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

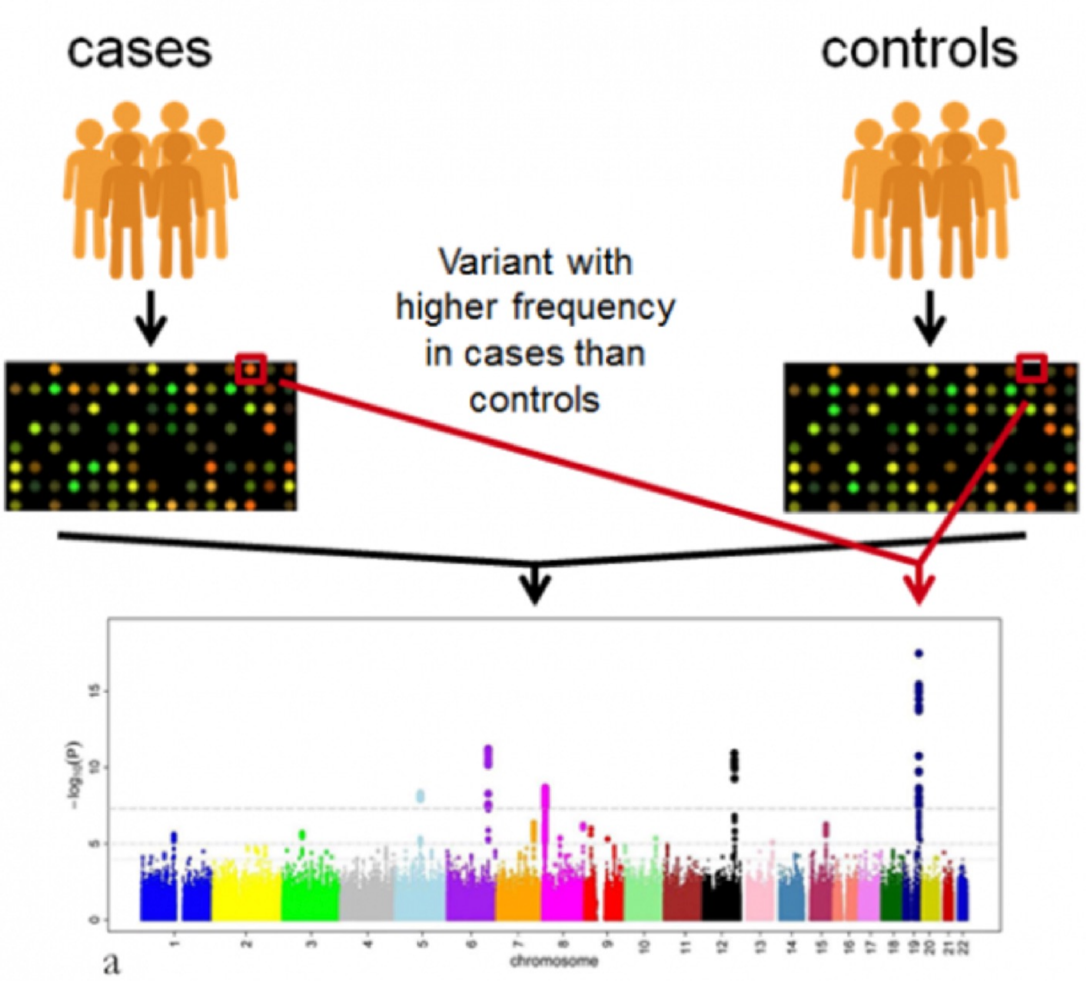
- Methods for GWAS with SNP or whole genome sequencing (WGS) chips
  - Integrating expression and SNP information

Many Shared Issues

- Most issues/choices/approaches discussed for microarray data are true across all “-omics”

- A few biological and technological issues that may make just “off the shelf” use of expression 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/

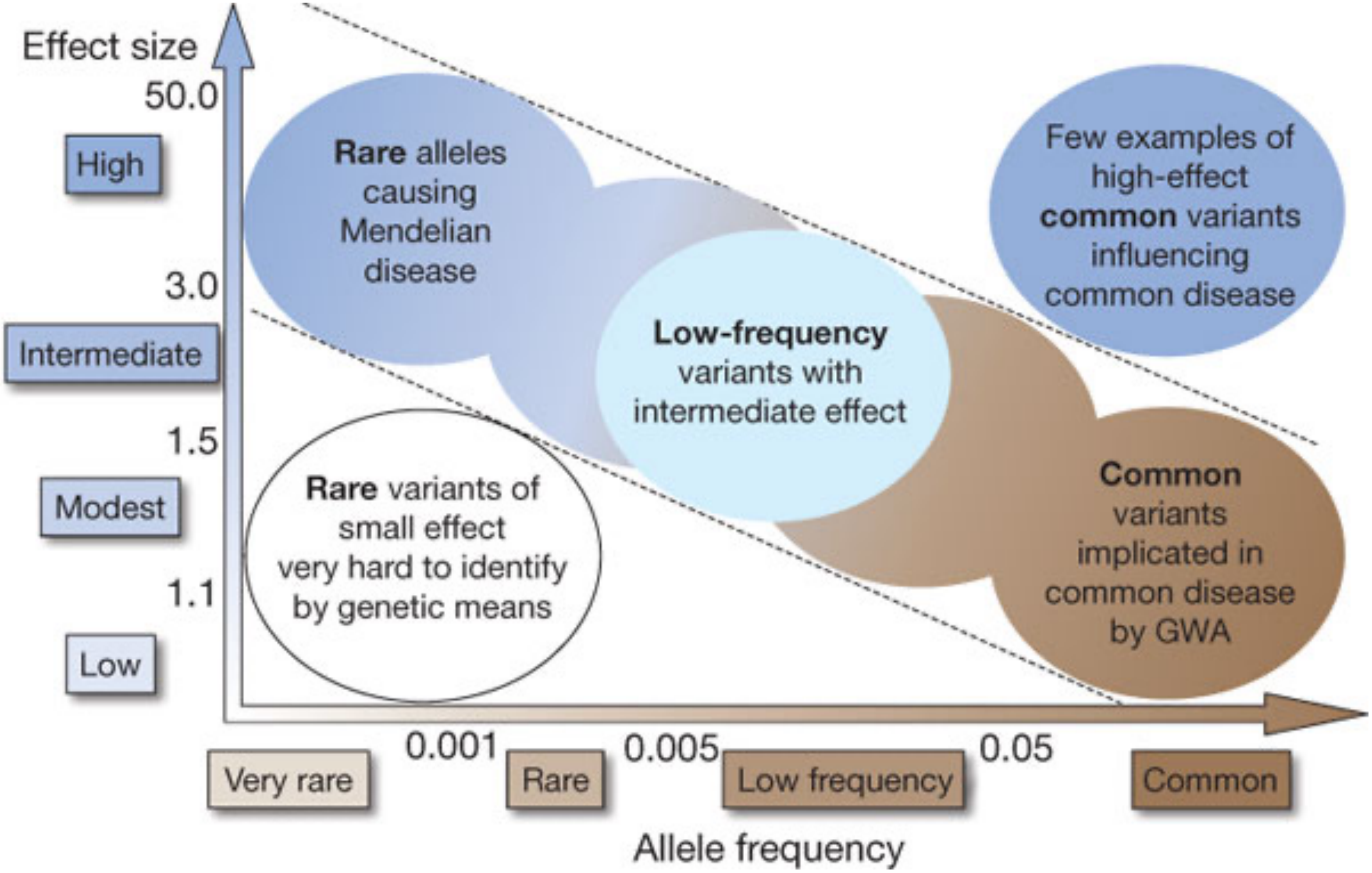
Advantages

- 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).


Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Few examples of high-effect common variants influencing common disease
Low-frequency variants with intermediate effect
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
0.001
0.005
0.05
Very rare
Rare
Low frequency
Common
Allele frequency

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

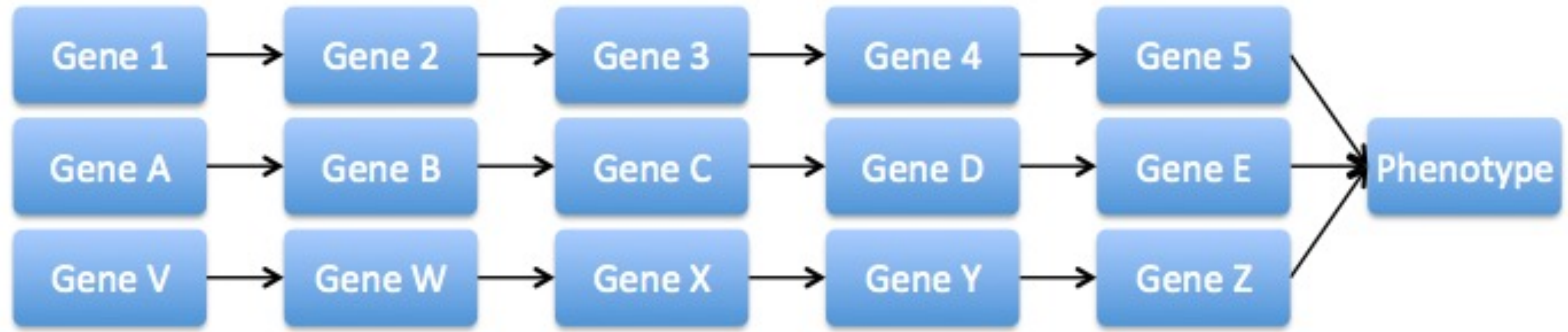
Possible Association Models

1. Each of several genes may have a variant that confers increased risk of disease independent of other genes

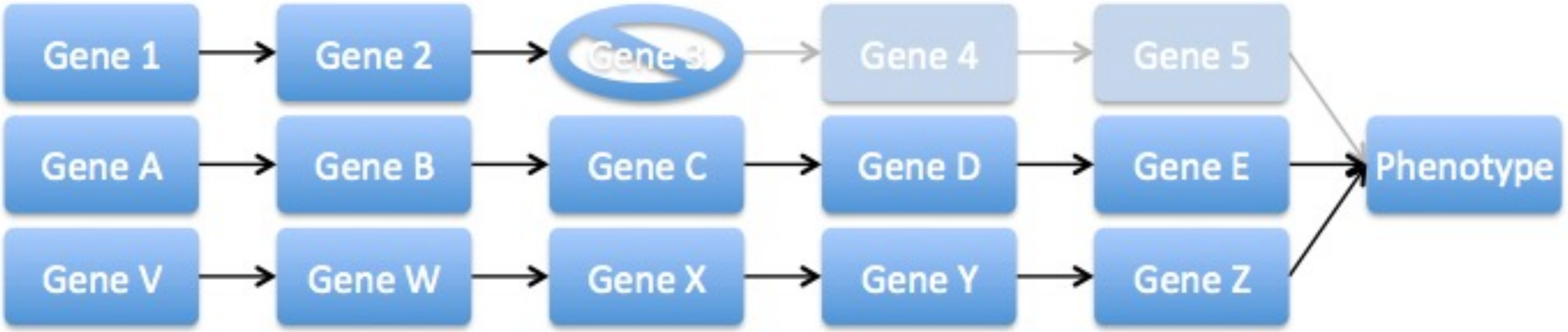
2. Several genes in contribute additively to the malfunction of the pathway

3. There are several distinct combinations of gene variants that increase relative risk but only modest increases in risk for any single variant

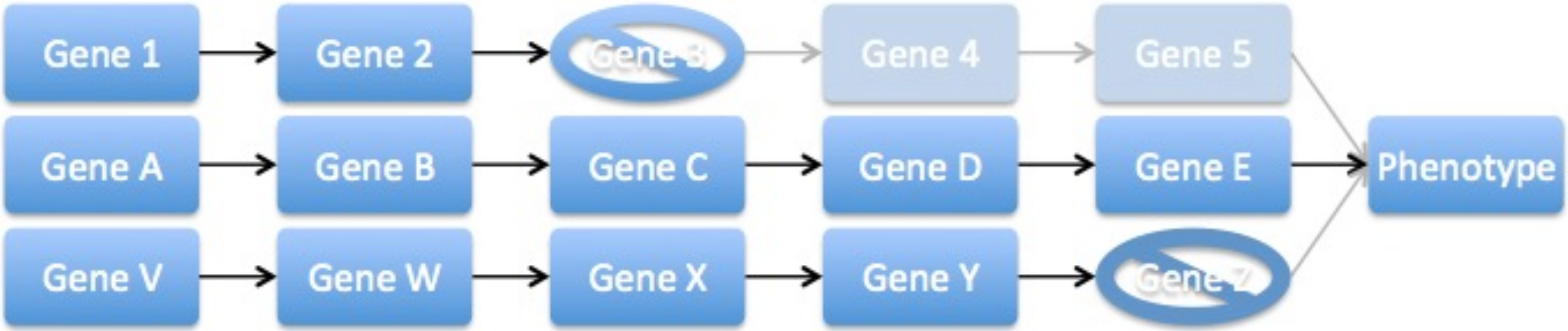
Hypothetical Disease Mechanism



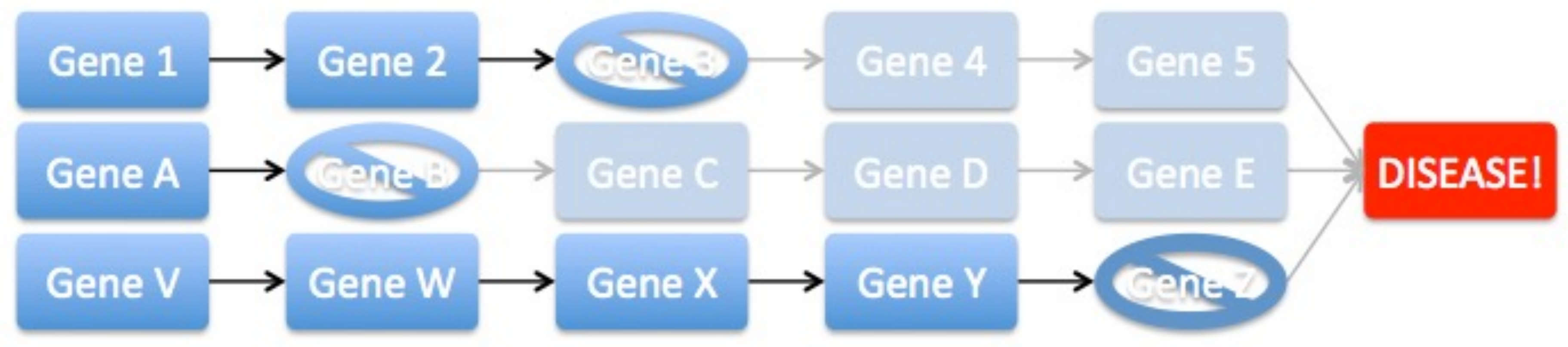
Hypothetical Disease Mechanism



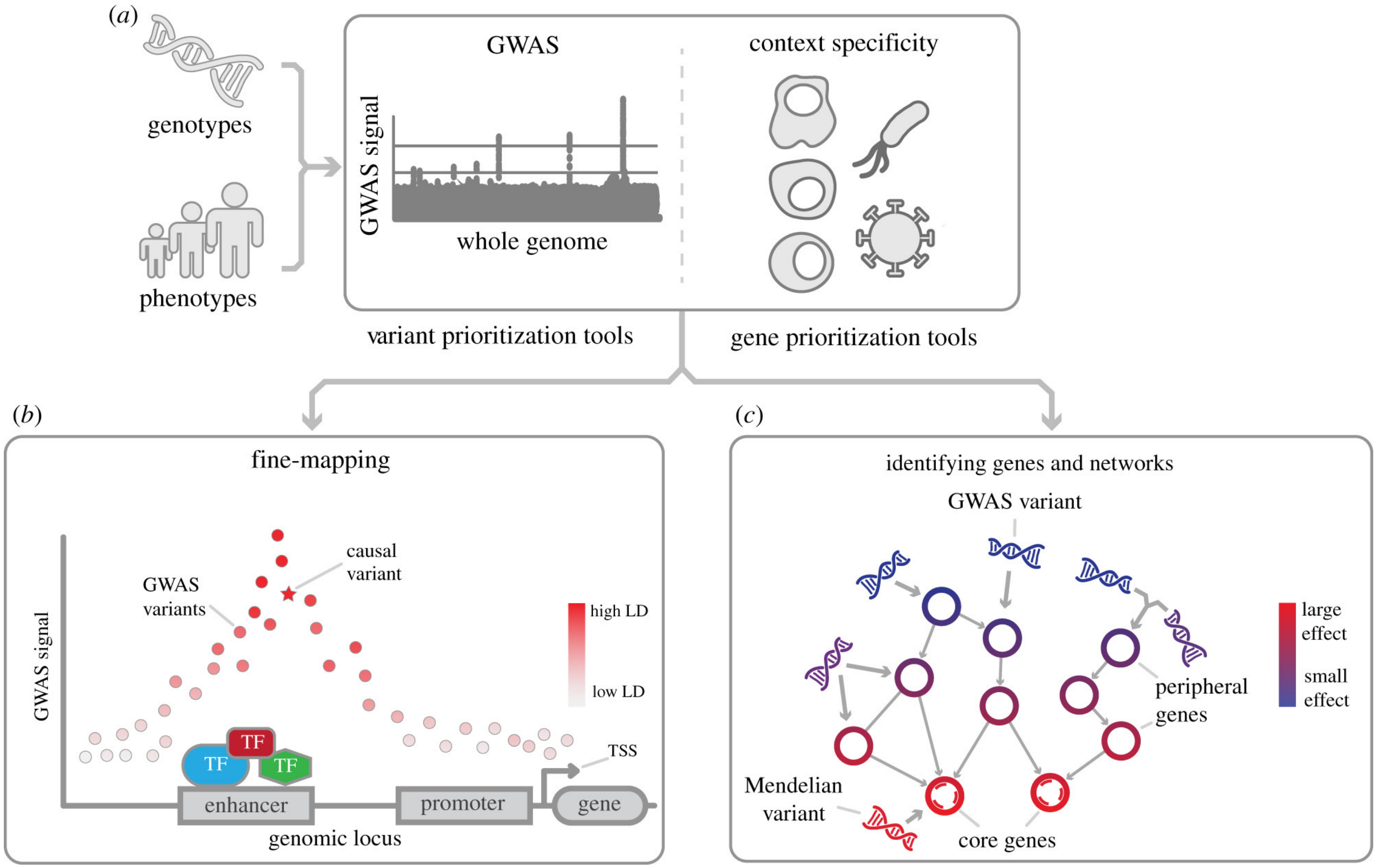
Hypothetical Disease Mechanism



Hypothetical Disease Mechanism



Post GWAS Era Workflow



Broekema et al. Open Biology 2020

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
  - Correcting for LD, especially across ancestral populations
  - Correcting for gene size, pathway size, and number of variants

Candidate Gene vs. Agnostic

- Choices dependent on study goals
  → **scientific question and available data**

- Candidate pathway analysis
  - Hypothesis driven questions
    - Ex: “are oxidative stress pathways genetically different in individuals with cancer?
    - More common in SNPs than expression

- Agnostic
  - Exploratory analyses across knowledge base
  - Many choices for input
    - Gene-level replication of previously implicated variants
    - Polygenic risk score
    - Ex: “what pathways are these genetic associations aggregating in?”

Mapping SNPs to Genes

- All SNPs in physical proximity of each gene
  - Pros:
    - All/most genes represented
    - Some approaches use LD to help define boundaries
  - Cons:
    - Varying number of SNPs per gene
    - Many of the SNPs may dilute signal
    - Defining gene proximity can affect results
    - LD is population dependent
      - Need to match LD panel to study population
      - Need raw values or post hoc analysis with summary statistics

Incorporating Functional Info

- 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

- Hi-C (3D folding)
  - relationship between chromosome organization and genome activity
  - Pros:
    - Helps understand transcription and translation
  - Cons
    - Limited annotation resources
    - Short vs. long range assumptions

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) *(sentinel SNP)*
  - $1-(1-\text{Min(pvalue)})^{N}$
  - $1-(1-\text{Min(pvalue)})^{(N+1)/2}$

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?

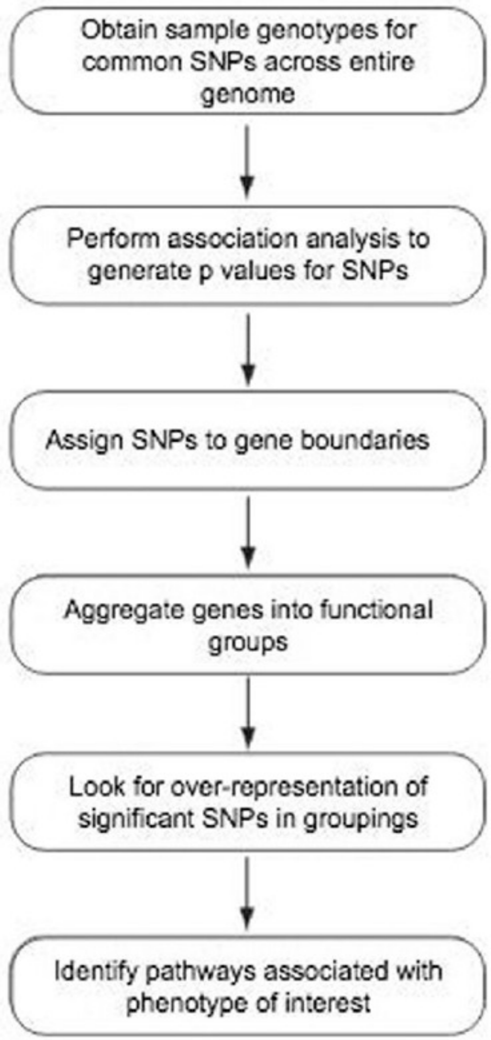
Choice of Pathways/Gene Sets

- 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

Overall Workflow



Obtain sample genotypes for common SNPs across entire genome

Perform association analysis to generate p values for SNPs

Assign SNPs to gene boundaries

Aggregate genes into functional groups

Look for over-representation of significant SNPs in groupings

Identify pathways associated with phenotype of interest

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

Current approaches

- Adding information about functional annotation to prioritize/select SNPs:
  - ICSNPathway (Zhang et al 2011)

- Use kth best SNP as the representative p-value combined with permutation testing
  - GSA-SNP2 (Nam et al. 2010, Yoon et al 2018)
  - Avoids bias from randomly significant SNPs
  - Loses power if functional SNPs are included

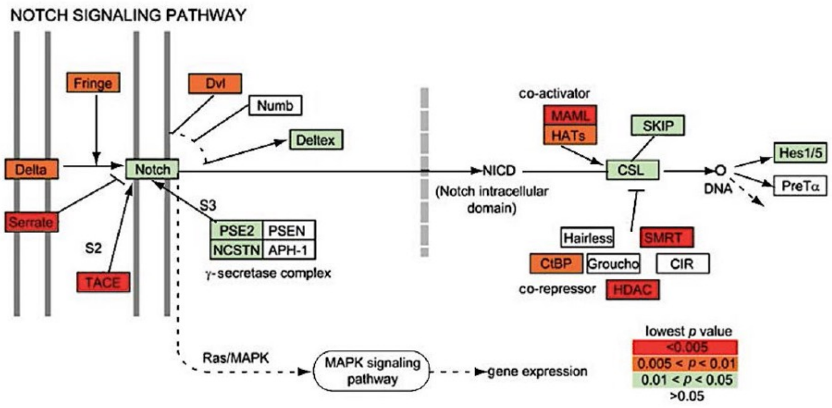
-

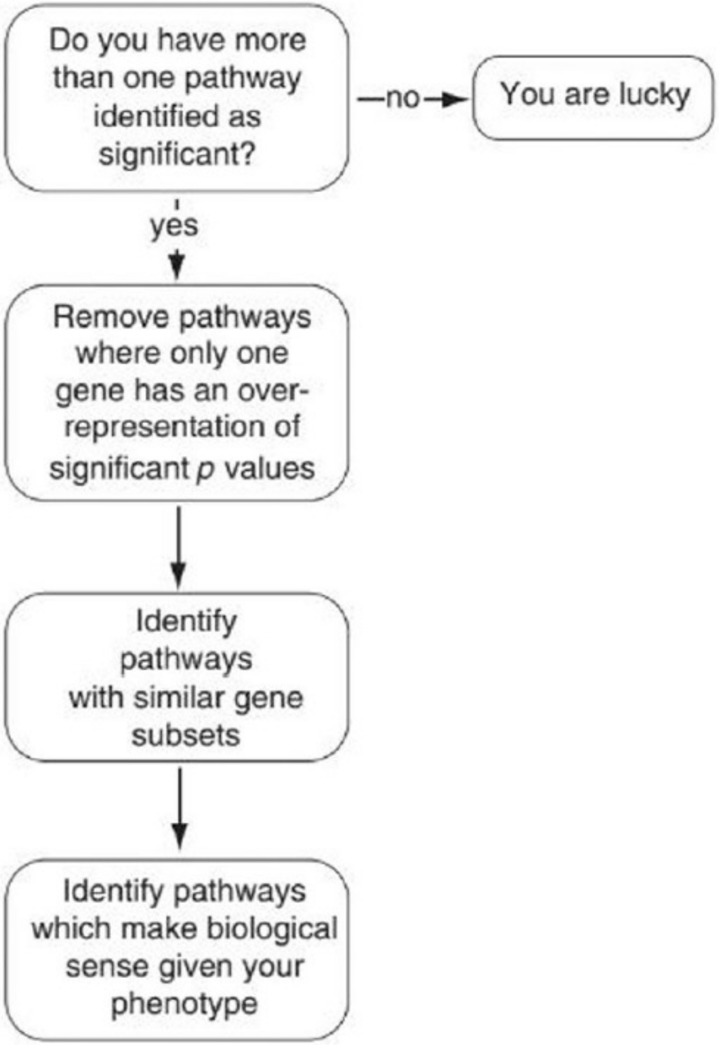
Current approaches

- Using ranked enrichment tests
  - Test the enrichment of a gene or gene set's SNPs at the significant end of a list of ranked SNP p-values
  - VEGAS2 (Liu et al 2010; Mishra & MacGregor 2015; Mishra & MacGregor 2017)
  - Accounts for all SNPs
  - Reduces effects of false positive GWAS through enrichment of moderately associate SNPs
  - Doesn't negate strongly significant associations
  - Computationally intensive

Follow-up

- Visualization
- Interpretation





Do you have more than one pathway identified as significant?

—no→ You are lucky

yes

Remove pathways where only one gene has an over-representation of significant *p* values

Identify pathways with similar gene subsets

Identify pathways which make biological sense given your phenotype

Some Specific Methods

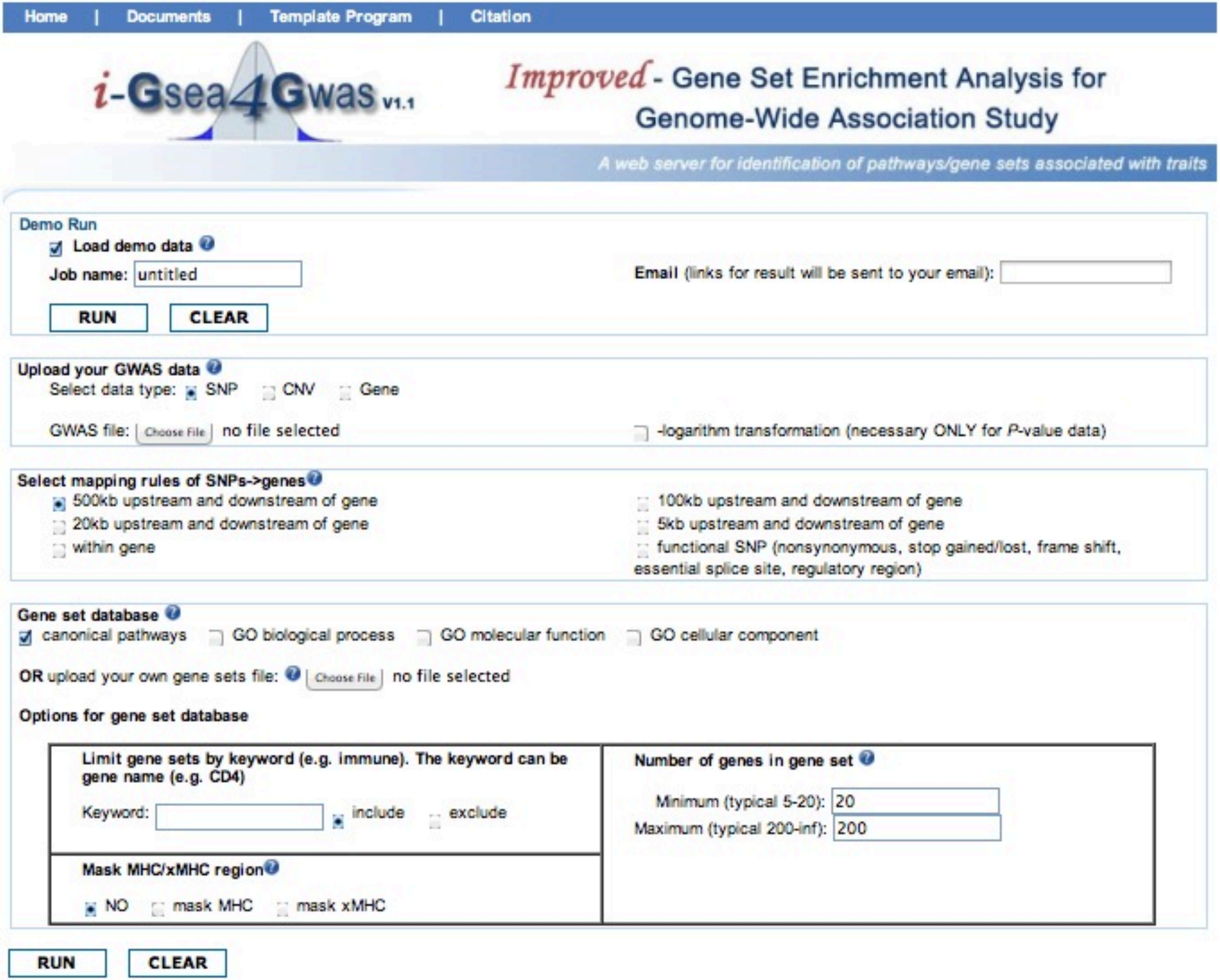
- i-GSEA4GWAS
- MAGENTA/FUMA
  - Meta-Analysis Gene-set Enrichment of variant Associations

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



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):

Upload your GWAS data

Select data type: ● SNP ○ CNV ○ Gene

GWAS file: no file selected

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

Select mapping rules of SNPs->genes

● 500kb upstream and downstream of gene
○ 100kb upstream and downstream of gene
○ 20kb upstream and downstream of gene
○ 5kb upstream and downstream of gene
○ within 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: 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

Number of genes in gene set

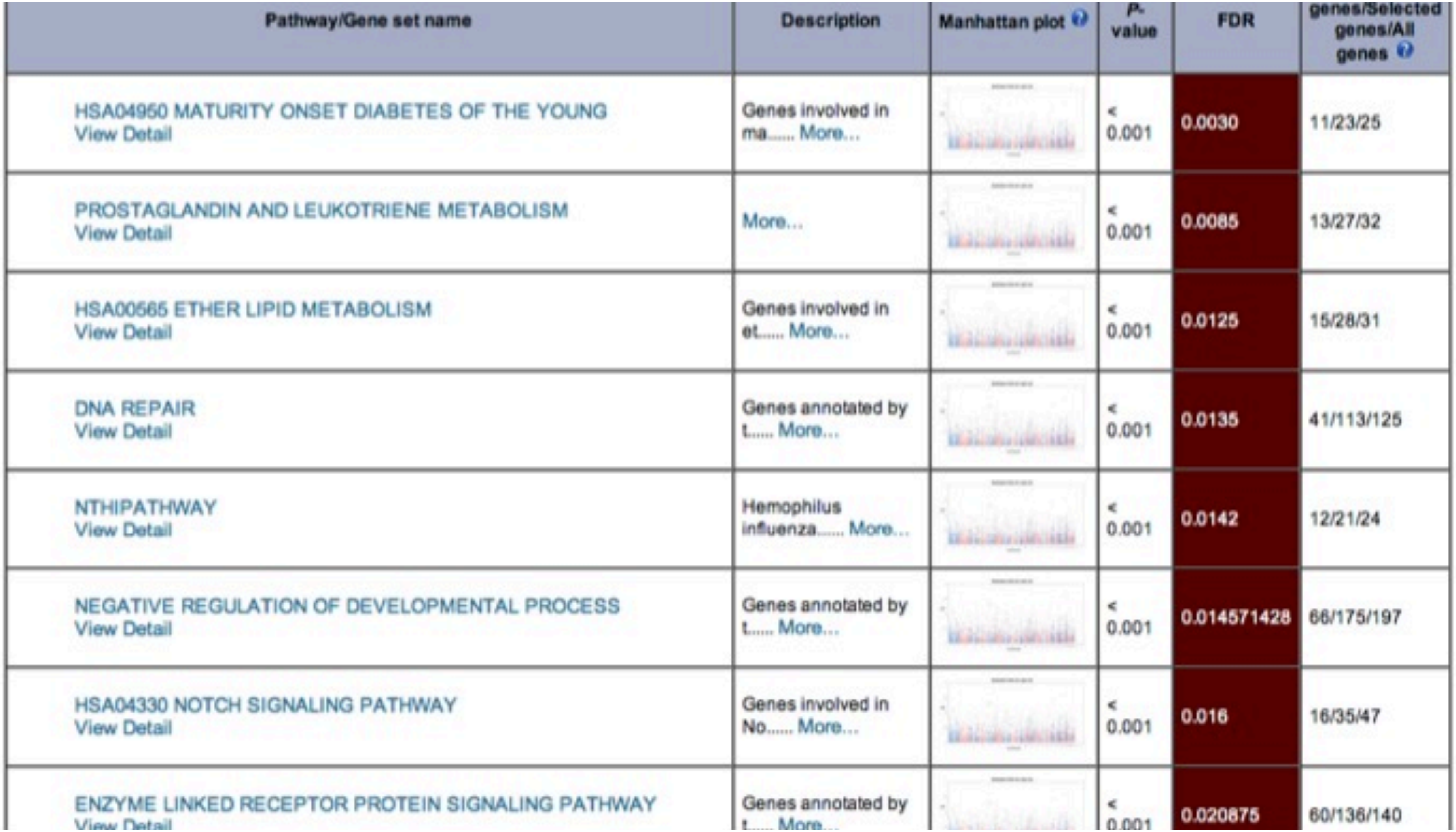
Minimum (typical 5-20): 20

Maximum (typical 200-inf): 200

Mask MHC/xMHC region

● NO ○ mask MHC ○ mask xMHC

Results

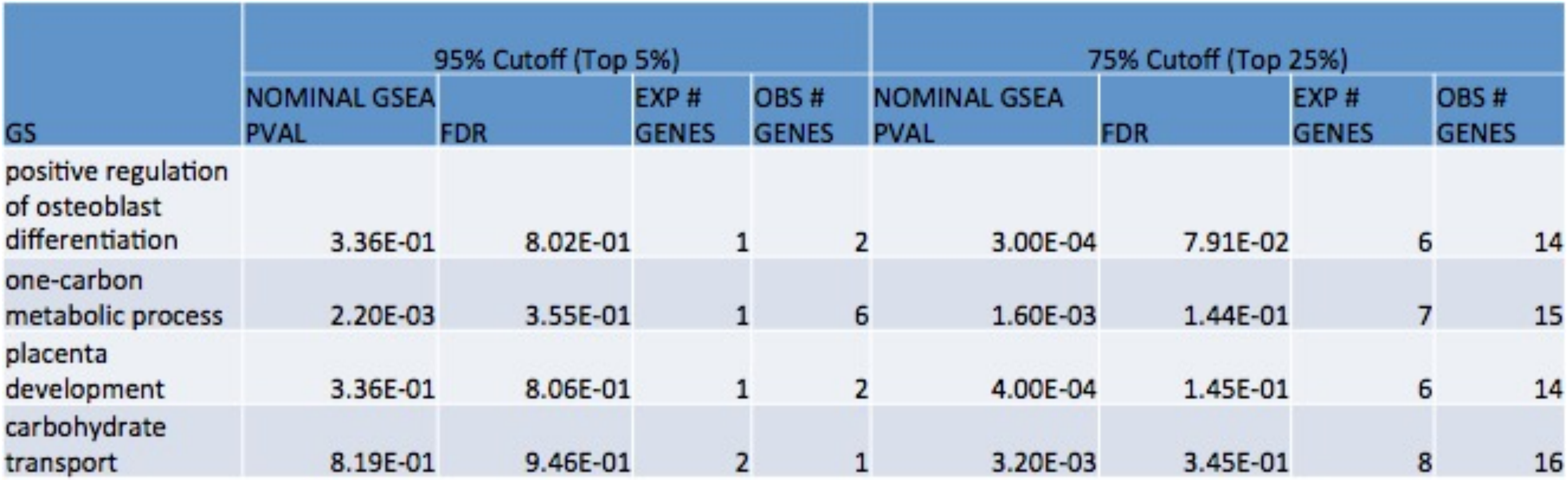


Pathway/Gene set name	Description	Manhattan plot	p-value	FDR	genes/Selected genes/All genes
HSA04950 MATURITY ONSET DIABETES OF THE YOUNG View Detail	Genes involved in ma...... More...		< 0.001	0.0030	11/23/25
PROSTAGLANDIN AND LEUKOTRIENE METABOLISM View Detail	More...		< 0.001	0.0085	13/27/32
HSA00565 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/FUMA

- 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

FUMA

- Watanabe et al, Nature Comm 2017

- Implementation of MAGENTA
- Functional annotation
- Pathway analysis
- Visualization
- ……

Other 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

Pathway Analysis for Rare Variants

- Low frequency (1% - 5% MAF) and rare variants (<1%) require additional considerations

- Off the shelf use of GWAS pathway methods may not be appropriate
  - Generally, rare variants need to be weighted to have any power
  - One and two stage options
    - Using variant level data
    - Collapsing variant level data into genes/regions/pathways

Pathway Analysis for Rare Variants

- Power is highly dependent on how closely the analysis plan matches the true underlying etiology
- Rare variant common disease (RVCD) hypothesis
  - Generally assumed RVs will have high effect sizes and/or direct functional consequences
  - Not always true
    - Ex: Missense mutations can have small effect sizes, with weak selective pressure

Pathway Analysis for Rare Variants

- Example methods:
  - aSPU
    - Pan et al American Journal of Human Genetics 2015
  - Smoothed functional principal components analysis
    - Zhao et al European Journal of Human Genetics 2015
  - Bayesian methods
    - Han et al 2019 bioRxiv
      doi: https://doi.org/10.1101/828061

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”
- Methods for rare variants are evolving

Questions?