Available online at www.sciencedirect.com

SciVerse ScienceDirect

Current Opinion in Biotechnology

Transcriptome analysis using next-generation sequencing

Kai-Oliver Mutz, Alexandra Heilkenbrinker, Maren Lönne, Johanna-Gabriela Walter and Frank Stahl

Up to date research in biology, biotechnology, and medicine requires fast genome and transcriptome analysis technologies for the investigation of cellular state, physiology, and activity. Here, microarray technology and next generation sequencing of transcripts (RNA-Seq) are state of the art. Since microarray technology is limited towards the amount of RNA, the quantification of transcript levels and the sequence information, RNA-Seq provides nearly unlimited possibilities in modern bioanalysis. This chapter presents a detailed description of next-generation sequencing (NGS), describes the impact of this technology on transcriptome analysis and explains its possibilities to explore the modern RNA world.

Address
Leibniz Universität Hannover, Institute for Technical Chemistry,
Callinstrasse 5, 30167 Hannover, Germany

Corresponding author: Stahl, Frank (stahl@iftc.uni-hannover.de)

Current Opinion in Biotechnology 2013, **24**:22–30

This review comes from a themed issue on **Analytical biotechnology**

Edited by **Susann Müller** and **Karsten Hiller**

For a complete overview see the Issue and the Editorial

Available online 25th September 2012

http://dx.doi.org/10.1016/j.copbio.2012.09.004

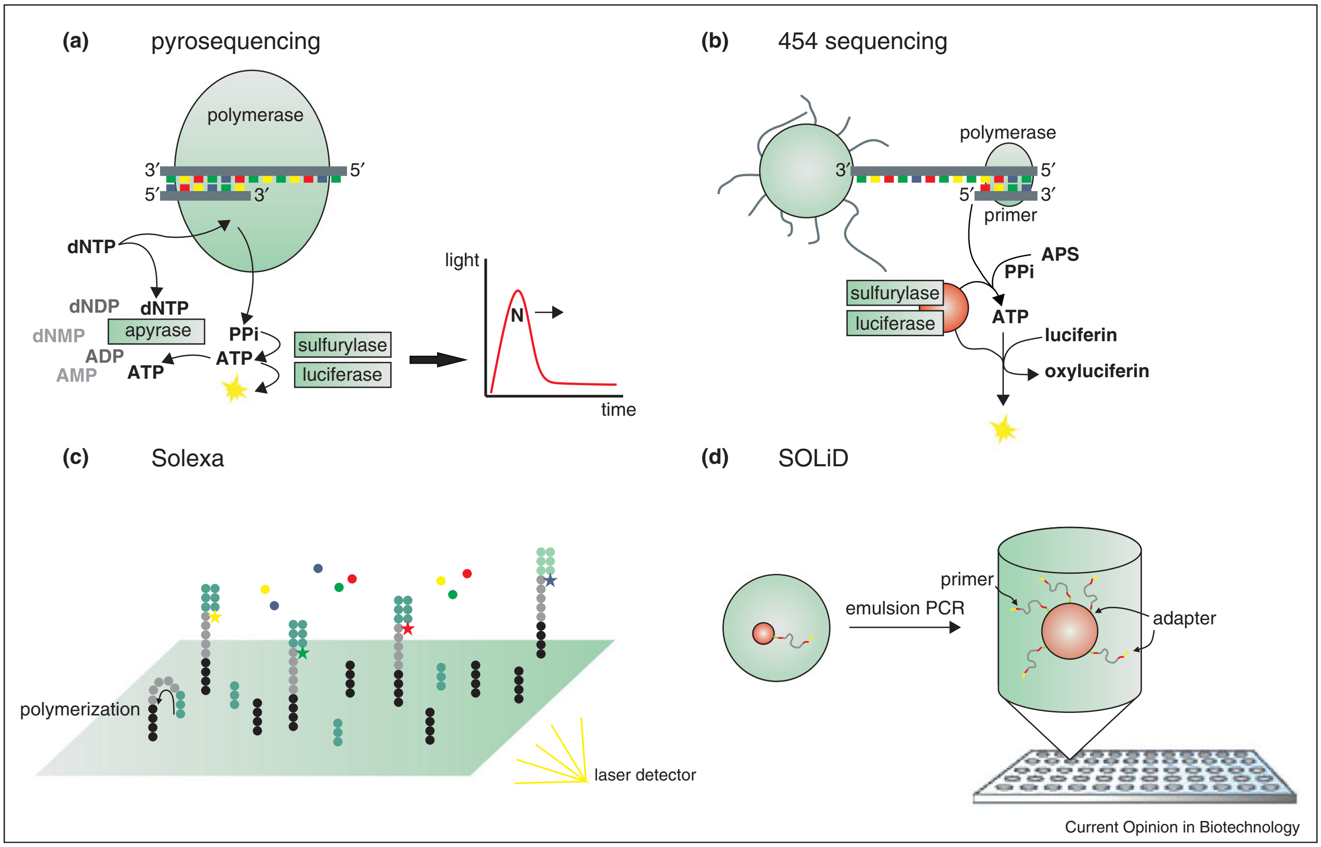
Introduction

Methodologically sound technologies for transcriptome analysis are available and widely used since the development of microarray technology and the complete sequencing of the human genome. Formerly mRNA expression was measured by microarray techniques or real-time PCR techniques. The first method does not have an exquisite sensitivity, the latter is quite expensive and cannot be used for a genome-wide survey of gene expression [1]. In contrast, the rapid and inexpensive next-generation sequencing NGS methods offer high-throughput gene expression profiling, genome annotation or discovery of non-coding RNA. In general, DNA sequencing is one of the most important platforms for molecular biological studies. Sequence decoding is usually performed using dideoxy chain termination technology [2]. With increasing importance of DNA sequencing in research and diagnostics [3], new methods were developed allowing a high-throughput sample treatment. While the first determination of the human genome took more than 15 years and millions of dollars, today it is possible to sequence it in about eight days for approximately $100 000 [4]. The important technological developments, summarized under the term next-generation sequencing, are based on the sequencing-by-synthesis (SBS) technology called pyrosequencing. The transcriptomics variant of pyrosequencing technology is called short-read massively parallel sequencing or RNA-Seq [5]. In recent years, RNA-Seq is rapidly emerging as the major quantitative transcriptome profiling system [6,7]. Different companies developed different sequencing platforms all based on variations of pyrosequencing.

Next-generation sequencing

To achieve sequence information, various methods are well established today. The classical dideoxy method, invented by Friedrich Sanger in the 1970th uses an enzymatic reaction. Starting with short random primers, DNA polymerase elongates the complementary strands. The addition of one of the four differently labeled dideoxynucleotides results in a detectable chain termination and therefore enables the identification of the unknown DNA. In contrast to the Sanger method, which is similar to natural DNA replication [8], pyrosequencing uses sequencing by synthesis technology. Here, the incorporation of nucleotides during DNA sequencing is monitored by luminescence. Therefore, a luziferase-based multi-enzyme system generates a lightning after nucleotide binding. The four different nucleotides are added sequentially and only incorporated nucleotides cause a signal. Pyrosequencing provides the analysis of single nucleotide polymorphisms (SNPs) as well as whole-genome sequencing [9]. Beyond this, NGS is pushing transcriptomics further into the digital age [10]. Pyrosequencing technology was invented in 1986 [11] from the idea to follow nucleotide incorporation using delivered pyrophosphate (PPi) to determine a signal. This indicates whether a nucleotide is incorporated or not. The release of an equimolar amount of PPi is a natural process during the binding of a nucleotide to the 3′ end of the primer used as starting point for sequencing [12••]. A multi-enzyme complex consisting of DNA polymerase, ATP sulfurylase, luciferase and apyrase is responsible for the amplification reaction. After nucleotide incorporation, the release of PPi induces the sulfurylase reaction resulting in a quantitative conversion of PPi to ATP [13]. The luciferase uses the ATP in the conversion of luciferin to oxyluciferin. As a result of this reaction, visible light is emitted and can be detected by a CCD camera [9] (Figure 1). A computer program

Figure 1



Basic principles of NGS techniques. **(a)** pyrosequencing: the incorporation of a new nucleotide generates detectable light. **(b)** 454 sequencing: nucleotide incorporation is associated with the release of pyrophosphate resulting in a light signal. **(c)** Solexa: DNA fragments build double-stranded bridges and after the addition of the labeled terminators the sequencing cycle starts. **(d)** SOLiD: if the adapters are bound, emulsion PCR is carried out to generate so-called bead clones.

illustrates the recorded data as peaks in a diagram. Based on the order of the signal peaks the DNA sequence can be determined [14]. Free ATP and nucleotides are degraded by the apyrase. This disables light emission and regenerates the solution. The complete enzyme process can be performed in a single well, offering a fast reaction time of approximately 20 min per 96-well plate [9].

A lot of sequencing platforms are available today. All of them use short fragments, so-called reads, to investigate genome sequences [15], chromatin immunoprecipitation (ChIP) or mapping of DNA methylation. The two most common variants based on conventional pyrosequencing principle are the *454 Sequencing*, and the *PyroMark ID* system. The second developmental stage of pyrosequencing was the invention of surface-based systems. These systems also follow the principle to detect every single nucleotide incorporation step, but enzymatically induced lightnings are not necessary anymore. This recent generation provides options for high-throughput functional transcriptomics. Various capabilities like discovery of transcription factor bindings or non-coding RNA expression profiling could have been established by now [6]. Here, the *Genome analyzer* and *SOLiD* are widely spread variants. All these methods have various advantageous and drawbacks and access or local facilities will influence the choice of the according technology [16]. These four systems are presented in detail next.

The first large-scale adaption of the pyrosequencing technique, invented by 454 Life Sciences [17] and later commercialized by Roche, is a high-throughput system [18]. It can be described as pyrosequencing in high-density picoliter reactors. Fragmentized DNA is attached to streptavidin beads that are consequently captured into aqueous droplets in an oil solution. This so-called emulsion PCR separates DNA molecules along with the primer-coated beads. Thus, the droplets form small amplification reactors [19]. Every bead is transferred to a picoliter plate and analyzed by normal pyrosequencing. 454 instruments are sequencing up to 500 million bases within ten hours. The read length (250 nt) is shorter than with Sanger technology (600 nt) because it is limited by the used pyrosequencing chemistry. As a result of decrease in apyrase's efficiency in degrading excess nucleotides, non-synchronized extension is the main limiting factor [20]. Nonetheless, 454 sequencing is the faster technique because up to 400 000 reactions can be performed in parallel [21]. With almost 1000 research publications, it is the most widely published next-generation platform.

The medium-throughput *PyroMark ID* platform by Qiagen is the method of choice for small research laboratories. It is able to analyze 96 samples simultaneously. In comparison to classical sequencing methods the read length is shorter (40 nt) and limited by the number of flows in the instrument. Therefore, this hardware is prevalently used for SNP, mutation analyses or RNA sequencing.

The *Genome analyzer* was developed by Solexa (now part of Illumina) and can be used for conventional DNA sequencing as well as for transcriptome analysis. This method is based on a two-step mechanism and combines single molecule amplification technology and novel reversible terminator-based sequencing. First, DNA randomly fragmented by shear stress is ligated with adapters at both sides of their chain. Then the DNA is attached to the internal surface of a flow cell. This flow cell is derivatized with oligonucleotides forming a dense layer of primers, which are complementary to the adapters [22]. The DNA fragments hybridized with the primers in a bridging way initialise solid-phase bridge amplification immediately and the fragments become double-stranded. With further steps including denaturation, renaturation and synthesis, a high-density of equal DNA fragments is generated in a small area. Approximately one million double-stranded DNA copies are produced per cluster, which is representing one single fragment. This guarantees the required signal intensity for detection during sequencing [1]. The second step is the real sequencing reaction. All four dNTPs carrying a base-unique fluorescent dye are added and incorporated by a DNA polymerase gradually. Following each base incorporation step, an image is made by laser excitation for each cluster. The identity of the first base is recordable. Then the elimination of the chemically blocked 3′-OH group and the dye follows. Within every new cycle, the DNA chain is elongated and more images are recorded. A base-calling algorithm assigns the sequences and evaluates the analysis quality. With read length of 25–35 bases the reading frame is tenfold smaller than with common pyrosequencing. *Solexa* read length is limited by interference between DNA polymerase and the used fluorochromes, which results in a reduced enzyme activity. Most strands get terminated early. Otherwise, fluorochromes used by *Solexa* and also by *SOLiD* generate stronger signals than the luciferase in pyrosequencing allowing a much higher density of reactions.

SOLiD is a system by Applied Bioscience. DNA fragments are ligase-modified with adapters, coupled to microparticles and applied to an emulsion PCR system (same principle like 454 sequencing) [23]. The adapters are cleaved and DNA rings are formed by ligation of the adapter ends. The rings get split at defined positions at the left and right domain of the adapter again. Otherwise new adapters are ligated to the new ends. Consequently, further fragments can hybridize to the beads and thus be amplified [24]. After the DNA accumulation steps, different fluorescence-labeled 8mer oligonucleotides are ligated to the sequencing primers binding to the adapter sequence. The dye color is defined by the first two of the eight bases. If the first bases fit to the DNA sequence, their fluorescence signal can be measured. After the elimination of the last three bases and the dye, the fifth, tenth and 15th base will be identified in further cycles [1]. Other steps with shorter primers lead to the detection of the positions four, nine, 14, and so on. The process is repeated until reaching read length between 35 and 100 nucleotides. Hence, *SOLiD* is not running as fast as the other systems and the degradation of the template strand over time is the limiting factor [25].

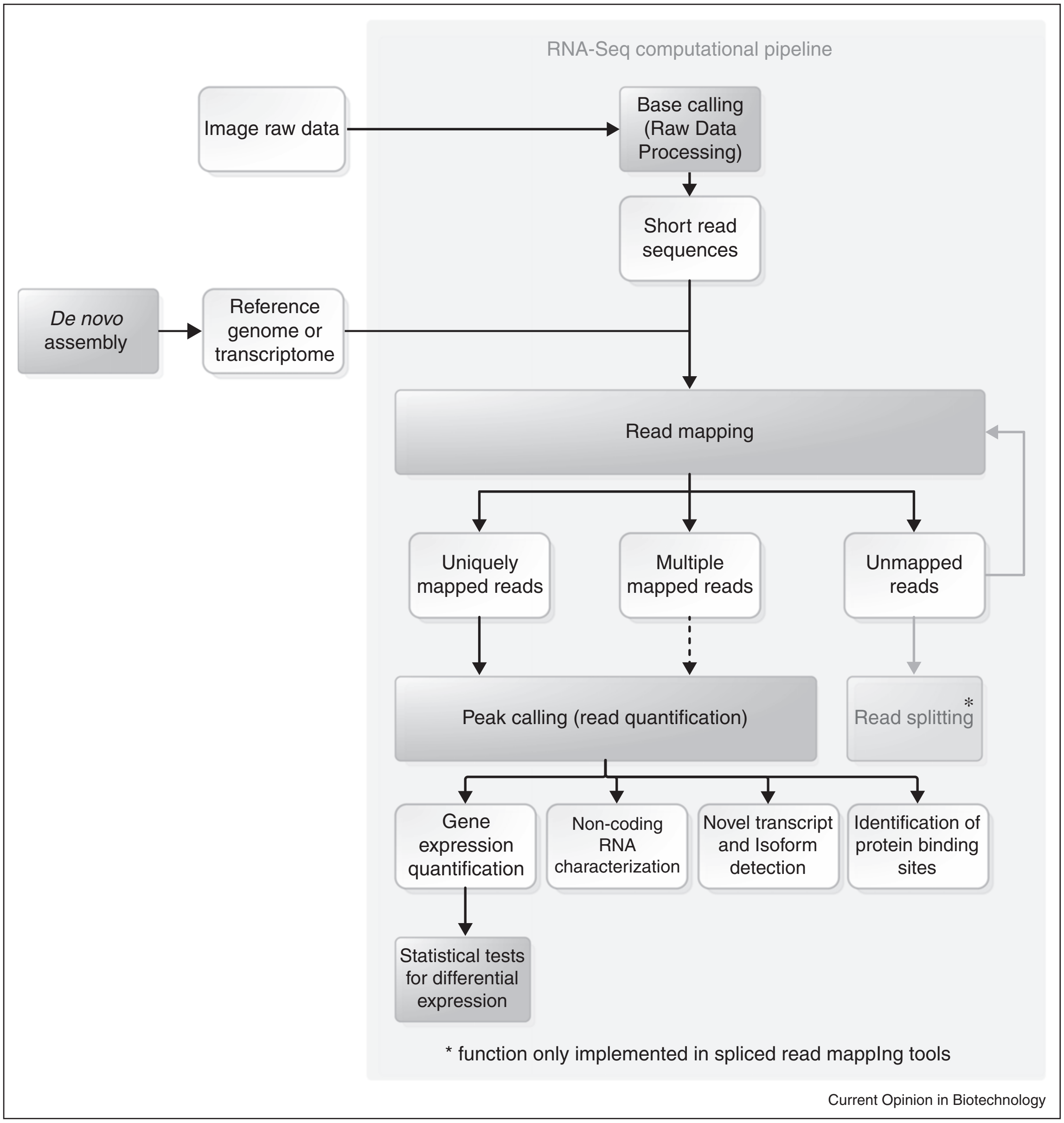
Data analysis

NGS analyzes millions of short DNA fragments during one sequencing run. The read lengths of those fragments depend on the type of NGS platform and can be in the range of 25–450 base pairs. While the reads are much shorter than those created by Sanger sequencing, NGS has a higher throughput and creates data sets with up to 50 gigabases per run [26]. This demands improved algorithms that are capable to process those huge amounts of raw data material.

The analysis pipeline of RNA-Seq (Figure 2) consists of four fundamental analysis steps, providing that an already sequenced reference genome or transcriptome is available for the reviewed organism. First of all, raw image data have to be converted into short read sequences, which are subsequently aligned to the reference genome or transcriptome. The amount of mapped reads is counted and the gene expression level is calculated by peak calling algorithms. Finally using statistical tests, differential gene expression is determined. During the past years several algorithms have been developed for each analysis step and afterwards adapted to specific applications. Today researchers can combine a variety of bioinformatic tools to obtain an appropriate analysis system optimized to their requirements [27]. Their principles of operation as well as advantages and limitations have been reviewed several times [27–29,30•].

The raw data generated by NGS consists of fluorescence signals, which have to be converted into base sequences by platform specific base calling-algorithms provided by the manufacturer. Additionally, a quality score for each base is calculated, which indicates the reliability of each base call. While the output of all NGS platforms is stored

Figure 2


RNA-Seq computational pipeline
Image raw data
Base calling (Raw Data Processing)
Short read sequences
De novo assembly
Reference genome or transcriptome
Read mapping
Uniquely mapped reads
Multiple mapped reads
Unmapped reads
Peak calling (read quantification)
Read splitting*
Gene expression quantification
Non-coding RNA characterization
Novel transcript and Isoform detection
Identification of protein binding sites
Statistical tests for differential expression
* function only implemented in spliced read mappIng tools
Current Opinion in Biotechnology

Computational pipeline for next-generation sequencing data.

in the standard FASTQ format, the calculation of quality score differs from manufacturer to manufacturer. This has to be considered when choosing an appropriate read mapping algorithm as the scores have to be interpreted in different manner. Besides the sequencing data, read mapping algorithms require reference sequences to which they can align the short reads. These reference genomes or transcriptomes, which arise from whole-genome sequencing projects, are available at Genome Online Database (GOLD) [27]. Currently, GOLD provides completely sequenced genomes of 3173 species and the number of registered sequencing projects raises continuously. Read mapping algorithms in contrast to conventional alignment algorithms use indexing strategies, which enable them to align millions of short reads in an appropriate period of time. As mapping tools have to balance between sensitivity and speed, no aligner can be best suited for all applications [31]. Currently, most algorithms either use hash look up tables, or Burrow Wheeler transformations for indexing. While hash table based tools show a high sensitivity, Burrow Wheeler methods are much faster [27]. Most mapping tools allow few mismatches during the sequence alignment considering sequencing errors, single nucleotide polymorphisms or mutations but they do not allow any large gaps. The exact number of allowed mismatches has to be chosen depending on the read length [31,32]. As this length is determined by the used sequencing platform, a certain read mapping tool is never suitable for all sequencing technologies. Moreover, every biological application demands for special mapping algorithms: For RNA-Seq analysis of eukaryotes, cDNA reads are preferably aligned to a reference transcriptome. Since these data are unfortunately rarely available so far, researchers have to use the genomic sequences instead. This requires spliced read mapping software, which considers the genomic intron–exon structure by splitting unmapped reads and aligning the read fragments independently [27]. To estimate the gene expression level using RNA-Seq, reads, which are mapped to a particular gene, have to be quantified. Therefore, bioinformatic tools count the number of reads in a window of defined size. By moving the window along the whole sequence, an expression profile is generated. Subsequently these expression scores have to be normalized because of inherent bias in read quantification: On the one hand the gene length influences the number of reads mapped to it, on the other hand this number also depends on the sequencing depth (total number of sequenced reads) [31]. Normalization of read counts enables the comparison of expression level between different genes as well as different experiments.

Genes, which are differentially expressed under different conditions, are detected by computational tools using normalized gene expression scores and statistical tests. These tools are classified as parametric or non-parametric algorithms. Parametric algorithms use common probability distributions such as Binomial or Poisson. Non-parametric ones model the noise distribution based on actual data [33]. It was demonstrated that non-parametric algorithms show a lower dependency on sequencing depth and consequently achieve more robust results [34].

Applications

One of the most obvious applications of NGS technology platforms is genome sequencing. The NGS revolution has tremendously reduced time and cost requirements for large genome sequencing. Therefore, NGS technologies became very interesting for *de novo* sequencing of eukaryotic genomes [35•]. The initial generation of genomic sequences enables a detailed genetic analysis of a particular organism and the analysis of genomic diversity among different individuals. During the past decade, even complete diploid human genomes have been sequenced with NGS technologies [36–39].

NGS is also used for whole-genome or targeted resequencing, which is currently one of the broadest applications of high-throughput sequencing [26]. Mapped to a reference genome resequenced sequences can be used to identify SNPs, small insertions or deletions (indels), copy number variations (CNVs), and other structural variations [26,40]. Thereby, next-generation resequencing enhances the understanding of how genetic differences affect phenotypic characteristics and diseases [32,41].

Recently, NGS platforms have developed a major significance in diagnostics for the detection of molecular mutations. The application of NGS enables the observation of genome-wide mutation patterns to identify aberrant DNA sequence changes and new disease genes [42,43]. This supports the prenatal and postnatal diagnosis of genetic diseases like Down syndrome [44] and Parkinson's disease [45,46]. The ability to sequence on a genome-wide scale allows many cancer-initiating mutations to be studied. These tumorgenesis-promoting mutations are able to effect gene inactivation (e.g. promoter silencing, gene deletion, nonsense mutations), gene overexpression/misexpression (e.g. copy number amplification, methylation), and dominant protein activation [42]. Up to now many individual cancer genomes have been successfully sequenced and a lot of new cancer-initiating mutations have been identified using NGS technologies [40].

NGS technologies have also been established to examine epigenetic modifications on a genome-wide scale [32]. Modifications, like DNA methylation, chromatin structure variations, and posttranslational modifications of histones, play an essential role in cellular processes including genome regulation [18,26,32], disease appearance, and oncogenetic development [47]. Therefore, it is reasonable to catalog epigenetic modifications on a genome wide scale using high-throughput sequencing technologies.

Furthermore, additional gene regulation events including DNA–protein interactions [35•,48], transcription factor bindings [32,49], and nucleosome positioning [18,48] have been analyzed with NGS.

Metagenomic studies are also a key application of next-generation sequencing technologies in ecological and environmental research. The microbial diversity can be monitored by analyzing environmental and clinical samples [3], including water, soil, sediments, and gut contents [50]. These NGS studies help to understand evolutionary development and ecological biodiversity of microbes to carry out species classifications [51].

NGS technologies have also been successfully applied to gene expression profiling by sequencing different mRNA species [3,52]. These high-throughput NGS analyses play an important role in molecular biology to measure the activity of thousands of genes in parallel. Moreover, NGS enables the absolute quantification of transcripts. This helps to detect changes of gene expression levels under different biological conditions or between different cell types or tissues [32,35•]. Gene expression profiling allows

Table 1

Applications of next-generation sequencing technologies

Classification	Applications	Sequencing principle[a]	NGS technology	Basis for sequencing	Reference
Genome	*de novo* genome sequencing	Pyrosequencing	454	DNA	[19]
	Whole-genome or targeted resequencing, detection of SNPs, indels, CNVs	Pyrosequencing	454	DNA	[36]
Transcriptome	Gene expression profiling, SNP, alternative splicing	Sequencing-by-ligation	SOLiD	RNA	[60]
	SNP discovery	Pyrosequencing	454	RNA	[63]
	Mapping and quantification of transcriptomes	RNA-Seq	Genome analyzer	RNA	[58]
Epigenome	DNA methylation patterns	MeDiP-Seq	Genome analyzer	DNA	[64]
	Histone modification	ChiP-Seq	Genome analyzer	DNA	[65]
Regulome	Nucleosome positioning	Sequencing-by-ligation	SOLiD	DNA	[66]
	Transcription factor binding	ChiP-Seq	Genome analyzer	DNA	[67]
Metagenome	Microbial diversity	Sequencing-by-synthesis	Genome analyzer	DNA	[68]
	Species classification	pyrosequencing	454	DNA	[69]
Diagnostics	Genetic diseases	Sequencing-by-ligation	SOLiD	DNA	[38]
	Prenatal diagnostics	Sequencing-by-ligation	SOLiD	DNA	[70]
	Cancer detection	Sequencing-by-ligation	SOLiD	RNA	[71]

[a] *Abbreviations*: RNA-Seq (RNA sequencing), ChIP-Seq (chromatin immunoprecipitation seqeuncing), MeDIP-Seq (methylated DNA immunoprecipitation sequencing).

the investigation of cellular functions, cellular states or cell physiology. The impact of environmental factors, chemical signals, or stress factors on gene expression can be analyzed on a genome-wide scale using NGS technologies.

To enhance the annotation of sequenced genomes, NGS has also been applied to small non-coding RNA (ncRNA) discovery and profiling [18,49]. Non-conding RNAs, including transfer RNA (tRNA), ribosomal RNA (rRNA), small nuclear and small nucleolar RNA, micro RNA (miRNA), and small interfering RNA (siRNA), are not translated into proteins. Several of these ncRNA species, like micro RNAs or small interfering RNAs, are of major interest to transcriptomic research, since they are implicated in post-transcriptional regulation of numerous biological processes [18]. The availability of NGS technologies has led to the discovery of several new species of ncRNAs [35•,53]. High-throughput sequencing technologies offer a greater potential to discover novel ncRNA molecules than microarray-based methods. Small RNA profiling studies with *454 Sequencing* technology contributed to the discovery of a novel class of small RNAs, so-called Piwi-interacting RNAs (piRNAs). These piRNA molecules are presumably required for germ cell development in mammalian organisms [54–56]. In contrast to microarray technologies, NGS has the potential to detect variants in mature small ncRNA length and ncRNA editing [57]. Moreover, NGS technologies enable the profiling of known and novel small RNA genes [35•]. Besides gene expression and ncRNA profiling, the applications of NGS in the field of transcriptional profiling also include transcript annotation studies [35•]. These analyses are carried out to detect novel transcribed regions, splice events [53], additional promoters, exons, or 3′ untranscribed regions [58]. Transcript annotation studies also help to analyze the impact of transcriptional complexity on current models of key signaling pathways. Identification of allele-specific expression [59] and discovery of transcriptional activity of repeated elements [60] are additional applications of next-generation sequencing technologies. High-throughput sequencing can also provide aberrant transcription events, like pseudogenes, fusion genes, and genome rearrangements. Genome rearrangements resulting in aberrant transcriptional events are an indication of human cancer [18]. Therefore, NGS technologies prove to be useful to detect such transcriptional events on a genome-wide scale.

The *454 Sequencing* technology is the leading next-generation method in transcriptomics. Aside, the Illumina sequencers are used for profiling microRNAs [57]. To date, small RNA profiling studies involving *454 Sequencing* have been reported. Illumina or *SOLiD* platforms provide a ten times deeper coverage of small RNAs and are also used to identify novel miRNA genes [57]. Thus, next-generation sequencing technologies have broad applicability in many fields of research. They offer new high-throughput sequencing techniques that prove to be useful for many applications, including genomic, transcriptomic, epigenomic, regulomic, metagenomic, and diagnostic research (Table 1).

Conclusion

In the last few years functional trancriptomics has been progressed by both microarray technology as well as RNA-Seq and the possibilities to perform transcriptome profiling provide a resolution that would have been inconceivable some years ago. Certainly microarray technology has achieved its technical limits and is more and more complemented by high-throughput next-generation sequencing technologies. Unlike microarrays, transcriptome sequencing (RNA-Seq) can evaluate absolute transcript levels of sequenced and unsequenced organisms, detect novel transcripts and isoforms, identify previously annotated 5′ and 3′ cDNA ends, map exon/intron boundaries, reveal sequence variations (e.g. SNPs) and splice variants and many more. During the past five years next-generation sequencing was established for almost every DNA-based molecular research field and therefore could be considered as an ''all in one'' platform. However, the economical benefits seem not to be sufficient enough. Different research groups and companies are working on the third generation of sequencing. In the near future, for example, sequencing the human genome will be available for less than $1000 [41,61,62•].

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Mardis ER: **Next-generation DNA sequencing methods**. *Annu Rev Genomics Hum Genet* 2008, **9**:387-402.

2. Ronaghi M: **Pyrosequencing sheds light on DNA sequencing**. *Genome Res* 2001, **11**:3-11.

3. Voelkerding KV, Dames SA, Durtschi JD: **Next-generation sequencing: from basic research to diagnostics**. *Clin Chem* 2009, **55**:641-658.

4. Lohr S: *Jobs Tried Exotic Treatments to Combat Cancer, Book Says*. The New York Times; 2011.

5. Denoeud F, Aury JM, Da Silva C, Noel B, Rogier O, Delledonne M, Morgante M, Valle G, Wincker P, Scarpelli C *et al.*: **Annotating genomes with massive-scale RNA sequencing**. *Genome Biol* 2008, **9**:R175.

6. Wang RL, Biales A, Bencic D, Lattier D, Kostich M, Villeneuve D, Ankley GT, Lazorchak J, Toth G: **DNA microarray application in ecotoxicology: experimental design, microarray scanning, and factors affecting transcriptional profiles in a small fish species**. *Environ Toxicol Chem* 2008, **27**:652-663.

7. Wang L, Feng Z, Wang X, Zhang X: **DEGseq: an R package for identifying differentially expressed genes from RNA-seq data**. *Bioinformatics* 2010, **26**:136-138.

8. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors**. *Proc Natl Acad Sci USA* 1977, **74**:5463-5467.

9. Marsh (Ed): *Pyrosequencing Protocols*. New York City: Humana Press; 2007.

10. Blow N: **Transcriptomics: the digital generation**. *Nature* 2009, **458**:239-242.

11. Nyren P: **The history of pyrosequencing**. In *Pyrosequencing Protocols.* Edited by Marsh S. Humana Press; 2007:1-14.

12. Novais RC, Thorstenson YR: **The evolution of pyrosequencing for microbiology: from genes to genomes**. *J Microbiol Methods* 2011, **86**:1-7.
•• This review gives detailed insights into the mechanisms of pyrosequencing. In this context, the authors highlight the possibilities to quantify via next-generations sequencing, which is the prerequisite for transcriptome analysis described in this review.

13. Agah A, Aghajan M, Mashayekhi F, Amini S, Davis RW, Plummer JD, Ronaghi M, Griffin PB: **A multi-enzyme model for pyrosequencing**. *Nucleic Acids Res* 2004, **32**:e166.

14. Shen S, Qin D: **Pyrosequencing data analysis software: a useful tool for EGFR, KRAS, and BRAF mutation analysis**. *Diagn Pathol* 2012, **7**:e56.

15. Korbel JO, Urban AE, Grubert F, Du J, Royce TE, Starr P, Zhong G, Emanuel BS, Weissman SM, Snyder M *et al.*: **Systematic prediction and validation of breakpoints associated with copy-number variants in the human genome**. *Proc Natl Acad Sci USA* 2007, **104**:10110-10115.

16. van Vliet AH: **Next generation sequencing of microbial transcriptomes: challenges and opportunities**. *FEMS Microbiol Lett* 2010, **302**:1-7.

17. Imelfort M, Edwards D: **De novo sequencing of plant genomes using second-generation technologies**. *Brief Bioinform* 2009, **10**:609-618.

18. Morozova O, Marra MA: **Applications of next-generation sequencing technologies in functional genomics**. *Genomics* 2008, **92**:255-264.

19. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z *et al.*: **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature* 2005, **437**:376-380.

20. Mashayekhi F, Ronaghi M: **Analysis of read length limiting factors in pyrosequencing chemistry**. *Anal Biochem* 2007, **363**:275-287.

21. Droege M, Hill B: **The Genome Sequencer FLX system – longer reads, more applications, straight forward bioinformatics and more complete data sets**. *J Biotechnol* 2008, **136**:3-10.

22. Koboldt DC, Larson DE, Chen K, Ding L, Wilson RK: **Massively parallel sequencing approaches for characterization of structural variation**. In *Genomic Structural Variants: Methods and Protocols.* Edited by Feuk L. Humana Press; 2012:369-384.

23. Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S *et al.*: **Evaluation of next generation sequencing platforms for population targeted sequencing studies**. *Genome Biol* 2009, **10**:R32.

24. Pandey V, Nutter RC, Prediger E: **Applied Biosystems SOLiD system: ligation-based sequencing**. In *Next-Generation Genome Sequencing: Towards Personalized Medicine.* Edited by Janitz M. Wiley-VCH; 2008:29-42.

25. Hossain MS, Azimi N, Skiena S: **Crystallizing short-read assemblies around seeds**. *BMC Bioinform* 2009, **10(Suppl 1)**:S16.

26. Nowrousian M: **Next-genaration sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems**. *Eukaryot Cell* 2010, **9**:1300-1310.

27. Pagani I, Liolios K, Jansson J, Chen IM, Smirnova T, Nosrat B, Markowitz VM, Kyrpides NC: **The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata**. *Nucleic Acids Res* 2011, **40**:D571-D579.

28. Garber M, Grabherr MG, Guttman M, Trapnell C: **Computational methods for transcriptome annotation and quantification using RNA-seq**. *Nat Methods* 2011, **8**:469-477.

29. Pepke S, Wold B, Mortazavi A: **Computation for ChIP-seq and RNA-seq studies**. *Nat Methods* 2009, **6**:S22-S32.

30. Treangen TJ, Salzberg SL: **Repetitive DNA and next-generation**
• **sequencing: computational challenges and solutions**. *Nat Rev Genet* 2011, **13**:36-46.

The authors give an overview of different computational tools for NGS genome alignment and assembly as well as transcriptome analysis. Supplementary to our review, this article introduces examples of open source software tools.

31. Park PJ: **ChIP-seq: advantages and challenges of a maturing technology**. *Nat Rev Genet* 2009, **10**:669-680.

32. Cullum R, Alder O, Hoodless PA: **The next generation: using new sequencing technologies to analyse gene regulation**. *Respirology* 2011, **16**:210-222.

33. Li J, Tibshirani R: **Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data**. *Stat Methods Med Res* 2011 http://dx.doi.org/10.1177/0962280211428386.

34. Tarazona S, Garcia-Alcalde F, Dopazo J, Ferrer A, Conesa A: **Differential expression in RNA-seq: a matter of depth**. *Genome Res* 2011, **21**:2213-2223.

35. Zhou X, Ren L, Meng Q, Li Y, Yu Y, Yu J: **The next-generation**
• **sequencing technology and application**. *Protein Cell* 2010, **1**:520-536.

The authors compare the commercially available next-generation sequencing platforms 454, Illumina, and SOLiD and their sequencing techniques. Moreover, current next-generation sequencing applications in genomics, transcriptomics, epigenomics and metagenomics are represented in detail.

36. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT *et al.*: **The complete genome of an individual by massively parallel DNA sequencing**. *Nature* 2008, **452**:872-876.

37. Pushkarev D, Neff NF, Quake SR: **Single-molecule sequencing of an individual human genome**. *Nat Biotechnol* 2009, **27**:847-850.

38. Lupski JR, Reid JG, Gonzaga-Jauregui C, Deiros DR, Chen DCY, Nazareth L, Bainbridge M, Dinh H, Jing C, Wheeler DA *et al.*: **Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy**. *N Engl J Med* 2010, **362**:1181-1191.

39. Zhang J, Chiodini R, Badr A, Zhang G: **The impact of next-generation sequencing on genomics**. *J Genet Genomics* 2011, **38**:95-109.

40. Jia P, Zhao Z: **Personalized pathway enrichment map of putative cancer genes from next generation sequencing data**. *PLoS ONE* 2012, **7**:e37595.

41. Kohlmann A, Grossmann V, Haferlach T: **Integration of next-generation sequencing into clinical practice: are we there yet?** *Semin Oncol* 2012, **39**:26-36.

42. Ma QC, Ennis CA, Aparicio S: **Opening Pandora's box – the new biology of driver mutations and clonal evolution in cancer as revealed by next generation sequencing**. *Curr Opin Genet Dev* 2012, **22**:3-9.

43. Zhang W, Cui H, Wong LJ: **Application of next generation sequencing to molecular diagnosis of inherited diseases**. In *Topics in Current Chemistry.* Edited by Wong CH, Houk KN, Hunter CA, Krische MJ, Lehn JM, Ley SV, Olivucci M, Thiem J, Venturi M, Vogel P.*et al.*: Springer-Verlag; 2012:1-27.

44. Palomaki GE, Deciu C, Kloza EM, Lambert-Messerlian GM, Haddow JE, Neveux LM, Ehrich M, van den Boom D, Bombard AT, Grody WW *et al.*: **DNA sequencing of maternal plasma reliably identifies trisomy 18 and trisomy 13 as well as Down syndrome: an international collaborative study**. *Genet Med* 2012, **14**:296-305.

45. Gasser T, Hardy J, Mizuno Y: **Milestones in PD genetics**. *Mov Disord* 2011, **26**:1042-1048.

46. Vilarino-Guell C, Wider C, Ross OA, Dachsel JC, Kachergus JM, Lincoln SJ, Soto-Ortolaza AI, Cobb SA, Wilhoite GJ, Bacon JA *et al.*: **VPS35 mutations in Parkinson disease**. *Am J Hum Genet* 2011, **89**:162-167.

47. Zhang Y, Jeltsch A: **The application of next generation sequencing in DNA methylation analysis**. *Genes* 2010, **1**:85-101.

48. Park PJ: **Epigenetics meets next-generation sequencing**. *Epigenetics* 2008, **3**:318-321.

49. Marguerat S, Wilhelm BT, Bahler J: **Next-generation sequencing: applications beyond genomes**. *Biochem Soc Trans* 2008, **36**:1091-1096.

50. Shokralla S, Spall JL, Gibson JF, Hajibabaei M: **Next-generation sequencing technologies for environmental DNA research**. *Mol Ecol* 2012, **21**:1794-1805.

51. Wong K, Fong TT, Bibby K, Molina M: **Application of enteric viruses for fecal pollution source tracking in environmental waters**. *Environ Int* 2012, **45**:151-164.

52. Barzon L, Lavezzo E, Militello V, Toppo S, Palu G: **Applications of next-generation sequencing technologies to diagnostic virology**. *Int J Mol Sci* 2011, **12**:7861-7884.

53. MacLean D, Jones JD, Studholme DJ: **Application of 'next-generation' sequencing technologies to microbial genetics**. *Nat Rev Microbiol* 2009, **7**:287-296.

54. Girard A, Sachidanandam R, Hannon GJ, Carmell MA: **A germline-specific class of small RNAs binds mammalian Piwi proteins**. *Nature* 2006, **442**:199-202.

55. Houwing S, Kamminga LM, Berezikov E, Cronembold D, Girard A, van den Elst H, Filippov DV, Blaser H, Raz E, Moens CB *et al.*: **A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish**. *Cell* 2007, **129**:69-82.

56. Lau NC, Seto AG, Kim J, Kuramochi-Miyagawa S, Nakano T, Bartel DP, Kingston RE: **Characterization of the piRNA complex from rat testes**. *Science* 2006, **313**:363-367.

57. Morin RD, O'Connor MD, Griffith M, Kuchenbauer F, Delaney A, Prabhu AL, Zhao Y, McDonald H, Zeng T, Hirst M *et al.*: **Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells**. *Genome Res* 2008, **18**:610-621.

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

59. Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, Pritchard JK: **Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data**. *Bioinformatics* 2009, **25**:3207-3212.

60. Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G *et al.*: **Stem cell transcriptome profiling via massive-scale mRNA sequencing**. *Nat Methods* 2008, **5**:613-619.

61. Check Hayden E: **Genome sequencing: the third generation**. *Nature* 2009, **457**:768-769.

62. Schadt EE, Turner S, Kasarskis A: **A window into third-**
• **generation sequencing**. *Hum Mol Genet* 2010, **19**:227-240.

The study shows first concepts for new sequencing strategies and techniques coming up next. The reader gains an impression of how easy transcriptome analysis could be performed in future.

63. Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS: **SNP discovery via 454 transcriptome sequencing**. *Plant J* 2007, **51**:910-918.

64. Li N, Ye M, Li Y, Yan Z, Butcher LM, Sun J, Han X, Chen Q, Zhang X, Wang J: **Whole genome DNA methylation analysis based on high throughput sequencing technology**. *Methods* 2010, **52**:203-212.

65. Pauler FM, Sloane MA, Huang R, Regha K, Koerner MV, Tamir I, Sommer A, Aszodi A, Jenuwein T, Barlow DP: **H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome**. *Genome Res* 2009, **19**:221-233.

66. Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, Zeng K, Malek JA, Costa G, McKernan K *et al.*: **A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning**. *Genome Res* 2008, **18**:1051-1063.

67. Jothi R, Cuddapah S, Barski A, Cui K, Zhao K: **Genome-wide identification of in vivo protein–DNA binding sites from ChIP-Seq data**. *Nucleic Acids Res* 2008, **36**:5221-5231.

68. Qin JJ, Li RQ, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T *et al.*: **A human gut microbial gene catalogue established by metagenomic sequencing**. *Nature* 2010, **464** 59-U70.

69. Nossa CW, Oberdorf WE, Yang LY, Aas JA, Paster BJ, DeSantis TZ, Brodie EL, Malamud D, Poles MA, Pei ZH: **Design of 16S rRNA gene primers for 454 pyrosequencing of the human foregut microbiome**. *World J Gastroenterol* 2010, **16**:4135-4144.

70. Chiu RWK, Sun H, Akolekar R, Clouser C, Lee C, McKernan K, Zhou DX, Nicolaides KH, Lo YMD: **Maternal plasma DNA analysis with massively parallel sequencing by ligation for noninvasive prenatal diagnosis of trisomy 21**. *Clin Chem* 2010, **56**:459-463.

71. Wu Q, Lu Z, Li H, Lu J, Guo L, Ge Q: **Next-generation sequencing of microRNAs for breast cancer detection**. *J Biomed Biotechnol* 2011, **2011**:597145.