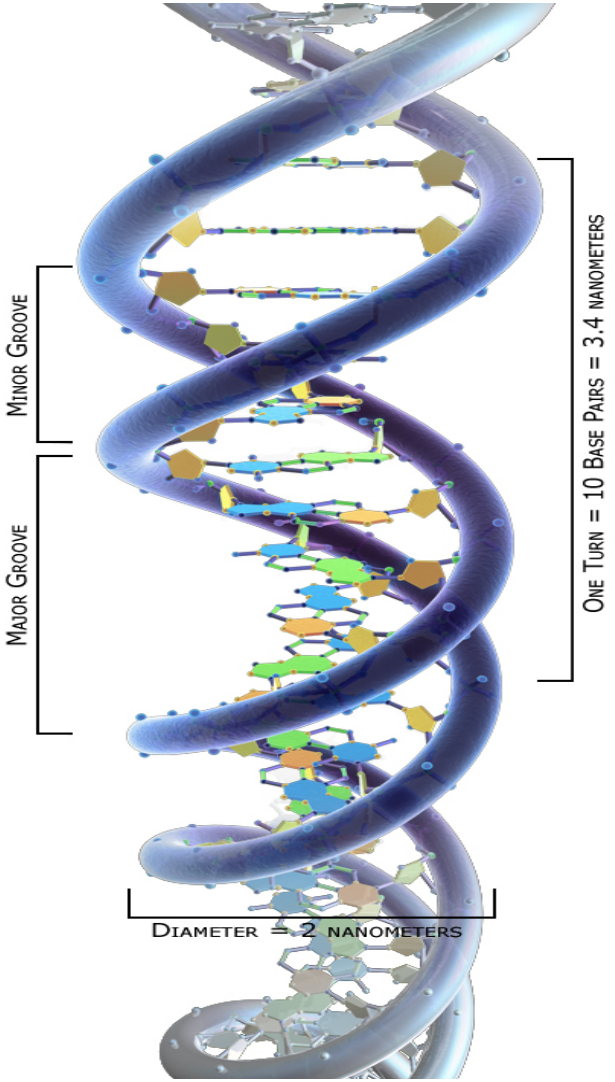


Pathways Analysis: databases and methods for genetics.

DR. GEROME BREEN

UPDATED BY: HELENA A. GASPAR

MRC Social, Genetic & Developmental Psychiatry Centre

Institute of Psychiatry at the Maudsley

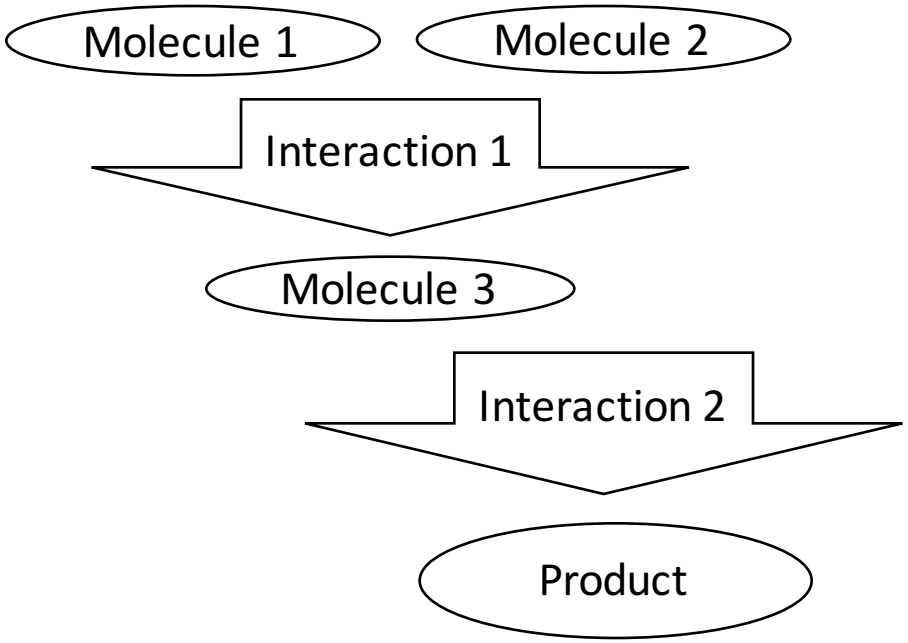
King's College London
University of London

Outline

1. What are pathways?
2. Pathway Resources
3. Pathways Analysis and GWAS
4. PGC CDG GWAS Pathway Analysis
5. Drug/GWAS Pathway Analysis

1. What are pathways?

A series of actions among molecules leading to a product or a change.



A typical pathway: WNT Signalling

A typical pathway: WNT Signalling

Wnt ligand binds to to a Fz receptor

↓

(…)
Unphosphorylated β-catenin translates into the nucleus

↓

Regulation of target genes

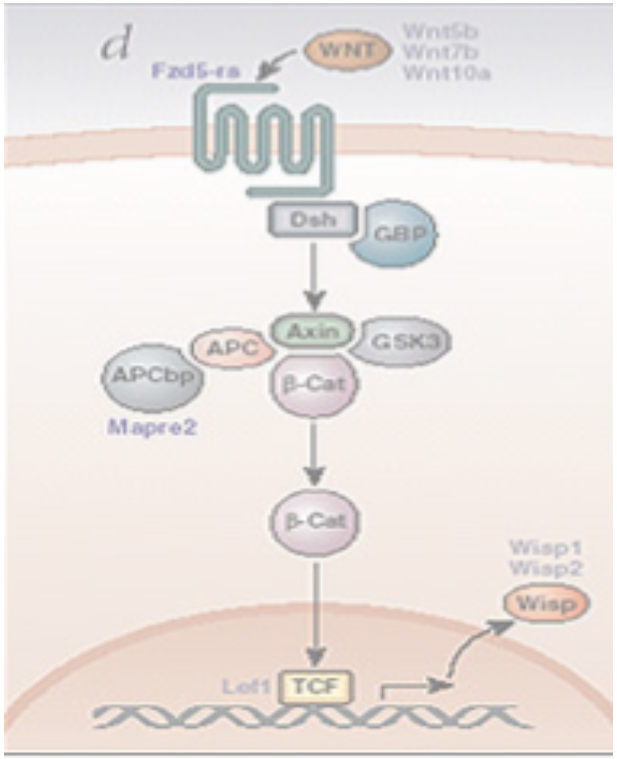
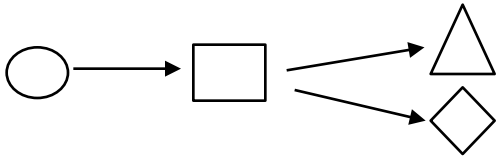


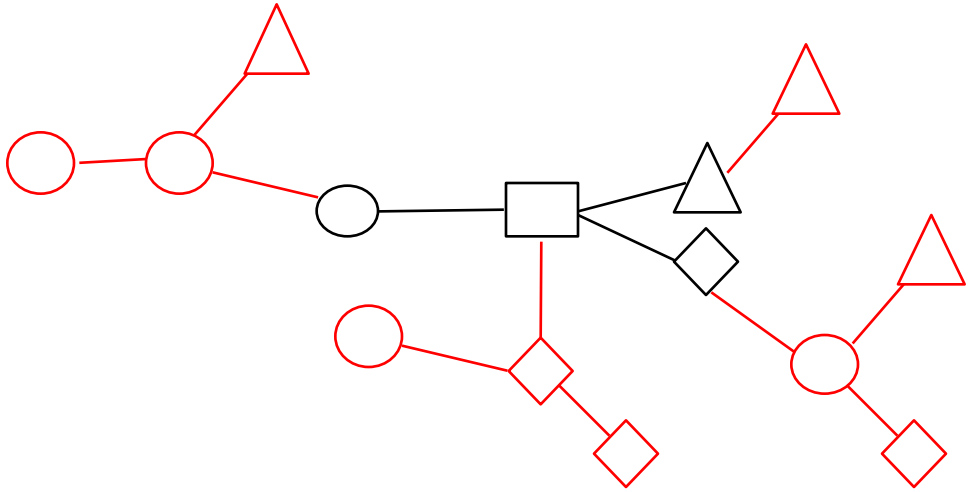
Image from Suzuki et al 2002 Nature Genetics 32 166

Pathways v.s. Networks

Pathway: series of actions among molecules leading to a product or a change.

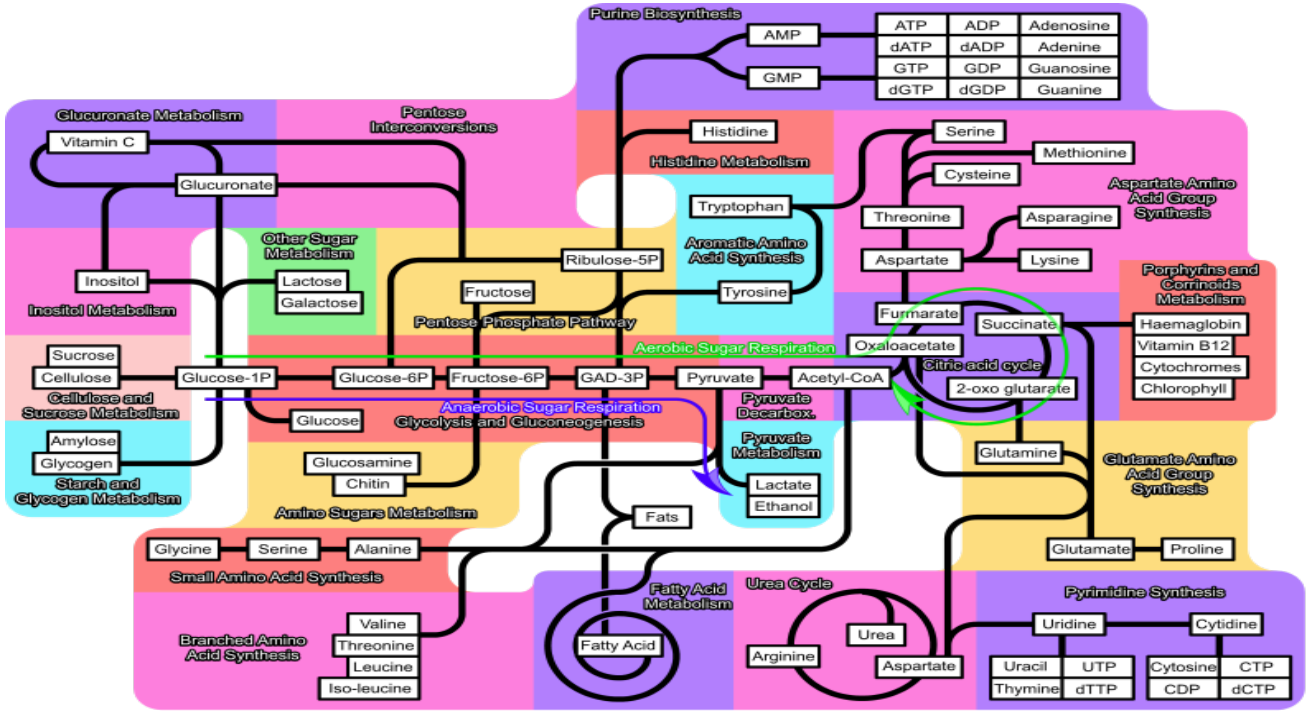


Network: a collection of nodes (molecules) and edges (type of interaction).



Pathways v.s. Networks

In biology, everything is interconnected in a network. But pathways provide intuitive views.



Pathways v.s. Networks

Early pathways were built by studying biochemical reactions of individual proteins, measuring the activity, learning the substrates and products, joining them to the next enzyme.

New technologies allow simultaneous measurement of *tens of thousands of different molecules*.

This reveals that many biological pathways are interconnected. They can work together or against each other.

Scientists who study these large scale interactions sometimes refer to the field as 'systems biology'.

When multiple biological pathways interact = a biological network.

Biologists often work with canonical pathways, those that represent the well-understood part of the entire network.

Different types of pathways

Metabolic pathways:

- provide the energy and materials

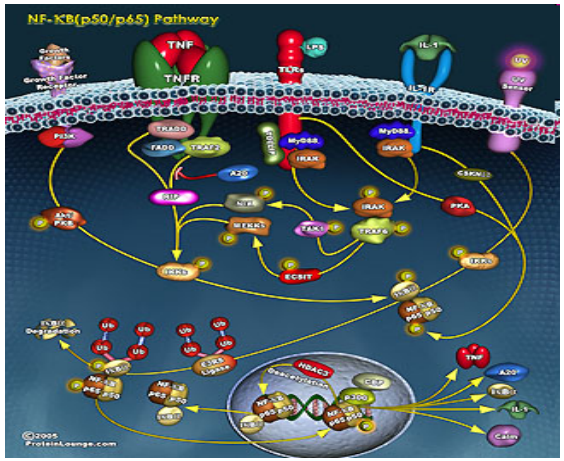
Signalling pathways:

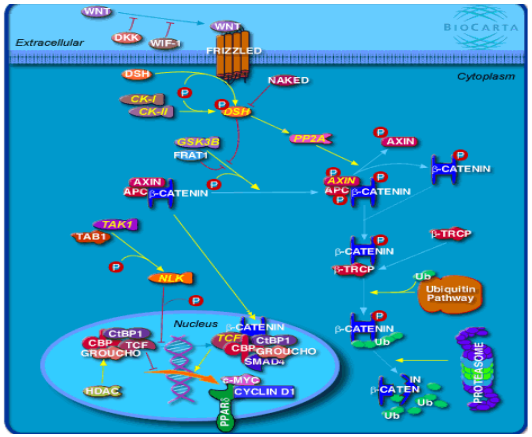
- sense the outside, coordinate activities within and between cells

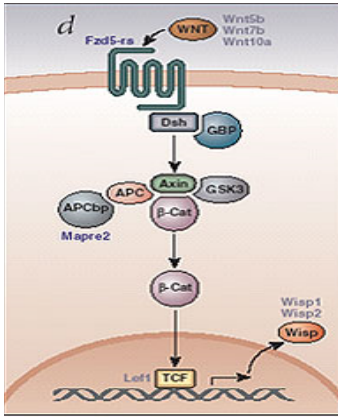
Gene regulatory pathways:

- control processes, set limits, control the molecular composition of cells

Different representations for one pathway







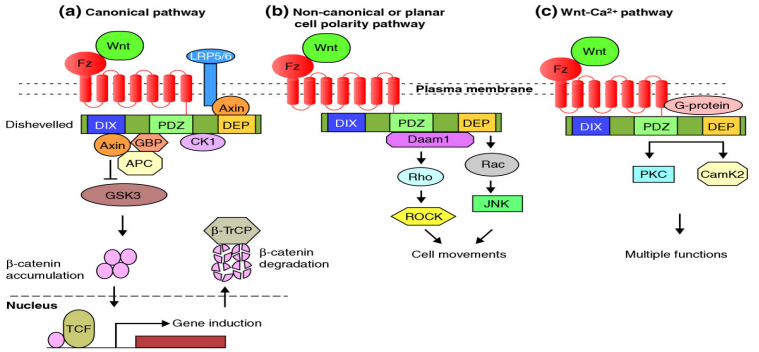


Image from Suzuki et al 2002 Nature Genetics 32 166, Habas & Dawid 2005 J Biol 4:2, BioCarta, SigmaAldrich Pathfinder

2. Pathway resources: private companies

e.g. Ingenuity Pathway Analysis & GeneGO

Manually curated

Subset of literature covered

Pre-generated networks

Make own networks

Pathway resources: free databases

Metabolic pathways:

- Reactome, GO, KEGG, Pathway Commons (…)

Signalling pathways:

- Reactome, GO, Panther (…)

Gene regulatory networks:

- ConsensusPath-DB, GeneMania (…)

Diagrams:

- WikiPathways, KEGG, BioCarta, (…)

Other databases for protein-protein interactions, protein-compound interactions…

Pathway resources: free databases

Main resources:

- REACTOME
- Gene Ontology (GO)

Reactome

Free, online & open-source

Curated resource of core pathways and reactions in human biology

Authored by expert biological researchers

Maintained by the Reactome editorial staff

Cross-referenced to

- NCBI Entrez Gene, Ensembl and UniProt databases
- UCSC and HapMap Genome Browsers
- KEGG Compound and ChEBI small molecule databases
- PubMed, and GO

Human data used to infer orthologous events in 22 non-human species

Tools for data analysis include Skypainter and Biomart

Encoding Journal Information

Nature 407(6805):770-6.The Biochemistry of Apoptosis.

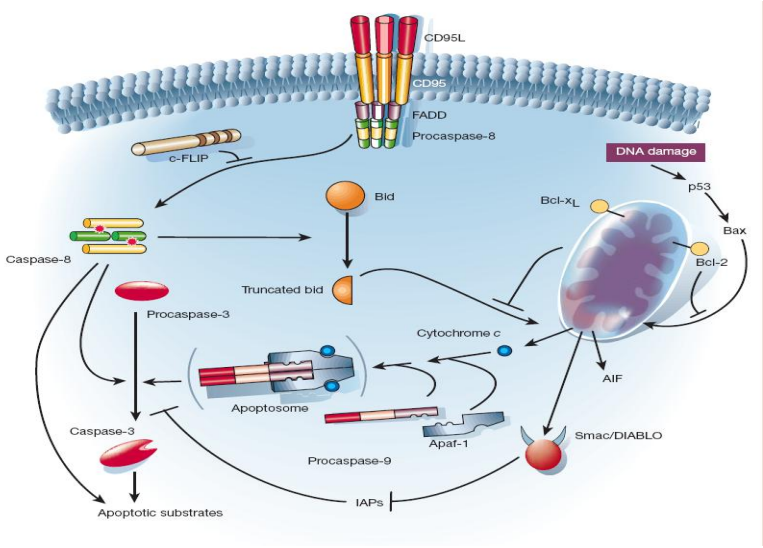
"Caspase-8 is the key initiator caspase in the death-receptor pathway. Upon ligand binding, death receptors such as CD95 (Apo-1/Fas) aggregate and form membrane-bound signalling complexes (Box 3). These complexes then recruit, through adapter proteins, several molecules of procaspase-8, resulting in a high local concentration of zymogen. The induced proximity model posits that under these crowded conditions, the low intrinsic protease activity of procaspase-8 (ref. 20) is sufficient to allow the various proenzyme molecules to mutually cleave and activate each other (Box 2). A similar mechanism of action has been proposed to mediate the activation of several other caspases, including caspase-2 and the nematode caspase CED-3 (ref. 21)."

How can I access the pathway described here and reuse it?

Reactome

Encoding Figures

A picture:



- Just pixels!
- Omits key details
- Facts or hypothesis?

→ ***Reactome***

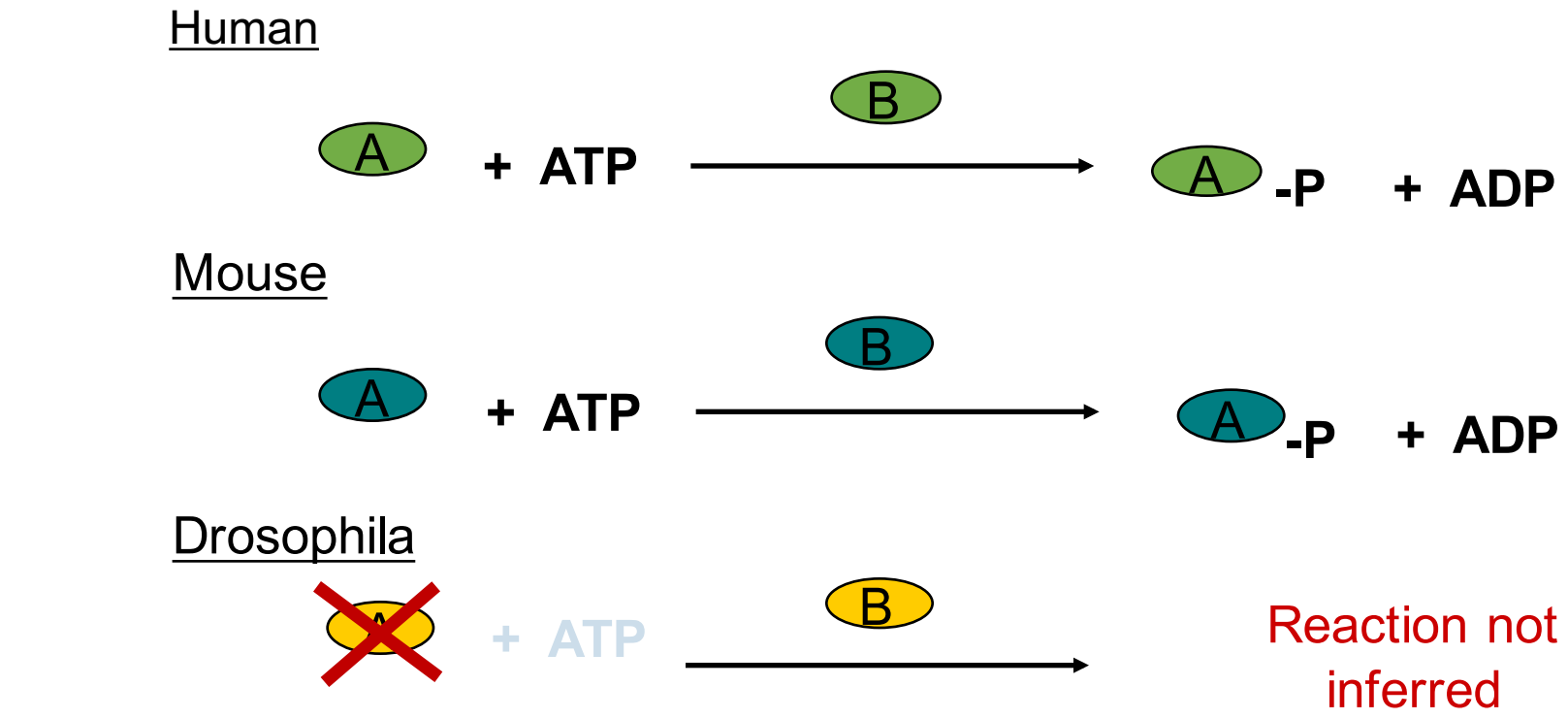
Reactome is like many other databases…

Extensively cross-referenced

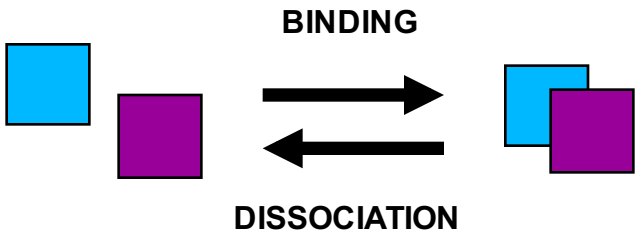
Tools for data analysis – Pathway Analysis, Expression Overlay, Species Comparison, Biomart…

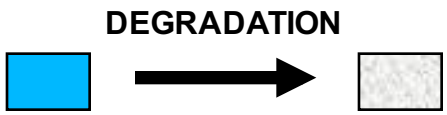
Used to infer orthologous events in 20 other species.

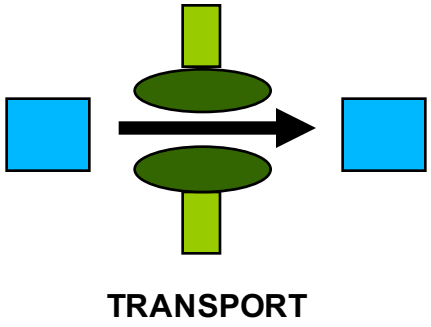
Data Expansion – Projecting to Other Species

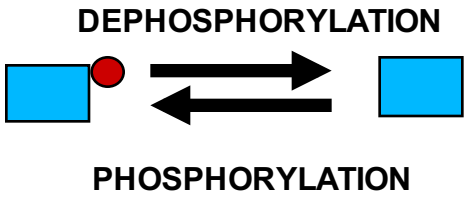


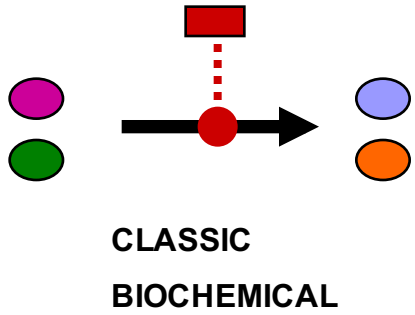
Reactome: Theory - Reactions











Reactome: Exportable Protein-Protein Interactions (PPIs)

Inferred from complexes and reactions

Interactions between proteins in the same complex, reaction, or adjoining reaction

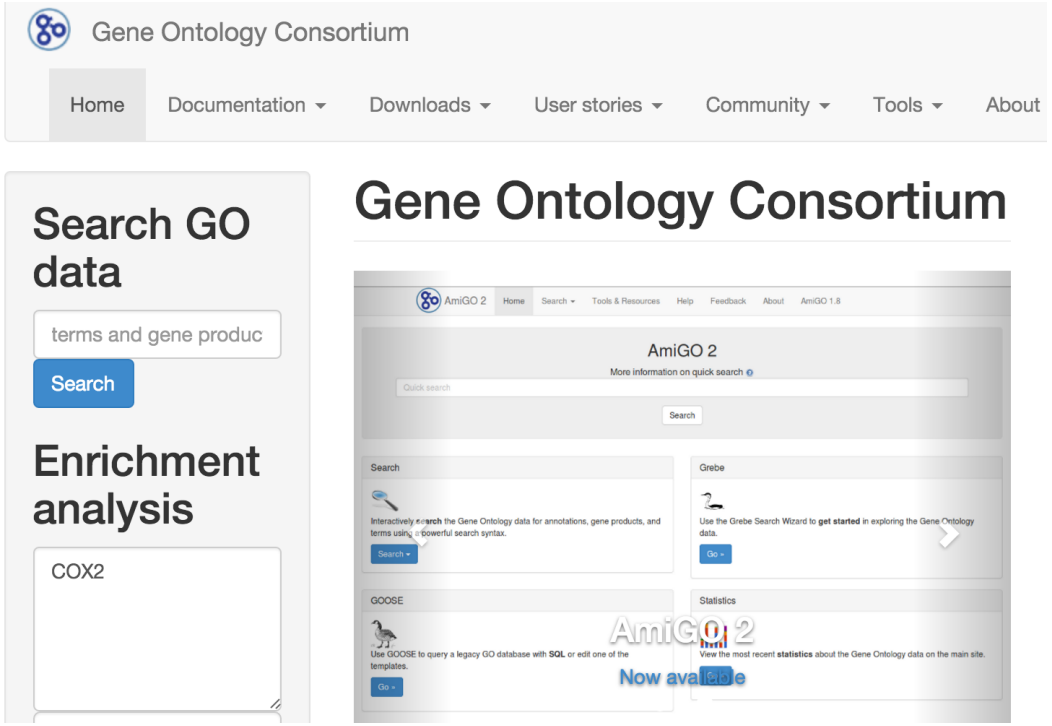
Lists available from ***Downloads***

Gene Ontology Consortium

http://geneontology.org/

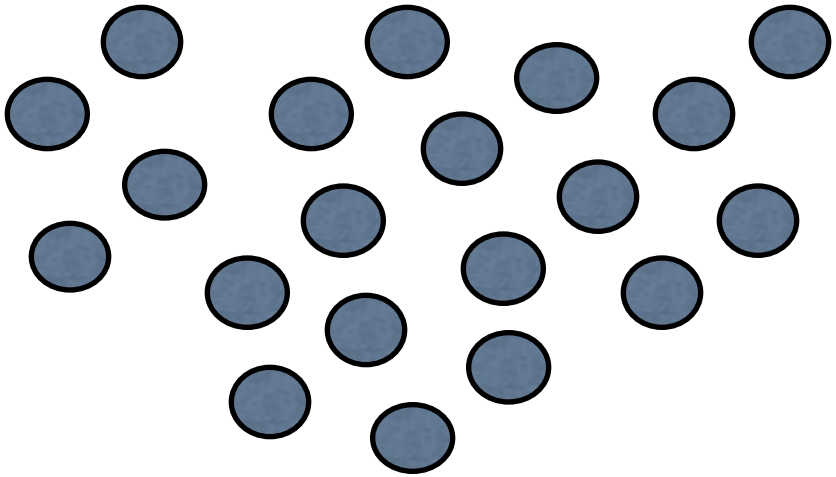
Largest and best known (at least in the USA).

Reactome and GO probably the two best.



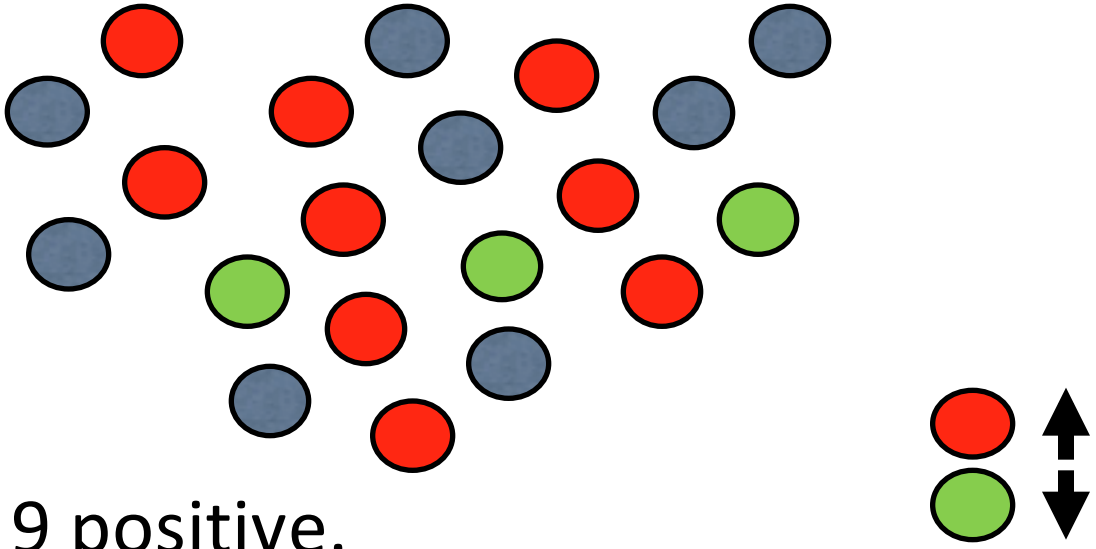
3. Pathway Analysis

Your favorite gene set



Pathway Analysis

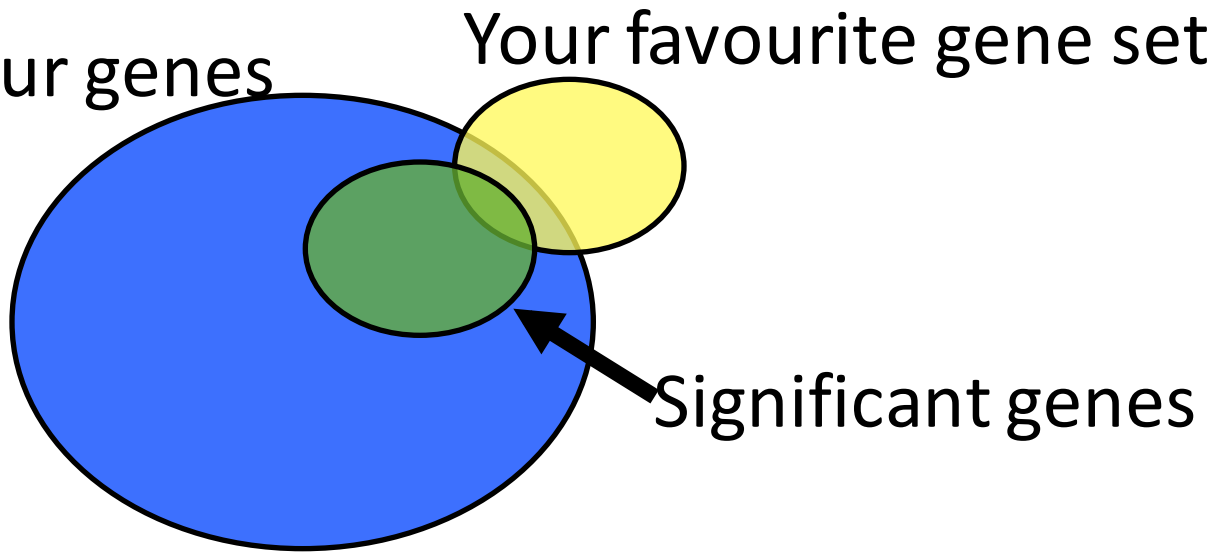
Your favorite gene set



20 genes, 9 positive.

Is that finding significant?

Pathway Analysis

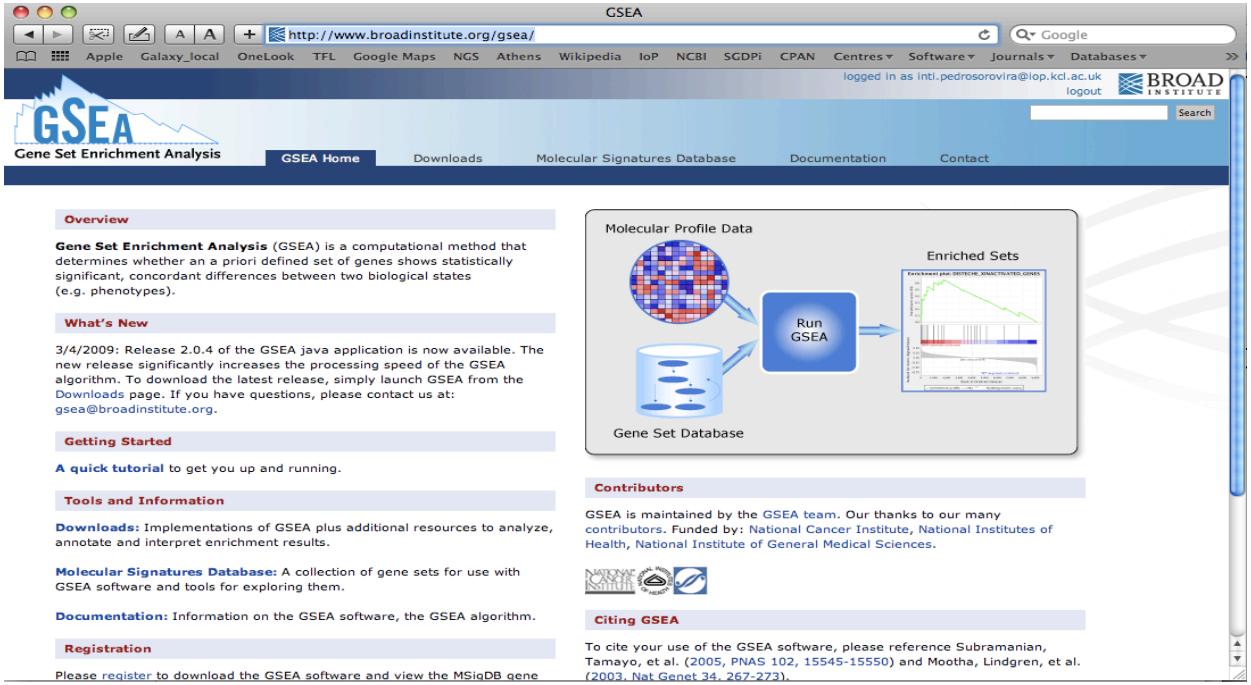


Contingency Table

11	9
1950	50

$P = 1x\ 10^{-9}$

Gene Set Enrichment Analysis



http://www.broadinstitute.org/gsea/

GWAS Pathway Analysis

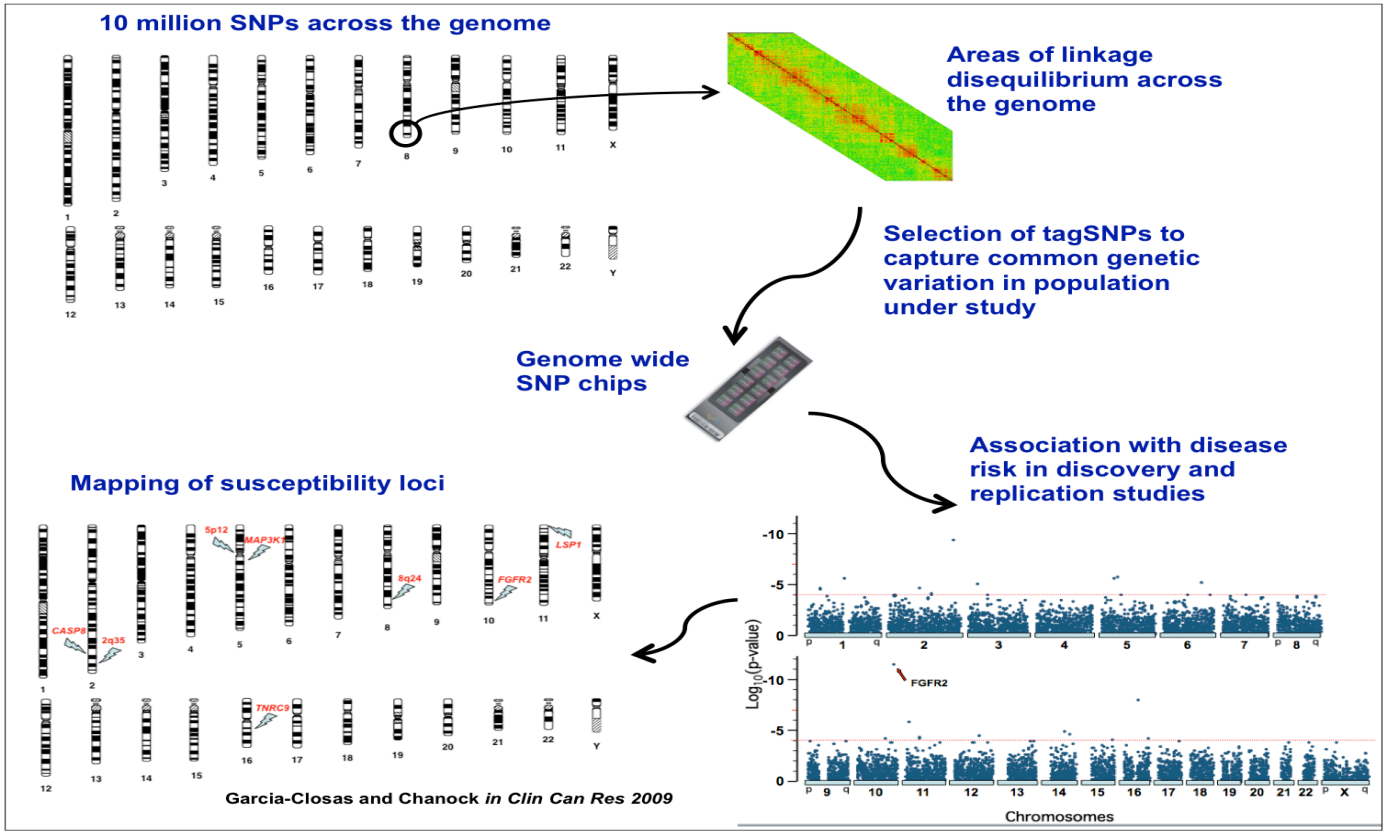
Classical GWAS studies focus on the analysis of single genetic variants:

- *Is a genetic variant associated to a trait of interest?*

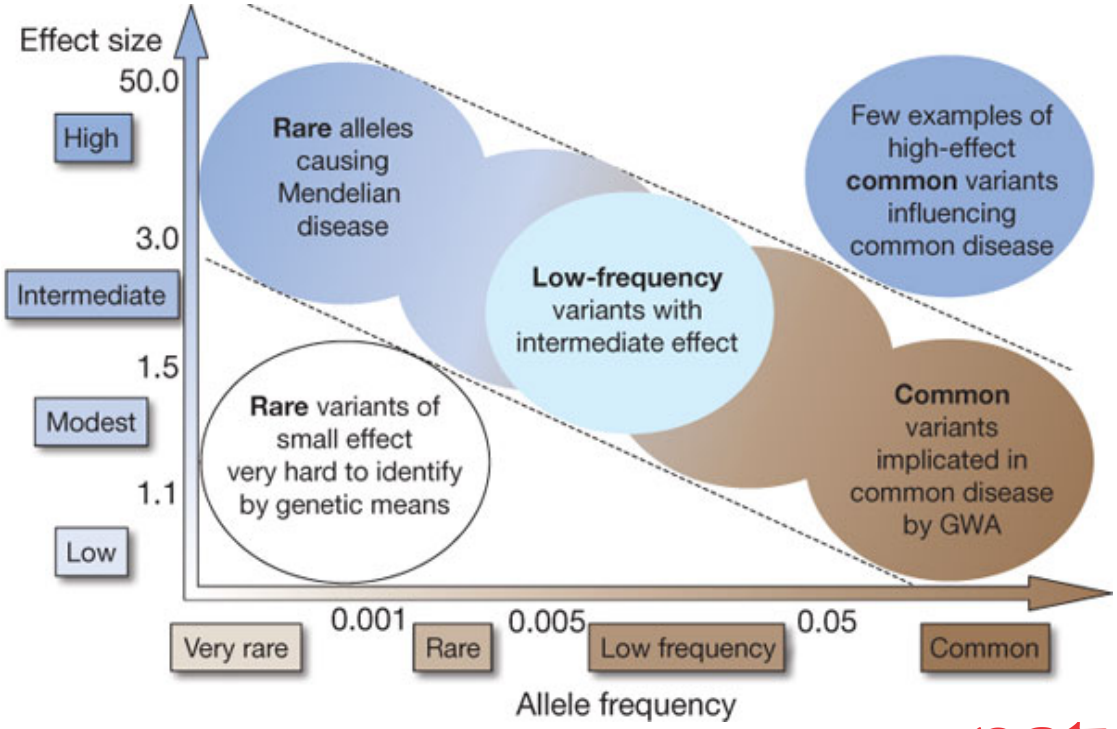
The more powerful pathway-based approaches also use information on known biological pathways

- *Are some genes in the same pathway associated to the trait of interest?*

Genome-Wide Association Studies (GWAS)



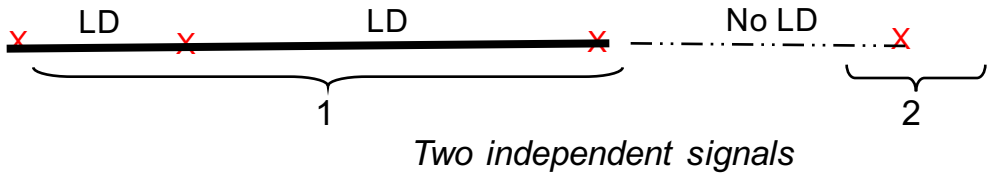
Genome-Wide Association Studies (GWAS)



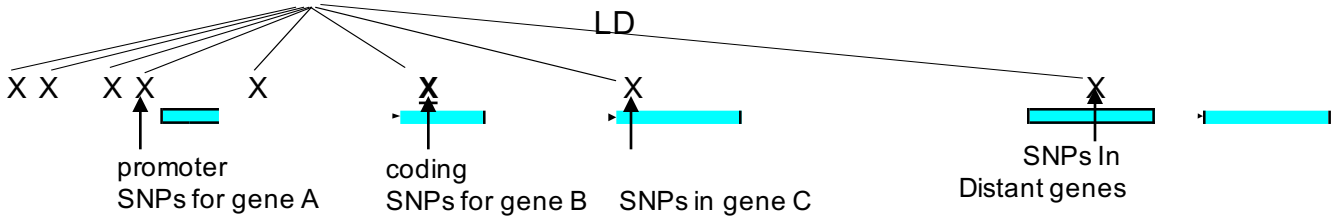
nature

The design of GWAS is based on information capture via LD tagging

Linkage desequilibrium (LD) = nonrandom association of alleles at different loci



Several genes can be mapped to one association



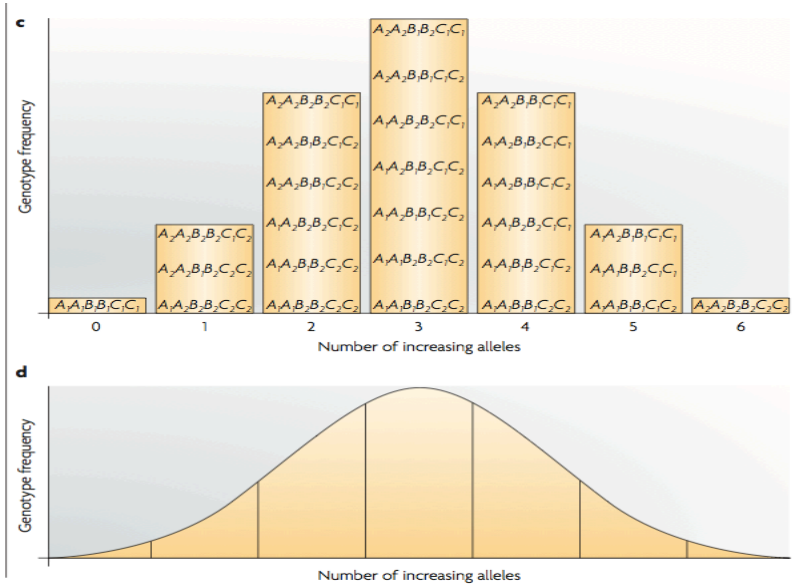
The design of GWAS is based on information capture via LD tagging

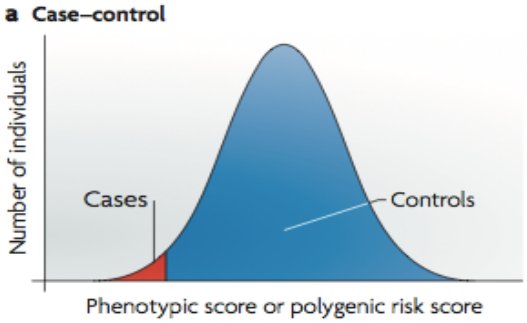
Although only Single Nucleotide Polymorphisms (SNP) are assayed, much information about common

- Variable Number Tandem Repeats (VNTR)
- Copy Number Polymorphisms (CNPs)
- Insertion/Deletions (Indel)
- Inversions

is captured by GWAS arrays.

GWAS research suggests that complex diseases are affected by many variants with small effects





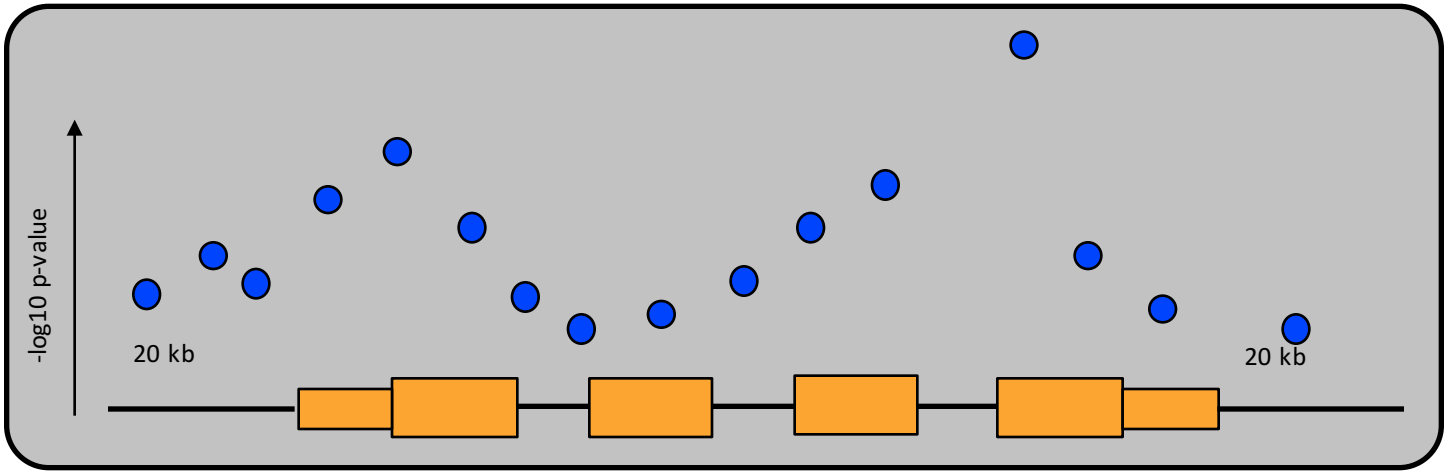
GENOME-WIDE ASSOCIATION STUDIES — OPINION

Common disorders are quantitative traits

Robert Plomin, Claire M. A. Haworth and Oliver S. P. Davis

Abstract | After drifting apart for 100 years, the two worlds of genetics — quantitative genetics and molecular genetics — are finally coming together in genome-wide association (GWA) research, which shows that the heritability of complex traits and common disorders is due to multiple genes of small effect size. We highlight a polygenic framework, supported by recent GWA research, in which qualitative disorders can be interpreted simply as being the extremes of quantitative dimensions. Research that focuses on quantitative traits — including the low and high ends of normal distributions — could have far-reaching implications for the diagnosis, treatment and prevention of the problematic extremes of these traits.

Genes in GWAS Data show evidence of multiple signals and high correlation between SNPs



GWAS pathway analysis: two main routes

Gene-wise

Derive a gene-wide statistic.

Then assess gene sets using many possible methods:

- Correct min p-value (e.g. Sidak)
- Threshold
- Combine p-values (e.g. Fisher's method)

Direct gene-set analyses

Treat a pathway as one large gene.

GWAS Pathway Analysis Software

MAGMA

INRICH

ALIGATOR

FORGE

MAGENTA

…

Using a method designed for gene expression data on genetic variation (GWAS data). Axon Guidance comes up strongly.

WTCCC Bipolar Disorder

Best p-value per gene

Pathway	Nominal P	FDR	
GO0007411	< 1e-3	0.003	Axon growth cone guidance
hsa04510	< 1e-3	0.005	Focal adhesion
hsa00040	< 1e-3	0.006	Pentose and glucuronate interconversions
GO0019198	< 1e-3	0.006	The catalysis of phosphate removal from a phosphotyrosine using aspartic acid as a nucleophile in a metal-dependent manner
GO0003779	< 1e-3	0.028	Membrane associated actin binding
hsa04512	< 1e-3	0.031	ECM-receptor interaction

Sklar et al.

Pathway	Nominal P	FDR	
GO0007411	< 1e-3	0.025	Axon growth cone guidance

Axon Guidance significant across all human disorders and diseases!

	T1D	T2D	RA	HT	CAD	CD	BP
GO0007411	0.004	0	0.004	0.001	0	0.002	0
hsa04510	0.081	0.037	0.143	0.001	0.021	0.003	0.095

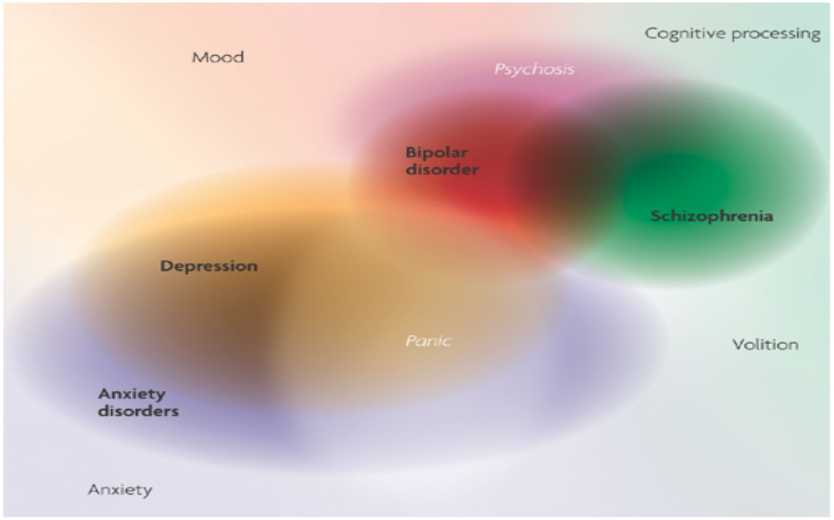
- The results were being driven by pathways having very large average gene sizes.
- Lesson: ***Always use software written for GWAS pathway analysis.***
- Reliable methods correct for LD, gene size, and other issues.

4. PGC CDG GWAS Pathway Analysis

5 methods

5 diseases

- Schizophrenia
- Manic Depression
- Major Depression
- Autism
- ADHD



Nature Reviews | Genetics

Psychiatric Disorders

Table 1 | **Defining features of nine psychiatric disorders***

Name	Life prevalence	Heritability	Essential characteristics	Notable feature
Alzheimer's disease	0.132	0.58	Dementia, defining neuropathology	Of the top ten causes of death in the United States, Alzheimer's disease alone has increasing mortality
Attention-deficit hyperactivity disorder (ADHD)	0.053	0.75	Persistent inattention, hyperactivity, impulsivity	Costs estimated at ~$US100 × 10^9 per year
Alcohol dependence (ALC)	0.178	0.57	Persistent ethanol use despite tolerance, withdrawal, dysfunction	Most expensive psychiatric disorder (total costs exceed US$225 × 10^9 per year)
Anorexia nervosa	0.006	0.56	Dangerously low weight from self-starvation	Notably high standardized mortality ratio
Autism spectrum disorder (ASD)	0.001	0.80	Markedly abnormal social interaction and communication beginning before age 3	Huge range of function, from people requiring complete daily care to exceptional occupational achievement
Bipolar disorder (BIP)	0.007	0.75	Manic-depressive illness, episodes of mania, usually with major depressive disorder	As a group, nearly as disabling as schizophrenia
Major depressive disorder (MDD)	0.130	0.37	Unipolar depression, marked and persistent dysphoria with physical and cognitive symptoms	Ranks number one in the burden of disease in the world
Nicotine dependence (NIC)	0.240	0.67	Persistent nicotine use with physical dependence (usually cigarettes)	Major preventable risk factor for many diseases
Schizophrenia (SCZ)	0.004	0.81	Long-standing delusions and hallucinations	Life expectancy decreased by 12–15 years

Samples

- The combined GWAS dataset of the five disorders comprised 60K case and controls.
- Major Depression (9,227 / 7,383)
- Manic Depression (Bipolar Disorder) (6,990/ 4,820)
- Schizophrenia (9,370 / 7,736)
- Autism (4,949 / 5,314) Trios
- Attention Deficit Hyperactivity Disorder (ADHD) (2,787/ 2,635) Trios

Pathway Analysis Methods

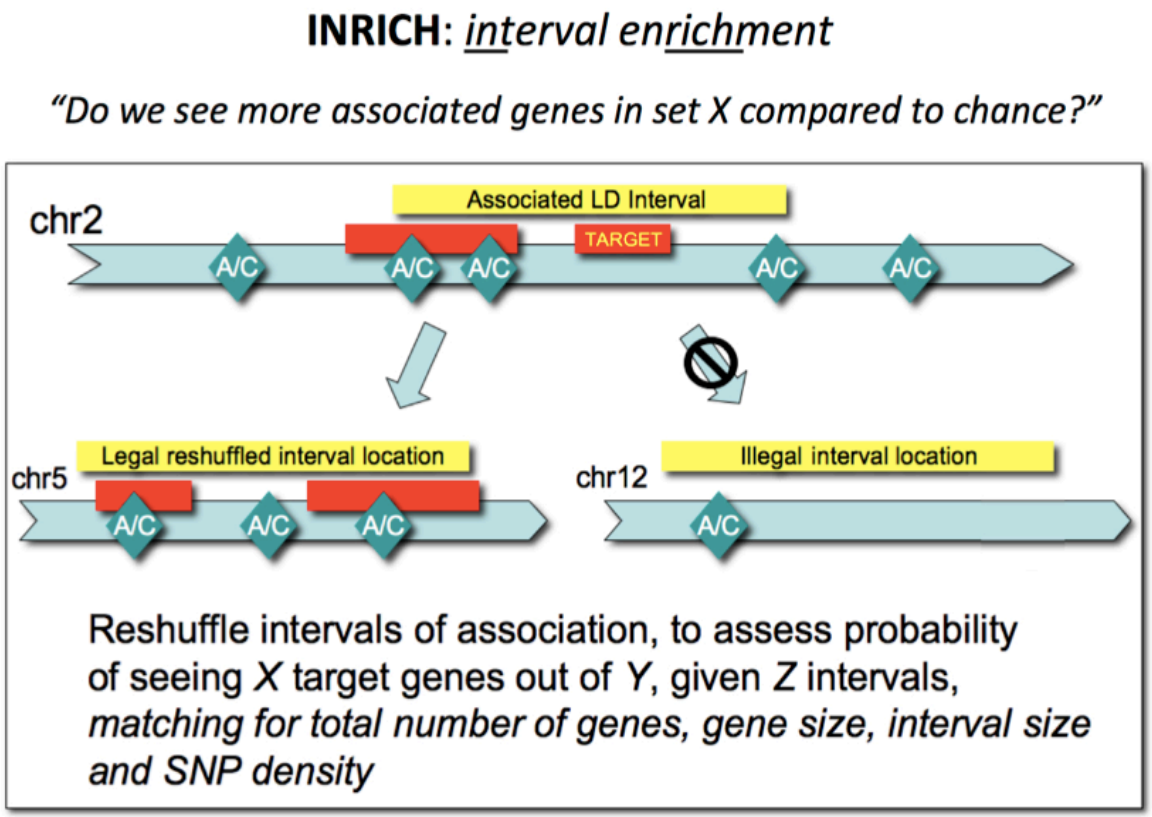
Thresholded best/number in gene/region methods

- ALIGATOR
- INRICH
- MAGENTA

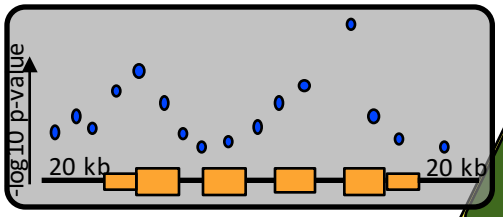
Gene-wide/Pathway-wide methods

- Set-screen test
- FORGE

INRICH


INRICH: interval enrichment
"Do we see more associated genes in set X compared to chance?"
chr2
Associated LD Interval
TARGET
A/C
A/C
A/C
A/C
A/C
chr5
Legal reshuffled interval location
A/C
A/C
A/C
chr12
Illegal interval location
A/C
Reshuffle intervals of association, to assess probability of seeing X target genes out of Y, given Z intervals, matching for total number of genes, gene size, interval size and SNP density

FORGE: gene p-values by LD corrected meta-analysis and GSA
Pedroso….Breen et al., Biol Psych 2012



$$M_F = -2\sum_{i=1}^{m} \ln(p_i)$$

$$Z = \frac{\sum_{i=1}^{k} w_i z_i}{\sqrt{\sum_{i=1}^{k} w_i^2}}$$

Meta-analysis

Gene-set Analyses

$$Z = \sqrt{m} \cdot \frac{(S_m - \mu)}{\sigma}$$

Stephen Newhouse

Databases – lots of them and lots of work

Colm O`Dushlaine

Lizzy Rossin

Database	# Genes covered	# Pathways	Median Pathway Size (min-max)
KEGG	5952	232	52 (1-1131)
GO	8589	7112	2 (1-2407)
Reactome	4539 5077	3526 (Reaction) 1086 (Pathway)	3 (1-434) 14 (1-934)
PANTHER	2170	140	16 (1-287)
OMIM	6983	4712	2 (1-22)
TargetScan	11095	162	173 (1-1240)

Comparing methods and disorders

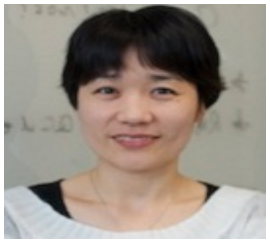
Each method has its advantages…

Comparing Disorders:

- BIPOLAR, MDD, SCZ
- Autism and ADD (PGC1 versions) are less well powered, making comparisons difficult

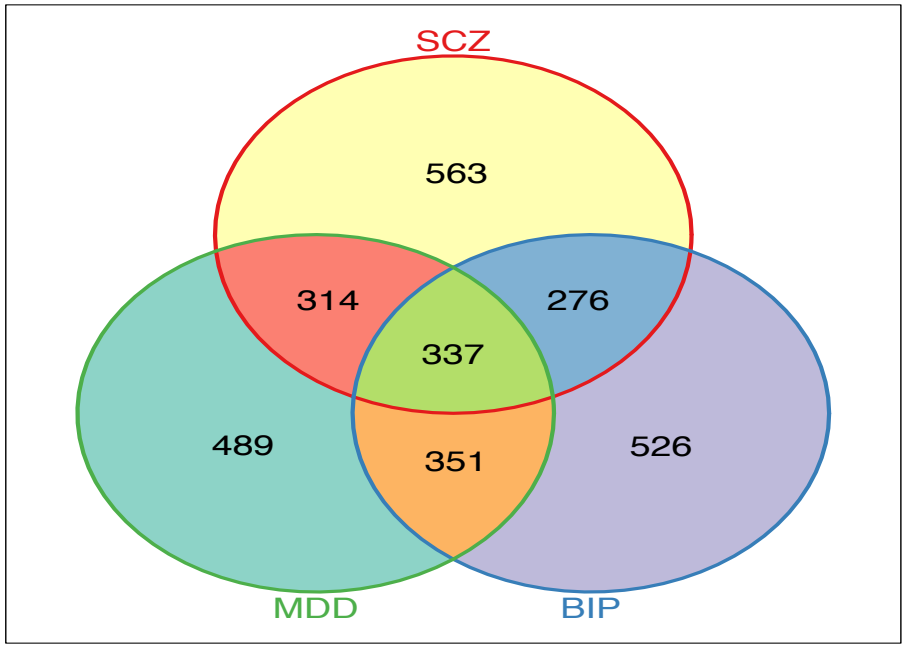
PGC Disorders

Phil Lee

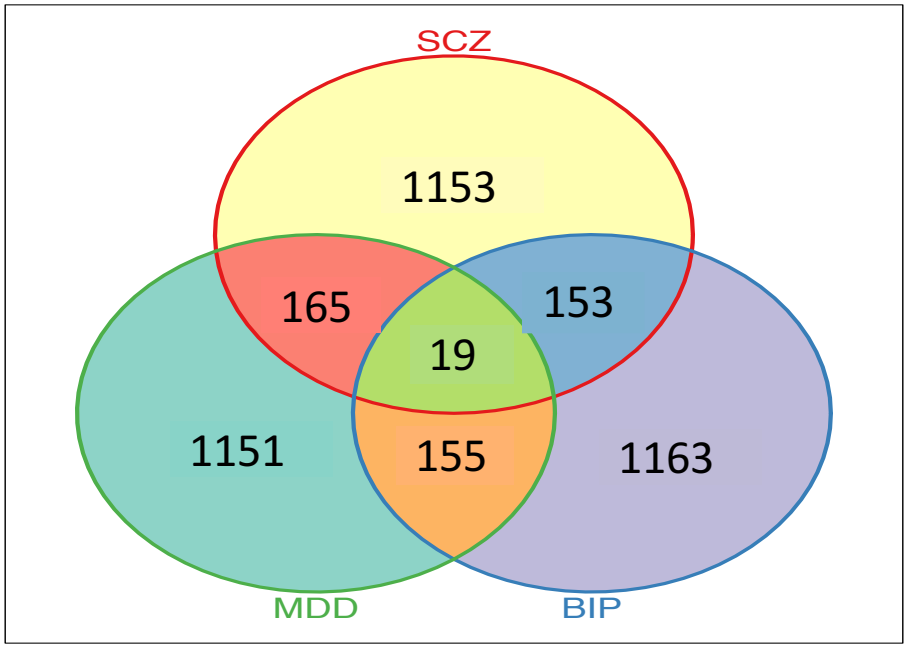


SCZ, BIPOLAR and MDD are well-powered

Overlap in the top 10% of pathways? (INRICH)

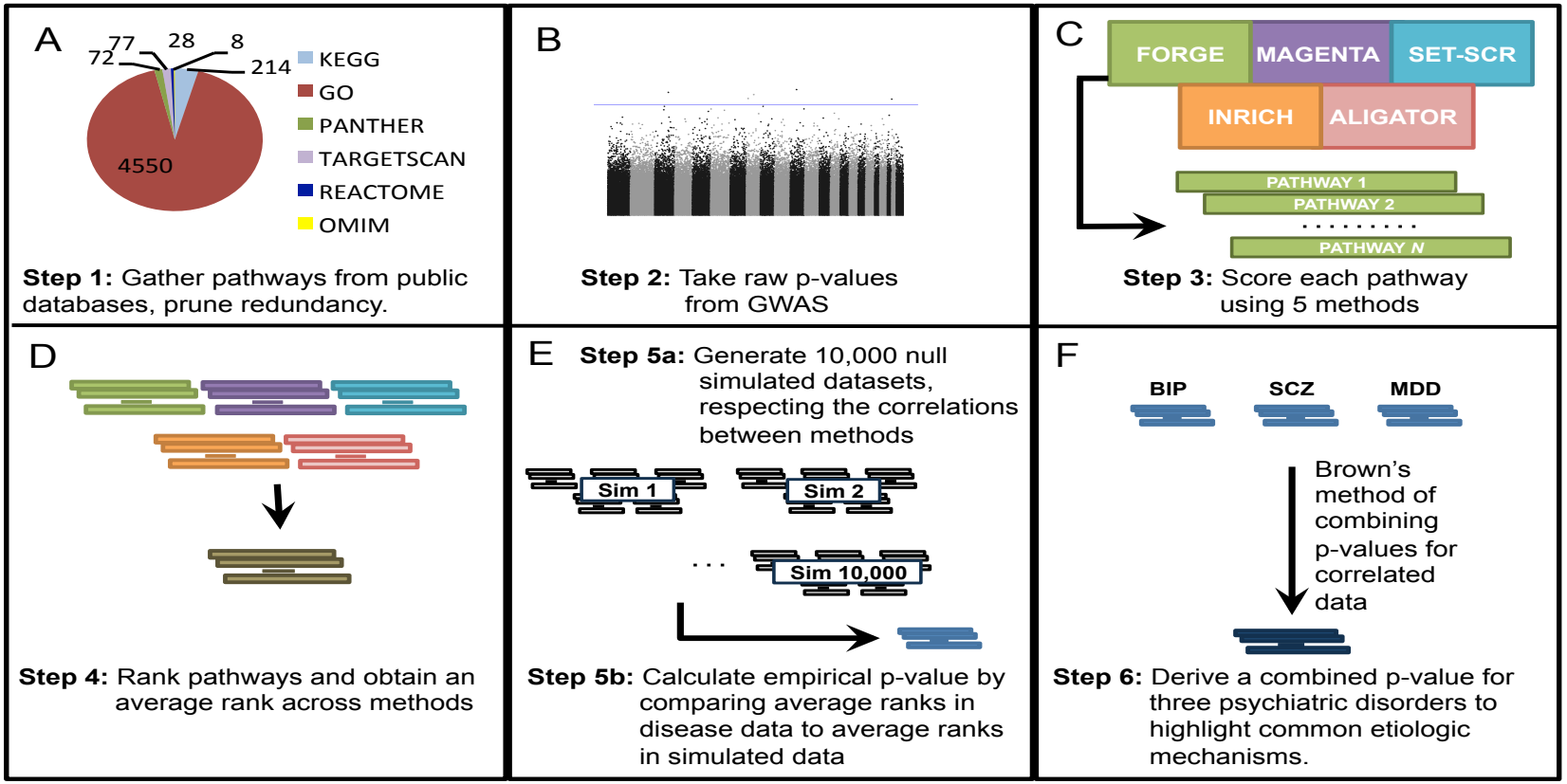


OBSERVED



EXPECTED

Derive a combined p-value for three psychiatric disorders



Top Pathways

- Results from a combined analysis of 5 methods on the 5 PGC1 CDG datasets but we focused on the three adult disorders.
- Thus, the results reflect the average across the 5 methods used.
- Significant q-values (= FDR adjusted p-values) defined as < 0.05 and suggestive < 0.5.
- MHC region excluded (high-LD genes also in that region on chromosome 6).

Pathway results SCZ, BIP, MDD

# methods	Av. rank	p rank	q-value	Pathway ID	Description
BIP					
5	17	1.01E-06	0.005	GO:51568	histone H3-K4 methylation
5	50.4	3.82E-05	0.093	path:hsa05218	Melanoma
5	79.2	1.16E-04	0.093	GO:7129	(chromosomal) synapsis
5	81.8	1.27E-04	0.093	path:hsa05213	Endometrial cancer
5	83.3	1.34E-04	0.093	P00003	Alzheimer_disease-amyloid_secretase_pathway
5	83.4	1.35E-04	0.093	path:hsa05215	Prostate cancer
5	87	1.50E-04	0.093	path:hsa05216	Thyroid cancer
4	89.5	1.59E-04	0.093	GO:90066	regulation of anatomical structure size
5	95.6	1.81E-04	0.093	path:hsa05214	Glioma
5	96.9	1.87E-04	0.093	GO:70192	chromosome organization involved in meiosis
SCZ					
5	38.4	1.58E-05	0.078	GO:14069	postsynaptic density
5	68.6	7.15E-05	0.160	GO:45211	postsynaptic membrane
5	76.8	9.67E-05	0.160	GO:43197	dendritic spine
5	85.4	1.36E-04	0.168	GO:51568	histone H3-K4 methylation
5	95.8	1.74E-04	0.173	GO:33267	axon part
MDD					
5	25.4	2.63E-06	0.012	GO:8601	protein phosphatase type 2A regulator activity
5	54.6	3.88E-05	0.092	GO:34330	cell junction organization
5	68.8	7.70E-05	0.094	GO:43297	apical junction assembly
5	70	7.92E-05	0.094	GO:45216	cell-cell junction organization
5	99.8	1.97E-04	0.186	GO:31056	regulation of histone modification

What are key differences in meta-analysis of pathway results?

Each disorder gave promising but not statistically compelling evidence for pathway association.

Analyse each disorder's pathways and then combine and meta-analyse.

May be much more powerful than SNP meta analysis.

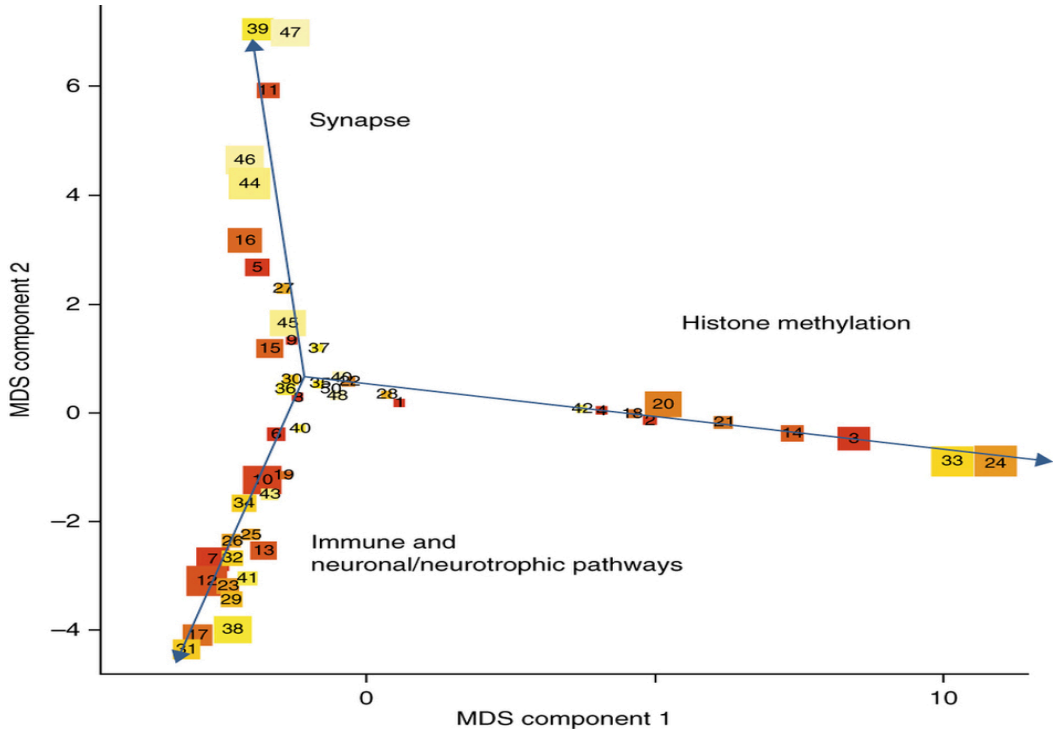
Robust to

- Allelic heterogeneity within GENES and within PATHWAYS across diseases.
- Allows for a multitude of weaker effects.
- Modulation of the pathways can differ across disorders.

Meta-analysis of the pathway results SCZ, BIP, MDD 16 pathways with FDR < 0.05, 49 with FDR < 0.1

Pathway Level Meta Analysis across three disorders.						
BIP	**MDD**	**SCZ**	**Combined P**	**Q Value**	**Pathway ID**	**Description**
0	0.0592	0.0001	5.75E-08	0.0003	GO:51568	histone H3-K4 methylation
0.0004	0.05	0.0006	1.46E-05	0.0362	GO:16571	histone methylation
0.0004	0.1462	0.0011	4.73E-05	0.0414	GO:43414	macromolecule methylation
0.0008	0.063	0.0014	5.10E-05	0.0414	GO:34968	histone lysine methylation
0.42	0.0001	0.0023	5.58E-05	0.0414	GO:45216	cell-cell junction organization
0.0001	0.091	0.0064	5.69E-05	0.0414	P00003	Alzheimer_disease-amyloid_secretase_pathway
0.0007	0.0495	0.0024	5.86E-05	0.0414	P04393	Ras_Pathway
0.312	0	0.1286	7.12E-05	0.0422	GO:8601	protein phosphatase type 2A regulator activity
0.898	0.0001	0.0017	7.83E-05	0.0422	GO:43297	apical junction assembly
0.0013	0.0207	0.0055	9.25E-05	0.0422	P00052	TGF-beta_signaling_pathway
0.489	0.0203	0	9.53E-05	0.0422	GO:14069	postsynaptic density
0.0085	0.0009	0.0239	0.0001	0.0422	GO:32869	cellular response to insulin stimulus
0.0188	0.0054	0.0022	0.0001	0.045	P00010	B_cell_activation
0.0023	0.2988	0.0003	0.0001	0.045	GO:8757	S-adenosylmethionine-dependent methyltransferase activity
0.0073	0.008	0.0044	0.0001	0.0454	GO:23061	signal release
0.459	0	0.0168	0.0002	0.0473	GO:34330	cell junction organization

Multidimensional scaling plot of top 50 pathways with suggestive significance (FDR < 0.5)



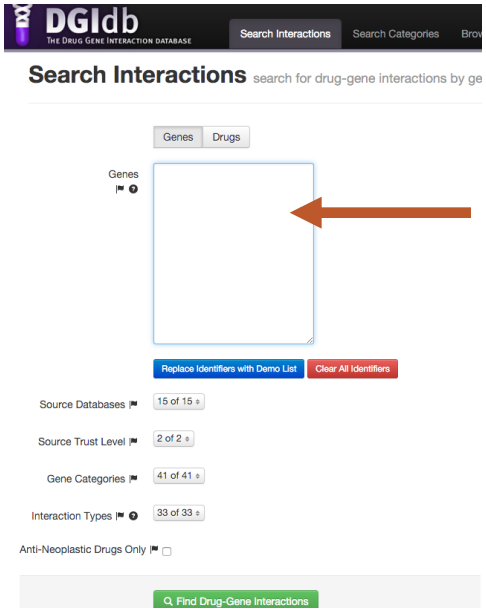
5- Drug/GWAS Pathway Analysis

Using GWAS summary statistics, drug/gene information and pathway analysis tools to:

- Find new purposes for known drugs
- Find new potential leads for specific disorders

Mine Drug/Gene Interactions: DGIdb

http://dgidb.genome.wustl.edu/



Paste gene or drug list here

The whole database can be downloaded.

Mine Drug/Gene Interactions: DGIdb

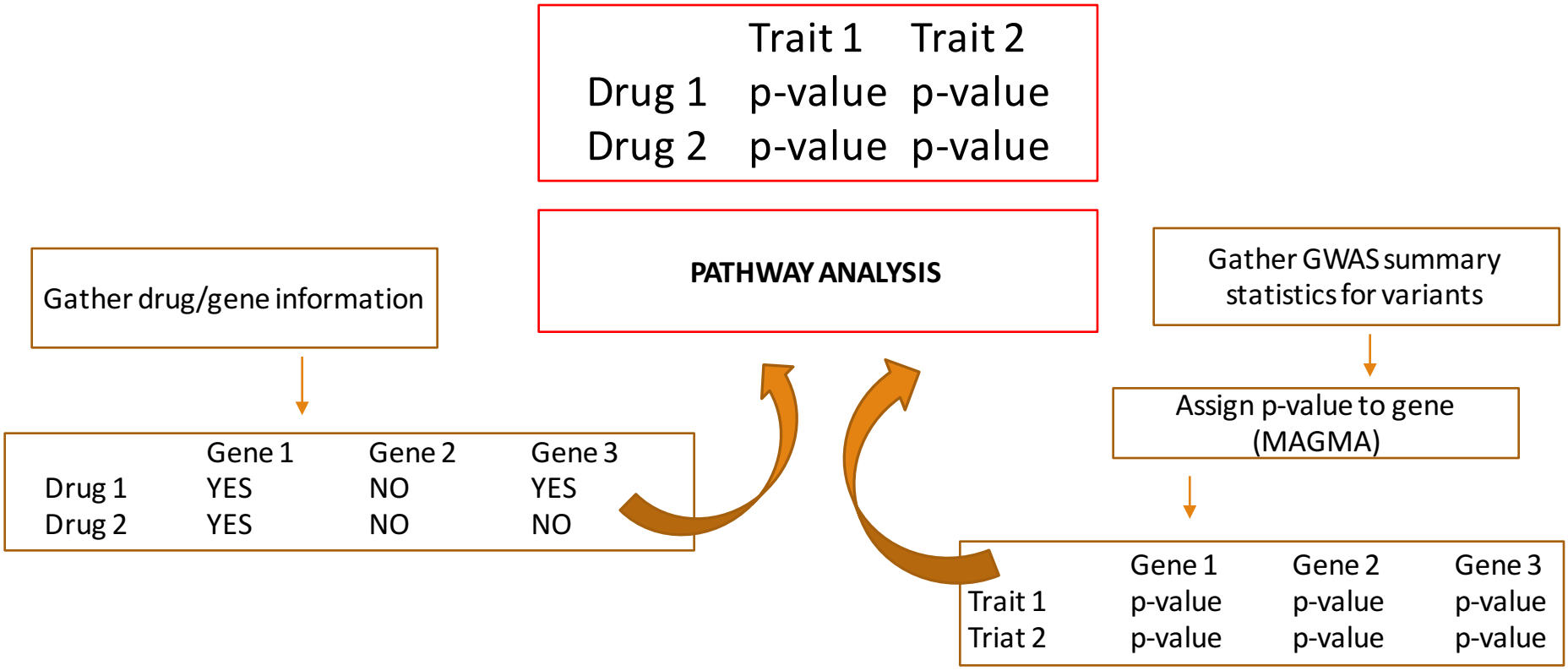
Drug-gene interactions mined from 15 databases:

- DrugBank
- therapeutic target database (TTD)
- PharmGKB
- Targeted agents in lung cancer (TALC)
- TdgClinicalTrial
- ChEMBL
- CancerCommons
- MyCancerGenome, MyCancerGenomeClinicalTrial
- CIViC
- ClearityFoundationBiomarkers
- ClearityFoundationClinicalTrail
- DoCM
- GuideToPharmacologyInteractions
- Trends in the Exploration of Novel Drug targets

GWAS Summary Statistics

AMD, Age-related macular degeneration
CARDIoGRAM, Coronary Artery Disease
CHIC, Childhood intelligence
DIAGRAM, Diabetes
EGG, Early growth
ENIGMA, Brain volume
GCAN, Anorexia Nervosa
GEFOS, Osteoporosis
GIANT, Anthropometric traits
MAGIC,Glycaemic traits
GLGC, Lipids
PGC, Psychiatric disorders
ReproGen, Reproductive ageing
SSGAC, Social science outcomes
TAG, Tobacco
IMSGC, Multiple sclerosis
GPC, Personality traits

Drug/GWAS Pathway Analysis



Some key references

http://www.ncbi.nlm.nih.gov/pubmed/25599223

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3378813/