# The mrbase\_grs function

The mrbase\_grs function creates genetic risk scores (GRS) using UK Biobank data given a list of studies or traits in MR Base. It creates 4 files that need to be copied to the Blue Crystal RDSF, and additional files are required if using proxies or clumping is required. The SNPs in the GRS are weighted by the beta coefficient reported by the GWAS found in MR Base.

|  |  |
| --- | --- |
| Argument | Description |
| **output=NULL,** | The output of the function, only mandatory argument. One of:   * subcategories * traits * studies * snps * code |
| category=NULL  subcategory=NULL  trait=NULL  population=NULL  sex=NULL  mr=NULL  samplesize=0  notstudies=NULL  studies=NULL  p=5e-08 | These options limit the retrieved studies. Lists as vectors can be specified.  *Category*, *subcategory* and *trait* limit study focus, and *population*, *sex* (male, female, male and female), *mr* (MR Base – does the study have data for all SNPs that allow MR?) and *sample size* (>= specified number) limit the study population.  *Notstudies* and *studies* limit by individual studies.  *P* limits SNPs to those with a P value less than specified (when downloading from MR Base, not in specified data frames) |
| exposure\_dat=NULL  exposure\_file = NULL | *exposure\_dat* is a data frame containing all SNP data required to create a GRS, which needs to be in the correct format for MR Base. The *exposure\_dat* file is created using MR Base if not specified, so this option is only relevant if you have a list of SNPs and effects etc.  *exposure\_file* is a csv containing all the *exposure\_dat* data, meaning you don’t need to read the data into R for the function to create the relevant code. Note the headings need to be in the correct format for MR Base, the same as *exposure\_dat* |
| proxies=TRUE  r2=0.8 | If *proxies* are specified, the program will search for proxies for SNPs not in UK Biobank. The *r2* gives a lower limit for the *r2*value for proxies to be allowed (SNPs not in UK Biobank without a proxy above this threshold will not be used in the GRS). The lower *r2* limit is 0.6 |
| clump=TRUE  clumped=TRUE | *clump* alters behaviour of the *extract­\_instruments* function in MR Base. If *TRUE*, MR Base will download pre-clumped SNPs.  *clumped* tells the program that data specified with *exposure\_dat* or *exposure\_file* has been clumped. Set to false if the data needs clumping.  The code will automatically set *clumped* to false if it is forced to download unclumped data from MR Base, which happens when too many SNPs are below the P value threshold – you don’t need to do anything if this happens, but it will slow down the code |
| gwas="biggest", | An option for determining which studies to use when gathering SNPs. Options are “**biggest**” or “**all**”. *Biggest* extracts SNPs from the study with the most participants. *All* uses all studies for a particular trait, removing duplicate SNPs by selecting the SNP from the largest study. Note: when more than one study is used, the data will be re-clumped |
| suffix=”” | Adds a suffix to any files created, allowing for a) more informative file names, e.g. suffix = “\_05” and “\_08” to distinguish between P value thresholds. Allows running of several scripts simultaneously in Blue Crystal |
| snpstats\_file="SNPstats.txt"  ld\_file="1000\_genomes\_ld.csv.gz"  eur\_file = “eur” | File names of required files. The SNP stats file will be created if not present, but the LD and EUR files must be copied from the project folder, or otherwise created/downloaded. The LD file is used for finding proxies, and the EUR files are used in clumping. Both come from 1000 genomes (EUR is Europeans only). |
| keep\_files=FALSE | If *TRUE*, will not delete selected intermediate files. If you want ALL intermediate files, you can always edit the code generated by the program. If *FALSE*, only necessary files will be kept |
| ipd=FALSE | If *TRUE*, generates a csv containing dosages of all SNPs requested for all participants in UK Biobank. Allows for MR analyses in addition to GRS analyses. IPD = individual participant data |
| maf = 0.42 | Used when harmonising MR Base with 1000 genomes and UK Biobank. This is the lower limit for minor allele frequency can be to determine intermediate frequency palindromic SNPs. Intermediate frequency palindromic SNPs cannot be reliably harmonised, EXCEPT when the strand can be determined for all other SNPs from the same study, in which case strand can be inferred and the MAF becomes irrelevant |
| plink\_grs=FALSE | If *TRUE*, the program will use plink to generate the GRS, not R. This gets around memory issues if using a large (>10,000) number of SNPs. This option ONLY WORKS for single traits – do not use this option when specifying more than one trait, you’ll still only get one (incomprehensible) GRS out. If you have specified many traits/studies and run into memory issues, split the traits/studies between more than one mrbase\_grs command and use prefixes to distinguish between them.  The absolute values generated by plink are not equal to those generated using R, but the standard deviations are the same, and therefore regression results will be the same (as will standardised versions of the GRS). |
| bgen\_folder=”… " | The folder to take the UK Biobank data from. This code will likely work fine if you specify a different folder containing the SAME type of data - .bgen and .bgen.bgi for 22 chromosomes. Note – you need access to this folder to use this program. |

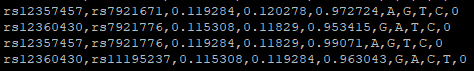
## Additional Files

To run the code fully, additional files are required in your home directory. The first two files can be copied from my project folder (email [sean.harrison@bristol.ac.uk](mailto:sean.harrison@bristol.ac.uk) for access), and can possibly be downloaded, but I’m not sure where from. The final file can be created by the code, or copied with the other two files.

### 1. 1000 genomes LD file – searching for proxies

This file shows the SNPs in linkage disequilibrium (R2 above 0.6) for all SNPs in 1000 genomes. It’s used to find proxies for SNPs that are in the original GWAS, but aren’t in the target dataset. It needs to show the original SNP, proxy SNP, EAF of both SNPs, R2 between the SNPs, and the alleles of both SNPs.

Called **1000\_genomes\_ld.csv.gz** by default.



### 2. European 1000 genomes data (bed, bim, fam) – clumping SNPs

These 3 files (bed, bim and fam) are raw genetic data for the European subsample of 1000 genomes. They are used to clump SNPs, removing any SNPs with R2 values of more than a threshold (0.001).

Called **eur.bed, eur.bim** and **eur.fam** by default.

### 3. SNPstats file

This file is a record of all SNPs in UK Biobank, and includes rsID, major/minor alleles, major allele frequency and info score, among other things. This file is used to see whether a SNP is in UK Biobank, and if so, what the major/minor alleles are for harmonisation.

Called **SNPstats.txt** by default. If not present, it will be created from the individual chromosome SNPstats files.

## Code procedures

### Step 0. Creating a list of relevant studies in MR Base given options specified

Creates a list of studies, part of most outputs.

1. Grab the available outcomes from MR BASE:
   1. **mrbase\_grs\_ao = available\_outcomes()**
2. Replace some missing values in sample size and consortium
3. Restrict the available outcomes by specified options (e.g. sex = “male”)

### Output 1. subcategories

Returns a list of subcategories, and the number of traits and studies in each subcategory.

1. Step 0
2. Create a list of subcategories from the outcomes
3. Return a dataframe showing the number of studies and traits for each subcategory

### Output 2. traits

Returns a list of traits, the number of studies in each trait, and the subcategory to which the trait belongs.

1. Step 0
2. Create a list of traits from the outcomes
3. Return a dataframe showing the number of studies for each trait, and which subcategory the trait is within

### Output 3. studies

Returns a list of studies. The information is a slimmer version of the output of *available\_outcomes* in MR BASE.

1. Step 0.
2. Return the outcomes dataframe, removing selected columns

### Output 4. snps

Returns a list of SNPs from relevant studies. This includes a merger of the info in *available\_outcomes* and *extract\_instruments* in MR Base.

1. If studies are NOT specified:
   1. Step 0
   2. Create a list of studies to be included
2. Else:
   1. Check if the studies listed are numeric, if so that is the list of studies
3. If the list of studies is not empty, extract the instruments using the P and clumping options specified:
   1. **mrbase\_grs\_instruments = extract\_instruments(outcomes=mrbase\_grs\_studies, p1 = p)**
4. Return these instruments, merged with study level information

### Output 5. code

This generates and saves four files necessary to create a GRS for each unique trait

1. If *exposure\_dat* or *exposure\_file* are specified, nothing is downloaded from MR Base
2. Else if *traits* are specified:
   1. Step 0
   2. Create a list of studies to be included
   3. If GWAS option “biggest” is specified, use the biggest study only, otherwise include all studies for each trait
   4. Create **exposure\_dat.csv** using the resulting SNPs
3. Else if *studies* are specified:
   1. Check the list of studies is numeric
   2. Step 0
   3. Limit the available outcomes to specified studies
   4. Create **exposure\_dat.csv** using the resulting SNPs
4. The **script.R** script is created, which is the main part of this process and will be described below
5. The **run.sh** script is created, which adds relevant modules to Linux to allow script.R to run, and gives instructions to Blue Crystal if not running on the login node
6. The **harmonise.R** script is created, which is used to help the harmonisation step(s)
7. Note: all files will have suffixes if specified, except **harmonise.R** which can be used for all programs, unless you change the value of maf

## Running the program

To use the mrbase\_grs function, you first need to copy it from my project folder (M:\projects\ieu2\\_working\IEU2\_P6\_006\data\R) or Github (<https://github.com/sean-harrison-bristol/UK_Biobank_PRS>) to a folder of your choice.

Once you have the code, import it into R using, for example:

**source(“M:\\projects\\ieu2\\\_working\\IEU2\_P6\_006\\data\\Code\\mrbase\_grs\_v2.02.r”)**

The syntax for the command is at the start of this document, but some example code is given at the bottom of this document. The first time you use the code, you’ll need to login to google – this is a check the MR Base uses, not mrbase\_grs, and will create an authorisation file in the folder you called the program from. I think your university email works fine for this. Any additional packages will be installed and added if not already in your version of R.

Once you have the four files from running mrsbase\_grs in R, follow these steps to complete the process. You must have logins for WinSCP and Putty, and access to the UK Biobank genetic data.

1. Copy the files in “Additional files” in my project data folder to your home directory on Blue Crystal, as well as the four files created using the mrsbase\_grs function. You can use WinSCP for this.
2. Type these commands into Putty (when in your home directory, type **pwd** to find out where you are, and **cd** to change directory):
   1. **chmod +x ru\* sc\***
   2. **dos2unix ru\* sc\***
3. This will give the **run.sh** and **script.R** files executable permission, and convert them from DOS (Windows) to Unix (Linux). These steps are mandatory if working on a Windows computer, since carriage returns are differently specified between the systems.
4. Type **qsub run.sh** (or whatever the run script is called if you used a suffix). This will send the run script to the Blue Crystal queuing system. I have entered reasonable values for the runtime, number of nodes and number of processors for the job, but if you are finding the code needs more time, increase the value of “walltime”.
5. You can run the script from the login node, but it isn’t recommended by HPC (if it’s a small job, I would say it’ll probably be fine to just run the script on the login node, type: **./run.sh**).
6. The scripts will produce a file called **grs.csv**, which contains a list of IDs and a list of GRS for all traits you specified. A file called **exposure\_dat\_harmonised.csv** is also created, and this contains details for all SNPs used in the GRS, including whether or not a proxy was used, the effect allele, betas etc. If specified, an individual participant data file will also be produced containing dosage data for all SNPs specified, **snp\_ipd.csv**.
7. All participants in UK Biobank with genetic data are included – remember to exclude individuals if you only want, for example, only White British individuals, or the set of minimally related individuals. See here for more information:
   1. <https://data.bris.ac.uk/datasets/3074krb6t2frj29yh2b03x3wxj/UK%20Biobank%20Genetic%20Data_MRC%20IEU%20Quality%20Control%20version%201.pdf>

## Additional Notes

### Proxies

The program checks which SNPs in *exposure\_dat* are in UK Biobank using the SNP stats file. For any SNPs not in UK Biobank, proxies are searched for using a file from 1000 genomes. The R2 value for these proxies are checked against the minimum (0.8 by default), and then these are searched for in UK Biobank. Proxy SNPs remaining that are in UK Biobank are ranked by their R2, and the top hit is selected as the proxy.

Proxies are marked in the **exposure\_dat\_harmonised.csv** file, with the proxy SNP rs ID in the “SNP” column, and original SNP rs ID in the “Original SNP” column. The r2 value is also included. To not search for proxies, use the proxies=FALSE option.

### Harmonisation

MR Base SNPs must be harmonised with SNPs in 1000 genomes and UK Biobank. There are three main issues in harmonisation, all of which are dealt with in the program:

1. The effect allele could be the other allele in different studies, e.g. in one study, the SNP A/C could have A as the effect allele, whereas in another the C could have been the effect allele
2. The strand could be different between studies, i.e. the forward strand could be used in one study, and the reverse in another. For example, on the forward strand the SNP is A/C, but on the reverse strand, the SNP is T/G
3. Palindromic SNPs make it impossible to tell whether SNPs are on the forward or negative strand, and thus which is the effect allele. Palindromic SNPs are either A/T or C/G, and thus appear the same whether on the forward or reverse strands. However, when the minor allele frequency is low (e.g. <0.3) you can think of the SNPs as [A/C/T/G]minor and [A/C/T/G]Major, so the forward and reverse strands are now identifiable. So long as the frequencies are relatively similar between studies, you can work out the effect allele.

The program gets around the final problem by seeing whether all the non-palindromic SNPs in one study are on the same strand as the other study (or whether they are all on the other strand). If so, then the minor allele frequency is not required to determine whether a palindromic SNP is on a different strand. If not, then any palindromic SNPs with an intermediate minor allele frequency (>0.3 by default) will be discarded.

### Clumping

Clumping is done using plink 1.90b4.1 and the European participants from 1000 genomes:

**plink -bfile eur --clump unclumped\_snps.txt --clump-p1 1 --clump-p2 1 --clump-r2 0.001 --out clumped\_snps**

**tr -s [:blank:] < clumped\_snps.clumped | cut -f 4 -d " " > clumped\_snps.txt**

### Bgenix and Plink

Bgenix is used to select SNPs from the UK Biobank .bgen files, before passing them to Plink 2.00 to create dosages:

**bgenix -g $inbgen -i $inbgenidx -incl-rsids $snp\_list > $temp\_geno\_prefix.$chrom\_padd.bgen**

**plink2 –bgen instruments.bgen --hard-call-threshold 0.4999 --export A --out instruments**

### Triallelic SNPs

About 50,000 SNPs are triallelic, so have two different other alleles or effect alleles. These mess with the code too much to be included, so are excluded immediately. Also note that Plink doesn’t work with triallelic SNPs.

### Field names

If you are specifying an *exposure\_dat* or *exposure\_file*, then ensure the headings are identical to those required by MR Base. Specifically, you require:

* SNP (string)
* effect\_allele.exposure (string)
* other\_allele.exposure (string)
* eaf.exposure (numeric)
* beta.exposure (numeric)
* pval.exposure (numeric, optional unless clumping is necessary)
* trait (string, optional but desirable)

### Errors

Memory errors can occur if the number of SNPs is too great. I think this is an issue with R on Blue Crystal, but can’t confirm. In any case, if using a single trait, use *plink\_grs=TRUE*, otherwise split the traits into manageable chunks.

Other errors may result from using this program in a way I haven’t anticipated. I’ve put in some controls to stop obvious errors in R, but can’t account for everything. You can always email me if you are having trouble, but if you’re doing something advanced, feel free to edit any code to suit your purposes, just let me know whether it’s something I should implement.

## Examples

Suppose that initially, I didn’t know what traits were available in MR Base, but I knew I only wanted to look at Europeans, so I ran this code:

**traits = mrbase\_grs(output="traits", population = "European")**

**write.csv(traits,file="traits.csv")**

This gave a CSV file containing all the traits in MR Base, and I selected traits from there to include. For this example, let’s say I chose “Body mass index” and “Height”. I now want to have a look at which studies are in MR Base that look at these traits, and I want them to not be in UK Biobank:

**traits = c("Body mass index", "Height")**

**studies = mrbase\_grs(output="studies", trait = traits, population = "European")**

**studies = studies[studies$consortium != "UK Biobank",]**

**write.csv(studies,file="Studies.csv")**

I can now look and see there are 13 studies looking at height and BMI. However, most of them are from the same consortia (GIANT and ECC). I should check those consortia online to see whether UK Biobank was included, but in these studies, it wasn’t. Now, I can either choose which study to include (the *id* field in the *studies* data frame), use the biggest studies by default, or use the option *gwas=”all”* to use all the SNPs from both the height and BMI studies (preferentially taking the data from the biggest studies if SNPs are in more than one study). In this case, I’ll go with the default options, and let the program choose the biggest studies in each of height and BMI.

**dat8 = mrbase\_grs(output="code",traits = traits, population = “European”, ipd=TRUE)**

I’ve specified that ipd=TRUE because I want the dosage data for all SNPs, so I can do some MR sensitivity analyses along with the GRS analysis.

This code generates the four standard files, which I copy across to Blue Crystal using Win SCP. I add executable permission to the script and run files, use dos2unix on them, then submit the run script to the qsub. This produces the grs.csv and snp\_ipd.csv files I can then use in my analysis. I remember to get rid of any participants I don’t want because I wrote the program, and hopefully everyone else will too.

If I want all SNPs with a P value below 5e-05, I can specify this instead:

**dat5 = mrbase\_grs(output="code",traits = traits, population = “European”, ipd=TRUE, p=5e-05,suffix = “\_05”)**

If I then change my mind and don’t want to create the IPD file, I can change the code without redownloading everything:

**dat5x = mrbase\_grs(output="code",exposure\_dat = dat5,suffix = “\_05x”)**

If I then find that I want to do a P=1 GRS, which uses ALL the SNPs measured in a GWAS, I can download the relevant genetic data from the GIANT website, rename any fields that don’t match, and specify:

**mrbase\_grs(output="code",exposure\_file = “exposure\_dat”,plink\_grs=TRUE)**

## Statement for papers

Information for SNPs associated with [trait[s]] in previous GWAS were downloaded from MR-Base [ref] using the R statistical package, including rs IDs, effect estimates, effect alleles and relevant study information. [If **proxies** used] The 1000 genomes project was used to find proxy SNPs in LD with SNPs not found in UK Biobank. The SNPs from MR-Base were harmonised with the SNPs from UK Biobank, aligning the effect estimates and alleles. The genetic risk score[s] were created by multiplying the number of effect alleles for each participant in UK Biobank by the effect estimate of the SNP from MR-Base, then summing across all SNPs associated with each trait.

**--Feel free to edit the above, or ask me any question,** [**sean.harrison@bristol.ac.uk**](mailto:sean.harrison@bristol.ac.uk)