

# Gene-set Analyses of GWAS data

---

PEILIN JIA

BEIJING INSTITUTE OF GENOMICS



# Outline

---

## Gene-based p-values

- Rationale
- Methods

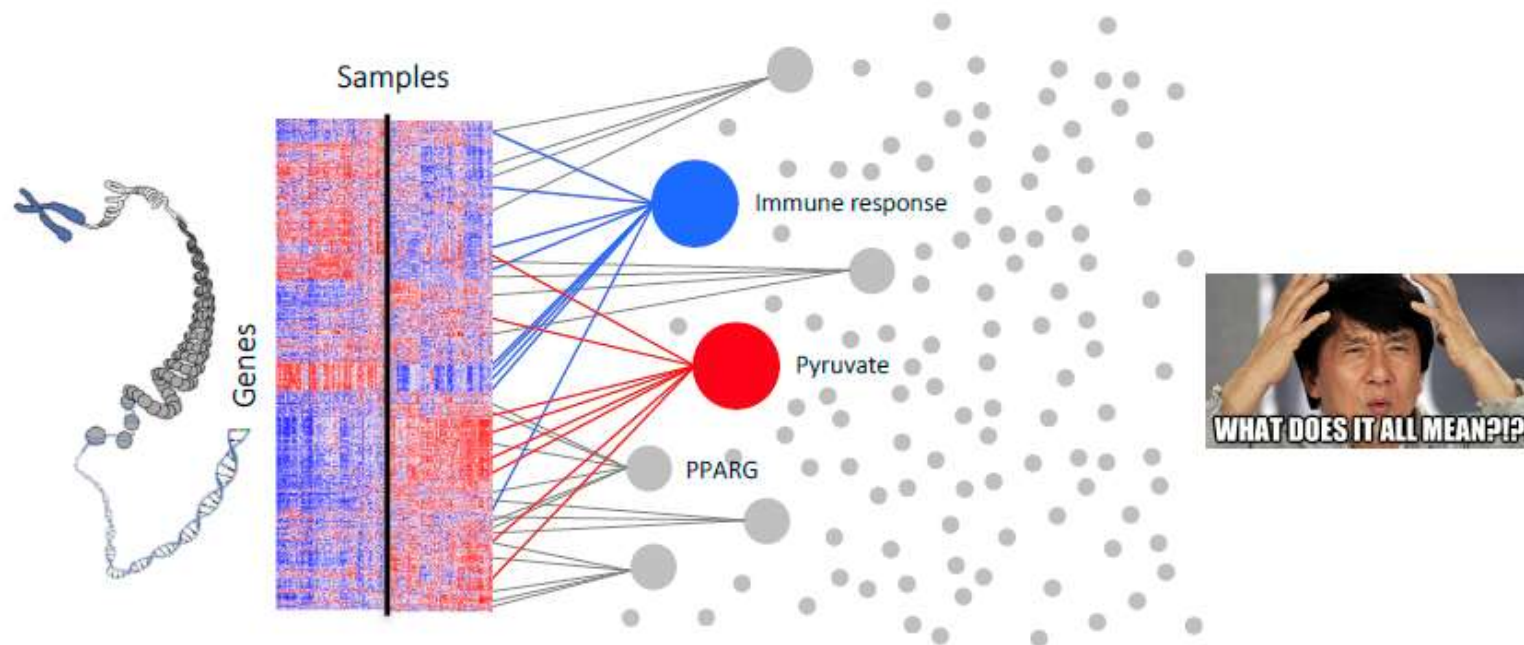
## Set-based analysis of GWAS data

- Pathway
- Network

## Tissue and cell type specific enrichment analysis of GWAS data

# What is gene-set analysis (GSA)?

---



# Rationale to conduct set-based analyses

---

The polygenic nature of most complex phenotypes

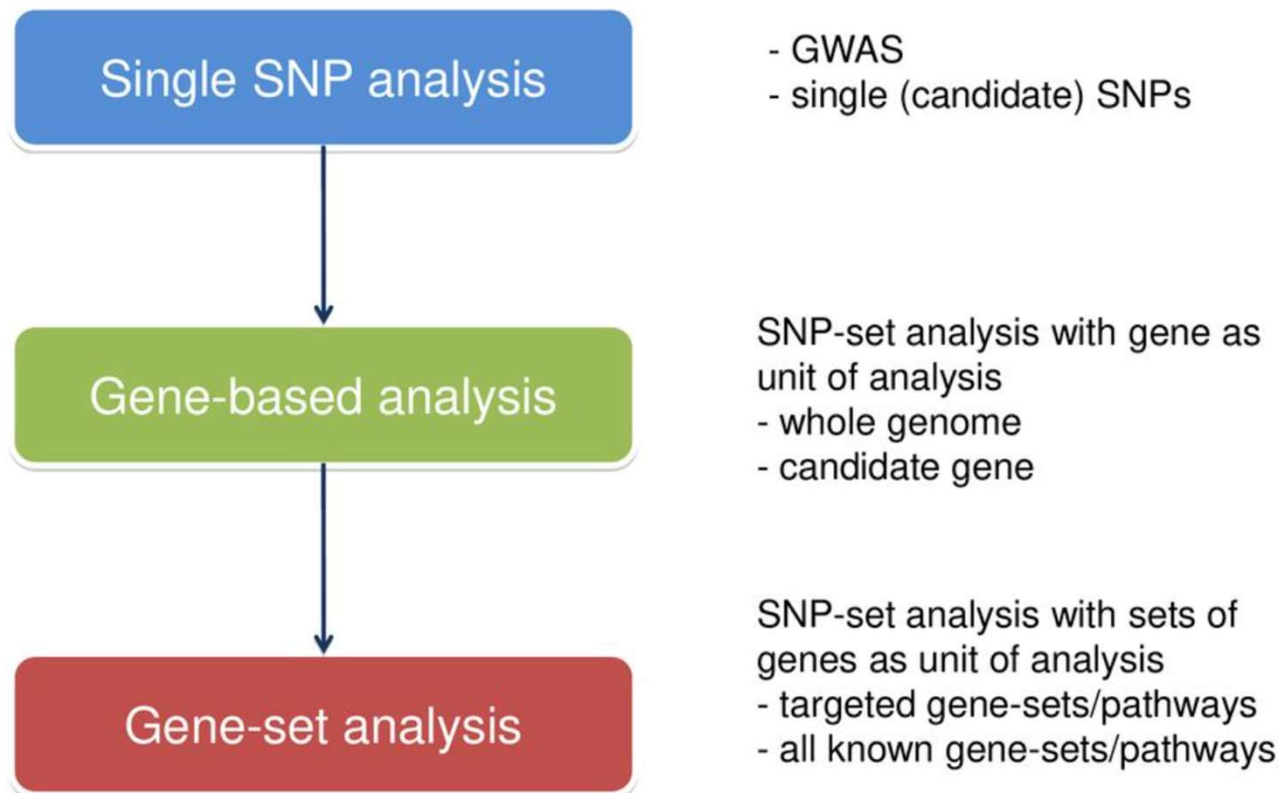
Focus on the combined effects of many loci, each making a small contribution to overall disease susceptibility

The reported significant SNPs at  $5e-8$  explained limited heritability and phenotype variation

~85 – 90% of GWAS reported significant SNPs are located in non-coding regions and have low to moderate effect sizes

# Testing for functional clustering of SNP associations

---



# Aggregation at different levels

---

Gene-based

Region-based

Pathway- or gene-set-based analysis: aims to provide insight into the biological processes involved by aggregating the association signal observed for a collection of SNPs into a pathway level signal.

This is generally carried out in two steps: first, individual SNPs are mapped to genes and their association p-values are combined into gene scores; second, genes are grouped into pathways and their gene scores are combined into pathway scores.

Existing tools vary in the methods used for each step and the strategies employed to correct for correlation due to linkage disequilibrium.

- Pre-defined, canonical pathways
- Define gene sets from scratch or by guidance of genetic signals

# Why gene-set analysis (GSA)?

---

Interpretation of genome-wide results

Gene-sets are (typically) fewer than all the genes and have more descriptive names

Difficult to manage a long list of significant genes

Detect patterns that would be difficult to discern simply by manually going through e.g. the list of differentially expressed genes

Integrates external information into the analysis

Less prone to false-positives on the gene-level

Top genes might not be the interesting ones, several coordinated smaller changes

# Outline

---

## Gene-based p-values

- Rationale
- Methods

## Set-based analysis of GWAS data

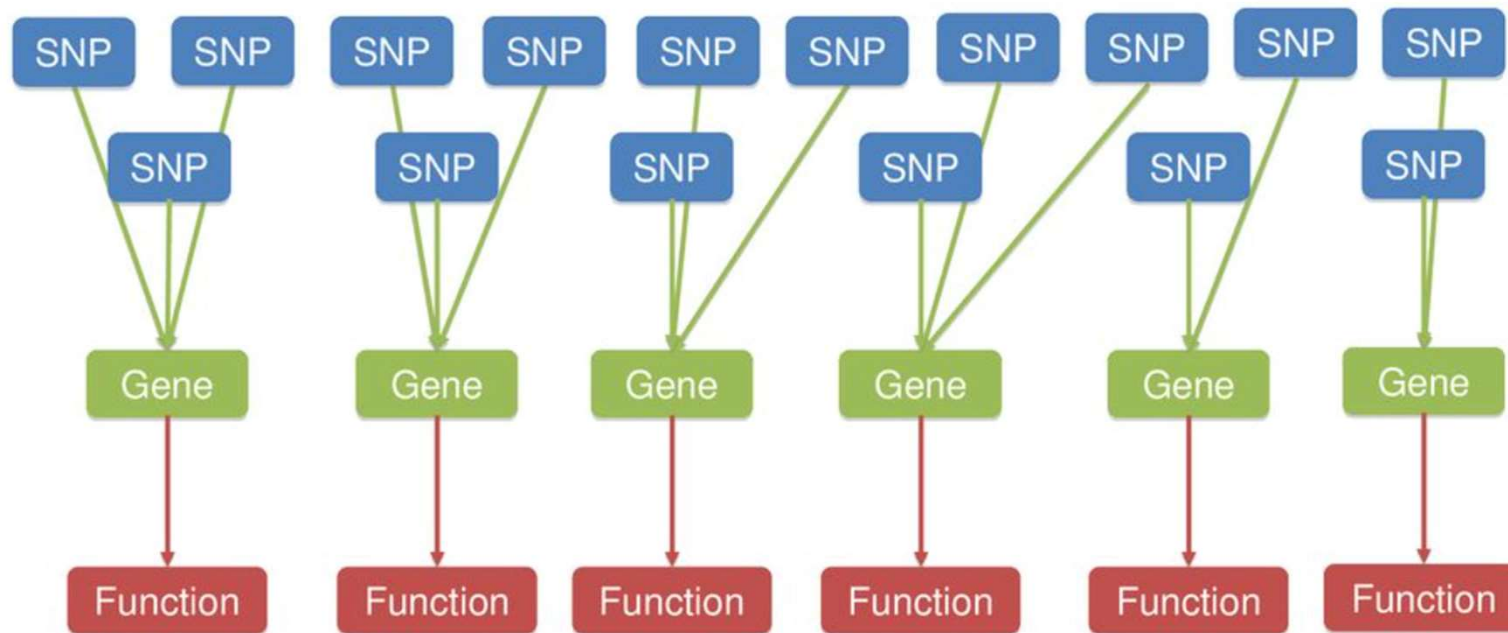
- Pathway
- Network

## Tissue and cell type specific enrichment analysis of GWAS data

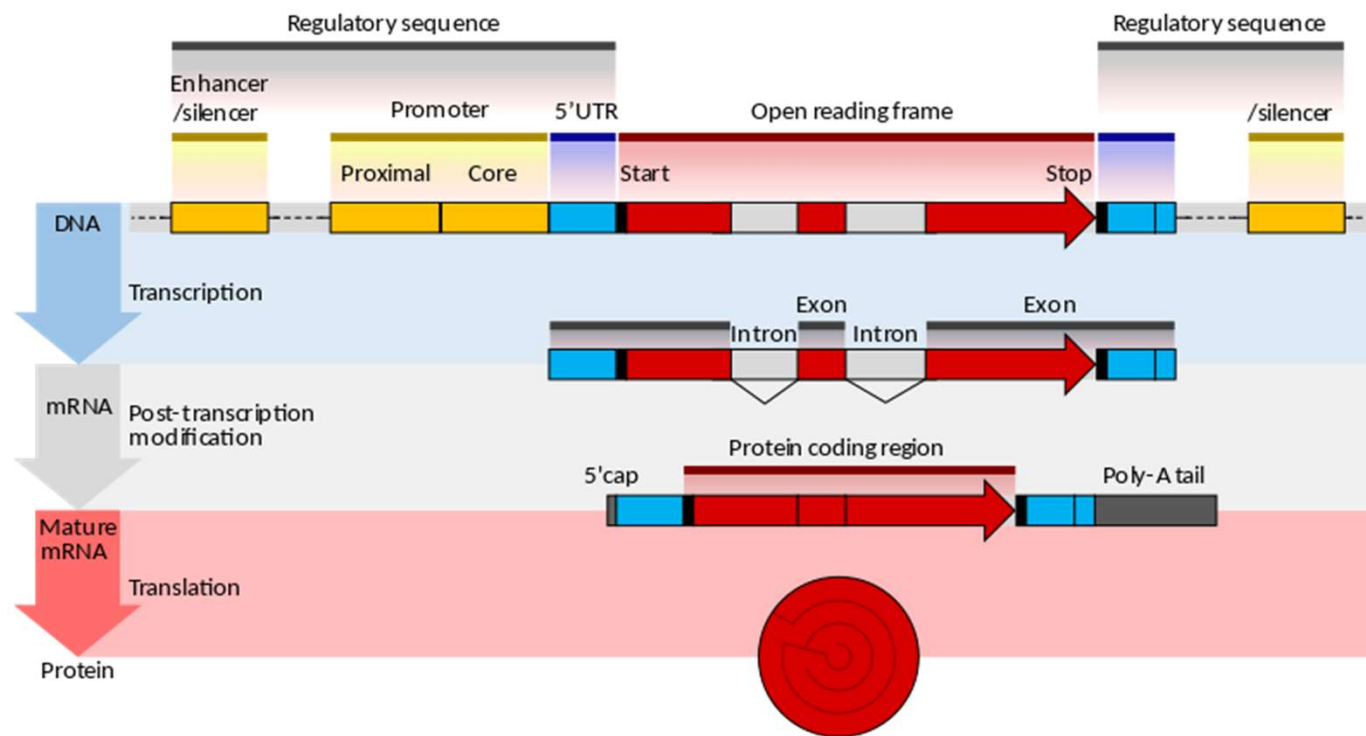


# Map SNPs to genes

---



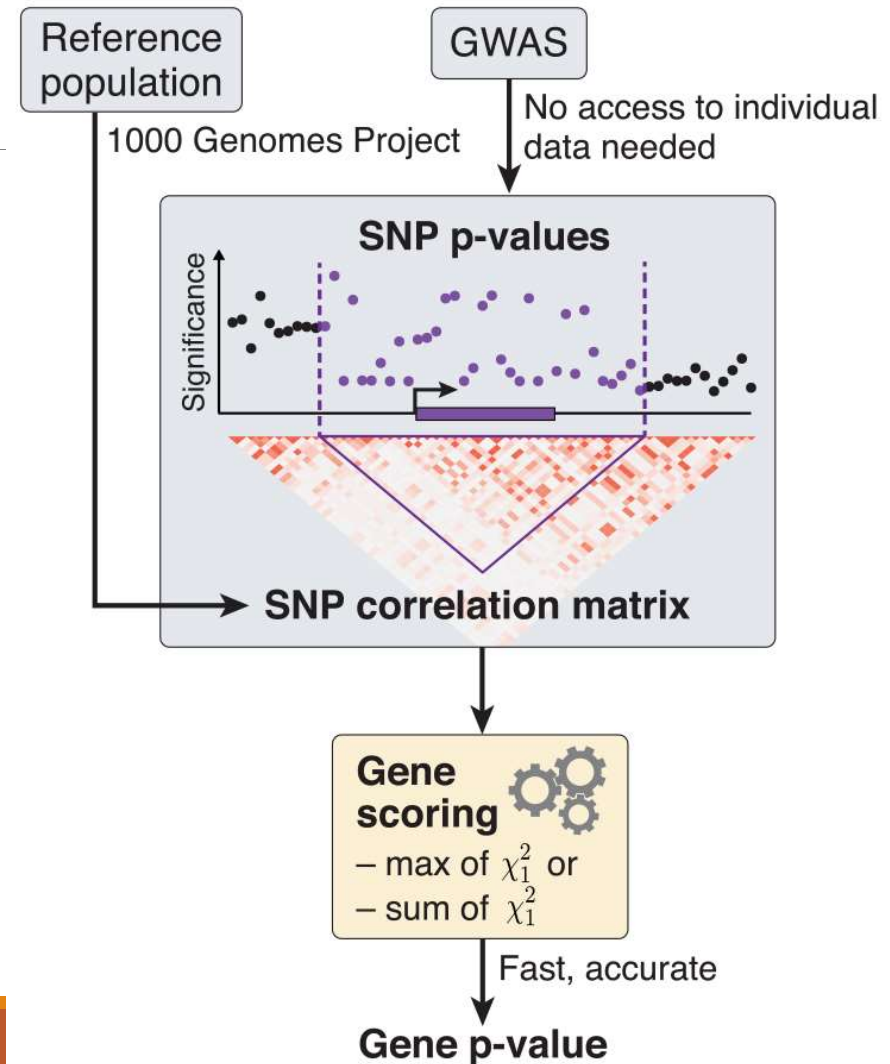
# Structure of a gene



# Gene-based analysis

Instead of testing single SNPs and annotating GWAS-significant ones to genes, gene-based analyses test for the joint association effect of all SNPs in a gene, taking into account LD (correlation between SNPs)

No single SNP needs to reach genome-wide significance, yet if multiple SNPs in the same gene have a lower p-value than expected under the null, the gene-based test can result in low p-values.

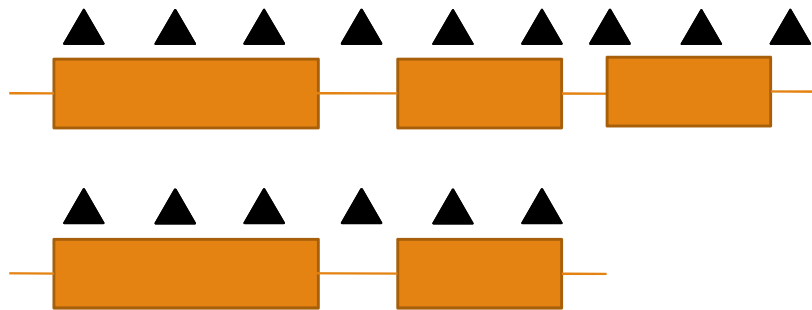


# Different statistical algorithms test different alternative hypotheses

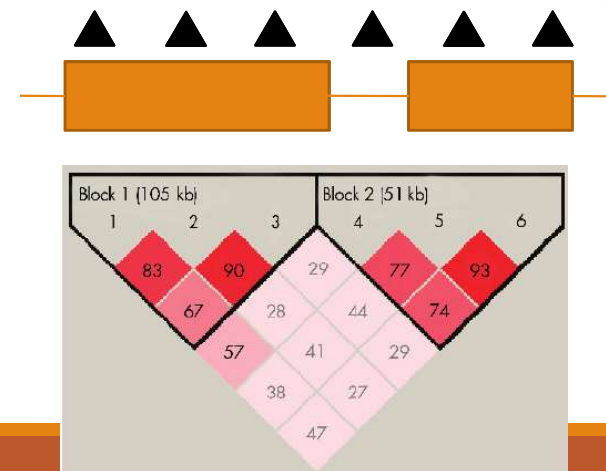
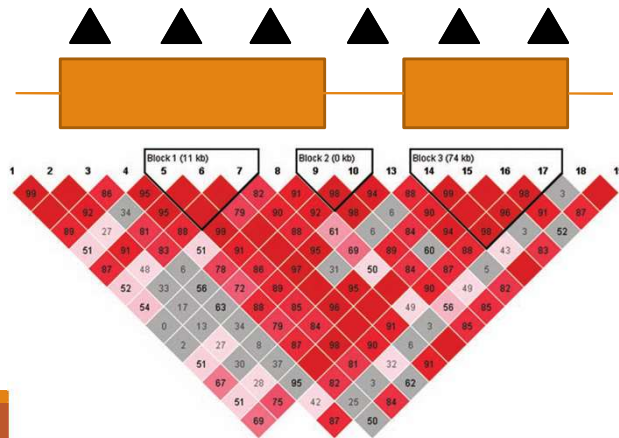
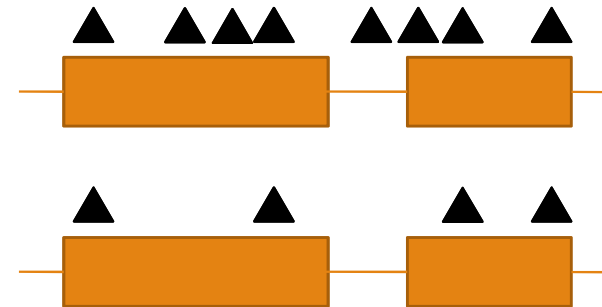
Strategy	Alternative hypothesis
Minimal P-value	At least one SNP in the gene or gene-set is associated with the trait
Combined P-value	The combined pattern of individual P-values provides evidence for association with the trait

# Factors impacting gene-based p-values

Gene length



Genotyping density



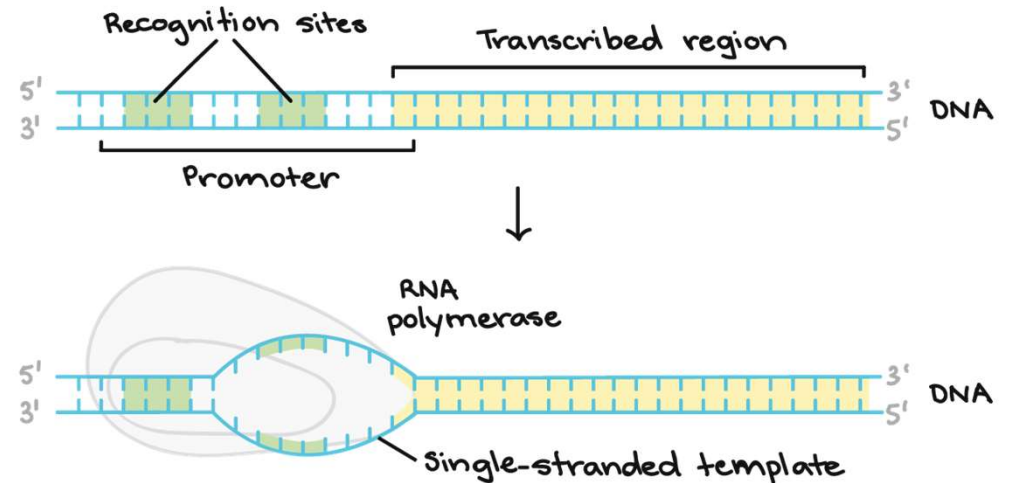
# Gene region definition

**Transcription** is the process in which a gene's DNA sequence is copied (transcribed) to make an RNA molecule.

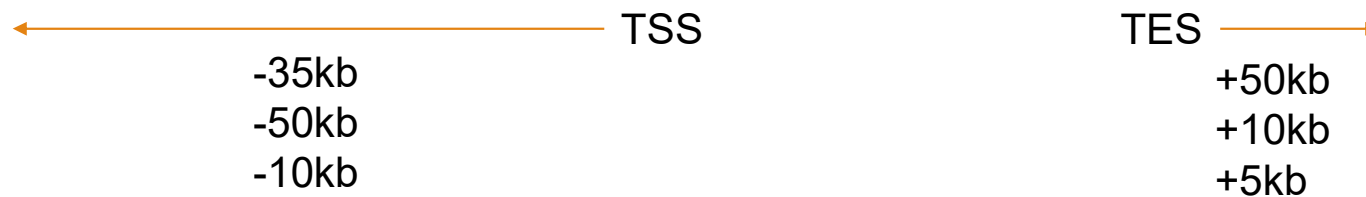
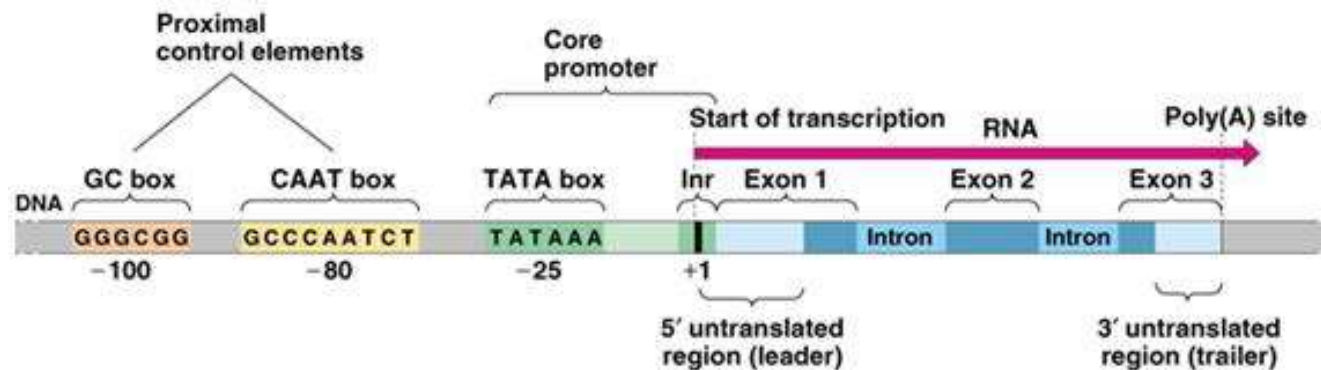
**RNA polymerase** is the main transcription enzyme.

Transcription begins when RNA polymerase binds to a **promoter** sequence near the beginning of a gene (directly or through helper proteins).

Transcription ends in a process called **termination**. Termination depends on sequences in the RNA, which signal that the transcript is finished.



# Gene region definition



# Available methods

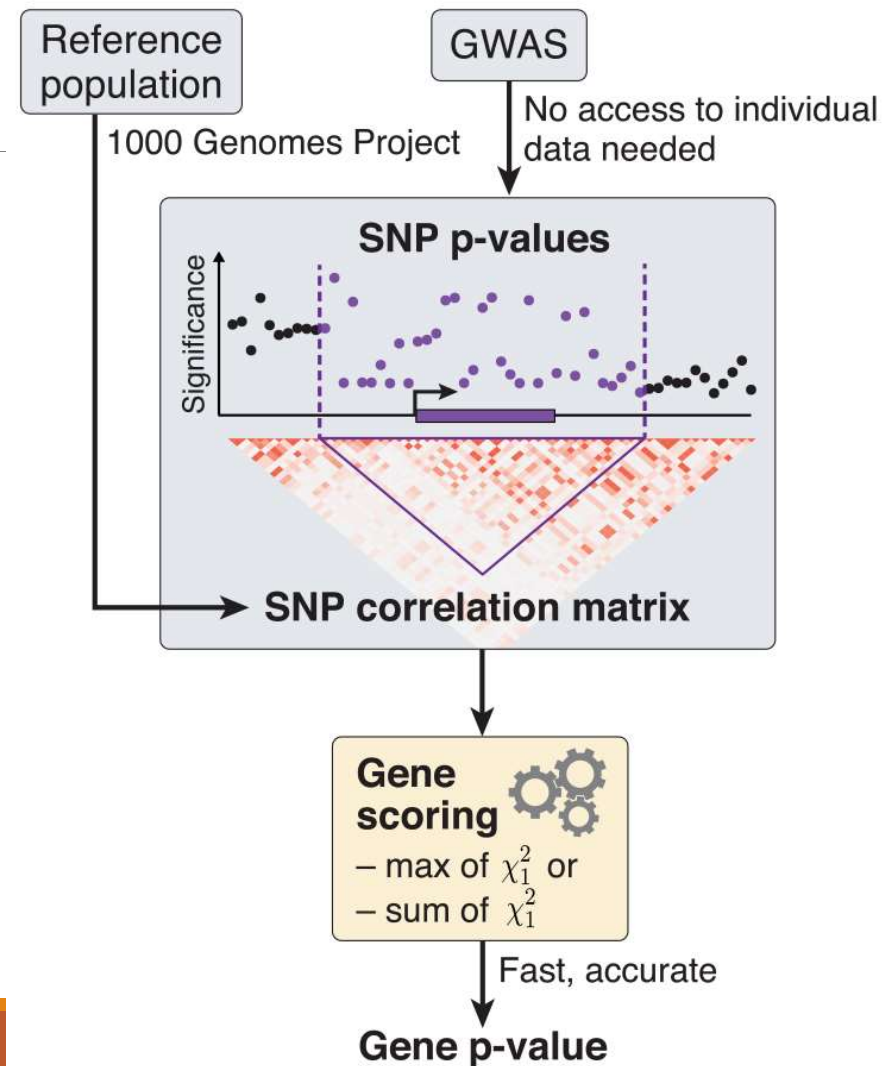
Pascal: Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics

Magma: Multi-marker Analysis of GenoMic Annotation

- Uses a multiple regression approach to properly incorporate LD between markers and to detect multi-marker effects

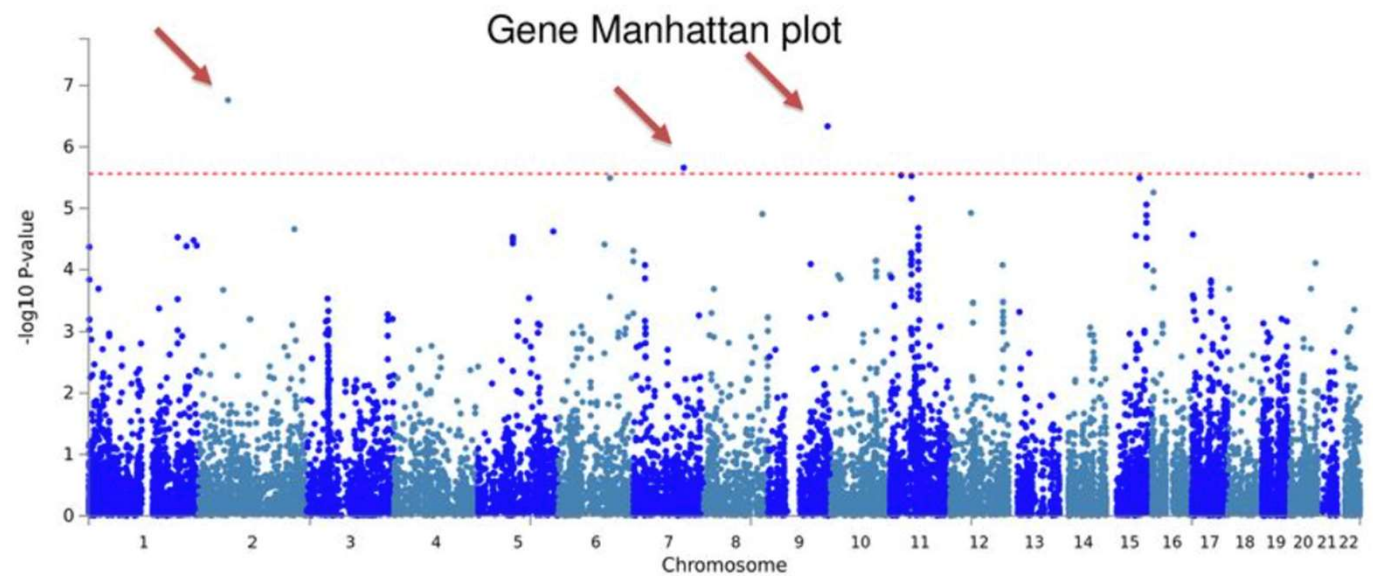
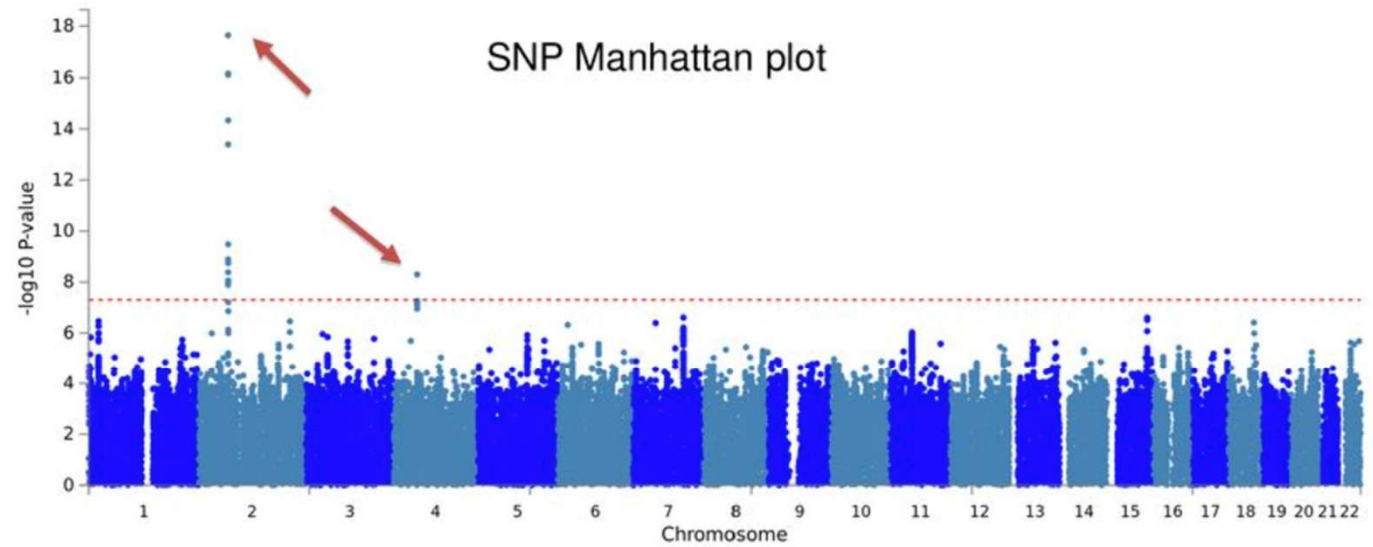
Use LD information from reference population

Allow for summary statistics data (no requirement on raw genotyping data)





# Example Manhattan



# Gene-based analysis

---

Unit of analysis is the gene

## Pros

- Reduce multiple testing correction from 2.5M SNPs to 22k genes
- Accounts for heterogeneity in genes
- Immediate gene-level interpretation

## Cons

- Disregards regulatory (often non-genic) information when based on location based annotation
- Still a lot of tests

# Outline

---

## Gene-based p-values

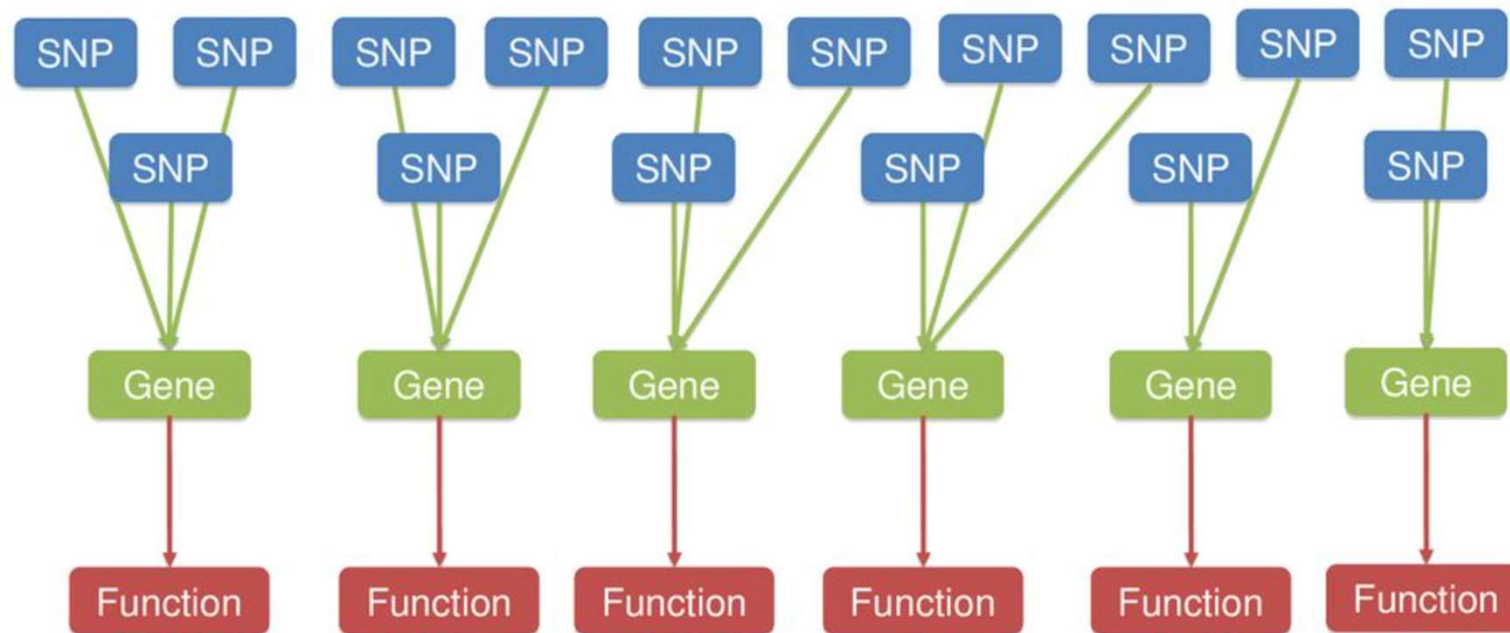
- Rationale
- Methods

## **Set-based analysis of GWAS data**

- **Pathway**
- **Network**

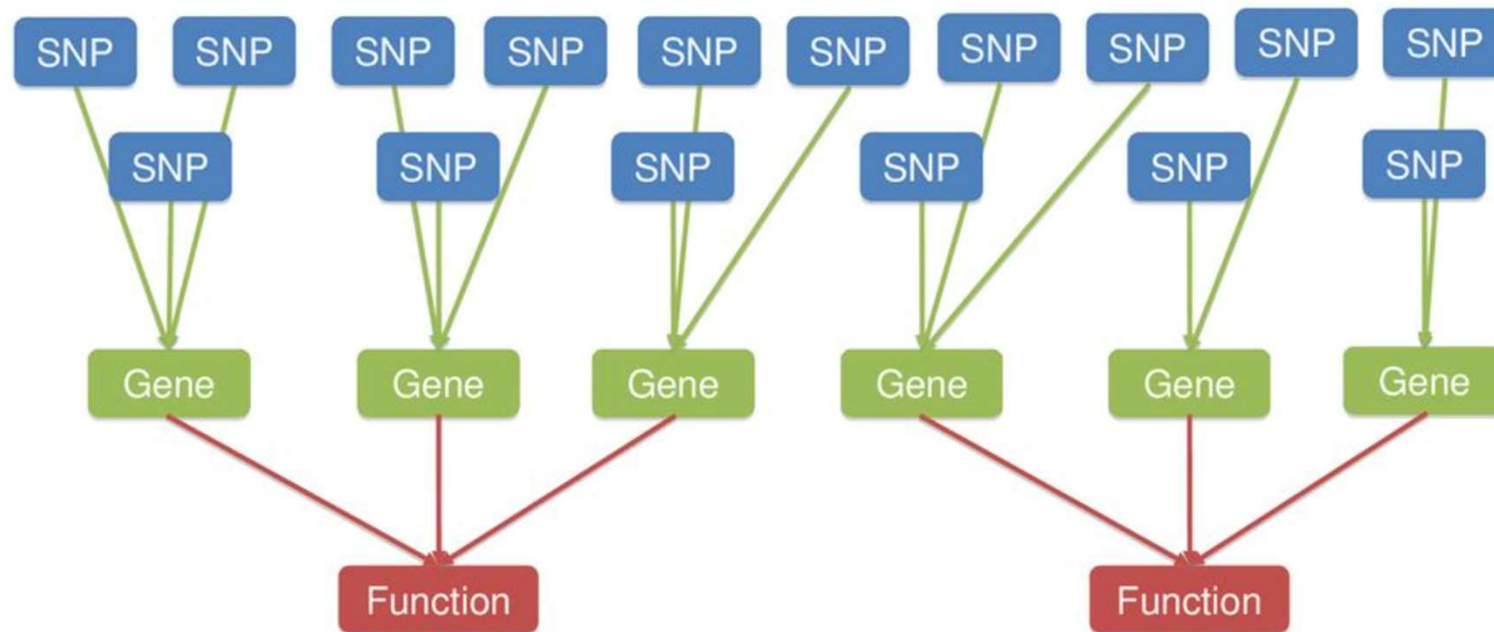
## Tissue and cell type specific enrichment analysis of GWAS data

# Do all implicated genes have different functions or are they functionally related?



# Do all implicated genes have different functions or are they functionally related?

---



# Gene-set analysis

---

Unit of analysis is a set of functionally related genes

## Pros

- Reduce multiple testing by prioritizing genes in biological pathways or in groups of (functionally) related genes
- Increases statistical power
- Deals with genic heterogeneity
- Provides immediate biological insight

## Cons

- Crucial to select reliable sets of genes
- Different levels of information
- Different quality of information

# Choosing gene-sets

---

Depends on the research question

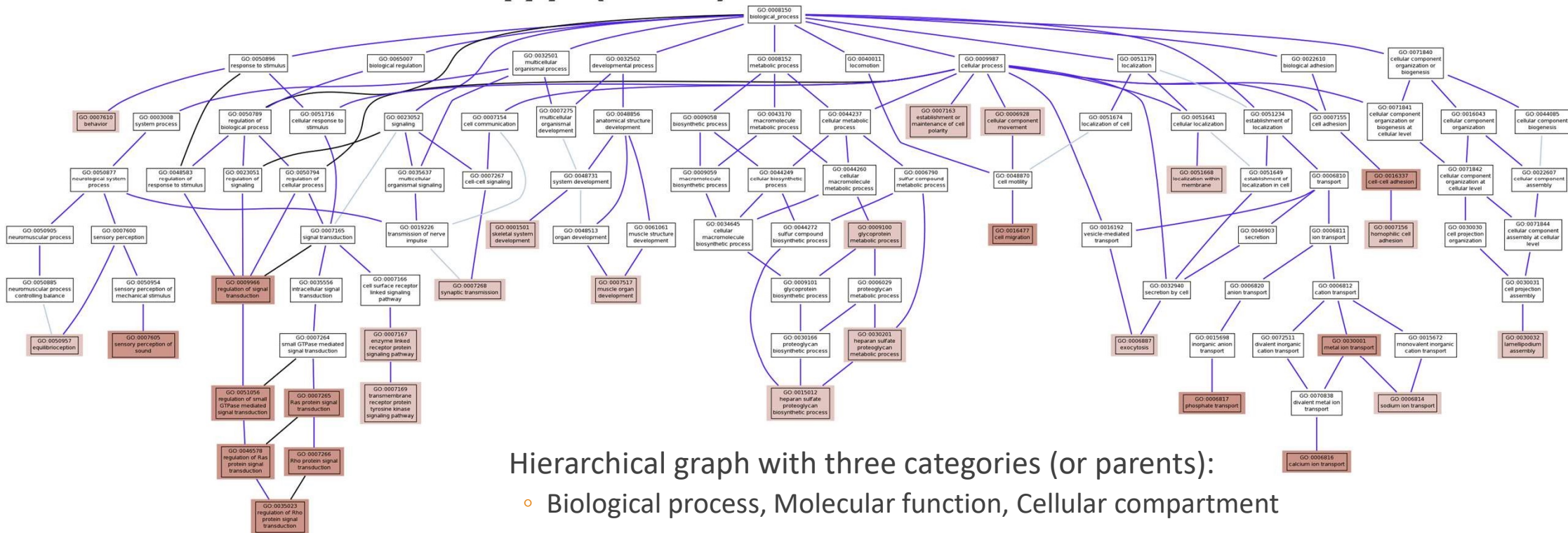
Several databases/resources available providing gene-set collections (e.g. MSigDB, Enrichr)

Included directly in some analysis tools

Gene-sets can be based on

- GO-terms (probably one of the most widely used gene-sets)
- Pathways (KEGG)
- Chromosomal locations
- Protein-protein interaction subnetworks
- Co-expression network
- Transcription regulatory network
- Known disease genes
- etc...

# Gene-set example: Gene ontology (GO) terms



Hierarchical graph with three categories (or parents):

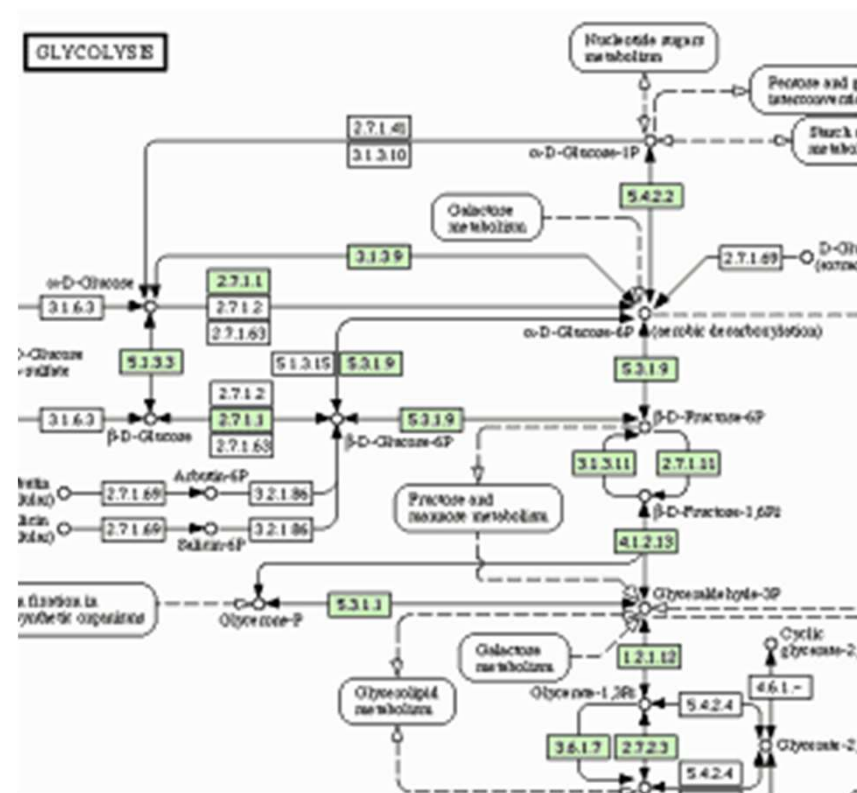
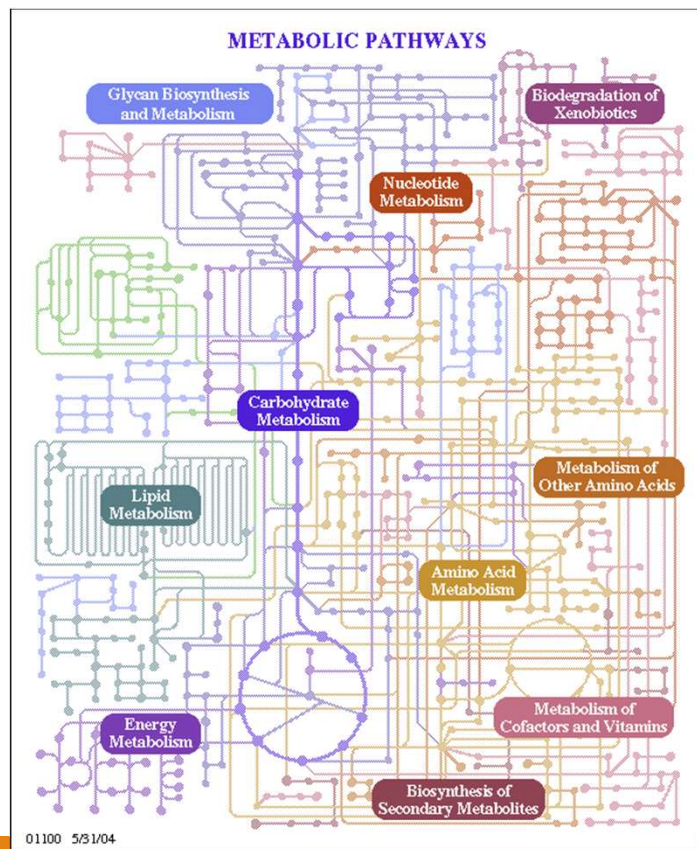
- Biological process, Molecular function, Cellular compartment

Terms get more and more detailed moving down the hierarchy

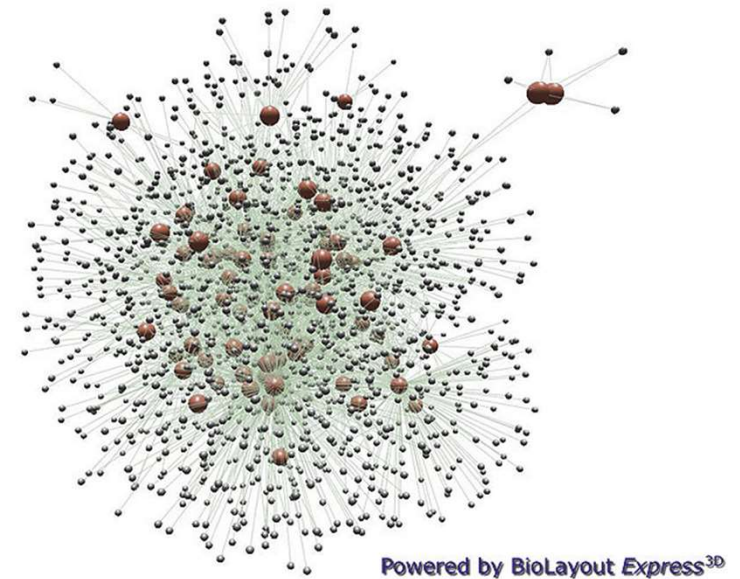
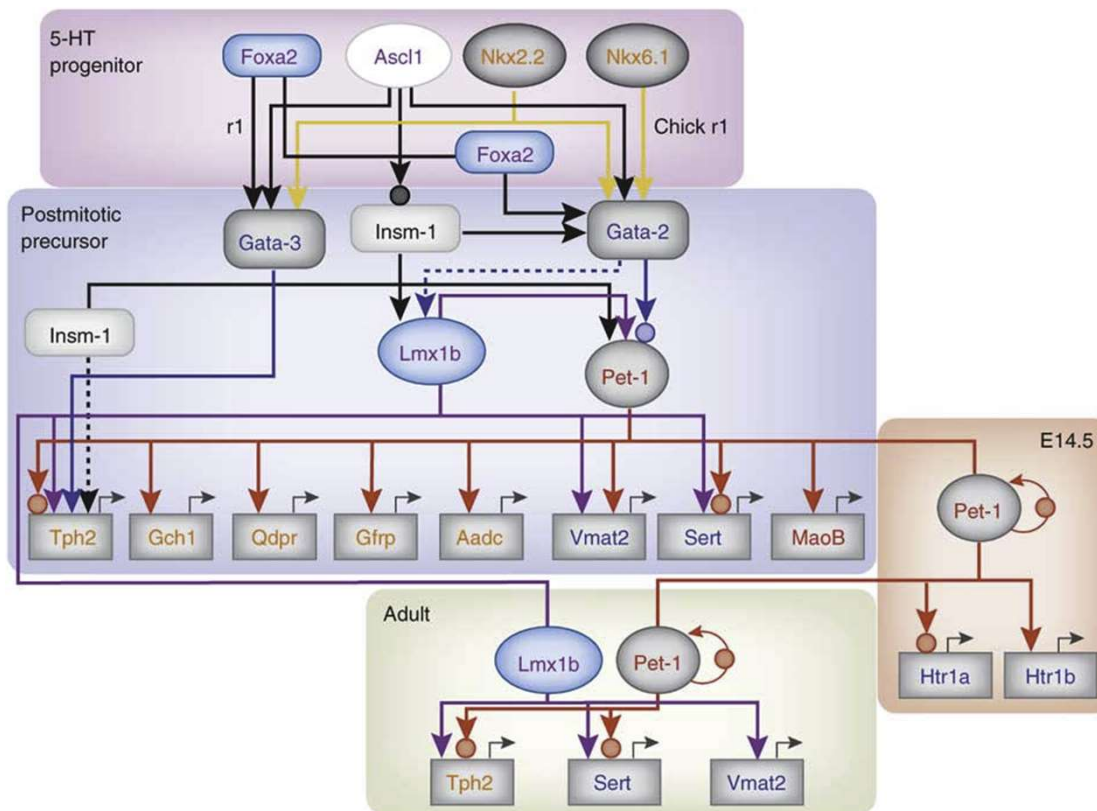
Genes can belong to multiple GO terms



# Gene-set example: Metabolic pathways or metabolites



# Gene-set example: Transcription factor targets



# Where to get gene-set collections?

<http://software.broadinstitute.org/gsea/msigdb/index.jsp>



Molecular Signatures Database v5.1

## Overview

The Molecular Signatures Database (MSigDB) is a collection of annotated gene sets for use with GSEA software. From this web site, you can

- ▶ **Search** for gene sets by keyword.
- ▶ **Browse** gene sets by name or collection.
- ▶ **Examine** a gene set and its annotations. See, for example, the [ANGIOGENESIS](#) gene set page.
- ▶ **Download** gene sets.
- ▶ **Investigate** gene sets:
  - ▶ **Compute overlaps** between your gene set and gene sets in MSigDB.
  - ▶ **Categorize** members of a gene set by gene families.
  - ▶ **View the expression profile** of a gene set in any of the three provided public expression compendia.

## Registration

Please [register](#) to download the GSEA software and view the MSigDB gene sets. After registering, you can log in at any time using your email address. Registration is free. Its only purpose is to help us track usage for reports to our funding agencies.

## Current Version

MSigDB database v5.1 updated January 2016. Release notes. GSEA/MSigDB web site v5.0 released March 2015

## Contributors

The MSigDB is maintained by the GSEA team with the support of our MSigDB Scientific Advisory Board. We also welcome and appreciate contributions to this shared resource and encourage

## Collections

The MSigDB gene sets are divided into 8 major collections:

- H** **hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.
- C1** **positional gene sets** for each human chromosome and cytogenetic band.
- C2** **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.
- C3** **motif gene sets** based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.
- C4** **computational gene sets** defined by mining large collections of cancer-oriented microarray data.
- C5** **GO gene sets** consist of genes annotated by the same GO terms.
- C6** **oncogenic signatures** defined directly from microarray gene expression data from cancer gene perturbations.
- C7** **immunologic signatures** defined directly from microarray gene expression data from immunologic studies.

<http://amp.pharm.mssm.edu/Enrichr/#stats>



[Login](#) | [Register](#)

1,052,595 lists analyzed

[Analyze](#) [What's New?](#) [Libraries](#) [Find a Gene](#) [About](#) [Help](#)

Gene-set Library	Terms	Gene Coverage	Genes per Term
Achilles_fitness_decrease	216	4271	128.0
Achilles_fitness_increase	216	4320	129.0
Aging_Perturbations_from_GEO_down	286	16129	292.0
Aging_Perturbations_from_GEO_up	286	15309	308.0
Allen_Brain_Atlas_down	2192	13877	304.0
Allen_Brain_Atlas_up	2192	13121	305.0
BioCarta_2013	249	1295	18.0
BioCarta_2015	239	1678	21.0
BioCarta_2016	237	1348	19.0
Cancer_Cell_Line_Encyclopedia	967	15797	176.0
ChEA_2013	353	47172	1370.0
ChEA_2015	395	48230	1429.0
Chromosome_Location	386	32740	85.0
CORUM	1658	2741	5.0
dbGaP	345	5613	36.0
Disease_Perturbations_from_GEO_down	839	23939	293.0
Disease_Perturbations_from_GEO_up	839	23561	307.0
Disease_Signatures_from_GEO_down_2014	142	15406	300.0
Disease_Signatures_from_GEO_up_2014	142	15057	300.0
Drug_Perturbations_from_GEO_2014	701	47107	509.0
Drug_Perturbations_from_GEO_down	906	23877	302.0
Drug_Perturbations_from_GEO_up	906	24350	299.0
ENCODE_and_ChEA_Consensus_TFs_from_ChIP-X	104	15562	887.0
ENCODE_Histone_Modifications_2013	109	15852	912.0
ENCODE_Histone_Modifications_2015	412	29065	2123.0

# Online databases vs. manual

---

Information in online databases tends to be

Somewhat biased

- Not all genes included – disease genes tend to be investigated more often
- Genes that are investigated more often will have more interactions

Not always reliable

- Interactions often not validated, sometimes only predicted. If experimentally seen, unknown how reliable that experiment was

# Statistical issues in gene-set analyses

---

Different statistical algorithms test different alternative hypotheses

Different statistical algorithms have different sensitivity to LD, number of genes, number of SNPs, background  $h^2$

Self-contained vs. competitive tests

Self-contained:

- $H_0$ : the gene sets are not associated with the trait

Competitive:

- $H_0$ : the genes in the gene-set are not more strongly associated with the trait than the genes not in the gene-set



# Schematic of the two tier structure of GSA

**a** | A measure of association with the phenotype is computed per gene from the genotype data.

**b** | This results in a gene-level data matrix, with each row corresponding to a gene and the gene set encoded as a binary indicator variable (coding genes in the gene set as 1 and the rest as 0). The gene-set analysis (GSA) then takes the form of a bivariate test with the genes as units of analysis, testing whether the joint association of genes in the gene set is greater than if those genes were not associated at all (self-contained analysis) or whether it is greater than the association of genes not in the gene set (competitive analysis). SNPs, single nucleotide polymorphisms.

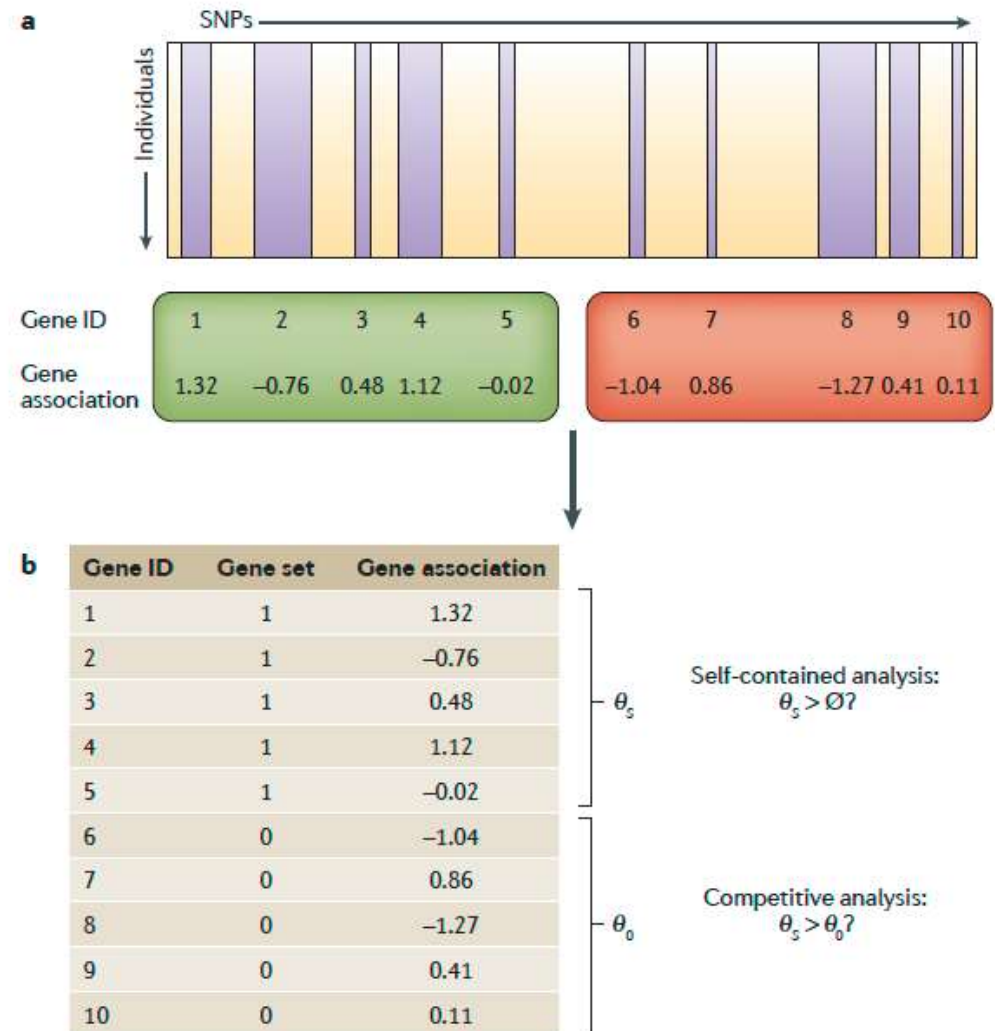
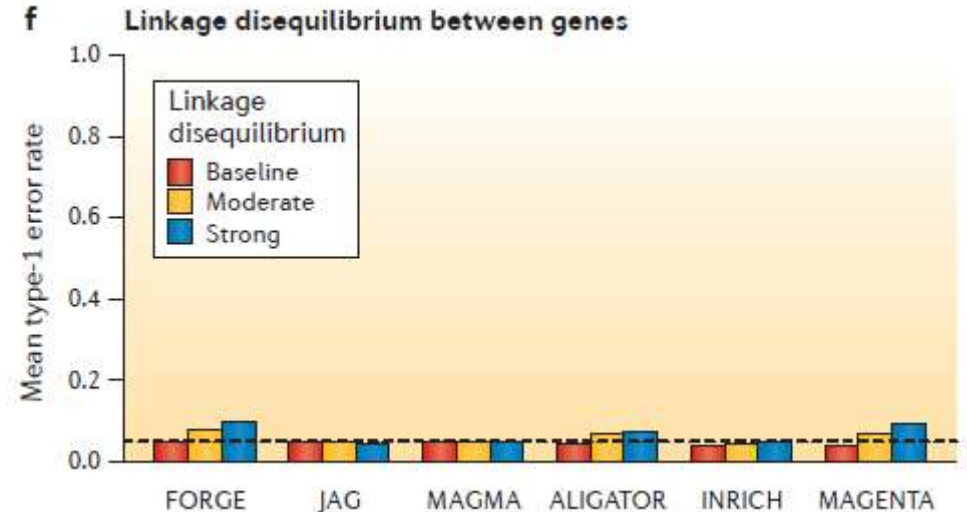
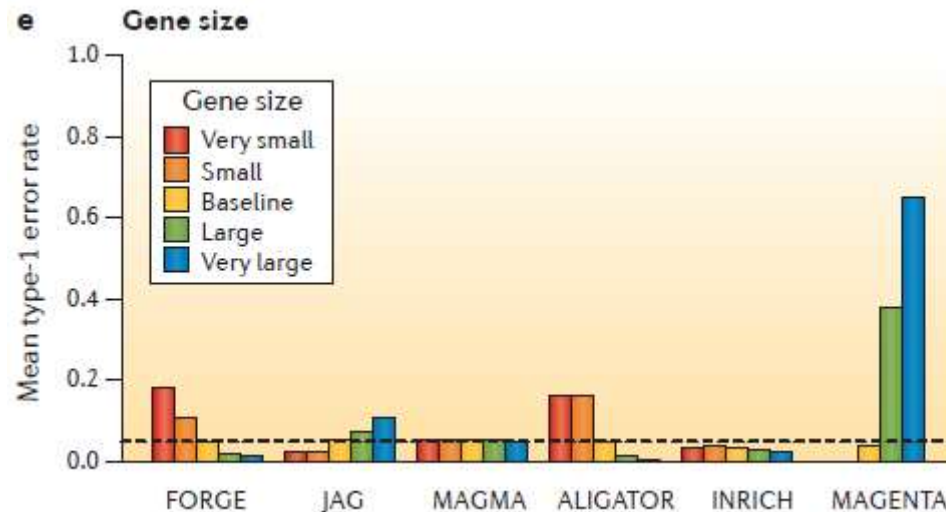


Table 1 | General classification of GSA methods\*

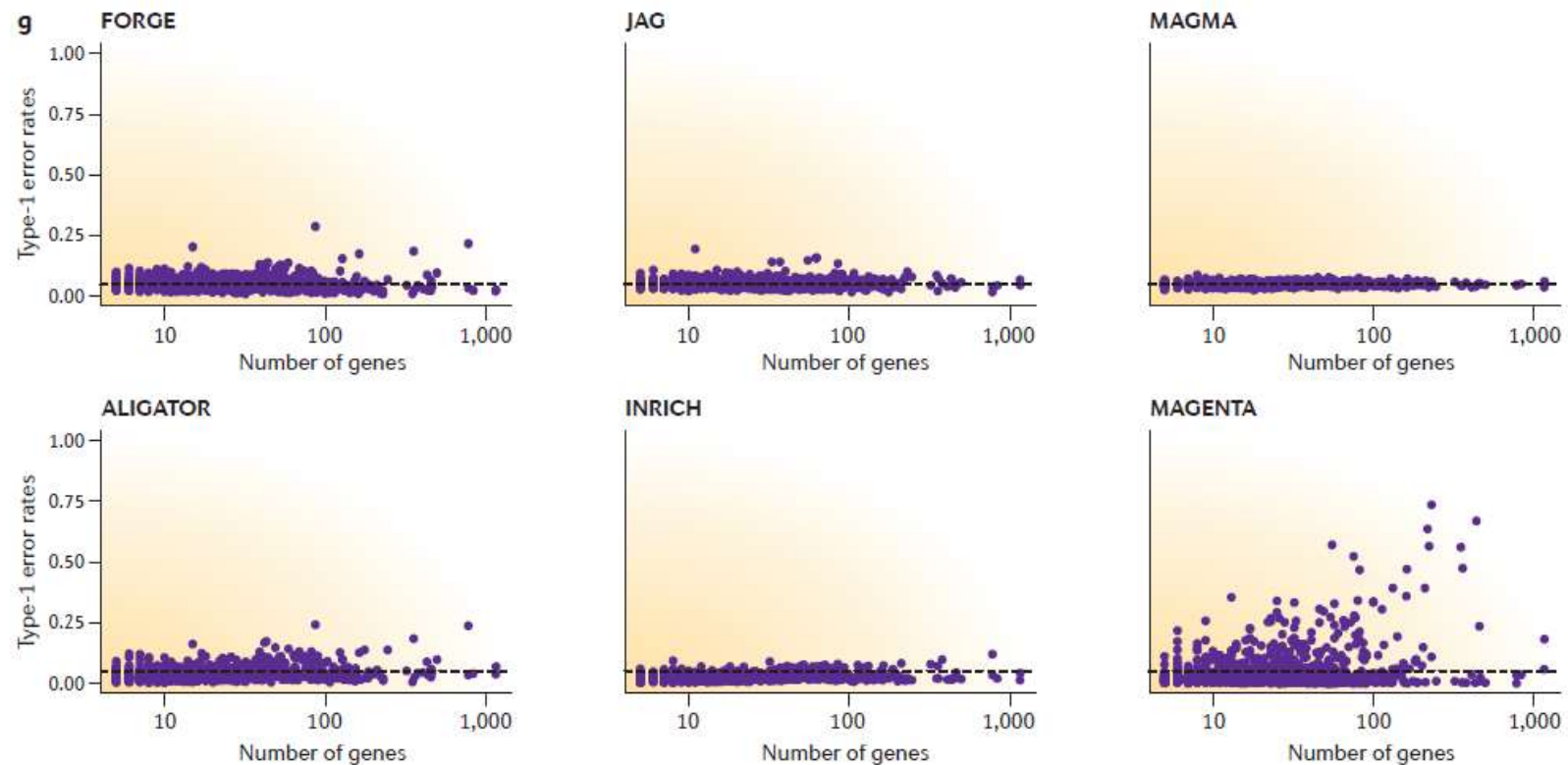
Method	Type	Description	Example tools
<b>Mean-based</b>			
Fisher's method	Self-contained	Tests mean of $-\log$ or transformed $P$ values in the set against the null <sup>‡</sup> mean	KGG-HYST <sup>33</sup> , PLINK <sup>29</sup> , SetScreen <sup>30</sup> and JAG <sup>32</sup>
Fisher's method	Competitive	Tests mean of $-\log$ or transformed $P$ values in the set against mean outside of the set	JAG
Single sample Z-test	Self-contained	Tests mean of probit transformed $P$ values in the set against the null <sup>‡</sup> mean	FORGE <sup>34</sup> and MAGMA <sup>35</sup>
Two-sample t-test	Competitive	Tests mean of probit transformed $P$ values in the set against mean outside of the set	FORGE
Linear regression <sup>§</sup>	Competitive	Tests whether being in the set or not is a predictor of having higher probit transformed $P$ values	MAGMA
<b>Count-based<sup>  </sup></b>			
Binomial test	Self-contained	Tests whether proportion of $P$ values in the set below the threshold is greater than the null <sup>‡</sup> proportion	SNP Ratio Test <sup>31</sup>
Hypergeometric test	Competitive	Tests whether proportion of $P$ values below the threshold in the set is greater than the proportion outside the set	ALIGATOR <sup>37</sup> , INRICH <sup>38</sup> and MAGENTA <sup>39</sup>
Logistic regression <sup>§</sup>	Competitive	Tests whether being in the set or not is predictor of having $P$ values below the threshold	—
<b>Rank-based</b>			
Two-sample KS test	Competitive	Tests whether genes in the set are overrepresented at the top of the list of all genes ranked by $P$ value	—
<b>Rank + mean-based</b>			
GSEA	Self-contained or competitive	Modified KS test, weight ranks by $-\log$ or transformed $P$ values	GenGen <sup>36</sup>

# Different algorithms: LD and #genes





# Different algorithms: LD and #genes



# Outline

---

## Gene-based p-values

- Rationale
- Methods

## Set-based analysis of GWAS data

- Pathway
- **Network**

## Tissue and cell type specific enrichment analysis of GWAS data

# Biological networks

Protein-protein interaction (physical interaction)

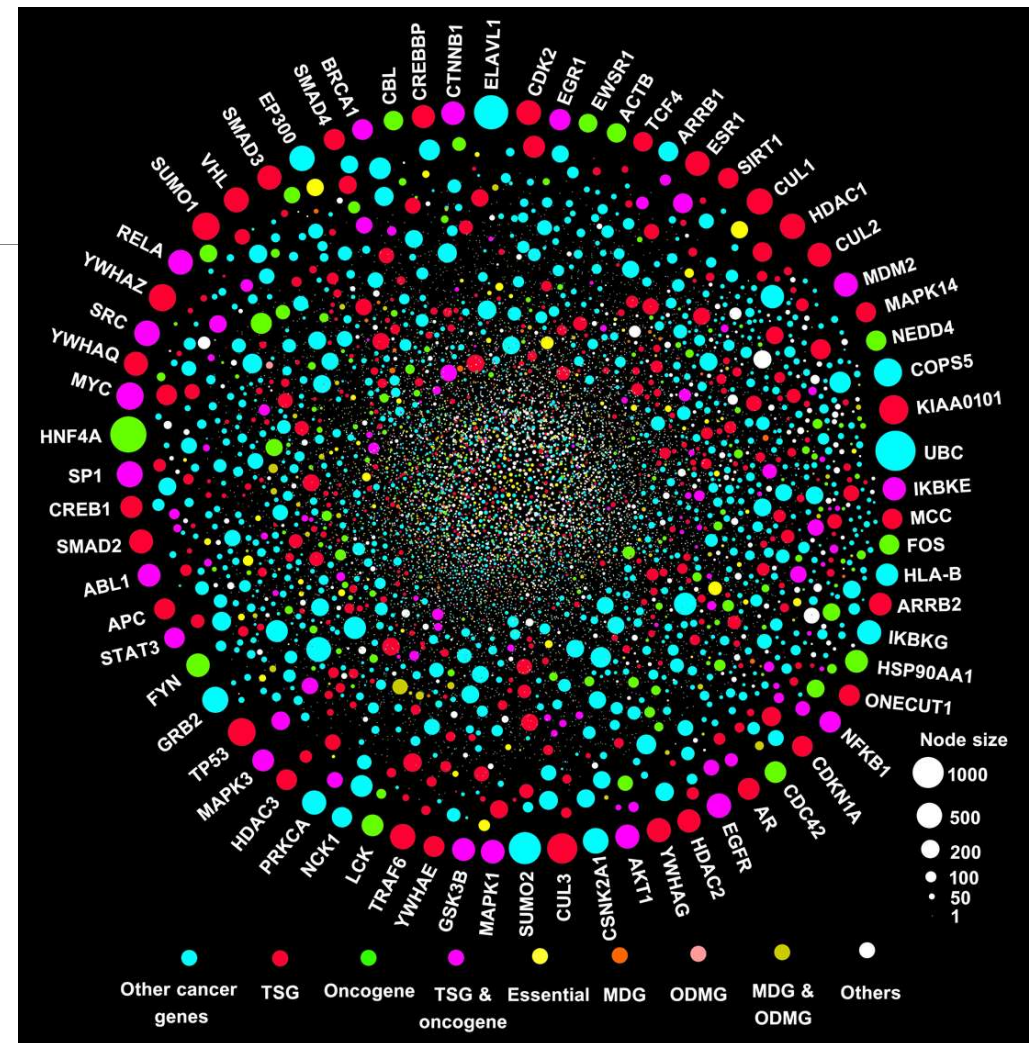
Kinase-substrate & phosphorylation-related interaction network (KSIN)

Protein 3D structure-based PPIN (3DPPIN)

Innate immunity PPIN (INPPIN)

Protein functional relations in pathways

Co-expression network



Cheng F, Jia P, Wang Q, Lin CC, Li WH, Zhao Z (2014) Studying tumorigenesis through network evolution and somatic mutational perturbations in the cancer interactome. *Molecular Biology and Evolution*, 31(8):2156-2169

# Resources for human biological networks

Networks	Nodes	Edges	Links and description
HPRD	~10,000	~50,000	Human Protein Reference Database
BioGRID	21,008	280,910	Interaction repository with curated data.
PINA	17,109	166,776	Integrative PPI data from six public curated databases: IntAct, BioGRID, MINT, DIP, HPRD, and MIPS/MPact.
MAGI	10,349	52,663	Physical PPIs: A combined dataset of STRING and HPRD
HumanNet, base	5,236	269,410	A reference set of gene pairs sharing Gene Ontology biological process annotations
HumanNet join	16,117	474,714	Predicted gene association pairs based on a Naïve Bayes approach
Pathway Common	16,305	369,884	Physical PPIs: An aggregated repository of gene interactions from several sources including BioGrid, HPRD, IntAct, and the NCI cancer specific pathways.

# Characteristics of biological networks

Degree: how many edges it is connected to

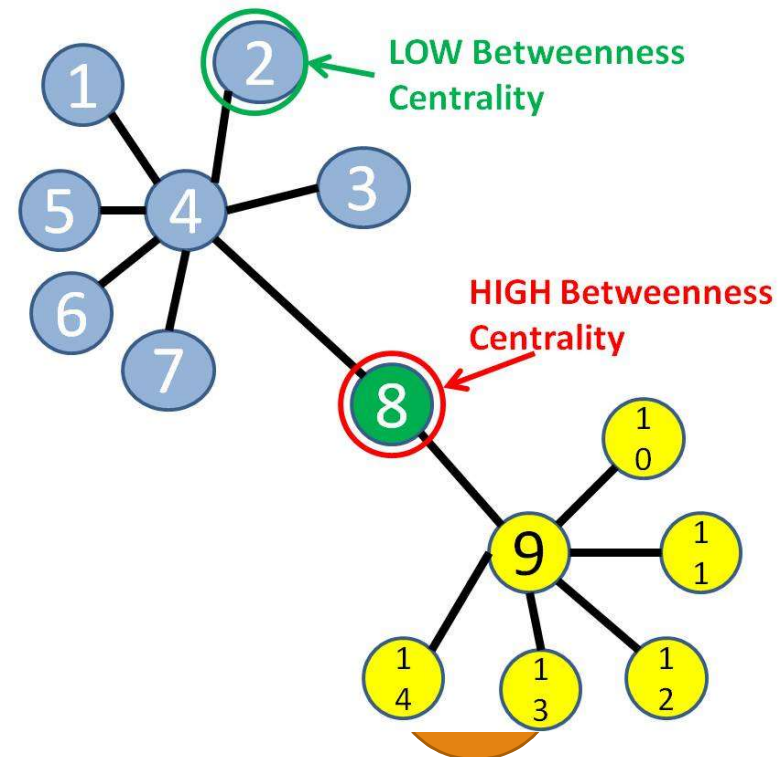
- E.g., degree of node A = 5
- Important genes tend to have large degree

Path and shortest path

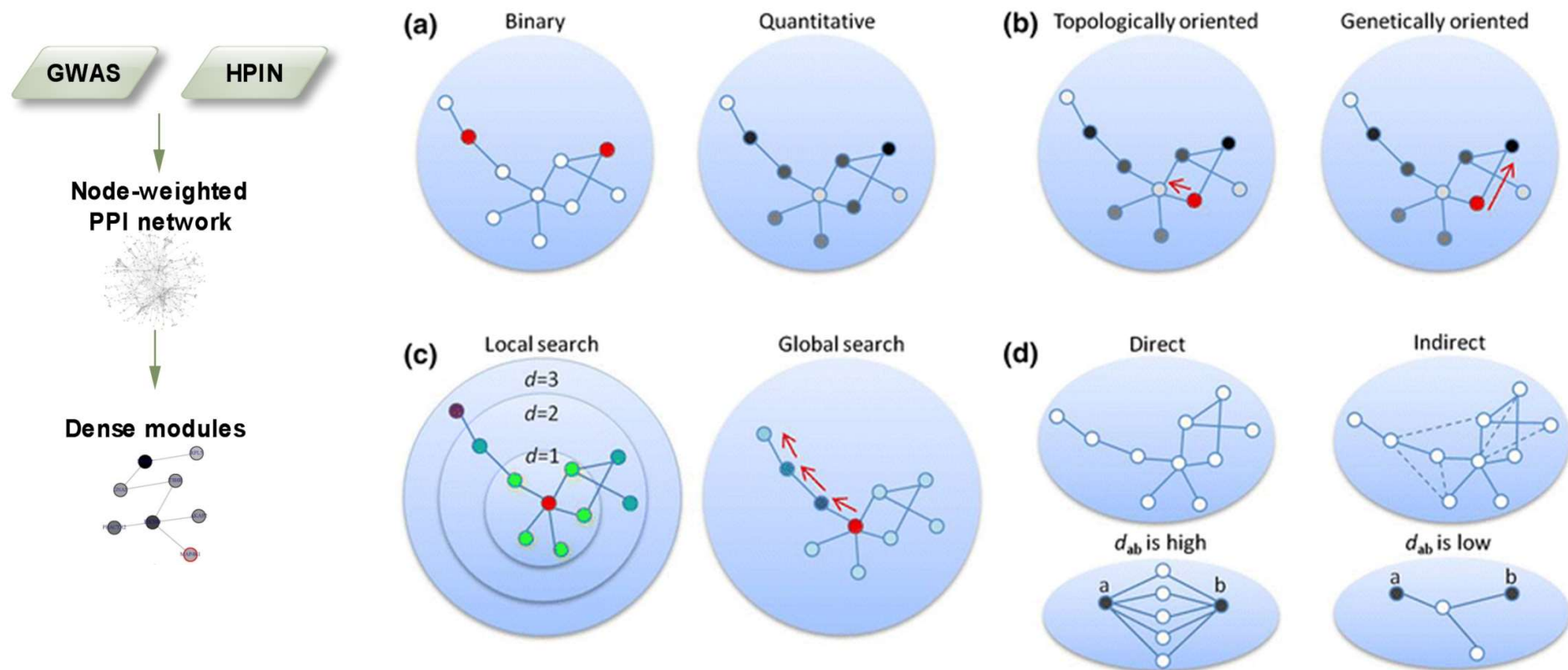
- E.g., shortest path between nodes B and F = (B-A and A-F)

Betweenness

- Defined for both nodes and edges

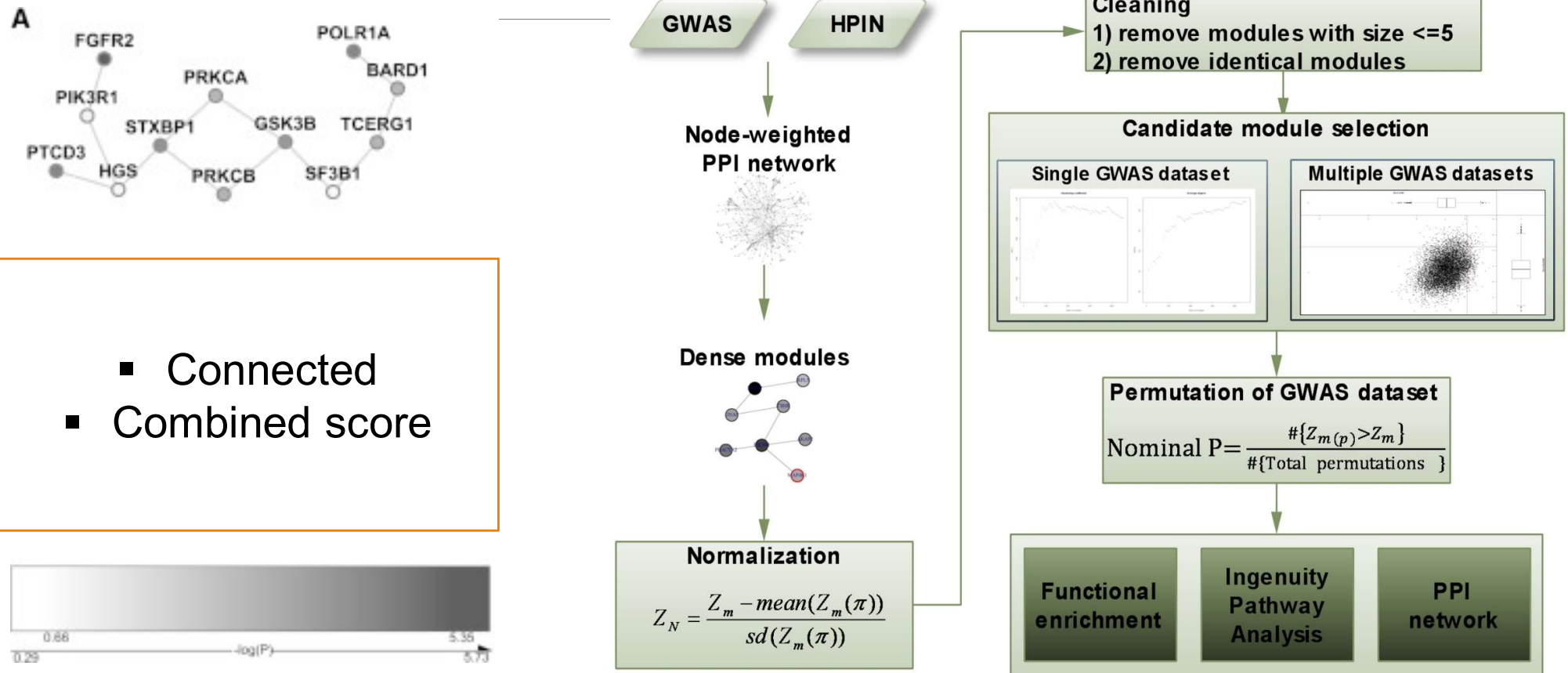


# Algorithms in network-assisted analysis of GWAS data





# dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks



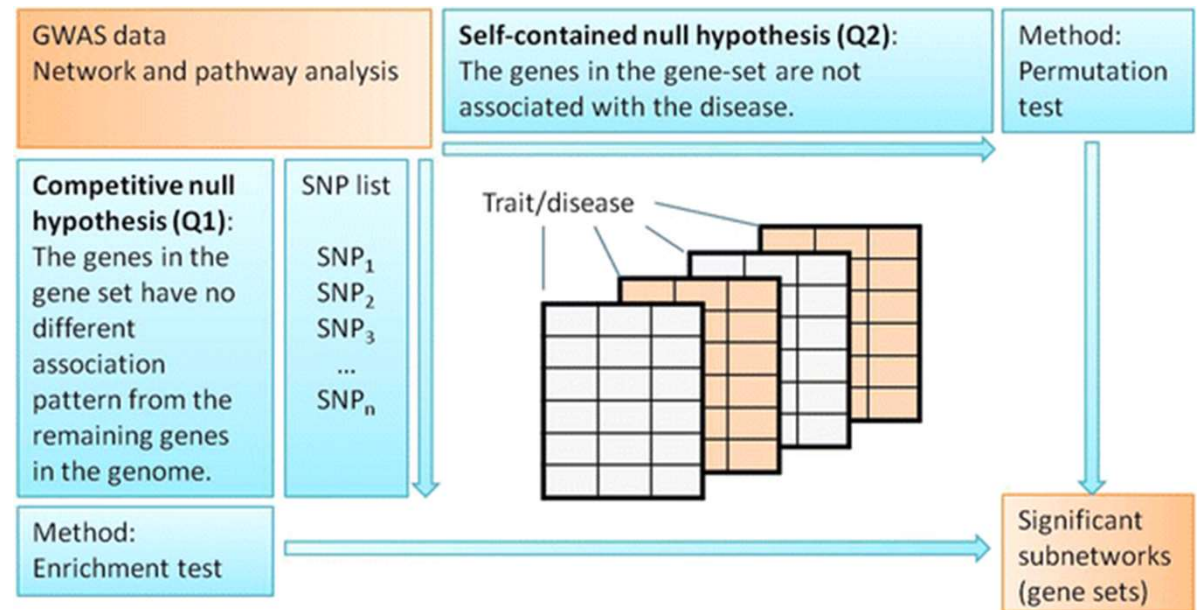
# Application

## Gene-based score

- Minimum p-value
- Combining multiple SNPs

## Significance

- Permutation: randomize case/control sample IDs





[illegible]

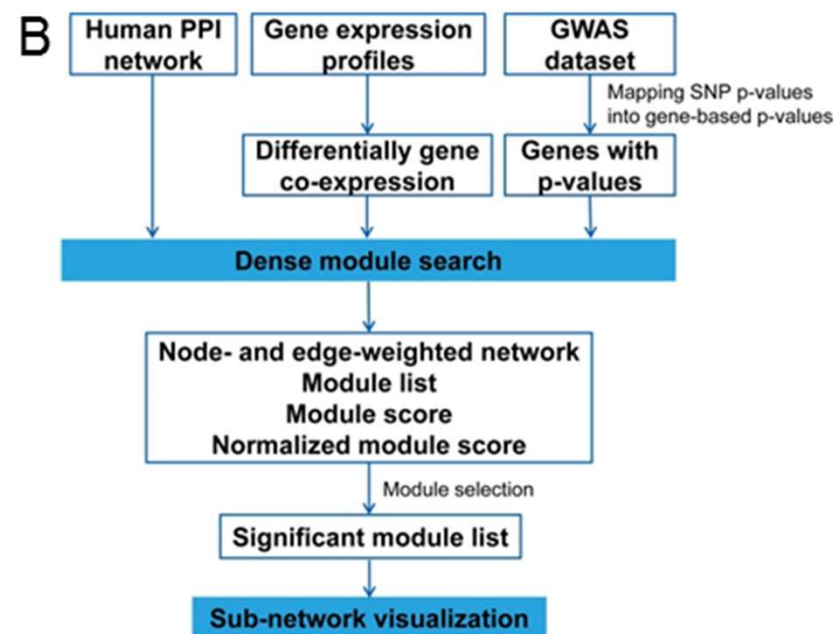
A subnetwork enriched with schizophrenia associated genes.

# EW-dmGWAS: **n**ode- and **e**dge-**w**weighted dense module search for GWAS and gene expression profile

Node weight: z-score from GWAS p-value

Edge weight: Fisher's transformation on Pearson Correlation Coefficient (PCC)

Gene expression profile: a set of disease samples and a set of control, suggested to use the disease-relevant tissues.



$$S = \lambda \frac{\sum_{e \in E} \text{edgeweight}(e)}{\sqrt{\# E}} + (1 - \lambda) \frac{\sum_{v \in V} \text{nodeweight}(v)}{\sqrt{\# V}}$$

# Outline

---

## Gene-based p-values

- Rationale
- Methods

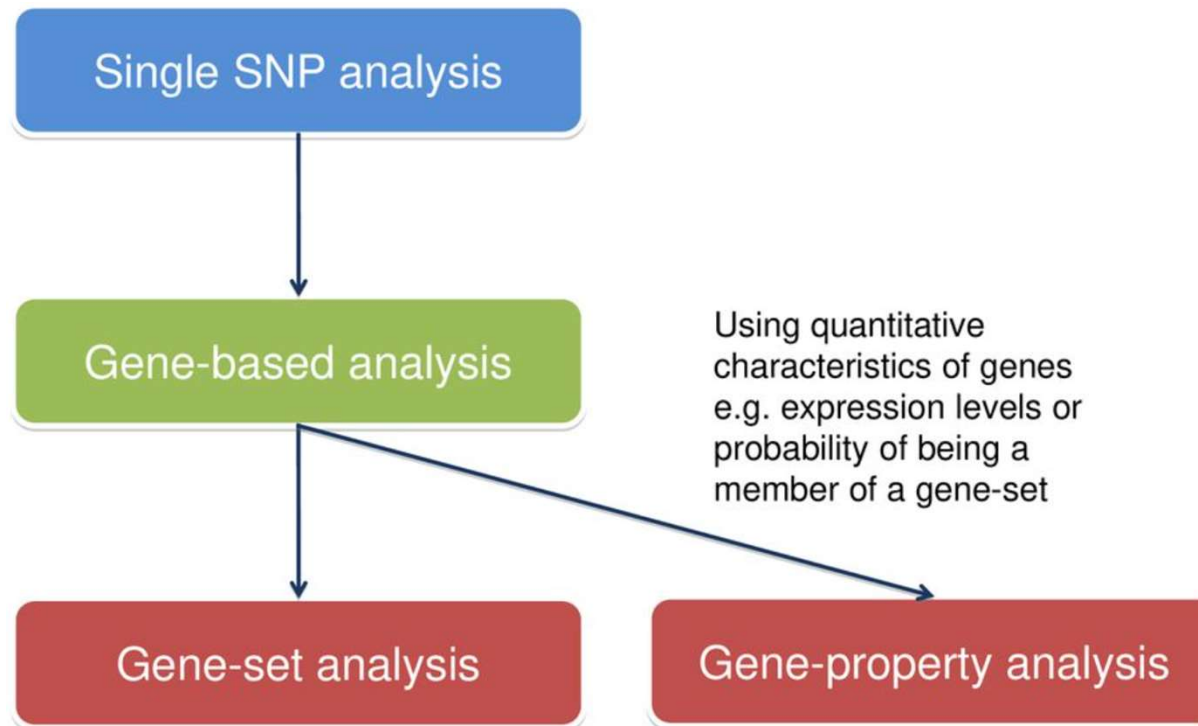
## Set-based analysis of GWAS data

- Pathway
- Network

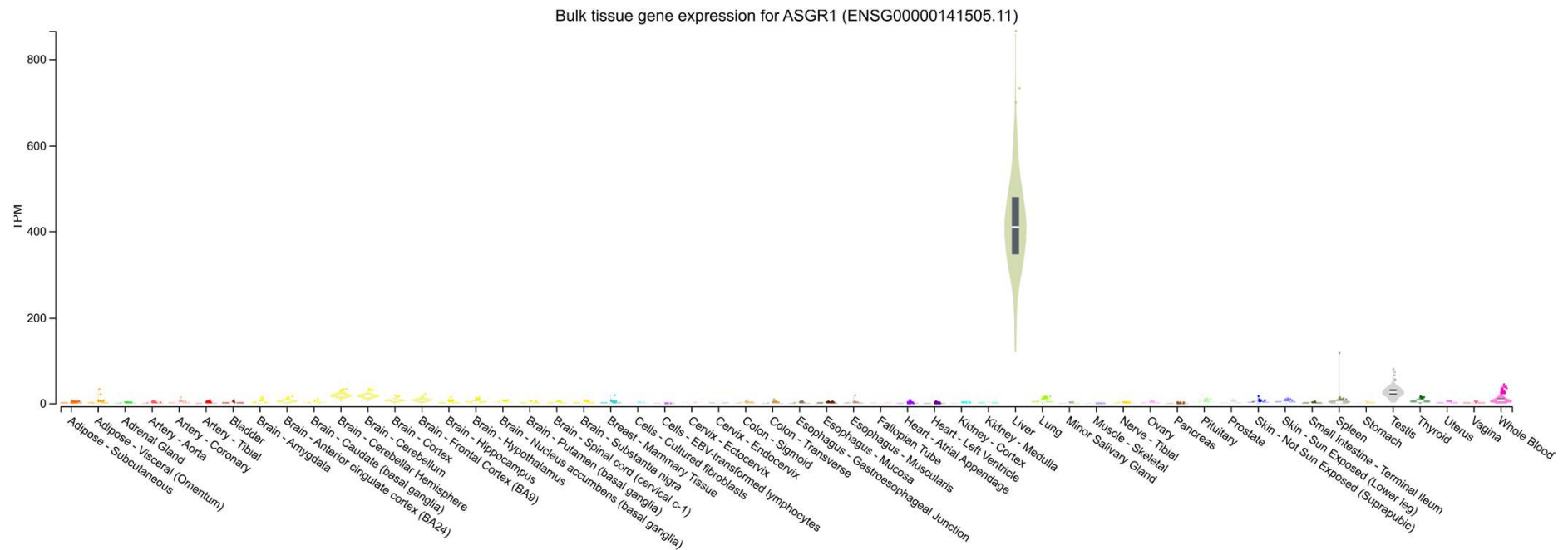
## **Tissue and cell type specific enrichment analysis of GWAS data**

# Gene-property analysis

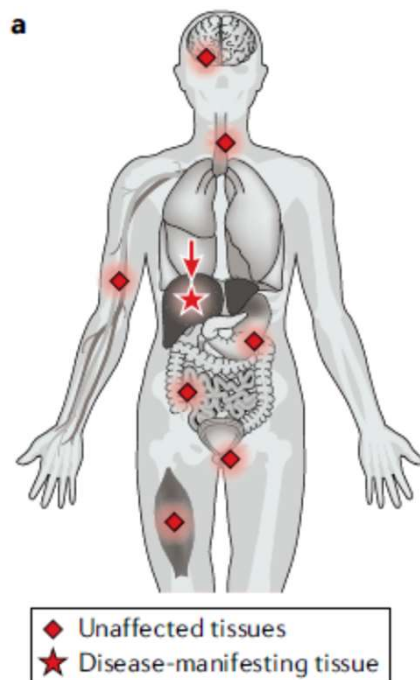
---



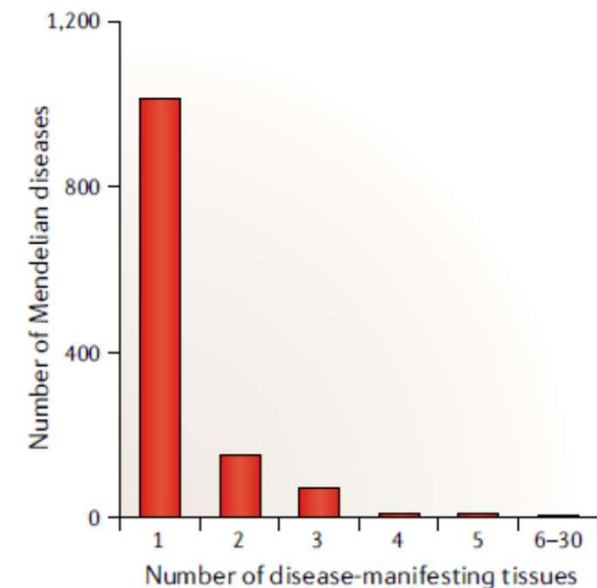
# What is tissue-specificity



# The tissue selectivity of heritable diseases and traits and their associated genes



Germline aberrations that underlie diseases are present throughout the body (red diamonds) yet often manifest clinically in few tissues



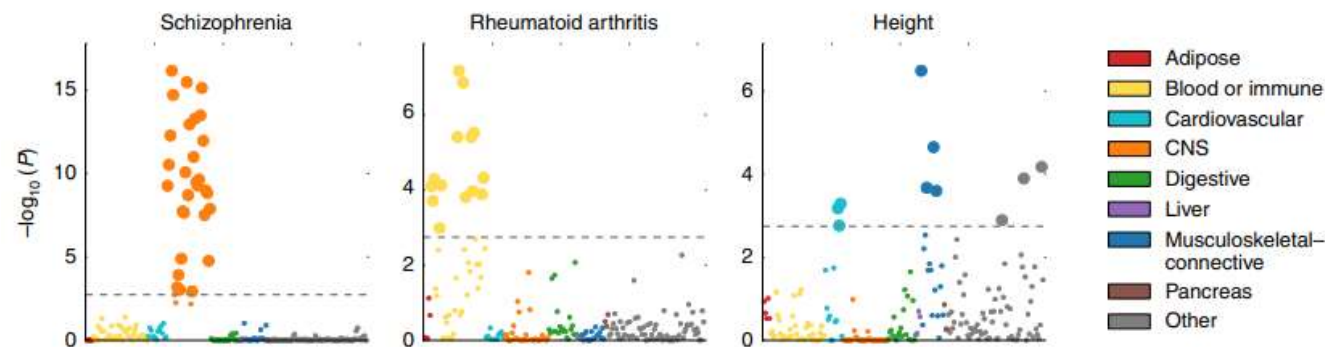
Data from 1,252 Mendelian diseases

# Can we find in which tissues that the GWAS-implied genes are actively expressed



## Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types

Hilary K. Finucane<sup>1,2,3\*</sup>, Yakir A. Reshef<sup>4</sup>, Verner Anttila<sup>1,5</sup>, Kamil Slowikowski<sup>1,6,7</sup>, Alexander Gusev<sup>8</sup>, Andrea Byrnes<sup>1,5</sup>, Steven Gazal<sup>9</sup>, Po-Ru Loh<sup>3</sup>, Caleb Lareau<sup>1,8</sup>, Noam Shores<sup>1</sup>, Giulio Genovese<sup>1</sup>, Arpiar Saunders<sup>9</sup>, Evan Macosko<sup>9</sup>, Samuela Pollack<sup>3</sup>, The Brainstorm Consortium<sup>10</sup>, John R. B. Perry<sup>11</sup>, Jason D. Buenrostro<sup>1,12</sup>, Bradley E. Bernstein<sup>1,13</sup>, Soumya Raychaudhuri<sup>1,7,14,15,16</sup>, Steven McCarroll<sup>1,9</sup>, Benjamin M. Neale<sup>1,5</sup>, and Alkes L. Price<sup>1,3\*</sup>





# Different mechanisms underlying tissue-selective manifestation of heritable traits and diseases

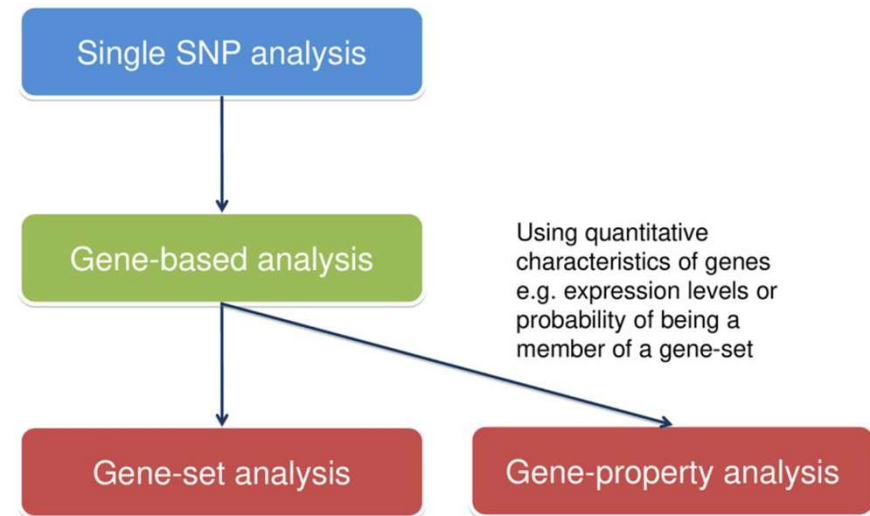
---

**1. *Expression-based mechanisms***

**2. *Regulatory mechanisms***

**3. *Tissue-disrupted networks***

**4. *Non-cell-autonomous mechanisms***



# Different mechanisms underlying tissue-selective manifestation of heritable traits and diseases

## 1. Expression-based mechanisms

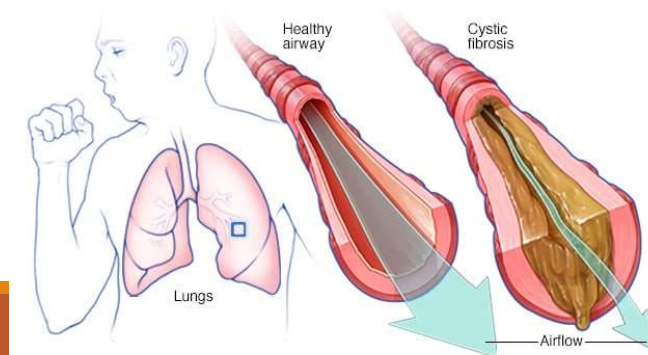
Exclusive expression of the causal gene in susceptible tissues

- CAV3 (caveolin 3) encodes the muscle-specific form of the caveolin protein family and is expressed primarily in heart and skeletal muscle. When mutated, causes cardiomyopathies and skeletal muscle disorders.

Preferential expression of the causal gene

- CFTR, causal for Cystic fibrosis, is highly expressed in a rare cell type found in the mouse respiratory system.
- Cystic fibrosis (CF) is an inherited disorder that causes severe damage to the lungs, digestive system and other organs in the body.

Mechanism	Affected tissue	Unaffected tissue
<b>a Expression-based mechanisms</b>		
① Exclusive expression		
② Preferential expression		
③ Post-transcriptional processes		
④ Reduced functional redundancy		






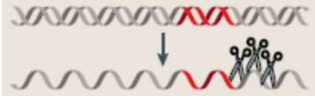
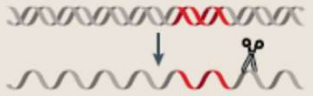

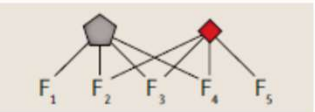
© MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH. ALL RIGHTS RESERVED.

# 1. *Expression-based mechanisms*

## Tissue-specific post-transcriptional processes

- Posttranscriptional processing affects the halflife and function of genes, making the relationship between transcript levels, protein levels and protein activity complicated and tissue-specific.
- COL10A1 gene encoding collagen X expressed in cartilage cells.

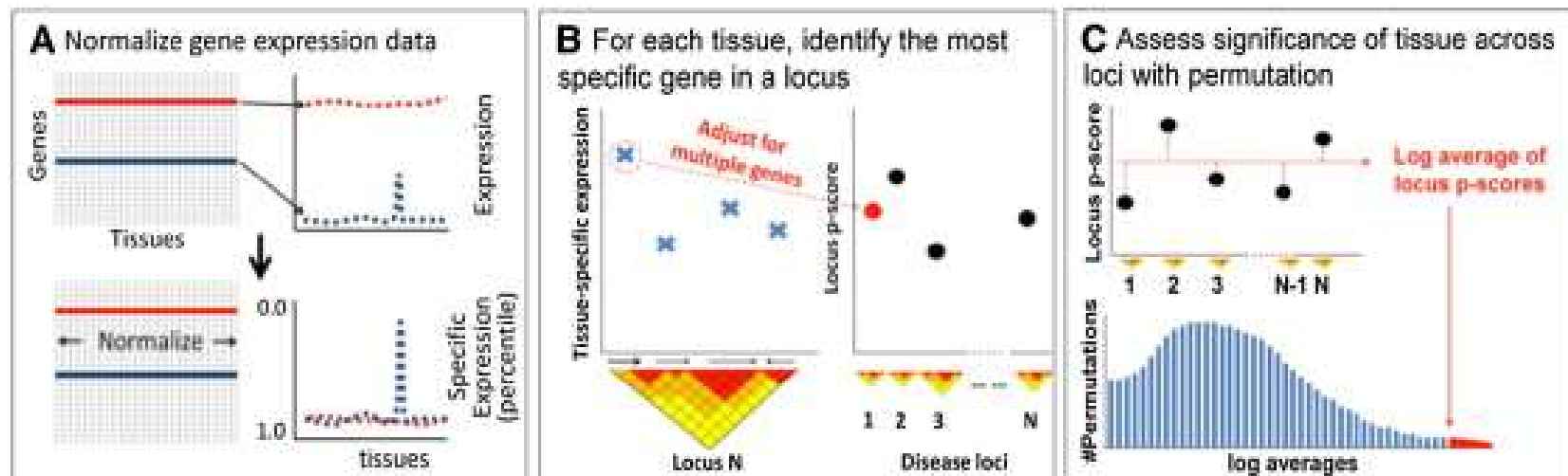
## Tissue-restricted functional redundancy

Mechanism	Affected tissue	Unaffected tissue
<b>a Expression-based mechanisms</b>		
① Exclusive expression		
② Preferential expression		
③ Post-transcriptional processes		
④ Reduced functional redundancy		

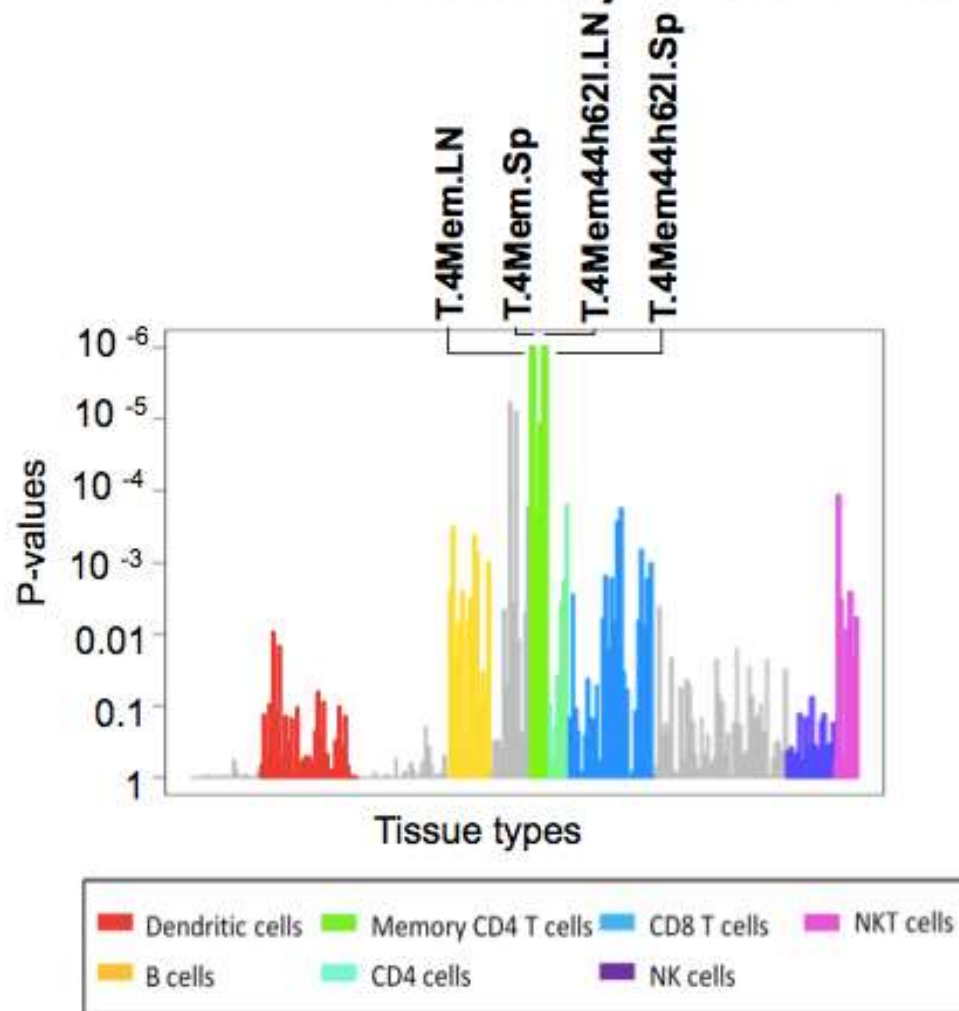
# Identifying Causal Cell-type for Complex Disease From Expression

Negative control: simulation from random set of SNPs

P-value: proportion of simulations exceeding the observed enrichment



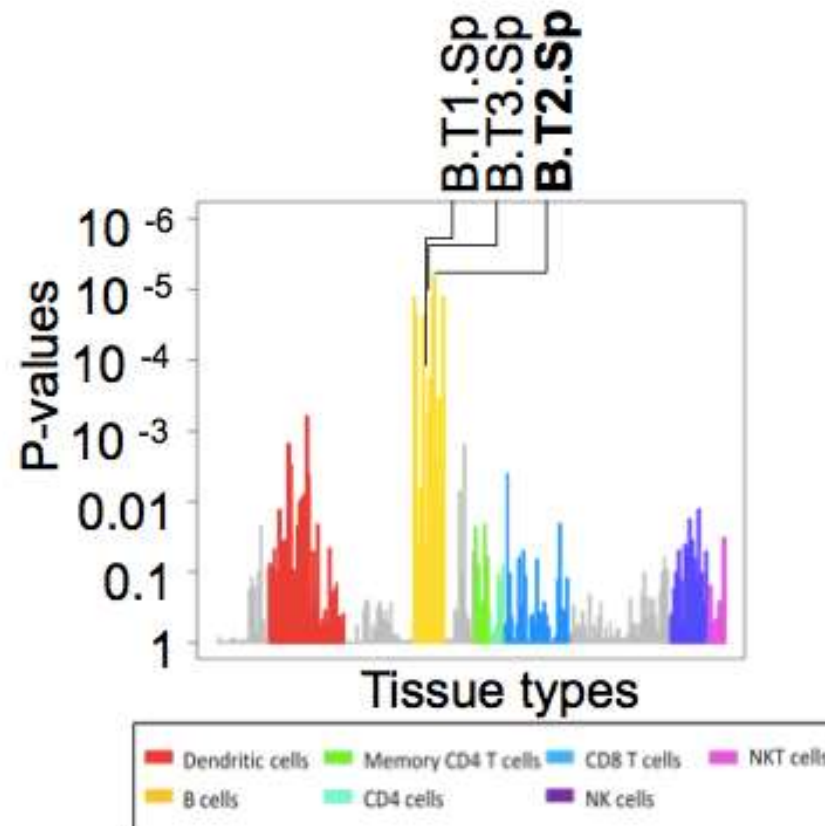
# Rheumatoid Arthritis implicates CD4+ Effector Memory T-cell Subsets



Examined 39 RA  
risk loci -  
implicating 170  
genes in  
aggregate

# Application to Immune Diseases

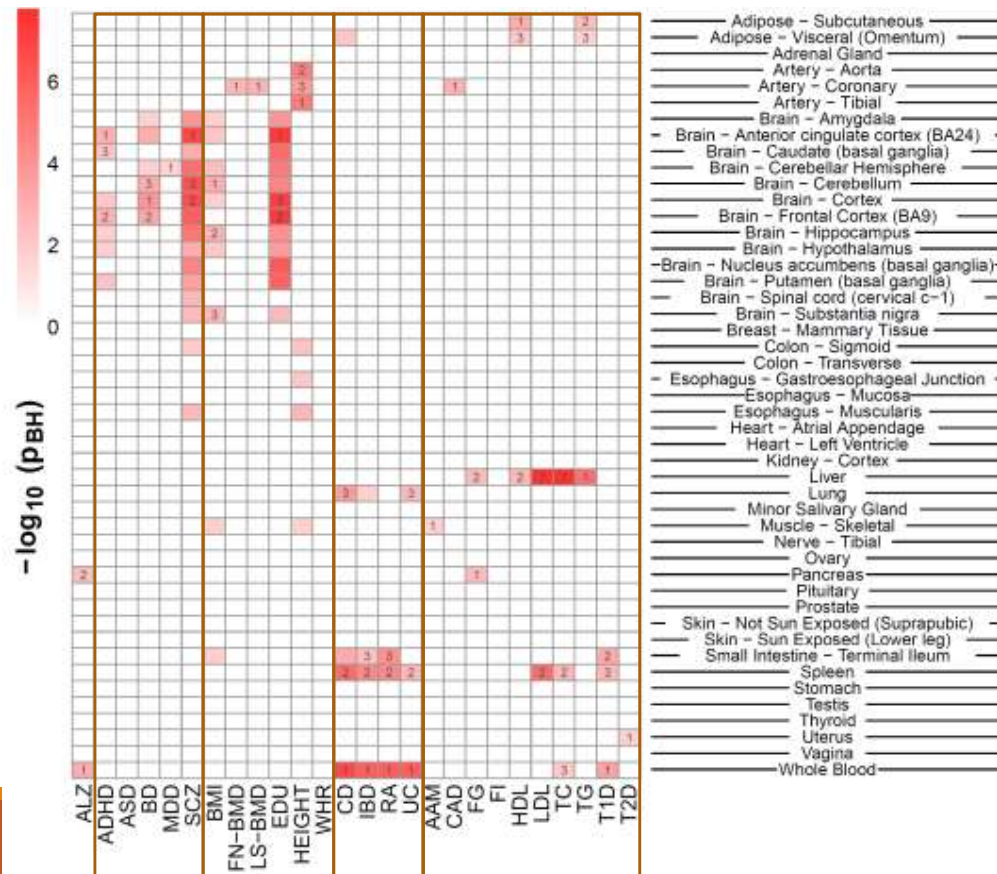
Different immune diseases implicate different immune cell-types!



Systemic Lupus Erythematosus  
- Transitional B-cells!



# Application: multi-trait GWAS summary statistics



Trait	Trait Full Name
ALZ	Alzheimer's disease
ADHD	Attention deficit-hyperactivity disorder
ASD	Autism spectrum disorder
BD	Bipolar disorder
MDD	Major depressive disorder
SCZ	Schizophrenia
BMI	Body mass index
FN-BMD	Bone mineral density (femoral neck)
LS-BMD	Bone mineral density (lumbar spine)
EDU	Educational attainment
HEIGHT	Height
WHR	Waist-hip ratio
CD	Crohn's disease
IBD	Inflammatory bowel disease
RA	Rheumatoid arthritis
UC	Ulcerative colitis
AAM	Age at menarche
CAD	Coronary artery disease
FG	Fasting glucose
FI	Fasting insulin
HDL	High-density lipoproteins

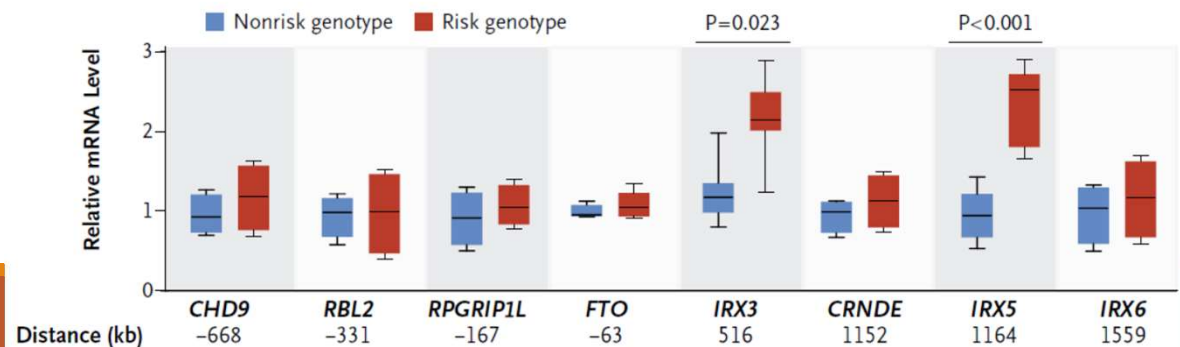
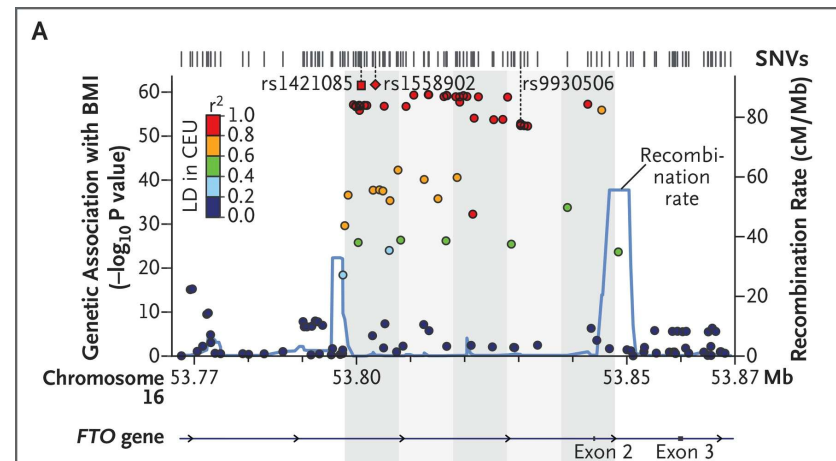
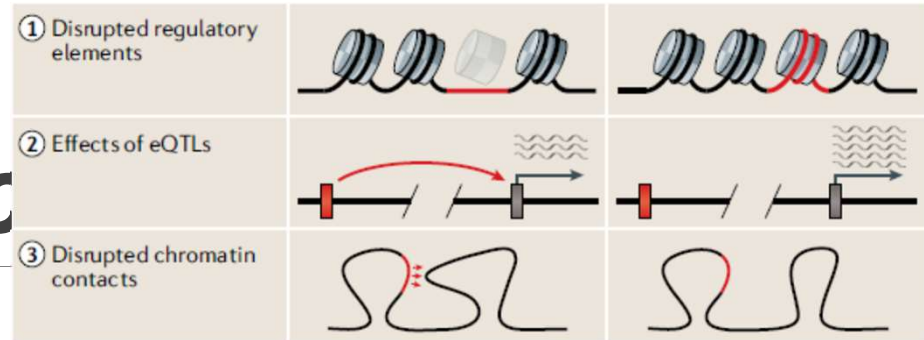


## 2. Regulatory mechanisms

### Disruption of gene-specific regulatory elements

- Obesity-associated locus *FTO*:
- Analysis of the chromatin state of *FTO* across 127 human cell types revealed that it harbors an enhancer that is specific to pre-adipocyte cells.
- The rs1421085 T-to-C disrupts a conserved motif for the ARID5B repressor, which leads to derepression of a potent preadipocyte enhancer and a doubling of *IRX3* and expression during early adipocyte differentiation.

#### b Regulatory mechanisms



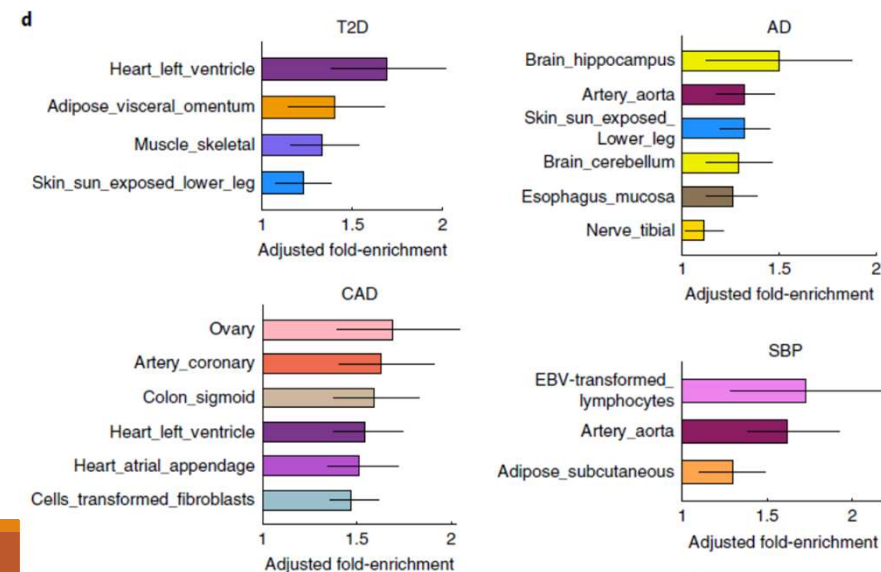
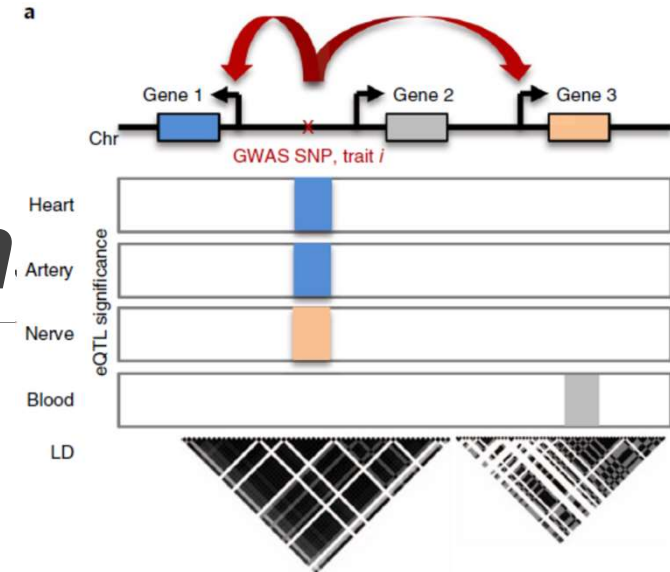
## 2. Regulatory mechanism

### Effects of eQTLs

- A study testing enrichment for 18 complex traits using 44 tissue eQTLs
- Example results: aortic artery for systolic blood pressure (SBP), coronary artery for coronary artery disease (CAD), skeletal muscle for type 2 diabetes (T2D), colon for Crohn's disease, and hippocampus for Alzheimer's disease

### Disruptions of chromatin contacts

- Genetic aberrations that disrupt TADs were shown to rewire enhancer–promoter interactions and lead to the misexpression of genes, resulting in pathogenic phenotypes such as limb malformations<sup>88</sup>



# 3. Tissue-disrupted networks

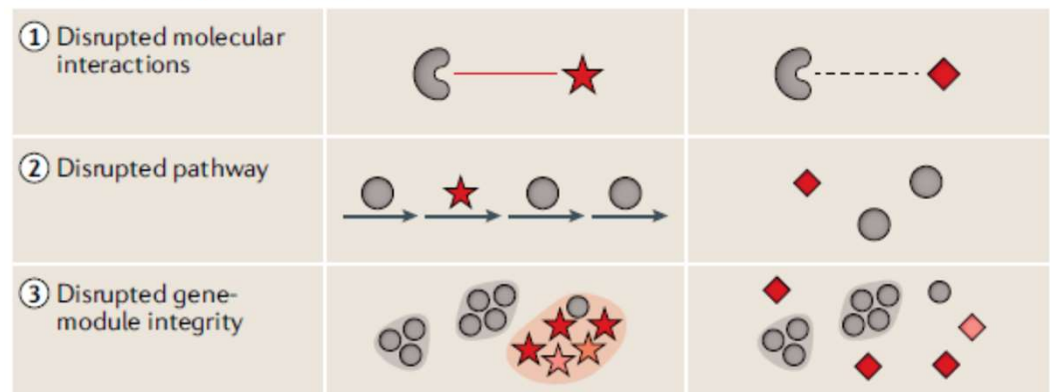
## Disrupted tissue-specific molecular interactions

- Disease-causing genes have a significantly higher tendency for tissue-specific interactions in their disease tissues
- A large-scale yeast two-hybrid analysis of 220 causal genes revealed that 61% of the disease-causing alleles exhibited partial loss of their wild-type PPIs

## Disrupted tissue-specific pathways and gene modules

- Disease module connectivity was significantly high particularly in disease-related tissues
- A study of 37 GWAS traits showed that risk alleles often perturb regulatory gene modules that were highly specific to disease-relevant tissues

**C** Tissue-disrupted networks



Pathways are biological processes performed by subsets of gene products and other molecules, and may involve metabolic reactions, as in glycolysis, or regulatory and protein interactions, as in signaling cascades.

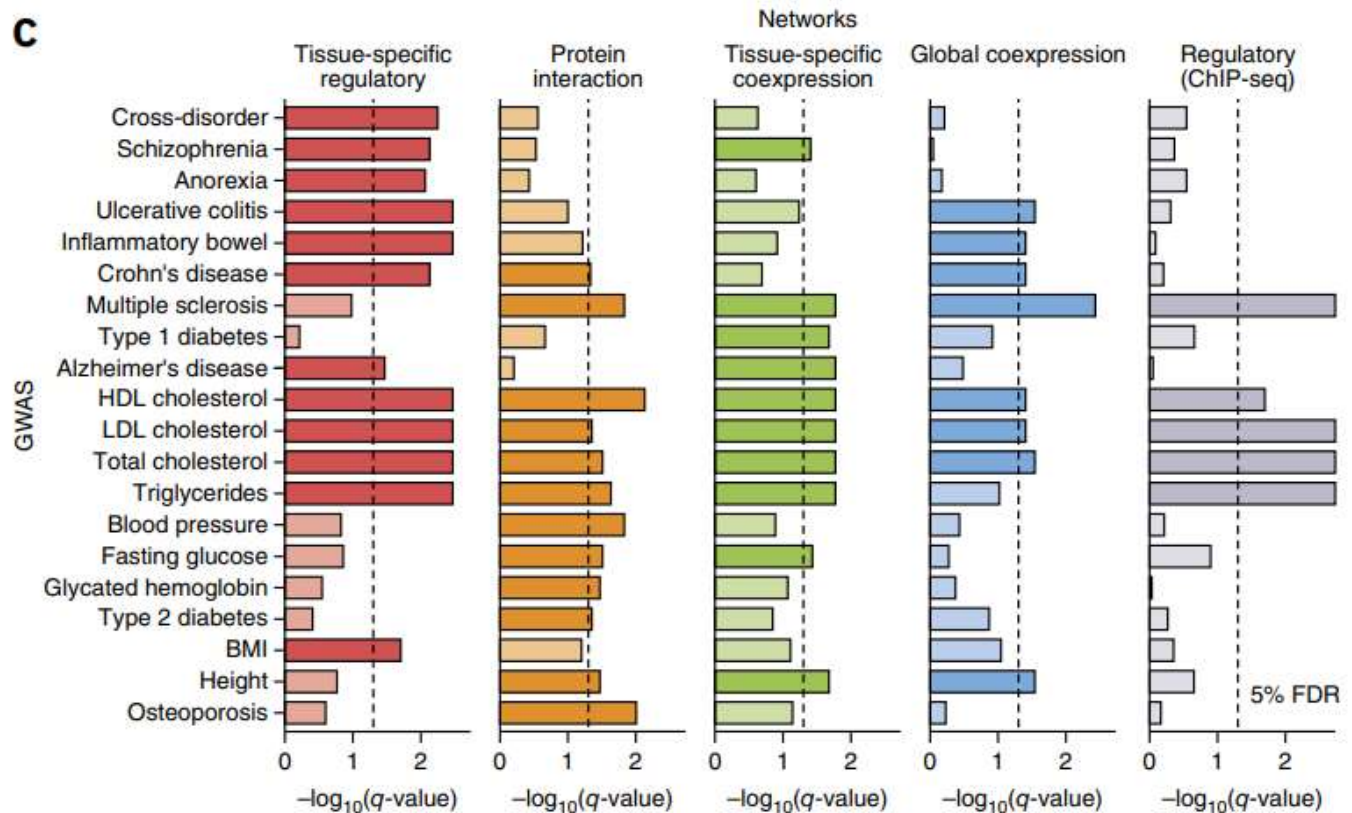
E.g., the glycogenolysis pathway, whereby the polysaccharide glycogen is degraded, takes place mainly in liver and muscle, and thus glycogen storage diseases manifest primarily in those tissues.

# Trait-associated genes tend to cluster in modules for different types of networks and GWAS traits

Five types of networks are compared

- cell type- and tissue-specific regulatory networks
- protein-protein interaction networks
- tissue-specific coexpression networks
- a global coexpression network
- a global regulatory network based on ChIP-seq

**Tissue-specific regulatory networks showed the strongest connectivity enrichment overall.**



# ***4. Non-cell-autonomous mechanisms***

---

*Tissue-specific responsiveness to signals*

*Tissue-specific microenvironment*

- Cells continuously interact with their microenvironment by cell–cell physical contacts and by processing chemical signals in the form of diffusible molecules.
- E.g., blood–brain barrier leakage was observed in patients with Alzheimer disease
- Tissue microenvironment is also critical for therapy.

# A schematic flow chart illuminating tissue-selectivity mechanisms

