

**Methods for gene network reconstruction and
issues related to the proper interpretation of
genetic/gene expression data**

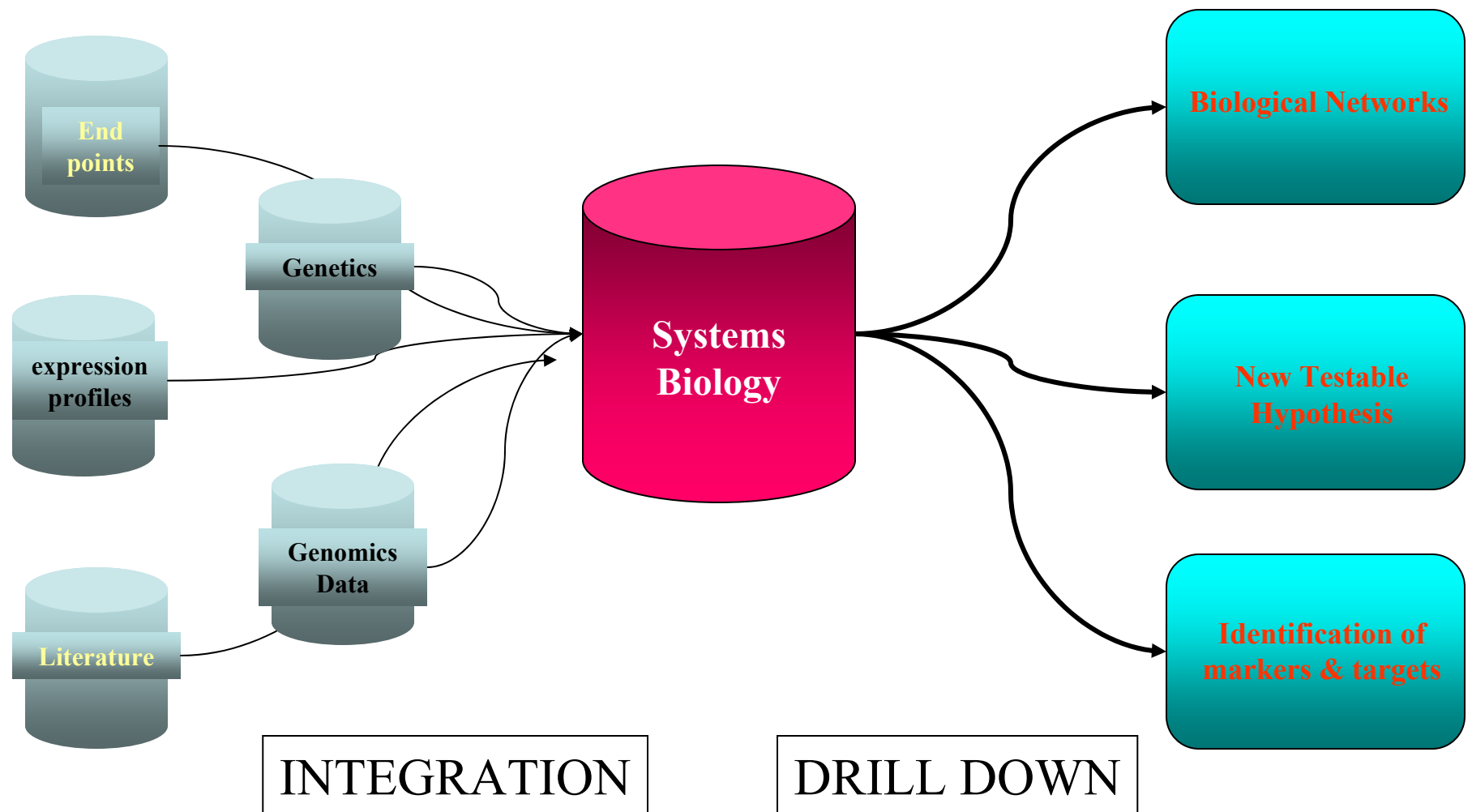
Eric Schadt, Ph.D.

Research Genetics

Rosetta Inpharmatics/Merck Research Labs

11 May 2005

Integrative Genomics: Era of large-scale, high-throughput analysis pipelines

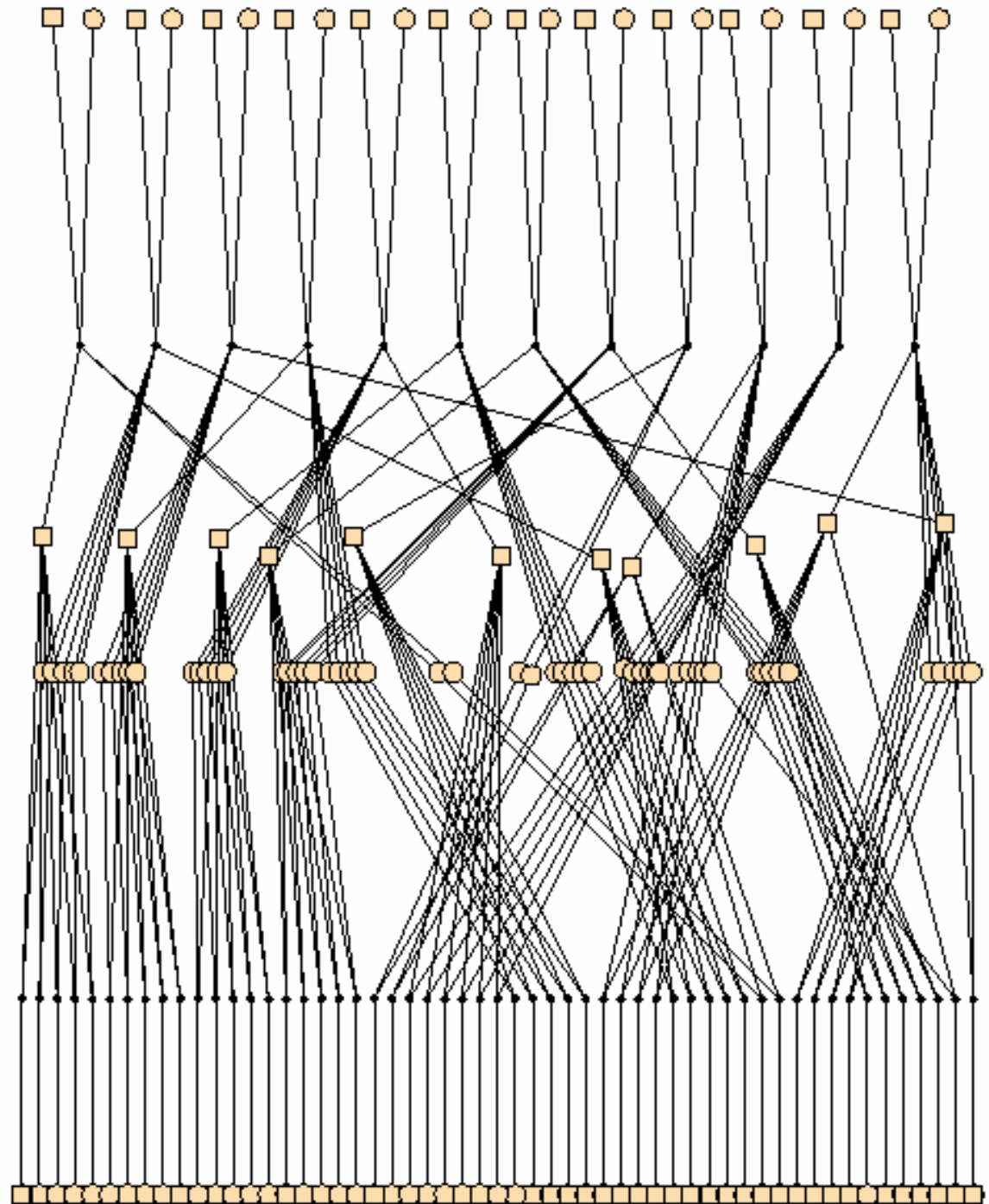


Sex Matters

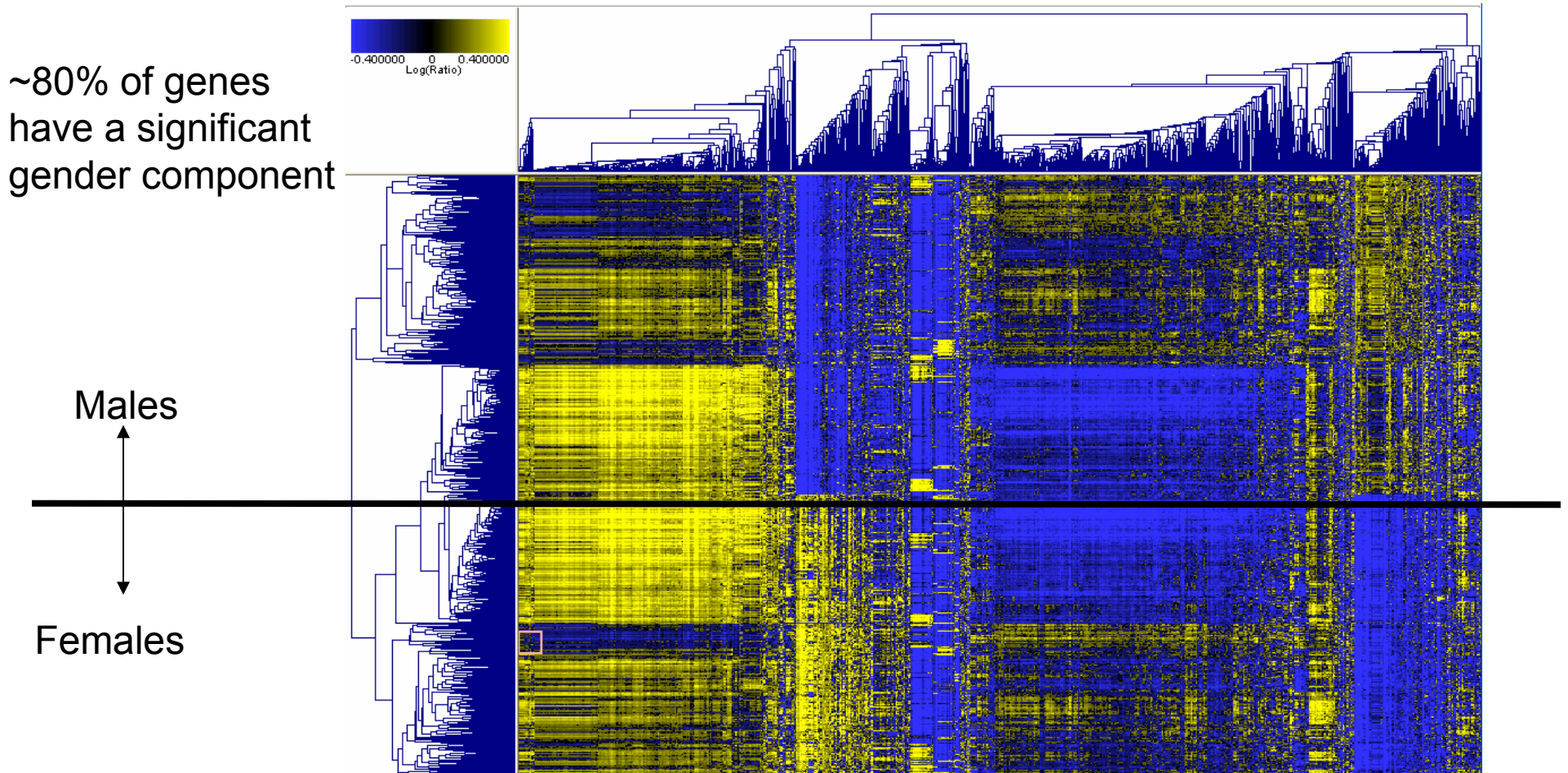
- Widespread differences observed between sexes with respect to clinical traits
- Widespread differences observed between sexes with respect to gene expression traits
- Differences driven by genetic and environmental factors
- All of our software for QTL analysis, “correlation” analysis, network reconstruction, accounts for sex and age covariates with interactions

More sophisticated mouse pedigree:

- UNL M16 X ICR Cross
- 24 outbred founders (12 families)
- M16 mice generated from selective breeding for rapid weight gain
- Males develop diabetes-like complications



Hierarchical Cluster of ~1500 of the Most Transcriptionally Active Liver Genes



- Must take gender into account
- Overall eQTL signature similar to what others have reported
- Focus here is on patterns of expression enriched for pathways associated with clinical traits and driven by genetic loci driving clinical traits

Straightforward models that take sex into account

- Regression-based approach to detect QTL:

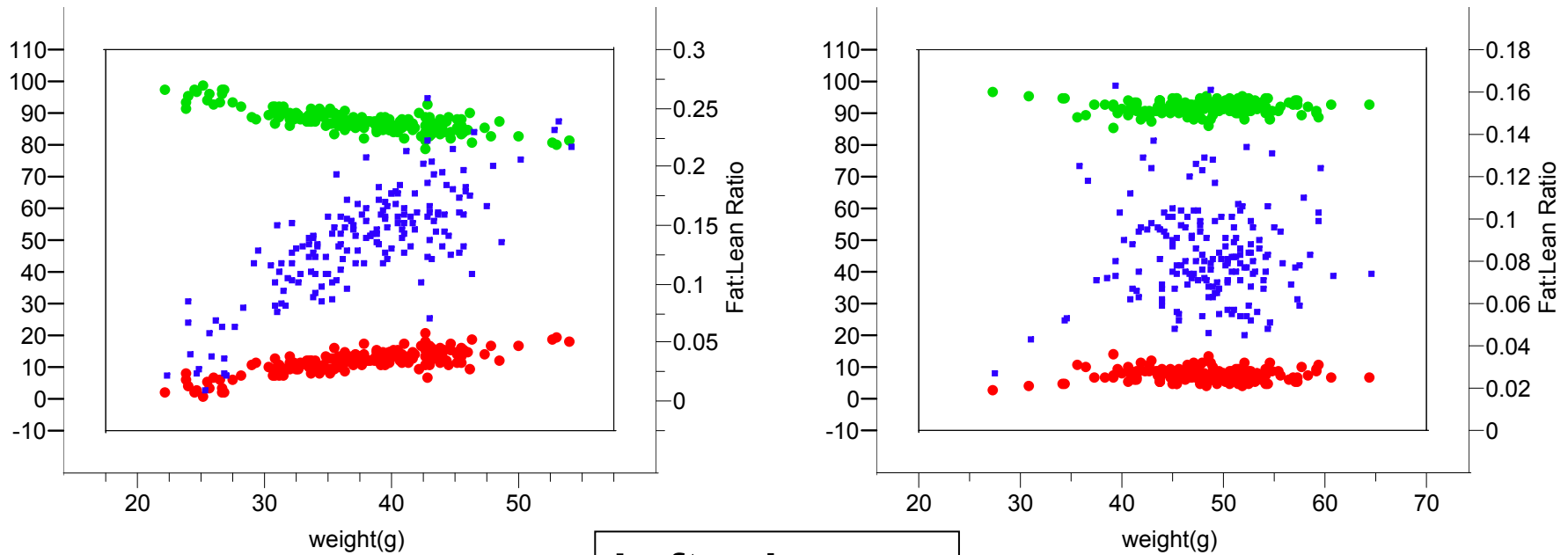
$$y = \mu + \beta_1(\text{additive}) + \beta_2(\text{dominant}) + \beta_3(\text{sex}) + \beta_4(\text{additive*sex}) + \beta_5(\text{dominant*sex}) + \varepsilon$$

- Regression-based approach to detect associations between traits:

$$y = \mu + \beta_1x + \beta_3(\text{sex}) + \beta_4(x*\text{sex}) + \varepsilon$$

- Fit sub-models from these “full” models and use likelihood ratio statistics to assess significance of models of interest
- From these simple models we see that:
 - ~80% of expression traits in adipose and liver have a significant sex component
 - 15-20% of expression traits in adipose and liver have significant sex*gene interaction
 - ~60% of expression traits in muscle have a significant sex component
 - Interestingly, only ~15% of genes expressed in brain have a significant gender component
- Experiments ongoing to determine if sex effects are driven by environmental conditions (e.g., different sexes living in different hormone pools) or genetic (genes differentially expressed off of X chromosome between sexes drives broader expression differences)

Relationship between lean and fat mass percent as a function of overall weight



Female

Female

Weight vs fat:lean ratio:

R_{sq} 0.6

ANOVA p value <0.0001

Left axis:

Lean mass (%)

Fat mass (%)

Right axis:

Fat/Lean Ratio

Male

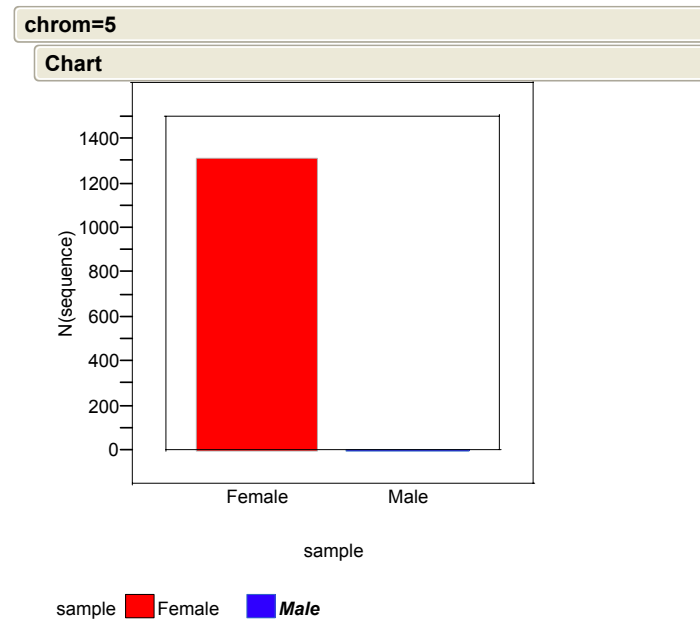
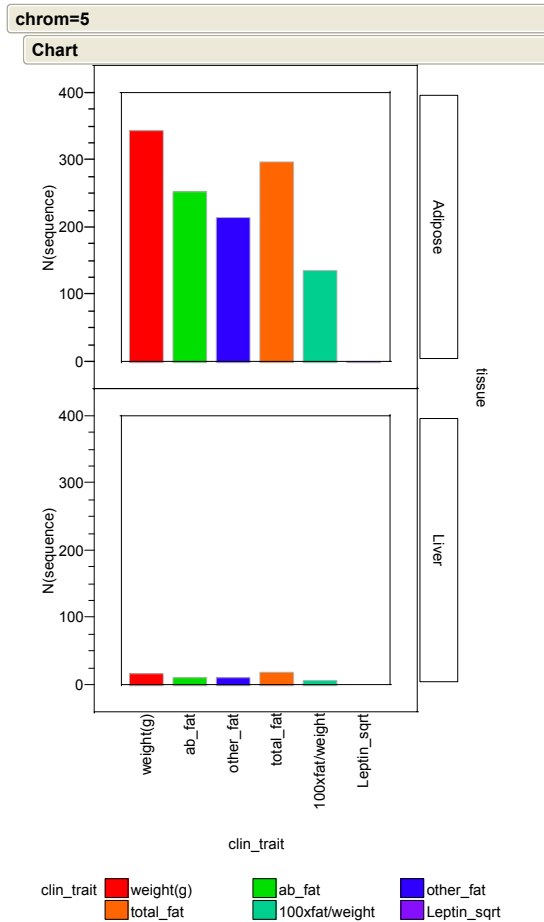
Male

Weight vs fat:lean ratio:

R_{sq} 0.0000

ANOVA p value <0.91

Genes correlated to obesity: Chromosome 5



Environment Matters

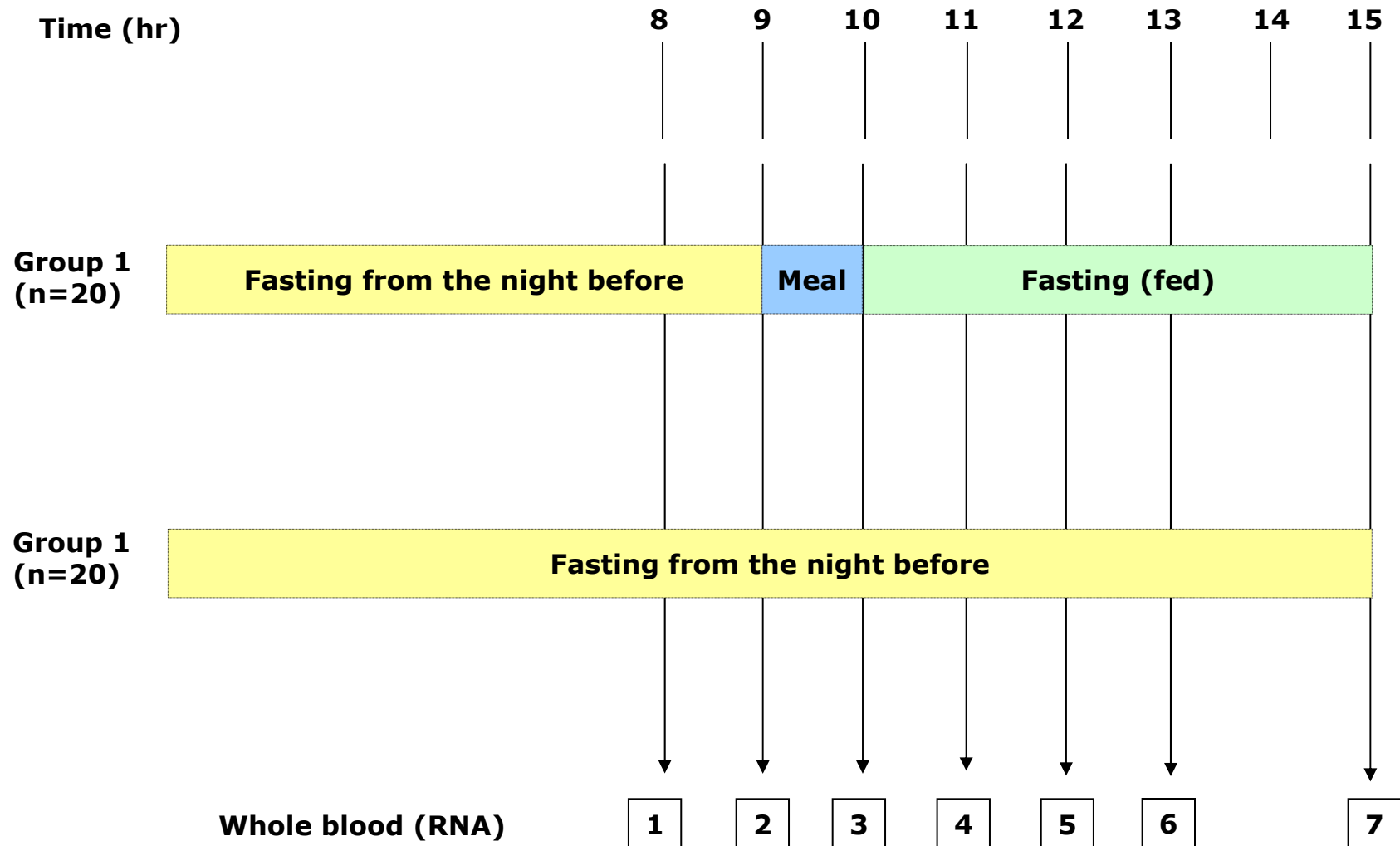
Example: Taking samples in the fasted or fed state

- Does fasting/fed state significantly impact expression profile
- When studying obesity (and related diseases) will it be more informative to collect samples for gene expression profiling from individuals that are in a fasted state or a fed state?

Fasting vs. Feeding Experiment

- 20 Individuals.
- Seven time points at which blood was collected and profiled.
- Two Arm Experimental Design : Each individual participated in both the fasting arm and the feeding arm (1 week separation)
- ~50 Clinical traits were scored at time point #1 in each of the two arms

Experimental Design



How to discriminate fasting / feeding status?

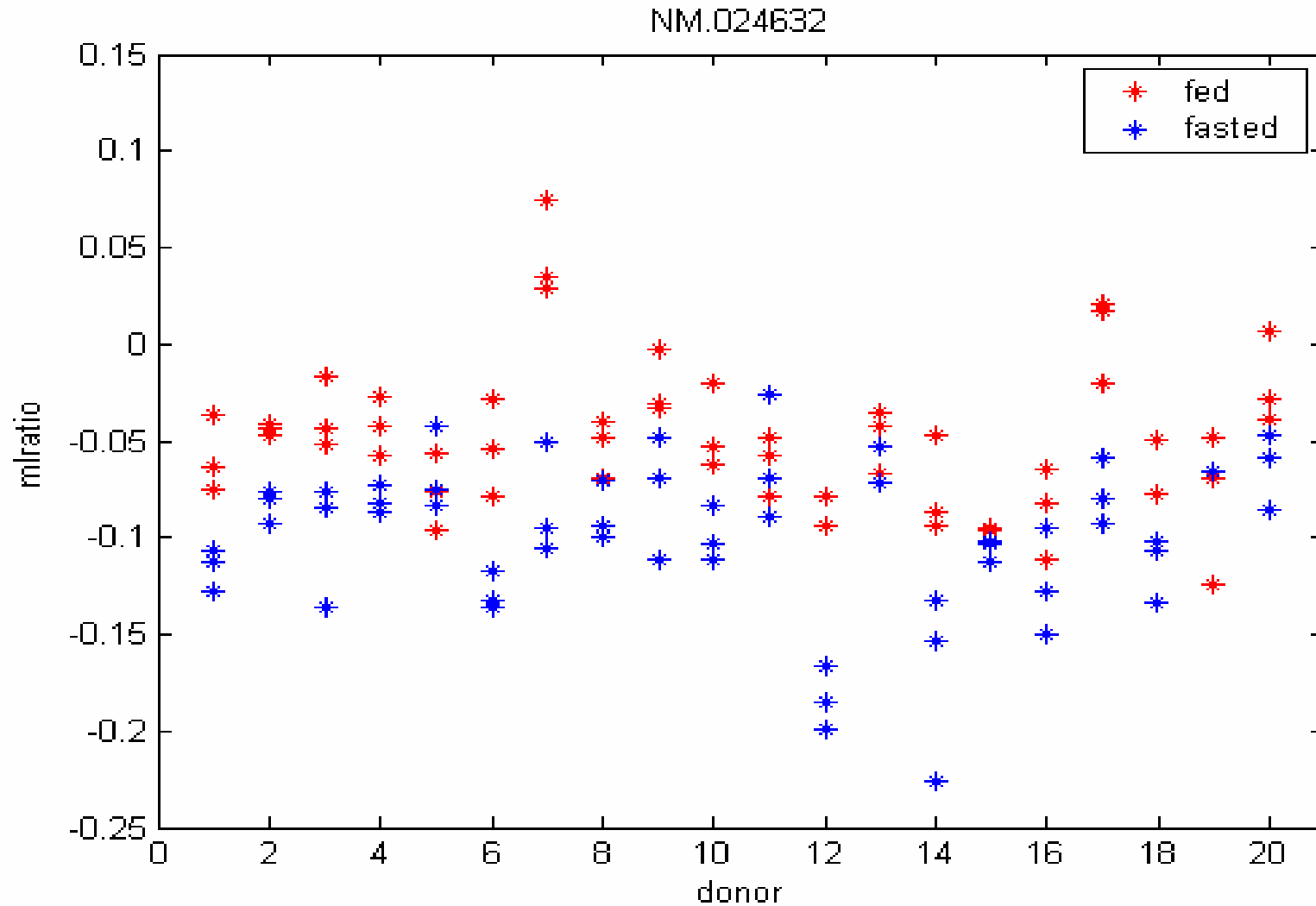
- Paired t-test or ARMAX-like procedures for “trend”
- Classification Algorithms
 - SVM (Support Vector Machine)
 - Random Forest

Differences among time points, conditional on previous time points

Genes that discriminate fast/fed status

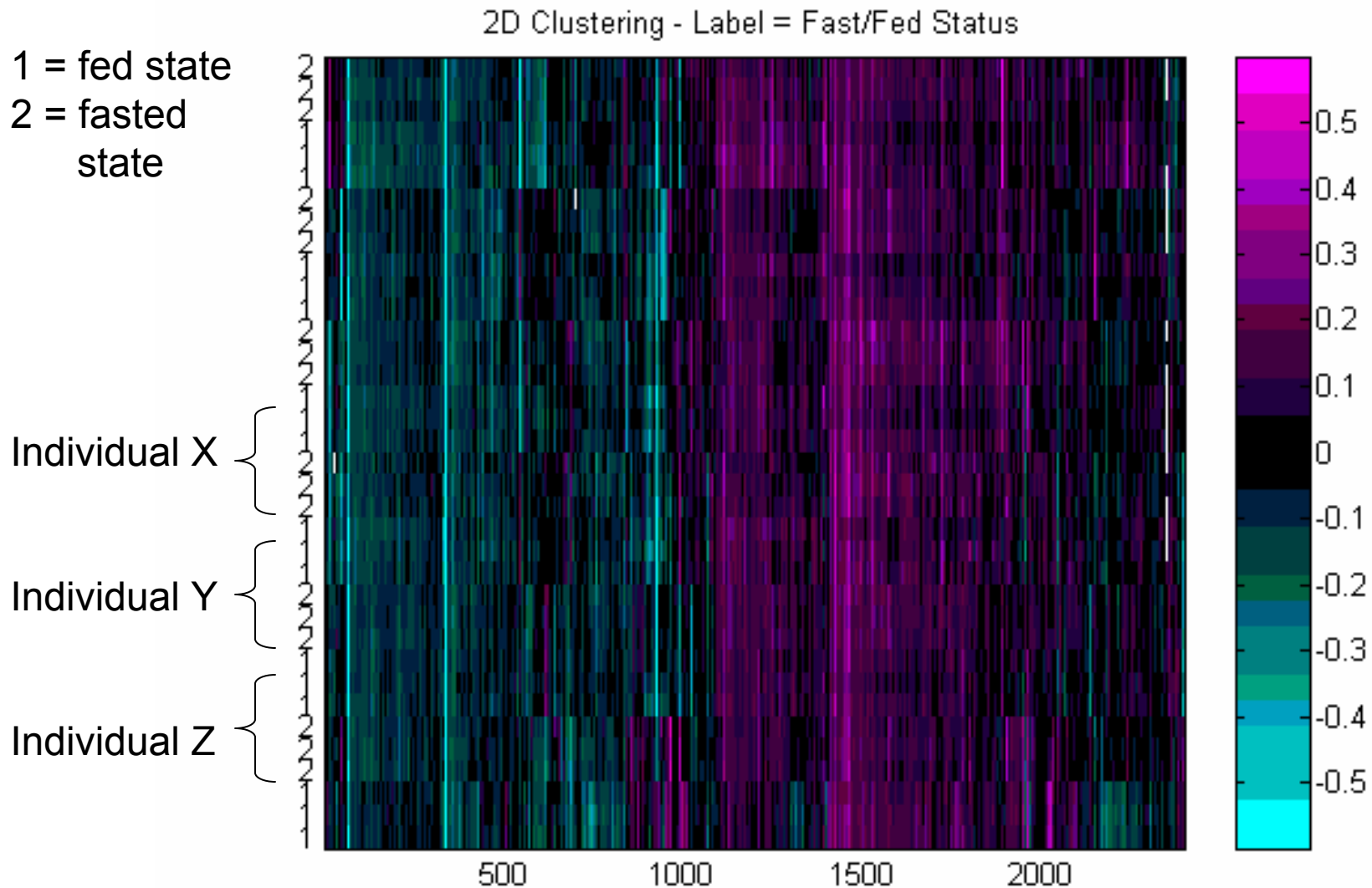
Time point	Number of Genes that Discriminate Fast/Fed Status		
	pval < .05	pval < .01	permuted (pval < .01)
1	216	39	23
2	320	74	54
3	769	253	20
4	1063	353	18
5	1748	848	26
1 and 2	436	185	70
3,4,and 5	3823	2474	86

Sample Gene: Donor vs. mlratio



There is a clear difference between fasting and feeding status.

2D Clustering of Fasting/Feeding signatures are dominated by donor specific differences



Standard classification packages are used to build classifiers

- Two Classification algorithms were explored to try to classify fast/fed status based on gene-expression.
 - SVM was found to perform best with a classification accuracy rate of 94%.
 - Random Forest was tried but did not perform as well, with 88% classification accuracy.
 - Classification accuracy on randomized data was 52.5% (almost exactly what was expected by chance)
- The SVM classifier identified 361 genes that were most predictive of fast/fed status
- This classifier was trained on the fast/fed data and then tested in a completely independent dataset
 - The classification accuracy in this independent dataset was 91%.
 - Accuracy on randomized data was 52.9%
- The 361 genes were rank ordered according to their ability to predict fast/fed status, and characterized in

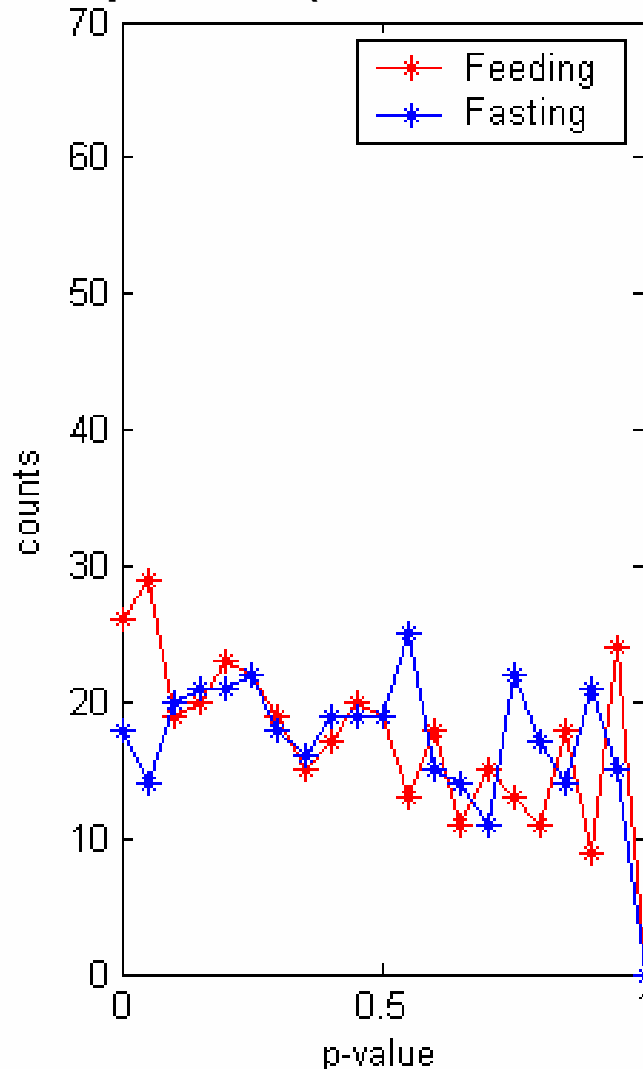
Characterization of 361 Genes : Heritability

Number of Significant Genes

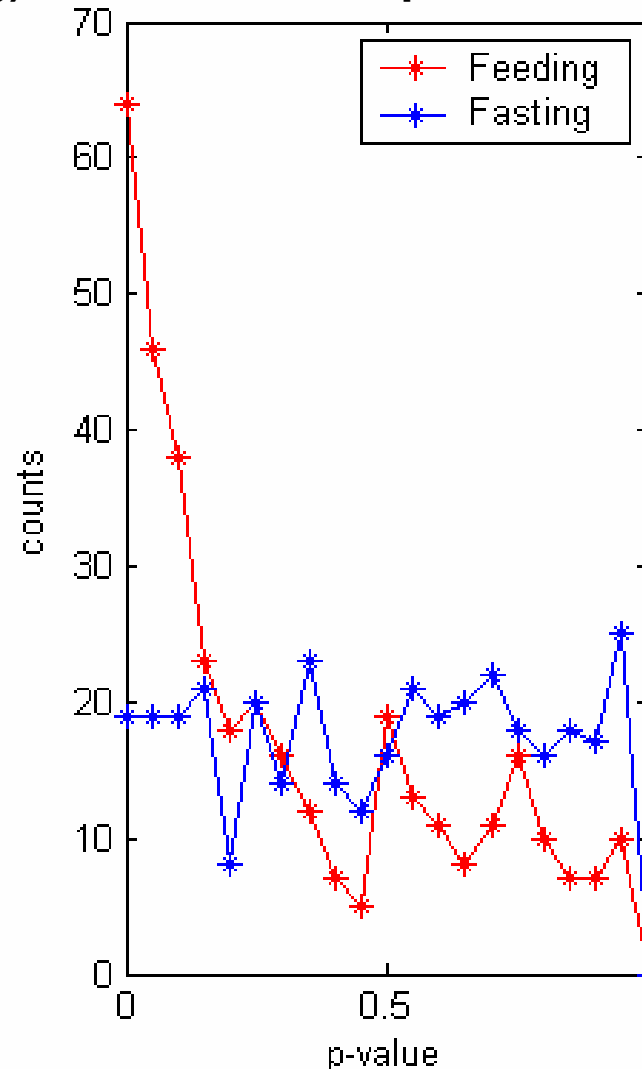
	pvalue < .05	pvalue < .01	pvalue < .001
361 genes that classify fast/fed status	252 (70%)	198 (55%)	113 (31%)
All (23720) Genes	8591 (36%)	5957 (25%)	3600 (15%)

Density of P-value Correlations of gene expression with PBF (Percent Body Fat) : Two time points

Time point #1 (both arms fasting)

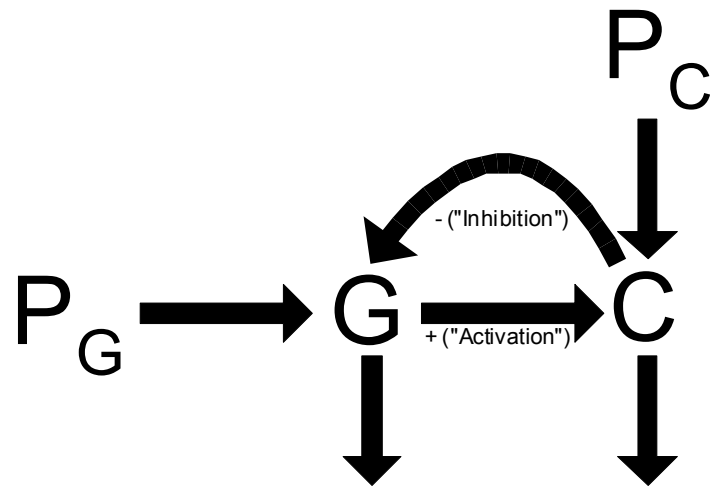


Time point #5



Modeling Feedback Control

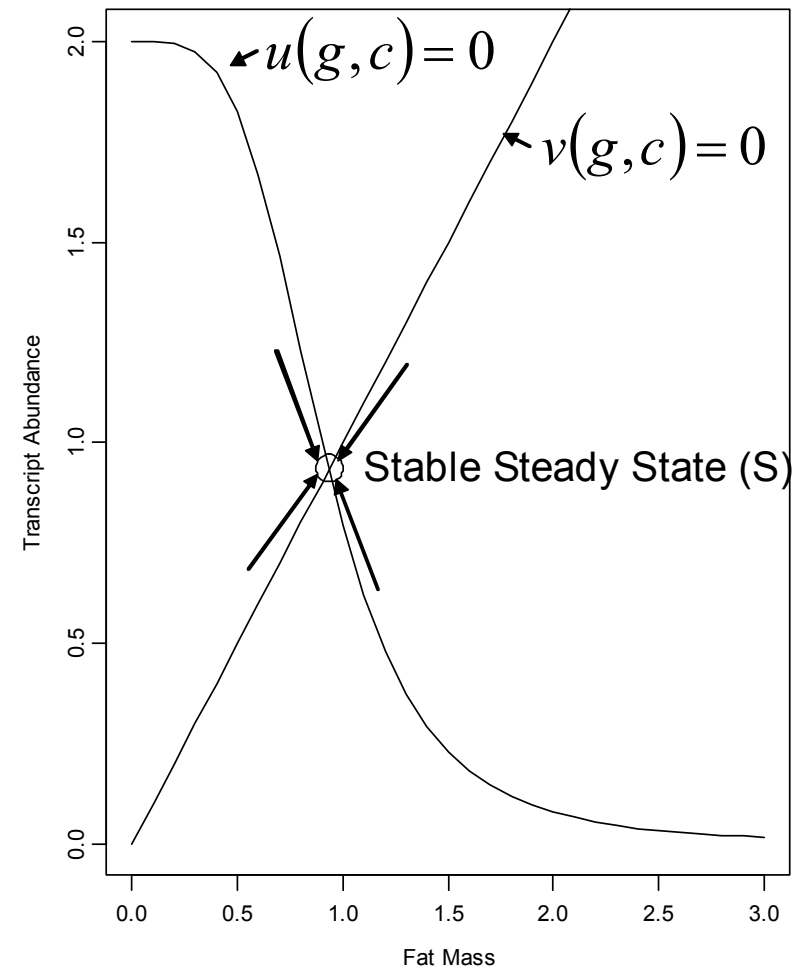
A)



$$\frac{dg}{dt} = \frac{\theta^n}{\theta^n + c^n} - \alpha g = u(g, c)$$

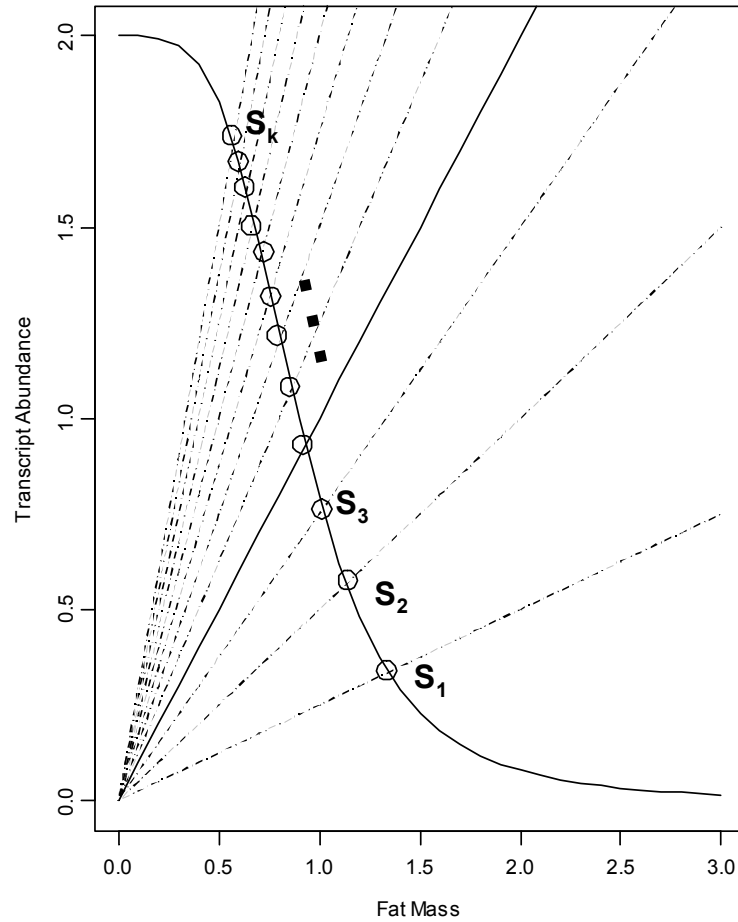
$$\frac{dc}{dt} = \beta g - \delta c = v(g, c)$$

B)

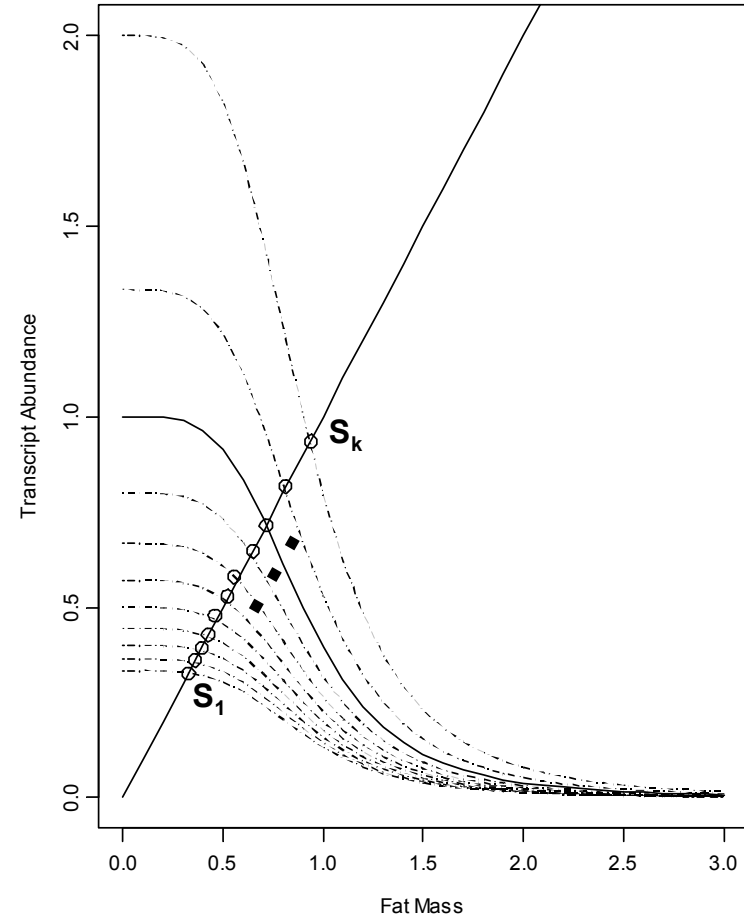


Location of the perturbations drives the significance and sign of the correlation between traits

A)

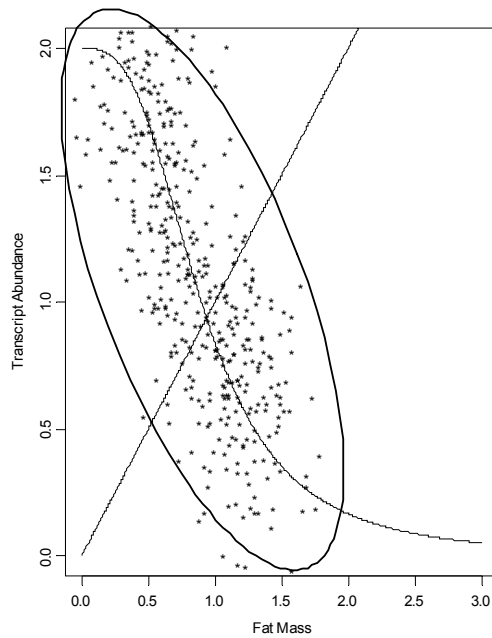


B)



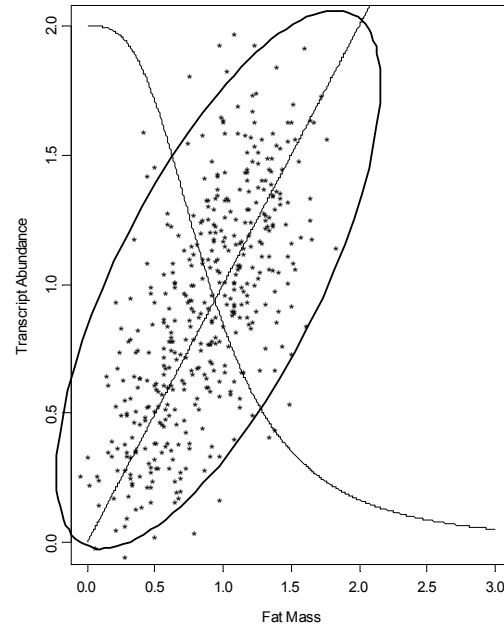
Perturbations on
C dominate

A)



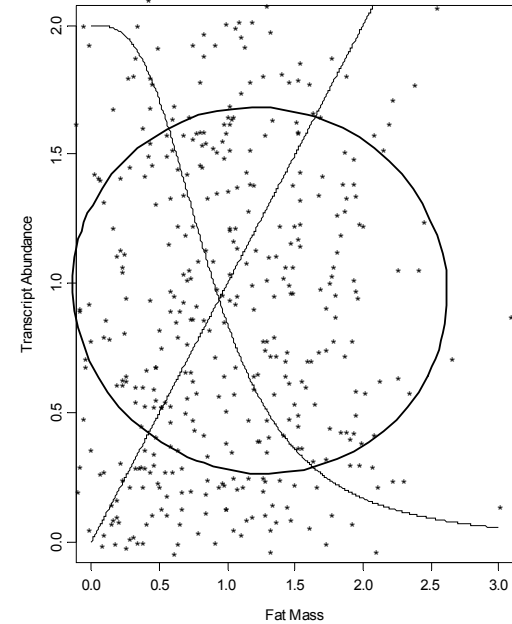
Perturbations on
G dominate

B)



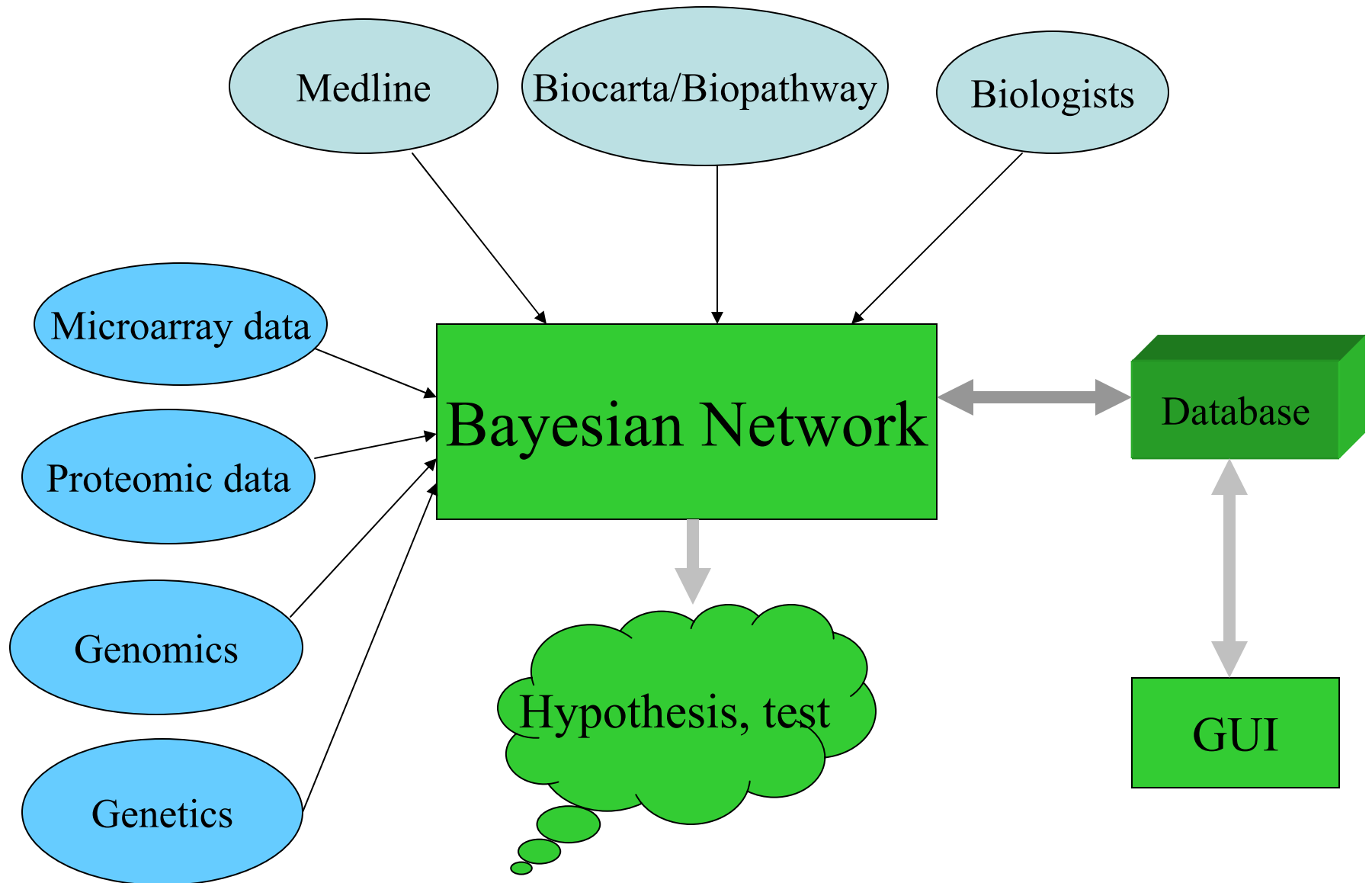
No dominant
perturbation pattern

C)

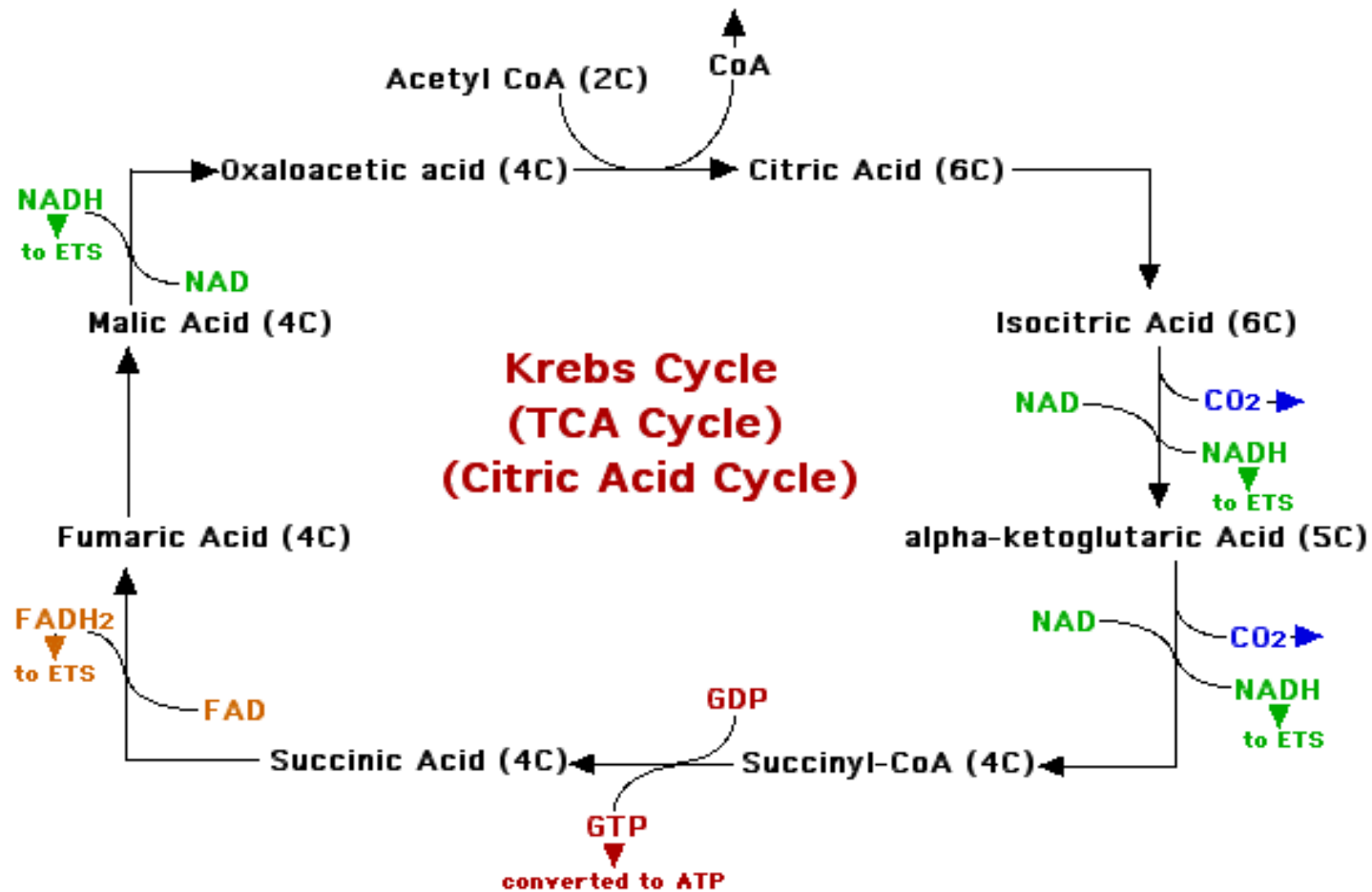


- Suggests a couple of different ways to identify traits that may be related in this way:
 - Look for reversal in sign of correlation in extremes of a given population
 - Look for reversal in sign of correlation between different populations

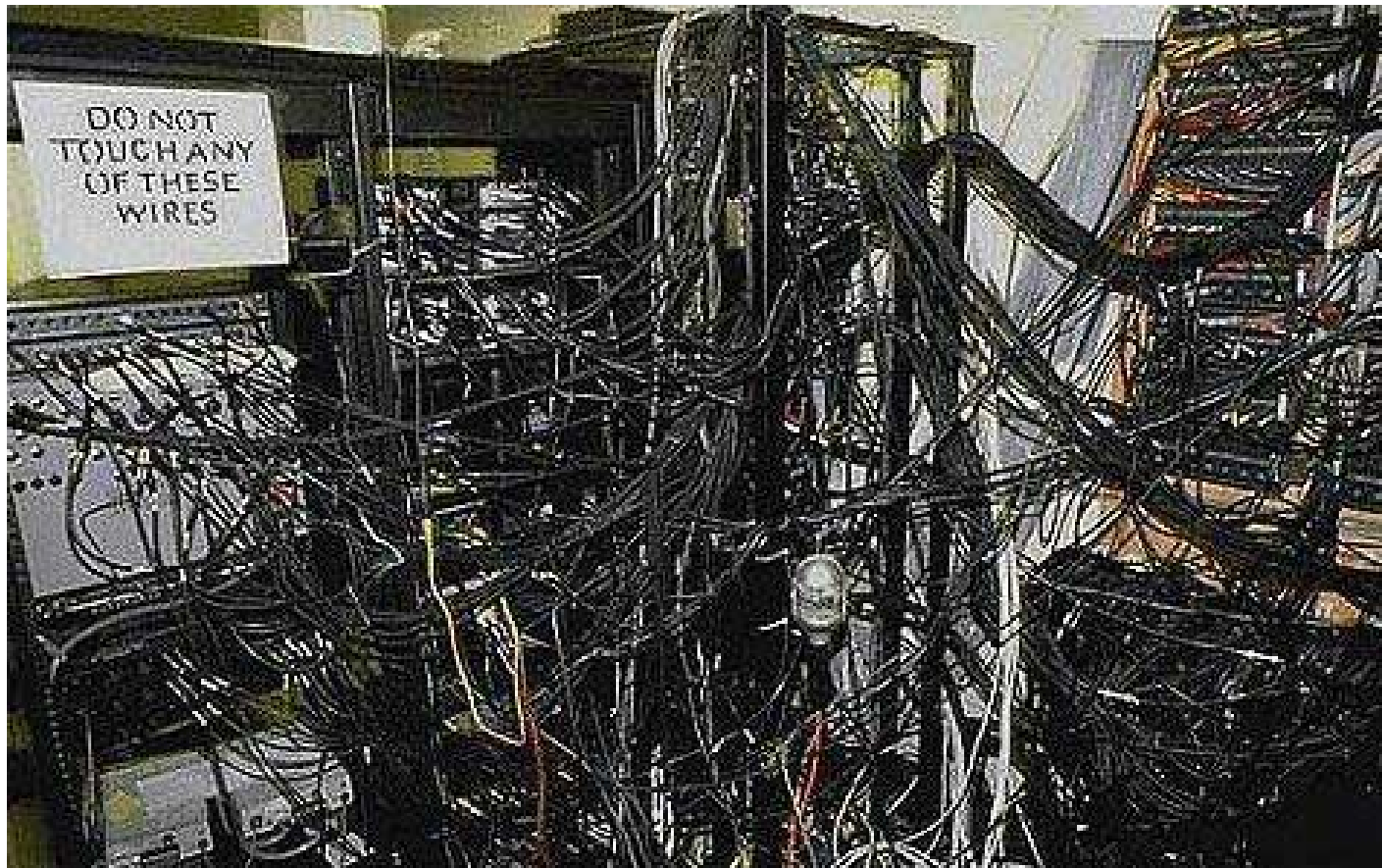
Genetics Network as a Frame Work



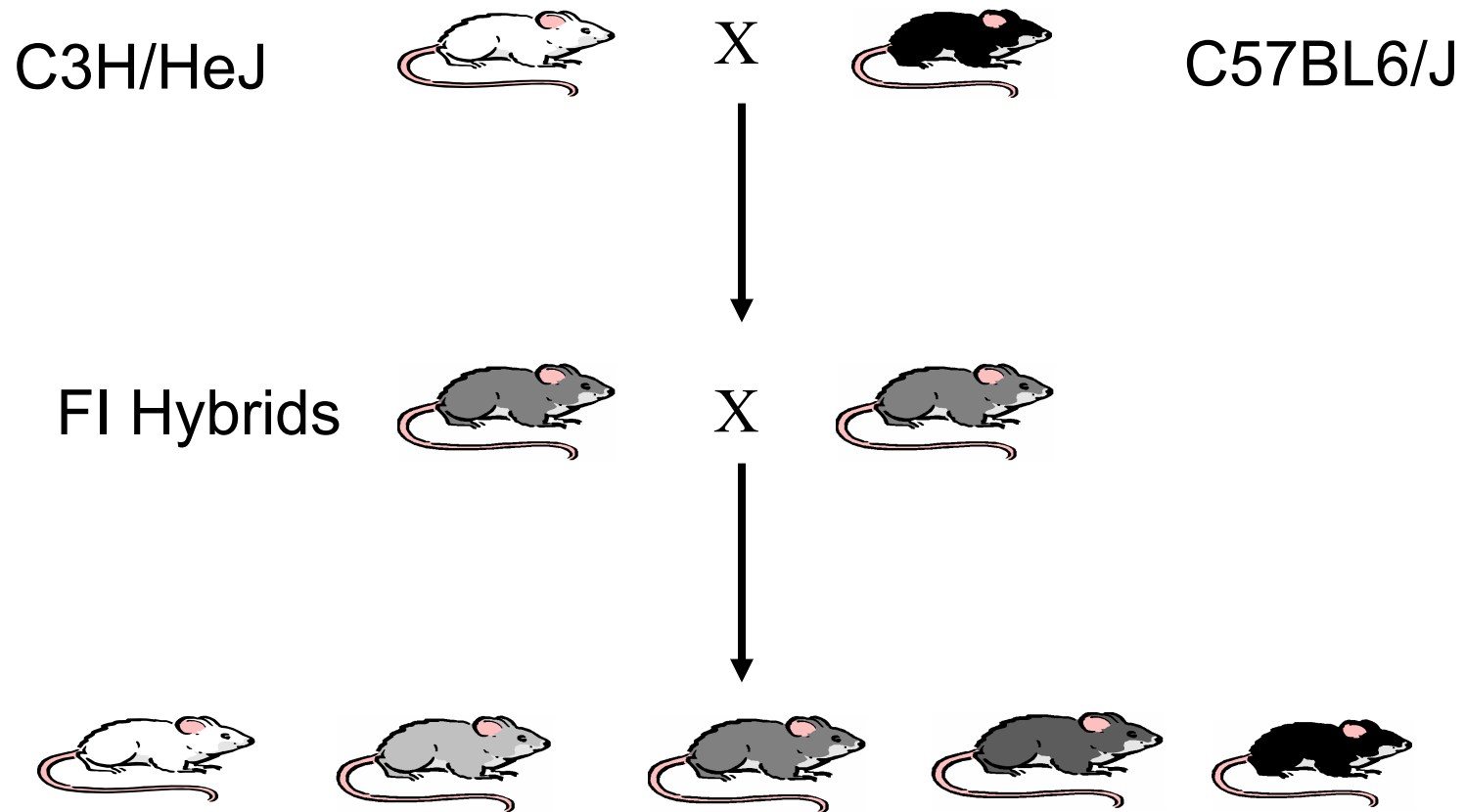
What is the true structure of biological pathways?
Could they be as “simple” as the TCA cycle?



OR...



QTL mapping of genes for obesity related traits in a standard F2 intercross



F2 Intercross Mice

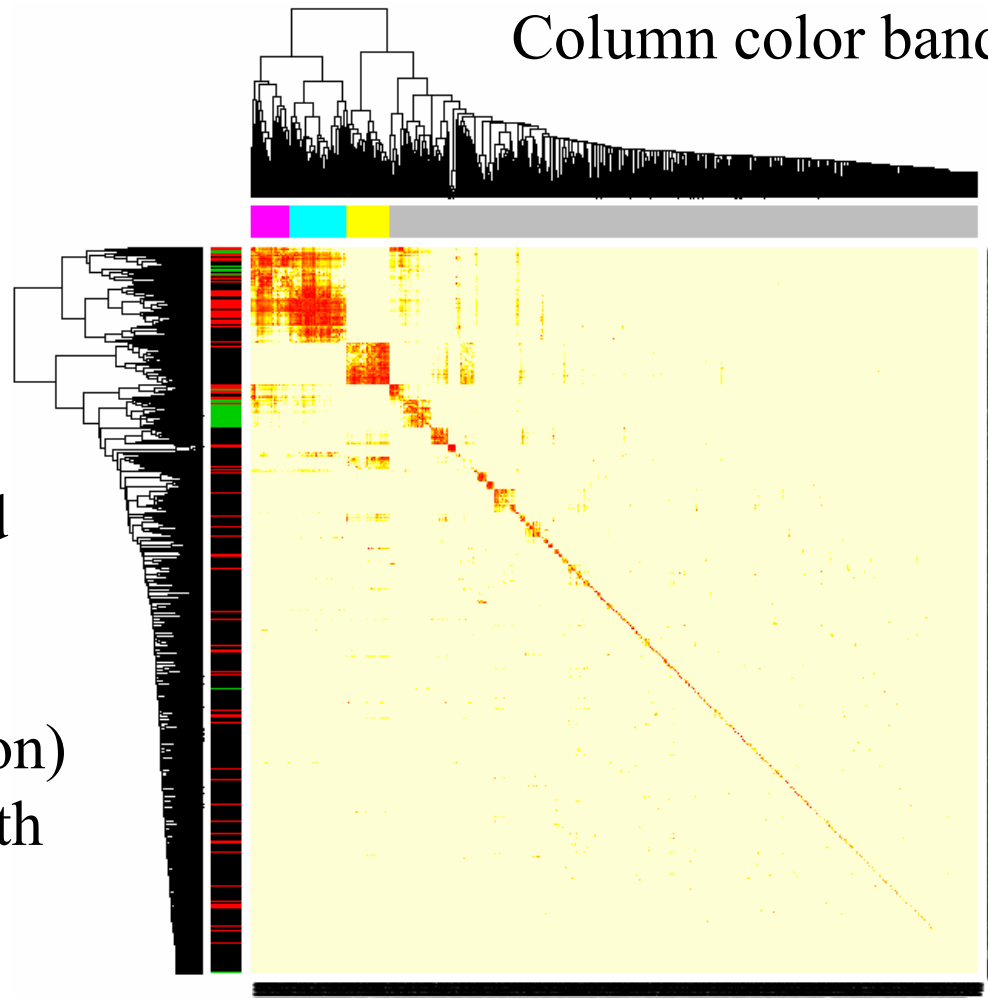
All mice put on a high-fat diet at 8 weeks of age for 14 weeks

Co-expression networks

- Gene co-expression networks have two important properties:
 - Scale-free topology (existence of few highly connected hub genes)
 - A high degree of clustering
- The co-expression networks can be decomposed into modules that have their own hubs
- Within each module, we find that the hub genes are most relevant for predicting the clinical traits
- Genetic markers offer the opportunity to add directions to the gene networks, i.e. causality
- Software implemented in Matlab, C++, and R
 - Separate CoExpress NetGen software packages developed in house
 - Collaborations with many others to integrate: Jun Liu at Harvard, Steve Horvath at UCLA, many interactions with ISMB

Again scripts written in matlab are used to explore topological properties of gene networks

Column color band indicates modules



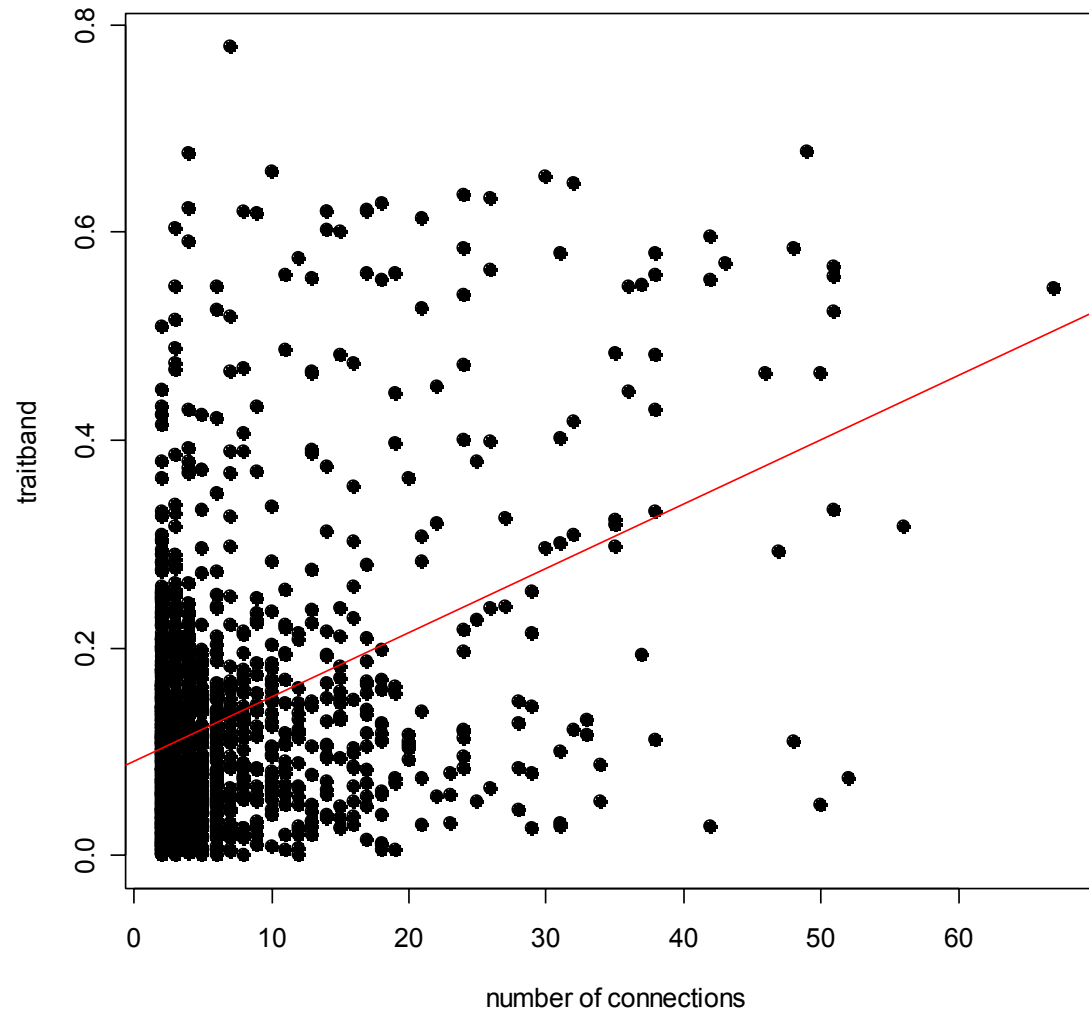
Row color band indicates how significant Gene (expression) is correlated with the clinical trait

Total
1930
= 105(pink)
+151(turquoise)
+113(yellow)
+1564(grey)

Relationship between trait band and connectivity: grey group

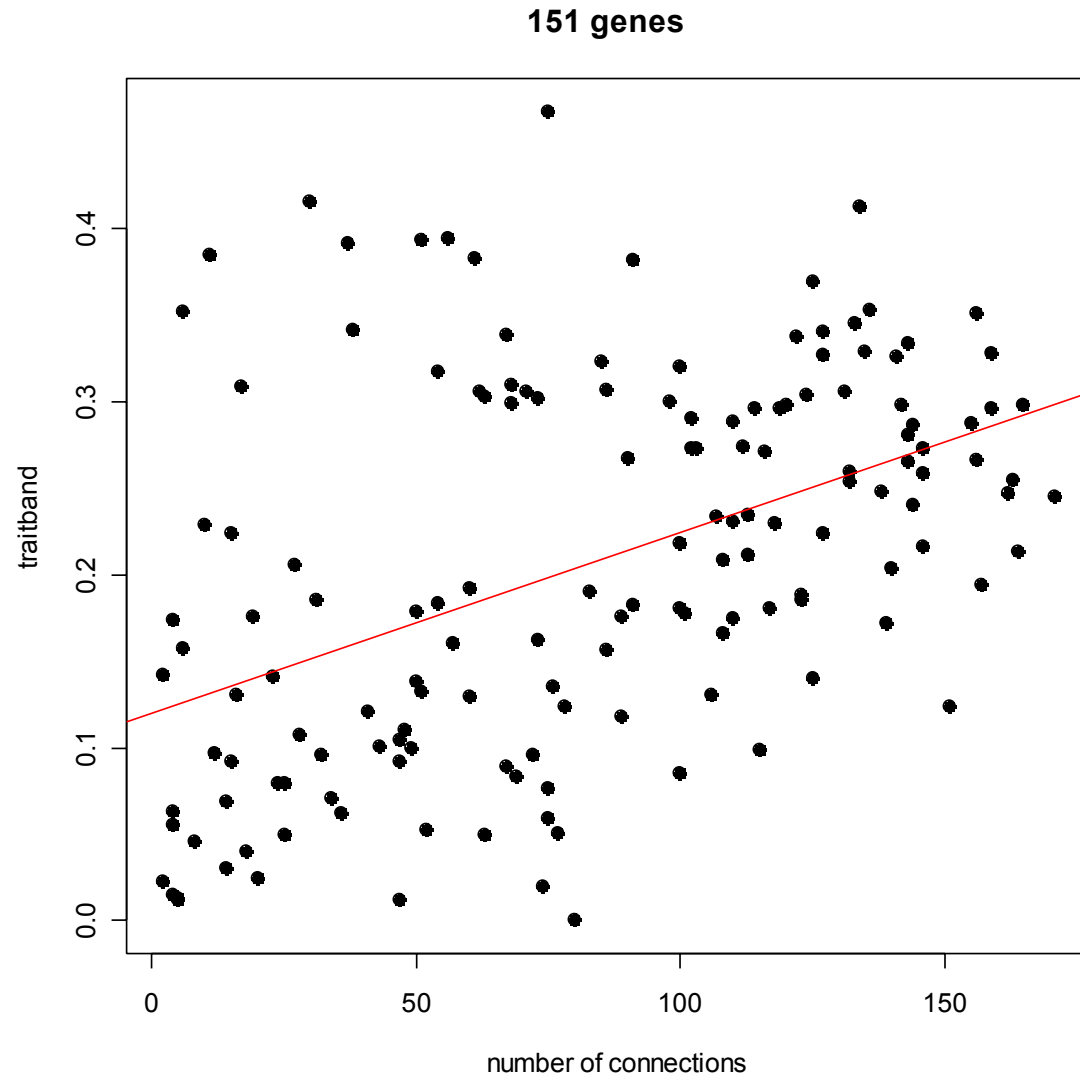
1564 genes in gray

Adjusted R-square
= 0.16;
p-value=2.2e-16



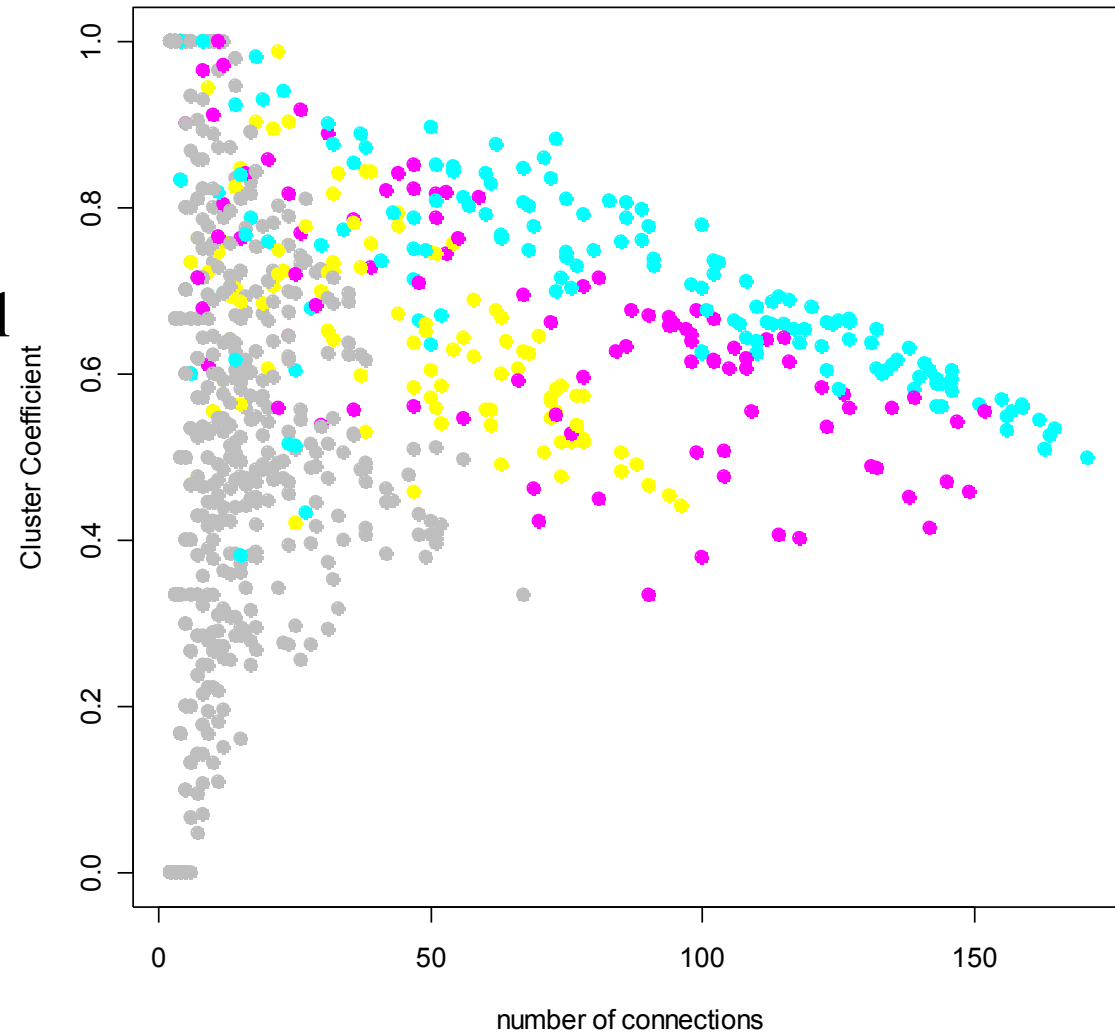
Relationship between trait band and connectivity: turquoise

Adjusted R-square = 0.21;
p-value=2.1e-9



Cluster coefficient versus connectivity colored by module

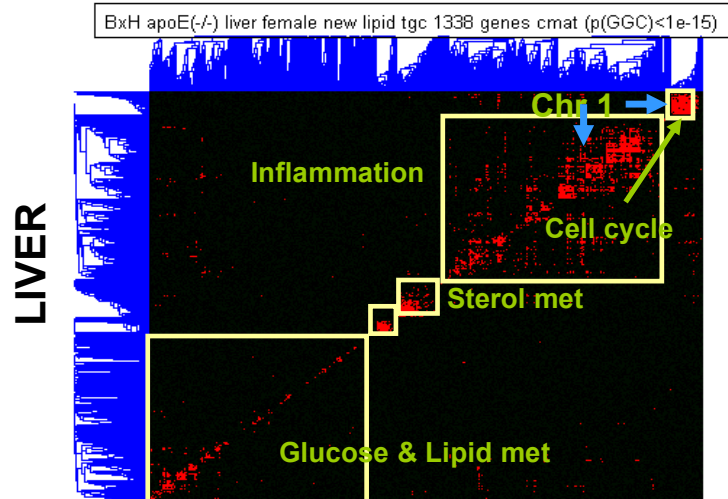
all 1930 genes



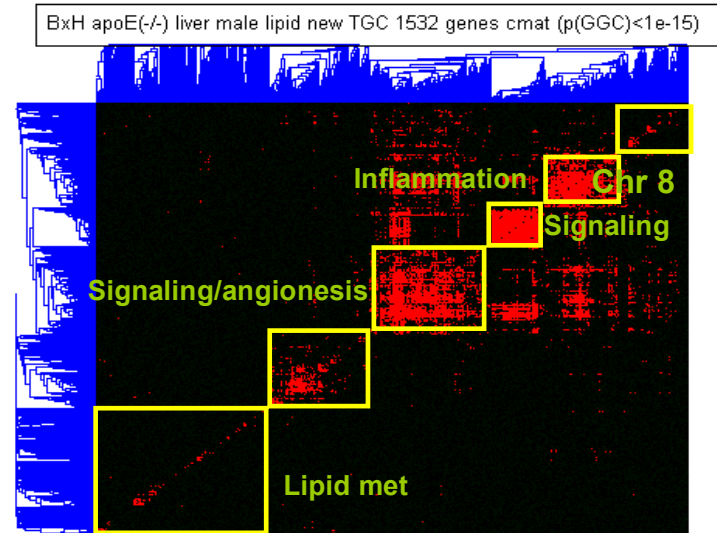
Message: hierarchical organization of the biological modules

Genes in modules identified are then intersected with “known” pathway sets (large database, includes Ingenuity, GeneGo, GO, and others) to search for enrichment using naïve methods (e.g., the Venn Diagram method or slightly improved rank ordering schemes

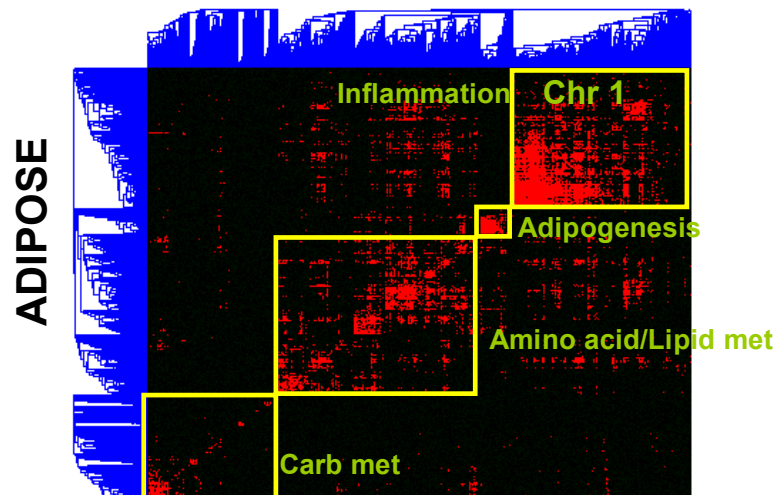
FEMALES



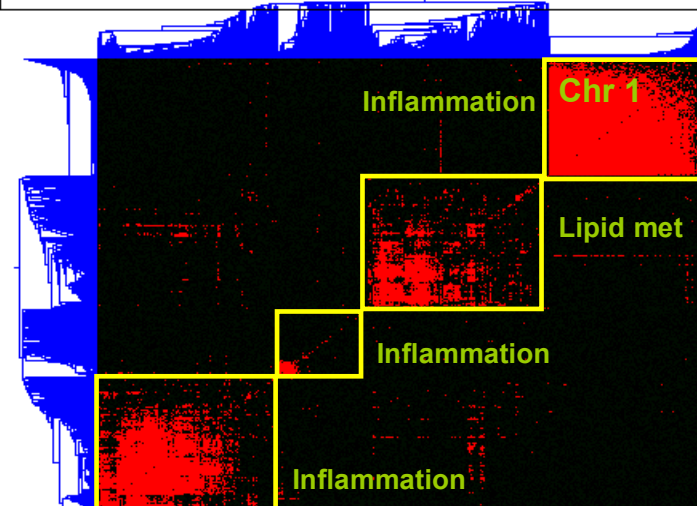
MALES



BxH ApoE(-/-) lipid trait correlated 3414 gene cmat (p(GGC)<1e-15)

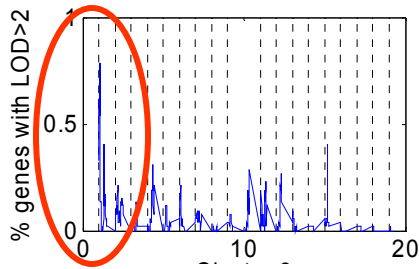


BxH apoE(-/-) adipose male lipid new TGC 2352 genes cmat (p(GGC)<1e-15)

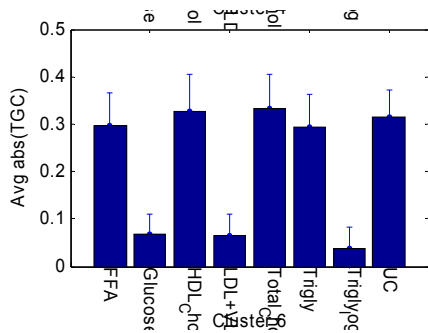


Genes with correlations to lipid traits form modules that are enriched for linkage to specific chromosomes: Determined by sets of SQL queries intersecting QTL results with co-expression results

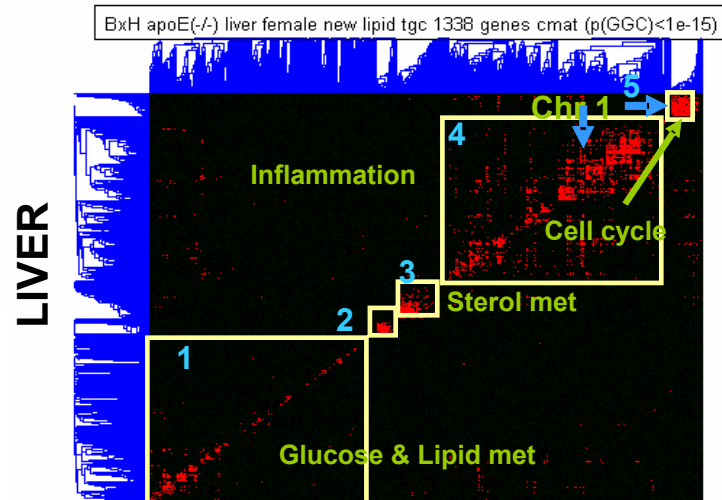
Inflammation cluster 4



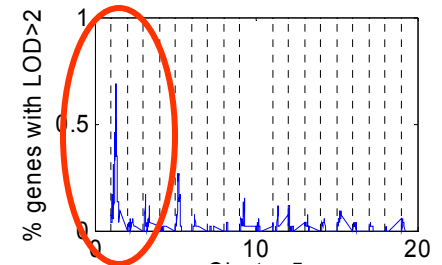
Chr 1



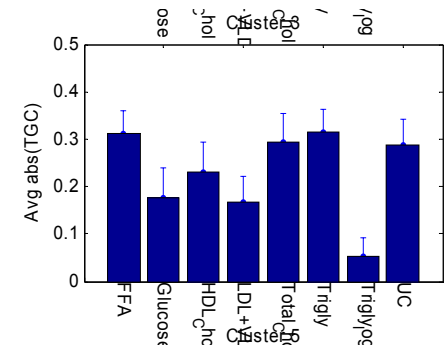
FEMALES

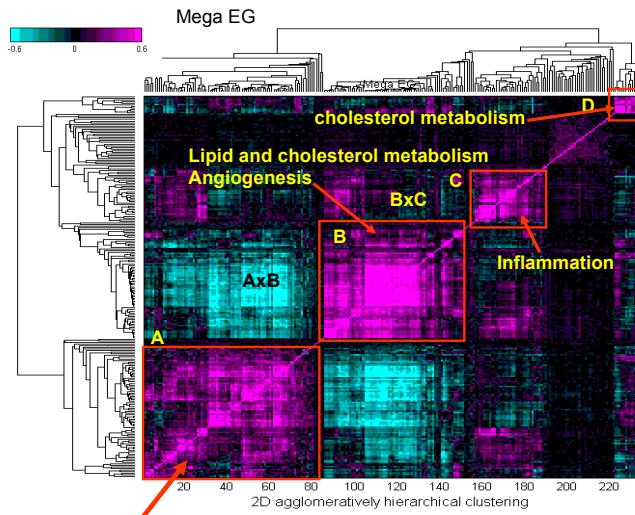


Cell cycle cluster 5

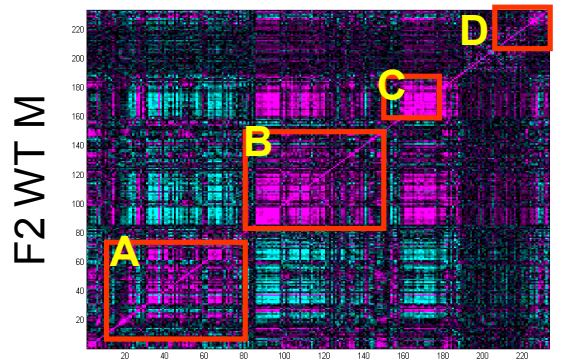
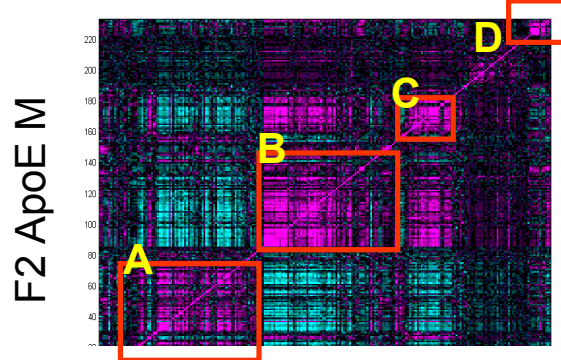


Chr 1

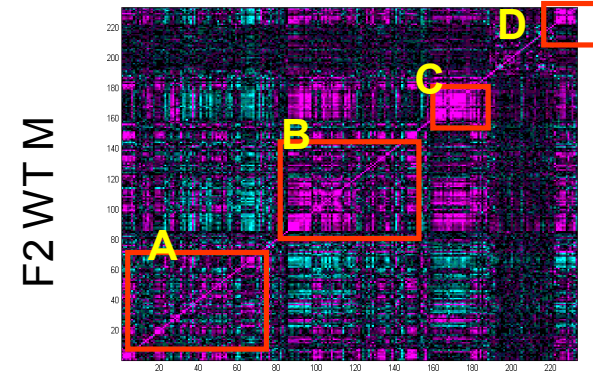
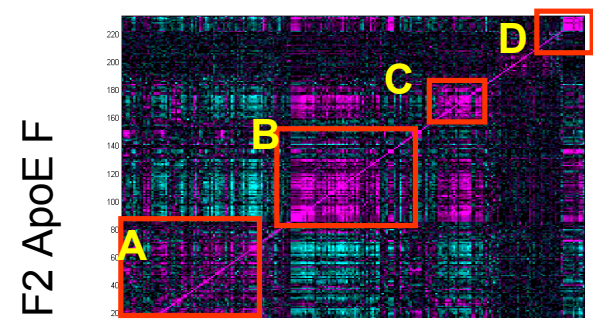




Inflammation and steroid metabolism



These quasi-automated procedures allow for extremely large-scale mining of the data (the search for patterns that mean something). Here we see sub-networks that are stable and do not rewire across large numbers of perturbations



The Bayesian network approach

- Bayesian networks are directed acyclic graph
- Imposes constraints on the joint probability distribution of the nodes (random variables) in the graph so that it decompose as (based on conditional independence):

$$p(G) = p(X_1, X_2, \dots, X_n) = \prod_i p(X_i | Pa(X_i))$$

- Aim is to find the graph G that maximizes the likelihood given the data $D - p(D|G)$

BN: Markov equivalent

- The Bayesian network itself does not reveal causal information

T1→T2

T2→T1

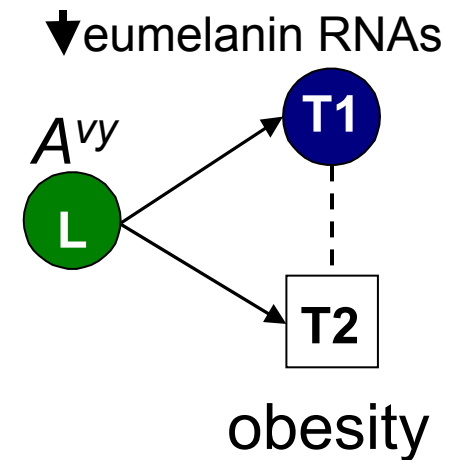
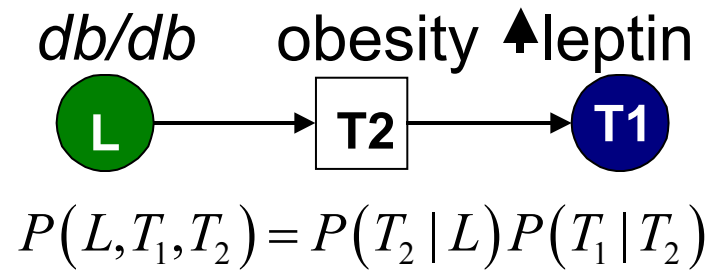
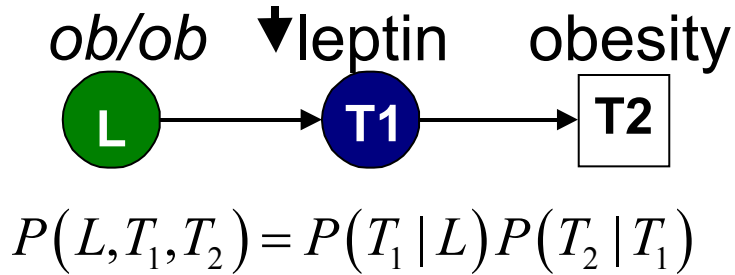
$$p(T_1, T_2) = p(T_2 / T_1) p(T_1) = p(T_1 / T_2) p(T_2)$$

Distinguishing Causal from Reactive Genes

Causative Model (M1)

Reactive Model (M2)

Independent Model (M3)



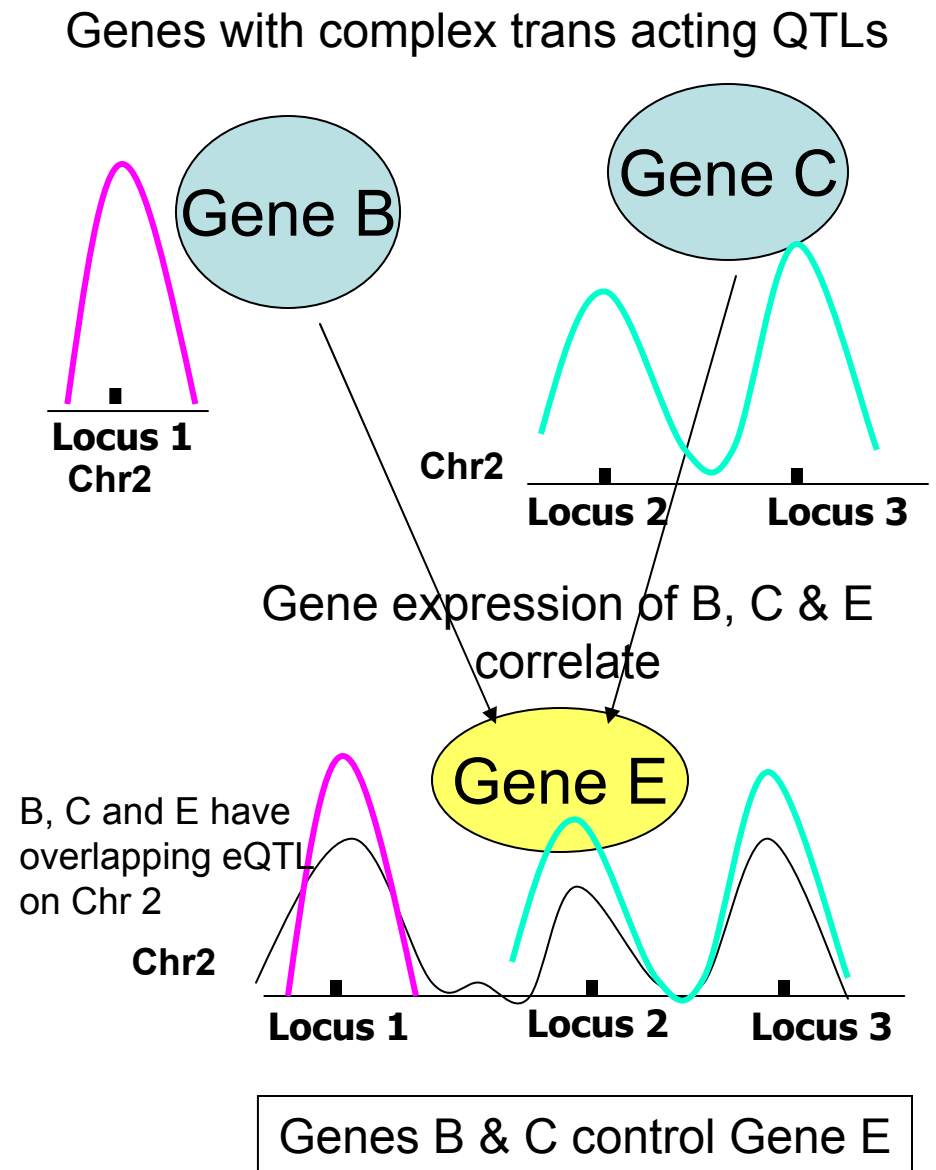
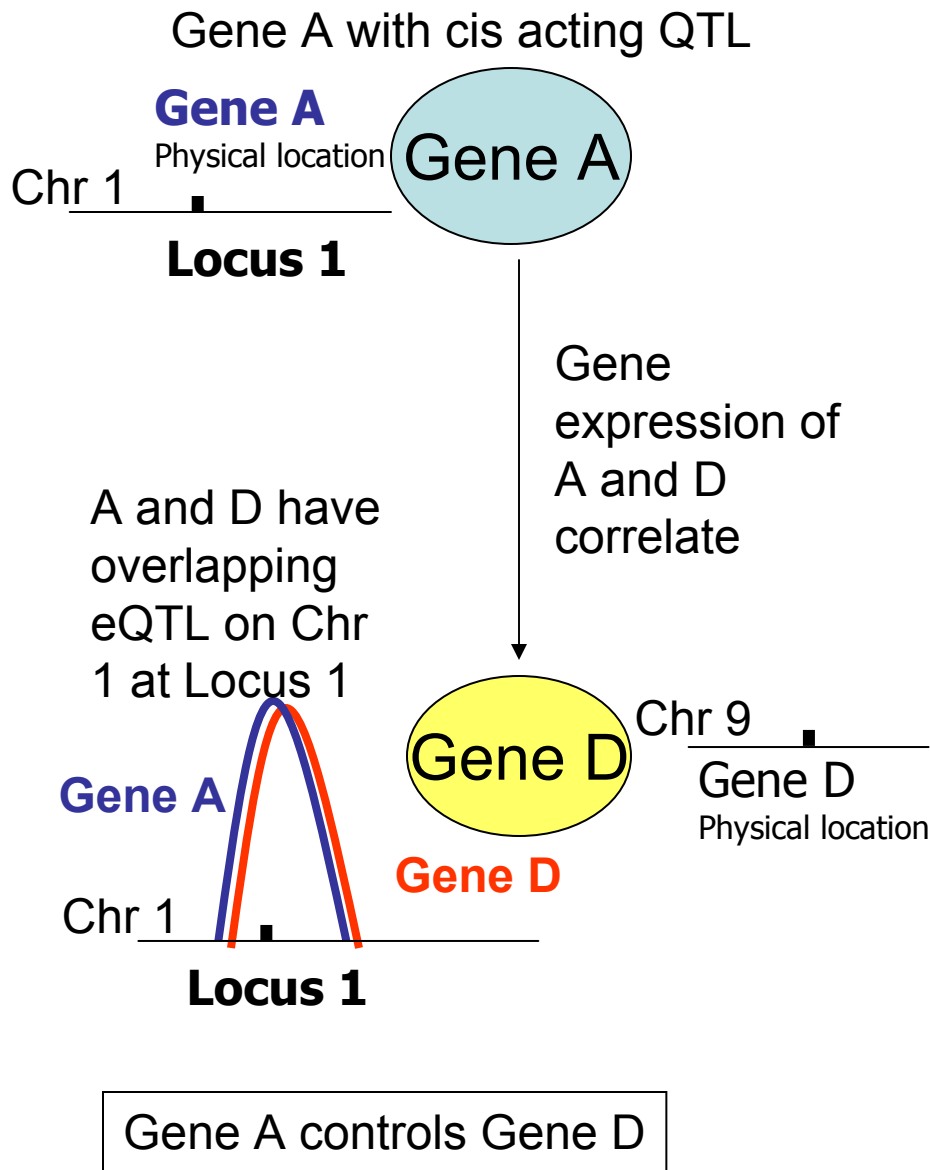
$$P(L, T_1, T_2) = P(T_2 | L)P(T_1 | L)$$

L DNA Locus controlling RNA levels and/or clinical traits

T1 Quantitative trait 1

T2 Quantitative trait 2

Network Reconstruction using Gene Expression and Genetics in a nutshell



For the different joint probability distributions for the different models:

$$\text{M1)} \quad P(L, T_1, T_2) = P(L)P(T_1 | L)P(T_2 | T_1)$$

$$\text{M2)} \quad P(L, T_1, T_2) = P(L)P(T_2 | L)P(T_1 | T_2)$$

$$\text{M3)} \quad P(L, T_1, T_2) = P(L)P(T_1 | L)P(T_2 | T_1, L)$$

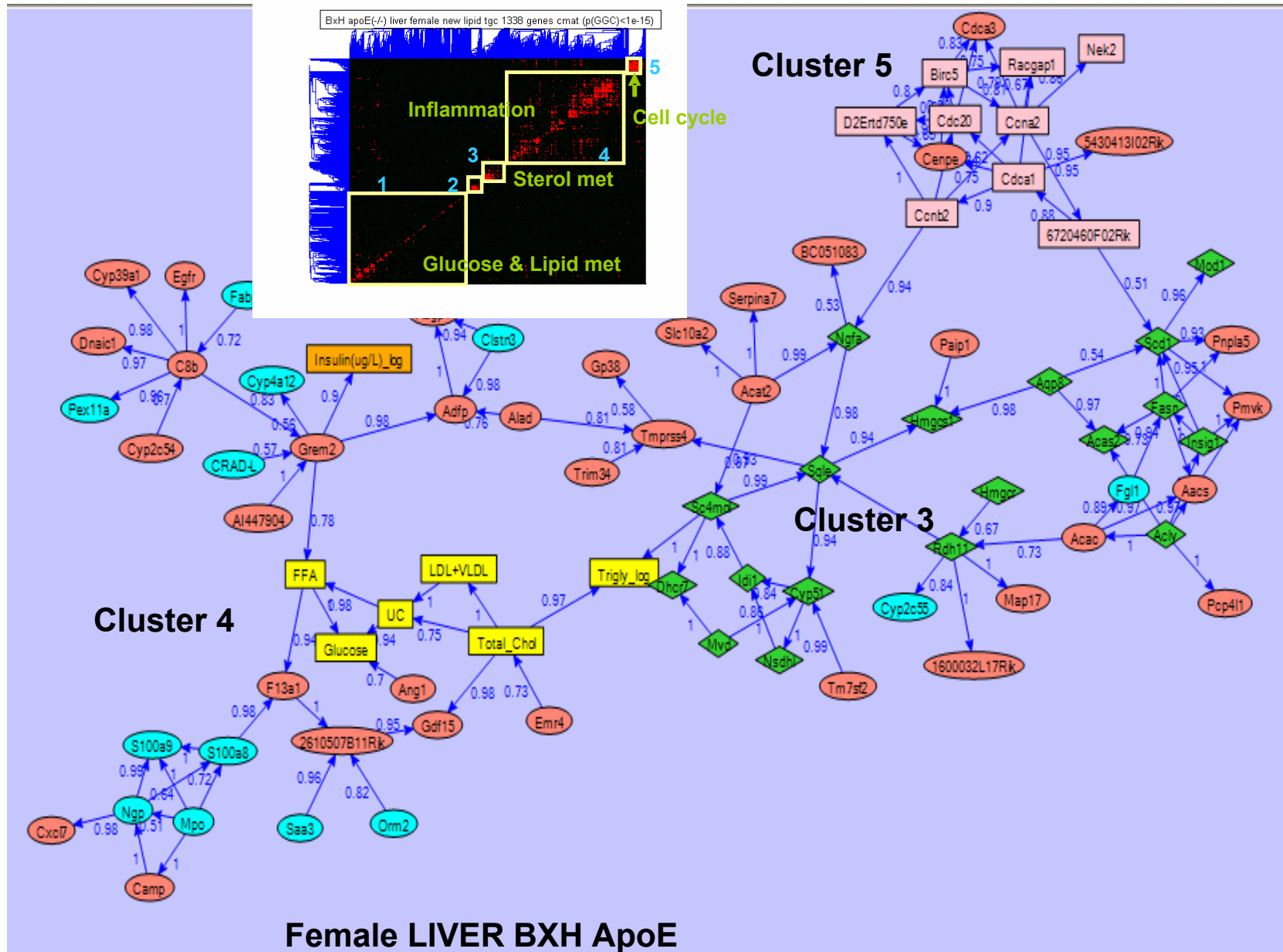
We assume the traits are normally distributed about each genotypic mean at the common locus L, with mean and variance for each component given by:

	$P(T_1 L)$	$P(T_2 L)$	$P(T_1 T_2)$
Mean	$E(T_1 L) = \mu_{T_{1L}}$	$E(T_2 L) = \mu_{T_{2L}}$	$E(T_1 T_2) = \mu_{T_1} + \rho \frac{\sigma_{T_1}}{\sigma_{T_2}} (T_2 - \mu_{T_2})$
Variance	$Var(T_1 L) = \sigma_{T_1}^2$	$Var(T_1 L) = \sigma_{T_2}^2$	$Var(T_1 T_2) = (1 - \rho^2) \sigma_{T_1}^2$

BN: reconstruction and prediction

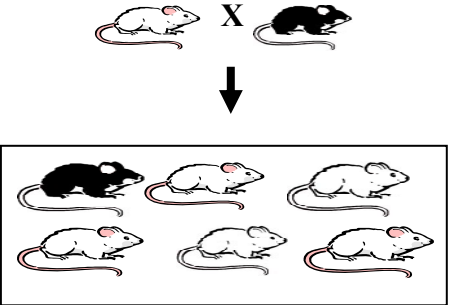
- 10,000 to 100,000 networks are reconstructed using random seeds
 - Computationally intense procedure
 - Presently runs on a 500 CPU IBM Blade Cluster
- Common features are then extracted (e.g., connections seen in $> 60\%$ of the networks are extracted) and probability tables are updated
- Prediction is then based on maximum likelihood state
- Our algorithm to search through possible models is similar to the local maximum search algorithm implemented by Friedman et al. (2000), *J. Comput. Biol.* 7:601-620
- Software internally developed (earlier version based on more heuristic use of genetic data was described in *Cytogenet. Genome Research* 105:363-374 (2004))

Visualization of the network structures is a significant effort; Tom Sawyer software and Cytoscape are being extensively modified for our purposes



Simulations are an important component of our validation procedures

Segregating Mouse Populations

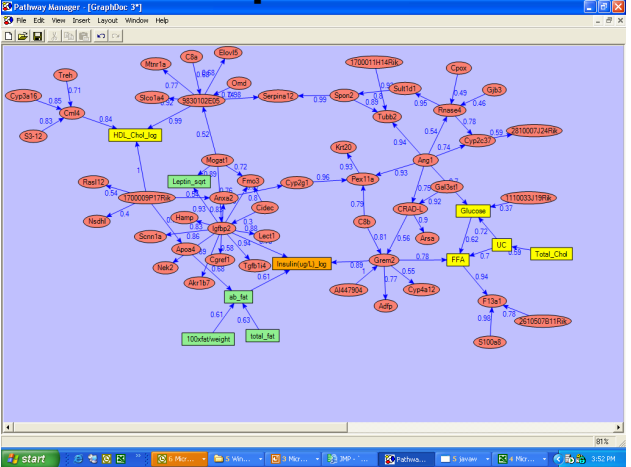


Disease Tissue

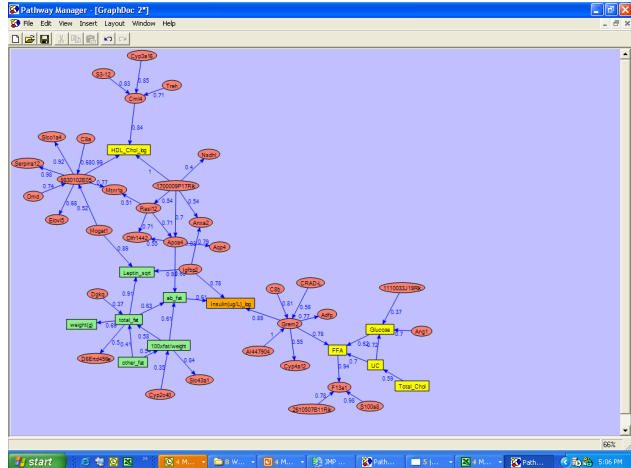
Normal Tissue

Genetics allows for increased power to reconstruct networks

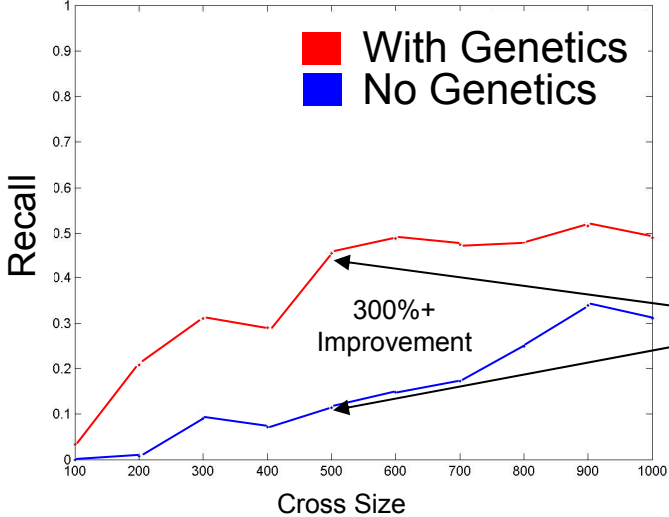
Disease-specific Network



Network for Normal Tissue



Recall Given at 80% Precision



Genetics has potential to dramatically improve the accuracy of the reconstructed network (greater than 300% improvement in recall with genetics vs. without)