entropy

MDPI

Article

A Method Based on Differential Entropy-Like Function for Detecting Differentially Expressed Genes Across Multiple Conditions in RNA-Seq Studies

Zhuo Wang †, Shuilin Jin *,† and Chiping Zhang *,†

Department of Mathematics, Harbin Institute of Technology, Harbin 150006, China; zhuowang@hit.edu.cn

* Correspondence: jinsl@hit.edu.cn (S.J.); zcp@hit.edu.cn (C.Z.);
Tel.: +86-451-8641-4216 (S.J.); +86-451-8640-2875 (C.Z.)

† These authors contributed equally to this work.

Received: 4 January 2019; Accepted: 27 February 2019; Published: 4 March 2019

Abstract: The advancement of high-throughput RNA sequencing has uncovered the profound truth in biology, ranging from the study of differential expressed genes to the identification of different genomic phenotype across multiple conditions. However, lack of biological replicates and low expressed data are still obstacles to measuring differentially expressed genes effectively. We present an algorithm based on differential entropy-like function (DEF) to test for the differential expression across time-course data or multi-sample data with few biological replicates. Compared with limma, edgeR, DESeq2, and baySeq, DEF maintains equivalent or better performance on the real data of two conditions. Moreover, DEF is well suited for predicting the genes that show the greatest differences across multiple conditions such as time-course data and identifies various biologically relevant genes.

Keywords: differential entropy-like function; differential expressed genes; multiple condition data; time-course data

1. Introduction

Next-generation sequencing (NGS) technology has rapidly become the tool for many genome-wide transcription studies. Production of millions or billions of short sequences from individual RNA molecules together with lower costs and higher budgets have enabled many methodologies, such as RNA sequencing (RNA-Seq) [1]. RNA-Seq technology empowers thorough recognition of gene isoforms [2], translocations [3], nucleotide variations [4], time-course gene expression analysis [5], and cap analysis of gene expression (CAGE) [6]. One reason for the growing popularity of RNA-Seq technology is its ability to detect differentially expressed genes between two or more conditions (e.g., different races of human populations). However, the expression of most genes is intrinsically stochastic [7], so several methods have been introduced for exploring it [8–10], despite the variability in RNA-Seq data, which poses challenges for differential expression and other relative analysis [11].

The theoretical methods for the study of RNA-Seq data can be grouped into two major categories. The first representative methods are limma [10], edgeR [12–14], DESeq2 [15], baySeq [16], PoissonSeq [17,18], DSS [19] and DGEclust [20]. These methods rely on the accuracy of distribution assumptions and parameters estimation. Once the reads are mapped at the gene, exon or transcript level, the problem is naturally summarized into matrices where rows represent exons, genes or transcripts and columns represent samples or replicates. Therefore, many of the earliest statistical methods are based

on Poisson or negative binomial distribution to model read counts. Some of these methods represent the current state-of-the-art of the field for the study of RNA-Seq data. However, the methods based on the negative binomial and Poisson model either fail to perform multiple conditions comparison or are excessively conservative [21,22]. It is noteworthy that assessing differential low expressed count features and biological experiments with few replicates is a challenge for the estimation of model parameters. The methods in the second category such as NOIseq [23,24], rSeqNP [25] and SAMseq [26] attempt to make use of non-parametric approaches. These methods rely less on the probability distribution of exon, gene and transcript counts. An advantage is that this class of methods aims to be data-adaptive and is suitable for differential expression analysis without prior information. Whereas when we teased apart this kind of methods, we found that they may tend to overestimate differentially expressed genes with high variability among replicates.

In this paper, we introduce an approach based on the differential entropy-like function (DEF), an algorithm for discovering the differentially expressed genes across multiple conditions. From the theoretical point of view, the importance of the proposed method derives from its information-theoretic background. On the one hand, DEF allows the recognition of differentially expressed genes on multiple conditions such as time-course data and multiple tissues data. Compared with the popular alternative methods, DEF obtains a wider application. On the other hand, DEF is effective in detecting the differentially expressed genes with the characteristic that it does not rely on the probability distribution. Another characteristic of DEF is its adaptability on zero expressed gene counts, which could give rise to the other troubling aspect that many hypothesis test methods fail. The paper is organized into the following sections. Section 2 presents the performance of DEF compared with the other methods on the datasets of two conditions. Section 3 presents the evaluation of DEF on the two-condition data. Section 4 presents the performance of the DEF on the time-course RNA-Seq data. Section 5 presents the effective analysis of DEF on the datasets of multiple conditions.

2. Results

2.1. DEF Shares Many Genes with Limma, DESeq2, baySeq and edgeR

We started by analyzing the actual RNA-Seq datasets with limma, DESeq2, baySeq, edgeR and DEF. Details of the datasets are described in Table 8 of the "Materials and Methods" Section. In all cases with "Sultan" and "Katz" datasets from R package "recount" [27], there were two-condition data with two technical replicates per condition. The Venn diagram for each of the cases is shown in Figures 1 and 2. All compared methods ranked each gene by providing *P* values (limma, DESeq2 and edgeR), FDR (baySeq) or an entropy-like value (DEF). *P* Value or FDR less than 0.05 was considered to indicate statistical significance. The cut-off values of DEF were 0.05 (Figures 1a and 2a) and 0.01 (Figures 1b and 2b). As indicated by the Venn diagrams constructed from differentially expressed genes, sharing 792 genes in the "Sultan" case demonstrated a significant overlap by the five methods (Figure 1a). In this figure, we note that the differentially expressed genes found by DEF were to a large extent also found by limma, edgeR, baySeq and DESeq2. Simultaneously, our DEF method found a fair amount of unique differentially expressed genes, which were not shared with the other methods (Figure 1b). We further investigated two possibilities for the additional differentially expressed genes in Figure 1b. Tables 1 and 2 list the raw read counts of ten unique differentially expressed genes detected by DEF (Figure 1b). The top five and last five genes were with the largest and smallest DEF values among the 1907 additional genes. As can be seen in Table 1, all five genes had only one non-zero expressed value across four replicates. These genes were true positives, which DEF detected better than the other methods. As shown in Table 2, the differences between these replicates were not so clear. Some gene such as "ENSG00000065357" had extreme read counts (10), while some genes had moderate read counts. These genes could be false positives detect by DEF,

which was why the other methods did not identify them. In "Katz" case, 41 and 100 overlapping genes were found by each method with different cut-off values of DEF (Figure 2). Figure 2b displays that DEF detected 1134 additional differentially expressed genes. Tables 3 and 4 demonstrate the raw read counts of "Katz" dataset, which further confirms the effectiveness of DEF method. The two tables list the top five and last five unique differentially expressed detected by DEF from 1134 genes (Figure 2b). In Table 3, five genes with largest DEF values expressed in Condition B. However, they all had zero expressed values in Condition A. These genes should be different genes and DEF successfully detected these genes, which failed to be identified by the other methods. In Table 4, the gene "ENSMUSG00000036977" was highly expressed in Condition A and the genes "ENSMUSG00000057924" and "ENSMUSG00000067203" had higher expression in Replicate 1 of Condition B. These genes were true positives detected by DEF. The difference performance of the other genes was not clear. These genes could be false positives.

The reasons for better performance of DEF compared to other methods in all datasets stem from its adaptability for zero expressed gene counts. Low replicates in each sample cause inaccuracy on the distribution-dependent methods and zero expressed gene counts give rise to the other troubling aspect that many hypothesis test methods failed.

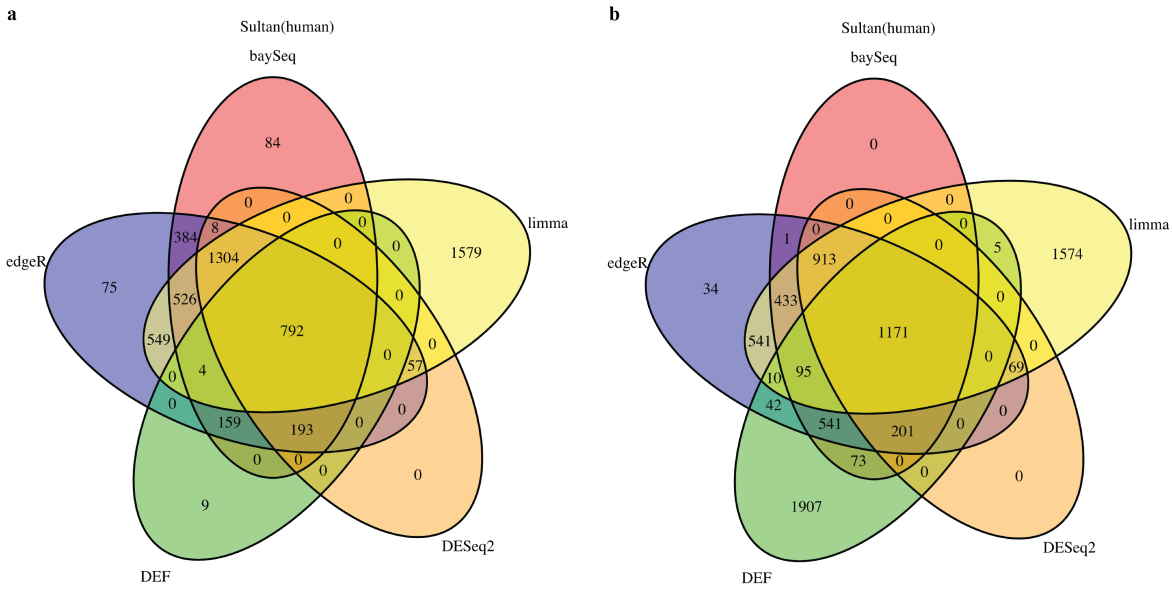

Figure 1. Venn diagram of differentially expressed genes obtained from limma, baySeq, DESeq2, edgeR and DEF: (**a**) read counts from "Sultan" dataset with the threshold for DEF entropy-like value of 0.05; and (**b**) read counts from "Sultan" dataset with the threshold for DEF entropy-like value of 0.01.

Table 1. Read counts of top five unique differentially expressed genes detected by DEF in "Sultan" dataset.

Ensembl ID	Condition A Replicate 1	Replicate 2	Condition B Replicate 1	Replicate 2	DEF Value
ENSG00000164002	5	0	0	0	0.0617
ENSG00000104833	0	0	7	0	0.0608
ENSG00000124920	0	0	7	0	0.0608
ENSG00000182310	0	0	7	0	0.0608
ENSG00000197608	0	0	0	6	0.0581

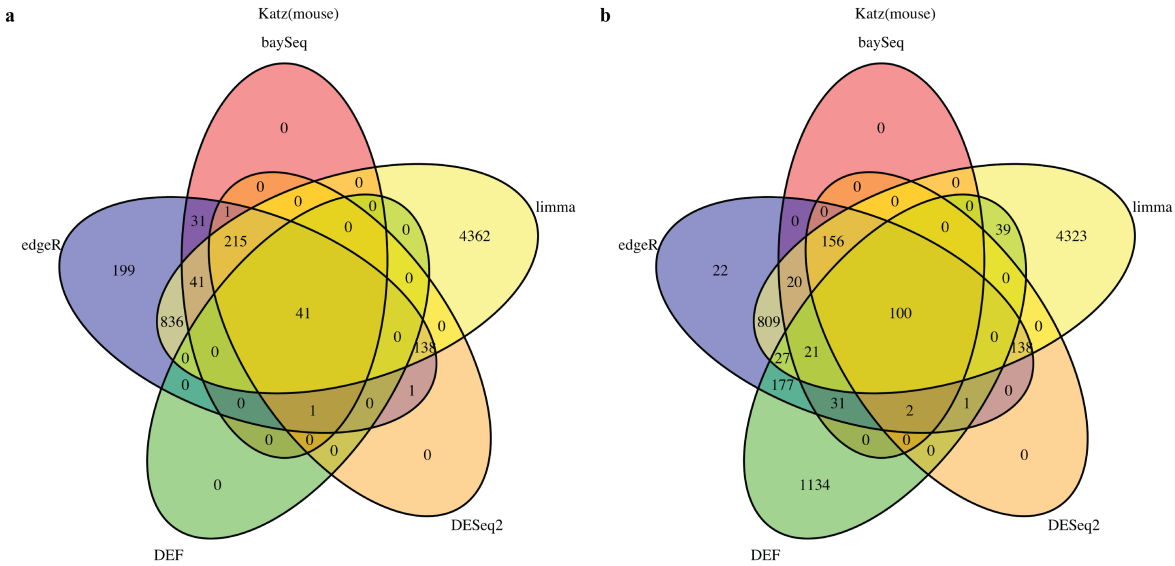

Figure 2. Venn diagram of differentially expressed genes obtained from limma, baySeq, DESeq2, edgeR and DEF: (**a**) read counts from "Katz" dataset with the threshold for DEF value of 0.05; and (**b**) read counts from "Katz" dataset with the threshold for DEF value of 0.01.

Table 2. Read counts of last five unique differentially expressed genes detected by DEF in "Sultan" dataset.

Ensembl ID	Condition A Replicate 1	Replicate 2	Condition B Replicate 1	Replicate 2	DEF Value
ENSG00000111325	4	8	6	2	0.0101
ENSG00000141431	1	1	5	4	0.0100
ENSG00000065357	4	10	3	4	0.0100
ENSG00000179021	2	6	6	2	0.0100
ENSG00000215301	2	6	6	2	00100

Table 3. Read counts of top five unique differentially expressed genes detected by DEF in "Katz" dataset.

Ensembl ID	Condition A Replicate 1	Replicate 2	Condition B Replicate 1	Replicate 2	DEF Value
ENSMUSG00000051920	0	0	5	0	0.0437
ENSMUSG00000029683	0	0	0	4	0.0436
ENSMUSG00000069301	0	0	0	4	0.0436
ENSMUSG00000070691	0	0	1	4	0.0432
ENSMUSG00000079332	0	0	3	2	0.0412

Table 4. Read counts of last five unique differentially expressed genes detected by DEF in "Katz" dataset.

Ensembl ID	Condition A Replicate 1	Replicate 2	Condition B Replicate 1	Replicate 2	DEF Value
ENSMUSG00000038593	6	3	4	0	0.0101
ENSMUSG00000036977	5	11	0	1	0.0101
ENSMUSG00000057924	2	2	5	1	0.0101
ENSMUSG00000067203	2	2	5	1	0.0101
ENSMUSG00000002205	2	13	10	3	0.0100

2.2. DEF Successfully Identified the Differentially Expressed Genes under the Real Dataset

In differential expression analysis, an important task is to identify the genes that are differentially expressed at higher variability between experimental conditions without prior information. We compared the performance of DEF to identify differentially expressed genes under the experimental conditions encapsulated by the actual dataset. Details of the dataset are in the "Materials and Methods" Section. We analyzed the 100 top-ranking genes with box plots, as shown in Figure 3a for "Sultan" dataset and Figure 3b for "Katz" dataset. These box plots show the variance across different samples. Medians of the box plots varied widely across samples. Hence, DEF is an approach for the identification of differentially expressed genes from count data. Essentially, our method creates a measurement for gene-wise counts by DEF and evaluates the absolute expression differences for the genes in all the samples. We also utilized the method generalized log-cpm to obtain the normalized matrix of counts, which can remove potentially library variation and prevent bias and mean squared error in downstream analyses. We evaluated various methods for differential expression analysis and found that our method performed equivalent to the classical methods. The main difference between DEF method and other methods based on statistics is the ability to handle low expression counts, especially zero counts, an issue of great importance when investigating differential expression in the context of RNA-Seq. When both samples have zero reads, clearly nothing can be said about differential expression and we have already filtered these genes. Presumably, this represents an interesting biological phenomenon, where a gene in all samples is completely non-expressed according to sequencing. For genes with zero counts in either sample, many methods failed, except DEF method (Tables 1 and 3). Because many methods cannot handle zero-count genes, their methods failed to detect many easy cases of differentially expressed genes (i.e., genes with zero-count in one condition while non-zero-count in the other condition). Because the use of parameters in the normalization step of our model successfully handled zero counts, the larger DEF value was particularly pronounced for genes expressed at a more obvious difference across samples (Tables 1–4).

2.3. DEF Is Applicable for the Real Time-Course Data

With the development of sequencing technology, researchers pay more attention to timeliness. More and more time-course data come from the advanced experiment and more methods are needed for analyzing these data. DEF also works effectively on the time-course data. "Trappnell" dataset was obtained from online resources of "recount". The mouse samples of the dataset were from four time points following differentiation. We applied DEF on the time-course data and took the top five differentially expressed genes with the largest DEF values for further analysis. Figure 4a displays dramatic changes over the four time points of the top five differentially expressed genes. The genes decreased rapidly in the first two time courses and increased in the last time course. These genes could be used for further analysis in the biological progress. Table 5 lists the gene symbols and gene functions of these five genes. The most differentially expressed gene, Kart9, plays an essential role in the correct development of sperm [28]. The Lce1g is one of the differentiation-related genes [29]. All these findings were consistent with the characteristic of the time-course data. Particularly, our results predicted that Avpr1a, Pcdh20 and Npas4 may play an important role in the shaping of the samples from time-course data. Figure 4b shows the box plot of the top 100 differentially expressed genes of the four time points separately. Comparison of the four box plots clearly indicates the expression variation across four time points, which proved the effectiveness of DEF.

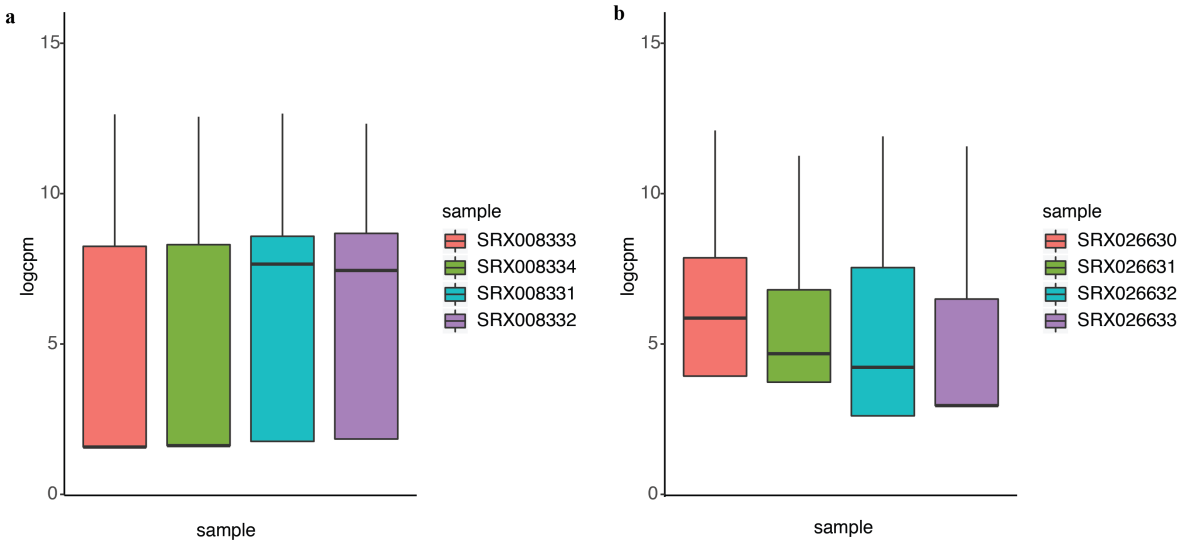

Figure 3. The performance of DEF on the two-group data: (**a**) box plots of the top 100 differentially expressed genes from "Sultan" dataset; and (**b**) box plots of the top 100 differentially expressed genes from "Katz" dataset.

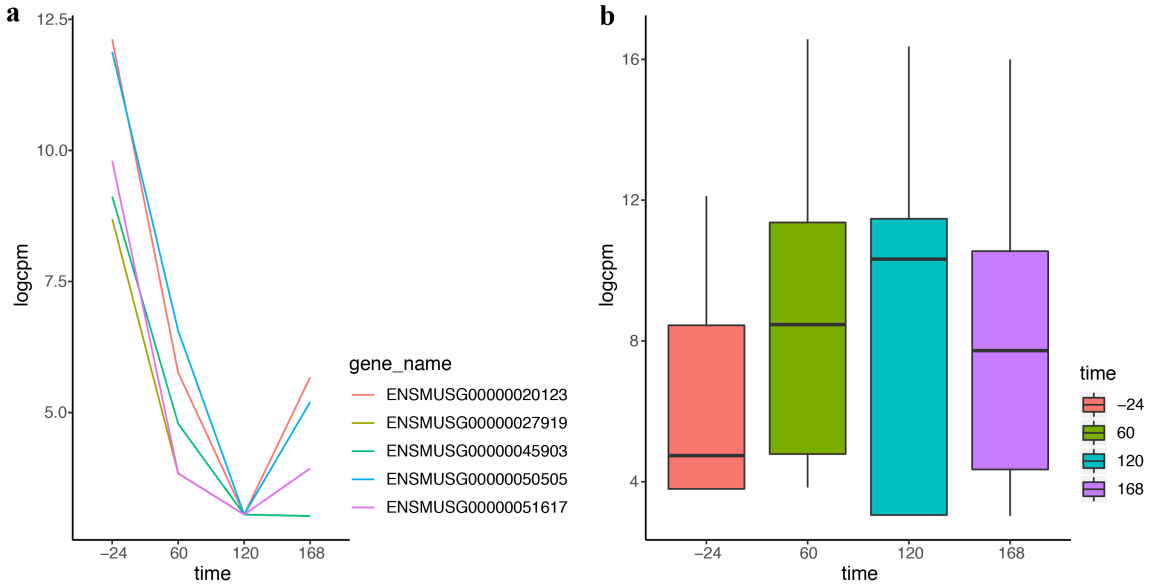

Figure 4. The performance of DEF on time-course data: (**a**) normalized read counts of the top five differentially expressed genes over four time points; and (**b**) box plots of the top 100 differentially expressed genes on the first time point compared with every time point.

Table 5. Top five differentially expressed genes obtained from DEF.

Ensembl ID	Gene Symbol	Gene Function
ENSMUSG00000051617	Krt9	An important special function in the mature palmar Plays an essential role in the correct development of sperm
ENSMUSG00000020123	Avpr1a	Receptor for arginine vasopressin
ENSMUSG00000050505	Pcdh20	Potential calcium-dependent cell-adhesion protein.
ENSMUSG00000027919	Lce1g	keratinocyte differentiation
ENSMUSG00000045903	Npas4	A key role in the structural and functional plasticity of neurons Transcription factor expressed in neurons of the brain

2.4. DEF Is Applicable for the Real Multiple Condition Data

Most existing methods deal with a two-condition comparison, while DEF was designed to be an effective tool for the quantification of differentially expressed genes across multiple conditions. We show that our method is applicable in terms of differential expression analysis on multiple condition data. A multiple condition dataset was considered as an example to test the feasibility of our method. We chose the "Cheung" dataset, which contains samples of 41 CEPH HapMap (CEU) samples [30], and applied DEF method for differential expression analysis across the 41 samples. DEF identified several genes that are differentially expressed across these samples (Table 6). We searched for some testified findings and these findings were consistent with our results. Our observation about the differentially expressed genes implicated that DEF is well suited for predicting which genes show the greatest differences in expression between biological samples. One of the differentially expressed genes is ZFP57, which is confirmed as highly variably expressed gene from 1000 Genomes CEU phase 1 [31]. The "Cheng" dataset consists of 17 female samples and 24 male samples. Another gene RPS4Y1 is also a differentially expressed gene detected by DEF. The gene RPS4Y1 is located in chromosome Y. In particular, our results predict that PRSS21, MKRN and GTSF1 may play an important role in the shaping of specificity of the CEU from HapMap (Table 6). Figure 5a shows the box plot of the top 100 differentially expressed genes across all the samples separately. Figure 5b displays 100 genes that DEF identified as non-differentially expressed genes. Clearly, medians change greatly in Figure 5a while the medians in Figure 5b are robust. Comparison of 41 box plots in Figure 5a,b clearly indicates the expression variation across samples, which proved the effectiveness of DEF. We also took additional multi-tissue data as an example to test the feasibility of our method. We chose the data "Wang" from online resources "recount" of RNA sequencing on human cell line with diverse tissues. The tissues included in the dataset are cerebellum, breast, brain, adipose, T47D, MCF7, MB435, HME and BT474. We compared the dataset from these tissues and applied DEF method for differential expression analysis between those samples. As there are limited "gold-standard" data with which to evaluate the accuracy of RNA-Seq quantification methods, and because real differentially expressed genes are difficult to confirm, we connected some gene functions of our analyzed DE genes with the real biological differential traits between different tissues. TCL1A [32,33] , POU2AF1, ARHGDIB [34–36], LRMP and IRS4 [37–39] were the most variable genes (Table 7).We also took the top 100 differentially expressed genes of every sample for further analysis. The result presented in Figure 6 show that our method found the obvious differences across the different samples. We found that there is limited "gold-standard" data with which to evaluate the accuracy of RNA-Seq quantification methods, calling into question how to thoroughly evaluate the DEF method. However some published findings ar e consistent with our results.

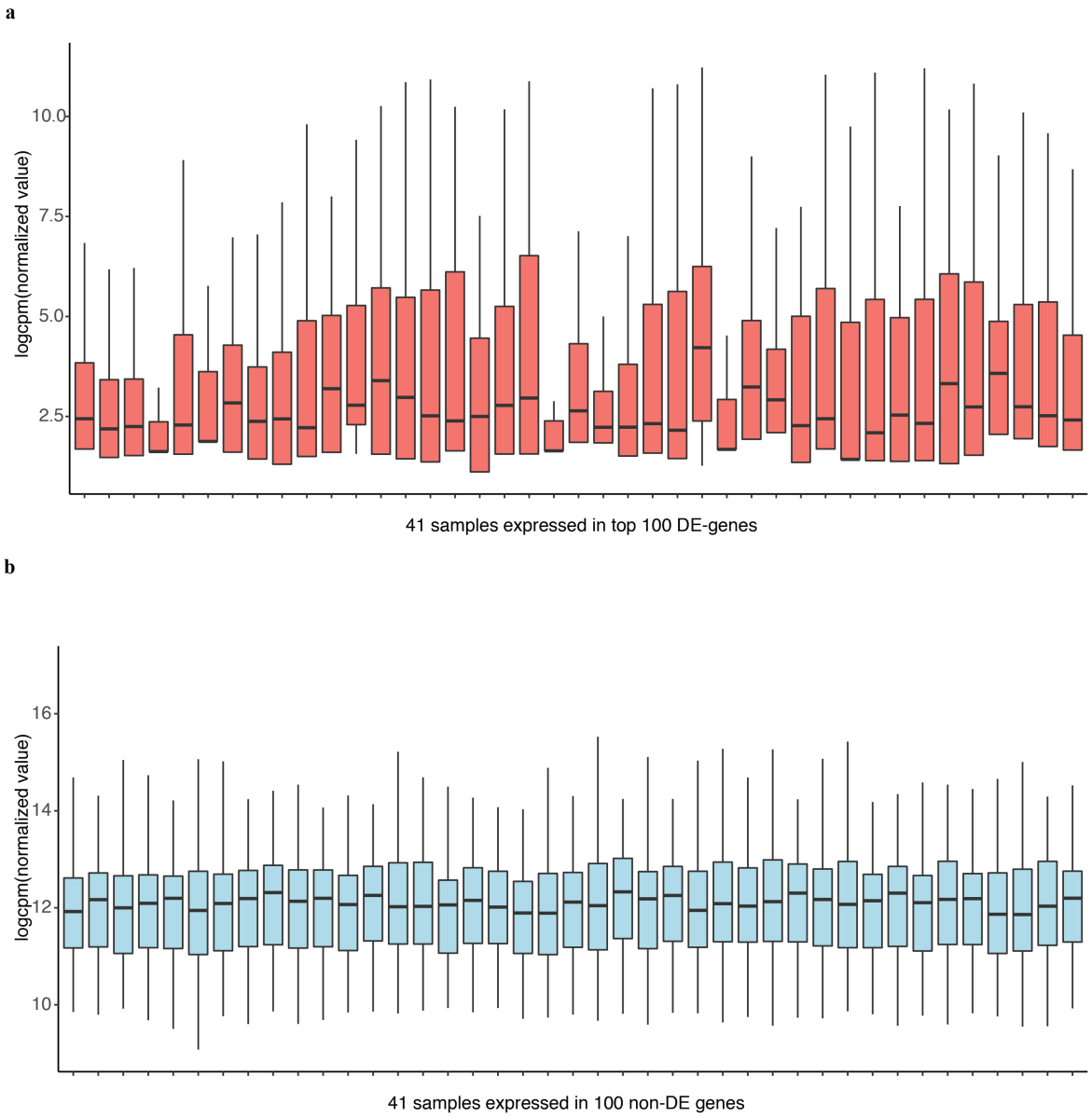

Figure 5. Box plot of the top 100 differentially expressed genes and last 100 non DE genes of 41 samples separately.

Table 6. Top five differentially expressed genes obtained from DEF in the "Cheung" dataset.

Ensembl ID	Gene Symbol	Gene Function
ENSG00000204644	ZFP57	May serve an important special function either in the mature palmar Plays an essential role in the correct development of sperm
ENSG00000007038	PRSS21	Receptor for arginine vasopressin
ENSG00000179455	MKRN	Potential calcium-dependent cell-adhesion protein.
ENSG00000170627	GTSF1	Protein coding
ENSG00000129824	RPS4Y1	multicellular organism development nuclear-transcribed mRNA catabolic process, nonsense-mediated decay

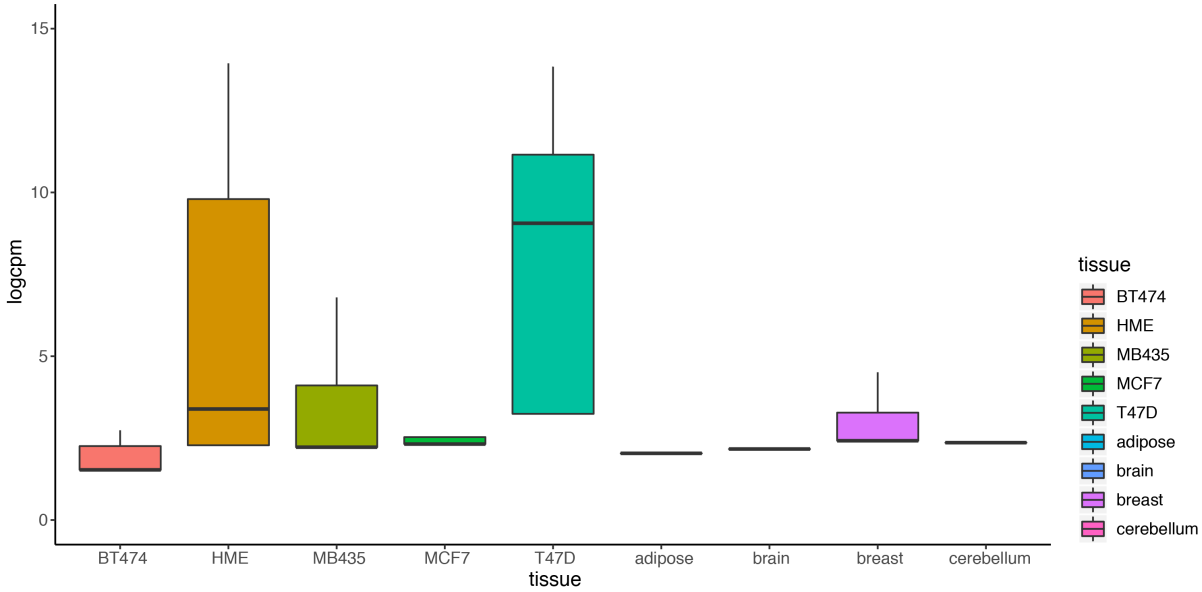

Figure 6. Box plot of the top 100 differentially expressed of nine tissues separately.

Table 7. Top five differentially expressed genes obtained from DEF in the "Wang" dataset.

Ensembl ID	Gene Symbol	Gene Function
ENSG00000100721	TCL1A	Enhances cell proliferation, stabilizes mitochondrial membrane potential and promotes cell survival Enhances the phosphorylation and activation of AKT1, AKT2 and AKT3.
ENSG00000110777	POU2AF1	It is essential for the response of B-cells to antigens and required for the formation of germinal centers
ENSG00000111348	ARHGDIB	Regulates the GDP/GTP exchange reaction of the Rho proteins by inhibiting the dissociation of GDP from them, and the subsequent binding of GTP to them
ENSG00000118308	LRMP	Plays a role in the delivery of peptides to major histocompatibility complex (MHC) class I molecules May play a role during fertilization in pronucleus congression and fusion
ENSG00000133124	IRS4	Acts as an interface between multiple growth factor receptors possessing tyrosine kinase activity Plays a pivotal role in the proliferation/differentiation of hepatoblastoma cell Plays a role in growth, reproduction and glucose homeostasis

3. Discussion

In this study, we performed a detailed comparative analysis of DEF method with several other methods for differential expression analysis by RNA-Seq data. For these methods, we focused on the specialty of generalized log-cpm normalization, especially conditions on low expression read counts. Small sample statistical analysis is still a tough task, while our proposed method could avoid the bias to some extent. Low biological or technical replicates in each sample could cause inaccuracy on the distribution-dependent methods. In contrast to other approaches, our model implies that the numbers of read counts are not entirely distinct, but they are connected between samples of replicates. This is a form of information sharing between genes and between samples, which are made possible by calculating the differential entropy-like function of the proposed model. The important contribution of this study is the solution of zero counts when performing differential expression analysis. Zero expressed gene counts give rise to the further troubling aspect that many hypothesis test methods fail. However, our model successfully

handles the two dilemmas. We emphasize that the difference does exist when one sample has zero read counts while the other does not. DEF not only shares many differentially expressed genes but also detects additional differentially expressed genes. Additionally, we studied the effect of the standard deviation on the gene-wise mean expression level. DEF is also faster and more convenient, and converts RNA-Seq data into a form that can be analyzed using a single value of a gene. It is demonstrated that combining generalized log-cpm normalization with provided DEF function can lead to a more powerful analysis to other alternatives, such as either parameter-dependent method or some other distribution-dependent method on their own. The last and the most meaningful contribution of our DEF method is its application on variable biological conditions with multi-sample and time-course data. The biological basis for diversity in gene expression between different conditions is likely to be complex. Analysis of multiple samples and time-course data helps to expose the biological bases underlying tissue diversity. The analysis of data from the same or different populations studies that each contained some population replicates showed us that the differentially expressed genes made by our strategy were likely to be biologically meaningful, as their phenotypes do have to distinguish among the gene set we test.

4. Conclusions

DEF performed as well as existing RNA-Seq methods, especially when the gene-wise expression was low. Meanwhile, DEF performed effectively on multiple conditions such as multi-sample and time-course RNA-Seq data. One characteristic of DEF is its independence of the statistical distribution of feature counts. Most methods for gene expression differential analysis are based on the negative binomial model, which seems unreasonable sometimes. Low replicates in each group cause inaccuracy in the distribution-dependent methods and zero expressed gene counts give rise to the other troubling aspect that many hypothesis test methods fail. However, DEF based on information theory successfully deals with the two issues. Following the previous studies, in differential expression analysis, an important task is to identify the genes that are expressed at higher variability across multiple samples without prior information. Accumulations of comparative studies for multi-sample data are desired, which is what we have always been focused on. DEF could work well with multiple conditions even with few replicates. The different degree of genes could be ordered by their relative DEF values. Moreover, we expect our work to inspire and support further theoretical research on modeling gene expression data and we believe that our software, DEF, will prove to be a useful addition to the existing methods for the statistical analysis of RNA-Seq and similar types of data. Some cell types will be more similar to each other, which will pose more challenges for meta-data analysis. Nevertheless, we speculate that the current method can be applied to single-cell data and comprehensive evaluations are the subsequent tasks.

5. Materials and Methods

5.1. Dataset

All datasets were obtained from the "recount" online resource http://bowtie-bio.sourceforge.net/recount/. We used the R package "recount" to get each count table combined with sample phenotype data. Table 8 lists the details of the four datasets analyzed in this article.

Table 8. Five datasets information.

Abbreviate in the Article	Number of Samples	Note
Sultan (human) [40]	4	cell type comparison
Katz (mouse) [41]	4	case and control comparison
Trappnel l(mouse) [2]	4	time course comparison
Cheung (human) [30]	41	individual comparison
Wang (human) [42]	9	tissue comparison

5.2. Normalization

Normalization of the count data is a crucial step in the analysis of RNA-Seq data, which has a strong impact on the detection of differentially expressed genes. A normalization strategy called generalized log-cpm was used. The log-cpm method is a well-accepted normalization step of gene expression by dividing the corresponding library size (in millions) of each read count shown in voom method [43]. Specifically, the entry point of the normalization algorithm is a set of n RNA samples whose sequence reads have been summarized according to the number mapping to each gene. The raw matrix Y with elements $y_{ij}(i = 1, ..., m; j = 1, ..., n)$ indicates the number of sequencing reads that have been mapped to a gene in a sample. Write Y_j for the total number of mapped reads for sample j, $Y_j = \sum_{j=1}^{n} \widetilde{y}_{ij}$.

$$\overline{y}_{ij} = log_2(\frac{y_{ij} + 0.5}{Y_j + 1} \times 10^6) \tag{1}$$

where y_{ij} is the number of reads mapped to gene i in sample j. However, log-cpm values could be negative, which is a problem in some specific conditions. Besides, the normalization size 10^6 was fixed, which may cause the log-cpm value to be too large or too small. To avoid these problems, a generalized log-cpm value was given:

$$\widetilde{y}_{ij} = log_2(\frac{y_{ij} + 1}{Y_j + 1} \times 10^k) \tag{2}$$

where $k = [log_{10}(\underset{j}{max} Y_j)] + 1$. Note that the positive $\widetilde{y}_{ij}$ was normalized by the maximal reads of each row. It is worth noting that the denominator counts Y_j was offset by one to avoid the meaningless condition when Y_j equals zero. At the same time, the numerator counts y_{ij} were augmented by a small positive value (one read) to avoid taking the logarithm of zero. Such operation not only ensures no missing generalized log-cpm values but also allows that $\frac{y_{ij} + 1}{Y_j + 1}$ is less than one as well as greater than zero. To avoid the phenomenon that it is highly possible that $\widetilde{y}_{ij}$ is negative, the parameter k played an important role to ensure that $\frac{y_{ij} + 1}{Y_j + 1} \times 10^k$ was strictly greater than one. As a result, $\widetilde{y}_{ij}$ was strictly greater than zero. Moreover, the benefit of generalized log-cpm is not only dealing with zero expression read counts but also decreasing the variance of the genes with larger RNA-Seq counts. In particular, the generalized log-cpm ensures no missing values and reduces the variability at high expressed count values. In summary, generalized log-cpm method for normalization is crucial in terms of differentiating between gene expression changes with low expressed or extremely high expressed read counts.

5.3. DEF Function

Subsequently, a novel method based on the differentially entropy-like function was used for detecting the differentially expressed genes across multiple samples in RNA-Seq data. Internally, DEF uses the normalization method generalized log-counts per million, which is a simple and reasonable scale for

normalization by providing scale factors to make counts comparable between different samples. For the normalized read counts matrix $\widetilde{Y}$, the differential entropy-like function is defined as

$$H = 1 - \frac{-\sum_{j=1}^{n}[(\frac{\widetilde{y}_{ij}}{Y_j}) \cdot log(\frac{\widetilde{y}_{ij}}{Y_j})]}{logn} \tag{3}$$

where $Y_j = \sum_{j=1}^{n} \widetilde{y}_{ij}$. Note that the value of DEF is a reasonable measurement of the differentially expressed genes. For example, gene i is expressed across all samples with reads count 1 after normalization, that is $\widetilde{y}_{i1} = \widetilde{y}_{i2} = \cdots = \widetilde{y}_{in} = 1$, then

$$H = 1 - \frac{-\sum_{j=1}^{n}[(\frac{1}{n}) \cdot log(\frac{1}{n})]}{logn} = 1 - \frac{-log(\frac{1}{n})}{logn} = 0 \tag{4}$$

which shows gene i is not differentially expressed by the differential entropy-like function. This agrees with the fact gene i is not differentially expressed across the samples. For another example, gene i is expressed differentially with read counts 1, 2, 3 after normalization across three samples. Then,

$$H = 1 - \frac{-\frac{1}{6} \times log\frac{1}{6} - \frac{2}{6} \times log\frac{2}{6} - \frac{3}{6} \times log\frac{3}{6}}{log3} = 0.079 \tag{5}$$

which shows the difference exists across these three samples. The bigger the value H is, the greater the expression difference among n samples is. For each gene, we calculated a differential entropy-like function value, which permits the direct estimation of the degree of expression. For each gene, its difference degree across multiple samples can be quantified by DEF value. This gene was defined as a differential expression when H was larger than a reasonable threshold; otherwise, it was judged as non-differential expression gene. We implemented the method presented in this article in the software package DEF, which was written in R language. A software implementation is available from https://github.com/xiaoxiaoxier/DEF. DEF expects a matrix of unnormalized count data as input and the output of the analysis is the IDs of the differentially expressed genes. When using DESeq2, edgeR and bayseq for comparison, all parameters were left at their default values. For baySeq, we took 5000 samples for estimating the priors with the quasi-likelihood approach.

Author Contributions: Conceptualization, Z.W. and S.J.; Methodology, Z.W.; Software, Z.W.; formal analysis, Z.W.; investigation, Z.W.; Writing–original draft preparation, Z.W.; Supervision, S.J. and C.Z; project administration, C.Z.

Funding: This research received no external funding.

Acknowledgments: The authors thank the editorial staff and the anonymous reviewers for insightful comments and suggestions that helped improve the quality of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DEF	Differential Entropy-like Function
NGS	Next Generation Sequencing
DEG	Differentially Expressed Gene

References

1. Wang, Z.; Gerstein, M.; Snyder, M. RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **2009**, *10*, 57–63. [CrossRef] [PubMed]
2. Trapnell, C.; Williams, B.A.; Pertea, G.; Mortazavi, A.; Kwan, G.; van Baren, M.J.; Salzberg, S.L.; Wold, B.J.; Pachter, L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **2010**, *28*, 511–515. [CrossRef] [PubMed]
3. Clavijo, B.J.; Venturini, L.; Schudoma, C.; Accinelli, G.G.; Kaithakottil, G.; Wright, J.; Borrill, P.; Kettleborough, G.; Heavens, D.; Chapman, H.; et al. An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome Res.* **2017**, *27*, 885–896. [CrossRef] [PubMed]
4. Chepelev, I.; Wei, G.; Tang, Q.; Zhao, K. Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. *Nucleic Acids Res.* **2009**, *37*, e106–e106. [CrossRef] [PubMed]
5. Velculescu, V.E.; Zhang, L.; Vogelstein, B.; Kinzler, K.W. Serial Analysis of Gene Expression. *Science* **1995**, *270*, 484–487, [CrossRef] [PubMed]
6. Kodzius, R.; Kojima, M.; Nishiyori, H.; Nakamura, M.; Fukuda, S.; Tagami, M.; Sasaki, D.; Imamura, K.; Kai, C.; Harbers, M.; et al. CAGE: Cap analysis of gene expression. *Nat. Methods* **2006**, *3*, 211–212. [CrossRef] [PubMed]
7. Little, S.C.; Tikhonov, M.; Gregor, T. Precise Developmental Gene Expression Arises from Globally Stochastic Transcriptional Activity. *Cell* **2013**, *154*, 789–800. [CrossRef] [PubMed]
8. Pertea, M.; Kim, D.; Pertea, G.M.; Leek, J.T.; Salzberg, S.L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **2016**, *11*, 1650–1667. [CrossRef] [PubMed]
9. Anders, S.; Pyl, P.T.; Huber, W. HTSeq—A Python framework to work with high-throughput sequencing data. *Bioinformatics* **2015**, *31*, 166–169. [CrossRef] [PubMed]
10. Ritchie, M.E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C.W.; Shi, W.; Smyth, G.K. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **2015**, *43*, e47–e47. [CrossRef] [PubMed]
11. Valerio, C. RNA-Seq and Human Complex Disease. *Eur. J. Hum. Genet.* **2013**, *21*, 134–142, [CrossRef]
12. Robinson, M.D.; Smyth, G.K. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **2007**, *23*, 2881–2887. [CrossRef] [PubMed]
13. Chen, Y.; Lun, A.T.L.; Smyth, G.K. Differential Expression Analysis of Complex RNA-seq Experiments Using edgeR. In *Statistical Analysis of Next Generation Sequencing Data*; Datta, S.; Nettleton, D., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 51–74. [CrossRef]
14. Lun, A.T.L.; Chen, Y.; Smyth, G.K. It's DE-licious: A Recipe for Differential Expression Analyses of RNA-seq Experiments Using Quasi-Likelihood Methods in edgeR. In *Statistical Genomics: Methods and Protocols*; Mathé, E., Davis, S., Eds.; Springer: New York, NY, USA, 2016; pp. 391–416. [CrossRef]
15. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **2014**, *15*, 550, [CrossRef] [PubMed]
16. Hardcastle, T.J.; Kelly, K.A. baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinform.* **2010**, *11*, 422. [CrossRef] [PubMed]
17. Li, J.; Witten, D.M.; Johnstone, I.M.; Tibshirani, R. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics* **2012**, *13*, 523–538. [CrossRef] [PubMed]
18. Clark, N.M.; Fisher, A.P.; Sozzani, R. Identifying Differentially Expressed Genes Using Fluorescence-Activated Cell Sorting (FACS) and RNA Sequencing from Low Input Samples. In *Computational Cell Biology: Methods and Protocols*; von Stechow, L., Santos Delgado, A., Eds.; Springer: New York, NY, USA, 2018; pp. 139–151. [CrossRef]
19. Wu, H.; Wang, C.; Wu, Z. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics* **2013**, *14*, 232–243, [CrossRef] [PubMed]
20. Vavoulis, D.V.; Francescatto, M.; Heutink, P.; Gough, J. DGEclust: Differential expression analysis of clustered count data. *Genome Biol.* **2015**, *16*, 39, [CrossRef] [PubMed]
21. Bullard, J.H.; Purdom, E.; Hansen, K.D.; Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinform.* **2010**, *11*, 94, [CrossRef] [PubMed]

22. Lin, Y.; Golovnina, K.; Chen, Z.X.; Lee, H.N.; Negron, Y.L.S.; Sultana, H.; Oliver, B.; Harbison, S.T. Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual Drosophila melanogaster. *BMC Genom.* **2016**, *17*, 28, [CrossRef] [PubMed]
23. Tarazona, S.; García-Alcalde, F.; Dopazo, J.; Ferrer, A.; Conesa, A. Differential expression in RNA-seq: A matter of depth. *Genome Res.* **2011**, *21*, 2213–2223, [CrossRef] [PubMed]
24. Tarazona, S.; Furió-Tarí, P.; Turrà, D.; Pietro, A.D.; Nueda, M.J.; Ferrer, A.; Conesa, A. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res.* **2015**, *43*, e140–e140, [CrossRef] [PubMed]
25. Shi, Y.; Chinnaiyan, A.M.; Jiang, H. rSeqNP: A non-parametric approach for detecting differential expression and splicing from RNA-Seq data. *Bioinformatics* **2015**, *31*, 2222–2224, [CrossRef] [PubMed]
26. Li, J.; Tibshirani, R. Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. *Stat. Methods Med. Res.* **2013**, *22*, 519–536, [CrossRef] [PubMed]
27. Frazee, A.C.; Langmead, B.; Leek, J.T. ReCount: A multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinform.* **2011**, *12*, 449, [CrossRef] [PubMed]
28. Rivkin, E.; Eddy, E.M.; Willis, W.D.; Goulding, E.H.; Suganuma, R.; Yanagimachi, R.; Kierszenbaum, A.L. Sperm tail abnormalities in mutant mice with neor gene insertion into an intron of the keratin 9 gene. *Mol. Reprod. Dev.* **2005**, *72*, 259–271, [CrossRef] [PubMed]
29. Cui, C.Y.; Klar, J.; Georgii-Heming, P.; Fröjmark, A.S.; Baig, S.M.; Schlessinger, D.; Dahl, N. Frizzled6 Deficiency Disrupts the Differentiation Process of Nail Development. *J. Investig. Dermatol.* **2013**, *133*, 1990–1997. [CrossRef] [PubMed]
30. Cheung, V.G.; Nayak, R.R.; Wang, I.X.; Elwyn, S.; Cousins, S.M.; Morley, M.; Spielman, R.S. Polymorphic cis- and trans-regulation of human gene expression. *PLoS Biol.* **2010**, *8*, e1000480, [CrossRef] [PubMed]
31. Plant, K.; Fairfax, B.P.; Makino, S.; Vandiedonck, C.; Radhakrishnan, J.; Knight, J.C. Fine mapping genetic determinants of the highly variably expressed MHC gene ZFP57. *Eur. J. Hum. Genet.* **2013**, *22*, 568–571. [CrossRef] [PubMed]
32. Laine, J.; Künstle, G.; Obata, T.; Sha, M.; Noguchi, M. The Protooncogene TCL1 Is an Akt Kinase Coactivator. *Mol. Cell* **2000**, *6*, 395–407, [CrossRef]
33. Pekarsky, Y.; Koval, A.; Hallas, C.; Bichi, R.; Tresini, M.; Malstrom, S.; Russo, G.; Tsichlis, P.; Croce, C.M. Tcl1 enhances Akt kinase activity and mediates its nuclear translocation. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 3028–3033, [CrossRef] [PubMed]
34. Scherle, P.; Behrens, T.; Staudt, L. Ly-GDI, a GDP-dissociation inhibitor of the RhoA GTP-binding protein, is expressed preferentially in lymphocytes. *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 7568–7572, [CrossRef] [PubMed]
35. Adra, C.; Ko, J.; Leonard, D.; Wirth, L.; Cerione, R.; Lim, B. Identification of a novel protein with GDP dissociation inhibitor activity for the ras-like proteins CDC42Hs and rac I. *Genes Chromosomes Cancer* **1993**, *8*, 253–261, [CrossRef] [PubMed]
36. Leffers, H.; Nielsen, M.; Andersen, A.; Honoré, B.; Madsen, P.; Vandekerckhove, J.; Celis, J. Identification of two human Rho GDP dissociation inhibitor proteins whose overexpression leads to disruption of the actin cytoskeleton. *Exp. Cell Res.* **1993**, *209*, 165–174, [CrossRef] [PubMed]
37. Fantin, V.; Sparling, J.; Slot, J.; Keller, S.; Lienhard, G.; Lavan, B. Characterization of insulin receptor substrate 4 in human embryonic kidney 293 cells. *J. Biol. Chem.* **1998**, *273*, 10726–10732, [CrossRef] [PubMed]
38. Qu, B.; Karas, M.; Koval, A.; LeRoith, D. Insulin receptor substrate-4 enhances insulin-like growth factor-I-induced cell proliferation. *J. Biol. Chem.* **1999**, *274*, 31179–31184, [CrossRef] [PubMed]
39. Cuevas, E.P.; Escribano, O.; Chiloeches, A.; Ramirez Rubio, S.; Román, I.D.; Fernández-Moreno, M.D.; Guijarro, L.G. Role of insulin receptor substrate-4 in IGF-I-stimulated HEPG2 proliferation. *J. Hepatol.* **2007**, *46*, 1089–1098, [CrossRef] [PubMed]
40. Marc, S.; Schulz, M.H.; Hugues, R.; Alon, M.; Andreas, K.; Matthias, S.; Martin, S.; Tatjana, B.; Aleksey, S.; Dmitri, P. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **2008**, *321*, 956–960.

41. Katz, Y.; Wang, E.T.; Airoldi, E.M.; Burge, C.B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **2010**, *7*, 1009–1015. [CrossRef] [PubMed]
42. Wang, E.T.; Rickard, S.; Shujun, L.; Irina, K.; Lu, Z.; Christine, M.; Kingsmore, S.F.; Schroth, G.P.; Burge, C.B. Alternative isoform regulation in human tissue transcriptomes. *Nature* **2008**, *456*, 470–476. [CrossRef] [PubMed]
43. Law, C.W.; Chen, Y.; Shi, W.; Smyth, G.K. Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **2014**, *15*, R29, [CrossRef] [PubMed]