Pathway analysis

Douglas Wightman

And thanks to Christiaan de Leeuw and Danielle Posthuma

CTG lab – VU Amsterdam

VU VRIJE UNIVERSITEIT AMSTERDAM

BRAINSCAPES

CNCR CTGLAB

To avoid straining the system:

- mkdir thursday_magma
- cd thursday_magma
- cp /home/douglasw/Boulder2025/magma_session.zip .

- Practical https://vuamsterdam.eu.qualtrics.com/jfe/form/SV_3f5d2iC6AvneNr8

Outline

- What is pathway analysis in a GWAS context?
- Why is it useful?
- Different pathways
- What type of pathway analyses are there?
    - MAGMA
    - GSEA
    - LDSC
- Self-contained vs competitive
- Conditional gene-set analysis
- Applications of gene-set analysis

Pathway analysis

- Pathway analysis (in our context) is a way to identify pathways relevant to our data using:
  - A pre-defined set of genes based on some functional/biological grouping (e.g. genes in the citric acid cycle)
  - A set of genes identified in our data (e.g. significant genes from a GWAS)

- Often this is called gene-set enrichment analysis
  - Where you define a set of genes from your data and identify the probability of overlap between your set and a pre-defined set
- If a disproportionally large portion of your genes in your set (e.g. genes significant in a diabetes GWAS) were also present in a gene-set defined by genes involved in insulin production.
  - Then there is evidence for insulin production being relevant to diabetes

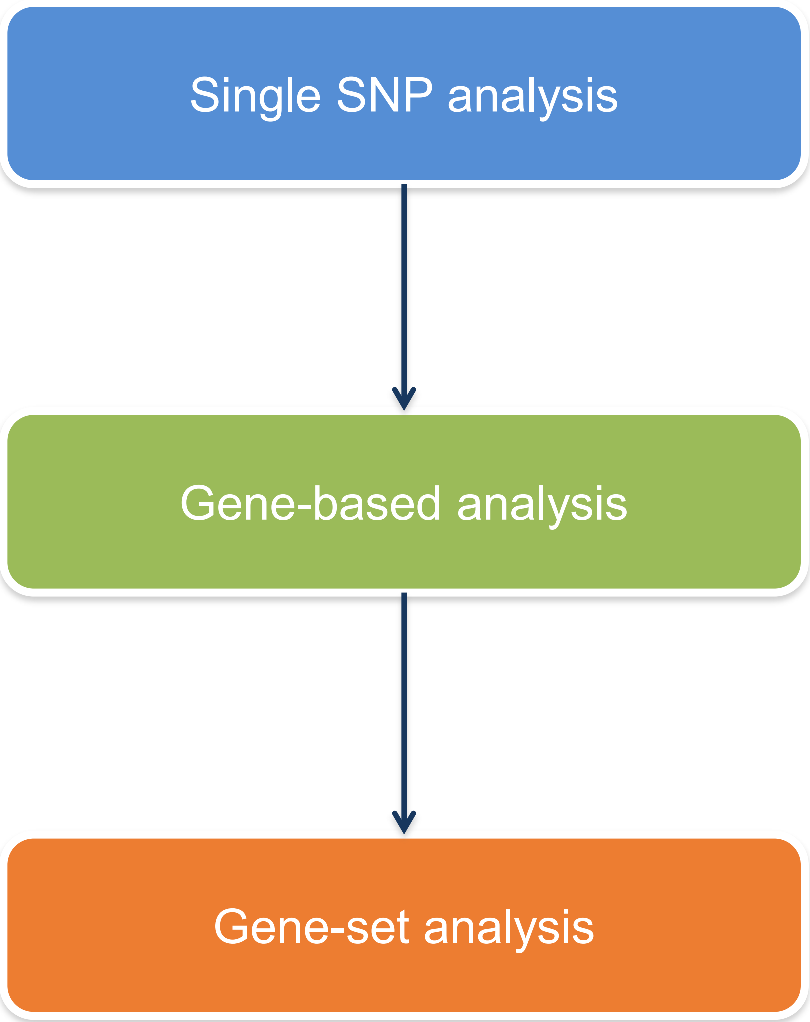
Why perform gene-set enrichment?

*Many traits are **polygenic** (many variants contribute to the trait)*

These variants can be aggregated together to highlight higher order biological processes

This may allow for easier translation to functional experiments where pathways can be targeted rather than specific variants

Testing for functional clustering of SNP associations



Single SNP analysis

- GWAS
- single SNPs

Gene-based analysis

SNP-set or gene-based analysis with gene as unit of analysis

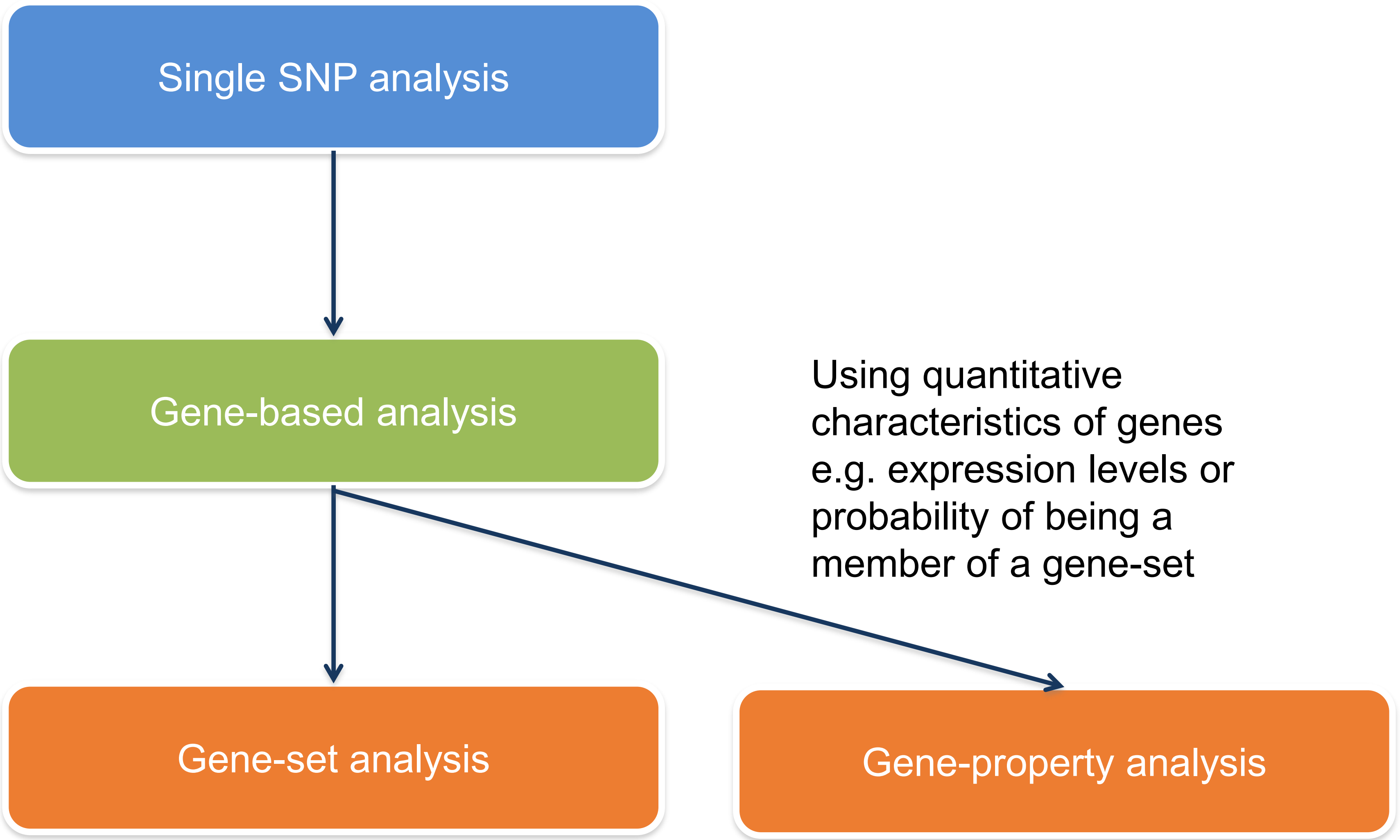
- whole genome

Gene-set analysis

Gene-set analysis with sets of genes as unit of analysis

- targeted gene-sets/pathways
- all known gene-sets/pathways

Testing for functional clustering of SNP associations



Single SNP analysis

Gene-based analysis

Gene-set analysis

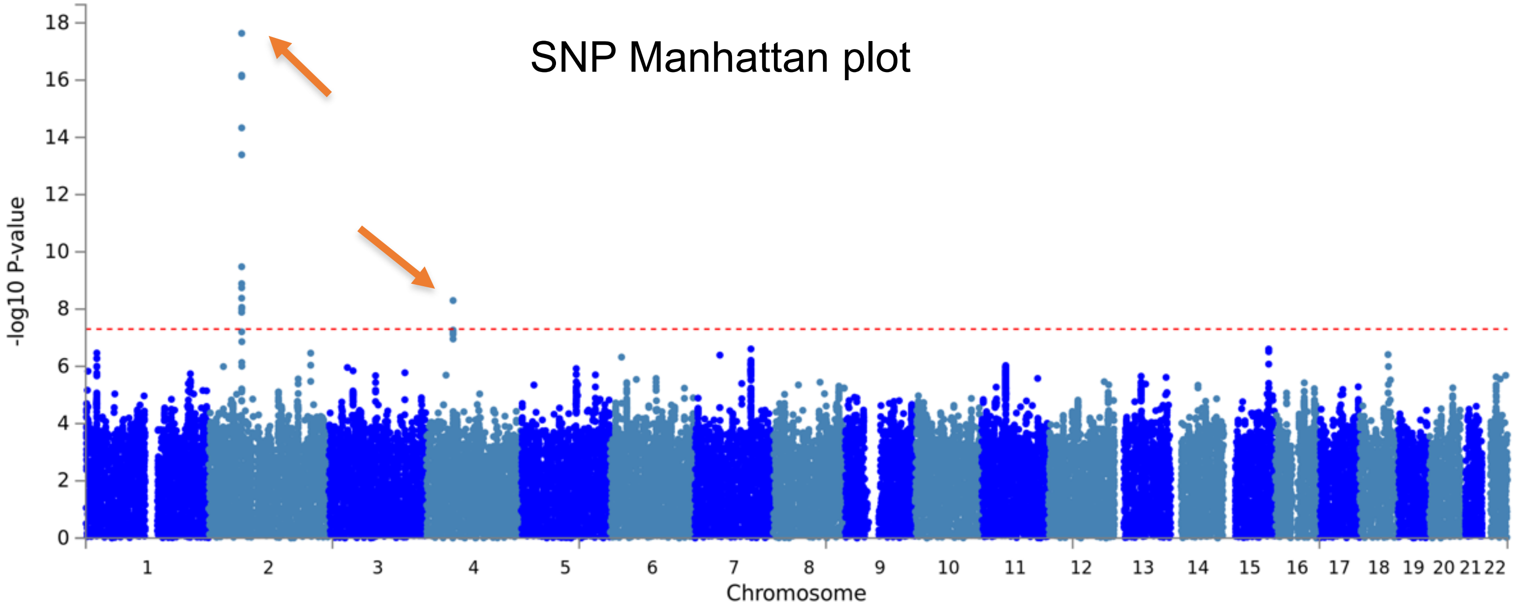
Gene-property analysis

Using quantitative characteristics of genes e.g. expression levels or probability of being a member of a gene-set

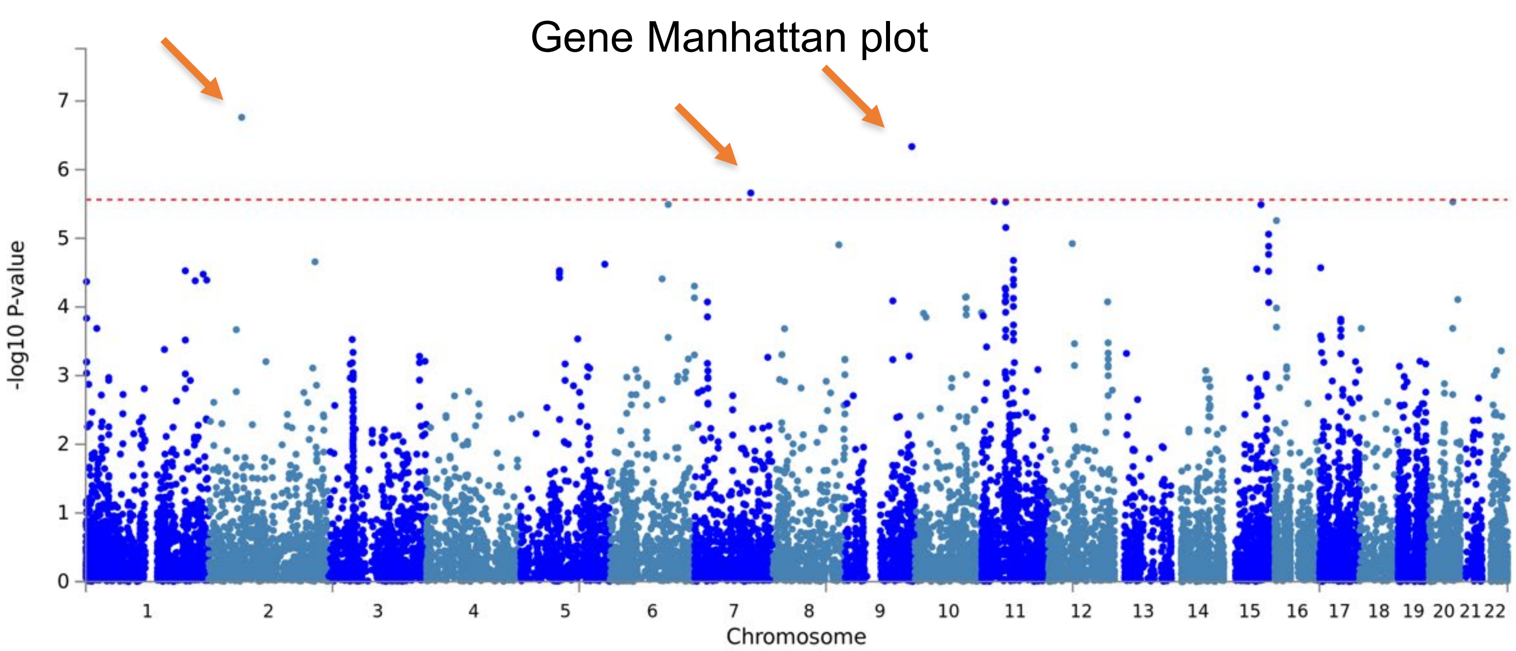
Gene-based analysis

- Instead of testing single SNPs and annotating GWAS-significant ones to genes, we 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



SNP Manhattan plot



Gene Manhattan plot

Gene-based analysis

Unit of analysis is the gene

- 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 when based on location-based annotation
  - Still a lot of tests

Gene-set analysis

Unit of analysis is a **set** of functionally related genes

Pro’s:

- Genes below significance threshold can converge on the same gene-set
- Provides biological insight

Cons

- Crucial to select reliable sets of genes!

Choosing gene-sets

Gene-sets can be based on e.g.

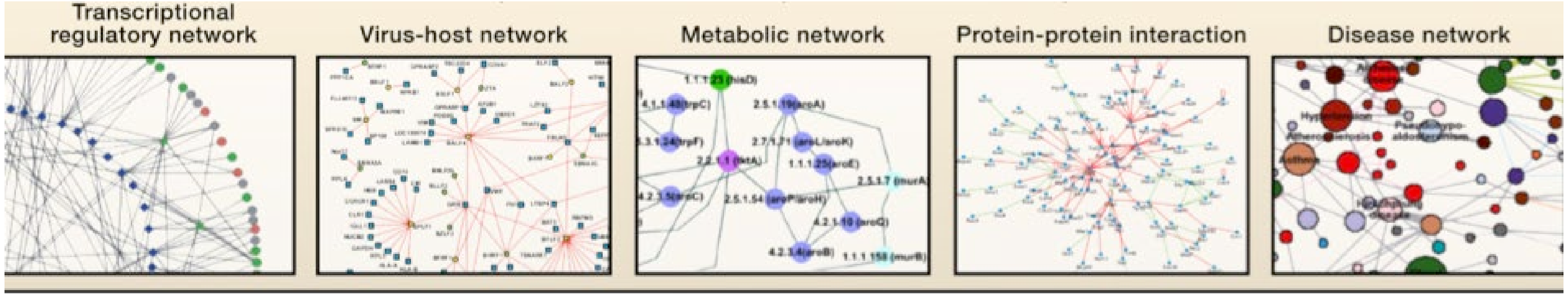
-protein-protein interaction

-co-expression

-transcription regulatory network

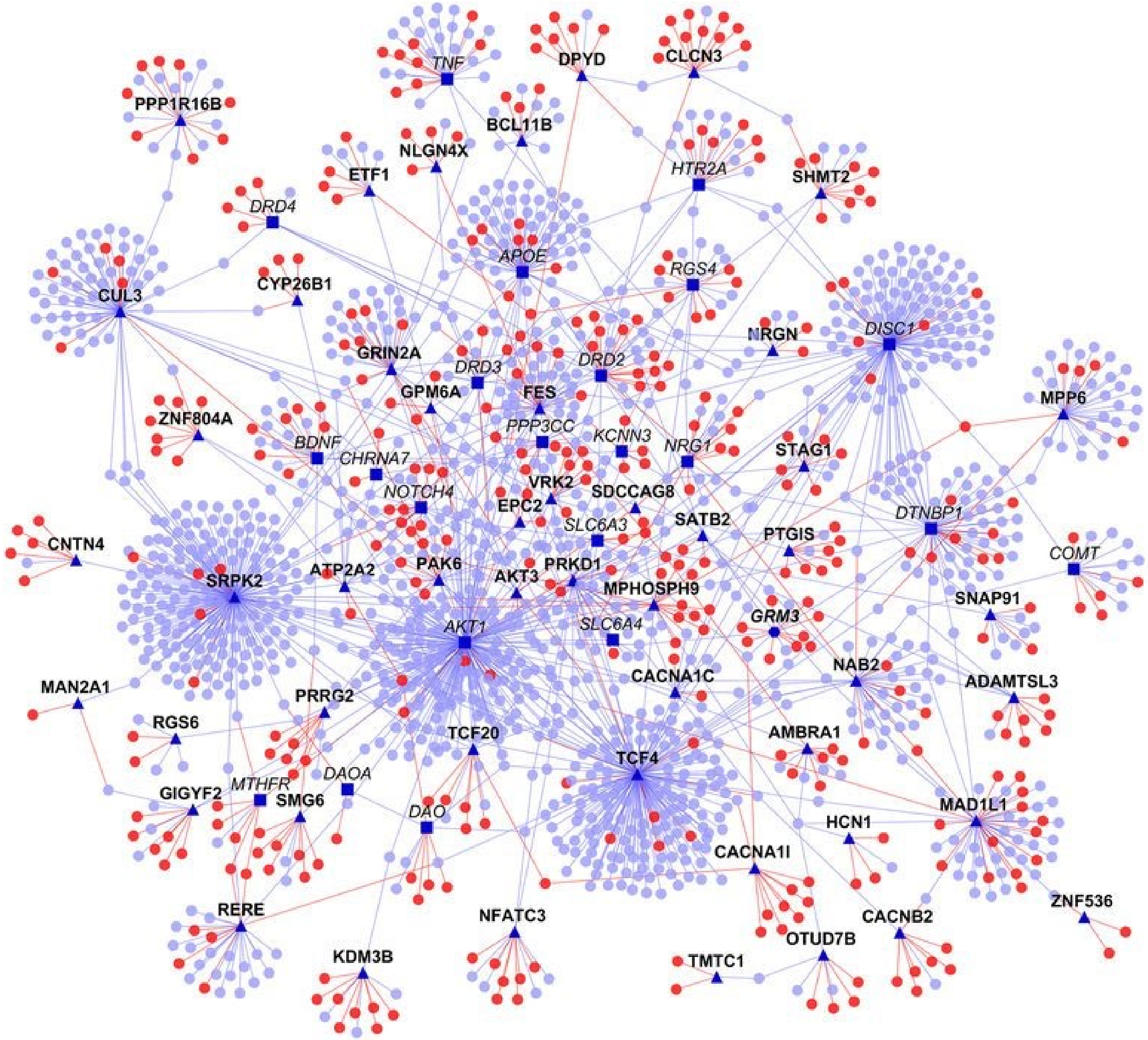
-biological pathway

-Functional relations

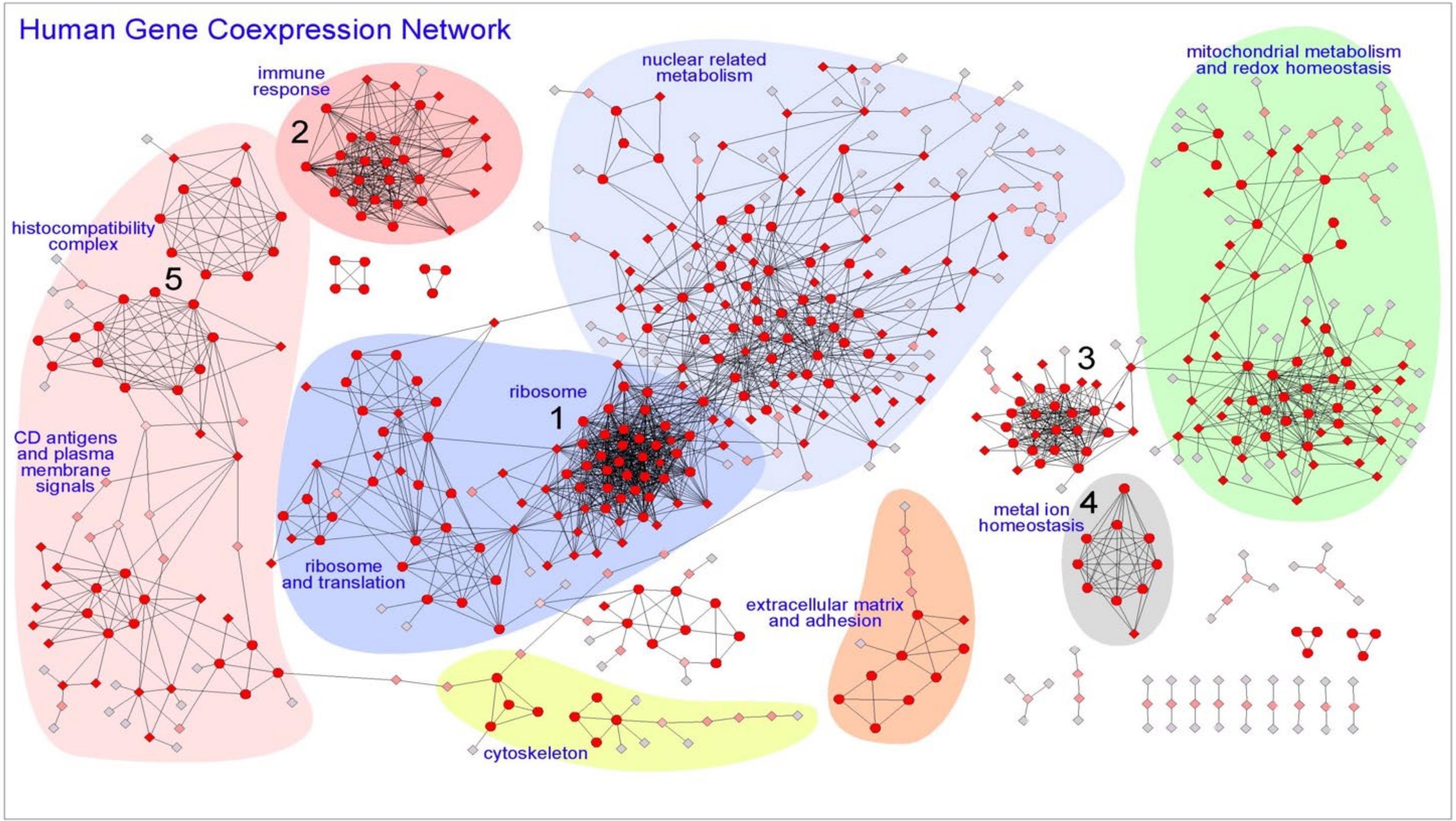

Transcriptional regulatory network
Virus-host network
Metabolic network
Protein-protein interaction
Disease network

Protein interaction networks

Using Y2H or Immunoprecipitations

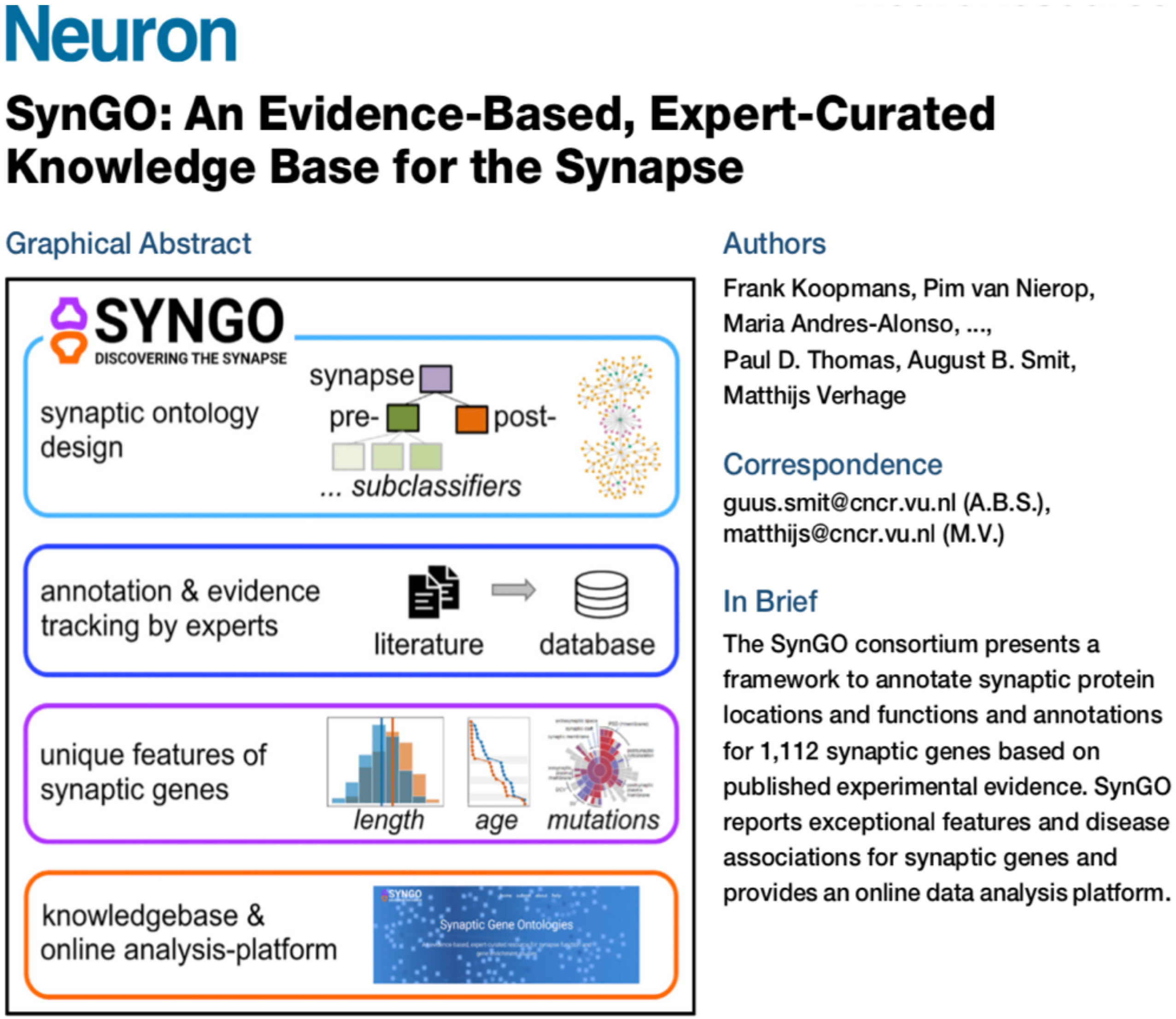


Co-expression networks


Human Gene Coexpression Network
immune response
2
nuclear related metabolism
mitochondrial metabolism and redox homeostasis
histocompatibility complex
5
ribosome
1
3
CD antigens and plasma membrane signals
metal ion homeostasis
4
ribosome and translation
extracellular matrix and adhesion
cytoskeleton

https://doi.org/10.1371/journal.pone.0003911.g006

Based on function - SYNGO



Neuron

SynGO: An Evidence-Based, Expert-Curated Knowledge Base for the Synapse

Graphical Abstract

Authors

Frank Koopmans, Pim van Nierop,
Maria Andres-Alonso, ...,
Paul D. Thomas, August B. Smit,
Matthijs Verhage

Correspondence

guus.smit@cncr.vu.nl (A.B.S.),
matthijs@cncr.vu.nl (M.V.)

In Brief

The SynGO consortium presents a framework to annotate synaptic protein locations and functions and annotations for 1,112 synaptic genes based on published experimental evidence. SynGO reports exceptional features and disease associations for synaptic genes and provides an online data analysis platform.

Selecting cell types based on GWAS results

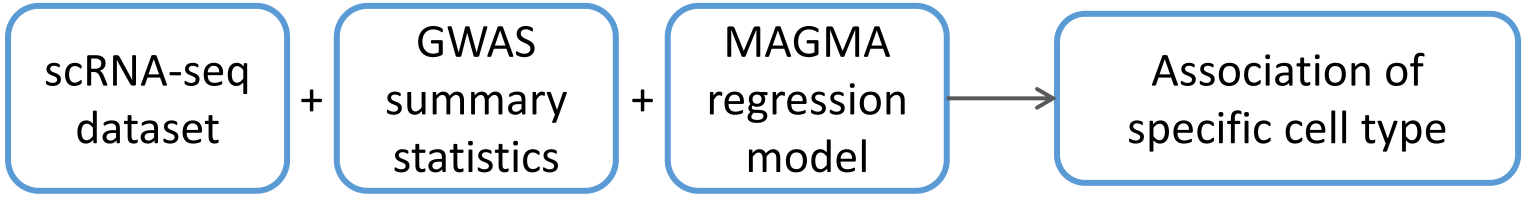
- GWAS-based gene P values can be combined with single cell expression values to imply cell types in complex traits

- Basically it tests whether there is an association between the association strength of genes with a trait and their expression levels in specific cell types

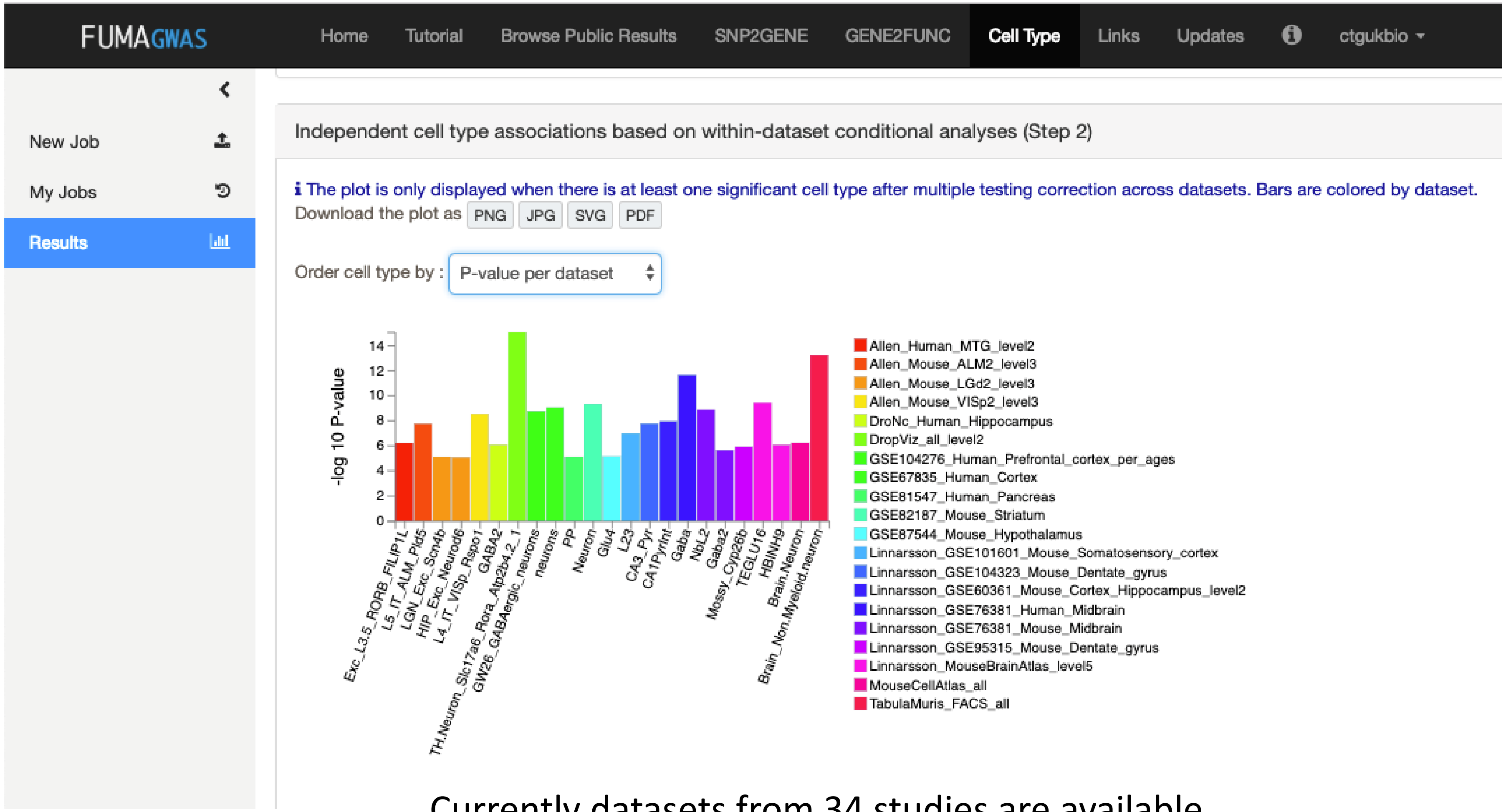
- *FUMA includes cell type enrichment analyses based on GWAS results*

(Watanabe, Mirkov, de Leeuw, Heuvel, Posthuma Nat Comm, 2019)

Cell type specificity analysis



scRNA-seq dataset + GWAS summary statistics + MAGMA regression model → Association of specific cell type



Independent cell type associations based on within-dataset conditional analyses (Step 2)

ℹ The plot is only displayed when there is at least one significant cell type after multiple testing correction across datasets. Bars are colored by dataset.

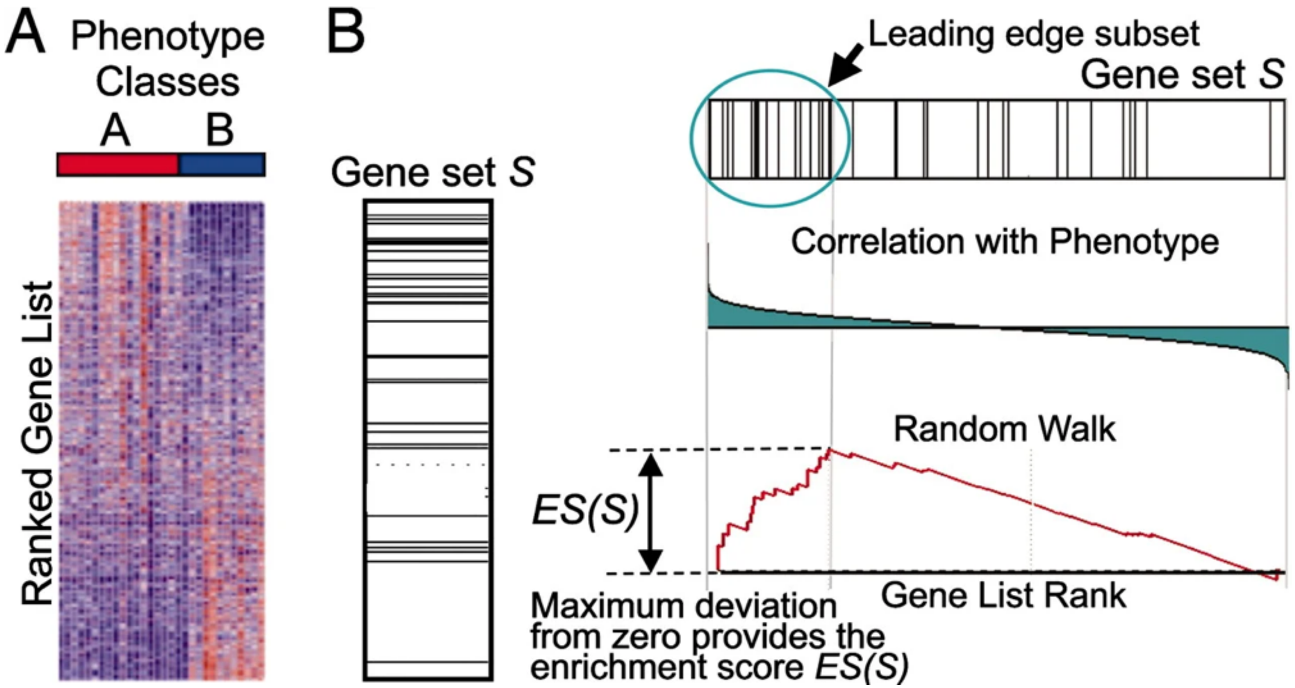
Download the plot as PNG JPG SVG PDF

Order cell type by : P-value per dataset

Currently datasets from 34 studies are available

Tools for statistical analysis of gene-sets

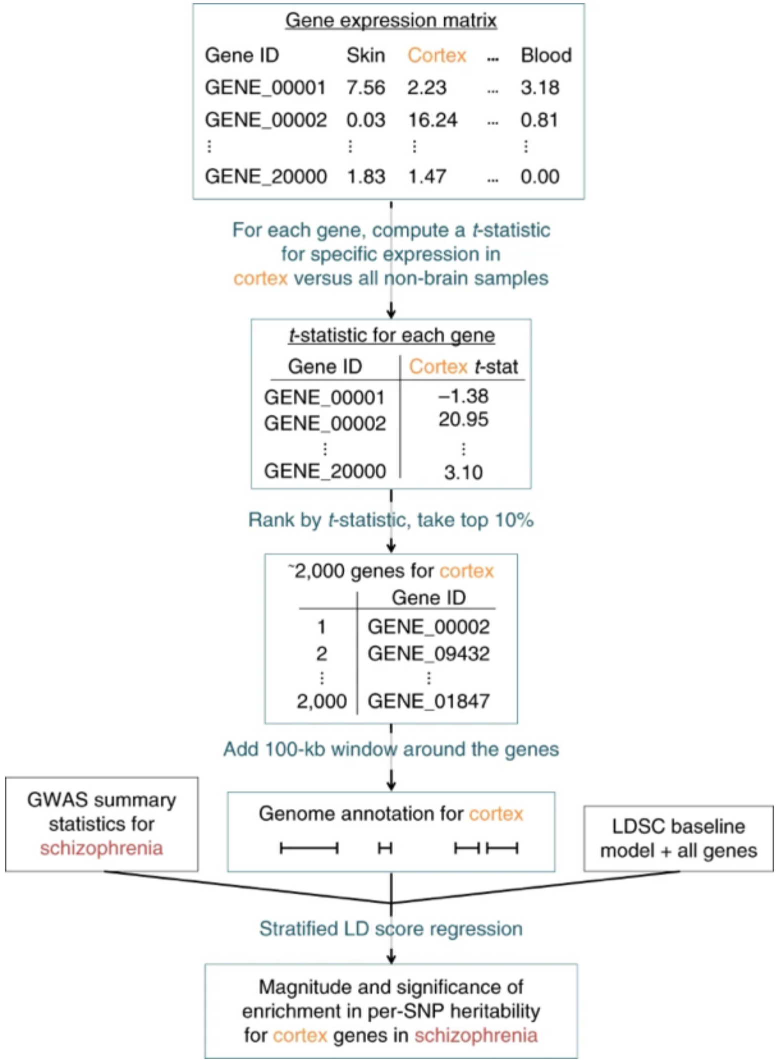
- GSEA – Gene set enrichment analysis
  - A user supplies a ranked set of genes from their analysis (e.g. ranked by P-value or effect size)
  - Then a random walk algorithm assesses the deviation from a null enrichment



https://doi.org/10.1073/pnas.0506580102

Tools for statistical analysis of gene-sets

- LD Score regression – partitioned heritability
  - Assess whether the heritability of a set of genes within an annotation (e.g. highly expressed genes in a specific cell type) is significantly different from 0 after conditioning on the baseline model
  - Baseline model includes 53 functional categories:
    - Includes regions expected to have more heritability (coding, UTR, promoter regions, histone marks etc)



Gene expression matrix

Gene ID	Skin	Cortex	...	Blood
GENE_00001	7.56	2.23	...	3.18
GENE_00002	0.03	16.24	...	0.81
⋮	⋮	⋮		⋮
GENE_20000	1.83	1.47	...	0.00

For each gene, compute a *t*-statistic for specific expression in cortex versus all non-brain samples

t-statistic for each gene

Gene ID	Cortex *t*-stat
GENE_00001	−1.38
GENE_00002	20.95
⋮	⋮
GENE_20000	3.10

Rank by *t*-statistic, take top 10%

~2,000 genes for cortex

	Gene ID
1	GENE_00002
2	GENE_09432
⋮	⋮
2,000	GENE_01847

Add 100-kb window around the genes

GWAS summary statistics for schizophrenia

Genome annotation for cortex

LDSC baseline model + all genes

Stratified LD score regression

Magnitude and significance of enrichment in per-SNP heritability for cortex genes in schizophrenia

https://doi.org/10.1038/s41588-018-0081-4

Tools for statistical analysis of gene-sets

- MAGMA – competitive gene set analysis
  - Regression based model
  - First, SNP P-values are used to estimate a gene Z-score for association with a trait
  - Then the vector of gene Z-scores is the outcome variable in a regression model
  - The predictors are either
    - a vector of membership for all genes in gene-set where included=1 and excluded=0
    - Or some quantitative gene-property (expression)
  - The regression framework is flexible so allows for conditional analyses
  - Approach accounts for LD between SNPs and genes, gene size, and number of SNPs
  - Compares enrichment of association signal in genes within a gene-set against genes not in the gene-set
    - Prevent inflation for traits with wide spread of signal

$$Z = \beta_{0,S} + S_S\beta_S + \epsilon$$

- S_S : indicator (if the gene is in a specified gene set)
- β_S : difference in effects between genes in the specified set and genes outside the set.

Statistical issues in gene-set analyses

- Self-contained vs. competitive tests

- Different statistical algorithms test different alternative hypotheses

- Different statistical algorithms have different sensitivity to LD, ngenes, nSNPs, background h^2

Self-contained vs. competitive tests

Null hypothesis:

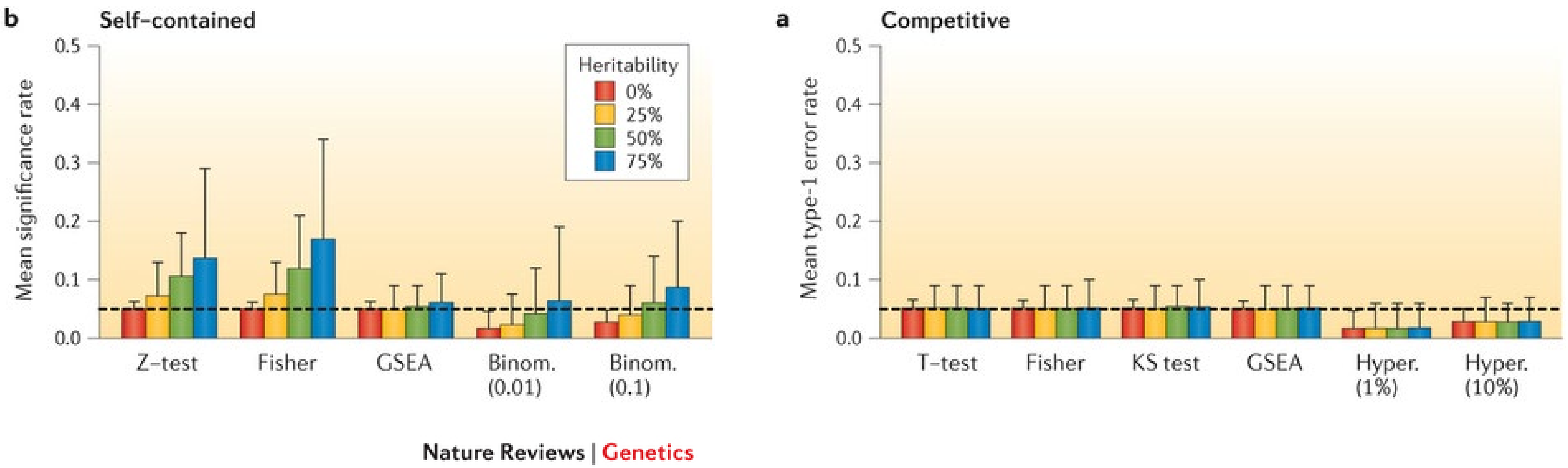
Self-contained:
H0: The genes in the gene-set are not associated with the trait

Competitive:
H0: The genes in the gene-set are not more strongly associated with the trait than the genes not in the gene-set

Why use competitive tests

- Polygenic traits influenced by thousands of SNPs in hundreds of genes
- Very likely that many combinations (i.e. gene-sets) of causal genes are significantly related
- Competitive tests define which combinations are biologically most interpretable

Polygenicity and number of significant gene-sets in self-contained versus competitive testing



For self-contained methods, rates increase with heritability, whereas they are constant for competitive methods.

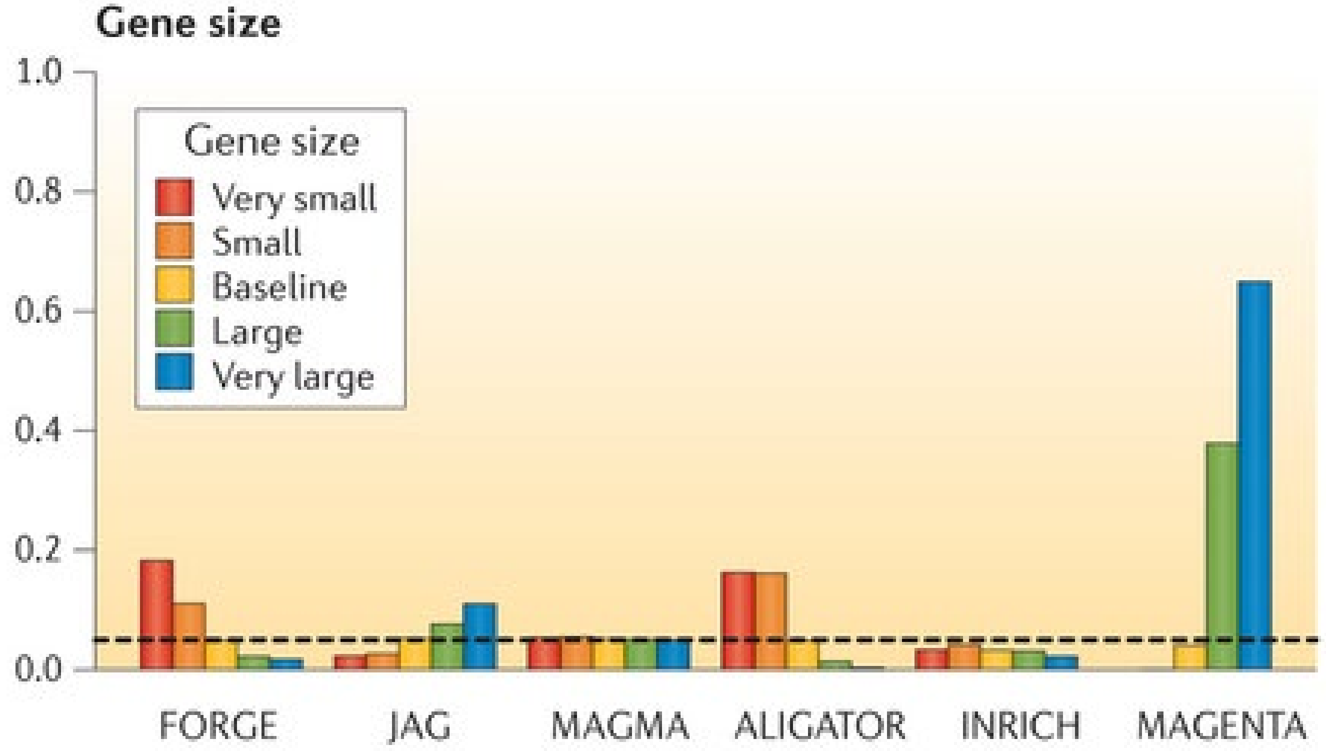
De Leeuw, Neale, Heskes, Posthuma. Nat Rev Genet, 2016

Different statistical algorithms test different alternative hypotheses

How to estimate gene association from SNP associations?

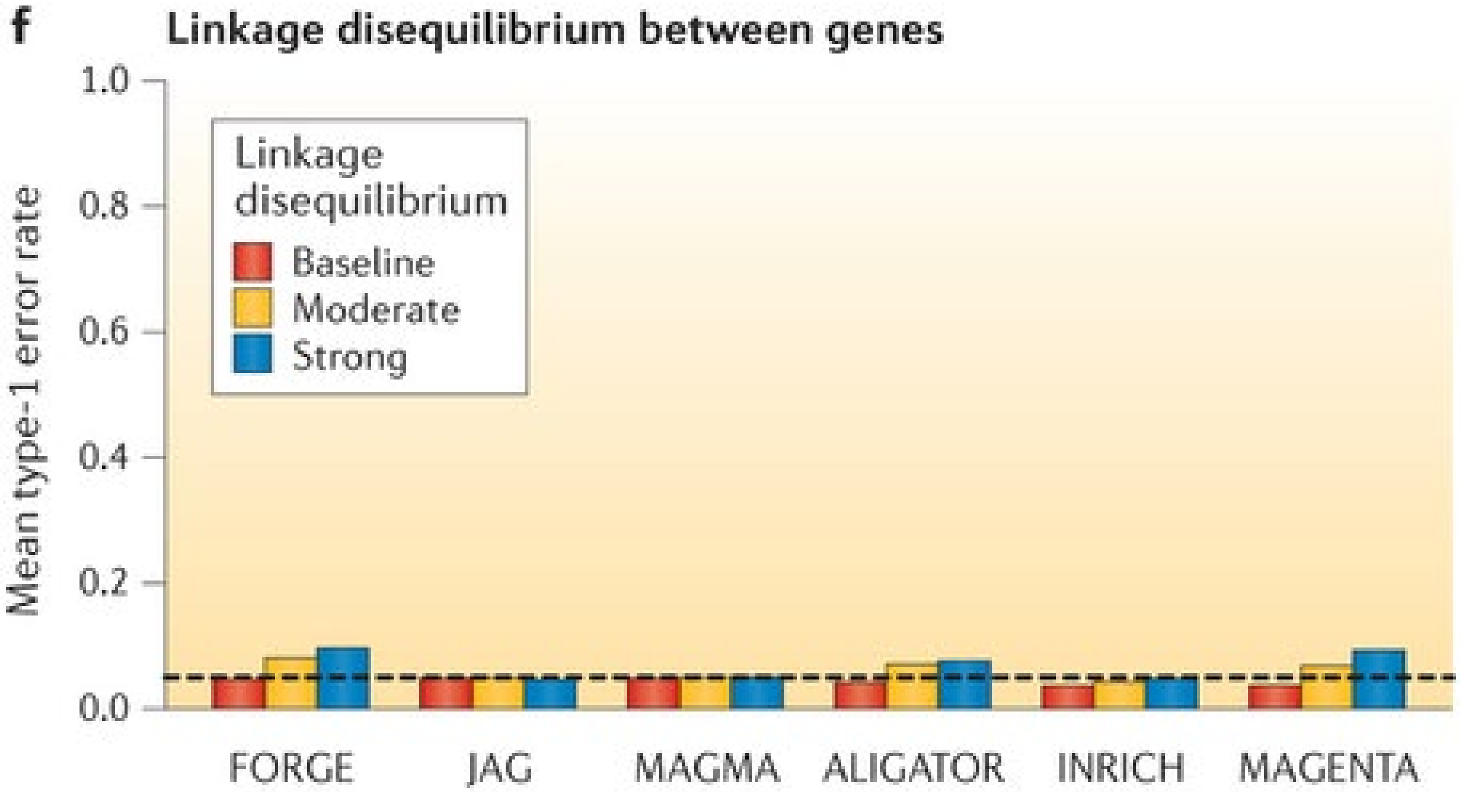
Strategy	Alternative hypothesis
Minimal P-value	At least one SNP in the gene or gene-set is associated with the trait
Combined P-value	The combined pattern of individual P-values provides evidence for association with the trait

Different tools are differentially affected by gene size



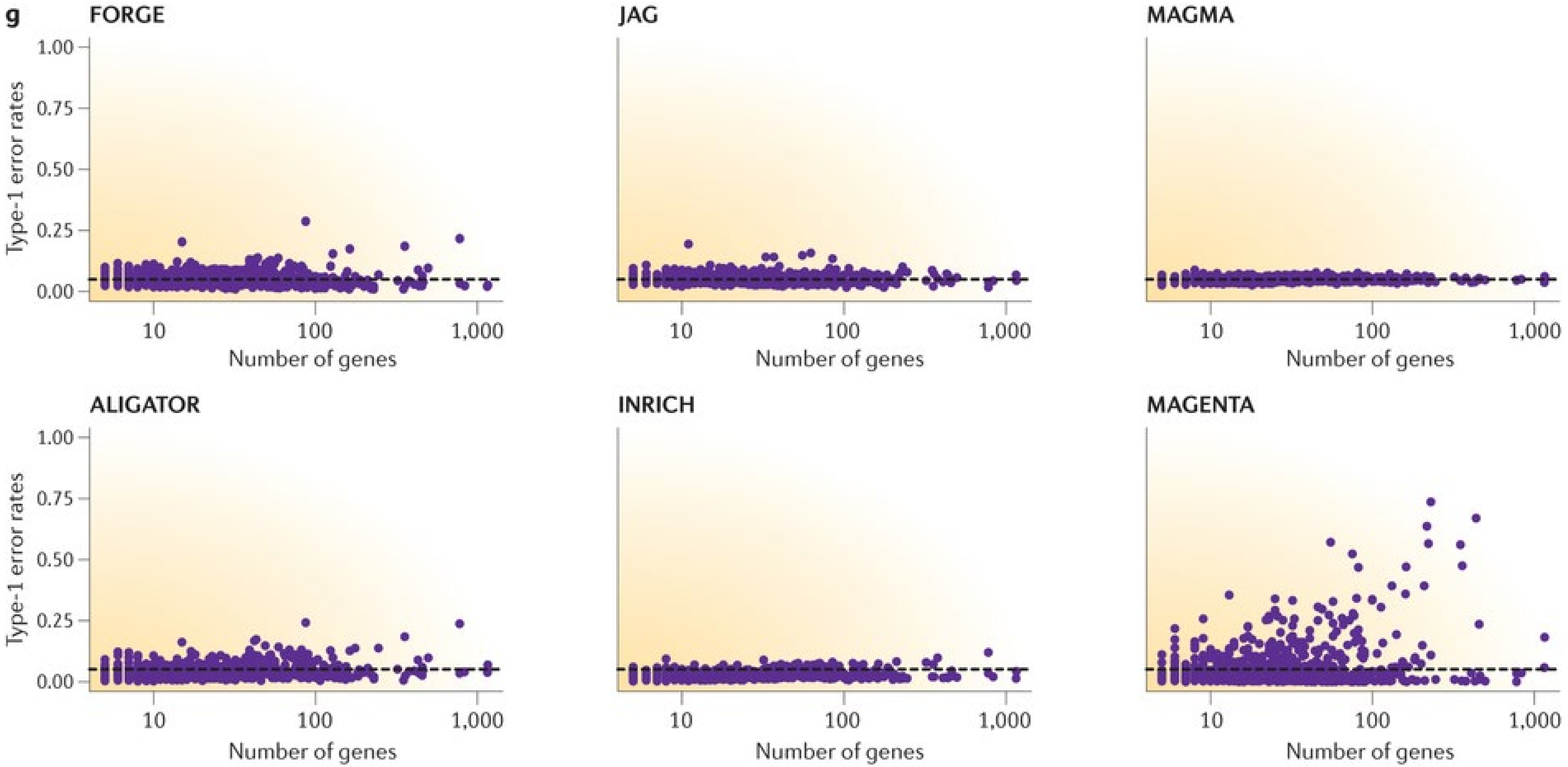
De Leeuw, Neale, Heskes, Posthuma. Nat Rev Genet, 2016

Different tools are differentially affected by LD between genes



De Leeuw, Neale, Heskes, Posthuma. Nat Rev Genet, 2016

Different tools are differentially affected by the number of genes

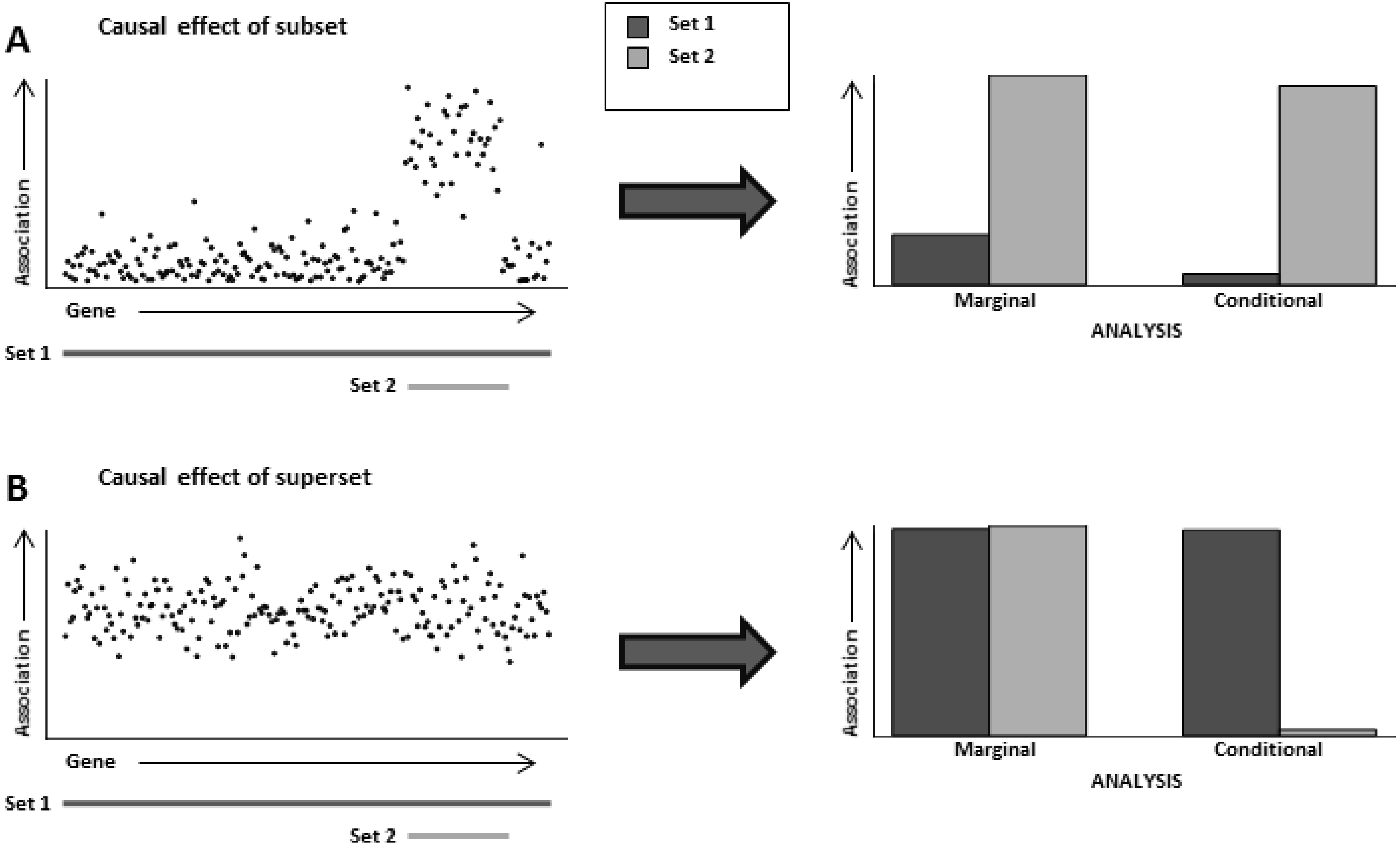


De Leeuw, Neale, Heskes, Posthuma. Nat Rev Genet, 2016

Issues of interpretation in gene-set analyses

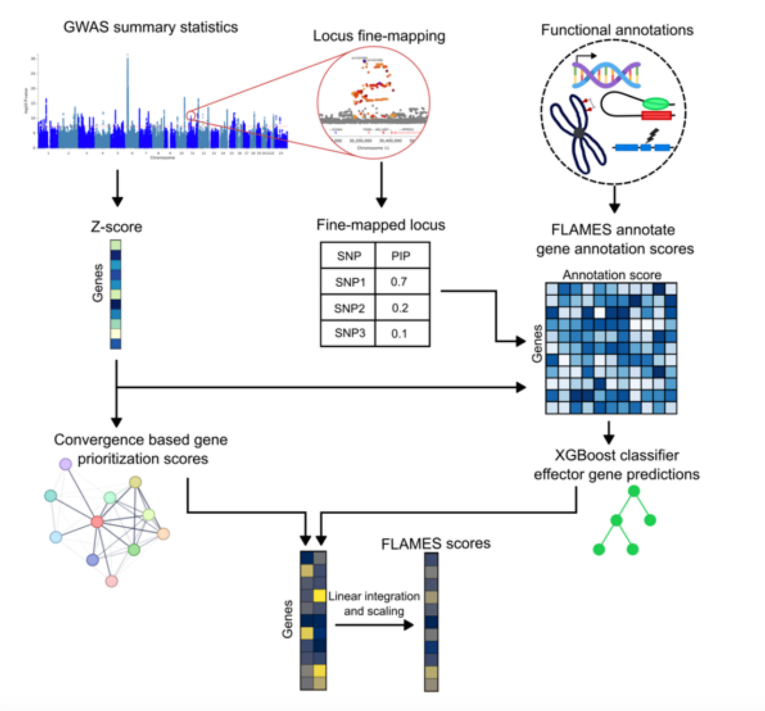
GSA tests for accumulation of genetic association in the set, which may be because:

- **Direct effect:** the set (or biological function) itself is involved
- **Confounding:** the set itself is not involved, but many genes in the set overlap with genes in another set that is involved

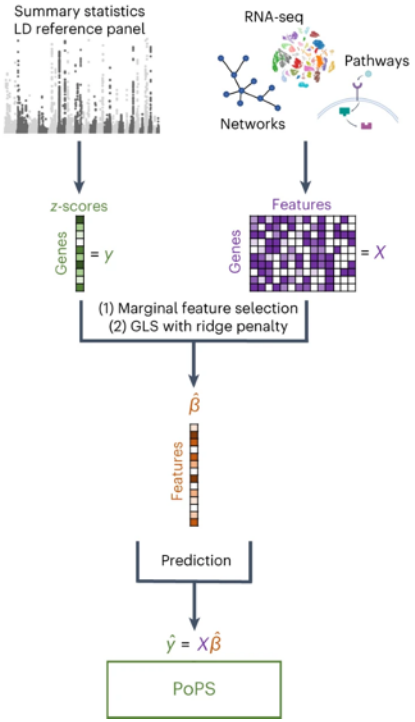

A
Causal effect of subset
Association
Gene
Set 1
Set 2
Set 1
Set 2
Association
Marginal
Conditional
ANALYSIS
B
Causal effect of superset
Association
Gene
Set 1
Set 2
Association
Marginal
Conditional
ANALYSIS

Applications of gene-set analysis

- Gene prediction tools
  - PoPS https://doi.org/10.1038/s41588-023-01443-6
  - PIGEAN https://youtu.be/b1fmzhgE3lI?si=2EWKQ7Y9U77uemH2
  - FLAMES https://doi.org/10.1038/s41588-025-02084-7



https://doi.org/10.1038/s41588-025-02084-7



https://doi.org/10.1038/s41588-023-01443-6

Practical

Developed and maintained by
Christiaan de Leeuw

Practical

1. Annotate SNPs to genes
2. Perform gene analysis (with 10 PCs as covariates)
3. Perform gene-set analysis
4. Perform tissue expression analysis
5. Perform joint gene-set / tissue expression analysis

Practical

1. Annotate SNPs to genes
2. Perform gene analysis (with 10 PCs as covariates)
3. Perform gene-set analysis
4. Perform tissue expression analysis
5. Perform joint gene-set / tissue expression analysis

Data

- Simulated GWAS data and phenotype; 400K SNPs, N = 2,500
- 1011 Reactome gene sets
- Tissue-specific expression data for 11 tissues
  - Simulated, but based on real expression data

Practical

- Open terminal window
- Make folder for practical and copy files
  - `mkdir thursday_magma`
  - `cd thursday_magma`
  - `cp /home/douglasw/Boulder2025/magma_session.zip .`
  - `unzip magma_session.zip`
- Questions/instructions are here https://vuamsterdam.eu.qualtrics.com/jfe/form/SV_3f5d2iC6AvneNr8
- Instructions are also in `instructions.txt` file

Practical - key points

- Step 1: annotation
  - Out of 19,427 protein-coding genes in the gene location file, only 13,772 had any SNPs annotated to them
    - Restricts any conclusions to the annotated genes, we cannot be sure whether the same relations hold in the other genes

- Step 2: gene analysis
  - Two genes are genome-wide significant
    - Threshold = 0.05/13,772 = 3.63e-6
  - Only 6.22% of genes have a p-value below 0.05
    - Would expect 5% by chance, so only modest genetic signal in data

Practical - key points

- Step 3a: basic competitive gene-set analysis
  - Out of 1013, there are 10 significant gene sets
    - Suggests that the underlying properties (known pathway, cell function, biological process, etc.) may play a role in the phenotype
    - Looking at the names, probably overlap between these gene sets
      - Use conditional gene-set analysis to improve specificity

  - For first significant gene-set (SIGNALING_BY_NOTCH1_T)
    - Lowest gene p-value is 0.00035, so not genome-wide significant
    - But: 28.3% of genes have a p-value below 0.05
      - Much higher than the 6.22% genome-wide
    - Gene-set association is driven by larger number of modestly associated genes

Practical - key points

- Step 3b: conditional competitive gene-set analysis
  - 6 out of 9 gene-sets are no longer significant after conditioning on the Critical Pathway gene-set

Set	P (step 3a)	P (step 3b)
Signaling by Notch1 T	1.08e-6	9.32e-7
Constitutive Signaling by Notch1 HD + Pest Domain Mutants	1.02e-5	9.02e-6
Elastic Fibre Formation	6.71e-7	0.135
Activation of the Phototransduction Cascade	8.20e-6	0.052
The Phototransduction Cascade	4.27e-9	0.143
Notch1 Intracellular Domain Regulates Transcription	3.65e-5	3.27e-5
Inactivation Recovery And Regulation of the Phototransduction Cascade	1.18e-9	0.058
Molecules Associated with Elastic Fibres	4.86e-5	0.857
Another Critical Pathway	3.05e-12	0.153
Critical Pathway	3.17e-12	-

Practical - key points

- Step 3b: conditional competitive gene-set analysis
  - 6 out of 9 gene-sets are no longer significant after conditioning on the Critical Pathway gene-set
  - Conversely, for 5 of these 6 sets, Critical Pathway remains significant when conditioning on that set, suggesting that
    - Of these sets, the Critical Pathway set is most likely to be the true ‘causal’ gene set
    - The originally observed associations of the 5 sets that are no longer significant are driven entirely by their overlapping with this causal set
  - For Another Critical Pathway, both it and Critical Pathway no longer significant
    - Likely a single underlying signal, but too much overlap to determine which of the two sets is more likely the relevant one

Practical - key points

- Step 4a: basic tissue expression analysis
  - All the tissue expression levels are significant, as is the mean expression level across tissues
    - In all likelihood, the associations per tissue are driven by the more general relation between gene expression and genetic association; not very informative

- Step 4b: conditional tissue expression analysis
  - Only the brain-specific expression level remains significant after conditioning on average gene expression level
    - More strongly (specifically) brain-expressed genes also tend to be more strongly associated with our phenotype; suggests that brain expression plays a role in (the genetics of) our phenotype

Practical - key points

- Step 5: joint gene set and gene expression analysis
  - The p-values remain effectively the same when conditioning on the average gene expression level, as well as when additionally conditioning brain-specific expression level
  - This suggests that the gene-set associations are not driven merely by gene expression effects (at least of the tissues we tested), which helps strengthen our interpretation of the gene-set associations

Practical - conclusion

- Full answer file and all output:
  - `/home/douglasw/Boulder2025/magma_answers.zip`

- Any further questions?
  - MAGMA program, manual and auxiliary files can be found on the MAGMA site: http://ctglab.nl/software/magma
  - Contact for questions, suggestions, etc. at d.p.wightman@vu.nl