STAT 540 2015
Analysis of gene function I: Gene set analysis

Paul Pavlidis

Outline

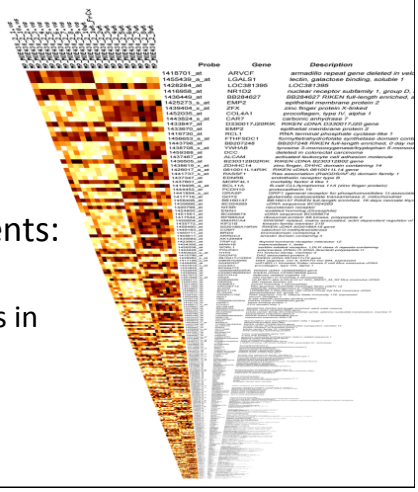
- Motivation: gene function and gene lists
- The Gene Ontology
- Enrichment analysis methods
- Difficulties, pitfalls and partial solutions

- Next lecture:
  - Gene networks
  - Graph clustering and 'guilt by association'

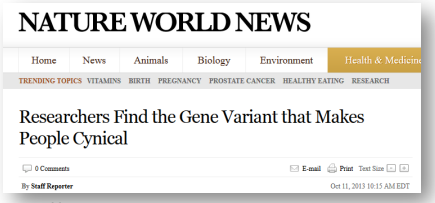
Gene function: a central concept of reductionist biology

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"

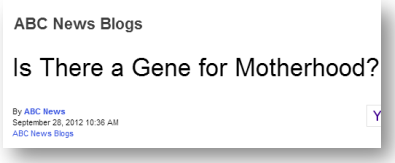


http://www.dailymail.co.uk/



http://www.natureworldnews.com

http://www.torontosun.com



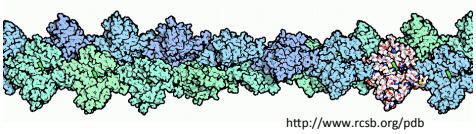
http://www.sciencedaily.com/

A useful distinction (?): "Proximal" function

About what the gene product does more or less directly.

Other "functions" require the proximal functions

Example: **Actin**



http://www.rcsb.org/pdb

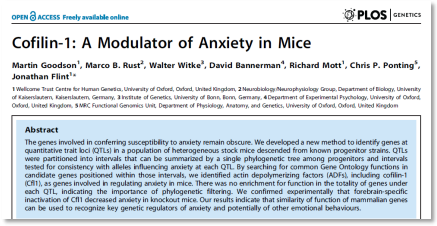
Actin forms filaments which help give cells shape and structure, and helps them move or change shape. (review: http://www.sciencemag.org/content/326/5957/1208.full)

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.



OPEN ACCESS Freely available online PLOS | GENETICS

Cofilin-1: A Modulator of Anxiety in Mice

Martin Goodson[1], Marco B. Rust[2], Walter Witke[3], David Bannerman[4], Richard Mott[1], Chris P. Ponting[5], Jonathan Flint[1]*

1 Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom, 2 Neurobiology/Neurophysiology Group, Department of Biology, University of Kaiserslautern, Kaiserslautern, Germany, 3 Institute of Genetics, University of Bonn, Bonn, Germany, 4 Department of Experimental Psychology, University of Oxford, Oxford, United Kingdom, 5 MRC Functional Genomics 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 summarized by a single phylogenetic tree among progenitors and intervals tested for consistency with alleles influencing anxiety at each QTL. By searching for common Gene Ontology functions in candidate genes positioned within those intervals, we identified actin depolymerizing factors (ADFs), including cofilin-1 (Cfl1), as genes involved in regulating anxiety in mice. There was no enrichment for function in the totality of genes under 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. Rubenstein[1]

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.

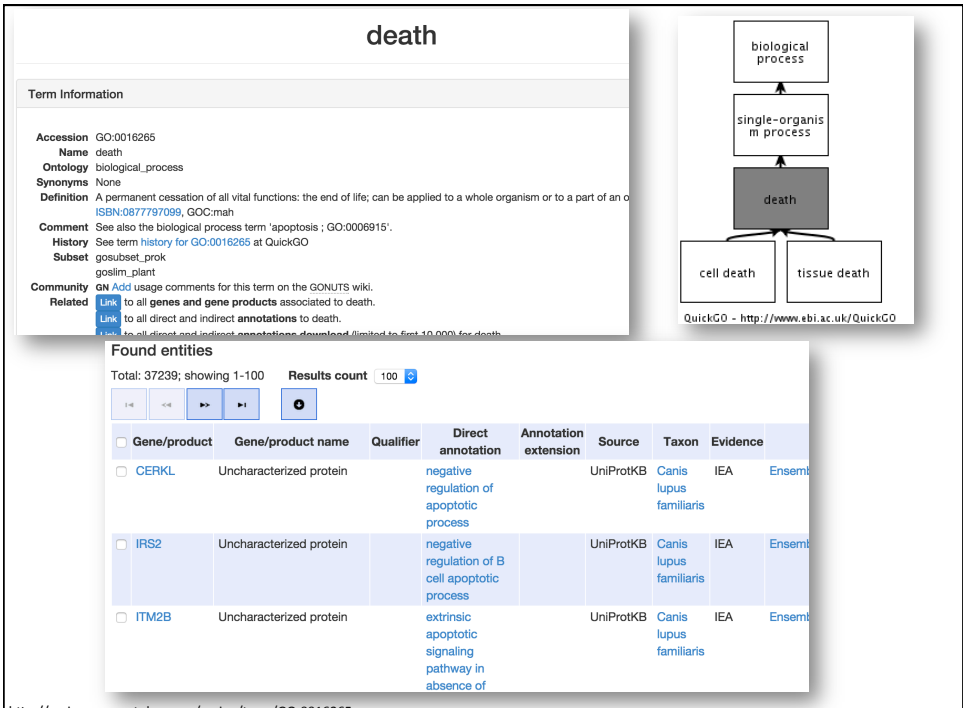
This is an important part of what we mean by a gene having a function – it's *specifically* involved, relative to other genes.

This gets trickier when the gene 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

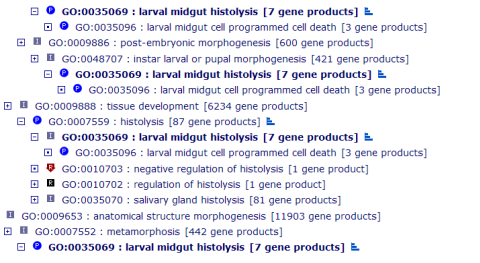
- Controlled vocabulary for function
- Motivated in part by need for species comparisons
  - Genome sequencing made uniform annotations ever more urgent
- Started ~1999 (mouse, fly, worm)
  - Built in part on earlier efforts such as Enzyme Commission
- Database curators use these terms to describe genes.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

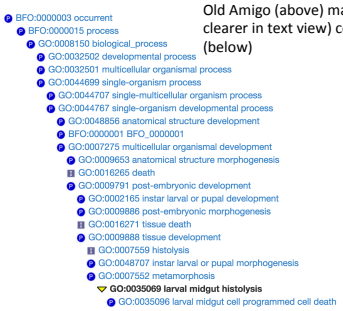


http://amigo.geneontology.org/amigo/term/GO:0016265

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.



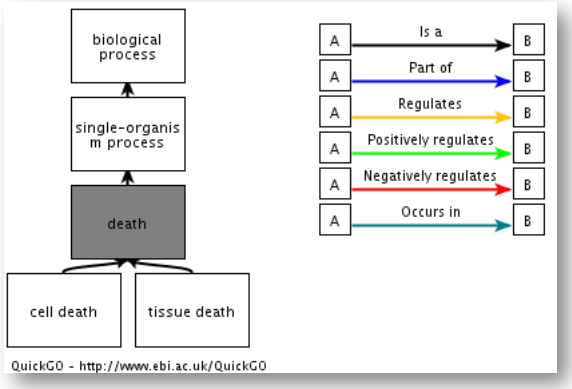


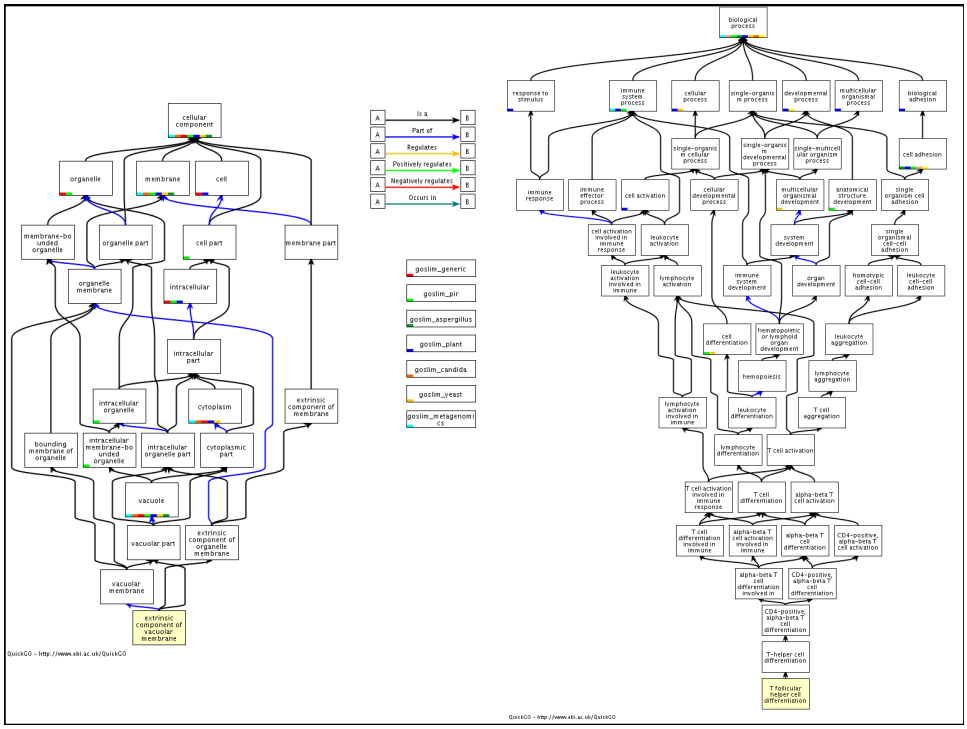
Old Amigo (above) made multiple paths clearer in text view) compared to Amigo2 (below)

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

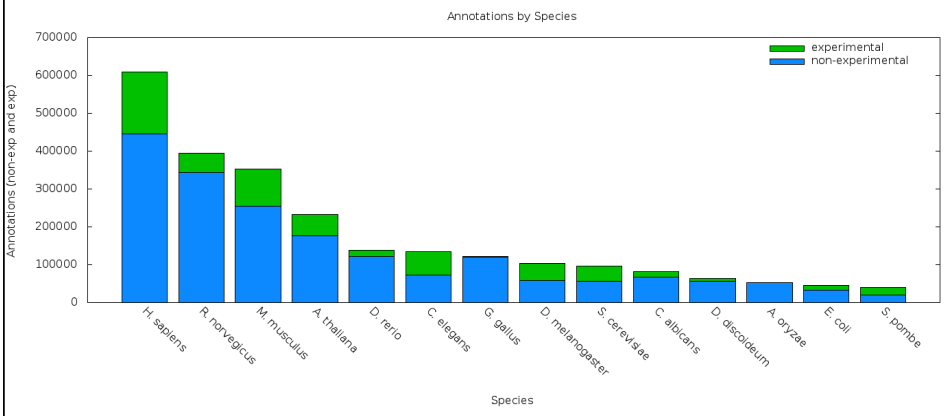
*October 2007 release

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)



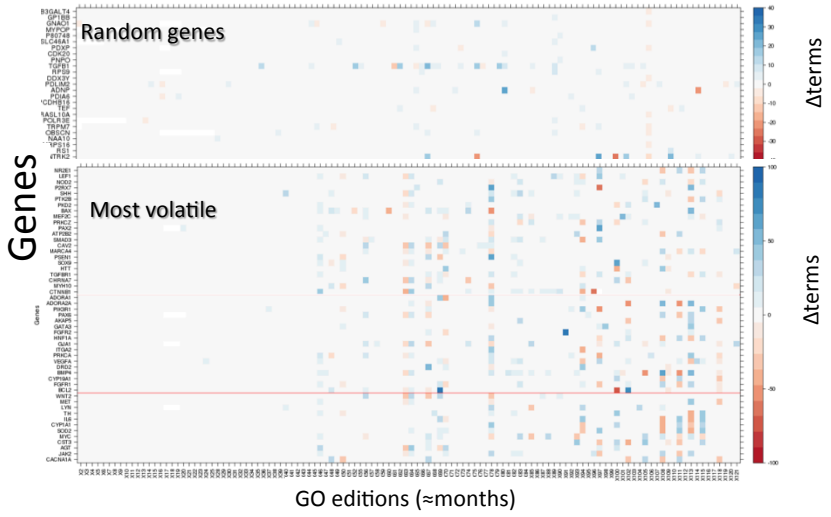
This is somewhat misleading because of varying genome size; annotations per gene would be a bit flatter (also omits inferred annotations)

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
- Inaccurate
  - Evaluations suggest ~50% reliable on average
    - http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1002533
- 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

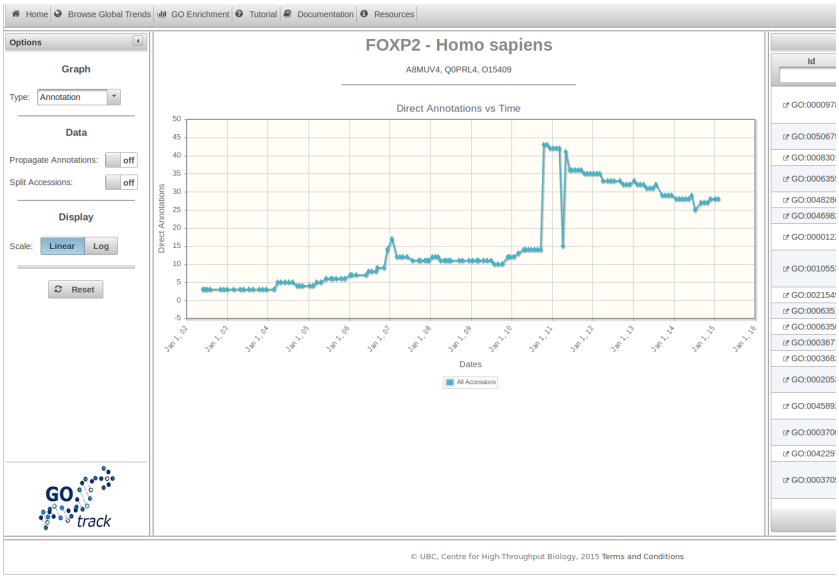
GO annotations are volatile



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

Adri Sedeño

GOTrackWeb



Adri Sedeño and Matt Jacobson

Multifunctionality

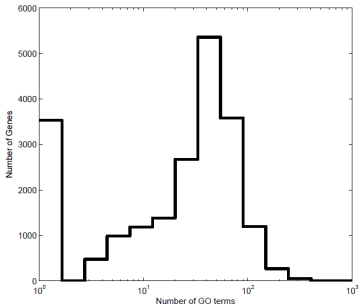
- "Some genes have more functions than others"
- Operational definition: Number of Gene Ontology terms (or other annotations)
  - This definition is appropriate inasmuch it is what we actually use in our analyses.

22

Distribution of annotations per gene

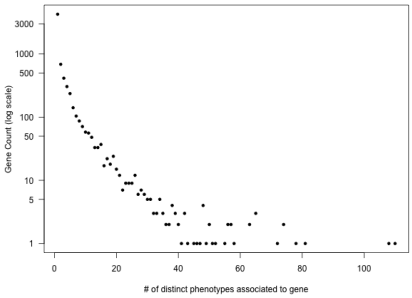
- Extremely uneven – many genes have no terms, others have hundreds.
- Unclear whether this is "popularity" or biological reality

GO annotations



20710 human genes
J Gillis

Disease annotations



http://www.chibi.ubc.ca/Gemma/phenotypes.html
Carolyn Ch'ng et al.

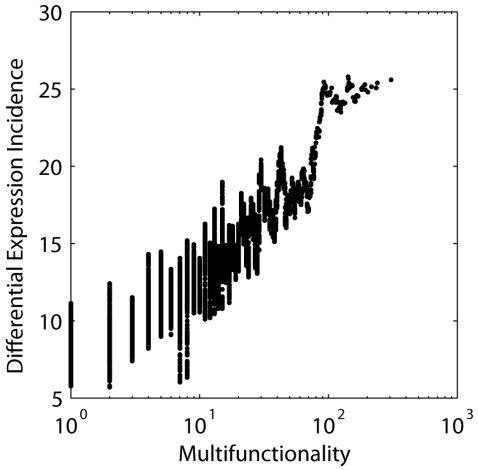
Some of the heavy hitters

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 BM… … … … … … … … … CDK5
CDKN2A … … … … … … … … … DPYD
DRD1 … … … … … … … … … ENTPD4
FASN … … … … … … … … … GH1
GHRL … … … … … … … … … HPRT1
HPX … … … … … … … … … IL6
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

No matter what you are studying, simply guessing that one of these genes is involved will get you surprisingly far.

24

Multifunctional genes tend to show up in genomics studies



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

Analysis of MolSigDB gene lists; Spearman correlation = 0.48, shown is sliding window of 100 genes

Gillis J and Pavlidis P, In preparation

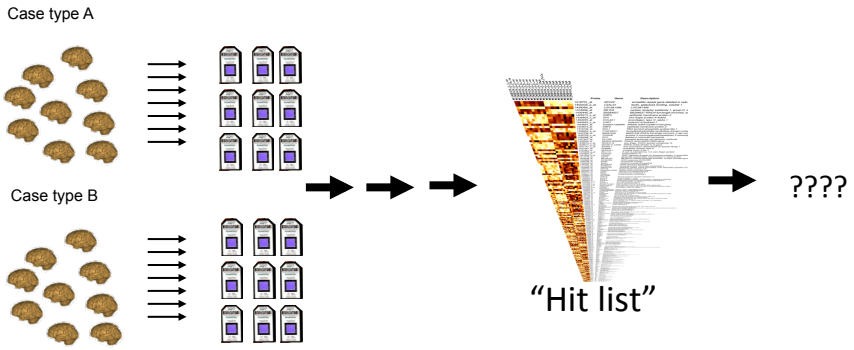
25

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?

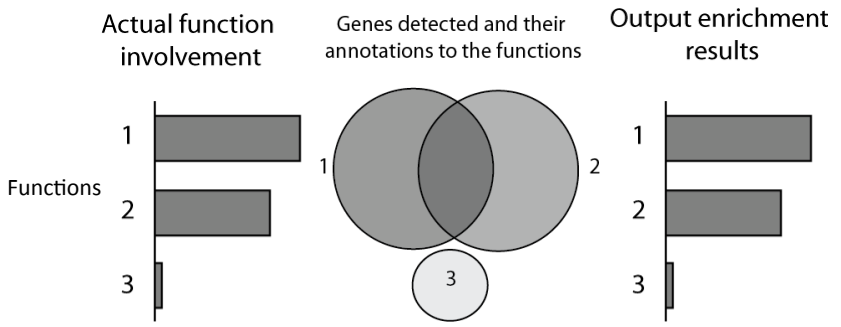
- In many of the studies we talk about in class, the goal is to figure out function of genes, but without a specific hypothesis: a "screen"
- Interpreting the results of the screen requires prior knowledge.



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

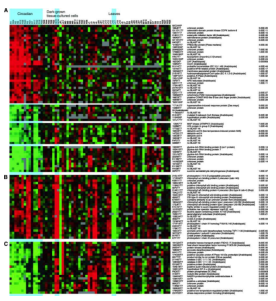


Table 1.

Categories of Genes That Cycle[a]

Category	A.M.	P.M.
Unknown	18	17
No similarity	16	23
Transcription factors	5	3
Kinase/phosphatase	4	3
Photosynthesis/carbon metabolism	7	4
Auxin response	1	3
Stress (cold/pathogen/water)	0	9
Membrane proteins	0	2
Glycine-rich RNA binding	0	3
Other	8	14

Microarray Analysis of Diurnal and Circadian-Regulated Genes in Arabidopsis
Robert Schaffer, Jeff Landgraf, Monica Accerbi, Vernadette Simon, Matt Larson, and Ellen Wisman[2]
Plant Cell. 2001 January; 13(1): 113–124.

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC102203/

"When the potential functions of these genes were identified, the genes could be classified into one of two categories: those that are regulated only by the circadian clock and those that may play a role in clock function. Some genes, including *APS* reductase (73F9T7), formamidase (91H14T7), and catalase (clone 154N18T7), would fall into the first category, whereas *LHY*, *CCA1*, and *GI* have already been shown to play some part in circadian clock function. Among members of this group are many genes of unknown function. However, a number of potential regulators were identified, including putative transcription factors and post-translational modifiers. *CCR1*, *CCR2*, and genes involved in dehydration were also identified in this cluster…"

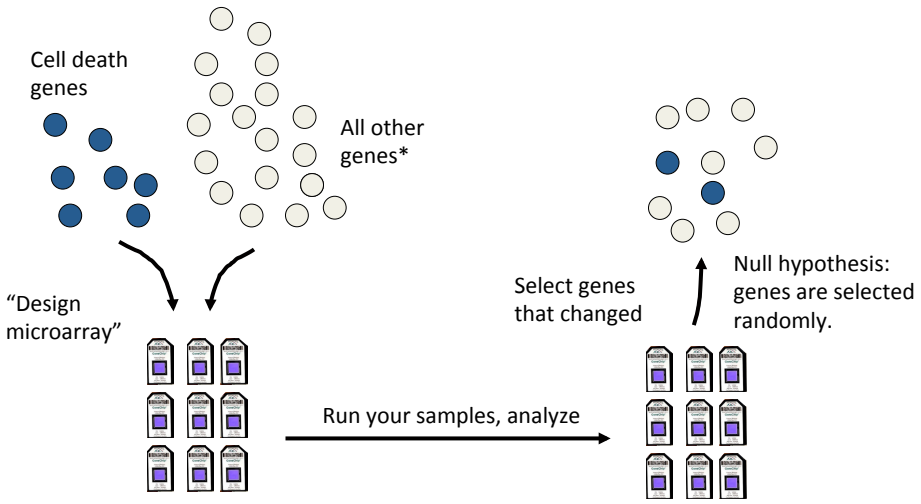
Systematic gene set analysis

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")
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.

Over-representation analysis (ORA)

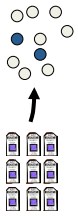
Balls and urn → Genes and Experiment



* It is important that this be "all other genes that could have been selected in principle", e.g., the genes on the microarray, not every gene in the genome

Hypergeometric distribution

- Choose *N* balls from an urn of *n* +*m* balls, of which *n* are red and *m* are black. *i* of the selected balls are red, *N-i* are black. (red = success)
- What is probability of this happening?



$$P(x=i) = \frac{[\#\text{ ways for } i \text{ successes}][\#\text{ ways for } N-i \text{ failures}]}{[\text{total number of ways to select}]} = \frac{\binom{n}{i}\binom{m}{N-i}}{\binom{n+m}{N}}$$

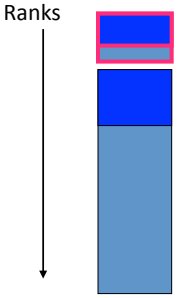
- Cumulative distribution: probability of getting *i or more* successes. Calculated by sum over the above.
- Approximate with the binomial distribution

Number of ways of picking *n* unordered outcomes from *m* possibilities (where m > n)

$$\binom{m}{n} = \frac{m!}{(m-n)!n!}$$

Toy example of over-representation analysis

- **100 genes total**, 20 genes are in the 'class'.
- Select **10 genes**. 6 are in the GO class.



P(success per trial) = 20/100 = 0.2
E(in class) = 0.2 * 10 = 2
• 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

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

"GSEA"

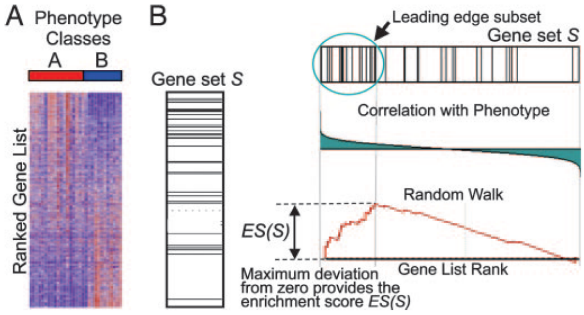
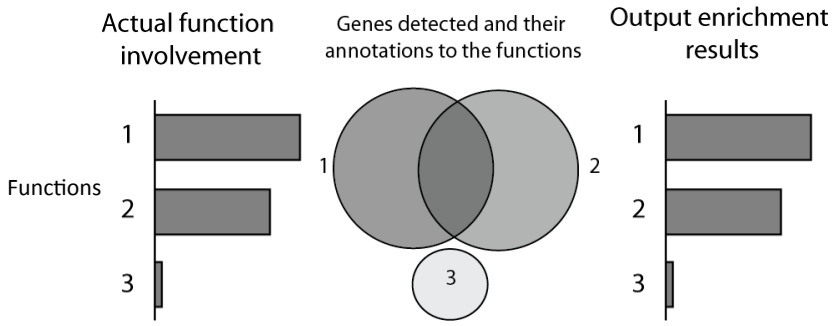


Fig. 1. A GSEA overview illustrating the method. (*A*) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the "gene tags," i.e., location of genes from a set *S* within the sorted list. (*B*) Plot of the running sum for *S* in the data set, including the location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

Subramanianan 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, $P_{gene} < 0.001$*

BA47, ORA		BA9, ORA	
Class†	*P* value‡	Class†	*P* value‡
[a]**Complement activation GO:0006956**	7×10^{-8}	[a]*Muscle development GO:0007517*	2.45×10^{-5}
[a]**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}	[a]**Complement activation, classical pathway GO:0006958**	0.0001
Regulation of muscle contraction GO:0006937	0.000141	[a]**Humoral immune response GO:0006959**	0.000125
Intramolecular isomerase activity, interconverting aldoses and ketoses GO:0016861	0.000177	[a]**Complement activation GO:0006956**	0.000126
Circadian rhythm GO:0007623	0.000264	Di-, trivalent inorganic cation transport GO:0015674	0.000143
[a]*Complement activation, alternative pathway GO:0006957*	0.000264	[a]Calmodulin binding activity GO:0005516	0.000187
Complement activity GO:0003811	0.000389	[a]Calcium ion transport GO:0006816	0.000528
Negative regulation of adenylate 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:

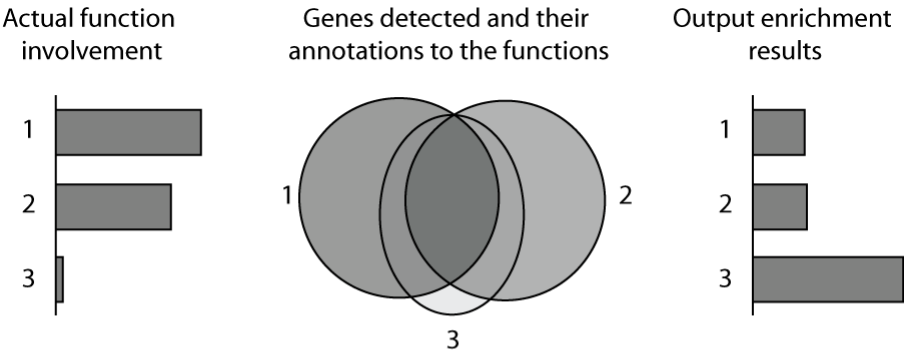
> Among the 1,411 database-derived gene sets that served as input, MAGENTA identified multiple testing-corrected significant enrichment ($P_{FDR} < 0.05$) for two gene sets: ''voltage-gated cation channel activity'' (Gene Ontology ID [GO]:0022843) and ''transport of glucose and other sugars, bile salts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

Two-Back Makes Step Forward in Brain Imaging Genomics

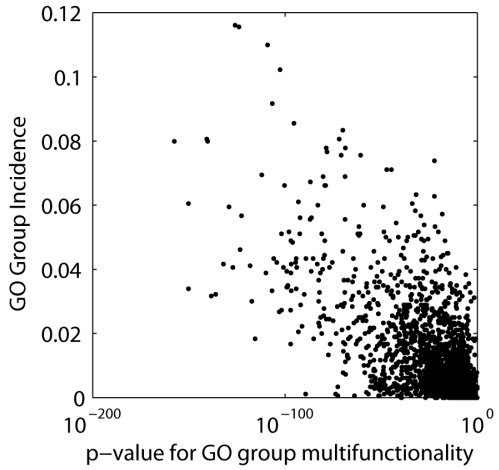
Andrew McIntosh,[1,2] Ian Deary,[2] and David J. Porteous[2,3,*]

"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."

How multifunctionality can affect findings



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

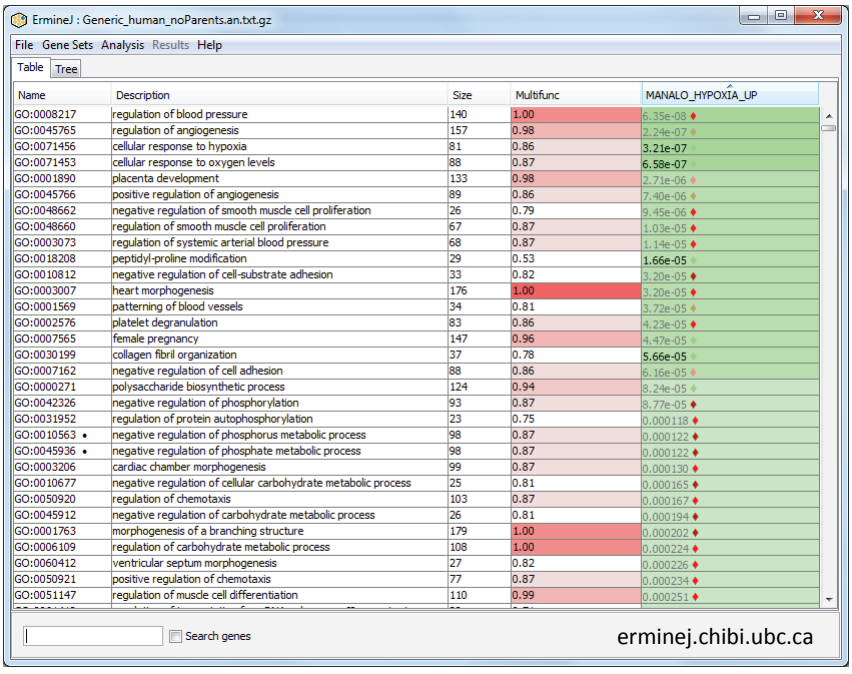


p value on ROCs for GO functions using gene multifunctionality list vs GO group incidence (corrected p<0.05) in MolSigDB (unsmoothed, rank r -0.59)

Jesse Gillis

Testing for multifunctionality effects

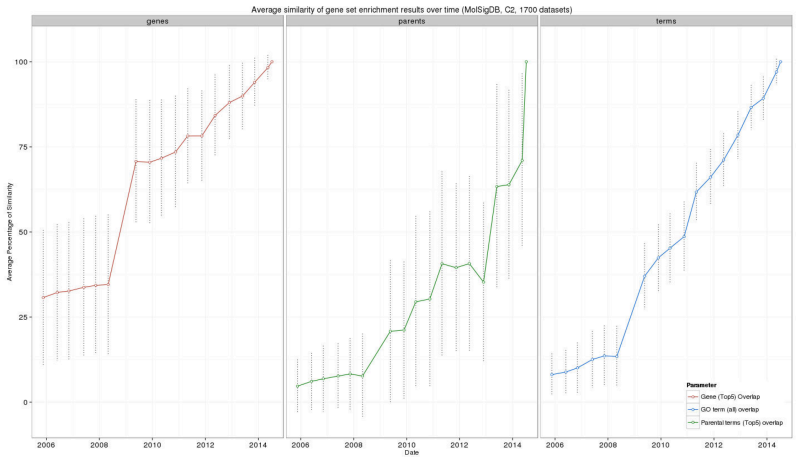
- Only necessary if your 'hit list' is biased towards multifunctionality
  - Corrections that simply reduce redundancy of gene groups don't fix it
  - Removing multifunctional genes by default overcorrects
- Two approaches:
  - Iteratively remove multifunctional genes; test for sensitivity of enrichment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



Name	Description	Size	Multifunc	MANALO_HYPOXIA_UP
GO:0008217	regulation of blood pressure	140	1.00	6.35e-08
GO:0045765	regulation of angiogenesis	157	0.98	2.24e-07
GO:0071456	cellular response to hypoxia	81	0.86	3.21e-07
GO:0071453	cellular response to oxygen levels	88	0.87	6.58e-07
GO:0001890	placenta development	133	0.98	2.71e-06
GO:0045766	positive regulation of angiogenesis	89	0.86	7.40e-06
GO:0048662	negative regulation of smooth muscle cell proliferation	26	0.79	9.45e-06
GO:0048660	regulation of smooth muscle cell proliferation	67	0.87	1.03e-05
GO:0003073	regulation of systemic arterial blood pressure	68	0.87	1.14e-05
GO:0018208	peptidyl-proline modification	29	0.53	1.66e-05
GO:0010812	negative regulation of cell-substrate adhesion	33	0.82	3.20e-05
GO:0003007	heart morphogenesis	176	1.00	3.20e-05
GO:0001569	patterning of blood vessels	34	0.81	3.72e-05
GO:0002576	platelet degranulation	83	0.86	4.23e-05
GO:0007565	female pregnancy	147	0.96	4.47e-05
GO:0030199	collagen fibril organization	37	0.78	5.66e-05
GO:0007162	negative regulation of cell adhesion	88	0.86	6.16e-05
GO:0000271	polysaccharide biosynthetic process	124	0.94	8.24e-05
GO:0042326	negative regulation of phosphorylation	93	0.87	8.77e-05
GO:0031952	regulation of protein autophosphorylation	23	0.75	0.000118
GO:0010563 •	negative regulation of phosphorus metabolic process	98	0.87	0.000122
GO:0045936 •	negative regulation of phosphate metabolic process	98	0.87	0.000122
GO:0003206	cardiac chamber morphogenesis	99	0.87	0.000130
GO:0010677	negative regulation of cellular carbohydrate metabolic process	25	0.81	0.000165
GO:0050920	regulation of chemotaxis	103	0.87	0.000167
GO:0045912	negative regulation of carbohydrate metabolic process	26	0.81	0.000194
GO:0001763	morphogenesis of a branching structure	179	1.00	0.000202
GO:0006109	regulation of carbohydrate metabolic process	108	1.00	0.000224
GO:0060412	ventricular septum morphogenesis	27	0.82	0.000226
GO:0050921	positive regulation of chemotaxis	77	0.87	0.000234
GO:0051147	regulation of muscle cell differentiation	110	0.99	0.000251

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 are focuses 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 Boehmann (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