**Title:**

# Phenotypes associated with genetic determinants of type I interferon regulation in the UK Biobank: a protocol

**Authors and affiliations:**

Bastien Rioux[1], Michael Chong[2,3,4], Rosie Walker[5], Sarah McGlasson[1], Kristiina Rannikmäe[6], Daniel McCartney[7], John McCabe[8,9], Robin Brown[10], Yanick J Crow[11,12], David Hunt[1*], William Whiteley[1,13*]

[1] Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, United Kingdom

[2] Population Health Research Institute, McMaster University, Hamilton, Ontario, Canada

[3] Thrombosis and Atherosclerosis Research Institute, McMaster University, Hamilton, Ontario, Canada

[4] Department of Pathology and Molecular Medicine, McMaster University, Hamilton, Ontario, Canada

[5] Department of Psychology, University of Exeter, Exeter, United Kingdom

[6] Centre for Medical Informatics, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom

[7] Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, United Kingdom

[8] School of Medicine, University College Dublin, Dublin, Ireland

[9] Department of Medicine for the Elderly, Mater Misericordiae University Hospital, Dublin, Ireland

[10] Department of Clinical Neurosciences, University of Cambridge, Cambridge, United Kingdom

[11] MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, United Kingdom

[12] Laboratory of Neurogenetics and Neuroinflammation, Institut Imagine, Université de Paris, Paris, France

[13] MRC Population Health Unit, Nuffield Department of Population Health, University of Oxford

[*] Last authors with equal contribution

**Corresponding author:**

Prof David Hunt

Centre for Clinical Brain Sciences, University of Edinburgh

Chancellor's Building, 49 Little France Cres, EH16 4SB

Edinburgh, UK

david.hunt@ed.ac.uk

**Keywords:** stroke, dementia, interferonopathy, lupus, inflammation, type I interferon, genetics, variants, UK Biobank

**Word count:** 3,987

## ABSTRACT

**Introduction:** Type I interferons are cytokines involved in innate immunity against viruses. Genetic disorders of type I interferon regulation are associated with a range of autoimmune and cerebrovascular phenotypes. Carriers of pathogenic variants involved in genetic disorders of type I interferons are generally considered asymptomatic. Preliminary data suggests, however, that genetically determined dysregulation of type I interferon responses is associated with autoimmunity, and may also be relevant to sporadic cerebrovascular disease and dementia. We aim to determine whether functional variants in genes involved in type I interferon regulation and signalling are associated with the risk of autoimmunity, stroke, and dementia in a population cohort.

**Methods and analysis:** We will perform a hypothesis-driven candidate pathway association study of type I interferon-related genes using rare variants in the UK Biobank (UKB). We will manually curate type I interferon regulation and signalling genes from a literature review and Gene Ontology, followed by clinical and functional filtering. Variants of interest will be included based on pre-defined clinical relevance and functional annotations (using LOFTEE, M-CAP and a minor allele frequency <0.1%). The association of variants with 15 clinical and three neuroradiological phenotypes will be assessed with a rare variant genetic risk score and gene-level tests, using a Bonferroni-corrected p-value threshold from the number of genetic units and phenotypes tested. We will explore the association of significant genetic units with 196 additional health-related outcomes to help interpret their relevance and explore the clinical spectrum of genetic perturbations of type I interferon.

**Ethics and dissemination:** The UKB has received ethical approval from the North West Multicentre Research Ethics Committee, and all participants provided written informed consent at recruitment. This research will be conducted using the UKB Resource under application number 93,160. We expect to disseminate our results in a peer-reviewed journal and at an international cardiovascular conference.

## STRENGTHS AND LIMITATIONS OF THIS STUDY

- The UK Biobank is the largest whole-exome sequencing project to date, with marked power to detect associations from a limited number of rare, functional variants.
- Our study will leverage current knowledge of interferon biology and genotype-phenotype correlations in Mendelian diseases of type I interferon to test biologically plausible hypotheses.
- The UK Biobank includes phenotypes from multiple sources, which improves classification accuracy for several health outcomes such as stroke and dementia.
- We will carefully select genes and variants with strong evidence of biological relevance to optimize the power of our analyses, which is particularly relevant for less common phenotypes in the UK Biobank such as systemic lupus erythematosus.
- We will increase the specificity of predicted loss-of-function variants by using stringent sample quality control and filtering criteria.

## INTRODUCTION

Interferons are a family of innate inflammatory cytokines primarily secreted by host cells in response to viruses (type I: mainly interferon-α and -β; type II: interferon-γ; type III: interferon-λ). Interferon-stimulated genes are involved in a wide range of processes, namely cellular defence against pathogens, apoptosis, nucleic acid degradation, and cell-to-cell communication [1]. Defects in type I interferon homeostasis are associated with autoimmunity, being implicated in the pathogenesis of systemic lupus erythematosus and other autoimmune disorders such as rheumatoid arthritis, Sjögren's syndrome, and scleroderma [2]. Low-grade type I interferon upregulation may also contribute to sporadic cerebrovascular disease and dementia. Preclinical data suggest type I interferon-related vascular inflammation is an essential contributor to atherosclerosis and may be involved in cerebral small vessel disease [3, 4]. Stroke risk is increased after long-term exposure to exogenous recombinant type I interferon [4, 5], whereas white matter hyperintensities (a radiological manifestation of cerebral small vessel disease), large vessel disease and stroke are more frequent in people with systemic lupus erythematosus as compared to the general population [6, 7].

Genetic type I interferonopathies are a group of rare Mendelian autoinflammatory diseases hypothesised to be caused by an upregulation of type I interferons. Affected individuals with Aicardi-Goutières syndrome, the first type I interferonopathy described, most frequently present in early childhood with progressive encephalopathy, skin vasculopathy, and autoimmunity [8], in addition to prominent white matter hyperintensities, calcifications and large vessel disease (aneurysms, arterial calcifications, stenoses) on brain imaging [9]. Most, albeit not all (e.g., mutations in *IFIH1*, *STING* and *COPA*), pathogenic variants associated with type I interferonopathies result in a loss-of-function (LOF) of key interferon negative regulators inherited as autosomal recessive traits. Carriers of such pathogenic variants are generally considered asymptomatic, although growing evidence from case series suggests they may also exhibit high expression of interferon-stimulated genes [10] and have mild

interferonopathy-related traits [11, 12]. Uncertainty remains, however, as to whether carriers of pathogenic variants in genes involved in type I interferon signalling and regulation have an increased risk of interferonopathy-related phenotypes such as autoimmunity, cerebrovascular disease, and dementia. Moreover, the causal role of type I interferon in sporadic cerebrovascular disease and dementia has not been comprehensively assessed in a population-based study [13], and whether findings from preclinical studies and observations in conditions with impaired interferon homeostasis translate to the general population is unclear.

We will apply a candidate pathway approach to determine whether functional variants in genes involved in type I interferon regulation and signalling are associated with clinical and neuroradiological interferonopathy phenotypes in the general population. We hypothesize that a subset of rare functional variants that result in an upregulation of the type I interferon cascade are associated with core interferonopathy phenotypes.

## METHODS AND ANALYSIS

We will report our results using guidance from the Strengthening the Reporting of Genetic Association Studies (STREGA) initiative [14], and present the protocol checklist in **Supplemental methods 1**. We present a graphical abstract of our protocol in **Figure 1**.

### Study population and exome extraction

We will use data from the UK Biobank (UKB), a large population-based cohort of 502,650 participants mostly of white British ancestry who were aged 40-69 years when recruited from UK patient registries between 2006 and 2010 (response rate: 5.5%) [15, 16]. We will consider individuals with whole-exome sequencing based on the final exome data release (July 2022; n=469,807; 93.5% of participants). The exome was sequenced in two batches composed of the first 50k participants (phase 1) and all other

samples (phase 2). Participants in the first phase were selected to enrich certain phenotypes, which may lead to spurious associations given time-varying sequencing coverage if this batch effect is not controlled (see below).

Genomic DNA samples were sent to the Regeneron Genetics Centre (Tarrytown, New York, USA) as part of a collaboration with the UKB and stored at -80ºC. Genomic libraries with a mean fragment size of 200 base pairs (bp) were created enzymatically and tagged with barcodes of 10 bp before capture. Exome was obtained by next generation sequencing using the Illumina NovaSeq 6000 platform (S2 and S4 flow cells for the first and second phase, respectively) and a target-enrichment probe kit (IDT xGen® Exome Research Panel v1.0) to enable deep and uniform coverage of ~39 Mbp (19,396 genes).

**Whole-exome sequencing data**

We will use the multi-sample project-level VCF (pVCF) files made available by the UKB [17]. To obtain these joint genotype data, raw sequencing outputs (FASTQs) were initially processed into sample-level aligned sequences (CRAMs) with a standard protocol (the Original Quality Functionally Equivalent; OQFE), which maps short sequences to the GRCh38 reference genome with alternate loci and marks duplicate segments [18]. DeepVariant (v0.10.0) was used to call variants from sample-level CRAMs and produce variant call data (gVCF) for each participant [19]. This calling approach uses a deep convolutional neural network to determine the most likely genotype at each locus from the reference genome, base reads and quality scores [20]. It outperforms existing state-of-the-art tools to call single nucleotide variants (SNVs) and small insertions or deletions (indels; up to 50 bp by definition), achieving high overall accuracy (>99.5%) [20, 21]. The variant call data set includes exome capture targets and their immediate flanking regions (100 bp upstream and downstream of each target).

Sample-level variants were aggregated into joint genotype pVCF files with a standard analysis pipeline (GLnexus v1.2.6) [19, 22].

For quality control, we will exclude participants with a mismatch between their genetically recorded and self-reported sex or with sex chromosome aneuploidy (~0.2%). We will apply a set of per-variant quality control metrics as previously employed for the UKB exome to analyse variants with [23]:

i) individual and variant missingness <10%;

ii) Hardy Weinberg equilibrium p-value $>10^{-15}$;

iii) at least one sample per site with allele balance threshold >0.15 for SNVs and >0.20 for small indels;

iv) minimum read coverage depth of seven for SNVs and 10 for indels.

We will also use a sequencing depth ≥10x in 90% of samples for our rare variant analysis, to prevent spurious associations that may result from batch effect [24]. We will resolve haplotype phase with the Segmented HAPlotype Estimation and Imputation Tools version 5 (SHAPEIT5 v1.0.0), which phases rare variants from the UKB with high accuracy (switch error rate <5% with minor allele count >5) [25].

**Genes of interest**

We will apply a hypothesis-driven candidate pathway approach of type I interferon-related genes by adapting a previously described methodology [26]. We will consider for inclusion any gene encoding a protein of interest belonging to one of the three following categories:

1. A negative regulator, positive regulator, or effector along the main signalling pathway of type I interferon (**Figure 2**);
2. A protein directly affecting the activity of an interferon regulator or effector (e.g., E3 ubiquitin-protein ligase TRIM21 inhibits interferon regulatory factor 3, a transcription factor that controls

multiple type I interferon-inducing pathways; both proteins are therefore considered for inclusion);

3. A protein involved in genetic type I interferonopathies (see **Supplemental table 1**).

We did not consider regulatory proteins acting beyond the second order of regulation (e.g., regulators of E3 ubiquitin-protein ligase TRIM21) to adequately balance the need to include important regulators of type I interferon, while maintaining their specificity to type I interferon signalling.

We will produce a preliminary list of genes from i) recently published reviews on type I interferon biology and ii) annotations in Gene Ontology [27]. We present herein both completed and upcoming steps. First, we searched Ovid MEDLINE to identify reviews describing ≥2 proteins of interest in physiological conditions. We used interferons (of any type to increase the sensitivity of our search) and regulation as main concepts, in addition to a previously published hedge for reviews (**Table 1**) [28]. We queried MEDLINE from January 2000 onwards to only include recent reviews, and conducted our search in English as we expected reviews in other languages to present similar information. Our strategy yielded 194 records. A single author (BR) will screen records by title and abstract, and include relevant articles after full-text reading. We will manually add four recent reviews [8, 9, 29, 30] on genetic interferonopathies to ensure these genes are captured. A single author (BR) will extract relevant proteins, their corresponding genes, and their presumed functions.

Second, we have queried the Gene Ontology resource to validate and enrich our gene set. Gene Ontology provides curated gene-specific knowledge with functional annotation and hierarchical relationships [27, 31, 32]. We extracted a list of 194 genes pertaining to 31 ontology terms relevant to type I interferon (**Supplemental table 2**). We will validate presumed gene product function from reviews and Gene Ontology on the UniProt platform [33] and the National Center for Biotechnology Information (NCBI) Gene database [34] before assigning their function (e.g., negative regulator) and

level of action (e.g., downstream to receptors). Discrepancies will be resolved through consensus by three authors with expertise in interferon biology (BR, SM, DH).

From this preliminary list of genes, we will only include those with ≥1 variant associated with a Mendelian disease through any effect on protein function to strengthen their biological relevance. We will search the Online Mendelian Inheritance in Man (OMIM) [35] and the NCBI ClinVar [36] clinical annotation databases for genotype-phenotype correlations. We will validate that all top 21 type I interferon-inducible genes in systemic lupus erythematosus are included in our list, and add missing items [37].

**Variants of interest**

We will include both SNVs and small indels in genes of interest with ≥1 of the following protein effects: i) LOF, dominant-negative, or gain-of-function (GOF) disease-causing variants through an autosomal dominant, recessive or X-linked inheritance [38], or ii) predicted LOF variants from functional annotations. We will define disease-causing variants as those reported in ClinVar (as pathogenic or likely pathogenic, excluding variants with conflicting interpretations of pathogenicity), OMIM (as disease-causing), and from discussion with clinical experts in interferonopathies (DH, SM, YC). The protein-level effect will be determined through comments and linked publications in ClinVar, descriptions in OMIM or, if undetermined, inferred from resulting phenotype.

We will also define a second set of putative functional variants identified in UKB participants to increase our statistical power [39]. We will assess the functional impact of these variants on Ensembl with the Variant Effect Predictor (VEP), an online resource that returns annotations on the effect of variants on transcripts and proteins [40]. We will interpret variant pathogenicity with the Loss-Of-Function Transcript Effect Estimator (LOFTEE) and the Mendelian Clinically Applicable Pathogenicity score (M-CAP v1.4). The LOFTEE filtering criteria will be used to annotate non-

missense predicted LOF variants, as it provides a conservative filtering strategy to increase specificity (e.g., removal of variants predicted to escape nonsense-mediated decay) and was used to annotate variants in the Genome Aggregation Database (gnomAD; a public resource of ~126k high-quality exomes from around the world that does not include UKB data) [41] and Genebass (a public resource of exome-based genotype-phenotype associations within the UKB) [42]. The M-CAP score will be used to interpret pathogenicity and nominate missense variants for inclusion [43]. This supervised learning classifier incorporates nine established pathogenicity likelihood scores (namely SIFT and PolyPhen-2) and achieves substantial reduction in the misclassification rate of known pathogenic variants (<5%) as compared to other existing methods (26-38%) [43]. We will define predicted LOF variants as either i) variants that inactivate a protein-coding gene through a premature stop codon, a shift in the transcriptional frame or an alteration of essential splice-site nucleotides (from LOFTEE), or ii) missense variants classified as likely pathogenic (from M-CAP). We will apply a minor allele frequency (MAF) threshold <0.1% in both the UKB and gnomAD to lower the probability of including benign variants and improve our statistical power. Using a more liberal MAF threshold of <1%, 8.03 million SNVs were identified in ~200k UKB participants, of which 5.4% (~450k) were predicted LOF variants [23]. In gnomAD, which used LOFTEE without MAF threshold, about 40% of genes had >10 predicted LOF variants [41].

**Phenotypes of interest**

We will test the association of selected variants with a set of 15 clinical and three neuroradiological phenotypes of interest in the UKB. These phenotypes were selected based on their frequency in the general population and the UKB, the plausibility of their association with type I interferon upregulation, and from type I interferonopathy clinical presentations (including Mendelian and sporadic diseases). The International Classification of Diseases (ICD) diagnostic codes and UKB fields

for each phenotype are presented in **Supplemental table 3**. Genes associated with ≥1 phenotype of interest will be assessed for their association with 196 clinical phenotypes to help interpret their relevance (**Supplemental table 4**). We manually grouped ICD-coded diagnoses by pathophysiology to reduce multiple testing and improve power for less common conditions. As part of the phenome analysis, we will test two stroke definitions developed by Rannikmäe et al [44] to help explain potential misclassifications.

Health-related outcomes were captured through self-completed questionnaires followed by a nurse-led interview on past medical history (at baseline in all and during follow-up for some participants), as well as data linkage with ICD-coded hospital admissions from National Health Service (NHS) registries (primary or secondary diagnoses; ICD v9 and v10) and national death registries (primary and secondary causes of death; ICD v10). Diagnostic codes from primary care (Read codes v2 and v3) are available in a subset of participants (~45.8%). Cancer diagnoses (ICD v9 and v10) are available through data linkage with national cancer registries. Stroke diagnoses from hospital and death registries have a high sensitivity (point estimate range: 88-94%) and specificity (>99%) [45]. Most stroke cases in the UKB are from hospital and death registries, although ~27% are self-reported without coded diagnosis [44]. Self-reported strokes have a lower sensitivity (79%) but maintain a high specificity (99%) [46]. In-hospital and death records for all-cause dementia in the UKB have a positive predicted value of ~85% [47].

We will define phenotypes in the UKB using algorithmically defined (or adjudicated) outcomes (v2.0), first diagnostic occurrences, and cancer registries. Algorithmically defined outcomes are custom diagnostic classification schemes developed by the UKB from self-reports, hospital admissions and death registries to optimize their positive predictive value. First diagnostic occurrences map clinical terms from all available sources into ICD v10 codes (apart from cancer registries). Algorithmically defined outcomes and first diagnostic occurrences will be combined to identify any stroke, ischemic

stroke, intracerebral haemorrhage, subarachnoid haemorrhage, all-cause dementia, Alzheimer's disease, vascular dementia, Parkinson's disease and myocardial infarction [48]. We chose to combine these two fields to capture primary care events (not included in algorithmically defined outcomes), which is expected to increase the number of cases from 0.7% (n=55) for all-cause dementia to 11.4% (n=1,376) for ischemic stroke. First diagnostic occurrences will be used alone for other non-cancerous conditions. Data from cancer registries will be used to define malignant neoplasms.

We will define three neuroradiological phenotypes: total white matter hyperintensity (WMH) volume, total brain (grey plus white matter) volume, and hippocampal grey matter (mean) volume [49]. Brain magnetic resonance imaging (MRI) scans were obtained in ~42k participants on 3T Siemens Skyra scanners running VD13A SP4 with a standard Siemens 32-channel radio-frequency receiver head coil. The UKB MRI quality control pipeline includes a pre-processing step to correct for head motion and other artifacts followed by automated identification of equipment failure and excessive artifacts [50]. We will normalize WMH and hippocampal grey matter volumes for head size using the UKB scaling factor derived from the external surface of the skull. The normalized total brain volume is available as an imaging-derived phenotype. We will log-transform WMH volumes given their right-skewed (log-normal) distribution.

The total WMH volume of presumed vascular origin per individual was generated by an image-processing pipeline [50] followed by a segmentation algorithm (the Brain Intensity Abnormality Classification Algorithm tool; BIANCA) using both T1- and T2-weighted/fluid-attenuated inversion recovery (FLAIR) sequences [51]. The algorithm results in high volumetric agreement (intraclass correlation coefficient = 0.99) and very good spatial overlap index (dice similarity index = 0.76) with manual segmentation. Total brain (including the cerebellum and the brainstem, as low as space-based brain masking allows) and regional brain volumes were extracted using tissue-segmented images obtained from an automated algorithm (FMRIB's Automated Segmentation Tool; FAST) [52] and

289 passed on to the SIENAX analysis pipeline to accurately measure volumetric phenotypes (relative
290 mean error in brain volume = 0.4%) [53]. We will use the mean hippocampal grey matter volume (from
291 right and left hippocampi) as this radiological marker of hippocampal atrophy is associated with
292 memory loss and progression to Alzheimer's disease [54].
293
294 **Statistical analyses**
295 Our primary (score-based) analysis will test the association of all selected genes modelled into a rare
296 variant genetic risk score (RVGRS) with individual phenotypes. We will regress each phenotype on
297 standardized scores using logistic and linear regressions for binary and continuous outcomes,
298 respectively. We will adapt a previously described methodology [55] to define our score as the
299 weighted sum of the number of variants per individual $i$ and gene $g$ ($V_{i,g}$), given a set of $M$ genes:

300 $$RVGRS_i = \sum_{g=1}^{M} \beta_g V_{i,g}$$

301 Gene-level weights will be allocated from theoretical variant effects on the type I interferon cascade.
302 For example, LOF variants in genes encoding negative regulators will receive a positive weight (+1) as
303 they are expected to upregulate the cascade, whereas those in genes encoding positive regulators or
304 effectors will receive a negative weight (-1). We chose this conservative weighting method given the
305 technical limitations of weighting variants from a transcriptomic signature (unavailable in the UKB) or
306 a proteomic profile (no measurement of type I interferon in the UKB Olink proteomics).
307 Our secondary analysis will test gene-level associations with individual phenotypes using the
308 optimal sequence kernel association test (SKAT-O) framed into SAIGE-GENE+. The SKAT-O test
309 leverages the advantages of burden tests and SKAT through a linear combination of their test statistics,
310 the relative contribution of which are estimated by a correlation term [56]. We chose this method to
311 balance the need to maximise power for genes that have a higher proportion of causal variants
312 satisfying the burden test assumption, while preserving power for genes that may have fewer causal

variants (or variants with heterogeneous effects) despite our filtering strategy. The SAIGE-GENE+ method builds upon SKAT-O to reduce type I error inflation for very rare variants in large biobanks with unbalanced case-control ratios, reduce computational resources and account for sample relatedness [57]. We will use a relatedness coefficient cut-off of ≥0.125 (up to third-degree relatedness) in SAIGE-GENE+, and perform our analyses using the open-source R package *SAIGE* (https://github.com/saigegit/SAIGE). We will include the first 10 genetic principal components in the gene-level (combined with the generalized mixed model approach in SAIGE-GENE+) and RVGRS models to control for population structure [58], in addition to adjusting for age and sex [59]. We will also adjust for scanner site in neuroradiological analyses to control for potential technical confounding [60]. We will run separate gene-level tests for LOF/dominant negative and GOF variants to account for their anticipated opposite effect directions. As SKAT-O is designed to test the overall gene-trait association and does not produce effect sizes, variant-level effects will be obtained through separate logistic and linear regressions to help interpret p-values (as in Genebass). Genetic units with <10 carriers of any variant in the UKB will not be analysed to preserve power.

Our score and individual genes will be tested for their association with each phenotype of interest (n=18), and those with ≥1 statistically significant association with any phenotype of interest will be tested for associations across the phenome (n=196). We will interpret statistical significance in our score-based analysis with a Bonferroni-corrected p-value threshold to account for multiple testing across phenotypes (phenotypes of interest: $0.05/18=2.78\times10^{-3}$; phenome: $0.05/196=2.55\times10^{-4}$). We will interpret statistical significance in our gene-level analysis similarly, with a more stringent correction to account for multiple testing across genes and phenotypes (0.05/[# phenotypes x # gene-level units]) [61]. Our analyses will be conducted on the UK Biobank Research Analysis Platform [62]. We present our pre-planned sensitivity analyses in **Supplemental methods 2**.

337 **Power calculations**

338 We performed a statistical power analysis for our gene-level tests and phenotypes of interest with
339 SKAT-O using the *SKAT* package (v2.2.5) for R. Our results and analysis parameters are presented in
340 **Figure 3** and detailed in **Supplemental methods 3**. Gene-level tests for lupus and vascular dementia
341 have the lowest powers overall although they increase to reasonable values in more optimistic
342 scenarios. Other cardiovascular and inflammatory outcomes have the highest power throughout all
343 scenarios.

344

345 **DISCUSSION**

346 Comprehensive phenotyping of interferonopathy variant carriers may expand the clinical spectrum of
347 genetic type I interferonopathies and help understand the biological relevance of type I interferon
348 dysregulation in the general population. Importantly, large population-based assessments of
349 interferonopathy carriers are lacking. Our study will leverage knowledge of Mendelian diseases of type
350 I interferon to develop an informed, hypothesis-driven candidate pathway approach to investigate the
351 frequency and phenotype associations of low-grade type I interferon dysregulation in the UKB. Our
352 results will help understand the clinical spectrum of genetic type I interferonopathies, and will provide
353 insights into the role of type I interferon in sporadic conditions.

354 Recent meta-analyses of genome-wide association studies (GWASs) have strengthened the case
355 of inflammatory contributors to stroke and dementia. The largest cross-ancestry GWAS meta-analysis
356 on stroke to date identified 89 independent genomic risk loci, of which two newly reported loci were
357 located near or within genes involved in type I interferon regulation or signalling (*PTPN11* and *TAP1*)
358 [63]. A recent large GWAS meta-analysis on Alzheimer's disease and related dementias identified 33
359 known and 42 new genomic loci, for which a pathway analysis exposed significant gene sets related to
360 immunity, including macrophage and microglia activation [64]. The nearest genes of two new lead

variants, *SHARPIN* and *RBCK1*, encode essential components of the linear ubiquitin chain assembly complex (LUBAC), involved in NF-κB activation. Despite these discoveries, GWASs are unable to identify rarer alleles that may carry important information on the biology of complex traits, while most variants in genomic risk loci are mapped outside coding regions and have unknown regulatory functions [65, 66].

**Strengths and limitations**

The UKB is the largest whole-exome sequencing project to date, markedly improving power to detect associations from a limited number of rare, functional variants [67]. Our informed approach will leverage current knowledge on type I interferon biology to reduce noise and test biologically plausible hypotheses. This contrasts with prior hypothesis-free phenome-wide association studies using rare variants in the UKB such as Genebass [42] and PheWAS [68], which did not include clinical annotations, used uncurated phenotypes, introduced greater multiple-testing burden (~4.5k and ~17k phenotypes tested in Genebass and PheWAS, respectively), and often used small sample sizes (as few as 30 cases/phenotype in PheWAS). The UKB also enables phenotyping from multiple sources, improving the classification accuracy for stroke and dementia as compared to studies using minimal phenotyping (e.g., case definition from self-reported dementia in relatives) [64].

Our study, however, will have some limitations. First, we expect that our weighting strategy based on theoretical knowledge will introduce noise into our score. We were technically unable to reliably assign empirical weights because of the lack of relevant transcript or protein measurements in the UKB. We anticipate this noise will be reduced by carefully selecting genes for which variants have a higher likelihood of functional and clinical consequences. We will also test genes individually as an alternative that does not mandate weights. Second, we anticipate some degree of residual pleiotropy through overlapping inflammatory and non-inflammatory pathways despite our careful curation of

genes to increase specificity to the type I interferon cascade. We will, however, explore the relevance of pleiotropic effects on our results with a proteomic sensitivity analysis (**Supplemental methods 2**). Third, although we will optimize our overall power by carefully selecting clinically relevant genes and functional variants, our power will likely remain lower for rarer phenotypes.

## ETHICS AND DISSEMINATION

The UKB has received ethical approval from the North West Multicentre Research Ethics Committee, and all participants provided written informed consent at recruitment. This research will be conducted using the UKB Resource under application number 93,160. We expect to disseminate our results in a peer-reviewed journal and at an international cardiovascular conference.

## FIGURE LEGENDS

**Figure 1. Graphical summary of the study methodology.**

This summary illustrates the three main steps of the study: i) genes of interest will be identified from a literature review and Gene Ontology, followed by clinical and functional filtering, ii) variants of interest will be included based on their clinical relevance and functional annotations, and iii) the association of variants and phenotypes will be tested with a rare variant genetic risk score and gene-level tests. Abbreviations: LOFTEE, Loss-Of-Function Transcript Effect Estimator; M-CAP, Mendelian Clinically Applicable Pathogenicity score; NCBI, National Center for Biotechnology Information; OMIM, Online Mendelian Inheritance in Man; OQFE, Original Quality Functionally Equivalent; pLOF, predicted loss-of-function; SKAT-O, optimal sequence kernel association test; VEP, Variant Effect Predictor. Created with BioRender.com.

**Figure 2. Overview of the interferon cascade.**

Graphical overview of the main steps involved in interferon regulation and signalling. Endogenous nucleases (blue circle sectors) remove nucleic acids (red confetti) that can trigger interferon production. Abnormal accumulation of endogenous material through impaired regulation (box 1) and viral nucleic acids (not shown) can trigger interferon production through linkage to i) toll-like receptor sensors at the cell membrane surface (not shown) and at endosomes, and ii) cytoplasmic sensors (box 2). Interferons are sensed by cell surface receptors specific to types I (heterodimer with subunits IFNAR1 and IFNAR2), II (heterotetramer with two IFNGR1 and two IFNGR2 subunits) and III (heterodimer with subunits IFNLR1 and IL-10R2) ligands. Signal transduction and intracellular signalling through JAK-STAT activates the transcription of interferon-stimulated genes (box 3). Abbreviations: GAS, gamma-activated sequence; IFN, interferon; IRF, interferon regulatory factor; ISG, interferon-stimulated gene; ISRE, interferon-stimulated response element; TLR, toll-like receptor. Created with BioRender.com.

**Figure 3. Power calculations for gene-level tests with phenotypes of interest using SKAT-O.**

The power calculation assumes an $\alpha=1.11x10^{-5}$, a genetic sampling length of 2,962 bp, a MAF <0.1%, an empirical optimal correlation coefficient, and sample sizes observed in the UKB. Abbreviations: AD, Alzheimer's disease; AF, atrial fibrillation; bp, base pairs; BrV, total brain (grey plus white matter) volume; CKD, chronic kidney disease; Dem, all-cause dementia; HipV, hippocampal grey matter volume (average); IBD, inflammatory bowel disease; ICH, intracerebral haemorrhage; IHD, ischemic heart disease; IS, ischemic stroke; MAF, minor allele frequency; PAD, peripheral artery disease; RA, rheumatoid arthritis; SAH, subarachnoid haemorrhage; SCTD, systemic connective tissue disorder; SLE, systemic lupus erythematosus; VascD, vascular dementia; WMHV, total white matter hyperintensity volume.

## TABLES

**Table 1. Ovid MEDLINE search strategy.**

| Line | Entry | Records |
|---|---|---|
| Interferon concept | | |
| 1 | (cytokine* adj (inflammat* or proinflammat*)).tw. | 422 |
| 2 | IFN*.ti. | 15433 |
| 3 | Interferons/ | 25640 |
| 4 | 1 or 2 or 3 | 40607 |
| Regulation concept | | |
| 5 | (regulat* or metabolism or biology or function).ti | 1191801 |
| Review design hedge | | |
| 6 | meta analysis.mp,pt. or review.pt. or search:.tw. | 3563440 |
| Combine concepts | | |
| 7 | 4 and 5 and 6 | 390 |
| 8 | 7 not ((exp animal/ or nonhuman/) not exp human/) | 357 |
| 9 | 8 not (case study/ or case report/) | 356 |
| 10 | limit 9 to dt=20000101-20230110 | 214 |
| 11 | limit 10 to English language | 194 |

This table presents the search strategy conducted in Ovid MEDLINE on 10 January 2023. Abbreviations: adj, adjacent; dt, create date; exp, explode; mp, multi-purpose fields; pt, publication type; ti, text word in title; tw, text word in title and abstract.

## REFERENCES

1. Schneider WM, Chevillotte MD, Rice CM. Interferon-stimulated genes: a complex web of host defenses. *Annu Rev Immunol* 2014;32:513-45. doi: 10.1146/annurev-immunol-032713-120231 [published Online First: 2014/02/22]

2. Crow MK, Olferiev M, Kirou KA. Type I Interferons in Autoimmune Disease. *Annual Review of Pathology: Mechanisms of Disease* 2019;14(1):369-93. doi: 10.1146/annurev-pathol-020117-043952

3. Boshuizen MC, de Winther MP. Interferons as Essential Modulators of Atherosclerosis. *Arterioscler Thromb Vasc Biol* 2015;35(7):1579-88. doi: 10.1161/ATVBAHA.115.305464 [published Online First: 2015/05/09]

4. Kavanagh D, McGlasson S, Jury A, et al. Type I interferon causes thrombotic microangiopathy by a dose-dependent toxic effect on the microvasculature. *Blood* 2016;128(24):2824-33. doi: 10.1182/blood-2016-05-715987 [published Online First: 2016/09/25]

5. de Jong HJI, Kingwell E, Shirani A, et al. Evaluating the safety of beta-interferons in MS: A series of nested case-control studies. *Neurology* 2017;88(24):2310-20. doi: 10.1212/WNL.0000000000004037 [published Online First: 20170512]

6. Gao N, Kong M, Li X, et al. Systemic Lupus Erythematosus and Cardiovascular Disease: A Mendelian Randomization Study. *Front Immunol* 2022;13:908831. doi: 10.3389/fimmu.2022.908831 [published Online First: 2022/06/24]

7. de Amorim LC, Maia FM, Rodrigues CE. Stroke in systemic lupus erythematosus and antiphospholipid syndrome: risk factors, clinical manifestations, neuroimaging, and treatment. *Lupus* 2017;26(5):529-36. doi: 10.1177/0961203316688784 [published Online First: 2017/04/11]

8. Rodero MP, Crow YJ. Type I interferon-mediated monogenic autoinflammation: The type I interferonopathies, a conceptual overview. *J Exp Med* 2016;213(12):2527-38. doi: 10.1084/jem.20161596 [published Online First: 2016/11/09]

9. Crow YJ, Manel N. Aicardi-Goutieres syndrome and the type I interferonopathies. *Nat Rev Immunol* 2015;15(7):429-40. doi: 10.1038/nri3850 [published Online First: 2015/06/09]

10. Rice GI, Kasher PR, Forte GM, et al. Mutations in ADAR1 cause Aicardi-Goutieres syndrome associated with a type I interferon signature. *Nat Genet* 2012;44(11):1243-8. doi: 10.1038/ng.2414 [published Online First: 2012/09/25]

11. Briggs TA, Rice GI, Adib N, et al. Spondyloenchondrodysplasia Due to Mutations in ACP5: A Comprehensive Survey. *Journal of Clinical Immunology* 2016;36(3):220-34. doi: 10.1007/s10875-016-0252-y

12. Gunther C, Kind B, Reijns MA, et al. Defective removal of ribonucleotides from DNA promotes systemic autoimmunity. *J Clin Invest* 2015;125(1):413-24. doi: 10.1172/JCI78001 [published Online First: 20141215]

13. Kelly PJ, Lemmens R, Tsivgoulis G. Inflammation and Stroke Risk: A New Target for Prevention. *Stroke* 2021:STROKEAHA121034388. doi: 10.1161/STROKEAHA.121.034388 [published Online First: 2021/06/25]

14. Little J, Higgins JP, Ioannidis JP, et al. Strengthening the reporting of genetic association studies (STREGA): an extension of the STROBE statement. *Eur J Epidemiol* 2009;24(1):37-55. doi: 10.1007/s10654-008-9302-y [published Online First: 2009/02/04]

15. Hewitt J, Walters M, Padmanabhan S, et al. Cohort profile of the UK Biobank: diagnosis and characteristics of cerebrovascular disease. *BMJ Open* 2016;6(3):e009161. doi: 10.1136/bmjopen-2015-009161 [published Online First: 2016/03/24]

16. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015;12(3):e1001779. doi: 10.1371/journal.pmed.1001779 [published Online First: 2015/04/01]

17. UK Biobank. UK Biobank Whole Exome Sequencing Protocol (September 2021) 2022 [Available from: https://biobank.ndph.ox.ac.uk/showcase/refer.cgi?id=915] accessed 10 October, 2022.

18. Regier AA, Farjoun Y, Larson DE, et al. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nat Commun* 2018;9(1):4038. doi: 10.1038/s41467-018-06159-4 [published Online First: 2018/10/04]

19. Yun T, Li H, Chang P-C, et al. Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics* 2021;36(24):5582-89. doi: 10.1093/bioinformatics/btaa1081

20. Poplin R, Chang P-C, Alexander D, et al. A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology* 2018;36(10):983-87. doi: 10.1038/nbt.4235

21. Lin Y-L, Chang P-C, Hsu C, et al. Comparison of GATK and DeepVariant by trio sequencing. *Scientific Reports* 2022;12(1):1809. doi: 10.1038/s41598-022-05833-4

22. Lin MF, Rodeh O, Penn J, et al. GLnexus: joint variant calling for large cohort sequencing. *bioRxiv* 2018:343970. doi: 10.1101/343970

23. Szustakowski JD, Balasubramanian S, Kvikstad E, et al. Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. *Nature Genetics* 2021;53(7):942-48. doi: 10.1038/s41588-021-00885-0

24. UK Biobank. UK Biobank Whole Exome Sequencing 300k Release: Analysis Best Practices 2021 [Available from: https://www.ukbiobank.ac.uk/media/najcnoaz/access_064-uk-biobank-exome-release-faq_v11-1_final-002.pdf] accessed 10 October, 2022.

25. Hofmeister RJ, Ribeiro DM, Rubinacci S, et al. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. *bioRxiv* 2022:2022.10.19.512867. doi: 10.1101/2022.10.19.512867

26. Sharafeldin N, Slattery ML, Liu Q, et al. A Candidate-Pathway Approach to Identify Gene-Environment Interactions: Analyses of Colon Cancer Risk and Survival. *J Natl Cancer Inst* 2015;107(9) doi: 10.1093/jnci/djv160 [published Online First: 20150613]

27. Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res* 2021;49(D1):D325-D34. doi: 10.1093/nar/gkaa1113

28. Health Information Research Unit (McMaster University). Hedges 2016 [Available from: https://hiru.mcmaster.ca/hiru/HIRU_Hedges_MEDLINE_Strategies.aspx#Reviews] accessed 10 January, 2023.

29. Eleftheriou D, Brogan PA. Genetic interferonopathies: An overview. *Best Pract Res Clin Rheumatol* 2017;31(4):441-59. doi: 10.1016/j.berh.2017.12.002 [published Online First: 2018/05/19]

30. Crow YJ, Stetson DB. The type I interferonopathies: 10 years on. *Nature Reviews Immunology* 2022;22(8):471-83. doi: 10.1038/s41577-021-00633-9

31. Carbon S, Ireland A, Mungall CJ, et al. AmiGO: online access to ontology and annotation data. *Bioinformatics* 2009;25(2):288-9. doi: 10.1093/bioinformatics/btn615 [published Online First: 20081125]

32. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;25(1):25-9. doi: 10.1038/75556

33. UniProt C. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* 2022 doi: 10.1093/nar/gkac1052 [published Online First: 20221121]

34. National Center for Biotechnology Information. Gene 2022 [Available from: https://www.ncbi.nlm.nih.gov/gene] accessed October 14, 2022.

35. McKusick-Nathans Institute of Genetic Medicine - Johns Hopkins University. Online Mendelian Inheritance in Man (OMIM) [Available from: https://omim.org/] accessed November 1, 2022.

36. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* 2018;46(D1):D1062-D67. doi: 10.1093/nar/gkx1153 [published Online First: 2017/11/23]

37. Yao Y, Higgs BW, Morehouse C, et al. Development of Potential Pharmacodynamic and Diagnostic Markers for Anti-IFN-α Monoclonal Antibody Trials in Systemic Lupus Erythematosus. *Hum Genomics Proteomics* 2009;2009 doi: 10.4061/2009/374312 [published Online First: 20091117]

38. Backwell L, Marsh JA. Diverse Molecular Mechanisms Underlying Pathogenic Protein Mutations: Beyond the Loss-of-Function Paradigm. *Annu Rev Genomics Hum Genet* 2022;23:475-98. doi: 10.1146/annurev-genom-111221-103208 [published Online First: 20220408]

39. Shameer K, Tripathi LP, Kalari KR, et al. Interpreting functional effects of coding variants: challenges in proteome-scale prediction, annotation and assessment. *Brief Bioinform* 2016;17(5):841-62. doi: 10.1093/bib/bbv084 [published Online First: 2015/10/24]

40. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol* 2016;17(1):122. doi: 10.1186/s13059-016-0974-4 [published Online First: 2016/06/09]

41. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020;581(7809):434-43. doi: 10.1038/s41586-020-2308-7 [published Online First: 2020/05/29]

42. Karczewski KJ, Solomonson M, Chao KR, et al. Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. *Cell Genomics* 2022;2(9):100168. doi: https://doi.org/10.1016/j.xgen.2022.100168

43. Jagadeesh KA, Wenger AM, Berger MJ, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nature Genetics* 2016;48(12):1581-86. doi: 10.1038/ng.3703

44. Rannikmae K, Rawlik K, Ferguson AC, et al. Physician-Confirmed and Administrative Definitions of Stroke in UK Biobank Reflect the Same Underlying Genetic Trait. *Front Neurol* 2021;12:787107. doi: 10.3389/fneur.2021.787107 [published Online First: 2022/02/22]

45. Woodfield R, Grant I, Group UKBSO, et al. Accuracy of Electronic Health Record Data for Identifying Stroke Cases in Large-Scale Epidemiological Studies: A Systematic Review from the UK Biobank Stroke Outcomes Group. *PLoS One* 2015;10(10):e0140533. doi: 10.1371/journal.pone.0140533 [published Online First: 2015/10/27]

46. Jones WJ, Williams LS, Meschia JF. Validating the Questionnaire for Verifying Stroke-Free Status (QVSFS) by neurological history and examination. *Stroke* 2001;32(10):2232-6. doi: 10.1161/hs1001.096191

47. Wilkinson T, Schnier C, Bush K, et al. Identifying dementia outcomes in UK Biobank: a validation study of primary care, hospital admissions and mortality data. *Eur J Epidemiol* 2019;34(6):557-65. doi: 10.1007/s10654-019-00499-1 [published Online First: 2019/02/27]

48. UK Biobank Outcome Adjudication Group. Algorithmically-defined outcomes (version 2.0): UK Biobank, 2022:28.

49. Shang X, Zhang X, Huang Y, et al. Association of a wide range of individual chronic diseases and their multimorbidity with brain volumes in the UK Biobank: A cross-sectional study. *eClinicalMedicine* 2022;47 doi: 10.1016/j.eclinm.2022.101413

50. Alfaro-Almagro F, Jenkinson M, Bangerter NK, et al. Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *Neuroimage* 2018;166:400-24. doi: 10.1016/j.neuroimage.2017.10.034 [published Online First: 2017/10/29]

51. Griffanti L, Zamboni G, Khan A, et al. BIANCA (Brain Intensity AbNormality Classification Algorithm): A new tool for automated segmentation of white matter hyperintensities. *Neuroimage* 2016;141:191-205. doi: 10.1016/j.neuroimage.2016.07.018 [published Online First: 2016/07/13]

52. Zhang Y, Brady M, Smith S. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans Med Imaging* 2001;20(1):45-57. doi: 10.1109/42.906424

53. Smith SM, Zhang Y, Jenkinson M, et al. Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. *Neuroimage* 2002;17(1):479-89. doi: 10.1006/nimg.2002.1040

54. Mielke MM, Okonkwo OC, Oishi K, et al. Fornix integrity and hippocampal volume predict memory decline and progression to Alzheimer's disease. *Alzheimers Dement* 2012;8(2):105-13. doi: 10.1016/j.jalz.2011.05.2416

55. Lali R, Chong M, Omidi A, et al. Calibrated rare variant genetic risk scores for complex disease prediction using large exome sequence repositories. *Nat Commun* 2021;12(1):5852. doi: 10.1038/s41467-021-26114-0 [published Online First: 2021/10/08]

56. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* 2012;13(4):762-75. doi: 10.1093/biostatistics/kxs014 [published Online First: 20120614]

57. Zhou W, Bi W, Zhao Z, et al. SAIGE-GENE+ improves the efficiency and accuracy of set-based rare variant association tests. *Nat Genet* 2022;54(10):1466-69. doi: 10.1038/s41588-022-01178-w [published Online First: 20220922]

58. Zhang Y, Pan W. Principal component regression and linear mixed model in association analysis of structured samples: competitors or complements? *Genet Epidemiol* 2015;39(3):149-55. doi: 10.1002/gepi.21879 [published Online First: 20141223]

59. Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology* 2009;20(4):488-95. doi: 10.1097/EDE.0b013e3181a819a1 [published Online First: 2009/06/16]

60. Alfaro-Almagro F, McCarthy P, Afyouni S, et al. Confound modelling in UK Biobank brain imaging. *Neuroimage* 2021;224:117002. doi: 10.1016/j.neuroimage.2020.117002 [published Online First: 20200602]

61. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* 2003;100(16):9440-5. doi: 10.1073/pnas.1530509100 [published Online First: 2003/07/29]

62. UK Biobank. Research Analysis Platform: Tools Library 2022 [Available from: https://dnanexus.gitbook.io/uk-biobank-rap/working-on-the-research-analysis-platform/tools-library] accessed November 16, 2022.

63. Mishra A, Malik R, Hachiya T, et al. Stroke genetics informs drug discovery and risk prediction across ancestries. *Nature* 2022 doi: 10.1038/s41586-022-05165-3 [published Online First: 2022/10/01]

64. Bellenguez C, Küçükali F, Jansen IE, et al. New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nature Genetics* 2022;54(4):412-36. doi: 10.1038/s41588-022-01024-z

65. Uffelmann E, Huang QQ, Munung NS, et al. Genome-wide association studies. *Nature Reviews Methods Primers* 2021;1(1):59. doi: 10.1038/s43586-021-00056-9

612 66. Tam V, Patel N, Turcotte M, et al. Benefits and limitations of genome-wide association studies. *Nat*
613 *Rev Genet* 2019;20(8):467-84. doi: 10.1038/s41576-019-0127-1 [published Online First:
614 2019/05/10]
615 67. UK Biobank. Final data release from the world's largest whole exome sequencing project 2022
616 [Available from: https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/news/final-data-
617 release-from-the-world-s-largest-whole-exome-sequencing-project] accessed 10 October, 2022.
618 68. Wang Q, Dhindsa RS, Carss K, et al. Rare variant contribution to human disease in 281,104 UK
619 Biobank exomes. *Nature* 2021;597(7877):527-32. doi: 10.1038/s41586-021-03855-y [published
620 Online First: 2021/08/11]
621

## AUTHORS' CONTRIBUTIONS

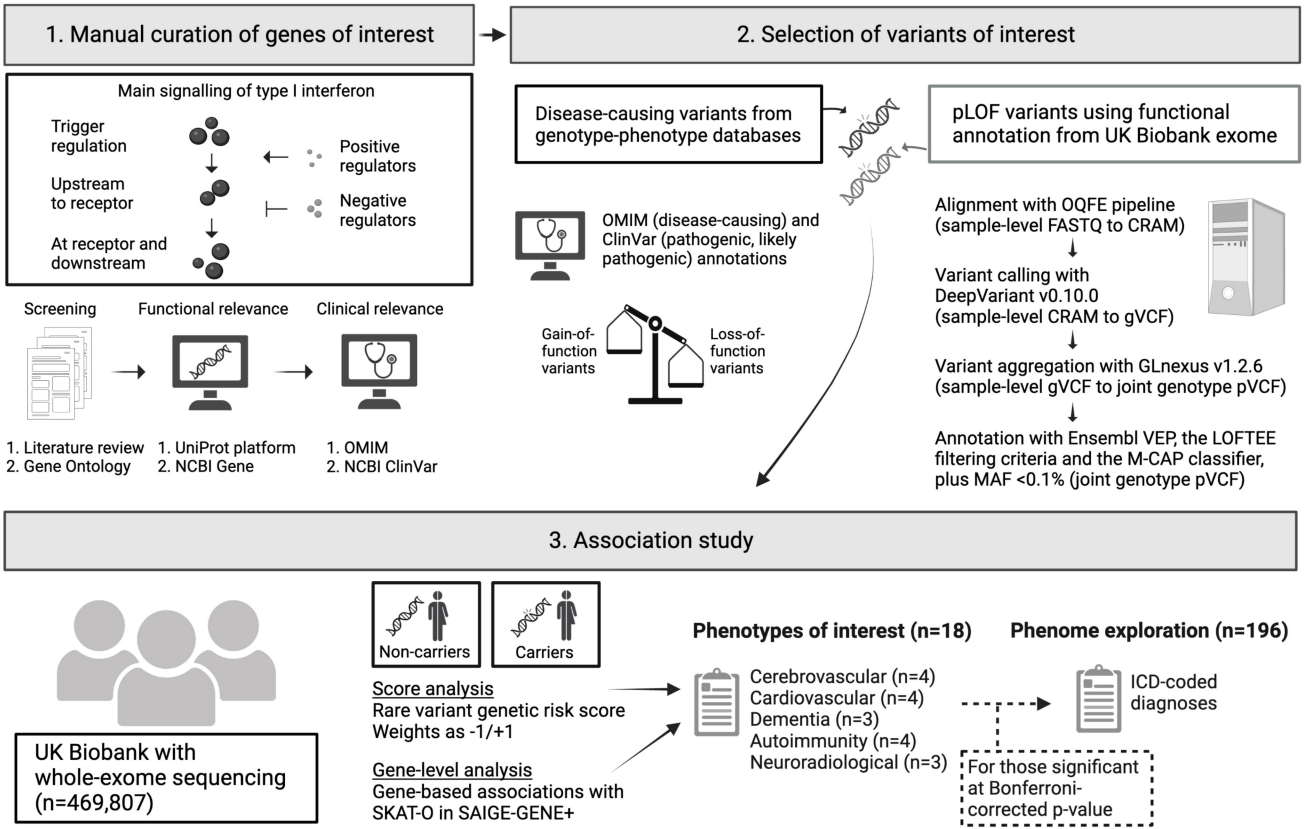
BR planned the study, performed the analyses, and drafted the protocol. WW and DH had a major role in the conception of the study, provided methodological and content expertise, and revised the manuscript. MC, RW, SM, KR, DM, JM, RB and YC provided methodological and content expertise, and had a major role in revising the manuscript. BR is the guarantor of the content of the study.

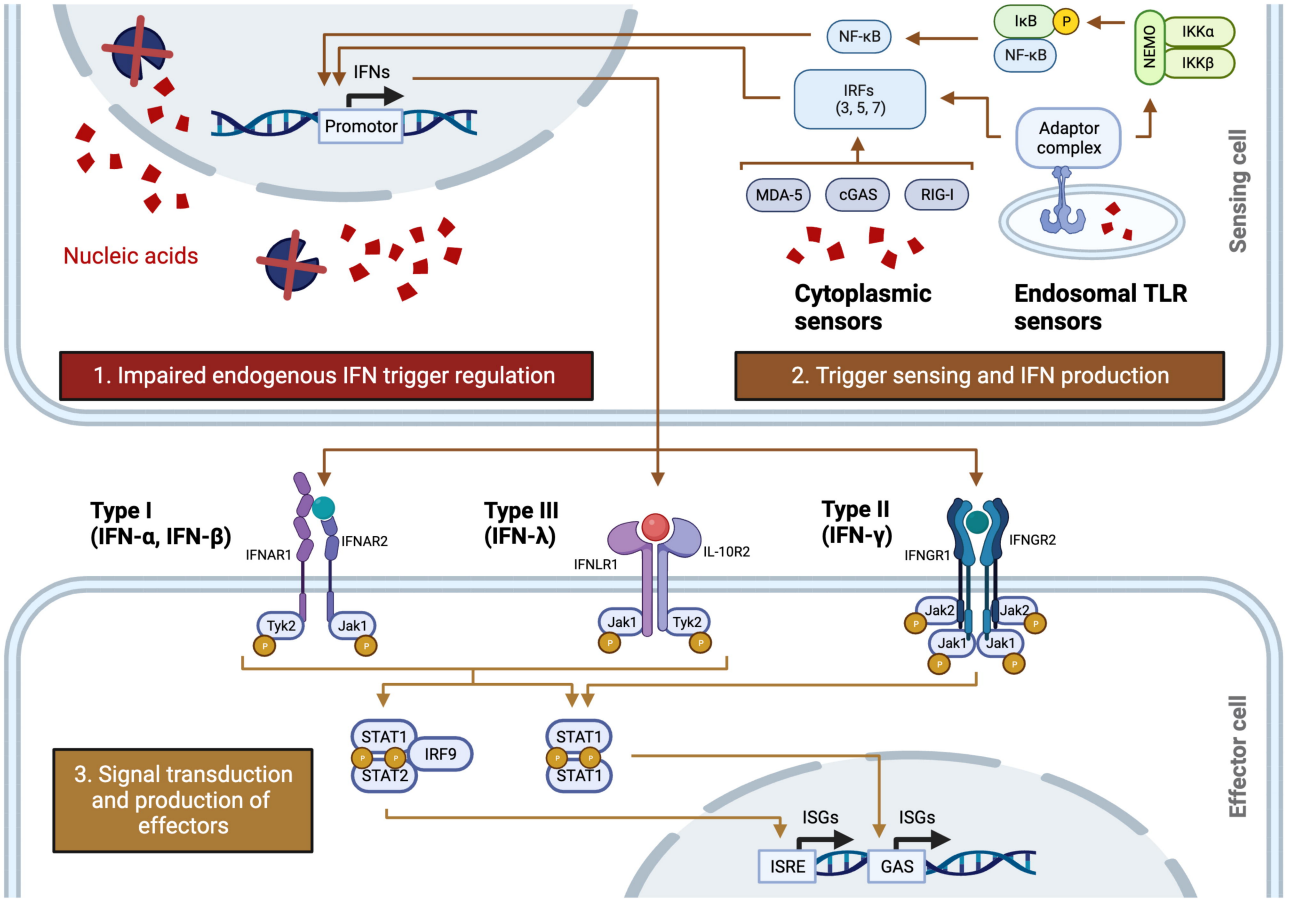
## FUNDING STATEMENT

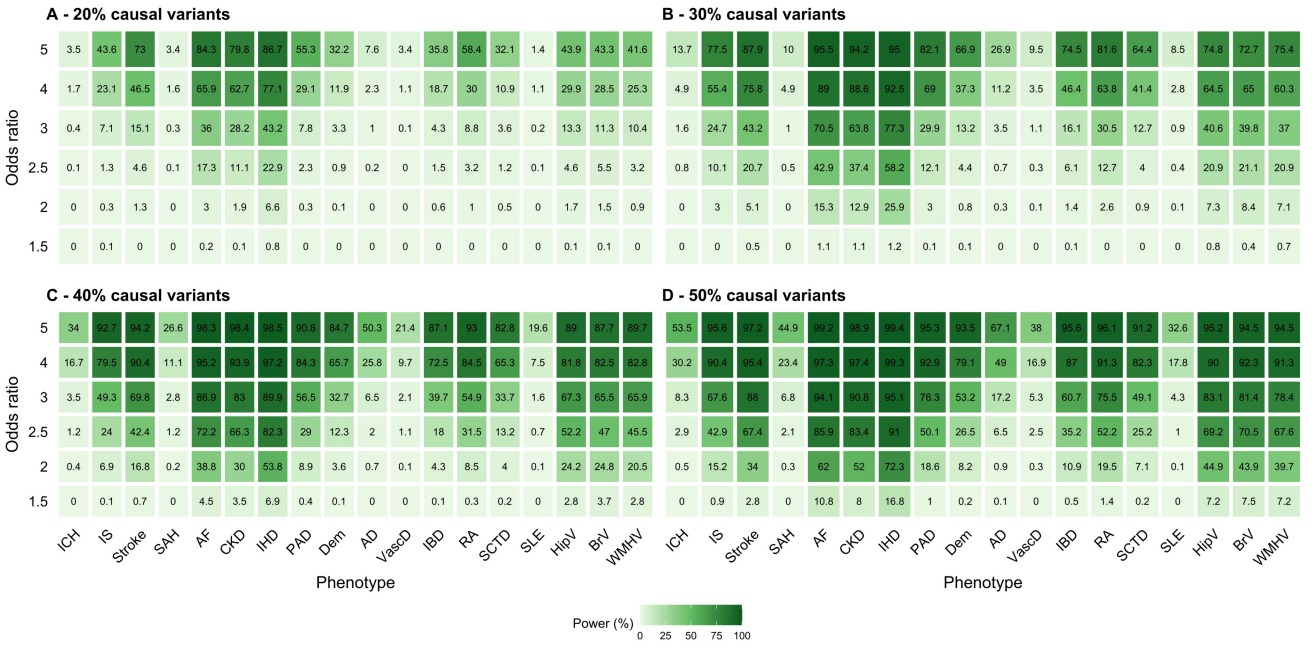
BR is supported by the Centre for Clinical Brain Sciences of the University of Edinburgh (Rowling & Dr Hugh S P Binnie scholarship), the Canadian Institutes of Health Research (CIHR; Doctoral Foreign Study Award, DFD-187711), the *Fonds de recherche du Québec – Santé* and the *Ministère de la Santé et des Services sociaux du Québec* (joint clinician-investigator fellowship), and the Power Corporation of Canada Chair in Neurosciences of the University of Montreal (research scholarship). KR is supported by Health Data Research UK (Rutherford fellowship MR/S004130/1), and the Wellcome Trust-University of Edinburgh Institutional Strategic Support Fund. SM is supported by the Clayco Foundation for RVCL research. DM is supported by the Wellcome Trust (216767/Z/19/Z). RB is supported by an Association of British Neurologists Clinical Research Training Fellowship funded by the Guarantors of Brain. DH is supported by a Wellcome Trust Senior Research Fellowship (215621/Z/19/Z) and the Medical Research Foundation. WW is supported by the Chief Scientist Office of the Scottish Government (CAF/17/01), the UK Alzheimer's Society and the Stroke Association, the National Institute for Health and Care Research (NIHR) and the National Institutes of Health (NIH). Funding sources had no role in the design or conduct of the study.

## COMPETING INTERESTS STATEMENT

All authors have nothing to declare.


1. Manual curation of genes of interest
2. Selection of variants of interest
Main signalling of type I interferon
Trigger regulation
Upstream to receptor
At receptor and downstream
Positive regulators
Negative regulators
Screening
Functional relevance
Clinical relevance
1. Literature review
2. Gene Ontology
1. UniProt platform
2. NCBI Gene
1. OMIM
2. NCBI ClinVar
Disease-causing variants from genotype-phenotype databases
pLOF variants using functional annotation from UK Biobank exome
OMIM (disease-causing) and ClinVar (pathogenic, likely pathogenic) annotations
Gain-of-function variants
Loss-of-function variants
Alignment with OQFE pipeline (sample-level FASTQ to CRAM)
Variant calling with DeepVariant v0.10.0 (sample-level CRAM to gVCF)
Variant aggregation with GLnexus v1.2.6 (sample-level gVCF to joint genotype pVCF)
Annotation with Ensembl VEP, the LOFTEE filtering criteria and the M-CAP classifier, plus MAF <0.1% (joint genotype pVCF)
3. Association study
UK Biobank with whole-exome sequencing (n=469,807)
Non-carriers
Carriers
Score analysis
Rare variant genetic risk score
Weights as -1/+1
Gene-level analysis
Gene-based associations with SKAT-O in SAIGE-GENE+
Phenotypes of interest (n=18)
Cerebrovascular (n=4)
Cardiovascular (n=4)
Dementia (n=3)
Autoimmunity (n=4)
Neuroradiological (n=3)
Phenome exploration (n=196)
ICD-coded diagnoses
For those significant at Bonferroni-corrected p-value


IFNs
Promotor
Nucleic acids
NF-κB
IκB
P
NF-κB
NEMO
IKKα
IKKβ
IRFs
(3, 5, 7)
Adaptor
complex
MDA-5
cGAS
RIG-I
Cytoplasmic
sensors
Endosomal TLR
sensors
Sensing cell
1. Impaired endogenous IFN trigger regulation
2. Trigger sensing and IFN production
Type I
(IFN-α, IFN-β)
IFNAR1
IFNAR2
Tyk2
Jak1
Type III
(IFN-λ)
IFNLR1
IL-10R2
Jak1
Tyk2
Type II
(IFN-γ)
IFNGR1
IFNGR2
Jak2
Jak2
Jak1
Jak1
STAT1
STAT2
IRF9
STAT1
STAT1
3. Signal transduction
and production of
effectors
ISGs
ISGs
ISRE
GAS
Effector cell


A - 20% causal variants
B - 30% causal variants
C - 40% causal variants
D - 50% causal variants
Odds ratio
5
4
3
2.5
2
1.5
3.5 43.6 73 3.4 84.3 79.8 86.7 55.3 32.2 7.6 3.4 35.8 58.4 32.1 1.4 43.9 43.3 41.6
1.7 23.1 46.5 1.6 65.9 62.7 77.1 29.1 11.9 2.3 1.1 18.7 30 10.9 1.1 29.9 28.5 25.3
0.4 7.1 15.1 0.3 36 28.2 43.2 7.8 3.3 1 0.1 4.3 8.8 3.6 0.2 13.3 11.3 10.4
0.1 1.3 4.6 0.1 17.3 11.1 22.9 2.3 0.9 0.2 0 1.5 3.2 1.2 0.1 4.6 5.5 3.2
0 0.3 1.3 0 3 1.9 6.6 0.3 0.1 0 0 0.6 1 0.5 0 1.7 1.5 0.9
0 0.1 0 0 0.2 0.1 0.8 0 0 0 0 0 0 0 0 0.1 0.1 0
13.7 77.5 87.9 10 95.5 94.2 95 82.1 66.9 26.9 9.5 74.5 81.6 64.4 8.5 74.8 72.7 75.4
4.9 55.4 75.8 4.9 89 88.6 92.5 69 37.3 11.2 3.5 46.4 63.8 41.4 2.8 64.5 65 60.3
1.6 24.7 43.2 1 70.5 63.8 77.3 29.9 13.2 3.5 1.1 16.1 30.5 12.7 0.9 40.6 39.8 37
0.8 10.1 20.7 0.5 42.9 37.4 58.2 12.1 4.4 0.7 0.3 6.1 12.7 4 0.4 20.9 21.1 20.9
0 3 5.1 0 15.3 12.9 25.9 3 0.8 0.3 0.1 1.4 2.6 0.9 0.1 7.3 8.4 7.1
0 0 0.5 0 1.1 1.1 1.2 0.1 0.1 0 0 0.1 0 0 0 0.8 0.4 0.7
34 92.7 94.2 26.6 98.3 98.4 98.5 90.8 84.7 50.3 21.4 87.1 93 82.8 19.6 89 87.7 89.7
16.7 79.5 90.4 11.1 95.2 93.9 97.2 84.3 65.7 25.8 9.7 72.5 84.5 65.3 7.5 81.8 82.5 82.8
3.5 49.3 69.8 2.8 86.9 83 89.9 56.5 32.7 6.5 2.1 39.7 54.9 33.7 1.6 67.3 65.5 65.9
1.2 24 42.4 1.2 72.2 66.3 82.3 29 12.3 2 1.1 18 31.5 13.2 0.7 52.2 47 45.5
0.4 6.9 16.8 0.2 38.8 30 53.8 8.9 3.6 0.7 0.1 4.3 8.5 4 0.1 24.2 24.8 20.5
0 0.1 0.7 0 4.5 3.5 6.9 0.4 0.1 0 0 0.1 0.3 0.2 0 2.8 3.7 2.8
53.5 95.6 97.2 44.9 99.2 98.9 99.4 95.3 93.5 67.1 38 95.6 96.1 91.2 32.6 95.2 94.5 94.5
30.2 90.4 95.4 23.4 97.3 97.4 99.3 92.9 79.1 49 16.9 87 91.3 82.3 17.8 90 92.3 91.3
8.3 67.6 88 6.8 94.1 90.8 95.1 76.3 53.2 17.2 5.3 60.7 75.5 49.1 4.3 83.1 81.4 78.4
2.9 42.9 67.4 2.1 85.9 83.4 91 50.1 26.5 6.5 2.5 35.2 52.2 25.2 1 69.2 70.5 67.6
0.5 15.2 34 0.3 62 52 72.3 18.6 8.2 0.9 0.3 10.9 19.5 7.1 0.1 44.9 43.9 39.7
0 0.9 2.8 0 10.8 8 16.8 1 0.2 0.1 0 0.5 1.4 0.2 0 7.2 7.5 7.2
ICH IS Stroke SAH AF CKD IHD PAD Dem AD VascD IBD RA SCTD SLE HipV BrV WMHV
Phenotype
Power (%)
0 25 50 75 100