A note on esApply

ExpressionSets are complex objects. \texttt{exprs(ExpressionSet)} produces $G \times N$, where $G$ is the number of genes on a chip and $N$ is the number of tissues analyzed, and \texttt{pData(ExpressionSet)} produces $N \times p$, where $p$ is the number of phenotypic or demographic, etc., variables collected.

Abstractly, we are often interested in evaluating functions $f(y; x)$ where $y$ is an $N$-vector of expression results for a specific gene and $x$ is an $N$-dimensional structure, coordinated with $y$, that distinguishes elements of $y$ for processing in the function $f$. A basic problem is to guarantee that the $j$th element of $y$ is correctly associated with the $j$th component of $x$.

As an example, let’s consider \texttt{sample.ExpressionSet}, which is an \texttt{ExpressionSet} supplied with Biobase. We will print a little report, then the first $N$-vector of gene expressions and some covariate data:

\begin{verbatim}
> print(sample.ExpressionSet)
ExpressionSet (storageMode: lockedEnvironment)
assayData: 500 features, 26 samples
  element names: exprs, se.exprs
protocolData: none
phenoData
  sampleNames: A B ... Z (26 total)
  varLabels: sex type score
  varMetadata: labelDescription
featureData: none
experimentData: use 'experimentData(object)'
Annotation: hgu95av2

> print(exprs(sample.ExpressionSet)[1,])
   A     B     C     D     E     F     G
192.7420 85.7533 176.7570 135.5750 64.4939 76.3569 160.5050
   H     I     J     K     L     M     N
  65.9631 56.9039 135.6080  63.4432  78.2126  83.0943  89.3372
   O     P     Q     R     S     T     U
  91.0615 95.9377 179.8450 152.4670 180.8340 157.9890
   V     W     X     Y     Z
 146.8000 93.8829 103.8550  64.4340 175.6150

> print(pData(sample.ExpressionSet)[1:2,1:3])

    sex   type  score
   A Female Control  0.75
   B   Male   Case  0.40
\end{verbatim}
Now let's see how expressions and a covariate are related:

```r
> print(rbind(exprs(sample.ExpressionSet[1,]),
+   sex <- t(pData(sample.ExpressionSet))[1,]))

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>192.742</td>
<td>85.7533</td>
<td>176.757</td>
<td>135.575</td>
<td>64.4939</td>
</tr>
<tr>
<td>F</td>
<td>Female</td>
<td>Male</td>
<td>Male</td>
<td>Male</td>
<td>Female</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>76.3569</td>
<td>160.505</td>
<td>65.9631</td>
<td>56.9039</td>
<td>135.608</td>
</tr>
<tr>
<td>G</td>
<td>Male</td>
<td>Male</td>
<td>Male</td>
<td>Female</td>
<td>Male</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>63.4432</td>
<td>78.2126</td>
<td>83.0943</td>
<td>89.3372</td>
<td>91.0615</td>
</tr>
<tr>
<td>H</td>
<td>Male</td>
<td>Female</td>
<td>Male</td>
<td>Male</td>
<td>Female</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>95.9377</td>
<td>179.845</td>
<td>152.467</td>
<td>180.834</td>
<td>85.4146</td>
</tr>
<tr>
<td>I</td>
<td>Female</td>
<td>Female</td>
<td>Male</td>
<td>Male</td>
<td>Female</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>157.989</td>
<td>146.8</td>
<td>93.8829</td>
<td>103.855</td>
<td>64.434</td>
</tr>
<tr>
<td>J</td>
<td>Male</td>
<td>Female</td>
<td>Male</td>
<td>Male</td>
<td>Female</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>175.615</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>Female</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

A function that evaluates the difference in median expression across strata defined using an abstract covariate \( x \) is

```r
> medContr <- function( y, x ) {
+   ys <- split(y,x)
+   median(ys[[1]]) - median(ys[[2]])
+ }
```

We can apply this to a small `ExpressionSet` that gives back the data listed above:

```r
> print(apply(exprs(sample.ExpressionSet[1,,drop=F]), 1,
+   medContr, pData(sample.ExpressionSet)[["sex"]]))

AFFX-MurIL2_at
-12.7935
```

That's a bit clumsy. This is where `esApply` comes in. We pay for some simplicity by following a strict protocol for the definition of the statistical function to be applied.
> medContr1 <- function(y) {
+  ys <- split(y,sex)
+  median(ys[[1]]) - median(ys[[2]])
+ }
> print(esApply( sample.ExpressionSet, 1, medContr1)[1])

AFFX-MurIL2_at
-12.7935

The manual page on esApply has a number of additional examples that show how applicable functions can be constructed and used. The important thing to note is that the applicable functions know the names of the covariates in the pData dataframe.

This is achieved by having an environment populated with all the variables in phenoData(ExpressionSet) put in as the environment of the function that will be applied. If that function already has an environment we retain that but in the second position. Thus, there is some potential for variable shadowing.

1 Session Information

The version number of R and packages loaded for generating the vignette were:

R version 3.4.0 (2017-04-21)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 16.04.2 LTS

Matrix products: default
BLAS: /home/biocbuild/bbs-3.5-bioc/R/lib/libRblas.so
LAPACK: /home/biocbuild/bbs-3.5-bioc/R/lib/libRlapack.so

locale:
[1] LC_CTYPE=en_US.UTF-8   LC_NUMERIC=C
[3] LC_TIME=en_US.UTF-8    LC_COLLATE=C
[5] LC_MONETARY=en_US.UTF-8 LC_MESSAGES=en_US.UTF-8
[7] LC_PAPER=en_US.UTF-8   LC_NAME=C
[9] LC_ADDRESS=C           LC_TELEPHONE=C

attached base packages:
[1] parallel  stats    graphics  grDevices utils     datasets
[7] methods   base

other attached packages:

3
Biobase_2.36.0    BiocGenerics_0.22.0

loaded via a namespace (and not attached):

[1] compiler_3.4.0 tools_3.4.0