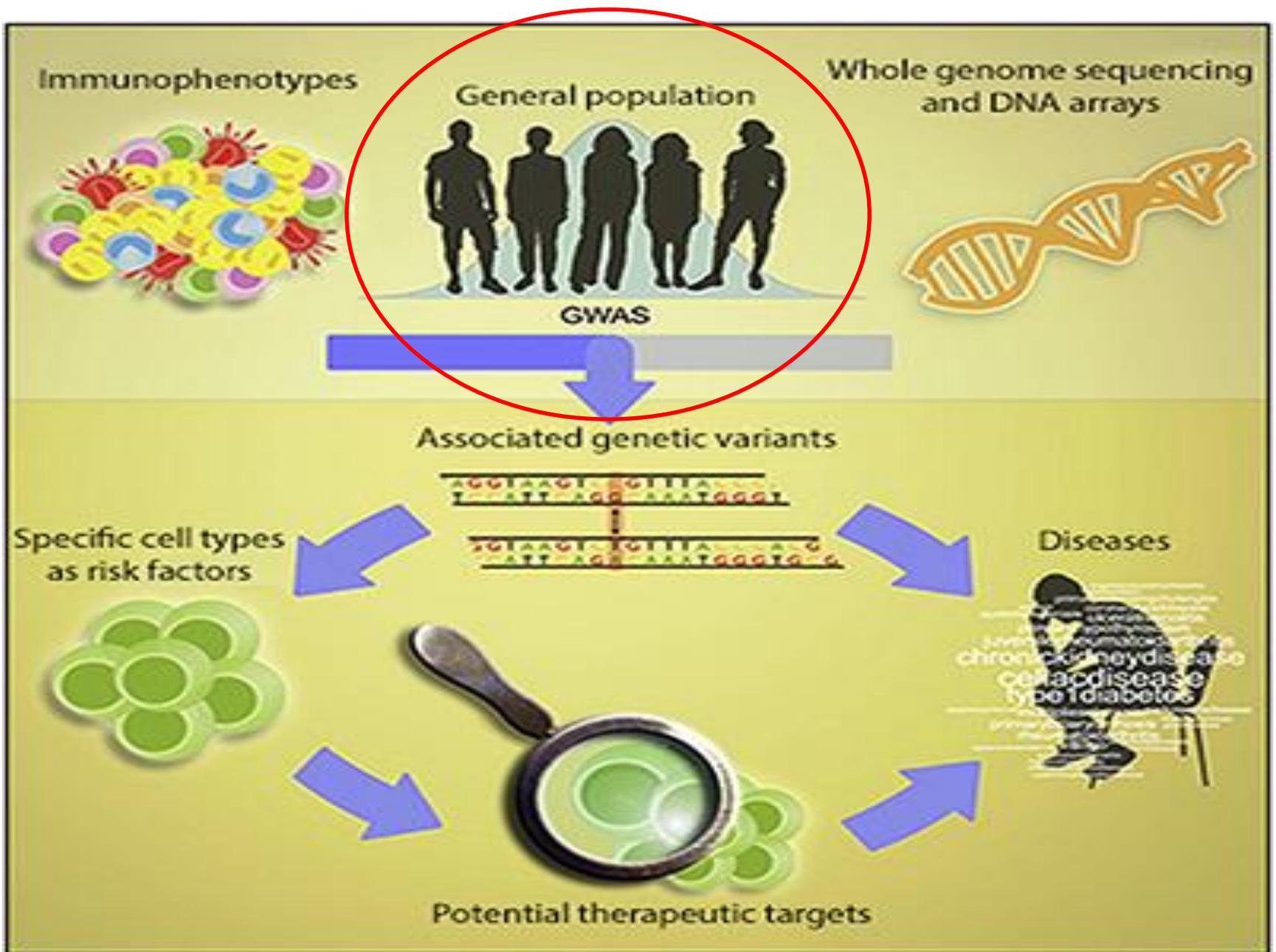# Genetic Mapping

GBIO0002

Archana Bhardwaj
University of Liege

Genetic Mapping

▪As the name suggests, it refers to locating genes or genetic information on a genetic map (possibly subdivided by chromosomal regions)

▪A genetic map describes the order of genes or genetic markers, and the spacing between each, on chromosomes

▪Scientists isolate DNA and use genetic markers to find genomic locations that can be linked / associated with a trait of interest (e.g., in humans: disease trait, in plants: yield);
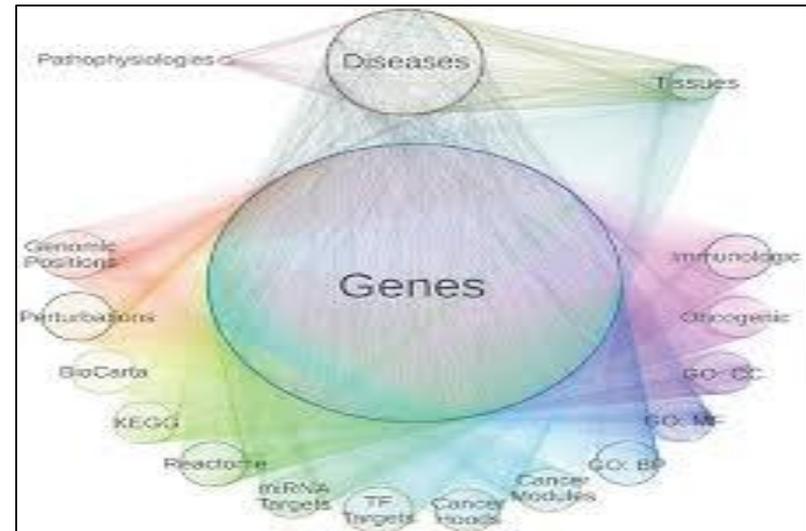
Uses of Genetic Mapping

➤ Identify genes that are responsible for traits of interest:

- ▪ Humans
- ▪ Animals
- ▪ Plants

➤ Understanding biological
mechanisms related to
the trait of interest

Genetic Mapping in Human Disease

David Altshuler,[1,2,3,4,5*] Mark J. Daly,[1,2,5*] Eric S. Lander[1,6,7,8*]

Genetic mapping provides a powerful approach to identify genes and biological processes underlying any trait influenced by inheritance, including human diseases. We discuss the intellectual foundations of genetic mapping of Mendelian and complex traits in humans, examine lessons emerging from linkage analysis of Mendelian diseases and genome-wide association studies of common diseases, and discuss questions and challenges that lie ahead.

By the early 1900s, geneticists understood that Mendel's laws of inheritance underlie the transmission of genes in diploid organisms. They noted that some traits are inherited according to Mendel's ratios, as a result of alterations in single genes, and they developed methods to map the genes responsible. They also recognized that most naturally occurring trait variation, while showing strong correlation among relatives, involves the action of multiple genes and nongenetic factors.

Although it was clear that these insights applied to humans as much as to fruit flies, it took most of the century to turn these concepts into practical tools for discovering genes contributing to human diseases. Starting in the 1980s, the use of naturally occurring DNA variation as markers to trace inheritance in families led to the discovery of thousands of genes for rare Mendelian diseases. Despite great hopes, the approach proved unsuccessful for common forms of human diseases—such as diabetes, heart disease, and cancer—that show complex inheritance in the general population.

Over the past year, a new approach to genetic mapping has yielded the first general progress toward mapping loci that influence susceptibility to common human diseases. Still, most of the genes and mutations underlying these findings

by Sturtevant for fruit flies in 1913 (1). Linkage analysis involves crosses between parents that vary at a Mendelian trait and at many polymorphic variants ("markers"); because of meiotic recombination, any marker showing correlated segregation ("linkage") with the trait must lie nearby in the genome.

In the 1970s, the ability to clone and sequence DNA made it possible to tie genetic linkage maps in model organisms to the underlying DNA sequence, and thereby to molecularly clone the genes responsible for any Mendelian trait solely on the basis of their genomic position (2, 3). Such studies typically involved three steps: (i) identifying the locus responsible through a genome-wide search; (ii) sequencing the region in cases and controls to define causal mutation(s); and (iii) studying the molecular and cellular functions of the genes discovered. So-called "positional cloning" became a mainstay of experimental genetics, identifying pathways that are crucial in development and physiology.

Linkage analysis in humans. For most of the 20th century, genome-wide linkage mapping was impractical in humans: Family sizes are small, crosses are not by design, and there were too few classical genetic markers to systematically trace inheritance. Progress in identifying the genes

of previous knowledge. (ii) Disease-causing mutations often cause major changes in encoded proteins. (iii) Loci typically harbor many disease-causing alleles, mostly rare in the population. (iv) Mendelian diseases often revealed great complexity, such as locus heterogeneity, incomplete penetrance, and variable expressivity.

Geneticists were eager to apply genetic mapping to common diseases, which also show familial clustering. Mendelian subtypes of common diseases [such as breast cancer (15), hypertension (16), and diabetes (17)] were elucidated, but mutations in these genes explained few cases in the population. In common forms of common disease, risk to relatives is lower than in Mendelian cases, and linkage studies with excellent power to detect a single causal gene yielded equivocal results.

These features were consistent with, but did not prove, a polygenic model. The idea that commonly varying traits might be polygenic in nature was offered by East in 1910 (18). By 1920, linkage mapping was used to identify multiple unlinked factors influencing truncate wings in *Drosophila* (19), and Fisher had developed a mathematical framework for relating Mendelian factors and quantitative traits (20). In the late 1980s, linkage mapping of complex traits was made feasible for experimental organisms through the use of genetic mapping in large crosses (21). But there was little success in humans.

Genetic association in populations. A possible path forward emerged from population genetics and genomics. Instead of mapping disease genes by tracing transmission in families, one might localize them through association studies—that is, comparisons of frequencies of genetic variants among affected and unaffected individuals. Genetic association studies were not a new

Human Genome Statistics

- Number of Chromosomes : 23 pairs
- Genome Size : 3,079,843,747 Base pairs
- No of Genes : 32,185

autosomes

sex chromosomes
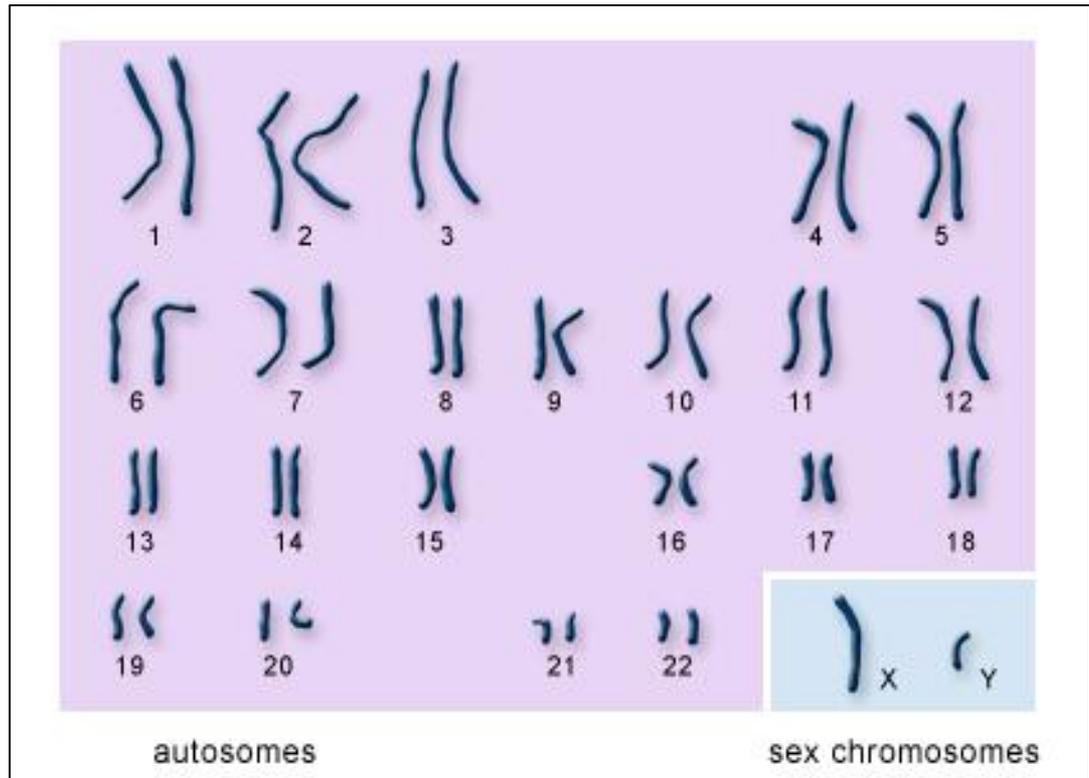
Tools For GWAS Analysis

1. R Package - LEA

2. PLINK

3. DEPICT

4. Association Viewer

Running Structure-like Population Genetic Analyses with R

Olivier François
olivier.francois@grenoble-inp.fr

June 2016

Summary

This short tutorial explains how population structure analyses reproducing the results of the widely-used computer program STRUCTURE can be performed using commands in the R language. The method works for any operating systems, and it does not require the installation of STRUCTURE or additional computer programs. The R program allows running population structure inference algorithms, choosing the number of clusters, and showing admixture coefficient bar-plots using a few commands. The methods used by R are fast and accurate, and they are free of standard population genetic equilibrium hypotheses. In addition, these methods allow their users to play with a large panel of graphical functions for displaying pie-charts and interpolated admixture coefficients on geographic maps.

R package: LEA

- First install dependency package
  install.packages(c ("fields","RColorBrewer","mapplots"))

- Now install LEA R package
  - *source("http://bioconductor.org/biocLite.R")*
  - *biocLite("LEA")*

- *Now check if LEA is properly installed and type*
  - *library(LEA)*
- *Install format conversion R script*
  source("http://membres-timc.imag.fr/Olivier.Francois/Conversion.R")

LEA: Parameters Descriptions

▪(TESS = TRUE) - Genotypic information is available

▪(TESS = FALSE) – No Genotypic information

▪FORMAT = 1 must be used when markers are encoded using one row of data for each individual.

▪ FORMAT = 2 means that markers are encoded using two rows of data for each individual

▪The extra.row and extra.col variables are integers indicating the number of extra rows and columns. Extra columns include all the columns that do not contain genotypic .

▪Output defines name of output file to store the output

LEA: Running R Package

- Genetic data for populations in Europe (Durand et al. 2009). The data consists of 60 population samples of 10 diploid individuals that were genotyped at 100 multi-allelic markers.

- Download the secondary_contact.str file from course website

- Define secondary_contact.str as input file in struct2geno function

  - *struct2geno(file= "/fullpath/secondary_contact.str", TESS = TRUE, diploid = TRUE, FORMAT = 2, extra.row = 0, extra.col = 0, output = "secondary_contact.geno")*

  - **Check if output file has been created in working directory**

- let us run a population structure analysis that assumes K = 3 clusters. This can be done by using the snmf function of the LEA package.

  - *obj.snmf = snmf("secondary_contact.geno", K = 3, alpha = 100, project = "new")*

- The qmatrix object (by function Q) contains the matrix of ancestry coefficients for each individual and for K = 3 clusters
  - *qmatrix = Q(obj.snmf, K = 3)*

- Create bar plot to see samples data populations structure

- *barplot(t(qmatrix), col = c("orange","violet","lightgreen"), border = NA, space = 0, xlab = "Individuals", ylab = "Admixture coefficients")*
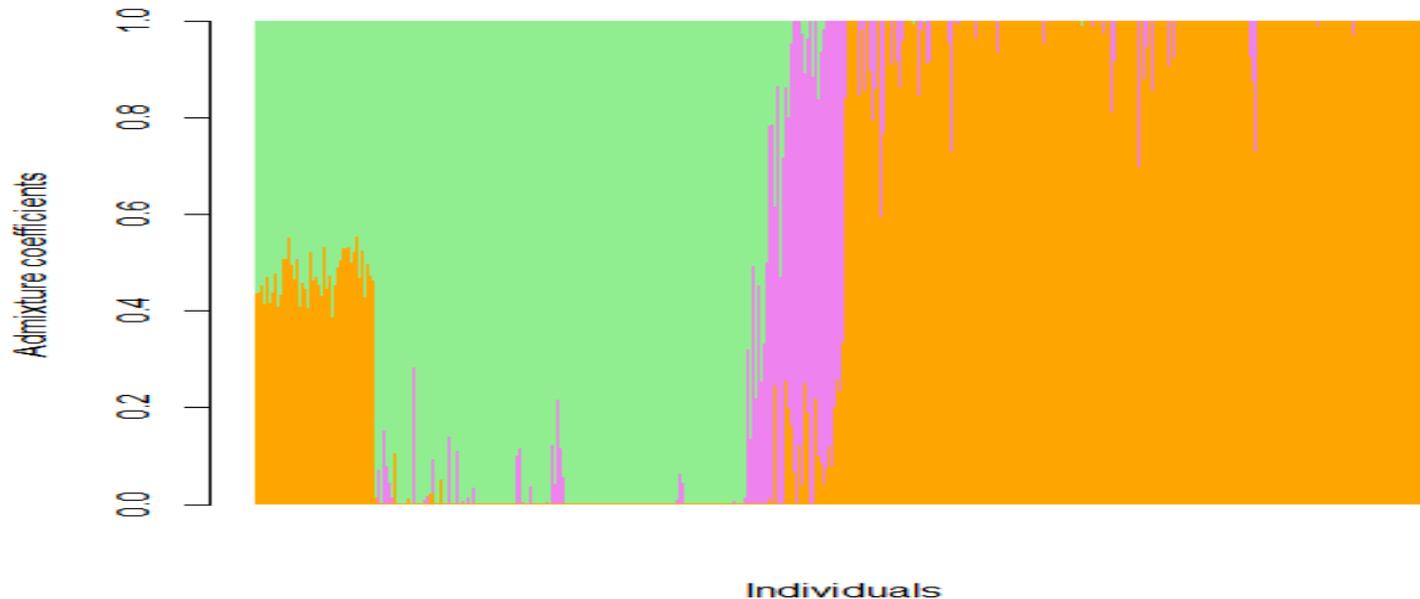
▪ **Let us try K=4**

▪ Now, we will change value of K=4 and perform the structure analysis

- *obj.snmf = snmf("secondary_contact.geno", K = 4, alpha = 100, project = "new")*
- *qmatrix = Q(obj.snmf, K = 4)*
- *barplot(t(qmatrix), col = c("orange","violet","lightgreen","blue"), border = NA, space = 0,xlab = "Individuals", ylab = "Admixture coefficients")*
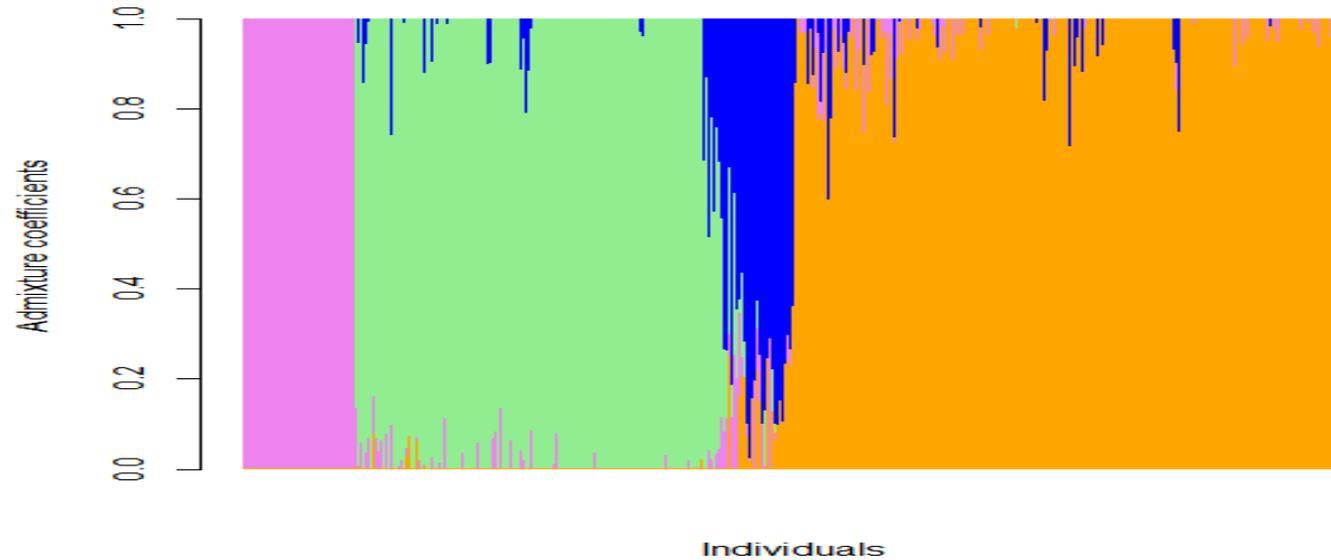
- **Comparison of plots K=3 and K=4??**

- **How to analyse better K value ?**

Cross-entropy

- Choosing the number of clusters is based on the cross-entropy criterion.

- The cross-entropy criterion is based on the prediction of a fraction of masked genotypes (matrix completion), and on the cross-validation approach.

- **Smaller values of the cross-entropy criterion usually mean better runs**

- Let us calculate cross-entropy value at K ranged from 1:8 for our input file *"secondarycontact.geno"*

  - *obj.snmf = snmf("secondarycontact.geno", K = 1:8, ploidy = 2, entropy = T, alpha = 100, project = "new")*

- Plot the data for visualization

  plot(obj.snmf, col = "blue4", cex = 1.4, pch = 19)

- **Here, we will choose the value of K for which the cross-entropy curve exhibits a plateau**
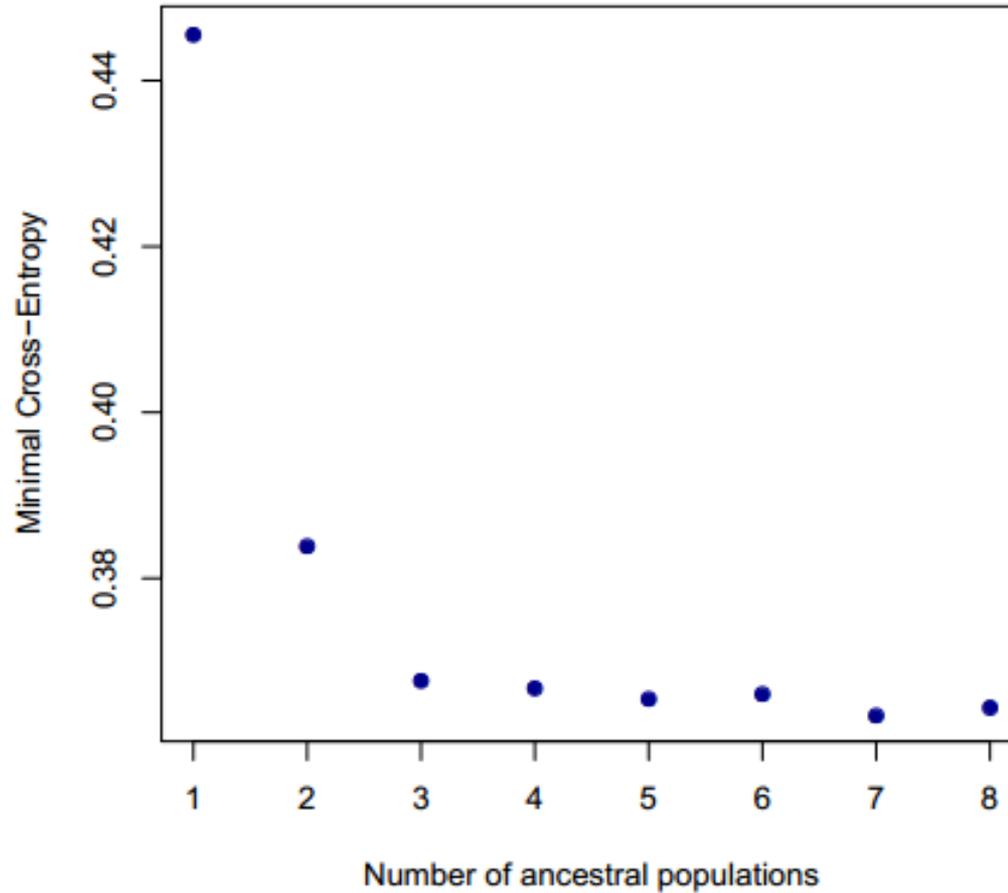
Retained value of K is 3

Let us Try Another Dataset

▪Download file from A_thaliana_chr1.geno from course website

▪Let us calculate cross-entropy value at K ranged from 1:10 for our input file "A_thaliana_chr1.geno"

obj.at = snmf("A_thaliana_chr1.geno", K = 1:10, ploidy = 1, entropy = T, CPU = 1, project = "new")

▪Plot the cross entropy values
plot(obj.at, col = "blue4", cex = 1.4, pch = 19)
What is best K value ??

QUESTIONS?

PLINK: Introduction

- PLINK is whole genome association analysis tool
- PLINK has a well-documented manual to explain all features
- PLINK is available for Linux, Mac OS and MS-DOS
- gPLINK is the other version of PLINK that provides the graphical user interface
- Command line version is faster than graphical PLINK

PLINK: Download

- To download PLINK:
- http://zzz.bwh.harvard.edu/plink/dist/plink-1.07-dos.zip
- Uncompress the PLINK-1.07-dos.zip
- Click on the folder. There are two files
    - – test.map contains the marker information

    - – test.ped contains genotype data and sample information

PLINK: File Formats(1/2)

■MAP Format

Each line of the MAP file describes a single marker and must contain exactly 4 columns:
* chromosome (1-22, X, Y or 0 if unplaced)
* rs# or snp identifier
* Genetic distance (morgans)
* Base-pair position (bp units)

PLINK: File Formats(2/2)

▪PED Format. This file is a white-space (space or tab) delimited file: the first six columns are mandatory:

Family ID, Individual ID,Paternal ID,Maternal ID,Sex (1=male; 2=female; other=unknown), Phenotype

▪Binary format:
  -> BED, BIM, and FAM

▪ Transposed text format :
  ->TPED and TFAM

PLINK: Command Line Run

- Type command :

    plink   --file test

- For binary format (BED, BIM, and FAM)

        plink --bfile test

- For transposed text format(TPED, and TFAM).
Note that all files must have the same name,
otherwise we need to clearly indicate by using
--tped and –tfam

        plink --tfile test

Format Conversion

- To convert or to indicate output as text format (PED and MAP)

     Plink --file test --recode --out test_ped

- To convert or to indicate output as TPED and TFAM

  Plink --file test --transpose --recode --out test_tp

- To convert or to indicate output as Binary format TPED and TFAM

    Plink –file –make-bed –out test_bin

Example data

▪ Download the example data from the course website

– TSI_JPT_chr20_case_control.bed
– TSI_JPT_chr20_case_control.bim
– TSI_JPT_chr20_case_control.fam
– TSI_JPT_chr20_pheno_header.txt
– TSI_JPT_chr20_pheno.txt

Data processing for SNPs(1/2)

▪ To get a set of SNPs, specify a single SNP and optionally, also ask for all SNPs in surrounding region, within the –window option

  plink --bfile mydata --snp rs652423 --window 20

▪ It will extract only SNPs within +/- 20kb of rs652423 based on multiple SNPs and ranges (--snps)

▪ To exclude some sets of SNPs

    Plink --bfile data –extract mysnps.txt

▪ Here, the file is mysnp.txt and –extract option will extract defined SNPs, one per line.

Data processing for SNPs(2/2)

▪ The –snps command will accept a comma-delimited list of SNPs, including ranges based on physical position. For example,

  plink --bfile mydata –snps rs273744, rs89883,rs12345- rs67890,rs999,rs222

▪Based on physical position (--from-kb, etc)

plink --bfile mydata --chr 2 --from-kb 5000 --to-kb 10000

  It will select all SNPs within this 5000kb region on chromosome 2.

Quality control processes

- Missing genotype

- Hardy-Weinberg Equilibrium

- Minor Allele frequency

- Linkage disequilibrium pruning

Missing Genotypes

▪To generate a list genotyping/missingness rate statistics:

 plink --bfile data --missing This option creates two files:

- plink.imiss
- plink.lmiss

▪It provides the detail missingness by individual and by SNP (locus), respectively.

Clustering based on Missing Genotypes

▪Systematic batch effects that induce missingness in parts of the sample will induce correlation between the patterns of missing data that different individuals display.

▪One approach to detect correlation in these patterns, that might possibly identify such biases, is to cluster individuals based on their identity-by-missingness (IBM).

- plink --bfile data --cluster-missing

▪which creates the files:

- *plink.matrix.missing*
- *plink.cluster3.missing*

which have similar formats to the corresponding IBS clustering files.

Missing Rate Per Person

▪The initial step in all data analysis is to exclude individuals with too much missing genotype data. This option is set as follows:

• plink --bfile mydata --mind 0.1

which means exclude with more than 10% missing genotypes.

▪A line in the terminal output will appear, indicating how many individuals were removed due to low genotyping. If any individuals were removed, a file called **plink.irem** will be created, listing the Family and Individual IDs of these removed individuals.

Missing Rate Per SNP

▪Subsequent analyses can be set to automatically exclude SNPs on the basis of missing genotype rate, with the --geno option: the default is to include all SNPS (i.e. --geno 1).

▪To include only SNPs with a 90% genotyping rate (10% missing) use

- *plink --bfile mydata --geno 0.1*

▪As with the --maf option, these counts are calculated after removing individuals with high missing genotype rates.

Hardy-Weinberg Equilibrium (1/2)

▪To generate a list of genotype counts and Hardy-Weinberg test statistics for each SNP, use the option:

• *plink --bfile data --hardy*

which creates a file: **plink.hwe.** The file has the following format

SNP SNP identifier

TEST Code indicating sample

A1   Minor allele code

A2   Major allele code

GENO Genotype counts:11/12/22

O(HET) observed hetrozygosity

E(HET) Expected hetrozygosity

P   H-W p-value

Hardy-Weinberg Equilibrium (2/2)

▪To exclude markers that failure the Hardy-Weinberg test at a specified significance threshold, use the option:

- plink --file mydata --hwe 0.001

▪By default this filter uses an exact test. The standard asymptotic (1 df genotypic chi-squared test) can be requested with the --hwe2 option instead of --hwe.

▪The following output will appear in the console window and in **plink.log,** detailing how many SNPs failed the Hardy-Weinberg test, for the sample as a whole, and (when PLINK has detected a disease phenotype) for cases and controls separately:

Writing Hardy-Weinberg tests (founders-only) to [plink.hwe]

30 markers failed HWE test (p <= 0.05) and have been excluded

    34 markers failed HWE test in cases

    30 markers failed HWE test in controls

Allele Frequency

▪To generate a list of minor allele frequencies (MAF) for each SNP, based on all founders in the sample:

- plink --file data --freq

▪ This will create a file: **plink.frq** with five columns:

CHR   Chromosome
SNP   SNP identifier
A1   Allele 1 code (minor allele)
A2   Allele 2 code (major allele)
MAF   Minor allele frequency
NCHROBS Non-missing allele count

Minor Allele Frequency

▪Once individuals with too much missing genotype data have been excluded, subsequent analyses can be set to automatically exclude SNPs on the basis of MAF (minor allele frequency):

- *plink --file mydata --maf 0.05*

▪ It means only include SNPs with MAF >= 0.05. The default value is 0.01. This quantity is based only on founders (i.e. individuals for whom the paternal and maternal individual codes and both 0).

▪This option is appropriately counts alleles for X and Y chromosome SNPs.

Linkage disequilibrium pruning (1/4)

▪Sometimes it is useful to generate a pruned subset of SNPs that are in approximate linkage equilibrium with each other. This can be achieved via two commands:
--indep which prunes based on the variance inflation factor (VIF), which recursively removes SNPs within a sliding window;
 -- indep-pairwise which is similar, except it is based only on pairwise genotypic correlation.

▪The VIF pruning routine is performed:
 plink --file data --indep 50 5 2
will create files   **plink.prune.in** and   **plink.prune.out**

Linkage disequilibrium pruning (2/4)

▪Each is a simlpe list of SNP IDs; both these files can subsequently be specified as the argument for a --extract or --exclude command.

▪The parameters for --indep are: window size in SNPs (e.g. 50), the number of SNPs to shift the window at each step (e.g. 5), the VIF threshold. The VIF is $1/(1-R^2)$ where R^2 is the multiple correlation coefficient for a SNP being regressed on all other SNPs simultaneously.

▪That is, this considers the correlations between SNPs but also between linear combinations of SNPs.

Linkage disequilibrium pruning (3/4)

▪The second procedure is performed:

• plink --file data --indep-pairwise 50 5 0.5

▪This generates the same output files as the first option; the only difference is that a simple pairwise threshold is used.

▪The first two parameters (50 and 5) are the same as above (window size and step); the third parameter represents the r^2 threshold.

Linkage disequilibrium pruning (4/4)

▪To give a concrete example: the command above that specify, 50 5 0.5 would

a) consider a window of 50 SNPs

b) calculate LD between each pair of SNPs in the window

c) remove one of a pair of SNPs if the LD is greater than 0.5

d) shift the window 5 SNPs forward and repeat the procedure.

Association Analysis

- Case/control

- Fisher's exact Full model

- Multiple-testing correction

Basic case/control association test

To perform a standard case/control association analysis, use the option:

    plink --file mydata --assoc

which generates a file

    plink.assoc

which contains the fields:

CHR	Chromosome
SNP	SNP ID
BP	Physical position (base-pair)
A1	Minor allele name (based on whole sample)
F_A	Frequency of this allele in cases
F_U	Frequency of this allele in controls
A2	Major allele name
CHISQ	Basic allelic test chi-square (1df)
P	Asymptotic p-value for this test
OR	Estimated odds ratio (for A1, i.e. A2 is reference)

Fisher's Exact test (allelic association)

To perform a standard case/control association analysis using Fisher's exact test to generate significance, use the option:

    plink --file mydata --fisher

which generates a file

    plink.fisher

which contains the fields:

CHR	Chromosome
SNP	SNP ID
BP	Physical position (base-pair)
A1	Minor allele name (based on whole sample)
F_A	Frequency of this allele in cases
F_U	Frequency of this allele in controls
A2	Major allele name
P	Exact p-value for this test
OR	Estimated odds ratio (for A1)

As described below, if --fisher is specified with --model as well, PLINK will perform genotypic tests using Fisher's exact test.

Adjustment for multiple testing

To generate a file of adjusted significance values that correct for all tests performed and other metrics, use the option:

    plink --file mydata --assoc --adjust

which generates the file

    plink.adjust

which contains the fields

CHR	Chromosome number
SNP	SNP identifer
UNADJ	Unadjusted p-value
GC	Genomic-control corrected p-values
BONF	Bonferroni single-step adjusted p-values
HOLM	Holm (1979) step-down adjusted p-values
SIDAK_SS	Sidak single-step adjusted p-values
SIDAK_SD	Sidak step-down adjusted p-values
FDR_BH	Benjamini & Hochberg (1995) step-up FDR control
FDR_BY	Benjamini & Yekutieli (2001) step-up FDR control

This file is sorted by significance value rather than genomic location, the most significant results being at the top.

QUESTIONS?

DEPICT: Why to use DEPICT

1. **Prediction of associated loci**

    Association is just referring to relationship. It does not impose a direction to this relationship (as in causality) nor does it give information whether this relationship can be used to make predictions in new samples about the trait. For the latter one will typically build a relationship model on "training" data and then validate this model on "validation" data. Optimal models are also usually referring to those models with maximal predictive power.
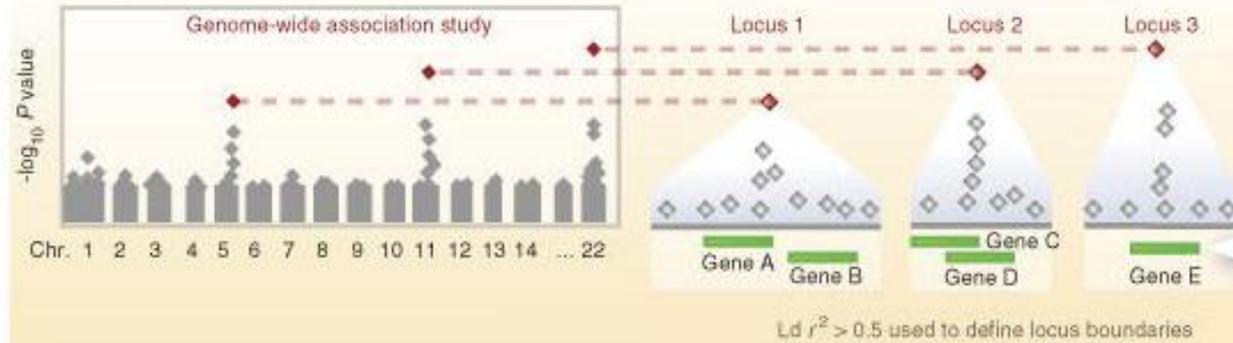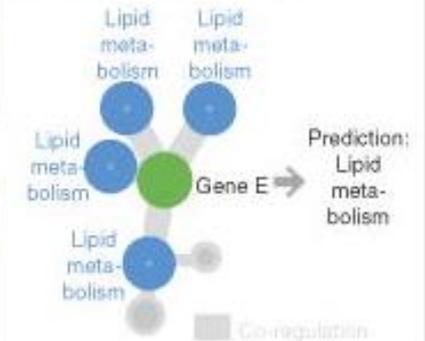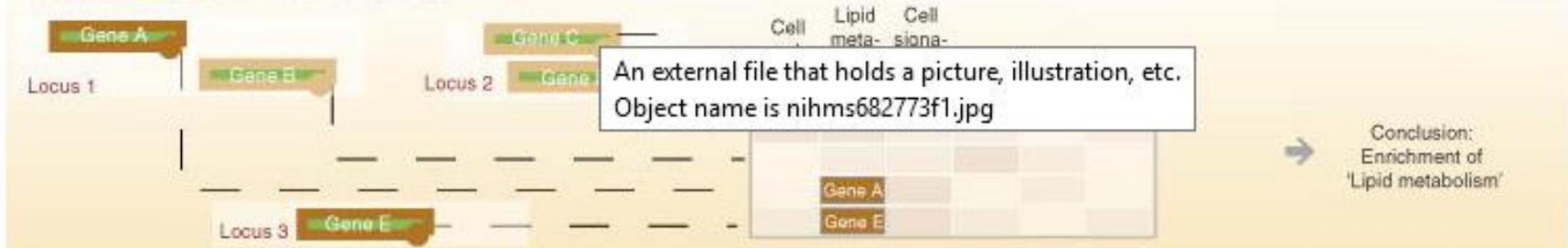
Find over-represented/enriched genes/biological patterns
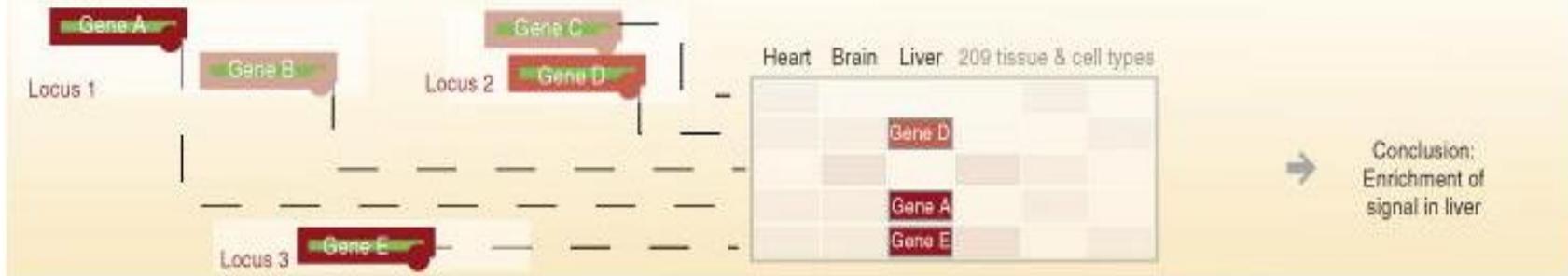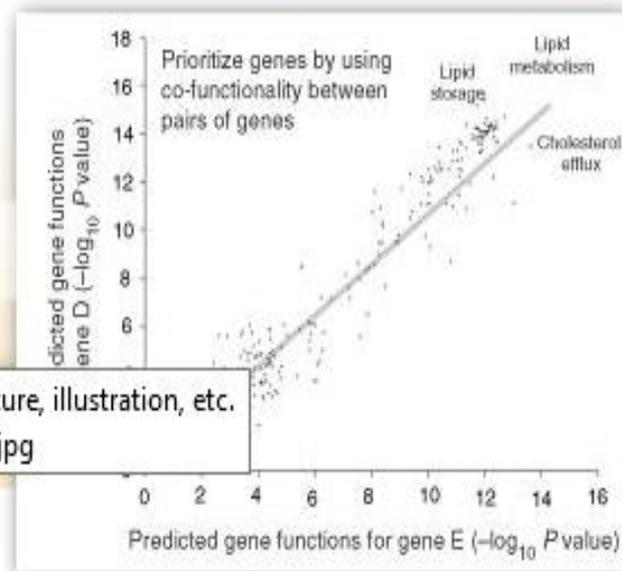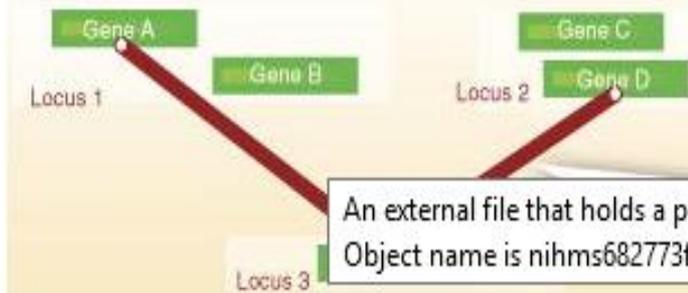
DEPICT : Download

- DEPICT Stands for Data-driven Expression-Prioritized Integration for Complex Traits

- Download DEPICT from following link
  https://data.broadinstitute.org/mpg/depict/depict_140721.tar.bz2

- Uncompress the depict_140721.tar.bz2 using
  tar xvfj depict_140611.tar.bz2

- There are multiple folders with depict_140721.Main file is depict.py with consist of changeable parameters.

- Let us run ./depict.py on window

- **What you can see on the terminal?**

DEPICT: Specific Parameters

▪param_analysis label - Set this to the label you want to appear in the result filenames.

▪path_snpfile - Path to your file with associated SNPs (rsIDs must be used to specifiy SNPs).

▪flag_loci - Construct loci based on your associated SNPs? (Yes, 1; No, 0). This parameter must be set to 1 the first time the analysis is run.

▪flag_genes - Should genes be prioritized? (Yes, 1; No, 0).

▪flag_genesets - Conduct reconstituted gene set enrichment analysis? (Yes, 1; No, 0).

▪flag_tissues - Conduct tissue/cell type enrichment analysis? (Yes, 1; No, 0).

▪param_ncores - Number of CPU cores used by DEPICT.

▪path_locus generator_jar- Path to JAR file used to constructed loci (Should not be changed).

▪path_depict_jar - Path to DEPICT JAR file (Should not be changed).

- The DEPICT locus specifies the nearest genes. Merged loci do contain as many nearest genes as SNPs that were merged. Consequently the number of nearest gene can be different to overall number of (merged) loci.
- Nearest genes are always listed in the locus file, but only included into the DEPICT analysis if there are no genes in the given associated loci.

DEPICT: Let us work on test data

- Open depict.py
- Download snps_list.txt from course website
- Open depict.py and provide snp_list.txt as input in front of variable path_snpfile
- Save the file as depict_custome.py and type ./depict_custome.py

Open result directory and check
- **snp_list.log ,**
- **snps_list_loci.txt ,**
- **_snp_list_genesetenrichment.txt,**
- **_snp_list_geneprioritization.txt**
- **_snp_list_tissueenrichment.txt**

What next ????????

- What does this gene and its protein product do?

- How and where does it do it?

- Does it make sense to see it on this list?

- Does it interact with other genes/proteins?

- Does its behavior change during disease, disorder or therapy

DEPICT: RESULT INTERPRETATION

1. **Open _geneprioritization.txt and count locus crossed FDR threshold**
   14 genes crossed the False discovery rate < 5%

2. **Provide the chromosome number of above selected locus**
   chr 1, chr 2, chr 3, chr 5

3. **Can you provide summarised inference based on the on "Gene bio_type" column description?**

DEPICT: RESULT INTERPRETATION

1. **Open _geneenrichment.txt and count locus crossed**
   FDR threshold

2. **Provide the chromosome number of above selected locus**

   Prepare separate list of Genes ID present in "Ensembl Gene ID " column.

Go to web browser http://agbase.msstate.edu/

Paste Gene names as shown in figure and click "search"

Click on ENSG00000206493 gene and see how many PFAM domains are presents
1?
2?
3?

Advance Functional interpretation

Open browser http://amp.pharm.mssm.edu/Enrichr/ and paste gene list

- Click on pathways
    - Draw the network
- Click on ontologies
    - Cellular , Molecular and Biological functions??

Let us try Unknown dataset

1. Download gene _ist .txt from course website.

2. Predict Gene ontology

3. Predict different pathways

4. Draw Network

5. Any idea, genes playing specific biological functional ??

(Just look into the Network and make inference)

Network connectivity of Identified pathways
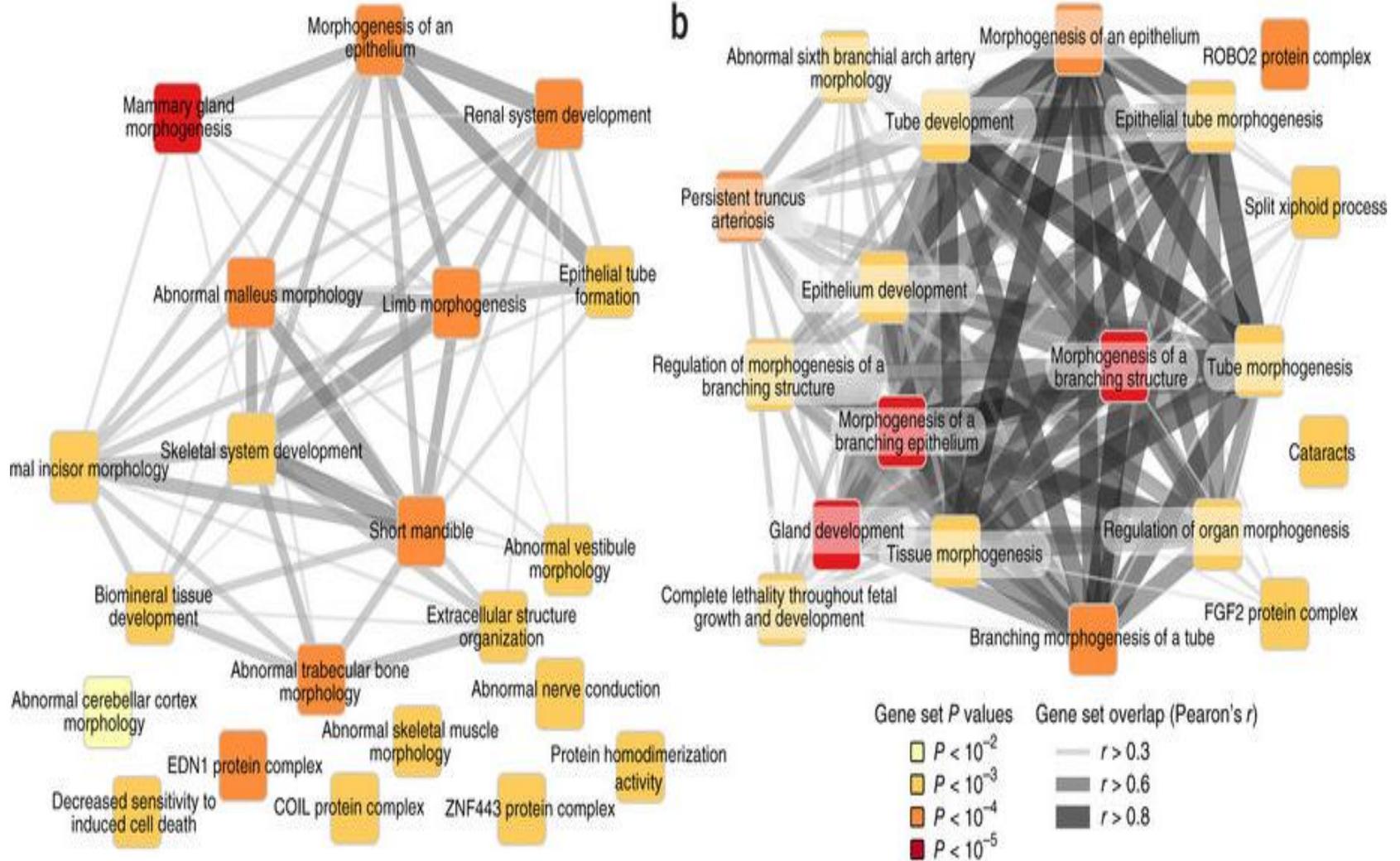
QUESTIONS?

Association Viewer

- Display SNPs and their p-values in a genetic context, similar to usual genome browsers
- Automatically download supplementary information from Ensembl/Biomart
- Display Hapmap LD plots
- Print and export the display to various data formats
- Import external data file such as BED or Wiggle as supplementary tracks
- Users can scroll and browse through their data in real time

Association Viewer: Download

■   Download Associationviewer from :
https://sourceforge.net/projects/associationview/files/latest/download?source=files

■Download the example data from the course website
sample_scores_ncbi35.txt.zip
use_case_1_ncbi36.wig
genes_chromosome21_ncbi35.bed

Association Viewer:Run

▪Double click on the associationviewer-2.0.jar and check different options

▪Add file by clicking option "Add track" as shown in figure below: