UNIVERSITÄT LEIPZIG

Classification and Clustering of RNAseq data

Verena Zuber

IMISE, University of Leipzig

5th June 2012

The presented publication

The Annals of Applied Statistics
2011, Vol. 5, No. 4, 2493–2518
DOI: 10.1214/11-AOAS493

CLASSIFICATION AND CLUSTERING OF SEQUENCING DATA USING A POISSON MODEL

BY DANIELA M. WITTEN

University of Washington

Author of the publication: Daniela Witten

Table of contents

Biological Background: Transcriptomics I

Technologies to "measure" the transcriptome:

- Microarrays
- Next or second generation RNA sequencing (RNAseq)

Limitations of microarrays:

- High levels of background noise due to cross-hybridization
- Only transcripts for which a probe is present on the array can be measured. Therefore, it is not possible to discover novel mRNAs in a typical microarray experiment.

Biological Background: Transcriptomics II

Promises of RNAseq

- Less noisy than microarray data, since the technology does not suffer from cross-hybridization.
- Detection of novel transcripts and coding regions
- "It seems certain that RNA sequencing is on track to replace the microarray as the technology of choice for the characterization of gene expression."

Challenges in the analysis:

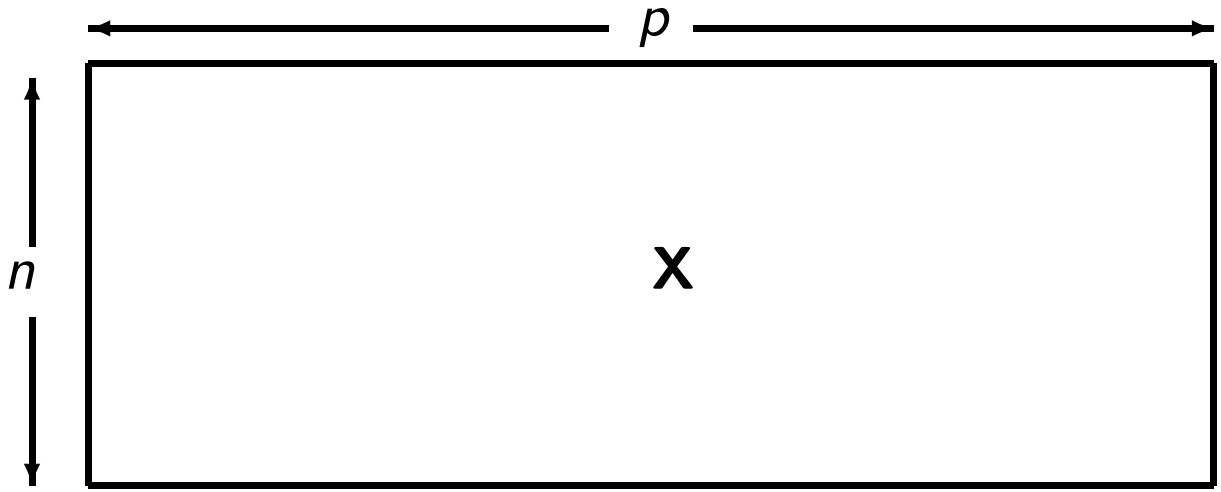
- Normalization
- Count data, integer valued and non-negative

Statistical Framework

Data structure

X $n \times p$ matrix of sequencing data

- $i \in 1, ..., n$ samples
- $j \in 1, ..., p$ features or regions of interest



- s_i sample-specific constant
- g_j gene-specific constant

Distributions

- Poisson distribution

$$X_{ij} \sim \text{Poisson}(N_{ij}), \qquad N_{ij} = s_i g_j$$

  - expectation: $\text{E}(X_{ij}) = N_{ij}$
  - variance: $\text{Var}(X_{ij}) = N_{ij}$

- Negative binomial distribution

$$X_{ij} \sim \text{NB}(N_{ij}, \phi_j), \qquad N_{ij} = s_i g_j$$

  - ϕ_j is a gene-specific over-dispersion
  - expectation: $\text{E}(X_{ij}) = N_{ij}$
  - variance: $\text{Var}(X_{ij}) = N_{ij} + N_{ij}^2 \phi_j$

Distributions dependent on class k

- $y_i = k \in 1, ..., K$: factor indicating the membership of sample i to class k
- Poisson distribution

$$X_{ij} \mid y_i = k \sim \text{Poisson}(N_{ij} d_{kj}), \qquad N_{ij} = s_i g_j$$

- Negative binomial distribution

$$X_{ij} \mid y_i = k \sim \text{NB}(N_{ij} d_{kj}, \phi_j), \qquad N_{ij} = s_i g_j$$

- d_{kj}: gene-specific, class-specific factor
  - $d_{kj} > 1$ indicates that the jth feature is over-expressed in class k relative to the baseline
  - $d_{kj} < 1$ indicates that the jth feature is under-expressed in class k relative to the baseline
- C_k comprises all samples belonging to class k

Poisson Log Linear Model

Assumptions:

- Poisson distribution
- Independence of features

Poisson log linear model

$$X_{ij} \mid y_i = k \sim \text{Poisson}(\hat{N}_{ij}\hat{d}_{kj}), \qquad \hat{N}_{ij} = \hat{s}_i\hat{g}_j$$

Estimation of the gene-specific constant g_j:

- $\hat{g}_j = \sum_{i=1}^{n} X_{ij}$

Poisson Log Linear Model

Estimation of the sample-specific constant s_i
(under identifiability constraint $\sum_{i=1}^{n} \hat{s}_i = 1$):

- Total count (ML-estimate):
  $\hat{s}_i = \sum_{j=1}^{p} X_{ij} / \sum_{i=1}^{n} \sum_{j=1}^{p} X_{ij}$
- Median ratio (Anders and Huber (2010)):
  $\hat{s}_i = m_i / \sum_{i=1}^{n} m_i$

$$m_i = \text{median}_j \left(\frac{X_{ij}}{(\Pi_{i'}^{n} X_{i'j})^{1/n}} \right)$$

- Quantile (Bullard et al. (2010)):
  $\hat{s}_i = q_i / \sum_{i=1}^{n} q_i$, where q_i is the 75th percentile of the counts for each sample

Poisson Log Linear Model

Estimation of the (gene and) class-specific factor d_{kj}:

- Maximum likelihood estimate

$$\hat{d}_{kj} = X_{C_k j} / \sum_{i \in C_k} \hat{N}_{ij}$$

- If $X_{C_k j} = 0$, then $\hat{d}_{kj} = 0$.
  "This can pose a problem for downstream analyses, since this estimate completely precludes the possibility of a nonzero count for feature j arising from an observation in class k."
- Bayesian estimate: Gamma(β, β) prior on d_{kj} results in the following posterior mean

$$\hat{d}_{kj} = \frac{X_{C_k j} + \beta}{\sum_{i \in C_k} \hat{N}_{ij} + \beta}$$

Transformation for overdispersed data

- Biological replicates of sequencing data tend to be overdispersed relative to the Poisson model (variance is larger than the expectation)
- Power transformation $X'_{ij} \leftarrow X_{ij}^{\alpha}$ where $\alpha \in (0, 1]$ is chosen so that

$$\sum_{i=1}^{n}\sum_{j=1}^{p}\frac{(X'_{ij}-\chi')^2}{\chi'} \approx (n-1)(p-1)$$

with $\chi' = \left(\frac{\sum_{j=1}^{p} X'_{ij} \sum_{i=1}^{n} X'_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{p} X'_{ij}}\right)$ (Goodness of fit test!)

- "Though the resulting transformed data are not integer-valued, we nonetheless model them using the Poisson distribution."

Supervised Learning: Classification

Poisson linear discriminant analysis

- Rather diagonal discriminant analysis (DDA)
  due to the independence assumption
- Bayes' rule to define the probability of belonging to class k
  depending on the test data $\mathbf{x}^\star$

$$\begin{aligned} \text{prob}(k|\mathbf{x}^\star) &= \frac{\pi_k f(\mathbf{x}^\star|k)}{f(\mathbf{x}^\star)} \\ &\propto \pi_k f(\mathbf{x}^\star|k) \end{aligned}$$

- where $f(\mathbf{x}^\star|k)$ is given by

$$X_{ij} \mid y_i = k \sim \text{Poisson}(N_{ij} d_{kj}), \qquad N_{ij} = s_i g_j$$

- π_k represents the a priori mixing probability for class k

Discriminant scores

- Poisson discriminant analysis

$$\begin{aligned} \log\{\text{prob}(k|\mathbf{x}^\star) &= \frac{\pi_k f(\mathbf{x}^\star|k)}{f(\mathbf{x}^\star)} \\ &\propto \sum_{j=1}^{p} X_j^\star \log \hat{d}_{kj} - \hat{s}^\star \sum_{j=1}^{p} \hat{g}_j \hat{d}_{kj} + \log \hat{\pi}_k \end{aligned}$$

- For comparison: Fisher's DDA (Gaussian Distribution)

$$\begin{aligned} \log\{\text{prob}(k|\mathbf{x}^\star)\} &= \frac{\pi_k f(\mathbf{x}^\star|k)}{f(\mathbf{x}^\star)} \\ &\propto \boldsymbol{\mu}_k^T \mathbf{V}^{-1} \mathbf{x}^\star - \frac{1}{2} \boldsymbol{\mu}_k^T \mathbf{V}^{-1} \boldsymbol{\mu}_k + \log(\pi_k) \end{aligned}$$

where $\boldsymbol{\mu}_k$ is the expectation in group k, and $\mathbf{V}$ is the diagonal variance matrix equal in all K groups

The sparse PLDA classifier

- Standard estimates $\hat{d}_{kj}$ are unequal 1 for all p features
- But for high-dimensional transcriptomics data classifiers build on a smaller subset of features are desirable
- Soft-threshold estimate (similar to PAM)

$$\hat{d}_{kj} = 1 + S(a/b - 1, \rho/\sqrt{b})$$

  - Soft-threshold operator with penalization-parameter ρ

$$S(a/b - 1, \rho/\sqrt{b}) = \text{sign}(a/b - 1)(\mid a/b - 1 \mid - \rho/\sqrt{b})_+$$

  - $a = X_{C_k j} + \beta$ (numerator of the Bayesian estimate $\hat{d}_{kj}$)
  - $b = \sum_{i \in C_k} \hat{N}_{ij} + \beta$ (denominator of the Bayesian estimate $\hat{d}_{kj}$)
- Shrinks $\hat{d}_{kj}$ towards 1 if $\mid a/b - 1 \mid < \rho/\sqrt{b}$, and thus excludes feature j from the classification rule

Unsupervised Learning: Clustering

Poisson dissimilarity

- Aim: Clustering based on a $n \times n$ dissimilarity matrix between observations
- Connection of Euclidean distance and log likelihood ratio statistic under a Gaussian model

$$X_{ij} \sim \mathrm{N}(\mu_{ij}, \sigma^2) \qquad X_{i'j} \sim \mathrm{N}(\mu_{i'j}, \sigma^2)$$

Testing $H_0 : \mu_{ij} = \mu_{i'j}$ against
H_1: "μ_{ij} and $\mu_{i'j}$ are unrestricted"
results in the following log likelihood ratio statistic

$$\begin{aligned} \sum_{j=1}^{p}(X_{ij} - \frac{X_{ij} + X_{i'j}}{2}) + \sum_{j=1}^{p}(X_{i'j} - \frac{X_{ij} + X_{i'j}}{2}) &= \sum_{j=1}^{p}(X_{ij} - X_{i'j})^2 \\ &\propto || \mathbf{x}_i - \mathbf{x}_j ||^2 \end{aligned}$$

Poisson dissimilarity

- Poisson distribution "restricted to $\mathbf{x}_i$ and $\mathbf{x}_{i'}$"

$$X_{ij} \sim \text{Poisson}(\hat{N}_{ij}\hat{d}_{ij}) \qquad X_{i'j} \sim \text{Poisson}(\hat{N}_{i'j}\hat{d}_{i'j})$$

- Testing $H_0 : d_{ij} = d_{i'j} = 1$ against
  H_1: "d_{ij} and $d_{i'j}$ are unrestricted results"
  results in the following log likelihood ratio statistic

$$\sum_{j=1}^{p}\left((\hat{N}_{ij} + \hat{N}_{i'j}) - (\hat{N}_{ij}\hat{d}_{ij} + \hat{N}_{i'j}\hat{d}_{i'j}) + (X_{ij}\log\hat{d}_{ij} + X_{i'j}\log\hat{d}_{i'j})\right)$$

- Can be used as dissimilarity of $\mathbf{x}_i$ and $\mathbf{x}_{i'}$; is nonnegative and equals zero if $\mathbf{x}_i = \mathbf{x}_{i'}$

Results

Simulation set up

Data is generated by the negative binomial distribution

$$X_{ij} \mid y_i = k \sim \text{NB}(s_i g_j d_{kj}, \phi)$$

- Overdispersion
  - $\phi = 0.01$: very slight overdispersion
  - $\phi = 0.1$: substential overdispersion
  - $\phi = 1$: very high overdispersion
- $s_i \sim \text{Unif}(0.2, 2.2)$
- $g_j \sim \text{Exp}(1/25)$
- $K = 3$ classes
- $p = 10,000$ features and 30% are differentially expressed
- $d_{1j} = d_{2j} = d_{3j} = 1$: feature j is not differentially expressed
- otherwise $\log(d_{kj}) \sim N(0, \sigma^2)$

Real sequencing data sets

- Liver and kidney
  The data are available as a Supplementary File associated with Marioni et al. (2008)
- Yeast
  The data are available as a Supplementary File associated with Anders and Huber (2010)
- Cervical cancer (Witten et al. (2010))
  The data are available from Gene Expression Omnibus [Barrett et al. (2005)] under accession number GSE20592
- Transcription factor binding
  The data are available as a Supplementary File associated with Anders and Huber (2010)

Competitors

1. Classification
   - Nearest Shrunken Centroid (NSC)
   - Nearest Shrunken Centroid with sqrt error transformation
2. Clustering
   - EdgeR (Robinson, McCarthy, and Smyth (2010))
   - Variance Stabilizing Transformation (VST) according to Anders and Huber (2010)
   - Euclidean distance

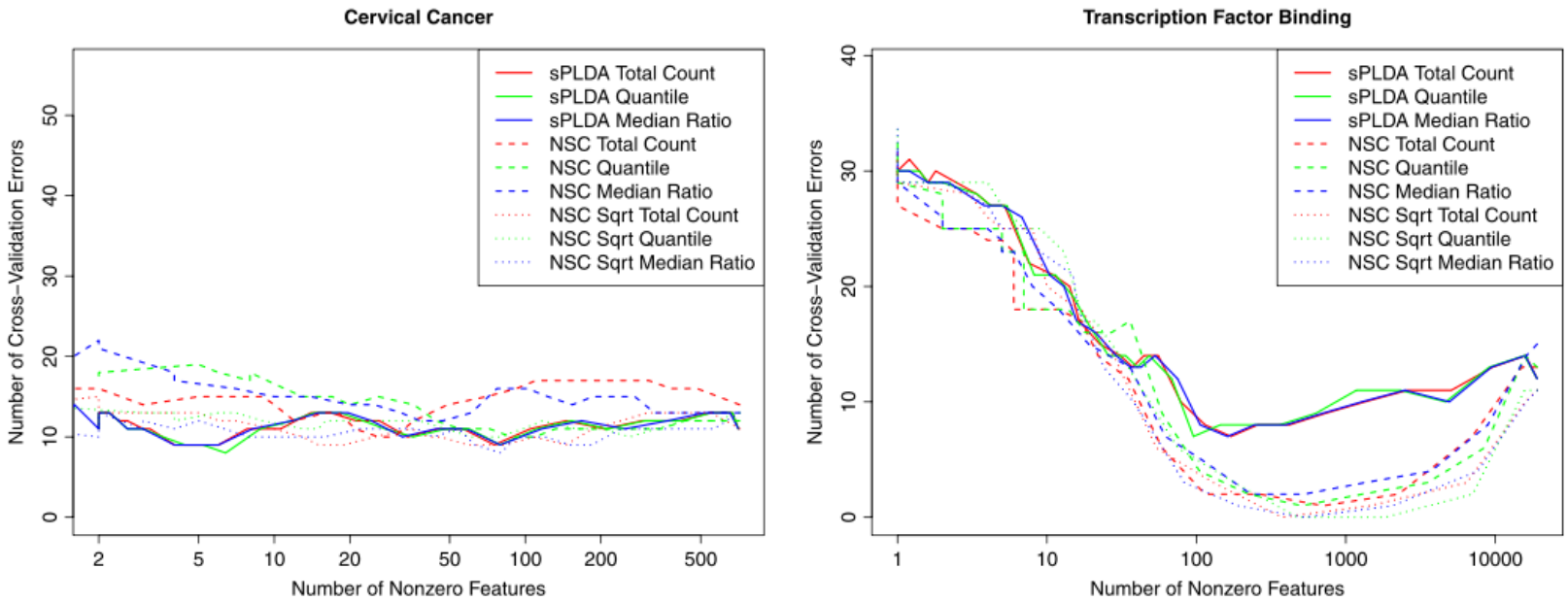
Simulation: Classification results

TABLE 2

Simulation results: nine classification methods. NSC, NSC on $\sqrt{X_{ij} + 3/8}$*, and sPLDA were performed, using three different size factor estimates: total count (TC), quantile (Q), and median ratio (MR). Cross-validation was performed on a training set of n observations, and error rates were computed on n test observations. We report the mean numbers of test errors and nonzero features over 50 simulated data sets. Standard errors are in parentheses*

n	ϕ	σ	Method	NSC err.	NSC sqrt err.	sPLDA err.	NSC nonzero	NSC sqrt nonzero	sPLDA nonzero
12	0.01	0.05	TC	4.18 (0.34)	5.74 (0.28)	2.24 (0.26)	1947.6 (441.3)	2217.9 (509.0)	791.4 (111.7)
			Q	4.38 (0.34)	5.82 (0.26)	2.26 (0.25)	1670.6 (394.5)	2010.1 (478.2)	782.3 (110.0)
			MR	4.28 (0.34)	5.78 (0.27)	2.20 (0.24)	1731.8 (402.6)	2327.8 (517.9)	795.4 (110.8)
50	0.01	0.025	TC	19.14 (0.67)	24.06 (0.70)	16.84 (0.55)	2316.6 (398.8)	3122.5 (516.6)	1830.7 (217.9)
			Q	20.32 (0.71)	24.82 (0.70)	17.14 (0.56)	1870.7 (335.2)	3380.9 (519.4)	1860.2 (229.6)
			MR	19.66 (0.69)	24.48 (0.69)	16.88 (0.60)	2488.7 (437.8)	2698.7 (513.4)	1934.9 (224.5)
12	0.1	0.1	TC	2.52 (0.31)	2.66 (0.26)	1.58 (0.25)	5143.2 (527.2)	2738.5 (461.2)	3878.2 (369.4)
			Q	2.12 (0.27)	2.68 (0.26)	1.62 (0.26)	5207.0 (536.8)	2879.8 (456.7)	3927.2 (371.7)
			MR	2.28 (0.29)	2.88 (0.28)	1.60 (0.26)	4849.4 (531.2)	2932.0 (477.3)	3889.2 (368.6)
50	0.1	0.05	TC	16.80 (0.54)	17.76 (0.61)	17.94 (0.70)	3802.5 (408.1)	3785.5 (418.1)	3308.5 (355.1)
			Q	17.08 (0.64)	17.16 (0.60)	17.88 (0.65)	4293.2 (479.1)	3921.5 (371.8)	3284.0 (352.8)
			MR	16.78 (0.59)	17.34 (0.66)	17.96 (0.71)	3475.3 (392.4)	4398.3 (457.0)	3489.3 (371.9)
12	1	0.2	TC	3.24 (0.25)	4.28 (0.35)	4.26 (0.32)	8846.2 (380.7)	6127.8 (524.6)	4502.0 (509.9)
			Q	3.20 (0.23)	4.04 (0.33)	4.08 (0.29)	8991.8 (318.8)	6342.1 (557.6)	4551.7 (512.1)
			MR	3.22 (0.26)	3.60 (0.30)	4.00 (0.31)	8389.0 (396.4)	7082.7 (515.5)	4518.9 (514.6)
50	1	0.1	TC	25.56 (0.61)	25.80 (0.55)	25.66 (0.50)	4237.8 (503.5)	4293.5 (495.9)	3150.5 (433.0)
			Q	25.82 (0.61)	25.90 (0.64)	26.02 (0.55)	4629.1 (516.2)	4170.5 (491.7)	3131.2 (406.6)
			MR	25.92 (0.68)	25.86 (0.59)	25.52 (0.51)	4427.5 (524.0)	4362.6 (498.0)	3156.8 (410.4)

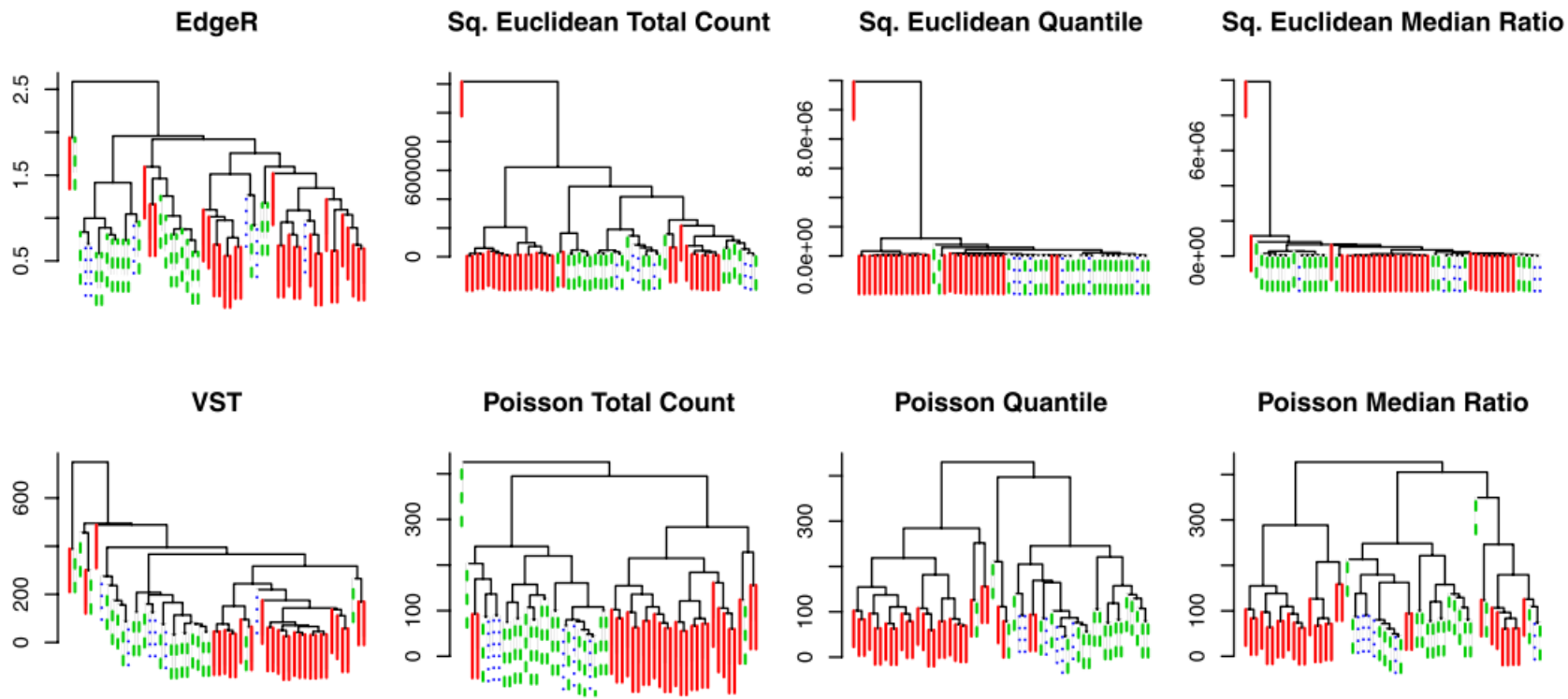
Sequencing data: Classification results



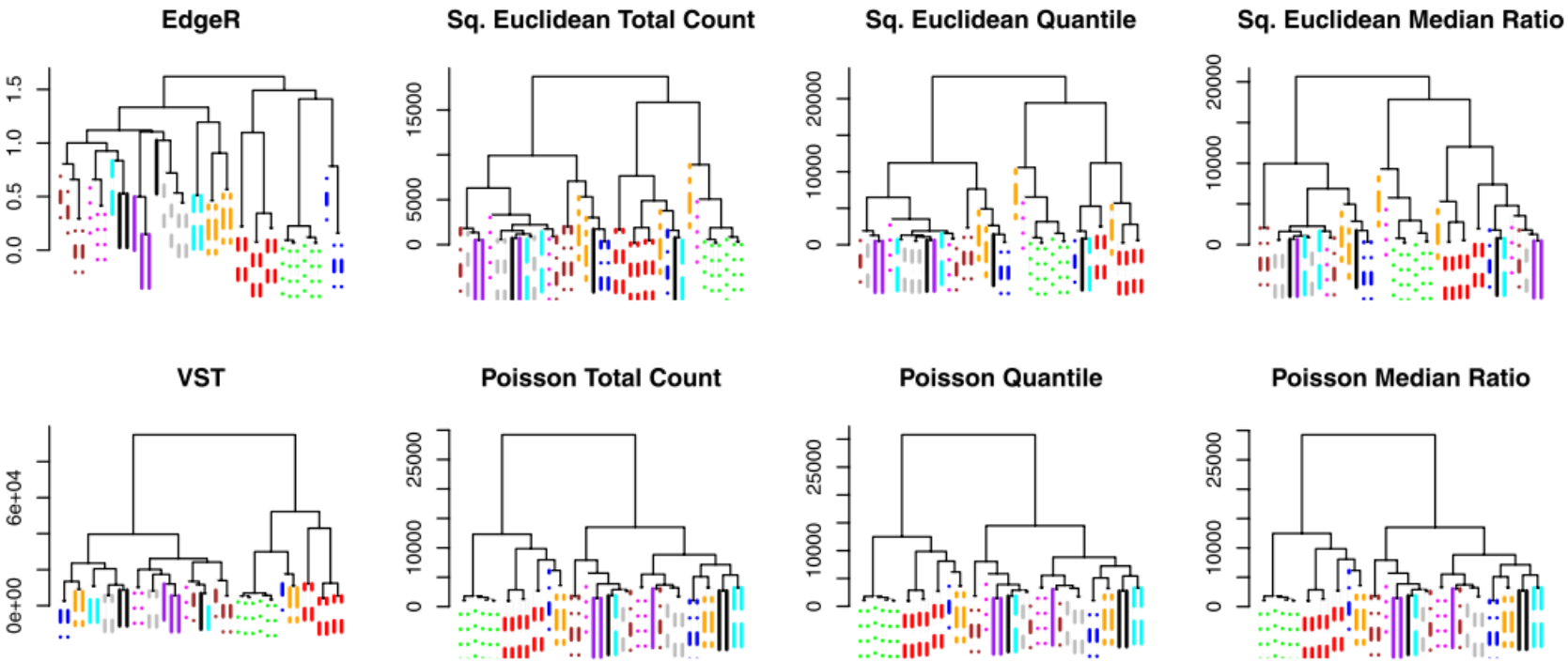
Simulation: Clustering results

ϕ	σ	Method	Clustering error rate
0.01	0.15	Cai	0.3592 (0.0071)
		Berninger	0.5704 (0.0173)
		EdgeR	0.0000 (0.0000)
		VST	0.6201 (0.0029)
		Squared Euclidean total count	0.5675 (0.0191)
		Squared Euclidean quantile	0.5662 (0.0215)
		Squared Euclidean median ratio	0.5755 (0.0178)
		Poisson total count	0.0045 (0.0045)
		Poisson quantile	0.0057 (0.0047)
		Poisson median ratio	0.0045 (0.0045)
0.1	0.2	Cai	0.3803 (0.0058)
		Berninger	0.1905 (0.0258)
		EdgeR	0.0000 (0.0000)
		VST	0.6204 (0.0029)
		Squared Euclidean total count	0.3051 (0.0327)
		Squared Euclidean quantile	0.2875 (0.0325)
		Squared Euclidean median ratio	0.3297 (0.0350)
		Poisson total count	0.2053 (0.0225)
		Poisson quantile	0.2067 (0.0228)
		Poisson median ratio	0.2006 (0.0219)

Sequencing data: Normal vs cancer



Sequencing data: Technical replicates of $n = 10$



Discussion I

- Transcript length bias
  "It seems clear that bias due to the total number of counts per feature is undesirable for the task of identifying differentially expressed transcripts, since it makes it difficult to detect differential expression for low-frequency transcripts. However, it is not clear that such a bias is undesirable in the case of classification or clustering, since we would like features about which we have the most information—namely, the features with the highest total counts—to have the greatest effect on the classifiers and dissimilarity measures that we use."

Discussion II

- Normalization
  " It has been shown that the manner in which samples are normalized is of great importance in identifying differentially expressed features on the basis of sequencing data [Bullard et al. (2010), Robinson and Oshlack (2010), Anders and Huber (2010)]. However, in Sections 5 and 6, the normalization approach appeared to have little effect on the results obtained. This seems to be due to the fact that the choice of normalization approach is most important when a few features with very high counts are differentially expressed between classes. In that case, identification of differentially expressed features can be challenging, but classification and clustering are quite straightforward."

Discussion III

- Poisson or Negative binomial distribution?
  "The methods proposed seem to work very well if the true model for the data is Poisson or if there is mild overdispersion relative to the Poisson model. Performance degrades in the presence of severe overdispersion. Most sequencing data seem to be somewhat overdispersed relative to the Poisson model. It may be that extending the approaches proposed here to the negative binomial model could result in improved performance in the presence of overdispersion."

Discussion IV

- Independence assumption?
- Transformation into non-integer values?
- Simulation results?
  - Classification: No clear superiority over NSC in overdispersed simulated or in real data
  - Clustering: EdgeR performs best