Introduction to Pathway and Network Analysis

Alison Motsinger-Reif, PhD
Associate Professor
Bioinformatics Research Center
Department of Statistics
North Carolina State University

Pathway and Network Analysis

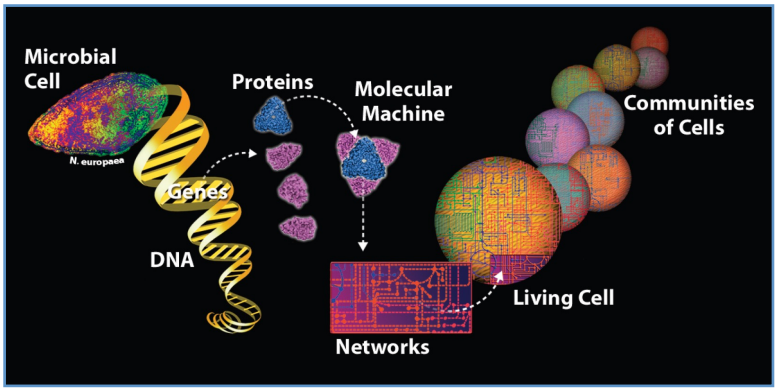
- High-throughput genetic/genomic technologies enable comprehensive monitoring of a biological system
- Analysis of high-throughput data typically yields a list of differentially expressed genes, proteins, metabolites…
  - Typically provides lists of single genes, etc.
  - Will use "genes" throughout, but using interchangeably mostly
- This list often fails to provide mechanistic insights into the underlying biology of the condition being studied
- How to extract meaning from a long list of differentially expressed genes → pathway/network analysis

What makes an airplane fly?

Chas' Stainless Steel, Mark Thompson's Airplane Parts, About 1000 Pounds of Stainless Steel Wire, and Gagosian's Beverly Hills Space

From components to networks

A biological function is a result of many interacting molecules and cannot be attributed to just a single molecule.



Pathway and Network Analysis

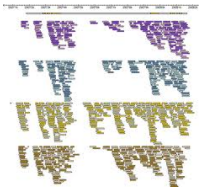
- One approach: simplify analysis by grouping long lists of individual genes into smaller sets of related genesreduces the complexity of analysis.
  - a large number of knowledge bases developed to help with this task
- Knowledge bases
  - describe biological processes, components, or structures in which individual genes \are known to be involved in
  - how and where gene products interact with each other

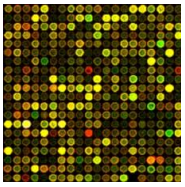
Pathway and Network Analysis

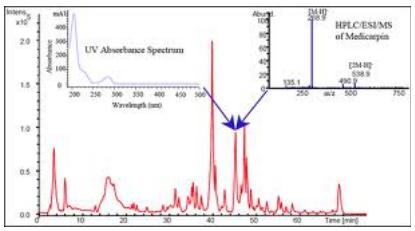
- Analysis at the functional level is appealing for two reasons:
  - First, grouping thousands of genes by the pathways they are involved in reduces the complexity to just several hundred pathways for the experiment
  - Second, identifying active pathways that differ between two conditions can have more explanatory power than a simple list of genes

Pathway and Network Analysis

- What kinds of data is used for such analysis?
  - Gene expression data
    - Microarrays
    - RNA-seq
  - Proteomic data
  - Metabolomics data
  - Single nucleotide polymorphisms (SNPs)
  - ….







Pathway and Network Analysis

- What kinds of questions can we ask/answer with these approaches?

Pathway and Network Analysis

- The term "pathway analysis" gets used often, and often in different ways
  - applied to the analysis of Gene Ontology (GO) terms (also referred to as a "gene set")
  - physical interaction networks (e.g., protein–protein interactions)
  - kinetic simulation of pathways
  - steady-state pathway analysis (e.g., flux-balance analysis)
  - inference of pathways from expression and sequence data
- May or may not actually describe biological pathways

Pathway and Network Analysis

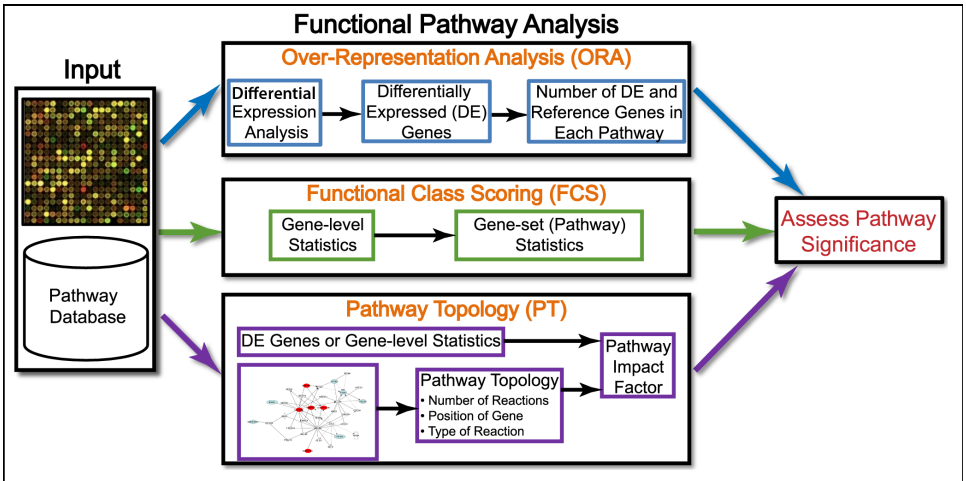
- For the first part of this module, we will focus on methods that exploit pathway knowledge in public repositories rather than on methods that infer pathways from molecular measurements
  - Use repositories such as GO or Kyoto Encyclopedia of Genes and Genomes (KEGG)

→ *knowledge base–driven pathway analysis*

A History of Pathway Analysis Approaches

- Over a decade of development of pathway analysis approaches
- Can be *roughly* divided into three generations:
  - 1st: Over-Representation Analysis (ORA) Approaches
  - 2nd : Functional Class Scoring (FCS) Approaches
  - 3rd : Pathway Topology (PT)-Based Approaches

Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012;8(2):e1002375.



- The data generated by an experiment using a high-throughput technology (e.g., microarray, proteomics, metabolomics), along with functional annotations (pathway database) of the corresponding genome, are input to virtually all pathway analysis methods.
- ORA methods require that the input is a list of differentially expressed genes
- FCS methods use the entire data matrix as input
- PT-based methods additionally utilize the number and type of interactions between gene products, which may or may not be a part of a pathway database.
- The result of every pathway analysis method is a list of significant pathways in the condition under study.

Over-Representation Analysis (ORA) Approaches

- Earliest methods → over-representation analysis (ORA)
- Statistically evaluates the fraction of genes in a particular pathway found among the set of genes showing changes in expression
- It is also referred to as "2×2 table method" in the literature

Over-Representation Analysis (ORA)

- Uses one or more variations of the following strategy:
  - First, an input list is created using a certain threshold or criteria
    - For example, may choose genes that are differentially over- or under-expressed in a given condition at a false discovery rate (FDR) of 5%
  - Then, for each pathway, input genes that are part of the pathway are counted
  - This process is repeated for an appropriate background list of genes
    - (e.g., all genes measured on a microarray)
  - Next, every pathway is tested for over- or under-representation in the list of input genes
    - The most commonly used tests are based on the hypergeometric, chi-square, or binomial distribution

ORA tools	
Onto-Express	Web (http://vortex.cs.wayne.edu)
GenMAPP	Standalone (http://www.genmapp.org)
GoMiner	Standalone, Web (http://discover.nci.nih.gov/gominer)
FatiGO	Web (http://babelomics.bioinfo.cipf.es)
GOstat	Web (http://gostat.wehi.edu.au)
FuncAssociate	Web (http://llama.mshri.on.ca/funcassociate/)
GOToolBox	Web (http://genome.crg.es/GOToolBox/)
GeneMerge	Standalone, Web (http://genemerge.cbcb.umd.edu/)
GOEAST	Web (http://omicslab.genetics.ac.cn/GOEAST/)
ClueGO	Standalone (http://www.ici.upmc.fr/cluego/)
FunSpec	Web (http://funspec.med.utoronto.ca/)
GARBAN	Web
GO:TermFinder	Standalone (http://search.cpan.org/dist/GO-TermFinder/)
WebGestalt	Web (http://bioinfo.vanderbilt.edu/webgestalt/)
agriGO	Web (http://bioinfo.cau.edu.cn/agriGO/)
GOFFA	Standalone, Web (http://edkb.fda.gov/webstart/arraytrack/)
WEGO	Web (http://wego.genomics.org.cn/cgi-bin/wego/index.pl)

Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012;8(2):e1002375.

Limitations of ORA Approaches

- First, the different statistics used by ORA are independent of the measured changes
  - (e.g., hypergeometric distribution, binomial distribution, chi-square distribution, etc.)
- Tests consider the number of genes alone but ignore any values associated with them
  - such as probe intensities
- By discarding this data, ORA treats each gene equally
  - Information about the extent of regulation (e.g., fold-changes, significance of a change, etc.) can be useful in assigning different weights to input genes/pathways
  - This can provide more information

Limitations of ORA Approaches

- Second, ORA typically uses only the most significant genes and discards the others
  - input list of genes is usually obtained using an arbitrary threshold (e.g., genes with fold-change and/or p-values)
- Marginally less significant genes are missed, resulting in information loss
  - (e.g., fold-change = 1.999 or p-value = 0.051)
  - A few methods avoiding thresholds
    - They use an iterative approach that adds one gene at a time to find a set of genes for which a pathway is most significant

Limitations of ORA Approaches

- Third, ORA assumes that each gene is independent of the other genes
- However, biology is a complex web of interactions between gene products that constitute different pathways
  - One goal might be to gain insights into how interactions between gene products are manifested as changes in expression
  - A strategy that assumes the genes are independent is significantly limited in its ability to provide insights
- Furthermore, assuming independence between genes amounts to "competitive null hypothesis" testing (more later), which ignores the correlation structure between genes
  - the estimated significance of a pathway may be biased or incorrect

Limitations of ORA Approaches

- Fourth, ORA assumes that each pathway is independent of other pathways → NOT TRUE!
- Examples of dependence:
  - GO defines a biological process as a series of events accomplished by one or more ordered assemblies of molecular functions
  - The cell cycle pathway in KEGG where the presence of a growth factor activates the MAPK signaling pathway
    - This, in turn, activates the cell cycle pathway
- No ORA methods account for this dependence between molecular functions in GO and signaling pathways in KEGG

Functional Class Scoring (FCS) Approaches

- *The hypothesis of functional class scoring (FCS) is that although large changes in individual genes can have significant effects on pathways, weaker but coordinated changes in sets of functionally related genes (i.e., pathways) can also have significant effects*
- With few exceptions, all FCS methods use a variation of a general framework that consists of the following three steps.

Step 1

- First, a gene-level statistic is computed using the molecular measurements from an experiment
  - Involves computing differential expression of individual genes or proteins
- Statistics currently used at gene-level include correlation of molecular measurements with phenotype
  - ANOVA
  - Q-statistic
  - signal-to-noise ratio
  - t-test
  - Z-score

Step 1

- Choice of a gene-level statistic generally has a negligible effect on the identification of significantly enriched gene sets
  - However, when there are few biological replicates, a regularized statistic may be better
- Untransformed gene-level statistics can fail to identify pathways with up- and down-regulated genes
  - In this case, transformation of gene-level statistics (e.g., absolute values, squared values, ranks, etc.) is better

Step 2

- Second, the gene-level statistics for all genes in a pathway are aggregated into a single pathway-level statistic
  - can be multivariate and account for interdependencies among genes
  - can be univariate and disregard interdependencies among genes
- The pathway-level statistics used include:
  - Kolmogorov-Smirnov statistic
  - sum, mean, or median of gene-level statistic
  - Wilcoxon rank sum
  - maxmean statistic

Step 2

- Irrespective of its type, the power of a pathway-level statistic depends on
  - the proportion of differentially expressed genes in a pathway
  - the size of the pathway
  - the amount of correlation between genes in the pathway
- Univariate statistics show more power at stringent cutoffs when applied to real biological data, and equal power as multivariate statistics at less stringent cutoffs

Step 3

- Assessing the statistical significance of the pathway-level statistic
- When computing statistical significance, the null hypothesis tested by current pathway analysis approaches can be broadly divided into two categories:
  - i) competitive null hypothesis
  - ii) self-contained null hypothesis
- A self-contained null hypothesis permutes class labels (i.e., phenotypes) for each sample and compares the set of genes in a given pathway with itself, while ignoring the genes that are not in the pathway
- A competitive null hypothesis permutes gene labels for each pathway, and compares the set of genes in the pathway with a set of genes that are not in the pathway

FCS tools	
GSEA	Standalone (http://www.broadinstitute.org/gsea/)
sigPathway	Standalone (BioConductor)
Category	Standalone (BioConductor)
SAFE	Standalone (BioConductor)
GlobalTest	Standalone (BioConductor)
PCOT2	Standalone (BioConductor)
SAM-GS	Standalone (http://www.ualberta.ca/~yyasui/software.html)
Catmap	Standalone (http://bioinfo.thep.lu.se/catmap.html)
T-profiler	Web (http://www.t-profiler.org)
FunCluster	Standalone (http://corneliu.henegar.info/FunCluster.htm)
GeneTrail	Web (http://genetrail.bioinf.uni-sb.de)
GAzer	Web

Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012;8(2):e1002375.

Advantages of FCS Methods

FCS methods address three limitations of ORA

1. Don't require an arbitrary threshold for dividing expression data into significant and non-significant pools.

   Rather, FCS methods use all available molecular measurements for pathway analysis.

2. While ORA completely ignores molecular measurements when identifying significant pathways, FCS methods use this information in order to detect coordinated changes in the expression of genes in the same pathway

3. By considering the coordinated changes in gene expression, FCS methods account for dependence between genes in a pathway

Limitations of FCS Methods

- First, similar to ORA, FCS analyzes each pathway independently
  - Because a gene can function in more than one pathway, meaning that pathways can cross and overlap
  - Consequently, in an experiment, while one pathway may be affected in an experiment, one may observe other pathways being significantly affected due to the set of overlapping genes
- Such a phenomenon is very common when using the GO terms to define pathways due to the hierarchical nature of the GO

Limitations of FCS Methods

- Second, many FCS methods use changes in gene expression to rank genes in a given pathway, and discard the changes from further analysis
  - For instance, assume that two genes in a pathway, A and B, are changing by 2-fold and 20-fold, respectively
  - As long as they both have the same respective ranks in comparison with other genes in the pathway, most FCS methods will treat them equally, although the gene with the higher fold-change should probably get more weight
- Importantly, however, considering only the ranks of genes is also advantageous, as it is more robust to outliers.
  - A notable exception to this scenario is approaches that use gene-level statistics (e.g., t-statistic) to compute pathway-level scores.
  - For example, an FCS method that computes a pathway-level statistic as a sum or mean of the gene-level statistic accounts for a relative difference in measurements (e.g., Category, SAFE).

Pathway Topology (PT)-Based Approaches

- A large number of publicly available pathway knowledge bases provide information beyond simple lists of genes for each pathway
  - KEGG
  - MetaCyc
  - Reactome
  - RegulonDB
  - STKE
  - BioCarta
  - PantherDB
  - ….
- Unlike GO and MSigDB, these knowledge bases also provide information about gene products that interact with each other in a given pathway, how they interact (e.g., activation, inhibition, etc.), and where they interact (e.g., cytoplasm, nucleus, etc.)

Pathway Topology (PT)-Based Approaches

- ORA and FCS methods consider only the number of genes in a pathway or gene coexpression to identify significant pathways, and ignore the additional information available from these knowledge bases
  - Even if the pathways are completely redrawn with new links between the genes, as long as they contain the same set of genes, ORA and FCS will produce the same results
- Pathway topology (PT)-based methods have been developed to use the additional information
  - PT-based methods are essentially the same as FCS methods in that they perform the same three steps as FCS methods
  - The key difference between the two is the use of pathway topology to compute gene-level statistics

Pathway Topology (PT)-Based Approaches

- Rahnenfuhrer et al. proposed ScorePAGE, which computes similarity between each pair of genes in a pathway (e.g., correlation, covariance, etc.)
  - similarity measurement between each pair of genes is analogous to gene-level statistics in FCS methods
  - averaged to compute a pathway-level score
- Instead of giving equal weight to all pairwise similarities, ScorePAGE divides the pairwise similarities by the number of reactions needed to connect two genes in a given pathway

Pathway Topology (PT)-Based Approaches

- Impact factor (IF) analysis
  - IF considers the structure and dynamics of an entire pathway by incorporating a number of important biological factors, including changes in gene expression, types of interactions, and the positions of genes in a pathway

Ali will talk more about these approaches in detail!!!

IF Analysis

- Briefly…
  - Models a signaling pathway as a graph, where nodes represent genes and edges represent interactions between them
  - Defines a gene-level statistic, called perturbation factor (PF) of a gene, as a sum of its measured change in expression and a linear function of the perturbation factors of all genes in a pathway
  - Because the PF of each gene is defined by a linear equation, the entire pathway is defined as a linear system
    - addresses loops in the pathways
  - The IF of a pathway (pathway-level statistic) is defined as a sum of PF of all genes in a pathway

Pathway Topology (PT)-Based Approaches

- FCS methods that use correlations among genes implicitly assume that the underlying network, as defined by the correlation structure, does not change as the experimental conditions change
- This assumption may be inaccurate → PT approaches improve on this

Pathway Topology (PT)-Based Approaches

- NetGSA accounts for the the change in correlation as well as the change in network structure as experimental conditions change
  - like IF analysis, models gene expression as a linear function of other genes in the network
- it differs from IF in two aspects
  - First, it accounts for a gene's baseline expression by representing it as a latent variable in the model
  - Second, it requires that the pathways be represented as directed acyclic graphs DAGs
    - If a pathway contains cycles, NetGSA requires additional latent variables affecting the nodes in the cycle.
    - In contrast, IF analysis does not impose any constraint on the structure of a pathway

Limitations of PT-based Approaches

- True pathway topology is dependent on the type of cell due to cell-specific gene expression profiles and condition being studied
  - information is rarely available
  - fragmented in knowledge bases if available
  - As annotations improve, these approaches are expected to become more useful
- Inability to model dynamic states of a system
- Inability to consider interactions between pathways due to weak inter-pathway links to account for interdependence between pathways

PT-based tools	
ScorePAGE	No implementation available
Pathway-Express	Web (http://vortex.cs.wayne.edu)
SPIA	Standalone (BioConductor)
NetGSA	No implementation available

Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012;8(2):e1002375.

Outstanding Challenges

- Broad Categories:

1. annotation challenges
2. methodological challenges

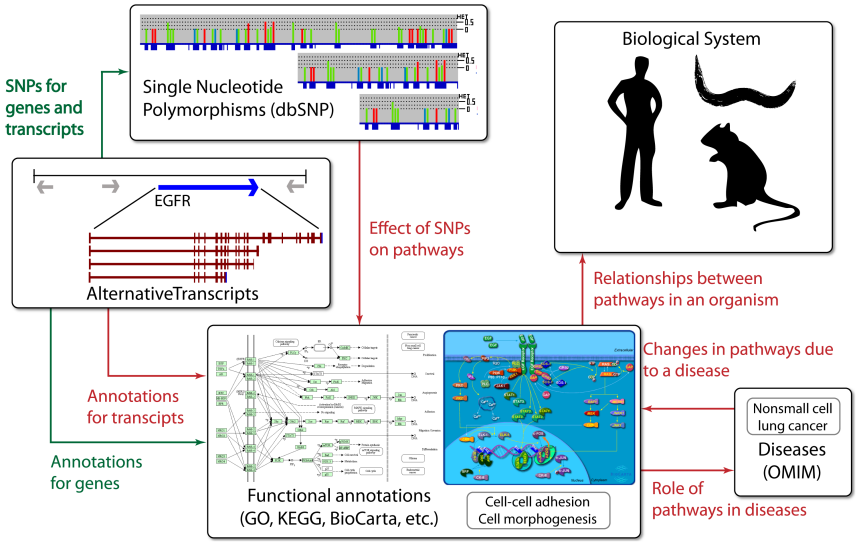
Outstanding Challenges

- Next generation approaches will require improvement of the existing annotations
  - necessary to create accurate, high resolution knowledge bases with detailed condition-, tissue-, and cell-specific functions of each gene
    - PharmGKB ….
  - these knowledge bases will allow investigators to model an organism's biology as a dynamic system, and will help predict changes in the system due to factors such as mutations or environmental changes

Annotation Challenges

- Low resolution knowledge bases
- Incomplete and inaccurate annotations
- Missing condition- and cell-specific information

Green arrows represent abundantly available information, and red arrows represent missing and/or incomplete information. The ultimate goal of pathway analysis is to analyze a biological system as a large, single network. However, the links between smaller individual pathways are not yet well known. Furthermore, the effects of a SNP on a given pathway are also missing from current knowledge bases. While some pathways are known to be related to a few diseases, it is not clear whether the changes in pathways are the cause for those diseases or the downstream effects of the diseases.



Low Resolution Knowledge Bases

- Knowledge bases not as high resolution as technologies
  - using RNA-seq, more than 90% of the human genome is estimated to be alternatively spliced
  - multiple transcripts from the same gene may have related, distinct, or even opposing functions
  - GWAS have identified a large number of SNPs that may be involved in different conditions and diseases.
  - However, current knowledge bases only specify which genes are active in a given pathway
  - Essential that they also begin specifying other information, such as transcripts that are active in a given pathway or how a given SNP affects a pathway

Low Resolution Knowledge Bases

- Because of these low resolution knowledge bases, every available pathway analysis tool first maps the input to a non-redundant namespace, typically an Entrez Gene ID
  - this type of mapping is advantageous, although it can be non-trivial, as it allows the existing pathway analysis approaches to be independent of the technology used in the experiment
  - However, mapping in this way also results in the loss of important information that may have been provided because a specific technology was used
    - XRN2a, a variant of gene XRN2, is expressed in several human tissues, whereas another variant of the same gene, XRN2b, is mainly expressed in blood leukocytes
    - Although RNA-seq can quantify expression of both variants, mapping both transcripts to a single gene causes loss of tissue-specific information, and possibly even condition-specific information

Low Resolution Knowledge Bases

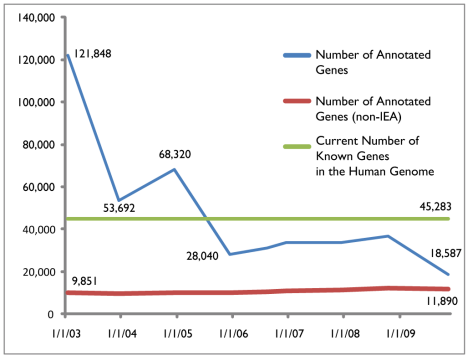
- Therefore, before pathway analysis can exploit current and future technological advances in biotechnology, it is critically important to annotate exact transcripts and SNPs that participate in a given pathway
- While new approaches are being developed in this regard, they may not yet be adequate
  - Braun et al. proposed a method for analyzing SNP data from a GWAS
  - Still relies on mapping multiple SNPs to a single gene, followed by gene-to-pathway mapping

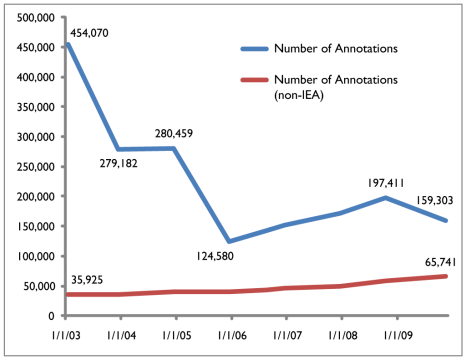
Incomplete and Inaccurate Annotation

- A surprisingly large number of genes are still not annotated
- Many of the genes are hypothetical, predicted, or pseudogenes
  - Although the number of protein-coding genes in the human genome is estimated to be between 20,000 and 25,000, according Entrez Gene, there are 45,283 human genes, of which 14,162 are pseudogenes
  - One could argue that the pseudogenes should not be included when evaluating functional annotation coverage
  - pseudogene-derived small interfering RNAs have been shown to regulate gene expression in mouse oocytes
  - GO provides annotations for 271 pseudogenes
  - A widely used DNA microarray, Affymetrix HG U133 plus 2.0, contains 1,026 probe sets that correspond to 823 pseudogenes
  - Should pseudogenes be included in the count when estimating annotation coverage for the human genome?

Incomplete and Inaccurate Annotation

Number of GO-annotated genes (left panel) and number of GO annotations (right panel) for human from January 2003 to November 2009.As the estimated number of known genes in the human genome is adjusted (between January 2003 and December 2003) and annotation practices are modified (between December 2004 and December 2005, and between October 2008 and November 2009), one can argue that, although the number of annotated genes and the annotations are decreasing (which is mainly due to the adjusted number of genes in the human genome and changes in the annotation process), the quality of annotations is improving, as demonstrated by the steady increase in non-IEA annotations and the number of genes with non-IEA annotations. However, the increase in the number of genes with non-IEA annotations is very slow. In almost 7 years, between January 2003 and November 2009, only 2,039 new genes received non-IEA annotations. At the same time, the number of non-IEA annotations increased from 35,925 to 65,741, indicating a strong research bias for a small number of genes.doi:10.1371/journal.pcbi.1002375.g003





Incomplete and Inaccurate Annotation

- Additionally, many of the existing annotations are of low quality and may be inaccurate
  - >95% of the annotations in the October 2007 release of GO had the evidence code "inferred from electronic annotations (IEA)"
  - the only ones in GO that are not curated manually
  - Annotations inferred from indirect evidence are considered to be of lower quality than those derived from direct experimental evidence
  - If the annotations with IEA code are removed, the number of genes with good quality annotations in the November 2009 release of human GO annotations is reduced from 18,587 to 11,890

Incomplete and Inaccurate Annotation

- It is very likely that the reduced number of annotations and annotated genes since January 2003 is an indicator of improving quality
- This is due in part to the fact that the number of genes in a genome are continuously being adjusted and the functional annotation algorithms are being improved
  - the number of non-IEA annotations is continuously increasing
- However, the rate of increase for non-IEA annotations is very slow (approximately 2,000 genes annotated in 7 years)

Incomplete and Inaccurate Annotation

- Manual curation of the entire genome is expected to take a very long time (~13–25 years)
- Entire research community could participate in the curation process
- One approach to facilitate participation of a large number of researchers is to adopt a standard annotation format similar to Minimum Information About a Microarray Experiment (MIAME)
  - should this be required like GEO?
- A format for functional annotation can be designed or adopted from the existing formats (e.g., BioPAX, SBML)
  - Such a format could allow researchers to specify an experimentally confirmed role of a specific transcript or a SNP in a pathway along with experimental and biological conditions

Missing Condition and cell-specific information

- Most pathway knowledge bases are built by curating experiments performed in different cell types at different time points under different conditions
- These details are typically not available in the knowledge bases!
- One effect of this omission is that multiple independent genes are annotated to participate in the same interaction in a pathway
- This effect is so widespread that many pathway knowledge bases represent a set of distinct genes as a single node in a pathway

Missing Condition and cell-specific information

- Example: *Wnt/beta-catenin pathway in STKE*
  - the node labeled "Genes" represents 19 genes directly targeted by Wnt in different organisms (Xenopus and human) in different cells and tissues (colon carcinoma cells and epithelial cells
  - these non-specific genes introduce bias for these pathways in all existing analysis approaches
  - For instance, any ORA method will assign higher significance (typically an order of magnitude lower p-value) to a pathway with more genes
  - Similarly, more genes in a pathway also increase the probability of a higher pathway-level statistic in FCS approaches, yielding higher significance for a given pathway.

Missing Condition and cell-specific information

- This contextual information is typically not available from most of the existing knowledge bases
- A standard functional annotation format discussed above would make this information available to curators and developers
  - For instance, the recently proposed Biological Connection Markup Language (BCML) allows pathway representation to specify the cell or organism in which each pathway interaction occurs.
  - BCML can generate cell-, condition-, or organism-specific pathways based on user-defined query criteria, which in turn can be used for targeted analysis

Missing Condition and cell-specific information

- Existing knowledge bases do not describe the effects of an abnormal condition on a pathway
  - For example, it is not clear how the Alzheimer's disease pathway in KEGG differs from a normal pathway
  - Nor it is clear which set of interactions leads to Alzheimer's disease
- We are now understanding that context plays an important role in pathway interactions
- Information about how cell and tissue type, age, and environmental exposures affect pathway interactions will add complexity that is currently lacking

Methodological Challenges

- Benchmark data sets for comparing different methods
- Inability to model and analyze dynamic response
- Inability to model effects of an external stimuli

Comparing Different Methods

- How do we compare different pathway analysis methods?
- Simulated data
  - Advantages:
    - Real signal is simulated, so "true" answer is known
  - Disadvantages
    - Cannot contain all the complexity of real data
    - The success of the methods can reflect the similarity of how well the simulation matches the knowledgebase structure used

Comparing Different Methods

- Benchmark data
  - Advantages:
    - Can compare sensitivity and specificity
    - Several datasets have been consistently used in the literature
    - Includes all the complexity of real biological data
  - Disadvantages
    - Affected by confounding factors
      - absence of a pure division into classes
      - presence of outliers
      - ….
    - No true answer known for grounded comparisons – actual biology isnt known

Comparing Different Methods

- A general challenge: *Different definitions of the same pathway in different knowledge bases can affect performance assessment*
  - GO defines different pathways for apoptosis in different cells
    - (e.g., cardiac muscle cell apoptosis, B cell apoptosis, T cell apoptosis)
    - Further distinguishes between induction and regulation of apoptosis
  - KEGG defines a single signaling pathway for apoptosis
    - does not distinguish between induction and regulation
  - An approach using KEGG would identify a single pathway as significant, whereas GO could identify multiple pathways, and/ or specific aspects of a single apoptosis pathway

Inability to model and analyze dynamic response

- No existing approach can collectively model and analyze high-throughput data as a single dynamic system
- Current approaches analyze a snapshot assuming that each pathway is independent of the others at a given time
  - measure expression changes at multiple time points, and analyze each time point individually
  - Implicitly assumes that pathways at different time points are independent
- Need models that accounts for dependence among pathways at different time points
  - Much of this limitation is due to technology/experimental design → not all bioinformatics limitations

Inability to model effects of an external stimuli

- Gene set–based approaches often only consider genes and their products
- Completely ignore the effects of other molecules participating in a pathway
  - such as the rate limiting step of a multi-step pathway.
- Example:
  - The amount/strength of Ca^{2+} causes different transcription factors to be activated
  - This information is usually not available.

Summary

- In the last decade, pathway analysis has matured, and become the standard for trying to dissect the biology of high throughput experiments.
- Many similarities across the three main generations of pathway analysis tools.
- Will discuss more details of some of these choices, knowledge bases, and specific approaches next.
- Many open methods development challenges!

Overview of Module

- First Half:
  - Overview of gene set and pathway analysis
    - Commonly used databases and annotation issues
    - 1st and 2nd generation tools
      - Basic differences in methods
      - Details on very popular methods
    - Issues with different "omics" datatypes
- Second Half
  - "3rd generation" methods
  - Network analysis modeling

Questions?

motsinger@stat.ncsu.edu

Pathway and Gene Set Analysis Part 1

Alison Motsinger-Reif, PhD
Bioinformatics Research Center
Department of Statistics
North Carolina State University
motsinger@stat.ncsu.edu

The early steps of a microarray study

- Scientific Question (biological)
- Study design (biological/statistical)
- Conducting Experiment (biological)
- Preprocessing/Normalizing Data (statistical)
- Finding differentially expressed genes (statistical)

A data example

- Lee et al (2005) compared adipose tissue (abdominal subcutaenous adipocytes) between obese and lean Pima Indians
- Samples were hybridised on HGu95e-Affymetrix arrays (12639 genes/probe sets)
- Available as GDS1498 on the GEO database
- We selected the male samples only
  - 10 obese vs 9 lean

Diabetologia (2005) 48: 1776–1783
DOI 10.1007/s00125-005-1867-3

ARTICLE

Y. H. Lee · S. Nair · E. Rousseau · D. B. Allison · G. P. Page · P. A. Tataranni · C. Bogardus · P. A. Permana

Microarray profiling of isolated abdominal subcutaneous adipocytes from obese vs non-obese Pima Indians: increased expression of inflammation-related genes

Received: 10 December 2004 / Accepted: 28 April 2005 / Published online: 30 July 2005

Abstract *Aims/hypothesis:* Obesity increases the risk of developing major diseases such as diabetes and cardiovascular disease. Adipose tissue, particularly adipocytes, may play a major role in the development of obesity and its comorbidities. The aim of this study was to characterise, in adipocytes from obese people, the most differentially expressed genes that might be relevant to the development of obesity. *Methods:* We carried out microarray gene profiling of isolated abdominal subcutaneous adipocytes from 20 non-obese (BMI 25±3 kg/m^2) and 19 obese (BMI 55±8 kg/m^2) non-diabetic Pima Indians using Affymetrix HG-U95 GeneChip arrays. After data analyses, we measured the transcript levels of selected genes based on their biological functions and chromosomal positions using quantitative real-time PCR. *Results:* The most differentially expressed genes in adipocytes of obese individuals consisted of 433 upregulated and 244 downregulated genes. Of these, 410 genes could be classified into 20 functional Gene Ontology categories. The analyses indicated that the inflammation/immune response category was over-represented, and that most inflammation-related genes were upregulated in adipocytes of obese subjects. Quantitative real-time PCR confirmed the transcriptional upregulation of representative inflammation-related genes (*CCL2* and *CCL3*) encoding the chemokines monocyte chemoattractant protein-1 and macrophage inflammatory protein 1α. The differential expression levels of eight positional candidate genes, including inflammation-related *THY1* and *C1QTNF5*, were also confirmed. These genes are located on chromosome 11q22–q24, a region with linkage to obesity in the Pima Indians. *Conclusions/interpretation:* This study provides evidence supporting the active role of mature adipocytes in obesity-related inflammation. It also provides potential candidate genes for susceptibility to obesity.

Electronic supplementary material Supplementary material is available in the online version of this article at http://dx.doi.org/10.1007/s00125-005-1867-3.

The "Result"

Probe Set ID	log.ratio	pvalue	adj.p
73554_at	1.4971	0.0000	0.0004
91279_at	0.8667	0.0000	0.0017
74099_at	1.0787	0.0000	0.0104
83118_at	-1.2142	0.0000	0.0139
81647_at	1.0362	0.0000	0.0139
84412_at	1.3124	0.0000	0.0222
90585_at	1.9859	0.0000	0.0258
84618_at	-1.6713	0.0000	0.0258
91790_at	1.7293	0.0000	0.0350
80755_at	1.5238	0.0000	0.0351
85539_at	0.9303	0.0000	0.0351
90749_at	1.7093	0.0000	0.0351
74038_at	-1.6451	0.0000	0.0351
79299_at	1.7156	0.0000	0.0351
72962_at	2.1059	0.0000	0.0351
88719_at	-3.1829	0.0000	0.0351
72943_at	-2.0520	0.0000	0.0351
91797_at	1.4676	0.0000	0.0351
78356_at	2.1140	0.0001	0.0359
90268_at	1.6552	0.0001	0.0421

What happened to the Biology???

Slightly more informative results

Probe Set ID	Gene Symbo	Gene Title	go biological process term	go molecular function term	log.ratio	pvalue	adj.p
73554_at	CCDC80	coiled-coil domain contair	---	---	1.4971	0.0000	0.0004
91279_at	C1QTNF5 ///	C1q and tumor necrosis fi	visual perception /// embr	---	0.8667	0.0000	0.0017
74099_at	---	---	---	---	1.0787	0.0000	0.0104
83118_at	RNF125	ring finger protein 125	immune response /// mod	protein binding /// zinc ion	-1.2142	0.0000	0.0139
81647_at	---	---	---	---	1.0362	0.0000	0.0139
84412_at	SYNPO2	synaptopodin 2	---	actin binding /// protein bi	1.3124	0.0000	0.0222
90585_at	C15orf59	chromosome 15 open rea	---	---	1.9859	0.0000	0.0258
84618_at	C12orf39	chromosome 12 open rea	---	---	-1.6713	0.0000	0.0258
91790_at	MYEOV	myeloma overexpressed	---	---	1.7293	0.0000	0.0350
80755_at	MYOF	myoferlin	muscle contraction /// blo	protein binding	1.5238	0.0000	0.0351
85539_at	PLEKHH1	pleckstrin homology dom	---	binding	0.9303	0.0000	0.0351
90749_at	SERPINB9	serpin peptidase inhibitor,	anti-apoptosis /// signal tr	endopeptidase inhibitor a	1.7093	0.0000	0.0351
74038_at	---	---	---	---	-1.6451	0.0000	0.0351
79299_at	---	---	---	---	1.7156	0.0000	0.0351
72962_at	BCAT1	branched chain aminotrar	G1/S transition of mitotic	catalytic activity /// branch	2.1059	0.0000	0.0351
88719_at	C12orf39	chromosome 12 open rea	---	---	-3.1829	0.0000	0.0351
72943_at	---	---	---	---	-2.0520	0.0000	0.0351
91797_at	LRRC16A	leucine rich repeat contair	---	---	1.4676	0.0000	0.0351
78356_at	TRDN	triadin	muscle contraction	receptor binding	2.1140	0.0001	0.0359
90268_at	C5orf23	chromosome 5 open read	---	---	1.6552	0.0001	0.0421

If we are lucky, some of the top genes mean something to us

But what if they don't?

And how what are the results for other genes with similar biological functions

How to incorporate biological knowledge

- The type of knowledge we deal with is rather simple:

  We know groups/sets of genes that for example
  - Belong to the same pathway
  - Have a similar function
  - Are located on the same chromosome, etc…

- We will assume these groupings to be given, i.e. we will not yet discuss methods used to detect pathways, networks, gene clusters
  - We will later!

What is a pathway?

- No clear definition
  - Wikipedia: "In biochemistry, **metabolic pathways** are series of chemical reactions occurring within a cell. In each pathway, a principal chemical is modified by chemical reactions."
  - These pathways describe enzymes and metabolites
- But often the word "pathway" is also used to describe gene regulatory networks or protein interaction networks
- In all cases a pathway describes a biological function very specifically

What is a Gene Set?

- Just what it says: a set of genes!
  - All genes involved in a pathway are an example of a Gene Set
  - All genes corresponding to a Gene Ontology term are a Gene Set
  - All genes mentioned in a paper of Smith et al might form a Gene Set
- A Gene Set is a much more general and less specific concept than a pathway
- Still: we will sometimes use two words interchangeably, as the analysis methods are mainly the same

Where Do Gene Sets/Lists Come From?

- Molecular profiling e.g. mRNA, protein
  - Identification → Gene list
  - Quantification → Gene list + values
  - Ranking, Clustering (biostatistics)
- Interactions: Protein interactions, Transcription factor binding sites (ChIP)
- Genetic screen e.g. of knock out library
- Association studies (Genome-wide)
  - Single nucleotide polymorphisms (SNPs)
  - Copy number variants (CNVs)
  - ……..

What is Gene Set/Pathway analysis?

- The aim is to give one number (score, p-value) to a Gene Set/Pathway
  - Are many genes in the pathway differentially expressed (up-regulated/downregulated)
  - Can we give a number (p-value) to the probability of observing these changes just by chance?

Goals

- Pathway and gene set data resources
  - Gene attributes
  - Database resources
    - GO, KeGG, Wikipathways, MsigDB
  - Gene identifiers and issues with mapping
- Differences between pathway analysis tools
  - Self contained vs. competitive tests
  - Cut-off methods vs. global methods
  - Issues with multiple testing

Goals

- Pathway and gene set data resources
  - Gene attributes
  - Database resources
    - GO, KeGG, Wikipathways, MsigDB
  - Gene identifiers and issues with mapping

- Differences between pathway analysis tools
  - Self contained vs. competitive tests
  - Cut-off methods vs. global methods
  - Issues with multiple testing

Gene Attributes

- Functional annotation
  - Biological process, molecular function, cell location
- Chromosome position
- Disease association
- DNA properties
  - TF binding sites, gene structure (intron/exon), SNPs
- Transcript properties
  - Splicing, 3' UTR, microRNA binding sites
- Protein properties
  - Domains, secondary and tertiary structure, PTM sites
- Interactions with other genes

Gene Attributes

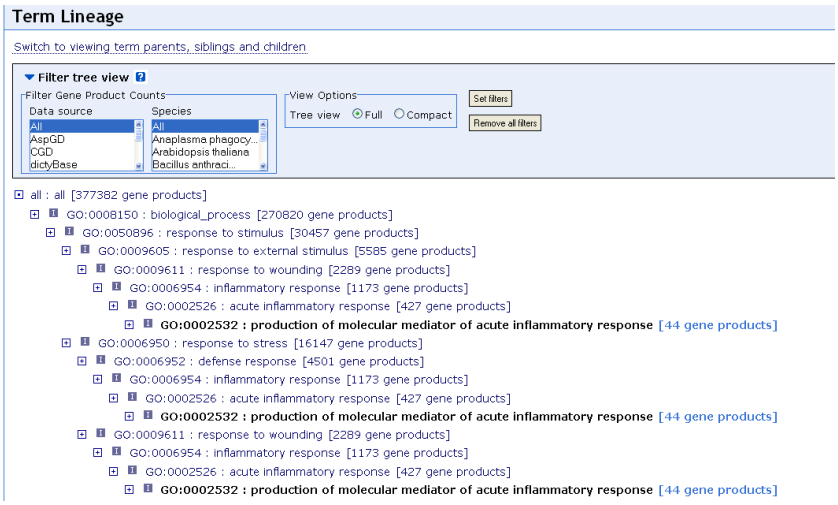
- Functional annotation
  - Biological process, molecular function, cell location
- Chromosome position
- Disease association
- DNA properties
  - TF binding sites, gene structure (intron/exon), SNPs
- Transcript properties
  - Splicing, 3' UTR, microRNA binding sites
- Protein properties
  - Domains, secondary and tertiary structure, PTM sites
- Interactions with other genes

Database Resources

- Use functional annotation to aggregate genes into pathways/gene sets
- A number of databases are available
  - Different analysis tools link to different databases
  - Too many databases to go into detail on every one
  - Commonly used resources:
    - GO
    - KeGG
    - MsigDB
    - WikiPathways

Pathway and Gene Set data resources

- The Gene Ontology (GO) database
  - http://www.geneontology.org/
  - GO offers a relational/hierarchical database
  - Parent nodes: more general terms
  - Child nodes: more specific terms
  - At the end of the hierarchy there are genes/proteins
  - At the top there are 3 parent nodes: biological process, molecular function and cellular component
- Example: we search the database for the term "inflammation"



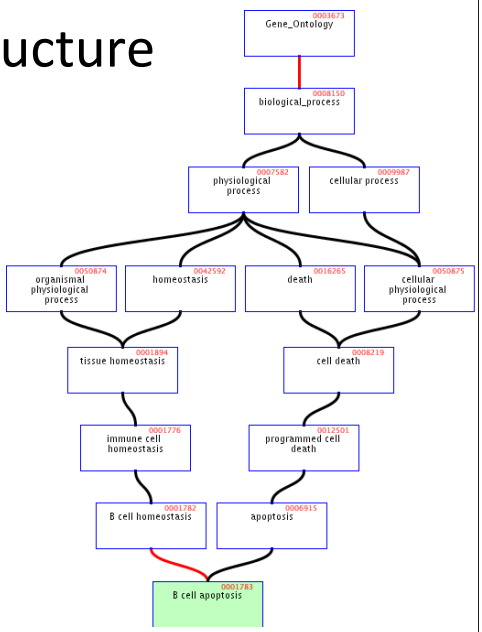
The genes on our array that code for one of the 44 gene products would form the corresponding "inflammation" gene set

What is the Gene Ontology (GO)?

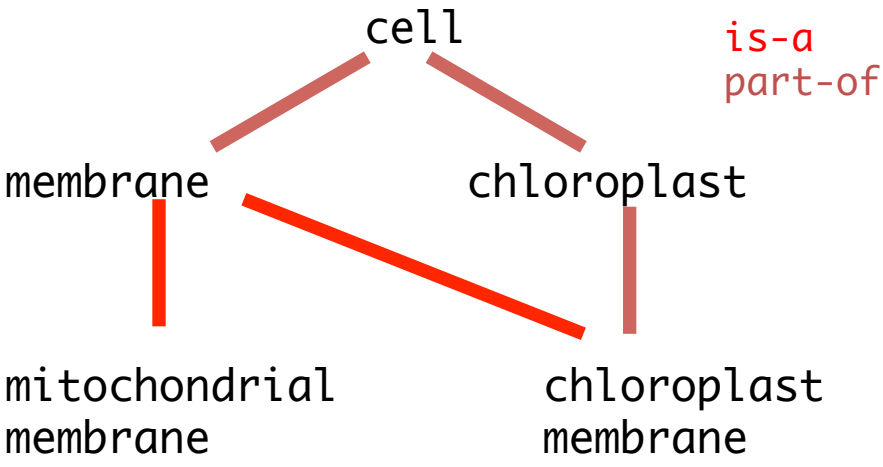
- Set of biological phrases (terms) which are applied to genes:
  - protein kinase
  - apoptosis
  - membrane
- Ontology: A formal system for describing knowledge

GO Structure

- Terms are related within a hierarchy
  - is-a
  - part-of
- Describes multiple levels of detail of gene function
- Terms can have more than one parent or child



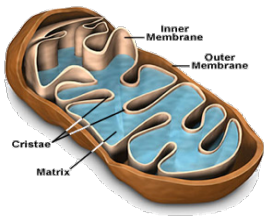
GO Structure

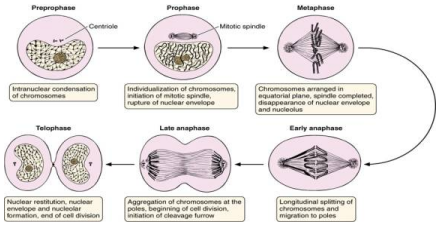


Species independent. Some lower-level terms are specific to a group, but higher level terms are not

What GO Covers?

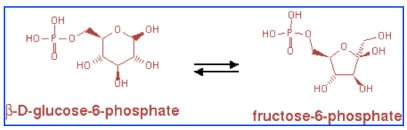
- GO terms divided into three aspects:
  - cellular component
  - molecular function
  - biological process





Cell division

glucose-6-phosphate isomerase activity



Terms

- Where do GO terms come from?
  - GO terms are added by editors at EBI and gene annotation database groups
  - Terms added by request
  - Experts help with major development
  - 27734 terms, 98.9% with definitions.
    - 16731 biological_process
    - 2385 cellular_component
    - 8618 molecular_function

Annotations

- Genes are linked, or associated, with GO terms by trained curators at genome databases
  - Known as 'gene associations' or GO annotations
  - Multiple annotations per gene
- Some GO annotations created automatically

Annotation Sources

- Manual annotation
  - Created by scientific curators
    - High quality
    - Small number (time-consuming to create)
- Electronic annotation
  - Annotation derived without human validation
    - Computational predictions (accuracy varies)
    - Lower 'quality' than manual codes
- Key point: be aware of annotation origin

Evidence Types

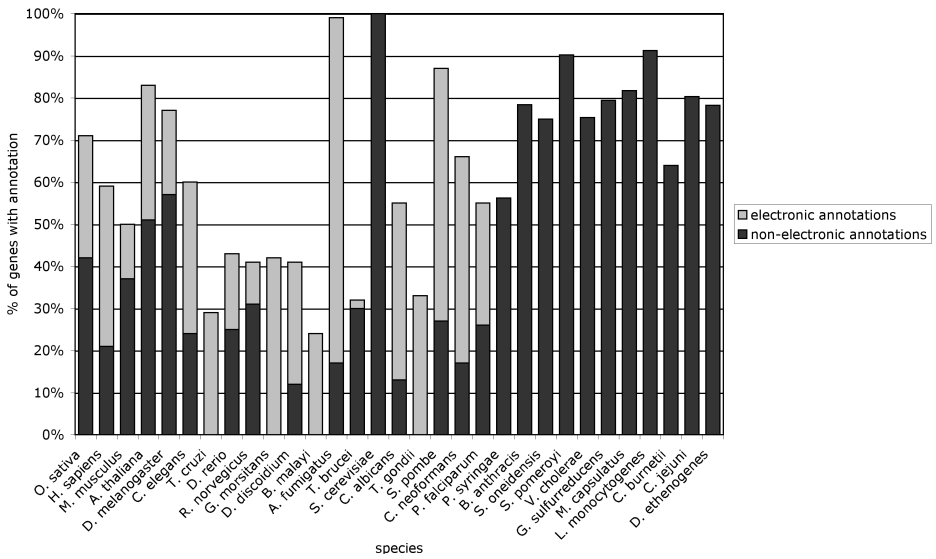
- **ISS**: Inferred from Sequence/Structural Similarity
- **IDA**: Inferred from Direct Assay
- **IPI**: Inferred from Physical Interaction
- **IMP**: Inferred from Mutant Phenotype
- **IGI**: Inferred from Genetic Interaction
- **IEP**: Inferred from Expression Pattern
- **TAS**: Traceable Author Statement
- **NAS**: Non-traceable Author Statement
- **IC**: Inferred by Curator
- **ND**: No Data available

- **IEA**: Inferred from electronic annotation

Species Coverage

- All major eukaryotic model organism species
- Human via GOA group at UniProt
- Several bacterial and parasite species through TIGR and GeneDB at Sanger
- New species annotations in development

Variable Coverage



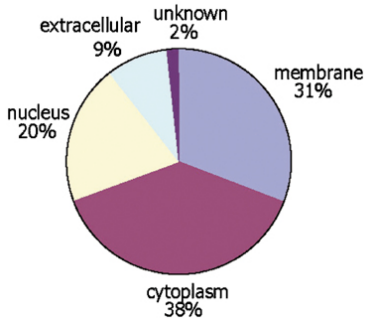
Lomax J. Get ready to GO! A biologist's guide to the Gene Ontology. Brief Bioinform. 2005 Sep;6(3):298-304.

Contributing Databases

- Berkeley *Drosophila* Genome Project (BDGP)
- dictyBase (*Dictyostelium discoideum)*
- FlyBase (*Drosophila melanogaster)*
- GeneDB (*Schizosaccharomyces pombe, Plasmodium falciparum*, *Leishmania major* and *Trypanosoma brucei)*
- UniProt Knowledgebase (Swiss-Prot/TrEMBL/PIR-PSD) and InterPro databases
- Gramene (grains, including rice, *Oryza*)
- Mouse Genome Database (MGD) and Gene Expression Database (GXD) (*Mus musculus)*
- Rat Genome Database (RGD) (*Rattus norvegicus)*
- Reactome
- *Saccharomyces* Genome Database (SGD) (*Saccharomyces cerevisiae)*
- The *Arabidopsis* Information Resource (TAIR) (*Arabidopsis thaliana)*
- The Institute for Genomic Research (TIGR): databases on several bacterial species
- WormBase (*Caenorhabditis elegans)*
- Zebrafish Information Network (ZFIN): (*Danio rerio)*

GO Slim Sets

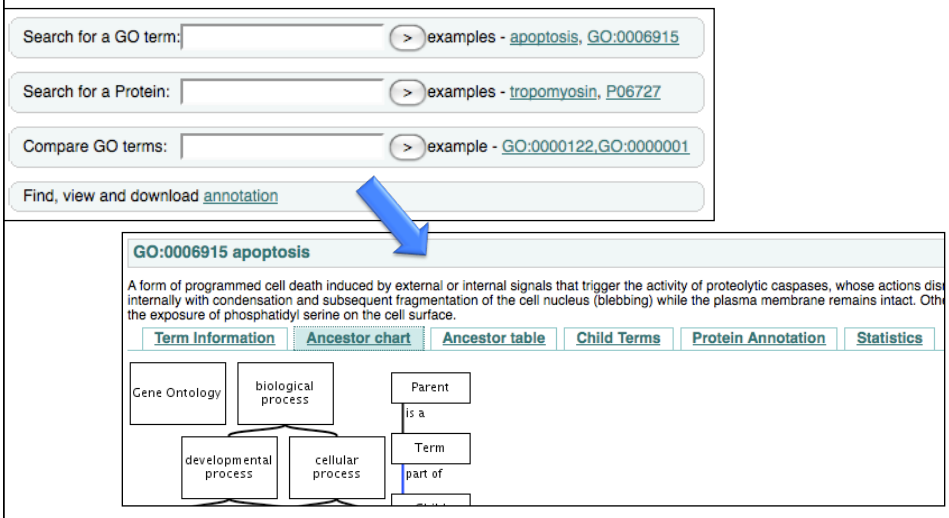
- GO has too many terms for some uses
  - Summaries (e.g. Pie charts)
- GO Slim is an official reduced set of GO terms
  - Generic, plant, yeast



GO Software Tools

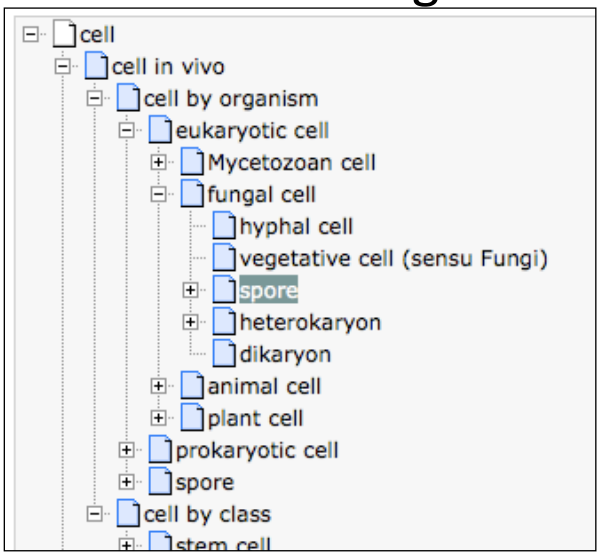
- GO resources are freely available to anyone without restriction
  - Includes the ontologies, gene associations and tools developed by GO
- Other groups have used GO to create tools for many purposes
  - http://www.geneontology.org/GO.tools

Accessing GO: QuickGO



http://www.ebi.ac.uk/ego/

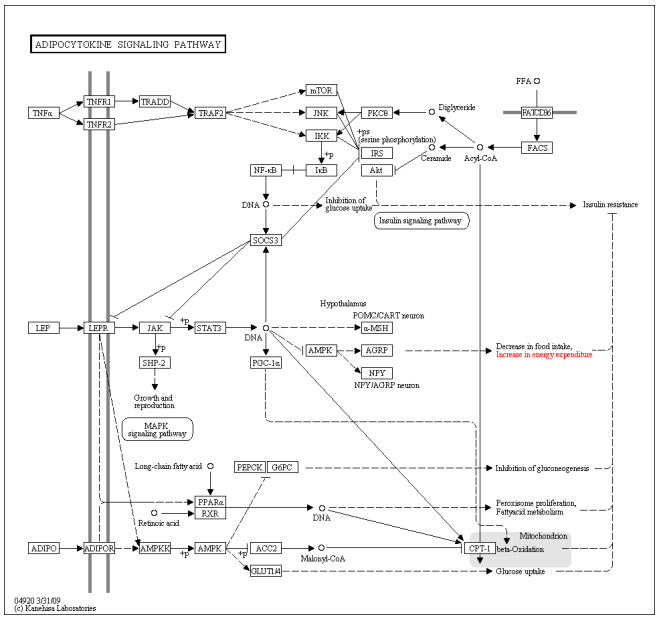
Other Ontologies



http://www.ebi.ac.uk/ontology-lookup

KEGG pathway database

- KEGG = Kyoto Encyclopedia of Genes and Genomes
  - http://www.genome.jp/kegg/pathway.html
  - The pathway database gives far more detailed information than GO
    - Relationships between genes and gene products
  - But: this detailed information is only available for selected organisms and processes
  - Example: Adipocytokine signaling pathway

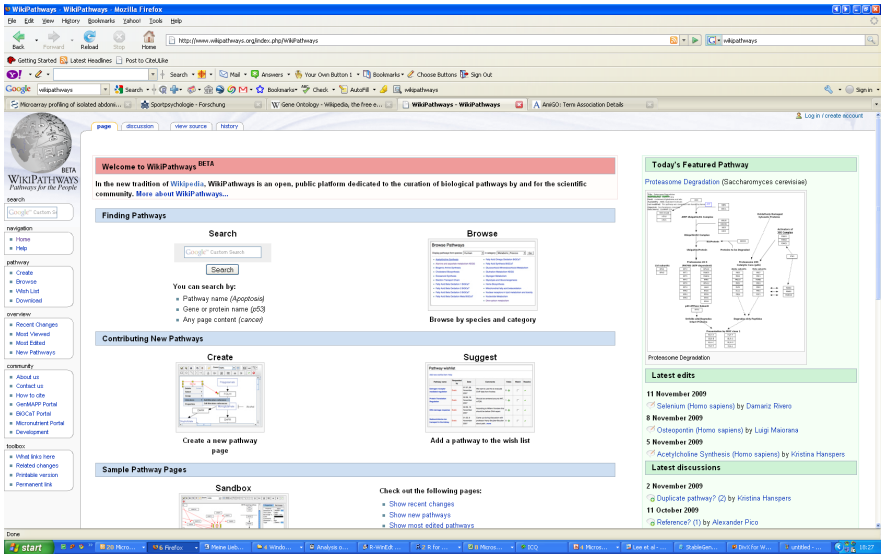


KEGG pathway database

- Clicking on the nodes in the pathway leads to more information on genes/proteins
  - Other pathways the node is involved with
  - Entries in Gene/Protein databases
  - References
  - Sequence information
- Ultimately this allows to find corresponding genes on the microarray and define a Gene Set for the pathway

Wikipathways

- http://www.wikipathways.org
- A wikipedia for pathways
  - One can see and download pathways
  - But also edit and contribute pathways
- The project is linked to the GenMAPP and Pathvisio analysis/visualisation tools

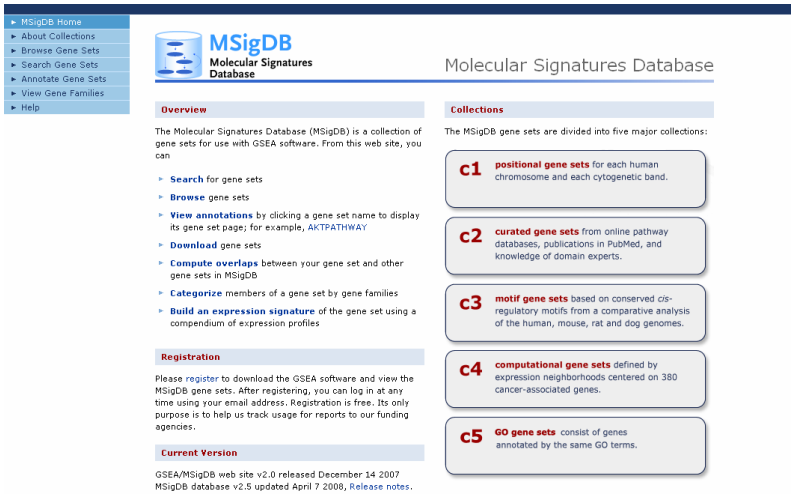


MSigDB

- MSigDB = Molecular Signature Database

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

- Related to the the analysis program GSEA
- MSigDB offers gene sets based on various groupings
  - Pathways
  - GO terms
  - Chromosomal position,…



Some Warnings

- In many cases the definition of a pathway/gene set in a database might differ from that of a scientist
- The nodes in pathways are often proteins or metabolites; the activity of the corresponding gene set is not necessarily a good measurement of the activity of the pathway
- There are many more resources out there (BioCarta, BioPax)
- Commercial packages often use their own pathway/gene set definitions (Ingenuity, Metacore, Genomatix,…)
- Genes in a gene set are usually not given by a Probe Set ID, but refer to some gene data base (Entrez IDs, Unigene IDs)
    - Conversion can lead to errors!

Some Warnings

- In many cases the definition of a pathway/gene set in a database might differ from that of a scientist
- The nodes in pathways are often proteins or metabolites; the activity of the corresponding gene set is not necessarily a good measurement of the activity of the pathway
- There are many more resources out there (BioCarta, BioPax)
- Commercial packages often use their own pathway/gene set definitions (Ingenuity, Metacore, Genomatix,…)
- Genes in a gene set are usually not given by a Probe Set ID, but refer to some gene data base (Entrez IDs, Unigene IDs)
    - Conversion can lead to errors!

Gene Attributes

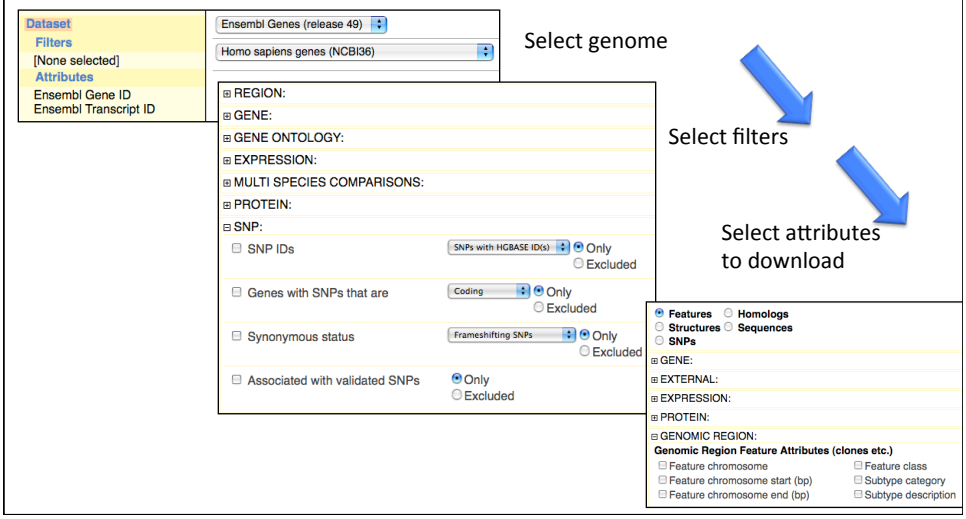
- Functional annotation
  - Biological process, molecular function, cell location
- Chromosome position
- Disease association
- DNA properties
  - TF binding sites, gene structure (intron/exon), SNPs
- Transcript properties
  - Splicing, 3' UTR, microRNA binding sites
- Protein properties
  - Domains, secondary and tertiary structure, PTM sites
- Interactions with other genes

Sources of Gene Attributes

- Ensembl BioMart (eukaryotes)
  - http://www.ensembl.org
- Entrez Gene (general)
  - http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene
- Model organism databases
  - E.g. SGD: http://www.yeastgenome.org/
- Many others…..

Ensembl BioMart

- Convenient access to gene list annotation



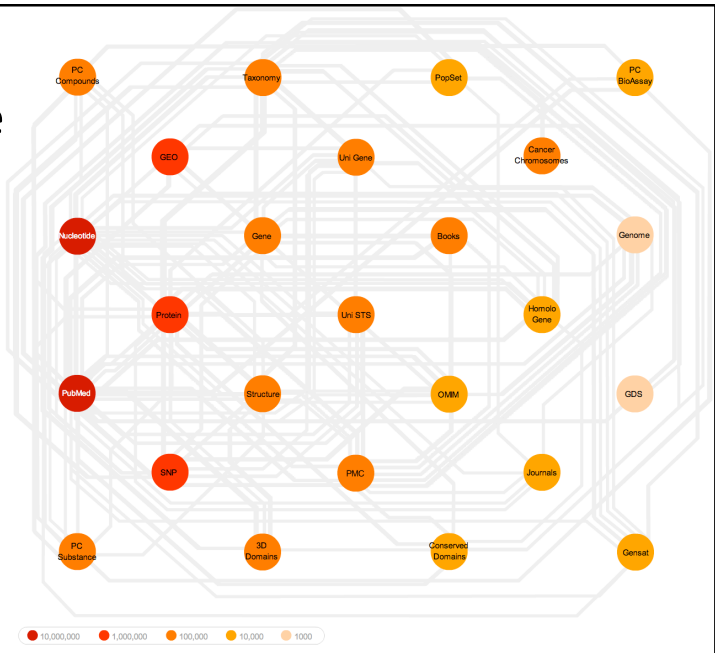
Gene and Protein Identifiers

- Identifiers (IDs) are ideally unique, stable names or numbers that help track database records
  - E.g. Social Insurance Number, Entrez Gene ID 41232
- Gene and protein information stored in many databases
  - → Genes have many IDs
- Records for: Gene, DNA, RNA, Protein
  - Important to recognize the correct record type
  - E.g. Entrez Gene records don't store sequence. They link to DNA regions, RNA transcripts and proteins.

NCBI Database Links

NCBI:
U.S. National Center for Biotechnology Information

Part of National Library of Medicine (NLM)



http://www.ncbi.nlm.nih.gov/Database/datamodel/data_nodes.swf

Common Identifiers

Gene
Ensembl ENSG00000139618
Entrez Gene 675
Unigene Hs.34012

RNA transcript
GenBank BC026160.1
RefSeq NM_000059
Ensembl ENST00000380152

Protein
Ensembl ENSP00000369497
RefSeq NP_000050.2
UniProt BRCA2_HUMAN or A1YBP1_HUMAN
IPI IPI00412408.1
EMBL AF309413
PDB 1MIU

Species-specific
HUGO HGNC BRCA2
MGI MGI:109337
RGD 2219
ZFIN ZDB-GENE-060510-3
FlyBase CG9097
WormBase WBGene00002299 or ZK1067.1
SGD S000002187 or YDL029W

Annotations
InterPro IPR015252
OMIM 600185
Pfam PF09104
Gene Ontology GO:0000724
SNPs rs28897757

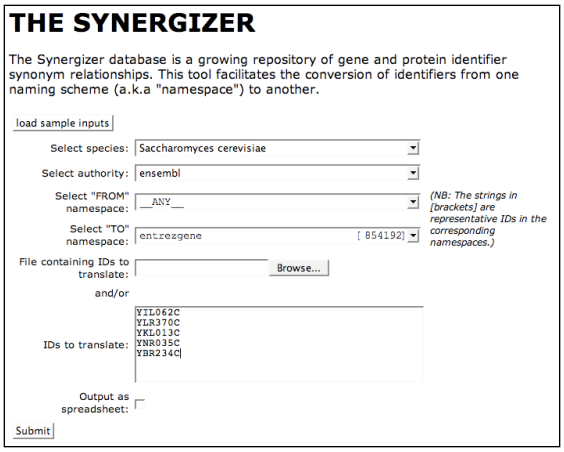
Experimental Platform
Affymetrix 208368_3p_s_at
Agilent A_23_P99452
CodeLink GE60169
Illumina GI_4502450-S

Red = Recommended

Identifier Mapping

- So many IDs!
  - Mapping (conversion) is a headache
- Four main uses
  - Searching for a favorite gene name
  - Link to related resources
  - Identifier translation
    - E.g. Genes to proteins, Entrez Gene to Affy
  - Unification during dataset merging
    - Equivalent records

ID Mapping Services



*	entrezgene
YIL062C	854748
YLR370C	851085
YKL013C	853856
YNR035C	855771
YBR234C	852536

- Synergizer
  - http://llama.med.harvard.edu/synergizer/translate/
- Ensembl BioMart
  - http://www.ensembl.org
- UniProt
  - http://www.uniprot.org/

ID Mapping Challenges

- Avoid errors: map IDs correctly
- Gene name ambiguity – not a good ID
  - e.g. FLJ92943, LFS1, TRP53, p53
  - Better to use the standard gene symbol: TP53
- Excel error-introduction
  - OCT4 is changed to October-4
- Problems reaching 100% coverage
  - E.g. due to version issues
  - Use multiple sources to increase coverage

Zeeberg BR et al. Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics BMC Bioinformatics. 2004 Jun 23;5:80

Goals

- Pathway and gene set data resources
  - Gene attributes
  - Database resources
    - GO, KeGG, Wikipathways, MsigDB
  - Gene identifiers and issues with mapping
- Differences between pathway analysis tools
  - Self contained vs. competitive tests
  - Cut-off methods vs. global methods
  - Issues with multiple testing

Goals

- Pathway and gene set data resources
  - Gene attributes
  - Database resources
    - GO, KeGG, Wikipathways, MsigDB
  - Gene identifiers and issues with mapping

- Differences between pathway analysis tools
  - Self contained vs. competitive tests
  - Cut-off methods vs. global methods
  - Issues with multiple testing

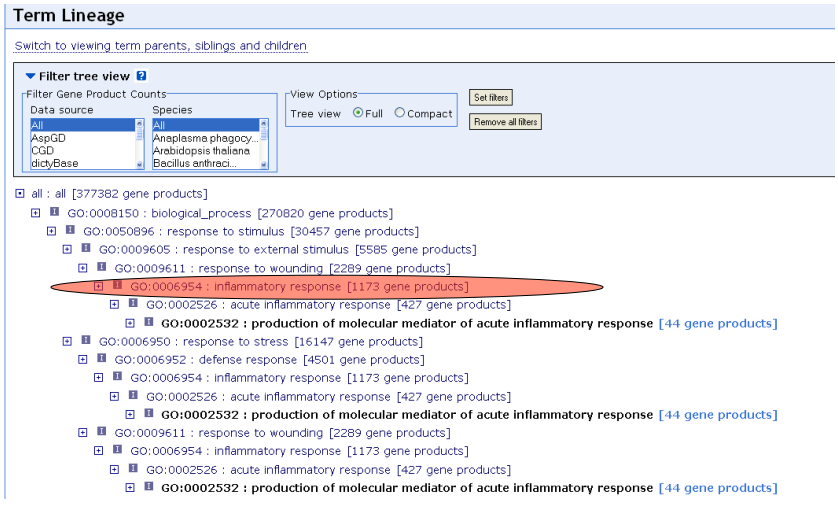
Aims of Analysis

- Reminder: The aim is to give one number (score, p-value) to a Gene Set/Pathway
  - Are many genes in the pathway differentially expressed (up-regulated/downregulated)?
  - Can we give a number (p-value) to the probability of observing these changes just by chance?
  - Similar to single gene analysis statistical hypothesis testing plays an important role

General differences between analysis tools

- Self contained vs competitive test
  - The distinction between “self-contained” and “competitive” methods goes back to Goeman and Buehlman (2007)
  - A self-contained method only uses the values for the genes of a gene set
    - The null hypothesis here is: H = {“No genes in the Gene Set are differentially expressed”}
  - A competitive method compares the genes within the gene set with the other genes on the arrays
    - Here we test against H: {“The genes in the Gene Set are not more differentially expressed than other genes”}

Example: Analysis for the GO-Term “inflammatory response” (GO:0006954)



Back to the Real Data Example

- Using Bioconductor software we can find 96 probesets on the array corresponding to this term
- 8 out of these have a p-value < 5%
- How many significant genes would we expect by chance?
- Depends on how we define "by chance"

The "self-contained" version

- By chance (i.e. if it is NOT differentially expressed) a gene should be significant with a probability of 5%
- We would expect 96 x 5% = 4.8 significant genes
- Using the binomial distribution we can calculate the probability of observing 8 or more significant genes as p = 10.8%, i.e. not quite significant

The "competitive" version

- Overall 1272 out of 12639 genes are significant in this data set (10.1%)
- If we randomly pick 96 genes we would expect 96 x 10.1% = 9.7 genes to be significant "by chance"
- A p-value can be calculated based on the 2x2 table
- Tests for association: Chi-Square-Test or Fisher's exact test

	In GS	Not in GS
sig	8	1264
non-sig	88	11 279

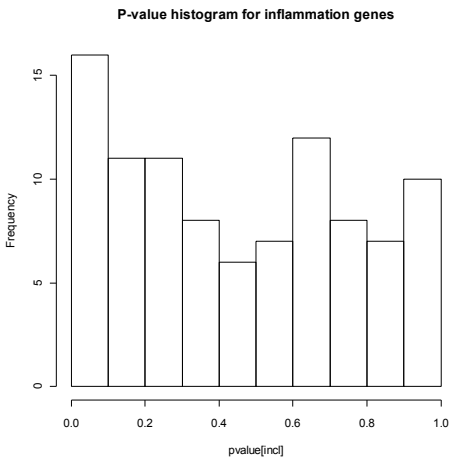
P-value from Fisher's exact test (one-sided): 73.3%, i.e very far from being significant

Competitive Tests

- Competitive results depend highly on how many genes are on the array and previous filtering
  - On a small targeted array where all genes are changed, a competitive method might detect no differential Gene Sets at all
- Competitive tests can also be used with small sample sizes, even for n=1
  - BUT: The result gives no indication of whether it holds for a wider population of subjects, the p-value concerns a population of genes!
- Competitive tests typically give less significant results than self-contained (as seen with the example)
- Fisher's exact test (competitive) is probably the most widely used method!

Cut-off methods vs whole gene list methods

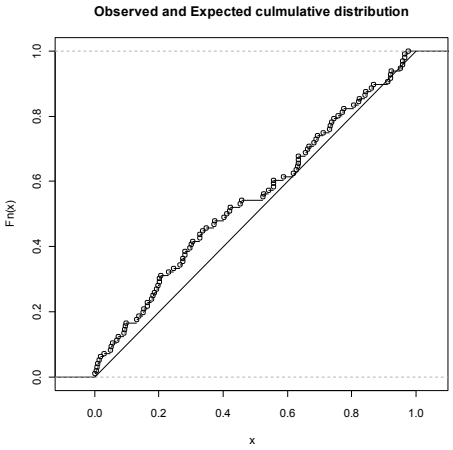
- A problem with both tests discussed so far is, that they rely on an arbitrary cut-off
- If we call a gene significant for 10% p-value threshold the results will change
  - In our example the binomial test yields p= 2.2%, i.e. for this cut-off the result is significant!
- We also lose information by reducing a p-value to a binary ("significant", "non-significant") variable
  - It should make a difference, whether the non-significant genes in the set are nearly significant or completely unsignificant

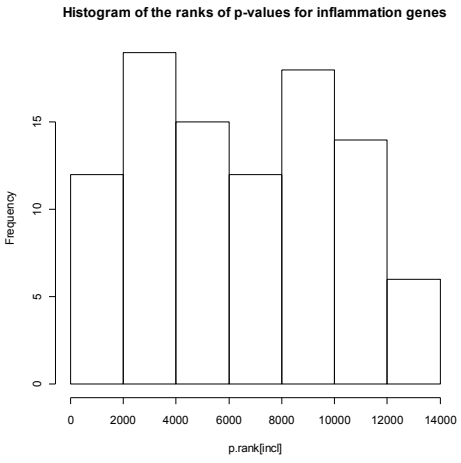


- We can study the distribution of the p-values in the gene set
- If no genes are differentially expressed this should be a uniform distribution
- A peak on the left indicates, that some genes are differentially expressed
- We can test this for example by using the Kolmogorov-Smirnov-Test
- Here p = 8.2%, i.e. not quite significant
- This would be a "self-contained" test, as only the genes in the gene set are being used

Kolmogorov-Smirnov Test

- The KS-test compares an observed with an expected cumulative distribution
- The KS-statistic is given by the maximum deviation between the two





• Alternatively we could look at the distribution of the RANKS of the p-values in our gene set

• This would be a competitive method, i.e we compare our gene set with the other genes

• Again one can use the Kolmogorov-Smirnov test to test for uniformity

• Here: p= 85.1%, i.e. very far from significance

Other general issues

- Direction of change
  - In our example we didn't differentiate between up or down-regulated genes
  - That can be achieved by repeating the analysis for p-values from one-sided test
    - Eg. we could find GO-Terms that are significantly up-regulated
  - With most software both approaches are possible
- Multiple Testing
  - As we are testing many Gene Sets, we expect some significant findings "by chance" (false positives)
  - Controlling the false discovery rate is tricky: The gene sets do overlap, so they will not be independent!
    - Even more tricky in GO analysis where certain GO terms are subset of others
  - The Bonferroni-Method is most conservative, but always works!

Multiple Testing for Pathways

- Resampling strategies (dependence between genes)
  - The methods we used so far in our example assume that genes are independent of each other…if this is violated the p-values are incorrect
  - Resampling of group/phenotype labels can correct for this
  - We give an example for our data set

Example Resampling Approach

1. Calculate the test statistic, e.g. the percentage of significant genes in the Gene Set
2. Randomly re-shuffle the group labels (lean, obese) between the samples
3. Repeat the analysis for the re-shuffled data set and calculate a re-shuffled version of the test statistic
4. Repeat 2 and 3 many times (thousands…)
5. We obtain a distribution of re-shuffled % of significant genes: the percentage of re-shuffled values that are larger than the one observed in 1 is our p-value

Resampling Approach

- The reshuffling takes gene to gene correlations into account
- Many programs also offer to resample the genes: This does NOT take correlations into account
- Roughly speaking:
  - Resampling phenotypes: corresponds to self-contained test
  - Resampling genes: corresponds to competitive test

Resampling Approaches

- Genes being present more than once
  - Common approaches
    - Combine duplicates (average, median, maximum,…)
    - Ignore (i.e treat duplicates like different genes)
- Using summary statistics vs using all data
  - Our examples used p-values as data summaries
  - Other approaches use fold-changes, signal to noise ratios, etc…
  - Some methods are based on the original data for the genes in the gene set rather than on a summary statistic

Resampling Approaches

- The resampling approaches are highly computationally intensive
- New methods are being developed to speed this up
  - Empirical approximations of permutations
  - Empirical pathway analysis, without permutation.
    - Zhou YH, Barry WT, Wright FA.Biostatistics. 2013 Jul; 14(3):573-85. doi: 10.1093/biostatistics/kxt004. Epub 2013 Feb 20.

Summary

- Databases
- Choice makes a difference
- Not all use the same IDs – watch out ☺
- Major differences between methods
- Issues with multiple testing

- Next lecture, will go into more detail on a few methods

Questions?

Pathway and Gene Set Analysis Part 2

Alison Motsinger-Reif, PhD
Bioinformatics Research Center
Department of Statistics
North Carolina State University
motsinger@stat.ncsu.edu

Goals

Some methods in more detail

- TopGO
- Global Ancova
- Pathvisio/Genmapp
- Impact Factor Analysis
- GSEA

Some methods in detail

- There are far too many methods to give a comprehensive overview

BRIEFINGS IN BIOINFORMATICS. VOL 9. NO 3. 189–197
Advance Access publication January 17, 2008
doi:10.1093/bib/bbn001

Gene-set approach for expression pattern analysis

Dougu Nam and Seon-Young Kim

Submitted: 7th November 2007; Received (in revised form): 28th December 2007

Abstract

Recently developed gene set analysis methods evaluate differential expression patterns of gene groups instead of those of individual genes. This approach especially targets gene groups whose constituents show subtle but coordinated expression changes, which might not be detected by the usual individual gene analysis. The approach has been quite successful in deriving new information from expression data, and a number of methods and tools have been developed intensively in recent years. We review those methods and currently available tools, classify them according to the statistical methods employed, and discuss their pros and cons. We also discuss several interesting extensions to the methods.

Keywords: *gene set analysis; DNA microarray; differential expression of genes*

Table of methods (from Nam & Kim)

Table 1: Cutoff-free gene set analysis methods

Authors	Year	Name	Statistical test	Self-contained versus competitive	Gene versus ample randomization	Reference
Virtaneva *et al.*	2001		sample randomization	self-contained	sample	[8]
Pavlidis *et al.*	2002		gene randomization	competitive	gene	[9]
Mootha *et al.*	2003	GSEA	sample randomization	mixed	sample	[7]
Breslin *et al.*	2004	Catmap	gene randomization	competitive	gene	[3]
Goeman *et al.*	2004	globaltest	sample randomization	self-contained	sample	[17]
Smid *et al.*	2004	GO-Mapper	z-test	competitive	gene	[38]
Volinia *et al.*	2004	GOAL	gene randomization	competitive	gene	[39]
Barry *et al.*	2005	SAFE	sample randomization	competitive	sample	[19]
Beh-Shaul *et al.*	2005		Kolmogorov–Smirnov test	competitive	gene	[5]
Boorsma e al.	2005	T-profiler	*t*-test	competitive	gene	[15]
Kim *et al.*	2005	PAGE	z-test	competitive	gene	[14]
Lee *et al.*	2005	ErmineJ	sample randomization	competitive	gene	[16]
Subramanian *et al.*	2005	GSEA	sample randomization	mixed	gene	[25]
Tian *et al.*	2005	Q1, Q2	gene or sample randomization	competitive or self-contained	gene or sample	[10]
Tomfohr *et al.*	2005	PLAGE	sample randomization	self-contained	sample	[20]
Edelman *et al.*	2006	ASSESS	sample randomization	competitive	sample	[28]
Kong *et al.*	2006		Hotelling's T squared	self-contained	sample	[21]
Nam *et al.*	2006	ADGO	z-test	competitive	gene	[29]
Saxena *et al.*	2006	AE	sample randomization	competitive	sample	[31]
Scheer *et al.*	2006	JProGO	Fisher's exact test, Kolmogorov–Smirnov test, *t*-test, unpaired Wilcoxon's test	competitive	gene	[40]
Al-Shahrour *et al.*	2007	Fatiscan	Fisher's exact test, hypergeometric test	competitive	gene	[41]
Backes *et al.*	2007	GeneTrail	Fisher's exact test, hypergeometric test, sample randomization	competitive	gene or sample	[42]
Cavalieri *et al.*	2007	Eu.Gene Analyzer	Fisher's exact test, sample randomization	competitive	gene or sample	[43]
Dinu *et al.*	2007	SAM-GS	sample randomization	self-contained	sample	[22]
Efron *et al.*	2007	GSA	sample randomization	mixed	sample	[26]
Newton *et al.*	2007	Random set	z-test	competitive	gene	[44]

Table of software (from Nam & Kim)

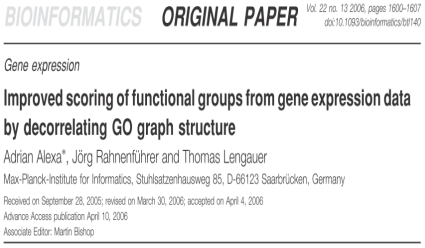
Table 2: Gene set analysis tools

Name	Organism[a]	Application Type	URL	Reference
ADGO	H, M, R, Y	Web server	http://array.kobic.re.kr/ADGO	[29]
ASSESS	H, M, R	Octave/Java standalone	http://people.genome.duke.edu/~jhg9/assess/	[28]
Babelomics	H, M, R, DM, S, C	Web server	http://www.babelomics.org	[45]
Catmap	H	Perl script	http://bioinfo.thep.lu.se/catmap.html	[3]
ErmineJ	H, M, R	Java standalone	http://www.bioinformatics.ubc.ca/ermineJ/	[16]
Eu.Gene Analyzer	H, M, R, Y	Windows/Unix standalone	http://www.ducciocavalieri.org/bio/Eugene.htm	[43]
FatiScan	H, M, R, Y, B, D, G, C, A, S, DM	Web server	http://fatiscan.bioinfo.cipf.es/	[41]
GAZER	H, M, R, Y	Web server	http://integromics.kobic.re.kr/GAzer/index.faces;	[13]
GeneTrail	H, M, R, Y, SA, CG, AT	Web server	http://genetrail.bioinf.uni-sb.de/	[42]
Global test	NA	R package	http://bioconductor.org/packages/2.0/bioc/html/globaltest.html	[17]
GOAL	H, M	Web server	http://microarrays.unife.it	[39]
GO-Mapper	H, M, R, Z, DM, Y	Windows standalone, Perl script	http://www.gatcplatform.nl/	[38]
GSA	H	R package	http://www-stat.stanford.edu/~tibs/GSA/	[26]
GSEA	H	Java standalone, R package	http://www.broad.mit.edu/gsea/	[25]
JProGO	Various prokaryotes	Web server	http://www.jprogo.de/	[40]
MEGO	H	Windows standalone	http://www.dxy.cn/mego/	[46]
PAGE	H, M, R, Y	Python script	From the author (kimsy@kribb.re.kr)	[14]
PLAGE	H, M	Web server	http://dulci.biostat.duke.edu/pathways/	[20]
SAFE	NA	R package	http://bioconductor.org/packages/2.0/bioc/html/safe.html	[19]
SAM-GS	NA	Windows Excel Add-In	http://www.ualberta.ca/~yyasui/homepage.html	[22]
T-profiler	Y, CA	Web server	http://www.t-profiler.org/	[15]

[a]H: *Homo sapiens*; M: *Mus musculus*; R: *Rattus norvegicus*; Y: *Saccharomyces cerevisiae*; B: *Bos Taurus*; D: *Daniel rerio*; G: *Gallus gallus*; C: *Caenorhabditis elegans*; A: *Arabidopsis thaliana*; DM: *Drosophila melanogaster*; Z: *Zebra fish*; CA: *Candida albicans*; SA: *Staphylococcus aureus*; CG: *Corynebacterium glutamicum*; AT: *Arabidopsis thaliana*.

TopGO

- TopGO is a GO term analysis program available from Bioconductor
- It takes the GO hierarchy into account when scoring terms
- If a parent term is only significant because of child term, it will receive a lower score
- TopGO uses the Fisher-test or the KS-test (both competitive)
- TopGO also gives a graphical representation of the results in form of a tree



BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 13 2006, pages 1600–1607
doi:10.1093/bioinformatics/btl140

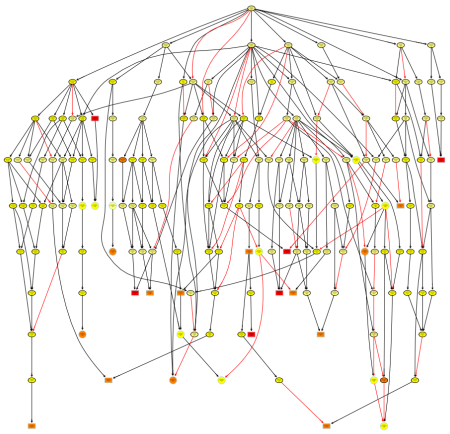
Gene expression

Improved scoring of functional groups from gene expression data by decorrelating GO graph structure

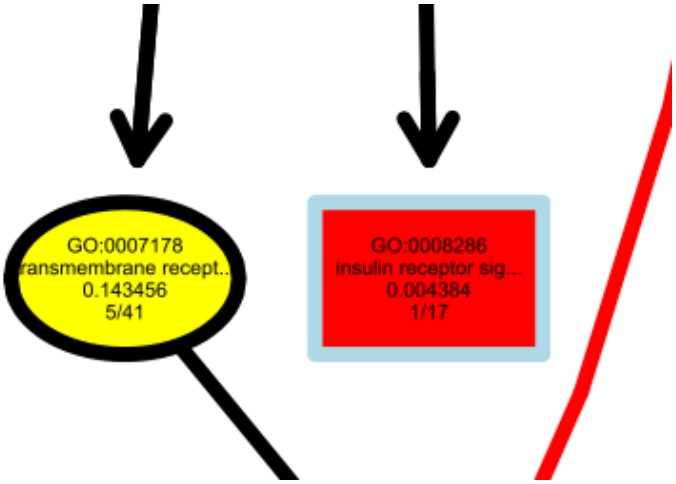
Adrian Alexa*, Jörg Rahnenführer and Thomas Lengauer
Max-Planck-Institute for Informatics, Stuhlsatzenhausweg 85, D-66123 Saarbrücken, Germany

Received on September 28, 2005; revised on March 30, 2006; accepted on April 4, 2006
Advance Access publication April 10, 2006
Associate Editor: Martin Bishop

Tree showing the 15 most significant GO terms



Zooming in



Global Ancova

- Uses all data (instead of summary statistics)
- NOT a multivariate method (MANOVA)
- One linear model for all genes within the gene set
  - Gene is a factor in the model that interacts with other factors
- Full model (e.g. including difference between lean and obsese) is compared with restricted model (no difference)
- P-values are calculated by group label resampling
- Algorithm allows for complex linear models including covariates
- Related to Goeman's Globaltest, which reverses roles of gene expression and groups: Goeman uses gene expression to explain groups (logistic regression)


Testing Differential Gene Expression in Functional Groups

Goeman's Global Test versus an ANCOVA Approach

U. Mansmann[1], R. Meister[2]

[1]IBE, Biometry and Bioinformatics, University of Munich, Munich, Germany

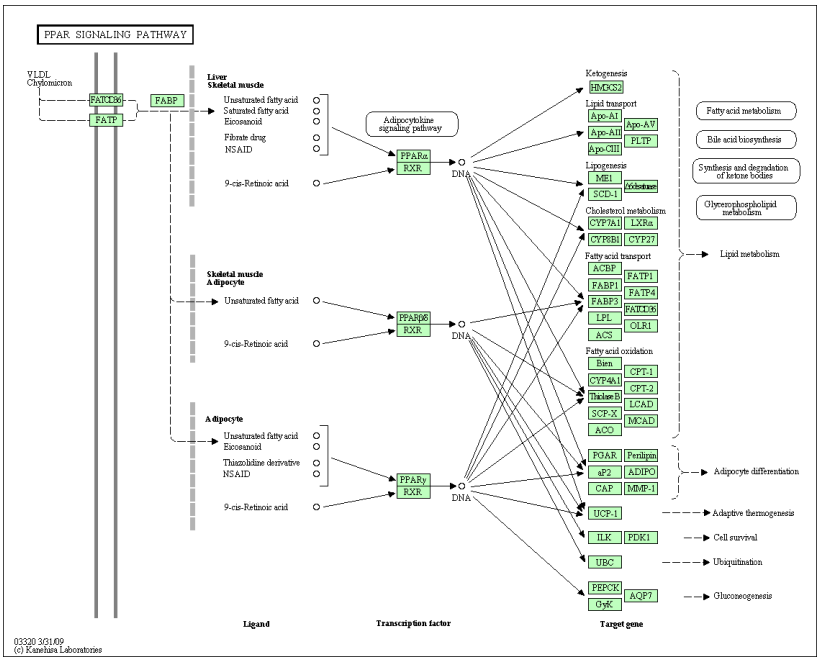
[2]Fachbereich II, University of Applied Sciences, Berlin, Germany

10 most significant KEGG pathways according to Global Ancova

Pathway Name	path.size	sig.genes	perc.sig	p.gs	p.fisher	p.globaltest	p.globalAncova
Pantothenate and CoA biosynthesis	11	3	27.27%	7.05%	9.08%	0.55%	0.01%
Valine, leucine and isoleucine biosynthesis	4	2	50.00%	4.10%	5.29%	0.22%	0.02%
Cell Communication	60	10	16.67%	8.77%	7.51%	1.02%	0.03%
PPAR signaling pathway	37	10	27.03%	11.01%	0.28%	1.64%	0.07%
Inositol metabolism	1	1	100.00%	8.46%	10.06%	0.19%	0.10%
Valine, leucine and isoleucine degradation	35	7	20.00%	49.56%	5.65%	1.42%	0.11%
Fatty acid metabolism	27	6	22.22%	49.59%	4.81%	1.54%	0.31%
ECM-receptor interaction	49	8	16.33%	4.91%	11.45%	1.47%	0.83%
Focal adhesion	122	16	13.11%	76.63%	16.40%	2.59%	0.87%
Purine metabolism	78	14	17.95%	26.82%	2.26%	3.42%	1.21%

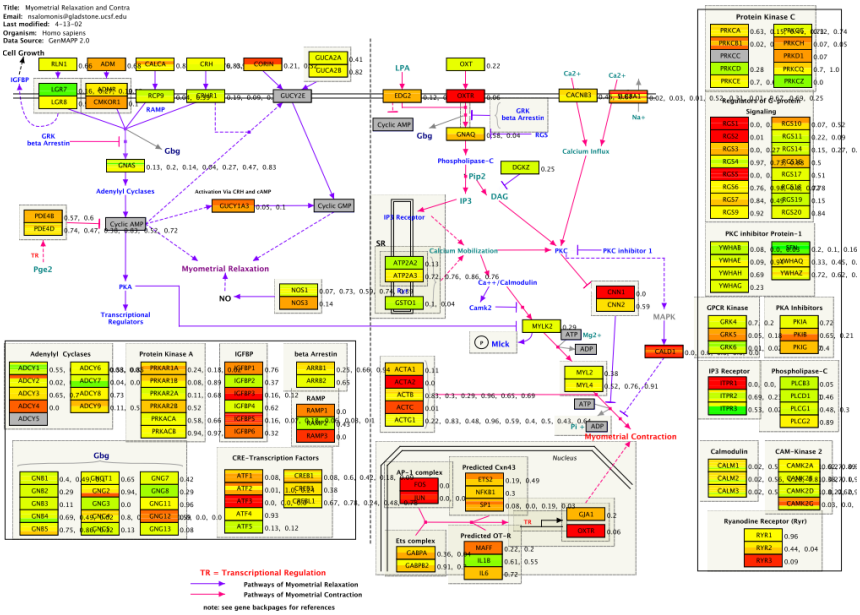
p.gs = A GSEA related competitive method (available in Limma)

p.fisher = Fisher-Test (competitive)



Genmapp/Pathvisio

- These are two pathway visualisation tools that collaborate
  - http://www.genmapp.org
  - http://www.pathvisio.org
- Both do some basic statistical analysis too (Fisher-Test with normal approximation)
- Main focus is on visually displaying pathways
  - Genes/nodes can be color-coded according to the data
  - Results (p-values, fold changes) can be displayed next to genes/nodes



Impact Factor Analysis

- Impact Factor (IF) analysis combines both ORA and FCS approach, while accounting for the topology of the pathway
- IF analysis computes Perturbation Factor (PF) for each gene in each pathway, which is a gene-level statistic, as follows:

$$PF(g_i) = \Delta F(g_i) + \sum_{j=1}^{n} \beta_{ji} \cdot \frac{PF(g_j)}{N_{ds}(g_j)}$$

- The first term, $\Delta F(g_i)$, represents the signed normalized measured expression change (i.e., fold change) of the gene g_i
- The second term accounts for the topology of the pathway, where gene g_j is upstream of gene g_i
- In the second term, β_{ji} represents the type and strength of interaction betweer g_j and g_i
- If g_j activates g_i, $\beta_{ji} = 1$, and if g_j inhibits g_i, $\beta_{ji} = -1$
- Note that the PF of the upstream gene g_j is normalized by the number of downstream genes it interacts with, $N_{ds}(g_i)$
- The second term is repeated for every gene g_j that is upstream of gene g_i

Impact Factor Analysis

- Next, Impact Factor (IF), is computed:

$$IF(P_i) = log\left(\frac{1}{p_i}\right) + \frac{\left|\sum_{g \in P_i} PF(g)\right|}{N_{de}(P_i)}$$

Impact Factor Analysis

- Next, Impact Factor (IF), is computed:

$$IF(P_i) = log\left(\frac{1}{p_i}\right) + \frac{\left|\sum_{g \in P_i} PF(g)\right|}{N_{de}(P_i)}$$

The 1st term captures the significance of the given pathway P_i as provided by ORA, where p_i corresponds to the probability of obtaining a value of the statistic used at least as extreme as the one observed when the null hypothesis is true

Impact Factor Analysis

- Next, Impact Factor (IF), is computed:

$$IF(P_i) = log\left(\frac{1}{p_i}\right) + \frac{\left|\sum_{g \in P_i} PF(g)\right|}{N_{de}(P_i)}$$

Because IF should be large for severely impacted pathways (i.e., small p-values), the 1st term uses $1/p_i$ rather than p_i

Impact Factor Analysis

- Next, Impact Factor (IF), is computed:

$$IF(P_i) = log\left(\frac{1}{p_i}\right) + \frac{\left|\sum_{g \in P_i} PF(g)\right|}{N_{de}(P_i)}$$

Log function is necessary to map the exponential scale of the p-values to a linear scale in order to keep the model linear

Impact Factor Analysis

- Next, Impact Factor (IF), is computed:

$$IF(P_i) = log\left(\frac{1}{p_i}\right) + \frac{\left|\sum_{g \in P_i} PF(g)\right|}{N_{de}(P_i)}$$

The 2nd term sums up the values of the PFs for all genes *g* on the given pathway P_i, and is normalized by the number of differentially expressed genes on the given pathway P_i

Impact Factor Analysis

- Note that Eq. 1 essentially describes the perturbation factor PF for a gene g_i as a linear function of the perturbation factors of all genes in a given pathway
- Therefore, the set of all equations defining the PFs for all genes in a given pathway P_i form a system of simultaneous equations
- Expanding and re-arranging Equation 1 for all genes g_1, g_2, …., g_n in a pathway Pi can be re-written as follows:

$$\begin{pmatrix} PF(g_1) \\ PF(g_2) \\ \cdots \\ PF(g_n) \end{pmatrix} = \begin{pmatrix} 1 - \frac{\beta_{11}}{N_{ds(g_1)}} & -\frac{\beta_{21}}{N_{ds(g_2)}} & \cdots & -\frac{\beta_{n1}}{N_{ds(g_n)}} \\ -\frac{\beta_{12}}{N_{ds(g_1)}} & 1 - \frac{\beta_{22}}{N_{ds(g_2)}} & \cdots & -\frac{\beta_{n2}}{N_{ds(g_n)}} \\ \cdots & \cdots & \cdots & \cdots \\ -\frac{\beta_{1n}}{N_{ds(g_1)}} & -\frac{\beta_{2n}}{N_{ds(g_2)}} & \cdots & 1 - \frac{\beta_{nn}}{N_{ds(g_n)}} \end{pmatrix}^{-1} \begin{pmatrix} \alpha(g_1) \cdot \Delta E(g_1) \\ \alpha(g_2) \cdot \Delta E(g_2) \\ \cdots \\ \alpha(g_n) \cdot \Delta E(g_n) \end{pmatrix}$$

Impact Factor Analysis

$$\begin{pmatrix} PF(g_1) \\ PF(g_2) \\ \cdots \\ PF(g_n) \end{pmatrix} = \begin{pmatrix} 1-\frac{\beta_{11}}{N_{ds(g_1)}} & -\frac{\beta_{21}}{N_{ds(g_2)}} & \cdots & -\frac{\beta_{n1}}{N_{ds(g_n)}} \\ -\frac{\beta_{12}}{N_{ds(g_1)}} & 1-\frac{\beta_{22}}{N_{ds(g_2)}} & \cdots & -\frac{\beta_{n2}}{N_{ds(g_n)}} \\ \cdots & \cdots & \cdots & \cdots \\ -\frac{\beta_{1n}}{N_{ds(g_1)}} & -\frac{\beta_{2n}}{N_{ds(g_2)}} & \cdots & 1-\frac{\beta_{nn}}{N_{ds(g_n)}} \end{pmatrix}^{-1} \begin{pmatrix} \alpha(g_1)\cdot\Delta E(g_1) \\ \alpha(g_2)\cdot\Delta E(g_2) \\ \cdots \\ \alpha(g_n)\cdot\Delta E(g_n) \end{pmatrix}$$

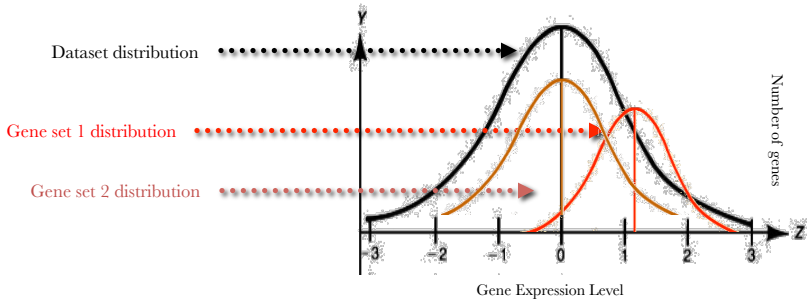
- After computing the PFs of all genes in a given pathway as the solution of this linear system, Eq. 2 is used to calculate the impact factor of each pathway
- The impact factor of each pathway is then used as a score to assess the impact of a given gene expression data set on all pathways (the higher the impact factor the more significant the pathway)

Gene Set Enrichment Analysis (GSEA)

- GSEA can be used with any gene set
- It is available as a standalone program, and versions of GSEA available within R/Bioconductor
- GSEA has many options and is a mix of a competitive and self-contained method
  - Default methods is to use a Kolmogorov Smirnov-type statistic to test the distribution of the gene set in the ranked gene list (competitive)
  - Typically that statistic ("enrichment score") is tested by permuting/ reshuffling the group labels (self-contained)
- Two Key Papers
  - Mootha et al., Nature Genetics 34, 267–273 (2003)
  - Subramanian et al., PNAS 102(43), 15545–15550 (2005).
    - Note - the description of GSEA changed between the two papers.

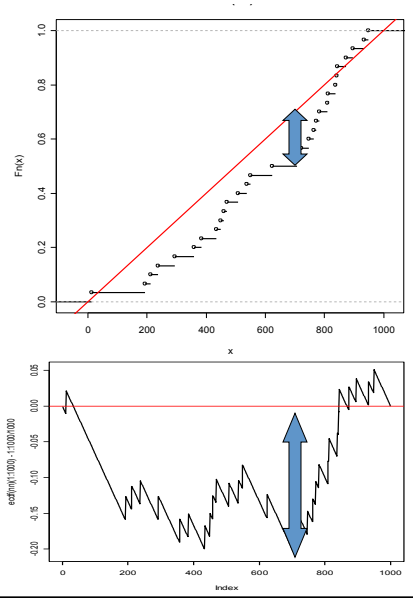
K-S Test

The Kolmogorov–Smirnov test is used to determine whether two underlying one-dimensional probability distributions differ, or whether an underlying probability distribution differs from a hypothesized distribution, in either case based on finite samples.

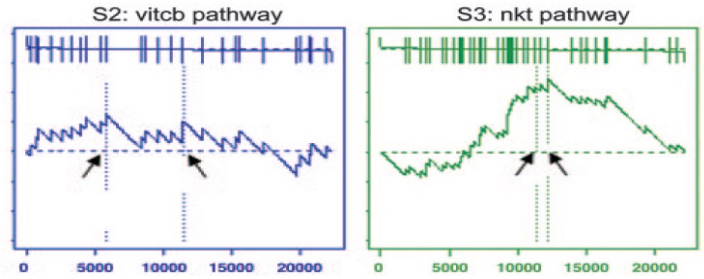


Kolmogorov-Smirnov Test

- Based on statistics of 'Brownian Bridge'
  - random walk fixed end
- Maximum difference is test statistic
  - Null distribution known
- Reformulated by GSEA as difference of CDF – uniform from axis



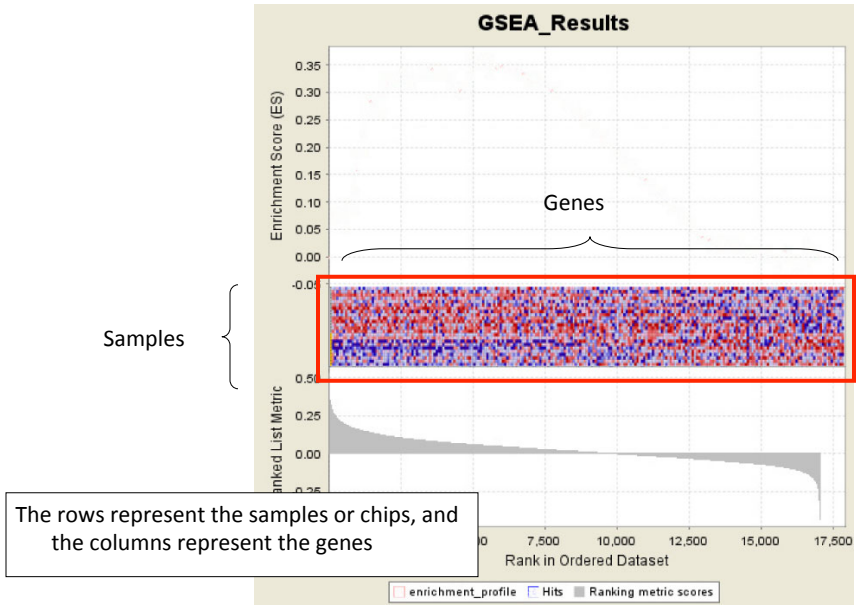
K-S Test Finds Irrelevant Sets



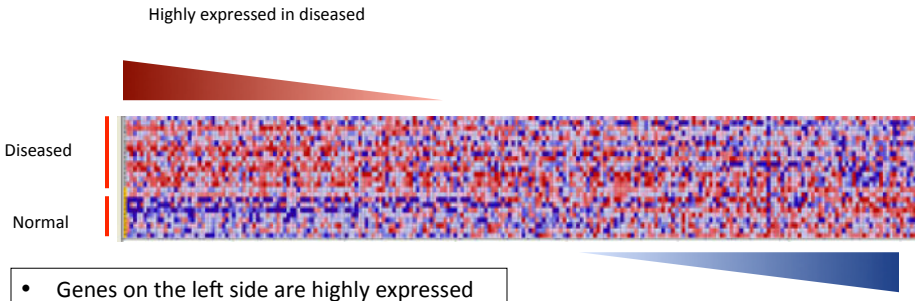
- Sometimes ranks concentrated in middle
  - K-S statistic high, but not meaningful for path change
- Fix: ad-hoc weighting by actual t-scores emphasizes departures at extreme ends
- No theory
- Generate null distribution by permutation

GSEA Algorithm: Step 1

- Calculate an Enrichment Score:
  - Rank genes by their expression difference
  - Compute cumulative sum over ranked genes:
    - Increase sum when gene in set, decrease it otherwise
    - Magnitude of increment depends on correlation of gene with phenotype.
- Record the maximum deviation from zero as the enrichment score

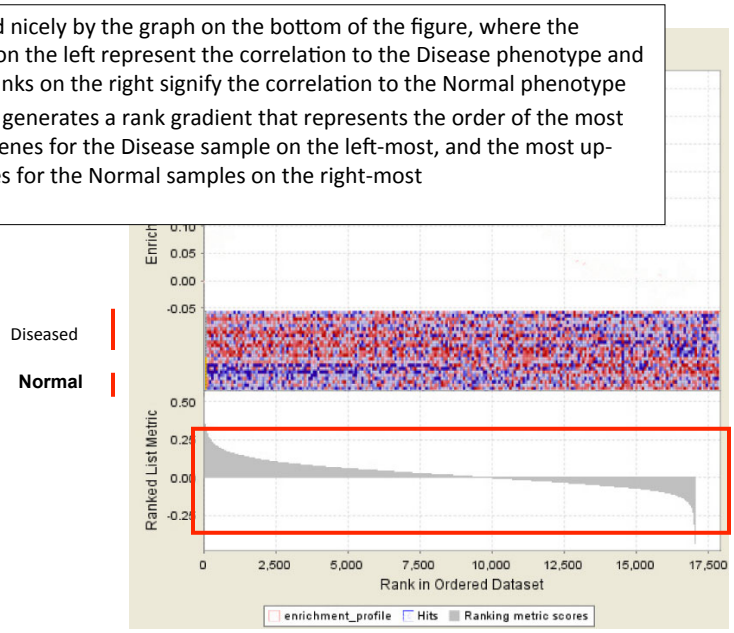


The rows represent the samples or chips, and the columns represent the genes

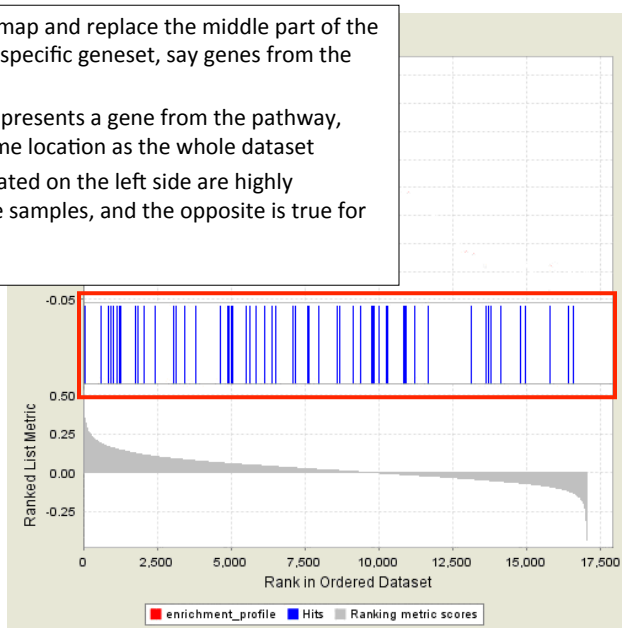


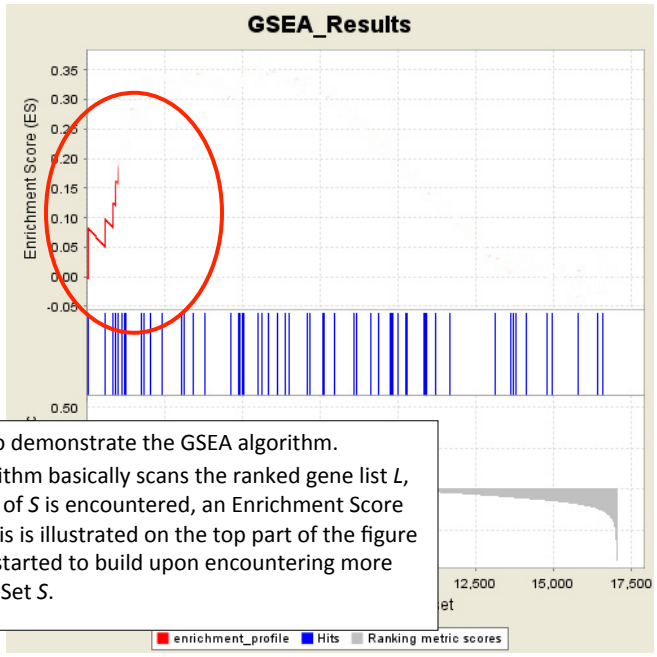
- Genes on the left side are highly expressed on the top half (indicated by red color) and lowly expressed on the bottom half (indicated by blue color). The reverse is shown on the right-most genes
- Created a gradient or ranked list corresponding to the degree of correlation with the two phenotypes

- This is depicted nicely by the graph on the bottom of the figure, where the positive ranks on the left represent the correlation to the Disease phenotype and the negative ranks on the right signify the correlation to the Normal phenotype
- The graph also generates a rank gradient that represents the order of the most up-regulated genes for the Disease sample on the left-most, and the most up-regulated genes for the Normal samples on the right-most

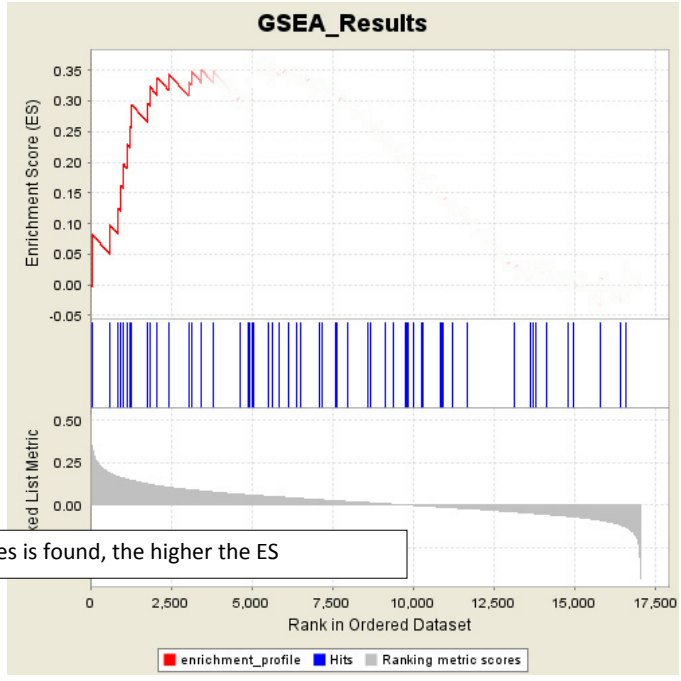


- Now, let's hide the heatmap and replace the middle part of the figure with genes from a specific geneset, say genes from the Glycolysis pathway.
- Each vertical blue bars represents a gene from the pathway, being mapped on the same location as the whole dataset
- Again, genes that are located on the left side are highly expressed on the Disease samples, and the opposite is true for the right-most genes

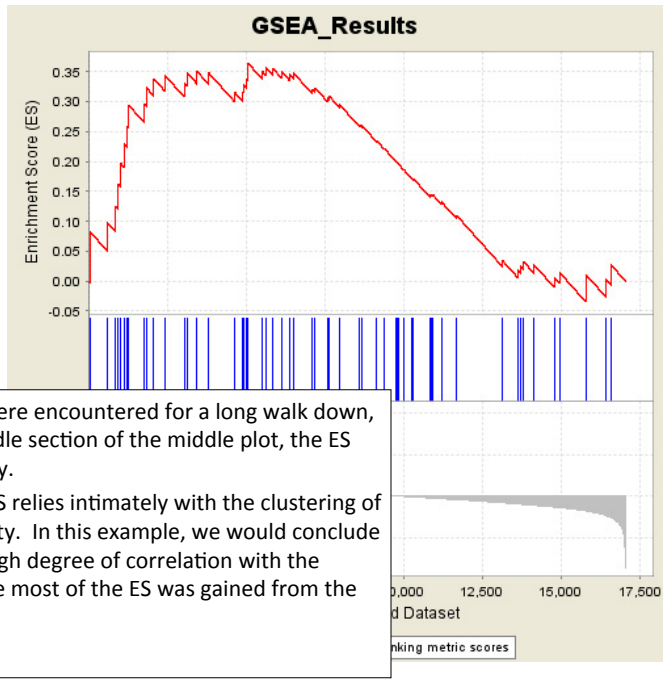




- Now, we are ready to demonstrate the GSEA algorithm.
- The walk down algorithm basically scans the ranked gene list *L*, and when a member of *S* is encountered, an Enrichment Score (*ES*) is registered. This is illustrated on the top part of the figure below; when the ES started to build upon encountering more genes from the GeneSet *S*.



- The more *S* genes is found, the higher the ES

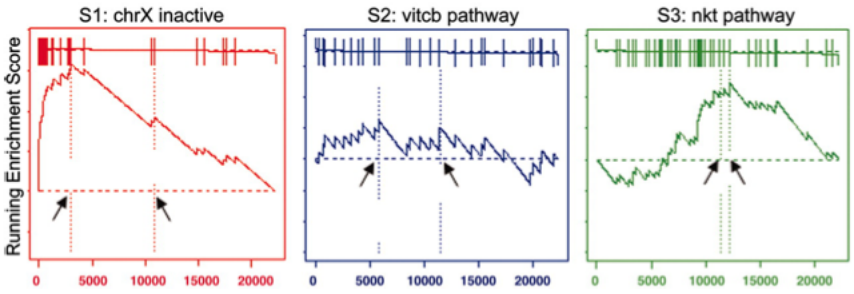


- But, when no *S* genes were encountered for a long walk down, as indicated on the middle section of the middle plot, the ES will decrease accordingly.
- In other words, a high ES relies intimately with the clustering of *S* genes in close proximity. In this example, we would conclude that the *S* genes have high degree of correlation with the Disease phenotype since most of the ES was gained from the left portion of the plot

GSEA Algorithm: Step 1

- Calculate an Enrichment Score:
  - Rank genes by their expression difference
  - Compute cumulative sum over ranked genes:
    - Increase sum when gene in set, decrease it otherwise
    - Magnitude of increment depends on correlation of gene with phenotype
- Record the maximum deviation from zero as the enrichment score

GSEA Algorithm: Step 1



Subramanian et al., PNAS 102(43), 15545–15550 (2005).

GSEA Algorithm: Step 2

- Assess significance:
  - Permute phenotype labels 1000 times
  - Compute ES score as above for each permutation
  - Compare ES score for actual data to distribution of ES scores from permuted data
- Permuting the phenotype labels instead of the genes maintains the complex correlation structure of the gene expression data

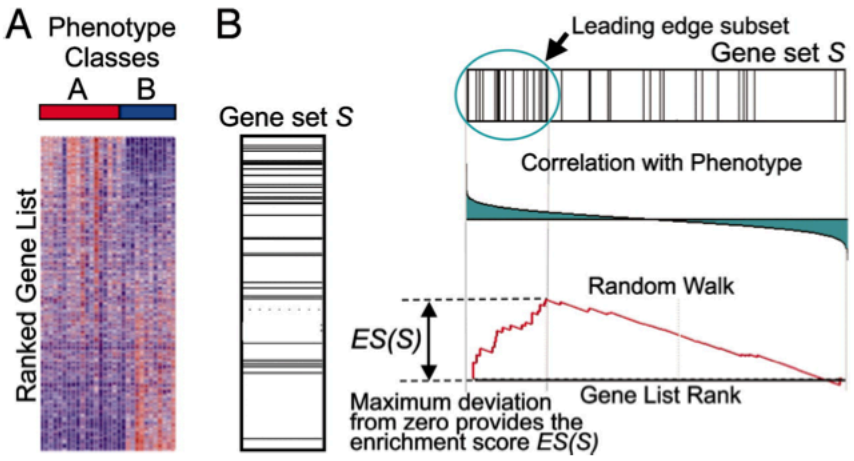
GSEA Algorithm: Step 3

- Adjustment for multiple hypothesis testing:
  - Normalize the ES accounting for size of each gene set, yielding normalized enrichment score (NES)
  - Control proportion of false positives by calculating FDR corresponding to each NES, by comparing tails of the observed and null distibutions for the NES

GSEA Algorithm: Step 4

- The original method used equal weights for each gene
  - The revised method weighted genes according to their correlation with phenotype
  - This may cause an asymmetric distribution of ES scores if there is a big difference in the number of genes highly correlated to each phenotype
- Consequently, the above algorithm is performed twice: one for the positively scoring gene sets and once for the negatively scoring gene sets

Overview of GSEA



Subramanian et al., PNAS 102(43), 15545–15550 (2005).

GSEA results for our data set (using pathway gene sets)

Enrichment in phenotype: lean (10 samples)

- 19 / 44 gene sets are upregulated in phenotype **lean**
- 0 gene sets are significant at FDR < 25%
- 0 gene sets are significantly enriched at nominal pvalue < 1%
- 1 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

Enrichment in phenotype: obese (9 samples)

- 25 / 44 gene sets are upregulated in phenotype **obese**
- 0 gene sets are significantly enriched at FDR < 25%
- 0 gene sets are significantly enriched at nominal pvalue < 1%
- 3 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

Dataset details

- The dataset has 12639 native features
- After collapsing features into gene symbols, there are: 6465 genes

Gene set details

- Gene set size filters (min=25, max=500) resulted in filtering out 595 / 639 gene sets
- The remaining 44 gene sets were used in the analysis
- List of gene sets used and their sizes (restricted to features in the specified dataset)

List of most significant up-regulated gene sets

Table: Gene sets enriched in phenotype lean (10 samples) [plain text format]

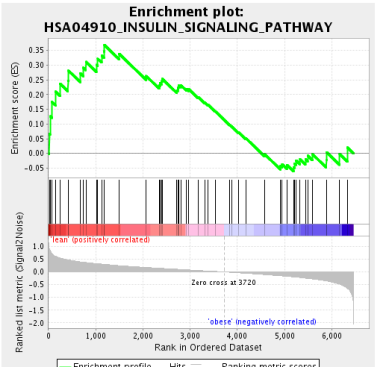
	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX
1	HSA04910_INSULIN_SIGNALING_PATHWAY	Details ...	51	0.37	1.41	0.036	0.960	0.620	1184
2	CALCINEURIN_NF_AT_SIGNALING	Details ...	32	0.39	1.33	0.074	0.833	0.800	2413
3	HSA04514_CELL_ADHESION_MOLECULES	Details ...	41	0.36	1.26	0.188	0.805	0.880	2038
4	HSA04310_WNT_SIGNALING_PATHWAY	Details ...	52	0.29	1.13	0.278	1.000	0.970	1086
5	HSA04350_TGF_BETA_SIGNALING_PATHWAY	Details ...	29	0.33	1.11	0.302	1.000	0.970	647
6	HSA05215_PROSTATE_CANCER	Details ...	28	0.36	1.11	0.291	0.914	0.970	1360
7	HSA04010_MAPK_SIGNALING_PATHWAY	Details ...	73	0.28	1.03	0.477	1.000	0.990	1482

Table: GSEA Results Summary

Dataset	Pimaunlog2_collapsed_to_symbols.Pima
Phenotype	Pima.cls
Upregulated in class	lean
GeneSet	HSA04910_INSULIN_SIGNALING_PATHWAY
Enrichment Score (ES)	0.3685702
Normalized Enrichment Score (NES)	1.4148982
Nominal p-value	0.035714287
FDR q-value	0.96008633
FWER p-Value	0.62

The Enrichment score is based on the difference of the cumulative distribution of the gene-set minus the expected

This plot is basically the Kolmogorov-Smirnov plot rotated by 45 degrees



Zoom In on Enrichment Plot

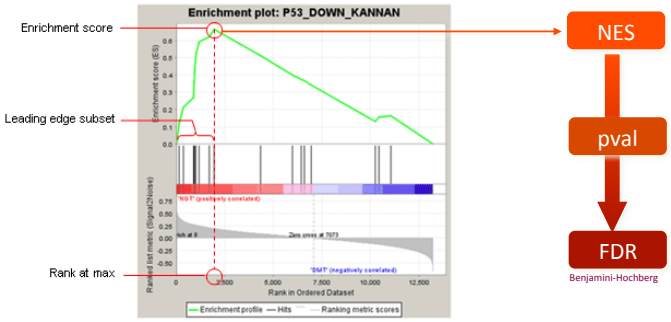
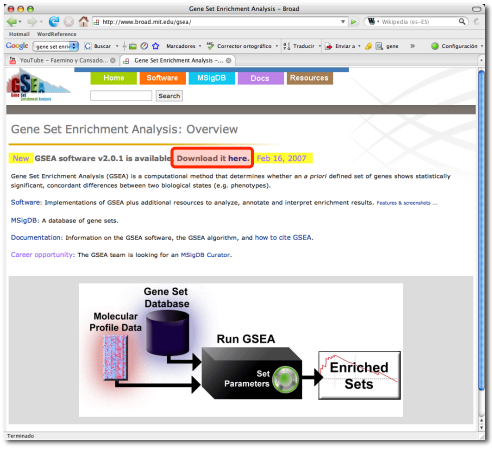


Fig 1: Enrichment plot: P53_DOWN_KANNAN
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

GSEA Software



http://www.broad.mit.edu/gsea/

Outlook

- Gene Set and Pathway Analysis is a very active field of research: new methods are published all the time!
- One important aspect: taking pathway structure into account
  - All methods we discuss ignored this structure
  - New methods use and "Impact Factor" (IF), which gives more weight to gene that are key regulators in the pathway (Draghici et al (2007))
- Other Aspects:
  - Study the behavior of pathways across experiments in microarray databases like GEO or Array Express
  - Incorporate other data into the analysis (proteomics, metabolomics, sequence data)

Summary

- There are many popular databases/internet resources for pathways and gene sets
- Many important analysis issues
- It is impossible to explain all existing approaches but many of them are some combinations of the methods we discussed
- This is an active field: improvements and further developments are a really active area of research

Questions?

Pathway/ Gene Set Analysis in Genome-Wide Association Studies

Alison Motsinger-Reif, PhD
Associate Professor
Bioinformatics Research Center
Department of Statistics
North Carolina State University

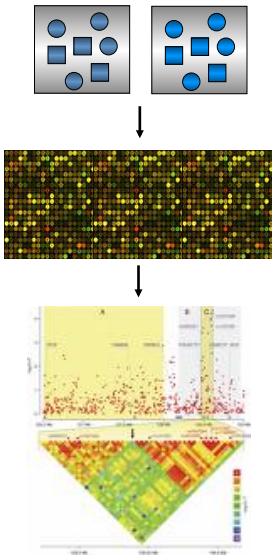
Goals

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

Many Shared Issues

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

Genome-Wide Association Studies



Population resources
- trios
- case-control samples

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

Genome-wide Association
- single SNP alleles
- genotypes
- multimarker haplotypes

Advantages of GWAS

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

Concerns with GWAS

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

Successes in GWAS Studies

- Over 400 GWAS papers published to date
- Big Finds:
  - In 2005, it was learned through GWAS that age-related macular degeneration is associated with variation in the gene for complement factor H, which produces a protein that regulates inflammation (Klein et al. (2005) Science, 308, 385–389)
  - In 2007, the Wellcome Trust Case-Control Consortium (WTCCC) carried out GWAS for the diseases coronary heart disease, type 1 diabetes, type 2 diabetes, rheumatoid arthritis, Crohn's disease, bipolar disorder and hypertension. This study was successful in uncovering many new disease genes underlying these diseases.

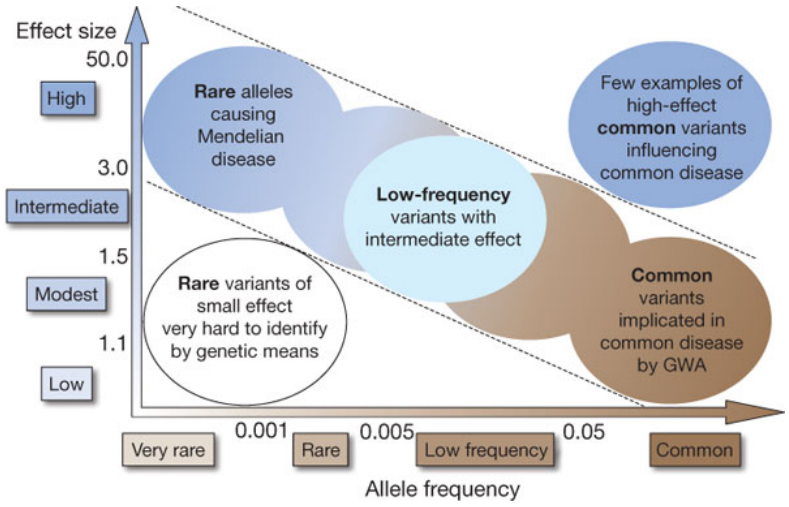
More Successes

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

Limitations

- For many diseases, the amount of trait variation explained by even the successes is way below the estimated heritability.
- Recently, GWAS are under a lot of criticism for relatively few translatable findings given the investment and hype.
- Assumptions underlying GWAS are not true for all diseases.

Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio).



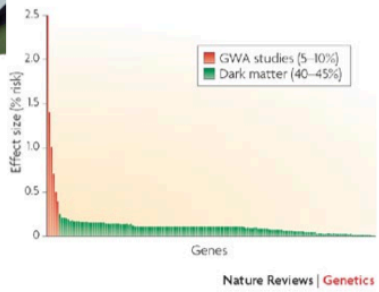
TA Manolio *et al. Nature* **461**, 747-753 (2009) doi:10.1038/nature08494

Reasons GWAS Can Fail

even if well-powered and well-designed….

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

Missing Heritability



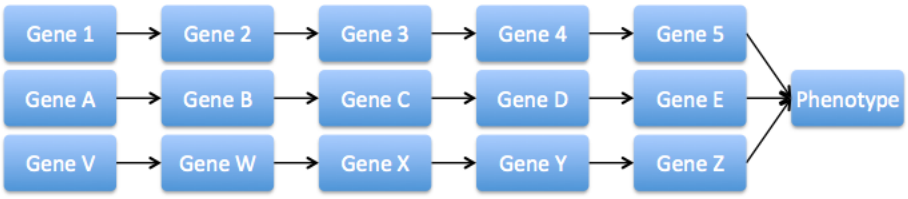
Nature Reviews | Genetics

Lusis et al, 2008

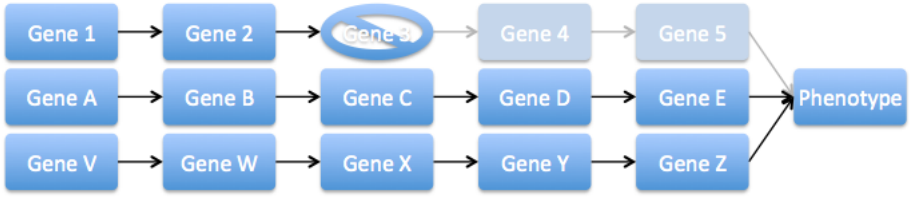
Possible Association Models

1. Each of several genes may have a variant that confers increased risk of disease independent of other genes
2. Several genes in contribute additively to the malfunction of the pathway
3. There are several distinct combinations of gene variants that increase relative risk but only modest increases in risk for any single variant

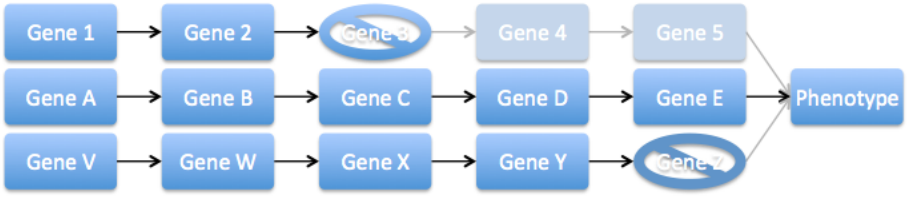
Hypothetical Disease Mechanism



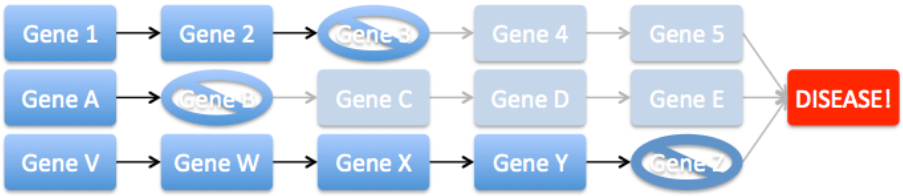
Hypothetical Disease Mechanism



Hypothetical Disease Mechanism



Hypothetical Disease Mechanism



Hypothetical Disease Mechanism

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

Best SNP		Pathway	
Significant	Suggestive	0.001	0.005
0.001	0.05	0.42	0.69

Linear Pathway

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

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

*Tested pathway includes 15 genes not in simulated pathway

Enrichment Testing in GWAS

- Testing pathway enrichment is possible in GWAS data
  - Many of the same issues that exist in gene expression enrichment testing occur in GWAS enrichment testing (e.g. choice of statistics, competitive vs self-contained)
- Primary difference:
  - In expression data the unit of testing is a gene
  - In GWAS data the unit of testing is a SNP
- Challenges:
  - Identifying the SNP (set) -> Gene mapping
  - Summarizing across individual SNP statistics to compute a per-gene measure

Mapping SNPs to Genes

- All SNPs in physical proximity of each gene
  - Pros:
    - All/most genes represented
  - Cons:
    - Varying number of SNPs per gene
    - Many of the SNPs may dilute signal
    - Defining gene proximity can affect results
- eSNPs (Expression associated SNPs)
  - Pros:
    - 1 SNP per gene
    - SNPs functionally associated
  - Cons:
    - Assumes variants effect expression
    - Not all genes have eSNPs
    - eSNPs may be study and tissue dependent

Gene summaries

- Initial studies propose different statistics for summarizing the overall gene association prior to enrichment analysis
  - Number/proportion of SNPs with pvalue < 0.05
  - Mean(-log10(pvalue))
  - Min(pvalue)
  - $1-(1-\text{Min(pvalue)})^{N}$
  - $1-(1-\text{Min(pvalue)})^{(N+1)/2}$

First approaches: combining p-values

- Compute gene-wise p-value:
  - Select most likely variant - 'best' p-value
  - Selected minimum p-value is biased downward
  - Assign 'gene-wise' p-value by permutations (Westfall-Young)
    - Permute samples and compute 'best' p-value for each permutation
    - Compare candidate SNP p-values to this null distribution of 'best' p-values
- Combine p-values by Fisher's method, across SNPs (biased in the presence of correlation)

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

$$p = \mathrm{P}(\chi^2_{(2k)} > 2V)$$

Next approaches

- Additive model: $\log(\frac{p}{1-p}) = \sum_{g_i \in G} \beta_i n_i$
  - Where n_i indexes the number of allele Bs of a SNP in gene *i* in the gene set *G*
  - Select subset of most likely SNP's
  - Fit by logistic regression (glm() in R)
- Significance by permutations
  - Permute sample outcomes
  - Select genes and fit logistic regression again
    - Assess goodness of fit each time
  - Compare observed goodness of fit

Competitive vs. Self-Contained Tests

- Competitive cutoff tests
  - Require only permuting SNP or Gene labels
  - May only allow to assess relative significance
- Self-contained distribution tests
  - Require permuting phenotype-genotype relationships
  - Resource intensive, may be difficult for large meta-analyses
  - Allow to assess overall significance

Competitive vs. Self-Contained Tests

- Self-contained null hypothesis
  - no genes in gene set are differentially expressed
- Competitive null hypothesis
  - genes in gene set are at most as often differentially expressed as genes not in gene set

What does this mean for SNP data?

Choice of Pathways/Gene Sets

- Relatively less "signal" in GWAS than in gene expression (GE)
  - GE enrichment typically test *which* gene sets/pathways show enrichment
  - GWAS enrichment typically test *if* there is enrichment
- Typically want to be conservative about selecting the number of pathways to test, otherwise will be difficult to overcome multiple testing
- Prioritized Approach:
  - Limited number of specific hypotheses (e.g. gene sets from experiment, co-expression modules, disease-specific pathways/ ontologies)
  - Exploratory analyses such as all KEGG/GO sets

Some Specific Methods

- SSEA
  - SNP Set Enrichment Analysis
- i-GSEA4GWAS
- MAGENTA
  - Meta-Analysis Gene-set Enrichment of variant Associations

SSEA

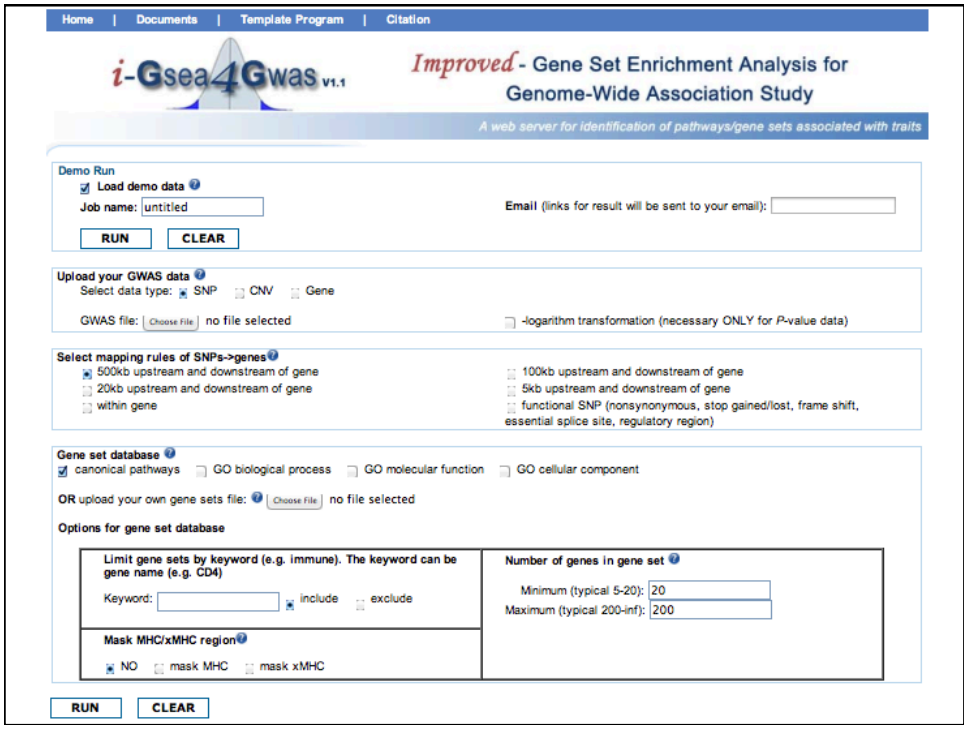
- Zhong et al. AJHG (2010)
- eSNP analysis to map SNPs to genes
  - More on this later…..
- Pathway statistic = one-sided Kolmogorov-Smirnov test statistic
- Pathway p-value assessed by permuting genotype-phenotype relationship
- FDR used to control error due to the number of pathways tested

i-GSEA4GWAS

- Zhang et al. *Nucl Acids Res* (2010)
- http://gsea4gwas.psych.ac.cn/

- Categorizes genes as significant or not significant
  - Significant: At least 1 SNP in the top 5% of SNPs
  - Does not adjust for gene size

- Pathway score: k/K
  - k = Proportion of significant genes in the geneset
  - K = Proportion of significant genes in the GWAS
- FDR assessed by permuting SNP labels



Results

Pathway/Gene set name	Description	Manhattan plot	P-value	FDR	genes/selected genes/All genes
HSA04950 MATURITY ONSET DIABETES OF THE YOUNG View Detail	Genes involved in ma..... More...		< 0.001	0.0030	11/23/25
PROSTAGLANDIN AND LEUKOTRIENE METABOLISM View Detail	More...		< 0.001	0.0085	13/27/32
HSA00565 ETHER LIPID METABOLISM View Detail	Genes involved in et..... More...		< 0.001	0.0125	15/28/31
DNA REPAIR View Detail	Genes annotated by t..... More...		< 0.001	0.0135	41/113/125
NTHIPATHWAY View Detail	Hemophilus influenza..... More...		< 0.001	0.0142	12/21/24
NEGATIVE REGULATION OF DEVELOPMENTAL PROCESS View Detail	Genes annotated by t..... More...		< 0.001	0.014571428	66/175/197
HSA04330 NOTCH SIGNALING PATHWAY View Detail	Genes involved in No..... More...		< 0.001	0.016	16/35/47
ENZYME LINKED RECEPTOR PROTEIN SIGNALING PATHWAY View Detail	Genes annotated by t..... More...		< 0.001	0.020875	60/136/140

MAGENTA

- Segre et al. *PLoS Genetics* (2010)
- Software download:
  - http://www.broadinstitute.org/mpg/magenta/
  - Requires MATLAB!!
  - Less convenient, but more customizable than iGSEA4GWAS
- Customizable proportion of "significant" genes
- Customizable gene window (upstream & downstream)
- Option for Rank-Sum test
- Gene Summary = min(p)
  - Uses stepwise regression to adjust for multiple possible factors: e.g. gene size, SNP density

MAGENTA Results

	95% Cutoff (Top 5%)				75% Cutoff (Top 25%)			
GS	NOMINAL GSEA PVAL	FDR	EXP # GENES	OBS # GENES	NOMINAL GSEA PVAL	FDR	EXP # GENES	OBS # GENES
positive regulation of osteoblast differentiation	3.36E-01	8.02E-01	1	2	3.00E-04	7.91E-02	6	14
one-carbon metabolic process	2.20E-03	3.55E-01	1	6	1.60E-03	1.44E-01	7	15
placenta development	3.36E-01	8.06E-01	1	2	4.00E-04	1.45E-01	6	14
carbohydrate transport	8.19E-01	9.46E-01	2	1	3.20E-03	3.45E-01	8	16

Adaptations of GSEA

- Order log-odds ratios or linkage p-values for all SNPs
- Map SNPs to genes, and genes to groups
- Use linkage p-values in place of t-scores in GSEA
  - Compare distribution of log-odds ratios for SNPs in group to randomly selected SNP's from the chip

Summary Points for GWAS

- In GWAS, few SNPs typically reach genome-wide significance
- Biological function of those that do can take years of work to unravel
- Incorporating biological information (expression, pathways, etc) can help interpret and further explore GWAS results
- Enrichment tests can be used to explore biological pathway enrichment
  - Different tests tell you different things
- Annotation choices very different that in gene expression data, though still rely on the same resources.... not necessarily so for other 'omics"

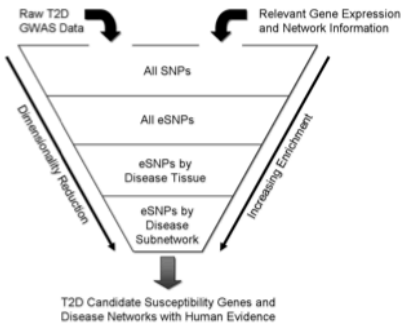
Adding in Gene Expression Data

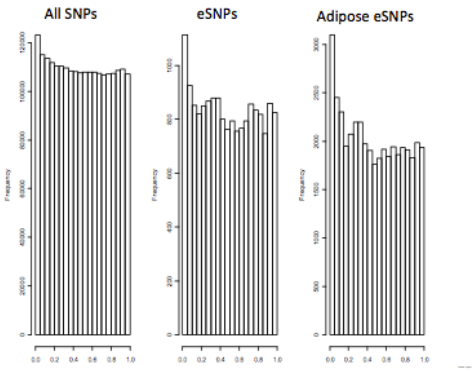
- Many motivating reasons to combine/integrate data from multiple "-omes"
- Expression and SNP data is most commonly done
  - Though methods could be applied to combine other "-omics"
- Generally make assumptions about central dogma

Genetics of Gene Expression

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

eSNPs can enrich p-values in GWAS studies
Example: T2D Data





Zhong et al. (2010) Elucidating Networks of eSNPs associated with Type 2 Diabetes.

Considerations on Filtering/Mining Data

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

Considerations: Multiple Test Correction

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

5% P-value vs 5% FDR

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

Partitioned SNP Testing (p-value)

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

Partitioned Testing (FDR)

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

eSNPs: Computing your own

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

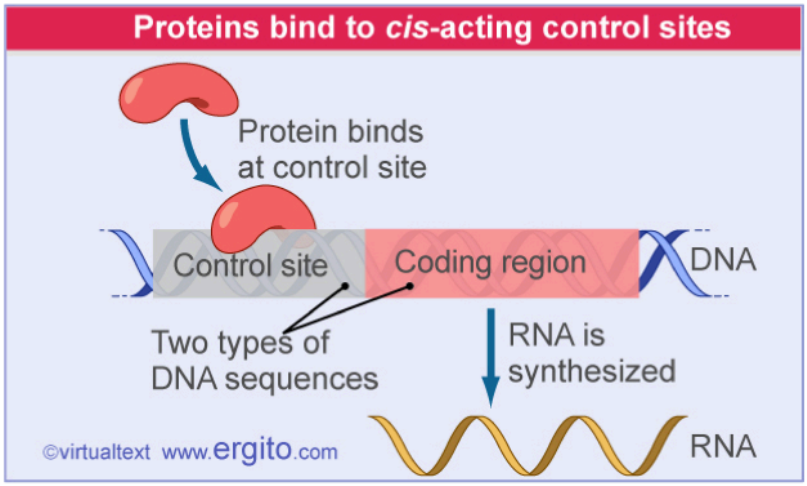
eSNPs: A note on computation

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

eSNP Discovery

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

Cis vs Trans Regulation



Aside: Cis/Trans vs Proximal/Distal

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

eSNPs: Publically Available

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

Motivation for Integrated Analysis

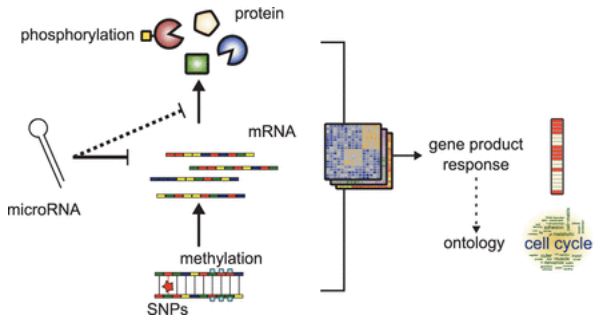
- Newer approaches will allow you to not do partitioned/filtered analysis, and leverage information across datatypes
- New technologies allow for more ready integration
  - Ex. RNA-Seq
  - Dropping costs allow for more datatypes to be collected simultaneously
  - Biobanking effort are storing more tissues

Motivation for Integrated Analysis

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

Motivation for Integrated Analysis

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



Summary

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

Summary

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

Questions?

motsinger@stat.ncsu.edu

Pathway Analysis in other data types

Alison Motsinger-Reif, PhD
Associate Professor
Bioinformatics Research Center
Department of Statistics
North Carolina State University

New "-Omes"

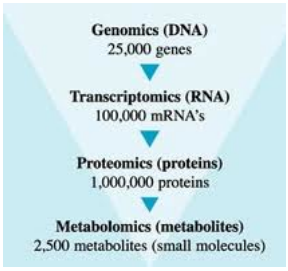
- Genome
- Transcriptome
- Metabolome
- Epigenome
- Proteome
- Phenome, exposome, lipidome, glycome, interactome, spliceome, mechanome, etc...

Goals

- Pathway analysis in metabolomics
- Pathway analysis in proteomics
- Issues, concerns in other data types
  - Methylation data
  - aCGH
  - Next generation sequencing technologies
- Many approaches generalize, but there are always specific challenges in different data types
- Weighted co-expression analysis

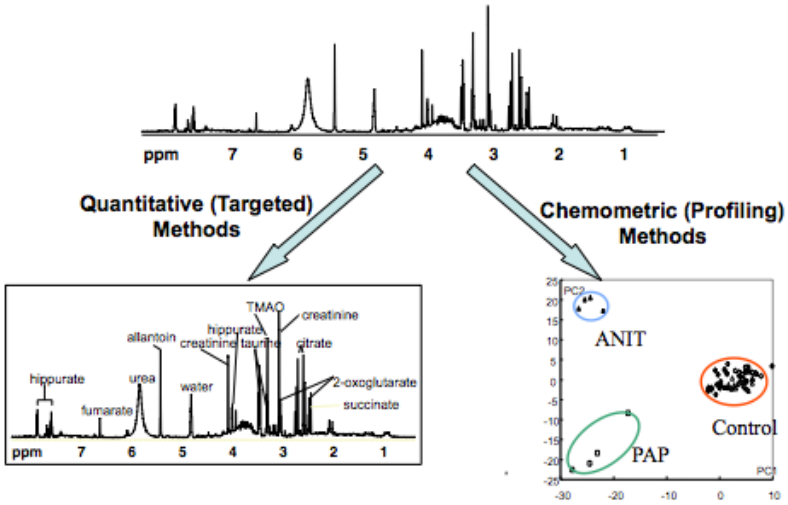
Metabolomics

- While many proteins interact with each other and the nucleic acids, the real metabolic function of the cell relies on the enzymatic interconversion of the various small, low molecular weight compounds (metabolites)
- Technology is rapidly advancing
- The frequent final product of the metabolomics pipeline is the generation of a list of metabolites who's concentrations have been (significantly) altered which must be interpreted in order to derive biological meaning



→ Perfect for pathway analysis

2 routes to Metabolomics



Data processing and annotation

- Preprocessing and the level of annotation is VERY different than in genomic and transcriptomic data
- Many steps in overall experimental design that greatly influence interpretation
- Will breifly cover some of the main issues

Analytical Platform

- Likely GC/LC-MS or NMR as they are the most common
- Choice is normally based more on available equipment, etc. more than experimental design
- GC-MS is an extremely common metabolomics platform, resulting in a high frequency of tools which allow for the direct input of GC-MS spectra.
  - Popularity is due to its relatively high sensitivity, broad range of detectable metabolites, existence of well-established identification libraries and ease of automation
  - separation-coupled MS data requires much processing and careful handling to ensure the information it contains is not artifactual

Targeted vs. Untargeted

- Scientists have been quantifying metabolite levels for over 50 years through targeted analysis…
- With new technologies, the focus can be on untargeted metabolomics
  - Really hard to annotate and interpret
  - Integrated –omics analysis being used to help annotate and understand untargeted metabolites
  - Analogous to candidate gene vs. genome wide testing

Key Issues in Metabolomics

- All of the metabolites within a system cannot be identified with any one analytical method due to chemical heterogeneity, which will cause downstream issues as all metabolites in a pathway have not been quantified
- Not all metabolites have been identified and characterized and so do not exist in the standards libraries, leading to large number of unannotated and/or unknown metabolites of interest
- Organism specific metabolic databases/networks only exist for the highest use model organisms making contextual interpretations difficult for many researchers
- Interpreting the huge datasets of metabolite concentrations under various conditions with biological context is an inherently complex problem requiring extremely in depth knowledge of metabolism.
- The issue of determining which metabolites are actually important in the experimental system in question.

Metabolomic Databases

- Two types of data-bases:
  - top-down (gene to protein to metabolite)
  - bottom-up (chemical entity to biological function) approaches
  - www.metabolomicssociety.org/database
- Most commonly used in biomedical applications:
  - MetaCyc
  - KEGG
    - Subdatabases LIGAND, REACTION PAIR and PATHWAY

Metabolomic Databases

- KEGG and MetaCyc are largest (in terms of number of organisms and most in depth comprehensive (i.e. contains linked information from metabolite to gene)
- Others that are rapidly growing:
  - Reactome (human)
  - KNApSAcK (plants)
  - Model SEED (diverse)
  - BiG [40] (6 model organisms)
  - can be more useful than the large databases if a specific organism is desired

Metabolomic Databases

- KEGG and MetaCyc databases each contain a generalized 'conserved' set of pathways based on metabolic pathways that are more or less the same throughout life in general
  - For KEGG, organism specific annotations are available to query
  - For MetaCyc, individual 'Cyc' databases have been generated for a number of organisms,
    - some just computationally
    - others extensively manually curated such as AraCyc for Arabidopsis
- More recent development are the cheminformatic databases like PubChem
  - provide a chemically ontological approach to cataloguing the ill-defined category of 'small molecules' active in biological systems
  - can provide additional non-biology specific information as well alternative formatting options for datasets *(watch for errors!)*

Enrichment analysis

- These databases are used to create "metabolite sets" for enrichment analysis
- Majority of available tools do early generation over-representation analysis
  - With all the advantages and caveats!
  - For more up to date analysis, will need to work to merge databases, etc. to correctly use more up-to-date approaches

Metabolomics Analysis Tools

- Comprehensive platforms
  - Provide a suite of utilities allowing comprehensive analysis from raw spectral data to pathway analysis
    - MetaboAnalyst
    - MeltDB
- Enrichment Analysis
  - Only works with processed data
    - PAPi
    - MBRole
    - MPEA
    - TICL
    - IMPaLA
- Metabolite Mapping
  - Connects metabolites to genetic/proteomic, etc. resources
    - MetaMapp
    - Masstrix
    - Paintomics
    - VANTED
    - Pathos

Metaboanalyst

- A number of utilities:
  - Data quality checking (useful for batch effects)
  - metabolite ID converter among others are also included.
  - If beginning from raw GC or LC-MS data MetaboAnalyst uses XCMS for peak fitting, identification etc.
  - Once at the peak list (NMR or MS) stage, various preprocessing options such as data-filtering and missing value estimation can be used.
  - A number of normalization, transformation and scaling operations can be performed.
  - Suite of statistical analyses including metabolomics standards like PCA, PLS-DA and hierarchically clustered heatmaps, among many other options.
  - *All these things can be done in other programs, but this is a great tool to get started if you're new to metabolomics!*

Metaboanalyst

- Enrichment Analysis tool of MetaboAnalyst was one of the earliest implementations of GSEA for metabolomics datasets (MSEA)
  - quite biased towards human metabolism unless you make custom background pathways/sets
- Three options for input
  - a single column list of compounds (Over Representation Analysis, ORA)
  - a two column list of compounds AND abundances (Single Sample Profiling, SSP)
  - a multi-column table of compound abundances in classed samples (Quantitative Enrichment Analysis, QEA).

Metaboanalyst

- ORA will calculate whether a particular set of metabolites is statistically significantly higher in the input list than a random list, which can be used to examine ranked or threshold cut-off lists
- SSP is aimed at determining whether any metabolites are above the normal range for common human biofluids
- QEA is the most canonical and will determine which metabolite sets are enriched within the provided class labels, while providing a correlation value and p-value

PAPi

- Pathway Activity Profiling is an R-based tool
- As input it takes a list with abundances (normalized and scaled)
- Works on the assumptions that the detection (i.e. presence in the list) of more metabolites in a pathway and that lower abundances of those metabolites indicates higher flux and therefore higher pathway activity
  - Assumption may not always be true
  - Ex. TCA cycle intermediates can have high abundance even when flux through the reactions in this pathway is also high

PAPi

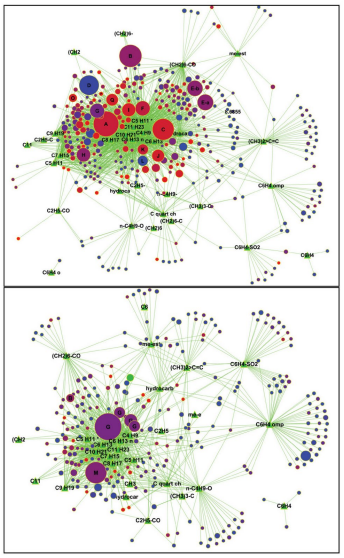
- PAPi calculates an activity score (AS) for each pathway
- The metabolic pathways are taken from the general KEGG database
- The AS indicates the probability of this pathway being active in the cell
- These scores can then be used to compare experimental and control conditions by performing ANOVA or a t-test to compare two sample types.

MetaMapp

- Performs metabolic mapping for unknown and unannotated metabolites
- Since biochemistry is the interconversion of chemically similar entities, compounds can be clustered solely by their chemical similarity
  - Highly beneficial for metabolites without reaction annotation
- Also uses KEGG reactant pair information
  - chemical similarity misclustered some obviously biologically-related metabolites

MetaMapp

- Can also map metabolites based on their mass spectral similarity (for unknowns)
- Can be used to make custom/novel sets for pathway analysis



Summary on Metabolomics Pathway Analysis

- Metabolomics is a maturing area
- "Easy" implementations of tools often behind best practices in pathway approaches
- Issues with time dependencies, tissue dependencies, etc. are more exaggerated in metabolomics
- As the technology is maturing, we are just getting to understand the biases, sources of variation, etc.
  - Data quality control best practices are evolving
  - Will have major impact on the pathway analysis

Specific Issues for other -omics

- Will consider some issues that are both specific to the "-ome" and to particular technologies

- Proteomics
- Epigenomics
- Array CGH data
- RNA seq
- Next generation sequencing
- …….

Proteomics

- After genomics and transcriptomics, proteomics is the next step in central dogma
- Genome is more or less constant, but the proteome differs from cell to cell and from time to time
- Distinct genes are expressed in different cell types, which means that even the basic set of proteins that are produced in a cell needs to be identified

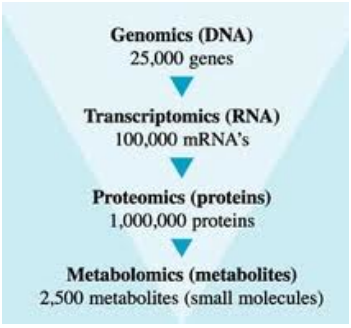
- It was assumed for a long time that microarrays would capture much of this information → NO!

Proteomics vs. Transcriptomics

- mRNA levels do not correlate with protein content
- mRNA is not always translated into protein
- The amount of protein produced for a given amount of mRNA depends on the gene it is transcribed from and on the current physiological state of the cell
- Many proteins are also subjected to a wide variety of chemical modifications after translation
  - Affect function
  - Ex: phosphorylation, ubiquitination
- Many transcripts give rise to more than one protein, through alternative splicing or alternative post-translational modifications

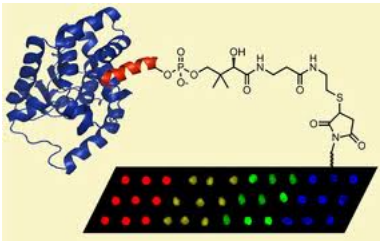
Proteomics

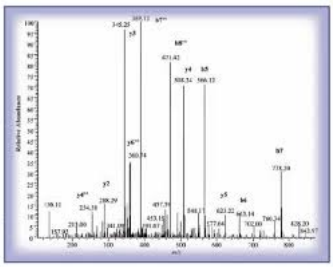
- Technological advances for proteomics has slowed
  - Like metabolomics, the lack of any PCR-like amplification is limited
  - Unlike metabolomics that has a reasonable search space, there estimated to be more than a million transcripts



Proteomics

- Available technologies have different challenges
  - Protein microarrays vs. mass spec based methods
  - General concerns with reproducibility dampened initial excitement





Proteomics

- The high complexity and technical instability mean that the level of annotation is often quite low
- Same challenges as with metabolomics, but more exaggerated given the large annotation space
- Many of the same issues …..

Epigenomics

- "Complete" set of epigenetic modifications on the genetic material of a cell
  - epigenetic modifications are reversible modifications on a cell's DNA or histones that affect gene expression without altering the DNA sequence
  - DNA methylation and histone modification most commonly assayed
- Rapidly advancing technologies
  - Histone modification assays
  - CHIP-CHIP and CHIP-Seq
  - Methylation arrays

Epigenomics

- Recent studies have focused on issues related to differential numbers of probes in genes
  - Most microarrays were designed with the same number
  - For methylation data, this is not the case, and extreme bias can be seen
  - Bias results in a large number of false positives
- Can be corrected by applying methods that models the relationship between the number of features associated with a gene and its probability of appearing in the foreground list
  - CpG probes in the case of microarrays
  - CpG sites in the case of high-throughput sequencing
  - Chip annotation
- Can also be corrected with careful application of permutation approaches

Next Generation Sequencing

- Variant calling in NGS can detect single nucleotide variants (SNVs) and SNPs
- For SNPs, the exact same pathway methods can be used as designed for GWAS studies (assuming genotyping in genome wide)
- For rare variants, standard approaches are a challenge
  - highly inflated false-positive rates and low power in pathway-based tests of association of rare variants
  - due to their lack of ability to account for gametic phase disequilibrium
  - New area of methods development

Next Generation Sequencing

- RNA-seq data
  - Not truly quantitative
  - With experience, know that there are very different variance distributions at different levels of expression
  - Will matter for methods that test for differences in variance as well as mean
    - Two sided K-S tests….

Summary on Integrated Analysis

- Technology advances across the "omics" is an exciting opportunity for better understanding complexity
- Technologies have unique properties that need to be understood and accounted for in analysis
- Metabolomics resources are rapidly maturing

Summary on Integrated Analysis

- Database development, curation, editing, etc. always lags behind technology
- Issues with incomplete and inaccurate annotation accumulate as more "omes" are considered
- With more complex data, this complexity is not readily captured in the databases the gene set analysis relies on
  - Differences in cell types, exposure, time, etc.
  - Major needs for methods development…..

Questions?

motsinger@stat.ncsu.edu