Analysis of Microarray Data with Methods from Machine Learning and Network Theory

Summer Lecture 2015

Prof. Dr. A. B. Cremers

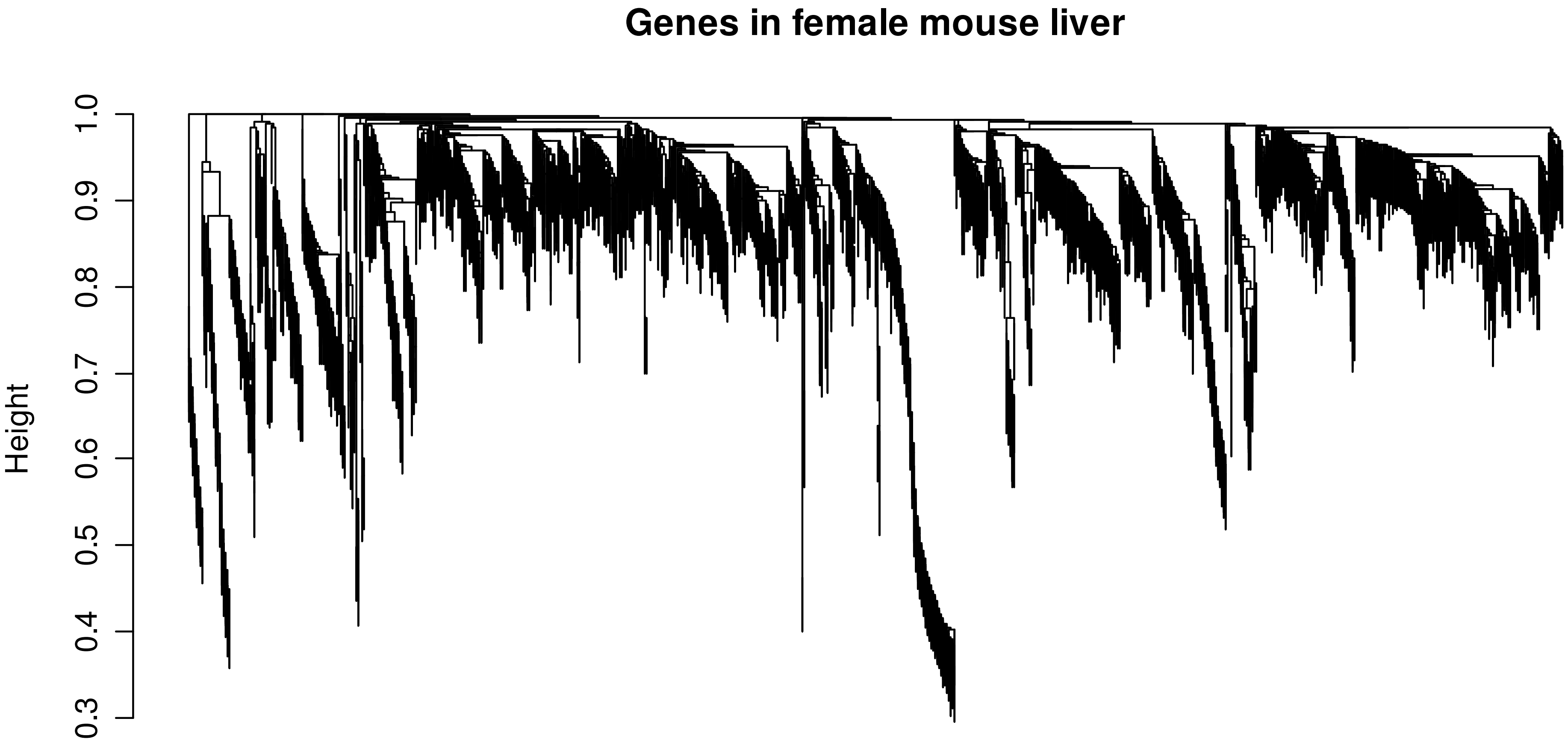
Dr. Jörg Zimmermann

Cutting branches from a cluster tree: the `dynamicTreeCut` R library

Identification of clusters (modules) in hierarchical clustering trees (dendrograms)

- A.k.a. branch or tree cutting, pruning
- General aim: find biologically meaningful groups of genes (terminology: network modules)
- Hypothesis: highly correlated (that is, connected) genes are functionally related
- Look for groups of highly connected genes
- These correspond to branches in the hierarchical clustering tree (dendrogram)

Example:



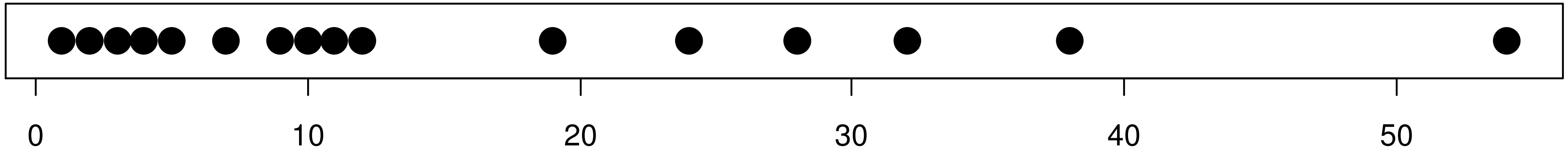
From: Ghazalpour et al (2006), *PLoS Genetics Volume 2 Issue 8*

Two types of branch cutting methods

- Constant height (static) cut
  - `cutreeStatic(dendro,cutHeight,minsize)`
  - based on R function `cutree`
- Adaptive (dynamic) cut
  - `cutreeDynamic(dendro, ...)`
- Getting more information about the dynamic tree cut:
  - `library(dynamicTreeCut)`
  - `help(cutreeDynamic)`
- More details: www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/BranchCutting/

Toy example of branch cutting

Data: 1,2,3,4,5, 7, 9,10,11,12, 19,24,28,32,38, 54

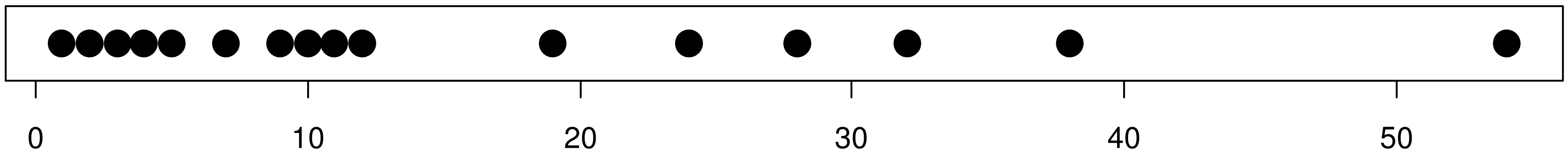


Dissimilarity:

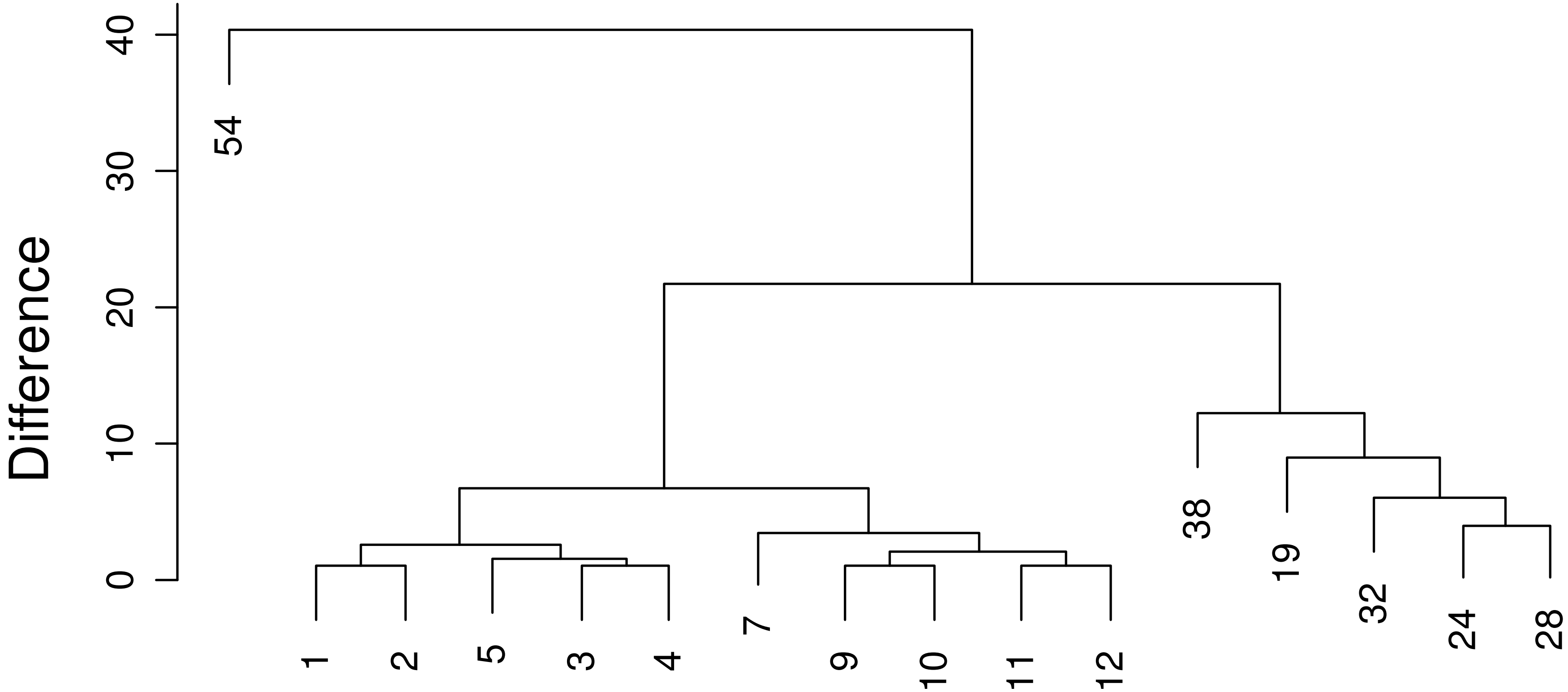
$$diss_{ij} = | x_i - x_j |$$

Example: Dissimilarity (1, 9) = 8

Clustering:

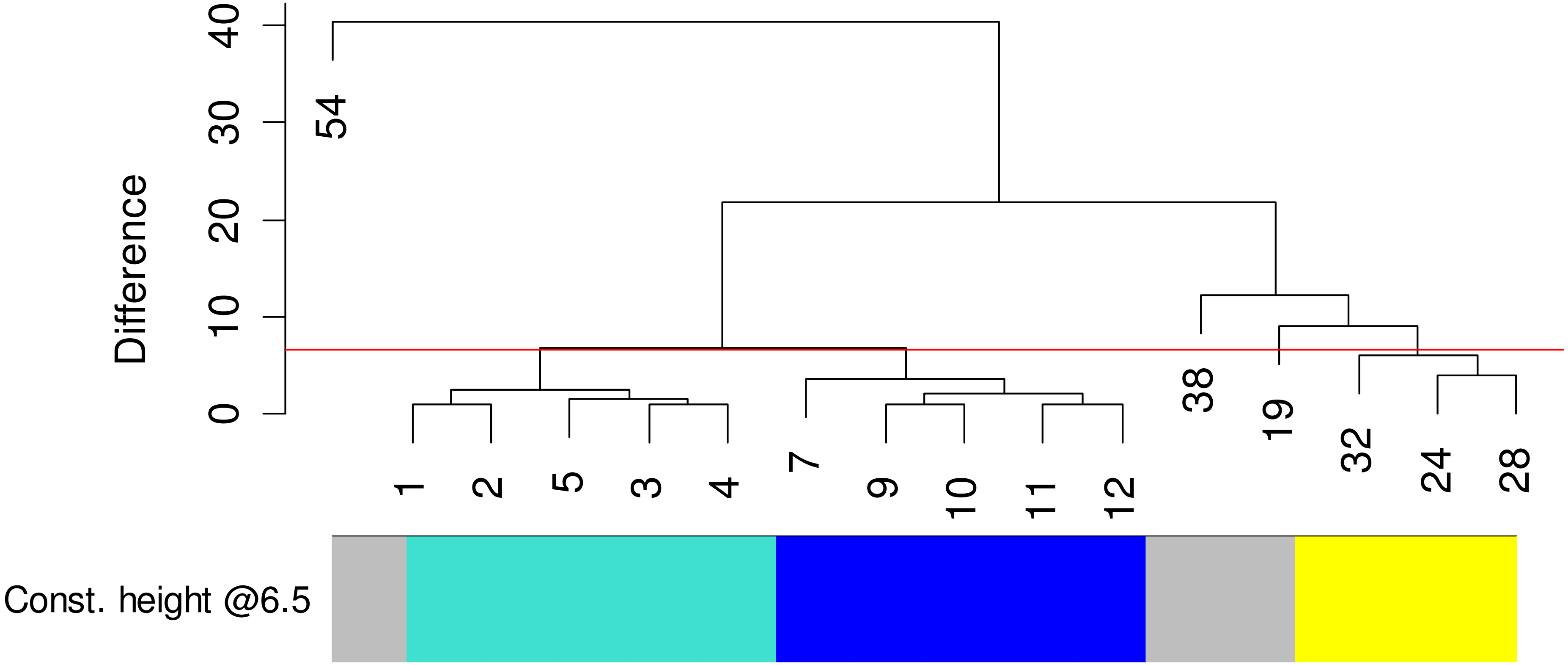


Dendrogram (average linkage):

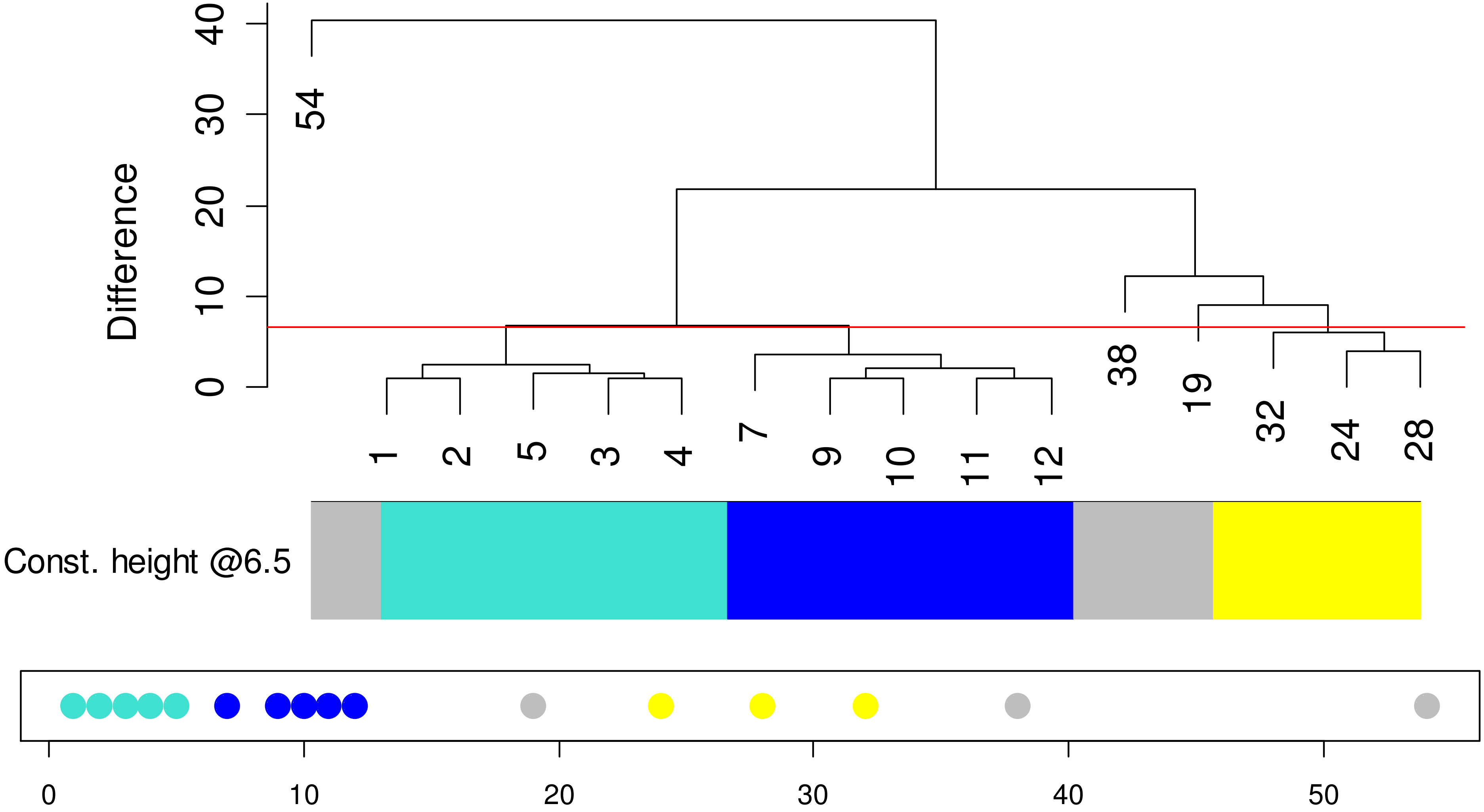


Constant height cut (a.k.a. static cut)

Pick a height (in this case 6.5) and minimum size (in this case 3). Draw a line (red) at the chosen height. Look at all branches cut off by the line. Those that have at least 3 objects on them are modules. Label each module by a color to simplify identification. Objects outside of any module are labeled grey.

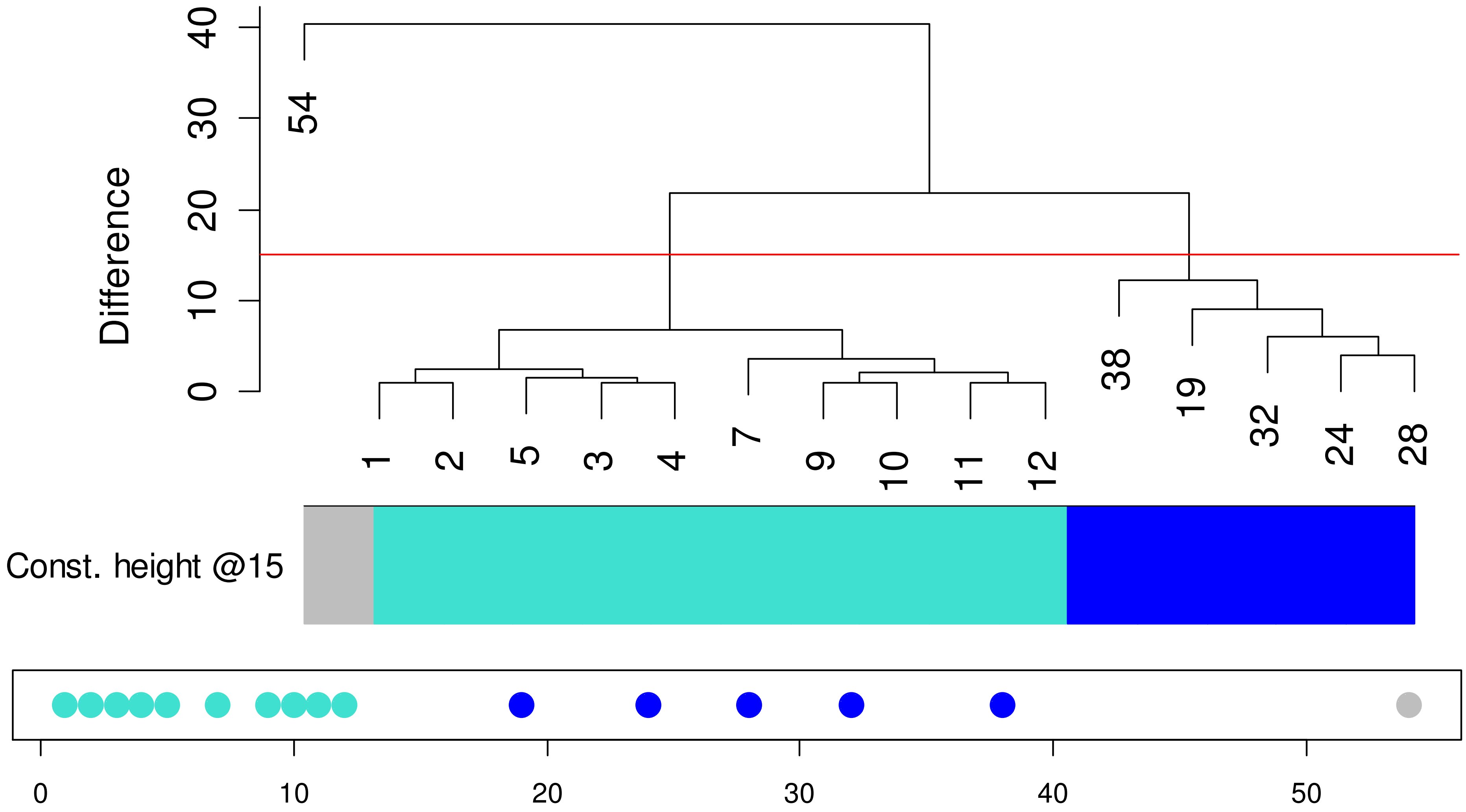


How do the clusters look like on the data?



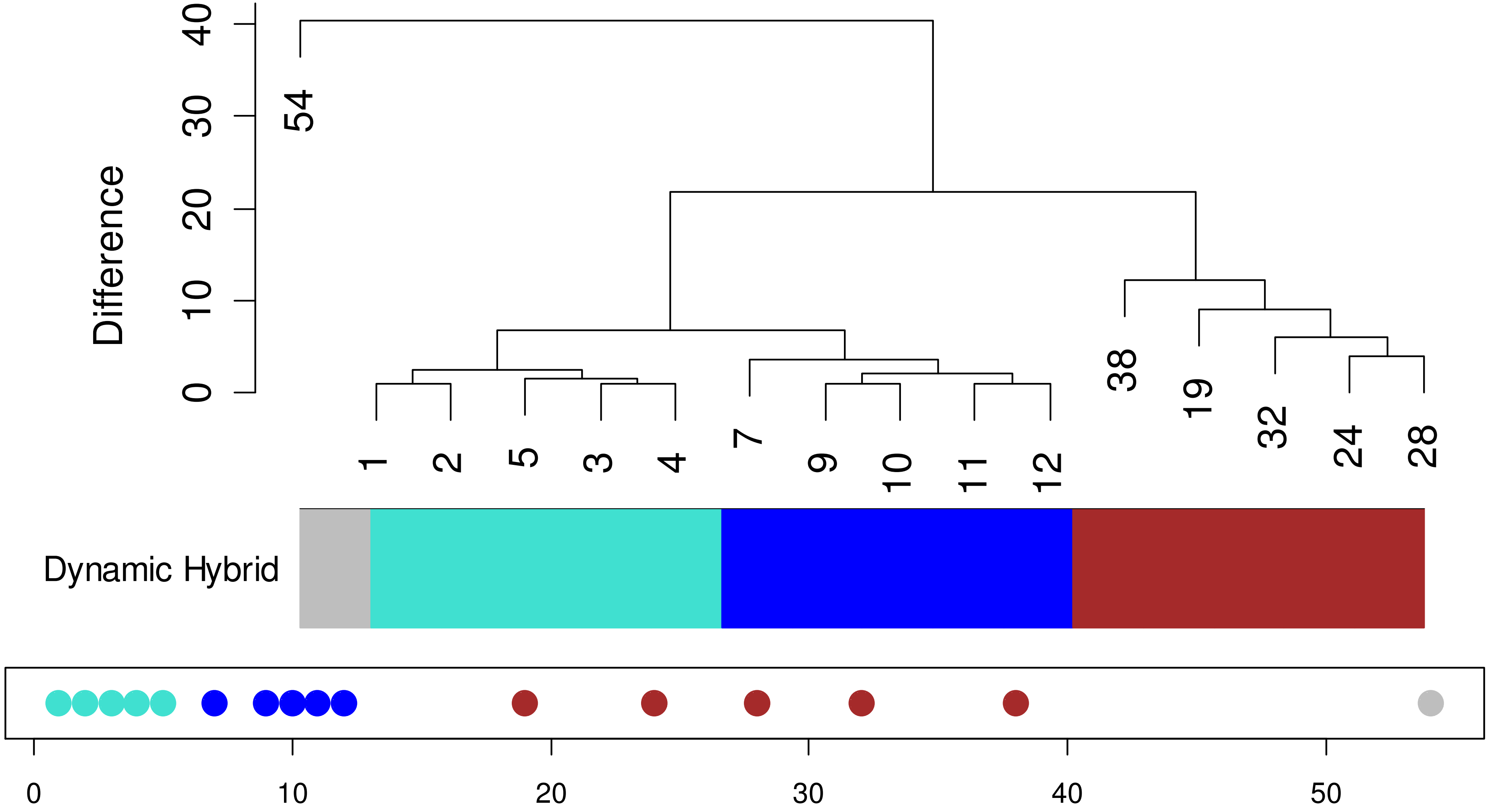
Yellow module appears to be missing its outer objects! Increase cut height?

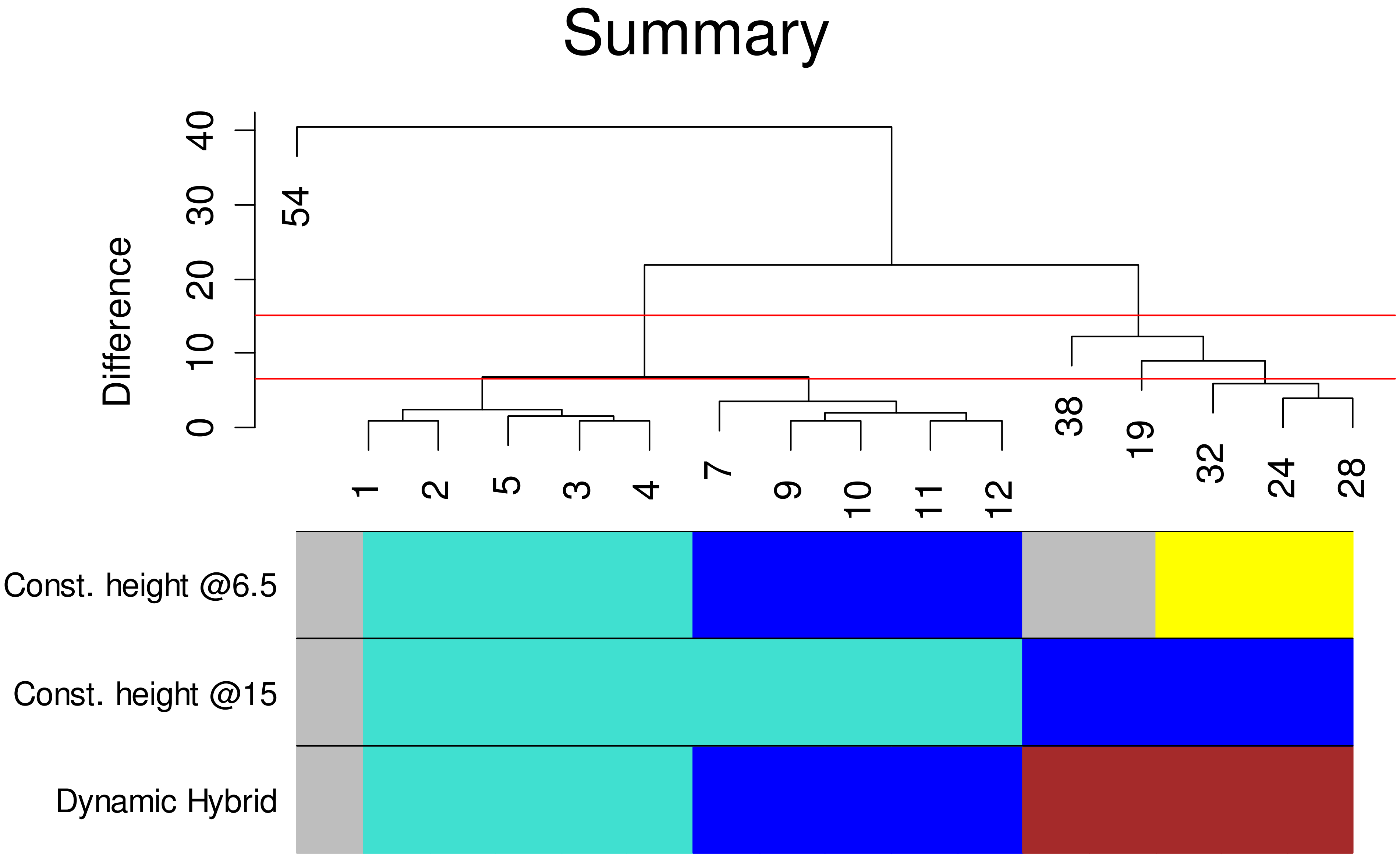
Constant height cut at height = 15:



Cut height is now too high: turquoise module swallowed its neighbor!
Lesson: constant-height cut cannot identify tight and loose modules at the same time.

Adaptive tree cut (“Dynamic Hybrid” method):





Reference: Langfelder, Zhang, and Horvath, Bioinformatics 2007

Using the singular value decomposition to define (module) eigengenes

Scale the gene expressions profiles (columns)

$$datX = scale(datX)$$

$$datX = UDV^T$$

$$U = (u_1 \quad u_2 \qquad\qquad u_m)$$

$$V = (v_1 \quad v_2 \qquad\qquad v_m)$$

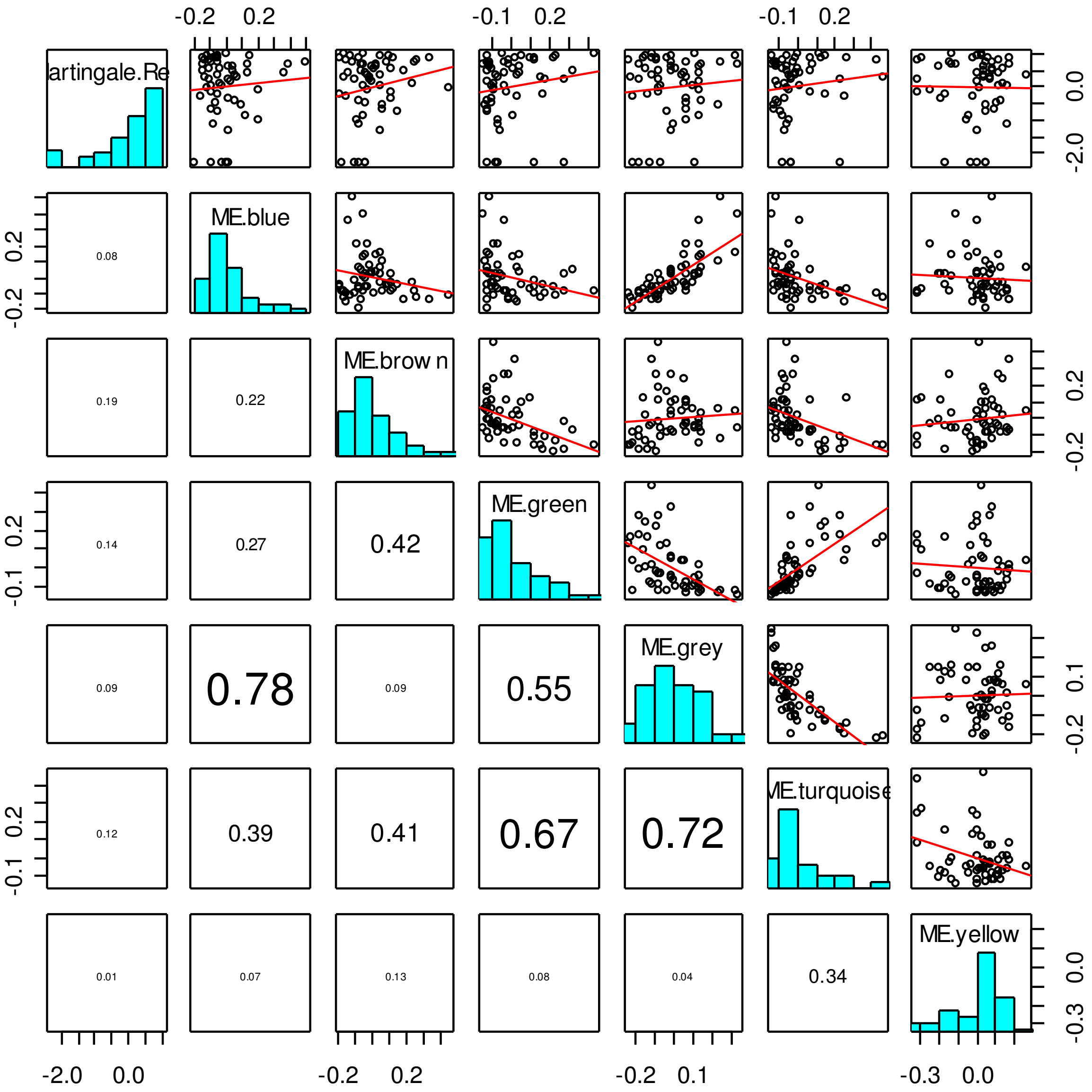
$$D = diag(|d_1|, |d_2|,^{--}, |d_m|)$$

Message: u_1 is the (first) eigengene E

If $\mathrm{datX}^{(\mathrm{q})}$ corresponds to the q-th module then

$\mathrm{E}^{(\mathrm{q})}$ is the q-th module eigengene.

Module eigengenes can be used to determine whether 2 modules are correlated. If correlation of MEs is high-> consider merging.



Eigengene networks Langfelder, Horvath (2007) BMC Systems Biology

Module eigengenes are very useful

- 1) They allow one to relate modules to each other
  - Allows one to determine whether modules should be merged
  - Or to define eigengene networks
- 2) They allow one to relate modules to clinical traits and SNPs
  - -> avoids multiple comparison problem
- 3) They allow one to define a measure of module membership: kME=cor(x,ME)

How to relate modules to external data?

Clinical trait (e.g. case-control status) 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 for the i-th gene

Concrete definition

- GS.ClinicalTrait(i) = |cor(x(i),ClinicalTrait)| where x(i) is the gene expression profile of the i-th gene

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

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

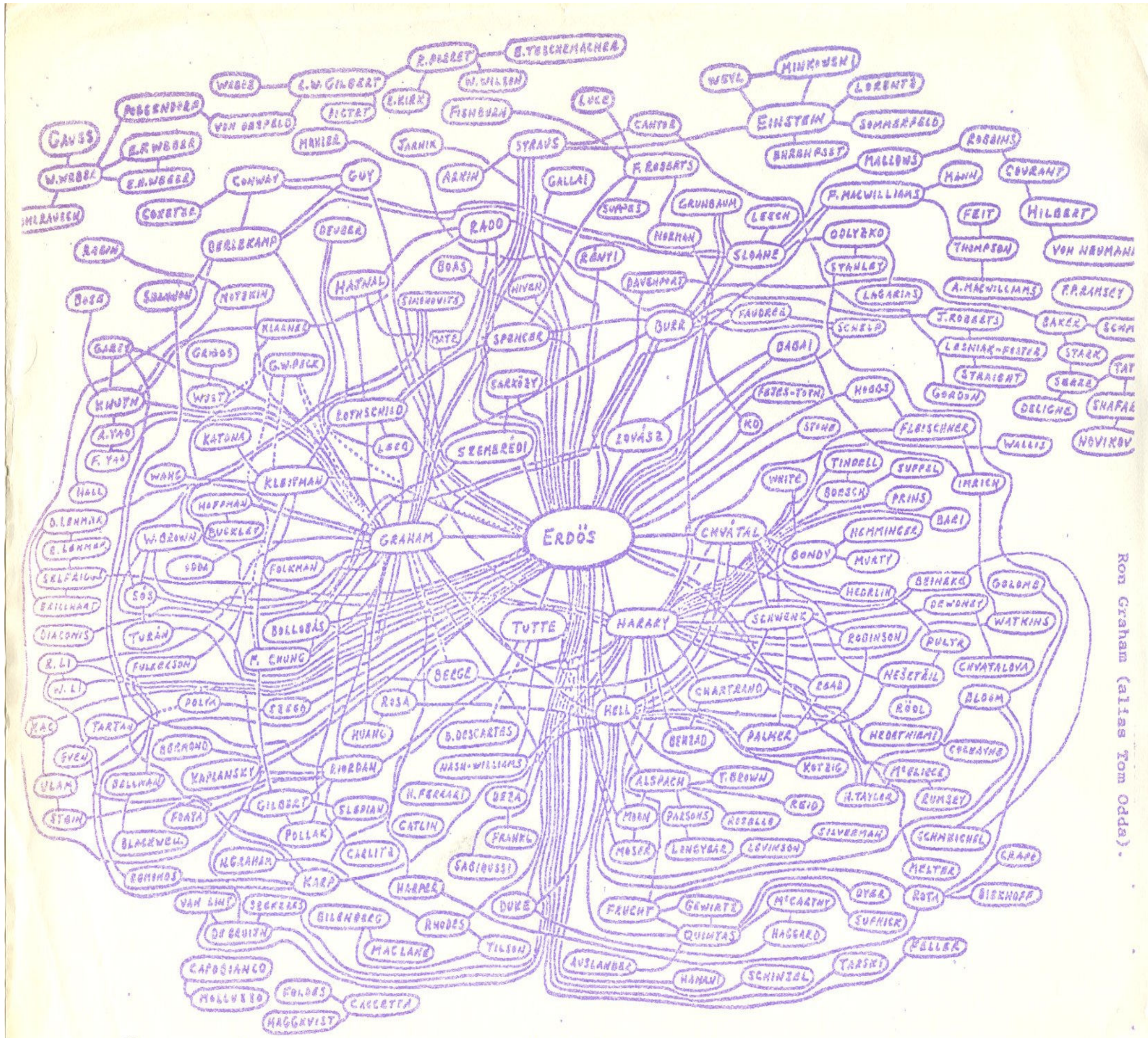
- Additive SNP marker coding: AA->2, AB->1, BB->0
- Absolute value of the correlation ensures that this is equivalent to AA->0, AB->1, BB->2
  - Dominant or recessive coding may be more appropriate in some situations

A gene significance naturally gives rise to a module significance measure

- Define module significance as mean gene significance
- Often highly related to the correlation between module eigengene and trait

Important Task in
Many Genomic Applications:
Given a network (pathway) of interacting genes how to find the central players?

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

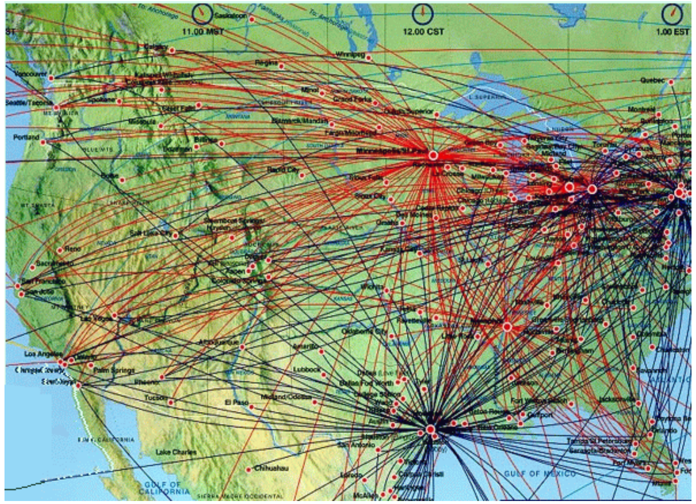


Ron Graham (alias Tom Odda).

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

Flight connections and hub airports



The nodes with the largest number of links (connections) are most important!

****Slide courtesy of A Barabasi**

Q: What is a hub gene?
Answer: it depends on the measure of node connectivity

Connectivity measure

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

$$Connectivity_i = k_i = \sum_{j \neq i} a_{ij}$$

$$\text{Scaled connectivity=K}_\text{i} = \frac{k_i}{\max(k)}$$

Define 2 alternative measures of intramodular connectivity and describe their relationship.

Intramodular Connectivity

- Intramodular connectivity kIN with respect to a given module (say the Blue module) is defined as the sum of adjacencies with the members of this module.
  - For unweighted networks=number of direct links to intramodular nodes
  - For weighted networks= sum of connection strengths to intramodular nodes

$$kIN_i^{BlueModule} = \sum_{\{j \in BlueModule\}} a_{ij}$$

Eigengene based connectivity, also known as kME or module membership measure

$$kME_i = ModuleMembership(i) = cor(x_i, ME)$$

kME(i) is simply the correlation between the i-th gene expression profile and the module eigengene.

Very useful measure for annotating genes with regard to modules.

Module eigengene turns out to be the most highly connected gene

Intramodular hubs

- Defined as nodes (genes) with high kME (or high kIM)
- Study intramodular hubs in
  - Single network analysis: Intramodular hubs in biologically interesting modules are often very interesting
  - Differential network analysis: Genes that are intramodular hubs in one condition but not in another are often very interesting