RNA-Seq: a revolutionary tool for transcriptomics

Biotechnology seminar, 27.05.2015

Sebastian Groß

The Transcriptome is the whole set of transcripts and their quantity in a cell, tissue, organ or a whole organism at a particular time, in a specific developmental stage or under special physiological conditions. The analysis of the transcriptome is useful for the interpretation of functional elements of the genome, to understand development and diseases or to determine the transcriptional structure of genes, like transcription start sites, 5' and 3' ends, posttranslational modifications and splicing processes. Therefore, the key aim of transcriptomics is to catalogue all species of transcripts, such as mRNAs, small interfering RNAs or non-coding RNAs. One technique for transcriptomics is RNA-Seq, which is praised as a revolutionary tool (Wang *et al.*, 2009).

RNA-Seq is a whole transcriptome shotgun sequencing method that uses next-generation sequencing (NGS) methods, like either Illumina sequencing by synthesis, 454-Pyrosequencing or SOLiD sequencing by ligation. The NGS step follows several preparation steps, as for example RNA isolation and preparation.

First, RNA has to be isolated (see Fig. 1; attachment). Thereby, after cell harvesting, guanidinium thiocyanate are added to prevent RNA degradation due to RNAse inhibition (Chomczynski & Sacchi, 1987). After the addition of chloroform and phenols, which lower the risk of RNA degradation as well, and a centrifugation step, the RNA will be solved in the aqueous phase. The RNA isolation process is finished by several washing steps, e.g. with alcohols like isopropanol.

Second, the isolated RNA has to be prepared for cDNA library construction. The RNA preparation steps depend on the following NGS method. Typically rRNA is depleted, reverse transcribed, indexed (adaptor attachment to one or both ends) and pooled before library construction and final sequencing steps. Thus, ones have to

decide whether to first index and pool the samples, which is then followed by the aforementioned steps (see Fig. 2; attachment). Alternatively, ones could treat every sample separately for every step before pooling, which may lead to an increased technical variation between the samples (Mortazavi, 2015).

Third, cDNA library construction is followed by NGS, whereat in this talk exclusively the SOLiD (sequencing by oligonucleotide ligation and detection; LifeTechnologies (Applied Biosystems)) technology was discussed. SOLiD bases on sequencing by ligation. In a first step cDNA constructs are clonally amplified in an emulsion PCR similar to 454-Pyrosequencing (see Fig. 3; attachment). After cross-linking the beads to a glass-slide, the proper sequencing step takes place. Therefore, interrogation probes (di-base probes) are used. Di-base probes consist of eight nucleotides whereof the first two nucleotides are known (see Fig. 4; attachment). Additionally, not only three degenerate bases, but also three universal bases (e.g. 5-nitroindol), coupled with specific fluorescent dyes, are part of di-base probes. The fluorescent dye color depends on the first two known nucleotides (see "Color-Space Coding" in Fig. 4). The enzyme ligase ligates di-base probes that fit to the unknown template cDNA with the help of a primer that binds the beads adaptor sequence. After fluorescence detection, the universal bases are cleaved together with the fluorescent marker. As a result, the 5' phosphate group becomes accessible for the next di-base probe ligation. The fact that three out of five bases in the ligated di-base probes are unknown requires the denaturation of the existing cDNA strands after a specific number of di-base probe ligations. A second ligation cycle is then performed with a primer that is one nucleotide shorter compared to the primer one ligation round before. Therefore, one nucleotide position further is sequenced in the following round. These ligation rounds are repeated either until the whole template is sequenced or until a specific number of di-base probes are ligated.

RNA-Seq has several benefits compared to other techniques in transcriptomics, which is described in the following section. Thus, RNA-Seq does not rely upon knowledge about the genome sequence (Wang *et al.*, 2009). In comparison, DNA microarray, which is the most frequently used transcriptomic technique, requires genome information to enable oligonucleotide synthesis for microarray chip

production. Consequently, RNA-Seq is attractive for non-model organisms too. Furthermore, RNA-Seq shows a low background signal due to the possibility to map the cDNA sequences to unique regions in the genome (Wang *et al.*, 2009). Likewise, RNA-Seq enables the identification of alternative spliced RNA isoforms, antisense transctipts and fusion genes by mapping the transcripts to the genome sequence (Ozsolak & Milos, 2011). As a result, RNA-Seq does not rely upon genome sequence information, but, nevertheless, genomic information are useful for these purposes. Moreover RNA-Seq has no upper quantification limit and a large dynamic detection range (Wang *et al.*, 2009). In contrast, DNA microarray shows a detection limitation due to the number of fixed oligonucleotides on a microarray chip. Finally, RNA-Seq shows a high reproducibility both for technical and biological replicates (Wang *et al.*, 2009).

In spite of all the benefits, some challenges still remain for RNA-Seq. Although there are just a few steps in RNA-Seq, there are still some manipulation stages, such as PCR amplification, RNA fragmentation and reverse transcription (Ozsolak & Milos, 2011). An ideal method for transcriptomics should be able to directly identify and quantify all RNAs. The direct RNA sequencing (Ozsolak *et al.*, 2009) method avoids the reverse transcription and PCR amplification, but the problem with fragmentation of large RNA molecules (e.g. to identify splicing patterns in eukaryotic transcriptomes) still remains. Another point which should be mentioned is that an increase in sequencing depth, which could be reached by a longer read length, would lead to a greater coverage and more significant data. Finally, RNA-Seq also faces bioinformatical challenges (Wang *et al.*, 2009). High-throughput sequencing approaches generate a high amount of data that have to be processed. Therefore, two challenges are to reduce errors in image analysis and to remove low-quality reads.

In the end, RNA-Seq is one possible method for transcriptomics that shows a high potential to improve the understanding of development and deseases. Nevertheless, it is assumed that RNA-Seq does not replace techniques like DNA microarrays, because it takes (at the moment) a much longer time to perform a RNA-Seq experiment.

Sources

Z Wang, M Gerstein, M Snyder (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics: Vol. 10, 57-63.

F Ozsolak, M Milos (2011) RNA sequencing: advances, challanges and opportunities. Nature Reviews Genetics: Vol. 12, 87-98.

A Mortazavi (2015) Better together: multiplexing samples to improve the preparation and reliability of gene expression studies. Nature Methods: Vol. 12, No. 4, 304-305.

ML Metzker (2010) Sequencing technologies - the next generation. Nature Reviews Genetics: Vol. 11, 21-46.

F Ozsolak, AR Platt, DR Jones, JG Reifenberger, LE Sass, P Mclnemey, JF Thompson, J Bowers, MJ Milos, PM Milos (2009) Direct RNA sequencing. Nature: Vol. 468, 814-818.

P Chomczynski, Nicoletta Sacchi (1987) Single-Step Method of RNA Isolation by Acid Guanidinium Thiocyanate-Phenol-Chloroform Extraction. Analytical Biochemistry: 162, 156-159.

HE Peckham, SF McLaughlin, JN Ni, MD Rhodes, JA Malek, KJ McKernan, AP Blanchard. SOLiD Sequencing and 2-Base Encoding. Applied Biosystems (Poster 2624).

http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/documents/generaldocuments/cms_057810.pdf (27.05.2015)

http://www.biochem.arizona.edu/classes/bioc471/pages/Lecture9/AMG5.1a.gif (19.05.15)

http://image.slidesharecdn.com/htsmasterstalk-111109074414-phpapp02/95/highthroughput-sequencing-10-728.jpg%3Fcb%3D1320824865 (20.05.15)

http://ll1.workcast.net/10403/4934292334051602/images/10403_rna_img.jpg (22.05.15)

Attachment

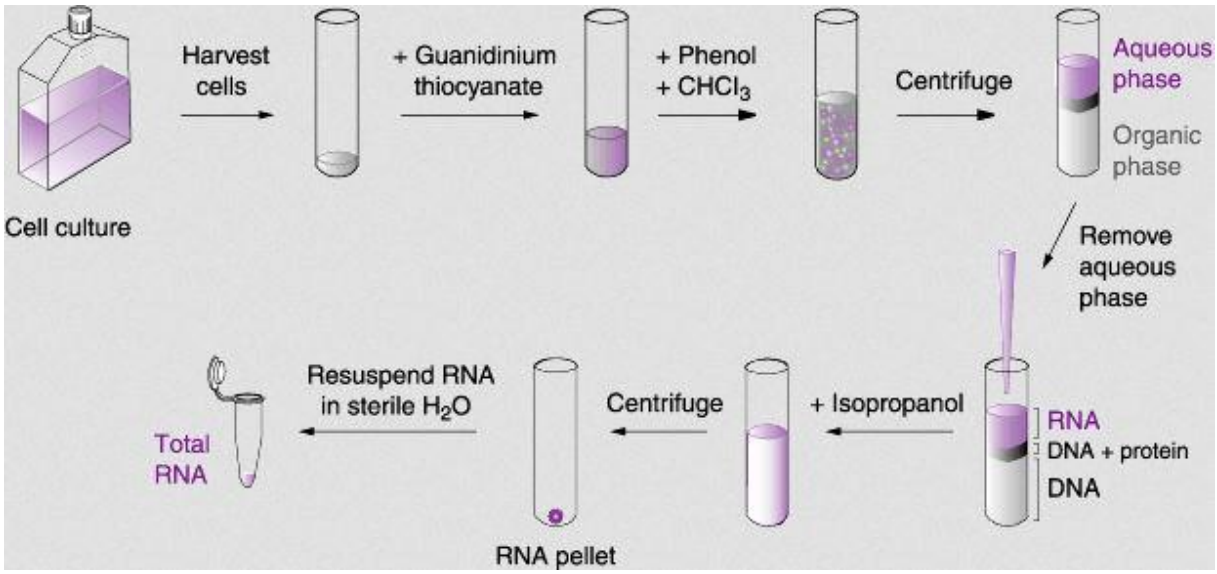

Fig. 1: Scheme of working steps in RNA isolation.

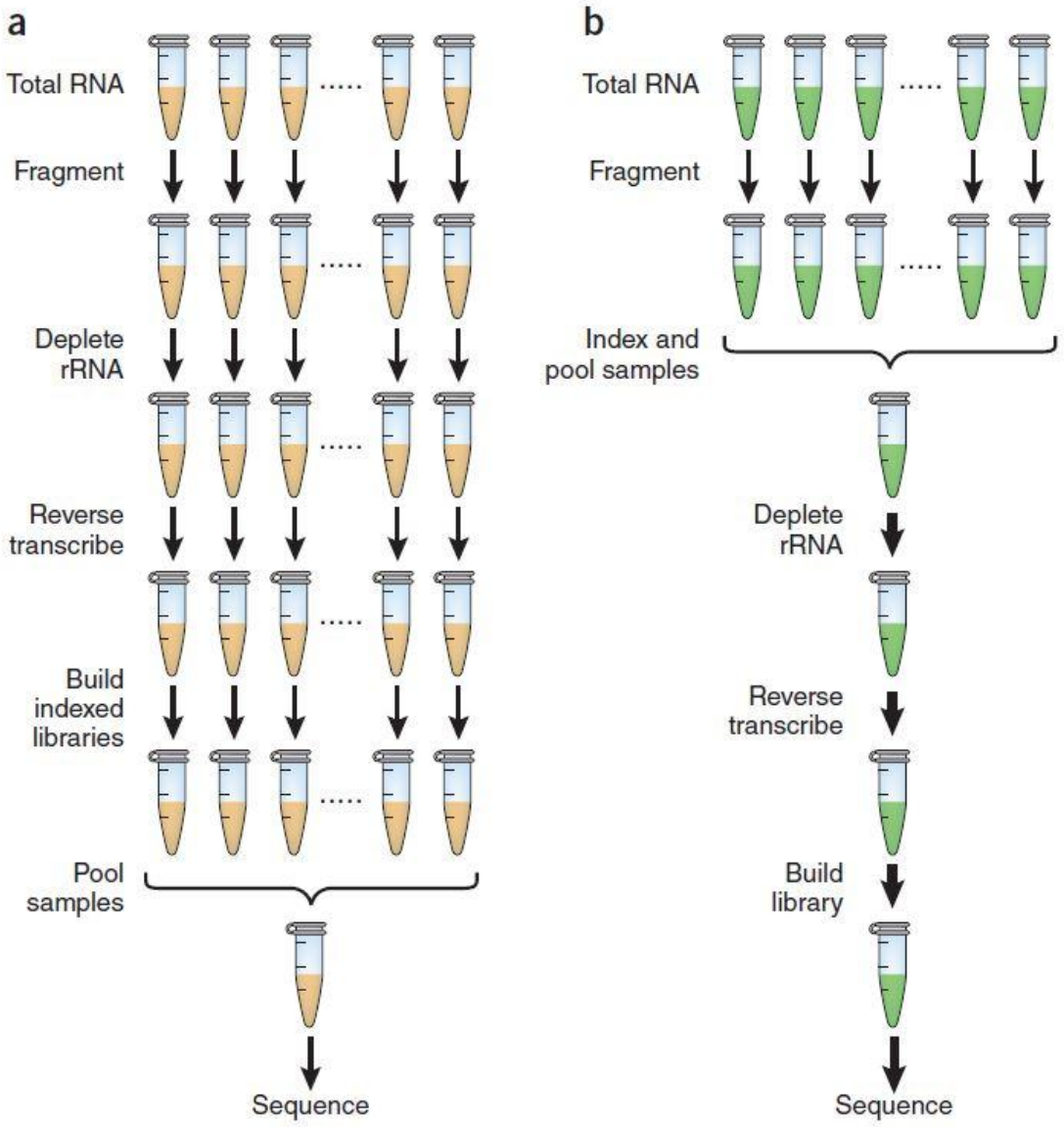

Fig. 2: Different ways to treat before RNA-Seq experiments.

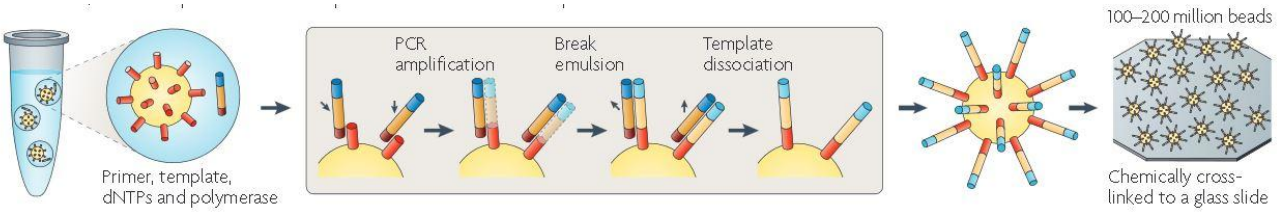

Fig. 3: Emulsion PCR before SOLiD.

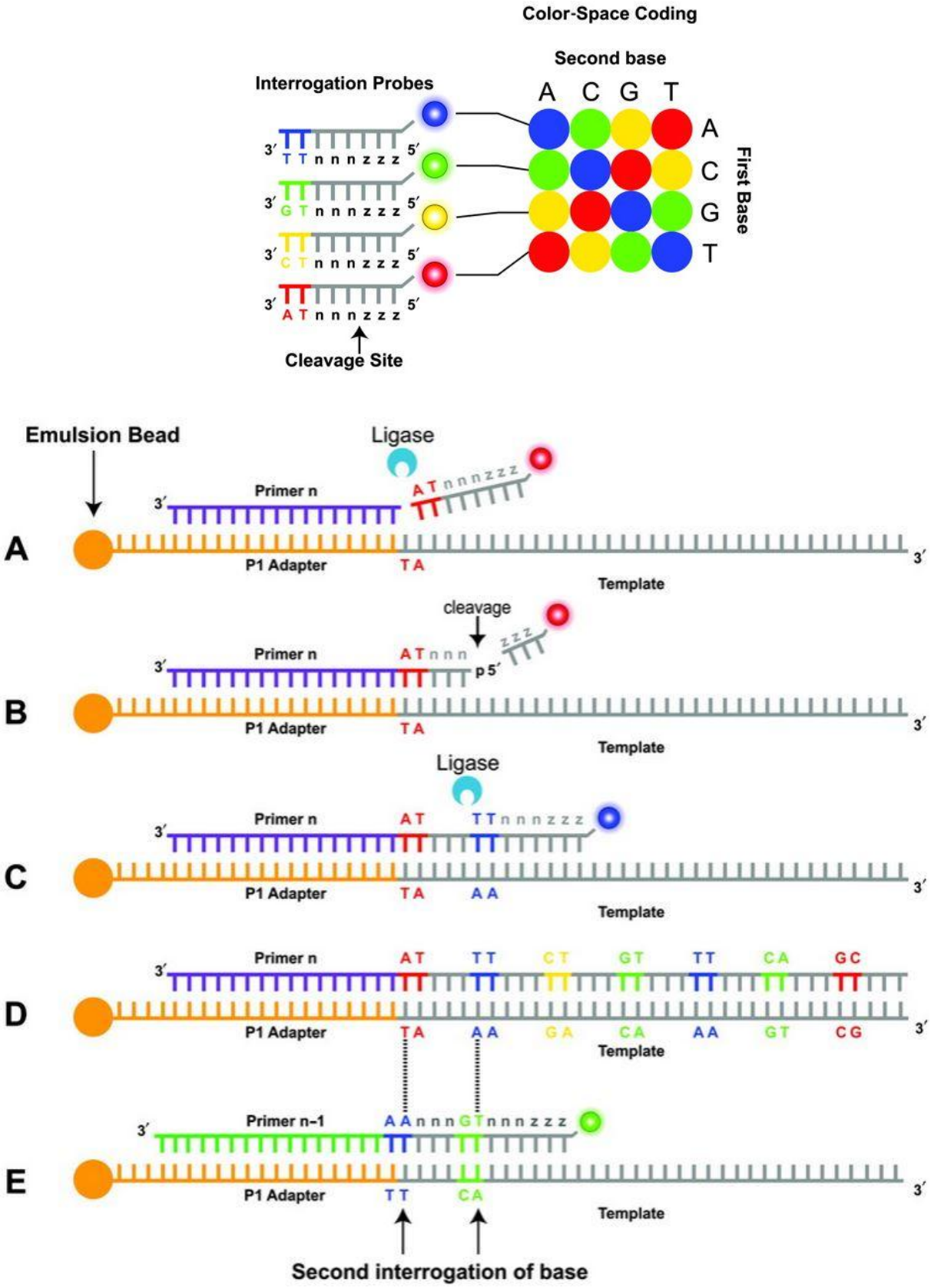

Fig. 4: SOLiD sequencing process.