BIOS6606: Analysis of RNA Expression Levels

Laura Saba, PhD

Department of Pharmaceutical Sciences

Skaggs School of Pharmacy and Pharmaceutical Sciences

University of Colorado Anschutz Medical Campus

Laura.Saba@UCDenver.edu

Many slides adapted from Katerina Kechris' presentation for BIOS6606 in the Fall of 2007

BMC Biology

BioMed Central

Research article

Open Access

Genetical genomic determinants of alcohol consumption in rats and humans

Boris Tabakoff*[1], Laura Saba[1], Morton Printz[2], Pam Flodman[3], Colin Hodgkinson[4], David Goldman[4], George Koob[5], Heather N Richardson[5,16], Katerina Kechris[6], Richard L Bell[7]

Neuropharmacology 60 (2011) 1269–1280

Contents lists available at ScienceDirect

Neuropharmacology

journal homepage: www.elsevier.com/locate/neuropharm

A systems genetic analysis of alcohol drinking by mice, rats and men: Influence of brain GABAergic transmission

Laura M. Saba [a], Beth Bennett [a], Paula L. Hoffman [a], Kelsey Barcomb [a], Takao Ishii [a], Katerina Kechris [b], Boris Tabakoff [a,*]

[a] Department of Pharmacology, University of Colorado Denver School of Medicine, PO Box 6511, Mail Stop 8303, Aurora, CO 80045, USA
[b] Colorado School of Public Health, Campus Box B119, Aurora, CO 80045, USA

Addiction Biology

PRECLINICAL STUDY

doi:10.1111/j.1369-1600.2010.00254.x

Using the Phenogen website for *'in silico'* a morphine-induced analgesia: identifying candidate genes

Paula L. Hoffman[1], Beth Bennett[1], Laura M. Saba[1], Sanjiv V. Bhave[1], Phyll Cheryl K. Hornbaker[1], Katerina J. Kechris[1], Robert W. Williams[2] & Boris

University of Colorado Denver, School of Medicine, Department of Pharmacology, Aurora CO 80045, USA[1], University o Department of Anatomy and Neurobiology, Memphis, TN 38163, USA[2]

BMC Genetics

BioMed Central

Database

Open Access

The PhenoGen Informatics website: tools for analyses of complex traits

Sanjiv V Bhave[1], Cheryl Hornbaker[1], Tzu L Phang[1], Laura Saba[1], Razvan Lapadat[1], Katherina Kechris[1,2], Jeanette Gaydos[1], Daniel McGoldrick[1], Andrew Dolbey[1], Sonia Leach[1], Brian Soriano[1], Allison Ellington[1], Eric Ellington[1], Kendra Jones[1], Jonathan Mangion[3], John K Belknap[4], Robert W Williams[5], Lawrence E Hunter[1], Paula L Hoffman[1] and Boris Tabakoff*[1]

Topics

I. RNA Expression Technologies
II. Example - Finding Genomic Drivers of Alcohol Consumption
III. Resources
IV. Glimpse into Co-Expression Networks

Definitions: Genes & Genomes

- Each cell contains a copy of the genome in its nucleus
  - complete set of inherited genetic material (human: ~3 billion DNA bases A,C,G,T)
  - blueprint for all cellular structures and activities
- Genes: segments of the genome that carry the directions a cell uses to perform a specific function (human: ~20,000-25,000 protein-coding genes, ~3% the genome)

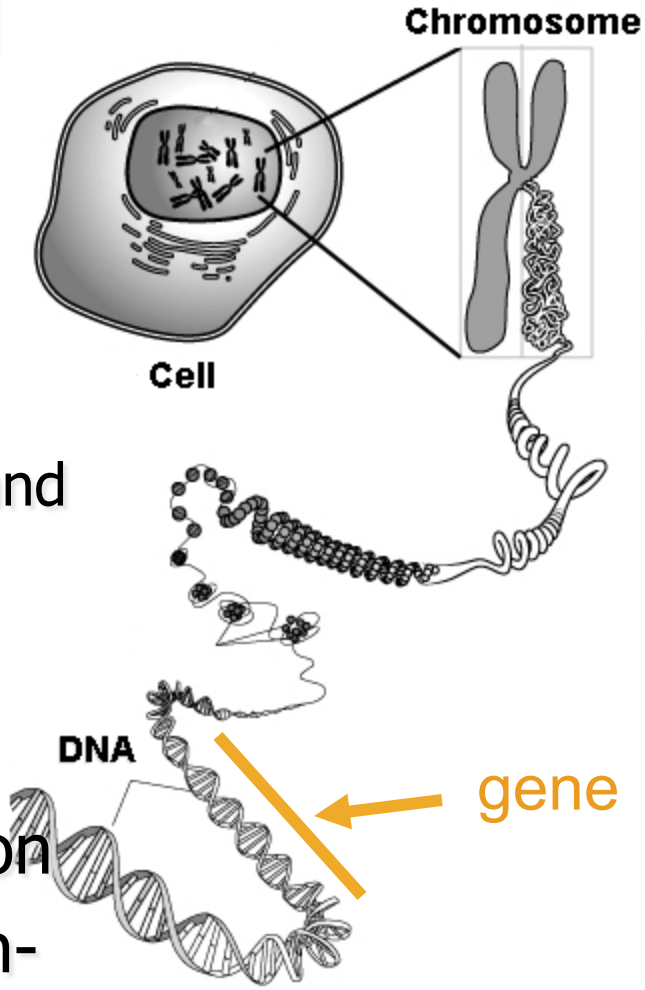


Image from http://www.koshland-science-museum.org/exhibitdna/

Definition of Gene Expression

Depending on cell type, environment, and life cycle, different genes are turned "on" or "off" with a dimmer switch

DNA

↓

TRANSCRIPTION

↓

RNA

↓

TRANSLATION

↓

PROTEIN

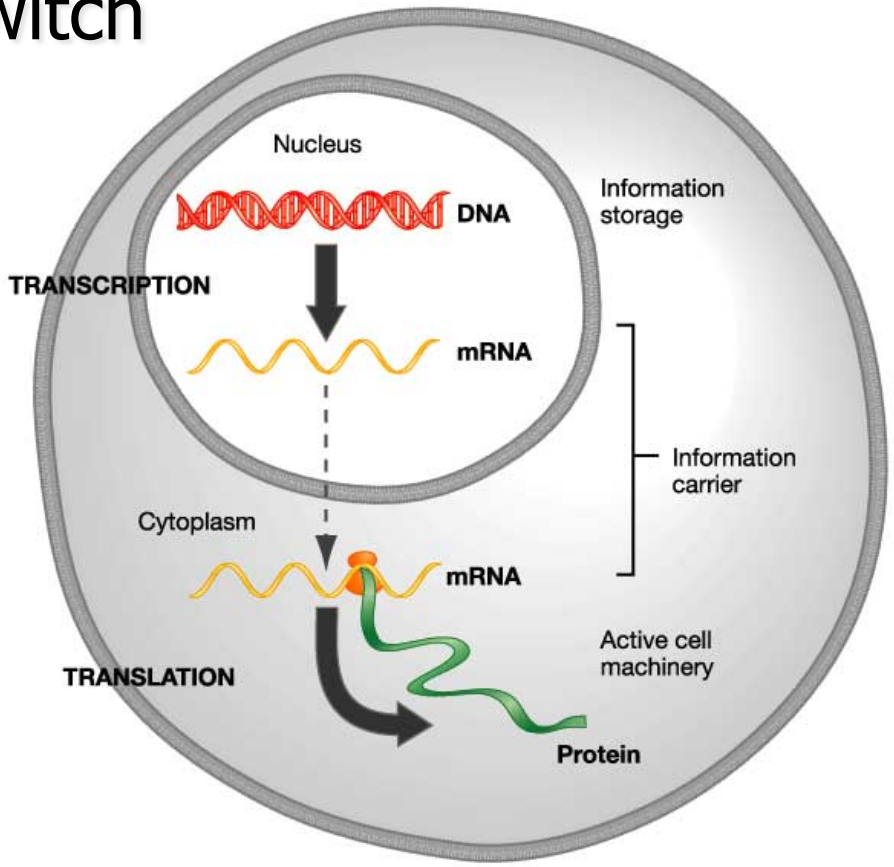
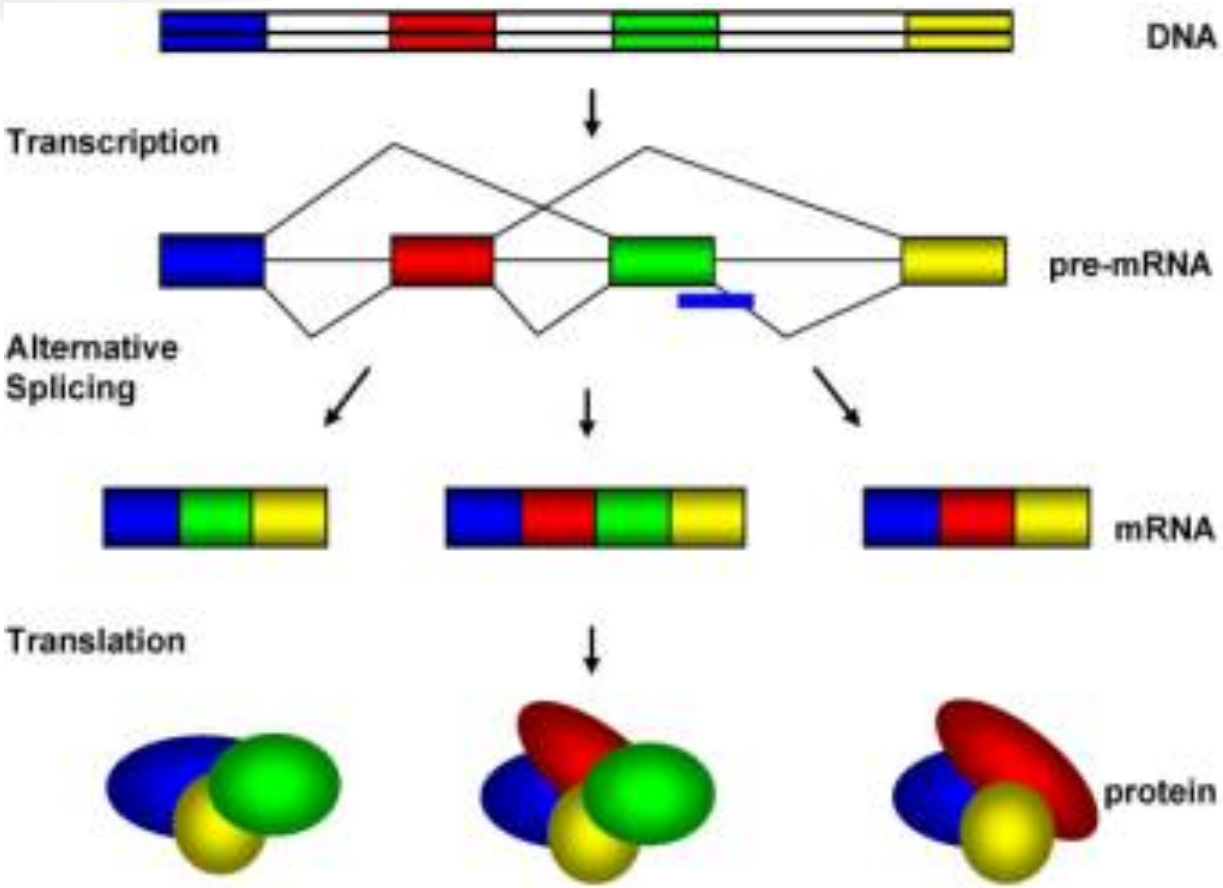


Image from http://campus.queens.edu/

Alternative Splicing



Examples of Biological Questions

- What genes are involved in a particular biological process?
- What genes are turned-on?
- What genes are turned-off?
- What genes work together?
- Do similar samples share similar gene expression profiles (*e.g.*, biomarkers)?

Hypothesis Generating VS Hypothesis Testing

How can we measure RNA expression levels?

- Microarrays:
  - How many times does this particular nucleotide sequence from a gene show up in our pool of RNA?
- High Throughput RNA Sequencing (RNA-Seq)
  - How many "pieces" of this gene show up in our pool of RNA?

Advantages of Microarrays

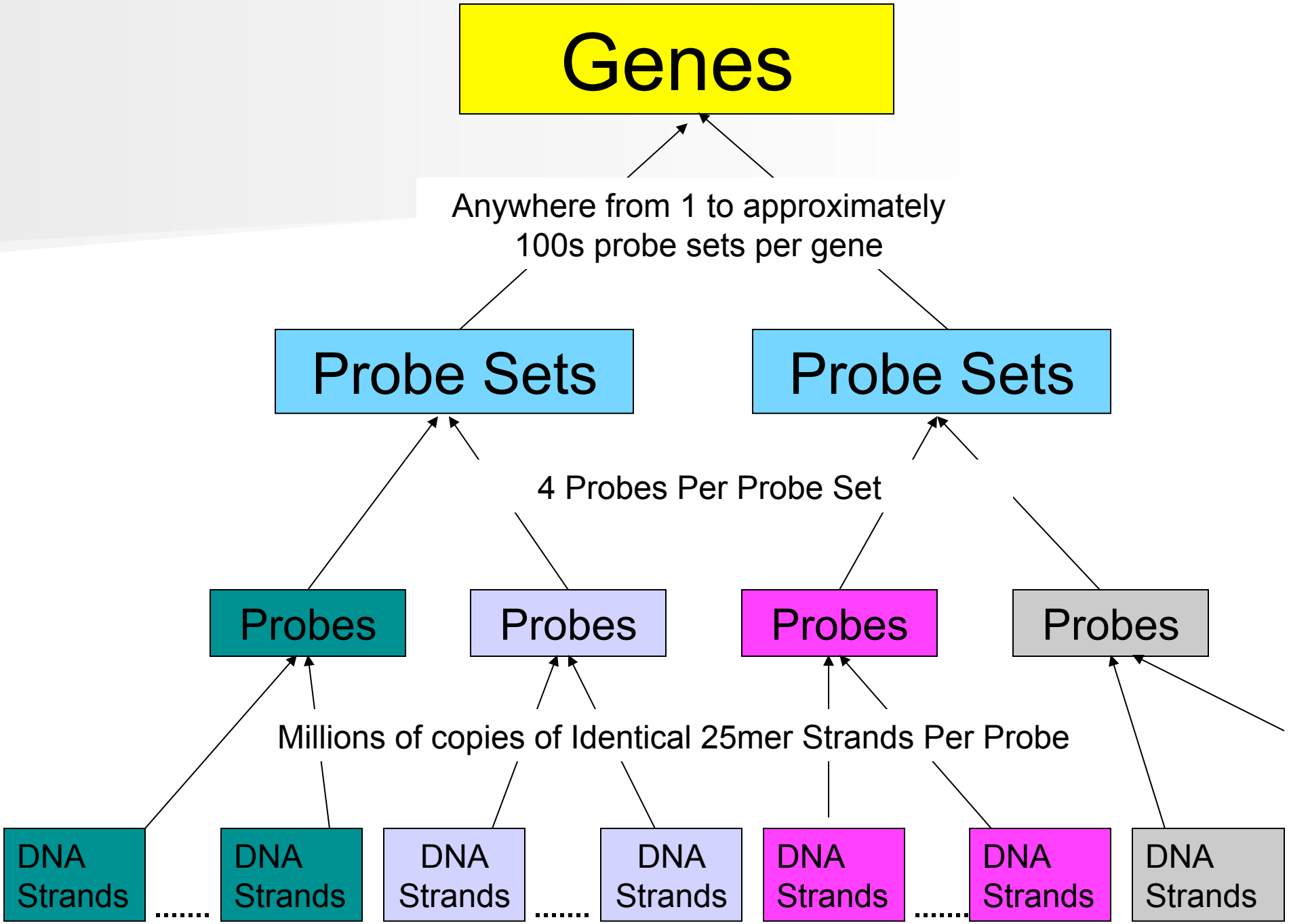
- **Multiplexing:** Multiple samples at same time
- **Automation:** Chip manufacturing, reagents
- **Cost:** relatively cheap (compared to RNA-Seq)
- **Ease of Analysis:** established pipelines that can be executed on most desktops

Disadvantages of Microarrays

- **Limited Interpretation:** only probing one area, dependent on current annotation
- **Probe Design:** probes are designed based on a reference sample, SNP can interfere with hybridization
- **Differences in Hybridization Efficiency:** prevent the comparison between genes, can only compare between samples

Affymetrix Technology

- 3’ Gene Expression
  - Original design
  - “Gene” level
- Exon Array Expression
  - New standard
  - Alternative splicing
  - Transcript or exon level
- Tiling Arrays
  - Target regions not genes
  - Unbias look at genome


Genes
Anywhere from 1 to approximately
100s probe sets per gene
Probe Sets
Probe Sets
4 Probes Per Probe Set
Probes
Probes
Probes
Probes
Millions of copies of Identical 25mer Strands Per Probe
DNA Strands
DNA Strands
DNA Strands
DNA Strands
DNA Strands
DNA Strands
DNA Strands

Affymetrix Technology

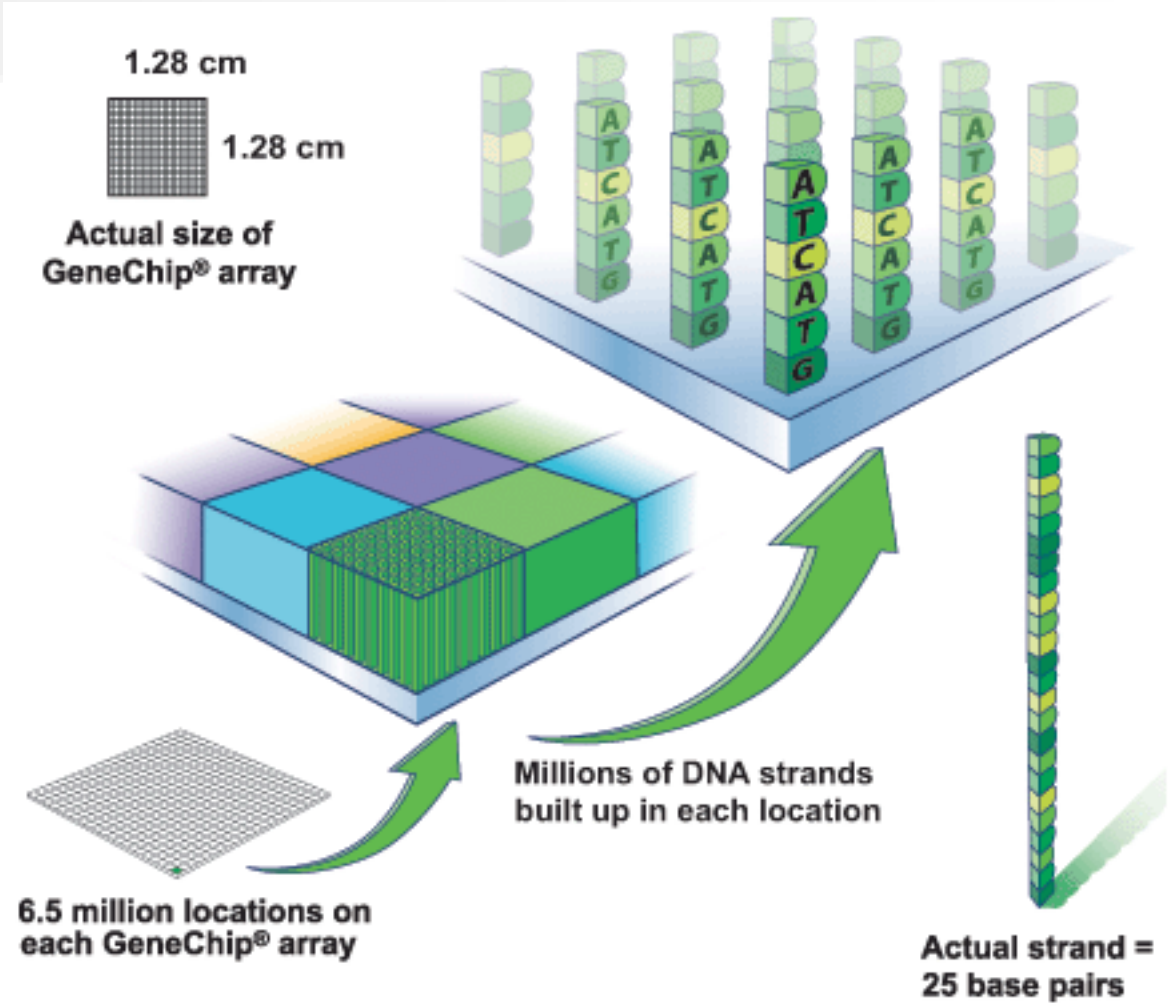


Image courtesy of Affymetrix.

Affymetrix Technology

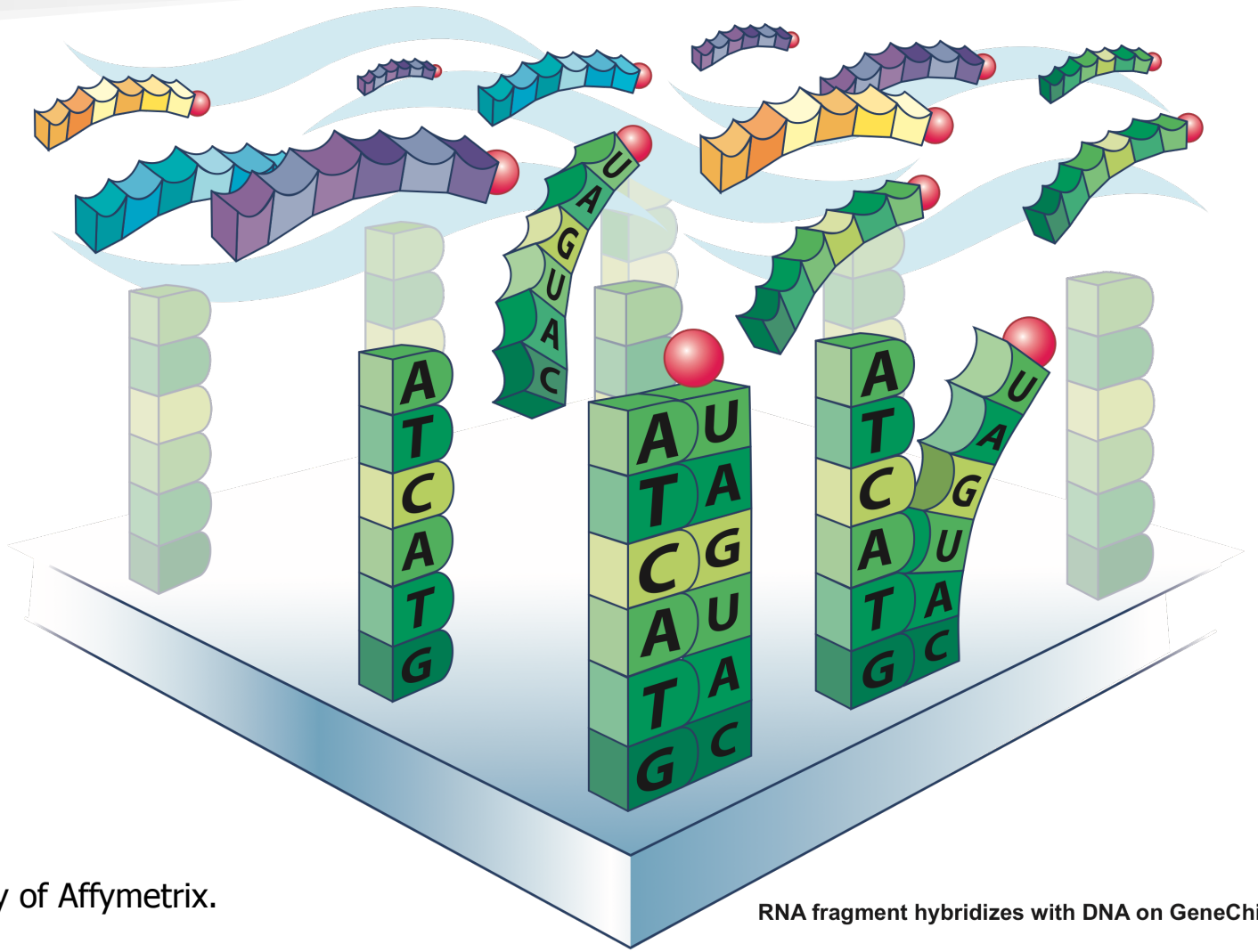


Image courtesy of Affymetrix.

Affymetrix Technology

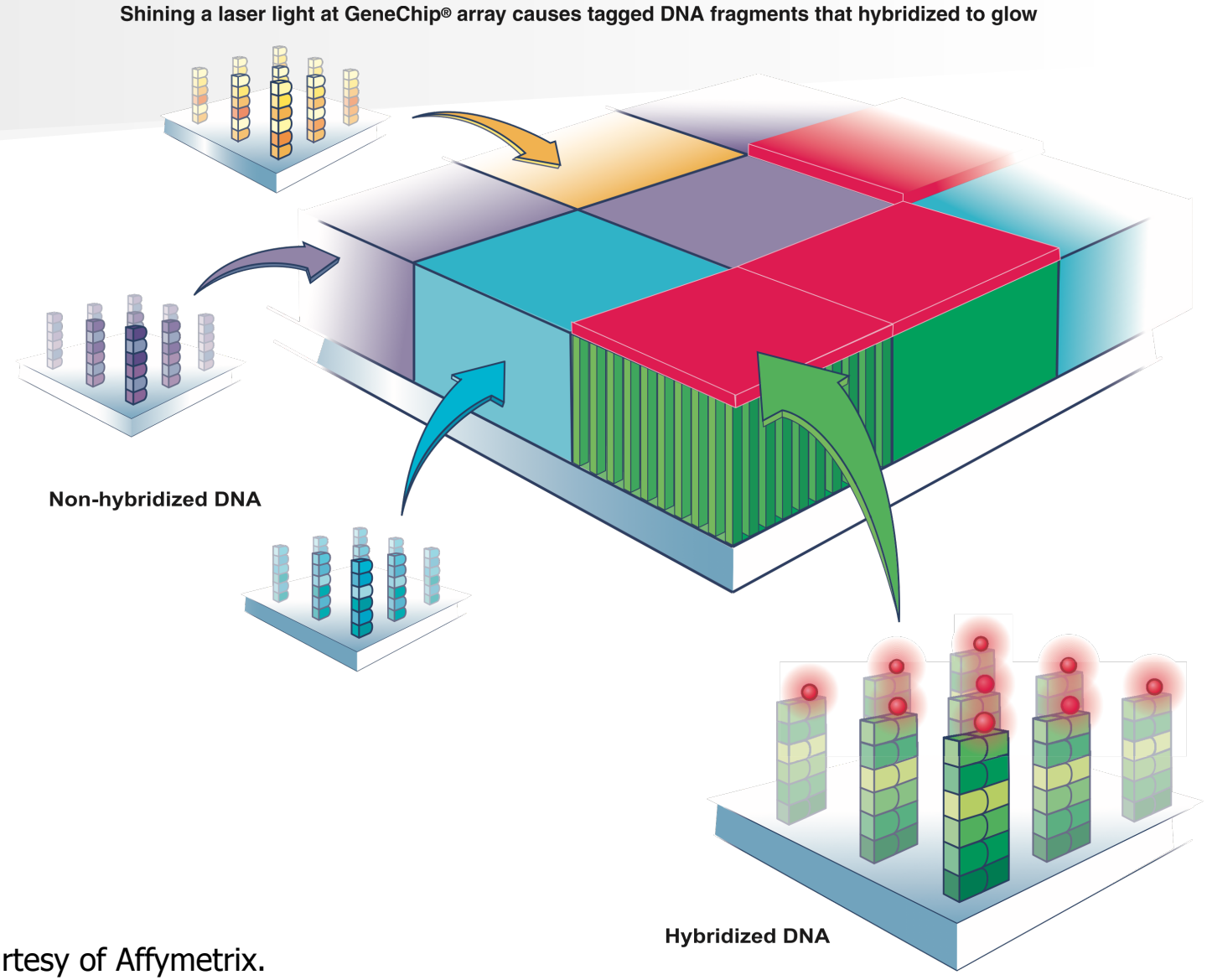
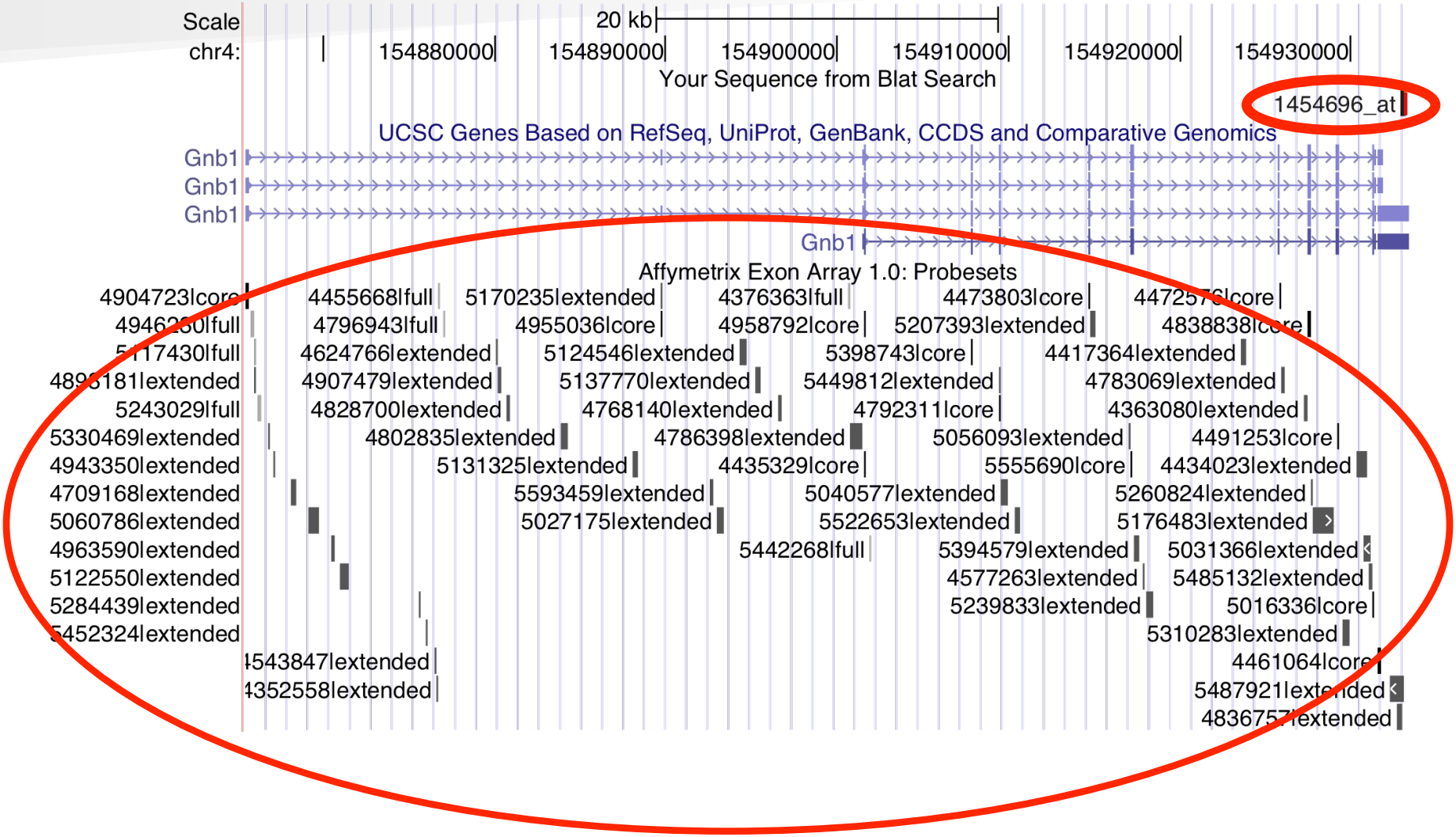
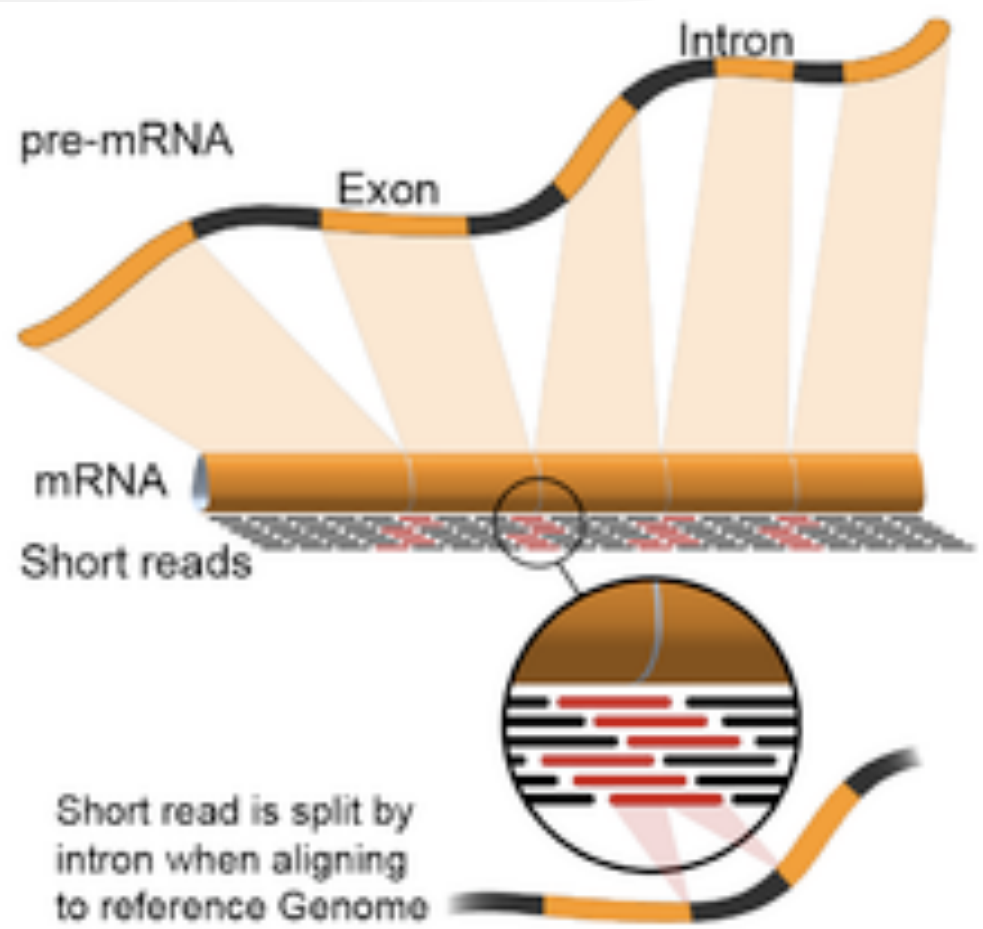


Image courtesy of Affymetrix.

Expansion of Microarray Technology

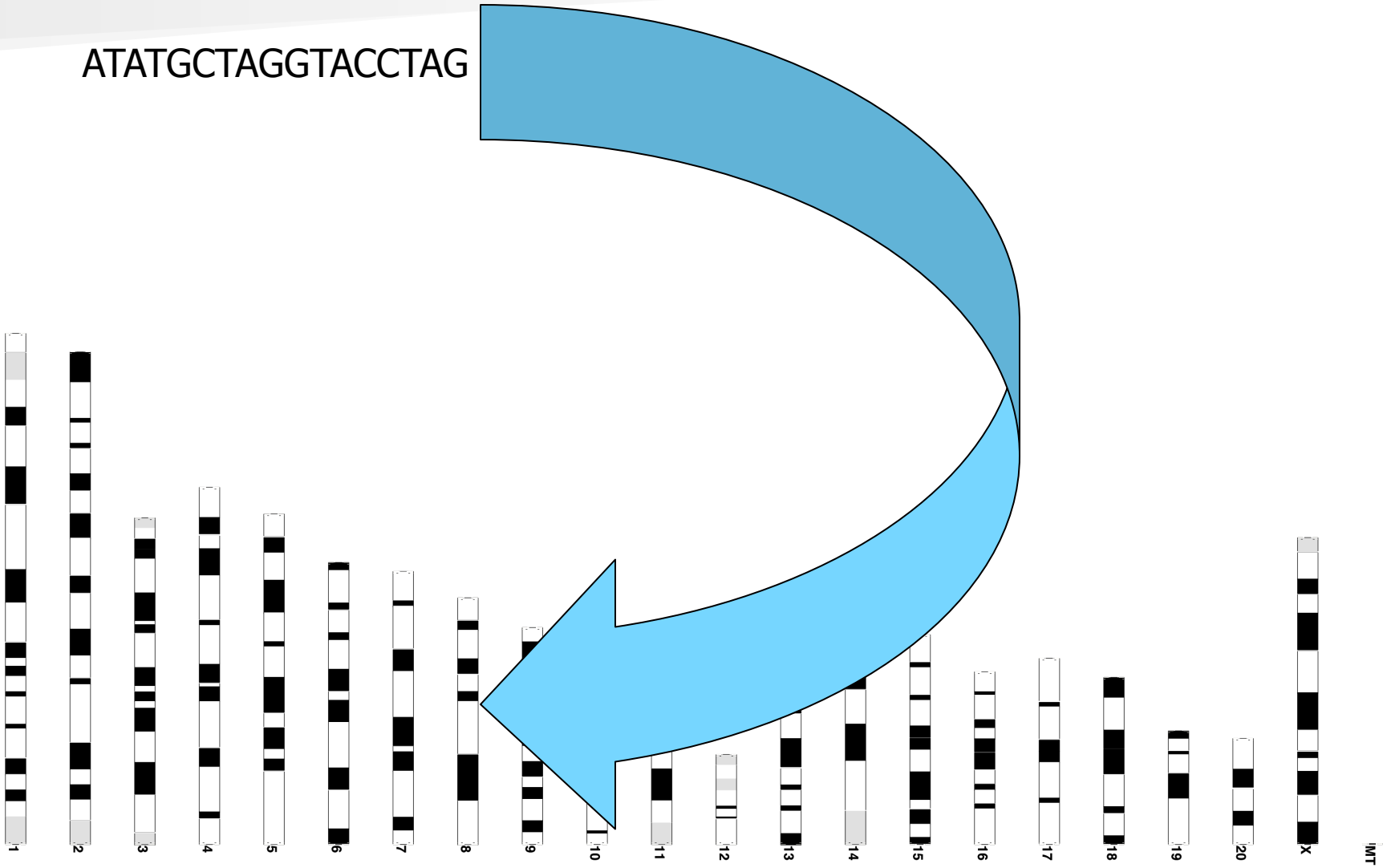


RNA-Seq

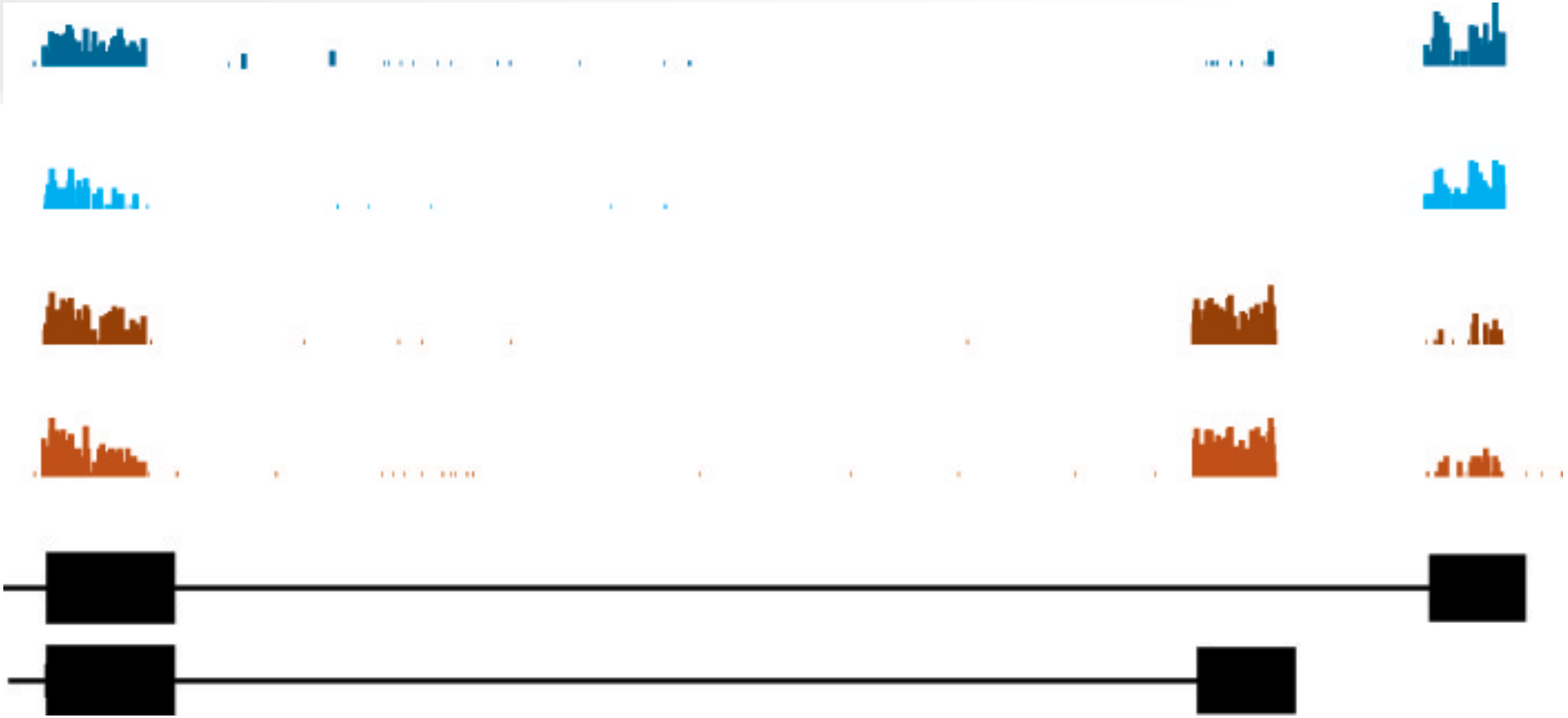

Intron
pre-mRNA
Exon
mRNA
Short reads
Short read is split by
intron when aligning
to reference Genome

Alignment

ATATGCTAGGTACCTAG



Quantification



Next Generation RNA Sequencing

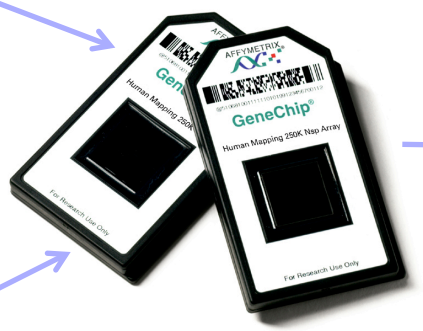
- Benefits
  - Unlimited dynamic range
  - Not restricted to "targeted" genomic regions
  - Can compare expression of one gene to another
- Disadvantages
  - Large amounts of data
  - Large amounts of data
  - $$ and throughput
  - Informatics tools

Topics

I. RNA Expression Technologies
II. Example - Genomic Drivers of Alcohol Consumption
III. Resources
IV. Glimpse into Co-Expression Networks

Microarray Analysis

Genomic Drivers of Alcohol Consumption



Genes that predispose to differing levels of alcohol consumption

The (Hypothetical) Order of Things

1. Preprocessing of Expression Data
2. Quality Control
3. Pre-Analysis Filtering
4. Differential Expression Analysis
5. Multiple Testing Correction
6. Gene List Interpretation

Preprocessing - Masking

- Array designs are OLD (at least in genome annotation time)
- Bad Probes:
  - Do not match to the reference genome
  - Match the reference genome in several spots
  - Contain SNPs that effect hybridization
- ELIMINATE THEM!!!

Pre-Processing Steps

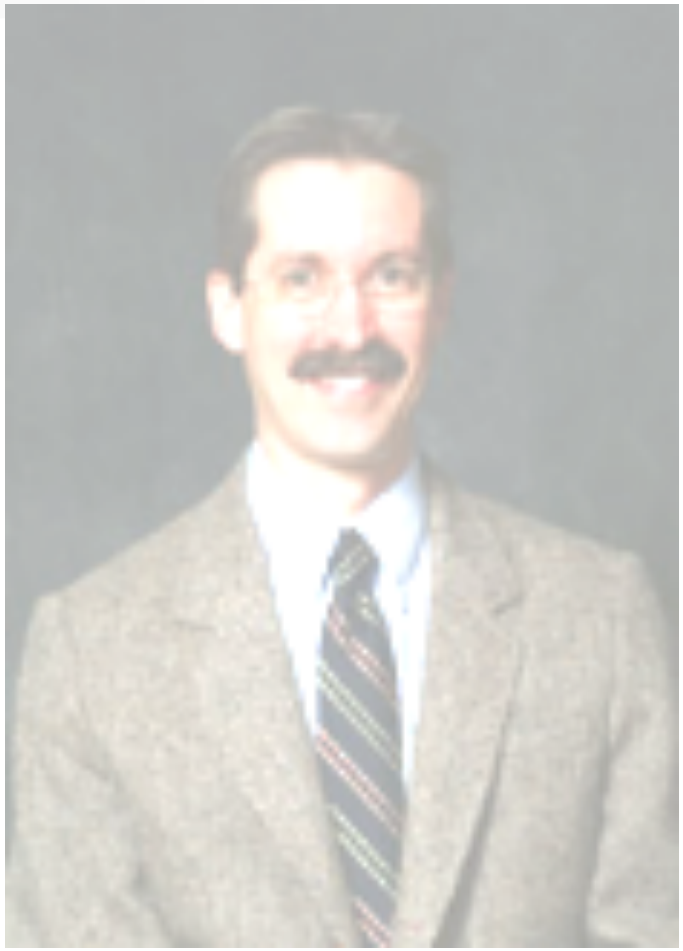
1. Background correction.
2. Data normalization.
3. Data summary methods (Affymetrix arrays only).

Most common methods include PLIER (Affymetrix's in-house algorithm) and RMA.

Preprocessing - Background Correction

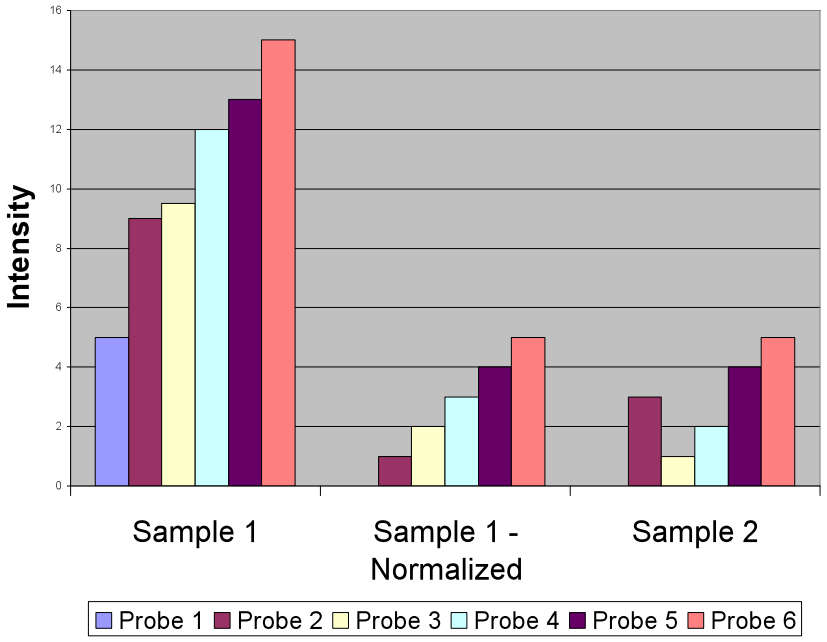
- Previous generation – perfect match versus mismatch
- Exon Array – mismatch probe identical in GC contents not necessarily identical sequence
- The popular RMA method uses neither

Preprocessing - Data Normalization



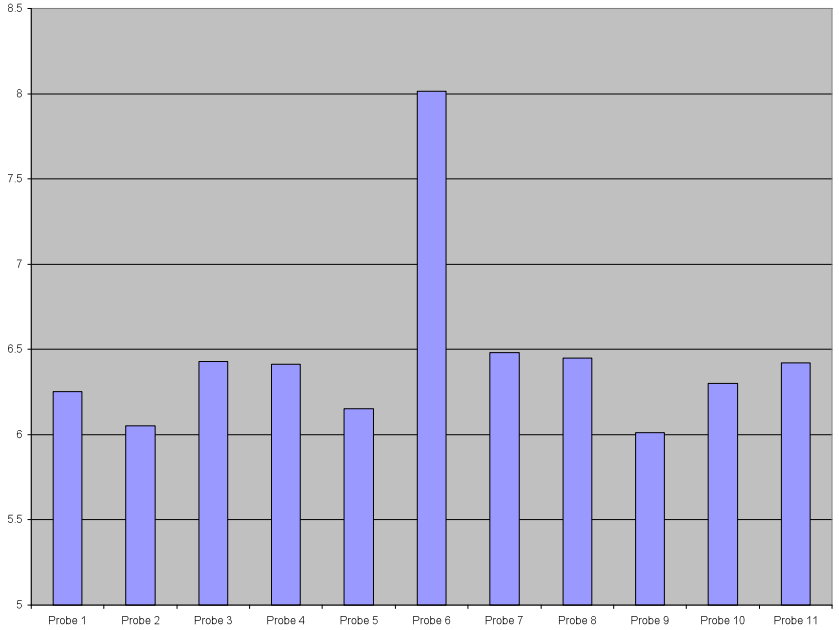
Preprocessing - Data Normalization

- Assumes that the large majority of transcripts on the array will not be different among subjects
- Most normalization is done on the log base 2 values of expression
- Common - Quantile Normalization



Preprocessing - Data Summary Methods

- 4 - 16 probes make up 1 probe set
- Robust summary lets us account for ROGUE probes

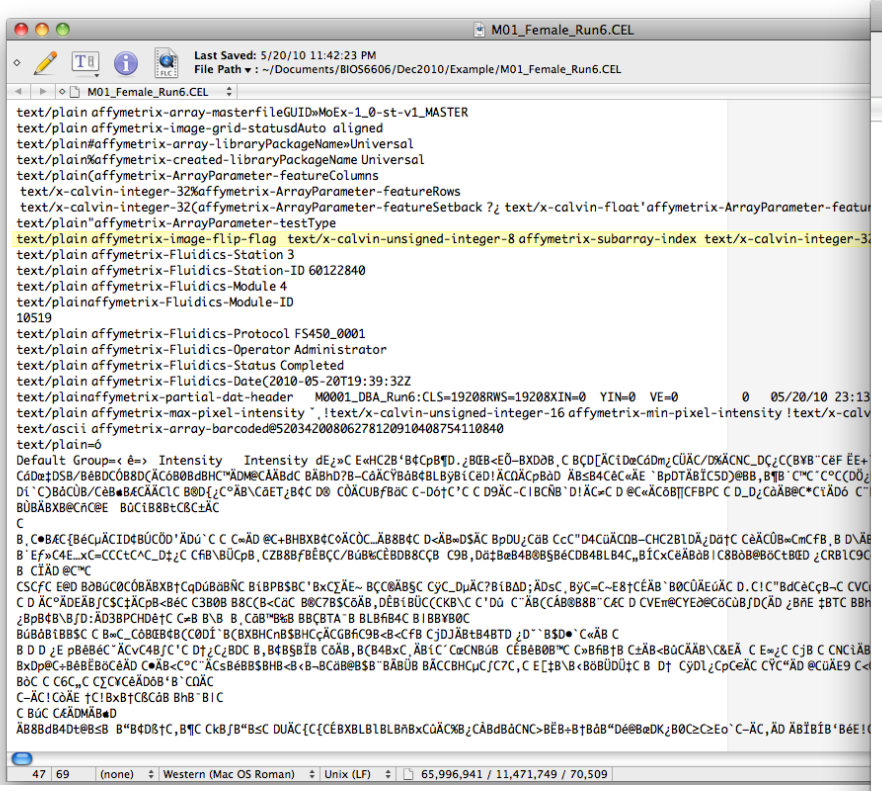


Preprocessing – Data Summary Methods

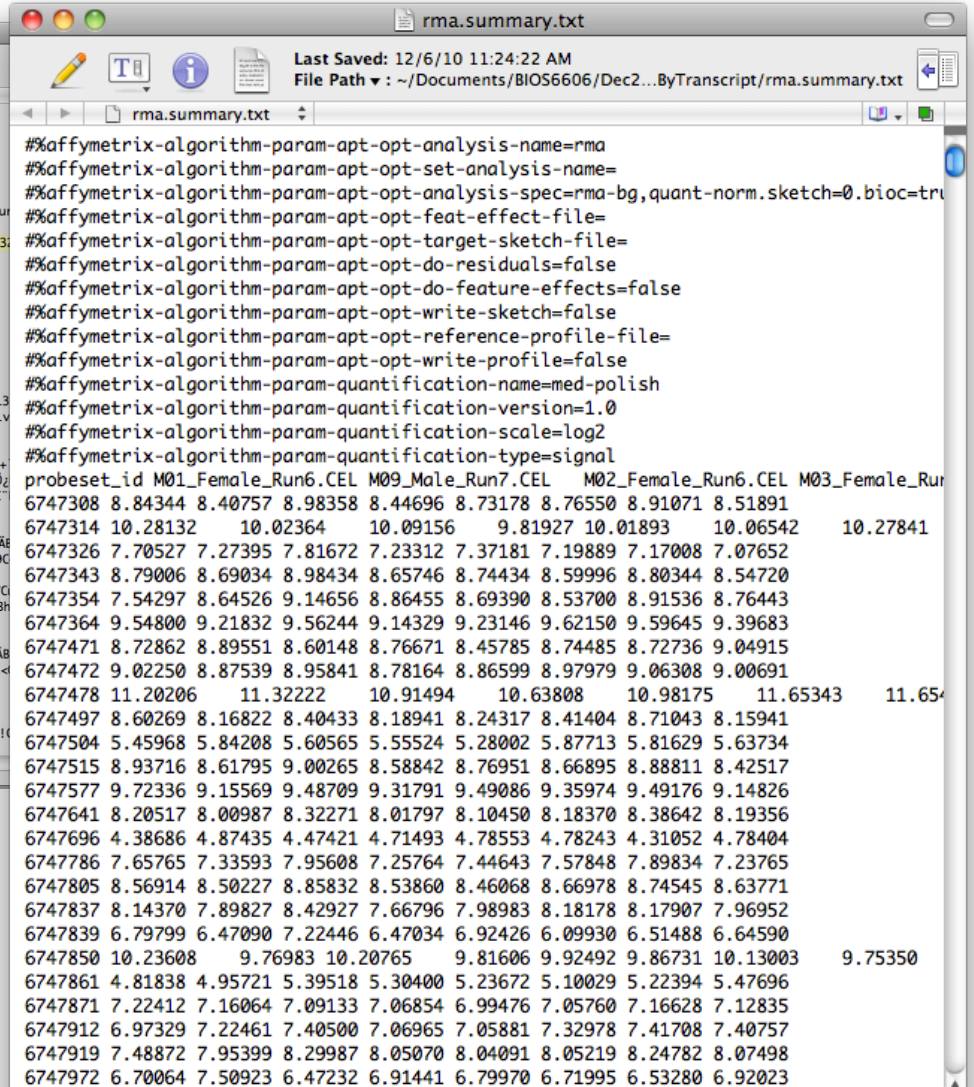
- Level of Analysis
  - Probe Set (approximately equivalent to exon)
  - Transcript Cluster (supposedly exons that are contained in all known isoforms of that gene)
- Confidence in Annotation
  - Core
  - Extended
  - Full

Example – Affymetrix Power Tools

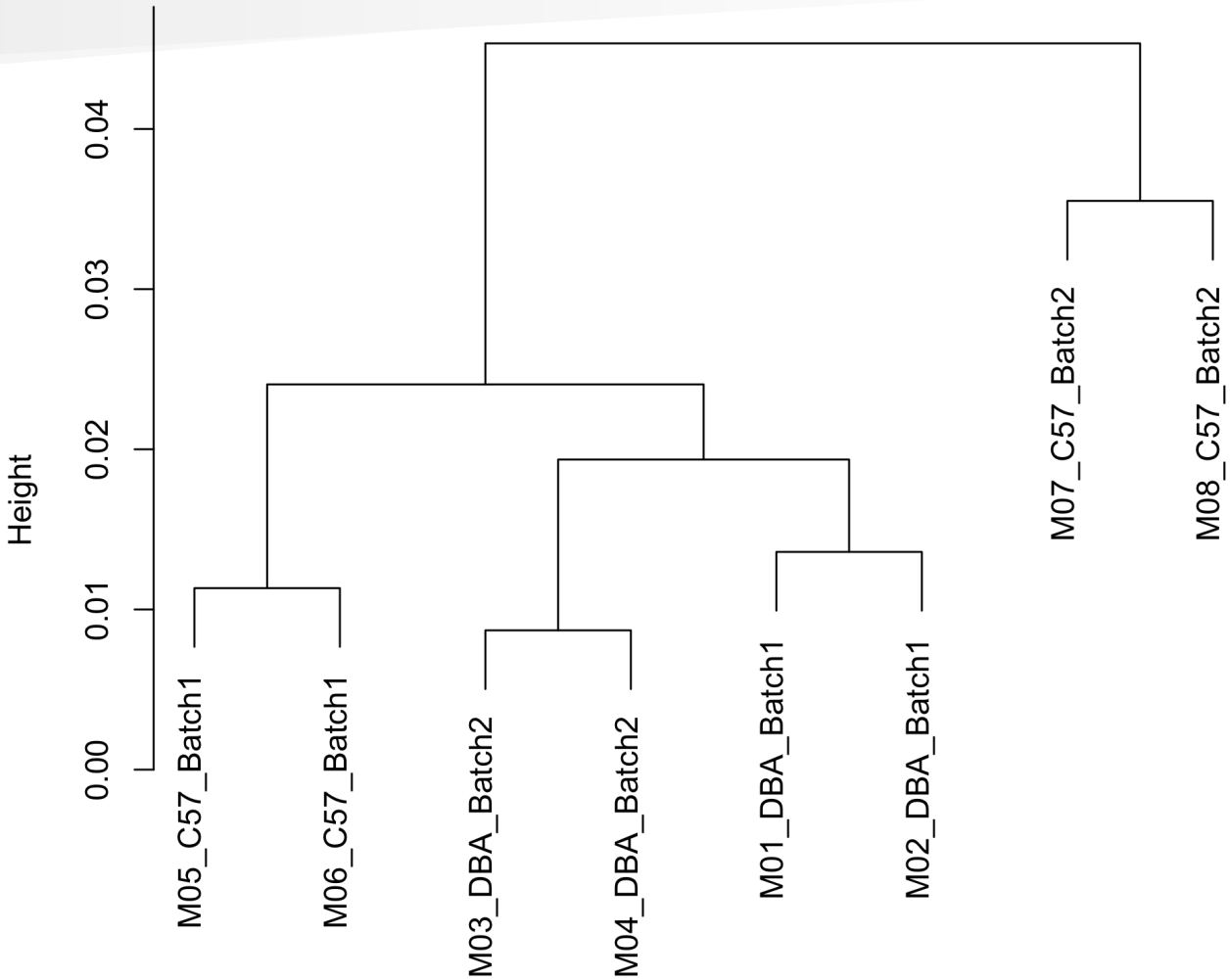
- Before



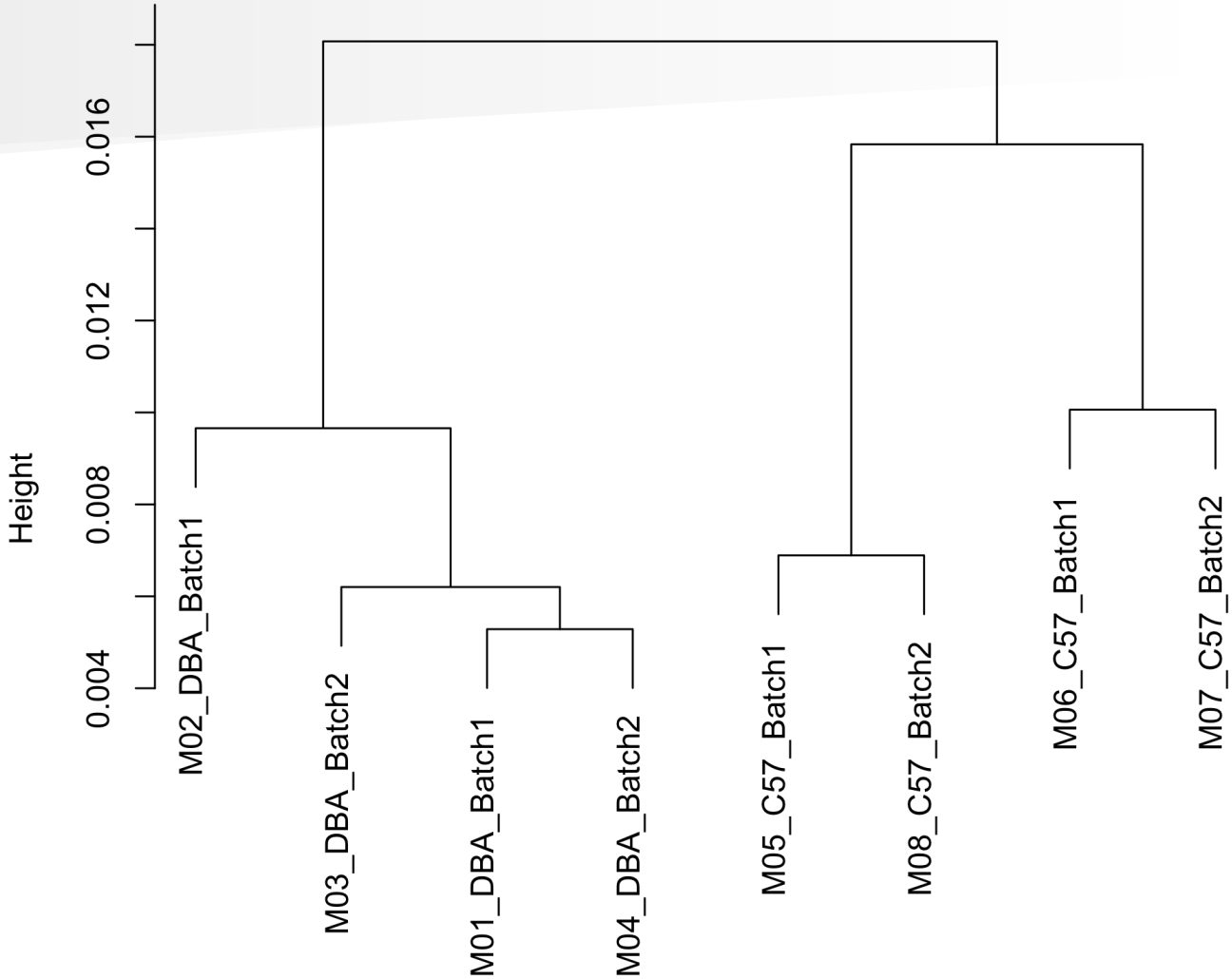
- After



Quality Control – Based on Study Design



Quality Control – Batch Adjustment



Pre-Analysis Filtering

- Eliminate Control Probes
- Present/Absent Calls (DABG)
- Quality Probe Sets
- Genes of Interest

Differential Expression

- Transcript by transcript basis
  - Can use any of the methods you have learned (as long as assumptions are met; keep in mind you can use tools such as permutation and bootstrapping)
  - Specialized methods for "borrowing" information across genes
  - Multiple Testing Consequences!!
- By groups of transcripts
  - Summarize probe sets into groups before analysis
  - Multiple testing not quite as harsh

Multiple Testing

- Definition of p-value: probability of observing data this extreme when the null hypothesis is true
- In other words, probability of declaring a significant difference when there isn't one (FALSE POSITIVE).

Multiple Testing Example

- 10 index cards within a bowl
  - 9 cards have simple words of encouragement
  - 1 card says STARBUCKS and can be turned in for a $5 Starbucks gift card.

- What is the probability that you get a Starbucks gift card?
- If everyone in the class draws one card (with replacement), how many cards will I have to give away?

Multiple Testing

- Assuming there are 30 students and the probability of winning a gift card is 10%, then we would expect I will have to give out 3 gift cards.
- We test 30,000 genes for differential expression
  - if all 30,000 genes are not truly differentially expressed, 1,500 genes will have p-values less than 0.05 by chance alone.

False Discovery Rate

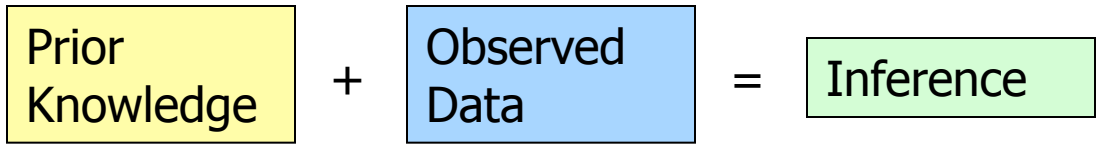
- Many times for genetic studies, we use a false discovery rate (FDR) rather than a traditional p-value to help account for multiple comparisons.
- FDR is the proportion of "significant" tests that are false positives.
- An FDR value is calculated for each test (e.g., gene), but it is dependent on the distribution of the other test results (e.g., other genes).
- When we use a 5% FDR threshold for significance, we are estimating that 5% of the significant genes are false positives.

Transcript by Transcript

- Pros
  - Super easy to implement
  - May find that causal gene (good luck)
  - Can apply a variety of specialized statistics
- Cons
  - Sooo many comparisons
  - Limited samples (do I believe my estimates?)
  - No multiple testing correction perfectly fits the microarray data

Borrowing Information

- Empirical Bayes
  - Calculating variance on three samples per group?!
  - "Shrinks" transcript level variance towards the common variance across transcripts



Grouping Transcripts

- Group Prior to Statistical Analysis
  - Cluster based on expression profile or some biological reason (Gene Ontology)
  - Create of summary expression profile for group and run analysis on summary measure
- Group After Analysis But Before Interpretation
  - summarize p-values among transcripts in same group

How do I decide which probes/ probesets are differentially expressed?

- Strict statistical evaluation – which genes are significant after multiple testing correction?
- Add some quantitative criteria – must have a fold change above 2
- But what if I just need 10 genes to take on to further testing? Depends.

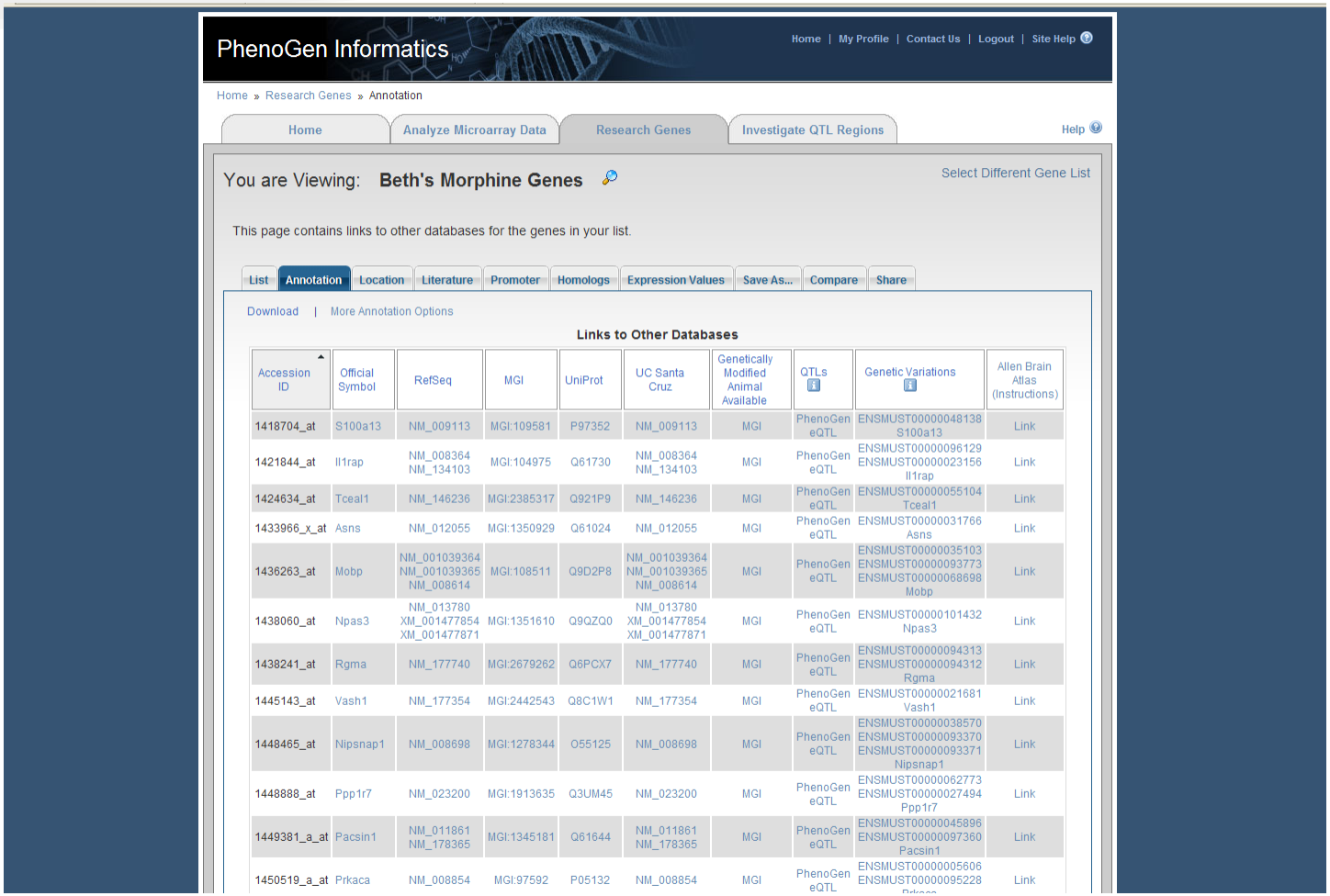
Hypothesis Generating NOT Hypothesis Testing

Example – Differences Between Mouse Strains

Transcript Cluster ID	log2 Mean - C57BL/6J	log2 Mean - DBA/2J	log2 Fold Change	unadjusted p-value	FDR
6775762	6.47	9.11	-2.65	3.29E-09	1.03E-05
6804268	8.01	9.92	-1.92	8.06E-09	2.00E-05
6830174	7.81	8.75	-0.94	1.65E-08	2.73E-05
6854540	8.40	9.45	-1.05	7.07E-08	8.08E-05
6854990	7.92	6.55	1.37	1.23E-08	2.61E-05
6860163	8.09	8.77	-0.68	8.63E-08	9.16E-05
6860165	6.98	8.59	-1.61	1.81E-09	8.97E-06
6890290	7.10	8.27	-1.17	1.52E-08	2.73E-05
6901353	8.59	7.32	1.27	5.74E-08	7.11E-05
6904367	7.37	5.46	1.90	3.01E-10	2.24E-06
6939005	9.12	10.45	-1.33	2.83E-08	4.20E-05
6955169	8.03	6.19	1.84	3.48E-09	1.03E-05
6975235	8.77	9.69	-0.92	3.17E-08	4.28E-05
6986000	6.89	10.96	-4.07	7.11E-13	1.06E-08

Which genes go to which probe sets?

- See handout for lots of websites



PhenoGen Informatics

Home | My Profile | Contact Us | Logout | Site Help

Home » Research Genes » Annotation

Home | Analyze Microarray Data | Research Genes | Investigate QTL Regions | Help

You are Viewing: **Beth's Morphine Genes** Select Different Gene List

This page contains links to other databases for the genes in your list.

List | Annotation | Location | Literature | Promoter | Homologs | Expression Values | Save As... | Compare | Share

Download | More Annotation Options

Links to Other Databases

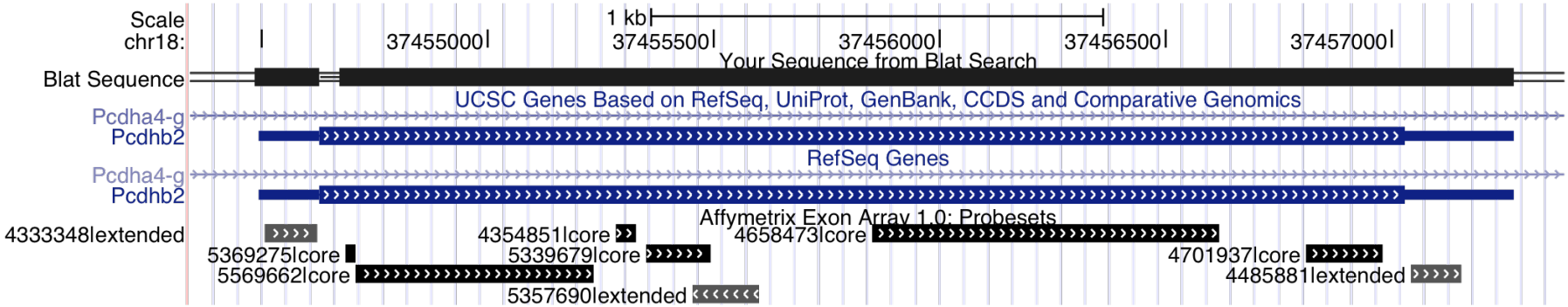
Accession ID	Official Symbol	RefSeq	MGI	UniProt	UC Santa Cruz	Genetically Modified Animal Available	QTLs	Genetic Variations	Allen Brain Atlas (Instructions)
1418704_at	S100a13	NM_009113	MGI:109581	P97352	NM_009113	MGI	PhenoGen eQTL	ENSMUST00000048138 S100a13	Link
1421844_at	Il1rap	NM_008364 NM_134103	MGI:104975	Q61730	NM_008364 NM_134103	MGI	PhenoGen eQTL	ENSMUST00000096129 ENSMUST00000023156 Il1rap	Link
1424634_at	Tceal1	NM_146236	MGI:2385317	Q921P9	NM_146236	MGI	PhenoGen eQTL	ENSMUST00000055104 Tceal1	Link
1433966_x_at	Asns	NM_012055	MGI:1350929	Q61024	NM_012055	MGI	PhenoGen eQTL	ENSMUST00000031766 Asns	Link
1436263_at	Mobp	NM_001039364 NM_001039365 NM_008614	MGI:108511	Q9D2P8	NM_001039364 NM_001039365 NM_008614	MGI	PhenoGen eQTL	ENSMUST00000035103 ENSMUST00000093773 ENSMUST00000068698 Mobp	Link
1438060_at	Npas3	NM_013780 XM_001477854 XM_001477871	MGI:1351610	Q9QZQ0	NM_013780 XM_001477854 XM_001477871	MGI	PhenoGen eQTL	ENSMUST00000101432 Npas3	Link
1438241_at	Rgma	NM_177740	MGI:2679262	Q6PCX7	NM_177740	MGI	PhenoGen eQTL	ENSMUST00000094313 ENSMUST00000094312 Rgma	Link
1445143_at	Vash1	NM_177354	MGI:2442543	Q8C1W1	NM_177354	MGI	PhenoGen eQTL	ENSMUST00000021681 Vash1	Link
1448465_at	Nipsnap1	NM_008698	MGI:1278344	O55125	NM_008698	MGI	PhenoGen eQTL	ENSMUST00000038570 ENSMUST00000093370 ENSMUST00000093371 Nipsnap1	Link
1448888_at	Ppp1r7	NM_023200	MGI:1913635	Q3UM45	NM_023200	MGI	PhenoGen eQTL	ENSMUST00000062773 ENSMUST00000027494 Ppp1r7	Link
1449381_a_at	Pacsin1	NM_011861 NM_178365	MGI:1345181	Q61644	NM_011861 NM_178365	MGI	PhenoGen eQTL	ENSMUST00000045896 ENSMUST00000097360 Pacsin1	Link
1450519_a_at	Prkaca	NM_008854	MGI:97592	P05132	NM_008854	MGI	PhenoGen eQTL	ENSMUST00000005606 ENSMUST00000095228	Link

Affymetrix Annotation

Transcript Cluster ID	Evidence Level	Gene Title	Gene Symbol	Cytoband	Number of Exon Clusters	Transcript Classification
6775762	core	stabilin 2	Stab2	10q23.1	92	full-length
6804268	core	cytochrome c oxidase subunit VIIa polypeptide 2-like	Cox7a2l	12p12.1	1	full-length
6830174	core	Smg-5 homolog, nonsense mediated mRNA decay factor pseudogene	A930017M01Rik	15q15.3	22	full-length
6854540	core	expressed sequence AI413582	AI413582	17q11.2	8	full-length
6854990	core	histocompatibility 2, K1, K region	H2-K1	17q12	19	full-length
6860163	core	protocadherin beta 2	Pcdhb2	18q12.3	2	full-length
6860165	core	protocadherin beta 3	Pcdhb3	18q12.3	1	full-length
6890290	core	phospholipase A2, group IVE	Pla2g4e	2q14.2	35	full-length
6901353	core	alanine-glyoxylate aminotransferase 2-like 1	Agxt2l1	3q22.1	38	full-length
6904367	core	predicted gene 5148	Gm5148	3p22.2	6	partial
6939005	core	gamma-aminobutyric acid (GABA) A receptor, subunit alpha 2	Gabra2	5q13.2	34	full-length
6955169	core	camello-like 2	Cml2	6q14.3	8	full-length
6975235	core	glutathione reductase	Gsr	8p12	22	full-length
6986000	core	melanoma antigen	Mela	8q24.13	5	full-length

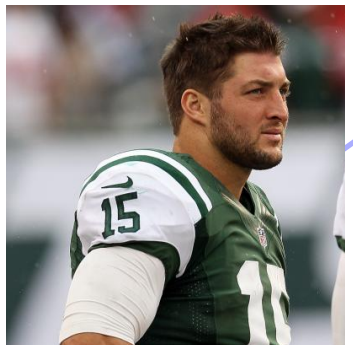
To Do Before Interpreting Gene List

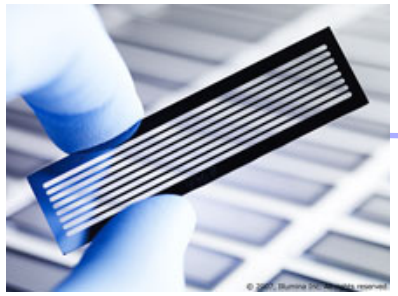
- Look up probe set ID in another database, e.g. UCSC Genome Browser



RNA-Seq Analyses

Genomic Drivers of Alcohol Consumption



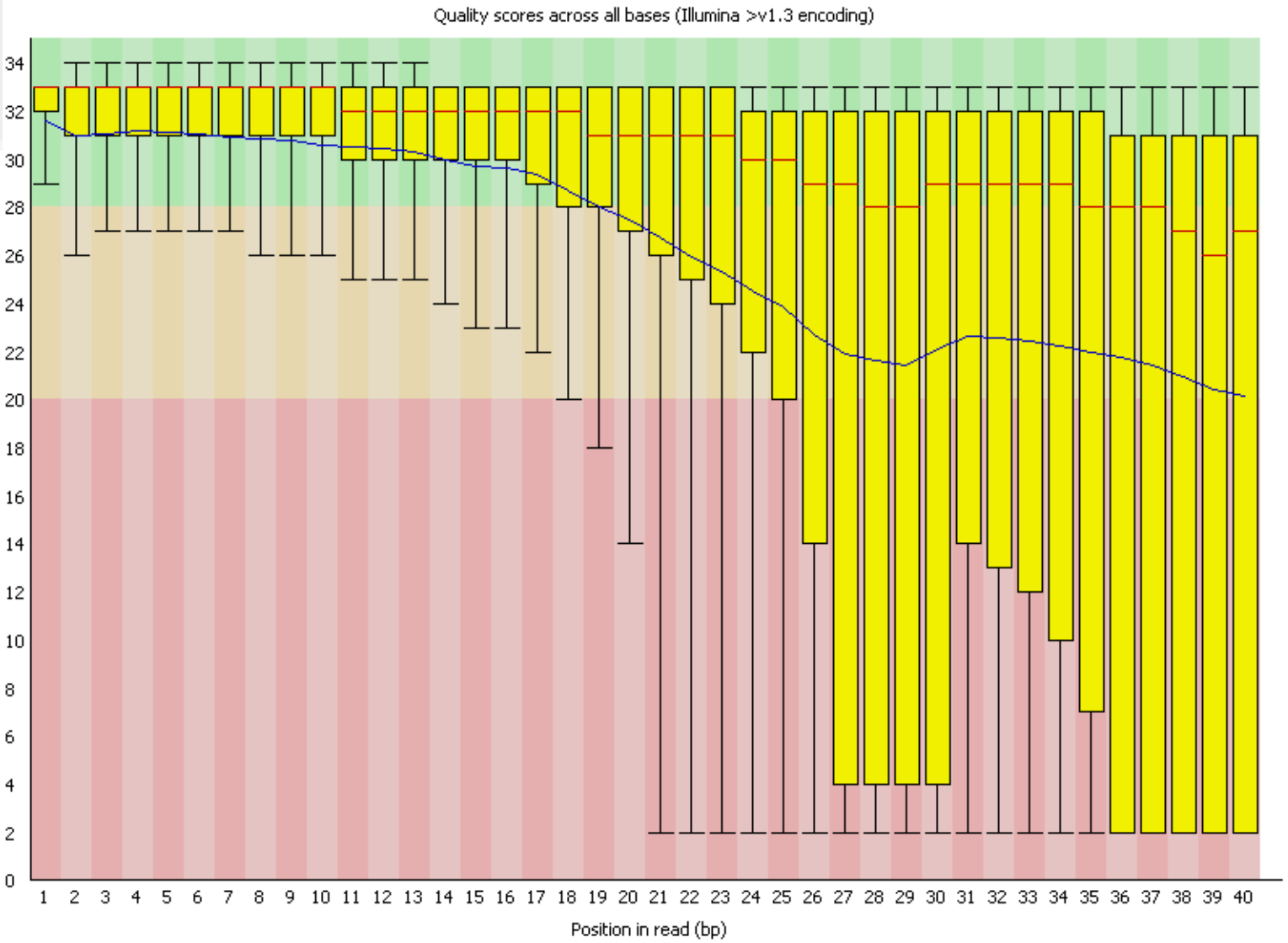


Genes that predispose to differing levels of alcohol consumption

The (Hypothetical) Order of Things

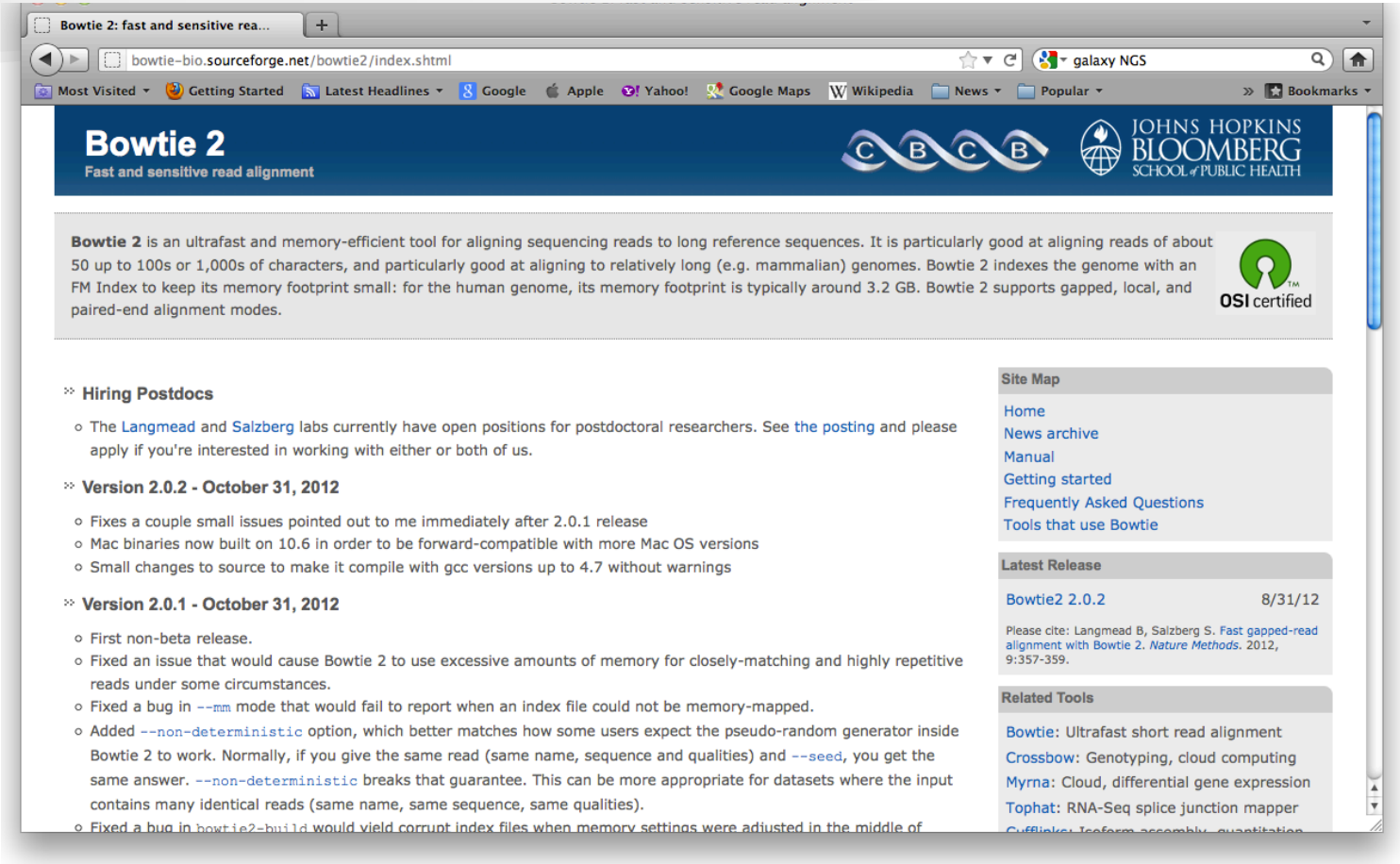
- Trim reads
- Align reads to genome
- Identify Transcripts
- Quantify Expression
- Differential Expression Analysis
- Multiple Testing Correction
- Gene List Interpretation

Trimming Reads

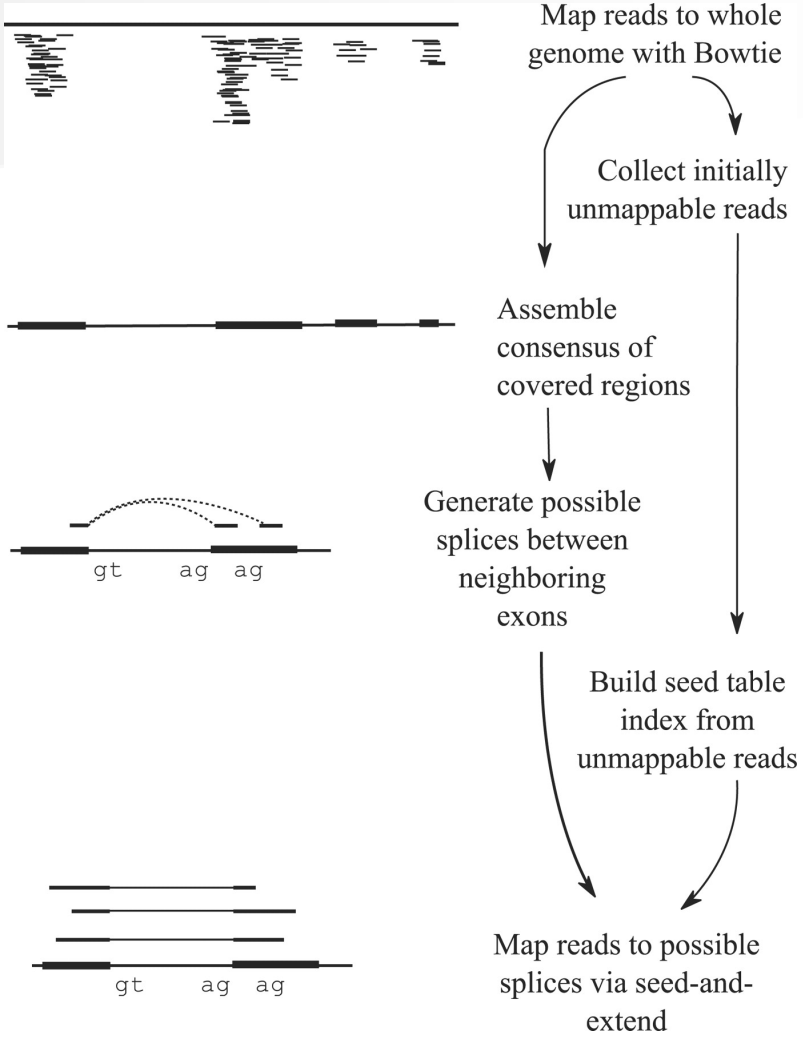


http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Aligning Reads

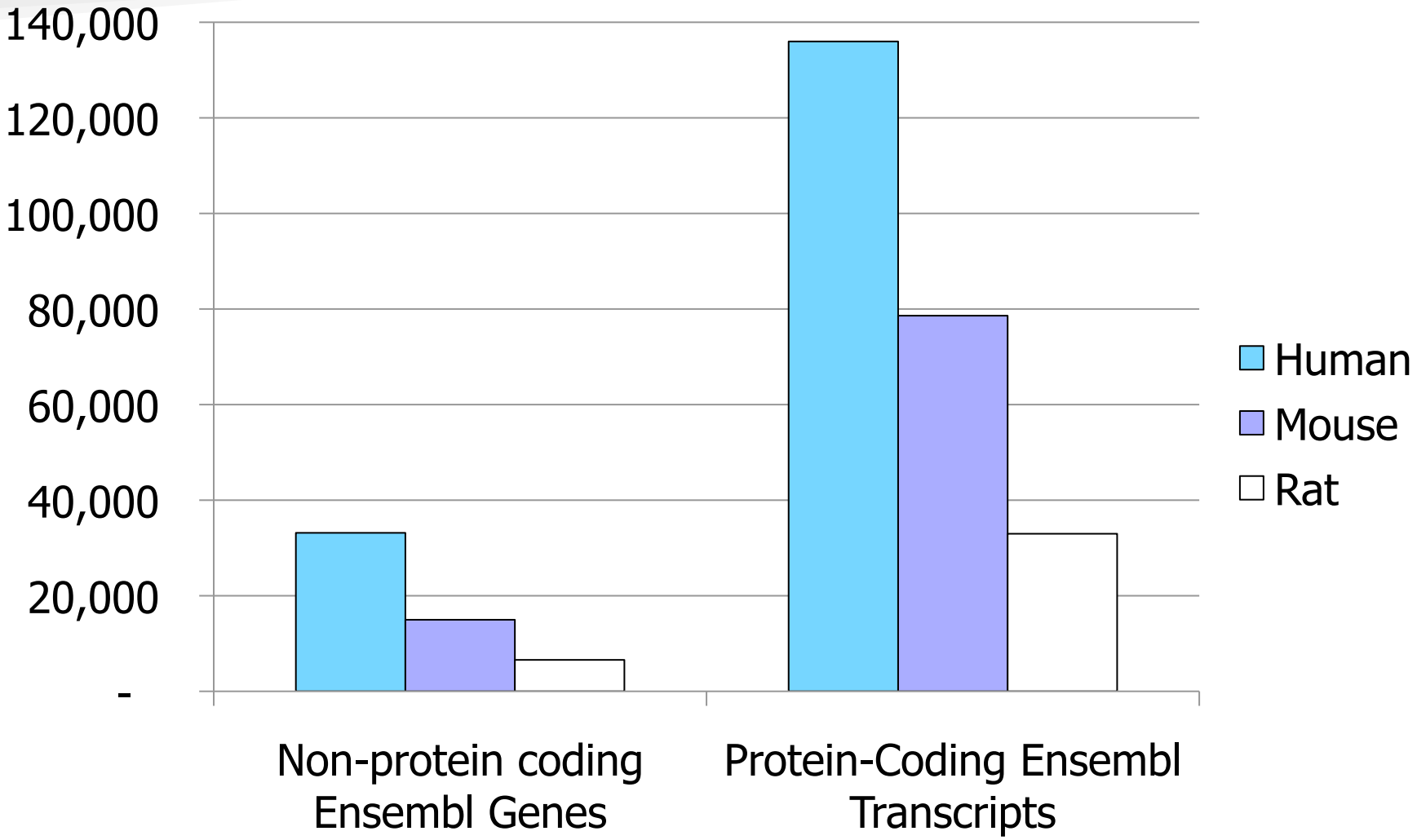


The TopHat pipeline.

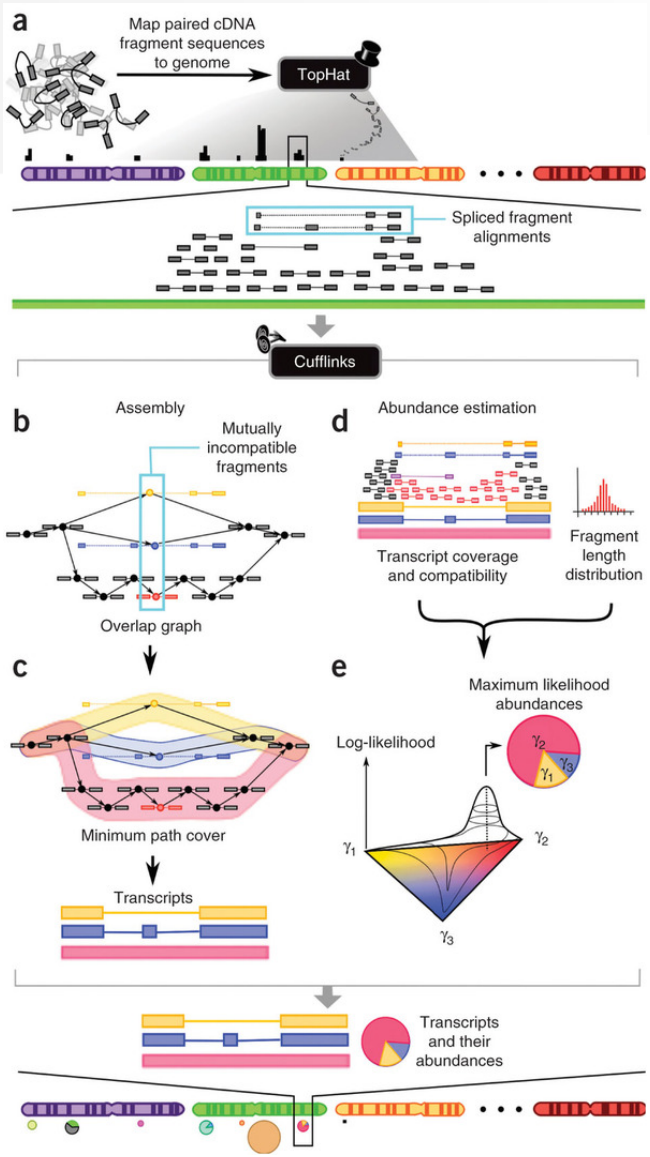


Trapnell C et al. Bioinformatics 2009;25:1105-1111

Why Transcriptome Reconstruction?



Transcriptome Reconstruction



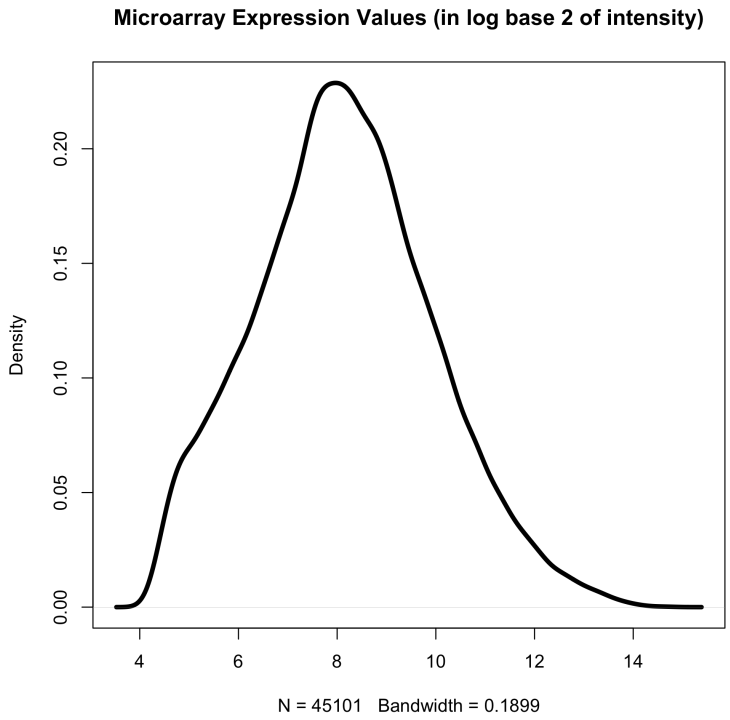
CuffLinks – Trapnell et al (2010), Nature Biotechnology 28(5): 511-5.

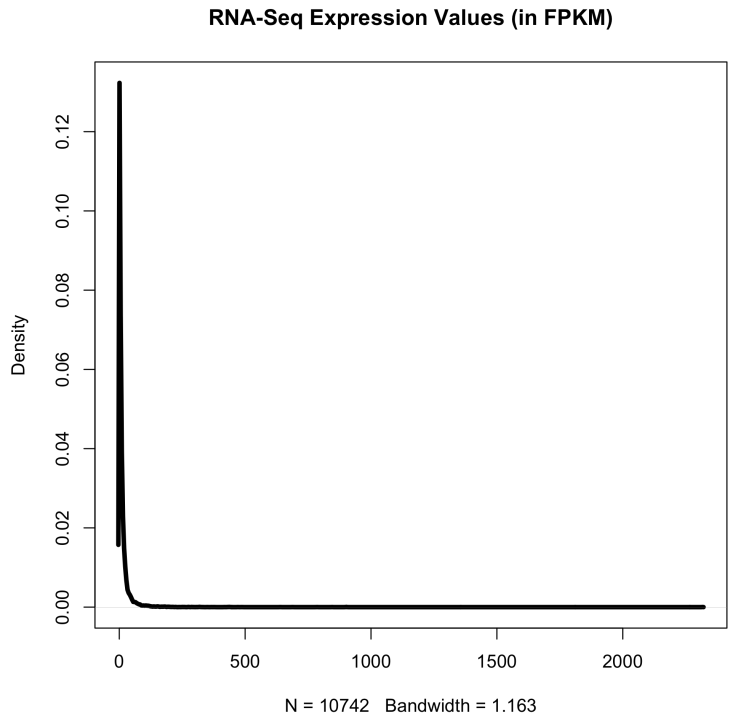
Quantifying Expression

- Count Number of Reads
  - multiple transcripts per read (Bayesian Deconvulsion methods – CuffLinks)
  - Different number of total reads per sample
    - original – divide by total mappable reads (RPKM - CuffLinks)
    - ongoing research – what should the denominator be?

Quantifying Expression

- No longer normal…





Differential Expression

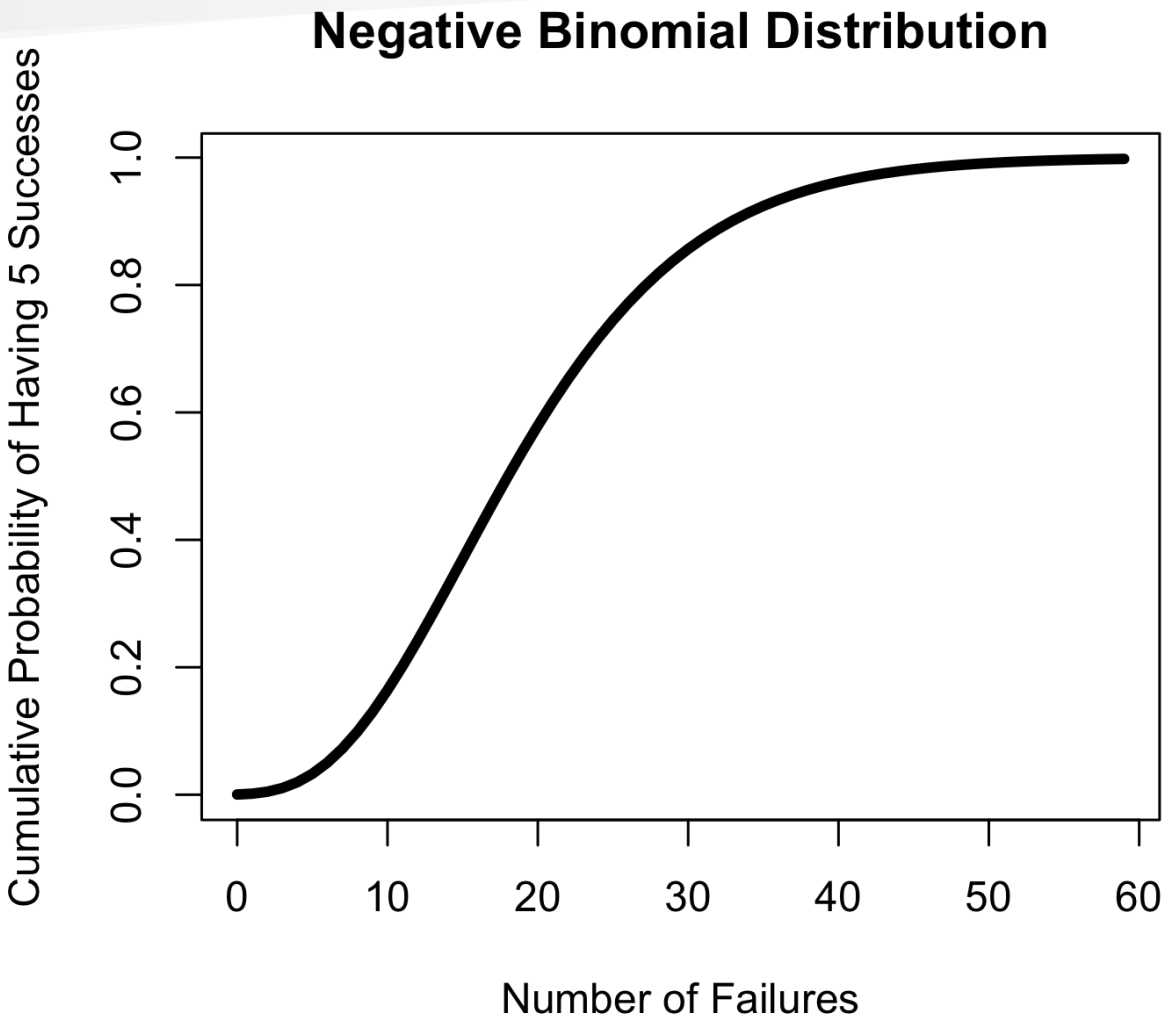
Sample A	Sample B
1	2
100	200
1100	1200
0	2
0	2000

Negative Binomial Distribution

- Goal – Find 5 people that have seen the movie Office Space
  - Probability of having seen Office Space is 20%

How many people will I have to ask before I find 5 people?

Negative Binomial Distribution



Methods For Differential Expression

- Can we estimate the "proportion of people who have seen Office Space" from how many failures before we have 5 success?
- Goal in RNA-Seq – identify genes expressed in different "proportions" in to populations

CuffDiff – uses a negative binomial distribution for error estimates and accounts for uncertainty in reads counts

Many methods out there with no clear WINNER!

Problem – Defining a "Failure"

Interpreting Gene Lists

Interpreting Gene Lists

- Location in Genome
- Gene Ontology Similarities
- Literature Search (co-citation)
- Shared Transcription Factor
- Pathway Analysis (e.g. PathwayExpress)

Affymetrix Annotation

Transcript Cluster ID	Evidence Level	Gene Title	Gene Symbol	Cytoband	Number of Exon Clusters	Transcript Classification
6775762	core	stabilin 2	Stab2	10q23.1	92	full-length
6804268	core	cytochrome c oxidase subunit VIIa polypeptide 2-like	Cox7a2l	12p12.1	1	full-length
6830174	core	Smg-5 homolog, nonsense mediated mRNA decay factor pseudogene	A930017M01Rik	15q15.3	22	full-length
6854540	core	expressed sequence AI413582	AI413582	17q11.2	8	full-length
6854990	core	histocompatibility 2, K1, K region	H2-K1	17q12	19	full-length
6860163	core	protocadherin beta 2	Pcdhb2	18q12.3	2	full-length
6860165	core	protocadherin beta 3	Pcdhb3	18q12.3	1	full-length
6890290	core	phospholipase A2, group IVE	Pla2g4e	2q14.2	35	full-length
6901353	core	alanine-glyoxylate aminotransferase 2-like 1	Agxt2l1	3q22.1	38	full-length
6904367	core	predicted gene 5148	Gm5148	3p22.2	6	partial
6939005	core	gamma-aminobutyric acid (GABA) A receptor, subunit alpha 2	Gabra2	5q13.2	34	full-length
6955169	core	camello-like 2	Cml2	6q14.3	8	full-length
6975235	core	glutathione reductase	Gsr	8p12	22	full-length
6986000	core	melanoma antigen	Mela	8q24.13	5	full-length

Literature Review

GABRG1 and *GABRA2* Variation Associated with Alcohol Dependence in African Americans

Chupong Ittiwut, Bao-Zhu Yang, Henry R. Kranzler, Raymond F. Anton,
Rung

Role of *GABRA2* in Moderating Subjective Responses to Alcohol

Sungwon Roh, Sachio Matsushita, Sachiko Hara, Hitoshi Maesato, Toshifumi Matsui,
Go Suzuki, Tomohiro Miyakawa, Vijay A. Ramchandani, Ting-Kai Li, and Susumu Higuchi

Genetic association study of *GABRA2* single nucleotide polymorphisms and electroencephalography in alcohol dependence

G.J. Lydall[a], J. Saini[b], K. Ruparelia[b], S. Montagnese[b,1], A. McQuillin[a], I. Guerrini[c], H. Rao[e], G. Reynolds[f], D. Ball[d], I. Smith[g], A.D. Thomson[a,c], M.Y. Morgan[b], H.M.D. Gurling[a,*]

I found the Alcoholism Gene(s)!

- Well maybe not…
  - Could be genes for
    - Gender
    - Size
    - Hair color
    - Depression
    - Anxiety
    - Aggression
    - Logic…

Topics

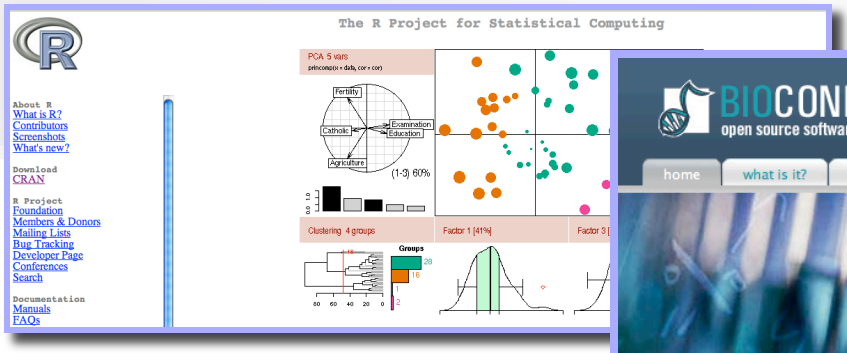
I. RNA Expression Technologies
II. Example - Genomic Drivers of Alcohol Consumption
III. Resources
IV. Glimpse into Co-Expression Networks

Microarray Resources

Most Popular Microarray Software

- R and BioConductor
- BRB-ArrayTools (Excel plug-in)
- Partek Genomics Suite
- Affymetrix Expression Console / Affymetrix Power Tools
- PhenoGen Website

R and BioConductor



www.r-project.org

www.bioconductor.org

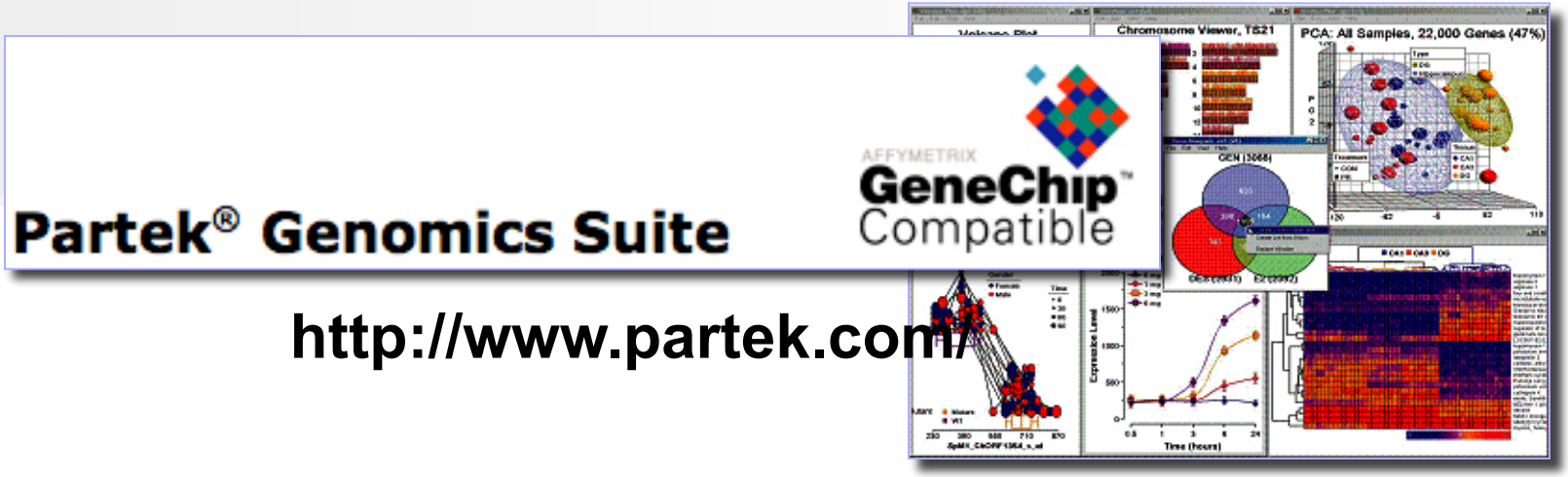
- ✓ Open-source software, free for the public
- ✓ Platform independent
- ✓ Unlimited statistical and programming flexibility
- ✗ Need command line usage experience
- ✗ Requires significant R programming experience

BRB-ArrayTools

http://linus.nci.nih.gov/BRB-ArrayTools.html

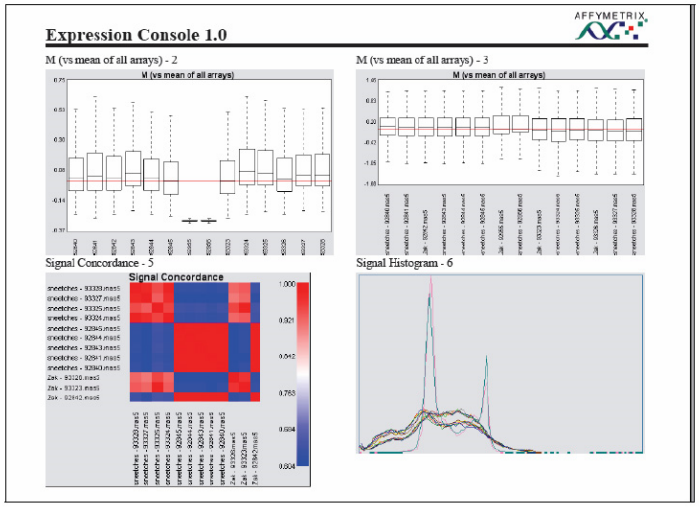
- ✓ Open-source software, free for the public
- ✓ Microsoft Excel plug-in
- × Only works on Windows platform
- × Imposed by all Excel limitations

Partek Genomics Suite



- ✓ Commercial software
- ✓ Run on Window and Linux machines
- ✓ Ideal for running large dataset (>100 chips)
- × Not free, share licensing
- × Doesn't have a good file management system

Affymetrix Expression Console

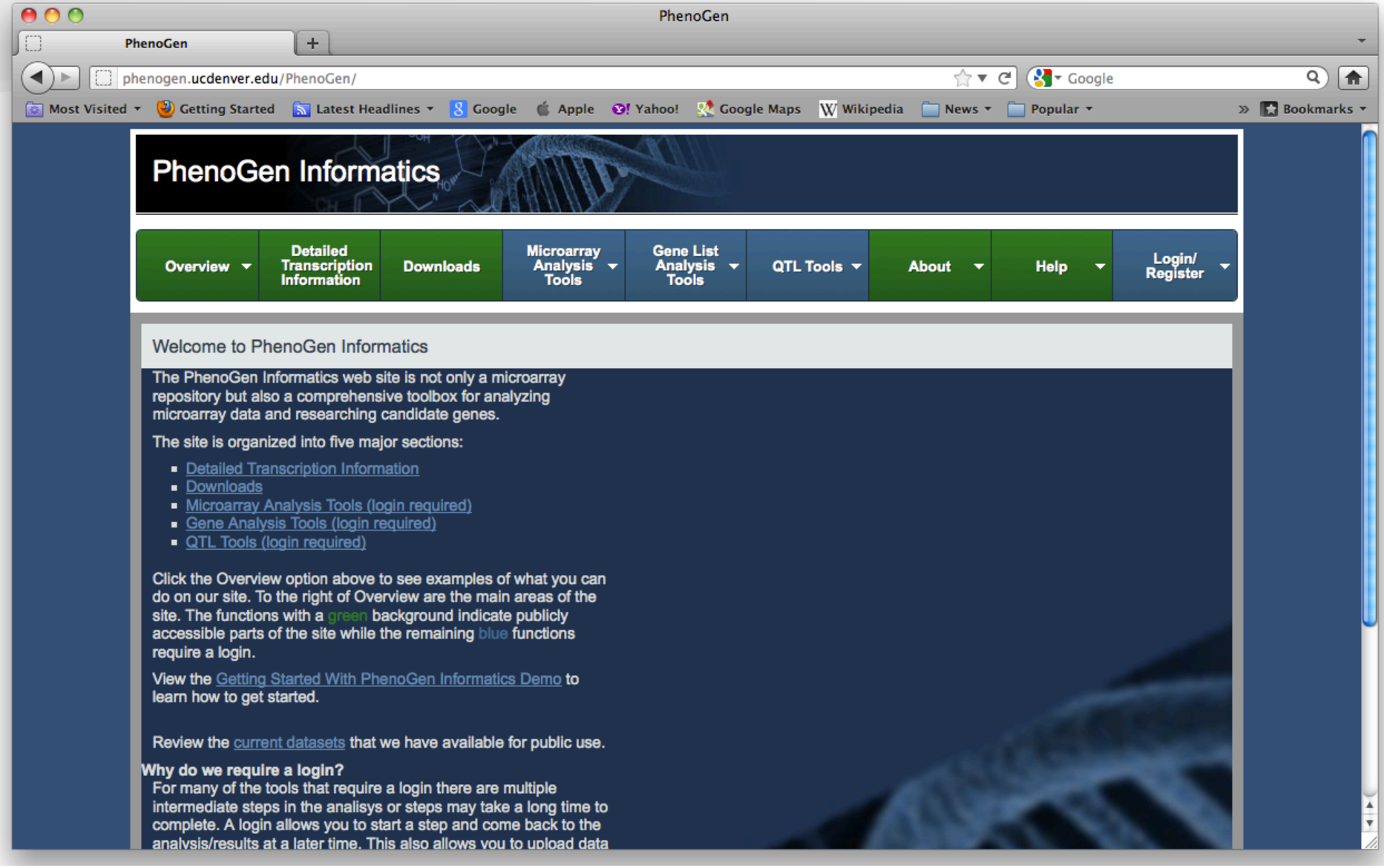


www.affymetrix.com/

- ✓ Commercial software
- ✓ Free
- × Not flexible
- × Does not replace software for secondary analyses

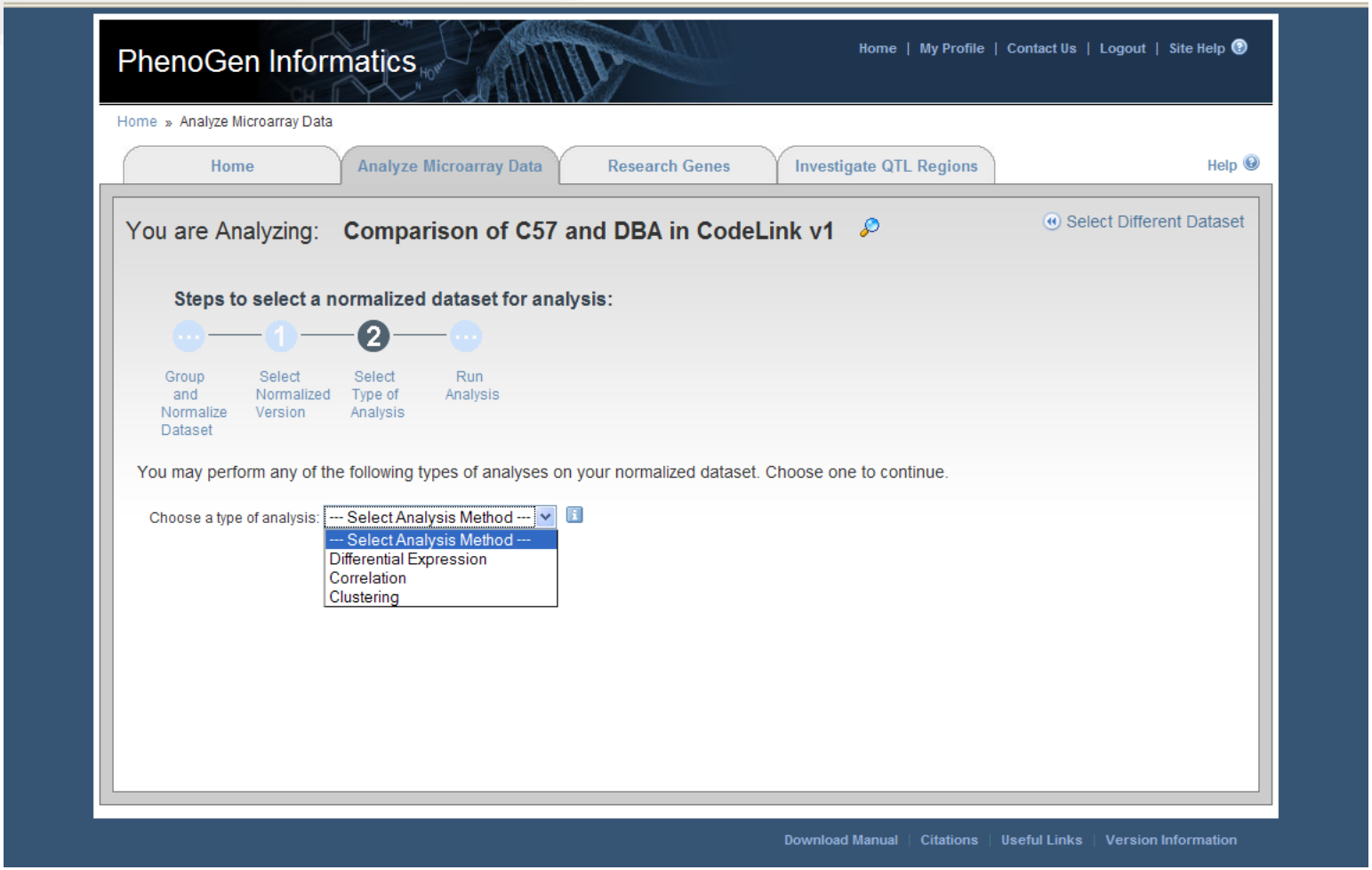
Our Resource

http://phenogen.ucdenver.edu/



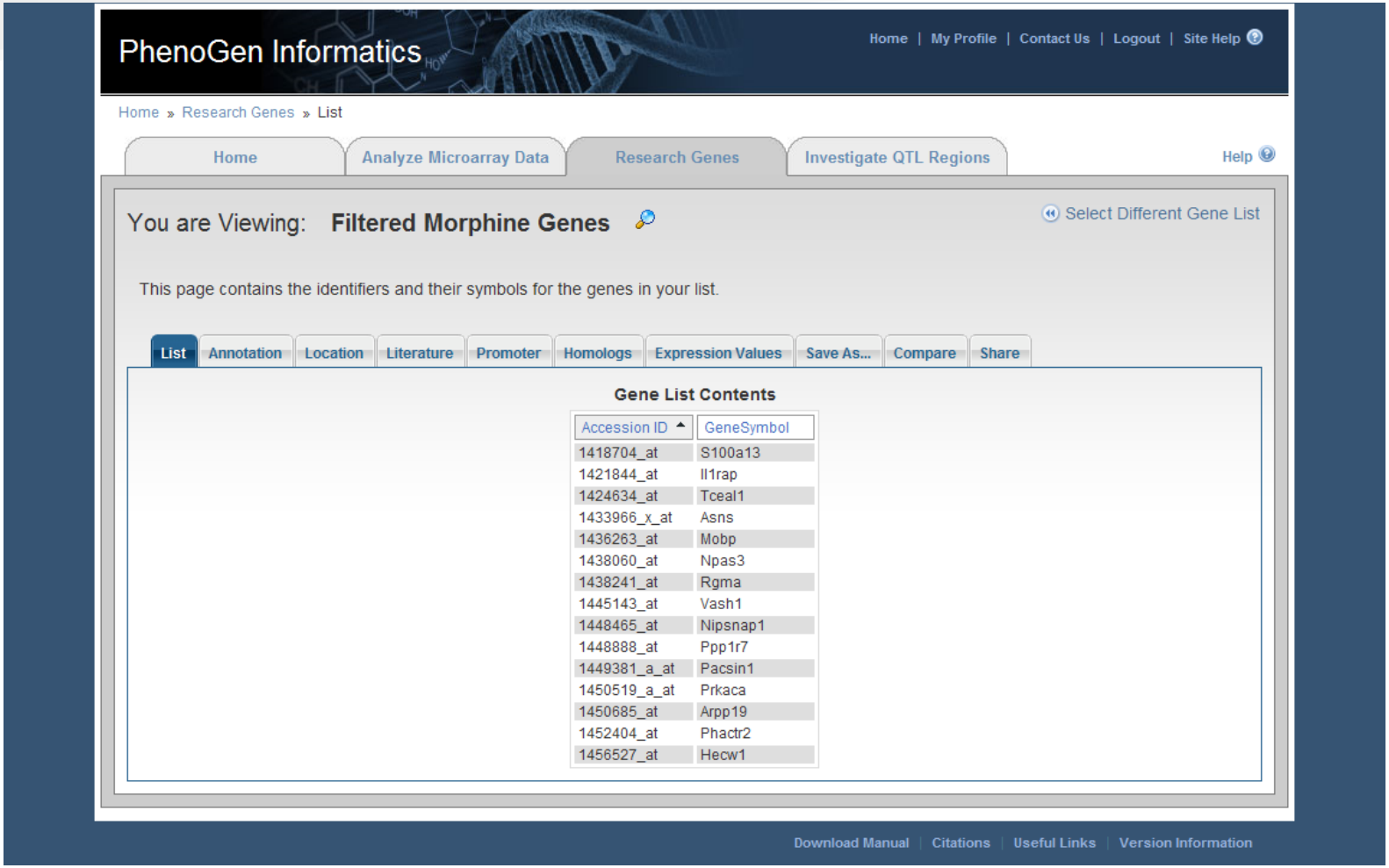
Our Resource

http://phenogen.ucdenver.edu/



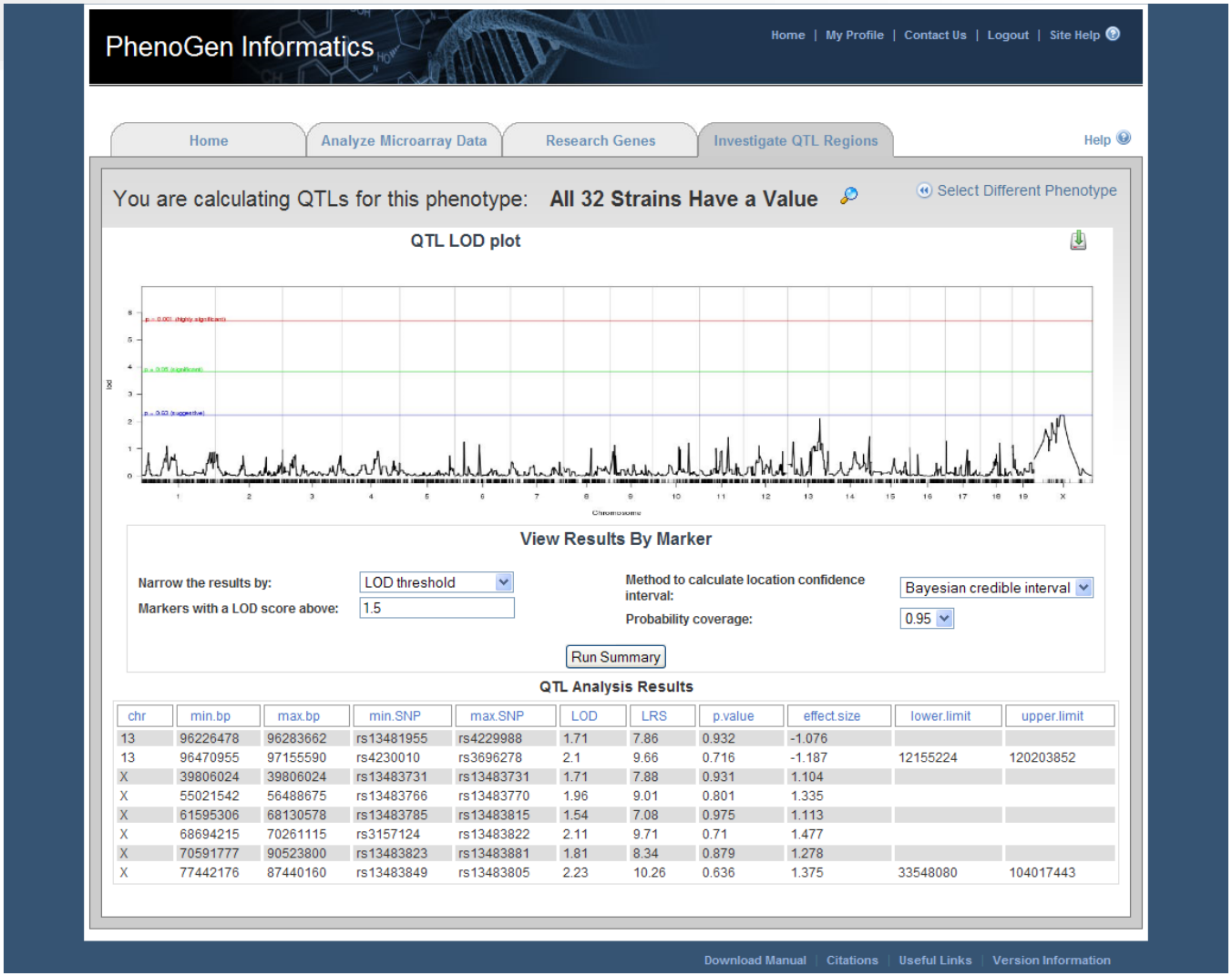
Our Resource

http://phenogen.ucdenver.edu/



Our Resource

http://phenogen.ucdenver.edu/



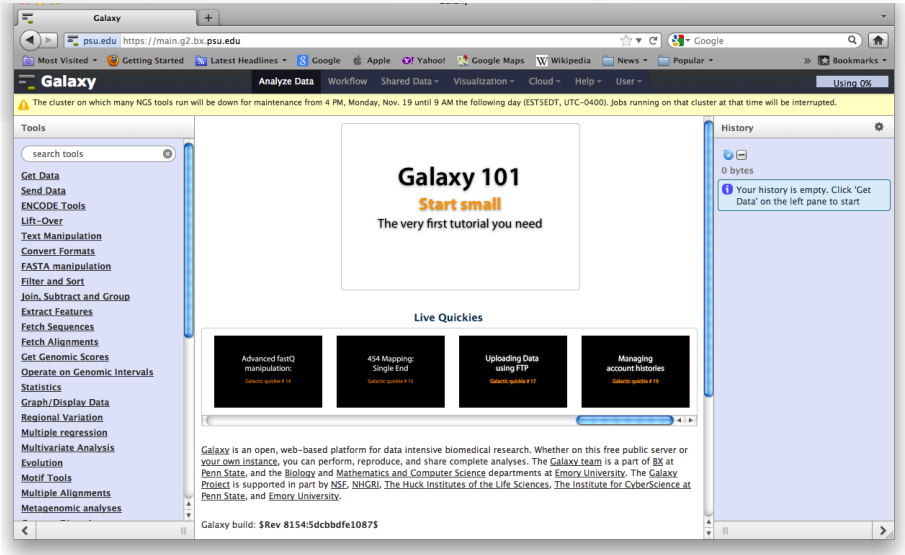
chr	min.bp	max.bp	min.SNP	max.SNP	LOD	LRS	p.value	effect.size	lower.limit	upper.limit
13	96226478	96283662	rs13481955	rs4229988	1.71	7.86	0.932	-1.076		
13	96470955	97155590	rs4230010	rs3696278	2.1	9.66	0.716	-1.187	12155224	120203852
X	39806024	39806024	rs13483731	rs13483731	1.71	7.88	0.931	1.104		
X	55021542	56488675	rs13483766	rs13483770	1.96	9.01	0.801	1.335		
X	61595306	68130578	rs13483785	rs13483815	1.54	7.08	0.975	1.113		
X	68694215	70261115	rs3157124	rs13483822	2.11	9.71	0.71	1.477		
X	70591777	90523800	rs13483823	rs13483881	1.81	8.34	0.879	1.278		
X	77442176	87440160	rs13483849	rs13483805	2.23	10.26	0.636	1.375	33548080	104017443

RNA-Seq Resources

Popular RNA-Seq Software

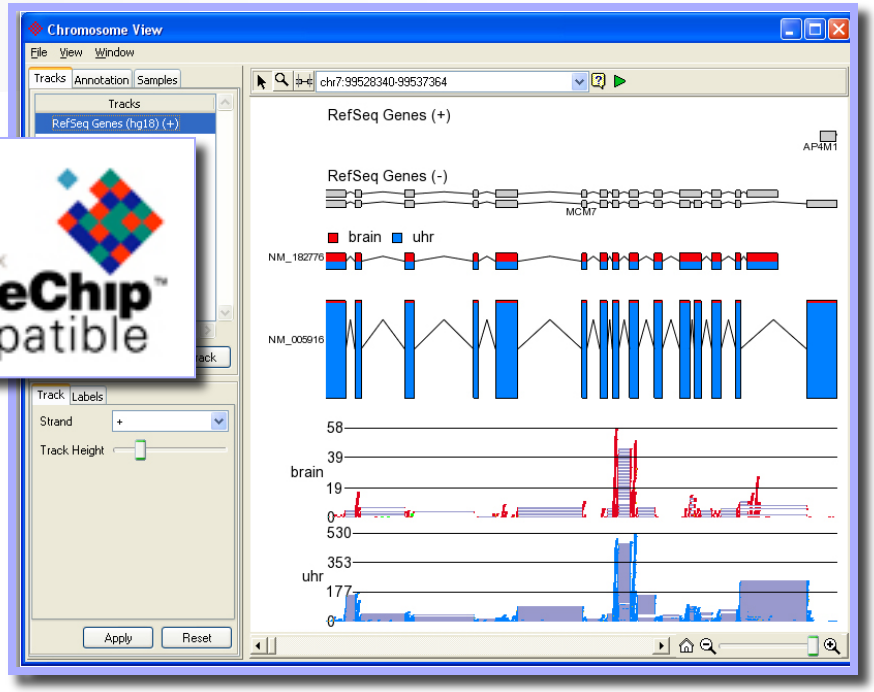
- Galaxy
- Partek Genomics Suite
- R and BioConductor
- TopHat/CuffLinks Suite
- samTools

Galaxy



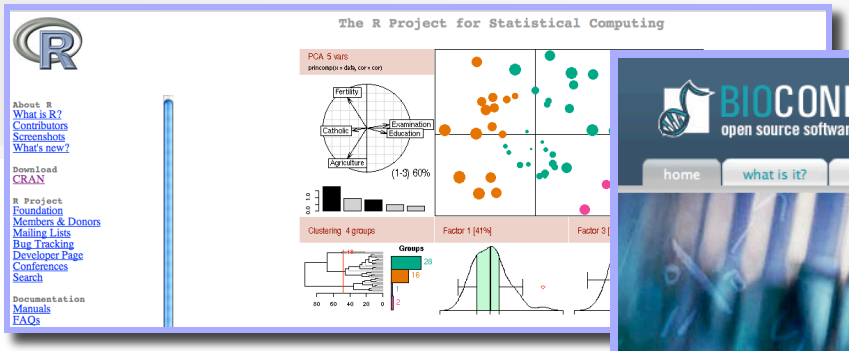
- ✓ Free internet-based software
- ✓ One-stop shop
- ✗ Limited storage
- ✗ Slow, highly utilized

Partek Genomics Suite



- ✓ Commercial software
- ✓ Run on Window and Linux machines
- ✓ Runs quickly
- ✓ Great visuals
- × Not free, share licensing
- × Black box, little flexibility

R and BioConductor

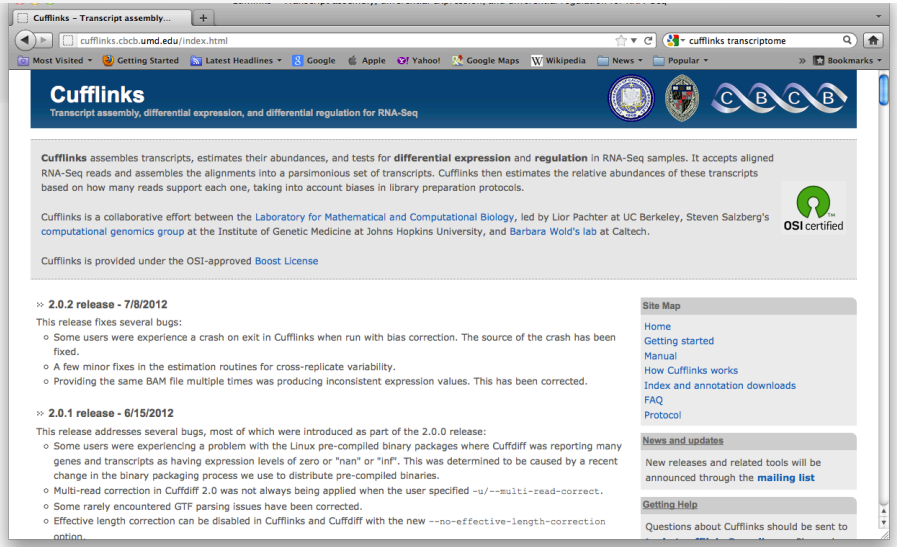


www.r-project.org

www.bioconductor.org

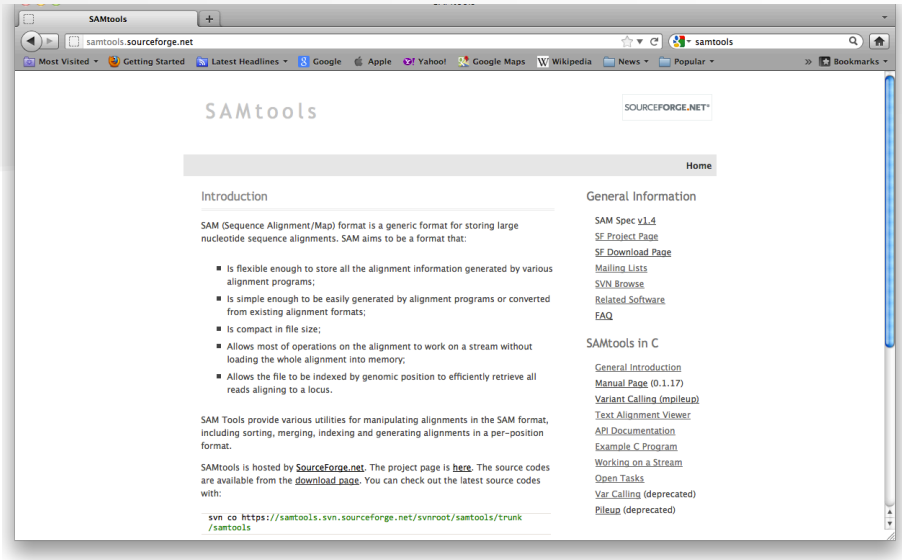
- ✓ Open-source software, free for the public
- ✓ Platform independent
- ✓ Unlimited statistical and programming flexibility
- ✗ Need command line usage experience
- ✗ Requires significant R programming experience
- ✗ Can't do everything

CuffLinks



- ✓ Free for the public
- ✓ Can go from raw reads to differential expression
- ✓ Datasets are formatted to go from task to the next
- ✗ Everything is done through the command line
- ✗ Limited flexibility for differential expression

SAMtools

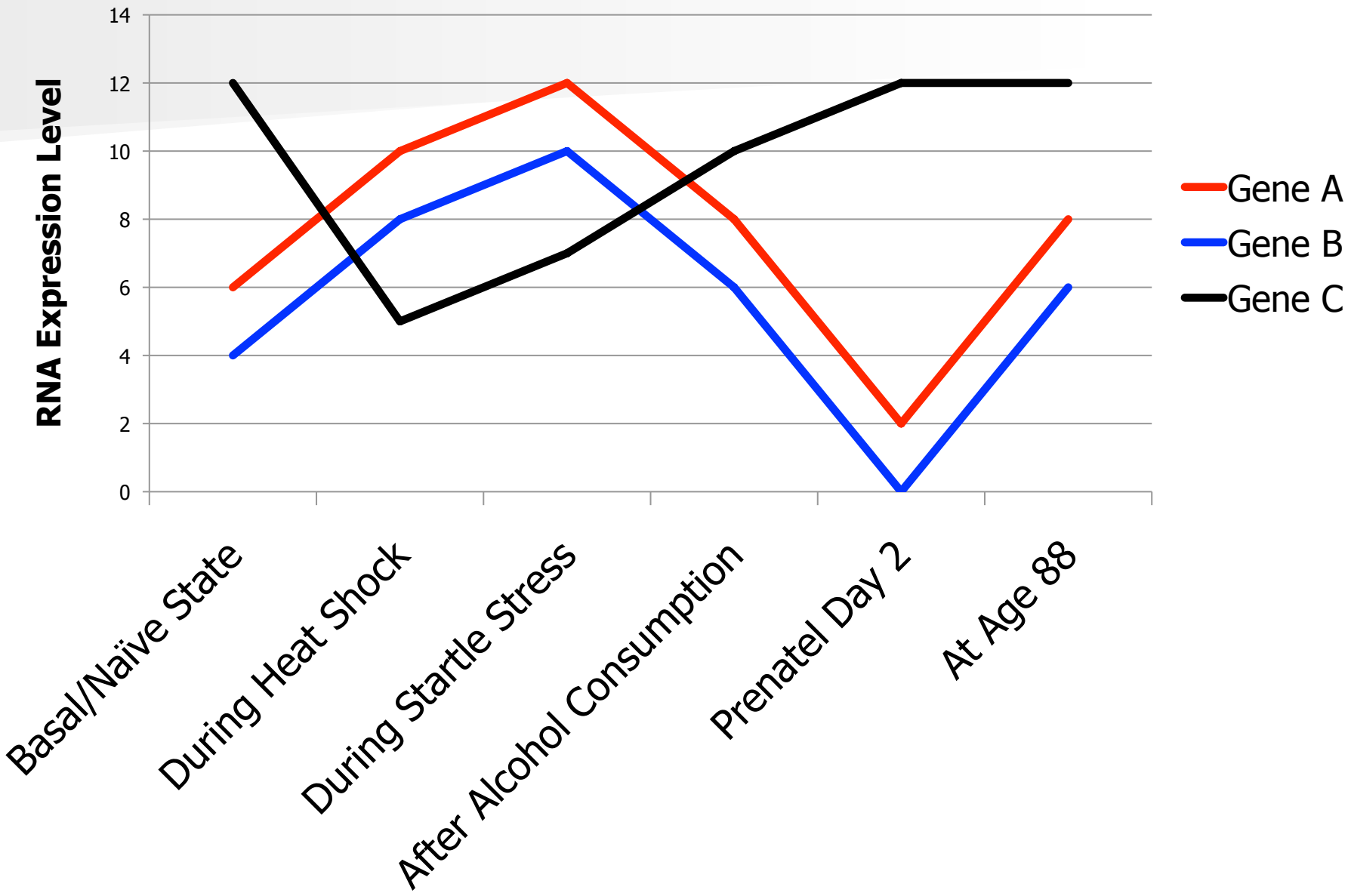


- ✓ Free for the public
- ✗ Mainly used for manipulating SAM files and genotype calls

Topics

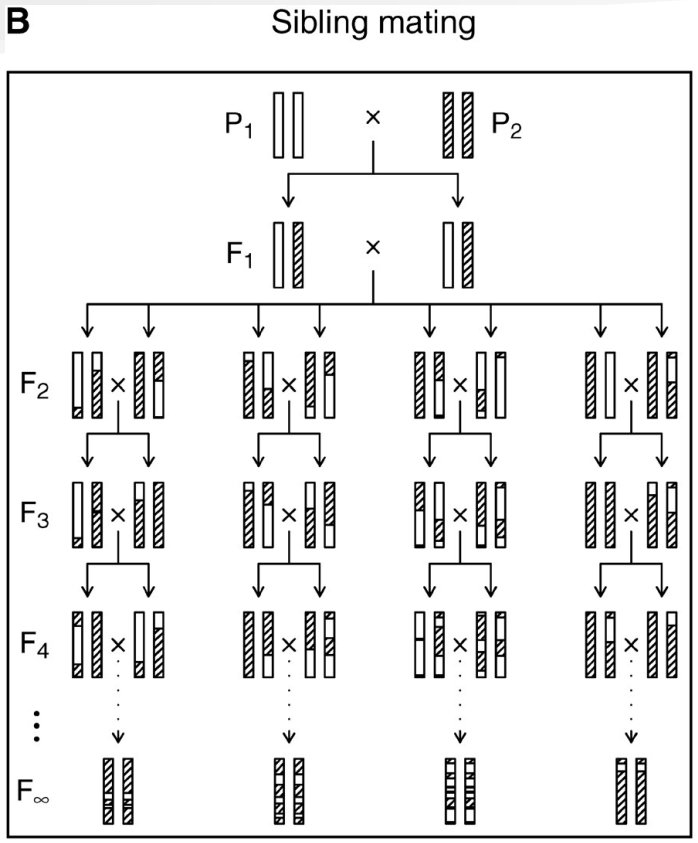
I. RNA Expression Technologies
II. Example - Genomic Drivers of Alcohol Consumption
III. Resources
IV. Glimpse into Co-Expression Networks

Characterize Relationships Among Transcripts



What if Lohan and Tebow a baby?

Recombinant Inbred Rodent Panels



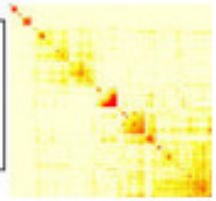
Broman, K. W. Genetics 2005;169:1133-1146

- Genetic identity is retained over generations
- Accumulative genetic and phenotype data across labs
- Ideal genetic controls for studying interventions/ environmental effects
- Biological replicates allow for the reduction of environmental variances

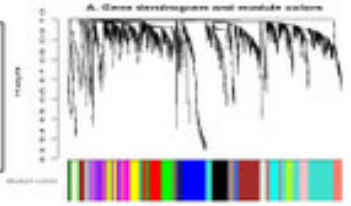
GENETICS

Weighted Gene Co-Expression Analysis

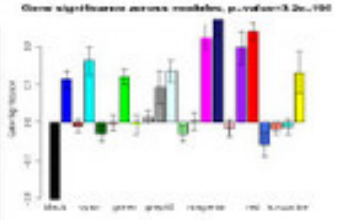
Construct a gene co-expression network
Rationale: make use of interaction patterns among genes
Tools: correlation as a measure of co-expression



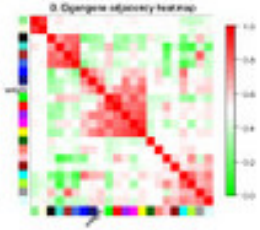
Identify modules
Rationale: module (pathway) based analysis
Tools: hierarchical clustering, Dynamic Tree Cut



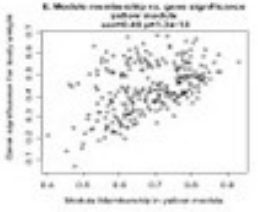
Relate modules to external information
Array Information: clinical data, SNPs, proteomics
Gene Information: ontology, functional enrichement
Rationale: find biologically interesting modules



Study module relationships
Rationale: biological data reduction, systems-level view
Tools: Eigengene Networks



Find the key drivers in *interesting* modules
Rationale: experimental validation, biomarkers
Tools: intramodular connectivity, causality testing



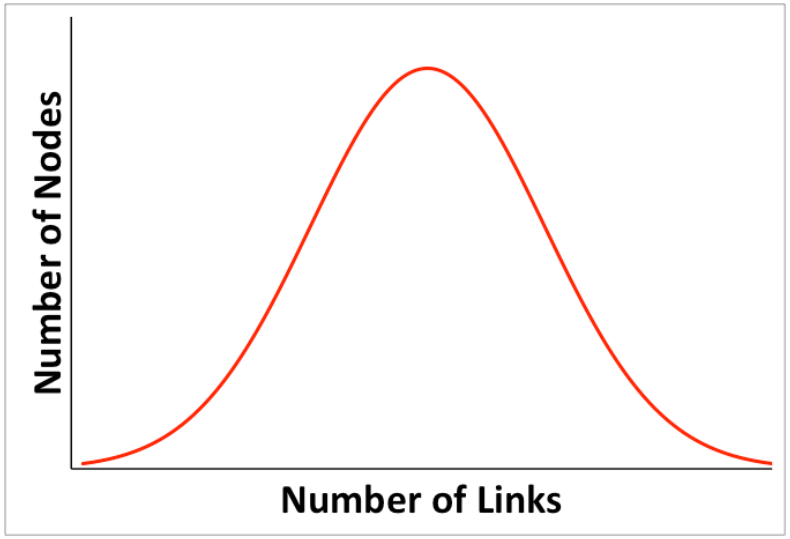
Langfelder and Horvath BMC Bioinformatics 9(1):559 (2008)

Weighted Gene Co-Expression Network Analysis
Why Not Just Use Correlation?

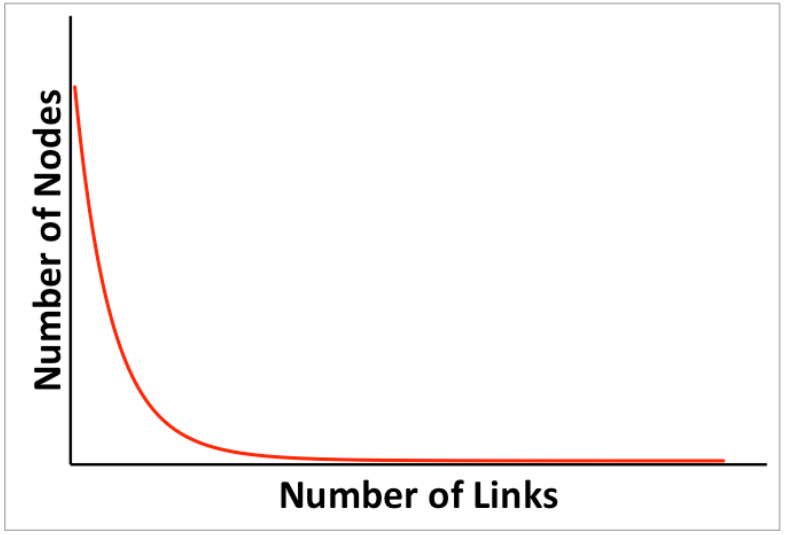
1. **How are we measuring co-expression?**
   - **Scale-Free Network**
     - Network has few highly connected genes rather then each gene have similar connectivity
     - Biologically motivated, fewer highly connected genes means that a system is more robust to failure of any one gene
2. **How do we get a robust measure of connectivity for identifying modules?**
   - **Topological Overlap Measure**
     - Includes a measure of how many "friends" two genes have in common
     - Protects against spurious correlations among genes

Scale-Free Networks

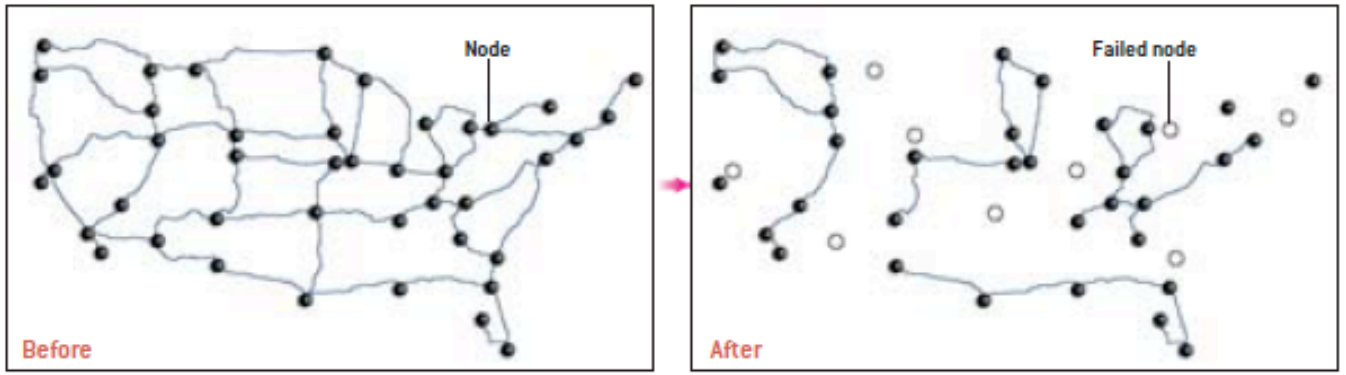
Random Network



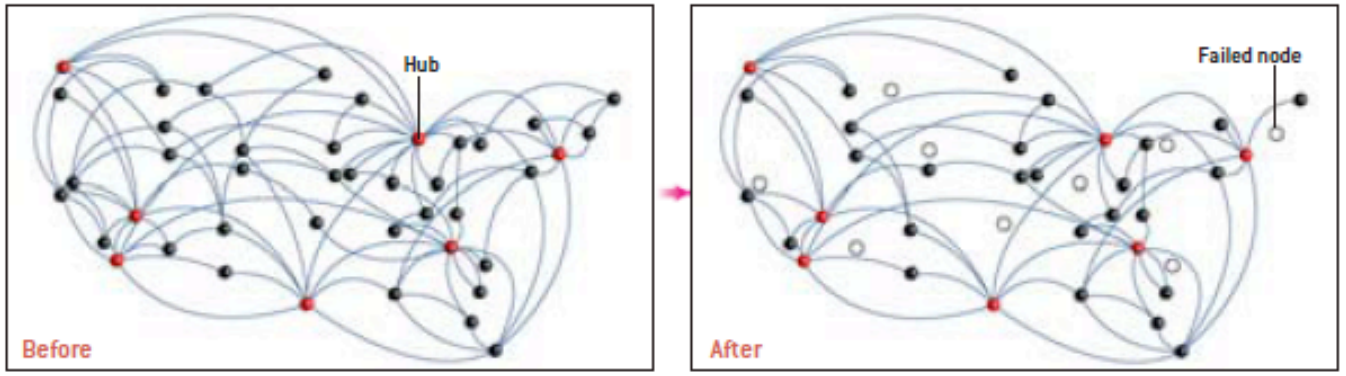
Scale-Free Network



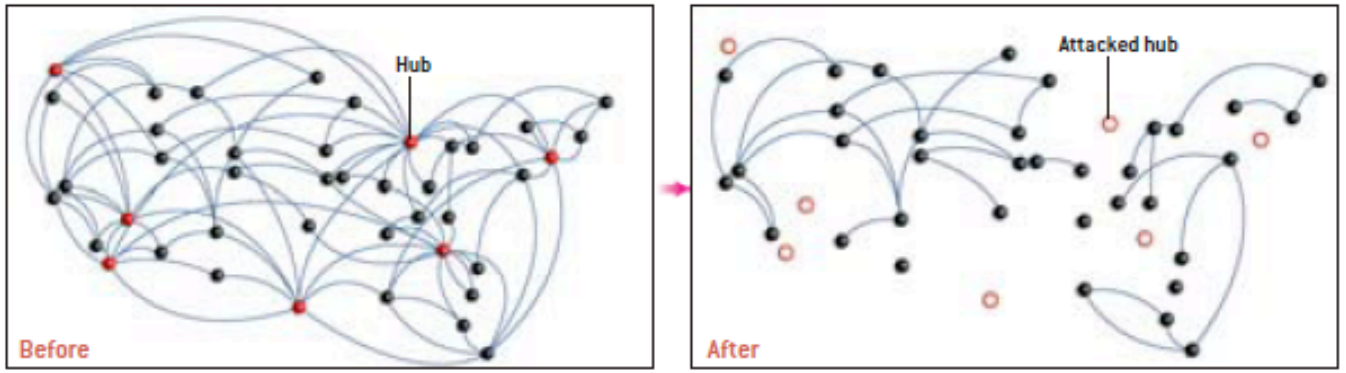
Random Network, Accidental Node Failure



Scale-Free Network, Accidental Node Failure

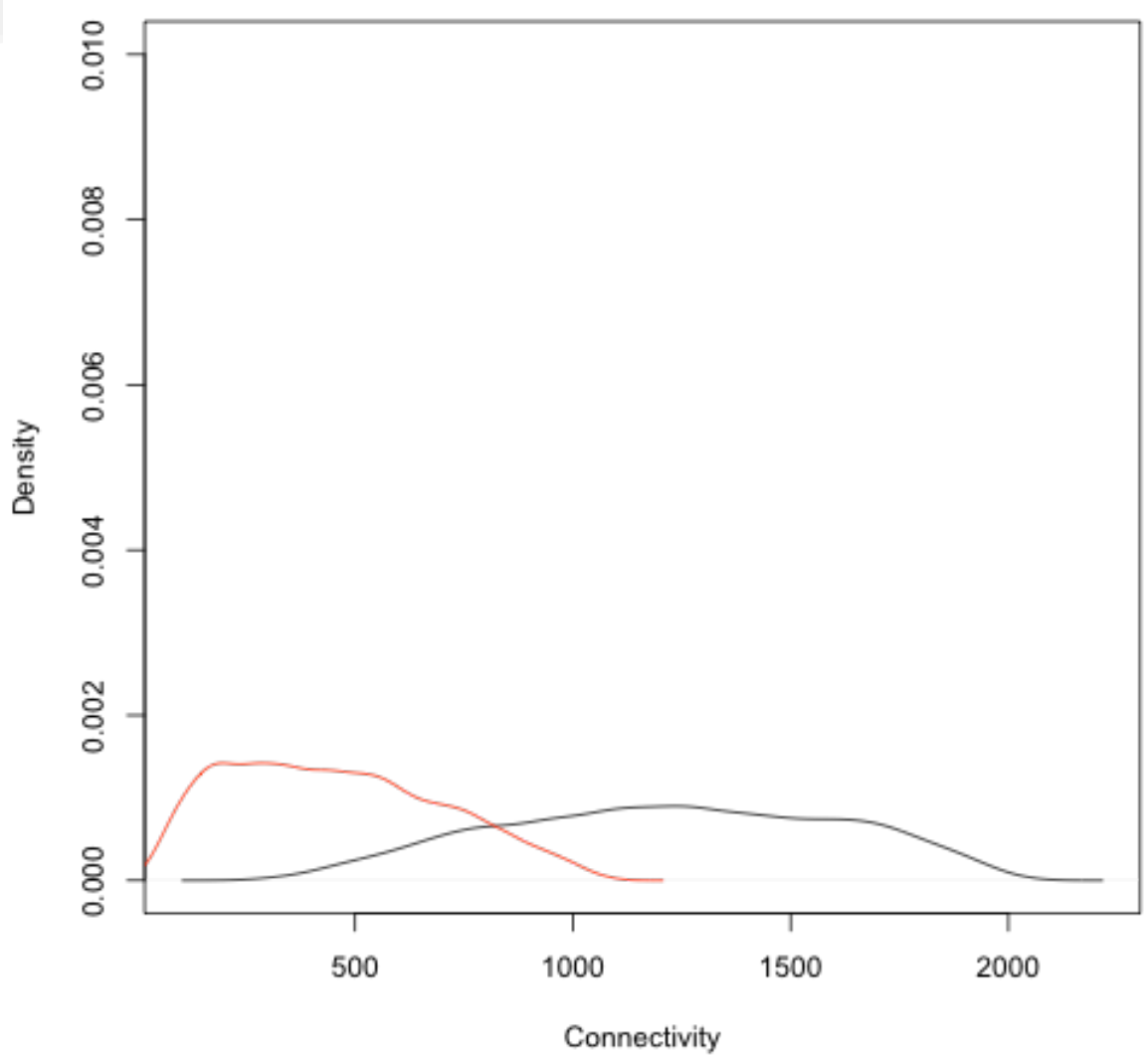


Scale-Free Network, Attack on Hubs



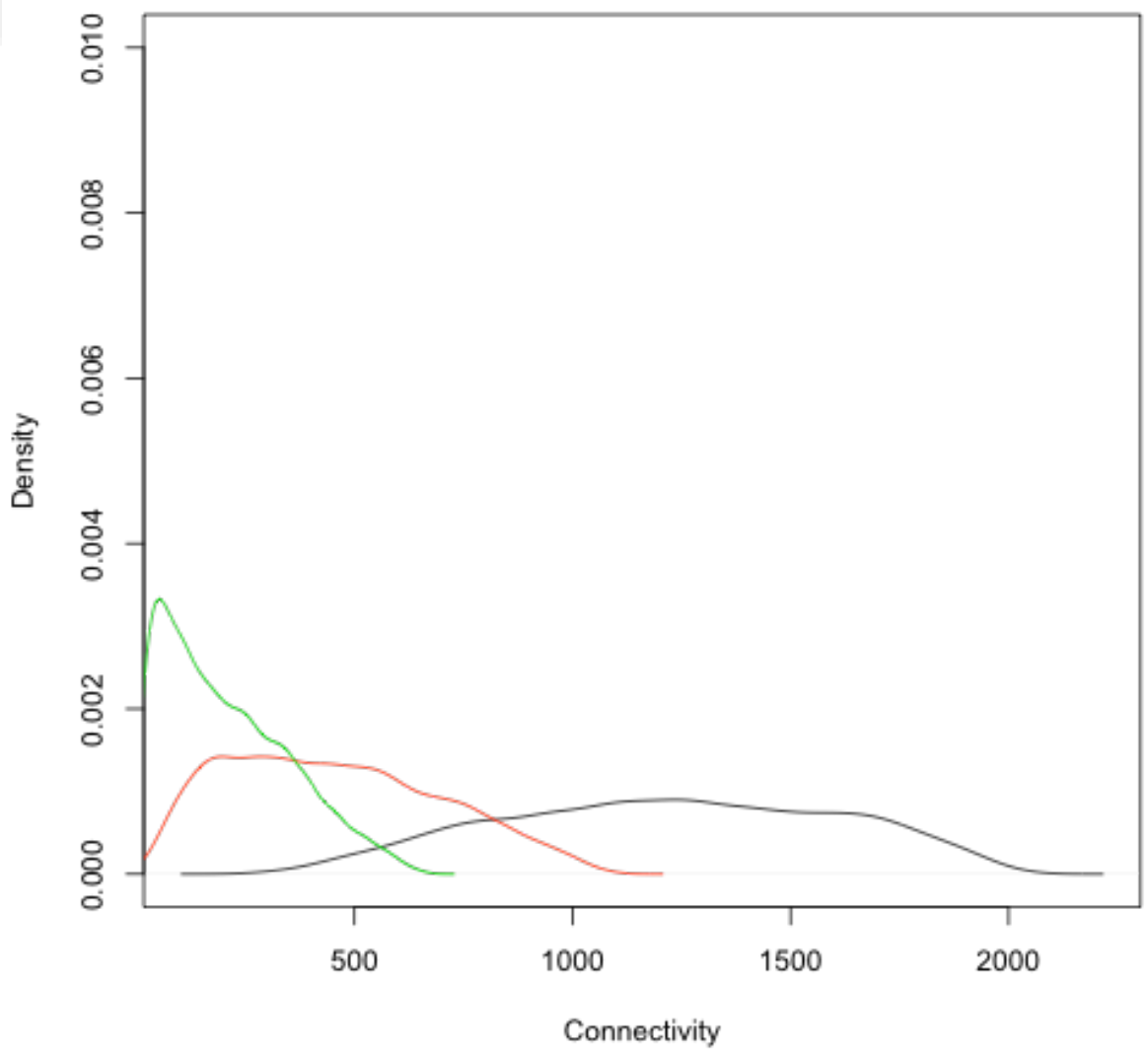
Scale-Free Network

$$\text{Connectivity}_i = \sum_{j \neq i} |\rho^2_{ij}|$$



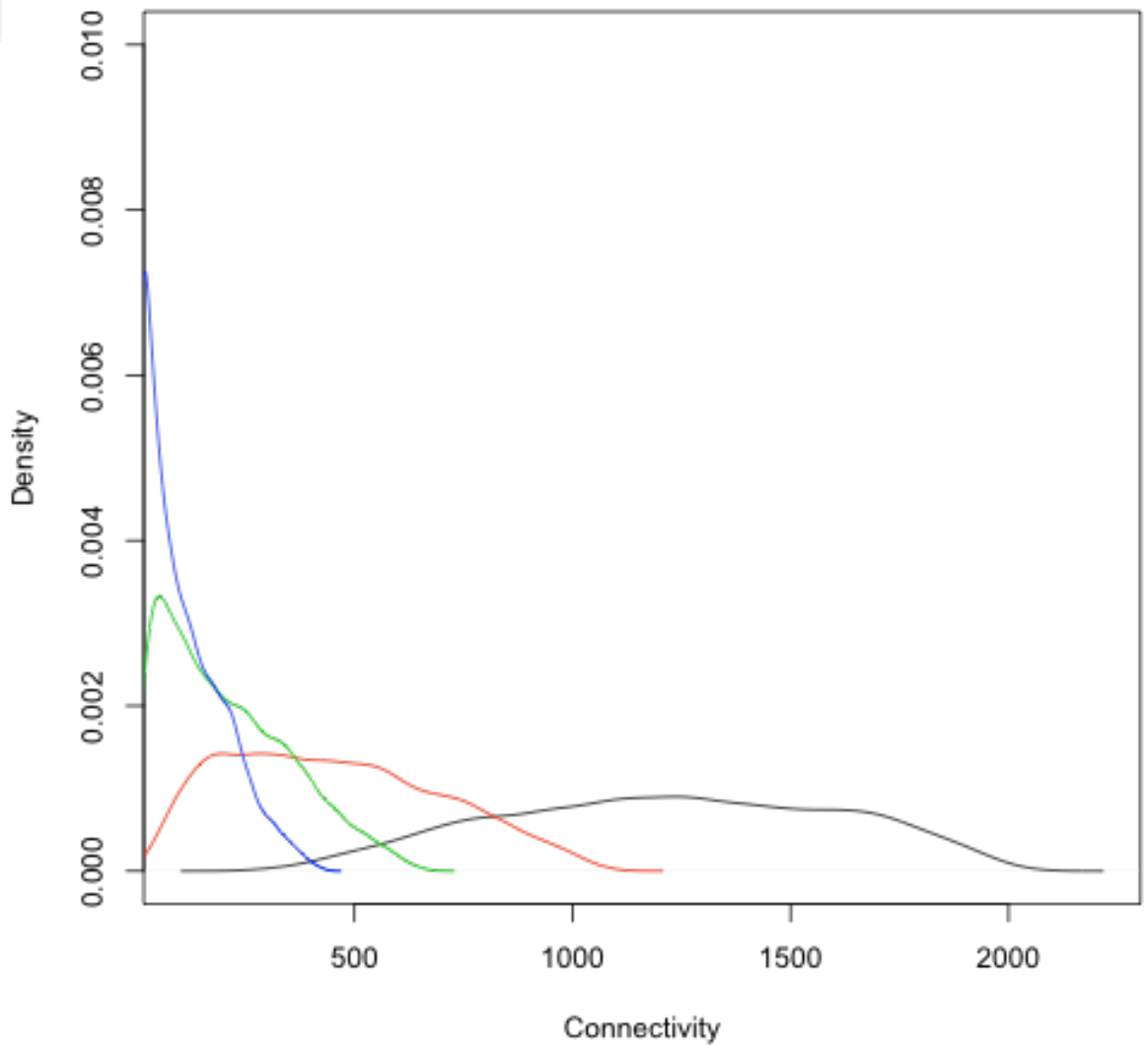
Scale-Free Network

$$\text{Connectivity}_i = \sum_{j \neq i} |\rho^3_{ij}|$$



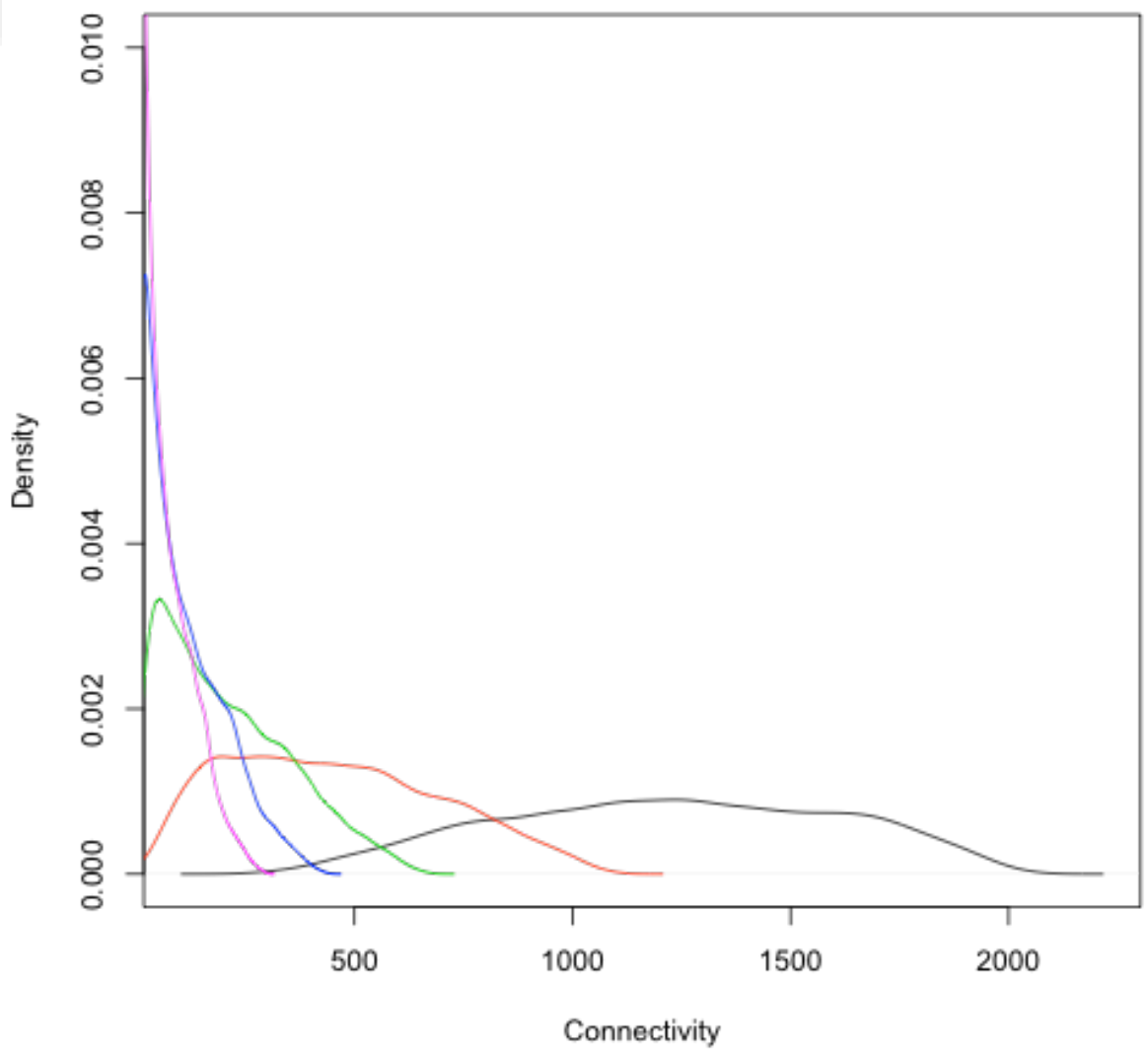
Scale-Free Network

$$\text{Connectivity}_i = \sum_{j \neq i} |\rho^4_{ij}|$$



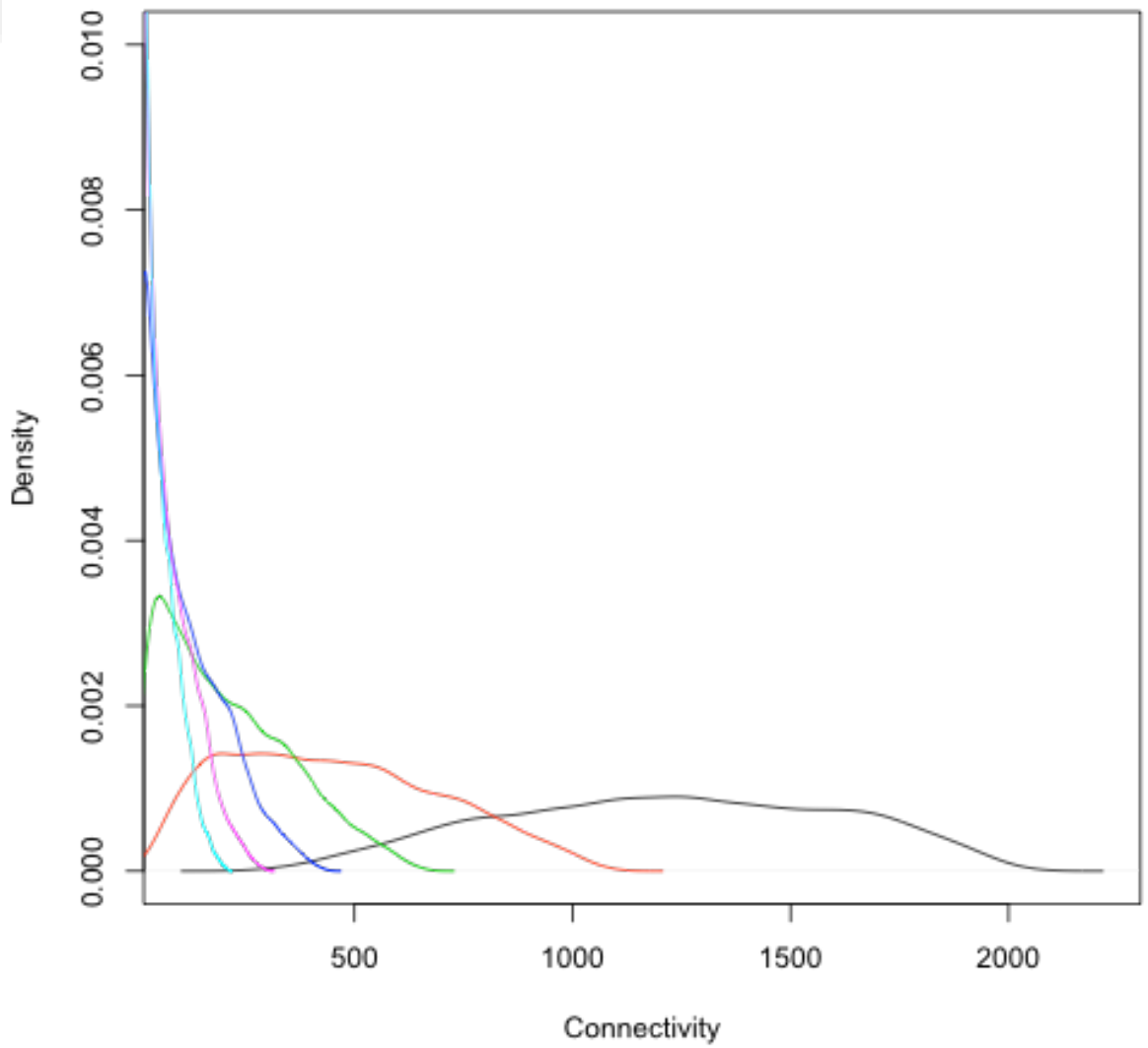
Scale-Free Network

$$\text{Connectivity}_i = \sum_{j \neq i} |\rho^5_{ij}|$$



Scale-Free Network

$$\text{Connectivity}_i = \sum_{j \neq i} |\rho^6_{ij}|$$



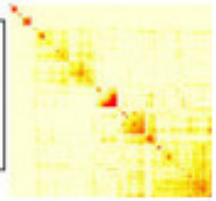
Topological overlap measure

$$TOM_{ij} = \frac{a_{ij} + \sum_u a_{iu} a_{uj}}{\min\left\{\sum_u a_{iu}, \sum_u a_{uj}\right\} + 1 - a_{ij}}$$

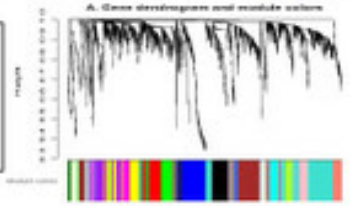
where $a_{ij} = \left|\rho_{ij}\right|^{\beta}$ is the connectivity of transcript *i* and transcript *j*

Weighted Gene Co-Expression Analysis

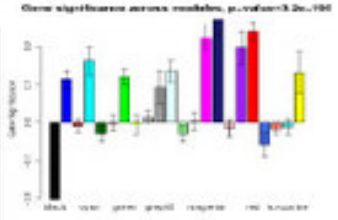
Construct a gene co-expression network
Rationale: make use of interaction patterns among genes
Tools: correlation as a measure of co-expression



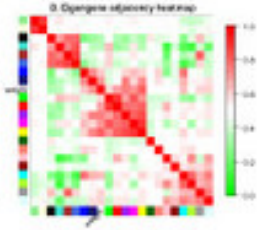
Identify modules
Rationale: module (pathway) based analysis
Tools: hierarchical clustering, Dynamic Tree Cut



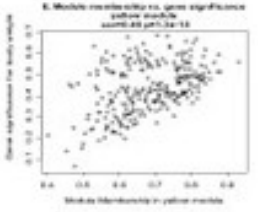
Relate modules to external information
Array Information: clinical data, SNPs, proteomics
Gene Information: ontology, functional enrichement
Rationale: find biologically interesting modules



Study module relationships
Rationale: biological data reduction, systems-level view
Tools: Eigengene Networks

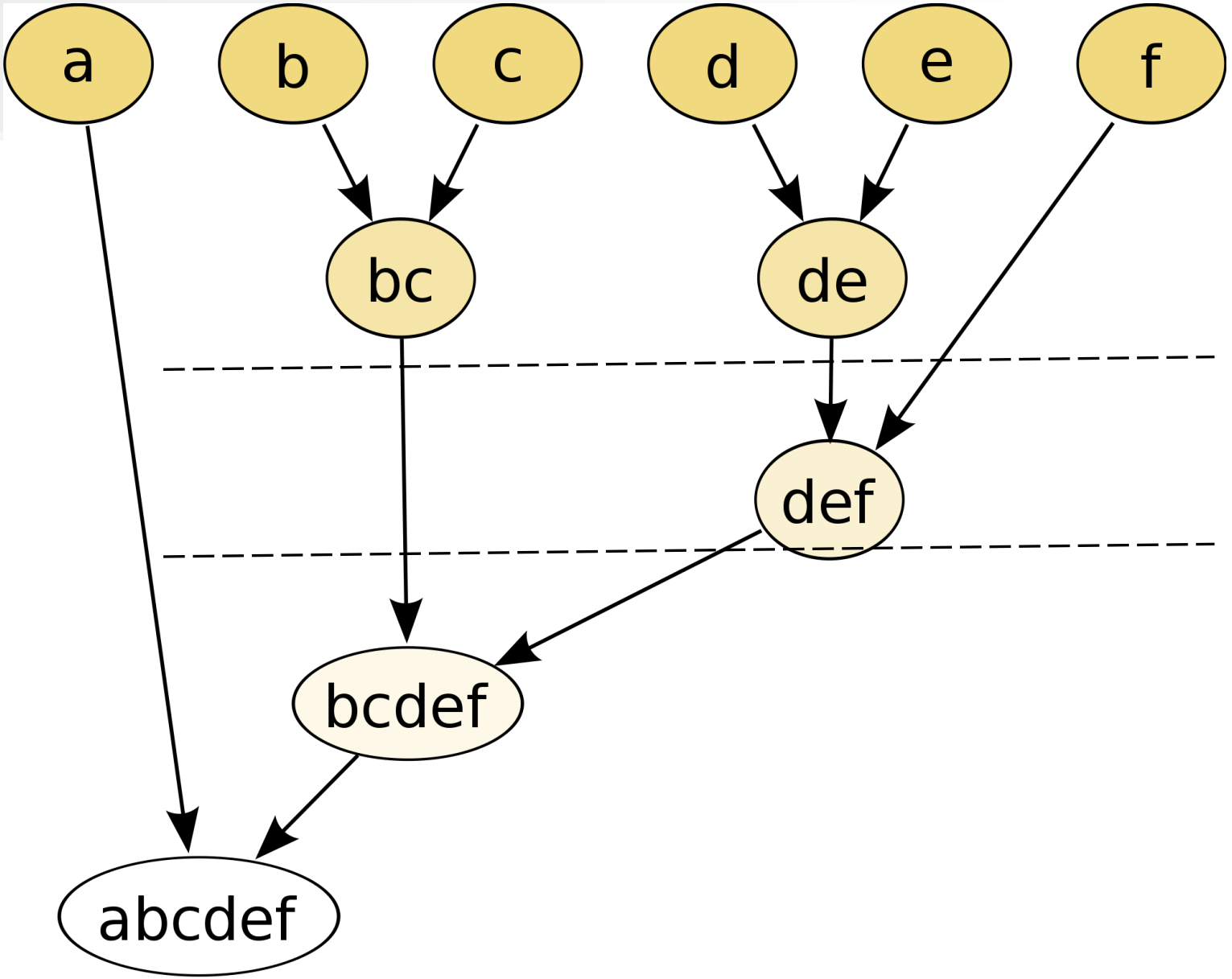


Find the key drivers in *interesting* modules
Rationale: experimental validation, biomarkers
Tools: intramodular connectivity, causality testing



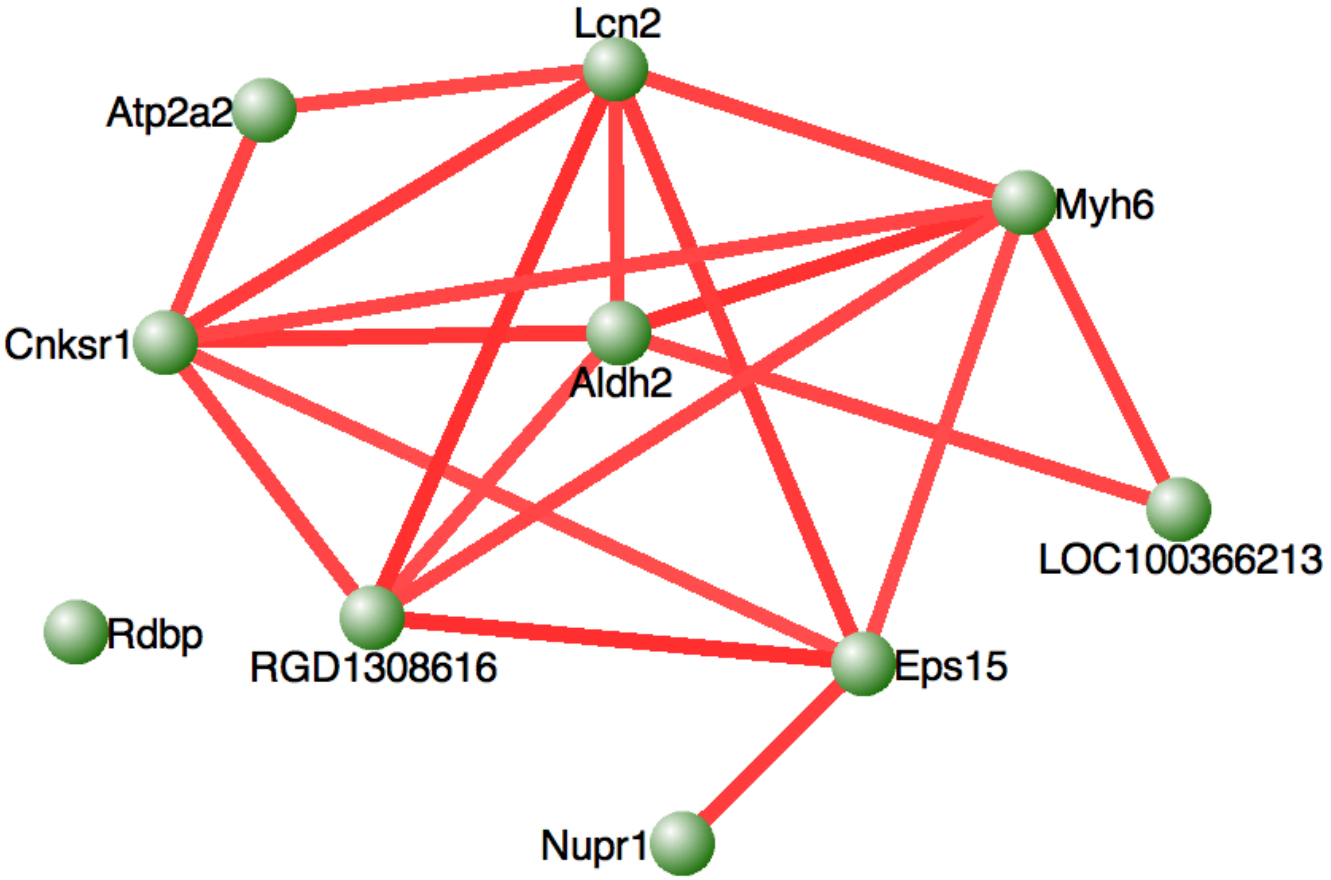
Langfelder and Horvath BMC Bioinformatics 9(1):559 (2008)

Hierarchical Clustering



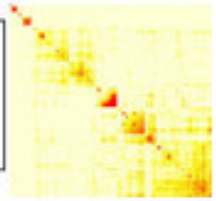
Example - Aldh2 Module

Enriched with genes from KEGG pathways:
Cardiac muscle contraction, proteasome, tight junction

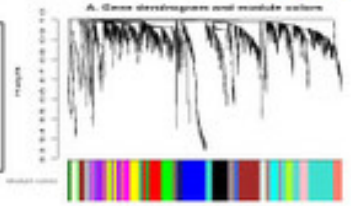


Weighted Gene Co-Expression Analysis

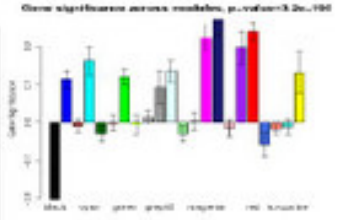
Construct a gene co-expression network
Rationale: make use of interaction patterns among genes
Tools: correlation as a measure of co-expression



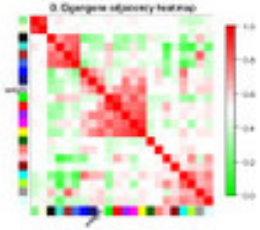
Identify modules
Rationale: module (pathway) based analysis
Tools: hierarchical clustering, Dynamic Tree Cut



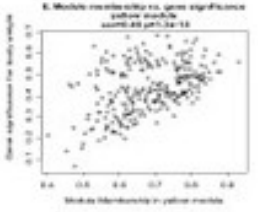
Relate modules to external information
Array Information: clinical data, SNPs, proteomics
Gene Information: ontology, functional enrichement
Rationale: find biologically interesting modules



Study module relationships
Rationale: biological data reduction, systems-level view
Tools: Eigengene Networks



Find the key drivers in *interesting* modules
Rationale: experimental validation, biomarkers
Tools: intramodular connectivity, causality testing



Langfelder and Horvath BMC Bioinformatics 9(1):559 (2008)