Biological Procedures Online

RESEARCH Open Access

Silhouette Scores for Arbitrary Defined Groups in Gene Expression Data and Insights into Differential Expression Results

Shitao Zhao, Jianqiang Sun, Kentaro Shimizu and Koji Kadota*

Abstract

Background: Hierarchical Sample clustering (HSC) is widely performed to examine associations within expression data obtained from microarrays and RNA sequencing (RNA-seq). Researchers have investigated the HSC 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.

Results: To objectively evaluate the degree of separation between groups of interest in the HSC dendrogram, we propose to use *Silhouette* scores. Silhouettes was originally developed as a graphical aid for the validation of data clusters. It provides a measure of how well a sample is classified when it was assigned to a cluster by according to both the tightness of the clusters and the separation between them. It ranges from 1.0 to − 1.0, and a larger value for the average silhouette (AS) over all samples to be analyzed indicates a higher degree of *cluster* separation. The basic idea to use an AS is to replace the term *cluster* by *group* when calculating the scores. We investigated the validity of this score using simulated and real data designed for differential expression (DE) analysis. We found that larger (or smaller) AS values agreed well with both higher (or lower) degrees of separation between different groups and higher percentages of differentially expressed genes (P_{DEG}). We also found that the AS values were generally independent on the number of replicates (N_{rep}). Although the P_{DEG} values depended on N_{rep}, we confirmed that both AS and P_{DEG} values were close to zero when samples in the data showed an intermingled nature between the groups in the HSC dendrogram.

Conclusion: Silhouettes is useful for exploring data with predefined group labels. It would help provide both an objective evaluation of HSC dendrograms and insights into the DE results with regard to the compared groups.

Keywords: Hierarchical sample clustering, Bioinformatics, Differential expression analysis, Silhouettes

Background

High-throughput technologies, including microarrays and RNA-seq, are widely used to monitor genome-wide expression levels in samples of interest and to compare expression patterns in different groups or conditions (e.g., healthy vs. tumor tissue samples) [1–6]. The latter, comparative analyses are often termed differential expression (DE) analyses and the identification of differentially expressed genes or transcripts (DEGs) is a common approach in studies of the molecular basis of traits [7, 8]. RNA-seq is now the main method used to obtain expression data, but microarrays have provided important insights (e.g., [9]). A main difference between the two technologies is the nature of the expression data: microarrays yield continuous signal intensities, while RNA-seq data provides discrete counts [10, 11]. To appropriately manipulate these expression data, several specialized models (e.g., the negative binomial (NB) model for RNA-seq count data [12–18]) have been proposed.

Another common approach for expression analyses is sample clustering (SC) based on similarity in expression patterns [19–21]. Utilizing its unsupervised characteristic, SC has been used to (i) detect previously unrecognized subtypes of cancer [22, 23], (ii) detect outliers (i.e., outlying samples) [24], (iii) represent overall similarities in expression among various organs [25, 26], and (iv) perform sanity

* Correspondence: kadota@bi.a.u-tokyo.ac.jp
Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, Japan

checks to verify expected clustering patterns [27]. 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 DE analysis (say, "DE data") roughly corresponds to the DE result when the groups for the DE analysis are evaluated with respect to the SC result [8]. If individual groups form distinct sub-clusters, where each sub-cluster consists only of members (or samples) in the particular group, DE 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 DE data and the percentages of DEGs (P_{DEG}) remains lacking [8].

Silhouettes is a graphical aid for the interpretation and validation of cluster analysis [28]. In SC, silhouettes provide 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. Silhouettes has been successfully used after clustering as a cluster validity measure [20, 29–31].

In this paper, we propose to use *Silhouette* 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. We here focus on single-factor DE data where only one grouping criterion is primarily of interest in relation to the DE result. We evaluated the relationship among SC results, DE results, and AS values, using both simulated and real expression data (RNA-seq and microarrays). We found silhouettes (i.e., AS values) to provide a relevant measure for the degrees of separation between groups of interest in SC results. We also found a positive correlation between AS values and DE results.

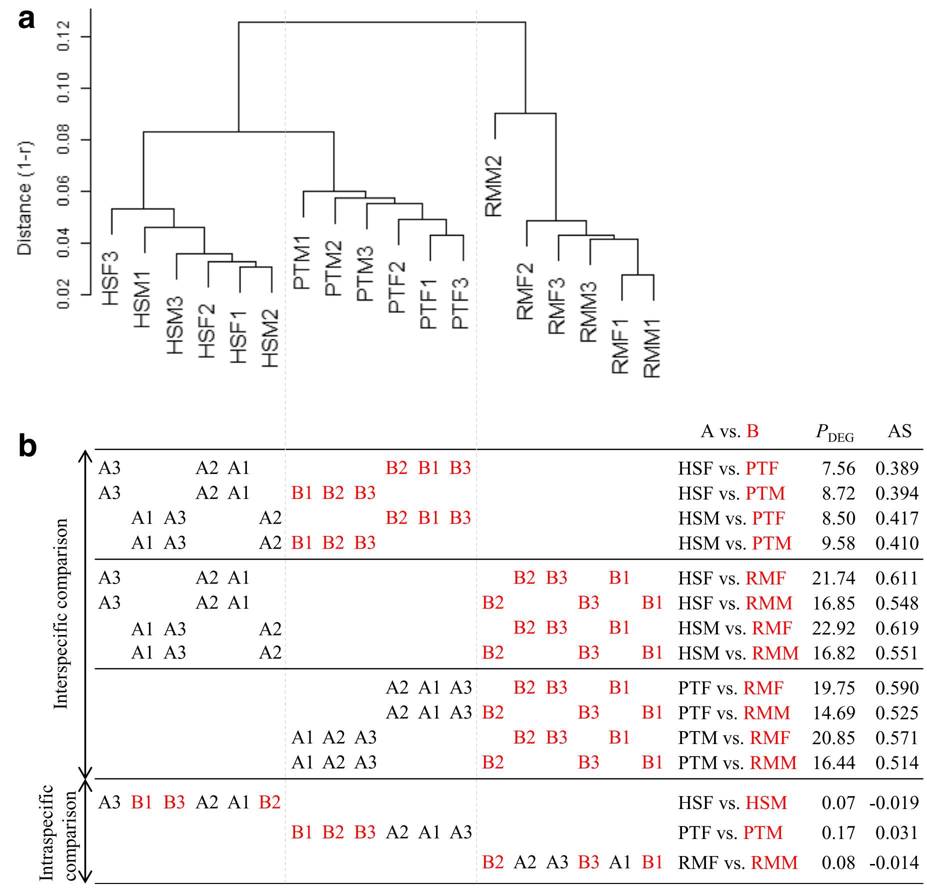
Results

In DE 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 (i) counts for RNA-seq data or (ii) the signal intensity for microarray data. Our previous observation of the positive correlation between SC and DE results [8] 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 biological replicates (BRs)) [32]. The analysis was performed using a hierarchical SC (HSC) algorithm and a DE pipeline, both of which are provided in the R/Bioconductor package TCC [33–35]. TCC implements a robust normalization strategy (called DEGES [36]) that uses functions provided in four widely used packages (baySeq [37], edgeR [38, 39], DESeq [40], and DESeq2) [15]. For simplicity and/or the algorithmic advantage [41, 42], we only used TCC for the DE analysis of RNA-seq data. Specifically, we used the default DE pipeline (*iDEGES/edgeR-edgeR* in [33] and *EEE-E* in [8]). When performing HSC for all input data, we 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, we filtered out genes with zero counts (or signals) in all samples. 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 data yielded 17,886 genes after the zero-count filtering and DE analyses were performed. After *unique* filtering, 16,560 genes were obtained, and HSC was performed using these genes. For simplicity, we 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, we use the terms *samples* and *replicates* interchangeably. Our primary interest was to investigate the applicability of Silhouette for the objective evaluation of gene expression data based on arbitrary grouping criteria. By using silhouettes (i.e., AS values) as a relevant measure for the group differentiation in the HSC results, we re-evaluated our previous observations (i.e., the positive correlation between HSC and DE results) [8].

Representative Relationship between HSC and DE Results with AS

We first demonstrate the relationship between HSC and DE 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) [32]. 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 1a shows the HSC dendrogram based on a correlation distance (1 - *r*) metric and average-linkage



Sample positions	A vs. B	P_{DEG}	AS
A3 A2 A1 B2 B1 B3	HSF vs. PTF	7.56	0.389
A3 A2 A1 B1 B2 B3	HSF vs. PTM	8.72	0.394
A1 A3 A2 B2 B1 B3	HSM vs. PTF	8.50	0.417
A1 A3 A2 B1 B2 B3	HSM vs. PTM	9.58	0.410
A3 A2 A1 B2 B3 B1	HSF vs. RMF	21.74	0.611
A3 A2 A1 B2 B3 B1	HSF vs. RMM	16.85	0.548
A1 A3 A2 B2 B3 B1	HSM vs. RMF	22.92	0.619
A1 A3 A2 B2 B3 B1	HSM vs. RMM	16.82	0.551
A2 A1 A3 B2 B3 B1	PTF vs. RMF	19.75	0.590
A2 A1 A3 B2 B3 B1	PTF vs. RMM	14.69	0.525
A1 A2 A3 B2 B3 B1	PTM vs. RMF	20.85	0.571
A1 A2 A3 B2 B3 B1	PTM vs. RMM	16.44	0.514
A3 B1 B3 A2 A1 B2	HSF vs. HSM	0.07	-0.019
B1 B2 B3 A2 A1 A3	PTF vs. PTM	0.17	0.031
B2 A2 A3 B3 A1 B1	RMF vs. RMM	0.08	-0.014

Fig. 1 Relationship between the shape of HSC and DE 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** DE results from a total of 15 two-group comparisons with three replicates. The DE 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

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), we 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 1b shows 15 DE 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 [32] and can primarily be explained by the interspecific distances shown in Fig. 1a. 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 (Fig. 1a).

Silhouette [28] has been successfully employed to estimate the appropriate number of clusters for gene expression data [20, 29–31]. Although Silhouette is generally used for the validation of clustering results, we here employ it independently from clustering. Technically, the term *cluster* is replaced with *group* in the silhouette calculation procedure. For each sample *i*, let u_i be the average distance between *i* and all other samples within the same *group* (e.g., group A). Let v_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 (v_i - u_i)/max(u_i, v_i). The index s_i ranges from − 1 to 1; it is positive if $u_i < v_i$, zero if $u_i = v_i$, and negative if $u_i > v_i$. 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 (Additional file 1a; right hand in Fig. 1b). The potential applicability of the silhouette unrelated to clustering has been described in the original study [28]. However, to the best of our 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.

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, $P_{\mathrm{DEG}} = 21.74\%$ and AS = 0.611, while for the HSF vs. RMM comparison, $P_{\mathrm{DEG}} = 16.85\%$ 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, we confirmed that the general trends for the 15 two-group comparisons were the same at 1%, 5%, 10%, 20%, 30%, and 40% FDR thresholds (Additional file 1b). 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, $P_{\mathrm{DEG}} = 0.78\%$ at a 40% FDR for the PTF vs. PTM comparison indicates that 0.78 × 0.4 = 0.31% are non-DEGs, and the remaining 0.78 × (1.0–0.4) = 0.47% are, at least statistically, true DEGs. In our 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% (Additional file 1c). These results indicate that differences in P_{DEG} values with respect to the FDR threshold are not important.

Based on our visual evaluation, the AS values effectively represented the overall relationship between groups of interest in the HSC analysis (shown in Fig. 1a). We 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 (N_{rep}) increases [43], suggesting that the correlation is influenced by N_{rep}.

Effects of the Number of Replicates (N_{rep}) on Parameter Estimates

We next investigated the effects of N_{rep} 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 ($N_{\mathrm{rep}} = 40$), 10,000 total genes, of which 20% were DEGs (2000 DEGs and 8000 non-DEGs; $P_{\mathrm{simDEG}} = 20\%$), the levels of DE were four-fold in individual groups, and the proportions of DEGs up-regulated in individual groups were the same (i.e., 1000 DEGs are up-regulated in group A). For a total of 80 samples (A1, A2, ..., A40, B1, B2, ..., B40), we obtained $P_{\mathrm{DEG}} = 21.0\%$ 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 DE pipeline [7, 8, 33, 36]. 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 DE 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 $P_{\mathrm{simDEG}} = 20\%$).

The DE pipeline was used to examine subsets from the *baseline* matrix with 40 replicates per group ($N_{\mathrm{rep}} =$ 40). Bootstrap resampling was performed 100 times at $N_{\mathrm{rep}} = 3, 6$, ..., and 30 (without replacement). Consistent the previous observations [43], the average P_{DEG} values increased as a function of N_{rep} (Fig. 2a). However, such an increasing trend was not observed for AS (Fig. 2b). This result indicates that the the silhouette (i.e., AS) is independent of N_{rep}. Note that the P_{DEG} value approached to the true value ($P_{\mathrm{simDEG}} = 20\%$) as N_{rep} increased (Fig. 2a). In general, the DE 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} (Fig. 2c), this interpretation can be trusted in this case.

Next, the effects of N_{rep} under different P_{simDEG} conditions ($P_{\mathrm{simDEG}} = 10\%$, 5%, 2%, 1%, 0.5%, 0.1%, and 0.02%) were investigated. We confirmed that P_{DEG}, but not on AS, is dependent on N_{rep} (Additional file 2). Different from the condition shown in Fig. 2 ($P_{\mathrm{simDEG}} = 20\%$), however, we observed a transition in the distribution of P_{DEG} values at around $P_{\mathrm{simDEG}} = 1\%$. Although the P_{DEG} value monotonously increased as N_{rep} 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 N_{rep} increased. These results indicate that more accurate DE results can be obtained as N_{rep} increases, irrespective of the true percentages of DEGs in the data.

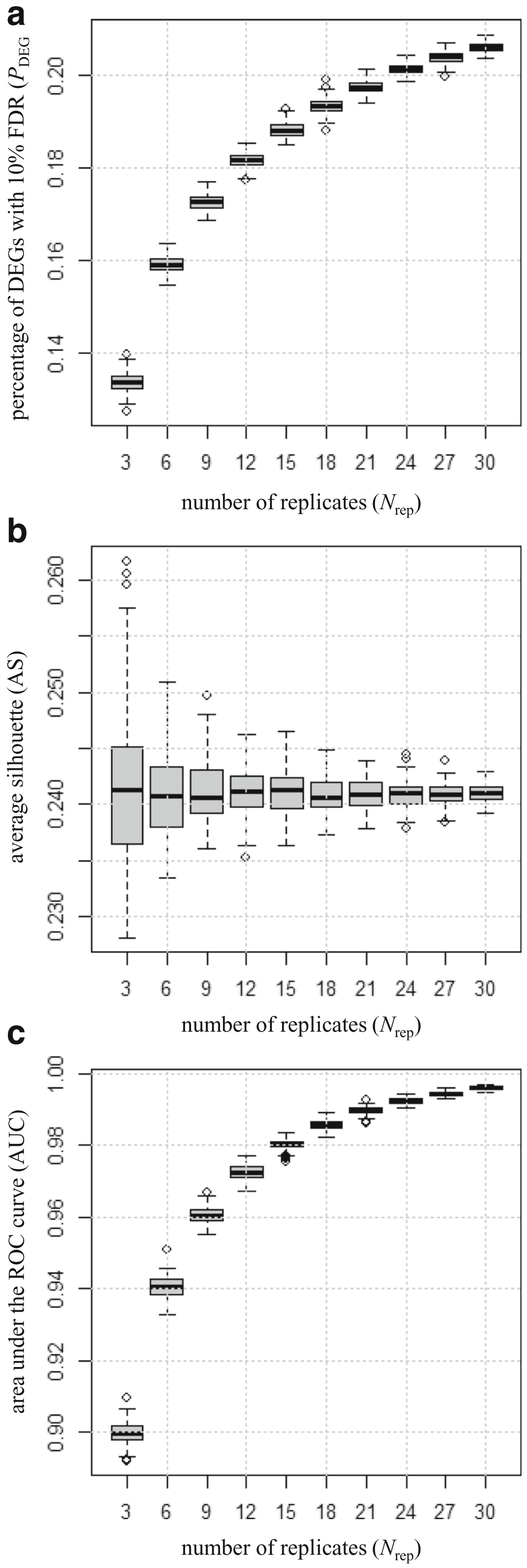


Fig. 2 Effects of N_{rep} on parameter estimates (simulated data). Bootstrapping results (100 iterations) from simulated RNA-seq data consisting of 10,000 genes × 80 samples with $P_{\mathrm{simDEG}} = 20\%$ are shown. Vertical axes for the boxplots indicate: (**a**) P_{DEG}, (**b**) AS, and (**c**) AUC values. Horizontal axes indicate the N_{rep} values (3, 6, …, 30). It can be seen that P_{DEG} and AUC values increase as a function of N_{rep}, but AS values do not

A similar analysis was performed using another real RNA-seq dataset consisting of 7126 genes × 96 samples [43, 44]. Ten outlier samples were rejected, following the original study [43], and subsequent zero-count filtering of the original data yielded 6885 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), we obtained $P_{\mathrm{DEG}} = 78.1\%$ and AS = 0.7289. Note that the AUC value could not be calculated for the data because, different from simulated data, we do not know which genes are true DEGs. We investigated the effects of N_{rep} on parameter estimates. The results were quite similar to those obtained using simulated data (shown in Fig. 2), i.e., P_{DEG} was dependent on N_{rep}, 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 [39] (Fig. 1a in [43]). This is quite reasonable because the DE pipeline implemented in TCC can be viewed as an iterative edgeR pipeline [8].

Relationships between P_{DEG} and AS Values

Next, we investigated the relationships between P_{DEG} and AS values under a fixed N_{rep} of 3. Figure 3 shows the results for (a) Schurch, (b) simulated, and (c) the mixture. For simulated data, we examined 19 P_{simDEG} conditions from 5% (black in Fig. 3b) to 0.95 (red in Fig. 3b). Overall, there was a strong positive correlation between P_{DEG} and AS values in this condition (Fig. 3c). 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 N_{rep}, while AS does not (Fig. 2; Additional file 3). In other words, the regression coefficients depend on N_{rep}. 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 (H_0: A = B). The acceptance of H_0 (AS = 0) indicates there are no or few DEGs in the two-group data ($P_{\mathrm{DEG}} = 0$). In this sense, AS could be used as

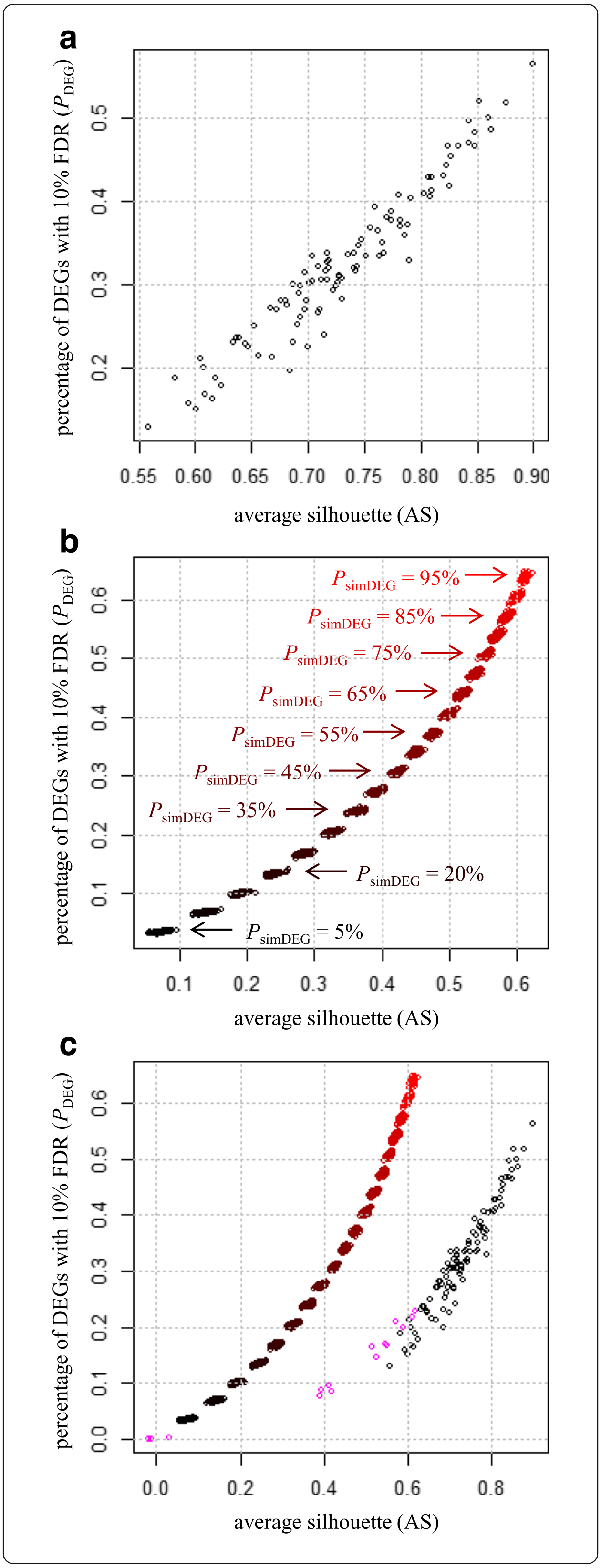


Fig. 3 Relationship between P_{DEG} and AS values. Scatter plots of P_{DEG} vs. AS at $N_{rep} = 3$ are shown. (**a**) Schurch data. The scatter plot shows a detailed relationship between P_{DEG} and AS values for Schurch data at $N_{rep} = 3$ (Additional file 3a and 3b). (**b**) simulated data under $P_{simDEG} = 5\%$, …, 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 $N_{rep} = 3$ shown in Fig. 2b. (**c**) the results for the mixture as well as the Blekhman data including 15 two-group comparisons shown in Fig. 1b (magenta)

helpful information for the interpretation of DE results, especially when only a few statistically significant DEGs are obtained.

It should be noted that the distribution shown in Fig. 3c (right panel) differs substantially from the distribution for real data (Blekhman [32] and Schurch [43]) 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 (Fig. 3c). 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; Fig. 1a). According to a previous study [43], 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.

Analyses of two Additional Real RNA-Seq Datasets

We further investigated two other real RNA-seq datasets available at the ReCount website [45]. The first dataset (called Bottomly [46]) consisted of 36,536 genes × 21 samples. Briefly, Bottomly et al. studied the expression levels of two common inbred mouse strains used in neuroscience research, i.e., 10 C57BL/6J strains (A1, A2 …, A10) and 11 DBA/2J strains (B1, B2, …, B11). DE analyses (i.e., estimates of P_{DEG} values) were performed using 13,932 genes after zero-count filtering. AS calculations and HSC were performed using 13,133 genes after *unique* filtering.

The results for this dataset comparing 10 vs. 11 samples were $P_{\mathrm{DEG}} = 11.0\%$ at a 10% FDR threshold and AS = 0.1872. Regarding the effects of $N_{\mathrm{rep},}$ we observed similar trends to those obtained using the Schurch data (Fig. 3a), i.e., P_{DEG} increased as a function of N_{rep}, while AS did not (Additional file 4a–b). Despite similar trends, the values obtained for the Bottomly data were clearly lower than those for the Schurch data. For example, the Bottomly data (and Schurch data) showed on average $P_{\mathrm{DEG}} =$ 3.81% (32.4%) and AS = 0.1874 (0.7306) at $N_{\mathrm{rep}} = 3$ (Additional files 3 and 4). These findings suggested that the P_{DEG} values were lower for Bottomly than for Schurch because the AS value for Bottomly was lower than that of Schurch.

In general, high (or low) AS values indicate clear (or unclear) separation of groups. The HSC dendrogram for the Bottomly data showed relatively unclear separation of two groups (Additional file 4c; AS = 0.1872) compared to the separation of groups for the Schurch data (Additional file 3c; AS = 0.7289). Nevertheless, as also implied by the positive AS value, the degree of inter-group separation for the Bottomly data was not random. For example, by visual inspection, we identified four clusters, each of which consisted only of samples within the same group (Additional file 4c). These clusters primarily explained the estimated P_{DEG} values and positive AS values. The highest values for P_{DEG} (=24.7%) and AS (= 0.3524) among 100 trials at $N_{\mathrm{rep}} = 3$ were obtained when comparing (A3, A4, A6) vs. (B1, B3, B8). This is reasonable because five of the six samples (except for A3) were members of either the *A2* or *B1* cluster. Relatively high values can be obtained by comparing two groups in which all members belong to either the *A2* or *B1* clusters. Indeed, we observed $P_{\mathrm{DEG}} = 34.8\%$ and AS = 0.4623 when comparing (A2, A4, A6) vs. (B1, B2, B3), though this comparison was not included in the 100 original trials shown in Additional file 4a–b.

Different from the Schurch data where the impact of sampling effects shrunk as N_{rep} increased (Additional file 3a), we did not observe shrinkage for the Bottomly data around $N_{\mathrm{rep}} = 3$–7 (Additional file 4a). This can also be explained by the four clusters mentioned above. For example, the highest values for P_{DEG} (=23.5%) and AS (= 0.2699) among 100 trials at $N_{\mathrm{rep}} = 6$ were obtained when comparing (A2, A3, A4, A6, A7, A9) vs. (B1, B2, B3, B4, B8, B10). All eight samples in the *A2* and *B1* clusters was included in the comparison. Additionally, a comparison between the two clusters, i.e., (A2, A4, A6, A7) vs. (B1, B2, B3, B8), yielded $P_{\mathrm{DEG}} = 31.8\%$ and AS = 0.3701. Accordingly, the decreases in P_{DEG} (=31.8% to 23.5%) and AS (0.3701 to 0.2699) values by the addition of four samples (A3, A9, B4, and B10) not included in the two clusters are reasonable. We observed that the impact of sampling effects tends to shrink as N_{rep} (> 7) increases. This is probably because the maximum number of samples in the four clusters is seven for the *B4* cluster; the addition of samples not included in the cluster can contribute to decreases in the P_{DEG} and AS values.

The second dataset (called Cheung [47]) consisted of 52,580 genes × 41 samples. Briefly, Cheung et al. studied the expression levels of human B-cells using 17 females (A1, A2, ..., A17) and 24 males (B1, B2, ..., B24). The DE analyses (i.e., estimates of P_{DEG} values) were performed using 12,410 genes after zero-count filtering. AS calculation and HSC were performed using 11,738 genes after *unique* filtering. The results for this dataset comparing 17 vs. 24 samples were $P_{\mathrm{DEG}} = 0.169\%$, SNR = 1.023, and AS = 0.0118. The values were considerably lower than those obtained for both the Schurch and Bottomly data and were similar to those for the three sex comparisons (Fig. 1b). This result is intuitively reasonable, as gene expression levels in B-cells are not expected to differ greatly between females and males.

We did not observe an increasing trend for P_{DEG} values as N_{rep} increased (Additional file 5a). The average P_{DEG} values for 100 trials at $N_{\mathrm{rep}} = 3$, 5, 7, 9, 11, 13, and 15 were 0.631%, 0.291%, 0.399%, 0.254%, 0.492%, 0.325%, and 0.219%, respectively. These values as well as the distribution were quite similar to those obtained from simulated data with $P_{\mathrm{simDEG}} = 0.5\%$ (Page 5 in Additional file 2a). This result suggests that the increase of N_{rep} does not contribute to the increase of P_{DEG} when AS is near 0. Since AS is independent of N_{rep}, no or few DEGs ($P_{\mathrm{DEG}} < 1\%$) would be obtained when AS < 0.1 for count data (Additional file 5b). The intermingled nature of the HSC dendrogram for the Cheung data (Additional file 5c) also supports this inference; AS can be utilized to interpret the DE results.

Analysis of two Microarray Datasets

We finally investigated two microarray datasets obtained using the Affymetrix Rat Genome 230 2.0 Array (GPL1355). The first dataset (called Nakai [4]) 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. We 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), we applied three algorithms (MAS [48], RMA [49], and RobLoxBioC [50]). Different from RNA-seq data represented as integer counts, microarray data are expressed as continuous signals and in most cases are log-transformed. We therefore applied a specialized DE pipeline for microarray data provided in the package limma [51], 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; Additional file 6a). Since the experimental design and the HSC dendrogram were very similar to those of the Blekhman data (Fig. 1), these microarray data can be regarded as the counterpart.

We performed 15 two-group comparisons with four BRs for each group, i.e., (A1, A2, A3, A4) vs. (B1, B2, B3, B4). Overall, we observed highly similar trends for the Nakai data and the Blekhman data (Additional file 6b). 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), we 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.

We observed similar results for RobLoxBioC-quantified data and relatively dissimilar results for RMA-quantified data. In particular, for the RMA-quantified data, we 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 [52]. 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 [53, 54]. Nevertheless, we 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 [55]) 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 (Additional file 7). 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 DE results for both RNA-seq and microarray data, especially when no or few DEGs were obtained.

We 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 (Additional file 7). *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 [52]. 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.

HSC dendrograms of the merged data provided several insights (Additional file 8). 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) [25–27]. 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 [21, 27]. 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 ∞. We 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.

Discussion

In this study, we proposed to use silhouettes (i.e., AS values) as an objective measure for the degrees of separation between groups of interest based on expression data. To our knowledge, the use of AS independent from HSC is the first practical application in the field of expression analysis. Our 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 N_{rep}, while P_{DEG} values obtained from DE analysis are fundamentally dependent on N_{rep}; and (iii) there is a positive correlation between AS and P_{DEG} values under a fixed N_{rep}. It is not necessary to estimate P_{DEG} from AS values because DE results (including P_{DEG}) can directly be obtained via the DE pipeline. The AS provides helpful information for interpreting DE results as well as HSC results.

Based on the current results, we conclude that our 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 Additional files 7 and 8). 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 [43, 44]. We manually eliminated the ten outliers as determined in the original study [44] and analyzed 86 *clean* samples in this dataset (Fig. 3a; Additional file 3). The values obtained without outliers ($P_{\mathrm{DEG}} = 78.1\%$ and AS = 0.7289) were clearly higher than those with outliers ($P_{\mathrm{DEG}} = 74.7\%$ and AS = 0.6530), indicating the importance of developing methods for the automatic detection of outliers [55, 56]. Our preliminary analysis for the original data using an existing method [57] successfully detected nine of the ten *true* outliers as well as three false positives. We obtained a promising result ($P_{\mathrm{DEG}} = 77.6\%$ and AS = 0.7301) using the remaining 84 samples. Rational removal of outlying samples would yield better DE results. We expect that AS would help objective evaluation of the changes in the DE results accompanying outlier removal.

In practice, Silhouettes can be utilized as supporting information to interpret DE results, especially when no or few DEGs are obtained. As demonstrated by several examples (e.g., Additional file 7), we actually encounter such expression data. Silhouettes enables us to discuss the DE results as well as SC dendrograms more objectively.

Conclusion

Silhouettes is useful for exploring data with predefined group labels. It would help provide both an objective evaluation of SC dendrograms and insights into the DE 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.

Methods

Most of the analyses were performed using R (ver. 3.3.2) [34] and Bioconductor [35]. The versions of major R packages used in the study were TCC ver. 1.14.0, edgeR ver. 3.16.5, ROC ver. 1.50.0, cluster ver. 2.0.5, affy ver. 1.44.0, and RobLoxBioC ver. 0.9. R-codes are provided in Additional file 10.

Simulated Data

The two-group simulated data were produced using the "simulateReadCounts" function in TCC [33]. The variance (V) of the NB distribution can generally be modeled as $V = \mu + \phi\mu^2$. 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 BRs for both treated and non-treated samples) in [58]. 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 DE analysis and HSC. These data were used to obtain Fig. 2, Fig. 3, and Additional file 2.

Four RNA-Seq Data

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) [32]. 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). These data were used to obtain Fig. 1, Fig. 3, Table 1, and Additional file 1.

Schurch's yeast data were obtained from the GitHub website (https://github.com/bartongroup/profDGE48/tree/master/Preprocessed_data) [43]. 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. These data were used to obtain Fig. 3 and Additional file 3.

Bottomly's mouse data were [46] obtained from the ReCount website (http://bowtie-bio.sourceforge.net/recount/countTables/bottomly_count_table.txt) [45] and consisted of 36,536 genes × 21 samples. These data were used to obtain Additional file 4.

Cheung's human data [47] were obtained from the ReCount website (http://bowtie-bio.sourceforge.net/recount/countTables/cheung_count_table.txt) [45] and consisted of

52,580 genes × 41 samples. These data were used to obtain Additional file 5.

Two Rat Microarray Data

Nakai's probe-level data (.CEL files) were obtained from the ArrayExpress website [59] through an R package ArrayExpress [60] by applying "GSE7623." The MAS-quantified data were obtained using the *mas5* function in the R/Bioconductor package affy [61]. Expression signals less than 1 were set to 1 and were subsequently $\log_2$-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 $\log_2$-transformed. The RobLoxBioC-quantified data were obtained using the *robloxbioc* function in the R package RobLoxBioC [50]. The expression signals less than 1 were set to 1 and were subsequently $\log_2$-transformed. These data were used to obtain Additional files 6 and 8.

Kamei's probe-level data (.CEL files) were obtained from the ArrayExpress website [59] using the R package ArrayExpress [60] by applying "GSE30533." The subsequent procedures were the same as those described for the Nakai data. These data were used to obtain Additional files 7 and 8.

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).

HSC and DE Analyses

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 [33]. The DE analysis was performed using three functions (calcNormFactors, estimateDE, and getResult) with default options which use functions in the package edgeR [39]. The genes were ranked in ascending order according to *p*-values. The ranks were used to calculate AUC values when analyzing simulated data (Fig. 2 and Additional file 2). 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%).

Calculation of Average Silhouette (AS) Values

The AS values were calculated using the *silhouette* function in the package cluster. Examples of the procedure to estimate AS values are given in Additional file 9.

Additional files

Additional file 1: Detailed results for Blekhman's RNA-seq count data. (a) Silhouette indices (s_i) 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 Fig. 1b. (b) 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 Fig. 1b. (c) Percentages of true DEGs (P_{trueDEG}), defined as $P_{\mathrm{DEG}} \times (1 - \mathrm{FDR\ threshold})$, at corresponding FDR thresholds shown in (b). (XLSX 19 kb)

Additional file 2: Effects of N_{rep} on parameter estimates (simulated count data). Bootstrapping results for simulated data under different P_{simDEG} values are shown: $P_{\mathrm{simDEG}} = 10\%$ (Page 1), 5% (Page 2), 2% (Page 3), 1% (Page 4), 0.5% (Page 5), 0.1% (Page 6), and 0.02% (Page 7). Other legends are the same as those in Fig. 2. (PPTX 110 kb)

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. 2. (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 $N_{\mathrm{rep}} = 3$ (black), 6 (blue), and 9 (sky blue). (PPTX 65 kb)

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 $N_{\mathrm{rep}} = 3$ (black), 6 (blue), and 9 (sky blue). (PPTX 55 kb)

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 $N_{\mathrm{rep}} = 3$ (black), 6 (blue), and 9 (sky blue). (PPTX 58 kb)

Additional file 6: 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 $N_{\mathrm{rep}} = 4$ are shown: MAS-quantified data (Page 1), RMA-quantified data (Page 2), and RobLoxBioC-quantified data (Page 3). (PPTX 76 kb)

Additional file 7: 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. (PPTX 49 kb)

Additional file 8: 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). (PPTX 62 kb)

Additional file 9: Examples of AS estimates for two- and three-group data. The procedures for analyzing Nakai's MAS-quantified data consisting of 31,099 probesets × 24 samples are provided. Example 1 compares three-group data with four BRs, 4 *BAT_fed* samples vs. 4 *WAT_fed* samples vs. 4 *LIV_fed* samples, with AS = 0.460. Example 2 compares three-group data with two BRs, "*BAT_fed1* and *2*" vs. "*WAT_fed1* and *2*" vs. "*LIV_fed1* and *2*," with AS = 0.438. Example 3 compares three-group data with two BRs, "*BAT_fed1* and *BAT_fas1*" vs. "*BAT_fed2* and *BAT_fas2*" vs. "*BAT_fed3* and *BAT_fas3*," with AS = – 0.185. Example 4 compares two-group data with four BRs, 4 *BAT_fed* samples vs. 4 *WAT_fed* samples, with AS = 0.374. Example 5 compares two-group data with four BRs, 4 *BAT_fed* samples vs. 4 *LIV_fed* samples, with AS = 0.657. (R 3 kb)

Additional file 10: R-codes for analyses. This zipped file includes a total of 23 R-code files. Results can be obtained by executing scripts in the order of the serial numbers *XX* in the filename "rcode_*XX*_…" Note that two files ("rcode_08_Add6_pre.R" and "rcode_10_Add7_pre.R") must be executed using R ver. 3.1.3 (affy ver. 1.44.0) instead of R ver. 3.3.2 (affy ver. 1.52.0). (ZIP 33 kb)

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; F: Female; FDR: False discovery rate; H_0: null hypothesis; HS: *Homo sapiens*; HSC: Hierarchical sample clustering; LIV: Liver tissue; M: Male; NB: Negative binomial (distribution or model); N_{rep}: 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 $P_{\mathrm{DEG}} \times (1.0 - \mathrm{FDR\ threshold})$; PT: *Pan troglodytes*; RM: *Rhesus macaques*; ROC: Receiver operating characteristic; SC: Sample clustering; TCC: Tag Count Comparison; *V*: Variance; WAT: White adipose tissue

Acknowledgments

Not applicable.

Funding

This work was supported by JSPS KAKENHI Grant Number JP15K06919.

Availability of Data and Materials

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). Schurch's yeast data were obtained from the GitHub website (https://github.com/bartongroup/profDGE48/tree/master/Preprocessed_data). Bottomly's mouse data were obtained from the ReCount website (http://bowtie-bio.sourceforge.net/recount/countTables/bottomly_count_table.txt). Cheung's human data were obtained from the ReCount website (http://bowtie-bio.sourceforge.net/recount/countTables/cheung_count_table.txt). Nakai's probe-level data (.CEL files) were obtained from the ArrayExpress website through an R package ArrayExpress by applying "GSE7623." Kamei's probe-level data (.CEL files) were obtained using the R package ArrayExpress by applying "GSE30533." The R-codes for obtaining current results are provided in Additional file 10.

Authors' Contributions

SZ performed the analysis and drafted the manuscript. JS maintained the package TCC and provided the critical comments. KS supervised the critical discussion and refined the paper. KK refined the paper, confirmed the analysis results, and led this project. All the authors read and approved the final manuscript.

Ethics Approval and Consent to Participate

Not applicable.

Consent for Publication

Not applicable.

Competing Interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 14 September 2017 Accepted: 12 January 2018
Published online: 01 March 2018

References

1. Miki R, Kadota K, Bono H, Mizuno Y, Tomaru Y, Carninci P, Itoh M, Shibata K, Kawai J, Konno H, Watanabe S, Sato K, Tokusumi Y, Kikuchi N, Ishii Y, Hamaguchi Y, Nishizuka I, Goto H, Nitanda H, Satomi S, Yoshiki A, Kusakabe M, DeRisi JL, Eisen MB, Iyer VR, Brown PO, Muramatsu M, Shimada H, Okazaki Y, Hayashizaki Y. Delineating developmental and metabolic path ways in vivo by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays. Proc Natl Acad Sci U S A. 2001;98(5):2199–204.
2. Ichikawa Y, Ishikawa T, Takahashi S, Hamaguchi Y, Morita T, Nishizuka I, Yamaguchi S, Endo I, Ike H, Togo S, Oki S, Shimada H, Kadota K, Nakamura S, Goto H, Nitanda H, Satomi S, Sakai T, Narita I, Gejyo F, Tomaru Y, Shimizu K, Hayashizaki Y, Okazaki Y. Identification of genes regulating colorectal carcinogenesis by using the ADMS (algorithm for diagnosing malignant state) method. Biochem Biophys Res Commun. 2002;296(2):497–506.
3. Shimoji T, Kanda H, Kitagawa T, Kadota K, Asai R, Takahashi K, Kawaguchi N, Matsumoto S, Hayashizaki Y, Okazaki Y, Shinomiya K. Clinico-molecular study of dedifferentiation in well-differentiated liposarcoma. Biochem Biophys Res Commun. 2004;314(4):1133–40.
4. Nakai Y, Hashida H, Kadota K, Minami M, Shimizu K, Matsumoto I, Kato H, Abe K. Up-regulation of genes related to the ubiquitin-proteasome system in the brown adipose tissue of 24-h-fasted rats. Biosci Biotechnol Biochem. 2008;72(1):139–48.
5. Kawaoka S, Kadota K, Arai Y, Suzuki Y, Fujii T, Abe H, Yasukochi Y, Mita K, Sugano S, Shimizu K, Tomari Y, Shimada T, Katsuma S. The silkworm W chromosome is a source of female-enriched piRNAs. RNA. 2011;17(12):2144–51.
6. Lin Y, Golovnina K, Chen ZX, Lee HN, Negron YL, Sultana H, Oliver B, Harbison ST. Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual Drosophila Melanogaster. BMC Genomics. 2016;17:28.
7. Kadota K, Shimizu K. Evaluating methods for ranking differentially expressed genes applied to MicroArray quality control data. BMC Bioinformatics. 2011;12:227.
8. Tang M, Sun J, Shimizu K, Kadota K. Evaluation of methods for differential expression analysis on multi-group RNA-seq count data. BMC Bioinformatics. 2015;16:361.
9. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. 2010;11:94.
10. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5(7):621–8.
11. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008;18(9):1509–17.
12. Robinson MD, Smyth GK. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics. 2008;9(2):321–32.
13. Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W, Robinson MD. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat Protoc. 2013;8(9):1765–86.
14. Zhou X, Lindsay H, Robinson MD. Robustly detecting differential expression in RNA sequencing data using observation weights. Nucleic Acids Res. 2014;42(11):e91.
15. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
16. Ching T, Huang S, Garmire LX. Power analysis and sample size estimation for RNA-Seq differential expression. RNA. 2014;20(11):1684–96.
17. Sun X, Dalpiaz D, Wu D, S Liu J, Zhong W, Ma P. Statistical inference for time course RNA-Seq data using a negative binomial mixed-effect model. BMC Bioinformatics. 2016;17(1):324.
18. Hardcastle TJ. Generalized empirical Bayesian methods for discovery of differential data in high-throughput biology. Bioinformatics. 2016;32(2):195–202.
19. de Souto MC, Costa IG, de Araujo DS, Ludermir TB, Schliep A. Clustering cancer gene expression data: a comparative study. BMC Bioinformatics. 2008;9:497.
20. Jaskowiak PA, Campello RJ, Costa IG. On the selection of appropriate distances for gene expression data clustering. BMC Bioinformatics. 2014;15(Suppl 2):S2.
21. Reeb PD, Bramardi SJ, Steibel JP. Assessing dissimilarity measures for sample-based hierarchical clustering of RNA sequencing data using Plasmode datasets. PLoS One. 2015;10(7):e0132310.
22. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling**.** Nature 2000, 403(6769):503–511.
23. Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, Sampas N, Dougherty E, Wang E, Marincola F, Gooden C, Lueders J, Glatfelter A, Pollock P, Carpten J, Gillanders E, Leja D, Dietrich K, Beaudry C, Berens M, Alberts D, Sondak V. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature. 2000;406(6795):536–40.
24. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. Broad patterns of gene expression revealed by clustering analysis of tumor and

normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci U S A. 1999;96(12):6745–50.
25. Kadota K, Miki R, Bono H, Shimizu K, Okazaki Y, Hayashizaki Y. Preprocessing implementation for microarray (PRIM): an efficient method for processing cDNA microarray data. Physiol Genomics. 2001;4(3):183–8.
26. Qin Y, Pan J, Cai M, Yao L, Ji Z. Pattern genes suggest functional connectivity of organs. Sci Rep. 2016;6:26501.
27. Danielsson F, James T, Gomez-Cabrero D, Huss M. Assessing the consistency of public human tissue RNA-seq data sets. Brief Bioinform. 2015;16(6):941–9.
28. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65.
29. Gat-Viks I, Sharan R, Shamir R. Scoring clustering solutions by their biological relevance. Bioinformatics. 2003;19(18):2381–9.
30. Bandyopadhyay S, Mukhopadhyay A, Maulik U. An improved algorithm for clustering gene expression data. Bioinformatics. 2007;23(21):2859–65.
31. Lord E, Diallo AB, Makarenkov V. Classification of bioinformatics workflows using weighted versions of partitioning and hierarchical clustering algorithms. BMC Bioinformatics. 2015;16:68.
32. Blekhman R, Marioni JC, Zumbo P, Stephens M, Gilad Y. Sex-specific and lineage-specific alternative splicing in primates. Genome Res. 2010;20(2): 180–9.
33. Sun J, Nishiyama T, Shimizu K, Kadota K. TCC: an R package for comparing tag count data with robust normalization strategies. BMC Bioinformatics. 2013;14:219.
34. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2016. https://www.r-project.org/.
35. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5:R80.
36. Kadota K, Nishiyama T, Shimizu K. A normalization strategy for comparing tag count data. Algorithms Mol Biol. 2012;7(1):5.
37. Hardcastle TJ, Kelly KA. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics. 2010;11:422.
38. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11:R25.
39. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
40. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11:R106.
41. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloë D, Le Gall C, Schaëffer B, Le Crom S, Guedj M, Jaffrézic F, French StatOmique Consortium. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform. 2013;14(6):671–83.
42. Maza E. 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. 2016;7:164.
43. Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, Wrobel N, Gharbi K, Simpson GG, Owen-Hughes T, Blaxter M, Barton GJ. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA. 2016;22(6):839–51.
44. Gierliński M, Cole C, Schofield P, Schurch NJ, Sherstnev A, Singh V, Wrobel N, Gharbi K, Simpson G, Owen-Hughes T, Blaxter M, Barton GJ. Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment. Bioinformatics. 2015;31(22):3625–30.
45. Frazee AC, Langmead B, Leek JT. ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. BMC Bioinformatics. 2011;12:449.
46. Bottomly D, Walter NA, Hunter JE, Darakjian P, Kawane S, Buck KJ, Searles RP, Mooney M, McWeeney SK, Hitzemann R. Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. PLoS One. 2011;6(3):e17820.
47. Cheung VG, Nayak RR, Wang IX, Elwyn S, Cousins SM, Morley M, Spielman RS. Polymorphic cis- and trans-regulation of human gene expression. PLoS Biol. 2010;8(9):e1000480.
48. Hubbell E, Liu WM, Mei R. Robust estimators for expression analysis. Bioinformatics. 2002;18(12):1585–92.
49. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4(2):249–64.
50. Kohl M, Deigner HP. Preprocessing of gene expression data by optimally robust estimators. BMC Bioinformatics. 2010;11:583.
51. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
52. Giorgi FM, Bolger AM, Lohse M, Usadel B. Algorithm-driven artifacts in median polish summarization of microarray data. BMC Bioinformatics. 2010;11:553.
53. Kadota K, Nakai Y, Shimizu K. A weighted average difference method for detecting differentially expressed genes from microarray data. Algorithms Mol Biol. 2008;3:8.
54. Kadota K, Nakai Y, Shimizu K. Ranking differentially expressed genes from Affymetrix gene expression data: methods with reproducibility, sensitivity, and specificity. Algorithms Mol Biol. 2009;4:7.
55. Pimentel MA, Clifton DA, Clifton L, Tarassenko L. A review of novelty detection. Signal Process. 2014;99:215–49.
56. Kadota K, Tominaga D, Akiyama Y, Takahashi K. Detecting outlying samples in microarray data: a critical assessment of the effect of outliers on sample classification. Chem-Bio Informatics J. 2003;3(1):30–45.
57. Kadota K, Ye J, Nakai Y, Terada T, Shimizu K. ROKU: A novel method for identification of tissue-specific genes. BMC Bioinformatics. 2006;7:294.
58. Di Y, Schafer DW, Cumbie JS, Chang JH. The NBP negative binomial model for assessing differential gene expression from RNA-Seq. Stat Appl Genet Mol Biol. 2011;10:art24.
59. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, Dylag M, Kurbatova N, Brandizi M, Burdett T, Megy K, Pilicheva E, Rustici G, Tikhonov A, Parkinson H, Petryszak R, Sarkans U, Brazma A. ArrayExpress update–simplifying data submissions. Nucleic Acids Res. 2015;43(Database issue):D1113–6.
60. Kauffmann A, Rayner TF, Parkinson H, Kapushesky M, Lukk M, Brazma A, Huber W. Importing ArrayExpress datasets into R/Bioconductor. Bioinformatics. 2009;25(16):2092–4.
61. Gautier L, Cope L, Bolstad BM, Irizarry RA. Affy–analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004;20(3):307–15.