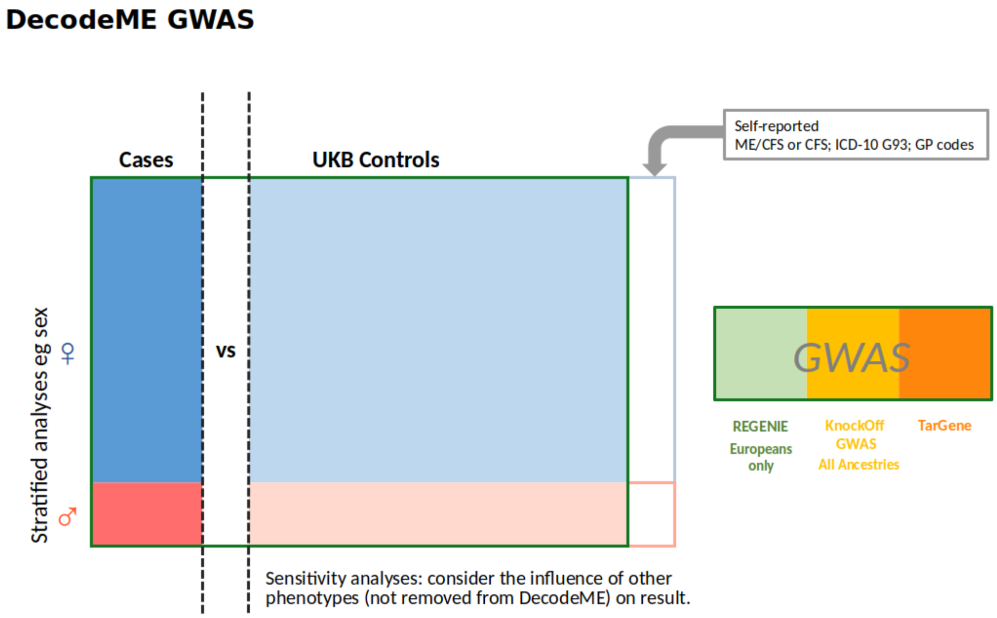
Data Analysis Plan

DecodeME Genetics Delivery Team

2024-03-17

Lay Summary: We will analyse the DNA of 18,069 people with ME/CFS and compare this with controls (people without ME/CFS 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 cases and, or, controls (other than the UK Biobank).

Technical Summary: Genome-wide analysis study (GWAS) analysis will be performed successively as data from ~4,800 sample batches is delivered. The main 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. The genome-wide significance threshold used will be $p \leq 5x10^{-8}$ per analysis using variants with minor allele frequency (MAF) ≥1%, e.g., for the first batch (~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. For replicating findings, data from various large biobanks will be used, where the electronic health record G93.3 will be used as proxy for an ME/CFS diagnosis.


DecodeME GWAS
Cases
UKB Controls
Self-reported
ME/CFS or CFS; ICD-10 G93; GP codes
vs
Stratified analyses eg sex
♀
♂
GWAS
REGENIE
Europeans only
KnockOff GWAS
All Ancestries
TarGene
Sensitivity analyses: consider the influence of other phenotypes (not removed from DecodeME) on result.

1 Case and control definition

1.1 Cases

1.1.1 DecodeME

Cases are (≥ 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 on here). The DecodeME algorithm selects participants based on three sets of answers on: 1) IOM/NAM criteria; 2) CCC criteria; and 3) 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 if possible. An aliquot of the extracted DNA is then sent by batch ($N \sim 5{,}000$) 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 was 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. Ultimately, DecodeME recruited 18,069 DNA participants. The main analysis will be initially conducted with these samples.

1.1.2 Limitations

Recruitment

The DecodeME project has completed recruitment of participants including those to be genotyped. Several rounds of GWAS will be performed as genotype batches (about 5,000 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

Questionnaire data collected from the first 17,074 DecodeME participants indicates a strongly biased sex-ratio towards women (5: 1)(1) as previously seen in some studies[1]. This might limit the discovery power for the male-only analysis (see Section 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 (less than 3%). With such low numbers, it will not be possible to use REGENIE (2) in a 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 (3) and TarGene (4) methods which allows us to analyse diverse and admixed individuals altogether.

1.2 Controls

1.2.1 UK Biobank

The controls are selected from the UK Biobank (UKB) (5,6) (UKB project 76173), as a general population cohort, but excluding UKB participants with ME/CFS as defined below. Furthermore, to limit the risk of spurious associations, UKB controls will be matched to DecodeME cases with regards to genetic sex and to genetically determined ancestry by excluding ancestry outliers in either cases or controls and by fitting ancestry informative covariates in the analysis model.

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

Potential UKB ME/CFS cases are defined as those who self-reported chronic fatigue syndrome ("CFS") in the baseline questionnaire (data field 20002), those who self-reported CFS in the verbal interview at any of the 4 visits to the UKB assessment centre" 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 at baseline but answered "No" to this question in the pain questionnaire (N = 88) will be excluded from being UKB controls, as will those who provided an ambiguous response ('Do not know' N=2,065 or 'Prefer not to answer'; N = 19). Potential ME/CFS cases who, in the baseline questionnaire, report both CFS diagnosis and good health will be used for neither cases nor controls.

Electronic health records (EHR) available in the UKB, such as hospital in-patient data based on ICD-10 codes and GP primary care record information, will be used to 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 as cases in UKB. However, among the UKB participants linked to the G93.3 code (N=1,278) there is a majority (65%) of individuals self-reporting having ME/CFS in the baseline or pain questionnaires. Similarly, these individuals are also more frequently (45%, N = 548) linked to a ME/CFS-relevant primary care code (Table 1 in Appendix) 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 in Appendix) 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, UK Biobank volunteers were aged between 40 and 69 years (y) old (median 56y) while the DecodeME DNA participants' ages span from 16 to 93y old (median 50y; mean 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. At minimum we will use a ratio of 1-to-3 cases-to-controls.

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 are returned by ThermoFisher in the format of CEL files (one per sample). The genotypes will be called in batch of up to 5,000individuals (53x 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 (Fig. 2.). Upon completion, this workflow returns a set of "best and recommended" variants prioritised by category (polyHighResolution, NoMinorMono, etc) as detailed in the AxAS manual. However, to ensure accuracy and reliability of the results analysis, 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 that are more likely to be less well genotyped or present problematic clustering. Therefore, extra QC steps need to be performed to maximize the number of samples and variants. These steps are presented below.

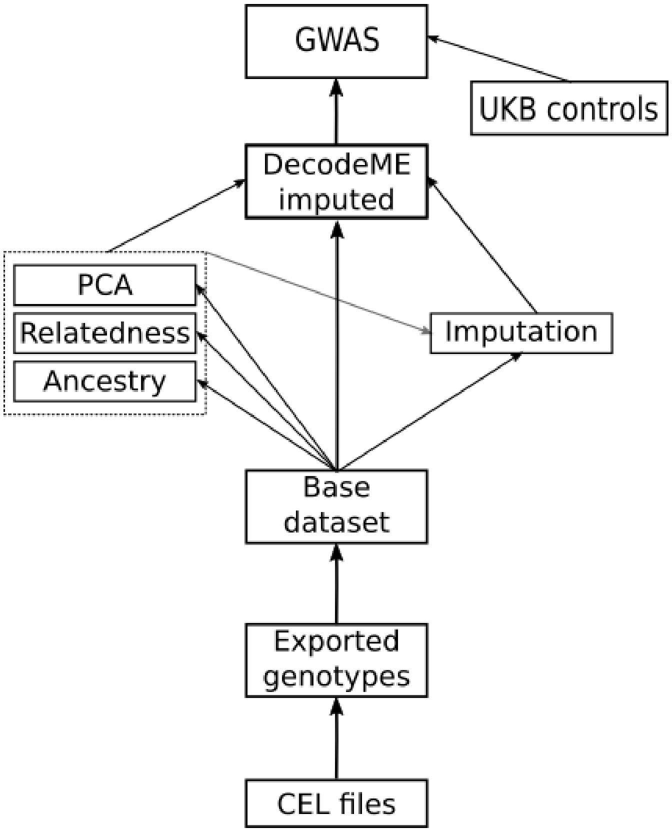
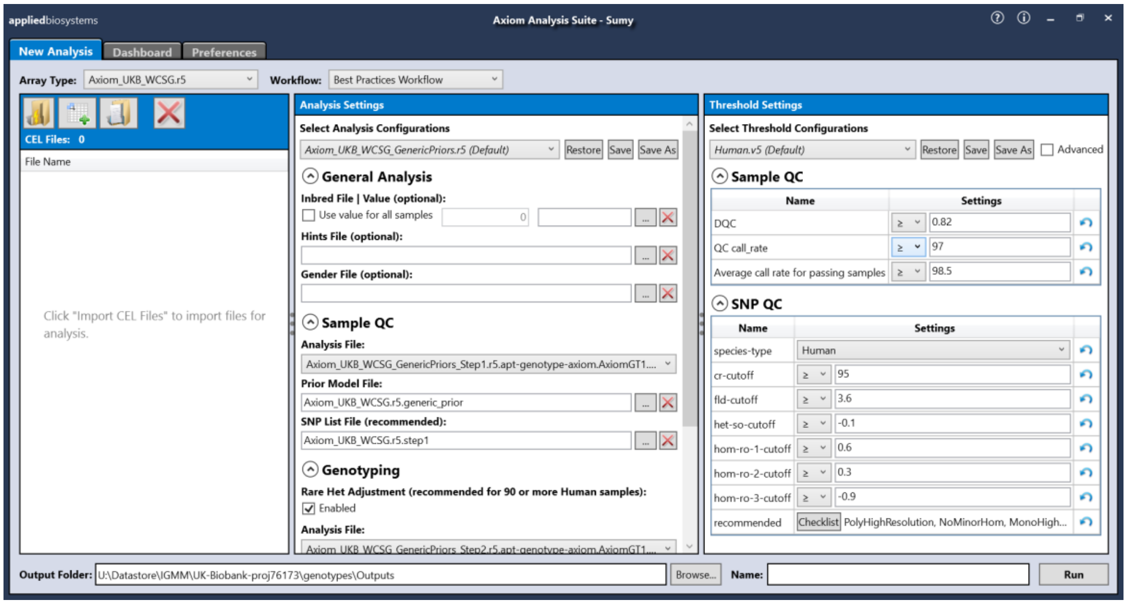


Figure 1: Quality control flowchart overview



Array Type: Axiom_UKB_WCSG.r5 **Workflow:** Best Practices Workflow

CEL Files: 0

File Name

Click "Import CEL Files" to import files for analysis.

Analysis Settings

Select Analysis Configurations

Axiom_UKB_WCSG_GenericPriors.r5 (Default)

General Analysis

Inbred File | Value (optional):

□ Use value for all samples 0

Hints File (optional):

Gender File (optional):

Sample QC

Analysis File:

Axiom_UKB_WCSG_GenericPriors_Step1.r5.apt-genotype-axiom.AxiomGT1....

Prior Model File:

Axiom_UKB_WCSG.r5.generic_prior

SNP List File (recommended):

Axiom_UKB_WCSG.r5.step1

Genotyping

Rare Het Adjustment (recommended for 90 or more Human samples):

☑ Enabled

Analysis File:

Axiom_UKB_WCSG_GenericPriors_Step2.r5.apt-genotype-axiom.AxiomGT1....

Threshold Settings

Select Threshold Configurations

Human.v5 (Default) □ Advanced

Sample QC

Name	Settings	
DQC	≥	0.82
QC call_rate	≥	97
Average call rate for passing samples	≥	98.5

SNP QC

Name	Settings	
species-type	Human	
cr-cutoff	≥	95
fld-cutoff	≥	3.6
het-so-cutoff	≥	-0.1
hom-ro-1-cutoff	≥	0.6
hom-ro-2-cutoff	≥	0.3
hom-ro-3-cutoff	≥	-0.9
recommended	Checklist	PolyHighResolution, NoMinorHom, MonoHigh...

Output Folder: U:\Datastore\IGMM\UK-Biobank-proj76173\genotypes\Outputs **Name:**

Figure 2: Screen capture of the Axiom Suite Analysis software

2.1 Extra sample-based quality control steps

Sex of participants are inferred during the automated genotype calling process in AxAS based on X and Y linked variants. Samples failing this inference are flagged as “unknown” sex. This can reflect underlying sex-chromosome aneuploidy and mosaicism. Such conditions can be identified after conducting a Copy Number Variant (CNV) analysis using the AxAS. The probeset intensities across whole sex chromosomes of “unknown” sex samples are visualised and compared to those of male or female references (7). Samples with an “unknown” sex that remain unresolved or presenting sex-chromosome aneuploidy will be flagged and removed. Additionally, samples showing a discrepancy between the self-reported sex in questionnaire at recruitment and the genetically inferred sex are also removed as indicative of potential sample mix-ups.

2.2 Extra marker-based quality control steps

2.2.1 Autosomal and female X chromosome variants

Variants with the lowest call rate or showing the strongest departure from HWE 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.

Visual inspection of genotyping clusters

The AxAS provides visualisation tools to inspect the genotype clusters for further QC (i.e., variant not flagged with default/recommended settings of metrics). A cluster plot “displays the probeset calls for the selected samples as a set of points in the clustering space used for making the calls” with the contrast on the x-axis and the size on the y-axis.

A non-problematic probeset should display a cluster plot with 3 or 2 (if no minor allele homozygotes are present among the samples) clusters that are well defined and separate, as shown below (Fig. 3) in an example of polymorphic biallelic variant from DecodeME batch 1.

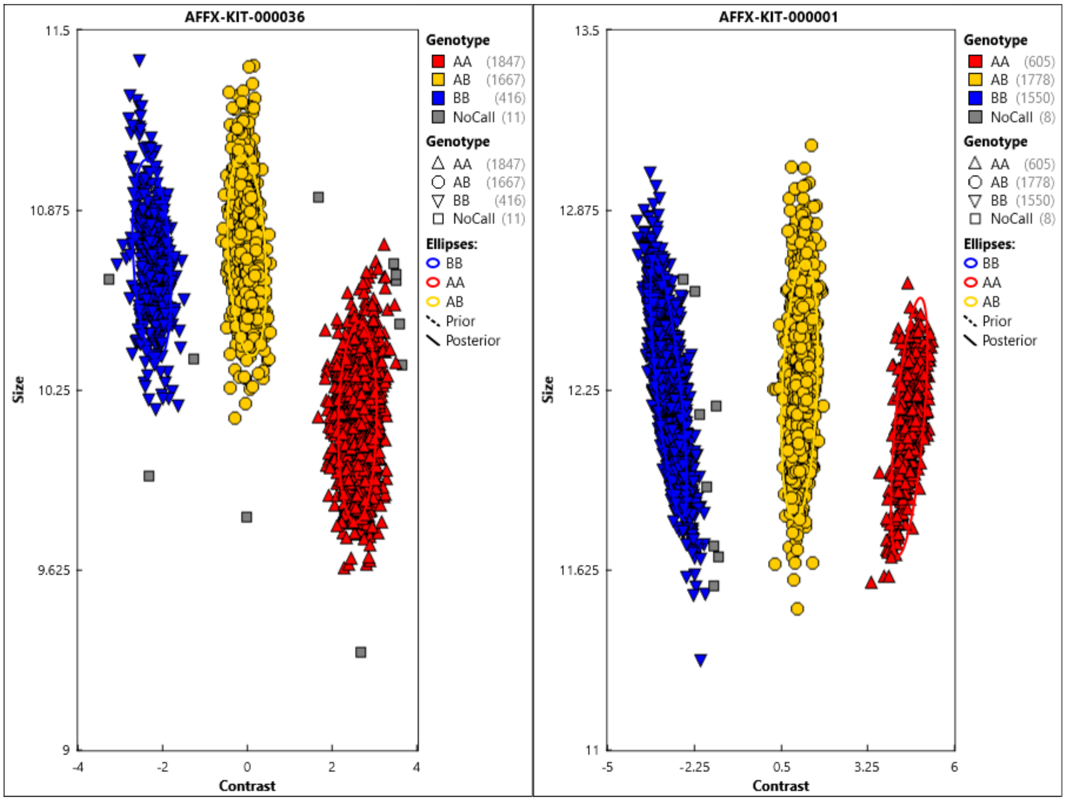
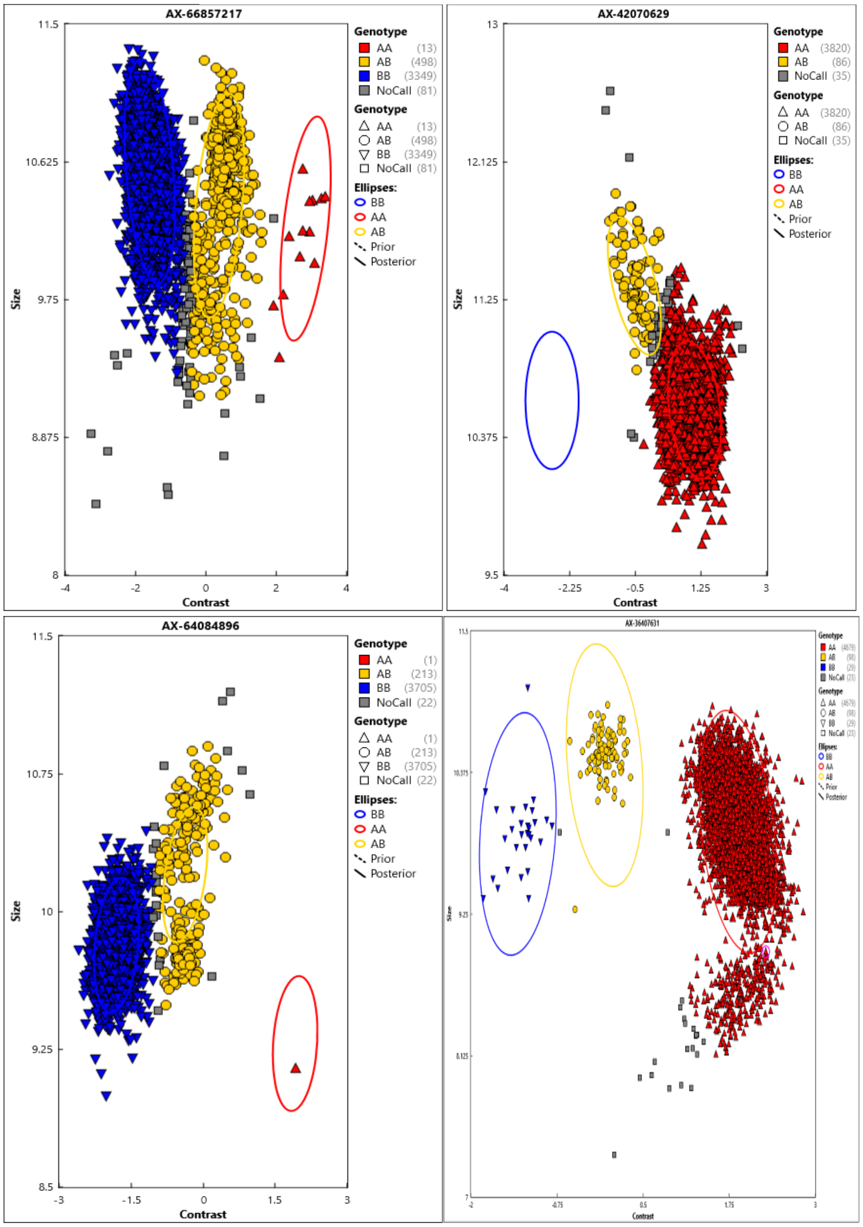


Figure 3: Example of expected clusters

Problematic probesets show cluster plots with equivocal patterns (overlapping clusters, or bimodal cluster) as in the examples below (Fig. 4):

Figure 4: Examples of problematic clusters



The problematic probesets detected are systematically removed from further analyses.

The following rules are applied:

Obvious cases of miscalled genotypes (e.g., bottom right panel, Figure 4), or unusually shaped clusters (top right and bottom left panels) render calls uncertain, possibly due to multiallelic rather than bi-allelic variants. These are flagged for removal.

For some probesets, clusters are touching or overlapping (e.g., Figure 4 far left panel). For these, probes were recalled manually or set to "Unknown" or "No call". Generating "No call" or "Unknown" increases genotype call missingness. If the fixed standard missingness threshold (2%) is reached or exceeded, then the variant is flagged for removal.

However, in some cases, adding missed calls ("Unknown" or "No call") does not change the overall genotype call, or the allele frequency. This would be the case, for example, for a homozygous cluster whose minor allele is rare. These cases are kept and are similar to the "NoMinorHom" category.

Hardy-Weinberg Equilibrium dependent visual inspection

Strong departure from HWE can highlight underlying genotyping issues. The Axiom guide indicates HWEp $<10^{-9}$ as indicative of poorly performing probes. For UKB batch reprocessing we applied a laxer threshold of HWEp $< 10^{-12}$ (cf UKB applied p $< 10^{-50}$). This accounts for a mixed ancestry of samples. A more stringent threshold will be applied in downstream analyses as required. For DecodeME cases, it is more challenging as these are selected samples rather than from a general population hence a departure from HWE can occur from the applied ascertainment. All probesets with HWEp $< 10^{-7}$ will be visually inspected (as above).

Plate and batch effect

It is not expected to see differences in genotype frequencies of a given variant between plates. Nevertheless, such a plate effect can occur when the intensities for a variant in one plate shift relative to the intensities of other plates within the same batch. Following the UKB methodology (4), we test if a given plate yields the same genotype frequencies as all other plates within a batch. This is done by using a Fisher's exact test on the 2×3 contingency table of genotype counts for each variant. The null hypothesis (no genotype frequencies difference) is rejected (i.e., there is an effect) when the smallest p-value is lower than the defined threshold as illustrated in Figure 5. This test is applied to all the best and recommended variants (~700,000) returned after genotype calling with AxAS. Any variant displaying a significant plate effect is flagged and removed from downstream analysis.

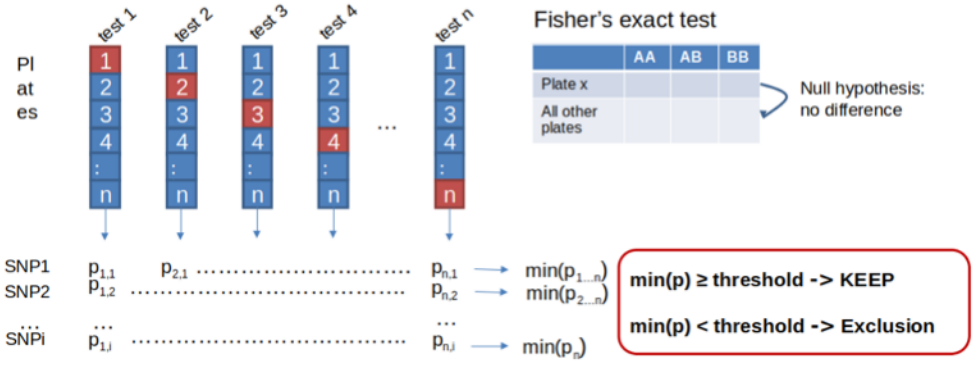


Figure 5: Schema of plate effect test

Similarly to the plate effect, it is not expected that we observe a significant difference of genotype frequencies for a given variant between different batches. The methods described above will be implemented at the batch level when multiple batches are available.

The significance threshold corresponds to a nominal p-value corrected by the number of tests carried out. The nominal p-value has been set at 0.005 and the number of tests is the product of the number of variants, number of plates, number of batch and the number of hypothesis (i.e., batch and plate effects, i.e., 2). So for 1 batch, 53 plates, 700,000 variant and 2 hypotheses the p-value significance threshold is $7x10^{-11}$.

Concordance analyses

Discordance between genotypes, for a given variant, between sets of genotype data from the same samples would suggest a genotyping issue which could lead to both wrong imputation and/or spurious associations.

Each plate processed by ThermoFisher has two control wells assigned to two individuals (NA19315, NA19318) from the AFR group of the 1000G. The first DecodeME batch of genotypes has 53 plates, and 51 and 50 genotyped replicates.

For each marker the discordance can be defined by (4) as:

$$d = 1 - \max(nAA,nAB,nBB)/(nAA+nAB+nBB)$$

where nAA, nAB, nBB is the count of each genotype. If d ≥ 0.05 for at least one of the controls then the marker is flagged and excluded.

Concordance analysis will also be done with known DecodeME duplicates present in different batches. Discordant variant between batches will be flagged and excluded.

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.
- Polymorphic Mitochondrial variants are also generally associated with two clusters in population samples.

Probesets for which call clusters aren't meeting these criteria or cannot be corrected manually to do so are 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 (8)

2.3.2 Relatedness

Genetic relatedness will be inferred using KING (9) 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 ancestry representation 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 principal component analysis (see below).

2.3.3 Ancestry

Ancestry will be inferred by principal component analysis (PCA) by projecting samples onto the principal components (PC) 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 (10) which identifies and removes long range linkage disequilibrium (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 6) to remove ancestry-informative variants (PC loading > 0.3) as was done for the UKB (8). The ancestry matching between the DecodeME cases and the UKB controls will be further refined using the PCA-based ancestry grouping method proposed in (11). 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.

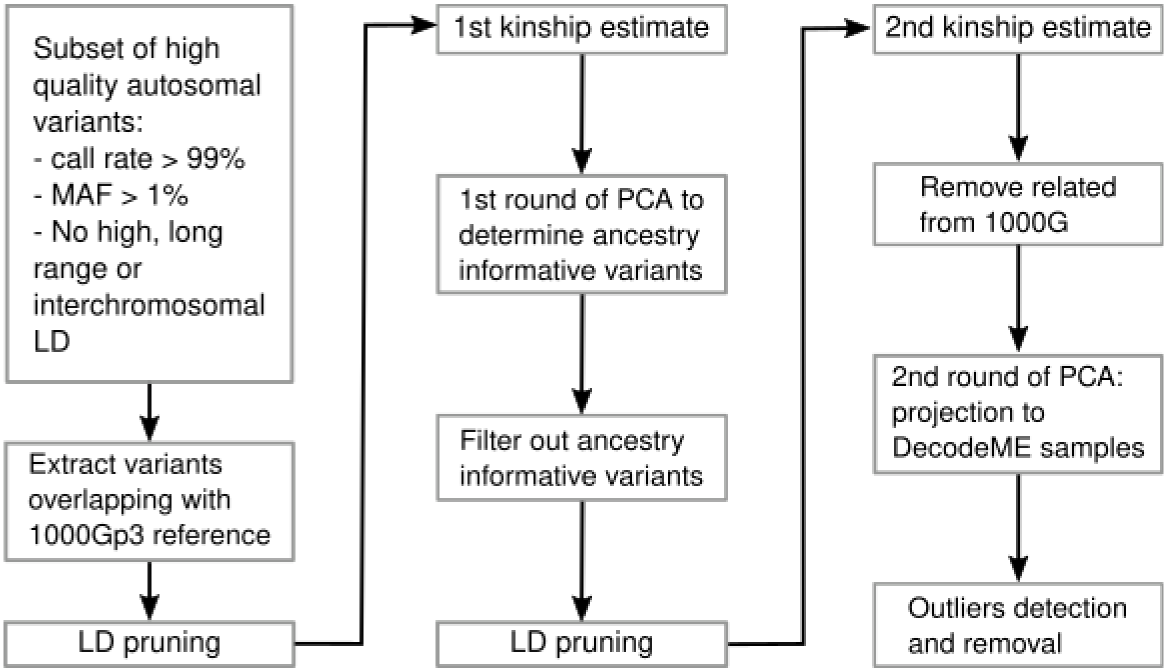

Subset of high quality autosomal variants:
- call rate > 99%
- MAF > 1%
- No high, long range or interchromosomal LD
Extract variants overlapping with 1000Gp3 reference
LD pruning
1st kinship estimate
1st round of PCA to determine ancestry informative variants
Filter out ancestry informative variants
LD pruning
2nd kinship estimate
Remove related from 1000G
2nd round of PCA: projection to DecodeME samples
Outliers detection and removal

Figure 6: Flowchart overview of the PCA-based ancestry grouping.

Once major ancestry groups are defined, we apply further quality checks within each group on all markers taken forward: individuals with outlying heterozygosity rate and variants departing significantly from HWE are identified.

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

Flag variants with $HWEp - value < 10^{-12}$ within each of the ancestry groups and remove them from downstream analyses.

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, albeit with the same array design and manufacturer. A batch of UKB controls' (N=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 UKB's variant calls; only markers passing QC in the independently performed processes and with concordant genotype calls will be kept.

Additionally, using the largest and homogeneous European ancestry UKB subgroup, allele frequencies (AF) will be checked for discrepancies from those of matched-ancestry references from GnomAD (v2.1.1), as shown in Figure 7. Under the assumption of a binomial distribution of allele count, we calculate the mean expected allele frequency using either the UKB or the GnomAD reference mean as true. Any variants with 6sd away from the expected mean is removed.

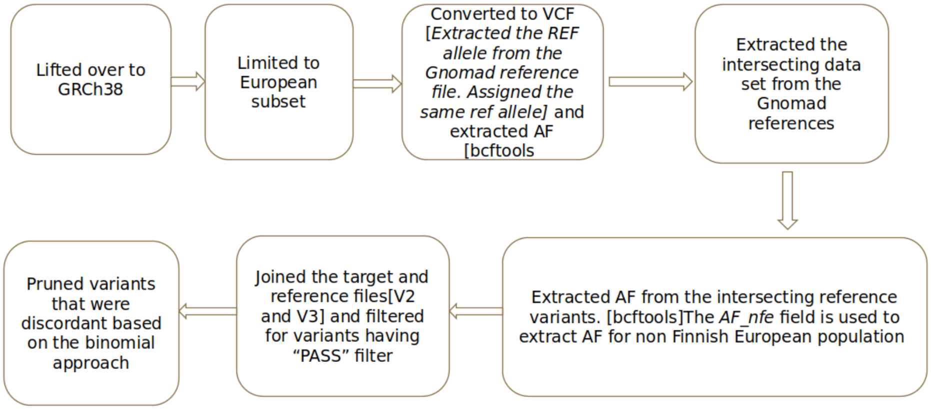

Lifted over to GRCh38
Limited to European subset
Converted to VCF [Extracted the REF allele from the Gnomad reference file. Assigned the same ref allele] and extracted AF [bcftools
Extracted the intersecting data set from the Gnomad references
Extracted AF from the intersecting reference variants. [bcftools]The AF_nfe field is used to extract AF for non Finnish European population
Joined the target and reference files[V2 and V3] and filtered for variants having "PASS" filter
Pruned variants that were discordant based on the binomial approach

Figure 7: Allele frequency concordance

In addition, the concordance between genotype calls from arrays and those newly generated from sequencing – as the whole genome sequences (WGS) available for entire UKB – will also be checked using BCFtools, variants with concordance < 95% will be removed.

3.2 Quality control on the merged set of case and controls

The DecodeME cases and the matched UKB controls (Section 1.2.1) are merged and evaluated together.

- Duplicate markers, inconsistencies of chromosome, position, strand or alleles have been identified. These issues were resolved whenever possible and if not the problematic variants were flagged and removed.
- The degree of relatedness between all the individuals was estimated using KING (Section 2.3.2). The relatedness estimation revealed duplicated samples between the DecodeME cases and the UKB cases. This allowed us to carry out a concordance analysis for merged genotyped variant between each pair of duplicates using the *sample-diff* function in Plink2. All the discordant variants were flagged and removed from analyses if discordant in more than one pair of duplicates.
- Following LD pruning and removal of regions of high LD ($r^2 > 0.8$), long-range LD (10) and inter-chromosomal LD (2), 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 PCs, variants driving discrimination will be flagged and removed, and PCA will be (re)performed until satisfactory outcomes.
- Following trial test of association using PLINK: variants with low p-value ($p<10^{-6}$) will be checked for departure from GnomAD population allele frequencies in controls, consistency of association p-values with LD structure, and genotype calls checked by visual inspection of probeset clusters (see 2.2.1) in cases, and controls if necessary, with the following rules:
    - If clearly dubious call, as defined in section 2.2.1, then blacklist and remove
    - If apparent good call (non-problematic cluster) for cases then check i) controls' genotyping clusters (only one batch); ii) for multiallelicity, which can induce discordant allele frequencies, in repositories (e.g., UKB WGS, GnomAD). In either scenario, identified dubious variants will be removed from the downstream analysis with UKB controls but nonetheless they will be kept in list of DecodeME genotypes passing QC
    - Associated variant is kept if no explanation is found for the genotype calls difference in cases and controls

4 Genotype imputation for cases

Genotype imputation is an important step prior to any GWAS. This process allows us to densify the genetic data for all individuals by inferring millions of unobserved genotypes from a 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.

4.1 Autosomes and X chromosome

Following UKB best practice (8), 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).

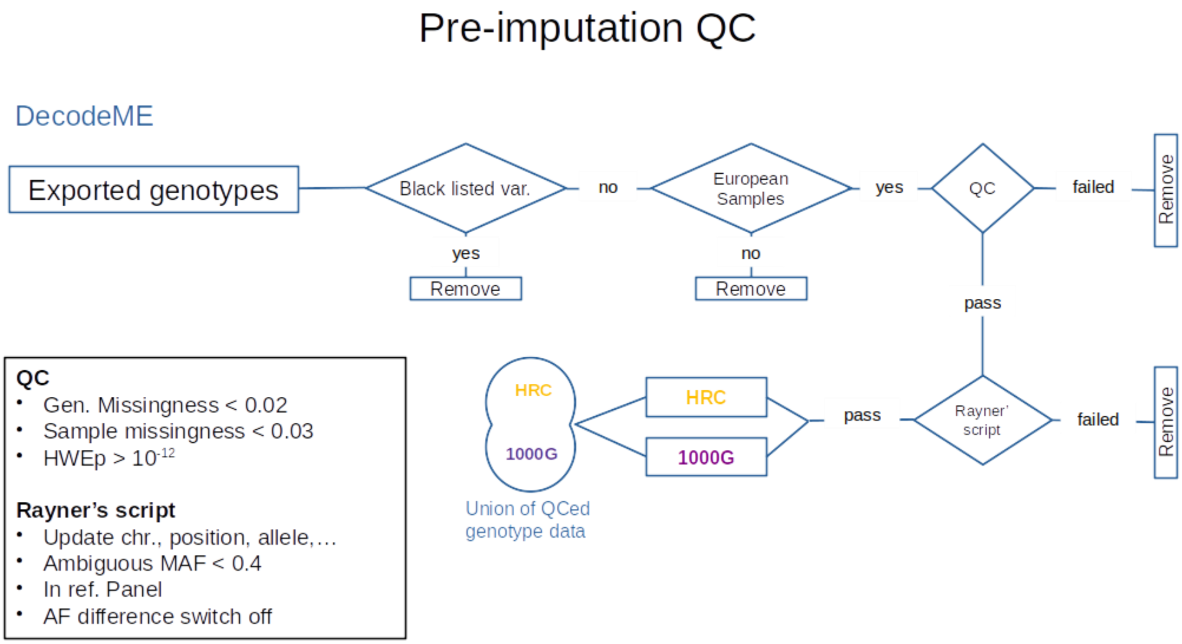
4.1.1 Data QC and preparation

Prior to imputation, it is essential to carry out specific quality checks pertaining to the imputation reference panel used. For this task, we will use the HRC/1000G 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 (muted for the cases and set to 10% for the controls). The strand, position, ref/alt assignment can be updated if discrepancy is flagged. The following criteria are used to filter out DNA variants:

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

For cases, an allele frequency difference test between genotypes and the reference is not applied.

We added to the pre-imputation QC by now imposing a reciprocal check with the UKB genotypes as described in Figures 8a, 8b and 9, below, to avoid artificial discrepancies between cases and controls genotypes which could lead to downstream spurious association. Only variants that passes QC in both cohorts will be imputed.



Pre-imputation QC

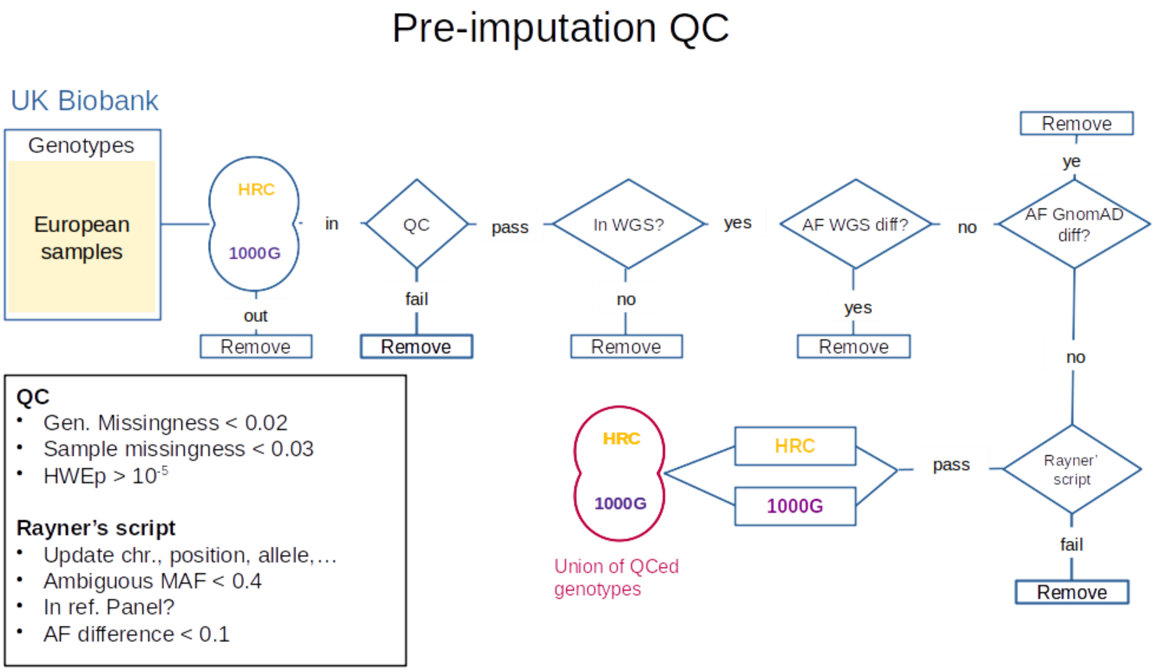
QC

- Gen. Missingness < 0.02
- Sample missingness < 0.03
- HWEp > 10^{-12}

Rayner's script

- Update chr., position, allele,…
- Ambiguous MAF < 0.4
- In ref. Panel
- AF difference switch off

Figure 8a: Pre-imputation QC for DecodeME genotypes



Pre-imputation QC

QC

- Gen. Missingness < 0.02
- Sample missingness < 0.03
- HWEp > 10^{-5}

Rayner's script

- Update chr., position, allele,…
- Ambiguous MAF < 0.4
- In ref. Panel?
- AF difference < 0.1

Figure 8b: Pre-imputation QC for the UKB genotypes

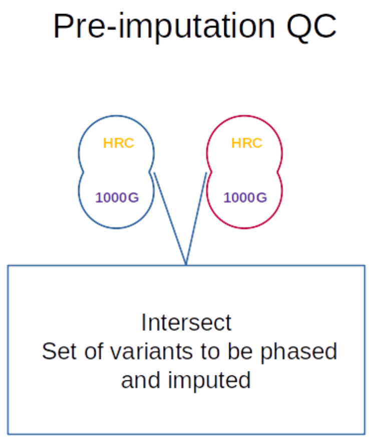


Figure 9: Pre-imputation QC

4.1.2 Phasing and Imputation

Phasing

This is a critical step before imputation which improves both efficiency and accuracy (12). 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 (13) 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. The phasing will be done on all the batches available altogether.

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 to 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, rare variants with a minor allele count (MAC) below 20, or 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 (8): the HRC variants will be combined with the UK10K+1000G variants absent from HRC. Each new batch of DecodeME genotypes will be pooled with the previous ones to allow them to be imputed together.

Thereafter, the DecodeME imputed genotypes are merged with the UK Biobank variants to create a single case-control dataset (see Section 5.1).

4.2 Human Leukocyte Antigen Complex

Classical human leukocyte antigen (HLA) alleles will be imputed using the HLA*IMP:02 algorithm as previously done for the UKB (8).

4.3 Mitochondrial DNA

Imputation of mitochondrial DNA (mtDNA) will follow the methodology proposed in (14).

4.4 Copy Number Variants

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 (15).

5 Association analysis

5.1 Merge the cases and control imputed data

The DecodeME cases and the UKB controls will be imputed separately (as the latter is already 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.

First, as shown in section 3.2, duplicate markers, inconsistencies of chromosome, position, strand, alleles or ambiguous variants with MAF > 0.4 will be flagged and removed if the issue cannot be resolved. Second, taking advantage of known duplicated samples between the DecodeME cases and the UKB controls (Section3.2), a concordance analysis will be carried out on the merged imputed data to identify discrepancies. Applying the aforementioned rule (Section 3.2), non-singleton discordant variants will be flagged and removed from downstream analyses. Third, we will stratify by caseness to extract the following metrics: genotype call missingness, HWE p-value, minor allele frequency and imputation quality. This will allow us to apply filters using these metrics before (e.g., only variants with genotype call missingness <2% after rounding of posterior genotypes probability using Plink in cases and controls will be put forward) or after GWAS. We will use the following lower thresholds: call rate > 98%, sample missingness > 97%, INFO > 0.4 and $MAF > 1\%$ for the first analysis with a batch of 4,800 cases and to 10^{-8} if $MAF > 0.5\%$ for larger sets of multiple batches. Fourth, we will check imputed variants' allele frequencies in controls against those in the UKB whole genome sequencing (WGS) and GnomAD. Imputed variants not present in the UKB WGS or failing the allele frequency (controls only) difference binomial test (see Section 3.1) with either UKB WGS or GnomAD will be flagged and removed.

The ancestry PCs will be calculated on the merged set using bigsnpr (10). We will carry out further quality control of the DecodeME genotyped and imputed data by doing genome-wide association studies on the blood type provided by the questionnaire (Q10).

5.2 Analysis plan

Different GWAS will be performed:

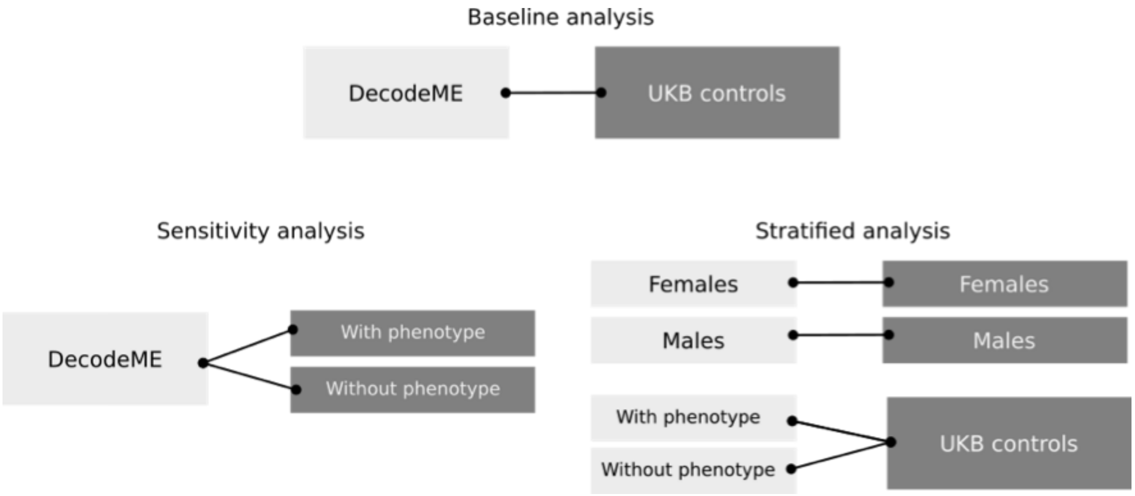


Figure 10: Overview of the different analyses

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 (d) have ME/CFS relevant primary care records. Furthermore, the controls will be sex-matched to the cases to account for the biased sex ratio (see Section 1.1.2) reducing therefore the total number of controls.

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. 10) 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 UKB (Table 2 in Appendix).

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. 10). First, both analyses (above) will be stratified by sex-at-birth (Part 2 Q8 in the Baseline 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 stratum.

5.2.4 Combined analysis

As previously mentioned, DecodeME participants will have their DNA genotyped and imputed following the UKB's standard procedure. This gives us the possibility to combine into a single set DecodeME cases with UKB participants who have evidence of a ME/CFS diagnosis (see 1.1.2). We will perform analysis with this combined set (against UKB controls; Fig. 11) which would boost the statistical power of discovery for variants enriched in both sets.

5.3 Association testing method

The following three association testing methods will be applied. All the GWAS will be performed with REGENIE which is our gold-standard for linear mixed model methods accounting for both relatedness and case-control imbalance. However, two other tools will be tested for some GWAS (see below).

5.3.1 REGENIE

REGENIE (2) 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 (16).

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.

5.3.3 TarGene

TarGene (4) is a statistical workflow that performs targeted estimation of effect sizes, as well as two-point (and higher) interactions. This algorithm has advantages of guaranteeing an optimal bias-variance trade-off, accounting for covariates, relatedness, population structure, ancestry and admixture, and detecting genetic non-linearities (i.e., the effect size of two alternate alleles is not twice that of one). It has been used primarily for testing at single loci, but for this project will be applied genome-wide.

5.4 Statistical significance threshold of association

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 Validation and replication

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) and matching UKB controls.
- Test for validation of variants that are significant (see thresholds in Section 5.4) in subsequently collected cases. A lower p-value of association with an increased number of cases would reinforce previous findings.
- 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 true associations. It also helps highlight potential technical issues such as genotyping or genotype-calling error.

A FinnGen team (https://www.finngen.fi/en) investigating ME/CFS, using only EHR available in the FinnGen, Estonian and Mass General Brigham biobanks, has agreed to collaborate with DecodeME to do mutual replication of our respective findings. This provides us with the best opportunity for an independent replication to date.

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 prioritise 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.

5.6.1 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) (17) then changes are considered meaningful.

5.6.2 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 UKB. Therefore, it will not be 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

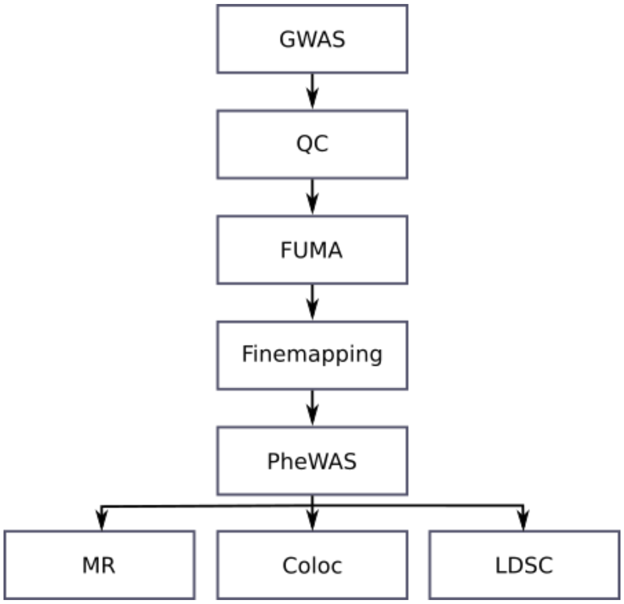


Figure 11: Flowchart overview of the post-GWAS analysis

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 strength of association (-log10 p-values) for each tested variant along the genome (chromosomal position). 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 significant association induced by causal effect(s). However, a strong deviation (inflation or deflation) might also indicate an enrichment in type I error (i.e., false positive association) caused by potential underlying issues such as unaccounted population stratification or covariates.

Due to its convenience, we will use the online plotting platform LocusZoom.org which generates both QQ and Manhattan plots that are contextual (gene annotations and local linkage disequilibrium patterns shown), interactive, zoomable and shareable.

6.2 Quality control

6.2.1 HWE departure

No HWE p-value filtering is applied to the DecodeME and UKB merged imputed data because a departure from HWE can reflect biological processes: increased number of homozygotes in cases or natural selection in a region (e.g., HLA region). It has also previously been noted that significant hits are often enriched with variants that are out of HWE (18). Nevertheless, variants showing a large departure from HWE, in controls, need further scrutiny. The HWE p-value of such variants should be checked in other control populations of the same ancestry. Keeping variants departing from HWE does not impact the GWAS procedure but only the outcome which can be filtered post-GWAS.

6.2.2 LD-based QC

A GWAS hit is defined by a lead SNP and other SNPs with which it is in LD. It is expected that the strength of association (p-value) of a variant is proportional to the genetic correlation (LD) between this variant and the SNP leading the association signal. The higher the LD correlation, the lower the p-value. Therefore, if there is an inconsistency of association p-values with LD then the association is likely spurious. LD-based QC, as implemented in DENTIST (19) is an efficient way to detect errors in GWAS summary statistics.

6.2.3 Visual inspection of genotyping cluster

Any genotyped variants that are significant SNPs or that are LD-clumped with a significant lead SNP will have its genotyping cluster inspected primarily in cases and, if needed, in controls (but limited to one UKB batch only) following the rules mentioned in section 2.2. A miscalled variant might affect imputation locally. Therefore, if such possibility is identified, a new local imputation (chromosome-wide) could be redone with the recalled or removed variant(s). Subsequently, the GWAS will be rerun.

Any variant failing these QC tests will be flagged and/or removed which could lead to an initial hit to disappear.

6.3 Functional annotation

Functional variant annotation is a crucial step for interpreting GWAS results and to 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 (20) 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 implements MAGMA (21) a gene-based test. Subsequent gene set enrichments could hint at potential biological functions, tissue expression, or overlaps with associations for other traits and conditions.

6.4 LD score regression

LD score regression (LDSC) (22) is a tool using GWAS summary statistics to estimate the tested trait heritability. It will 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.

Additionally, if heritability is non-zero, LDSC can also be used to partition heritability in different components such as that contributed by regulatory regions in specific tissues as a mean to determine the most relevant biological tissue where genetic variants exert their effects (23,24)

6.5 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 (25) or SuSie (26) that can handle multiple causal variants in proximity. Regulatory annotations obtained from FUMA for causal variants within the credible sets defined by fine-mapping can be complemented by using specialized platforms with most up to date tissue and cell-type expression evidences, such as FORGEdb (27).

6.6 Phenome-wide association studies

Phenome-wide association studies (PheWAS) test whether a variant has been previously associated with others traits or diseases. PheWAS can be carried out on curated genotype-phenotype databases like the NHGRI-EBI GWAS catalog (https://www.ebi.ac.uk/gwas/) with tools such as PhenoScanner (28) or LDtrait (29). Other more specialized databases can be used such as the GeneATLAS (30) for UKB based genotype-phenotype associations or drug target dedicated platform (OpenTarget, https://www.opentargets.org/). Level of significance for associations is set by Bonferroni correction for multiple (phenotypes) testing taking into account the non-independence of phenotypes.

6.7 Colocalisation and Mendelian randomization

Colocalisation (31) is a statistical method that tests if two traits share a genetic cause. Mendelian randomization (MR) tests whether an exposure might have a causal effect on an outcome using one or more genetic variants as instrumental variables. Colocalisation tests whether significant 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 (32).

To test if the GWAS signals are shared with expression quantitative trait locus (eQTL) data from blood (https://www.eqtlgen.org/) and multiple tissues (eQTL catalogue, https://www.ebi.ac.uk/eqtl/) we will use the summary data-based Mendelian randomization (SMR) (33), and the heterogeneity in dependent instruments will be tested with (HEIDI) (33). Other MR tools such as GSMR (34), will also be used. Colocalisation will be performed with the R package coloc (35) using the fine mapping results (see Section 6.5) which allow us to relax the single causal variant hypothesis. This approach is applicable to other available molecular quantitative trait locus (molQTL) data, such as protein (plasma) quantitative trait locus (pQTL) from the UK Biobank Pharma Proteomics Project (UKB-PPP) (36).

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/ UKB-PPP: https://metabolomips.org/ukbbpgwas/ Sanger imputation server: https://imputation.sanger.ac.uk/

UKB project 76173: https://www.ukbiobank.ac.uk/enable-your-research/approved-research/genome-wide-association-study-of-myalgic-encephalomyelitis-chronic-fatigue-syndrome-me-cfs

References

1. Bretherick AD, McGrath SJ, Devereux-Cooke A, Leary S, Northwood E, Redshaw A, et al. Typing myalgic encephalomyelitis by infection at onset: A DecodeME study. NIHR Open Res. 2023 Aug 21;3:20.

2. 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.

3. 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.

4. Pabet OL, Tetley-Campbell K, Laan MJ van der, Ponting CP, Beentjes SV, Khamseh A. Dispensing with unnecessary assumptions in population genetics analysis [Internet]. bioRxiv; 2023 [cited 2024 Mar 12]. p. 2022.09.12.507656. Available from: https://www.biorxiv.org/content/10.1101/2022.09.12.507656v2

5. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, 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 Mar 31;12(3):e1001779.

6. Allen NE, Lacey B, Lawlor DA, Pell JP, Gallacher J, Smeeth L, et al. Prospective study design and data analysis in UK Biobank. Sci Transl Med. 2024 Jan 10;16(729):eadf4428.

7. 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.

8. 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.

9. 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.

10. Privé F, Aschard H, Ziyatdinov A, Blum MGB. Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. Bioinformatics. 2018 Aug 15;34(16):2781–7.

11. 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.

12. 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.

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

14. 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.

15. 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.

16. Morris AP. Transethnic meta-analysis of genomewide association studies. Genet Epidemiol. 2011;35(8):809–22.

17. 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.

18. Truong VQ, Woerner JA, Cherlin TA, Bradford Y, Lucas AM, Okeh CC, et al. Quality Control Procedures for Genome-Wide Association Studies. Curr Protoc. 2022;2(11):e603.

19. Chen D, Tashman K, Palmer DS, Neale B, Roeder K, Bloemendal A, et al. A data harmonization pipeline to leverage external controls and boost power in GWAS. Hum Mol Genet. 2022 Feb 1;31(3):481–9.

20. 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.

21. Leeuw CA de, Mooij JM, Heskes T, Posthuma D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. PLOS Comput Biol. 2015 Apr 17;11(4):e1004219.

22. 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.

23. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015 Nov;47(11):1228–35.

24. Finucane HK, Reshef YA, Anttila V, Slowikowski K, Gusev A, Byrnes A, et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nat Genet. 2018 Apr;50(4):621–9.

25. 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.

26. 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.

27. Breeze CE, Haugen E, Gutierrez-Arcelus M, Yao X, Teschendorff A, Beck S, et al. FORGEdb: a tool for identifying candidate functional variants and uncovering target genes and mechanisms for complex diseases. Genome Biol. 2024 Jan 2;25(1):3.

28. Kamat MA, Blackshaw JA, Young R, Surendran P, Burgess S, Danesh J, et al. PhenoScanner V2: an expanded tool for searching human genotype–phenotype associations. Bioinformatics. 2019 Nov 15;35(22):4851–3.

29. Lin SH, Brown DW, Machiela MJ. LDtrait: An Online Tool for Identifying Published Phenotype Associations in Linkage Disequilibrium. Cancer Res. 2020 Aug 14;80(16):3443–6.

30. Canela-Xandri O, Rawlik K, Tenesa A. An atlas of genetic associations in UK Biobank. Nat Genet. 2018 Nov;50(11):1593–9.

31. 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.

32. 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.

33. 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.

34. Zhu Z, Zheng Z, Zhang F, Wu Y, Trzaskowski M, Maier R, et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. Nat Commun. 2018 Jan 15;9(1):224.

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

36. Sun KY, Bai X, Chen S, Bao S, Kapoor M, Backman J, et al. A deep catalog of protein-coding variation in 985,830 individuals [Internet]. bioRxiv; 2023 [cited 2023 May 23]. p. 2023.05.09.539329. Available from: https://www.biorxiv.org/content/10.1101/2023.05.09.539329v1

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 .F12Z	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 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

Code	Diagnostic
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
XaRAz 8HIL.	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

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 syndrome

Palpitations

Urinary frequency and bladder dysfunction

exertional dyspnoea

lightheadness

coeliac disease

fibromyalgia

Mast cell activation syndrome (MCAS)

Q fever

Narcolepsy

Sjogren’s syndrome

Shingles

Document change history:

Date	Version	Description
2023 03 06	1	New document created. Archived file can be accessed here - OSF | DecodeME
2024 03 17	2	Section 2.2 was expanded to include genotyping visual inspection of clusters rules, plate/batch effect and concordance analysis Section 4.1 was expanded to include improvement of pre-imputation QC Section 5.1 was expanded to include post-imputation QC Section 5.3.3 was added to include information about TarGene analysis. Section 5.5.3 was updated to reference the replication cohort used in the study. Section 5.3.3 was added to include information about TarGene analysis. Section 6 was expanded to include a new subsection about phenome-wide association studies and complement the other subsections (tools and databases).