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 perform simulations with both model-generated and real-world data using *BioQC*. We show that *BioQC* is a fast and sensitive method to detect tissue heterogeneity from high-throughput gene expression data. The source code to produce this document can be found at `https://github.com/Accio/BioQC-example`.

Contents

BioQC is a R/Bioconductor package to detect tissue heterogeneity from high-throughput gene expression profiling data. It implements an efficient Wilcoxon-Mann-Whitney test, and offers tissue-specific gene signatures that are ready to use 'out of the box'.

1 Experiment setup

In this document, we perform three simulation studies with *BioQC*:

- **Time benchmark** tests the time-efficiency of the Wilcoxon test implemented in *BioQC*, compared with the native implementation in *R*;

- **Sensitivity benchmark** tests the sensitivity and specificity of *BioQC* detecting tissue heterogeneity using model-generated simulated data;

- **Mixing benchmark** tests the sensitivity and specificity of *BioQC* using simulated contamination with real-world data.

All source code that is needed to reproduce the results can be found in the *Rnw* file generating this document.

2 Time benchmark

In the first experiment, we setup expression matrices of 20155 human protein-coding genes of 1, 5, 10, 50, or 100 samples. Genes are *i.i.d* distributed following $\mathcal{N}(0,1)$. The Wilcoxon-Mann-Whitney test implemented in BioQC and the native R implementation are applied to the matrices respectively.

The numeric results of both implementations, *bioqcNumRes* (from BioQC) and *rNumRes* (from R), are equivalent, as shown by the next command.

```
> print(expect_equal(bioqcNumRes, rNumRes))
```

```
As expected: bioqcNumRes equals rNumRes
```

The *BioQC* implementation is more than 500 times much faster (Figure 1): while it takes about one second for BioQC to calculate enrichment scores of all 155 signatures in 100 samples, the native R implementation about 20 minutes.

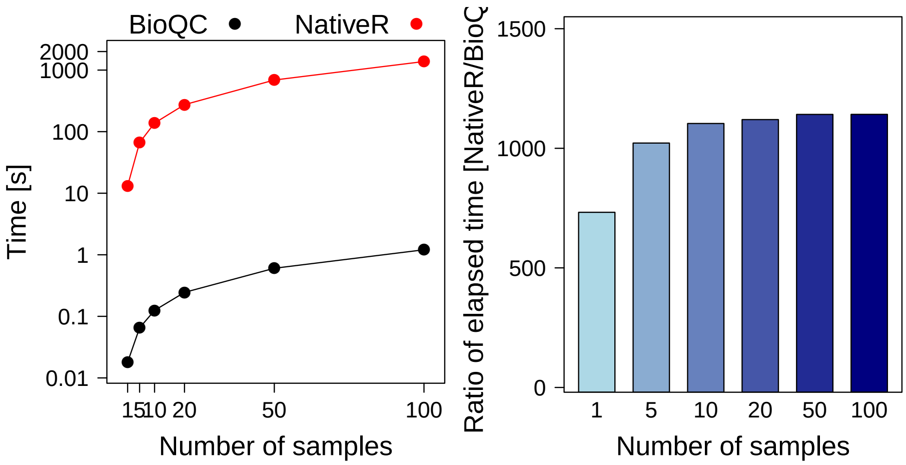

Figure 1: Time benchmark results of BioQC and R implementation of Wilcoxon-Mann-Whitney test. Left panel: elapsed time in seconds (logarithmic Y-axis). Right panel: ratio of elapsed time by two implementations. All results achieved by a single thread on in a RedHat Linux server.

The main reason underlying the low performance of R implementation is that the `wilcox.test` function sorts two numeric vectors that are to be compared. When the function is repeatedly applied to gene expression data, it performs many expensive sorting operations which are futile, because the sort of genes outside of the gene set (*background genes*) does not change between samples. *BioQC* sorts the background genes only once for each gene set, independent of how many samples are tested.

In addition, *BioQC* implements an approximate Wilcoxon test instead of the exact version, because the difference between the two is negligible for high-throughput gene expression data. Last but not least, *BioQC* implements its core algorithm in C so as to maximize the time- and memory-efficiency.

Putting these tweaks together, *BioQC* achieves identical results as the native implementation with two order of magnitude less time. This renders *BioQC* an highly efficient tool for quality control of large-scale high-throughput gene expression data.

3 Sensitivity benchmark

We next asked the question how sensitive is *BioQC* to expression changes of tissue signature genes. Similar to the previous simulation, while keeping all other genes *i.i.d* normally distributed following $\mathcal{N}(0,1)$, we dedicatedly increase the expression of genes in one randomly selected tissue signature (ovary, with 43 genes) by different amplitudes: these genes' expression levels are randomly drawn from different normal distributions with varying expection and constant variance between $\mathcal{N}(0,1)$ and $\mathcal{N}(3,1)$. To test the robustness of the algorithm, 10 samples are generated for each mean expression difference value.

Figure 2 visualizes the distribution of enrichment scores and their ranks dependent on the mean expression difference between ovary signature genes and background genes. As soon as the expression of signature genes increases by a very moderate ampltiude (1σ), BioQC will identify the gene set as the highest-ranking signature. A even stronger difference in expression will lead to higher enrichment scores but no change in the rank.

The results suggest that *BioQC* is sensitive even to moderate changes in the average expression of a gene set.

4 Mixing benchmark

The sensitivity benchmark above suffers from the limitation that the distributions of gene expression are not physiological. To overcome this, we designed and performed a benchmark by *in silico* mixing expression profiles with weighted linear combination, thereby mimicking tissue contamination.

Given the expression profile of a sample of tissue A ($\mathbf{Y}_A$), and that of a sample of tissue B ($\mathbf{Y}_B$), the weighted linear mixing produces a new profile $\mathbf{Y} = \omega\mathbf{Y_A} + (1-\omega)\mathbf{Y_B}$, where $\omega \in [0,1]$. In essence the profiles of two tissue types are linearly mixed in different proportions, which simulates varying severities of contaminations. We asked whether BioQC could detect such mixings, and if so, how sensitive is the method.

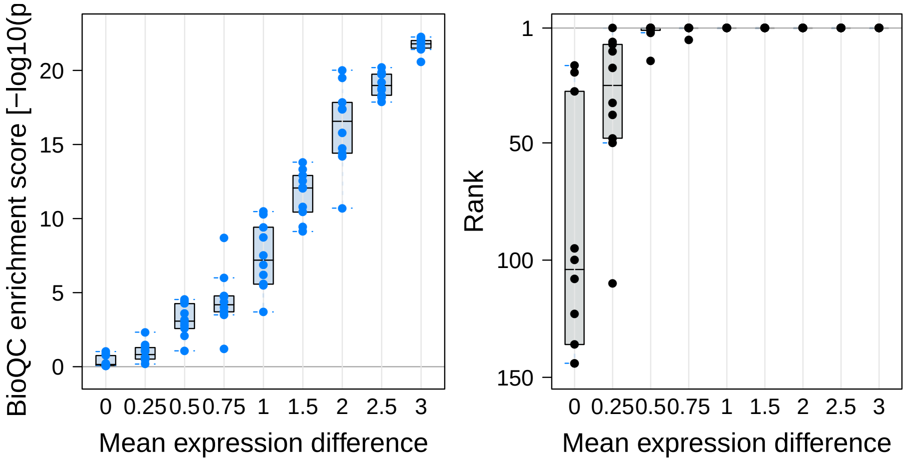

Figure 2: Sensitivity benchmark. Expression levels of genes in the ovary signature are dedicately sampled randomly from normal distributions with different mean values. Left panel: enrichment scores reported by *BioQC* for the ovary signature, plotted against the differences in mean expression values; Right panel: rank of ovary enrichment scores in all 155 signatures plotted against the difference in mean expression values.

4.1 Dataset selection and quality control

In order to avoid over-fitting of signatures derived from human expression data, we decided to use a normal tissue expression dataset from a non-human mammal species, because it has been shown that tissue-preferential expression patterns tend to be conserved between mammal species. We identified a dataset of *Canis lupus familiaris* (dog), which is publicly available in Gene Expression Omnibus (GDS4164).

In this study, the authors examined 39 samples from 10 pathologically normal tissues (liver, kidney, heart, lung, brain, lymph node, spleen, jejunum, pancreas, and skeletal muscle) of four dogs (with one pancreas sample missing). We downloaded the data, and performed minimal preprocessing: for multiple probesets that map to same genes, we kept the one with the highest average expression level and removed the rest. The resulting dataset contained expression of 16797 genes. BioQC was applied to the dataset to test whether there are major contamination issues. The results, including tissues reported by the authors, and the BioQC tissue signatures with the highest and second-highest scores, are reported in Table 1.

	Label	BioQC.best	BioQC.second
1	Brain	Spinal_cord	Nodose_nucleus
2	Brain	Brain_Cortex_prefrontal	Brain_Amygdala
3	Brain	Brain_Cortex_prefrontal	Brain_Amygdala
4	Brain	Brain_Cortex_prefrontal	Brain_Amygdala
5	Heart	Muscle_cardiac	Muscle_skeletal

6	Heart	Muscle_cardiac	Muscle_skeletal
7	Heart	Muscle_cardiac	Muscle_skeletal
8	Heart	Muscle_cardiac	Muscle_skeletal
9	Jejunum	Intestine_small	Intestine_Colon
10	Jejunum	Intestine_small	Intestine_Colon
11	Jejunum	Intestine_small	Intestine_Colon
12	Jejunum	Intestine_small	Intestine_Colon
13	Kidney	Kidney	Kidney_Renal_Cortex
14	Kidney	Kidney	Kidney_Renal_Cortex
15	Kidney	Kidney	Kidney_Renal_Cortex
16	Kidney	Kidney	Kidney_Renal_Cortex
17	Liver	Liver	Liver
18	Liver	Liver	Liver
19	Liver	Liver	Liver
20	Liver	Liver	Liver
21	Lung	Lung	Monocytes
22	Lung	Monocytes	Lung
23	Lung	Lung	Monocytes
24	Lung	Monocytes	Lung
25	LymphNode	Lymphocyte_B_FOLL	Lymphocytes_T_H
26	LymphNode	Lymphocyte_B_FOLL	Lymphocytes_T_H
27	LymphNode	Lymphocyte_B_FOLL	Lymphocytes_T_H
28	LymphNode	Lymphocyte_B_FOLL	Lymphocytes_T_H
29	Pancreas	Pancreas_Islets	Pancreas
30	Pancreas	Pancreas_Islets	Pancreas
31	Pancreas	Pancreas_Islets	Pancreas
32	SkeletalMuscle	Muscle_skeletal	Muscle_cardiac
33	SkeletalMuscle	Muscle_skeletal	Muscle_cardiac
34	SkeletalMuscle	Muscle_skeletal	Muscle_cardiac
35	SkeletalMuscle	Muscle_skeletal	Muscle_cardiac
36	Spleen	Monocytes	Lymphocyte_B_FOLL
37	Spleen	Monocytes	Lymphocyte_B_FOLL
38	Spleen	Monocytes	Erythroid_cells
39	Spleen	Monocytes	Myeloblast

Table 1: Quality control of the mixing benchmark input data with *BioQC*. Four columns (f.l.t.r.): sample index; tissue reported by the authors; the tissue signature with the highest enrichment score reported by *BioQC*; the tissue signature with the second-highest enrichment score.

By comparing the tissue labels provided by the authors and the predictions of *BioQC*, we notice that in most cases the two match well (despite of ontological differences). In three cases (sample 1, 22, and 24) though there seem to be intriguing differences, which might be explained by different sampling procedures or immune cell infiltration. We will however in this vignette not further explore them. These three samples are removed from the simulation procedures.

4.2 An example of weighted mixing: heart and jejunum

As an example, we take average expression of heart and jejunum samples, and mix them by different compositions.

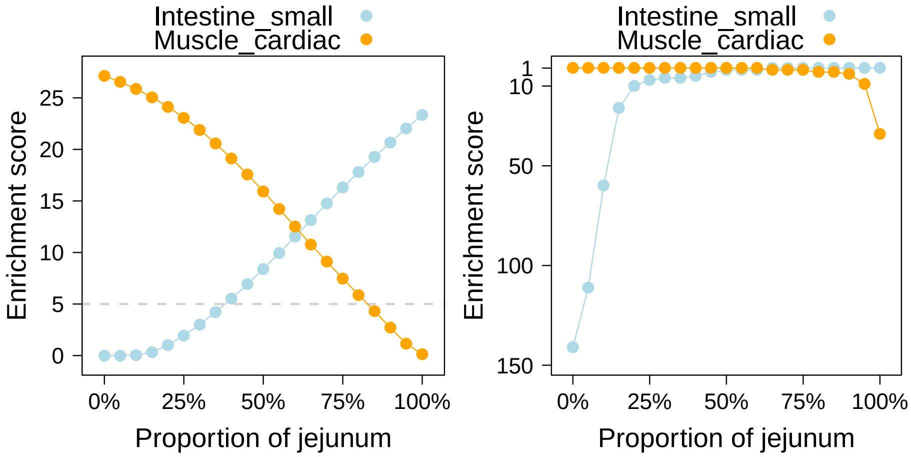

Figure 3: Results of a mixing case study. Left panel: *BioQC* enrichment scores of small intestine and cardiac muscle varying upon different proportions of jejunum; Right panel: ranks of enrichment scores varying upon different proportions of jejunum.

Figure 3 allows us comparing enrichment scores and their ranks when the expression profiles of heart and jejunum are mixed *in silico*. We observe that with as little as 5% contamination of heart tissue in jejunum samples (rightmost in the right panel), the rank of heart signature jumps from 34 to 9; 10% and 20% contamination will further enhance the rank to 4 and 3 respectively. If we start from the other end, namely assuming jejunum contamination in heart samples, the BioQC algorithms ranks jejunum the 7th only when there are more than 25% contamination. If we set enrichment score equal or over 5 as the threshold of calling a suspected contamination event, it takes about 20% heart in jejunum tissue or about 35% jejunum tissue in heart to make a call. It means the sensitivity of contamination detection is not symmetric between tissues: contamination by tissues with distinct expression patterns (such as heart) are easier to be detected than contamination by tissues with less distinct expression patterns (such as small intestine).

While it is difficult to quantify the absolute sensitivity of contamination detection, it is apparent that if the enrichment score of a unforeseen tissue is very high (or ranked high), one may suspect potential contamination. Also, if there are replicates of samples from the same tissue, a higher value in one sample compared with the other samples suggests a contamination or infiltration incident.

4.3 Pairwise mixing

Following the heart-jejunum example, we performed all 45 pairwise mixing experiments, producing weighted linear combinations of gene expression profiles of each pair of tissues (excluding self-mixing). The results are summaried in a heatmap in Figure 4.

The heatmap visualization summarizes the detection limit of contamination of each pair of tissues. Take the cell in row 1 column 2 from top left: its value (0.20) means that if there are 20% or more contamination by heart in the brain sample, *BioQC* will be able to detect it, because the enrichment score is equal to or larger than 5, or the heart tissue signature ranks in the top 3 of all tissue signatures.

Take another cell in row 2 column 1 from top left: its value (0.75) means that if there are 75% or more contanmination by brain in a heart sample, *BioQC* will be able to detect it. Here we observe the asymmetry again that we observed before with the heart/jejenum example: while it is relative easy to identify heart contamination of a brain sample, it is more difficult to identify brain contamination of a heart sample in this dataset.

Interestingly, brain samples are a special case: if they contaminate other tissues, it is more difficult to identify (but not other way around). We suspect this is due to the fact that many genes are specifically expressed in brain but with relative low levels; therefore although the prefrontal cortex signatures could correctly predict the brain sample in the quality control (Table 1), they are less powerful detecting brain-causing contaminations. It remains to be tested whether this phenomenon is observed in other datasets and in species other than dog.

Otherwise, most *in silico* contamination events are detectable in this dataset, with median detection limit around 30%. This suggests that *BioQC* works well in physiological settings.

5 Conclusions

Benchmark studies with similated and real-world data demonstrate that *BioQC* is a efficient and sensitive method to detect tissue heterogeneity from high-throughput gene expression data.

6 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
```

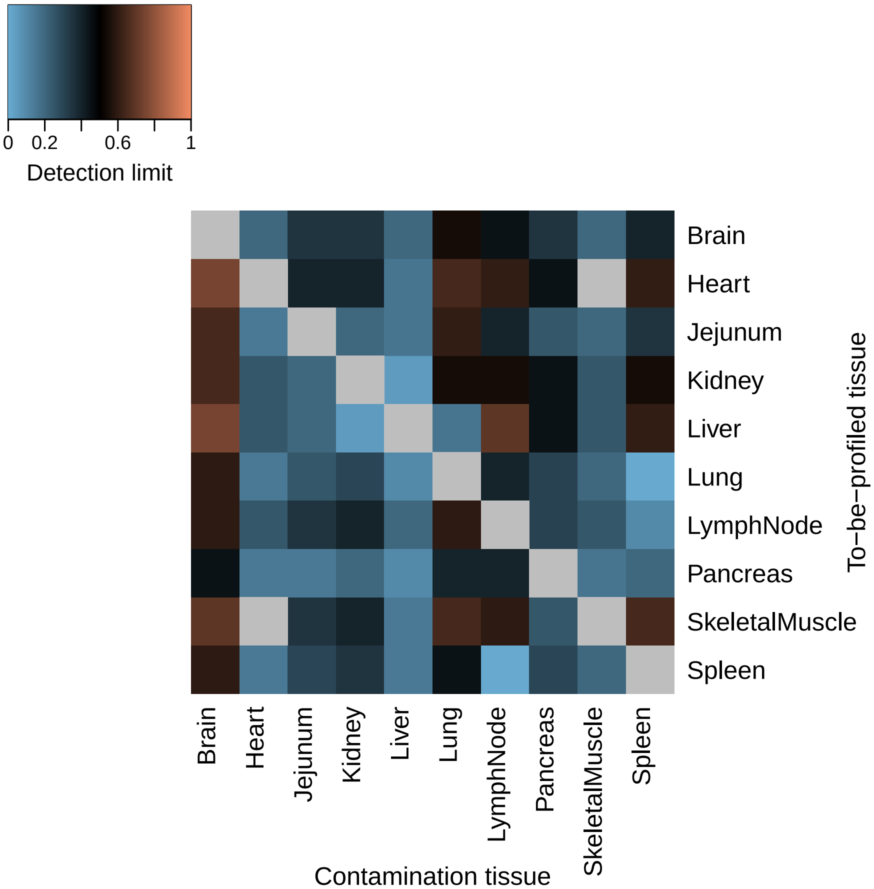

Figure 4: Results of the pairwise mixing experiment. Each cell represents the minimal percentage of tissue of the column as contamination in the tissue of the row that can be detected by *BioQC*. No values are available for cells on the diagonal because self-mixing was excluded. Heart and skeletal muscle are very close to each other and therefore their detection limit is not considered.

```
 [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] stats4    parallel  stats     graphics  grDevices utils     datasets
[8] methods   base

other attached packages:
 [1] gplots_2.17.0        xtable_1.8-2         GEOquery_2.32.0
 [4] latticeExtra_0.6-28  RColorBrewer_1.1-2   lattice_0.20-33
 [7] hgu133plus2.db_3.0.0 org.Hs.eg.db_3.0.0   RSQLite_1.0.0
[10] DBI_0.3.1            AnnotationDbi_1.28.2 GenomeInfoDb_1.2.5
[13] IRanges_2.0.1        S4Vectors_0.4.0      BioQC_0.99.4
[16] Biobase_2.26.0       BiocGenerics_0.12.1  Rcpp_0.12.0
[19] testthat_0.11.0

loaded via a namespace (and not attached):
 [1] bitops_1.0-6       caTools_1.17.1     crayon_1.3.1       digest_0.6.9
 [5] gdata_2.17.0       grid_3.1.3         gtools_3.5.0       KernSmooth_2.23-15
 [9] memoise_1.0.0      RCurl_1.95-4.8     tools_3.1.3        XML_3.98-1.3
```

7 Appendix

Comparing BioQC with Principal Component Analysis (PCA)

In the context of the dog transcriptome dataset, we can compare the results of principal component analysis (PCA, Figure 5) with that of *BioQC*.

PCA sugggests that samples from each tissue tend to cluster together, in line with the *BioQC* results. In addition, PCA reveals that tissues with cells of similar origins cluster together, such as skeletal muscle and heart. As expected, one brain sample and two lung samples seem to be different from other samples of the same cluster, which are consistent with the *BioQC* findings; however, unlike BioQC, PCA does not provide information on what are potential contamination/infiltration casues.

Therefore, we believe *BioQC* complements existing unsupervised methods to inspect quality of gene expression data.

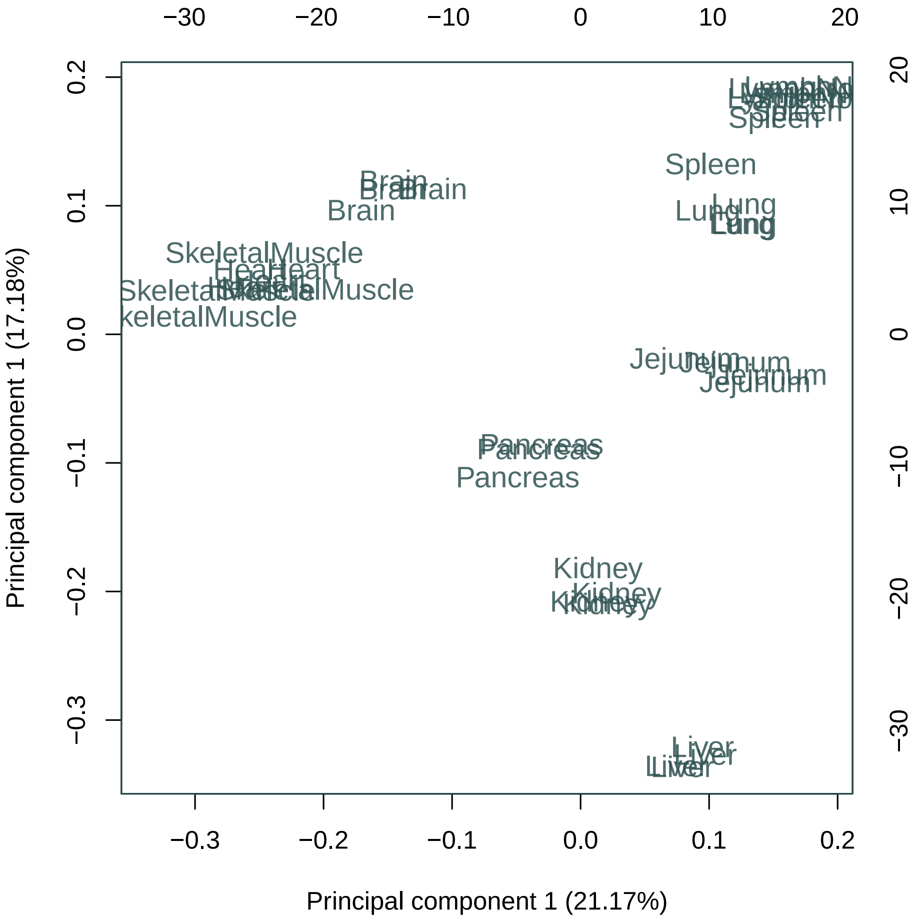

Figure 5: Sample separation revealed by principal component analysis (PCA) of the dog transcriptome dataset.