BioTechnologia vol. 92(4) • pp. 311-319 • 2011
Journal of Biotechnology, Computational Biology and Bionanotechnology PERSPECTIVES

Transcriptome sequencing: next generation approach to RNA functional analysis

AGNIESZKA ŻMIEŃKO [1*], PAULINA JACKOWIAK [1], MAREK FIGLEROWICZ [1,2]

[1] Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznań, Poland

[2] Institute of Computing Science, Poznan University of Technology, Poznań, Poland

* Corresponding author: akisiel@ibch.poznan.pl

Abstract

The most general definition of transcriptome assumes that it is a complete set of RNA molecules (transcripts) that are formed in a given organism. In response to exogenous and endogenous stimuli, transcriptome undergoes constant qualitative and quantitative changes. Monitoring these changes can provide information on how the living organisms function and deal with the changing environmental conditions. Recently, a new approach, called RNA sequencing or RNA-Seq has been applied to transcriptome-oriented studies. It is based on the next-generation sequencing and allows transcript discovery and quantification. This new approach permits identification of novel transcripts and products of their processing or partial degradation. Since RNA-Seq does not require information about the sequence of the reference genome, it can be used for transcriptome analysis in any organism. This feature combined with the high sensitivity and broad dynamic range of RNA-Seq experiments, make them much more versatile for high-throughput studies as compared to DNA microarray technology.

Key words: RNA-Seq, transcriptome, next-generation sequencing, isoforms

Introduction

Transcriptome is a complete set of RNA molecules (transcripts) that are, at a given moment, present in a given organism/organ/tissue/cell. It is a composition of coding and non-coding RNAs and their degradation products. The amount of each RNA molecule is a resultant of the balance among the transcription, the RNA processing and the RNA degradation events. Transcriptome undergoes constant qualitative and quantitative changes which reflect natural physiological processes or are triggered by external stimuli. The analysis of transcriptome composition can therefore provide an insight into the way in which organisms function.

Over the years, several techniques that allow high-throughput identification and/or quantification of transcriptome components have been developed. They include: expressed sequence tag (EST) sequencing, serial analysis of gene expression (SAGE) and DNA microarrays. The last method is based on hybridization of a fluorescently labeled sample that represents transcriptome under investigation to cDNA or – more often – oligonucleotide probes immobilized on a glass (Schena et al., 1995). EST and SAGE methods permit partial determination of a transcript sequence. Provided that a sufficiently large number of sequences are generated, these techniques also enable to digitally measure the level of gene expression (Adams et al., 1991, Velculescu et al., 1995). These methods, however, are laborious and relatively expensive, as they require DNA cloning. Recently, technologies of next-generation sequencing (NGS) have been applied for transcriptome analysis. This approach, called RNA-sequencing or RNA-Seq, has instantly revolutionized the field. The RNA-Seq technique does not involve cloning steps and enables simultaneous sequencing of millions of molecules. It produces unprecedented amounts of data, providing the possibility to investigate many transcriptomes at once, within days or weeks. As transcriptome sequencing does not necessarily require prior knowledge of the genome sequence, it is applicable to practically any organism and allows transcript discovery and not only analysis.

The aim of this short review is to present the basics of RNA-Seq methodology and its possible applications. We will briefly describe a typical RNA-Seq experiment,

focusing on mRNA, as well as highlight some important issues that should be taken into consideration during the study design. We will also discuss the advantages and drawbacks of the RNA-Seq methods referring to microarrays. It should be mentioned that although we mainly focus on Illumina sequencing by synthesis (SBS) technology, several aspects raised here apply to other sequencing platforms as well.

Overview of an RNA-Seq experiment

A typical RNA-Seq experiment workflow is presented in Figure 1. The process can be divided into three main stages: library construction, DNA sequencing and sequence data analysis. A first critical issue of the RNA-Seq experiment, however, is the preparation of high quality RNA. Before continuing with the library preparation, RNA purity and integrity should be checked. As the process of generating RNA-Seq library includes reverse transcription, any reverse transcriptase inhibitors remaining in the RNA sample after isolation, will affect the downstream steps of the protocol. Therefore, it is crucial to select a method that is proven to produce high-quality RNA, based on literature and/or user's own experience.

Library construction

Most current next-generation sequencing instruments generate reads significantly shorter than standard Sanger sequencing. In case of Illumina SBS technology, reads are 35-150 nucleotide long. The template length must be therefore adjusted accordingly to permit uniform transcript coverage by the sequence reads. This is typically achieved by chemical fragmentation of RNA prior to cDNA synthesis (Mortazavi et al., 2008). Fragmented RNA is then reverse transcribed into cDNA, followed by a second strand synthesis and end-modification steps. The latter permits subsequent ligation of oligonucleotide adapters to both DNA ends (Fig. 2A). The adapters include annealing sites recognized by primers used for cluster generation (P5 and P7) and sequencing (Seq P1 and Seq P2). They also contain a short region, called index, which can be sequenced independently of the template DNA, using the Index primer. The index identifies the adapter used for library preparation. If different adapters (with different index sequences) are ligated to different DNA libraries, those libraries can be mixed and subjected to multiplexed sequencing in one lane of a flow cell (see DNA sequencing section). Optionally, the DNA library may be PCR-amplified to selectively enrich the DNA fragments that have adapter molecules ligated to both ends (Fig. 2B). Finally, the DNA quantity and quality is validated by capillary electrophoresis, spectrophotometry and/or qPCR, before proceeding to the next stage.

This basic Illumina protocol allows to generate a library that does not maintain information on the orientation of the original RNA strand. As a consequence, it is impossible to distinguish between sense and antisense transcripts generated from the same genomic locus. However, in many applications, evaluating the rate of antisense transcription is essential to precisely determine the amount of actual sense transcript, to define transcription initiation and termination points or to get insight into the antisense transcription rate itself. In such cases, information on the orientation of RNA template is a prerequisite. Similarly, RNA-Seq experiments focused on RNA from prokaryotes and lower eukaryotes, whose genes are densely located and often overlapping, will benefit from such knowledge. Therefore, numerous strand-specific protocols have been developed (reviewed and evaluated in Levin et al., 2010). Often, they rely on the sequential attachment of specific adapters to each end of an RNA template in a known orientation (Lister, et al. 2008). Other methods are based on strand marking by chemical modifications, for example by incorporating dUTPs into the second cDNA strand, followed by its selective removal (Parkhomchuk et al., 2009). Although, non-standard and still in development, the strand-specific RNA-Seq methods have already proved to be more accurate in transcript structure analysis and expression quantification (Wang et al. 2011).

DNA sequencing

In the Illumina sequencing workflow, all further steps of DNA preparation and sequencing occur on a flow cell – a plate that is physically divided into eight separate channels, called "lanes". The surface of each lane is densely and uniformly covered by covalently attached short oligonucleotide primers P1 and P7 that are complementary to binding sites in the adapters, previously ligated to the DNA (Fig. 2C). The templates are hybridized to these primers, copied and removed. The immobilized DNA copies are further amplified in the process called isothermal bridge amplification (Bent

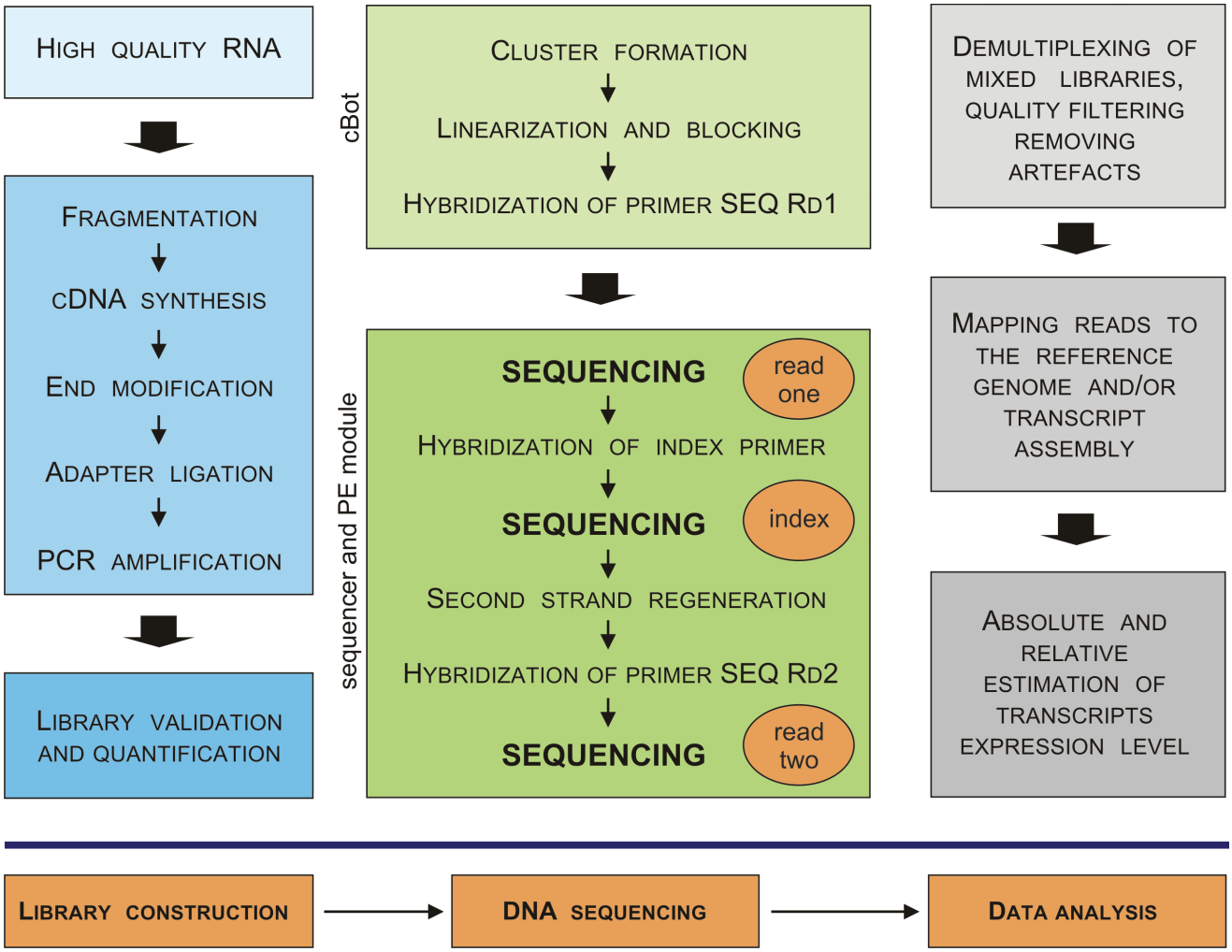

Fig. 1. Overview of RNA-seq experiment, according to typical Illumina sequencing workflow. DNA library is prepared from RNA templates and used for cluster formation. Clusters are generated on the surface of a flow cell, in a cBot instrument. The flow cell is then transferred to the sequencer where the template DNA is sequenced (read one, followed by index read in case of multiplexing, followed by read two in case of paired-end sequencing). The flow cell is kept in the sequencer along all sequencing steps; all DNA template preparation procedures for index read and read two are performed in there, using attached Paired-End (PE) Module. Generated data are then analyzed using appropriate software

ley et al., 2008, www.illumina.com) – Figure 2D. Bridge amplification is performed in an automated system called cBot. High fidelity DNA polymerase is used in this reaction to minimize the rate of mis-incorporation errors. As a result, each template DNA molecule gives rise to about two thousand identical double-stranded DNA copies, attached to the flow cell surface and tightly clustered in a very small area (about one micron in diameter). In this way, millions of dense clonal clusters are generated on each lane. In the next step, DNA is linearized by denaturation, followed by cleavage and removal of the DNA strand with P5 binding site on its 5′ end (Fig. 2E). The 3′ ends of the remaining DNA strand and the flow cell-bound oligonucleotides are blocked to prevent unintended extension. After hybridization of the sequencing primer (Seq P1, Fig. 3A) to the complementary adapter located at the unbound end of each template, the flow cell is ready for sequencing in the sequencer.

The sequencing process lasts for a predefined number of cycles, which corresponds to the expected length of a sequence read. In each SBS cycle, the flow cell lanes are flushed with the equimolar mixture of four deoxyribonucleotide triphosphates (dNTPs) and the hybridized sequencing primer is extended by high fidelity DNA polymerase (Bentley et al., 2008). Each dNTP (A, C, G, T) is labeled with a different fluorescent dye which also

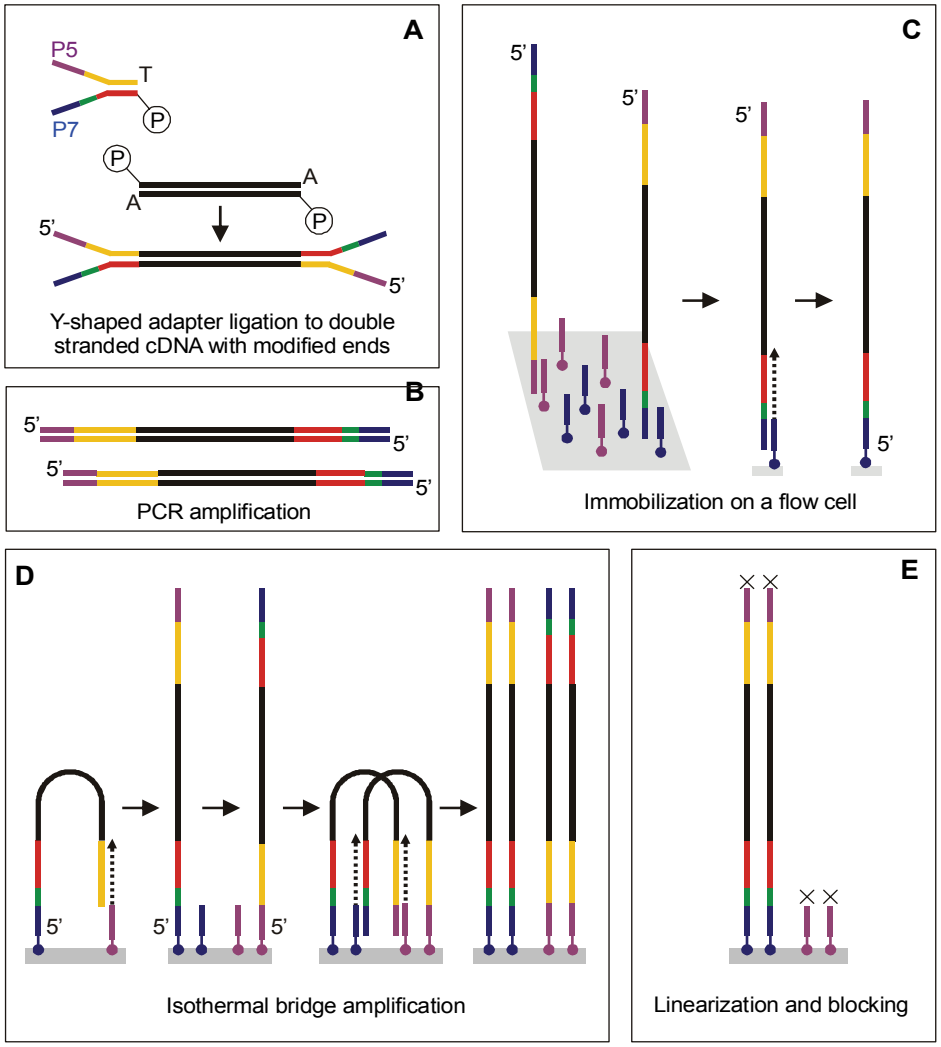

Fig. 2. Preparation of DNA templates for sequencing. A) Y-shaped adapters are ligated to double-stranded cDNA with modified ends. Resulting DNA contains hybridization sites for all sequencing primers. B) Optionally, the PCR amplification-based enrichment for DNA molecules with adapters on both ends may be performed. C) After denaturation, each single-stranded DNA molecule is hybridized to the flow cell-bound primers and primer is extended. D) The immobilized DNA copies are amplified by isothermal bridge amplification, resulting in formation of dense clonal clusters. E) After denaturation and removal of the reverse strand, the 3′ ends of the remaining strand and flow cell-bound primers are blocked. Each cluster contains identical DNA sequences that are ready for sequencing

serves as an elongation terminator. As a result, the newly synthesized DNA strands can be elongated by only one nucleotide per cycle and the added nucleotides can be identified by imaging of laser-excited dye fluorescence and measuring signal intensities. After the imaging step, the dyes are enzymatically cleaved and removed, so the synthesized strand can be extended by one base again in the next cycle. The process of base-by-base data collection enables robust DNA sequencing, including homopolymers and repetitive sequences. Depending on the application, up to 150 sequencing cycles may be performed and the data are collected simultaneously for each of millions of clusters present on the flow cell. If mixed libraries are sequenced, the extended primer Seq P1 is removed after the read is complete, and Index primer is hybridized to enable index sequencing (Fig. 3B).

For some applications, it is beneficial to sequence the DNA templates from both ends, which is called paired-end sequencing. In this case, after generation of the read one (and optionally, index read) products, the synthesized fragments are removed, the flow cell-bound oligonucleotides are de-blocked and the comple-

mentary strand of each DNA is resynthesized. After removal of the original strand, read two is generated with the SEQ P2 primer (Fig. 3C).

Data analysis

The sequences are deciphered in real-time by appropriate software, therefore soon after the run is finished, data analysis can be commenced. The dataset should be filtered for high quality sequences. Although the obtained data do not always meet the required standards, still, millions or billions of sequences (reads), each generated from an individual DNA cluster, are kept for downstream analysis. Typically, the reads are mapped to the reference genome or transcriptome. Next, the reads that

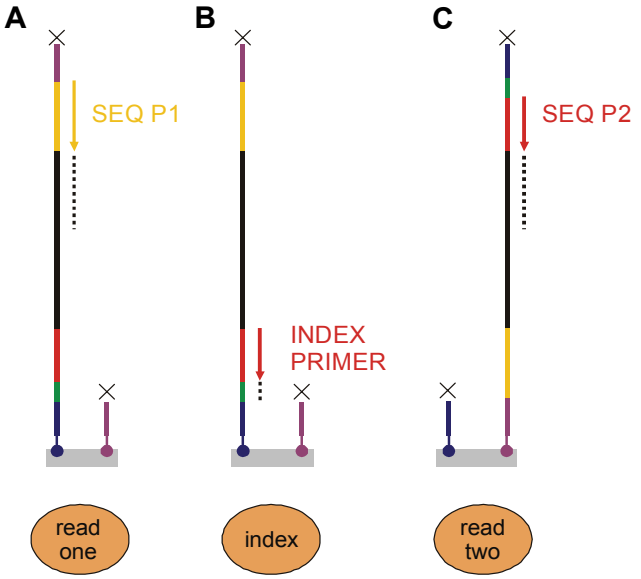

Fig. 3. Multiplexed paired-end sequencing process. A) Read one is generated from one end of the DNA template by primer Seq P1 extension. B) Read one product is removed and Index primer is annealed to the binding site in the same strand, to produce 6-nt index read. C) The existing template is used to regenerate the complementary strand, after which the original strand is removed. The new strand is used as a template for read two, primed by primer Seq P2

align to the same locus are assembled into transcripts. Finally, the expression level of each transcript is estimated by counting the number of reads that aligned to this transcript. To enable data comparison, the values are normalized by taking into account the fact that the number of reads per transcript depends on the transcript length and the total number of reads obtained for a particular library (Mortazavi et al., 2008). Specific software exists for each of the analysis steps and many methods are available as free or open-source programs. The current state of knowledge regarding RNA-Seq data analysis has recently been reviewed and the efficiencies of most popular methods have been compared (Garber et al., 2011; Martin and Wang, 2011) therefore this problem will not be discussed in detail here.

RNA-Seq: applications and comparison with microarrays

Assuming that: (i) the sequencing library has been prepared from high-quality RNA, (ii) a large enough number of short reads has been generated, (iii) the method is free from sequence biases and provides uniform coverage of all transcripts, and that (iv) adequate computational tools are available, it is theoretically possible to fully reconstruct the whole transcriptome in particular conditions, using the sequence data. This promises to make RNA-Seq a powerful and versatile technology with several possible applications, some of which are briefly described below. In this context, advantages of RNA-Seq over DNA microarrays are also discussed.

Analysis of gene expression

Intuitively, the first RNA-Seq utilization that comes into mind is a simple comparison of gene expression. Owing to the large amount of sequence data, reads mapping to particular genes can be counted and compared. A comparison of the normalized number of reads between the samples allows to estimate the relative transcript abundance. If a reference genome is available and there is no need to distinguish various transcript isoforms, this approach is pretty straightforward and similar to profiling gene expression with DNA microarrays (Żmieńko et al., 2008). Yet there are several differences between these two methods that must be carefully considered. NGS enables direct quantification of transcripts, represented by short reads. In the case of microarrays, the quantification of transcripts is indirect and based on the assumption that the intensity of the signal emitted from the individual DNA spot corresponds to the amount of labeled sample that hybridized to a microarray probe. This presumption is an obvious simplification and often makes accurate transcript quantification a challenge. A dynamic range of microarrays is far narrower than that of NGS and a simultaneous data collection for genes with

large differences in expression levels is often impossible with this technique. Adjusting microarray scanner settings to higher sensitivity will cause the level of abundant transcripts to be underestimated due to signal saturation. On the other hand, lowering sensitivity can make rare transcripts undetectable or incorrectly quantified, due to background fluorescence. Moreover, the position of the probe target site in the transcript and its actual sequence may affect hybridization strength and specificity, influencing the microarray results. Last but not least, the microarray analysis is always limited to a predefined number of genes or isoforms, represented by the microarray probes. A great advantage of NGS is the possible identification and quantification of new transcripts (Wang et al., 2009; Trapnell et al., 2010). Still, DNA microarrays are much cheaper than NGS and may be advantageous in some cases. Also, what is often stressed, the biases and limitations of the microarray technology are well recognized and standard computational solutions are available to overcome these limitations. In contrast, NGS-related problems are just being discovered. Some platform-specific biases have already been described in the scientific literature. In the case of Illumina sequencers, unequal coverage has been reported (Harrismendy et al., 2009; Kozarewa et al., 2009) as well as increased rates of sequence errors at first and especially last cycles (Kircher et al., 2009) or in the GC-rich regions (Dohm et al., 2008). Nakamura et al. (2011) observed that some miscalls are sequence-specific and are triggered by inverted repeats and GGC sequences. The authors propose that the presence of such sequences prevents base incorporation at a given cycle, which results in read dephasing. As the number of sequencing data in the public databases grows rapidly, it can be expected that some general rules will be applied to such observations in the near future and eventually some standard methods of bias correction will be developed.

De novo transcriptome assembly

Design of oligonucleotide microarrays depends on the knowledge of a reference sequence. For several organisms, the microarray usage is restricted to cross-species hybridization only, where the probes represent genes of a related species. This approach adds complexity to the microarray data analysis and is far less sensitive than single-species hybridization (Bar-Or et al., 2006). The lack of reference genome is not an issue for RNA-Seq, provided that a sufficiently large number of reads are available. Transcripts can be assembled *de novo*, by direct joining of the overlapping reads into longer contigs. However, it requires a lot of computing resources, especially in the case of large eukaryotic transcriptomes. Furthermore, the amount of sequence data needed for *de novo* assembly is about three times larger than the one needed for the reference genome-guided assembly. Thus, the use of *de novo* assembly significantly raises sequencing costs. Despite these disadvantages, successful *de novo* transcriptome assembly has been recently performed for lower and higher eukaryotes (Martin et al., 2010; Garg et al., 2011; Grabherr et al., 2011; Jäger et al., 2011). Often, the two strategies (reference genome-based and *de novo*) are combined by first mapping the data to the reference sequence and further *de novo* assembly of reads that failed to align. Alternatively, shorter *de novo* assembled transcripts can be aligned to the reference genome or to a closely related genome, to attempt reconstructing full-length transcripts.

Transcriptome reconstruction can be aided by paired-end reads. In this sequencing approach, each DNA fragment is sequenced from both ends and the information on the paired sequences as well as expected distance between them is used in transcriptome mapping and assembly (Fig. 4). This information proves to be especially helpful for mRNA isoform detection and for estimating the relative frequency of their occurrence (other than expected, the distance between the paired reads can indicate an occurrence of a new splicing variant, even when the reads do not directly map to this site). However, with the paired-end protocol, a lower read coverage is obtained than with a single-read protocol, at the same cost. It can affect transcript quantification as large number of reads has been shown to give more accurate transcript abundance estimates (Li and Dewey, 2011).

Alternative splicing analysis

The ability to distinguish isoforms of the same transcripts and to detect new splicing variants is a key difference between microarrays and RNA-Seq. While genome-wide tiling arrays can provide a lot of information on this issue, they are available only for a limited number of organisms or represent just a part of the genome. Splicing arrays, having their probes positioned at

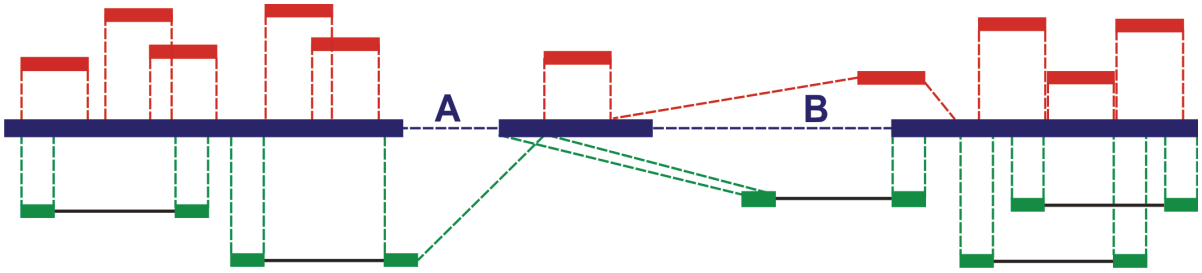

Fig. 4. Detection of splice sites and transcript quantification by single-read and paired-end sequencing approaches. A hypothetical transcript with two junction sites (denoted A and B) is shown (blue), sequenced by single-read (100 base-reads, shown in read) and paired-end (2 × 50-base reads, generated from both ends of DNA fragments, shown in green) approaches. Identification of splice sites and actual transcript sequence requires more single-reads for the splice sites to be covered. With paired ends, both read one and read two are used in the transcript assembly, as well as the information on the expected distance between them (shown as black solid line). The total number of sequenced bases is equal in two approaches but the number of reads/fragments is different. In this example, splice site A is detected only with the paired-end approach. On the other hand, a single-read approach results in higher read coverage, allowing more precise transcript quantification

the exon-exon junctions, are designed to identify and discriminate predefined splicing events, but not new ones. RNA-Seq, supported by sophisticated computational tools, like *de novo assembly* methods, permits not only transcript quantification, but also isoform identification, detection of new splicing sites and previously unknown exon combinations as well as gene fusions. A recent work, aimed at creating a comprehensive catalogue of alternative splicing sites in yeast, is a great illustration of substantial differences between the two technologies. The experiment included a comparison of RNA-Seq and exon junction microarray data (the microarrays were designed using available mRNA/EST sequences) (Ramani et al., 2011). RNA-Seq was shown to reliably detect known splicing variants – more than 90% of identified variants with only a single read spanning the spliced site were confirmed to be real. Moreover, RNA-Seq analysis allowed discovery of more than 8600 putative new splicing variants encoded by more than 2600 genes. Most of them were generated on the basis of new combination of previously known splice sites or contained one new splice site. However, about 12% of detected junctions connected two novel splice sites. A fraction of them (25 junctions) was assayed by RT-PCR and all were validated.

SNP detection

Single nucleotide polymorphisms (SNPs) are one-nucleotide variances throughout genome sequence observed among individuals. They are used as markers in population and genome-wide association studies. There are numerous examples of SNPs (alleles) being associated with genomic disorders. The observed associations may result from the physical proximity of SNP position to the disrupted gene. However, another possibility is that SNP is located in the gene coding sequence or it can itself account for the disease. Therefore, one can attempt to narrow down the analyzed region to the gene coding parts, as the variances located there are more likely to serve as trait-associated markers for complex phenotypes (Nicolae et al., 2010). Focusing on the transcribable portion of the genome only significantly reduces the amount of sequence data. The saved "sequencing space" can be used for increasing the read coverage (which helps to distinguish real SNPs from sequencing errors) and/or number of individuals tested, at comparable costs. Unlike DNA microarrays, NGS enables SNP discovery and analysis in species whose genomes are still not sequenced. Actually, SNP discovery by RNA-Seq is an effective preliminary step towards the construction of a SNP array for a non-model species (Liu et al., 2011). Consequently, transcriptome sequencing has been successfully applied to characterize single nucleotide markers in several organisms, using various NGS platforms (Vera et al., 2008, Trick et al., 2009, Parchman et al., 2010).

Conclusion and perspectives

Due to rapid technological development, researchers working in the field of functional genomics have been provided a tool that enables them to perform a detailed investigation of practically any transcriptome. Free from

many limitations of DNA microarray-based methods, RNA-Seq opens the way to analyze transcriptomes at a scale and resolution that were unattainable just a few years ago. As sequencing costs drop and personal sequencers are introduced to the market, RNA-Seq becomes not only popular but also, frequently, an indispensable analytical tool. Therefore, it can be expected that new RNA-Seq applications will be developed by growing NGS-users community to support transcriptome-oriented studies.

Acknowledgments

We are grateful to dr Zbigniew Michalski for help with the figure preparation. We acknowledge funding from the National Science Centre, grant 2011/01/B/NZ2/04816.

References

Adams M.D., Kelley J.M., Gocayne J.D., Dubnick M., Polymeropoulos M.H., Xiao H., Merril C.R., Wu A., Olde B., Moreno R.F. et al. (1991) *Complementary DNA sequencing: expressed sequence tags and human genome project*. Science 252: 1651-1656.

Bar-Or C., Bar-Eyal M., Gal T.Z., Kapulnik Y., Czosnek H., Koltai H. (2006) *Derivation of species-specific hybridization-like knowledge out of cross-species hybridization results*. BMC Genomics 7: 110.

Bentley D.R., Balasubramanian S., Swerdlow H.P., Smith G.P., Milton J., Brown C.G., Hall K.P., Evers D.J., Barnes C.L., Bignell H.R. et al. (2008) *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature 456: 53-59.

Delseny M., Cooke R., Raynal M., Grellet F. (1997) *The Arabidopsis thaliana cDNA sequencing projects*. FEBS Lett. 405: 129-132.

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

Garber M., Grabherr M.G., Guttman M., Trapnell C. (2011) *Computational methods for transcriptome annotation and quantification using RNA-seq*. Nat. Meth. 8: 469-477.

Garg R., Patel R.K., Tyagi A.K., Jain M. (2011) *De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification.* DNA Res. 18: 53-63.

Grabherr M.G., Haas B.J., Yassour M., Levin J.Z., Thompson D.A., Amit I., Adiconis X., Fan L., Raychowdhury R., Zeng Q. et al. (2011) *Full-length transcriptome assembly from RNA-Seq data without a reference genome.* Nat. Biotechnol. 29: 644-652.

Harismendy O., Ng P.C., Strausberg R.L., Wang X., Stockwell T.B., Beeson K.Y., Schork N.J., Murray S.S., Topol E.J., Levy S., Frazer K.A. (2009) *Evaluation of next generation sequencing platforms for population targeted sequencing studies.* Genome Biol. 10: R32.

Jäger M., Ott C.E., Grünhagen J., Hecht J., Schell H., Mundlos S., Duda G.N., Robinson P.N., Lienau J. (2011) *Composite transcriptome assembly of RNA-seq data in a sheep model for delayed bone healing*. BMC Genomics 12: 158.

Kircher M., Stenzel U., Kelso J. (2009) *Improved base calling for the Illumina Genome Analyzer using machine learning strategies.* Genome Biol. 10: R83.

Kozarewa I., Ning Z., Quail M.A., Sanders M.J., Berriman M., Turner D.J. (2009) *Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes.* Nat. Meth. 6: 291-295.

Levin J.Z., Yassour M., Adiconis X., Nusbaum C., Thompson D.A., Friedman N., Gnirke A., Regev A. (2010) *Comprehensive comparative analysis of strand-specific RNA sequencing methods.* Nat. Meth. 7: 709-715.

Li B., Dewey C.N. (2011) *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome*. BMC Bioinform. 12: 323.

Lister R., O'Malley R.C., Tonti-Filippini J., Gregory B.D., Berry C.C., Millar A.H., Ecker J.R. (2008) *Highly integrated single-base resolution maps of the epigenome in Arabidopsis*. Cell 133: 523-536.

Liu S., Zhou Z., Lu J., Sun F., Wang S., Liu H., Jiang Y., Kucuktas H., Kaltenboeck L., Peatman E., Liu Z. (2011) *Generation of genome-scale gene-associated SNPs in catfish for the construction of a high-density SNP array*. BMC Genomics. 12: 53.

Martin J., Bruno V.M., Fang Z., Meng X., Blow M., Zhang T., Sherlock G., Snyder M., Wang Z. (2010) *Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads.* BMC Genomics. 11: 663.

Martin J.A., Wang Z. (2011) *Next-generation transcriptome assembly.* Nat. Rev. Genet. 12: 671-682.

Mortazavi A., Williams B.A., McCue K., Schaeffer L., Wold B. (2008) *Mapping and quantifying mammalian transcriptomes by RNA-Seq*. Nat. Meth. 5: 621-628.

Nakamura K., Oshima T., Morimoto T., Ikeda S., Yoshikawa H., Shiwa Y., Ishikawa S., Linak M.C., Hirai A., Takahashi H. et al. (2011) *Sequence-specific error profile of Illumina sequencers*. Nucl. Acids Res. 39: e90.

Nicolae D.L., Gamazon E., Zhang W., Duan S., Dolan M.E. et al. (2010) *Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS.* PLoS Genet. 6: e1000888.

Parchman T.L., Geist K.S., Grahnen J.A., Benkman C.W., Buerkle C.A. (2010) *Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery.* BMC Genomics. 11: 180.

Parkhomchuk D., Borodina T., Amstislavskiy V., Banaru M., Hallen L., Krobitsch S., Lehrach H., Soldatov A. (2009) *Transcriptome analysis by strand-specific sequencing of complementary DNA.* Nucl. Acids Res. 37: e123.

Ramani A.K., Calarco J.A., Pan Q., Mavandadi S., Wang Y., Nelson A.C., Lee L.J., Morris Q., Blencowe B.J., Zhen M., Fraser A.G. (2011) *Genome-wide analysis of alternative splicing in Caenorhabditis elegans.* Genome Res. 21: 342-348.

Schena M., Shalon D., Davis R.W., Brown P.O. (1995) *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science 270: 467-470.

Trapnell C., Williams B., Pertea G., Mortazavi A., Kwan G., van Baren M., Salzberg S., Wold B., Pachter L. (2010) *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation*. Nature Biotechnol. 28: 511-515.

Trick M., Long Y., Meng J., Bancroft I. (2009) *Single nucleotide polymorphism (SNP) discovery in the polyploid Brassica napus using Solexa transcriptome sequencing.* Plant Biotechnol. J. 7: 334-346.

Velculescu V.E., Zhang L., Vogelstein B., Kinzler K.W. (1995) *Serial analysis of gene expression.* Science. 270: 484-487.

Vera J.C., Wheat C.W., Fescemyer H.W., Frilander M.J., Crawford D.L., Hanski I., Marden J.H. (2008) *Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing.* Mol. Ecol. 17: 1636-1647.

Wang Z., Gerstein M., Snyder M. (2009) *RNA-Seq: a revolutionary tool for transcriptomics*. Nature Rev. Genet. 10: 57-63.

Wang L., Si Y., Dedow L.K., Shao Y., Liu P., Brutnell T.P. (2011) *A Low-Cost Library Construction Protocol and Data Analysis Pipeline for Illumina-Based Strand-Specific Multiplex RNA-Seq*. PLoS One. 6: e26426.

Żmieńko A., Handschuh L., Góralski M., Figlerowicz M. (2008) *Zastosowanie mikromacierzy DNA w genomice strukturalnej i funkcjonalnej.* Biotechnologia 4: 39-53.