CHAPTER NINETEEN

Single-Molecule Sequencing: Sequence Methods to Enable Accurate Quantitation

Christopher Hart, Doron Lipson, Fatih Ozsolak, Tal Raz, Kathleen Steinmann, John Thompson, *and* Patrice M. Milos

Contents

Abstract

Helicos® Single-Molecule Sequencing provides a unique view of genome biology through direct sequencing of cellular and extracellular nucleic acids in an unbiased manner, providing both quantitation and sequence information. Using

Helicos BioSciences Corporation, One Kendall Square, Cambridge, Massachusetts, USA

Methods in Enzymology, Volume 472
ISSN 0076-6879, DOI: 10.1016/S0076-6879(10)72002-4

a simple sample preparation, involving no ligation or amplification, genomic DNA is sheared, tailed with poly-A and hybridized to the flow-cell surface containing oligo-dT for initiating sequencing-by-synthesis. RNA measurements involving direct RNA hybridization to the flow cell allows for the direct sequencing and quantitation of RNA molecules. From these methods, a diverse array of applications has now been successfully demonstrated with the Helicos® Genetic Analysis System, including human genome sequencing for accurate variant detection, ChIP Seq studies involving picogram quantities of DNA obtained from small cell numbers, copy number variation studies from both fresh tumor tissue and formalin-fixed paraffin-embedded tissue and archival tissue samples, small RNA studies leading to the identification of new classes of RNAs, and the direct capture and sequencing of nucleic acids from cell quantities as few as 400 cells with our end goal of single cell measurements. Helicos methods provide an important opportunity to researchers, including genomic scientists, translational researchers, and diagnostic experts, to benefit from biological measurements at the single-molecule level. This chapter will describe the various methods available to researchers.

1. Introduction

The revolution in genomic sequencing that is currently occurring in the scientific community is heralding an exciting era of biology where experiments can be performed at a scale that fully elucidates the genome, its corresponding architecture, and the resulting transcriptome (all RNA molecules transcribed from the genome), revealing amazing new findings (Kahvejian *et al.*, 2008). This revolution is continuing as we move into the era of single-molecule sequencing where, for the first time, we are sequencing and measuring the actual molecules present in cells and tissues. This new era offers the promise of a better understanding of the fundamental basis of health and disease.

Helicos single-molecule sequencing offers the opportunity to examine billions of DNA or RNA molecules in a highly parallel fashion, scalable to sequencing of an entire human genome (Harris *et al.*, 2008; Ozsolak *et al.*, 2009; Pushkarev *et al.*, 2009). While other technologies may offer similar approaches, the simplicity and the scalability of single-molecule sequencing sets it distinctly apart from next-generation sequencing technologies. Further, these same principles contribute directly to the absolute quantitative nature of the technology. By eliminating cumbersome sample preparation steps, including complex ligations and polymerase chain reactions for amplification, single-molecule sequencing offers both sequence information and reliable quantitation for many different applications. Often referred to as ''third-generation'' sequencing (Hayden, 2009), the methods involved in single-molecule sequencing demonstrate these unique principles.

This chapter describes the methodological details for a variety of genomic applications used by the research and translational biology communities, including preparation of genomic DNA for complete genomic sequencing, copy number variation detection and chromatin immunoprecipitation (ChIP) studies. Quantitative aspects of single-molecule measurements for RNA are also described for methods associated with digital gene expression.

2. Basic Principles of Single-Molecule Sequencing

Helicos single-molecule sequencing utilizes sequencing-by-synthesis methodology, involving individual nucleic acid molecules that are initially fragmented in the case of genomic DNA, melted into single strands of DNA, and poly-A tailed. These DNA molecules are then captured as individual strands of DNA through deposition onto a glass Helicos® Flow Cell (Fig. 19.1B) surface coated with oligo-dT-50 oligonucleotides, which are then filled with dTTP and polymerase for the purpose of filling in any remaining nucleotides complementary to the poly-A tail. Following the fill, nucleic acid templates are locked in place by the addition of fluorescently labeled dCTP, dGTP, and dATP Virtual Terminator™ nucleotides, which incorporate as a single complementary nucleotide and prohibit subsequent extension prior to terminator cleavage. This "fill and lock" step ensures that each template become available for the sequencing-by-synthesis reaction (Bowers *et al.*, 2009; Harris *et al.*, 2008).

Following the fill and lock step, sequencing-by-synthesis is initiated through the addition of fluorescently labeled Virtual Terminator™ nucleotides added one at a time. Nucleotide incorporation occurs at the complementary position in the individual growing strands of DNA, using a DNA polymerase. After incorporation, unincorporated nucleotides are rinsed through the flow cell. The flow-cell surface is then illuminated with a laser and incorporation is detected by the fluorescent emission of light. The HeliScope Sequencer captures the images via a CCD camera and records which strands have incorporated a nucleotide and records positional information as well as cycle information to ensure conversion of the image to the individual DNA molecules as well as the A, C, G, or T nucleotide sequence information. After the images are captured, the terminator moiety is cleaved from the incorporated nucleotide, allowing subsequent addition of the next complementary nucleotide.

Once the thousands of images, which correspond to all the channels of the flow cell, have been recorded, the fluorescent label is cleaved from the nucleotide, allowing the instrument to continue incorporation of the next nucleotide in the addition cycle. In a standard run, the HeliScope Sequencer

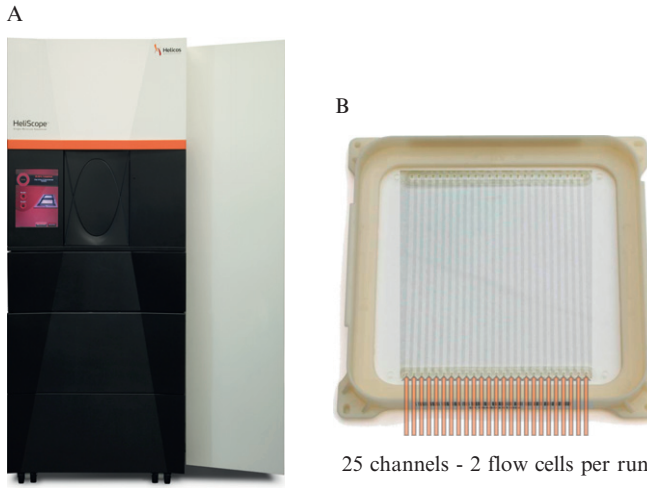

Figure 19.1 (A) The HeliScope® Single-Molecule Sequencer. A powerful genetic analyzer that performs single-molecule sequencing chemistry and captures images of single-molecule incorporation of fluorescently labeled nucleotides, producing accurate sequences of billions of templates at a time. (B) The Helicos® Flow Cell. Specifically designed for sequencing chemistry used with the Sequencer, two flow cells – each with **25** channels enable a multitude of applications all benefiting from Helicos proprietary chemistry.

completes 120 cycles of individual nucleotide additions. A representative visual image taken from the HeliScope Sequencer is shown in Fig. 19.2. At the end of the run, real-time image processing has converted all the images into a complete sequence file, recording both the DNA strand position and nucleotide string addition; scientists are then able to download the sequence file and begin the alignment to appropriate reference genomic or transcriptomic sequences. To date, numerous genomes have been sequenced using the HeliScope Sequencer, including genomes from M13 virus (Harris *et al.*, 2008), bacterial species, yeast, and *Caenorhabditis elegans* (Bowers *et al.*, 2009), culminating in the world's first sequencing of a human genome using single-molecule sequencing (Pushkarev *et al.*, 2009). The following will describe the basic methodologies one requires in order to prepare genomic templates for single-molecule DNA and cDNA sequencing.

3. Preparation of Genomic DNA for Single-Molecule Sequencing

The basic principles involved in the preparation of genomic DNA for subsequent sequencing-by-synthesis involve DNA fragmentation and quantitation, poly-A tailing, and 3′ end blocking to ensure that sequence obtained

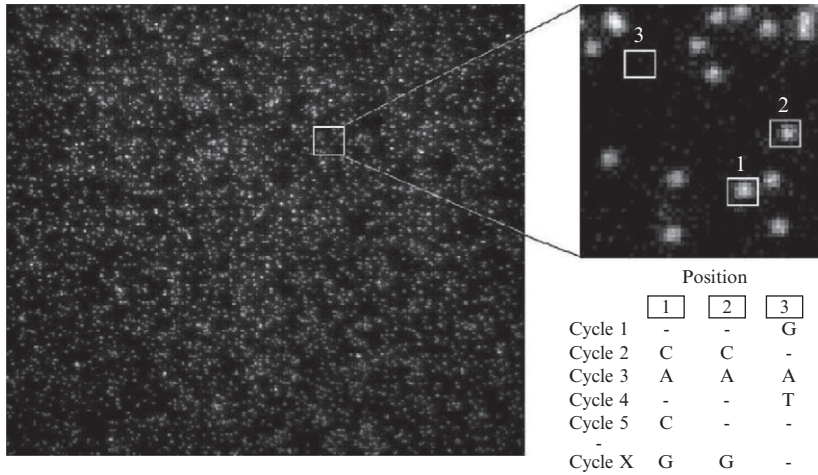

Figure 19.2 Helicos Flow Cell Image and Virtual Terminator® Nucleotide Incorporation. A true image derived from a section of the sequencing flow cell showing a close-up image of single molecules of DNA which have incorporated fluorescent Virtual Terminator® nucleotides. The right insert shows a close-up view of the single molecules and the definition of the nucleotide incorporated at the positions **1**, **2**, or **3** during the cyclic addition of nucleotides.

from the 3′ end of surface-bound oligoT is not contaminated with sequence from the 3′ end of the hybridized DNA strands. Figure 19.3 outlines the process described in detail below.

3.1. DNA fragmentation and quantitation

When quantities are not limiting, 1–3 μg of genomic DNA is typically used for single-molecule DNA sequencing of whole genomes, although much smaller quantities are also possible (see subsequent ChIP DNA Sequencing section). When the amount of DNA is low, care should be taken to use low-loss tubes and pipette tips. Addition of any type of carrier nucleic acid should be done cautiously as it could become a significant contaminant in sequencing.

3.1.1. DNA shearing

1. Prepare 1–3 μg of genomic DNA in a final volume of 120 μl 10 m*M* Tris 1 m*M* EDTA (1 × TE).
2. Any method of DNA shearing can be used; however, if complete coverage is desired, the method chosen should cleave the DNA randomly and provide a 3′ hydroxyl end for subsequent tailing. In the current protocol, ultrasonic shearing of the DNA is achieved using the Covaris

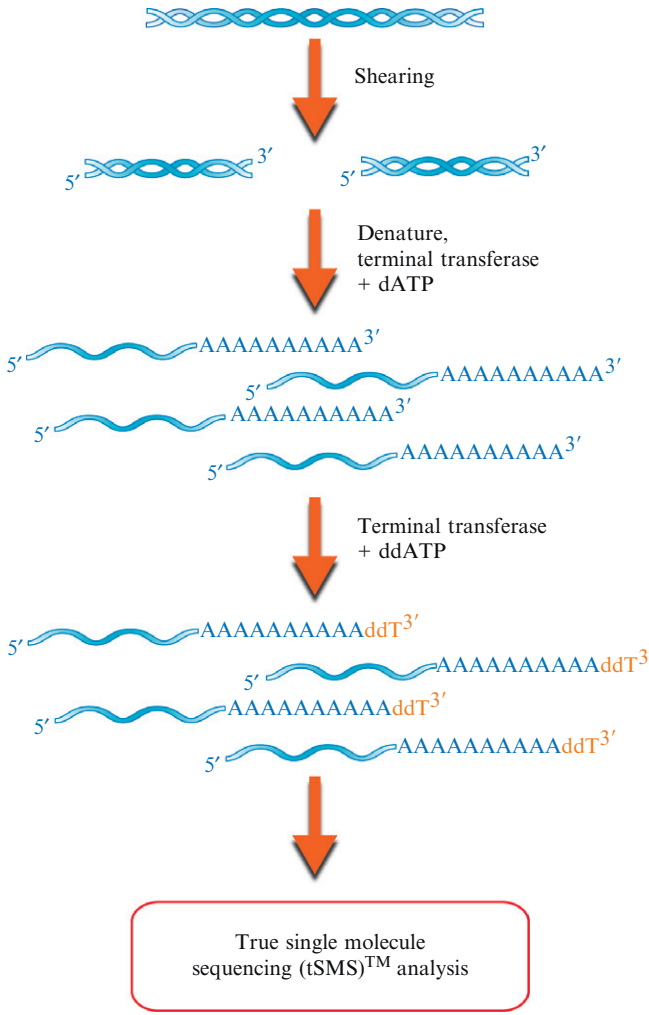

Figure 19.3 Depiction of DNA sequencing methodology. Illustration of the basic sample preparation steps of genomic DNA for single-molecule sequencing.

S2 instrument, resulting in fragmentation suitable for sequencing of the entire genomic sample. Conditions have been optimized by Covaris to allow for the use of genomic DNA ranging in length, at present, from 100 to 3000 base pairs (bp) so that the researcher can select the desired fragment size (Covaris, Woburn, MA; http://www.covarisinc.com). For typical genomic DNA sequencing using single reads, DNA is fragmented to an average size of 200–300 bp. For paired read sequencing in

which two or more regions of the same DNA fragment are sequenced, fragmentation of 1500 bp is optimal to provide spacer gaps ranging from 100 to 700 nucleotides in length.
3. Conditions vary for the various fragmentation sizes: For example, to shear DNA to 200 bp, the DNA is sheared in Covaris microTubes using 3 cycles of 60 s, 10% duty cycle, intensity 5, and 200 cycles per burst.
4. Transfer the DNA to a clean 1.5-ml microtube. At this point, the DNA sample can be stored at −20 °C.

3.1.2. DNA size selection

The DNA sample is subsequently cleaned using Agencourt AMPure® beads to remove small nucleic acids, nucleotides, and salts that may be present in the sheared sample.

1. Adjust the DNA volume to 100 μl.
2. Warm AMPure Bead solution to room temperature (RT). Vortex to resuspend.
3. Transfer DNA sample to 1.5-ml tube and add water to bring each sample to 100 μl. Vortex the beads again and add 300 μl AMPure Bead slurry.
4. Incubate at RT for 30 min. Shake the tube every 10 min.
5. Briefly centrifuge at low speed, capture beads on Dynal® magnetic stand for 5 min and carefully aspirate supernatant.
6. Wash beads twice with 700 μl freshly prepared 70% (v/v) ethanol.
7. Briefly centrifuge, place on magnet, remove ethanol, and dry pellet completely at RT for 5–7 min. Cracks will form when the pellet is dried sufficiently.
8. To elute the sheared DNA from the AMPure beads add 20 μl of water, pipette the beads and water up and down 20 times and place the tube back on the Dynal magnet.
9. Collect the 20 μl volume and transfer to a new 1.5-ml tube.
10. Repeat this process again to remove any remaining DNA on the AMPure beads. DNA will now be in the 40 μl volume.

3.1.3. Concentration estimation of 3′ ends for subsequent poly-A tailing

1. In order to effectively tail the 3′ ends of the genomic DNA, one must determine the approximate concentration of 3′ ends, which requires a determination of the average fragment size of the sheared DNA obtained by running a 2-μl DNA aliquot on a 4–20% gradient Tris Borate EDTA (TBE) polyacrylamide gel.
2. DNA standards of 1000 and 25 bp ladders are included for size comparison.

3. To estimate the size of the sheared product, compare the middle of the DNA smear to the size standards. An example gel is shown in Fig. 19.4.
4. Determine the double-stranded DNA concentration using a NanoDrop 1000 or 8000 spectrophotometer. Calculate the pmoles of the ends in the sample using the following formula.

$$\begin{aligned}\text{pmol } 3'\text{termini}/\mu\text{l} = {} & \text{XXng DNA}/\mu\text{l} \times (10^3\text{pg/ng}) \\ & \times (\text{pmole}/660\,\text{pg}) \\ & \times (1/\text{average fragment size as determined from gel}) \\ & \times 2(3'\text{termini/dsDNA molecule})\end{aligned}$$

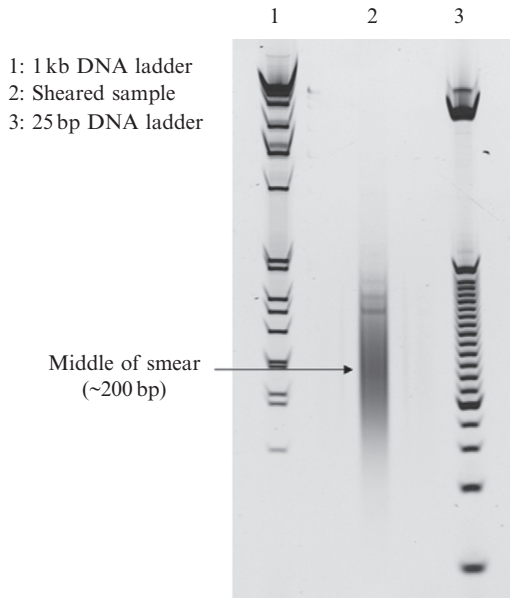

Figure 19.4 Example of gel sizing for sheared DNA. A 4–20% TBE gradient gel is used to assess the successful fragmentation of genomic DNA for subsequent poly-A tailing. Size standards of 1 kilobase ladder and 25 bp ladder are used to estimate average fragmentation length. Compare the average size in the middle of the sample smear to the size standards.

3.2. Poly-A tailing

The DNA fragments must be modified at their 3′ ends with a poly-A tail to allow for efficient hybridization onto the oligonucleotide-coated Helicos Flow Cell. Conditions are optimized to allow the addition of 90–200 poly-As to the single-stranded DNA molecules.

1. Prepare a sample DNA Tailing Mix assuming a 3.0-pmole sample reaction. For one reaction—4 μl 10× Terminal Transferase buffer, 4 μl 2.5 m*M* $CoCl_2$, 2 μl Terminal Transferase Enzyme (20 U/μl), 3.9 μl Helicos supplied Poly-A Tailing dATP and 1.1 μl deionized water (dH_2O). Please note: For the Tailing Control Tube, adjust the Poly-A Tailing dATP to 1.3 μl and the dH_2O to 3.7 μl.
2. Place the 3.0 pmole sheared DNA sample into a 200-μl PCR tube.
3. At the same time, prepare a separate 200-μl PCR tube with tailing control sample which consists of 0.8 pmoles of your DNA sample and 0.2 pmoles of tailing oligo control supplied by Helicos to monitor efficiency of tailing.
4. Denature the DNA by placing the sheared DNA and tailing oligo control tubes in a PCR Thermocycler, at 95 °C for 5 min. Snap cool by placing tubes in an aluminum block prechilled on an ice slurry for 2 min to prevent reannealing of the denatured single-stranded DNA.
5. Add 15 μl of Sample Tailing Mix or Control Tailing Mix to each DNA tube. Mix by pipetting up and down 10 times. Collect liquid contents by centrifuging briefly.
6. Place the tubes in the thermocycler using the following conditions: 37 °C for 60 min, 70 °C for 10 min, maintain at 4 °C until ready to proceed to next step.
7. Success of tailing is determined by monitoring the oligo control tailing. Twenty microliters of the oligo control is run on a 4–20% polyacrylamide gel in TBE alongside 100 and 25 bp ladders. An example of successfully sheared and tailed DNA is shown in Fig. 19.4.
8. Control-tailed oligos should migrate anywhere between 250 and 600 bp, indicating the sample is properly poly-A tailed with a desired tail length of between 90 and 200 dA.

3.3. 3′ end blocking

During flow-cell hybridization, the poly-A tail on the DNA sequencing templates may align imperfectly to the oligo-dT surface on the Helicos Flow Cell surface. This may result in the generation of a recessed 3′ end that can serve as a substrate for the sequencing-by-synthesis reaction. To prevent the incorporation of fluorescent Virtual Terminator nucleotides at that end

of the duplex, the 3′ ends of sheared DNA molecules are modified with a dideoxy terminator, using the following protocol.

1. Following the poly-A tailing, heat denature the DNA at 95 °C for 5 min in the thermocycler. Immediately remove and snap cool for a minimum of 2 min by placing in the ice-cooled aluminum block.
2. Add 0.3 μl of 500 μ*M* Biotin ddATP to each tube.
3. Add 2 μl Terminal Transferase (20 U/μl) to each tube. Mix thoroughly by pipetting up and down 10 times.
4. Collect contents by brief centrifugation.
5. Return to the thermocycler and run the following conditions: 37 °C for 60 min, 70 °C for 10 min, maintain at 4 °C until ready to proceed to next step.

Samples are now ready for hybridization to the Helicos Flow Cell for subsequent sequencing-by-synthesis. DNA concentrations in the range of 150–300 pM are utilized for each Helicos Flow Cell Channel typically in a 20-μl loading volume.

4. Bacterial Genome Sequencing

Helicos BioSciences has applied the above DNA sample preparation methodology to the sequencing of three bacterial genomes to demonstrate the principles of single-molecule sequencing—the simplicity of the sample preparation, the lack of amplification requirement, and the corresponding lack of G + C biases (Dohm *et al.*, 2008), as well as the evenness of coverage across a broad range of bacterial genomes, including *Escherichia coli* K12 MG1655, *Staphylococcus aureus* USA 3000, and *Rhodobacter sphaeroides* 2.4.1. The percentage of guanine and cytosine nucleotides (%GC) content of the genomes of these organisms represents the entire range of %GC content of kilobase-sized windows found in the human genome (Table 19.1). They have therefore been employed as reference genomes to test the ability of sequencing platforms to sequence the human genome. Achieving accurate and even coverage across these bacterial genomes demonstrates an absence of sequence content bias, which thus provides both sequence information as well as quantitative accuracy.

4.1. Preparation and sequencing of bacterial DNA

1. Shear and prepare 1 μg of bacterial DNA obtained from each species to 250–300 bp as described in Section 3.
2. Following sample preparation, load 150–300 pM of each bacterial DNA into individual flow-cell channels in a volume of 20 μl and

Table 19.1 Reference bacterial genomes containing diverse genome sequence content

	Genome size (Mb)	GC content (%)
E. coli	4.6	50.8
R. sphaeroides	4.3	68.8
S. aureus	2.8	32.7

Genomic size and G + C content of bacterial genomes used for demonstration of high-throughput sequencing methodologies due to diverse genomic content.

sequence-by-synthesis for 120 nucleotide cycle additions via the HeliScope Sequencer in an 8-day run in which both flow cells are utilized.

3. Align the single-molecule sequence reads obtained at run completion to the corresponding bacterial reference genomes using the Helicos IndexDP Genomic aligner (available at Helicos HeliSphere Technology Center http://open.helicosbio.com/mwiki/index.php/Main_Page).
4. The resulting throughput yields 12–20 million aligned reads per flow-cell channel or, given the two flow cells totaling 50 channels per run, 0.6–1 billion alignable reads per run.
5. A single channel provides upward of 80–120× coverage for these bacterial genomes, depending on the genome size, and represents some 3–4× more coverage than is required for accurate consensus calling.

Figure 19.5 shows the alignment view of reads and coverage within a selected region of the *E. coli* genome, which allows one to compare the sequence reads mapped to the region of a 5-kilobase pairs (kbp) window against the background of varying GC content in this same region. Coverage of sequence reads remains evenly distributed. Figure 19.5 also shows the read alignment, demonstrating the accuracy of the sequence information obtained.

4.2. Assessment of coverage and lack of bias

The ability to achieve consistent coverage across these genomes with special emphasis on regions of highly varying GC content is a hallmark of single-molecule sequencing. To demonstrate consistent coverage across the genomic regions of the three bacteria, we have plotted in Fig. 19.6 the coverage depth of single-molecule sequence reads binned across the bacterial genomic sequence and similarly plotted their known GC content in the same windows alongside the reference genomes. The HeliScope Sequencer produces even coverage across the entire span of genomic sequence content within a genome, even in the case of very G + C rich (*R. sphaeroides*) and highly A + T rich (*S. aureus*) genomes.

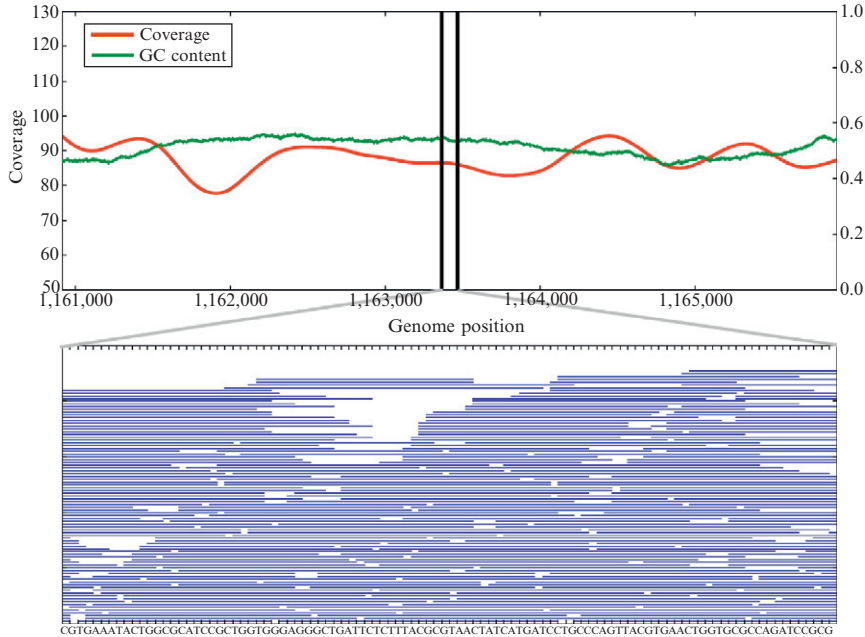

Figure 19.5 Detailed view of reads and coverage within an arbitrary position within the *E. coli* genome. Top panel shows sequence read coverage across each 5 kb region of the *E. coli* genome along (red smooth line) with the regional GC coverage (green jagged line). In each case positional statistics are derived from a sampling of the 500 bp upstream and 500 bp downstream regions. Bottom panel shows sequence reads as they aligned to the genome within the region demarcated by the vertical black lines in the top panel.

5. Human Genome Sequencing and Quantitation

Whole genome sequencing has been successfully achieved by scientists at Stanford University using Helicos single-molecule sequencing methods and the HeliScope Sequencer. Pushkarev *et al.* (2009) utilized 200 pM of poly-A tailed human genomic DNA per Helicos Flow Cell channel and loaded some 170 channels with the genomic DNA. The researchers obtained 148 Gigabases of raw sequence of, on average, 33-nt read length to achieve, on average, a 28× coverage of a human genome. Some 90% of the human genome was covered using this initial genome sequencing methodology. Sequence variants were identified as described in Pushkarev *et al.* (2009), which included data on copy number variation found within the human genome sequence.

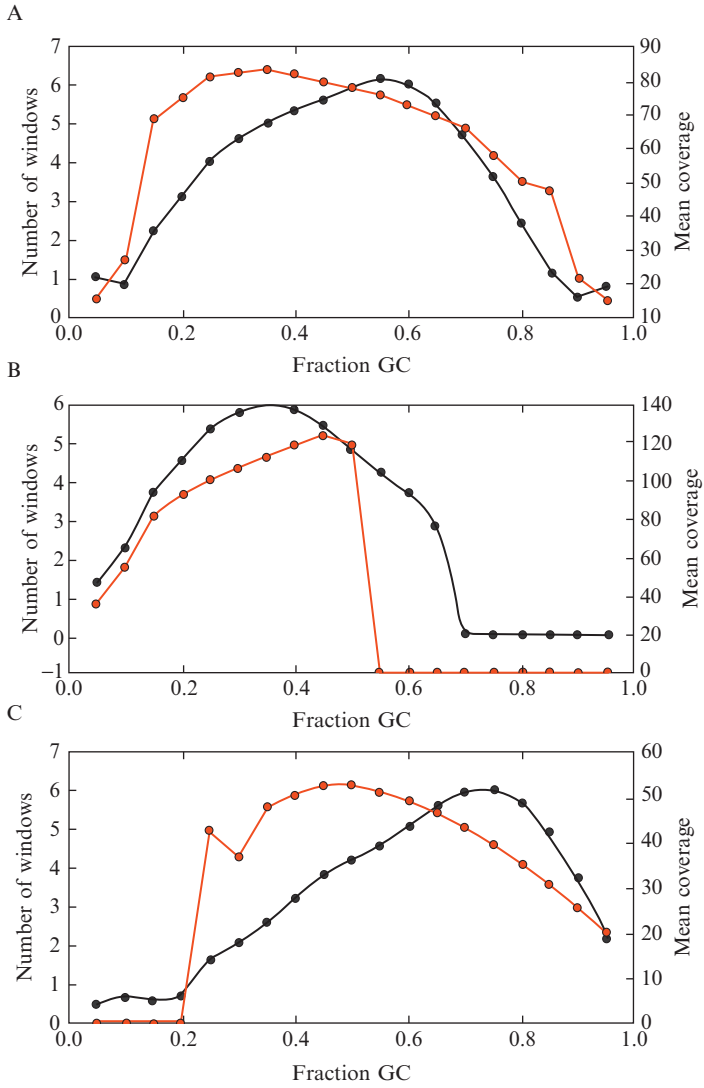

Figure 19.6 Comparison of genomic sequence coverage across differing G + C content within bacterial genomes. (A) *E. coli* (51% G + C). (B) *S. aureus* (32% G + C). (C) *R. sphaeroides* (70% G + C). Single molecule DNA sequencing provides minimal sequence bias across diverse genomic content. Sequence reads were mapped to each genome and the number of reads which map in discrete bins of the genome are plotted (red line) versus the expected bins if the mapping was perfect. Obtaining a signal as nearly identical to each other demonstrates the unique ability to sequence across diverse GC and AT rich regions. The analysis utilized a 200-bp sliding window, the local GC content and observed mean sequencing coverage were tabulated. Windows were then aggregated into GC-content bins ranging from 0 to 1 with a step size of 0.1. Plotted is the mean coverage (RED; Right Y axis) for each window within each of the aggregated GC content bins (BLACK; Left Y axis). A distinguishing feature of the Helicos SMS approach appears to be the minimal shifts in coverage across the vast majorities of sequence contexts.

5.1. Copy number variation

Copy number variation studies provide a methodology for detecting amplification and deletion of genomic regions across the human genome and often represent critically important knowledge of mutational events occurring in cancer genomes. Given the demonstration of evenness of sequence coverage in the bacterial strains representing the diversity of sequence found in the human genome, the use of single-molecule sequencing with the HeliScope Sequencer for an assessment of copy number variation represents an important, cost-effective method.

1. When available, prepare 1–2 μg of genomic DNA as described in Section 3. Less material may be utilized if sample is limited. This material may be obtained as genomic DNA prepared from tissue, blood, and formalin-fixed paraffin-embedded (FFPE) genomic DNA.
2. In the case of FFPE DNA, visualize the isolated DNA on a 1% agarose gel to determine the size of the genomic DNA. It is possible that, depending on the fixation of the tissue from which the DNA was obtained, the DNA still consists of high-molecular weight DNA and, if above 2–3 kbp, will require additional shearing as described in Section 3.1.1.
3. If the FFPE DNA falls below the size range of 2–3 kbp, proceed directly to Section 3.1.2 (DNA size selection) to ensure removal of small molecular weight DNA that can interfere with DNA sequence yields.
4. Following preparation of poly-dA tailed genomic DNA, load 150–300 pM of genomic DNA on each Helicos Flow Cell channel for the HeliScope Sequencer.
5. Depending on the desired level of resolution required for localization of the regions of amplification and duplication, a decision will be required regarding the depth of sequence coverage desired. At present performance, one channel of the HeliScope Sequencer provides 0.2–0.3× coverage of the human genome. This allows you to group sequence reads by using "bins" which can be between 10 and 50-kilobase-sized segments of the human genome. This resolution allows sufficient coverage for detection of amplification and duplication events, including loss of heterozygosity and two- to threefold amplification across the entire human genome.

Figure 19.7 summarizes the copy number variation data obtained from a human cancer cell line in which approximately 100 Mio sequence reads were mapped to the genome at a read bin size of 1 kbp intervals. These data are compared to existing comparative genomic hybridization data using an array technology. Peaks of amplification are easily detected, and the peak intensities reflect the extent of amplification. We refer also to the copy number variation data obtained from the first single-molecule human genome sequence (Pushkarev *et al.*, 2009). To further demonstrate the power of single-molecule sequencing technology, data used for the comprehensive

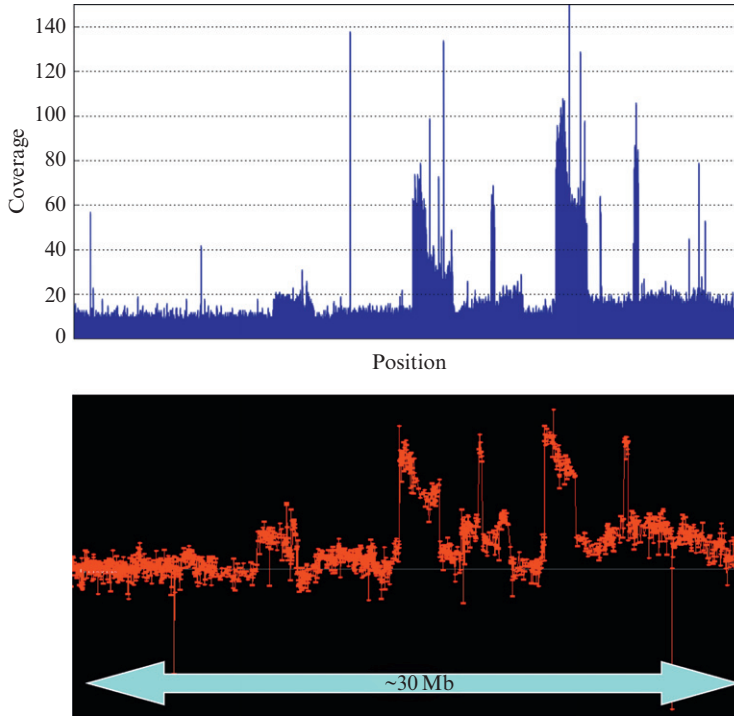

Figure 19.7 Detection of genomic amplification comparing Helicos single-molecule sequencing to comparative genomic hybridization. Genomic DNA from a breast cancer cell line was isolated, sheared, tailed, and sequenced according to the methods described in Section 5.1. Following sequencing, sequence reads were aligned to the human genome, binned into genomic bin sizes of 10 kb and bin sizes are plotted along the a 30 Megabase region chromosomal 20 (top panel). Regions of amplification are clearly detected in well described regions of Chr 20 previously identified using CGH arrays (bottom panel). (*CGH Data: Courtesy of Genome Institute of Singapore*).

view shown in Fig. 19.7 are replotted as individual channels of HeliScope Sequencer data and displayed in a 14-kbp region of the genome with 1-kbp smoothing of the read peaks (Fig. 19.8). These data reflect the consistency as well as the resolution achieved in single channels, allowing one to detect a region of five- to sevenfold amplification in this region.

6. Chromatin Immunoprecipitation Studies

Helicos single-molecule sequencing technology is ideally suited also for another area of genomic science where accurate quantitation is key, ChIP studies (Goren *et al.*, 2010). This method requires no ligation, amplification,

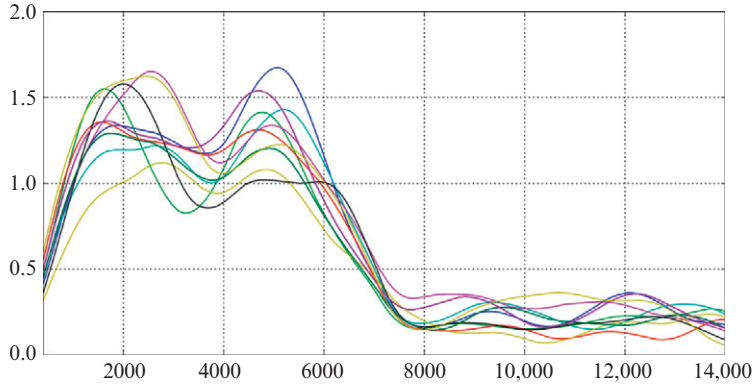

Figure 19.8 **Visualization of genomic amplification obtained from single HeliScope Sequencer Channels. Genomic reads used in Fig. 19.7 shown aligned to a segment of Chr 20 defining the boundary of amplification. Ten independent HeliScope Sequencer channels demonstrate the ability of a single channel of HeliScope Sequencer reads to clearly define the boundary of genomic amplification.**

or complicated cleanup steps—all of which have the potential to induce sample loss and bias. The Helicos ChIP Seq methodology consists of a 1-h 3′ poly-A tailing step followed by a 1-h 3′ dideoxy-blocking step. Recommended starting material consists of 6–9 ng ChIP DNA (average fragment size 400–500 bp), although as little as 1–3 ng DNA prepared using this same method can be successfully employed. Typical yields obtained with the recommended 6–9 ng ChIP DNA from mouse or human studies allow one to load 3–6 Helicos Flow Cell Channels with a yield of 7–12 Million aligned sequence reads per channel.

6.1. Preparation of ChIP DNA

1. The quantity of ChIP DNA should first be determined with the Quant-iT^{TM} PicoGreen dsDNA Reagent Kit (Invitrogen).
2. Samples should be free of RNA contamination and the use of the Qiagen Reaction Cleanup Kit (Qiagen) is recommended.
3. The micrococcal nuclease treatment used for fragmentation in some selected ChIP methods will generate phosphate groups on the 3′ ends and thus will require end repair prior to initiating the ChIP tailing protocol.
4. One must also consider the alternative types of shearing used for fragmentation to ensure the 3′ ends are amenable to direct poly-A tailing. Recommendations vary with shearing devices and one must check with the manufacturer on their advice for subsequent end repair.

6.2. ChIP DNA poly-A tailing

1. Prepare a mix of 2 μl 10× Terminal Transferase buffer (NEB), 2 μl 2.5 m*M* $CoCl_2$, ChIP DNA and Nuclease-free water 10.8 μl in a 200 μl PCR tube.
2. Place mixture in a thermocylcer and heat to 95 °C for 5 min to denature the DNA.
3. Remove tube from the thermocylcer and quickly chill in an aluminum block held in an icy slurry to prevent renaturation.
4. Prepare a mix of 1 μl Terminal Transferase (1:4 diluted, 5 U/μl; NEB), 4 μl 50 μM dATP, and 0.2 μl BSA (NEB).
5. Add 5.2 μl mix to the denatured DNA on ice to bring total volume to 20 μl.
6. Place tube in the thermocycler and run the following program: 37 °C for 1 h, 70 °C for 10 min, maintain at 4 °C until ready to proceed to next step.

6.3. ChIP DNA 3′ blocking

1. Denature the 20 μl poly-A tailed ChIP DNA at 95 °C for 5 min in the thermocycler, followed by immediate transfer to a prechilled aluminum block kept in an ice and water slurry.
2. Prepare a 10 μl mixture of 1 μl 10× Terminal Transferase buffer (NEB), 1 μl 2.5 m*M* $CoCl_2$, 1 μl Terminal Transferase (1:4 diluted, 5 U/μl), 0.5 μl 200 μM Biotin-ddATP and 6.5 μl Nuclease-free water.
3. Add the 10 μl mixture to the denatured, poly-adenylated ChIP DNA mixture for a final volume of 30 μl.
4. Place the tube in a thermocycler and run the following program: 37 °C for 1 h, 70 °C for 20 min, followed by 4 °C until ready to proceed to next step.
5. Add 2 pmol of a 50–80 nucleotide carrier oligonucleotide to the above terminal transferase reaction to minimize ChIP DNA loss during the sample loading steps. Since it does not contain a poly-A tail, the oligonucleotide will not hybridize to the Helicos Flow Cell.
6. Hybridize ChIP DNA sample to Helicos Flow Cell and sequence.

7. Digital Gene Expression for Transcriptome Quantitation

Full transcriptome sequencing using high-throughput sequencing platforms (RNA Seq) has increased the sensitivity and accuracy of gene expression analysis. However, RNA Seq results in an inherent bias as a result of more reads from longer transcripts and thus has reduced the

sensitivity for quantification of shorter transcripts (Oshlack and Wakefield, 2009). Further, assessing expression levels requires prior knowledge of transcript length for count normalization, which will not always be a reasonable demand, say in the case where there may be alternative splicing variants. Single-molecule sequencing digital gene expression (smsDGE) answers these difficulties and provides a route to quantitative analyses.

smsDGE differs from RNA Seq in that only a single sequence read is generated per transcript molecule, regardless of its length. This permits short transcripts to be detected with the same sensitivity as long ones. Thus, whereas it would require 50 million RNA Seq reads to quantify 95% of the human transcriptome, with smsDGE 10 million reads will suffice (Lipson *et al.*, 2009).

7.1. Methodology for single-molecule sequencing digital gene expression

Sample preparation for smsDGE is minimal, requiring neither PCR amplification nor ligation. A summary of the method is shown in Fig. 19.9. Single-stranded cDNA is made directly from total RNA or poly-A + RNA using poly-U primed reverse transcription. The RNA is then digested away using RNase, and a poly-A tail is added to the cDNA's 3′ end using terminal transferase. The sample can then be hybridized to the HeliScope flow-cell surface and sequenced (Lipson *et al.*, 2009).

7.1.1. Single-stranded cDNA preparation

7.1.1.1. cDNA synthesis

1. Thaw RNA on ice (1–8 μg total RNA or 100–200 ng poly-A+ RNA) preferably in 8 μl volume.
2. For sample: Prepare Master Mix A stock of 1 μl poly-U primer dTU25V (50 μM) and 1 μl dNTP nucleotide mix. Keep on ice. Prepare Master Mix B from Invitrogen SuperscriptIII kit as follows: 2 μl 10× Reverse Transcriptase buffer, 4 μl 25 mM $MgCl_2$, 2 μl 0.1 mM DTT, 1 μl RNaseOUT^{TM} and 1 μl Superscript III Reverse Transcriptase.
3. Aliquot 2 μl Master Mix A into a PCR tube.
4. Pipette 8 μl of RNA Sample into the PCR tube. Mix thoroughly by pipetting up and down.
5. Incubate the RNA at 65 °C for 5 min. Snap cool by placing in aluminum block held in an ice water bath.
6. Add 10 μl Master Mix B Reverse Transcriptase enzyme and buffer to each tube. Mix well and spin down.
7. Place PCR tubes in the thermocycler. Incubate at 40 °C for 5 min, 55 °C for 50 min, 85 °C for 5 min, maintain at 4 °C until ready to proceed to next step.

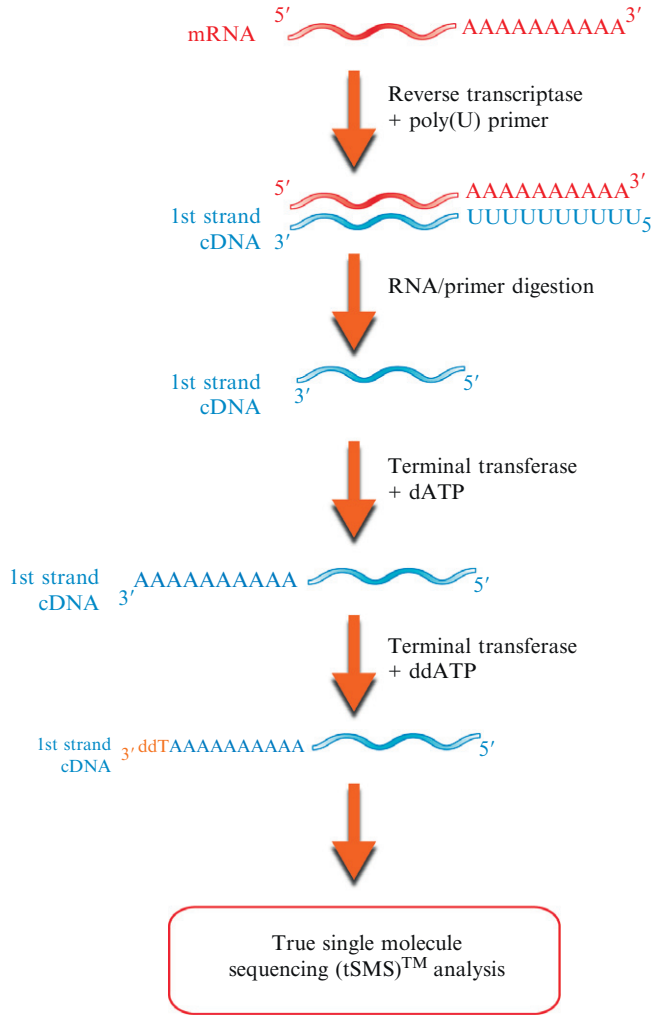

Figure 19.9 Overview of method utilized for single molecule sequencing digital gene expression. Principles employed in the single-molecule sequencing digital gene expression methodology described in Section 7.

7.1.1.2. RNA digestion

1. Add 1 μl RNase H (2 U/μl) to the cDNA Synthesis reaction. Mix well and incubate at 37 °C for 15 min, maintain at 4 °C until ready to proceed to next step.

2. Add 1 μl USER^{TM} (1 U/μl) enzyme. Mix well and incubate at 37 °C for 15 min, maintain at 4 °C until ready to proceed to next step.
3. Add 1 μl RNase I (50 U/μl) enzyme. Mix well and incubate at 37 °C for 15 min, maintain at 4 °C until ready to proceed to next step.

7.1.1.3. cDNA sample cleanup

1. Warm AMPure Bead solution to RT. Vortex to resuspend.
2. Transfer cDNA sample to a 1.5-ml tube and add water to bring each sample to 50 μl. Vortex the beads again and add 65 μl AMPure Bead slurry.
3. Incubate at RT for 30 min. Shake tube every 10 min.
4. Briefly centrifuge at low speed, capture beads on Dynal magnetic stand for 5 min and carefully aspirate supernatant.
5. Wash beads twice with 200 μl freshly prepared 70% (v/v) ethanol.
6. Briefly centrifuge, place on magnet, remove ethanol, and dry pellet completely at RT for 5–7 min.
7. Elute the cDNA sample from the beads with 20 μl distilled water twice.
8. Repeat entire cDNA sample cleanup once more. Final product will be in 40 μl volume.

7.1.1.4. cDNA quantification

1. Determine the concentration and yield for each cDNA sample preparation using a small volume spectrophotometer. If the sample concentration is likely below 2 ng/μl use the Quant-iT^{TM} OliGreen® ssDNA Reagent Kit and obtain spectrofluorometer reads accordingly.
2. Store samples at −20 °C to continue sample preparation the next day if desired.

7.1.2. Poly-A tailing of the cDNA

7.1.2.1. Poly-A tailing reaction

1. Obtain control oligonucleotides from Helicos® Digital Gene Expression Assay Reagent Kit.
2. Place 20–60 ng of cDNA into a PCR tube. Add water to bring each to 28 μl.
3. Add 7.5 μl of Helicos® Control Oligonucleotide. Mix well and store on ice.
4. Incubate at 95 °C for 5 min. Snap cool on ice. Briefly centrifuge.
5. Prepare poly-A Tailing mix of 5 μl 10× Terminal Transferase buffer, 5 μl $CoCl_2$, 2.5 μl Helicos® poly-A Tailing dATP, and 1.5 μl Terminal Transferase. Mix well.
6. Add 14 μl of poly-A Tailing Mix to the cDNA and pipette up and down.

7. Incubate at 42 °C for 60 min, 70 °C for 10 min, and maintain at 4 °C until ready to proceed to next step.

7.1.2.2. Determining the success of the tailing reaction

1. Success of the tailing reaction is determined by monitoring the oligo control tailing. Run an aliquot of the control oligonucleotide without poly-A tail addition and the control poly-A tailed oligonucleotide alongside your cDNA tailing reaction on a 4–20% gradient polyacrylamide gel in 1× TBE, together with a 25-bp ladder.
2. Since the cDNA molecules will be of a very broad size range, assess the length of the tail added to the control oligonucleotide as a measure of the tail added to the cDNA molecules.
3. Control-tailed oligos should migrate anywhere between 225 and 450 bp of the 25-bp ladder to ensure a proper poly-A tail with a desired length between 90 and 140 dA.

7.1.3. cDNA blocking

7.1.3.1. cDNA blocking reaction

1. Incubate the cDNA sample at 95 °C for 5 min. Snap cool on ice to denature.
2. Add 0.3 μl biotin-ddATP and 1.5 μl of Terminal Transferase enzyme. Mix well and spin down.
3. Incubate at 37 °C for 30 min, then 70 °C for 10 min, and maintain at 4 °C until ready to proceed to next step.

7.1.3.2. Poly-A tailing control oligonucleotide digestion

1. Add 1 μl USER Enzyme (1 U/μl) to the cDNA sample. Mix well and spin down.
2. Incubate at 37 °C for 30 min, maintain at 4 °C until ready to proceed to next step.

7.1.3.3. Sample cleanup

1. Transfer cDNA from digestion step above to a 1.5-ml tube. Add water to bring volume to 60 μl.
2. Mix cDNA with 60 μl AMPure Bead slurry and incubate at RT for 30 min. Shake every 10 min.
3. Capture the beads on Dynal magnetic stand for 5 min and carefully aspirate supernatant.
4. Wash beads twice with 200 μl freshly prepared 70% (v/v) ethanol.

5. Dry pellet completely at RT for 5–7 min.
6. Elute cDNA sample from beads with 20-μl distilled water twice.

Hybridize 150–300 pmol smsDGE cDNA in 20 μl volume to Helicos Flow Cell and sequence.

7.2. Demonstration of DGE counting reproducibility

To assess smsDGE reproducibility, we independently prepared three brain samples from the same RNA (poly-A RNA, Ambion, Austin TX) and sequenced each sample in a single HeliScope flow-cell channel. The three channels yielded 15, 14, and 12 million transcriptome-aligned reads. Transcript abundance ranged from 0 to 370,000 transcripts per million (tpm) with the highest seen for mitochondrial transcripts (chromosome M). Of the 28,800 transcripts included in our reference (UCSC genome database), 18,700 were present at a level higher than 1 tpm (>12 mapped reads). Transcript count reproducibility between samples was high ($r = 0.99$) with coefficient of variation (%CV) ranging from 4% at 100 tpm to 20% at 1 tpm (Fig. 19.10).

8. Summary

Methods for single-molecule sequence analysis of nucleic acids provide a diverse repertoire for quantitative and qualitative investigation of the genome and transcriptome. As such, we have attempted to describe many of the simple sample preparation methods offered to the research community. We will continue to optimize our sample preparation protocols to allow preparation and sequencing from picogram quantities of nucleic acid (Ozsolak *et al.*, 2010)—all important for maximizing researchers abilities to perform important biological experiments with limiting biological sample amounts. These methods will serve as the starting point for the next edition of methods for single-molecule sequencing.

ACKNOWLEDGMENTS

Special thanks to the many individuals who have contributed to the success of Helicos BioSciences technology—for their scientific excellence and passions to develop a remarkable new technology and all its broad applications.

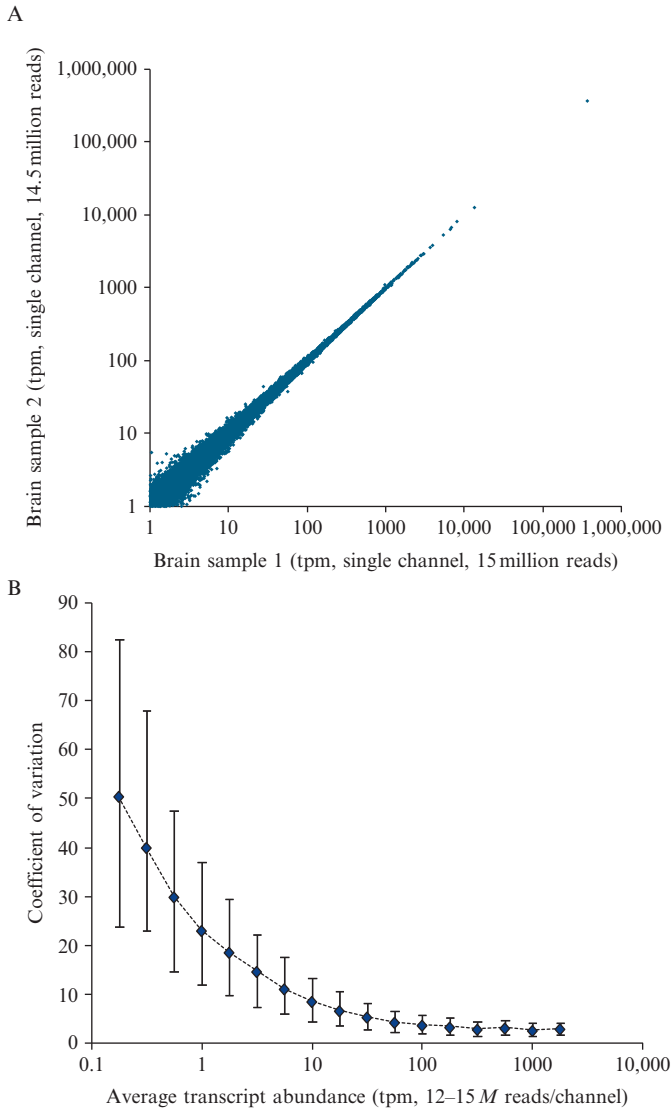

Figure 19.10 Demonstration of transcript counting reproducibility obtained using smsDGE methods with human brain RNA. smsDGE transcript quantification of independently prepared human brain samples. (A) Transcript count comparison between two samples run on a single flow-cell channel each. Each sample represents a single transcript ($r = 0.99$). (B) Coefficient of variation (%CV) across transcript abundance levels between three samples at 12, 14, and 15 million transcriptome-aligned reads per channel.

REFERENCES

Bowers, J., Mitchell, M., Beer, E., Buzby, P. R., Causey, M., Efcavitch, J. W., Jarosz, M., Krzymanska-Olejnik, E., Kung, L., Lipson, D., Lowman, G. M., Marappan, S., *et al.* (2009). Virtual terminator nucleotides for next generation DNA sequencing. *Nat. Methods* **6,** 593–595.

Dohm, J. C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36** (16), e105.

Goren, A., Ozsolak, F., Shoresh, N., Ku, M., Adli, M., Hart, C., Gymrek, M., Zuk, O., Regev, A., Milos, P. M., and Bernstein, B. E. (2010). Chromatin profiling by directly sequencing small quantities of immunoprecipitated DNA. *Nat. Methods* **7**(1), 47–49.

Harris, T. D., Buzby, P. R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., Colonell, J., Dimeo, J., Efcavitch, J. W., Giladi, E., Gill, J., *et al.* (2008). Single-molecule DNA sequencing of a viral genome. *Science* **320**(5872), 106–109.

Hayden, E. (2009). Genome sequencing: The third generation. *Nature* **457,** 768–769.

Kahvejian, A., Quackenbush, J., and Thompson, J. F. (2008). What would you do if you could sequence everything? *Nat. Biotechnol.* **26,** 1125–1133.

Lipson, D., Raz, T., Kieu, A., Jones, D. R., Giladi, E., Thayer, E., Thompson, J. F., Letovsky, S., Milos, P., and Causey, M. (2009). Quantification of the yeast transcriptome by single-molecule sequencing. *Nat. Biotechnol.* **27,** 652–658.

Oshlack, A., and Wakefield, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct* **4,** 14.

Ozsolak, F., Platt, A., Jones, D., Reifenberger, J., Sass, L. E., McInerney, P., Thompson, J. F., Bowers, J., Jarosz, M., and Milos, P. (2009). Direct RNA sequencing. *Nature* **461,** 814–818.

Ozsolak, F., Goren, A., Gymrek, M. A., Guttman, M., Regev, A., Bernstein, B. E., and Milos, P. M. (2010). Digital transcriptome profiling from attomole-level RNA samples. *Genome Res.* [Epub ahead of print].

Pushkarev, D., Neff, N. F., and Quake, S. R. (2009). Single-molecule sequencing of an individual human genome. *Nat. Biotechnol.* **27,** 847–850.