R script: 14\_Fig8b\_LDplot.R

Need: 8kb region on Chromosome 16 from337285 To 355042

on linux:

bgzip Documents/GitRepos/BcSolGWAS/data/genome/big\_set\_v97iso\_SNPs\_filtered\_qual30\_dp6\_maf20\_recode.vcf

install vcftools

vcftools --gzvcf Documents/GitRepos/BcSolGWAS/data/genome/big\_set\_v97iso\_SNPs\_filtered\_qual30\_dp6\_maf20.recode.vcf.gz --chr Chromosome16 --recode --out Documents/GitRepos/BcSolGWAS/data/genome/chr16\_analysis/chr16\_analysis

this keeps only 12309 sites, much more manageable.

I’ll also make a shorter fragment with only 8kb around the gene of interest:

vcftools --gzvcf Documents/GitRepos/BcSolGWAS/data/genome/big\_set\_v97iso\_SNPs\_filtered\_qual30\_dp6\_maf20.recode.vcf.gz --chr Chromosome16 --from-bp 342785 --to-bp 349542 --recode --out Documents/GitRepos/BcSolGWAS/data/genome/chr16\_analysis/chr16\_analysis\_seg

then go from abbreviated vcf to DNAbin (with full vcf, this crashes Rstudio on my PC)

<https://knausb.github.io/vcfR_documentation/dnabin.html>

then extract haplotype from DNAbin file

but, pegas has a dependency problem on ubuntu → run on PC

Population and Evolutionary Genetics Analysis System

<https://www.rdocumentation.org/packages/pegas/versions/0.10/topics/haplotype>

Plotting: snp.plotter

<https://cran.r-project.org/web/packages/snp.plotter/vignettes/using_snp_plotter.html>

<https://cran.r-project.org/web/packages/snp.plotter/index.html>

[https://academic.oup.com/bioinformatics/article/23/6/774/415503/snp-plotter-an-R-based-SNP-haplotype-association#6145357](https://academic.oup.com/bioinformatics/article/23/6/774/415503/snp-plotter-an-R-based-SNP-haplotype-association" \l "6145357)

or: LDheatmap

or: LocusZoom

<http://locuszoom.sph.umich.edu/genform.php?type=yourdata>

Input type: PLINK data

Another, simple option using PLINK: <http://www.molecularecologist.com/2015/04/visualizing-linkage-disequilibrium-in-r/>

Or: SNAP (SNP Annotation and Proxy Search)

<http://archive.broadinstitute.org/mpg/snap/>

Trying again: to get files for SNP.plotter

https://cran.r-project.org/web/packages/snp.plotter/vignettes/using\_snp\_plotter.html

Need:

SNP.FILE

* **SNP.FILE:** SNP.FILE includes four necessary columns ASSOC, SNP.NAME, LOC, and SS.PVAL corresponding to positive or negative association (indicating protective or susceptibility alleles, a SNP label, the location, and a p-value for each SNP. SNP labels may not start with numbers. In the figure, SNPs are indicated by the symbols chosen for the data, if symbol type 'NA' is specified, the SNP.FILE ASSOC column is read and an up-triangle and down-triangle are used to indicate positive and negative association (indicating susceptibility or protective alleles), respectively.

ASSOC SNP.NAME LOC SS.PVAL

+ rs10\_8 126272509 0.065

- rs11\_8 126274467 0.029

+ rs12\_8 126275017 0.046

- rs13\_8 126275750 0.005

I can make this from: HEM.topSNPsSM in my 14\_Fig8b plot file

Also Need:

* **GENOTYPE.FILE:** GENOTYPE.FILE is a modified Linkage PED file. Each row should have the following information: family ID, individual ID, father ID, mother ID, sex, and affection status followed by marker loci coded as binary factors, as shown in the example below. This file should not have column headers.

1 1 0 0 1 1 1 1 1 1

2 1 0 0 2 1 1 1 1 1

3 1 0 0 1 1 1 1 1 2

4 1 0 0 2 1 1 1 2 2

5 1 0 0 1 1 0 0 1 1

I can make this from: my PLINK / GEMMA work files. Get off Linux computer? From script: 01\_TABtoPED.R

Also Need:

* **HAP.FILE:** HAP.FILE includes three necessary columns ASSOC, G.PVAL, and I.PVAL corresponding to positive or negative association (indicating protective or susceptibility alleles, a global p-value and an individual p-value for each haplotype followed by a set of columnns of SNPs with corresponding haplotypes. Haplotypes are presented in a step-wise fashion with the major allele given as 1 and the minor allele as 2; haplotype variants for a set of SNPs should be grouped. SNP labels in HAP.FILE must be the same as in SNP.FILE, and only SNPs with corresponding haplotypes need to be included. In the figure, unfilled symbols connected by solid lines are used to indicate global haplotype p-values, (a circle is used if no symbol is specified for the dataset). Unfilled and filled symbols are used to indicate alleles 1 and 2, respectively connected by solid lines and dashed lines for positive and negative association (indicating susceptibility or protective haplotypes) when using indivudal haplotype p-values.

ASSOC G.PVAL I.PVAL rs10\_8 rs11\_8 rs12\_8 rs13\_8 rs14\_8 rs15\_8 rs16\_8 rs17\_8

- 0.015 0.004 1 1 1

+ 0.015 0.062 1 2 2

+ 0.075 0.079 1 1 1

+ 0.075 0.039 2 2 2

- 0.032 0.121 1 1 1

+ 0.032 0.153 1 2 2

+ 0.425 0.474 1 1 1

+ 0.425 0.003 2 2 2

+ 0.1 0.077 1 1 1

+ 0.1 0.1 1 2 2

- 0.003 0.341 1 1 1

+ 0.003 0.001 2 2 2

I should be able to get this through PLINK. This is the trickiest bit…

<http://zzz.bwh.harvard.edu/plink/haplo.shtml>

First need to generate haplotypes from sliding window, which requires –bfile

<https://www.cog-genomics.org/plink2/formats>

this means: first generate a matching .bed, .bim, and .fam file set, then run it through the haplotype sliding window

## Specification of haplotypes to be estimated

Haplotype testing in PLINK requires that the user supplies a file listing the haplotypes to be tested (Some precomputed lists are given [below](http://zzz.bwh.harvard.edu/plink/haplo.shtml" \l "precomputed) which might be useful in some circumstances.) The formats of these files are described below. An alternative is to specify a simple, sliding window of fixed haplotype size (also described below).

The command

##### plink --file mydata --hap myfile.hlist

will read the file myfile.hlist, each row of which is expected to have one of the three following formats:

…

This seems easiest (automated 3-SNP window analysis)

***4) Sliding window specification***

Finally, instead of specifying a haplotype file with the --hap option, you can use the --hap-window option to specifty all haplotypes in sliding windows of a fixed number of SNPs (shifting 1 SNP at a time).

plink --bfile mydata --hap-window 3 --hap-assoc

to form all 3-SNP haplotypes across the entire dataset (respecting chromosome boundaries, however). In this case the windows will be automatically named WIN1, WIN2, etc. This command can take a comma-delimited list of values, e.g.

--hap-window 1,2,3

to perform all single SNP tests (1-SNP haplotypes) as well as sliding windows of all 2-SNP and 3-SNP haplotypes.

BUT PLINK 1.9 does not support this, have to go via BEAGLE.

Download BEAGLE:

[https://faculty.washington.edu/browning/beagle/beagle.html#download](https://faculty.washington.edu/browning/beagle/beagle.html" \l "download)

copy to Documents/

give it executable permissions:

chmod u+x beagle…

recommend using BEAGLE instead of PLINK for case/control haplotype association

analysis.

--recode beagle to

GitRepos/BcSolGWAS/data/genome/chr16\_analysis/myCHR16\_A.chr-\*.dat +

GitRepos/BcSolGWAS/data/genome/chr16\_analysis/myCHR16\_A.chr-\*.map

Also going to try PLINK 1.07

THEN

plink --file mydata --hap myfile.hlist --hap-assoc

which will generate haplotype-specific tests (1df) for both disease and quantitative traits; for disease traits only, an omnibus association statistic will also be computed. This option generates the file

plink.assoc.hap

which contains the following fields:

LOCUS Haplotype locus / window name

HAPLOTYPE Haplotype identifer / "OMNIBUS"

F\_A Frequency in cases

F\_U Frequency in controls

CHISQ Test for association

DF Degrees of freedom

P Asymptotic p-value

SNPS SNPs forming the haplotype

So use LOCUS / HAPLOTYPE for each row in HAP.FILE

And p for I.PVAL

Not sure what to use for G.PVAL

And can use SNPS for filling in the SNP lists in HAP.FILE

FYI, for multiple phenotypes in PLINK:

## Can I analyse multiple phenotypes in a single run (e.g. for gene expression datasets)?

For most association commands, you can specify the --all-pheno option to automatically loop over all phenotypes in an alternate phenotype file:

##### plink --bfile mydata --pheno phenos.raw --all-pheno --linear --covar covar.dat

If there are N phenotypes, this will generate N separate output files. If a header row was supplied in the alternate phenotype file, then each file will have the phenotype name appended (it is up to the user therefore to ensure that the phenotype names are unique). If not, the output files are simply numbered, P1, P2, etc, (e.g. plink.P1.assoc, etc).

This works for most basic association commands that consider all SNPs (e.g. --assoc, --logistic, --fisher, --cmh, etc) but currently not for any haplotype analysis or epistasis options.

CITATIONS

The citation for Beagle’s phasing algorithm is: S R Browning and B L Browning (2007). Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. Am J Hum Genet 81:1084-97. doi:10.1086/521987

The citation for Beagle’s imputation algorithm is: S R Browning and B L Browning (2016). Genotype imputation with millions of reference samples. Am J Hum Genet 98:116-126. doi:10.1086/j.ajhg.2015.11.020

The citation for Beagle’s IBD detection algorithm is: B L Browning and S R Browning (2013). Improving the accuracy and efficiency of identity by descent detection in population data. Genetics 194(2):459-71. doi:10.1534/genetics.113.150029

Package: PLINK (including version number)

Author: Shaun Purcell

URL: http://pngu.mgh.harvard.edu/purcell/plink/

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR,

Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC (2007)

PLINK: a toolset for whole-genome association and population-based

linkage analysis. American Journal of Human Genetics, 81.