Sample Size Estimation for Microarray Experiments Using the `ssize` package.

Gregory R. Warnes
email:greg@random-technologies-llc.com

July 22, 2006

Abstract

mRNA Expression Microarray technology is widely applied in biomedical and pharmaceutical research. The huge number of mRNA concentrations estimated for each sample make it difficult to apply traditional sample size calculation techniques and has left most practitioners to rely on rule-of-thumb techniques. In this paper, we briefly describe and then demonstrate a simple method for performing and visualizing sample size calculations for microarray experiments as implemented in the `ssize` R package.

Note

This document is a simplified version of the manuscript

> Warnes, G. R., Liu, P. (2006) Sample Size Estimation for Microarray Experiments, Technical Report, Department of Biostatistics and Computational Biology, University of Rochester.

which has been available as a pre-publication manuscript since 2004. Please refer to that document for a detailed discussion of the sample size estimation method and an evaluation of its performance.

1 Introduction

High-throughput microarray experiments allow the measurement of expression levels for tens of thousands of genes simultaneously. These experiments have been used in many disciplines of biological research, including neuroscience (Mandel *et al.*, 2003), pharmacogenomic research, genetic disease and cancer diagnosis (Heller, 2002). As a tool for estimating gene expression and single nucleotide polymorphism (SNP) genotyping, microarrays produce huge amounts of data which can providing important new insights.

Microarray experiments are rather costly in terms of materials (RNA sample, reagents, chip, etc), laboratory manpower, and data analysis effort. It is critical, therefore, to perform proper experimental design, including sample size estimation, before carrying out these experiments. Since tens of thousands of variables (gene expressions) may be measured on each individual chip, it is essential appropriately take into account multiple testing and dependency among variables when calculating sample size.

2 Method

2.1 Overview

Warnes and Liu (2006) provide a simple method for computing sample size for microarray experiments, and reports on a series of simulations demonstrating its performance. Surprisingly, despite its simplicity, the method performs exceptionally well even for data with very high correlation between measurements.

The key component of this method is the generation of a cumulative plot of the proportion of genes achieving a desired power as a function of sample size, based on simple gene-by-gene calculations. While this mechanism can be used to select a sample size numerically based on pre-specified conditions, its real utility is as a visual tool for understanding the trade off between sample size and power. In our consulting work, this latter use as a visual tool has been exceptionally valuable in helping scientific clients to make the difficult trade offs between experiment cost and statistical power.

2.2 Assumptions

In the current implementation, we assume that a microarray experiment is set up to compare gene expressions between one treatment group and one control group. We further assume that microarray data has been normalized and transformed so that the data for each gene is sufficiently close to a normal distribution that a standard 2-sample pooled-variance t-test will reliably detect differentially expressed genes. The tested hypothesis for each gene is:

$$H_0 : \mu_T = \mu_C$$

versus

$$H_1 : \mu_T \neq \mu_C$$

where μ_T and μ_C are means of gene expressions for treatment and control group respectively.

2.3 Computations

The proposed procedure to estimate sample size is:

1. Estimate standard deviation (σ) for each gene based on *control samples* from existing studies performed on the same biological system. (While samples from the study to be performed are not, of course, generally available, control samples from other studies using the same biological system are often readily available.)
2. Specify values for
   (a) minimum effect size, Δ, (log of fold-change for log-transformed data)
   (b) maximum family-wise type I error rate, α
   (c) desired power, $1 - \beta$.
3. Calculate the per-test Type I error rate necessary to control the family-wise error rate (FWER) using the Bonferroni correction:
$$\alpha_G = \frac{\alpha}{G} \tag{1}$$
where G is the number of genes on the microarray chip.
4. Compute sample size separately for each gene according to the standard formula for the two-sample t-test with pooled variance:
$$\begin{aligned} 1 - \beta & \\ &= 1 - T_{n_1+n_2-2}\left(t_{\alpha_G/2,n_1+n_2-2} | \frac{\Delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right) \\ &\quad + \ T_{n_1+n_2-2}\left(-t_{\alpha_G/2,n_1+n_2-2} | \frac{\Delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right) \end{aligned} \tag{2}$$
where $\mathrm{T}_d(\bullet|\theta)$ is the cumulative distribution function for non-central t-distribution with d degree of freedom and the non-centrality parameter θ.
5. Summarize the necessary sample size across all genes using a cumulative plot of required sample size verses power. An example of such a plot is given in Figure 3 for which we assume equal sample size for the two groups, $n = n_1 = n_2$.

On the cumulative plot, for a point with x coordinate n, the y coordinate is the proportion of genes which require a sample size smaller than or equal to n, or equivalently the proportion of genes with power greater than or equal to the specified power $(1 - \beta)$ at sample size n. This plot allows users to visualize the relationship between power for all genes and required sample size in a single display. A sample size can thus be selected for a proposed microarray experiment based on user-defined criterion. For the plot in Figure 3, for example, requiring 80% of genes to achieve the 80% power yields a sample size of 10.

Similar plots can be generated by fixing the sample size and varying one of the other parameters, namely, significance level (α), power $(1 - \beta)$, or minimum effect size (Δ). Two such plots are shown in Figures 2 and 4.

2.4 Functions

There are three pairs of functions available in the `ssize` package.

```
pow(sd, n, delta, sig.level,
  alpha.correct = "Bonferonni")
power.plot(x, xlab = "Power",
  ylab = "Proportion of Genes with"
```

```
        " Power >= x",
  marks = c(0.7, 0.8, 0.9), ...)

ssize(sd, delta, sig.level, power,
  alpha.correct = "Bonferonni")
ssize.plot(x,
  xlab = "Sample Size (per group)",
  ylab = "Proportion of Genes Needing Sample"
        " Size <= n",
  marks = c(2, 3, 4, 5, 6, 8, 10, 20), ...)

delta(sd, n, power, sig.level,
  alpha.correct = "Bonferonni")
delta.plot (x, xlab = "Fold Change",
  ylab = "Proportion of Genes with "
        "Power >= 80\% at\\n"
 "Fold Change=delta",
  marks = c(1.5, 2, 2.5, 3, 4, 6, 10), ...)
```

pow, power.plot compute and display a cumulative plot of the fraction of genes achieving a specified power for a fixed sample size (`n`), effect size (`delta`), and significance level (`sig.level`).

ssize,ssize.plot compute and display a cumulative plot of the fraction of genes for which a specified sample size is sufficient to achieve a specified power (`power`), effect size (`delta`), and significance level (`sig.level`).

delta,delta.plot compute and display a cumulative plot of the fraction of genes which can achieve a specified power (`power`), for a specified sample size (`n`), and significance level (`sig.level`) for a range of effect sizes.

3 Example

First, we need to load the `ssize` library:

```
> library(ssize)
> library(xtable)
> library(gdata) # for nobs()
> options(width=30)
```

The `ssize` library provides an example data set containing gene expression values for smooth muscle cells from a control group of untreated healthy volunteers processed using Affymetrix U95 chips and normalized per the Robust Multi-array Average (RMA) method of Irizarry *et al.* (2003).

```
> # Load the example data
> data(exp.sd)
> # Use only the first 1000,
> # so examples run faster
> exp.sd <- exp.sd[1:1000]
```

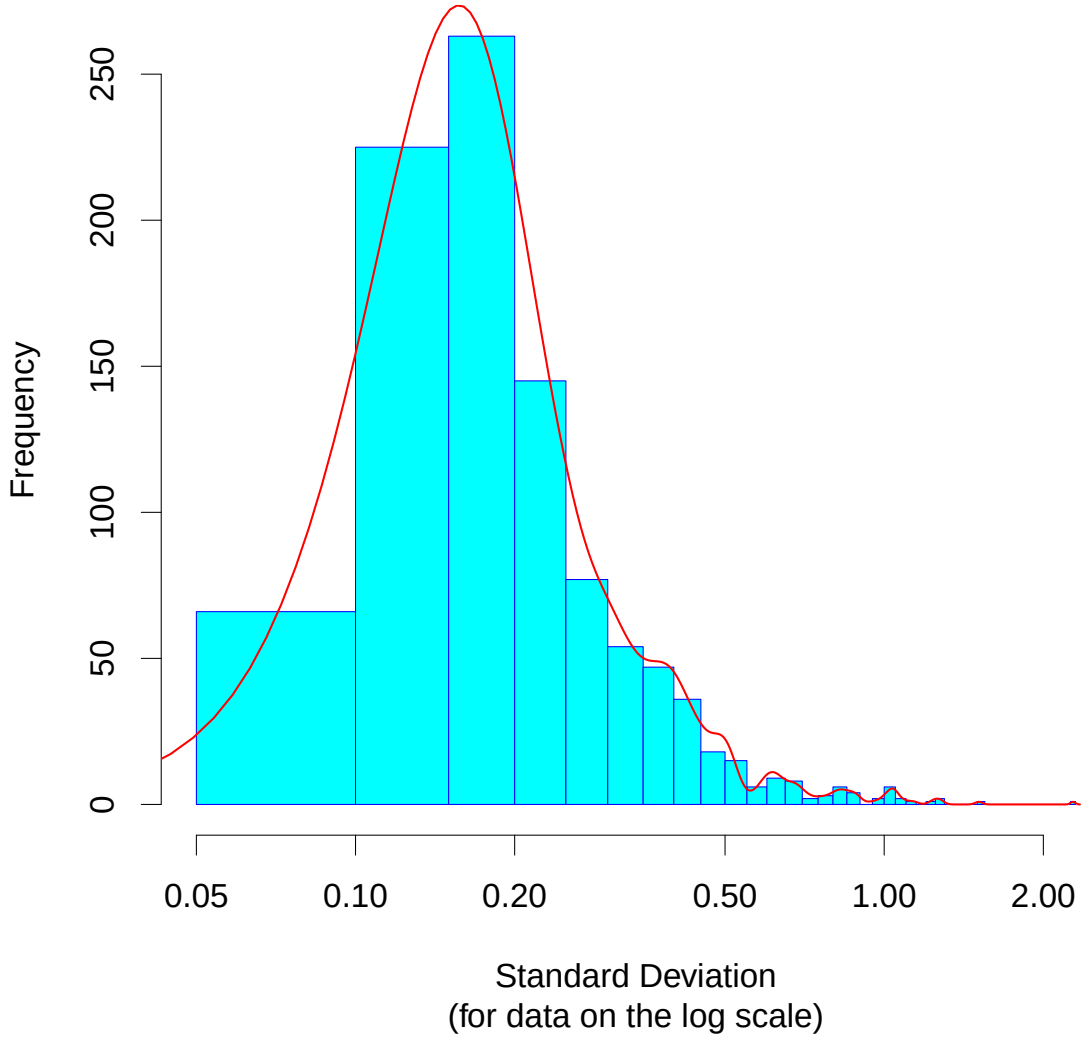
This data was calculated via:

```
library(affy)
load("probeset_data.Rda")
expression.values <- exprs(probeset.data)
covariate.data <- pData(probeset.data)
controls <- expression.values[,
            covariate.data$GROUP=="Control"] #$
exp.sd <- apply(controls, 1, sd)
```

Lets see what the distribution looks like:

```
> par(cex=2)
> xlab <- c("Standard Deviation", "(for data on the log scale)")
> hist(exp.sd,n=40, col="cyan", border="blue", main="", xlab=xlab, log="x")
> dens <- density(exp.sd)
> scaled.y <- dens$y*par("usr")[4]/max(dens$y)
> lines(dens$x,scaled.y ,col="red",lwd=2) #$
```

Figure 1: Standard deviations for of logged example data



As is often the case, this distribution is extremely right skewed, even though the standard deviations were computed on the $\log_2$ scale.

So, now lets see the functions in action. First, define the parameter values we will be investigating:

```
> n<-6
> fold.change<-2.0
> power<-0.8
> sig.level<-0.05
```

Now, the functions provided by the **ssize** package can be used to address several questions:

1. What is the necessary per-group sample size for 80% power when $\delta = 1.0$, and $\alpha = 0.05$?

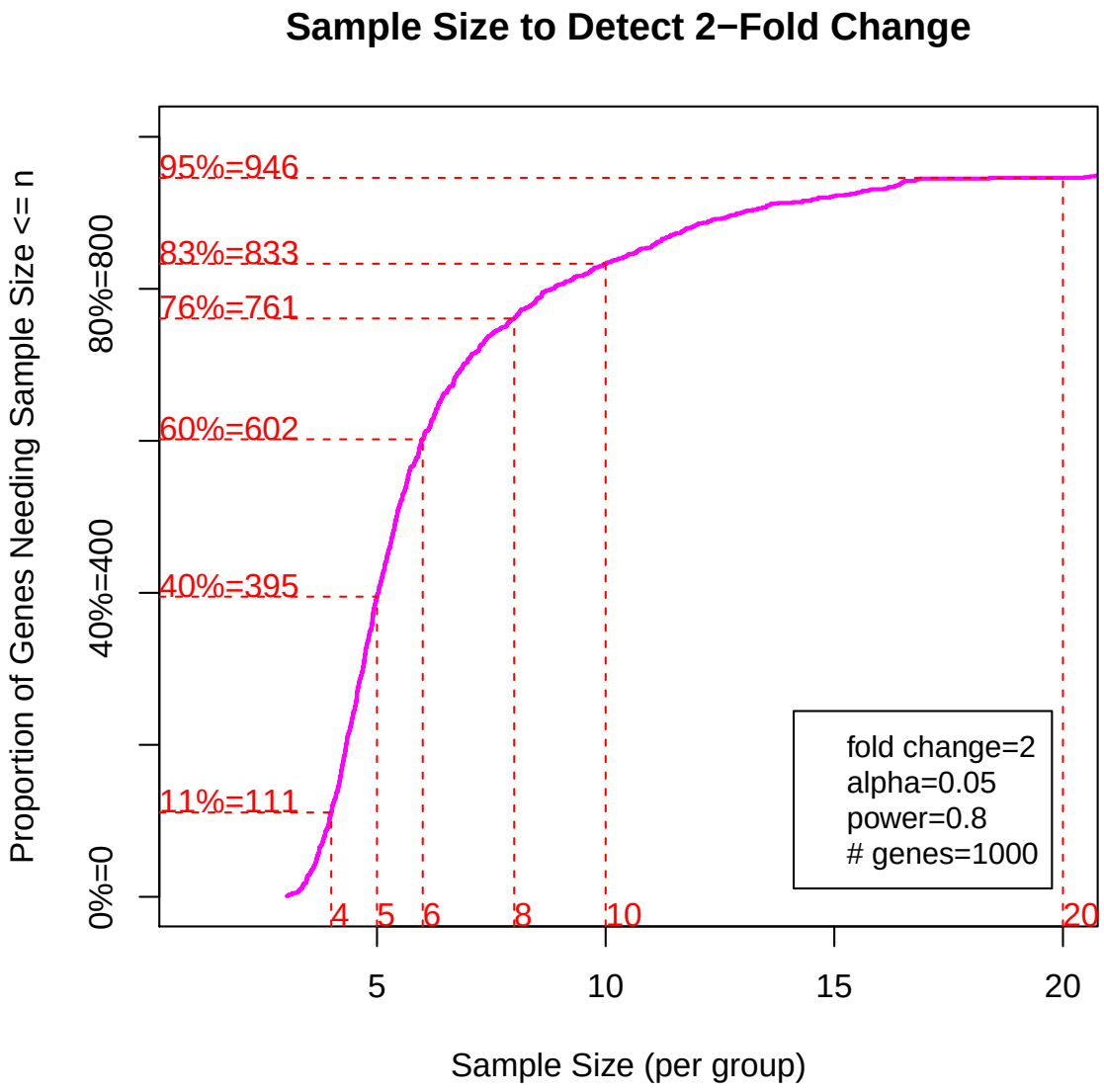
```
> all.size  <- ssize(sd=exp.sd, delta=log2(fold.change),
+                    sig.level=sig.level, power=power)

.....................................................................................................

> par(cex=1)
> ssize.plot(all.size, lwd=2, col="magenta", xlim=c(1,20))
```

```
> xmax <- par("usr")[2]-1;
> ymin <- par("usr")[3] + 0.05
> legend(x=xmax, y=ymin,
+        legend= strsplit( paste("fold change=",fold.change,",",
+          "alpha=", sig.level, ",",
+          "power=",power,",",
+          "# genes=", nobs(exp.sd), sep=''), "," )[[1]],
+        xjust=1, yjust=0, cex=0.90)
> title("Sample Size to Detect 2-Fold Change")
```

Figure 2: Sample size required to detect a 2-fold treatment effect.



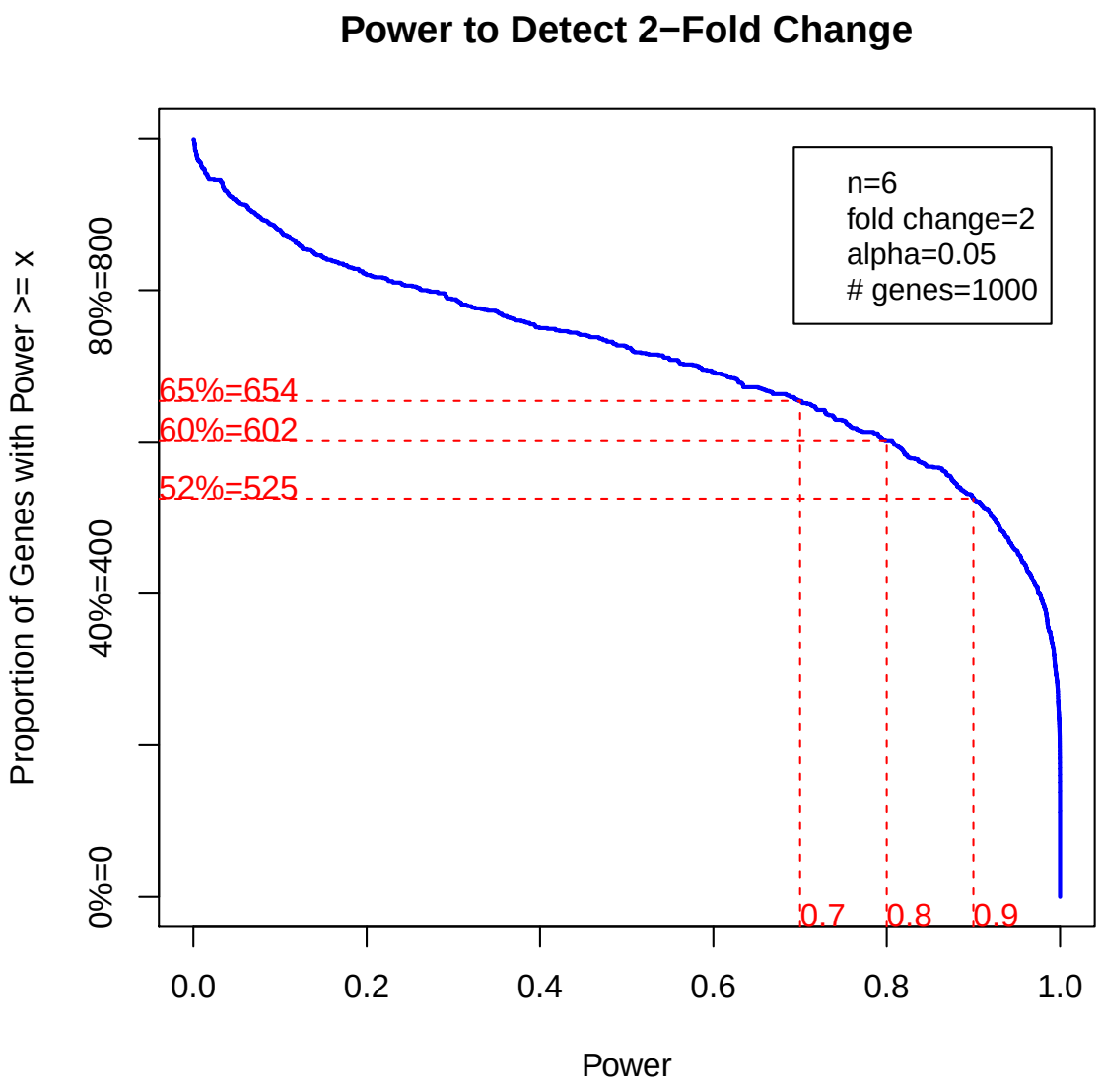
This plot illustrates that a sample size of 10 is required to ensure that at least 80% of genes have power greater than 80%. It also shows that a sample size of 6 is sufficient if only 60% of the genes need to achieve 80% power.

2. What is the power for 6 patients per group with $\delta = 1.0$, and $\alpha = 0.05$?

```
> all.power <- pow(sd=exp.sd, n=n, delta=log2(fold.change),
+                  sig.level=sig.level)
```

```
> par(cex=1)
> power.plot(all.power, lwd=2, col="blue")
> xmax <- par("usr")[2]-0.05; ymax <- par("usr")[4]-0.05
> legend(x=xmax, y=ymax,
+        legend= strsplit( paste("n=",n,",",
+          "fold change=",fold.change,",",
+          "alpha=", sig.level, ",",
+          "# genes=", nobs(exp.sd), sep=''), "," )[[1]],
+        xjust=1, yjust=1, cex=0.90)
> title("Power to Detect 2-Fold Change")
```

Figure 3: Effect of Sample Size on Power



This plot shows that only 52% of genes achieve at 80% power at this sample size and significance level.

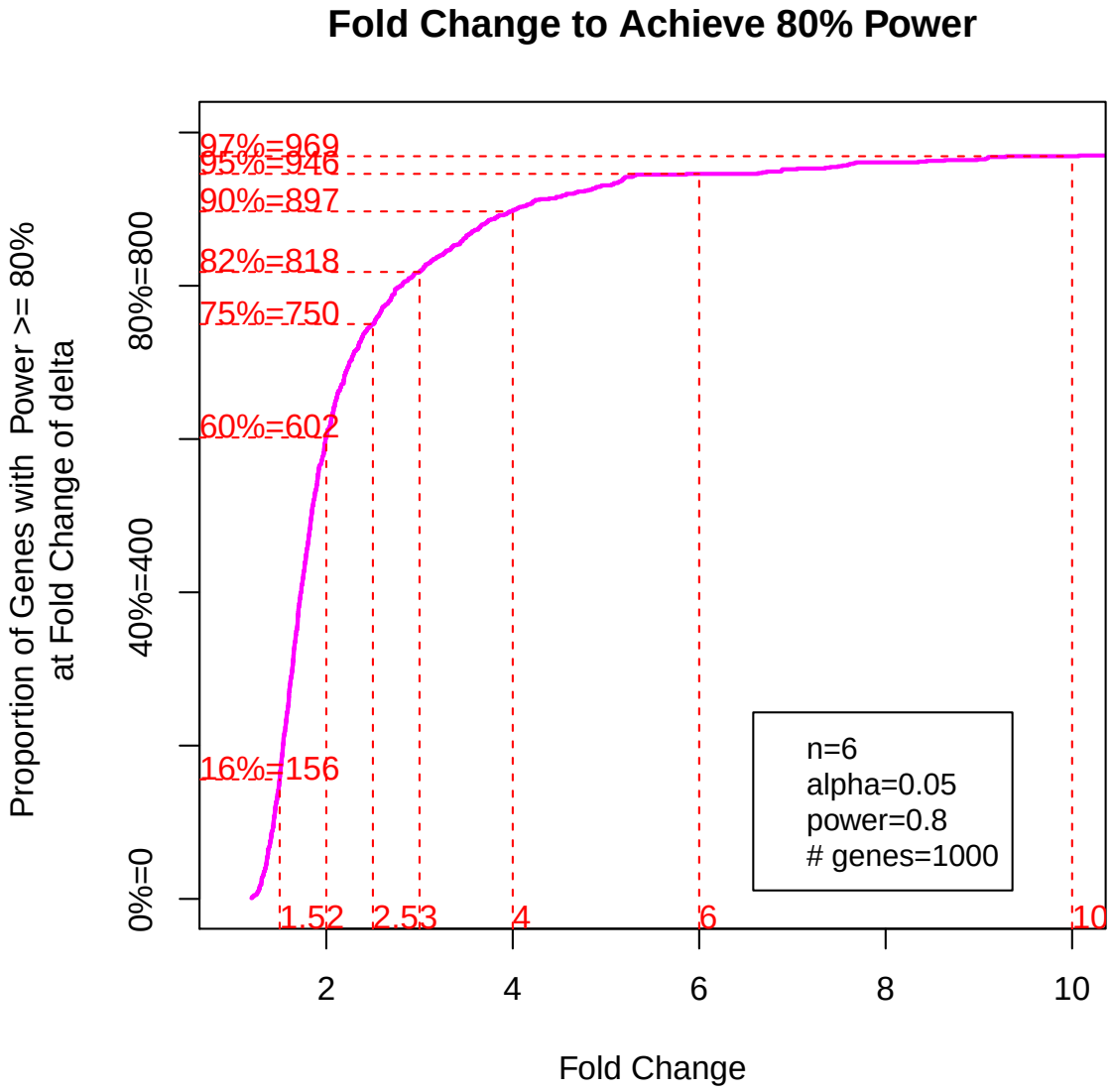
3. How large does a fold-change need to be for 80% of genes to achieve 80% power for an experiment for $n = 6$ patients per group and $\alpha = 0.05$?

```
> all.delta  <- delta(sd=exp.sd, power=power, n=n,
+                     sig.level=sig.level)
```

.....

```
> par(cex=1, mar=c(5.1,5.1,4,2))
> delta.plot(all.delta, lwd=2, col="magenta", xlim=c(1,10),
+            ylab = paste("Proportion of Genes with ",
+                         "Power >= 80% \n",
+                         "at Fold Change of delta")
+ )
> xmax <- par("usr")[2]-1; ymin <- par("usr")[3] + 0.05
> legend(x=xmax, y=ymin,
+        legend= strsplit( paste("n=",n,",",
+          "alpha=", sig.level, ",",
+          "power=",power,",",
+          "# genes=", nobs(exp.sd), sep=''), "," )[[1]],
+        xjust=1, yjust=0, cex=0.90)
> title("Fold Change to Achieve 80% Power")
```

Figure 4: Given sample size, this plot allows visualization of the fraction of genes achieving the specified power for different fold changes.



This plot shows that for a fold change of 2.0, only 60% of genes achieve 80% power, while a fold change of 3.0 will be detected with 80% power for 80% of genes.

4 Modifications

While the `ssize` package has been implemented using the simple 2-sample pooled t-test, you can easily modify the code for other circumstances. Simply replace the call to `power.t.test` in each of the functions `pow`,`ssize`,`delta` with the appropriate computation for the desired experimental design.

5 Future Work

Peng Liu is currently developing methods and code for substituting False Discovery Rate for the Bonferonni multiple comparison adjustment.

6 Contributions

Contributions and discussion are welcome.

7 Acknowledgment

This work was supported by Pfizer Global Research and Development.

References

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of Royal Statistical Society B*, **57:1**, 289-300.

Dow,G.S. (2003) Effect of sample size and p-value filtering techniques on the detection of transcriptional changes induced in rat neuroblastoma (NG108) cells by mefloquine, *Malaria Journal*, **2**, 4.

Heller, M. J. (2002) DNA microarray technology: devices, systems, and applications, *Annual Review in Biomedical Engineering*, **4**, 129-153.

Hwang,D., Schmitt, W. A., Stephanopoulos, G., Stephanopoulos, G. (2002) Determination of minimum sample size and discriminatory expression patterns in microarray data, *Bioinformatics*, **18:9**, 1184-1193.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., Speed, T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data, *Biostatistics*, **4:2**, 249-264.

Mandel, S., Weinreb, O., Youdim, M. B. H. (2003) Using cDNA microarray to assess Parkinson's disease models and the effects of neuroprotective drugs, *TRENDS in Pharmacological Sciences*, **24:4**, 184-191.

Yang, Y. H., Speed, T. Design and analysis of comparative microarray experiments *Statistical analysis of gene expression microarray data*, Chapman and Hall, 51.

Storey, J., (2002) A direct approach to false discovery rates, *Journal of Royal Statistical Society B*, **64:3**, 479-498.

Warnes, G. R., Liu, P. (2006) Sample Size Estimation for Microarray Experiments, Technical Report, Department of Biostatistics and Computational Biology, University of Rochester.

Yang, M. C. K., Yang, J. J., McIndoe, R. A., She, J. X. (2003) Microarray experimental design: power and sample size considerations, *Physiological Genomics*, **16**, 24-28.

Zien, A., Fluck, J., Zimmer, R., Lengauer, T. (2003) Microarrays: how many do you need?, *Journal of Computational Biology*, **10:3-4**, 653-667.