Imputation Protocol  
Michigan Imputation Server

# Step 1: Preparing data for the imputation server

Genotype QC

Typical pre-imputation QC criteria:

These are some steps that we recommend for sample QC. Additional QC steps may be needed and should be determined by the local analysts for each study. Studies should provide a brief description of QC criteria in the Excel sheet.

• Sample call rate (cut off >95% threshold recommended)

• Exclude samples with heterozygosity > median + 3\*IQR

• Remove gender mismatches

• Remove duplicates

• Remove PCA outliers using a PCA projection of the study samples onto 1KG reference samples.

• Hardy-Weinberg p>10-6, SNP call rate ≥98%

• Remove monomorphic markers

Additional QC will be conducted centrally at the meta-analysis stage.

Our starting point is a QC’ed plink dataset in binary format (bim, bed, fam-build 37). In this first step you use the perl script *HRC-check-bim.pl* (version 4.2.5, found also in this folder) to check the format of your QC’ed plink files. You need to provide to this command the \*.bim and \*.frq file from your cohort. The tool also requires site lists from the reference panels for comparison, which is available also in this folder (HRC.r1-1.GRCh37.wgs.mac5.sites.tab.gz (unzipped with gunzip))

The frq file can be generated as:

*plink --bfile <binary plink prefix> --freq --out input\_file\_prefix*

Ths perl script creates a set of plink commands to fix any errors in your files.

*perl HRC-1000G-check-bim.pl -b <bim file> -f <frequency file> -r HRC.r1-1.GRCh37.wgs.mac5.sites.tab –h*

A shell script *Run-plink.sh* wil be created containing the plink commands needed to correct your plink files to be suitable for the Michigan imputation pipeline. (> chmod a+x *Run-plink.sh* )

Then you need to transforms the created files to vcf.gz format as required for the Michigan Imputation Server. Do this in 2 steps, for autosomal chromosomes please run the following shell script

*#!/bin/bash*

*for i in `seq 1 22`;*

*do*

*plink --bfile myfile-updated-chr${i} --recode vcf-iid --out myfile-updated-chr${i}*

*bgzip myfile-updated-chr${i}.vcf*

*tabix –p vcf myfile-updated-chr${i}.vcf.gz*

*done*

While for chromosome X please run:

*plink --bfile myfile-updated-chr23 --recode vcf-iid --out myfile-updated-chr23*

*sed 's/^23/X/g' myfile-updated-chr23.vcf >  myfile-updated-chrX.vcf*

*bggzip myfile-chrX.vcf*

# Step 2: Running the imputations

Now it is time to run the imputations. First you must make an user account for Michigan Imputation Server (<https://imputationserver.sph.umich.edu/index.html#!pages/home>). After this you can log in. When you log in you will automatically get to the upload page: We recommend you submit all autosomal chromosomes in one job. And the X chromosome in a different one.

For autosomal chromosomes we recommend using phasing algorithm Eagle which is implemented by the imputation server). If you would like to provide phased haplotypes please convert them to vcf format.

Thus, select the following options:

* Reference Panel: HRC r1.1 2016
* Phasing: Eagle v2.3
* Population: EUR (this parameter is for quality control purposes)
* Mode: Quality Control & Imputation

As for X chromosome currently only SHAPEIT is able to do the phasing

Thus, select the following options:

* Reference Panel: HRC r1.1 2016
* Phasing: SHAPEIT
* Population: EUR (this parameter is for quality control purposes)
* Mode: Quality Control & Imputation

# Step 3: Downloading the imputation results

After imputations have finished you will receive an email with the instructions to download the imputations. Please note that the imputation results will be deleted after 7 days, so you must download them before the deadline. The email you receive will also contain a decryption code, DO NOT DELETE IT!. Download the imputations and save them somewhere safe. The best way is to download straight to your server, by using the download links located under the curved arrows at the right of the imputation results.

NOTE on chromosome X: You will receive two output files for chromosome X, one for males and one for females. Please merge these files into a single vcf for chromosome X using vcf-merge from vcftools.

*vcf-merge males.chrX.imputed.vcf.gz females.chrX.imputed.vcf.gz | bgzip –c > combined.chrX.imputed.vcf.gz*

# Step 4: Prune to remove monomorphic variants

Each of these panels will impute a great many variants into your samples, the vast majority of which are rare, and many will not be polymorphic in your sample. To minimize the size of output files in the association analyses below, we ask that you remove all monomorphic sites from input files prior to analysis. In conjunction, we ask that you upload a list of the variants that you drop in this process. The following UNIX code for plink2, bgzip and tabix, will both generate the list of monomorphic SNPs and the pruned vcf files.

*for i in `seq 1 22` X;*

*do*

*plink2 --vcf chr${i}.imputed.vcf.gz --freq --recode vcf-iid --out chr${i}*

*cat chr${i}.frq | awk -v FS=” “ ‘{if($5 == 0) print $2}’ > chr${i}.exclude*

*plink2 --vcf chr${i}.imputed.vcf.gz --exclude chr${i}.exclude --recode vcf-iid --out chr${i}.imputed.poly*

*bgzip chr${i}.imputed.poly.vcf*

*tabix -p vcf chr${i}.imputed.poly.vcf.gz*

*cat chr${i}.exclude >> PANEL\_STUDY\_ANALYST\_imputed.monomorphic.txt*

*done*