Geneset analysis

Danielle Posthuma & Douglas Wightman

And thanks to Christiaan de Leeuw

CTGlab – VU Amsterdam

shared drive Danielle/MAGMA_new2023.ppt

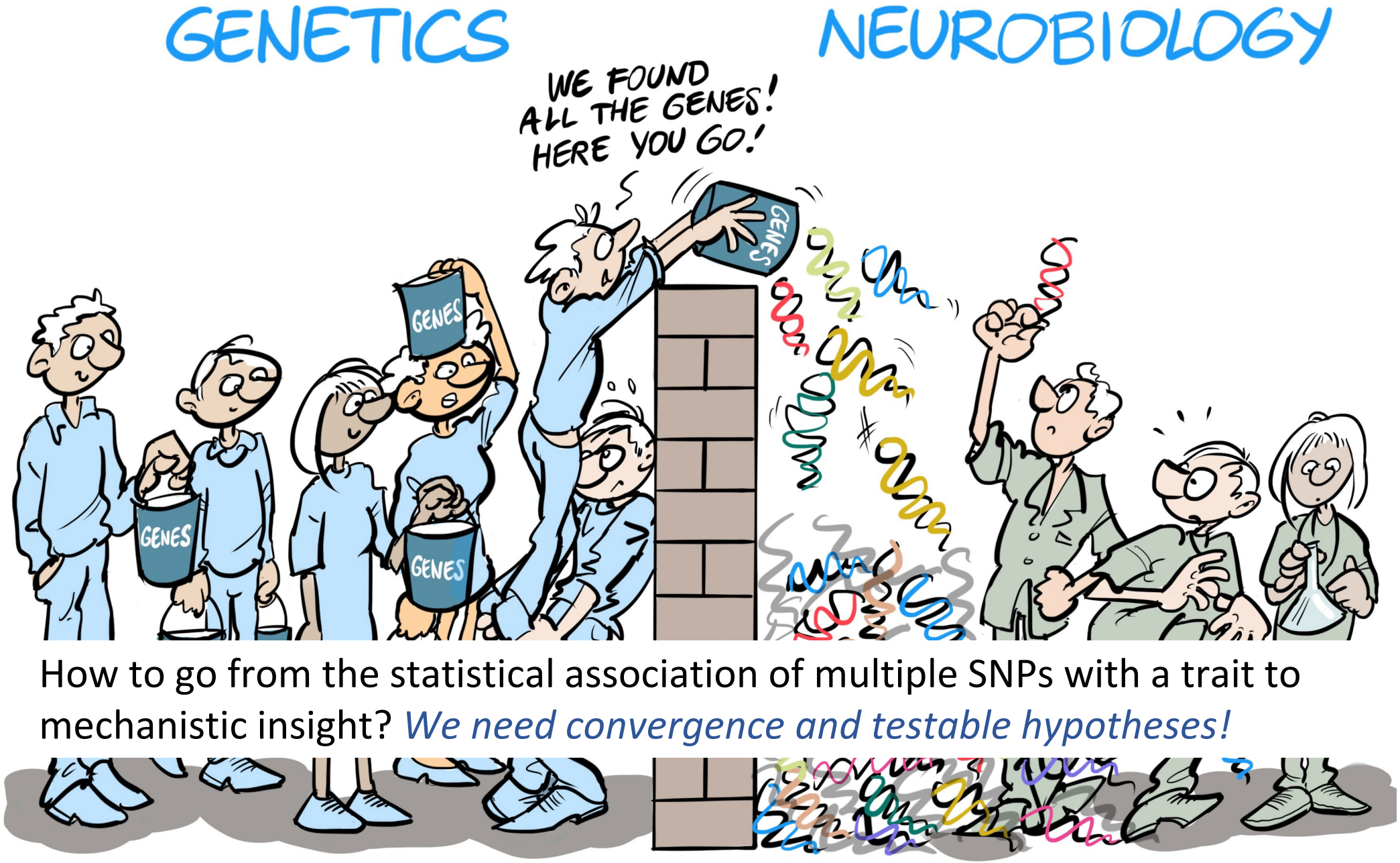
To avoid straining the system:

- `mkdir thursday_magma`
- `cd thursday_magma`
- `cp /home/christiaan/Boulder2023/magma_session.zip .`

Challenges in interpreting GWAS outcome

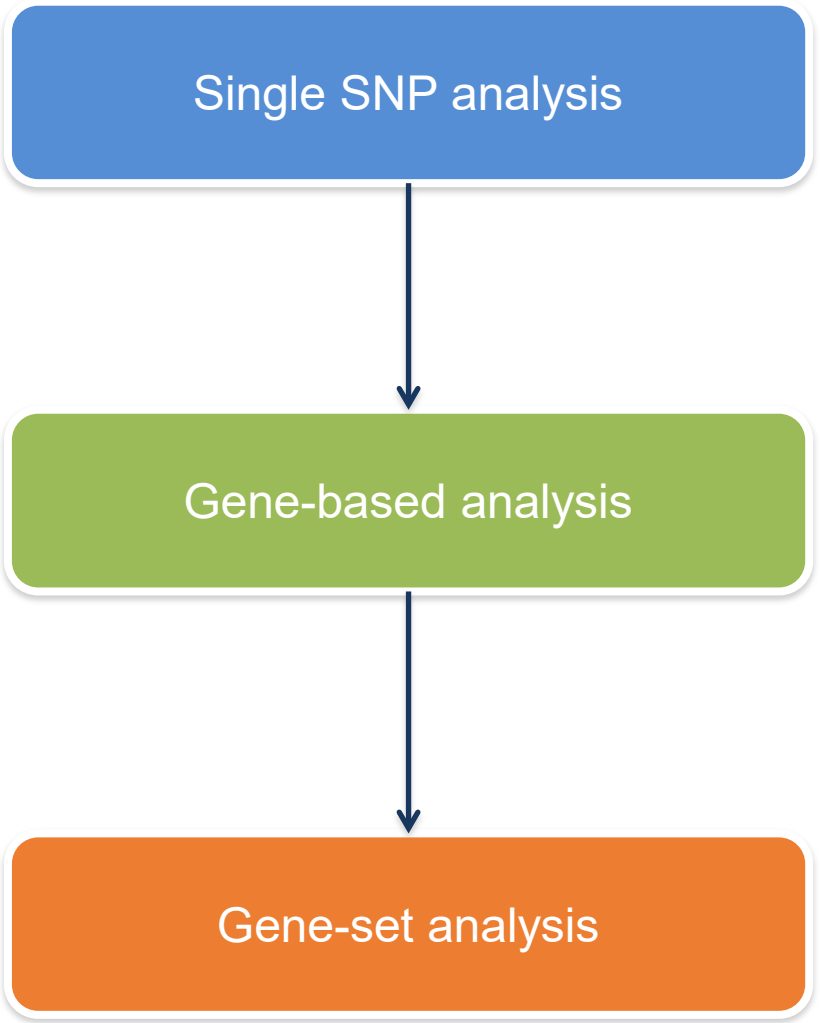
3. *Many traits are **polygenic***
multiple genetic variants of small effect contribute. A single genetic variant, even if it is known to be causal, is usually not informative for biology

Solution: map associated SNPs to genes and look for **convergence** in biological pathways, shared cellular or synaptic function, co-localization, co-expression in tissue or cell types (e.g. tools MAGMA, Ldscore regression, DEPICT)



How to go from the statistical association of multiple SNPs with a trait to mechanistic insight? *We need convergence and testable hypotheses!*

Testing for functional clustering of SNP associations

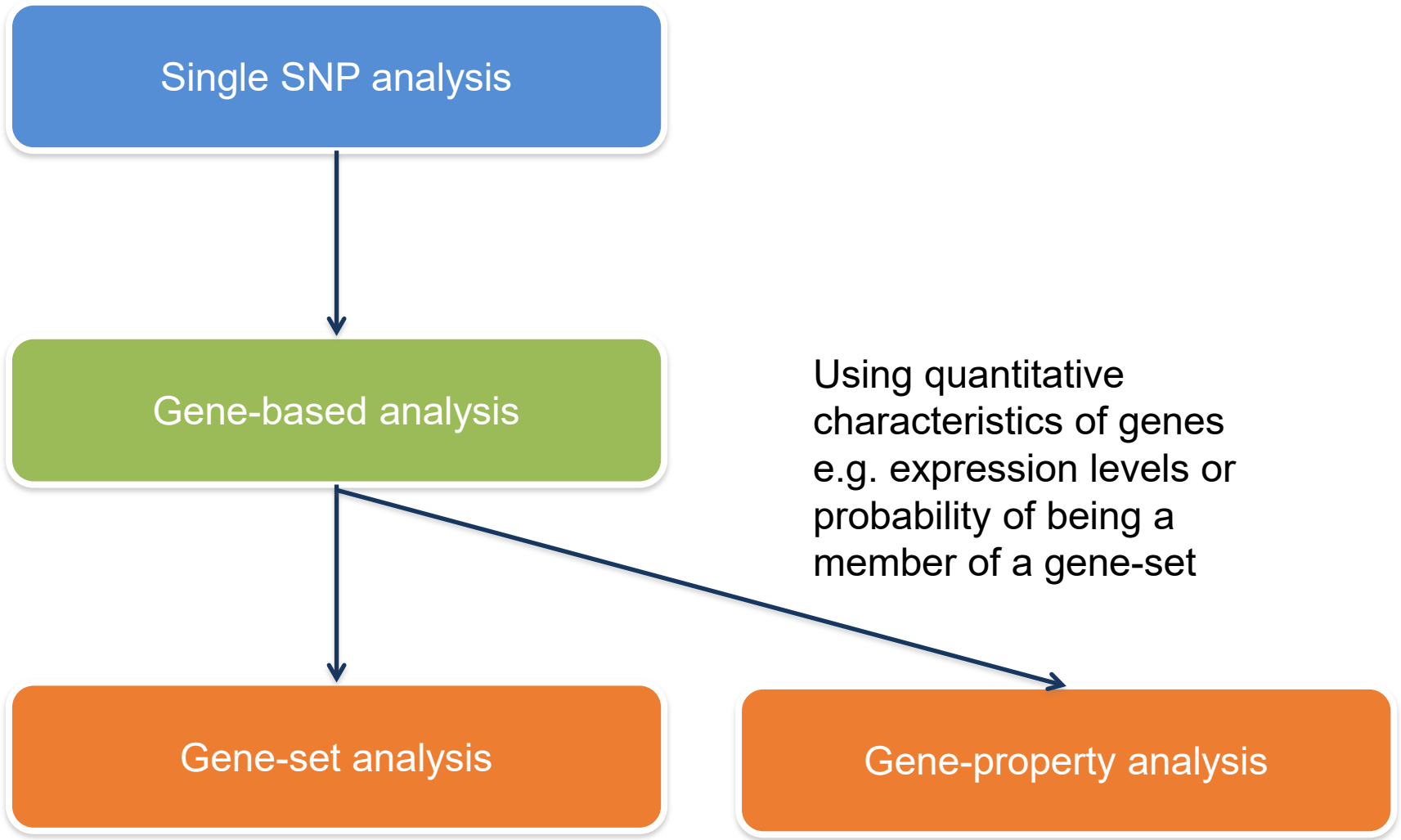


- GWAS
- single SNPs

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

Gene-set analysis with sets of genes as unit of analysis
- targeted gene-sets/pathways
- all known gene-sets/pathways

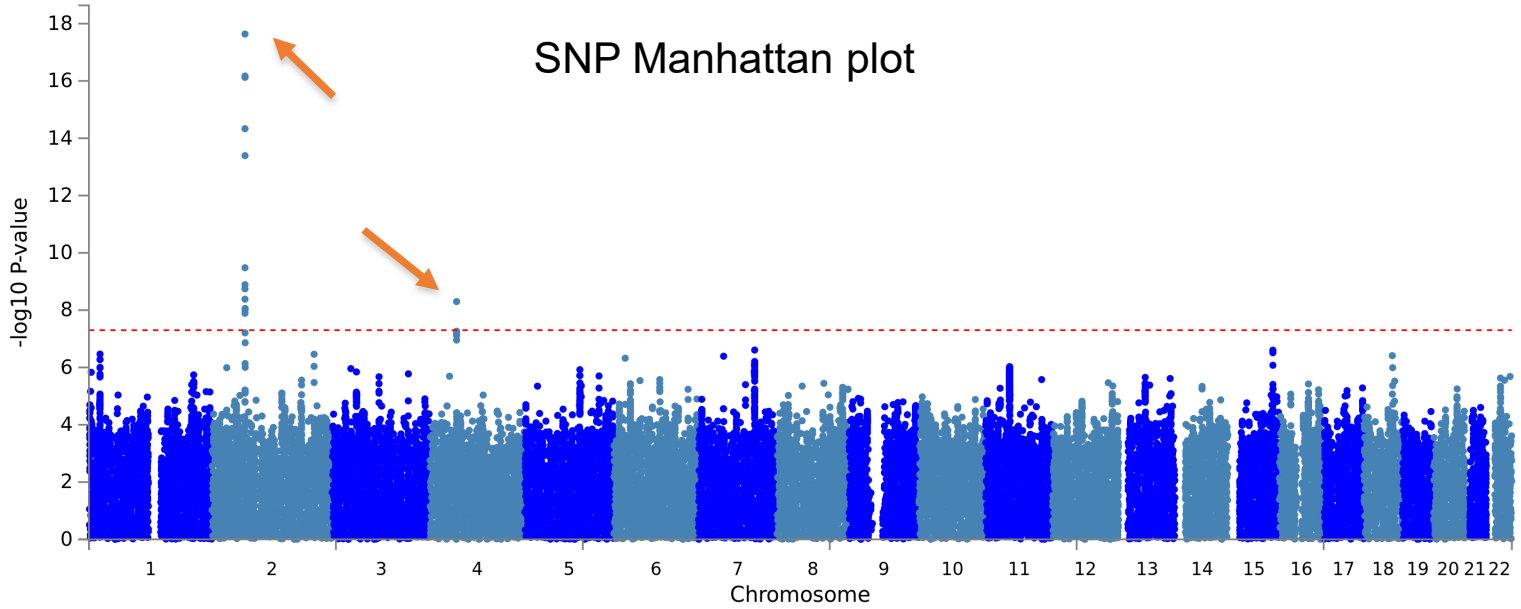
Testing for functional clustering of SNP associations

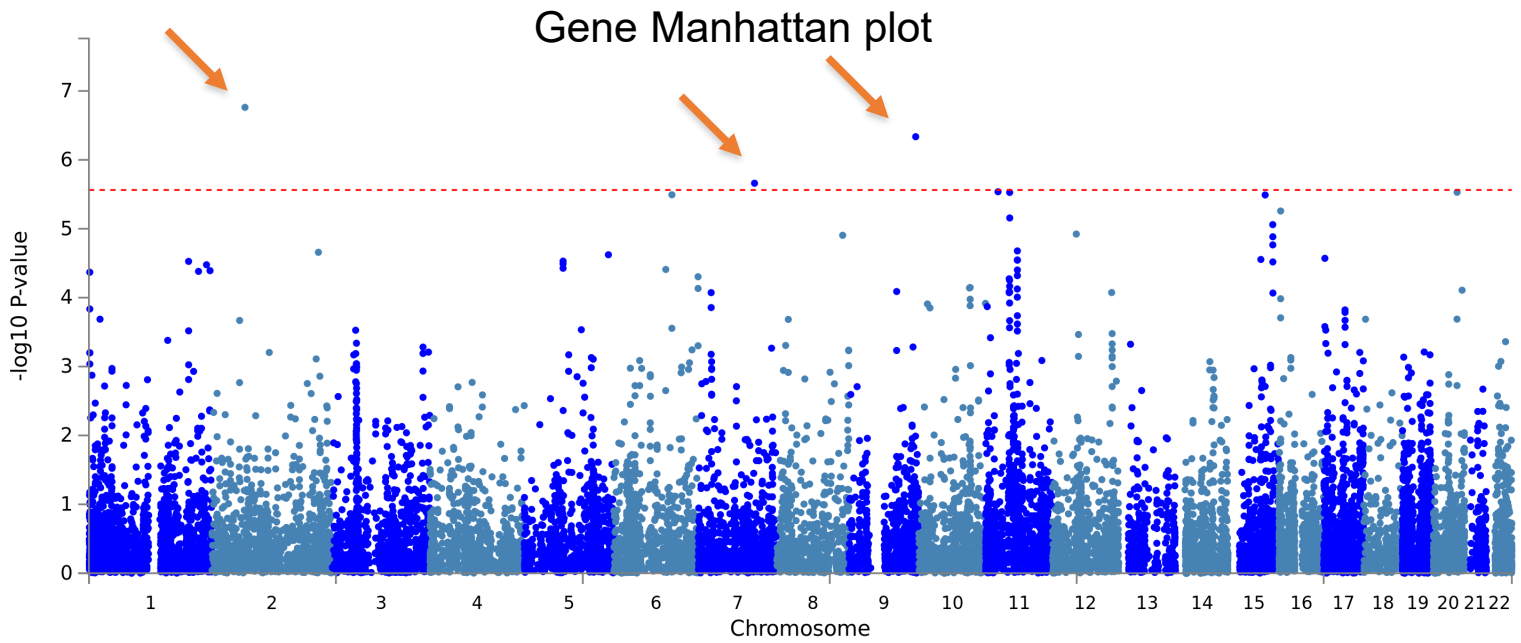


Gene-based analysis

- 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
-log10 P-value
0
2
4
6
8
10
12
14
16
18
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome


Gene Manhattan plot
-log10 P-value
0
1
2
3
4
5
6
7
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome

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:

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

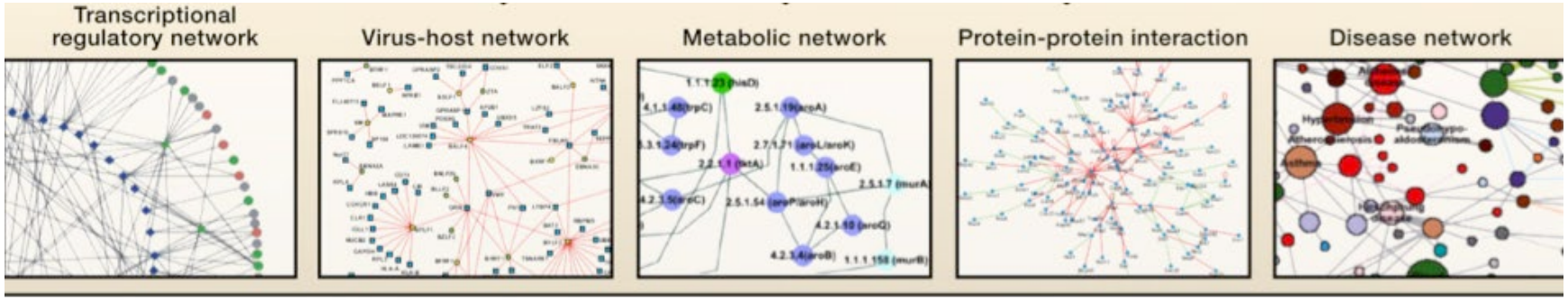
Gene-set analysis

Cons

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

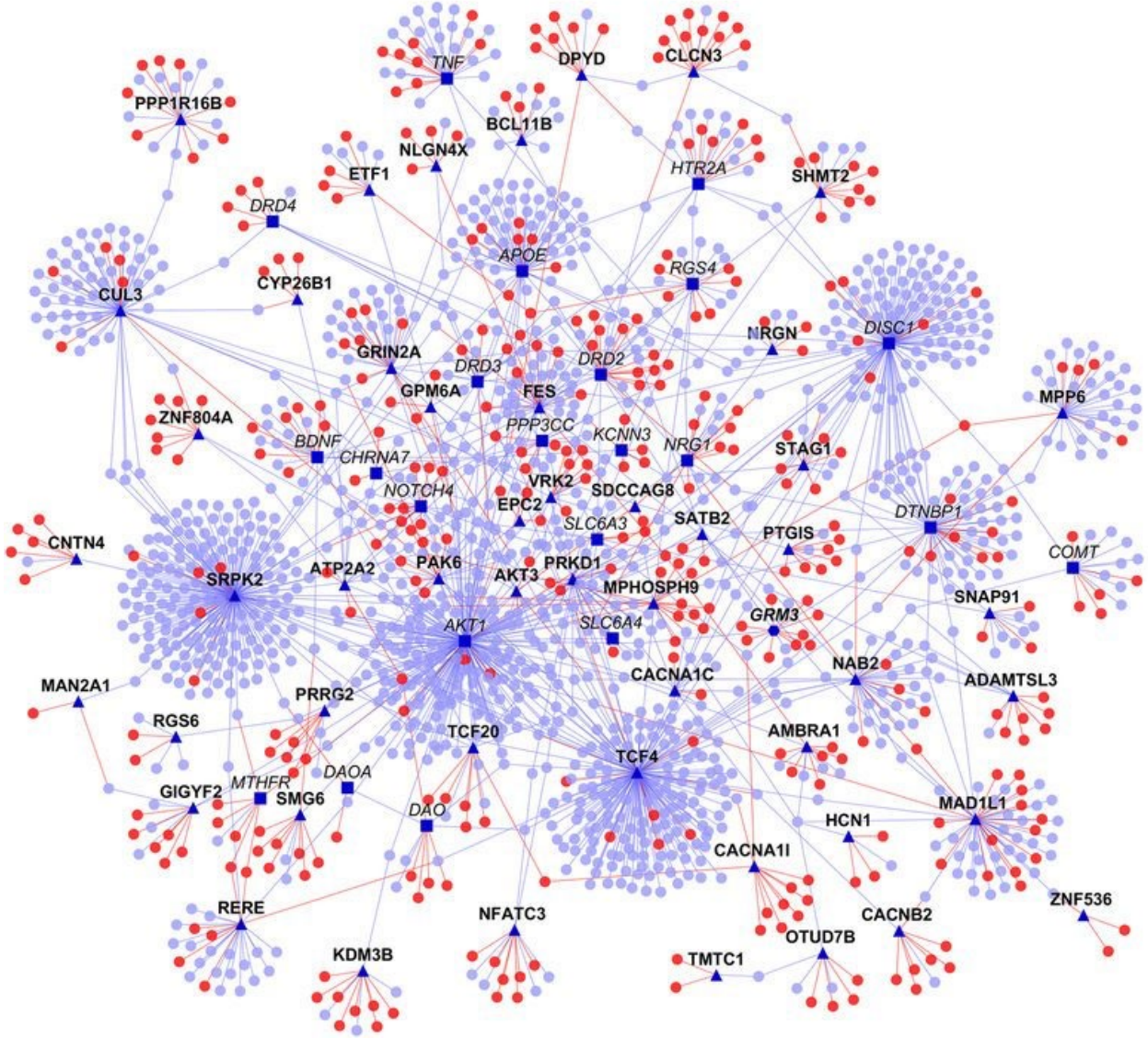
Choosing gene-sets

Gene-sets can be based on e.g.
-protein-protein interaction
-co-expression
-transcription regulatory network
-biological pathway
-Functional relations

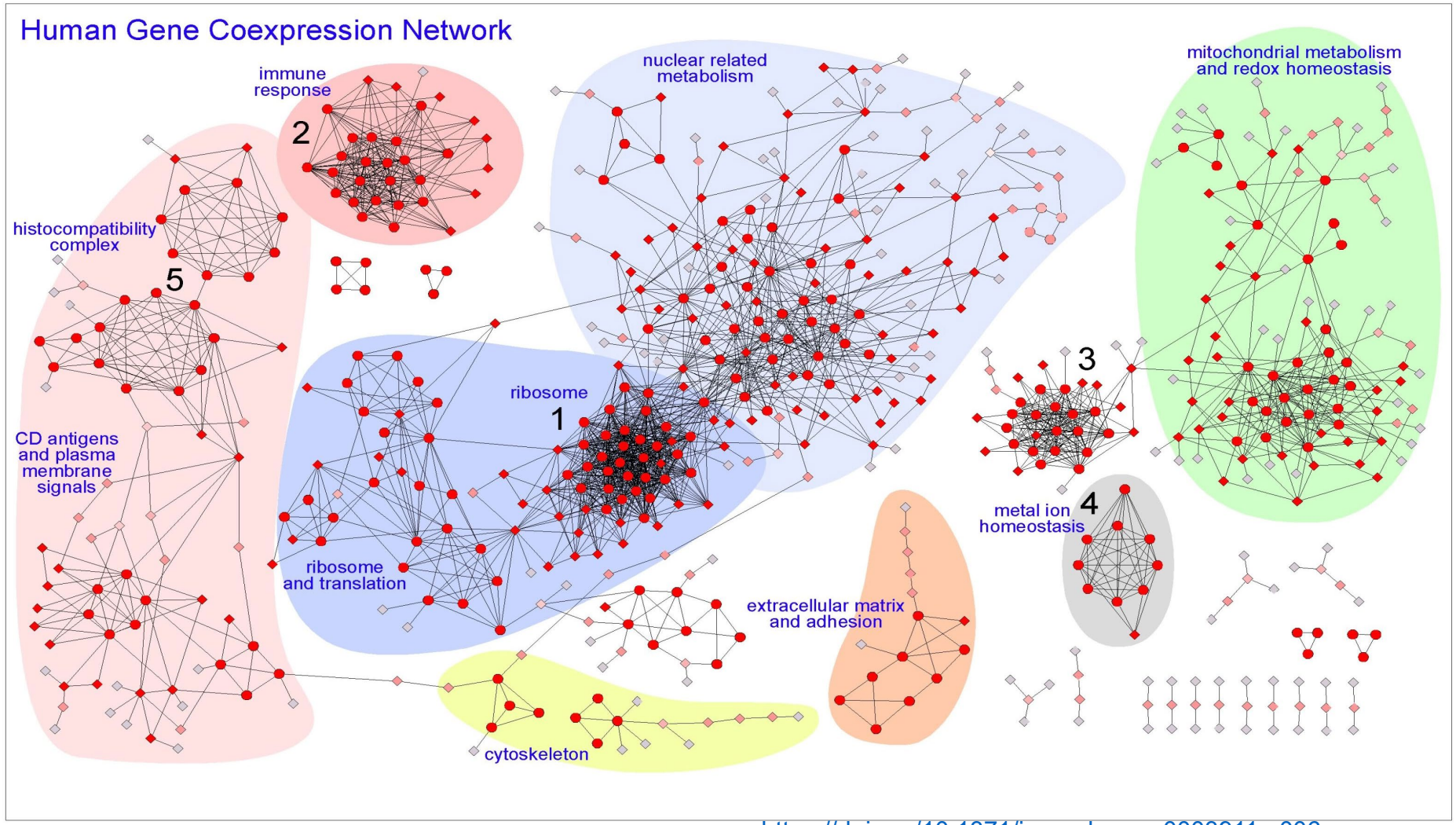


Protein interaction networks

Using Y2H or Immunoprecipitations



Co-expression networks



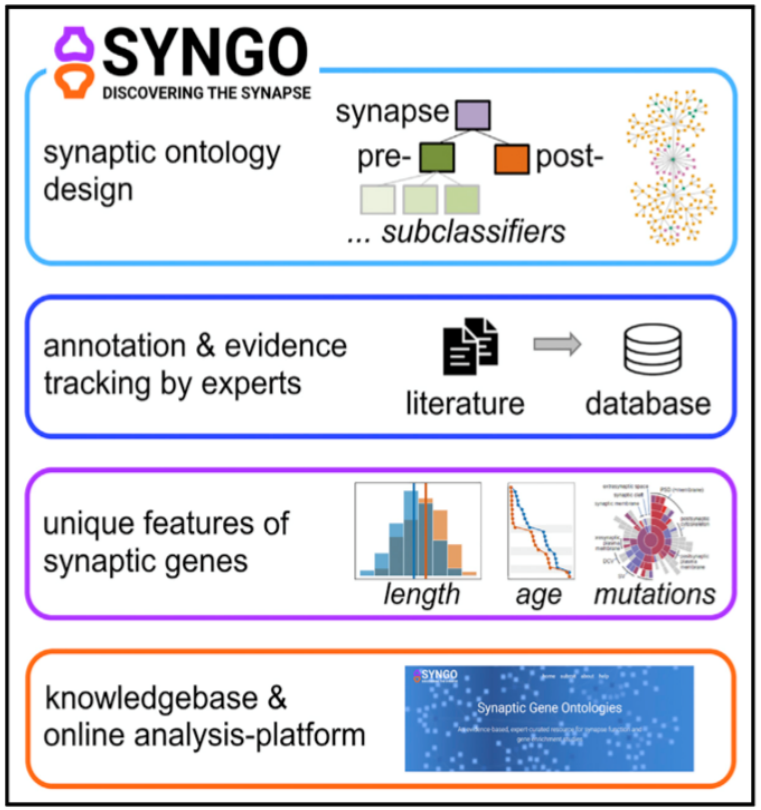
https://doi.org/10.1371/journal.pone.0003911.g006

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

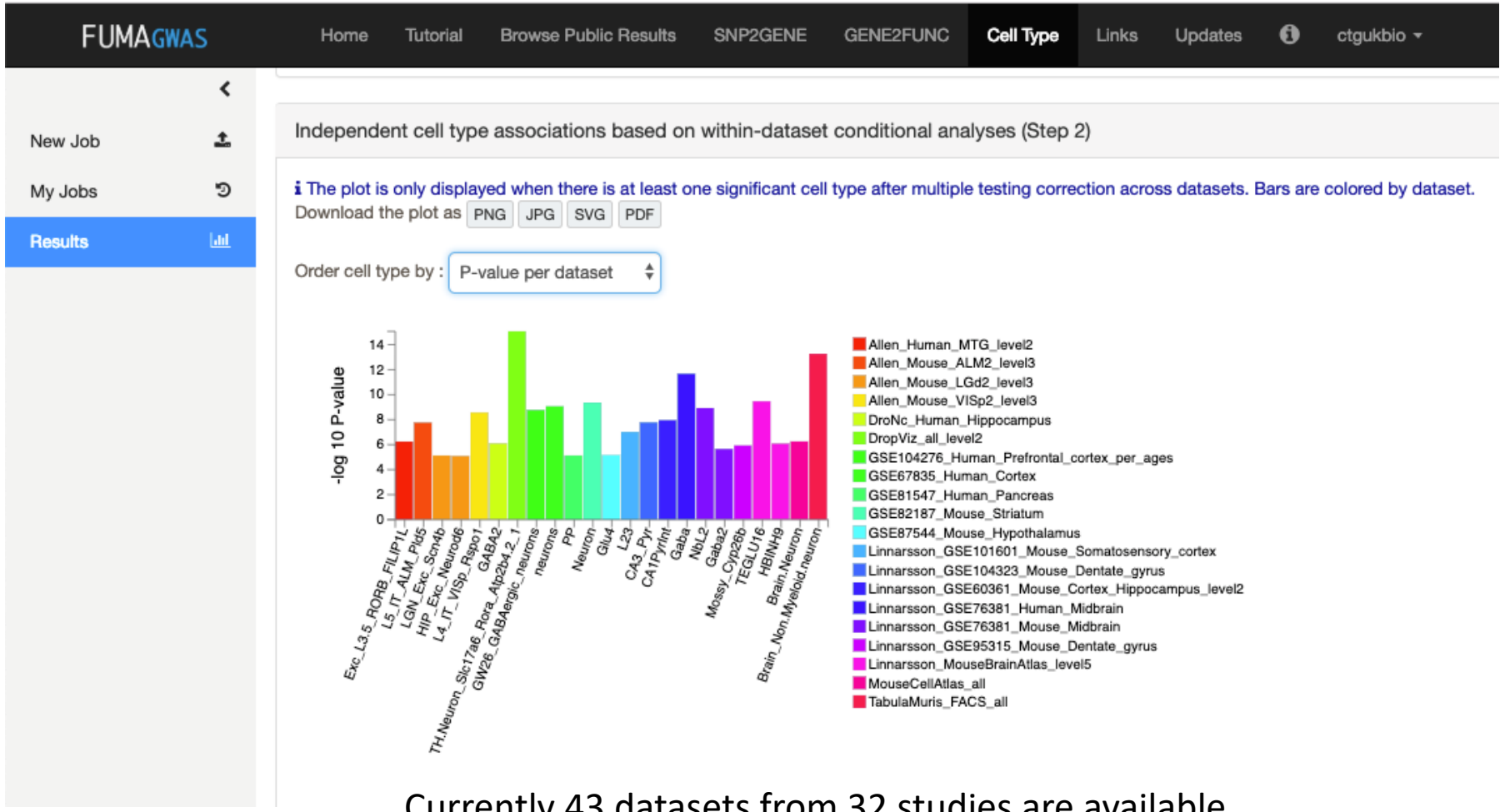
- Basically it tests whether there is an association between the association strength of genes with a trait and their expression levels in specific cell types

- *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



Currently 43 datasets from 32 studies are available

Tools for statistical analysis of gene-sets

INRICH, ALIGATOR, MAGENTA, FORGE, SETSCREEN, DAPPLE, DEPICT, MAGMA etc etc

-> do they all provide the same answer..?

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:

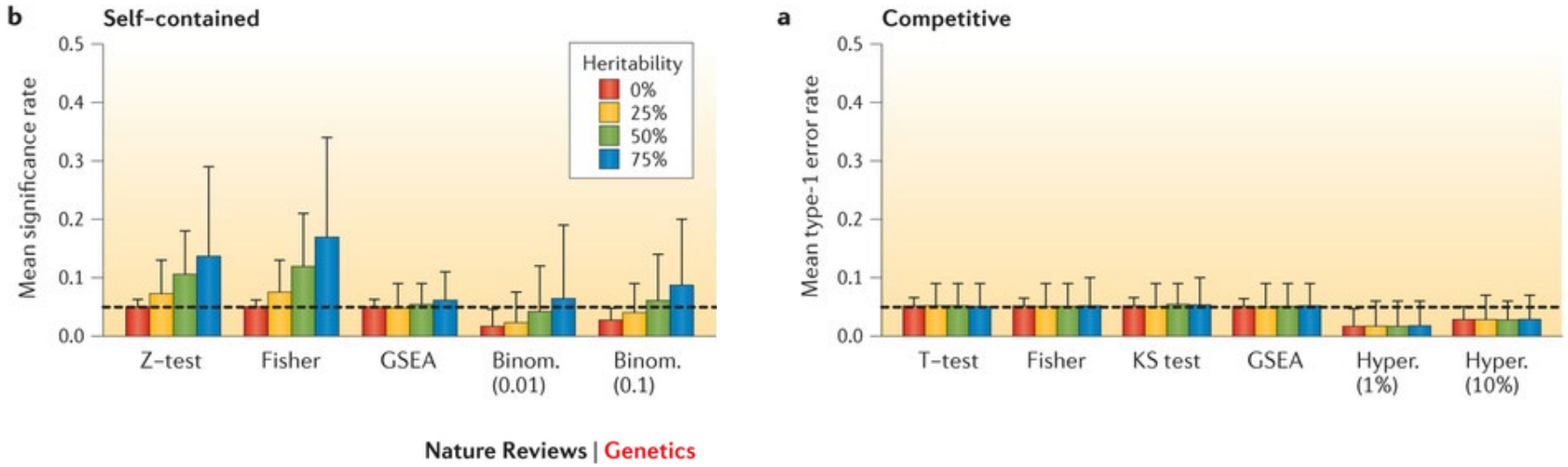
Self-contained:
H0: The genes in the gene-set are not associated with the trait

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



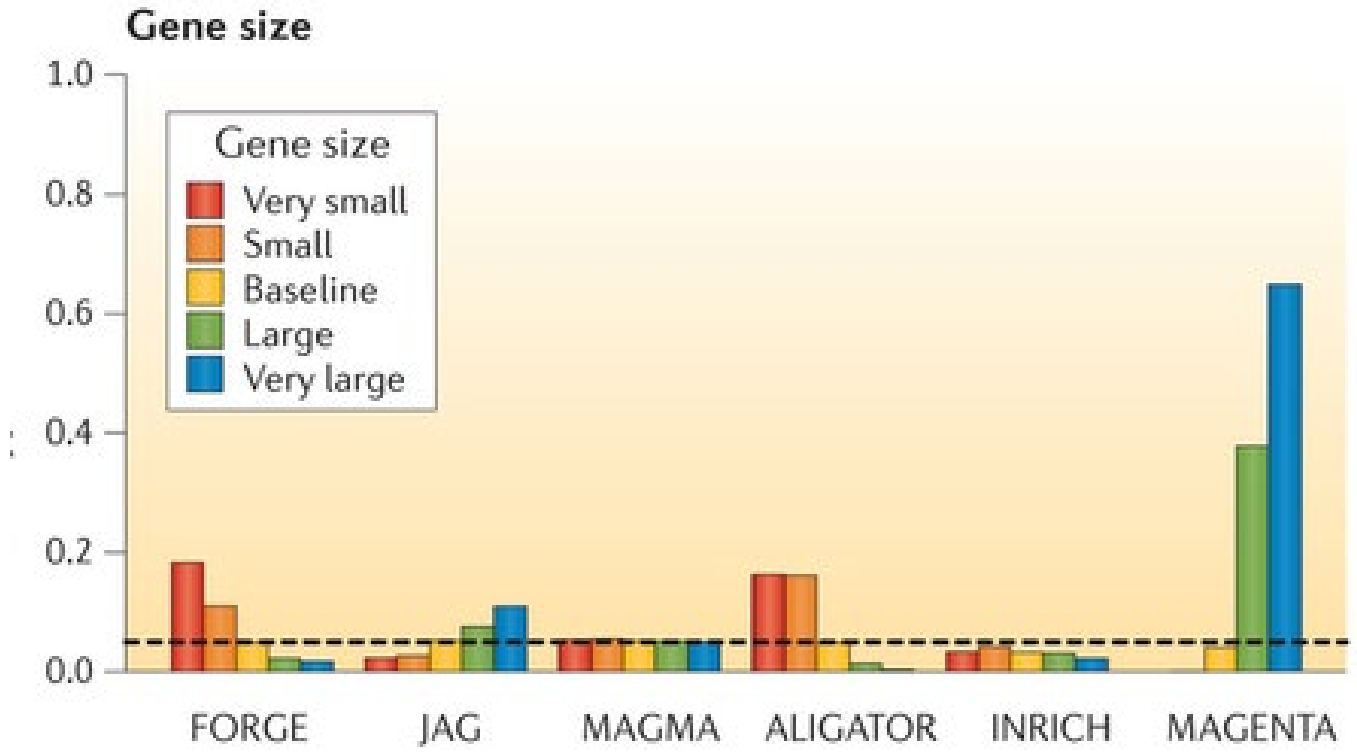
For self-contained methods, rates increase with heritability, whereas they are constant for competitive methods.

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

Different statistical algorithms test different alternative hypotheses

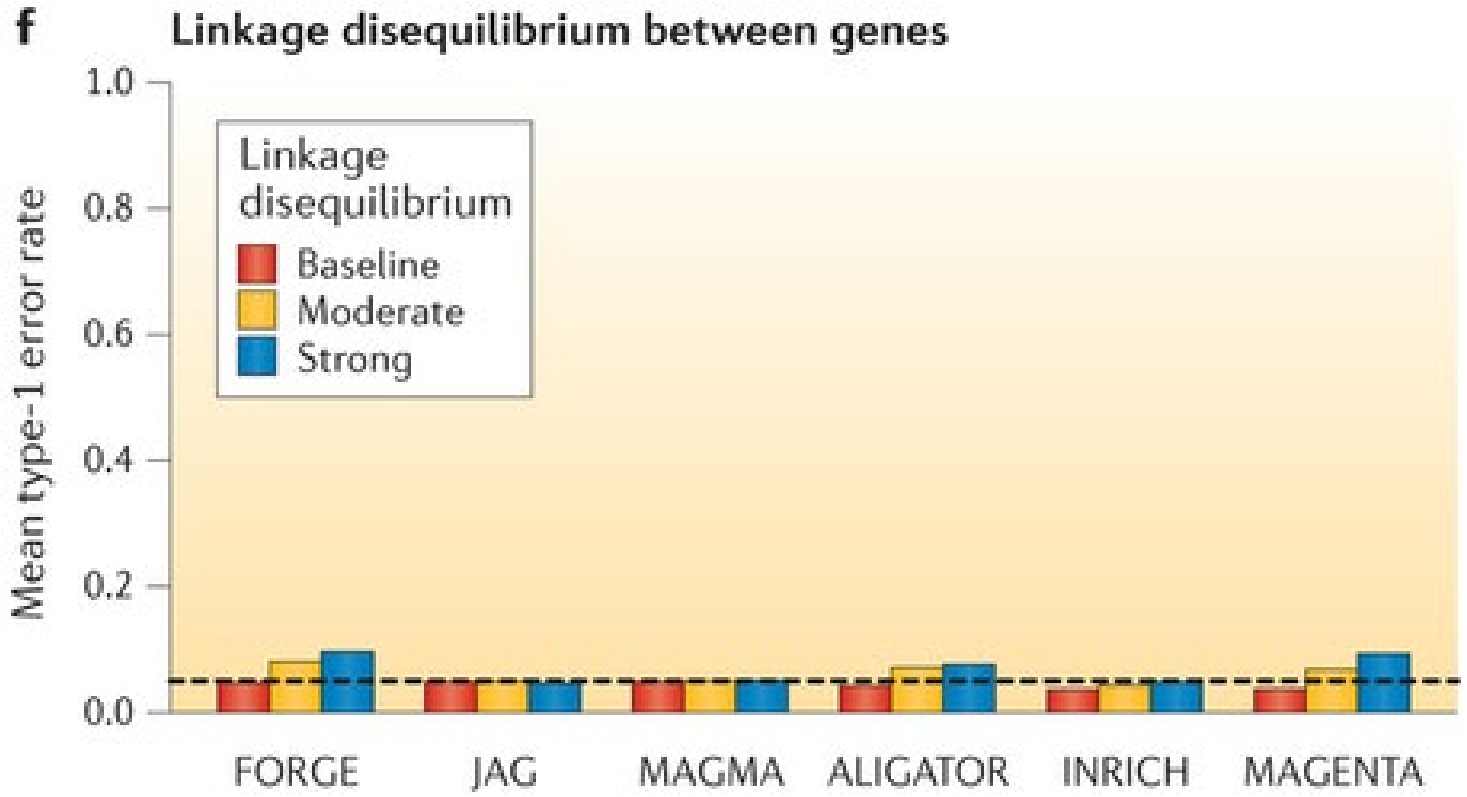
Strategy	Alternative hypothesis
Minimal P-value	At least one SNP in the gene or gene-set is associated with the trait
Combined P-value	The combined pattern of individual P-values provides evidence for association with the trait

Different tools are differentially affected by gene size



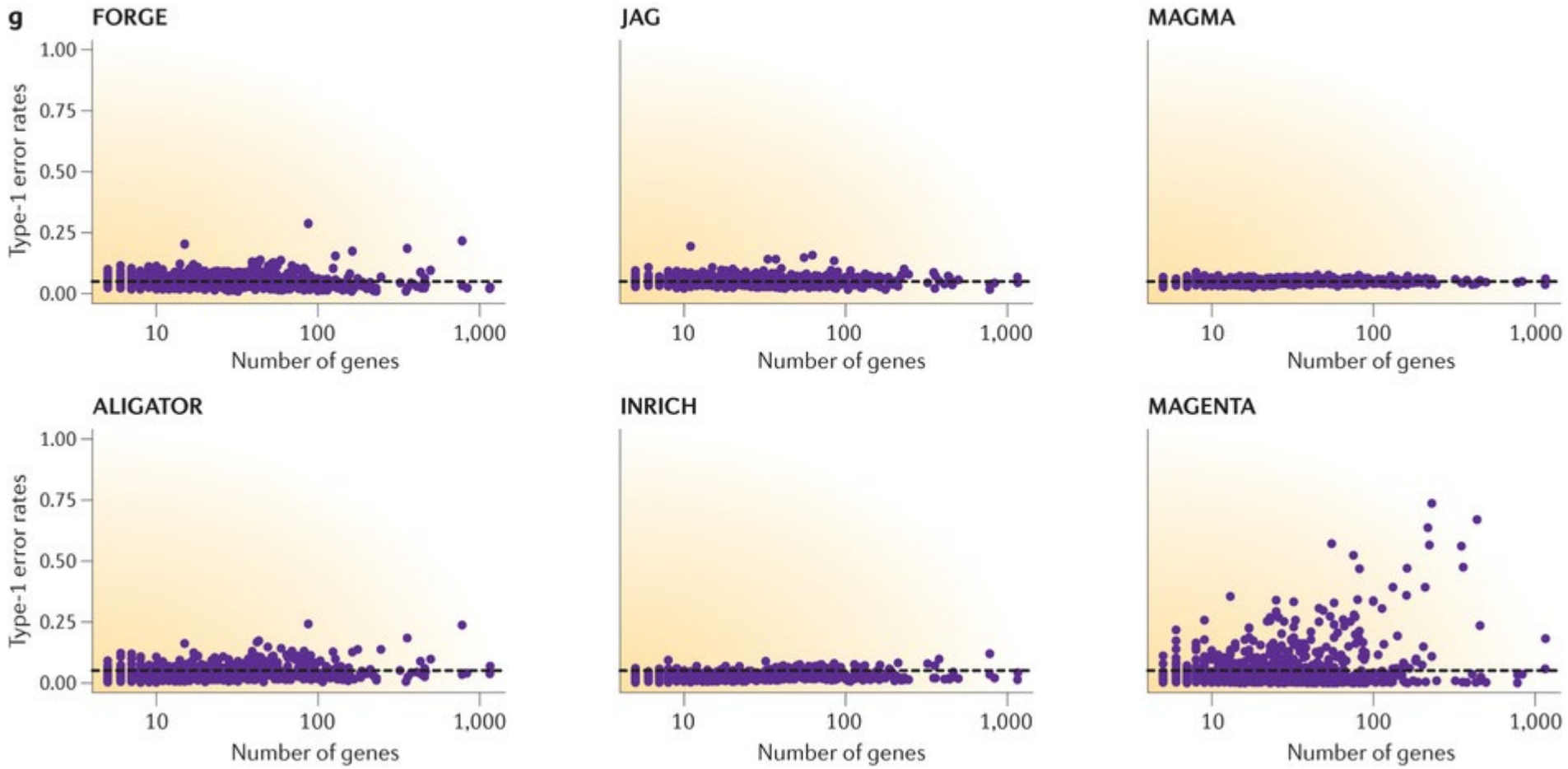
De Leeuw, Neale, Heskes, Posthuma. Nat Rev Genet, 2016

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

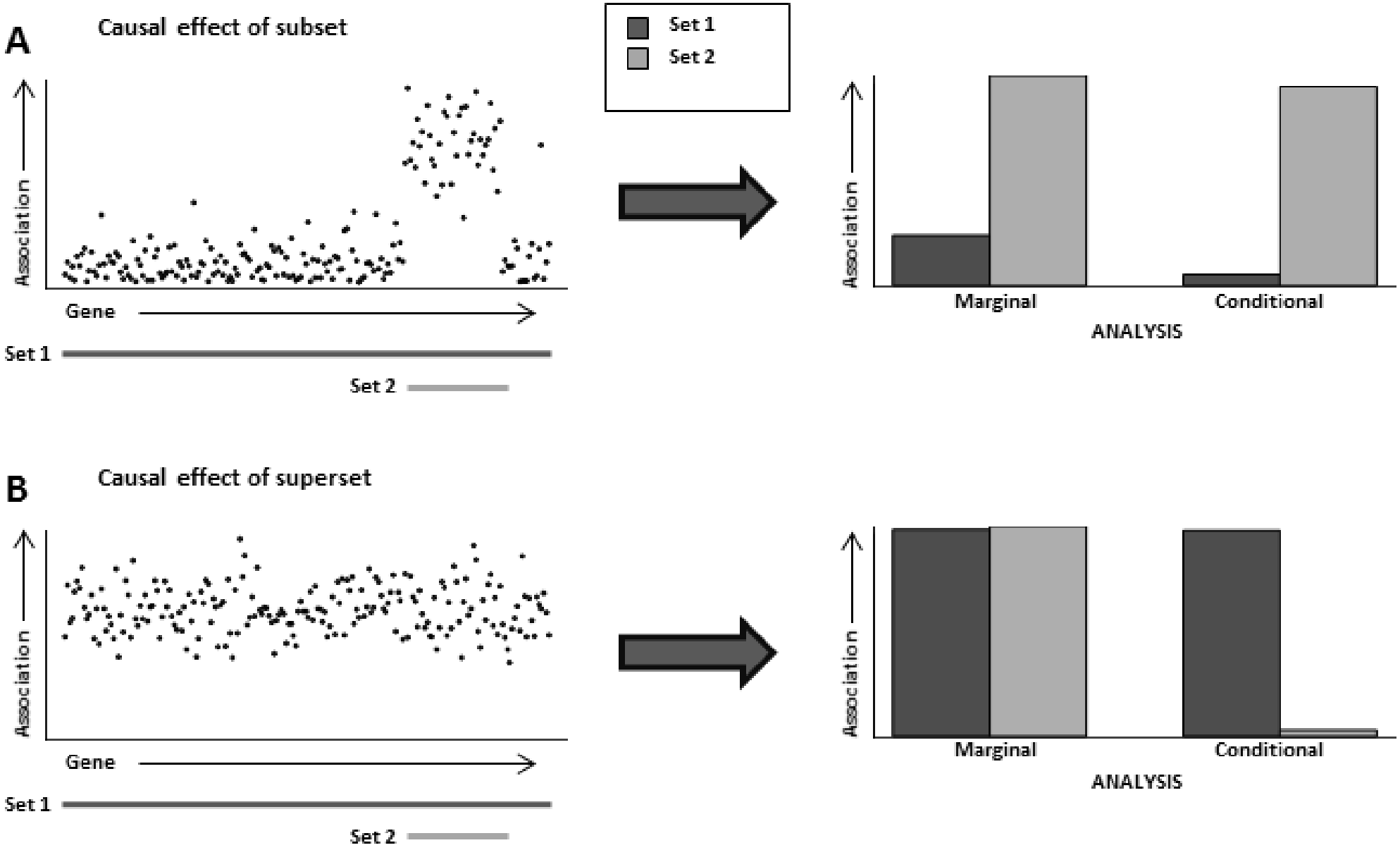


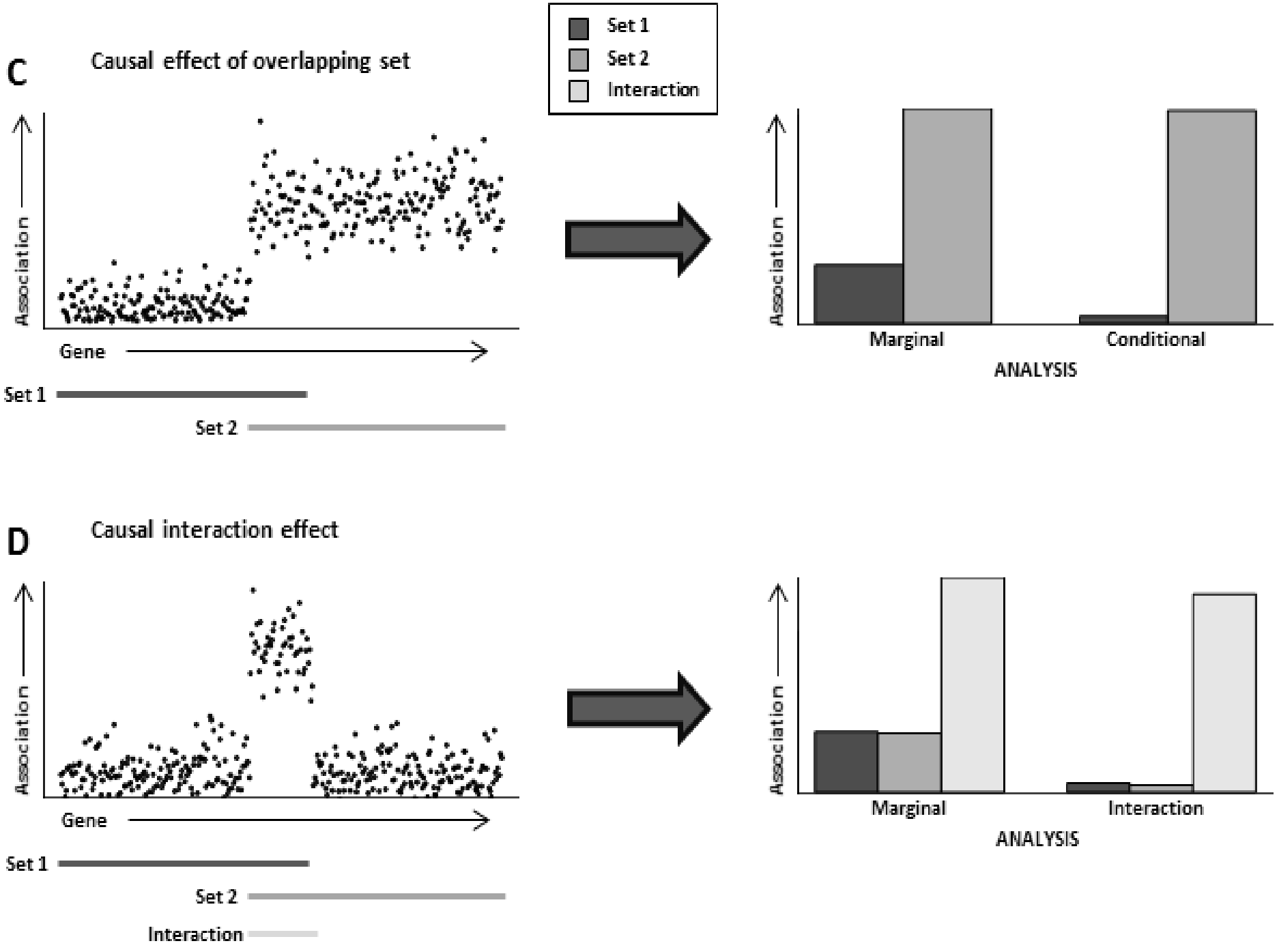
De Leeuw, Neale, Heskes, Posthuma. Nat Rev Genet, 2016

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
- **Interaction:** the set itself is partially involved, with the effect specific to a subset defined by another gene set


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

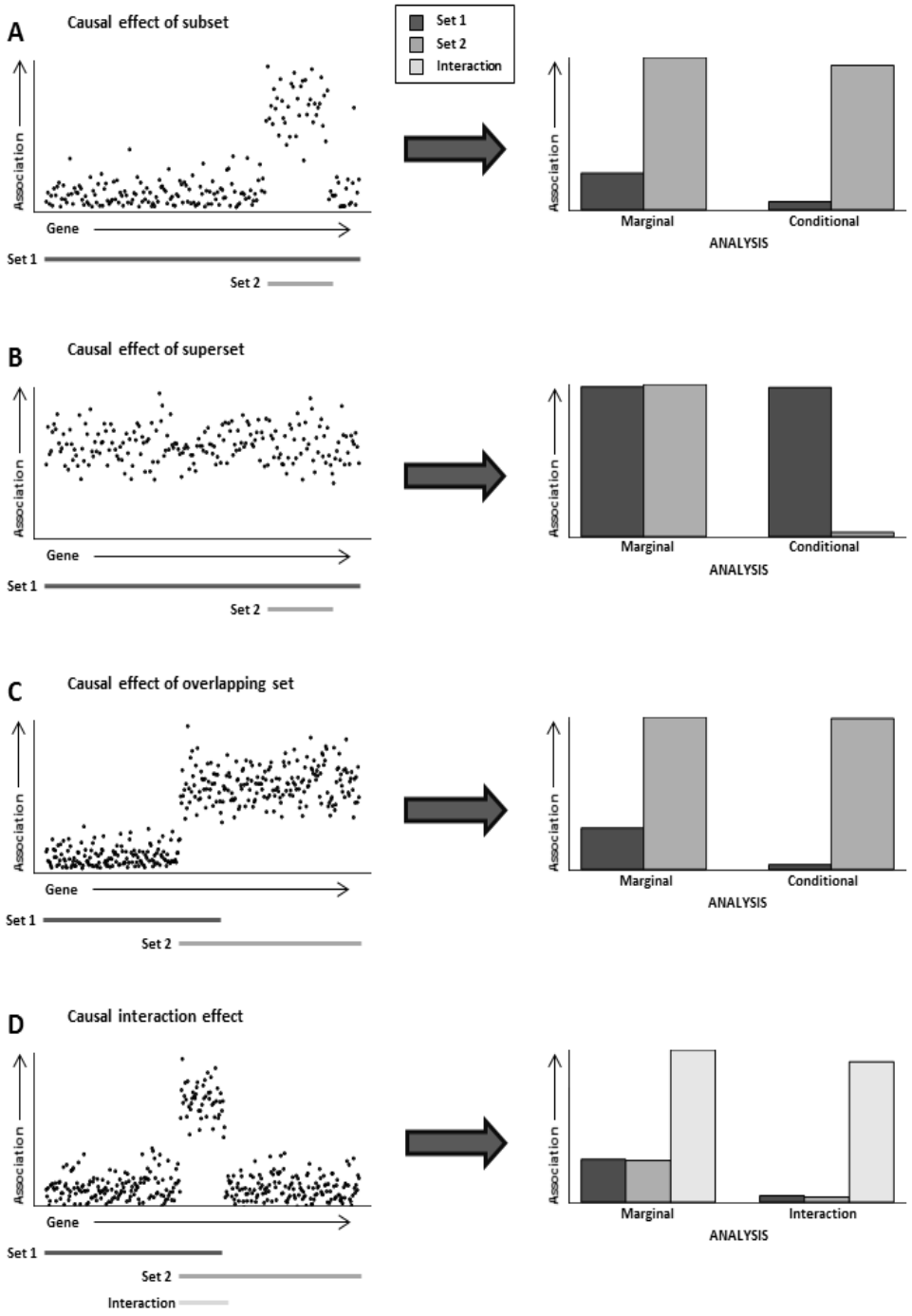

C
Causal effect of overlapping set
Set 1
Set 2
Interaction
Association
Gene
Set 1
Set 2
Association
Marginal
Conditional
ANALYSIS
D
Causal interaction effect
Association
Gene
Set 1
Set 2
Interaction
Association
Marginal
Interaction
ANALYSIS

Four general confounding scenarios (A-D)

- Overlap with actually associated set induces spurious association
- Interaction can be seen as special instance of subset confounding

Example:

- Brain-expressed genes are strongly enriched for schizophrenia-associated genes
- Gene sets reflecting brain-specific processes and pathways predominantly contain brain-expressed genes
- Such gene sets will therefore show increased association with SZ even if completely irrelevant to SZ



Conditional gene-set analysis - recap

Confounding among gene sets can be tested using a conditional analysis

- In MAGMA: linear regression framework, can add potential confounders as covariates in the analysis to evaluate their influence

When analysing a ‘causal’ set A and an overlapping set B:

- Conditioning set B (on A) will make its association disappear, whereas conditioning set A (on B) will only reduce its association

Confounding remains problematic if ‘causal’ set not available

Interaction gene-set analysis- recap

- Interaction between gene sets A and B can be tested as an extension to the conditional analysis model in MAGMA
  - The interaction term is the set AB of genes shared by A and B
  - The interaction can be evaluated by testing AB conditional on A and B
- A gene set interaction arises if the genetic associations are specific to genes that share the same multiple functions

Practical

MAGMA

Developed and maintained by
Christiaan de Leeuw

MAGMA can be used to run gene-based and gene-set analyses

Practical

1. Annotate SNPs to genes
2. Perform gene analysis (with 10 PCs as covariates)
3. Perform gene-set 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
6. Perform interaction 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
6. Perform interaction 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/christiaan/Boulder2023/magma_session.zip .`
  - `unzip magma_session.zip`
- All instructions are in `instructions.txt` file
  - Can copy-paste commands directly from file
  - Make sure to set up the DATA variable before running MAGMA commands
- `The answer file can be found: /home/christiaan/Boulder2023/magma_answers.zip)`

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 - key points

- Step 6: interaction analysis
  - The I_LOVE_BRAINS gene set is showing a significant (positive) interaction with brain expression, but was not significant ($p = 0.38$) in earlier gene-set analysis
  - Suggests that this pathway is relevant for our phenotype, but only when genes are also more strongly expressed in the brain

Practical - conclusion

- Full answer file and all output:
  - `/home/christiaan/Boulder2023/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 c.a.de.leeuw@vu.nl