Nucleic Acids Research Advance Access published April 14, 2010

Nucleic Acids Research, 2010, 1–7
doi:10.1093/nar/gkq224

Biases in Illumina transcriptome sequencing caused by random hexamer priming

Kasper D. Hansen[1,*], Steven E. Brenner[2] and Sandrine Dudoit[1,3]

[1]Division of Biostatistics, School of Public Health, UC Berkeley, 101 Haviland Hall, Berkeley, CA 94720-7358, [2]Department of Plant and Microbial Biology, UC Berkeley, 461 Koshland Hall, Berkeley, CA 94720-3102 and [3]Department of Statistics, UC Berkeley, 367 Evans Hall, Berkeley, CA 94720-3860, USA

Received December 1, 2009; Revised March 16, 2010; Accepted March 17, 2010

Downloaded from http://nar.oxfordjournals.org at Univ of California Library on April 16, 2010

ABSTRACT

Generation of cDNA using random hexamer priming induces biases in the nucleotide composition at the beginning of transcriptome sequencing reads from the Illumina Genome Analyzer. The bias is independent of organism and laboratory and impacts the uniformity of the reads along the transcriptome. We provide a read count reweighting scheme, based on the nucleotide frequencies of the reads, that mitigates the impact of the bias.

INTRODUCTION

Transcriptome analysis using next-generation sequencing technologies, or RNA-Seq, is a promising new form of analysis already yielding breakthrough discoveries. It is well-known that spatial biases along the genome exist, resulting in a non-uniform coverage of expressed transcripts (1). These spatial biases hinder comparisons between genomic regions and will therefore adversely affect any analysis where such a comparison is integral. Two important such types of analyses, on which RNA-Seq is expected to have great impact, are the study of alternative splicing, where an alternatively spliced exon is compared with a constitutively spliced exon, and *de novo* inference of gene structure.

General biases in DNA sequencing using the Illumina platform have been studied in (2). Dohm *et al*. found that there was no fragmentation bias in DNA sequencing reads, but uneven coverage as a result of locally varying GC content.

The standard library preparation protocol (1) for transcriptome sequencing using the Illumina Genome Analyzer platform starts with extraction of total RNA, followed by poly(A) enrichment using oligo(dT) beads, RNA fragmentation and reverse transcription into double-stranded complementary DNA (dscDNA) primed by random hexamers. The resulting dscDNA is then sequenced using the DNA sequencing protocol, starting with end repair, addition of an A nucleotide and adapter ligation. Random hexamer priming is utilized to generate reads across the entire length of all expressed transcripts, although the resulting sequence coverage is far from uniform (1). Fragmenting the RNA prior to reverse transcription has been shown to make coverage more uniform within a transcript (1).

Here, we demonstrate that the use of random hexamer priming results in a bias in the nucleotide composition at the start of sequencing reads and that this bias influences the uniformity of the location of the reads along expressed transcripts. We also propose a reweighting scheme that adjusts for the bias and makes the distribution of sequencing reads more uniform.

MATERIALS AND METHODS

Data

Data were obtained according to information in the references. Additional data have been deposited to the Short Read Archive under accession number SRA009901.

The Supplementary Data section contains precise details on the number of (mapped, unmapped) reads (Supplementary Tables S1 and S2), data files and their origin (Supplementary Table S3) and an overview of experimental protocols (Supplementary Table S4). One lane worth of data was used for each sample depicted in Figures 1–3. One lane was selected at random from each experiment, except for Nagalakshmi where we chose a lane each from a randomly primed sample (RH) and an oligo(dT) primed sample (DT) and Wang where we chose a lane each from heart and brain tissue.

Mapping

Reads were mapped using Bowtie (3), requiring perfect and unique matching to the genome. See Supplementary Data for additional information.

Free energies

Binding energies were computed using the program 'DNA', from the Mobyle webserver hosted at the

*To whom correspondence should be addressed. Tel: +1 510 642 3241; Fax: +1 510 643 5163; Email: khansen@stat.berkeley.edu

© The Author(s) 2010. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Pasteur Institute, which provides access to the EMBOSS (4) suite of programs. Free energies from RNA–DNA duplexes were used.

Reweighting scheme

Each read is assigned a weight $w(h)$ based on its first heptamer h (seven bases) and computed as follows. Given a set of mapped reads, let $\hat{p}_{hep:i}$ be the observed distribution of heptamers starting at position i (hence spanning positions i to $i+6$) of the reads. Specifically, $\hat{p}_{hep:i}(h)$ is the proportion of reads which have heptamer h at position i and $\hat{p}_{hep:1}(h)$ is the proportion of reads starting with heptamer h. Define the weights w by

$$w(h) = \frac{\frac{1}{6}\sum_{i=24}^{29} \hat{p}_{hep:i}(h)}{\frac{1}{2}(\hat{p}_{hep:1}(h)+\hat{p}_{hep:2}(h))} \quad (1)$$

(Here, we assume a read length of at least 35 nt.)

At each (stranded) genomic location l, we compute the number of reads $c(l)$ with 5′-end mapping to l. Each (stranded) genomic position is identified with a single heptamer $h(l)$ and the reweighted counts are computed as $c_w(l) = c(l)w(h(l))$. A specific example is provided in Table 1.

The method is implemented in the R package Genominator (version 1.1.5), released through the Bioconductor Project (5).

Evaluation of the reweighting scheme

Regions of constant expression (ROCE) were defined based on existing annotation as maximal genomic regions for which each position is annotated as belonging to the same set of transcripts (for example, two overlapping transcripts are split into three ROCEs). Highly expressed ROCEs (heROCEs) were defined as ROCEs longer than 100 (*Saccharomyces cerevisiae*) or 50 (*Homo sapiens*) mappable bases, with more than one mapped read per mappable base. For the experiments considered here, roughly 10% of all ROCEs were highly expressed (Supplementary Table S2).

For each heROCE, we computed χ^2 goodness-of-fit statistics to the Poisson distribution with a constant intensity and coefficients of variation for the unadjusted $c(l)$ and the reweighted $c_w(l)$ counts (Figure 4 and Supplementary Figures S6–S8). Either of these two measures being lower for the reweighted counts than for the unadjusted counts was taken as evidence that reweighting improved the uniformity of the location of the reads within heROCEs. For five different data sets (four from *S. cerevisiae* and one from *H. sapiens*), we observed a consistent decrease in both measures for almost all heROCEs.

RESULTS

Positional nucleotide bias

To explore potential biases in the sequencing system, we analyzed data from a number of recently published and unpublished RNA-Seq experiments conducted using the Illumina Genome Analyzer (1,6–9). For each experiment, we determined a set of stringently mapped reads and computed the nucleotide frequencies for each position of the reads (see 'Materials and Methods' section).

There is a strong distinctive pattern in the nucleotide frequencies of the first 13 positions at the 5′-end of mapped RNA-Seq reads (Figure 1a). The nucleotide frequencies vary from position to position, but are almost the same across different experiments. It is striking that while the exact nucleotide frequencies differ slightly between experiments, their relative changes are nearly identical. After the first 13 positions, the nucleotide frequencies become independent of position, but are different between experiments, presumably reflecting the base content of the different transcriptomes. The pattern is thus reproducible across experiments in different organisms and laboratories. A similar pattern is also present when all unmapped (versus. only mapped) reads are considered (Supplementary Figure S1).

It is not only the frequencies of single nucleotides at the 5′-end of reads that are very similar across experiments, but also the frequencies of longer runs of nucleotides, such as hexamers. As an example, in Figure 2, we compare the distributions of hexamers corresponding to read positions 1–6 as well as positions 25 to 30 for *S. cerevisiae* and *H. sapiens*. Presumably, the sizable correlation ($r = 0.77$) at the beginning of reads primarily reflects a bias in nucleotide content, whereas the limited correlation ($r = 0.35$) at the end reflects some similarity of hexamer composition between the transcriptomes of the two organisms.

This dependence of nucleotide frequency on position can occur either because of a reproducible bias originating from the sequencing platform or because of a spatial bias of the reads across the transcriptome.

In contrast to RNA-Seq, no strong distinctive pattern in nucleotide frequencies is observed for DNA resequencing and ChIP-Seq experiments (10–14) (Figure 1b and Supplementary Figure S2). This shows that the 5′ nucleotide bias from RNA-Seq is not caused by the Illumina Genome Analyzer DNA sequencing protocol, but rather by additional steps in the RNA-Seq library preparation, namely RNA extraction and reverse transcription into dscDNA. The standard protocol described above was used in all experiments in Figure 1a, except for two experiments that omitted RNA fragmentation.

Based on Figure 1c, we conclude that the pattern is caused by the use of random hexamers to prime the reverse transcription of RNA into dscDNA. The figure shows a number of RNA-Seq experiments employing alternative library preparation protocols (15–17). Two of these experiments used oligo(dT) priming followed by fragmentation of dscDNA using nebulization and sonication; both these experiments show no dependence of the nucleotide frequencies on position. The other two experiments employed oligo(dT) priming and random hexamer priming, both followed by fragmentation of dscDNA using DNase I. The nucleotide frequencies for these latter two experiments have similar patterns, but compared with the pattern of the RNA-Seq

Downloaded from http://nar.oxfordjournals.org at Univ of California Library on April 16, 2010

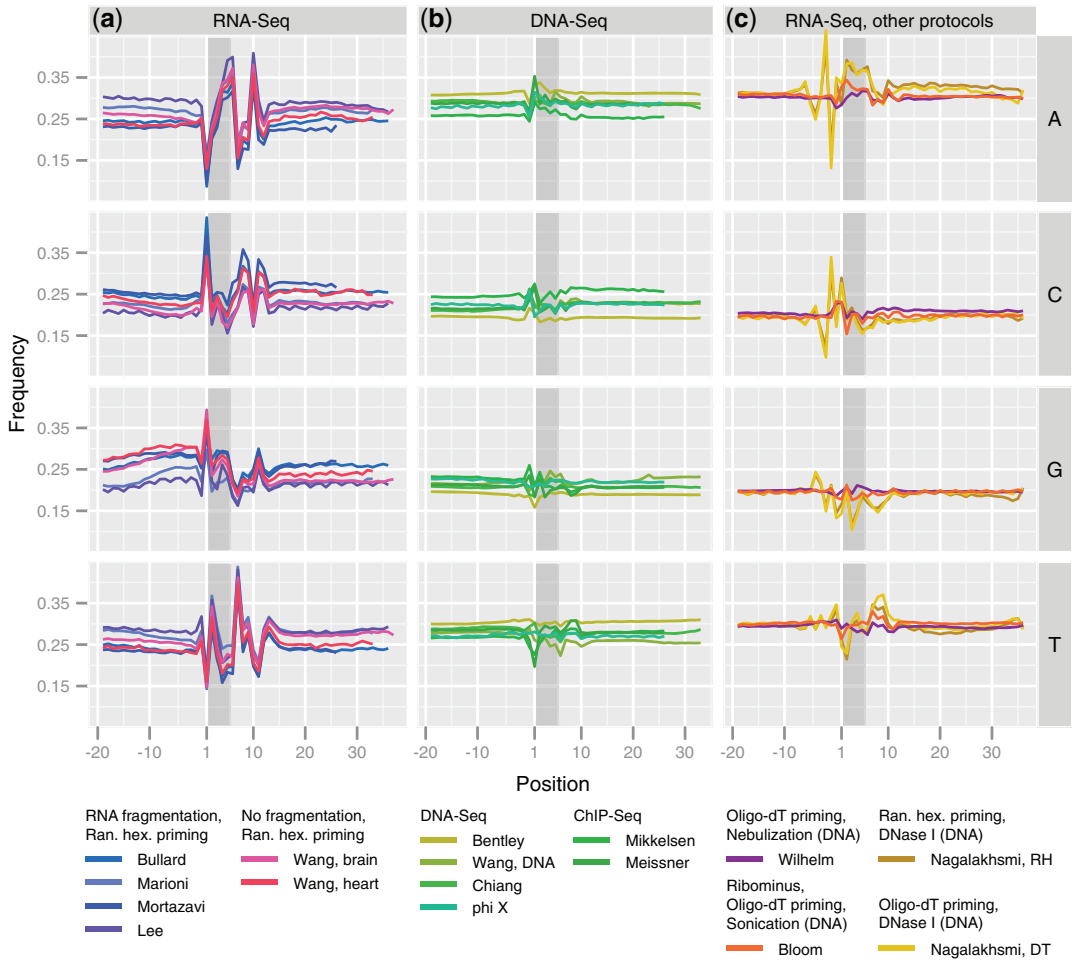

Figure 1. Nucleotide frequencies versus position for stringently mapped reads. For each experiment, mapped reads were extended upstream of the 5′-start position, such that the first position of the actual read is 1 and positions 0 to −20 are obtained from the genome. The first hexamer of the read is shaded. Brief experimental protocols are indicated in the key. **(a)** RNA-Seq experiments conducted using priming with random hexamers, with and without RNA fragmentation. **(b)** DNA resequencing and ChIP-Seq experiments. **(c)** RNA-Seq experiments with alternative library preparation protocols, including priming with random hexamers followed by fragmentation using DNase I and priming with oligo(dT) followed by fragmentation using either DNase I, nebulization or sonication.

experiments of Figure 1a, this pattern is smaller in magnitude and extends upstream of the start position of the reads. Because the two different priming methods used in these experiments result in the same pattern, we conclude that the pattern is caused by DNase I digestion.

We find that computationally predicted binding energies associated with the random hexamers do not explain the observed hexamer frequencies at the beginning of the reads (see Supplementary Data, Supplementary Figure S3). Rather, we find that any relationship between binding energies and hexamer frequencies is a feature of the particular transcriptome and is not related to the use of random hexamers for priming.

In the standard model of second-strand synthesis of dscDNA, the second DNA strand is synthesized by DNA polymerase primed from nicks in the original RNA strand, where the nicks are created by lightly treating the RNA–DNA duplexes with RNase. This model implies that the 5′ bias caused by random hexamer priming will only affect reads from the 5′-end

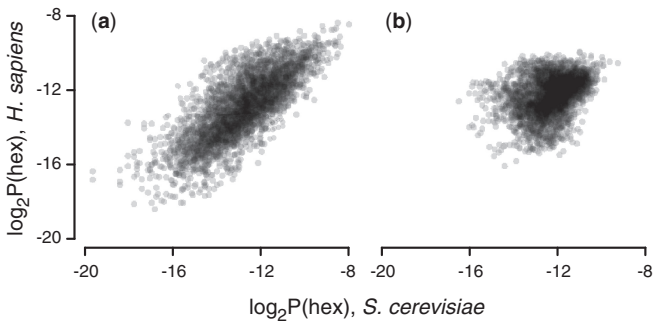

Figure 2. Hexamer frequencies. (**a**) The logarithm (base 2) of all (4096) observed hexamer frequencies computed using positions 1–6 of the aligned reads for an experiment in *H. sapiens* (8) versus an experiment in *S. cerevisiae* (9). The two distributions have a correlation of 0.77. (**b**) As in (a), but the hexamers correspond to positions 25–30 of the aligned reads, with a correlation of 0.35.

Downloaded from http://nar.oxfordjournals.org at Univ of California Library on April 16, 2010

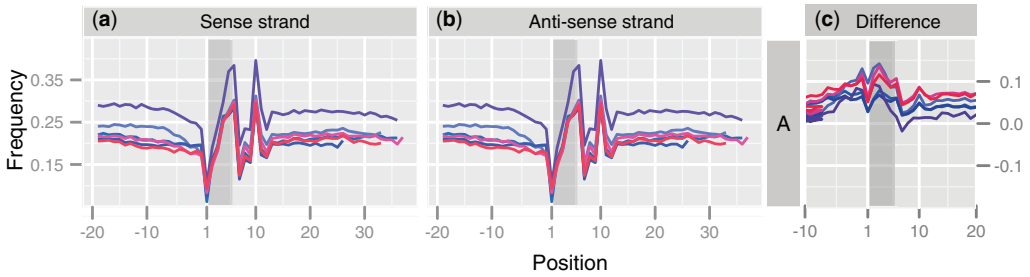

Figure 3. Nucleotide frequencies versus position for stringently mapped stranded reads for the A nucleotide. (**a** and **b**) As in Figure 1a, but split according to whether reads map to the sense or antisense strand. (**c**) Difference between the frequencies in (a and b).

Table 1. Data from a small genomic region in the sense strand of the YOL086C gene in *S. cerevisiae*

Strand	Location	Heptamer	Count	Weight	Reweighted count
	l	$h(l)$	$c(l)$	$w(h(l))$	$c_w(l)$
...					
−1	159792	TTGGTCG	17	1.39	23.6
−1	159793	TTTGGTC	17	0.25	4.3
−1	159794	TTTTGGT	65	0.31	20.4
−1	159795	GTTTTGG	72	0.32	23.3
−1	159796	CGTTTTG	10	1.66	16.6
...					

$c(l)$ denotes the number of mapped reads starting at a particular (stranded) location l and $h(l)$ is the unique heptamer associated with this location. $w(h(l))$ are weights such as in Equation (1) and $c_w(l) = c(l)w(h(l))$ are the location-specific reweighted counts. For this particular small genomic region, reweighting makes the counts more comparable between different locations. Data from the WT experiment.

of the first strand. We have used the annotated coding regions of the genome to infer whether reads are likely to have originated from the sense strand (second-strand synthesis by DNA polymerase) or the antisense strand (first-strand synthesis by reverse transcriptase) (Figure 3a–b and Supplementary Figure S4). A 5′ pattern similar to the one depicted in Figure 1a is visible on both strands, suggesting that the second-strand DNA synthesis is not solely primed by nicked RNA, but also by random hexamers remaining in the solution. There is a small, but consistent difference between the patterns on the sense and antisense strands (Figure 3c and Supplementary Figure S4), perhaps reflecting different sequence specificity of reverse transcriptase and DNA polymerase or the effect of nick priming.

A model for the bias

If the reads were uniformly sampled from the transcriptome, there would be no dependence of nucleotide frequencies on position. Based on this observation, we now posit the following model: priming using random hexamers generates a biased sample of dscDNA fragments, not uniformly distributed across the transcriptome. Specifically, dscDNA fragments beginning with certain 13-mers are over- and underrepresented compared with the frequency of the 13-mers in the transcriptome. Starting with such a biased sample of fragments, the resulting reads produced by the Genome Analyzer reflect the bias in the nucleotide frequencies of these fragments. This interpretation implies that the pattern is not caused by sequencing errors and hence cannot be removed by trimming the 5′-end of the reads. Indeed, we have consistently found that more reads map by trimming their 3′-end than their 5′-end, in agreement with a higher error rate at the 3′-end of the reads. Trimming the 5′-end of the reads would thus remove the pattern only from the part of the reads that is mapped to the genome, but the pattern would be visible in the nucleotides upstream of the mapped trimmed portion of the reads. This model also explains why dscDNA fragmentation substantially alters the 5′ pattern (Figure 1c), as the fragmentation occurs after random priming and generates a new set of fragments with ends affected by DNase I.

It is surprising that the pattern extends well beyond the hexamer primer, out to 13 bases. The length of the pattern could potentially be explained by a strong bias in the first 6 bases of the reads, coupled with dependencies between adjacent nucleotides in the transcriptome. Two observations contradict this explanation. First, the pattern in the nucleotide frequencies ends immediately upstream of the first base of the reads, indicating that the dependence between adjacent nucleotides in the transcriptome is weak (Figure 1a). Note that it is possible for a pattern to extend upstream of the reads, as seen with DNase I fragmentation (Figure 1c). Second, dinucleotide transition probabilities appear biased throughout all 13 initial bases (Supplementary Figure S5). The fact that the 5′ bias extends over 13 bases could be explained by the sequence specificity of the polymerase. Alternately, due to the end repair performed as part of the standard DNA sequencing protocol, the first sequenced base of a read may not be where the primer binds.

Adjusting for the bias

In order to investigate the effect of the priming bias on the distribution of reads along expressed transcripts and adjust for this bias, we developed the following read count re-weighting scheme. Each read is assigned a weight based on its first heptamer, reflecting the fact that certain heptamers are overrepresented in the biased distribution at the start of the reads, relative to the unbiased distribution at the end. Reads beginning with a heptamer overrepresented in the heptamer distribution at

Downloaded from http://nar.oxfordjournals.org at Univ of California Library on April 16, 2010

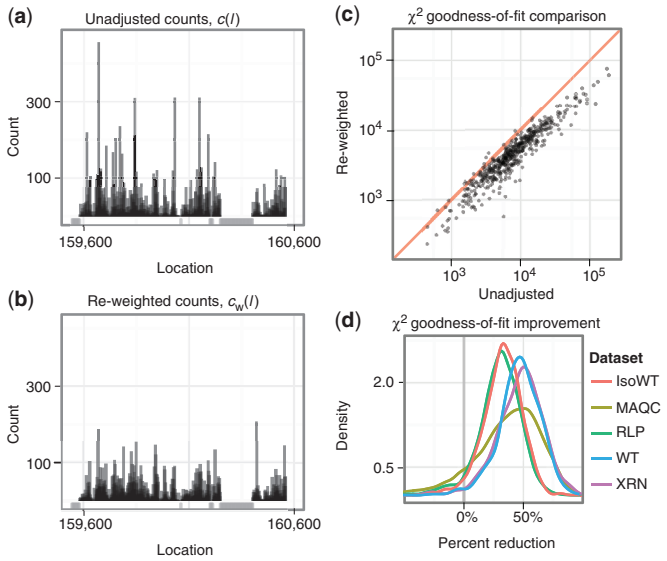

Figure 4. Evaluation of the reweighting scheme. (**a** and **b**) Unadjusted and re-weighted base-level counts for reads from the WT experiment mapped to the sense strand of a 1-kb coding region in *S. cerevisiae* (YOL086C). The graey bars near the *x*-axis indicate unmappable genomic locations. (**c**) The χ^2 goodness-of-fit statistics based on unadjusted and reweighted counts for 552 highly expressed regions of constant expression. (**d**) Smoothed histograms of the reduction in χ^2 goodness-of-fit statistics when using the re-weighting scheme, evaluated in five different experiments. Values greater than zero indicate that the re-weighting scheme improves the uniformity of the read distribution.

the beginning relative to the end are downweighted and vice versa. Accordingly, weights are of the following general form

$$w(h) \approx \frac{\hat{p}_{end\ of\ the\ reads}(h)}{\hat{p}_{start\ of\ the\ reads}(h)} \quad (2)$$

where $\hat{p}(h)$ is the proportion of reads starting/ending with heptamer h [in practice, we use a slight variation of this basic idea to compute the weights, see Equation (1) in 'Materials and Methods' section for a precise definition]. The weights are used in the following way: instead of counting the reads mapping to a genomic region (for example, an exon), the weights of the reads mapping to the region are added, see Table 1 for an example. Simply counting the reads mapping to a genomic region is equivalent to a reweighting scheme where all weights are equal to one. The weights are constructed in such a way that the distribution of the first heptamer of the re-weighted reads is equal to the distribution of the last heptamer of the unweighted reads.

Defining the weights based on 7 nt requires computing $4^7-1 = 16\,383$ frequencies. While the bias extends over the first 13 positions of the reads, we have found that weights based on only the first 7 nt perform the best (with weights based on either 6 or 8 nt performing almost as well). The reason that weights based on longer oligomers do not perform better is likely to be that the number of required oligomer frequencies increases exponentially while the amount of data stays constant. In addition, as longer sequences are used, certain sequences are only observed at the beginning or at the end of the reads, leading to weights being either zero or infinity. If sequencing depth is substantially increased, it might be possible to base the weights on >7 nt. Note that one of the data sets used for evaluation has >80 million mapped reads.

As detailed in 'Materials and Methods' section, we define the weights in a slightly different manner than the intuitive approach outlined above. Essentially, we average frequencies over several locations in the reads. Interestingly, we find that the performance of the reweighting scheme is substantially improved by averaging over the heptamer distributions starting at positions 1 and 2 (Supplementary Figure S8). Figure 1 shows that these two heptamer distributions are very different, since the marginal distributions of single nucleotides are very different. We propose two explanations for this improvement. First, is the well-known observation that the Illumina sequencer tends to have a higher error rate at the first base of the read (2). Second, the end repair performed as part of the standard protocol may shift the start position of the read relative to the binding of the random hexamer.

Using data from experiments in *S. cerevisiae* (8) and *H. sapiens* (9), we found that reweighting the reads improves their uniformity along expressed transcripts, although substantial heterogeneity remains (Figure 4 and Supplementary Figures S6–S8). We concentrated on non-short, highly expressed regions of constant expression, defined based on existing gene annotation (roughly equal to coding sequences in yeast and exons in human) and taking mappability into account. We used highly expressed regions (>1 read per base), because evaluating the effect of the methodology on base-level spatial heterogeneity requires a reasonable number of reads per base. For measuring the uniformity of the reads, we used the χ^2 goodness-of-fit statistic and the coefficient of variation. Both measures were substantially reduced (up to 50%) when re-weighting was used, although the statistics as well as qualitative evaluation suggest that the reweighted counts are still not uniformly distributed within an expressed transcript.

One possible concern with this methodology is the use of the same data set for computing the weights and evaluating the performance improvement, especially since we focus on highly expressed genes where most of the reads come from. We have also computed the weights using only reads that mapped outside of highly expressed genes and the performance improvement did not change. For convenience, we suggest using all mapped reads to compute the weights.

DISCUSSION

We have shown that priming using random hexamers biases the nucleotide content of RNA sequencing reads and that this bias also affects the uniformity of the locations of the reads along expressed transcripts. Despite this bias, we believe that priming using random hexamers is preferable to using oligo(dT) priming, as the latter is highly biased toward the 3′-end of the expressed transcripts.

Downloaded from http://nar.oxfordjournals.org at Univ of California Library on April 16, 2010

Mamanova *et al.* recently described an alternative protocol for sequencing RNA using the Illumina Genome Analyzer (18), in which reverse transcription takes place directly on the flow cell and which yields stranded reads and avoids polymerase chain reaction amplification. RNA is not converted to dscDNA using random priming, instead sequencing adapters are ligated directly onto RNA fragments. The ligated RNA library is then reverse transcribed on the flow cell. Data from their study does not show the nucleotide biases reported here (Supplementary Figure S9).

As an alternative to poly(A) enrichment, Armour *et al.* describe a protocol for ribosomal depletion, where only a subset of the 4096 possible hexamers are used to prime the reverse transcription (19). There are currently no publicly available data sets generated using this protocol.

A natural question is whether the read biases observed with the Illumina Genome Analyzer also occur with other sequencing platforms. We have begun investigating SOLiD data and our preliminary results based on data from (20) indicate that similar, but smaller random priming biases may be present in addition to more important SOLiD-specific biases.

We have provided a method for adjusting the nucleotide bias and have shown that this method makes the start positions of the reads closer to being uniformly distributed across the transcriptome. The method is implemented in the R package Genominator available from Bioconductor (5).

During the review of the present manuscript, we became aware of recent related work by Li *et al.* ('Modeling non-uniformity in short-read rates in RNA-Seq data', submitted for publication). These authors consider a different approach for handling the non-uniformity of the read distribution, which involves modeling the base-level read counts of a given gene using a Poisson model with parameter that is a function of the neighboring nucleotides. Compared to our reweighting procedure, the Li *et al.* method relies on gene annotation and does not relate to the library preparation protocol (i.e. random hexamer priming). In a similar spirit, Bullard (21) considers generalized linear models for base-level read counts in a simulation study assessing the performance of the differential allele-specific expression method proposed in Bullard *et al.* (22). However, preliminary attempts in the explicit modeling of base-level read counts have had limited success and more research and benchmarking is needed before one can confidently derive expression measures based on such models.

Alleviating the positional bias induced by random hexamer priming is one step toward enabling comparisons within samples, between distinct genomic regions. Such comparisons are crucial to several aims of transcriptome sequencing, including the study of alternative splicing, transcript assembly and allele specific expression.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We benefited from illuminating discussions with Nicholas T. Ingolia, Terence P. Speed, Margaret Taub and James H. Bullard.

FUNDING

U.S. National Institutes of Health, (grant U01 HG004271 to S. E. Celniker and grant R01 GM071655 to S.E.B.). Funding for open access charge: U.S. National Institutes of Health (grant U01 HG004271 to S. E. Celniker).

Conflict of interest statement. None declared.

REFERENCES

1. Mortazavi,A., Williams,B., McCue,K., Schaeffer,L. and Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.
2. Dohm,J.C., Lottaz,C., Borodina,T. and Himmelbauer,H. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.*, **36**, e105.
3. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
4. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.
5. Gentleman,R., Carey,V.J., Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.
6. Marioni,J., Mason,C., Mane,S., Stephens,M. and Gilad,Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.
7. Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.
8. Lee,A., Hansen,K.D., Bullard,J., Dudoit,S. and Sherlock,G. (2008) Novel low abundance and transient RNAS in yeast revealed by tiling microarrays and ultra high-throughput sequencing are not conserved across closely related yeast species. *PLoS Genet.*, **4**, e1000299.
9. Bullard,J.H., Purdom,E.A., Hansen,K.D. and Dudoit,S. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, **11**, 94.
10. Wang,J., Wang,W., Li,R., Li,Y., Tian,G., Goodman,L., Fan,W., Zhang,J., Li,J., Zhang,J. *et al.* (2008) The diploid genome sequence of an Asian individual. *Nature*, **456**, 60–65.
11. Bentley,D.R., Balasubramanian,S., Swerdlow,H.P., Smith,G.P., Milton,J., Brown,C.G., Hall,K.P., Evers,D.J., Barnes,C.L., Bignell,H.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.
12. Meissner,A., Mikkelsen,T.S., Gu,H., Wernig,M., Hanna,J., Sivachenko,A., Zhang,X., Bernstein,B.E., Nusbaum,C., Jaffe,D.B. *et al.* (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, **454**, 766–770.
13. Chiang,D., Getz,G., Jaffe,D., O'Kelly,M., Zhao,X., Carter,S., Russ,C., Nusbaum,C., Meyerson,M. and Lander,E. (2008) High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods*, **6**, 99–103.
14. Mikkelsen,T.S., Hanna,J., Zhang,X., Ku,M., Wernig,M., Schorderet,P., Bernstein,B.E., Jaenisch,R., Lander,E.S. and Meissner,A. (2008) Dissecting direct reprogramming through integrative genomic analysis. *Nature*, **454**, 49–55.

Downloaded from http://nar.oxfordjournals.org at Univ of California Library on April 16, 2010

15. Wilhelm,B., Marguerat,S., Watt,S., Schubert,F., Wood,V., Goodhead,I., Penkett,C., Rogers,J. and Bahler,J. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, **453**, 1239–1243.
16. Bloom,J.S., Khan,Z., Kruglyak,L., Singh,M. and Caudy,A.A. (2009) Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. *BMC Genomics*, **10**, 221.
17. Nagalakshmi,U., Wang,Z., Waern,K., Shou,C., Raha,D., Gerstein,M. and Snyder,M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–1349.
18. Mamanova,L., Andrews,R.M., James,K.D., Sheridan,E.M., Ellis,P.D., Langford,C.F., Ost,T.W.B., Collins,J.E. and Turner,D.J. (2010) FRT-seq: amplification-free, strand-specific transcriptome sequencing. *Nat. Methods*, **7**, 130–132.
19. Armour,C.D., Castel,C.C., Chen,R., Babak,T., Loerch,P., Jackson,S., Shah,J.K., Dey,J., Rohl,C.A., Johnson,J.M. *et al.* (2009) Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. *Nat. Methods*, **6**, 647–649.
20. Passalacqua,K.D., Varadarajan,A., Ondov,B.D., Okou,D.T., Zwick,M.E. and Bergman,N.H. (2009) Structure and complexity of a bacterial transcriptome. *J. Bacteriol.*, **191**, 3203–3211.
21. Bullard,J.H. (2009) Statistical methods and software for high-throughput gene expression experiments. *Ph.D. thesis*, Graduate Group in Biostatistics, University of California, Berkeley, Berkeley, CA 94720-7358.
22. Bullard,J.H., Mostovoy,Y., Dudoit,S. and Brem,R.B. (2010) Polygenic and directional regulatory evolution across pathways in Saccharomyces. *Proc. Natl Acad. Sci. USA*, **107**, 5058–5063.

Downloaded from http://nar.oxfordjournals.org at Univ of California Library on April 16, 2010