nature biotechnology

Quantification of the yeast transcriptome by single-molecule sequencing

Doron Lipson[1,2], Tal Raz[1,2], Alix Kieu[1], Daniel R Jones[1], Eldar Giladi[1], Edward Thayer[1], John F Thompson[1], Stan Letovsky[1], Patrice Milos[1] & Marie Causey[1]

We present single-molecule sequencing digital gene expression (smsDGE), a high-throughput, amplification-free method for accurate quantification of the full range of cellular polyadenylated RNA transcripts using a Helicos Genetic Analysis system. smsDGE involves a reverse-transcription and polyA-tailing sample preparation procedure followed by sequencing that generates a single read per transcript. We applied smsDGE to the transcriptome of *Saccharomyces cerevisiae* strain DBY746, using 6 of the available 50 channels in a single sequencing run, yielding on average 12 million aligned reads per channel. Using spiked-in RNA, accurate quantitative measurements were obtained over four orders of magnitude. High correlation was demonstrated across independent flow-cell channels, instrument runs and sample preparations. Transcript counting in smsDGE is highly efficient due to the representation of each transcript molecule by a single read. This efficiency, coupled with the high throughput enabled by the single-molecule sequencing platform, provides an alternative method for expression profiling.

Analysis of gene expression has been a primary tool in the study of cellular mechanisms. Large-scale sequencing of cDNA clones and comparisons of transcript abundance between samples have provided valuable insights into the gene content and tissue-specific and developmental expression patterns of a wide range of organisms. More recently, microarray expression profiling has provided gene expression information at relatively low cost and increased throughput[1,2]. Although microarrays are now widely used for monitoring transcript expression, hybridization-based technologies have several important limitations[2]. First, low-abundance transcripts cannot be measured accurately. Second, discovery of novel transcripts is limited. Third, direct comparison of transcripts within an individual sample is inaccurate because hybridization kinetics for individual mRNAs are sequence dependent, necessitating ratiometric comparison between paired samples.

To overcome these limitations, researchers have developed digital gene expression (DGE) technologies, such as serial analysis of gene expression (SAGE) and massively parallel signature sequencing (MPSS)[3–8], with the goal of discovering new transcripts as well as performing complete transcriptome profiling over the full dynamic range of cellular mRNA expression. In general, DGE methods use high-throughput sequencing of short cDNA fragments (tags) that are matched to a reference transcriptome to identify the corresponding gene. Individual transcript abundances are then inferred from the relative tag counts for each gene in a 'digital' manner, in contrast to the 'analog' nature of microarray intensity-based quantification. To date, most SAGE-like strategies rely on restriction digestion, adaptor ligation and additional steps. This extensive sample manipulation and the generation of tags from limited sequence contexts per transcript are likely sources of transcript quantification biases[9–13].

Recent studies have demonstrated that high-throughput short-read sequencing platforms can be used to generate high-resolution maps of complete transcriptomes by sequencing a significant fraction of the transcriptome at depth[14–16]. Because these 'RNA-Seq' methods generate variable numbers of reads from each mRNA molecule, extraction of quantitative measurements requires an assessment of coverage depth for each transcript. Although this approach yields informative transcript quantification, it is costly in terms of the sheer number of reads that are required to completely cover an entire transcriptome (several tens of millions of reads per sample). Because the number of reads generated from each transcript is dependent on its length, additional normalization steps are required[15], and quantification accuracy for shorter transcripts is lower[17].

We present a DGE technology based on single-molecule sequencing[18]. Because no amplification is employed, sample preparation does not involve adding adaptors to cDNA, thus enabling a simple procedure free of restriction digestion, ligation or amplification steps. This methodology generates strand-specific, accurate transcript counts covering the complete cellular dynamic range. Single-molecule sequencing DGE (smsDGE) is optimized for mRNA quantification rather than full transcriptome sequencing. The effectiveness of counting by smsDGE is driven by the fact that only a single read is generated from each cDNA molecule, thereby maintaining a faithful representation of transcript distribution. In contrast to RNA-Seq, quantification is independent of transcript length, and sequence read counts are directly proportional to transcript abundance. smsDGE generates sequence reads from the 3′ ends of first-strand cDNA molecules, which usually corresponds to the 5′ ends of mRNAs, depending on the completeness of the reverse transcription (**Fig. 1a**). It does not require

[1]Helicos Biosciences Corporation, Cambridge, Massachusetts, USA. [2]These authors contributed equally to this work. Correspondence should be addressed to T.R. (traz@helicosbio.com).

Received 30 March; accepted 9 June; published online 5 July 2009; doi:10.1038/nbt.1551

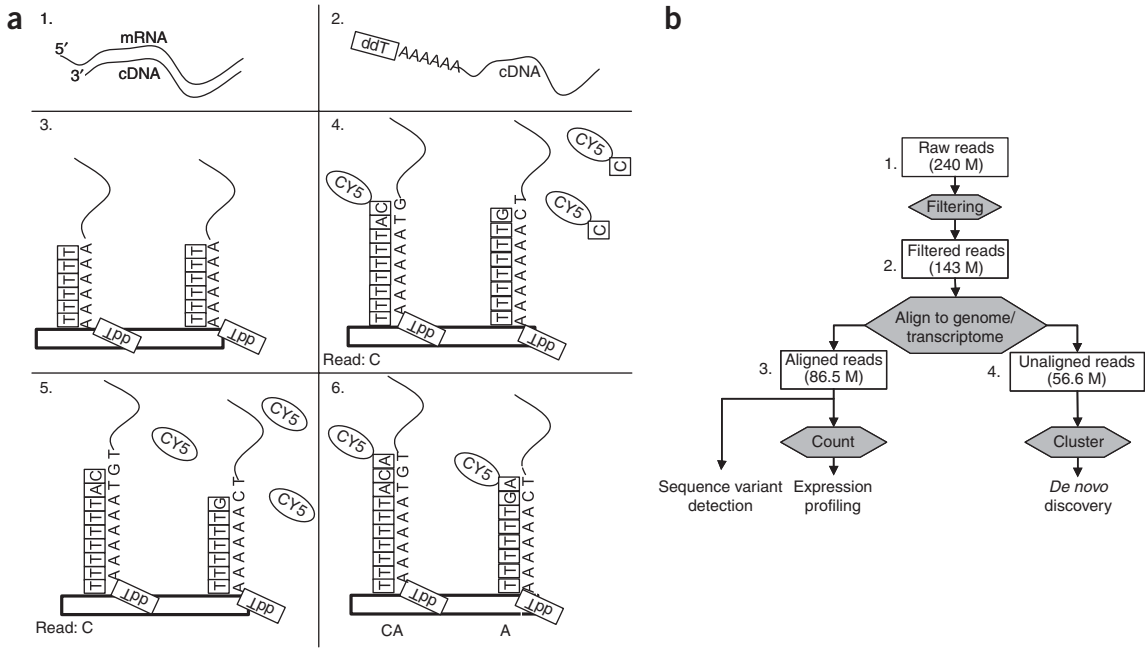

Figure 1 Sample preparation, sequencing and analysis workflow. (**a**) Sample preparation and sequencing. (1) Prepare first-strand cDNA from *S. cerevisiae* mRNA and an oligo-dU primer. (2) Add 3′ tail of dATP followed by dideoxy-TTP (ddT) blocking. Remove RNA and oligo dU sequence used for the cDNA synthesis using RNase enzymes and the USER reagent. (3) Hybridize tailed sample to poly-dT oligonucleotides covalently attached to the flow-cell's channel surface. Surface oligonucleotides act as primers for the sequencing reaction. (4) Sequence a single base by adding a Cy5-labeled nucleotide, washing the chemistry away and imaging the flow-cell channel. (5) Cleave off Cy5 dye label and wash it away. (6) Add and image next nucleotide. (**b**) Data analysis workflow. (1) Filter raw sequencing reads by length and sequence context. (2) Align filtered reads to the SGD yeast transcriptome and genome reference libraries. (3) Count transcripts using aligned reads, and generate transcript sequence information. (4) Cluster unaligned reads to identify unannotated transcripts. M, million.

the cDNA to be full length, and therefore should work well with short cDNAs generated by incomplete reverse transcription or partial mRNA degradation.

smsDGE involves the hybridization of 3′ poly-A tailed, first-strand cDNA molecules to oligonucleotide primers attached to the flow-cell surface. The cDNA is sequenced by single-molecule imaging of the stepwise addition of fluorescently labeled nucleotides into the surface-captured strand. The sequencing reaction does not require any amplification steps, allowing strands to be densely packed resulting in extremely high throughput (tens of millions of strands per channel). To demonstrate the application of smsDGE, we report the profiling of the *S. cerevisiae* DBY746 transcriptome.

RESULTS

Sample preparation and sequencing

Sequencing was performed on a Helicos Genetic Analysis system whose basic principles have been described[18]. This system allows separate sequencing reactions to take place in 50 channels, thereby enabling 50 samples to be sequenced in parallel (**Fig. 1a**). mRNA from *S. cerevisiae* strain DBY746 was used for first-strand cDNA synthesis and a poly-dA tail was added to the 3′ end of the single-stranded cDNA. The sample was then hybridized to poly-dT oligonucleotides covalently attached to the flow-cell surface. This allows the attached oligonucleotides to be used as primers for the subsequent sequencing-by-synthesis reaction. Fluorescently labeled Virtual Terminator (VT) nucleotides[19] are incorporated only once each cycle onto the growing strand. The surface is illuminated by laser and imaged to record nucleotide incorporations at each DNA strand location. The serial incorporation and imaging of four nucleotides is termed a 'quad-cycle'; 30 quad-cycles were used for the transcriptome profiling described here. Two independently prepared samples from a single source of mRNA were run in three separate flow-cell channels each.

Data and alignment

The data analysis workflow is outlined in **Figure 1b**. An initial 240 million raw reads were collected from six channels of a single run. Filtering by length and sequence complexity (**Supplementary Methods**) yielded a final count of 143 million reads of 24–60 nt (loss is mostly attributable to the minimum length criteria). Reads were aligned to a *S. cerevisiae* genome reference and to a transcriptome reference library consisting of single-stranded 5′ untranslated region (UTR) and open reading frame (ORF) sequences of 6,719 verified, uncharacterized and dubious ORFs from the *Saccharomyces* Genome Database (SGD)[20]. Short-read alignment was performed at a stringent threshold using a Smith-Waterman–based alignment algorithm, which is tolerant of indel errors. In total, 86 million (60%) of the filtered reads could be stringently mapped to the yeast genome, and 78 million (55%) to at least one yeast transcript. The high fraction of genome-aligned reads mapping to the transcriptome (91%) is indicative of the relative completeness of the yeast transcript annotation, where the remaining 9% of genome-mapped reads are attributable to reads derived from unannotated transcripts, noncoding RNAs and spurious reverse strands of known transcripts. Over 99% of reads were 24–50 nt in length with a median length of 33 nt (**Fig. 2a**). The average error rates were in the range of 4.4–4.8% per base across the six channels. The set of reads generated is provided within the supplement.

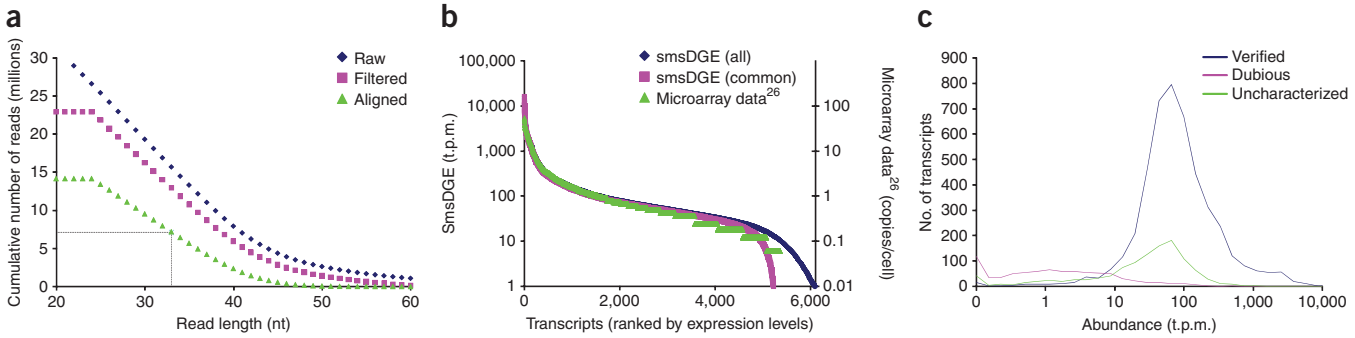

Figure 2 Data description. (**a**) Read-length distribution. Distributions of raw reads (blue), filtered reads (pink) and transcriptome-aligned reads (green) in a single channel. Median length of aligned reads is 33 nt. (**b**) Transcript abundance profile. Comparison of smsDGE to previously published transcript distribution[25] by oligonucleotide array measurement. For each set, transcripts are ranked by abundance on the *x* axis. smsDGE counts (left axis) are depicted for both the entire reference set (6,711 transcripts), as well as for the subset of genes that are common to both studies (5,460 transcripts). Estimated number of copies per cell given for array data[25] (right axis). (**c**) Transcript abundance by type. Comparing transcript abundance levels to transcript type, as annotated in SGD.

Transcript counting

Methods for estimating transcript distributions from short-tag sequencing data typically assign each read to a single unique transcript. However, owing to the occurrence of natural transcript sequence homologies, sequence variation and read errors, unique assignments based on maximal alignment scores may lead to miscounting, as assignments may be ambiguous or incorrect. A method for assigning reads that match equally well to several sites ('multireads') has been reported[15], but does not account for suboptimal-scoring alignments, which can be especially important when considering transcripts of radically different abundances. Read misassignment to abundant transcripts will not substantially skew transcript counts. However, low abundance (or nonexistent) transcripts will be overcounted.

To achieve maximal assay specificity, we used read misassignment–corrected counting (**Supplementary Methods**), a probability-based method, for assignment of reads to transcripts. Briefly, suboptimal alignments between each read and the entire reference library are considered, and the probability of assignment of each read to each transcript is assessed based on both the alignment score and the transcript abundance. The latter value is estimated iteratively based on an initial assessment. Ambiguously aligned reads are distributed among transcripts based on their assessed abundances. Final counts assigned to each transcript are reported as transcripts per million (t.p.m.). Because only one read is generated per transcript molecule, transcript length normalization is not needed. Transcripts per million from DGE should be directly comparable to length-normalized counts from RNA-Seq.

The smsDGE profile of the *S. cerevisiae* transcriptome is depicted in **Figure 2b**. We measured 6,086 transcripts of the 6,711 putative ORFs in the reference set at an abundance of 1–16,000 t.p.m., and 5,376 at >10 t.p.m. (**Supplementary Table 1**). This profile demonstrates high agreement with a transcript level profile previously measured for 5,460 genes using oligonucleotide arrays[15]. smsDGE transcript counts span at least four orders of magnitude (0.01–100 transcripts per cell[15]) with higher resolution of low abundance transcripts (<10 t.p.m.) than was demonstrated in the microarray study. The remaining 625 ORFs in the reference set were detected at <1 t.p.m., signifying no or extremely low expression. Among these are 393 ORFs annotated in SGD as "dubious," and only 62 ORFs annotated as "verified." This infrequent detection of dubious transcripts provides additional confirmation of the high specificity attainable by this method (**Fig. 2c**).

To demonstrate accurate quantification of low abundance transcripts and to assess the dynamic range of transcript detection, we serially diluted five synthetically generated RNAs across four orders of magnitude, and mixed them with two yeast samples. Each was then

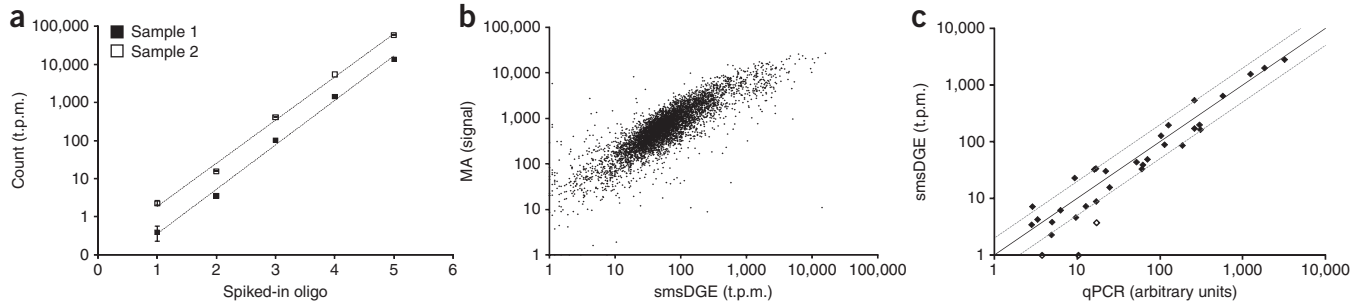

Figure 3 Reproducibility and counting accuracy. (**a**) Measurement of spiked-in RNA. Normalized counts of five spiked-in RNAs in two separate yeast samples. To demonstrate accurate quantification of low abundance transcripts and to assess the dynamic range of transcript detection, we synthetically generated five RNAs serially diluted in the range of 50 ng–5 pg (that is, spanning four orders of magnitude), and mixed with 1 μg of *S. cerevisiae* poly-A selected RNA (sample 1). An additional RNA sample was prepared by mixing a 5× dilution of the spiked-in mix with the same *S. cerevisiae* RNA (sample 2). Each RNA sample was then prepared separately and sequenced in three channels. Error bars denote s.d. in the three channels (only noticeable in lowest data point). (**b**) Comparison of smsDGE counts to measurement by microarray. Microarray signal values from a single array were compared to smsDGE counts of the same mRNA sample. Linear correlation, 0.70; rank correlation, 0.85. (**c**) Comparison of smsDGE counts to qPCR. We quantified 33 transcripts by qPCR and compared to smsDGE counts. qPCR arbitrary units were normalized by the mean. Dashed lines denote twofold difference. Linear correlation, 0.98. Open squares are outliers.

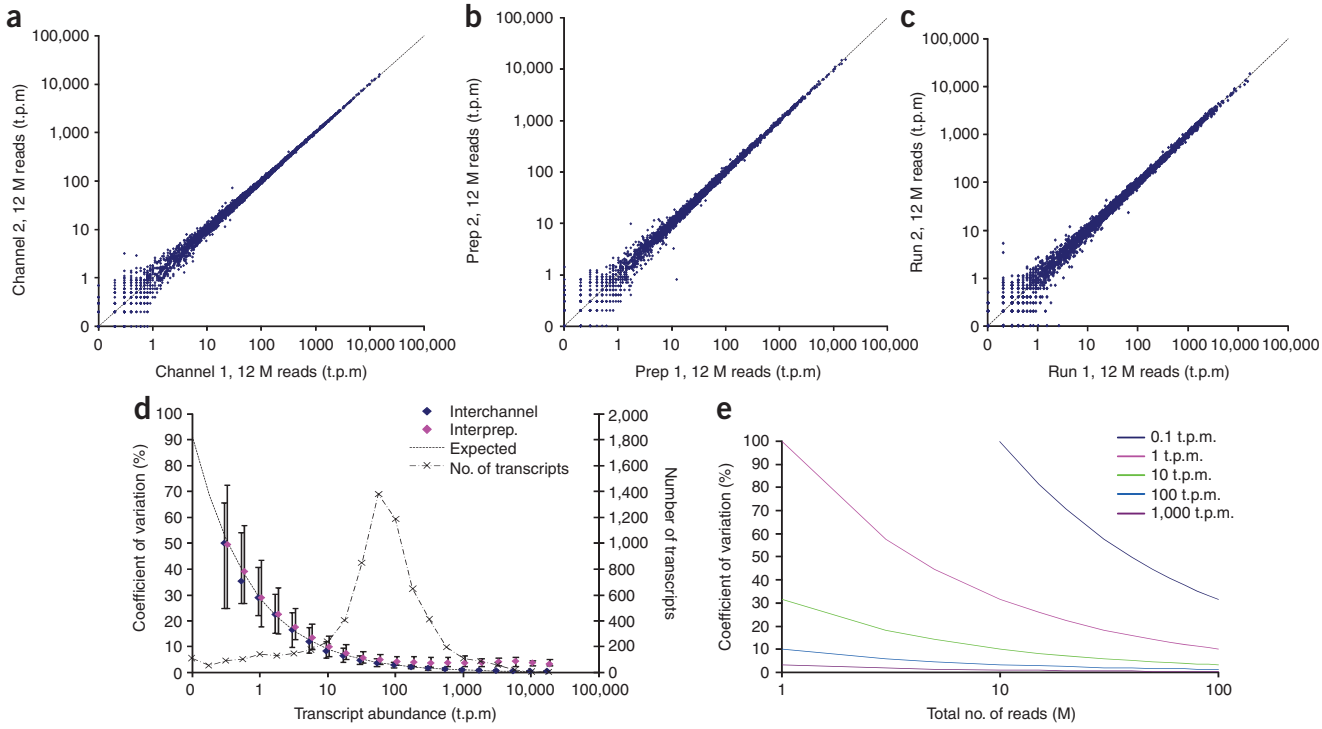

Figure 4 Count reproducibility. (**a**–**c**) Comparison of counts for all yeast transcripts between the same sample in (**a**) two different channels (Pearson correlation, $r > 0.9995$), (**b**) different sample preps of the same original mRNA ($r > 0.998$), (**c**) the same sample on two different runs ($r > 0.994$). (**d**) Observed variance in transcript abundance across three flow-cell channels containing the same (interchannel) and different preparations (interprep.) of the same mRNA. Dotted line shows the expected variance based on counting stochasticity (variance based on Poisson distribution corresponding to 12 million counts in each channel). Blue and purple dots and error bars denote median and quantile values within each bin, respectively. (**e**) Predicted variance in counts based on total number of reads and transcript abundance. M, million.

prepared separately and sequenced in three channels. Quantification of the mixed spiked-in RNAs was linear from 0.5–50,000 t.p.m. demonstrating accurate quantification within each channel with a dynamic range of four orders of magnitude (**Fig. 3a**).

We compared smsDGE counts to a microarray analysis of the identical sample (**Fig. 3b**). The array intensity signal and smsDGE measurements have an overall correlation of 0.70 (rank correlation of 0.85). smsDGE counts also had high agreement to published transcript counts in a different yeast strain assessed using RNA-Seq[15] ($r = 0.7$, **Supplementary Fig. 1**) In addition, we compared smsDGE counts to qPCR measurements of the same mRNA sample on a panel of 33 transcripts at a wide range of transcription levels (**Fig. 3c** and **Supplementary Table 2**). This comparison demonstrates a particularly high correlation ($r > 0.98$, $P < 10^{-20}$) over three orders of magnitude. Thirty out of 33 transcripts fell within a 2.5-fold range of their qPCR measurements. The three outliers were measured by smsDGE at lower levels than qPCR measurements, at low abundance levels (<4 t.p.m.). Interestingly, all outliers were found to overlap with a higher number of reads found on the opposing DNA strand, suggesting that the higher abundance measured by qPCR may be a result of the inability of qPCR to distinguish between transcripts on both strands.

Counting reproducibility

Counting results were highly correlated between different flow-cell channels for each sample (Pearson correlation, $r > 0.9995$ for all channel pairs, **Fig. 4a**). Correlation was only marginally lower between the two different sample preparations in the same run ($r > 0.998$; **Fig. 4b**), and the same sample in two separate runs ($r > 0.994$; **Fig. 4c**).

To assess counting variability across independently prepared samples, we prepared a third *S. cerevisiae* sample from the same RNA. Inter-sample variability is only slightly greater than the expected sampling stochasticity (of a Poisson sampling process), and is mostly observable at high expression levels, as high abundance transcripts have negligible sampling-based variance (**Fig. 4d**). Using 12 million reads per channel, the median coefficient of variation is 4% at 100 t.p.m., 10% at 10 t.p.m. and 30% at 1 t.p.m., with the bulk of the transcripts falling between 10 and 1,000 t.p.m. The predictability of the count variance allows us to forecast the effect of additional sequencing on counting accuracy, and to determine the minimal number of reads required to reliably detect changes in transcripts of given abundance (**Fig. 4e**).

Transcription start site mapping

This study provides the opportunity to map transcriptional start sites (TSS) of yeast genes, owing to the large number of reads sequenced from the 5′ end of complete transcripts. To allow mapping of reads to 5′ UTR regions, we included in our reference transcriptome library sequence up to 250 bp upstream of the ORF start codon. smsDGE reads are sampled from the point where reverse transcription of a strand stopped, either because it reached the 5′ end of the mRNA strand or dissociated from it before the end. In total, 55% of the reads uniquely mapped to 5′ UTR regions, 72% of which begin at the region 50 bp upstream of the ORF start codon—the assumed TSS position of most yeast transcripts[15,21,22]. As expected, owing to reverse transcriptase processivity and/or mRNA degradation, the fraction of reads reaching the 5′ end of a transcript is inversely proportional to length (**Fig. 5a**). These results, including hundreds of reads per TSS for many transcripts, could be used in future work to obtain physical maps of

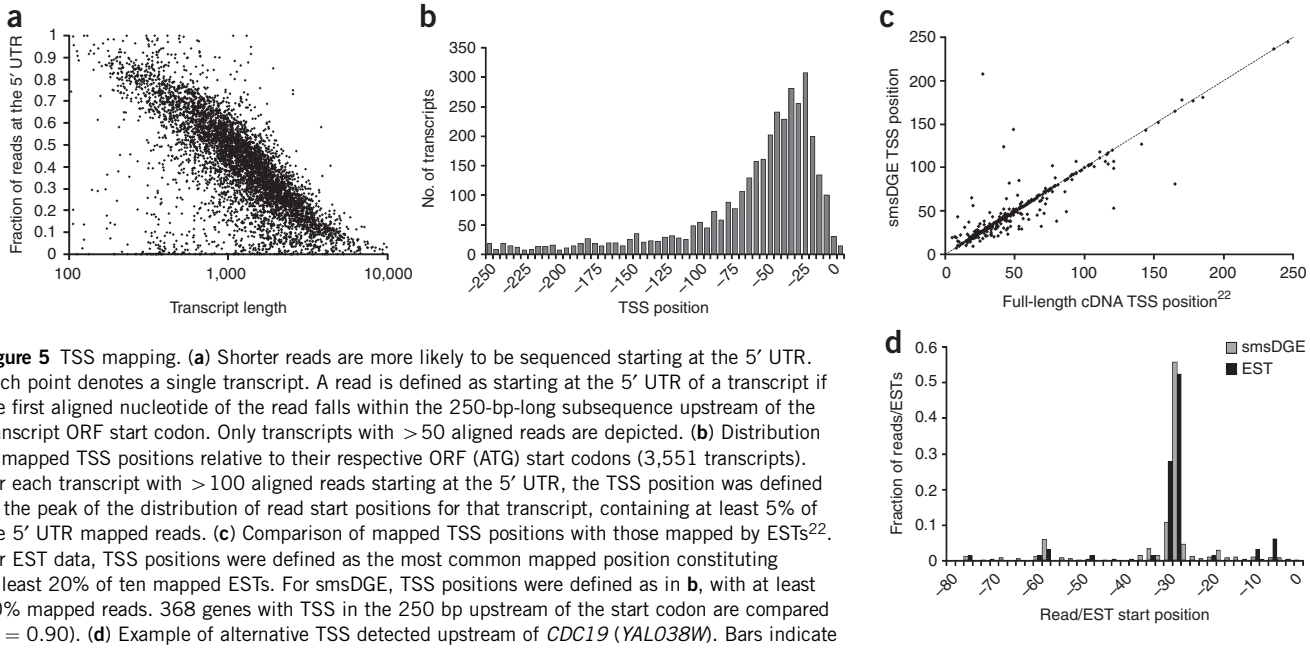

Figure 5 TSS mapping. (**a**) Shorter reads are more likely to be sequenced starting at the 5′ UTR. Each point denotes a single transcript. A read is defined as starting at the 5′ UTR of a transcript if the first aligned nucleotide of the read falls within the 250-bp-long subsequence upstream of the transcript ORF start codon. Only transcripts with >50 aligned reads are depicted. (**b**) Distribution of mapped TSS positions relative to their respective ORF (ATG) start codons (3,551 transcripts). For each transcript with >100 aligned reads starting at the 5′ UTR, the TSS position was defined as the peak of the distribution of read start positions for that transcript, containing at least 5% of the 5′ UTR mapped reads. (**c**) Comparison of mapped TSS positions with those mapped by ESTs[22]. For EST data, TSS positions were defined as the most common mapped position constituting at least 20% of ten mapped ESTs. For smsDGE, TSS positions were defined as in **b**, with at least 20% mapped reads. 368 genes with TSS in the 250 bp upstream of the start codon are compared ($R = 0.90$). (**d**) Example of alternative TSS detected upstream of *CDC19* (*YAL038W*). Bars indicate fraction of smsDGE reads (gray) or ESTs (black) starting at each position of the transcript's reference sequence. Position 0 denotes ORF start codon. 15,797 smsDGE reads were assigned to this 5′ UTR region of this transcript in comparison to 65 ESTs.

yeast TSS that are more accurate than maps derived from 5′ SAGE[21] and EST sequencing[22]. **Figure 5b** demonstrates the distribution of mapped TSS positions, relative to ORF start codons. **Figure 5c** demonstrates agreement between mapped TSS positions and those in a previous study[22], where **Figure 5d** illustrates agreement with EST data in a transcript with multiple alternative TSS positions.

Additional transcript characterization

Although the primary goal of this study was to provide accurate abundance levels of all yeast transcripts, the variability in read start sites provides a wealth of transcriptome sequence information. This variability is the result of cDNAs that were incompletely reverse transcribed. The sequence coverage of transcripts varies substantially as a result of their relative sizes and abundance levels. However, uniquely aligned reads from a single channel covered 7.6 Mbp of the ~9 Mbp of *S. cerevisiae* transcriptome coding sequence, with 4.6 Mbp covered at a depth of ≥5×, and 2.7 Mpb at ≥10× (**Supplementary Fig. 2**). Applying a single nucleotide polymorphism–discovery tool to these reads (**Supplementary Methods**) identified >3,000 single-base substitutions between the strain being sequenced (DBY746) and the strain in the reference database (S288C; **Fig. 6a**). A full list of variations is provided (**Supplementary Table 3**). Ten arbitrarily selected variations were verified by Sanger sequencing (**Supplementary Table 4**).

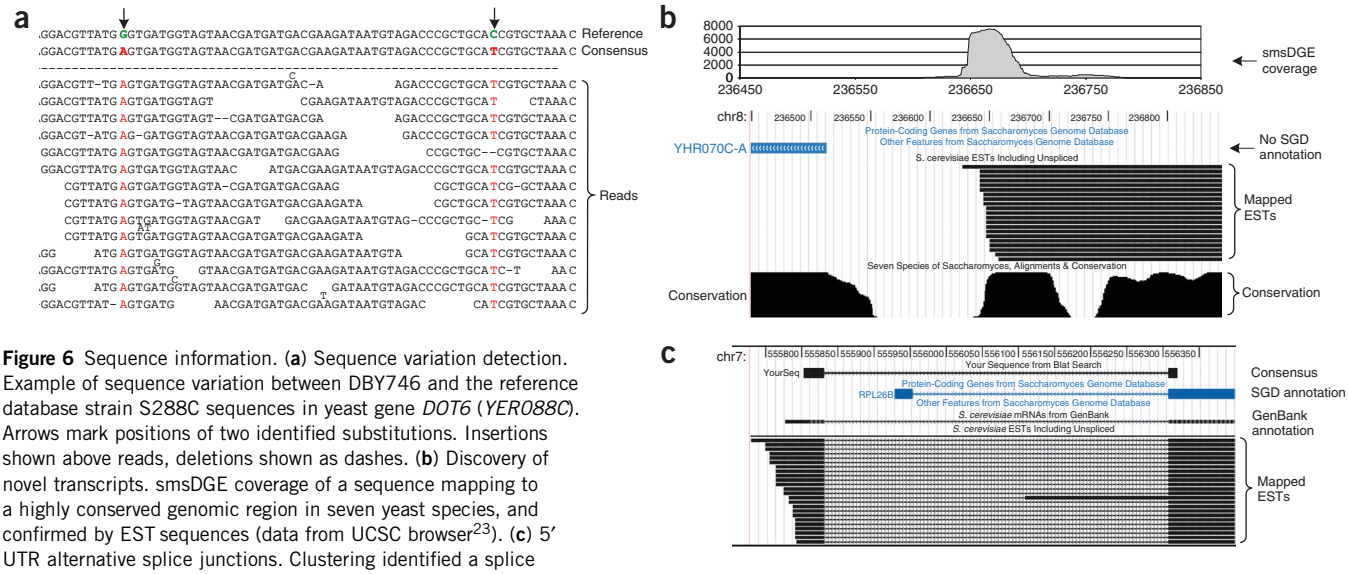

Figure 6 Sequence information. (**a**) Sequence variation detection. Example of sequence variation between DBY746 and the reference database strain S288C sequences in yeast gene *DOT6* (*YER088C*). Arrows mark positions of two identified substitutions. Insertions shown above reads, deletions shown as dashes. (**b**) Discovery of novel transcripts. smsDGE coverage of a sequence mapping to a highly conserved genomic region in seven yeast species, and confirmed by EST sequences (data from UCSC browser[23]). (**c**) 5′ UTR alternative splice junctions. Clustering identified a splice variant in *RPL26B* (*YGL189C*) that is different from the variant annotated in the SGD transcriptome database. Alternative splice variant is confirmed by GenBank and EST sequences (data from UCSC browser[23]). Consensus, smsDGE cluster consensus sequence.

To identify unknown transcripts in the sample, we also aligned our reads against the complete genome. 700 K reads from a single channel mapped to intergenic regions that are >250 bp from annotated ORFs. 370 K reads could be grouped into 1,049 peaks with expression levels ≥5 t.p.m., many of which were annotated as distant 5′ UTRs, noncoding RNAs and known ESTs. **Figure 6b** depicts one peak mapping to an unannotated genomic sequence in agreement with known ESTs[22] in a region highly conserved among yeast species[23] (**Supplementary Table 5**). An mRNA sample may include additional transcripts that cannot be aligned to the reference genome, such as contaminants, spliced or edited RNA.

To examine smsDGE for *de novo* characterization of unknown sequences, we employed a read-clustering strategy to a subset of reads poorly aligned to either the genome or transcriptome libraries. Consensus sequences (**Supplementary Table 6**) could be mapped to 5′ UTR splice junctions that were not annotated in the reference library (e.g., **Fig. 6c**), and matched sequences previously discovered by EST mapping[22] and tiling arrays[24], and to the 2-μm circle plasmid.

DISCUSSION

smsDGE is a transcriptome profiling method that utilizes the unique attributes of high-throughput single-molecule sequencing. Over 12 million usable reads (that is, ≥24 nt long and transcriptome aligned), generated in each of six channels, were used to quantify the complete range of transcripts expressed in the *S. cerevisiae* DBY746 strain. Quantification accuracy was assessed using spiked-in RNA, demonstrating accurate counts across more than four orders of magnitude to an abundance level below 1 t.p.m. using a single channel (**Fig. 3a**). Counting correlation was demonstrated across different channels, sample preparations and runs (**Fig. 4**).

Expression profiling by smsDGE overcomes many of the limitations of array-based methods. Specifically, it allows accurate quantification of a wide range of expression levels, including low abundance transcripts. It enables detection of sequence variants and generates counts that are comparable between different transcripts, sample preparations and runs. In addition, it provides the ability to discover novel transcripts by detecting reads that do not align to the known transcriptome reference. Like array-based methods, smsDGE may also be useful for mapping and identifying alternative TSSs, especially in short- to average-sized transcripts. smsDGE uses a sample preparation method free of amplification, restriction digest or ligation steps, thereby reducing biases related to preparation steps inherent in previous DGE methods such as SAGE and MPSS[9–13].

Recently, short-read sequencing technologies have been demonstrated to generate accurate measurement of gene expression by means of full transcriptome sequencing (RNA-Seq)[14–16]. RNA-Seq differs from smsDGE because multiple reads are generated from each transcript molecule, where long transcripts generate more reads in proportion to their length. The variance in transcript abundance measurements is driven mostly by read number. Whereas this variance depends only on transcript abundance in smsDGE, in RNA-Seq it is dependent on both transcript abundance and length, making short transcripts harder to count accurately[17]. We compared the number of reads per transcript in smsDGE with the expected number of reads per transcript that would be generated by RNA-Seq. It would be necessary to generate 40 million reads in RNA-Seq to get as many counts for 95% of transcripts that 10 million smsDGE reads would provide (**Supplementary Fig. 3**). A similar analysis of human transcriptome data suggests that more than five times as many reads would be needed. An additional complexity of RNA-Seq is that transcript counts must be derived by a normalization process that assumes uniform transcript coverage (e.g., RPKM[15]). smsDGE, on the other hand, uses the raw counts directly and is likely to be more accurate in the presence of 3′ biased mRNA material. An additional useful aspect of smsDGE data is that all reads are generated from single-stranded cDNA molecules and are, therefore, strand specific relative to the genome. This is especially advantageous in cases where ORFs overlap on the forward and reverse DNA strands.

In this study, the simplicity of the yeast transcriptome enabled a demonstration of the counting accuracy of smsDGE covering a large cellular dynamic range. The capacity of smsDGE to provide accurate transcript quantification for a single sample promises to simplify comparison between independently prepared and measured samples. This ability, combined with the efficiency of transcript counting and the high throughput of the SMS platform, portends cost-efficient expression profiling for large multisample studies.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Accession numbers. Short Read Archive (SRA): accession no. SRA 008810.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We thank all of the past and present colleagues at Helicos who have contributed to this work.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturebiotechnology/.

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Lockhart, D.J. & Winzeler, E.A. Genomics, gene expression and DNA arrays. *Nature* **405**, 827–836 (2000).
2. Churchill, G.A. Fundamentals of experimental design for cDNA microarrays. *Nat. Genet.* **32** Suppl, 490–495 (2002).
3. Velculescu, V.E., Zhang, L., Vogelstein, B. & Kinzler, K.W. Serial analysis of gene expression. *Science* **270**, 484–487 (1995).
4. Brenner, S. *et al.* Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* **18**, 630–634 (2000).
5. Saha, S. *et al.* Using the transcriptome to annotate the genome. *Nat. Biotechnol.* **20**, 508–512 (2002).
6. Shiraki, T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. USA* **100**, 15776–15781 (2003).
7. Hashimoto, S. *et al.* 5′-end SAGE for the analysis of transcriptional start sites. *Nat. Biotechnol.* **22**, 1146–1149 (2004).
8. Kim, J.B. *et al.* Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science* **316**, 1481–1484 (2007).
9. Chen, J. & Rattray, M. Analysis of tag-position bias in MPSS technology. *BMC Genomics* **7**, 77 (2006).
10. Siddiqui, A.S. *et al.* Sequence biases in large scale gene expression profiling data. *Nucleic Acids Res.* **34**, e83 (2006).
11. Gilchrist, M.A., Qin, H. & Zaretzki, R. Modeling SAGE tag formation and its effects on data interpretation within a Bayesian framework. *BMC Bioinformatics* **8**, 403 (2007).
12. Hene, L. *et al.* Deep analysis of cellular transcriptomes - LongSAGE versus classic MPSS. *BMC Genomics* **8**, 333 (2007).
13. So, A.P., Turner, R.F. & Haynes, C.A. Minimizing loss of sequence information in SAGE ditags by modulating the temperature dependent 3′ → 5′ exonuclease activity of DNA polymerases on 3′-terminal isoheptyl amino groups. *Biotechnol. Bioeng.* **94**, 54–65 (2006).
14. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344–1349 (2008).
15. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
16. Cloonan, N. *et al.* Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* **5**, 613–619 (2008).
17. Oshlack, A. & Wakefield, M.J. Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct* **4**, 14 (2009).

18. Harris, T.D. *et al.* Single-molecule DNA sequencing of a viral genome. *Science* **320**, 106–109 (2008).
19. Bowers, J. *et al.* Novel virtual terminator nucleotides for next generation DNA sequencing. *Nat. Methods* (in the press).
20. Fisk, D.G. *et al. Saccharomyces cerevisiae* S288C genome annotation: a working hypothesis. *Yeast* **23**, 857–865 (2006).
21. Zhang, Z. & Dietrich, F.S. Mapping of transcription start sites in *Saccharomyces cerevisiae* using 5′ SAGE. *Nucleic Acids Res.* **33**, 2838–2851 (2005).
22. Miura, F. *et al.* A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. *Proc. Natl. Acad. Sci. USA* **103**, 17846–17851 (2006).
23. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
24. Juneau, K., Palm, C., Miranda, M. & Davis, R.W. High-density yeast-tiling array reveals previously undiscovered introns and extensive regulation of meiotic splicing. *Proc. Natl. Acad. Sci. USA* **104**, 1522–1527 (2007).
25. Holstege, F.C. *et al.* Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**, 717–728 (1998).

ONLINE METHODS

cDNA preparation. mRNA from *S. cerevisiae* strain DBY746 (his3Δ1 leu2-3 leu2-112 ura3-52 trp1-289), grown under standard conditions (yeast peptone dextrose, 30 °C) was obtained from Clontech. We mixed 1 μg *S. cerevisiae* RNA with 6 *in vitro* transcribed *Arabidopsis thaliana* RNAs at 40 ng to 400 fg as described in **Figure 3** legend (Stratagene, Agilent Technologies). In addition three assay replicates were prepared independently from the same RNA for assay reproducibility studies. We used 1 to 2 μg yeast poly A selected RNA to make first-strand cDNA. First-strand cDNA was prepared using a SuperScript III first-strand cDNA synthesis kit (Invitrogen) according to manufacturer's instructions except that 5 μM of a 50-nucleotide deoxyuracil primer (IDT) was used in place of the recommended primer. mRNA was removed by RNase H (Invitrogen) digestion for 20 min at 37 °C followed by removal of the deoxyuracil primer sequence by USER reagent (New England Biolabs) digestion for 20 min at 37 °C. A final incubation with RNase I (New England BioLabs) for 15 min at 37 °C was then performed to remove any remaining RNA. The sample was purified using the AMPure kit (Agencourt Biosciences) at a 1:1.8 sample to bead ratio according to manufacturer's instructions. The above preparations yielded ~500 and 1,000 ng cDNA for 1 and 2 μg preparations, respectively. 60 ng of this prepared cDNA was then poly dA tailed and loaded on 1–3 channels of the Helicos Genetic Analysis system.

Poly dA tailing. A poly dA tail of 90 ± 20 nucleotides on average was added to the 3′ end of the cDNA by terminal deoxynucleotidyl transferase (New England BioLabs). A 60 ng cDNA sample was combined with terminal deoxynucleotidyl transferase reaction buffer (potassium acetate (50 mM), Tris acetate (20 mM), magnesium acetate (10 mM), pH 7.9) $CoCl_2$ (250 μM), dATP (170 pmoles) and a control oligo used to assess the tailing efficiency (1.5 pmoles). We added 24 units terminal transferase after denaturation and snap cooling on ice, followed by 1 h incubation at 42 °C and 10 min heat inactivation at 70 °C. Tailed samples were then labeled and 3′ blocked by dideoxy TTP (600 pmoles). The sample was then denatured and snap cooled on ice, 24 units of terminal transferase were added followed by a 1 h incubation at 37 °C and a final heat inactivation step. The control oligo and excess nucleotides were removed from the sample by Ampure purification at a 1:1.3 sample to bead ratio (Agencourt Bioscience).

Template capture and sequencing. Each sequencing reaction takes place in one of 50 channels of the sequencing flow-cell. Each channel's surface is lined with a covalently attached poly-dT oligonucleotide. This surface oligonucleotide has the dual role of facilitating the template capture and priming the sequencing reaction. For capture, the cDNA template's poly-dA 3′ tail is hybridized to the poly-dT surface oligonucleotide. The sequencing reaction can then be initiated at the surface oligo's 3′ end (**Fig. 1**). To avoid sequencing the template poly-dA tail, before sequencing we use a 'fill and lock' procedure in which the surface oligo is extended against the template's 3′ poly-dA tail by a dTTP fill. dGTP, dCTP and dATP VTs are also included in the reaction to 'lock' the surface oligo against the sample template after the dTTP fill is complete.

Sequencing by synthesis is performed following the 'fill and lock' procedure by introducing one of four Cy5 labeled VT nucleotides in the presence of a polymerase reaction mix (Helicos BioSciences). Incorporated nucleotides are imaged after which the Cy5 dye is chemically cleaved off the incorporated nucleotide and rinsed away. This process is repeated for each of the next three nucleotides to complete a sequencing quad cycle. Because read growth rate varies somewhat by sequence context (resulting from the order in which bases are added in the sequencing reaction), 30 quad-cycles were used to ensure that slow growing reads could reach the threshold length. The process of sequence base calling was previously described and used here with the exception that no intensity-based homopolymer length calling was performed in this study because VT nucleotides do not run through homopolymer sequences[18].

Quantitative PCR. We selected 33 *S. cerevisiae* transcripts spanning a large range of expression levels for comparison of smsDGE counts against qPCR quantification (18 Taqman and 15 SYBR green assays; **Supplementary Table 2**). Thirteen of these 33 transcripts were selected from transcripts with smsDGE counts <10 t.p.m. to test accuracy at low abundance levels. qPCR reactions were denatured at 95 °C for 10 min followed by 40 cycles of 95 °C for 15s and 57 °C for 30s. Taqman assays had forward and reverse primers at 0.3 μM each, a Taqman probe at 0.25 μM and 1× Taqman reaction mix (Taqman universal PCR mix, Applied Biosystems). SYBR green assays had forward and reverse primer at 0.15 μM each and 1× SYBR green mix.

qPCR normalization was done in two steps. (i) Each transcript was first quantified using a yeast genomic DNA standard. (ii) Quantification was then standardized against an arbitrarily selected reference transcript–YDL047W. In 13 out of 33 of the more abundant transcripts quantification was done against YDL047W alone.

Data analysis. Data analysis and computational methods, including processing of reads, alignment, transcript counting, detection of sequence variants and clustering, are described in detail in the **Supplementary Methods**.