MRC IEU UK Biobank GWAS pipeline, version 1, 14/12/2017

Ben Elsworth, Ruth Mitchell, Chris Raistrick, Lavinia Paternoster, Gibran Hemani, Tom Gaunt

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

1. Introduction

The MRC-IEU UK Biobank genome wide association study (GWAS) pipeline has been optimized to perform GWAS on the imputed genetic dataset of the full 500 000 from UK Biobank quickly, efficiently and in a standardized manner. The imputed data has been quality controlled[1] for the appropriate samples and SNPs to be included in the GWAS as detailed in this document. This pipeline offers the options of performing your GWAS of your trait of interest using either PLINK or BOLT-LMM software.

PLINK allows for an analysis to be performed in a homogeneous and unrelated population.

BOLT-LMM uses a linear mixed model (LMM) to account for both relatedness and population stratification, therefore allowing a wider range of individuals to be included in terms of relatedness and ancestry at a cost of slightly longer running time.

All that is required to use this pipeline is for the user to supply a phenotype file containing their phenotype of interest, a covariates file and the script used to generate these files (see below for detailed instructions).

2. Job submission

The job submission happens through a submission sheet which you will sign in to using your university email address. Each GWAS submission corresponds to a single row on the sheet with multiple columns to fill in. The procedure for the submission of a GWAS to the pipeline is detailed in the steps below (These details are also on the MRCIEU/Biobank Phenotypes Github wiki).

1) To avoid duplication, double check that your phenotype of interest has not already been analysed on the submission sheet.

2) Contact the IEU/ICEP Data Manager (Chris Raistrick, chris.raistrick@bristol.ac.uk) who will make a directory under your username in the GWAS project on the Research Data Storage Facility (RDSF) containing input and output directories.

3) Add your phenotype files, covariate file and scripts to input directory created above (see section 3 for file formats). Please do not have spaces in the file names.

A standard covariate file containing sex, chip and the first ten principal components have been provided within the MRC-IEU research RDSF space in the following directory:

```
./data/ukbiobank/software/gwas_pipeline/dev/release_candidate/data/covariates
```

These will need to be copied into your input directory adding any additional covariates necessary for your analysis. See the section on 'phenotype and covariate files' below for more details.

4) Check that the permissions on the input files are set to read for all (chmod 744 input/*)

5) Email Chris and ask for your input files to be copied to BlueCrystal4 (BC4) (this pipeline is implemented on BC4 and as RDSF is not mounted there, the files need to be manually copied over).

6) Once Chris has told you that he has moved the data **(and only then!)**, fill in the submission sheet with one row per phenotype, using the file names that are now in your input folder. **Do not edit the final four red fields.**

- **Name** - Your name
- **Location** - A unique name that matches that in the config file of the pipeline (don't worry too much about this).
- **Username** - Your University user name
- **Email** - The email address that alerts will be sent to
- **Method** – Choose either BOLT-LMM or PLINK
- **Model** - For PLINK only (linear or logistic)
- **Phenotype name** – An unique identifier for the phenotype which will be used to create a folder for the output
- **Phenotype description** - A description of your phenotype of interest
- **Biobank column ID(s)** - The Biobank column IDs used to create the phenotype
- **Phenotype file** - The name of the file containing your phenotype of interest (see below for details)
- **Phenotype file generation script** - The script used to generate the phenotype and covariate files, no spaces please.
- **Phenotype column name** - The column containing the phenotype data to be tested in the association analysis
- **Covariates file** - The file to be used for the covariates data (can be contained in the same file as the phenotype data so long as the covariates are specified)
- **Categorical covariates** - The column(s) to be used for the categorical covariates
- **Quantitative covariates** - The column(s) to be used for the quantitative covariates. Default covariates, if none are specified, for BOLT-LMM are sex and genotyping array; for PLINK, sex, genotyping array and the first 10 principal components.
- **Job status** - Set to 'Hold' to hold the job back from submission, or 'Run' to submit the job to the queue.

7) Check the status of the job on the sheet. You will be notified by email when the job finishes or of any errors to the address that you have specified. An email will also be sent to Chris Raistrick and he will copy your results back to your directory on RDSF.

PLEASE NOTE - as soon as you click run and the jobs are submitted to the queue, pressing hold or deleting the fields in red will not cancel the GWAS on BlueCrystal. Each chromosome runs as a separate job so the 22 jobs will still be running. In the event of needing to cancel the GWAS please contact Ben Elsworth or Tom Gaunt.

3. Phenotype and covariate files

3.1. IDs

The pipeline uses a filtered version of the UK Biobank genetic data and therefore both the phenotype and covariate files need to have the genetic ids of UK Biobank project approval 8786 (PI Neil Davies). It is recommended that when you apply to use the genetic dataset, you request to link to this application. **The linker file provided will link all your phenotype ids to the 500,000 genetic data release.**

3.2. File format

- These files should be space delimited text files.
- The first two columns must be FID and IID (the PLINK identifiers of an individual); any number of columns may follow.
- Values in the column should be numeric.
- Case/control phenotypes should be encoded as 1=unaffected (control), 2=affected (case).

An example of a combine phenotype and covariate file:

```
FID IID phenotype1 covariate1 covariate2
123456 123456 100 1 1
234567 234567 20 2 1
345678 345678 50 3 NA
```

Default covariates, if none are specified, for BOLT-LMM are sex and genotyping array; for PLINK, sex, genotyping array and the first 10 principal components.

There is evidence of differential array effect on markers scattered across the genome and therefore we recommend adjusting for genotyping array ('chip'). 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 recommended performing analyses with and without adjustment for genotyping array as sensitivity analyses.

3.3. BOLT-LMM binary trait analysis

BOLT-LMM performs a linear regression and therefore the output betas will represent an absolute ‘risk difference’ scale. To transform to obtain log odds ratios you can use the following formula:

$$\log OR = \frac{\beta_{bolt}}{\mu(1-\mu)}$$

with $\mu = n_{case}/(n_{case} + n_{control})$ being the case prevalence, which can be obtained from the BOLT output files. The standard errors are adjusted using the same transformation

$$se(\log OR) = \frac{se_{bolt}}{\mu(1-\mu)}$$

4. Quality control of UK Biobank Genetic Data for GWAS pipeline

***** We are currently running the pipeline with only the imputed SNPs within the HRC – this will be updated once UK Biobank re-release the UK10UK imputation *****

The quality control (QC) of the input genetic data used in the GWAS has been described elsewhere[1]. The genetic data provided by UK Biobank has been filtered to include 11,511,739 SNPs in the analysis.

4.1 Mixed model to account for population stratification and relatedness: BOLT-LMM

Please see the BOLT-LMM website and accompanying paper for full details of the BOLT-LMM algorithm for mixed model association testing. This pipeline performs that default BOLT-LMM analysis which consists of calculating the BOLT-LMM-inf statistic (Standard (infinitesimal) mixed model association) and only computing the BOLT-LMM statistic (association test on residuals from Bayesian modelling using a mixture-of-normals prior on SNP effect sizes) if an increase in power is expected based on cross-validation.

4.1.1. Sample QC

The BOLT-LMM analysis includes 463,013 individuals after standard exclusions and the population has been restricted to a European subset.

4.1.2. Kinship modelling

The random effects component of the BOLT-LMM analysis is computed using only a subset of the SNPs that were directly genotyped. Sample QC for these SNPs was performed as described above. The number of SNPs used in the model has been optimised for timely running. Following testing and simulations, 143,006 SNPs are included in the model after selection using the following criteria:

- MAF > 0.01
- genotyping rate > 0.015
- Hardy-Weinberg equilibrium p-value< 0.0001
- r^2 threshold of 0.1

4.2. Restriction to homogeneous and unrelated population: PLINK

4.2.1. Sample QC

The PLINK analysis includes 334,977 individuals after standard exclusions, the population has been restricted to the 'White British' subset and related individuals have been excluded.

5. Acknowledgement

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

If you use this pipeline, you should reference both this documentation (stating the DOI) as well as the ‘UK Biobank Genetic Data: MRC-IEU Quality Control’ documentation[1] 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.Paternoster as described in the published protocol (doi:10.5523/bris.3074krb6t2frj29yh2b03x3wxj). The MRC IEU UK Biobank GWAS pipeline was developed by B.Elsworth, R.Mitchell, C.Raistrick, L.Paternoster, G.Hemani, T.Gaunt (doi:…..) ”

For full referencing details please see the catalogue record on data.bris.

You may wish to consider if it would be appropriate to include any of the authors of this pipeline 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) or Tom Gaunt (tom.gaunt@bristol.ac.uk).

6. 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[2]. 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[3]. 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[4]. 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 within the HRC site list 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[1].

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).

PLINK:

In the current study, 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[4]. Estimated kinship coefficients using the KING toolset[8] identified 107,162 pairs of related individuals[4]. 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[1]. These individuals were excluded (n=79,448). Additionally, 2 individuals were removed due to them relating to a very large number (>200) of individuals.

BOLT-LMM:

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) [1].

Association analysis: statistical methods

PLINK:

Genome-wide association analysis (GWAS) was conducted using linear/logistic regression, implemented using the software PLINKv2.00. Genotype array, sex and the first 10 PCs (out of 40) supplied by UKBiobank were fitted as covariates in the model.

BOLT-LMM:

Genome-wide association analysis (GWAS) was conducted using linear mixed model (LMM) association method as implemented in BOLT-LMM (v2.3)[9]. To model population structure in the sample we used 143,006 directly genotyped SNPs, obtained after filtering on MAF > 0.01; genotyping rate > 0.015; Hardy-Weinberg equilibrium p-value < 0.0001 and LD pruning to an r^2 threshold of 0.1 using PLINKv2.00. Genotype array and sex were adjusted for in the model. BOLT-LMM association statistics are on the linear scale. As such, test statistics (betas and their corresponding standard errors) were transformed to log odds ratios and their corresponding 95% confidence intervals on the liability scale using a Taylor transformation expansion series[10].

1. Mitchell, R., Hemani, G., Dudding, T. & Paternoster, L. UK Biobank Genetic Data: MRC-IEU Quality Control, Version 1. *doi.org* doi:10.5523/bris.3074krb6t2frj29yh2b03x3wxj
2. 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).
3. Collins, R., Peakman, T., Alegre-Diaz, J., al., et & al., et. What makes UK Biobank special? *Lancet (London, England)* **379,** 1173–4 (2012).
4. Bycroft, C. *et al.* Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv* (2017).
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).

9. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47,** 284–290 (2015).

10. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed model association for biobank-scale data sets. *doi.org* 194944 (2017). doi:10.1101/194944