The CTC saga: C4 comparison

Stelios Sfakianakis

Date: 2014-10-07

Datasets

C4 is the comparison between cancer blood and normal tissue. We will merge an number of datasets to create a big one and then perform common differential expression algorithms.

For the normal tissue samples we take the following datasets from the C1 comparison:

- 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*). The GEO dataset contains **176 primary breast cancer** and **10 normal patients**.
- GSE9574: "Gene expression abnormalities in histologically normal breast epithelium of breast cancer patients", PMID: 18058819 (*Affymetrix Human Genome U133A Array*). We selected **15 normal samples** that have been identified as "disease-state: normal (reduction mammoplasty)"
- GSE18672: "Mammographic density and genetics - A study of breast biopsies in relation to mammoraphic density", PMID: 20799965 (Agilent Whole Human Genome Oligo Microarray G4112A). Here we used the the **79 women with no malignancy** (healthy women)

Additionally the following datasets from the C2 comparison are used for the cancer blood samples:

- 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)

Filtering Genes per Dataset

A "non specific" filtering is performed in each data set:

```
> nsFilter(gse, require.entrez=TRUE,remove.dupEntrez=TRUE,
>              var.func=IQR, var.cutoff=0.5, var.filter=FALSE)
```

where `gse` is the dataset. This means that probes/transcripts that do not map in Entrez genes identifiers are removed and in the case that multiple probes map to the same identifier we remove all but one that has the largest IQR (i.e. the largest variance).

Now loading the data and performing the merge:

Merging

I am using the inSilicoMerging package of Bioconductor that features some of the most well known cross platform normalization methods (COMBAT is Empirical Bayes, DWD stands for Distance Weighted Discrimination, and GENENORM means gene "z-score" normalization) for the merger of these datasets:

```
> require(inSilicoMerging)
> mgse_COMBAT <- merge(datasets, method="COMBAT");
```

The distribution of the different classes in the initial datasets is shown in the next table:

	GSE16443	GSE18672	GSE22820	GSE27562	GSE9574
Cancer	67	64	176	131	14
Control	54	0	0	31	0
Control Tissue	0	79	10	0	15

For each dataset we keep specific samples in the final (merged) dataset.

```
> ind1 = grep("Malignant", gse27562$characteristics_ch1)
> gse27562_samples = sampleNames(gse27562[,ind1])
> ind2 = grep("Cancer", gse16443$Disease)
> gse16443_samples = sampleNames(gse16443[,ind2])
>
> blood_samples = c(gse16443_samples, gse27562_samples)
>
> idx = mgse_COMBAT$Disease == 'Control Tissue'
> tissue_samples = sampleNames(mgse_COMBAT[,idx])
>
> samples = c(blood_samples, tissue_samples)
>
> c4.data = mgse_COMBAT[,samples]
> pData(c4.data)$Disease <- factor(c4.data$Disease)
```

After the selection of samples their distribution is as follows:

	GSE16443	GSE18672	GSE22820	GSE27562	GSE9574
Cancer	67	0	0	57	0
Control Tissue	0	79	10	0	15

We can compare the output (how the different data sets are mixed together) by computing a Multidimensional Scaling (MDS) plot:

```
> plotMDS(c4.data,  "Disease", "Study", main="Distribution of samples in the merged datasets")
```

The plot is shown in Figure 1.

After the merging and selection of samples what is the proportion of normal versus tumor cases?

```
> dim(c4.data);
```

```
Features  Samples
    5358      228
```

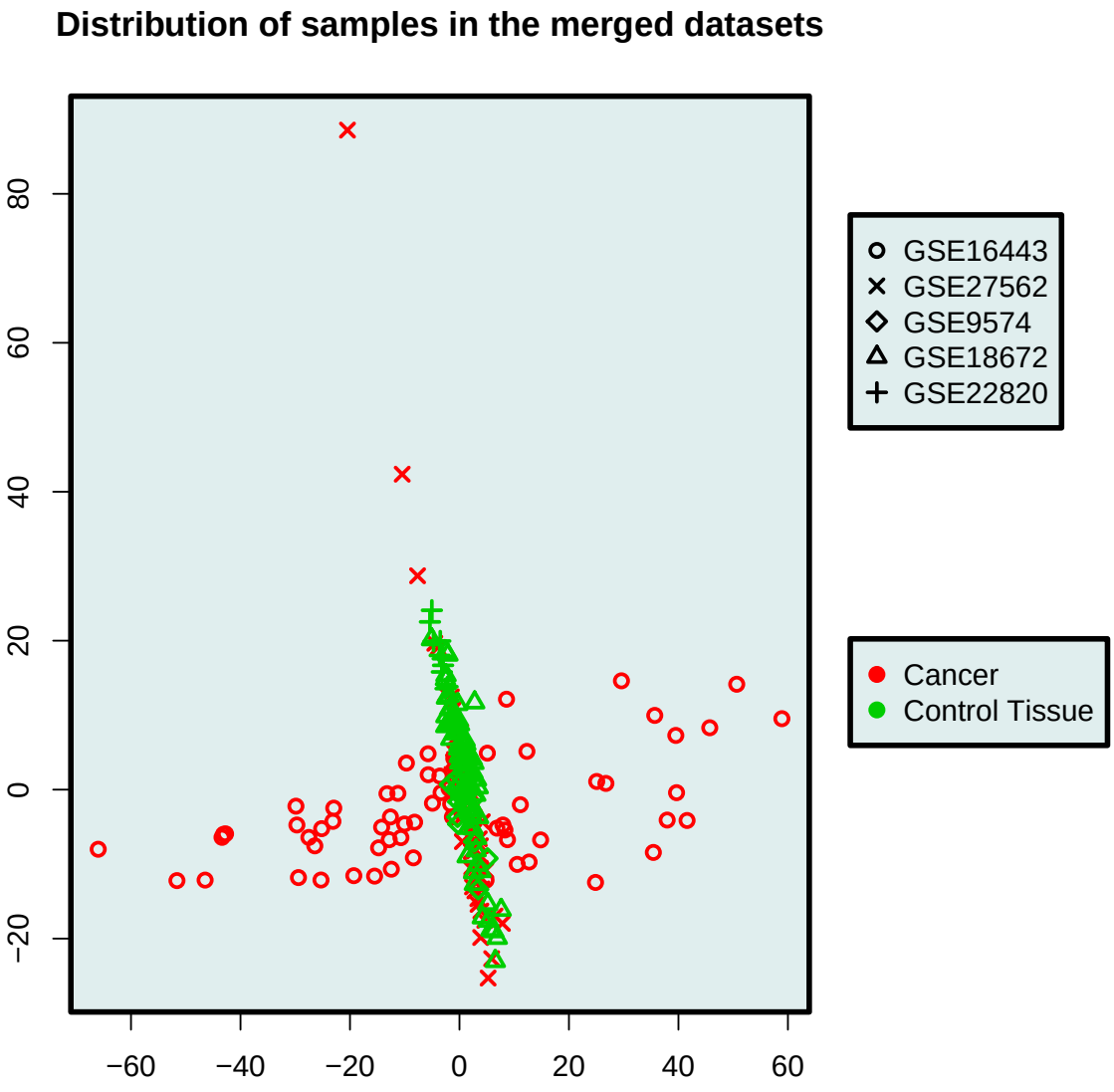

Figure 1: 2D (MDS) plot of the merged dataset

```
> table(c4.data$Disease);


        Cancer Control Tissue
           124            104
```

Let's cache the dataset for subsequent analyses:

```
> saveRDS(c4.data, file=file.path("intermediate", "c4.data.rds"))
```

Gene selection

I am using Significance Analysis of Microarrays (SAM) for finding differential expressed genes. We try to select the differentially expressed genes so that the estimated False Discovery Rate (FDR) to be bounded by `0.05` (i.e. to expect 5% falsely assumed differentially expressed genes):

```
> cl = ifelse(c4.data$Disease == 'Control Tissue', 0, 1)
> sam.out <- sam(exprs(c4.data), cl, B=1000, rand=0xDEAD);
> dd <- findDelta(sam.out, fdr=0.05, verbose=FALSE)[2,1]

The threshold seems to be at
  Delta Called     FDR
5 1.183   1615 0.05004
6 1.183   1614 0.04977

> sam.sum <- summary(sam.out, delta=dd)
> ee=exprs(c4.data)[sam.sum@row.sig.genes,]
> num.diff.exp = dim(ee)[1]
```

So by choosing 1.1827 as our "delta" we find 1614 differential expressed genes:

```
> plot(sam.out, dd)
```

Let's find out which of the genes are over-expressed in cancer samples:

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

[1] 1076

> siggenes.all <- list.siggenes(sam.out, dd)
> siggenes.over <- list.siggenes(sam.out, dd)[w]
```

So there are 1076 genes over-expressed in cancer state. They can be found in the `c4_siggenes.txt` file.

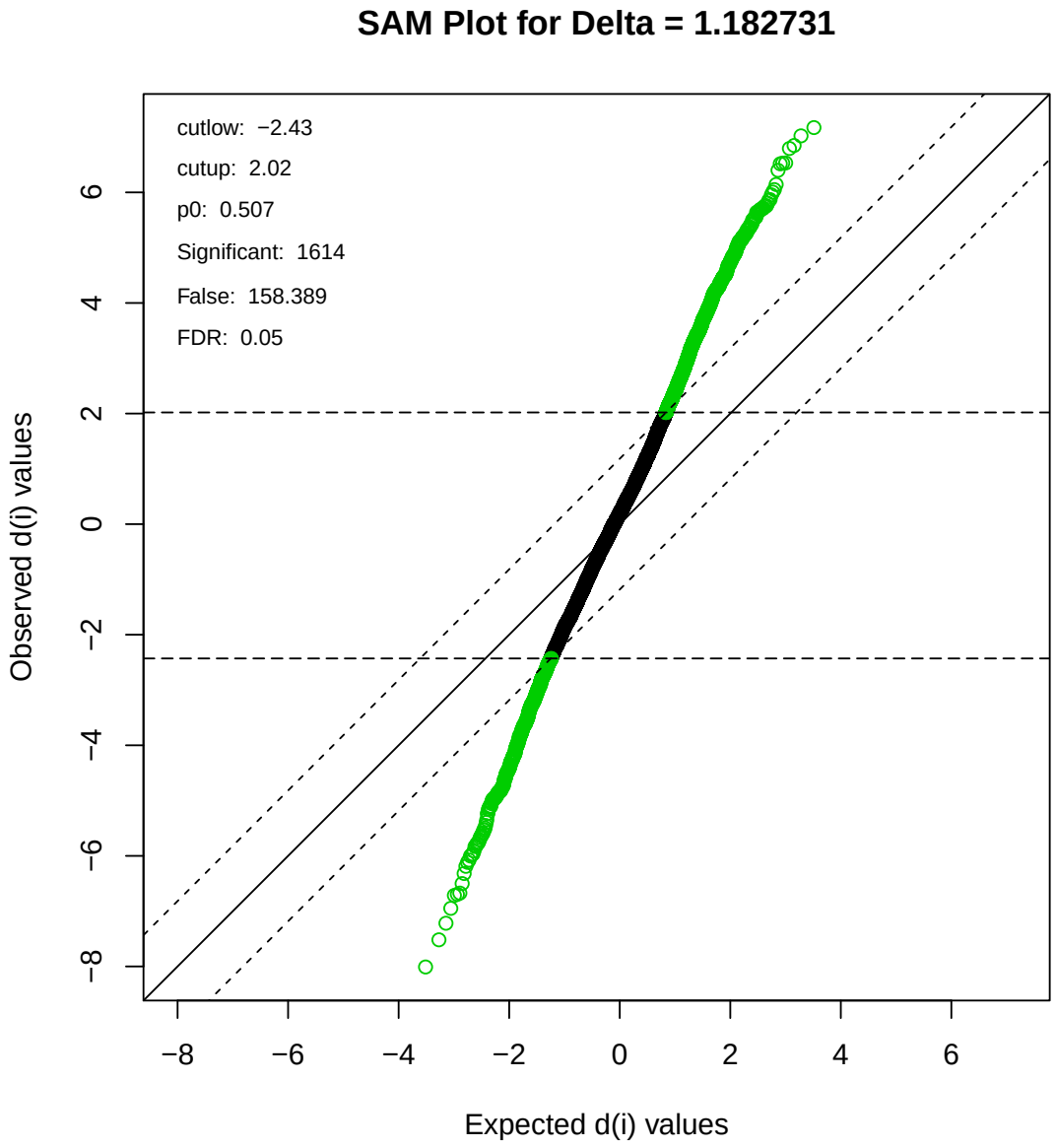

Figure 2: SAM plot

Common genes with C1, C2

```
> ez_to_genes <- function(ez) {
+     require(org.Hs.eg.db)
+     cl <- mget(as.character(ez), org.Hs.egSYMBOL)
+
+     m <- sapply(cl, function(x) x[[1]])
+     df <- data.frame(ezgene=names(m), symbol=m)
+     df
+ }
>
> c1 <- read.table("c1_siggenes.txt",col.names="g")$g
> c2 <- read.table("c2_siggenes.txt",col.names="g")$g
> c3 <- read.table("c3_siggenes.txt",col.names="g")$g
> c4 <- read.table("c4_siggenes.txt",col.names="g")$g
>
> old.common <- intersect(c1, intersect(c2, c3))
> common <- intersect(c1, intersect(c2, c4))
```

The following are the common genes between C1, C2, and C4:

```
> kable(ez_to_genes(common), row.names=FALSE)
```

ezgene	symbol
2553	GABPB1
7529	YWHAB
5052	PRDX1
54968	TMEM70
1032	CDKN2D
4800	NFYA
56681	SAR1A
9991	PTBP3
7305	TYROBP
3151	HMGN2
6636	SNRPF
10972	TMED10
3840	KPNA4
55062	WIPI1
7852	CXCR4
6431	SRSF6
3726	JUNB
10221	TRIB1
79132	DHX58
9935	MAFB

ezgene	symbol
2585	GALK2
2739	GLO1
51398	WDR83OS
5573	PRKAR1A
3192	HNRNPU
8678	BECN1

The following are the "new" genes not contained in the old intersection of C1, C2, C3:

ezgene	symbol
6636	SNRPF
5573	PRKAR1A
3192	HNRNPU

while in the table below are the genes in the old intersection of C1, C2, C3 that are removed from the new intersection:

ezgene	symbol
3692	EIF6

Session information

```
> sessionInfo()

R version 3.1.1 (2014-07-10)
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   grDevices datasets  parallel  stats     graphics  utils
[8] methods   base

other attached packages:
 [1] plyr_1.8.1            ggplot2_1.0.0         xtable_1.7-3
 [4] inSilicoMerging_1.8.6 DWD_0.11              Matrix_1.1-4
 [7] siggenes_1.38.0       multtest_2.20.0       knitr_1.6
```

```
[10] Biobase_2.24.0        BiocGenerics_0.10.0   magrittr_1.1.0

loaded via a namespace (and not attached):
 [1] codetools_0.2-9  colorspace_1.2-4 digest_0.6.4     evaluate_0.5.5
 [5] formatR_0.10     grid_3.1.1       gtable_0.1.2     lattice_0.20-29
 [9] MASS_7.3-34      munsell_0.4.2    proto_0.3-10     Rcpp_0.11.2
[13] reshape2_1.4     scales_0.2.4     stats4_3.1.1     stringr_0.6.2
[17] survival_2.37-7  tools_3.1.1
```