Biochem 158 Final Project

Kris Sankaran

June 7, 2013

1 Introduction

In this project, we first review considerations for designing microRNA (miRNA) profiling experiments that have appeared in recent literature and then explain and reproduce some of the statistical inferential techniques proposed in [14] for working with the high-throughput data extracted from these experiments. In addition to supplying `R` code within the text of this paper, all preprocessed data and `R` analysis can be downloaded from http://github.com/krisrs1128.

miRNAs are small (∼22 nucleotide) noncoding RNAs that facilitate regulation of gene expression through translational inhibition [5, 11]. Further, though it is estimated that there are only several thousand human miRNAs (there are 2042 mature human miRNA sequences indexed in mirBASE as of June 2013), they control expression patterns many coding genes; of particular interest, in the last decade, it has been found that miRNAs regulate transitions in plant and animal development, including tissue differentiation [9, 15, 3]. Further, there are relevant links between the study of miRNA function and research involving stem cells and cancer [1, 2, 13].

Hence, the quantification of miRNAs expression levels between diverse biological samples has become an important activity both for fundamental biology, especially work to understand the mechanisms of gene regulation, and medical research, such as the identification of

disease biomarkers. The diverse methodologies developed towards these ends are referred to as miRNA profiling. There is no standard, one-size-fits-all approach to the design of profiling experiments [5, 11]. However, the assortment of profiling techniques share the same overall pipeline: RNA extraction, quantification and quality control, and profiling [11]. The specific choices at stages in this pipeline, along with the research and sample considerations that guide these choices, are the focus of section 2. In section 3, we transition to the problem of inferring interesting patterns from profiling data collected from RNA-seq profiling platforms. In particular, we focus on explaining and illustrating the high-dimensional statistical learning machinery found useful in [14].

2 miRNA profiling pipeline and considerations

2.1 RNA extraction and sample types

The first step of any miRNA profiling experiment is the collection of RNA for subsequent statistical inspection. miRNAs can be extracted from a number of sample types, including cell lines, fresh tissues, FACS, FFPE tissues, plasma, serum, and urine [11]. Cell lines and fresh tissues typically have the highest yield, but can be more difficult to prepare; body fluids typically have the smallest yield, but are easy to collect. Fixed tissues can be convenient to work with, and it has been reported that, unlike the longer mRNAs, miRNAs are stable in fixed tissues.

Once a sample type has been selected, standard RNA extraction procedures can be applied, though they are typically modified to promote enrichment of small RNAs. A number of commercially available kits are available to isolate miRNAs; these are based on the appropriate chemical extraction protocols. To increase representation of small RNAs, size fractionation is common. However, this step is not necessary, as profiling technologies, briefly reviewed in section 2.3, are increasingly capable of distinguishing miRNAs from other RNAs.

Indeed, most profilers are capable of resolving the smaller mature miRNAs from the larger precursor and primary transcript miRNAs, despite the overlaps in base pairs [11, 5].

2.2 Quantification and quality assurance

Quantification is generally achieved through spectrophotometry or automated capillary electrophoresis. Certain well-characterized miRNAs can be taken as endogenous controls. Furthermore, standardization can be achieved by spiking in control miRNAs. For example, for miRNA extracted from body fluids, miRNAs from *C. elegans* have been artificially introduced in known amounts, and the measured values for these controls guide normalization [6]. These methods facilitate the reproducibility of study findings by separating true biological signal from variations emerging from assay preparation [8, 10].

2.3 Profiling technologies

miRNA profiling technologies provide measurements of the relative abundances of miRNAs between different samples. Two main approaches have emerged: direct hybridization without sample amplification and variations of PCR-like amplification [5].

The most common direct hybridization method employs oligo-microarrays. Modifications have been designed to differentiate mature miRNAs from their primary transcript and hairpin precursor miRNAs [7]. These methods require larger initial total RNA samples; however, since amplification is avoided, the measurements obtained tend to be more stable and robust to variations in extract preparation [5]. Further, since many labs already have microarray equipment, this approach to miRNA profiling is more generally accessible.

As alternatives to direct hybridization, quantitative RNA sequencing PCR (qRT-PCR) and RNA sequencing (RNA-seq) have been employed [5, 11]. Since these methods involve first amplifying the miRNA content in samples, it is possible to use then with samples

with smaller RNA content, such as plasmas or needle-biopsies. In qRT-PCR, total RNA undergoes a modified reverse transcriptase (RT) reaction to produce cDNA where, instead of the standard poly-A reverse primers, a miRNA-specific reverse primer is utilized. In the PCR step, mature miRNA-specific forwards primers are introduced to selectively amplify miRNA sequences. As a consequence, qRT-PCR can only be used to quantify abundances of known miRNAs; it cannot identify novel miRNAs.

Conversely RNA-seq methods *can* identify novel miRNAs; further, they finely resolve between similar miRNA sequences [11]. The specific approaches differ between platforms, but they all involve both hybridization and selective amplification. For example, in miRAGE (miRNA serial analysis of gene expression), magnetic linkers added to enriched small RNAs, whose sequences are then reverse transcribed into cDNA. This cDNA is then amplified by PCR; however, the PCR product is filtered for the magnetic linkers. Only small RNA remains, which is then sequenced. This sequencing step allows for identification of novel miRNAs.

3 Statistical inference

Section 2 describes the overall workflow and options possible for performing a miRNA profiling study. This overall pipeline has been the subject of a number of recent reviews, including [11, 5] cited above. Once data is appropriately collected, focus shifts to performing valid and powerful statistical inference. In this section, we review the high-dimensional statistical learning methods described in [14] and reproduce their analysis using the data available in their additional files. Since the measurements in [14] was collected from the RNA-seq profiling approach of 2.3, our discussion will be tailored to data collected through such amplification and sequencing. Their experiment was designed to identify miRNAs (known and novel) that are differentially abundant in normal and cervical cancer tumor tissue.

3.1 Preliminary observations

Before applying specific inferential methods, it is worth previewing the miRNA sequencing data. We read in the data below.

```
X <- read.csv("miExpress2.csv", header = T)
rownames(X) <- X[, 1]
X <- X[, -1]
nGenes <- nrow(X)
nSamples <- ncol(X)
X <- X^(1/3)  # Cube-root transform

sample.type <- substr(colnames(X), start = 6, stop = 6)
sample.type[59:60] <- c("N", "T")  # two come from second run on same individual
sample.type <- factor(sample.type)

nGenes

[1] 714

nSamples

[1] 60


X[1:5, 1:4]  # Preview first five miRNAs, first four samples

        G547.N G547.T G576.N G576.T
let-7a   9.528 14.952  9.322 17.088
let-7a*  1.442  2.289  2.289  4.626
let-7b   9.916 16.156 14.078 34.654
let-7b*  2.466  2.000  2.621  5.290
let-7c   9.390  8.036 10.775 16.708


summary(sample.type)  # Number of normal and tumor samples.

 N  T
30 30
```

Observe that we work with expression of 714 miRNAs in paired normal and cancerous tissues over 60 individuals, with 2 sequencing runs (those for G699N and G761T). Since

the data is right-skewed (a few measurements are very large, and most are close to zero), a cube-root transformation is applied.

Figure 3.1 displays overall distribution of counts between samples and miRNAs. The plot "Counts per Sample" shows the total number of (cube-rooted) counts in different samples, color coded according to whether the samples are normal or cancer tissue. The plot "Counts per miRNA" shows the cube rooted counts per miRNA over all samples. Even after cube rooting, the distribution is right-skewed. The plot "Gene by Sample Counts" displays the same information without aggregating over either samples or miRNAs – it shows the distribution of individual elements in the data matrix.

3.2 Unsupervised methods

The first class of methods relevant to our inference are called unsupervised methods. In these methods, the relation between samples is characterized without using the labels associated with them (in this case, the label is whether or not the sample comes from normal or cancerous tissue). The nature of the characterization depends on the specific method, but often involves finding appropriate lower-dimensional representations of or significant separations within samples.

If these algorithms, despite being trained independently of class labels, nonetheless identify separations corresponding to class labels, then we can deduce that the most important variations in the data are associated with these labels. We find this to be the case with our miRNA data, indicating that collected miRNA expression levels exhibit strong associations with the presence or cervical cancer. It is in the domain of the supervised methods of section 3.5 to identify which of the 714 are responsible for these associations.

```
library(ggplot2)
library(reshape2)
mX <- melt(X)
mX$sample.type <- rep(sample.type, each = nGenes)
colnames(mX) <- c("miRNA", "sample", "count", "sample.type")
p1 <- qplot(x = colSums(X), geom = "histogram", fill = sample.type) + ggtitle("Counts pe
    scale_x_continuous("Cube root of Sample Counts") + scale_y_continuous("Frequency")
p2 <- ggplot(mX) + geom_histogram(aes(x = count, fill = sample.type)) + ggtitle("Gene by
    scale_y_continuous("Frequency")
p3 <- qplot(x = rowSums(X), geom = "histogram") + ggtitle("Counts per miRNA") +
    scale_x_continuous("Cube root of Sample Counts") + scale_y_continuous("Frequency")
multiplot(p1, p2, p3, cols = 2)
```

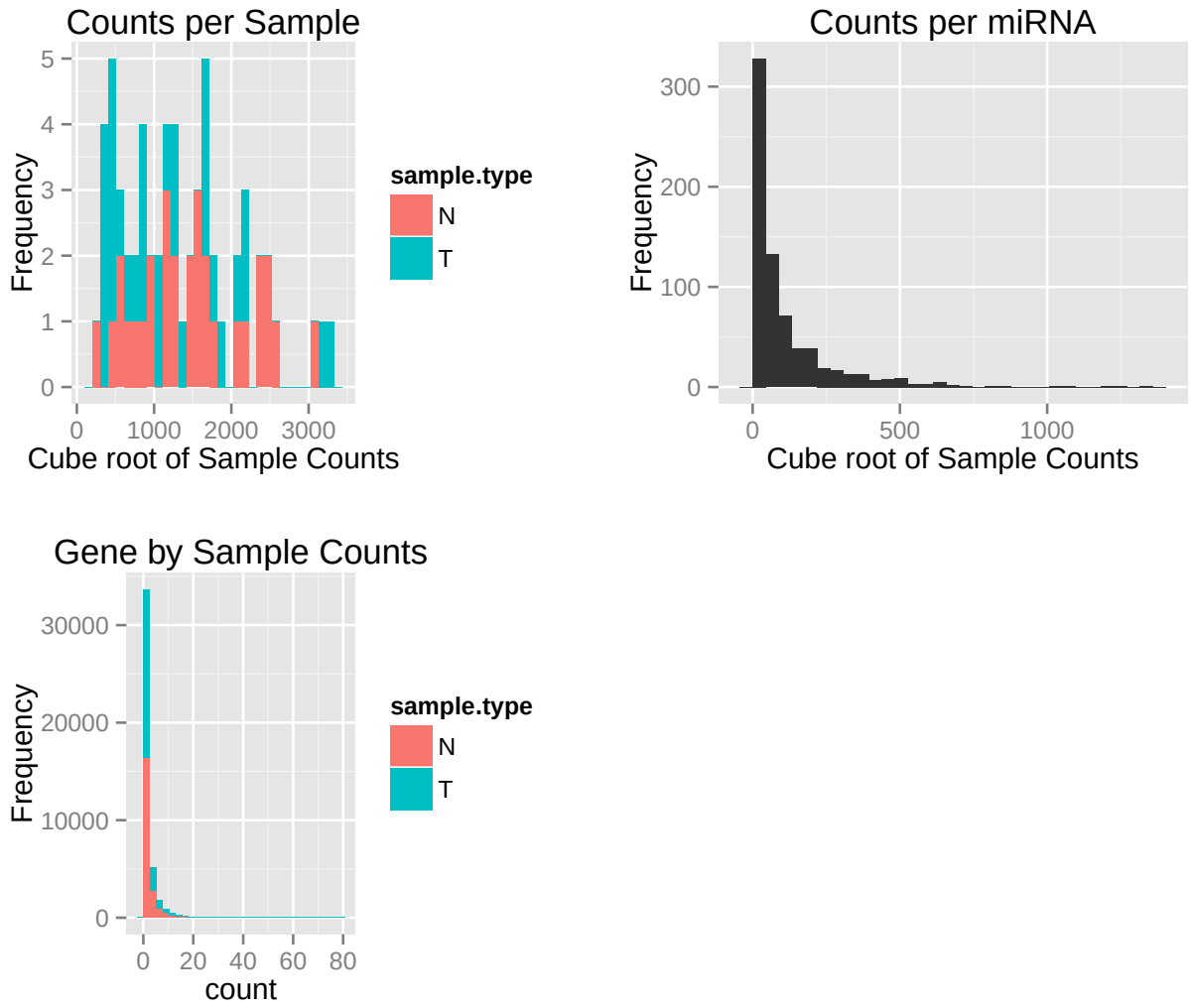

Figure 3.1: Histograms of sample aggregated, miRNA aggregated, and unaggregated counts in the data provided by [14].

3.3 Principal components analysis

A first thought is to apply principal components analysis (PCA). Conceptually, if we think about each sample as a point in 714-dimensional space (each coordinate corresponding to the expression level of a single miRNA), then PCA finds a lower d-dimensional plane (we take $d = 2$ for visualization purposes) that best approximates all 60 sample points in our data. Mathematically, it can be demonstrated that this plane also maximizes the variation of the projection of points onto a d-dimensional plane, so we can interpret the principal components as the linear combinations of covariates that explain the most variation in the data.

```
# Code for PCA
pc.X <- function(X) {
    X.tilde <- scale(X, center = T, scale = F)
    svdX <- svd(X.tilde)
    pca.X <- svdX$u %*% diag(svdX$d)
    pca.approx <- pca.X[, 1:2]
    return(pca.approx)
}

pca.labeled <- data.frame(pc.X(t(X)), pc.X(1 - cor(X)), label = colnames(X))
colnames(pca.labeled) <- c("PC1", "PC2", "PC1.cor", "PC2.cor", "label")
```

The most direct approach to PCA for our miRNA data is to work with the cube-rooted data visualized before. This is displayed on the left of figure 3.2. Alternatively, we can construct a distance based on the correlations between samples. We then find the best 2-dimensional approximation for this data. The result is displayed on the right in 3.2.

Notice that both approaches effectively separate samples obtained from tumorous and normal tissues, despite being constructed without any label of the state of samples. This is corroborates the claim above that the miRNA data supplied contains strong associations between expression level and tumor-state.

```
pca.plot <- ggplot(pca.labeled) + geom_text(aes(x = PC1, y = PC2, label = label,
    col = sample.type)) + ggtitle("PCA from rooted data")
pca.plot.cor <- ggplot(pca.labeled) + geom_text(aes(x = PC1.cor, y = PC2.cor,
    label = label, col = sample.type)) + ggtitle("PCA from correlation distance")

multiplot(pca.plot, pca.plot.cor)
```

Figure 3.2: 2-dimensional approximation obtained by PCA. Notice that the sample labels are distinguished even though the algorithm does not use them in its development. The correlation approach yields less skewed groups.

3.4 Hierarchical clustering

The second unsupervised method we consider is hierarchical clustering. Given pairwise distances between samples, hierarchical clustering recursively pairs closest samples, resulting in a tree allowing visualization of the distances between all samples. In a sense, it serves as a function to map local pairwise distance information into a global representation of the relationship between all samples.

More precisely, given all pairwise distances between samples, we first merge the two points closest to each other, calling this a cluster. At the next iteration, we recompute pairwise distances, where the distance of points to this new cluster is defined as the distance to the farthest point in this cluster. This procedure is repeated, where the distance between two clusters is defined to be the maximum of the pairwise distances between points contained within either cluster. The sequences of merges is displayed in a tree. The heights of the merges in the tree are the distances between the clusters when they are merged. Notice that no where in this algorithm is a labeling of points used.

For our application, we define the distance between two samples with estimated miRNA abundances correlation $\hat{\rho}$ to be $1-|\hat{\rho}|$. This is intuitively reasonable: points with perfect correlation should have zero distance, while if they are uncorrelated, they should have maximum distance to each other. The result of the clustering is displayed in figure 3.3. Amazingly, with the exception of samples G428_T and G701_T, the clustering perfectly separates the normal samples from the tumor samples. Again, this corroborates the idea that miRNA abundance effectively distinguishes between cancer and normal samples.

3.5 Supervised methods

Next, we consider the application of supervised learning methods. The distinguishing feature of these algorithms is that they need to be trained with sample labels. These methods can

```
source("http://addictedtor.free.fr/packages/A2R/lastVersion/R/code.R")  # Plotting Packa
miRNA.clust <- hclust(as.dist(1 - cor(X)))
A2Rplot(miRNA.clust, k = 3, col.down = c("#FF6B6B", "#4ECDC4", "#556270"), main = "Hierc
```

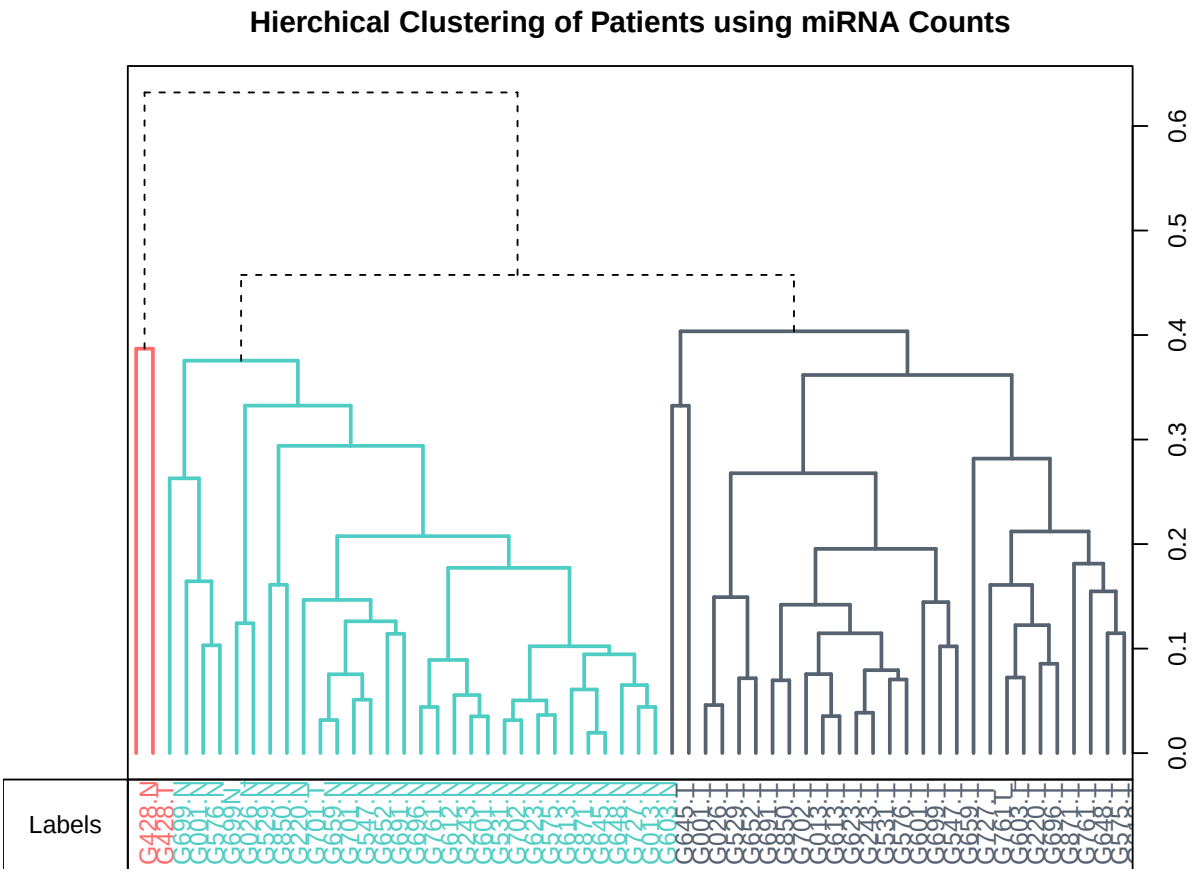

Figure 3.3: Hierarchical clustering of samples using correlation obtained using miRNA expression levels.

be trained to perform classification and identify differential expression of miRNAs between tumor and normal and tissue samples. Hence, while the methods of section 3.2 were able to identify strong associations between global miRNA expression and cancer state, the methods below will be able to resolve the particular miRNAs that contribute to these differences.

3.5.1 NSC classification

Again, we consider each sample to be a point in 714-dimensional space where each coordinate is the expression level of a particular miRNA. The goal is to train a classifier such that, given a vector of expression levels of these miRNAs for a new tissue, we can decide whether the sample came from a cervical cancer or normal tissue. Further, we would like the classifier to output which are the differentially expressed miRNAs that contribute to the classification. An approach designed specifically for this kind of high-dimensional setting is nearest shrunken centroids (NSC) [12, 4].

Heuristically, the NSC algorithm is a version of linear discriminant analysis regularized to be appropriate for the high-dimension of samples. Specifically, given a new sample, we can compute the distance to the known average expression levels for the tumorous and normal tissues in our training data (the distance is taken with respect to the estimated variances of individual miRNA expression levels).

The naive approach would be to classify the new sample to the closest of these two averages. However, this presumes that every miRNA contributes to the differences between normal and tumorous samples. We don't expect this to be the case, so we soft-threshold the averages until only the miRNAs contributing the most to the classification are included. We then classify the new sample to the closest of these thresholded averages.

Of course, we don't know in advance the appropriate level to threshold these averages. Hence, we apply cross-validation. The idea is to fix a thresholding level, train the classifier on a subset of samples, and calculate the error rate of the classification of the remaining

samples. A range of thresholding levels is considered, and that yielding the minimum error rate is selected. This classifier can now be used on entirely new data.

We apply the R package pamr to perform NCS classification on our miRNA data. We display the cross-validation results in figure 3.4, the expression levels of miRNAs used in the classifier in figure 3.5, and the corresponding centroids in 3.6. The cross-validation results are typical: the under and overthresholded models have higher misclassification rates. We choose the threshold 3.3021, because it is the largest (among those tested) that still yields a minimum cross-validation error; intuitively, this is the simplest model achieving minimum estimated average test error, which we find to be about 0.1667 (compared to 0.5 for random classifications).

The classifier uses 15 of the 714 total miRNAs. From figures 3.5 and 3.6, we find that all of these miRNAs, with the exception of miR-205 are less expressed in tumor samples than in the normal tissues. We can now look up these miRNAs on miRBase. For example, miR-125b has been found to be associated with cell differentiation, and miR-143 has been implicated in both cardiac morphogenesis and certain types of cancer.

```
library("pamr")
nsc.train.data <- list(x = X, y = sample.type, genenames = rownames(X))
training.results <- pamr.train(nsc.train.data)
set.seed("672013")  # Cross-validation includes randomness
nsc.cv.fit <- pamr.cv(training.results, nsc.train.data)
cv.thresh.ix <- max(which(nsc.cv.fit$error == min(nsc.cv.fit$error)))
cv.thresh <- nsc.cv.fit$threshold[cv.thresh.ix]
nMisclassified <- nsc.cv.fit$error[cv.thresh.ix] * 60
nGenesUsed <- nsc.cv.fit$size[cv.thresh.ix]
```

3.5.2 Testing differential expression

Finally, we can formally test for the differential expression of miRNAs between samples and estimate the associated false discovery rate (FDR). We describe a simple approach; a more sophisticated technique can be read in [14].

```
pamr.plotcv(nsc.cv.fit)
```

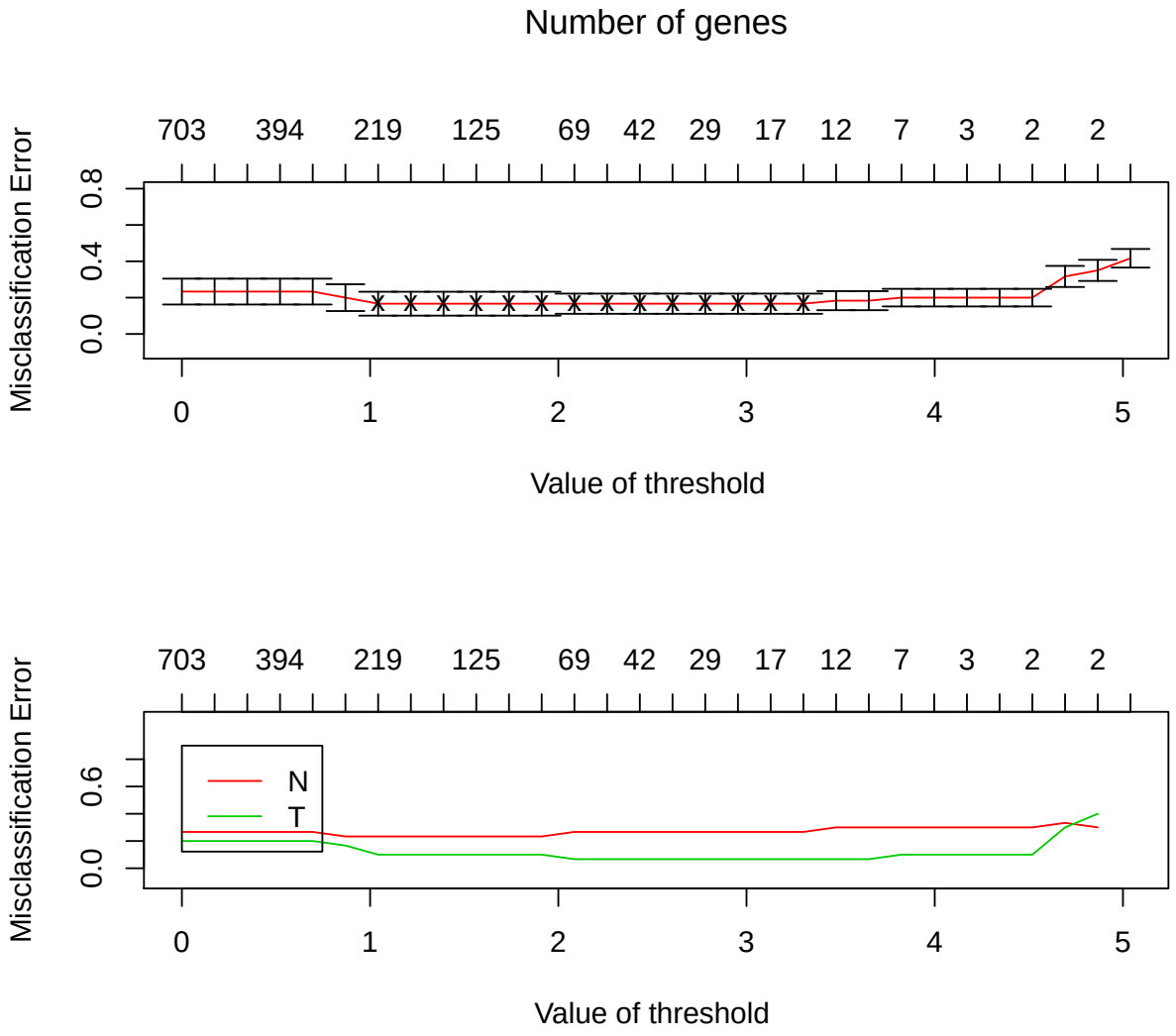

Figure 3.4: Misclassification rates estimated for different values of the thresholding parameter, using cross-validation.

```
pamr.geneplot(training.results, nsc.train.data, threshold = cv.thresh)
```

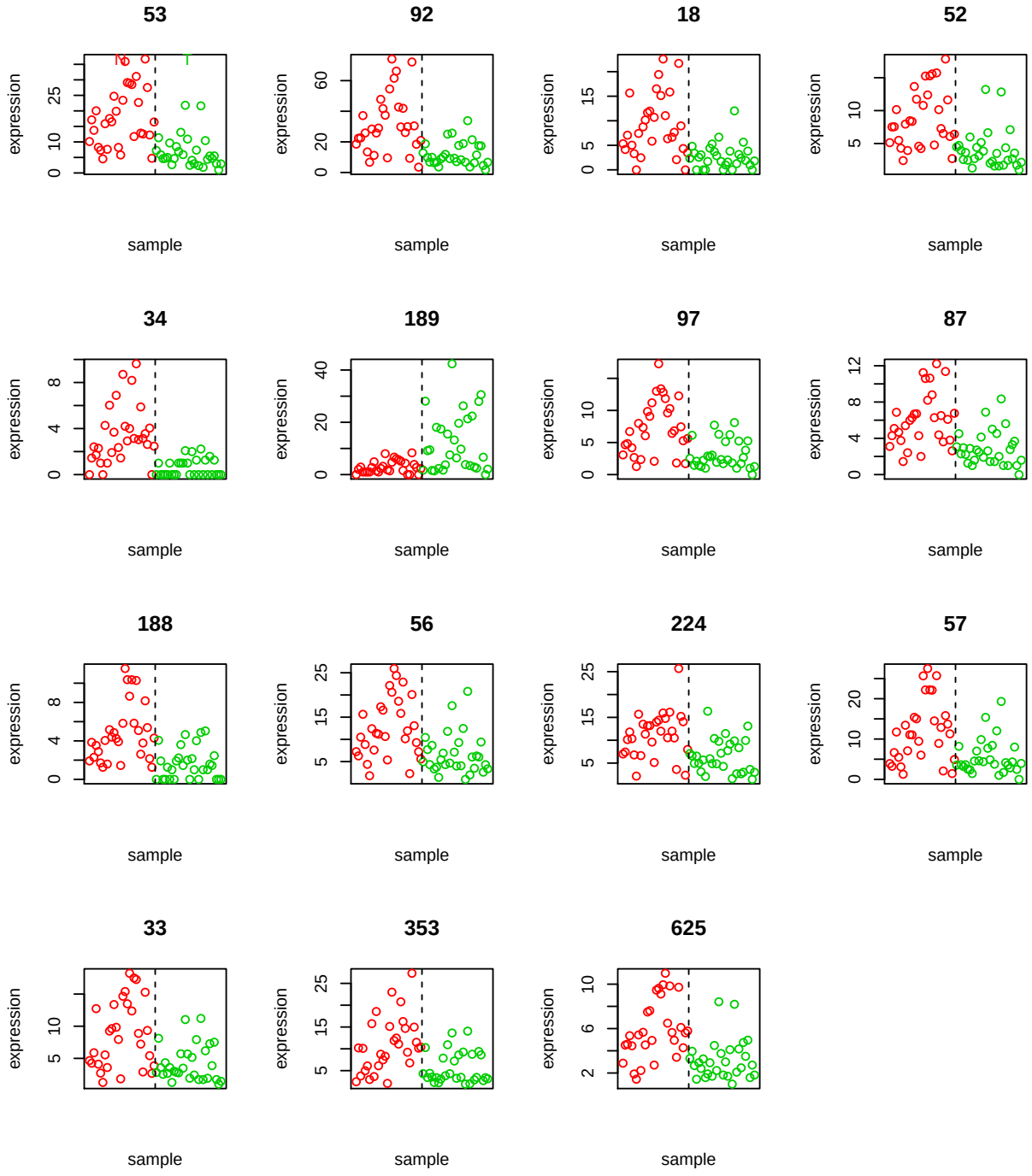

Figure 3.5: microRNA expression levels in tumor and normal samples. Reading from left to right, top to bottom, the indices in the titles correspond to the miRNA names, read top to bottom, in figure 3.6. Green gives expression levels in tumorous samples while red gives expression levels in normal tissues.

```
pamr.plotcen(training.results, nsc.train.data, threshold = cv.thresh)
```

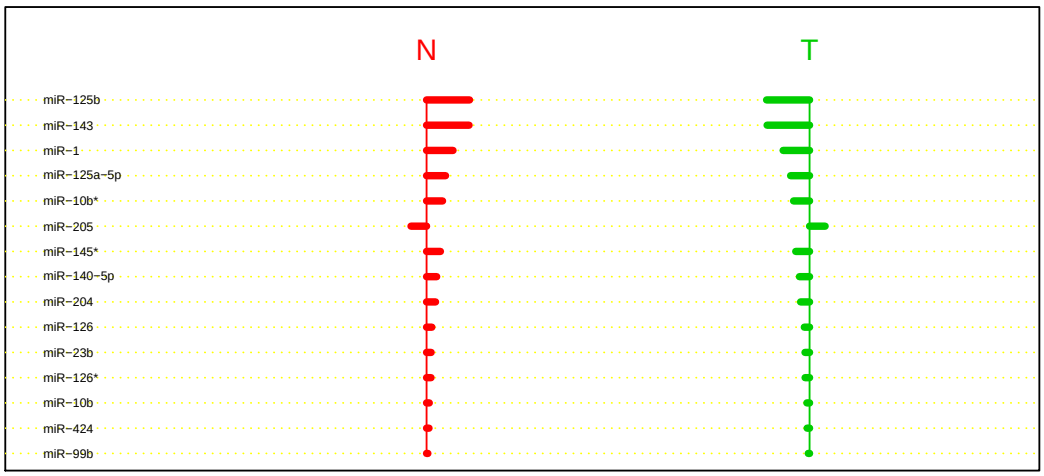

Figure 3.6: These are the centroids employed in classification after a threshold has been chosen. The larger the centroid coordinate for a particular RNA, the stronger the association.

The basic idea is to perform a t-test for differential expression of each miRNA. Due to the multiple testing, we cannot directly use the analytic p-values for the t-test to determine which results are significant at any prespecified level. Instead, we nonparametrically estimate the FDR via permutations.

Recall that the FDR is defined as the expected proportion of false positives among all hypotheses rejected. To estimate this quantity, we observe that under the null hypothesis of equal miRNA expression between tumor and normal tissue classes, we can permute class labels without changing the distribution of test statistics. Hence, to estimate the FDR, assuming a fixed number of rejected hypotheses, we can permute the class labels to calculate the null distribution of test statistics. We can compare the test statistics calculated on our real data with this null distribution; the FDR associated with k rejections is then the proportion of test statistics in the null distribution larger than the statistic $t_{(k)}$ (the k^{th} largest statistic in our tests on true data).

The estimated FDRs are displayed in figure 3.7. 91 miRNAs are deemed significant at

an FDR level of 0.05; they are stored in the object `significant.genes`. The first few are displayed in the code chunk below.

```
t.test.miRNA <- function(X, sample.type) {
    # Normalize each individual's mean miRNA counts
    X.tilde <- apply(X, MARGIN = 2, FUN = function(v) {
        v/(mean(v))
    })
    # Two sample t-test
    t.statistics <- apply(X.tilde, MARGIN = 1, FUN = function(gene) t.test(gene ~
        sample.type)$statistic)
    sort.genes <- sort(t.statistics, decreasing = T)

    return(sort.genes)
}

perm.statistics <- function(X, sample.type, n.perm = 100) {
    nGenes <- nrow(X)
    t.statistics.perm <- matrix(nrow = n.perm, ncol = nGenes)
    for (i in 1:n.perm) {
        # Generate random permutations
        sample.type.perm <- sample(sample.type)
        t.statistics.perm[i, ] <- t.test.miRNA(X, sample.type.perm)
    }
    return(t.statistics.perm)
}

fdr.rates <- function(true.statistics, perm.stats) {
    nGenes <- length(true.statistics)
    fdr.rate <- vector(length = nGenes)
    for (i in 1:nGenes) {
        thresh <- true.statistics[i]
        fdr.rate[i] <- length(which(permuted.stats >= thresh))/(nrow(perm.stats) *
            i)
    }
    return(fdr.rate)
}

genes.t.test <- t.test.miRNA(X, sample.type)
permuted.stats <- perm.statistics(X, sample.type)
t.test.fdr <- fdr.rates(genes.t.test, permuted.stats)
```

```
qplot(x = 1:200, t.test.fdr[1:200]) + scale_x_continuous("Number of Significant miRNAs D
    scale_y_continuous("FDR") + ggtitle("Estimated FDR for t-tests") + geom_line(y = 0.0
    col = "red")
```

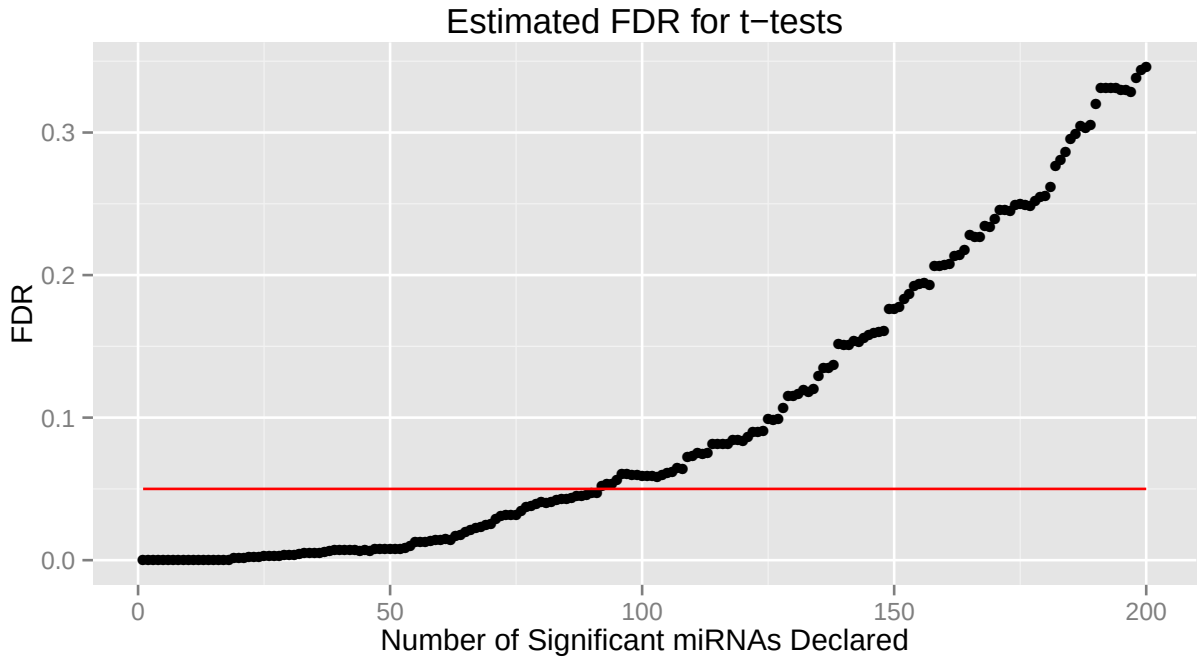

Figure 3.7: FDR.

```
n.rejected.05level <- max(which(t.test.fdr <= 0.05))
significant.genes <- genes.t.test[1:n.rejected.05level]
head(significant.genes)

##     miR-143    miR-125b    miR-145*    miR-10b* miR-125a-5p     miR-204
##       8.424       8.337       7.598       7.132       7.100       7.076
```

4 Conclusion

In this report, we explored current techniques in experimental preparation and statistical analysis of miRNA profiling data. We found that there is no universal approach to collecting miRNA profiling data, but that the necessary choices can be guided by sample and research

considerations. Further, in our reproduction of statistical analysis of RNA-seq pipeline profiling data, we found strong associations between specific miRNA expression levels and the tumor-status of tissues; indeed, associations were detected even via unsupervised methods.

More broadly, this project has given a chance to work in the intersection between modern genetics and statistics. We have seen how an array of creative experimental and statistical techniques have been designed to both deepen our understanding of miRNAs and the regulation of gene expression as well as provide potentially powerful biomarkers for diseases. As personalized genomics and medicine continues to expand and the associated data and research questions grow in complexity, we anticipate challenging problems whose solutions can be informed by the themes that we have encountered in this report and which present rich research and healthcare opportunities.

References

[1] Ines Alvarez-Garcia and Eric A Miska. Microrna functions in animal development and human disease. *Development*, 132(21):4653–4662, 2005.

[2] George Adrian Calin and Carlo Maria Croce. Microrna-cancer connection: the beginning of a new tale. *Cancer research*, 66(15):7390–7394, 2006.

[3] Helge Großhans, Ted Johnson, Kristy L Reinert, Mark Gerstein, and Frank J Slack. The temporal patterning microrna< i> let-7</i> regulates several transcription factors at the larval to adult transition in< i> c. elegans</i>. *Developmental cell*, 8(3):321–330, 2005.

[4] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

[5] William Kong, Jian-Jun Zhao, Lili He, and Jin Q Cheng. Strategies for profiling microrna expression. *Journal of cellular physiology*, 218(1):22–25, 2009.

[6] Evan M Kroh, Rachael K Parkin, Patrick S Mitchell, and Muneesh Tewari. Analysis of circulating microrna biomarkers in plasma and serum using quantitative reverse transcription-pcr (qrt-pcr). *Methods*, 50(4):298–301, 2010.

[7] Chang-Gong Liu, Riccardo Spizzo, George Adrian Calin, and Carlo Maria Croce. Expression profiling of microrna using oligo dna arrays. *Methods (San Diego, Calif.)*, 44(1):22, 2008.

[8] Pieter Mestdagh, Pieter Van Vlierberghe, An De Weer, Daniel Muth, Frank Westermann, Frank Speleman, Jo Vandesompele, et al. A novel and universal method for microrna rt-qpcr data normalization. *Genome Biol*, 10(6):R64, 2009.

[9] Yusuke Ohnishi, Yasushi Totoki, Atsushi Toyoda, Toshiaki Watanabe, Yasuhiro Yamamoto, Katsushi Tokunaga, Yoshiyuki Sakaki, Hiroyuki Sasaki, and Hirohiko Hohjoh. Small rna class transition from sirna/pirna to mirna during pre-implantation mouse development. *Nucleic acids research*, 38(15):5141–5151, 2010.

[10] Heidi J Peltier and Gary J Latham. Normalization of microrna expression levels in quantitative rt-pcr assays: identification of suitable reference rna targets in normal and cancerous human solid tissues. *Rna*, 14(5):844–852, 2008.

[11] Colin C Pritchard, Heather H Cheng, and Muneesh Tewari. Microrna profiling: approaches and considerations. *Nature Reviews Genetics*, 13(5):358–369, 2012.

[12] Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Class prediction by nearest shrunken centroids, with applications to dna microarrays. *Statistical Science*, pages 104–117, 2003.

[13] Stefano Volinia, George A Calin, Chang-Gong Liu, Stefan Ambs, Amelia Cimmino, Fabio Petrocca, Rosa Visone, Marilena Iorio, Claudia Roldo, Manuela Ferracin, et al. A microrna expression signature of human solid tumors defines cancer gene targets. *Proceedings of the National Academy of Sciences of the United States of America*, 103(7):2257–2261, 2006.

[14] Daniela Witten, Robert Tibshirani, Sam G Gu, Andrew Fire, and Weng-Onn Lui. Ultra-high throughput sequencing-based small rna discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls. *BMC biology*, 8(1):58, 2010.

[15] Beiyan Zhou, Stephanie Wang, Christine Mayr, David P Bartel, and Harvey F Lodish. mir-150, a microrna expressed in mature b and t cells, blocks early b cell development when expressed prematurely. *Proceedings of the National Academy of Sciences*, 104(17):7080–7085, 2007.