Pathway/ Gene Set Analysis in Genome-Wide Association Studies

Alison Motsinger-Reif, PhD
Branch Chief, Senior Investigator
Biostatistics and Computational Biology Branch
National Institute of Environmental Health Sciences

alison.motsinger-reif@niehs.nih.gov

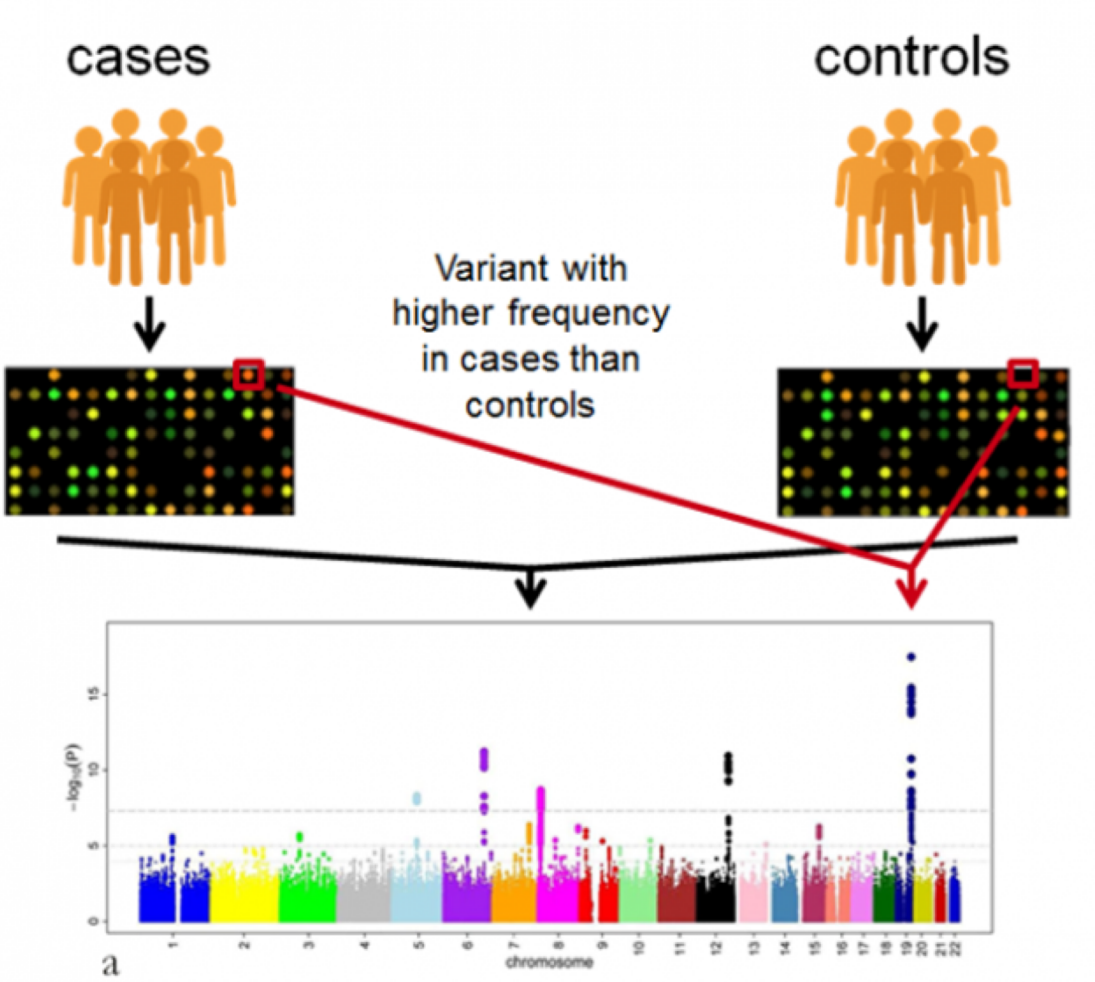
Goals

- Methods for GWAS pathway analysis with SNP chips

Many Shared Issues

- Many of the issues/choices/methodological approaches discussed for microarray data are true across all “-omics”
- Many methods have been readily extended for other omic data
- There are several biological and technological issues that may make just “off the shelf” use of pathway analysis tools inappropriate

Genome-Wide Association Studies



https://www.ebi.ac.uk/training-beta/online/courses/gwas-catalogue-exploring-snp-trait-associations/what-is-gwas-catalog/what-are-genome-wide-association-studies-gwas/

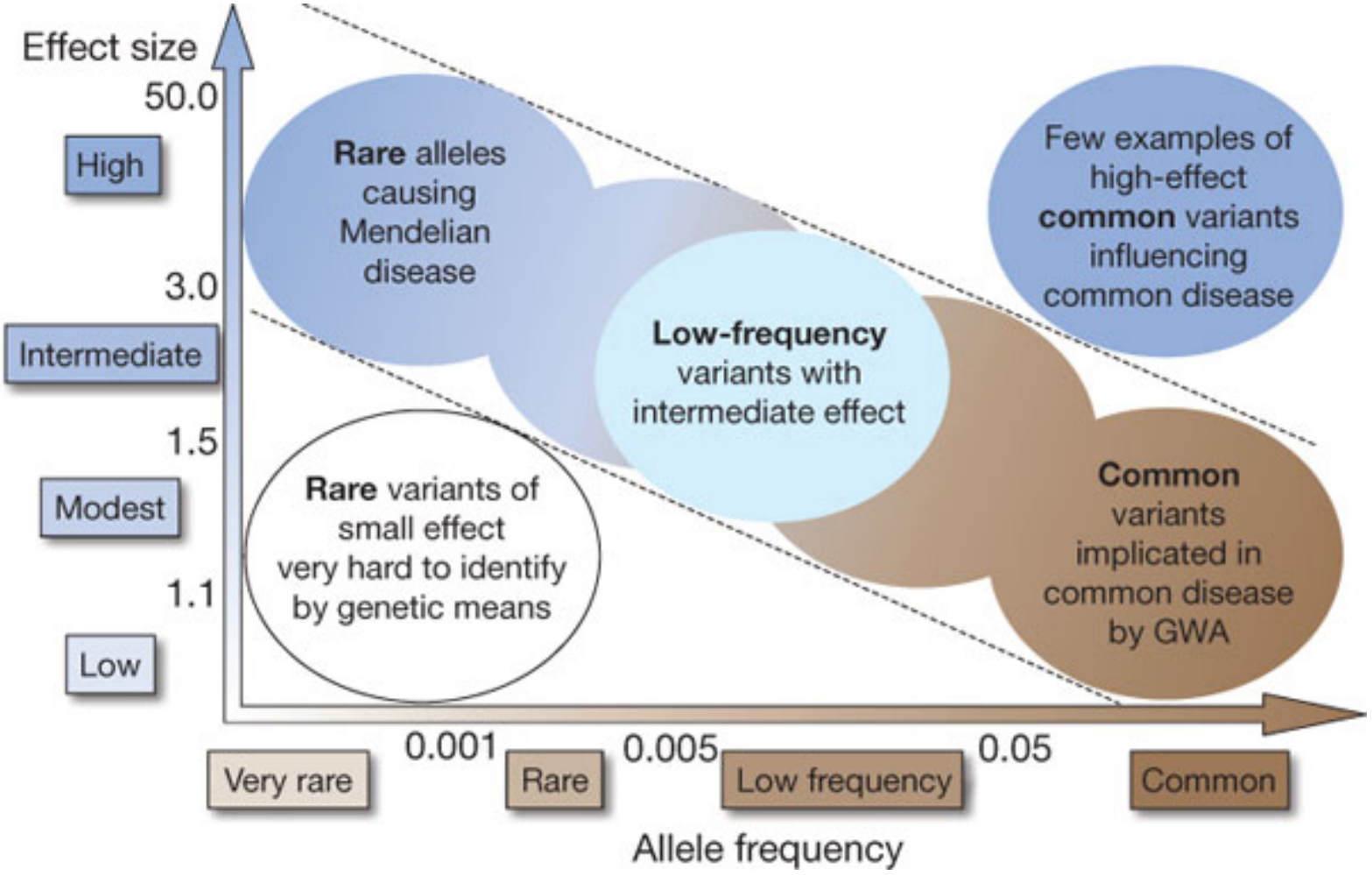
Advantages

- relatively unbiased, covers most of genome
- current cost is reasonable
- Fine mapping compared to linkage

Concerns

- missing heritability
  - Single SNPs explain little variation
- underlying assumptions not always true
  - CDCV…..
- Standard analysis looks variant-by-variant

Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio).

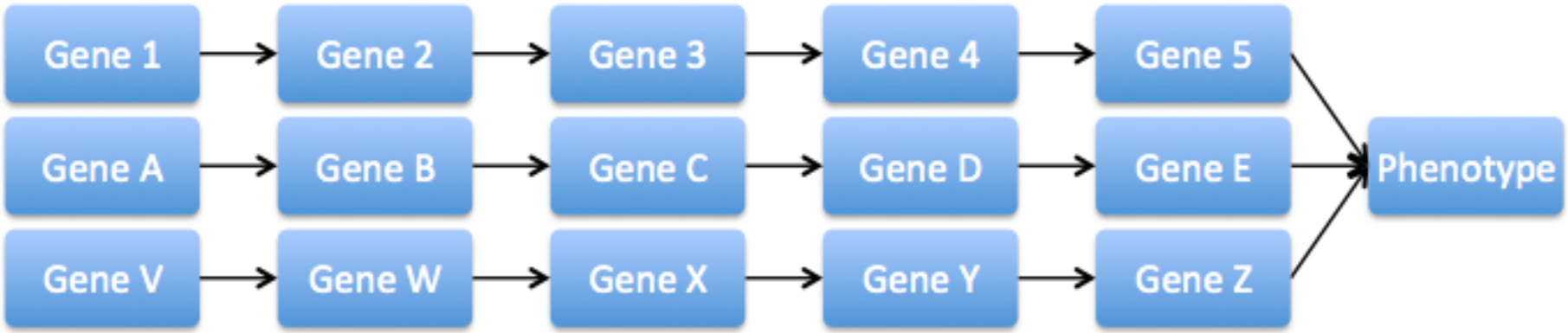


TA Manolio *et al. Nature* **461**, 747-753 (2009) doi:10.1038/nature08494

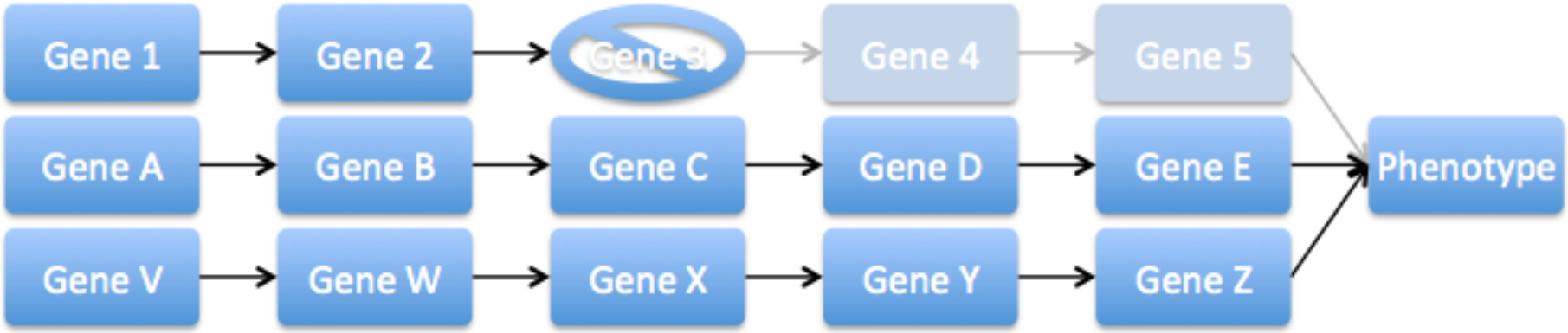
Possible Association Models

1. Each of several genes may have a variant that confers increased risk of disease independent of other genes
2. Several genes in contribute additively to the malfunction of the pathway
3. There are several distinct combinations of gene variants that increase relative risk but only modest increases in risk for any single variant

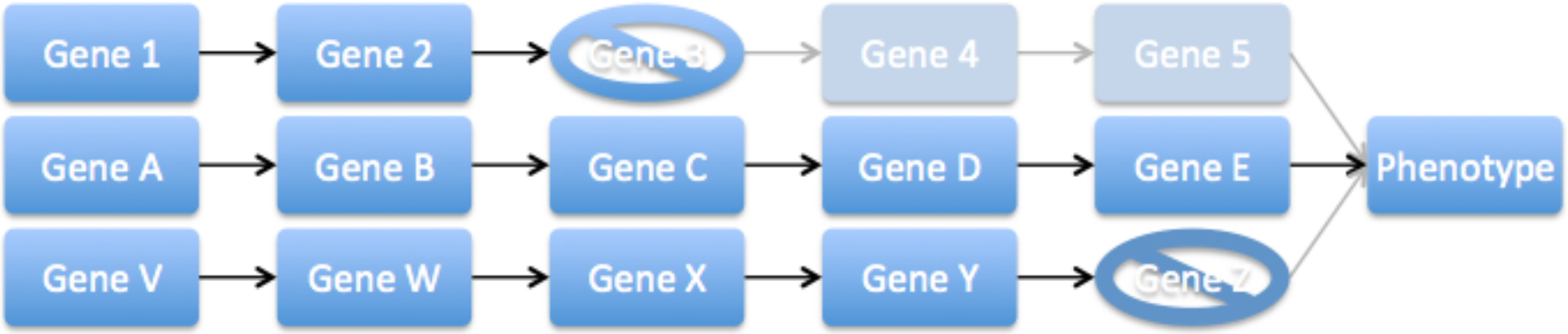
Hypothetical Disease Mechanism



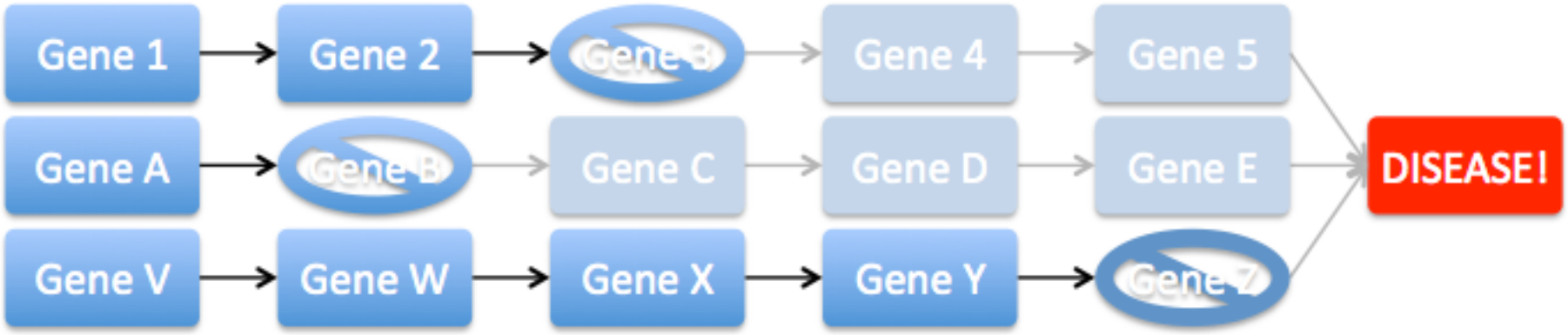
Hypothetical Disease Mechanism



Hypothetical Disease Mechanism



Hypothetical Disease Mechanism



Enrichment Testing in GWAS

- Testing pathway enrichment is possible in GWAS data
  - Many of the same issues that exist in gene expression enrichment testing occur in GWAS enrichment testing (e.g. choice of statistics, competitive vs self-contained)

- Primary difference:
  - In expression data the unit of testing is a gene
  - In GWAS data the unit of testing is a SNP

- Challenges:
  - Identifying the SNP (set) -> Gene mapping
  - Summarizing across individual SNP statistics to compute a per-gene measure

Mapping SNPs to Genes

- All SNPs in physical proximity of each gene
  - Pros:
    - All/most genes represented
  - Cons:
    - Varying number of SNPs per gene
    - Many of the SNPs may dilute signal
    - Defining gene proximity can affect results

- eSNPs (Expression associated SNPs)
  - Pros:
    - 1 SNP per gene
    - SNPs functionally associated
  - Cons:
    - Assumes variants effect expression
    - Not all genes have eSNPs
    - eSNPs may be study and tissue dependent

Gene summaries

- Initial studies propose different statistics for summarizing the overall gene association prior to enrichment analysis
  - Number/proportion of SNPs with pvalue < 0.05
  - Mean(-log10(pvalue))
  - Min(pvalue)
  - $1-(1-\text{Min(pvalue)})^{N}$
  - $1-(1-\text{Min(pvalue)})^{(N+1)/2}$

First approaches: combining p-values

- Compute gene-wise p-value:
  - Select most likely variant - ‘best’ p-value
  - Selected minimum p-value is biased downward
  - Assign ‘gene-wise’ p-value by permutations (Westfall-Young)
    - Permute samples and compute ‘best’ p-value for each permutation
    - Compare candidate SNP p-values to this null distribution of ‘best’ p-values

- Combine p-values by Fisher’s method, across SNPs (biased in the presence of correlation)

$$V = -\sum_{g_i \in G} \log(p_i)$$

$$p = \mathrm{P}(\chi^2_{(2k)} > 2V)$$

Next approaches

- Additive model: $$\log(\frac{p}{1-p}) = \sum_{g_i \in G} \beta_i n_i$$
  - Where n_i indexes the number of allele Bs of a SNP in gene *i* in the gene set *G*
  - Select subset of most likely SNP's
  - Fit by logistic regression (glm() in R)

- Significance by permutations
  - Permute sample outcomes
  - Select genes and fit logistic regression again
    - Assess goodness of fit each time
  - Compare observed goodness of fit

Competitive vs. Self-Contained Tests

- Competitive cutoff tests
  - Require only permuting SNP or Gene labels
  - May only allow to assess relative significance
- Self-contained distribution tests
  - Require permuting phenotype-genotype relationships
  - Resource intensive, may be difficult for large meta-analyses
  - Allow to assess overall significance

Competitive vs. Self-Contained Tests

- Self-contained null hypothesis
  - no genes in gene set are differentially expressed

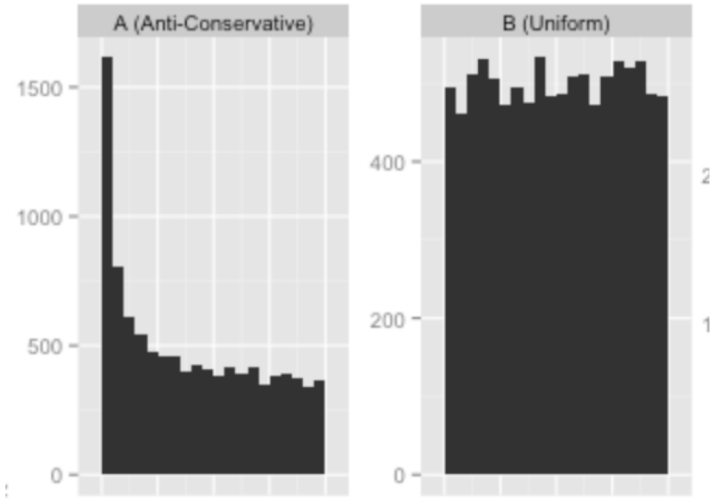
- Competitive null hypothesis
  - genes in gene set are at most as often differentially expressed as genes not in gene set

What does this mean for SNP data?

Choice of Pathways/Gene Sets

- Relatively less “signal” in GWAS than in gene expression (GE)
  - GE enrichment typically test *which* gene sets/pathways show enrichment
  - GWAS enrichment typically test *if* there is enrichment

- Typically want to be conservative about selecting the number of pathways to test, otherwise will be difficult to overcome multiple testing



- Prioritized Approach:
  - Limited number of specific hypotheses (e.g. gene sets from experiment, co-expression modules, disease-specific pathways/ontologies)
  - Exploratory analyses such as all KEGG/GO sets

Some Specific Methods

- SSEA
  - SNP Set Enrichment Analysis
- i-GSEA4GWAS
- MAGENTA
  - Meta-Analysis Gene-set Enrichment of variant Associations

SSEA

- Zhong et al. AJHG (2010)
- eSNP analysis to map SNPs to genes
  - More on this later…..
- Pathway statistic = one-sided Kolmogorov-Smirnov test statistic
- Pathway p-value assessed by permuting genotype-phenotype relationship
- FDR used to control error due to the number of pathways tested

i-GSEA4GWAS

- Zhang et al. *Nucl Acids Res* (2010)
- http://gsea4gwas.psych.ac.cn/

- Categorizes genes as significant or not significant
  - Significant: At least 1 SNP in the top 5% of SNPs
  - Does not adjust for gene size

- Pathway score: k/K
  - k = Proportion of significant genes in the geneset
  - K = Proportion of significant genes in the GWAS
- FDR assessed by permuting SNP labels

i-Gsea4Gwas V1.1

Improved - Gene Set Enrichment Analysis for Genome-Wide Association Study

A web server for identification of pathways/gene sets associated with traits

Demo Run

☑ **Load demo data**

Job name: untitled

Email (links for result will be sent to your email):

RUN CLEAR

Upload your GWAS data

Select data type: ◉ SNP ○ CNV ○ Gene

GWAS file: Choose File no file selected

☐ -logarithm transformation (necessary ONLY for *P*-value data)

Select mapping rules of SNPs->genes

◉ 500kb upstream and downstream of gene
○ 20kb upstream and downstream of gene
○ within gene
○ 100kb upstream and downstream of gene
○ 5kb upstream and downstream of gene
○ functional SNP (nonsynonymous, stop gained/lost, frame shift, essential splice site, regulatory region)

Gene set database

☑ canonical pathways ☐ GO biological process ☐ GO molecular function ☐ GO cellular component

OR upload your own gene sets file: Choose File no file selected

Options for gene set database

Limit gene sets by keyword (e.g. immune). The keyword can be gene name (e.g. CD4)

Keyword: ◉ include ○ exclude

Mask MHC/xMHC region

◉ NO ○ mask MHC ○ mask xMHC

Number of genes in gene set

Minimum (typical 5-20): 20

Maximum (typical 200-inf): 200

RUN CLEAR

Results

Pathway/Gene set name	Description	Manhattan plot	P-value	FDR	genes/Selected genes/All genes
HSA04950 MATURITY ONSET DIABETES OF THE YOUNG View Detail	Genes involved in ma..... More...		< 0.001	0.0030	11/23/25
PROSTAGLANDIN AND LEUKOTRIENE METABOLISM View Detail	More...		< 0.001	0.0085	13/27/32
HSA00565 ETHER LIPID METABOLISM View Detail	Genes involved in et..... More...		< 0.001	0.0125	15/28/31
DNA REPAIR View Detail	Genes annotated by t..... More...		< 0.001	0.0135	41/113/125
NTHIPATHWAY View Detail	Hemophilus influenza..... More...		< 0.001	0.0142	12/21/24
NEGATIVE REGULATION OF DEVELOPMENTAL PROCESS View Detail	Genes annotated by t..... More...		< 0.001	0.014571428	66/175/197
HSA04330 NOTCH SIGNALING PATHWAY View Detail	Genes involved in No..... More...		< 0.001	0.016	16/35/47
ENZYME LINKED RECEPTOR PROTEIN SIGNALING PATHWAY View Detail	Genes annotated by t..... More...		< 0.001	0.020875	60/136/140

MAGENTA

- Segre et al. *PLoS Genetics* (2010)
- Software download:
  - http://www.broadinstitute.org/mpg/magenta/
  - Requires MATLAB!!
  - Less convenient, but more customizable than iGSEA4GWAS
- Customizable proportion of "significant" genes
- Customizable gene window (upstream & downstream)
- Option for Rank-Sum test
- Gene Summary = min(p)
  - Uses stepwise regression to adjust for multiple possible factors: e.g. gene size, SNP density

MAGENTA Results

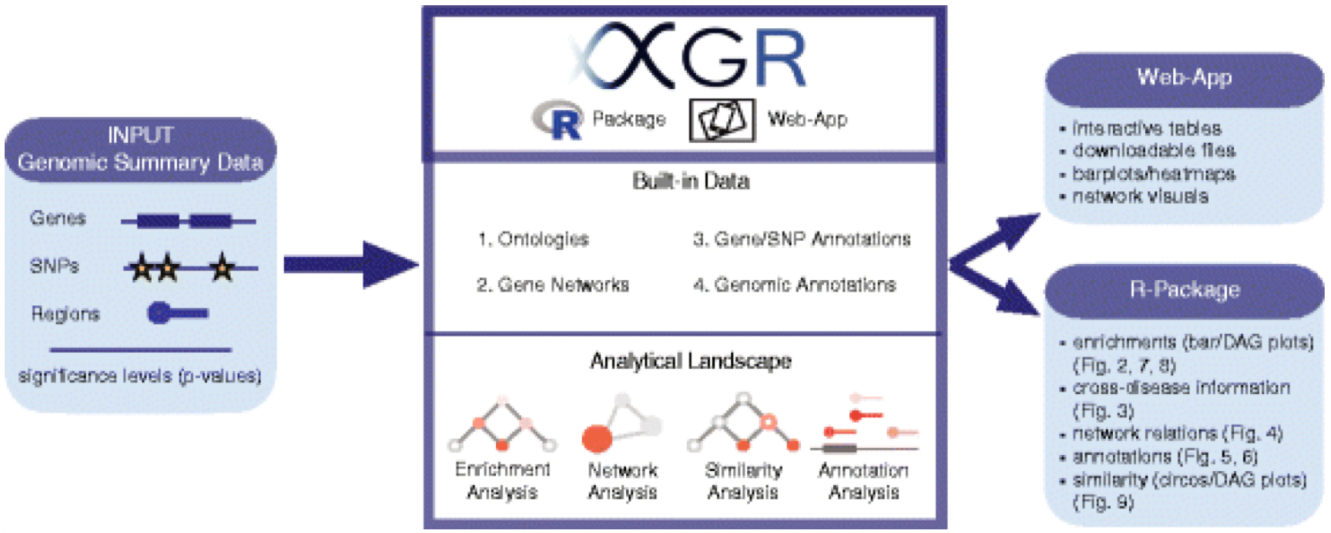
	95% Cutoff (Top 5%)				75% Cutoff (Top 25%)			
GS	NOMINAL GSEA PVAL	FDR	EXP # GENES	OBS # GENES	NOMINAL GSEA PVAL	FDR	EXP # GENES	OBS # GENES
positive regulation of osteoblast differentiation	3.36E-01	8.02E-01	1	2	3.00E-04	7.91E-02	6	14
one-carbon metabolic process	2.20E-03	3.55E-01	1	6	1.60E-03	1.44E-01	7	15
placenta development	3.36E-01	8.06E-01	1	2	4.00E-04	1.45E-01	6	14
carbohydrate transport	8.19E-01	9.46E-01	2	1	3.20E-03	3.45E-01	8	16

Adaptations of GSEA

- Order log-odds ratios or linkage p-values for all SNPs
- Map SNPs to genes, and genes to groups
- Use linkage p-values in place of t-scores in GSEA
  - Compare distribution of log-odds ratios for SNPs in group to randomly selected SNP's from the chip

XGR

- Fang H, Knezevic B, Burnham KL, Knight JC. XGR software for enhanced interpretation of genomic summary data, illustrated by application to immunological traits. Genome Med. 2016 Dec 13;8(1):129.



- Schematic workflow of XGR: achieving enhanced interpretation of genomic summary data. This flowchart illustrates the basic concepts behind XGR. The user provides an input list of either genes, SNPs, or genomic regions, along with their significance levels (collectively referred to as genomic summary data). XGR, available as both an R package and a web-app, is then able to run enrichment, network, similarity, and annotation analyses based on this input. The analyses themselves are run using a combination of ontologies, gene networks, gene/SNP annotations, and genomic annotation data (built-in data). The output comes in various forms, including bar plots, directed acyclic graphs (DAG), circos plots, and network relationships. Furthermore, the web-app version provides interactive tables, downloadable files, and other visuals (e.g. heatmaps)

XGR Functions

Functions	Tasks achieved	Runtime[a]
Enrichment analysis		
xEnricher	A template for enrichment analysis	~40
xEnricherGenes	Gene-based enrichment analysis using a wide variety of ontologies[b]	~40
xEnricherSNPs	SNP-based enrichment analysis using Experimental Factor Ontology on GWAS traits	~70
xEnricherYours	Custom-based enrichment analysis using user-defined ontologies	~5
xEnrichConciser	Removing redundant ones from enrichment outputs	~15
xEnrichBarplot	Barplot of enrichment outputs	<1
xEnrichCompare	Side-by-side barplots of comparative enrichment outputs	<1
xEnrichDAGplot	DAG plot of enrichment outputs	<1
xEnrichDAGplotAdv	DAG plot of comparative enrichment outputs	<1
Annotation analysis		
xGRviaGeneAnno	Annotation analysis using nearby gene annotations by a wide variety of ontologies[b]	~60
xGRviaGenomicAnno	Annotation analysis using a wide variety of genomic annotations[c]	~30
Similarity analysis		
xSocialiser	A template for similarity analysis	~60
xSocialiserGenes	Gene-based similarity analysis using structured ontologies on functions, diseases, and phenotypes	~70
xSocialiserSNPs	SNP-based similarity analysis using Experimental Factor Ontology on GWAS traits	~60
xCircos	Circos plot of similarity outputs	~10
xSocialiserDAGplot	DAG plot of one set of terms used for similarity analysis	<1
xSocialiserDAGplotAdv	DAG plot of two sets of terms used for similarity analysis	<1
Network analysis		
xSubneterGenes	Gene-based network analysis	~60
xSubneterSNPs	SNP-based network analysis	~60
xVisNet	Network visualisation	<1

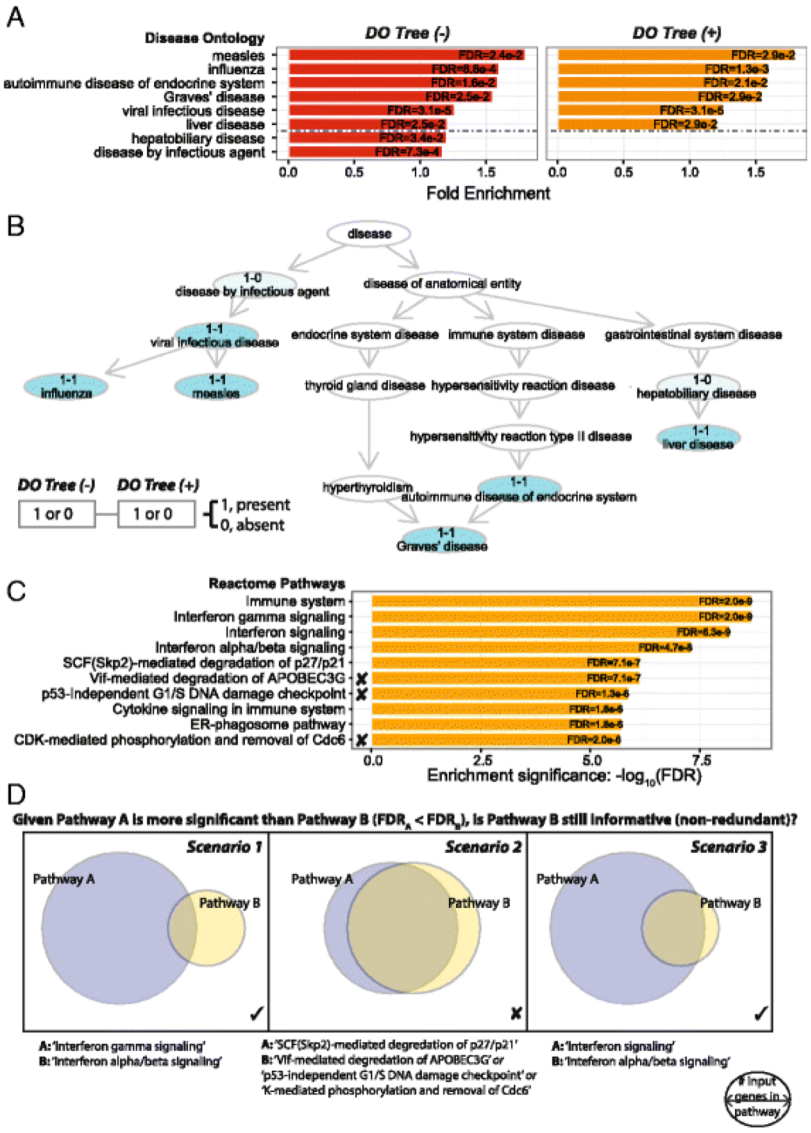


Fig. 2

Necessity of respecting ontology tree-like structure and of removing redundant non-structured pathways in enrichment analysis. This is demonstrated by analysing differentially expressed genes induced by 24-h interferon gamma in monocytes. The effect of taking ontology tree-like structure into account is demonstrated using Disease Ontology (DO) and the removal of redundant non-structured ontologies using Reactome pathways. **a** Side-by-side bar plots comparing the significant DO terms between the analysis without considering the tree structure (*DO Tree(-)*) versus the analysis considering the tree structure (*DO Tree(+)*). The horizontal dotted line separates commonly identified terms (*top section*) and redundant terms in the *DO Tree(-)* analysis. **b** DAG plot comparing commonly identified terms (coloured in *cyan*) and redundant terms from the *DO Tree(-)* analysis (coloured in *light cyan*). The term name (if significant) is prefixed in the form 'x1-x2'. x1 represents '*DO Tree (-)*' and x2 '*DO Tree (+)*'. The value of x1 (or x2) can be '1' or '0', denoting whether this term is identified (present) or not (absent). **c** The top pathway enrichments, with the redundant pathways to be removed indicated (*X*). **d** Illustrations of whether a less significant pathway B is redundant considering a more significant pathway A. Pathway B is counted redundant if it meets both criteria. Criterion 1: more than 90% of input genes annotated with pathway B are also covered by pathway A. Criterion 2: more than 50% of input genes annotated with pathway A are also covered by pathway B. Scenario 1 does not meet either criteria, scenario 2 meets both, and scenario 3 meets criterion 1 but not criterion 2. Notably, criterion 2 ensures the resulting pathways (as shown in scenario 3) are informative in capturing knowledge spheres of different granularities; otherwise, pathway B would be considered redundant in scenario 3, leading to loss of information. *FDR*: false discovery rate

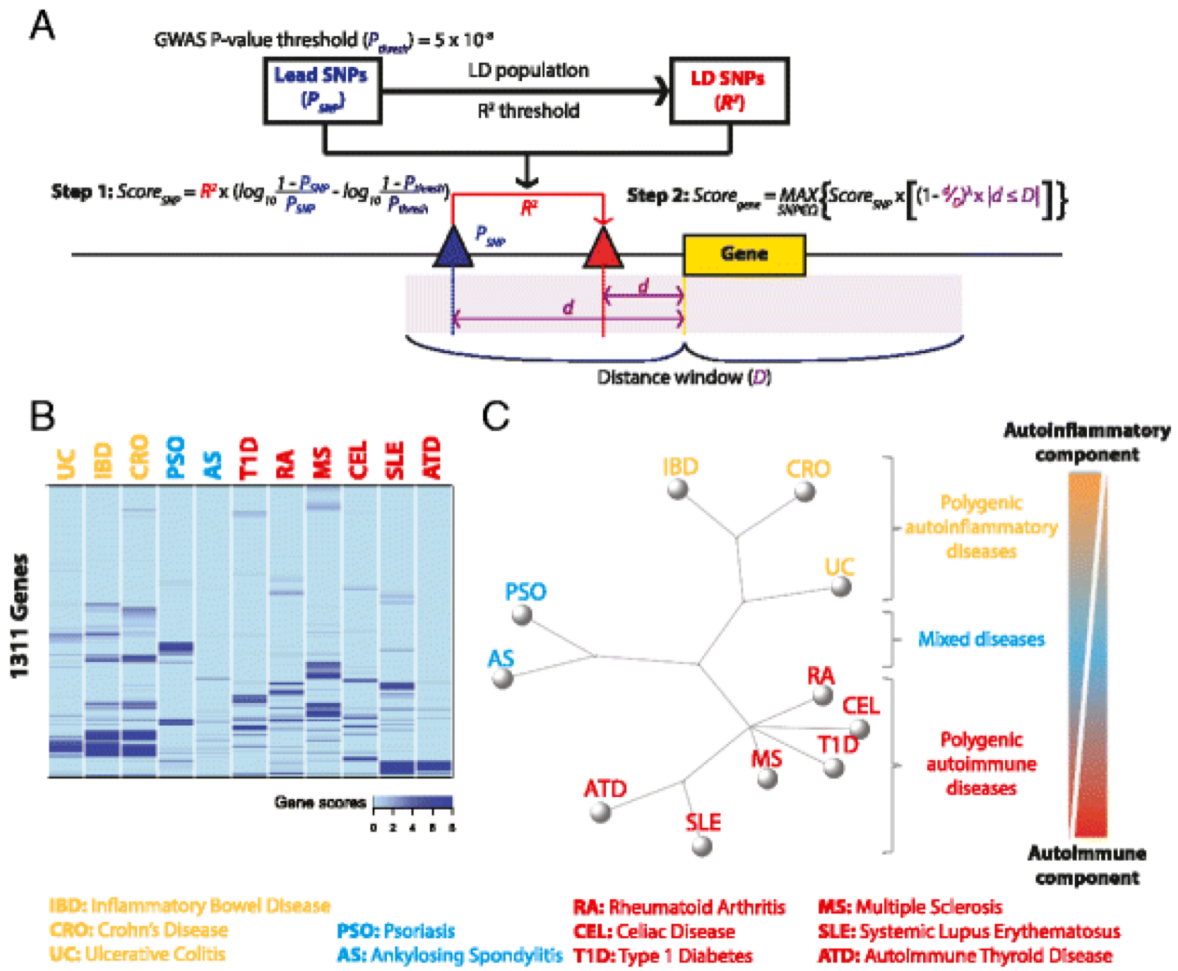


Fig. 3

Informativeness of using cross-disease GWAS summary data in characterising relationships between immunological disorders. **a** Gene scoring from GWAS SNPs prior to network analysis. **b** Heatmap of cross-disease gene scores for 11 common immunological disorders based on ImmunoBase GWAS summary data. **c** Consensus neighbour-joining tree based on the gene-scoring matrix resolves disease classification/taxonomy according to the genetic and cellular basis of autoinflammation and autoimmunity. Subdivided into 1) polygenic autoinflammatory diseases with a prominent autoinflammatory component, 2) polygenic autoimmune diseases with a prominent autoimmune component, and 3) mixed diseases having both components. Inter-disease distance is defined as the cumulative difference in gene scores

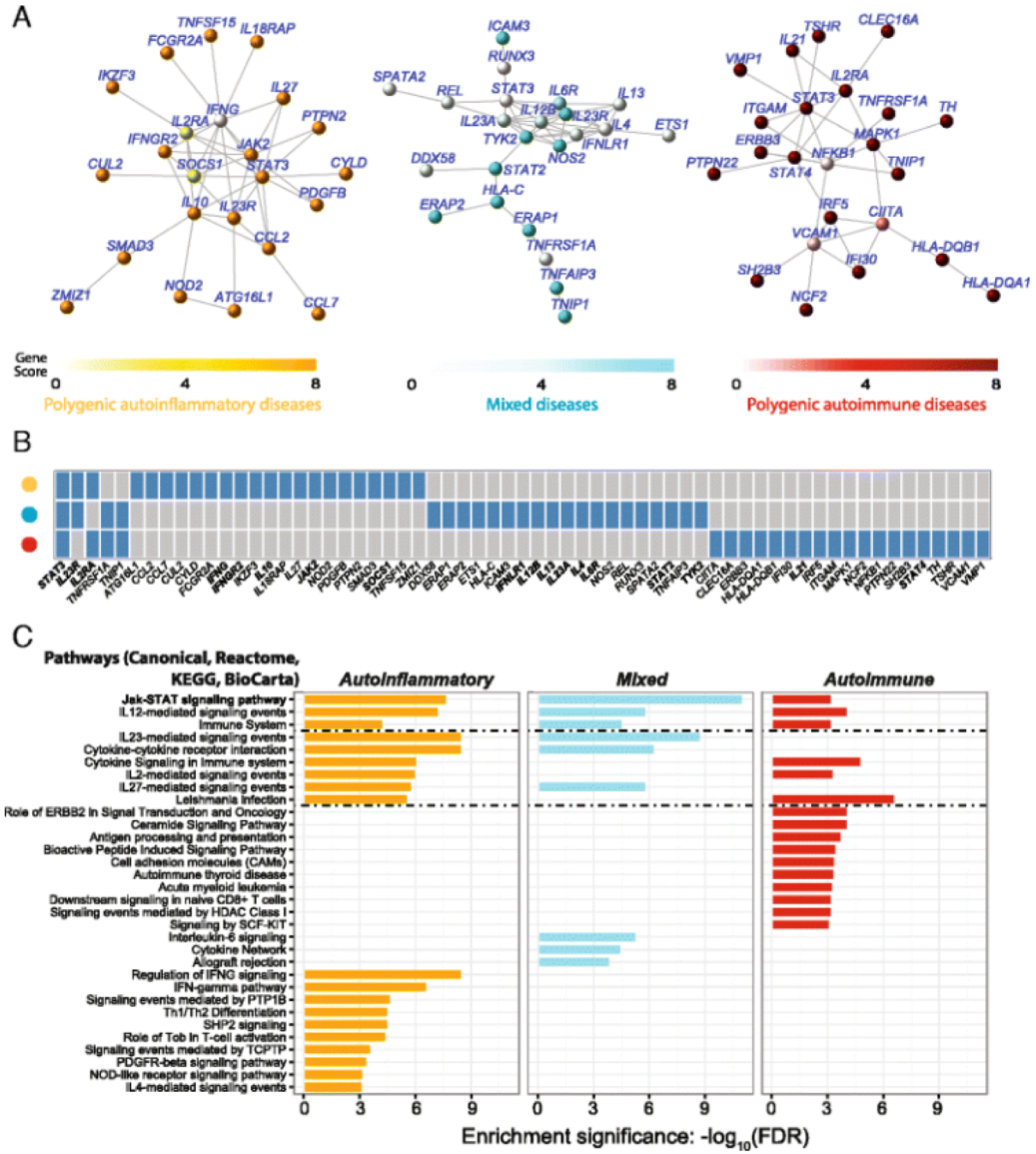


Fig. 4

SNP-modulated gene networks underlying three immunological disease categories. **a** The top-scoring gene network for the three disease categories: autoinflammatory diseases (*orange*), mixed diseases (*cyan*), and autoimmune diseases (*red*). **b** Network genes shared by and unique to disease categories. Genes involved in the Jak-STAT signalling pathway are in *bold text*. **c** Pathway enrichment analysis of network genes using all pathway ontologies and eliminating redundant pathways. The *horizontal dotted line* separates pathways common to all three disease categories (*top section*; e.g. Jak-STAT signalling pathway), those shared by any two categories (*middle*), and those only enriched in one category (*bottom*)

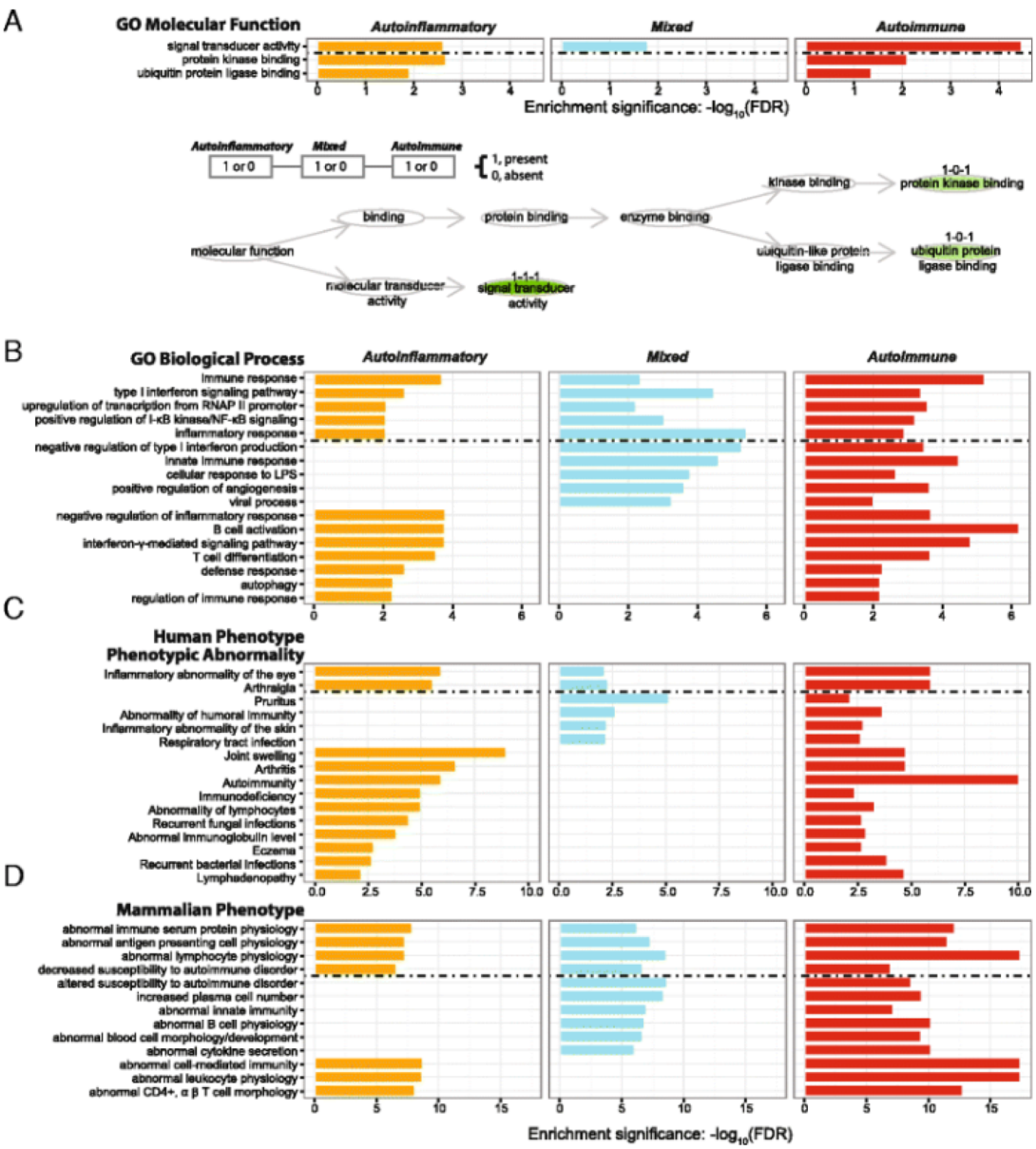


Fig. 5

Functional and phenotypic annotation analysis of genes harbouring GWAS SNPs for three immunological disease categories. Visualised in aside-by-side bar plot and/or DAG plot using functional ontologies, including **a** GO molecular function and **b** GO biological process; and using phenotype ontologies in human and mouse, including **c** human phenotype phenotypic abnormality, and **d** mammalian phenotype

Other Functionalities

- Cross-condition comparative enrichment analysis
- SNP similarity analysis based on disease trait profiles
  - eQTLs
- Epigenetic annotation/enrichment

Summary Points for GWAS

- In GWAS, few SNPs typically reach genome-wide significance
- Biological function of those that do can take years of work to unravel
- Incorporating biological information (expression, pathways, etc) can help interpret and further explore GWAS results
- Enrichment tests can be used to explore biological pathway enrichment
  - Different tests tell you different things
- Annotation choices very different that in gene expression data, though still rely on the same resources.... not necessarily so for other ‘omics”

Questions?