

STAT 540 2016

Analysis of gene function I:

Gene set analysis

Paul Pavlidis

Outline

- Motivation: gene function and gene lists
- The Gene Ontology
- Enrichment analysis methods
- Difficulties, pitfalls and partial solutions
- Next lecture:
 - Gene networks
 - Graph clustering and ‘guilt by association’

Gene function: a central concept of reductionist biology

Determining function is often the goal:

- From the standpoint of genes:
 - Many genes still have “unknown function”
 - We can’t say that we know “all the functions” of the others.
- From the standpoint of experiments:
 - We don’t know enough to easily interpret the results of experiments in terms of “functions”
 - Challenge in interpreting “hit lists”



The mean gene: The DNA strand that makes people stingy with their cash

By FIONA MACRAE
Last updated at 4:34 AM on 4th November 2010

[Comments \(45\)](#) | [Add to My Stories](#)

If you have a friend who never buys a round or who rarely pays their fair share, try not to get too angry. Because being mean could be in their genes. Scientists have pinpointed a stretch of DNA that makes people stingy with their cash.

<http://www.dailymail.co.uk/>

SCIENTIFIC AMERICAN™

[Sign In | Register](#) | [Search](#)

[Subscribe](#) | [News & Features](#) | [Topics](#) | [Blogs](#) | [Videos & Podcasts](#) | [Education](#)

[Health > News](#)

Rare Genetic Mutation Lets Some People Function with Less Sleep

Ever wished you could get by with less sleep? Some people can—and don't seem to be any worse off for it—thanks, possibly, to one unusual mutation

Aug 13, 2009 | By Katherine Harmon

ABC News Blogs

Is There a Gene for Motherhood?

By ABC News
September 28, 2012 10:36 AM
[ABC News Blogs](#)

NATURE WORLD NEWS

[Home](#) | [News](#) | [Animals](#) | [Biology](#) | [Environment](#) | **Health & Medicine**

TRENDING TOPICS: VITAMINS | BIRTH | PREGNANCY | PROSTATE CANCER | HEALTHY EATING | RESEARCH

Researchers Find the Gene Variant that Makes People Cynical

0 Comments | [Email](#) | [Print](#) | [Test Size](#) | Oct 11, 2013 10:15 AM EDT

<http://www.natureworldnews.com>

Scientists find how 'obesity gene' makes people fat

BEN HIRSCHLER, REUTERS

FIRST POSTED: MONDAY, JULY 15, 2013 12:54 PM EDT | UPDATED: MONDAY, JULY 15, 2013 01:01 PM EDT

<http://www.torontosun.com>

Featured Research

Gene mutation for excessive alcohol drinking found

Date: November 26, 2013
Source: Newcastle University

Summary: Researchers have discovered a gene that regulates alcohol consumption and when faulty may cause excessive drinking. They have also identified the mechanism underlying this phenomenon. The study showed that normal mice show no interest in alcohol and drink little or no alcohol when offered a free choice between a bottle of water and a bottle of diluted alcohol. However, mice with a genetic mutation to the gene Gabob1 overwhelmingly preferred drinking alcohol over water, choosing to consume almost 85% of their daily fluid as drinks containing alcohol – about the strength of wine.

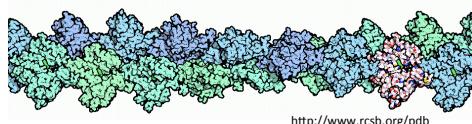
<http://www.sciencedaily.com>

A useful distinction (?): “Proximal” function

About what the gene product does more or less directly.

Other “functions” require the proximal functions

Example: Actin



Actin forms filaments which help give cells shape and structure, and helps them move or change shape. (review: <http://www.sciencemag.org/content/326/5957/1208.full>)

For a related discussion see Schrager J, Bioinformatics 2003 19:1934–1936

“Distal” function

- The gene can be “present” without the function being present, and vice versa
- Often defined due to mutant phenotypes

Actin function: makes mice anxious.

PLOS GENETICS

Cofilin-1: A Modulator of Anxiety in Mice

Martin Goodson¹, Marco B. Rust², Walter Witke³, David Bannerman⁴, Richard Mott⁵, Chris P. Ponting⁶, Jonathan Flint^{1*}

¹ Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom, ² Neurodegeneration Group, Department of Biology, University of Oxford, United Kingdom, ³ Institute of Genetics, University of Edinburgh, Edinburgh, United Kingdom, ⁴ Department of Experimental Psychology, University of Oxford, United Kingdom, ⁵ MRC Functional Genetics Unit, Department of Physiology, Anatomy, and Genetics, University of Oxford, Oxford, United Kingdom

Abstract
The genes involved in conferring susceptibility to anxiety remain obscure. We developed a new method to identify genes at quantitative trait loci (QTLs) in a population of heterogeneous stock mice descended from known progenitor strains. QTLs were partitioned into intervals that can be explained by a single phylogenetic tree among predators and prey intervals. When considering the effect of anxiety on anxiety-related phenotypes, we found that QTLs containing anxiety loci are often positioned within those intervals. We used this phylogenetic information to filter candidate genes positioned within those intervals, we identified actin depolymerizing factors (ADFs), including cofilin-1 (Cfl1). Cfl1 is a member of the ADF family of proteins, which are involved in actin dynamics. We found that Cfl1 is expressed in each QTL, indicating the importance of phylogenetic filtering. We confirmed experimentally that forebrain-specific inactivation of Cfl1 decreased anxiety in knockout mice. Our results indicate that similarity of function of mammalian genes can be used to recognize key genetic regulators of anxiety and potentially of other emotional behaviours.

(Cofilin regulates actin)

Actin function: lets use hear sounds

Two Deafness-causing (DFNA20/26)
Actin Mutations Affect Arp2/3-dependent
Actin Regulation *

Karina A. Kruth and Peter A. Rubenstein¹

<http://www.jbc.org/content/287/32/27217>

Functional specificity*: assess in a comparative way

- “Actin is involved in cell shape”
- “RNA polymerase is involved in cell shape”
 - True, because it is needed to transcribe actin.
- Actin is “more involved” in cell shape than RNA polymerase.

This is an important part of what we mean by a gene having a function – it's *specifically* involved, relative to other genes.

This gets trickier when the function is not likely to be proximal for any gene (like “hearing”).

*The word ‘specific’ is overloaded, just like ‘function’ – I’m open to suggestions for a better term.

Codifying function: Gene Ontology

- Controlled vocabulary for function
- Motivated in part by need for species comparisons
 - Genome sequencing made uniform annotations ever more urgent
- Started ~1999 (mouse, fly, worm)
 - Built in part on earlier efforts such as Enzyme Commission
- Database curators use these terms to describe genes.
 - GO ≠ “GO annotations”
- Over 40,000 terms
- Mostly species-agnostic
- Three aspects
 - Biological Process
 - Cellular Component
 - Molecular Function

death

Term Information

Accession GO:0016265
Name death
Ontology biological_process
Synonyms None
Definition A permanent cessation of all vital functions: the end of life; can be applied to a whole organism or to a part of an organism.
ISBN:0877797059, GOC:mah
Comment See also the biological process term 'apoptosis' ; GO:0006915'.
History See term history for GO:0016265 at QuickGO
Subset gosubset_prok
gosubset_plant
Community [GMA](#) Add usage comments for this term on the GONUTS wiki.
Related [Link to all genes and gene products associated to death.](#)
[Link to all direct and indirect annotations to death.](#)
[Link to all direct and indirect annotations download \(limited to first 10,000\) for death](#)

Found entities
Total: 37239; showing 1-100 Results count 100

Gene/product	Gene/product name	Qualifier	Direct annotation	Annotation extension	Source	Taxon	Evidence
CERKL	Uncharacterized protein		negative regulation of apoptotic process		UniProtKB	Canis lupus familiaris	IEA Ensembl
IRS2	Uncharacterized protein		negative regulation of B cell apoptotic process		UniProtKB	Canis lupus familiaris	IEA Ensembl
ITM2B	Uncharacterized protein		extrinsic apoptotic signaling pathway in absence of		UniProtKB	Canis lupus familiaris	IEA Ensembl

<http://amigo.geneontology.org/amigo/term/GO:0016265>

QuickGO - <http://www.ebi.ac.uk/QuickGO>

The GO structure: semantic relationships among concepts

- Structure is an acyclic graph, not a tree: terms **commonly** have multiple parents.
- Annotations in the database and most 3rd party resources do not list all the parents: you must infer these in computations.**
- Gene might have 10 "directly annotated" terms but >100 "inferred" terms.
- The inferred terms are just as valid as the directly annotated ones.**

Old Amigo (above) made multiple paths clearer in text view compared to Amigo2 (below)

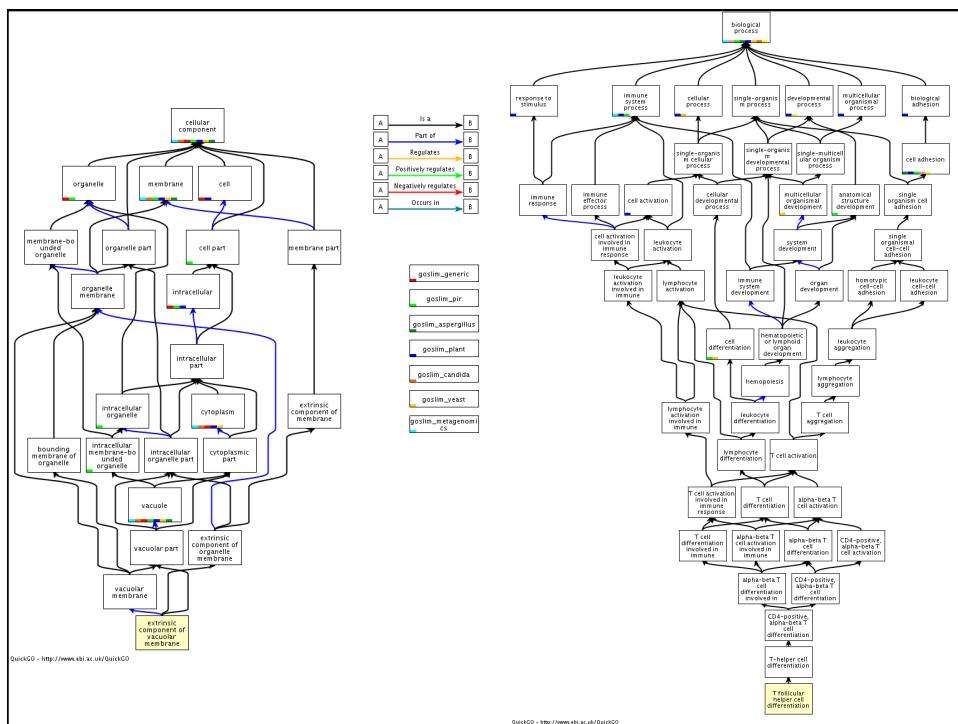
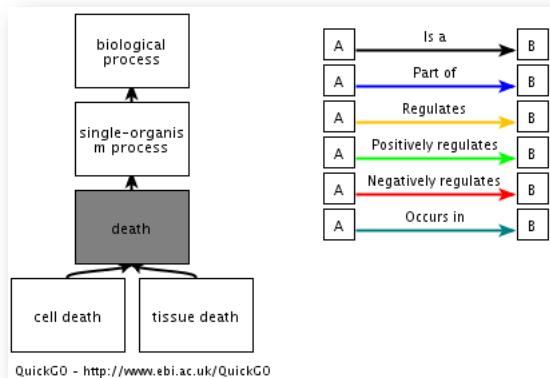
```

GO:0035069 : larval midgut histolysis [7 gene products]
  □ GO:005096 : larval midgut cell programmed cell death [3 gene products]
  □ GO:009886 : post-embryonic morphogenesis [600 gene products]
  □ □ GO:0048707 : instar larval or pupal morphogenesis [421 gene products]
    □ □ GO:0035069 : larval midgut histolysis [7 gene products]
      □ □ GO:0035096 : larval midgut cell programmed cell death [3 gene products]
  □ □ GO:009888 : tissue development [6234 gene products]
  □ □ GO:0007559 : histolysis [87 gene products]
    □ □ □ GO:0035069 : larval midgut histolysis [7 gene products]
      □ □ □ GO:0035096 : larval midgut cell programmed cell death [3 gene products]
    □ □ □ GO:0010703 : negative regulation of histolysis [1 gene product]
    □ □ □ GO:0010702 : regulation of histolysis [1 gene product]
    □ □ □ GO:0035070 : salivary gland histolysis [81 gene products]
  □ □ GO:009653 : anatomical structure morphogenesis [11903 gene products]
  □ □ GO:0007552 : metamorphosis [442 gene products]
  □ □ □ GO:0035069 : larval midgut histolysis [7 gene products]

  BFO:000003 occurrence
  BFO:0000015 process
  □ GO:0008150 biological_process
  □ GO:0005907 developmental_process
  □ GO:0005906 morphogenetic_process
  □ GO:0044899 single-organism_process
  □ GO:0044707 single-multicellular organism_process
  □ GO:0044787 single-organism developmental_process
  □ GO:0048856 anatomical structure development
  □ BFO:0000001 BFO_0000001
  □ GO:0007275 multicellular organismal development
  □ GO:0009653 anatomical structure morphogenesis
  □ GO:0016265 death
  □ GO:0009791 post-embryonic development
  □ GO:0002165 instar larval or pupal development
  □ GO:0001659 post-embryonic morphogenesis
  □ GO:0016271 tissue development
  □ GO:0009888 tissue development
  □ GO:0007559 histolysis
  □ GO:0048707 instar larval or pupal morphogenesis
  □ GO:0007552 metamorphosis
  □ GO:0035069 larval midgut histolysis
  □ GO:0035096 larval midgut cell programmed cell death
  
```

Relationship types

There are different types of parents
 Basic ones: **is-a** and **part-of**
 There are other relationship types that GO supports



Molecular Function (MF)

- “Elemental activities, such as catalysis or binding, describing the actions of a gene product at the molecular level.”
- Primarily enzymatic activities
- Often end in “activity”
- Examples:
 - “helicase activity”
 - “translation repressor activity”

Cellular Component (CC)

- The part of a cell of which a gene product is a component
- “For purpose of GO includes the extracellular environment of cells”
- “This term includes gene products that are parts of macromolecular complexes, by the definition that all members of a complex normally copurify under all except extreme conditions.”
- Examples:
 - “apical complex”
 - “cell wall”
 - “condensed chromosome”

Biological Process (BP)

- “A phenomenon marked by changes that lead to a particular result, mediated by one or more gene products.”
- Closest GO concept to “pathways”*
- Examples:
 - “exocytosis”, “vesicle docking during exocytosis”
 - “death”, “programmed cell death”, “apoptosis”

*GO relations exist for “regulates” etc., but these are currently less used

Redundancy of GO aspects

BP: “translation”

MF: “structural constituent of ribosome”

CP: “ribosome”

- A gene annotated to one will often be annotated to the other.
- Partial solution: only work with one of these three
- Overlaps are also extensive *within* aspects

GO Annotations and evidence codes

Evidence code	Evidence code description	Source of evidence	Manually checked	Current number of annotations*
IDA	Inferred from direct assay	Experimental	Yes	71,050
IEP	Inferred from expression pattern	Experimental	Yes	4,598
IGI	Inferred from genetic interaction	Experimental	Yes	8,311
IMP	Inferred from mutant phenotype	Experimental	Yes	61,549
IPI	Inferred from physical interaction	Experimental	Yes	17,043
ISS	Inferred from sequence or structural similarity	Computational	Yes	196,643
RCA	Inferred from reviewed computational analysis	Computational	Yes	103,792
IGC	Inferred from genomic context	Computational	Yes	4
IEA	Inferred from electronic annotation	Computational	No	15,687,382 ↗
IC	Inferred by curator	Indirectly derived from experimental or computational evidence made by a curator	Yes	5,167
TAS	Traceable author statement	Indirectly derived from experimental or computational evidence made by the author of the published article	Yes	44,564
NAS	Non-traceable author statement	No 'source of evidence' statement given	Yes	25,656
ND	No biological data available	No information available	Yes	132,192
NR	Not recorded	Unknown	Yes	1,185

*October 2007 release

Most annotations are "IEA" – derived by computational analysis, not an experiment, and most not reviewed by a curator

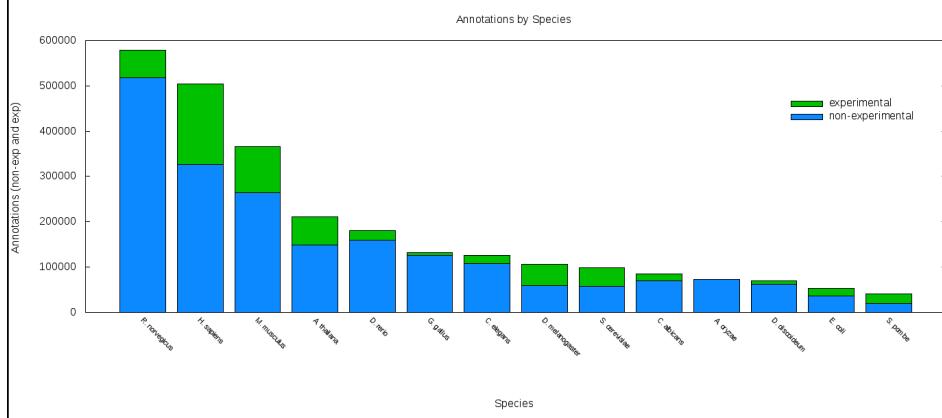
However, IEA annotations are roughly as good as the manual annotations.

Numbers here are old; direct annotations only (not inferred)
Over 98% of annotations are computational;

Also more evidence codes have been added.

Nat Rev Genet. 2008 Jul;9(7):509-15

GO annotations (updated view)



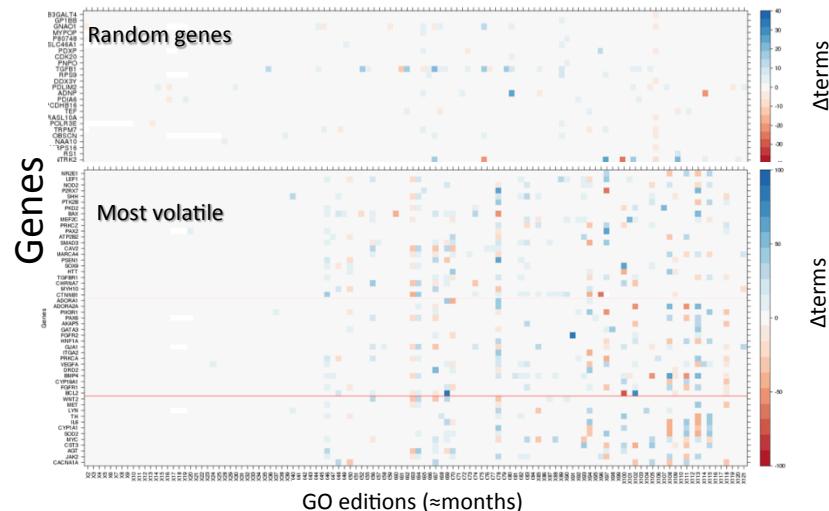
This is somewhat misleading because of varying genome size; annotations per gene would be a bit flatter (also omits inferred annotations)

<http://geneontology.org/page/current-go-statistics>

Complaints about GO annotations

- Fails at an intended purpose: species comparisons
 - Now accepted to be difficult, at best
 - <http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1002386>
- Incomplete
 - Information about experiments from 1980s is still being entered
- Inconsistent/confusing
 - Insufficient distinction between “proximal” and “distal” function; this is left up to curators; annotation guidelines vague
- Unstable
 - The function of a gene can change over time even without new data
 - <http://bioinformatics.oxfordjournals.org/content/29/4/476.abstract>

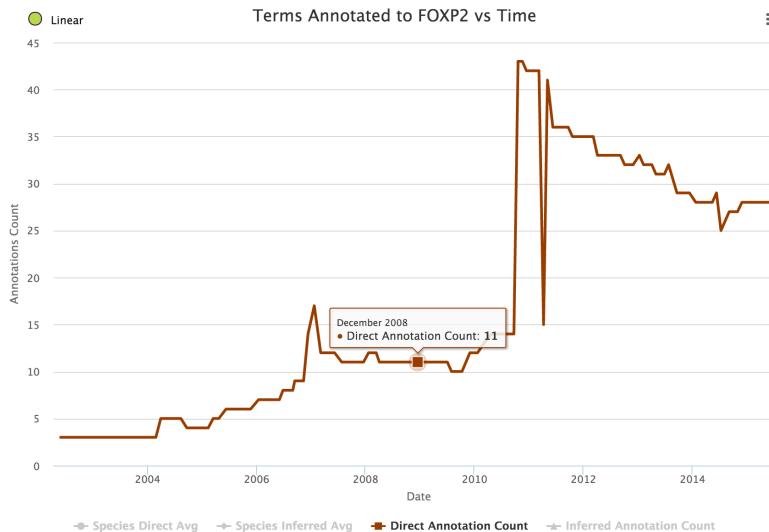
GO annotations are volatile



Genes gain or lose terms over time as GO changes and annotation practices are revised

Adri Sedeño

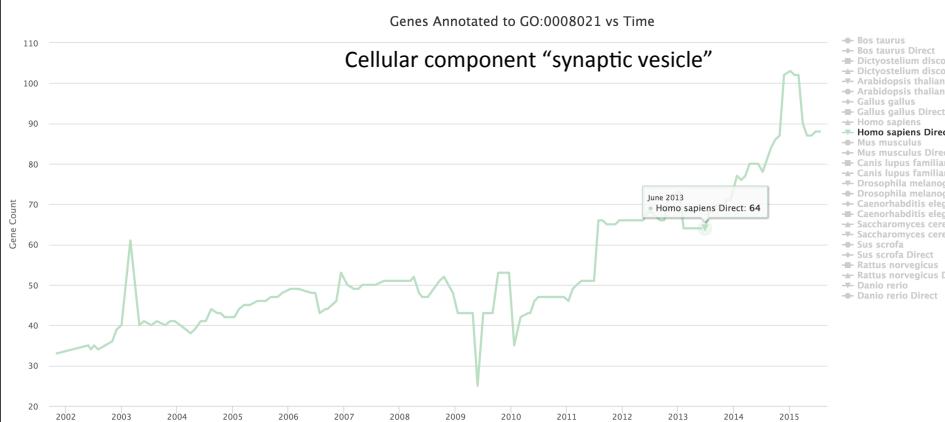
GOTrackWeb: Track gene annotations



Adri Sedeño and Matt Jacobson

<http://www.chibi.ubc.ca/gotrack/>

Tracking use of a GO term



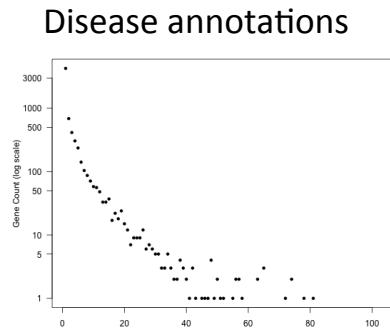
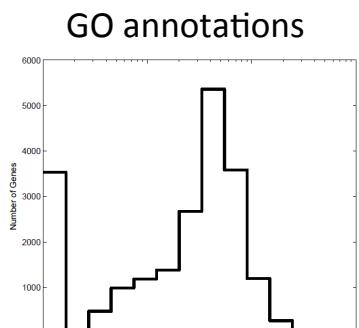
Multifunctionality

- “Some genes have more functions than others”
- Operational definition: Number of Gene Ontology terms (or other annotations)
 - This definition is appropriate inasmuch it is what we actually use in our analyses.

23

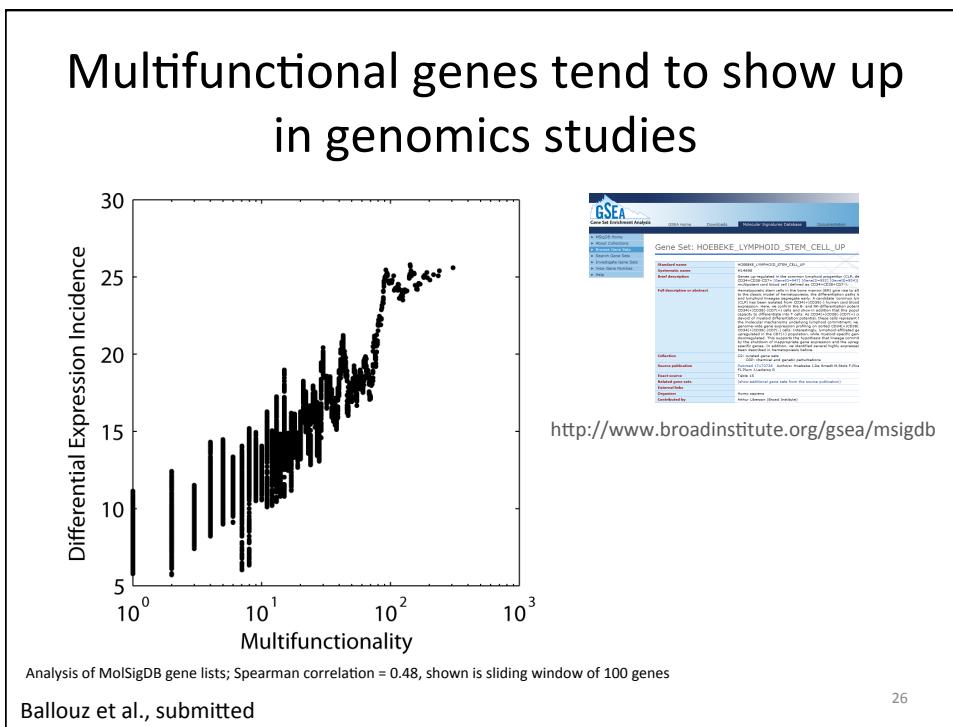
Distribution of annotations per gene

- Extremely uneven – many genes have no terms, others have hundreds.
- Unclear whether this is “popularity” or biological reality



Some of the heavy hitters									
ABCA1	ABCG1	ADA	ADAM17	ADH5	ADIPOQ	ADORA2B	ADRA1B		
ADRB2	AGT	AGTR1	AGTR2	AK5	AKT1	ALB	APC	APOA1	APOE
ASCL1	ATP2A1	ATP7A	AZU1	B4GALT1	BAX	BBS4	BCL2	BCL3	BCL6
BLM	BMF1	BRCA1	CC	CALCA	CANX	CANX	CD20	CD24	CDK5
CDKN2A									DPYD
DRD1									ENTPD4
FASN									GK
GHRL									GH1
HPX	H								HPRT1
IL6ST	INS	INSR	IRAK3	JMJD6	KCNMA1	LBP	LIG4	MSH2	MSH6
MYH10	MYH9	NEDD4	NF1	NFKBIA	NGFR	NKX2-5	NOD2	NOS3	NP
NUDT9	P2RX4	PAPSS1	PLEK	PPARG	PROX1	PTEN	PVRL2	RBP4	
RYR2	SHH	SIRT1	SLC11A2	SLC1A3	SLC1A4	SLC22A5	SMAD6	SMAD7	
SOD1	SOD2	SORD	SPN	SRR	STAT5A	STAT5B	SYK	TERF2	TGFB1
TGFB2	TGFB3	TGFBR2	TGFBR3	THBS1	TNF	TP53	TYMS		

25



26

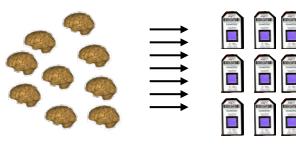
Summary so far

- Defining gene function is hard, but it will help to think about what you mean
 - “Distal” function is a slippery slope
 - Functional “specificity” can help clarify
- Systematizations such as GO are only an approximation (or crutch)
- Multifunctionality will have influence on interpretation

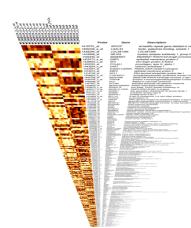
Can we use function to discover function?

- In many of the studies we talk about in class, the goal is to figure out function of genes, but without a specific hypothesis: a “screen”
- Interpreting the results of the screen requires prior knowledge.

Case type A



→ → →



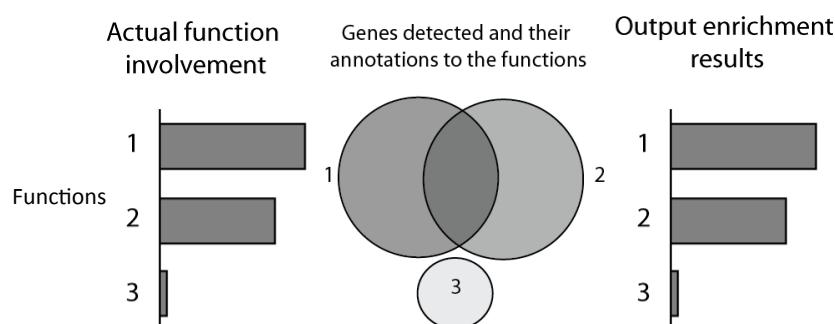
→

????

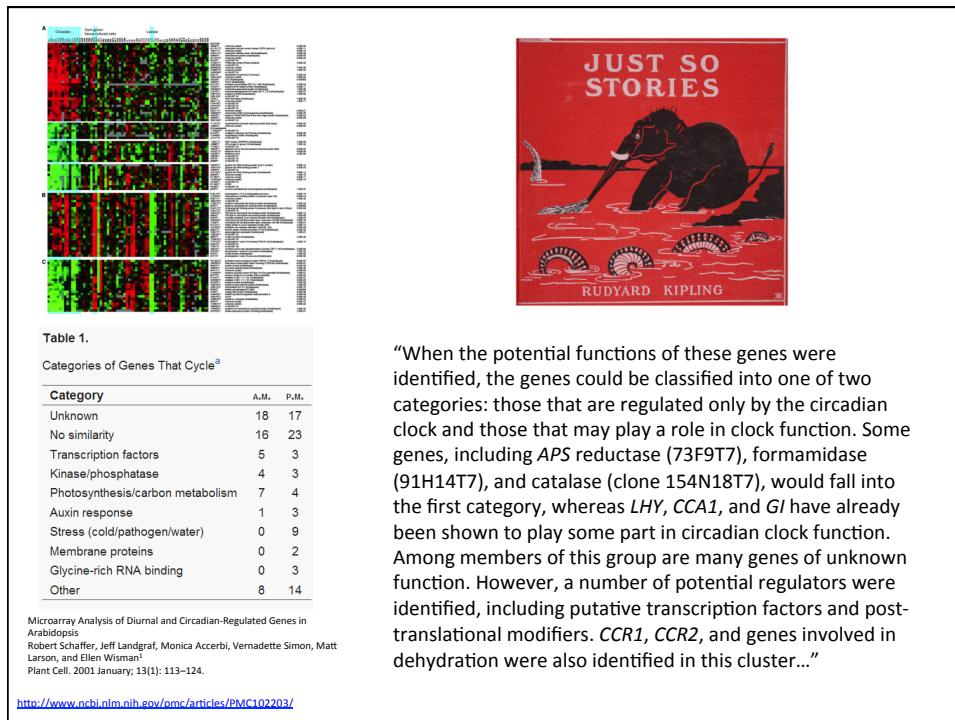
Goals of Gene set analysis

- “Is my hit list involved in any **functions**”
(typical imprecise statement of the problem)
- Instead of analyzing genes one at a time, analyze them in **biologically-motivated groupings**
 - Not necessarily functions, but we’ll focus on that case.
- Ideally have some statistical framework to guide our decision making.

What we hope gene sets can tell us



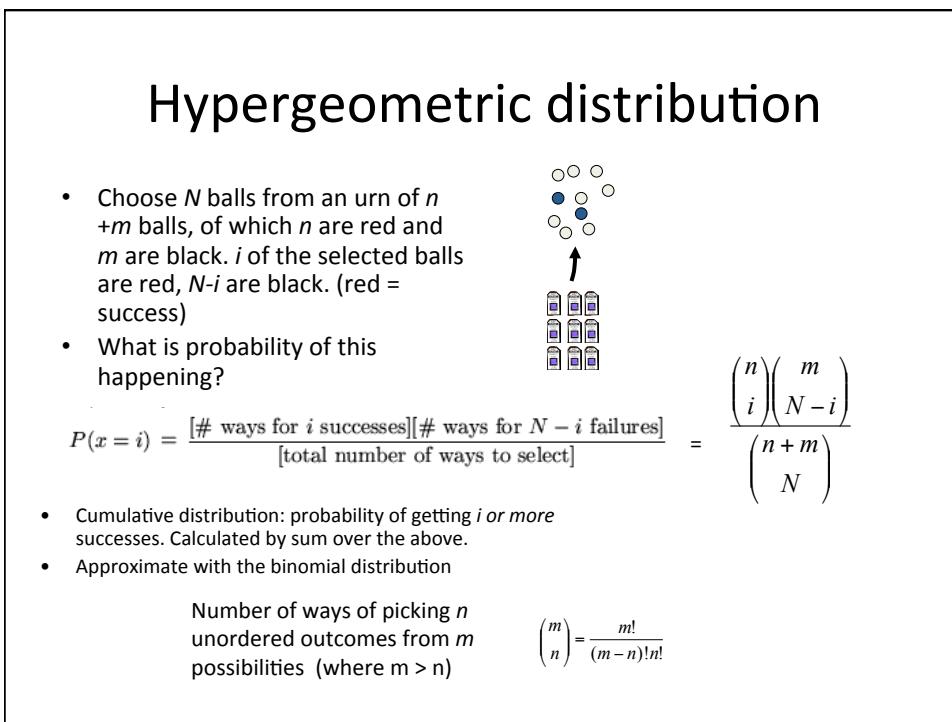
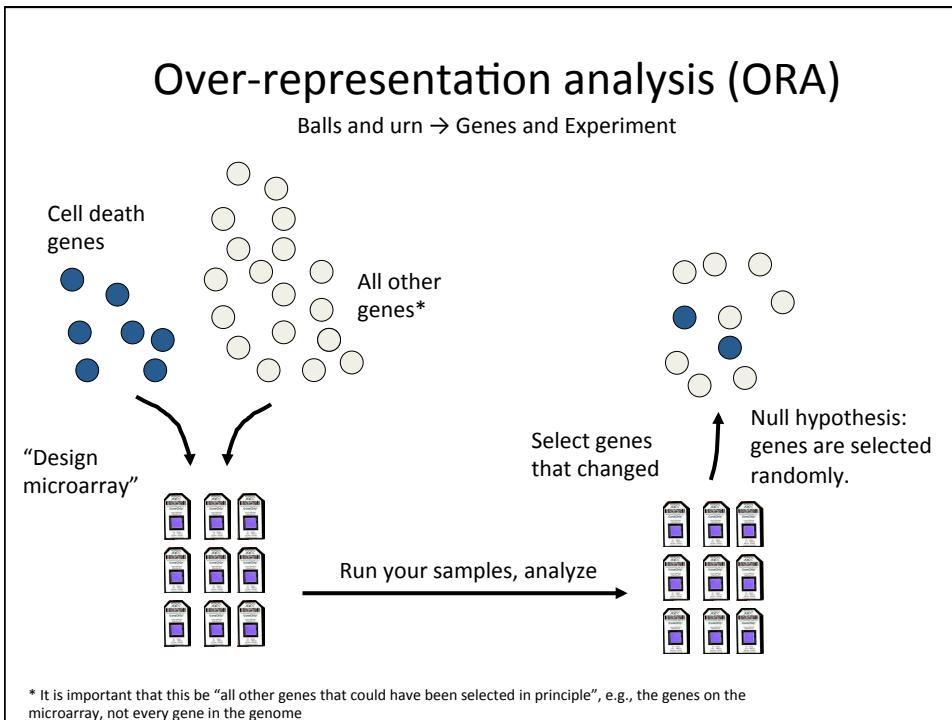
Jesse Gillis



Systematic gene set analysis

1. Get a score for each gene in your data set
 - Categorical ("Cluster number 10" or "Significant genes")
 - Continuous ("Fold-change" or "limma p-value")
 - Make sure you only count each gene once!
2. Assign each gene to one or more gene sets.
3. For each gene set, ask:
 - Is it enriched with my significant genes?
 - Is it enriched with high-scoring genes?

Numerous algorithms, break down to a smaller number of basic possibilities.



Toy example of over-representation analysis

- **100 genes total**, 20 genes are in the “class”.
- Select **10 genes**. 6 are in the GO class. Is this “surprising”?

$$P(\text{success per trial}) = 20/100 = 0.2$$

$$E(\text{in class}) = 0.2 * 10 = 2 \text{ (null model)}$$

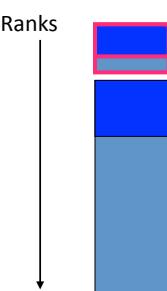
- Binomial probability is

$$\binom{10}{6} 0.2^6 0.8^4 = 210 * 0.00026 = 0.0055$$

Cumulative probability (6 or more successes) is 0.0063

$$\text{Hypergeometric: } \frac{\binom{20}{6} \binom{80}{4}}{\binom{100}{10}} = \frac{38760 * 1581580}{17310309456440} = 0.0035$$

Cumulative probability (6 or more successes) is 0.0039



Issues with the ORA approach

- What if no genes are “significant”?
- What if some of the “involved” genes are not “significant” (e.g., just miss the threshold)
- In general, the results can be sensitive to the choice of the threshold for gene selection

Non-thresholding alternatives

- ROC or other rank-based: GSEA, ErmineJ, Catmap
 - General idea: think Mann-Whitney ‘U’ test.
- Gene score resampling: ErmineJ
- However, be careful using if you have “no signal”
e.g. a uniform p-value distribution

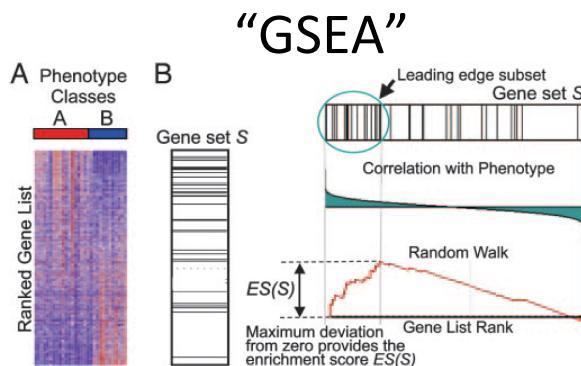
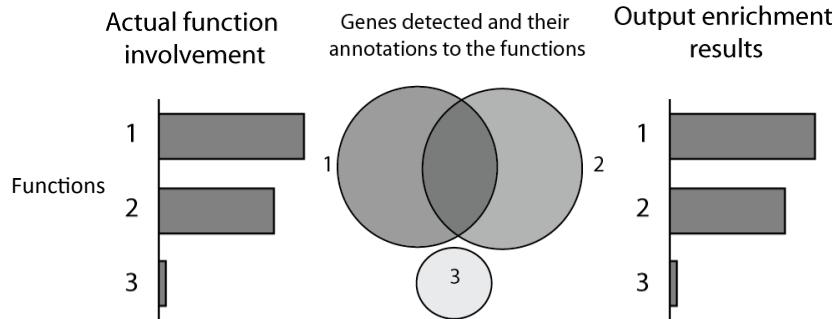


Fig. 1. A GSEA overview illustrating the method. (A) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the “gene tags,” i.e., location of genes from a set S within the sorted list. (B) Plot of the running sum for S in the data set, including the location of the maximum enrichment score (ES) and the leading-edge subset.

Subramanian et al. 2005 PNAS

- Modification of the Kolmogorov-Smirnov test
- Significance established by resampling (Requires gene scores can be represented as correlations, simple experimental designs only)
- Vigorous argument in the literature whether this method is good/appropriate/sensible:
<http://www.ncbi.nlm.nih.gov/pubmed/20048385>
<http://smm.sagepub.com/content/early/2012/09/24/0962280212460441.abstract>

What we hope gene sets can tell us



Jesse Gillis

A high-level summary? Or a discovery?

- The list of “significant” GO groups can be very long. How are we to make sense of this?
- Look at confirmatory findings only? (“That makes sense”); Take the top 10?

Table II. Top 10 Classes Selected by Overrepresentation Analysis Results, $P_{\text{gene}} < 0.001^*$

BA47, ORA		BA9, ORA	
Class [†]	P value [‡]	Class [†]	P value [‡]
[§] Complement activation GO:0006956	7×10^{-8}	[§] Muscle development GO:0007517	2.45×10^{-5}
[§] Complement activation, classical pathway GO:0006958	7.3×10^{-7}	Regulation of cell proliferation GO:0042127	6.02×10^{-5}
Humoral immune response GO:0006959	1.07×10^{-5}	[§] Complement activation, classical pathway GO:0006958	0.0001
Regulation of muscle contraction GO:0006937	0.000141	[§] Humoral immune response GO:0006959	0.000125
Intramolecular isomerase activity, interconverting aldoses and ketoses GO:0016861	0.000177	[§] Complement activation GO:0006956	0.000126
Circadian rhythm GO:0007623	0.000264	Di-, trivalent inorganic cation transport GO:0015674	0.000143
[§] Complement activation, alternative pathway GO:0006957	0.000264	[§] Calmodulin binding activity GO:0005516	0.000187
Complement activity GO:0003811	0.000389	[§] Calcium ion transport GO:0006816	0.000528
Negative regulation of adenylyl cyclase activity GO:0007194	0.000523	Nutritional response pathway GO:0007584	0.000794
Neuropeptide hormone activity GO:0005184	0.000523	Tubulin binding activity GO:0015631	0.000845

Example of use for discovery

Converging Genetic and Functional Brain Imaging Evidence Links Neuronal Excitability to Working Memory, Psychiatric Disease, and Brain Activity

<http://www.cell.com/neuron/abstract/S0896-6273%2814%2900015-4>

Rank genes by their genetic association with memory performance; nothing is statistically significant, so do GO enrichment analysis:

Among the 1,411 database-derived gene sets that served as input, MAGENTA identified multiple testing-corrected significant enrichment ($P_{FDR} < 0.05$) for two gene sets: “voltage-gated cation channel activity” (Gene Ontology ID [GO]:0022843) and “transport of glucose and other sugars, bile salts and organic acids, metal ions, and amine compounds” (gene set database: Reactome; $P_{\text{nominal}} = 9.9 \times 10^{-5}$ and 9.9×10^{-5} , respectively;

Two-Back Makes Step Forward in Brain Imaging Genomics

Andrew McIntosh,^{1,2} Ian Deary,² and David J. Porteous^{2,3,*}

“Given the objective of the study, you have to like the look of the first class, because these channels are known to regulate neuronal excitability, but what can you make of the second? This is both the blessing and the curse of the general approach—gene ontologies reflect the state of knowledge about gene function, which is both limited and biased by virtue of the foregoing research. Even those that survive the false discovery tests will inevitably include false positives.”

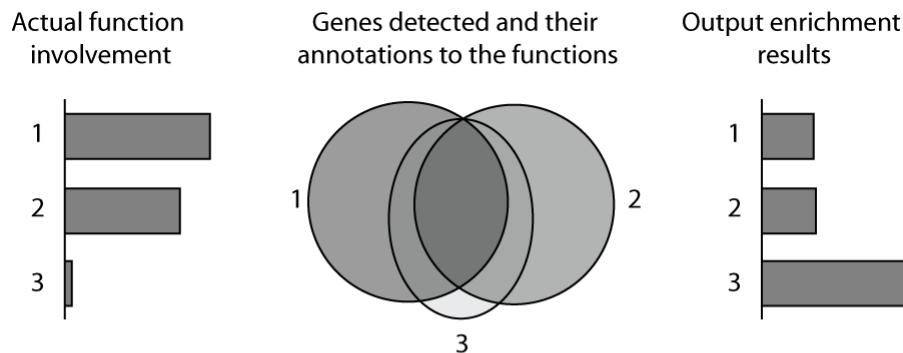
Table 1. G SYSTEMS BIOLOGY

Symbol	Retraction for “Identification of gene ontologies linked to pre-frontal-hippocampal functional coupling in the human brain,” by Luanna Dixson, Henrik Walter, Michael Schneider, Susanne Erk, Axel Schäfer, Leila Haddad, Oliver Grimm, Manuel Mattheisen, Markus M. Nöthen, Sven Cichon, Stephanie H. Witt, Marcella Rietschel, Sebastian Mohnke, Nina Seiferth, Andreas Heinz, Heike Tost, and Andreas Meyer-Lindenberg, which appeared in issue 26, July 1, 2014, of <i>Proc Natl Acad Sci USA</i> (111:9657–9662; first published June 16, 2014; 10.1073/pnas.1404082111).
<i>NLGN1</i>	The authors wish to note the following: “In this paper we report an association of the ‘synapse organization and biogenesis’ gene set with a neuroimaging phenotype, using gene set enrichment methodology. The methods and results of the paper, as described, have been conducted after consultation with experts in the field and support this conclusion. However, a potential confound relating to statistical inference has been brought to our attention that arises from the fact that several clustered genes, all of which are included in this gene set, have been tagged by the same SNP. This problem, which concerns only a small fraction of our tested gene sets (unfortunately including our top finding), belongs to a known category of potential pitfalls in gene set association analyses, and we are sorry that this problem was not detected earlier. Our reanalyses suggest that if adjustments for this confound are applied, the results for our top finding no longer reach experiment-wide significance. Therefore, we feel that the presented findings are not currently sufficiently robust to provide definitive support for the conclusions of our paper, and that an extensive reanalysis of the data is required. The authors have therefore unanimously decided to retract this paper at this time.”
<i>KLK8</i>	
<i>CACNB2</i>	
<i>PCDHB9</i>	
<i>PCDHB10</i>	
<i>PCDHB11</i>	
<i>PCDHB6</i>	
<i>PCDHB5</i>	
<i>PCDHB16</i>	
<i>PCDHB14</i>	
<i>PCDHB13</i>	
<i>GHRL</i>	
<i>NRCAM</i>	
<i>PCDHB4</i>	
<i>COL4A4</i>	
<i>POU4F1</i>	
<i>AGRN</i>	
<i>ACHE</i>	
<i>UBB</i>	
<i>CHRNA1</i>	
<i>NRD1</i>	
<i>PCDHB2</i>	
<i>PCDHB3</i>	

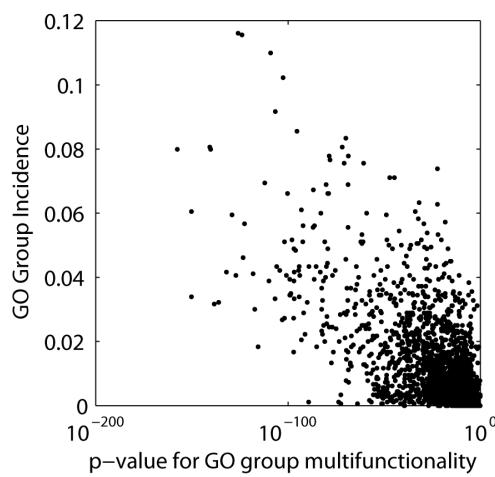
ene list	P
4	.0017
5	.0090
8	.0106
9	.0114
0	.0114
1	.0114
4	.0114
4	.0114
9	.0114
3	.0114
5	.0114
1	.0208
1	.0306
3	.0310
9	.0343
2	.0724
3	.1080
4	.1222
3	.1372
5	.2309
6	.2723
3	.3875
4	.3875

(from left). The fourth columns.

How multifunctionality can affect findings



Multifunctional GO groups are more likely to show up in enrichment analysis



p value on ROCs for GO functions using gene multifunctionality list vs GO group incidence (corrected p<0.05) in MolSigDB
(unsmoothed, rank r -0.59)

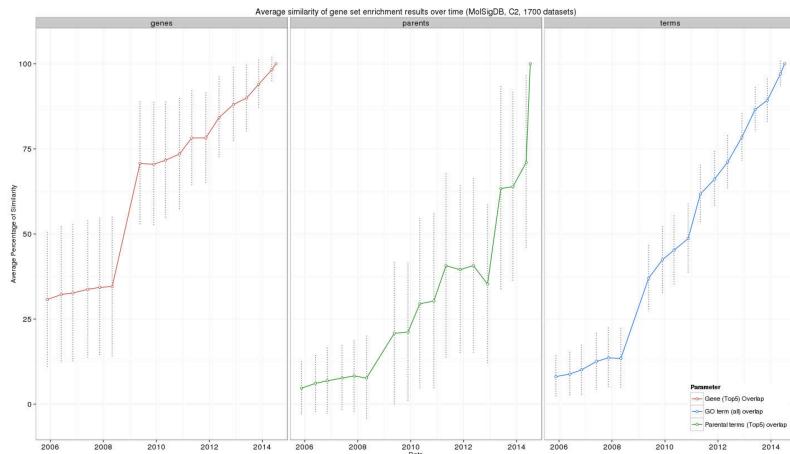
Jesse Gillis

Testing for multifunctionality effects

- Only necessary if your ‘hit list’ is biased towards multifunctionality
 - Corrections that simply reduce redundancy of gene groups don’t fix it
 - Removing multifunctional genes “by default” overcorrects
- Two approaches:
 - Iteratively remove multifunctional genes; test for sensitivity of enrichment results (hit-list tests)
 - Regress out the multifunctionality bias (rank-based tests)
- Implementation: ErmineJ 3.0 (erminej.chibi.ubc.ca)

Reanalysis of hit list reported in: Blood 105: 659-669.

Annotations changing can affect enrichment analysis



Adri Sedeño

Summary for gene set analysis

- Treat it as exploratory
- Do not use enrichment analysis as your primary result.
- Beware of potential for annotations to change; they are not “truth”
- Don’t forget that ontology structure implies additional annotations (“inferred”)

Also:

- You don’t have to use GO – could test for enrichment in disease genes, by chromosome, pathway etc.
- Not that much agreement on how to do enrichment analysis (see additional slides for some discussion)

Additional topics

Self-contained vs. competitive tests

- Most methods focus on comparing “genes in the set to all genes” (competitive)
 - Null hypothesis is “the genes in set X are at most as often differentially expressed as all the genes”
- Goeman and Boehm (2007) suggest “self-contained tests”
 - Null hypothesis is “There is no differential expression in the gene set X”.
 - Basic idea: look at the p-value distribution for the genes in the set. If it is non-uniform, reject null. Amounts to doing FDR-type analysis on just the set.
(That's not how they do it in the paper, but that's the gist)

This has not really caught on (though the paper is justifiably influential)

Bioinformatics, Vol. 23 no. 8 2007, pages 980–987

Clarifying the distinction

- Imagine a GO term that has only one gene
 - Self-contained test reduces to simply the hypothesis test for diff. ex. on that one gene.
 - Competitive test breaks down: “background” makes no sense.
- Imagine all genes are differentially expressed
 - All groups will be sig. by self-contained test.
 - None by competitive test.
- Which is “right”? Depends.

Multiple test correction

- **Problem:** gene sets are correlated (overlap)
- Simple approaches (e.g. Benjamini-Hochberg) likely to yield conservative results.
- Numerous attempts to fix this.
- Reducing numbers of groups tested is helpful.
 - GO Slim, etc.
- The structure of the GO hierarchy could be exploited.
 - See paper by Goeman and Mansmann on “focus level” analysis (Bioinformatics 2008 24:537)
 - User selects a level in the GO tree to start
 - ‘globaltest’ package in Bioconductor
- However: Often the problem is too many results, not too few.

Gene correlation

- Most methods assume genes are independent, but expression is correlated with GO categories (in some cases).
 - Example: Ribosomal proteins
- In those cases, null hypothesis that genes are randomly distributed is not reasonable.
- Problem is worst when no genes are statistically significantly changed on their own.
 - A small (chance) bias in the rankings for the ribosomal protein genes can result in a top score.

See Goeman and Buhlmann Bioinformatics 2007 23: 980 for discussion

Addressing gene correlation: Sample permutations

- Conceptually want multiple realizations of the experiment under the null with respect to differential expression (or whatever we are studying).
- One solution: **resampling over the subjects** to get a null
 - Shuffle the subject labels, repeat the linear modeling etc. (gene ranking)
 - Compute GO enrichment analysis (however you like)
 - Repeat many times
 - Estimate p-values for real results from these empirical distributions
- Problem: not easily applicable if you have complex experimental design.
- Alternative: “rotations”, suggested by Smyth
 - See Wu et al. Bioinformatics 2010 26:2176 for “self-contained” version; “competitive” version implemented in limma::romer