

**Computationally intensive statistical
methods for analysis of gene expression
data**

Katherine S. Pollard & Mark J. van der Laan

Division of Biostatistics, U.C. Berkeley

www.stat.berkeley.edu/~laan

Computational Biology

Computational fields (statistics, mathematics) and biology have been linked for a long time.

What's new?

1. Technologies (PCR, sequencing, microarrays)
2. Databases (genomes, protein classification, TRANSFAC, SNPs)
3. Computers and computer science

So we have:

- very high dimensional data
- vast quantities of data
- the potential to answer complex questions

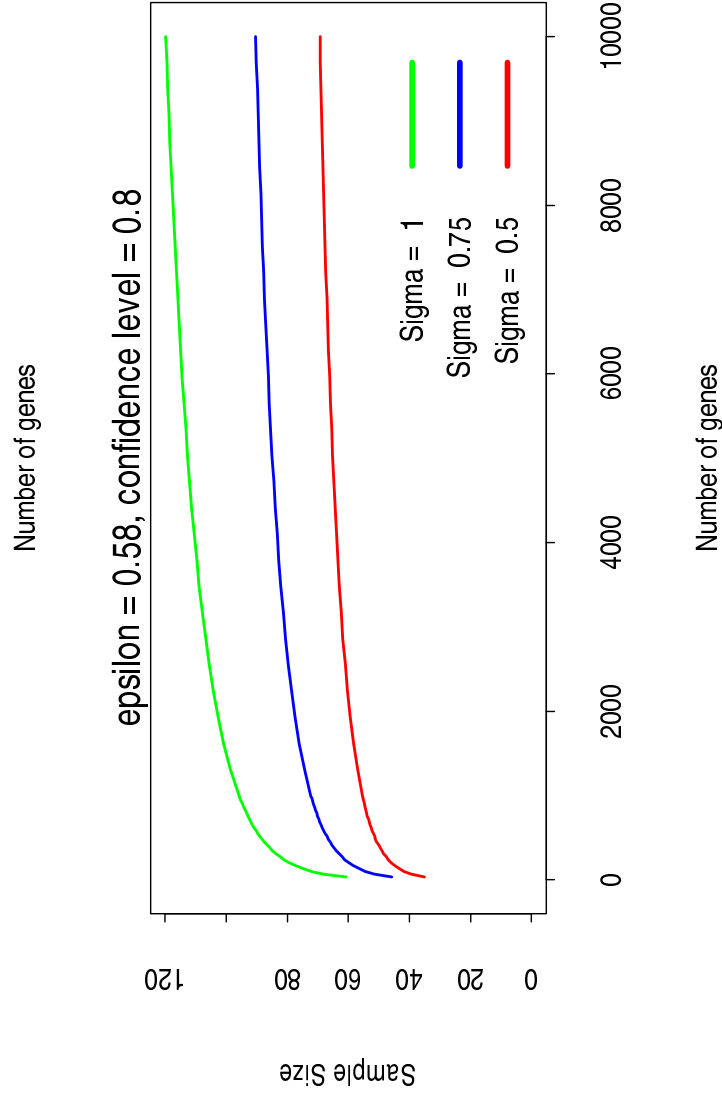
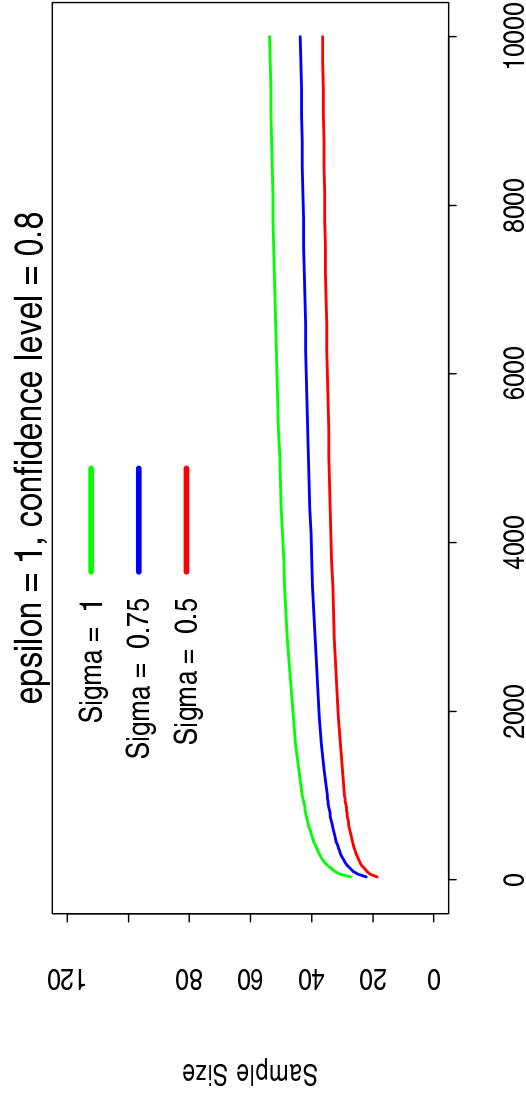
Statistical Framework

Typically these data sets can be viewed as n observations of a random variable.

The dimension of the data (p) often far exceeds the sample size ($n \ll p$). So, statistical rigor is particularly important.

- Clear definitions of statistical notions:
 - parameter
 - parameter estimate
 - consistency
 - confidence
- New asymptotics for $p \rightarrow \infty$ (at what rate relative to n ?)
- Sample size formulas
- Methods to assess reliability of exploratory techniques
 - bootstrap
 - cross-validation

Sample Size Requirements



Statistical Framework

Typically these data sets can be viewed as n observations of a random variable.

The dimension of the data (p) often far exceeds the sample size ($n \ll p$). So, statistical rigor is particularly important.

- Clear definitions of statistical notions:
 - parameter
 - parameter estimate
 - consistency
 - confidence
- New asymptotics for $p \rightarrow \infty$ (at what rate relative to n ?)
- Sample size formulas
- Methods to assess reliability of exploratory techniques
 - bootstrap
 - cross-validation

Gene Expression Data

- Observe n copies of a p -dimensional vector $X \sim P$ of gene expression measurements, plus possibly covariates (e.g.: sequence data) and outcomes (e.g.: survival), which can be censored.
- Each gene expression measurement is a ratio, calculated from the intensities of two fluorescently labeled mRNA samples hybridized to an array spotted with known cDNA sequences.
- Data preprocessing may include background subtraction, normalization, log transformation.

Example: Gene expression in tumors vs. healthy tissues plus clinical data of n cancer patients.

Gene Expression Data Analysis

Objective	Methods	Contributions
Subset genes	multiple testing	permutations biased, t-statistics hard
Identify genes (and samples) with similar expression patterns	clustering	PAMSIL, HOPACH, MSS criteria, simultaneous framework
Identify genes with similar association between expression and outcomes	supervised clustering	transformations, IPCW estimators
Prediction e.g.: of gene expression from sequence	classification, regression	cross-validation optimal
Assess reliability of subsets, clusters and predictors	bootstrap	asymptotic validity, cluster probability plots, joint reappearance probs

* Visualization of results is important for all methods.

Multiple Testing for Gene Expression Data

www.bepress.com/ucbbiostat/paper121

Two Sample Problem

Suppose we have n_1 observations from Population 1 with mean μ_1 and n_2 observations from Population 2 with mean μ_2 .

Parameter of Interest: $\mu(P) = \mu_2 - \mu_1 \in \mathfrak{R}$

Null Hypotheses: $H_{0,j} : \mu_j = \mu_{2,j} - \mu_{1,j} = 0, j = 1, \dots, p$

Test Statistics: for $j = 1, \dots, p$

$$D_j = \hat{\mu}_j = \bar{X}_{2,j} - \bar{X}_{1,j}$$
$$\text{or } T_j = \hat{\mu}_j / \text{sd}(\hat{\mu}_j) = \frac{\bar{X}_{2,j} - \bar{X}_{1,j}}{\sqrt{\hat{\sigma}_{1,j}^2/n_1 + \hat{\sigma}_{2,j}^2/n_2}}$$

Reject $H_{0,j}$ if e.g.: $|D_j| > c_j$, where c_j is chosen to control a multiple testing error rate, such as family-wise error rate.

Null Distributions

1. Tables
 - Ignores correlation
 - Same distribution for all genes
2. Resampling (*resample as many times as possible!*)
 - Permutations
 - Can't be used for one sample problems
 - Assumes equality of distributions, not just the parameters of interest
 - Parametric Bootstrap
 - Good choice *if* believe a parametric model
 - Can be computationally intensive (e.g.: decomposing Σ)
 - Non-parametric Bootstrap
 - Good choice when the data generating model is unknown
 - Problems with ties for small sample sizes

Comparison of Null Distributions: Covariance

Let $COV(X_j, X_{j'})$ be ϕ_1 in population 1 and ϕ_2 in population 2.

Distribution	$Var(D_j)$	$Cov(D_j, D_{j'})$
Permutations	$\frac{\sigma_{1,j}^2}{n_2} + \frac{\sigma_{2,j}^2}{n_1}$	$\frac{\phi_1}{n_2} + \frac{\phi_2}{n_1}$
Bootstrap	$\frac{\sigma_{1,j}^2}{n_1} + \frac{\sigma_{2,j}^2}{n_2}$	$\frac{\phi_1}{n_1} + \frac{\phi_2}{n_2}$

Note:

- $VAR(T_j) = 1$ for both distributions.
- But $COV(T_j, T_{j'})$ is not equivalent unless $n_1 = n_2$.

Comparison of Null Distributions: Bias

Suppose $\bar{X}_1 \neq \bar{X}_2$.

Distribution	$E(D_j)$	$E(T_j)$
Permutations	0	$f(\bar{X}_1, \bar{X}_2, \sigma_{1,j}^2, \sigma_{2,j}^2, n_1, n_2) \neq 0$
Bootstrap	0	0

Note:

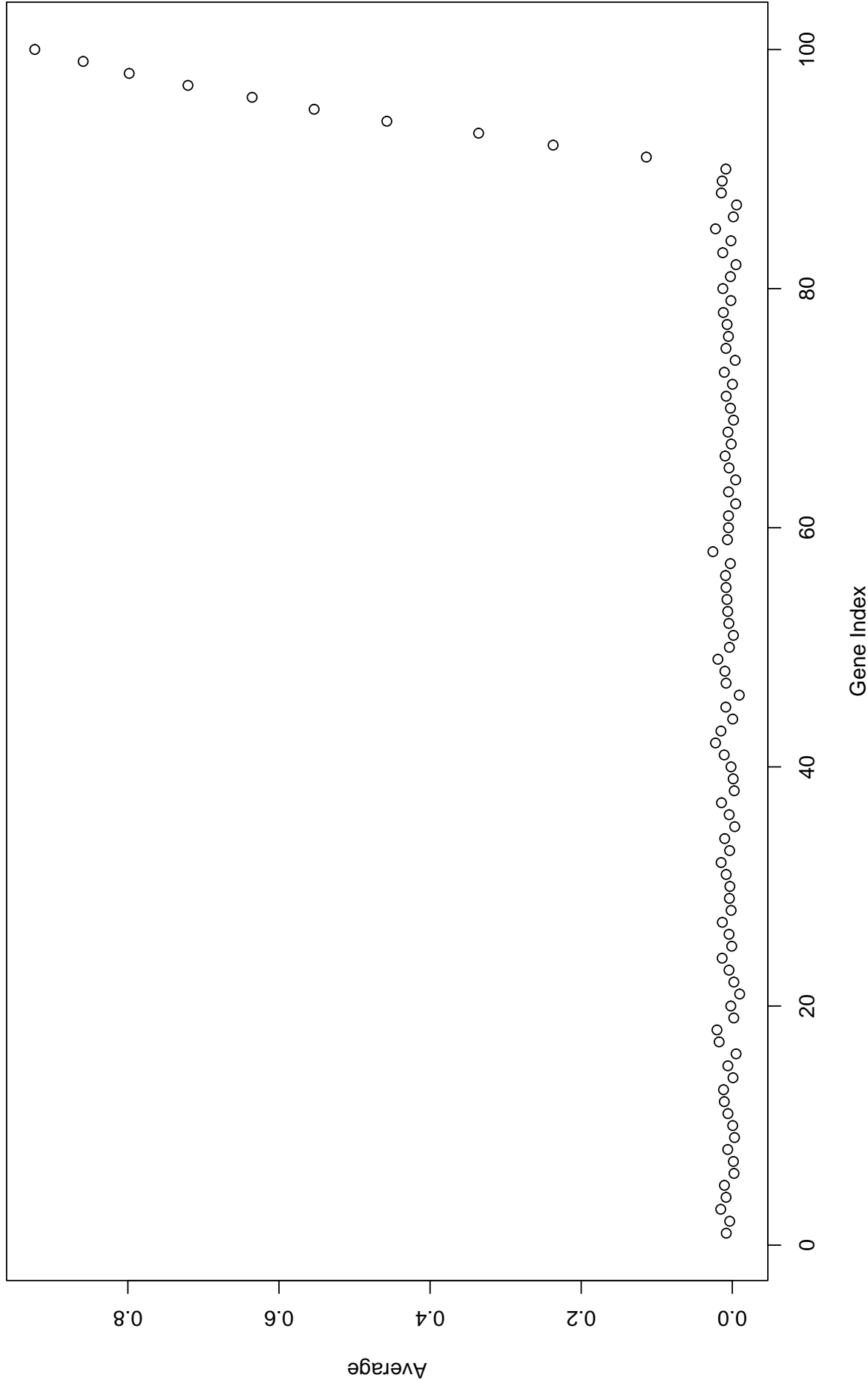
- $E(T_j) = 0$ with permutations if $n_1 = n_2$.
- In practice, also need $n_1 = n_2 > 20$.

Example: Suppose $n_1 = 2, n_2 = 50$ and for gene j Population 1= $(1, 3)$, Population 2= $(0, \dots, 0)$. Then,

$$E(T_j) = \frac{\binom{2}{2} * 2 + \binom{50}{1} * 0.87 + \binom{50}{1} * 0.99 - \binom{50}{2} * 1.27}{\binom{52}{2}} = -1.104 \neq 0.$$

Bias of the Permutation Distribution

Mean of the T-statistic Permutation Distribution for 100 Genes
(genes 91 – 100 are differently expressed)



Multiple Testing: Conclusions

1. Permutation null distribution
 - produces bias with t-statistics whenever $\bar{x}_1 \neq \bar{x}_2$.
 - has the wrong variance for differences in means and the wrong covariance for both statistics.
 - hence, often fails to control the error rate.
 - may be appropriate in some cases, e.g.: balanced samples sizes with t-statistics.
2. Parametric bootstrap performs well when the model is known, but is inferior to the non-parametric bootstrap in real data simulations.
3. In general, the t-statistic's null distribution is harder to estimate than that of the difference in means (e.g.: variance overestimated with the non-parametric bootstrap due to ties when one sample size is small).

HOPACH Algorithm for Clustering Gene Expression Data

www.stat.Berkeley.edu/~laan/Research

www.bepress.com/ucbbiostat/paper107

Background

Clustering algorithms map a $p \times p$ dissimilarity matrix D into parameter estimates (e.g.: p cluster labels, orderings).

Approaches:

- Supervised (COBWEB, SVMs, CART, gene-shaving) vs. Unsupervised
 - Model-based (AUTOCLASS, SNOB) vs. Nonparametric
 - Partitioning (SOMs, PAM, PAMSIL, MASLOC, KMEANS) vs. Hierarchical (HOPACH)
1. Agglomerative (single, complete, and average linkage CLUSTER, AGNES)
 2. Divisive (SOTA, DIANA, TSVQ)
- Graphical approaches (CAST)

Motivation

Simultaneous Clustering: Usually genes and patients clustered separately, but with some two-way visualization methods.

Refs: Tibshirani *et al.* (1999), Getz, Levine and Dommanay (2000), Fellenberg *et al.* (2001), Pollard & van der Laan (2001).

Nice Properties for Clustering Algorithms:

- General distance metric.
- Robust cluster profiles.
- Sensible ordering (hierarchical).
- Non-binary splits (hierarchical).
- Allow simultaneous clustering.
- Identify parameters of biological interest.

Hierarchical Ordered Partitioning and Collapsing Hybrid (HOPACH)

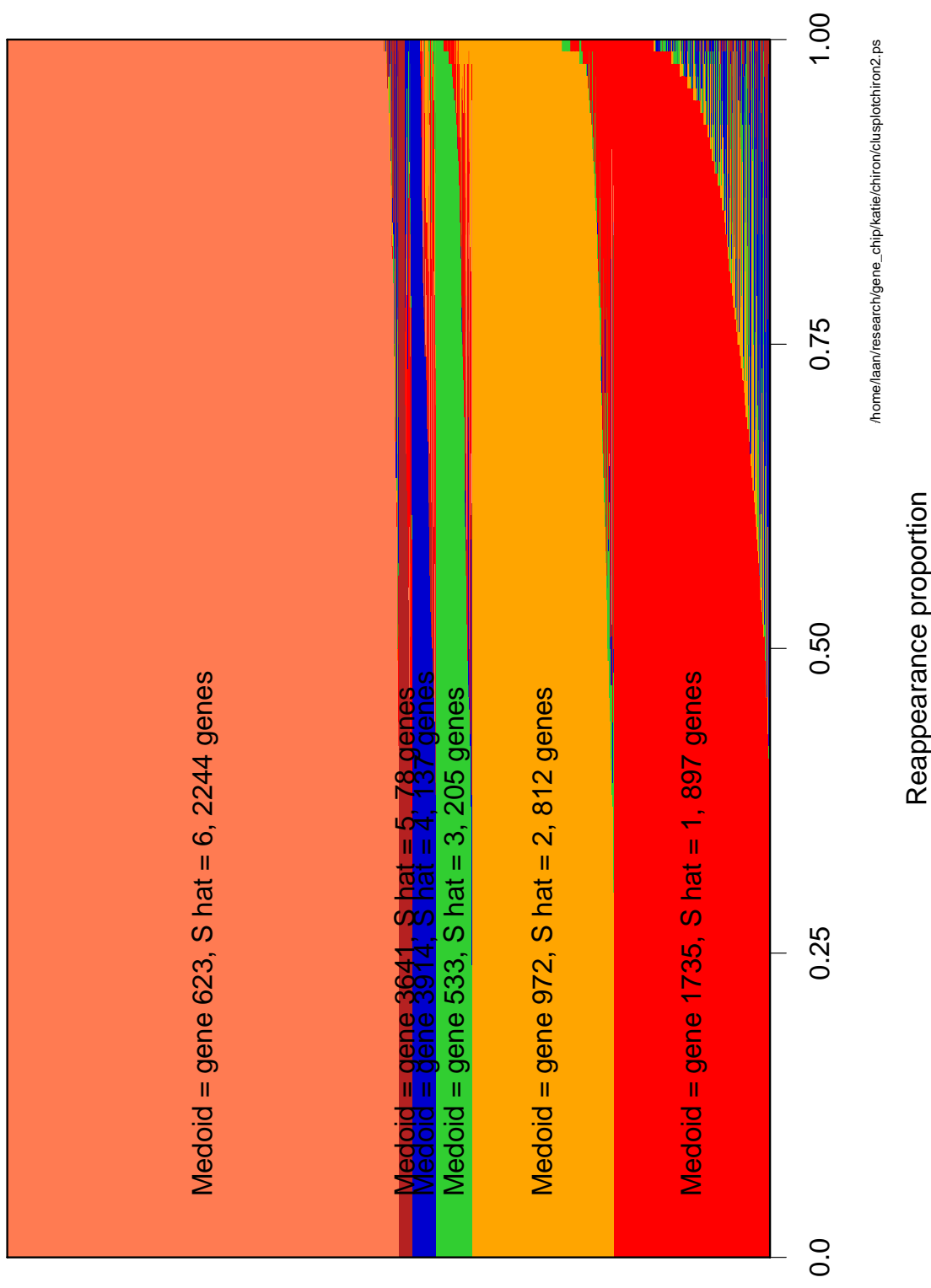
- Builds a tree of clusters.
- At every level, each cluster is split into two or more smaller clusters with a choice of partitioning algorithm (e.g.: PAM).
- Clusters are ordered deterministically based on D .
- Collapsing steps correct errors.
- Mean Split Silhouette (MSS) criteria is used to automate:
 1. Number of clusters in each split
 2. Collapsing steps
 3. Stopping rule (to identify main clusters)
- Produces a final ordering that improves on current algorithms.

Ordered Data and Distance Matrices: Chiron

Cluster Probability Plot: Chiron

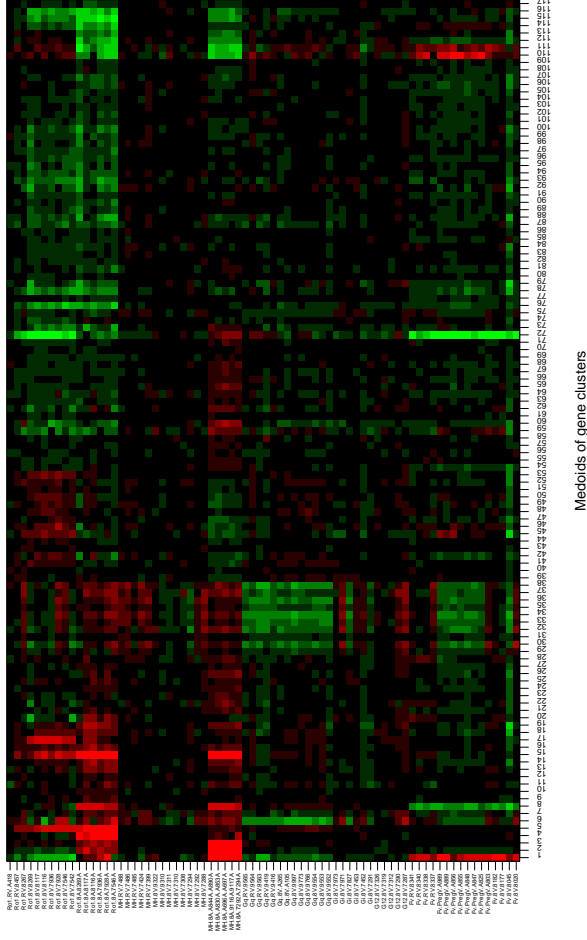
Reappearance proportions and cluster reproducibility

Subset contains 4373 genes.



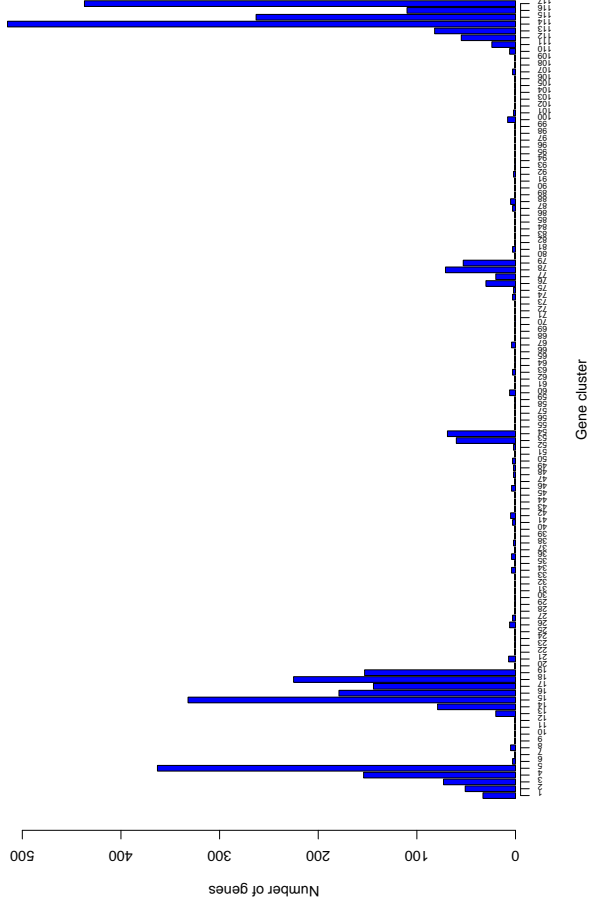
Ordered HOPACH Medoids: Conklin Lab

Heat map of G74 data set showing relative expression of the 117 medoids



Medoids of gene clusters

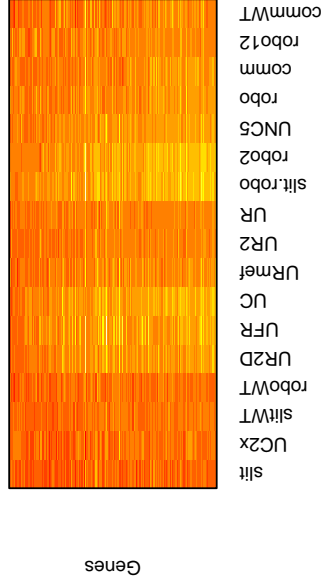
Sizes of 117 gene clusters in G74 data set



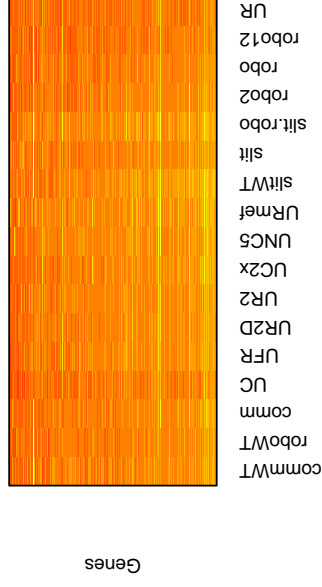
Simultaneous Clustering: *Drosophila*

Clustering of cell lines separately for each gene cluster

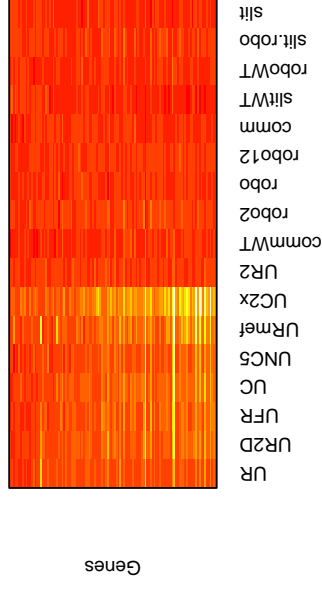
Gene Cluster 1



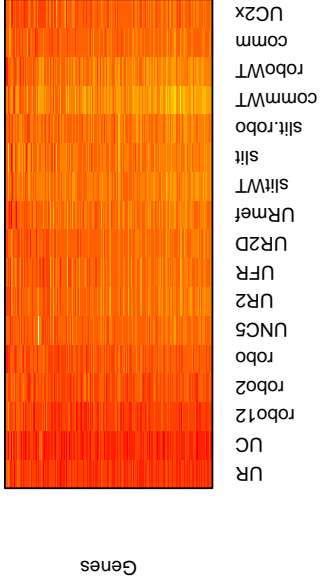
Gene Cluster 2



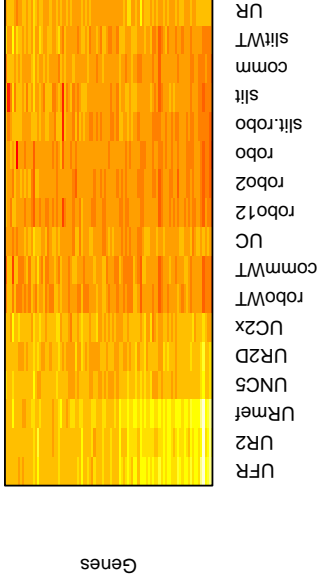
Gene Cluster 3



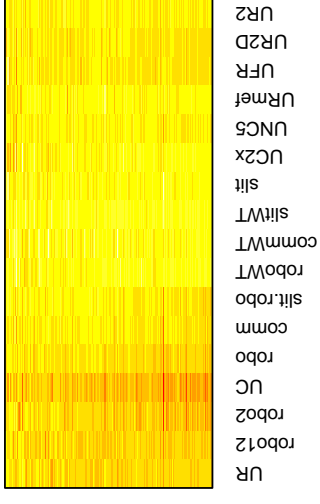
Gene Cluster 4



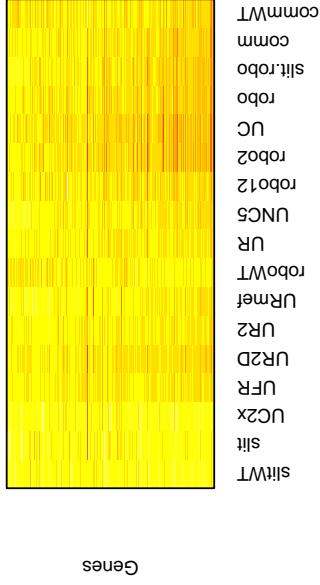
Gene Cluster 5



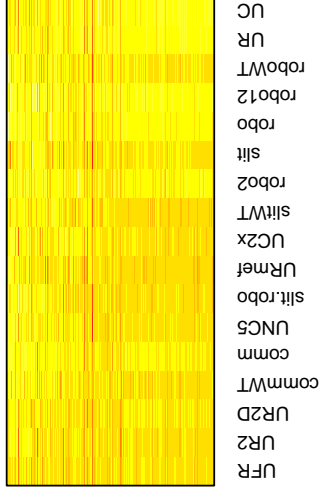
Gene Cluster 6



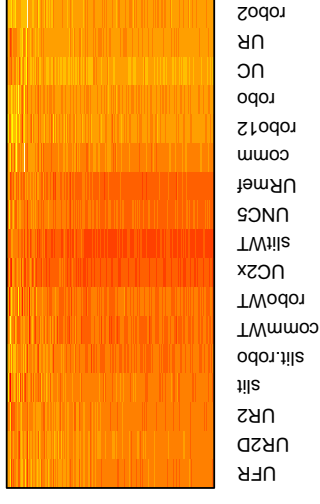
Gene Cluster 7



Gene Cluster 8



Gene Cluster 9



Choosing the Number of Clusters

Problem: Most existing criteria identify global structure only.

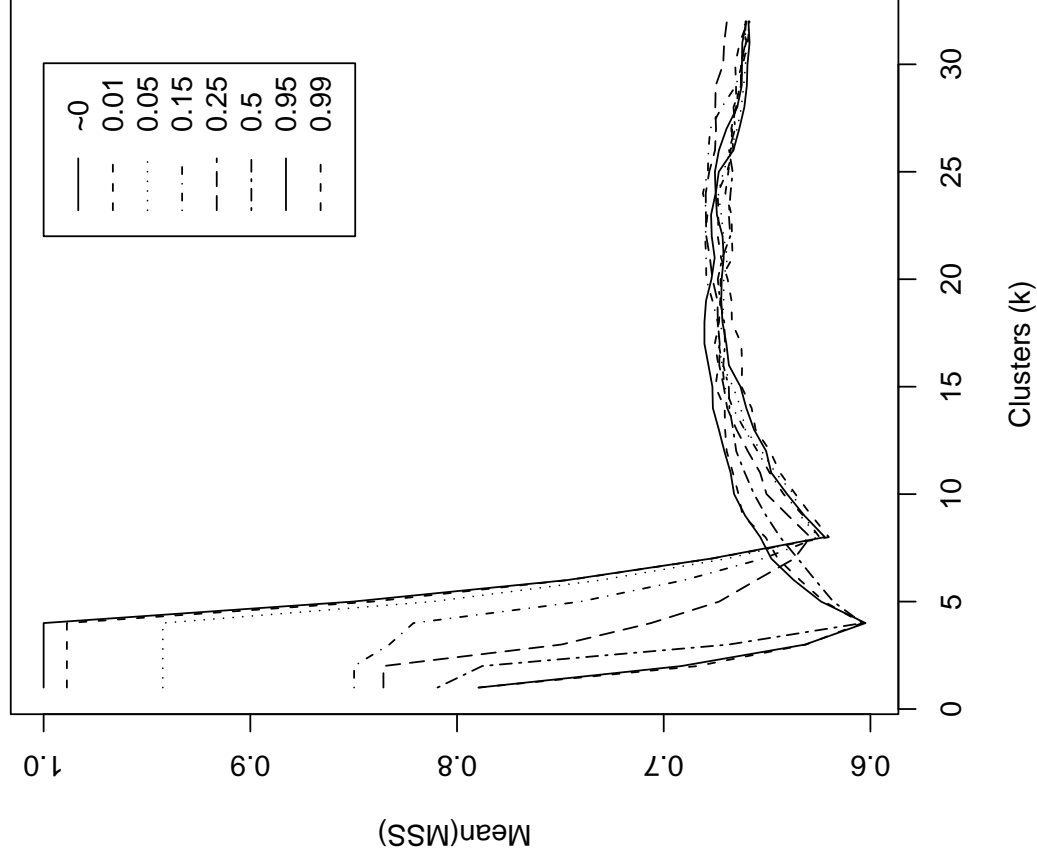
Solutions:

1. Mean Split Silhouette (MSS) criteria
 - Identifies finer structure in the data
 - Chooses fewer clusters if they overlap
 - Can be used to choose 1 cluster
 - Computationally easy
2. Resampling from an appropriate null distribution

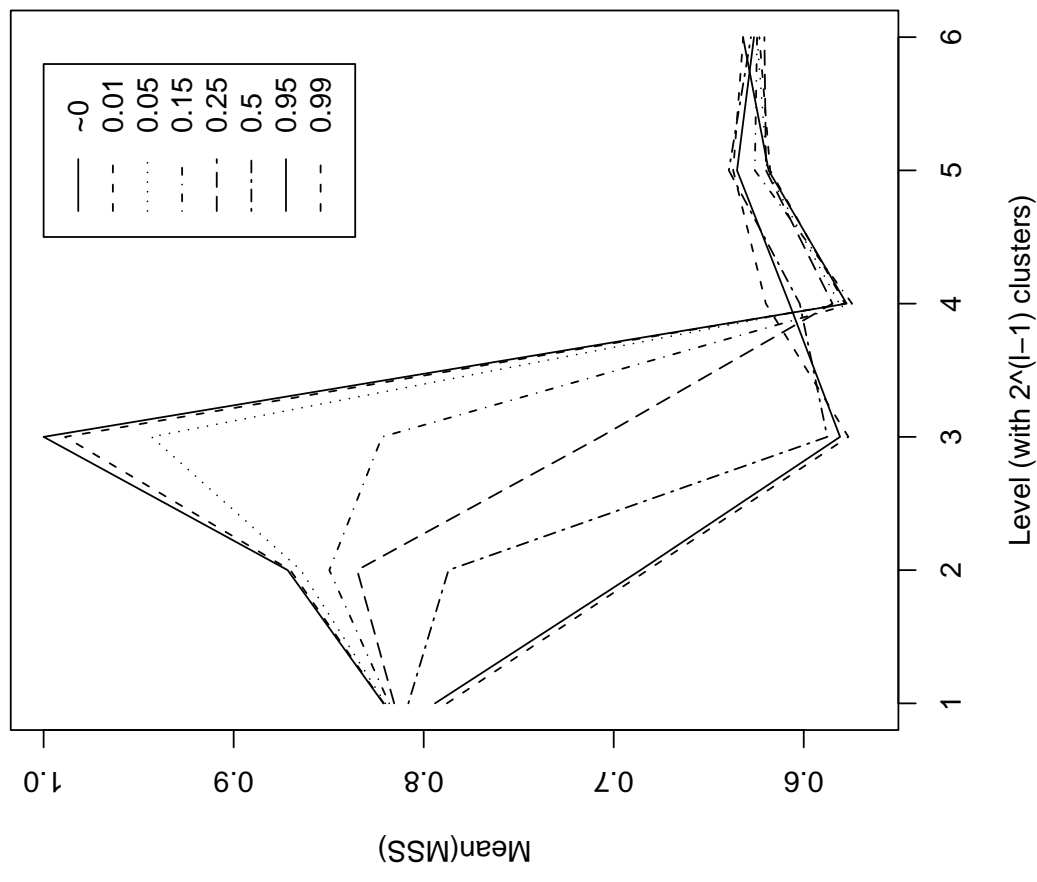
Number of Clusters with MSS

MSS versus number of clusters for eight values of sigma
n=360, mu=(1,2,5,6,14,15,18,19)

PAM (MSS for k clusters)



HOPACH (MSS for l levels)



Choosing the Number of Clusters

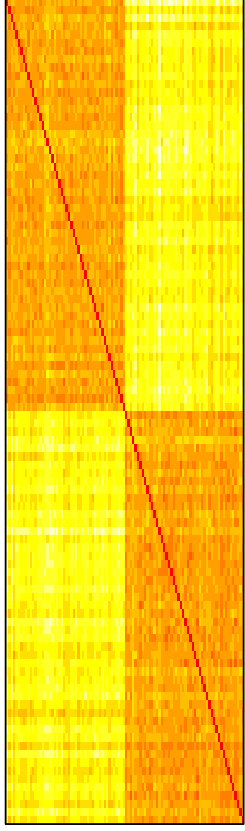
Problem: Most existing criteria identify global structure only.

Solutions:

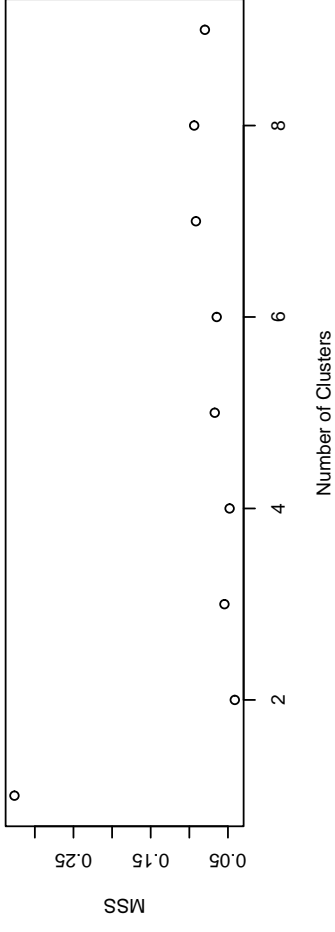
1. Mean Split Silhouette (MSS) criteria
 - Identifies finer structure in the data
 - Chooses fewer clusters if they overlap
 - Can be used to choose 1 cluster
 - Computationally easy
2. Resampling from an appropriate null distribution

MSS with Mixture Distributions

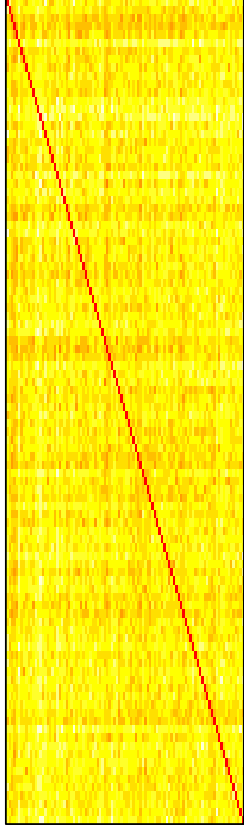
D: $\mu_1=6$, $\mu_2=3$, $\sigma_1=\sigma_2=2$, $q=0.5$



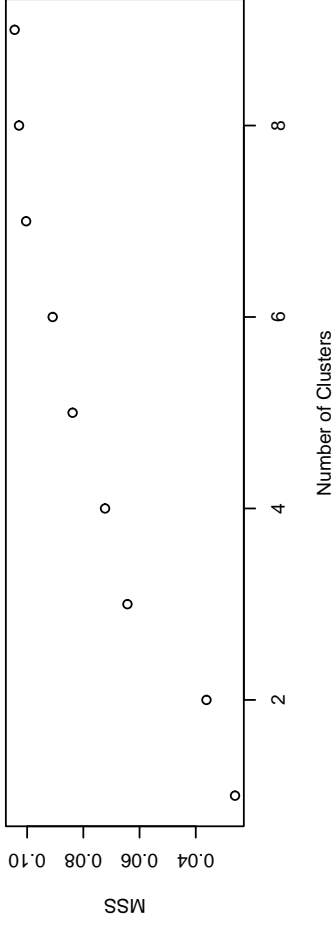
Mean Split Silhouette



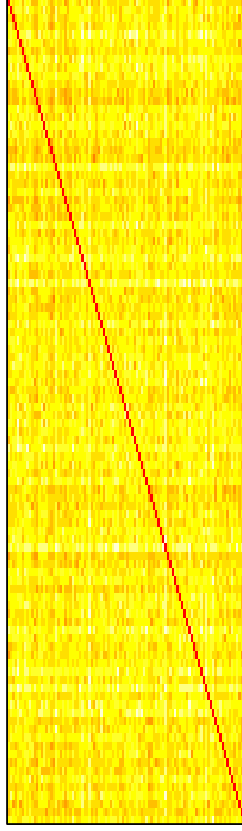
D: $\mu_1=6$, $\mu_2=3$, $\sigma_1=\sigma_2=10$, $q=0.5$



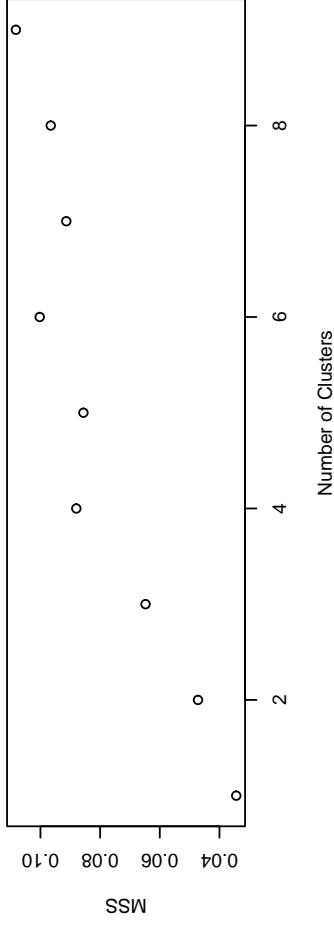
Mean Split Silhouette



D: $\mu_1=6$, $\sigma_1=10$, $q=1$



Mean Split Silhouette



Choosing the Number of Clusters

Problem: Most existing criteria identify global structure only.

Solutions:

1. Mean Split Silhouette (MSS) criteria
 - Identifies finer structure in the data
 - Chooses fewer clusters if they overlap
 - Can be used to choose 1 cluster
 - Computationally easy
2. Resampling from an appropriate null distribution

Clustering: Conclusions

- Simultaneous clustering reveals interesting patterns in gene expression data, for example, subpopulations with different gene expression patterns.
- Statistical inference is possible with gene expression data using the bootstrap and appropriate null distributions.
- New clustering algorithms help us to find biologically meaningful groups of genes and samples and to produce sensible ordered lists.
- Visualization of ordered data and distance matrices reveals the underlying cluster structure.
- MSS criteria identifies the finer structure in gene expression data.

Summary: Statistical Approach

Given n observations from a data generating distribution P :

1. What is the parameter of interest $S(P)$?
2. Is the empirical estimate $S(P_n)$ consistent?
3. Do different algorithms estimate the same parameter?
 - If not, which do we prefer?
 - If so, what are their relative efficiencies?
4. How reliable is the observed estimate $S(P_n)$?
 - Resampling-based estimates of the distribution of $S(P_n)$.
 - Sample size formulas for a given accuracy.

Interesting Directions

1. Combining gene expression with other types of data:
 - Regulatory motif detection by regression of sequence words on gene expression
 - Refinement of QTL mapping with gene expression
 - Gene clustering supervised by (clinical) outcomes
 - Relating gene expression and chromosome location
 - Functional annotation by gene clustering
 - Refining gene prediction with oligo expression arrays
 - Connecting genotype to disease susceptibility based on changes in gene expression
2. Applying methods to other high-dimensional data:
 - Comparative genomic hybridization (CGH) data
 - Proteomics
 - Detecting selection in the genome by haplotype structure
 - Phylogenetics, phylogenomics, comparative genomics