



Gene- and pathway-levels association analyses

Miaoxin Li

Zhongshan School of Medicine, Sun Yat-Sen University, Guangzhou

Email: limiaoxin@mail.sysu.edu.cn

Homepage: <http://pmglab.top/>



Outline

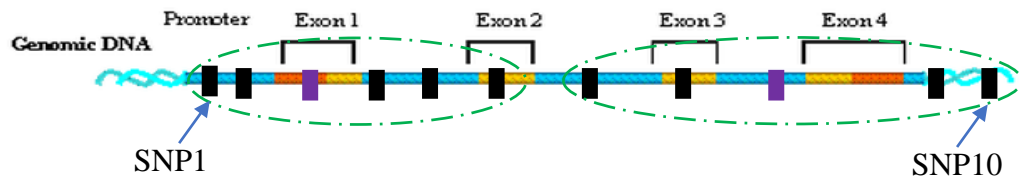
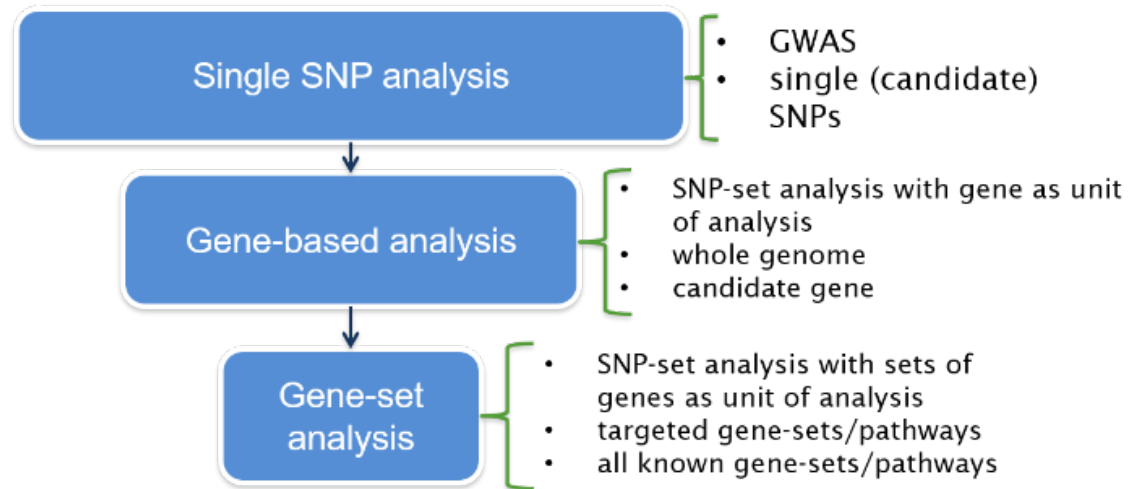
- Gene-based association analyses
- Conditional gene-based association analyses
- Gene pair-based association analyses
- Gene-set based association analyses
- Practical on KGGSEE

Limitation of variant level analysis in GWAS

- Low statistical power (due to multiple testing $p \leq 5E-8$)
- Large heterogeneity (allele frequency different in different populations)
- Hard to link biological functions (variants are not well-annotated)

Alternative strategies

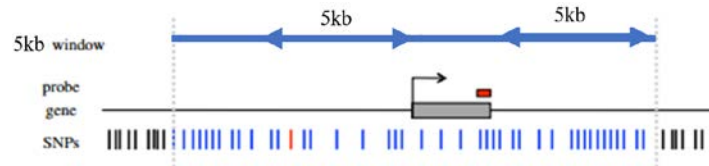
- Test OVERALL genetic associations of multiple SNPs in a single gene or gene-set



Advantage's of gene level analysis in GWAS

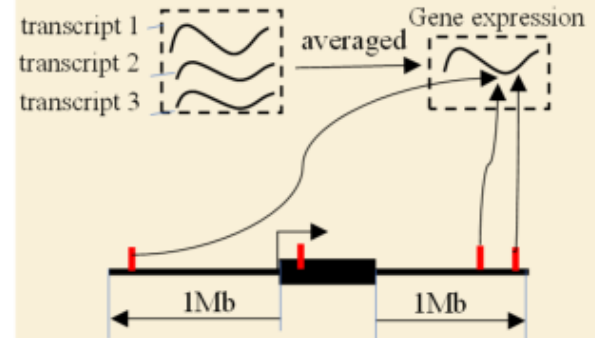
- Improved statistical power (Fewer multiple testing $p \leq 2.5E-6$)
- Robust to population heterogeneity (More stable than allele frequencies in different populations)
- Easier to link biological functions (Genes are basic function unit of genome)

Map SNPs to genes

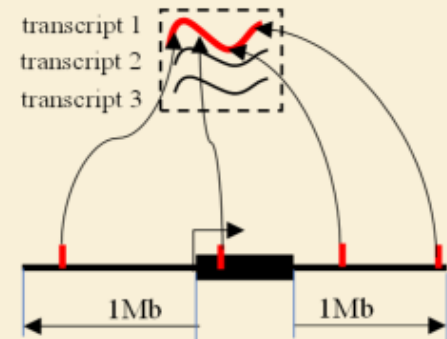


The three strategies of mapping SNPs to genes

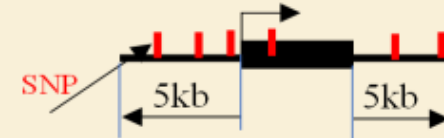
Gene-level eQTLs



Isoform-level eQTLs

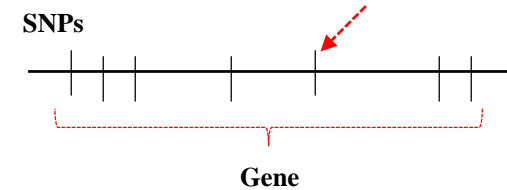


Physical distance

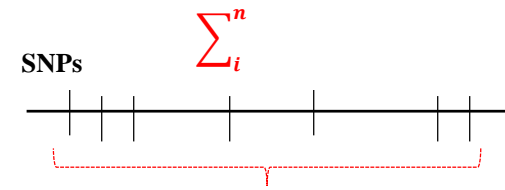


Ways of combining SNP's association signals in a gene

- Perform multiple testing of SNPs for a gene (-multiple testing based combination)---a single representative SNP



- Accumulate association signals at all SNPs of a gene (-sum based combination)---the averaged significance of SNPs



Multiple testing based combination

First of all, sort p-values of SNPs in a gene: $P_{(1)}, \dots, P_{(i)}, \dots, P_{(m)}$

- Bonferroni correction:

$$P_G = mP_{(1)}$$

- Sida correction:

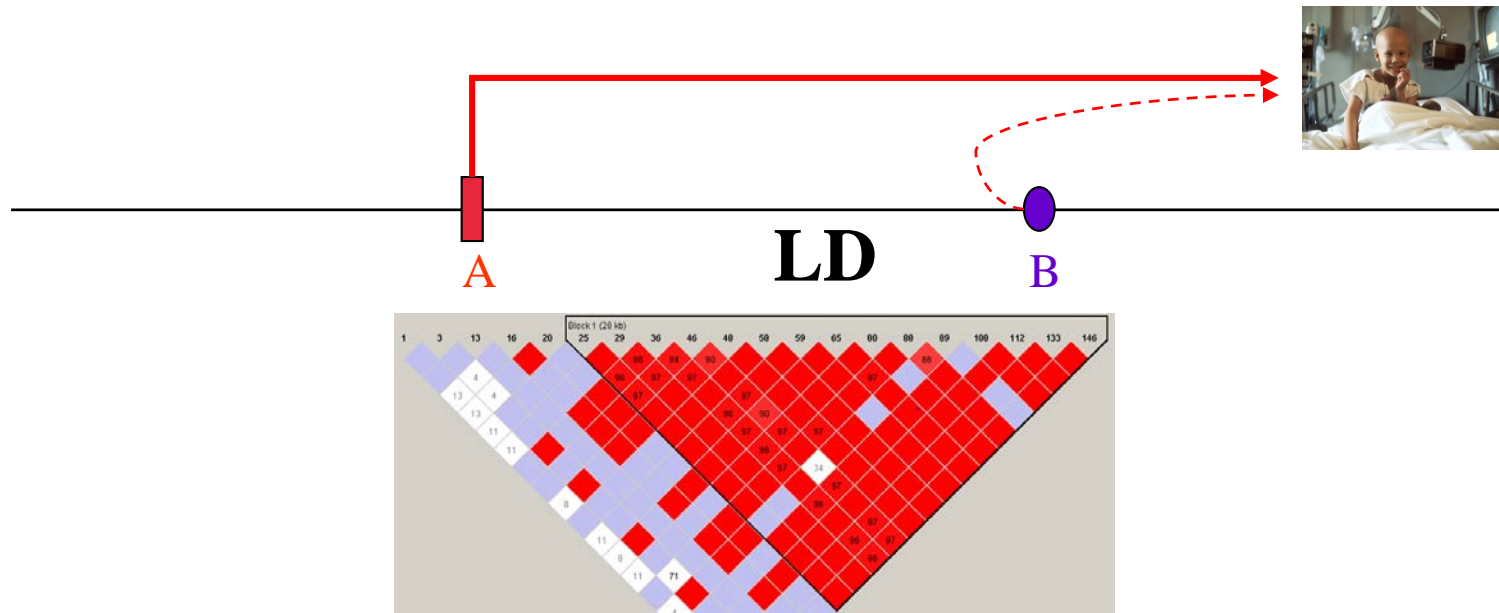
$$P_G = 1 - (1 - P_{(1)})^m$$

- Simes correction

$$P_G = \min\left\{\frac{mP_{(i)}}{i}\right\}$$

Problems in combination

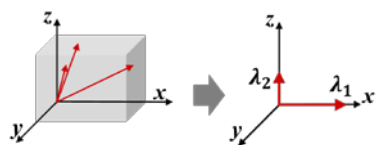
- Redundant association due to linkage disequilibrium
- **Over-correction** in multiple testing



A new measure of the effective number of independent tests

The formula:

<https://pmglab.top/gec>



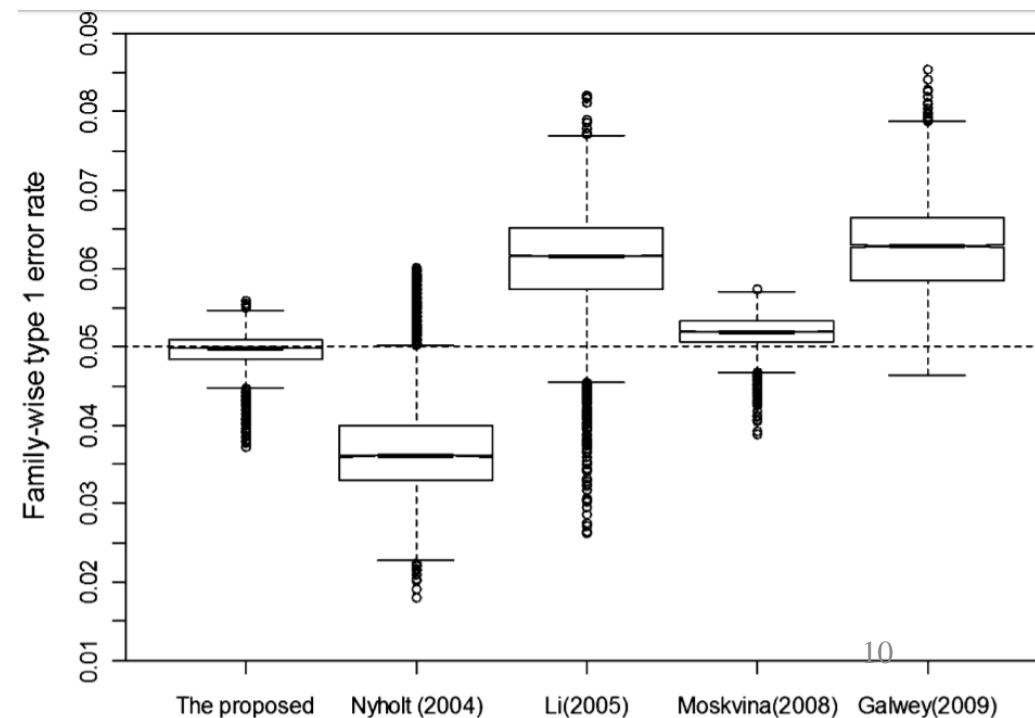
$$m_e = \sum_{i=1}^M [I(\lambda_i \geq 1) + \lambda_i I(\lambda_i < 1)] = M - \sum_{i=1}^M [I(\lambda_i > 1)(\lambda_i - 1)]$$

Hum Genet.
2012;131(5):747-56
(>400 citations)

where $I(x)$ is an indicator function and λ_i is the i th Eigenvalue of the p-value correlation coefficient matrix

Bonferroni correction: α/m_e

Box plot of MVN derived FWERs for five different methods. For each method, the nominal FWER was set to be 0.05. The bottom and top of each box mark the 25th and 75th percentile, respectively, and the band in the box denotes the 50th percentile. The lines above and below each box denote the upper and lower 1.5 interquartile range (IQR)

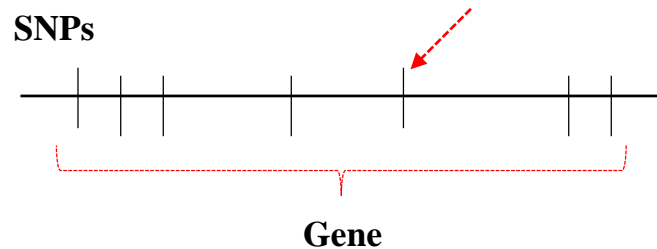


Gene-based association analysis by GATES

- Extended Simes' test (known as GATES)

$$P_S = \min\left(\frac{mP_{(j)}}{j}\right) \Rightarrow P_S = \min\left(\frac{m_e P_{(j)}}{\sum_{k=1}^j w_{(k)}}\right) \quad \text{where}$$

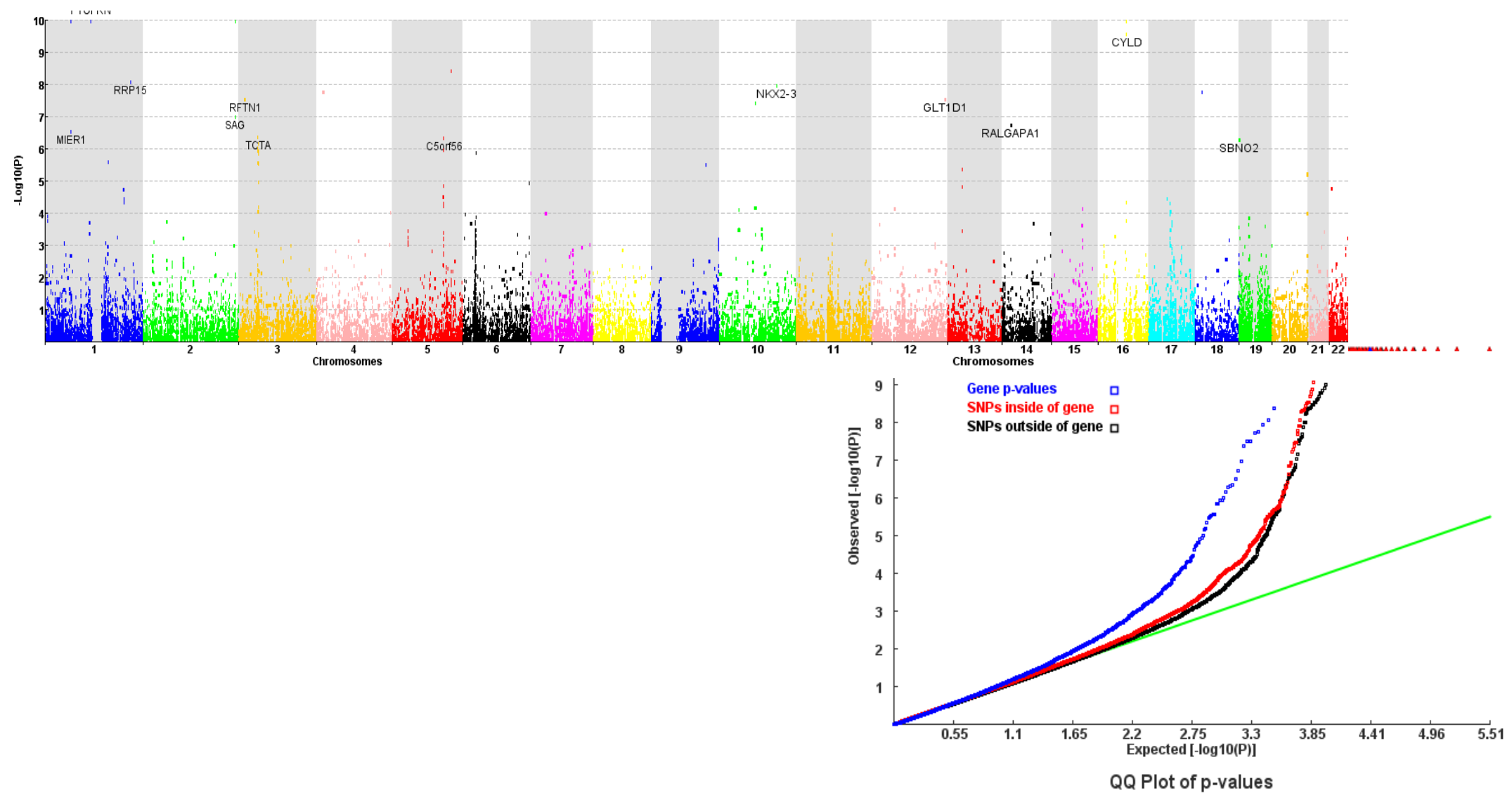
$P_{(j)}$ = sorted SNP p value
 m_e = effective number of test
 $w_{(k)}$ = weight for SNP j



Desirable properties

- ✓ Dependent p values
- ✓ Not resort to permutation and simulation

Gene-based association test by GATES Crohn's Disease GWAS data set



Simulation: type 1 error and power comparison

	#SNP (#DSL)	Logistic Regression	Original Simes	VEGAS – Max	GATES
Moderate LD					
Error Rate (no disease)	3(0)	4.86	4.54	4.81	4.98
	10(0)	4.88	4.55	4.92	5.00
	30(0)	5.63	4.97	5.29	5.56
Power (additive model)	3(1)	44.59	49.71	50.51	51.23
	10(2)	56.25	58.93	59.12	60.72
	30(6)	65.47	53.29	52.24	55.65
Power (multiplicative model)	3(1)	46.52	51.19	52.00	52.65
	10(2)	68.42	70.66	70.9	72.4
	30(6)	93.68	86.07	84.34	87.52
Strong LD					
Error Rate (no disease)	3(0)	4.96	3.88	5.22	5.35
	10(0)	5.33	3.37	4.88	5.34
	30(0)	5.57	3.38	4.89	5.64
Power (additive model)	3(1)	45.03	53.88	58.2	60.43
	10(2)	57.2	66.39	71.71	74.3
	30(6)	65.56	62.84	66.80	72.75
Power (multiplicative model)	3(1)	47.13	56.28	60.74	62.77
	10(2)	68.45	77.14	80.59	83.00
	30(6)	93.4	91.42	92.24	95.38
Data are given as percentages. Abbreviations are as follows: LE, linkage equilibrium; LD, linkage disequilibrium; and DSL, disease susceptibility loci.					

Li et al. Am J Hum Genet
. 2011 Mar 11;88(3):283-93. doi:
10.1016/j.ajhg.2011.01.019

$$\ln \left(\frac{P(y = 1)}{1 - P(y = 1)} \right) = \log \left(\frac{P(y = 1)}{P(y = 0)} \right) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

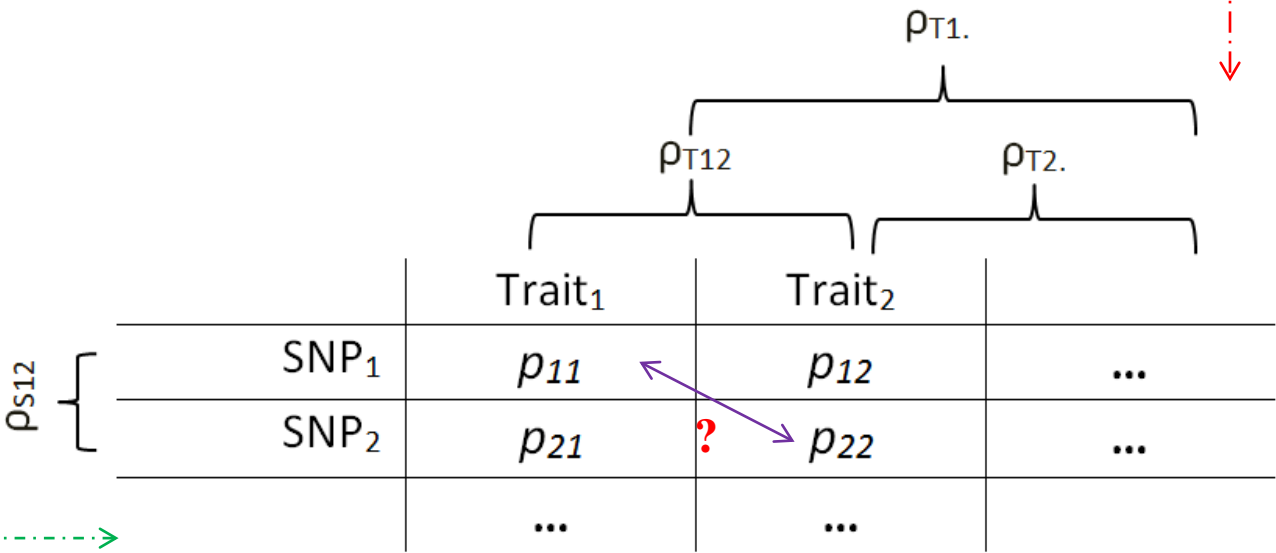
MGAS: multivariate gene-based association test by extended Simes procedure

Bioinformatics, 31(7), 2015, 1007–1015
doi: 10.1093/bioinformatics/btu783
Advance Access Publication Date: 26 November 2014
Original Paper

Genetics and population analysis

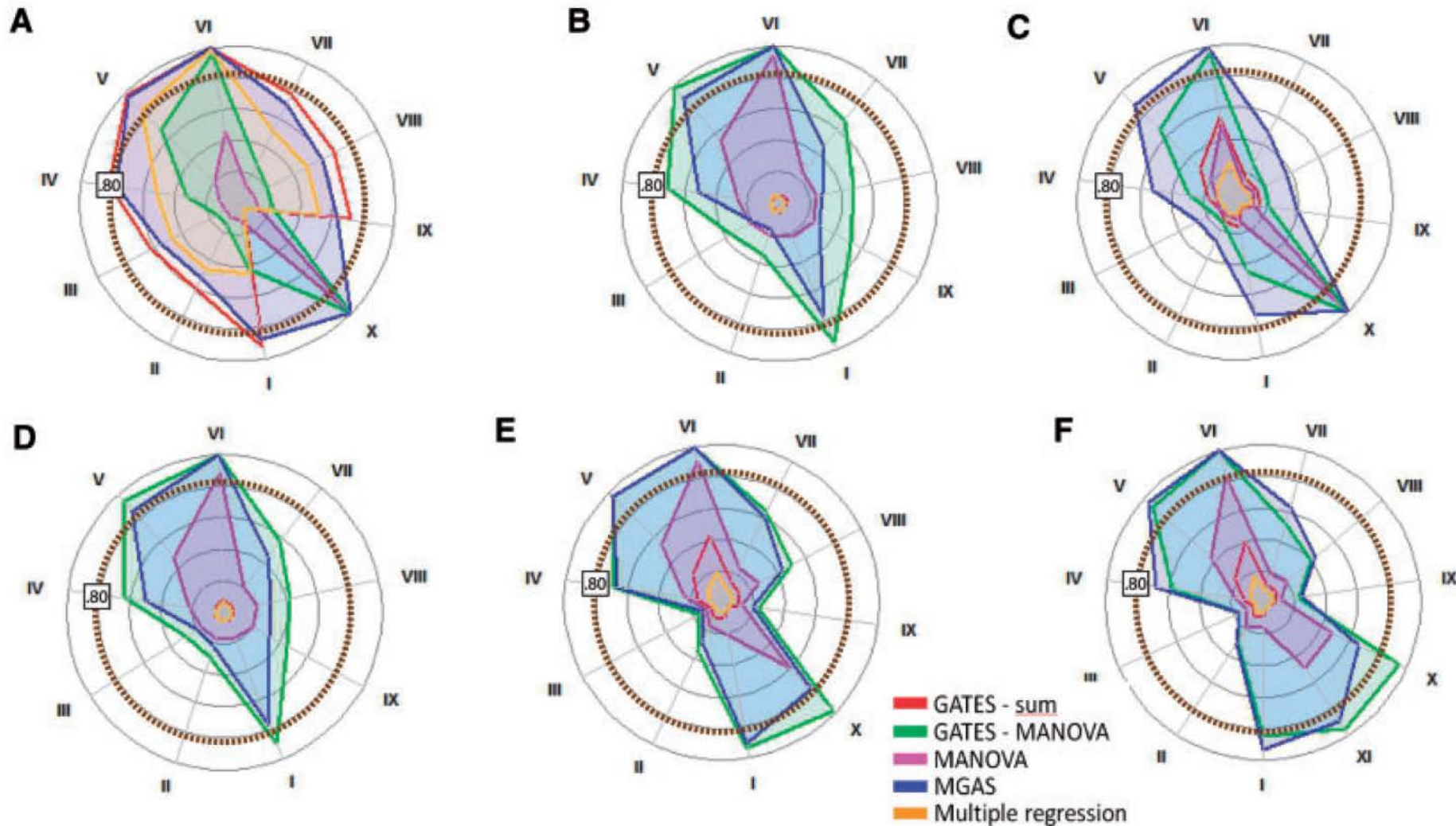
MGAS: a powerful tool for multivariate gene-based genome-wide association analysis

Sophie Van der Sluis¹, Conor V. Dolan², Jiang Li³, Youqiang Song^{3,4,5,6}, Pak Sham^{4,5,6,7}, Danielle Posthuma^{1,8} and Miao-Xin Li^{4,5,6,7,*}

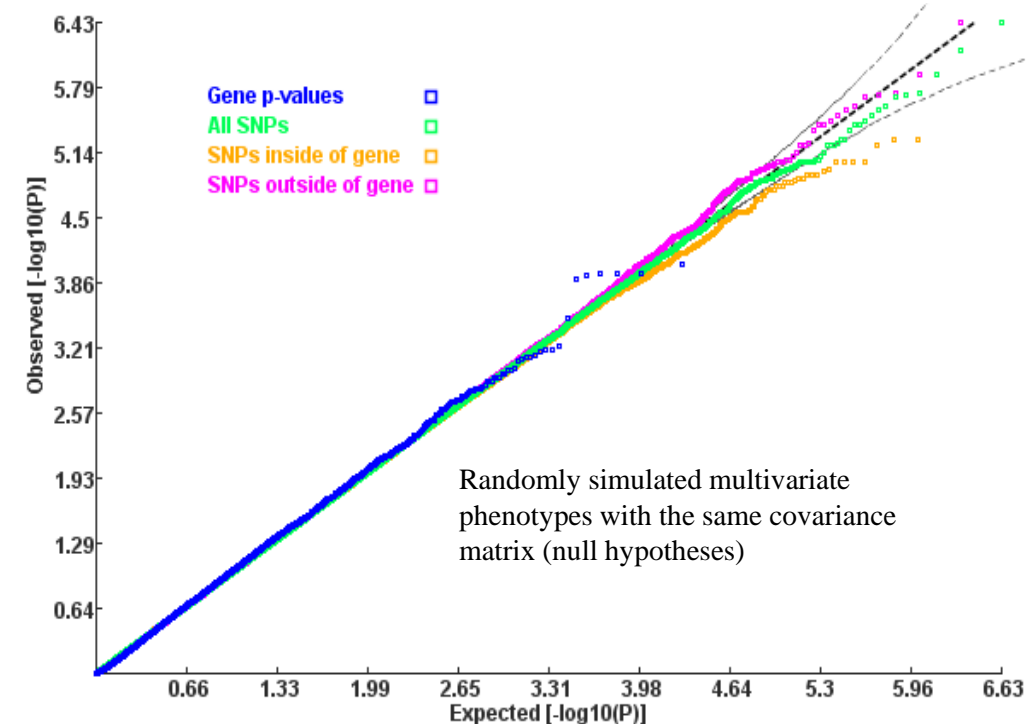
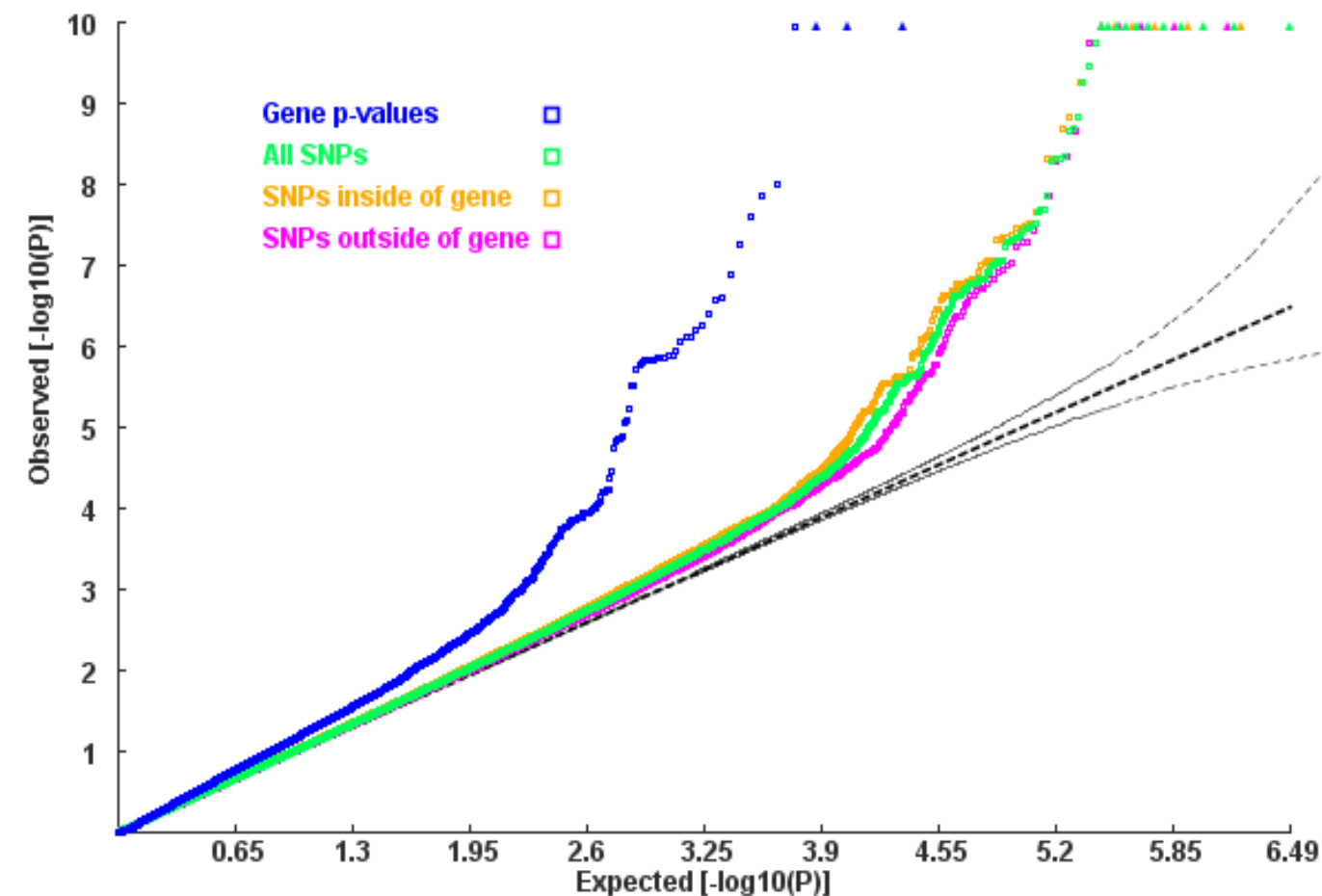


$$P_{\text{MGAS}} = \min \left(\frac{q_e p_j}{q_{ej}} \right)$$

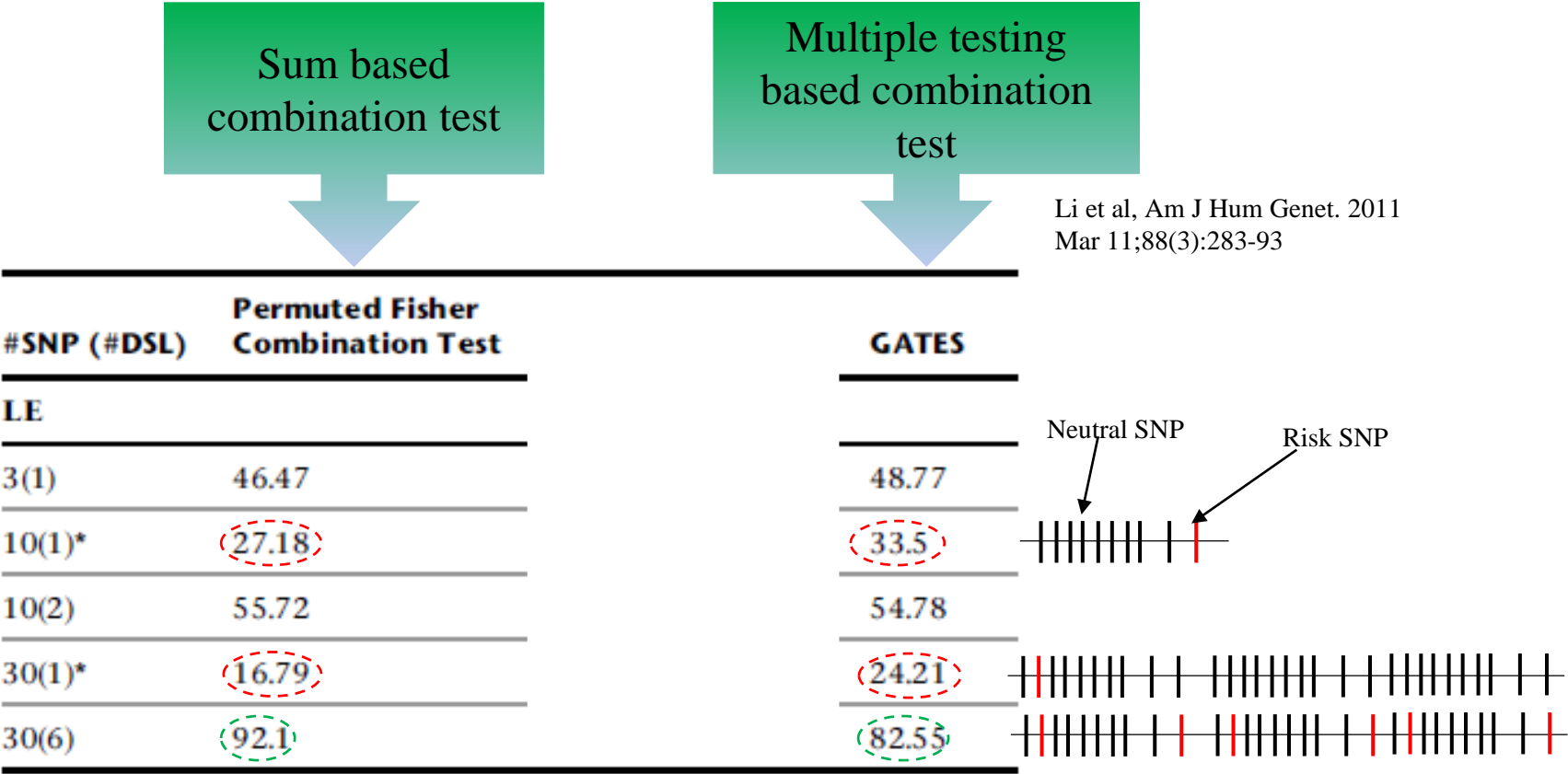
Power plots for six trait-generating genotype–phenotype models



Multivariate gene-based test for Nine quantitative metabolic traits



Sometimes, multiple testing based combination test is under powered!



Sum based combination

- Fisher's combination method combines extreme value probabilities from each test

$$X_{2k}^2 \sim -2 \sum_{i=1}^k \log(p_i) \quad , \text{where } p_i \text{ is a p value at } i\text{th SNP.}$$

- Weighted z-tests

$$p_Z = 1 - \Phi \left(\frac{\sum_{i=1}^k w_i Z_i}{\sqrt{\sum_{i=1}^k w_i^2}} \right)$$

where Z_i is a converted quantile from p_i under normal distribution

Inflated
type 1
error due
to LD

	#SNP (#DSL)	Logistic Regression	Fisher
Moderate LD			
Error Rate (no disease)	3(0)	4.86	7.17
	10(0)	4.88	9.8
	30(0)	5.63	11.09
Power (additive model)	3(1)	44.59	55.8
	10(2)	56.25	72.38
	30(6)	65.47	83.04
Power (multiplicative model)	3(1)	46.52	57.5
	10(2)	68.42	81.73
	30(6)	93.68	98.04
Strong LD			
Error Rate (no disease)	3(0)	4.96	11.49
	10(0)	5.33	15.68
	30(0)	5.57	17.9
Power (additive model)	3(1)	45.03	72.29
	10(2)	57.2	89.82
	30(6)	65.56	96.04
Power (multiplicative model)	3(1)	47.13	74.28
	10(2)	68.45	94.41
	30(6)	93.4	99.92

Li et al. Am J Hum Genet
. 2011 Mar 11;88(3):283-93. doi:
10.1016/j.ajhg.2011.01.019

Combine **correlated** chi-square statistics

Approximation of summation of correlated chi-square statistics

- Weighted Fisher's combination statistic (MAGMA)

$$X = \sum_{i=1}^k w_i (-2 \log P_i) \quad \text{Statistics \& Probability Letters 73(2),2005, P179-187}$$

- Simulation based on multi-variate normal distribution (VEGAS)

$$Q = (q_1, q_2 \dots q_n), \quad q_i = z_i^2 \quad Z = (z_1, z_2 \dots z_n)$$

$$Z \sim N_n(0, \Sigma)$$

Am J Hum Genet. 2010 Jul 9; 87(1): 139–145

Quadratic form of normal variables

- Test statistics

$$T = \sum_i^N Z_i^2 = \mathbf{Z}^T \mathbf{Z}$$

where N is the number of SNPs mapped in a gene

- And T has the distribution

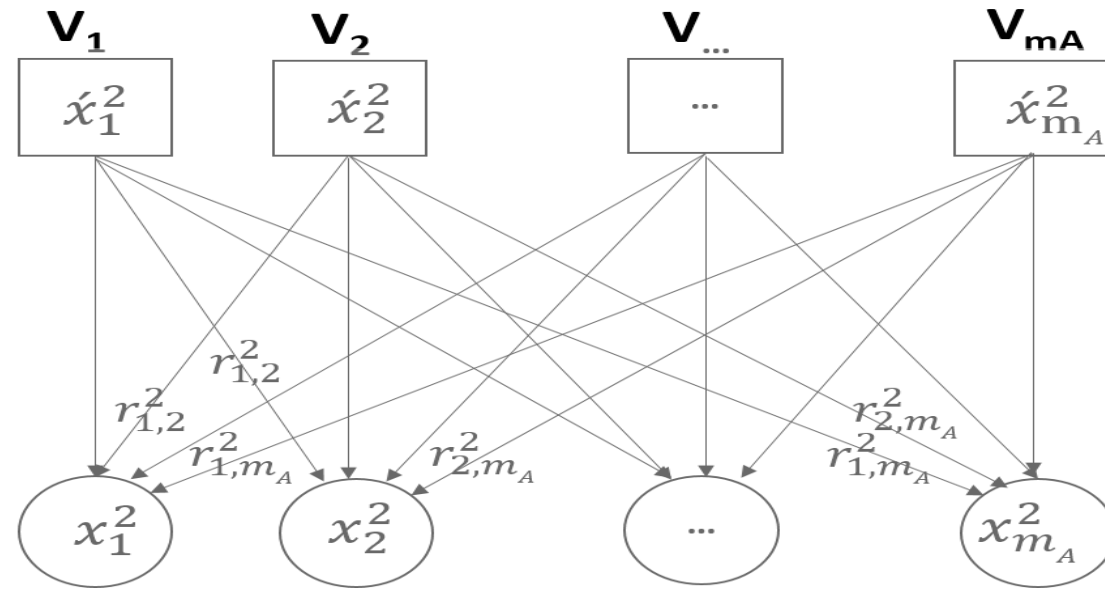
$$T = \mathbf{Z}^T \mathbf{Z} = (\mathbf{Q} \mathbf{A}^{0.5} \mathbf{D})^T \mathbf{Q} \mathbf{A}^{0.5} \mathbf{D} = \mathbf{D}^T \mathbf{A} \mathbf{D} = \sum_i^N \lambda_i D_i^2$$

with $D_i \sim N(0, 1)$ and $D_i^2 \sim \chi_1^2$

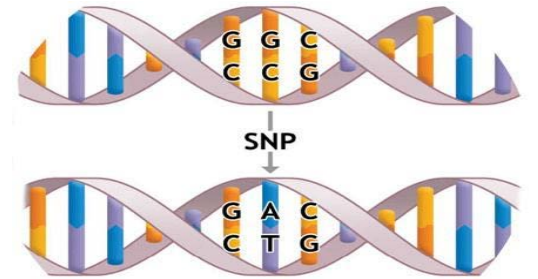
\mathbf{Q} is an orthogonal matrix and $\mathbf{A} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_N)$

ECS: effective chi-squared statistic for gene-based association

Hidden independent chi-square statistics



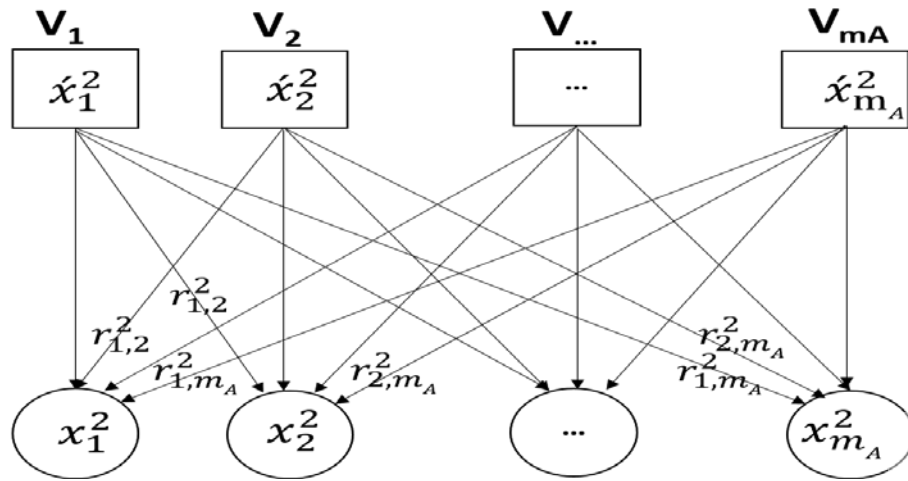
Observed chi-square statistics



$$x_1^2 = 1 * \acute{x}_1^2 + r_{1,2}^2 * \acute{x}_2^2 + \dots + r_{1,m_A}^2 * \acute{x}_{m_A}^2$$

$$\begin{bmatrix} 1 & r_{1,2}^2 & \dots & r_{1,m_A}^2 \\ r_{2,1}^2 & 1 & \dots & r_{2,m_A}^2 \\ \vdots & \vdots & \ddots & \vdots \\ r_{m_A,1}^2 & r_{m_A,2}^2 & \dots & 1 \end{bmatrix} * \begin{bmatrix} \acute{x}_1^2 \\ \acute{x}_2^2 \\ \vdots \\ \acute{x}_{m_A}^2 \end{bmatrix} = \begin{bmatrix} x_1^2 \\ x_2^2 \\ \vdots \\ x_{m_A}^2 \end{bmatrix}$$

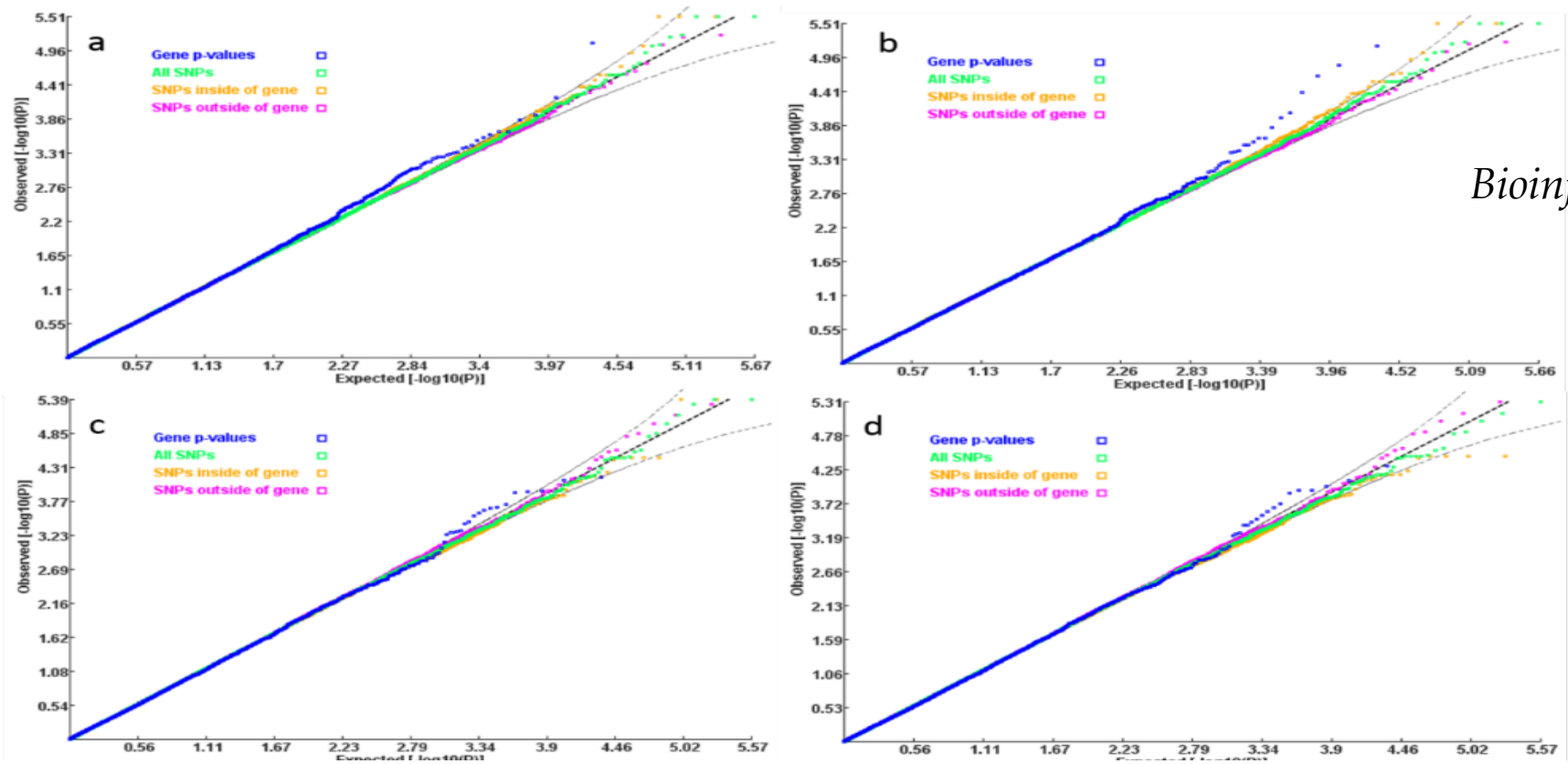
ECS: effective chi-squared statistic for gene-based association



$$\begin{bmatrix} \acute{x}_1^2 \\ \acute{x}_2^2 \\ \dots \\ \acute{x}_{m_A}^2 \end{bmatrix} = \begin{bmatrix} 1 & r_{1,2}^2 & \dots & r_{1,m_A}^2 \\ r_{2,1}^2 & 1 & \dots & r_{2,m_A}^2 \\ \vdots & \vdots & \ddots & \vdots \\ r_{m_A,1}^2 & r_{m_A,2}^2 & \dots & 1 \end{bmatrix}^{-1} * \begin{bmatrix} x_1^2 \\ x_2^2 \\ \dots \\ x_{m_A}^2 \end{bmatrix}$$

$$\acute{S}_A = \sum_{i=1}^{m_A} \acute{x}_i^2 \quad \acute{S}_A \sim \chi_{d_A}^2$$

Simulation: Reasonable type 1 error rates and power



Bioinformatics 2019 Feb 15;35(4):628-635.

#QTL	Cutoff of p	Effective chi-square		VEGAS-sum		Linear regression		SKAT	
		0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01
1	LOC64713(#5)	0.9948	0.9754	0.9934	0.9702	0.9858	0.9446	0.9904	0.9512
	SIPA1L2(#29)	0.9494	0.8532	0.9204	0.7852	0.8346	0.6355	0.8568	0.7048
2	LOC64713(#5)	0.9011	0.8052	0.8741	0.7602	0.8741	0.7722	0.8640	0.7230
	SIPA1L2(#29)	0.9271	0.8581	0.9201	0.8322	0.8581	0.7023	0.8950	0.7940

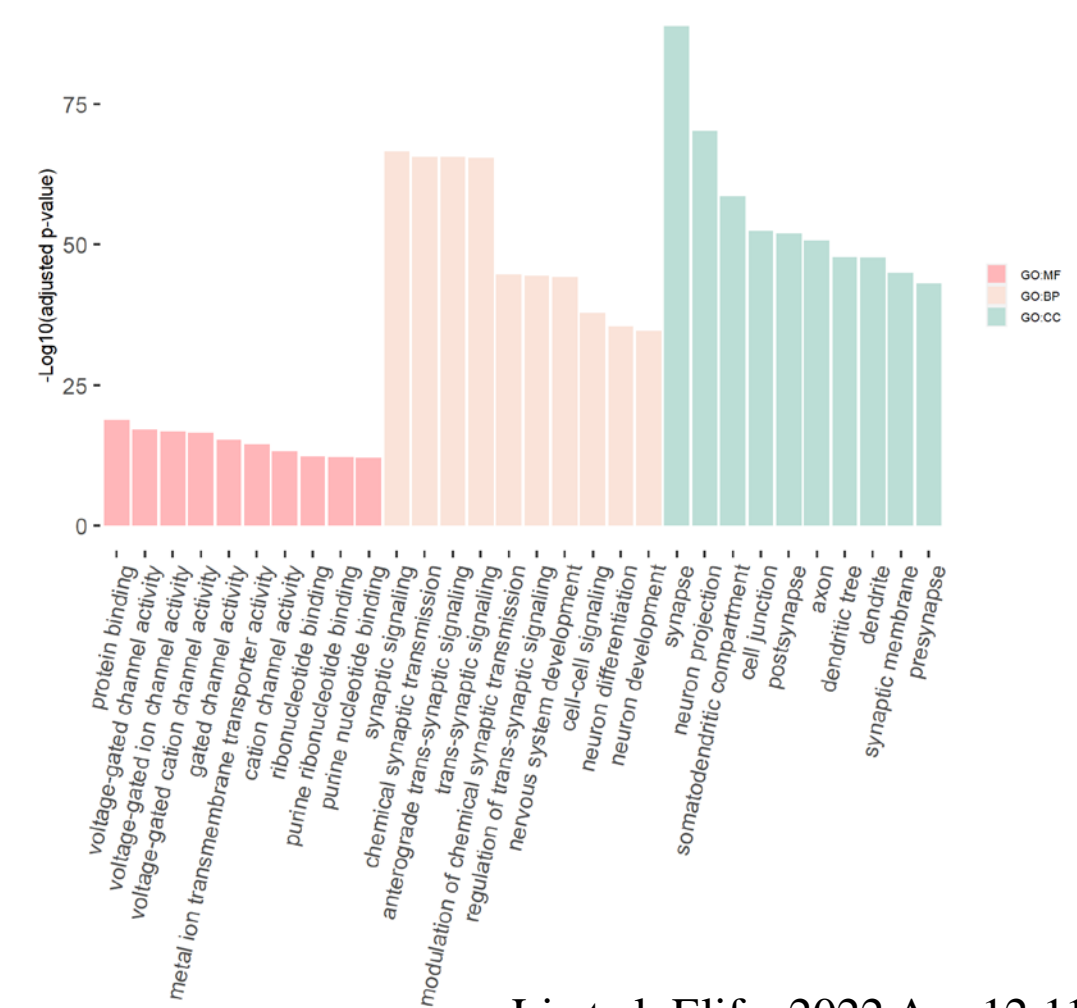
In addition, tests based on transcript-level eQTLs are more powerful than gene-level eQTLs!

Simulation

- Type I error < 0.1%;
- Power: transcript-level eQTLs > gene-level eQTLs

Scenarios	Important Parameters			Binary Trait					Continuous Trait				
	*Eg	Vgp	Vge	AllVar	IsoeQTL	GeneQTL	Gen3eQTL	Gen6eQTL	AllVar	IsoeQTL	GeneQTL	Gen3eQTL	Gen6eQTL
				Type I error									
1	0	0	0.05	0	0	0	0	0.002	0	0	0.001	0.002	0.003
2	0	0	0.15	0	0.002	0.001	0	0	0	0	0	0.001	0
3	0	0	0.3	0	0	0	0.002	0	0	0.001	0.001	0.002	0.002
				Power									
4	0	0.005	0.05	0.251	0.036	0.022	0.034	0.031	0.246	0.032	0.019	0.032	0.038
5	0	0.005	0.15	0.219	0.021	0.013	0.023	0.032	0.301	0.025	0.017	0.037	0.043
6	0	0.005	0.3	0.229	0.028	0.017	0.021	0.034	0.282	0.024	0.017	0.025	0.039
7	0.1	0	0.05	0	0.017	0.019	0.006	0.001	0	0.017	0.027	0.009	0.001
8	0.1	0	0.15	0	0.213	0.221	0.113	0.054	0.002	0.245	0.245	0.132	0.068
9	0.1	0	0.3	0.018	0.704	0.659	0.581	0.388	0.027	0.72	0.686	0.607	0.446
10	0.1	0.005	0.05	0.288	0.052	0.076	0.043	0.043	0.313	0.063	0.091	0.05	0.041
11	0.1	0.005	0.15	0.403	0.33	0.302	0.199	0.134	0.46	0.357	0.334	0.229	0.136
12	0.1	0.005	0.3	0.569	0.778	0.738	0.677	0.485	0.62	0.805	0.774	0.712	0.512

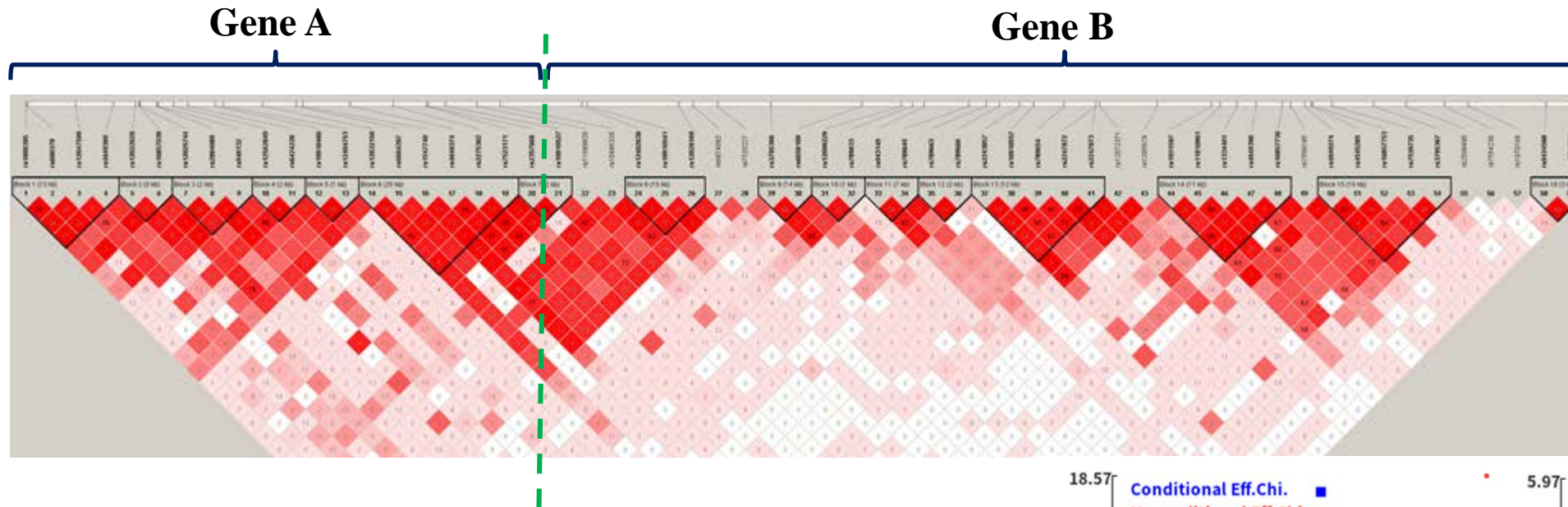
Gene-based association analysis with transcript-level eQTLs revealed many susceptibility genes of schizophrenia!



Gene Name	# of hits in PubMed
RGS4	> 100
TCF4	> 100
RANGAP1	> 100
GRIA1	80
GRM3	76
TSPO	39
TPH2	35
FEZ1	31
ZDHHC8	24
VRK2	23
KCNN3	20
NCAM1	20
MIP	15
SLC39A8	14
DLG1	14
BDNF-AS	13
FGA	13
ADRA1A	12
MAPT	10

138 of the 524
(26.3%)
≥1 PubMed hits

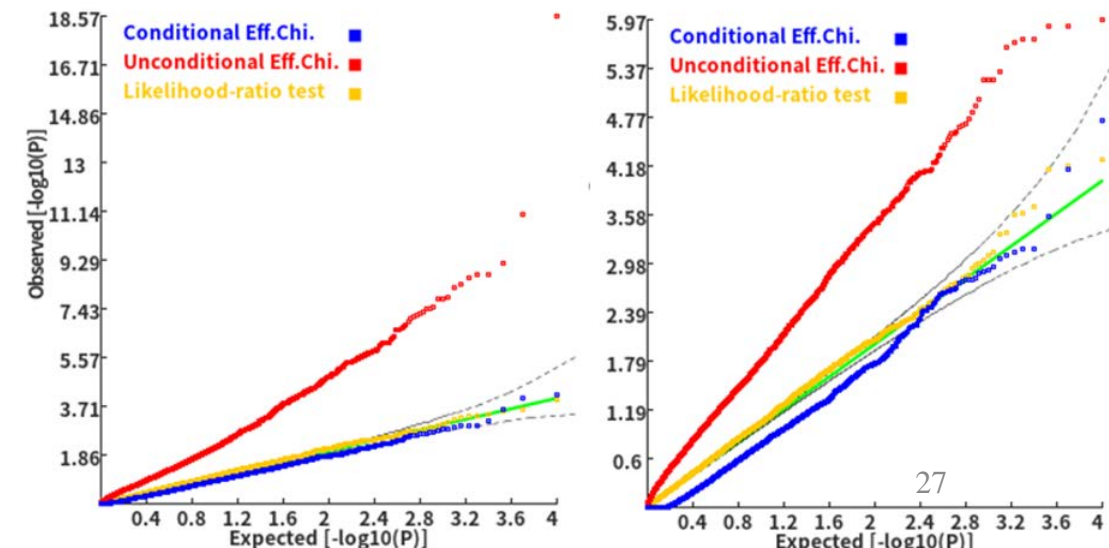
Conditional gene-based association test by ECS



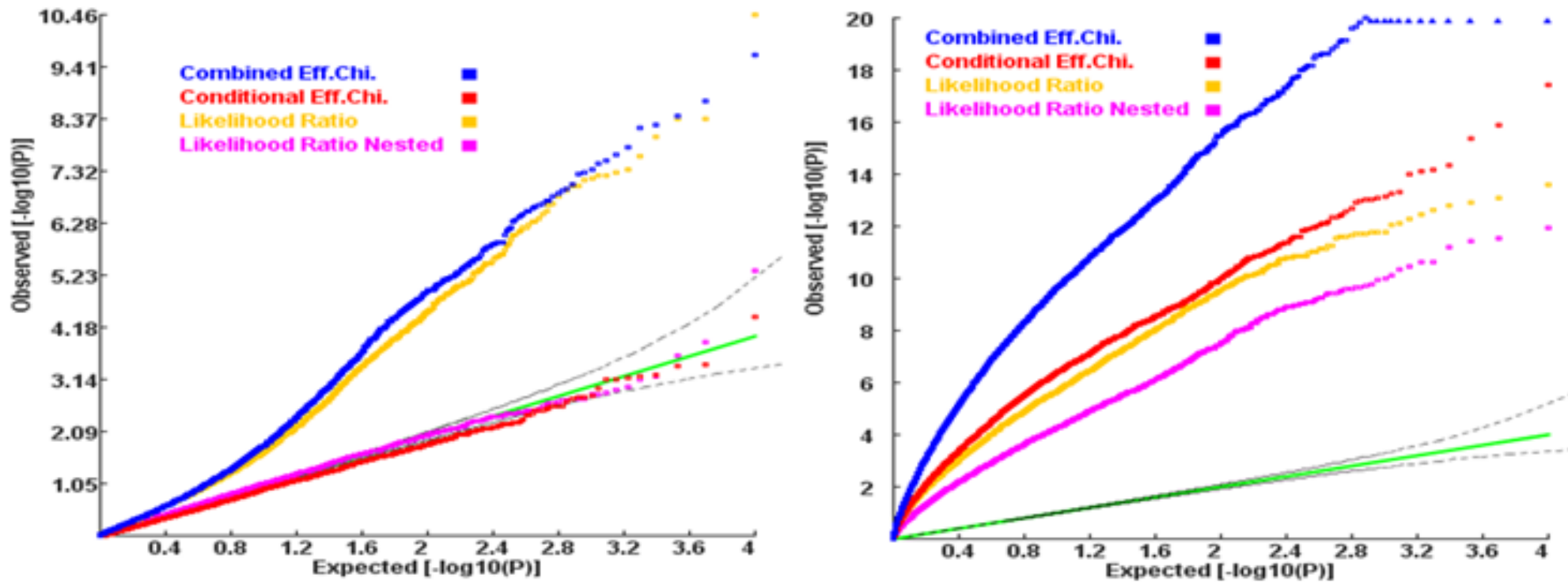
$$\hat{S}_{B|A} = (\hat{S}_{AB} - \hat{S}_A) \sim \chi^2_{d_{AB}-d_A}$$

Do the conditional analysis one by one for multiple genes!

Validation by simulations



Simulation: Improved statistical power while having reasonable type 1 error rate



Applied to meta-analysis dataset of Schizophrenia (PGC2)

- **674** significant genes ($p \leq 1.95 \times 10^{-6}$) → 150 conditionally significant genes (22%)

[nature](#) > [articles](#) > [article](#)

[Published: 22 July 2014](#)

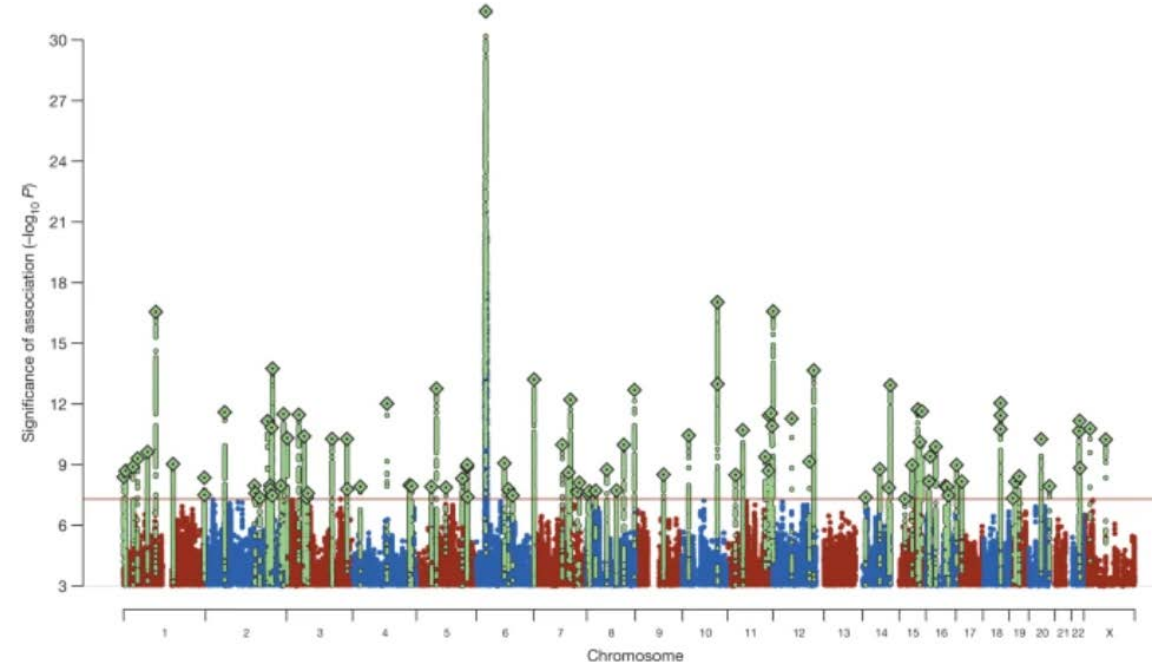
Biological insights from 108 schizophrenia-associated genetic loci

[Schizophrenia Working Group of the Psychiatric Genomics Consortium](#)

[Nature](#) **511**, 421–427 (2014) | [Cite this article](#)

297k Accesses | **4667** Citations | **1109** Altmetric | [Metrics](#)

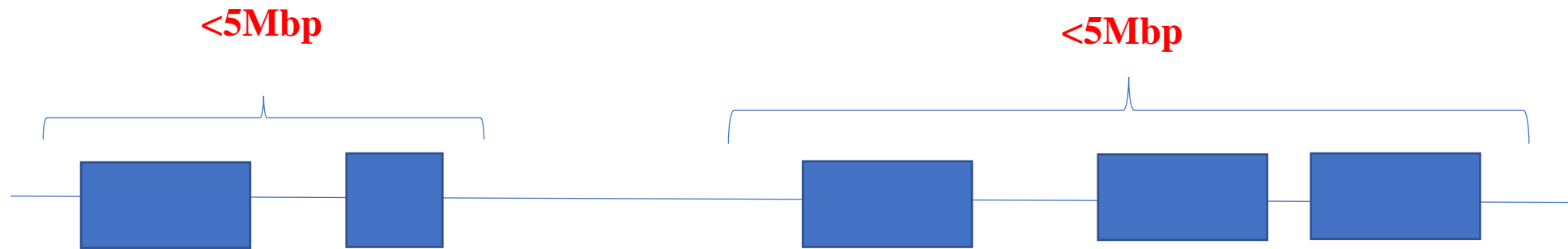
34,241 cases and 45,604 controls



Conditional gene-based association analysis in human genome

Step1: Ordinary gene-based association analysis

Step2: Conditional gene-based association analysis among significant and nearby genes



Which gene should enter the condition procedure with higher priority?

- Genes with smaller p-values
- Genes with larger **functional relevance**

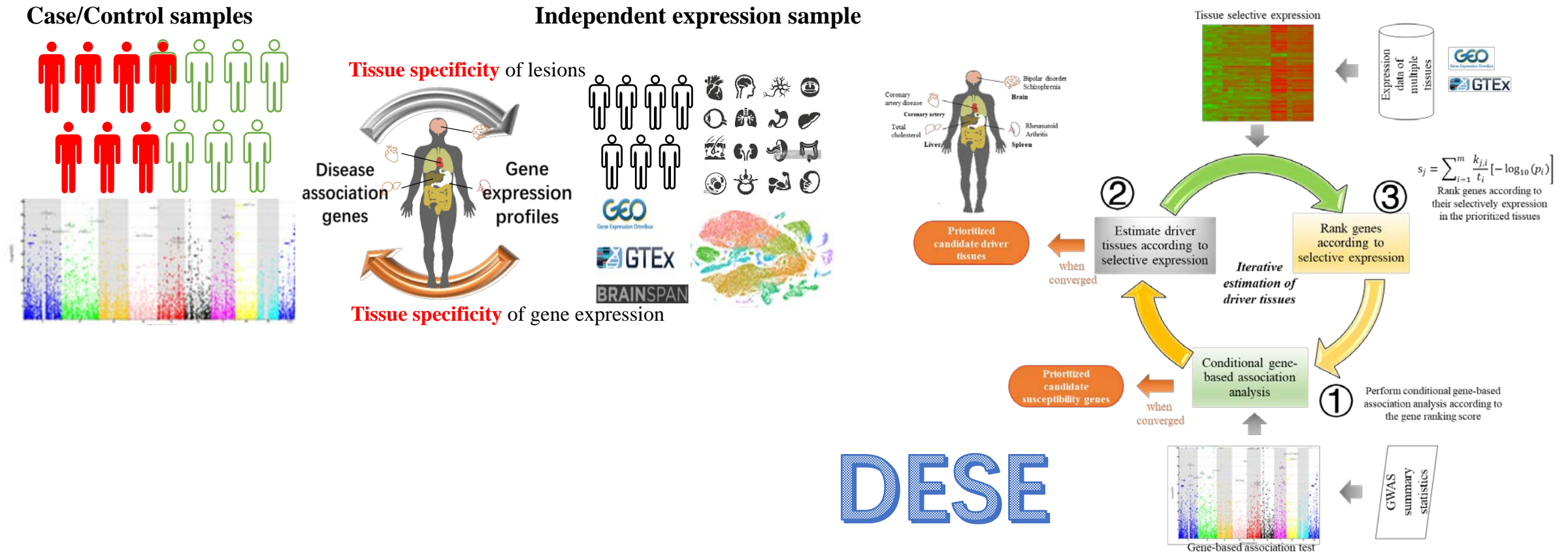
Hypotheses about functional relevance

Causal or susceptibility genes of diseases tend to have **higher selective expression** in disease relevant tissue or cell types.

- Single-gene disorder (✓)
- Complex diseases (?)

Difficulties:

- Disease susceptibility genes are **unknown**
- Disease relevant tissues are **unknown**



Conditional gene-based association analysis with selective expression led to more supported genes!

		SCZ	BD	CAD	RA	Height	TC
DESE	Different genes	32	19	13	118	54	54
	PubMed hits	16	5	11	40	15	23
	Hit ratio	50%	26%	85%	34%	28%	43%
Cond-ECS	Different genes	29	16	8	42	34	47
	PubMed hits	13	3	6	25	14	12
	Hit ratio	45%	19%	75%	60%	41%	26%

SCZ: Schizophrenia; BP: Bipolar disorder; RA: Rheumatoid arthritis; CAD: Coronary artery disease; TC: Total cholesterol

PubMed Hits: the number of papers mentioning the gene and the phenotype in their abstracts

An example in Rheumatoid arthritis

Conditional analysis based on p-values (P1) vs. based on selective expression (P2)

Group	Gene	Chromosome	StartPosition	OriginalP	#SNP	RankingScore	Select	Cond.P1*	Cond.P2*
6	PHTF1	1	114248388	4.79E-127	165	126.32	true	0.43106	0.14144
6	RSBN1	1	114304453	2.59E-140	109	139.59	true	2.59E-140	7.41E-06
6	PTPN22	1	114356432	2.34E-121	150	120.63	true	1	2.34E-121
6	AP4B1-AS1	1	114399256	6.39E-105	230	104.19	true	0.99999	0.48387
6	BCL2L15	1	114419435	4.7E-40	49	39.33	true	1	1
6	AP4B1	1	114436816	3.93E-26	53	25.41	true	1	1.77E-08
6	DCLRE1B	1	114447914	1.08E-29	43	28.97	true	0.10325	0.01665
6	HIPK1-AS1	1	114466622	1.56E-46	42	45.81	true	1	1
6	HIPK1	1	114496498	1.04E-32	125	31.98	true	0.92996	0.25159
6	OLFML3	1	114522012	3.06E-25	56	24.51	true	0.0303	0.00132
6	SYT6	1	114631913	2.02E-11	331	10.69	true	5.36E-06	0.00242

DESE: **PTPN22**

Ordinary ECS: **RSBN1**

PubMed Search:

PTPN22 >100 papers

RSBN1 0 papers

PTPN22 's selective expression

Tissue/Cell type	Mean	SE	Z-score	p-value
Cells-EBV-transformedlymphocytes	52.7688	1.00628	70.9546	0.0
Spleen	15.9522	0.415350	20.9685	0.0
SmallIntestine-TerminalIleum	10.5153	0.607467	13.5868	0.0
Lung	10.3931	0.282981	13.4208	0.0
Cells-Transformedfibroblasts	6.57372	0.191014	8.23525	2.22045E-16
Colon-Transverse	3.69515	0.144851	4.32699	1.51162E-5

Immune
tissues

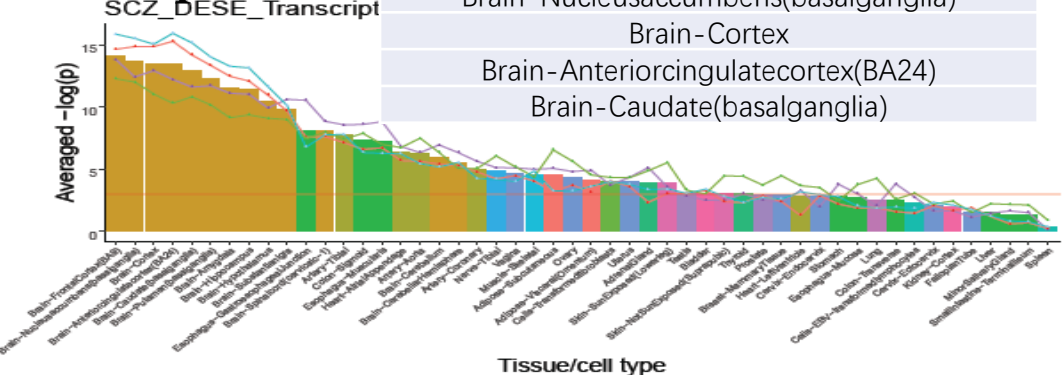
RSBN1 's selective expression

Tissue/Cell type	Mean	SE	Z-score	p-value
Brain-CerebellarHemisphere	22.7930	0.670436	3.27436	0.00105901
Ovary	18.0370	0.361491	2.24677	0.0246544
Uterus	17.2777	0.393087	2.08272	0.0372770
Cervix-Endocervix	16.3900	0.399812	1.89093	0.0586342
Cells-EBV-transformedlymphocytes	16.3061	0.359064	1.87279	0.0610971
Brain-Cerebellum	14.8991	0.316516	1.56880	0.116695

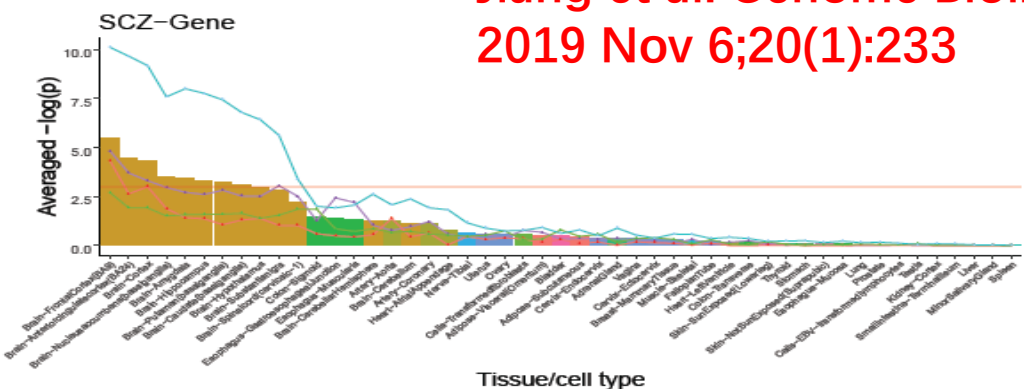
Brain regions

(<https://pmglab.top/rez>)

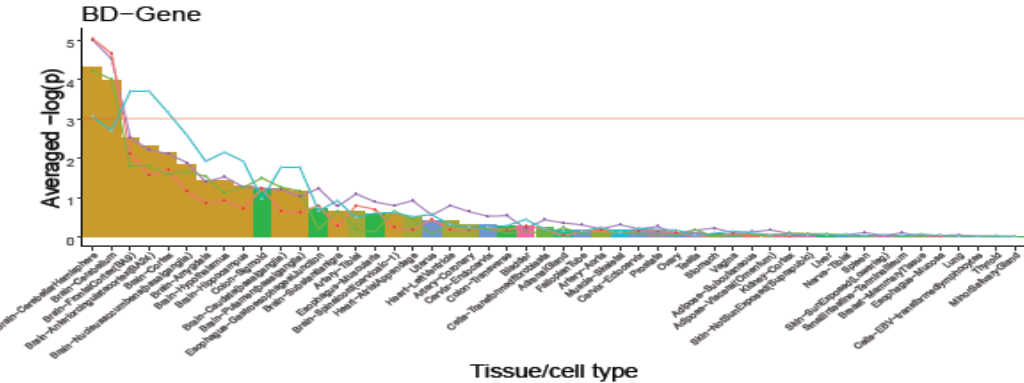
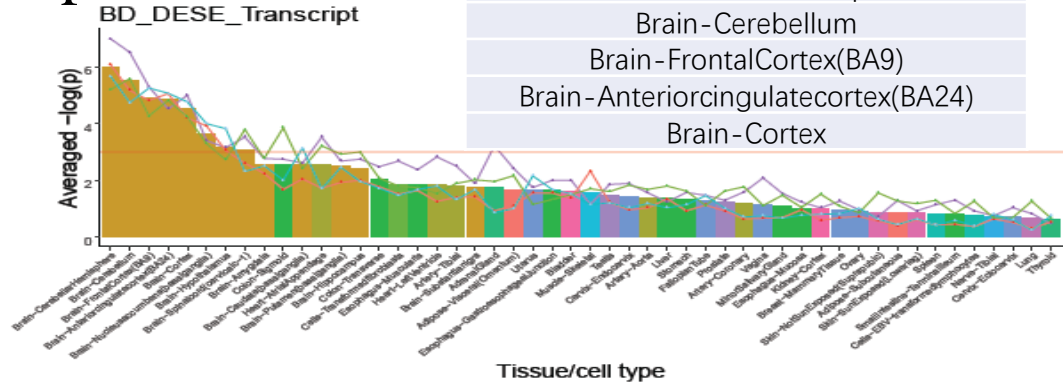
Schizophrenia



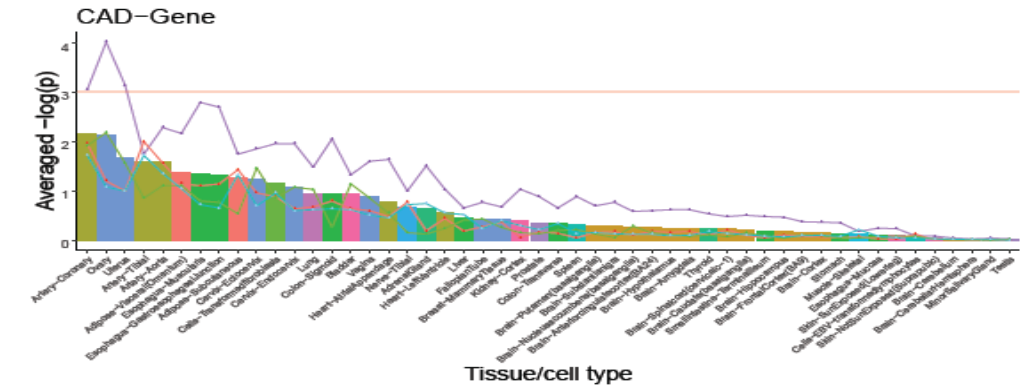
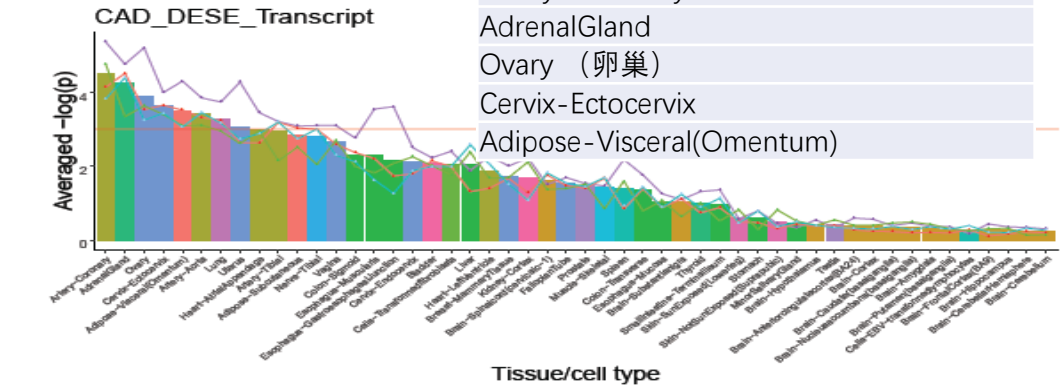
Jiang et al. Genome Biol.
2019 Nov 6;20(1):233



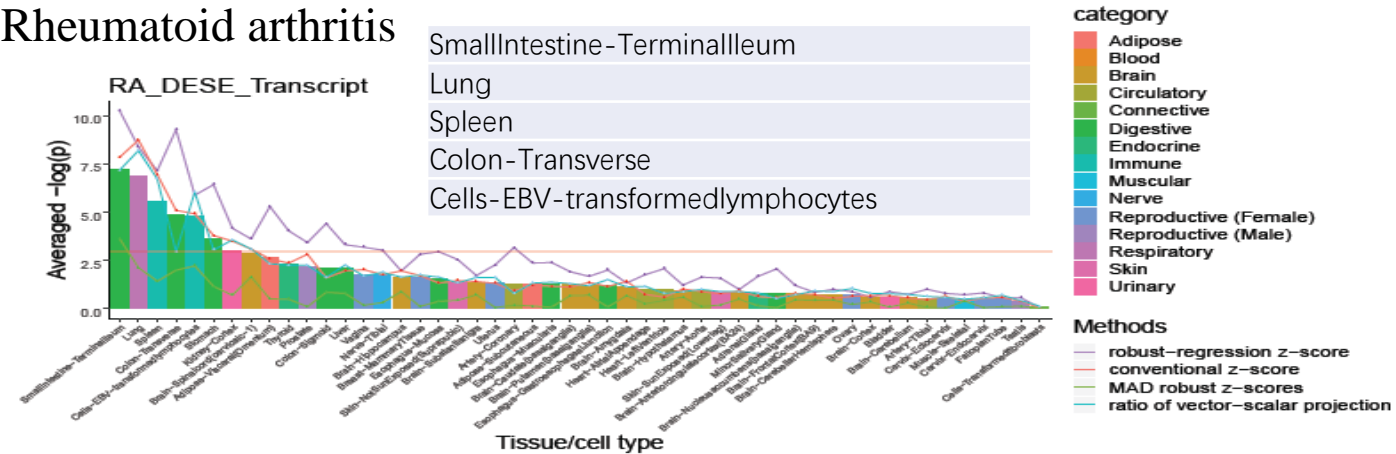
Bipolar disorder



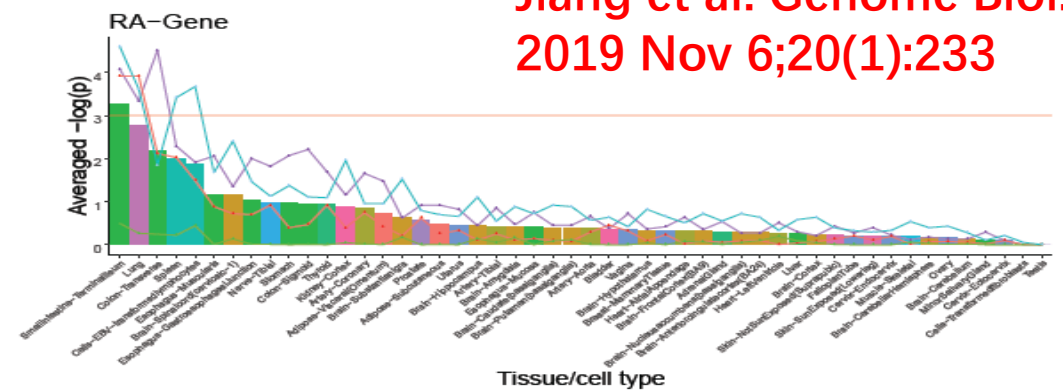
Coronary artery disease



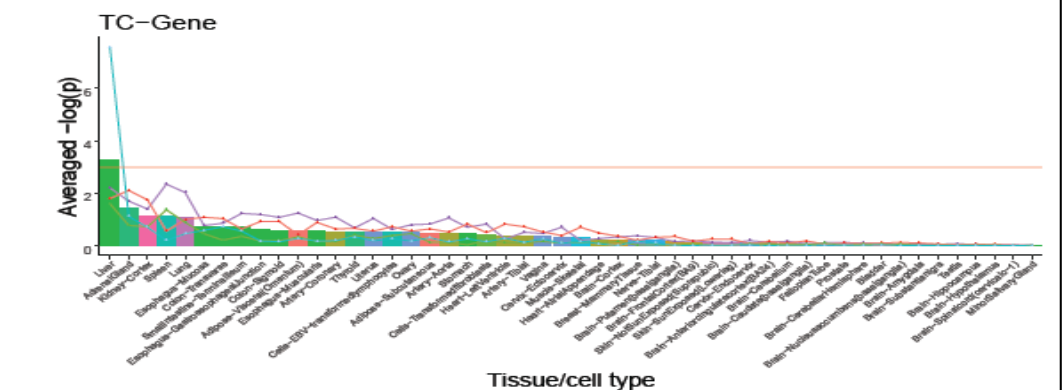
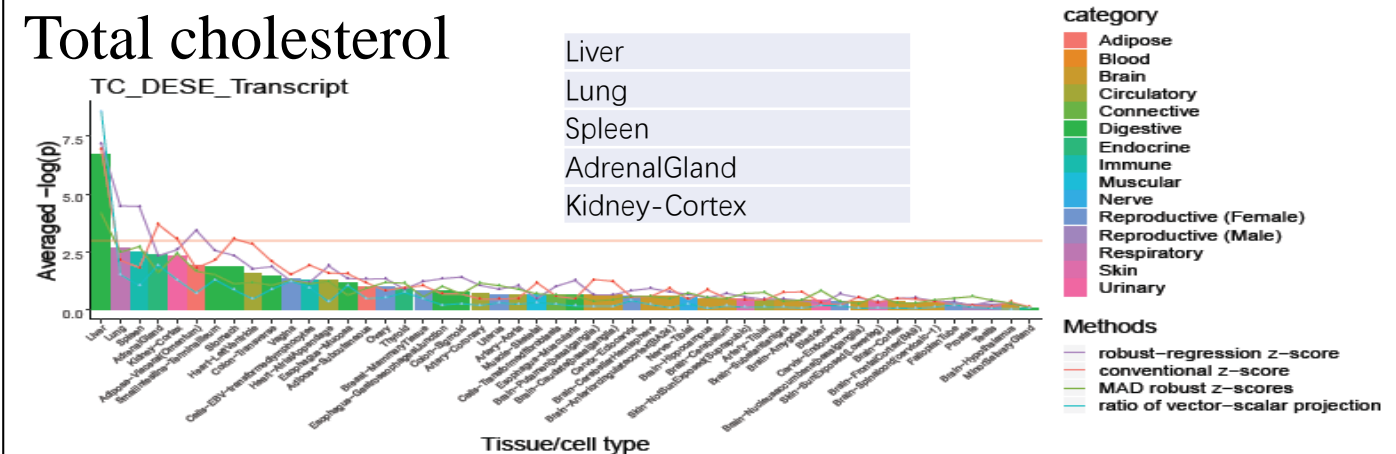
Rheumatoid arthritis



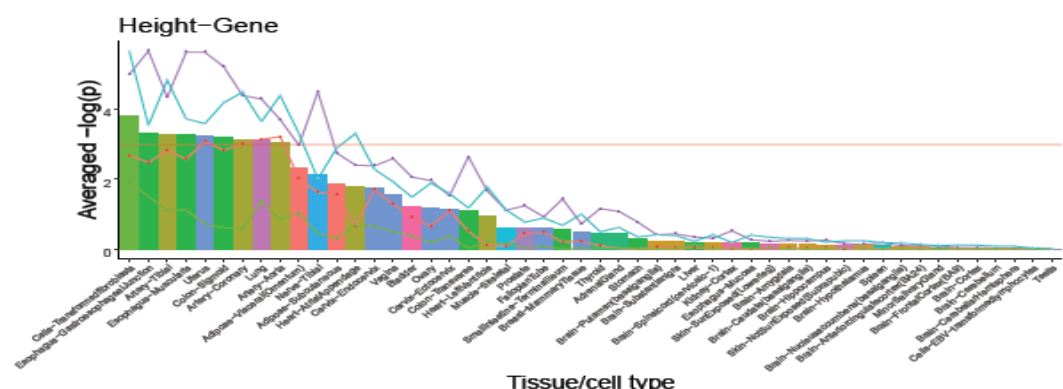
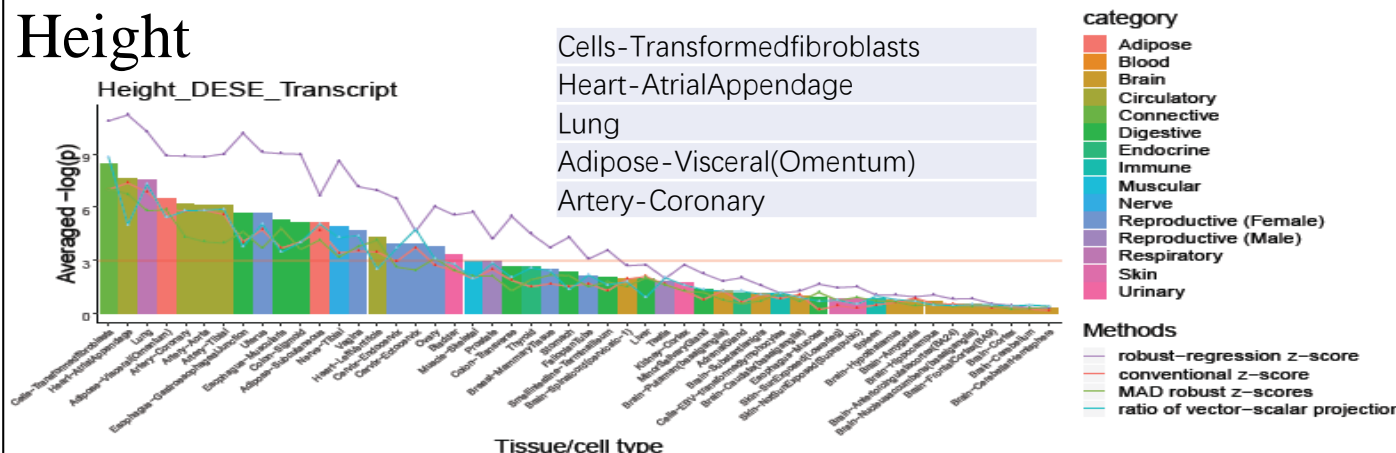
Jiang et al. Genome Biol.
2019 Nov 6;20(1):233



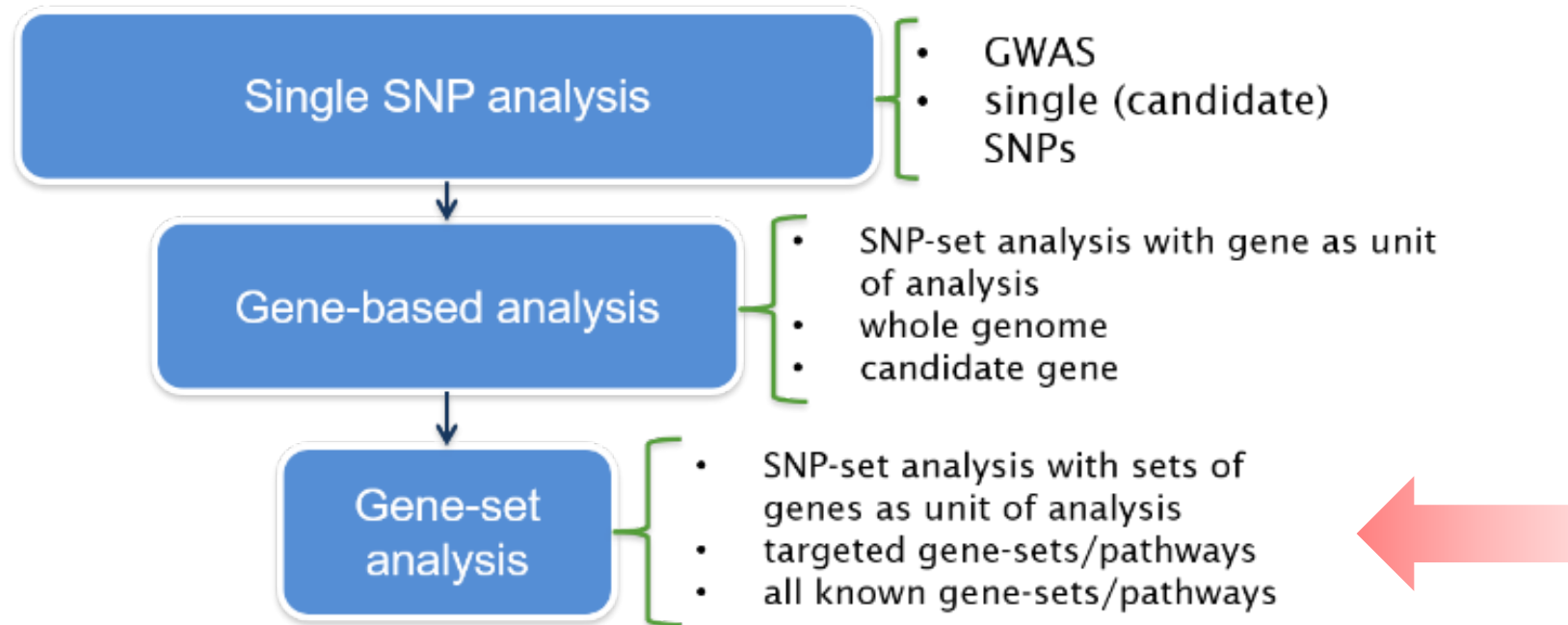
Total cholesterol



Height



Moving to multiple genes



Pros and cons of gene based analysis

- Pro's:

- reduce multiple testing (from 2.5M SNPs to 23k genes)
- accounts for heterogeneity in gene
- Immediate gene-level interpretation

- Cons:

- disregards regulatory (often non-genic) information
- Still a lot of tests

Gene-set analysis

Unit of analysis is a set of functionally related genes

Pro's:

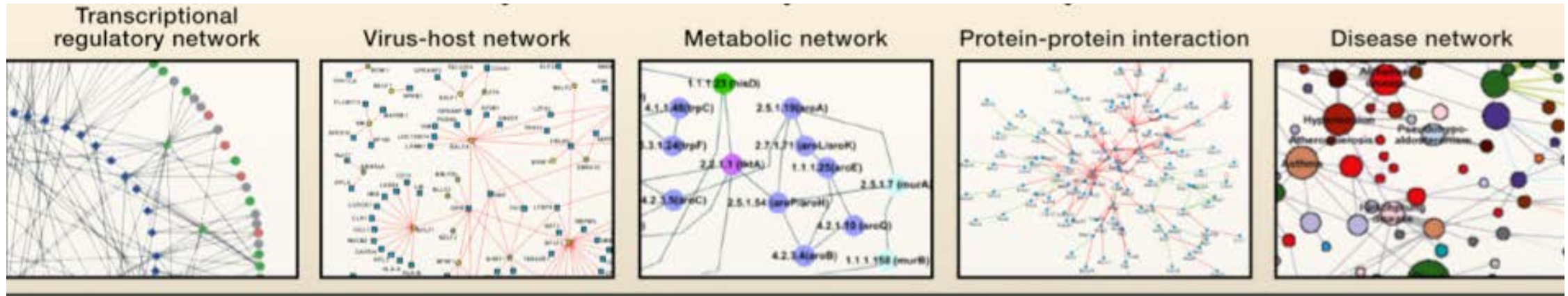
- ✓ Reduce multiple testing by prioritizing genes in biological pathways or in groups of (functionally) related genes
- ✓ Increased power
- ✓ Deals with genetic heterogeneity
- ✓ Provides immediate biological insight

Gene-set analysis

Cons

- crucial to select reliable sets of genes!
 - Different levels of information
 - Different quality of information

How to define gene-sets?



Create gene-sets based on

- protein-protein interaction
- co-expression
- transcription regulatory network
- biological pathway

How to collect gene-sets or pathways

Use public or commercial databases:

- KEGG
- Gene Ontology
- Ingenuity
- Biocarta
- String database
- Human Protein Interaction database
- HIPPIE database
- .. and any more

vs.

Create manually, expert curated lists

Online databases vs. manual

Information in online databases is

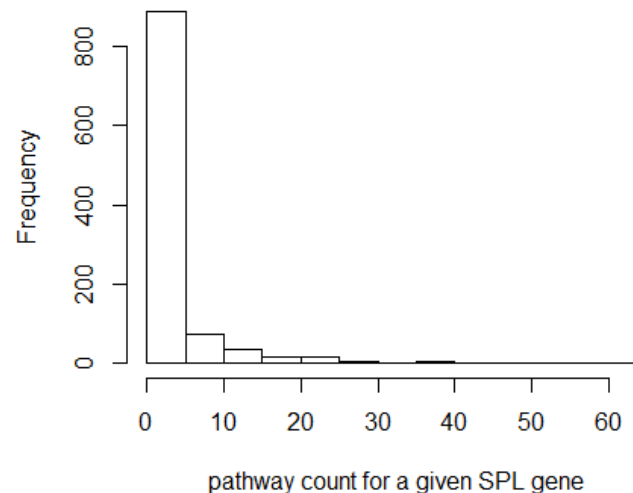
- somewhat biased
 - not all genes included, disease genes tend to be investigated more often
 - genes that are investigated more often will have more interactions
- not always reliable
 - interactions often not validated, sometimes only predicted. If experimentally seen, unknown how reliable that experiment was

Public databases vs. manual curation

Manually curated vs. KEGG:

- List of 1043 genes known to be active in the synapse, with known synaptic function, based on expert knowledge, verified by repeated lab experiments (Y2H, IP) – the synaptic particle list (SPL).
- Of these <42% have a known KEGG pathway:

438 / 1043 SPL genes have no KEGG pathway



Public databases vs. manual curation

Manually curated vs. GO:

- SPL genes: N=1043
- GO contains 117 terms that contain “synaptic” or “synapse”. These terms include 934 gene IDs.
- Of the 1043 SPL genes, only 388 are synaptic according to GO-terms, 655 are NOT synaptic according to GO.
- Of the 934 synaptic genes according to GO, 546 have not been experimentally validated as playing a role in the synapse

-> Thus: be aware of the biases and incompleteness in current information in online databases, if possible use well-annotated expert-curated gene-sets.

Molecular Signatures Database

<https://www.gsea-msigdb.org/gsea/msigdb>



The screenshot shows the GSEA Molecular Signatures Database website. At the top is the GSEA logo with the text 'Gene Set Enrichment Analysis'. Below it are navigation links: 'GSEA Home', 'Downloads', and 'Molecular Signatures'. On the left is a sidebar menu with links: 'MSigDB Home', 'About Collections', 'Browse Gene Sets', 'Search Gene Sets', 'Investigate Gene Sets', 'View Gene Families', and 'Help'. The main content area features the 'MSigDB Molecular Signatures Database' logo and the word 'Molecular' partially visible.

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 [HALLMARK_APOPTOSIS](#) gene set page.

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

regulatory target gene sets based on gene target predictions for microRNA seed sequences and predicted transcription factor binding sites.

C4

computational gene sets defined by mining large collections of cancer-oriented microarray data.

C5

ontology gene sets consist of genes annotated by the same ontology term.

C6

oncogenic signature gene sets defined directly from microarray gene expression data from cancer gene perturbations.

C7

immunologic signature gene sets represent cell states and perturbations within the immune system.

C8

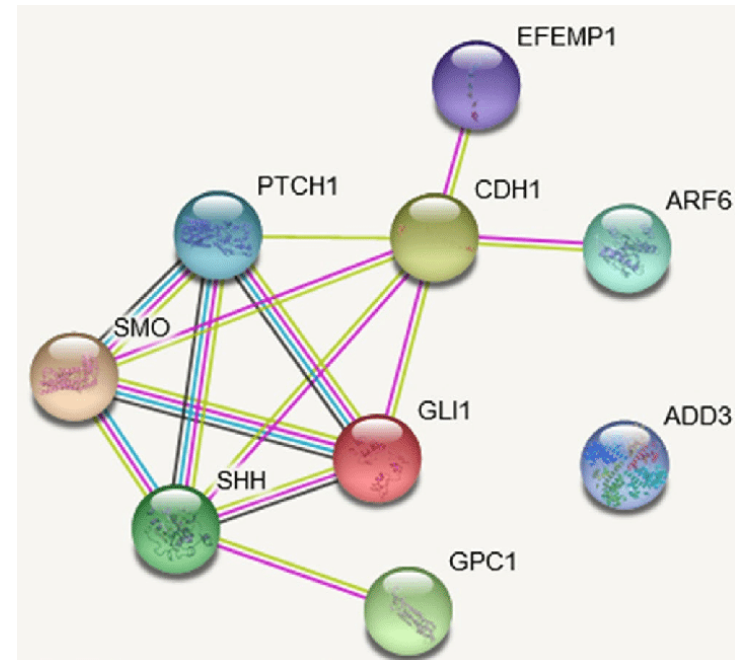
cell type signature gene sets curated from cluster markers identified in single-cell sequencing studies of human tissue.

UC San Diego



BROAD
INSTITUTE

Gene-pair based association analysis




STRING is a database of known and predicted protein-protein interactions. The interactions include direct (physical) and indirect (functional) associations; they stem from computational prediction, from knowledge transfer between organisms, and from interactions aggregated from other (primary) databases.






Known Interactions

-  *from curated databases*
-  *experimentally determined*

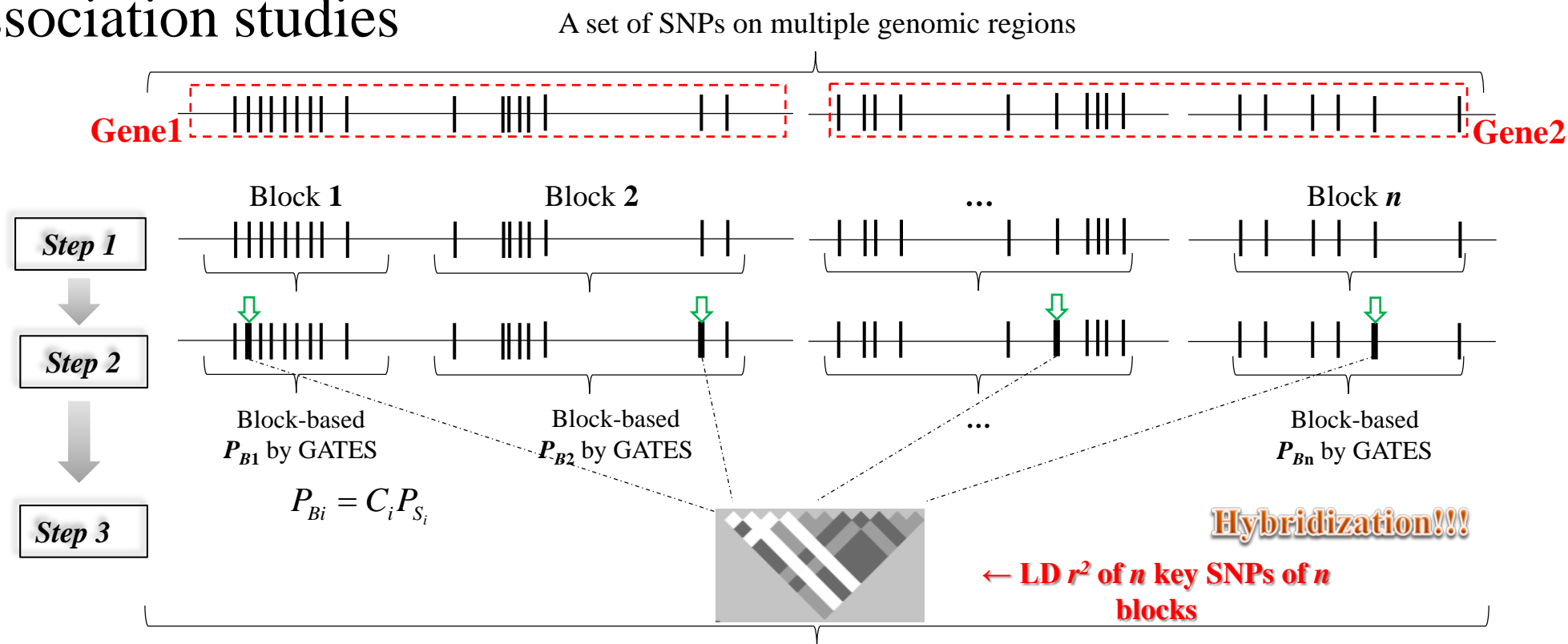
Predicted Interactions

-  *gene neighborhood*
-  *gene fusions*
-  *gene co-occurrence*

Others

-  *textmining*
-  *co-expression*
-  *protein homology*

HYST: A HYbrid Set-based Test for genome-wide association studies



Set-based P by scaled chi-square test, accounting for r^2

$$X = -2 \sum_{i=1}^n \ln P_{Bi}$$

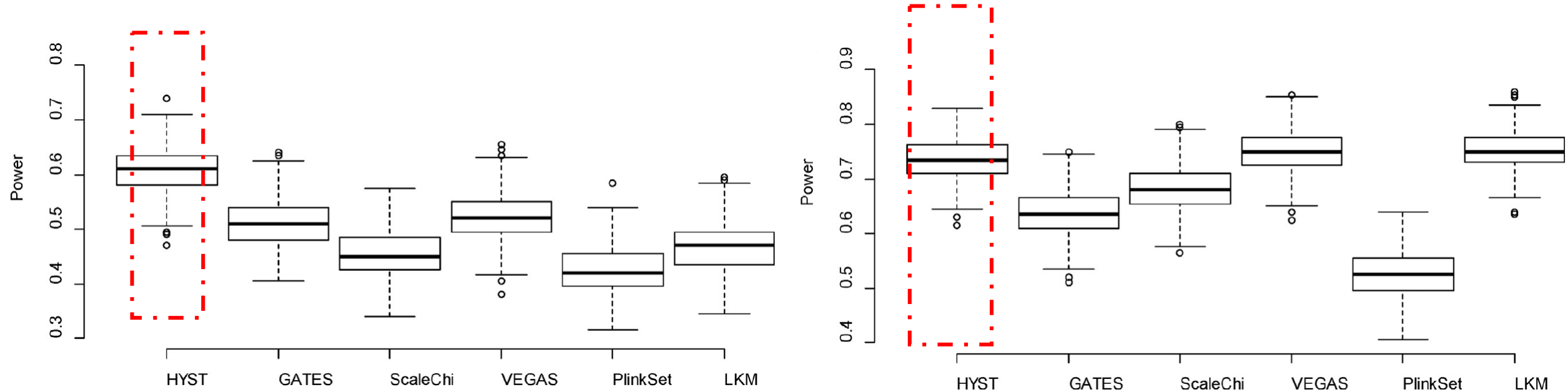
$$X \sim c \chi_f^2$$

$$c = 1 + \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{cov}(-2 \ln P_{Bi}, -2 \ln P_{Bj}) / 2n \quad f = 2n / c$$

$$\text{cov}(-2 \ln P_{Bi}, -2 \ln P_{Bj}) \approx 8.6 r^2 \quad (\text{Improved})$$

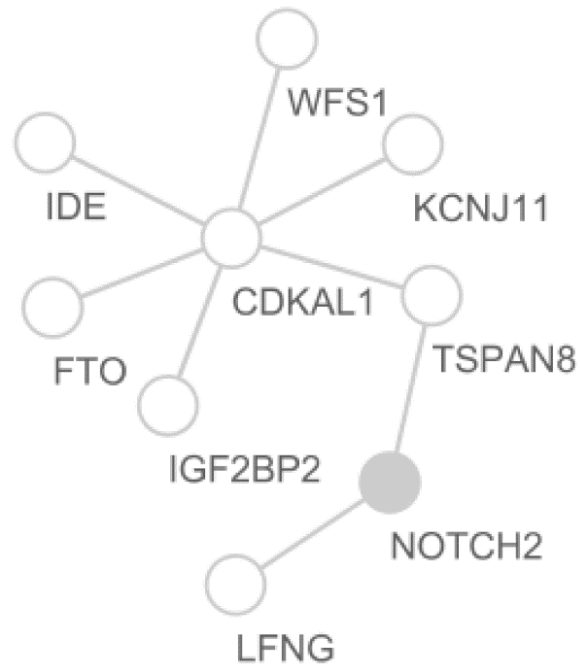
Li et al. Am J Hum Genet. 2012; 7;91(3):478-88.

Power in simulation under Alternative Hypotheses



- ScaleChi, scaled chis-square test;
- VEGAS: A versatile gene-based test for genome-wide association studies
- LKM: the Logistic Kernel Machine Test. **(SKAT)**
- GATES: A Rapid and Powerful Gene-Based Association Test Using Extended Simes Procedure

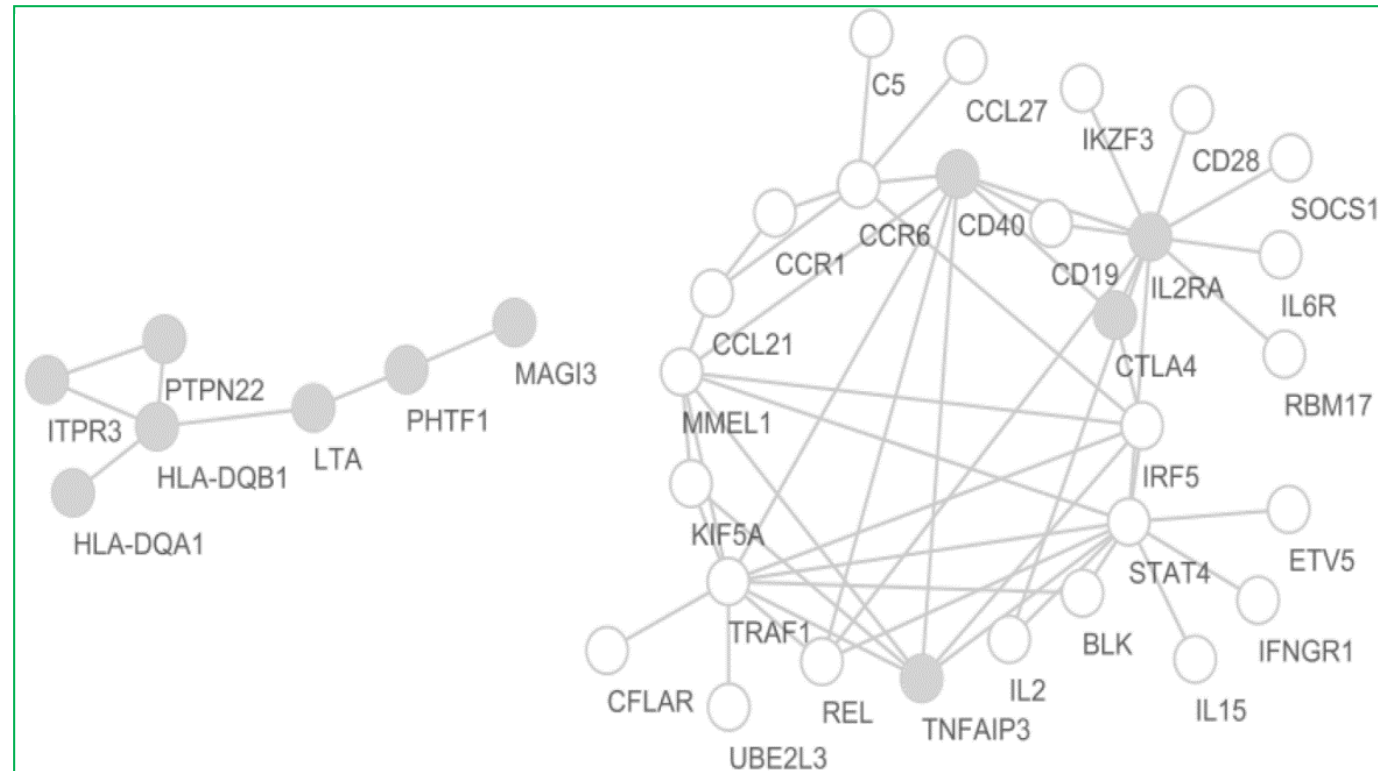
The significant PPI pairs in Type-2 diabetes GWAS data set



Seven(IGF2BP2, WFS1, CDKAL1, IDE, KCNJ11, TSPAN8 and FTO) out of the **Eight** were confirmed susceptibility genes

WFS1 even has a SNP-based and gene-based p value as large as as 8.2×10^{-3} and 2.4×10^{-2}

The significant PPI pairs in Rheumatoid arthritis GWAS data set



Filled: genes significant in either the SNP-based, the gene-based analysis, or both tests

Gene-set based association analysis for multiple genes

Null and alternative hypotheses:

- **Self-contained:**

- *H₀: The gene-set is not associated with the trait*
- *H_a: The gene-set is associated with the trait*

- **Competitive:**

- *H₀: The gene-set is not more strongly associated with the trait than other (matched) gene-sets*
- *H_a: The gene-set is more strongly associated with the trait than other (matched) gene-sets*

Nat Rev Genet. 2010, 11: 843-854; BMC Res Notes
. 2011 Oct 7;4:386

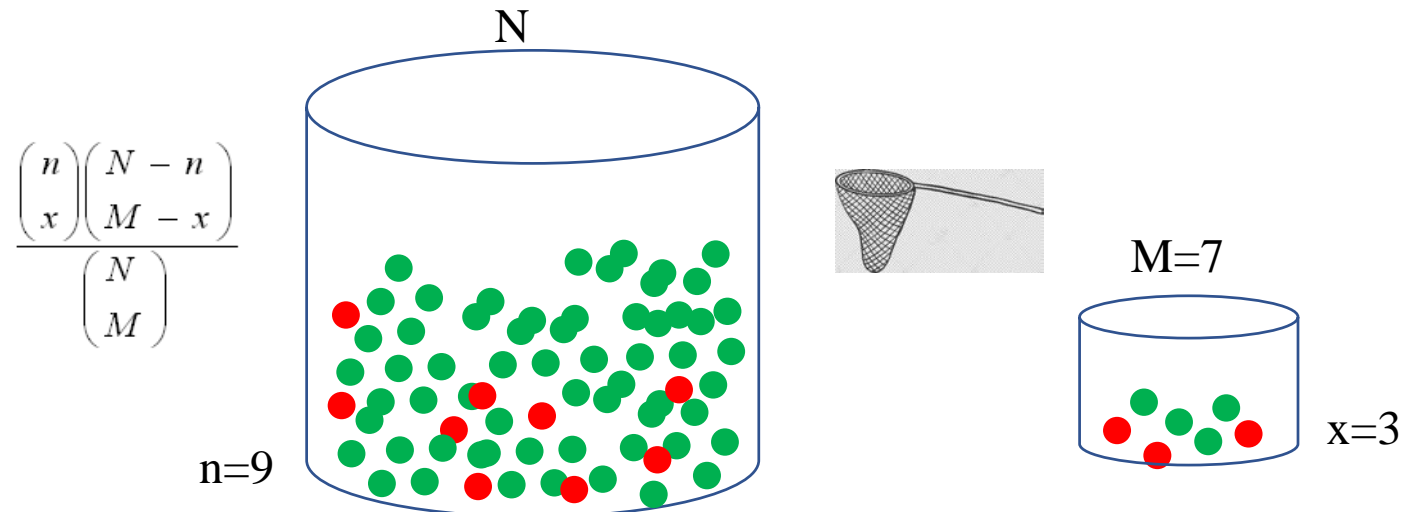
Competitive test (enrichment analysis):

- Hypergeometric test (Fisher's exact test):

$$P(X \geq q) = 1 - \sum_{x=1}^q \frac{\binom{n}{x} \binom{N-n}{M-x}}{\binom{N}{M}}$$

	Observed	Background
Inside Pathway	x	n
Outside Pathway	M-x	N-n
Total	M	N

Important: Use conditional independent genes for the analysis



Pros and cons in gene-set based association tests

- **Self-contained:**

- *Pros: independent of other genes; more powerful*
- *Cons: sensitive to outliers; hard to explain*

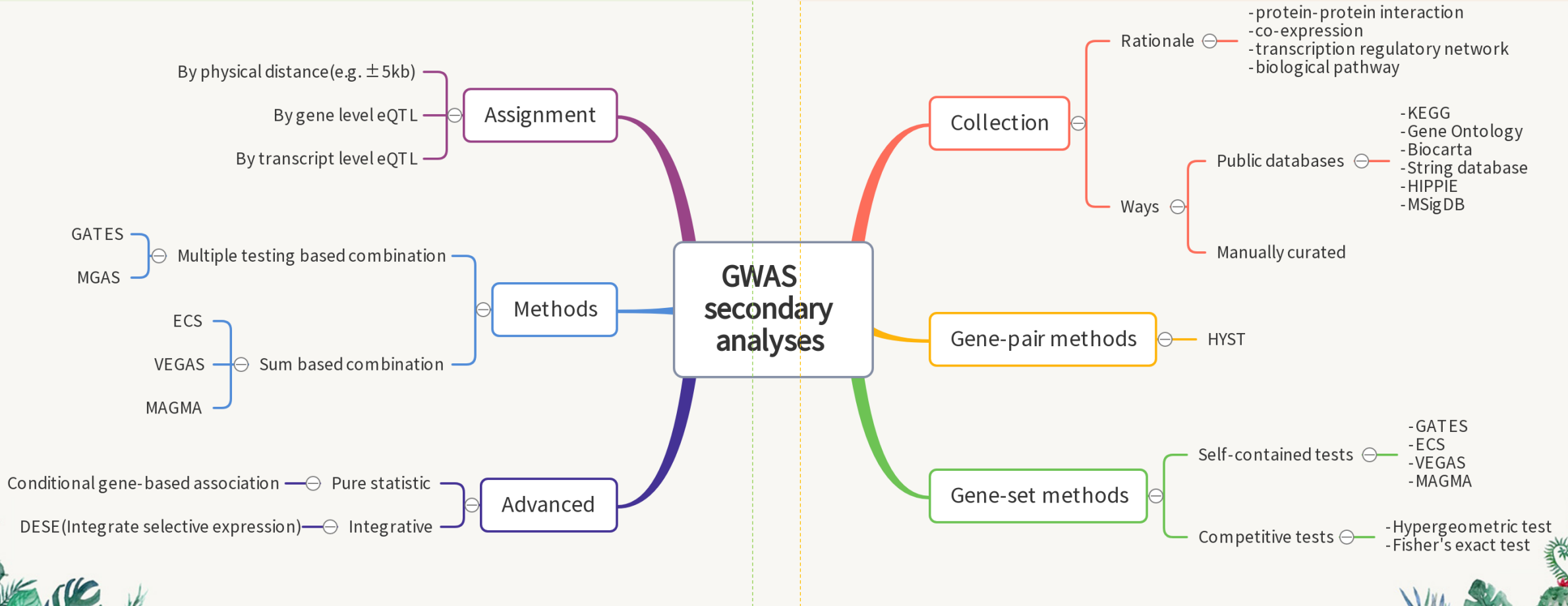
- **Competitive:**

- *Pros: insensitive to outliers; easier to explain*
- *Cons: sensitive to background sample sizes; lower power*

To recapitulate

Gene-based association

Gene-set-based association



Tools to perform gene- or pathway-based association analysis

- Stand alone
 - KGGSEE (<http://pmglab.top/kggsee>)
 - MAGMA (<https://ctg.cncr.nl/software/magma>)
 - ...
- Online
 - PCGA (<http://pmglab.top/pcga>)
 - FUMA (https://ctg.cncr.nl/software/fuma_gwas)
 - ...

Functions in KGGSEE and MAGMA

Functions	KGGSEE	MAGMA
Gene-based association	GATE & ECS	Best chi-square & mean chi-square
Conditional gene-based association	ECS & DESE	NO
Gene-set based association	Competitive tests	Competitive tests + self-contained tests
Gene-based causation	Effective median MR	NO
Gene-based heritability	ECS	NO

Differences in PCGA and FUMA

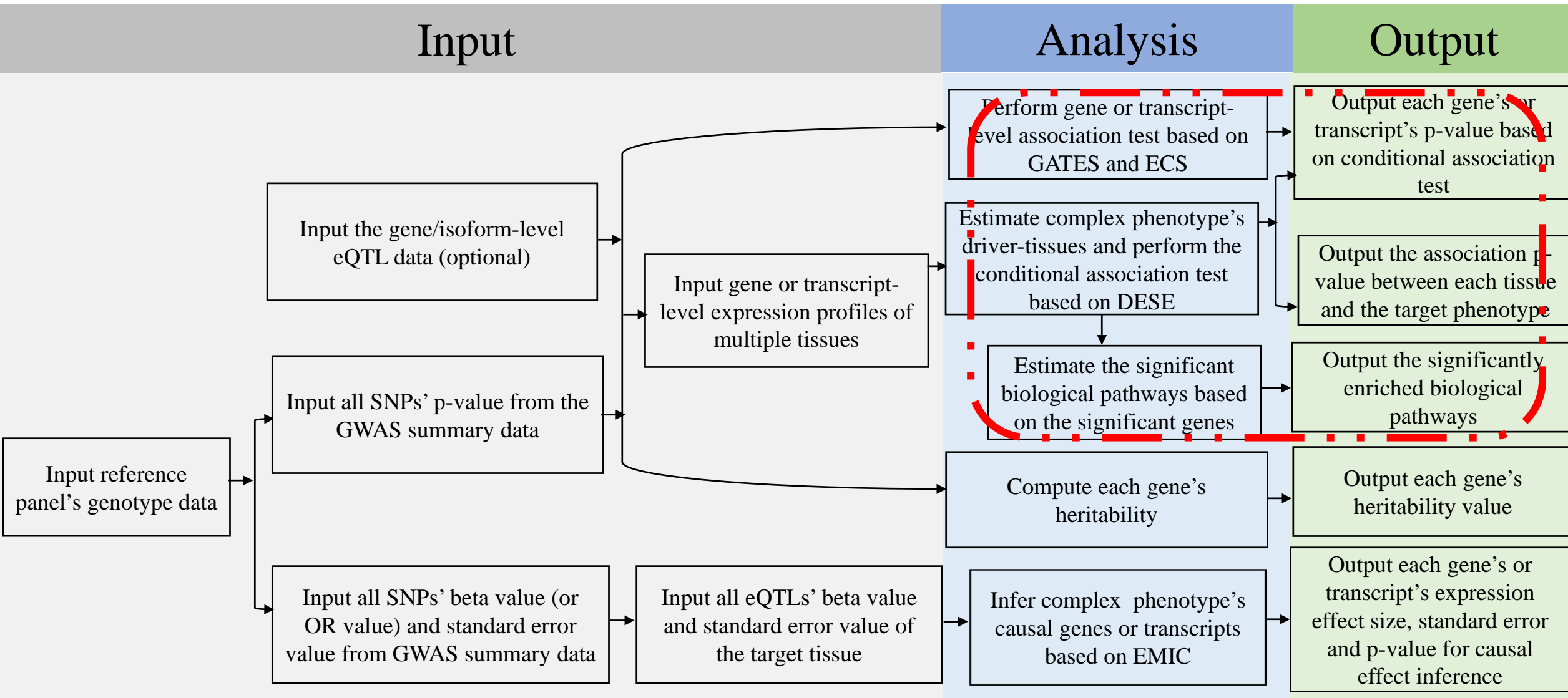
	PCGA	FUMA
Core Method	DESE, subtly allows the estimation of associated cell types and prioritization of susceptibility genes to help each other	Regression model based on associated genes of MAGMA and expression
Expression Dataset	6,598 cell types (human and mouse) and 54 human tissues Manually unified cell type label and sampling tissue/organ	2,679 cell types (human and mouse) Raw cell type label
Selection of cell types' expression profile	(1) Automatically select by associated tissues results (2) Manually select by unified tissues (organs)	Manually select by datasets
Phenotype-cell-gene association landscape	Associated tissues, cell types and genes of 1,871 public GWASs	No

Practical with KGGSEE

KGGSEE (a biological Knowledge-based mining platform for Genomic and Genetic association Summary statistics using gEne Expression) ,
<http://pmglab.top/kggsee>

- Main functions
 - Phenotype-gene/transcript association test;
 - Phenotype gene set association test;
 - Phenotype's driver cell type/tissue estimation;
 - Gene/transcript expression causal effect estimation;
 - Gene/transcript heritability estimation.

KGGSSEE's main workflow



The main input data of KGGSEE

- Input 1: Key GWAS summary statistics of SNPs

CHR	BP	P
1	205053219	0.000329
1	113657233	0.000356
1	46874246	0.000519
1	46877284	0.000519
1	236749649	0.001261
1	236721660	0.001309

CHR: chromosome

BP: position on the chromosome

P: association p-value

Input 2: eQTL summary statistics

```

◊ #symbol id      chr  pos   ref   alt   altfreq beta  se  p
◊ >WASH7P ENSG00000227232 1
◊ 19391 G      A      0.11  -0.879 0.283  1.89E-3 72  0.181
◊ 49243 G      A      0.09  0.76  0.264  4.03E-3 96  0.125
◊ 54380 T      C      0.247 0.549 0.165  8.52E-4 94  0.15
◊ 101674 C     A      0.079 0.929 0.354  8.64E-3 64  0.147
◊ 104339 T     C      0.086 -0.749 0.213  4.22E-4 129 0.145
◊ 245455 C     G      0.08  -1.154 0.418  5.76E-3 58  0.169
◊ 247506 T     C      0.078 -0.792 0.255  1.87E-3 147 0.112
◊ 276482 T     G      0.139 -0.694 0.242  4.07E-3 119 0.093
◊ 294107 A     C      0.43  0.344 0.122  4.71E-3 127 0.11
◊ 322762 A     G      0.552 0.278 0.108  9.91E-3 156 0.084
◊ 533291 G     A      0.128 0.645 0.249  9.62E-3 67  0.133
◊ 535952 T     C      0.098 -0.597 0.21  4.49E-3 126 0.118
◊ 535968 A     G      0.059 -0.807 0.259  1.82E-3 133 0.109
◊ 535971 A     T      0.073 -0.788 0.26  2.48E-3 105 0.154
◊ 540975 G     A      0.071 -0.719 0.26  5.59E-3 141 0.078
◊ 569427 C     T      0.06  -0.772 0.223  5.24E-4 155 0.115
◊ 572444 G     A      0.071 0.882 0.318  5.51E-3 122 0.108
◊ 578738 G     A      0.136 1.06  0.316  8.09E-4 61  0.245
◊ 616154 C     T      0.102 0.503 0.184  6.18E-3 145 0.099
◊ 770505 A     C      0.127 0.84  0.288  3.50E-3 56  0.159
◊ 924603 T     C      0.927 -0.799 0.305  8.75E-3 154 0.092
◊ 924629 A     G      0.929 -0.805 0.306  8.48E-3 155 0.087
◊ 943250 C     T      0.925 -0.795 0.283  4.96E-3 156 0.088
◊ >OR4F5 ENSG00000186092 1
◊ 14522 G      A      0.051 0.889 0.331  7.29E-3 136 0.059
◊ 14542 A      G      0.101 0.876 0.237  2.21E-4 121 0.113
◊ 103756 G     T      0.103 -0.578 0.203  4.34E-3 132 0.077
◊ 108573 C     T      0.054 -0.578 0.216  7.33E-3 156 0.058

```

symbol: HGNC gene symbol;
 id: genes' ENSEMBL ID;
 chr: the chromosome;
 pos: SNPs' position;
 ref: the allele on reference genome;
 alt: the alternative allele;
 beta: the effect size of the SNP to gene expression;
 se: the standard error of beta ;
 p: the association p-value the SNP with the expression;
 neff: the effective sample sizes;
 r²: the determination coefficient of the regression

Input 3: Expression profile of multiple tissues or cell types

Name	Adipose-Subcutaneous.mean	Adipose-Subcutaneous.SE	Artery-Aorta.mean	Artery-Aorta.SE	Bladder.mean	Bladder.SE	Brain-Amygdala.mean	Brain-Amygdala.SE	...
ENSG00000223972.5	0.0038016	0.00036668	0.0037934	0.00047474	0.003839	0.001373	0.010042	0.001372	...
ENSG00000227232.5	1.9911	0.030021	1.8168	0.035621	2.1628	0.20917	1.4575	0.052624	...
ENSG00000278267.1	0.00049215	0.00010645	0.00032336	0.0001084	0	0	0.000282	0.000162	...
ENSG00000243485.5	0.0047772	0.00038018	0.003716	0.00041081	0.00384	0.001369	0.016675	0.001954	...
ENSG00000237613.2	0.0030462	0.00027513	0.0024075	0.00034102	0.003002	0.001391	0.008903	0.001378	...
ENSG00000268020.3	0.011766	0.00061769	0.0069227	0.00055835	0.009265	0.00234	0.046975	0.005307	...
ENSG00000240361.1	0.017913	0.00093294	0.012114	0.00071561	0.009961	0.003628	0.067864	0.006905	...
ENSG00000186092.4	0.024515	0.001612	0.011889	0.00078072	0.013971	0.003836	0.12368	0.021212	...
ENSG00000238009.6	0.02339	0.0014725	0.018208	0.0015718	0.016282	0.004774	0.05095	0.005741	...

Basic options

The **best way** to learn how to use KGGSEE is to look up our **online manual** <https://kggsee.readthedocs.io/en/latest/index.html> !!!

```
java -Xmx4g -jar ../kggsee.jar \  
--sum-file scz_gwas_eur_chr1.tsv.gz \  
--vcf-ref 1kg_hg19_eur_chr1.vcf.gz \  
--keep-ref \  
--gene-assoc \  
--out t1
```

Flag	Description
--sum-file	Specifies a whitespace delimited file of GWAS summary statistics. In this analysis, columns of SNP coordinates and p-values (CHR, BP, and P by default) are needed.
--vcf-ref	Specifies a VCF file of genotypes sampled from a reference population. These genotypes are used to estimate LD correlation coefficients among SNPs.
--keep-ref	Keep the parsed VCF file (KGGSEE object format) in a folder named VCFRefhg19 under the output folder. KGGSEE will read these files in the following tutorials, which will be faster than parsing VCF files.
--gene-assoc	Triggers gene-based association tests.
--out	Specifies the prefix of output files.

Notes

- Update kggsee.jar
java -jar kggse.jar --lib-update

Goals in the present tutorial:

- Infer associated genes of schizophrenia based on coordinates of variants;
- Infer associated transcripts of schizophrenia based on eQTLs;
- Infer associated tissues and prioritize susceptibility genes of schizophrenia;
- Infer associated gene-sets or pathways of schizophrenia