ENIGMA 1000 Genomes phase 3 version 5 cookbook

(Adapted from the Michigan Imputation Server and MiniMaCH3 instructions)

Please note this protocol was developed for the ENIGMA consortium. If you are not part of the ENIGMA consortium and wish to use this protocol please register on the ENIGMA mailing list so that we can contact you regarding any updates or issues that arise relating to this protocol. You can register at http://enigma.ini.usc.edu/members/join/

If you use data generated by this protocol for non-ENIGMA projects please include the following citation for the protocol in your work:

ENIGMA Genetics Support Team. ENIGMA 1KGP_p3v5 Cookbook [Online]. The Enhancing Neuroimaging Genetics through Meta-Analysis (ENIGMA) Consortium. http://enigma.ini.usc.edu/genetics-protocols/ [13 July 2017]

Please let us know if you run into problems along the way -The ENIGMA Genetics Support Team (enigma.genetics@gmail.com).

What reference set are we using?
We will be using the all-ethnicities 1KGP reference set (phase 3 release v5). This reference set has been show to impute well in European ancestry samples and provides the good imputation in non-European ancestry samples.

If you have already imputed your data to the 1KGP phase 3 release v5 references using the all ethnicities references you should not have to run this protocol. Please contact the helpdesk to let us know how your data were imputed to confirm this.

Programs required for this protocol

It is assumed that you have access to a server or computer running linux/unix. If this is not the case please contact the enigma helpdesk. Before we start, you need to download and install some required programs (which you may already have). The required programs are: Plink1.9, R, bgzip, tabix and VCFtools. Links to the download sites are available below. Please address any questions to: enigma.genetics@gmail.com.

Plink1.9 can be downloaded here: https://www.cog-genomics.org/plink2

R can be downloaded here: http://cran.stat.ucla.edu/

Bgzip and tabix can be downloaded from here: http://www.htslib.org/doc/tabix.html

VCFtools can be downloaded from here: http://vcftools.sourceforge.net/

Paste the blue lines below into a terminal window or shell script, substitute the correct values where text is highlighted in yellow.

Multi-dimensional Scaling (MDS) Protocol

Please note if you ran the MDS Protocol for ENIGMA1, 2 or 3 and you have not added any new individuals or new genotyping you do not need to re-run the MDS analysis!

The MDS protocol assumes you are using a bash shell. To check which shell you are using on your server type:

```
echo $SHELL
```

If you are not using bash please change to the bash shell to run the protocol by typing:

```
bash
```

when you are finished to go back to your usual shell (if you weren't already in bash) type:

```
exit
```

##

Download the customized reference data from the following webpages to your working directory which we will call /enigma/genetics/:

##

```
cd /enigma/genetics/
```

Download the following 3 files from https://genepi.qimr.edu.au/staff/sarahMe/enigma/MDS/ and put them in the directory created above.

```
wget "http://genepi.qimr.edu.au/staff/sarahMe/enigma/MDS/HM3_b37.bed.gz"

wget "http://genepi.qimr.edu.au/staff/sarahMe/enigma/MDS/HM3_b37.bim.gz"

wget "http://genepi.qimr.edu.au/staff/sarahMe/enigma/MDS/HM3_b37.fam.gz"
```

If the wget command errors out, you can also download directly from the website.

##

Filter SNPs out from your dataset which do not meet Quality Control criteria (Minor Allele Frequency < 0.01; Genotype Call Rate < 95%; Hardy-Weinberg Equilibrium < 1x10-6). Directions assume your data are in binary plink format (bed/bim/fam), if this is not the case try to convert to plink format and contact enigma.genetics@gmail.com with questions.

##

```
export datafileraw=yourrawdata # replace yourrawdata with the name of the local plink file name

plink --bfile $datafileraw --hwe 1e-6 --geno 0.05 --maf 0.01 --noweb --make-bed --out
${datafileraw}_filtered
```

##

Unzip the HM3_b37 genotypes. Prepare the HM3_b37 and the raw genotype data by extracting only snps that are in common between the two genotype data sets - this avoids exhausting the system memory. We are also removing the strand ambiguous snps from the genotyped data set to avoid strand mismatch among these snps. Your genotype files should be filtered to remove markers which do not satisfy the quality control criteria above.

##

```
cd /enigma/genetics #change directory to a folder with you plink dataset and downloaded HM3_b37 files
```

```
gunzip HM3_b37*.gz
```

```
export datafile=${datafileraw}_filtered # ${datafileraw}_filtered should give you the name of the local plink file name that has been filtered of SNPs not meeting QC criteria (see above)
```

```
awk '{print $2}' HM3_b37.bim > HM3_b37.snplist.txt
```

```
plink --bfile $datafile --extract HM3_b37.snplist.txt --make-bed --noweb --out local
```

```
awk '{ if (($5=="T" && $6=="A")||($5=="A" && $6=="T")||($5=="C" && $6=="G")||($5=="G" && $6=="C")) print $2, "ambig" ; else print $2 ;}' $datafile.bim | grep -v ambig > local.snplist.txt
```

```
plink --bfile HM3_b37 --extract local.snplist.txt --make-bed --noweb --out external
```

##

Merge the two sets of plink files – In merging the two files plink will check for strand differences. If any strand differences are found plink will crash with the following error (ERROR: Stopping due to mis-matching SNPs -- check +/- strand?)

Ignore warnings regarding different physical positions

##

```
plink --bfile local --bmerge external.bed external.bim external.fam --make-bed --noweb --out HM3_b37merge
```

##

If plink crashed with a strand error (ERROR: Stopping due to mis-matching SNPs -- check +/- strand?) run the following two lines of alternate code

```
plink --bfile local --flip HM3_b37merge-merge.missnp --make-bed --noweb --out flipped
```

```
plink --bfile flipped --bmerge external.bed external.bim external.fam --make-bed --noweb --out HM3_b37merge
```

##

Run the MDS analysis

##

```
plink --bfile HM3_b37merge --cluster --mind .05 --mds-plot 4 --extract local.snplist.txt --noweb --out HM3_b37mds
```

###

Plot the MDS results using R into a file called mdsplot.eps and mdsplot.pdf (Note: type R to start R in unix and q() followed by n to close the R session after the plot has been made)

###

```
awk 'BEGIN{OFS=","};{print $1, $2, $3, $4, $5, $6, $7}' >> HM3_b37mds2R.mds.csv HM3_b37mds.mds #This formats the plink output into an R compatible format.

R

#From this point until the end of the section, you are working in the R statistical package

library(calibrate)

#If you don’t have calibrate package, install it using

#install.packages("calibrate")

mds.cluster = read.csv("HM3_b37mds2R.mds.csv", header=T);

colors=rep("red",length(mds.cluster$C1));
colors[which(mds.cluster$FID == "CEU")] <- "lightblue";
colors[which(mds.cluster$FID == "CHB")] <- "brown";
colors[which(mds.cluster$FID == "YRI")] <- "yellow";
colors[which(mds.cluster$FID == "TSI")] <- "green";
colors[which(mds.cluster$FID == "JPT")] <- "purple";
colors[which(mds.cluster$FID == "CHD")] <- "orange";
colors[which(mds.cluster$FID == "MEX")] <- "grey50";
colors[which(mds.cluster$FID == "GIH")] <- "black";
colors[which(mds.cluster$FID == "ASW")] <- "darkolivegreen";
colors[which(mds.cluster$FID == "LWK")] <- "magenta";
colors[which(mds.cluster$FID == "MKK")] <- "darkblue";
pdf(file="mdsplot.pdf",width=7,height=7)

plot(rev(mds.cluster$C2), rev(mds.cluster$C1), col=rev(colors), ylab="Dimension 1", xlab="Dimension 2",pch=20)

legend("topright", c("My Sample", "CEU", "CHB", "YRI", "TSI", "JPT", "CHD", "MEX", "GIH", "ASW","LWK", "MKK"), fill=c("red", "lightblue", "brown", "yellow", "green", "purple", "orange", "grey50", "black", "darkolivegreen", "magenta", "darkblue"))
```

#label your sample points, if you want to know the subject ID label of each sample on the graph, uncomment the value below (this is optional and you can choose not to do this if you are worried about patient information being sent; when you send us your MDS plot please make sure the subject ID labels are NOT on the graph)

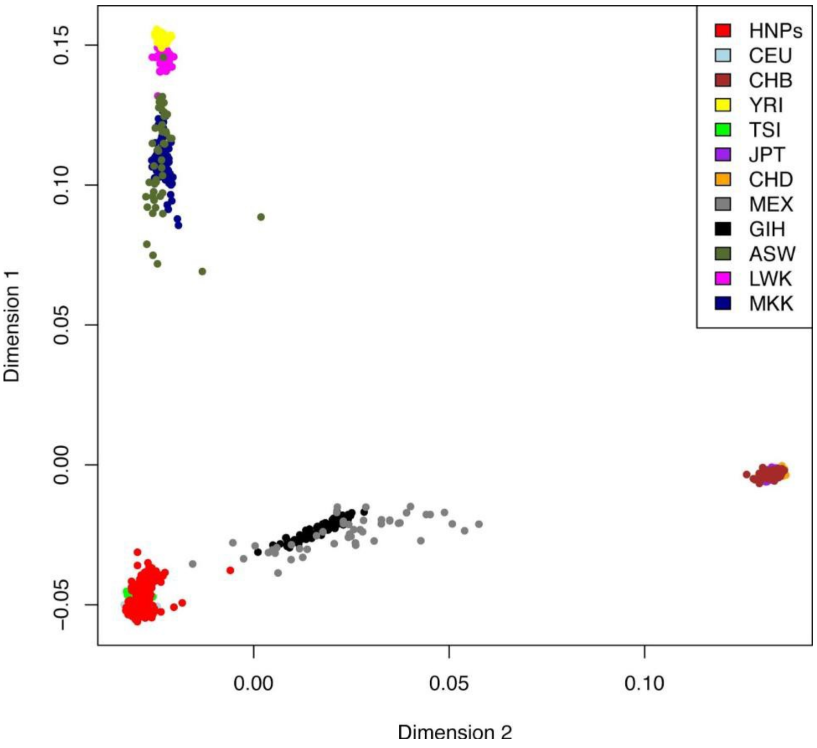
```
#FIDlabels <- c("CEU", "CHB", "YRI", "TSI", "JPT", "CHD", "MEX", "GIH", "ASW","LWK", "MKK");

#textxy(mds.cluster[which(!(mds.cluster$FID %in% FIDlabels)), "C2"], mds.cluster[which(!(mds.cluster$FID %in% FIDlabels)), "C1"], mds.cluster[which(!(mds.cluster$FID %in% FIDlabels)), "IID"])

dev.off();
```

##

Please send the newly created mdsplot.pdf and HM3_b37mds2R.mds.csv files to the authors of this protocol at the ENIGMA helpdesk: enigma.genetics@gmail.com.

##

Your output will look something like this when viewed as a PDF file:



Imputation Protocol

It is assumed that your data are stored using build 37 positions. If this is not the case you will need to remap your data. Contact the enigma help desk if you are unsure what build your data are mapped to or you need help with this.

The imputation protocol assumes you are using a bash shell. To check which shell you are using on your server type:

```
echo $SHELL
```

If you are not using bash please change to the bash shell to run the protocol by typing:

```
bash
```

when you are finished to go back to your usual shell (if you weren't already in bash) type:

```
exit
```

##

Step 0: Sign up for a free account on the Michigan Imputation Server

The Michigan Imputation Server is a free NIH funded service which provides researchers with a fast and secure way to impute their data (https://imputationserver.sph.umich.edu/).

You can sign up for a free account at
https://imputationserver.sph.umich.edu/index.html#!pages/register

If you are unable to use the imputation server please contact the helpdesk and we will discuss alternate imputation methods with you.

Step 1: Re QC your data

Before starting the imputation process you need to drop any strand ambiguous SNPs and rescreen for low MAF, missingness and HWE in your PLINK-format genotype files. Copy your PLINK-format genotype files (*.bed, *.bim, *.fam, files into your working directory) and then run the following code customising the sections that are highlighted.

```
awk '{ if (($5=="T" && $6=="A")||($5=="A" && $6=="T")||($5=="C" && $6=="G")||($5=="G" && $6=="C")) print $2, "ambig" ; else print $2 ;}' $datafile.bim | grep ambig | awk '{print $1}' > ambig.list
```

```
plink --bfile $datafile --exclude ambig.list --make-founders --out lastQC --maf 0.01 --hwe 0.000001 --make-bed --noweb
```

If your SNP data are stored with rs numbers change the SNP names from rs numbers to CHR:BP names as follows.

```
awk '{print $2,  $1":"$4}' lastQC.bim > updateSNPs.txt
```

```
plink --bfile lastQC --update-name  updateSNPs.txt --make-bed --out lastQCb37 --noweb --list-duplicate-vars
```

After modifying the SNP identifiers and running the last command, you might encounter duplicate markers, you need to remove those before going on. If there are no duplicate SNPs you may skip this step:

```
plink --bfile lastQCb37 --exclude lastQCb37.dupvar --out lastQCb37_noduplicates --make-bed --noweb
```

Step 2: Preparing your data for imputation

To convert your ped/map file into a VCF file (using either lastQCb37 or lastQCb37_noduplicates)

```
for i in {1..22}
do
plink --bfile lastQCb37_noduplicates --chr $i --recode vcf --out StudyName_chr"$i"
vcf-sort StudyName_chr"$i".vcf | bgzip -c > StudyName_chr"$i".vcf.gz
done
```

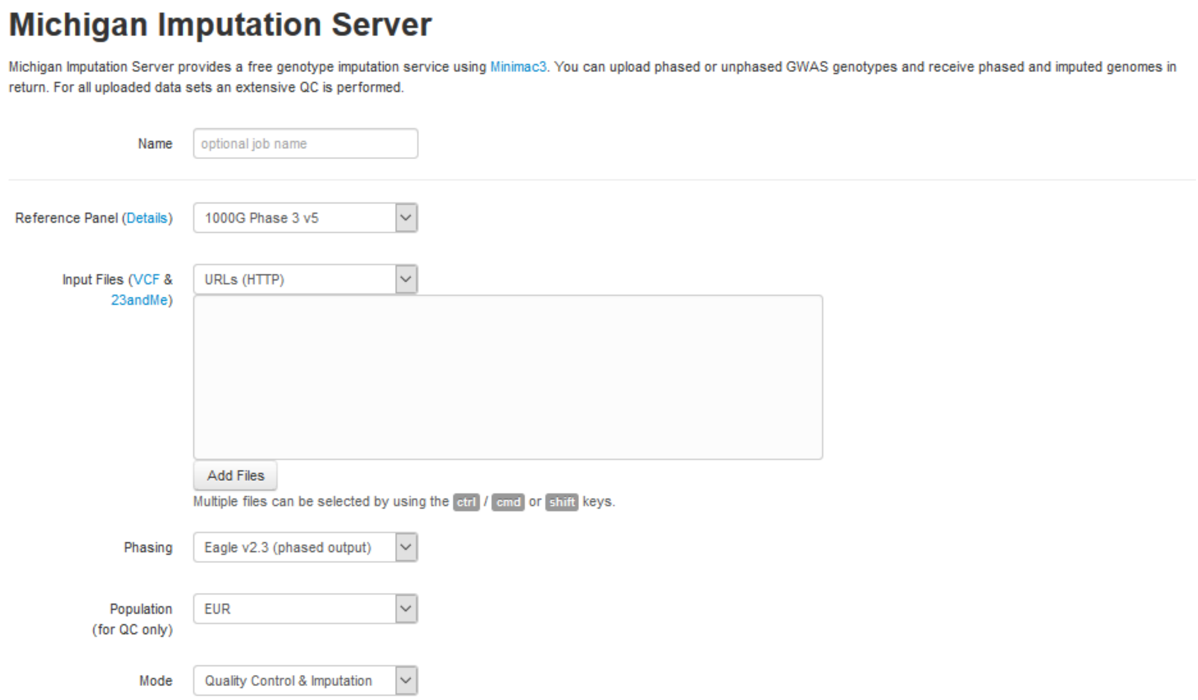
Step 3: Upload your data to the imputation server

Upload your data to the server using the instructions provided in section 3 of this page: https://imputationserver.sph.umich.edu/index.html#!pages/help

Step 4: Run the imputation

Select the following options

- **Reference panel: 1000G Phase 3 v5**
- **Phasing: Eagle v2.3**
- **Population: EUR** if your sample is of European ancestry. If not use the most appropriate reference group. This is used for the allele frequency check in quality control only.



Michigan Imputation Server

Michigan Imputation Server provides a free genotype imputation service using Minimac3. You can upload phased or unphased GWAS genotypes and receive phased and imputed genomes in return. For all uploaded data sets an extensive QC is performed.

Name: optional job name

Reference Panel (Details): 1000G Phase 3 v5

Input Files (VCF & 23andMe): URLs (HTTP)

Add Files

Multiple files can be selected by using the ctrl / cmd or shift keys.

Phasing: Eagle v2.3 (phased output)

Population (for QC only): EUR

Mode: Quality Control & Imputation

The imputation server will first check the validity of the VCF, then the QC of the data, the run the imputation. You will be able to check the progress of each stage using the JOBS tab in your account.

If you run into any problems that you can’t fix please email us for assistance.

Step 5: Download the imputed data

Once the imputation is finished you will be sent an email with an encrypted link to allow you to download the imputed data.

This link will only be valid for one week!

After this time your data will be deleted from the imputation server.