

# 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](mailto:alison.motsinger-reif@niehs.nih.gov)

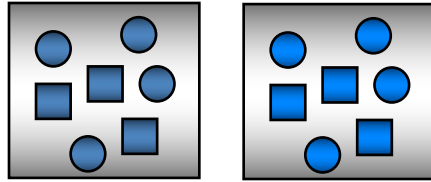
# Goals

- Methods for GWAS with SNP chips
  - Integrating expression and SNP information

# Many Shared Issues

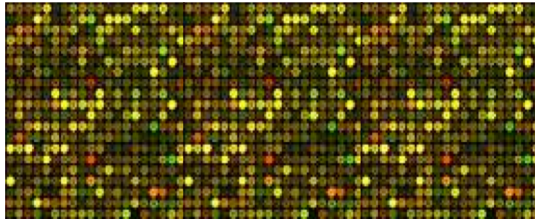
- Many of the issues/choices/methodological approaches discussed for microarray data are true across all “-omics”
- Many methods have been readily extended for other omic data
- There are several biological and technological issues that may make just “off the shelf” use of pathway analysis tools inappropriate

# Genome-Wide Association Studies



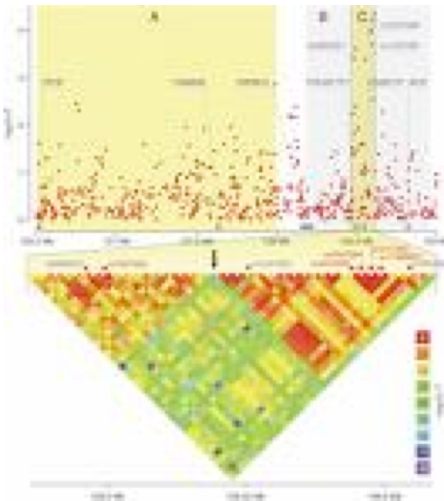
Population resources

- trios
- case-control samples



Whole-genome genotyping

- hundreds of thousands or million(s) of markers, typically SNPs



Genome-wide Association

- single SNP alleles
- genotypes
- multimarker haplotypes

# Advantages of GWAS

- Compared to candidate gene studies
  - unbiased scan of the genome
  - potential to identify totally novel susceptibility factors
- Compared to linkage-based approaches
  - capitalize on all meiotic recombination events in a population
    - Localize small regions of the chromosome
    - enables rapid detection causal gene
  - Identifies genes with smaller relative risks

# Concerns with GWAS

- Assumes CDCV hypothesis
- Expense
- Power dependent on:
  - Allele frequency
  - Relative risk
  - Sample size
  - LD between genotyped marker and the risk allele
  - disease prevalence
  - .ultiple testing
  - .....
- Study Design
  - Replication
  - Choice of SNPs
- Analysis methods
  - IT support, data management
  - Variable selection
  - Multiple testing

# Successes in GWAS Studies

- 33989 GWAS papers published to date (GWAS catalogue)
- Big Finds:
  - In 2005, it was learned through GWAS that age-related macular degeneration is associated with variation in the gene for complement factor H, which produces a protein that regulates inflammation (Klein et al. (2005) *Science*, 308, 385–389)
  - In 2007, the Wellcome Trust Case-Control Consortium (WTCCC) carried out GWAS for the diseases coronary heart disease, type 1 diabetes, type 2 diabetes, rheumatoid arthritis, Crohn's disease, bipolar disorder and hypertension. This study was successful in uncovering many new disease genes underlying these diseases.

# More Successes

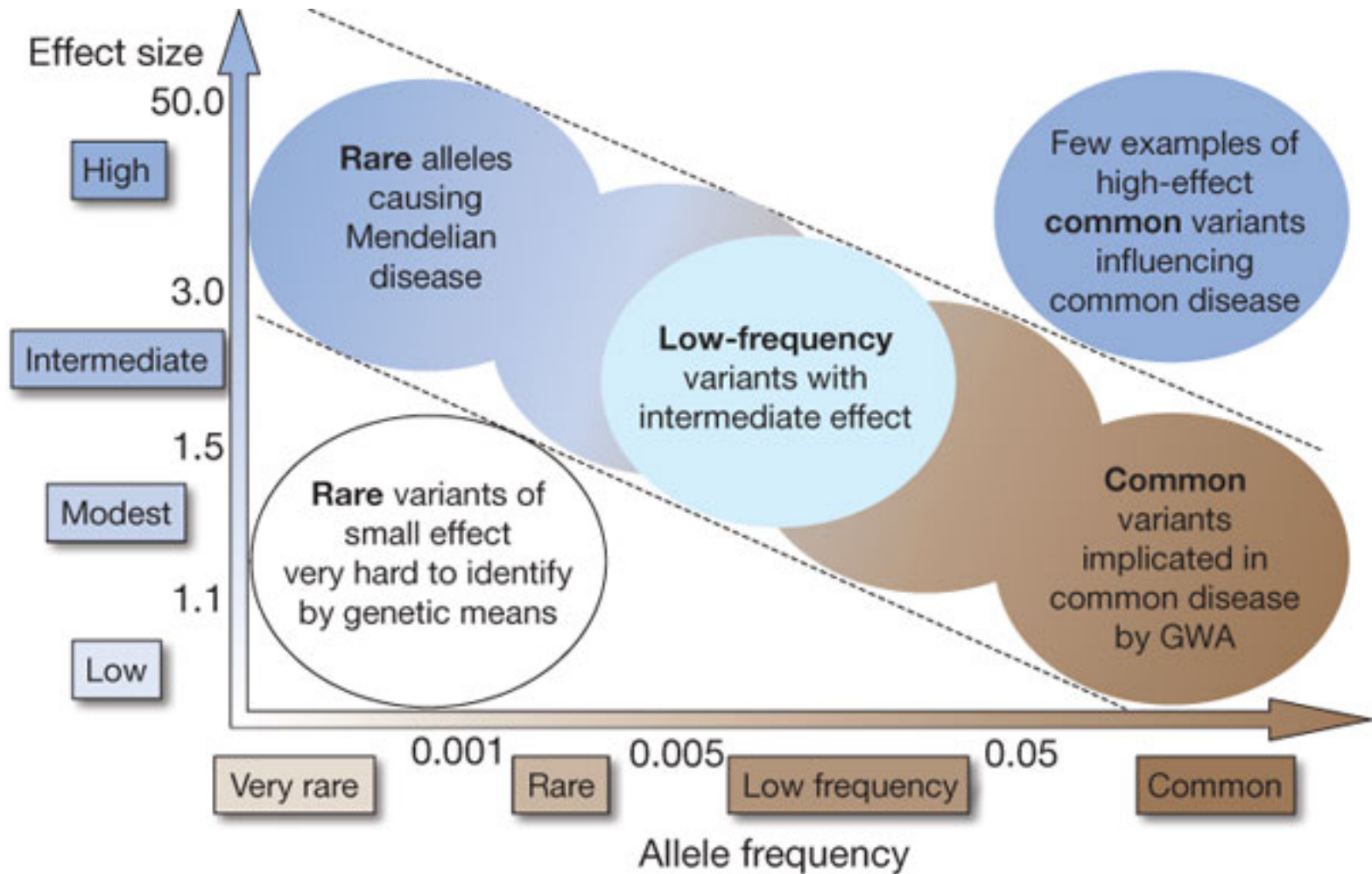
- Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nat Genet.* 2007
- Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Wellcome Trust Case Control Consortium Nature.* 2007;447;661-78
- Genomewide association analysis of coronary artery disease. *Samani et al. N Engl J Med.* 2007;357;443-53
- Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Parkes et al. Nat Genet.* 2007;39;830-2
- Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Todd et al. Nat Genet.* 2007;39;857-64
- A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Frayling et al. Science.* 2007;316;889-94
- Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Zeggini et al. Science.* 2007;316;1336-41
- Scott et al. (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science,* 316, 1341–1345.
- .....



# Limitations

- For many diseases, the amount of trait variation explained by even the successes is way below the estimated heritability.
- Assumptions underlying GWAS are not true for all diseases.

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

# Reasons GWAS Can Fail

even if well-powered and well-designed....

- Alleles with small effect sizes
- Rare variants
- Population differences
- Epistatic interactions
- Copy number variation
- Epigenetic inheritance
- Disease heterogeneity
- .....

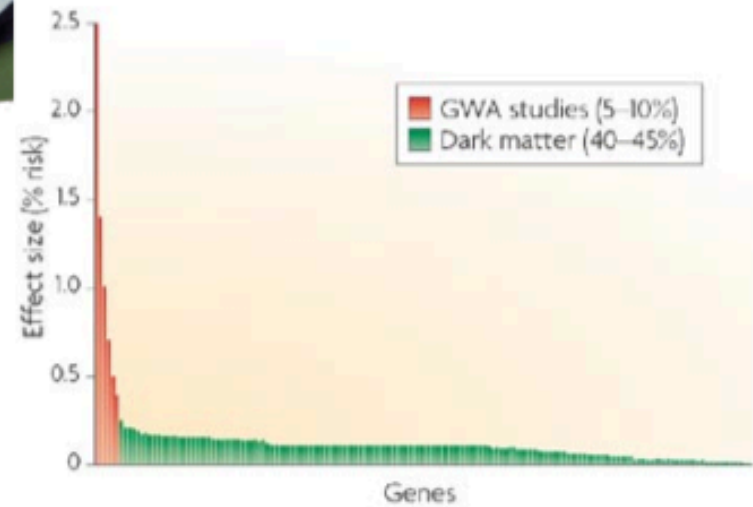
# Missing Heritability

NEWS FEATURE PERSONAL GENOMES

NATURE | 454 | November 2008



## The case of the missing heritability



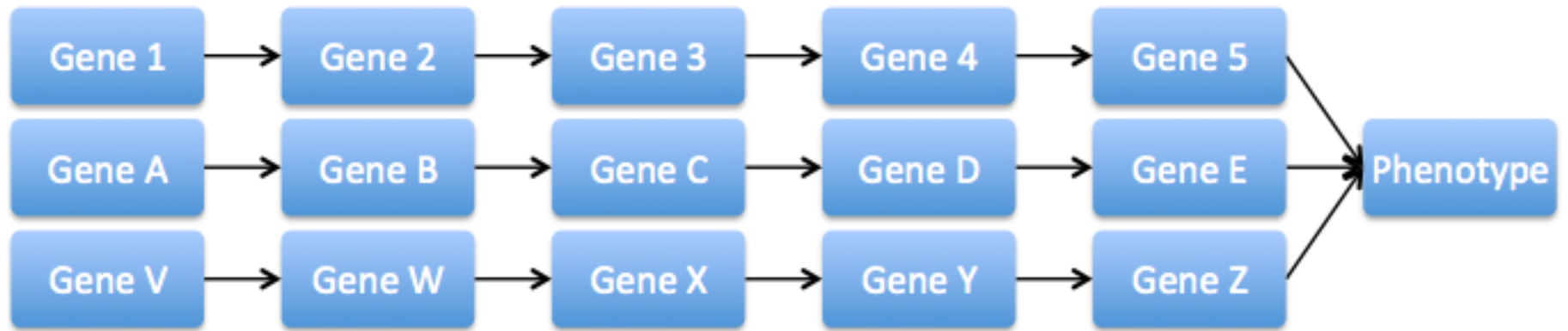
Nature Reviews | **Genetics**

Lusis et al, 2008

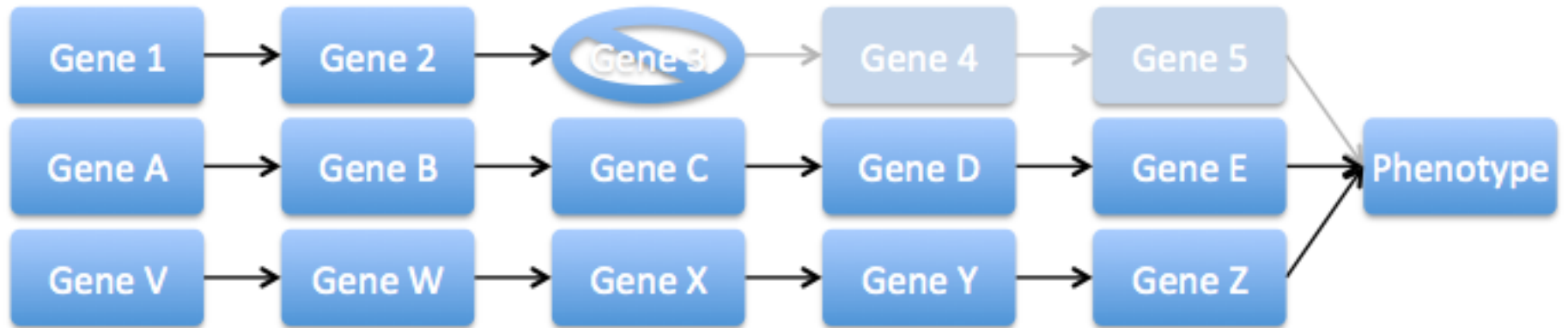
# Possible Association Models

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

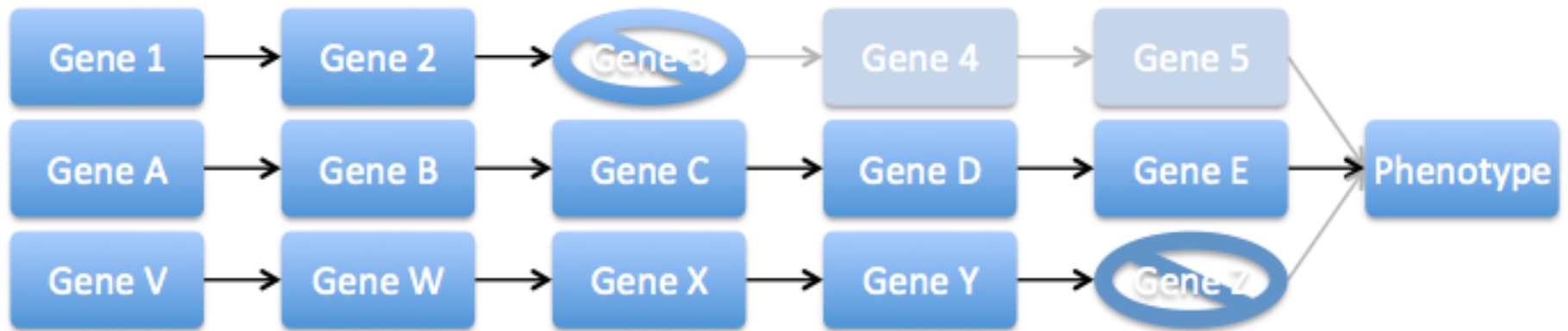
# Hypothetical Disease Mechanism



# Hypothetical Disease Mechanism

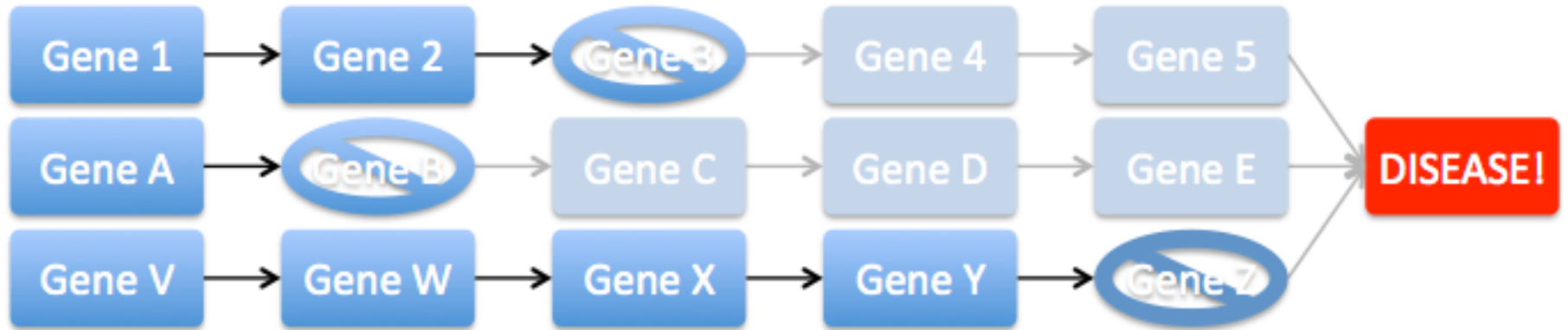


# Hypothetical Disease Mechanism





# Hypothetical Disease Mechanism

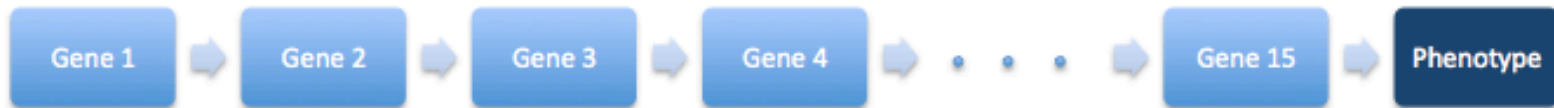


# Hypothetical Disease Mechanism

- For each gene probability of knockout =  $0.2^2 = 0.04$
- Probability of disease:
  - Pathway knocked out = 0.4
  - Pathway in tact = 0.2
- Sample Size = 2000 cases, 2000 controls
- Power:

Best SNP		Pathway	
Significant	Suggestive	0.001	0.005
0.001	0.05	0.42	0.69

# Linear Pathway



- For each gene probability of knockout =  $0.2^2 = 0.04$
- Probability of disease:
  - Pathway knocked out = 0.4
  - Pathway in tact = 0.2
- Sample Size = 2000 cases, 2000 controls
- Power:

Best SNP		Pathway		Pathway (mis-specified)*	
Significant	Suggestive	0.001	0.005	0.001	0.005
0.002	0.02	0.94	0.98	0.51	0.73

\*Tested pathway includes 15 genes not in simulated pathway

# 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  $pvalue < 0.05$
  - $\text{Mean}(-\log_{10}(pvalue))$
  - $\text{Min}(pvalue)$
  - $1-(1-\text{Min}(pvalue))^N$
  - $1-(1-\text{Min}(pvalue))^{(N+1)/2}$

# 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)

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

$$p = \mathbf{P}(\chi_{(2k)}^2 > 2V)$$

# Next approaches

- Additive model: 
$$\log\left(\frac{p}{1-p}\right) = \sum_{g_i \in G} \beta_i n_i$$
  - 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



## 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:

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

**RUN**

**CLEAR**

### 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  
 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: [?](#)  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):

Maximum (typical 200-inf):





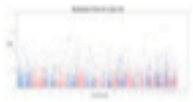





Mask MHC/xMHC region [?](#)

NO  mask MHC  mask xMHC

**RUN**

**CLEAR**

# Results

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

GS	95% Cutoff (Top 5%)				75% Cutoff (Top 25%)			
	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

# 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 than in gene expression data, though still rely on the same resources.... not necessarily so for other 'omics"

# Adding in Gene Expression Data

- Many motivating reasons to combine/integrate data from multiple “-omes”
- Expression and SNP data is most commonly done
  - Though methods could be applied to combine other “-omics”
- Generally make assumptions about central dogma

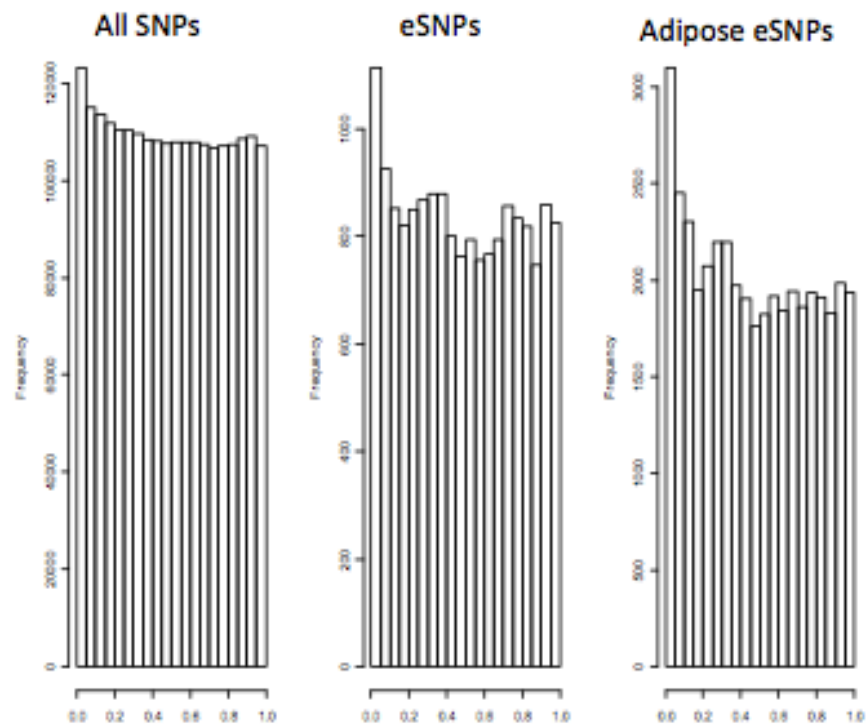
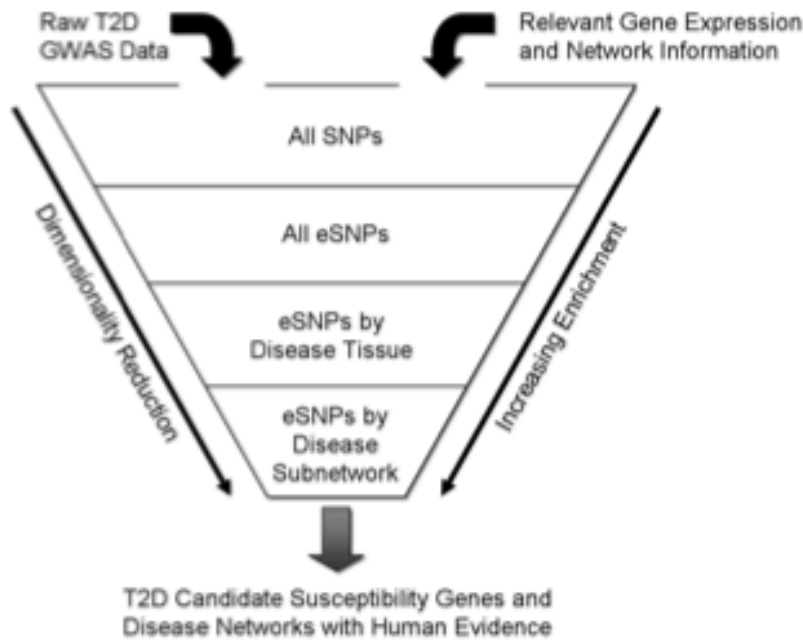


# Genetics of Gene Expression

- Schadt, Monks, et al. (*Nature* 2003) & Morley, Molony, et al. (*Nature* 2004) showed that gene expression is a heritable trait under genetic control
- Identifying expression-associated SNPs (eSNPs) can identify SNPs which are associated with biological function
- For significant GWAS “hits” eSNPs can suggest candidate genes and possibly information about direction of association

# eSNPs can enrich p-values in GWAS studies

## Example: T2D Data



# Considerations on Filtering/Mining Data

- Trade-off between un-biased discovery and improving power (improving enrichment)
- Gold standard for publication is  $p\text{-value} < 5e-8$  PLUS replication
- For hypothesis generation or biological data mining might be willing to accept more Type I error
- Possible approaches:
  - Gold standard only
  - Gold standard then mining “biological” SNPs (e.g. all SNPs near genes, eSNPs, eSNPs by tissue, etc)
  - Partitioning SNPs into sets by prior information



# Considerations: Multiple Test Correction

- Can be valid to test hypotheses in a partitioned fashion if:
  1. The partitions are specified **before** you look at the data
  2. Your multiple testing procedure controls the overall error rate

## 5% P-value vs 5% FDR

- P-value -> Over a large number of times the experiment is repeated, 5% of the time we'll identify 1 or more false positive SNPs
- FDR -> 5% of identified SNPs are false positives

# Partitioned SNP Testing (p-value)

- Can be beneficial if you have a small number of high(er)-confidence SNPs
- Genomewide significance threshold:  $5e-8 = 0.05/1,000,000$
- Example: 10,000 eSNPs
  - eSNP threshold:  $0.025/10,000 = 2.5e-6$
  - Remaining SNP threshold:  $0.025/990,000 = 2.53e-8$

# Partitioned Testing (FDR)

- Simple way to control error over multiple partitions
- Controlling FDR at level  $\xi$  in each (non-overlapping) set, results in overall FDR  $\xi$



# eSNPs: Computing your own

- eSNP analyses are just GWAS's with continuous traits, but 1000's of them
- Approaches:
  - Frequentist:
    - Linear Regression
      - Outlier sensitive, can adjust for covariates
    - Robust Regression
      - Outlier resistant, can adjust for covariates, more computationally demanding
    - Kruskal-Wallis
      - Nonparametric (outlier resistant), difficult to adjust for covariates
  - Bayesian:
    - More resistant to outlier effects than linear regression, but require setting priors on each parameter
    - Some software available:
      - Bimbam
      - SNPTEST

# eSNPs: A note on computation

- eSNP analysis is extremely resource intensive in both processor time and storage
- Computation requires a cluster (not possible on a desktop machine)
- Storage:  $N_{\text{markers}} \times N_{\text{expression traits}}$  is typically large
  - One approach is to store only results with pvalue < some threshold

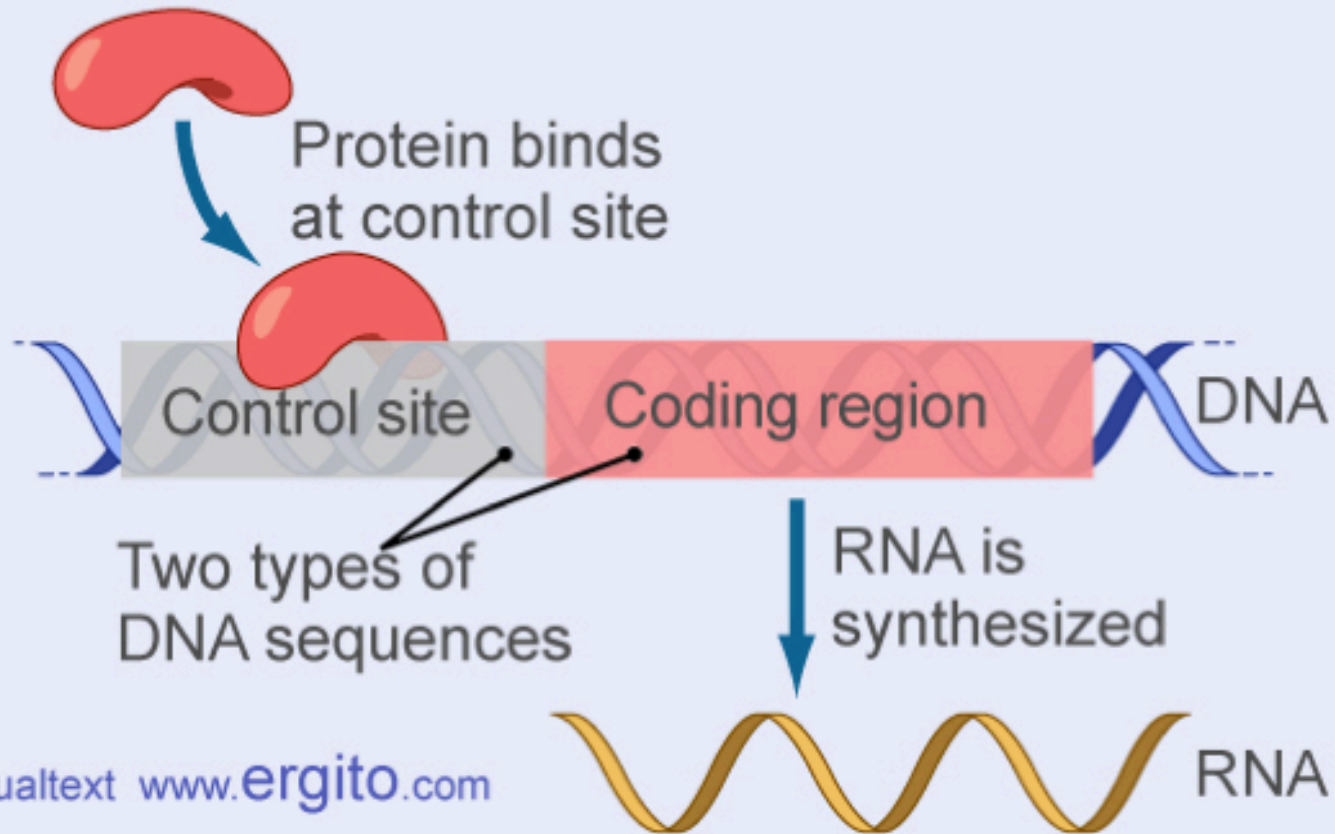
# eSNP Discovery

- eSNPs near gene location are easier to find
  - Real biological effects (*cis* regulation)
  - Fewer hypothesis tests relative to genomewide
- Typical approach is to identify local (proximal) eSNPs and distant (distal) eSNPs in separate steps
- Controlling each at fixed FDR,  $\xi$ , controls the overall FDR at  $\xi$
- Choice of proximal window can effect eSNP discovery



# Cis vs Trans Regulation

Proteins bind to *cis*-acting control sites





## Aside: Cis/Trans vs Proximal/Distal

- *Cis* element -> Regulates transcription only of copy sharing same DNA strand
- *Trans* element -> Regulates transcription of both DNA strands
- *Trans* elements can be near the gene, *cis* elements can be far from gene (on MB scale)
- Proximal (near) and distal (far) more accurate when referring variants associated with expression

# eSNPs: Publically Available

- Databases:
  - [www.scandb.org](http://www.scandb.org)
  - <http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/>
- Available in Synapse ([synapse.sagebase.org](http://synapse.sagebase.org)):
  - Harvard Brain- Brain, multiple disease
  - Kronos Phase I- Brain, alzheimer's
  - Human Liver Cohort- Liver, population sample
- ...

# Motivation for Integrated Analysis

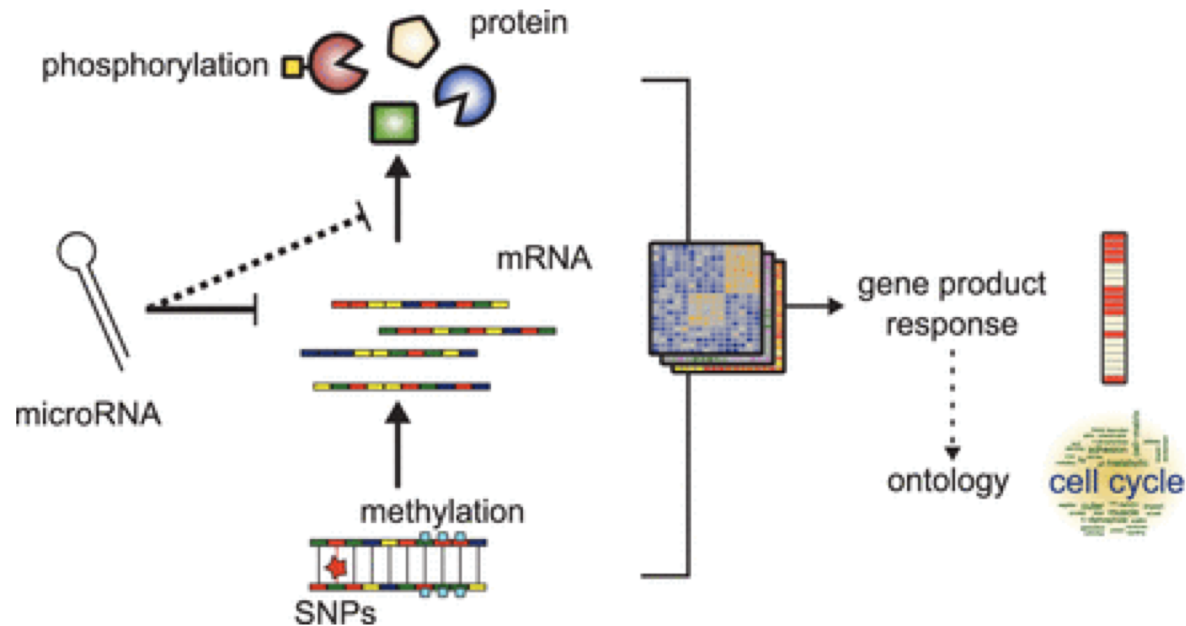
- Newer approaches will allow you to not do partitioned/filtered analysis, and leverage information across datatypes
- New technologies allow for more ready integration
  - Ex. RNA-Seq
  - Dropping costs allow for more datatypes to be collected simultaneously
  - Biobanking effort are storing more tissues

# Motivation for Integrated Analysis

- Naturally allow Bayesian approaches for identifying priors or jointing modeling data
- Several new approaches proposed
  - Methods that were developed for eSNPs are readily extended across data types
  - Other approaches take into account similarities between/withing phenotypes
    - Several an ontology jointly representing disease risk factors and causal mechanisms based on GWAS results
    - Proposed ontology is disease-specific (nicotine addiction and treatment) and only applicable to very specific research questions
  - More later on “different issues for –omics”

# Motivation for Integrated Analysis

- Methods are largely relying on central dogma assumptions that do not always hold



# Summary

- Pathway and gene set analysis has been extended to SNP and SNV data
- Some annotation resources are readily adapted, but a new series of choices are available
- Software packages for GWAS pathway analysis are maturing
- Advances in approximation for permutation testing will make these tools more computationally tractable
- Many of the same issues with missing annotation, etc. are still a concern

# Summary

- Integration of SNP level and eSNP data has been highly successful, and helps motivate the integration of other “-omes” in analysis
- Such integration will be dependent on the quality of the annotation that it relies on
- Next, we will talk about specific concerns for different datatypes
- Issues will compound in integrated analysis...

Questions?