Specificity using Pickrell et al dataset

Michael Love

October 14, 2014

1 Load the benchmarking results

We load the benchmarking results, which were produced by the script `/inst/script/pickrellDiffExpr.R`. The object `resList` is a list, one element for each random replicate, of data frames which contain a column for each algorithm giving the p-values for each gene. The *SummarizedExperiment* object used for this analysis is contained in `/data/pickrell_sumexp.RData`.

We make note if any combination of algorithm and replicate had an estimated false positive rate larger than 5% as we will truncate the plot at this point for visibility.

```
library("DESeq2paper")
data("specificity")
alpha <- 0.01
# for cuffdiff2: set the genes with zero row sum to have p-value = NA as we
# do not count these to the denominator of the FPR for other algos
for (i in seq_len(length(resList))) {
    resList[[i]]$cuffdiff2[is.na(resList[[i]]$DESeq2)] <- NA
}
resMat <- t(sapply(resList, function(z) colMeans(z < alpha, na.rm = TRUE)))
colnames(resMat) <- namesAlgos
resMat <- subset(resMat, select = -EBSeq)  # EBSeq does not produce p-values
library("ggplot2")
d <- data.frame(fpr = as.vector(resMat), algorithm = factor(rep(colnames(resMat),
    each = nrow(resMat)), levels = colnames(resMat)))
# these points are outliers
d[d$fpr >= 0.05, ]

##         fpr    algorithm
## 101 0.05339 edgeR-robust
## 161 0.13324         voom
## 178 0.05923         voom
## 191 0.06140       SAMseq
```

The following function helps to rename algorithms.

```
renameAtoB <- function(f, a, b) {
    levels(f)[levels(f) == a] <- b
    f
}
```

```
d$algorithm <- renameAtoB(d$algorithm, "DESeq", "DESeq (old)")
d$algorithm <- renameAtoB(d$algorithm, "cuffdiff2", "Cuffdiff 2")
```

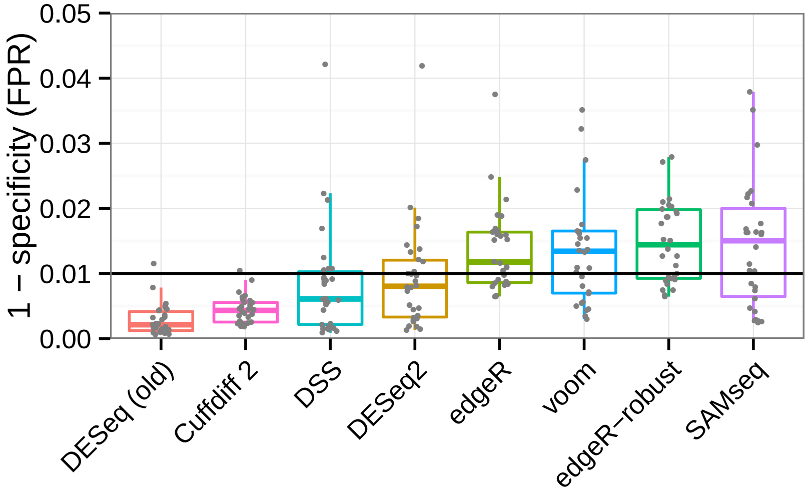

Figure 1: Estimate of E(p-value $< .01$), constructed by counting the number of p-values less than .01 and dividing by the total number of tests. Genes with all zero counts not included. Plot cropped to remove outliers printed in code above.

2 Plot

For the specificity, we simply plot the number of genes with p-value less than .01, as we have created mock comparisons of 5 vs 5 randomly chosen samples from a population with no known condition dividing them.

```
p <- ggplot(d, aes(x = reorder(algorithm, fpr, median), y = fpr, color = algorithm))
p + geom_boxplot(outlier.colour = rgb(0, 0, 0, 0)) + theme_bw() + geom_point(position = position_jitter(w =
    h = 0), color = "grey50", size = 1) + geom_hline(aes(yintercept = alpha)) +
    ylab("1 - specificity (FPR)") + theme(axis.text.x = element_text(angle = 45,
    hjust = 1)) + xlab("") + scale_colour_discrete(guide = "none") + coord_cartesian(ylim = c(0,
    0.05))
```

3 Session information

- R version 3.1.0 (2014-04-10), `x86_64-unknown-linux-gnu`
- Locale: `LC_CTYPE=en_US.UTF-8`, `LC_NUMERIC=C`, `LC_TIME=en_US.UTF-8`, `LC_COLLATE=C`, `LC_MONETARY=en_US.UTF-8`, `LC_MESSAGES=en_US.UTF-8`, `LC_PAPER=en_US.UTF-8`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_US.UTF-8`, `LC_IDENTIFICATION=C`
- Base packages: base, datasets, grDevices, graphics, grid, methods, parallel, splines, stats, utils
- Other packages: Biobase 2.24.0, BiocGenerics 0.10.0, DESeq2 1.4.0, DESeq2paper 1.3, Formula 1.1-1, GenomeInfoDb 1.0.0, GenomicRanges 1.16.0, Hmisc 3.14-4, IRanges 1.21.45, LSD 2.5, MASS 7.3-31, RColorBrewer 1.0-5, Rcpp 0.11.1, RcppArmadillo 0.4.200.0, abind 1.4-0, colorRamps 2.3, ellipse 0.3-8, ggplot2 0.9.3.1, gplots 2.13.0, gridExtra 0.9.1, gtools 3.3.1, hexbin 1.27.0, knitr 1.5, lattice 0.20-29, reshape 0.8.5, schoolmath 0.4, survival 2.37-7, vsn 3.32.0, xtable 1.7-3
- Loaded via a namespace (and not attached): AnnotationDbi 1.26.0, BiocInstaller 1.14.2, DBI 0.2-7, KernSmooth 2.23-12, RSQLite 0.11.4, XML 3.98-1.1, XVector 0.4.0, affy 1.42.2, affyio 1.32.0, annotate 1.42.0, bitops 1.0-6, caTools 1.16, cluster 1.15.2, codetools 0.2-8, colorspace 1.2-4, dichromat 2.0-0, digest 0.6.4, evaluate 0.5.5, formatR 0.10, gdata 2.13.3, genefilter 1.46.0, geneplotter 1.42.0, gtable 0.1.2, highr 0.3, labeling 0.2, latticeExtra 0.6-26, limma 3.20.1, locfit 1.5-9.1, munsell 0.4.2, plyr 1.8.1, preprocessCore 1.26.1, proto 0.3-10, reshape2 1.4, scales 0.2.3, stats4 3.1.0, stringr 0.6.2, tools 3.1.0, zlibbioc 1.10.0