Supplementary information for

Detect tissue heterogeneity in gene expression data with *BioQC*

Jitao David Zhang, Klas Hatje, Clemens Broger, Martin Ebeling, and Laura Badi

June 24, 2016

Abstract

In this vignette, we demonstrate the use of *BioQC* with a case study where mouse kidney samples were profiled for gene expression. Results of BioQC pointed to potential tissue heterogeneity caused by pancreas contamination which was confirmed by qRT-PCR experiments. Source code and data needed to reproduce this document can be found at https://github.com/Accio/BioQC-example.

This is a supplementary documentation of *BioQC*, a R/Bioconductor package used to detect tissue heterogeneity from high-throughput gene expression profiling data with tissue-specific gene signatures. For its basic use please refer to the documentation and vignettes shipped along with the package, or to the other vignette *bioqc-simulation.Rnw* which applies the algorithm to simulated datsets. Here we demonstrate its use with a real biological data set, which is not included in the package distribution due to size limitations.

1 Importing data

First we load the package, the tissue-specific gene signatures, and the expression data into the R session.

```
> library(BioQC)
> gmtFile <- system.file("extdata/exp.tissuemark.affy.roche.symbols.gmt",
+                        package="BioQC")
> gmt <- readGmt(gmtFile)
> file <- "bioqc-nephrectomy.RData"
> load(file)
> print(eset)
```

```
ExpressionSet (storageMode: lockedEnvironment)
assayData: 34719 features, 25 samples
  element names: exprs
protocolData: none
phenoData
  sampleNames: FSP5.FVB.NJ.Sham.Placebo FS2.FVB.NJ.Sham.Control ...
```

```
    FN6.FVB.NJ.Nephrectomy.Control (25 total)
  varLabels: Experiment.name INDIVIDUALNAME ... Elastase (7 total)
  varMetadata: labelDescription
featureData
  featureNames: 1415670_at 1415671_at ... AFFX-TransRecMur/X57349_M_at
    (34719 total)
  fvarLabels: GeneID GeneSymbol OrigGeneID OrigGeneSymbol
  fvarMetadata: labelDescription
experimentData: use 'experimentData(object)'
Annotation:
```

The dataset contains expression of 34719 genes in 25 samples. The expression profile was normalized with RMA normalization. The signals were also `log2`-transformed; however, this step does not affect the result of *BioQC* since it is essentially a non-parametric statistical test.

2 Running BioQC

Next we run the core function of the *BioQC* package, `wmwTest`, to perform the analysis.

```
> system.time(bioqcRes <- wmwTest(eset, gmt,
+                                 alternative="greater"))
```

```
   user  system elapsed
  1.615   0.005   1.620
```

The function returns *one-sided* p-values of Wilcoxon-Mann-Whitney test. We next visualize this metric after transformation.

```
> bioqcResFil <- filterPmat(bioqcRes, 1E-8)
> bioqcAbsLogRes <- absLog10p(bioqcResFil)
```

By closer examination (e.g. using heatmaps such as the one shown in Fig 1), we found expression of pancreas and adipose specific genes is significantly enriched in samples 23-25.

```
> library(RColorBrewer)
> heatmap(bioqcAbsLogRes, Colv=NA, Rowv=TRUE,
+         cexRow=0.85, margin=c(3,12),
+         col=rev(brewer.pal(7, "RdBu")),
+         labCol=1:ncol(bioqcAbsLogRes))

> filRes <- bioqcAbsLogRes[c("Kidney_NGS_RNASEQATLAS_0.6_3",
+                            "Pancreas_Islets_NR_0.7_3"),]
> matplot(t(filRes), pch=c("K", "P"), type="b", lty=1L,
+         ylab="BioQC score", xlab="Sample index")
```

Visual inspection (Figure 2) reveals that there might be contaminations in samples 23-25, potentially by pancreas and adipose tissues.

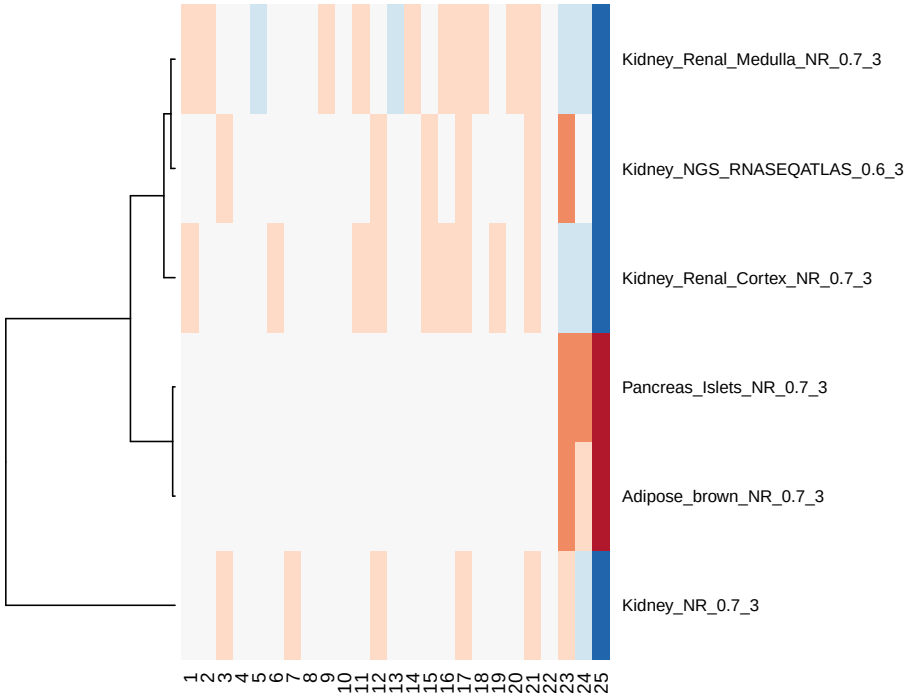

Figure 1: BioQC scores (defined as $abs(log10(p))$) of the samples visualized in heatmap. Red and blue indicate high and low scores, respectively.

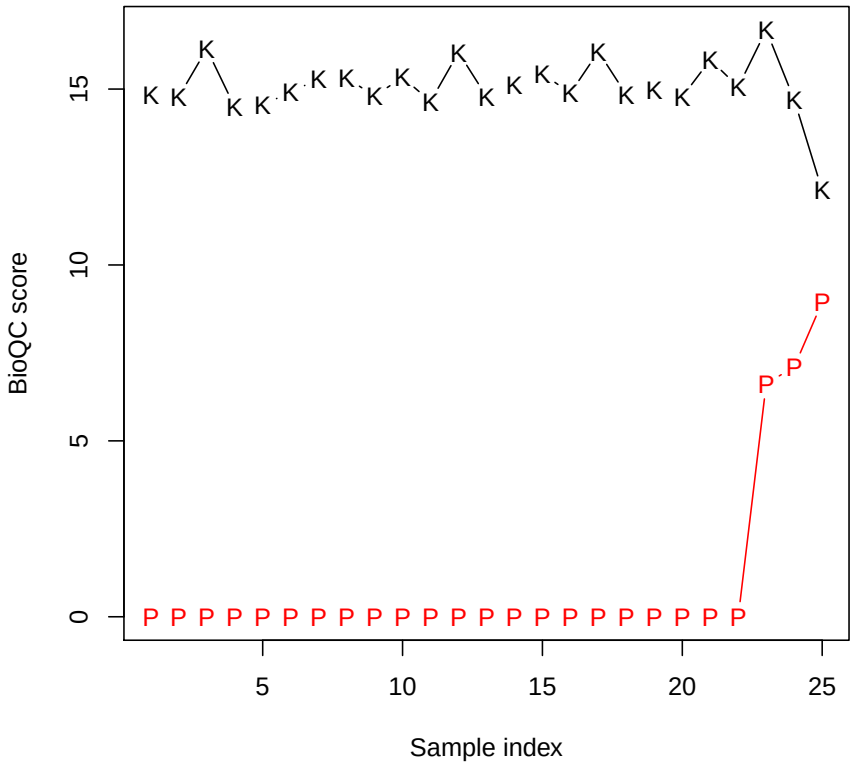

Figure 2: BioQC scores (defined as $abs(log10(p))$) of the samples. K and P represent kidney and pancreas signature scores, respectively.

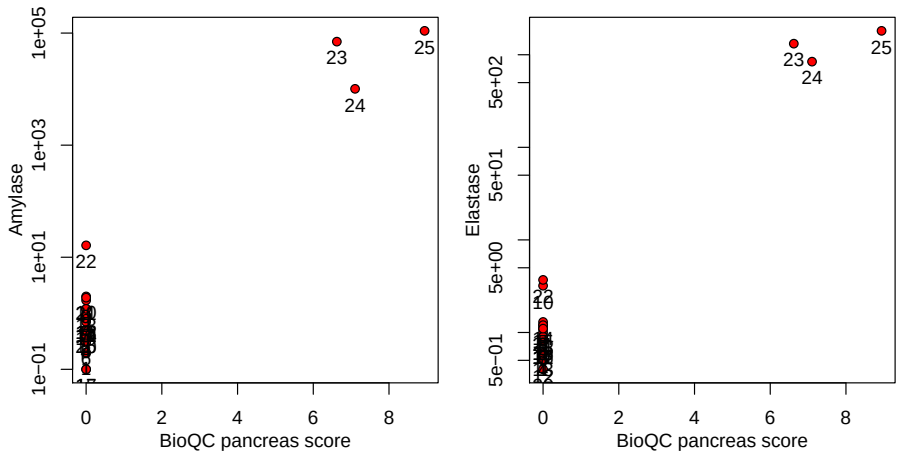

Figure 3: Correlation between qRT-PCR results and BioQC pancreas score

3 Validation with quantitative RT-PCR

To confirm the hypothesis generated by *BioQC*, we performed qRT-PCR experiments to test two pancreas-specific genes' expression in the same set of samples. Note that the two genes (amylase and elastase) are not included in the signature set provided by BioQC.

```
> amylase <- eset$Amylase
> elastase <- eset$Elastase
> pancreasScore <- bioqcAbsLogRes["Pancreas_Islets_NR_0.7_3",]
> par(mfrow=c(1,2), mar=c(3,3,1,1), mgp=c(2,1,0))
> plot(amylase~pancreasScore, log="y", pch=21, bg="red",
+      xlab="BioQC pancreas score", ylab="Amylase")
> text(pancreasScore,amylase, 1:ncol(eset), pos=1)
> plot(elastase~pancreasScore, log="y", pch=21, bg="red",
+      xlab="BioQC pancreas score", ylab="Elastase")
> text(pancreasScore,elastase, 1:ncol(eset), pos=1)
```

The results are shown in Figure 4. It seems likely that sample 23-25 are contaminated by nearby pancreas tissues when the kidney was dissected. Potential contamination by adipose tissues remains to be tested.

4 Impact of sample removal on differential gene expression analysis

In this study, four mice of the *FVB/NJ* strain received nephrectomy operation and treatment of Losartan, an angiotensin II receptor antagonist drug, and four mice reveiced an sham operation and Losartan. Among the Nephrectomy+Losartan group, one sample (index 24) is possibly contaminated by pancreas. Suppose now we are interested in the differential gene expression between the conditions. We now run the analysis twice, once with and once without the contaminated sample, to study the impact of removing heterogenous samples detected by *BioQC*.

```
> library(limma)
> isNeph <- with(pData(eset), Strain=="FVB/NJ" & TREATMENTNAME %in% c("Nephrectomy-Losartan", "Sham-
> isContam <- with(pData(eset), INDIVIDUALNAME %in% c("BN7", "FNL8", "FN6"))
> esetNephContam <- eset[,isNeph]
> esetNephExclContam <- eset[, isNeph & !isContam]
> getDEG <- function(eset) {
+   group <- factor(eset$TREATMENTNAME, levels=c("Sham-Losartan","Nephrectomy-Losartan"))
+   design <- model.matrix(~group)
+   colnames(design) <- c("ShamLo", "NephLo")
+   contrast <- makeContrasts(contrasts="NephLo", levels=design)
+   exprs(eset) <- normalizeBetweenArrays(log2(exprs(eset)))
+   fit <- lmFit(eset, design=design)
+   fit <- contrasts.fit(fit, contrast)
+   fit <- eBayes(fit)
+   tt <- topTable(fit, n=nrow(eset))
+   return(tt)
+ }
> esetNephContam.topTable <- getDEG(esetNephContam)
> esetNephExclContam.topTable <- getDEG(esetNephExclContam)
> esetFeats <- featureNames(eset)
> esetNephTbl <- data.frame(feature=esetFeats,
+                           GeneSymbol=esetNephContam.topTable[esetFeats,]$GeneSymbol,
+                           OrigGeneSymbol=esetNephContam.topTable[esetFeats,]$OrigGeneSymbol,
+                           Contam.logFC=esetNephContam.topTable[esetFeats,]$logFC,
+                           ExclContam.logFC=esetNephExclContam.topTable[esetFeats,]$logFC)
> par(mfrow=c(1,1), mar=c(3,3,1,1)+0.5, mgp=c(2,1,0))
> with(esetNephTbl, smoothScatter(Contam.logFC~ExclContam.logFC,
+                                 xlab="Excluding one contaminating sample [logFC]",
+                                 ylab="Including one contaminating sample [logFC]"))
> abline(0,1)
> isDiff <- with(esetNephTbl, abs(Contam.logFC-ExclContam.logFC)>=2)
> with(esetNephTbl, points(Contam.logFC[isDiff]~ExclContam.logFC[isDiff], pch=16, col="red"))
```

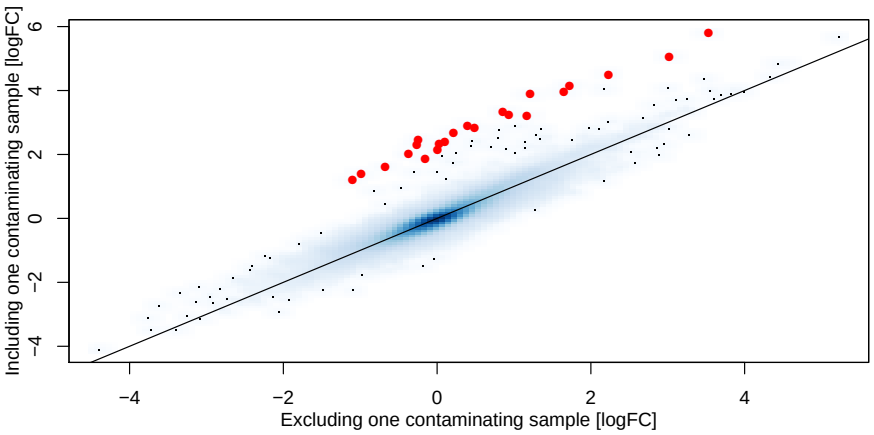

Figure 4: Log2 fold change (logFC) values reported by *limma* with one contaminated sample included (y-axis) or excluded (x-axis). Genes strongly affected by the contamination are indicated by red dots.

```
> diffTable <- esetNephTbl[isDiff,]
> diffGenes <- unique(diffTable[,"GeneSymbol"])
> diffGenesPancreas <- diffGenes %in% gmt[["Pancreas_Islets_NR_0.7_3"]]$genes
```

We found that 22 probesets mapped to 17 are associated with very strong expression changes if the contaminated sample is not excluded (Table 4). Not surprisingly 13 genes among them are highly enriched in pancreas but not kidney and belong to the pancreas signature used by *BioQC*.

In summary, we observe that tissue heterogeneity can impact down-stream analysis results and negatively affect reproducibility of gene expression data if it remains overlooked. It underlines again the value of applying *BioQC* as a first-line quality control tool.

5 Session Info

The script runs within the following session:

```
R version 3.3.0 (2016-05-03)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Red Hat Enterprise Linux Server release 6.3 (Santiago)

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
```

feature	GeneSymbol	OrigGeneSymbol	Contam.logFC	ExclContam.logFC
1415805_at	CLPS	Clps	1.87	-0.15
1415883_a_at	CELA3B	Cela3b	1.61	-0.68
1415905_at	REG1A	Reg1	2.46	-0.24
1416139_at	REG1B	Reg2	2.82	0.49
1417257_at	CEL	Cel	4.16	1.72
1417413_at	CUZD1	Cuzd1	1.39	-0.99
1418287_a_at	DMBT1	Dmbt1	3.34	0.85
1421868_a_at	PNLIP	Pnlip	5.05	3.02
1422434_a_at	PRSS1	2210010C04Rik	4.50	2.23
1422435_at	PRSS1	2210010C04Rik	3.23	0.93
1428062_at	CPA1	Cpa1	2.91	0.40
1428102_at	CPB1	Cpb1	3.95	1.65
1428358_at	ZG16	Zg16	1.20	-1.10
1428359_s_at	ZG16	Zg16	2.38	0.10
1433431_at	PNLIP	Pnlip	3.21	1.17
1437015_x_at	PLA2G1B	Pla2g1b	2.14	0.01
1437438_x_at	PNLIPRP2	Pnliprp2	2.30	-0.26
1438612_a_at	CLPS	Clps	3.88	1.21
1448186_at	PNLIPRP2	Pnliprp2	2.03	-0.37
1448220_at	CTRB1	Ctrb1	5.82	3.53
1451228_a_at	SYCN	Sycn	2.67	0.21
1454623_at	CPA2	Cpa2	2.32	0.03

Table 1: Genes that are identified as strongly changed ony if the contaminated sample is included.

```
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

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

other attached packages:
[1] xtable_1.8-2        limma_3.28.5        RColorBrewer_1.1-2
[4] BioQC_1.0.0         Biobase_2.32.0      BiocGenerics_0.18.0
[7] Rcpp_0.12.5

loaded via a namespace (and not attached):
[1] tools_3.3.0        KernSmooth_2.23-15
```