Introduction to antiProfiles

Héctor Corrada Bravo hcorrada@gmail.com

Modified: March 13, 2013. Compiled: October 24, 2023

Introduction

This package implements the gene expression anti-profiles method in [1]. Anti-profiles are a new approach for developing cancer genomic signatures that specifically takes advantage of gene expression heterogeneity. They explicitly model increased gene expression variability in cancer to define robust and reproducible gene expression signatures capable of accurately distinguishing tumor samples from healthy controls.

In this vignette we will use the companion `antiProfilesData` package to illustrate some of the analysis in that paper.

```
> # these are libraries used by this vignette
> require(antiProfiles)
> require(antiProfilesData)
> require(RColorBrewer)
```

Colon cancer expression data

The `antiProfilesData` package contains expression data from normal colon tissue samples and colon cancer samples from two datasets in the Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE8671 and http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE4183. Probesets annotated to genes within blocks of hypo-methylation in colon cancer defined in [2]. Let's load the data and take a look at its contents.

```
> data(apColonData)
> show(apColonData)
```

```
ExpressionSet (storageMode: lockedEnvironment)
assayData: 5339 features, 68 samples
  element names: exprs
protocolData: none
phenoData
  sampleNames: GSM95473 GSM95474 ... GSM215114 (68 total)
  varLabels: filename DB_ID ... Status (7 total)
  varMetadata: labelDescription
featureData: none
experimentData: use 'experimentData(object)'
Annotation: hgu133plus2
```

```
> # look at sample types by experiment and status
> table(apColonData$Status, apColonData$SubType, apColonData$ExperimentID)
```

```
, ,  = GSE4183


    adenoma colorectal_cancer normal tumor
  0       0                 0      8     0
  1      15                15      0     0

, ,  = GSE8671


    adenoma colorectal_cancer normal tumor
  0       0                 0      7     0
  1       0                 0      0    23
```

The data is stored as an `ExpressionSet`. This dataset contains colon adenomas, benign but hyperplastic growths along with the normal and tumor tissues. Let's remove these from the remaining analysis.

```
> drop=apColonData$SubType=="adenoma"
> apColonData=apColonData[,!drop]
```

Building antiprofiles

The general anti-profile idea is to find genes with hyper-variable expression in cancer with respect to normal samples and classify new samples as normal vs. cancer based on deviation from a normal expression profile built from normal training samples. Anti-profiles are built using the following general algorithm:

1. Find candidate differentially variable genes (anti-profile genes): rank by ratio of cancer to normal variance
2. Define region of normal expression for each anti-profile gene: normal median $\pm$ 5 * normal MAD
3. For each sample to classify:
   (a) count number of antiProfile genes outside normal expression region (anti-profile score)
   (b) if score is above threshold, then classify as cancer

We will use data from one of the experiments to train the anti-profile (steps 1 and 2 above) and test it on the data from the other experiment (step 3).

Computing variance ratios

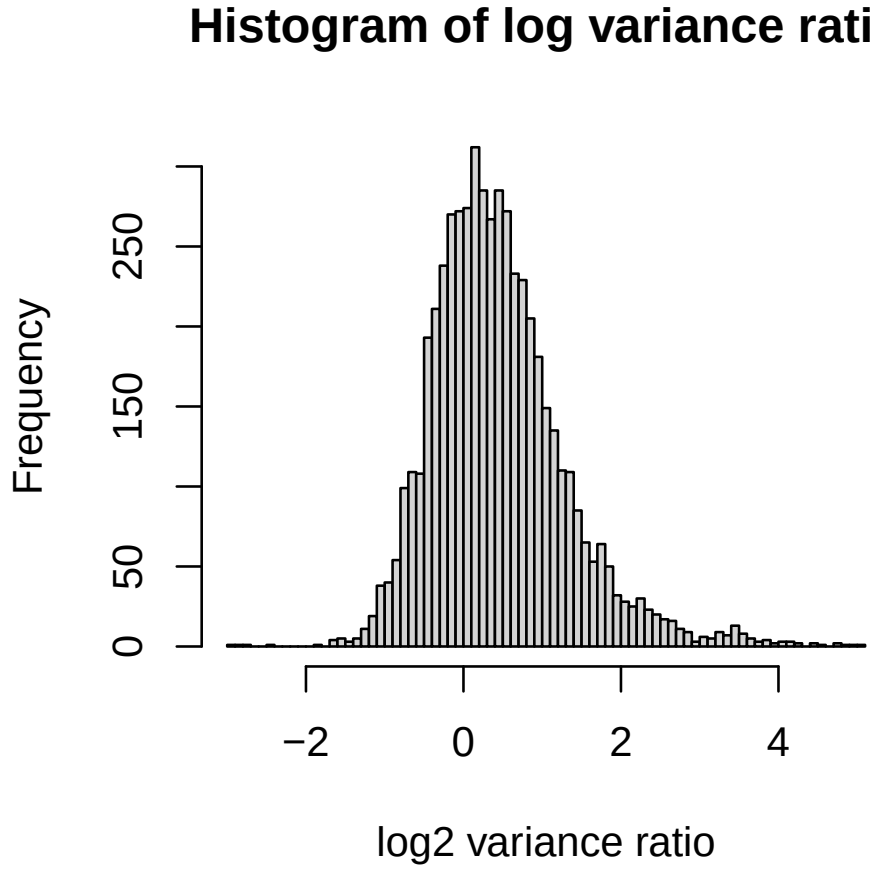
The first step in building an antiprofile is to calculate the ratio of normal variance to cancer variance. This is done with the `apStats` function.

```
> trainSamples=pData(apColonData)$ExperimentID=="GSE4183"
> colonStats=apStats(exprs(apColonData)[,trainSamples],
+                    pData(apColonData)$Status[trainSamples],minL=5)
> head(getProbeStats(colonStats))
```

```
                affyid        SD0      SD1     stat      meds0      mads0
214974_x_at 214974_x_at 0.23369554 7.777167 5.056543 -1.5182409 0.20537749
210118_s_at 210118_s_at 0.12684950 3.991524 4.975750 -0.7449072 0.08599552
205719_s_at 205719_s_at 0.09529811 2.817550 4.885850 -0.8532822 0.08670758
205863_at     205863_at 0.12430282 3.431335 4.786839 -0.5508941 0.15574925
215101_s_at 215101_s_at 0.24540888 6.578072 4.744406 -0.9257671 0.15043660
227140_at     227140_at 0.24976921 5.718607 4.516996 -1.8026488 0.13726594
```

We can see how that ratio is distributed for these probesets:

```
> hist(getProbeStats(colonStats)$stat, nc=100,
+      main="Histogram of log variance ratio", xlab="log2 variance ratio")
```



Building the anti-profile

Now we construct the anti-profile by selecting the 100 probesets most hyper-variable probesets

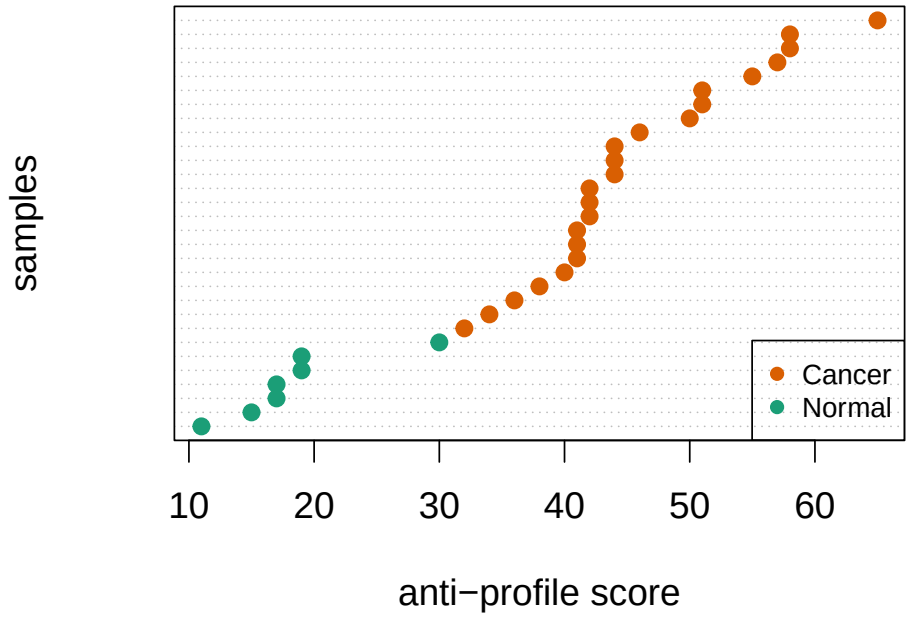
```
> ap=buildAntiProfile(colonStats, tissueSpec=FALSE, sigsize=100)
> show(ap)
```

```
AntiProfile object with 100 probes
Normal medians
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.8026 -0.8542 -0.5102  0.6835 -0.1187 23.6885
Using cutoff 5
```

Computing the anti-profile score

Given the estimated anti-profile, we can get anti-profile scores for a set of testing samples.

```
> counts=apCount(ap, exprs(apColonData)[,!trainSamples])
> palette(brewer.pal(8,"Dark2"))
> # plot in score order
> o=order(counts)
> dotchart(counts[o],col=pData(apColonData)$Status[!trainSamples][o]+1,
+          labels="",pch=19,xlab="anti-profile score",
+          ylab="samples",cex=1.3)
> legend("bottomright", legend=c("Cancer","Normal"),pch=19,col=2:1)
```



The anti-profile score measures deviation from the normal expression profile obtained from the training samples. We see in this case that the anti-profile score can distinguish the test samples perfectly based on deviation from the normal profile.

References

[1] Hector Corrada Bravo, Vasyl Pihur, Matthew McCall, Rafael A Irizarry, and Jeffrey T Leek. Gene expression anti-profiles as a basis for accurate universal cancer signatures. *BMC Bioinformatics*, 13(1):272, October 2012.

[2] Kasper Daniel Hansen, Winston Timp, Héctor Corrada Bravo, Sarven Sabunciyan, Benjamin Langmead, Oliver G McDonald, Bo Wen, Hao Wu, Yun Liu, Dinh Diep, Eirikur Briem, Kun Zhang, Rafael A Irizarry, and Andrew P Feinberg. Increased methylation variation in epigenetic domains across cancer types. *Nature Genetics*, 43(8):768–775, August 2011.

SessionInfo

- R version 4.3.1 (2023-06-16 ucrt), `x86_64-w64-mingw32`
- Locale: `LC_COLLATE=C`, `LC_CTYPE=English_United States.utf8`, `LC_MONETARY=English_United States.utf8`, `LC_NUMERIC=C`, `LC_TIME=English_United States.utf8`
- Time zone: `America/New_York`
- TZcode source: `internal`
- Running under: `Windows Server 2022 x64 (build 20348)`
- Matrix products: default
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: Biobase 2.62.0, BiocGenerics 0.48.0, RColorBrewer 1.1-3, antiProfiles 1.42.0, antiProfilesData 1.37.0, locfit 1.5-9.8, matrixStats 1.0.0
- Loaded via a namespace (and not attached): compiler 4.3.1, grid 4.3.1, lattice 0.22-5, tools 4.3.1