genotyped by Omni2.5) as previously described[32] and available in dbGaP under study accession number phs000672.v1.p1.

Genetic data for schizophrenia

The schizophrenia analysis made use of genotype data from 40 cohorts of European ancestry (28,799 cases, 35,986 controls) made available by the Psychiatric Genetics Consortium (PGC) as previously described[43]. Genotyping chips used for each cohort are listed in supplementary table 3 of that study.

Imputation of C4 alleles

The reference haplotypes described above were used to extend the SLE, Sjögren's syndrome or schizophrenia cohort SNP genotypes by imputation. SNP data in VCF format were used as input for Beagle v.4.1[45,46] for imputation of C4 as a multi-allelic variant. Within the Beagle pipeline, the reference panel was first converted to bref format. From the cohort SNP genotypes, we used only those SNPs from the MHC region (chr6:24–34 Mb on hg19) that were also in the haplotype reference panel. We used the conform-gt tool to perform strand-flipping and filtering of specific SNPs for which strand remained ambiguous. Beagle was run using default parameters with two key exceptions: we used the GRCh37 PLINK recombination map, and we set the output to include genotype probability (that is, GP field in VCF) for correct downstream probabilistic estimation of *C4A* and *C4B* joint dosages.

Imputation of HLA alleles

For HLA allele imputation, sample genotypes were used as input for the R package HIBAG[47]. For both European ancestry and African American cohorts, publicly available multi-ethnic reference panels generated for the most appropriate genotyping chip (that is, Immunochip for European ancestry SLE cohort, Omni 2.5 for the European ancestry Sjögren's syndrome cohort, and OmniExpress for African American SLE cohort) were used[48]. Default parameters were used for all settings. All class I and class II HLA genes were imputed. Output haplotype posterior probabilities were summed per allele to yield diploid dosages for each individual.

Associating single and joint C4 structural allele dosages to SLE and Sjögren's syndrome in European ancestry individuals

The analysis described above yields dosage estimates for each of the common C4 structural haplotypes (for example, AL-BS or AL-AL) for each genome in each cohort. In addition to performing association analysis on these structures (Fig. 1b), we also performed association analysis on the dosages of each underlying C4 gene isotype (that is, *C4A*, *C4B*, *C4L* and *C4S*). These dosages were computed from the allelic dosage (DS) field of the imputation output VCF simply by multiplying the dosage of a C4 structural haplotype by the number of copies of each C4 isotype that haplotype contains (for example, AL-BL contains one *C4A* gene and one *C4B* gene).

C4 isotype dosages were then tested for disease association by logistic regression, with the inclusion of four available ancestry covariates derived from genome-wide principal component analysis (PCA) as additional independent variables, $PC_{c,}$

$$\text{logit}(\theta) = \beta_0 + \beta_1 \text{C4} + \sum_c \beta_c \text{PC}_c + \varepsilon \quad (1)$$

where $\theta = \text{E}[\text{SLE}|\mathbf{X}]$, C4 is dosage of one of the isotypes per individual, β_0 is the fit intercept, other β values associated with each independent variable are best fit coefficients across the cohort, and ε is residual error. For Sjögren's syndrome, the model instead included two available multiethnic ancestry covariates from dbGaP that correlated strongly with European-specific ancestry covariates (specifically, PC5 and PC7) and smoking status as independent variables. Coefficients for relative weighting of *C4A* and *C4B* dosages (C4A and C4B) were obtained from a joint logistic regression,

$$\text{logit}(\theta) = \beta_0 + \beta_1 \text{C4A} + \beta_2 \text{C4B} + \sum_c \beta_c \text{PC}_c + \varepsilon \quad (2)$$

where terms are as in equation (1) except both *C4A* and *C4B* isotype dosages are included.

The values per individual of $\beta_1\text{C4A} + \beta_2\text{C4B}$ were used as a combined C4 risk term for estimating both association strength (Extended Data Fig. 3a, b) as well as evaluating the relationship between the strength of nearby variants' association with SLE or Sjögren's syndrome and linkage with C4 variation (Extended Data Fig. 4a–c).

Joint dosages of *C4A* and *C4B* for each individual in the same cohort were estimated by summing across their genotype probabilities of paired structural alleles that encode for the same diploid copy numbers of both *C4A* and *C4B* (Extended Data Fig. 2a, b). For each individual or genome, this yields a joint dosage distribution of *C4A* and *C4B* gene copy number, reflecting any possible imputed haplotype-level dosages with non-zero probability. Joint dosages for *C4A* and *C4B* diploid copy numbers were tested for association with SLE in a joint model with the same ancestry covariates (Fig. 1a),

$$\text{logit}(\theta) = \beta_0 + \sum_{i,j} \beta_{i,j} P(\text{C4A} = i, \text{C4B} = j) + \sum_c \beta_c \text{PC}_c + \varepsilon \quad (3)$$

where terms are as in equation (1) except $P(\text{C4A} = i, \text{C4B} = j)$ which represents the probability that an individual has i integer copies of *C4A* and j integer copies of *C4B*.

Calculation of composite C4 risk for SLE

SLE risk was strongly associated with *C4A* and *C4B* copy numbers (Fig. 1a) in an initial, simple model in which their contributions were treated as linear and independent. In specific subsequent analyses (for example, to map C4-independent effects), to account for the possibility of nonlinear or interacting contributions, a composite C4 risk score was derived by taking the weighted sum of joint *C4A* and *C4B* dosages multiplied by the corresponding effect sizes from the aforementioned model of the joint *C4A* and *C4B* diploid copy numbers. The weights for calculating this composite C4 risk term were computed from the data from the European ancestry cohort, and then applied unchanged to analysis of the African American cohort.

Associations of variants across the MHC region to SLE and Sjögren's syndrome

Genotypes for non-array SNPs were imputed with IMPUTE2 using the 1,000 Genomes reference panel; separate analyses were performed for the European-ancestry and African American cohorts. Unless otherwise stated, all subsequent SLE analyses were performed identically for both European ancestry and African American cohorts. Dosage of each variant, v_i, was tested for association with SLE or Sjögren's syndrome in a logistic regression including available ancestry covariates (and smoking status for Sjögren's syndrome) first alone (Extended Data Fig. 3a, b),

$$\text{logit}(\theta) = \beta_0 + \beta_1 v_i + \sum_c \beta_c \text{PC}_c + \varepsilon \quad (4)$$

then with C4 composite risk (Extended Data Fig. 3c),

$$\text{logit}(\theta) = \beta_0 + \beta_1 v_i + \beta_1 \text{C4} + \sum_c \beta_c \text{PC}_c + \varepsilon \quad (5)$$

where other terms are as in equation (1). For Sjögren's syndrome, the simpler weighted (2.3)C4A + C4B model was used instead of composite risk term, as the cohort's size gave poor precision to estimates of risk for many joint (C4A, C4B) copy numbers (Extended Data Fig. 3d). The Pearson correlation between the C4 composite risk term and each other variant was computed and squared (r^2) to yield a measure of LD between C4 composite risk and that variant in that cohort.

Association analyses for specific C4 structural alleles

The C4 structural haplotypes were tested for association with disease (Figs. 1b, 2a) in a joint logistic regression that included (1) terms for dosages of the five most common C4 structural haplotypes (AL-BS, AL-BL,

AL-AL, BS and AL), (2) (for SLE and Sjögren's syndrome) rs2105898 genotype, and (3) ancestry covariates and (for Sjögren's syndrome) smoking status,

$$\text{logit}(\theta) = \beta_0 + \beta_1 \text{BS} + \beta_2 \text{AL} + \beta_3 \text{ALBS} + \beta_4 \text{ALBL} + \beta_5 \text{ALAL} + \beta_6 \text{rs2105898} + \sum_c \beta_c \text{PC}_c + \varepsilon \quad (6)$$

where other terms are as in equation (1). Several of these common C4 structural alleles arose multiple times on distinct haplotypes; we term the set of haplotypes in which such a common allele appeared as haplogroups. The haplogroups can be further tested in a logistic regression model in which the structural allele appearing in all member haplotypes is instead encoded as dosages for each of the SNP haplotypes in which it appears. These association analyses (Figs. 1b, 2a) were performed as in equation (6), with structural allele dosages for ALBS, ALBL and ALAL replaced by multiple terms for each distinct haplotype.

To delineate the relationship between C4-BS and *DRB1**03:01 alleles—which are highly linked in European ancestry haplotypes—allelic dosages per individual in the African American SLE cohort were rounded to yield the most likely integer dosage for each. Although genotype dosages for each are reported by BEAGLE and HIBAG respectively, probabilities per haplotype are not linked and multiplying possible diploid dosages could yield incorrect non-zero joint dosages. Joint genotypes were tested as individual terms in a logistic regression model (Fig. 2b),

$$\text{logit}(\theta) = \beta_0 + \sum_{i,j} \beta_{i,j} P(\text{C4–BS} = i, \text{DRB1*03 : 01} = j) + \sum_c \beta_c \text{PC}_c + \varepsilon \quad (7)$$

where terms are as in equation (1) except P(C4-BS = i,DRB1*03:01 = j) which represents the probability that an individual has i haplotypes with C4-BS allele and j haplotypes with *DRB1**03:01 allele.

Sex-stratified associations of C4 structural alleles and other variants with SLE, Sjögren's syndrome and schizophrenia

Determination of an effect from sex on the contribution of overall C4 variation to risk for each disorder was done by including an interaction term between sex and C4; that is, (2.3)C4A + C4B for SLE and Sjögren's syndrome and estimated C4A expression for schizophrenia:

$$\text{logit}(\theta) = \beta_0 + \beta_2 \text{C4} + \beta_3 I_{\text{sex}} + \beta_4 I_{\text{sex}} \text{C4} + \sum_c \beta_c \text{PC}_c + \varepsilon \quad (8)$$

where terms are as in equation (1) except the term C4 = (2.3)C4A + C4B and I_{sex} which is an indicator variable for whether an individual is male.

Each variant in the MHC region was tested for association with among European ancestry cases and cohorts in a logistic regression as in equations (4)–(6) using only male cases and controls, and then separately using only female cases and controls (Extended Data Fig. 6a–c). Likewise, allelic series analyses were performed as in equation (7), but in separate models for men and women (Fig. 3a, b).

To assess the relationship between sex bias in the risk associated with a variant and linkage to C4 composite risk (as non-negative r^2), male and female log-odds were multiplied by the sign of the Pearson correlation between that variant and C4 composite risk before taking the difference.

Analyses of CSF

CSF from healthy individuals was obtained from two research panels. The first panel, consisting of 533 donors (327 male, 126 female) from hospitals around Utrecht, Netherlands, was described previously[49,50]. The donors were generally healthy research participants undergoing spinal anaesthesia for minor elective surgery. The same donors were previously genotyped using the Illumina Omni SNP array. To estimate C4 copy numbers, we used SNPs from the MHC region (chr6:24-34 Mb on hg19) as input for C4 allele imputation with Beagle, as described above in 'Imputation of C4 alleles'.

The second CSF panel sampled specimens from 56 donors (14 male, 42 female) from Brigham and Women's Hospital (BWH) under a protocol approved by the institutional review board at BWH (IRB protocol ID no. 1999P010911) with informed consent. These samples were originally obtained to exclude the possibility of infection, and clinical analyses had revealed no evidence of infection. Donors ranged from 18 to 64 years of age. Blood samples from the same individuals were used for extraction of genomic DNA, and C4 gene copy number was measured by droplet digital PCR (ddPCR) as previously described[7]. Samples were excluded from measurements if they lacked C4 genotypes, sex information, or contained visible blood contamination.

C4 measurements were performed by sandwich ELISA of 1:400 dilutions of the original CSF sample using goat anti-sera against human C4 as the capture antibody (Quidel, A305, used at 1:1,000 dilution), FITC-conjugated polyclonal rabbit anti-human C4c as the detection antibody (Dako, F016902-2, used at 1:3,000 dilution), and alkaline phosphatase–conjugated polyclonal goat anti-rabbit IgG as the secondary antibody (Abcam, ab97048, used at 1:5,000 dilution). C3 measurements were performed using the human complement C3 ELISA kit (Abcam, ab108823).

Because C4 gene copy number had a large and proportional effect on C4 protein concentration in these CSF samples (Extended Data Fig. 7a), we corrected for C4 gene copy number in our analysis of relationship between sex and C4 protein concentration, by normalizing the ratio of C4 protein (in CSF) to C4 gene copies (in genome). Therefore, these analyses included only samples for which DNA was available or C4 was successfully imputed. In total, 495 (332 male, 163 female) C4 and 304 (179 male, 125 female) C3 concentrations were obtained across both cohorts. log concentrations of C3 (in ng ml^{-1}) and C4 (in ng ml^{-1}, per C4 gene copy number) protein were then used separately in linear regression models to estimate a sex-unbiased cohort-specific offset for each protein,

$$\log_{10}(\text{C3 or C4 concentration}) = \beta_0 + \beta_1 I_{\text{sex}} + \beta_2 I_{\text{cohort}} + \varepsilon \quad (9)$$

to be applied to all concentrations for that protein, where I_{sex} is an indicator variable for whether an individual is male, I_{cohort} is an indicator variable for whether an individual was in the second cohort, β_0 is the fit intercept, other β associated with each independent variable are best fit coefficients across the cohort, and ε is residual error. Estimation of average measurements by age for each sex was done by LOESS (Fig. 3c, d). To evaluate the significance of sex effects, we used these cohort-corrected concentrations estimates and analysed them with the non-parametric unsigned Mann–Whitney rank-sum test comparing concentration distributions for males and females.

Analyses of blood plasma

Blood plasma was collected and immunoturbidimetric measurements of C3 and C4 protein in 1,844 individuals (182 men, 1662 women) by Sjögren's International Collaborative Clinical Alliance (SICCA) from individuals with and without Sjögren's syndrome as previously described[51]. C4 copy numbers for these individuals were previously imputed for use in logistic regression of Sjögren's syndrome risk. As C4 copy number has an effect on measured C4 protein similar to CSF (Extended Data Fig. 7b), we normalized C4 levels to them in all following analyses. Estimation of average measurements by age for each sex was done by local polynomial regression smoothing (LOESS) on log-concentrations of C3 (mg dl^{-1}) and C4 (mg dl^{-1}, per C4 gene copy number) protein (Extended Data Fig. 7c, d). To evaluate the significance of sex bias within age ranges displaying the greatest difference (informed by LOESS), we analysed individuals in these bins with the

non-parametric unsigned Mann–Whitney rank-sum test comparing concentration distributions for males and females.

Difference in C4 protein levels between individual with and without Sjögren's syndrome was done by performing a non-parametric unsigned Mann–Whitney rank-sum test on C4 protein levels with and without normalization to C4 genomic copy number (Extended Data Fig. 7e, f).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Individual genotype data for Sjögren's syndrome cases and controls and individual plasma concentrations for C4 and C3 are available in dbGaP under accession number phs000672.v1.p1. Individual genotype data for schizophrenia cases and controls are available by application to the Psychiatric Genomics Consortium (PGC). Questions regarding individual genotype data for SLE cases and controls of European and/or African American ancestry can be directed to T.J.V. Data resources are available on the McCarroll lab website at http://mccarrolllab.org/resources/resources-for-c4/. We have deposited the haplotype reference panel we created for C4 imputation in dbGaP under accession number phs001992.v1.p1. Genotype and protein concentration data for CSF samples are available upon request.

Code availability

Software scripts and instructions for imputing C4 alleles into SNP datasets are available on the McCarroll laboratory website at http://mccarrolllab.org/resources/resources-for-c4/.

44. Handsaker, R. E. et al. Large multiallelic copy number variations in humans. *Nat. Genet.* **47**, 296–303 (2015).
45. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
46. Browning, B. L. & Browning, S. R. Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.* **98**, 116–126 (2016).
47. Zheng, X. et al. HIBAG—HLA genotype imputation with attribute bagging. *Pharmacogenomics J.* **14**, 192–200 (2014).
48. Zheng, X. Imputation-based HLA typing with SNPs in GWAS studies. *Methods Mol. Biol.* **1802**, 163–176 (2018).
49. Luykx, J. J. et al. A common variant in ERBB4 regulates GABA concentrations in human cerebrospinal fluid. *Neuropsychopharmacology* **37**, 2088–2092 (2012).
50. Albersen, M. et al. Vitamin B-6 vitamers in human plasma and cerebrospinal fluid. *Am. J. Clin. Nutr.* **100**, 587–592 (2014).
51. Malladi, A. S. et al. Primary Sjögren's syndrome as a systemic disease: a study of participants enrolled in an international Sjögren's syndrome registry. *Arthritis Care Res. (Hoboken)* **64**, 911–918 (2012).
52. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
53. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

Acknowledgements This work was supported by the National Human Genome Research Institute (HG006855), the National Institute of Mental Health (MH112491, MH105641, MH105653), the Stanley Center for Psychiatric Research, and the National Institute for Health Research Biomedical Research Centre (NIHR BRC) at Guy's and St Thomas' NHS Foundation and King's College London. We thank C. Usher and C. Patil for contributions to the figures and manuscript text, and M. Florio for suggestions regarding figure display.

Author contributions N.K., A.S., T.J.V. and S.A.M. conceived the genetic studies. M.T.P., C.N.P. and M.B. collected and contributed WGS data for the Genomic Psychiatry Cohort. R.E.H. and C.W.W. genotyped C4 structural variation in the Genomic Psychiatry Cohort and optimized variant selection for use as a reference panel in the imputation of C4 variation into lupus and schizophrenia cohorts (Extended Data Fig. 1). T.J.V., R.R.G., L.A.C., C.D.L., R.P.K., J.B.H., K.M.K., D.L.M. and P.T. contributed genotype data and imputation of non-C4 variation for analysis of SLE cohorts. K.E.T. and L.A.C. contributed genotype and phenotype data along with imputation of non-C4 variation for analysis of the Sjögren's syndrome cohort. Investigators in the Schizophrenia Working Group of the Psychiatric Genomics Consortium collected and phenotyped cohorts and contributed genotype data for analysis of schizophrenia cohorts. N.K. did the imputation and association analysis (Figs. 1, 2, 3a, b, Extended Data Figs. 2–6). T.J.V., R.R.G. and D.L.M. provided valuable advice on the analysis and interpretation of SLE-association results. R.A.O. and L.M.O.L. collected and provided CSF samples composing the group from Utrecht, Netherlands. C.E.S. collected and provided CSF samples composing the Brigham & Women's Hospital group. H.d.R and K.T. performed the C4 and C3 immunoassay experiments on CSF samples (Fig. 3c, d, Extended Data Fig. 7a). N.K. did the analysis of plasma C4 and C3 concentrations (Extended Data Fig. 7b–f). S.A.M and N.K. wrote the manuscript with contributions from all authors. Management Committee of Wellcome Trust Case–Control Consortium 2: P.D., I.B., J.M.B., E.B., M.A.B., J.P.C., A.C., P.D., A.D., J.J., H.S.M., C.G.M., C.N.A.P., R.P., A.R., S.J.S., R.C.T., A.C.V. and N.W.W; Data and Analysis Group of Wellcome Trust Case–Control Consortium 2: C.C.A.S., G.B., C.B., P.D., C.F., E.G., G.H., R.P., M.P., A.S., Z.S., D.V.; DNA, Genotyping, Data QC, and Informatics of Wellcome Trust Case–Control Consortium 2: C.L., I.B., H.B., S.J.B., P.D., S.D., S.E., M.G., E.G., R.G., N.H., S.E.H., A.J., J.L., O.T.M., S.C.P., R.R., M.R., A.T.-G., M.W., P.W., P.W., S.W.; Publications Committee of Wellcome Trust Case–Control Consortium 2: C.G.M., J.M.B., M.A.B., A.C., M.I..M. and C.C.A.S.

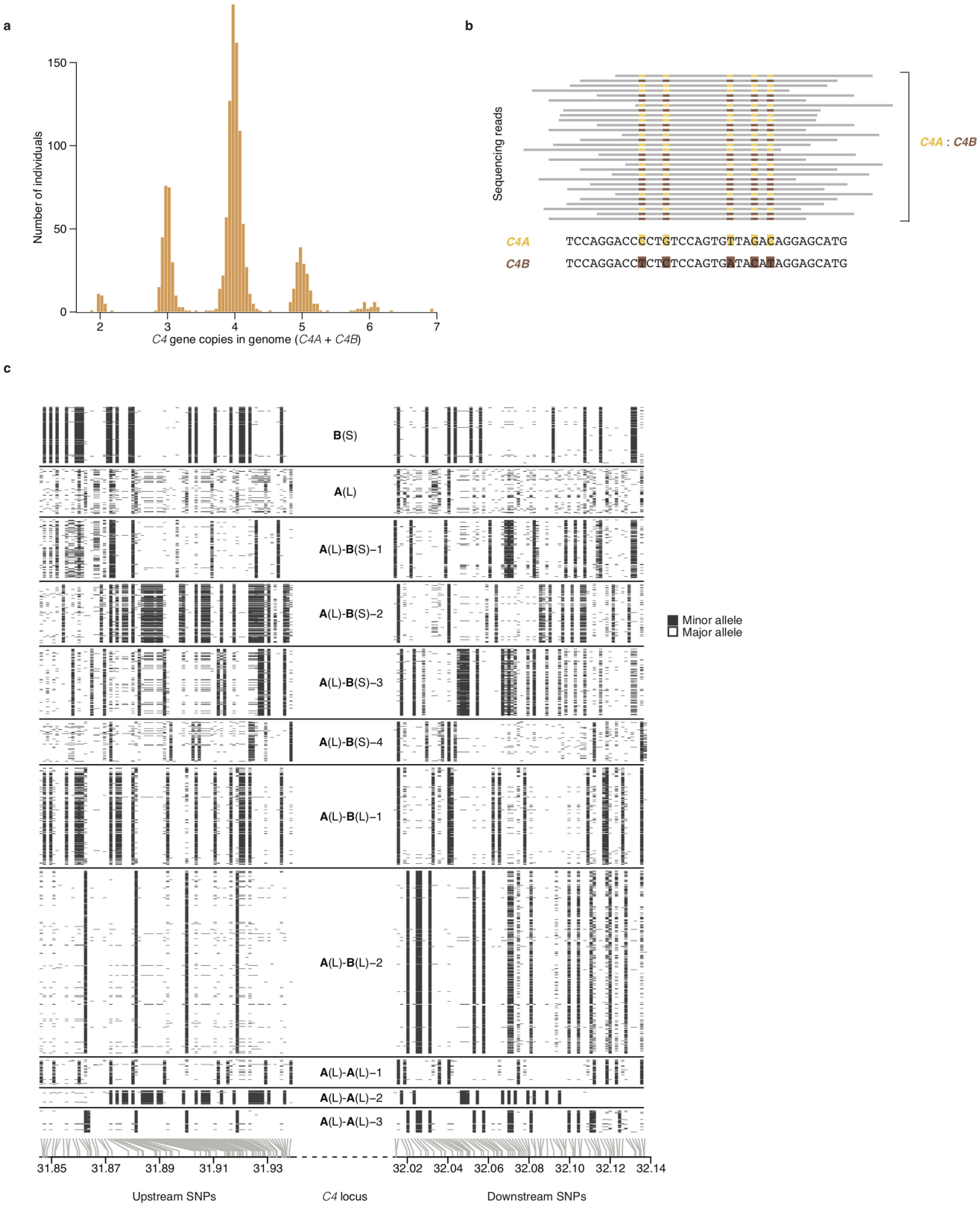
Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2277-x.

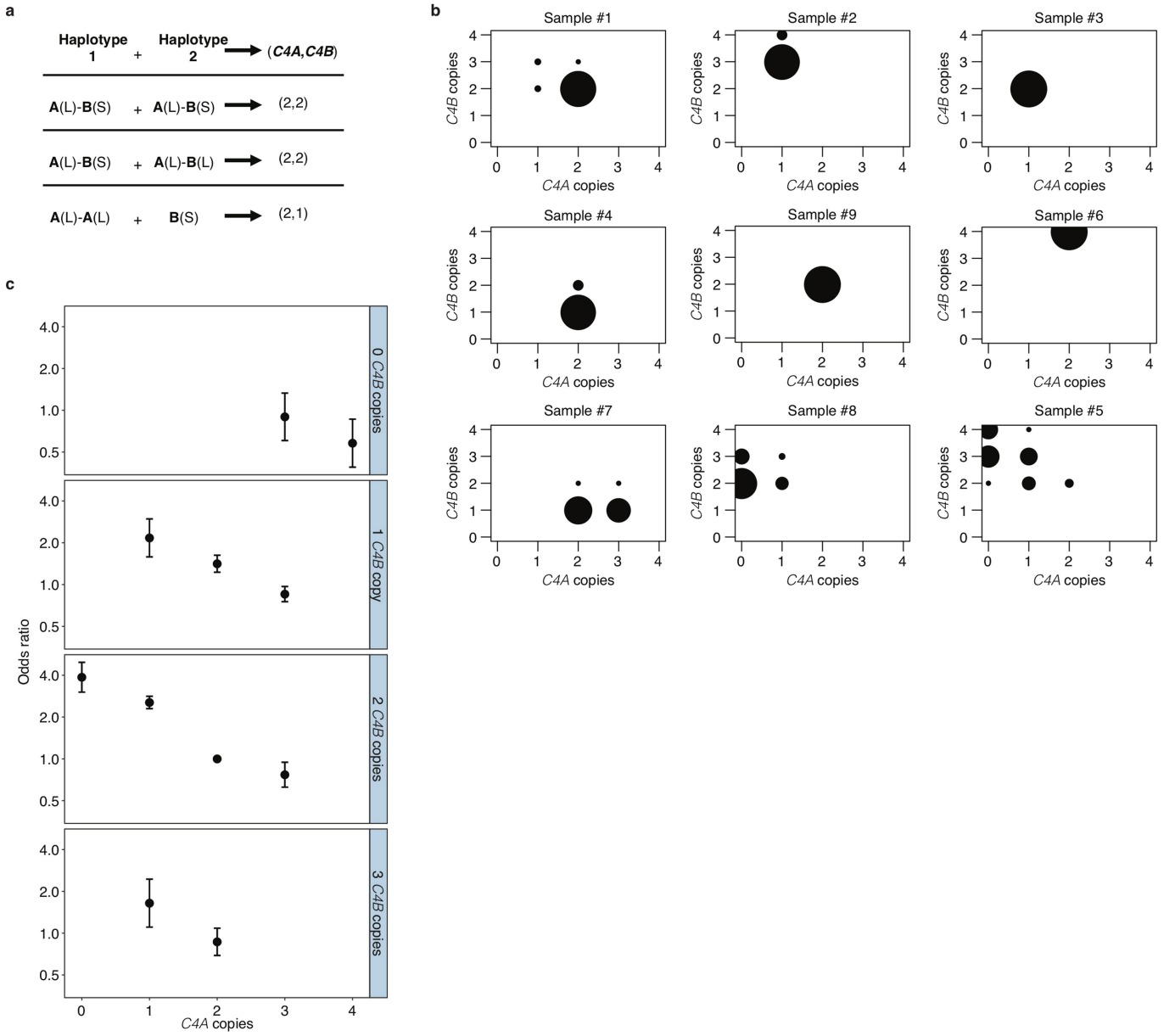
Correspondence and requests for materials should be addressed to N.K., T.J.V. or S.A.M.

Peer review information *Nature* thanks John Armour and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

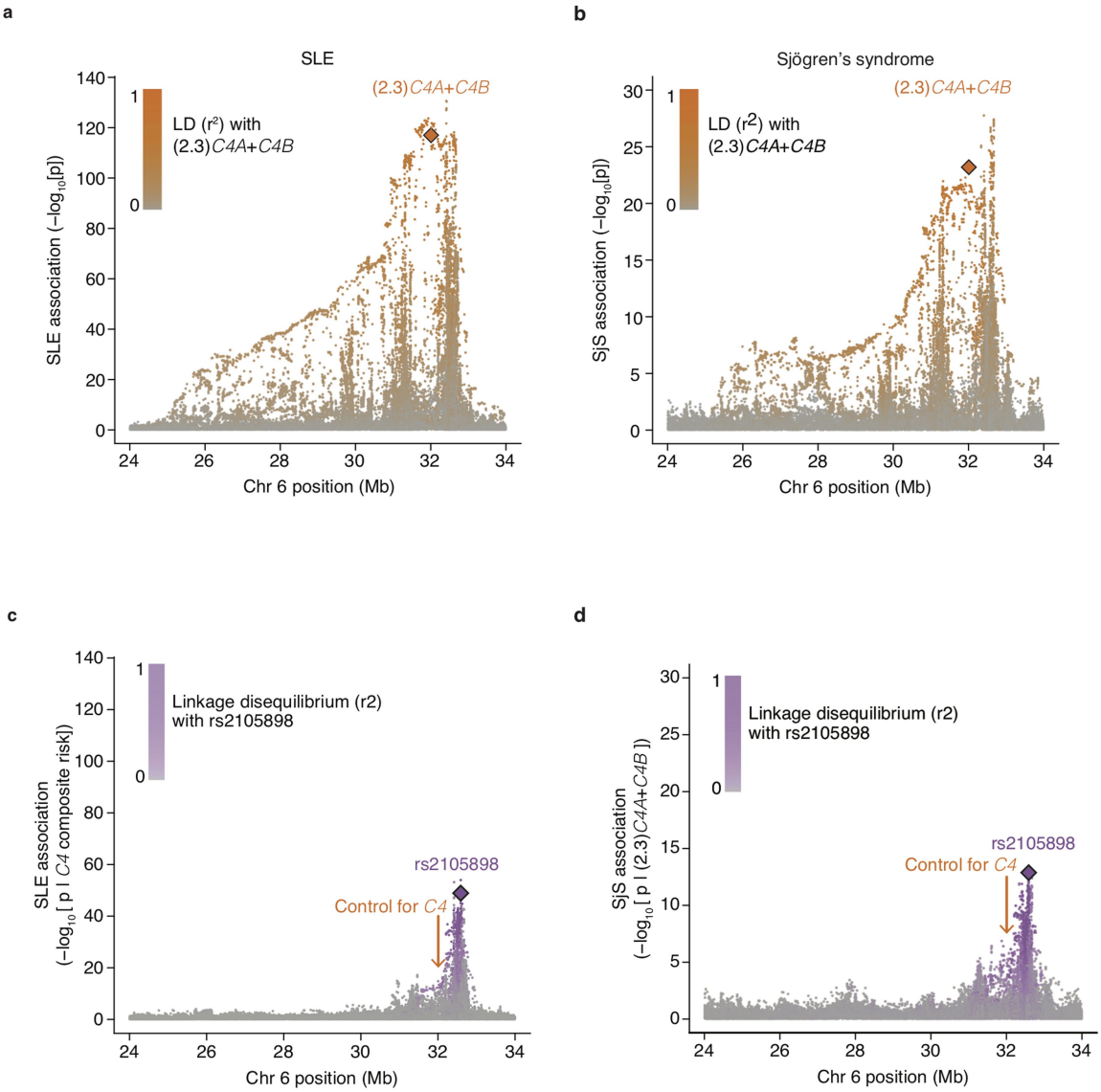


Extended Data Fig. 1 | See next page for caption.

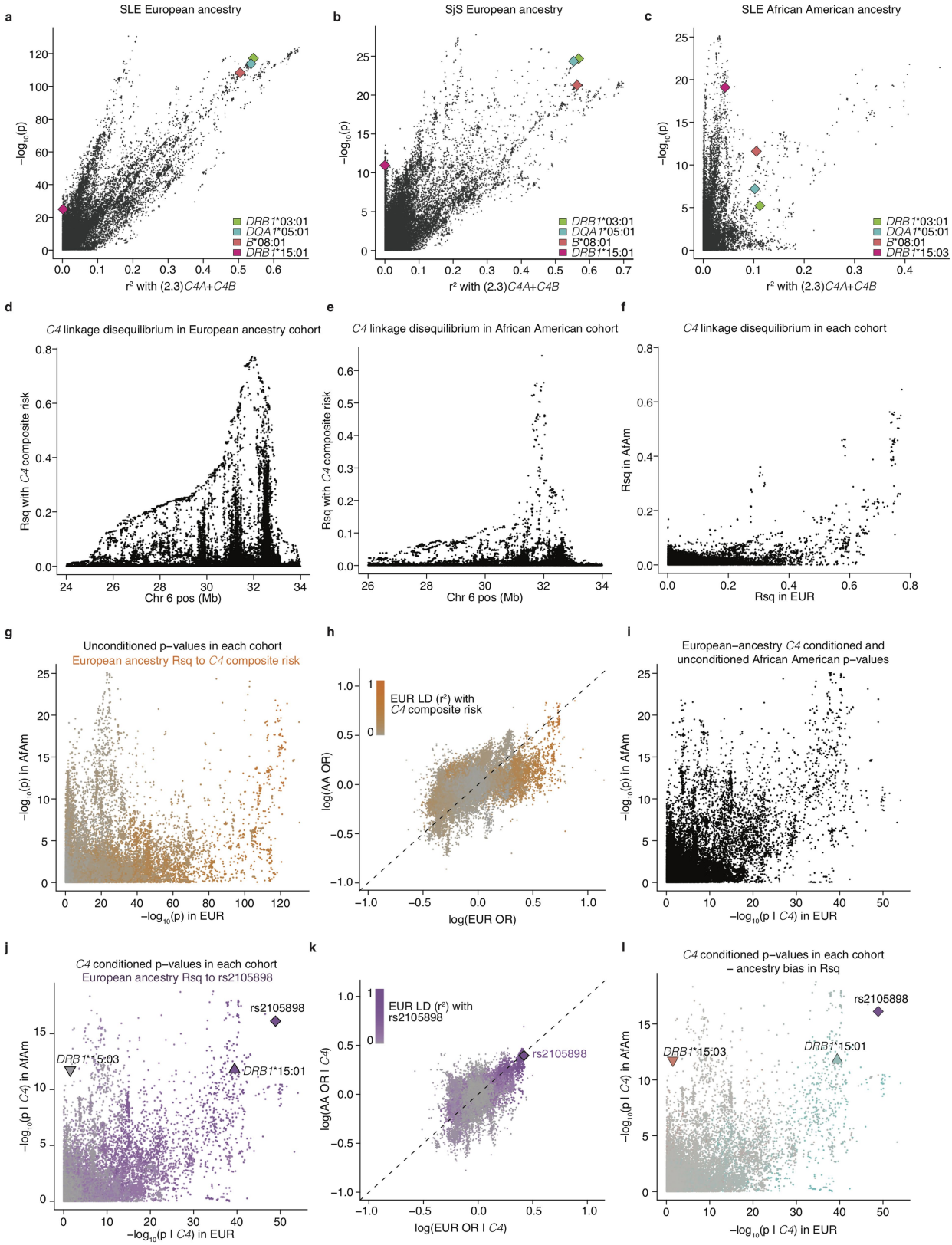
Extended Data Fig. 1 | A panel of 2,530 reference haplotypes (created from WGS data) containing C4 alleles and SNPs across the MHC genomic region enables imputation of C4 alleles into SNP data. a, Distributions (across 1,265 individuals) of total C4 gene copy number (*C4A* + *C4B*), as measured from read depth of coverage across the C4 locus, in WGS data. **b**, The relative numbers of reads that overlap sequences specific to *C4A* or *C4B* (together with the total C4 gene copy number as in **a**) are used to infer the underlying copy numbers of the *C4A* and *C4B* genes. For example, in an individual with four C4 genes, the presence of equal numbers of reads specific to *C4A* or *C4B* suggests the presence of two copies each of *C4A* and *C4B*. Precise statistical approaches (including inference of probabilistic dosages) and further approaches for phasing C4 allelic states with nearby SNPs to create reference haplotypes, are described in Methods. **c**, The SNP haplotypes flanking each C4 allele are shown as rows (SNPs as columns), with white and black representing the major and minor allele of each SNP. Grey lines at the bottom indicate the physical location of each SNP along chromosome 6. The differences among the haplotypes are most pronounced closest to C4 (towards the centre of the plot), as historical recombination events in the flanking megabases will have caused the haplotypes to be less consistently distinct at greater genomic distances from C4. The patterns indicate that many combinations of *C4A* and *C4B* gene copy numbers have arisen recurrently on more than one SNP haplotype, a relationship that can be used in association analyses (Fig. 1b).



Extended Data Fig. 2 | Aggregation of joint *C4A* and *C4B* genotype probabilities per individual across imputed C4 structural alleles for estimation of SLE risk for each combination. a, An individual's joint *C4A* and *C4B* gene copy number can be calculated by summing the *C4A* and *C4B* gene contents for each possible pair of two inherited alleles. Many pairings of possible inherited alleles result in the same joint *C4A* and *C4B* gene copy number. **b**, Each individual's *C4A* and *C4B* gene copy number was imputed from their SNP data, using the reference haplotypes summarized in Extended Data Fig. 1c. For more than 95% of individuals (exemplified by samples 1–6 in the figure), this inference can be made with >90% certainty or confidence (the areas of the circles represent the posterior probability distribution over possible *C4A/C4B* gene copy numbers). For the remaining individuals (exemplified by samples 7–9 in the figure), greater statistical uncertainty persists about C4 genotype. To account for this uncertainty, in downstream association analysis, all C4 genotype assignments are handled as probabilistic gene dosages—analogous to the genotype dosages that are routinely used in large-scale genetic association studies that use imputation. **c**, Odds ratios and 95% confidence intervals underlying each of the C4-genotype risk estimates in Fig. 1a presented as a series of panels for each observed copy number of *C4B*, with increasing copy number of *C4A* for that *C4B* dosage (*x*-axis). Data are from analysis of 6,748 SLE cases and 11,516 controls of European ancestry.



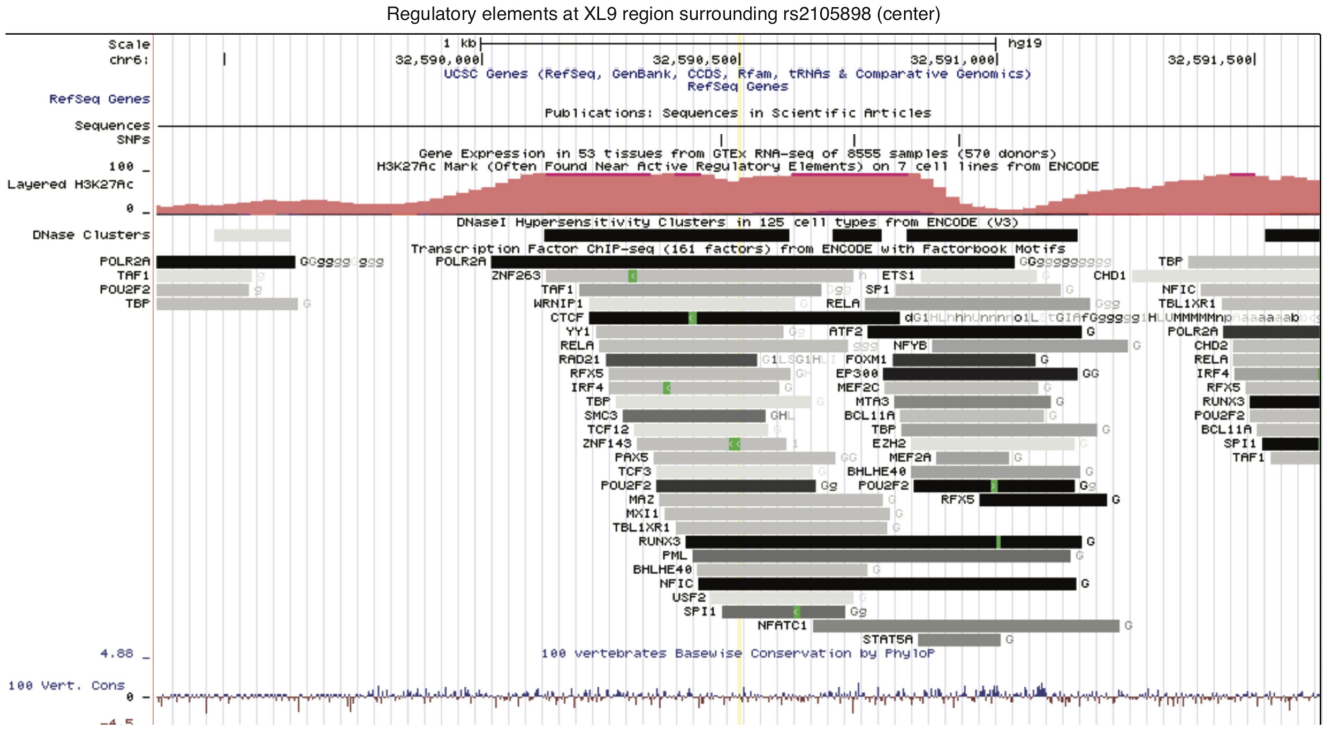
Extended Data Fig. 3 | Conditional association analyses for genetic markers across the extended MHC genomic region within the European-ancestry SLE and Sjögren's syndrome cohorts. a, Association of SLE with genetic markers (SNPs and imputed HLA alleles) across the extended MHC locus within the European-ancestry SLE cohort (6,748 cases and 11,516 controls). Orange diamond: an initial estimate of C4-related genetic risk, calculated as a weighted sum of the number of *C4A* and *C4B* gene copies: (2.3)*C4A*+*C4B*, with the weights derived from the relative coefficients estimated from logistic regression of SLE risk versus *C4A* and *C4B* gene dosages. This risk score is imputed with an accuracy (r^2) of 0.77. Points representing all other genetic variants in the MHC locus are shaded orange according to their level of LD-based correlation to this C4-derived risk score. **b**, As in **a**, but for a European-ancestry Sjögren's syndrome (SjS) cohort (673 cases and 1,153 controls). The orange diamond here also represents (2.3)*C4A* + *C4B*, with this weighting derived from the relative coefficients estimated from logistic regression of Sjögren's syndrome risk versus *C4A* and *C4B* gene dosages. **c**, Association of SLE with genetic markers (SNPs and imputed HLA alleles) across the extended MHC locus within the European-ancestry SLE cohort controlling for C4 composite risk (weighted sum of risk associated with various combinations of *C4A* and *C4B*). Variants are shaded in purple by their LD with rs2105898, an independent association identified from trans-ancestral analyses. **d**, As in **c**, but in association with a European-ancestry Sjögren's syndrome cohort. Here a simpler linear model of risk contributed by *C4A* and *C4B* was used instead of a weighted sum across all possible combinations.



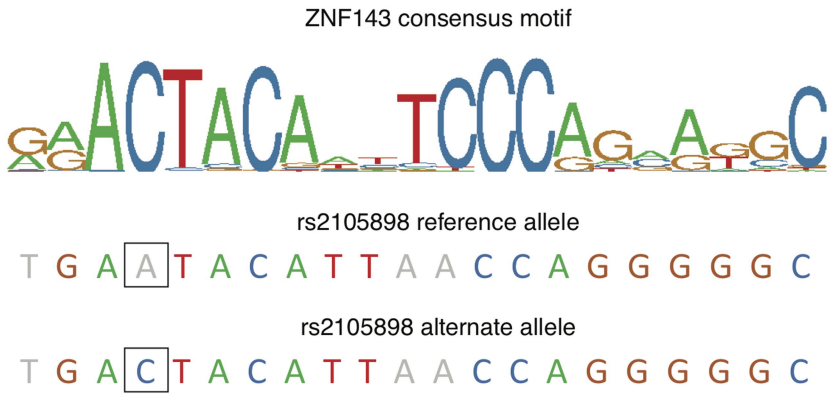
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Using C4 gene variation to understand the appearance of trans-ancestral disparity in MHC association signals, and to fine-map an additional genetic effect. Association signals (for SLE and Sjögren's syndrome) for variants in a multi-megabase region of human chromosome 6 containing the MHC region including the HLA and C4 genes. **a**, Relationship between SLE association ($-\log_{10}(p)$, *y*-axis) and LD to the weighted C4 risk score (*x* axis) for genetic markers and imputed HLA alleles across the extended MHC locus. In this European-ancestry cohort, it is unclear (from this analysis alone) whether the association with the markers in the predominant ray of points (at an angle of ~45° from the *x* axis) is driven by variation at C4 or by the long haplotype containing *DRB1**03:01 (green), *DQA1**05:01 (blue), *B**08:01 (red) and many other SNPs (black). In addition, at least one independent association signal (a ray of points at a higher angle in the plot, with strong association signals and only weak linkage disequilibirum-based correlation to C4 and *DRB1**0301) with some LD to *DRB1**15:01 (maroon) is also present. **b**, Analysis as in **a**, but for associations to Sjögren's syndrome in a cohort of European ancestry. As in SLE, it is initially unclear whether the genetic association signal is driven by variation at C4 or by linked HLA alleles, *DRB1**03:01 (green), *DQA1**05:01 (blue), and *B**08:01 (red). There is also an independent association signal with LD to *DRB1**15:01 (maroon). **c**, Analysis as in **a**, but of an African American SLE case–control cohort (in which LD in the MHC region is more limited). Many MHC-region SNPs associate with SLE in proportion to their LD with the weighted C4 risk score inferred from the earlier analysis of the European-ancestry cohort; this C4-derived risk score itself associates with SLE at $P = 4.3 \times 10^{-19}$ in a logistic regression on 1,494 SLE cases and 5,908 controls. No similarly strong association is observed for *DRB1**03:01, *DQA1**05:01 or *B**08:01, HLA alleles which are in strong LD with C4 risk on European-ancestry (but not African American) haplotypes. An independent association signal is also present in this cohort, more clearly in LD with the *DRB1**15:03 allele (maroon). **d**, LD in the European-ancestry SLE cohort between the composite C4 risk term (weighted sum of risk associated with various combinations of *C4A* and *C4B* from Fig. 2a) and variants in the MHC region as r^2 (*y*-axis). **e**, As in **d**, but for the African American SLE cohort. **f**, LD (to C4 composite risk) for the same variants in European-ancestry individuals (*x* axis) and African Americans (*y* axis). Note the abundance of variants that have greater LD with C4 risk among European-ancestry individuals than among African Americans. Also, several groups of variants have equivalent LD (to C4 risk) in European ancestry individuals but exhibit a range of LD to C4 risk among African Americans. **g**, Associations with SLE ($-\log_{10} P$ values) for the same variants in European ancestry (*x* axis) and African American (*y* axis) case–control cohorts. Orange shading represents the extent of LD with C4 risk in European ancestry individuals. Variants with strong European-specific association to SLE are generally in strong LD with C4 risk among European-ancestry individuals. **h**, Comparison of the inferred effect size from association of genetic markers with SLE (unconditioned log odds ratios) among European-ancestry (*x* axis) and African American (*y* axis) research participants. As also seen in **g**, variants with discordant associations to SLE (across populations) tend also to be in strong LD to C4 risk among European-ancestry individuals. **i**, As in **g**, but now controlling for the effect of C4 variation in analysis of the European-ancestry cohort (*x* axis). Note that controlling for C4 risk in European-ancestry individuals alone greatly aligns (relative to **g**) the patterns of association between European ancestry and African American cohorts. **j**, As in **i**, but now also controlling for the effect of C4 in associations of the African American cohort. Note that due to the lack of strong LD relationships between C4 and variants in the MHC region in African Americans (**e**), this further adjustment does not change results strongly (relative to **i**). The independent signal, rs2105898, and HLA alleles, *DRB1**15:01 and *DRB1**15:03, are also highlighted. LD with rs2105898 in European-ancestry individuals is indicated by purple shading. **k**, Comparison of the inferred effect sizes from association of genetic markers with SLE (log odds ratios) controlling for C4-derived risk among European-ancestry (*x* axis) and African American (*y* axis) research participants. Two SNPs (rs2105898 and rs9271513) that form a short haplotype common to both ancestry groups are among the strongest associations in both cohorts. (Their association to SLE in the European-ancestry cohort was initially much less remarkable than that of other SNPs that are in strong LD with C4.) LD with rs2105898 in European-ancestry individuals is indicated by purple shading. **l**, As in **i**, but with variants shaded by whether they exhibit greater LD to rs2105898 in Europeans (blue) or African Americans (red).

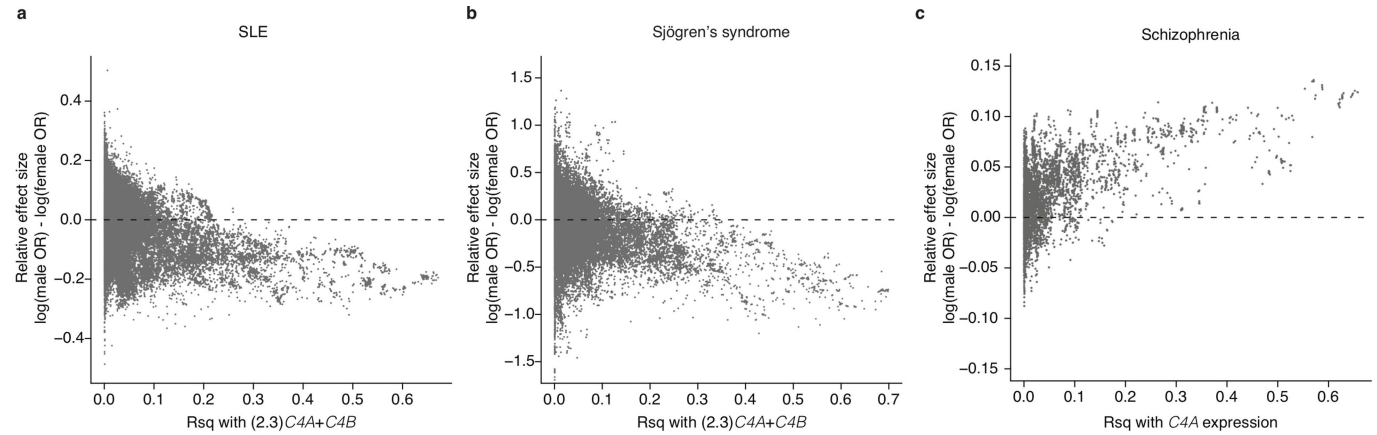
a



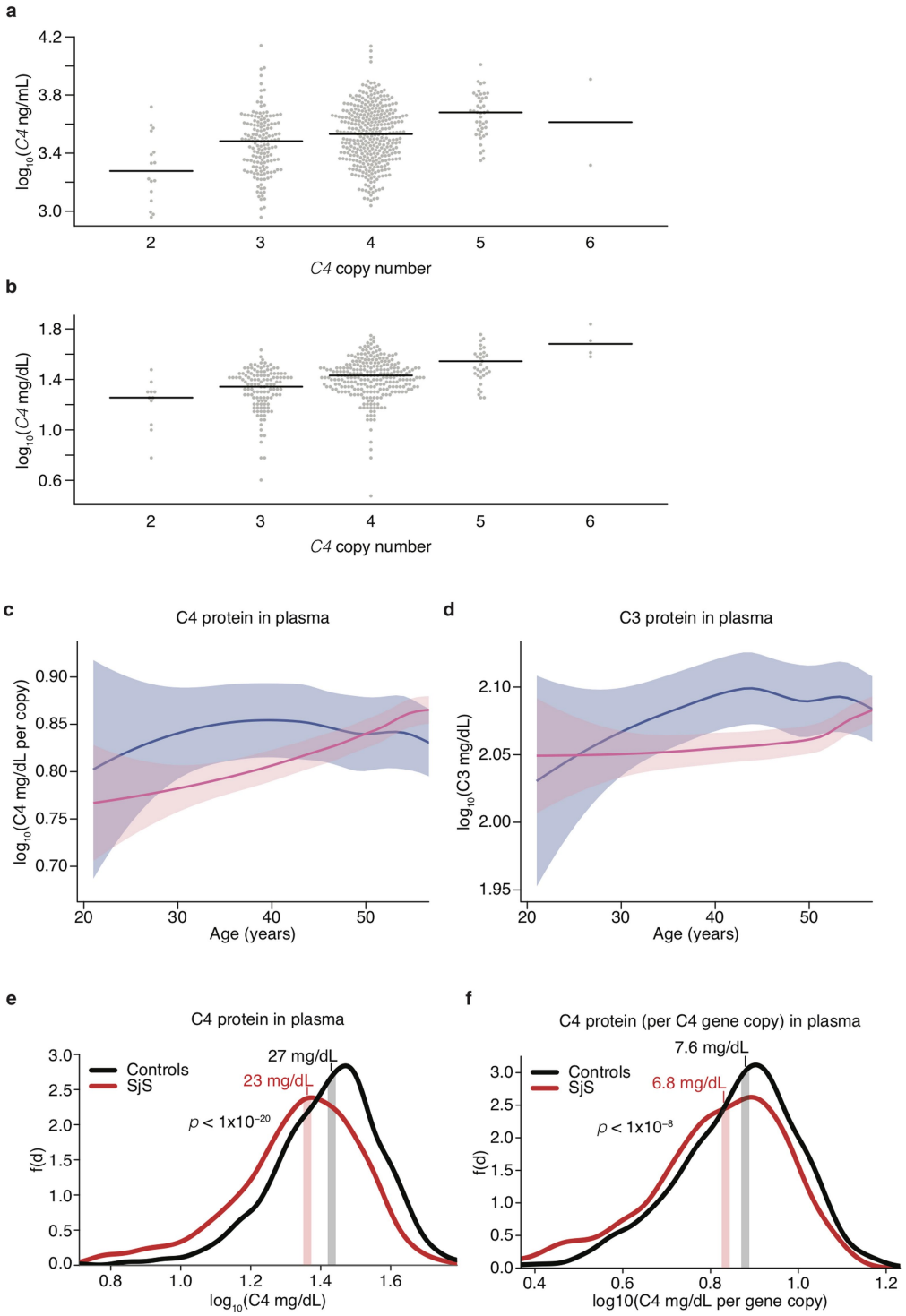
b



Extended Data Fig. 5 | Relationship of rs2105898 alleles to a known ZNF143 binding motif in the XL9 region of the MHC class II locus. a, Location of rs2105898 (yellow line at centre) within the XL9 region, with relevant tracks showing overlapping histone marks and transcription factor binding peaks (from ENCODE[52]), visualized with the UCSC genome browser[53]. **b**, ZNF143 consensus binding motif as a sequence logo, with the letters coloured if the base is present in more than 5% of observed instances. The alleles of rs2105898 are indicated by outlined box surrounding the base.



Extended Data Fig. 6 | Relationships between sex bias of disease associations and LD to C4 risk for variants in the MHC region. **a**, Relationship between male bias in SLE risk (difference between male and female log–odds ratios) and LD with C4 risk for common (minor allele frequency (MAF) > 0.1) genetic markers across the extended MHC region (6,748 cases and 11,516 controls). For each SNP, the allele for which sex risk bias is plotted is the allele that is positively correlated (via LD) with C4-derived risk score. **b**, Relationship between male bias in Sjögren's syndrome risk (log-odds ratios) and LD with C4 risk for common (MAF > 0.1) genetic markers across the extended MHC region (673 cases and 1,153 controls). For each SNP, the allele for which sex risk bias is plotted is the allele that is positively correlated (via LD) with C4-derived risk score. **c**, Relationship of male bias in schizophrenia risk (log odds ratios) and LD to *C4A* expression for common (MAF >0.1) genetic markers across the extended MHC region (28,799 cases and 35,986 controls). For each SNP, the allele for which sex risk bias is plotted is the allele that is positively correlated (via LD) with imputed *C4A* expression, as previously described[7].



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Correlation of C4 protein measurements in CSF and blood plasma with imputed C4 gene copy number and relationship of plasma complement to sex and Sjögren's syndrome status. **a**, Measurements of C4 protein in CSF obtained by ELISA ($n = 507$ total) are presented as $\log_{10}$[concentration (ng ml^{-1})] (y axis) for each observed or imputed copy number of total C4 (x axis, here showing most likely copy number from imputation). Because C4 gene copy number affects C4 protein levels so strongly, we normalized C4 protein measurements to each donor's C4 gene copy number in subsequent analyses (Fig. 3c). Bars indicate median values for each C4 copy number. **b**, Measurements of C4 protein in blood plasma obtained by immunoturbidimetric assays are presented as $\log_{10}$[concentration (mg dl^{-1})] (y axis) for each imputed most-likely copy number of C4 genes (x axis). Because C4 gene copy number affects C4 protein levels so strongly, we normalized C4 protein measurements by C4 gene copy number in subsequent analyses as in **c**. Due to the number of observations ($n = 1{,}844$ total), the plot is downsampled to 500 points; the median bars shown are for all individuals (before downsampling). **c**, Levels of C4 protein in blood plasma from 182 adult men and 1,662 adult women as a function of age. Concentrations are normalized to the number of C4 gene copies in an individual's genome (a strong independent source of variance) and shown on a $\log_{10}$ scale as a LOESS curve. Shaded regions represent 95% confidence intervals derived during LOESS. **d**, Levels of C3 protein in blood plasma as a function of age from the same individuals in panel **c**. Concentrations are shown on a $\log_{10}$ scale as a LOESS curve. Shaded regions represent 95% confidence intervals derived during LOESS. **e**, C4 protein in blood plasma was measured in 670 individuals with Sjögren's syndrome (red) and 1,151 individuals without Sjögren's syndrome (black) and is shown on a $\log_{10}$ scale (x axis). Vertical stripes represent median levels for cases and controls separately. Comparison of the two sets was done with a non-parametric two-sided Mann–Whitney rank-sum test ($P = 4.8 \times 10^{-21}$). **f**, As in **e**, but concentrations are normalized to the number of C4 gene copies in an individual's genome; this per-copy amount is shown on a $\log_{10}$ scale (x axis). Comparison of the two sets was done with a non-parametric two-sided Mann–Whitney rank-sum test ($P = 7.6 \times 10^{-9}$).

Extended Data Table 1 | Imputation accuracy for C4 copy numbers in European ancestry and African American haplotypes

	Imputation accuracy (r^2)	
Gene copy number	**European ancestry**	**African Americans**
C4	0.80	0.58
C4A	0.78	0.65
C4B	0.74	0.61
C4-HERV	0.91	0.76
2.3(*C4A*)+*C4B*	0.77	0.64

Imputation accuracy was evaluated by correlation of imputation results to C4 gene copy numbers directly inferred from WGS data. Aggregated copy numbers imputed from each round of leaving ten individuals out were correlated with the directly-typed measurements and are reported as r^2 for each feature of C4 structural variation for European ancestry (693 individuals) and African American (250 individuals) members of the reference panel separately.

Extended Data Table 2 | Frequency of common C4 alleles and their LD-based correlation with HLA alleles in European ancestry and African American cohorts

European ancestry

A			*B*			*C*			*C4* allele	Allele Frequency	*DRB1*			*DQA1*			*DQB1*		
allele	%	r^2	allele	%	r^2	allele	%	r^2			allele	%	r^2	allele	%	r^2	allele	%	r^2
01:01	69	0.27	08:01	93	**0.75**	07:01	93	**0.57**	B(S)	13.7%	03:01	94	**0.71**	05:01	94	**0.7**	02:01	94	**0.7**
									A(L)	4.8%									
						06:02	69	0.31	A(L)-B(S)-1	6.1%	07:01	74	0.25	02:01	74	0.25			
			44:03	54	0.28	16:01	53	0.39	A(L)-B(S)-2	4.5%	07:01	57	0.1	02:01	57	0.1	02:02	55	0.14
									A(L)-B(S)-3	3.8%									
									A(L)-B(S)-4	4.5%									
			07:02	64	**0.42**	07:02	63	0.35	A(L)-B(L)-1	15.5%	15:01	73	**0.49**	01:02	74	0.32	06:02	70	**0.47**
									A(L)-B(L)-2	23.1%									
			35:01	55	0.2	04:01	57	0.09	A(L)-A(L)-1	3.2%	01:01	65	0.14	01:01	65	0.11	05:01	64	0.1
									A(L)-A(L)-2	2.1%	13:01	67	0.16	01:03	65	0.13	06:03	67	0.15
02:01	65	0.03	44:02	74	0.24	05:01	72	0.23	A(L)-A(L)-3	4.5%	04:01	80	0.29	03:03	79	0.37	03:01	82	0.15

African American

A			*B*			*C*			*C4* allele	Allele Frequency	*DRB1*			*DQA1*			*DQB1*		
allele	%	r^2	allele	%	r^2	allele	%	r^2			allele	%	r^2	allele	%	r^2	allele	%	r^2
									B(S)	5.0%				01:02	51	0.01			
									A(L)	7.5%									
									A(L)-B(S)-1	14.1%									
									A(L)-B(S)-2	18.1%									
									A(L)-B(S)-3	17.7%									
									A(L)-B(S)-4	6.5%									
									A(L)-B(L)-1	4.4%	15:01	67	0.2	01:02	72	0.04	06:02	59	0.06
									A(L)-B(L)-2	4.5%									
									A(L)-A(L)-1	0.7%	01:01	57	0.07	01:01	53	0.01			
									A(L)-A(L)-2	0.8%									
02:01	72	0.03	44:02	86	0.31	05:01	78	0.17	A(L)-A(L)-3	0.8%	04:01	93	0.27	03:03	86	0.14	03:01	87	0.03

For each common C4 allele and HLA gene, the allele with strongest LD (r^2) is listed if present on more than half of the haplotypes with that C4 allele (for 36,528 European ancestry and 14,804 African American haplotypes separately, with exact fraction as a percentage). r^2 values greater than 0.4 are highlighted to point out particularly strong C4-HLA allele correlations, such as for several HLA alleles with the C4-B(S) allele in European ancestry individuals. Some common C4 alleles are further subdivided into distinct haplotypes used in imputation (and in Fig. 1b), as defined by shared alleles from variants flanking C4. Note that some alleles such as C4-A(L)-A(L)-3 are present at a low frequency in African Americans that might reflect their presence on admixed European-origin haplotypes spanning this region, whereas others such as C4-B(S) are likely to also exist on African haplotypes – these differences between C4 alleles are also reflected in the similarity of LD with HLA alleles to the corresponding row of the European ancestry section.

Extended Data Table 3 | Results of association analyses of SLE risk against C4 variation, HLA alleles, and/or rs2105898 in European ancestry and African American cohorts

European ancestry

	C4			*C4A*			*C4B*			*DRB1**03:01			*B**08:01			rs2105898				
Model	beta	se	-log10(p)	beta	se	-log10(p)	beta	se	-log10(p)	beta	se	-log10(p)	beta	se	-log10(p)	beta	se	-log10(p)	AIC	LRT -log10(p)
C4	-0.55	0.027	92.7																22855.26	260.2
C4A				-0.53	0.024	105.3													22790.05	274.3
C4A+C4B				-0.62	0.028	112	-0.27	0.037	12.3										22739.8	284.4
*DRB1**03:01										0.7	0.03	117.1							22748.33	283.3
*B**08:01													0.69	0.031	108.4				22790.65	274.2
rs2105898																-0.32	0.027	30.7	23153.86	195.5
*C4A + C4B + DRB1**03:01				-0.35	0.041	17.2	-0.11	0.041	2.3	0.4	0.046	17.5							22666.1	299.6
*C4A + C4B + B**08:01				-0.41	0.039	24.6	-0.17	0.039	4.7				0.35	0.044	14.4				22680.53	296.4
C4A + C4B + rs2105898				-0.67	0.028	122.8	-0.32	0.038	16.4							-0.38	0.028	41.1	22558.42	322.8

African American

	C4			*C4A*			*C4B*			*DRB1**03:01			*B**08:01			rs2105898				
Model	beta	se	-log10(p)	beta	se	-log10(p)	beta	se	-log10(p)	beta	se	-log10(p)	beta	se	-log10(p)	beta	se	-log10(p)	AIC	LRT -log10(p)
C4	-0.51	0.059	17.3																7358.65	19.7
C4A				-0.43	0.062	11.2													7385.17	14
C4A+C4B				-0.62	0.068	18.7	-0.41	0.068	8.6										7351.45	20.9
*DRB1**03:01										0.41	0.091	5.2							7413.36	8
*B**08:01													0.78	0.11	11.6				7387.33	13.6
rs2105898																-0.46	0.047	21.9	7339.35	23.9
*C4A + C4B + DRB1**03:01				-0.59	0.073	15	-0.38	0.071	7.1	0.1	0.099	0.5							7352.34	20.4
*C4A + C4B + B**08:01				-0.51	0.073	11.7	-0.37	0.069	7.2				0.49	0.12	4.4				7337.24	23.6
C4A + C4B + rs2105898				-0.52	0.07	13.2	-0.43	0.069	9.4							-0.42	0.048	17.8	7277.78	36.2

Coefficients (beta, standard error) and *P* values (as $-\log_{10}(P)$) for individual terms composing several relevant logistic regression models for predicting SLE risk in a European ancestry cohort of 6,748 SLE cases and 11,516 controls and an African American cohort of 1,494 SLE cases and 5,908 controls. Each analysis also included ancestry-specific covariates. For each model, the Akaike information criterion (AIC) and overall *P* value (as determined by Chi-squared likelihood-ratio test) are given on the right to indicate the relative strengths of similar models for each ancestry cohort.

nature research

Corresponding author(s): Steven A. McCarroll

Last updated by author(s): Feb 19, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	BioTek ELx800 Gen 5 software was used to collect ELISA absorbance readout on microplates for CSF samples.
Data analysis	Genome STRiP 2.0 was used for C4 copy number calling on whole genome-sequenced samples. BEAGLE v4.1 (21Jan17.6cc) was used for imputation of C4 variation. HIBAG v1.4 was used for HLA allele imputation. R v3.6 was used for downstream analyses and functions were derived largely from default packages (e.x. stats) with the exception of third-party HIBAG and ggplot2 packages.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Individual genotype data for Sjögren's syndrome cases and controls and individual plasma concentrations for C4 and C3 are available in dbGaP under accession number phs000672.v1.p1. Individual genotype data for schizophrenia cases and controls are available by application to the Psychiatric Genomics Consortium (PGC). Questions regarding individual genotype data for SLE cases and controls of European and/or African American ancestry can be directed to Timothy J. Vyse (timothy.vyse@kcl.ac.uk). Data resources (reference haplotypes), software scripts and instructions for imputing C4 alleles into SNP data sets are available at http://mccarrolllab.org/resources/resources-for-c4/. C4 genotype and protein concentration data for CSF samples are available upon request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For genetic analyses, we were maximally inclusive of human-genetic datasets available at the time of analyses, and collaborated internationally to achieve the largest sample sizes we could: e.g. 6,748 SLE cases and 11,516 controls of European ancestry; 1,494 SLE cases and 5,908 controls for African Americans. A strong pre-analysis indicator that these sample sizes would be sufficient, came from the fact that earlier work on these same data sets had already established extremely strong associations to genetic markers at the MHC locus ($p < 1e-100$ among Europeans; $p < 1e-25$ among African Americans). For analyses of the relationship of CSF complement protein levels to sex and age, we sampled from a larger panel of CSF samples so as to include sufficient numbers of samples within the age ranges (20-50) that correspond to sex-biased disease incidence. We used sample sizes that were comparable to or larger than those in previous CSF studies. Evidence that these sample sizes were sufficient came from the strong statistical significance of the results.
Data exclusions	For human-genetic analyses, pre-established QC metrics standard in the field were used to exclude some samples or genotypes for analysis, as described in Methods; these were pre-established criteria similar to those used in most human genetics studies. These included: (i) exclusion of SNPs based on genotyping rate and Hardy-Weinberg equilibrium; (ii) relatedness (genotyped individuals were excluded if we found them to be related to one another, based on predetermined cutoffs for relatedness, such as excluding duplicate samples and close relatives); (iii) any disagreements of annotated characteristics (such as sex or ancestry) with the inference of these same characteristics from genotype data. It was also pre-determined (before ELISA assays) that CSF samples were to be excluded if they had any visually apparent blood contamination.
Replication	Genetic findings were first critically evaluated by analyses finding that results were consistent across two distinct levels of analysis: (i) the copy number of C4A and C4B genes (Fig. 1a); and (ii) the haplotypes formed by C4 structural alleles and flanking SNPs (Fig. 1b). We then replicated the results for SLE by an independent analysis in another cohort. We found that the findings on C4-associated risk levels were consistent (Fig. 2a) across populations (European-ancestry and African American research cohorts) with different ancestries and different patterns of linkage disequilibrium. We further replicated these results by finding the results to be consistent with those in an independent cohort of patients with a closely related illness (Sjogren's, Fig. 1b). Finally, one of the most surprising results (the finding that C4 alleles associated with larger effects in men) was replicated in a distinct illness, schizophrenia (Fig. 3ab). For analyses of complement protein concentrations in men and women, we analyzed two panels of CSF samples which had been collected by different investigators at different hospitals. We found that the finding of sex bias (higher levels in males than females) was consistent across these cohorts and significant in each cohort independently. We also replicated the CSF results in plasma by re-analyzing data from an earlier study.
Randomization	Individuals genotyped for disease associations had been previously organized into cohorts (with matched controls) by disease status and ancestry. SNP genotyping was done in batches, as described in the original publications in which the SNP genotype data were generated. To address the possibility that population stratification or batches could contribute to any results, we utilized a practice (standard in well-powered human-genetic studies that have access to genome-wide SNP data) of addressing such potential influences by calculating the principal components (PCs) of the genotype matrix for each cohort, then using the PC scores as covariates in logistic-regression association analysis. For schizophrenia analyses, for which multiple cohorts of European ancestry had been collected, the sample's collection site was encoded as an additional indicator covariate in logistic regression, to control for variability in diagnostic thresholds.
Blinding	Blinding was accomplished by the use of an ID number for each sample, which was only re-associated with metadata (e.g. donor sex) in the final statistical analysis. Thus, for example, laboratory analyses of CSF proteins occurred in a manner blinded to donor characteristics.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

Polyclonal Antiserum to Human C4 Protein; supplier: Quidel; catalog number: A305; lot number: 903556-1; dilution: 1:1000
Polyclonal Rabbit Anti-Human C4c Complement/FITC; supplier: Dako; catalog number: F016902-2; lot number: 89152; dilution: 1:3000
Goat Anti-Rabbit IgG H&L (Alkaline Phosphatase); supplier: abcam; catalog number: ab97048; lot number: GR166802-2; dilution: 1:5000
Human Complement C3 ELISA Kit; supplier: abcam; catalog number: ab108823

Validation

All antibodies are validated as described in their respective technical data sheets or similar; these statements along with citations can be found on supplier webpages for the product. We also quote highlights from these documents here. For polyclonal antiserum to human C4 protein (Quidel, A305), "Highly purified human C4 was isolated from normal serum and used to immunize goats. The anti-human C4 polyclonal antisera was tested against normal human plasma by double immunodiffusion, one-dimensional immunoelectrophoresis,quantitative radial immunodifussion, and quantitative rocket immunoelectrophoresis. The antiserum was determined to be monospecific for C4 at varying concentrations. Applications of the C4 polyclonal antisera have been evaluated by various research facilities, and include, Western Blot, IHC, Immunofluorescence, and ELISA." For polyclonal rabbit anti-human C4c Complement/FITC (Dako, F016902-2), "The antibody reacts with C4, C4b and C4c, but does not react with the C4d fragment. Traces of contaminating anti-bodies have been removed by solid-phase absorption with human plasma proteins. The specificity has been ascertained as follows:Crossed immunoelectrophoresis:Only reactivity with C4 complement and its C4c-containing fragments is observed when using unconjugated antibody corresponding to 40 uL F-0169 per square cm gel area against 2 uL human plasma. Staining: Coomassie Brilliant Blue. In rocket immunoelectrophoresis the antibody cross-reacts with C4c complement from all of 11 animal species tested so far: Cat, cow, dog, goat, guinea pig, horse, mink, mouse, rat, sheep and swine."

Human research participants

Policy information about studies involving human research participants

Population characteristics

Our analyses included patients of both sexes and multiple ancestries (European and African American). These cohorts have been described in previously published studies (cited in the current work); we summarize their most analysis-relevant characteristics here. We addressed the effects of cryptic (unseen) ancestry by calculating principal components of the genotype matrix and using these as additional covariates in association analysis; this is standard practice in well-powered human genetics studies that have access to genome-wide data. Additional key covariates included sex (men comprised 27% of the European-ancestry SLE cohort, 29% of African American SLE cohort, 10% of the SjS cohort, and 61% of the schizophrenia cohort), collection site/cohort (used in schizophrenia analyses, to account for variation in diagnostic thresholds and ascertainment strategies at sites contributing data to the Psychiatric Genomics Consortium analyses; this was encoded for analysis as a set of indicator variables for membership in each of 40 cohorts), and smoking status (used in Sjögren's syndrome analysis; 12% of the cohort were current smokers). For each disease, genotyped case-control cohorts were as described in prior publications (cited in the current work), in which detailed definitions of phenotypes and associated covariates can be found. Relevant metadata for plasma samples were age (22-89 years old), sex (10% men), and disease status (670 patients were clinically diagnosed with Sjögren's syndrome by meeting the American College of Rheumatology classification criteria for Sjögren's syndrome) and were as described in the dbGaP study with accession number phs000672.v1.p1. CSF sample metadata of age (18-64 years old) and sex (67% men) were recorded upon collection.

Recruitment

For previously-collected samples – including genomic DNA for genotyping (from >40 sites), plasma complement measurements, and one CSF sample panel – recruitment was as described in the previously published studies (cited in the current work). For one set of CSF samples that has not been described in previous papers, CSF was drawn in a hospital context to evaluate for the possibility of CNS infection (cases of confirmed infection were excluded from the collection).

Ethics oversight

Statistical analyses at Harvard Medical School received an NHSR determination from the Harvard Medical School IRB. For previously-collected samples – including genomic DNA for genotyping (from >40 sites), plasma complement measurements, and one CSF sample panel – local IRBs at each institution had approved the collections and patient-consent materials, as described in the earlier papers on these cohorts (cited in the current work). For one set of CSF samples that has not been described in previous papers, the IRB at Brigham and Womens Hospital approved this under protocol #1999P010911.

Note that full information on the approval of the study protocol must also be provided in the manuscript.