Published online 30 December 2011

Nucleic Acids Research, 2012, Vol. 40, No. 9 ***e63***
doi:10.1093/nar/gkr1249

Preferred analysis methods for single genomic regions in RNA sequencing revealed by processing the shape of coverage

Michał J. Okoniewski[1,*], Anna Leśniewska[1,2], Alicja Szabelska[3], Joanna Zyprych-Walczak[3], Martin Ryan[1], Marco Wachtel[4], Tadeusz Morzy[2], Beat Schäfer[4] and Ralph Schlapbach[1]

[1]Functional Genomics Center Zurich, UNI ETH Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland, [2]Institute of Computer Science, Poznan University of Technology, ul. Piotrowo 2, 60-965 Poznan, [3]Department of Mathematical and Statistical Methods, Poznan University of Life Sciences, Wojska Polskiego 28, 60-637 Poznan, Poland and [4]University Children's Hospital, Steinwiesstrasse 75, CH-8032 Zurich, Switzerland

Received October 7, 2011; Revised November 11, 2011; Accepted December 1, 2011

ABSTRACT

The informational content of RNA sequencing is currently far from being completely explored. Most of the analyses focus on processing tables of counts or finding isoform deconvolution via exon junctions. This article presents a comparison of several techniques that can be used to estimate differential expression of exons or small genomic regions of expression, based on their coverage function shapes. The problem is defined as finding the differentially expressed exons between two samples using local expression profile normalization and statistical measures to spot the differences between two profile shapes. Initial experiments have been done using synthetic data, and real data modified with synthetically created differential patterns. Then, 160 pipelines (5 types of generator $\times 4$ normalizations $\times 8$ difference measures) are compared. As a result, the best analysis pipelines are selected based on linearity of the differential expression estimation and the area under the ROC curve. These platform-independent techniques have been implemented in the Bioconductor package rnaSeqMap. They point out the exons with differential expression or internal splicing, even if the counts of reads may not show this. The areas of application include significant difference searches, splicing identification algorithms and finding suitable regions for QPCR primers.

INTRODUCTION

The advances in the throughput of next-generation sequencers have recently enabled the sequencing of transcriptomes of many higher species. In contrast to the microarray data the whole transcriptome sequencing does not have any pre-assumptions on what transcripts are being measured. There is also a middle-of-the road solution, i.e. sequencing with enrichment of sequences of interest. Either way, RNA sequencing produces a lot of data, which is currently not fully explored.

According to Garber *et al.* (1) there are three main classes of RNA sequencing software for the secondary analysis, after the mapping of reads. These are, the finding of differential expression from read counts, finding novel regions of expression and the discovering of exonic composition of transcripts.

The first type of analysis is represented by inter alia (2–4). It is in fact similar to the analysis of microarrays, with the difference that the input count table can be adjusted to any annotation expressed by genomic ranges. In addition, the tests based upon the negative binomial distribution are expected to be at least a partial solution to the issue of few replicates, since the RNA sequencing experiments are still quite expensive in comparison to microarrays. Either way, the important condition is ensuring a proper depth of coverage (5).

The transcriptome reconstruction by discovering expressed regions and deconvolution of isoforms can be performed by (6–10). In most cases the tools use junction reads or paired reads to discover splice junctions and compose appropriate isoforms using graph methods. Those methods rely greatly on the quality of the

*To whom correspondence should be addressed. Tel: +41 44 635 39 24; Fax: +41 44 635 39 22; Email: michal@fgcz.ethz.ch

sequencing itself as junctions are still hard to quantitate precisely.

This article proposes a new way of exploring the informational value of RNA sequencing data, based upon different methods of comparison of coverage shapes. It goes beyond the first type of analysis (tests based upon counts of reads) because it takes into account not only a number, but also a distribution of reads within a genomic region. It should also be complementary to exon junction analysis, as we show methods of analysing exons where the differences in the splice sites can be discovered without using the junction data. According to the classification in (1) the novel method described here is probably closest to 'differential expression analysis' but not count based, like many current methods: (6,11–15). The goal of the research described is to devise a novel set of methods that can be used to differentiate the coverage function of reads in genomic regions. To achieve this we applied a number of transformations that process and quantify the coverage shape. The methods described below are designed to be independent of any hardware platform and mapping algorithm, so should be applicable to any type of RNA sequencing project: with an available coverage function. The testing methodology itself was inspired by the paper (16) on comparing microarray processing pipelines.

In the conclusions, it is pointed out how the new transformations and measures can be used to find alternative splicing or novel types of genomic signatures. The methods have been published as functions in the Bioconductor package *rnaSeqMap* (17) and the code is also available in the Supplementary File S1.

MATERIALS AND METHODS

Definition of the coverage difference measure

This article presents pipelines that consist of a data generator, normalization method and a function for comparing two profiles of a coverage function. The coverage function itself can be defined in the context of RNA sequencing experiment as follows:

Definition 1 Coverage function

For a genome range, defined as $D(chr,st,en,strand)$ of length $l = en - st + 1$, and set of reads G_k mapped to the region, the coverage function C_D is defined for each nucleotide in D as the count of reads that have been aligned to this nucleotide.

The coverage function is represented in the R code as the *NucleotideDistr* object (17).

Synthetic and semi-synthetic data generators

Generators are, in this context, functions that convert a coverage on a given region into another coverage function by imposing a specific type of degeneration, measured by the level of degeneration d (see Figure 1).

The synthetic data have the form of a single cycle of absolute value of a sinusoid, given by:

$$C_{sin}(k) = \left\lfloor mm * \left| \sin\left(\frac{2\pi(k - st + 1)}{l}\right)\right| \right\rfloor, \quad \forall_{k\in D}. \tag{1}$$

where $mm = \max_{k\in D}(C_D(k))$, which is the maximum of the coverage in the real biological sample.

For the second sample, one 'hump' of this coverage function is modified into a profile given by:

$$C^d_{\text{synth}}(k) = \begin{cases} C_{sin}(k) & \text{if } k \le l/2 + st - 1, \\ C_{sin}(k) * (d+1) & \text{if } k > l/2 + st - 1, \end{cases} \quad \forall_{k\in D} \tag{2}$$

where d is the 'degeneration coefficient' between 0 and 1 of the coverage profile.

In the case of semi-synthetic data, the $C(k)$ is a real coverage function from an RNA sequencing experiment, and the generators of the modified coverage are as follows:

The 'additive generator' adds a proportion of the maximum coverage to a part of the coverage function, defined by the parameter s. It is in the range $(0, l)$. By default the number of nucleotides modified $s = 0.5*l$.

$$C^{d,s}_{\text{add}}(k) = \begin{cases} C(k) & \text{if } k \le l - s, \\ C(k) + mm * d & \text{if } k > l - s, \end{cases} \quad \forall_{k\in D} \tag{3}$$

The 'multiplicative generator' scales the coverage of s nucleotides to a factor of $d+1$.

$$C^{d,s}_{\text{mult}}(k) = \begin{cases} C(k) & \text{if } k \le l - s, \\ C(k) * (d+1) & \text{if } k > l - s. \end{cases} \quad \forall_{k\in D} \tag{4}$$

where $s \in (0, l)$

The 'truncation generator' simulates a 'truncated' coverage function—in biology this could represent the case of an alternative transcription start site.

$$C^{d}_{\text{trunc}}(k) = \begin{cases} 0 & \text{if } k \le l * d, \\ C(k) & \text{if } k > l * d. \end{cases} \quad \forall_{k\in D} \tag{5}$$

The 'peak generator' simulates a peak of coverage caused by many identical reads of length rl, aligned to the region starting at position s within the region, and $mm = \max_{k\in D} C$.

$$C^{d,s,rl}_{\text{peak}}(k) = \begin{cases} C(k) + mm * d & \text{if } k \in (s, l - s - rl), \\ C(k) & \text{in all other cases.} \end{cases} \quad \forall_{k\in D} \tag{6}$$

where $s \in (0, l - rl)$ and rl is a single read length in base pairs (for example 50 base pairs).

In all these cases, the comparison is between the original coverage function C and the modified one, $C^d_{\text{generator}}$, where d is a chosen value of the degeneration coefficient between 0 and 1.

Normalization of coverage function

Normalizations of the coverage functions are used only on a particular shape in a defined genome range $D(chr,st,en,strand)$. All the normalizations presented below are local ones, which can be performed on a single coverage profile, as opposed to the global normalization methods between samples or between genes, described e.g. by ref. (11). The local normalizations can be applied for the real data after the global

normalizations, e.g. balancing the coverage values to the total sequencing output in the file.

Thus we have the following methods of normalizing a coverage shape:

Min-Max normalization.

$$N_{mM}(C(k)) = \frac{C(k) - \min\limits_{k\in D}(C(k))}{\max\limits_{k\in D}(C(k)) - \min\limits_{k\in D}(C(k))}, \quad \forall_{k\in D} \tag{7}$$

This normalization takes into account the minimal and maximal values of the coverage, scales the profile according to these and fits this into the range <0, 1>

Density normalization.

$$N_D(C(k)) = \frac{C(k)}{\sum_{st}^{en} C(k)}, \quad \forall_{k\in D} \tag{8}$$

This transformation divides each value by the sum of all reads for all the nucleotides within range, so gives scaling by a fixed factor that also moves the values into the $<0, 1>$ range. It is required for the case where the coverage function is supposed to be treated as a density function of specific nucleotide expression. With this transformation, the coverage function will fulfill the assumptions of being a density function.

It is possible to combine the normalizations one after another or not to use normalization at all. Then the notation is N_{mMD} and N_{none}, respectively.

Difference measures

In this study, a number of difference measures have been used to calculate the distance between coverage shapes. The domain of the coverage function is a set of natural numbers within the contiguous range of nucleotides, so it is not possible to apply operators like derivatives or integration from calculus. That is way we use the operator *int* (pseudo-integral) and *diff* (pseudo-derivative) of the coverage function C. The first one is defined in the range <*a*,*b*> as follows:

$$\text{int}_a^b(C) = \sum_{k=a}^{b} \frac{C(k)}{b-a+1}, \tag{9}$$

where a and b are some values from range D that $a < b$. This operator has similar interpretation as the integral of the function.

By analogy, *diff* operator of the coverage function C is defined as:

$$diff(C(k)) = C(k) - C(k-1), \quad \forall_{k\in(st,en]} \tag{10}$$

It is defined on the discrete domain and gives the information about changes in the shape of function C.

The following measures have been considered:

Area under the curve of differences 1 (DA). The first difference measure has following form:

$$M_{\text{DA}} = \text{int}_{st}^{en}(|C_1 - C_2|), \tag{11}$$

where C_1, C_2 are the coverage functions to be compared. It does not need any normalization. However, if coverage functions are normalized to the range <0, 1> the the values of M_{DA} are in this range as well.

Area under the curve of differences 2 (DDA). This measure is similar to the previous one. However, it uses $diff(C_1)$ and $diff(C_2)$ instead of C_1 and C_2, respectively. It can be written as follows:

$$M_{\text{DDA}} = \text{int}_{st+1}^{en}(|diff(C_1) - diff(C_2)|). \tag{12}$$

QQ measure 1 (QQ). In this case it is assumed that the data coming from C_1, C_2—two considered coverage functions after normalization are normally distributed. Based on this assumption, quantiles of the data are derived. The difference measure in this case is computed as follows:

$$M_{\text{QQ}} = \sqrt{\sum_{k=st}^{en} \frac{(x_k - y_k)^2}{2l}}, \tag{13}$$

where x_k, y_k are the quantiles of $C_1(k)$ and $C_2(k)$, respectively.

QQ measure 2 (QQD). To determine the next difference measure, first $diff(C_1)$ and $diff(C_2)$ are computed. Similarly to the previous measure, it is assumed that the data coming from the considered functions $diff(C_1)$ and $diff(C_2)$ after normalization are normally distributed and the appropriate quantiles are derived. The difference measure is then of the following form:

$$M_{\text{QQD}} = \sqrt{\sum_{k=st+1}^{en} \frac{(x'_k - y'_k)^2}{2l}}, \tag{14}$$

where x'_k, y'_k are the quantiles of $diff(C_1(k))$ and $diff(C_2(k))$, respectively.

PP measure 1 (PP). This difference measure requires density normalization. Coverage functions C_1 and C_2 after this normalization fulfill the conditions of being the probability density function of the expression level of nucleotide k from range D. After transformation to the cumulative distribution functions values of $C_1(k)$ and $C_2(k)$ are considered as the coordinates for the pp-plot. Based on this, the difference measure M_{PP} is derived as follows:

$$M_{\text{PP}} = \sqrt{\sum_{k=st}^{en} \frac{(C_1(k) - C_2(k))^2}{2l}}, \tag{15}$$

PP measure 2 (PPD). This measure is similar to the M_{PP}. However, it is based on $diff(C_1)$ and $diff(C_2)$ functions instead of the coverage functions C_1 and C_2. This means that after density normalization $diff(C_1)$ and $diff(C_2)$ it can be treated as the density functions of differences in the expression level between adjacent nucleotides. After transformation to the cumulative distribution functions, values

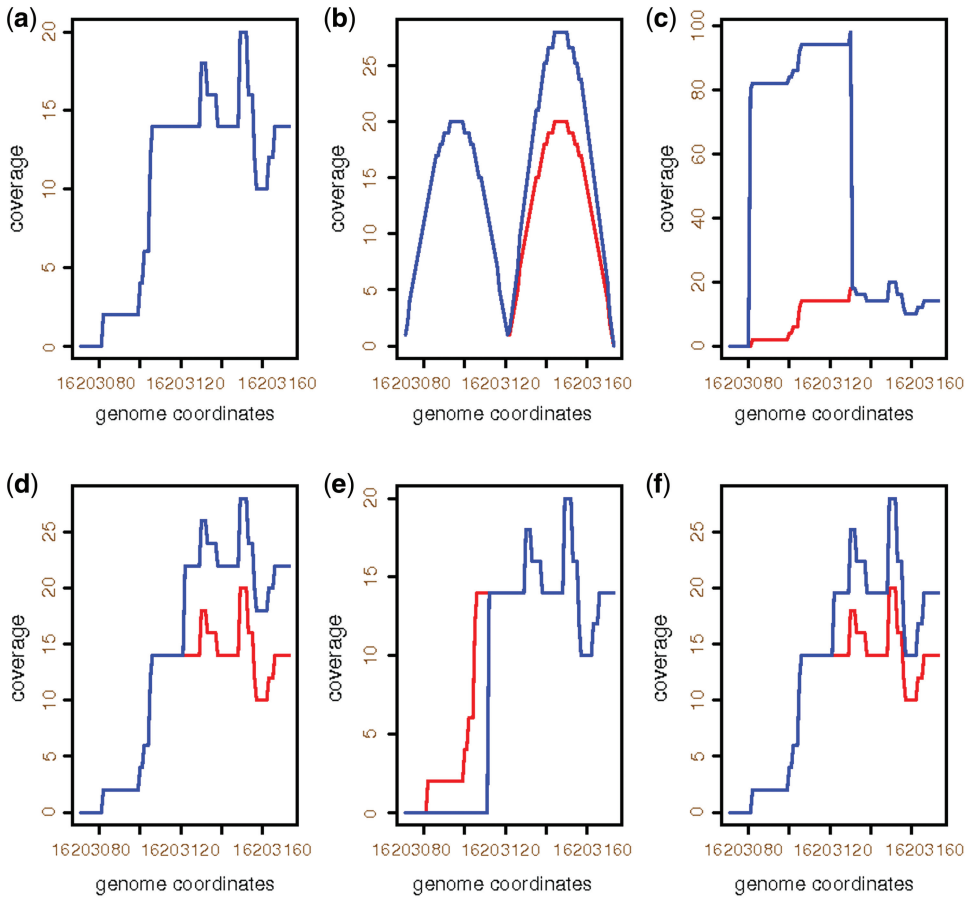

Figure 1. RNA seq coverage profiles for a single exon, transformed by data generators with the degeneration coefficient $d = 0.4$. The red profile is the original one, while blue (partially overlapping with the red) is the modified profile. (**a**) Original coverage function (**b**) Synthetic data of the same domain length (**c**) Peak generator, $s = 0.5$, $rl = 50$ (**d**) Additive generator, $s = 0.5$ (**e**) Truncation generator (**f**) Multiplicative generator, $s = 0.5$.

of $diff(C_1(k))$ and $diff(C_2(k))$ are considered as the coordinates for the pp-plot. Based on this, the difference measure M_{PPD} is derived as follows:

$$M_{\mathrm{PPD}} = \sqrt{\sum_{k=st+1}^{en} \frac{(diff(C_1(k)) - diff(C_2(k)))^2}{2l}}, \quad (16)$$

Local extrema heuristics 1 (HD1). This measure is called the 'hump difference' as it operates on the extrema of coverage profiles that often have a shape reminiscent of camels (although with more than two humps). For this measure normalization that results with the values of coverage function in the range $<0,1>$ is needed. We denote L_1 and L_2 as sets of nucleotides for which all the local maxima of coverage functions C_1 and C_2 appear, respectively. Let $L = L_1 \cup L_2$. In that case the M_{HD1} difference measure is defined as follows:

$$M_{\mathrm{HD1}} = \frac{\sum_{k \in L} |C_1(k) - C_2(k)|}{\#L}, \quad (17)$$

Since the coverage function after normalization has values in range $<0,1>$, then M_{HD1} measure is the range $<0,1>$ as well. The notation # here means the count of the set of extrema.

Local extrema heuristics 2 (HD2). The last difference measure is similar to the previous one, but with a different normalization factor in the denominator. Using the same notation it has following form:

$$M_{\mathrm{HD2}} = \frac{\sum_{k \in L} |C_1(k) - C_2(k)|}{2 * \min\{\#L_1, \#L_2\}}, \quad (18)$$

If some $k \in L_1$ also belongs to L_2, then M_{HD2} results in lower values compared to M_{HD1}. On the other hand, the difference in the counts of L_1 and L_2 increases the value of this measure compared to the previous one. In the case when $\#L_1 = \#L_2$ and $L_1 \cap L_2 =$, M_{HD2} gives the same results as M_{HD1}.

Numeric experiments processing flow

Numeric experiments have been conducted using synthetic, semi-synthetic and real data (Figure 2). In all the cases, a combination of normalization and statistical measure has been tested. In the case of synthetic and semi-synthetic data, appropriate data generators, as described above, have been used. In all the cases,

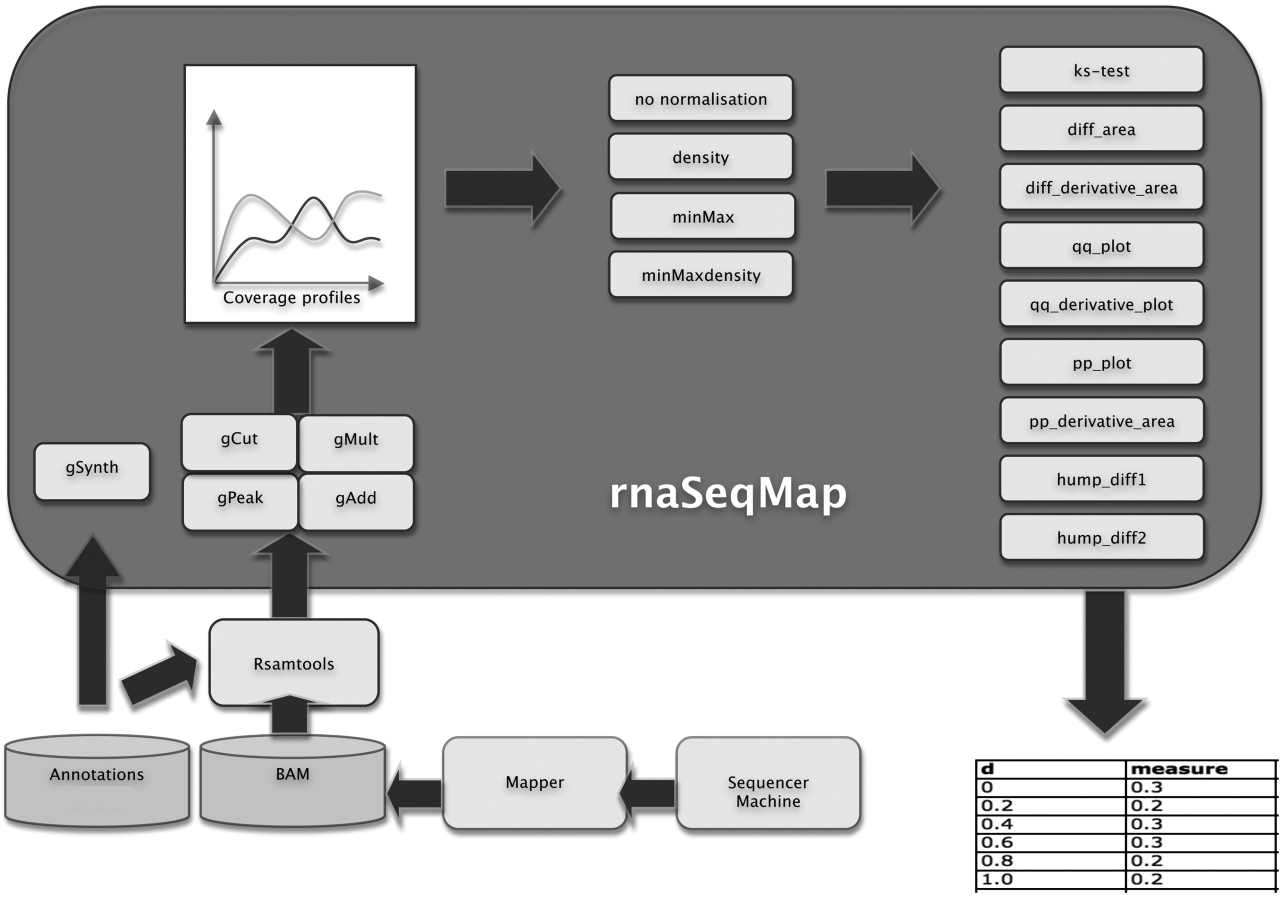

Figure 2. Pipeline for processing the coverages. The data from a short read sequencer may be mapped by any mapper and processed into BAM files with known genomic annotation. Then, using the Bioconductor libraries RSamtools and rnaSeqMap, they are processed as coverage profiles using generators of modifications, normalizations and statistical measures. Finally, the output of the measures and their matching degeneration levels are checked using correlations and ROC curves.

3000 randomly selected exonic regions from human chromosome 1 have been analysed.

Synthetic data. In this case, only the regions' genome coordinates and maximal coverage levels have been used to construct the profiles with the generators C_{sin} and C_{synth}. For each of the 3000 regions both profiles have been generated, with a random level of degeneration d, ranging from 0 to 1. Then all the combinations of the normalization and measure have been calculated for all the pairs.

Semi-synthetic data. This case took the first profile in the pair from real coverage in a rhabdomyosarcoma sample. Then, using the generators C_{peak}, C_{add}, C_{trunc} and C_{mult} the second profile was created, using the fixed parameters $s = 0.5$, $rl = 50$ as described in equations (3), (4) and (6) and the random d level vector, as for synthetic data. Once again, for all the pairs of real and generated profile, the normalizations were performed and measures calculated.

For both synthetic and semi-synthetic datasets, the relationship between the values of the measures and the degeneration level d has been taken into account. The processing pipelines (consisting of generator, normalization and measure) were compared based on the linearity of the measures as a function of d, according to the Pearson correlation. In addition, the measures are treated as binary classifiers using the cutoff level of $d = (0.2, 0.4, 0.6, 0.8)$. ROC curves were calculated and best measures selected using the area under the curves.

Real data. In the case of real data, the same genomic ranges have been used, but coverage profile pairs were taken from two samples of an alveolar and an embryonal rhabdomyosarcoma sample (BAM files with genes available in Supplementary File S2). Initially, the second sample has been normalized to the total number of reads in the sequencing output by multiplying all the coverage values by a factor of 2.216128. Then all the normalizations and measures described above were performed on the two samples. The coverage pipelines have been also compared to the count-based fold change and P-values from DESeq test (2).

RESULTS

Synthetic data experiment

For the synthetic data, there is a group of combinations of the normalization and difference measures that can distinguish well between the original symmetric bimodal coverage and the coverage with one of the maxima

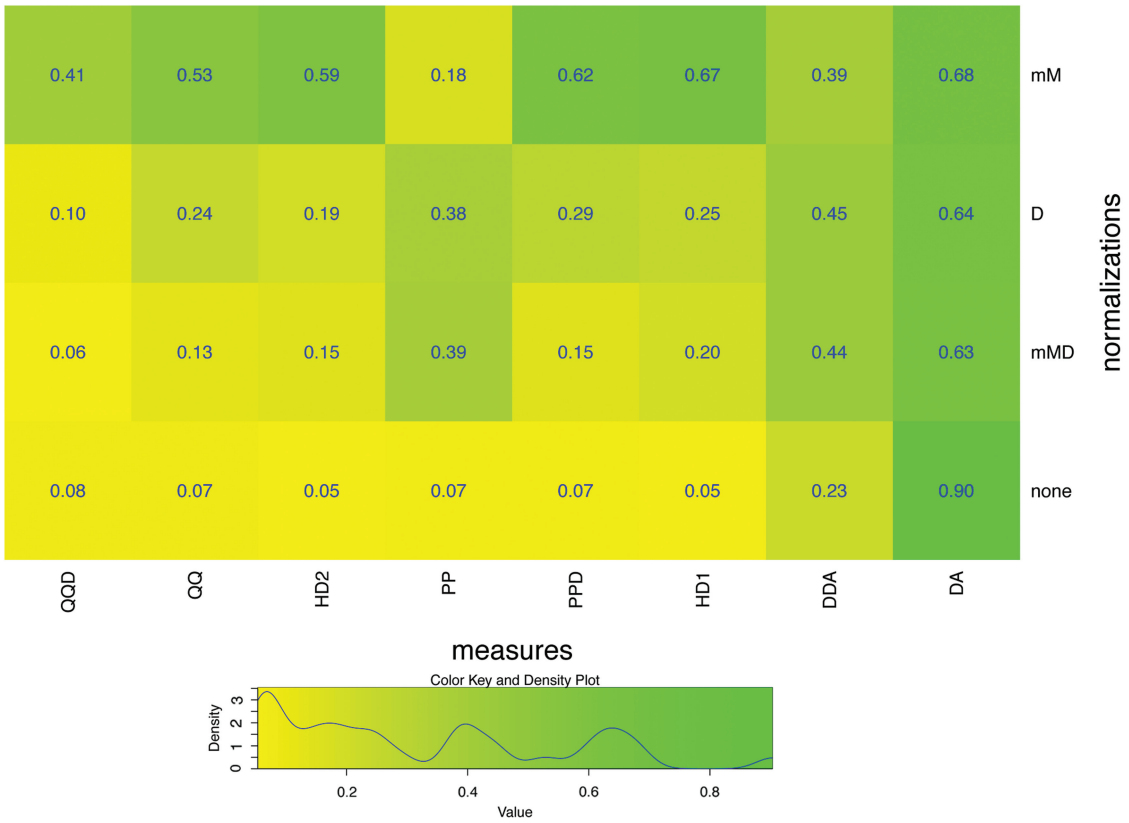

Figure 3. Heatmap of correlations for the synthetic data with normalizations in rows and measures in columns. The best correlation between M and d is observed for M_{DA} and the measures normalized by Min-Max. This heatmap table presents the values for the combinations of normalizations and measures.

increased. There are 10 combinations that have correlation of d and a measure higher than 0.8 (Figure 3), and all the values of AUC higher than 0.8 (for all the thresholds of d, see Figure 4). These are: $N_D M_{\mathrm{PP}}$, $N_{mM} M_{\mathrm{PP}}$, $N_{mM} M_{\mathrm{DA}}$, $N_D M_{\mathrm{DA}}$, $N_{mMD} M_{\mathrm{DA}}$, $N_{\mathrm{none}} M_{\mathrm{DA}}$, $N_{mM} M_{\mathrm{PPD}}$, $N_{mM} M_{\mathrm{QQ}}$, $N_{mM} M_{\mathrm{HD1}}$, $N_{mM} M_{\mathrm{HD2}}$.

Among those are all the pipelines for the M_{DA}. The measure M_{PP} performs well only in the case of density or min-max-density normalizations. The other four measures, including local extrema heuristics perform well only with min-max normalization. All the other pipelines have much worse results of both correlation and AUC (see Figure 4 and Supplementary File S3).

Semi-synthetic data

In the case of semi-synthetic data, the results differ according to the generator applied, as each of them simulates different transcriptomic phenomena.

Additive generator. This generator modifies the real coverage function in such a way that d of the maximum coverage value is added to the part of the genome region. This is the way in which splicing within the exon would occur due to alternative transcription start or end sites.

In the case of the additive generator, the M_{DA} measure with no normalization, by far outperforms all the other pipelines (Figures 5 and 6). The M_{DA}, with any of the normalizations, is still one of best measures, especially for low levels of threshold d. The next best measure is M_{HD1} after Min-Max normalization.

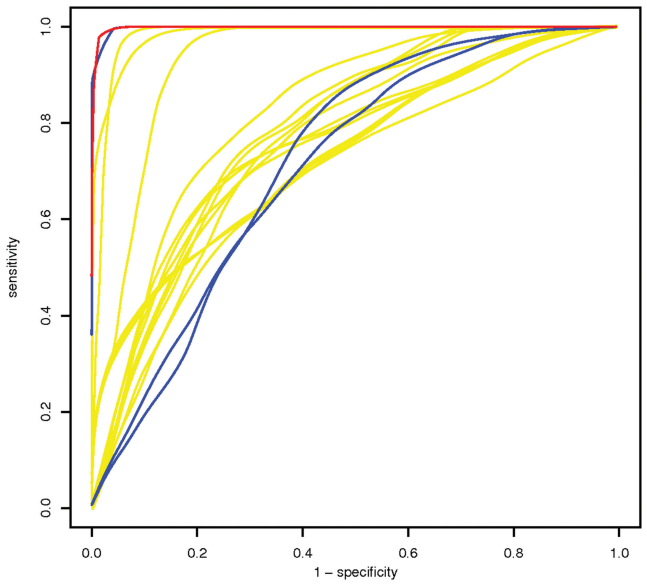

Figure 4. ROC curves for the synthetic data. The curve for the $N_{\mathrm{none}} M_{\mathrm{DA}}$ method is marked in red, while those in blue are for the M_{PP} measure with different normalizations. Curves for all other pipelines are yellow.

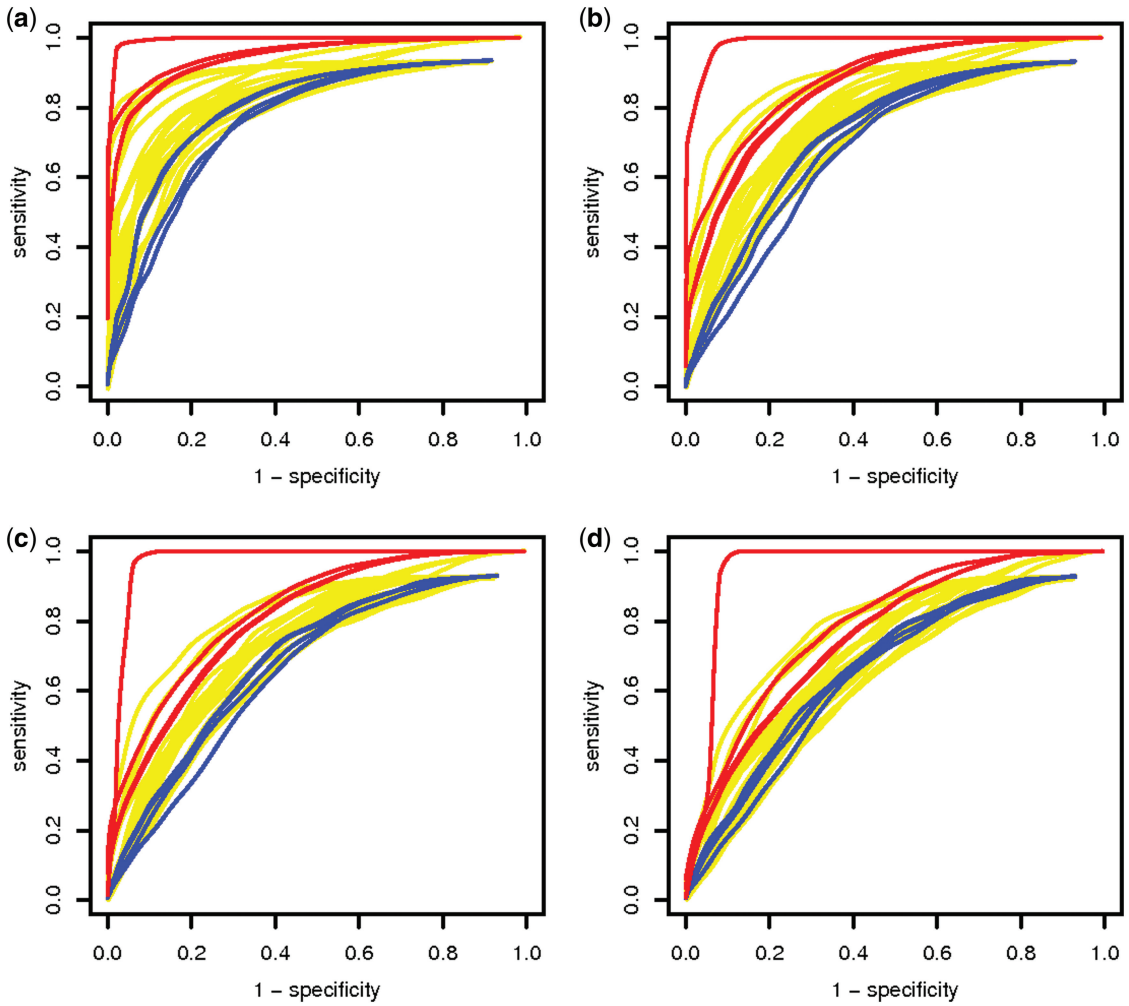

Figure 5. ROC curves for the additive generator for the thresholds of d level 0.2, 0.4, 0.6, 0.8 (**a**, **b**, **c**, **d**, respectively). In red are marked the curves for the M_{DA} method, while in blue are marked those for M_{PP}.

Truncation generator. The generator C_{trunc}, like the additive generator, also simulates the influence of alternative transcription start sites in the studied region. However, in this case, the effect of an alternative transcription start site is not a mixture of two exon effects, but the switching on of the transcription in a place not defined as an exon boundary.

The M_{DA} measure performs very well in this case too, especially for small levels of d threshold. The classifying efficiency becomes worse for the normalized M_{DA} and higher d. For higher d, the M_{PP} with density normalization has the highest AUC result, also M_{QQ} performs relatively well.

Multiplicative generator. This generator is expected to simulate the situation where the increase in expression in the part of the region is proportional to its value for each nucleotide. It is assumed that the transcription machinery produces longer and shorter versions of the exon by doing multiple runs over the DNA. The shape of coverage is therefore mainly the result of phenomena such as GC content-related sequence amplification in the sequencer.

In this case, methods with no normalization perform better than others. Best in terms of AUC is again not normalized M_{DA}, closely followed by density normalized M_{PP}. However, the shape of ROC curves is highly different for various pipelines. The M_{PP} with density normalization reaches a high level of true positives very fast, but then gets flat almost asymptotically, while other measures can reach almost 100% of sensitivity for higher levels of false negatives.

Peak generator. This generator simulates a peak of the width of a single read in the coverage profile, so it has a different interpretation. The measures that perform well on the data obtained with this generator, find artifacts rather than real biological phenomena. As expected, the M_{DA} does not find the difference as efficiently as other measures. Still, the other measures have the best predictive power here, when used without normalization.

For the full set of correlations and AUC values and for the full set of plots of analytic pipelines performance, see the Supplementary File 3.

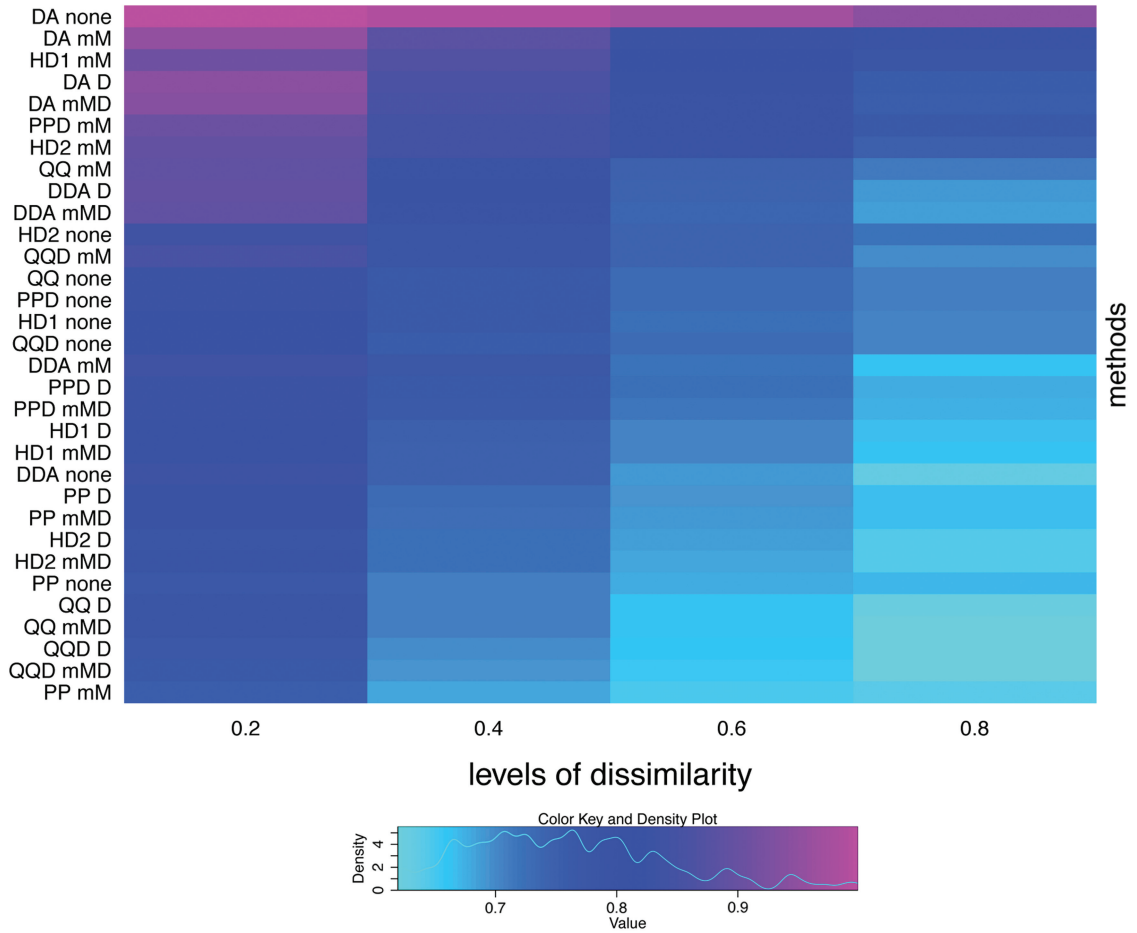

Figure 6. Heatmap of the area under the ROC curves for the additive generator, for the thresholds of d level 0.2, 0.4, 0.6, 0.8. In the top rows are those pipelines that are good classifiers in terms of the AUC.

Real data analysis

The results of testing all the pipelines on two samples of real data are presented as a heatmap (Figure 7). Although the pipelines have different ranges of the results on the log scale, they tend to agree on most of the regions. There is some 10% of the regions (right side of the heatmap) where the pipelines give highly spurious numeric results. In particular, the M_{HD1} and M_{HD2} tend to give results contrary to the other measures for this fraction of genomic regions.

It can be also observed that the pipelines happen to cluster together; there is a cluster that all the non-normalized pipelines fall into, except the M_{DA}—which clusters with most of the normalized pipelines.

The comparison to the count-based methods with the best performing pipelines (M_{DA} without normalization and M_{PP} with density normalization) is shown in the Figure 8. Although the correlation between count-based fold changes and the measures reaches 0.4 in some cases, there is no clear correspondence between the count-based methods and the studied pipelines. This proves that there is always a group of exons that will not be found as differentially expressed according to counts, but will be clearly different in terms of the shape of coverage. Examples of such exon coverages from the two real data samples are presented in the Supplementary Figure S1.

DISCUSSION

All of the pipelines tested differentiate between real and modified coverage profiles in most of the cases. However, the efficiency of this classification varies by generator and by the shape of the profile. In particular, several pipelines such as M_{DA} without normalization and M_{PP} with density normalization have proven to be useful for finding the differences in more than one type of data generator.

One of the difficulties in applying the measures described in this article is that they do not have a predictable range and distribution of values. As can be seen in the Supplementary File S3, the measures without normalization especially, tend to have high values for those genomic regions where the coverage and its differences are high. This makes it difficult to combine the measures into heuristics by averaging or weighting. Still, their predictive value to find significant differences of expression remains. For this reason, the correlation check of d versus M was performed—as the ideal measure is also a linear one.

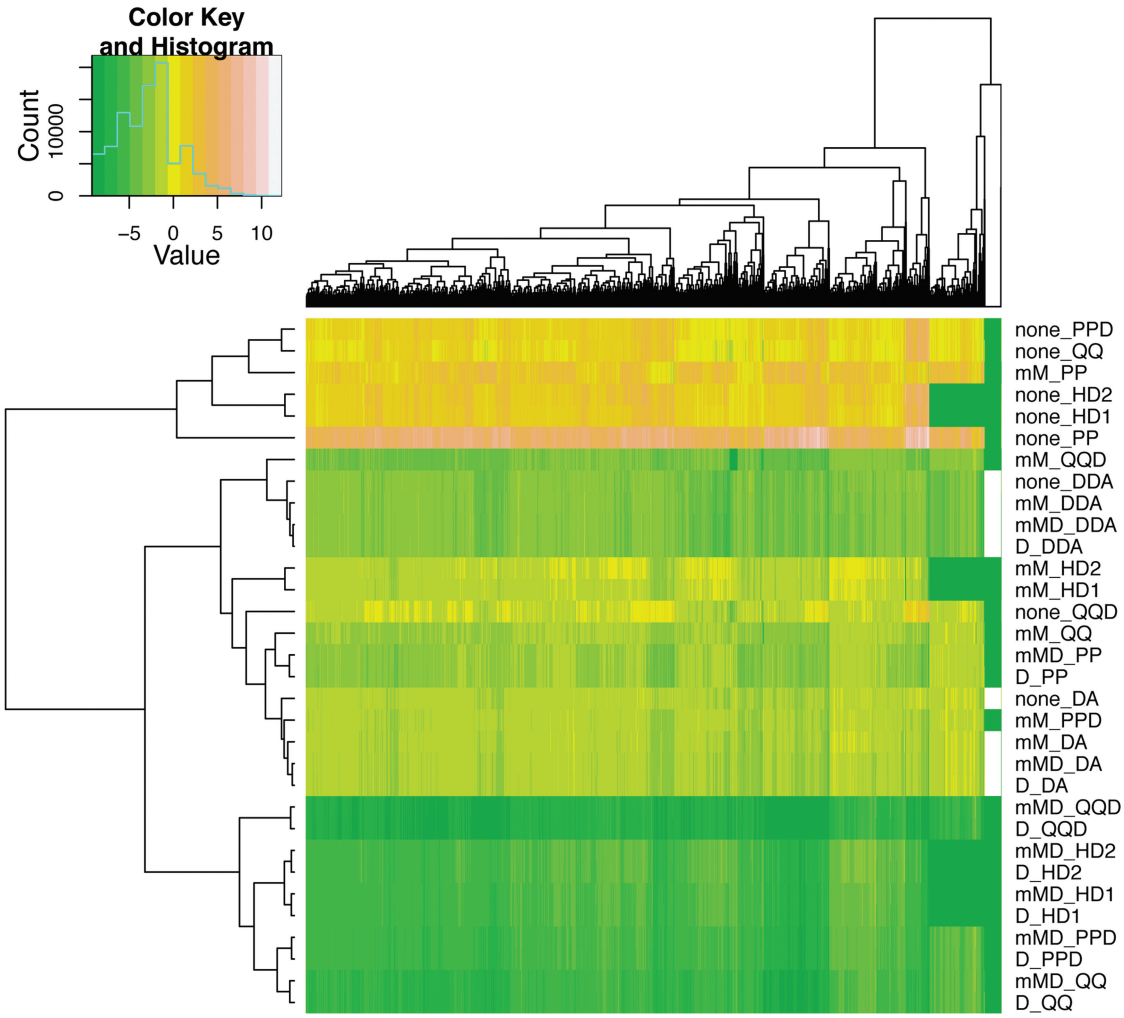

Figure 7. Heatmap for all the pipelines run on 3000 real exons. Rows represents pipelines, columns represent exonic regions, the color depicts the log2 of the difference between two real samples given by the specific pipeline.

In classic statics and optimization theory there are various test and formulae for measuring a goodness of fit of the functions such as Kolmogorov–Smirnov test. However in most cases they involve specific assumptions for instance about normality of the data or continuos domain, which do not hold in the case of coverage function.

There are several advantages of such novel problem formulation involving the use of these local measures that differentiate the samples in every region:

- it is possible to 'work with only partial information' about the sequencing experiment—i.e. the minimum analysis is a single region measured for two samples with no replicates. There is no global model needed [like in (2–4)] to differentiate between the expression shapes in a region, i.e. there is no need to know the global number of reads in the samples.
- For the above reason, the analysis 'does not need high processing power' to get the results. The complexity of computation is linear with the number of genes and linear with the average size of the region. There are no special memory requirements—it is possible to process a single region at a time in the same object slot.
- it is 'platform independent' as there are no assumptions about the sequencing machine and mapping algorithm. Nevertheless, the methods may behave differently in the case of coverage shapes very different from those tested here and may need additional tuning. However, most RNA sequencing experiments seem to have similar shapes and similar artifacts of coverage. Further cross-platform research may be needed.
- All the measures can point out the differences in the expressed regions, even when they cannot be spotted by the difference of counts, because the analysis of the shape is far more involved than just comparing two numbers. Cases where the count numbers are similar and coverage shapes are different are easy to spot in the data sets.

The pipelines and measures may be applicable in several critical applicability areas of RNA sequencing:

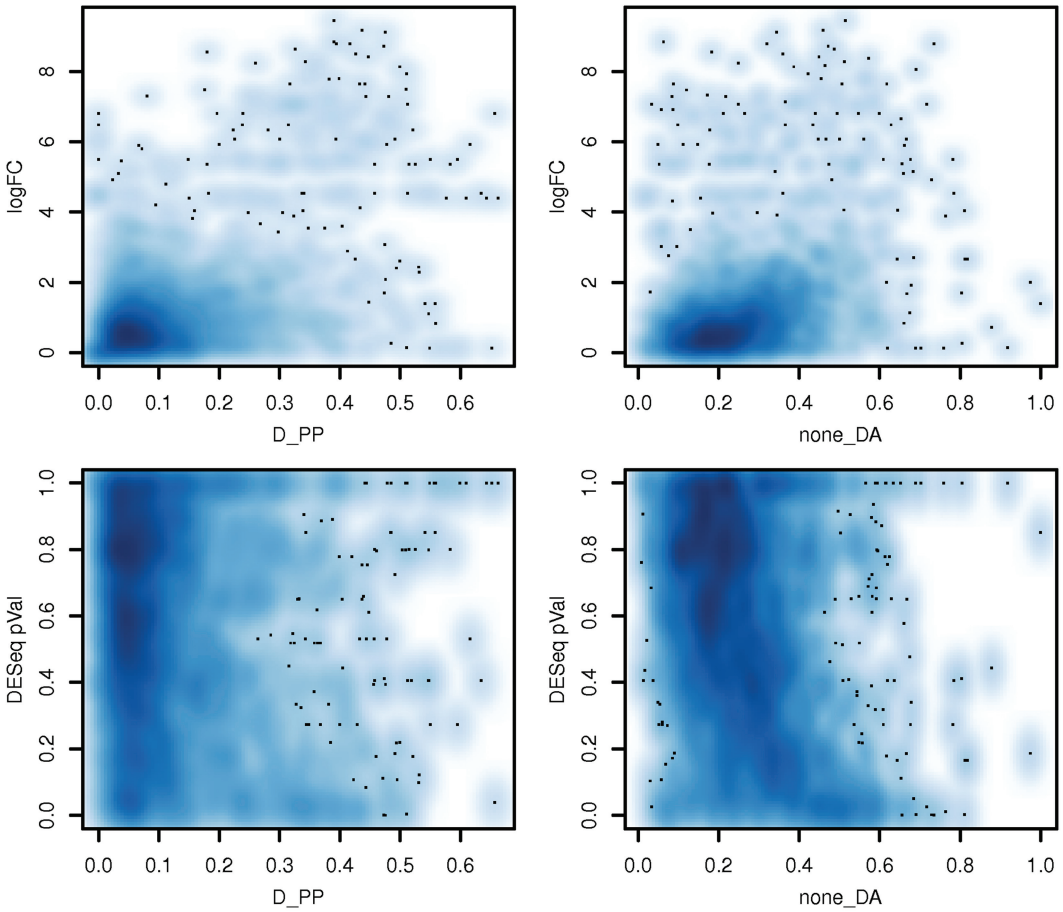

Figure 8. Scatterplots of M_{DA} without normalization and M_{PP} with density normalization against the log2 fold change and *P*-value from the DESeq test for the 3000 exons in the real data experiment.

- **Significance search**—for both well defined exons and newly discovered expressed regions.
- In **splicing analysis**, the measures may be a base for algorithms of splicing assessment—eg. replace the exon expression proportion in the classic splicing index (18).
- The discrepancies in the results of the measures that cannot be explained by overlapping of exon variants may 'suggest novel transcription start/end sites' and, therefore, new isoforms.
- The results of the pipelines may point out good and improper places to design the primers for a 'QPCR verification of RNA sequencing' results.

Additionally, the findings of this article could be applied to analyzing other types of sequencing data in transcriptomics, such as chip-seq or exome enrichment sequencing. In the case of chip-seq, there is already a publication considering the shape of coverage (19), but it describes an unsupervised method for discovering the peaks.

Like the paper of Choe *et al.* (16), this study gives indications as to which of the pipelines may be most useful for particular types of significance search. To avoid the controversy in testing the methods only with synthetic data (20), semi-synthetic and real data have also been applied for the tests—showing that there is a link between the findings in all three approaches. The point of the experiments presented in this article is not just to show the best method, but by extensive data mining to understand the relationships between the biological phenomena of the transcriptome, coverage profiles splicing and their possible artifacts.

CONCLUSION

The article consists of problem formulation and, based upon it, experimental evaluation of a novel set of methods for RNA sequencing data analysis, using the comparison of coverage profiles in genomic regions. To show the utility of those methods, a considerable amount of statistical experiments have been performed.

The methodology may be applied to find transcript variants not limited to the well-known ones, and to be used for local searches for significant RNA expression difference. This is possible even in the case of those genomic sequences that do not have established annotation e.g. non-coding RNA. In the biological experiment context, it may be applied to find the exons with stable expression in order to define the QPCR primers. The further development of these methods may help in the research on constantly evolving and increasingly complex field of deciphering the transcriptional code of nature.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Files 1–3, Supplementary Figure 1.

ACKNOWLEDGEMENTS

The authors would like to thank Remy Bruggmann and Marzanna Künzli for all the valuable discussions during the preparation of the experiments and manuscript.

FUNDING

Scientific Exchange Programme NMS-CH (sciex.ch) grant nr 09.025. Funding for open access charge: University of Zurich.

Conflict of interest statement. None declared.

REFERENCES

1. Garber,M., Grabherr,M.G., Guttman,M. and Trapnell,C. (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods*, **8**, 469–477.
2. Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.
3. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
4. Wang,L., Feng,Z., Wang,X., Wang,X. and Zhang,X. (2010) DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, **26**, 136–138.
5. Tarazona,S., García-Alcalde,F., Dopazo,J., Ferrer,A. and Conesa,A. (2011) Differential expression in RNA-seq: A matter of depth. *Genome Res.*, **21**, 2213–2223.
6. Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., van Baren,M.J., Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnol.*, **28**, 511–515.
7. Roberts,A., Pimentel,H., Trapnell,C. and Pachter,L. (2011) Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics*, **27**, 2325–2329.
8. Guttman,M., Garber,M., Levin,J.Z., Donaghey,J., Robinson,J., Adiconis,X., Fan,L., Koziol,M.J., Gnirke,A., Nusbaum,C. *et al.* (2010) Ab initio reconstruction of cell typespecific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnol.*, **28**, 503–510.
9. Robertson,G., Schein,J., Chiu,R. and Corbett,R. (2010) De novo assembly and analysis of RNA-seq data. *Nat. Methods*, **7**, 909–912.
10. Bohnert,R. and Rätsch,G. (2010) rQuant.web: a tool for RNA-Seq-based transcript quantitation. *Nucleic Acids Res.*, **38(Web Server issue)**, W348–W351.
11. Robinson,M.D. and Oshlack,A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, **11**, R25.
12. Li,B., Ruotti,V., Stewart,R.M., Thomson,J.A. and Dewey,C.N. (2010) RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, **26**, 493–500.
13. Jiang,H. and Wong,W.H. (2009) Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, **25**, 1026–1032.
14. Wu,Z., Wang,X. and Zhang,X. (2011) Using non-uniform read distribution models to improve isoform expression inference in RNA-Seq. *Bioinformatics*, **27**, 502–508.
15. Langmead,B., Hansen,K. and Leek,J. (2010) Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biol.*, **11**, R83.
16. Choe,S.E., Boutros,M., Michelson,A.M., Church,G.M. and Halfon,M.S. (2005) Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biol.*, **6**, R16.
17. Leśniewska,A. and Okoniewski,M.J. (2011) rnaSeqMap: a Bioconductor package for RNA sequencing data exploration. *BMC Bioinformatics*, **12**, 200.
18. Gardina,P., Clark,T., Shimada,B., Staples,M., Yang,Q., Veitch,J., Schweitzer,A., Awad,T., Sugnet,C., Dee,S. *et al.* (2006) Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. *BMC Genomics*, **7**, 325.
19. Hower,V., Evans,S.N. and Pachter,L. (2011) Shape-based peak identification for ChIP-Seq. *BMC Bioinformatics*, **12**, 15.
20. Dabney,A.R. and Storey,J.D. (2006) A reanalysis of a published Affymetrix GeneChip control dataset. *Genome Biol*, **7**, 401.