

The CTC saga: C2 comparison

Stelios Sfakianakis

Date: 2013/03/15

This is the comparison between peripheral blood of cancer patients and normal individuals, which identifies genes expressed in cancer blood. We are using the following datasets:

- [GSE27562](#): “In total, we collected blood from **57** women with a diagnosis of breast cancer and **37** with a benign diagnosis” (PMID: [21781289](#)) Platform: Affymetrix
- [GSE16443](#): “Blood samples were collected from **121** females referred for diagnostic mammography following an initial suspicious screening mammogram. Diagnostic work-up revealed that **67** of these women had breast cancer while **54** had no malignant disease. Additionally, 9 samples from 6 healthy female controls (three pregnant women, one breast-feeding woman and two healthy controls at different timepoints in their menstrual cycle) were included.” (PMID: [20078854](#)). Platform: Applied Biosystems (ABI)

Loading the data and performing the merge:

```
> require(inSilicoMerging)
> require(siggenes)
> load("data/gse27562_c2.rda")
> load("data/gse16443.rda")
> datasets = list(GSE27562 = gse27562, GSE16443 = gse16443)
> mgse_COMBAT <- merge(datasets, method = "COMBAT")
```

We select the samples to be used for this comparison:

```
> ind <- gse27562$characteristics_ch1 == "phenotype: Normal" |
+   gse27562$characteristics_ch1 == "phenotype: Malignant" |
+   gse27562$characteristics_ch1 == "phenotype: Pre-Surgery (aka Malignant)"
> gse27562_samples = sampleNames(gse27562[, ind])
```

```
> gse16443_samples = sampleNames(gse16443)
> samples = c(gse16443_samples, gse27562_samples)
> c2.data = mgse_COMBAT[, samples]
```

The distribution of the samples is shown in Table 1. % latex table generated in R 2.15.1 by xtable 1.7-1 package % Thu Jun 13 14:53:25 2013

	GSE16443	GSE27562
Cancer	67	57
Control	54	31

Table 1: Sample distribution

```
> plotMDS(c2.data, "Study", "Disease")

> cl = ifelse(c2.data$Disease == "Control", 0, 1)
> sam.out <- sam(exprs(c2.data), cl, B = 500, rand = 57005)
> summary(sam.out)
```

SAM Analysis for the Two-Class Unpaired Case Assuming Unequal Variances

s0 = 0

Number of permutations: 500

MEAN number of falsely called variables is computed.

	Delta	p0	False	Called	FDR	cutlow	cutup	j2	j1
1	0.1	0.589	4545.612	5392	0.49653	-0.241	0.573	3331	4558
2	0.4	0.589	2379.358	3885	0.36072	-0.631	1.299	2716	5450
3	0.7	0.589	693.006	2179	0.18732	-1.375	2.043	1645	6085
4	1.0	0.589	162.486	1188	0.08056	-2.032	2.757	974	6405
5	1.3	0.589	26.558	553	0.02829	-2.702	3.546	497	6563
6	1.5	0.589	6.686	315	0.01250	-3.148	4.059	295	6599
7	1.8	0.589	0.822	137	0.00353	-3.755	Inf	137	6619
8	2.1	0.589	0.066	54	0.00072	-4.386	Inf	54	6619
9	2.4	0.589	0.002	13	9.06e-05	-5.168	Inf	13	6619
10	2.7	0.589	0	0	0	-Inf	Inf	0	6619

```
> findDelta(sam.out, fdr = 0.05)
```

The threshold seems to be at

	Delta	Called	FDR
5	1.147	835	0.05030
6	1.147	826	0.04999

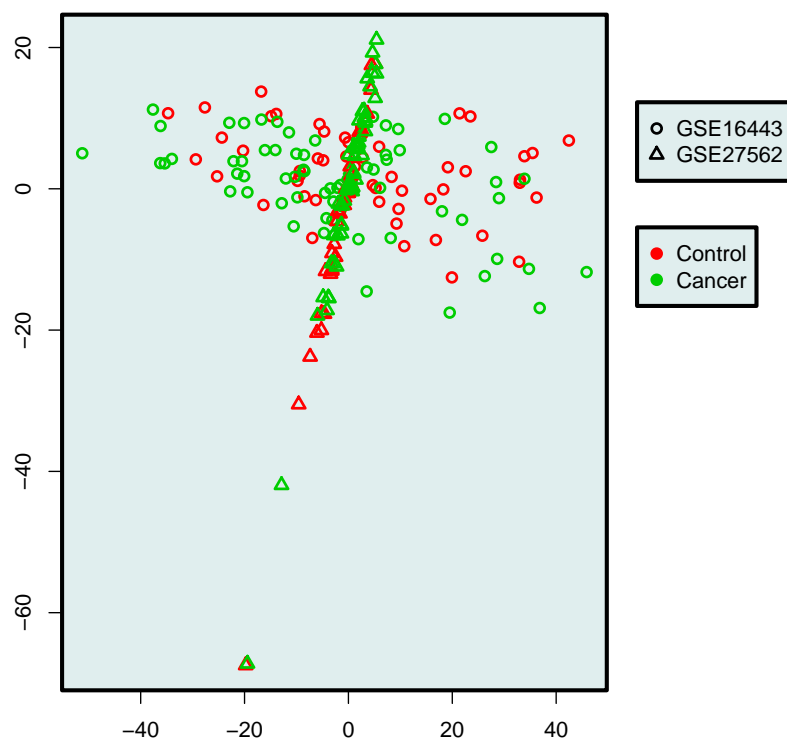


Figure 1: MDS plot of the merged samples

```
> delta <- findDelta(sam.out, fdr = 0.05, verbose = FALSE)[2, 1]
```

The threshold seems to be at

	Delta	Called	FDR
5	1.147	835	0.05030
6	1.147	826	0.04999

```
> sam.sum <- summary(sam.out, delta)
> w <- which(sam.sum@mat.sig$d.value > 0)
> num.genes.over <- length(w)
> num.genes.over
```

```
[1] 103
```

So we find 103 genes overexpressed in cancer peripheral blood.

```
> siggenes.all <- list.siggenes(sam.out, delta)
> siggenes.over <- list.siggenes(sam.out, delta)[w]
> ee = exprs(c2.data)[sam.sum@row.sig.genes, ]
> dim(ee)
```

```
[1] 826 209
```