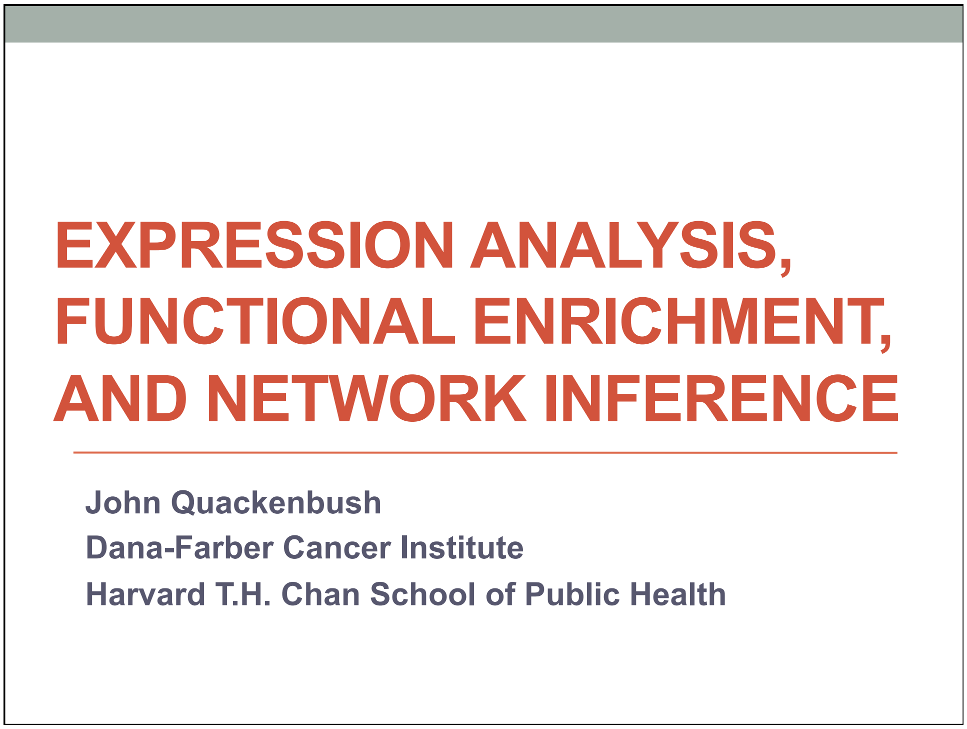
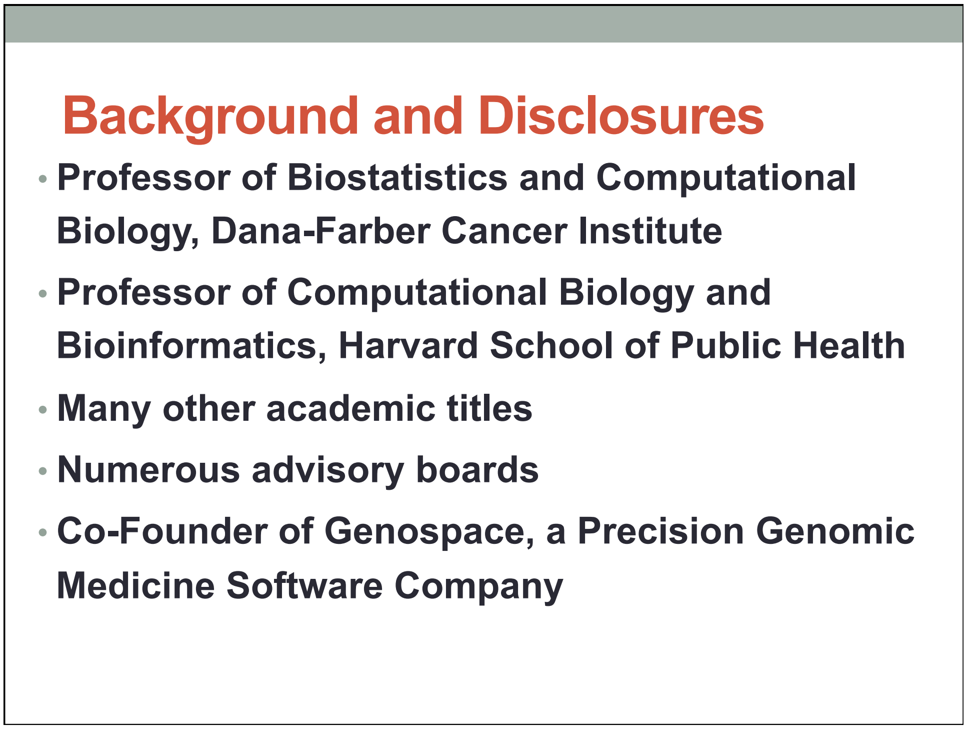


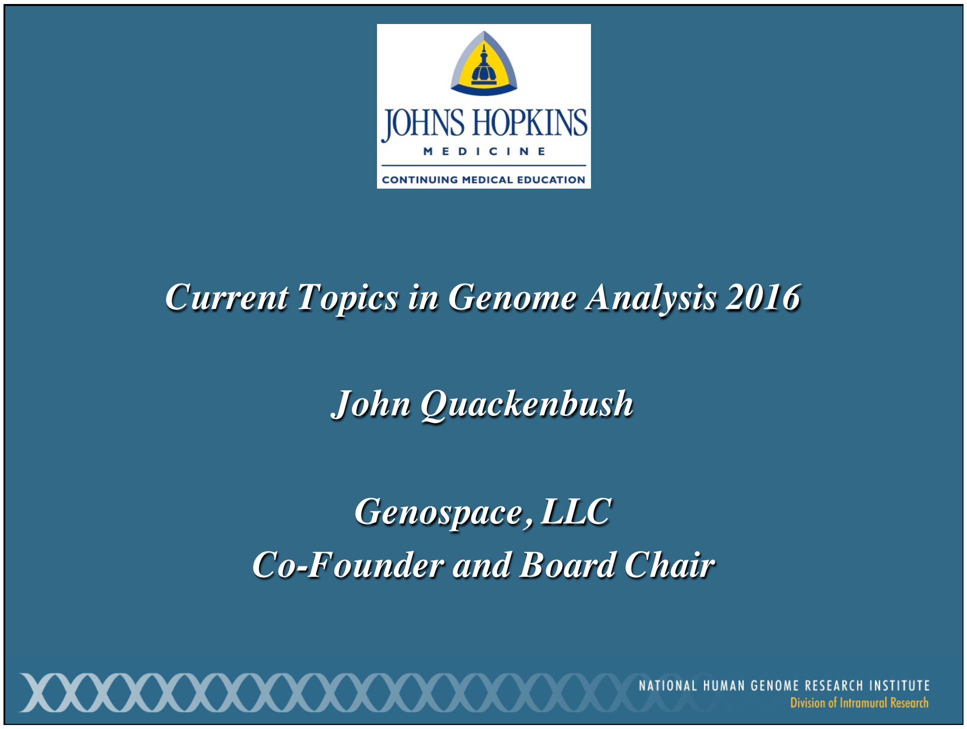
EXPRESSION ANALYSIS, FUNCTIONAL ENRICHMENT, AND NETWORK INFERENCE

John Quackenbush
Dana-Farber Cancer Institute
Harvard T.H. Chan School of Public Health



Background and Disclosures

- Professor of Biostatistics and Computational Biology, Dana-Farber Cancer Institute
- Professor of Computational Biology and Bioinformatics, Harvard School of Public Health
- Many other academic titles
- Numerous advisory boards
- Co-Founder of Genospace, a Precision Genomic Medicine Software Company



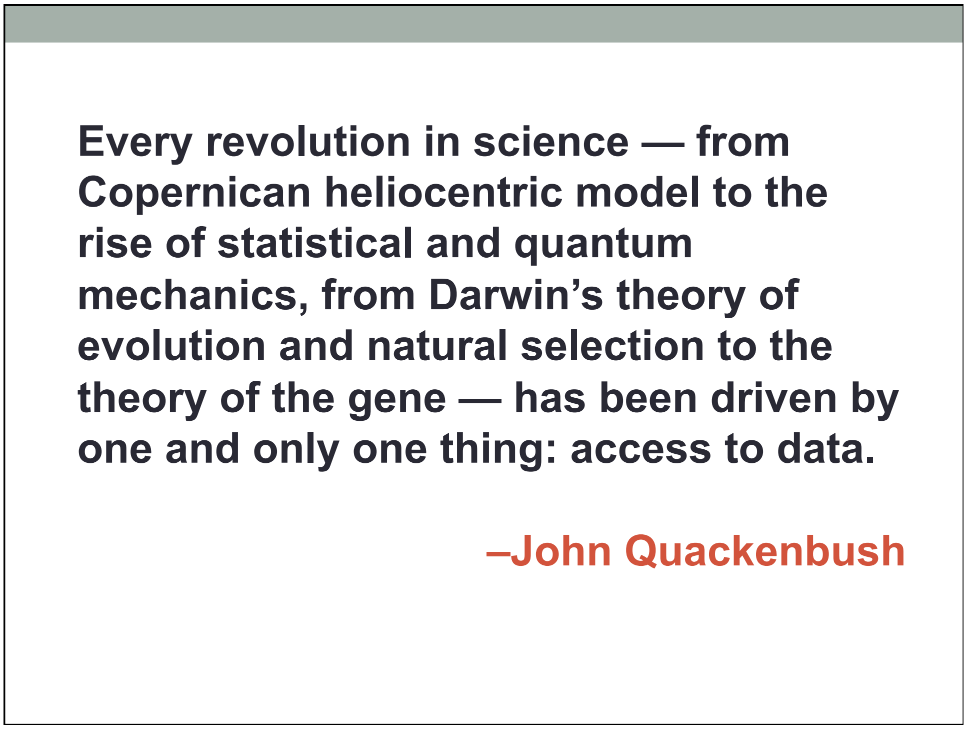
JOHNS HOPKINS
MEDICINE
CONTINUING MEDICAL EDUCATION

Current Topics in Genome Analysis 2016

John Quackenbush

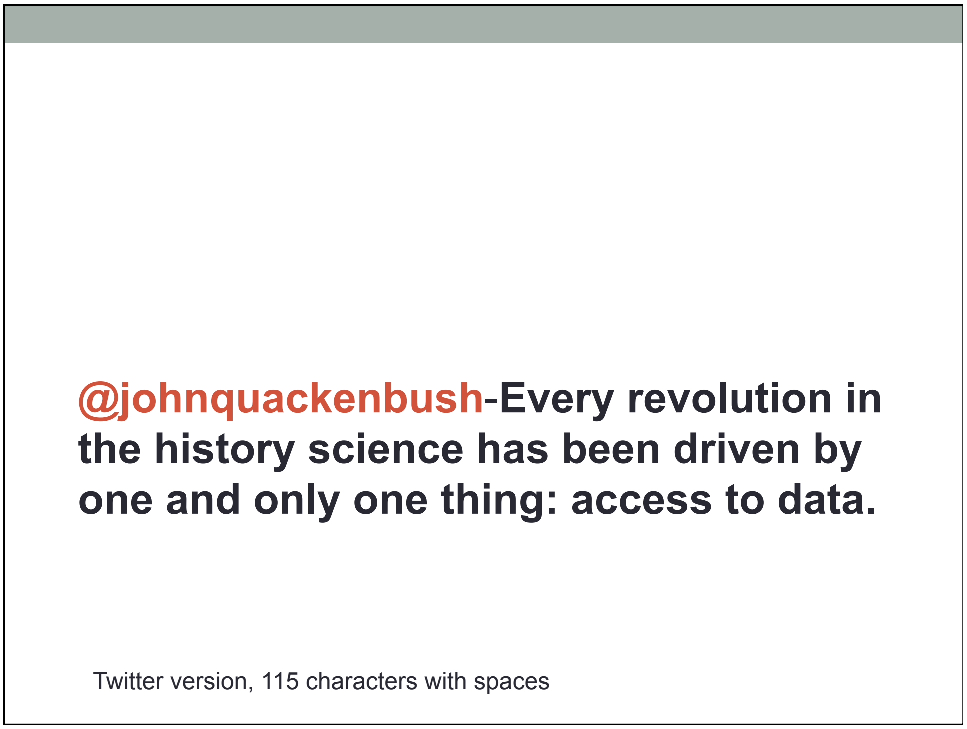
Genospace, LLC
Co-Founder and Board Chair

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



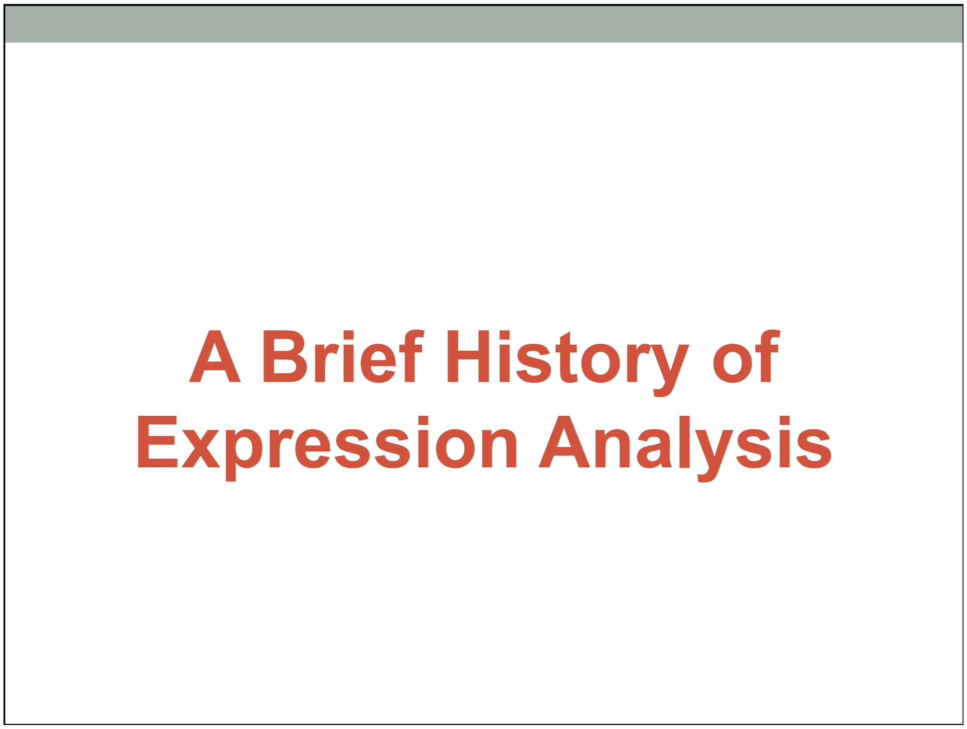
Every revolution in science — from Copernican heliocentric model to the rise of statistical and quantum mechanics, from Darwin's theory of evolution and natural selection to the theory of the gene — has been driven by one and only one thing: access to data.

–John Quackenbush



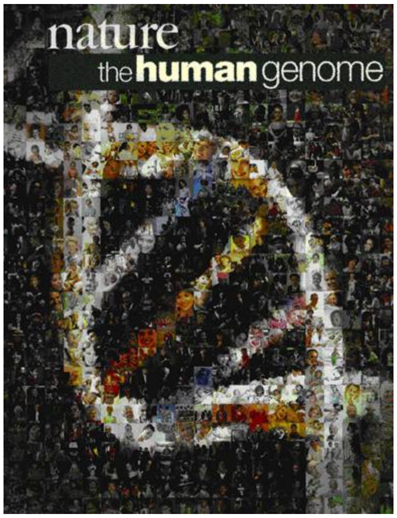
@johnquackenbush-Every revolution in the history science has been driven by one and only one thing: access to data.

Twitter version, 115 characters with spaces

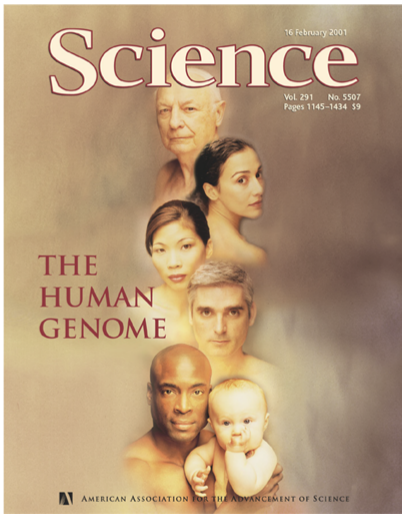


A Brief History of Expression Analysis

February 2001: Completion of the Draft Human Genome

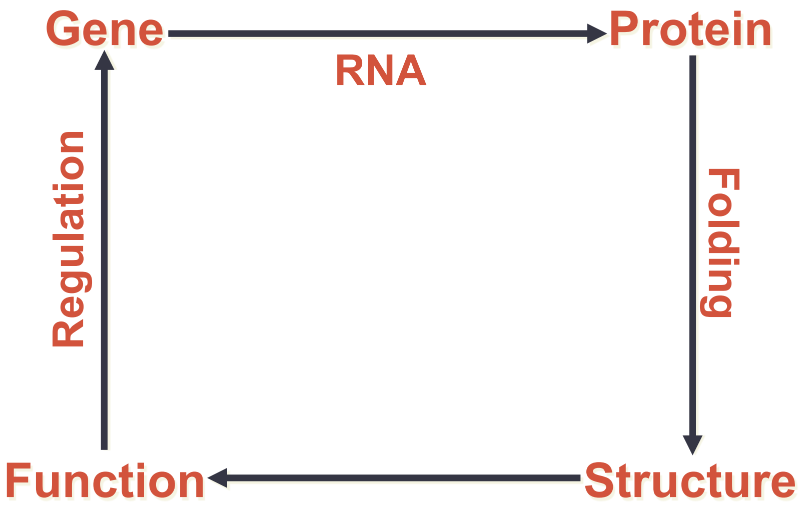


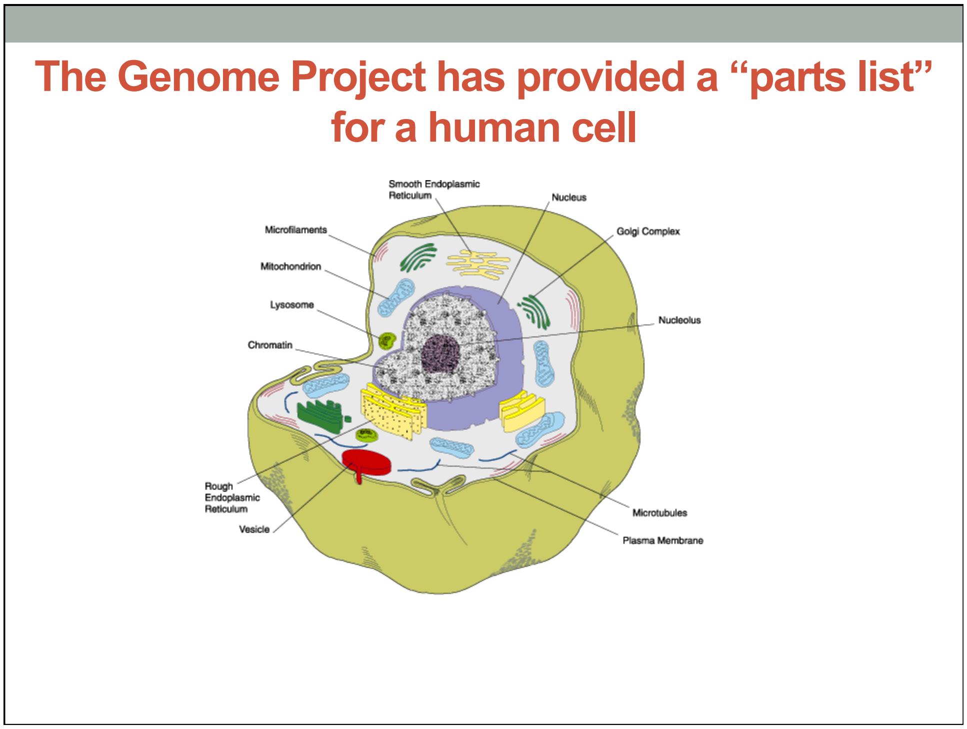
Public HGP



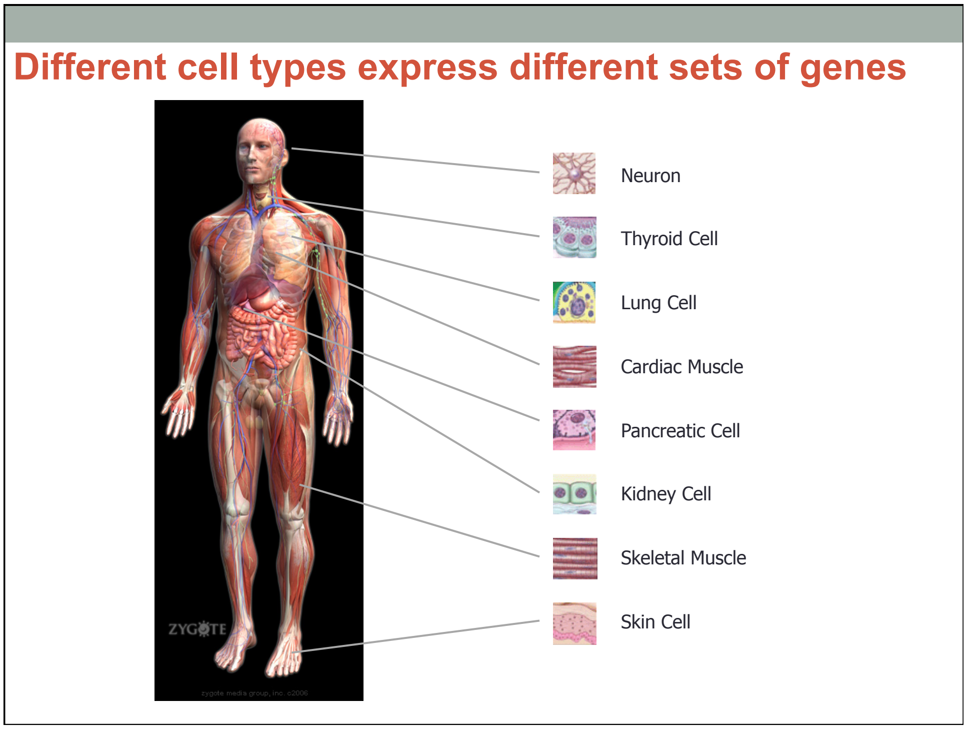
Celera Genomics

Molecular Biology in 7 Words





The Genome Project has provided a "parts list" for a human cell



Different cell types express different sets of genes

Neuron

Thyroid Cell

Lung Cell

Cardiac Muscle

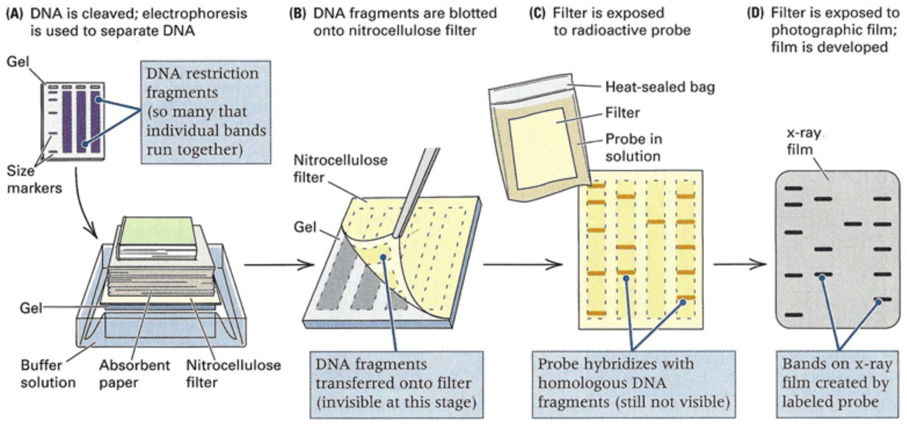
Pancreatic Cell

Kidney Cell

Skeletal Muscle

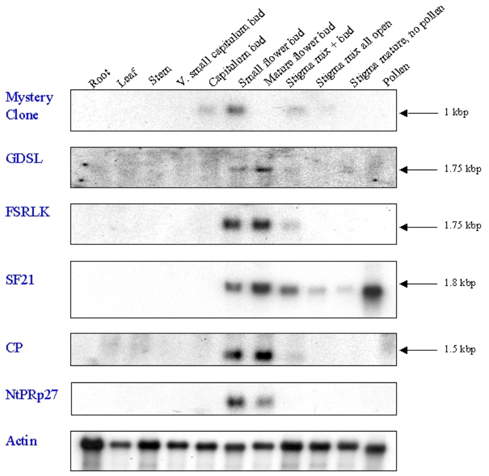
Skin Cell

Northern Blots: Before the Dawn of Time


(A) DNA is cleaved; electrophoresis is used to separate DNA
(B) DNA fragments are blotted onto nitrocellulose filter
(C) Filter is exposed to radioactive probe
(D) Filter is exposed to photographic film; film is developed
Gel
DNA restriction fragments (so many that individual bands run together)
Size markers
Gel
Buffer solution
Absorbent paper
Nitrocellulose filter
Nitrocellulose filter
Gel
DNA fragments transferred onto filter (invisible at this stage)
Heat-sealed bag
Filter
Probe in solution
Probe hybridizes with homologous DNA fragments (still not visible)
x-ray film
Bands on x-ray film created by labeled probe

http://http://www.lookfordiagnosis.com/mesh_info.php?term=Blotting%2C+Northern&lang=1

Northern Blots

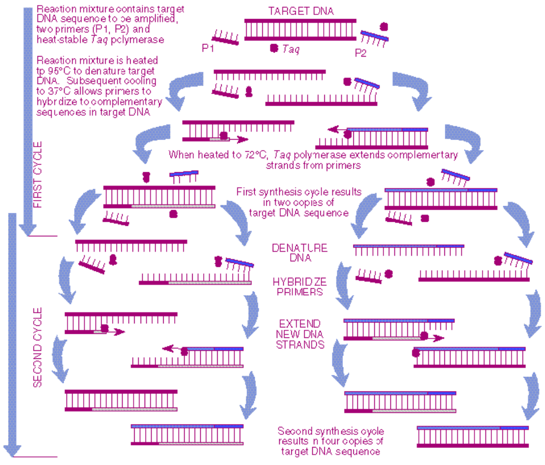


- The size of the amplification product is the filter on the data
- Quantitation requires normalization
- Normalization is based on assumptions

Quantitative RT-PCR: The Ancient World

ORNL-DWG 91M-17476

DNA Amplification Using Polymerase Chain Reaction



Reaction mixture contains target DNA sequence to be amplified, two primers (P1, P2) and heat-stable *Taq* polymerase

Reaction mixture is heated tp 95°C to denature target DNA. Subsequent cooling to 37°C allows primers to hybridize to complementary sequences in target DNA

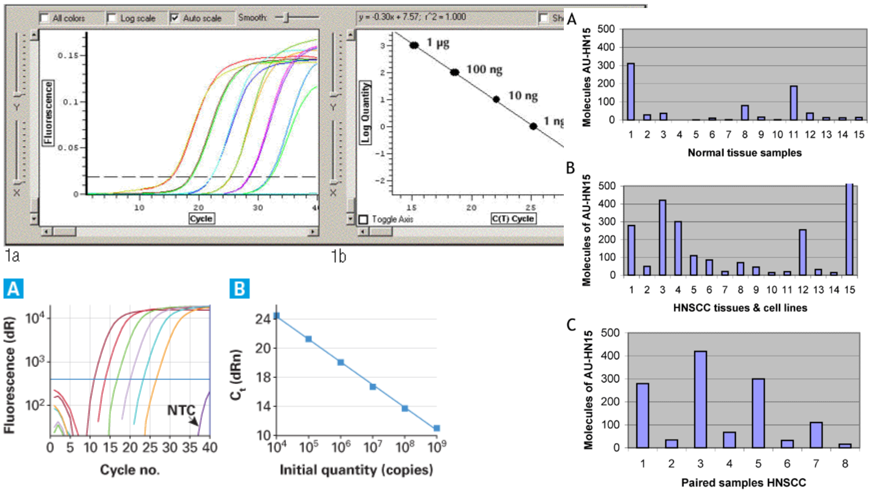
When heated to 72°C, *Taq* polymerase extends complementary strands from primers

First synthesis cycle results in two copies of target DNA sequence

Second synthesis cycle results in four copies of target DNA sequence

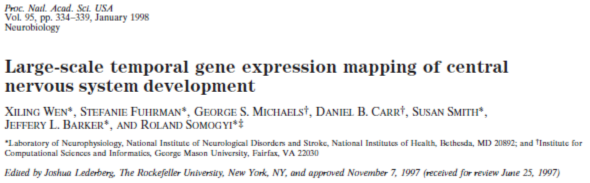
Source: *DNA Science*, see Fig. 13.

Quantitative PCR



- Paired hybridization of two primers is the filter on the data
- Quantitation requires normalization (comparison to standard curves)
- Normalization is based on assumptions

Large-scale Quantitative RT-PCR: The Dawn of the Modern Age



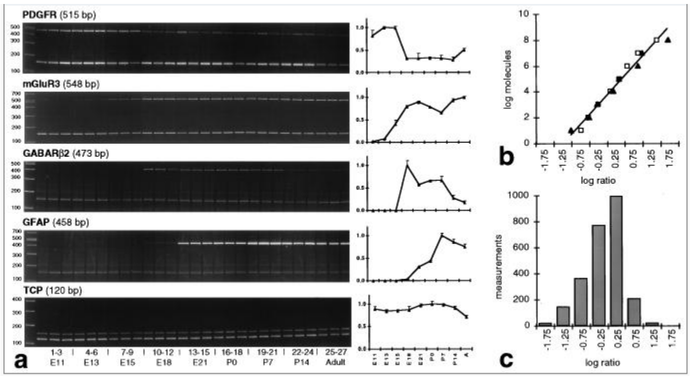
Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 334–339, January 1998
Neurobiology

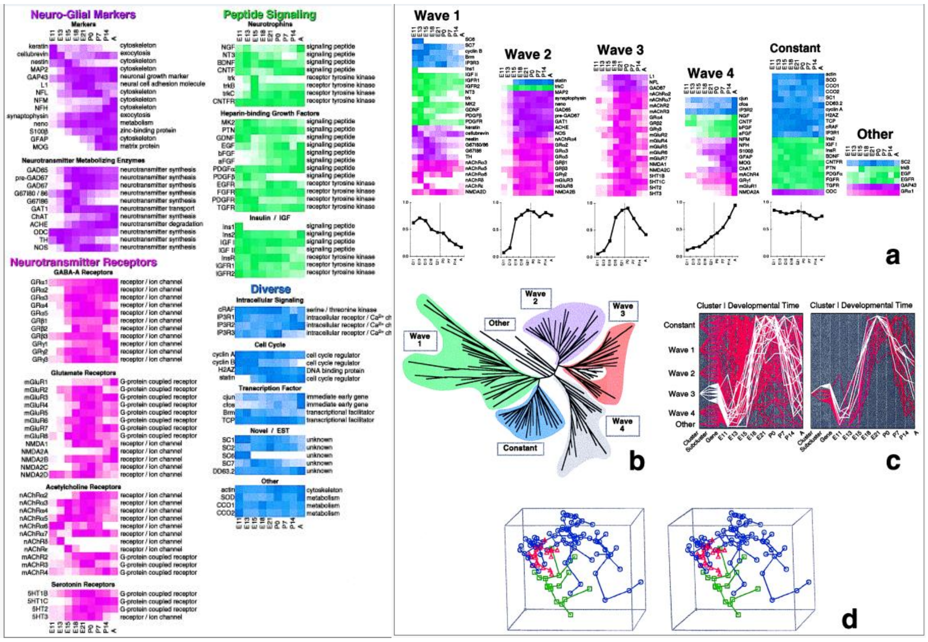
Large-scale temporal gene expression mapping of central nervous system development

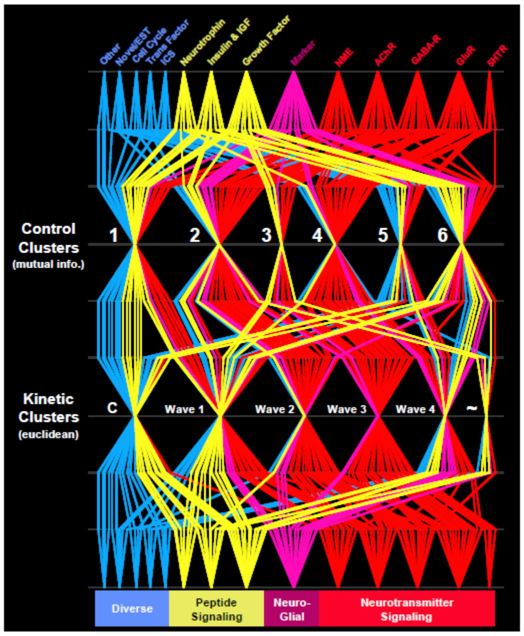
Xiling Wen*, Stefanie Fuhrman*, George S. Michaels†, Daniel B. Carr†, Susan Smith*, Jeffery L. Barker*, and Roland Somogyi*‡

*Laboratory of Neurophysiology, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD 20892; and †Institute for Computational Sciences and Informatics, George Mason University, Fairfax, VA 22030

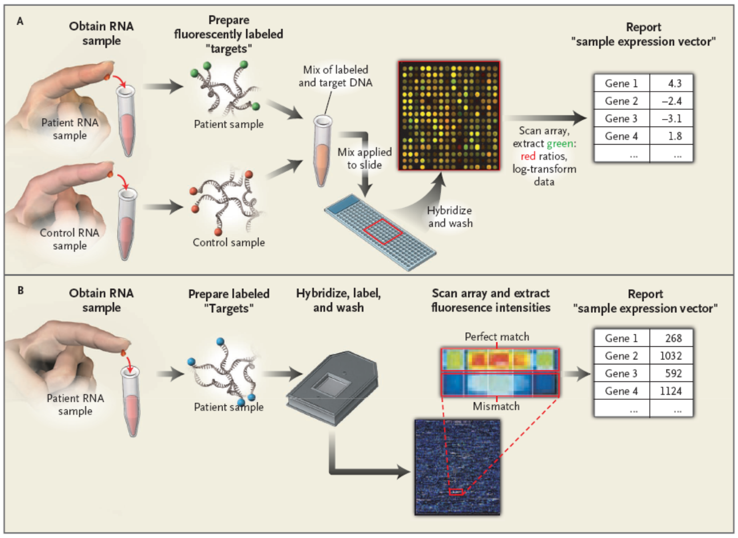
Edited by Joshua Lederberg, The Rockefeller University, New York, NY, and approved November 7, 1997 (received for review June 25, 1997)







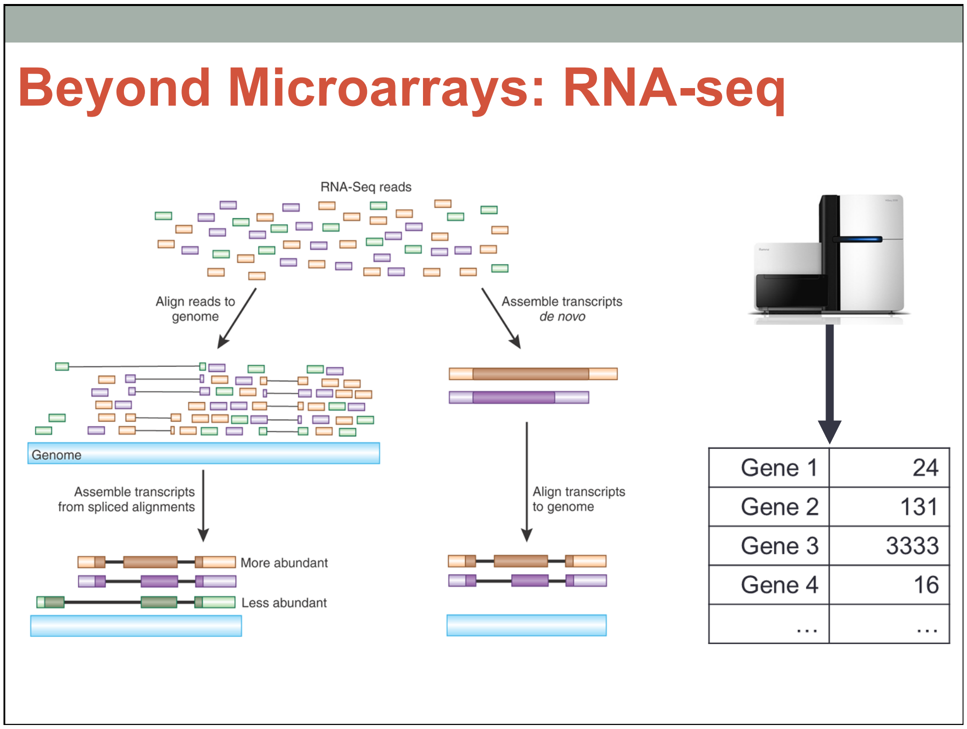
Beyond qRT-PCR: Microarrays


A
Obtain RNA sample
Prepare fluorescently labeled "targets"
Report "sample expression vector"
Patient RNA sample
Patient sample
Mix of labeled and target DNA
Mix applied to slide
Scan array, extract green: red ratios, log-transform data
Control RNA sample
Control sample
Hybridize and wash
B
Obtain RNA sample
Prepare labeled "Targets"
Hybridize, label, and wash
Scan array and extract fluorescence intensities
Report "sample expression vector"
Perfect match
Mismatch
Patient RNA sample
Patient sample

Gene 1	4.3
Gene 2	−2.4
Gene 3	−3.1
Gene 4	1.8
...	...

Gene 1	268
Gene 2	1032
Gene 3	592
Gene 4	1124
...	...

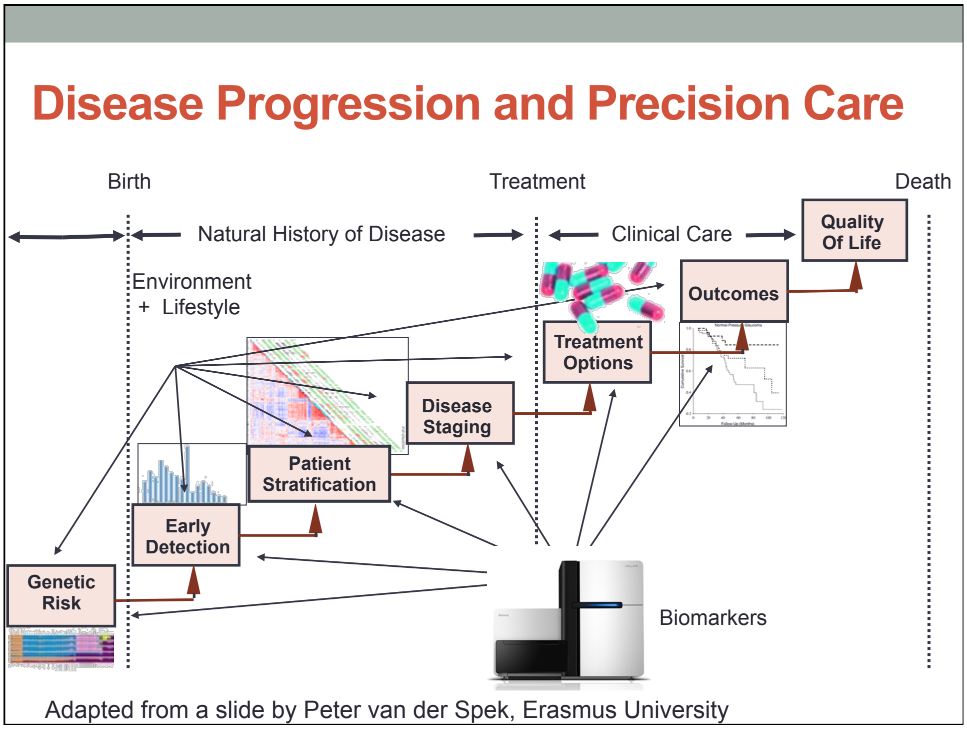
Spatial position is the filter on the data.



Beyond Microarrays: RNA-seq

RNA-Seq reads
Align reads to genome
Assemble transcripts *de novo*
Genome
Assemble transcripts from spliced alignments
Align transcripts to genome
More abundant
Less abundant

Gene 1	24
Gene 2	131
Gene 3	3333
Gene 4	16
…	…



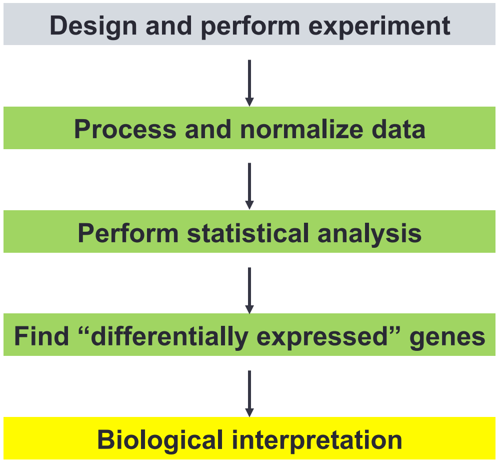
Disease Progression and Precision Care

Birth
Treatment
Death
Natural History of Disease
Clinical Care
Quality Of Life
Environment + Lifestyle
Outcomes
Treatment Options
Disease Staging
Patient Stratification
Early Detection
Genetic Risk
Biomarkers

Adapted from a slide by Peter van der Spek, Erasmus University

Experimental Overview

Expression Analysis Pipeline: Microarrays



Design and perform experiment

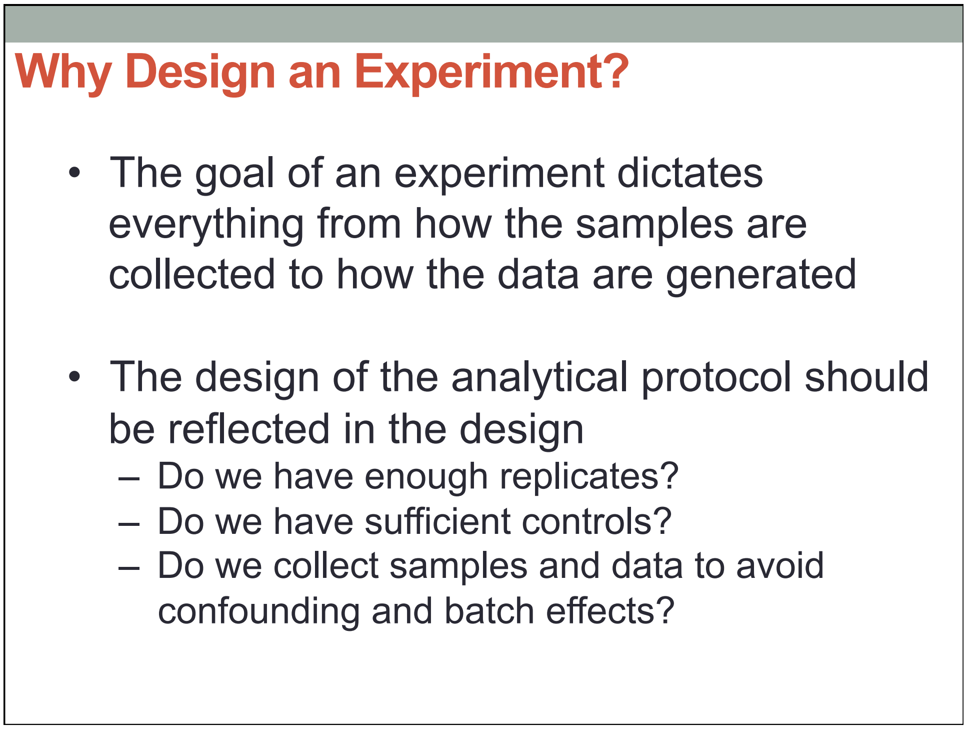
Process and normalize data

Perform statistical analysis

Find “differentially expressed” genes

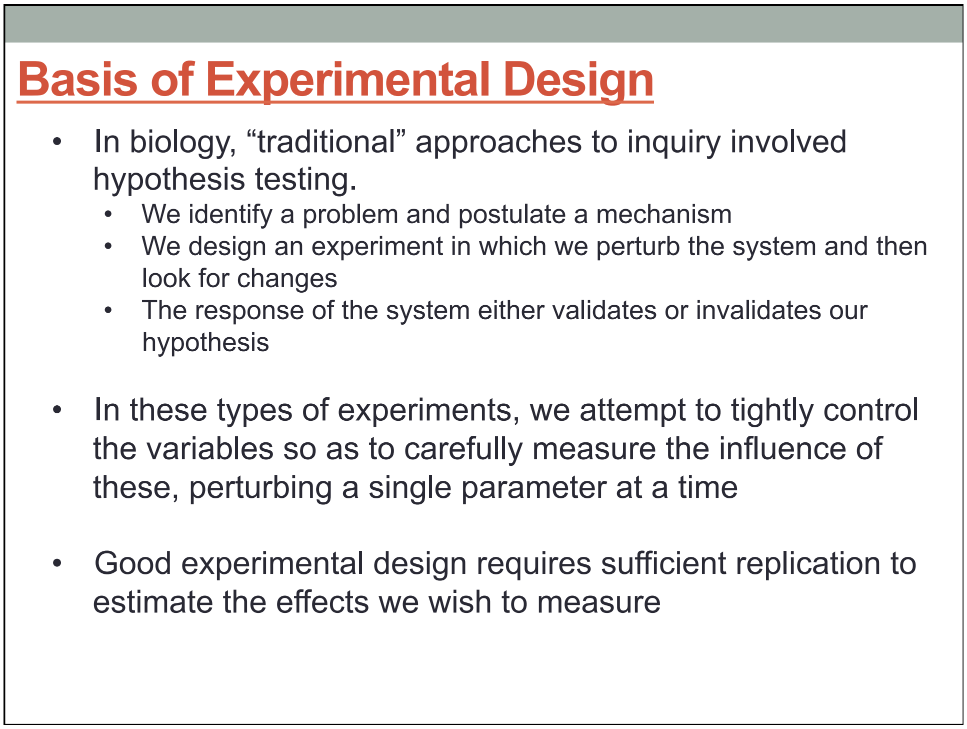
Biological interpretation

Design the Experiment



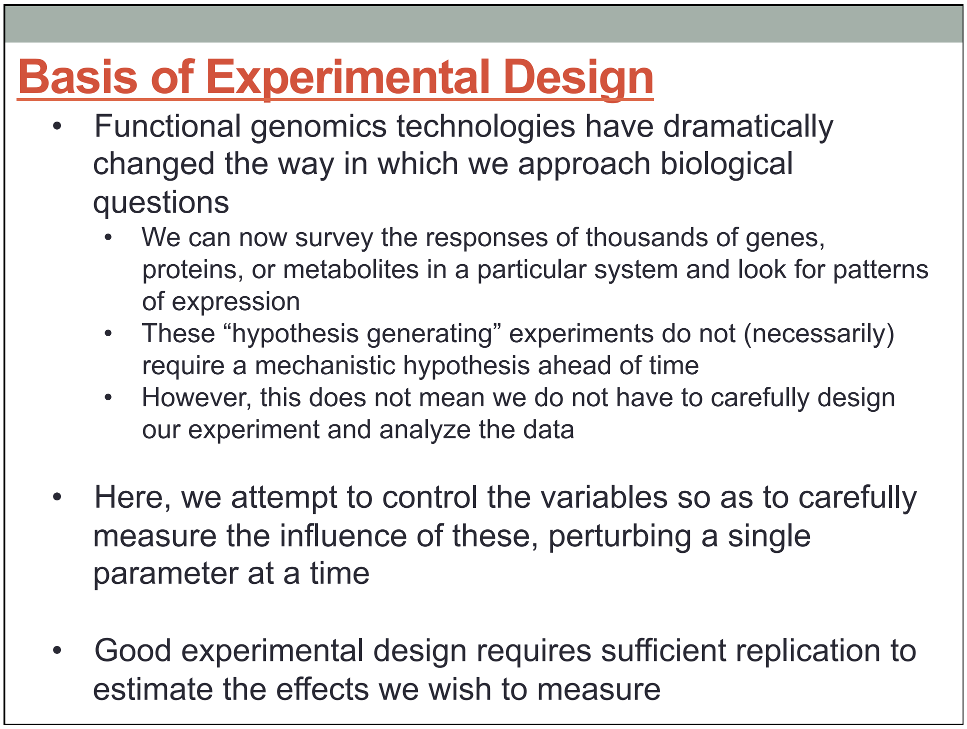
Why Design an Experiment?

- The goal of an experiment dictates everything from how the samples are collected to how the data are generated
- The design of the analytical protocol should be reflected in the design
  - Do we have enough replicates?
  - Do we have sufficient controls?
  - Do we collect samples and data to avoid confounding and batch effects?



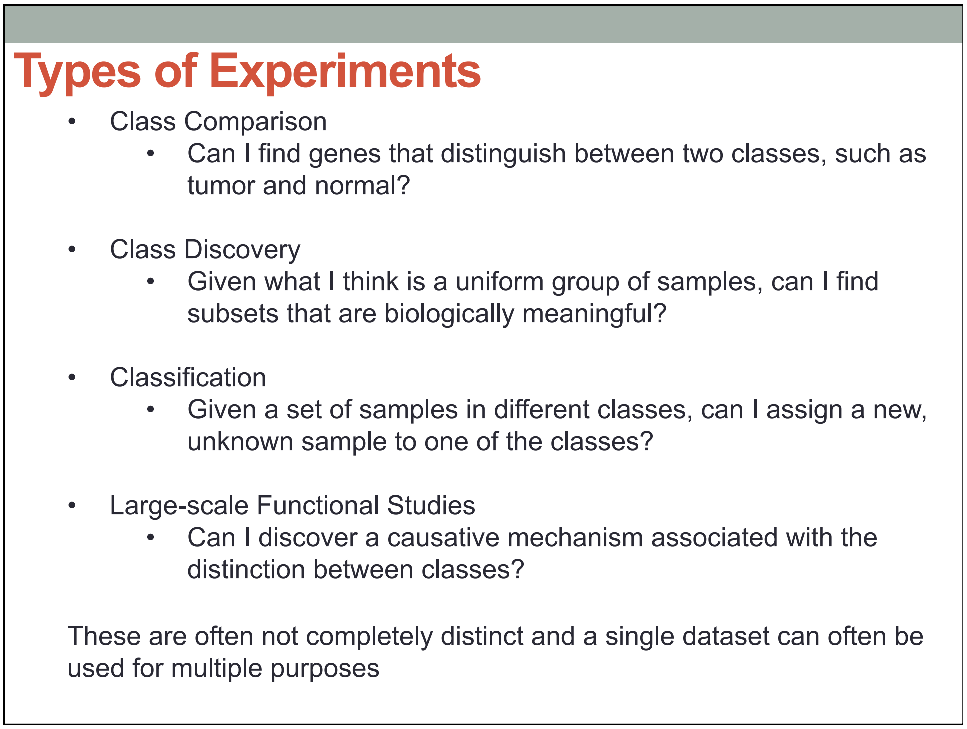
Basis of Experimental Design

- In biology, "traditional" approaches to inquiry involved hypothesis testing.
  - We identify a problem and postulate a mechanism
  - We design an experiment in which we perturb the system and then look for changes
  - The response of the system either validates or invalidates our hypothesis
- In these types of experiments, we attempt to tightly control the variables so as to carefully measure the influence of these, perturbing a single parameter at a time
- Good experimental design requires sufficient replication to estimate the effects we wish to measure



Basis of Experimental Design

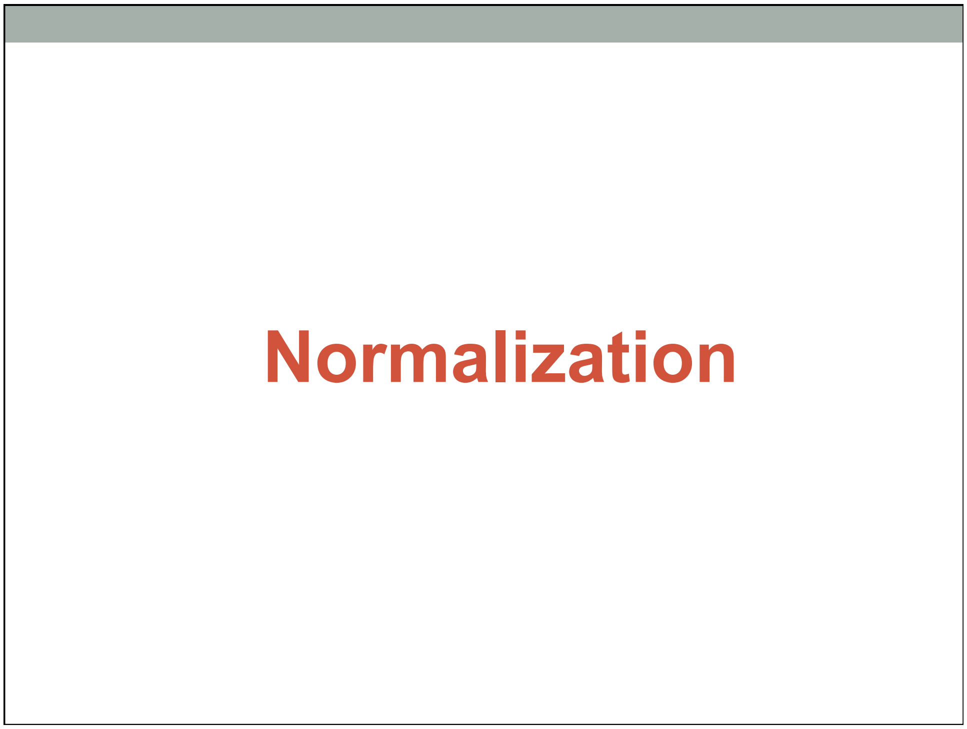
- Functional genomics technologies have dramatically changed the way in which we approach biological questions
  - We can now survey the responses of thousands of genes, proteins, or metabolites in a particular system and look for patterns of expression
  - These "hypothesis generating" experiments do not (necessarily) require a mechanistic hypothesis ahead of time
  - However, this does not mean we do not have to carefully design our experiment and analyze the data
- Here, we attempt to control the variables so as to carefully measure the influence of these, perturbing a single parameter at a time
- Good experimental design requires sufficient replication to estimate the effects we wish to measure



Types of Experiments

- Class Comparison
  - Can I find genes that distinguish between two classes, such as tumor and normal?
- Class Discovery
  - Given what I think is a uniform group of samples, can I find subsets that are biologically meaningful?
- Classification
  - Given a set of samples in different classes, can I assign a new, unknown sample to one of the classes?
- Large-scale Functional Studies
  - Can I discover a causative mechanism associated with the distinction between classes?

These are often not completely distinct and a single dataset can often be used for multiple purposes



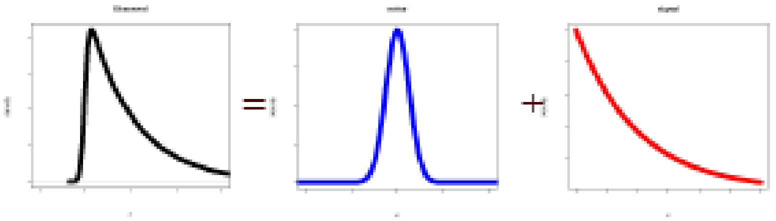
Normalization

Why Normalize Data?

- The goal of normalization is to remove systematic variation from the data and scale it so that comparisons can be made across studies

RMA Background correction

- Expression= Background ($N(0,\sigma^2)$) + Signal ($Exp(\alpha)$))

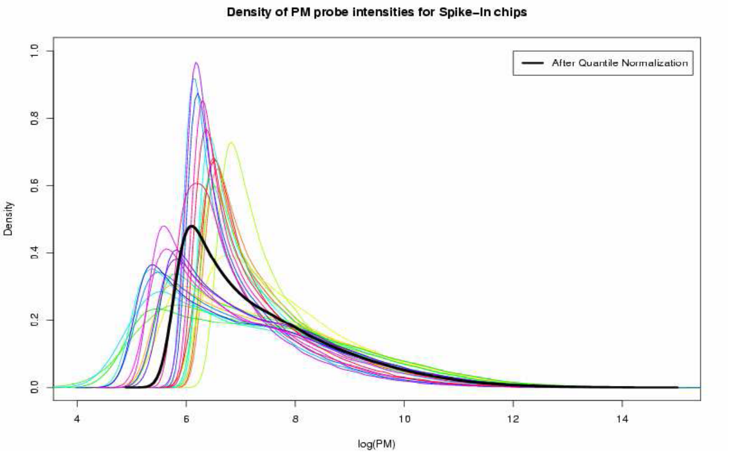


RMA Normalization

- Force the empirical distribution of probe intensities to be the same for every chip in an experiment
- The common distribution is obtained by averaging each *quantile* across chips:

  Quantile Normalization

One distribution for all arrays: the black curve

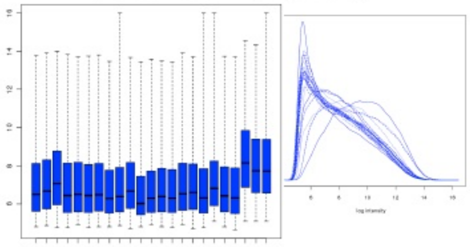


RMA: Probe set summary

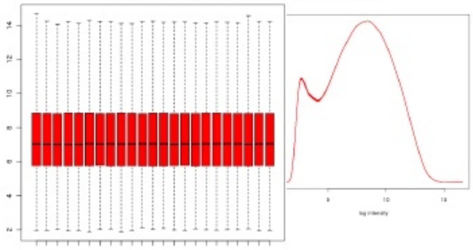
- Robustly fit a two-way model yielding an estimate of $\log_2$(signal) for each probe set
- Fit may be by
  - median polish (quick) or by
  - Mestimation (slower but yields standard errors and good quality
- RMA reduces variability without loosing the ability to detect differential expression

RMA: Before and After

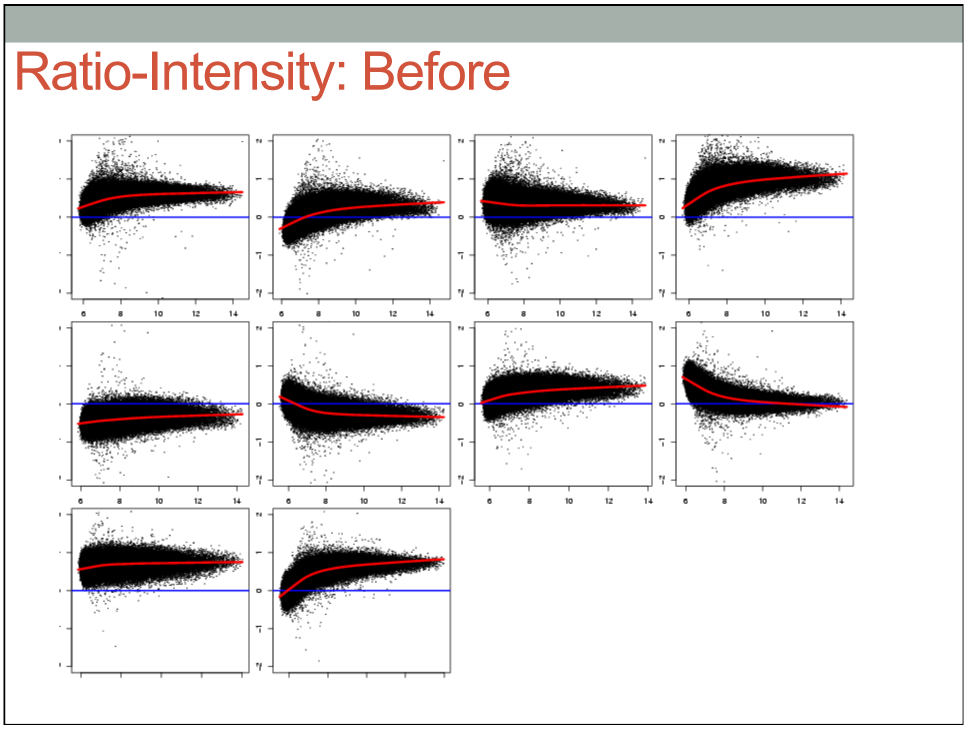
Boxplot and histogram of signal intensities before RMA pre-processing



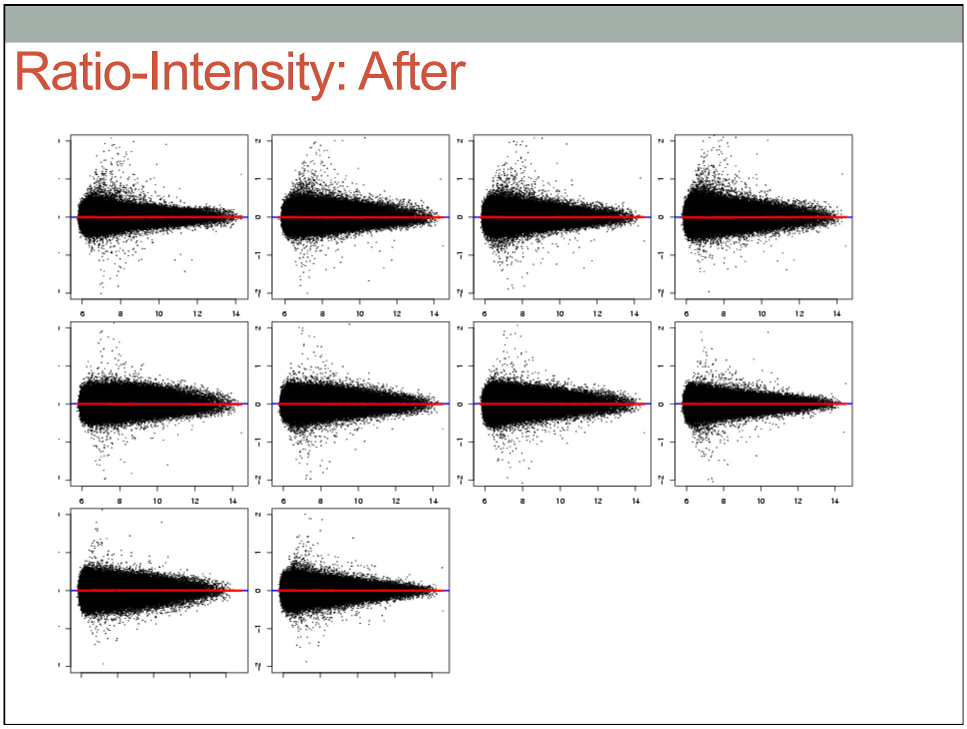
Boxplot and histogram of signal intensities after RMA pre-processing



http://www.slideshare.net/wjjessen/covance-talk



Ratio-Intensity: Before

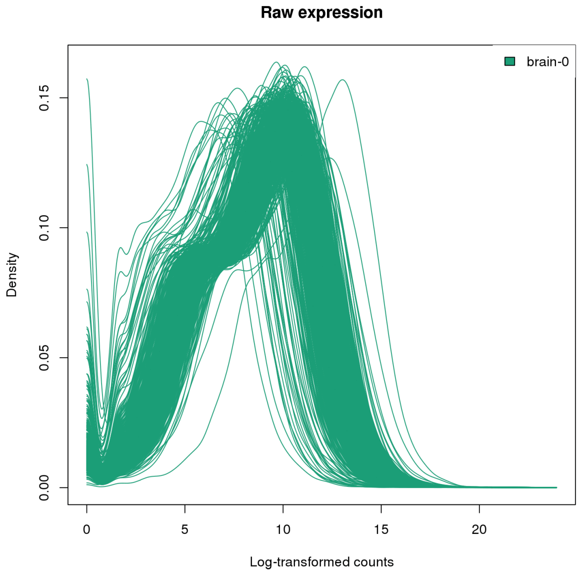


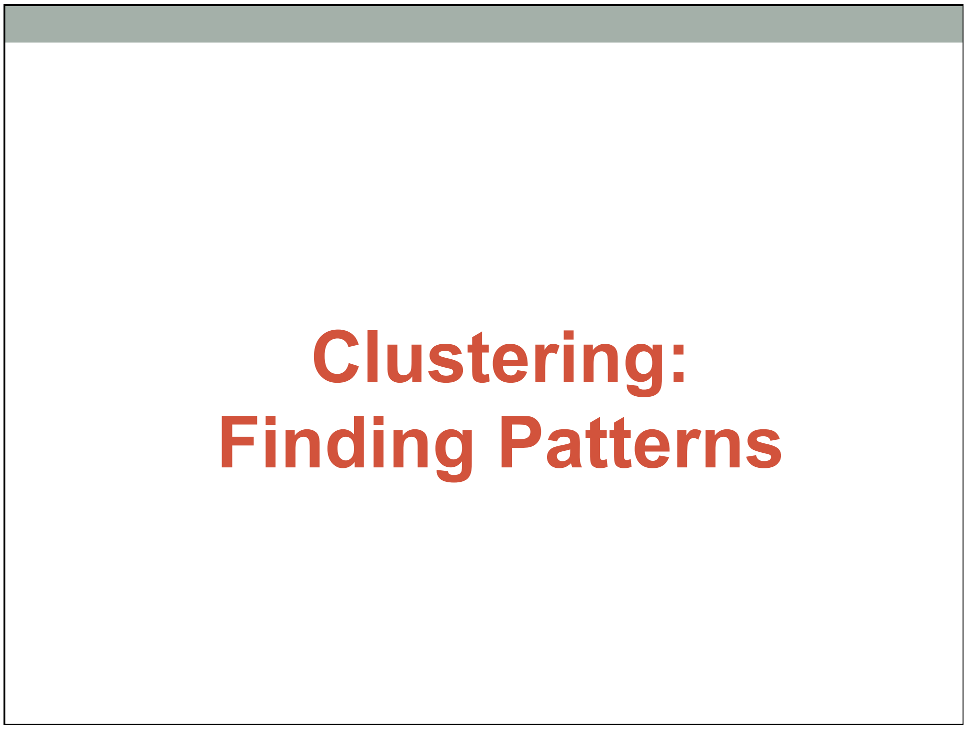
Ratio-Intensity: After

Normalization

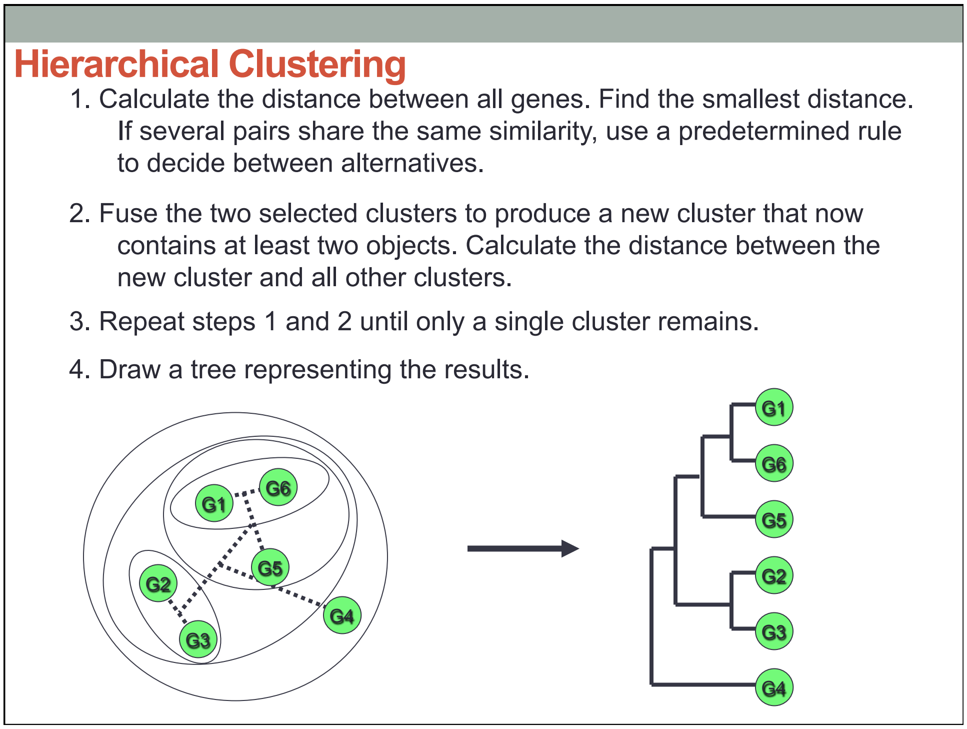
- There are many, many methods
- All attempt to do the same thing, but all have their own assumptions that may or may not be violated
- RMA is widely accepted as the standard for microarrays
- There is less consensus on what works best for RNA-seq
- We constantly have to test our assumptions, even with normalization

GTEx: Complex data requires complex methods





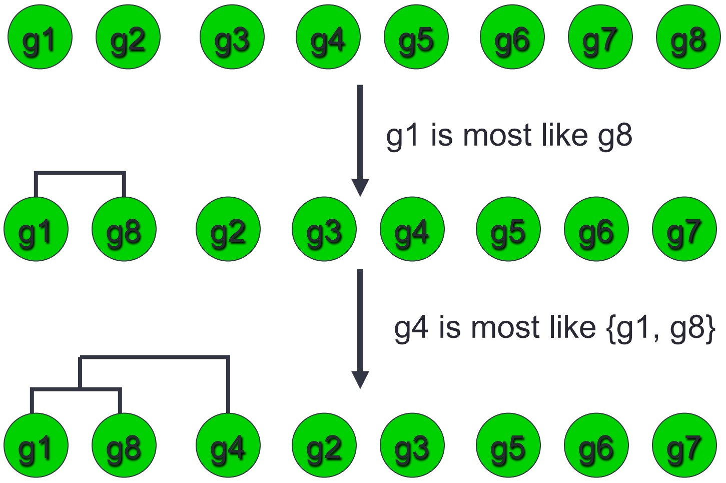
Clustering: Finding Patterns



Hierarchical Clustering

1. Calculate the distance between all genes. Find the smallest distance. If several pairs share the same similarity, use a predetermined rule to decide between alternatives.
2. Fuse the two selected clusters to produce a new cluster that now contains at least two objects. Calculate the distance between the new cluster and all other clusters.
3. Repeat steps 1 and 2 until only a single cluster remains.
4. Draw a tree representing the results.

Hierarchical Clustering

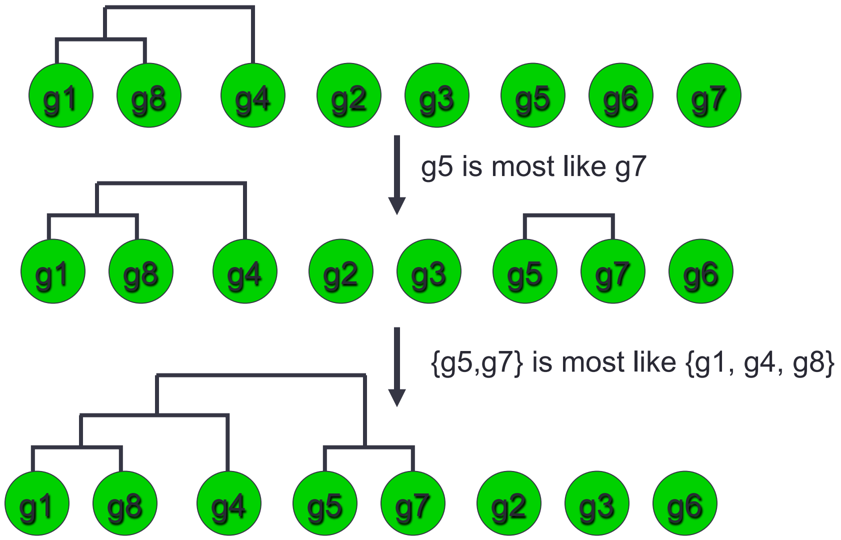


g1 is most like g8

g4 is most like {g1, g8}

(HCL2)

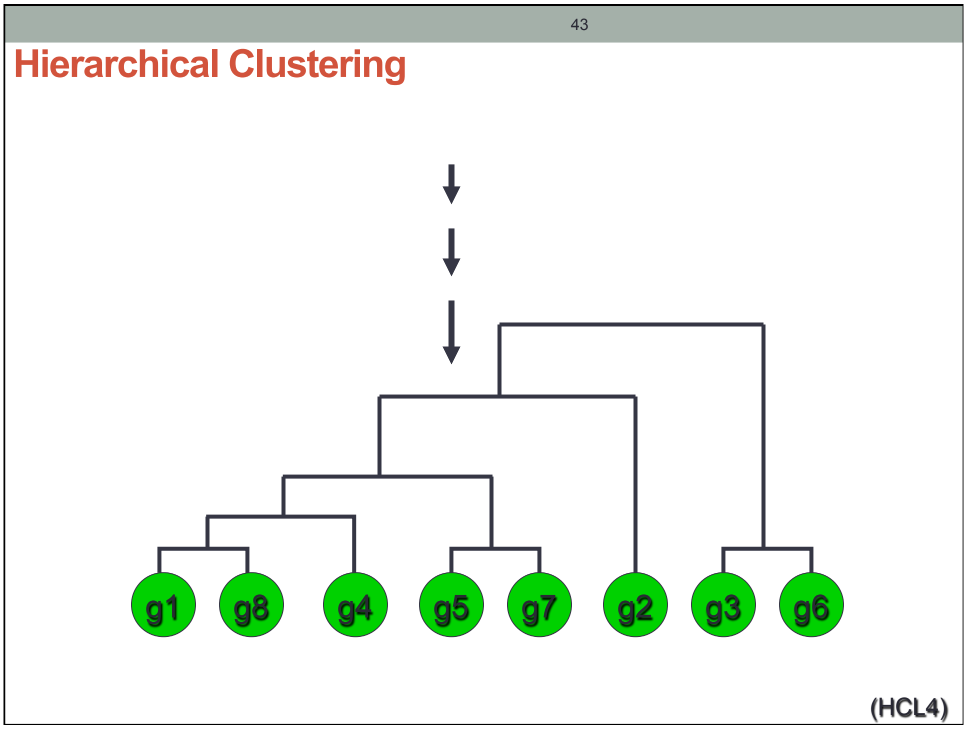
Hierarchical Clustering



g5 is most like g7

{g5,g7} is most like {g1, g4, g8}

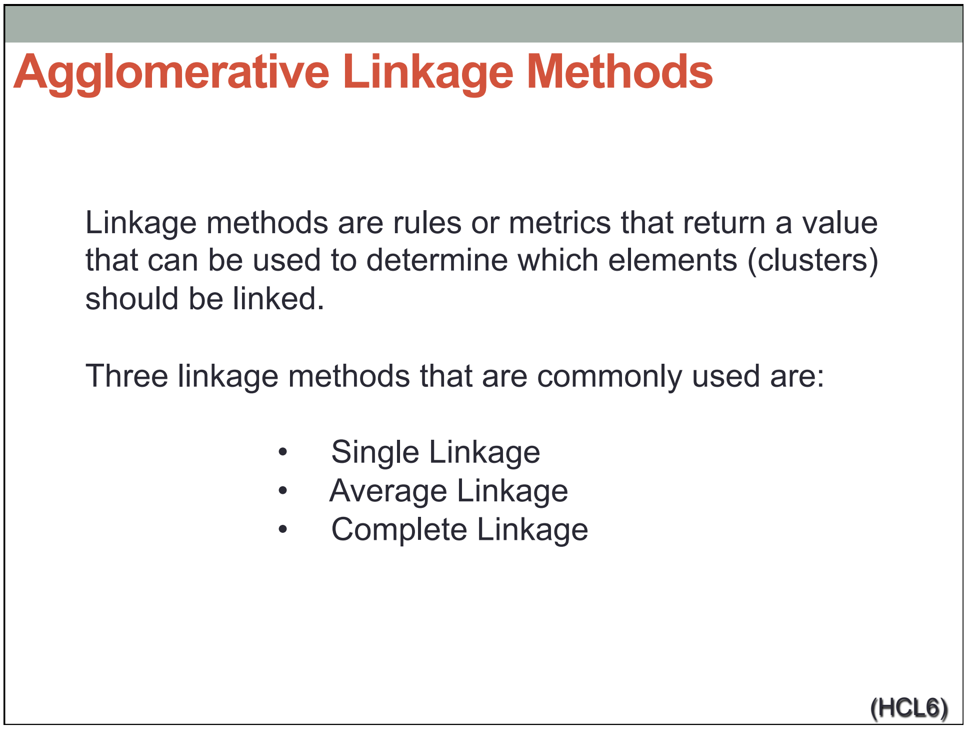
(HCL3)



43

Hierarchical Clustering

(HCL4)



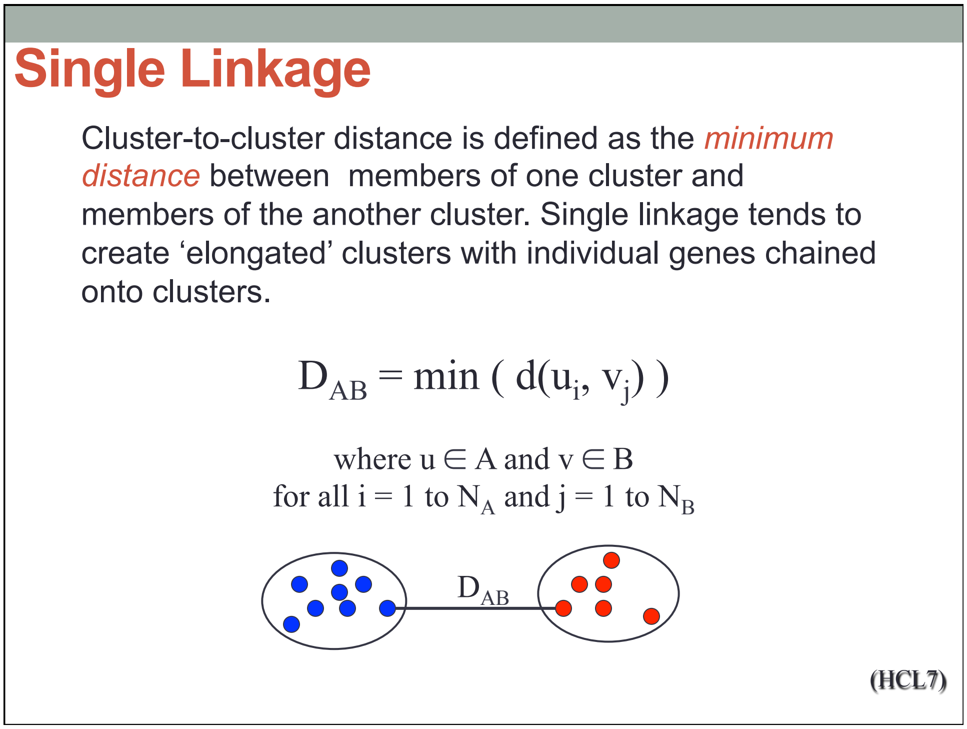
Agglomerative Linkage Methods

Linkage methods are rules or metrics that return a value that can be used to determine which elements (clusters) should be linked.

Three linkage methods that are commonly used are:

- Single Linkage
- Average Linkage
- Complete Linkage

(HCL6)



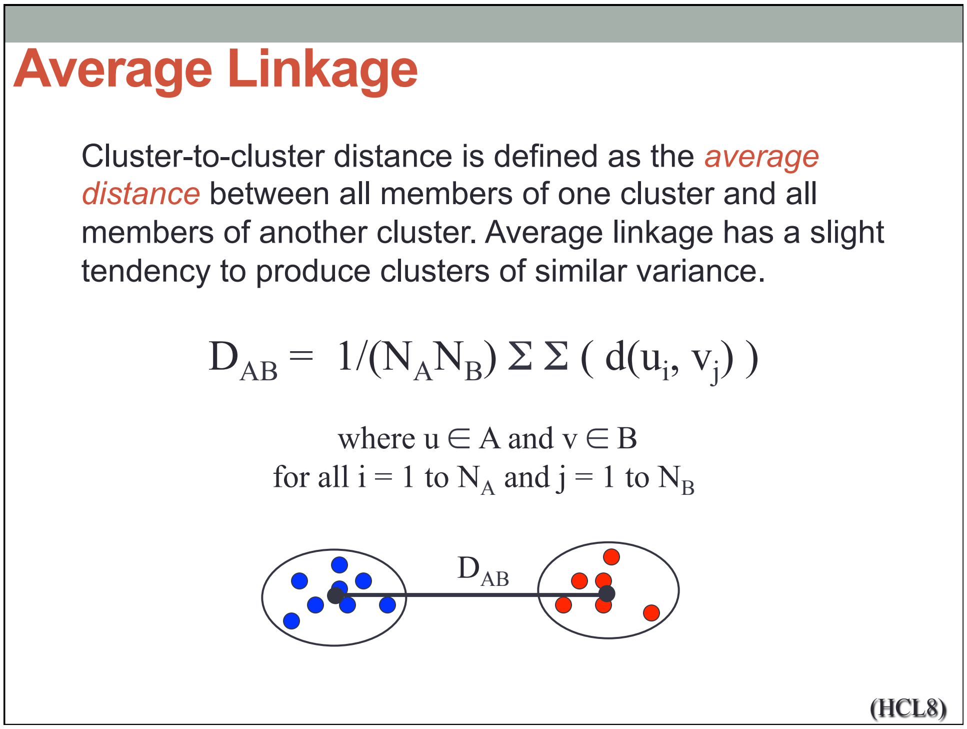
Single Linkage

Cluster-to-cluster distance is defined as the *minimum distance* between members of one cluster and members of the another cluster. Single linkage tends to create ‘elongated’ clusters with individual genes chained onto clusters.

$$D_{AB} = \min (d(u_i, v_j))$$

where $u \in A$ and $v \in B$
for all $i = 1$ to N_A and $j = 1$ to N_B

(HCL7)



Average Linkage

Cluster-to-cluster distance is defined as the *average distance* between all members of one cluster and all members of another cluster. Average linkage has a slight tendency to produce clusters of similar variance.

$$D_{AB} = 1/(N_A N_B) \Sigma \Sigma (d(u_i, v_j))$$

where $u \in A$ and $v \in B$
for all $i = 1$ to N_A and $j = 1$ to N_B

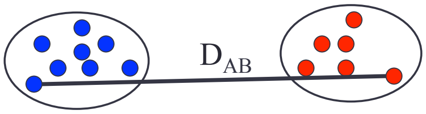
(HCL8)

Complete Linkage

Cluster-to-cluster distance is defined as the *maximum distance* between members of one cluster and members of the another cluster. Complete linkage tends to create clusters of similar size and variability.

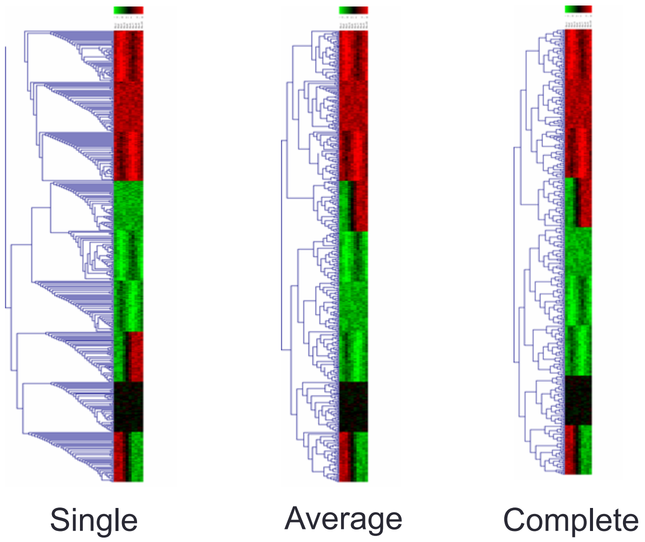
$$D_{AB} = \max\ (\ d(u_i,\ v_j)\ )$$

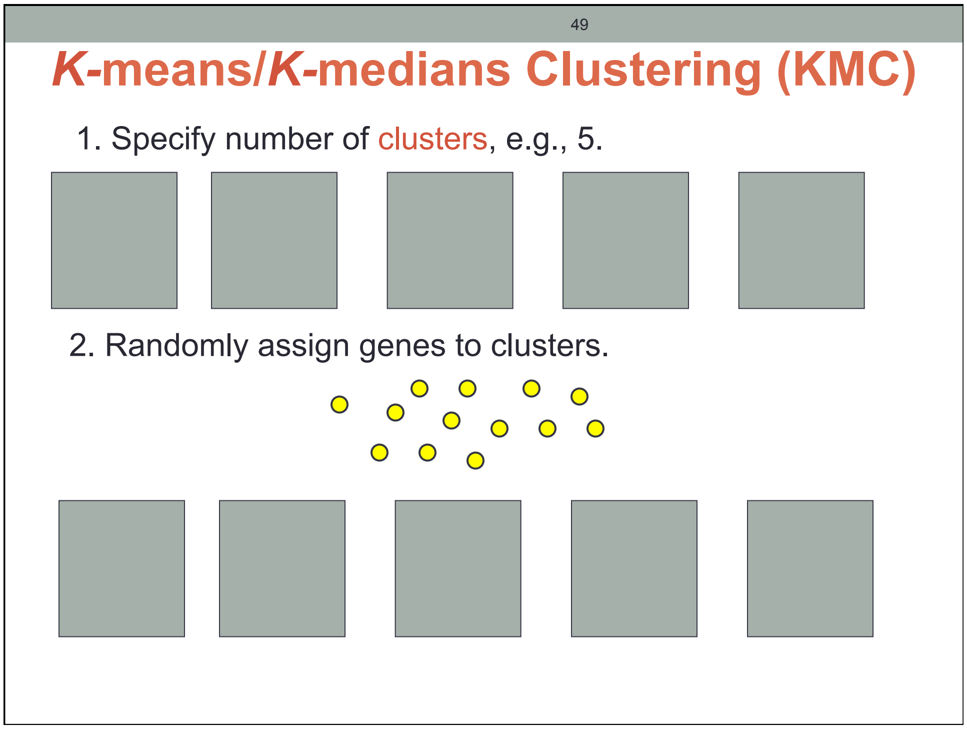
where $u \in A$ and $v \in B$
for all $i = 1$ to N_A and $j = 1$ to N_B



(HCL9)

Comparison of Linkage Methods



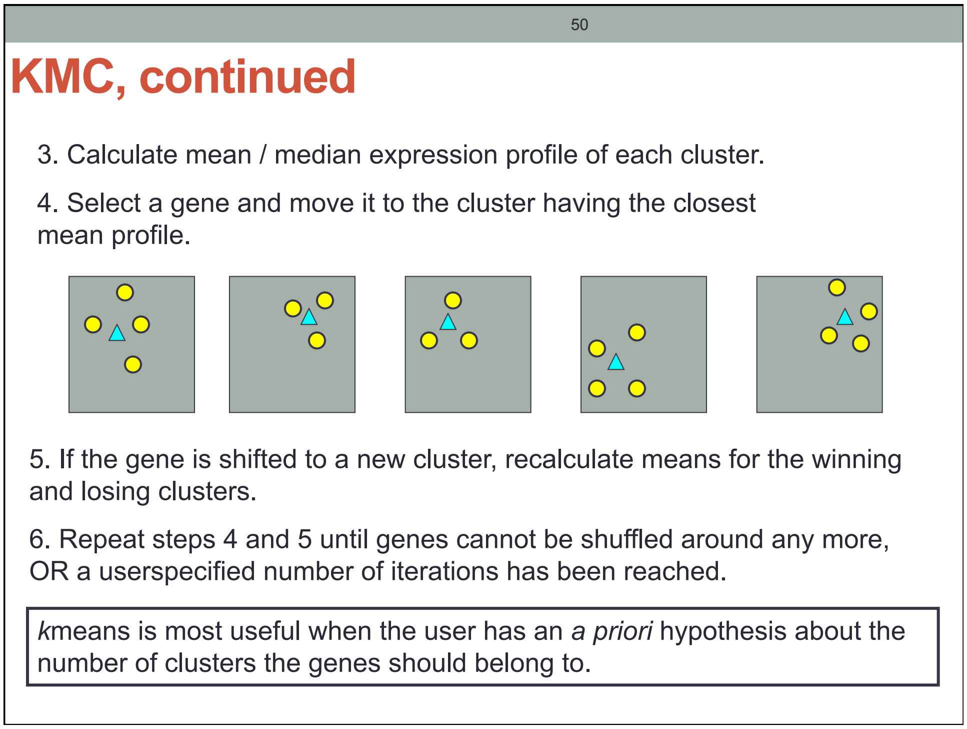


49

K-means/*K*-medians Clustering (KMC)

1. Specify number of clusters, e.g., 5.

2. Randomly assign genes to clusters.



50

KMC, continued

3. Calculate mean / median expression profile of each cluster.

4. Select a gene and move it to the cluster having the closest mean profile.

5. If the gene is shifted to a new cluster, recalculate means for the winning and losing clusters.

6. Repeat steps 4 and 5 until genes cannot be shuffled around any more, OR a userspecified number of iterations has been reached.

*k*means is most useful when the user has an *a priori* hypothesis about the number of clusters the genes should belong to.

Finding Differentially Expressed Genes

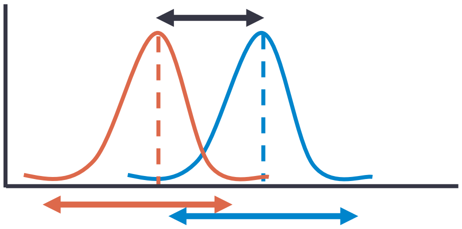
Lies, Damn Lies, and Statistics

Finding Significant Genes

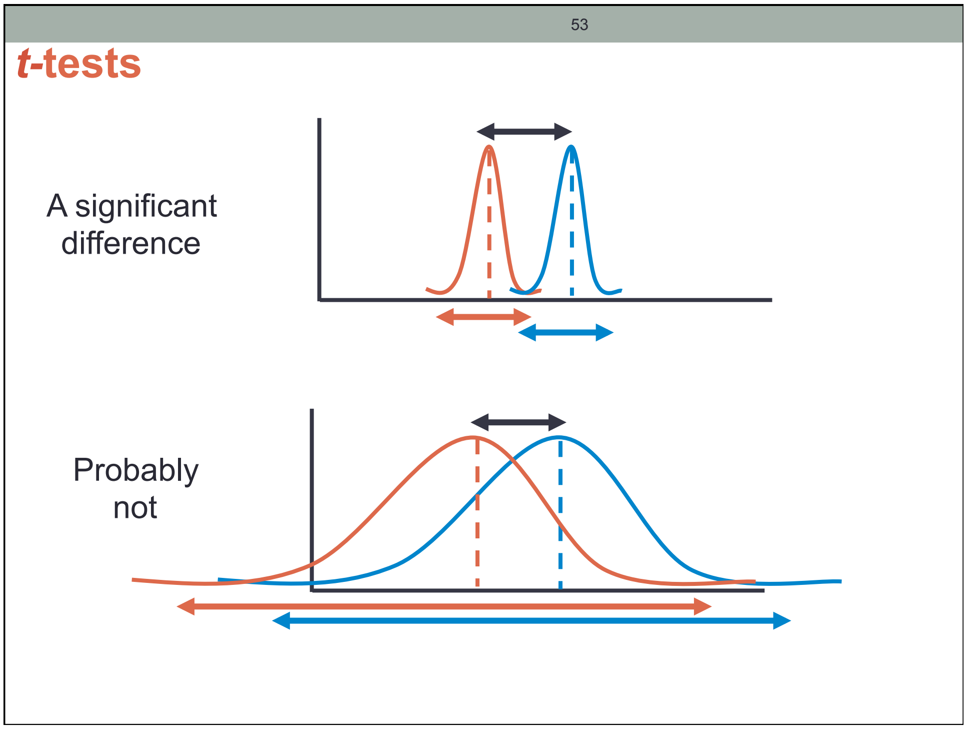
t-test for each gene

- Tests whether the difference between the mean of the query and reference groups are the same
- Essentially measures signal-to-noise
- Calculate *p*-value (permutations or distributions)
- May suffer from intensity-dependent effects

$$t = \frac{\text{signal}}{\text{noise}} = \frac{\text{difference between means}}{\text{variability of groups}} = \frac{\text{<Xq> – <Xc>}}{\text{SE(XqXc)}}$$



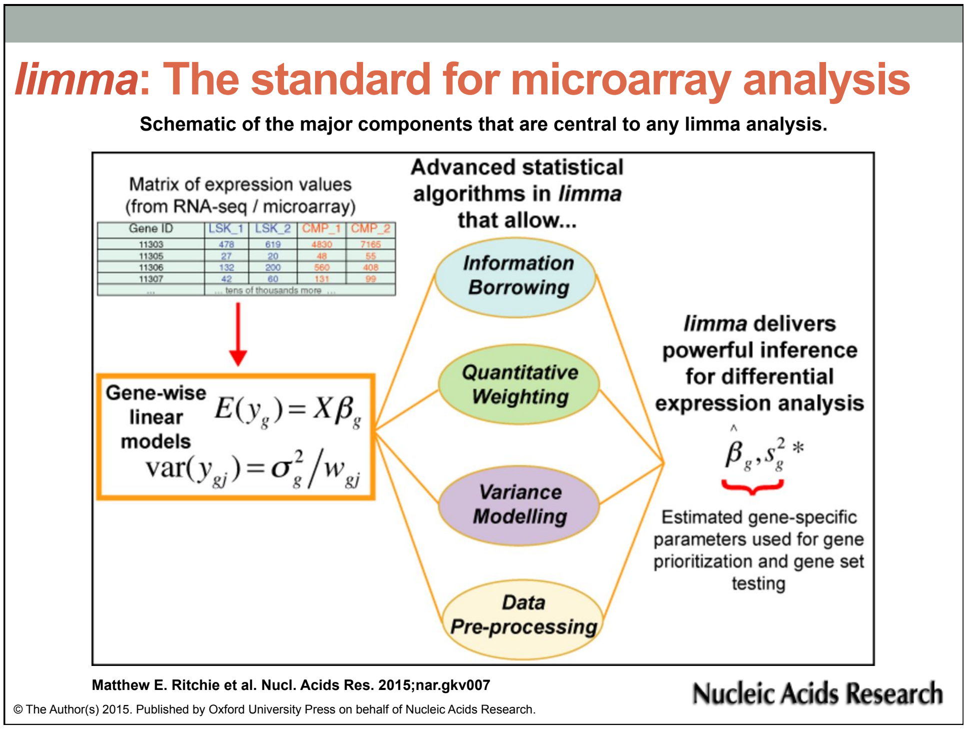
$$t = \frac{\langle Xq \rangle - \langle Xc \rangle}{\sqrt{\frac{\sigma_q^2}{n_q} + \frac{\sigma_c^2}{n_c}}}$$



t-tests

A significant difference

Probably not



limma: The standard for microarray analysis

Schematic of the major components that are central to any limma analysis.

Matrix of expression values
(from RNA-seq / microarray)

Gene ID	LSK_1	LSK_2	CMP_1	CMP_2
11303	478	619	4830	7165
11305	27	20	48	55
11306	132	200	560	408
11307	42	60	131	99
...	... tens of thousands more ...			

Gene-wise linear models

$$E(y_g) = X\beta_g$$

$$\mathrm{var}(y_{gj}) = \sigma_g^2 / w_{gj}$$

Advanced statistical algorithms in *limma* that allow...

Information Borrowing

Quantitative Weighting

Variance Modelling

Data Pre-processing

limma delivers powerful inference for differential expression analysis

$$\underbrace{\hat{\beta}_g, s_g^2}\,*$$

Estimated gene-specific parameters used for gene prioritization and gene set testing

Matthew E. Ritchie et al. Nucl. Acids Res. 2015;nar.gkv007

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic Acids Research.

Nucleic Acids Research

Biological Interpretation

What do the genes in this list do?

Tell me a story, Grampa

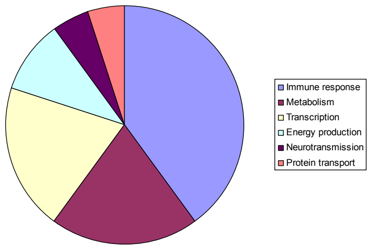
Biological Interpretation

- An obvious way to gain biological insight is to assess the differentially expressed genes in terms of their known function(s)
- Requires an automated and objective (statistical) approach
- Functional profiling or pathway analysis

Early functional analyses

- Manually annotate list of differentially expressed (DE) genes
- Extremely time-consuming, not systematic, user-dependent
- Group together genes with similar function
- Conclude functional categories with most DE genes important in disease/condition under study
- BUT… it may not be the right conclusion
- This is what we call “Biopoetry.”

GO and functional analysis



Functional category	Number of sig genes
Immune response	40
Metabolism	20
Transcription	20
Energy production	10
Neurotransmission	5
Protein transport	5
TOTAL	**100**

Immune response category contains 40% of all significant genes - by far the largest category.

Reasonable to conclude that immune response may be important in the condition being studied?

However …

- What if 40% of the genes on the array were involved in immune response?
- Only detected as many significant immune response genes as expected by chance
- Need to consider not only the number of significant genes for each category, but also total number on the array

Same example, relative to background

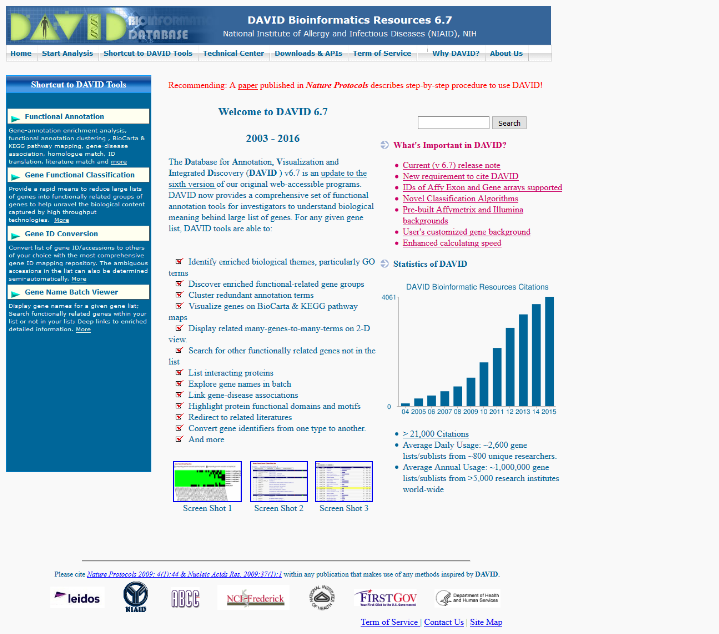
Functional category	Number of genes on array	Observed number of significant genes	Expected number of significant genes
Immune response	8000	40	40
Metabolism	4000	20	20
Transcription	2000	10	10
Energy production	4000	30	20
Neurotransmission	200	5	1
Protein transport	1800	5	9
ALL	20000	100	

Expected number of significant genes for category X is
(num sig genes ÷ total genes on array)*(num genes in category X on array)

Same example, relative to background

Functional category	Number of genes on array	Observed number of significant genes	Expected number of significant genes
Immune response	8000	40	40
Metabolism	4000	20	20
Transcription	2000	10	10
Energy production	4000	30	20
Neurotransmission	200	5	1
Protein transport	1800	5	9
ALL	20000	100	

- Now, energy production and neurotransmission categories appear more interesting as many more significant genes were observed than expected by chance
- Largest categories are not necessarily the most interesting!



DAVID Bioinformatics Resources 6.7
National Institute of Allergy and Infectious Diseases (NIAID), NIH

Home | Start Analysis | Shortcut to DAVID Tools | Technical Center | Downloads & APIs | Term of Service | Why DAVID? | About Us

Shortcut to DAVID Tools

Functional Annotation
Gene-annotation enrichment analysis, functional annotation clustering , BioCarta & KEGG pathway mapping, gene-disease association, homologue match, ID translation, literature match and more

Gene Functional Classification
Provide a rapid means to reduce large lists of genes into functionally related groups of genes to help unravel the biological content captured by high throughput technologies. More

Gene ID Conversion
Convert list of gene ID/accessions to others of your choice with the most comprehensive gene ID mapping repository. The ambiguous accessions in the list can also be determined semi-automatically. More

Gene Name Batch Viewer
Display gene names for a given gene list; Search functionally related genes within your list or not in your list; Deep links to enriched detailed information. More

Recommending: A paper published in ***Nature Protocols*** describes step-by-step procedure to use DAVID!

Welcome to DAVID 6.7

2003 - 2016

The **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery (**DAVID**) v6.7 is an update to the sixth version of our original web-accessible programs. DAVID now provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes. For any given gene list, DAVID tools are able to:

- ☑ Identify enriched biological themes, particularly GO terms
- ☑ Discover enriched functional-related gene groups
- ☑ Cluster redundant annotation terms
- ☑ Visualize genes on BioCarta & KEGG pathway maps
- ☑ Display related many-genes-to-many-terms on 2-D view.
- ☑ Search for other functionally related genes not in the list
- ☑ List interacting proteins
- ☑ Explore gene names in batch
- ☑ Link gene-disease associations
- ☑ Highlight protein functional domains and motifs
- ☑ Redirect to related literatures
- ☑ Convert gene identifiers from one type to another.
- ☑ And more

Screen Shot 1 Screen Shot 2 Screen Shot 3

Search

What's Important in DAVID?

- Current (v 6.7) release note
- New requirement to cite DAVID
- IDs of Affy Exon and Gene arrays supported
- Novel Classification Algorithms
- Pre-built Affymetrix and Illumina backgrounds
- User's customized gene background
- Enhanced calculating speed

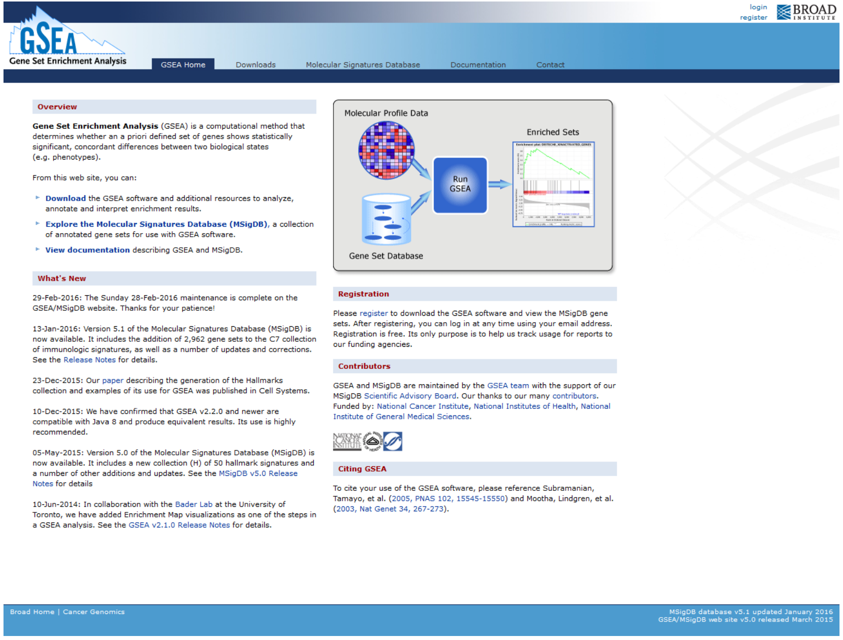
Statistics of DAVID

- > 21,000 Citations
- Average Daily Usage: ~2,600 gene lists/sublists from ~800 unique researchers.
- Average Annual Usage: ~1,000,000 gene lists/sublists from >5,000 research institutes world-wide

Please cite *Nature Protocols 2009; 4(1):44 & Nucleic Acids Res. 2009;37(1):1* within any publication that makes use of any methods inspired by **DAVID**.

Term of Service | Contact Us | Site Map

https://david.ncifcrf.gov/



Overview

Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes).

From this web site, you can:

- **Download** the GSEA software and additional resources to analyze, annotate and interpret enrichment results.
- **Explore the Molecular Signatures Database (MSigDB)**, a collection of annotated gene sets for use with GSEA software.
- **View documentation** describing GSEA and MSigDB.

What's New

29-Feb-2016: The Sunday 28-Feb-2016 maintenance is complete on the GSEA/MSigDB website. Thanks for your patience!

13-Jan-2016: Version 5.1 of the Molecular Signatures Database (MSigDB) is now available. It includes the addition of 2,962 gene sets to the C7 collection of immunologic signatures, as well as a number of updates and corrections. See the Release Notes for details.

23-Dec-2015: Our paper describing the generation of the Hallmarks collection and examples of its use for GSEA was published in Cell Systems.

10-Dec-2015: We have confirmed that GSEA v2.2.0 and newer are compatible with Java 8 and produce equivalent results. Its use is highly recommended.

05-May-2015: Version 5.0 of the Molecular Signatures Database (MSigDB) is now available. It includes a new collection (H) of 50 hallmark signatures and a number of other additions and updates. See the MSigDB v5.0 Release Notes for details

10-Jun-2014: In collaboration with the Bader Lab at the University of Toronto, we have added Enrichment Map visualizations as one of the steps in a GSEA analysis. See the GSEA v2.1.0 Release Notes for details.

Registration

Please register to download the GSEA software and view the MSigDB gene sets. After registering, you can log in at any time using your email address. Registration is free. Its only purpose is to help us track usage for reports to our funding agencies.

Contributors

GSEA and MSigDB are maintained by the GSEA team with the support of our MSigDB Scientific Advisory Board. Our thanks to our many contributors. Funded by: National Cancer Institute, National Institutes of Health, National Institute of General Medical Sciences.

Citing GSEA

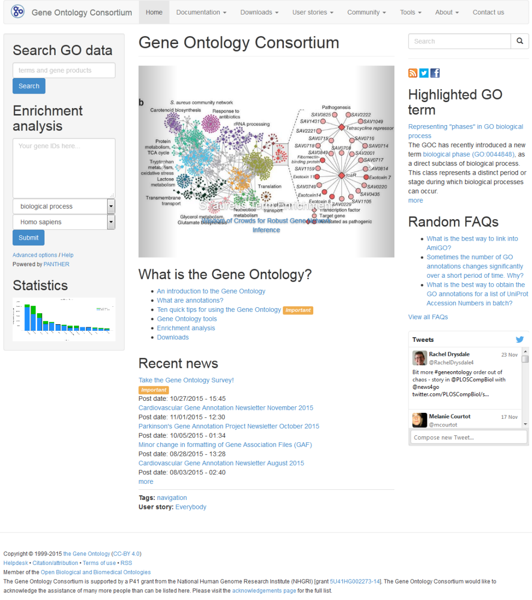
To cite your use of the GSEA software, please reference Subramanian, Tamayo, et al. (2005, PNAS 102, 15545-15550) and Mootha, Lindgren, et al. (2003, Nat Genet 34, 267-273).

Broad Home | Cancer Genomics

MSigDB database v5.1 updated January 2016
GSEA/MSigDB web site v5.0 released March 2015

http://software.broadinstitute.org/gsea/index.jsp

Gene Ontology



Gene Ontology Consortium

Search GO data

Enrichment analysis

Advanced options / Help
Powered by PANTHER

Statistics

What is the Gene Ontology?

- An introduction to the Gene Ontology
- What are annotations?
- Ten quick tips for using the Gene Ontology *Important*
- Gene Ontology tools
- Enrichment analysis
- Downloads

Recent news

Take the Gene Ontology Survey!
Important
Post date: 10/27/2015 - 15:45
Cardiovascular Gene Annotation Newsletter November 2015
Post date: 11/01/2015 - 12:30
Parkinson's Gene Annotation Project Newsletter October 2015
Post date: 10/05/2015 - 01:34
Minor change in formatting of Gene Association Files (GAF)
Post date: 08/28/2015 - 13:28
Cardiovascular Gene Annotation Newsletter August 2015
Post date: 08/03/2015 - 02:40
more

Tags: navigation
User story: Everybody

Highlighted GO term

Representing "phases" in GO biological process
The GOC has recently introduced a new term biological phase (GO:0044848), as a direct subclass of biological process. This class represents a distinct period or stage during which biological processes can occur.
more

Random FAQs

- What is the best way to link into AmiGO?
- Sometimes the number of GO annotations changes significantly over a short period of time. Why?
- What is the best way to obtain the GO annotations for a list of UniProt Accession Numbers in batch?

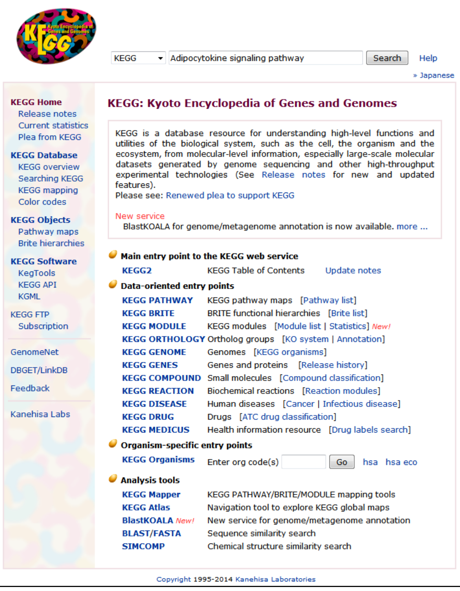
View all FAQs

Rachel Drysdale @RachelDrysdale4 23 Nov
Bit more #geneontology order out of chaos - story in @PLOSCompBiol with @news4go twitter.com/PLOSCompBiol/s...

Melanie Courtot @mcourtot 17 Nov

Copyright © 1999-2015 the Gene Ontology (CC-BY 4.0)
Helpdesk • Citation/attribution • Terms of use • RSS
Member of the Open Biological and Biomedical Ontologies
The Gene Ontology Consortium is supported by a P41 grant from the National Human Genome Research Institute (NHGRI) [grant 5U41HG002273-14]. The Gene Ontology Consortium would like to acknowledge the assistance of many more people than can be listed here. Please visit the acknowledgements page for the full list.

KEGG pathway database



KEGG ▾ | Adipocytokine signaling pathway | Search | Help

» Japanese

KEGG Home
Release notes
Current statistics
Plea from KEGG

KEGG Database
KEGG overview
Searching KEGG
KEGG mapping
Color codes

KEGG Objects
Pathway maps
Brite hierarchies

KEGG Software
KegTools
KEGG API
KGML

KEGG FTP
Subscription

GenomeNet

DBGET/LinkDB

Feedback

Kanehisa Labs

KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies (See Release notes for new and updated features).
Please see: Renewed plea to support KEGG

New service
BlastKOALA for genome/metagenome annotation is now available. more ...

Main entry point to the KEGG web service

KEGG2	KEGG Table of Contents	Update notes

Data-oriented entry points

KEGG PATHWAY	KEGG pathway maps [Pathway list]
KEGG BRITE	BRITE functional hierarchies [Brite list]
KEGG MODULE	KEGG modules [Module list | Statistics] *New!*
KEGG ORTHOLOGY	Ortholog groups [KO system | Annotation]
KEGG GENOME	Genomes [KEGG organisms]
KEGG GENES	Genes and proteins [Release history]
KEGG COMPOUND	Small molecules [Compound classification]
KEGG REACTION	Biochemical reactions [Reaction modules]
KEGG DISEASE	Human diseases [Cancer | Infectious disease]
KEGG DRUG	Drugs [ATC drug classification]
KEGG MEDICUS	Health information resource [Drug labels search]

Organism-specific entry points

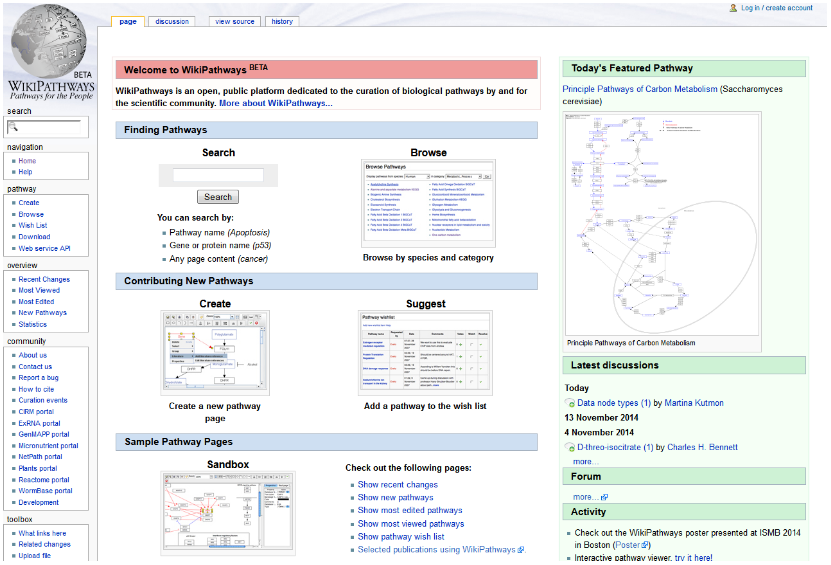
KEGG Organisms	Enter org code(s) □ Go hsa hsa eco

Analysis tools

KEGG Mapper	KEGG PATHWAY/BRITE/MODULE mapping tools
KEGG Atlas	Navigation tool to explore KEGG global maps
BlastKOALA *New!*	New service for genome/metagenome annotation
BLAST/FASTA	Sequence similarity search
SIMCOMP	Chemical structure similarity search

Copyright 1995-2014 Kanehisa Laboratories

WikiPathways



page | discussion | view source | history

Log in / create account

BETA
WIKIPATHWAYS
Pathways for the People

search

navigation
- Home
- Help

pathway
- Create
- Browse
- Wish List
- Download
- Web service API

overview
- Recent Changes
- Most Viewed
- Most Edited
- New Pathways
- Statistics

community
- About us
- Contact us
- Report a bug
- How to cite
- Curation events
- CIRM portal
- ExRNA portal
- GenMAPP portal
- Micronutrient portal
- NetPath portal
- Plants portal
- Reactome portal
- WormBase portal
- Development

toolbox
- What links here
- Related changes
- Upload file

Welcome to WikiPathways BETA

WikiPathways is an open, public platform dedicated to the curation of biological pathways by and for the scientific community. More about WikiPathways...

Finding Pathways

Search

Search

You can search by:
- Pathway name *(Apoptosis)*
- Gene or protein name *(p53)*
- Any page content *(cancer)*

Browse

Browse by species and category

Contributing New Pathways

Create

Create a new pathway page

Suggest

Add a pathway to the wish list

Sample Pathway Pages

Sandbox

Check out the following pages:
- Show recent changes
- Show new pathways
- Show most edited pathways
- Show most viewed pathways
- Show pathway wish list
- Selected publications using WikiPathways.

Today's Featured Pathway

Principle Pathways of Carbon Metabolism (Saccharomyces cerevisiae)

Principle Pathways of Carbon Metabolism

Latest discussions

Today
Data node types (1) by Martina Kutmon
13 November 2014
4 November 2014
D-threo-isocitrate (1) by Charles H. Bennett
more...

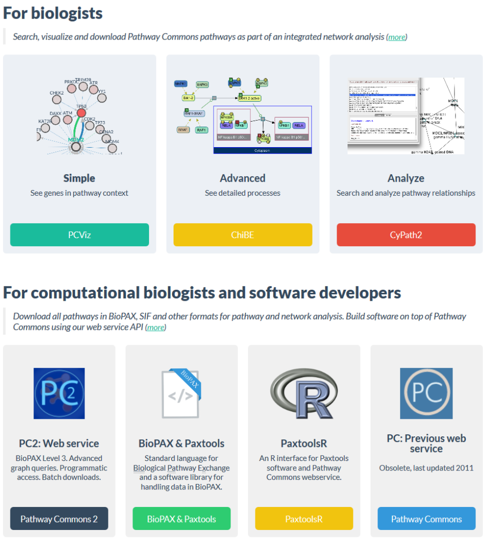
Forum

more...

Activity

- Check out the WikiPathways poster presented at ISMB 2014 in Boston (Poster)
- Interactive pathway viewer, try it here!

Pathway Commons



For biologists

Search, visualize and download Pathway Commons pathways as part of an integrated network analysis (more)

Simple
See genes in pathway context

Advanced
See detailed processes

Analyze
Search and analyze pathway relationships

For computational biologists and software developers

Download all pathways in BioPAX, SIF and other formats for pathway and network analysis. Build software on top of Pathway Commons using our web service API (more)

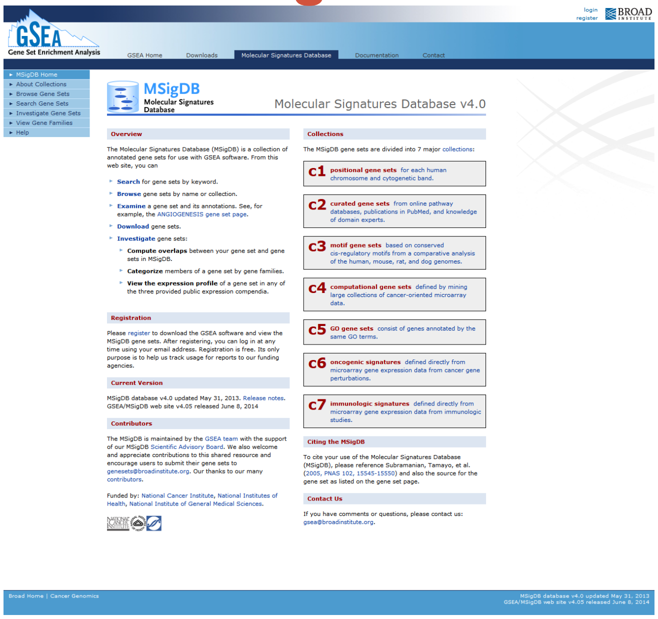
PC2: Web service
BioPAX Level 3. Advanced graph queries. Programmatic access. Batch downloads.

BioPAX & Paxtools
Standard language for Biological Pathway Exchange and a software library for handling data in BioPAX.

PaxtoolsR
An R interface for Paxtools software and Pathway Commons webservice.

PC: Previous web service
Obsolete, last updated 2011

MSigDB



MSigDB Molecular Signatures Database

Molecular Signatures Database v4.0

Overview

The Molecular Signatures Database (MSigDB) is a collection of annotated gene sets for use with GSEA software. From this web site, you can

- **Search** for gene sets by keyword.
- **Browse** gene sets by name or collection.
- **Examine** a gene set and its annotations. See, for example, the ANGIOGENESIS gene set page.
- **Download** gene sets.
- **Investigate** gene sets:
  - **Compute overlaps** between your gene set and gene sets in MSigDB.
  - **Categorize** members of a gene set by gene families.
  - **View the expression profile** of a gene set in any of the three provided public expression compendia.

Registration

Please register to download the GSEA software and view the MSigDB gene sets. After registering, you can log in at any time using your email address. Registration is free. Its only purpose is to help us track usage for reports to our funding agencies.

Current Version

MSigDB database v4.0 updated May 31, 2013. Release notes.
GSEA/MSigDB web site v4.05 released June 8, 2014

Contributors

The MSigDB is maintained by the GSEA team with the support of our MSigDB Scientific Advisory Board. We also welcome and appreciate contributions to this shared resource and encourage users to submit their gene sets to genesets@broadinstitute.org. Our thanks to our many contributors.

Funded by: National Cancer Institute, National Institutes of Health, National Institute of General Medical Sciences.

Collections

The MSigDB gene sets are divided into 7 major collections:

c1 **positional gene sets** for each human chromosome and cytogenetic band.

c2 **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

c3 **motif gene sets** based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.

c4 **computational gene sets** defined by mining large collections of cancer-oriented microarray data.

c5 **GO gene sets** consist of genes annotated by the same GO terms.

c6 **oncogenic signatures** defined directly from microarray gene expression data from cancer gene perturbations.

c7 **immunologic signatures** defined directly from microarray gene expression data from immunologic studies.

Citing the MSigDB

To cite your use of the Molecular Signatures Database (MSigDB), please reference Subramanian, Tamayo, et al. (2005, PNAS 102, 15545-15550) and also the source for the gene set as listed on the gene set page.

Contact Us

If you have comments or questions, please contact us: gsea@broadinstitute.org.

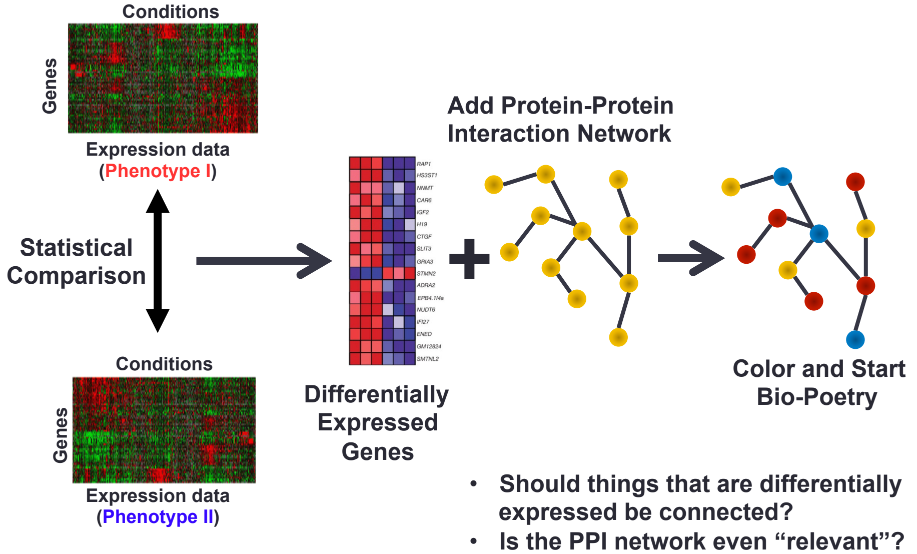
Broad Home | Cancer Genomics

MSigDB database v4.0 updated May 31, 2013
GSEA/MSigDB web site v4.05 released June 8, 2014

Biological Networks

Can we make this more complicated?

How *NOT* to do Network Analysis



Expression data (Phenotype I)

Statistical Comparison

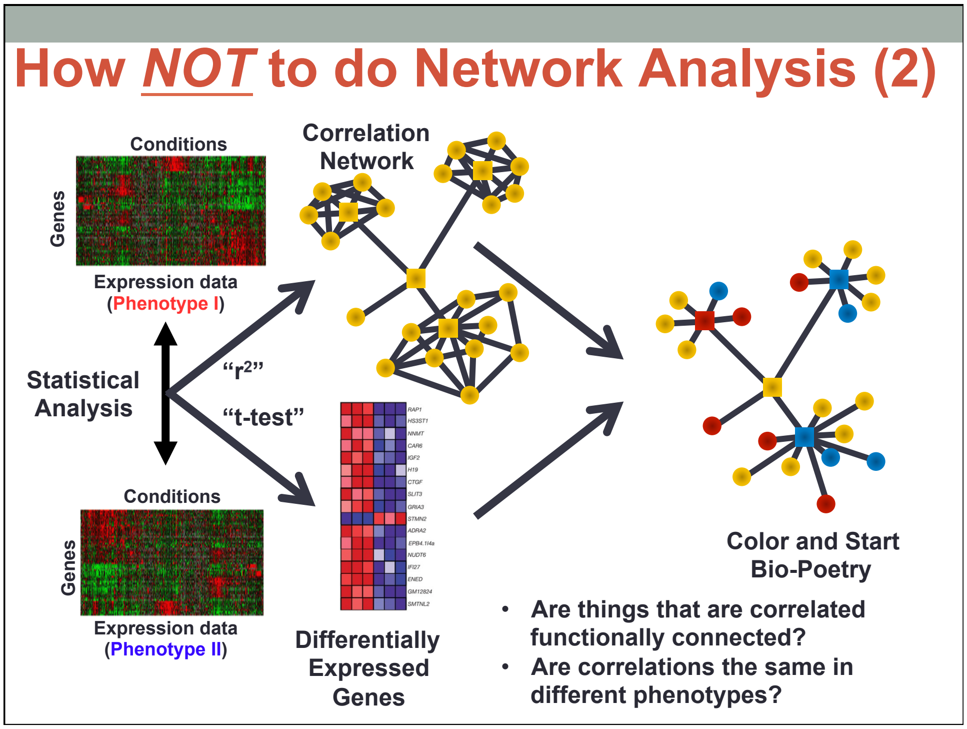
Expression data (Phenotype II)

Differentially Expressed Genes

Add Protein-Protein Interaction Network

Color and Start Bio-Poetry

- Should things that are differentially expressed be connected?
- Is the PPI network even “relevant”?



How *NOT* to do Network Analysis (2)

Conditions

Genes

Expression data (Phenotype I)

Statistical Analysis

"r²"

"t-test"

Conditions

Genes

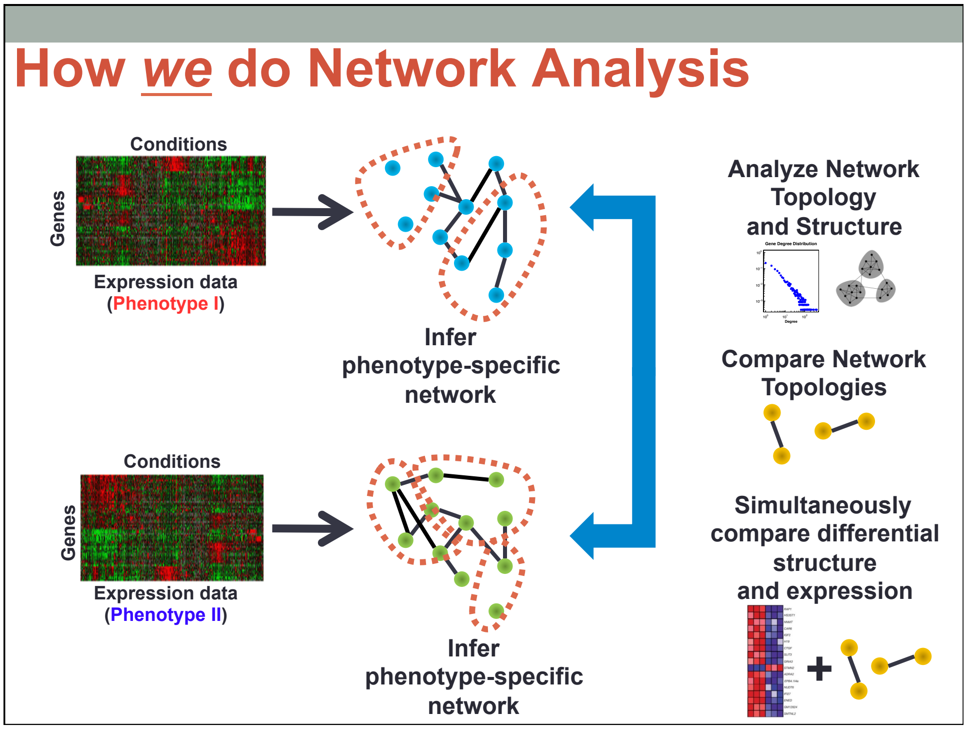
Expression data (Phenotype II)

Correlation Network

Differentially Expressed Genes

Color and Start Bio-Poetry

- Are things that are correlated functionally connected?
- Are correlations the same in different phenotypes?



How *we* do Network Analysis

Conditions

Genes

Expression data (Phenotype I)

Infer phenotype-specific network

Conditions

Genes

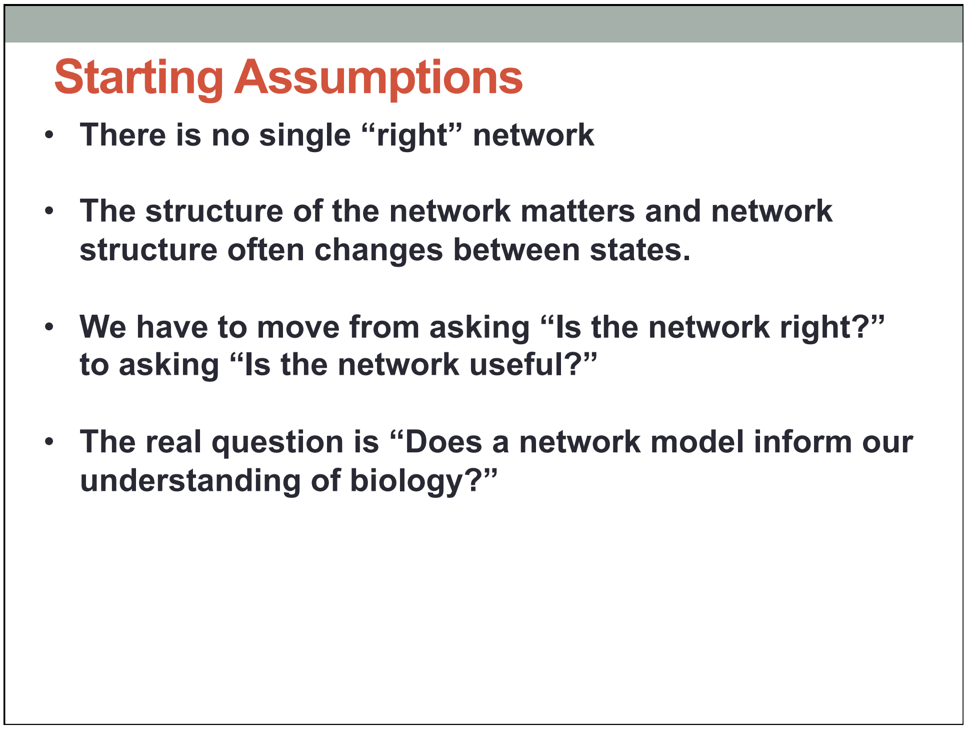
Expression data (Phenotype II)

Infer phenotype-specific network

Analyze Network Topology and Structure

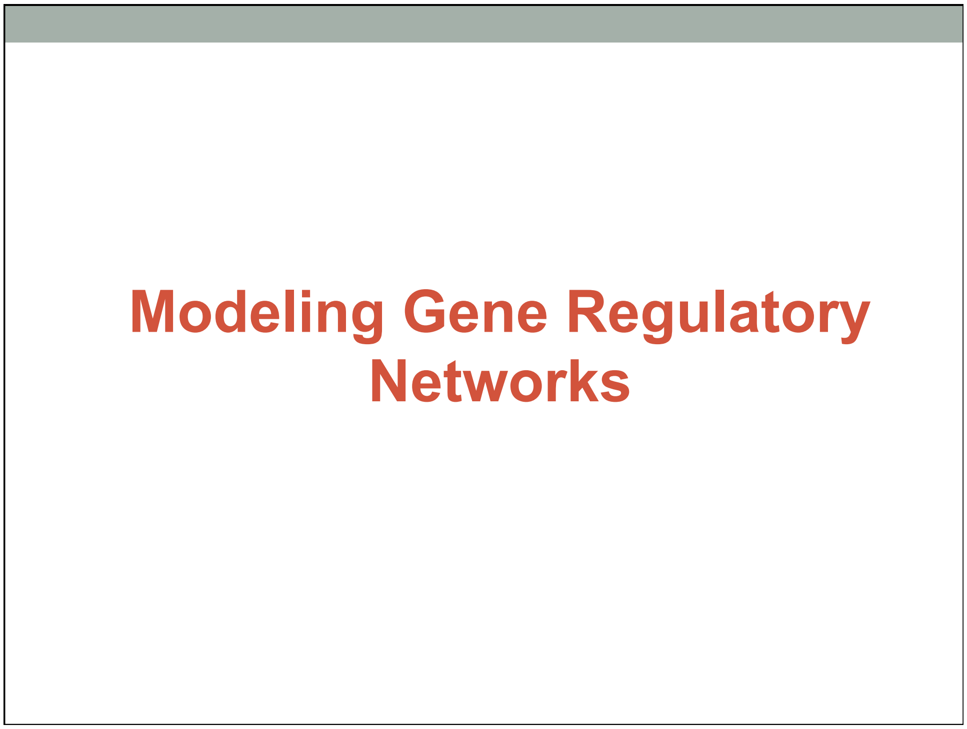
Compare Network Topologies

Simultaneously compare differential structure and expression



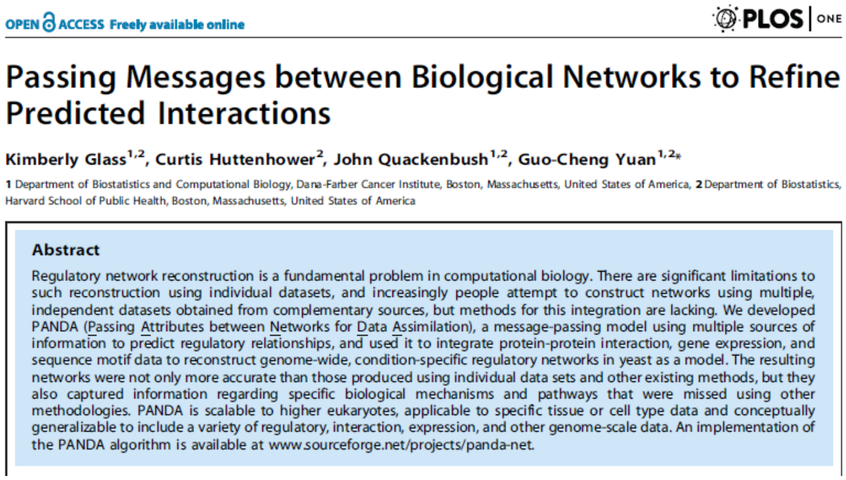
Starting Assumptions

- **There is no single “right” network**
- **The structure of the network matters and network structure often changes between states.**
- **We have to move from asking “Is the network right?” to asking “Is the network useful?”**
- **The real question is “Does a network model inform our understanding of biology?”**



Modeling Gene Regulatory Networks

Integrative Network Inference: PANDA



OPEN ACCESS Freely available online

PLOS | ONE

Passing Messages between Biological Networks to Refine Predicted Interactions

Kimberly Glass[1,2], Curtis Huttenhower[2], John Quackenbush[1,2], Guo-Cheng Yuan[1,2]*

1 Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, United States of America, 2 Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, United States of America

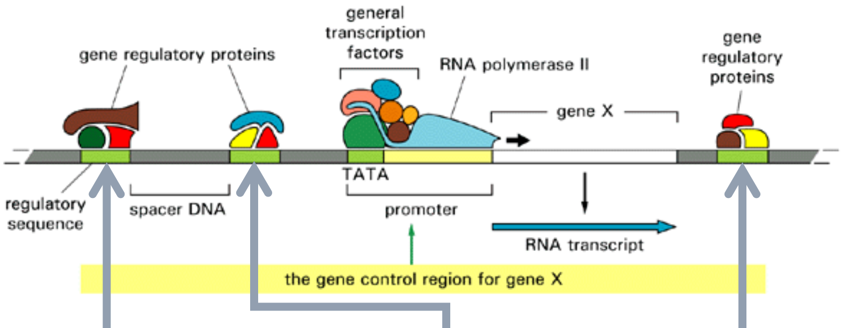
Abstract

Regulatory network reconstruction is a fundamental problem in computational biology. There are significant limitations to such reconstruction using individual datasets, and increasingly people attempt to construct networks using multiple, independent datasets obtained from complementary sources, but methods for this integration are lacking. We developed PANDA (Passing Attributes between Networks for Data Assimilation), a message-passing model using multiple sources of information to predict regulatory relationships, and used it to integrate protein-protein interaction, gene expression, and sequence motif data to reconstruct genome-wide, condition-specific regulatory networks in yeast as a model. The resulting networks were not only more accurate than those produced using individual data sets and other existing methods, but they also captured information regarding specific biological mechanisms and pathways that were missed using other methodologies. PANDA is scalable to higher eukaryotes, applicable to specific tissue or cell type data and conceptually generalizable to include a variety of regulatory, interaction, expression, and other genome-scale data. An implementation of the PANDA algorithm is available at www.sourceforge.net/projects/panda-net.

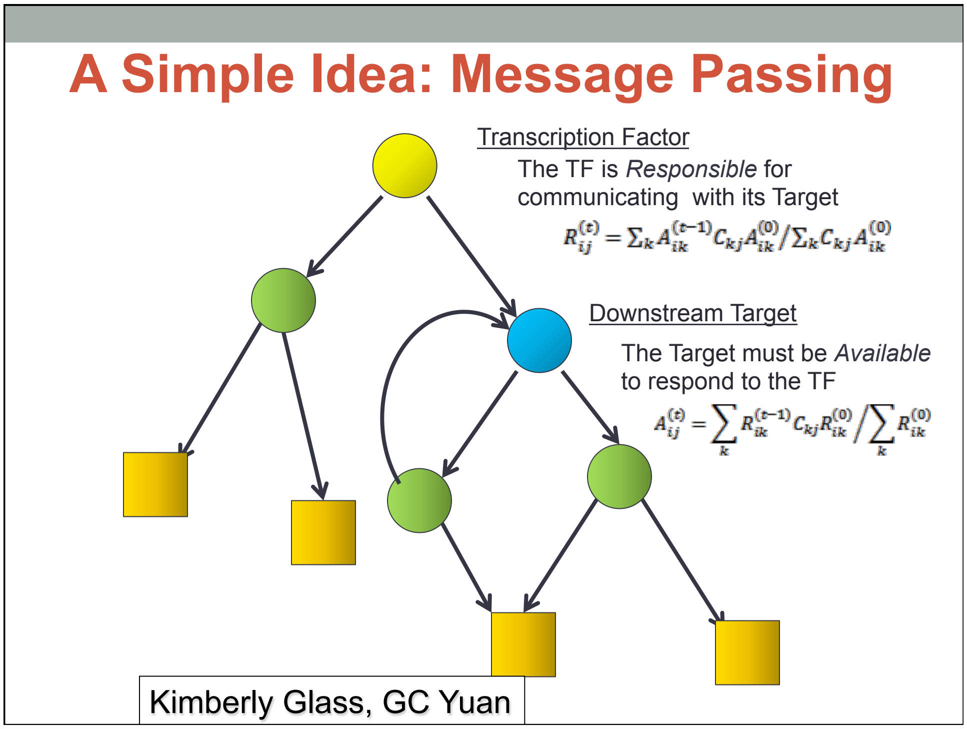
Regulation of Transcription

regulatory sequences

promoter



Specific transcription factors



A Simple Idea: Message Passing

Transcription Factor

The TF is *Responsible* for communicating with its Target

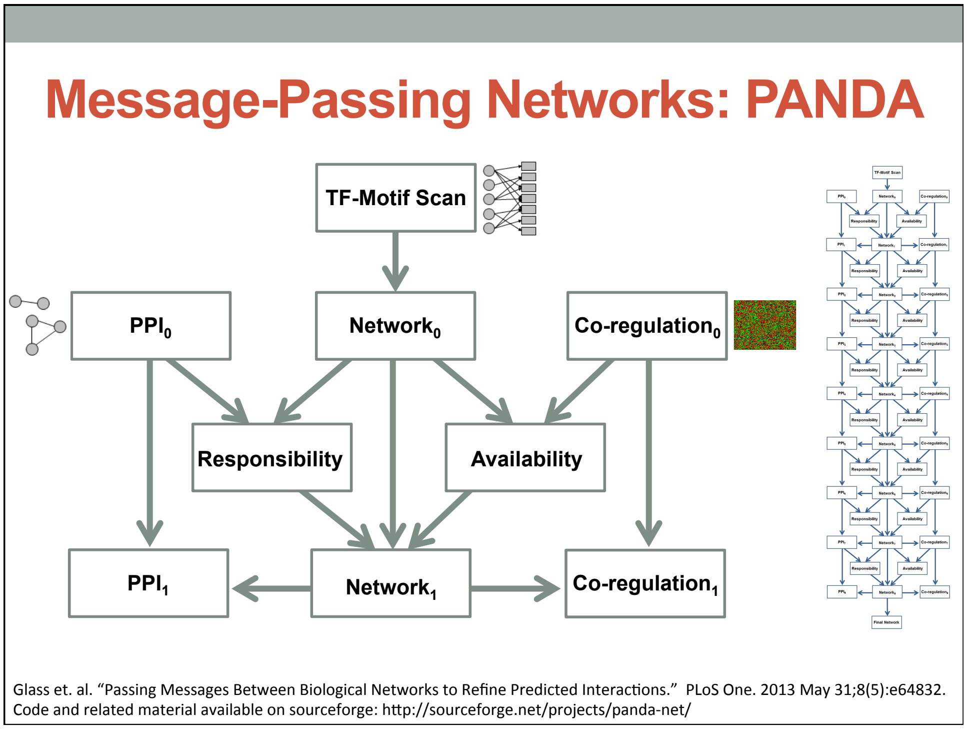
$$R_{ij}^{(t)} = \Sigma_k A_{ik}^{(t-1)} C_{kj} A_{ik}^{(0)} / \Sigma_k C_{kj} A_{ik}^{(0)}$$

Downstream Target

The Target must be *Available* to respond to the TF

$$A_{ij}^{(t)} = \sum_k R_{ik}^{(t-1)} C_{kj} R_{ik}^{(0)} \Big/ \sum_k R_{ik}^{(0)}$$

Kimberly Glass, GC Yuan



Message-Passing Networks: PANDA

TF-Motif Scan

PPI_0

Network_0

Co-regulation_0

Responsibility

Availability

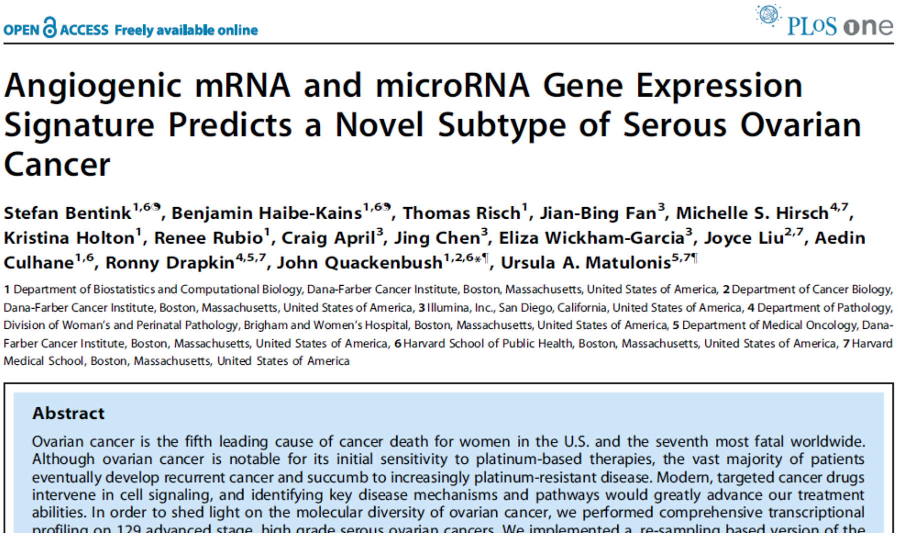
PPI_1

Network_1

Co-regulation_1

Glass et. al. "Passing Messages Between Biological Networks to Refine Predicted Interactions." PLoS One. 2013 May 31;8(5):e64832.
Code and related material available on sourceforge: http://sourceforge.net/projects/panda-net/

Subtypes of Ovarian Cancer



OPEN ACCESS Freely available online

PLoS one

Angiogenic mRNA and microRNA Gene Expression Signature Predicts a Novel Subtype of Serous Ovarian Cancer

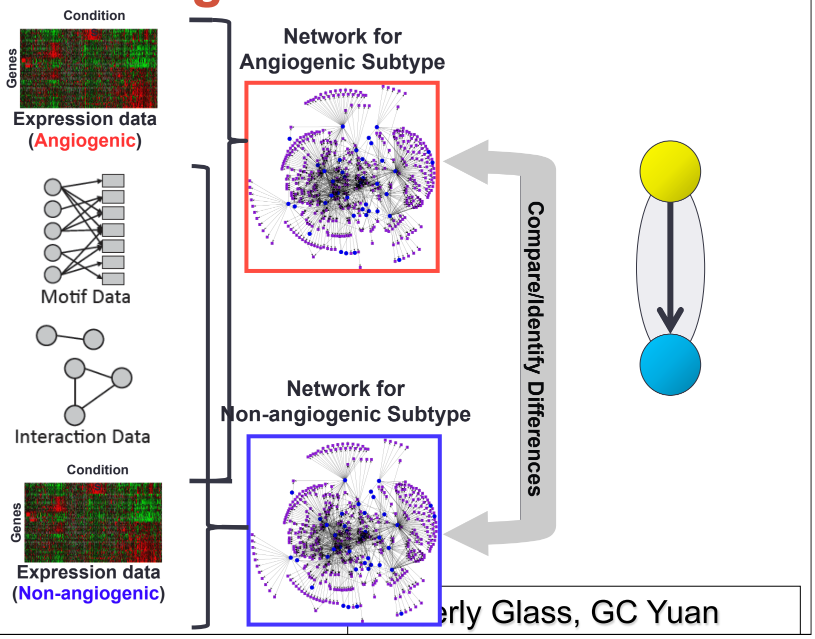
Stefan Bentink[1,6,9], Benjamin Haibe-Kains[1,6,9], Thomas Risch[1], Jian-Bing Fan[3], Michelle S. Hirsch[4,7], Kristina Holton[1], Renee Rubio[1], Craig April[3], Jing Chen[3], Eliza Wickham-Garcia[3], Joyce Liu[2,7], Aedin Culhane[1,6], Ronny Drapkin[4,5,7], John Quackenbush[1,2,6*¶], Ursula A. Matulonis[5,7¶]

1 Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, United States of America, 2 Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, United States of America, 3 Illumina, Inc., San Diego, California, United States of America, 4 Department of Pathology, Division of Woman's and Perinatal Pathology, Brigham and Women's Hospital, Boston, Massachusetts, United States of America, 5 Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, United States of America, 6 Harvard School of Public Health, Boston, Massachusetts, United States of America, 7 Harvard Medical School, Boston, Massachusetts, United States of America

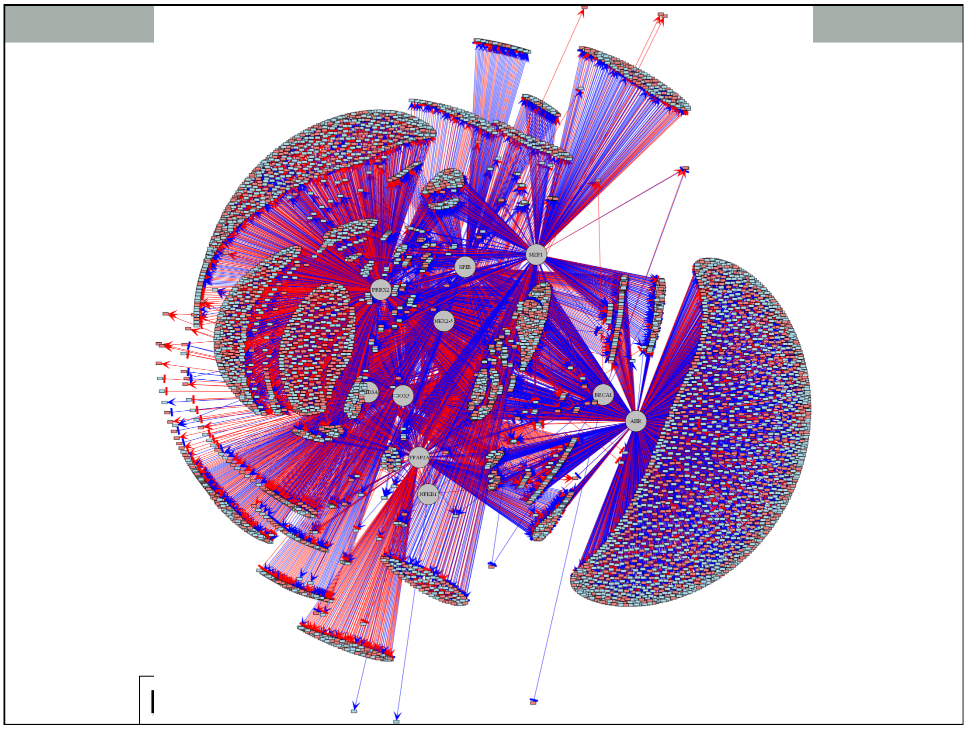
Abstract

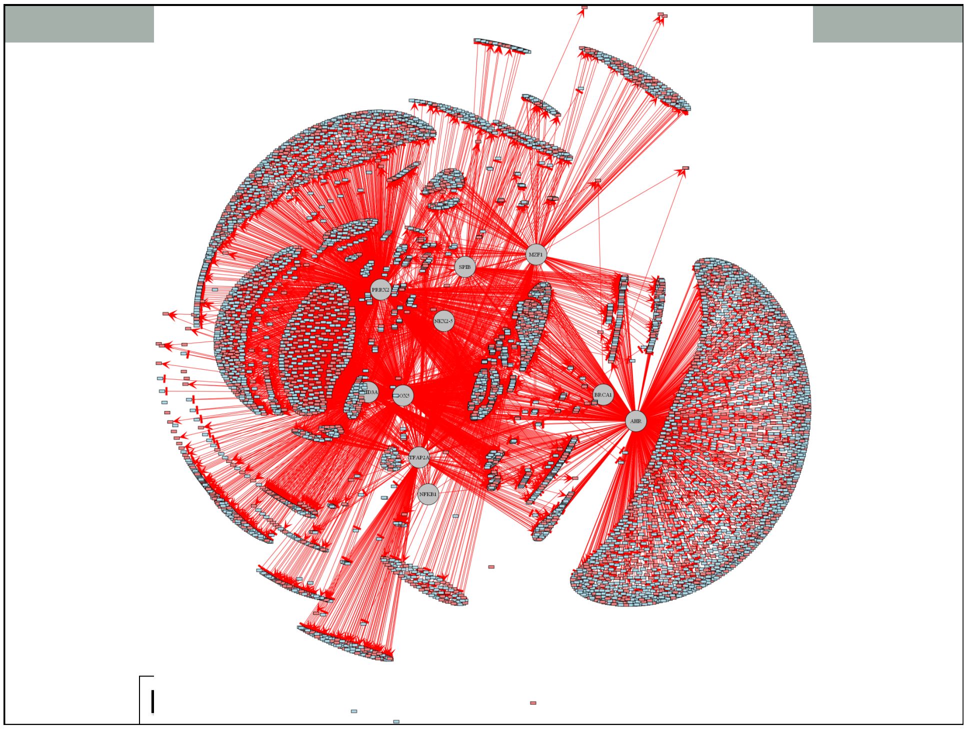
Ovarian cancer is the fifth leading cause of cancer death for women in the U.S. and the seventh most fatal worldwide. Although ovarian cancer is notable for its initial sensitivity to platinum-based therapies, the vast majority of patients eventually develop recurrent cancer and succumb to increasingly platinum-resistant disease. Modern, targeted cancer drugs intervene in cell signaling, and identifying key disease mechanisms and pathways would greatly advance our treatment abilities. In order to shed light on the molecular diversity of ovarian cancer, we performed comprehensive transcriptional profiling on 129 advanced stage, high grade serous ovarian cancers. We implemented a re-sampling based version of the

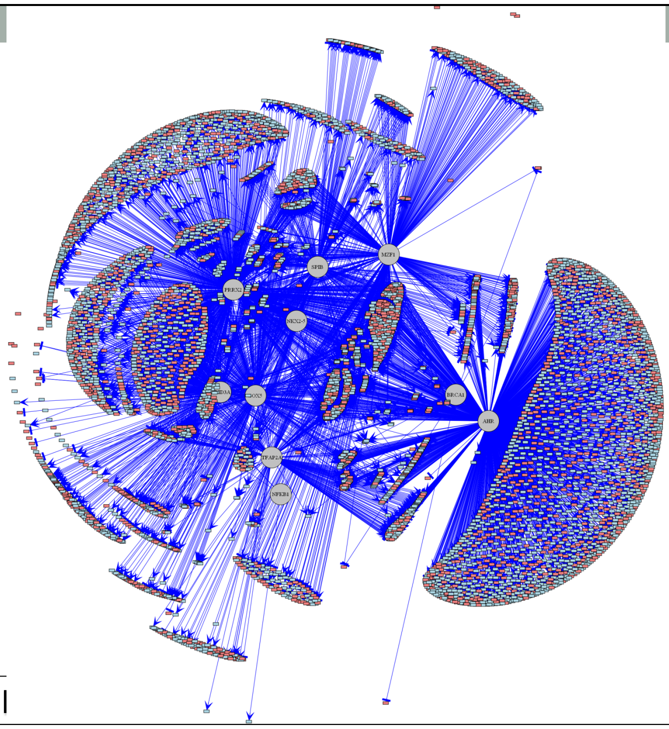
PANDA: Integrative Network Models



rly Glass, GC Yuan


MZF1
SPIB
PRRX2
NKX2-5
ID3A
OX5
BRCA1
AHR
TFAP2A
NFKB1


MZF1
SPIB
PRRX2
NKX2-5
ID3A
OX5
BRCA1
AHR
TFAP2A
NFKB1

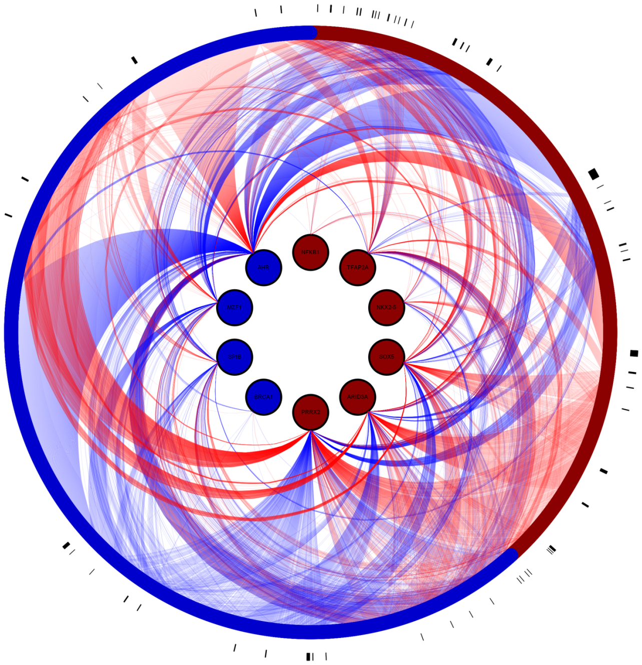


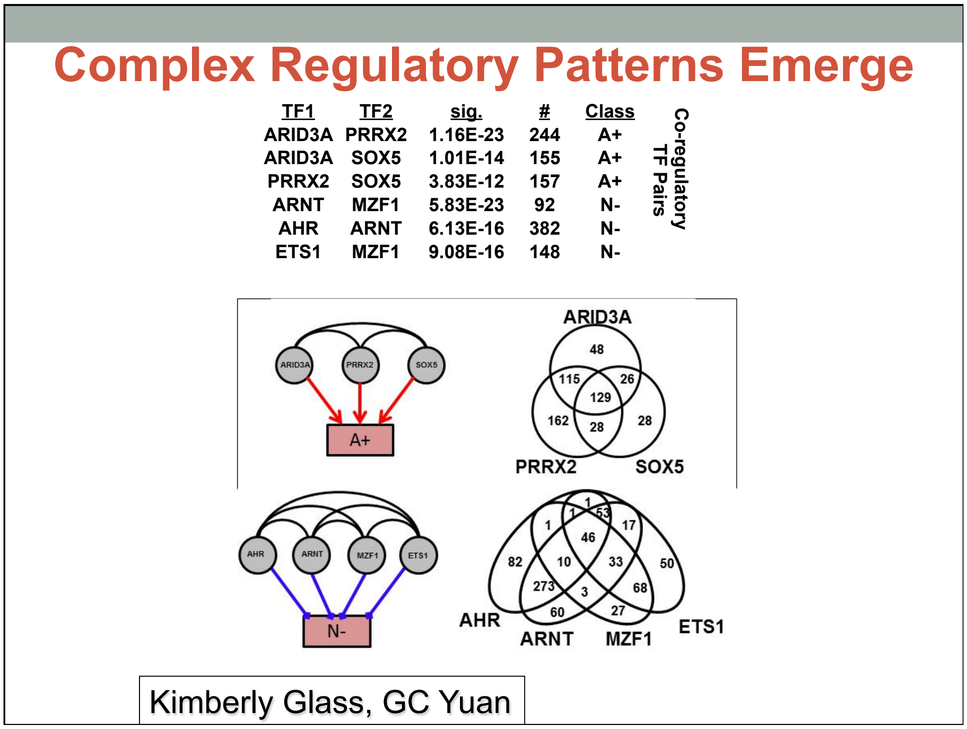
Inner ring: key TFs Colored by Edge Enrichment (A or N)

Outer ring: genes Colored by Differential Expression (A or N)

Interring Connections Colored by Subnetwork (A or N)

Ticks – genes annotated to “angiogenesis” in GO,



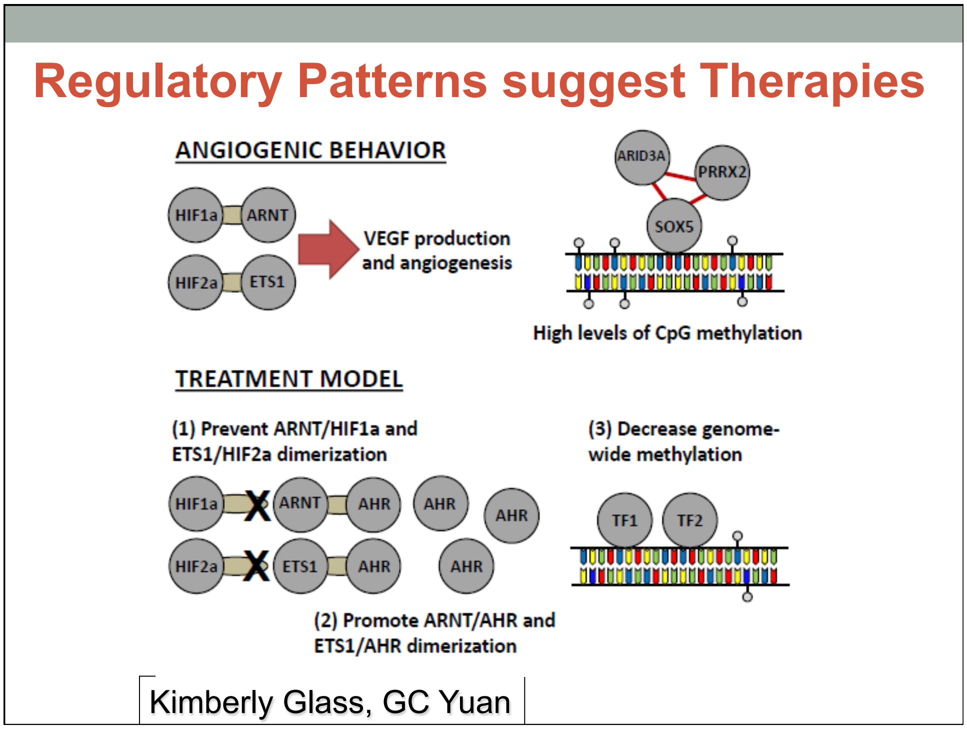


Complex Regulatory Patterns Emerge

TF1	TF2	sig.	#	Class
ARID3A	PRRX2	1.16E-23	244	A+
ARID3A	SOX5	1.01E-14	155	A+
PRRX2	SOX5	3.83E-12	157	A+
ARNT	MZF1	5.83E-23	92	N-
AHR	ARNT	6.13E-16	382	N-
ETS1	MZF1	9.08E-16	148	N-

Co-regulatory TF Pairs

Kimberly Glass, GC Yuan



Regulatory Patterns suggest Therapies

ANGIOGENIC BEHAVIOR

High levels of CpG methylation

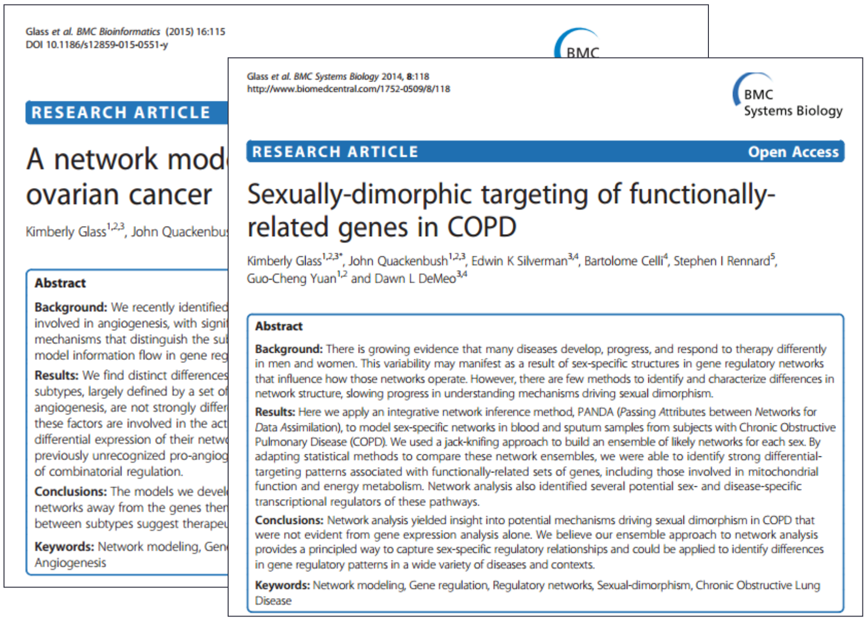
TREATMENT MODEL

(1) Prevent ARNT/HIF1a and ETS1/HIF2a dimerization

(2) Promote ARNT/AHR and ETS1/AHR dimerization

(3) Decrease genome-wide methylation

Kimberly Glass, GC Yuan



Glass *et al. BMC Bioinformatics* (2015) 16:115
DOI 10.1186/s12859-015-0551-y

RESEARCH ARTICLE

A network mod
ovarian cancer

Kimberly Glass[1,2,3], John Quackenbu

Abstract

Background: We recently identified
involved in angiogenesis, with signi
mechanisms that distinguish the su
model information flow in gene reg

Results: We find distinct differences
subtypes, largely defined by a set of
angiogenesis, are not strongly differ
these factors are involved in the act
differential expression of their netw
previously unrecognized pro-angiog
of combinatorial regulation.

Conclusions: The models we devel
networks away from the genes ther
between subtypes suggest therapeu

Keywords: Network modeling, Gen
Angiogenesis

Glass *et al. BMC Systems Biology* 2014, **8**:118
http://www.biomedcentral.com/1752-0509/8/118

BMC Systems Biology

RESEARCH ARTICLE Open Access

Sexually-dimorphic targeting of functionally-related genes in COPD

Kimberly Glass[1,2,3*], John Quackenbush[1,2,3], Edwin K Silverman[3,4], Bartolome Celli[4], Stephen I Rennard[5], Guo-Cheng Yuan[1,2] and Dawn L DeMeo[3,4]

Abstract

Background: There is growing evidence that many diseases develop, progress, and respond to therapy differently in men and women. This variability may manifest as a result of sex-specific structures in gene regulatory networks that influence how those networks operate. However, there are few methods to identify and characterize differences in network structure, slowing progress in understanding mechanisms driving sexual dimorphism.

Results: Here we apply an integrative network inference method, PANDA (*P*assing *A*ttributes between *N*etworks for *D*ata *A*ssimilation), to model sex-specific networks in blood and sputum samples from subjects with Chronic Obstructive Pulmonary Disease (COPD). We used a jack-knifing approach to build an ensemble of likely networks for each sex. By adapting statistical methods to compare these network ensembles, we were able to identify strong differential-targeting patterns associated with functionally-related sets of genes, including those involved in mitochondrial function and energy metabolism. Network analysis also identified several potential sex- and disease-specific transcriptional regulators of these pathways.

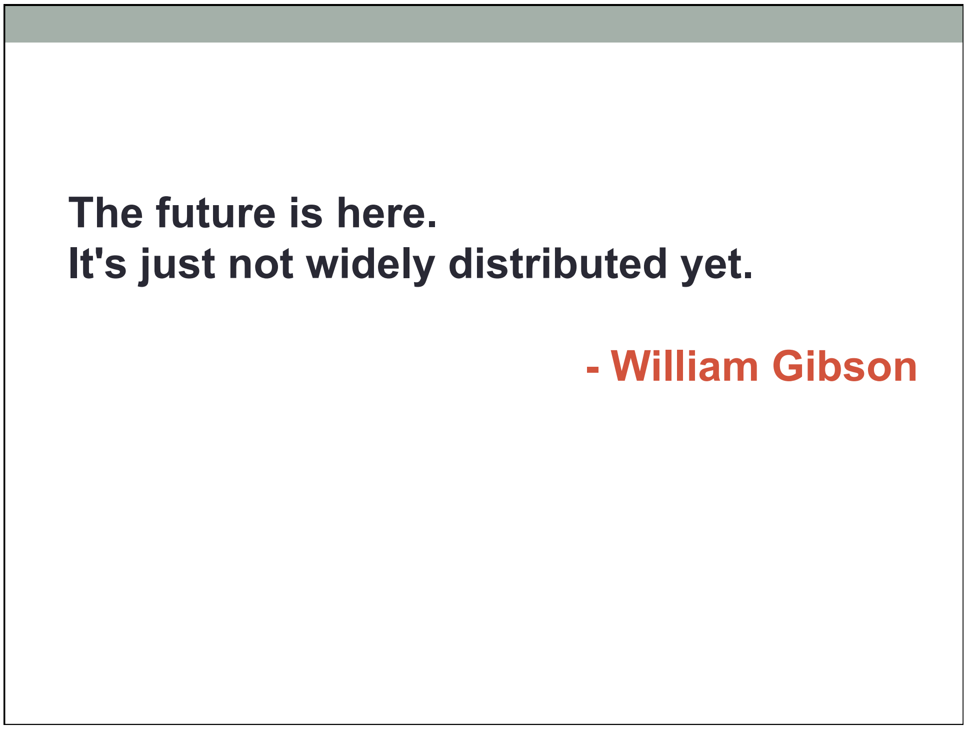
Conclusions: Network analysis yielded insight into potential mechanisms driving sexual dimorphism in COPD that were not evident from gene expression analysis alone. We believe our ensemble approach to network analysis provides a principled way to capture sex-specific regulatory relationships and could be applied to identify differences in gene regulatory patterns in a wide variety of diseases and contexts.

Keywords: Network modeling, Gene regulation, Regulatory networks, Sexual-dimorphism, Chronic Obstructive Lung Disease

More application papers coming….

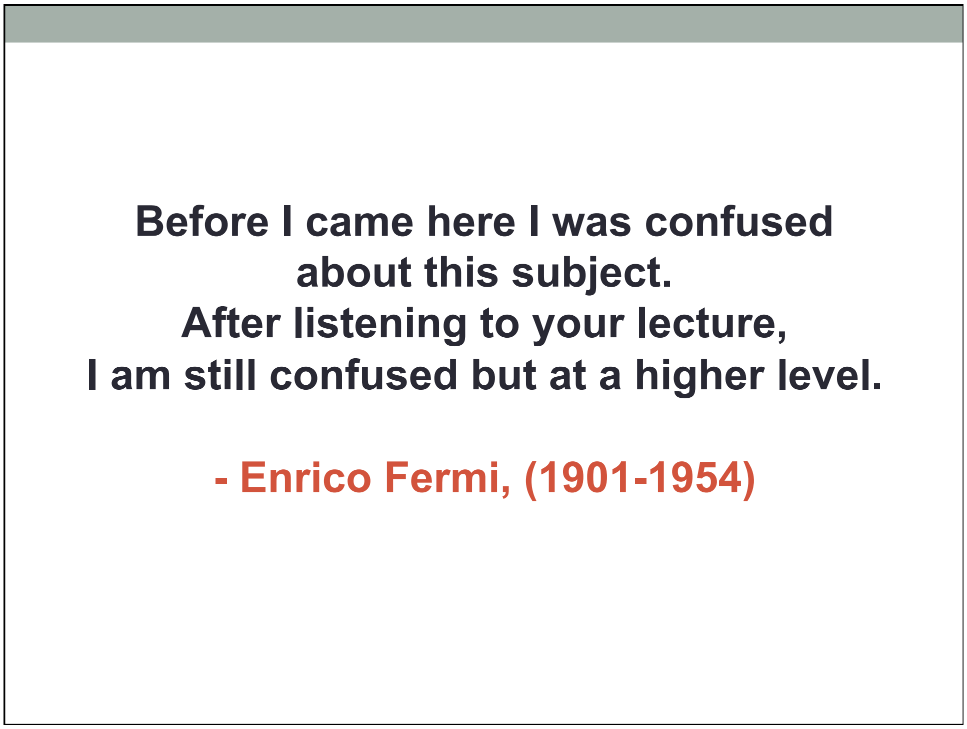
At the End of the Day

- The goal of an experiment is to discover new biology
- The challenge is sorting through lots of data
- Comparing groups of samples requires thorough annotation
- Making sense of the genes that are significant in such a comparison requires thorough gene annotation
- New technologies are giving us new ways of generating data, but the analysis approaches are more-or-less the same.



The future is here.
It's just not widely distributed yet.

- William Gibson



Before I came here I was confused
about this subject.
After listening to your lecture,
I am still confused but at a higher level.

- Enrico Fermi, (1901-1954)