

---

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

---

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

June 25, 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

1	Experiment setup	1
2	Time benchmark	2
3	Sensitivity benchmark	3
4	Mixing benchmark	4
5	Conclusions	9
6	Session Info	10
7	Appendix	10

*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 20534 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.

```
> expect_equal(bioqcNumRes, 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.

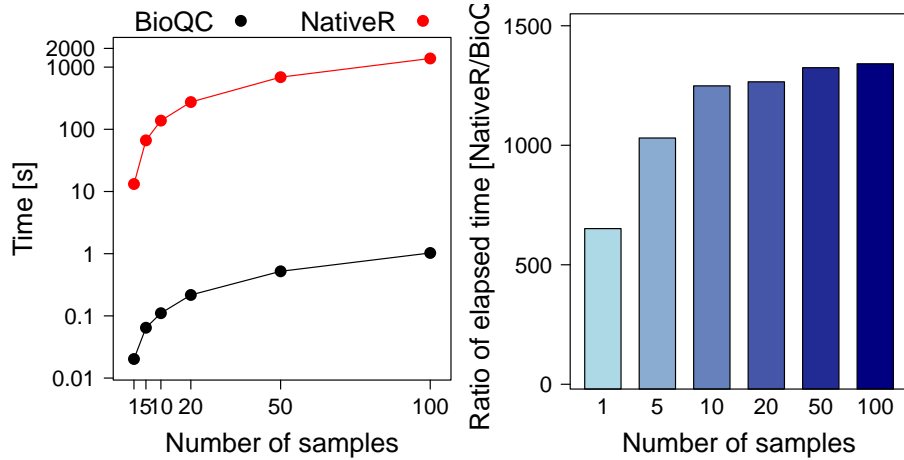


Figure 1: Time benchmark results of BioQC and R implementation of Wilcoxon-Mann-Whitney test. Left panel: elapsed time in seconds (logarithmic Y-axis). Right panel: ratio of elapsed time by two implementations. All results achieved by a single thread on in a RedHat Linux server.

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 expectation 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.

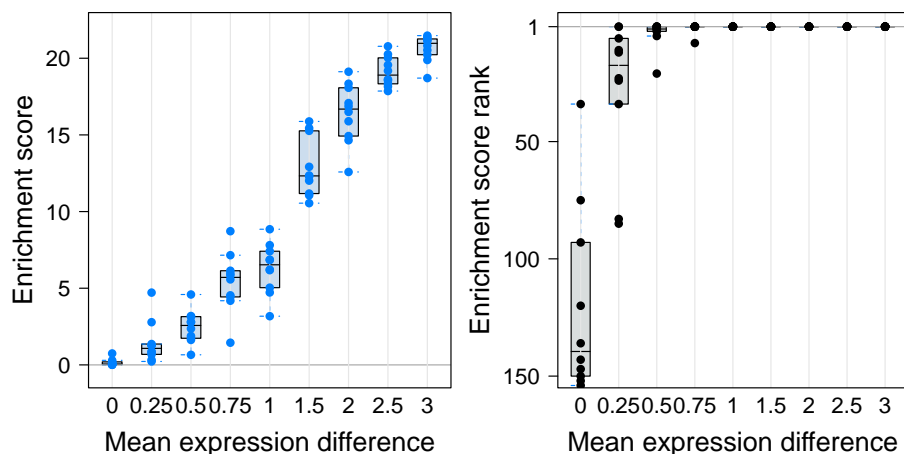


Figure 2: Sensitivity benchmark. Expression levels of genes in the ovary signature are dedicatedly sampled randomly from normal distributions with different mean values. Left panel: enrichment scores reported by *BioQC* for the ovary signature, plotted against the differences in mean expression values;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

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 amplitude ( $1\sigma$ ), 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.

### 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 pre-processing: 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
GSM502573	Brain	Spinal_cord	Nodose_nucleus
GSM502574	Brain	Brain_Cortex_prefrontal	Brain_Amygdala
GSM502575	Brain	Brain_Cortex_prefrontal	Brain_Amygdala
GSM502576	Brain	Brain_Cortex_prefrontal	Brain_Amygdala
GSM502577	Heart	Muscle_cardiac	Muscle_skeletal
GSM502578	Heart	Muscle_cardiac	Muscle_skeletal

GSM502579	Heart	Muscle_cardiac	Muscle_skeletal
GSM502580	Heart	Muscle_cardiac	Muscle_skeletal
GSM502581	Jejunum	Intestine_small	Intestine_Colon
GSM502582	Jejunum	Intestine_small	Intestine_Colon
GSM502583	Jejunum	Intestine_small	Intestine_Colon
GSM502584	Jejunum	Intestine_small	Intestine_Colon
GSM502585	Kidney	Kidney	Kidney_Renal_Cortex
GSM502586	Kidney	Kidney	Kidney_Renal_Cortex
GSM502587	Kidney	Kidney	Kidney_Renal_Cortex
GSM502588	Kidney	Kidney	Kidney_Renal_Cortex
GSM502589	Liver	Liver	Liver
GSM502590	Liver	Liver	Liver
GSM502591	Liver	Liver	Liver
GSM502592	Liver	Liver	Liver
GSM502593	Lung	Lung	Monocytes
GSM502594	Lung	Monocytes	Lung
GSM502595	Lung	Lung	Monocytes
GSM502596	Lung	Monocytes	Lung
GSM502597	LymphNode	Lymphocyte_B_FOLL	Lymphocytes_T_H
GSM502598	LymphNode	Lymphocyte_B_FOLL	Lymphocytes_T_H
GSM502599	LymphNode	Lymphocyte_B_FOLL	Lymphocytes_T_H
GSM502600	LymphNode	Lymphocyte_B_FOLL	Lymphocytes_T_H
GSM502601	Pancreas	Pancreas_Islets	Pancreas
GSM502602	Pancreas	Pancreas_Islets	Pancreas
GSM502603	Pancreas	Pancreas_Islets	Pancreas
GSM502604	SkeletalMuscle	Muscle_skeletal	Muscle_cardiac
GSM502605	SkeletalMuscle	Muscle_skeletal	Muscle_cardiac
GSM502606	SkeletalMuscle	Muscle_skeletal	Muscle_cardiac
GSM502607	SkeletalMuscle	Muscle_skeletal	Muscle_cardiac
GSM502608	Spleen	Monocytes	Lymphocyte_B_FOLL
GSM502609	Spleen	Monocytes	Lymphocyte_B_FOLL
GSM502610	Spleen	Monocytes	Erythroid_cells
GSM502611	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 ID GSM502573, GSM502594, and GSM502596) 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.

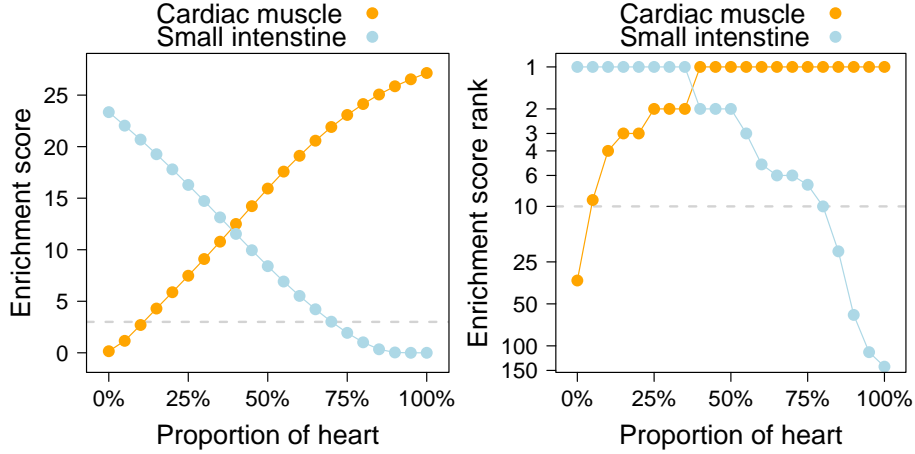


Figure 3: Results of a mixing case study. Left panel: *BioQC* enrichment scores of small intestine and cardiac muscle varying upon different proportions of jejunum; Right panel: ranks of enrichment scores varying upon different proportions of jejunum.

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 3 as the threshold of calling a suspected contamination event ( $p < 0.001$  in the one-sided Wilcoxon-Mann-Whitney test), it takes about 10% heart in jejunum tissue or about 30% 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-jejenum 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 summarized 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 (0.15) means that if there are 15% or more contamination by heart in the brain sample, *BioQC* will be able to detect it (with the threshold enrichment score  $\geq 3$  or the rank  $\leq 10$ ), because the enrichment score is equal to or larger than 3, 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 (0.5) means that if there are 50% or more contamination 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.

The average detection limits of tissues as contamination sources are listed in Table 2. The values are derived from median values of each column in the Figure 4 except for diagonal and missing elements.

	Tissue	MedianDetectionLimit
1	Brain	52.22%
2	Heart	10.62%
3	Jejunum	19.44%
4	Kidney	23.75%
5	Liver	11.87%
6	Lung	31.67%
7	LymphNode	30.63%
8	Pancreas	17.22%
9	SkeletalMuscle	15.00%
10	Spleen	13.57%

Table 2: Median lower detection limites of tissues as contamination sources.

Interestingly, brain samples are a special case: if they contaminate other tissues, it is more difficult to identify (but not other way around). It can be partially explained by the experiment

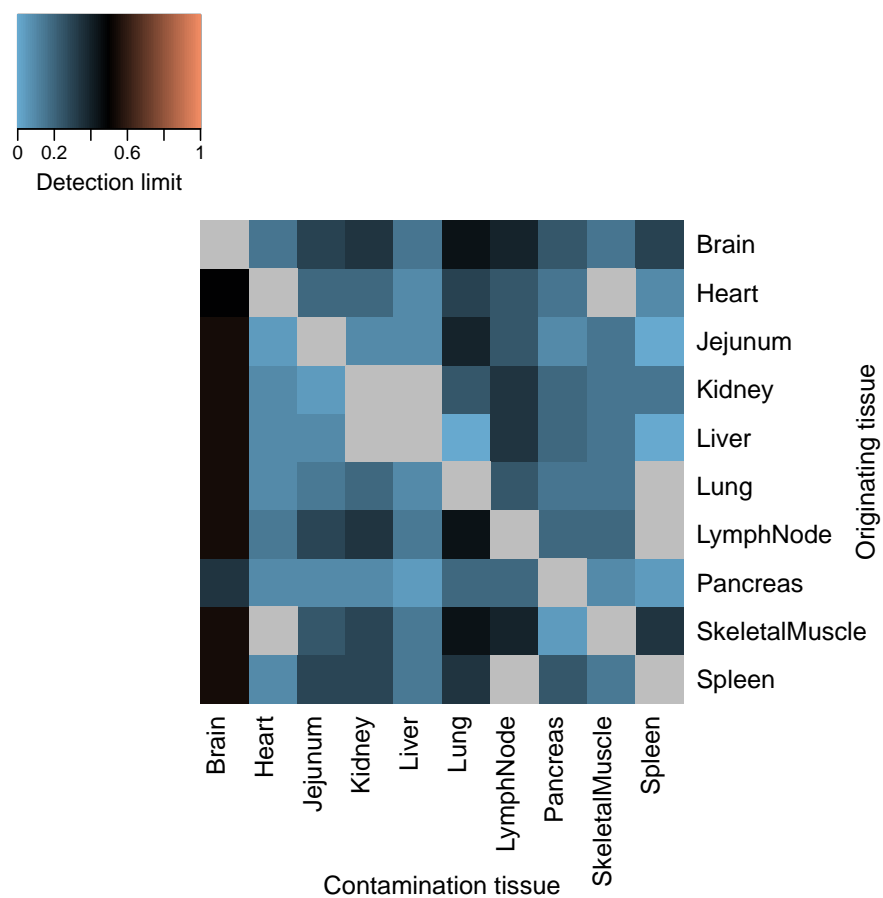


Figure 4: Results of the pairwise mixing experiment. Each cell represents the minimal percentage of tissue of the column as contamination in the tissue of the row that can be detected by *BioQC*. No values are available for cells on the diagonal because self-mixing was excluded. Heart and skeletal muscle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



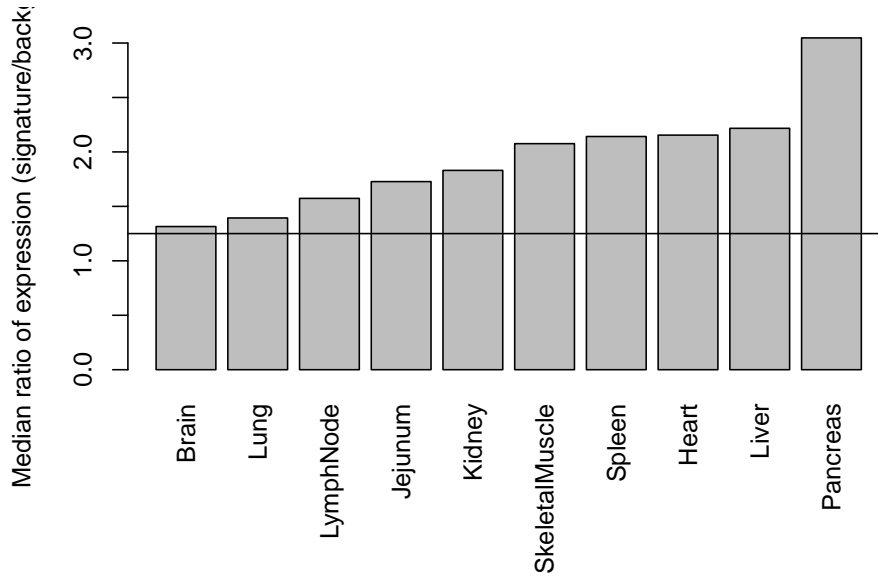


Figure 5: Tissue-specific genes’ expression in respective average tissue profiles. For each tissue (*e.g.* brain), we calculate the median ratio of gene expression level of specific genes over the median expression level of background genes. The value reflects the specificity of tissue-specific genes in respective tissues. Likely due to the sampling of different brain regions, the score of brain ranks the lowest.

design: Briggs *et al.* profiled the whole cerebrum, whereas in *BioQC* there are 22 distinct gene sets assigned to distinct brain regions. Though the prefrontal cortex signature scored highest in the canine brain samples, its score is relative low (7.45), and the genes in the signature are not too far away from the background genes (Figure 5). Therefore only a strong contamination by brain in this dataset will be detected by the given threshold. We expect that if prefrontal cortex instead of cerebrum sample was profiled, the mixing profile of brain will be similar to other organs. This needs to be tested in other datasets.

Otherwise, most *in silico* contamination events are detectable in this dataset, with median detection limit around 0.2. This suggests that *BioQC* is sensitive towards tissue heterogeneity in physiological settings.

## 5 Conclusions

Benchmark studies with simulated 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 Under development (unstable) (2016-06-24 r70830)

Platform: x86\_64-pc-linux-gnu (64-bit)

Running under: Ubuntu 14.04.4 LTS

locale:

```
[1] LC_CTYPE=de_DE.UTF-8      LC_NUMERIC=C
[3] LC_TIME=de_CH.UTF-8      LC_COLLATE=de_DE.UTF-8
[5] LC_MONETARY=de_CH.UTF-8  LC_MESSAGES=de_DE.UTF-8
[7] LC_PAPER=de_CH.UTF-8     LC_NAME=C
[9] LC_ADDRESS=C             LC_TELEPHONE=C
[11] LC_MEASUREMENT=de_CH.UTF-8 LC_IDENTIFICATION=C
```

attached base packages:

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

other attached packages:

```
[1] rbenchmark_1.0.0      gplots_3.0.1          xtable_1.8-2
[4] GEOquery_2.38.4       latticeExtra_0.6-28   RColorBrewer_1.1-2
[7] lattice_0.20-33       hgu133plus2.db_3.2.3  org.Hs.eg.db_3.3.0
[10] AnnotationDbi_1.34.3  IRanges_2.6.1         S4Vectors_0.10.1
[13] BioQC_1.0.0           Biobase_2.32.0        BiocGenerics_0.18.0
[16] Rcpp_0.12.5           testthat_1.0.2
```

loaded via a namespace (and not attached):

```
[1] magrittr_1.5          R6_2.1.2              httr_1.2.0            caTools_1.17.1
[5] tools_3.4.0           grid_3.4.0            KernSmooth_2.23-15    DBI_0.4-1
[9] gtools_3.5.0          digest_0.6.9          crayon_1.3.1          bitops_1.0-6
[13] RCurl_1.95-4.8        memoise_1.0.0         RSQLite_1.0.0         gdata_2.17.0
[17] XML_3.98-1.4
```

## 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 6) with that of *BioQC*.

PCA suggests 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

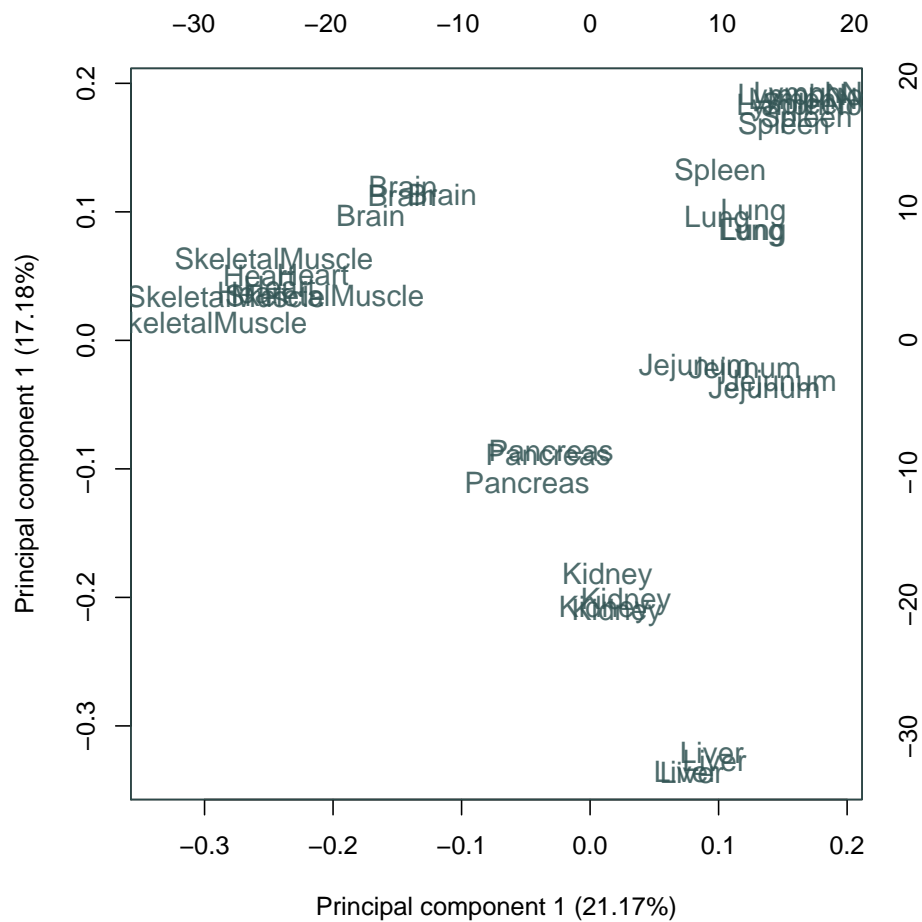


Figure 6: Sample separation revealed by principal component analysis (PCA) of the dog transcriptome dataset.

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.