Transcriptome sequencing with the Genome Sequencer FLX system

Transcriptome sequencing is a term that encompasses experiments including mRNA transcript-expression analysis (full-length mRNA, expressed sequence tags (ESTs) and ditags), novel gene discovery, gene space identification in novel genomes, assembly of full-length genes, single-nucleotide polymorphism (SNP), insertion-deletion and splice-variant discovery, as well as analyses of allele-specific expression and chromosomal rearrangement. The combination of long, accurate reads and high throughput makes 454 Sequencing™ analysis on the Genome Sequencer FLX ideally suited to detailed transcriptome investigation.

With the introduction of next-generation sequencing, transcriptome sequencing has become more encompassing, as it is now possible to detect and identify nearly every class of molecules that is transcribed—from the short microRNA to the longer 5′ and 3′ untranslated regions to the longest, full-length mRNA. The current challenges for researchers are to decide which class of transcript (or portion of the transcript) is to be studied to address their experimental goals and to select the appropriate input sample (total RNA or mRNA) quality and quantity. The choice of which sample protocol to use to isolate and purify the correct class of transcriptome molecules for sequencing is often confusing. To help clarify this complex situation, here we present an overview of full-length mRNA sequencing using the Genome Sequencer FLX platform (**Figs. 1** and **2**), including general guidance on experimental setup and design (see **Table 1** for selected publications). The general principles we present should serve as a starting point for aligning experimental goals, sample qualities and experimental guidelines. Here we limit our scope to mRNA and will not cover small RNAs such as microRNAs and Piwi-interacting RNAs. Information on how to prepare small RNAs for 454 Sequencing analysis is available (Jarvie, T. & Harkins, T. Small RNA analysis using the Genome Sequencer FLX system. *Nat. Methods* **4**, 2007).

Starting material

Transcriptome sequencing experiments require the conversion of mRNA into cDNA before the sequencing reaction. Double-stranded cDNA (the most common starting material for transcriptome sequencing experiments) can be generated from mRNA by a variety of methods. In all methods, experimental success depends upon the removal of ribosomal RNA, as rRNA will be the majority species sequenced if not carefully removed from the total RNA. For all protocols, a control dataset should be established using a high-quality mRNA (for example, rat liver mRNA purchased from a vendor). This

Figure 1 | The Genome Sequencer FLX instrument. The Genome Sequencer FLX supports a number of formats, allowing users to customize the number of samples per instrument run and the number of reads per sample. A single run can be physically divided into 2, 4, 8 or 16 samples. Multiplexing of samples is also supported by Multiplex Identifier tags.

Thomas Jarvie[1] & Timothy Harkins[2]

[1]454 Life Sciences, 20 Commercial Street, Branford, Connecticut 06405, USA. [2]Roche Diagnostics, Roche Applied Science, 9115 Hague Road, Indianapolis, Indiana 46250, USA. Correspondence should be addressed to T.J. (thomas.jarvie@roche.com).

dataset would enable the researcher to understand how the sample performs as it progresses through the different stages of the experiment, providing a tool to troubleshoot subsequent projects that use experimental samples.

To sequence full-length cDNA using the Genome Sequencer System, the first step is to procure, ideally, 3–5 µg of double-stranded cDNA. Although one can proceed using less material, this quantity will allow for the detection of low-abundance transcripts and provide sufficient material for quality-control steps throughout the sample preparation process. This cDNA will then be processed by the standard Genome Sequencer library-preparation method using the GS DNA Library Preparation Kit to generate single-stranded DNA ready for emulsion PCR (emPCR™). If quantitation of the transcripts is desired, then the input cDNA material should be unamplified (and not normalized).

Amplification for low-quantity samples

Often starting samples are limited in quantity, making it impractical to procure the required amount of total RNA or mRNA. One can still generate a sufficient quantity of cDNA for the standard Genome Sequencer library preparation; however, the cDNA must be amplified first. Several amplification methods can be used. One method is to use the Eberwine protocol to synthesize first-strand cDNA using an oligo(dT) that contains a T7 promoter. The first-strand cDNA can then be used as a template for *in vitro* cRNA synthesis by T7 RNA polymerase. After sufficient amplification, the cRNA can then be used to make the appropriate quantity of cDNA for library prep.

The other methods incorporate a universal adaptor onto each end of the mRNA, generate first-strand cDNA and then use the cDNA as a template for amplification. It is important to note that protocols that require a 5′ cap (only full-length mRNA will have a 5′ cap) will filter out mRNA that is not full-length. Therefore, these protocols are excellent for ensuring full-length cDNA if the sequencing of full-length template (or the 5′ and 3′ ends) is the goal. However, if the input material is degraded, templates will be lost and the downstream sequencing may be compromised, yielding an incomplete detection and identification of transcripts. In contrast, protocols that are not limited to the amplification of mRNA with a 5′ cap will amplify full-length and non–full-length mRNA. When the input material is full-length mRNA, these protocols generate high-quality template for sequencing. However, when the input mRNA is degraded (not full-length), one must be careful when using these protocols as the degraded mRNA may lead to short fragments that tend to amplify well in PCR. Furthermore, if not removed before the emPCR step, these short fragments will amplify more efficiently in emPCR. As a result, they have the potential to become a substantial fraction of fragments sequenced, thereby lessening the opportunity to get a comprehensive view of the full-length transcripts of primary interest.

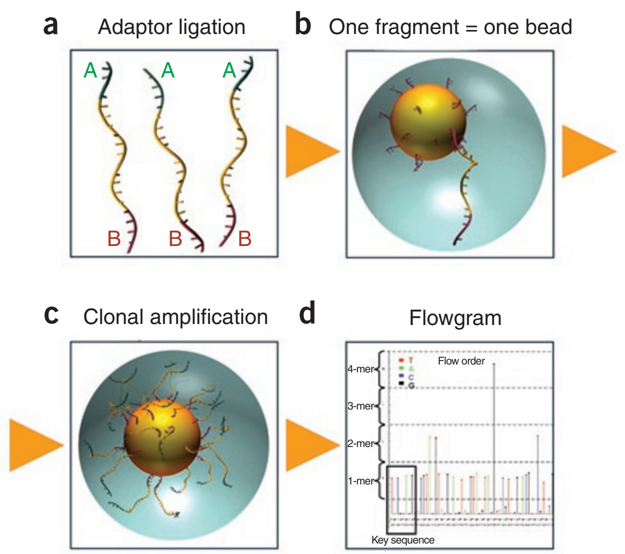

Figure 2 | One fragment = one bead = one read. (**a**–**d**) Genome Sequencer process–specific A and B adaptors are appended to each fragment or small RNA (**a**) to allow binding to DNA capture beads (**b**). The beads are placed into a water-in-oil emulsion that contains hundreds of thousands of PCR microreactors. Each microreactor contains all reagents necessary for PCR. The entire collection of millions of fragments and beads is amplified in parallel (**c**), with each fragment amplified without the introduction of competing or contaminating sequences. When sequenced on the Genome Sequencer FLX, each clonally amplified fragment generates its own unique sequence read, represented by a flowgram (**d**), with over 400,000 reads generated per instrument run.

Library preparation and sequencing

Nebulization, the fragmentation process used in the standard Genome Sequencer shotgun library preparation procedure, works best (yields the most unbiased and random fragmentation) on high-molecular-weight DNA. cDNA, particularly fragments shorter than 2 kb, may not fragment uniformly. This lack of uniformity may manifest as non-uniform coverage across the entire length of the cDNA (commonly seen as an increase in coverage near the ends and the middle of the cDNA). If the cDNA is less than 800 bp, one can ligate the GS DNA Library Preparation Kit adaptors directly to the fragments (no nebulization) with the Low Molecular Weight Protocol in the GS FLX Shotgun DNA Library Preparation Method Manual. If the cDNA is 1–2 kb, this will be too long to directly ligate the adaptors and too short for uniform nebulization. The length of these fragments will prevent complete sequence coverage—and, most importantly, they will not perform well in emPCR. If one is faced with cDNA in this size range, the best option for preparing it for sequencing may be an alternative fragmentation method such as cutting with a restriction enzyme. If the fragment distribution is known at the mRNA step, the Random Transcriptome Sequencing (TSEQ) protocol[1,2] is a suitable choice for the processing of materials in this size range.

Long poly(A:T) tails in cDNA may result in sequencing reads of low quality for the Genome Sequencer FLX. Two separate methods have been developed to either reduce or eliminate the poly(A:T) tail. The first poly(A) reduction method uses a type of universal primer when making the first strand of cDNA. In this approach, the poly(T) primer includes a site recognized by a type II restriction enzyme (16–22-nucleotide cutter) and is designed to ensure that the primer will bind

Table 1 | Transcriptome sequencing with the Genome Sequencer platform

Full-length cDNA	Bräutigam, A. *et al.* Low-coverage massively parallel pyrosequencing of cDNAs enables proteomics in non-model species: comparison of a species-specific database generated by pyrosequencing with databases from related species for proteome analysis of pea chloroplast envelopes. *J. Biotechnol.* published online 17 February 2008 (doi:10-.1016/j.jbiotec.2008.02.007). Emrich, S.J. *et al.* Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Res.* **17**, 69–73 (2007). Salehi-Ashtiani K. *et al.* Isoform discovery by targeted cloning, 'deep-well' pooling and parallel sequencing. *Nat. Methods* **5**, 597–600 (2008). Sugarbaker, D.J. *et al.* Transcriptome sequencing of malignant pleural mesothelioma tumors. *Proc. Natl. Acad. Sci. USA* **105**, 3521–3526 (2008). Vera, J.C. *et al.* Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Mol. Ecol.* **17**, 1636–1647 (2008). Weber, A.P.M. *et al.* Sampling the *Arabidopsis* transcriptome with massively-parallel pyrosequencing. *Plant Physiol.* **144**, 32–42 (2007).
Microbial transcriptomes	Frias-Lopez, J. *et al.* Microbial community gene expression in ocean surface waters. *Proc. Natl. Acad. Sci. USA* **105**, 3805–3810 (2008). Mao, C. *et al.* Identification of new genes in *Sinorhizobium meliloti* using the Genome Sequencer FLX system. *BMC Microbiol.* **8**, 72 (2008).
3′ untranslated regions	Eveland, A.L. *et al.* Transcript profiling by 3′-untranslated region sequencing resolves expression of gene families. *Plant Physiol.* **146**, 32–44 (2008). Torres, T.T. *et al.* Gene expression profiling by massively parallel sequencing. *Genome Res.* **18**, 172–177 (2008).

This is a selected list of publications. For a list of additional publications, see http://www.genome-sequencing.com/.

at the poly(A) tail and core mRNA sequence. An example is below, in which XXXXXXXXXXX is a type II enzyme recognition site:

5′-XXXXXXXXXXXTTTTTTTTTTTTTTTTTTTTVN-3′.

When choosing the enzyme and primer design, the objective is to cut within the poly(A) tail, reducing its length so that all subsequent sequences start with a short run of thymidines (TT to TTTTTT). A protocol is available in several publications[3,4].

The second method for minimizing the poly(A) tail is to use a modified poly(T) primer for first-strand cDNA synthesis that is not a straight run of thymidines, but thymidines interspersed with other nucleotides[5]. When performed properly, the modified poly(T) primer method will work well for cDNA synthesis while minimizing the run of poly(A:T). Sequencing through these modified poly(A:T) tails will not negatively impact sequence quality (presented in a poster by E. Meyer *et al. De novo* sequencing of coral larva transcriptome using 454 FLX. Annual meeting of the Society for Molecular Biology and Evolution, Barcelona, Spain (2008)).

The sequencing read lengths in transcriptome experiments are strongly dependent upon the input material. Unlike genomic DNA, where the Genome Sequencer FLX average read lengths are typically in the 250–300-bp range, average cDNA read lengths are often in the 200-bp range owing to a shorter fragment distribution that is being sequenced.

Data analysis

When sequencing a well-characterized genome (for example, the human genome or those of model organisms such as mouse, *Drosophila* or *Arabidopsis*), analysis typically proceeds by taking the FASTA reads that are generated by the Genome Sequencer FLX software and mapping them against the reference genome to detect SNPs and identify novel transcripts, or against a full-length transcriptome database to identify splice variants. When sequencing an unknown genome, the reads can be mapped against a closely related genome (or transcriptome) to identify homology between the two genomes/transcriptomes or the reads can be assembled and used to identify the gene space of the novel genome.

Summary

The Genome Sequencer FLX System is a powerful platform for transcriptome sequencing. The long, accurate read lengths enable a range of applications including genome annotation, novel-transcript identification, splice-variant detection, expression analysis (including allele-specific expression and the ability to distinguish paralogous genes), full-length gene assembly and SNP discovery (and discovery of other variations such as indels).

More information about the Genome Sequencer System is available from Roche Applied Science (http://www.genome-sequencing.com/). 454, 454 Life Sciences, 454 Sequencing and emPCR are trademarks of 454 Life Sciences Corporation, Branford, Connecticut, USA. For life science research use only. Not for use in diagnostic procedures. License disclaimer information is available online (http://www.genome-sequencing.com/).

1. Sugarbaker, D.J. *et al.* Transcriptome sequencing of malignant pleural mesothelioma tumors. *Proc. Natl. Acad. Sci. USA* **105**, 3521–3526 (2008).
2. Toth, A.L. *et al.* Wasp gene expression supports an evolutionary link between maternal behavior and eusociality. *Science* **318**, 441–444 (2007).
3. Ng, P. *et al.* 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 (2006).
4. Ng, P. *et al.* Paired-end diTagging for transcriptome and genome analysis. *Curr. Protoc. Mol. Biol.* 21.12.1–21.12.42 (2007).
5. Beldade, P. *et al.* A wing expressed sequence tag resource for *Bicyclus anynana* butterflies, an evo-devo model. *BMC Genomics* **7**, 130 (2006).

This article was submitted to *Nature Methods* by a commercial organization and has not been peer reviewed. *Nature Methods* takes no responsibility for the accuracy or otherwise of the information provided.