How do we go from genetic discoveries from GWAS/WGS/WES to mechanistic disease insight?

Danielle Posthuma

Part II – looking for convergence of gene functions – gene-set analysis

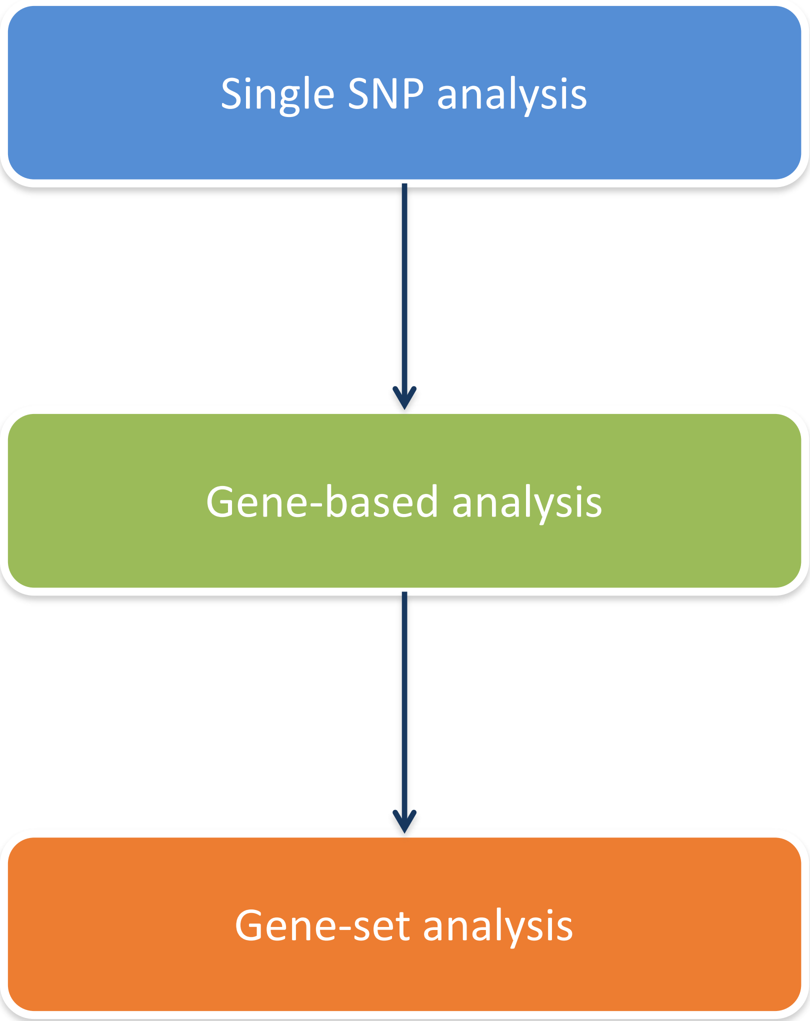
Making sense of GWAS results for complex traits

- Annotate SNPs to genes, based on physical location or regulatory relation

- Conduct gene-based analyses

- Conduct gene-set analyses

Testing for functional clustering of SNP associations



- GWAS
- single SNPs

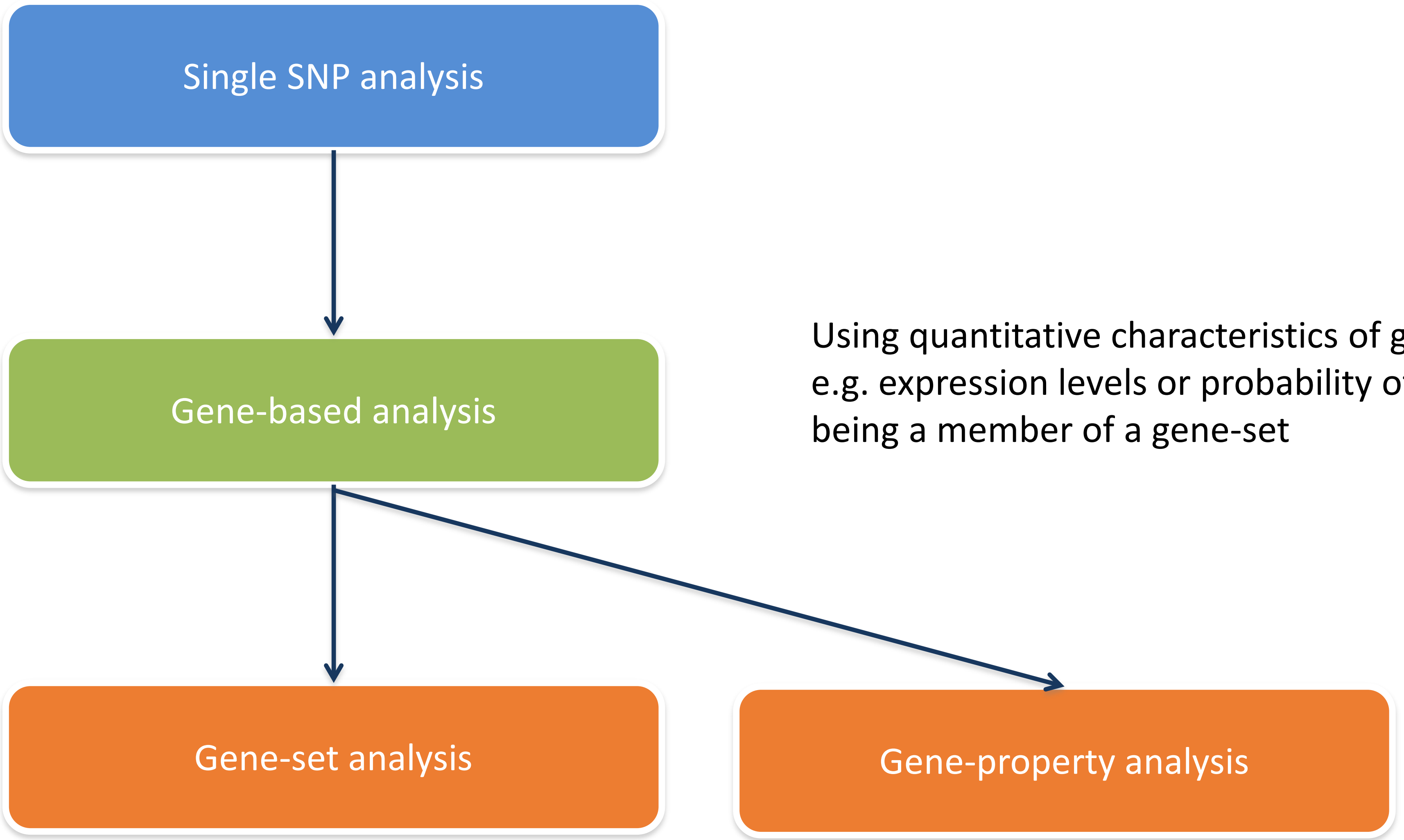
SNP-set or gene-based analysis with gene as unit of analysis

- SNPs annotated to genes based on e.g. position, eQTL association, or chromatin interaction
- 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

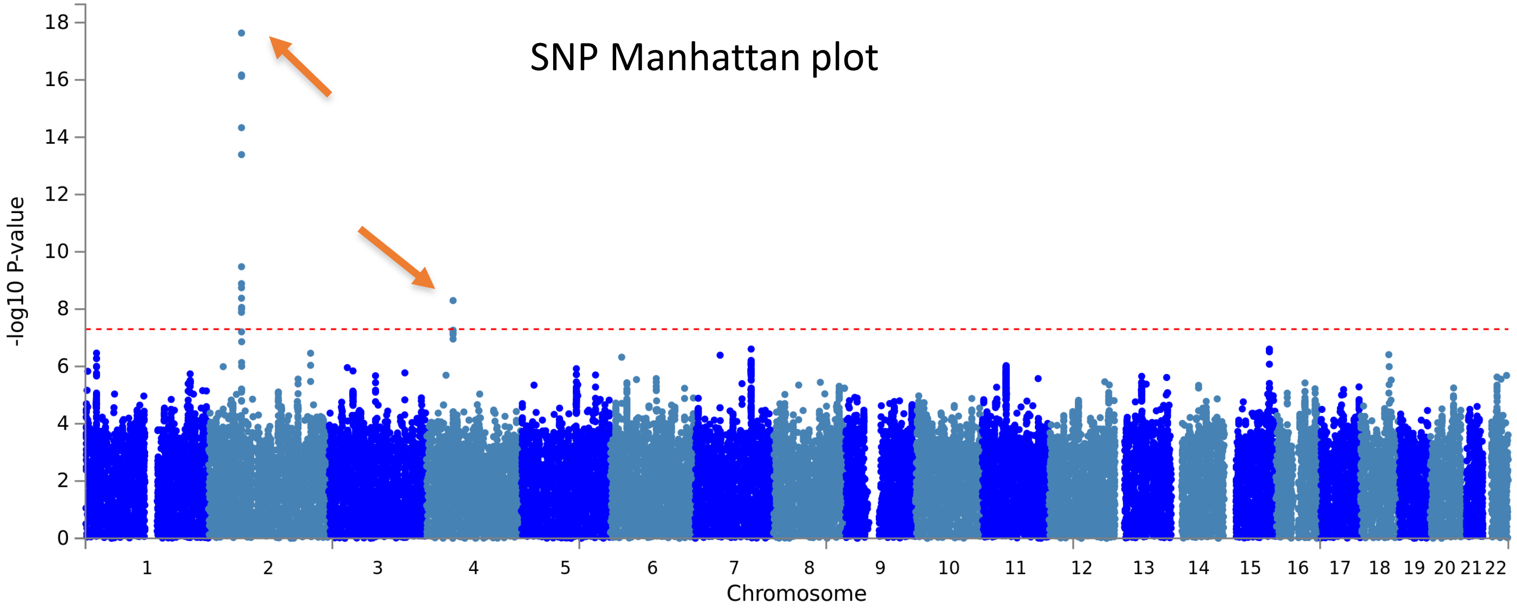


Using quantitative characteristics of genes e.g. expression levels or probability of being a member of a gene-set

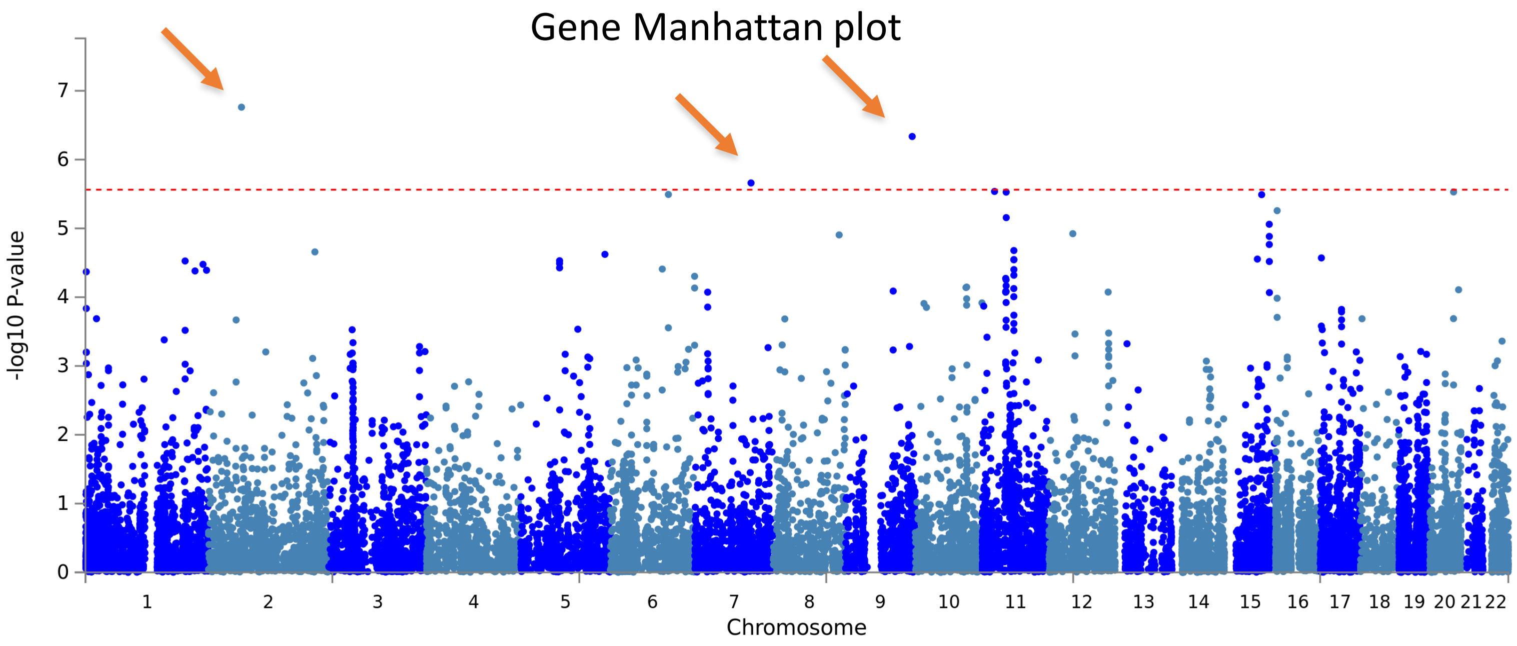
Gene-based analysis

- 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:
  - 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 information

Choosing gene-sets

Gene-sets can be based on e.g.

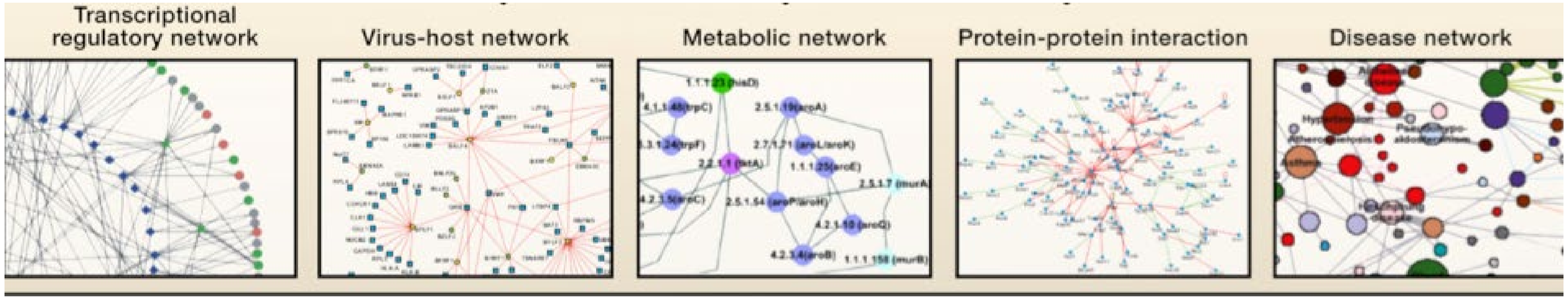
-protein-protein interaction

-co-expression

-transcription regulatory network

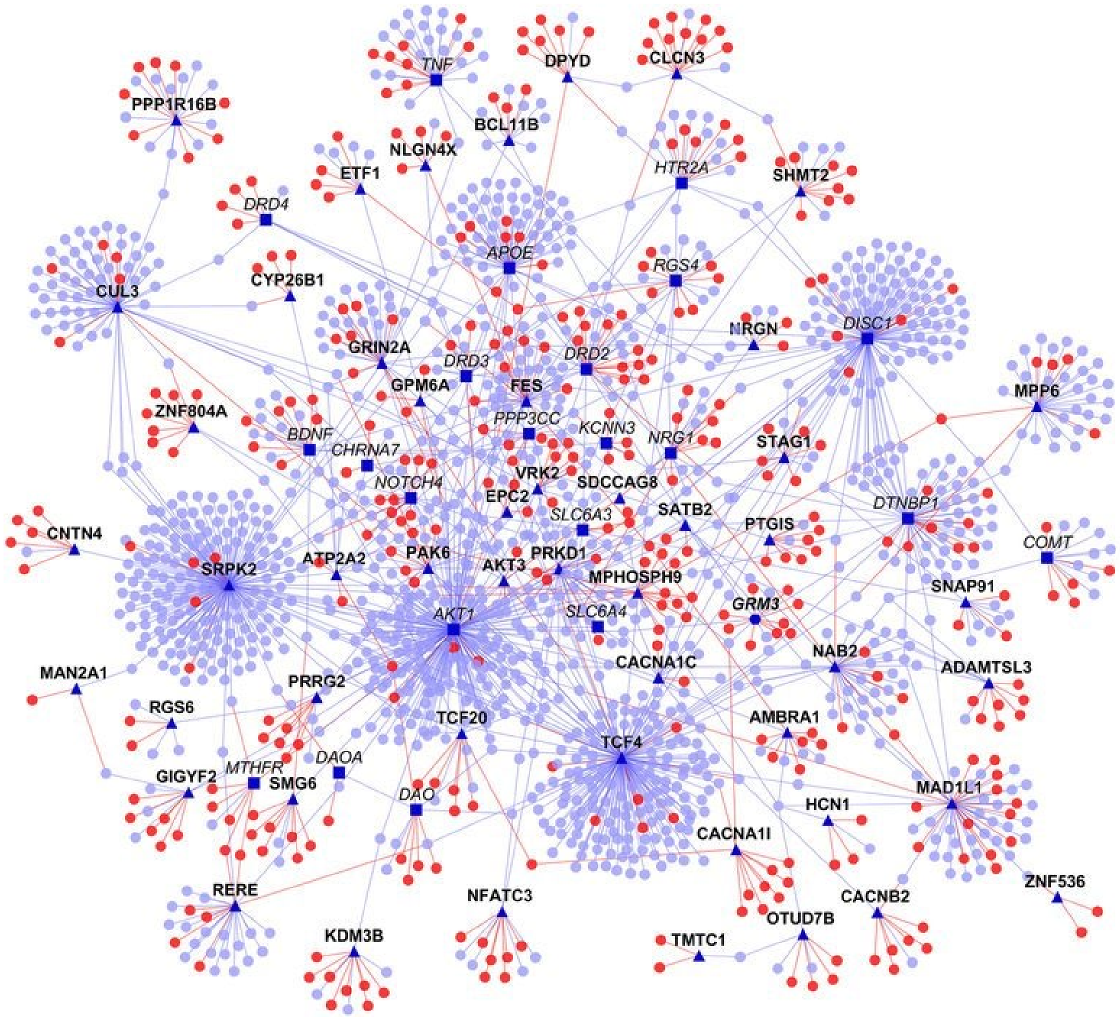
-biological pathway

-Functional relations

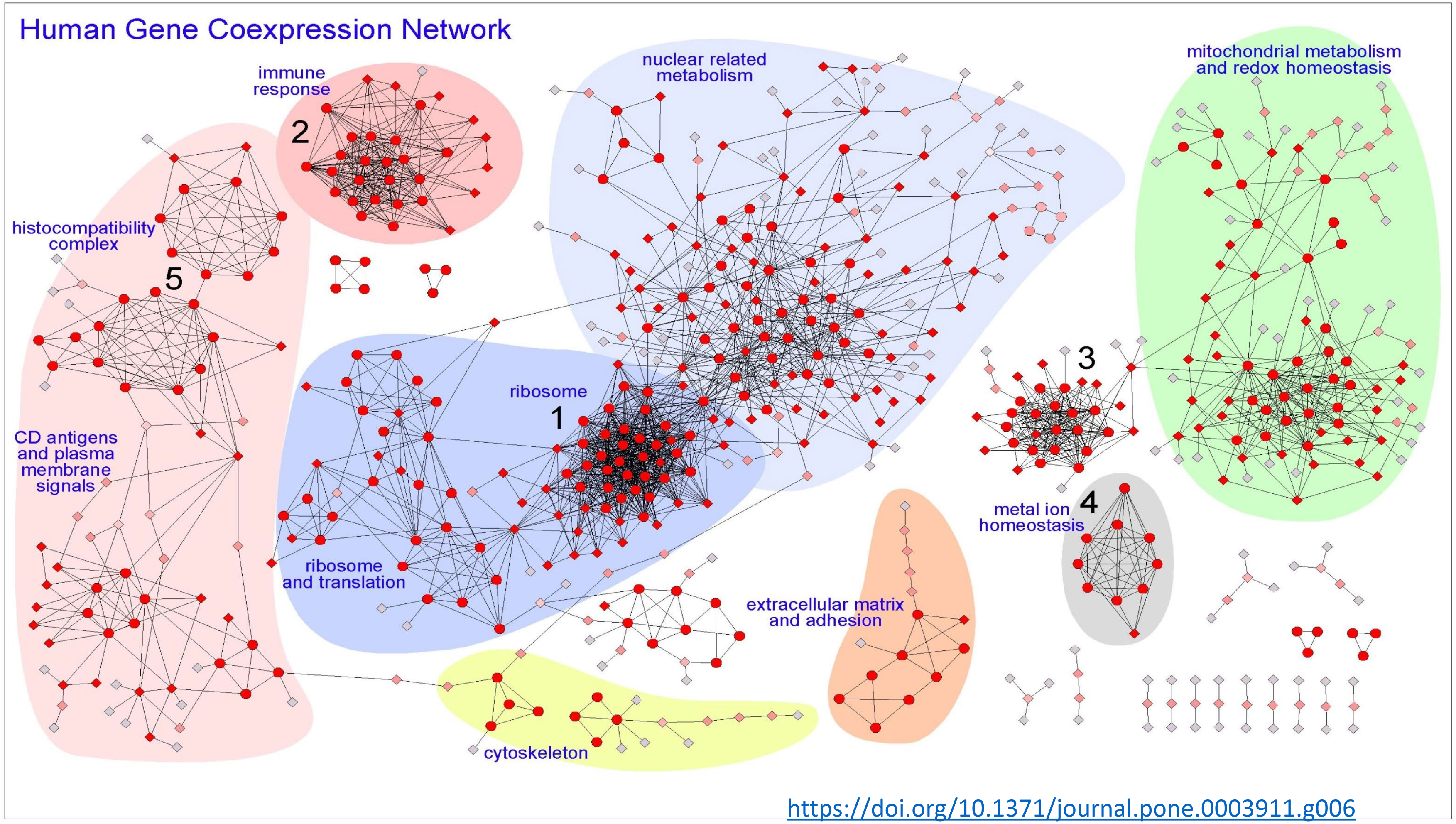


Protein interaction networks

Using Y2H or Immunoprecipitations

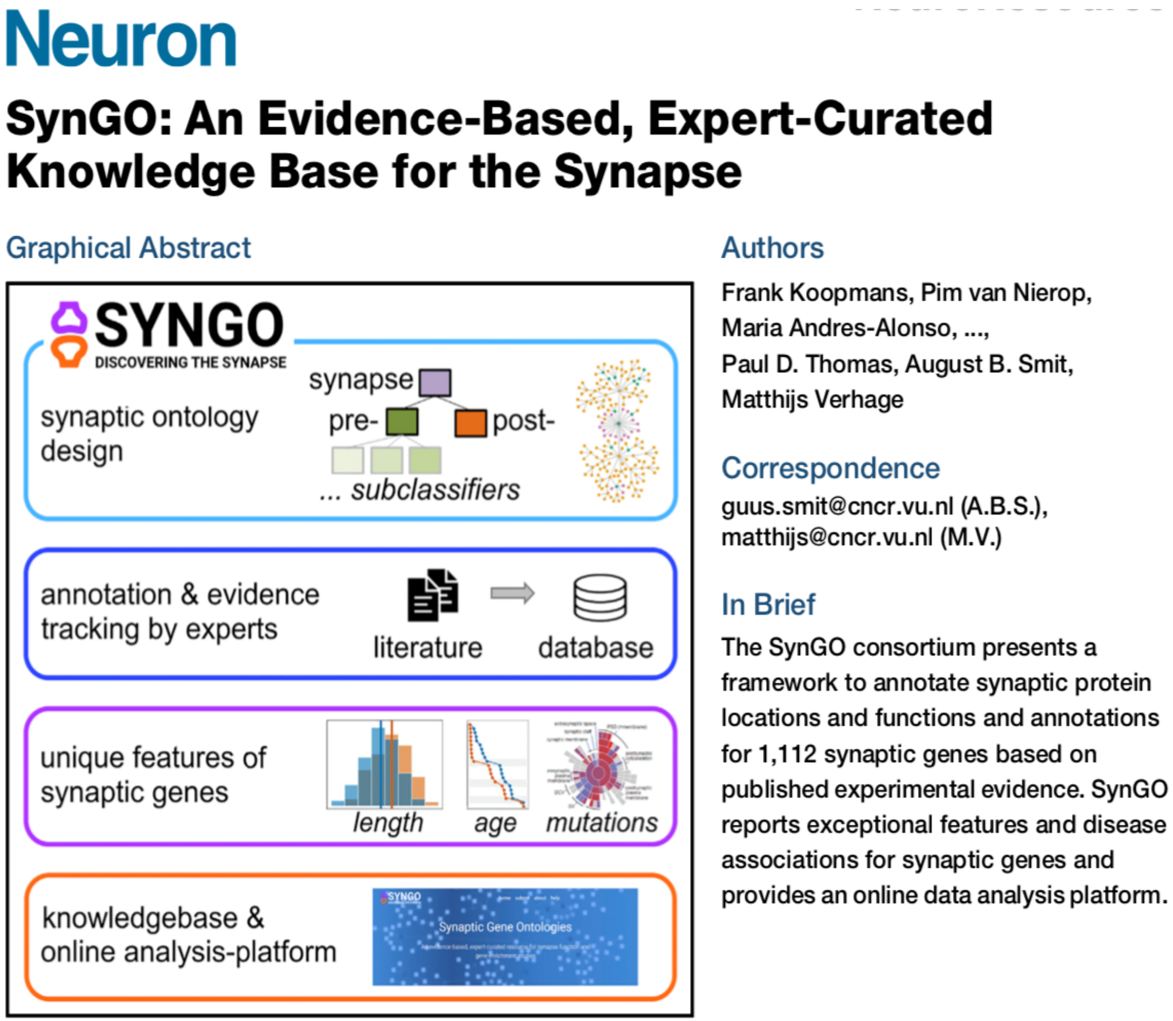


Co-expression networks


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, Gene Ontology, or 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.

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:

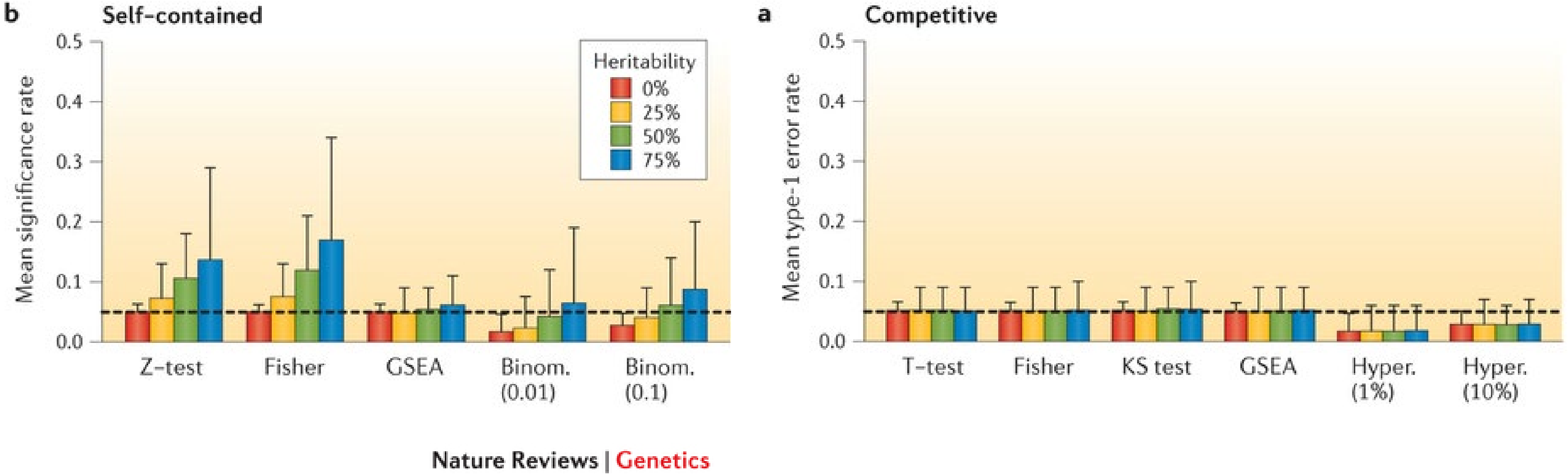
Self-contained:
H0: The genes in the gene-set are not associated with the trait

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



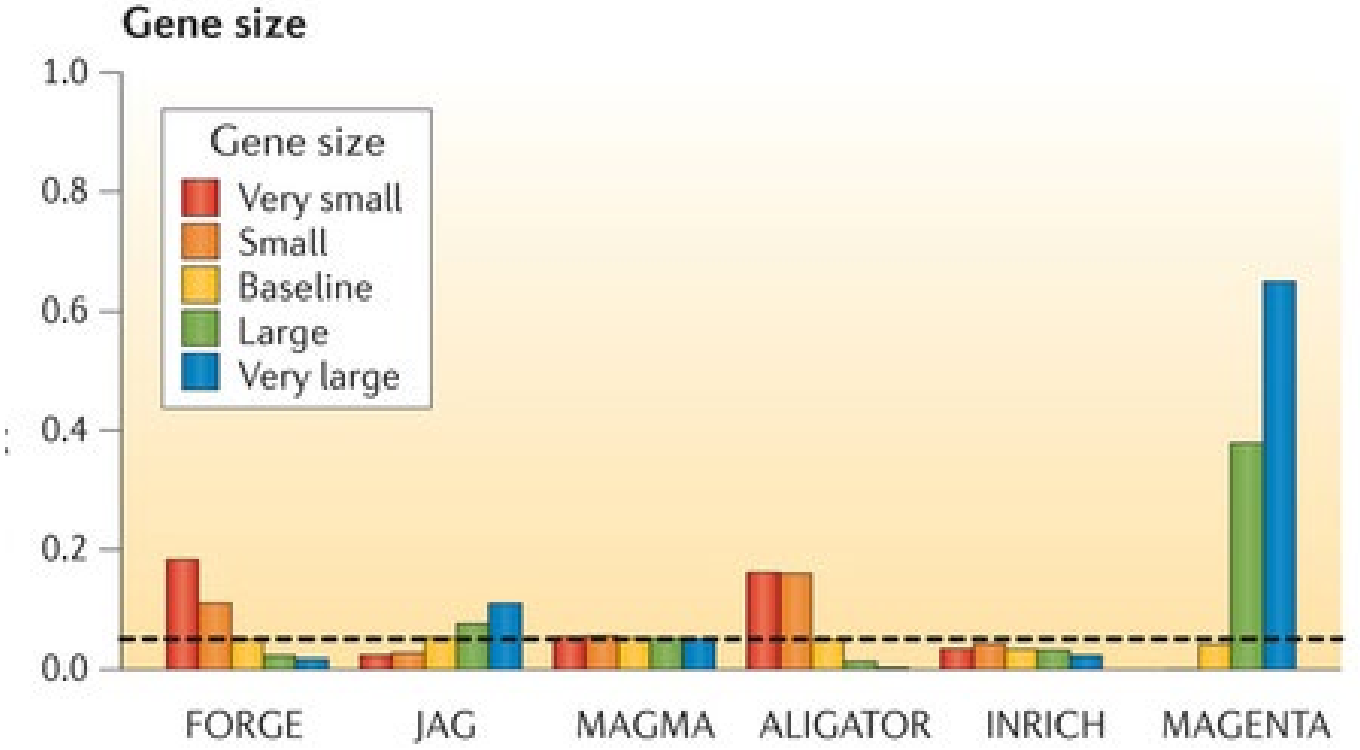
For self-contained methods, rates increase with heritability, whereas they are constant for competitive methods.

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

Different statistical algorithms test different alternative hypotheses

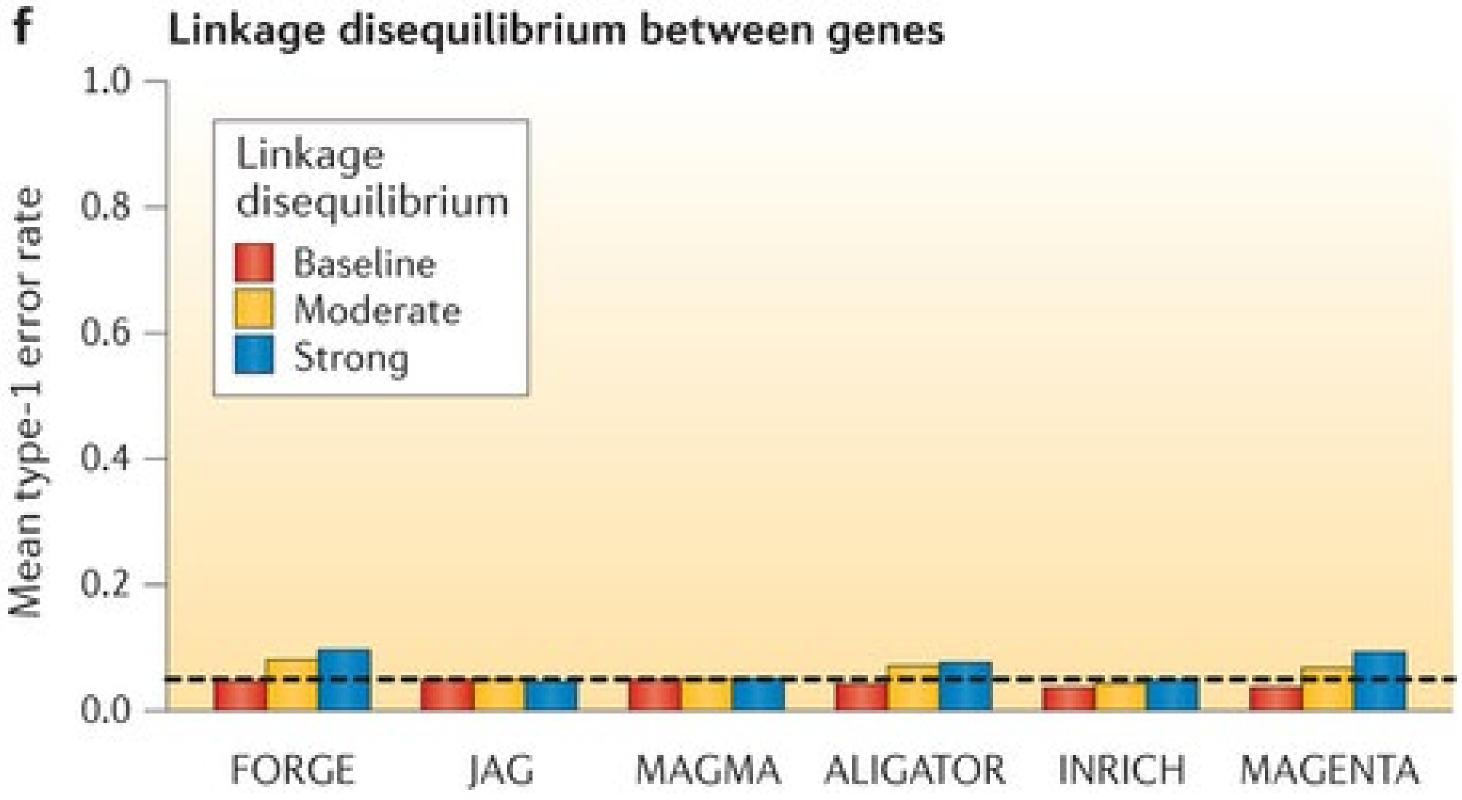
Strategy	Alternative hypothesis
Minimal P-value	At least one SNP in the gene or gene-set is associated with the trait
Combined P-value	The combined pattern of individual P-values provides evidence for association with the trait

Different tools are differentially affected by gene size



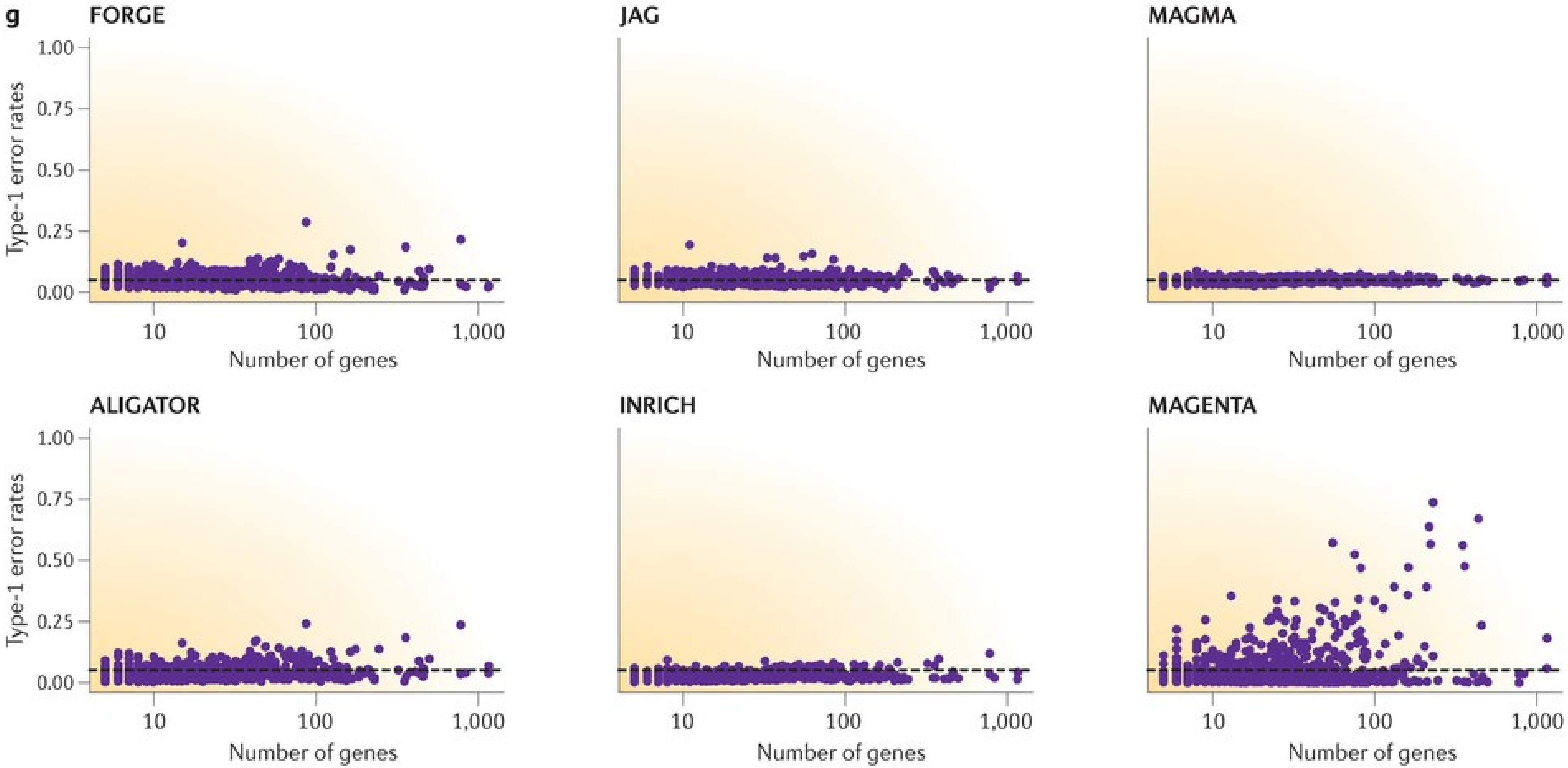
De Leeuw, Neale, Heskes, Posthuma. Nat Rev Genet, 2016

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

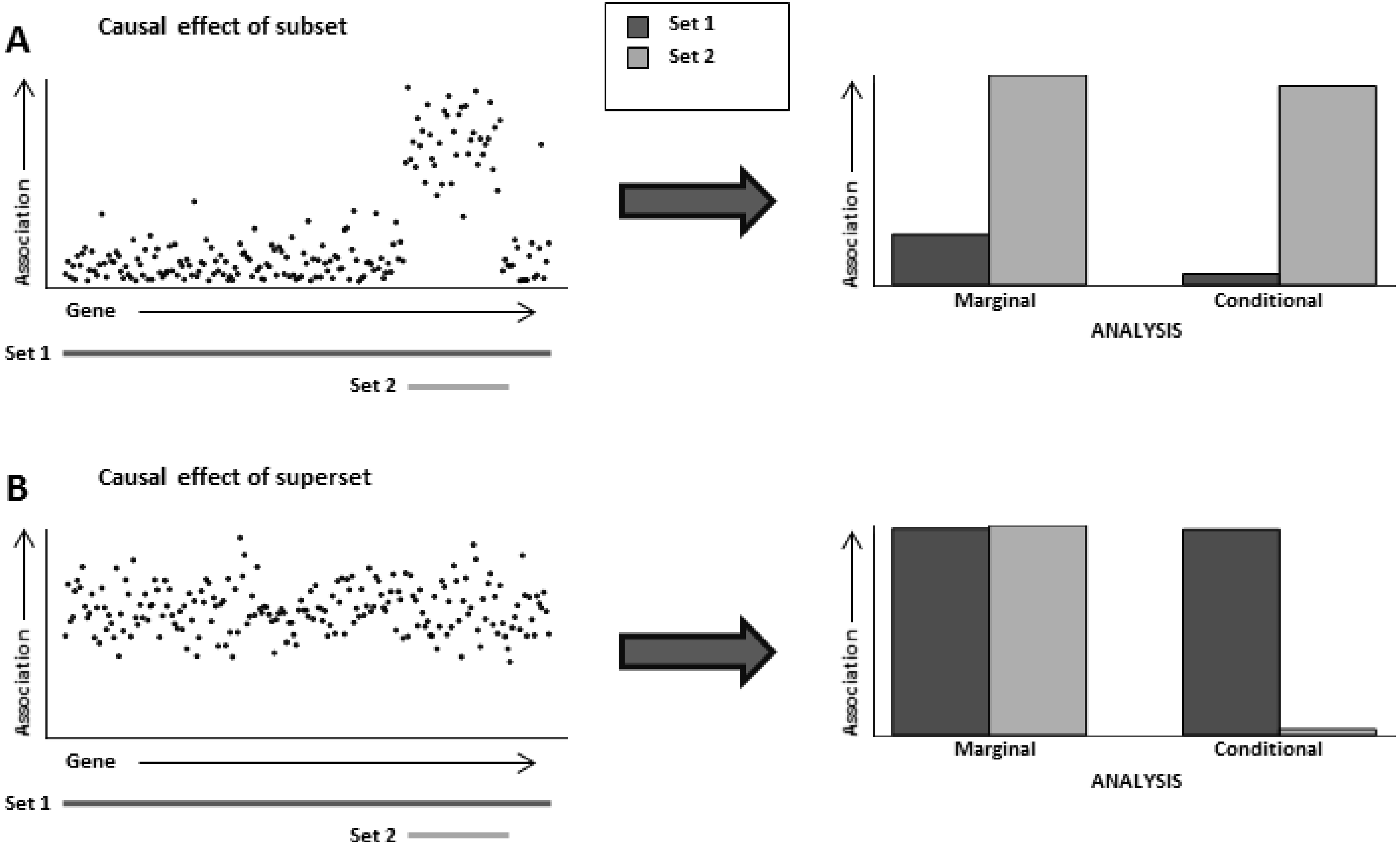


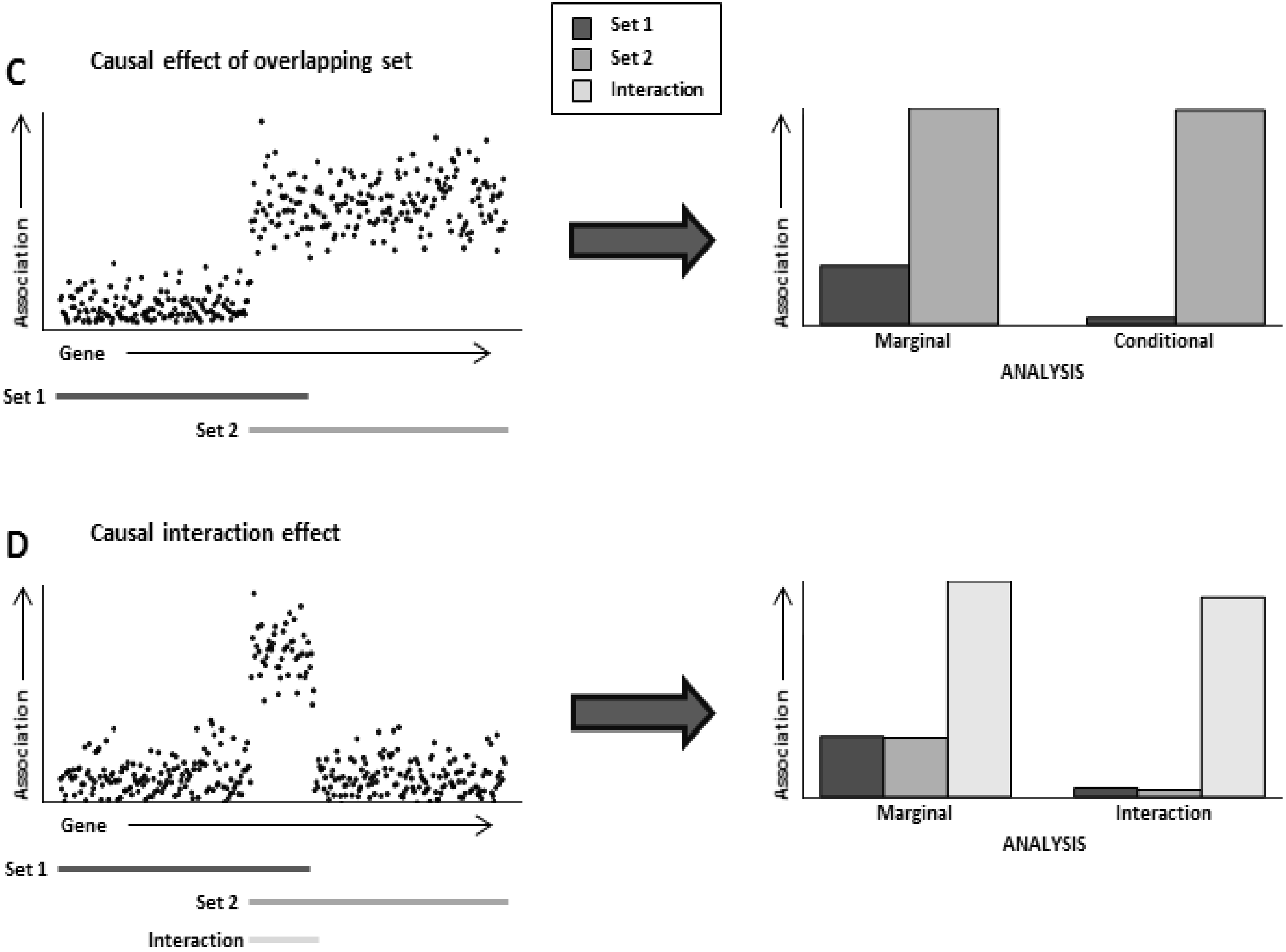
De Leeuw, Neale, Heskes, Posthuma. Nat Rev Genet, 2016

Issues of interpretation in gene-set analysis (GSA)

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

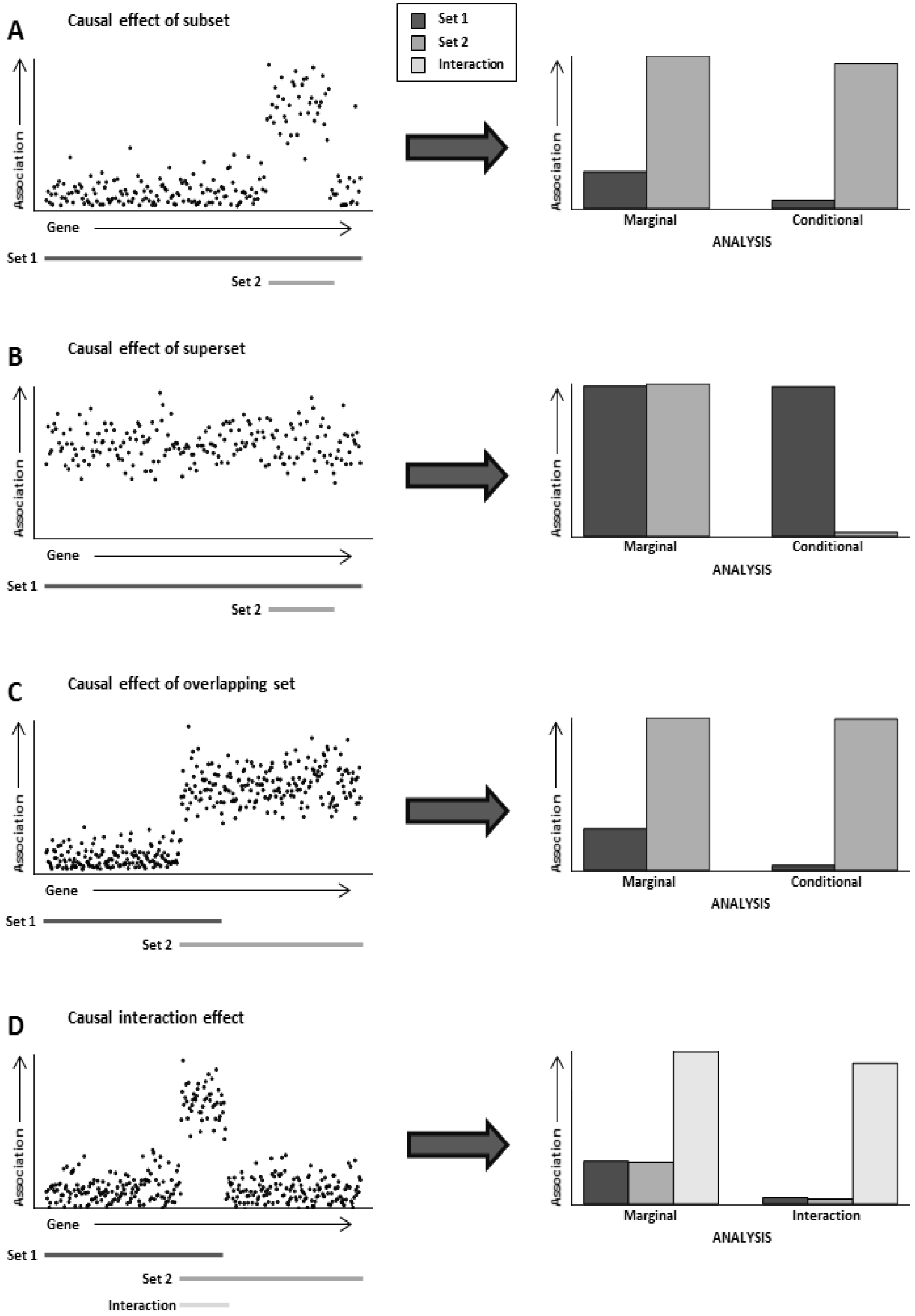

Set 1
Set 2
Interaction
C
Causal effect of overlapping set
Association
Gene
Set 1
Set 2
Association
Marginal
Conditional
ANALYSIS
D
Causal interaction effect
Association
Gene
Set 1
Set 2
Interaction
Association
Marginal
Interaction
ANALYSIS

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

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

- 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

Interpreting GWAS outcomes

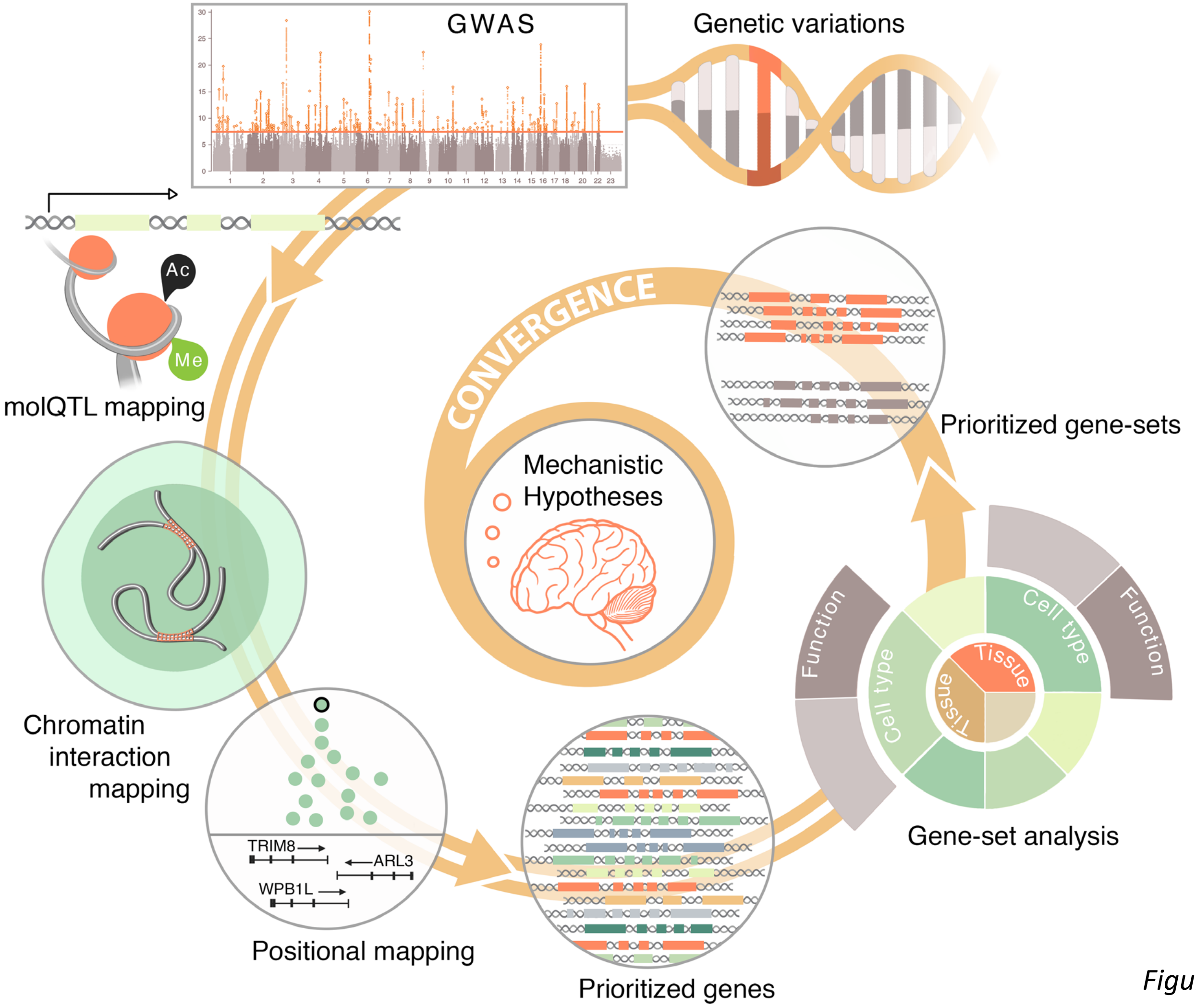


Figure from Uffelmann & Posthuma, Biol Psychiatry, 2020