The CTC saga: C3 comparison

Stelios Sfakianakis

Date: 2014-05-29

C3 comparison

This is the comparison between cancer tissue and normal blood. We are using the cancer tissue samples from the c1 comparison, i.e.:

- GSE22820:"Gene expression profiles were generated from **176** primary breast cancer patients and 12 normal breast samples", PMID: 21356353 (*Agilent Whole Human Genome Microarray 4x44K G4112F*): **176 cancer tissues** (we don't take 2 metastatic cancers)

- GSE19783:"mRNA profiling from **115** breast cancer samples was performed", PMID: 21364938 (*Agilent Whole Human Genome Microarray 4x44K G4112F*) : **113 cancer tissues**

- GSE31364:"Seventy-two primary breast cancer tumor have been analyzed against a breast cancer reference pool.", PMID: 22384245 (*Agendia_human_DiscoverPrint_v1* **custom platform**) : **72 cancer tissues**

- GSE18672: "Mammographic density and genetics - A study of breast biopsies in relation to mammoraphic density", PMID: 20799965 (Agilent Whole Human Genome Oligo Microarray G4112A): **64 cancer tissues**

Furthermore we are using the following datasets providing cancer samples:

- GSE15852: "Gene expression patterns distinguish breast carcinomas from normal breast tissues: the Malaysian context" (PMID: 20097481). Platform: Affymetrix genechip U133A. **43 cancer patients**

- GSE12763: "In vivo antitumor activity of MEK and phosphatidylinositol 3-kinase inhibitors in basal-like breast cancer models" (PMID: 19567590). Platform: Affymetrix Human Genome U133 Plus 2.0 Array **30 human breast cancers**

For the normal blood samples we are using the ones from (my) c2 comparison, that is:

- 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 : **31 normal blood samples**

- GSE16443: "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) : **54 normal blood samples**

The merge of these datasets result in 6421 (common) genes and 510 samples (85 normal blood samples and 425 cancer tissue).

```
> print.xtable(xtable(table(c3.data$Origin, c3.data$Study)), comment=FALSE)
```

	GSE12763	GSE15852	GSE16443	GSE19783	GSE22820	GSE27562	GSE31364
Cancer Tissue	30	43	0	113	176	0	72
Normal Blood	0	0	54	0	0	31	0

```
> plotMDS(c3.data, "Study", "Origin", main="Distribution of samples in the merged datasets")
```

Differential expression

```
> cl = ifelse(c3.data$Origin == 'Normal Blood', 0, 1)
> sam.out = sam(exprs(c3.data), cl, B=500, rand=0xDEAD)
> delta = findDelta(sam.out, fdr=0.05)[2,1]


The threshold seems to be at
  Delta Called      FDR
5 1.036     316 0.05024
6 1.036     312 0.04960


> sam.sum <- summary(sam.out, delta)
> w <- which(sam.sum@mat.sig$d.value > 0)
> num.genes.over <- length(w)
> siggenes.over <- list.siggenes(sam.out, delta)[w]
```
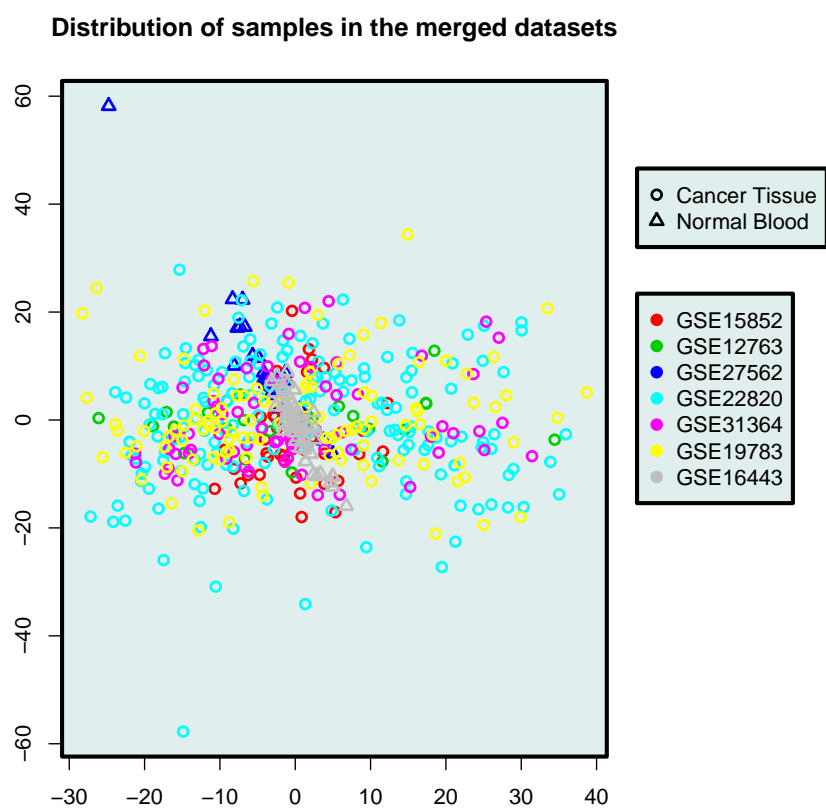
Figure 1: 2D (MDS) plot of the merged dataset

We find 241 genes over-expressed in cancer samples. They can be found in the
`c3_siggenes.txt` file.

```
> sessionInfo()

R version 3.0.3 (2014-03-06)
Platform: x86_64-unknown-linux-gnu (64-bit)

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
 [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
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

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

other attached packages:
 [1] plyr_1.8.1             ggplot2_1.0.0         xtable_1.7-3
 [4] inSilicoMerging_1.6.0 DWD_0.11              Matrix_1.1-3
 [7] siggenes_1.36.0       multtest_2.18.0       Biobase_2.22.0
[10] BiocGenerics_0.8.0    knitr_1.6

loaded via a namespace (and not attached):
 [1] colorspace_1.2-4 digest_0.6.4    evaluate_0.5.5   formatR_0.10
 [5] grid_3.0.3       gtable_0.1.2    lattice_0.20-29  MASS_7.3-33
 [9] munsell_0.4.2    proto_0.3-10    Rcpp_0.11.1      reshape2_1.4
[13] scales_0.2.4     stats4_3.0.3    stringr_0.6.2    survival_2.37-7
[17] tools_3.0.3
```