Various methods had been introduced for normalization and comparison of RNA-seq count data. However, they lacked objectivity because they based on ad hoc assumptions that were never verified their appropriateness. Here, we introduced a method that assumes parsimony models on data distribution; the assumptions were verified according to exploratory data analysis. As was expected, count data were lognormally distributed. The level of noise in recent data appeared to be much higher than those of microarrays. Still, the appropriate distribution model would improve certainty and accuracy of normalization, by finding out the reliable range of data. Primary cause of noise was not the principle of the methodology; that is, each read is a trial that which transcript is read. Rather, the cause would be overlooking of transcripts, and the overlooking occurred more often among lower range of data. To find out genes likely to be overlooked, number of replications would be more important than read depth, which will not prevent overlooking. Both signal and noise in the reliable range of data were distributed normally, showing the suitability to use generalized linear model to evaluate differences in expression levels. In the framework, normalized data can be compared and combined freely beyond studies.

Introduction

By counting every piece of cDNA fragment detected by next-generation sequencers, RNA-seq estimates state of transcriptome of a sample (Marioni et al. 2008; Wang et al. 2009; Kassahn et al. 2011). Each fragment of cDNA is compared with the genome and assigned to a corresponding gene; expression level of a gene is estimated using the count of reads. Whereas microarray measures expression levels by intensity of hybridization, RNA-seq measures the level digitally; therefore, the methodology was expected to be more accurate. Additionally, as it can recognize only a single read, it was also expected to have superior sensitivity.

Analysis of the count data consists of several processes, such as preprocess that removes uncertain reads, mapping the reads onto the genome, and integrating the counts of exons to cover a transcript (Song et al. 2014). Also, de novo transcriptome assembly is available (Robertson et al. 2010; Grabherr et al. 2011). Anyway, by finding pieces of transcripts and counting them, expression level of a gene is estimated as integers of the counts.

Before comparing to other measurements, the count data have to be normalized, as count data are not compatible to standard units such as molar in the International System of Units; although once it was expected that such sophisticated treatment is unnecessary for RNA-seq data (Wang et al. 2009), in a practical sense, we cannot compare any sets of data without a mathematical model. Normalization is a process that makes unitless data comparable among measurements. In the process, certain characters found in each data set are unified by a predetermined protocol.

So many methods have been produced for normalization of RNA-seq data; actually, studies that compared the effect of those methods have been published (Dillies et al. 2013; Zyprych-Walczak et al. 2015). The methods can be classified according to assumptions in their concepts as follows. Among measurements, the number of total counts should be common: total count, reads per kilobase per million mapped reads; a quantile of the data should be common: upper quartile, median; most genes are not differentially expressed: DESeq, trimmed mean of M-values; distribution should be common: quantile
Evaluation of the expression differences among normalized data requires other mathematical models. Discrete distribution models like Poisson distribution, negative binomial distribution, or beta binomial distribution are proposed for this purpose (Baggerly et al. 2003; Lu et al. 2005); those are versatile models frequently applied for rarely occurring events. In particular, the latter two were used for separating the equality of the mean and variance in Poisson distribution model. Furthermore, complex models were also introduced, assuming nonparametric model (Tarazona et al. 2011) or relationship of noise to count data (Love et al. 2014).

Many of those methods have been introduced by taking a common procedure. First, a new concept is made on an assumption; this concept may relate existing problem or estimate a problem in data analysis. Then, a method is made on the concept. The effect of the method will be evaluated by comparing to other methods, on a scale that was defined for the purpose. Although this procedure seems to be an exemplar (Kuhn 1970) of this field, it is not surprising whether the method get better scores than others. This could be a reason why there have been so many methods introduced. As the analytical results depend on the assumptions, those obtained using different methods are difficult to be integrated.

Those methods have another commonality; assumptions of their concepts are not verified. Although the assumptions may sound reasonable, whether real data will satisfy the assumptions or not is not self-evident. For example, were the counts of spike controls exactly made? Was samples really shared data distribution? Did data distributed as it was assumed?

To proof objectivity of a method, such assumptions should be verified. However, some of those methods even totally lack ability to be validated. For example, Quantile (Bullard et al. 2010) replaces all the data with commonly found one. DESeq2 replaces variances with those estimated from pooled data (Love et al. 2014). In fact, these methods artificially change data to have desired character. Therefore, the methods totally lack falsifiability (Thornton 2014).

This article introduces a parametric method that normalizes data and evaluates the difference in expression levels measured by RNA-seq, as like the case of microarray data (Konishi 2004). Here, we use the term parametric for attempts to find a parsimony model that represents a character of data with least numbers of parameters. Appropriateness of the model is verified on each data according to the philosophy of exploratory data analysis (Tukey 1977).

Results and discussion
Distribution model
The logical aspect of the method is that distribution of data is lognormal and measurement is binomial, as follows. According to the thermodynamic model that explains how a cell maintains transcriptome according to the genome, the distribution of transcripts will be lognormal (Konishi 2005); this character is repeatedly confirmed in microarray analyses (Konishi 2004). In mathematical meanings, the lognormal model has three parameters: \( \sigma \), \( \mu \), and \( \gamma \), which determine the scale, location, and the background of the distribution, respectively. As the background can deeply alter the distribution, it is problematic in microarray data (Konishi 2004). In RNA-seq data, positive and negative background could be derived from reads of residual genome DNA fragments contaminated in RNA sample, and RNA absorption by the genome DNA, respectively. Additionally, in measuring the transcripts by RNA-seq, the data will have binomial character as well; at a gene of the genome, each of the reads is a trial to be read or not.

Checking data distribution
Data distribution can be briefly observed on histograms, but it is much rigorous to compare corresponding quantiles between data and theoretical model on a quantile–quantile plot (Tukey 1977). However, in this case, theoretical distribution is difficult to be estimated for routine purpose, as the model is a mixture of lognormal and binomial. Therefore, effect of neglecting binomial character was estimated by comparing logarithms of simulated data (Methods) against normal distribution. Fortunately, they were quite similar, showing coincidence in the quantiles (Fig. 1A). Even with decreased number of total read such as some ten thousands, the straight relationship remained, although the range became narrower (not shown). The effect of trial noise on data distribution would be rather limited; therefore, we can check data using the normal distribution, of which theoretical values can be feasibly obtained.
Samples of real distribution of $\log_{10}(\text{count data})$ were compared against normal distribution (Fig. 1B). The data did not correspond to the model at lower data range, as always be found in cases of microarray data (Konishi 2004). The lower limit of model-accordance range was $z = 1$ regardless to the total reads of the samples, in the experiment showed in Fig. 1B. The lower limits in the other measurements were $1 - 0$ in $z$-score (Supplementary Figure in Supporting Information). Although the model-accordance range was much narrower than microarray analyses, which can cover $-1$ to $-3$ in $z$-score (Konishi 2004), the range was enough to find out the parameters required for normalization as slope and intercept of the linear relationship (the range $1 - 2.5$ in Fig. 1B). Using the parameters, the count data were normalized measurement wisely (Methods). Commonality of the slopes was also found in microarray data (Konishi 2004), and it may represent an essential character of transcriptome (Konishi 2005).

Being lognormally distributed, it is obvious that the total read count is not a robust estimator of data character. Rather, it can be sensitively affected by some dominantly transcribed genes. Therefore, methods based on it, such as total count, FPKM, and RPKM, are not suitable for the purpose.

Noise and level of data

In Fig. 1B, only a part of data coincided to the model. Is the discordance occurred due to the data or the model? It is more likely that it has been caused by data noise. Level of noise, caused by both biological and technical reasons, was estimated using within-group SD of normalized data for each gene (data sets GSE54795, $n = 4$) and compared against the mean expression level (Fig. 2A). The SD changed drastically according to the expression level (white line); they were kept low only within the model-accordance range (straight relationship found in Fig. 1B). This critically shows that the reproducibility of measurement is hardly expectable below the accordance range; such expansion of deviations was routinely found in other data sets (Supplementary Figure in Supporting Information). Especially, below $-1$ of the mean expression level, only a few genes showed SDs at the same level of the model-accordance range (Fig. 2A); it is obvious that the larger SDs in the lower half of data present total failure of measurements. It should be noted that validations for the previous methods might be carried out on likely noisy data.

A cause of the increasing SDs was overlooking of transcripts. This phenomenon became evident in
comparisons with two normalized measurements (Fig. 2B and Supplementary Figure in Supporting Information); from the bottom of the plot, zero (arrows), one, two, three read counts formed horizontal and vertical lines. The lines show genes that gave higher expression levels could be totally overlooked in other measurements. Such overlooking may not always be complete lost but failure of detection with certain possibility. As some of the genes, if not all of them, may be overlooked and hence skipped, the order of data distribution has become distorted, forming the downward bend appeared in Fig. 1B. The tight curve found in Fig. 1B suggests that the overlooking occurred in many of the corresponding genes.

The overlooking tended to occur in lower expression levels (Fig. 2). If it had been occurred randomly and independent to the expression levels, it did not alter the lognormal character of data distribution as like random sampling from data; if so, the overlooked reads as well as nonoverlooked reads will have similar distribution pattern that accords to the population distribution.

According to lognormality of data and the overlooking, it is obvious that the five assumptions used in previously proposed methods were all inappropriate. The total read count is not a robust estimator of data character; the total can be easily affected by outliers appeared in higher range of data and overlooking. Any quartile: upper quartile or median will not be robust characteristic value, as the model-accordance range may not cover the quartile; it will be much accurate and robust to find the parameters from the linear relationship appeared in quantile plot of data (Fig. 1B). Both the stability of gene expression and data distribution are difficult to be assumed with the noisy data. The level of noise was not stable but varied much (Fig. 2). Furthermore, the previously method could not find out the wide data range that would be affected by noise.

The tendency of overlooking can be explained biologically. Some of the RNA species have short half-lives (Konishi 2005). Such character of RNA is beneficial in rapid control of the levels, and because of faster degradation, the levels would become rather low. Such RNA species may attach to fewer protein factors that protect the RNA from attack of RNases. In handling of RNA, such as depleting of rRNA and fragmentation of RNA, both of which may not be applied for microarray methodologies, intactness of such unguarded RNA could be easily lost, and this may affect quality of count data.

Taken together, the departure of the model would be caused by data noise. The model-accordance range showed the reliable range of data (Figs 1B and 2A); below the level, at least some of the genes would be affected by overlooking, and we cannot predict which genes will tend to be overlooked. This could
be partly dissolved by making the assignments tolerant to mismatches, but this will be a trade-off between accuracy and sensitivity. Improvement in sensitivity may require refinement in wet procedures.

This parametric method of normalization can be applied to each of exons as well. However, to find out the parameters, the range of observations should be maintained as wide as possible, such as a whole chromosome. Otherwise, effect of activation and inactivation occurred by chromosome modifications could affect estimation of the parameters.

**Effect of by-chance noise**

There would be another source of noise that caused by chance of trial. For each read of data, any transcript $i$ has a chance to be read, and the probability can be estimated as $P = (\text{copy number of the transcript})/(\text{total copy number of RNA})$. With this probability, the count will be binomially distributed. In an opposite way, we can estimate the copy number from count data $x_i$; given the count, with any $P$, expectation of probability getting $x_i$ can be represented using $\text{dbinom}$ function of R (Fig. 3, solid line). Additionally, if $x_i$ is large enough, this distribution could be approximated using probability $P = x_i/r$ (Fig. 3, circles; in this case, the count was 50). Therefore, the level of noise can be estimated as variance of binomial distribution: $rp(1-P) = x_i (1-P)$. As for most of the genes $P$ is close to zero, the variance is $x_i$ in a practical sense.

Additionally, as we can clearly see the shape of distribution on the discrete data (Fig. 3), it would be unobvious that quantization noise (caused by the gaps between the circles) is much smaller than the by-chance noise.

Furthermore, effect of the by-chance noise will be much compressed upon normalization. When total read is $r$ and the probability at a gene is $P$, the expectation of the read will be $np$ and the variance will be $np(1-P) \approx np$. Hence, if a read had one-sigma positive noise, the read will be $np + \sqrt{np}$. In normalization process, logarithm of this value will be handled: $\log(np + \sqrt{np}) = \log(np) + \log(1 + 1/\sqrt{np})$. As the noise $\log(1 + 1/\sqrt{np})$, would be much smaller than $\log(np)$—when $np = 100$, $\log(np) = 2$ whereas the noise is 0.041—effect of the noise derived by the trial is rather frivolous.

**Distribution of signal and noise**

As the effect of by-chance noise is negligible, we should concentrate to noise that has been originated by the sample, such as caused by individual differences. In the model-accordance range, differences of two measurements were distributed normally (Fig. 4), both within a group and between groups as like the cases found in microarray (Konishi 2004, 2011). These clearly show normality of distributions of signal and noise; therefore, generalized linear models such as ANOVA are preferable to estimate expressional changes.

Based on the evidences presented here, it is now clear that the previously proposed models for evaluating expressional changes are inappropriate for RNA-seq data. The effect of trial is negligible opposed to the assumptions (Baggerly et al. 2003; Lu et al. 2005). When we can use parametric models, nonparametric ones, of which assumptions are difficult to be verified (Tarazona et al. 2011), should be avoided to keep objectivity of analysis. Furthermore, estimating noise level from pooled data, which was found in some methods (Tarazona et al. 2011; Love et al. 2014), could be pertinent to ‘making up data or results and recording or reporting them’, that is, definition of fabrication (U. S. Federal Policy on Research Misconduct 2002). There would be no doubt that those handling should not be attempted in analysis.
It should be noted that at the lower level of expression, many of the data could be seriously affected by noise that might be derived by insufficient quality of counting. For microarray analysis, I once pointed out the unnecessity to consider familywise error rate or false discovery rate (Konishi 2011). The idea is based on accuracy of microarray, and it would be effective in the model-accordance range of RNA-seq data as well. However, many of the genes below the model-accordance level are just ruled by noise; if within-group variance is estimated quite low, it could be an artifact given by chance. Statistical test cannot ensure conclusions on such noisy observations, even with severe corrections for multiplicity.

Comparison with other methods

For the purpose of reference, a difference in the result of normalization is shown in Fig. 5. As the previously introduced methods cannot distinguish model-accordance range of data, analyses will become rather noisy. In this figure, the level of noise, including biological one, was presented. As SD is dependent to the magnitude of normalized data, coefficient of variation was found for each gene of a group, and compared among methods (Fig. 5). The noise level was distinctively lower in data normalized parametrically.

Importance of replication

Larger number of total reads will not improve accuracy of data, as the departure from the model would
be a character of population distribution. In fact, in the presented measurement, the limitation of reliance was ca. \( z = 1 \) (Fig. 1B), and the read count at \( z = 1 \) was 5000. The by-chance noise of trial would be negligible with count of 100 as was discussed. Therefore, the count data of experiment, which had total 70 million reads for each measurement, were fifty times more than that required to cover the reliance range with enough read count.

To find out the genes tend to be overlooked, the number of replications is quite important. As was discussed, accuracy of analysis cannot be improved by larger number of total reads. Therefore, more replication would be beneficial than more sequence, as has been reported on an empirical bases (Tarazona et al. 2011; Liu et al. 2014). For this purpose, redoing of measurements is not sufficient; the replication requires another set of tissue samples, as the same RNA preparation may have the same predisposition of overlooking.

Although importance of count was also suggested (Lu et al. 2005), this idea could be derived by a fact that more abundant RNA species would have better intactness and more count. Incidentally, total read required to cover a gene at \( z = 0 \) and 1 with 100 reads would be 26 and 3.2 million, respectively, when \( \sigma = 0.9 \) and the number of genes is 3E4.

### Conclusion

In this article, two series of evidences were presented: RNA-seq count data were lognormally distributed, and both signal and noise among normalized data were normally distributed. Those give a firm base on parametric handling of data analysis. The parametric method seems to be the first procedure of which concept has been verified its appropriateness, and the verification will be repeated on each analysis. It is true that the concepts of the previous methods stand on presumable assumptions; however, if they do not fit to the data structure, they are useless. Therefore, the assumptions should have been verified (Thornton 2014). In this point of view, the parametric method is different from the previous ones rather qualitatively; it was not made in the present paradigm (Kuhn 1970) of bioinformatics; rather, it was found parsimony according to the philosophy of exploratory data analysis (Tukey 1977).

At present, RNA-seq did not show accuracy or sensitivity originally expected (Marioni et al. 2008; Wang et al. 2009; Kasahn et al. 2011). The main cause of the noise was the overlooking of reads, which could be derived by inappropriate handling of sample RNA. This tended to occur at lower expression level; the data range affected by this artifact could be found out at normalization process and confirmed by comparing SD against expression level. In comparison with noise caused by the overlooking that caused by the trials would have rather negligible effect. To accurately estimate within-group variance and find out genes that were frequently overlooked, more replications are desirable rather than irrelevant read counts. As we can expect normal character in changes of normalized data, evaluation of differences can be carried out in a parametric manner using the normal distribution model; however, this is not practical for genes that are seriously affected by noise.

### Experimental procedures

**RNA-seq data**

RNA-seq count data were obtained from the public database NCBI Gene Expression Omnibus (Barrett et al. 2013). The database were searched with keyword of ‘(rna-seq) AND transcriptome’, obtaining 1175 data sets (16 Sep. 2015). Among them, read count of 10 data sets that were released on August 2015, GSE54795, GSE57202, GSE63405, GSE64797, GSE66446, GSE66962, GSE69007, GSE69707, GSE70073, and GSE70484, was downloaded and used in this study.

**Calculation**

Calculations were made on the R (R Core Team 2015) using functions indicated. The complete scripts are described in supplementary data.

**Simulating a data set**

As distribution of transcriptome was deduced to be lognormal (Konishi 2005), ideal level of \( i \)th quantile \( q_i \) can be estimated using inverse cumulative density function of normal distribution (i.e., \( \text{qnorm} \) function of the R) as \( \log(q_i) = \text{qnorm}(i\ \text{th probability point, mean = 0, SD = 1}) \). Also, a series of data was simulated using \( \text{rbinom} \) function; using the ideal \( q_i \) and their sum total, \( r \), the probability of the trial for gene \( i \) can be estimated as \( p_i = q_i / r \). The simulated data were generated as \( x_i = \text{rbinom}(\text{number of observations} = 1, \text{number of trials} = 3E4, \text{probability of success} = p_i) \).

**Data normalization**

The count data \( x_i \) were normalized as \( z_i = (\log(x_i) - \gamma) / \sigma \) (Konishi 2004), where \( \gamma \) is the background, and \( \mu \) and \( \sigma \) are the location and scale parameters of normal distribution, respectively. In the presented cases, the parameter \( \gamma \) was set to zero. Sorted \( \log(x_i) \) was compared with the theoretical
quantiles of normal distribution; the values for $\mu$ and $\sigma$ were robustly estimated as intersect and slope, respectively, at the model-accordance range (Fig. 1B) using $\ln$ function of the $R$.

### Estimation of probability from read count

According to the total read $r$ and a count $x$, expectation of population probability density was estimated using probability density function of binomial distribution of the $R$, $\text{binom}$. Here, the density was estimated when the count was 50 and total read was $1E8$, with various $P$. This distribution was approximated with fixed $P = 50/1E8$ for various values of the count.

### Comparison with other methodologies

Count data of GSE54795 were normalized by methods indicated, and in the control group, standard deviation (SD) and mean were calculated for each gene. Total count, upper quartile, and quantile were carried out according to Zyprych-Walczak et al. 2015; Remove unwanted variation was carried out using R package supplemented (Risso et al. 2014); residuals were found as the most stable 200 genes and used for the process. Then coefficient of variance was found from each gene of a group. Coefficient of variance was estimated by dividing the SD by mean. Resulted values were presented in a box plot. For the parametric normalization, only the model-accordance range of data (z-scores above 1) were used.

### Acknowledgements

This work was supported by the President Project of Akita Prefectural University.

### References

Baggerly, K.A., Deng, L., Morris, J.S. & Aldaz, C.M. (2003) Differential expression in SAGE: accounting for normal between-library variation. *Bioinformatics* 19, 1477–1483.

Barrett, T., Wilhite, S.E., Ledoux, P. et al. (2013) NCBI GEO: archive for genomic data set–update. *Nucleic Acids Res.* 41, D991–D995.

Bullard, J.H., Purdom, E., Hansen, K.D. & Dudoit, S. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11, 94.

Dillies, M.A., Rau, A., Aubert, J. et al. (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* 14, 671–683.

Grabherr, M.G., Haas, B.J., Yassour, M. et al. (2011) Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat. Biotechnol.* 29, 644–652.

Kassahn, K.S., Waddell, N. & Grimmond, S.M. (2011) Sequencing transcriptomes in toto. *Integr Biol (Camb)* 3, 522–528.

Konishi, T. (2004) Three-parameter lognormal distribution ubiquitously found in cDNA microarray data and its application to parametric data treatment. *BMC Bioinformatics* 5, 5.

Konishi, T. (2005) A thermodynamic model of transcriptome formation. *Nucleic Acids Res.* 33, 6587–6592.

Konishi, T. (2011) Microarray test results should not be compensated for multiplicity of gene contents. *BMC Syst. Biol.* 5(Suppl 2), S6.

Kuhn, T.S. (1970) *The Structure of Scientific Revolution*, 2nd edn., IL: University of Chicago Press.

Liu, Y., Zhou, J. & White, K.P. (2014) RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics* 30, 301–304.

Love, M.I., Huber, W. & Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550.

Lu, J., Tomfohr, J.K. & Kepler, T.B. (2005) Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach. *BMC Bioinformatics* 6, 165.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. & Gilad, Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* 18, 1509–1517.

R Core Team. (2015) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria. [Cited 1 Nov 2015.] Available from URL: http://www.R-project.org

Risso, D., Ngui, J., Speed, T.P. & Dudoit, S. (2014) Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* 32, 896–902.

Robertson, G., Schein, J., Chiu, R. et al. (2010) De novo assembly and analysis of RNA-seq data. *Nat. Methods* 7, 909–912.

Song, L., Florea, L. & Langmead, B. (2014) Lighter: fast and memory-efficient sequencing error correction without genome. *Genome Biol.* 15, 509.

Tarazona, S., García–Alcalde, F., Dopazo, J. & Ferrer, A. (2011) Conesa A. Differential expression in RNA-seq: a matter of depth. *Genome Res.* 21, 2213–2223.

Thornton, S. *Karl Popper*. In: *The Stanford Encyclopedia of Philosophy*, (Summer 2014 ed.). [Cited 1 Nov 2015.] Available from URL: http://plato.stanford.edu/archives/sum2014/entries/popper/

Tukey, J.W. (1977) *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

U. S. Federal Policy on Research Misconduct (2002) [Cited 1 Nov 2015.] Available from URL: http://www.apo.org/policy/statements/federalpolicy.cfm

Wang, Z., Gerstein, M. & Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63.

Zyprych-Walczak, J., Szabelska, A., Handschuh, L., Górczak, K., Klamecka, K., Figlerowicz, M. & Siatkowski, I. (2015)
The impact of normalization methods on RNA-Seq data analysis. *Biomed Res Int* 2015, 621690.

Received: 25 December 2015
Accepted: 4 April 2016

**Supporting Information**

Additional Supporting Information may be found online in the supporting information tab for this article:

**Figure S1** QQ plots of other experiments that contained multiple measurements (relating to Figure 1B in the article).

**Figure S2** (A) Relationship with within-group SD and average mean of another group. (B) Scatter plot comparing normalized data of two of the repeated measurements in other combinations.

**Data S1** Parametric analysis of RNA-seq expression data.