GENOME RESEARCH

Gene expression profiling by massively parallel sequencing

Tatiana Teixeira Torres, Muralidhar Metta, Birgit Ottenwälder and Christian Schlötterer

Genome Res. 2008 18: 172-177; originally published online Nov 21, 2007;
Access the most recent version at doi:10.1101/gr.6984908

Gene expression profiling by massively parallel sequencing

Tatiana Teixeira Torres,[1] Muralidhar Metta,[1] Birgit Ottenwälder,[2] and Christian Schlötterer[1,3]

[1]Institut für Tierzucht und Genetik, Veterinärmedizinische Universität Wien, 1210 Vienna, Austria; [2]Eurofins Medigenomix GmbH, 82152 Martinsried, Germany

Massively parallel sequencing holds great promise for expression profiling, as it combines the high throughput of SAGE with the accuracy of EST sequencing. Nevertheless, until now only very limited information had been available on the suitability of the current technology to meet the requirements. Here, we evaluate the potential of 454 sequencing technology for expression profiling using *Drosophila melanogaster*. We show that short (< ~80 bp) and long (> ~300–400 bp) cDNA fragments are under-represented in 454 sequence reads. Nevertheless, sequencing of 3′ cDNA fragments generated by nebulization could be used to overcome the length bias of the 454 sequencing technology. Gene expression measurements generated by restriction analysis and nebulization for fragments within the 80- to 300-bp range showed correlations similar to those reported for replicated microarray experiments (0.83–0.91); 97% of the cDNA fragments could be unambiguously mapped to the genomic DNA, demonstrating the advantage of longer sequence reads. Our analyses suggest that the 454 technology has a large potential for expression profiling, and the high mapping accuracy indicates that it should be possible to compare expression profiles across species.

[Supplemental material is available online at www.genome.org. The EST sequences have been deposited in GenBank under accession nos. EV574767–EV600806.]

Gene expression technologies have greatly matured over the past years, but it has become clear that hybridization-based approaches have obvious limitations in cross-species comparisons (Gilad et al. 2005, 2006). Probably the most eminent problems are mismatches in heterologous probes and probe-specific hybridization kinetics, which complicate the design of species-specific oligonucleotide arrays. Alternatively, sequencing-based approaches could be used to measure gene expression if the sequence reads could be unambiguously mapped to the corresponding transcripts. While the short sequence reads of serial analysis of gene expression (SAGE) (Velculescu et al. 1995) and related techniques are severely limited by the requirement of a reliable genome annotation, the recently developed 454 sequencing technology (Margulies et al. 2005) may provide sufficient sequence information to overcome this limitation at moderate costs.

In this study, we evaluate the potential of 454 sequencing technology to serve as a reliable tool for expression profiling. We show that 454 sequencing technology has a biased representation of cDNA fragments with different length. However, in combination with random breakage of the cDNAs by nebulization, 454 sequencing provides an excellent tool for expression profiling. The high accuracy with which we could map the sequenced fragments onto the *Drosophila melanogaster* genome suggests that 454 sequencing has great potential for interspecific expression profiling.

Results

Conceptual design

Measuring gene expression by sequencing requires only that a proportion of the transcript be analyzed. We sequenced a 3′ region of the cDNA to avoid potential bias due to incomplete reverse transcription of the mRNAs. We used two different approaches to evaluate the potential of 454 sequencing for expression profiling. First, we generated well-defined 3′ cDNA fragments by restriction enzyme treatment (Fig. 1). Within the limitations of the available genome annotation, we could predict the expected 3′ cDNA fragments, as we used a *D. melanogaster* strain with a fully sequenced genome. This strategy allowed us to evaluate whether fragment size affected 454 sequencing efficiency. As a second strategy, we sequenced 3′ cDNA ends that were generated by random shearing of the cDNA (Fig. 1). Use of the same mRNA for both approaches allowed comparison between the two different strategies and thus a measurement for the reliability of 454 sequencing-based expression profiling.

Sequencing and mapping efficiency

The first prerequisite for a reliable measure of gene expression based on absolute counts is accurate identification of the transcript corresponding to the sequence read. Totals of 11,477 and 14,570 reads with average read lengths of 114 ± 20 and 116 ± 23 bp were collected from digested (DIG) and nebulized (NEB) samples, respectively. After raw data were processed to eliminate low-quality sequences and poly(A) tails, we obtained 11,437 (DIG) and 14,512 (NEB) high-quality short expressed sequence tags (ESTs) (Table 1).

The 454-ESTs were mapped to transcripts or genomic DNA

[3]**Corresponding author.**
E-mail christian.schloetterer@vu-wien.ac.at; fax 43-1-250775693.

Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6984908. Freely available online through the *Genome Research* Open Access option.

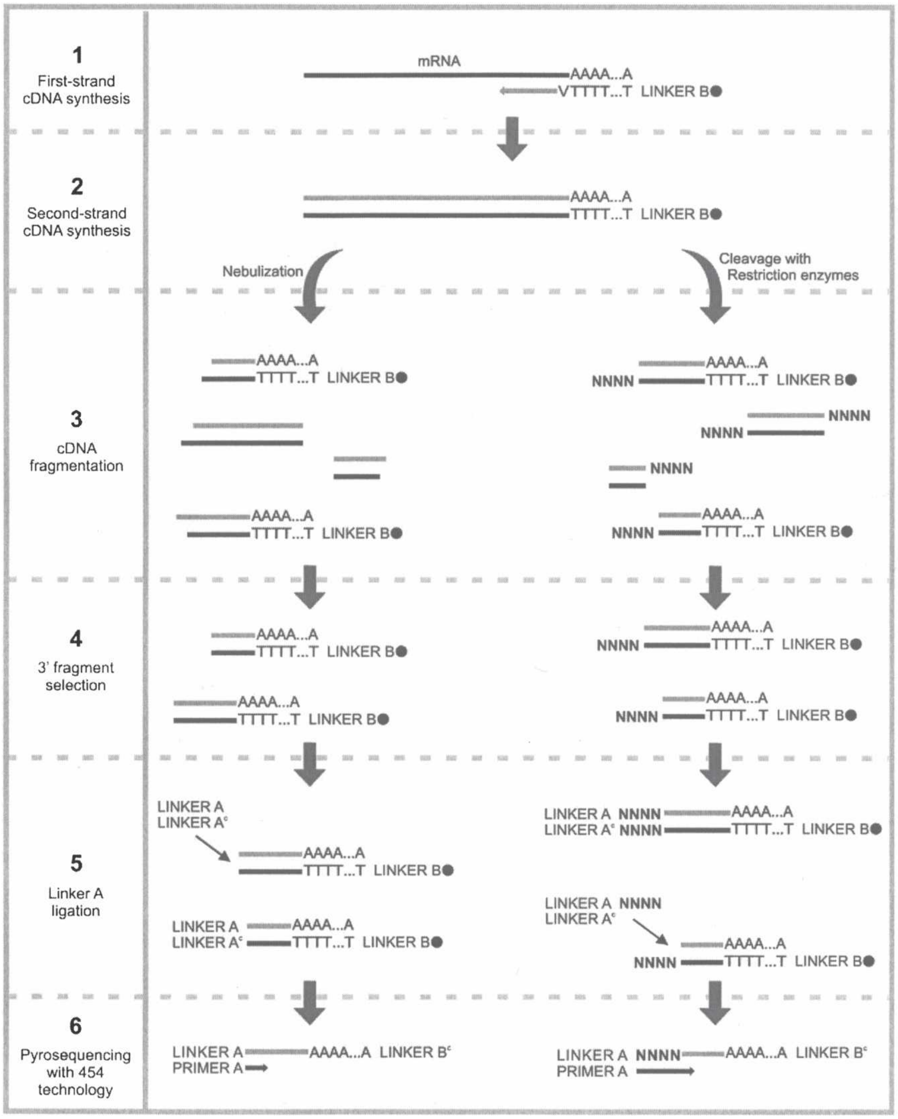

Figure 1. Overview of the methods used to generate 3′ cDNA fragments. Double-stranded cDNA was fragmented using one of the two different strategies: restriction enzyme treatment or nebulization (step *3*). 3′ Fragments were recovered (step *4*) and ligated to specific linkers (step *5*). cDNA fragments were then sequenced using 454 sequencing technology (step *6*).

of *D. melanogaster* (release 5.1) using BLAST with highly stringent criteria ($E < 10^{-10}$, >90% identity, >50% of read length included in the high-scoring segment pair [HSP]). About two-thirds of the ESTs could be unambiguously mapped to the database of protein-coding transcript: 97% were unambiguously mapped to the genome and 90% to annotated genes (Table 1). Thus, 7% of the ESTs were not mapped to annotated gene sequences in *D. melanogaster* collection despite having an unambiguous match in the genomic sequence. To test if this discrepancy could be explained by incomplete annotation of the transcripts in *D. melanogaster* (i.e., lack of 3′ UTR, or missing isoforms), we performed a BLAST search against portions of 3′ flanking sequences of 500 and 2000 bp. This improved the mapping efficiency to 95%, indicating that information on 3′ UTRs is still missing or incomplete even for the well-characterized *D. melanogaster* genome. Further support for incomplete 3′ UTR information is provided by the higher mapping efficiency of TaiI library, which harbored, on average, longer cDNA fragments and had a higher probability to overlap with the coding part of the transcript. Five percent of the ESTs that were not mapped to the transcript database could be located on intronic regions, suggesting the presence of new isoforms. We also found that 3%–6% of the hits to the transcript database consisted of antisense transcripts from 5%–6% of the genes sampled (Table 1).

Assessment of biases in transcript representation

Accurate measurement of gene expression with 454 sequencing technology requires an unbiased representation of the cDNA molecules, irrespective of length or sequence composition. As the expression intensity is not known, we designed a test that did not depend on the gene expression level. We used the sequenced *D. melanogaster* strain, hence it was possible to predict the restriction fragment length of every known transcript. We obtained an expected fragment-length distribution by an in silico restriction analysis of all annotated transcripts. To compare this expected distribution to the observed one, we considered every identified transcript only once. This procedure is expected to result in a good approximation of the fragment-length distribution for an unbiased sequencing procedure.

As 454 sequencing reads are frequently too short to cover the entire 3′cDNA fragment, we estimated the fragment length by matching the sequence read to the transcripts and determining the number of bases between the first base of the alignment and the 3′ end of the reference transcript. While the predicted 3′ cDNA fragment-length distribution differed among the three restriction enzymes used, we consistently observed a striking difference between the expected and observed distributions. For all enzymes tested, ESTs shorter than ~80 bp or longer than 300 bp were under-represented (Fig. 2). The under-representation of short fragments results in part from the filtering of the 454 sequencing software, which requires a minimum read length. Thus, these fragments are completely absent. The filtering, however, does not explain why fragments longer than the software threshold are under-represented. It is possible that this bias is produced during library preparation, but it is not entirely clear which step in the 454 sequencing procedure caused this effect. We can only speculate that short fragments are lost during enrichment of DNA capture beads carrying amplification products. Capture beads loaded with small fragments may not be recovered by the magnetic beads as effectively as those with longer frag-

Table 1. Summary statistics for 454-ESTs sequencing and mapping to *D. melanogaster* genomes and annotated transcripts (release 5.1)

		DIG ESTs			
	NEB ESTs	Total	MboI	NlaIII	Tail
454-ESTs collection, mapping, and gene coverage					
Raw reads	14,570	11,477			
After quality control	14,512 (100%)	11,437 (100%)	4,503 (100%)	3,292 (100%)	2,844 (100%)
Mapped to the genome[a]	14,037 (97%)	11,145 (97%)	4,410 (98%)	3,197 (97%)	2,781 (98%)
Mapped to the genome[b]	14,401 (99%)	11,358 (99%)	4,478 (99%)	3,273 (99%)	2,825 (99%)
Mapped to the transcripts[c]	9,560 (66%)	7,811 (68%)	2,806 (62%)	2,066 (63%)	2,429 (85%)
Antisense hits to transcripts	321	501	165	86	118
Genes sampled	2,555	2,472	1,156	1,007	1,059
Genes with antisense hits	146	146	54	50	66
Breakdown of ESTs not mapped to annotated transcripts					
Not mapped to transcripts[c]	4,952 (100%)	3,626 (100%)	1,697 (100%)	1,226 (100%)	415 (100%)
Mapped to the genome	4,588 (93%)	3,374 (93%)	1,619 (95%)	1,144 (93%)	361 (87%)
Mapped to annotated genes	3527 (71%)	2620 (72%)	1412 (83%)	919 (75%)	82 (20%)
Mapped to intergenic sequences	858 (17%)	651 (18%)	169 (10%)	178 (14%)	270 (65%)
Mapped to 2000 bp on the 3′[d]	554 (11%)	442 (12%)	152 (9%)	103 (8%)	166 (40%)
Mapped to 500 bp on the 3′[d]	378 (8%)	336 (9%)	122 (7%)	78 (6%)	119 (29%)
Mapped to intron sequences	284 (6%)	170 (5%)	40 (2%)	46 (4%)	74 (18%)
Mapped to transposon db	54 (1%)	36 (1%)	7 (0.4%)	8 (0.7%)	20 (5%)
Mapped to non-coding RNA	16 (0.3%)	25 (0.7%)	7 (0.4%)	5 (0.4%)	9 (2%)

[a] EST had at least 90% identity with a sequence in the database over at least 50% of its length.
[b] Mapping with decreased stringency (no identity cutoff).
[c] Transcripts from protein-coding genes.
[d] Flanking sequences from the 3′-most base of annotated gene sequence.

ments. The under-representation oflong fragments is probably caused by the inefficiency of the emulsion PCR for long PCR products.

Nebulization success

The undesirable effect of this apparent size bias in the 454 sequencing could be overcome if every transcript had a similar distribution of fragment sizes. Thus, randomly breaking cDNA fragments should overcome the size bias, as it affects all transcripts similarly. Shearing of DNA fragments by high-pressure nitrogen (nebulization) is frequently used to produce short DNA fragments for sequencing (Surzycki 2000). For expression profiling, it is essential that this procedure work for different cDNAs with the same efficiency.

We tested for a potential effect of cDNA length on nebulization efficiency. As for the DIG library, we estimated the 3′ cDNA fragment size by extending the aligned 454 sequencing ESTs to the 3′ end of the transcript and compared the distribution of the inferred fragment sizes among different cDNA length classes. Despite covering a wide range of size classes, we found the mean size of the nebulized cDNA fragments to be very similar among cDNAs of different length (Fig. 3). We further scrutinized the nebulization pattern by analyzing highly expressed genes for which at least 30 sequences were available. Genes that are not spliced and that have similar transcript lengths were found to have similar fragment sizes (data not shown). This observation suggests that there is no apparent effect of the DNA sequence on the nebulization procedure, but more data are required to corroborate this. Nevertheless, even if nebulization were to cause some differences between cDNA fragments, they may not translate into a biased measurement of gene expression due to the relatively broad size range for which 454 sequencing quantitatively operates.

Cross-method consistency

The above analyses suggested that 454 sequencing could be used for expression profiling when cDNAs are nebulized. To further validate this approach, we compared the results of the nebulized cDNAs to those obtained from cDNAs treated with restriction enzymes. When the expression levels of nebulized library were compared with the different digested libraries, the correlation coefficients ranged from 0.71 to 0.77 (Table 2). A recent study comparing the reproducibility for different microarray platforms reported intra-platform correlation coefficients ranging from 0.68 to 0.95 (Kuo et al. 2006). Thus, despite the apparent under-representation of short and long fragments in the digested libraries, the correlation coefficients fall within the range of correlation coefficients reported for microarrays. The correlation coefficients were markedly lower for the fragments longer than 300 bp (0.52–0.57), reflecting the under-representation of long fragments. Not only were transcripts represented by long fragments missed (Fig. 2), but read counts for long fragments were also extremely low (data not shown). If we limit our correlation analysis to those cDNAs, which result in fragments not suffering from an under-representation (80–300 bp), the correlation coefficients improved profoundly (0.83–0.91). Interestingly, we observed similar correlation coefficients for those cDNAs that resulted in fragments smaller than 80 bp (Table 2), suggesting that the purification step affected all cDNA fragments in this size class to a similar extent. Similar trends were observed when the size thresholds were varied by 10 or 20 bp (data not shown). As the nebulized library showed a high correlation coefficient with each of the three different restriction libraries, our results strongly indicate that the nebulization procedure is highly suitable to provide a reliable measurement of gene expression. Furthermore, the high correlation coefficients also suggest that 454 sequencing expression analysis is as reproducible as microarray experiments.

Discussion

Despite the large potential of 454 sequencing technology for transcriptome analysis, so far only a limited number of approaches using this technique have been published. One approach involved a modification of the paired-end ditagging (PET) technique (Ng et al. 2006). In this technique, 5′ and 3′ signatures of ~20 bp of each full-length transcript are simultaneously extracted and covalently-linked into the paired-end ditag. In the second approach, DeepSAGE (Nielsen et al. 2006), 21 bp are sequenced from each transcript. Both approaches greatly benefit from the high-throughput of 454 sequencing technology, but they still require the cloning of cDNAs to generate the tags/paired ditags. Given that some cDNA sequences are potentially refractory to cloning, this cloning step could introduce a bias. Furthermore, both techniques require the presence of a NlaIII recognition site. An in silico digestion of the *D. melanogaster* transcripts indicated that 4% of the sequences did not have a NlaIII recognition site. These transcripts would be entirely missed in both methods. If one considers that cDNA synthesis of long transcripts is less effective, the number of under-represented transcripts increases even more. For example, 6% of the NlaIII recognition sites required for SAGE are found >800 bp away from the poly(A) tail. Even stronger is the effect of incomplete cDNA synthesis for the paired ditag method, as this requires full-length cDNA synthesis. Furthermore, the dependence on restriction enzymes makes both methods sensitive to intraspecific polymorphism, which could generate/destroy restriction sizes. At the very extreme, this may result in a loss of the transcript due to the absence of the NlaIII recognition site. Finally, 454 sequencing has a higher error rate than does Sanger sequencing, which results in a reduced tag-to-gene mapping efficiency of short transcripts.

In our proof-of-principle study using *D. melanogaster*, we showed that the sequencing of randomly sheared 3′ cDNA provides a very good alternative to the previously suggested approaches for expression profiling using 454 sequencing technology. While it would be also possible to sequence full-length cDNAs (Bainbridge et al. 2006; Emrich et al. 2007; Weber et al. 2007) rather than 3′ ends, we consider our approach more cost-effective, as it requires only a single read per transcript. Furthermore, no adjustment for transcript length needs to be made.

As in a previous report using another technique of massively parallel sequencing (Chen and Rattray 2006), we found a severe bias against long fragments, possibly due to the inefficiency of the PCR amplification of long fragments. Furthermore, we showed that shorter fragments are also strongly under-represented. Hence, it is absolutely mandatory that the length distribution of the cDNA fragments generated by shearing be similar among transcripts. We found that shearing of cDNA molecules using high-pressure nitrogen (nebulization) results in a very similar distribution of sheared fragments among cDNA size

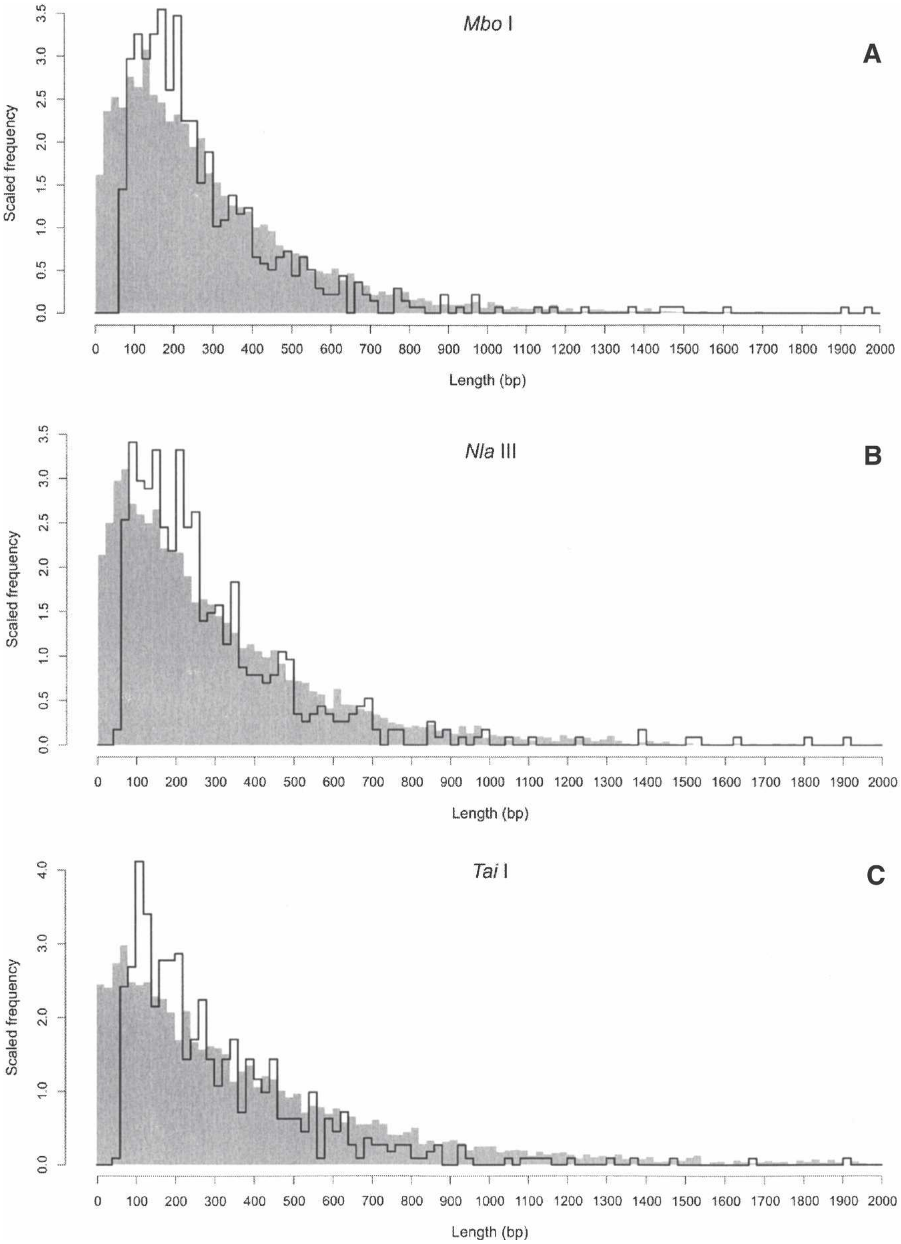

Figure 2. Under-representation of short and long 3′ cDNA fragments in 454 sequencing reads. The frequency distribution of 3′ cDNA fragment lengths obtained from in silico digestion of all *D. melanogaster* transcripts (release 5.1.) is shown in gray. The black line indicates the frequency distribution of 3′ cDNAs obtained from 454 sequencing reads. Independently of the actual counts obtained by the 454 sequencing, each transcript was considered only once. To compare the two datasets that are on different scales, the number of fragments in each class was divided by their root mean square (Becker et al. 1988). After scaling, both samples had a mean of zero and a standard deviation of one. Regardless of which restriction enzyme was used, we noted a pronounced under-representation of short (< ~80 bp) and long (> ~300 bp) fragments.

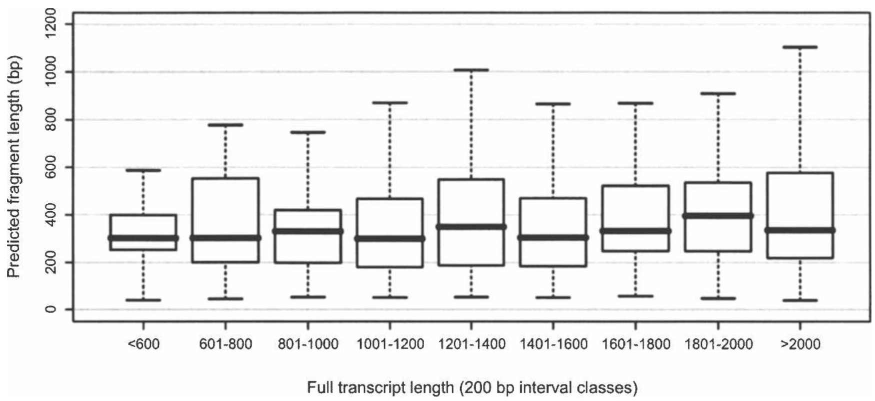

Figure 3. Length distribution of 3′ cDNA fragments after nebulization among different size classes of full-length transcripts (as inferred from the available genome annotation). The bold line indicates the median. The lower hinge gives the 25% quantile, and the upper hinge the 75% quantile. Whiskers (dashed lines) extend to the maximum and minimum sizes. Outliers are not shown.

classes and concluded that the fragmentation of the cDNAs during nebulization did not introduce a major bias in the representation of transcripts. Interestingly, recent studies also assessed the randomness of nebulization and found more reads in the 5′ end of the transcript (Bainbridge et al. 2006; Emrich et al. 2007; Weber et al. 2007). Although this bias was not very strong, our focus on the 3′ ends of the transcripts has probably resulted in an even higher homogeneity of the size distribution of the cDNA fragments analyzed. Consistent with the absence of a bias introduced by the nebulization process, our comparison of transcription profiles generated by nebulization and restriction fragments resulted in correlation coefficients that are similar to those that have been observed in intraplatform comparisons of microarray performance (Kuo et al. 2006). Thus, we conclude that 454 sequencing-based expression profiling is highly reproducible and that no strong bias is introduced by nebulization.

One difference of the 454 sequencing technology to other massively parallel sequencing techniques is the generation of longer read lengths. We evaluated the effect of read length on the mapping efficiency by truncating the obtained 454 reads to 20, 50, and 100 bp. Short read lengths result in many HSPs with scores very similar to the best one. About 20% of the 20-bp fragments had at least two perfect matches in the *D. melanogaster* genome (Fig. 4), whereas 50- and 100-bp fragments had substantially increased mapping accuracies, resulting in only 3% and 0.5% ambiguously mapped fragments, respectively. Furthermore, the difference in bit scores between the best and second-best hits is much more pronounced for longer fragments. Hence, as expected, longer fragments result in a higher proportion of unambiguously mapped sequences. In the presence of sequence polymorphism, the mapping of short sequence reads to the corresponding genes becomes an even more challenging task and introduces considerable uncertainty. For well-annotated genomes, restricting the analysis to the transcriptome rather than the genome can compensate to some extent for the low mapping accuracy of shorter sequence reads. This is a widely used approach for SAGE analyses, but for poorly annotated genomes this strategy is not efficient. For example, a recent gene expression study in *D. pseudoobscura* could map only 27% of the SAGE tags to transcripts (Metta et al. 2006). Hence, while massively parallel sequencing holds enormous potential for cross-species comparison of gene expression, our analyses also showed that sufficient read length is essential to ensure reliable identification of the corresponding transcript. Furthermore, our analyses also showed that, even for a well-studied species, such as *D. melanogaster*, we could identify new isoforms, UTRs, and antisense transcripts. Due to ability to identify SNPs in transcripts, we anticipate that this method will be also extremely powerful to measure allele-specific gene expression.

Table 2. Consistency between libraries

Fragment length range considered	MboI	NlaIII	TaiI
Full data set	**0.75**	**0.77**	**0.71**
<80 bp	0.89	0.94	0.93
80–300 bp	0.83	0.91	0.83
>300 bp	0.67	0.52	0.62

Correlation coefficients were calculated between the tag counts from the DIG library (as a whole and sorted) and counts from the NEB library.

Methods

RNA isolation and cDNA synthesis

The *D. melanogaster* genome strain (*y; cn bw sp*) was obtained from the Tucson Stock Center (stock no. 14021-0231.36). Flies were grown at 20°C in standard cornmeal medium. Total RNA was extracted with TRIzol (Invitrogen) from 30 virgin females aged 3–7 d (three replicates of 10 flies each). RNA samples were treated with DNase I (10 units/50 µg of total RNA), and absence of contaminating genomic DNA was confirmed by PCR of two low-expressed genes *CG11053* and *CG13272* (Supplemental Table 1) from total RNA. Furthermore, the primers were chosen such that the amplicon includes an intronic sequence, which permits the identification of genomic DNA contamination.

First-strand cDNA was generated from ~5 µg of total RNA using the RevertAid H Minus First Strand cDNA Synthesis Kit (Fermentas) according to manufacturer's instructions. The synthesis was carried out using a biotinylated oligo(dT) fused to the 454 sequencing primer B (5′-biotin-GCCTTGCCAGCCC GCTCAG$(T)_{17}$V-3′, where V stands for any base but T). Double-stranded cDNA was synthesized by addition of 30 U of *Escherichia coli* DNA polymerase I and 1 U of *E. coli* ribonuclease H to the first-strand synthesis reaction, following the manufacturer's (Fermentas) suggested protocol for second-strand cDNA synthesis.

Enzyme library preparation

Methods for the library preparation were based on previously described SAGE (Velculescu et al. 1995) and GLGI (Generation of Longer cDNA fragments for Gene Identification; Chen et al. 2002) protocols. Double-stranded cDNA was digested separately with the following restriction endonucleases: MboI (Sau3AI, DpnII), NlaIII, and TaiI. 3′ Fragments were recovered using M-270 Streptavidin beads (Dynal). After selection, specific linkers for each enzyme were ligated to the 3′ fragments. The linkers consisted of the double-stranded 454 sequencing primer A (5′-

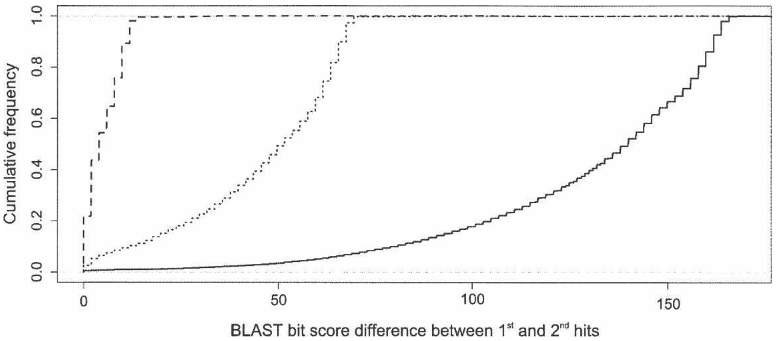

Figure 4. Cumulative distribution of the difference in BLAST bit scores of the best and second-best hits. The dashed, dotted, and solid lines show the cumulative distribution of 20, 50, and 100 bp, respectively. The plots are based on the 454 sequencing reads that provided at least 100 bp sequence. The BLAST searches were performed by using the 5′-most 20, 50, and 100 bp. The BLAST search was performed against the *D. melanogaster* genomic sequence without filtering regions of low complexity.

GCCTCCCTCGCGCCATCAG-3′) and a four-base overhang complementary to the enzyme restriction site (Supplemental Table 1). The three enzyme libraries were pooled before sequencing.

Shotgun library preparation

Approximately 5 µg of double-stranded cDNA was nebulized following previously described methods (Margulies et al. 2005) using 3 bar of nitrogen for 1 min. 3′ Nebulized fragments were recovered using M-270 Streptavidin beads (Dynal), blunt-ended with T4 DNA polymerase and ligated to the double-stranded linker A.

454 sequencing

To reduce the technical error, two libraries were produced with each methodology and pooled prior to sequencing. The libraries were purified and analyzed on a BioAnalyzer DNA 1000 LabChip to determine the concentration and quality of the fragmented cDNA. Sequencing was performed on a Genome Sequencer GS20 Instrument (Roche Diagnostics) following standard protocols (Margulies et al. 2005).

Bioinformatics

In-house Perl scripts (available on request) for the automated analysis of the 454-ESTs were used to (1) trim low-quality sequences, (2) remove poly(A) tails, (3) sort reads derived from the enzyme library into three different sample sets based on the 5′-most restriction site, and (4) map the ESTs to annotated transcripts and genome. We used as a reference data set the 5.1 release of transcript annotations and genomic sequence of *D. melanogaster*. Mapping of the ESTs to the available databases was undertaken using BLAST. The *E*-value cutoff was set at 1 × 10^{-10}, and only reads with ≥90% identity with a sequence in the database ≥50% of their length were considered. Statistical analyses were carried out using the statistical programming language R (R Development Core Team 2007).

Acknowledgments

We thank three anonymous reviewers for insightful comments. We also thank the C.S. laboratory members for helpful comments on earlier versions of this manuscript and S. Glinka for 454 sequencing. This work was supported by Fonds zur Förderung der wissenschaftlichen Forschung grants to C.S. and a fellowship of the Brazilian National Council for Scientific and Technological Development (CNPq) to T.T.T.

References

Bainbridge, M.N., Warren, R.L., Hirst, M., Romanuik, T., Zeng, T., Go, A., Delaney, A., Griffith, M., Hickenbotham, M., Magrini, V., et al. 2006. Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics* **7:** 246. doi: 10.1186/1471-2164-7-246.

Becker, R.A., Chambers, J.M., and Wilks, A.R. 1988. *The new S language: A programming environment for data analysis and graphics*. Wadsworth & Brooks/Cole, Pacific Grove, CA.

Chen, J., Lee, S., Zhou, G., and Wang, S.M. 2002. High-throughput GLGI procedure for converting a large number of serial analysis of gene expression tag sequences into 3′ complementary DNAs. *Genes Chromosomes Cancer* **33:** 252–261.

Chen, J. and Rattray, M. 2006. Analysis of tag-position bias in MPSS technology. *BMC Genomics* **7:** 77. doi: 10.1186/1471-2164-7-77.

Emrich, S.J., Barbazuk, W.B., Li, L., and Schnable, P.S. 2007. Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Res.* **17:** 69–73.

Gilad, Y., Oshlack, A., Smyth, G.K., Speed, T.P., and White, K.P. 2006. Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature* **440:** 242–245.

Gilad, Y., Rifkin, S.A., Bertone, P., Gerstein, M., and White, K.P. 2005. Multi-species microarrays reveal the effect of sequence divergence on gene expression profiles. *Genome Res.* **15:** 674–680.

Kuo, W.P., Liu, F., Trimarchi, J., Punzo, C., Lombardi, M., Sarang, J., Whipple, M.E., Maysuria, M., Serikawa, K., Lee, S.Y., et al. 2006. A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies. *Nat. Biotechnol.* **24:** 832–840.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z.T., et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376–380.

Metta, M., Gudavalli, R., Gibert, J.M., and Schlotterer, C. 2006. No accelerated rate of protein evolution in male-biased *Drosophila pseudoobscura* genes. *Genetics* **174:** 411–420.

Ng, P., Tan, J.J.S., Ooi, H.S., Lee, Y.L., Chiu, K.P., Fullwood, M.J., Srinivasan, K.G., Perbost, C., Du, L., Sung, W.K., et al. 2006. Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes. *Nucleic Acids Res.* **34:** e84. doi: 10.1093/nar/gkl444.

Nielsen, K.L., Hogh, A.L., and Emmersen, J. 2006. DeepSAGE—Digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples. *Nucleic Acids Res.* **34:** e133. doi: 10.1093/nar/gkl714.

R Development Core Team. 2007. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Surzycki, S.J. 2000. *Basic methods in molecular biology*. Springer-Verlag, New York.

Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. 1995. Serial analysis of gene expression. *Science* **270:** 484–487.

Weber, A.P., Weber, K.L., Carr, K., Wilkerson, C., and Ohlrogge, J.B. 2007. Sampling the *Arabidopsis* transcriptome with massively parallel pyrosequencing. *Plant Physiol.* **144:** 32–42.

Received August 1, 2007; accepted in revised form October 3, 2007.