Lognormality and oscillations in the coverage of high-throughput transcriptomic data towards gene ends

Nicolas Innocenti$^1$ and Erik Aurell$^{1,2}$

$^1$ Department of Computational Biology, KTH Royal Institute of Technology, AlbaNova University Center, SE-10691 Stockholm, Sweden
$^2$ Department of Information and Computer Science, Aalto University, PO Box 15400, FI-00076 Aalto, Finland
E-mail: njain@kth.se and eaurell@kth.se

Received 3 July 2013
Accepted 16 September 2013
Published 17 October 2013

Abstract. High-throughput transcriptomics experiments have reached the stage where the count of the number of reads alignable to a given position can be treated as an almost-continuous signal. This allows us to ask questions of biophysical/biotechnical nature, but which may still have biological implications. Here we show that when sequencing RNA fragments from one end, as is the case on most platforms, an oscillation in the read count is observed at the other end. We further show that these oscillations can be well described by Kolmogorov’s 1941 broken stick model. We investigate how the model can be used to improve predictions of gene ends (3′ transcript ends), but conclude that with present data the improvement is only marginal. The results highlight subtle effects in high-throughput transcriptomics experiments which do not have a biological origin, but which may still be used to obtain biological information.

Keywords: gene expression and regulation (theory), gene expression and regulation (experiment), bioinformatics, genomics

ArXiv ePrint: 1303.4229
1. Introduction

Next generation sequencing (NGS) is the common name used for most sequencing platforms currently (August 2013) available on the market and widely used in genomic and post-genomic research. In the present paper we investigate a common effect on most such platforms which is relevant in transcriptomics, i.e. when they are used to sequence RNA extracted from some organism. Let us first remark that the outcome of an NGS experiment is a number count for each position on the genome of the organism under investigation. This number count is called the coverage, and the average of the coverage over the genome, or over the expressed parts of a genome in a transcriptomics experiment, is called the sequencing depth. For the sequencing of the human genome a sequencing depth of 8–9× was used, i.e. the total length of all useable sequenced fragments was equivalent to eight to nine times the length of the studied genome. With current technologies one can go much higher. For the bacterial transcriptomics data which will be used in the present study the sequencing depth was about 450× and we can therefore treat our coverage signal as almost continuous. In a larger (more costly) experiment today, or in the future, one can certainly also envisage such sequencing depths on human data, which opens up the interesting perspective of using data analysis techniques from physics or signal processing in genomics more broadly.

All the common NGS platforms currently available on the market function by sequencing in parallel many short DNA molecules, typically 25–500 nucleotides (nt) in length. These short molecules are obtained by fragmenting DNA/RNA molecules using either physical or biochemical means such as nebulization, sonication or random enzymatic digestion. The result of sequencing one short molecule is called a read. In this work we will use data from the SOLiD platform, where fragmentation and size
Lognormality and oscillations in RNA-seq towards gene ends

selection are performed targeting fragments significantly longer than the read length\(^4\) [3, 4]. Consequently, only the first nucleotides of each fragment (5'-end) are read by the machine, leading to a systematic truncation of the fragment ends (3'-ends). This effect has often been ignored based on the assumption that the bias averages out for a large number of fragments coming from multiple identical molecules. The inspiration for this work is the observation that these effects in fact do not average out, but on the contrary lead to observable oscillations in the read coverage signal close to 3'-ends of RNA transcripts.

Perhaps surprisingly we were able to describe these effects, on average, by Kolmogorov’s 1941 broken stick model [5], which was, according to authoritative sources [6], one of the inspirations for Kolmogorov’s 1962 refined theory of turbulence. The model can be used to improve predictions of ends of transcripts (3'-ends), albeit we can only show a marginal improvement on a gene-per-gene basis using current data. The reasons for this, which arise from other problems or ‘artefacts’ in the sequencing, are discussed below.

The paper is structured as follows: in section 2 we recall briefly the broken stick model and show how oscillations near 3'-ends can be generated by such a process. In section 3 we discuss our data, and in section 4 we discuss how the theory combined with such data can be used to predict gene 3'-ends. In section 5 we show that the predicted oscillations are well displayed on average but that the model only marginally improves the prediction of single gene 3'-ends, for reasons unrelated to the present work, which we discuss. In section 6 we summarize and discuss our results.

2. Reads and coverage

The background to the theory is that biological RNA molecules very often are too long to be directly handled by the sequencing platform. For instance, a typical bacterial gene may have transcripts of length 1000 nucleotides or longer, while an eukaryotic transcript can be many times that length. Biologists have many ways to divide up a long RNA (or DNA) molecule into shorter fragments, but one of the main ones is sonication, or shaking by ultrasound until the molecule breaks. An unbiased RNA fragmentation can then be seen as a sequential random process that, for \(N\) fragments at a step \(t\) creates \(N+1\) fragments at step \(t+1\) by selecting an existing fragment independently of its length and breaking it in two not necessarily equal pieces. Such a process is a particular case of Kolmogorov’s broken stick model, and leads to fragments with lengths following a log-normal distribution [5, 7] with probability distribution function given by

\[
P(L) = \frac{1}{L\sqrt{2\pi}s^2} e^{-\left(ln(L/m)^2/2s^2\right)},
\]

where \(L\) is the fragment length and the parameters \(m\) and \(s\) are given by

\[
s = \sqrt{\ln\left(\frac{\sigma^4}{L_0^4} - 1\right)}, \quad m = \ln\left(\frac{L_0^2}{\sqrt{\sigma^4 + L_0^4}}\right),
\]

where \(L_0\) is the average and \(\sigma^2\) the variance of the fragment length distribution.

---

\(^4\) Applied Biosystems. Fragment Library Preparation 5500 Series SOLiD™ Systems. Life Technologies, Publication Part Number 4460960 Rev. A.
Lognormality and oscillations in RNA-seq towards gene ends

The effect we are interested in arises from each of these fragments being read from one end (the 5’-end). In particular, the very last fragment, which extends all the way to the gene end (3’-end), is read from its other end (its 5’-end). Therefore, unless it is shorter than the number of nucleotides read by the platform, the last fragment will typically not be read all the way to the end of the gene. This means that the coverage signal, the number of reads which can be obtained from a given genomic position, will tend to diminish towards the end of the gene, and at the very end the signal will be low. This leads to considerable error if one tries to use the signal to locate the 3’-end precisely.

However, a more subtle effect also gives rise to oscillations in the coverage signal. Consider the situation where there are preferred lengths \( l'' \) and \( l' \) of the last two fragments. If so, the sequencing platform will deliver more than average number of reads that align to genomic positions respectively \( l' \) and \( l' + l'' \) nucleotides before the gene end, while, in addition to the depletion at the 3’-end, there will also be less than average number of reads aligning to positions between \( l' \) and \( l' + l'' \). In other words, the coverage signal will display oscillations towards the gene end.

In reality, one would observe a distribution of values of \( l'' \) and \( l' \), and the oscillations will only show up on average. We now suppose that these distributions can be modelled using the broken stick model. We set up a Monte Carlo scheme that creates fragments with a distribution given by (1) from the last \( M \) nucleotides of numerous identical copies of an RNA molecule and build \( c_k(L_0, \sigma) \), the coverage \( k \) nucleotides upstream of the transcript 3’-end that one expects to obtain provided that only the first \( N \) base pairs of each fragment are sequenced, i.e.

\[
c_k(L_0, \sigma) = A \sum_{i \in I_{L_0,\sigma}} 1_N(X_i - k), \quad k = 1, 2, \ldots, M, \tag{3}
\]

where \( X_i \) corresponds to the distance from the first nucleotide of a fragment \( i \) to the 3’-end on the original RNA molecule, \( A \) is an arbitrary constant and \( 1_N(x) \) is an indicator function equal to one on \([0, N]\) and zero elsewhere. Each fragment \( i \) is taken from a set \( I_{L_0,\sigma} \) obtained by generating fragments with lengths distributed according to (1) and excluding those shorter than the read length \( N \).

This Monte Carlo scheme gives us the coverage signal ‘as it should be’. Provided that the parameters \( L_0 \) and \( \sigma \) are known for a given experiment and that the existence of a 3’-end is known within a genomic region \( D \) of reasonable size, its precise location \( z \) can then be obtained by maximizing the overlap between the predicted pattern and the experimental coverage, i.e. by solving

\[
\arg \max_{z \in D} \sum_{k=0}^{M} c_k(L_0, \sigma) \tilde{c}_{k,z}, \tag{4}
\]

where \( \tilde{c}_{k,z} \) denotes the coverage obtained from the experiments \( k \) nucleotides upstream of a genomic position \( z \).

3. High-throughput RNA sequencing of the human pathogen \textit{E. faecalis}

The data which we used were obtained as part of an ongoing effort to understand genes, their regulation and generally new biology in the human pathogen \textit{Enterococcus faecalis} [8]. This bacterium, which is commonly found in the lower parts of the human digestive
apparatus, is also an opportunistic pathogen and ranks approximately fourth or fifth as a cause of nosocomial infections, or ‘hospital disease’, world-wide [9]. It is especially medically important as it easily acquires antibiotic resistance, potentially leading to severe disease outcomes in a hospital environment. It is also interesting from a general scientific point of view as an example of an organism which has been known for a long time—and is probably much more important to human health than e.g. *Escherichia coli*—but which is still comparatively much less well known.

We proceed to outline what is meant by the last phrase in the preceding paragraph by following what a bioinformatician (or a physicist using bioinformatic tools) would do. First, we would like to know which are the genes in the organism and what their function could be. If the organism is sequenced, i.e. if the genome is available and deposited in a public database such as GenBank, there are standard tools which will take the genome as input and will give a (hypothetical) gene list as output.

The reference variant of *E. faecalis* is the strain v583, which is resistant to the antibiotic vancomycin, and was isolated in Barnes Hospital, St Louis, MO, USA, in 1987 [10]. Standard tools to annotate bacterial genomes such as Glimmer [11] will give a gene list where a putative function of each putative gene is given by analogy with similar functions of genes coding for a similar protein in other bacteria, in practice mainly in *E. coli*. Glimmer or other tools give in fact only what are known as open reading frames (ORFs), which are DNA sequences that may code for a protein, but they do not give the full gene from the start (the 5′-end) to end (the 3′-end). It has long been recognized that the residues in the flanking regions of DNA at the 5′ and 3′ ends, which do not code for amino acids, are crucial for regulation of transcription [12]. Unfortunately, determining 5′-ends of regulatory elements automatically from a sequence is more difficult than determining an ORF and cannot be done reliably. Hence the interest in deducing such genomic elements by other means.

We analysed a total of six datasets from RNA sequencing (RNA-seq) experiments performed on *E. faecalis* strain v583. Four of these originate from a single RNA extraction sequenced on the SOLiD v3 platform using the plain single-end strand-specific RNA sequencing protocol from the platform manufacturer. In experiments labelled here and below S3sr1A and S3sr1B all RNA was sequenced, while in S3sr0A and S3sr0B the ribosomal RNA (rRNA) was removed beforehand using an Ambion MICROBExpress bacterial mRNA enrichment kit. We note that in bacteria rRNA may constitute as much as 90% of total RNA in an extract, and that removing rRNA before analysis is a common practice. From the point of view of the present work, this difference is however not important, and the experiments with or without rRNA removal can therefore be considered as replicates. The different suffixes A or B correspond to independent RNA-seq performed on two samples from the same RNA extract separated after fragmentation, i.e. they are technical replicates of the experiments.

The two further datasets labelled S55sr1 and S55rr1 correspond to single-end strand-specific RNA-seq on the SOLiD 5500 platform from two RNA extractions obtained from bacteria grown in different growth conditions. Ahead of the manufacturer’s protocol, short RNA oligos were ligated to the 5′-ends as described in [8].

An important aspect of the preparation protocols used in all the experiments introduced above is that RNA (or DNA) molecules are fragmented and size selected targeting a length greater than 100 nt [3, 4] (see footnote 4). These fragments are then
amplified and attached to beads in an emulsion polymerase chain reaction (emulsion PCR) before being sequenced. The sequencing itself is performed on the 50 first nucleotides—or more precisely on the first 50 nucleotide transitions, as obtained with SOLiD 2-base colour encoding [13]—of each fragment, leaving the remaining part of the molecule unread.

Every one of the above introduced datasets contains 60–100 million reads, each with a length corresponding to 50 colour encoded nucleotide transitions, each position of each read being accompanied by a quality score indicating how well this particular position was read by the sequencer.

Once you have a read you have to find where on the genome it came from. This procedure is called alignment and, while it could often be done directly with high accuracy, it is better practice to use a pre-existing bioinformatic tool which avoids known pitfalls and which will indicate how likely the information given is to be correct. We aligned our data using Bowtie (version 0.12.7) [14], an open source alignment software widely used in RNA-seq data analysis. In the configuration we used, the software reports an alignment if the read matches the genome with fewer than two errors in the first 28 nucleotides of the read and if the sum of the read qualities at all mismatching positions is lower than a threshold. If multiple positions satisfy those conditions, all of them are reported as valid alignments. After alignment, we calculate the coverage and take multiply mapped reads into account by dividing their contribution to the read count by the number of matches. Due to the specificities of the aligner, reads are converted to 48 nt long fragments during the alignment, thus reducing the practical read length to this number.

For every dataset, we were able to align between 40 and 60% of reads to the genome, which corresponds to an average coverage over \( 450 \times \) of the 3.2 million nucleotides present in the \( E. \ faecalis \) v583 genome.

In order to confirm that RNA fragmentation results in the postulated log-normal distribution, we measured the fragment length distribution for another RNA sample prepared for future RNA-seq using an Agilent 2100 Bioanalyzer (figure 1). The match with a log-normal distribution is clear except for a peak corresponding to fragments of 13 nt, which corresponds to synthetic RNAs artificially added to this particular sample. Unfortunately, no measurements of this type are available for the samples used to generate our six datasets.

### 4. Transcript 3′-ends

Bacteria have two major mechanisms for transcription termination, the rho-dependent terminator and the rho-independent, or intrinsic, terminator. In \( E. \ coli \), each of the mechanisms accounts for the termination of about half of the genes, the rho-dependent termination playing a major role in transcription regulation [15].

In the rho-independent termination, due to nucleotide-pairing, the transcribed RNA folds into a stem–loop structure of 7–20 nt in length, rich in cytosine–guanine nucleotide pairs and followed by a poly-uracil (poly-U) tail. This stem–loop binds to the RNA-polymerase complex, which catalyses the synthesis of RNA from the DNA template, causing a pause in the transcription while the weak bonds between the poly-U tail and the adenines on the DNA allow for the complex to detach from the DNA strand [16]. Such terminators can be quickly and reliably predicted using TranstermHP [17], a now standard computational tool that searches whole genomes for occurrences of sequences that could...
Lognormality and oscillations in RNA-seq towards gene ends

Figure 1. Length distribution of an RNA sample measured with an Agilent 2100 Bioanalyzer after fragmentation and library preparation adding adapters with total constant length of 128 nt. The dashed blue line shows a log-normal distribution with mean 65 and standard deviation 60 shifted by 128 nt to the right. The peak observed in the measurement at 141 nt corresponds to an excess of synthetic short RNA sequences artificially added to the particular sample used.

lead to the terminators described above by systematically inspecting the folded structure of RNA transcripts ahead of all adenine-rich regions of the genome that could be coding for suitable poly-U tails. The software takes as input a whole genome and outputs a list of potential rho-independent terminators described by the beginning and end of the stem–loop as well as a confidence level corresponding to the probability that the candidate actually is a functional terminator.

In the rho-dependent termination, the mechanism is mediated by a protein called rho factor. Rho binds the freshly transcribed RNA in weakly folded regions at specific locations called rho utilization (rut) sites. Such sites are about 70 nt in length and cytosine rich, but no consensus sequence for a rut is known. After binding, rho moves downstream along the RNA strand through an energy consuming mechanism. When approaching the region where RNA synthesis takes place, rho unwinds the nascent RNA from the DNA strand and eventually terminates the transcription by causing the RNA polymerase to detach in a mechanism that is so far not well known [18]. The transcription termination typically takes place between 10 and 100 nt downstream of the rut site, its exact location depending on the relative speed of translocation of rho on the RNA strand and that of RNA synthesis of the RNA polymerase. Due to the complexity of this mechanism, there is no way described in the literature to reliably predict rho-dependent termination sites.

We predicted rho-independent terminators on the chromosome of E. faecalis v583 (NCBI reference AE016830) using TranstermHP (v2.08, with default options). Out of a total of 1851 predictions, only the 1229 terminators found with 100% confidence were retained.

Upstream of each selected terminator, we collected the read coverage obtained from RNA-seq on a window of 400 nt starting from the last nucleotide of the terminator stem–loop as predicted by TranstermHP. It is worth noting that the real 3’-end of the corresponding transcript is located downstream of this location, usually after a 4–9 nt long poly-uracil (poly-U) tail [15, 19]. The signal was normalized so that its integral over
the region of interest is unity. Averaging over all the 1229 terminators, for the case of the S3sr1B dataset, led to the signal presented in figure 2.

It can be easily seen that the signal falls to zero well before the location of the terminator and has a pronounced peak around position $-80$ nt. Upstream of the peak, one can notice two to three periods of a weak oscillation with a period of the order of 100 nt, with no plausible biological origin.

5. Results

We use (3) from the previously described Monte Carlo scheme with a read length $N = 48$ nt to generate the coverage expected from our theory. In every case, the scheme creates random fragments from $10^5$ copies of the original complete molecule, which appears to be sufficient for a relative accuracy better than $10^{-2}$. Additionally, we introduce a parameter $d$ that corresponds to a positional shift representing the average distance between the last nucleotide of the stem–loop in the rho-independent terminator as predicted by TranstermHP and the real 3′-end of the corresponding molecule, i.e. the length of the poly-U tail following the stem–loop. We fit $d$, the mean $L_0$ and the standard deviation $\sigma$ of the underlying log-normal distribution. We output the coverage at each position over 350 nt ahead of the last nucleotide of the stem–loop. We measure the error $\varepsilon$ between the theoretical coverage $c$, obtained using the fitted parameters, and $\tilde{c}$, the one obtained from the data, as

$$
\varepsilon = \frac{\sum_{i=1}^{350} (c_{i+d} - \tilde{c}_i)^2}{\sum_{i=1}^{350} \tilde{c}_i^2}.
$$

Fitting the two datasets from the S3sr0 type (figures 3(a) and (b)) results in an average fragment length of 110 nt and an estimated poly-U tail length $d$ of 6 nt, which is in good agreement with what is commonly expected for this tail [15, 19]. On the other hand, the
Figure 3. Comparison between the signals obtained as described in section Transcripts 3’-ends from the 6 RNA-seq experiments and those predicted by the theory after fitting the parameters $L_0$, $\sigma$ and $d$. (a) S3sr0A, (b) S3sr0B, (c) S3sr1A, (d) S3sr1B, (e) S55rr1 and (f) S55sr1.

datasets of the S3sr1 series (figures 3(c) and (d)) result in a shorter estimated average fragment length and a poly-U tail length of 17 nt, which seems unrealistically high. As there is no reason for the two pairs of experiments to differ in the length of the poly-U tail (the two samples were prepared using the same protocol from the same RNA extraction and differ only by the removal of rRNA), we attribute this difference to an issue with data fitting and note that underestimating $L_0$ causes an overestimate of $d$.

Furthermore, we note that the results for the ‘A’ and ‘B’ samples of S3sr0 and S3sr1 are very similar to each other in every aspect. Since the members of each pair were separated after the RNA fragmentation step, this indicates a good reproducibility of the
Figure 4. Coverage signal from the dataset S3sr0A between genes EF 0331 and EF 0332 where the transcription of EF 0332 seems to start within the terminator of the previous gene. The end of the ORF of EF 0331 and the beginning of that of EF 0332 are indicated by the green horizontal lines and the green arrows indicate the transcription direction. The stem–loop of the predicted rho-independent terminator, located between position 307 170 and 307 203 on the chromosome, is marked in blue. The horizontal axis shows the relative distance to the last nucleotide of the stem–loop of the predicted terminator and the numbers inserted in the plot indicate the position on the E. faecalis v583 chromosome.

sample preparation and sequencing, and implies that the observed effects depend only on the steps preceding and including the fragmentation.

The further datasets of the S55 series lead to much worse fits (figures 3(e) and (f)), where only the rightmost part of the signal corresponding to the fall before the terminator can be reproduced. Overall, the signal obtained from the data seems much more noisy and the pattern is less pronounced than in the examples previously described. This effect is probably related to the different protocol used in those experiments: the inclusion of tags at the beginning of RNA molecules modifies their lengths in a non-controlled way, causing a blurring of the pattern. Nevertheless, the parameters obtained from the fit are realistic and in line with those of the other four datasets.

Finally we note that in all cases the experimental signal increases close to the transcript end, starting 5–10 nt ahead of the last nucleotide of the stem–loop, which is at odds with the proposed theory. We observe that in a few cases the next gene downstream on the DNA strand seems to have its transcription start site within the genomic region corresponding to the rho-independent termination (example on figure 4), which is enough to explain the effect observed in the average signal. Whether such transcription start sites are biologically relevant or result from sequencing artefacts in our experiments is unclear.

We now focus on the dataset S3sr1B that gives the lowest relative error and use our theory to predict individual 3′-ends near the 1229 predicted rho-independent terminators. We use as initial guess the location of the last nucleotide of the stem–loop and search within a region of 100 nt centred around this guess. We compare on figure 5 and table 1 the predictions obtained by optimizing (4) to a na¨ıve approach that detects the locus closest to the initial guess where the coverage goes to zero. We notice that the number of predictions within a range of 100 nt around the initial guess is similar for both methods, while the number of predictions falling between 4 and 9 nt downstream of the last nucleotide of the
Figure 5. Distribution of distances between the last nucleotide in the stem–loop of the rho-independent terminator and predictions of 3′-end of the corresponding transcript obtained either by a naïve approach searching for the closest locus where the signal goes to zero (top) or using the maximum pattern overlap method from (4) (bottom) for the 1229 terminators predicted by TramertransermHP with 100% confidence. The vertical dashed lines delimit the region between 4 and 9 nt downstream of the terminator stem–loop where the actual 3′-end is expected.

Table 1. Summary of the data presented in figure 5 showing for both methods the number of predictions within certain ranges and the mean and standard deviation of the absolute distance between a prediction and the expected location of the 3′-end, assumed to be 7 nt downstream of the last nucleotide in the stem–loop of the rho-independent terminator, as obtained from the fit in figure 3(b).

|                  | Nb. in [−50, 50] | Nb. in [4, 9] | Mean   | Std dev. |
|------------------|------------------|--------------|--------|----------|
| Naïve approach   | 745              | 32           | 24.83  | 13.41    |
| Pattern overlap  | 744              | 67           | 22.80  | 16.08    |

terminator stem–loop (the location where the real 3′-end of the transcript is expected) is more than twice as high for the maximum pattern overlap method. We also note that the average absolute distance to the expected 3′-end is slightly improved by using our theory, but its standard deviation is worse. Finally, we point out that the naïve approach predicts many 3′-ends far downstream of the terminator (+30 nt and further) which are not biologically relevant. This effect is not present while using predictions based on our theory.

The reason for having only a relatively modest improvement is that, while we have shown that the proposed theory works well for the average signal ahead of gene ends, the pattern is not directly observed when considering the read coverage of one individual
Figure 6. Examples of the read coverage ahead of predicted terminators showing patterns very different from the one predicted by the theory. (a) Region ahead of terminator at 739,111–739,129. (b) Region ahead of terminator at 273,009–273,026. (c) Region ahead of terminator at 595,281–595,333.

transcript. Often, the signal shows one or several box-shaped regions of high signal, with a position and periodicity only very roughly in agreement with the theory (figures 6(a) and (b)). In other situations, the pattern may be completely absent (figure 6(c), in particular the blue curves). These effects are probably related to sequence-dependent biases in the fragmentation process that can induce preferential cutting sites in multiple copies of the same transcript, thus invalidating the assumptions of unbiased random fragmentation of our theory. While considering many transcripts with different sequences, such biases average out, leading to the previously observed agreement on average. The second observed effect may be related to RNA degradation from the 3′-end in the cell: at the moment of RNA extraction, multiple copies of the same transcript that have been engaged in degradation are likely to have different lengths and thus their 3′-ends at different but neighbouring genomic locations. Such a mechanism will blur the pattern described above and, by shortening RNA molecules, will shift the predictions upstream. Due to those shortcomings, a more detailed model taking fragmentation biases and potential RNA degradation into account will be necessary to provide reliable predictions of the location of the 3′-end of a transcript.

A direct consequence of the mechanisms described in this work is that SOLiD single-end strand-specific RNA-seq cannot be used for accurate determination of 3′-ends.
Lognormality and oscillations in RNA-seq towards gene ends

Figure 7. Visualization of the read coverage from the S3sr0A dataset in a region containing an ncRNA detected using a microarray [20]. The region of the ncRNA is marked with vertical bars. While the two methods are in perfect agreement for the 5'-end, a naïve interpretation of the coverage from RNA-seq would place the 3'-end of the transcript at the position indicated by the dashed line, well before the end of the transcript as measured using a microarray.

of transcripts. The potential truncation of transcripts ends can cause problems while comparing results from RNA-seq to other measurement methods such as northern blot, 3’ RACE or microarray (example in figure 7). Furthermore, RNA transcripts longer than the read length (here 50 nt) but significantly shorter than the targeted average length (here 100–120 nt) may simply be totally absent from the final results while one would naïvely expect them to be well visible.

6. Conclusions

We demonstrated that a simple theory based on Kolmogorov’s broken stick model explains well the type of signal observed on average near transcript 3’-ends in high-throughput sequencing experiments.

The use of this theory for prediction of the 3’-end of individual transcripts has shown only minor improvements compared to a naïve approach. The main reason for this is the presence of clear preferential cleavage sites on transcripts that break the assumption of unbiased fragmentation, essential in our theory.

Most importantly, we have demonstrated that drawing conclusions about transcript 3’-ends from single stranded RNA-seq experiments on SOLiD must be done with great care. A good understanding of all the mechanisms in the whole RNA-seq pipeline is necessary to give a correct interpretation of experimental results.

Acknowledgments

Many thanks to Sean P Kennedy (MetaQuant platform, INRA UMR1319 Micalis) for sequencing the samples of the S55 series as well as his valuable clarifications on the SOLiD protocols. We also thank Aymeric Fouquier d’Hérouël (Institute for Systems Biology) and Francis Repoila (INRA UMR1319 Micalis) for the preparation of RNA samples as well as Ingemar Ernberg (Karolinska Institutet) for his generous hospitality by providing

doi:10.1088/1742-5468/2013/10/P10013
laboratory space and equipment for Aymeric Fouquier d’Hérouël. This work was supported by the Academy of Finland as part of its Finland Distinguished Professor program, project 129024/Aurell, and by the Academy of Finland Centre of Excellence COIN.

References

[1] Metzker M L, Sequencing technologies—the next generation, 2010 Nature Rev. Genet. 11 31
[2] Knierim E, Lucke B, Schwarz J M, Schuelke M and Seelow D, Systematic comparison of three methods for fragmentation of long-range pcr products for next generation sequencing, 2011 PLoS ONE 6 e28240
[3] Hedges D J et al, Comparison of three targeted enrichment strategies on the solid sequencing platform, 2011 PLoS ONE 6 e18595
[4] Kennedy S P, 2012 private communication
[5] Kolmogorov A N, O logarifmicheski-normal’nom zakone raspredelenija razmerov chastici pri droblenii (On the logarithmic normal distribution of particle sizes under grinding), 1941 Dokl. Akad. Nauk SSSR 31 99
[6] Frisch U, 1996 Turbulence: The Legacy of A N Kolmogorov (Cambridge: Cambridge University Press)
[7] Siegel A F and Sugihara G, Moments of particle size distributions under sequential breakage with applications to species abundance, 1983 J. Appl. Probab. 20 158
[8] d’Hérouel A F, Wessner F, Halpern D, Joseph L-V, Kennedy S P, Serror P, Aurell E and Repoila F, A simple and efficient method to search for selected primary transcripts: non-coding and antisense RNAs in the human pathogen enterococcus faecalis, 2011 Nucleic Acids Res. 39 e46
[9] Palmer K L and Gilmore M S, Multidrug-resistant enterococci lack CRISPR-cas, 2010 mBio 1 e00227
[10] Sahm D F, Kissinger J, Gilmore M S, Murray P R, Mulder R, Solliiday J and Clarke B, In vitro susceptibility studies of vancomycin-resistant enterococcus faecalis, 1989 Antimicrob. Agents Chemother. 33 1588
[11] Delcher A L, Harmon D, Kasif S, White O and Salzberg S L, Improved microbial gene identification with glimmer, 1999 Nucleic Acids Res. 27 4636
[12] Ptashne M, 2004 A Genetic Switch: Phage Lambda Revisited 3rd edn (New York: Cold Spring Harbor Laboratory Press)
[13] Breu H, A theoretical understanding of 2 base color codes and its application to annotation, error detection, and error correction, 2010 Report Applied Biosystems (July)
[14] Langmead B, Trapnell C, Pop M and Salzberg S, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, 2009 Genome Biol. 10 R25
[15] Peters J M, Vangeloff A D and Landick R, Bacterial transcription terminators: the RNA 3′-end chronicles, 2011 J. Mol. Biol. 412 793
[16] Wilson K S and von Hippel P H, Transcription termination at intrinsic terminators: the role of the RNA hairpin, 1995 Proc. Nat. Acad. Sci. 92 8793
[17] Kingsford C, Ayanbule K and Salzberg S, Rapid, accurate, computational discovery of rho-independent transcription terminators illuminates their relationship to DNA uptake, 2007 Genome Biol. 8 R22
[18] Boudvillain M, Nara F-B and Lionello B, Terminator still moving forward: expanding roles for rho factor, 2013 Curr. Opin. Microbiol. 16 118
[19] d’Aubenton C Y, Brody E and Thermes C, Prediction of rho-independent escherichia coli transcription terminators: a statistical analysis of their RNA stem–loop structures, 1990 J. Mol. Biol. 216 835
[20] Shioya K, Michaux C, Kuenne C, Hain T, Verneuil N, Budin-Verneuil A, Hartsch T, Hartke A and Giard J-C, Genome-wide identification of small RNAs in the opportunistic pathogen enterococcus faecalis v583, 2011 PLoS One 6 e23948

doi:10.1088/1742-5468/2013/10/P10013 14