Tools for Spike-in Data Analysis and Visualization (spkTools)

Matthew N. McCall

May 13, 2008

```
> library(spkTools)
```

Load Affymetrix HGU133A spike-in dataset:

```
> library(affy)
> library(SpikeIn)
> data(SpikeIn133)
```

Use RMA for preprocessing:

```
> e <- rma(SpikeIn133)
```

Set up a matrix of expression values and a matrix of nominal concentrations:

```
> expr.affy <- exprs(e)
> spks <- t(pData(e))
> i <- sort(rownames(spks), index.return = T)$ix
> spks <- spks[i, ]
> nomin.affy <- matrix(NA, nrow = nrow(expr.affy), ncol = ncol(expr.affy))
> ind <- which(rownames(expr.affy) %in% rownames(spks))
> nomin.affy[ind, ] <- spks
> rownames(nomin.affy) <- rownames(expr.affy)
> colnames(nomin.affy) <- colnames(expr.affy)
> nomin.affy[nomin.affy == 0] <- NA
> nomin.affy <- log2(nomin.affy)
```

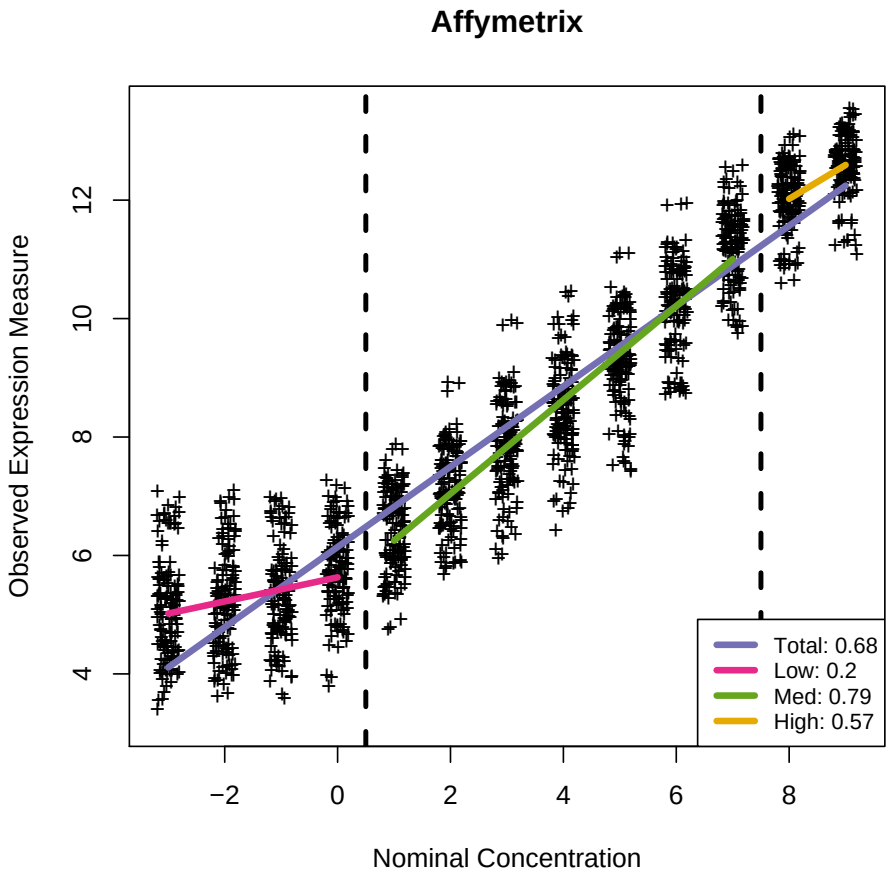
Create a new SpikeInExpressionSet object from the two matrices:

```
> object <- new("SpikeInExpressionSet", exprs = expr.affy, spikeIn = nomin.affy)
```

Set a few parameters:

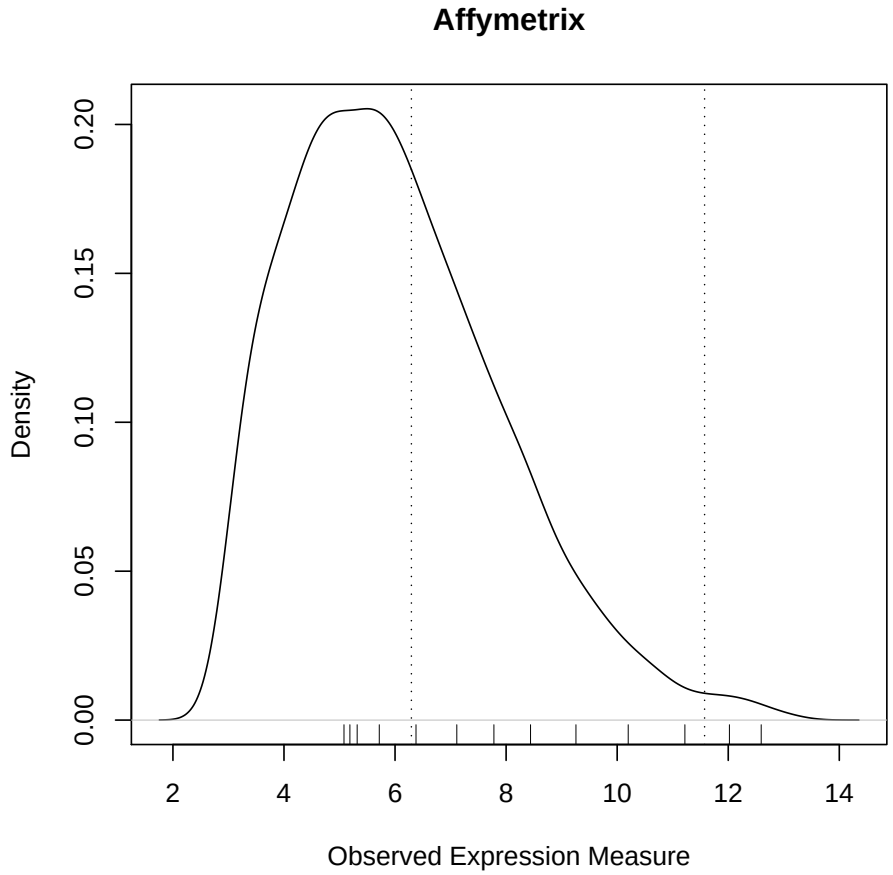
```
> fc = 2
> label = "Affymetrix"
> par(mar = c(3, 2.5, 2, 0.5), cex = 1.8)

> spkSlopeOut <- spkSlope(object, label, pch = "+")
```



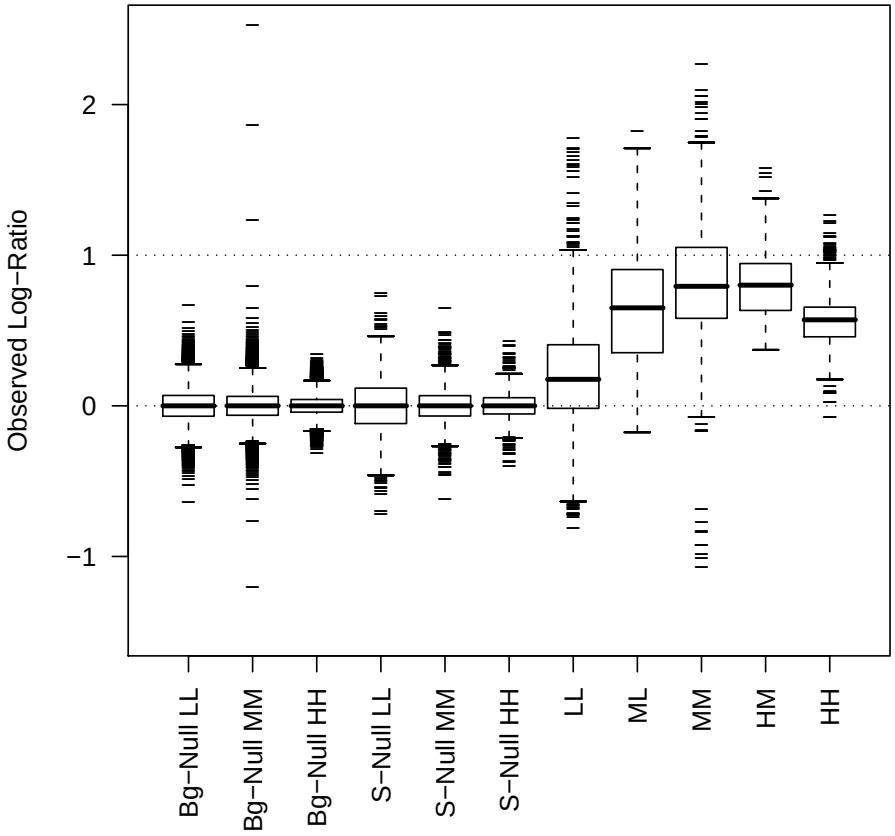
Observed versus nominal values: This plot depicts expression values plotted against the log (base 2) of the reported nominal concentration. The regression slope obtained utilizing all the data and the regression slopes obtained within each ALE value strata are shown. The slope of each line is reported in the legend. The vertical lines divide the ALE strata.

```
> spkDensity(object, spkSlopeOut, cuts = TRUE, label)
```



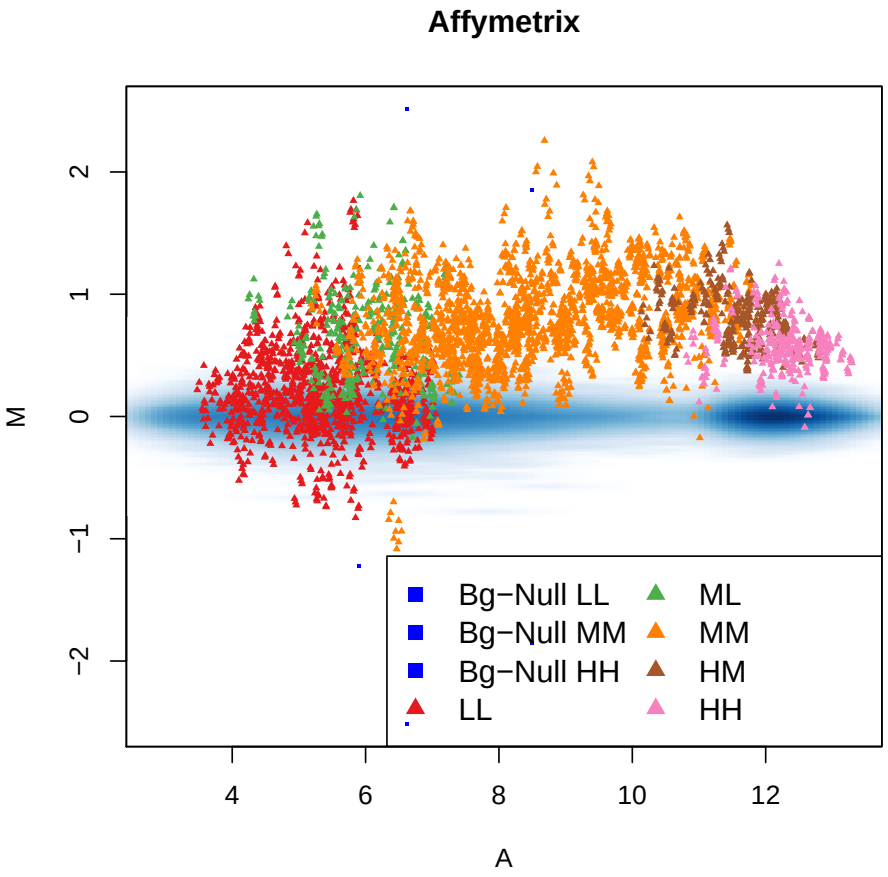
Empirical densities: This plot depicts the empirical density of the average (across arrays) expression values for the background RNA. The tick marks on the x-axis show the average expression at each nominal concentration. The dotted lines represent the cut points for low, medium, and high ALE values.

```
> spkBoxOut <- spkBox(object, spkSlopeOut, fc)
> plotSpkBox(spkBoxOut, fc, ylim = c(-1.5, 2.5))
> sbox <- summarySpkBox(spkBoxOut)
```



Log-ratio distributions: This plot depicts the distribution of observed log ratios for a given nominal fold change. The log ratios are stratified by the ALE strata into which the two nominal concentrations fall. The null distributions' log-ratios are divided into background RNA (Bg-Null) and spike-ins at the same nominal concentration (S-Null), for each bin. The dotted horizontal lines represent the expected or nominal log-ratios: zero for the null distribution and one for the other comparisons.

```
> spkMA(object, spkSlopeOut, fc, label = label, ylim = c(-2.5,
+     2.5))
```



MA plots: For each platform, we performed all pair-wise comparisons of the arrays. From each comparison we computed the log-ratio (M) and average expression value (A) for each gene. These plots show M plotted against A. To avoid drawing hundreds of points on top of each other we use a smooth scatter plot which shows the distribution of these points: dark and light shades of blue show high and low concentrations of points respectively. Points not associated with the spike-in transcripts (expected M=0) that achieved fold changes above 2 are shown as large blue dots. The points associated with spike-in transcripts with nominal fold changes of 2 are shown as triangles. The different colors denote the ALE groups.

```
> vtmp <- spkVar(object)
> sv <- as.numeric(vtmp[, 2][-nrow(vtmp)])
> bin <- c("Low", "Med", "High")
> bins <- bin[spkSlopeOut$breaks[2, ]]
> tab1 <- data.frame(NominalConc = 2^spkSlopeOut$breaks[1, ], AvgExp = round(spkSlope
+     1), PropGenesBelow = round(spkSlopeOut$prop, 2), ALEStrata = bins,
+     SD = round(sv, 2))
> colnames(tab1) <- c("Nominal Conc", "Avg Expression", "Prop of Genes Below",
+     "ALE Strata", "Std Dev")
```

	Nominal Conc	Avg Expression	Prop of Genes Below	ALE Strata	Std Dev
1	0.12	5.10	0.35	Low	0.87
2	0.25	5.20	0.37	Low	0.90
3	0.50	5.30	0.40	Low	0.74
4	1.00	5.70	0.48	Low	0.72
5	2.00	6.40	0.62	Med	0.82
6	4.00	7.10	0.73	Med	0.79
7	8.00	7.80	0.82	Med	0.68
8	16.00	8.40	0.88	Med	0.67
9	32.00	9.30	0.94	Med	0.79
10	64.00	10.20	0.97	Med	0.72
11	128.00	11.20	0.99	Med	0.54
12	256.00	12.00	0.99	High	0.49
13	512.00	12.60	1.00	High	0.51

Nominal concentration to ALE mapping: This table contains summary measures specific to each nominal spike-in level. The first column shows the nominal concentrations as originally reported. The second column shows the average of all observed expression values associated with the row's nominal concentration. The third column shows the proportion of background RNA with expression values less than the average expression value. The fourth column shows the ALE strata associated with the row's nominal concentration. Finally, the fifth column shows the standard deviation of all observed expression values associated with the row's nominal concentration.

```
> AccuracySlope <- round(spkSlopeOut$slopes[-1], digits = 2)
> AccuracySD <- round(spkAccSD(object, spkSlopeOut), digits = 2)
> pot <- spkPot(object, spkSlopeOut, AccuracySlope, AccuracySD,
+     precisionQuantile = 0.995)
> PrecisionSD <- round(sbox$madFC[1:3], digits = 2)
> PrecisionQuantile <- round(pot$quantiles, digits = 2)
> SNR <- round(AccuracySlope/PrecisionSD, digits = 2)
> POT <- round(pot$POTs, digits = 2)
> tab2 <- data.frame(AccuracySlope = AccuracySlope, AccuracySD = AccuracySD,
+     PrecisionSD = PrecisionSD, PrecisionQuantile = PrecisionQuantile,
+     SNR = SNR, POT = POT)
```

	AccuracySlope	AccuracySD	PrecisionSD	PrecisionQuantile	SNR	POT
Low	0.20	0.31	0.10	0.36	2.00	0.30
Med	0.79	0.35	0.09	0.40	8.78	0.87
High	0.57	0.15	0.06	0.22	9.50	0.99

Assessment results: For each of the ALE strata we report summary assessments for accuracy, precision, and overall performance. The first column shows the signal detection slope which can be interpreted as the expected observed difference when the true difference is a fold change of 2. In parenthesis is the standard deviation of the log-ratios associated with non-zero nominal log-ratios. The second column shows the standard deviation of null log-ratios. The SD can be interpreted as the expected range of observed log-ratios for genes that are not differentially expressed. The third column shows the 99.5th percentile of the null distribution. It can be interpreted as the expected minimum value that the top 100 non-differentially expressed genes will reach. The fourth column shows the ratio of the values in column 1 and column 2. It is a rough measure of signal to noise ratio. The fifth column shows the probability that, when comparing two samples, a gene with a true log fold change of 2 will appear in a list of the 100 genes with the highest log-ratios.

```
> bals <- round(spkBal(object))
> anv <- round(spkAnova(object), digits = 2)
> tab3 <- t(c(anv, bals))
```

	spike	probe	array	error	Probe Imbalance	Array Imbalance
1	2.48	0.54	0.17	0.47	0.00	0.00

ANOVA results: To understand the variability contributed by differences in nominal concentrations, probe effect, and array, we fitted a 3-way ANOVA model containing only main effects to the expression values from the spike-in transcripts. The estimated standard deviation of each effect is shown in the first three columns. The forth column shows the standard deviation of the error term. Finally, a measure of the amount of confounding between nominal concentration and the other two effects is included in columns five and six. We use the measure presented by Wu in Technometrics (1981), Volume 23, Number 1. An optimal design, such as a Latin Square, will have a measure of 0 for each imbalance. The more confounding the larger these values. Because Affymetrix using a latin square design, there is no imbalance.