# EMERGING RNA-SEQ APPLICATIONS IN FOOD SCIENCE

*Alberto Valdés, Carolina Simó, Clara Ibáñez, and Virginia García-Cañas**

Laboratory of Foodomics, CIAL (CSIC). Nicolás Cabrera 9, 28049 Madrid, Spain.

*Corresponding e-mail: virginia.garcia@csic.es

## Abstract

Groundbreaking research in food science is shifting from classical methods to novel analytical approaches in which high-throughput techniques have a key role. Among these techniques, RNA-Seq in combination with bioinformatics is applied to investigate topics in food science that were not approachable few years ago. Relevant applications of transcriptomics in modern food science include transcriptome characterization and analysis of gene expression levels in food crops, foodborne pathogens and fermenting microorganisms. The aim of the present chapter is to provide an overview of the recent progress in RNA-Seq techniques discussing their advantages and drawbacks. Besides, relevant applications of these technologies will be highlighted in the context of food science to illustrate their impressive potential. Besides, some ideas of the foreseen technological advances and potential applications of these fast-evolving techniques are also provided.

**Keywords:** Transcriptomics; RNA-Seq; next-generation sequencing; foodborne pathogens; crops; fermentations; gene expression.

## TABLE OF CONTENTS

1. INTRODUCTION

2. OVERVIEW OF RNA-Seq TECHNOLOGY
2.1 Library preparation and clonal amplification
2.2 Sequencing chemistry
2.3 Data analysis

3. APPLICATIONS OF RNA-Seq
2.1 De novo transcriptome assembly
2.2 Comparative transcriptomic analysis /Digital Gene Expression profiling
2.3 Splicing analysis
2.4 MicroRNA-Seq

3. RNA-Seq IN FOOD SCIENCE
3.1 Production of food crops
3.2 Foodborne pathogenic microorganisms
3.3 Food fermentations

4. FUTURE OUTLOOKS AND CONCLUSIONS

## 1. INTRODUCTION

For the last two decades, gene expression microarray has been regarded as one of the foremost technological advances in high-throughput analysis. The success of microarray in profiling gene expression has been remarkable in last years, and owing to the extensive optimization and standardization performed in instruments and protocol microarray has become a mature technology [1]. Gene expression microarray technique has been employed to obtain meaningful insights into the molecular mechanisms underlying complex biological processes relevant to food science. However, gene expression microarray fails to reach comprehensive and precise characterization of transcriptome due to some unsolved technical constraints. As alternative, recent unbiased sequencing methodologies, termed RNA-sequencing (RNA-Seq) are now available for genome-wide high-throughput transcriptomics. Such groundbreaking high-throughput technologies are changing the way we investigate the transcriptome landscapes in biological systems.

The transcriptome, defined as the complete set of RNA transcripts produced by the genome at any one time, can be considered as an important link between phenotype and information encoded in the Genome [2]. Comprehensive and precise characterization of transcriptome encompasses (i) the annotation of all species of transcript, including mRNAs, non-coding RNAS and small RNAs; (ii) the determination of the transcriptional structure of genes (start sites, 5' and 3' ends, splicing variants, etc); and the quantification of differential expression levels of each transcript under different conditions [3]. Revolutionary tools for transcriptomics analysis are providing new opportunities and prospects to investigate previously unanswered questions relevant to food science.  In opposition to gene expression microarray, RNA-Seq technologies are independent of any annotated sequence feature and rely on recent technical advances in high-density microarraying and various sequencing chemistries. They provide extraordinary opportunities to explore many different aspects of entire transcriptomes and also, they allow the determination of gene expression levels by robust digital quantitative analysis. Applications of RNA-Seq to the food and nutrition domain are relatively recent compared with their use in basic science applications. This chapter provides insight into recent progress in RNA-Seq technologies, discussing their main advantages and limitations. Innovative applications of RNA-Seq will be discussed in the context of food science to illustrate its extraordinary potential. Finally, some outlooks regarding forthcoming potential applications and technical advances will be drawn.

## 2. OVERVIEW OF RNA-Seq TECHNOLOGY

The development of groundbreaking DNA sequencing strategies known as next-generation sequencing (NGS) technologies has opened a new age in the study of biological systems. Over the last years, the extraordinary progress in NGS technologies has been driven by the strong competition between manufacturers. Thus, different generations of platforms for DNA sequencing have been developed with improved capabilities at costs that were unimaginable few years ago [4]. Besides, this novel sequencing technology has been expanded to the analysis of gene expression by specific techniques known as massively parallel sequencing of RNA or RNA-Seq. This technology has the potential to truly embrace the whole transcriptome, offering excellent possibilities for the discovery of new transcripts. In addition, RNA-Seq offers a larger dynamic range to measure RNA abundance with high reproducibility and in a nonrelative quantitative mode. These features, and the upcoming cost reductions, make RNA-Seq an increasingly attractive strategy, even for unknown genome organisms.

The earliest and currently most widely used NGS systems, referred to as second-generation sequencing (SGS) technologies, are characterized by the synchronous-controlled optical detection, together with cyclic reagent washes of a substrate where the extension of thousands DNA templates is carried out in massive and parallel fashion [5]. SG sequencers are open platforms designed for sequencing libraries of nucleic acids in a high-throughput and cost effective way. The three major SG sequencers are Illumina Genome Analyzer (GA), Roche 454 Genome Sequencer FLX system (FLX), and Applied Biosystems SOLiD (SOLiD) platform. Regardless the platform chosen for high-throughput sequencing, the analytical procedure shares common steps: (1) library preparation involving fragmentation of RNA molecules, cDNA synthesis and ligation to specific adaptors at both ends; (2) clonal amplification of each template since most imaging systems have not been designed to detect single chemiluminiscent or fluorescent events; (3) attachment of the amplified DNA templates to a solid support in a flow cell or a reaction chamber; and (4) iterative and synchronized flowing and washing off the reagents for DNA strand extension while signals are acquired by the detection system [6]. Following signal acquisition, the resulting raw image data have to be converted into short reads (nucleotidic sequences generated per DNA template) by a process named “base-calling”. Depending on the platform, DNA extension (or synthesis) can be attained enzymatically by ligation or polymerization. A brief description of the key features and some technical aspects of the main platforms employed in the food science applications is next provided.

### 2.1 Library preparation and clonal amplification

The selected method for sample (RNA or DNA) preparation prior to sequencing will depend on the NGS instrument used. Most of the sample preparation methods follow common procedures that are aimed to produce sets of short DNA molecules (library) with adapters ligated in their 3’- and 5’-ends. Owing to the sensitivity limitations of most detection systems in SGS platforms, library amplification is a previous requirement to the sequencing step. As the most widely used methods for amplification are capable to amplify every template from the library in several orders of magnitude in a very controlled way, the process has been commonly termed clonal amplification. Thus, clonal amplification can be performed using either solid-phase bridge amplification (employed in Illumina platforms, formerly Solexa) or emulsion PCR (emPCR, used in Roche and SOLiD platforms). In bridge amplification, an optically transparent surface in a flow cell is derivatized with forward and reverse primers that capture library DNA sequences by hybridizing with the adaptors [7]. Each single-stranded DNA molecule, immobilized at one end on the surface, bends over to hybridize with the complementary adapter on the support by its free end, forming a “bridge” structure that serves as template for amplification. The addition of amplification reagents initiates the generation of clusters of thousands of equal DNA fragments in small areas. After several amplification cycles, random clusters of about 1000 copies of single-stranded DNA (termed DNA polonies) are originated on the surface. By contrast, in emPCR, each DNA template from the library is captured onto a bead under conditions that favor one DNA molecule per bead [8]. Then, beads are emulsified along with PCR reagents in water-in-oil emulsions. During emPCR, each single template attached to a bead is then clonally amplified to obtain millions of copies of the same sequence. Next, beads can be immobilized in different supports or loaded into individual PicoTiterPlate (PTP) wells, which are made of fiber-optic bundle and is designed to house one single bead per well. This approach has been implemented in 454 Life Technologies/Roche and Applied Biosystems platforms.

## 2.2 Sequencing chemistry

The two most commonly used techniques employed in SGS platforms to obtain detectable signals can be classified as sequencing by synthesis and sequencing by ligation. Sequencing by synthesis techniques have as a common feature the use of DNA polymerase enzyme during the sequencing step. Sequencing by synthesis can be divided into cyclic-reversible termination technique and single-nucleotide addition technique.

Cyclic-reversible termination techniques have been implemented in Illumina and Helicos Bioscience sequencers. The general procedure involves the use of nucleotides (reversible terminators) that contain a removable fluorophore that blocks the incorporation of subsequent nucleotides to the newly synthetized chain until the fluorescent group is cleaved off. Thus, a complete cycle encompasses: (a) flowing the nucleotides to the sequencing surface, (b) washing-out the unbound nucleotides, (c) fluorescence signals acquisition and, (d) nucleotide deprotection to allow the incorporation of the next reversible terminator. The Illumina technology uses a unique removable fluorescent-dye group for each nucleotide type (**Fig. 1**). In this case, the reaction mixture for the sequencing reactions is supplied onto the amplified DNA created on a surface by bridge amplification. After each terminator nucleotide is incorporated into the growing DNA strand, a fluorescent signal is emitted and recorded by a CCD camera, which detects the position of the fluorescent signal in the support and identifies de nucleotide. By deprotection of blocking groups and washing, a new synthesis cycle starts. Similarly, Helicos Bioscience Technology has implemented this chemistry in the HeliScope sequencer, but using only a single-color fluorescence label. Another particular feature in HeliScope sequencer is the lack of clonal amplification step prior sequencing. This system has been the first commercial single-molecule DNA sequencing system and its detection system is capable of scanning millions of single molecules of DNA anchored to a glass cover slip flow cell.

Single-nucleotide addition techniques are based on adding limited amounts of an individual dNTP to the reaction media to synthesize DNA and, after washing, sequencing is resumed with the addition of another nucleotide. This technique has been combined with pyrosequencing by Roche and incorporated in the GS and GS FLX sequencers. In brief, after emPCR, the beads positioned onto the PTP wells are subjected to pyrosequencing process. This is initiated by the addition of ATP sulfurylase and luciferase enzymes, adenosine 5’ phosphosulfate (ASP), and luciferin substrates. Thus, in every sequencing cycle, a single species of dNTPs is flowed into the PTP (**Fig. 2**). The incorporation of a complementary nucleotide results in the release of pyrophosphate (PPi) which eventually leads to a burst of light [1]. Individual dNTPs are dispensed in a predetermined sequential order and the chemiluminescence is imaged with a charge-coupled device (CCD) camera. This process results in parallel sequencing of multiple DNA templates attached to a single bead in each well of the PTP. Based on a similar idea, the semiconductor technology, developed by Life Technologies, has developed a novel generation of sequencers (Ion Torrent and Ion Proton PGM) that employ emPCR for clonal amplification in combination with a highly dense microwell array in which each well acts as an individual DNA synthesis reaction chamber. Every time a nucleotide is incorporated in the complementary template, H+ is released. A layer composed of a highly dense field-effect transistor array is aligned underneath the microwell array to transform the change in pH to a recordable voltage change.

Sequencing by ligation has been implemented in the ABI SOLiD sequencer by Applied Biosystems. This sequencing technique requires the use of a DNA ligase instead of a DNA polymerase to generate the complementary DNA strands [9]. The ligation method involves the use of several fluorescent 8-mer oligonucleotides that hybridize with the template. More precisely, amplified templates by emPCR are subjected to a 3’ modification to ensure further attachment to the glass slide. Then, a universal (sequencing) primer hybridizes to the adapter on the templates and sequencing starts by ligation of any of the four competing fluorescent 8-mer oligonucleotides (**Fig. 3**). In each oligonucleotide, the dye color is defined by the first two of the eight bases. A ligation event is detected by the fluorescent label incorporated to the growing strand. After the detection of the fluorescence from the dye, bases 1 and 2 in the sequence can thus be determined. Then, the last three bases and the dye are cleaved to enable a further ligation event. Then, another hybridization-ligation cycle is initiated. After 10 cycles of ligation, the extended primer is removed from the template and the process is repeated with a universal primer that is shifted one base from the adaptor-fragment position. Shifting the universal primer in five rounds of 10 cycles enables the entire fragment to be sequenced, and provides an error correction scheme because each base position is interrogated twice. Color calls from the five ligation rounds are then ordered into a linear sequence and decoded in a process termed “two-base encoding”.

A common aspect in these NGS platforms, excepting HelisCope sequencer is that the observed signal is a consensus of the nucleotides or probes added to the identical templates (clones or clusters) in each cycle. In consequence, the sequencing process suffers from numerous biases due to imperfect clonal amplification and DNA extension failure. For example, a strand which has failed to incorporate a base in a given cycle will lag behind the rest of the sequencing run (phasing), whereas the addition of multiple nucleotides or probes in a given cycle, results in leading-strand dephasing (pre-phasing) [10]. Phasing and pre-phasing accumulate during sequencing leading to an increase of base-calling errors towards the end of reads. **Table 1** summarizes the comparison of the main NGS platforms including some of their advantages and limitations. For a broader overview of next-generation sequencing technology refer to [1,3,9].

### 2.3 Data analysis

Analysis of RNA-Seq data demands tailored bioinformatics strategies able to manage and process the huge amount of data generated by RNA-Seq methods [11]. Novel algorithms able to process the huge amount of RNA-Seq data are in continuous development for a variety of applications including sequence alignment, assembly, read annotation and quantitation, among others. This incessant bioinformatics progress provides improved resources, but in turn, is delaying the establishment of standard practice tools for analysis. For a comprehensive description of the bioinformatics tools and algorithms frequently used for analyzing NGS data refer to some articles in recent literature [12-14].

## 3. APPLICATIONS OF RNA-Seq

Current RNA-Seq methods offer some relevant advantages over the more mature microarray technology. In contrast to gene expression microarray, RNA-Seq provides more complete information about transcriptome because it allows the direct characterization of transcript sequences. As it will be discussed in next sections, this technology provides excellent opportunities to detect point mutations in expressed transcripts, discover new classes of RNA, identify fusion transcripts and unknown splice variants [15]. As detection of sequences does not rely on the availability of an annotated genome, RNA-Seq is particularly suitable for the investigation of organisms for which their genome has not been totally sequenced. Also, wider dynamic range (spanning over 5 orders of magnitude) and better sensitivity can be achieved using RNA-Seq, as signal-to noise ratios increase with the sequencing depth (the times that a particular base is sequenced) [16]. On the other side, increasing the sequencing depth is directly associated with an increase in the sequencing cost [2]. Hence, sequencing costs will vary depending on the sequencing depth required to effectively interrogate a given transcriptome. Also, another issue that complicates the RNA-Seq analysis is the heterogeneity of sequencing depth along the transcript length. This heterogeneity that might be originated during RNA enrichment, fragmentation, ligation, amplification and sequencing procedures, introduces an important bias in accurate quantification of gene expression. To the contrary, heterogeneity is not a concern in microarray analysis because of the fixed nature of probes that capture the transcripts by hybridization.

### 3.1 *De novo* transcriptome assembly

Depending on the analytical goals, sequencing reads can be either aligned to a reference genome (or transcriptome) or assembled *de novo* to produce a genome-scale transcription map that consists of both the transcriptional structure and level of expression for each gene [17]. The more basic outcome from RNA-Seq analysis is represented by the raw reads that should be subjected to quality evaluation in order to decide which portion of reads are suitable for downstream analysis. After this step, the alignment process is aimed at mapping the reads into the sequence of the reference genome or transcriptome. However, in many sequencing projects, a reference genome or transcriptome is not available. RNA-Seq technologies and potent computational tools can be combined to obtain the *de novo* assembly and annotation of a transcriptome for many organisms. Several strategies have been employed to that aim, including the generation of libraries with different fragmentation degrees; increase the sequencing coverage; use of paired-end reads technology, etc. The recent development of strategies based on sequencing paired-end (mate-paired) reads has improved the efficiency and accuracy of the assembly and mapping process [18]. Paired-end read approach involves the sequencing from both ends of the same molecule, thereby generating a larger read with known sequences at either end. Illumina has successfully implemented this technology by using two rounds of sequencing. Thus, once the first strand is sequenced, the template is regenerated to allow a second round of sequencing from the opposite end (complementary strand). On the other side, mate-paired sequencing in SOLiD technology involves joining the fragment ends by recircularization using oligonucleotide adaptors during library production to allow both ends to be determined in a single round of sequencing. The enormous interest on this topic has led to the development of dedicated bioinformatic tools to generate *de novo* assemblies. In spite of the high accuracy of these tools, validation of the generated assemblies is still challenging and requires intensive research.

## 3.2 Comparative transcriptomic analysis/Digital Gene Expression profiling

Providing that an already sequenced reference genome or transcriptome is available, reads can be aligned to the reference genome or transcriptome using mapping algorithms. In general, mapping algorithms can deal with single base differences owing to sequencing errors, mutations or SNPs; however, these tools cannot accommodate large gaps. In addition, the mapping process is time-consuming and faces several challenges which aggravate when working with complex eukaryote transcriptomes. In such cases, the use of splice-aware mapping tools capable of working with reads that represent alternative splicing patterns improves the mapping efficiency [12]. Another important issue that needs to be improved is mapping repetitive sequences (multimapped reads), especially when they match to more than one position in the genome.

After mapping step, the workflow for RNA-Seq data analysis for expression level determination involves the two general steps: calculation of gene expression levels by counting mapped reads (digital readout) and determination of differential gene expression using statistical tests. Regardless the method selected for quantifying gene expression, either sequencing full RNA molecules or the less expensive method based on sequencing the 3’end of each transcript, well suited bioinformatic tools are required for the estimation of transcript levels [19]. The first step for quantifying gene expression involves the conversion of sequence reads into a quantitative value for each transcript. In most cases the selection of a suitable approach will ultimately depend on the procedure used for library preparation. A common approach when sequencing full RNA molecules involves summarizing the number of reads for each transcript and then normalizing for the length of the transcript. In comparative transcriptomics, the central objective relies on the estimation of differential gene expression values from different tissues, treatments, conditions, developmental stages, varieties, etc. This can be achieved by an additional transformation of data in order to eliminate differences on sequencing depth between runs or libraries. The most frequently adopted method to that aim is based on the conversion of read counts to reads per kilobase per million mapped reads (RPKM) [20].

Sample preparation methods, specific for sensitive and accurate quantification of each transcript, are also gaining attention [21]. For instance, digital gene expression (DGE) tag profiling is a specific method based on Illumina RNA-Seq for accurate quantification of each transcript by sequencing short (20- or 21-bp) cDNA tags rather than the entire transcript. This approach offers enhanced sensitivity without the need for increasing the sequencing depth. The library preparation procedure involves sequential digestion steps of cDNA molecules alternated by enzymatic ligation of adaptors to finally build DNA templates consisting of a single 20- or 21-bp cDNA tag flanked by defined adapters that are further sequenced. The reads delivered by the sequencer are filtered and mapped to a reference genome (or transcriptome), to be subsequently counted and normalized with respect to transcript length and sequencing depth. Since DGE tag profiling does not attempt to sequence the entire length of each transcript, its sensitivity is higher with fewer total reads per run. Then, rare transcript discovery and quantification can be achieved by selecting the depth of coverage.

### 3.3 Splicing analysis

In addition to providing digital gene expression profiles, RNA-Seq can also assist on the investigation of many other aspects of the transcriptome such as alternative splicing isoform composition, gene fusion and nucleotide variations, unknown coding or non-coding transcripts and RNA editing [22]. Alternative splicing is a major mechanism of post-trsanscriptional regulation, by which the immature mRNA of a gene can be spliced into multiple isoforms after the transcription. For instance, in the human, more than 95% of multi-exon genes undergo alternative splicing. Considering that the alternative spliced isoforms can have relevant functional meaning and that they are not expressed equal, there has been great interest on the research of alternative splicing and its regulatory mechanisms over the last years. Specific bioinformatic approaches directed to analyze RNA-Seq data have been developed. Thus, junction between exons can be detected in RNA-Seq data even in those cases where the isoform in unknown using splicer aligners such as TopHat and SOAPsplice [23]. Those junction reads should be unique to isoforms and may provide information regarding the expression level of the isoform, whereas reads mapped within an exon will be redundant across isoforms sharing that particular exon. As mentioned, RNA-Seq is potentially well-suited for alternative splicing analysis, but it is not free of constraints. The success on mapping junctions partially depends on read length and the sequence depth; however, the latter increases the false positives.

### 3.4 MicroRNA-Seq

In addition to the aforementioned applications, RNA-Seq techniques have the potential to analyze non-coding RNA molecules, such as microRNAs (miRNAs) [24]. MiRNAs are short (15-25 nucleotides) RNA sequences that play an important role in the regulation of gene expression in a number of biological processes in plants [25]. These RNA sequences may act post-transcriptionally by hybridizing to specific 3’ untranslated region in mRNA transcripts, to induce their subsequent degradation, or to inhibit their translation. The RNA-Seq analysis of miRNA requires specialized protocols for preparation of libraries that allow capturing the short miRNA sequences from RNA samples [26].

## 3. RNA-Seq IN FOOD SCIENCE

### 3.1. Production of food crops

Global warming and the demand for food of an ever-growing world population are relevant issues attracting much attention worldwide. In this regard, the study of the links existing between gene function and traits relevant to agriculture in food crops is of great interest. Transcriptomic analysis of crops provides valuable information of how genome responds to cellular perturbations, and also reveals the expressed genes that control important traits (e.g., yield and tolerance to adverse environmental conditions) [27]. Moreover, the detection of differential transcription patterns and identification of novel transcripts at specific stages of development or conditions establishes the foundation for understanding the molecular mechanisms underlying production of proteins and metabolites relevant to food science (e.g., bioactive compounds and nutrients). These insights provide further directions for controlling gene expression to increase or decrease accumulation of the compounds of interest.

In recent years, the genomes of several food crops have been fully sequenced. The progress in NGS techniques has contributed to the generation of comprehensive gene expression data sets of cell-, tissue- and developmental-specific gene expression for many food crop species. Data derived from NGS provide the starting point to discover the function of unknown genes and to describe the transcriptome throughout the life cycle for crop species relevant in food production, including rice, soybean, maize, and wheat, to mention few. In crops, adverse growing conditions often result in lower yields that have a negative economic impact for producers and consumers. Understanding the mechanisms involved in the response to unfavorable conditions will help on producing crops with higher tolerance to stress. To this regard, various gene expression profiling studies have been completed using NGS to investigate a variety of responses to drought, salinity, cold and diseases. Novel technological advances as those directed to increase the read length and the total number of reads per run combined with novel strategies that utilize paired-end reads have prompted its widespread use to decipher several food crops' transcriptomes. As mentioned above, the generation of longer paired-end reads enables higher levels of mappability, better identification of reads from splice variants, and the assembly of transcriptomes in the absence of a reference genome using *de novo* assembly approaches [18,28]. D*e novo* transcriptome assembly represents an emerging application of RNA-Seq particularly interesting for those food crop species whose genome has not been fully sequenced. Nevertheless, due to the complexity and frequently incomplete representation of transcripts in sequencing libraries, the assembly of high-quality transcriptomes can be challenging. In early RNA-Seq studies, the short 30 bp average read length restricted transcriptome assembly, whereas with the 75 bp or longer reads now available, transcriptomes can be assembled more easily, allowing reads whose ends are anchored in different exons to define splice sites without relying on prior annotations [28]. A strategy used in many studies aimed at capturing the most comprehensive transcript representation in the novel assembly relies on pooling samples obtained from different tissues and developmental stages for subsequent sequencing. After transcriptome assembly, validation or quality control for the new assembly output is frequently addressed using bioinformatic tools and searching in databases. This process allows to ascertain the degree of similarity/conservation between the novel assembly and other closely related transcriptomes and also, to address transcript and functional annotation. Furthermore, the use of mining tools on *de novo* assemblies provides excellent opportunities for the discovery of novel transcripts, as well as short sequence repeats (SSR), also known as microsatellites. These microsatellites consist on repetitions of short (two to six) nucleotides and are extremely useful for gene mapping, marker-assisted selection, and comparative genome analysis.

As a general trend, *de novo* transcriptome assembly studies reveal greater transcriptome complexity than expected and provide a blueprint for further studies. *De novo* transcriptome assembly of black pepper is a representative example of the application of SOLiD RNA-Seq technology in non-model species [29]. In that study, 71 million short reads, representing a sequencing coverage per base (depth) of 62X, were used to assemble a total of 22,363 transcripts in root samples. Transcript and functional annotation was performed based on the sequence homology with other species.

As stated before, paired-end reads method from Illumina technology has been commonly used for *de novo* transcriptome assembly in non-model food crop species. For instance, Zhang et al. investigated *de novo* assembly of peanut transcriptome with the aim at identifying expressed genes during the fast accumulation period in seeds [30]. Prior sequencing step, libraries were prepared from three peanut varieties including low and high oil content varieties. An average of 26 million paired-end Illumina reads were combined to form longer fragments (i.e., contigs) that, in turn, were used to form longer sequences (i.e., scaffolds), and ultimately unigene sequences. A comparison of the assembly with the NCBI protein database indicated that 42% of the unigenes (24,814 sequences) did not significantly match the mRNA database and were, thus, considered putative novel transcribed sequences. In addition to these findings, data mining using Perl script MISA enabled the detection of 5,883 microsatellites in 4,993 unigenes, demonstrating the extraordinary potential of RNA-Seq for the discovery of this type of polymorphisms. The transcriptomic study on sesame samples is another example of RNA-Seq application to study non-model species [31]. Transcriptome assembly was achieved by sequencing 24 paired-end cDNA libraries using Illumina technology. Also, a survey of the new transcriptome assembly for the presence of SSRs revealed more than ten thousands microsatellites in 42,566 unitranscrip sequences, many of them showing an uneven distribution in the transcriptome. In some studies, more than one sequencing platform has been used in parallel to address a novel transcriptome assembly. This strategy has shown to be particularly helpful for improving transcriptome assembly of bread wheat using about 16 million reads provided by Illumina and Roche sequencers [32]. The assembly of this complex polyploidy eukaryotic transcriptome was performed following a two-stage approach. First, a rough assembly was produced using the Velvet/Oases assembler, and then, reads in each cluster were re-assembled using the high-precision assembler MIRA. Using the FLX platform, the date palm fruit transcriptome has been investigated at seven different developmental stages [33]. An average of 1 million reads with a median length of 399 bp was obtained for each sequenced library. Interestingly, date fruit transcriptome showed high homology with grapevine sequences. In a separate report, *de novo* hop transcriptome assembly using Illumina sequencing has provided relevant information regarding the lupulin gland gene expression [34]. Transcriptomic data in combination with metabolite analysis provided evidence for the lupulin gland-specific BCAA and isoprenoid metabolism to produce precursors for bitter acids, which are important in brewing industry. Turmeric transcriptome has been assembled using a similar RNA-Seq approach with an Illumina sequencer [35]. In that case, pathway annotation of transcripts indicated that a number of expressed genes were related to the biosynthesis of secondary metabolites, some of them have been suggested to exert potential health-promoting effects. Also recently, the complexity of tea transcriptome has been investigated by RNA-Seq [36]. In that study, approximate 2.5 Gb were obtained with Illumina technology from different tea plant tissues. Processing and aligning the near 34.5 million 75-bp paired-end reads enabled the construction of 127,094 unigenes, a number 10-fold higher than existing sequences for tea plant in GeneBank. Some of the unigenes were annotated and assigned to putative metabolic pathways that are important to tea quality, such as flavonoid, theanine and caffeine biosynthesis.

In addition to the unique capability for *de novo* transcriptome assembly and improved transcriptome annotation, RNA-Seq is also useful for comparative transcriptomic analysis. As an example, Ono et al. employed Illumina RNA-Seq technology to to determine gene expression levels in fruit peel after accomplishing *de novo* assembly of pomegranate transcriptome [37]. In this case, transcript annotation based on homologues grape and *Arabidopsis* genomes, allowed the identification of putative gene sequences involved in the metabolism of terpenoid and phenolic compounds with a role in pigmentation, flavor and nutritional value of fruits. In a separate report, Feng et al. followed a different approach to investigate the transcriptomic changes during development and ripening of Chinese bayberry fruit [38]. To achieve that, each of the RNA samples from various tissues and fruit of different development and ripening stages were ligated with a different adaptor and sequenced in the same run using Illumina sequencing. The data produced from the mixed samples were used to construct the whole transcriptome assembly. In the second part of the study, the generated assembly was used as the reference, and data from each separate sample, identified by the adaptor sequences, were used to estimate the global differential gene expression between samples. Moreover, organic acid and sugar profiling data obtained with mass spectrometry-based methods were obtained as complementary information to the transcriptomic profiles in order to gain some insights on the metabolic pathways involved in fruit ripening, color development and taste quality. Similarly, Illumina RNA-Seq has been used for the study of different aspects in major food crops, such as for example the studies on rice including the investigation of seed development [39,40], the transcriptional response to nematode infection [41] and to drought stress [42]; differential gene expression between aleurone and starchy wheat endosperm [43]; and transcript profiling in maize [44], chickpea [45], and soybean [46]. Also, Illumina RNA-Seq technology has been applied to study the miRNA fraction in common bean [47] and rapeseed [48]. González-Ibeas et al. used barcoding strategy to prepare 10 RNA libraries from different melon plant tissues for further sequencing using Roche technology [49]. This strategy allowed the identification of conserved miRNAs, small interfering RNAs, as well as the discovery of potential melon-specific sequences miRNAs.

In addition to the aforementioned studies, Illumina DGE technology has been successfully applied to identify transcriptome differences in seven tissues on sweet potato [50]. Using another approach, Kalavacharla et al. used FLX platform in combination with barcode tagging for the quantitative analysis of gene expression in common bean [51]. Moreover, tagging cDNA libraries was very helpful on verifying and validating global gene expression patterns, and detecting both shared and unique transcripts among the analyzed bean tissues. Also, the study by Garg et al. on chickpea has demonstrated the great potential of FLX sequencer for differential gene expression analysis [52]. In the initial stage of the study *de novo* transcriptome assembly of chickpea was obtained from nearly 2 million short reads. For the assembly process, authors assayed eight different programs highlighting the importance of optimizing the assembly procedure.

### 3.2 Foodborne pathogenic microorganisms

One of the main goals for the food industry is the production of safe foods with the desired quality using minimal processing technologies. Foodborne disease, commonly referred to as food poisoning, occurs when food becomes contaminated with harmful species. Although chemical species such as pesticides, among others, can originate important health problems, the vast majority of food poisonings are the direct result of microbiological hazards induced by bacteria, toxigenic molds and microalgae, viruses, and parasites. Over the past years, the availability of genome sequences of relevant food microorganisms has given rise to extraordinary possibilities for the study the molecular mechanisms in complex biological processes such as food spoilage and biofilm formation [53,54]. In this field, RNA-Seq offers great potential to investigate the activities of foodborne microorganisms under strictly controlled conditions in the laboratory as well as in industrial environments or in food products. Despite its good potential, RNA-Seq methods have been applied less frequently to study foodborne pathogens than to investigate food crops. Regarding sample preparation, enrichment for all transcripts other than the abundant rRNA and tRNA species in RNA samples can be challenging, especially for bacterial transcriptomes, lacking mRNAs with poly(A) tail. A frequent solution to this problem involves 16S and 23S rRNA depletion from total RNA fraction isolated from microbial cells. With regard to the investigation on foodborne pathogens, RNA-Seq has been highly valuable in providing global transcriptomic profiles of persistent and nonpersistent *Listeria monocytogenes* isolates [55]. This foodborne pathogen is the causative bacterium of serious invasive disease in animals and in humans. The contamination of food processing facilities and food products with this *L. monocytogenes* is of particular concern because it survives extreme environmental conditions and has the ability to form resistant biofilms. The RNA-Seq study by Fox et al. was focused on the comparison of both bacterial strains in response to the treatment with benzethonium chloride, a disinfectant used in food-processing industry. RNA-Seq data suggested that treatment induced a complex peptidoglycan biosynthesis response, which may play a key role in disinfectant resistance. Also, RNA-Seq has provided the information to generate transcription start site maps for pathogenic and non-pathogenic Listeria species [56]. In that work, the discovery of novel long antisense RNA species using this methodology suggested new mechanisms for the regulation of gene expression in bacteria. Also, the potential of RNA-Seq for the investigation of microorganisms in complex food matrices has been demonstrated in a recent study of *Salmonella* in peanut oil [57]. Interestingly, the study revealed that desiccated bacterial cells in peanut oil were in physiologically dormant state with a low portion (<5%) of its genome being transcribed.

### 3.3 Food fermentations

A major focus in food biotechnology research is directed to the investigation of cellular and molecular processes involved in industrially relevant microorganisms that are responsible for food fermentations. The acquired information obtained from such studies may ultimately help fermentation-based industries to enhance the quality of the final product, to improve their product yields, and even to develop novel foods. In this research field, high-throughput transcriptomic tools assist in elucidating the molecular mechanisms behind interesting metabolic transformations and functionalities in fermented food ecosystems [58]. Over the last years, the complete genomes of significant fermenting microorganisms, including yeasts and bacteria species, have been released into public databases. These genomic data resources have been essential to develop several species-specific microarrays that enable the study of gene expression under different conditions, providing new insight into important metabolic processes. For instance, transcriptome profiling of the yeast *Saccharomyces cerevisiae* and lactic acid bacteria (LAB) has contributed to improve our knowledge about cellular processes and responses of these organisms in different environments. Recent microarray applications in this field include the study of fermentation-related stress factors on the transcriptional response for laboratory or industrial wine, lager brewing and baker's yeast strains [59-

61], gene expression dynamics during different fermentation stages in synthetic media or natural substrates [62], and transcriptional differences between diverse strains and mutants [63]. In addition to yeasts, LAB have also industrial relevance since this group of microorganisms has the ability to provide the key flavor, texture, and preservative qualities to variety of fermented foods such as sourdoughs, dairy products, and fermented sausages [64].

Although gene expression microarray has become a powerful tool in this research field, RNA-Seq has taken gene expression analysis to a higher level in terms of improving the possibilities for investigating novel aspects of transcriptomes in fermenting microorganisms [65]. For instance, many aspects of the transcriptome structure of *S. cerevisiae* have been elucidated using Illumina RNA-Seq method. In their pioneer work, Ngalakshmi et al. identified alternative initiation codons, upstream open reading frames, the presence of several overlapping genes and unexpected 3'-end heterogeneity in the yeast transcriptome [66]. Interestingly, RNA-Seq data also indicated that about 75% of the nonrepetitive sequence of the yeast genome was transcribed. In another report, RNA-Seq has also provided remarkable insight into the transcriptome of Aspergillus oryzae, a mold used in various oriental fermented foods [67]. Among the discoveries using paired-end reads Illumina technology, authors highlighted the identification of novel transcripts, new exons, untranslated regions, alternative splicing isoforms, alternative upstream initiation codons and upstream open reading frames. Moreover, gene expression profiling indicated that the mold showed superior protein production grown under solid substrate than in liquid culture. Although the application of RNA-Seq to food metatranscriptomics studies is still lacking, this technique is well suited for the complex uncharacterized microbial ecosystems [68,69]. In a near future, it can be expected that RNA-Seq will be applied to investigate the activities involved in global metabolic processes in fermented foods, and to decipher the temporal contribution of each species in food ecosystems. Such analyses will constitute the foundation for constructing a system level understanding of microbial activity in complex food ecosystems.

## 4. FUTURE OUTLOOKS AND CONCLUSIONS

RNA-Seq techniques have demonstrated their impressive analytical potential for gene expression studies in the context of food science. RNA-Seq has the potential to quickly supersede microarray in many gene expression studies. However, RNA-Seq technology still remains evolving and several technical and bioinformatics challenges need to be overcome to realize the full potential of this technique in food science. Despite the extraordinary reductions in cost per sequenced base associated with SGS in comparison with more conventional sequencing methods, the application of RNA-Seq to survey the transcriptome (structure and expression levels) is still expensive. Thus, it might be expected that cheaper and faster library preparation methods will be developed to decrease the costs of generating sequencing data in the near future. Also, the development of novel high-throughput enrichment strategies, such as in-solution or in-microarray capture methods, aimed to target specific sequences of interest, will broaden the applicability of RNA-Seq in food science, since such strategies have the potential to improve the sequencing of low-abundance transcripts without the need for increasing the costly sequencing depth [70].

Given the evolution path of RNA-Seq technology, high-end instruments with higher sequencing throughput able to provide longer and accurate reads can be expected in the very near future. For instance, a new generation of sequencers (third-generation sequencing, TGS), based on single-molecule sequencing, is rapidly emerging. An outstanding advantage of these novel approaches is that do not require routine PCR amplification prior sequencing, thereby avoiding systematic amplification bias occurring in SGS. In addition to their capability for sequencing RNA directly with the corresponding savings in reagents and manpower, novel technologies can also sequence molecules in real-time, decreasing the time of analysis and allowing longer read lengths. TGS systems, including PacBio-, nanopore-sequencing technologies, and direct imaging of individual DNA molecules using advanced microscopy techniques have been recently reviewed [4]. In addition to the high-end sequencers, the recent commercialization of bench-top instruments, including the 454 GS Junior (Roche), Ion PGM (Life Technologies), and MiSeq (Illumina), seeking lower cost and time of analysis will bring about RNA-Seq to gain more popularity in food research laboratories in coming years [71].

As it has been discussed in this chapter, transcriptomic profiling of food crops is a hot application for RNA-Seq technologies. Nevertheless, the applicability of RNA-Seq in Food Science and Nutrition is still in its infancy and it is not fully exploited yet. In the near future, it can be anticipated that many relevant topics in Food Science will benefit from RNA-Seq techniques. For instance in Nutrigenomics studies, it can be expected that RNA-Seq studies will improve our limited understanding of the roles of nutritional compounds at molecular level. Also, RNA-Seq may be a suitable profiling tool to characterize and investigate different safety aspects related with genetically modified organisms (GMOs). Novel aspects related with food pathogens will be addressed such as for example, the link between pathogen and host, pathogenesis, and virulence factors. For microorganisms that produce toxins in food products, understanding the molecular mechanisms for toxins and their secondary metabolites production is a key point to limit the toxin contamination in food products and food processing facilities. In consequence, the availability of tools such RNA-Seq, capable to provide detailed information about the transcriptional regulation of the metabolic pathways implicated in the biosynthesis of toxins in toxigenic microorganism in real samples may be very valuable in ongoing and future research.

## Acknowledgments

This work was supported by AGL2011-29857-C03-01 project (Ministerio de Economía y Competitividad, Spain), and CSD2007-00063 FUN-C-FOOD (Programa CONSOLIDER, Ministerio de Educación y Ciencia, Spain). A.V. thanks the Ministerio de Economía y Competitividad for his FPI pre-doctoral fellowship.

## REFERENCES

[1] C.S. Pareek, R. Smoczynski and A. Tretyn, J. Appl. Genet. 52:413-435, 2011

[2] Z. Wang, M. Gerstein and M. Snyder, Nat. Rev. Genet. 10:57-63, 2009

[3] J.H. Malone and B. Oliver, BMC Biology 9:34, 2011

[4] T.P. Niedringhaus, D. Milanova, M.B. Kerby, M.P. Snyder and A.E. Barron, Anal. Chem. 83:4327-4341, 2011

[5] O. Morozova and M.A. Marra, Genomics 92:255-264, 2008

[6] E.R. Mardis, Annu. Rev. Genomics Hum. Genet. 9:387-402, 2008

[7] D.R. Bentley, S. Balasubramanian, H.P. Swerdlow, G.P. Smith, J. Milton, C.G. Brown, K.P. Hall, D.J. Evers, et al., Nature 456:53-59, 2008

[8] M. Margulies, M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, et al., Nature 435:376-380, 2005

[9] M.L. Metzker, Nat. Rev. Genet. 11:31-46, 2010

[10] S. Datta, S. Datta, S. Kim, S. Chakraborty and R.S. Gill, J. Proteomics Bioinform. 3:183-190, 2010

[11] P.A. McGettigan, Curr. Opin. Chem. Biol. 17:4-11, 2013

[12] L.D. Stein, Curr. Protoc. Bioinformatics. 36:11.1.1., 2011

[13] H. Hong, W. Zhang, J. Shen, Z. Su, B. Ning, T. Han, R. Perkins, L. Shi and W. Tong, Sci China Life Sci. 56:110-118, 2013

[14] M.B. Scholz, C.C. Lo and P.S. Chain, Curr. Opin. Biotechnol. 23:9-15, 2012

[15] N. Blow, Nature 458:239-242, 2009

[16] S. Marguerat, B.T. Wilhelm and J. Bähler, Biochem. Soc. Trans. 36:1091-1096, 2008

[17] K. Mutz, A. Heilkenbrinker, M. Lönne, J. Walter and F. Stahl, Curr. Opin. Biotech. 24:22-30, 2013

[18] F. Ozsolak and P.M. Milos, Nat. Rev. Genet. 12:87-98, 2011

[19] B.T. Wilhelm and J. Landry, Methods 48:249-257, 2009

[20] A. Mortazavi, B.A. Williams, K. McCue, L. Schaeffer and B. Wold, Nat. Methods 5:621-628, 2008

[21] X. Tao, Y. Gu, H. Wang, W. Zheng, X. Li, C. Zhao and Y. Zhang, PLoS ONE 7:e36234, 2012

[22] H. Feng, Z. Qin and X. Zhang, 340:179–191, 2013

[23] J.W. Malcom and J.H. Malone in C. Simó, A. Cifuentes, and V. García-Cañas (Eds.), Fundamentals of Advanced Omics Technologies: From Genes to Metabolites, Elsevier, In press.

[24] Q. Zhou, R. Gallagher, R. Ufret-Vincenty, X. Li, E.N. Olson and S. Wang, Proc. Natl. Acad. Sci. USA 108:8287-8292, 2011

[25] M.A. Edwards and R.J. Henry, J. Cereal Sci. 54:395-400, 2011

[26] D.W. Wegman and S.N. Krylov, Trend Anal. Chem. 44:121-130, 2013

[27] W.A. Rensink and R. Buell, Trends Plant Sci. 10:603-609, 2005

[28] M.K Iyer and A.M Chinnaiyan, Nat. Biotechnol. 29:599-600, 2011

[29] S.M. Gordo, D.G. Pinheiro, E.C. Moreira, S.M. Rodrigues, M.C. Poltronieri, O.F. de Lemos, I.T. da Silva, R.T. Ramos et al., BMC Plant Biol. 12:168, 2012

[30] H. Zhang, L. Wei, H. Miao, T. Zhang and C. Wang, BMC Genomics 13:316, 2012

[31] J. Zhang, S. Liang, J. Duan, J. Wang, S. Chen, Z. Cheng, Q. Zhang, X. Liang, et al., BMC Genomics 13:90, 2012

[32] A.W. Schreiber, M.J. Hayden, K.L. Forrest, S.L. Kong, P. Langridge and U. Baumann, BMC Genomics 13:492, 2012

[33] Y. Yin, X. Zhang, Y. Fang, L. Pan, G. Sun, C. Xin, M.M. Ba Abdullah, X. Yu, et al., Plant Mol. Biol. 78:617-626, 2012

[34] S.M. Clark, V. Vaitheeswaran, S.J. Ambrose, R.W. Purves and J.E. Page, BMC Plant Biol. 13:12, 2013

[35] R.S. Annadurai, R. Neethiraj, V. Jayakumar, A.C. Damodaran, S.N. Rao, M.A.V.S.K. Katta, S. Gopinathan, S.P. Sarma, et al., PLoS ONE 8:e56217, 2013

[36] C. Shi, H. Yang, C. Wei, O. Yu, Z. Zhang, C. Jiang, J. Sun, Y. Li, et al., BMC Genomics 12:131, 2011

[37] N.N. Ono, M.T. Britton, J.N. Fass, C.M. Nicolet, D. Lin and L. Tian, J. Integr. Plant Biol. 53:800-813, 2011

[38] C. Feng, M. Chen, C. Xu, L. Bai, X. Yin, X. Li, A.C. Allan, I.B. Ferguson, et al., BMC Genomics 13:19, 2012

[39] H. Xu, Y. Gao and J. Wang, PLoS ONE 7(2): e30646. doi:10.1371/journal.pone.0030646

[40] R.M. Davidson, M. Gowda, G. Moghe, H. Lin, B. Vaillancourt, S.-H. Shiu, N. Jiang and C.R. Buell, Plant J. 71:492–502, 2012

[41] T. Kyndt, S. Denil, A. Haegeman, G. Trooskens, L. Bauters, W. Van Criekinge, T. De Meyer and G. Gheysen, New Phytol. 196:887–900, 2012

[42] W. Zong, X. Zhong, J. You, and L. Xiong, Plant Mol. Biol. 81:175–188, 2013

[43] S.A. Gillies, A. Futardo, and R.J. Henry, Plant Biotechnol. J. 10:668–679, 2012

[44] X. Hui Xu, H. Chen, Y.L. Sang, F. Wang, J. Ma, X. Gao and X.S. Zhang, BMC Genomics 13:294, 2012

[45] R. Garg, R.K. Patel, S. Jhanwar, P. Priya, A. Bhattacharjee, G. Yadav, S. Bhatia, D. Chattopadhyay, A. K. Tyagi, and M. Jain, Plant Physiol 156:1661-1678, 2011

[46] H. Chen, F.W. Wang, Y.Y. Dong, N. Wang, Y.P. Sun, X.Y. Li, L. Liu, X.D. Fan, H.L. Yin, Y.Y. Jing, X.Y. Zhang, Y.L. Li, G. Chen and H.Y. Li, BMC Plant Biol. 12:122, 2012

[47] P. Peláez, M.S. Trejo, L.P. Iñiguez, G. Estrada-Navarrete, A.A. Covarrubias, J.L. Reyes and F. Sanchez, BMC Genomics 13:83, 2012

[48] A.P. Körbes, R.D. Machado, F. Guzman, M.P. Almerão, L.F.V. de Oliveira, G. Loss-Morais, A.C. Turchetto-Zolet, A. Cagliari, et al., PLoS ONE 7:e50663, 2012

[49] D. Gonzalez-Ibeas, J. Blanca, L. Donaire, M. Saladié, A. Mascarell-Creus, A. Cano-Delgado, J. Garcia-Mas, C. Llave, et al., BMC Genomics 12:393, 2011

[50] X. Tao, Y.H. Gu, H. Wang, W. Zheng, X. Li, C.W. Zhao and Y.Z. Zhang, PLoS ONE 7(4):e36234, 2012

[51] V. Kalavacharla, Z. Liu, B.C. Meyers, J. Thimmapuram and K. Melmaiee, BMC Plant Biol. 11:135, 2011

[52] R. Garg, R.K. Patel, S. Jhanwar, P. Priya, A. Bhattacharjee, G. Yadav, S. Bhatia, D. Chattopadhyay, et al., Genome Anal. 156:1661-1678, 2011

[53] S. Puttamreddy, M.D. Carruthers, M.L. Madsen and F.C. Minion, Foodborne Pathog. Dis. 5:517-529, 2008

[54] H.L. Andrews-Polymenis, C.A. Santiviago and M. McClelland, Curr. Opin. Biotech. 20:149-157, 2009

[55] E.M. Fox, N. Leonard and K. Jordan, Appl. Environ. Microbiol. 77:6559-6569, 2011

[56] O. Wurtzel, N. Sesto, J.R. Mellin, I. Karunker, S. Edelheit, C. Becavin, C. Archambaud, P. Cossart, et al., Mol. Syst. Biol. 8:583, 2012

[57] X. Deng, Z. Li and W. Zhang, Food Microbiol. 30:311-315, 2012

[58] N.A. Bokulich and D.A. Mills, BMB Rep. 45:377-389, 2012

[59] J. Shima, S. Kuwazaki, F. Tanaka, H. Watanabe, H. Yamamoto, R. Nakajima, T. Tokashiki and H. Tamura, Int. J. Food Microbiol. 102:63-71, 2005

[60] S.L. Tai, P. Daran-Lapujade, M.C. Walsh, J.T. Pronk and J. Daran, Mol. Biol. Cell 18:5100-5112, 2007

[61] T. Rossignol, O. Postaire, J. Storaï and B. Blondin, Appl. Microbiol. Biotechnol. 71:699-712, 2006

[62] V. Penacho, E. Valero and R. Gonzalez, Int. J. Food Microbiol. 153:176-182, 2012

[63] E. Bartra, M. Casado, D. Carro, C. Campama and B. Piña, J. Appl. Microbiol. 109:272-281, 2010

[64] T.R. Klaenhammer, M.A. Azcarate-Peril, E. Altermann and R. Barrangou, J. Nutr. 137:748S-750S, 2007

[65] L. Solieri, T.C. Dakal and P. Giudici, Ann. Microbiol. 63:21-37, 2013

[66] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein and M. Snyde, Science 320:1344-1349, 2008

[67] B. Wang, G. Guo, C. Wang, Y. Lin, X. Wang, M. Zhao, Y. Guo, M. He, Y. Zhang and L. Pan, Nucleic Acids Res. 38:5075-5087, 2010

[68] U. Mäder, P. Nicolas, H. Richard, P. Bessieres and S. Aymerich, Curr. Opin. Biotech. 22:32-41, 2011

[69] N.P. McNulty, T. Yatsunenko, A. Hsiao, J.J. Faith, B.D. Muegge, A.L. Goodman, B. Henrissat, R. Oozeer, et al., Sci. Transl. Med. 3:106ra106, 2011

[70] T.R. Mercer, D.J. Gerhardt, M.E. Dinger, J. Crawford, C. Trapnell, J.A. Jeddeloh, J.S. Mattick and J.L. Rinn, Nat. Biotechnol. 30:99-104, 2012

[71] N.J. Loman, C. Constantinidou, J.Z.M. Chan, M. Halachev, M. Sergeant, C.W. Penn, E.R. Robinson and M.J. Pallen, Nat. Rev. Microbiol. 10:599-606, 2012

## FIGURE LENGEDS

**Figure 1.** Sequencing scheme in Illumina Genome Analyzer platform.

**Figure 2.** Sequencing procedure in Roche 454 Genome sequencer FLX platform.

**Figure 3.** Sequencing strategy in Applied Biosystems SOLiD platform.

**Table 1.** Comparison of main NGS platforms.

| | **Illumina**<br>**GA (II)[a] and HiSeq[b]** | **Roche**<br>**454 GS FLX +** | **Applied Biosystems**<br>**SOLiD 5500xl** | **Life Technologies**<br>**Ion PGM (318)** |
|---|---|---|---|---|
| **Amplification** | Bridge amp. | Emulsion PCR | Emulsion PCR | Emulsion PCR |
| **Sequencing chemistry** | Reversible terminator | Pyrosequencing | Ligation | Proton detection |
| **Method of detection** | Fluorescence | Chemiluminescence | Fluorescence | Change in pH |
| **Run time** | 14 days[a] or 11 days[b] | 23 hours | 8 days | 4 hours |
| **Read length** | 75 bp[a] or 100 bp[b] | ~ 800 bp | 75 + 35 bp | 100 or 200 bp |
| **Millions of reads/run** | 40[a] or 3,000[b] | 1 | 1,400 | 4 |
| **Data generation/run** | 35 Gb[a] or 600 Gb[b] | 0.7 Gb | 155 Gb | 0.86 Gb |
| **Advantage** | -Cost-effectiveness<br>-Massive throughput<br>-Low hands-on time | -Long read length improves mapping in repetitive regions<br>-Short run time | -Low error rate<br>-Massive throughput | -Short run times<br>-No need for modified DNA bases |
| **Disadvantage** | -Long run time<br>-Short read lengths | -High reagent costs<br>-High error rate in homopolymer repeats<br>-High hands-on time | -Short read lengths<br>-Long run time | -High error rate in homopolymers repeats<br>-High hands-on time |

[a] Illumina/GA(II) instrument

[b] Illumina/HiSeq 2000 instrument