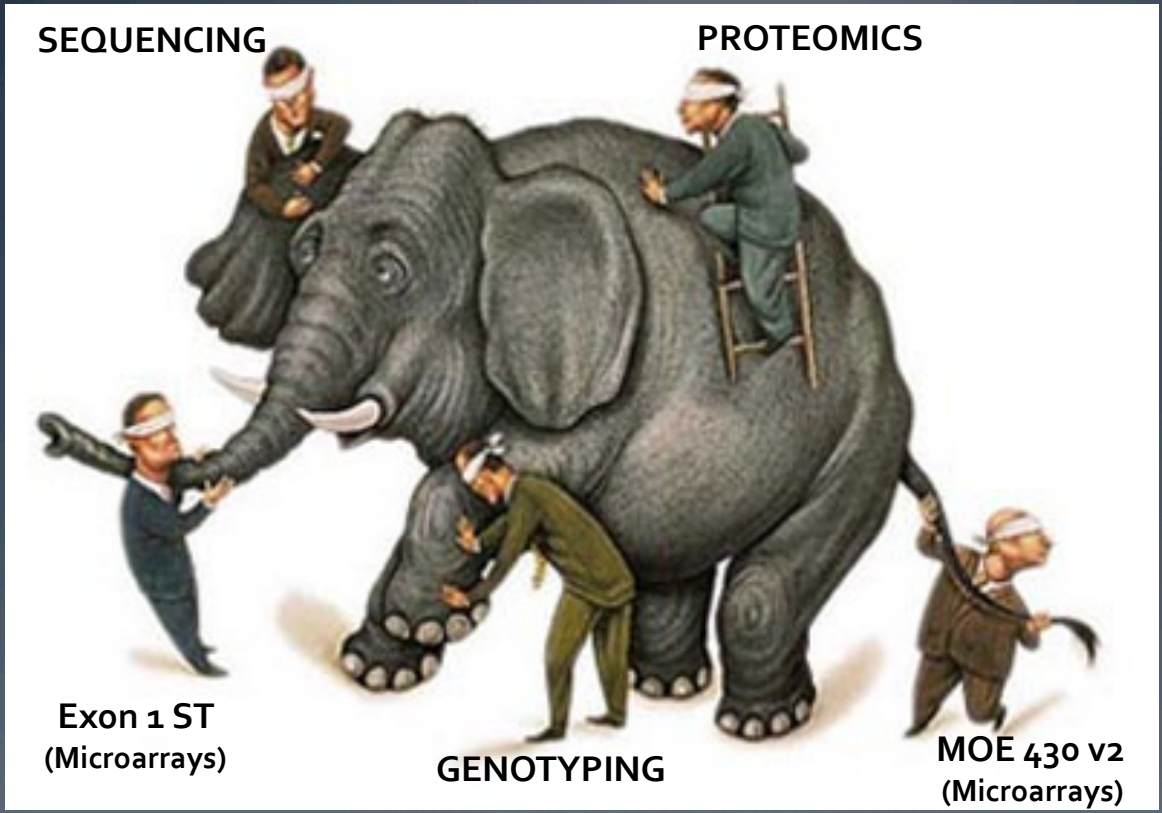
Bringing systems genetics to your RNA expression

Laura Saba, PhD
Research Assistant Professor
Department of Pharmaceutical Sciences
Skaggs School of Pharmacy and Pharmaceutical Sciences
University of Colorado, Anschutz Medical Campus
Laura.Saba@ucdenver.edu
GitHub: https://github.com/LauraSaba and https://github.com/TabakoffLab

CANDIDATE GENE SEARCH



...
And so these men of Indostan
Disputed loud and long,
Each in his own opinion
Exceeding stiff and strong.
Though each was partly in the right
They all were in the wrong!

-- John Godfrey Saxe (1816-1887)

...
And so these men of OMICstan
Disputed loud with might
Each in his own opinion
Exceeding stiff and bright.
Though each was partly in the wrong
They all "together" were in the RIGHT*!*

-- SB, PH & BT 2010

Objectives

- MAIN – Convince you of the value of systems genetics for studying the interactions between genetics and complex traits

1. ***De novo* hypothesis generating** - Demonstrate how whole genomic/whole transcriptome information can be combined with a phenotype to identify relevant genetic pathways, not just genes, that can uncover biological mechanisms in addition to diagnostic markers

2. **Guided hypothesis generating - Illustrate basic genomic/transcriptome information and basal (untreated) genetic networks related to genes typically studied for genetic risk of your complex trait**

Complex Traits

http://en.wikipedia.org/wiki/Eye_color

http://www.medicinethink.com/personal-genomics-why-23me-doesnt-work/

"What nature hath joined together, multiple regression cannot put asunder"

-- Richard Nisbett (via twitter @StatFact)

Why Study the RNA Dimension

Transcriptome links DNA and complex traits/diseases

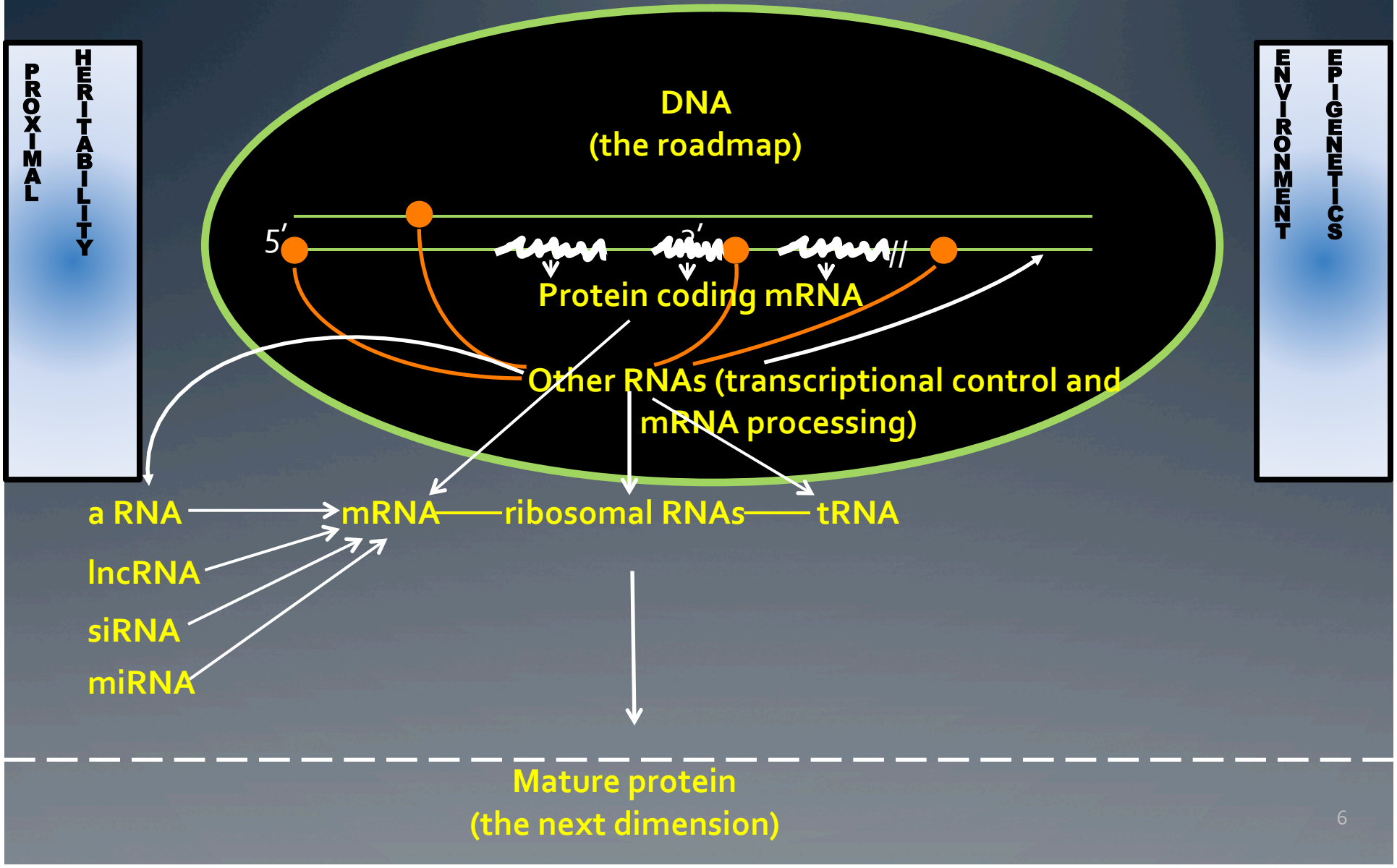
A. First quantitative link between DNA sequence and phenotype.

B. GWAS Gap: how does identified polymorphic locus contribute to disease?

C. First step where DNA sequence and environment interact.

D. Implementation of network analyses at the transcript level provides insight into genetic interactions that are the basis for susceptibility to complex diseases.

The RNA Dimension (the true intermediate phenotype)



Recombinant Inbred Rodent Panels And Why We Love Them

Difference Between Extremes

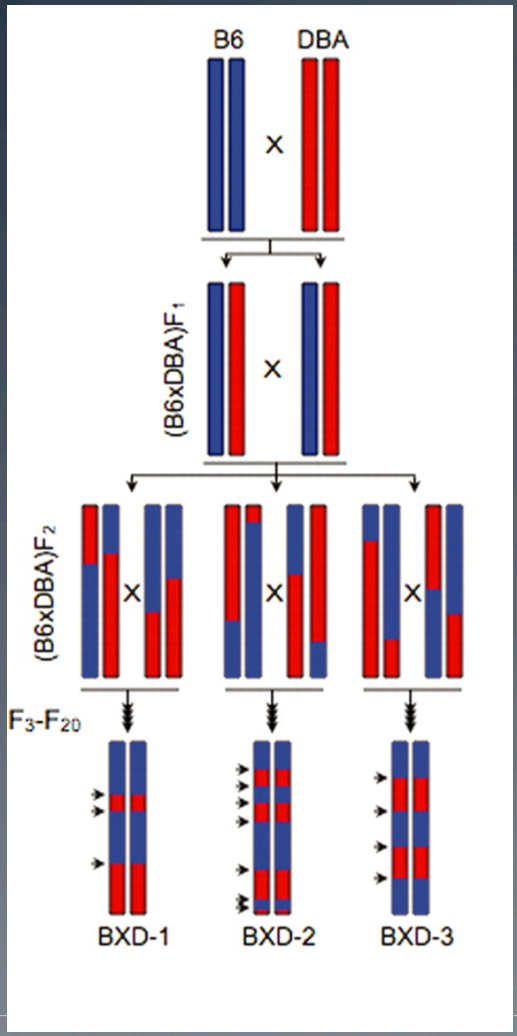
Alcohol Consumption…

- gender
- ethnicity
- hair color
- size
- anxiety
- depression
- logic…

What if Lohan and Tebow a baby?

What if their babies had babies?

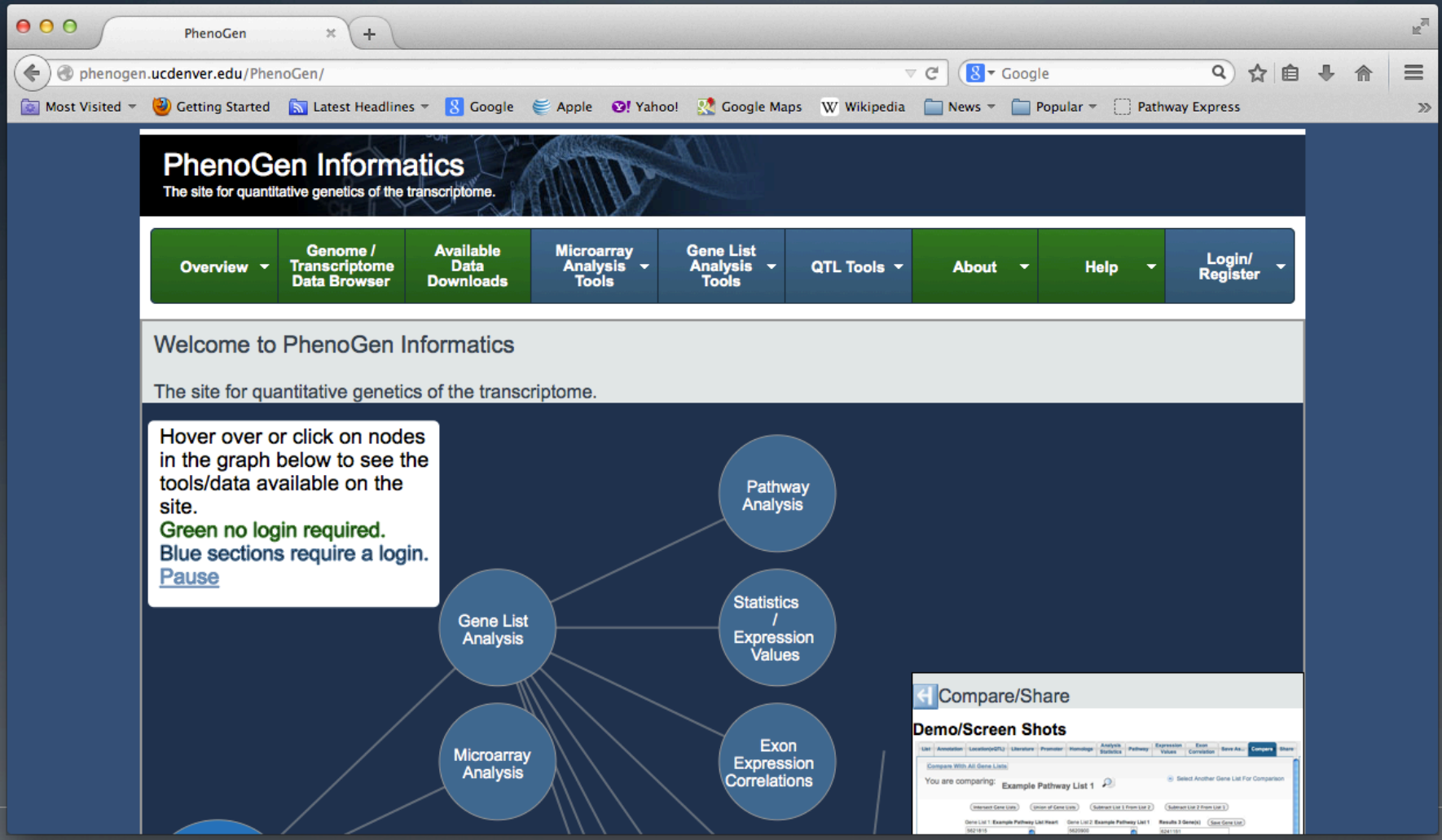
Recombinant Inbred Rodent Panel



- Genetic identity is retained over generations
- Cumulative genetic and phenotype data across labs
- Ideal genetic controls for studying interventions/ environmental effects

Wu Z K et al. Nucl. Acids Res. 2010;nar.gkp1251

http://phenogen.ucdenver.edu



PhenoGen Database

Mouse Rat Home

Mus Musculus

Inbred Strains

Number of Strains	Type of Samples	Number of Arrays Per Strain	Array	
28	whole brain from adult male mice	4 to 6	Affymetrix Mouse 430 version 2 (targeted to 3' UTR)	129P3/J, 129S1/SvImJ, 129X C3H/HeJ, C57BL/6J, C58/J, KK/H1J, LP/J, Molf/EiJ, NOD SJL/J, SPRET/EiJ, WSB/EiJ
4	whole brain from adult male mice	6 to 36	Affymetrix Mouse Exon Array	C57BL/6J, DBA/2J, ILS, ISS

Recombinant Inbred Panels

Number of Strains	Type of Samples	Number of Arrays Per Strain	Array	
32	whole brain from adult male mice	4 to 7	Affymetrix Mouse 430 version 2 (targeted to 3' UTR)	BXD1, BXD2, BXD5, BXD6, BXD15, BXD16, BXD18, BXD BXD28, BXD29, BXD31, BXD BXD40, BXD42, DBA, C57
61	whole brain from adult male mice	4 to 6	Affymetrix Mouse Exon Array	LXS100, LXS101, LXS102, LXS122, LXS123, LXS13, LX LXS26, LXS28, LXS3, LXS32 LXS46, LXS48, LXS5, LXS50 LXS72, LXS73, LXS75, LXS7 LXS89, LXS9, LXS90, LXS92 ILS, ISS

Selected Lines

Number of Pairs	Type of Samples	Number of Arrays Per Line	Array	
7	whole brain from adult male mice	4 to 6	Affymetrix Mouse 430 version 2 (targeted to 3' UTR)	HAFT1/LAFT1, HAFT2/LAFT iHAFT/iLAFT, ILS/ISS

Genetically Modified Animals

Number of Pairs	Type of Samples	Number of Arrays Per Group	Array	
8	multiple	3 to 6	Affymetrix Mouse 430 version 2 (targeted to 3' UTR)	Fyn KO, Maob KO, Maoa KO Spinophilin KO, AC7 transge

6	brown fat from adult male rats	4	Affymetrix Rat Exon Array	WKY/NCrlPrin
4	Kupffer cells from liver (fresh/cultured /MACS)	4	Affymetrix Rat Exon Array	BN-Lx/CubPrin, PD/CubPrin, SHR/NCrlPrin, SHR/OlaPrin
4	Hepatic Stellate cells from liver (fresh)	4	Affymetrix Rat Exon Array	BN-Lx/CubPrin, PD/CubPrin, SHR/NCrlPrin, SHR/OlaPrin
4	hepatocyte cells from liver (fresh/cultured)	4	Affymetrix Rat Exon Array	BN-Lx/CubPrin, PD/CubPrin, SHR/NCrlPrin, SHR/OlaPrin
4	sinusoidal edothelial cells from liver (fresh/cultured)	4	Affymetrix Rat Exon Array	BN-Lx/CubPrin, PD/CubPrin, SHR/NCrlPrin, SHR/OlaPrin

Recombinant Inbred Panels

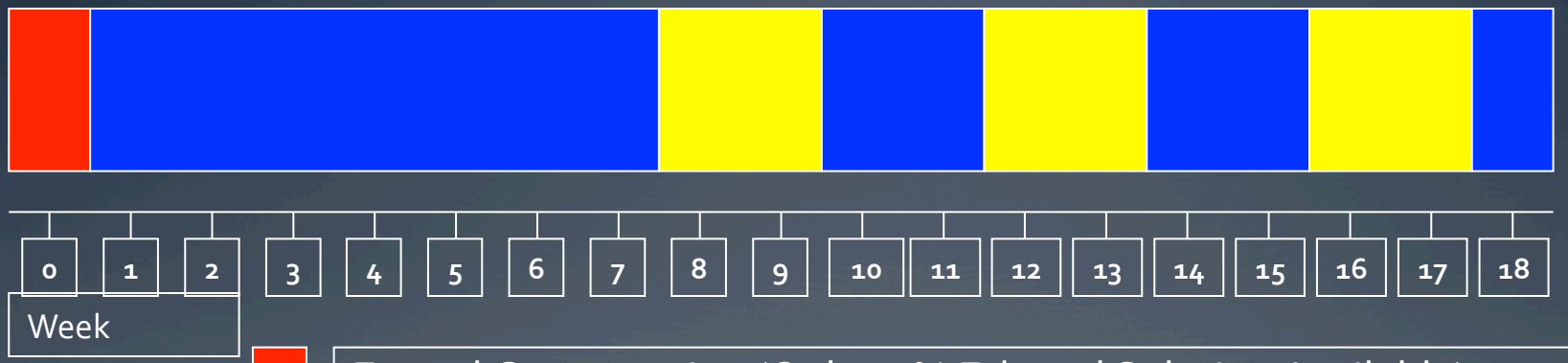
Number of Strains	Type of Samples	Number of Arrays Per Strain	Array	Strains
28	whole brain from adult male rats	4 to 6	CodeLink Whole Genome Rat Array	BN-Lx/CubPrin, BXH08, BXH10, BXH11, BXH12, BXH13, BXH5, BXH6, BXH8, BXH9, HXB1, HXB10, HXB13, HXB15, HXB17, HXB18, HXB2, HXB20, HXB23, HXB25, HXB26, HXB27, HXB29, HXB3, HXB31, HXB4, HXB7, SHR/OlaPrin
23	whole brain from adult male rats	3 to 4	Affymetrix Rat Exon Array	BN-Lx/CubPrin, BXH08, BXH10, BXH11, BXH13, BXH6, HXB1, HXB10, HXB13, HXB15, HXB17, HXB18, HXB2, HXB23, HXB25, HXB26, HXB27, HXB29, HXB3, HXB31, HXB4, HXB7, SHR/OlaPrin
23	heart - left ventrical from adult male rats	3 to 4	Affymetrix Rat Exon Array	BN-Lx/CubPrin, BXH08, BXH10, BXH11, BXH13, BXH6, HXB1, HXB10, HXB13, HXB15, HXB17, HXB18, HXB2, HXB23, HXB25, HXB26, HXB27, HXB29, HXB3, HXB31, HXB4, HXB7, SHR/OlaPrin
23	liver from adult male rats	3 to 4	Affymetrix Rat Exon Array	BN-Lx/CubPrin, BXH08, BXH10, BXH11, BXH13, BXH6, HXB1, HXB10, HXB13, HXB15, HXB17, HXB18, HXB2, HXB23, HXB25, HXB26, HXB27, HXB29, HXB3, HXB31, HXB4, HXB7, SHR/OlaPrin
21	brown fat from adult male rats	3 to 4	Affymetrix Rat Exon Array	BN-Lx/CubPrin, BXH08, BXH10, BXH11, BXH13, BXH6, HXB10, HXB13, HXB15, HXB17, HXB18, HXB2, HXB23, HXB25, HXB26, HXB27, HXB29, HXB3, HXB4, HXB7, SHR/OlaPrin

Selected Lines

Number of Pairs	Type of Samples	Number of Arrays Per Line	Array	Lines
8	whole brain from adult male rats	4 to 6	CodeLink Whole Genome Rat Array	AA/ANA, HAD1/LAD1, HAD2/LAD2, IHAS1/ILAS1, IHAS2/ILAS2, P/NP, sP/sNP, UChB/UChA
6	whole brain from adult male rats	4 to 6	Affymetrix Rat Exon Array	AA/ANA, HAD1/LAD1, HAD2/LAD2, P/NP, sP/sNP, UChB/UChA

Genetic Pathways Associated with Alcohol Consumption

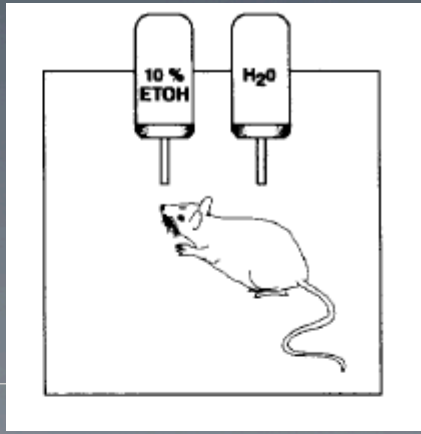
Alcohol Preference Procedure



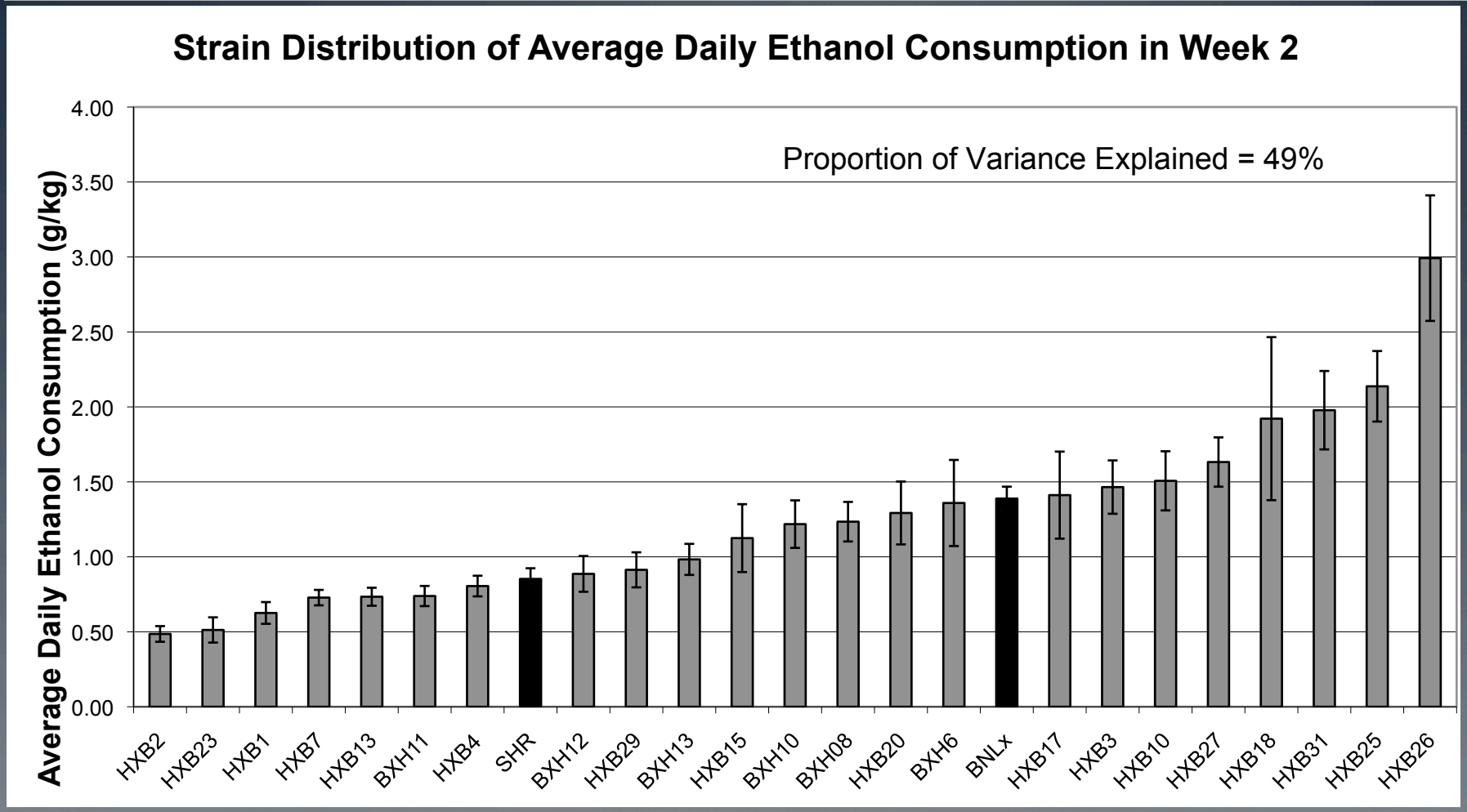
Rats

21 HXB/BXH Strains and 2 Progenitor Strains

223 Male Rats



Distribution of alcohol consumption in two bottle choice



Copied From Tabakoff B, Saba L et al 2009. BMC Biol. 7:170

Genetical Genomics/Phenomics Approach

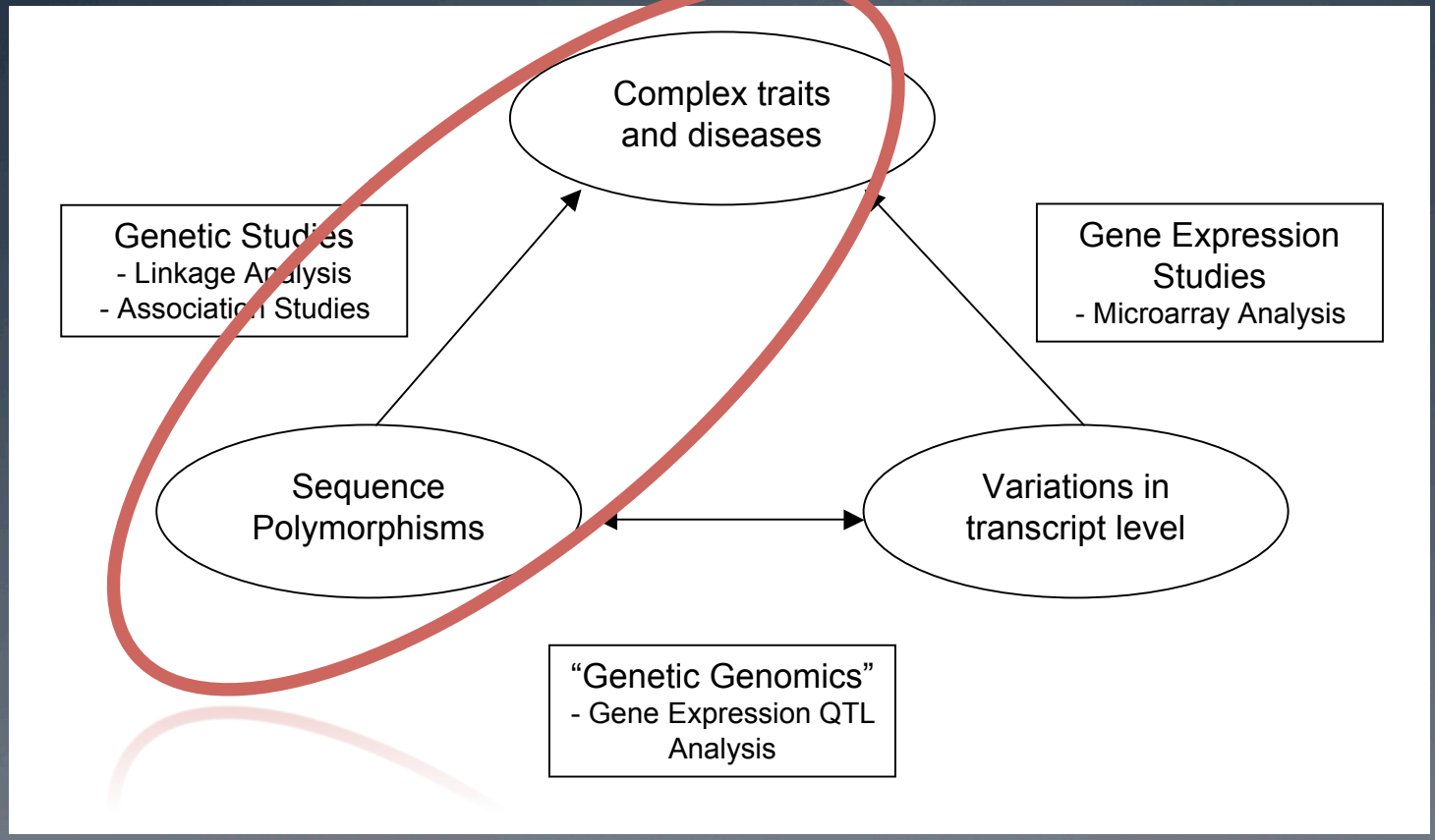
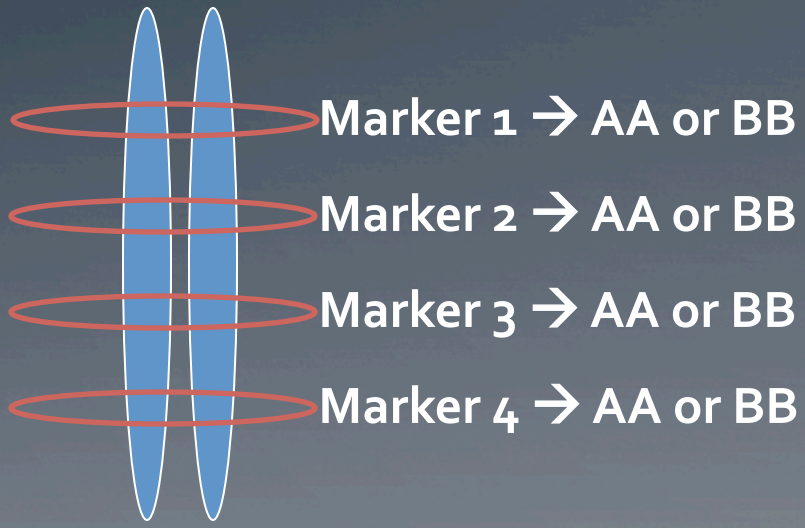


Image copied from "The Marriage of Phenomics and Genetical Genomics: A Systems Approach to Complex Trait Analysis" in Systems Biology in Psychiatric Research: From High-Throughput Data to Mathematical Modeling, edited by Tetter F, Winterer G, Gebicke-Haerter PG, and Mendoza E. Wiley-VCH 2010.

Quantitative Trait Loci

- Definition – area of the genome where polymorphisms are associated with a quantitative trait

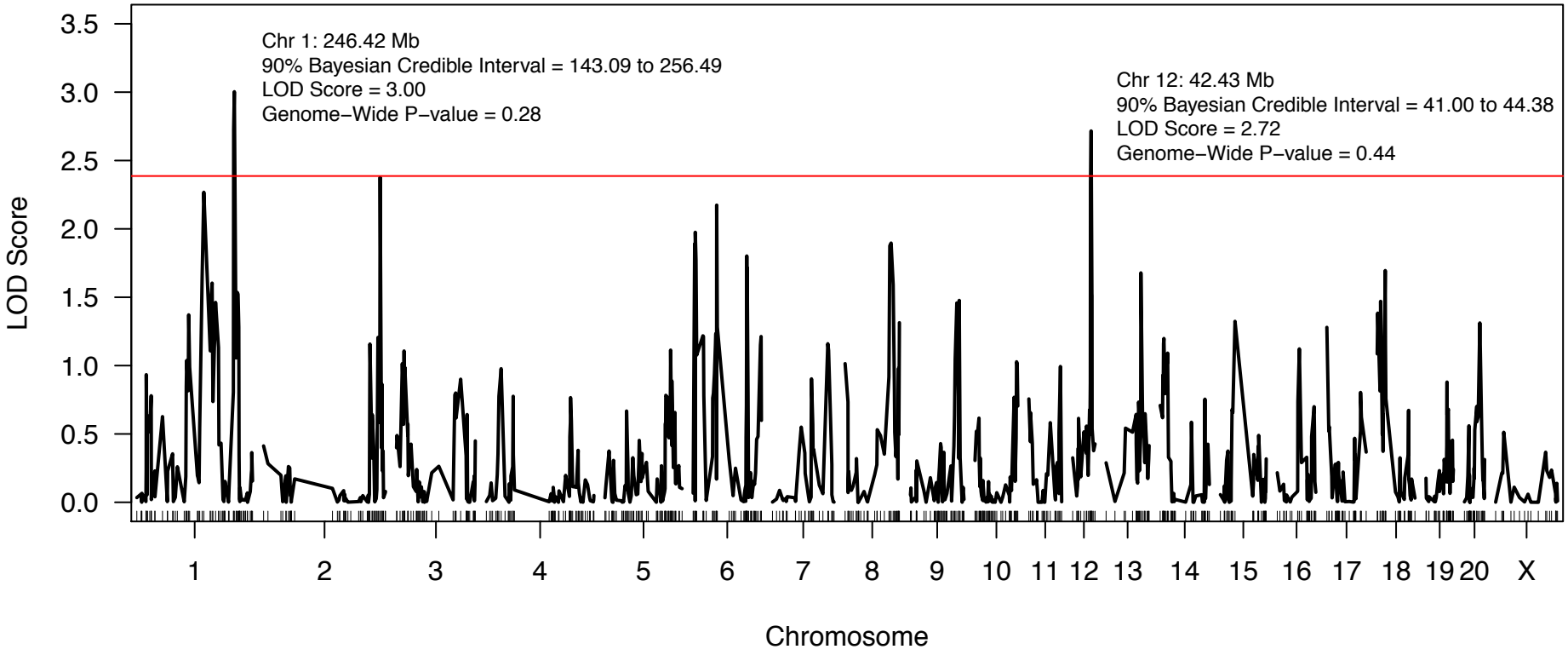


At each marker:

1. **Split population into two/three groups based on genotype**
2. **Compare mean values for the quantitative trait (phenotype) between the two groups**
3. **LARGE Statistically Significant Differences → QTL**

Finding Candidate DNA Polymorphism

QTL Analysis for Alcohol Consumption



Genetical Genomics/Phenomics Approach

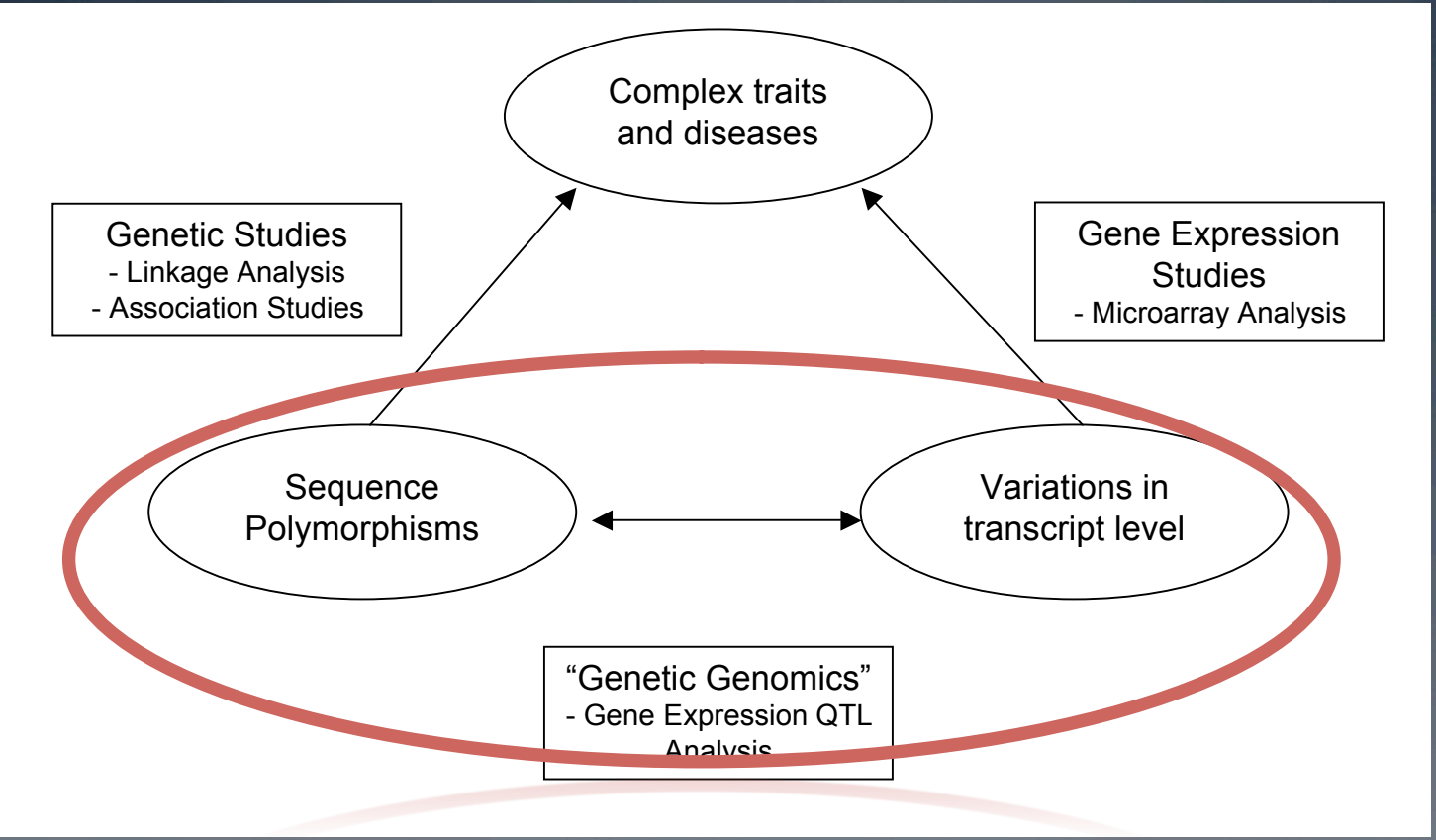
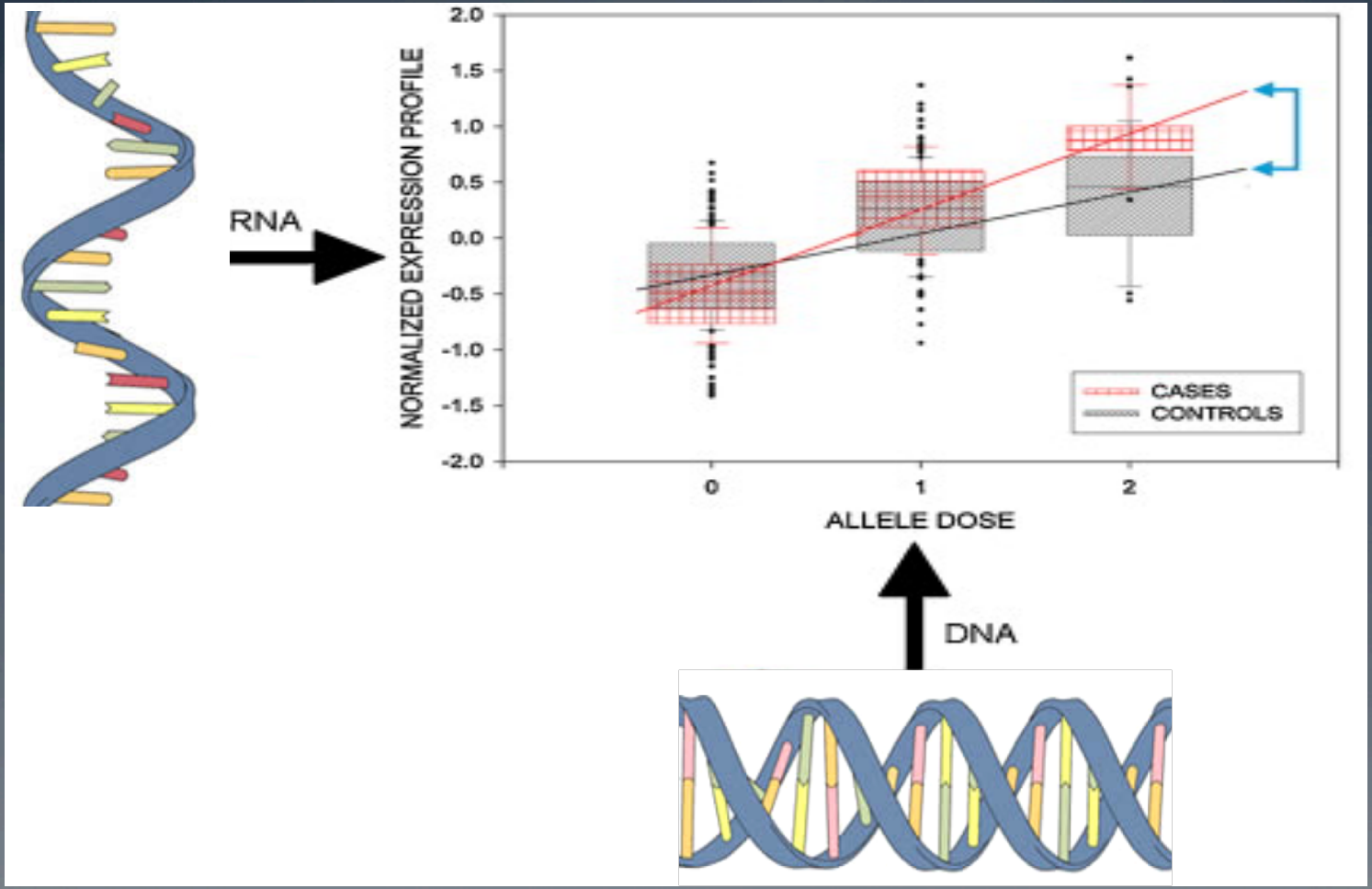


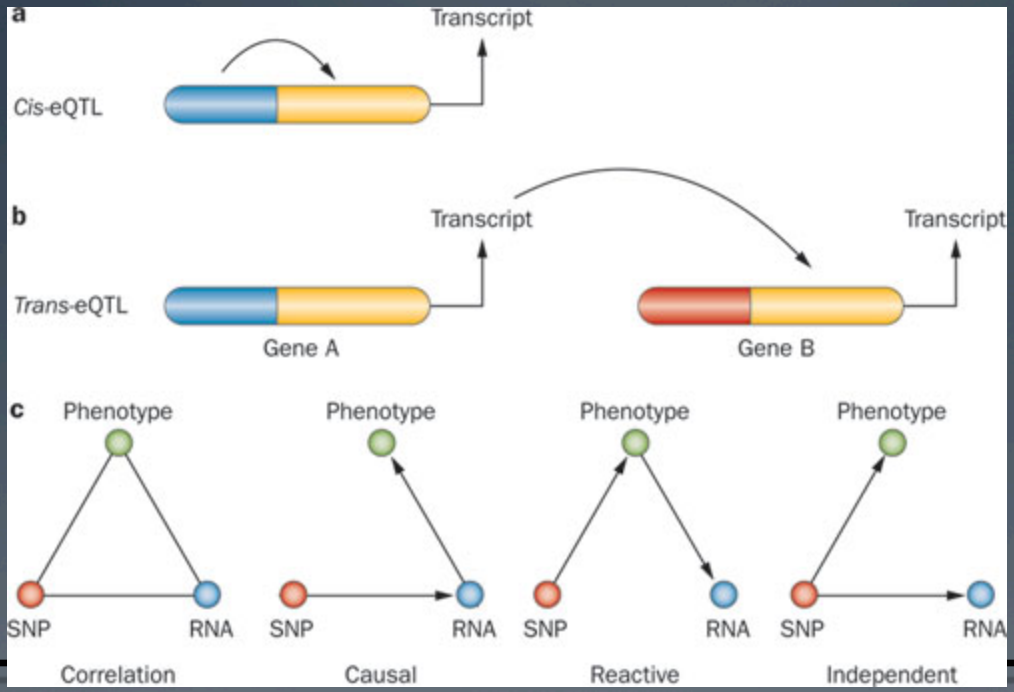
Image copied from "The Marriage of Phenomics and Genetical Genomics: A Systems Approach to Complex Trait Analysis" in Systems Biology in Psychiatric Research: From High-Throughput Data to Mathematical Modeling, edited by Tetter F, Winterer G, Gebicke-Haerter PG, and Mendoza E. Wiley-VCH 2010.

eQTL Definition



Myers, AJ. The age of the "ome": Genome, transcriptome and proteome data set collection and analysis. Brain Research Bulletin Volume 88, Issue 4 2012 294 - 301

Figure 2 Principles of eQTL analysis



MacLellan, W. R. *et al.* (2012) Systems-based approaches to cardiovascular disease
Nat. Rev. Cardiol. doi:10.1038/nrcardio.2011.208

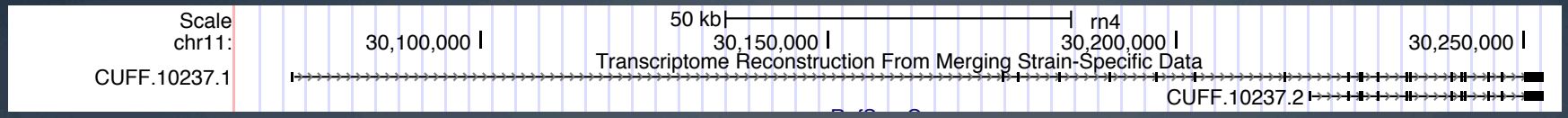
nature REVIEWS CARDIOLOGY

RNA Expression Estimates

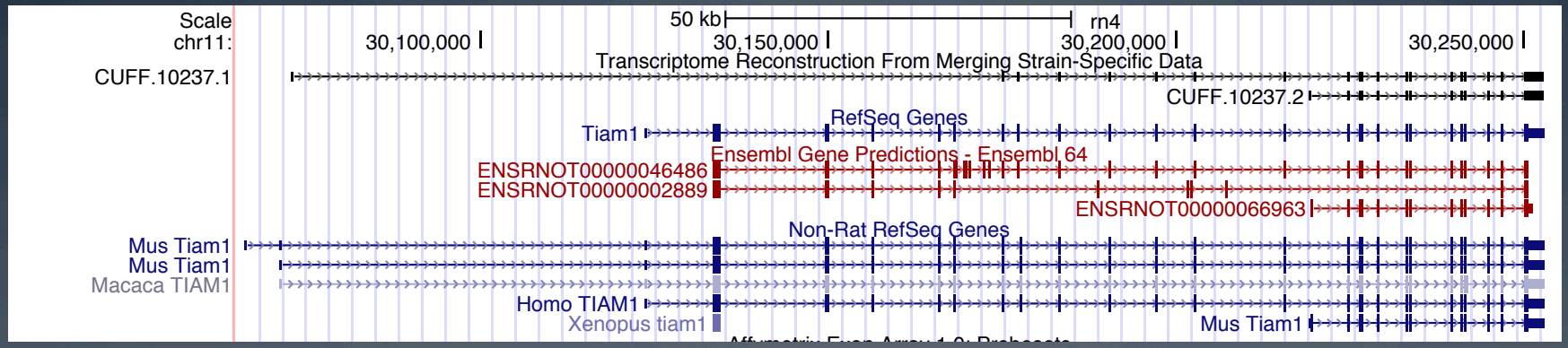
DNA/RNA-Seq Guided Microarray Analysis

- How good are the probes?
  - DNA-Seq of BN-Lx and SHR inbred strains
- What gene/transcript do the probes represent?
  - RNA-Seq of BN-Lx and SHR inbred strains

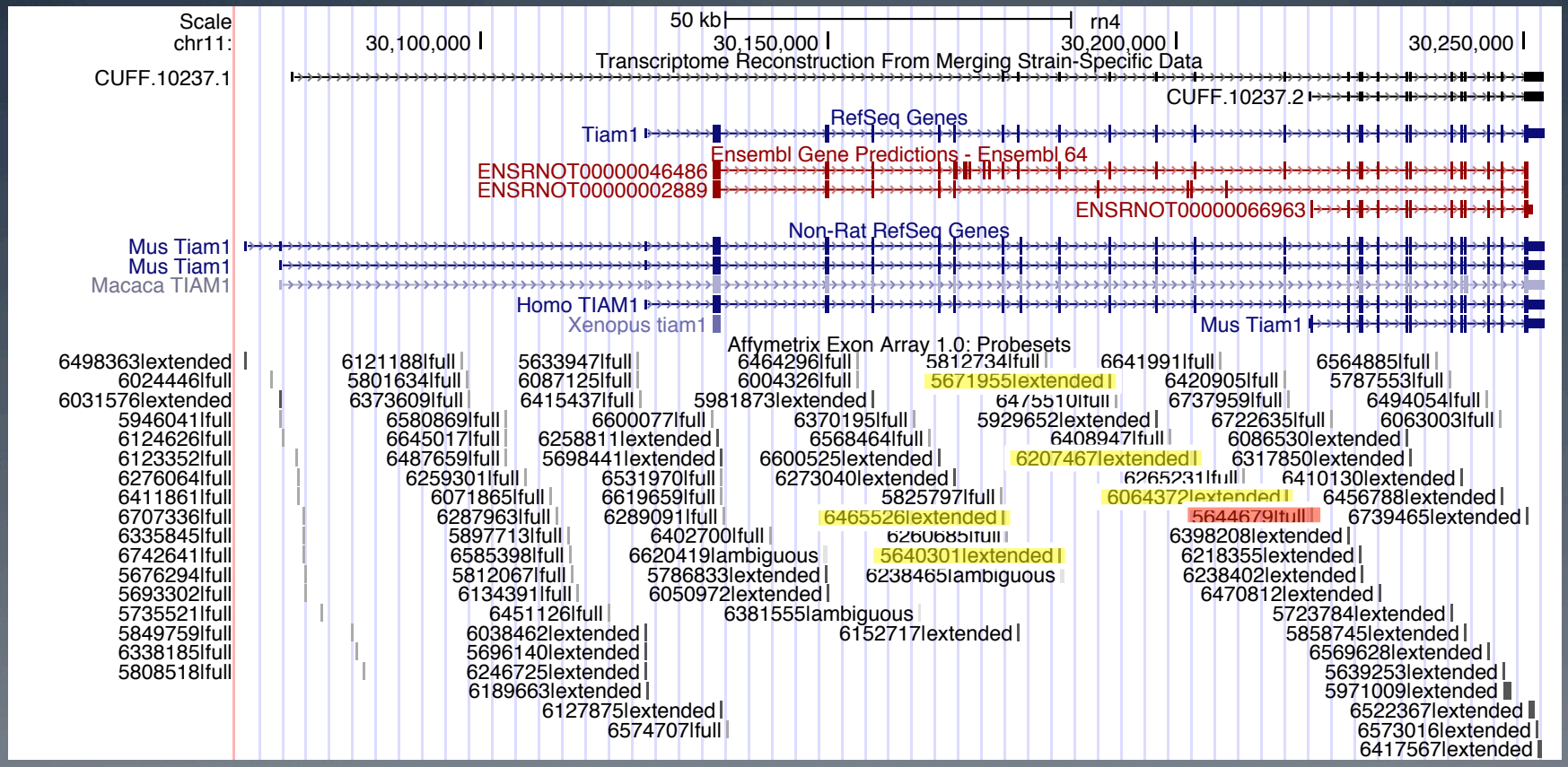
Transcriptome Reconstruction

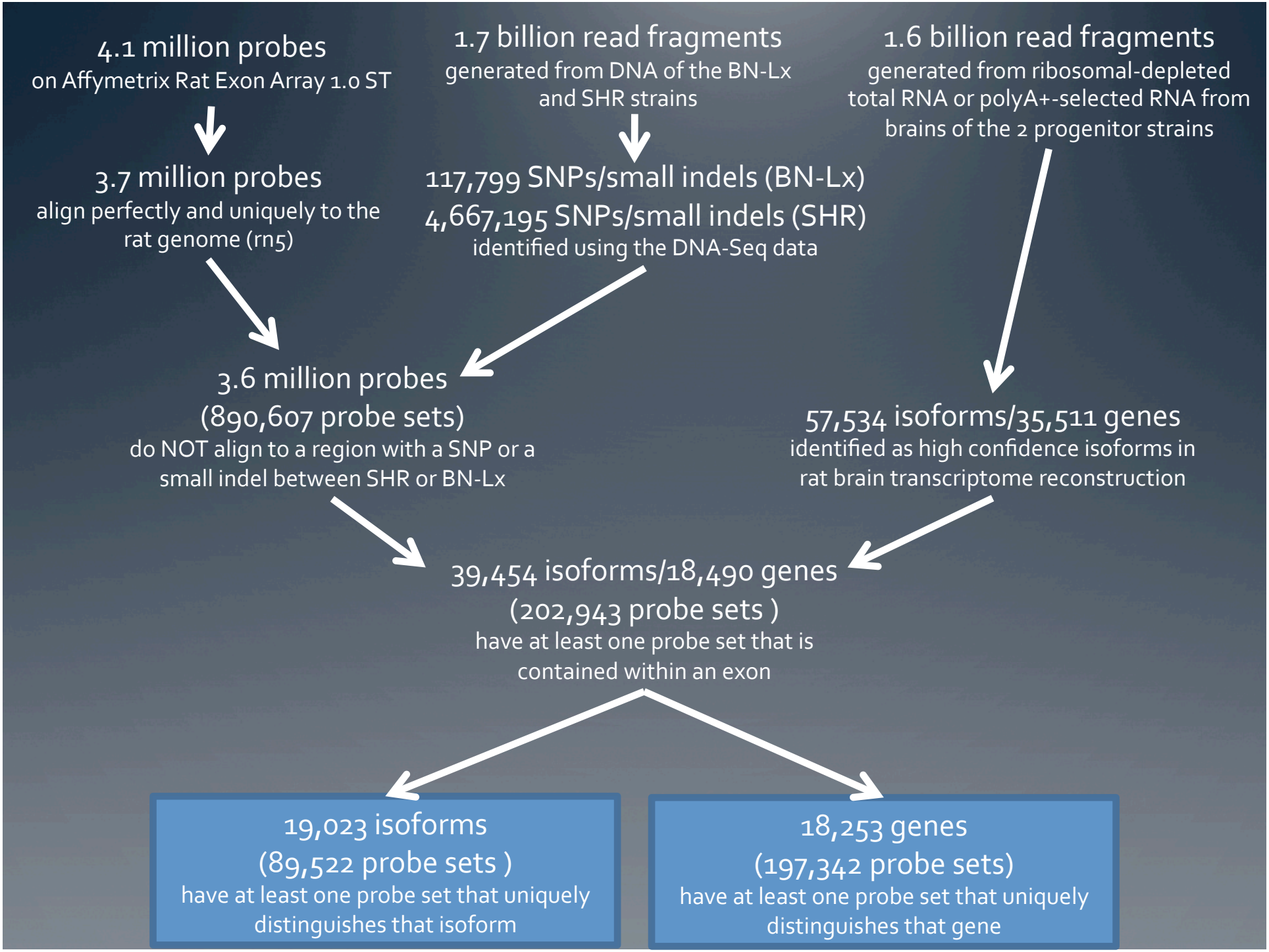


Transcriptome Reconstruction



Transcriptome Reconstruction



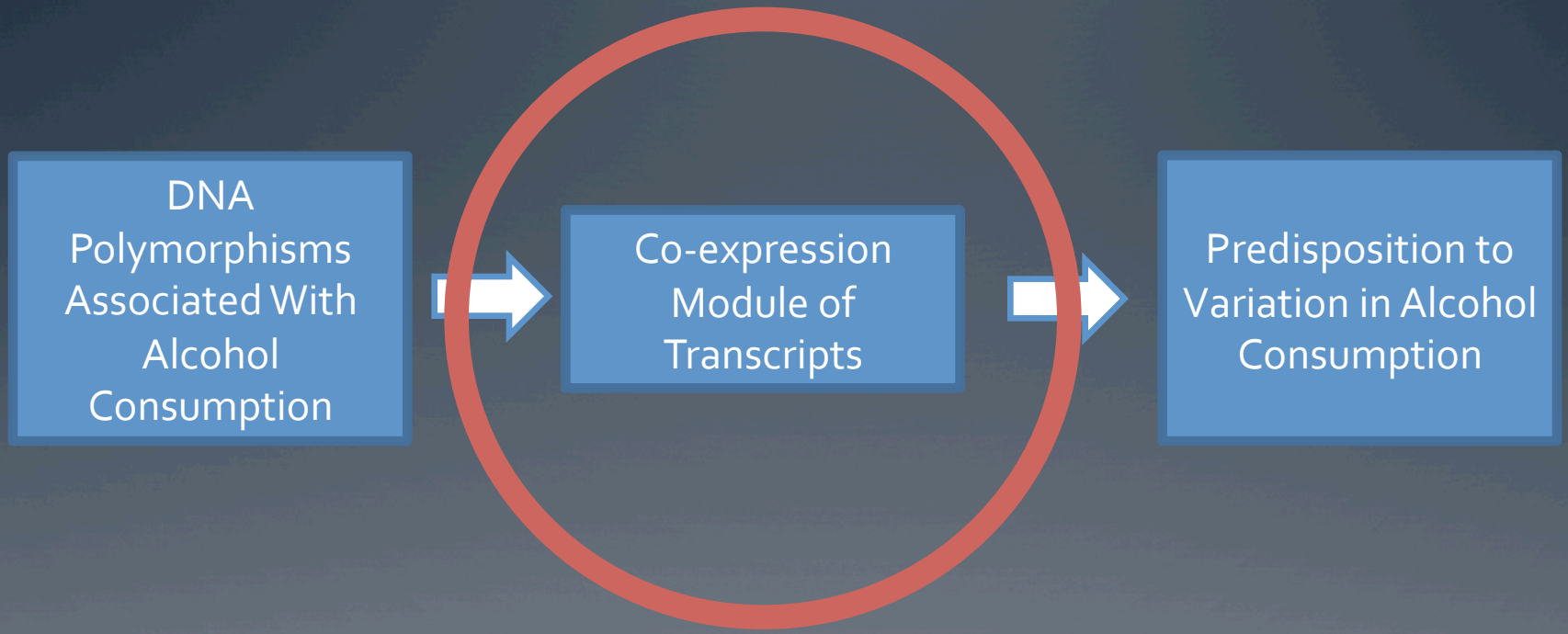

4.1 million probes
on Affymetrix Rat Exon Array 1.0 ST
1.7 billion read fragments
generated from DNA of the BN-Lx and SHR strains
1.6 billion read fragments
generated from ribosomal-depleted total RNA or polyA+-selected RNA from brains of the 2 progenitor strains
3.7 million probes
align perfectly and uniquely to the rat genome (rn5)
117,799 SNPs/small indels (BN-Lx)
4,667,195 SNPs/small indels (SHR)
identified using the DNA-Seq data
3.6 million probes
(890,607 probe sets)
do NOT align to a region with a SNP or a small indel between SHR or BN-Lx
57,534 isoforms/35,511 genes
identified as high confidence isoforms in rat brain transcriptome reconstruction
39,454 isoforms/18,490 genes
(202,943 probe sets)
have at least one probe set that is contained within an exon
19,023 isoforms
(89,522 probe sets)
have at least one probe set that uniquely distinguishes that isoform
18,253 genes
(197,342 probe sets)
have at least one probe set that uniquely distinguishes that gene

Co-expression Modules

Why Pathways and Not Individual Genes?

- Complex polygenic trait ≠ one gene
- Pathways add biological context to unannotated or under annotated genes.
- We hypothesize that different perturbations, i.e., changes in the expression of different genes, of the same genetic network related to voluntary alcohol consumption can produce similar phenotypic outcomes.

Transcriptional Pathway



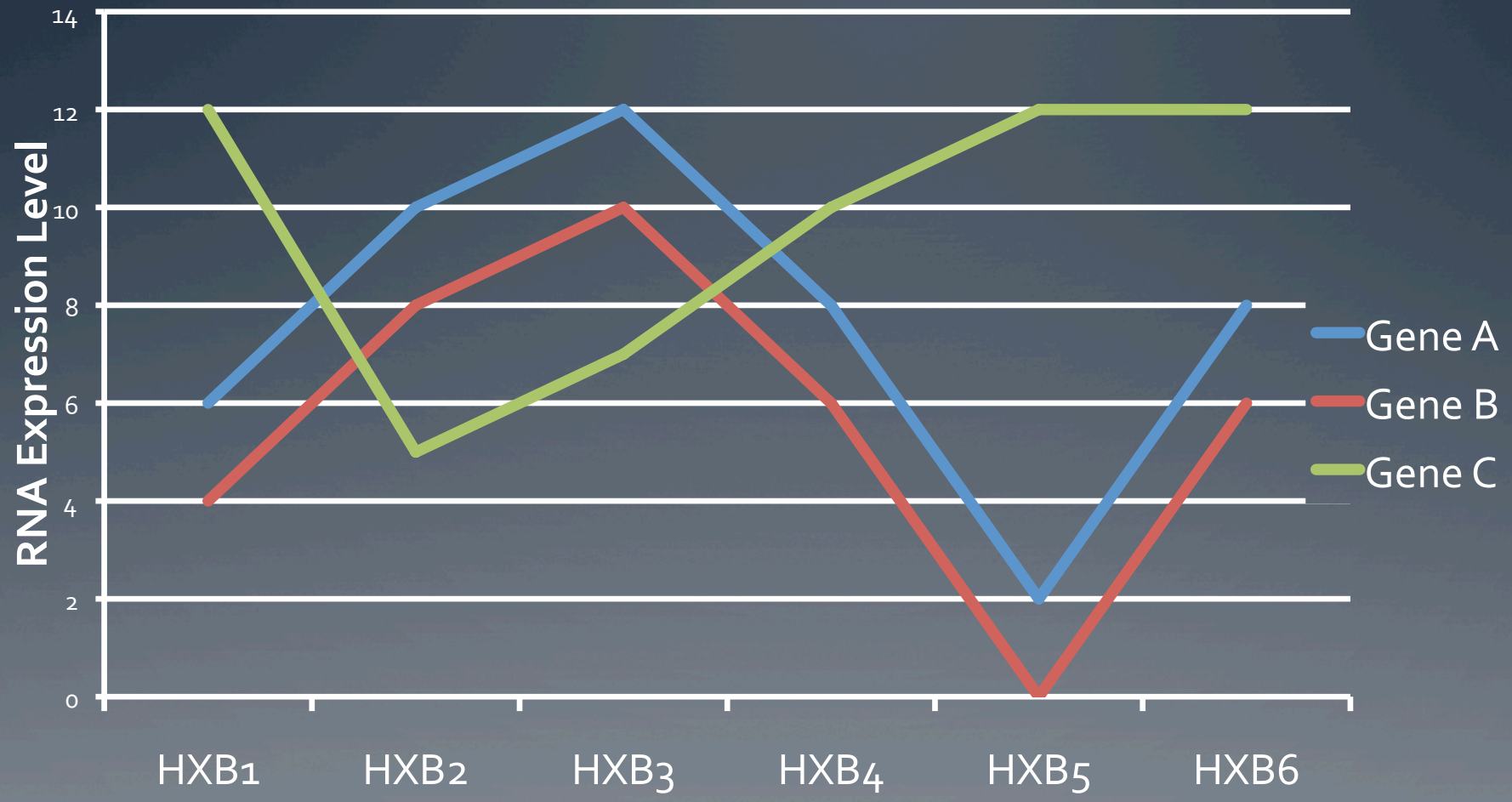
What are we missing by considering each candidate gene individually?

- Define relationships among genes
- Infer biological function from other co-expressed genes
- Define the context in which the gene exerts its effect
- Find multiple therapeutic targets within the same pathway

What do we gain by building networks and identifying modules?

- No gene product acts **independently** in the cell
- Information about biological function in cell
- Information on causes/ consequences of differential expression

How are transcripts related, one to another?



Co-expression as a measure of "interaction"

- **Theory** – if the magnitude of RNA expression of two transcripts correlates over multiple "environments", then the two transcripts are involved in similar biological processes
- Caveats when multiple environments = multiple genetic backgrounds
  - Linkage Disequilbrium – two genes are physically located near one another in the genome or the loci that control expression of two genes are located near one another in the genome
  - Environment-dependent correlation
  - Cell-type mixing proportions - in heterogeneous tissue, differences in the composition of cell types within a sample can present as correlations between transcripts that are cell type specific

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

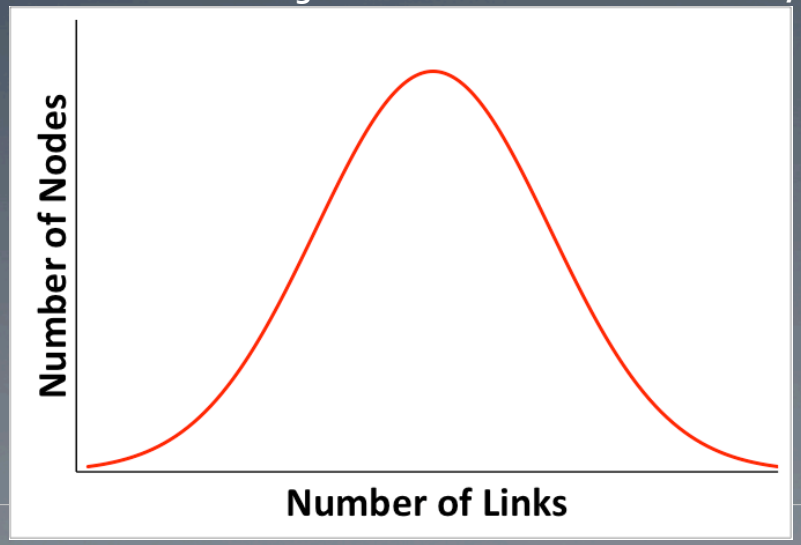
1. Simple correlation does not give connectivity.
2. How are we measuring co-expression?
   - Scale-Free Network
     - Network has few highly connected genes rather then each gene have similar connectivity
     - **Biologically motivated**, fewer highly connected genes means that a system is more robust to failure of any one gene
3. 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

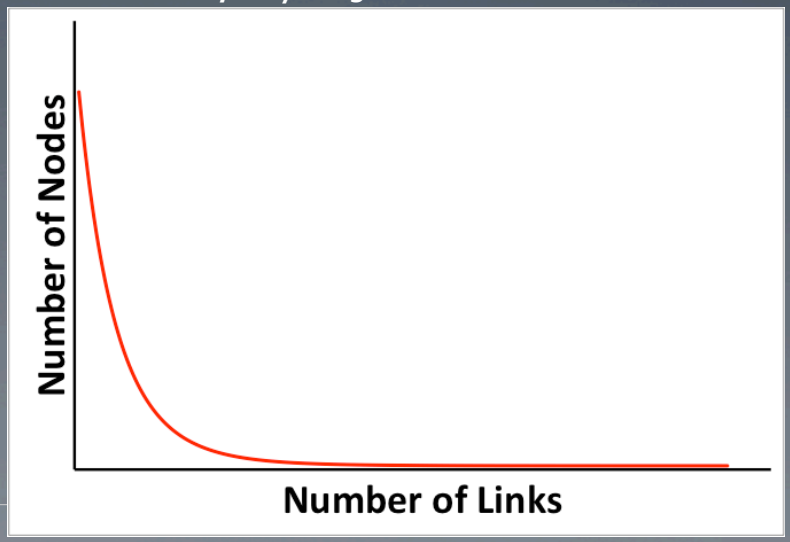
Scale-Free Networks

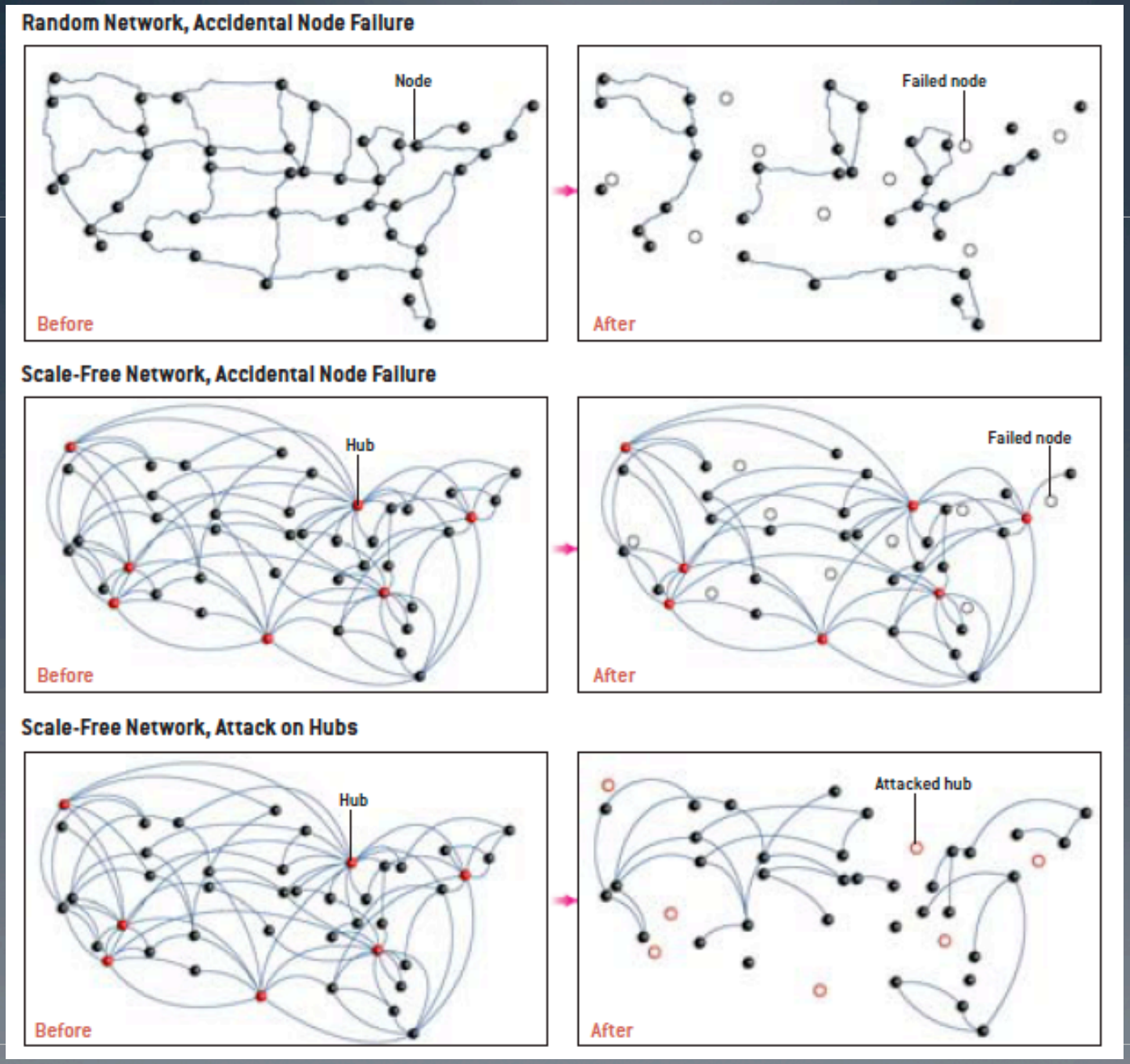
Random Network

Scale-Free Network

Images taken from *Scale-Free Networks*, Scientific American, May 2003



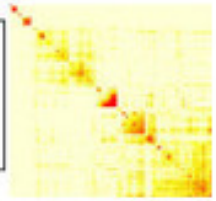




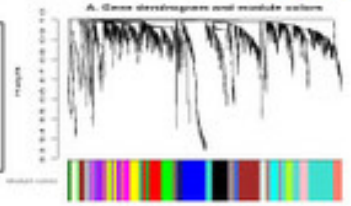
Images taken from *Scale-Free Networks*, Scientific American, May 2003

Typical WGCNA Workflow

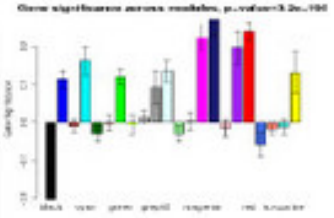
Construct a gene co-expression network
Rationale: make use of interaction patterns among genes
Tools: correlation as a measure of co-expression



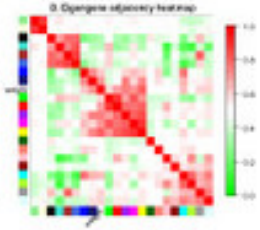
Identify modules
Rationale: module (pathway) based analysis
Tools: hierarchical clustering, Dynamic Tree Cut



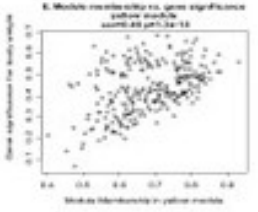
Relate modules to external information
Array Information: clinical data, SNPs, proteomics
Gene Information: ontology, functional enrichement
Rationale: find biologically interesting modules



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



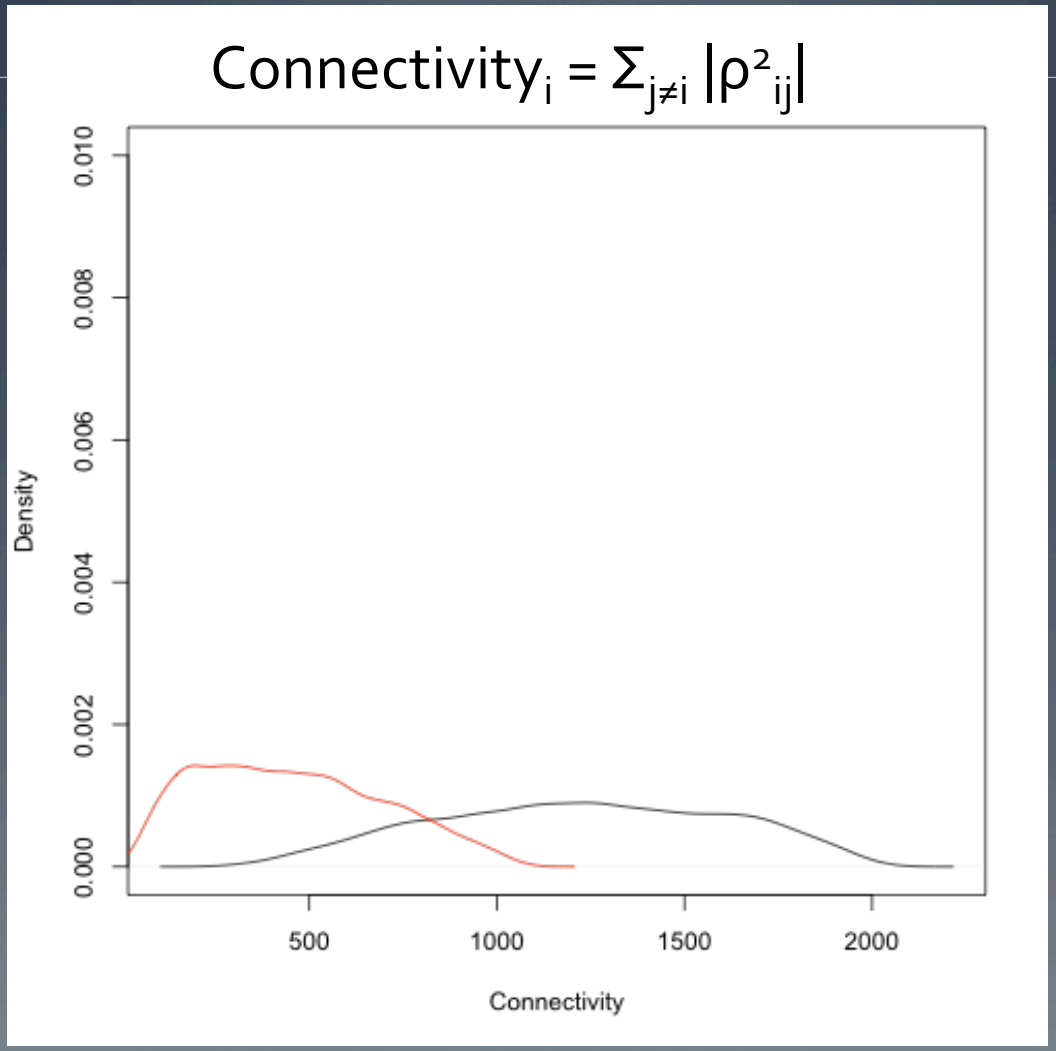
Find the key drivers in *interesting* modules
Rationale: experimental validation, biomarkers
Tools: intramodular connectivity, causality testing



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

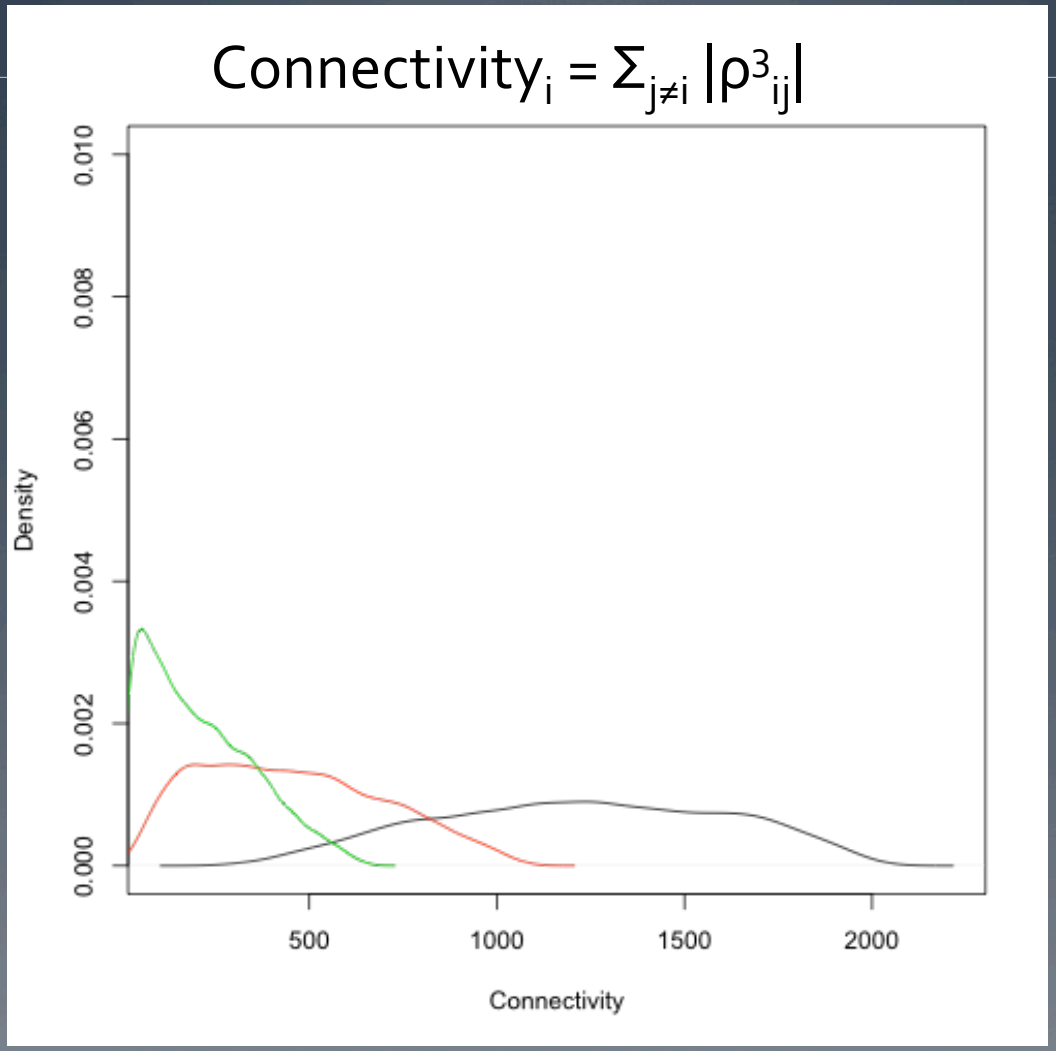
Scale-Free Network

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



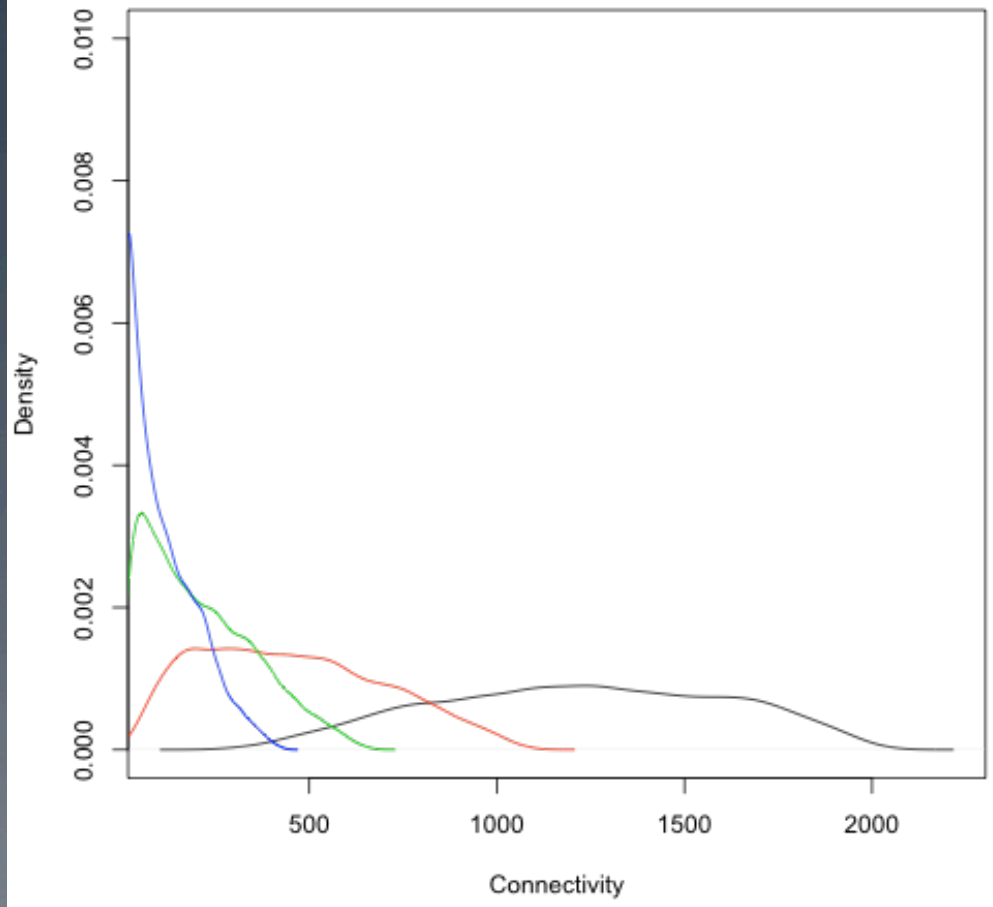
Scale-Free Network

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



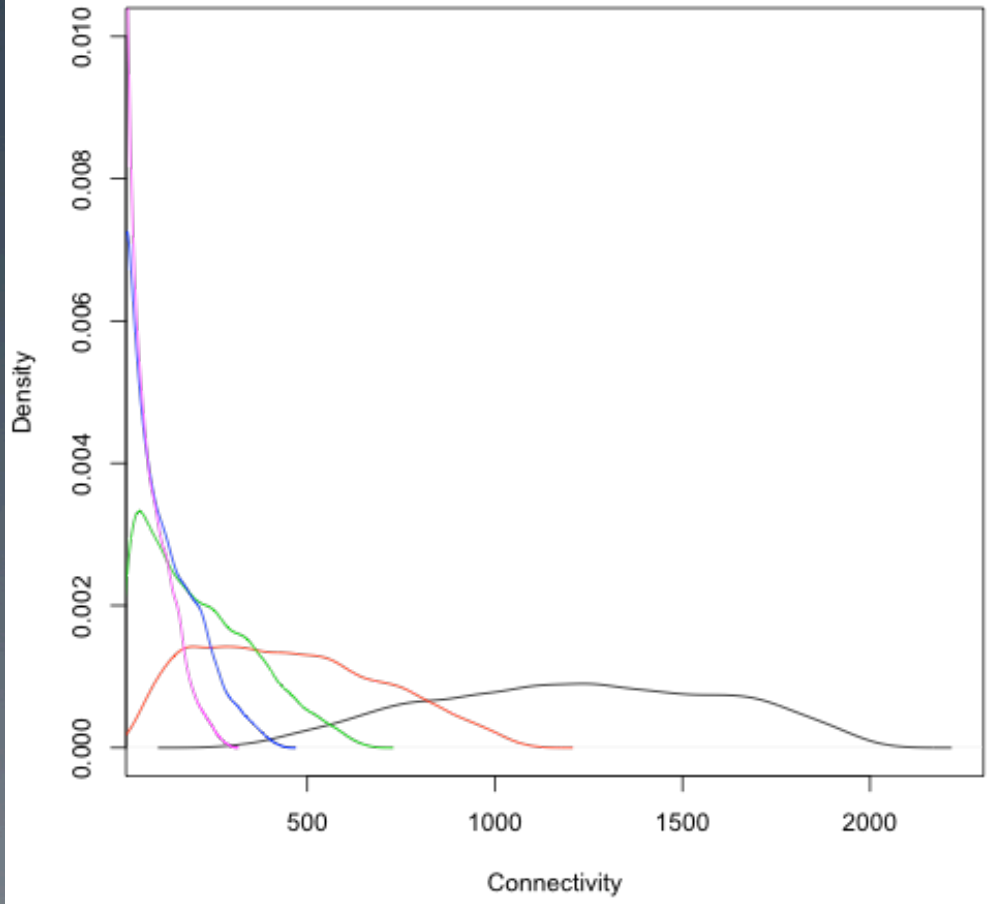
Scale-Free Network

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



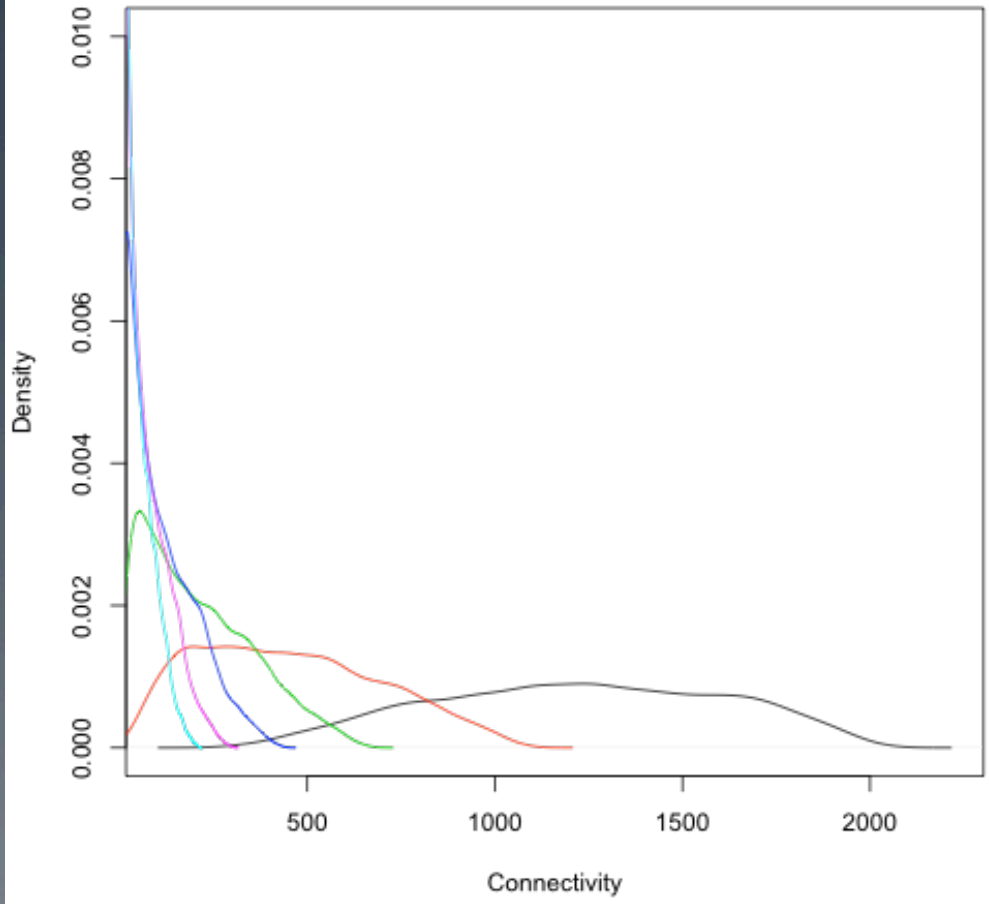
Scale-Free Network

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



Scale-Free Network

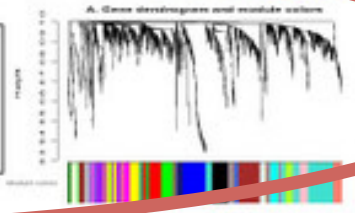
$$\text{Connectivity}_i = \Sigma_{j \neq i} |\rho^6_{ij}|$$



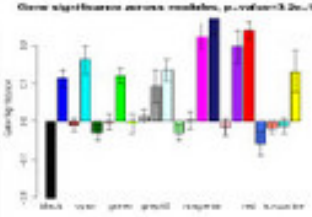
Typical WGCNA Workflow

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

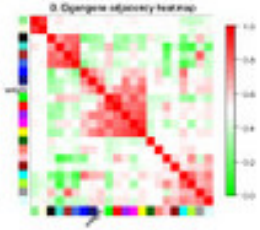
Identify modules
Rationale: module (pathway) based analysis
Tools: hierarchical clustering, Dynamic Tree Cut



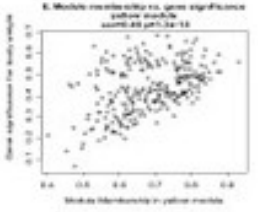
Relate modules to external information
Array Information: clinical data, SNPs, proteomics
Gene Information: ontology, functional enrichement
Rationale: find biologically interesting modules



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

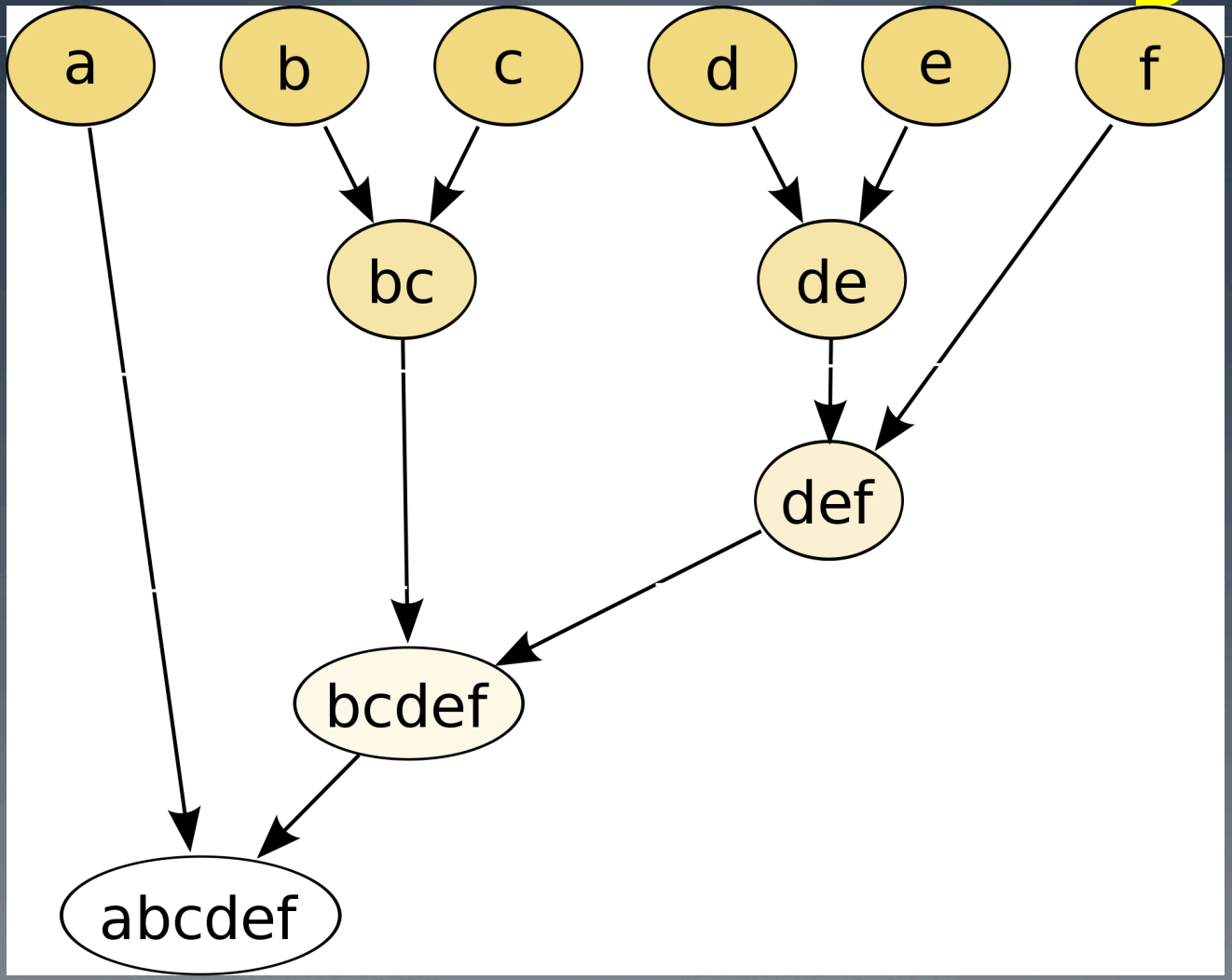
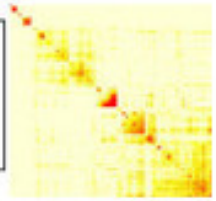


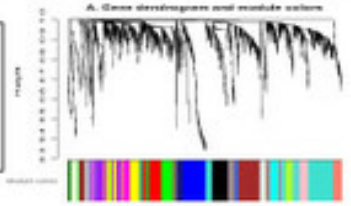
Figure reproduced from http://en.wikipedia.org/wiki/Hierarchical_clustering

Typical WGCNA Workflow

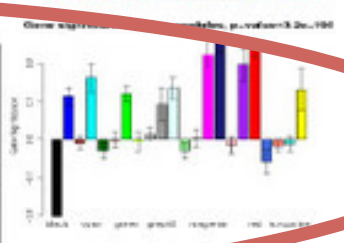
Construct a gene co-expression network
Rationale: make use of interaction patterns among genes
Tools: correlation as a measure of co-expression



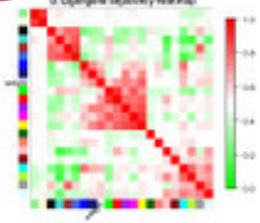
Identify modules
Rationale: module (pathway) based analysis
Tools: hierarchical clustering, Dynamic Tree Cut



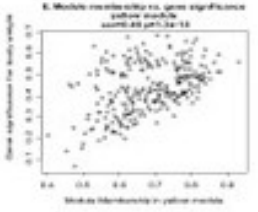
Relate modules to external information
Array Information: clinical data, SNPs, proteomics
Gene Information: ontology, functional enrichement
Rationale: find biologically interesting modules



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



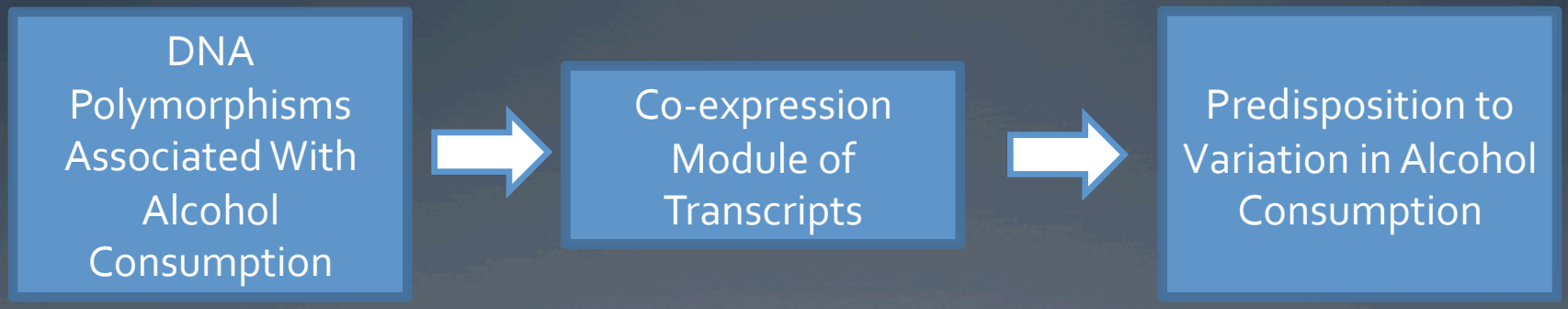
Find the key drivers in *interesting* modules
Rationale: experimental validation, biomarkers
Tools: intramodular connectivity, causality testing



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

Candidate Modules

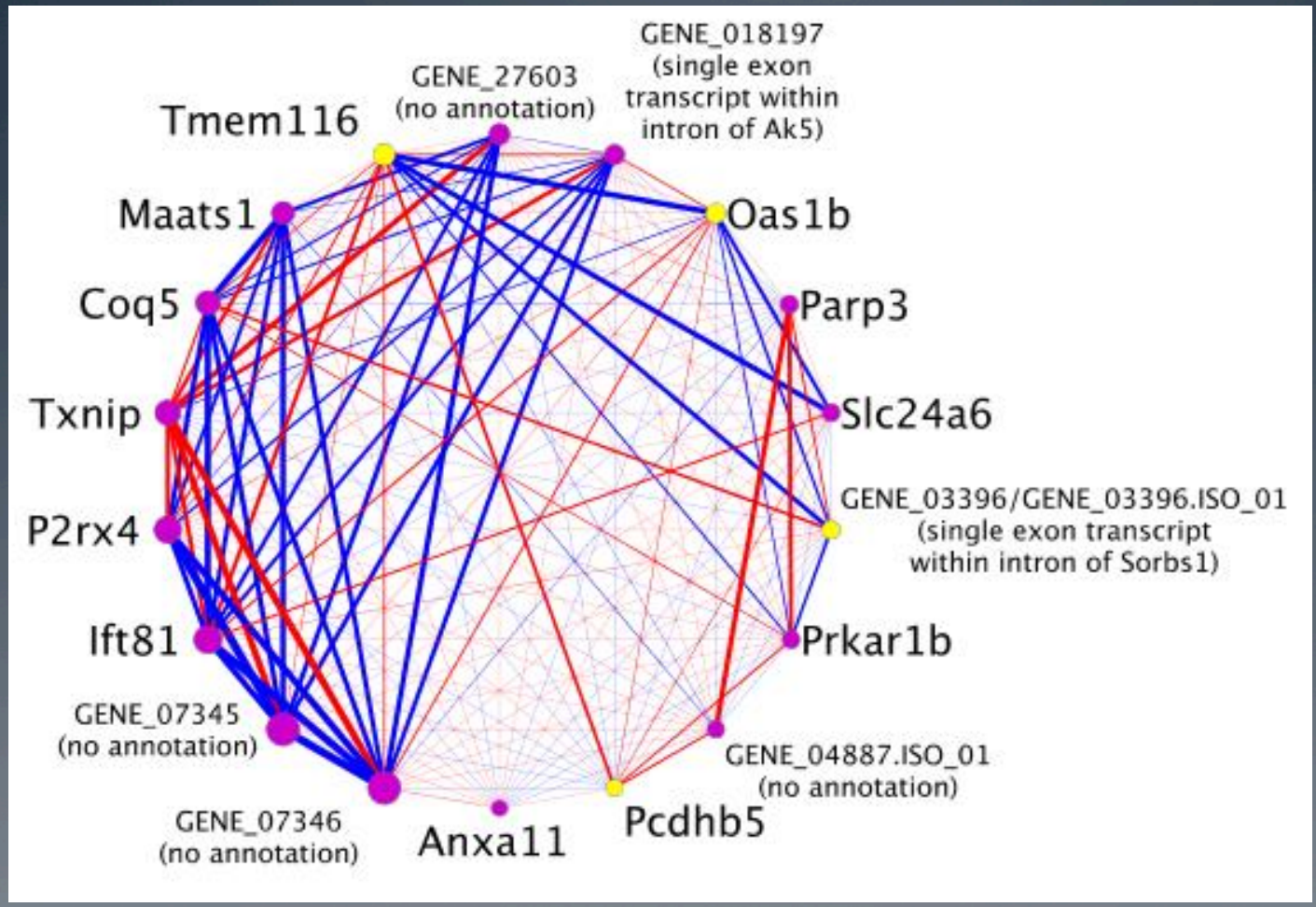
Transcriptional Pathway



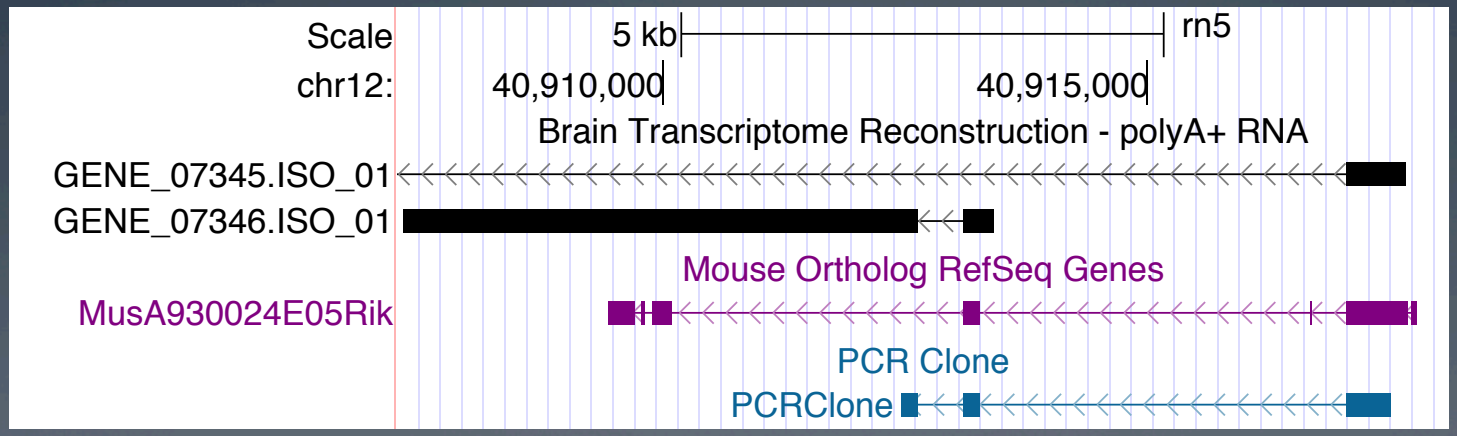
Candidate Modules

1. Significant eQTL for the co-expression module within a QTL for alcohol consumption
   - **DNA → RNA and DNA → phenotype**
2. Module expression correlated with alcohol consumption and/or genes within the co-expression module tend to be differential expressed in the selected lines
   - **RNA → phenotype**

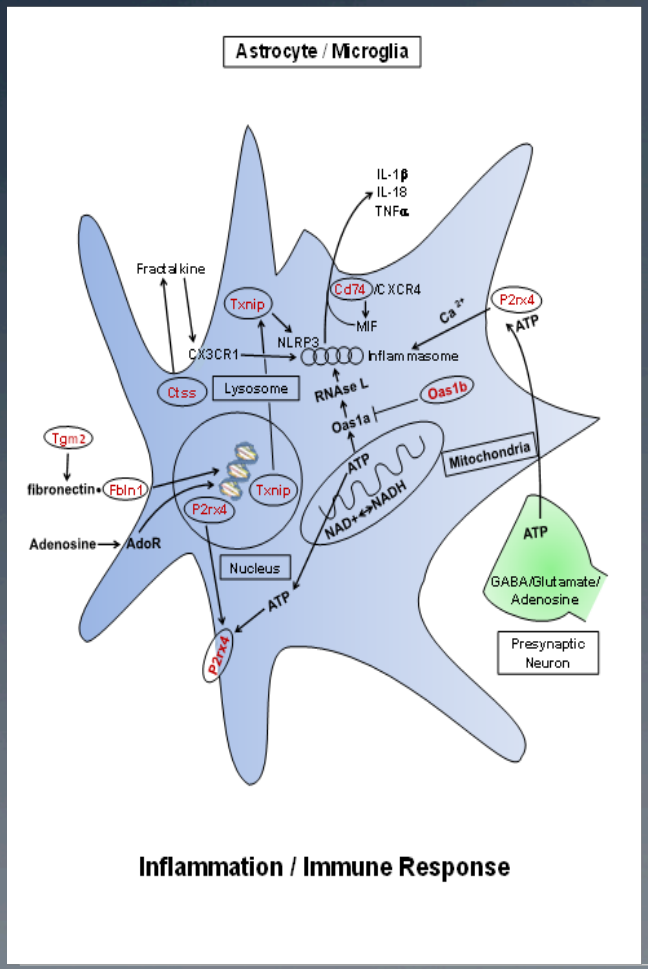
Transcriptional Pathway Associated With Alcohol Consumption

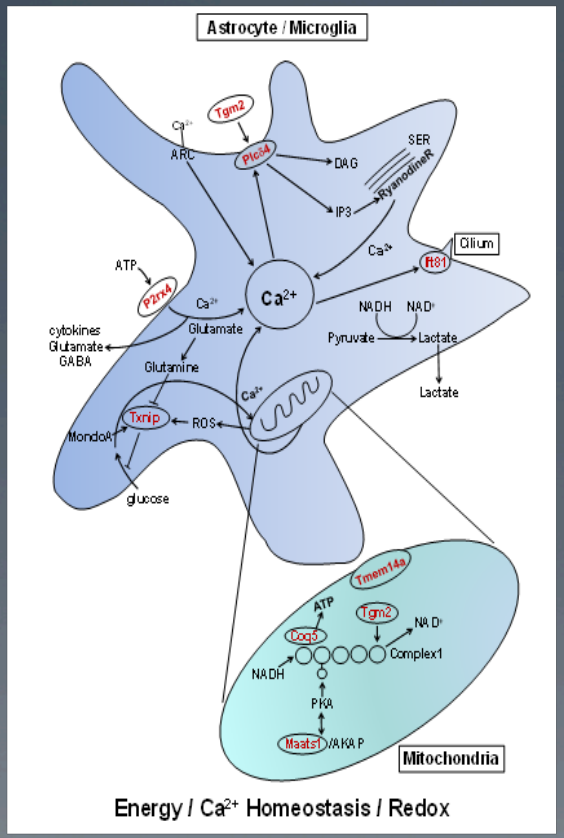


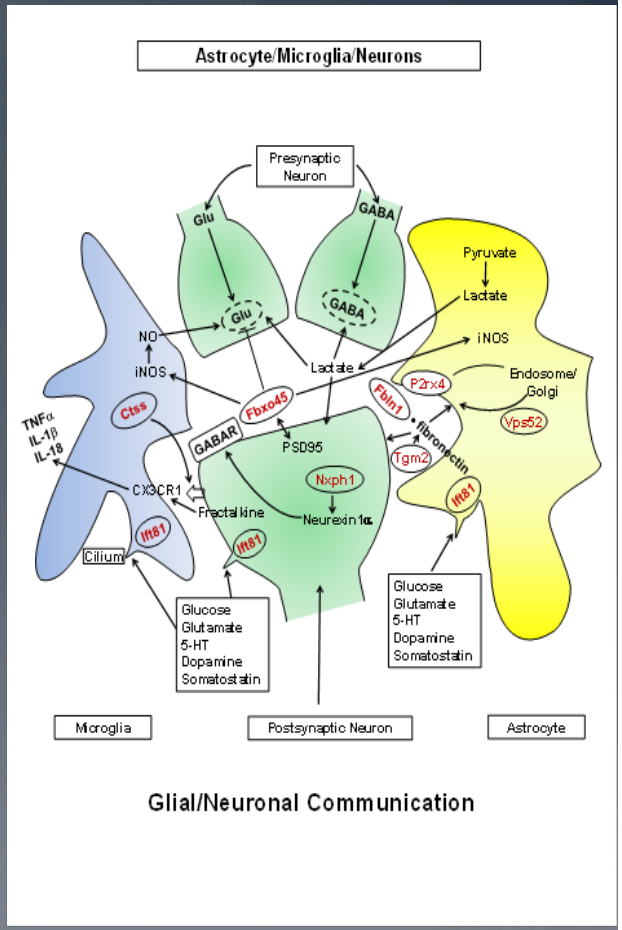
RNA-Seq and Unannotated Genes



Biological Context from Pathway





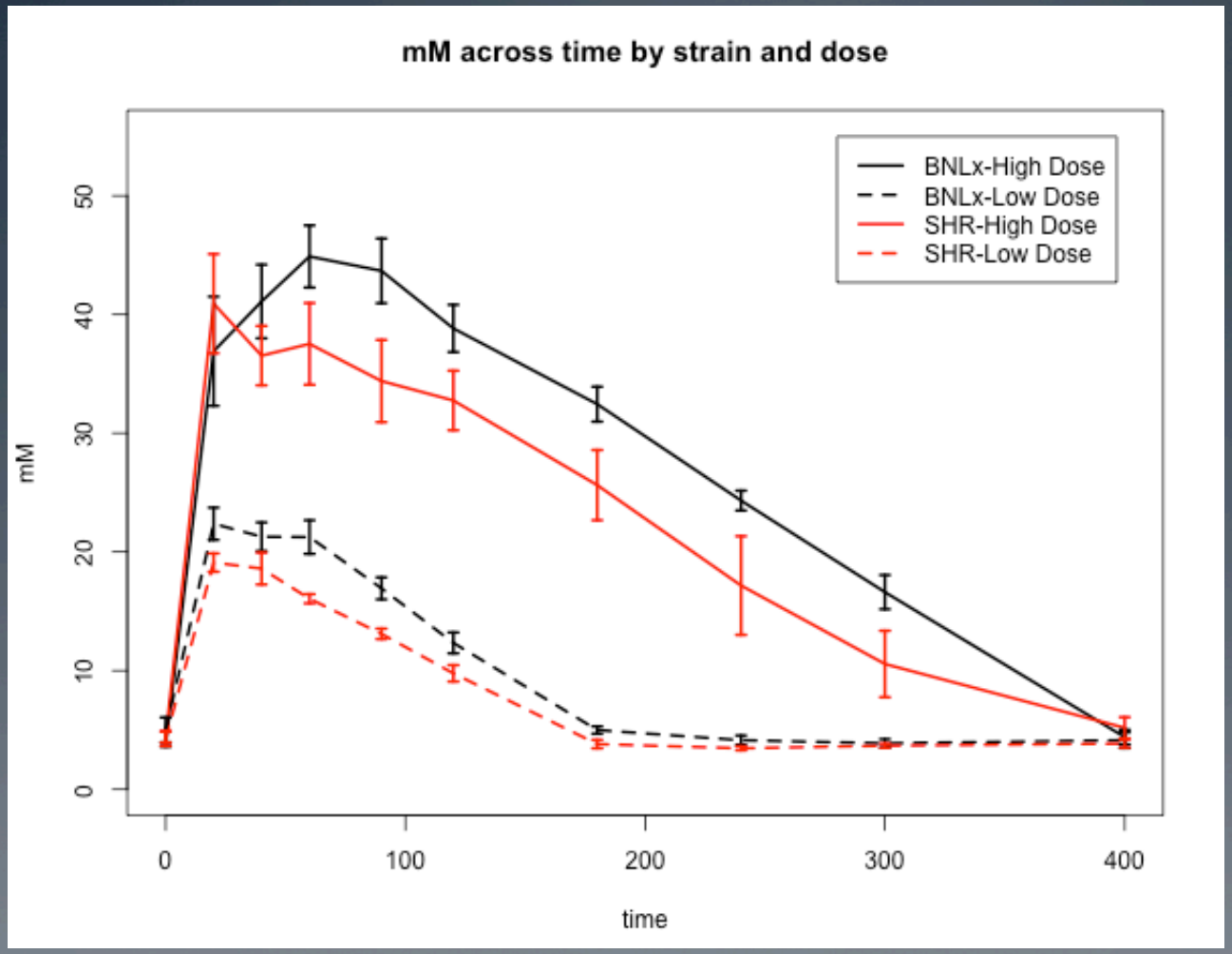


Genetic Risk for Alcohol-Induced Cancer

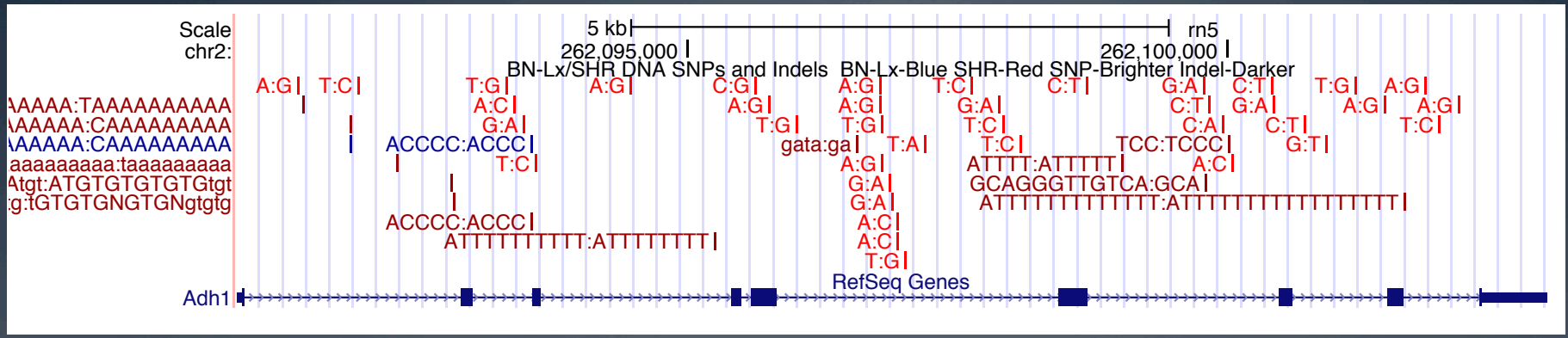
Disentangling genetic risk for high alcohol consumption from genetic susceptibility to alcohol-induced cancer

- In humans – genetic susceptibility to alcohol-induced cancer is confounded by the genetic susceptibility to high alcohol consumption
- Attempts to identify interactions between loci and alcohol consumption that significantly effect risk of alcohol-related cancer have given mixed results and even some results that are counterintuitive.
- Usual Suspects:
  - Alcohol Dehydrogenase 1B and 1C
  - Aldehyde Dehydrogenase 2
  - Methylenetetrahydrofolate reductase (NAD(P)H)
  - Cytochrome P450, family 2, subfamily E, polypeptide 1

Alcohol Metabolism in Rats

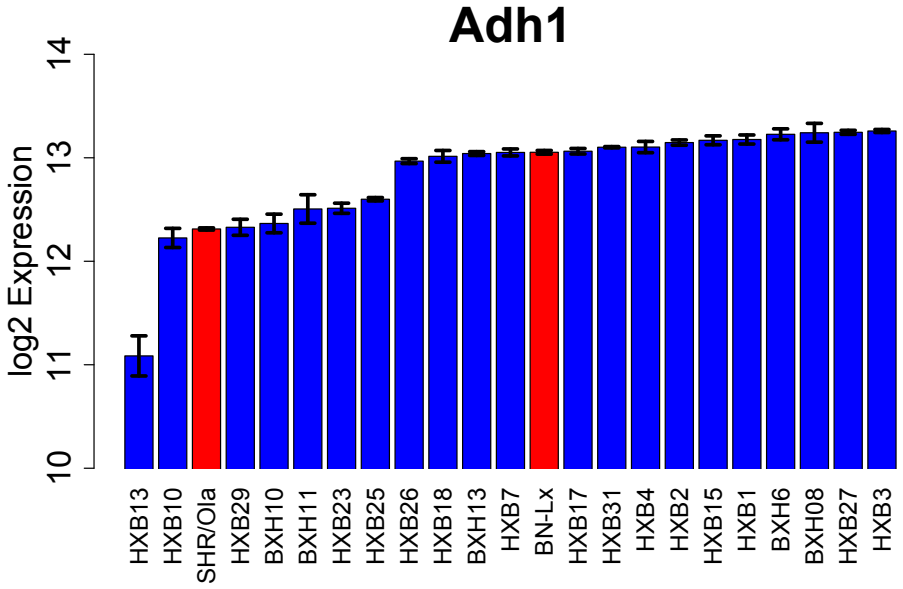


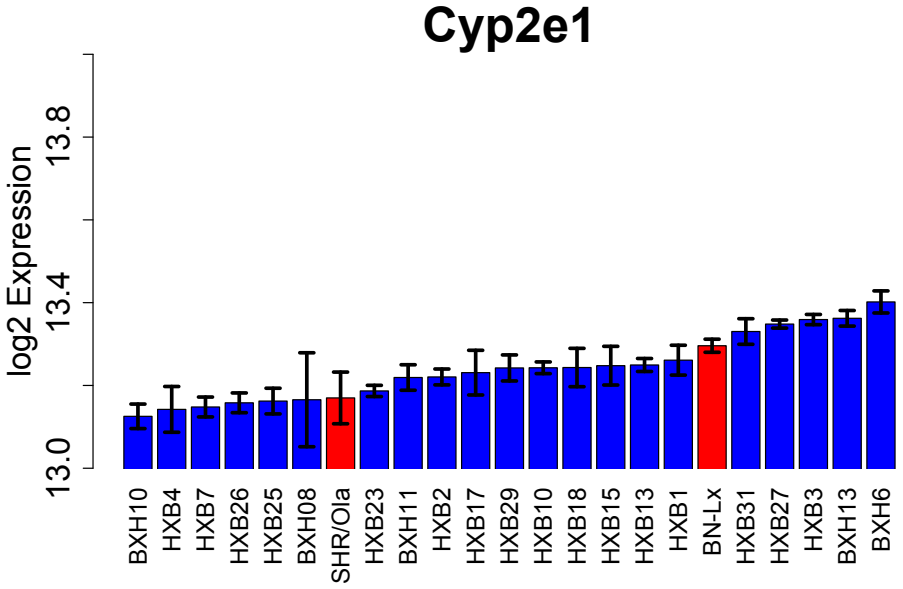
DNA Polymorphisms

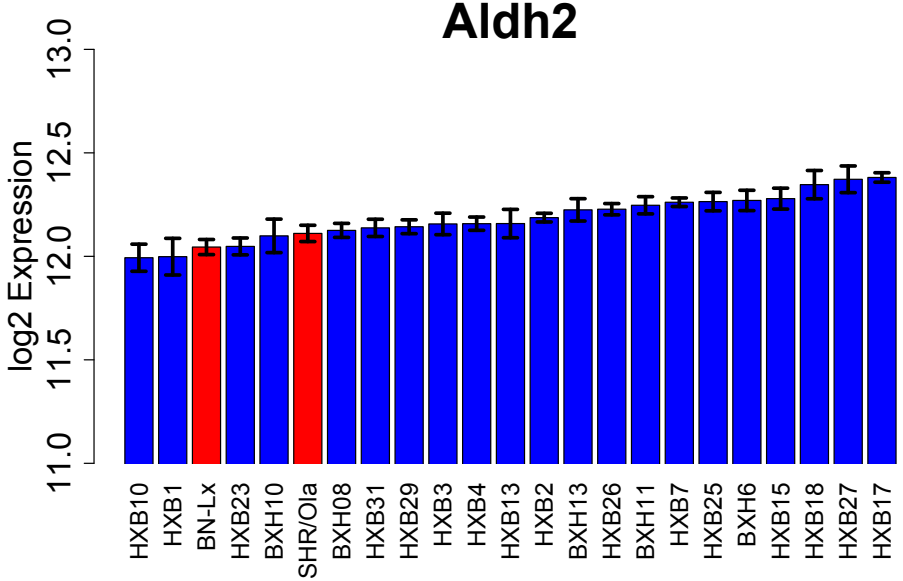


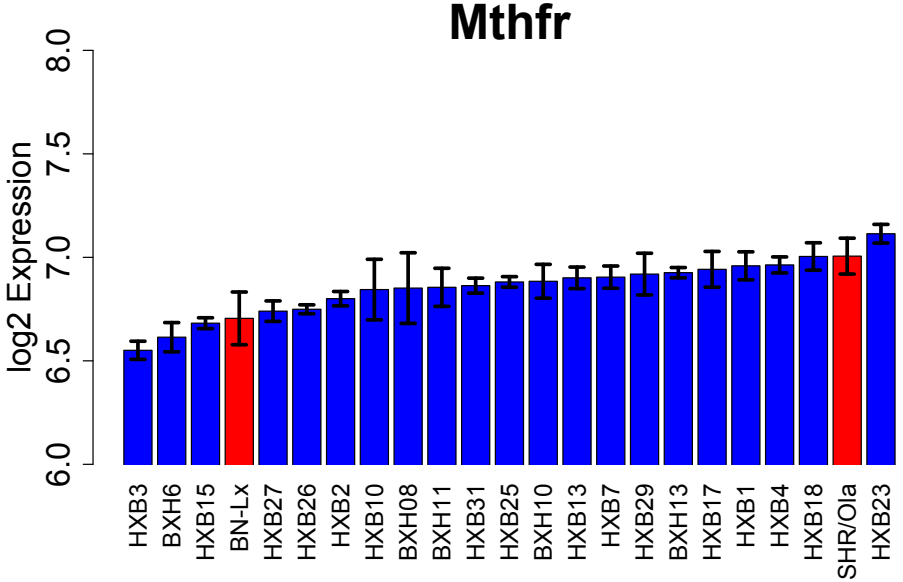
	Number of SNPs/Indels Within Exons	Number of SNPs / Indels Within Gene Region (introns and exons)
Adh1	4 / 0	38 / 10
Aldh2	6 / 0	71 / 24
Mthfr	0 / 0	0 / 1
Cyp2e1	0 / 0	8 / 3

RNA Expression - Liver

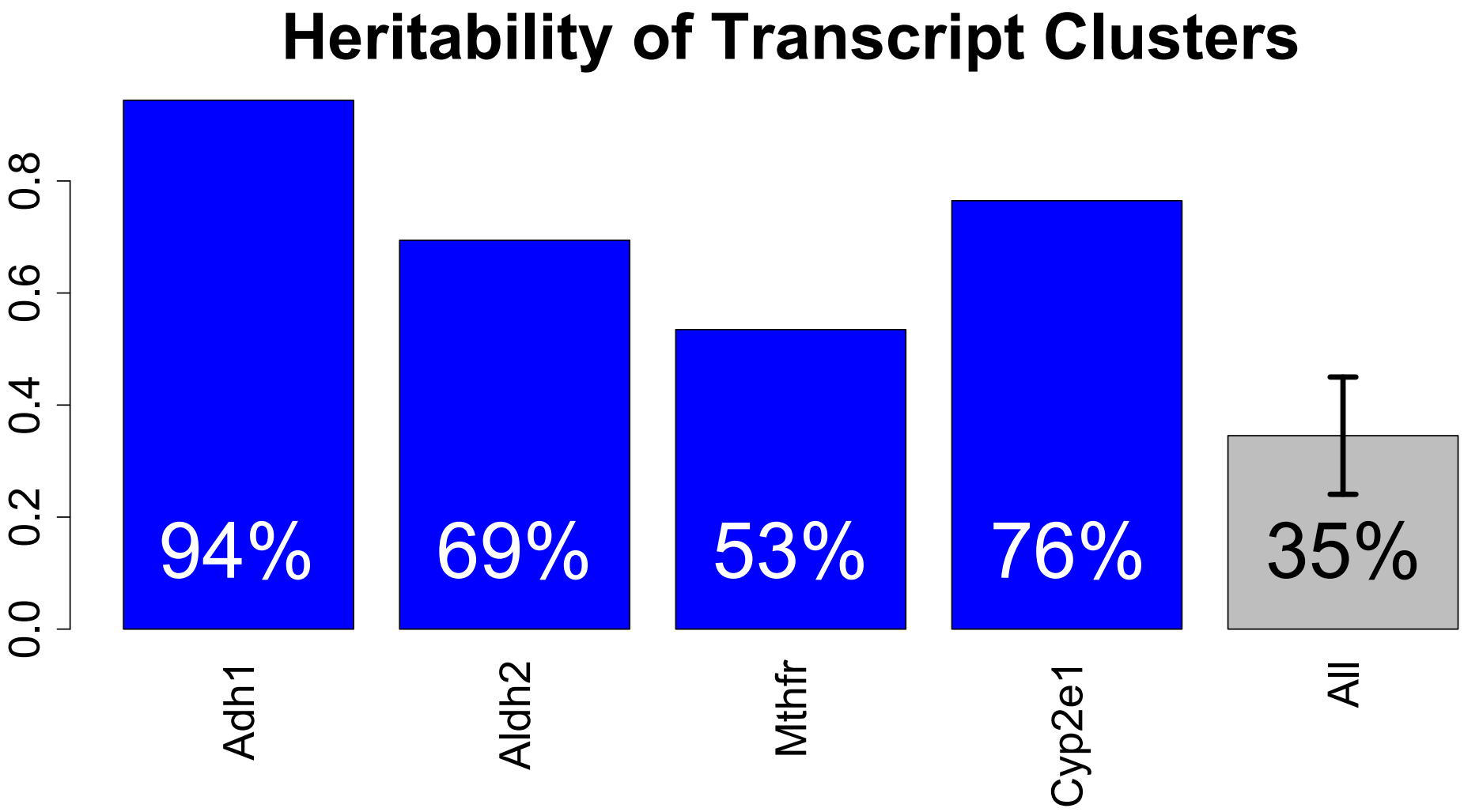




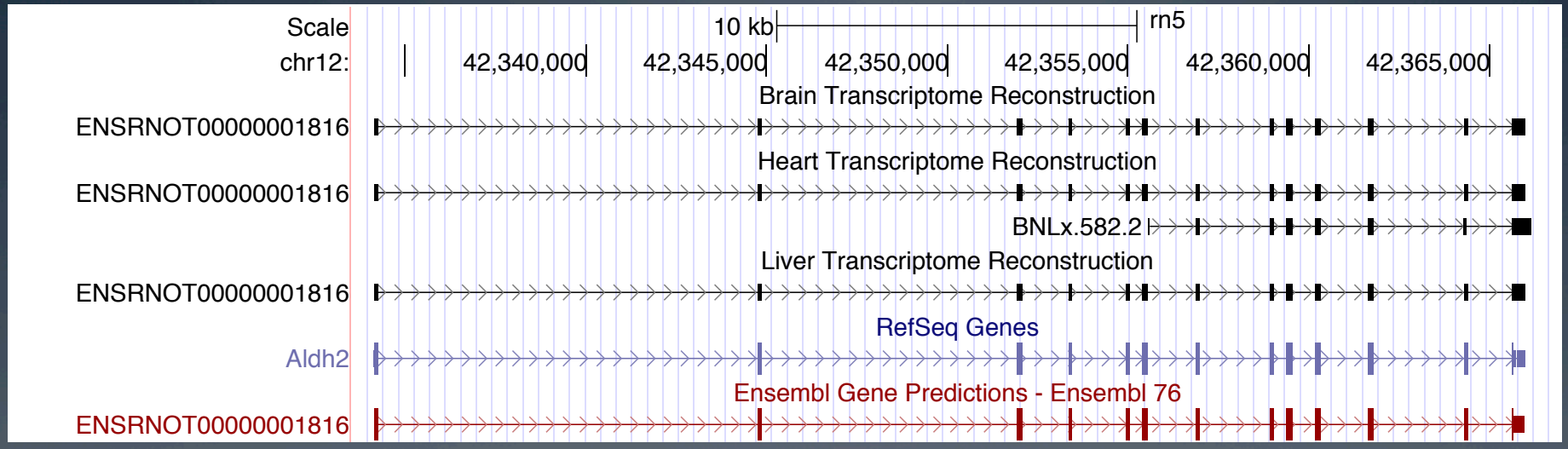


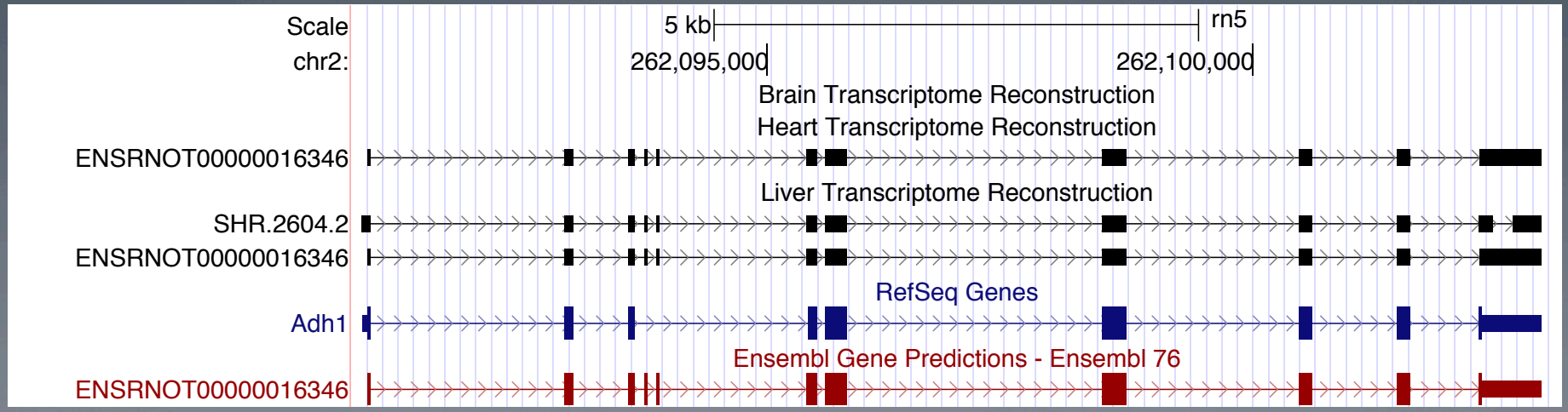


Heritability

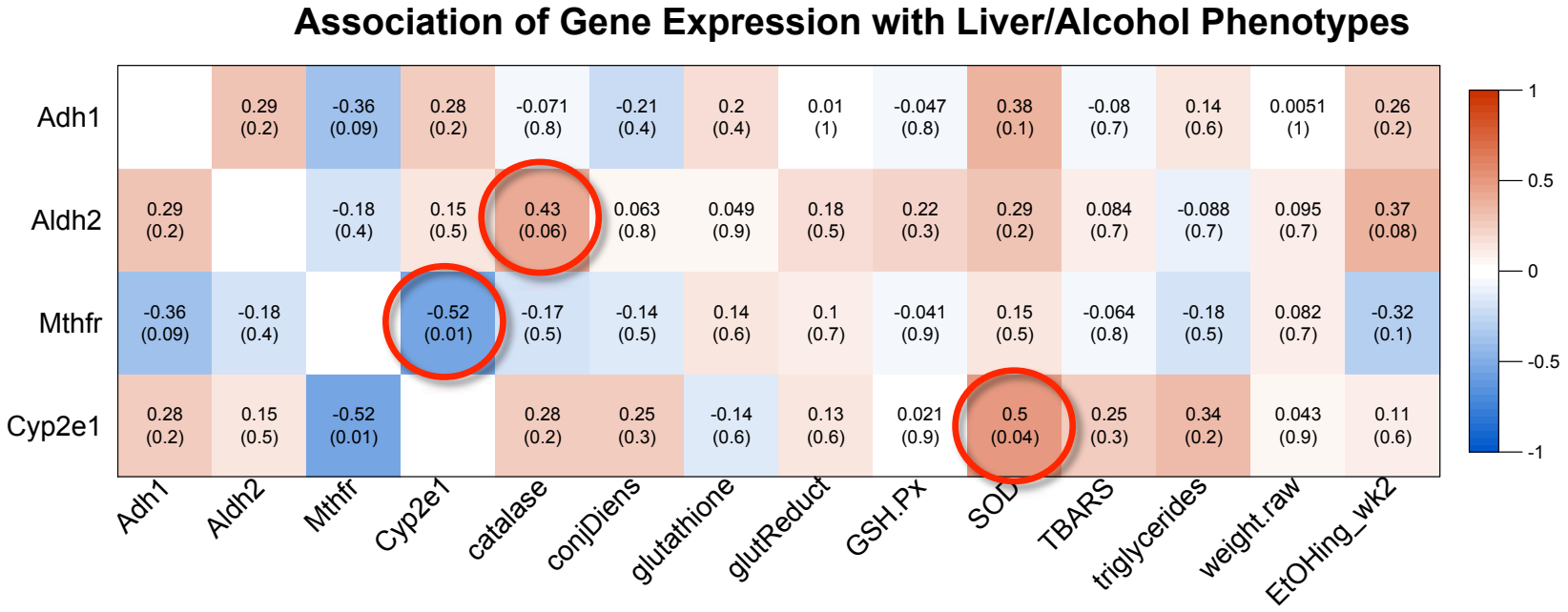


Transcriptome Reconstruction from RNA-Seq Data

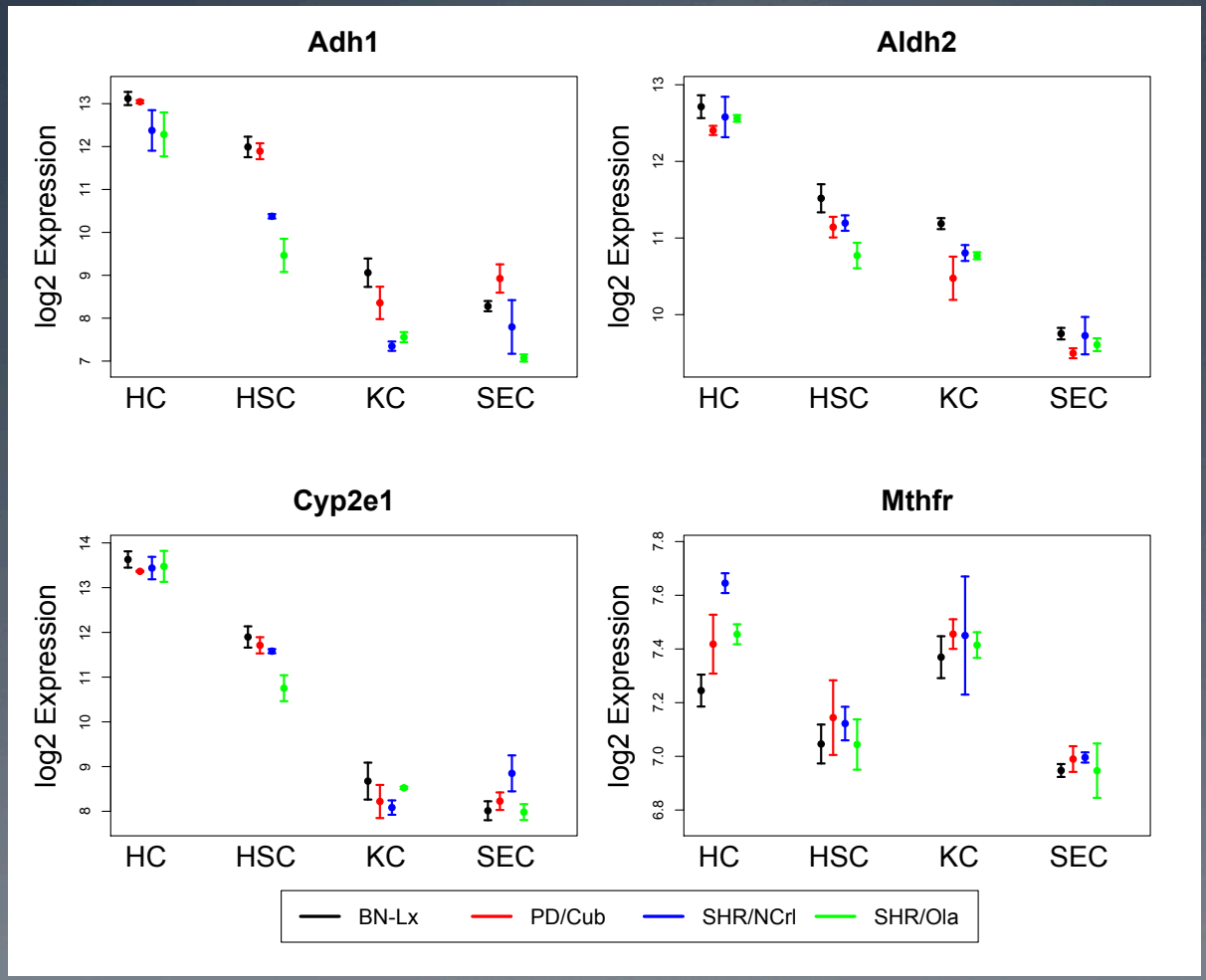




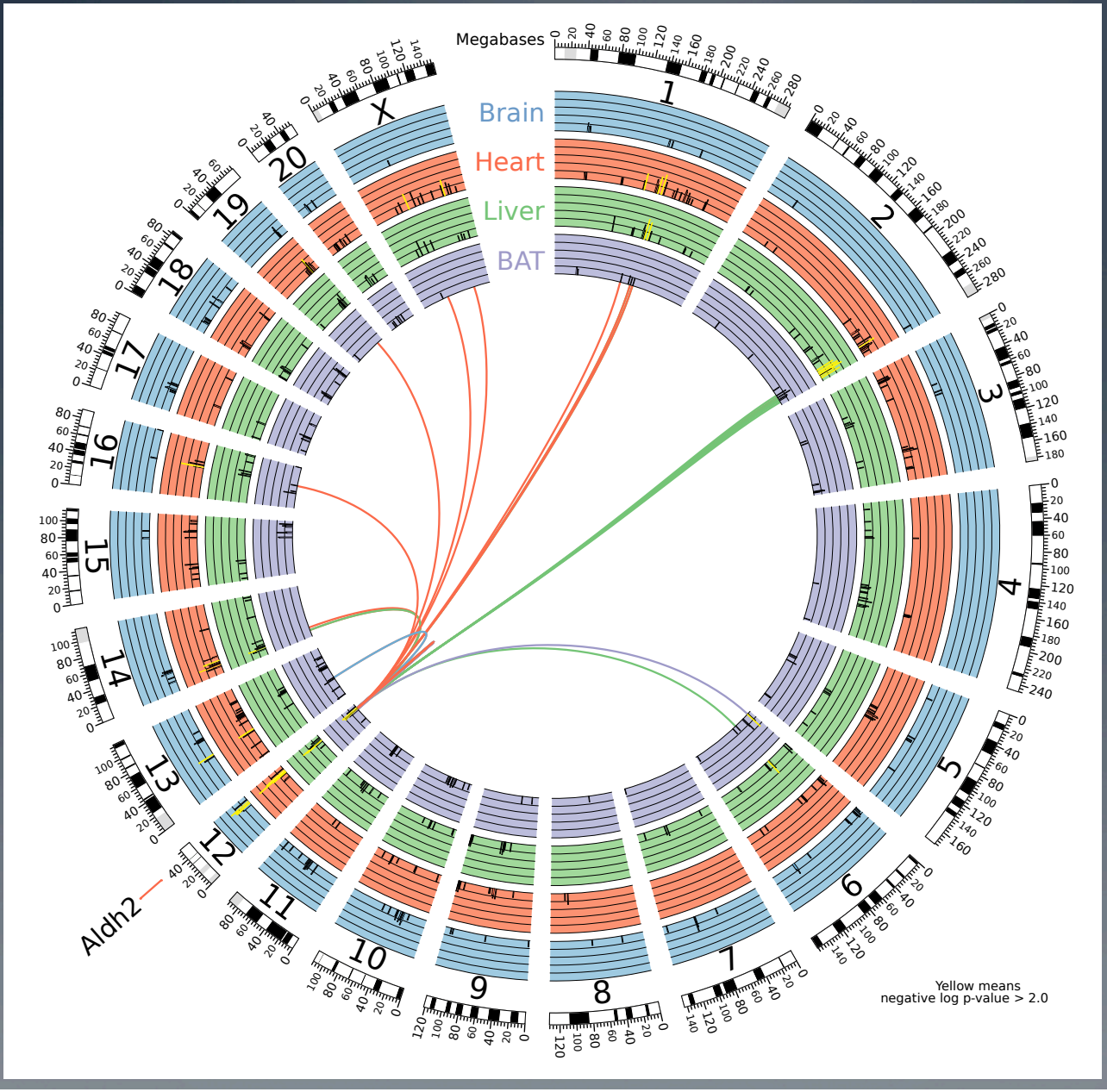
Correlation with each other and with liver traits



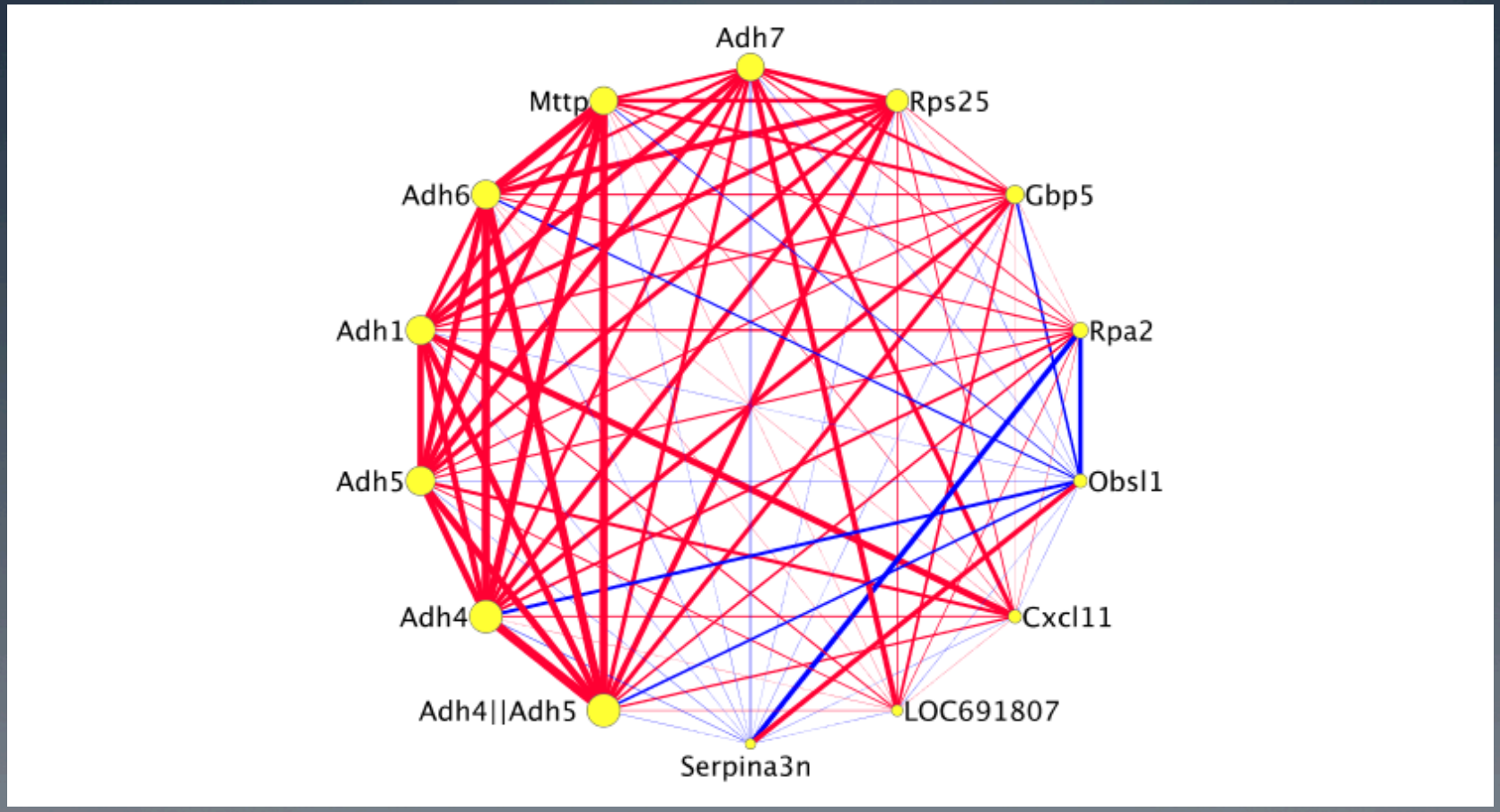
Cell-Specific Expression



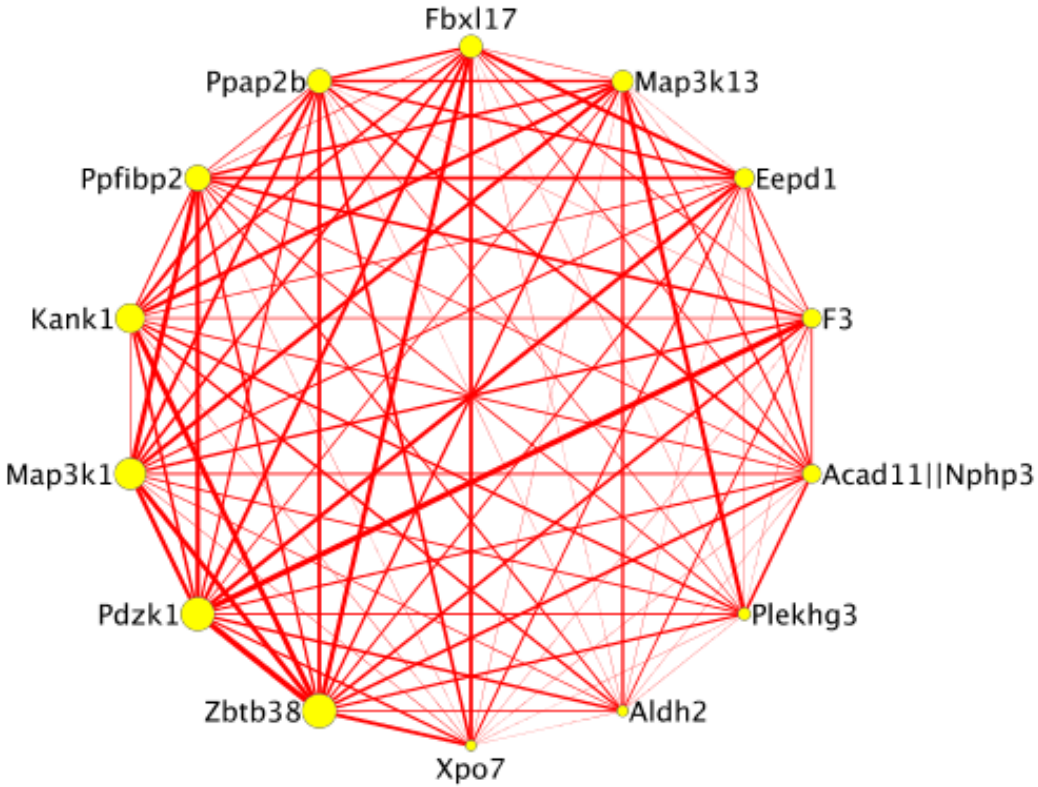
eQTL – Aldh2


Megabases
Brain
Heart
Liver
BAT
Aldh2
Yellow means
negative log p-value > 2.0

Adh1 Module


Adh7
Mttp
Rps25
Adh6
Gbp5
Adh1
Rpa2
Adh5
Obsl1
Adh4
Cxcl11
Adh4||Adh5
LOC691807
Serpina3n

Aldh2 Module



Future Directions

PhenoGen Database

- Fairly soon…
  - Improved transcriptome reconstructions for all tissues (brain, heart, liver)
  - Female Rat Brain Transcriptome Reconstruction – total RNA from BN-Lx and SHR strains
  - Quantification of brain RNA-Seq from 30 HXB/BXH RI
- Within the next few years…
  - RNA-Seq Data on Rat Diversity Inbred Panel (**60** strains of inbred and recombinant inbred rats)
    - Brain (total RNA and small RNA)
    - Liver (total RNA and small RNA)

Future Statistical Directions

- Identification and quantification for several large **RNA-Seq data** sets we are wadding through (http://phenogen.ucdenver.edu)
- Genetic **causal inference**, i.e., can we determine if the relationship between two 'correlated' transcripts and/or phenotypes is causal, reactive, or, completely independent
- **Predictive** network analysis when merging multiple 'omics' data sets in a single model
- Hierarchical Bayesian modeling to leverage the abundance of data on mouse/rat RNA expression in human eQTL and GWAS studies

Acknowledgements

- Supported by:
  - NIAAA (AA013162, AA013162-08S1, AA016663)
  - INIA-West Pilot Grant
  - National Foundation for Prevention of Chemical Dependency Disease (NFPCDD Career Development Award)
  - Banbury Fund
- All transcript expression data is available at:
  - http://phenogen.ucdenver.edu
- Special thanks to:
  - Dr. Boris Tabakoff, UCD
    - Dr. Stephen Flink
    - Spencer Mahaffey, ME
    - Lauren Vanderlinden, MS
    - Yinni Yu, MS
  - Dr. Paula Hoffman, UCD
    - Seija Tillanen
  - Dr. Katerina Kechris, CSPH
  - Dr Morton Printz, UCSD – HXB/BXH rats
    - Laura Breen
  - Dr. Hidekazu Tsukamoto, USC – liver cell extraction
  - UCD Microarray core – DNA-Seq
  - Dr. James Huntley, UCB – RNA-Seq