

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

Research article

Open Access

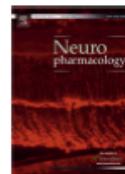
Genetical genomic determinants of alcohol consumption in rats and humans

Boris Tabakoff^{*1}, Laura Saba¹, Morton Printz², Pam Flodman³,
Colin Hodgkinson⁴, David Goldman⁴, George Koob⁵,
Heather N Richardson^{5,16}, Katerina Kechris⁶, Richard J Bell⁷
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,*}

^aDepartment of Pharmacology, University of Colorado Denver School of Medicine, PO Box 6511, Mail Stop 8303, Aurora, CO 80045, USA

^bColorado 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’ analysis of morphine-induced analgesia: identifying candidate genes

Paula L. Hoffman¹, Beth Bennett¹, Laura M. Saba¹, Sanjiv V. Bhave¹, Phyllis Cheryl K. Hornbaker¹, Katerina J. Kechris¹, Robert W. Williams² & Boris

University of Colorado Denver School of Medicine, Department of Pharmacology, Aurora CO 80045, USA¹, University of Tennessee Department of Anatomy and Neurobiology, Memphis, TN 38163, USA²

BMC Genetics

Open Access

The PhenoGen Informatics website: tools for analyses of complex traits

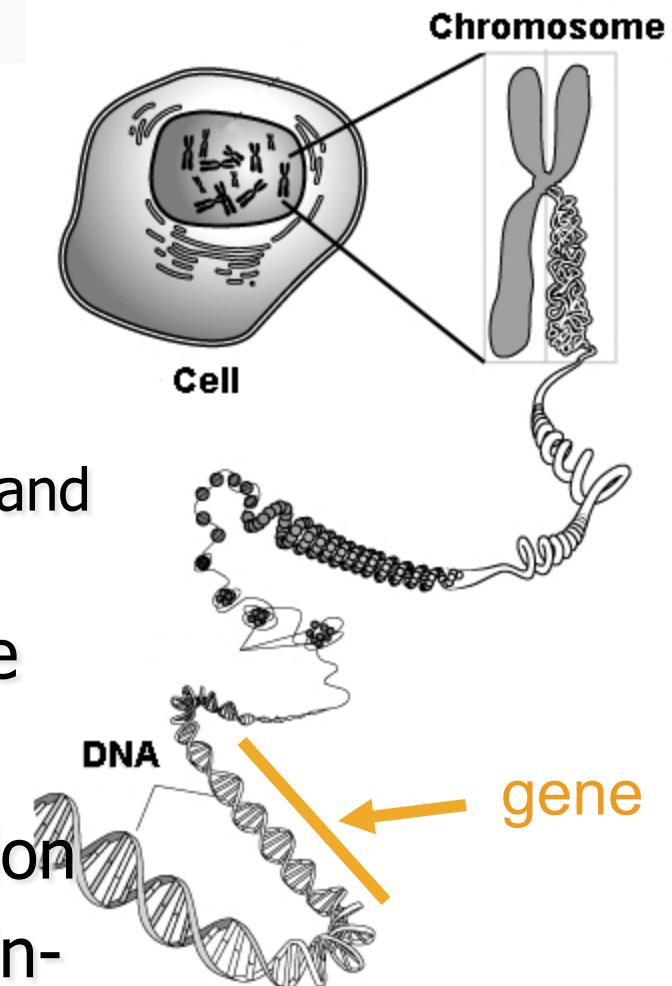
Sanjiv V Bhave¹, Cheryl Hornbaker¹, Tzu L Phang¹, Laura Saba¹,
Razvan Lapadat¹, Katherina Kechris^{1,2}, Jeanette Gaydos¹,
Daniel McGoldrick¹, Andrew Dolbey¹, Sonia Leach¹, Brian Soriano¹,
Allison Ellington¹, Eric Ellington¹, Kendra Jones¹, Jonathan Mangion³,
John K Belknap⁴, Robert W Williams⁵, Lawrence E Hunter¹,
Paula L Hoffman¹ 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)



Definition of Gene Expression

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

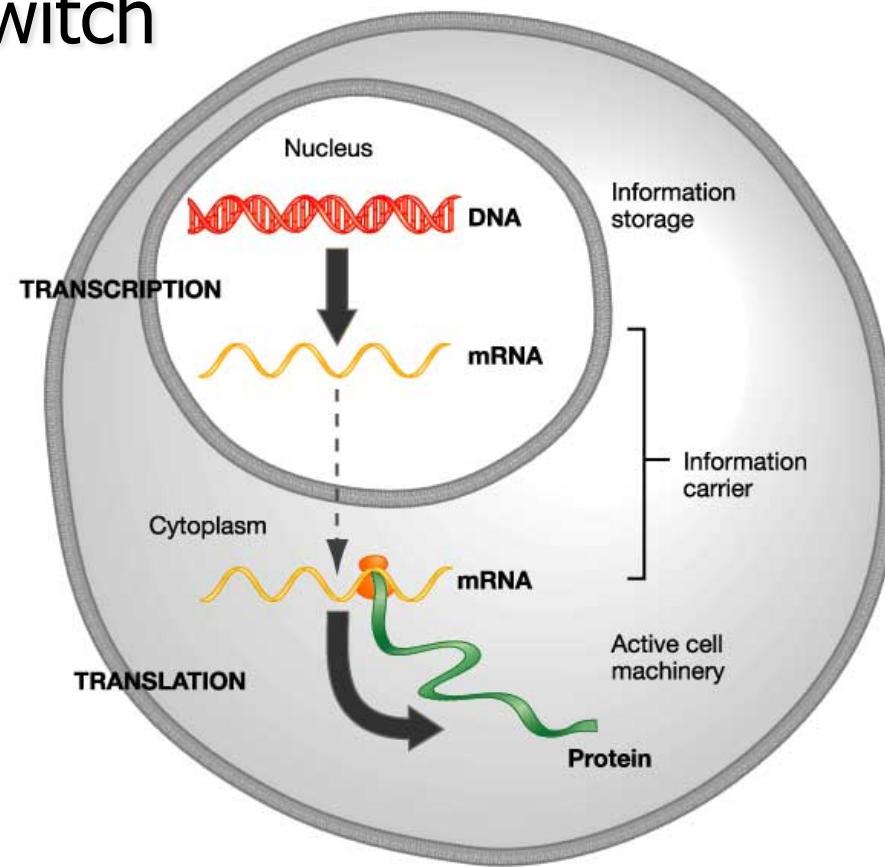
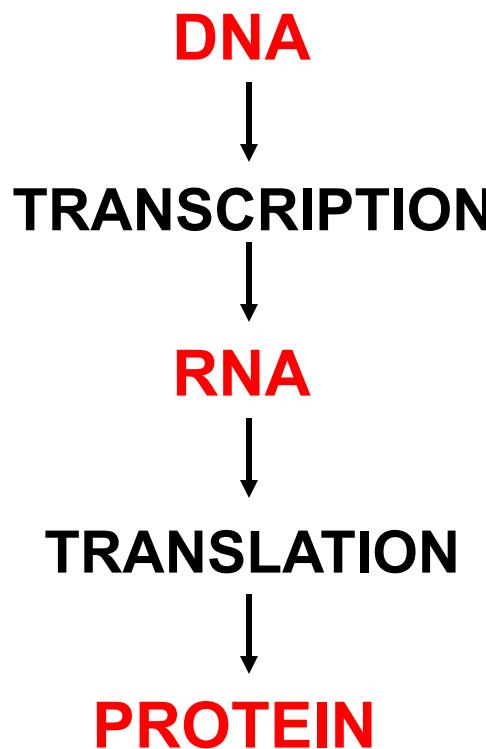
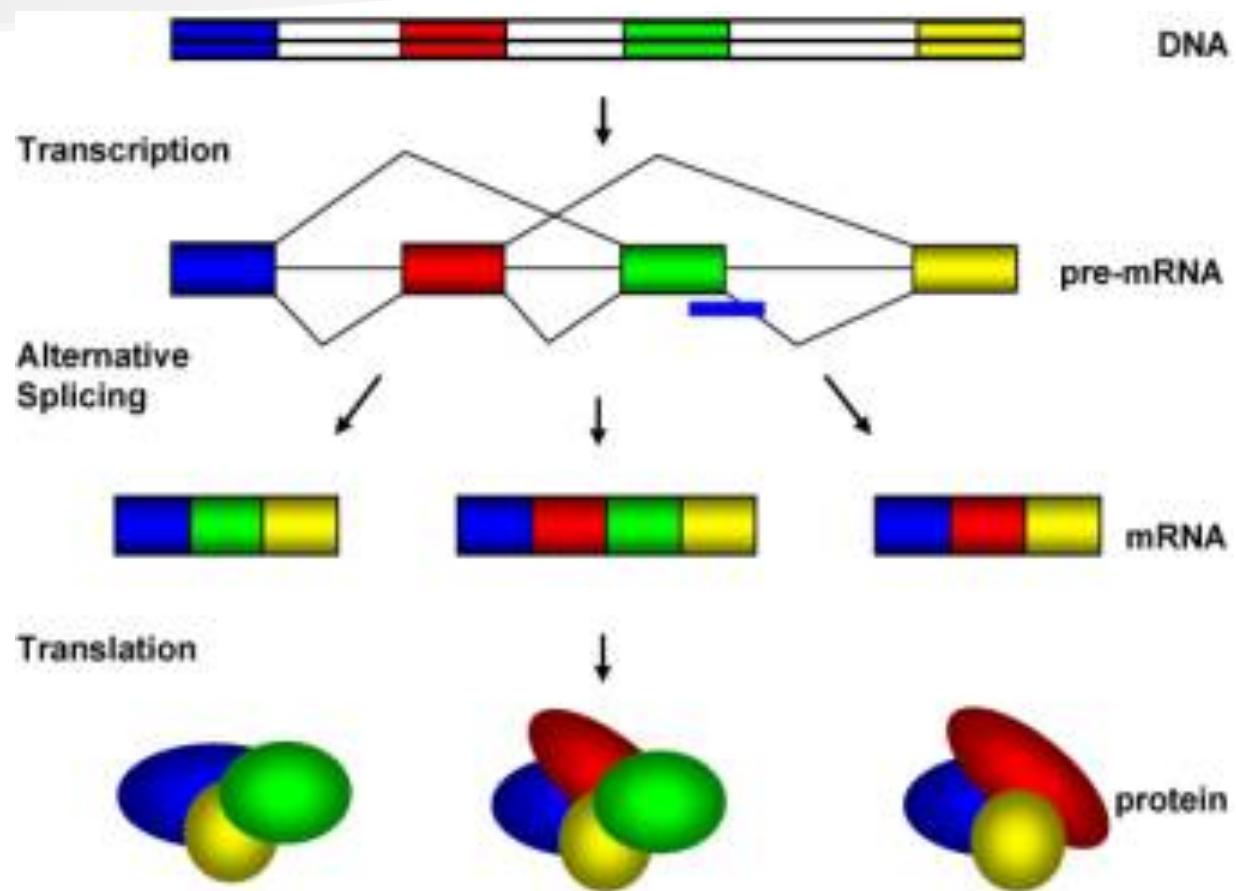


Image from <http://campus.queens.edu/>

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

Affymetrix Technology

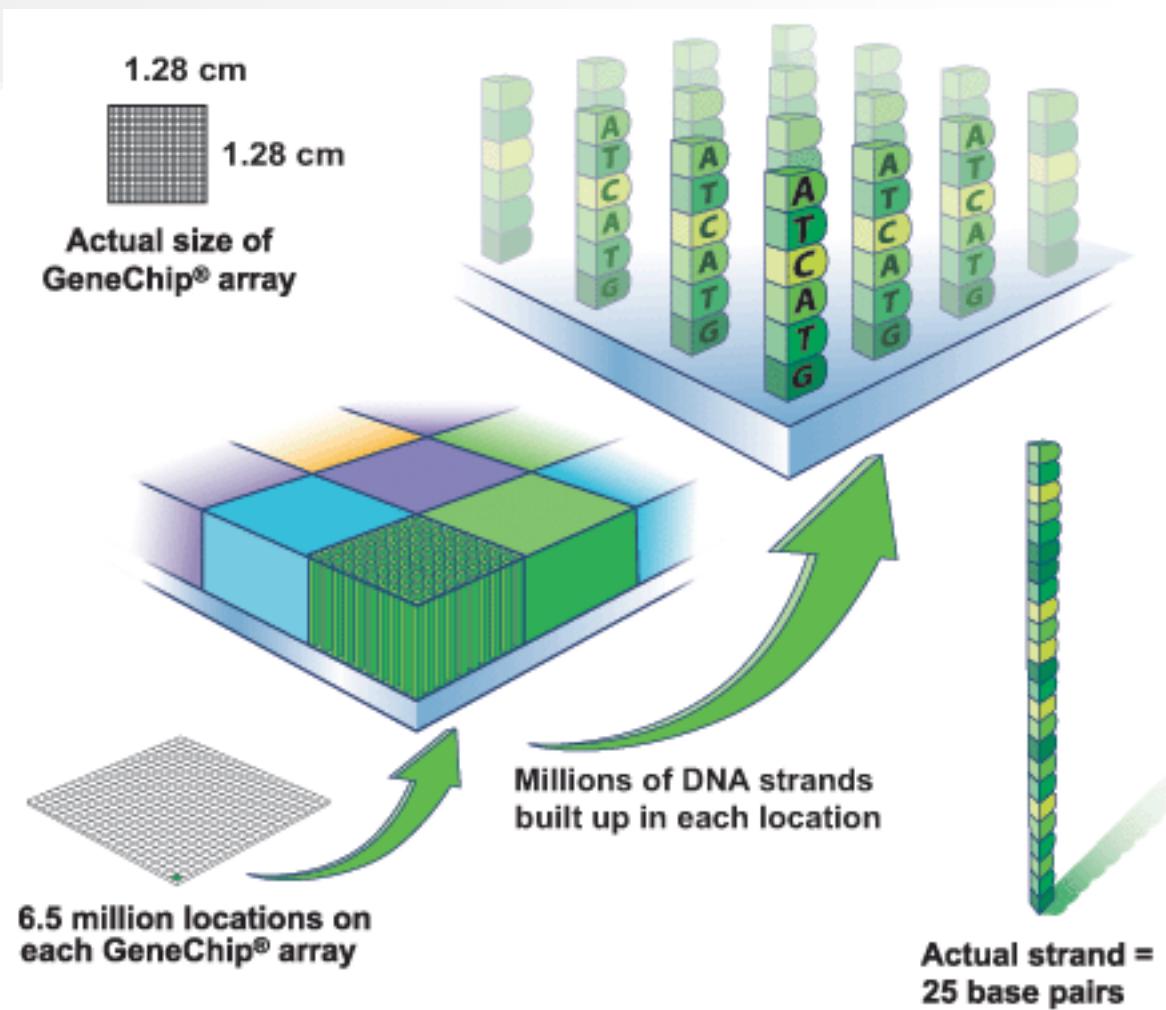


Image courtesy of Affymetrix.

Affymetrix Technology

RNA fragments with fluorescent tags from sample to be tested

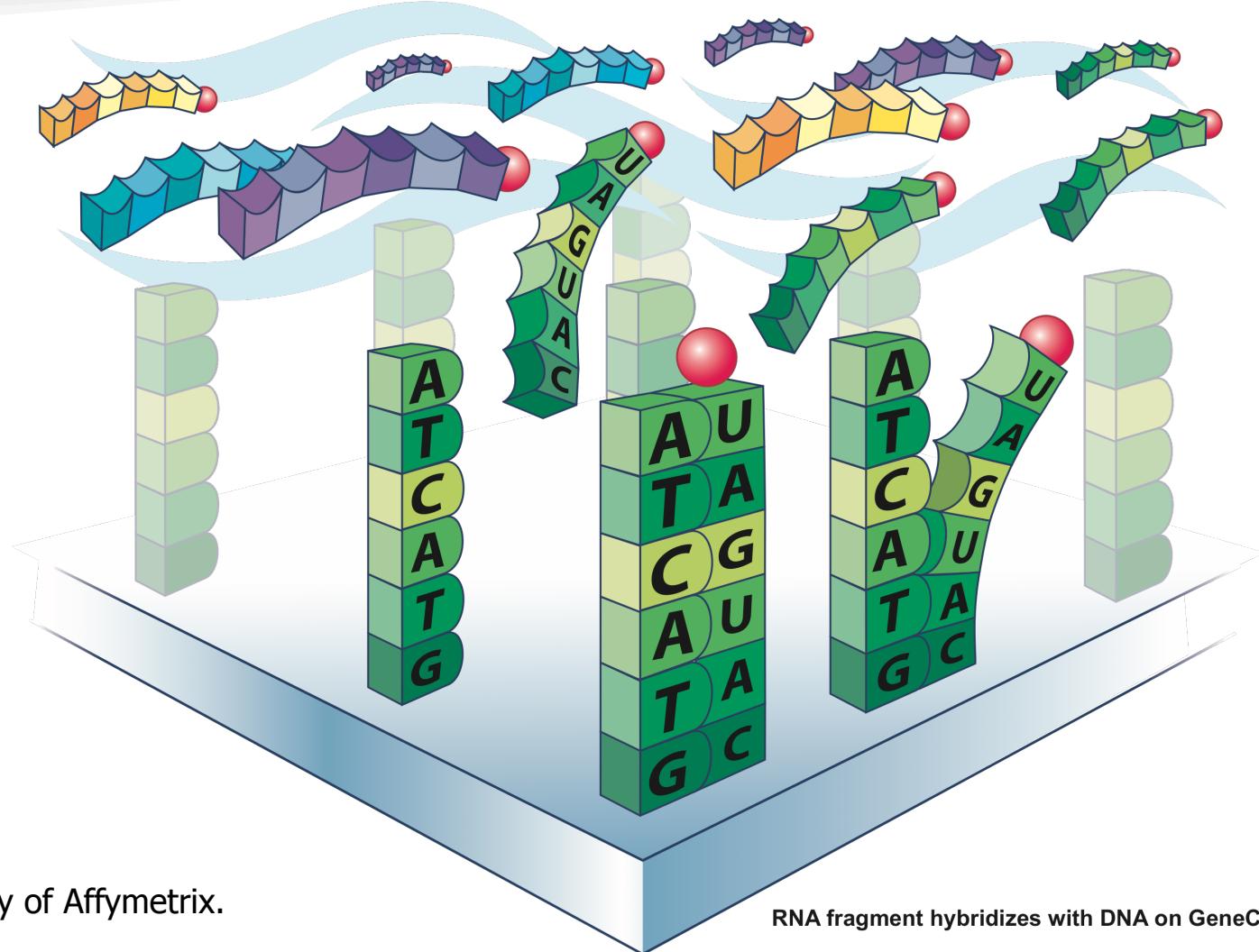


Image courtesy of Affymetrix.

RNA fragment hybridizes with DNA on GeneChip® array

Affymetrix Technology

Shining a laser light at GeneChip® array causes tagged DNA fragments that hybridized to glow

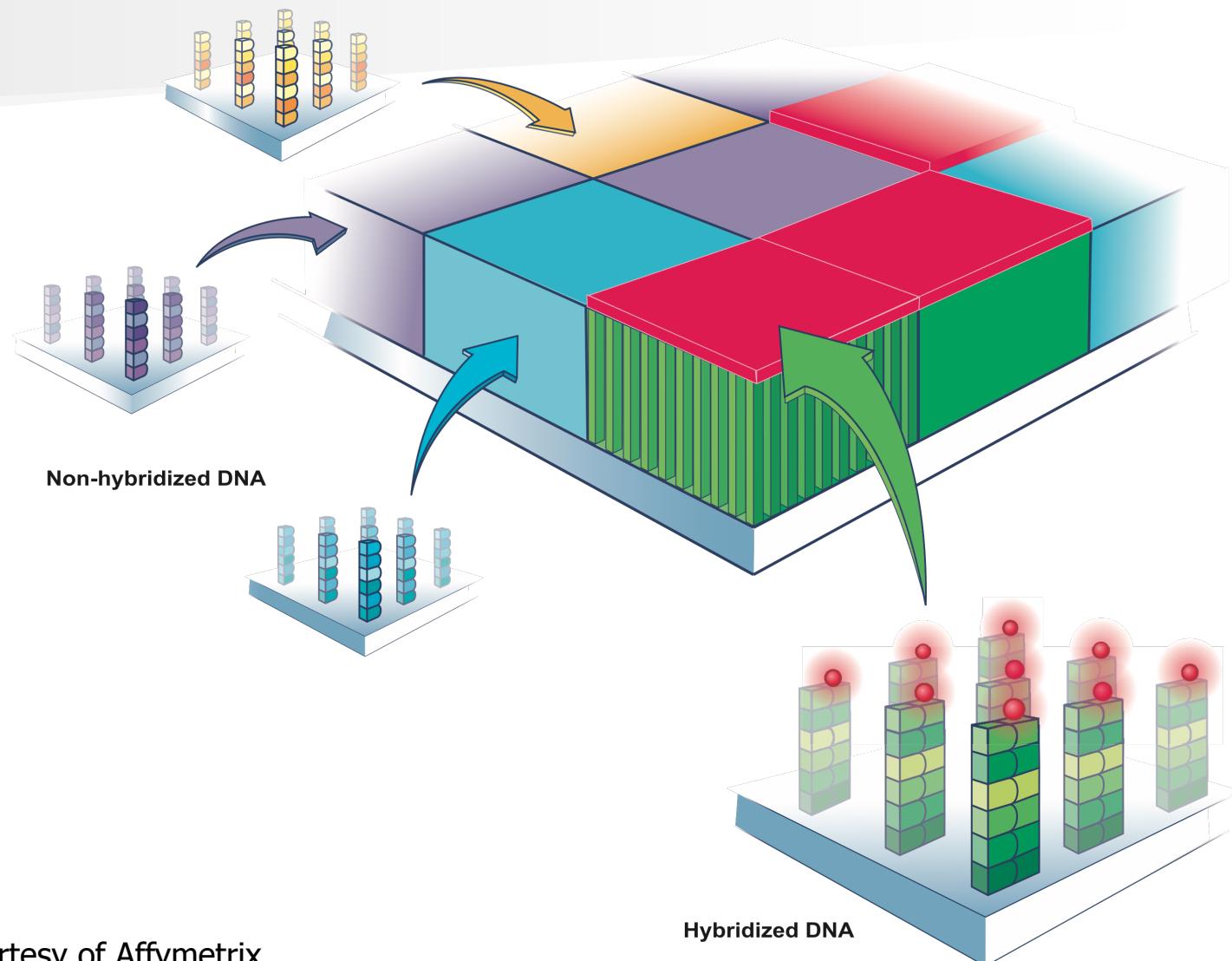
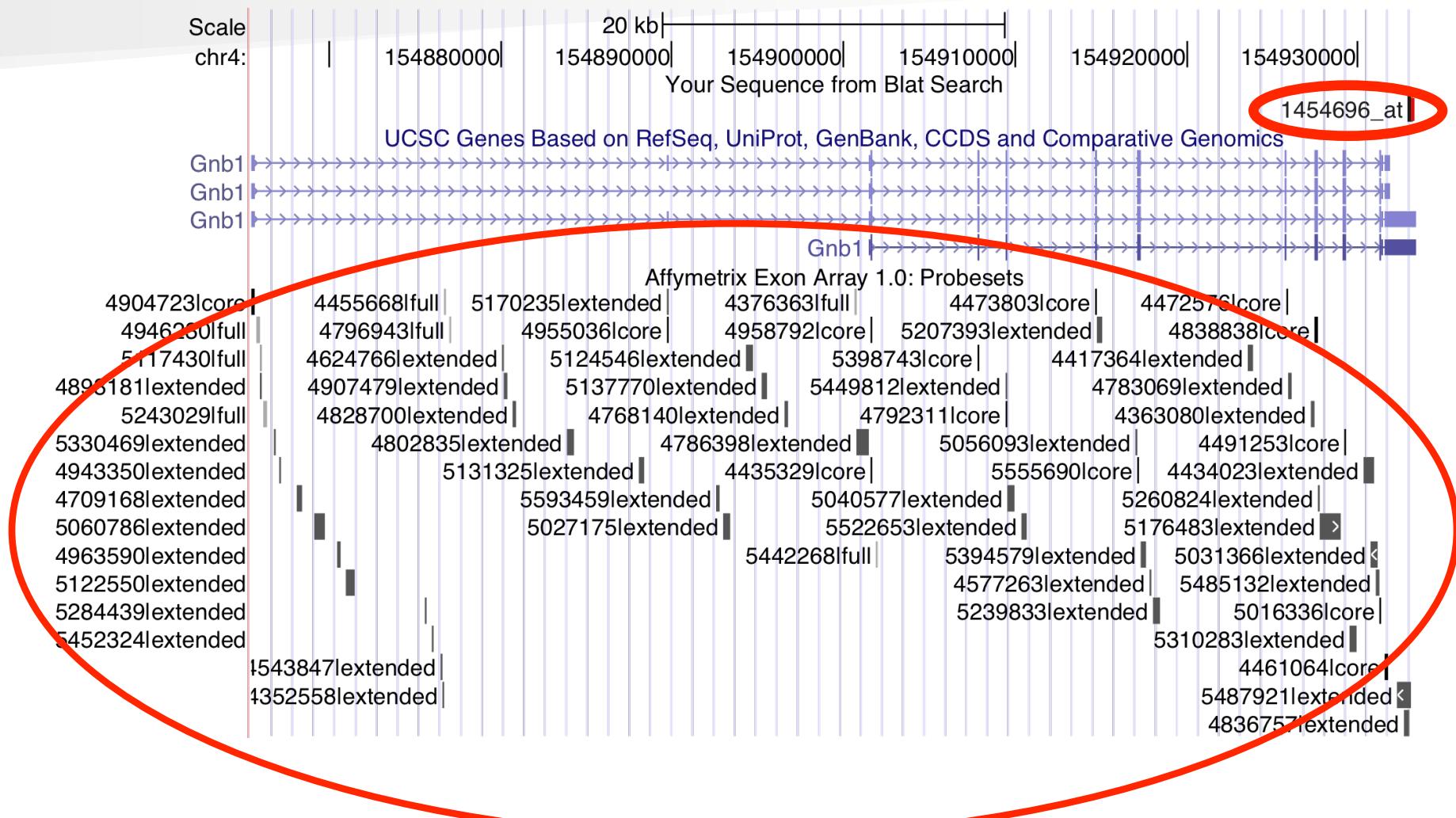
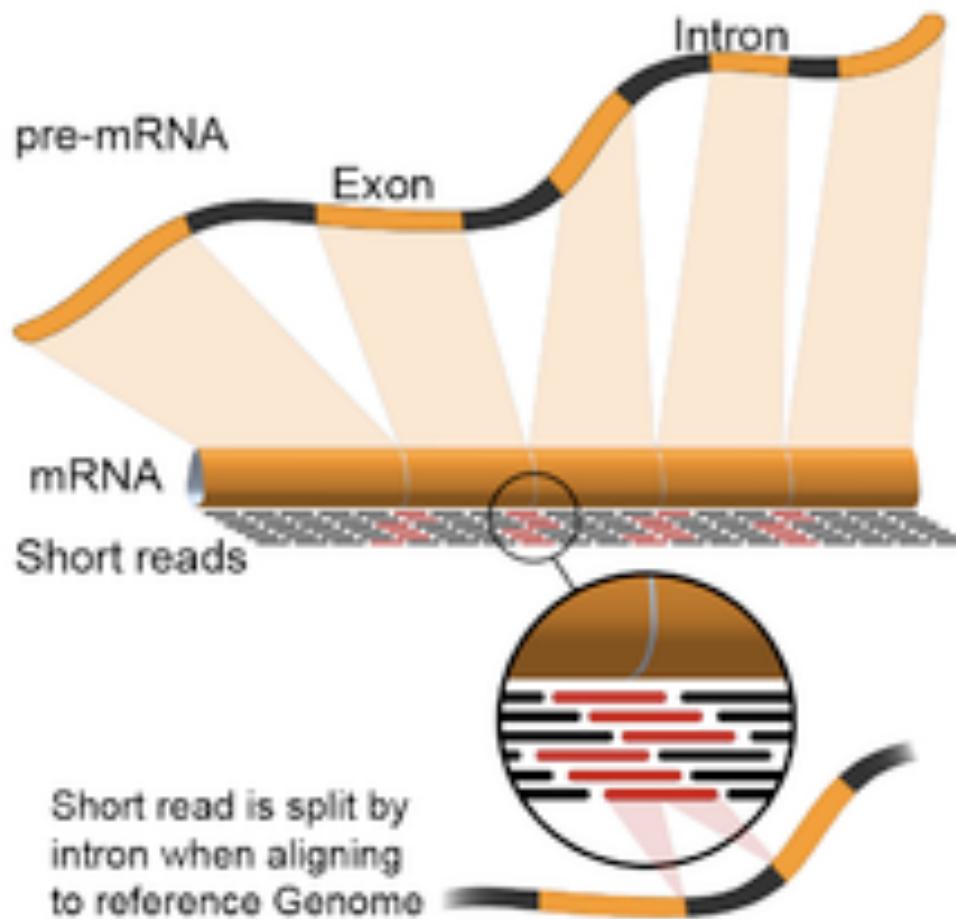


Image courtesy of Affymetrix.

Expansion of Microarray Technology

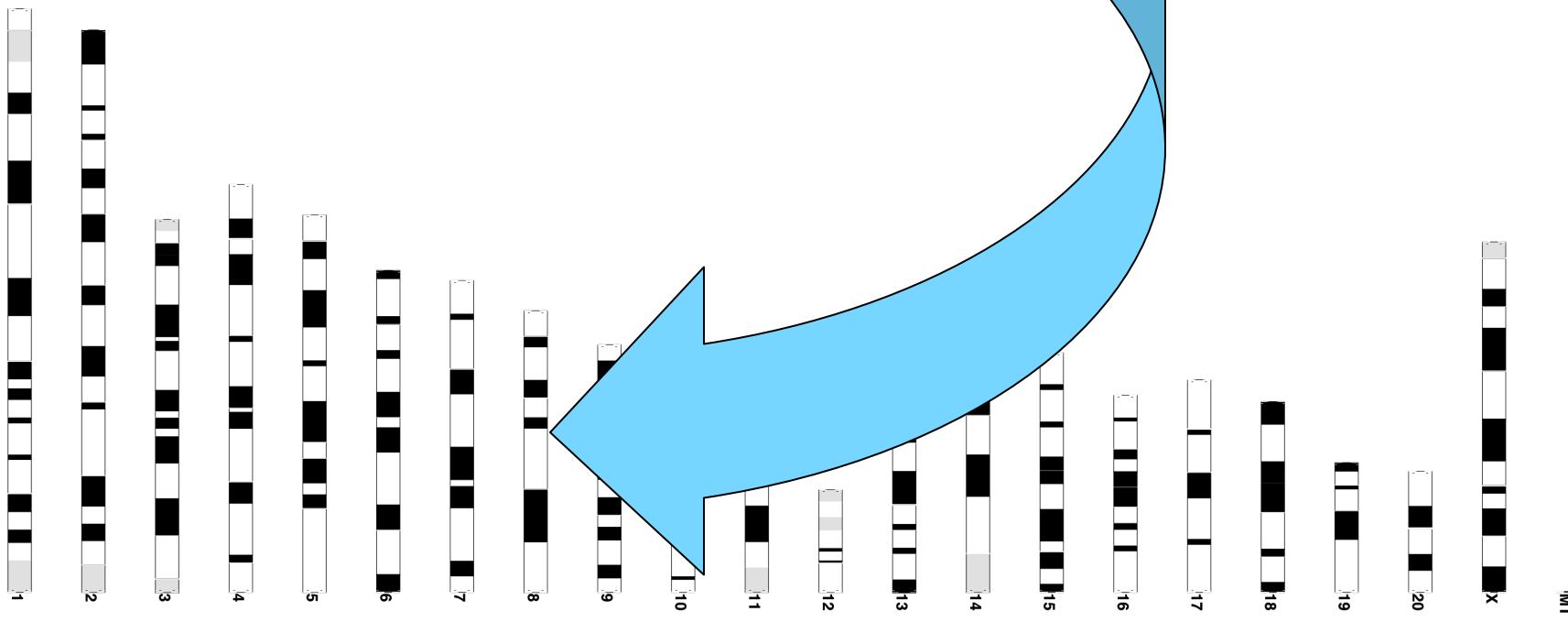


RNA-Seq

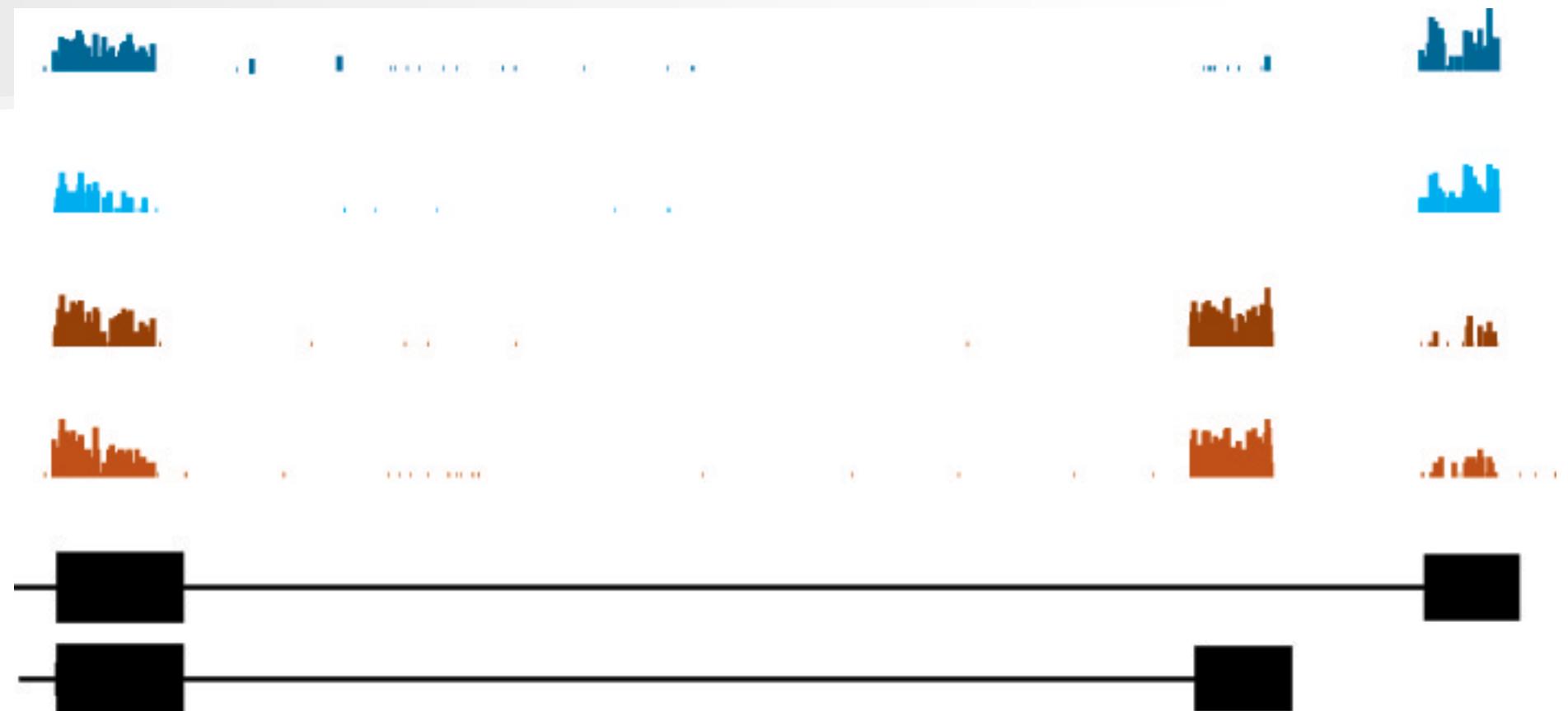


Alignment

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

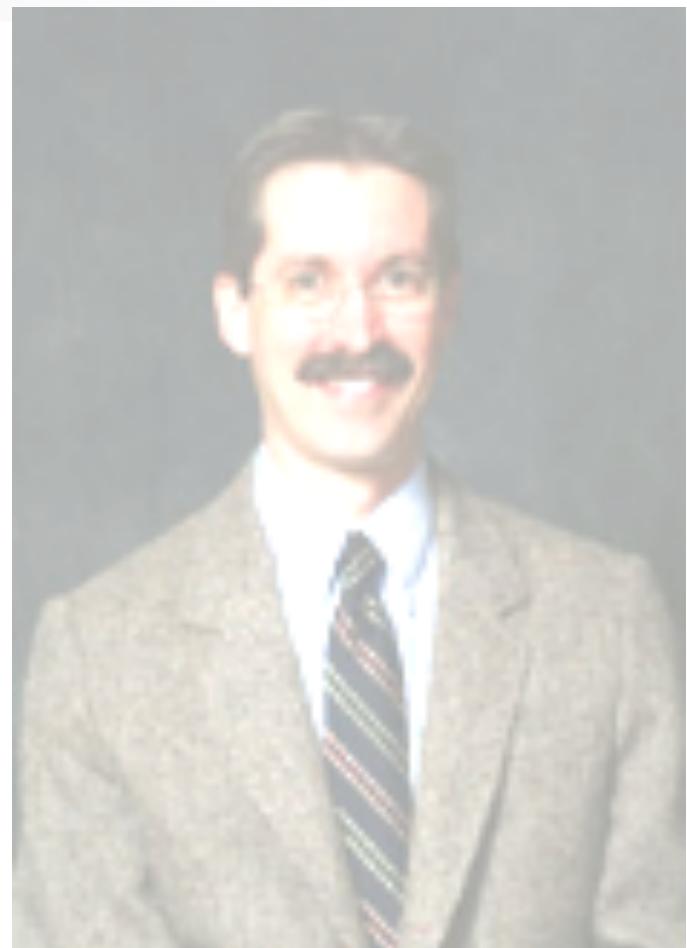
1. Background correction.
2. Data normalization.
3. Data summary methods (Affymetrix arrays only).

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

Preprocessing - Background Correction

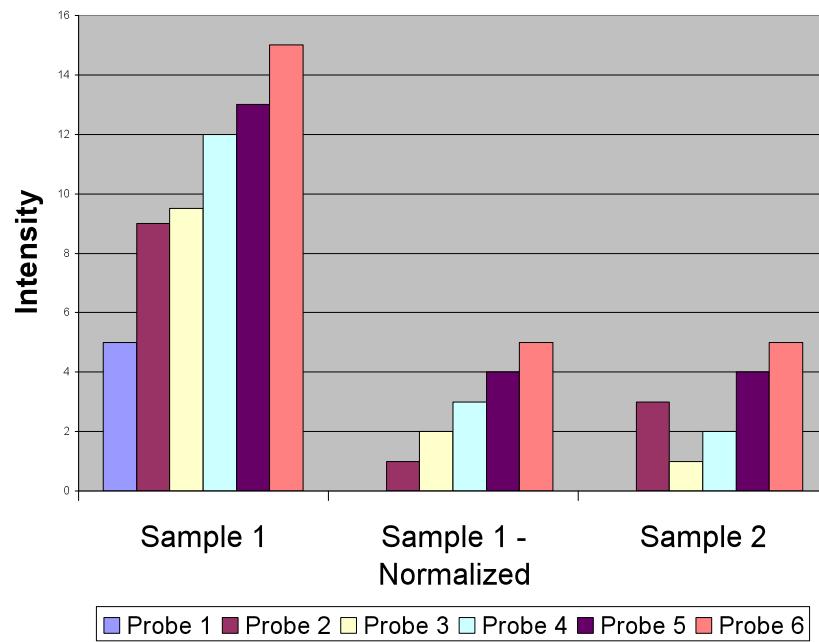
- Previous generation – perfect match versus mismatch
- Exon Array – mismatch probe identical in GC contents not necessarily identical sequence
- The popular RMA method uses neither

Preprocessing - Data Normalization



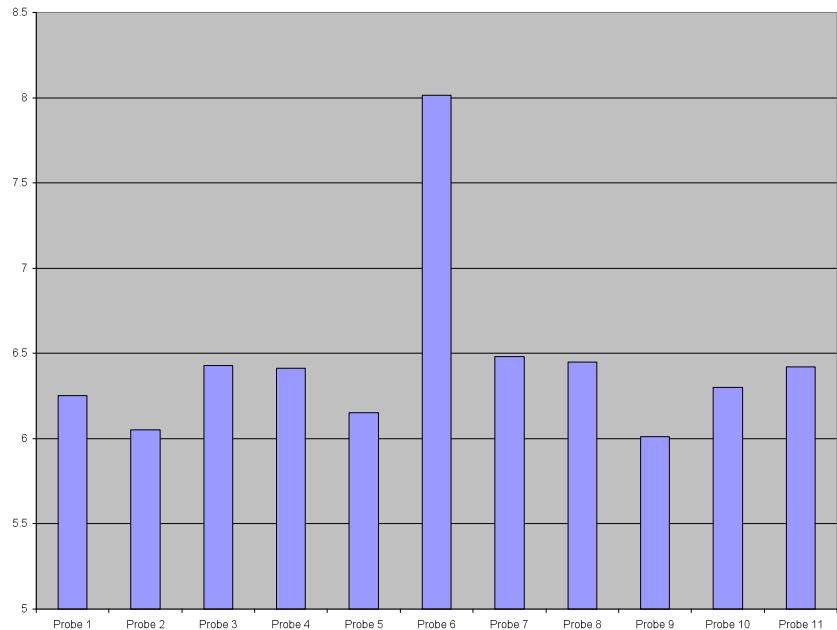
Preprocessing - Data Normalization

- 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

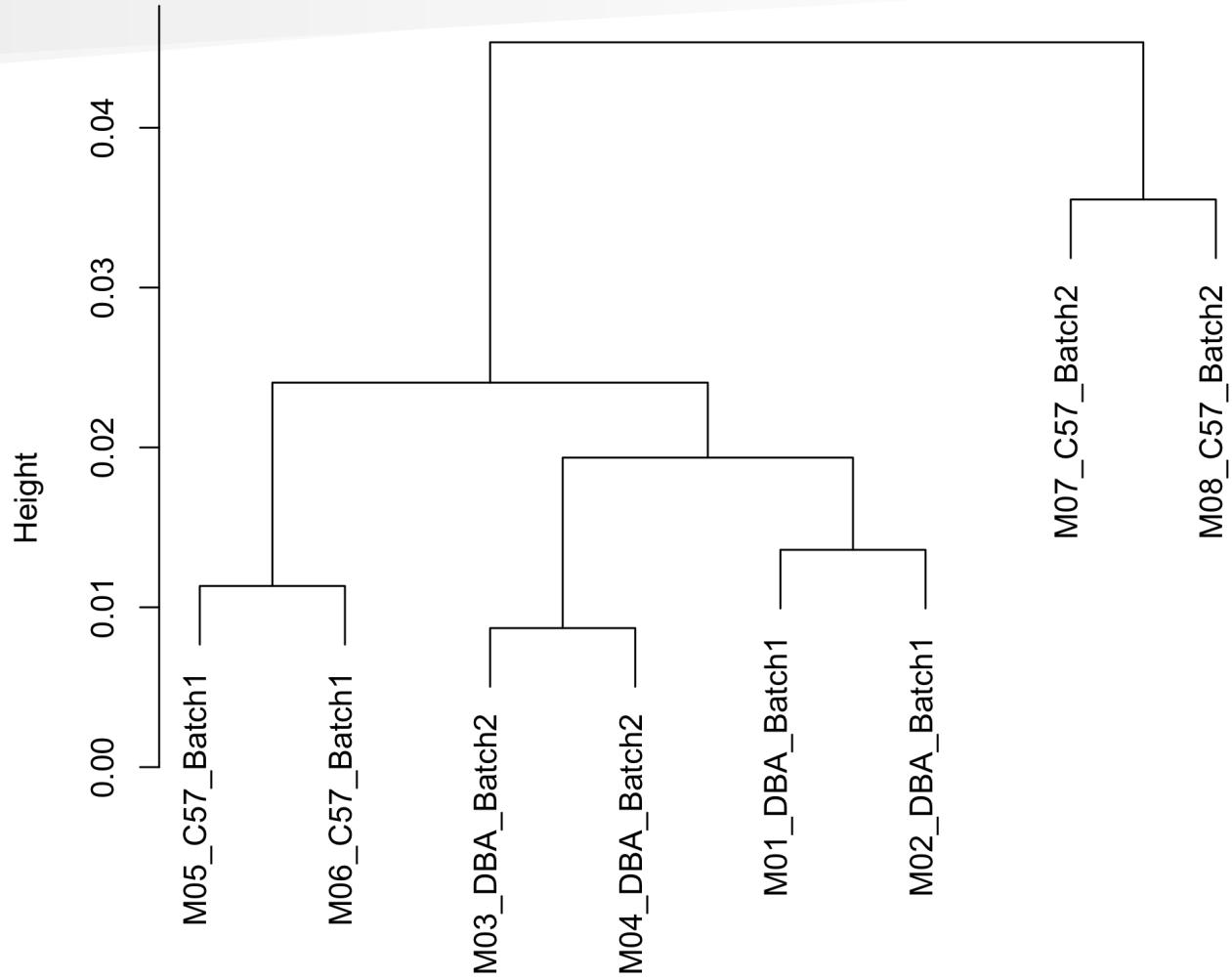
Before

After

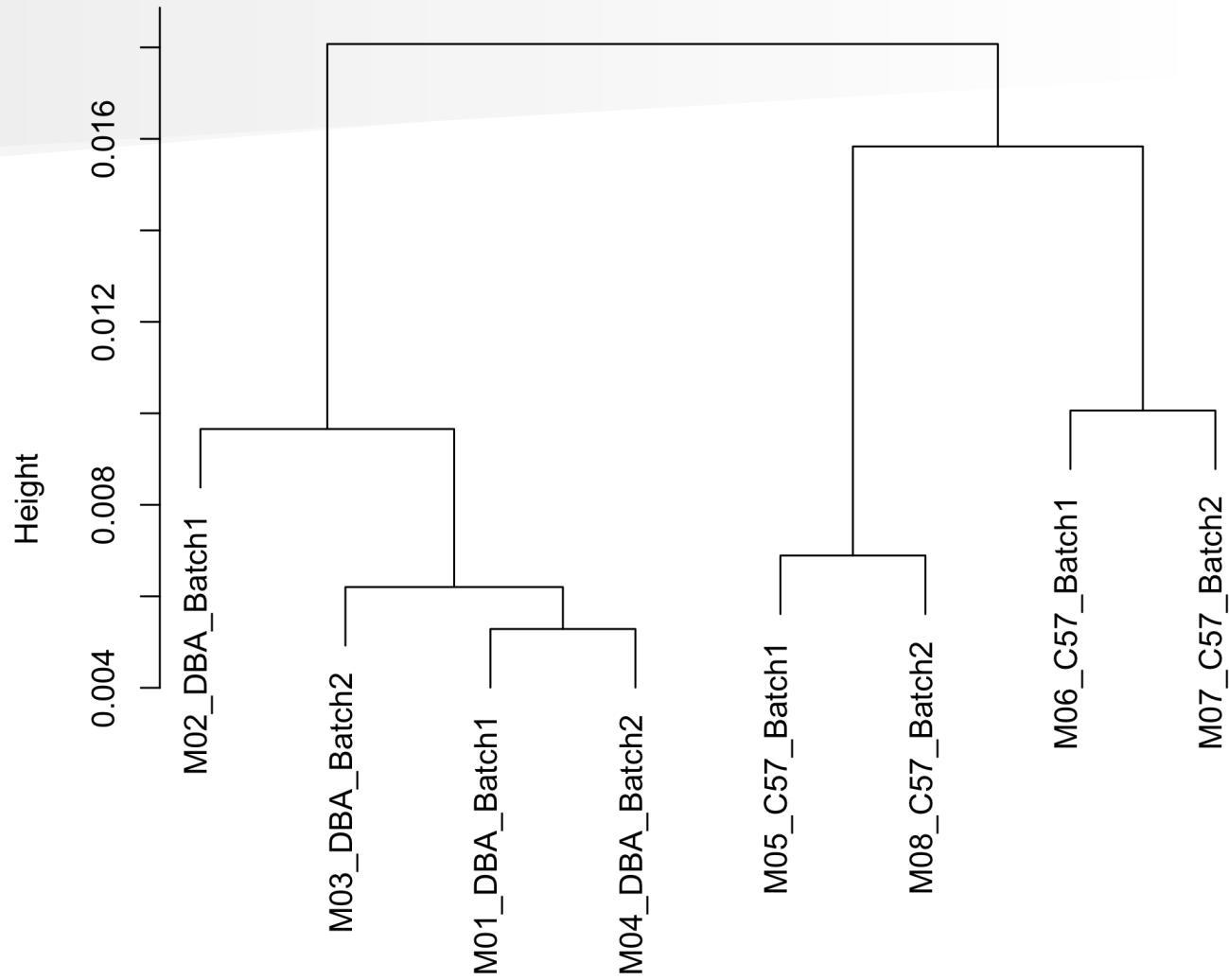
Last Saved: 12/6/10 11:24:22 AM
File Path ▾ ~/Documents/BIOS6606/Dec2...ByTranscript/rma.summary.txt

```
%affymetrix-algorithm-param-opt-analysis-name=rma
%affymetrix-algorithm-param-opt-set-analysis-name=
%affymetrix-algorithm-param-opt-opt-analysis-spec=rma-bg,quant-norm.sketch=0.bioc=true
%affymetrix-algorithm-param-opt-opt-feat-effect-file=
%affymetrix-algorithm-param-opt-opt-target-sketch-file=
%affymetrix-algorithm-param-opt-opt-do-residuals=false
%affymetrix-algorithm-param-opt-opt-do-feature-effects=false
%affymetrix-algorithm-param-opt-opt-write-sketch=false
%affymetrix-algorithm-param-opt-opt-reference-profile-file=
%affymetrix-algorithm-param-opt-opt-write-profile=false
%affymetrix-algorithm-param-quantification-name=med-polish
%affymetrix-algorithm-param-quantification-version=1.0
%affymetrix-algorithm-param-quantification-scale=log2
%affymetrix-algorithm-param-quantification-type=signal
probeset_id M01_Female_Run6.CEL M09_Male_Run7.CEL M02_Female_Run6.CEL M03_Female_Run7.CEL
6747308 8.84344 8.40757 8.98358 8.44696 8.73178 8.76550 8.91071 8.51891
6747314 10.28132 10.02364 10.09156 9.81927 10.01893 10.06542 10.27841
6747326 7.70527 7.27395 7.81672 7.23312 7.37181 7.19889 7.17008 7.07652
6747343 8.79006 8.69034 8.98434 8.65746 8.74434 8.59996 8.80344 8.54720
6747354 7.54297 8.64526 9.14656 8.86455 8.69390 8.53700 8.91536 8.76443
6747364 9.54800 9.21832 9.56244 9.14329 9.23146 9.62150 9.59645 9.39683
6747471 8.72862 8.89551 8.60148 8.76671 8.45785 8.74485 8.72736 9.04915
6747472 9.02250 8.87539 8.95841 8.78164 8.86599 8.97979 9.06308 9.00691
6747478 11.20206 11.32222 10.91494 10.63808 10.98175 11.65343 11.65491
6747497 8.60269 8.16822 8.40433 8.18941 8.24317 8.41404 8.71043 8.15941
6747504 5.45968 5.84208 5.60565 5.55524 5.28002 5.87713 5.81629 5.63734
6747515 8.93716 8.61795 9.00265 8.58842 8.76951 8.66895 8.88811 8.42517
6747577 9.72336 9.15569 9.48709 9.31791 9.49086 9.35974 9.49176 9.14826
6747641 8.20517 8.00987 8.32271 8.01797 8.10450 8.18370 8.38642 8.19356
6747696 4.38686 4.87435 4.47421 4.71493 4.78553 4.78243 4.31052 4.78404
6747786 7.65765 7.33593 7.95608 7.25764 7.44643 7.57848 7.89834 7.23765
6747805 8.56914 8.50227 8.85832 8.53860 8.46068 8.66978 8.74545 8.63771
6747837 8.14370 7.89827 8.42927 7.66796 7.98983 8.18178 8.17907 7.96952
6747839 6.79799 6.47090 7.22446 6.47034 6.92426 6.09930 6.51488 6.64590
6747850 10.23608 9.76983 10.20765 9.81606 9.92492 9.86731 10.13003 9.75350
6747861 4.81838 4.95721 5.39518 5.30400 5.23672 5.10029 5.22394 5.47696
6747871 7.22412 7.16064 7.09133 7.06854 6.99476 7.05760 7.16628 7.12835
6747912 6.97329 7.22461 7.40500 7.06965 7.05881 7.32978 7.41708 7.40757
6747919 7.48872 7.95399 8.29987 8.05070 8.04091 8.05219 8.24782 8.07498
6747972 6.70064 7.50923 6.47232 6.91441 6.79970 6.71995 6.53280 6.92023
```

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

$$\begin{array}{c} \text{Prior} \\ \text{Knowledge} \end{array} + \begin{array}{c} \text{Observed} \\ \text{Data} \end{array} = \begin{array}{c} \text{Inference} \end{array}$$

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

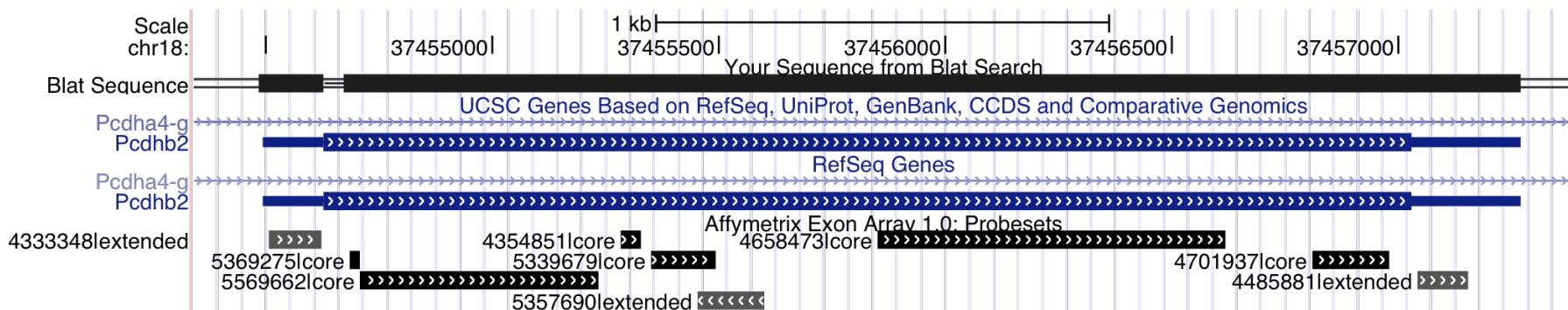
Accession ID	Official Symbol	RefSeq	MGI	UniProt	UC Santa Cruz	Genetically Modified Animal Available	QTLs 	Genetic Variations 	Allen Brain Atlas (Instructions)
1418704_at	S100a13	NM_009113	MGI:109581	P97352	NM_009113	MGI	PhenoGen eQTL	ENSMUST00000048138 S100a13	Link
1421844_at	Il1rap	NM_008364 NM_134103	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	Mobb	NM_001039364 NM_001039365 NM_008614	MGI:108511	Q9D2P8	NM_001039364 NM_001039365 NM_008614	MGI	PhenoGen eQTL	ENSMUST00000035103 ENSMUST00000093773 ENSMUST00000068698 Mobb	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	ENSMUST0000005606 ENSMUST00000095228 Prkaca	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

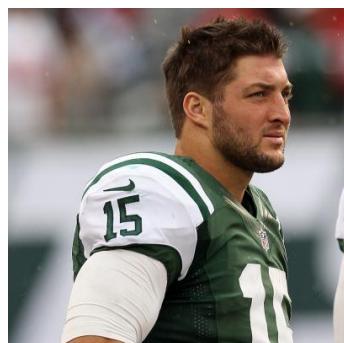
To Do Before Interpreting Gene List

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



RNA-Seq Analyses

Genomic Drivers of Alcohol Consumption

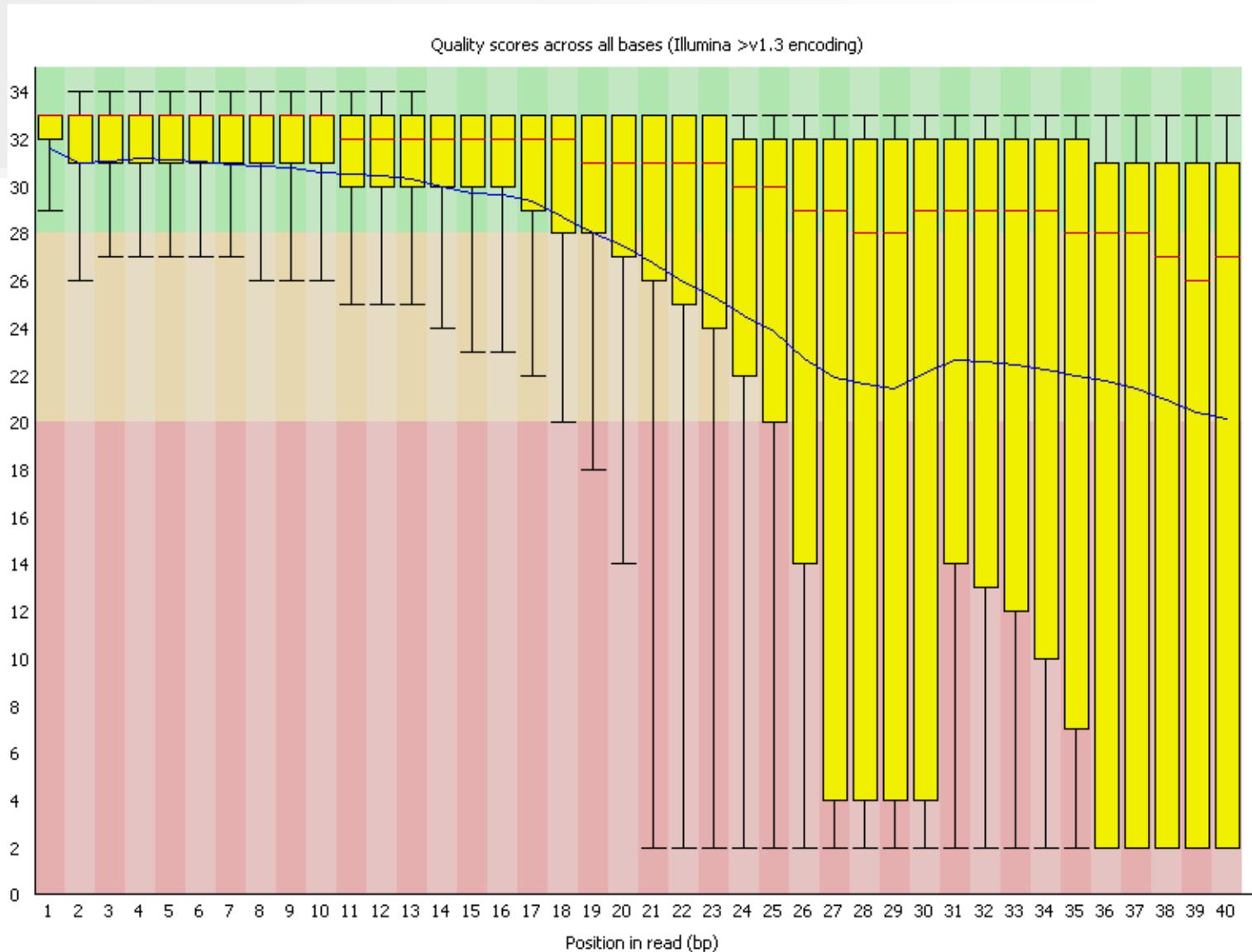


Genes that predispose to differing levels of alcohol consumption

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 screenshot shows a web browser window with the title "Bowtie 2: fast and sensitive re..." and the URL "bowtie-bio.sourceforge.net/bowtie2/index.shtml". The page content is as follows:

Bowtie 2
Fast and sensitive read alignment

Bowtie 2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. It is particularly good at aligning reads of about 50 up to 100s or 1,000s of characters, and particularly good at aligning to relatively long (e.g. mammalian) genomes. Bowtie 2 indexes the genome with an FM Index to keep its memory footprint small: for the human genome, its memory footprint is typically around 3.2 GB. Bowtie 2 supports gapped, local, and paired-end alignment modes.

Hiring Postdocs
The Langmead and Salzberg labs currently have open positions for postdoctoral researchers. See [the posting](#) and please apply if you're interested in working with either or both of us.

Version 2.0.2 - October 31, 2012
Fixes a couple small issues pointed out to me immediately after 2.0.1 release
Mac binaries now built on 10.6 in order to be forward-compatible with more Mac OS versions
Small changes to source to make it compile with gcc versions up to 4.7 without warnings

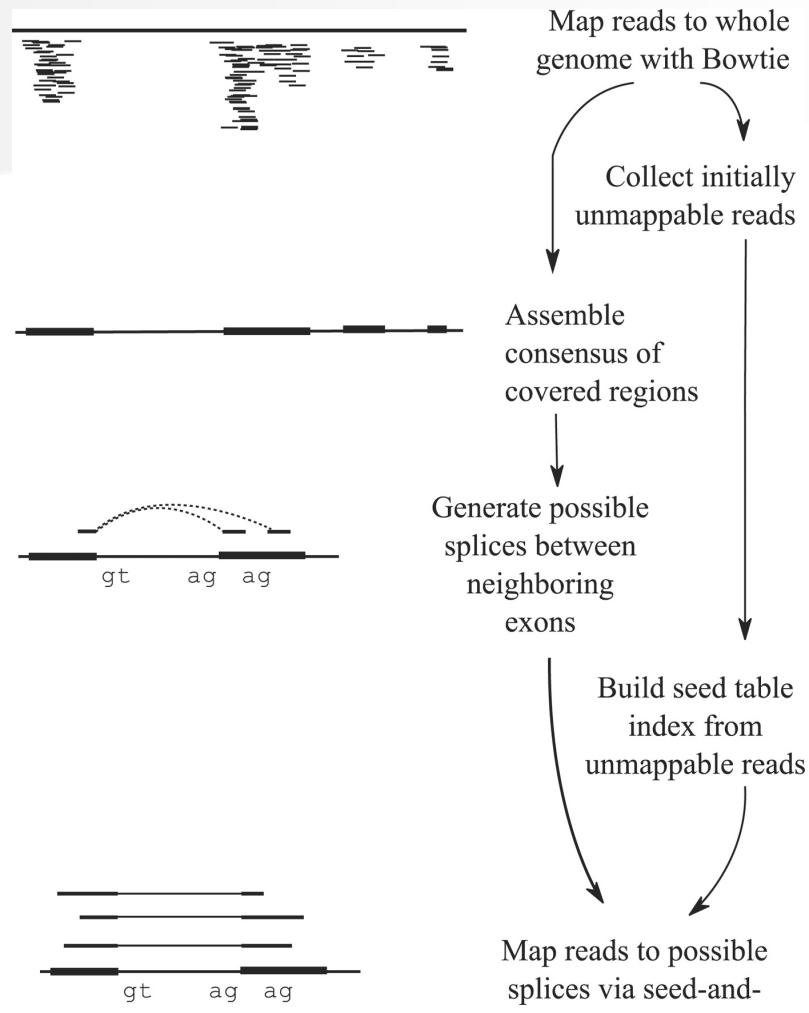
Version 2.0.1 - October 31, 2012
First non-beta release.
Fixed an issue that would cause Bowtie 2 to use excessive amounts of memory for closely-matching and highly repetitive reads under some circumstances.
Fixed a bug in `--mm` mode that would fail to report when an index file could not be memory-mapped.
Added `--non-deterministic` option, which better matches how some users expect the pseudo-random generator inside Bowtie 2 to work. Normally, if you give the same read (same name, sequence and qualities) and `--seed`, you get the same answer. `--non-deterministic` breaks that guarantee. This can be more appropriate for datasets where the input contains many identical reads (same name, same sequence, same qualities).
Fixed a bug in `bowtie2-build` would yield corrupt index files when memory settings were adjusted in the middle of

Site Map
Home
News archive
Manual
Getting started
Frequently Asked Questions
Tools that use Bowtie

Latest Release
Bowtie2 2.0.2 8/31/12
Please cite: Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 2012, 9:357-359.

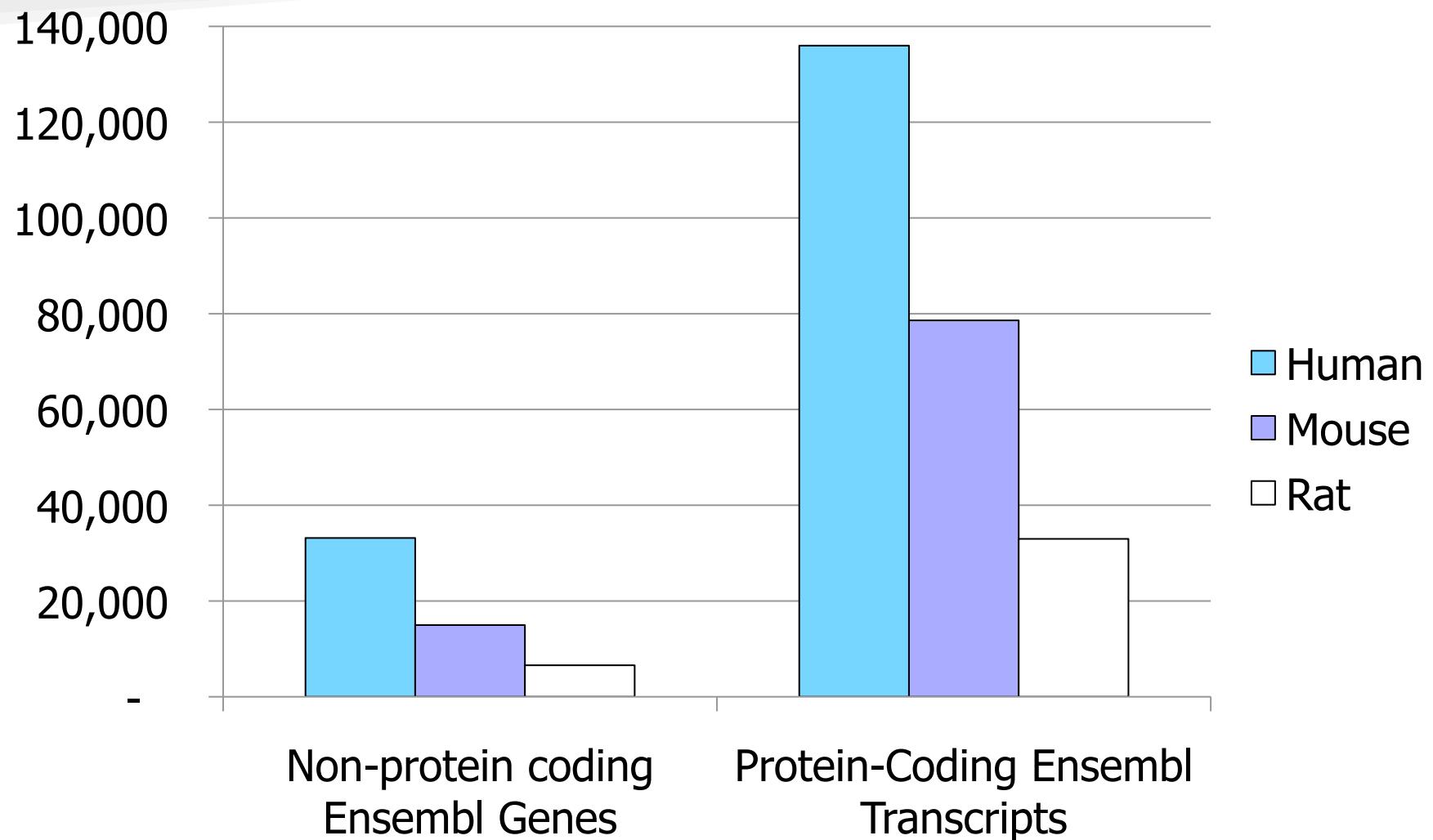
Related Tools
Bowtie: Ultrafast short read alignment
Crossbow: Genotyping, cloud computing
Myrna: Cloud, differential gene expression
Tophat: RNA-Seq splice junction mapper
Cufflinks: Transcript assembly, quantitation

The TopHat pipeline.

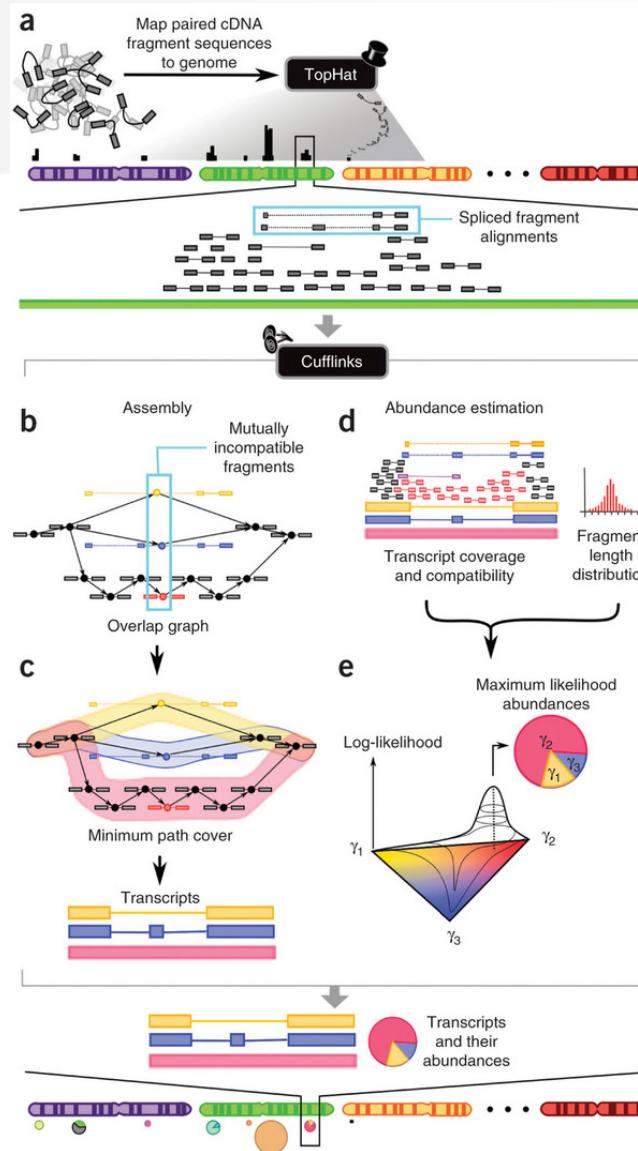


Trapnell C et al. Bioinformatics 2009;25:1105-1111

Why Transcriptome Reconstruction?



Transcriptome Reconstruction



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

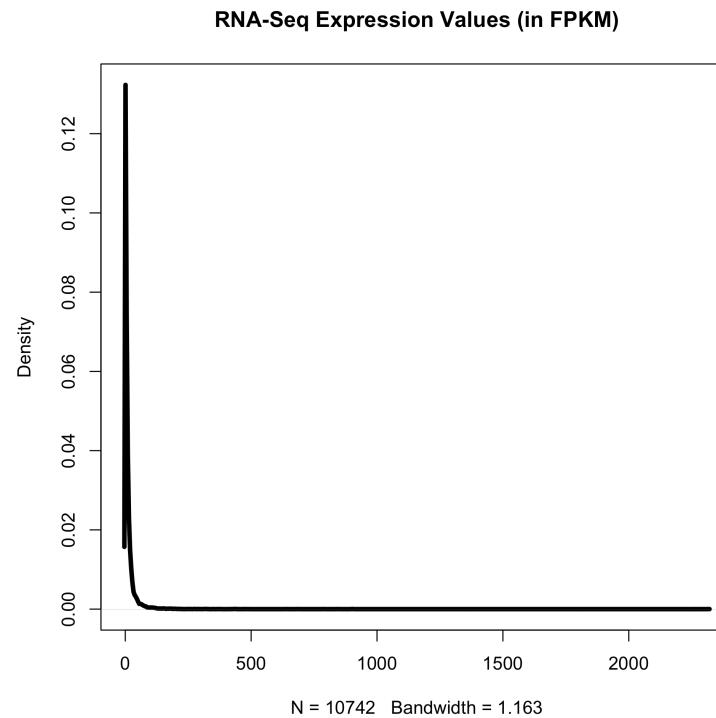
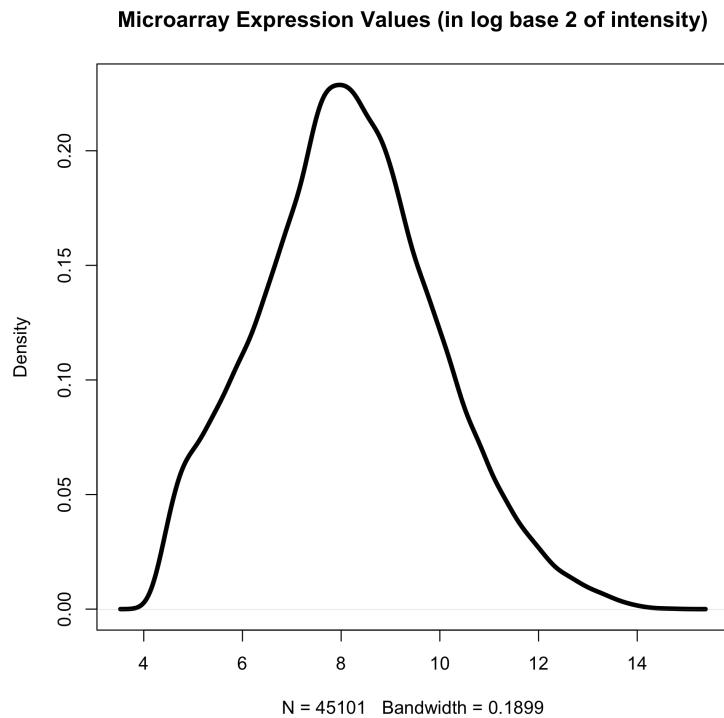
Quantifying Expression

■ Count Number of Reads

- multiple transcripts per read (Bayesian Deconvolution 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

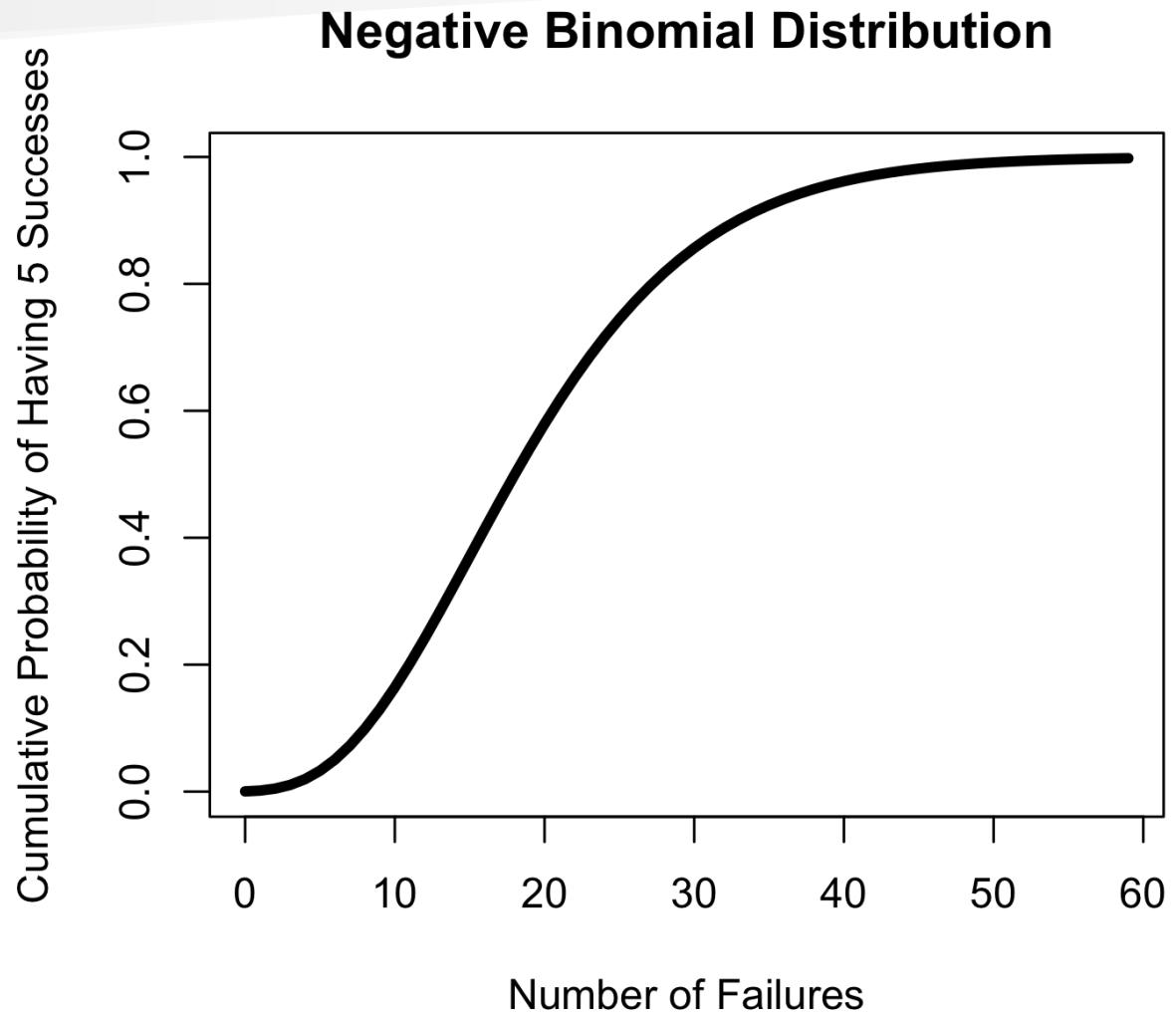
Negative Binomial Distribution

- 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 two 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

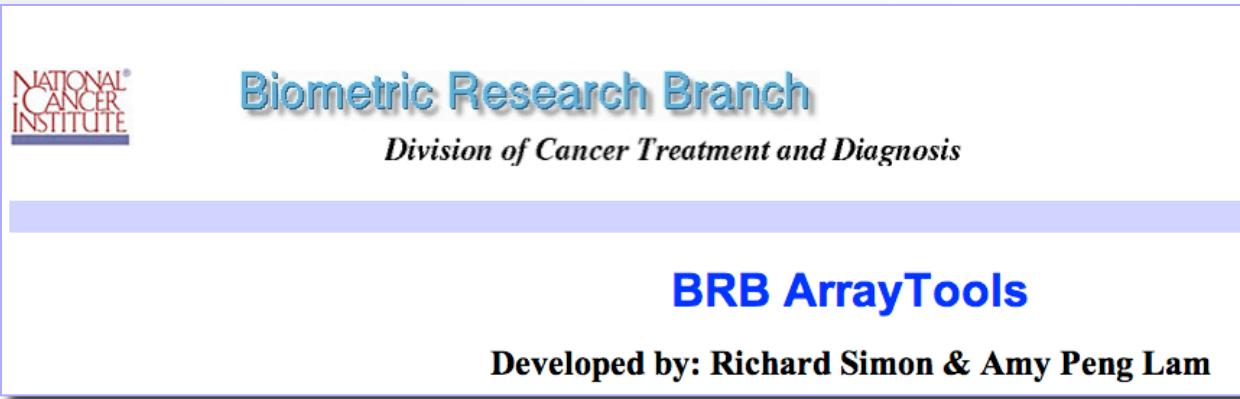
The image shows two side-by-side screenshots of scientific software websites. On the left is the 'The R Project for Statistical Computing' website, featuring a large R logo, a navigation menu with links like 'About R', 'What is R?', 'Contributors', 'Screenshots', 'What's new?', 'Download CRAN', 'R Project Foundation', 'Meetings & Donors', 'Mailing List', 'Bug Tracking', 'Developer Page', 'Conferences', and 'Search'. Below the menu are several data visualization plots, including a PCA plot titled 'PCA 5 vars' with axes 'Fertility' and 'Catholic', and a dendrogram titled 'Clustering 4 groups'. On the right is the 'BIOCONDUCTOR open source software for bioinformatics' website, featuring a large blue header with the BioConductor logo and the tagline 'Bioconductor is an open source and open development software project for the analysis and comprehension of genomic data.' Below the header is a navigation menu with links 'home', 'what is it?', 'download', 'documentation', 'publications', and 'workshops'. A central image shows laboratory glassware. To the right of the image is a 'project news' section with two entries: '2006-08-01 Changes in BioC Devel, August 2006' and '2006-07-07 Changes in BioC Devel, July 2006', followed by a 'More...' link.

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

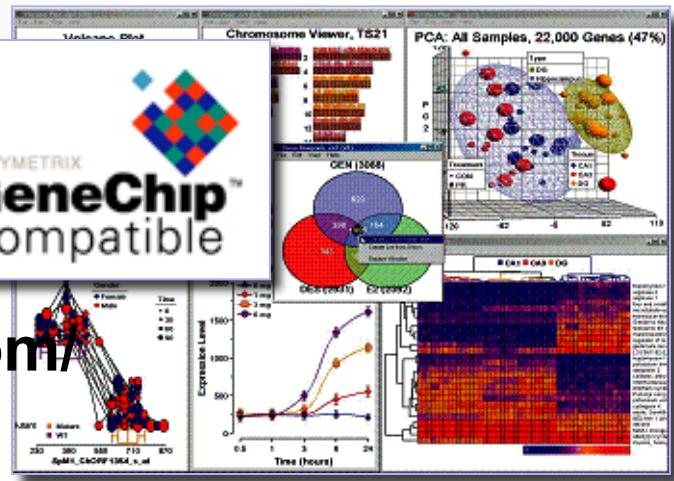
BRB-ArrayTools



<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



The screenshot displays the Partek Genomics Suite software interface. At the top, there is a banner with the text "Partek® Genomics Suite" and "AFFYMETRIX GeneChip™ Compatible". Below the banner, several data visualizations are shown: a chromosome viewer for T521, a PCA plot titled "PCA: All Samples, 22,000 Genes (47%)", a Venn diagram for GCN (3568), a scatter plot of expression levels over time (0.5 to 24 hours) for different concentrations (0.1, 0.5, 1, 5 mg), and a heatmap of expression data.

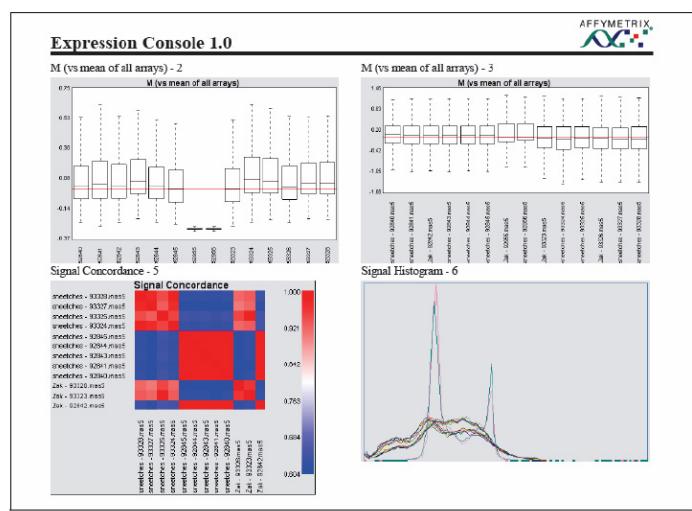
Partek® Genomics Suite

AFFYMETRIX GeneChip™ Compatible

http://www.partek.com/

- ✓ 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/>

The screenshot shows a web browser window titled "PhenoGen" displaying the homepage of the PhenoGen Informatics website. The URL in the address bar is "phenogen.ucdenver.edu/PhenoGen". The page features a dark blue header with the title "PhenoGen Informatics" and a background image of a DNA helix. Below the header is a navigation menu with green and blue buttons. The menu items include "Overview", "Detailed Transcription Information", "Downloads", "Microarray Analysis Tools", "Gene List Analysis Tools", "QTL Tools", "About", "Help", and "Login/Register". The main content area has a white background and contains text about the site's purpose, sections, and login requirements.

Welcome to PhenoGen Informatics

The PhenoGen Informatics web site is not only a microarray repository but also a comprehensive toolbox for analyzing microarray data and researching candidate genes.

The site is organized into five major sections:

- [Detailed Transcription Information](#)
- [Downloads](#)
- [Microarray Analysis Tools \(login required\)](#)
- [Gene Analysis Tools \(login required\)](#)
- [QTL Tools \(login required\)](#)

Click the Overview option above to see examples of what you can do on our site. To the right of Overview are the main areas of the site. The functions with a green background indicate publicly accessible parts of the site while the remaining blue functions require a login.

View the [Getting Started With PhenoGen Informatics Demo](#) to learn how to get started.

Review the [current datasets](#) that we have available for public use.

Why do we require a login?

For many of the tools that require a login there are multiple intermediate steps in the analysis or steps may take a long time to complete. A login allows you to start a step and come back to the analysis/results at a later time. This also allows you to upload data

Our Resource

<http://phenogen.ucdenver.edu/>

The screenshot shows the Phenogen Informatics website interface for analyzing microarray data. At the top, there is a dark header bar with the title "PhenoGen Informatics" and a DNA helix graphic. Below the header, a navigation menu includes links for "Home", "My Profile", "Contact Us", "Logout", and "Site Help". The main content area displays the current analysis project: "You are Analyzing: Comparison of C57 and DBA in CodeLink v1". A "Select Different Dataset" link is also present. The main steps for analysis are outlined: "Steps to select a normalized dataset for analysis: 1. Group and Normalize Dataset; 2. Select Normalized Version; 3. Select Type of Analysis; 4. Run Analysis". Step 2 is highlighted with a large number "2". Below this, a message says "You may perform any of the following types of analyses on your normalized dataset. Choose one to continue." A dropdown menu titled "Select Analysis Method" is open, showing options: "Select Analysis Method", "Differential Expression", "Correlation", and "Clustering". At the bottom of the page, there are links for "Download Manual", "Citations", "Useful Links", and "Version Information".

Our Resource

<http://phenogen.ucdenver.edu/>

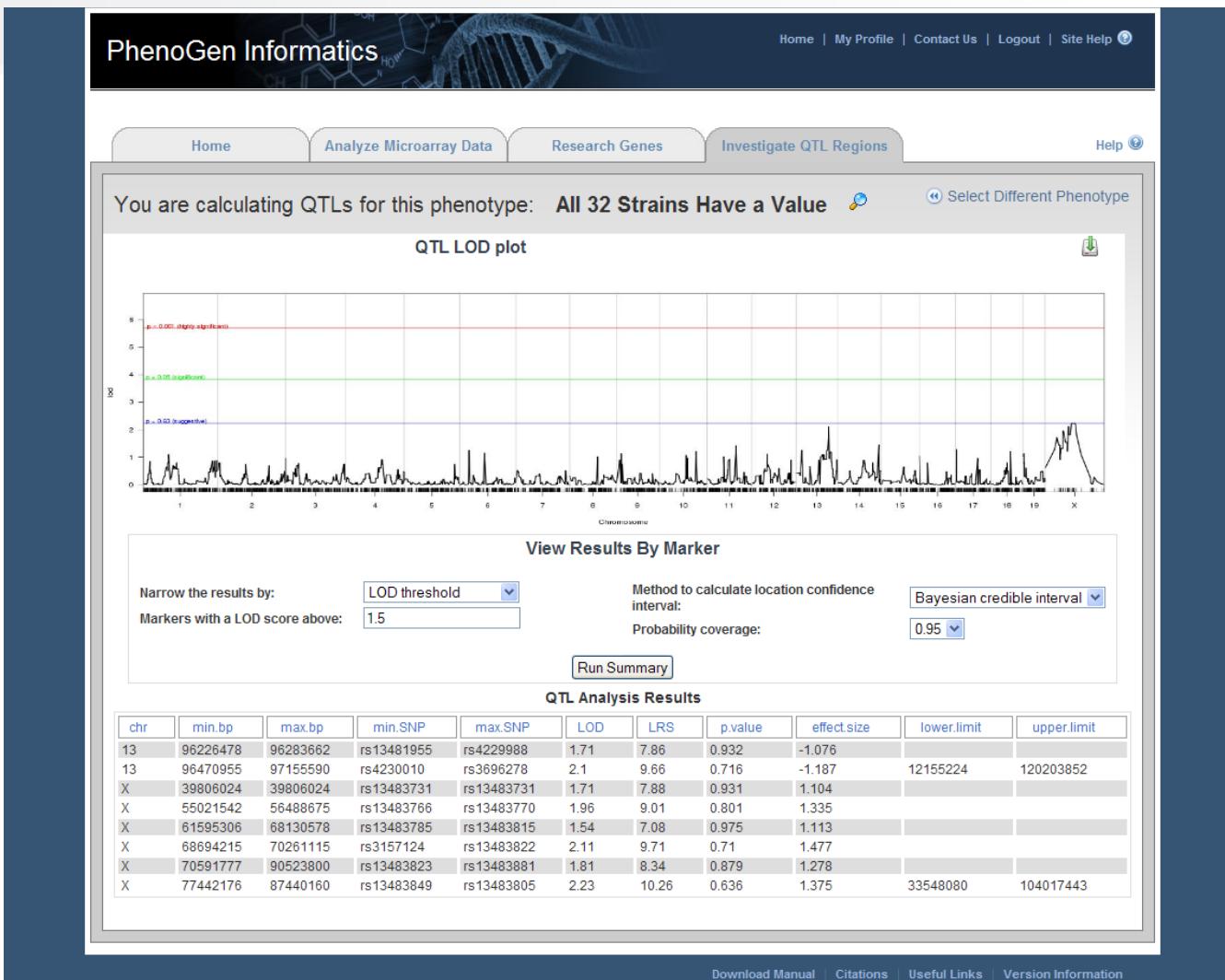
The screenshot shows the Phenogen Informatics website interface. At the top, there is a dark header bar with the title "PhenoGen Informatics" and a decorative background featuring a DNA helix and chemical structures. The header also includes links for "Home", "My Profile", "Contact Us", "Logout", and "Site Help". Below the header, a breadcrumb navigation shows "Home > Research Genes > List". A horizontal menu bar contains links for "Home", "Analyze Microarray Data", "Research Genes" (which is currently selected), and "Investigate QTL Regions". There is also a "Help" link with a question mark icon. The main content area displays a message "You are Viewing: Filtered Morphine Genes" with a blue gear icon, and a link to "Select Different Gene List". A note below states, "This page contains the identifiers and their symbols for the genes in your list." A toolbar below the message includes buttons for "List" (selected), "Annotation", "Location", "Literature", "Promoter", "Homologs", "Expression Values", "Save As...", "Compare", and "Share". A table titled "Gene List Contents" lists gene identifiers and symbols. The table has two columns: "Accession ID" and "GeneSymbol". The data is as follows:

Accession ID	GeneSymbol
1418704_at	S100a13
1421844_at	Il1rap
1424634_at	Tceal1
1433966_x_at	Asns
1436263_at	Mobp
1438060_at	Npas3
1438241_at	Rgma
1445143_at	Vash1
1448465_at	Nipsnap1
1448888_at	Ppp1r7
1449381_a_at	Pacsin1
1450519_a_at	Prkaca
1450685_at	Arpp19
1452404_at	Phactr2
1456527_at	Hecw1

At the bottom of the page, there are links for "Download Manual", "Citations", "Useful Links", and "Version Information".

Our Resource

<http://phenogen.ucdenver.edu/>

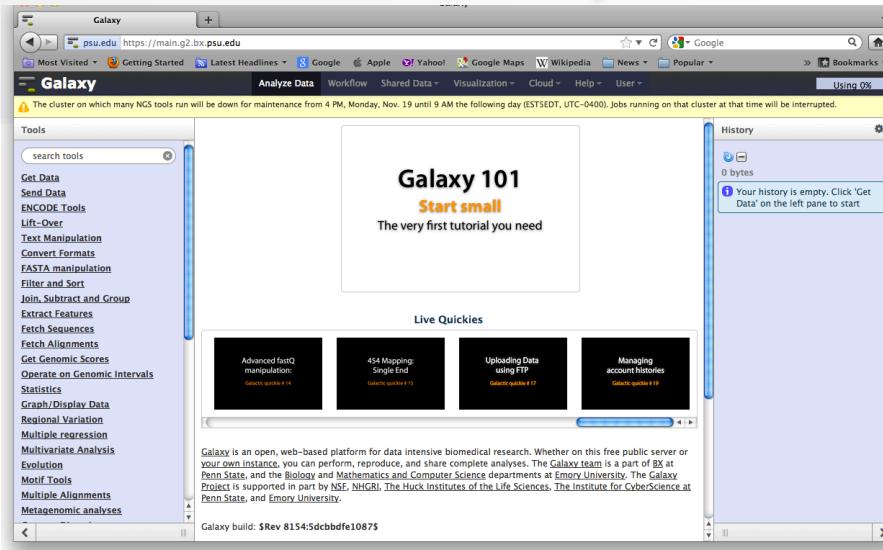


RNA-Seq Resources

Popular RNA-Seq Software

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

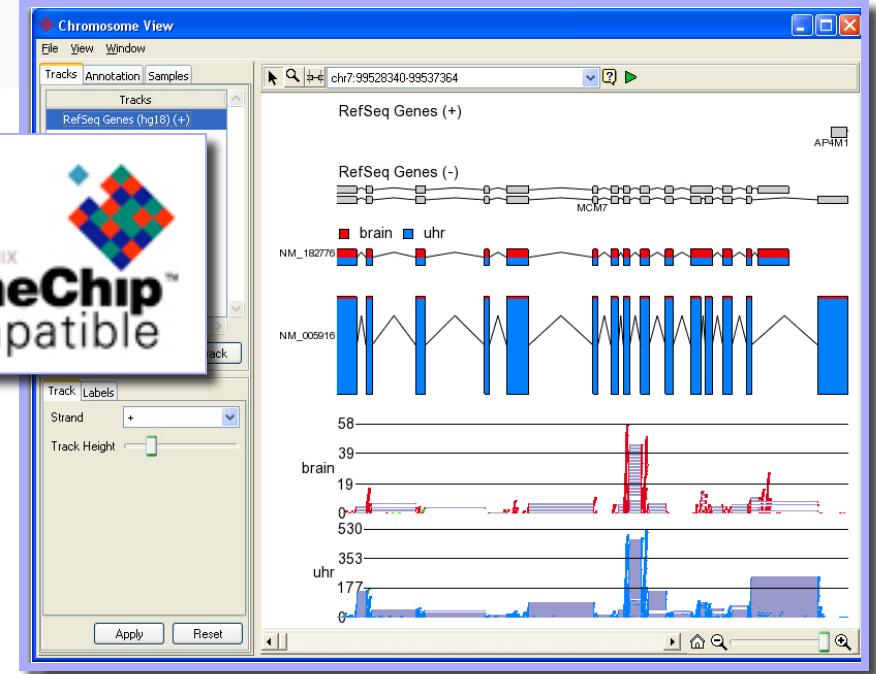
Galaxy



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

Partek Genomics Suite

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

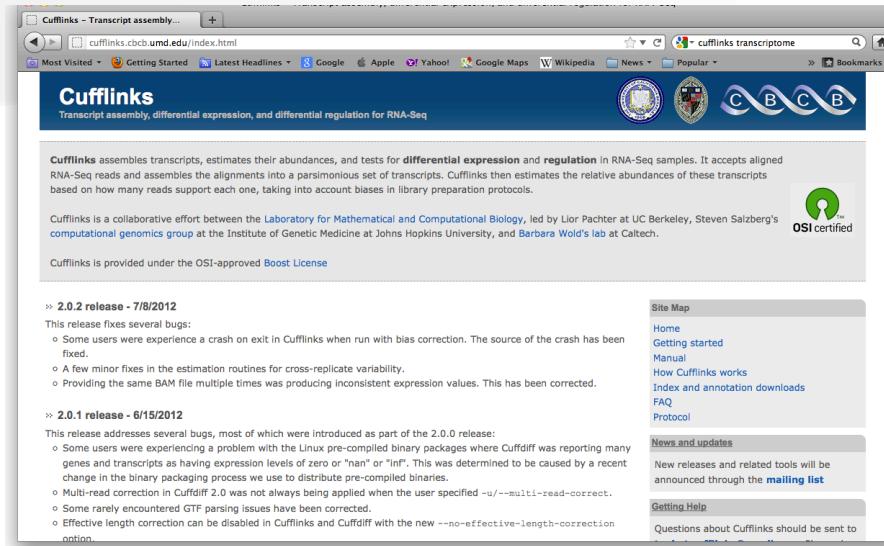
The image shows two side-by-side screenshots of scientific software websites. On the left is the 'The R Project for Statistical Computing' website, featuring a large R logo, a navigation menu with links like 'About R', 'What is R?', 'Contributors', 'Screenshots', 'What's new?', 'Download CRAN', 'R Project Foundation', 'Meetings & Donors', 'Mailing List', 'Bug Tracking', 'Developer Page', 'Conferences', and 'Search'. Below the menu are several data visualization plots, including a PCA plot titled 'PCA 5 vars' with axes 'Fertility' and 'Catholic', and a dendrogram titled 'Clustering 4 groups'. On the right is the 'BIOCONDUCTOR open source software for bioinformatics' website, featuring a large blue header with the BioConductor logo and the tagline 'Bioconductor is an open source and open development software project for the analysis and comprehension of genomic data.' Below the header is a navigation menu with links 'home', 'what is it?', 'download', 'documentation', 'publications', and 'workshops'. A central image shows laboratory glassware. To the right of the image is a 'project news' section with two entries: '2006-08-01 Changes in BioC Devel, August 2006' and '2006-07-07 Changes in BioC Devel, July 2006', followed by a 'More...' link.

www.r-project.org

www.bioconductor.org

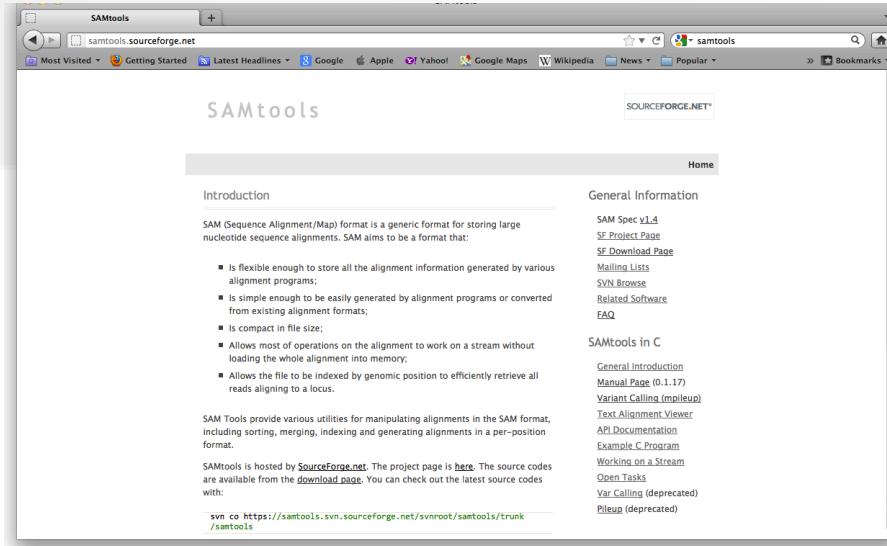
- ✓ Open-source software, free for the public
- ✓ Platform independent
- ✓ Unlimited statistical and programming flexibility
- ✗ Need command line usage experience
- ✗ Requires significant R programming experience
- ✗ Can't do everything

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

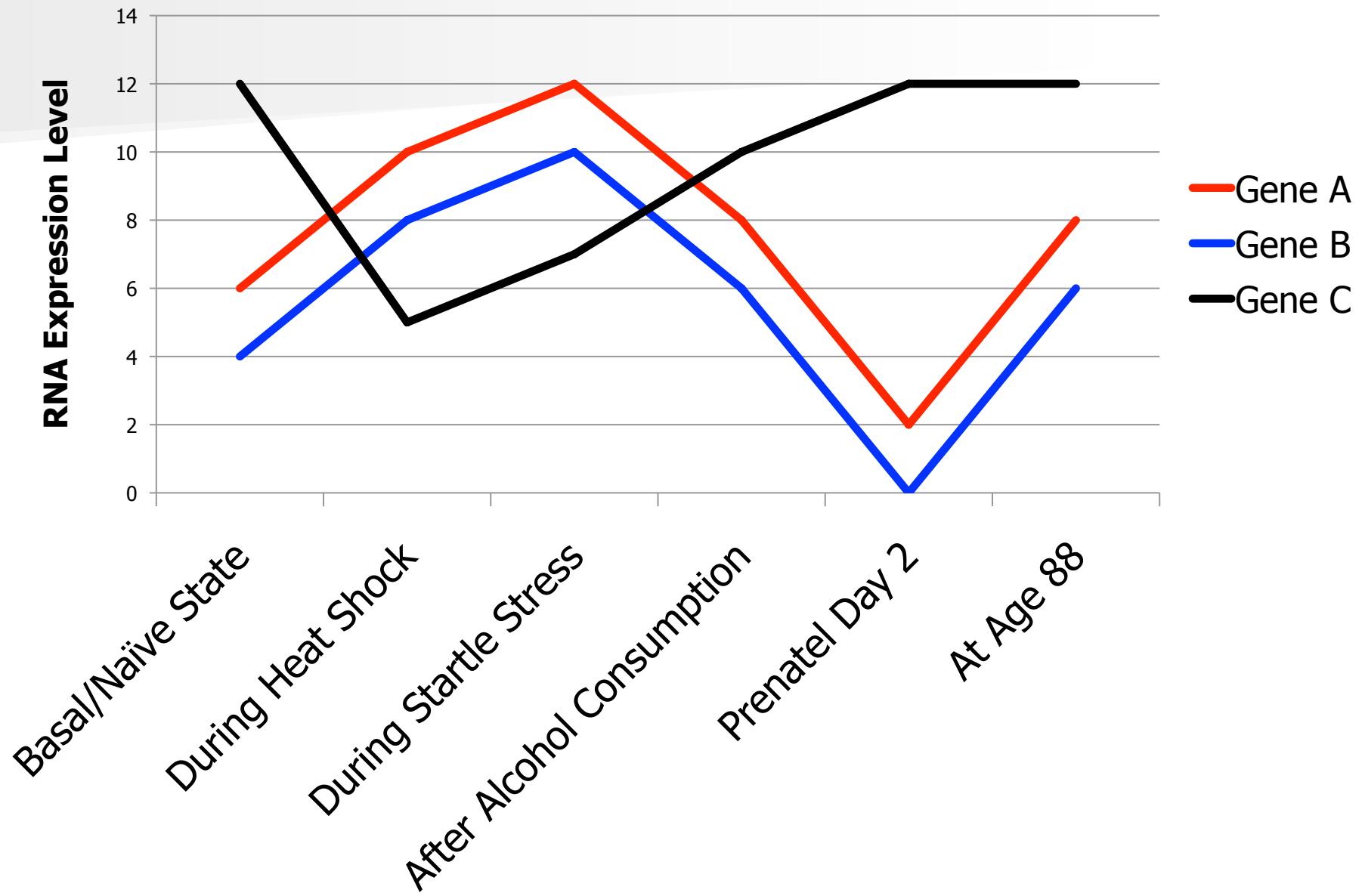


- ✓ Free for the public
- ✗ Mainly used for manipulating SAM files and genotype calls

Topics

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

Characterize Relationships Among Transcripts

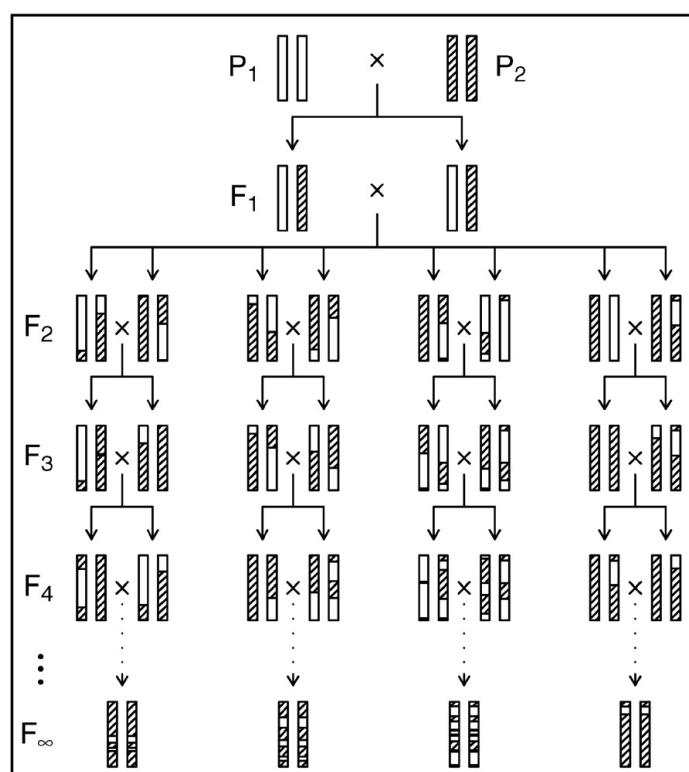


What if Lohan and Tebow a baby?



Recombinant Inbred Rodent Panels

B Sibling mating



- 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

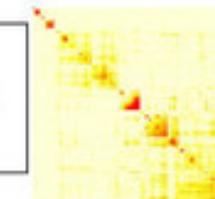
Broman, K. W. *Genetics* 2005;169:1133-1146

Weighted Gene Co-Expression Analysis

Construct a gene co-expression network

Rationale: make use of interaction patterns among genes

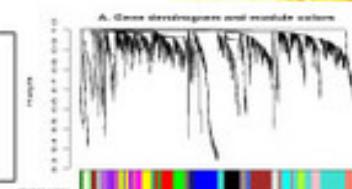
Tools: correlation as a measure of co-expression



Identify modules

Rationale: module (pathway) based analysis

Tools: hierarchical clustering, Dynamic Tree Cut

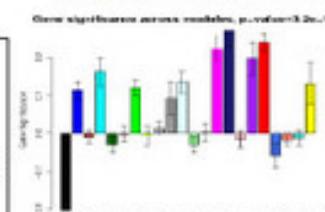


Relate modules to external information

Array Information: clinical data, SNPs, proteomics

Gene Information: ontology, functional enrichment

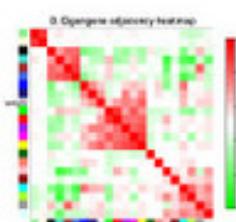
Rationale: find biologically interesting modules



Study module relationships

Rationale: biological data reduction, systems-level view

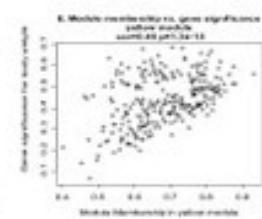
Tools: Eigengene Networks



Find the key drivers in *interesting* modules

Rationale: experimental validation, biomarkers

Tools: intramodular connectivity, causality testing



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 than 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**

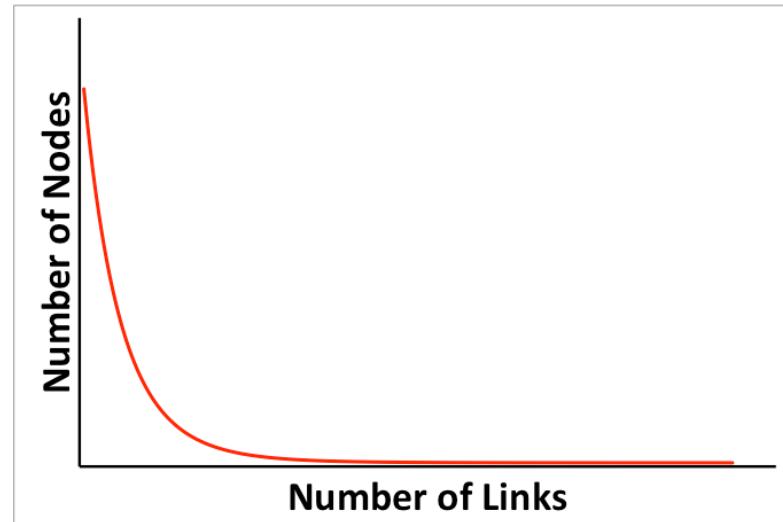
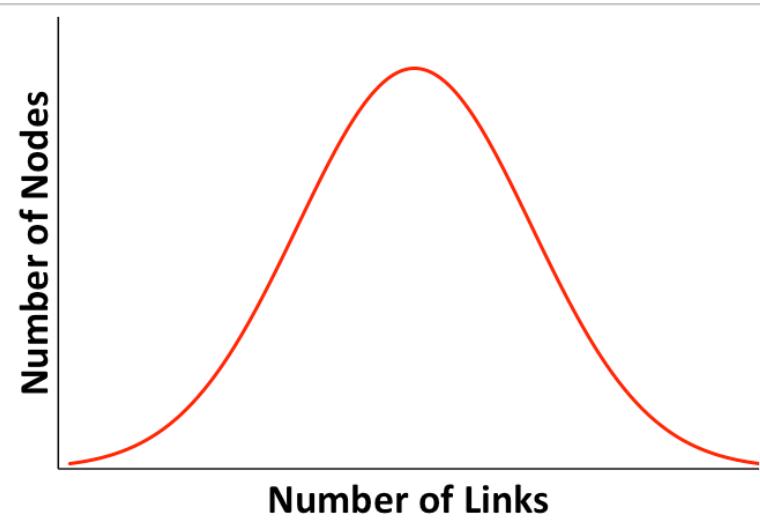
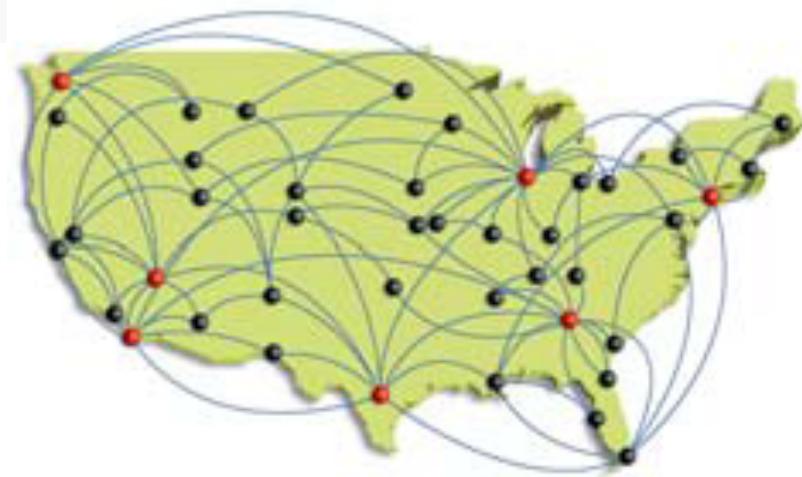
- Includes a measure of how many “friends” two genes have in common**
- Protects against spurious correlations among genes**

Scale-Free Networks

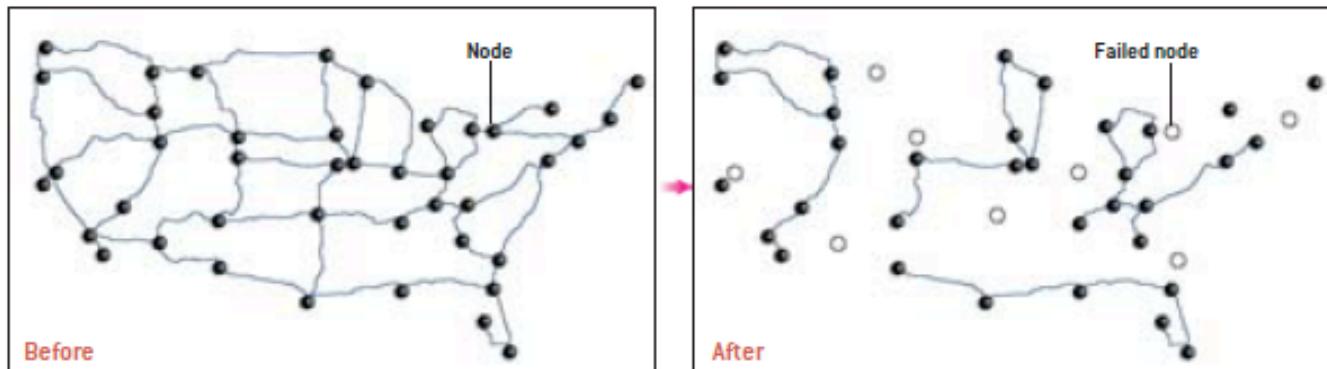
Random Network



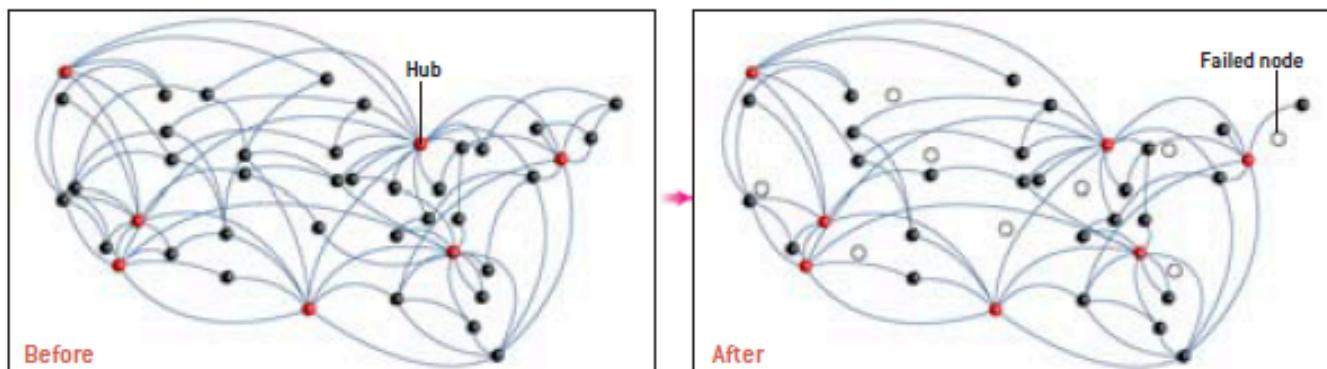
Scale-Free Network



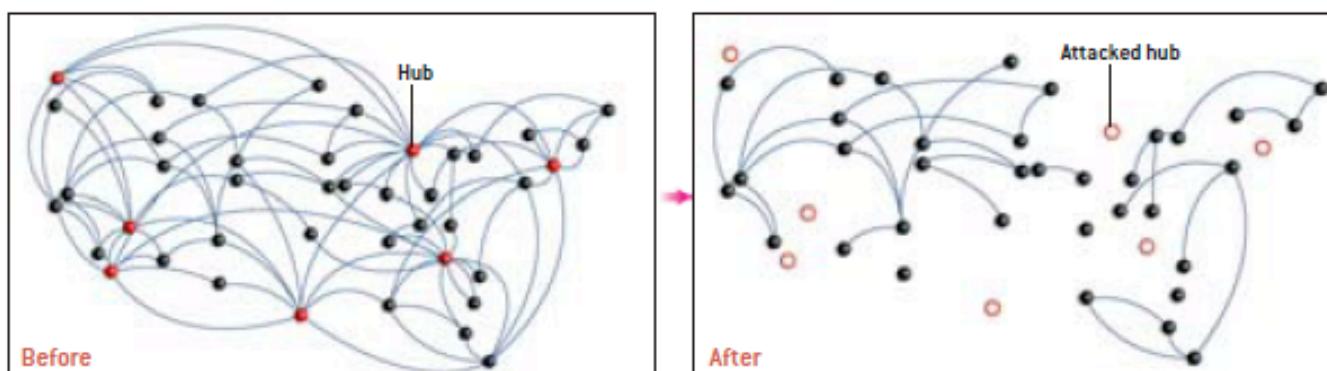
Random Network, Accidental Node Failure



Scale-Free Network, Accidental Node Failure

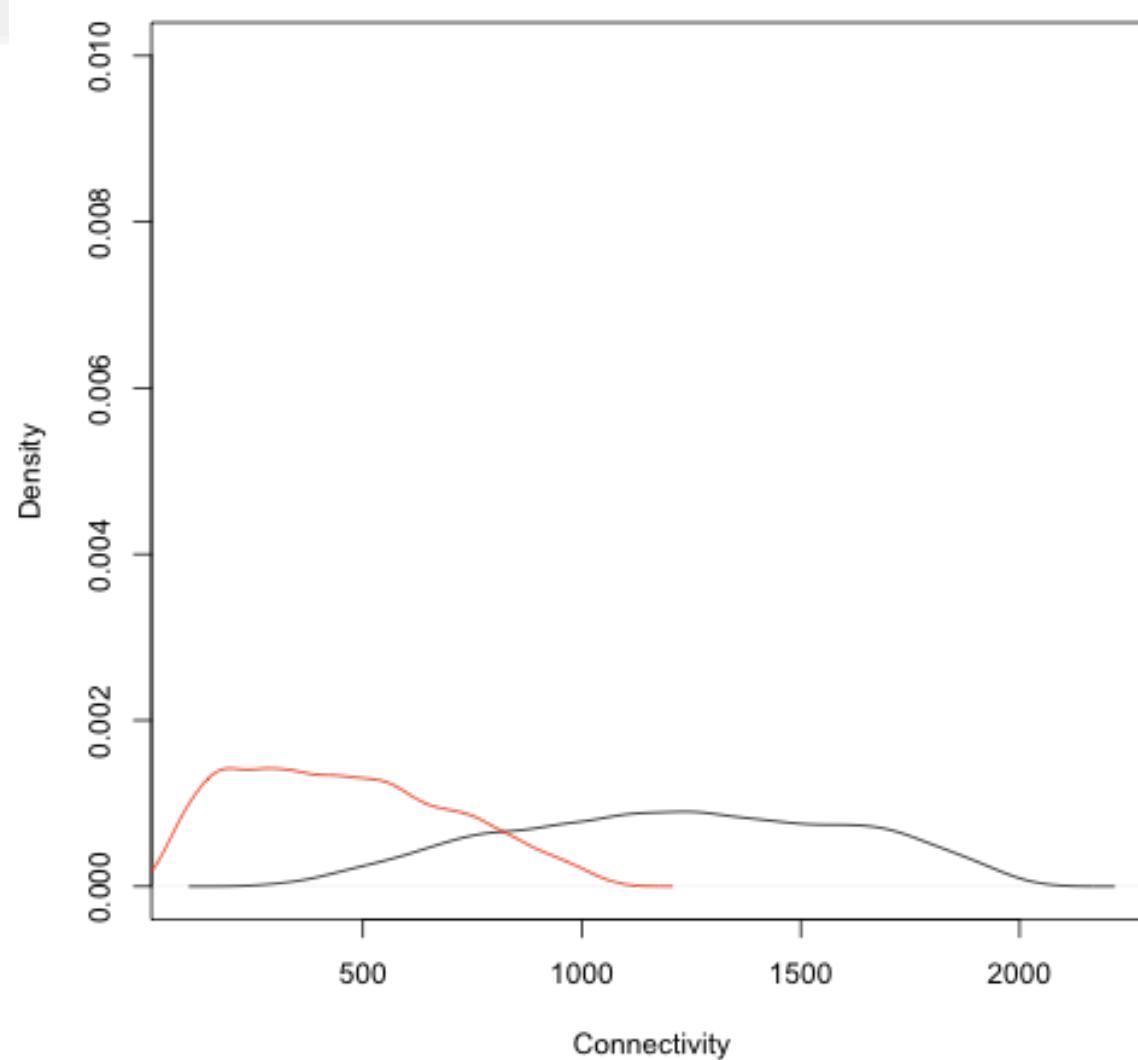


Scale-Free Network, Attack on Hubs



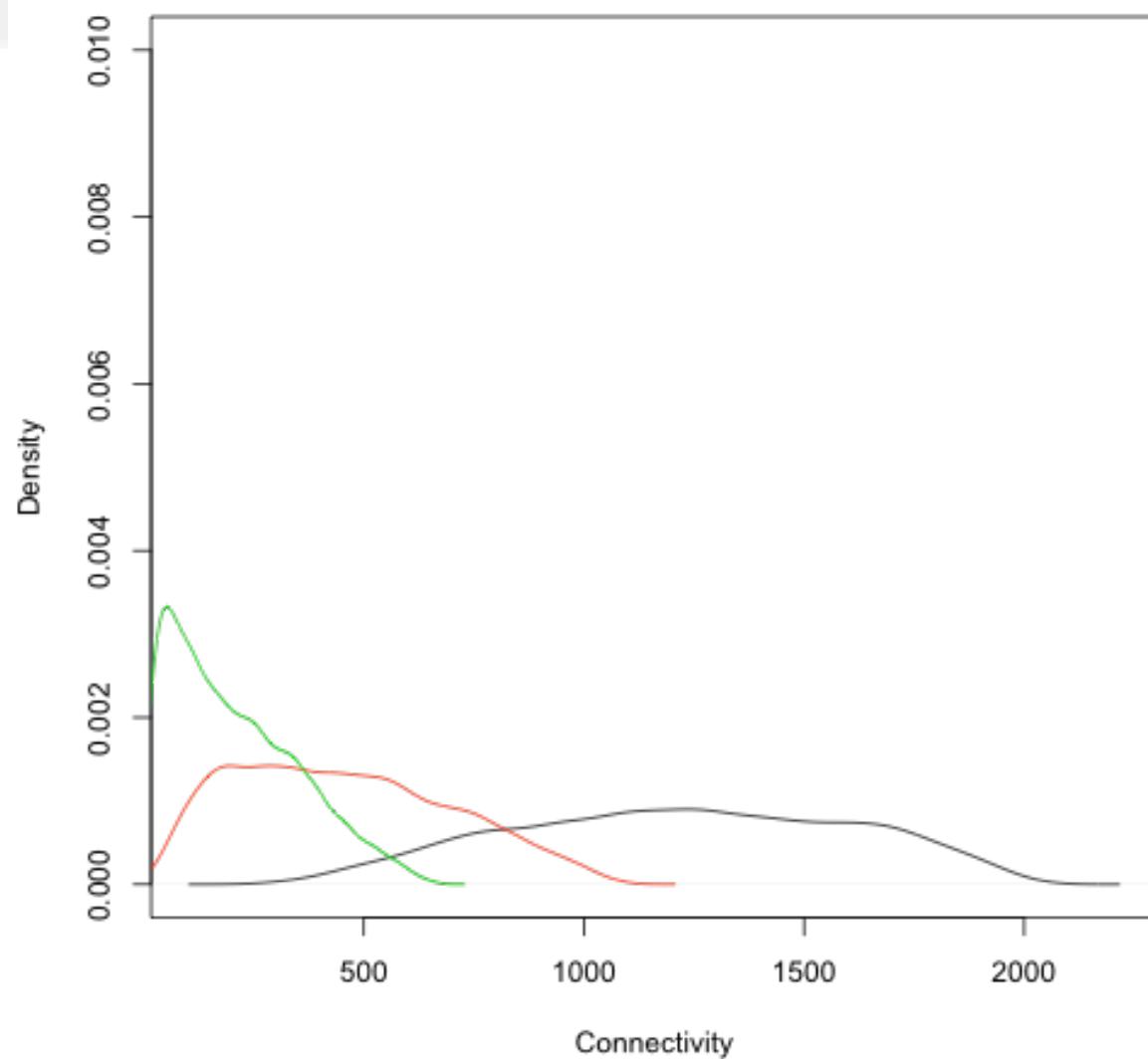
Scale-Free Network

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



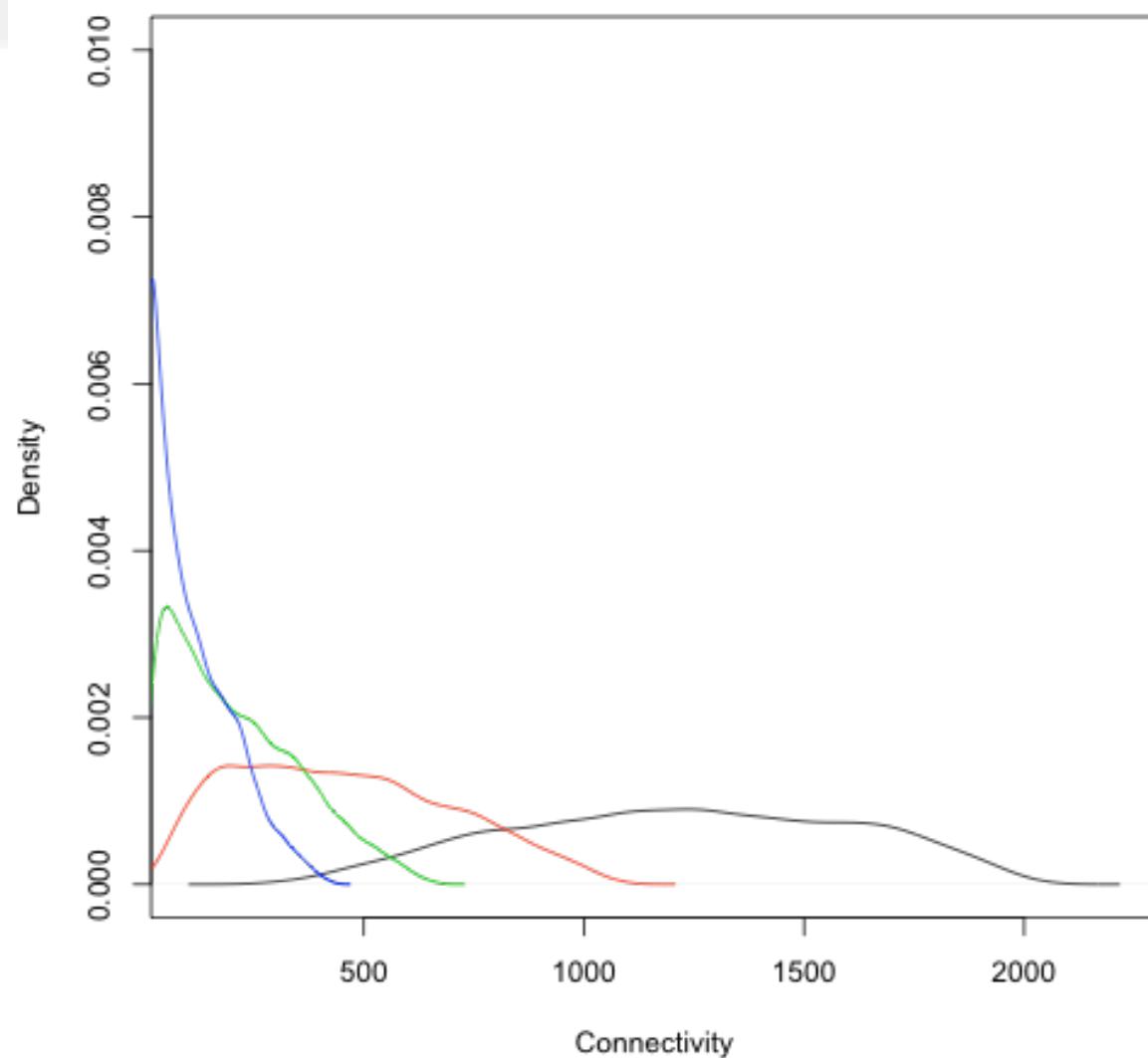
Scale-Free Network

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



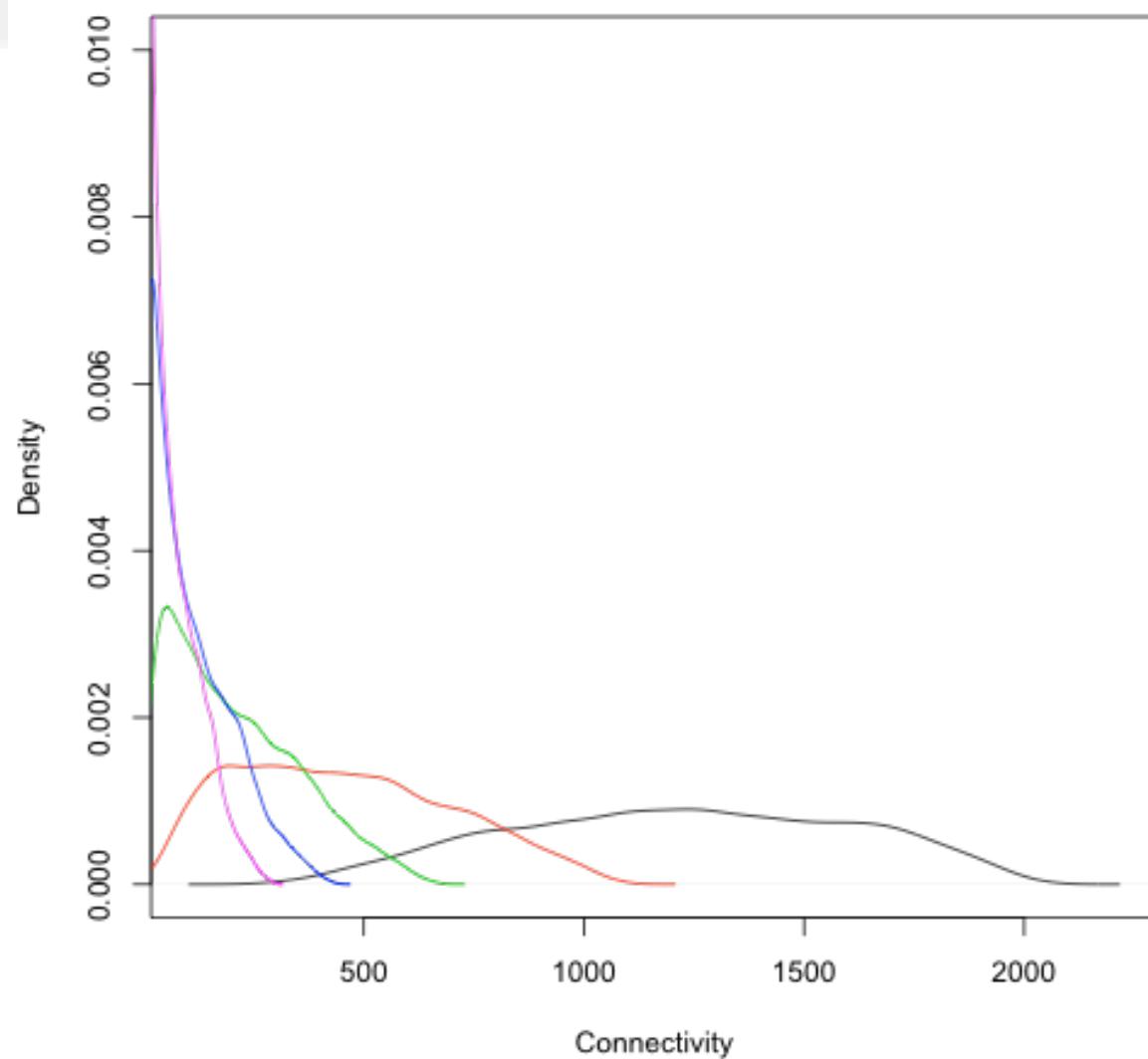
Scale-Free Network

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



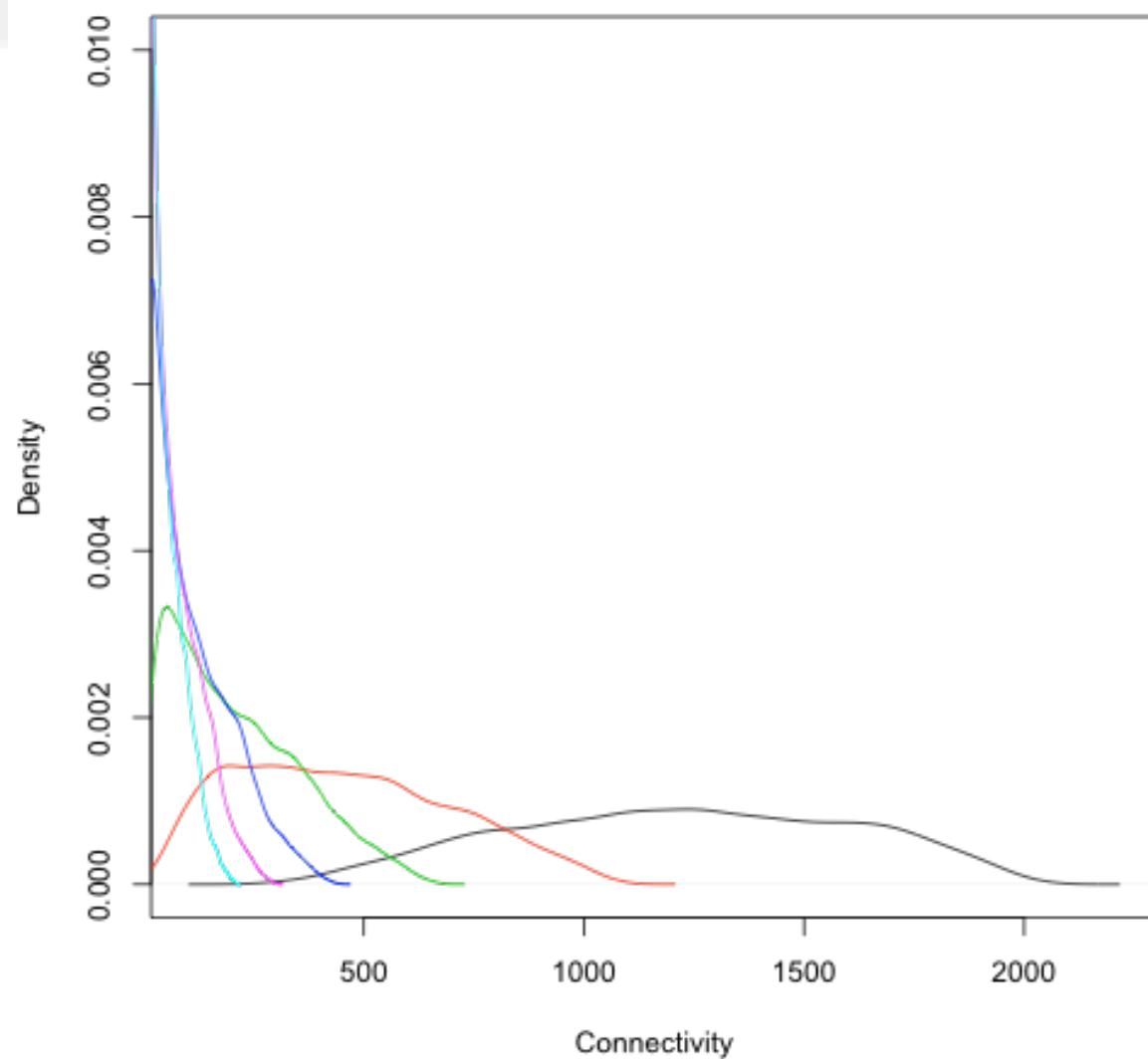
Scale-Free Network

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



Scale-Free Network

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



Topological overlap measure

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

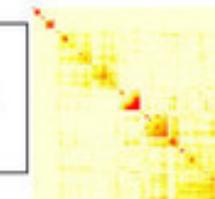
where $a_{ij} = |\rho_{ij}|^\beta$ is the connectivity of transcript i and transcript j

Weighted Gene Co-Expression Analysis

Construct a gene co-expression network

Rationale: make use of interaction patterns among genes

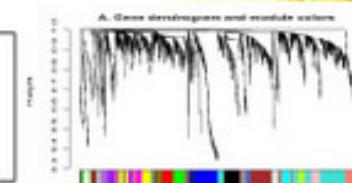
Tools: correlation as a measure of co-expression



Identify modules

Rationale: module (pathway) based analysis

Tools: hierarchical clustering, Dynamic Tree Cut

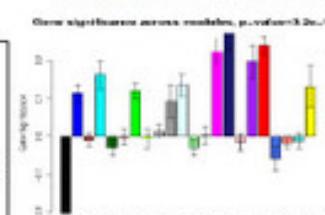


Relate modules to external information

Array Information: clinical data, SNPs, proteomics

Gene Information: ontology, functional enrichment

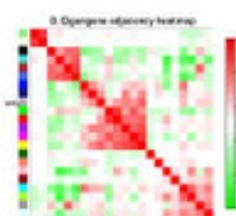
Rationale: find biologically interesting modules



Study module relationships

Rationale: biological data reduction, systems-level view

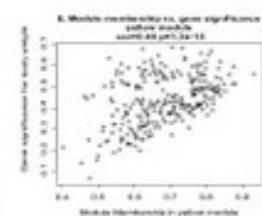
Tools: Eigengene Networks



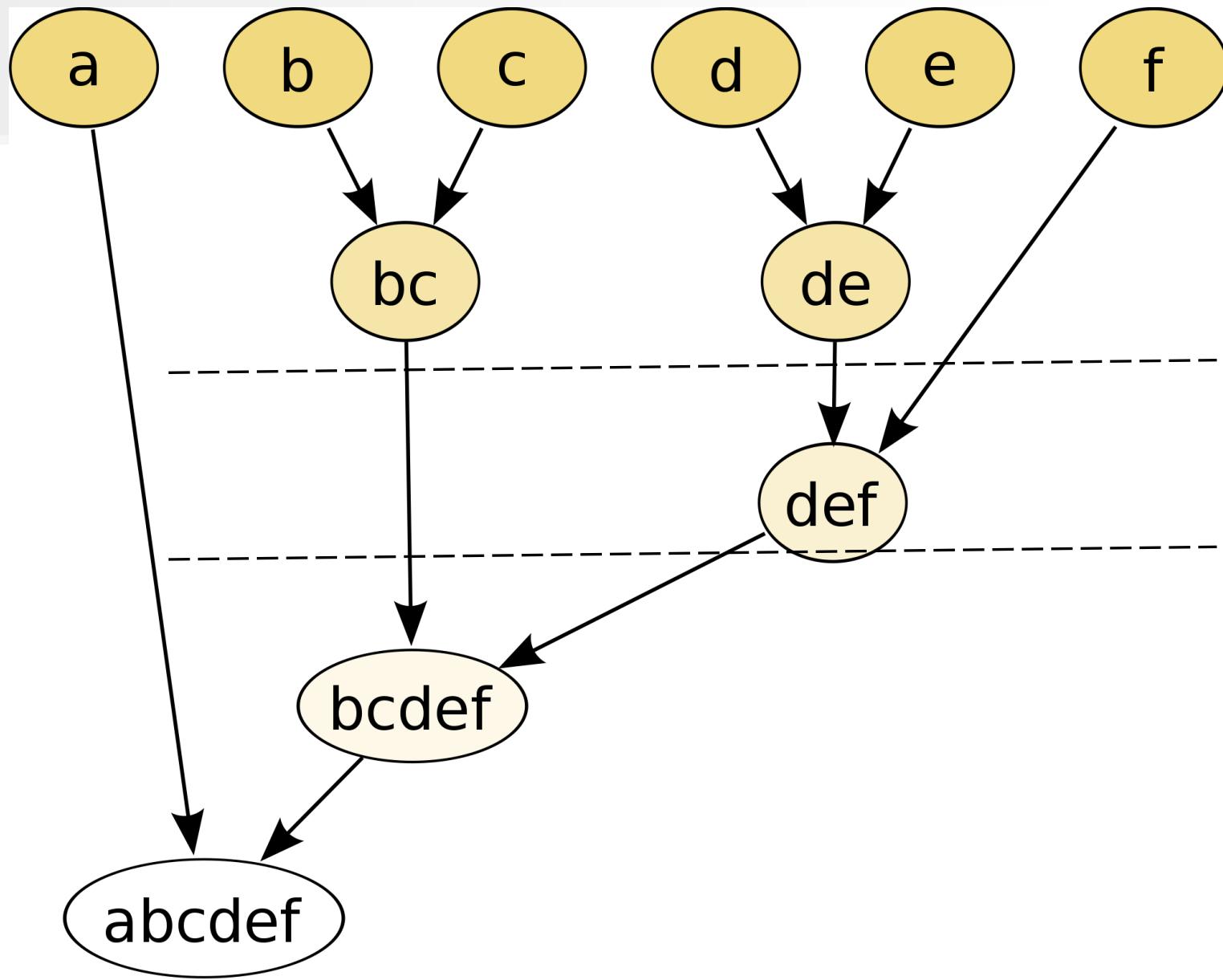
Find the key drivers in *interesting* modules

Rationale: experimental validation, biomarkers

Tools: intramodular connectivity, causality testing

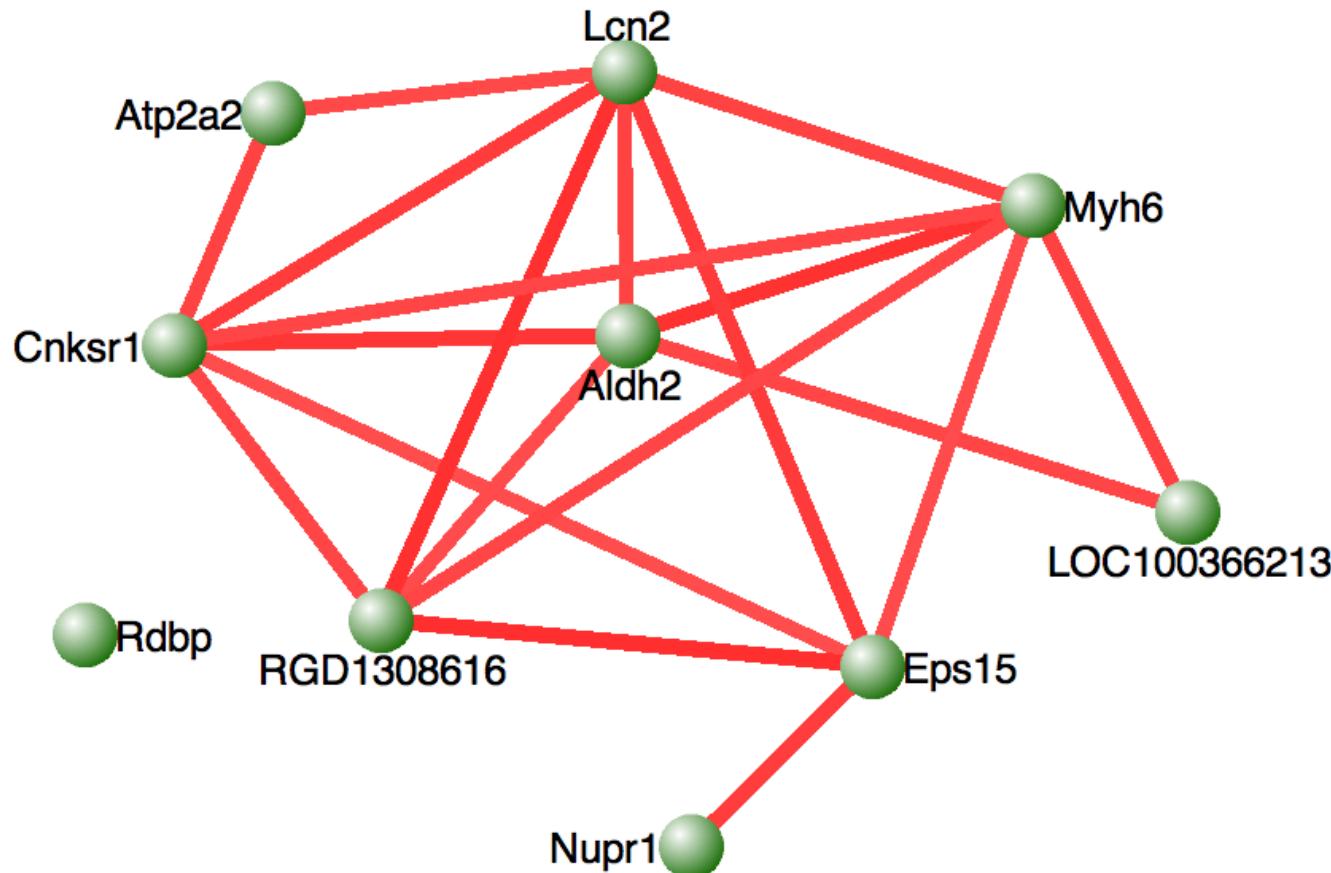


Hierarchical Clustering



Example - Aldh2 Module

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

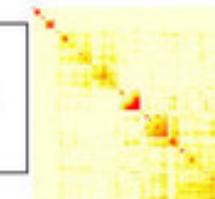


Weighted Gene Co-Expression Analysis

Construct a gene co-expression network

Rationale: make use of interaction patterns among genes

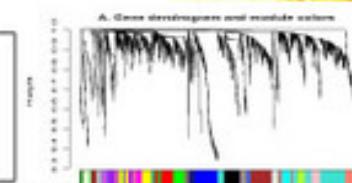
Tools: correlation as a measure of co-expression



Identify modules

Rationale: module (pathway) based analysis

Tools: hierarchical clustering, Dynamic Tree Cut

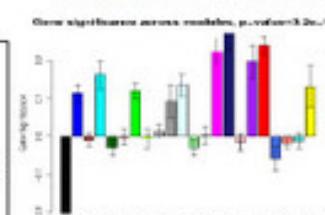


Relate modules to external information

Array Information: clinical data, SNPs, proteomics

Gene Information: ontology, functional enrichment

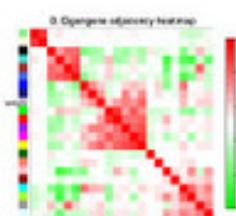
Rationale: find biologically interesting modules



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

