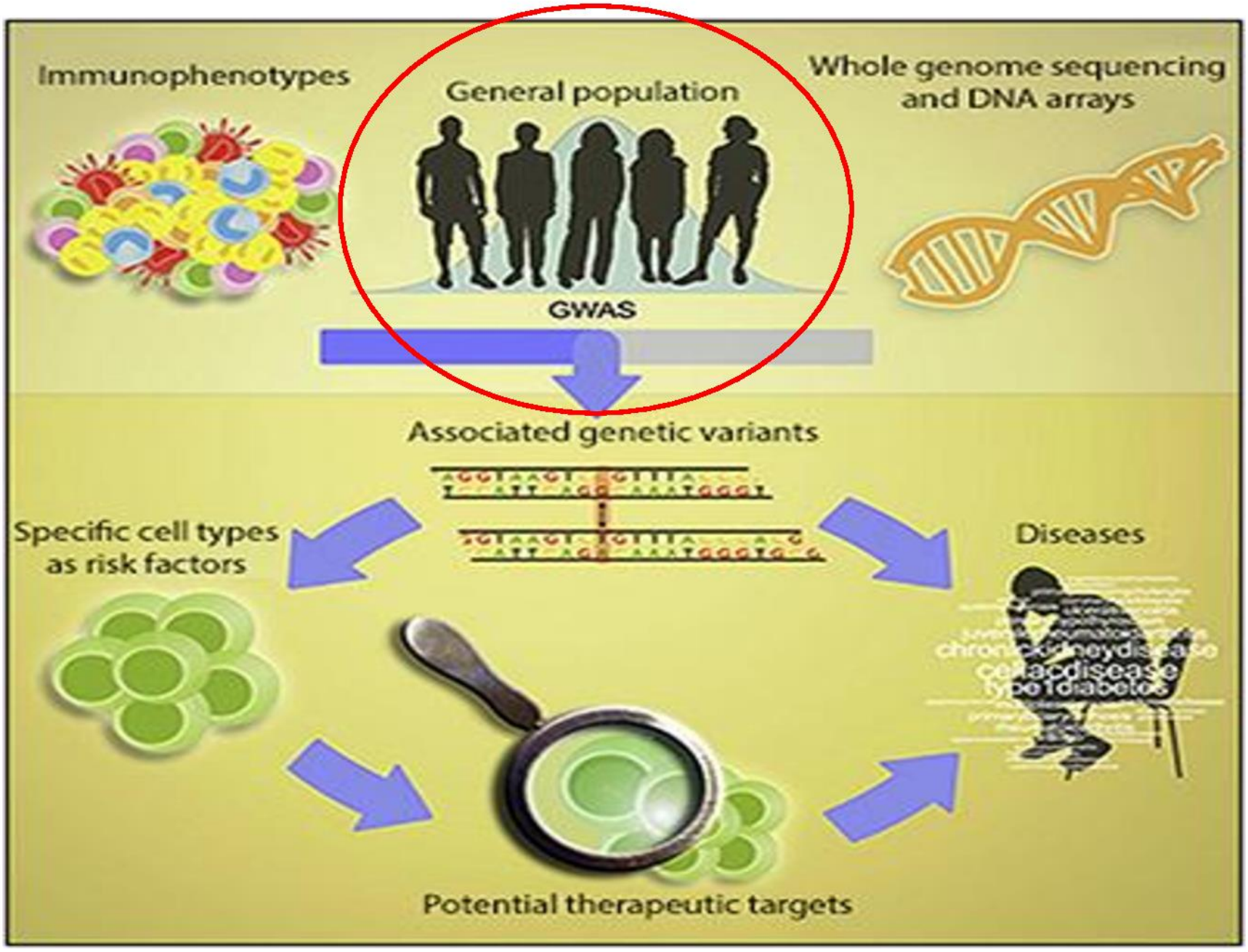
Pre and Post GWAS

Archana Bhardwaj


Immunophenotypes
General population
Whole genome sequencing and DNA arrays
GWAS
Associated genetic variants
Specific cell types as risk factors
Diseases
Potential therapeutic targets

Genetic Mapping

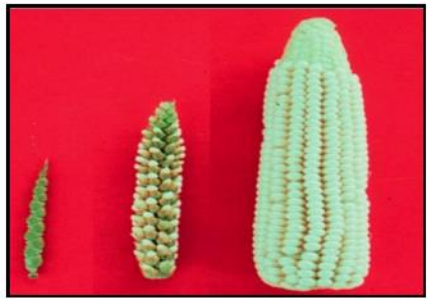
- As the name suggests, it refers to locating genes or genetic information on a genetic map (possibly subdivided by chromosomal regions)

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

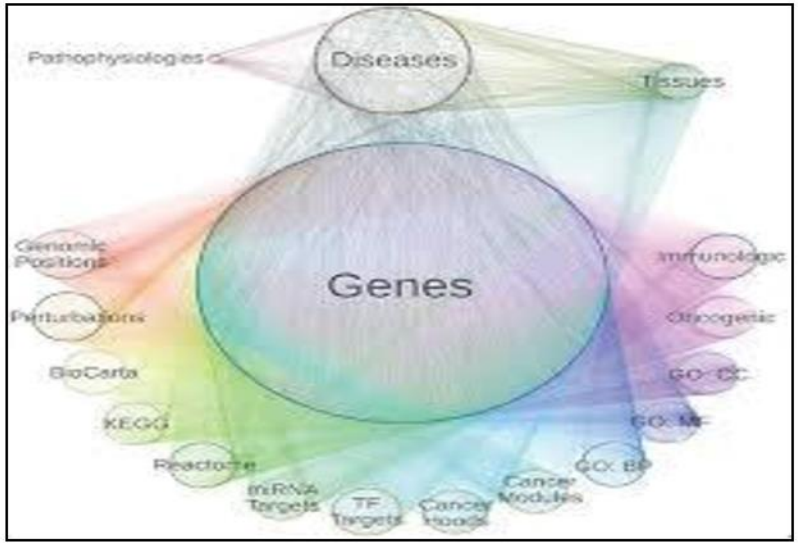
- Scientists isolate DNA and use genetic markers to find genomic locations that can be linked / associated with a trait of interest (e.g., in humans: disease trait, in plants: yield);

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.

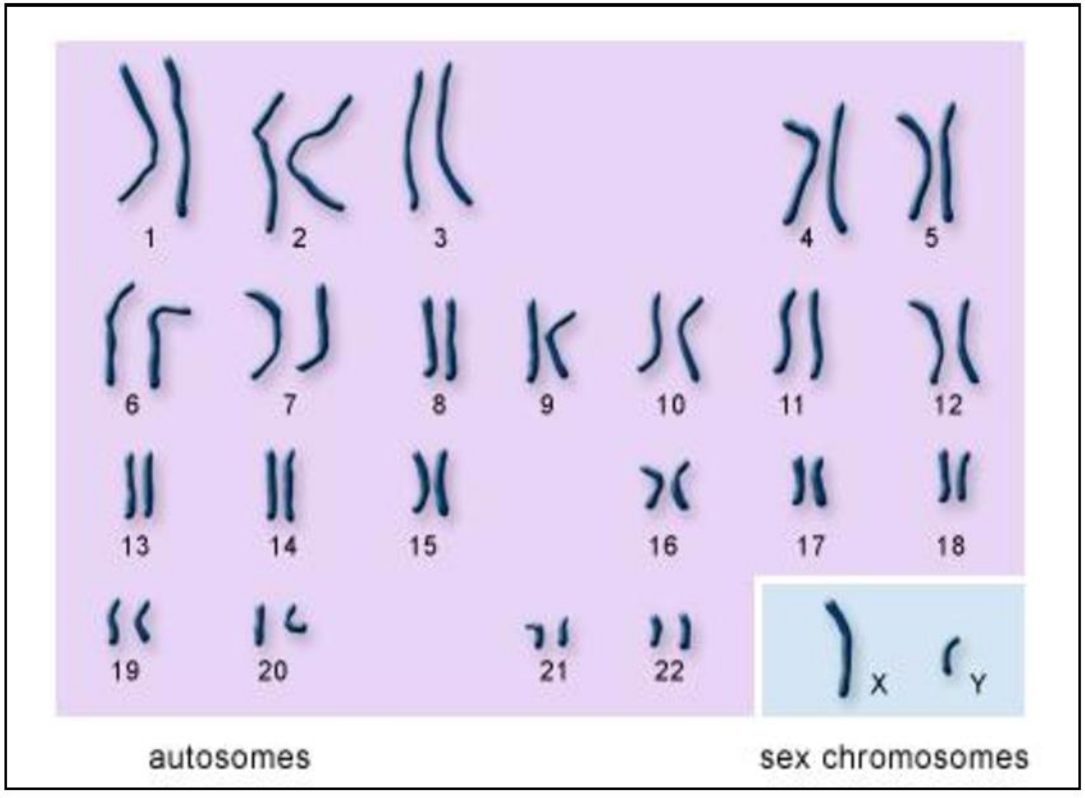
These features were consistent with, but did not prove, a polygenic model. The idea that commonly varying traits might be polygenic in nature was offered by East in 1910 (*18*). By 1920, linkage mapping was used to identify multiple unlinked factors influencing truncate wings in *Drosophila* (*19*), 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

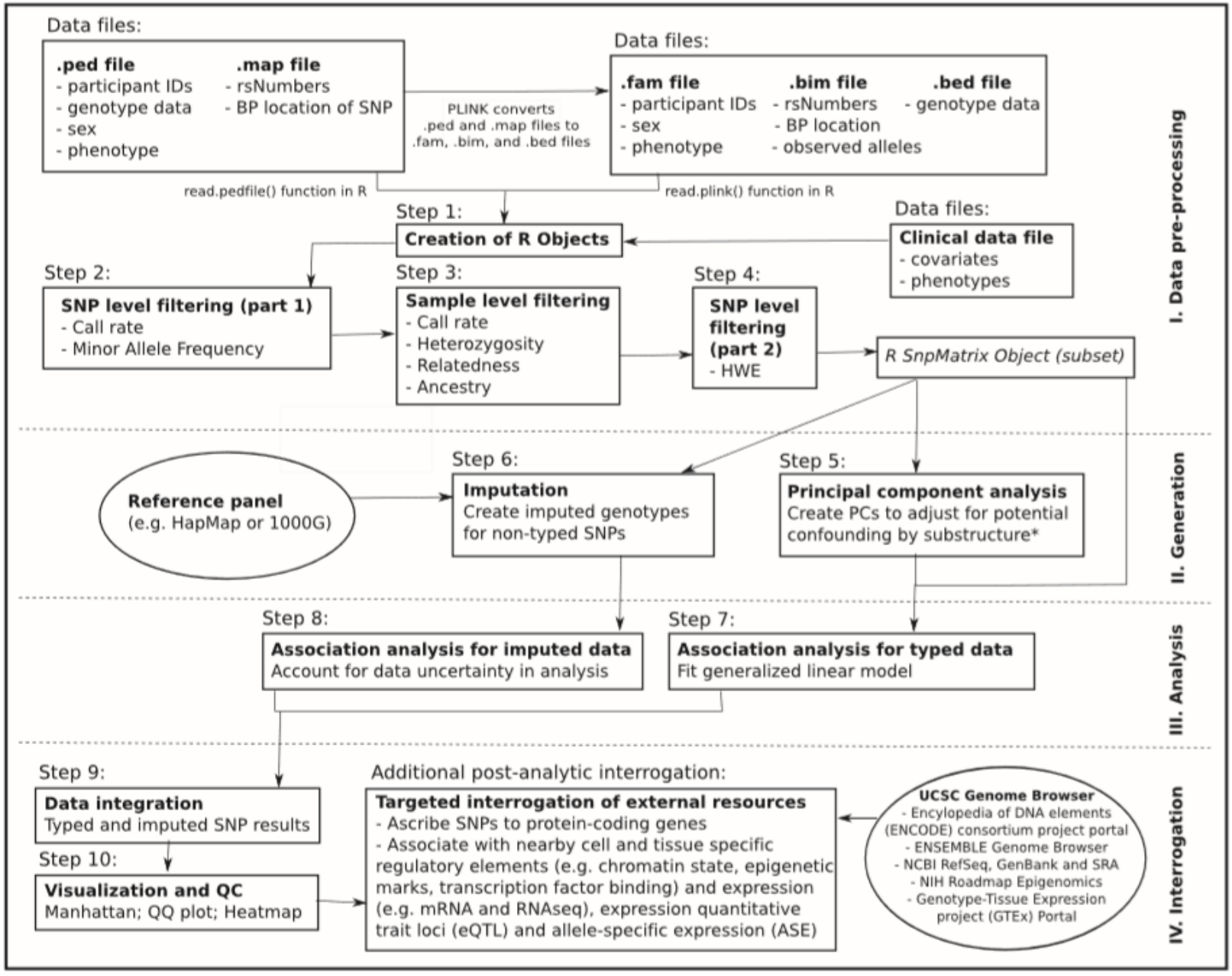


Tools for GWAS

PLINK

DEPICT

Association Viewer


Data files:
.ped file
- participant IDs
- genotype data
- sex
- phenotype
.map file
- rsNumbers
- BP location of SNP
PLINK converts .ped and .map files to .fam, .bim, and .bed files
Data files:
.fam file
- participant IDs
- sex
- phenotype
.bim file
- rsNumbers
- BP location
- observed alleles
.bed file
- genotype data
read.pedfile() function in R
read.plink() function in R
Step 1:
Creation of R Objects
Data files:
Clinical data file
- covariates
- phenotypes
Step 2:
SNP level filtering (part 1)
- Call rate
- Minor Allele Frequency
Step 3:
Sample level filtering
- Call rate
- Heterozygosity
- Relatedness
- Ancestry
Step 4:
SNP level filtering (part 2)
- HWE
R SnpMatrix Object (subset)
I. Data pre-processing
Reference panel
(e.g. HapMap or 1000G)
Step 6:
Imputation
Create imputed genotypes for non-typed SNPs
Step 5:
Principal component analysis
Create PCs to adjust for potential confounding by substructure*
II. Generation
Step 8:
Association analysis for imputed data
Account for data uncertainty in analysis
Step 7:
Association analysis for typed data
Fit generalized linear model
III. Analysis
Step 9:
Data integration
Typed and imputed SNP results
Step 10:
Visualization and QC
Manhattan; QQ plot; Heatmap
Additional post-analytic interrogation:
Targeted interrogation of external resources
- Ascribe SNPs to protein-coding genes
- Associate with nearby cell and tissue specific regulatory elements (e.g. chromatin state, epigenetic marks, transcription factor binding) and expression (e.g. mRNA and RNAseq), expression quantitative trait loci (eQTL) and allele-specific expression (ASE)
UCSC Genome Browser
- Encylopedia of DNA elements (ENCODE) consortium project portal
- ENSEMBLE Genome Browser
- NCBI RefSeq, GenBank and SRA
- NIH Roadmap Epigenomics
- Genotype-Tissue Expression project (GTEx) Portal
IV. Interrogation

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

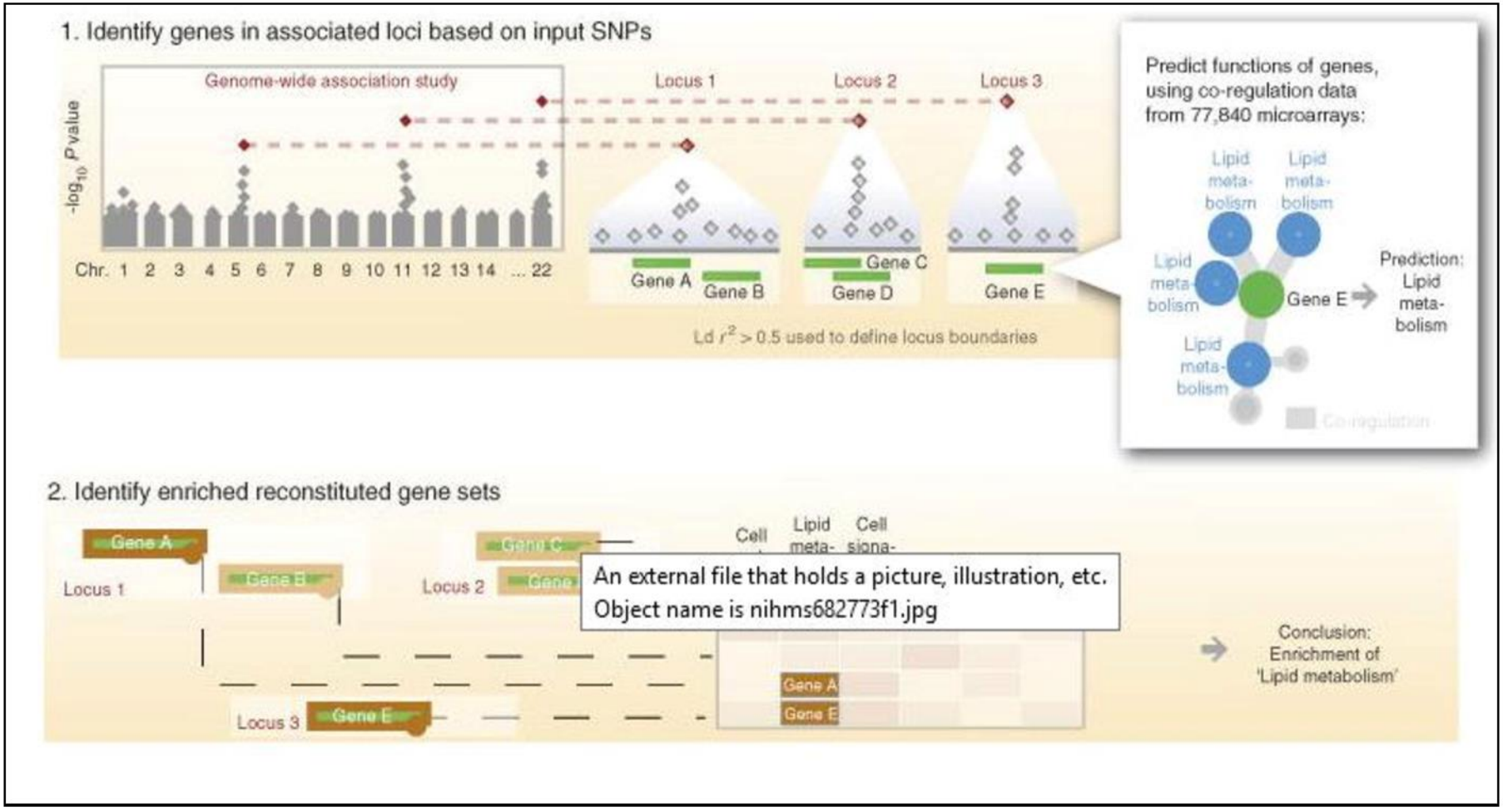
d) shift the window 5 SNPs forward and repeat the procedure.

QUESTIONS?

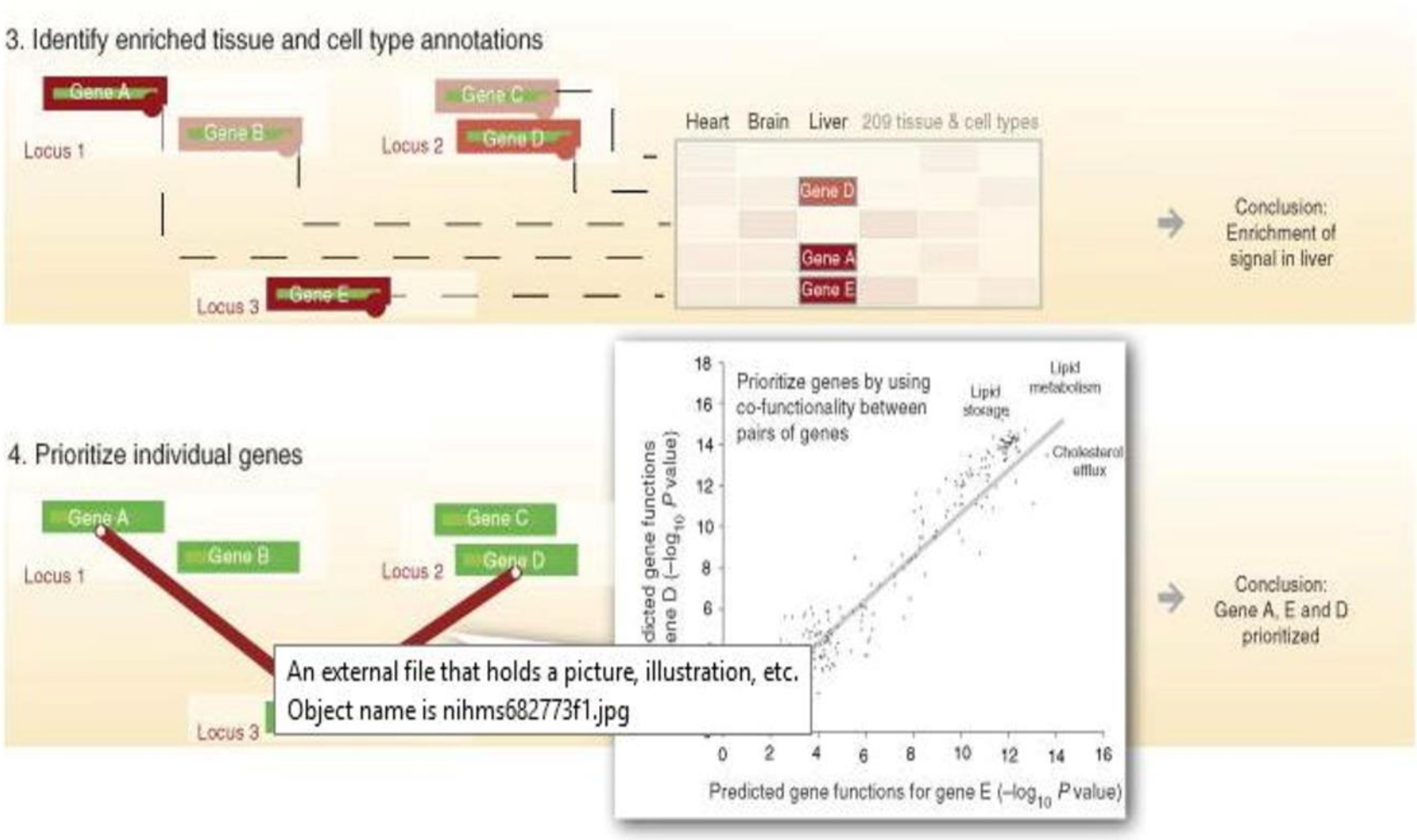
DEPICT: Why to use DEPICT

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


3. Identify enriched tissue and cell type annotations
Gene A
Gene B
Locus 1
Gene C
Gene D
Locus 2
Gene E
Locus 3
Heart
Brain
Liver
209 tissue & cell types
Gene D
Gene A
Gene E
Conclusion:
Enrichment of
signal in liver
4. Prioritize individual genes
Gene A
Gene B
Locus 1
Gene C
Gene D
Locus 2
Locus 3
Prioritize genes by using
co-functionality between
pairs of genes
Lipid
metabolism
Lipid
storage
Cholesterol
efflux
Predicted gene functions for gene E (−log10 P value)
Conclusion:
Gene A, E and D
prioritized
An external file that holds a picture, illustration, etc.
Object name is nihms682773f1.jpg

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

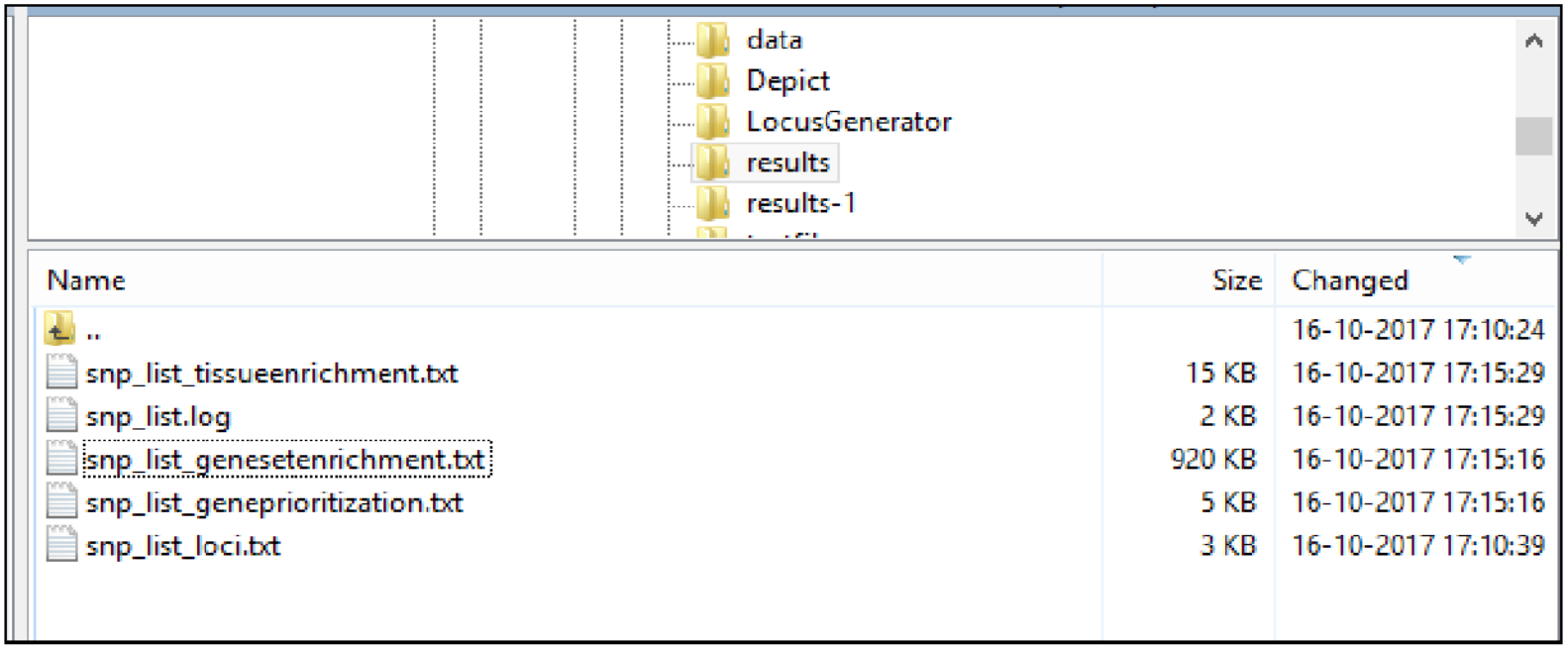
- param_analysis label - Set this to the label you want to appear in the result filenames.
- path_snpfile - Path to your file with associated SNPs (rsIDs must be used to specifiy SNPs).
- flag_loci - Construct loci based on your associated SNPs? (Yes, 1; No, 0). This parameter must be set to 1 the first 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?

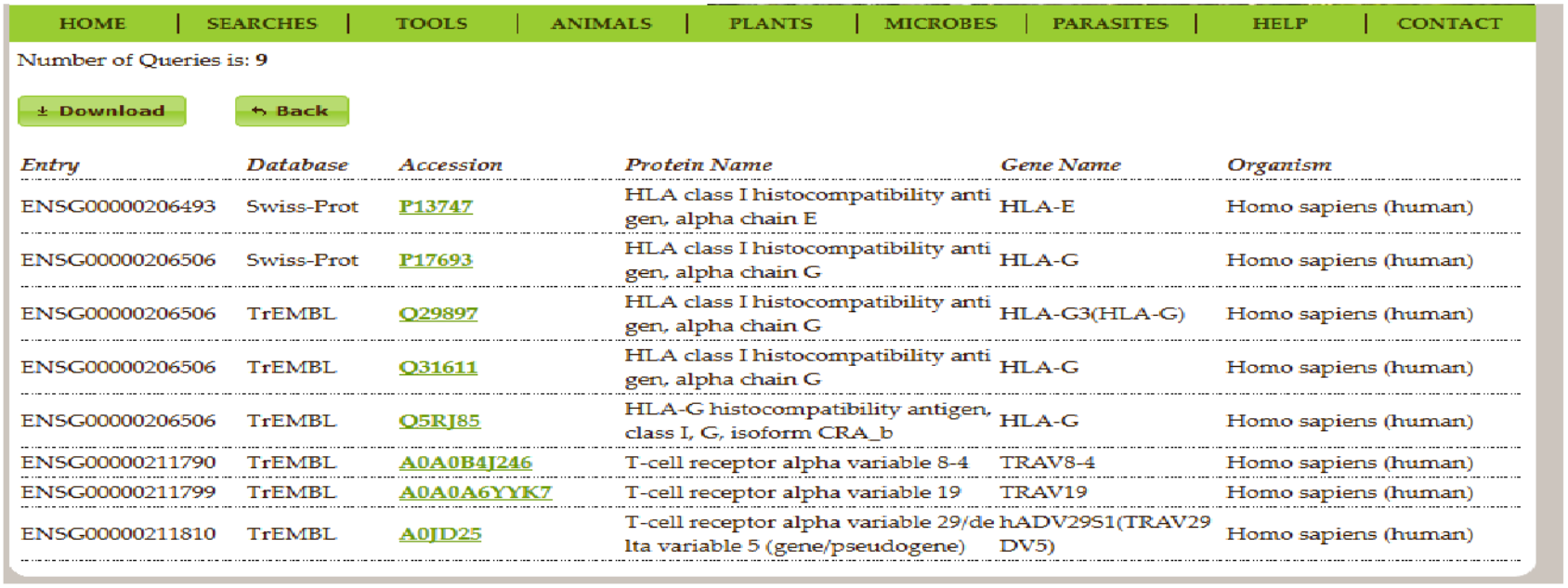
DEPICT: RESULT INTERPRETATION

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”



HOME | SEARCHES | TOOLS | ANIMALS | PLANTS | MICROBES | PARASITES | HELP | CONTACT

Number of Queries is: 9

Download Back

Entry	*Database*	*Accession*	*Protein Name*	*Gene Name*	*Organism*
ENSG00000206493	Swiss-Prot	P13747	HLA class I histocompatibility antigen, alpha chain E	HLA-E	Homo sapiens (human)
ENSG00000206506	Swiss-Prot	P17693	HLA class I histocompatibility antigen, alpha chain G	HLA-G	Homo sapiens (human)
ENSG00000206506	TrEMBL	Q29897	HLA class I histocompatibility antigen, alpha chain G	HLA-G3(HLA-G)	Homo sapiens (human)
ENSG00000206506	TrEMBL	Q31611	HLA class I histocompatibility antigen, alpha chain G	HLA-G	Homo sapiens (human)
ENSG00000206506	TrEMBL	Q5RJ85	HLA-G histocompatibility antigen, class I, G, isoform CRA_b	HLA-G	Homo sapiens (human)
ENSG00000211790	TrEMBL	A0A0B4J246	T-cell receptor alpha variable 8-4	TRAV8-4	Homo sapiens (human)
ENSG00000211799	TrEMBL	A0A0A6YYK7	T-cell receptor alpha variable 19	TRAV19	Homo sapiens (human)
ENSG00000211810	TrEMBL	A0JD25	T-cell receptor alpha variable 29/delta variable 5 (gene/pseudogene)	hADV29S1(TRAV29DV5)	Homo sapiens (human)

Click on ENSG00000206493 gene and see how many PFAM domains are presents

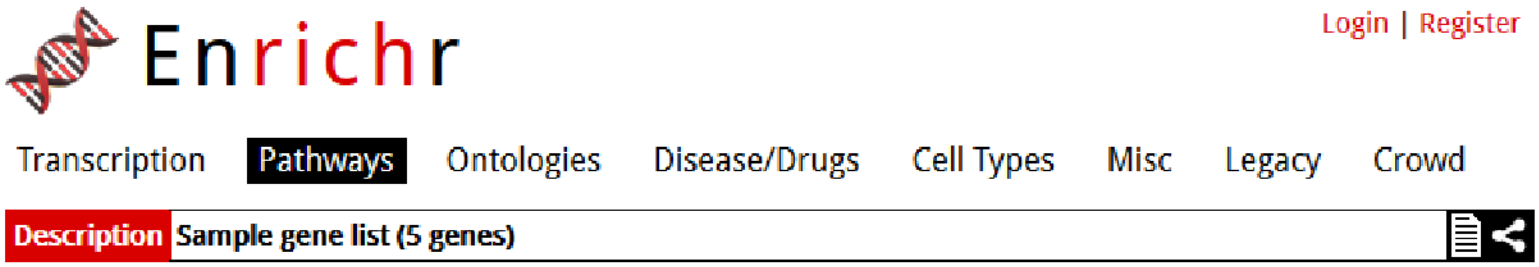
1?

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

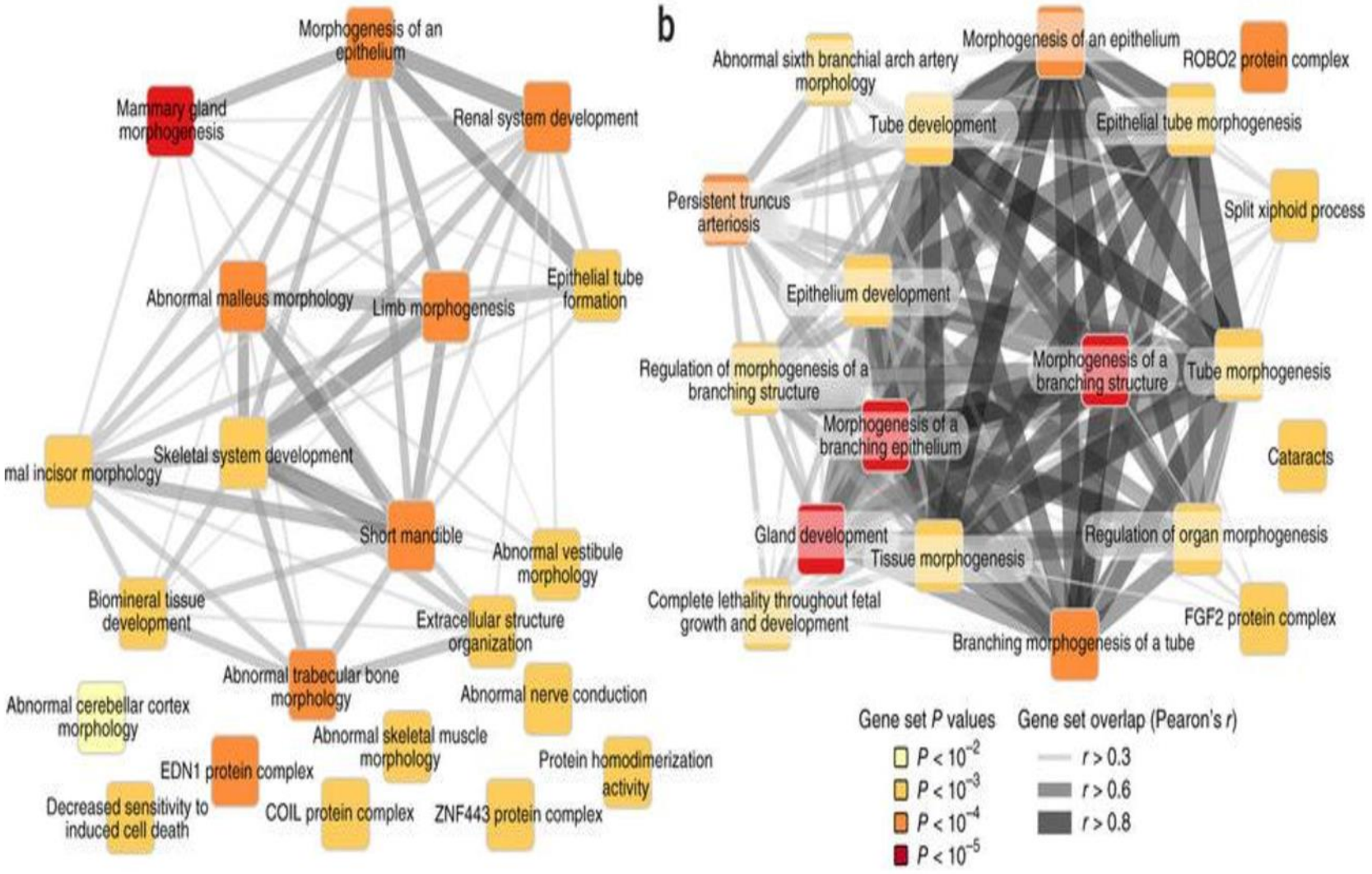
3. Predict different pathways

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

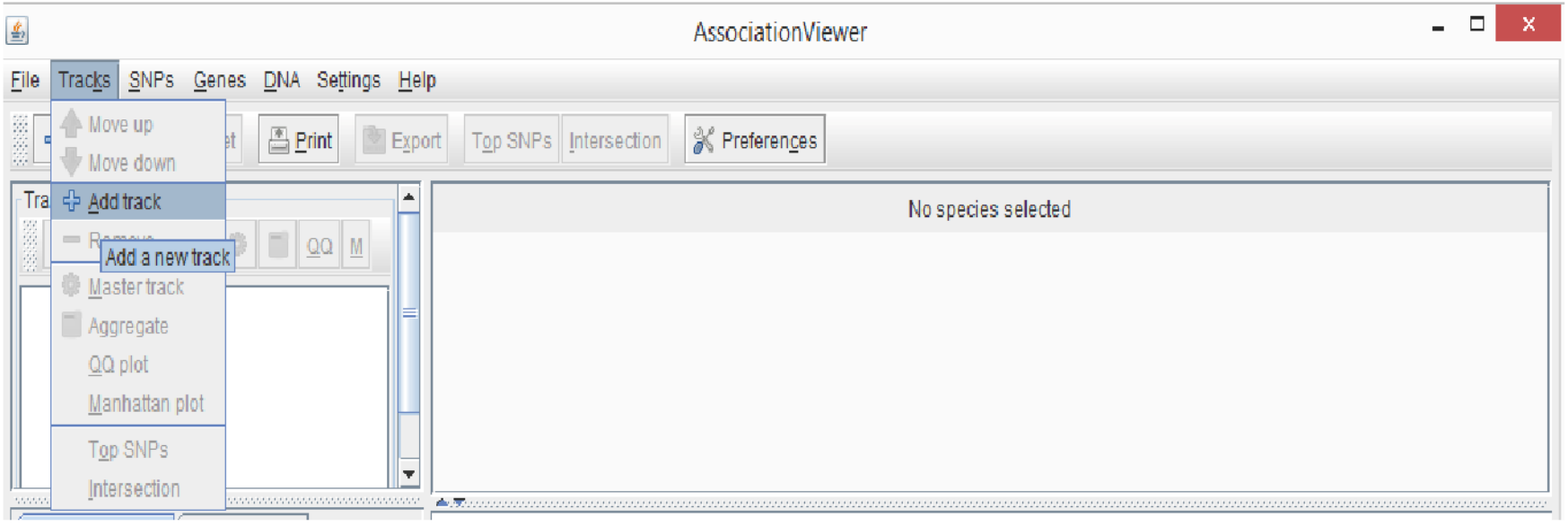
- Download Associationviewer from : https://sourceforge.net/projects/associationview/files/latest/download?source=files

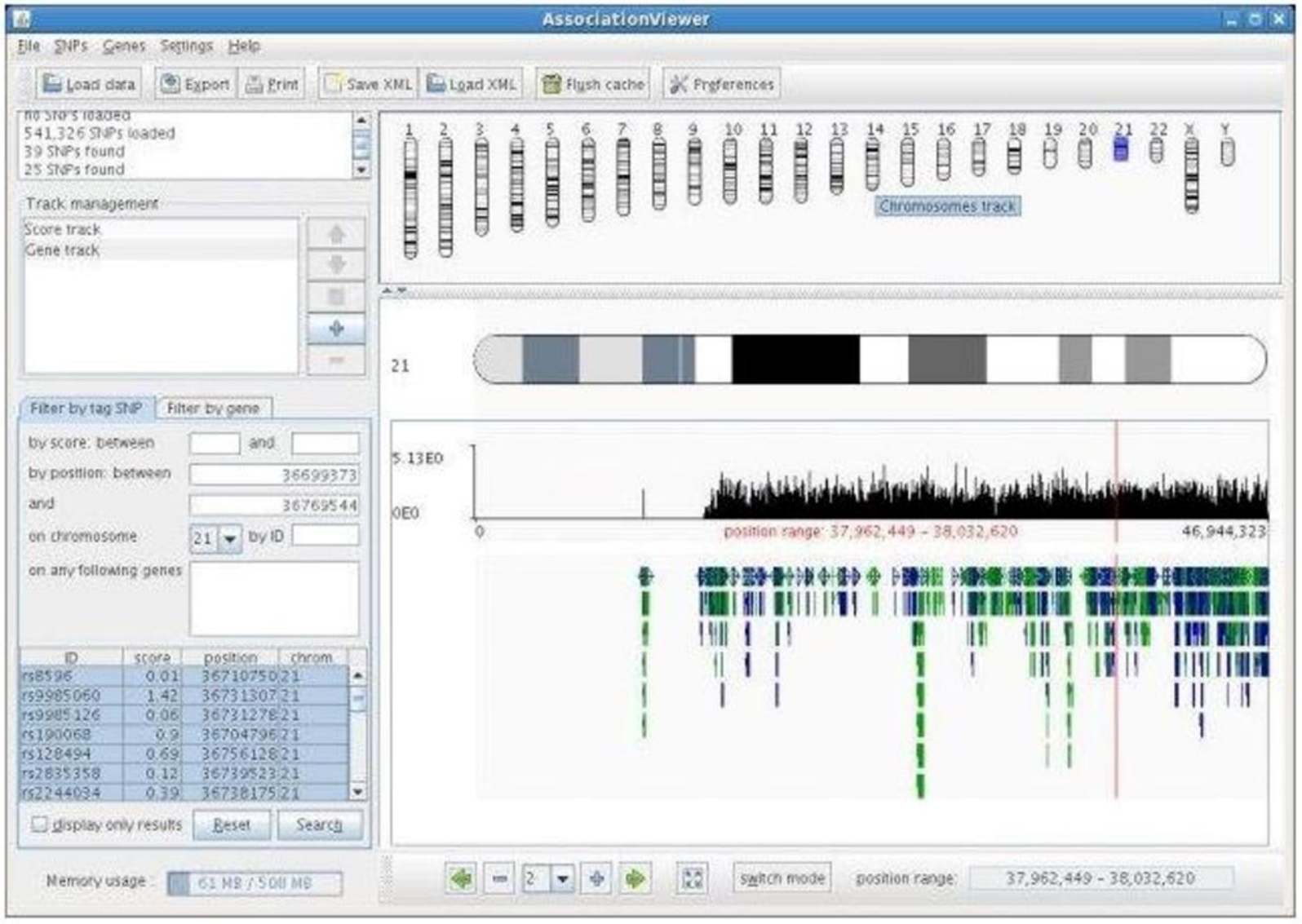
- Download the example data from the course website
  sample_scores_ncbi35.txt.zip
  use_case_1_ncbi36.wig
  genes_chromosome21_ncbi35.bed

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:




AssociationViewer
File SNPs Genes Settings Help
Load data
Export
Print
Save XML
Load XML
Flush cache
Preferences
541,326 SNPs loaded
39 SNPs found
25 SNPs found
Track management
Score track
Gene track
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
Chromosomes track
21
Filter by tag SNP
Filter by gene
by score: between
and
by position: between
36699373
and
36769544
on chromosome
21
by ID
on any following genes
ID score position chrom
rs8596 0.01 36710750 21
rs9985060 1.42 36731307 21
rs9985126 0.06 36731278 21
rs190068 0.9 36704796 21
rs128494 0.69 36756128 21
rs2835358 0.12 36739523 21
rs2244034 0.39 36738175 21
display only results
Reset
Search
5.13E0
0E0
0
position range: 37,962,449 - 38,032,620
46,944,323
Memory usage:
61 MB / 500 MB
2
switch mode
position range:
37,962,449 - 38,032,620