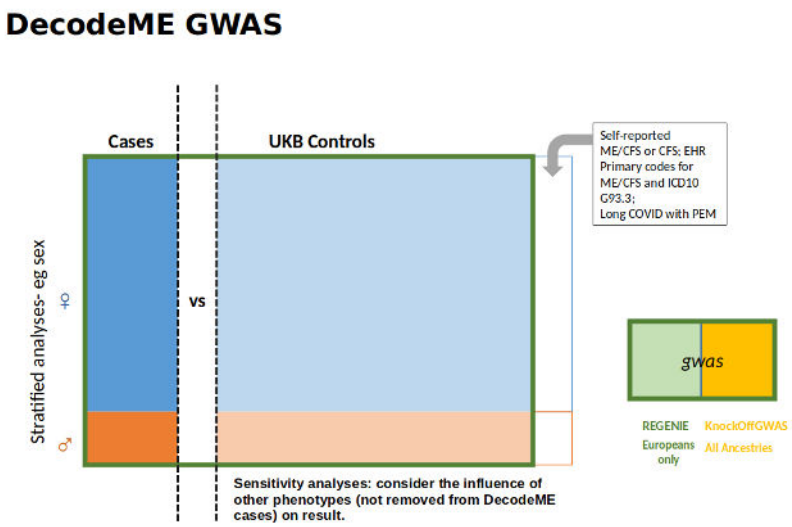
Data Analysis Plan

DecodeME Genetics Delivery Team

2023-03-06

Lay Summary: We will analyse the DNA of up to 25,000 people with ME/CFS and compare this with controls (people without ME/CFS, Long Covid (with PEM) or post-viral fatigue syndrome) from the UK Biobank, to look for genetic differences. Genetic data will arrive in batches of 4,800 samples, so initial analyses will be performed on a smaller sample but will increase over time as more samples become available. These data will be thoroughly quality controlled to minimise false-positive findings in the downstream analyses. We will analyse females and males separately, and combined, and will also stratify by disease onset (i.e. infectious vs non-infectious). To ensure that any observed differences in DNA between people with ME/CFS and controls are not down to 'chance', we set a 'genome-wide significance threshold' which will help us be confident of the validity of any findings. Additionally, to ensure any genetic associations (findings) are not being driven by common co-occurring conditions (including, but not limited to: IBS or Fibromyalgia), further sensitivity analyses will be performed to adjust for this. For further validation, we plan to replicate the findings of DecodeME with an independent group of controls (other than the UK Biobank) but we are yet to find a suitable cohort as of yet.

Technical Summary: Genome-wide analysis study (GWAS) analysis will be performed successively as data from ~4,800 sample batches is delivered. The GWAS will be performed using two methods (logistic regression and a knockoff framework) for all, and for females or males separately. In the main analysis, UK Biobank (UKB) controls will exclude individuals who self-reported ME/CFS or CFS, or are linked to GP codes for ME/CFS or those with ICD10 G93.3 code, or (potentially) those with Long COVID with post exertional malaise (PEM), when data become available. The genome-wide significance threshold used will be $p<5x10^{-8}$ per analysis using variants with minor allele frequency (MAF>1%), e.g. for the first analysis (~4,800 cases), and more stringently, $1x10^{-8}$ when MAF>0.5% are considered with large sample size analyses. Sensitivity and stratified analyses will be deployed to investigate: (a) if a genetic association ("a hit") is driven by a mismatch between cases and controls arising from DecodeME selecting on conditions that often co-occur with ME/CFS; and, (b) if a hit is associated with a particular co-occurring condition or else revealed when subsets of cases are considered. Currently, there is not an appropriate (e.g. sufficiently large) cohort available for replicating findings.



1 Case and control definition

1.1 Cases

1.1.1 DecodeME

Cases are ($\geq$ 16 years old) ascertained based on self-report of a clinical diagnosis of ME/CFS given by a health professional and responses to a set of questions (expanded here). The DecodeME algorithm selects participants based on four sets of answers on: 1) IOM/NAM criteria; 2) CCC criteria; 3) My conditions; and 4) DecodeME exclusionary criteria. The saliva DNA sampled from selected participants is extracted at the UK Biocentre using the Kingfisher DNA extraction platform. Failed extractions are repeated. An aliquot of the extracted DNA will be then sent by batch ($N < 5{,}000$) to ThermoFisher Scientific for genotyping using the UK Biobank Axiom v2 array. This genotyping platform was chosen to match that used in 450,000 UK Biobank participants, the primary convenience controls. A second DNA aliquot is stored for future whole genome sequencing when sufficient funds are found.

The questionnaire was designed to know the COVID-19 status before the clinical diagnosis of ME/CFS as per question 25: "Did you have a (COVID-19) infection when, or just before,

your first ME/CFS symptoms started?" Therefore, the cases can be segregated into two categories pre- or post-Covid. The aim of this study is to recruit up to 20,000 participants with ME/CFS diagnosed pre-Covid and up to 5,000 diagnosed with ME/CFS after contracting COVID-19. The main analysis will be initially conducted on the former subset.

1.1.2 Limitations

On-going recruitment

The DecodeME project is still recruiting participants including those to be genotyped. Therefore, several rounds of GWAS will be performed as genotype batches (about 4,800 individuals as recommended by ThermoFisher, see Section 2) become available. However, data from batches will be pooled if they are expected to arrive within one month of each other.

Sex-ratio

The data collected from the first ~17k DecodeME participants indicates a strongly biased sex-ratio towards women (5: 1) as previously seen in some studies[1]. This might limit the discovery power for the male-only analysis (see 5.1).

Ancestry

The participants can indicate their ethnic group in the questionnaire (question 9). The data collected, so far, suggests a very low participation of self-reported non-Whites. If that number does not increase, then it might not be possible to use REGENIE (1) in separate genome-wide association analysis per ancestry group due to an insufficient number of cases for a given ancestry. However, we will use the KnockOffGWAS (2) method which allows us to analyse diverse and admixed individuals altogether.

[1] ME/CFS is known to be more prevalent among females. Women are also slightly more likely to participate in population cohorts than men.

1.2 Controls

1.2.1 UK Biobank

The controls will be selected from the UK Biobank (UKB), as a general population cohort, but excluding UK Biobank participants with ME/CFS as defined below. Furthermore, to avoid spurious associations, UKB controls will be matched to DecodeME cases with regards to genetically determined ancestry by excluding ancestry outliers in either cases or controls and by fitting ancestry informative covariates in the analysis model.

Potential people in UKB who have ME/CFS will be first identified as those who self-reported chronic fatigue syndrome ("CFS") in the baseline questionnaire (data field 20002) or who answered "Yes" to "Ever had chronic Fatigue Syndrome or Myalgic Encephalomyelitis (M.E.)" in the pain questionnaire (data field 120010). Participants who self-reported CFS but answered "No" to the pain questionnaire (N = 88) will be excluded from acting as UKB controls as will those who provided an ambiguous response ('Do not know' or 'Prefer not to answer'; N = 13).

Electronic health records available in the UK Biobank, such as hospital in-patient data based on ICD-10 codes and GP primary care record information, will be used to perform exclude additional UKB individuals from acting as controls. The G93.3 ICD-10 code (Post-Viral Fatigue Syndrome) or R53 (Malaise and fatigue) are unlikely to be sufficiently specific to ascertain people with ME/CFS in UKB. However, among the UKB participants linked to the G93.3 code (N=1278) there is a majority (65%) of individuals identifying as having ME/CFS. Similarly, these individuals are also more frequently (13%, N = 548) linked to a ME/CFS-relevant primary care code (Table 1) than the remainder of the cohort (0.1%, N= 512). Therefore, UKB participants who present either an ICD-10 code G93.3 (N=449) or a ME/CFS relevant primary care code (Table 1) will be removed from controls (0.09% and 0.1%, respectively) as they report symptoms similar to those manifesting in people with ME/CFS.

1.2.2 Limitations

Age

At the time of recruitment the UK Biobank volunteers were aged between 37 and 73 years (y) old (median 56y) while the DecodeME participants' ages span from 16 to 92y old (median 49y). The narrower age range among the UK Biobank controls means that 24% of DecodeME cases cannot be age-matched. This age mismatch suggests that adjusting for age in the GWAS analysis would be meaningless.

Sex-ratio

As mentioned above, the sex-bias reduces the number of controls who can be used to sex-match with ME/CFS cases. Note, therefore, that sex-stratified analyses on females or on males will not be similarly powered.

1.2.3 Alternative controls

Given these limitations and that genotypes of cases and controls are acquired separately, performing the same analyses with another control set would allow some check of whether results are robust to the choice of controls. Options of alternative controls include from the Genomics England (GEL) 100,000 genomes project, specifically the parents of younger participants, and Genes for Good, imputation from Infinium CoreExome-24 v.1.0 or v.1.1 array data might be considered. Other options (e.g. NextGenScot and ALSPAC) are more limited in size and in geography but might also be considered.

2 Cases Genotype Data QC

The unprocessed genotype data will be returned by ThermoFisher in the format of CEL files (one per sample). The genotypes will be called in batch of up to 4,800 individuals (50x 96-array plates) using the Axiom Analysis suite (AxAS v5.1) following the *Best Practices Genotyping Analysis Workflow* option implemented in the inbuilt library Axiom_UKB_WCSG.r5 (Fig. 1). This workflow performs a series of quality checks on both samples and markers as presented in the Axiom™ Genotyping Solution Data Analysis USER

GUIDE (see here). We will apply the recommended ("default") filtering threshold unless mentioned otherwise. Upon completion, this workflow returns a set of recommended variants. However, special attention will be paid to the sex-linked markers, variants with strong departure from Hardy-Weinberg equilibrium (HWE), rarer variants and mitochondrial variants calls. Therefore, extra QC steps need to be performed to maximize the number of samples and variants. These steps are presented below.

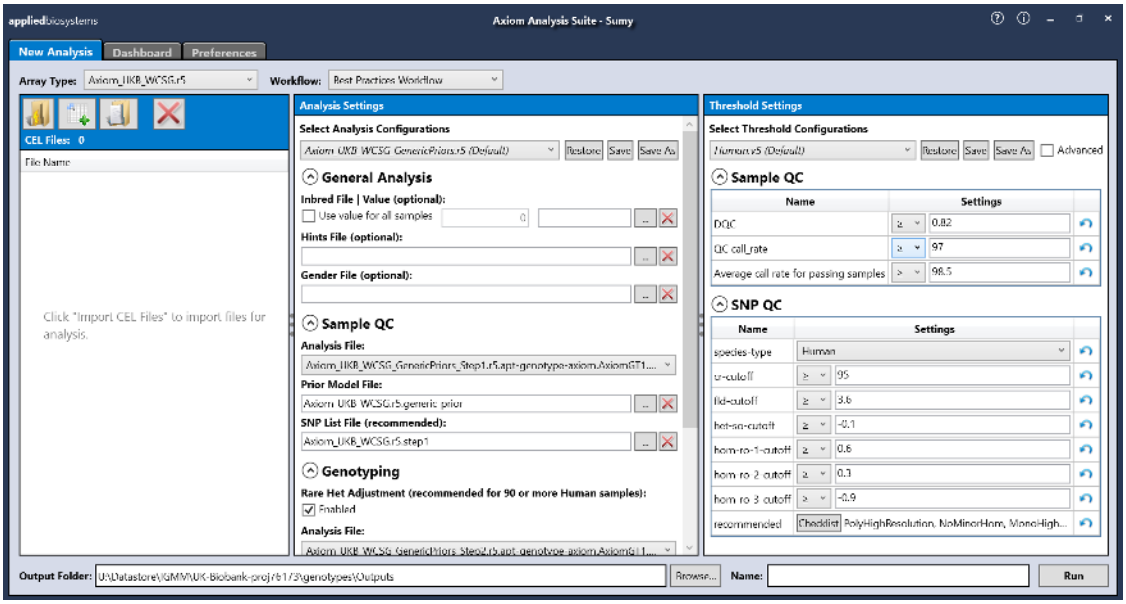


Figure 1: Capture of the Axiom Suite Analysis software

2.1 Extra sample-based quality control

Some samples are flagged as "unknown" sex. This highlights a discrepancy between the self-reported sex and the genetically inferred sex or any sex-chromosome aneuploidy. In the latter case, this can be investigated by conducting a Copy Number Variant (CNV) analysis using the AxAS. The whole genome views for the sex chromosomes of analysed samples are taken and compared to those of males or females with normal sets of chromosomes (3). Samples with an "unknown" sex that remain unresolved will be flagged and removed.

2.2 Extra marker-based quality control

2.2.1 Autosomal and female X chromosome variants

Variants with the lowest call rate or showing the strongest departure from Hardy-Weinberg equilibrium will be visually checked to confirm whether they show poor clustering of genotype calls; thresholds for variants to be exported are set to 90% call rate and HWE p $=10^{-12}$. These are lax thresholds used because ancestry grouping is not done at this stage. Further QC steps are performed outside the Affymetrix platform on the genotypes exported as text files.

2.2.2 Y chromosome and mitochondrial variants

For these variants, we follow the recommended Affymetrix protocol detailed in the Axiom™ Genotyping Solution Data Analysis USER GUIDE (see here). In summary, all the probesets need to be visually inspected to check the following criteria:

- Y chromosome: after all females are set to "No call", no heterozygote genotypes for non-pseudoautosomal Y–linked markers should be observed for each male individual; the expectation is that only two clusters are observed in a population sample for a polymorphic biallelic Y marker.
- Polymorphic Mitochondrial variants are also generally associated with two clusters in population samples.
- When these criteria are not met, the probeset is discarded.

2.3 Population structure

Understanding the genetic structure of a cohort is important to determine the degree of relatedness between participants and to determine their ancestry as these can confound downstream analyses.

2.3.1 Data filtering

To estimate close relatedness (up to first-degree cousins once removed) between individuals and their genetic ancestry, a subset of high quality variants is selected from DNA variants passing step 2.2.1 (see above) by removing:

- Non-autosomal variants
- (A/T or G/C) DNA variants that can introduce DNA strand ambiguity when merging genotype data with the ancestry reference panel
- Samples with call rate < 0.95
- DNA variants with call rate < 0.99
- DNA variants with MAF < 0.01
- DNA variants in high LD listed in (4)

2.3.2 Relatedness

Genetic relatedness will be inferred using KING (5) which implements a kinship measure that does not require population allele frequency estimates. Pairs of samples with a kinship coefficient over 0.04419 (1/2 to the power k+1, upper limit of expected sharing for first cousins once removed, k=4, degree of kinship) will be considered related. Identical samples (kinship coefficient = 0.5), monozygous twins or repeated samples, will be flagged and one of each pair removed. The determination of kinship in samples of uneven representation of ancestry groups can however be distorted (i.e. relatedness of individuals from minor groups will be inflated). Hence, kinship estimates need to be performed using non ancestry-informative variants. These are identified after a first round of PCs analysis (see below).

2.3.3 Ancestry

Ancestry will be inferred by principal component analysis (PCA) by projecting samples onto PCs of a reference population with representative of all major ancestry groups, the 1000 Genomes Phase 3 (https://www.internationalgenome.org/data-portal/data-collection/phase-3). This will be done using the R package bigSNPr (6) which identifies and removes long range LD as source of non-ancestry related discriminative features (e.g. chromosome inversion). The samples subjected to PCA need to be unrelated so that PCs' coordinates do not reflect family structure. However, ancestry-informative variants can confound the relatedness estimation within a given ancestry group. Therefore, an extra step is added into the pipeline (Figure 2) to remove ancestry-informative variants (PC loading > 0.3) as was done for the UK Biobank (4). The ancestry matching between the DecodeME

cases and the UKB controls will be further refined using the PCA-based ancestry grouping method proposed in (6). The PCs will be calculated for the merged case and control sets, and ancestry groups will be defined based on the self-reported country of birth with at least 1,000 persons per group.

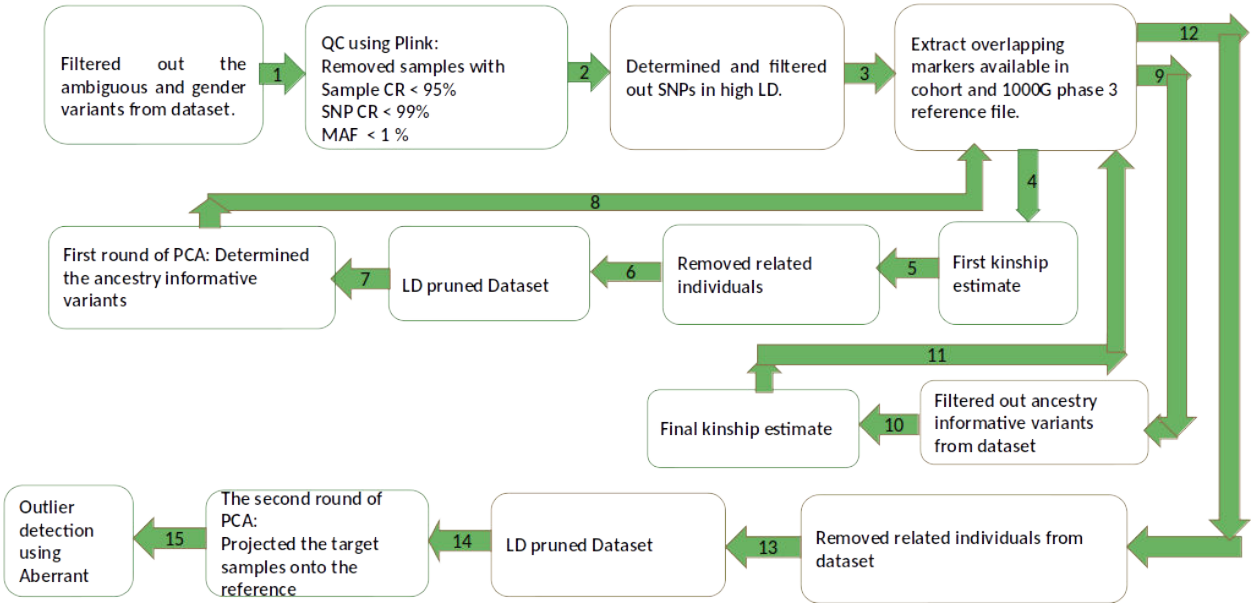


Figure 2: Overview of the PCA-based ancestry grouping. The numbered arrows indicate the step order in the workflow.

Once major ancestry groups are defined it is possible to apply further quality checks within each group to remove individuals with outlying heterozygosity rate or variants departing from Hardy-Weinberg Equilibrium.

Outlying heterozygosity

- Calculate the mean heterozygosity and its standard deviation (sd) in the group.
- Remove samples whose heterozygosity rate is 4sd away from the mean heterozygosity.

Departure from HWE

Remove variants with $HWE < 10^{-7}$ within each of the ancestry groups.

3 Case-Control quality control

3.1 Controls genotyped data recall

Extra care is required to avoid spurious associations arising from correlation structure in the data due to the project's separate genotyping of cases and controls. A batch of UKB controls' (4,700) image files has been reprocessed from scratch using the genotyping call pipeline described above (Section 2) that will be used to call the cases, in order to check the reproducibility between our and UK Biobank's variant calls; only markers passing QC in the independently performed processes and with concordant genotype calls will be kept.

In our QC, concordance with the largest and homogeneous European ancestry group populations' minor allele frequencies for controls will be checked against references from GnomAD, as shown in figure 3.

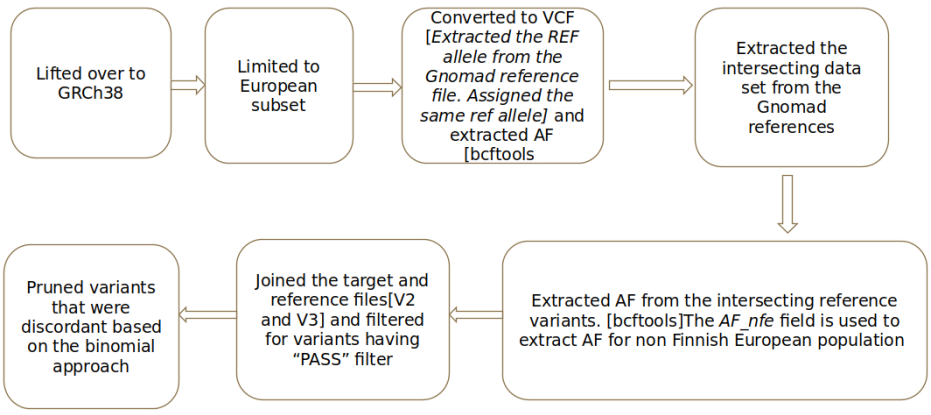


Figure 3: Allele frequency concordance

In addition, the concordance between genotype calls and the whole exome sequences (WES) will also be checked, as the WES data are now available for the entire UK Biobank.

3.2 Quality control on the merged set of case and controls

The DecodeME cases and the matched UK Biobank controls (Section 1.2.1) will be merged and evaluated together.

- Duplicate markers, inconsistencies of chromosome, position, strand or alleles will be identified.
- The degree of relatedness between all the individuals will be estimated using KING (Section 2.3.2)
- Following LD pruning and removal of regions of high LD ($r^2 > 0.8$), principal component analysis (PCA) using unrelated $kinship < 0.04419$ individuals will be used to check for good ancestry matches between cases and controls. If matching is poor along some principal components, variants driving discrimination will be flagged and removed, and PCA will be (re)performed until satisfactory outcomes.
- Trial test of association using PLINK: variants corresponding to top associations will be checked for genotype calls, GnomAD population allele frequencies and consistency of association p-values with linkage disequilibrium (LD) structure.

4 Genotype imputation for cases

Genotype imputation is an important step prior to any genome-wide association study. This process allows us to densify the genetic data for all individuals by inferring unobserved genotypes with a multi-million whole-genome-based reference panel. The imputed genotypes boost the statistical power of GWAS by increasing the chance of capturing the causal risk variant if there is a true association signal in a genomic region. In addition, it also

helps reduce the risk of spurious associations in case-control association studies with outsourced controls such as in the DecodeME study.

4.1 Autosomes and X chromosome

Following UKB best practice (4), autosomal and X-linked genotypes' imputation will be done using two complementary reference panels: the Haplotype Reference Consortium (HRC.r1.1) and the merged UK10K + 1000 Genomes Phase 3. The former helps to yield quality imputation for common and low frequency variants while the latter increases variant number (especially insertions/deletions; InDels) and diversity. Neither of these reference panels is publicly available and the imputation will need to be performed remotely (and securely) using the Sanger Imputation Service (https://www.sanger.ac.uk/tool/sanger-imputation-service/) provided by the Wellcome Sanger Institute (Hinxton, UK).

4.1.1 Data QC and preparation

Prior to imputation, it is essential to check genotype data against the imputation reference panel. For this task, we will use the HRC preparation checking tool developed by W. Rayner (https://www.well.ox.ac.uk/~wrayner/tools/). This tool checks the strand, alleles, position, ref/alt assignments and frequency differences between the genotyped data and the reference panel. The strand, position, ref/alt assignment are updated if necessary. The following criteria are used to filter out DNA variants:

- ambiguous (A/T or G/C) if MAF > 0.4
- differing alleles
- allele frequency difference > 0.2
- not in the reference panel

4.1.2 Phasing and Imputation

Phasing

This is a critical step before imputation which improves both efficiency and accuracy (7). This process estimates the haplotypes, blocks of variants inherited altogether either from the paternal or maternal genome, for each individual. The phasing will be done locally per chromosome using SHAPEIT 4 (8) which implements a reference-based haplotype estimation. Here, the genotyped data will be phased with the 1000 Genome Phase 3 reference panel which is publicly available.

Imputation

The imputation will be performed by the Sanger Imputation service using the PBWT imputation software (https://github.com/VertebrateResequencing/pbwt) as implemented in their server. The phased data for each chromosome will be sorted by genomic position using GRCh37 coordinates (to match reference panels build), then concatenated into a single VCF file, which will be uploaded into the Sanger Imputation server, using Globus (https://www.globus.org/) with encryption, in accordance with data privacy regulation. This process is transient: once the phased data are uploaded, sanity checked and imputed they will be downloaded in the University of Edinburgh secure server and automatically deleted from the Sanger imputation server. The data will not be shared or used for other purpose by the Sanger Imputation Service.

As previously mentioned two complementary reference panels will be used. The data will therefore be imputed with the following panels separately:

- Reference panel 1: HRC (version r1.1 on GRCh37) which contains about 40 million sites from 32,470 samples of predominantly European ancestry.
- Reference panel 2: UK10K+1000Gp3 (build GRCh37) contains 91 million variants from 6,285 diverse samples. It was built using the *-merge_ref_panels* option of IMPUTE2 to merge the two reference panels. The UK10K contains 24 million variants from 3,781 predominantly British samples. The 1000 Genomes phase 3 has 85 million variants provided by 2,504 samples from 26 different populations around the world.

After imputation, variants with a low imputation quality (INFO < 0.4) (9) will be filtered out. Then both imputed datasets will be combined into a single set of imputed genotypes following UKB methods [1]: after filtering a variant is kept if it is present in only one imputation, or if seen in both imputations, only the HRC one will be kept.

Before processing the DecodeME genotyped data, we will do these steps for the test batch of 4,700 UKB samples previously analysed (Section 3) to check for concordance between variants imputed in house and the imputed variants provided by UK Biobank. Any discordant markers will be flagged and removed. Subsequent to this, the genotype call rates might change, and additional variants or individuals not reaching the missingness thresholds for GWAS analysis (SNP < 99% call rate and Individual call rate < 95%) removed.

Each new batch of DecodeME genotypes will be pooled with the previous ones to allow them to be imputed together.

4.2 HLA

Classical HLA alleles will be imputed using the HLA*IMP:02 algorithm as previously done for the UK Biobank (4).

4.3 mtDNA

Imputation of mitochondrial DNA (mtDNA) will follow the methodology proposed in (10).

4.4 CNV

Known copy number variants (CNVs) will be called using the dedicated software PennCNV. After calling, the CNV will be quality controlled and analysed following the recommendations shown in (11).

5 Association analysis

5.1 Merge the cases and control imputed data

The DecodeME cases and the UKB controls will be both imputed separately (the latter done by the UKB) using the same reference panel (HRC combined with UK10K+1000Gp3) and the same methods (Section 4.1.2). Before performing any case and control association analysis these two sets will be merged and subsequent quality checks will be done as shown in section 3.2. The principal components will be calculated on the merge set using bigsnpr (6). We will carry out further quality control of the genotyped and imputed data by doing genome-wide association studies on the blood type provided by the questionnaire (Q10).

5.2 Analysis plan

Different genome-wide association studies (GWAS) will be performed:

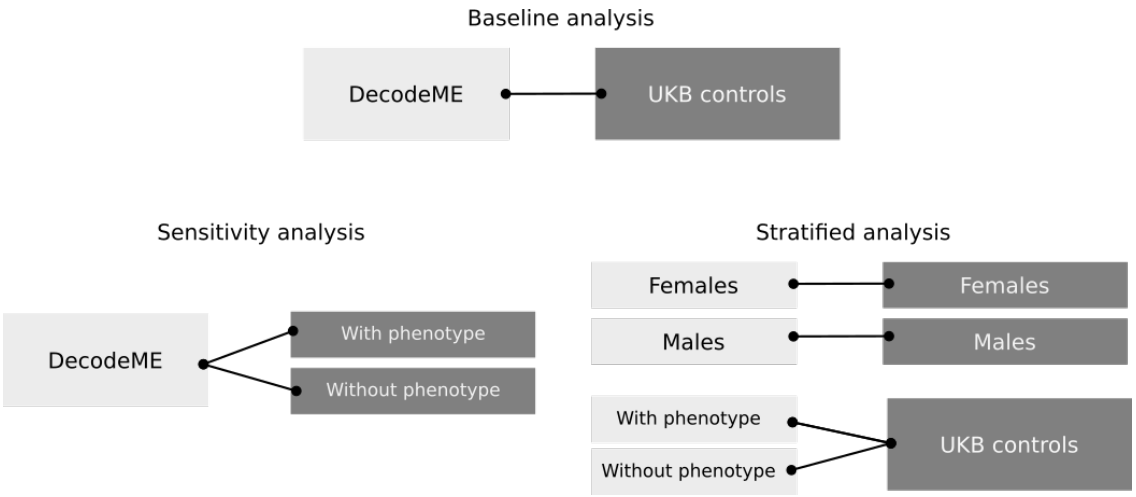


Figure 4: Overview of the different analyses

5.2.1 Main analyses

This GWAS ("gwas-1") will be our **main** analysis for which we would like to keep a general population setting by *only* removing from controls: (a) individuals who are ME/CFS cases in

the UK Biobank based on baseline and pain questionnaire responses, and/or (b) have the hospital record ICD-10 code G93.3 (Section 1.2), and/or (c) have a report of long Covid with Post-Exertional Malaise (information expected to be available in Q4 2023), and/or (d) have ME/CFS relevant primary care records.

5.2.2 Sensitivity analyses

Sensitivity analysis allows us to test whether a statistically significant GWAS signal arises only from subsets of controls. If so, then this subset of controls could wrongly induce association to phenotype A because of confounding by phenotype B. These analyses are performed by carrying out a GWAS with and without controls who match phenotype B (Fig. 4) while keeping the same cases. This will be done only:

1. If a significant variant or a signal lies in a genomic region that was previously associated with another trait measured in UK Biobank, or
2. On ME/CFS co-occurring conditions that (a) are defined in the DecodeME questionnaire and (b) has an equivalent phenotype in UK Biobank (Table 2).

For a subsequent analysis, we will exclude from controls any individual who has any less direct evidence for ME/CFS or post-viral illness. For example, this could be individuals linked with the ICD-10 code R53 (Malaise and fatigue).

5.2.3 Stratified analysis

Stratified GWAS are carried out to determine whether genetic variants are specifically associated with a subset of the samples (Fig.4). First, both analyses (above) will be stratified by sex-at-birth (Part 2 Q8 in the Questionnaire) for both cases and controls. Controls for the first analysis will be sex-matched, while for the second they will be split proportionally to the cases to have a similar case-control ratio across the different stratified GWAS. Any other stratification will be done on cases, only for features, such as infectious disease onset status (Q25 in the Questionnaire), or co-occurring conditions (e.g.: Irritable bowel syndrome, fibromyalgia), with at least 1000 samples per strata.

5.2.4 Combined analysis

As previously mentioned, DecodeME participants will have their DNA genotyped and imputed following the UK Biobank's standard procedure. This give us the possibility to combine into a single set DecodeME cases with UKB participants with evidence of a ME/CFS diagnosis. We will perform analysis with this combined set (against UKB controls; Fig. 5) which would boost the power of discovery for variants enriched in both sets.

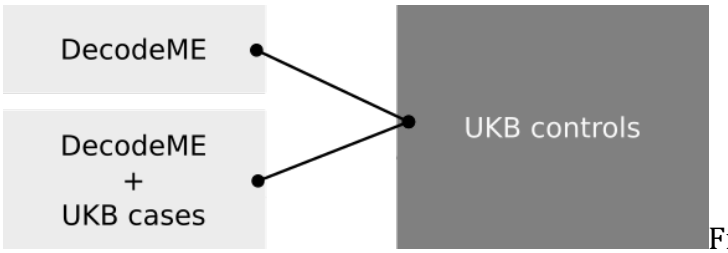


Figure 5: Overview of the combination analysis

5.3 Association testing method

The two following association testing methods will be applied.

5.3.1 REGENIE

REGENIE (1) is a machine-learning method performing whole-genome regression on both quantitative and binary phenotypes on data from a large number of individuals. The approaches implemented in REGENIE allow us to account for co-factors or covariates that could influence ME/CFS risk and/or confound case-control genetic associations (sex, the ancestry PCs explaining most of the inter-individual variation, multiple deprivation index etc), with additional fitting of a polygenic random effect that accounts for cryptic and non-cryptic relatedness. We will use the Firth logistic regression implemented in REGENIE which is robust to case-control imbalance.

The initial REGENIE analysis will be performed on cases and controls of European genetic ancestry only. This could be extended to other ancestries provided we have sufficient cases, but not to admixed samples. Then, the separate ancestry GWAS could be meta-analysed using a trans-ethnic approach as implemented in MANTRA (11).

5.3.2 KnockOffGWAS

KnockOffGWAS (2) is a multivariate knockoff-filters-based method using a sparse regression (lasso) of binary, or quantitative, phenotypes on individual haplotypes. This algorithm has the advantage of accounting for covariates, relatedness, population structure, ancestry and admixture. However, it remains limited to genotyped data as imputed genotypes' uncertainty is not well suited for this method.

5.4 Statistical significance threshold

In any genome-wide association study the statistical significance threshold is critical to differentiate true positive from false positive genotype-phenotype associations. The nominal accepted p-value (i.e. probability of association being a false positive) of 0.05 when only one test is carried out needs to be corrected for multiple testing (millions of variants) using the Bonferroni correction. Therefore, the significance threshold is set to $5x10^{-8}$ for the analysis using $MAF > 1\%$ for the first analysis with a batch of 4,800 cases and to 10^{-8} if $MAF > 0.5\%$ for larger sets.

5.5 Replication and Validation

5.5.1 Staged GWAS for the ME/CFS cohort

- Perform a discovery GWAS using the first batch of genotyped cases ($n = 4{,}800$: the maximum number of individuals per batch), for practical reasons, and matching UKB controls.
- Test for replication of variants that are significant (see thresholds in section 5.4) in subsequently collected cases accounting for multiple tests corresponding to the number of loci put forward for replication.
- Joint analysis of all DecodeME cases available and matched UKB controls.

5.5.3 Independent replication

A replication of significantly associated DNA variants using independent cohorts will be important to replicate, or not, true associations. It also helps highlight potential technical issues such as genotyping or genotype-calling error.

The Scottish Health Research Register and Biobank (SHARE) (12) (700 people with ME/CFS but unclear whether genotyped) could be a good cohort with matching ages. Alternatively, the USA-based cohort, All Of Us, could also provide a good solution but access is currently restricted. Other cohorts such as 23andMe, Genes for Good and Genes and Health that include a question on CFS diagnosis could be used although their case definitions might not be reliable. Currently, there is not an appropriately sized replication cohort available.

5.6 GWAS interpretation and limitation

For each of the hundreds of thousands of variants tested with the trait, the GWAS analysis outputs summary statistics (p-value, effect size and its standard error for each variant tested) that indicate what loci are associated with the trait. However, further analyses are required to prioritise causal variants from a large number of variants showing correlated associations, and to show how they might exert their function; to prioritize the target gene affected if they are regulatory variants; characterize the regulatory region affected in the locus, if any, to point to a possible relevant altered biological pathway. Before this analysis is described in Section 6, it is important to highlight how a *hit* is defined and the limitations of GWAS we could encounter in this project.

What is a hit?

A hit corresponds to a genomic region where a GWAS signal has been detected, with associations reaching the significant p-value threshold (as defined in Section 5.4). A hit may disappear (lying below the significance threshold) among the different analyses performed within this study, e.g. in stratified analyses where the power of GWAS is reduced due to a

lower sample size. To investigate whether the changes observed are due to power issue, we can compare the effect size of the lead DNA variants between the different GWAS. If the difference in effect size of a given DNA variant (lead DNA variants usually) between two analyses is significant (Student's t-test p-value < 0.05) (13) then changes are considered meaningful.

Limitations

Ascertainment of the UKB controls is patchy (different sources), limited (pain questionnaire on a fraction of participants, on-going curation of EHR) and low resolution in some cases. For example, the DecodeME questionnaire includes the active or inactive status of comorbidities but this level of information is not available in the UK Biobank. Therefore, it is not possible to fully match the cases and the outsourced controls. For this reason, it is considered better to carry out the main GWAS (here gwas-1) with a general population control to limit the risk of spurious association (false-positive).

6 Post-GWAS analysis

6.1 Visualisation

The first step of post-GWAS analysis is to visualise the analysis outcome in Manhattan and QQ plots. The former shows the genomic position and strength of association (-log10 p-values) for each tested variant. QQ plots show whether the observed p-values deviate from the expected p-value under the null hypothesis (no association); a deviation reflects the presence of causal effect(s). We will use the convenient online platform LocusZoom.org which generates QQ plots and Manhattan plots that are contextual (gene annotations and local linkage disequilibrium (LD) patterns), interactive, zoomable and shareable.

6.2 Functional annotation

Functional variant annotation is a crucial step for interpreting GWAS results and prioritize DNA variants. First, it contextualizes associated loci by mapping the surrounding genes and the local LD patterns (see above). Second, it can show the effect of variants on genes, transcripts, protein and regulatory region.

For that purpose, we will use FUMA (14) an integrative web platform that performs extensive functional annotation for all DNA variants in genomic areas identified by lead variants using multiple resources. FUMA also provides another convenient function for annotating genes according to their biological context.

Additional post GWAS analyses can be performed if ME/CFS shows sufficient genetic underpinnings (a heritability estimate will be valuable output from the GWAS analysis) and signals detected are strong.

6.3 LD score regression

LD score regression (LDSC) (15) is a tool using GWAS summary statistics to estimate the tested trait heritability. It can also be used to estimate the genetic correlation between the phenotype of interest with other traits. Genetic correlation between ME/CFS and relevant traits (hit driven or the ones use for stratification analysis) will be done whenever possible.

6.4 Fine-mapping

Fine-mapping is a statistical process for defining the credible set of variants, i.e. those that could cause the association signals, which also ranks these variants by statistical support for causality. Each significantly associated loci (i.e. hits; see Sections 5.4 and 5.6) will be systematically fine-mapped to pinpoint most likely causal variants using Bayesian tools such as FINEMAP (16) or SuSie (17) that can handle multiple causal variants in proximity.

6.5 Colocalisation and Mendelian randomization

Colocalisation (18) and Mendelian randomization (MR) are statistical methods aiming to test if two traits share a genetic cause. Mendelian randomization tests whether an exposure

might have a causal effect on an outcome using one or more genetic variants as instrumental variables. Colocalisation tests whether association signals shared by two traits are caused by the same variants. These two methods are based on different frameworks but share some similarity and are complementary (19).

To test if the GWAS signals are shared with expression quantitative trait locus (eQTL) data from multiple tissues we will use the summary data-based Mendelian randomization (SMR) (20), and the heterogeneity in dependent instruments will be tested with (HEIDI) (20). There are other MR tools available that could be used if necessary. Colocalisation will be performed using the R package coloc (21). This approach is applicable to other molecular quantitative trait locus (molQTL) data, such as splicing quantitative trait locus (sQTL) or protein quantitative trait locus (pQTL).

Links

UK Biobank: https://www.ukbiobank.ac.uk/ NIHR: https://www.ukbiocentre.com/ TOPMed server: https://imputation.biodatacatalyst.nhlbi.nih.gov/#! PennCNV: https://penncnv.openbioinformatics.org/en/latest/ PLINK: https://www.cog-genomics.org/plink/ REGENIE: https://rgcgithub.github.io/regenie/overview/ KnockOffGWAS: https://msesia.github.io/knockoffgwas/

References

1. Mbatchou J, Barnard L, Backman J, Marcketta A, Kosmicki JA, Ziyatdinov A, et al. Computationally efficient whole-genome regression for quantitative and binary traits. Nat Genet. 2021 Jul;53(7):1097–103.

2. Sesia M, Bates S, Candès E, Marchini J, Sabatti C. False discovery rate control in genome-wide association studies with population structure. Proc Natl Acad Sci. 2021 Oct 5;118(40):e2105841118.

3. Forgetta V, Li R, Darmond-Zwaig C, Belisle A, Balion C, Roshandel D, et al. Cohort profile: genomic data for 26 622 individuals from the Canadian Longitudinal Study on Aging (CLSA). BMJ Open. 2022 Mar 1;12(3):e059021.

4. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018 Oct;562(7726):203–9.

5. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010 Nov 15;26(22):2867–73.

6. Privé F, Aschard H, Carmi S, Folkersen L, Hoggart C, O'Reilly PF, et al. Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. Am J Hum Genet. 2022 Jan 6;109(1):12–23.

7. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet. 2012 Jul 22;44(8):955–9.

8. Delaneau O, Zagury JF, Robinson MR, Marchini JL, Dermitzakis ET. Accurate, scalable and integrative haplotype estimation. Nat Commun. 2019 Nov 28;10(1):5436.

9. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010 Jul;11(7):499–511.

10. Yonova-Doing E, Calabrese C, Gomez-Duran A, Schon K, Wei W, Karthikeyan S, et al. An atlas of mitochondrial DNA genotype–phenotype associations in the UK Biobank. Nat Genet. 2021 Jul;53(7):982–93.

11. Auwerx C, Lepamets M, Sadler MC, Patxot M, Stojanov M, Baud D, et al. The individual and global impact of copy-number variants on complex human traits. Am J Hum Genet. 2022 Apr 7;109(4):647–68.

12. McKinstry B, Sullivan FM, Vasishta S, Armstrong R, Hanley J, Haughney J, et al. Cohort profile: the Scottish Research register SHARE. A register of people interested in research participation linked to NHS data sets. BMJ Open. 2017 Feb 1;7(2):e013351.

13. Huffman JE, Albrecht E, Teumer A, Mangino M, Kapur K, Johnson T, et al. Modulation of Genetic Associations with Serum Urate Levels by Body-Mass-Index in Humans. PLOS ONE. 2015 Mar 26;10(3):e0119752.

14. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun. 2017 Nov 28;8(1):1826.

15. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015 Mar;47(3):291–5.

16. Benner C, Spencer CCA, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics. 2016 May 15;32(10):1493–501.

17. Zou Y, Carbonetto P, Wang G, Stephens M. Fine-mapping from summary data with the "Sum of Single Effects" model. PLOS Genet. 2022 Jul 19;18(7):e1010299.

18. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLOS Genet. 2014 May 15;10(5):e1004383.

19. Zuber V, Grinberg NF, Gill D, Manipur I, Slob EAW, Patel A, et al. Combining evidence from Mendelian randomization and colocalization: Review and comparison of approaches. Am J Hum Genet. 2022 May 5;109(5):767–82.

20. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet. 2016 May;48(5):481–7.

21. Wallace C. A more accurate method for colocalisation analysis allowing for multiple causal variants. PLOS Genet. 2021 Sep 29;17(9):e1009440.

Appendix

Table 1 Primary care codes for diagnoses associated with ME, CFS or Post-viral fatigue syndrome

Code	Diagnostic
F03y.	Other causes of encephalitis (& [myalgic encephalomyelitis] or [encephalomyelitis NOS]) Other causes of encephalitis Encephalomyelitis NOS Myalgic encephalomyelitis
XE17Z	Postinfective encephalitis (& [myalgic encephalitis] or [myalgic encephalomyelitis]) Post-infectious encephalitis Post-infectious encephalitis Myalgic encephalitis Postinfective encephalitis
XE17b	Encephalitis/myelitis: [NOS] or [encephalomyelitis & (myalgic)] Encephalomyelitis Myalgic encephalomyelitis Encephalitis/myelitis NOS
Xa01F	Chronic fatigue syndrome Myalgic encephalomyelitis ME Myalgic encephalomyelitis Myalgic encephalomyelitis syndrome Postviral fatigue syndrome PVFS - Postviral fatigue syndrome CFS - Chronic

Code	Diagnostic
	fatigue syndrome
.F122	Postinfective encephalitis (& [myalgic encephalitis] or [myalgic encephalomyelitis]) Post-infectious encephalitis Myalgic encephalomyelitis Myalgic encephalitis Postinfective encephalitis
.F38.	Chronic fatigue syndrome Myalgic encephalomyelitis ME - Myalgic encephalomyelitis Myalgic encephalomyelitis syndrome Postviral fatigue syndrome PVFS - Postviral fatigue syndrome CFS - Chronic fatigue syndrome
F286.	Chronic fatigue syndrome Myalgic encephalomyelitis Myalgic encephalomyelitis ME - Myalgic encephalomyelitis Myalgic encephalomyelitis syndrome Postviral fatigue syndrome PVFS - Postviral fatigue syndrome PVFS - Postviral fatigue syndrome CFS - Chronic fatigue syndrome
X75s8	Chronic fatigue syndrome Myalgic encephalomyelitis ME - Myalgic encephalomyelitis Myalgic encephalomyelitis syndrome Postviral fatigue syndrome PVFS - Postviral fatigue syndrome CFS - Chronic fatigue syndrome
XM06p	Chronic fatigue syndrome Myalgic encephalomyelitis ME Myalgic encephalomyelitis Myalgic encephalomyelitis syndrome Postviral fatigue syndrome PVFS - Postviral fatigue syndrome CFS - Chronic fatigue syndrome
	mild/mod/sev
F2860	Mild chronic fatigue syndrome
F2861	Moderate chronic fatigue syndrome
F2862	Severe chronic fatigue syndrome
XaPom	Mild chronic fatigue syndrome
XaPon	Moderate chronic fatigue syndrome
XaPoo	Severe chronic fatigue syndrome
	Activity management
XaPeC	Activity management for chronic fatigue syndrome Activity management for myalgic encephalopathy Actvty managm for myalg enceph
.8Q1.	Activity management for chronic fatigue syndrome Activity management for myalgic encephalopathy Actvty managm for myalg enceph
8Q1..	Activity management for chronic fatigue syndrome Activity management for myalgic encephalopathy Actvty managm for myalg enceph
	Referrals
XaR7C	Referral to chronic fatigue syndrome specialist team Referral to myalgic encephalomyelitis specialist team

Code	Diagnostic
XaRAz	Referral for chronic fatigue syndrome activity management Referral for myalgic encephalopathy activity management
8HkW.	Referral to chronic fatigue syndrome specialist team Referral to myalgic encephalomyelitis specialist team

Table 2 ME/CFS comorbidities or inclusion/exclusion criteria

Addison's Disease – Adrenal insufficiency

Cushing's syndrome – Overactive adrenal gland

Hypothyroidism – Underactive thyroid

Hyperthyroidism (overactive thyroid)

Anaemia requiring treatment or blood transfusion

Haemochromatosis (iron overload)

Diabetes

Cancer (including lymphoma, leukemia, melanoma, carcinoma, neuroendocrine tumours)

Upper airway resistance syndrome

Sleep apnoea

Rheumatoid arthritis

Lupus

Polymyositis

Polymyalgia rheumatica

HIV/AIDs

Multiple sclerosis

Parkinson's disease

Myasthenia gravis

B12 deficiency

Tuberculosis

Hepatitis

Lyme disease

Clinical Depression

Bipolar Disorder

Schizophrenia

Substance abuse

cerebral cyst

glandular fever

orthostatic intolerance

Post-exertional malaise

Sleep disorder

Pain

cognitive impairment

Fatigue

Extreme pallor

Nausea and irritable bowel syndrom

Palpitations

Urinary frequency and bladder dysfunction

exertional dyspnoea

lightheadness

coeliac disease

fibromyalgia

Mast cell activation syndrome (MCAS)

Q fever

Narcolepsy

Sjogren's syndrome

Shingles