Supplementary information for

Detect tissue heterogeneity from high-throughput gene expression data with BioQC

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

March 10, 2016

Abstract

In this vignette, we first demonstrate the use of *BioQC* with a biological dataset, where mouse kidney samples were profiled for gene expression. Results of BioQC pointed to potential tissue heterogeneity caused by pancreas, which was confirmed by qRT-PCR. Furthermore, we describe data preprocessing and analysis procedures used to generate tissue-specific gene signatures. Finally we discuss the applicability of the *BioQC* algorithm. Both the source code and the data to produce 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 Case study with a kidney expression profiling dataset

1.1 Data read in

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 non-parametric.

1.2 Run 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
  1.577   0.028   1.604
```

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))
```

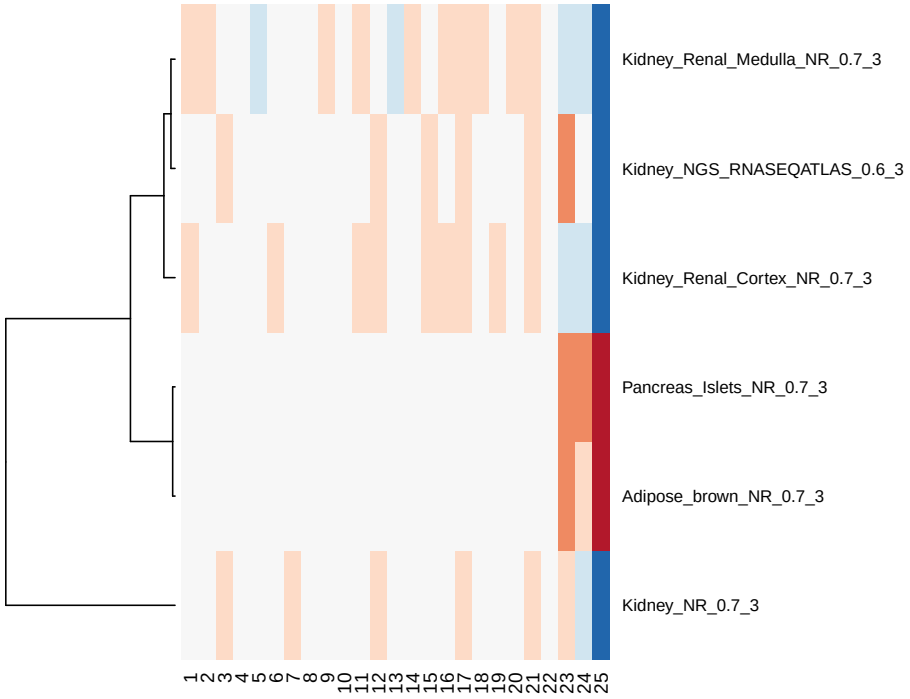

Figure 1: BioQC scores (defined as $abs(log10(p))$) of the samples visualized in heatmap. Red and blue indicate high and low scores, respectively.

```
> 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.

1.3 Biological validation

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 3. 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.

2 Methods

Data preprocessing of public datasets to generate tissue-specific gene signatures

Two microarray datasets (*NB* and *GNF*) were downloaded from the Gene Expression Omnibus database of NCBI. Chips of poor quality were removed. Raw signals were normalised with the MAS5 method.

The GTEx dataset was downloaded from the GTEx database (http://www.broadinstitute.org/gtex/). Gene expression levels were measured in RPKMs (reads per kilobase per million mapped reads).

In either case, per-tissue gene expression was estimated by averaging replicates of each tissue. Gini indices were calculated from the resulting matrix.

Currently, *BioQC* provides 155 sets of tissue-specific gene signatures.

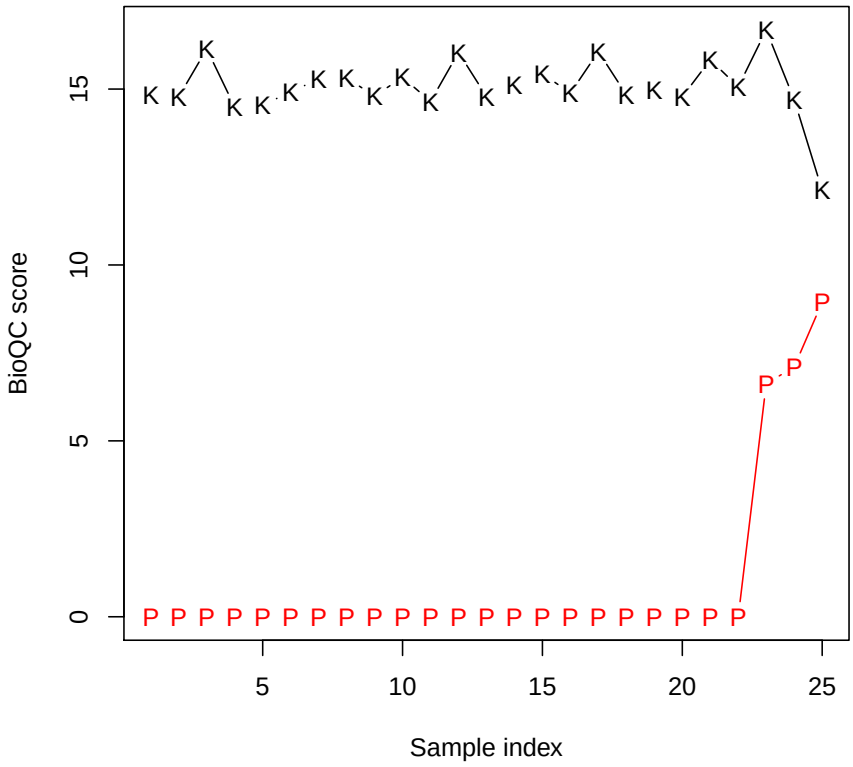

Figure 2: BioQC scores (defined as $abs(log10(p))$) of the samples. K and P represent kidney and pancreas signature scores, respectively.

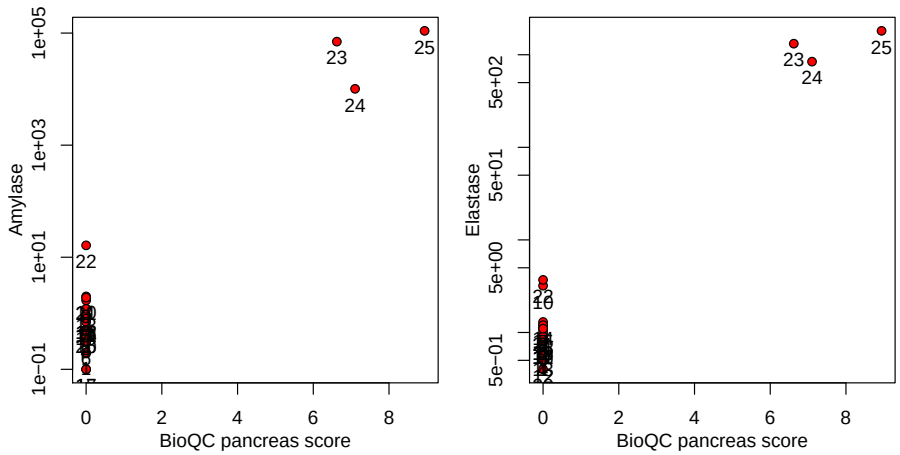

Figure 3: Correlation between qRT-PCR results and BioQC pancreas score

3 Discussions

We developed *BioQC* as a simple method to detect tissue heterogeneity in large-scale gene expression profiling datasets.

BioQC is conceptually related to GSEA (Subramanian *et al.*, 2005) and ssGSEA (Barbie *et al.*, 2009). Different from GSEA, BioQC can run on single samples since no between-sample statistics is required. BioQC differs from the ssGSEA algorithm in three aspects. First, BioQC applies Wilcoxon-Mann-Whitney tests instead of Kolmogorov-Smirnov tests that are used by ssGSEA, because we are interested in the difference of the locations between the empirical distribution of signature genes expression and the distribution of background genes expression, rather than the difference of shapes. At the same time, BioQC comes with a comprehensive list of normal human tissue signatures that can be used 'out of the box'.

BioQC implements Wilcoxon-Mann-Whitney test to generate a score for each tissue and each sample. The statistical test makes strong assumptions about the underlying data, such as (1) independence between the genes, (2) gene expression read-outs (from platforms such as microarray and NGS) preserves the ranks of the real expression level, and (3) tissue contamination leads to consistent increase of tissue signature genes identified from public large-scale expression datasets such as GTex. We are well aware that these assumptions are often violated in reality. However, we have observed good empirical performance of BioQC in real-life data. Confirming our experience, a recent study found that ssGSEA performed slightly inferior to but fairly comparable with GSVA, a GSEA variant that does not make the same assumption (Hanzelmann *et al.*, 2013). In clinical settings ssGSEA has also provided interesting insights into molecular mechanisms of diseases (Verhaak *et al.*, 2013). Therefore, while we notice the theoretical limitations, we believe BioQC can be useful for quality-control purposes, by generating hypotheses and triggering further analysis to identify tissue heterogeneity.

We note that BioQC is not a stringent statistical test for the purpose of hypothesis testing; instead, it provides a simple method, which if used and interpreted properly, can detect potential tissue contamination issues in large-scale expression datasets and provide hints of the source of contamination. It provides an alternative to commonly used unsupervised statistical methods such as principal component analysis.

Our experience suggests that there are a surprisingly non-trivial proportion of samples in public databases that are potentially contaminated by tissues that are not declared in the meta data. If not taken into consideration, they are likely to distort any downstream analysis results. Therefore, we sincerely hope to see tools such as (and better than) BioQC that are simple to use yet powerful enough to detect tissue contamination in expression profiling data.

4 Acknowledgment

We want to thank Guido Steiner, Gonzalo Durán Pacheco, and many other colleagues for trying out and providing feedbacks to improve the package, Laurent Essioux for discussions, Detlef Wolf for technical support, and Silvia Ines Pomposiello for expression and PCR data.

References

Barbie D, *et al.* (2009) Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1, *Nature*, **462**, 108–112.

Hanzelmann S, *et al.* (2013) GSVA: gene set variation analysis for microarray and RNA-Seq data, *BMC Bioinformatics*, **14**, 7.

Ma J, *et al.* (2011) Appearance frequency modulated gene set enrichment testing, *BMC Bioinformatics*, **12**, 81.

Subramanian A, *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles, *PNAS*, **102**, 15545–15550.

Verhaak R, *et al.* (2013) Prognostically relevant gene signatures of high-grade serous ovarian carcinoma, *J Clin Invest*, **123**, 517–525.

5 Session Info

The script runs within the following session:

```
R version 3.1.3 (2015-03-09)
Platform: x86_64-unknown-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
 [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] RColorBrewer_1.1-2  BioQC_0.99.4        Biobase_2.26.0
[4] BiocGenerics_0.12.1 Rcpp_0.12.0

loaded via a namespace (and not attached):
[1] tools_3.1.3
```