Gene 557 (2015) 82–87

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Diminishing returns in next-generation sequencing (NGS) transcriptome data

Rex Lei [a,b], Kaixiong Ye [a], Zhenglong Gu [a,*], Xuepeng Sun [a,c,**]

[a] Division of Nutritional Sciences, Cornell University, Ithaca, NY 14853, USA
[b] Ithaca High School, Ithaca, NY 14853, USA
[c] College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310058, PR China

ARTICLE INFO

Article history:
Received 14 February 2014
Received in revised form 5 December 2014
Accepted 8 December 2014
Available online 10 December 2014

Keywords:
RNA-seq efficiency

ABSTRACT

RNA-seq is increasingly used to study gene expression of various organisms. While it provides a great opportunity to explore genome-scale transcriptional patterns with tremendous depth, it comes with prohibitive costs. Establishing a minimal sequencing depth for required accuracy will guide cost-effective experimental design and promote the routine application of RNA-seq. To address this issue, we selected 36 RNA-seq datasets, each with more than 20 million reads from six widely-used model organisms: *Saccharomyces cerevisiae*, *Homo sapiens*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Mus musculus*, and *Arabidopsis thaliana*, and investigated statistical correlations between the sequencing depth and the outcome accuracy. To achieve this, we randomly chose reads from each dataset, mapped them to the reference genomes, and analyzed the accuracy achieved with varying coverage. Our results indicated that as low as one million reads can provide the same sequencing accuracy in transcript abundance ($r = 0.99$) as >30 million reads for highly-expressed genes in all six species. Because many metabolically and pathologically-relevant genes are highly expressed, our findings might be instructive for cost-effective experimental designs in NGS-based research and also provide useful guidance to similar research for other organisms.

1. Introduction

Regulatory diversity in the transcriptome, such as regulatory mutations or perturbed signaling pathways, often leads to phenotypic or functional differences including diseases (Sul et al., 2009; Marguerat and Bahler, 2010; Ozsolak and Milos, 2011). Characterizing transcriptomic regulation is essential to understanding the molecular mechanisms of basic biology and the pathogeneses of human diseases. Over the years, many methods have been developed to study gene expression and regulation. One that has been primarily used in the past two decades is the DNA microarray, a hybridization-based approach (DeRisi et al., 1996; Wang et al., 2009). While insightful knowledge of the transcriptome and its regulations has been obtained using this method, it has several drawbacks. First, microarray analysis is an indirect method that relies on probes to detect gene expression (Cassone et al., 2007). Second, microarrays must use individual, pre-prepared gene probes and have background, cross-reaction, and reproducibility issues (Draghici et al., 2006; Okoniewski and Miller, 2006). Third, microarrays cannot determine unidentified genes because they rely on well-known, preexisting probes for detection. Finally, microarrays have low sensitivities for lowly-expressed genes (Heller, 2002; Wang et al., 2009).

Recently, a new method called RNA-seq has emerged as a popular alternative to quantify mRNA abundance (Mortazavi et al., 2008; Nagalakshmi et al., 2008). As a digital process, RNA-seq analyzes the transcriptome by recording frequencies and alterations of transcripts in test samples. Comparatively, RNA-seq has many advantages over DNA microarrays. RNA-seq does not require prior knowledge of the target sequence and its use of digital detection reduces errors that rise from indirect probe hybridization in microarrays. Furthermore, RNA-seq is more sensitive than microarrays with respect to measuring low-abundance genes (Marioni et al., 2008).

However, RNA-seq's major disadvantage is its cost. Despite the rapid improvement in its efficiency, RNA-seq is still too expensive for most research laboratories and for routine applications. Although developing barcodes for multiplexing samples (Smith et al., 2010; Islam et al., 2011; Wang et al., 2011a) could decrease the cost, serious concern has risen regarding the sufficient depth for all the individual samples sequenced together. While deeper sequencing generally gives a more accurate picture of the whole transcriptome, especially for genes with low abundance (Tarazona et al., 2011), it remains unclear if appropriately reducing the sequencing depth for certain purposes could reduce the sequencing cost and time without sacrificing accuracy.

Abbreviations: RNA-seq, RNA sequencing

* Correspondence to: Z. Gu, 312 Savage Hall, Cornell University, Ithaca, NY 14853, USA.
** Correspondence to: X. Sun, 344 Savage Hall, Cornell University, Ithaca, NY 14853, USA.
E-mail addresses: zg27@cornell.edu (Z. Gu), xs57@cornell.edu (X. Sun).

http://dx.doi.org/10.1016/j.gene.2014.12.013
0378-1119/

Although RNA-seq datasets have been generated for many species, the balance between the sequencing depth and accuracy was only investigated in limited species such as chicken and bacteria (Wang et al., 2011b; Haas et al., 2012). We conducted this research to explore the relationship between sequencing depth and accuracy in various species, aiming at providing reference points for researchers by analyzing diverse model organisms. Interestingly, we found that, for high-abundance genes, one million reads basically provided information of similar quality regarding expression abundance to that of more than 30 million reads in all studied species. Implication and limitation of our results were further discussed.

2. Materials and methods

2.1. Selection of species

Six widely studied species, *Saccharomyces cerevisiae* (yeast), *Homo sapiens* (human), *Drosophila melanogaster* (fly), *Caenorhabditis elegans* (worm), *Mus musculus* (mouse), and *Arabidopsis thaliana* (plant) were selected in this project. These eukaryotes were chosen partially for their evolutionary distances, so that one could evaluate the differences and similarities among them and generate broader conclusions. In addition, these species are model organisms for the biomedical research and have numerous accessible experimental datasets.

2.2. Data sources

We analyzed a total of 36 different RNA-seq datasets. Among them, 35 were downloaded from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) database (http://www.ncbi.nlm.nih.gov/sra), and the remaining one (*S. cerevisiae*) was our laboratory-produced dataset generated by Illumina Hi-Seq 2000 Platform. The RNA-seq datasets downloaded from NCBI were all checked to fit the following criteria: they were all single-end, selected from cDNA, transcriptomic, and done on the platform Illumina. Of the NCBI datasets that fit these requirements, the five largest files that had a mapping accuracy of >65% were selected and used for all the species except for yeast (*S. cerevisiae*) and plant (*A. thaliana*). For the yeast data, the 7 largest files found in the NCBI, in addition to our own laboratory data (which fits the above criteria), were used. For the plant data, the 8 largest files found were used, and two of them had mapping accuracies of ~50%. The genome sequence and predicted gene model files for each species were downloaded from the species-specific databases: *Saccharomyces* genome database (SGD) (Cherry et al., 1998) for *S. cerevisiae*, University of California, Santa Cruz (UCSC)'s genome browser (Kent et al., 2002) for *H. sapiens* and *M. musculus*, Wormbase (Stein et al., 2001) for *C. elegans*, Flybase (Drysdale and Crosby, 2005) for *D. melanogaster*, and the *Arabidopsis* information resource (TAIR) for *A. thaliana*.

2.3. Links to data

C. elegans genome sequence and gene models:
ftp://ftp.wormbase.org/pub/wormbase/species/c_elegans/sequence/genomic/c_elegans.WS238.genomic.fa.gz; and ftp://ftp.wormbase.org/pub/wormbase/species/c_elegans/gff/c_elegans.WS238.annotations.gff3.gz

D. melanogaster genome sequence and gene models:
ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/current/fasta/dmel-all-chromosome-r5.52.fasta.gz; and ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/current/gff/dmel-all-r5.52.gff.gz

M. musculus genome sequence and gene models:
http://hgdownload.soe.ucsc.edu/goldenPath/mm10/bigZips/

H. sapiens genome sequence and gene models:
http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/

S. cerevisiae genome sequence and gene models:
http://www.yeastgenome.org/download-data/sequence

A. thaliana genome sequence and gene models:
ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR10_genome_release/TAIR10_gff3/TAIR10_GFF3_genes.gff
ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR10_genome_release/TAIR10_chromosome_files/TAIR10_chr_all.fas

2.4. Yeast growth condition and RNA-seq experiment

One set of in-house generated RNA-seq for *S. cerevisiae* was included in the study. To conduct the experiment, the wild type BY4741 (MATa, *his3Δ1*, *leu2Δ0*, *met15Δ0*, *ura3Δ0*) strain was cultured overnight in the YPD (yeast extract, peptone, and dextrose) medium, and transferred into fresh YPEG (yeast extract, peptone, and glycerol) medium for full growth. Total RNA was extracted from the yeast cells by the standard Trizol protocol (Life Technologies, Grand Island, NY), and the mRNA was then purified using oligo-dT DynaBeads. The cDNA sequencing library was constructed according to the protocol described by Wang et al. (2011a). After sequencing, the reads were mapped into *S. cerevisiae* genome by bowtie2 (Langmead and Salzberg, 2012) with no more than two mismatches. The number of mapped reads per kilobase per million (RPKM) was used to represent the expression level of each gene.

2.5. Computational analysis for the transcriptome data

The sequencing files downloaded from NCBI SRA database were initially converted from SRA format to FASTQ format using SRA toolkit (http://www.ncbi.nlm.nih.gov/Traces/sra/?view=software). Then, the raw data were filtered using the following criteria: (1) the number of unknown bases (N) was no more than two for each read; and (2) the fraction of low quality sites (Q < 5) was no more than 50% for each read. The data that passed this quality control were then used to map back to their respective genome sequences using bowtie2 (Langmead and Salzberg, 2012). Only uniquely mapped reads with no more than two mismatches were retained for further analysis. After mapping, the counts for each gene were summarized using HTSeq (http://www-huber.embl.de/users/anders/HTSeq/doc/overview.html). In the simulation, a predetermined-sized subset of reads was randomly selected from the original file. Using the same mapping procedure as mentioned above, the RPKM for each gene and depth of coverage were calculated and compared with those from original data. In-house Perl and R scripts were developed for data analysis and graphing (available upon request).

3. Results and discussions

3.1. Pearson correlation coefficients between simulated subsets and the original reads

To analyze the relationship between the sequencing depth and accuracy of the measurement of gene expression level, we compared the expression levels (RPKM) of all genes estimated with simulated reduced data subsets to those estimated with the original full dataset. Pearson correlation coefficients were calculated to measure the consistency of measurements. For each species, eight simulations were run for each data size and the standard error (SE) of mean correlation was obtained for each sampling.

As shown in Fig. 1, regardless of species and data size, correlation graphs displayed a similar, nonlinear relationship between correlation coefficients and number of reads in the selected subsets. The correlation coefficients between simulated small-sized subsets and the original, big data rapidly increase and then plateau at roughly 10–15% of the entire sample size. However, more importantly, the correlation coefficients for all sample sizes (>1 million reads) are larger than 0.99. Indeed, aside from the yeast dataset (170 million reads), each coverage graph displayed greater than 99.9% correlation with only one million reads, indicating that with this sequencing depth the relative level of transcription can be accurately estimated for those covered genes.

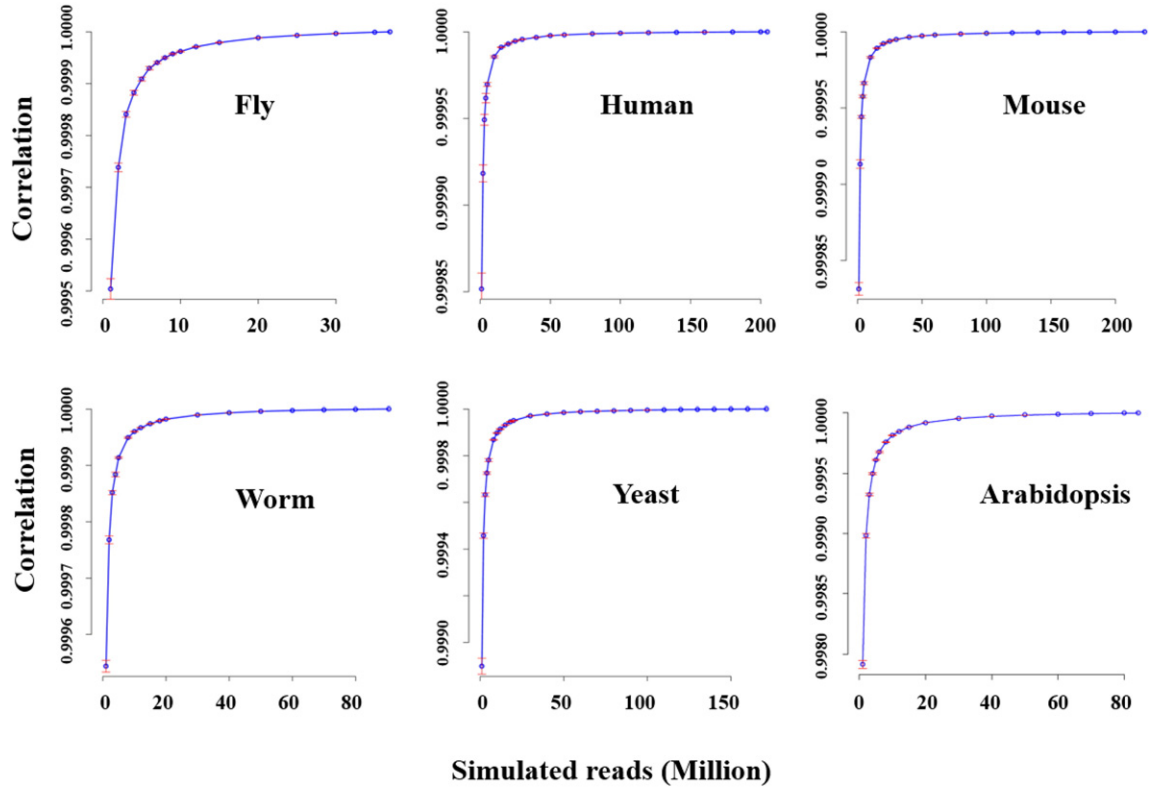

Fig. 1. Correlation coefficients of transcript abundance between the simulated reduced subsets and the original data. For each dataset, reduced subsets were generated by randomly selecting reads from the original data. The correlation between each subset and the entire dataset was calculated and plotted by simulated subset size (number of reads). All genes, regardless of whether they can be detected, were included in the analysis. For each species, we used the same annotation file for the original data and its derived (sampled) data, and counted the read number for all gene models in the annotation file. If a gene was not detected in some randomly sampled data, we marked the read number for this gene as zero. The simulation was run 8 times for each subset size and the standard error (SE) for each dataset was calculated.

3.2. Gene coverage for the simulated subsets of the RNA-seq data

To investigate how many genes can be covered by different sequencing depths, we conducted the following simulations. (Note: We defined a gene to be covered if 10 or more reads can be reliably mapped to it. The total gene coverage was the number of genes which satisfied this criterion.) As described above, we randomly selected a predetermined amount of reads, grouped them into a dataset, and calculated how many genes were covered in that dataset. Simulations were run 8 times for each sample size to obtain the mean value and SE for gene coverage estimates. As shown in Fig. 2, the gene coverage increased rapidly when the sizes of the subsets were small (e.g. <10 million reads) and approached plateaus. The specific number of reads at which coverage started to plateau was different among species. Unlike the correlation coefficient of the expression level, the gene coverage never fully flattened, as low-expression genes were continuing to be covered at higher depths. In both the coverage and correlation graphs, the margin of error in the random sampling was negligible in comparison to the differences between the coverage/correlation values. Combining Figs. 1 and 2, our results reaffirm that a diminishing-return relationship exists between sequencing depth and information regarding transcription abundance and coverage in most RNA-seq experimental designs.

3.3. Detection efficiency for genes with various levels of transcript abundance

To understand this diminishing-return relationship between sequencing depth and information on transcription, we plotted the distribution of reads on the genes (with genes sorted by number of reads). Fig. 3 shows a non-uniform distribution of gene transcript level in each species. In a one-million read sample, a significant amount of genes are already covered due to their relatively high expression. On the other hand, very deep coverage in sequencing may not reveal much more information because many genes in the genome are not expressed enough to be detected.

To further illustrate this point, we calculated the percentage of genes that were covered in each species for the smallest set of the simulated sample sizes (1–5 million reads). As shown in Fig. 4, for Yeast, Human, Mouse, and Fly data, a sample of 1 million reads had roughly half of the coverage for the entire sequenced dataset that typically includes more than 30 million reads, and sometime could be around 200 million reads (for human and mouse). The 5 million read samples had roughly 75% (or more) of the coverage of the entire sequenced samples. The worm samples' coverage seemed to grow more slowly from ~33% coverage at 1 million reads to ~50% coverage at 5 million reads.

3.4. Caveats and conclusions

The most noticeable finding from this study is that a low sequencing depth such as 1 million reads demonstrates very similar information regarding transcript abundance and coverage to that of a much higher sequencing depth (>30 million reads) for roughly half of the expressed genes (Fig. 4). The strength of this correlation was consistently shown in all six examined species, with a wide range of genome sizes. Our finding represents a general pattern instead of just a special case, and offers guidance for selecting the depth of RNA-seq for various organisms with different objectives. Some extremely low-abundant genes, despite expensive sample sizes (100 million or more), may still remain uncovered or barely covered (Wang et al., 2011b). So, direct RNA-seq is a cost-ineffective process for analyzing these genes. To better investigate and understand these genes, methods with certain way of enrichment before RNA-seq should be pursued.

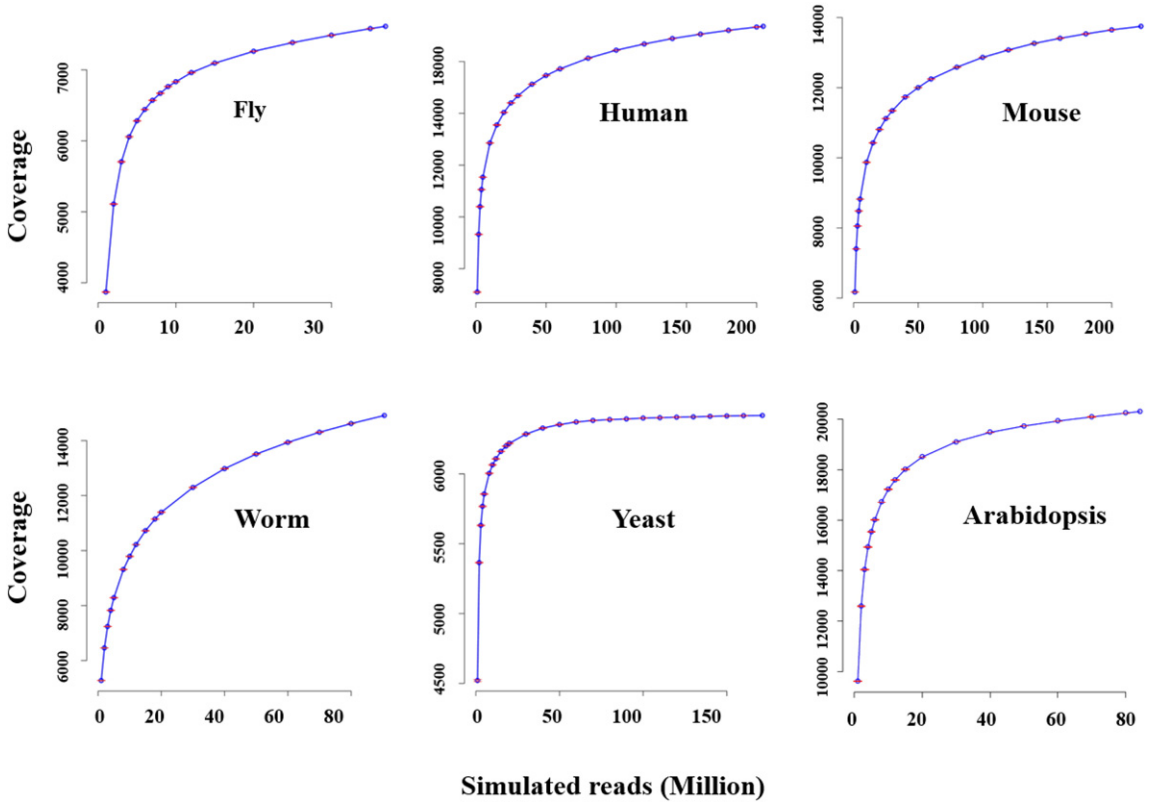

Fig. 2. The gene coverage of various sequencing depths. For each dataset, the simulated subsets were generated by randomly selecting sequencing reads. Every gene that has 10 or more reads is considered as covered in the simulated data. The total coverage for a subset is the number of the covered genes in that dataset. Data for each subset size was simulated 8 times to ensure accuracy. SE was calculated and plotted for each subset.

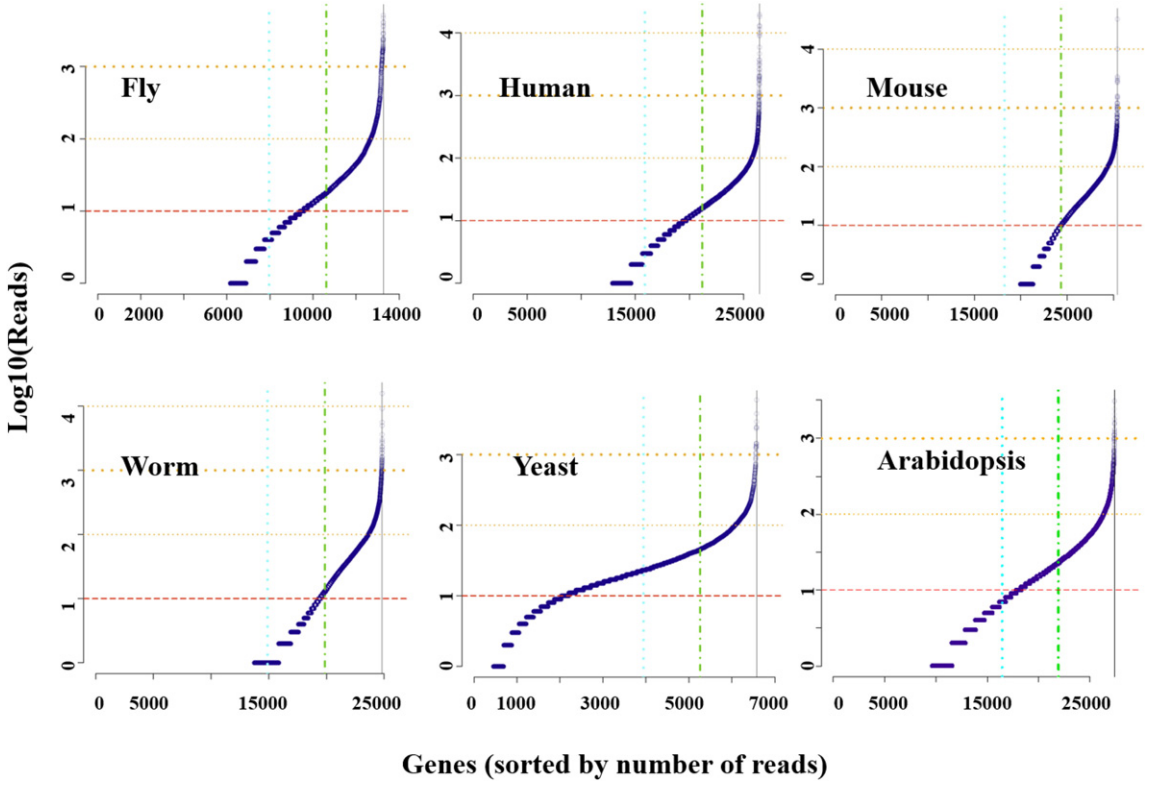

Fig. 3. Number of reads on genes in one million read samples. For each species, one million reads were randomly drawn from the largest dataset for the species. Genes, including those with zero coverage, were sorted by the number of mapped reads. The number of mapped reads in logarithmic form is indicated on the y axis. The red horizontal line indicates the number of reads a gene needed to be considered covered (10). The vertical green and blue lines indicate the 80th and 60th percentiles, respectively, of the total number of genes for each species. Genome sizes for Fly, Human, Mouse, Worm, Yeast, and Plant were 13,269, 26,468, 30,428, 24,831, 6575, and 27,416 genes, respectively.

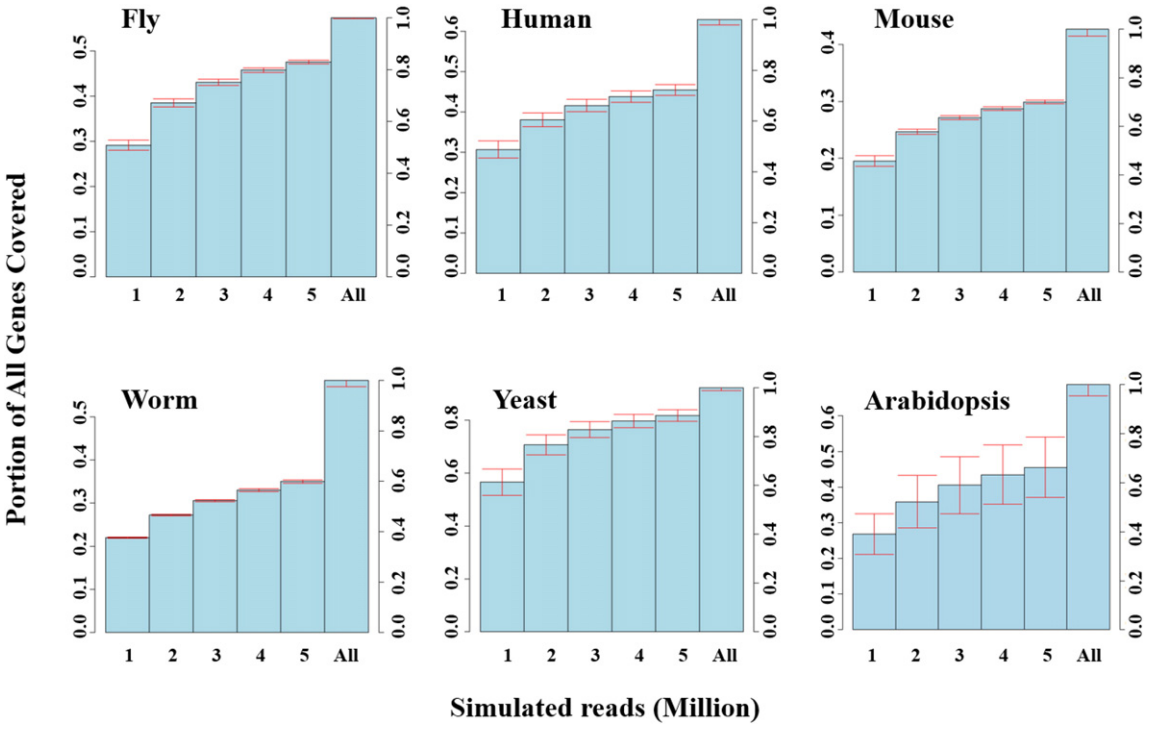

Fig. 4. Gene coverage with small sample sizes (1 to 5 million reads). For each species, the gene coverage from different datasets was calculated at randomly drawn small subset (1–5 million reads). The coverages for five simulated datasets were averaged to produce the bar graphs; one standard error was also calculated and plotted. The left y-axis indicates how many genes were covered in the subset compared to the entire genome (as a decimal), while the right y-axis shows how many genes were covered in the subset compared to the entire, large set used in the study (>30 million reads), which is denoted on the x-axis as "All".

One important factor that could potentially confound the RNA-seq accuracy is the read length. To take account of the read length bias in our analysis, we initially picked the data with a different read length for each species (Supplementary table S1). As observed from the error bars of each species, the read length difference did not cause a high variation for the corresponding estimates, indicating that the read length might not affect our general conclusions. In this study, we have focused on the single-end RNA-seq sequencing, which is dominant in the field of transcriptome profiling. However, we are aware that many data in the SRA database are paired-end. To extend our understanding for this type of data, we analyzed 24 paired-end datasets from these six species (Supplementary Table S2). As shown in Supplementary Figs. S1 and S2, we got the similar conclusions using the paired-end data as those inferred from the single-end data.

It is worth noting that our analysis did not consider other information of the transcriptome. For example, it is reasonable to speculate that the requirement for minimal total reads should be increased to investigate allelic expression in diploid cells, or alternative splicing patterns. And, even many genes are detected in one million reads, the statistical power for detecting the significance of gene transcription changes might be compromised due to relatively low sequencing counts. Nevertheless, our results indicate that sequencing depth as low as 1-million reads basically provided similar information of transcript abundance to that of more than 30 million reads for highly expressed genes. Because genes that are important in regulating metabolism and pathogenesis of diseases usually have high abundance of transcription, our results might enable researchers to conduct minimal depth sequencing while achieving satisfactory accuracy.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gene.2014.12.013.

Acknowledgments

We would like to thank other members of the Gu lab for our discussions and for reading this manuscript. Part of this research was supported by NSF MCB-1243588 (to ZG) and the Chinese Visiting Scholar Fellowship (No.201206320045) (to XS).

References

Cassone, M., Giordano, A., Pozzi, G., 2007. Bacterial DNA microarrays for clinical microbiology: the early logarithmic phase. Front. Biosci. 12, 2658–2669.
Cherry, J.M., Adler, C., Ball, C., Chervitz, S.A., Dwight, S.S., Hester, E.T., Jia, Y.K., Juvik, G., Roe, T., Schroeder, M., Weng, S.A., Botstein, D., 1998. SGD: Saccharomyces Genome Database. Nucleic Acids Res. 26, 73–79.
DeRisi, J., Penland, L., Brown, P.O., Bittner, M.L., Meltzer, P.S., Ray, M., Chen, Y.D., Su, Y.A., Trent, J.M., 1996. Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat. Genet. 14, 457–460.
Draghici, S., Khatri, P., Eklund, A.C., Szallasi, Z., 2006. Reliability and reproducibility issues in DNA microarray measurements. Trends Genet. 22, 101–109.
Drysdale, R.A., Crosby, M.A., 2005. FlyBase: genes and gene models. Nucleic Acids Res. 33, D390–D395.
Haas, B.J., Chin, M., Nusbaum, C., Birren, B.W., Livny, J., 2012. How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? BMC Genomics 13, 734.
Heller, M.J., 2002. DNA microarray technology: devices, systems, and applications. Annu. Rev. Biomed. Eng. 4, 129–153.
Islam, S., Kjallquist, U., Moliner, A., Zajac, P., Fan, J.B., Lonnerberg, P., Linnarsson, S., 2011. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res. 21, 1160–1167.
Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., Haussler, D., 2002. The human genome browser at UCSC. Genome Res. 12, 996–1006.
Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359.
Marguerat, S., Bahler, J., 2010. RNA-seq: from technology to biology. Cell. Mol. Life Sci. 67, 569–579.
Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., Gilad, Y., 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 18, 1509–1517.
Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., Wold, B., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods 5, 621–628.
Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., Snyder, M., 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320, 1344–1349.
Okoniewski, M.J., Miller, C.J., 2006. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. BMC Bioinformatics 7, 276.
Ozsolak, F., Milos, P.M., 2011. RNA sequencing: advances, challenges and opportunities. Nat. Rev. Genet. 12, 87–98.
Smith, A.M., Heisler, L.E., St Onge, R.P., Farias-Hesson, E., Wallace, I.M., Bodeau, J., Harris, A.N., Perry, K.M., Giaever, G., Pourmand, N., Nislow, C., 2010. Highly-multiplexed

barcode sequencing: an efficient method for parallel analysis of pooled samples. Nucleic Acids Res. 38, e142.
Stein, L., Sternberg, P., Durbin, R., Thierry-Mieg, J., Spieth, J., 2001. WormBase: network access to the genome and biology of *Caenorhabditis elegans*. Nucleic Acids Res. 29, 82–86.
Sul, J.Y., Wu, C.W., Zeng, F., Jochems, J., Lee, M.T., Kim, T.K., Peritz, T., Buckley, P., Cappelleri, D.J., Maronski, M., Kim, M., Kumar, V., Meaney, D., Kim, J., Eberwine, J., 2009. Transcriptome transfer produces a predictable cellular phenotype. Proc. Natl. Acad. Sci. U. S. A. 106, 7624–7629.
Tarazona, S., Garcia-Alcalde, F., Dopazo, J., Ferrer, A., Conesa, A., 2011. Differential expression in RNA-seq: a matter of depth. Genome Res. 21, 2213–2223.
Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63.
Wang, L., Si, Y.Q., Dedow, L.K., Shao, Y., Liu, P., Brutnell, T.P., 2011a. A low-cost library construction protocol and data analysis pipeline for Illumina-based strand-specific multiplex RNA-Seq. PLoS One 6, e26426.
Wang, Y., Ghaffari, N., Johnson, C.D., Braga-Neto, U.M., Wang, H., Chen, R., Zhou, H.J., 2011b. Evaluation of the coverage and depth of transcriptome by RNA-Seq in chickens. BMC Bioinformatics 12, S5.