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.