UK Biobank Genetic Data: MRC-IEU Quality Control, version 2, 18/01/2019

Ruth E Mitchell, Gibran Hemani, Tom Dudding, Laura Corbin, Sean Harrison, Lavinia Paternoster

MRC Integrative Epidemiology Unit (IEU), University of Bristol, UK

1. Introduction

This document describes the quality control procedure undertaken and the derived files produced by the MRC-IEU for the full UK Biobank (N=~500,000) genetic data, version 3, imputed with the Haplotypes Reference Consortium (HRC) and the UK10K haplotype resource. These files are suitable for the majority of uses (i.e. to analyse common and well imputed low frequency variants for a subset of individuals with White British or European ancestry). If you have any questions about the suitability of using these files for your analysis, please speak to one of the authors.

All these QC files use 'IEU' ids. These have been generated to match the order of UK Biobank project 16009 fam files. Therefore all applications will need to link to these 'IEU' ids in order to use the filtered genetic dataset. **You will need to generate a linker file to match these ids that will link all your phenotype ids to the filtered 500,000 genetic data release and any QC files required. The code to generate your linker file is available in:** `./scripts/id_mapping/linker.sh`

The QC'd and derived files are located within the MRC-IEU research RDSF space and on Blue Crystal phase 4 for use on the compute nodes in the following directory:
`./data/ukbiobank/genetic/variants/arrays/imputed/released/2018-09-18`

Please note that if you are running long jobs using this data, you will need to use the compute nodes on BlueCrystal4 (BC4). All your analysis and scripts using UK Biobank data should be stored in an appropriate project directory that you create within the MRC-IEU projects directory on BC4 (`./projects/`).

Documentation about these files and scripts used in their generation can be found in the following directories:

```
./data/ukbiobank/genetic/variants/arrays/imputed/released/2018-09-18/docs
./data/ukbiobank/genetic/variants/arrays/imputed/released/2018-09-18/scripts
```

All QC lists of ids are given in the format of FID and IID required by PLINK, without headers or as a single list of identifiers with a two-line header as required for qctools. If using other software, then please adjust as required.

The majority of these fields have been derived from the sample_qc file provided as part of the full genetic release from UKBiobank, this file can be found here:

```
./data/ukbiobank/genetic/variants/arrays/imputed/released/2018-09-
18/data/raw_downloaded/qc_docs/ukb_sqc_v2.txt
```

The sample and fam files for **all individuals** within the genetic data are located here:

```
./data/ukbiobank/genetic/variants/arrays/imputed/released/2018-09-
18/data/id_mapping
```

NB: The .fam file contains all genotyped individuals, i.e. those with array data (n=488,377). The .sample files contain the subset of individuals for whom imputed data is available and separate sample files apply to autosomes (n=487,409), chrX (n=486,757) and chrXY (n=486,443).

Full documentation about the genetic data released by UK Biobank has been detailed in this publication: Bycroft et al. The UK Biobank resource with deep phenotyping and genomic data, *Nature* volume 562, pages 203–209 (2018)[1]. You should read this manuscript before using the data. Appropriate sections of the Supplementary Information are referenced here (prefixed S, e.g. S3.6). Some of the QC steps have data fields within the category 100313 'Genotyping process and sample QC' on the Biobank Showcase. The relevant Field IDs are referred to below, prefixed "f.".

The QC files containing list of individuals described in this document can also be used with the genotype array data. A .fam file with 'IEU' ids has been generated for the directly genotyped data (available as .bed files). No SNP filtering or sample QC has been applied to these files which are available here:

```
./data/ukbiobank/genetic/variants/arrays/directly_genotyped
```

2. Individual/sample level QC

Withdrawn consent - any individuals who have withdrawn consent will be regularly excluded from genetic data. However, it is the **responsibility** of the individual analysts to ensure that these individuals have been excluded for each analysis that they perform.

QC information (ie. inclusion/exclusion lists) is available for all those with genotype array data available (not just those with imputed data). Therefore, inclusion/exclusion lists will sum to 488,377 individuals.

2.1 Standard exclusions

```
./data/ukbiobank/genetic/variants/arrays/imputed/released/2018-09-
18/data/derived/standard_exclusions/
```

- data.sex_mismatch.[software].txt – **378 individuals to exclude** – this list of individuals has been derived by comparing genetic sex ('Inferred.gender' as determined by affymetrix) (f.22001) with reported sex ('reported gender') of the participant (f.31). Refer to S3.6.

- data. putative_sex_chromosome_aneuploidy.[software].txt - **652 individuals to exclude** - individuals with sex chromosome karyotypes putatively different from XX or XY (f.22019). 181 individuals overlap with the sex mismatch list. Refer to S3.6.1.

- data. het_missing_outliers.[software].txt – **968 individuals to exclude** - individuals that are outliers in heterozygosity and missing rates (f.22027). These individuals have been excluded from the imputed data. Refer to S3.5.

- data.combined_recommended.[software].txt - **1812 non-overlapping individuals to exclude** - individuals in the above three files with duplicates removed. Please note that 849 individuals will be excluded from the imputed data.

2.2 Ancestry restrictions

```
./data/ukbiobank/genetic/variants/arrays/imputed/released/2018-09-
18/data/derived/ancestry/
```

UK Biobank has defined a White British subset:

- data.white_british.[software].txt – **409,703 individuals to include** – these individuals have self-reported as 'White' and 'British' and have very similar genetic ancestry based on a principal components analysis of the genotypes. They are referred to as the "white British ancestry subset". Refer to section S3.4.

- data.non_white_british.[software].txt – **78,674 individuals to exclude** - exclusion list of individuals not in the "white British ancestry subset".

  NB: These lists sum to n=488,377

Our QC procedure (e.g. re-estimating minor allele frequencies) has filtered based on a less stringent ancestral subset (named 'Europeans'). This is a group of individuals who cluster with Europeans in a 4-cluster model, but includes a small proportion of individuals who self-identify with non-white ethnic groups. You may wish to use this less stringent set, but should carefully consider accounting for population structure if you do so (such as by using a mixed model method):

- data.europeans.[software].txt – **464,708 individuals to include** -an in-house standard k-means clustering analysis was performed on the first 4 principal components provided by UK Biobank using the parameters of 4 centers and 150 random sets in the statistical software environment R. This generated 4 clusters of which the largest forms of the individuals in this list. The individuals that form the European cluster is larger than the stricter 'white British' subset.

- data.non_europeans.[software].txt – **23,669 individuals to exclude** - exclusion list of 'non-europeans' derived from the in-house k-means cluster analysis.

  NB: These lists sum to n=488,377

NB: Please note that there is evidence that there is still a degree of population structure even within the white British subset[2] and so any analysis should carefully consider if such structure might affect the results.

2.3 Relatedness

```
./data/ukbiobank/genetic/variants/arrays/imputed/released/2018-09-
18/data/derived/relateds_exclusions/
```

- data.highly_relateds.[software].txt - **9 individuals to remove** - these individuals appear to be related (3rd degree) to a very large number (>200) of individuals. 7 of these individuals are not in the 'white British' subset. This list is derived using the list of individuals excluded from the kinship inference. Refer to S3.7.1.

- data.minimal_relateds.[software].txt - **79,491 individuals to exclude** - once removed, the remaining subset is a maximal set of unrelated individuals based on the kinship coefficients provided by UK Biobank (please be aware of the individuals excluded from this inference - individuals in the list of outliers and heterozygosity and missing rates and also the above 9 highly related individuals). This exclusion list was derived in house using an algorithm applied to the list of all the related pairs provided by UK Biobank (3rd degree or closer, derived using the kinship matrix). It preferentially removes the individuals related to the greatest number of other individuals until no related pairs remain.

NB. In total (removing the non white_British subset, the minimal_relateds, the highly_relateds and recommended_exclusions) there are 151,301 individuals to exclude from the genotype data and 150,333 individuals to exclude from the imputed data as individuals that are outliers in heterozygosity and missing rates have already been excluded (337,076 individuals to include). These files and number of individuals do not take into account availability of phenotype data, so the exact number for a particular analysis may vary.

3. Principal components

```
./data/ukbiobank/genetic/variants/arrays/imputed/released/2018-09-
18/data/derived/principal_components/
```

- data.pca1-10.[software].txt - contains the first 10 principal components.
- data.pca1-40.[software].txt - contains all 40 principal components.

These have been calculated by UK Biobank for all genotyped individuals (n=488,377) using a set of 406,247 unrelated, high quality samples and 147,551 high quality markers pruned to minimise linkage disequilibrium (f.22009). Please refer to section S3.3 for full details.

4. Standard covariates

```
./data/ukbiobank/genetic/variants/arrays/imputed/released/2018-09-
18/data/derived/standard_covariates/
```

- data.covariates.[software].txt - contains standard covariates of sex and genotyping array ('chip').

See section S2.3.3 for more details. There is evidence of differential array effect on markers scattered across the genome and so you may wish to adjust for genotyping array ('chip') in your analysis. However, if your outcome of interest is likely to affect lung function or smoking behaviour you should be aware that such an adjustment may introduce collider bias (due to UKBiLEVE participants being genotyped on a different array) and so we would recommend performing analyses with and without adjustment for genotyping array as sensitivity analyses.

5. Imputed SNP level QC

5.1 Filtered bgen files

Filtered bgen files that contain only the less stringent ‘european’ subset based on an in house kmeans clustering algorithm and after removing the standard exclusions (n=463,005, described in sections 2.1 and 2.2 of this document) are available. MAF and imputation info scores were recalculated on this subset of individuals and the following graded filtering of SNPs was performed (with varying imputation quality for different allele frequency ranges):

- Info>0.3 for MAF >3%
- Info>0.6 for MAF 1-3%
- Info>0.8 for MAF 0.5-1%
- Info>0.9 for MAF 0.1-0.5%

We recommend using these filtered files for your analysis, due to their reduction in size. They are located here:

```
./data/ukbiobank/genetic/variants/arrays/imputed/released/2018-09-
18/data/dosage_bgen
```

This directory also contains the associated sample (.sample) files for use with qctools and the index files (.bgi) for use the bgenix software.

The accompanying sample-stats files and snp-stats are located alongside this directory in:
`./data/ukbiobank/genetic/variants/arrays/imputed/released/2018-09-18/data/sample-stats`
and
`./data/ukbiobank/genetic/variants/arrays/imputed/released/2018-09-18/data/snp-stats`

NB. In total there are 12,370,749 SNPs and 463,005 individuals present in these files. Please be aware of duplicate SNPs due to triallelic alleles that may not fit the Info and MAF filtering mentioned above.

Please see associated documentation (`/docs`) for full information about all files in the `/data` folder.

6. Acknowledgement

We encourage the wide use of these files, for ease and consistency of analysis across the MRC-IEU and the authors can be contacted to discuss the appropriate use of these files for specific analyses.

If you use the derived files, as a minimum you should reference this documentation (stating the DOI) in the methods and include the following acknowledgement statement:

"Quality Control filtering of the UK Biobank data was conducted by R.Mitchell, G.Hemani, T.Dudding, L.Corbin, S.Harrison, L.Paternoster as described in the published protocol (doi:…………….)"

For full referencing details please see the catalogue record on data.bris (https://data.bris.ac.uk/data/).

You may wish to consider if it would be appropriate to include any of the authors of these derived files as co-authors on individual publications, especially where they have made specific contributions to your own analysis.

If you have any questions, please contact Lavinia Paternoster (l.paternoster@bristol.ac.uk).

7. Paragraph for publication

UK Biobank is a population-based health research resource consisting of approximately 500,000 people, aged between 38 years and 73 years, who were recruited between the years 2006 and 2010 from across the UK [3]. Particularly focused on identifying determinants of human diseases in middle-aged and older individuals, participants provided a range of information (such as demographics, health status, lifestyle measures, cognitive testing, personality self-report, and physical and mental health measures) via questionnaires and interviews; anthropometric measures, BP readings and samples of blood, urine and saliva were also taken (data available at www.ukbiobank.ac.uk). A full description of the study design, participants and quality control (QC) methods have been described in detail previously[4]. UK Biobank received ethical approval from the Research Ethics Committee (REC reference for UK Biobank is 11/NW/0382).

Genotyping and imputation
The full data release contains the cohort of successfully genotyped samples (n=488,377). 49,979 individuals were genotyped using the UK BiLEVE array and 438,398 using the UK Biobank axiom array. Pre-imputation QC, phasing and imputation are described elsewhere [1]. In brief, prior to phasing, multiallelic SNPs or those with MAF ≤1% were removed. Phasing of genotype data was performed using a modified version of the

SHAPEIT2 algorithm [5]. Genotype imputation to a reference set combining the UK10K haplotype and HRC reference panels [6] was performed using IMPUTE2 algorithms [7]. The analyses presented here were restricted to autosomal variants using a graded filtering with varying imputation quality for different allele frequency ranges. Therefore, rarer genetic variants are required to have a higher imputation INFO score (Info>0.3 for MAF >3%; Info>0.6 for MAF 1-3%; Info>0.8 for MAF 0.5-1%; Info>0.9 for MAF 0.1-0.5%) with MAF and Info scores having been recalculated on an in-house derived 'European' subset.

Data quality control

Individuals with sex-mismatch (derived by comparing genetic sex and reported sex) or individuals with sex-chromosome aneuploidy were excluded from the analysis (n=814).

Ancestry:

We restricted the sample to individuals of white British ancestry who self-report as "White British" and who have very similar ancestral backgrounds according to the PCA (n=409,703), as described by Bycroft [1].

OR

We restricted the sample to individuals of 'european' ancestry as defined by an in-house k-means cluster analysis performed using the first 4 principal components provided by UK Biobank in the statistical software environment R. The current analysis includes the largest cluster from this analysis (n=464,708).

Degree of relatedness:

Estimated kinship coefficients using the KING toolset[8] identified 107,162 pairs of related individuals [1]. An in-house algorithm was then applied to this list and preferentially removed the individuals related to the greatest number of other individuals until no related pairs remain. These individuals were excluded (n=79,448). Additionally 2 individuals were removed due to them relating to a very large number (>200) of individuals.

1. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* (2018). doi:10.1038/s41586-018-0579-z
2. Haworth, S. *et al.* Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nat. Commun.* **10,** 333 (2019).
3. Allen, N. E., Sudlow, C., Peakman, T., Collins, R. & UK Biobank. UK Biobank Data: Come and Get It. *Sci. Transl. Med.* **6,** 224ed4-224ed4 (2014).
4. Collins, R., Peakman, T., Alegre-Diaz, J., al., et & al., et. What makes UK Biobank special? *Lancet (London, England)* **379,** 1173–4 (2012).
5. O'Connell, J. *et al.* Haplotype estimation for biobank-scale data sets. *Nat. Genet.* **48,** 817–820 (2016).
6. Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* **6,** 8111 (2015).
7. Howie, B., Marchini, J. & Stephens, M. Genotype Imputation with Thousands of Genomes. *G3&amp;#58; Genes|Genomes|Genetics* **1,** 457–470 (2011).
8. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26,** 2867–2873 (2010).