Linkage disequilibrium information calculation plan v1.1

To fine-map signals from ancestry specific meta-analyses, linkage disequilibrium (LD) information is estimated from the imputed genotypes (from either IMPUTE2 or MaCH/minimac) of the cohorts from the same ancestry for fine-mapping in FINEMAP (<http://www.christianbenner.com/)>.

**Please use the SAME 1000G imputed genotypes used for the MAGIC 1000G imputation.**

1. Fine-mapping regions

LD information is calculated within each fine-mapping region. The regions can be found in the attached **fine-mapping.bed** file with three columns chromosome index, start position and end position. For example:

2 1000001 2000000

X 2200001 5000000

1. Preparing input files for LD calculation

2.1 Imputed genotypes check (from IMPUTE2 or MaCH/minimac)

For single file of imputed genotypes including all chromosomes, separate it to separate files by chromosome with corresponding chromosome indices in the file name.

For files already separated by chromosome, go to section 2.2.

For imputed genotypes in imputation chunks (the same chunks used in the imputation), merge the imputation chunks from the same chromosome together.

**The order of the imputed variants in each chromosome should be ordered by position (smallest to largest).**

You should have two files for each chromosome. For **IMPUTE2**, one is for the imputed genotypes with suffix **gen.gz** and the other is for the imputation quality with suffix **info**. For **MaCH/minimac**, one is for the imputed dosages with suffix **dose.gz** and the other is for the imputation quality with suffix **info**.

2.2 Input file preparation

After generating files with imputed genotypes by chromosome, prepare the input files for LD calculation. The output files should be named as using the same format as shown in the examples below.

IMPUTE2:

In the info file for each chromosome, replace anything in column snp\_id with chromosome index

MaCH/minimac:

Generate file **imputed.i.vars** containing SNP ID, chromosome index and position based on the corresponding imputation info file for chromosome **i**, for example:

SNP Chr Position

2: 1000001 2 1000001

2: 1000101 2 1000101

**Make sure the dosages are calculated for Al1 in the info file**

1. LD Calculation

Download the R scripts attached

Calculate the LD in each region from the file **fine-mapping.bed** file, for example:

IMPUTE2:

Rscript computeCorrelationsImpute2forFINEMAP.r imputed.2\_info imputed.2.gen.gz 2 1000001 2000000 0 0 region2\_1000001\_2000000 20

where imputed.2\_info and imputed.2.gen.gz are the imputed results for chromosome 2 from IMPUTE2, and the R script computeCorrelationsImpute2forFINEMAP.r (with the instructions) is attached along with this file.

MaCH/minimac:

Rscript computeCorrelationsMinimac12forFINEMAP.r imputed.2.info imputed.2.dose.gz imputed.2.vars Position Chr 2 1000001 2000000 0 0 region2\_1000001\_2000000 20

where imputed.2.info and imputed.2.dose.gz are the imputed results for chromosome **2** from MaCH/minimac, and the R script computeCorrelationsMinimac12forFINEMAP.r (with the instructions) is attached along with this file.

The R scripts are modified based on that from Christian Benner the author of FINEMAP and LDstore.

**Name the results in the same way as shown in the examples above please.** The format is region**Chr**\_**start position**\_**end postion**, where **Chr**, **start position** and **end position** are those from the **fine-mapping.bed** file. The suffixes are ld.gz and map**.**

1. Upload results

Pack all the results in a single file before uploading as follows:

mkdir **COHORT\_ANCESTRY\_N\_LD\_DATE\_INITIAL**

where:

**COHORT** is replaced by the unique cohort identifier used for the GWAS.

**ANCESTRY** uses 'EA' for European Ancestry, 'AA' for African Ancestry, 'HA' for Hispanic Ancestry, 'EAA' for East Asian Ancestry, 'IAA' for Indian Asian Ancestry and ‘UAA’ for Uganda population.

**N** is the sample size used to calculate the LD.

**DATE** is replaced by the date of file generation (DDMMYYYY format).

**INITIAL** is replaced by the initials of the analyst submitting the results file.

mv region\*.ld COHORT\_ANCESTRY\_N\_LD\_DATE\_INITIAL

mv region\*.map COHORT\_ANCESTRY\_N\_LD\_DATE\_INITIAL

Then,

tar –czvf COHORT\_ANCESTRY\_N\_LD\_DATE\_INITIAL.tar.gz COHORT\_ANCESTRY\_N\_LD\_DATE\_INITIAL

md5sum COHORT\_ANCESTRY\_N\_LD\_DATE\_INITIAL.tar.gz > COHORT\_ANCESTRY\_N\_LD\_DATE\_INITIAL.md5

Upload COHORT\_ANCESTRY\_N\_LD\_DATE\_INITIAL.tar.gz and COHORT\_ANCESTRY\_N\_LD\_DATE\_INITIAL.md5 to the Sanger sftp at this location: /gluinsrelatedtraits/LDstore\_output

1. Appendix

For results from minimac3 in M3VCF format, use DosageConvertor in http://genome.sph.umich.edu/wiki/DosageConvertor to convert the VCFs to dose files. For example,

DosageConvertor --vcfDose myfile.dose.vcf.gz --info myfile.info --prefix mydosefile --type mach --format DS --buffer 10000

grep -v SNP [mydosefile.info](http://mydosefile.info)|awk 'BEGIN{print "SNP\tAl1\tAl2\tFreq1\tMAF\tQuality\tRsq"};{if($3==$5) print $1"\t"$3"\t"$2"\t"$6"\t"$6"\t.\t"$8; else print  $1"\t"$3"\t"$2"\t"1-$6"\t"$6"\t.\t"$8;}'  > mydosefile

mv mydosefile mydosefile.info

For any questions/issues with the analysis plan, please email Ji on [jc34@sanger.ac.uk](mailto:jc34@sanger.ac.uk)

For access to the sftp, please email Ellie on [ew2@sanger.ac.uk](mailto:ew2@sanger.ac.uk)