Introduction to Pathway and Network Analysis

Alison Motsinger-Reif, PhD
Branch Chief, Senior Investigator
Biostatistics and Computational Biology Branch
National Institute of Environmental Health Sciences

alison.motsinger-reif@niehs.nih.gov

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

Pathway and Network Analysis

- Simplify analysis by grouping long lists of individual genes into smaller sets of related genes to reduce complexity
  - Can be derived from the data (clustering approaches)
  - Knowledge bases

- Knowledge bases
  - describe biological processes, components, or structures in which individual genes are known to be involved in
  - how and where gene products interact

Pathway and Network Analysis

- Simplify analysis by grouping long lists of individual genes into smaller sets of related genes

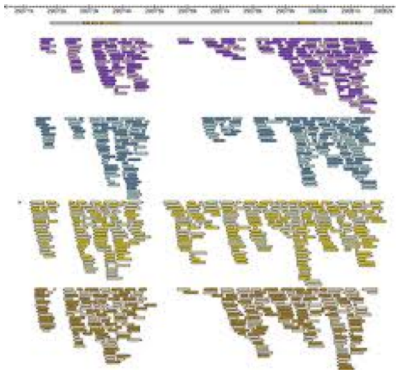
Rationale:

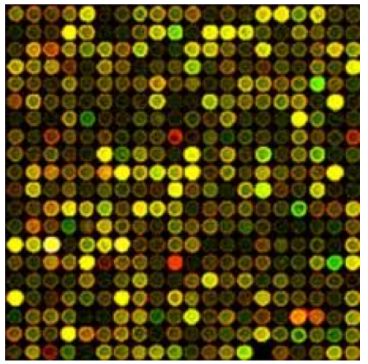
- Grouping genes reduces complexity from thousands to typically hundreds of pathways
- Pathways may have more explanatory power than single genes

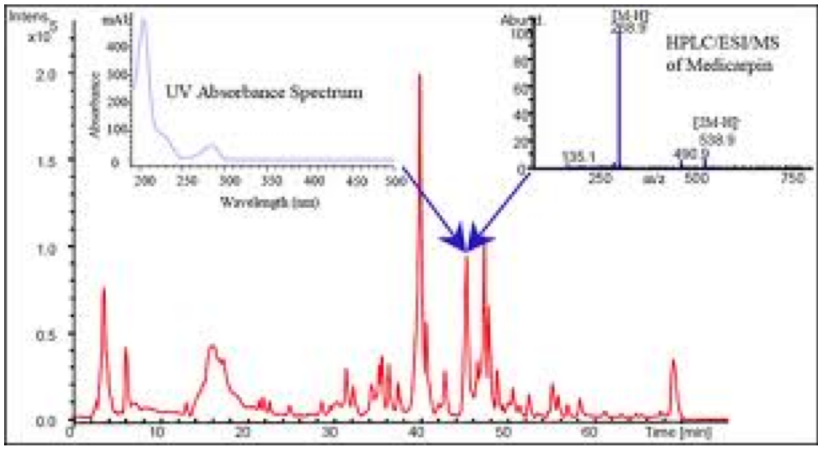
- – how and where gene products interact

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

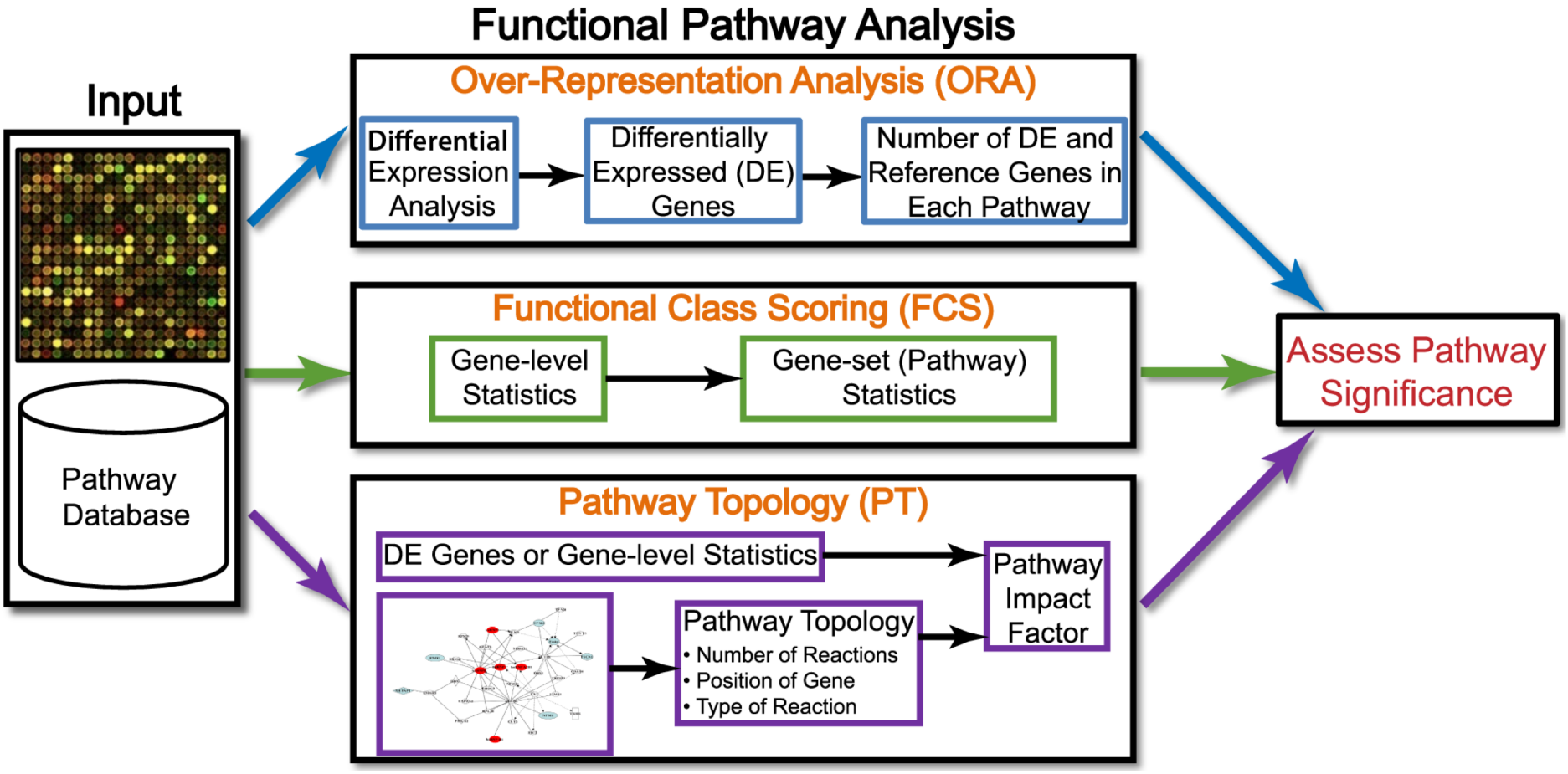
- For the first part of this module, we will focus on methods that exploit pathway knowledge in public repositories rather than on methods that infer pathways from molecular measurements
  - Use repositories such as GO or Kyoto Encyclopedia of Genes and Genomes (KEGG)

→ *knowledge base–driven pathway analysis*

A History of Pathway Analysis Approaches

- 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
- “2×2 table methods”

Over-Representation Analysis (ORA)

- Uses one or more variations of the following strategy:
  - Create an input list using a certain threshold or criteria (significance cutoff, etc.)
  - For each pathway, count input genes that are part of the pathway
  - Repeat for an appropriate background list of genes
    - (e.g., all genes measured on a microarray)
  - 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

Limitations of ORA Approaches

- Different statistics used are independent of the measured changes
  - Loses information, treats each gene equally
- Typically uses only the most significant genes and discards the others (information loss)
- Assumes that each gene is independent of the other genes
  - Counterintuitive to understanding of interactions amongst 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
- ORA assumes that each pathway is independent of other pathways → NOT TRUE!

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*

Step 1

- First, compute a gene-level statistic using the molecular measurements from an experiment
  - (i.e. differential expression)

- 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, aggregate gene-level statistics for all genes in a pathway 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

- Assess 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

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.
   - use all available molecular measurements for pathway analysis.

2. FCS methods use molecular measurement to detect coordinated changes in the expression of genes in the same pathway

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

Limitations of FCS Methods

- 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, while one pathway may be affected in an experiment, one may observe other pathways being significantly affected due to the set of overlapping genes
- Typically rank based statistics
  - Usual advantages and disadvantages of rank based statistics
  - There are notable exceptions to this scenario is approaches that use gene-level statistics (e.g., t-statistic) to compute pathway-level scores (SAFE, etc.)

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
  - ….
- 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
  - 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

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

Outstanding Challenges

- Broad Categories:

1. annotation challenges
2. methodological challenges

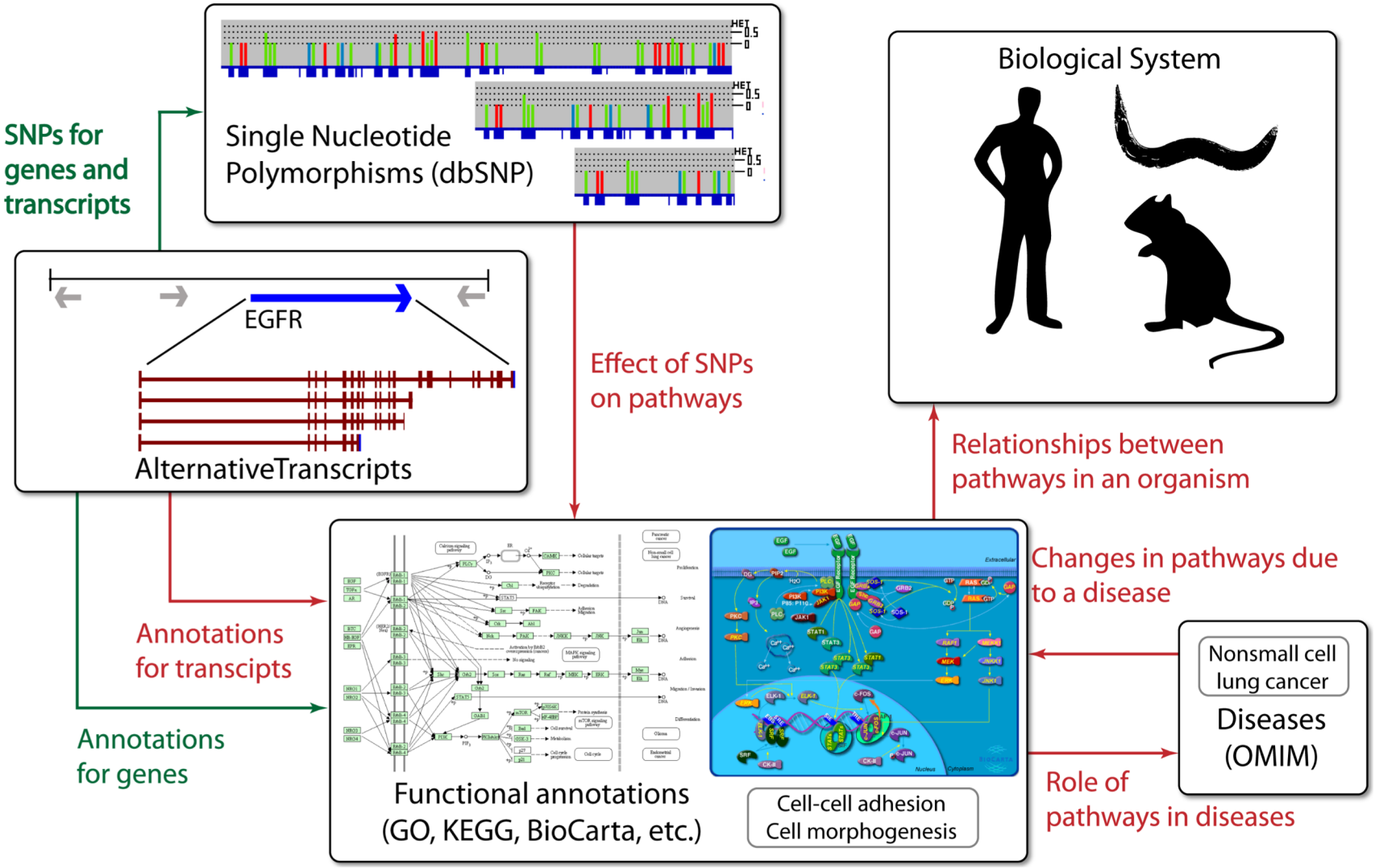
Outstanding Challenges

- 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

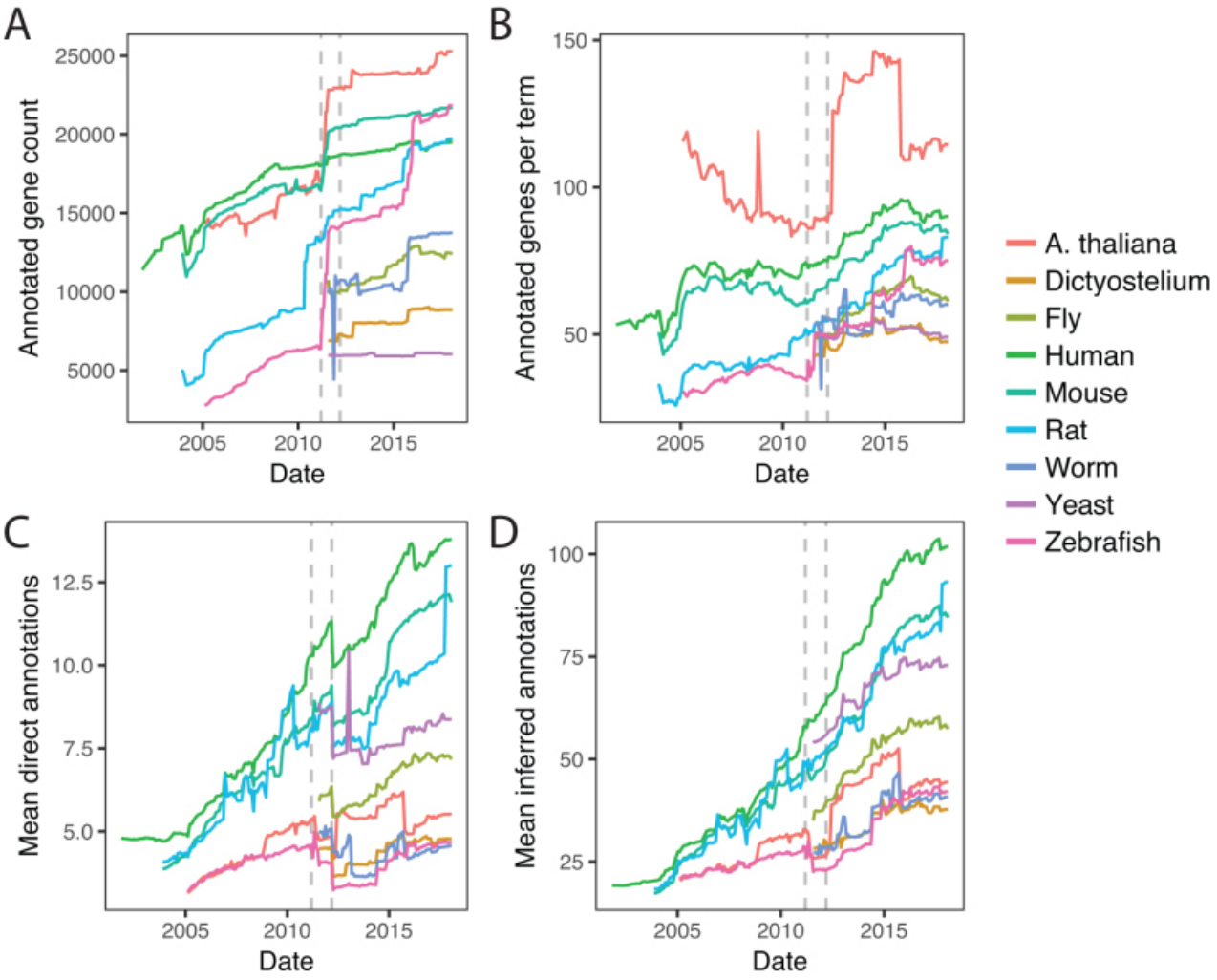
- While new approaches are being developed in this regard, they may not yet be adequate
  - Braun et al. proposed a method for analyzing SNP data from a GWAS
  - Still relies on mapping multiple SNPs to a single gene, followed by gene-to-pathway mapping

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



Jacobson et al, Monitoring changes in the Gene Ontology and their impact on genomic data analysis. Gigascience. 2018 Aug; 7(8); giy103

Trends in taxon-wide annotation statistics. **(A)** Number of annotated genes. **(B)** Mean annotations per term (inferred + direct). **(C)** Mean number of direct annotations per gene. **(D)** Mean number of inferred (including direct) annotations per gene. Times of prominent discontinuities affecting multiple species in A and C are marked by dashed gray lines in all four panels.

Incomplete and Inaccurate Annotation

- Additionally, many of the existing annotations are of low quality and may be inaccurate
  - >90% of the annotations in the October 2015 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 2015 release of human GO annotations is reduced from ~18K to ~12K

Incomplete and Inaccurate Annotation

- Likely that the reduced number of annotations and annotated genes since January 2003 is an indicator of improving quality

- 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

- 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
  - New methods emerging here

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

- Most current approaches 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 15 years, 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?