博士論文

Application of Silhouette Scores for Arbitrarily Defined Groups in Gene Expression Data

（遺伝子発現データにおける任意定義群に対するシルエットスコアの応用）

趙　世涛

Table of Contents

Table of Figure

Table List

Table of Addition files

Chapter 1 Introduction

1.1 Transcriptomics analysis

The central dogma of molecular biology outlines the flow of information that is stored in genes as DNA, transcribed into RNA, and finally translated into proteins [1]. Understanding the transcriptome is essential for interpreting the functional elements of the genome and revealing the molecular constituents of cells and tissues, and also for understanding development and disease [2]. Quantifying the expression levels of each transcript during development and under different conditions is a basic task in transcriptomics.

In the field of transcriptomics analysis, there are two key contemporary techniques: microarrays, which quantify a set of predetermined sequences, and RNA-seq, which capture all sequences using high-throughput sequencing. The emergence of two methods enable researchers to simultaneously interrogate tens of thousands of transcripts in a cell at the same time. This ability has led to important advances in a wide range of biological research fields, including the identification of differentially expressed genes (DEGs) between diseased and healthy tissues, new insights into developmental processes, pharmacogenomic responses, and the evolution of gene regulation in different species [3]–[6].

1.2 Microarray

Since the first use of DNA microarrays for gene expression analysis which explained how the expression of many genes could be monitored in parallel in 1995 by Schena et al. [7], it has been the choice for large-scale studies of gene expression. A microarray consists of

a solid surface on which strands of polynucleotide have been attached or synthesized in fixed positions called spots. A microarray may contain thousands of spots and each spot may contain a few million copies of identical DNA molecules that uniquely correspond to a gene. Two types of expression microarrays are the most popular between users; cDNA microarrays and oligonucleotide chips.

The experiment steps involved in a microarray experiment including: First, RNA is extracted from the cells. Next, RNA molecules in the extract are reverse transcribed into cDNA by using an enzyme reverse transcriptase and nucleotides labelled with different fluorescent dyes. Once the samples have been differentially labelled, cDNA sequence in the same sample will hybridize to specific spots on the glass slide containing its complementary sequence. The amount of cDNA bound to a spot is proportional to the level of expression of the gene represented by that probe in sample. To determine the amount of sample hybridized the microarray is illuminated by a laser and scanned at suitable wavelengths to detect the red and green dyes. This fluorescence intensity represented relative expression level for each gene (population of RNA in the two samples) finally be stored as an image.

Normalization plays an important role in microarray data analysis. It is common practice to transform to a logarithmic (usually base 2) scale. The principal motivation for this transformation is to make variation roughly comparable among measures that span several orders of magnitude. The MAS 5.0 software [8]-[9] developed by Affymetrix uses one-step Tukey biweight to combine the probe intensities in log scale to extract the signal from background noise in a single chip. Robust multiarray analysis

(RMA) [10] now is the most widely used preprocessing algorithm for Affymetrix gene expression microarrays. The normalization of RMA requires multiple arrays to be analyzed simultaneously. The ability to borrow information across samples provides RMA various advantages.

Differential expression analysis is the first field that involved high dimensional statistics methods since the introduction of microarray technologies. Numerous methods have been developed based on the various assumptions of data distribution or model selection. A comparative review of all methods has already been done elsewhere [11]-[12]. Among the methods, there are two most representative and popular models, the SAM method [13], a popular non-parametric approach, and the limma [14] method, a parametric approach using linear models and empirical Bayes.

1.3 RNA-seq

In the last few years, RNA-seq has clear advantages over microarrays, and became the best choice for genome-wide differential gene expression (DGE) experiments. Compared with microarrays, RNA-seq works well for investigating both known transcripts and exploring new ones and provides larger dynamic range in quantifying gene expression. It has been a routine tool in molecular biology, medicine, agriculture, and ecology research. In the aspects of wet lab experiments, A typical RNA-seq workflow includes [15]:

1. capture of cell or tissue samples of interest
2. isolation of RNA from a biological sample
3. reverse transcription into cDNA
4. sequencing of millions of short cDNA fragments (~200bp)

The whole flow chart is summarized in Figure 1.

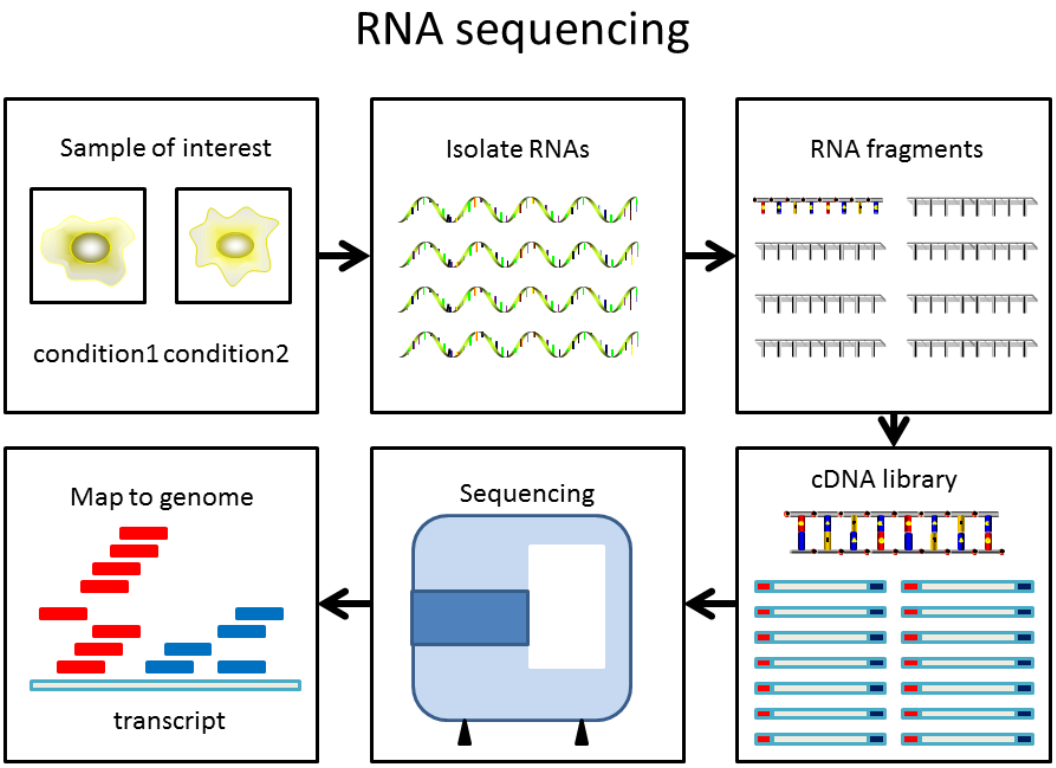


Figure 1 A typical RNA-seq experiment workflow

An overview of the typical RNA-seq pipeline for DGE analysis is outlined in Figure 2. RNA-seq data is stored in FASTQ format files (sequence and base quality). Quality control (QC) of raw data should be performed as the initial step of routine RNA-seq workflow. Tools such as FastQC [16] and Trimmomatic [17] can be applied in this step to assess the quality of raw data. The next computational step of the RNA-seq data analysis pipeline is read mapping: reads are aligned to a reference genome or transcriptome by identifying gene regions that match read sequences. So far, many alignment tools such as tophat [18], STAR [19], bowtie2 [20] and HISAT [21] have been proposed. After mapping, the reads aligned to each coding unit, such as exon, transcript or gene, are used to compute counts, so to give an estimate of its expression level. The most used approach for computing counts considers the total number of reads overlapping the exons of a gene. Two common used tools are featureCount [22] and HTSeq [23].

The 'sequencing depth' of a sample, defined as the total number of sequenced or mapped reads plays a vital role in the design of next generation sequencing experiments. It was revealed that higher sequencing depth generates more informational reads, which increases the statistical power to detect DEGs [24]. Most widely used normalization method called 'TMM' was proposed by Robinson and Oshlack [25] to account for differences in library composition between samples. The method of geometric mean implemented in the R package DESeq [26] is also an effective approaches for library size normalization. 'Reads Per Kilobase of exon model per Million mapped reads' (RPKM), and 'Fragments Per Kilobase of exon per Million fragments mapped' (FPKM) are proposed to reduce both differences in library size and length bias. In recent years, a lot of DGE analysis tools have been developed. Lamarre et al. [27] listed a table about

information on 29 R packages, methods, or pipelines, for DGE analysis of RNA-seq data. For small numbers of replicates as often encountered in RNA-seq count data, the negative binomial (NB) distribution [28] taking account for overdispersion and generalized linear model (GLM) framework [29] are considered to be better choice. Using the Benjamini and Hochberg procedure [30] to adjust the p-value is a routine work in multiple testing.

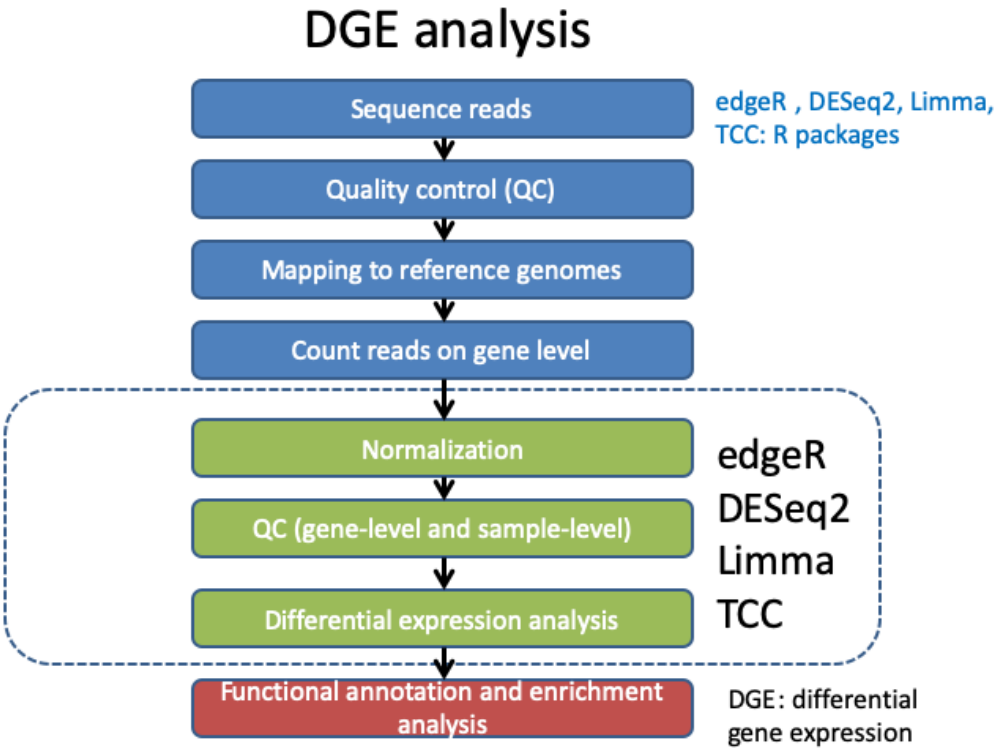


Figure 2 Typical workflow of DGE analysis

1.4 The relationship between HSC and DGE analyses

A common approach for expression analyses is sample clustering (SC) based on similarity in expression patterns [31]-[32]. Utilizing its unsupervised characteristic, SC has been used to (i) detect previously unrecognized subtypes of cancer [33]-[34], (ii) detect outliers (i.e., outlying samples) [35], (iii) represent overall similarities in expression among various organs [36]-[37], and (iv) perform sanity checks to verify expected clustering patterns [38]. When using this approach, researchers can investigate SC results with several possible criteria for grouping (e.g., sex, age, and disease types). However, the evaluation of arbitrary defined groups still counts in subjective visual inspection. Numerical scores indicating the degree of separation between predefined groups would help in the objective assessment of the SC results.

Some researchers empirically know that an SC result of data designed for DGE analysis (say, "DGE data") roughly corresponds to the DGE result when the groups for the DGE analysis are evaluated with respect to the SC result [39]. If individual groups form distinct sub-clusters, where each sub-cluster consists only of members (or samples) in the particular group, DGE analysis using such distinct groups would result in many DEGs. Conversely, if members (or samples) in each sub-cluster originate from multiple groups, no or few DEGs would be expected. However, objective evaluation of the relationship between SC results on DGE data and the percentages of DEGs (P_{DEG}) remains lacking [39].

1.5 Silhouette score

Silhouette score is a graphical aid for the interpretation and validation of cluster analysis [40]. In SC, silhouette score provides a measure of how well a sample is classified when it is assigned to a cluster according to both the tightness of the clusters and the separation between them. Therefore, the silhouette scores are calculated for individual samples. By taking the mean over all samples, the average silhouette (AS) value can be obtained. It ranges from 1.0 to −1.0: a higher (or lower) AS value indicates higher (or lower) degree of separation between clusters. Silhouette score has been successfully used after clustering as a cluster validity measure [31], [41]–[43].

1.6 The purpose of this study

In this study, I propose to use silhouette score for the objective evaluation of gene expression data based on arbitrary grouping criteria. Although they are independent of SC, silhouette scores measuring the degrees of separation between groups of interest would enable a more objective discussion about the SC result in terms of the groups. I here focus on single-factor gene expression data where only one grouping criterion is primarily of interest in relation to the DGE results. I evaluated the relationship among SC results, DGE results, and AS values, using both simulated and real expression data (RNA-seq and microarrays). I found silhouette score (i.e., AS values) to provide a relevant measure for the degrees of separation between groups of interest in SC results. I also found a positive correlation between AS values and DGE results. In the multiple comparison part, it was found to be a universal method that could be adapted to two-group and multiple-group comparison. In both conditions, it can offer promising results

through assessing the degrees of separation between groups of interest to estimate the DGE results.

Chapter 2 Materials and Methods

2.1 Methods

Most of the analyses were performed using R (ver. 3.3.2) [44] and Bioconductor [45]. The versions of major R packages used in the study were TCC ver. 1.14.0 [46], edgeR ver. 3.16.5 [47], ROC ver. 1.50.0, cluster ver. 2.0.5, affy ver. 1.44.0, and RobLoxBioC ver. 0.9.

2.1.1 Hierarchical sample clustering (HSC)

The HSC was performed using the *clusterSample* function with default options ("1 – Spearman's r" as the distance and unique expression patterns as an objective low-count filtering method) in TCC.

2.1.2 DGE analysis pipelines in TCC

The DGE analysis was performed using three functions (*calcNormFactors*, *estimateDE*, and *getResult*) with default options which use functions in TCC. The genes were ranked in ascending order according to p-values. The ranks were used to calculate AUC values when analyzing simulated data. The AUC values were calculated using the *AUC* function in the package ROC. The p-values were adjusted for multiple-testing with the Benjamini–Hochberg procedure. The adjusted p-values (i.e., q-values) were used to obtain the numbers of DEGs satisfying an arbitrarily defined FDR threshold (mainly 10%).

2.1.3 Calculation of Average Silhouette (AS) values

Silhouette score [40] has been successfully employed to estimate the appropriate number of clusters for gene expression data [31], [41]–[43]. Although Silhouette is generally used for the validation of clustering results, I here employ it independently from clustering.

Technically, the term cluster is replaced with group in the silhouette calculation procedure. For each sample *i*, let μ_i be the average distance between *i* and all other samples within the same group (e.g., group A). Let ν_i be the average distance between *i* and the other group (e.g., group B), of which *i* is not a sample member. The silhouette index S_i for sample *i* is calculated as

$$S_i = \frac{(\nu_i - \mu_i)}{\max(\nu_i, \mu_i)} \quad (1)$$

The index S_i ranges from − 1 to 1; it is positive if $\mu_i < \nu_i$, zero if $\mu_i = \nu_i$, and negative if $\mu_i > \nu_i$. A larger S_i value indicates increased group separation and vice versa. By taking the mean S_i over all samples, the average silhouette (AS) value for each comparison can be obtained. The potential applicability of the silhouette unrelated to clustering has been described in the original study [40]. However, to the best of my knowledge, the current study is the first practical application of the concept to estimate the degree of separation between groups (not clusters) using gene expression data. The AS values were calculated using the *silhouette* function in the package cluster.

2.2 Simulated data

The two-group simulated data were produced using the *simulateReadCounts* function in TCC. The variance (V) of the NB distribution can generally be modeled as $V = \mu + \varphi\mu^2$. The empirical distribution of read counts to obtain the mean (μ) and dispersion (φ) parameters of the NB model was obtained from *Arabidopsis* data (three biological replicates (BRs) for both treated and non-treated samples) in [48]. The output of the *simulateReadCounts* function is stored in the TCC class object with information about the simulated conditions and is therefore ready-to-analyze for both the DGE analysis and HSC. The three-group simulated data (n=3) were generated in a similar way as previously described the two-group simulated data.

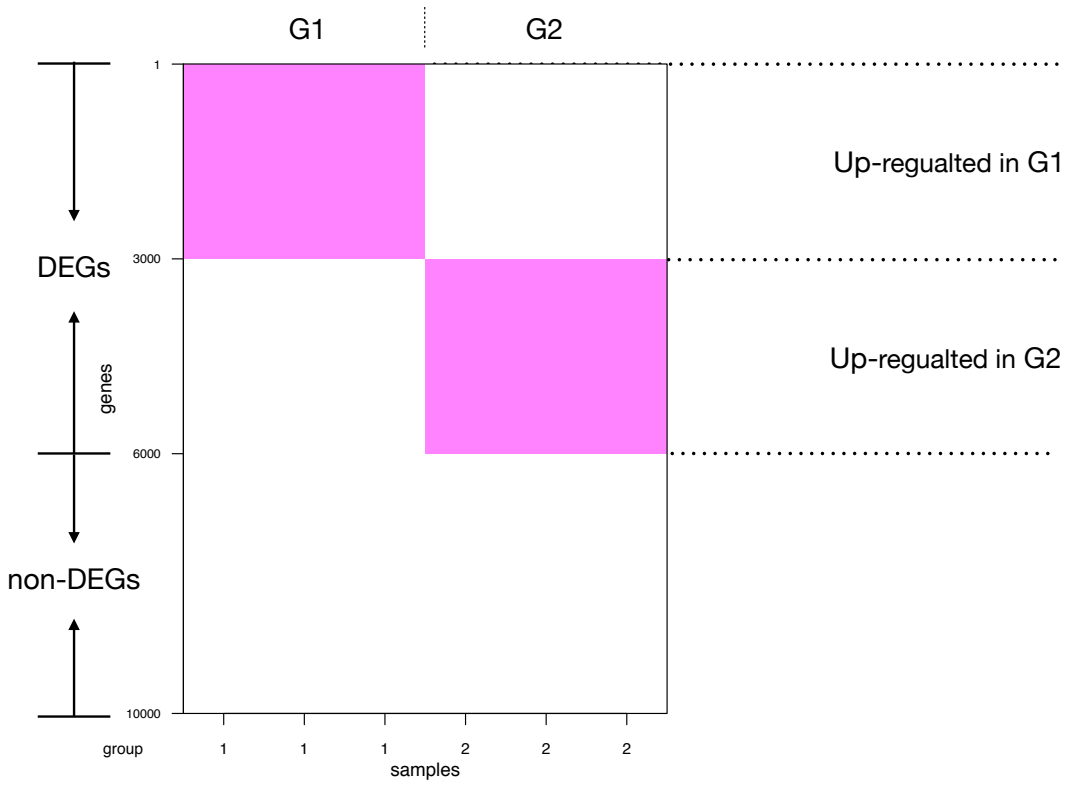


Figure 3 shows an example of generating the two-group count data.

The simulation condition is as follows: the total number of genes is 10,000 (*Ngene* = 10000), the number of replicates is 3 (*Nrep* = 3), 60% of the genes are DEGs (P_{DEG} = 60%), the level of DE is four-fold in individual groups, and the proportions of DEGs up-regulated in individual groups (P_{G1}, P_{G2}) are (0.5, 0.5) which means that there are 3,000 and 3,000 up-regulated genes in G1 and G2, respectively.

2.3 Real datasets

2.3.1 Blekhman's mammalian data (RNA-seq)

Blekhman's mammalian data were obtained from the supplementary website (http://genome.cshlp.org/content/suppl/2009/12/16/gr.099226.109.DC1/suppTable1.xls) [49]. The raw count matrix consisting of 20,689 genes × 36 samples (= 3 species × 2 sexes × 3 BRs × 2 technical replicates) was collapsed by summing the data for technical replicates, giving a reduced number of columns in the matrix (i.e., 18 samples; 3 species × 2 sexes × 3 BRs).

2.3.2 Schurch's yeast data (RNA-seq)

Schurch's yeast data were obtained from the GitHub website (https://github.com/bartongroup/profDGE48/tree/master/Preprocessed_data) [50]. After merging the count vectors for a total of 96 samples, data from 10 outlying samples (WT_rep21, WT_rep22, WT_rep25, WT_rep28, WT_rep34, WT_rep36, Snf2_rep06, Snf2_rep13, Snf2_rep25, and Snf2_rep35) were eliminated. Subsequent data eliminations (named no_feature, ambiguous, too_low_aQual, not_aligned, and alignment_not_unique) yielded a count matrix consisting of 7126 genes × 86 samples.

2.3.3 Bottomly's mouse data (RNA-seq)

Bottomly's mouse data [51] were obtained from the ReCount website (http://bowtie-bio.sourceforge.net/recount/countTables/bottomly_count_table.txt) [52] and consisted of 36,536 genes × 21 samples.

2.3.4 Cheung's human data (RNA-seq)

Cheung's human data [53] were obtained from the ReCount website (http://bowtie-bio.sourceforge.net/recount/countTables/cheung_count_table.txt) [52] and consisted of 52,580 genes × 41 samples.

2.3.5 Nakai's probe-level data (microarray)

Nakai's probe-level data (.CEL files) [64] were obtained from the ArrayExpress website [54] through an R package ArrayExpress [55] by applying "GSE7623." The MAS-quantified data were obtained using the *mas5* function in the R/Bioconductor package affy [56]. Expression signals less than 1 were set to 1 and were subsequently log2-transformed. The RMA-quantified data were obtained using the *rma* function in the same package, i.e., affy. The output of the function was already log2-transformed. The RobLoxBioC-quantified data were obtained using the *robloxbioc* function in the R package RobLoxBioC [57]. The expression signals less than 1 were set to 1 and were subsequently log2-transformed.

2.3.6 Kamei's probe-level data (microarray)

Kamei's probe-level data (.CEL files) [68] were obtained from the ArrayExpress website using the R package ArrayExpress by applying "GSE30533." The subsequent procedures were the same as those described for the Nakai's data. Note that the quantification procedure was performed using R ver. 3.1.3 (affy ver. 1.44.0) because we encountered an error when executing the functions *mas5* and *robloxbioc* in R ver. 3.3.2 (affy ver. 1.52.0).

Chapter 3 Results

3.1 RNA-seq (two groups)

In DGE analyses, a gene expression matrix is typically generated, where each row indicates the gene (or derivatives), each column indicates the sample, and each cell indicates counts for RNA-seq data. Previous observation of the positive correlation between SC and DGE results [39] was obtained from an RNA-seq dataset (referred to as Blekhman, for short) consisting of 20,689 genes × 18 samples (= 3 species × 2 sexes × 3 BRs) [49]. The HSC and DGE analyses were performed using TCC. TCC implements a robust normalization strategy (called DEGES [34]) that uses functions provided in four widely used packages (baySeq [58], edgeR [47], DESeq [26], and DESeq2 [59]). For simplicity and/or the algorithmic advantage [60]-[61], I only used TCC for the DGE analysis of RNA-seq data. Specifically, I used the default DGE pipeline (iDEGES/edgeR-edgeR in [46] and EEE-E in [39]). When performing HSC for all input data, I used the clustering function *clusterSample* with default options ("1 – Spearman's correlation coefficient (r)" as a distance estimate and average-linkage agglomeration) in TCC.

Throughout this study, I filtered out genes with zero counts (or signals) in all samples (RNA-seq). For HSC analyses, an additional filtering was performed where genes having identical expression patterns were collapsed. Expression data having those unique expression patterns were used for calculating distance defined as "1 – Spearman's r." This filtering procedure was intended to reduce the negative impact of genes with low expression levels when calculating the distance between samples. For example, the Blekhman's data yielded 17,886 genes after the zero-count filtering and DGE analyses were performed. After unique filtering, 16,560 genes were obtained, and HSC was

performed using these genes. For simplicity, I focus on two-group comparisons with three replicates for each group, i.e., (A1, A2, A3) vs. (B1, B2, B3), in most cases. In this study, I use the terms samples and replicates interchangeably. My primary interest was to investigate the applicability of silhouette score for the objective evaluation of gene expression data based on arbitrary grouping criteria. By using silhouette score (i.e., AS values) as a relevant measure for the group differentiation in the HSC results, I re-evaluated the previous observations (i.e., the positive correlation between HSC and DGE results) [39].

3.1.1 Representative relationship between HSC and DGE results with AS

I first demonstrate the relationship between HSC and DGE results using a representative dataset, the Blekhman data obtained for three species (i.e., the three-group data): humans (HS), chimpanzees (PT), and rhesus macaques (RM) [49]. Briefly, Blekhman et al. studied expression levels in liver samples from three males (M1, M2, and M3) and three females (F1, F2, and F3) from each species/group. Figure 4a shows the HSC dendrogram based on a correlation distance (1-*r*) metric and average-linkage agglomeration. There were three major clusters, each of which represented a particular species (HS, PT, and RM clusters) and the RM cluster was relatively distant from the other clusters. Different from the clear interspecific discrimination (i.e., high dissimilarity between species), I observed a very low degree of separation between sexes (F vs. M) within each of the three major clusters. That is, samples labelled female (F) and male (M) were intermingled within each species, except for the PTF sub-cluster comprising three female samples (PTF1, PTF2, and PTF3).

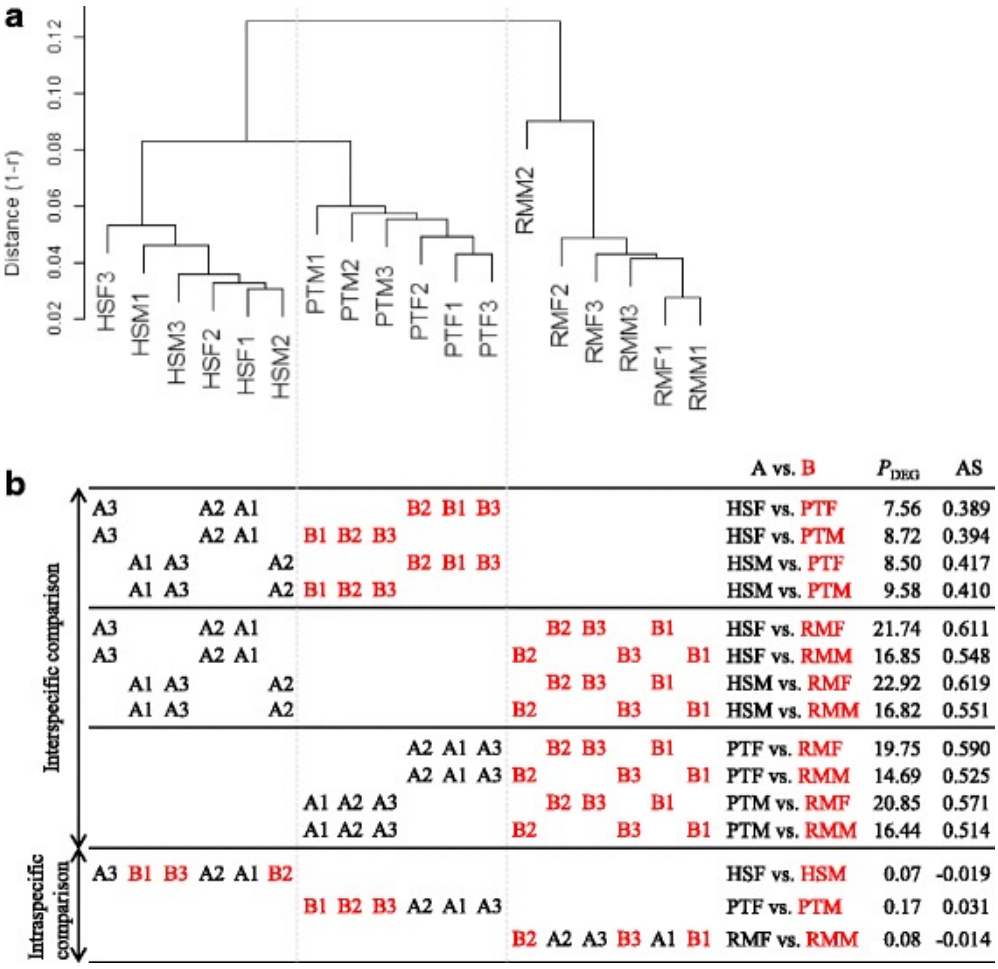


Figure 4 Relationship between the shape of HSC and DGE results.

(a) HSC dendrogram for Blekhman data consisting of 16,560 genes × 18 samples. The clustering was performed using the *clusterSample* function with default options in TCC. The unique filtering (from 17,886 genes to 16,560 genes with unique expression patterns across 18 samples) was internally performed in the function to reduce the negative effect on associations in low count regions when calculating Spearman's r as a distance measure. (b) DGE results from a total of 15 two-group comparisons with three replicates. The DGE pipeline provided in TCC was applied to the Blekhman's count matrix consisting of 17,886 genes after zero-count filtering. The P_{DEG} values and AS values for individual comparisons are provided on the right.

Figure 4b shows 15 DGE results for two-group comparisons. The percentages of DEGs (P_{DEG}) satisfying the 10% false discovery rate (FDR) threshold were obtained using TCC with default settings. The four P_{DEG} values for the HS vs. PT comparison (7.56–9.58%) were much smaller than those for either the HS vs. RM (16.82–22.92%) or the PT vs. RM comparison (14.69–20.85%). These results are consistent with those of the original study [49] and can primarily be explained by the interspecific distances shown in Figure 4a. Different from the interspecific comparisons, sex comparisons (F vs. M) showed extremely low P_{DEG} values (0.07–0.17%). This is consistent with the lack of separation between female and male samples within each species in the HSC analysis (Figure 4a).

It is noteworthy that, in the eight RM-related inter-group comparisons, both P_{DEG} and AS values obtained from four RMF-related comparisons were consistently larger than those from the four RMM-related comparisons. For example, for the HSF vs. RMF comparison, P_{DEG} = 21.74% and AS = 0.611, while for the HSF vs. RMM comparison, P_{DEG} = 16.85% and AS = 0.548. This difference is primarily explained by the smaller average distance of samples in RMF (0.0475) than in RMM (0.0722). Small P_{DEG} values (0.07–0.17%) obtained for the sex (i.e., intra-group) comparisons can be explained by the similarity between inter-group distances and intra-group distances. In other words, two-group comparisons showing AS ≈ 0 would result in few, if any, DEGs. The numbers of DEGs (or P_{DEG} values) can, of course, vary with FDR thresholds and generally increase when the threshold is less restrictive.

Nevertheless, I confirmed that the general trends for the 15 two-group comparisons were the same at 1%, 5%, 10%, 20%, 30%, and 40% FDR thresholds. Based on the definition

of FDR, an increase in the P_{DEG} value by loosening the FDR threshold does not necessarily indicate an increase in the true number of DEGs. For example, $P_{DEG} = 0.78\%$ at a 40% FDR for the PTF vs. PTM comparison indicates that $0.78 \times 0.4 = 0.31\%$ are non-DEGs, and the remaining $0.78 \times (1.0–0.4) = 0.47\%$ are, at least statistically, true DEGs. In my experience, the percentage of true DEGs (say $P_{trueDEG}$) generally approaches a constant value at a non-stringent FDR threshold, such as 30% or 40%. In this case, the maximum $P_{trueDEG}$ value for any sex comparison was ~ 0.5%. These results indicate that differences in P_{DEG} values with respect to the FDR threshold are not important.

Based on my visual evaluation, the AS values effectively represented the overall relationship between groups of interest in the HSC analysis (shown in Figure 4a). I think the expressive power in cases of few or no DEGs in the dataset (i.e., AS ≈ 0) is practically promising, but increasing the correlation between P_{DEG} (or $P_{trueDEG}$) and AS is not practical. This is simply because the P_{DEG} value tends to increase as the number of replicates (*Nrep*) increases [62], suggesting that the correlation is influenced by *Nrep*.

3.1.2 Effects of the number of replicates (*Nrep*) on parameter estimates

I next investigated the effects of *Nrep* on P_{DEG} and AS values, using both simulated and real RNA-seq data. The simulated data were constructed as follows: two-group comparison (A vs. B) with 40 replicates per group (*Nrep* = 40), 10,000 total genes, of which 20% were DEGs (2,000 DEGs and 8,000 non-DEGs; $P_{simDEG} = 20\%$), the expression levels of DEGs were four-fold in individual groups, and the proportions of DEGs up-regulated in individual groups were the same (i.e., 1,000 DEGs are up-regulated

in group A). For a total of 80 samples (A1, A2,…, A40, B1, B2, …, B40), I obtained P_{DEG} = 21.0% at a 10% FDR threshold, AS = 0.2409, and area under the ROC curve (AUC) = 0.9986. The AUC is a widely used measure of both the sensitivity and specificity of the DGE pipelines [34], [39], [46], [63]. The value (ranging from 0 to 1) can also be regarded as an overall indicator of the ability to distinguish true DEGs from non-DEGs. A larger AUC value indicates better DGE separation and vice versa. The AUC value of 0.9986 indicates nearly perfect separation and the estimated P_{DEG} value (21.0% at FDR = 0.1) is in good agreement with the true value (i.e., 20% DEGs or P_{simDEG} = 20%).

The DGE pipeline was used to examine subsets from the baseline matrix with 40 replicates per group (*Nrep* = 40). Bootstrap resampling was performed 100 times at *Nrep* = 3, 6,…, and 30 (without replacement). Consistent the previous observations [62], the average P_{DEG} values increased as a function of *Nrep* (Figure 5a). However, such an increasing trend was not observed for AS (Figure 5b). This result indicates that the silhouette score (i.e., AS) is independent of *Nrep*. Note that the P_{DEG} value approached to the true value (P_{simDEG} = 20%) as *Nrep* increased (Figure 5a). In general, the DGE pipeline does not necessarily produce a well-ranked gene list in which true DEGs are top-ranked and non-DEGs are bottom ranked. Given the increase in AUC values in conjunction with increases in P_{DEG} (Figure 5), this interpretation can be trusted in this case.

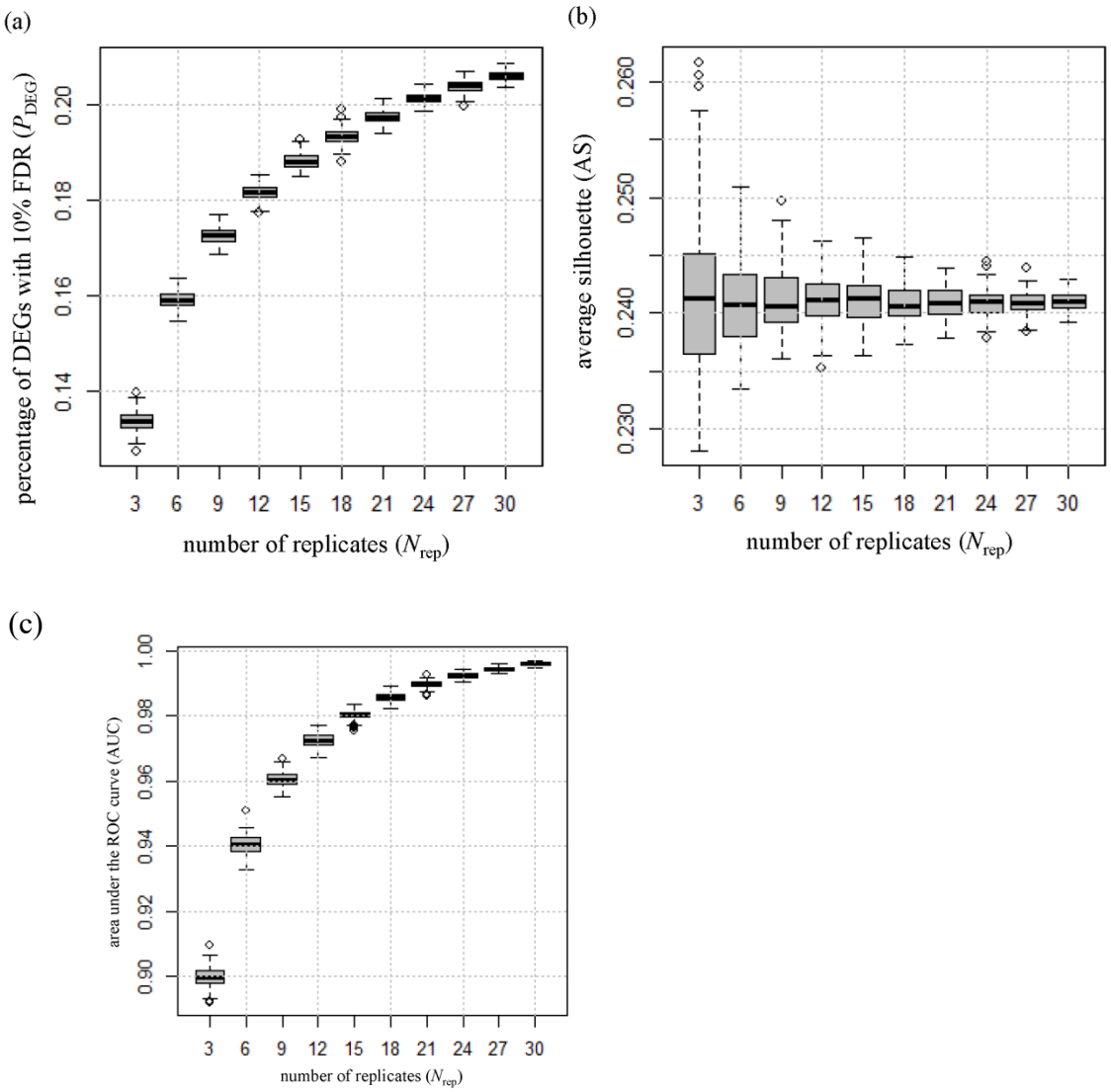


Figure 5 Effects of *Nrep* on parameter estimates (simulated data).

Bootstrapping results (100 iterations) from simulated RNA-seq data consisting of 10,000 genes × 80 samples with $P_{simDEG} = 20\%$ are shown. Vertical axes for the boxplots indicate: (a) P_{DEG}, (b) AS values, and (c) AUC values. Horizontal axes indicate the *Nrep* values (3, 6, …, 30). It can be seen that P_{DEG} and AUC values increase as a function of *Nrep*, but AS values do not.

Next, the effects of *Nrep* under different P_{simDEG} conditions (P_{simDEG} = 10%, 5%, 2%, 1%, 0.5%, 0.1%, and 0.02%) were investigated. I confirmed that P_{DEG}, but not on AS, is dependent on *Nrep* (Additional file 2). Different from the condition shown in Figure 5 (P_{simDEG} = 20%), however, I observed a transition in the distribution of P_{DEG} values at around P_{simDEG} = 1%. Although the P_{DEG} value monotonously increased as *Nrep* increases when P_{simDEG} was 20% or more, the P_{DEG} value switched to a monotonously decreasing trend when P_{simDEG} was 0.1% or less. Overall, the P_{DEG} values approached the true values (i.e., the P_{simDEG} values) as *Nrep* increased. These results indicate that more accurate DGE results can be obtained as *Nrep* increases, irrespective of the true percentages of DEGs in the data.

A similar analysis was performed using another real RNA-seq dataset consisting of 7,126 genes × 96 samples [50], [62]. Ten outlier samples were rejected, following the original study [62], and subsequent zero-count filtering of the original data yielded 6,885 genes × 86 samples (unique filtering did not have any effect for this dataset). For the data (called Schurch for short) comparing two groups (42 wild-type samples vs. 44 *Δsnf2* mutant samples), I obtained P_{DEG} = 78.1% and AS = 0.7289. Note that the AUC value could not be calculated for the data because, different from simulated data, I do not know which genes are true DEGs. I investigated the effects of *Nrep* on parameter estimates. The results were quite similar to those obtained using simulated data (shown in Figure 5), i.e., P_{DEG} was dependent on *Nrep*, but AS was not (Additional file 3). Note that the distribution of P_{DEG} values obtained using TCC (Additional file 3a) was also similar to that obtained using edgeR [47] (Figure.1a in [62]). This is quite reasonable because the DGE pipeline implemented in TCC can be viewed as an iterative edgeR pipeline [39].

3.1.3 Relationships between P_{DEG} and AS values

Next, I investigated the relationships between P_{DEG} and AS values under a fixed *Nrep* of 3. Figure 6 shows the results for (a) Schurch, (b) simulated, and (c) the mixture. For simulated data, I examined 19 P_{simDEG} conditions from 5% (black in Figure 6b) to 0.95 (red in Figure 6b). Overall, there was a strong positive correlation between P_{DEG} and AS values in this condition (Figure 6c). However, the accurate estimation of P_{DEG} using AS is not realistic and accordingly is not a goal of the current study. This is mainly because P_{DEG} increases as a function of *Nrep*, while AS does not (Figure 5). In other words, the regression coefficients depend on *Nrep*. Most importantly, if one wants to calculate P_{DEG}, there is no need to estimate the AS value; rather, it is only necessary to directly execute the DE pipeline. Nevertheless, as P_{DEG} approaches 0, AS also approaches 0. This suggests that P_{DEG} values near 0 can be interpreted as a mathematical explanation for AS near 0, i.e., the samples in the two groups (A vs. B) were completely mixed. In statistical terms, this situation is essentially the same as the null hypothesis (*H0*: A = B). The acceptance of *H0* (AS = 0) indicates there are no or few DEGs in the two-group data ($P_{DEG} = 0$). In this sense, AS could be used as helpful information for the interpretation of DGE results, especially when only a few statistically significant DEGs are obtained.

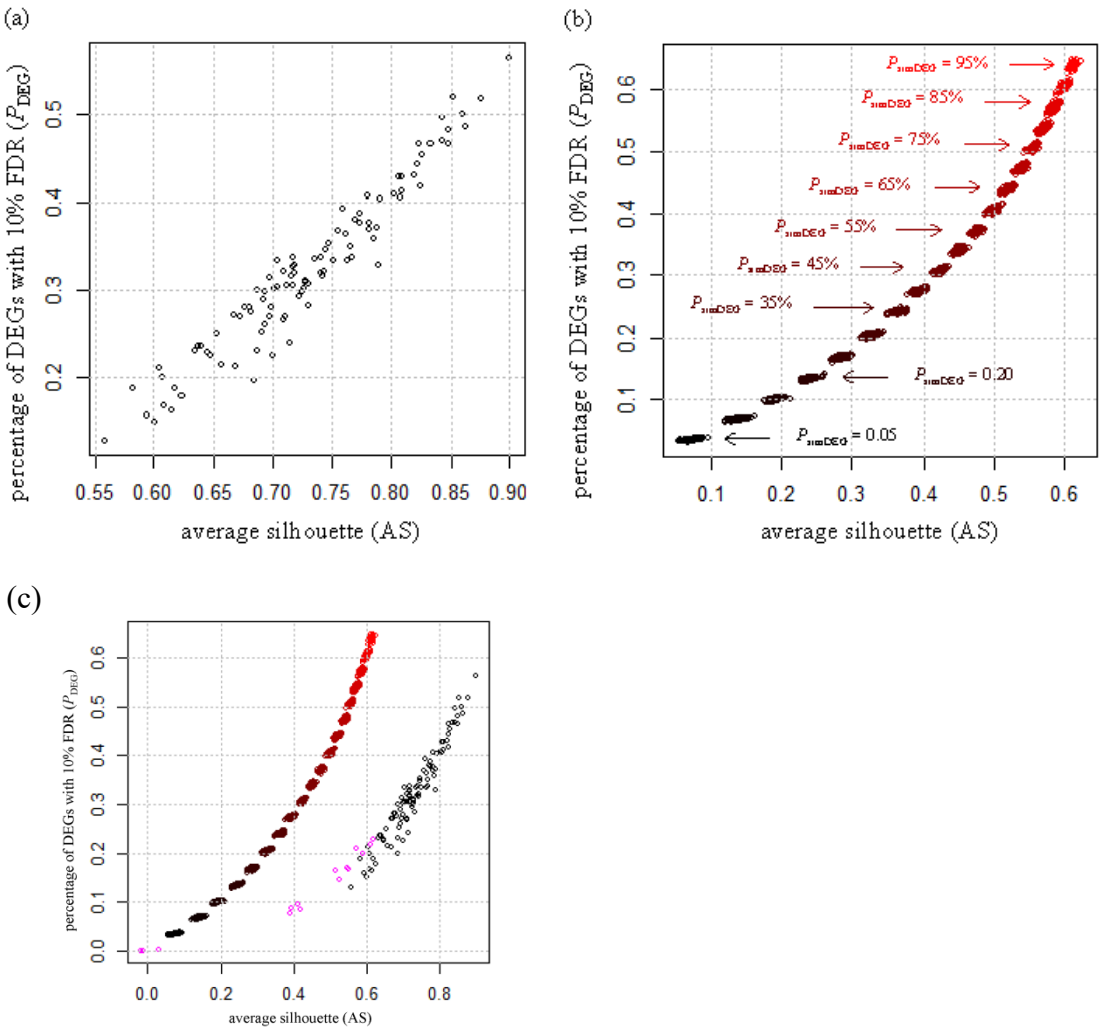


Figure 6 Relationship between P_{DEG} and AS values.

Scatter plots of P_{DEG} vs. AS at $Nrep = 3$ are shown. (a) Schurch data. The scatter plot shows a detailed relationship between P_{DEG} and AS values for Schurch data at $Nrep = 3$ (Additional file 3a and 3b). (b) Simulated data under $P_{simDEG} = 5\%, \ldots, 95\%$. The scatter plot for $P_{simDEG} = 20\%$ corresponds to the P_{DEG} (ranging from 0.1273 and 0.1397) and AS values (ranging from 0.2281 and 0.2617) for $Nrep = 3$ shown in Figure 5b. (c) The results for the mixture as well as the Blekhman data including 15 two-group comparisons shown in Figure 4b (magenta).

It should be noted that the distribution shown in Figure 6c differs substantially from the distribution for real data (Blekhman [49] and Schurch [62]) and simulated data, but the shapes of the distributions were similar. For example, the P_{DEG} value at AS = 0.6 was approximately 0.6 for the simulated data, while P_{DEG} for real data was approximately 0.2. Since the AS value for the simulated data at $P_{DEG} = 0.2$ was approximately 0.3, the difference for AS at $P_{DEG} = 0.2$ was 0.3. Similarly, the difference for P_{DEG} at AS = 0.6 was 0.4. It should also be noted that the distribution of values for Blekhman (magenta) and Schurch (black with AS > 0.5) was different (Figure 6c). While low AS values (-0.019~0.619) and low P_{DEG} values (0.07–22.92%) were obtained for the Blekhman data, high AS values (0.5585–0.8998) and high P_{DEG} values (13.03–56.34%) were obtained for the Schurch data. The difference can be explained by the intra-group distances. For the Schurch data, including 42 wild-type samples (group A) and 44 *Δsnf2* mutant samples (group B), the distances for groups A and B were 0.0144 and 0.0084, respectively. The values obtained for the Schurch data were clearly smaller than those obtained for the Blekhman data (> 0.04; Figure 4a). According to a previous study [62], the Schurch data represents a best-case scenario for DE pipelines, since the within-group biological variation (BV) is low. As the BVs roughly correspond to the intra-group distances, many other real RNA-seq data may display low P_{DEG} and AS values compared to those obtained for the Schurch data.

3.2 Microarray (two groups)

I also investigated two microarray datasets obtained using the Affymetrix Rat Genome 230 2.0 Array (GPL1355). The first dataset (called Nakai [64]) consisted of 31,099 probesets (which can be viewed as genes) × 24 samples (= 3 tissues × 2 conditions × 4 BRs). Briefly, Nakai et al. studied the expression levels of genes in brown adipose tissues (BAT), white adipose tissues (WAT), and liver tissues (LIV). They compared two conditions (fed vs. fasted for 24 h) for each tissue type. I here denoted the fed BAT samples BAT_fed, the 24 h–fasted LIV samples LIV_fas, and so on. To quantify expression from the probe-level data (i.e., Affymetrix CEL files), I applied three algorithms (MAS [65], RMA [10], and RobLoxBioC [57]). Different from RNA-seq data represented as integer counts, microarray data are expressed as continuous signals and in most cases are log-transformed. I therefore applied a specialized DE pipeline for microarray data provided in the package limma [66], instead of the DE pipeline used for RNA-seq data in TCC.

As expected based on the nature of microarray expression signals, zero signal values were not obtained for any genes in all samples and all genes displayed unique expression patterns. Accordingly, the subsequent analysis of microarray data was performed based on total set of genes (= 31,099). The HSC dendrogram for the Nakai data displayed three major clusters corresponding to the three tissue types (LIV, WAT, and BAT clusters) for all quantification algorithms (MAS, RMA, and RobLoxBioC; Figure 7a). Since the experimental design and the HSC dendrogram were very similar to those of the Blekhman data (Figure. 4), these microarray data can be regarded as the counterpart.

I performed 15 two-group comparisons with four BRs for each group, i.e., (A1, A2, A3, A4) vs. (B1, B2, B3, B4). Overall, I observed highly similar trends for the Nakai data and the Blekhman data (Figure 7b). For MAS-quantified data, for example, four P_{DEG} values in the BAT vs. WAT comparison (24.49–34.98%) were smaller than those in the BAT vs. LIV comparison (41.79–44.63%) or WAT vs. LIV comparison (39.74–44.05%). Different from the clear inter-tissue differentiation (i.e., high dissimilarity between tissues), I detected a relatively low degree of separation between conditions (fed vs. fasted) within each of the three major clusters. The P_{DEG} values for the fed vs. fasted comparison were 4.5–8.79%. Of these three comparisons, the intra-BAT comparison (i.e., BAT_fed vs. BAT_fas) showed the highest P_{DEG} (8.79%) and AS (0.207) values.

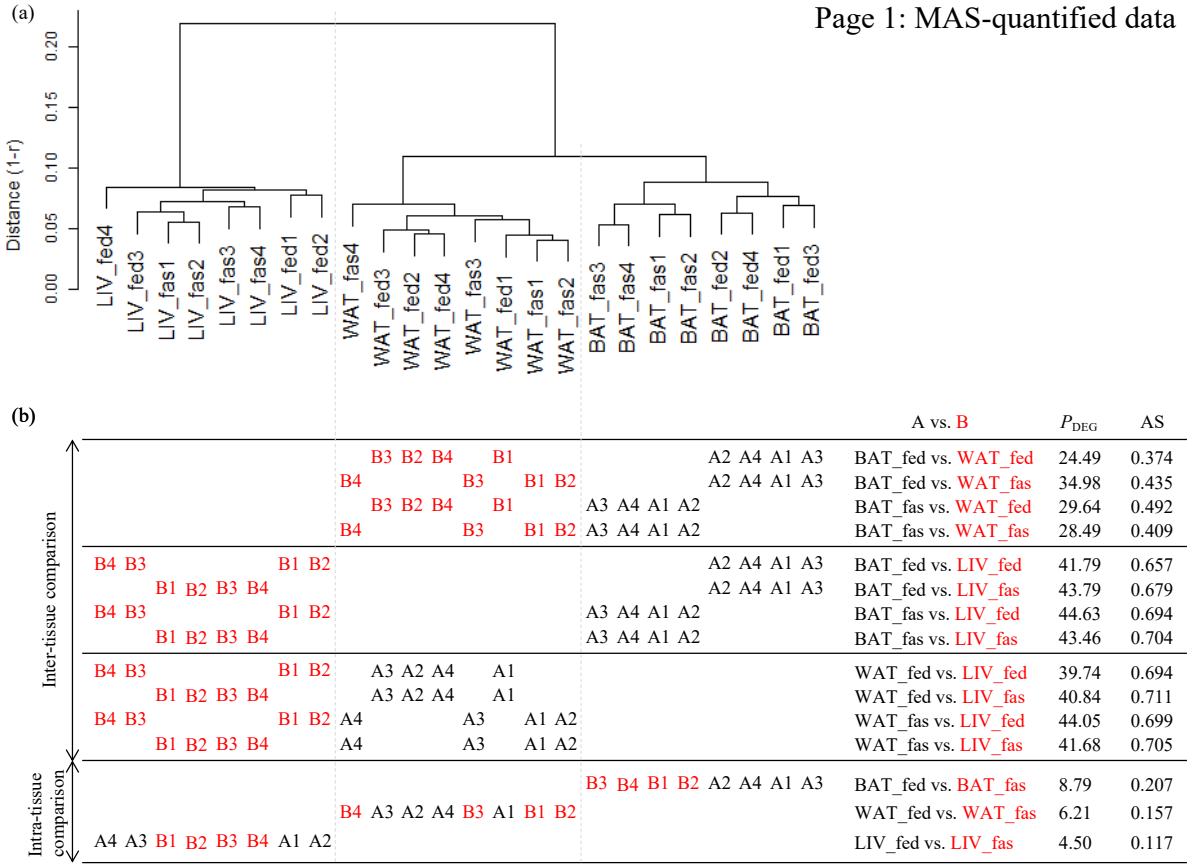


Figure 7 - Results for Nakai's microarray data.

(a) HSC dendrogram for Nakai data consisting of 31,099 genes × 24 samples and (b) P_{DEG} and AS values from a total of 15 two-group comparisons with $Nrep = 4$ are shown: MAS-quantified data (Page 1)

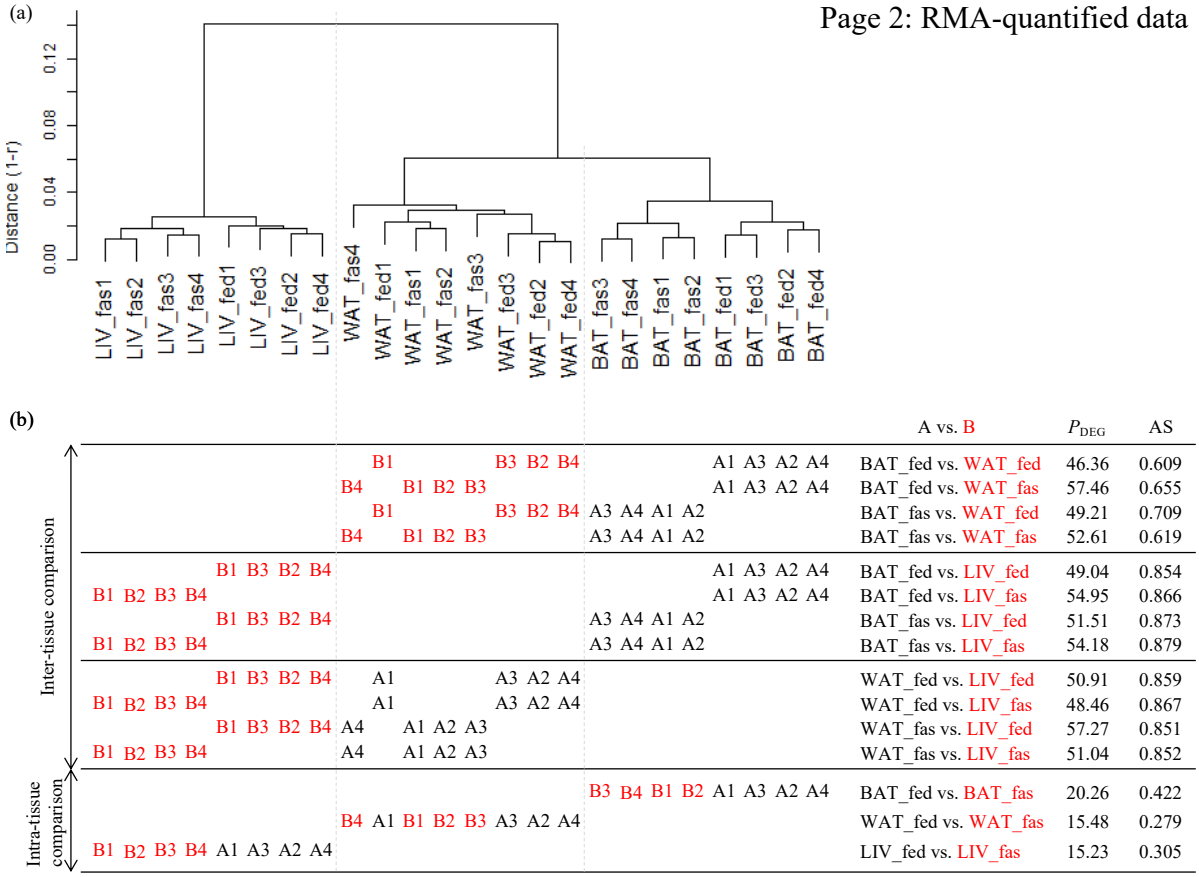
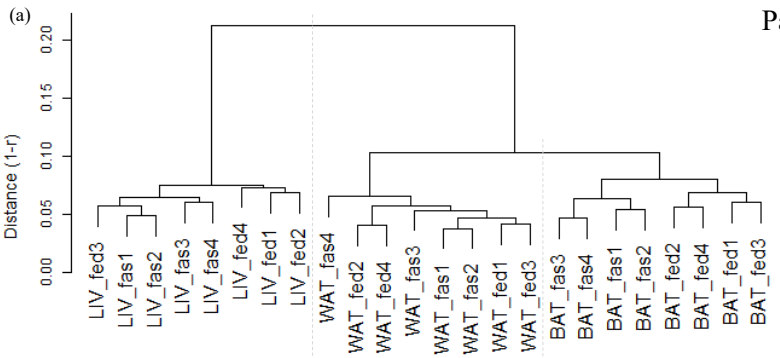


Figure 7 Results for Nakai's microarray data.

(a) HSC dendrogram for Nakai data consisting of 31,099 genes × 24 samples and (b) P_{DEG} and AS values from a total of 15 two-group comparisons with $Nrep = 4$ are shown: RMA-quantified data (Page 2).



(b)

	A vs. B	P_{DEG}	AS
Inter-tissue comparison	BAT_fed vs. WAT_fed	38.90	0.399
	BAT_fed vs. WAT_fas	36.07	0.460
	BAT_fas vs. WAT_fed	42.51	0.520
	BAT_fas vs. WAT_fas	35.19	0.435
	BAT_fed vs. LIV_fed	39.50	0.685
	BAT_fed vs. LIV_fas	43.42	0.706
	BAT_fas vs. LIV_fed	41.97	0.720
	BAT_fas vs. LIV_fas	43.11	0.729
	WAT_fed vs. LIV_fed	51.95	0.717
	WAT_fed vs. LIV_fas	51.66	0.734
	WAT_fas vs. LIV_fed	47.28	0.721
	WAT_fas vs. LIV_fas	46.07	0.727
Intra-tissue comparison	BAT_fed vs. BAT_fas	10.88	0.226
	WAT_fed vs. WAT_fas	5.48	0.168
	LIV_fed vs. LIV_fas	5.08	0.129

Figure 7 Results for Nakai's microarray data.

(a) HSC dendrogram for Nakai data consisting of 31,099 genes × 24 samples and (b) P_{DEG} and AS values from a total of 15 two-group comparisons with *Nrep* = 4 are shown: RobLoxBioC-quantified data (Page 3).

I observed similar results for RobLoxBioC-quantified data and relatively dissimilar results for RMA-quantified data. In particular, for the RMA-quantified data, I detected higher P_{DEG} and AS values compared to those of the other data. There are several potential explanations. RMA treats a batch of arrays simultaneously, while MAS and RobLoxBioC treat each array independently. RMA tends to overestimate sample similarity [67]. Combinations of DE pipelines with different quantification algorithms might also explain the higher P_{DEG} values observed in RMA-quantified data: limma is more compatible with MAS than RMA [8], [10]. Nevertheless, I observed a clear positive relationship between P_{DEG} and AS values, suggesting that AS is also applicable to microarray data.

The second dataset (called Kamei [68]) consisted of 31,099 genes × 10 samples (five BRs per group). Briefly, Kamei et al. compared gene expression in livers for rats fed a low-iron diet (approximately 3 ppm iron) for 3 days and a normal diet (48 ppm iron) as a control. The P_{DEG} and AS values obtained (Iron_def vs. Control) were close to zero and the HSC dendrogram showed an intermingled structure (Figure 8). These results indicate that the Kamei data can be regarded as a counterpart of the Cheung data (Additional file 5). AS can be utilized as supporting information to interpret DGE results for both RNA-seq and microarray data, especially when no or few DEGs were obtained.

One should note that one sample (Iron_def1) was a clear outlier in the HSC dendrogram for the RMA-quantified data, but not in the other dendrograms (Figure. 8). Iron_def3 was the most distant from the other samples in MAS- and RobLoxBioC-quantified data. This difference can also be explained by tendency of RMA to overestimate sample similarity

[67]. Indeed, the average distance (0.007) among samples in RMA-quantified data was considerably lower than those for the other datasets (0.043 for MAS and 0.037 for RobLoxBioC). The expression levels for the two microarray datasets (Nakai and Kamei) were obtained using the same device (i.e., the Affymetrix Rat Genome 230 2.0 Array), indicating that the datasets can be directly compared. The average distances among ten liver samples in the Kamei data were clearly lower than those among eight liver samples (LIV) in the Nakai data (0.078 for MAS, 0.022 for RMA, and 0.070 for RobLoxBioC). These results suggest that the differences in the most distant samples in the Kamei data (Iron_def1 in RMA data and Iron_def3 in the other data) are within the error range.

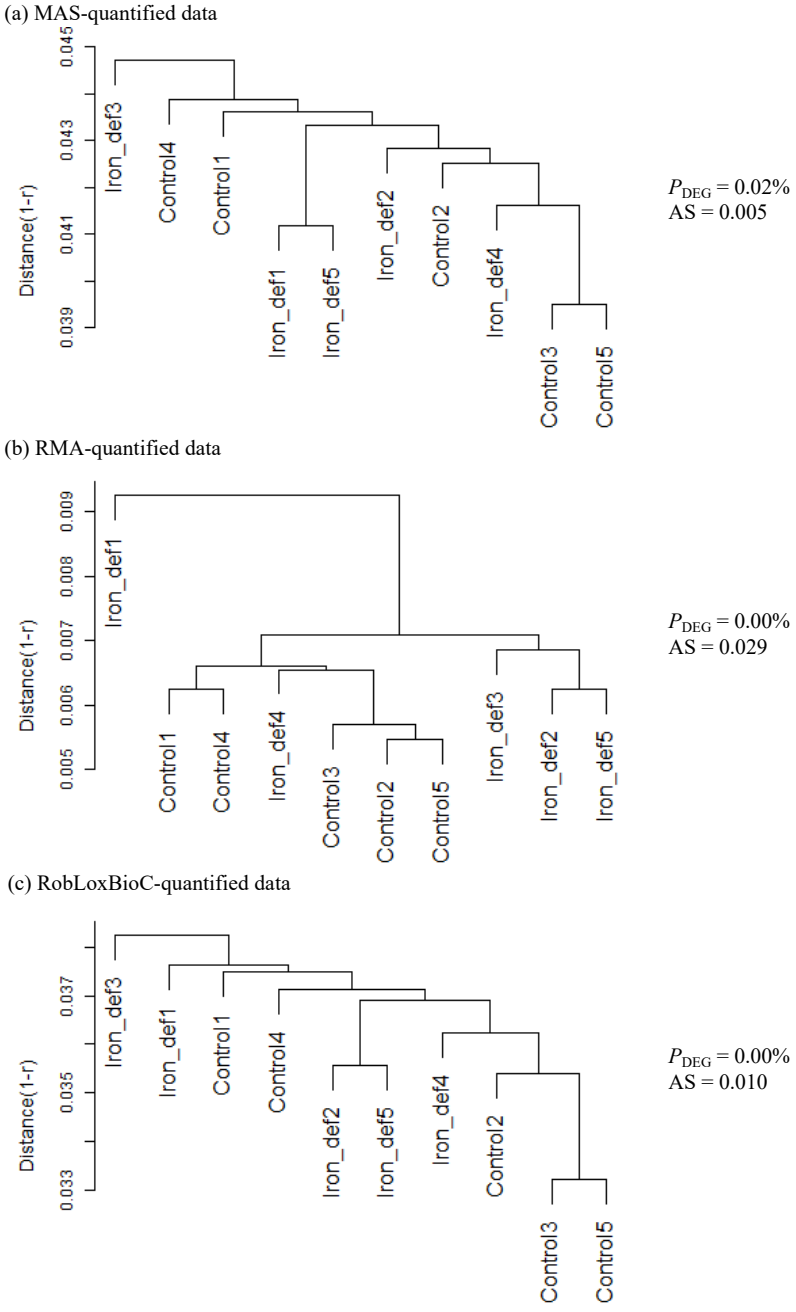
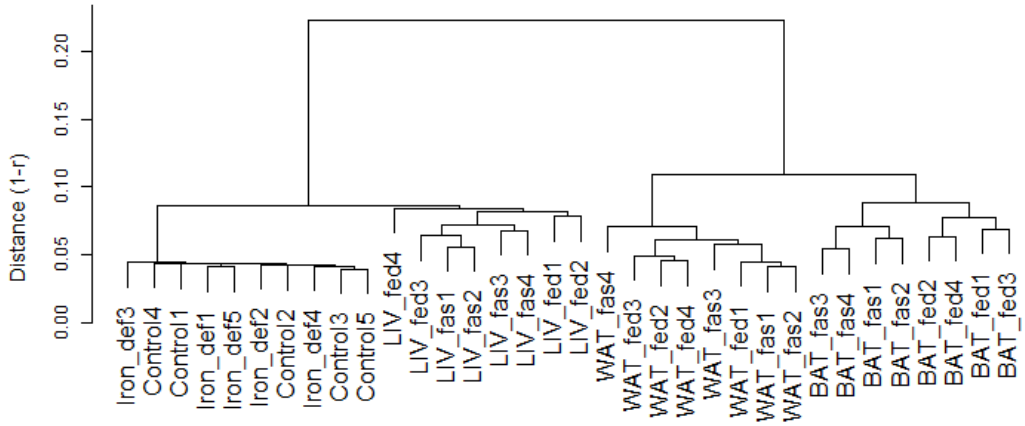


Figure 8 Results for Kamei's microarray data.

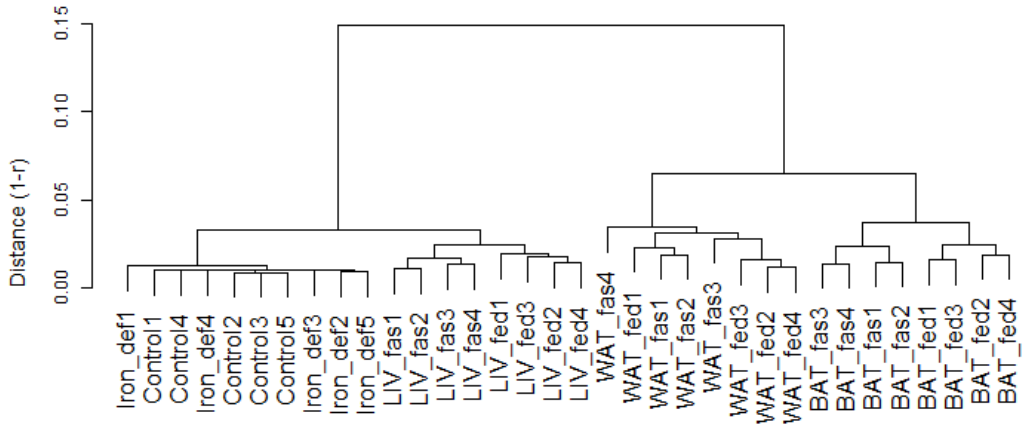
HSC dendrograms for (a) MAS-, (b) RMA-, and (c) RobLoxBioC-quantified data are shown. These data consist of 31,099 genes × 10 samples and compares two conditions (five Iron_def samples vs. five Control samples). The P_{DEG} and AS values are also shown on the right side of the dendrogram.

HSC dendrograms of the merged data provided several insights (Figure 9). First, the ten liver samples in Kamei data formed a tight cluster, even after adding the Nakai data, and formed a larger cluster when the eight liver samples from the Nakai data were included, confirming the overall similarities among various tissues (i.e., a sanity check) [36]–[38]. Second, compared to 24-h fasting, the short-term, iron-deficient diet might not result in significant differences in gene expression. This conclusion is supported by adding other publicly available dataset(s) for identical (or highly similar) tissues. It may be more important to add independent, publicly available datasets than to perform more detailed analyses using a single dataset. Third, an appropriate distance measure is important. The distance was defined here as (1 - Spearman's r); this definition is widely used [32], [38]. Since the distance ranges from 0 to 2, the interpretation is relatively easy compared to the interpretation of Euclidean distances, which range from 0 to ∞. I indeed understood the extremely high similarity among the ten liver samples in the Kamei data in the context of the very small distance values. In general, distance information is not interpreted so broadly in HSC analyses, but examinations of both the distance (1 - r) and AS may be useful.

(a) MAS-quantified data



(b) RMA-quantified data



(c) RobLoxBioC-quantified data

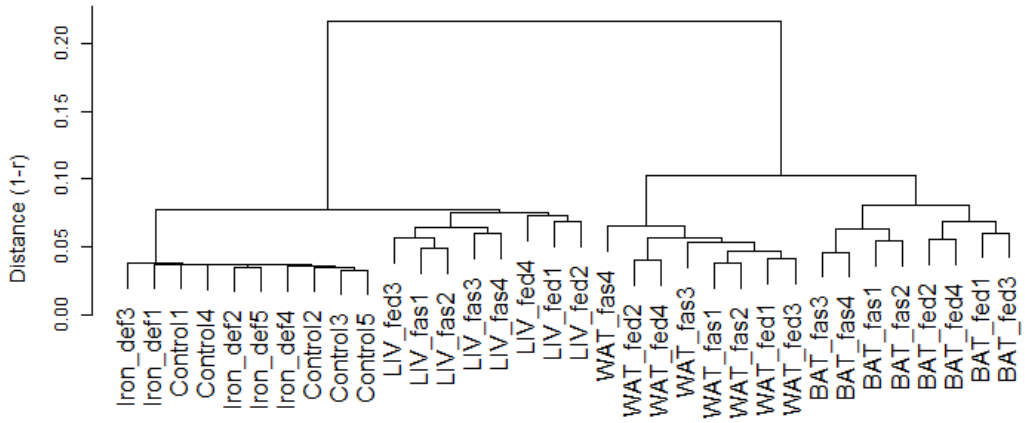


Figure 9 HSC dendrograms for merged microarray data (Nakai + Kamei).

HSC dendrograms for (a) MAS-, (b) RMA-, and (c) RobLoxBioC-quantified data are shown. These data consist of 31,099 genes × 34 samples (24 from Nakai and 10 from Kamei data).

3.3 Extension to multi-group comparison

In the previous section, a comprehensive study has been conducted on the relationships between HSC and DEGs in two-group comparison. In the study, silhouette score has shown its potential to objectively evaluate the degree of separation between groups of interest in the HSC dendrogram and estimate P_{DEG} values in DGE analysis. Actually, silhouette score was typically used to determine the best choice of number of clusters in unsupervised clustering like K-means [40]. The optimal number of clusters k is the one that maximizes the average silhouette scores over a range of possible values for k (k=1, 2, 3 and so on). In most case, the k is a value greater than two. Thus, it is natural to apply the criteria to multi-group condition. One thing should be noted that silhouette score is a measure of how close each point in its own cluster versus those in the neighbor cluster (the cluster that has the shortest distance from the centers of the chosen cluster). In the context, for a given $P_{DEG,}$ neighbor cluster or neighbor group is usually unstable due to the biased proportions of P_{DEG} assignment among each group. Hence, I proposed a more stable means termed AAS to quantify HSC dendrogram in the multi-group gene expression data. The AAS was obtained through calculating the average silhouette scores in a group pair-wise way. Equation was listed as follows:

$$AAS = \sum_{i \neq j}^{N} AS \Big/ \frac{n!}{r!(n-r)!} \tag{2}$$

in which, AS is average silhouette scores in pair wise comparisons could be obtained from the multiple groups, n is the number of whole groups under comparison, r is a fixed value 2.

In three-group condition, the AAS is equal to:

$$AAS = \frac{(AS_{1,2} + AS_{1,3} + AS_{2,3})}{3} \quad (3)$$

where: $AS_{1,2}$ represents the average silhouette scores obtained from the group1 and group2, $AS_{1,3}$ represents the average silhouette scores obtained from the group1 and group3 and $AS_{2,3}$ represents the average silhouette scores obtained from the group2 and group3.

For DGE analysis, TCC [46] package implements a generalized pipeline with the multi-step normalization can be described as X-(Y-X)n -Y . From an extensive evaluation among 12 pipelines [39], EEE-E pipeline was recommended on multi-group RNA-seq data (n=3) with few replicates as the best practices, in which X = TMM and Y = the DEG identification method, both were implemented in edgeR. The whole workflow of the EEE- E pipeline was summarized in the Figure 10. In this study, to reduce calculation cost, the parameter of iterations was set as a default value (n=3). Nevertheless, the DEG elimination strategy (called DEGES) can be repeated until the calculated normalization factors converge with a pretty large value($n \gg 3$). In the DEG identification step, the gene ranking (DGEList) was obtained through the likelihood ratio test under the GLM framework and a routine Benjamini-Hochberg procedure to adjust the p-value.

3.3.1 Simulation data with replicates

The generation of simulation data is intensively time consuming and laborious, since many parameters are involved in the process and have effects on eventual DGE analysis results. To perform the multi-group comparison as simply as possible, I focus here on the three-group data with the equal numbers of replicates (*Nrep*=3) per group. The simulation

conditions are summarized as follows: the total number of genes is 10,000 (*Ngene*=10,000), 5%, 10%, 15%, 20%, 25%, 30% in a sequential manner proportion of the genes are DEGs, the ratio of the four fold up-regulated DEGs divided in each group (G1, G2, G3) are (1/3, 1/3, 1/3), (0.5, 0.3, 0.2), (0.5, 0.4, 0.1), (0.6, 0.2, 0.2), (0.6, 0.3, 0.1), (0.7, 0.2, 0.1), and (0.8, 0.1, 0.1). All combinations of parameters produced 42 group data. In each group, the generation of simulation data, DGE analysis and the calculation of AS and AAS values were repeated in 100 trials.

I first assessed the performance of DEGs calling using the EEE-E pipeline in the simulation data. The AS and AAS values were also calculated simultaneously in every trial. Table 1 lists the average P_{DEG}, average AS values and average AAS values of 100 trials with *Nrep* =3. As shown in Figure 11-(A), the accuracy of the DEGs calling is decreasing when the true P_{DEG} setting versus the total number of genes across multi-condition increases. In the graph, the vertical axis is obtained P_{DEG} versus the corresponding true P_{DEG} setting in simulation data and horizontal axis is obtained AS and AAS value respectively. The low reproducibility maybe due to the strict FDR control (0.05). Since the AUC values of 100 trials between the ranked gene list and the truth for various simulation conditions with *Nrep*=3 are usually above 90%, loosening the FDR control to 0.1 may remedy the low DEGs calling efficiency. In the Figure 11-(B), the EEE-E pipeline shown its predominant that the P_{DEG} values keep consistency with low variance across the seven different proportions of DEGs up-regulated in individual groups (G1, G2, G3).

I also analyzed the performance of the AS and AAS values under various difference proportions of DEGs up-regulated. Comparing the two Figure 12-(A) and 12-(B), we can draw the conclusion that AAS is a more accurate and stable criteria than AS in multi-group condition. When the true P_{DEG} was set to 5%, the range between maximum and minimum value of AS is 0.0167, whilst the range between maximum and minimum value of AAS is 0.0025. When the true P_{DEG} was set to 30%, the range between maximum and minimum values of AS is 0.0948, whilst the range between maximum and minimum values of AAS is 0.0377. Across the seven conditions, the deviation from the mean of both AS and AAS values is increasing along with the global true P_{DEG} covering the whole three groups.

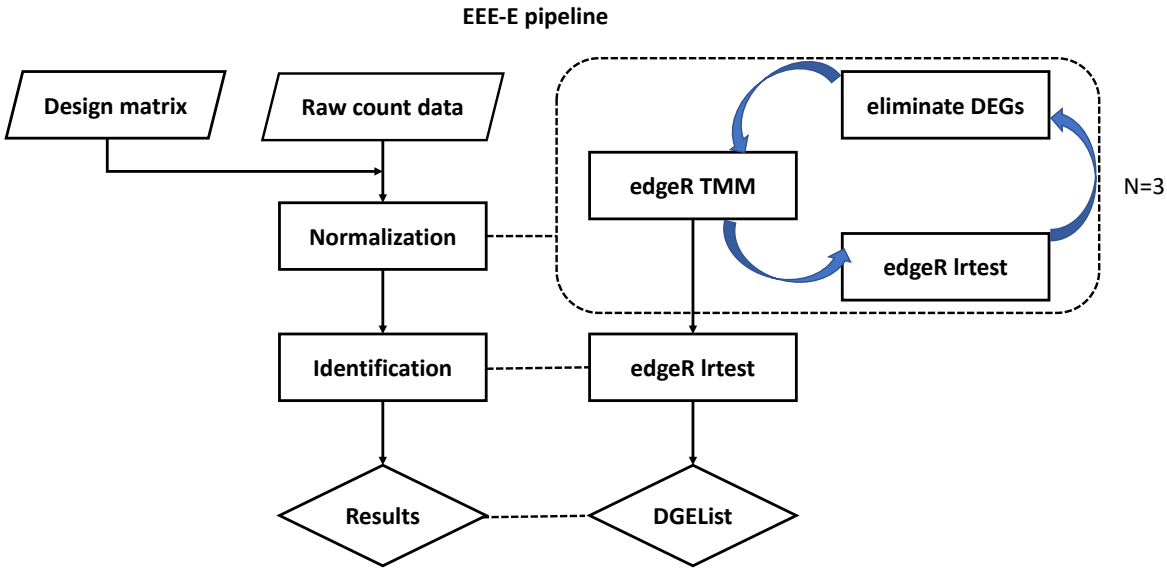


Figure 10 Schematic diagram of EEE-E pipeline in TCC package

There are two main steps in the RNA-seq data analysis: normalization and identification. In the context, I refer to the two steps as X for data normalization and Y for DEG identification. In the EEE-E pipeline, TMM is for X and the likelihood-ratio test (lrtest) in GLM model for Y in multi-group comparison. The pipeline with a multi-step normalization can be described as X-(Y-X)*n*-Y in which the X-(Y-X)*n* with n ≥ 2 (default setting *n*=3) corresponds to the iterative DEGES-based normalization. Finally, through the whole workflow, the expression levels of an individual gene between different groups can be compared and would be set into DEGs and non-DEGs groups.

Table 1- Silhouette score and DGE analysis results for simulation data

PG1	33%	50%	50%	60%	60%	70%	80%
PG2	33%	30%	40%	20%	30%	20%	10%
PG3	33%	20%	10%	20%	10%	10%	10%
P_{DEG} = 5%							
EEE-E (P_{DEG})	3.83%	3.84%	3.85%	3.83%	3.84%	3.83%	3.83%
AS	0.0499	0.0461	0.0433	0.0427	0.0413	0.0375	0.0332
AAS	0.0550	0.0545	0.0542	0.0544	0.0543	0.0534	0.0525
P_{DEG} = 10%							
EEE-E (P_{DEG})	7.22%	7.23%	7.28%	7.25%	7.29%	7.25%	7.21%
AS	0.0999	0.0902	0.0848	0.0828	0.0802	0.0736	0.0660
AAS	0.1047	0.1037	0.1034	0.1020	0.1025	0.1004	0.0981
P_{DEG} = 15%							
EEE-E (P_{DEG})	10.67%	10.65%	10.64%	10.64%	10.64%	10.58%	10.56%
AS	0.1458	0.1292	0.1208	0.1204	0.1142	0.1050	0.0935
AAS	0.1504	0.1473	0.1458	0.1454	0.1443	0.1402	0.1359
P_{DEG} = 20%							
EEE-E (P_{DEG})	14.09%	14.03%	14.02%	13.98%	14.01%	13.97%	13.96%
AS	0.1863	0.1652	0.1539	0.1526	0.1456	0.1338	0.1189
AAS	0.1908	0.1867	0.1838	0.1821	0.1819	0.1766	0.1695
P_{DEG} = 25%							
EEE-E (P_{DEG})	17.48%	17.43%	17.42%	17.41%	17.42%	17.38%	17.36%
AS	0.2235	0.1972	0.1839	0.1830	0.1731	0.1593	0.1409
AAS	0.2275	0.2217	0.2180	0.2168	0.2148	0.2078	0.1976
P_{DEG} = 30%							
EEE-E (P_{DEG})	20.85%	20.86%	20.81%	20.81%	20.77%	20.77%	20.71%
AS	0.2556	0.2275	0.2116	0.2111	0.1979	0.1815	0.1608
AAS	0.2596	0.2549	0.2497	0.2471	0.2437	0.2347	0.2219

P_{DEG}, AS and AAS of 100 trials for each simulation condition (P_{DEG} = 5%, P_{DEG} = 10%, P_{DEG} = 15%, P_{DEG} = 20%, P_{DEG} = 25%, P_{DEG} = 30%) are shown. Simulation data contains a total of 10,000 genes: P_{DEG} % of genes is for DEGs, PG1 % of P_{DEG} in G1 is higher than in the other groups, and each group has three BRs (*Nrep* = 3). Seven conditions are shown in total. EEE-E pipeline in TCC package was employed to call DEGs.

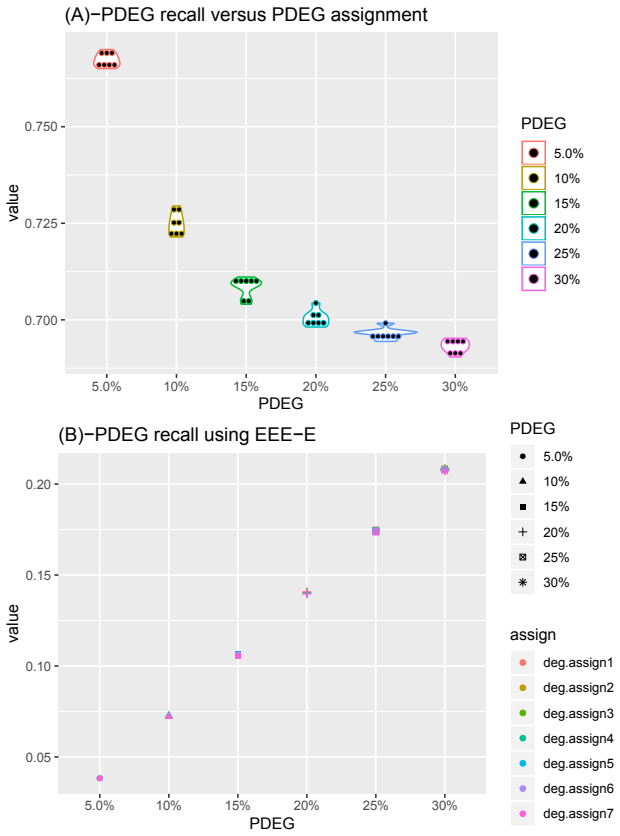


Figure 11 DGE analysis results in simulation data (*Nrep*=3)

The scatter plot for DGE analysis results of 100 trials for each simulation condition (P_{DEG} = 5%, P_{DEG} = 10%, P_{DEG} = 15%, P_{DEG} = 20%, P_{DEG} = 25%, P_{DEG} = 30%) are shown. (A) The horizontal axis represents ground truth P_{DEG} (parameter setting) of simulated data, and the vertical axis represents the recall of detected DEGs versus ground truth using EEE-E pipeline. The points representing different groups differ in their color. Violin plots were used to visualize variations in the data of each group. (B) The horizontal axis represents ground truth P_{DEG} (parameter setting) of simulated data, and the vertical axis represents the P_{DEG} of detected DEGs using EEE-E pipeline. The points representing different groups differ in their color and shape. The degs.assign1~7 correspond to the parameter setting among PG1, PG2 and PG3 in Table 1.

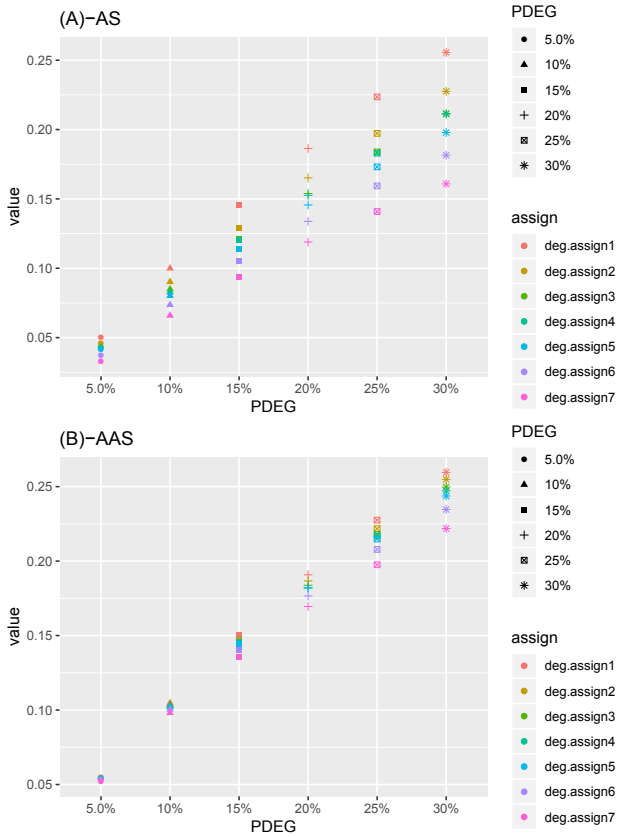


Figure 12 Silhouette score analysis results in simulation data (*Nrep*=3)

The scatter plot for AS and AAS calculation results of 100 trials for each simulation condition (P_{DEG} = 5%, P_{DEG} = 10%, P_{DEG} = 15%, P_{DEG} = 20%, P_{DEG} = 25%, P_{DEG} = 30%) are shown. (A) The horizontal axis represents ground truth P_{DEG} (parameter setting) of simulated data, and the vertical axis represents the AS value. (B) The horizontal axis represents ground truth P_{DEG} (parameter setting) of simulated data, and the vertical axis represents the AAS value. The points representing different groups differ in their color and shape. The degs.assign1~7 correspond to the parameter setting among PG1, PG2 and PG3 in Table 1

3.3.2 Real data with replicates

In addition to the simulation study, the Blekhman's data was also employed to evaluate the performance of two indices. To correctly estimate the biological variation and its effects on DGE analysis and silhouette score calculation, a reduced count matrix (i.e., 18 samples; 3 species × 2 sexes × 3 BRs) was used as input. Then, I regarded this dataset as a single-factor experimental design of three species where each has six biological replicates (i.e., HS_rep1-6 vs. PT_rep1-6 vs. RM_rep1-6). A total of 100 bootstrap trials were carried out in light of the AS and AAS as a function of the number of DEGs satisfying a FDR threshold (0.05 or 0.1). In each bootstrap iteration, three biological replicates were picked out from six biological replicates under the same group (the same species). The DEGs were obtained using the DGE analysis pipeline called EEE-E implemented in TCC. The AS and AAS were obtained as definition of equations 1 and 2.

The expression pattern of the DEGs obtained whether affect the calculation of AS score was also invested. I firstly classified the expression pattern of the DEGs obtained from the EEE-E pipeline in 100 trails. Sequentially, the DEGs results were assigned to one of ten possible pattern defined in baySeq [58]. The label information of these patterns to each gene was determined using the whole Belkhman's dataset (3 groups × 6 BRs). Finally, the change trends of each DEGs results assigned in definition groups was shown utilizing the parallel coordinate plots. The expression pattern of DEGs obtained by EBSeq was analyzed using the same procedure.

The results were summarized in Figure 13, additional file 6 and Table 2. When the FDR control was set to 0.05, the P_{DEG} falls within the range 0.2034 to 0.3779. When the FDR

control was set to 0.1, the P_{DEG} falls within the range 0.2585 to 0.4441. The results of P_{DEG} are in good agreement with the researchers' experience that the loose FDR control usually yields more DEGs. Corresponding to the number of DEGs calling changes, the AS value falls within the range 0.3807 to 0.5503 and the AAS value falls within the range 0.4633 to 0.6095. From the Figure 13 and additional file 6, I can draw a conclusion that there is a very little difference between the FDR = 0.05 and FDR = 0.1 which is reflected in the scatter plot and the trends (using a linear regression model fitting). In both graphs, there are a good linear relationship between the two variables (AS versus P_{DEG} or AAS versus P_{DEG}). In other words, the FDR control is not a major factor in the relationships between two the silhouette score and DGE analysis. Although in the simulation data, when the difference between the proportion of P_{DEG} assigned in each group became large, the AS cannot keep consistency under the same given P_{DEG} parameter setting. In the real data, although the bootstrap sampling was employed to increase the diversity of experimental sample for multiple-group comparison, not such a distinct difference of variance between AS and AAS were observed as expected in simulation data. A reasonable explanation is that the bootstrap sampling, in most case, just causes changes in the true P_{DEG} across the multiple conditions. In the sampling progress, proportion of P_{DEG} assigned in each group is floating up and down around a fixed value. In the previous study, I observed that the fraction of P_{DEG} among the multi-groups in blekhman's dataset is approximately equal to 1:1:1 [39]. The expression patterns of DEGs reported by EEE-E pipeline which is G1 up-regulated (G1>G2=G3,G1>G2>G3 and G1>G3>G2) : G2 up-regulated (G2>G1=G3,G2>G1>G3 and G2>G3>G1):G3 up-regulated (G3>G2=G1,G3>G2>G1 and G3>G1>G2) equal to 28.16% : 30.07% : 40.7%. According to my simulation data results (Figure 12), AS and AAS values both have the most least

variants, when the assignment of P_{DEG} is around the 1:1:1 (33% : 33% : 33%). In this study, I also invested expression patterns for DEGs in 100 bootstrap trials and assessed the effect of P_{DEG} assignment on silhouette score (AS and AAS). As shown in Figure 14, I traced the three potential outliers marked in Figure 13 and marked them using difference color against normal data. Both of them have an extreme value in group G3>G2>G1 and two of them were flagged out as outlier by boxplot statistics. It is considered to be the main reason for bad fitting results using linear regression. In Figure 15, it is more obviously that both of them were detected as outliers in pattern 1, pattern2 and pattern5 groups.

All in all, the AS is powerful assessment criteria to evaluate the clustering or degree of separation between groups predefined. It has potential to apply in DGE analysis filed. Especially in multi-group comparison condition, the calculation just requires distance matrix and commonly used group labels. Although in the simulation data, the AS is not so good as AAS that keep consistency even the proportion of P_{DEG} assigned in each group dramatically changes. In practice, it is easy to handle and offer reliable results making litter difference between AAS results.

Table 2- The statistics information about AS, AAS and P_{DEG} in difference FDR control

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
FDR= 0.05						
P_{DEG}	0.2034	0.2318	0.2632	0.2632	0.2866	0.2866
FDR= 0.1						
P_{DEG}	0.2585	0.2907	0.3248	0.325	0.3514	0.4441
AS	0.3807	0.4252	0.4695	0.4625	0.4933	0.5503
AAS	0.4633	0.5020	0.5361	0.5331	0.5592	0.6095

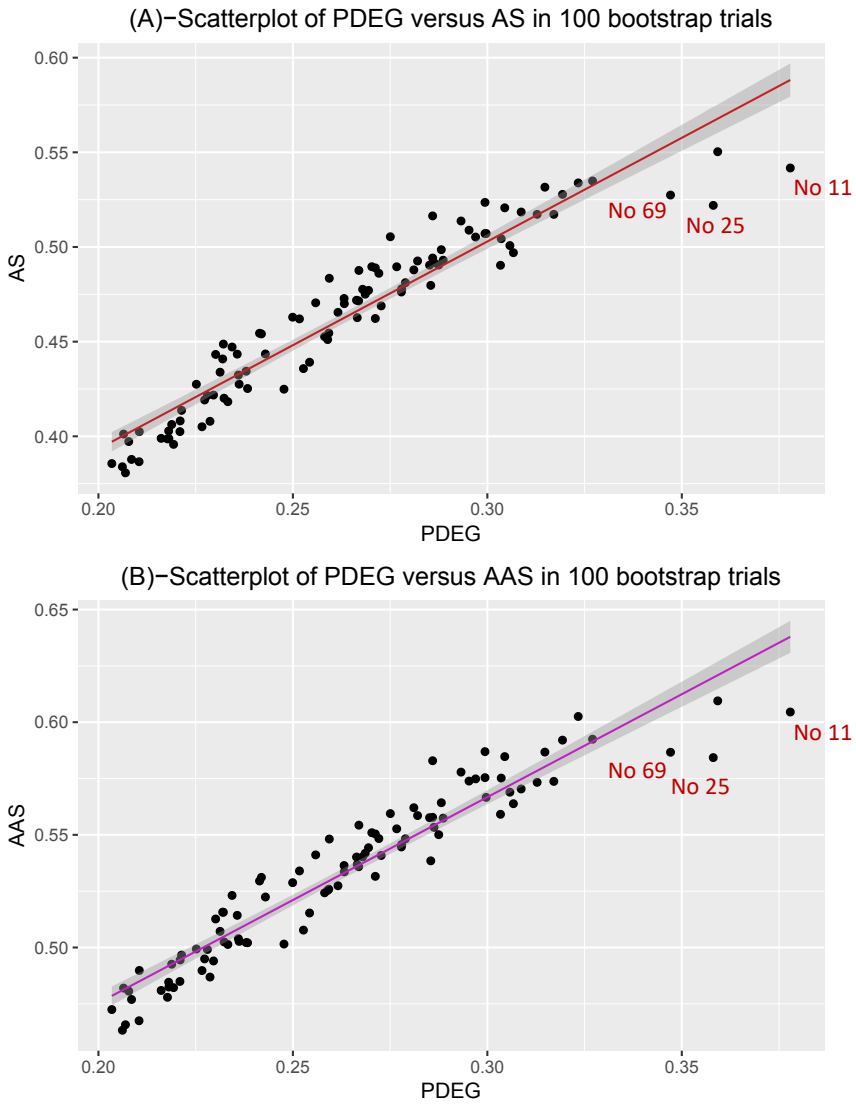


Figure 13 Silhouette score (AS and AAS) in bootstrap experiments (FDR=0.05).

The scatter plot for AS and AAS calculation results of 100 trials using bootstrap sampling without replacement in belkhman's dataset. A straight line of best fit through points was found via simple linear regression and three potential outliers were marked using red color annotations.(A) The horizontal axis represents AS value, and the vertical axis represents the P_{DEG} obtained via EEE-E pipeline.(B) The horizontal axis represents AAS value, and the vertical axis represents the P_{DEG} obtained via EEE-E pipeline.

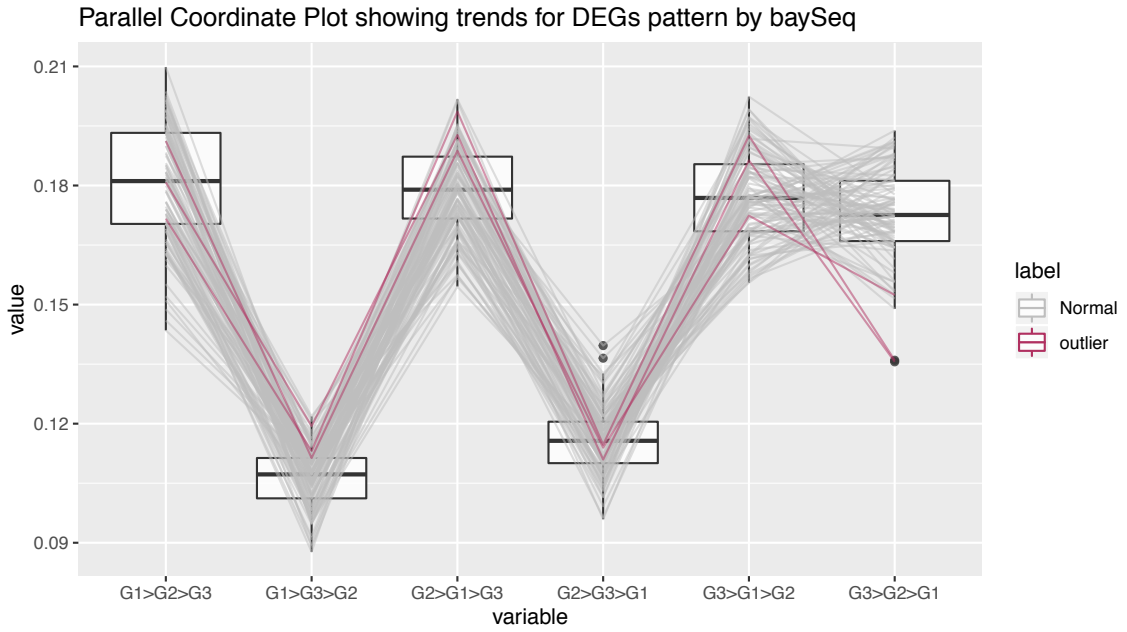


Figure 14 Parallel coordinate plot for DEGs pattern classified by baySeq.

The input data were DEGs results obtained through EEE-E pipeline under FDR control equals to 0.05. Among ten definition groups, six group with relative larger proportion of P_{DEG} were used to generate the graph. In the graph, gray color lines represent the clean normal data, while maroon color represent the potential outlier points (trial number 11, 25, 69)

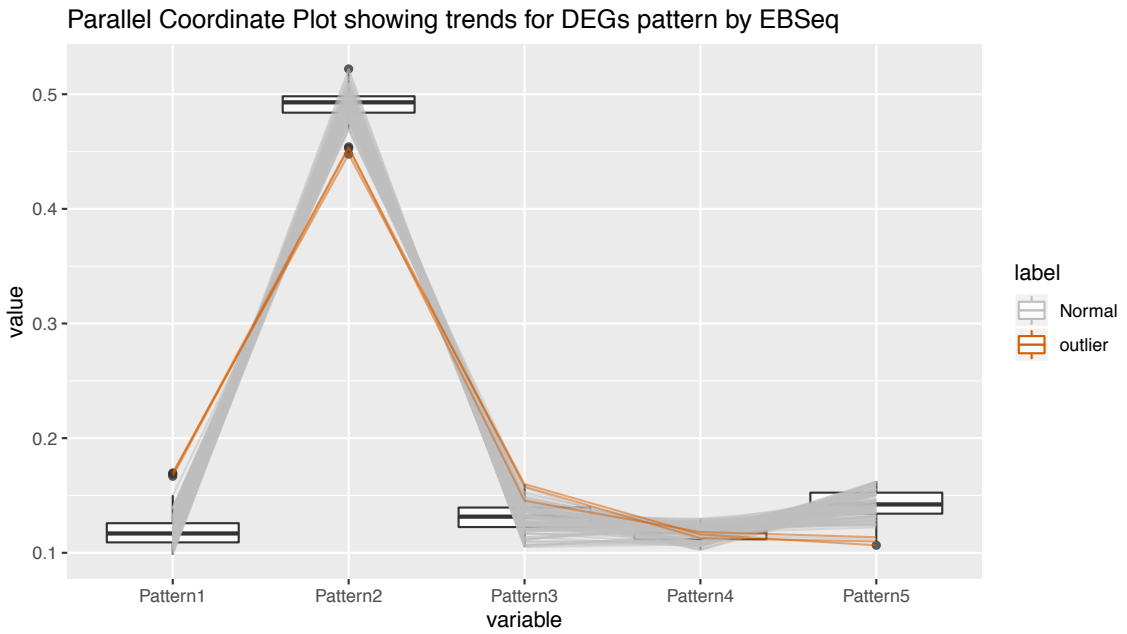


Figure 15 Parallel coordinate plot for DEGs pattern classified by EBSeq.

The input data were DEGs results obtained through EEE-E pipeline under FDR control equals to 0.05. The all five definition groups classified by EBSeq package were used to generate the graph. In the graph, gray color lines represent the clean normal data, while yellow color represent the potential outlier points (trial number 11, 25,69).

Chapter 4 Conclusions

In this study, I proposed to use silhouette score (i.e., AS values) as an objective measure for the degrees of separation between groups of interest based on expression data. To my knowledge, the use of AS independent from HSC is the first practical application in the field of gene expression analysis. My main findings are (i) AS is an effective indicator of the overall relationship in the HSC dendrogram based on arbitrary grouping criteria; (ii) AS values are independent of *Nrep*, while P_{DEG} values obtained from DGE analysis are fundamentally dependent on *Nrep*; and (iii) there is a positive correlation between AS and P_{DEG} values under a fixed *Nrep*. It is not necessary to estimate P_{DEG} from AS values because DGE results (including P_{DEG}) can be directly obtained via the DGE pipeline. The AS provides helpful information for interpreting DGE results as well as HSC results; (iv) The AS can be easy to adapt to multi-group comparison without additional calculation. In most cases, it offers promising results when the variation between proportions of DEGs in individual groups is not very large. The obtained AS value still keep an easy interpretable linear relationship with the DGE analysis results under the GLM framework.

Based on the current results, I conclude that my calculation procedure for AS is appropriate. The procedure consists of 1) filtering genes with low expression, 2) calculating distances among samples, and 3) calculating the AS values based on distance estimates. The high similarity among samples in the Kamei data could be detected by investigating the distances defined as (1 - Spearman's r). Considering this finding in addition to other data, some samples could be misidentified as outliers (e.g., Iron_def1 in Figure 8 and 9). In addition to the AS value obtained for the groups of interest, (i) the

investigations of distances among samples and/or groups in the dataset and (ii) comparison with other datasets obtained from the same or similar samples are practically important.

Of course, there are true outliers, e.g., ten outlying samples in the original Schurch data [62], [50]. I manually eliminated the ten outliers as determined in the original study [50] and analyzed 86 clean samples in this dataset. The values obtained without outliers (P_{DEG} = 78.1% and AS = 0.7289) were clearly higher than those with outliers (P_{DEG} = 74.7% and AS = 0.6530), indicating the importance of developing methods for the automatic detection of outliers [69]. My preliminary analysis for the original data using an existing method [70] successfully detected nine of the ten true outliers as well as three false positives. I obtained a promising result (P_{DEG} = 77.6% and AS = 0.7301) using the remaining 84 samples. Rational removal of outlying samples would yield better DGE results. I expect that AS would help objective evaluation of the changes in the DGE results accompanying outlier removal.

In practice, Silhouette score can be utilized as supporting information to interpret DGE results, especially when no or few DEGs are obtained. As demonstrated by several examples (e.g., Figure 8), I actually encounter such expression data. Silhouette score enables us to discuss the DGE results as well as SC dendrograms more objectively.

As shown in my study, Silhouette score is useful for exploring data with predefined group labels. It would help provide both an objective evaluation of SC dendrograms and insights

into the DGE results with regard to the compared groups. The use of this measure would enable a more objective discussion about the SC result in terms of the groups.

Acknowledgments

I am very grateful to Professor Shimizu for taking a Ph.D. course under his supervision. He is a very gentleman and kind person. Under his supervision, I can freely brainstorm and explore some interesting research topic. I also thank Associate Professor Kadota who enabled me to publish my first academic paper in a very short time. I learned a lot of knowledge about the RNA-seq analysis by taking his lecture and reading his homepage. Also, I thank Associate Professor Nakamura for giving me some basic bioinformatics training. Through this, I also have learned python programming and basic machine leaning knowledge. I thank Professor Terada for his advice in the seminar on my research. I did have benefited a lot from them. Thanks to Assistant Professor Moriwaki for creating a very comfortable laboratory environment. I can focus on my research.

I am very grateful to Dr. Sun Jianqiang, for giving me a lot of helpful suggestions when I determine the research field. He also helped me solve many bugs in the R programming. I feel pleasure to publish an article with him on the academic journals. Thanks to the many colleagues in the research lab. Through book reading seminars, I have learned a lot about machine learning and expanded my research horizons. Thanks to the former secretary, Miss Shoji, and the current secretary, Miss Watarai giving me much help on school affairs. Also thanks to Miss Miura and Miss Terada for helping me a lot on my lectures and TA in The Agricultural Bioinformatics Research Unit.

Finally, I sincerely appreciate the support and encouragement from my family and relatives, especially my parents. It is a very difficult decision to pursue a Ph.D. Thanks for their unconditional supports, I can stick to it now.

Additional files

Additional file 1-1 P_{DEG} results for Blekhman's RNA-seq count data

A vs. B	1%	5%	10%	20%	30%	40%
Interspecific comparison						
HSF vs. PTF	3.16%	5.56%	7.56%	11.06%	14.31%	17.95%
HSF vs. PTM	3.31%	6.16%	8.72%	12.79%	17.11%	21.08%
HSM vs. PTF	3.93%	6.58%	8.50%	12.00%	15.68%	18.98%
HSM vs. PTM	3.73%	6.71%	9.58%	13.69%	17.80%	22.27%
HSF vs. RMF	11.30%	17.59%	21.74%	28.10%	33.26%	38.02%
HSF vs. RMM	7.98%	12.98%	16.85%	22.80%	27.47%	32.44%
HSM vs. RMF	12.03%	18.75%	22.92%	29.35%	34.71%	39.58%
HSM vs. RMM	7.88%	12.78%	16.82%	22.44%	27.46%	32.68%
PTF vs. RMF	9.69%	15.25%	19.75%	25.74%	30.40%	35.33%
PTF vs. RMM	6.99%	11.38%	14.69%	19.94%	24.86%	29.37%
PTM vs. RMF	9.80%	15.95%	20.85%	27.43%	32.70%	37.45%
PTM vs. RMM	6.84%	12.15%	16.44%	22.24%	27.42%	32.47%
Intraspecific comparison						
HSF vs. HSM	0.02%	0.04%	0.07%	0.11%	0.22%	0.27%
PTF vs. PTM	0.02%	0.14%	0.17%	0.40%	0.49%	0.78%
RMF vs. RMM	0.03%	0.04%	0.08%	0.19%	0.34%	0.57%

(a) P_{DEG} values at various FDR thresholds (1%, 5%, 10%, 20%, 30%, and 40% FDR). The values at 10% FDR were the same as those shown in Figure 4.

Additional file 1-2 $P_{trueDEG}$ results for Blekhman's RNA-seq count data

A vs. B	1%	5%	10%	20%	30%	40%
Interspecific comparison						
HSF vs. PTF	3.13%	5.28%	6.80%	8.85%	10.02%	10.77%
HSF vs. PTM	3.28%	5.85%	7.84%	10.23%	11.98%	12.65%
HSM vs. PTF	3.89%	6.25%	7.65%	9.60%	10.97%	11.39%
HSM vs. PTM	3.70%	6.38%	8.62%	10.95%	12.46%	13.36%
HSF vs. RMF	11.19%	16.71%	19.56%	22.48%	23.28%	22.81%
HSF vs. RMM	7.90%	12.33%	15.17%	18.24%	19.23%	19.47%
HSM vs. RMF	11.91%	17.81%	20.63%	23.48%	24.30%	23.75%
HSM vs. RMM	7.80%	12.14%	15.14%	17.95%	19.22%	19.61%
PTF vs. RMF	9.60%	14.48%	17.78%	20.59%	21.28%	21.20%
PTF vs. RMM	6.92%	10.81%	13.22%	15.95%	17.40%	17.62%
PTM vs. RMF	9.70%	15.15%	18.76%	21.94%	22.89%	22.47%
PTM vs. RMM	6.77%	11.55%	14.80%	17.79%	19.20%	19.48%
Intraspecific comparison						
HSF vs. HSM	0.02%	0.04%	0.06%	0.09%	0.15%	0.16%
PTF vs. PTM	0.02%	0.13%	0.16%	0.32%	0.34%	0.47%
RMF vs. RMM	0.03%	0.04%	0.07%	0.15%	0.24%	0.34%

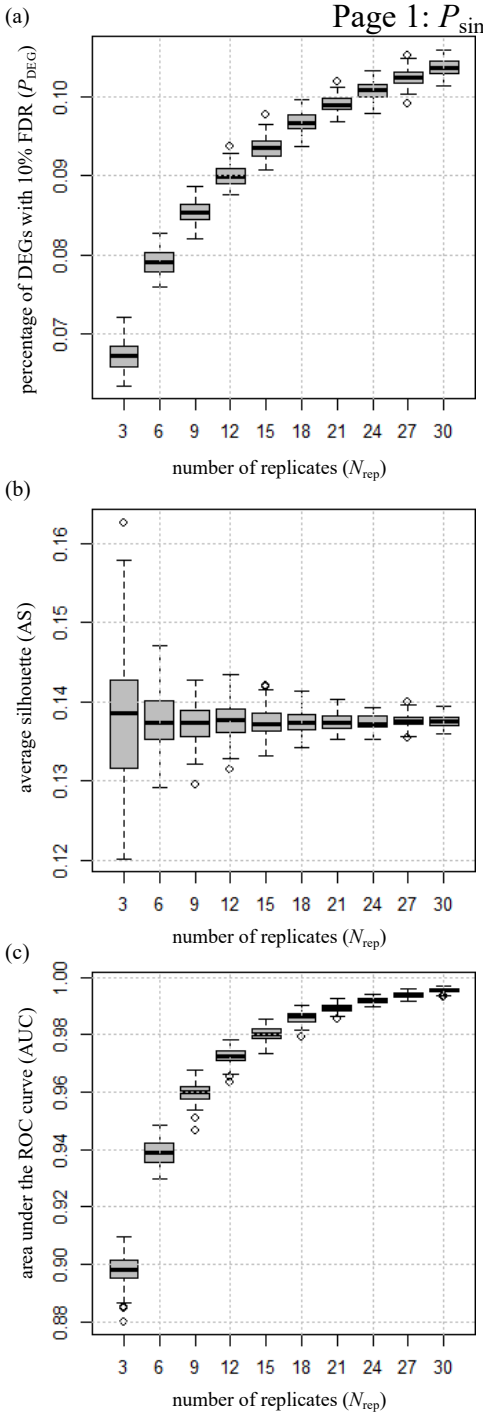
(b) Percentages of true DEGs ($P_{trueDEG}$), defined as $P_{DEG} \times (1 - \text{FDR threshold})$, at corresponding FDR thresholds shown in (a).

Additional file 1-3 Silhouette indices for Blekhman's RNA-seq count data.

A vs. B	i = A1	i = A2	i = A3	i = B1	i = B2	i = B3	average si (AS)
Interspecific comparison							
HSF vs. PTF	0.405	0.439	0.370	0.406	0.373	0.342	0.389
HSF vs. PTM	0.484	0.543	0.454	0.304	0.223	0.354	0.394
HSM vs. PTF	0.423	0.444	0.472	0.408	0.411	0.344	0.417
HSM vs. PTM	0.521	0.509	0.532	0.334	0.241	0.324	0.410
HSF vs. RMF	0.603	0.642	0.615	0.618	0.604	0.583	0.611
HSF vs. RMM	0.645	0.672	0.637	0.444	0.416	0.473	0.548
HSM vs. RMF	0.617	0.627	0.669	0.622	0.605	0.574	0.619
HSM vs. RMM	0.639	0.664	0.690	0.435	0.409	0.471	0.551
PTF vs. RMF	0.632	0.576	0.561	0.604	0.602	0.566	0.590
PTF vs. RMM	0.651	0.615	0.603	0.424	0.417	0.441	0.525
PTM vs. RMF	0.550	0.449	0.534	0.654	0.642	0.599	0.571
PTM vs. RMM	0.574	0.509	0.559	0.486	0.461	0.497	0.514
Intraspecific comparison							
HSF vs. HSM	-0.096	-0.111	0.062	0.024	-0.031	0.036	-0.019
PTF vs. PTM	0.172	0.117	0.193	-0.080	-0.123	-0.092	0.031
RMF vs. RMM	0.174	0.247	0.144	-0.436	0.050	-0.262	-0.014

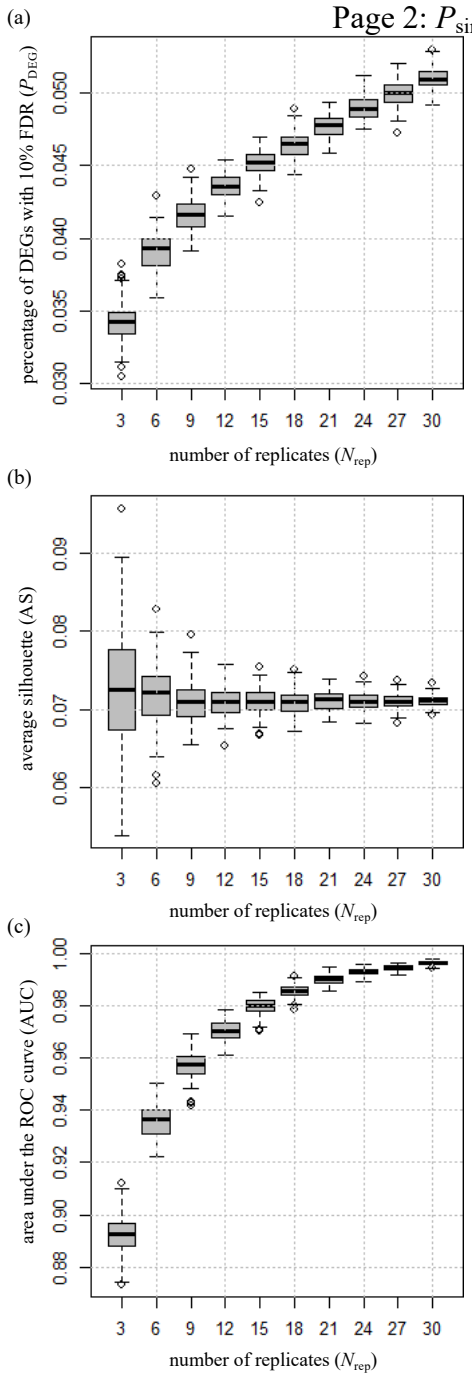
(c) Silhouette indices (si) for each sample i and the average (AS). The sample names (A1, A2, A3, B1, B2, or B3) for i correspond to those shown in Figure 4.

Additional file 2 Effects of *Nrep* on parameter estimates (simulated count data).



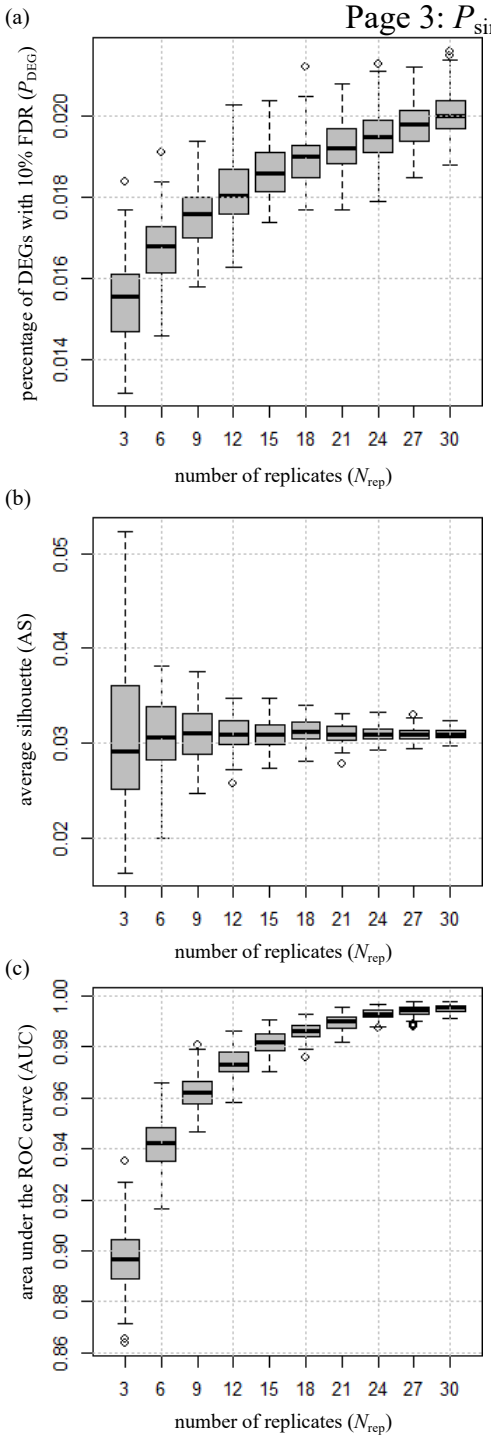
Bootstrapping results for simulated data under different P_{simDEG} values are shown: P_{simDEG} = 10% (Page 1).

Additional file 2 Effects of *Nrep* on parameter estimates (simulated count data).



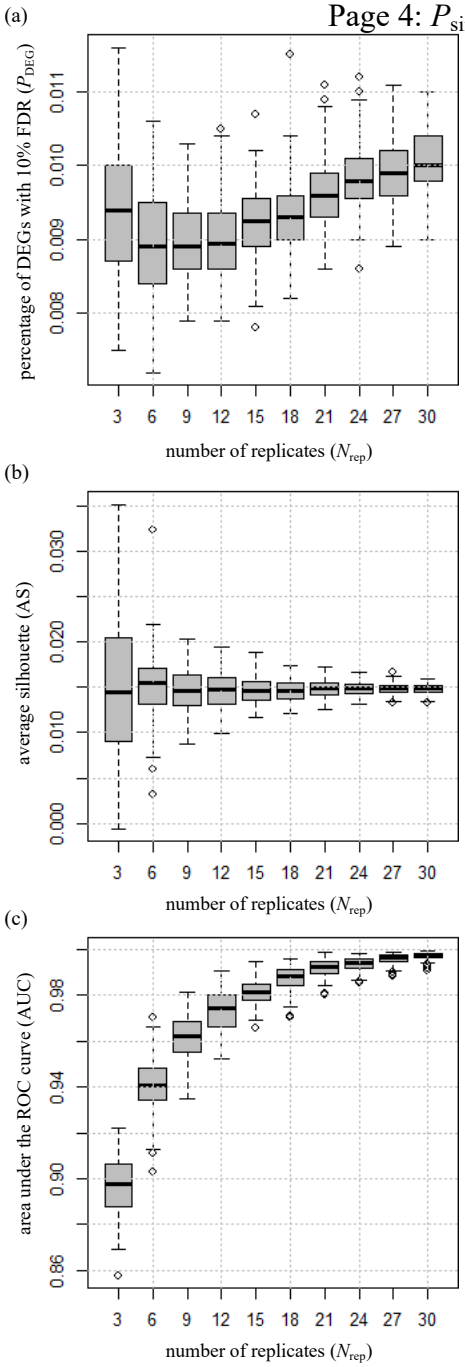
Bootstrapping results for simulated data under different P_{simDEG} values are shown: P_{simDEG} = 5% (Page 2).

Additional file 2 Effects of *Nrep* on parameter estimates (simulated count data)



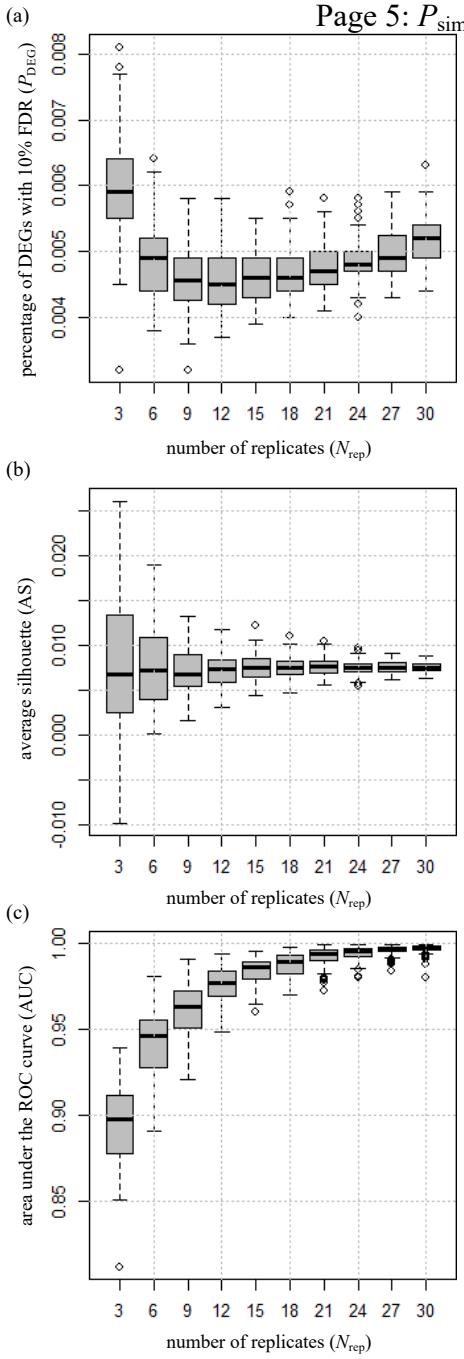
Bootstrapping results for simulated data under different P_{simDEG} values are shown: P_{simDEG} = 2% (Page 3).

Additional file 2 Effects of *Nrep* on parameter estimates (simulated count data)



Bootstrapping results for simulated data under different P_{simDEG} values are shown: P_{simDEG} = 1% (Page 4).

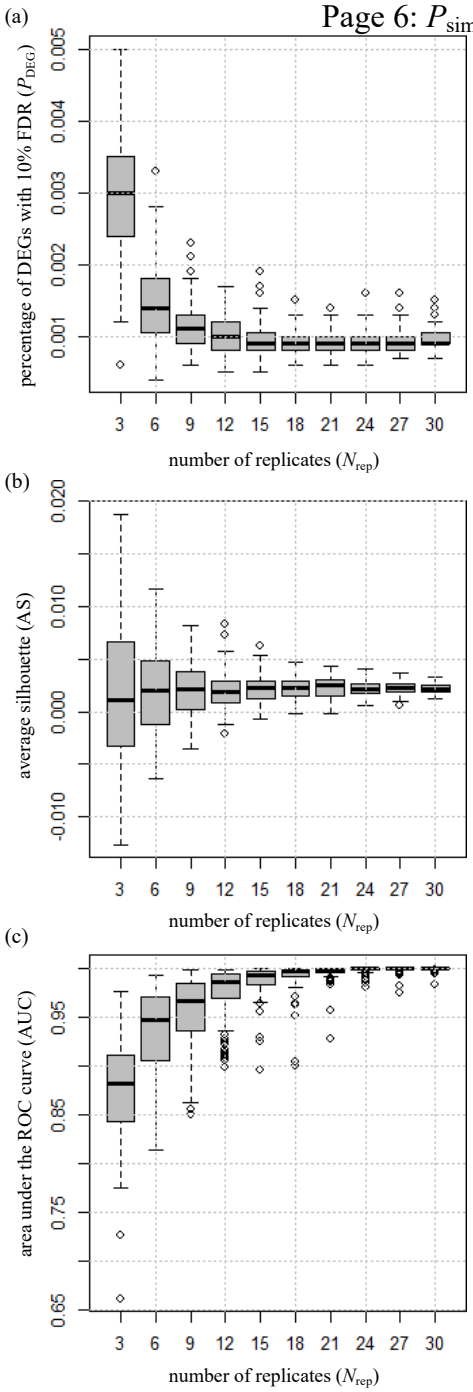
Additional file 2 Effects of *Nrep* on parameter estimates (simulated count data).



Bootstrapping results for simulated data under different P_{simDEG} values are shown:

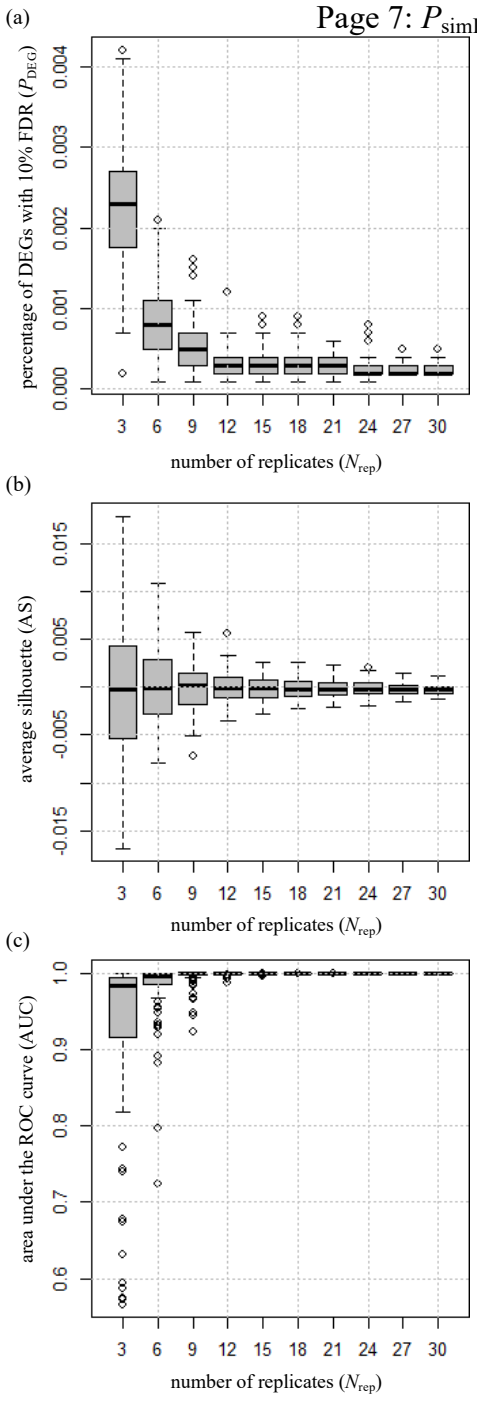
P_{simDEG} = 0.5% (Page 5).

Additional file 2 Effects of *Nrep* on parameter estimates (simulated count data).



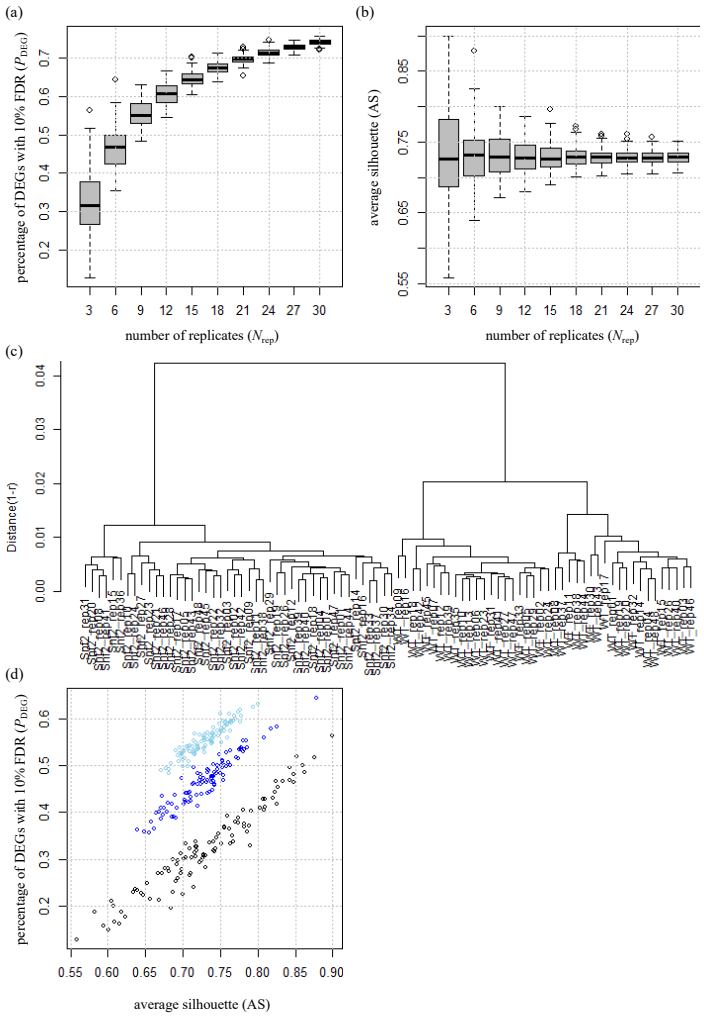
Bootstrapping results for simulated data under different P_{simDEG} values are shown: $P_{simDEG} = 0.1\%$ (Page 6).

Additional file 2 Effects of *Nrep* on parameter estimates (simulated count data).



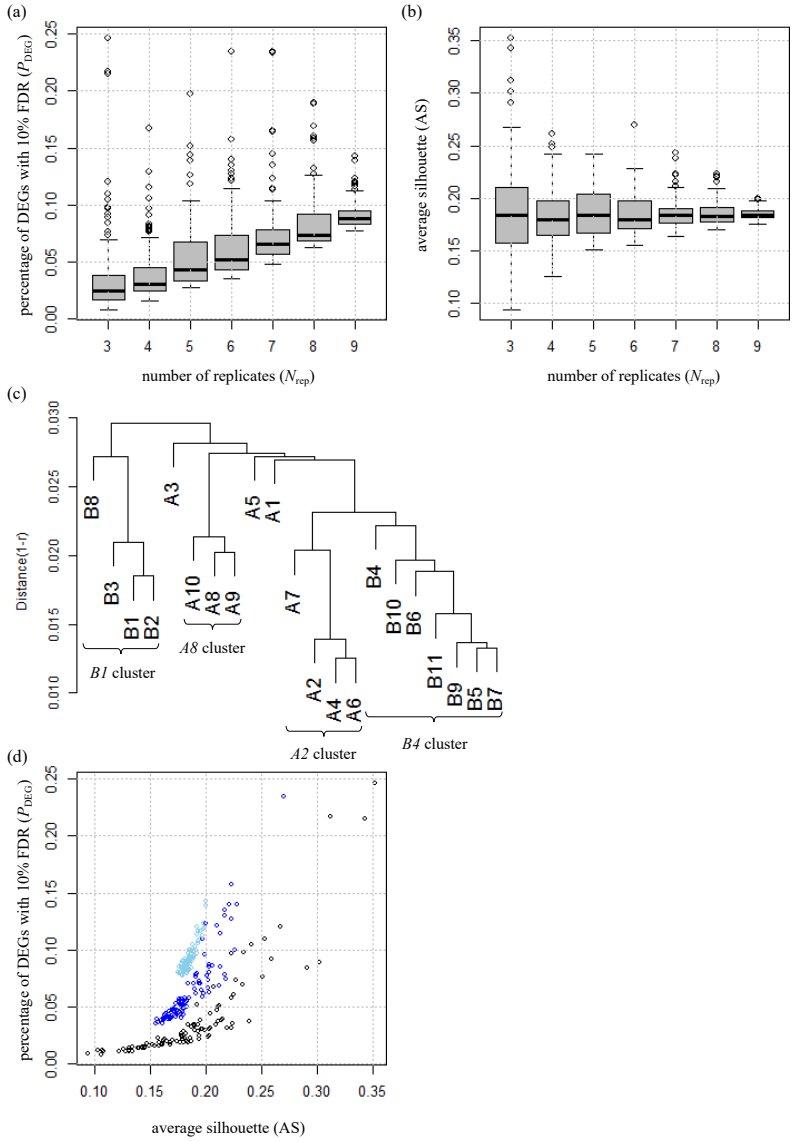
Bootstrapping results for simulated data under different P_{simDEG} values are shown: P_{simDEG} = 0.02% (Page 7).

Additional file 3 Results for Schurch's RNA-seq count data



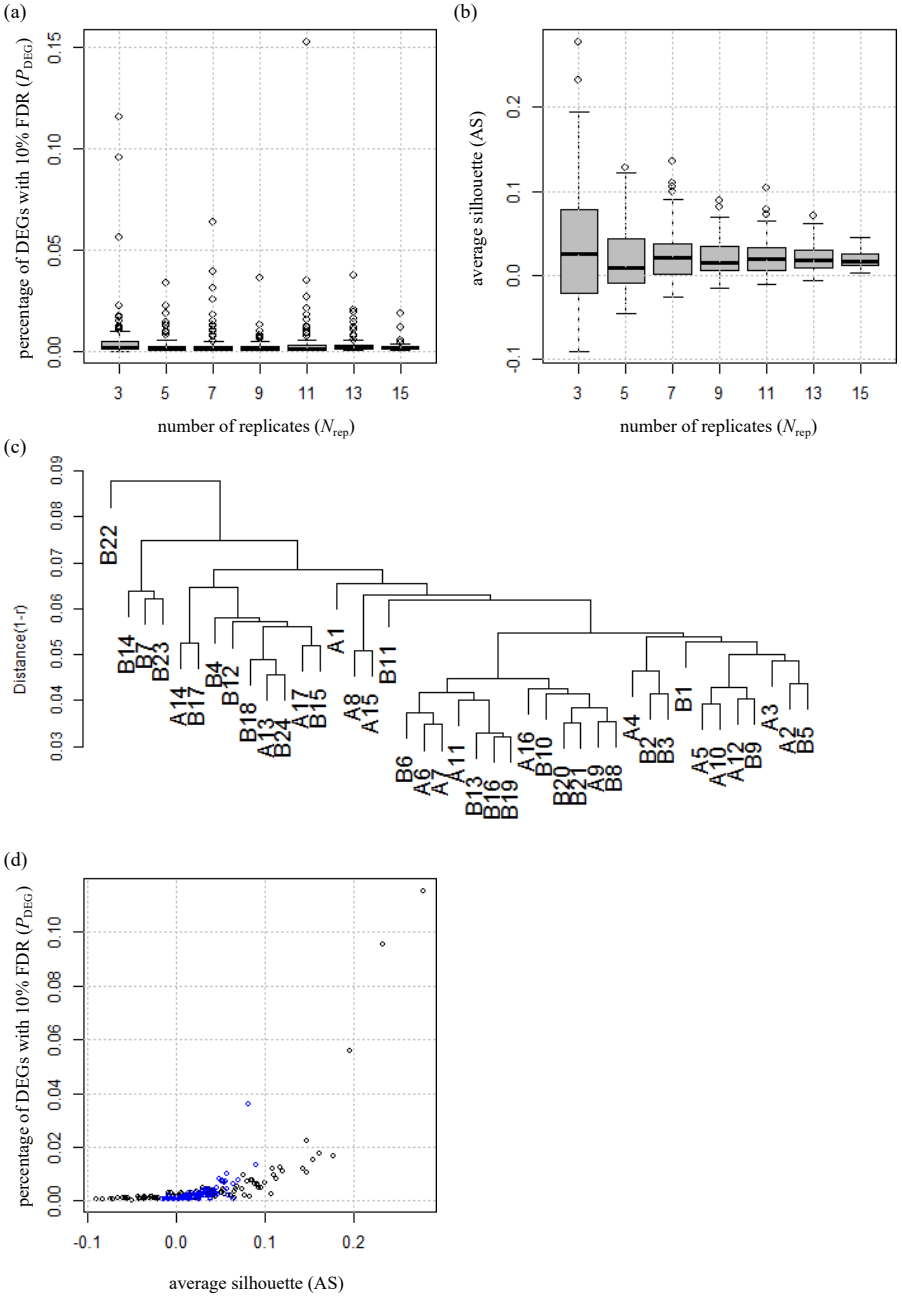
For (a–b), Bootstrapping results for Schurch data comparing 42 wild-type samples and 44 *Δsnf2* mutant samples are shown. Legends are the same as those in Fig 5. (c) HSC dendrogram. Two distinct clusters, a wild-type cluster (right side) and *Δsnf2* mutant cluster (left side), can be seen. The intra-group distances within 42 wild-type samples and 44 *Δsnf2* mutant samples were 0.0144 and 0.0084, respectively. (d) Scatter plots of P_{DEG} vs. AS at *Nrep* = 3 (black), 6 (blue), and 9 (sky blue).

Additional file 4 Results for Bottomly's RNA-seq count data.



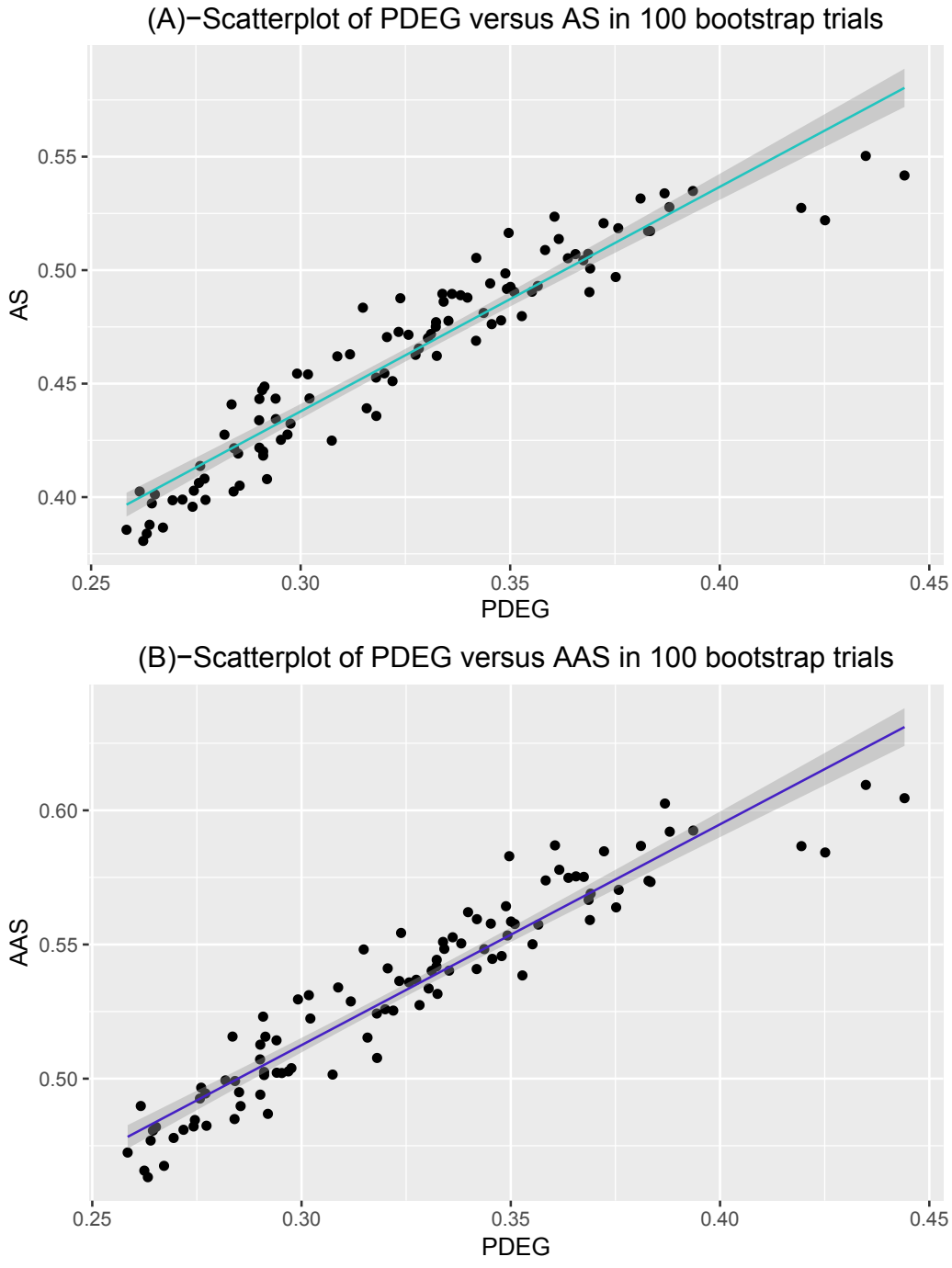
For (a–b), Bootstrapping results for Bottomly data comparing 10 C57BL/6J strains (A1, A2 …, A10) vs. 11 DBA/2 J strains (B1, B2, …, B11) are shown. (c) HSC dendrogram. For explanation, four clusters are defined in (d) the HSC dendrogram: the B1 cluster (consisting of B1, B2, B3, and B8), A8 cluster (A8, A9, and A10), A2 cluster (A2, A4, and A6), and B4 cluster (B4, B5, B6, B7, B9, B10, and B11). (d) Scatter plots of P_{DEG} vs. AS at $N_{rep} = 3$ (black), 6 (blue), and 9 (sky blue).

Additional file 5 Results for Cheung's RNA-seq count data



For (a–b), Bootstrapping results for Cheung data comparing 17 females (A1, A2, …, A17) vs. 24 males (B1, B2, …, B24) are shown. (c) HSC dendrogram. (d) Scatter plots of P_{DEG} vs. AS at *Nrep* = 3 (black), 6 (blue), and 9 (sky blue).

Additional file 6 Scatter plot of results in simulation data (FDR=0.1)



Abbreviations

AS	Average silhouette
AUC	the area under the ROC curve
BAT	Brown adipose tissue
BR	Biological replicate
BV	Biological variation
DE	Differential expression
DEG	Differentially expressed gene
DGE	Differential gene expression
F	Female
FDR	False discovery rate
H0	null hypothesis
HS	*Homo sapiens*
HSC	Hierarchical sample clustering
LIV	Liver tissue
M	Male
NB	Negative binomial (distribution or model)
Nrep	Number of (biological) replicates
P_{DEG}	Percentage of estimated DEGs (satisfying basically 10% FDR) by TCC
P_{simDEG}	Percentage of DEGs when generating simulated data
$P_{trueDEG}$	Percentage of true DEGs defined as $P_{DEG} \times$ (1.0 – FDR threshold)
PT	*Pan troglodytes*
RM	*Rhesus macaques*
ROC	Receiver operating characteristic

SC	Sample clustering
TCC	Tag Count Comparison
V	Variance
WAT	White adipose tissue
AAS	Average of Average Silhouette scores in group pair-wise way

References

[1] F. Crick, “Central dogma of molecular biology,” *Nature*, vol. 227, no. 5258, pp. 561–563, 1970.

[2] Z. Wang, M. Gerstein, and M. Snyder, “RNA-Seq: A revolutionary tool for transcriptomics,” *Nature Reviews Genetics*, vol. 10, no. 1. pp. 57–63, 2009.

[3] P. Baldi and G. W. Hatfield, *DNA Microarrays and Gene Expression*. Cambridge University Press, 2009.

[4] T. Speed, *Statistical analysis of gene expression microarray data*. Chapman and Hall/CRC, 2003.

[5] G. Kerr, H. J. Ruskin, M. Crane, and P. Doolan, “Techniques for clustering gene expression data,” *Computers in Biology and Medicine*, vol. 38, no. 3. pp. 283–293, 2008.

[6] G. Passador-Gurgel, W. P. Hsieh, P. Hunt, N. Deighton, and G. Gibson, “Quantitative trait transcripts for nicotine resistance in Drosophila melanogaster,” *Nat. Genet.*, vol. 39, no. 2, pp. 264–268, 2007.

[7] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, “Quantitative monitoring of gene expression patterns with a complementary DNA microarray,” *Science*, vol. 270, no. 5235, pp. 467–470, 1995.

[8] E. Hubbell, W.-M. Liu, and R. Mei, “Robust estimators for expression analysis,” *Bioinformatics*, vol. 18, no. 12, pp. 1585–1592, 2002.

[9] G. Liu *et al.*, “NetAffx: Affymetrix probesets and annotations,” *Nucleic Acids Res.*, vol. 31, no. 1, pp. 82–86, 2003.

[10] R. A. Irizarry *et al.*, “Exploration, normalization, and summaries of high density oligonucleotide array probe level data,” *Biostatistics*, vol. 4, no. 2, pp. 249–264,

2003.

[11] W. Pan, “A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments,” *Bioinformatics*, vol. 18, no. 4, pp. 546–554, 2002.

[12] X. Cui, J. T. G. Hwang, J. Qiu, N. J. Blades, and G. A. Churchill, “Improved statistical tests for differential gene expression by shrinking variance components estimates,” *Biostatistics*, vol. 6, no. 1, pp. 59–75, 2005.

[13] V. G. Tusher, R. Tibshirani, and G. Chu, “Significance analysis of microarrays applied to the ionizing radiation response,” *Proc. Natl. Acad. Sci. USA*, vol. 98, no. 9, pp. 5116–5121, 2001.

[14] G. K. Smyth, J. Michaud, and H. S. Scott, “Use of within-array replicate spots for assessing differential expression in microarray experiments,” *Bioinformatics*, vol. 21, no. 9, pp. 2067–2075, 2005.

[15] K. R. Kimberley and M. B. Stephen, “RNA Sequencing and Analysis,” *Cold Spring Harb. Protoc.*, vol. 11, no. 2, pp. 951–969, 2015.

[16] S. Andrews, “FastQC : A quality control tool for high throughput sequence data,” *Babraham Bioinformatics*, 2010. [Online]. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[17] A. M. Bolger, M. Lohse, and B. Usadel, “Trimmomatic: a flexible trimmer for Illumina sequence data,” *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, 2014.

[18] C. Trapnell, L. Pachter, and S. L. Salzberg, “TopHat: discovering splice junctions with RNA-Seq,” *Bioinformatics*, vol. 25, no. 9, pp. 1105–1111, 2009.

[19] A. Dobin *et al.*, “STAR: ultrafast universal RNA-seq aligner,” *Bioinformatics*, vol. 29, no. 1, pp. 15–21, 2013.

[20] B. Langmead and S. L. Salzberg, “Fast gapped-read alignment with Bowtie 2,” *Nat. Methods*, vol. 9, no. 4, pp. 357–359, 2012.

[21] D. Kim, B. Langmead, and S. L. Salzberg, “HISAT: A fast spliced aligner with low memory requirements,” *Nat. Methods*, vol. 12, no. 4, pp. 357–360, 2015.

[22] Y. Liao, G. K. Smyth, and W. Shi, “featureCounts: an efficient general purpose program for assigning sequence reads to genomic features,” *Bioinformatics*, vol. 30, no. 7, pp. 923–930, 2014.

[23] S. Anders, P. T. Pyl, and W. Huber, “HTSeq-a Python framework to work with high-throughput sequencing data,” *Bioinformatics*, vol. 31, no. 2, pp. 166–169, 2015.

[24] S. Tarazona, F. García-Alcalde, J. Dopazo, A. Ferrer, and A. Conesa, “Differential expression in RNA-seq: A matter of depth,” *Genome Res.*, vol. 21, no. 12, pp. 2213–2223, 2011.

[25] M. D. Robinson and A. Oshlack, “A scaling normalization method for differential expression analysis of RNA-seq data,” *Genome Biol.*, vol. 11, no. 3, p. R25, 2010.

[26] S. Anders and W. Huber, “Differential expression analysis for sequence count data,” *Genome Biol.*, vol. 11, no. 10, p. R106, 2010.

[27] S. Lamarre *et al.*, “Optimization of an RNA-Seq Differential Gene Expression Analysis Depending on Biological Replicate Number and Library Size,” *Front. Plant Sci.*, vol. 9, 2018.

[28] M. D. Robinson and G. K. Smyth, “Small-sample estimation of negative binomial dispersion, with applications to SAGE data,” *Biostatistics*, vol. 9, no. 2, pp. 321–332, 2007.

[29] D. J. McCarthy, Y. Chen, and G. K. Smyth, “Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation,” *Nucleic Acids Res.*, vol. 40, no. 10, pp. 4288–4297, 2012.

[30] B. Hochberg, “Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing,” *J. R. Stat. Soc.*, vol. 57, no. 1, pp. 289–300, 1995.

[31] P. A. Jaskowiak, R. J. G. B. Campello, and I. G. Costa, “On the selection of appropriate distances for gene expression data clustering,” *BMC Bioinformatics*, vol. 15, no. 2, p. S2, 2014.

[32] P. D. Reeb, S. J. Bramardi, and J. P. Steibel, “Assessing dissimilarity measures for sample-based hierarchical clustering of RNA sequencing data using plasmode datasets,” *PLoS One*, vol. 10, no. 7, p. e0132310, 2015.

[33] A. A. Alizadeh *et al.*, “Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling,” *Nature*, vol. 403, no. 6769, pp. 503–511, 2000.

[34] K. Kadota, T. Nishiyama, and K. Shimizu, “A normalization strategy for comparing tag count data,” *Algorithms Mol. Biol.*, vol. 7, p. 5, 2012.

[35] U. Alon *et al.*, “Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays,” *Proc. Natl. Acad. Sci.*, vol. 96, no. 12, pp. 6745–6750, 1999.

[36] K. Kadota, R. Miki, H. Bono, K. Shimizu, Y. Okazaki, and Y. Hayashizaki, “Preprocessing implementation for microarray (PRIM): an efficient method for processing cDNA microarray data,” *Physiol. Genomics*, vol. 4, no. 3, pp. 183–188, 2001.

[37] Y. Qin, J. Pan, M. Cai, L. Yao, and Z. Ji, “Pattern genes suggest functional connectivity of organs,” *Sci. Rep.*, vol. 6, p. 26501, 2016.

[38] F. Danielsson, T. James, D. Gomez-Cabrero, and M. Huss, “Assessing the consistency of public human tissue RNA-seq data sets,” *Brief. Bioinform.*, vol. 16, no. 6, pp. 941–949, 2015.

[39] M. Tang, J. Sun, K. Shimizu, and K. Kadota, “Evaluation of methods for differential expression analysis on multi-group RNA-seq count data,” *BMC Bioinformatics*, vol. 16, no. 1, p. 360, 2015.

[40] P. J. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.

[41] I. Gat-Viks, R. Sharan, and R. Shamir, “Scoring clustering solutions by their biological relevance,” *Bioinformatics*, vol. 19, no. 18, pp. 2381–2389, 2003.

[42] S. Bandyopadhyay, A. Mukhopadhyay, and U. Maulik, “An improved algorithm for clustering gene expression data,” *Bioinformatics*, vol. 23, no. 21, pp. 2859–2865, 2007.

[43] E. Lord, A. B. Diallo, and V. Makarenkov, “Classification of bioinformatics workflows using weighted versions of partitioning and hierarchical clustering algorithms,” *BMC Bioinformatics*, vol. 16, no. 1, p. 68, 2015.

[44] R Core Team, “R: A language and environment for statistical computing,” *Vienna: R Foundation for Statistical Computing*, 2016. [Online]. Available: http://www.r-project.org/.

[45] R. C. Gentleman *et al.*, “Bioconductor: open software development for computational biology and bioinformatics,” *Genome Biol.*, vol. 5, no. 10, p. R80, 2004.

[46] J. Sun, T. Nishiyama, K. Shimizu, and K. Kadota, “TCC: an R package for comparing tag count data with robust normalization strategies,” *BMC*

Bioinformatics, vol. 14, no. 1, p. 219, 2013.

[47] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: A Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2009.

[48] Y. Di, D. W. Schafer, J. S. Cumbie, and J. H. Chang, "The NBP negative binomial model for assessing differential gene expression from RNA-Seq," *Stat. Appl. Genet. Mol. Biol.*, vol. 10, no. 1, 2011.

[49] R. Blekhman, J. C. Marioni, P. Zumbo, M. Stephens, and Y. Gilad, "Sex-specific and lineage-specific alternative splicing in primates," *Genome Res.*, vol. 20, no. 2, pp. 180–189, 2010.

[50] M. Gierliński *et al.*, "Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment," *Bioinformatics*, vol. 31, no. 22, pp. 3625–3630, 2015.

[51] D. Bottomly *et al.*, "Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays," *PLoS One*, vol. 6, no. 3, p. e17820, 2011.

[52] A. C. Frazee, B. Langmead, and J. T. Leek, "ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets," *BMC Bioinformatics*, vol. 12, no. 1, p. 449, 2011.

[53] V. G. Cheung *et al.*, "Polymorphic cis-and trans-regulation of human gene expression," *PLoS Biol.*, vol. 8, no. 9, p. e1000480, 2010.

[54] N. Kolesnikov *et al.*, "ArrayExpress update-simplifying data submissions," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D1113–D1116, 2015.

[55] A. Kauffmann *et al.*, "Importing arrayexpress datasets into r/bioconductor,"

Bioinformatics, vol. 25, no. 16, pp. 2092–2094, 2009.

[56] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry, "affy-analysis of Affymetrix GeneChip data at the probe level," *Bioinformatics*, vol. 20, no. 3, pp. 307–315, 2004.

[57] M. Kohl and H.-P. Deigner, "Preprocessing of gene expression data by optimally robust estimators," *BMC Bioinformatics*, vol. 11, no. 1, p. 583, 2010.

[58] T. J. Hardcastle and K. A. Kelly, "baySeq: empirical Bayesian methods for identifying differential expression in sequence count data," *BMC Bioinformatics*, vol. 11, no. 1, p. 422, 2010.

[59] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biol.*, vol. 15, no. 12, p. 550, 2014.

[60] M.-A. Dillies *et al.*, "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis," *Brief. Bioinform.*, vol. 14, no. 6, pp. 671–683, 2012.

[61] E. Maza, "In Papyro comparison of TMM (edgeR), RLE (DESeq2), and MRN normalization methods for a simple two-conditions-without-replicates RNA-Seq experimental design," *Front. Genet.*, vol. 7, p. 164, 2016.

[62] N. J. Schurch *et al.*, "How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?," *RNA*, vol. 22, no. 6, pp. 839–851, 2016.

[63] K. Kadota and K. Shimizu, "Evaluating methods for ranking differentially expressed genes applied to microArray quality control data," *BMC Bioinformatics*, vol. 12, no. 1, p. 227, 2011.

[64] Y. Nakai *et al.*, “Up-Regulation of Genes Related to the Ubiquitin-Proteasome System in the Brown Adipose Tissue of 24-h-Fasted Rats,” *Biosci. Biotechnol. Biochem.*, vol. 72, no. 1, pp. 139–148, 2008.

[65] R. Miki *et al.*, “Delineating developmental and metabolic pathways in vivo by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays,” *Proc. Natl. Acad. Sci. USA*, vol. 98, no. 5, pp. 2199–2204, 2001.

[66] M. E. Ritchie *et al.*, “limma powers differential expression analyses for RNA-sequencing and microarray studies,” *Nucleic Acids Res.*, vol. 43, no. 7, pp. e47–e47, 2015.

[67] F. M. Giorgi, A. M. Bolger, M. Lohse, and B. Usadel, “Algorithm-driven artifacts in median polish summarization of microarray data,” *BMC Bioinformatics*, vol. 11, no. 1, p. 553, 2010.

[68] M. A. F. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, “A review of novelty detection,” *Signal Processing*, vol. 99, pp. 215–249, 2014.

[69] K. Kadota, D. Tominaga, Y. Akiyama, and K. Takahashi, “Detecting outlying samples in microarray data: A critical assessment of the effect of outliers on sample classification,” *Chem-Bio Informatics J.*, vol. 3, no. 1, pp. 30–45, 2003.

[70] K. Kadota, J. Ye, Y. Nakai, T. Terada, and K. Shimizu, “ROKU: a novel method for identification of tissue-specific genes,” *BMC Bioinformatics*, vol. 7, no. 1, p. 294, 2006.