Useful combination of massive parallel long-read technology (Roche/454) and short-read technology (Illumina) opens new ways for identification and quantitative analysis of mRNA transcripts and helps understanding gene expression.

The highly dynamic development of sequencing technologies, earlier powered from whole genome sequencing, opens new approaches to analysis. The wide differences of the currently available techniques provide a great opportunity for transcriptional analysis. In this application note we would like to introduce a technique developed in-house that combines short and long read technologies for expression profiling which will vastly improve your sequencing and analysis results.

**Qualitative Or Quantitative Transcription Analysis?**

The analysis of the RNA transcriptome of any specific organism distinguishes mostly between two principal questions. On the one hand, one would like to find out which transcripts are present at a given time and at a given physiological background. For this purpose, it is favourable to generate long assembled cDNA sequences, since these can be used later for the prediction of different protein functions ("annotation") (qualitative de novo transcriptome analysis).

On the other hand, it is of great interest to study the amount of specific transcripts in a given sample. This analysis would reveal if specific phenotypes correlate with an increased or decreased transcription rate of specific genes that have been induced by dedicated conditions (quantitative expression analysis).

---

**Fig.1** Scheme of the de novo expression analysis with combined sequencing of two different cDNA libraries on different systems. More explanation can be found in the text.
Currently, answering both questions in just one experiment is not possible with standard approaches. As long as third-generation instruments are not able to sequence RNA molecules in a high-throughput approach, we have to answer these tasks in a different way. For transcriptional analysis as well as identification of rare splice-variants, one would need long and well covered mRNA contigs. Using the Roche FLX/FLX+ technology that produces read lengths up to 450 bp is helpful to generate long and precise single reads optimal for the assembly approach (Table 1). In contrast you would need as many reads as possible and a technology that produces only one sequence per transcript if you are interested in a comprehensive and deep analysis of the copy number of given transcripts. Using the HiSeq 2000 system one receives those short reads, however this requires mapping of the reads to an excellent reference genome (Table 1). Mapping is defined as to seek and identify short stretches within the reference sequence by using bioinformatic algorithms. Consequently, the absence of a reliably annotated reference genome is cumbersome for this approach.

Combination Of Known System Strengths
Eurofins Genomics found a way to satisfy both objectives – de novo identification of transcripts and the quantitative expression analysis – in one project. This approach can also be applied to organisms whose annotated reference sequence is still unknown. In fact, the parallel sequencing with both technologies mentioned above (GS FLX/GS FLX+ & HiSeq 2000) and different cDNA synthesis steps permit this kind of analysis (Fig. 1) In order to prove the reliability of this new method we have analysed some closely related organisms. The transcriptomes produced from fungi, like the human pathogenic *Aspergillus fumigatus* can be highly complex. More importantly, it is well known that the automatically annotated reference genome is not reliable for other *Ascomycetes* strains. We could deliver comprehensive transcription profile data for different growth conditions. This strengthens our approach and the idea of combining two systematic approaches (unpublished data AG Krappman ZINF Würzburg).

Preparation Of A Normalised cDNA Reference
To begin with, any number of samples of total RNA will be used to prepare a "random-primed" cDNA gene library that is subsequently sequenced with the GS FLX/GS FLX+ technology (Steps 1-3, Fig. 1). The modal read length is around 450 – 650 bp. But the average read length of the RNA's of different organisms is approximately 1,500 – 2,500 bp. Hence we have to sequence the transcripts in different sub-steps and assemble the reads afterwards as accurately as possible to generate our own reliable cDNA reference data set (Step 3, Fig. 1). Due to the normalisation of the cDNA library preparation only a limited number of reads per transcript is needed, since the expression level of ribosomal, redundant and so-called "house-keeping" genes is strongly levelled (Step 2, Fig. 1). Such, we can also identify low expressed genes accurately.

Sequencing of up to 1.2 million single reads normally lead to more than 5,000 and up to 15,000 long cDNA contigs (> 1,000 bp) after the final de novo assembly. Routinely mRNAs up to 3,000 bp and more can be successfully assembled (Step 3, Fig. 1). In addition we will provide you with up to 100,000 EST contigs, which represent shorter, and uncompleted fragments of mRNA transcripts (Fig. 2).

Preparation Of A 3'-Fragment cDNA Library
Our original starting material is used again to prepare the so called "3'-fragment cDNA library" - without normalisation (Step 4, Fig. 1). During the enrichment the included poly(A)-fragment selection steps ensures that also fragments that are partially degraded at their 5'-end can still be processed successfully. The site-directed sequencing of the individually barcoded 3'-cDNA libraries is performed using a sample pool with the re-
spective number of sequences on the Illumina HiSeq 2000. The instrument will produce up to 180 million reads per channel (read length 1x 100 bp) enabling the distinct identification of the origin of transcripts as well as a high significance in respect to the copy number of each transcript (Step 5, Fig. 1).

**Bioinformatic Data Analysis**

The strategy implies using the assembled cDNA contigs from the FLX sequencing (so called unigene-set) as a reference for the mapping of the Illumina reads (Step 5, Fig. 1). The refinement of the available assembly and mapping algorithms lead to great improvements in the data analysis and therefore in the identification of the relevant transcript copy number. Since it is also possible to identify rare transcripts (GS FLX/GS FLX+) at the same time, the whole procedure is easily applicable without their own annotated reference sequence. Optionnally, we are able to further characterise the given FLX cDNA contigs by applying a protein prediction analysis. The bioinformatic normalisation of read numbers is essential for a comprehensive and comparative analysis of the mapping, since the quantitative deviation of single reads needs to be taken into account (Step 6, Fig. 1). In the final data analysis sheet the results of the 3’ fragment libraries will be compared and this analysis facilitates the interpretation of the data (Fig. 4). As explained in the example above we are able to perform direct comparisons of the expression levels of single samples.

**Reproductive Analysis Method**

The method described here has been already applied successfully for different groups of organisms, e.g. partially characterised marine algae (TransAlgae, Israel), economically interesting agricultural crops like sweet potato (*Ipomoea batatas*) and fungi without any specific or insufficient reference annotation. In the experiment described above we gained initial comprehensive insights in to the adaption of *A. fumigatus* to different nitrogen sources – this adaption goes hand in hand with extensive changes in the fungal transcriptome: in contrast to primary nitrogen sources we found that proteins induce an increased expression of various secrete proteolytic activities that enable the fungi to use nitrogen as a growth substrate.

**Contact**

Dr. Axel Strittmatter  
Eurofins Genomics  
Business Development Manager Next Generation Sequencing  
Tel. +49 8092 8289-972  
Email: axelstrittmatter@eurofins.com

PD Dr. Sven Krappmann  
Zentrum für Infektionsforschung, Würzburg  
Email: sven.krappmann@uni-wuerzburg.de

GS FLX/GS FLX+ are trademarks of Roche; Illumina and HiSeq are trademarks of Illumina, Inc.