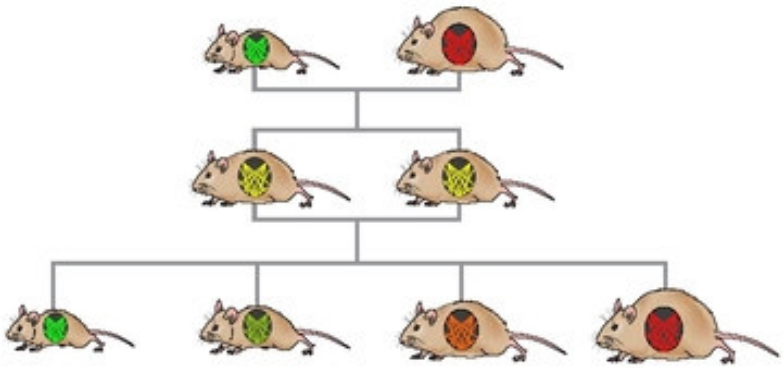
Integrating Genetic and Network Analysis to Characterize Genes Related to Mouse Weight

Steve Horvath
University of California, Los Angeles



Contents

- Brief review of gene network construction
- New terminology:
  - Gene significance based on body weight
  - Module quantitative trait locus (mQTL=eQTL hotspot for a given module)
  - Gene significance measure based on a SNP
- Characterize body weight related genes in mice

Important Task in Many Genomic Applications:

Given a network (pathway) of interacting genes (proteins) how to find the central players?

Which of the following mathematicians had the biggest influence on others?

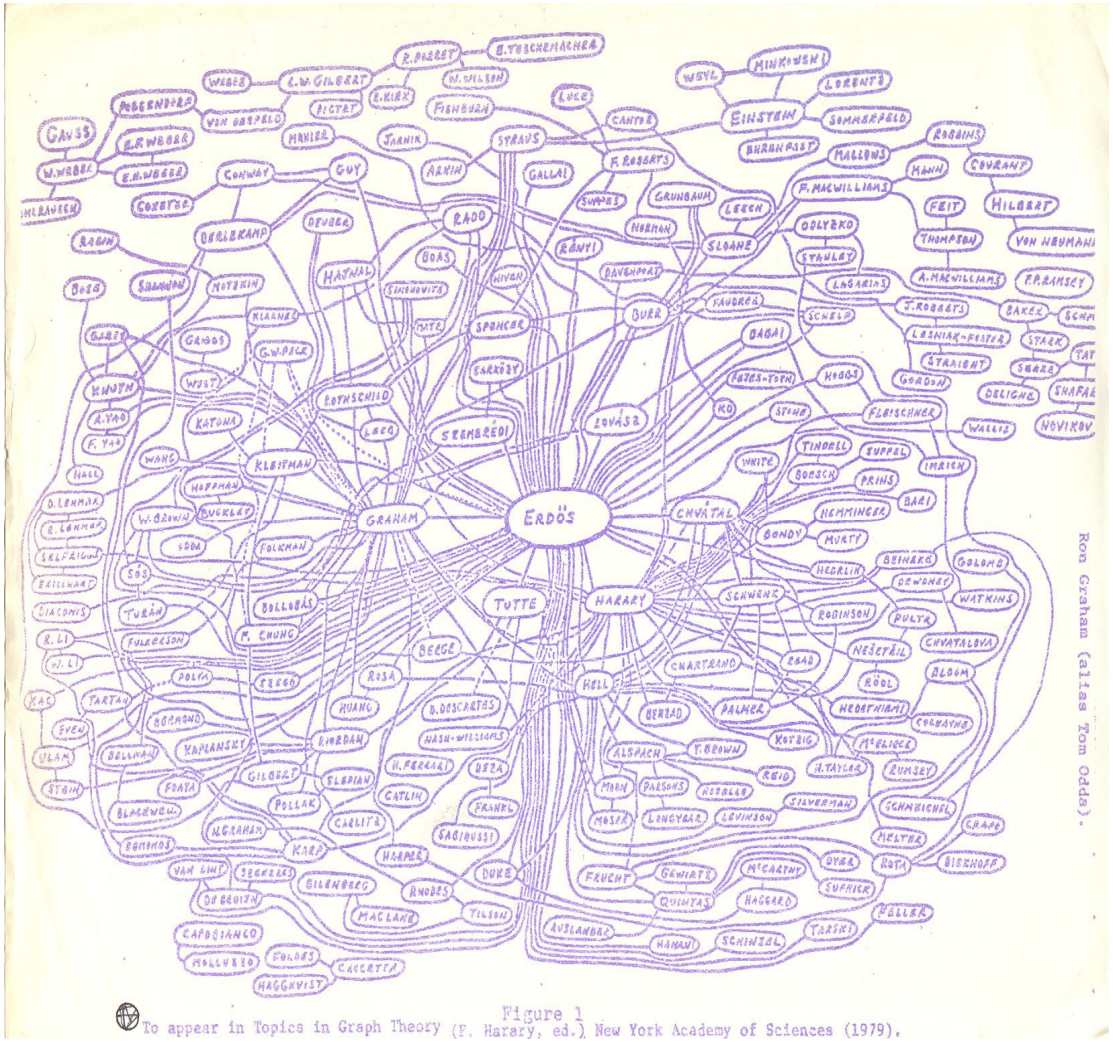


Figure 1
To appear in Topics in Graph Theory (F. Harary, ed.) New York Academy of Sciences (1979).

Connectivity can be an important variable for identifying important nodes

Network Construction

Bin Zhang and Steve Horvath (2005) "A General Framework for Weighted Gene Co-Expression Network Analysis", Statistical Applications in Genetics and Molecular Biology: Vol. 4: No. 1, Article 17.

Network=Adjacency Matrix

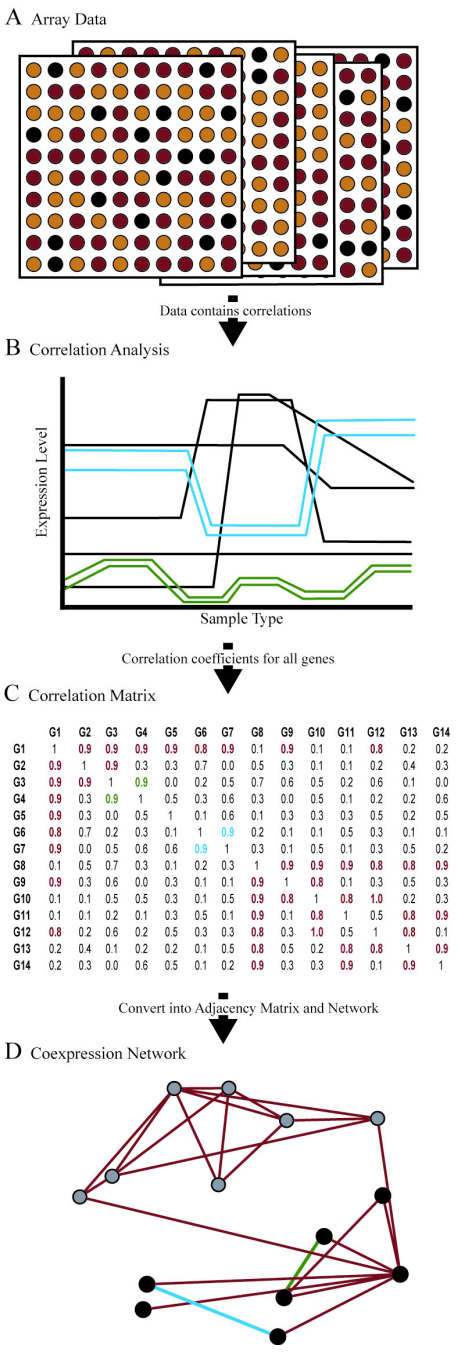
- A network can be represented by an adjacency matrix, $A=[a_{ij}]$, that encodes whether/how a pair of nodes is connected.
  - A is a symmetric matrix with entries in [0,1]
  - For unweighted network, entries are 1 or 0 depending on whether or not 2 nodes are adjacent (connected)
  - For weighted networks, the adjacency matrix reports the connection strength between gene pairs

Generalized Connectivity

- Gene connectivity = row sum of the adjacency matrix
  - For unweighted networks=number of direct neighbors
  - For weighted networks= sum of connection strengths to other nodes

$$k_i = \sum_j a_{ij}$$

Figure 1



	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14
G1	1	0.9	0.9	0.9	0.9	0.8	0.9	0.1	0.9	0.1	0.1	0.8	0.2	0.2
G2	0.9	1	0.9	0.3	0.3	0.7	0.0	0.5	0.3	0.1	0.1	0.2	0.4	0.3
G3	0.9	0.9	1	0.9	0.0	0.2	0.5	0.7	0.6	0.5	0.2	0.6	0.1	0.0
G4	0.9	0.3	0.9	1	0.5	0.3	0.6	0.3	0.0	0.5	0.1	0.2	0.2	0.6
G5	0.9	0.3	0.0	0.5	1	0.1	0.6	0.1	0.3	0.3	0.3	0.5	0.2	0.5
G6	0.8	0.7	0.2	0.3	0.1	1	0.9	0.2	0.1	0.1	0.5	0.3	0.1	0.1
G7	0.9	0.0	0.5	0.6	0.6	0.9	1	0.3	0.1	0.5	0.1	0.3	0.5	0.2
G8	0.1	0.5	0.7	0.3	0.1	0.2	0.3	1	0.9	0.9	0.9	0.8	0.8	0.9
G9	0.9	0.3	0.6	0.0	0.3	0.1	0.1	0.9	1	0.8	0.1	0.3	0.5	0.3
G10	0.1	0.1	0.5	0.5	0.3	0.1	0.5	0.9	0.8	1	0.8	1.0	0.2	0.3
G11	0.1	0.1	0.2	0.1	0.3	0.5	0.1	0.9	0.1	0.8	1	0.5	0.8	0.9
G12	0.8	0.2	0.6	0.2	0.5	0.3	0.3	0.8	0.3	1.0	0.5	1	0.8	0.1
G13	0.2	0.4	0.1	0.2	0.2	0.1	0.5	0.8	0.5	0.2	0.8	0.8	1	0.9
G14	0.2	0.3	0.0	0.6	0.5	0.1	0.2	0.9	0.3	0.3	0.9	0.1	0.9	1

Steps for constructing a co-expression network

A) Microarray gene expression data

B) Measure concordance of gene expression with a Pearson correlation

C) The Pearson correlation matrix is either dichotomized to arrive at an adjacency matrix → unweighted network

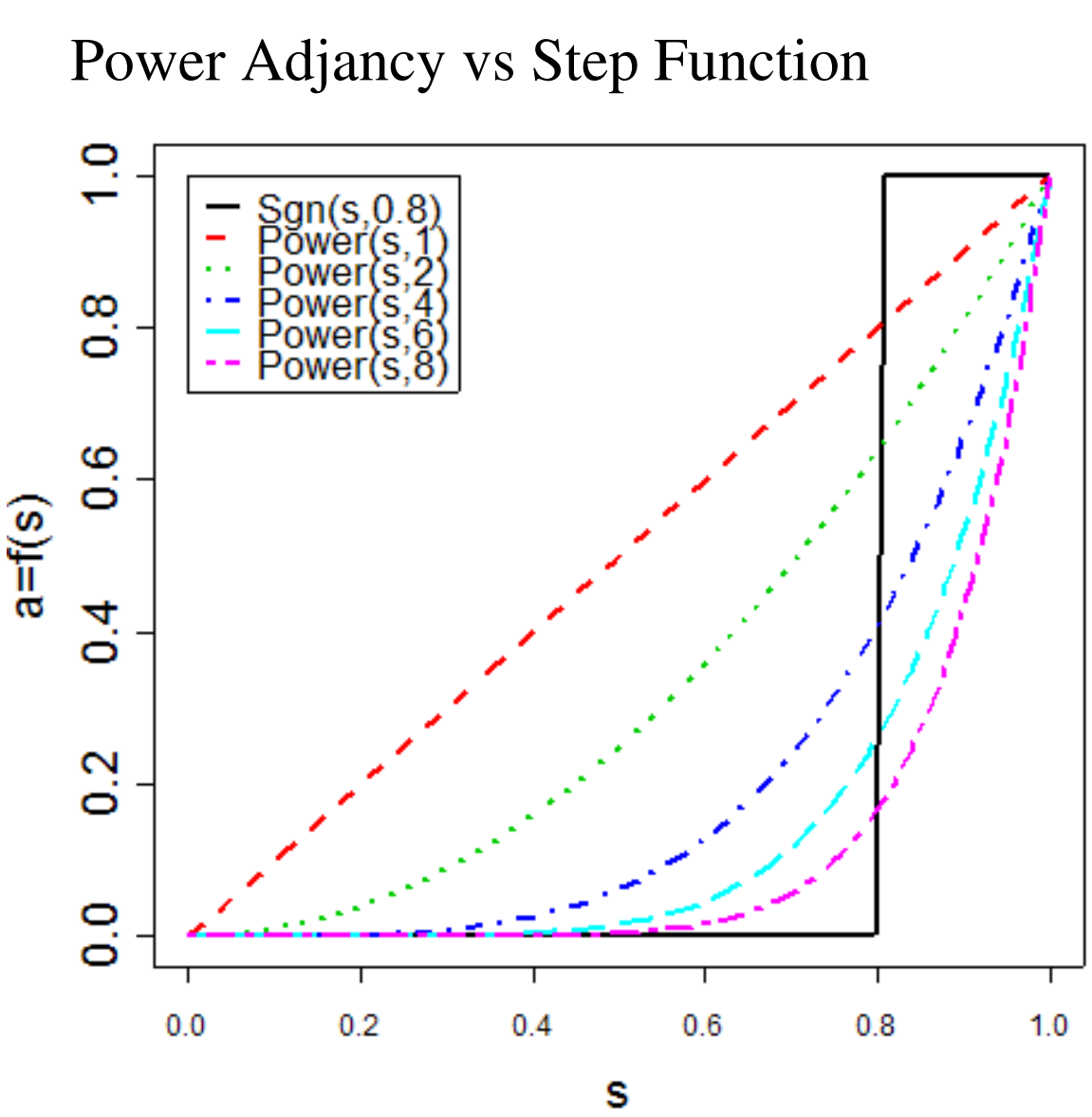
Or transformed continuously with the power adjacency function → weighted network

Power adjacency function results in a weighted gene network

$$a_{ij} = | cor(x_i, x_j) |^{\beta}$$

Often choosing beta=6 works well but in general we use the "scale free topology criterion" described in Zhang and Horvath 2005.

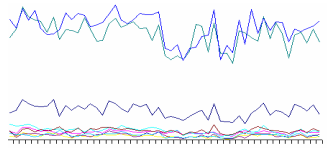
Comparing adjacency functions



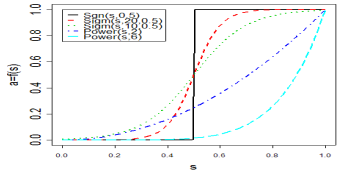
Comparing the power adjacency function to the step function

- While the network analysis results are usually highly robust with respect to the network construction method there are several reasons for preferring the power adjacency function.
  - Empirical finding: Network results are highly robust with respect to the choice of the power beta
  - Theoretical finding: Network Concepts make more sense in terms of the module eigengene.

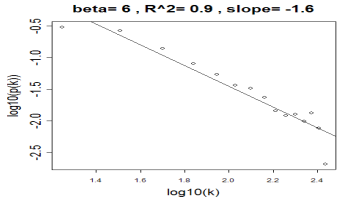
Define a Gene Co-expression Similarity



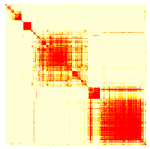
Define a Family of Adjacency Functions



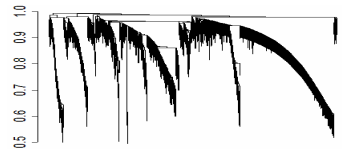
Determine the AF Parameters



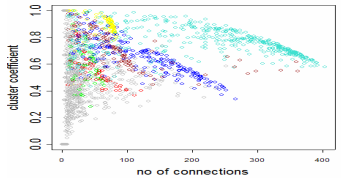
Define a Measure of Node Dissimilarity



Identify Network Modules (Clustering)

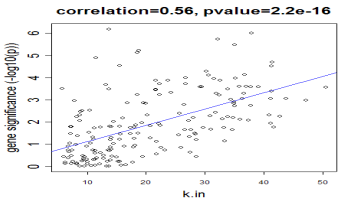


Relate Network Concepts to Each Other



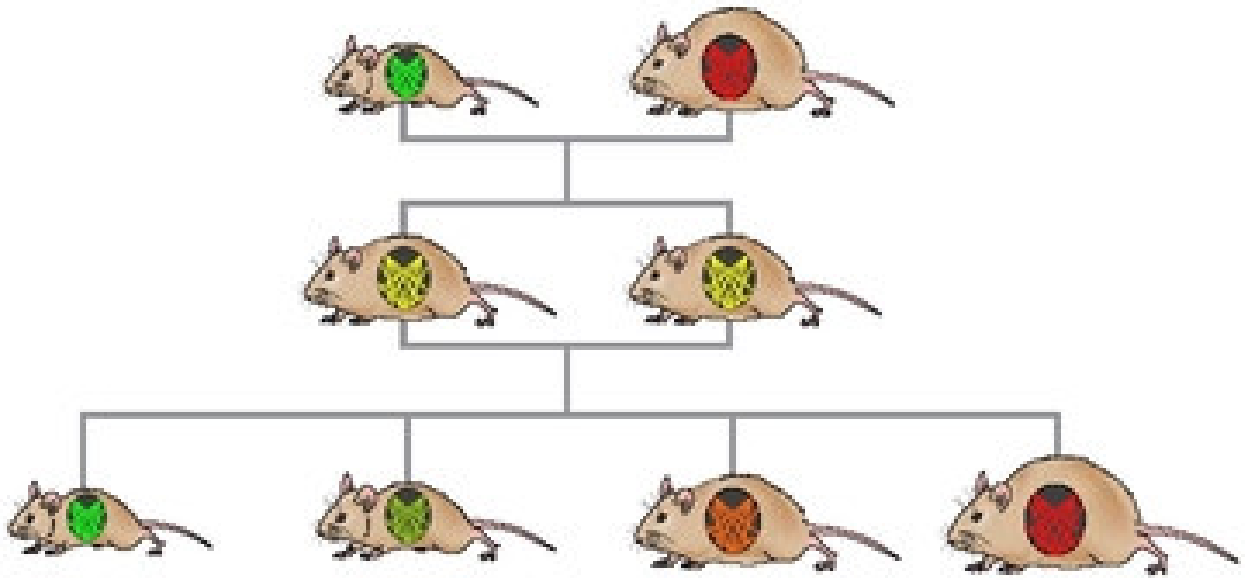
Focus of this talk:

Relate the Network Concepts to External Gene or Sample Information



Integrating Genetic and Network Analysis to Characterize Genes Related to Mouse Weight

A Ghazalpour, S Doss, B Zhang, C Plaisier, S Wang, EE Schadt, T Drake, AJ Lusis, S Horvath. PLoS Genetics August 2006



F2 mouse cross data

- We applied the network construction algorithm to a subset of gene expression data from an F2 intercross between inbred strains C3H/HeJ and C57BL/6J.
- Used liver gene expression data from 135 female mice (very different from male mice!)
- **Goal: *Characterize* genes whose expression profile are correlated with body weight**
- Statistical Method: Integrate network concepts with genetic concepts in a multivariate linear regression model

Defining Gene Modules
=sets of tightly co-regulated genes

Module Identification based on the notion of topological overlap

- One important aim of metabolic network analysis is to detect subsets of nodes (modules) that are tightly connected to each other.
- We adopt the definition of Ravasz et al (2002): modules are groups of nodes that have high topological overlap.

Topological Overlap leads to a network distance measure

- Generalized in Zhang and Horvath (2005) to the case of weighted networks
- Generalized in Yip and Horvath (2006) to higher order interactions

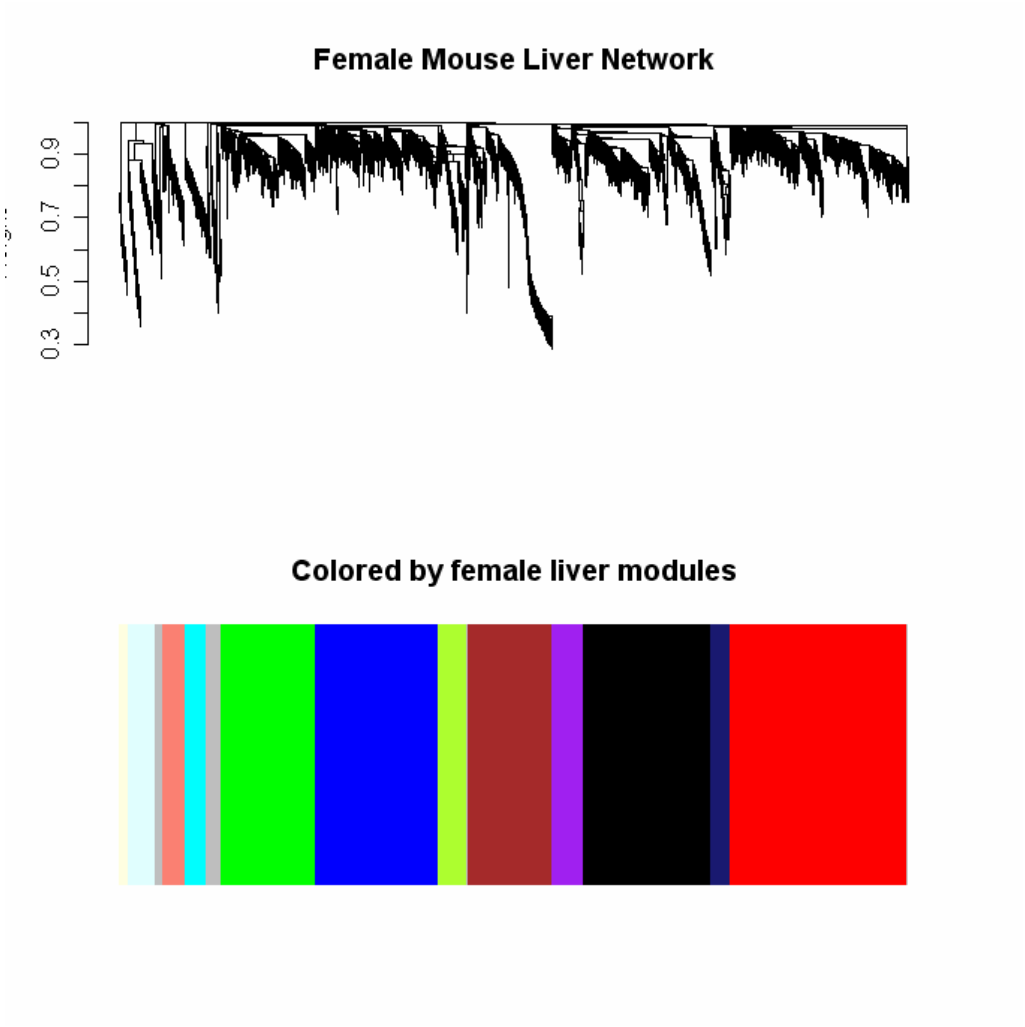
$$TOM_{ij} = \frac{\sum_u a_{iu} a_{uj} + a_{ij}}{\min(k_i, k_j) + 1 - a_{ij}}$$

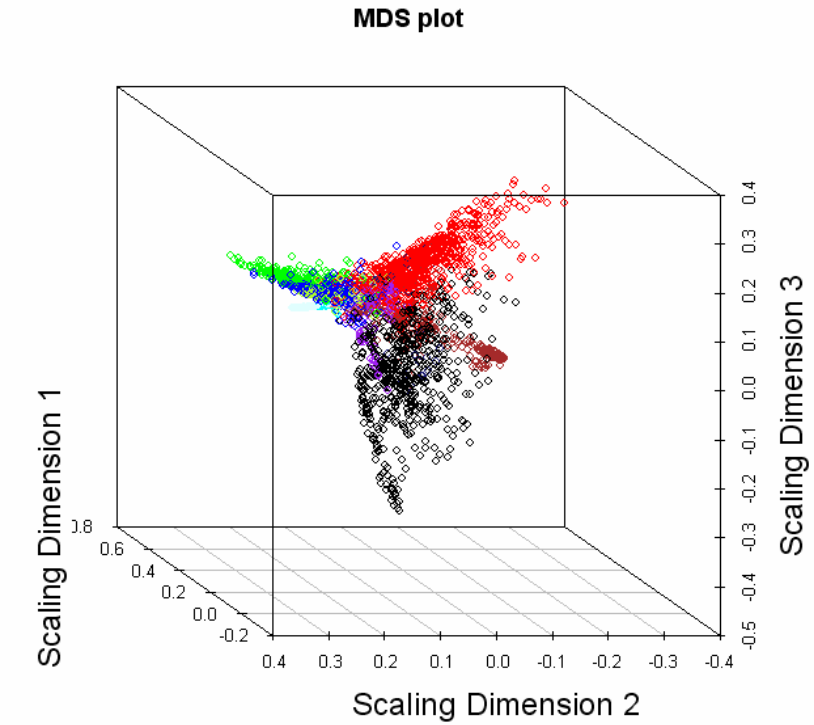
$$DistTOM_{ij} = 1 - TOM_{ij}$$

Using the topological overlap dissimilarity matrix to cluster genes

- To group nodes with high topological overlap into modules (clusters), we use average linkage hierarchical clustering coupled with the TOM dissimilarity measure.
- Modules correspond to branches of the dendrogram
- Once a dendrogram is obtained from a hierarchical clustering method, modules correspond to cut-off branches.
  - we use the "dynamic tree cut algorithm" since it allows for a flexible choice of height cut-offs.

Module plots for female liver expression data





Mouse body weight gives rise to a gene significance measure

- Abstract definition of a gene significance measure:
  - GS(i) is non-negative,
  - the bigger, the more *biologically* significant
  - Example: GS(i)=-log(p-value)

But here we use

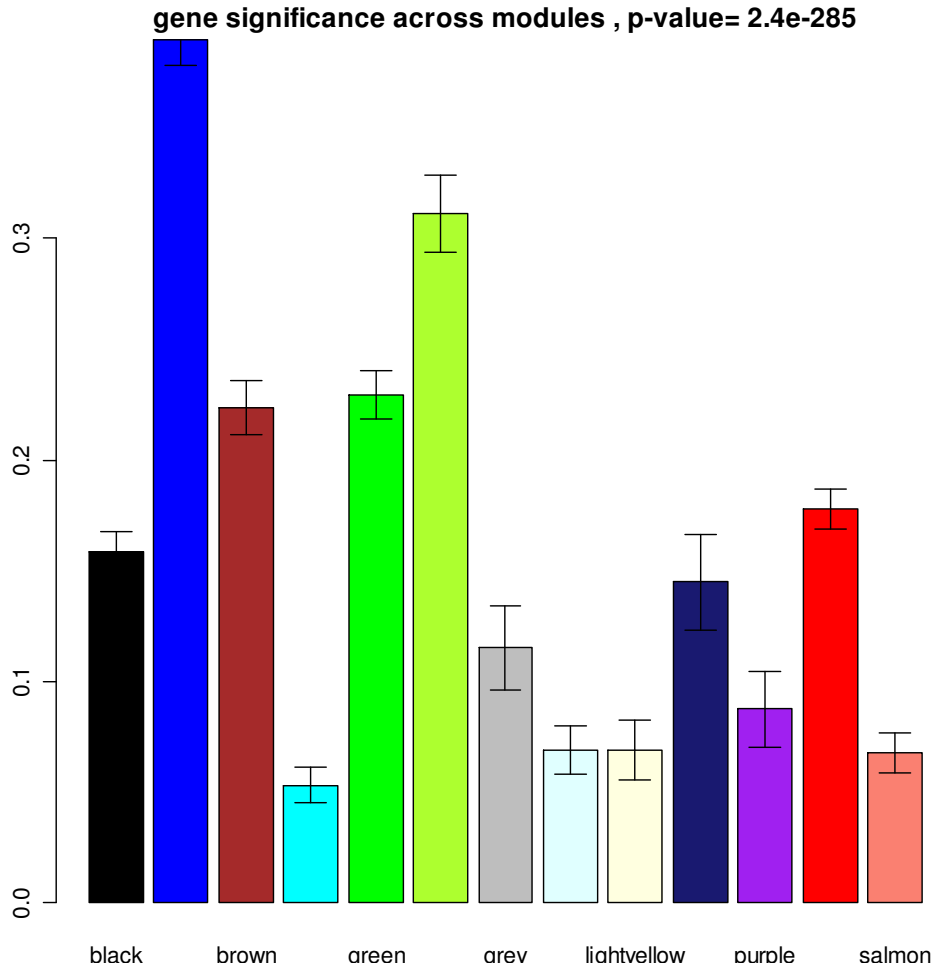
- GSweight(i) = |cor(x(i), weight)|
  - where x(i) is the gene expression profile of the ith gene.

A gene significance measure naturally gives rise to a module

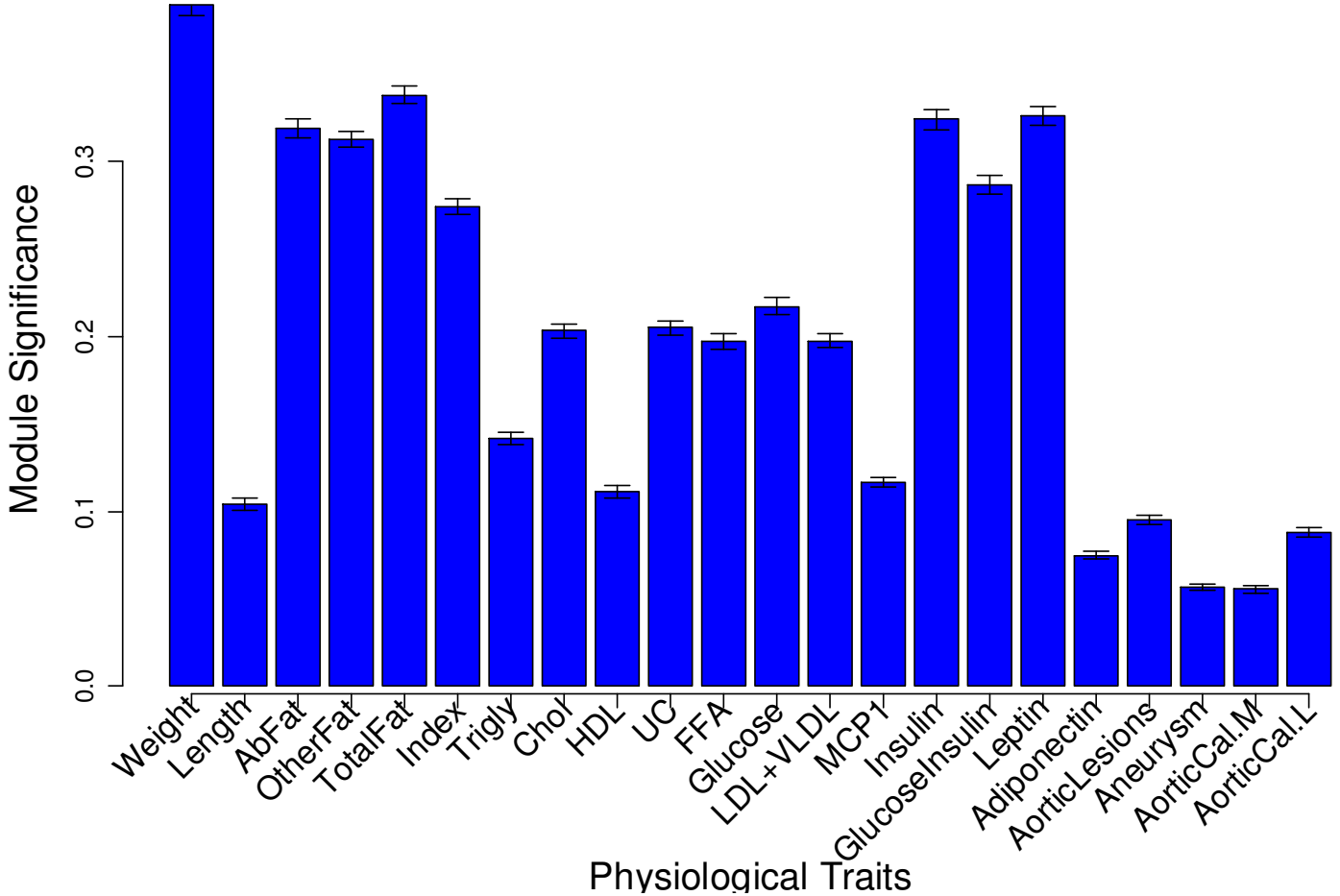
module significance measure

- Module Significance=mean gene significance

The blue module has high module significance with respect to body weight, i.e. it is highly enriched with genes that are correlated with weight



Relating the blue module genes to 22 physiological traits



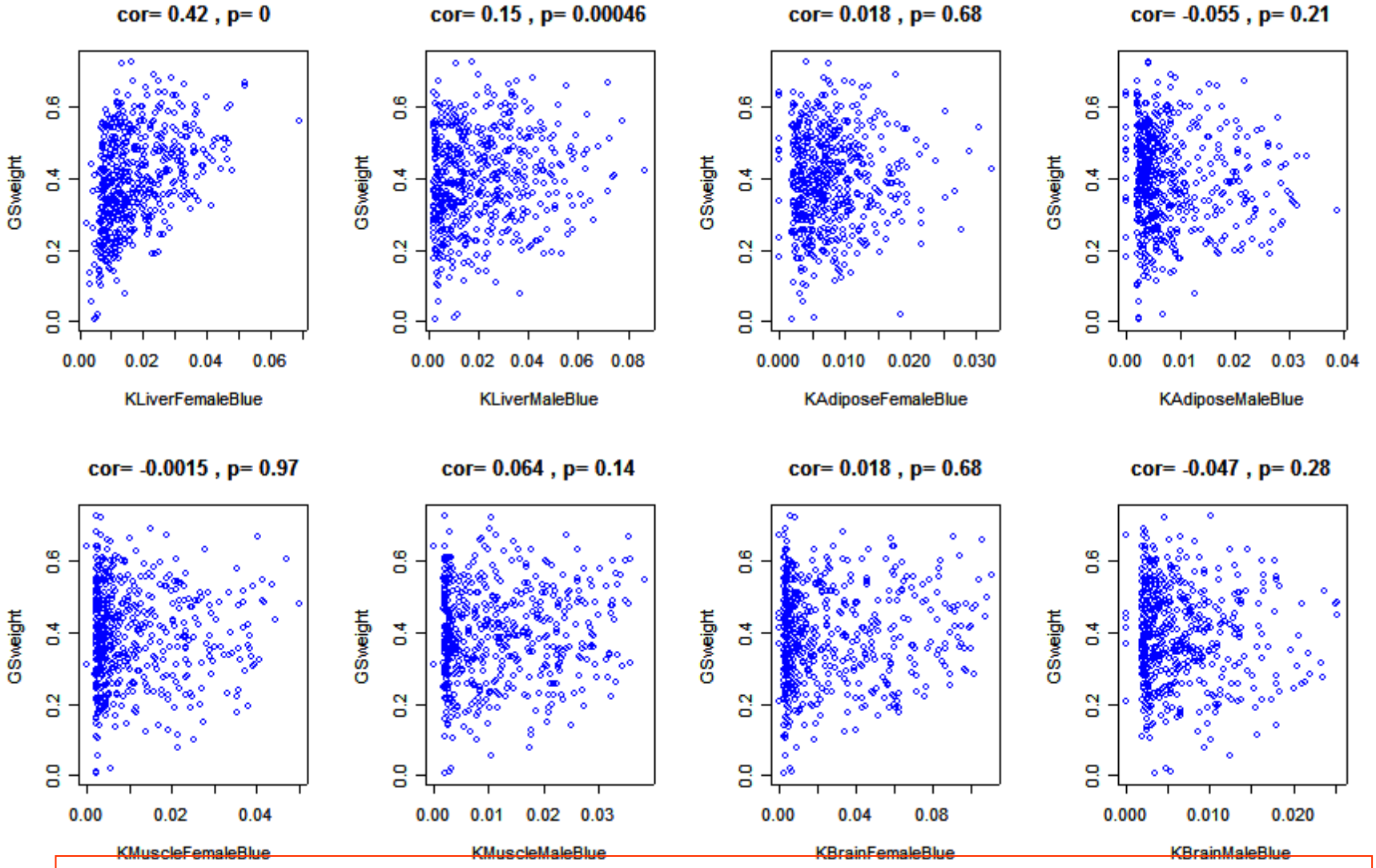
Message: unsupervised module detection method found a biologically interesting module

- The network modules were defined without regard to a physiological trait (unsupervised clustering of genes)
- The blue module is comprised of genes that relate to physiologically interesting traits, in particular body weight.
- Gene ontology: The blue module is enriched for genes in the ‘extra-cellular matrix (ECM) receptor interaction’ (p=2.3x10-9) and ‘complement and coagulant cascades’ (p=1.0x10-6) pathways.

Since highly connected `hub' genes have been found to be biologically important in other applications, it is natural to ask whether GSweight is related to intramodular connectivity in the blue module.

Further it is interesting to study the relationship between GSweight and k in different gender/tissue combinations.

Relating blue module connectivity to weight-based gene significance in different gender/tissue combinations.

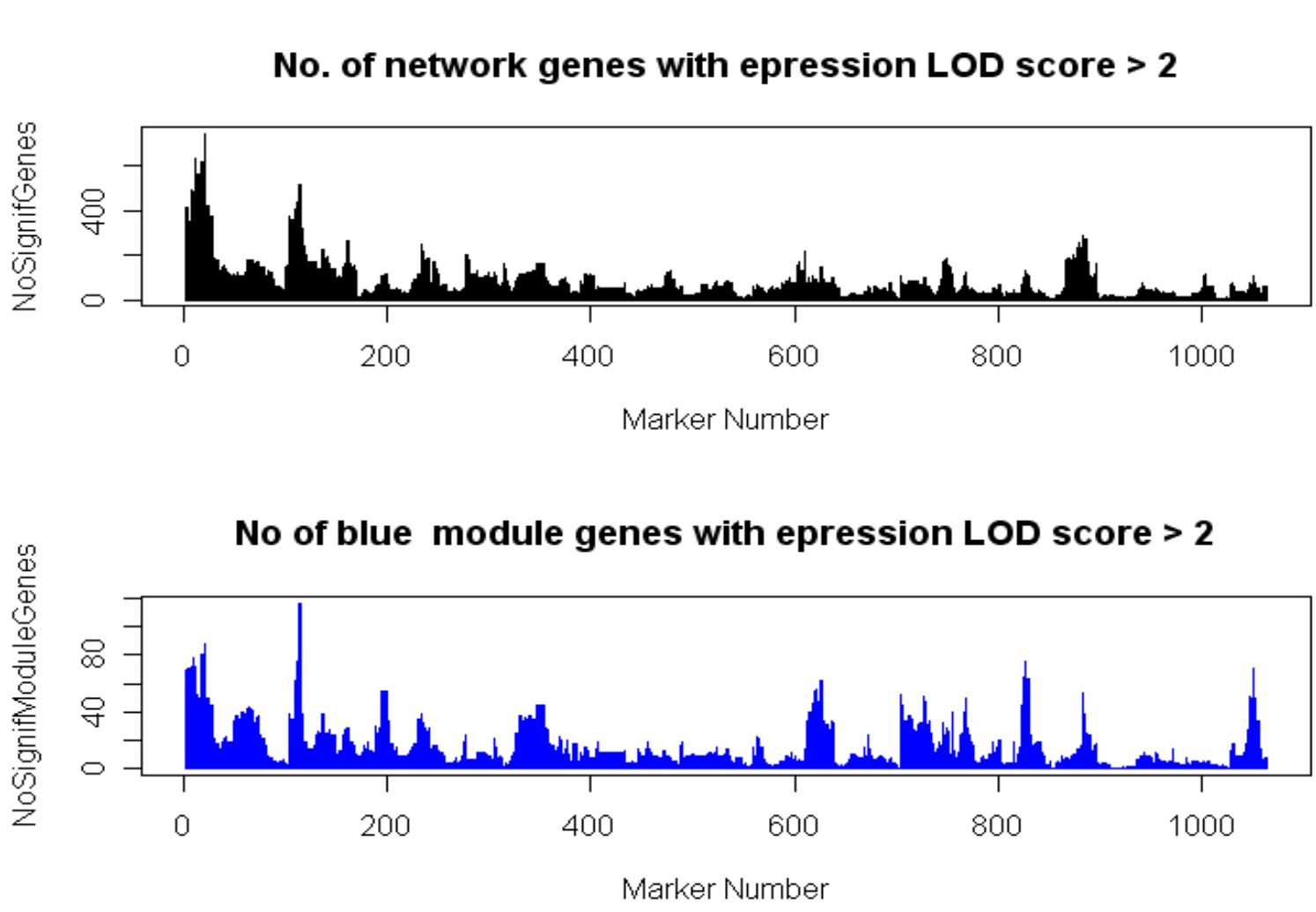


Message: there is a highly significant relationship between GSweight and k In the female liver network which cannot be found in other combinations.

Understanding the genetic drivers of the module genes

- Since genetic marker data were available for each mouse, it is natural to relate blue module gene expressions to the SNP markers. This could help identify the genetic drivers of the blue module pathway.
- Using 1065 single nucleotide polymorphism (SNP) markers that were evenly spaced across the genome (~1.5 cM density), we mapped the gene expression values and plotted the distribution of the expression quantitative trait loci (eQTL) for all genes within each gene module.

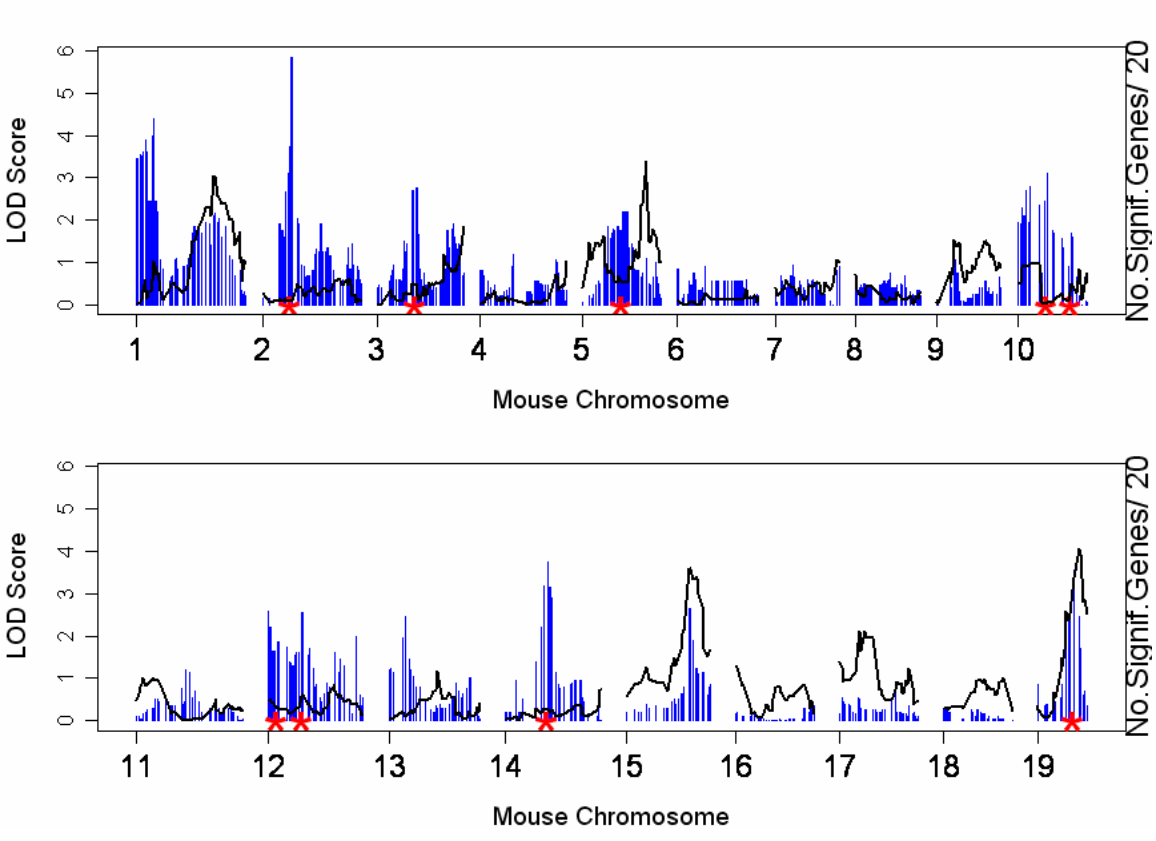
Comparing eQTL hotspots between the 3421 most connected genes (black) and the module genes (blue)



Module QTLs=mQTL =chromosomal location that affects module gene expressions.

- we hypothesized that there might also be genomic hot spots which coordinately regulate the transcript levels of the genes within each module.
- New Terminology:
- Module QTL (mQTL)=genomic “hotspot” that regulates transcript levels of the module genes.

Comparing the body weight LOD score curve (black curve) to distribution of module eQTLs (blue bars) of the blue module



Blue bar=
No. of genes whose expression LOD score at the marker >2
Red stars label mQTLs

Message:
While there is some overlap between the mQTLs and clinical traits (chromosome 19) there are also pronounced differences: see the blue spike (mQTL2) on chromosome 2.

A SNP marker naturally gives rise to a measure of gene significance

$$GS.SNP(i) = |cor(x(i), SNP)|.$$

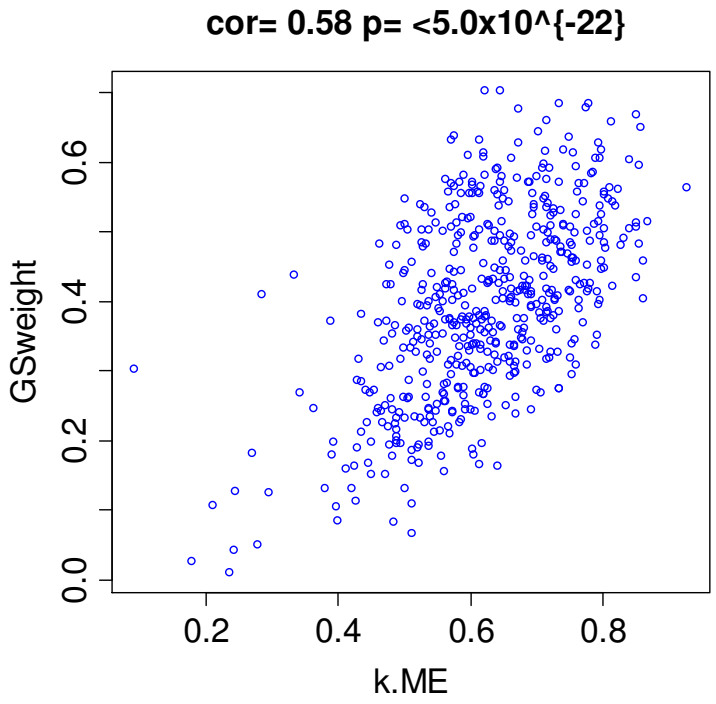
- Additive SNP marker coding: AA->2, AB->1, BB->0
- Absolute value of the correlation ensures that this is equivalent to AA->0, AB->1, BB->2
- Dominant or recessive coding may be more appropriate in some situations
- Conceptually related to a LOD score at the SNP marker for the i-th gene expression trait

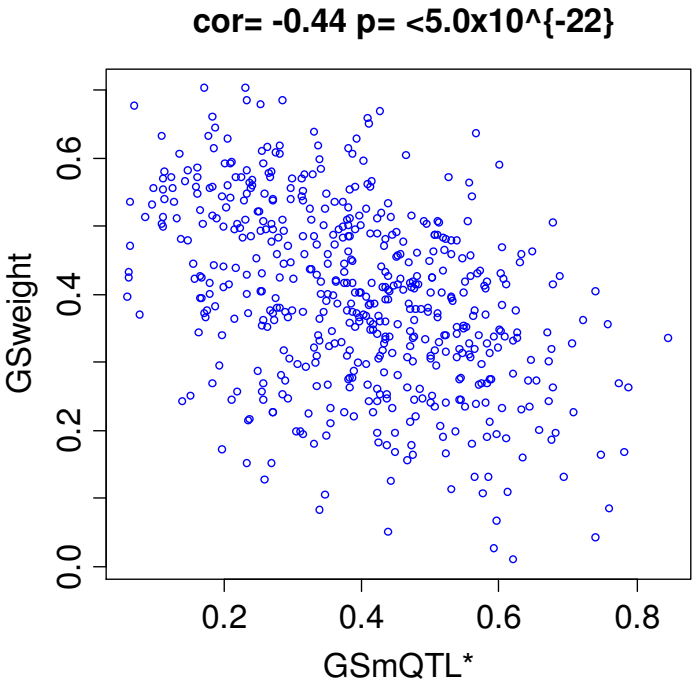
Using mQTLs to define gene significance measures

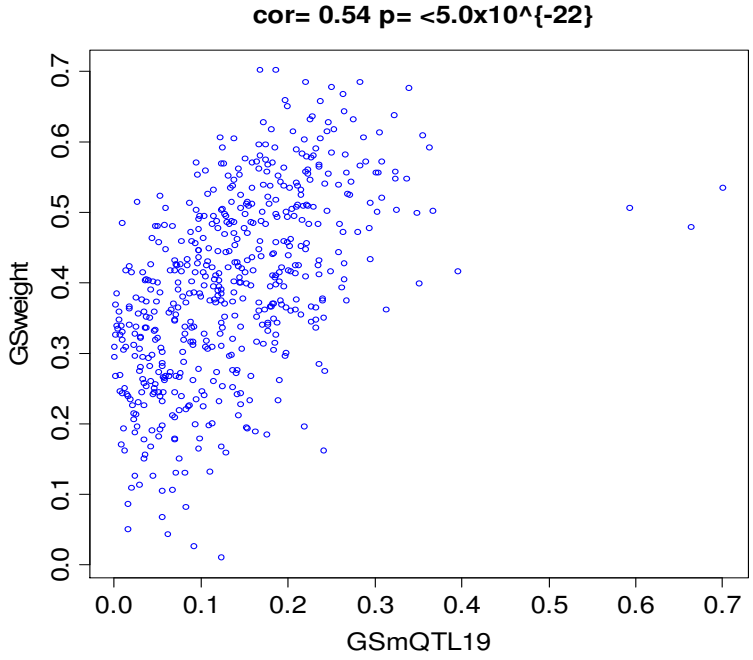
GSmQTL2(i) = |cor(x(i), mQTL2)|
GSmQTL5(i) = |cor(x(i), mQTL5)|
GSmQTL10(i) = |cor(x(i), mQTL10)|
GSmQTL19(i) = |cor(x(i), mQTL19)|
We also find it useful to define the following summary covariate since it is highly significant in our multivariate linear regression model
GSmQTL(i)=GSmQTL2+GSmQTL5+GSmQTL10*

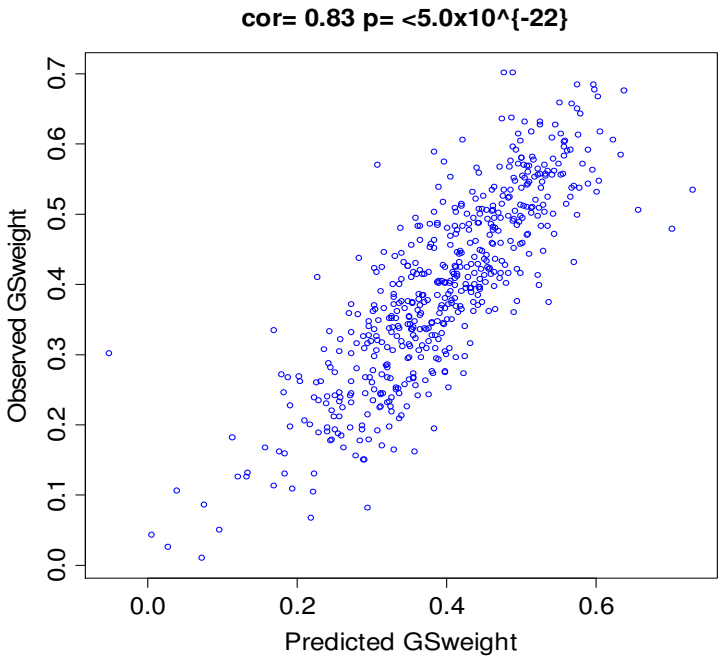
Multivariate Linear Regression Models for GSweight

Model	**Regression Model**	***R^2***	**Covariate**	**Co-efficient**	**Z**	***p*-Value**
Model 1: Genetic View	GSweight ~ GSmQTL* + GSmQTL19	0.37	GSmQTL*	–0.250	–8.05	5.00E–15
			GSmQTL19	0.652	12.30	<2E–16
			—	—	—	—
Model 2: Network View	GSweight ~ k_{me}	0.34	K_{me}	0.643	16.51	<2E–16
			—	—	—	—
			—	—	—	—
Model 3: Network + Genetics	GSweight ~ k_{me} + GSmQTL* + GSmQTL19	0.70	GSmQTL*	–0.304	–14.00	<2E–16
			GSmQTL19	0.552	14.87	<2E–16
			K_{me}	0.636	23.86	<2E–16

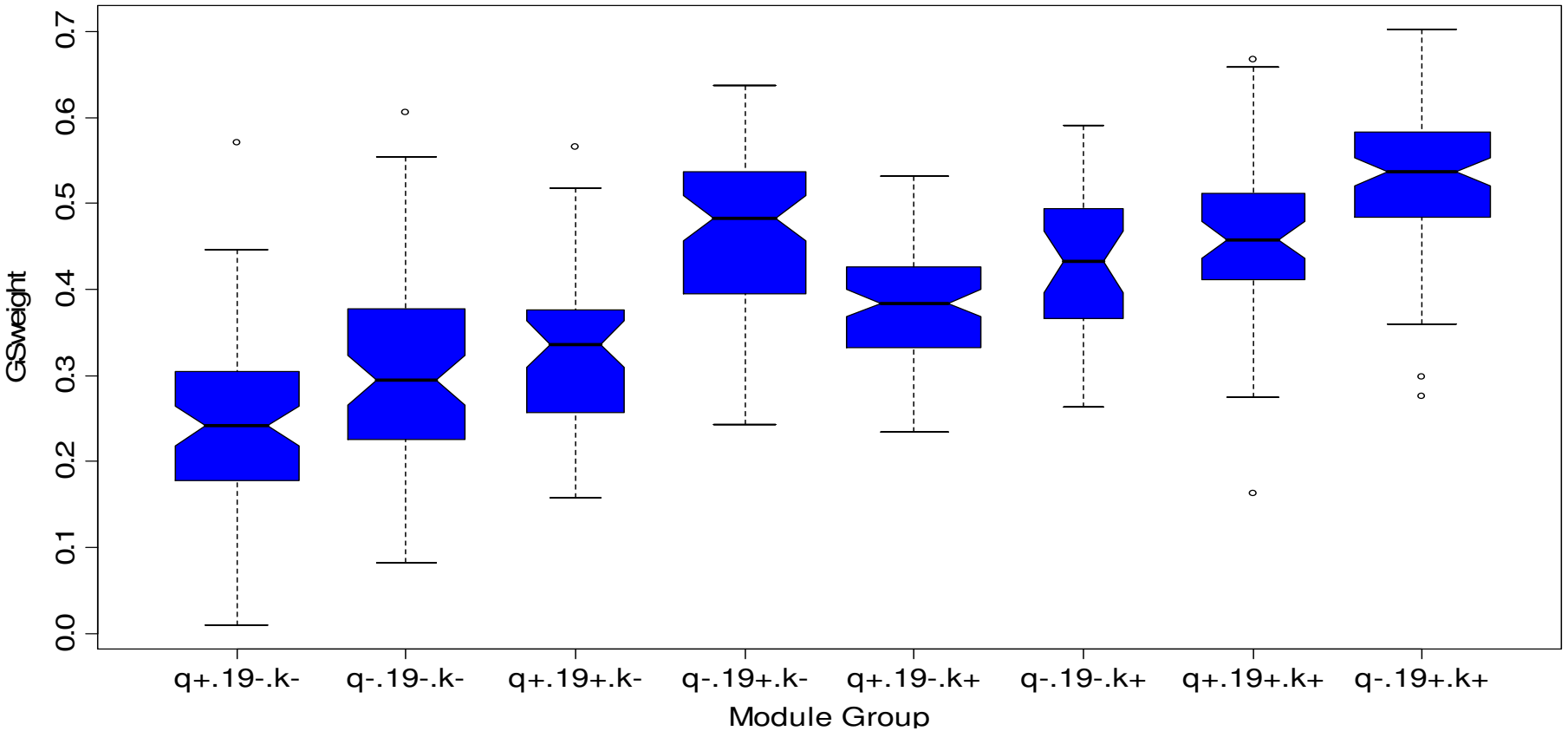

cor= 0.58 p= <5.0x10^{-22}
GSweight
k.ME
0.0
0.2
0.4
0.6
0.2
0.4
0.6
0.8


cor= -0.44 p= <5.0x10^{-22}
GSweight
GSmQTL*
0.0
0.2
0.4
0.6
0.2
0.4
0.6
0.8


cor= 0.54 p= <5.0x10^{-22}
GSweight
GSmQTL19
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7


cor= 0.83 p= <5.0x10^{-22}
Observed GSweight
Predicted GSweight
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

The integrated model allows us to characterize genes that are related to weight



Here the blue module genes are binned into 2^3=8 bins created by dichotomizing the covariates GSmQTL* (high=q+,low=q-), GSmQTL19(high19+), k(high=k+). (splits were chosen by the median)

Discussion

The multivariate regression models in the Table highlight the value of taking a network perspective. Model 3 integrates co-expression network concepts (connectivity) and genetic marker information (GSmQTL) to explain 70% of the variation in GSweight.

- This simple model is attractive since it illustrates that 3 biologically intuitive variables suffice to explain which genes of this pathway are related to body weight.
- Integrating gene co-expression networks with genetic marker information allows one to understand what factors influence the relationship between gene expression and weight.

Comparing our analyses to standard approaches

- Instead of modelling the relationship bodyweight~SNPs
  we find it advantageous to model
- GSweight~GS.mQTL+connectivity.
- While traditional mapping would take the mice as unit of observation, we consider the genes of a physiologically interesting network module.
- Major reason: intramodular connectivity turns out be a highly significant independent predictor.
- Related to modeling
  - weight~mQTL+module eigengene

The advantages of a correlation based analysis

We define simple and intuitive concepts that are based on the Pearson correlation (connectivity, GSweight, GSmQTL).

For example, GSmQTL19 measures to what extent a gene "maps" to the chromosome 19 location and it is highly related to a single point LOD score.

Using the same association measure (Pearson correlation) puts the disparate data sets (gene expression, physiological traits and SNPs) on the same footing and highlights that these very different data sets can be naturally integrated using weighted gene co-expression network methodology.

For example, a complex trait can be considered as "idealized" gene in a co-expression network. Thus the gene significance GSweight(i)^beta can be interpreted as adjacency between body weight and the i-th gene expression.

- A mathematical advantage of the Pearson correlation is that it allows one to study the relationship between the network concepts in terms of the module eigengene, see Horvath, Dong, Yip (2006).

Software and Data Availability

- This ppt presentation and detailed software tutorials can be found at the following webpage
- http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/MouseWeight/

Acknowledgement

- Mouse genetics
  - Anatole Ghazalpour, Sud Doss, Bin Zhang, Chris Plaisier, Susanna Wang, Eric E Schadt (Merck), Tom Drake, Jake Lusis

Lab members

- Jun Dong, Ai Li, Bin Zhang, Lin Wang, Wei Zhao

Other Collaborations

- Brain Cancer, Yeast Genetics
  - Paul Mischel, Stan Nelson, Marc Carlson
- Human/chimp brain
  - Mike Oldham, Dan Geschwind