FFPE Package Example (Version 1.2.0)

Levi Waldron

October 1, 2012

1 Introduction

Gene expression data derived for formalin-fixed, paraffin-embedded (FFPE) tissues tend to be noisier and more susceptible to experimental artefacts than data derived from fresh-frozen tissues. Microarray studies of FFPE tissues may also be of a larger scale than of fresh-frozen tissues. Both of these factors contribute to the need for new quality control and visualization techniques. This is an example of using the `ffpe` Bioconductor package for quality control of gene expression data derived from formalin-fixed, paraffin-embedded (FFPE) tissues.

Example data (of better quality than is typical for clinical FFPE specimens) are taken from the early study of the Illumina WG-DASL microarray assay for FFPE specimens by April et al. [1], using only the dilution series from Burkitts Lymphoma and Breast Adenocarcinoma cell lines. The dilution series provide a range of sample qualities from very high at most dilution levels, to low at the lowest dilution levels.

2 Initial inspection of raw data

The boxplot of raw log2 expression intensities is a useful first look at data quality. Samples can be ordered by extraction sequence, batch number, or Interquartile Range (IQR). When dealing with hundreds of samples, however, a boxplot can become difficult to view. The *sortedIqrPlot* function provides a convenient means to view only the 25th to 75th percentile of expression intensities, which is extensible to more than a thousand samples, and to sort samples by a specified quality metric. By default, samples are sorted from smallest to largest IQR, but batch ID or any string can be provided for ordering of the samples. In the case of duplicate IDs, for example with batches, samples are further sorted by IQR within each batch. An example of this simplified, sorted boxplot is shown for the April et al. dilution series in Figure 1.

```
> library(ffpe)
> library(ffpeExampleData)
> data(lumibatch.GSE17565)
> sortedIqrPlot(lumibatch.GSE17565,dolog2=TRUE)
```

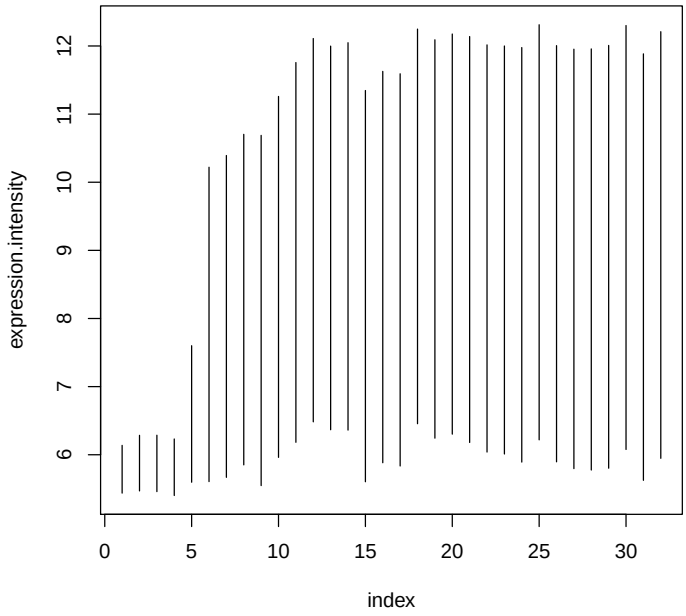


Figure 1: Simplified, sorted boxplot of the April et al. dilution series. Vertical lines indicate 25^{th} to 75^{th} percentile of raw log_2 intensities for each sample; ie, the box portion of a boxplot. Samples are sorted from smallest to largest Interquartile Range (IQR).

3 Sample Quality Control

Expression profiles with a low intrinsic measure of quality in addition to low similarity to other samples from the study tend to be less reliable and less reproducible. The *sortedIqrPlot* function is a flexible interface for identification low-quality samples with these attributes. The default intrinsic quality measure is IQR, and the default comparative measure is Spearman correlation to a median pseudochip (constructed from the median value of each probe). The default values are a reasonable choice, but other other measures can also be used for both intrinsic and comparative quality measures - see the help page for sampleQC for other options.

We can see that the samples rejected by this procedure (Figure 2) are those at the low concentration of end of the dilution series (Figure 3), and in fact, the same samples would be rejected if RNA concentration were chosen as the intrinsic quality control metric (Figure 4).

```
> QC <- sampleQC(lumibatch.GSE17565,xaxis="index",cor.to="pseudochip",QCmeasure="IQR")

[1] "Calculating Spearman correlation to median pseudochip using pairwise complete observati
[1] "Using samples: GSM437734, GSM437733, GSM437751, GSM437737, GSM437742, GSM437727, GSM437
```

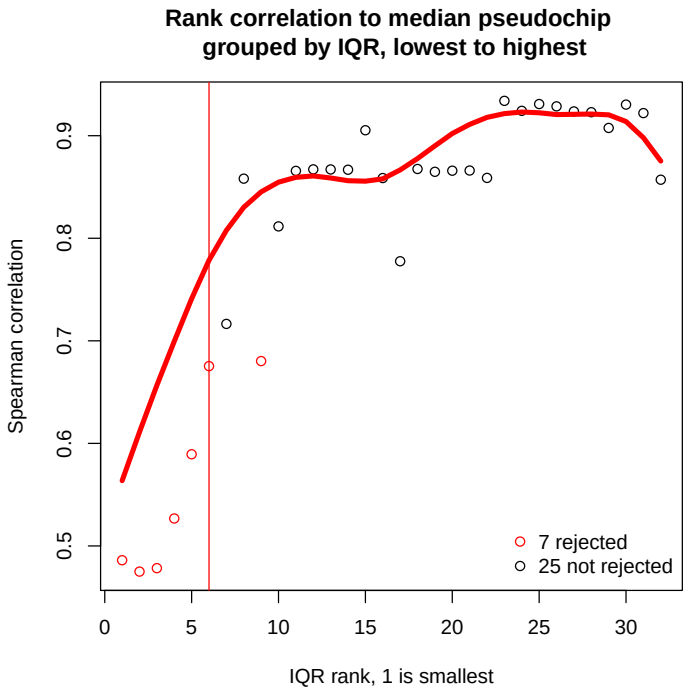


Figure 2: Sample Quality Control plot. In this example the plot is more readable if we use the rank of each sample on the x-axis (xaxis="index"). We use default IQR as the intrinsic quality measure, and the median pseudochip for the entire study as the comparative measure.

```
> QCvsRNA <- data.frame(inputRNA.ng=lumibatch.GSE17565$inputRNA.ng,
+                       rejectQC=QC$rejectQC)
> QCvsRNA <- QCvsRNA[order(QCvsRNA$rejectQC,-QCvsRNA$inputRNA.ng),]
> par(mgp=c(4,2,0))
> dotchart(log10(QCvsRNA$inputRNA.ng),
+          QCvsRNA$rejectQC,
+          xlab="log10(RNA conc. in ng)",
+          ylab="rejected?",
+          col=ifelse(QCvsRNA$rejectQC,"red","black"))
```

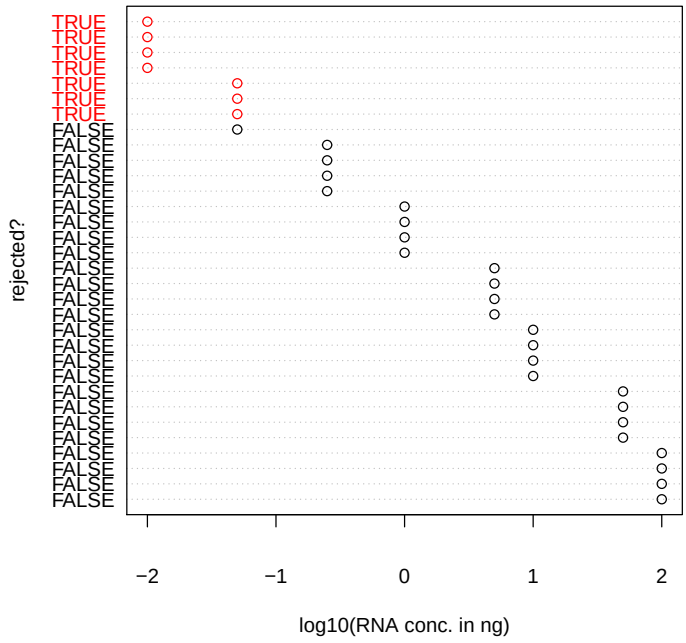


Figure 3: RNA concentration of samples whose expression profiles were rejected and not rejected by the above QC test.

```
> QC <- sampleQC(lumibatch.GSE17565,xaxis="index",cor.to="pseudochip",QCmeasure=log10(lumiba

[1] "Calculating Spearman correlation to median pseudochip using pairwise complete observati
[1] "Using samples: GSM437750, GSM437751, GSM437742, GSM437743, GSM437734, GSM437735, GSM437
```

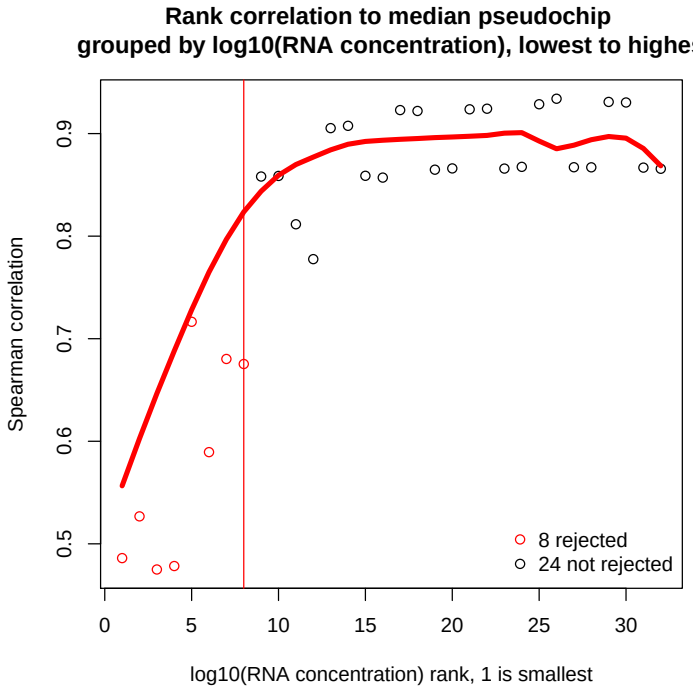


Figure 4: In this example, RNA concentration could have been used as an alternative intrinsic QC metrix.

4 Feature quality control

Features with high variance are likely to contain a higher proportion of signal to noise than features with low variance. This is the case with gene expression data from fresh-frozen tissues as well, but the fixation, storage, and gene expression assaying for FFPE tissues add more steps which may cause detection of a transcript to fail. Since technical replicates are available in this dataset, we can look at reproducibility of probe measurements between replicate measurements as a function of variance. First, we will use only samples which passed QC in the previous step:

```
> lumibatch.QC <- lumibatch.GSE17565[,!QC$rejectQC]
```

Now do normalization for each set of replicates independently:

```
> ##replicate 1
> lumibatch.rep1 <- lumibatch.QC[,lumibatch.QC$replicate==1]
> lumbiatch.rep1 <- lumiT(lumibatch.rep1,"log2")

Perform forcePositive background correction ...
Perform log2 transformation ...

> lumbiatch.rep1 <- lumiN(lumibatch.rep1,"quantile")

Perform quantile normalization ...

> ##replicate 2
> lumibatch.rep2 <- lumibatch.QC[,lumibatch.QC$replicate==2]
> lumibatch.rep2 <- lumiT(lumibatch.rep2,"log2")

Perform forcePositive background correction ...
Perform log2 transformation ...

> lumibatch.rep2 <- lumiN(lumibatch.rep2,"quantile")

Perform quantile normalization ...
```

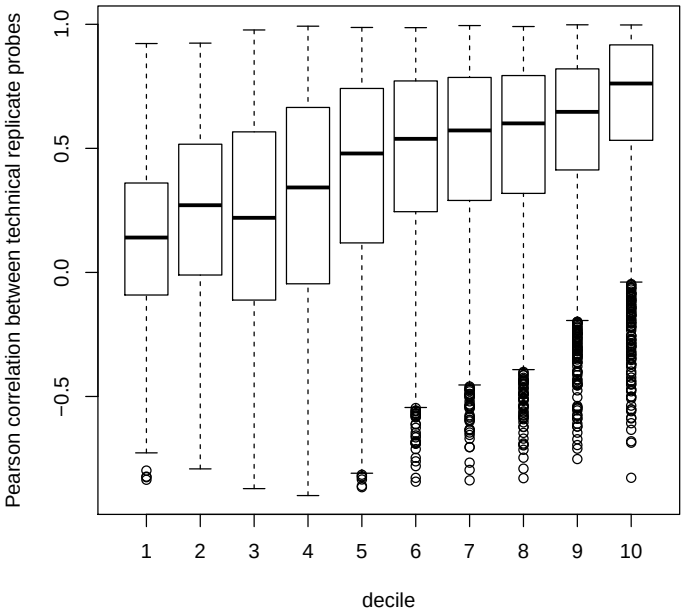
Keep samples which passed QC for both replicate sets:

```
> available.samples <- intersect(lumibatch.rep1$source,lumibatch.rep2$source)
> lumibatch.rep1 <- lumibatch.rep1[,na.omit(match(available.samples,lumibatch.rep1$source))]
> lumibatch.rep2 <- lumibatch.rep2[,na.omit(match(available.samples,lumibatch.rep2$source))]
> all.equal(lumibatch.rep1$source,lumibatch.rep2$source)

[1] TRUE
```

And finally, plot correlation of replicates as a function of probe variance in replicate 1. Note that reproducibility increases with probe variance; in the absence of technical replication.

```
> probe.var <- apply(exprs(lumibatch.rep1),1,var)
> rowCors = function(x, y) {  ##rowCors function borrowed from the arrayMagic Bioconductor p
+   sqr = function(x) x*x
+   if(!is.matrix(x)||!is.matrix(y)||any(dim(x)!=dim(y)))
+     stop("Please supply two matrices of equal size.")
+   x   = sweep(x, 1, rowMeans(x))
+   y   = sweep(y, 1, rowMeans(y))
+   cor = rowSums(x*y) /  sqrt(rowSums(sqr(x))*rowSums(sqr(y)))
+ }
> probe.cor <- rowCors(exprs(lumibatch.rep1),exprs(lumibatch.rep2))
> ##the plot will be easier to see if we bin variance into deciles:
> quants <- seq(from=0,to=1,by=0.1)
> probe.var.cut <- cut(probe.var,breaks=quantile(probe.var,quants),include.lowest=TRUE,label
> boxplot(probe.cor~probe.var.cut,
+         xlab="decile",
+         ylab="Pearson correlation between technical replicate probes")
```



A default filter which removes probes with less than the median variance is recommendable. Keeping only probes with variance greater than the median is simple:

```
> lumibatch.rep1 <- lumibatch.rep1[probe.var > median(probe.var),]
> lumibatch.rep2 <- lumibatch.rep2[probe.var > median(probe.var),]
```

5 Concordance at the Top

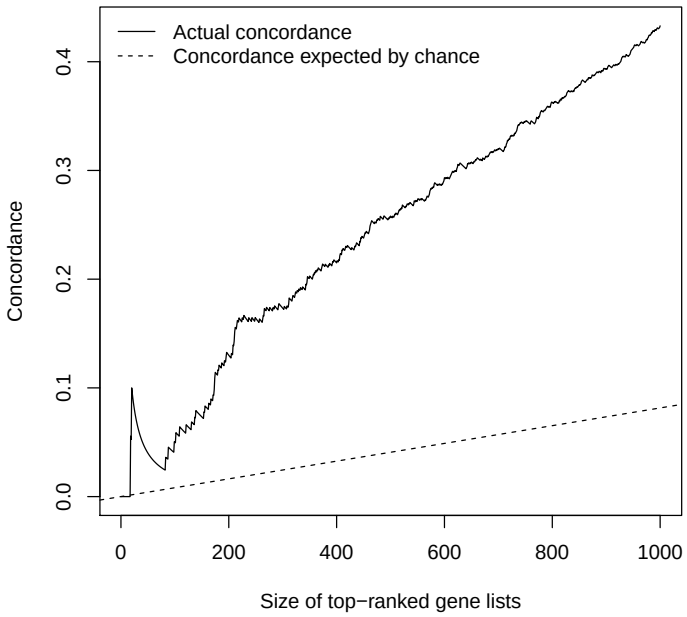
A common interim objective of gene expression studies is simply to identify differentially expressed genes with respect to a treatment or phenotype of interest, and to follow up on hypotheses generated from the top differentially expressed genes. Furthermore, Gene Set Enrichment Analysis depends on the ranking of a list of genes to identify gene sets enriched at the top (or bottom) of the list. The Concordance at the Top plot (CAT-plot)[2] measures the reproducibility of differentially expressed gene lists by the concordance of genes in the top n genes of each list (concordance = number of common genes divided by the number of genes in each list).

In this example we produce a CAT-plot for differentially expression with respect to cell type in the GSE17565 dataset, representing concordance between the replicate measurements. We calculate nominal p-values for differential expression between Burkitts Lymphoma samples and Breast Adenocarcinoma samples, using the fast rowttests function from the genefilter package:

```
> library(genefilter)
> ttests.rep1 <- rowttests(exprs(lumibatch.rep1),fac=factor(lumibatch.rep1$cell.type))
> ttests.rep2 <- rowttests(exprs(lumibatch.rep2),fac=factor(lumibatch.rep2$cell.type))
> pvals.rep1 <- ttests.rep1$p.value;names(pvals.rep1) <- rownames(ttests.rep1)
> pvals.rep2 <- ttests.rep2$p.value;names(pvals.rep2) <- rownames(ttests.rep2)
```

The CATplot can be made using the CATplot function:

```
> x <- CATplot(pvals.rep1,pvals.rep2,maxrank=1000,xlab="Size of top-ranked gene lists",ylab=
> legend("topleft",lty=1:2,legend=c("Actual concordance","Concordance expected by chance"),
```



Note: An extension to the CAT-plot, termed the CAT-boxplot, can be used in the absence of technical replicates (Waldron et al, under review). The samples are randomly split into two equal parts, each used to rank differentially expressed genes, and the splitting is repeated to generate a distribution of concordances. This function can facilitate generating these distributions by setting $make.plot = FALSE$.

References

[1] Craig April, Brandy Klotzle, Thomas Royce, Eliza Wickham-Garcia, Tanya Boyaniwsky, John Izzo, Donald Cox, Wendell Jones, Renee Rubio, Kristina Holton, Ursula Matulonis, John Quackenbush, and Jian-Bing Fan. Whole-genome gene expression profiling of formalin-fixed, paraffin-embedded tissue samples. *PloS One*, 4(12):e8162, 2009. PMID: 19997620.

[2] Rafael A Irizarry, Daniel Warren, Forrest Spencer, Irene F Kim, Shyam Biswal, Bryan C Frank, Edward Gabrielson, Joe G N Garcia, Joel Geoghegan, Gregory Germino, Constance Griffin, Sara C Hilmer, Eric Hoffman, Anne E Jedlicka, Ernest Kawasaki, Francisco Martinez-Murillo, Laura Morsberger, Hannah Lee, David Petersen, John Quackenbush, Alan Scott, Michael Wilson, Yanqin Yang, Shui Qing Ye, and Wayne Yu. Multiple-

laboratory comparison of microarray platforms. *Nat Meth*, 2(5):345–350, May 2005.