

# The CTC saga: C2 comparison

Stelios Sfakianakis

June 14, 2013

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)
> require(xtable)
> 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.

	GSE16443	GSE27562
Cancer	67	57
Control	54	31

Table 1: Sample distribution

And the distribution of samples in a 2D MDS plot are shown in Figure 1.

```

> 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	0.58	4588.124	5452	0.487998	-0.243	0.551	3332	4499
2	0.4	0.58	2441.588	3970	0.356633	-0.617	1.271	2755	5404
3	0.7	0.58	686.398	2167	0.183678	-1.378	2.052	1639	6091
4	1.0	0.58	168.1	1198	0.081367	-2.009	2.805	1010	6431
5	1.3	0.58	24.652	524	0.027281	-2.728	3.601	477	6572
6	1.6	0.58	3.476	234	0.008614	-3.344	Inf	234	6619
7	1.8	0.58	0.74	135	0.003179	-3.770	Inf	135	6619
8	2.1	0.58	0.06	54	0.000644	-4.386	Inf	54	6619
9	2.4	0.58	0.004	12	0.000193	-5.190	Inf	12	6619
10	2.7	0.58	0	7	0	-5.592	Inf	7	6619

```

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

```

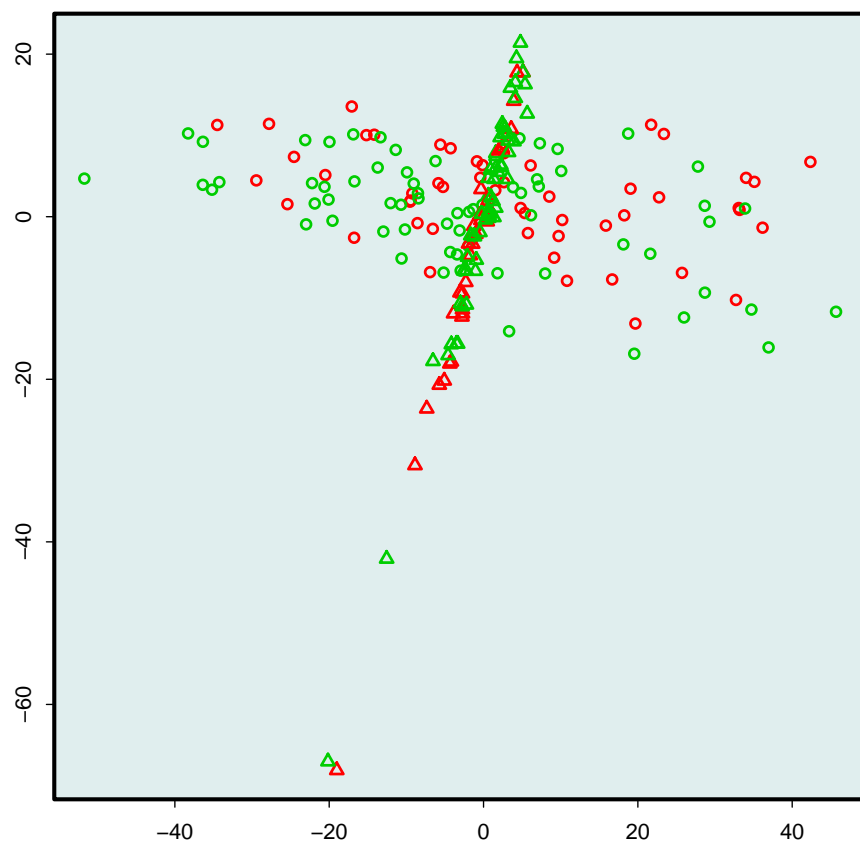


Figure 1: MDS plot of the merged samples

The threshold seems to be at

	Delta	Called	FDR
5	1.144	808	0.05052
6	1.144	777	0.04934

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

```
[1] 79
```

So we find 79 genes overexpressed in cancer peripheral blood.

```
[1] 777 209
```

They can be found in the `c2_siggenes.txt` file.