454
SEQUENCING

Comprehensive transcriptome analysis with the Genome Sequencer FLX System

Protein isoforms make the transcriptome complex and challenging to resolve. Different forms of a protein may be produced from related genes or may arise from the same gene by alternative splicing. With a combination of long (400–500 bp) 454 Sequencing™ reads, dedicated GS De Novo Assembler software and a straightforward protocol using just 200 ng of RNA as sample input, the Genome Sequencer FLX System offers a powerful solution to help unravel the complexities of the transcriptome.

Transcriptome sequencing offers a cost-effective, rapid alternative to *de novo* sequencing, which is time-consuming and expensive for large genomes. The cDNA analysis approach has expanded sequencing capabilities to support research in fields such as cytology and protein analysis. The applications of transcriptome sequencing are potentially unlimited and include mRNA transcript analysis and novel gene discovery, as well as identification of single nucleotide polymorphisms, insertions-deletions and chromosomal rearrangements. When investigating unsequenced organisms or mapping back to the human genome, a complex, challenging transcriptome or proteome can be sequenced, analyzed and compared within or between populations¹. For the investigation of cells and their pathways, cDNA sequencing data yields reconstructed isoforms that are accessible in polyploid data sets where assembling sequence differences can be challenging².

Compared to *de novo* sequencing, cDNA sequencing is an easier method to identify coding sequences in eukaryotes, as all intergenic sequences of the genome have been eliminated, making it easier to annotate genes. This approach also presents a challenge, however, as it yields approximately 450,000 sequences of expressed genes in one high-throughput GS FLX™ Titanium series sequencing run with no information on the genomic potential region. cDNA mapping's "gold standard" is a manual approach for functional annotation of the genome, but an increasing number of tools are being developed to accurately reflect both the functional assignments and the evidence supporting them. In order to assign functional annotation to uncharacterized cDNA sequences, software such as FANTOM³ and SEED⁴ are available to analyze the data and start making sense of cell characterizations, environmental influences and so on.

The Genome Sequencer FLX System (**Fig. 1**) is a powerful platform for transcriptome sequencing. Its long accurate reads and ease of transcript assembly enable a full range of applications including transcript annotation, identification of novel transcripts, assembly of full-length genes, splice variant detection, expression analysis, and discovery of SNPs or other variations such as insertions and deletions. With the Roche 454™ cDNA sequencing application and cDNA-specific analysis software, one can obtain an unbiased transcriptome survey by sequencing full-length cDNA libraries on the Genome Sequencer FLX System using GS FLX Titanium series reagents.

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 Identifiers (molecular barcodes).

Clotilde Teiling

Roche Diagnostics, Roche Applied Science, Indianapolis, Indiana, USA.
Correspondence should be addressed to Clotilde Teiling. E-mail: Clotilde.Teiling@roche.com.

APPLICATION NOTES

Figure 2 | Overview of the cDNA workflow, from sample preparation to sequencing and data analysis on the Genome Sequencer FLX Instrument.

Sample preparation

cDNA sequencing experiments require the conversion of mRNA into cDNA prior to preparation with the GS Rapid Library Prep Kit (Cat. No. 05 608 228 001), as illustrated in **Figure 2**. Utilizing the cDNA Synthesis System Kit (11 117 831 001) from Roche, transcription of RNA into cDNA is performed using 200 ng of RNA as starting material (OD 260/280 ≥ 1.8). mRNA is isolated and then treated with zinc chloride to fragment the mRNA into the desired 450-bp range. The sequencing read length is determined by the fragmentation of the RNA (first step, **Fig. 3**). RNA size varies; to utilize the full potential of the Genome Sequencer FLX System, the RNA length should be as close as possible to 450 bp on average.

The cDNA synthesis protocol, developed by 454 Life Sciences, utilizes random hexamer primers, which contain every possible 6-base single-stranded DNA and can therefore hybridize anywhere on the RNA. Once randomly fragmented, cDNA first-strand synthesis is primed utilizing the random hexamers, which diminishes the priming of the 3' poly(A) tail; it is then followed by second-strand synthesis. From this hybridization, reverse transcriptase utilizes the double-stranded sequence as a primer to start translation. The double-stranded cDNA is used directly as starting material for rapid library preparation, yielding blunt-ended cDNA with the addition of the overhanging A. The resulting cDNA fragments are then polished and prepared for 454 Adaptor ligation. Once ligated, the DNA is ready for emulsion

Figure 3 | cDNA Rapid Library Protocol. With 200 ng of messenger RNA as starting material, a complementary DNA (cDNA) library can be quickly prepared using a straightforward protocol consisting of four major steps using two Roche kits: the cDNA Synthesis System Kit and the GS Rapid Library Prep Kit.

PCR (emPCR™) (**Fig. 3**). 454 pyrosequencing is then performed using standard GS FLX Titanium series reagent kits.

Application

Depending on the experimental goal, researchers can select the appropriate input material (for instance, total RNA, mRNA) for sample preparation. Complementary DNA (cDNA) is DNA synthesized from an mRNA template in a catalyzed reaction. In higher eukaryotes, mRNA is a more useful predictor of a polypeptide sequence than is a genomic sequence, because introns have been spliced out. mRNA transcripts for a given cell reflect the genes that are highly expressed.

Transcriptomes are particularly important for understanding the processes of cellular differentiation. Analysis of mRNA expression levels can be complicated, and small changes in expression can have a large impact on protein levels present in the cell. The GS *De Novo* Assembler software configures sequencing reads into a single “isotig,” assembling reads into full-length contigs, similar to *de novo* genome assembly into a collection of full-length transcripts unique to that sample. Furthermore, the number of identified transcripts yields relative mRNA expression data. The *de novo* transcriptome assembly can also be used as a reference template for mapping reads from other experimental samples or species. Our premise is that without a full assembly of the transcriptome, counting or mapping exercises are error-prone⁵. Full transcriptome assembly is only possible using the 400–500 bp long reads unique to 454 Sequencing systems — an advantage that enables the resolution of isoforms and, consequently,

contigs that can be joined into isotigs representing full-length protein-coding transcripts. In addition, 454 Sequencing long reads enable the discovery of allele-specific expression in heterozygotes⁶.

Bioinformatics

The improved GS *De Novo* Assembler software enables straightforward transcriptome analysis. In polyploid data sets, the assembly of sequence differences can be a challenge. With the GS *De Novo* Assembler software, each isoform within the data set is reconstructed, enabling the differentiation of each individual within a sample. The understanding of isoforms offers the ability to discern the different proteins within a sample; from this information, one can categorize each protein to characterize the proteome of the same gene. In addition, the GS *De Novo* Assembler creates “isotigs” — combinations of selected contigs — that are aligned into a putative transcript. Full-length transcripts (that is, 400 to 500 bp) are also sequenced in single reads. By providing full-length transcript analysis, the GS *De Novo* Assembler offers a pioneering software solution that enables a growing number of non-model genome studies, some of which are highlighted in the included citations.

Summary

Comprehensive transcriptome analysis is a powerful tool for discovery in both novel and previously characterized genomes. cDNA sequencing data can be utilized in a broad range of applications, including annotation of unsequenced genomes, identification of genetic variations and phylogenetic analysis. Additionally, full-length transcriptome data is valuable for downstream analysis, such as identifying novel isoforms and their relative concentration, serving as a

reference for further mapping studies, linking isoforms to disease states, and generating relevant sequence data for an unsequenced species.

With the combination of long, accurate reads and the intuitive GS *De Novo* Assembler software, the Genome Sequencer FLX System is an ideal tool for comprehensive transcriptome sequencing.

For more information about the Genome Sequencer FLX System, visit www.454.com. 454, 454 Life Sciences, 454 Sequencing, GS FLX, and emPCR are trademarks of Roche. **For life science research use only. Not for use in diagnostic procedures.** License disclaimer information is available online (www.454.com).

1. Ferguson, L. *et al.* Characterization of a hotspot for mimicry: assembly of a butterfly wing transcriptome to genomic sequence at the HmYb/Sb locus. *Molecular Ecology* **19 Suppl 1**, 240–254 (2010).
2. Ueno, S. *et al.* Bioinformatic analysis of ESTs collected by Sanger and pyrosequencing methods for a keystone forest tree species: oak. *BMC Genomics* **11**, 650 (2010).
3. Katayama S. *et al.* Integrated analysis of the genome and the transcriptome by FANTOM. *Brief Bioinform.* **5**, 249–258 (2004).
4. Desnues, C. *et al.* Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature* **452**, 340–343 (2008).
5. Murchison, E.P. *et al.* The Tasmanian Devil transcriptome reveals Schwann cell origins of a clonally transmissible cancer. *Science* **327**, 84–87 (2010).
6. Salem, M. *et al.* Characterization of the rainbow trout transcriptome using Sanger and 454-pyrosequencing approaches. *BMC Genomics* **11**, 564 (2010).

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.