Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 853916, 19 pages
doi:10.1155/2010/853916

Review Article

Uncovering the Complexity of Transcriptomes with RNA-Seq

Valerio Costa,[1] Claudia Angelini,[2] Italia De Feis,[2] and Alfredo Ciccodicola[1]

[1] *Institute of Genetics and Biophysics "A. Buzzati-Traverso", IGB-CNR, 80131 Naples, Italy*
[2] *Istituto per le Applicazioni del Calcolo "Mauro Picone", IAC-CNR, 80131 Naples, Italy*

Correspondence should be addressed to Valerio Costa, costav@igb.cnr.it

Received 22 February 2010; Accepted 7 April 2010

Academic Editor: Momiao Xiong

In recent years, the introduction of massively parallel sequencing platforms for Next Generation Sequencing (NGS) protocols, able to simultaneously sequence hundred thousand DNA fragments, dramatically changed the landscape of the genetics studies. RNA-Seq for transcriptome studies, Chip-Seq for DNA-proteins interaction, CNV-Seq for large genome nucleotide variations are only some of the intriguing new applications supported by these innovative platforms. Among them RNA-Seq is perhaps the most complex NGS application. Expression levels of specific genes, differential splicing, allele-specific expression of transcripts can be accurately determined by RNA-Seq experiments to address many biological-related issues. All these attributes are not readily achievable from previously widespread hybridization-based or tag sequence-based approaches. However, the unprecedented level of sensitivity and the large amount of available data produced by NGS platforms provide clear advantages as well as new challenges and issues. This technology brings the great power to make several new biological observations and discoveries, it also requires a considerable effort in the development of new bioinformatics tools to deal with these massive data files. The paper aims to give a survey of the RNA-Seq methodology, particularly focusing on the challenges that this application presents both from a biological and a bioinformatics point of view.

1. Introduction

It is commonly known that the genetic information is conveyed from DNA to proteins via the messenger RNA (mRNA) through a finely regulated process. To achieve such a regulation, the concerted action of multiple cis-acting proteins that bind to gene flanking regions—"core" and "auxiliary" regions—is necessary [1]. In particular, core elements, located at the exons' boundaries, are strictly required for initiating the pre-mRNA processing events, whereas auxiliary elements, variable in number and location, are crucial for their ability to enhance or inhibit the basal splicing activity of a gene.

Until recently—less than 10 years ago—the central dogma of genetics indicated with the term "gene" a DNA portion whose corresponding mRNA encodes a protein. According to this view, RNA was considered a "bridge" in the transfer of biological information between DNA and proteins, whereas the identity of each expressed gene, and of its transcriptional levels, were commonly indicated as "transcriptome" [2]. It was considered to mainly consist of ribosomal RNA (80–90%, rRNA), transfer RNA (5–15%, tRNA), mRNA (2–4%) and a small fraction of intragenic (i.e., intronic) and intergenic noncoding RNA (1%, ncRNA) with undefined regulatory functions [3]. Particularly, both intragenic and intergenic sequences, enriched in repetitive elements, have long been considered genetically inert, mainly composed of "junk" or "selfish" DNA [4]. More recently it has been shown that the amount of noncoding DNA (ncDNA) increases with organism complexity, ranging from 0.25% of prokaryotes' genome to 98.8% of humans [5]. These observations have strengthened the evidence that ncDNA, rather than being junk DNA, is likely to represent the main driving force accounting for diversity and biological complexity of living organisms.

Since the dawn of genetics, the relationship between DNA content and biological complexity of living organisms has been a fruitful field of speculation and debate [6]. To date, several studies, including recent analyses performed during the ENCODE project, have shown the pervasive nature of eukaryotic transcription with almost the full length of nonrepeat regions of the genome being transcribed [7].

The unexpected level of complexity emerging with the discovery of endogenous small interfering RNA (siRNA) and microRNA (miRNA) was only the tip of the iceberg [8]. Long interspersed noncoding RNA (lincRNA), promoter- and terminator-associated small RNA (PASR and TASR, resp.), transcription start site-associated RNA (TSSa-RNA), transcription initiation RNA (tiRNA) and many others [8] represent part of the interspersed and crosslinking pieces of a complicated transcription puzzle. Moreover, to cause further difficulties, there is the evidence that most of the pervasive transcripts identified thus far, have been found only in specific cell lines (in most of cases in mutant cell lines) with particular growth conditions, and/or particular tissues. In light of this, discovering and interpreting the complexity of a transcriptome represents a crucial aim for understanding the functional elements of such a genome. Revealing the complexity of the genetic code of living organisms by analyzing the molecular constituents of cells and tissues, will drive towards a more complete knowledge of many biological issues such as the onset of disease and progression.

The main goal of the whole transcriptome analyses is to identify, characterize and catalogue all the transcripts expressed within a specific cell/tissue—at a particular stage—with the great potential to determine the correct splicing patterns and the structure of genes, and to quantify the differential expression of transcripts in both physio- and pathological conditions [9].

In the last 15 years, the development of the hybridization technology, together with the tag sequence-based approaches, allowed to get a first deep insight into this field, but, beyond a shadow of doubt, the arrival on the marketplace of the NGS platforms, with all their "*Seq*" applications, has completely revolutionized the way of thinking the molecular biology.

The aim of this paper is to give an overview of the RNA-Seq methodology, trying to highlight all the challenges that this application presents from both the biological and bioinformatics point of view.

2. Next Generation Sequencing Technologies

Since the first complete nucleotide sequence of a gene, published in 1964 by Holley [10] and the initial developments of Maxam and Gilbert [11] and Sanger et al. [12] in the 1970s (see Figure 1), the world of nucleic acid sequencing was a RNA world and the history of nucleic acid sequencing technology was largely contained within the history of RNA sequencing.

In the last 30 years, molecular biology has undergone great advances and 2004 will be remembered as the year that revolutionized the field; thanks to the introduction of massively parallel sequencing platforms, the *Next Generation Sequencing*-era, [13–15], started. Pioneer of these instruments was the Roche (454) Genome Sequencer (GS) in 2004 (http://www.454.com/), able to simultaneously sequence several hundred thousand DNA fragments, with a read length greater than 100 base pairs (bp). The current GS FLX Titanium produces greater than 1 million reads in excess of 400 bp. It was followed in 2006 by the Illumina Genome Analyzer (GA) (http://www.illumina.com/) capable to generate tens of millions of 32-bp reads. Today, the Illumina GAIIx produces 200 million 75–100 bp reads. The last to arrive in the marketplace was the Applied Biosystems platform based on Sequencing by Oligo Ligation and Detection (SOLiD) (http://www3.appliedbiosystems.com/AB_Home/index.htm), capable of producing 400 million 50-bp reads, and the Helicos BioScience HeliScope (http://www.helicosbio.com/), the first single-molecule sequencer that produces 400 millions 25–35 bp reads.

While the individual approaches considerably vary in their technical details, the essence of these systems is the miniaturization of individual sequencing reactions. Each of these miniaturized reactions is seeded with DNA molecules, at limiting dilutions, such that there is a single DNA molecule in each, which is first amplified and then sequenced. To be more precise, the genomic DNA is randomly broken into smaller sizes from which either fragment templates or mate-pair templates are created. A common theme among NGS technologies is that the template is attached to a solid surface or support (immobilization by primer or template) or indirectly immobilized (by linking a polymerase to the support). The immobilization of spatially separated templates allows simultaneous thousands to billions of sequencing reactions. The physical design of these instruments allows for an optimal spatial arrangement of each reaction, enabling an efficient readout by laser scanning (or other methods) for millions of individual sequencing reactions onto a standard glass slide. While the immense volume of data generated is attractive, it is arguable that the elimination of the cloning step for the DNA fragments to sequence is the greatest benefit of these new technologies. All current methods allow the direct use of small DNA/RNA fragments not requiring their insertion into a plasmid or other vector, thereby removing a costly and time-consuming step of traditional Sanger sequencing.

It is beyond a shadow of doubt that the arrival of NGS technologies in the marketplace has changed the way we think about scientific approaches in basic, applied and clinical research. The broadest application of NGS may be the resequencing of different genomes and in particular, human genomes to enhance our understanding of how genetic differences affect health and disease. Indeed, these platforms have been quickly applied to many genomic contexts giving rise to the following "Seq" protocols: RNA-Seq for transcriptomics, Chip-Seq for DNA-protein interaction, DNase-Seq for the identification of most active regulatory regions, CNV-Seq for copy number variation, and methyl-Seq for genome wide profiling of epigenetic marks.

3. RNA-Seq

RNA-Seq is perhaps one of the most complex next-generation applications. Expression levels, differential splicing, allele-specific expression, RNA editing and fusion transcripts constitute important information when comparing samples for disease-related studies. These attributes, not

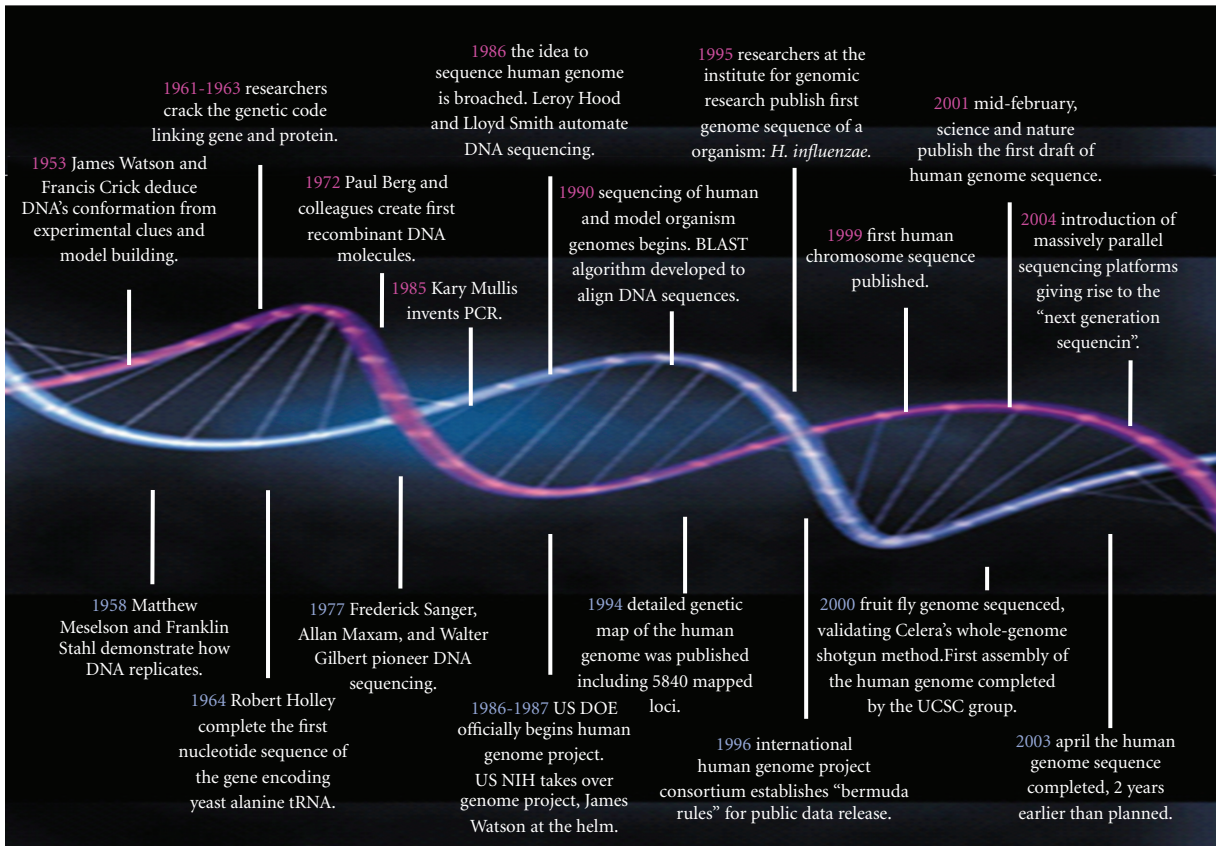

Figure 1: Evolution of DNA revolution.

readily available by hybridization-based or tag sequence-based approaches, can now be far more easily and precisely obtained if sufficient sequence coverage is achieved. However, many other essential subtleties in the RNA-Seq data remain to be faced and understood.

Hybridization-based approaches typically refer to the microarray platforms. Until recently, these platforms have offered to the scientific community a very useful tool to simultaneously investigate thousands of features within a single experiment, providing a reliable, rapid, and cost-effective technology to analyze the gene expression patterns. Due to their nature, they suffer from background and cross-hybridization issues and allow researchers to only measure the relative abundance of RNA transcripts included in the array design [16]. This technology, which measures gene expression by simply quantifying—via an indirect method—the hybridized and labeled cDNA, does not allow the detection of RNA transcripts from repeated sequences, offering a limited dynamic range, unable to detect very subtle changes in gene expression levels, critical in understanding any biological response to exogenous stimuli and/or environmental changes [9, 17, 18].

Other methods such as Serial, Cap Analysis of Gene Expression (SAGE and CAGE, resp.) and Polony Multiplex Analysis of Gene Expression (PMAGE), tag-based sequencing methods, measure the absolute abundance of transcripts in a cell/tissue/organ and do not require prior knowledge of any gene sequence as occurs for microarrays [19]. These analyses consist in the generation of sequence tags from fragmented cDNA and their following concatenation prior to cloning and sequencing [20]. SAGE is a powerful technique that can therefore be viewed as an unbiased digital microarray assay. However, although SAGE sequencing has been successfully used to explore the transcriptional landscape of various genetic disorders, such as diabetes [21, 22], cardiovascular diseases [23], and Downs syndrome [24, 25], it is quite laborious for the cloning and sequencing steps that have thus far limited its use.

In contrast, RNA-Seq on NGS platforms has clear advantages over the existing approaches [9, 26]. First, unlike hybridization-based technologies, RNA-Seq is not limited to the detection of known transcripts, thus allowing the identification, characterization and quantification of new splice isoforms. In addition, it allows researchers to determine the correct gene annotation, also defining—at single nucleotide resolution—the transcriptional boundaries of genes and the expressed Single Nucleotide Polymorphisms (SNPs). Other advantages of RNA-Seq compared to microarrays are the low "background signal," the absence of an upper limit for quantification and consequently, the larger dynamic range of expression levels over which transcripts can be detected. RNA-Seq data also show high levels of reproducibility for both technical and biological replicates.

Table 1: Selection of papers on mammalian RNA-Seq.

Reference	Organism	Cell type/tissue	NGS platform
Bainbridge et al., 2006 [27]	*Homo sapiens*	Prostate cancer cell line	Roche
Cloonan et al., 2008 [30]	*Mus musculus*	ES cells and Embryoid bodies	ABI
Core et al., 2008 [31]	*Homo sapiens*	Lung fibroblasts	IIlumina
Hashimoto et al., 2008 [32]	*Homo sapiens*	HT29 cell line	ABI
Li et al., 2008 [33]	*Homo sapiens*	Prostate cancer cell line	IIlumina
Marioni et al., 2008 [34]	*Homo sapiens*	Liver and kidney samples	IIlumina
Morin et al., 2008 [35]	*Homo sapiens*	ES cells and Embryoid bodies	IIlumina
Morin et al., 2008 [36]	Homo sapiens	HeLa S3 cell line	IIlumina
Mortazavi et al., 2008 [37]	*Mus musculus*	Brain, liver and skeletal muscle	IIlumina
Rosenkran et al., 2008 [38]	*Mus musculus*	ES cells	IIlumina
Sugarbaker et al., 2008 [39]	*Homo sapiens*	Malignant pleural mesothelioma, adenocarcinoma and normal lung	Roche
Sultan et al., 2008 [40]	*Homo sapiens*	Human embryonic kidney and B cell line	IIlumina
Asmann et al., 2009 [41]	*Homo sapiens*	Universal and brain human reference RNAs	IIlumina
Chepelev et al., 2009 [42]	*Homo sapiens*	Jurkat and $GD4^+$ T cells	IIlumina
Levin et al., 2009 [43]	*Homo sapiens*	K562	IIlumina
Maher et al., 2009 [44]	*Homo sapiens*	Prostate cancer cell lines	Roche IIlumina
Parkhomchuk et al., 2009 [45]	*Mus musculus*	Brain	IIlumina
Reddy et al., 2009 [46]	*Homo sapiens*	A549 cell line	IIlumina
Tang et al., 2009 [47]	*Mus musculus*	Blastomere and oocyte	ABI
Blekhman et al., 2010 [48]	*Homo sapiens,* *Pan troglodytes,* *Rhesus macaca.*	Liver	IIlumina
Heap et al., 2010 [49]	*Homo sapiens*	Primary $GD4^+$ T cells	IIlumina
Raha et al., 2010 [50]	*Homo sapiens*	K562 cell line	IIlumina

Recent studies have clearly demonstrated the advantages of using RNA-Seq [27–50]. Table 1 provides a short description of recent and more relevant papers on RNA-Seq in mammals.

Many research groups have been able to precisely quantify known transcripts, to discover new transcribed regions within intronic or intergenic regions, to characterize the antisense transcription, to identify alternative splicing with new combinations of known exon sequences or new transcribed exons, to evaluate the expression of repeat elements and to analyze a wide number of known and possible new candidate expressed SNPs, as well as to identify fusion transcripts and other new RNA categories.

3.1. Sample Isolation and Library Preparation. The first step in RNA-Seq experiments is the isolation of RNA samples; further RNA processing strictly depends on the kind of analysis to perform. Indeed, as "transcriptome" is defined as the complete collection of transcribed elements in a genome (see [2]), it consists of a wide variety of transcripts, both mRNA and non-mRNA, and a large amount (90–95%) of rRNA species. To perform a whole transcriptome analysis, not limited to annotated mRNAs, the selective depletion of abundant rRNA molecules (5S, 5.8S, 18S and 28S) is a key step. Hybridization with rRNA sequence-specific 5′-biotin labeled oligonucleotide probes, and the following removal with streptavidin-coated magnetic beads, is the main procedure to selectively deplete large rRNA molecules from total isolated RNA. Moreover, since rRNA—but not capped mRNAs—is characterized by the presence of 5′ phosphate, an useful approach for selective ribo-depletion is based on the use of an exonuclease able to specifically degrade RNA molecules bearing a 5′ phosphate (mRNA-ONLY kit, Epicentre). Compared to the polyadenylated (polyA+) mRNA fraction, the ribo-depleted RNA is enriched in non-polyA mRNA, preprocessed RNA, tRNA, regulatory molecules such as miRNA, siRNA, small ncRNA, and other RNA transcripts of yet unknown function (see review [8]).

How closely the RNA sequencing reflects the original RNA populations is mainly determined in the library preparation step, crucial in the whole transcriptome protocols. Although NGS protocols were first developed for the analysis of genomic DNA, these technical procedures have been rapidly and effectively adapted to the sequencing of double-strand (ds) cDNA for transcriptome studies [51].

A double-stranded cDNA library can be usually prepared by using: (1) fragmented double-stranded (ds) cDNA and (2) hydrolyzed or fragmented RNA.

The goal of the first approach is to generate high-quality, full-length cDNAs from RNA samples of interest to be fragmented and then ligated to an adapter for further amplification and sequencing. By the way, since the primer adaptor is ligated to a fragmented ds cDNA, any information on the transcriptional direction would completely be lost. Preserving the strandedness is fundamental for data analysis; it allows to determine the directionality of transcription and gene orientation and facilitates detection of opposing and overlapping transcripts. To take into account and thus to avoid this biologically relevant issue, many approaches, such as pretreating the RNA with sodium bisulphite to convert cytidine into uridine [52], have been so far developed. Other alternative protocols, differing in how the adaptors are inserted into ds cDNA, have been recently published: direct ligation of RNA adaptors to the RNA sample before or during reverse transcription [30, 31, 53], or incorporation of dUTP during second strand synthesis and digestion with uracil-Nglycosylase enzyme [45]. For instance, SOLiD Whole Transcriptome Kit contains two different sets of oligonucleotides with a single-stranded degenerate sequence at one end, and a defined sequence required for sequencing at the other end, constraining the orientation of RNA in the ligation reaction. The generation of ds cDNA from RNA involves a number of steps. First, RNA is converted into first-strand cDNA using reverse transcriptase with either random hexamers or oligo(dT) as primers. The resulting first-strand cDNA is then converted into double-stranded cDNA, further fragmented with DNAse I and then ligated to adapters for amplification and sequencing [54]. The advantage of using oligo dT is that the majority of cDNA produced should be polyadenylated mRNA, and hence more of the sequence obtained should be informative (nonribosomal). The significant disadvantage is that the reverse transcriptase enzyme will fall off of the template at a characteristic rate, resulting in a bias towards the 3′ end of transcripts. For long mRNAs this bias can be pronounced, resulting in an under representation (or worse in the absence) of the 5′ end of the transcript in the data. The use of random primers would therefore be the preferred method to avoid this problem and to allow a better representation of the 5′ end of long ORFs. However, when oligo dT primers are used for priming, the slope which is formed by the diminishing frequency of reads towards the 5′ end of the ORF can, in some cases, be useful for determining the strand of origin for new transcripts if strand information has not been retained [28, 37].

Fragmenting RNA, rather than DNA, has the clear advantage of reducing possible secondary structures, particularly for tRNA and miRNA, resulting in a major heterogeneity in coverage and can also lead to a more comprehensive transcriptome analysis (Figure 2). In this case, the RNA sample is first fragmented by using controlled temperature or chemical/enzymatic hydrolysis, ligated to adapters and retro-transcribed by complementary primers. Different protocols have been so far developed. Indeed, the adaptor sequences may be directly ligated to the previously fragmented RNA molecules by using T4 RNA ligase, and the resulting library can be reverse transcribed with primer pairs specifically suited on the adaptor sequences, and then sequenced. Another approach, recently described in [55], consists in the *in vitro* polyadenilation of RNA fragments in order to have a template for the next step of reverse transcription using poly(dT) primers containing both adaptor sequences (linkers), separated back-to-back by an endonuclease site. The resulting cDNAs are circularized and then cleaved at endonuclease site in the adaptors, thus leaving ss cDNA with the adaptors at both ends [55]. A third protocol described by [33], named double random priming method, uses biotinylated random primers (a sequencing primer P1 at the 5′ end, and a random octamer at the 3′ end). After a first random priming reaction, the products are isolated by using streptavidin beads and a second random priming reaction is performed on a solid phase with a random octamer carrying the sequencing primer P2. Afterwards, second random priming products are released from streptavidin beads by heat, PCR-amplified, gel-purified, and finally subjected to sequencing process from the P1 primer. Moreover, as already mentioned, in [45] the authors used dUTP—a surrogate for dTTP—during the second-strand synthesis to allow a selective degradation of second cDNA strand after adaptor ligation using a uracil-N-glycosylase. The use of engineered DNA adaptors, combined to the dUTP protocol, ensures that only the cDNA strand corresponding to the "real" transcript is used for library amplification and sequencing, reserving the strandedness of gene transcription [45].

However, independently on the library construction procedure, particular care should be taken to avoid complete degradation during RNA fragmentation.

The next step of the sequencing protocols is the clonally amplification of the cDNA fragments.

Illumina, 454 and SOLiD use clonally amplified templates. In particular, the last two platforms use an innovative procedure, emulsion PCR (emPCR), to prepare sequencing templates in a cell-free system. cDNA fragments from a fragment or paired-end library are separated into single strands and captured onto beads under conditions that favour one DNA molecule per bead. After the emPCR and beads enrichment, millions of them are chemically crosslinked to an amino-coated glass surface (SOLiD) or deposited into individual PicoTiterPlate (PTP) wells (454) in which the NGS chemistry can be performed. Solid-phase amplification (Illumina) can also be used to produce randomly distributed, clonally amplified clusters from fragment or mate-pair templates on a glass slide. High-density forward and reverse primers are covalently attached to the slide, and the ratio of the primers to the template defines the surface density. This procedure can produce up to 200 million spatially separated template clusters, providing ends for primer hybridization, needed to initiate the NGS reaction. A different approach is the use of single molecules templates (Helicos BioScience) usually immobilized on solid supports, in which PCR amplification is no more required, thus avoiding the insertion of possible confounding mutations in the templates. Furthermore, AT- and GC-rich sequences present amplification issues, with over- or under-representation bias

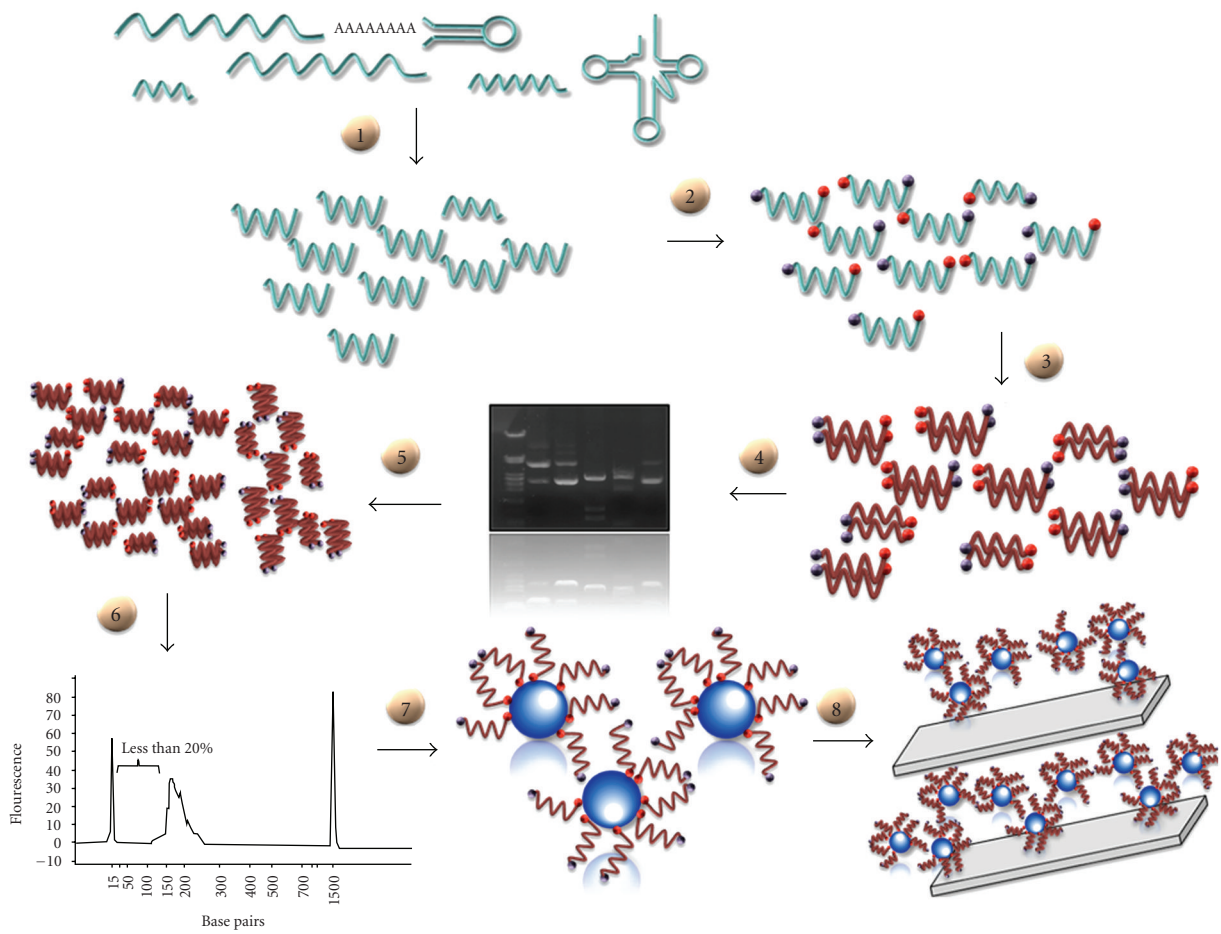

Figure 2: *Library preparation and clonal amplification.* Schematic representation of a workflow for library preparation in RNA-Seq experiments on the SOLiD platform. In the figure is depicted a total RNA sample after depletion of rRNA, containing both polyA and non-polyA mRNA, tRNAs, miRNAs and small noncoding RNAs. Ribo-depleted total RNA is fragmented (1), then ligated to specific adaptor sequences (2) and retro-transcribed (3). The resulting cDNA is size selected by gel electrophoresis (4), and cDNAs are PCR amplified (5). Then size distribution is evaluated (6). Emulsion PCR, with one cDNA fragment per bead, is used for the clonal amplification of cDNA libraries (7). Purified and enriched beads are finally deposited onto glass slides (8), ready to be sequenced by ligation.

in genome alignments and assemblies. Specific adaptors are bound to the fragmented templates, then hybridized to spatially distributed primers covalently attached to the solid support [56].

3.2. Sequencing and Imaging. NGS platforms use different sequencing chemistry and methodological procedures.

Illumina and HeliScope use the Cyclic Reversible Termination (CRT), which implies the use of reversible terminators (modified nucleotide) in a cyclic method. A DNA polymerase, bound to the primed template, adds one fluorescently modified nucleotide per cycle; then the remaining unincorporated nucleotides are washed away and imaging capture is performed. A cleavage step precedes the next incorporation cycle to remove the terminating/inhibiting group and the fluorescent dye, followed by an additional washing. Although these two platforms use the same methodology, Illumina employs the four-colour CRT method, simultaneously incorporating all 4 nucleotides with different dyes; HeliScope uses the one-colour (Cy5 dye) CRT method.

Substitutions are the most common error type, with a higher portion of errors occurring when the previous incorporated nucleotide is a G base [57]. Under representation of AT-rich and GC-rich regions, probably due to amplification bias during template preparation [57–59], is a common drawback.

In contrast, SOLiD system uses the Sequencing by Ligation (SBL) with 1, 2-nucleotide probes, based on colour space, which is an unique feature of SOLiD. It has the main advantage to improve accuracy in colour and single nucleotide variations (SNV) calling, the latter of which requires an adjacent valid colour change. In particular, a universal primer is hybridized to the template beads, and a library of 1, 2-nucleotide probes is added. Following four-colour imaging, the ligated probes are chemically cleaved to generate a 5′-phosphate group. Probe hybridization and ligation, imaging, and probe cleavage is repeated ten times to yield ten colour calls spaced in five-base intervals. The extended primer is then stripped from the solid-phase-bound templates. A second ligation round is performed with

a $n - 1$ primer, which resets the interrogation bases and the corresponding ten colour calls one position to the left. Ten ligation cycles ensue, followed by three rounds of ligation cycles. Colour calls from the five-ligation rounds are then ordered into a linear sequence (the csfasta colour space) and aligned to a reference genome to decode the sequence. The most common error type observed by using this platform are substitutions, and, similar to Illumina, SOLiD data have also revealed an under representation of AT- and GC-rich regions [58].

Another approach is pyrosequencing (on 454), a non-electrophoretic bioluminescence method, that unlike the above-mentioned sequencing approaches is able to measure the release of pyrophosphate by proportionally converting it into visible light after enzymatic reactions. Upon incorporation of the complementary dNTP, DNA polymerase extends the primer and pauses. DNA synthesis is reinitiated following the addition of the next complementary dNTP in the dispensing cycle. The enzymatic cascade generates a light recorded as a flowgram with a series of picks corresponding to a particular DNA sequence. Insertions and deletions are the most common error types.

An excellent and detailed review about the biotechnological aspects of NGS platforms can be found in [15].

3.3. From Biology to Bioinformatics. The unprecedented level of sensitivity in the data produced by NGS platforms brings with it the power to make many new biological observations, at the cost of a considerable effort in the development of new bioinformatics tools to deal with these massive data files.

First of all, the raw image files from one run of some next generation sequencers can require terabytes of storage, meaning that simply moving the data off the machine can represent a technical challenge for the computer networks of many research centers. Moreover, even when the data are transferred from the machine for subsequent processing, common desktop computer will be hopelessly outmatched by the volume of data from a single run. As a result, the use of a small cluster of computers is extremely beneficial to reduce computational bottleneck.

Another issue is the availability of software required to perform downstream analysis. Indeed after image and signal processing the output of a RNA-Seq experiment consists of 10–400 millions of short reads (together with their base-call quality values), typically of 30–400 bp, depending on the DNA sequencing technology used, its version and the total cost of the experiments.

NGS data analysis heavily relies on proper mapping of sequencing reads to corresponding reference genomes or on their efficient *de novo* assembly. Mapping NGS reads with high efficiency and reliability currently faces several challenges. As noticed by [60], differences between the sequencing platforms in samples preparation, chemistry, type and volume of raw data, and data formats are very large, implying that each platform produces data affected by characteristic error profiles. For example the 454 system can produce reads with insertion or deletion errors during homopolymer runs and generate fewer, but longer, sequences in fasta like format allowing to adapt classical alignment algorithms; the Illumina has an increased likelihood to accumulate sequence errors toward the end of the read and produce fasta reads, but they are shorter, hence requiring specific alignment algorithms; the SOLiD also tends to accumulate bias at the end of the reads, but uses di-base encoding strategy and each sequence output is encoded in a colour space csfasta format. Hence, some sequence errors are correctable, providing better discrimination between sequencing error and polymorphism, at the cost of requiring analysis tools explicitly built for handling this aspect of the data. It is not surprising that there are no "box standard" software available for end-users, hence the implementation of individualized data processing pipelines, combining third part packages and new computational methods, is the only advisable approach. While some existing packages are already enabling to solve general aspects of RNA-Seq analysis, they also require a time consuming effort due to the lack of clear documentation in most of the algorithms and the variety of the formats. Indeed, a much clear documentation of the algorithms is needed to ensure a full understanding of the processed data. Community adoption of input/output data formats for reference alignments, assemblies and detected variants is also essential for ease the data management problem. Solving these issues may simply shift the software gap from sequence processing (base-calling, alignment or assembly, positional counting and variant detection) to sequence analysis (annotation and functional impact).

3.4. Genome Alignment and Reads Assembly. The first step of any NGS data analysis consists of mapping the sequence reads to a reference genome (and/or to known annotated transcribed sequences) if available, or *de novo* assembling to produce a genome-scale transcriptional map. (see Figure 3 for an illustration of a classical RNA-Seq computational pipeline). The decision to use one of strategies is mainly based on the specific application. However, independently on the followed approach, there is a preliminary step that can be useful to perform which involves the application of a quality filtering to remove poor quality reads and to reduce the computational time and the effort for further analysis.

Analyzing the transcriptome of organisms without a specific reference genome requires *de novo* assembling (or a guided assembly with the help of closely related organisms) of expressed sequence tags (ESTs) using short-read assembly programs such as [61, 62]. A reasonable strategy for improving the quality of the assembly is to increase the read coverage and to mix different reads types. However RNA-Seq experiments without a reference genome propose specific features and challenges that are out of the scope of the present paper; we refer the readers to [63, 64] for further details.

In most cases, the reference genome is available and the mapping can be carried out using either the whole genome or known transcribed sequences (see, e.g., [28–30, 32, 34, 37, 40, 46, 47]). In both cases, this preliminary but crucial step is the most computationally intensive of the entire process and strongly depends on the type of available sequences (read-length, error profile, amount of data and data format). It is

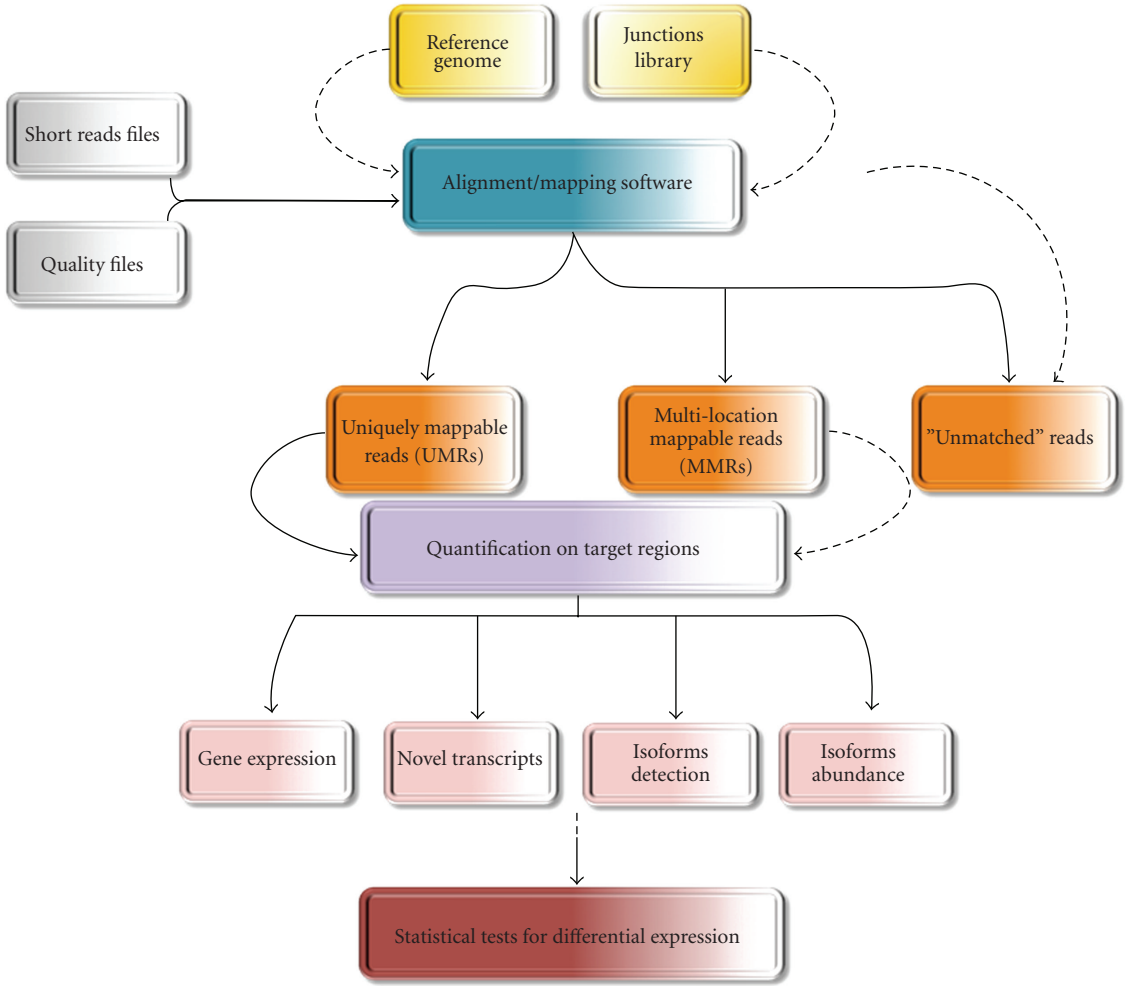

Figure 3: RNA-Seq computational pipeline.

not surprising that such nodal point still constitutes a very prominent area of research (see, e.g., [65–67] for a review) and has produced a great number of different algorithms in the last couple of years (e.g., [68–78]). Clearly, not all of them completely support the available platforms or are scalable for all amount of throughput or genome size. Nevertheless, the sequencing technologies are still in a developing phase with a very fast pace of increase in throughput, reads length and data formats after few months. Consequently, the already available mapping/assembly software are continuously under evolution in order to adapt themselves to the new data formats, to scale with the amount of data and to reduce their computational demand. New softwares are also continuously complementing the panorama. Moreover, the alignment phase of reads from RNA-Seq experiments presents many other subtleties to be considered; standard mapping algorithms are not able to fully exploit the complexity of the transcriptome, requiring to be modified or adapted in order to account for splicing events in eucaryotes.

The easiest way to handle such difficulty is to map the reads directly on known transcribed sequences, with the obvious drawback of missing new transcripts. Alternatively, the reads can be mapped continuously to the genome, but with the added opportunity of mapping reads that cross splice junctions. In this case, the algorithms differ from whether they require or not junctions's model. Algorithms such as Erange [37] or RNA-mate [79] require library of junctions constructed using known splice junctions extracted from data-bases and also supplemented with any set of putative splice junctions obtained, for instance, using a combinatorial approach on genes' model or ESTs sequences. Clearly, such approaches do not allow to map junctions not previously assembled in the junctions' library. On the other hand, algorithms like the WT [69], QPALMA [80], TopHat [81], G.Mo.R-Se [63], and PASS [78] potentially allow to detect new splice isoforms, since they use a more sophisticated mapping strategy. For instance, WT [69] splits the reads in left and right pieces, aligns each part to the genome, then attempts to extend each alignment on the other side to detect the junction. Whereas TopHat [81] first maps the reads against the whole reference genome using [77], second aggregates the mapped reads in islands of candidate exons on which compute a consensus measure, then generates potential donor/acceptor splice sites using

neighboring exons, and finally tries to align the reads, unmapped to the genome, to these splice junction sequences.

Most of the RNA-Seq packages are built on top of optimized short read *core* mappers [68, 69, 72, 77] and the mapping strategy is carried out by performing multiple runs or cycles. At the end of each cycle the unmatched reads are trimmed from one extreme and another step of alignment is attempted (see, e.g., [79]). Specific tolerances can be set for each alignment in order to increase the amount of mappable data. Obviously the simplest *core* approach is to map the sequence reads across the genome allowing the user to specify only the number of tolerated mismatches, although other methods allow to use also gapped alignment. Such flexibility can be beneficial for the rest of the analysis since both sequencing errors, that usually increase with the length of the sequence, and SNPs may cause substitutions and insertion/deletion of nucleotides in the reads. On the other hand, increasing the mapping flexibility also introduces a higher level of noise in the data. The compromise between the number of mapped reads and the quality of the resulting mapping is a very time consuming process without an optimal solution.

At the end of the mapping algorithm one can distinguish between three types of reads: reads that map uniquely to the genome or to the splice junctions (Uniquely Mappable Reads, UMR), reads with multiple (equally or similarly likely) locations either to the genome or to the splice junctions (Multilocation Mappable Reads, MMR) and reads without a specific mapping location. MMRs arise predominantly from conserved domains of paralogous gene families and from repeats. The fraction of mappable reads that are MMRs depends on the length of the read, the genome under investigation, and the expression in the individual sample; however it is typically between 10–40% for mammalian derived libraries [30, 37]. Most of the studies [28, 34] usually discarded MMRs from further analysis, limiting the attention only to UMRs. Clearly, this omission introduces experimental bias, decreases the coverage and reduces the possibility of investigating expressed regions such as active retrotransposons and gene families. An alternative strategy for the removal of the MMRs is to probabilistically assign them to each genomic location they map to. The simplest assignment considers equal probabilities. However, far better results have been obtained using a guilt-by-association strategy that calculates the probability of a MMRs originating from a particular locus. In [82], the authors proposed to proportionally assign MMRs to each of their mapping locations based on unique coincidences with either UMRs and other MMRs. Such a technique was later adopted in [79]. By contrast, in [83], the authors computed the probability as the ratio between the number of UMRs occurring in a nominal window surrounding each locus occupied by the considered MMR and the total number of UMRs proximal to all loci associated with that MMR. Similarly, in [37] the MMRs were fractionally assigned to their different possible locations considering the expression levels of their respective gene models. All these rescue strategies lead to substantially higher transcriptome coverage and give expression estimates in better agreement with microarrays than those using only UMRs (see, [37, 83]). Very recently, a more sophisticated approach was proposed in [84]. The authors introduced latent random variables representing the true mappings, with the parameters of the graphical model corresponding to isoform expression levels, read distributions across transcripts, and sequencing error. They allocated MMRs by maximizing the likelihood of the expression levels using an Expectation-Maximization (EM) algorithm. Additionally, they also showed that previous rescue methods introduced in [37, 82] are roughly equivalent to one iteration of EM. Independently on the specific proposal, we observe that all the above mentioned techniques work much better with data that preserve RNA strandedness. Alternatively, the use of paired-end protocols should help to alleviate the MMRs problem. Indeed, when one of the paired reads maps to a highly repetitive element in the genome but the second does not, it allows both reads to be unambiguously mapped to the reference genome. This is accomplished by first matching the first nonrepeat read uniquely to a genomic position and then looking within a size window, based on the known size range of the library fragments, for a match for the second read. The usefulness of this approach was demonstrated to improve read matching from 85% (single reads) to 93% (paired reads) [70], allowing a significant improvement in genome coverage, particularly in repeat regions. Currently, all of the next generation sequencing technologies are capable for generating data from paired-end reads, but unfortunately, till now only few RNA-Seq software support the use of paired-end reads in conjunction with the splice junctions mapping.

One of the possible reasons for reads not mapping to the genome and splice junctions is the presence of higher sequencing errors in the sequence. Other reasons can be identified in higher polymorphisms, insertion/deletion, complex exon-exon junctions, miRNA and small ncRNA: such situations could potentially be recovered by more sophisticated or combined alignment strategy.

Once mapping is completed, the user can display and explore the alignment on a genome browser (see Figure 4 for a screen-shot example) such as UCSC Genome Browser [85] (http://genome.ucsc.edu/) or the Integrative Genomics Viewer (IGV) (http://www.broadinstitute.org/igv), or on specifically devoted browsers such as EagleView [86], MapView [87] or Tablet [88], that can provide some highly informative views of the results at different levels of aggregations. Such tools allow to incorporate the obtained alignment with database annotations and other source of information, to observe specific polymorphism against sequence error, to identify well documented artifacts due to the DNA amplifications, as well as to detect other source of problems such as the not uniformity of the reads coverage across the transcript. Unfortunately, in many cases the direct visualization of the data is hampered by the lack of a common format for the alignment algorithm, causing a tremendous amount of extra work in format conversion for visualization purposes, feature extraction and other downstream analysis. Only recently, the SAM (Sequencing Alignment/Map) format [89] has been proposed as a possible standard for storing read alignment against reference sequences.

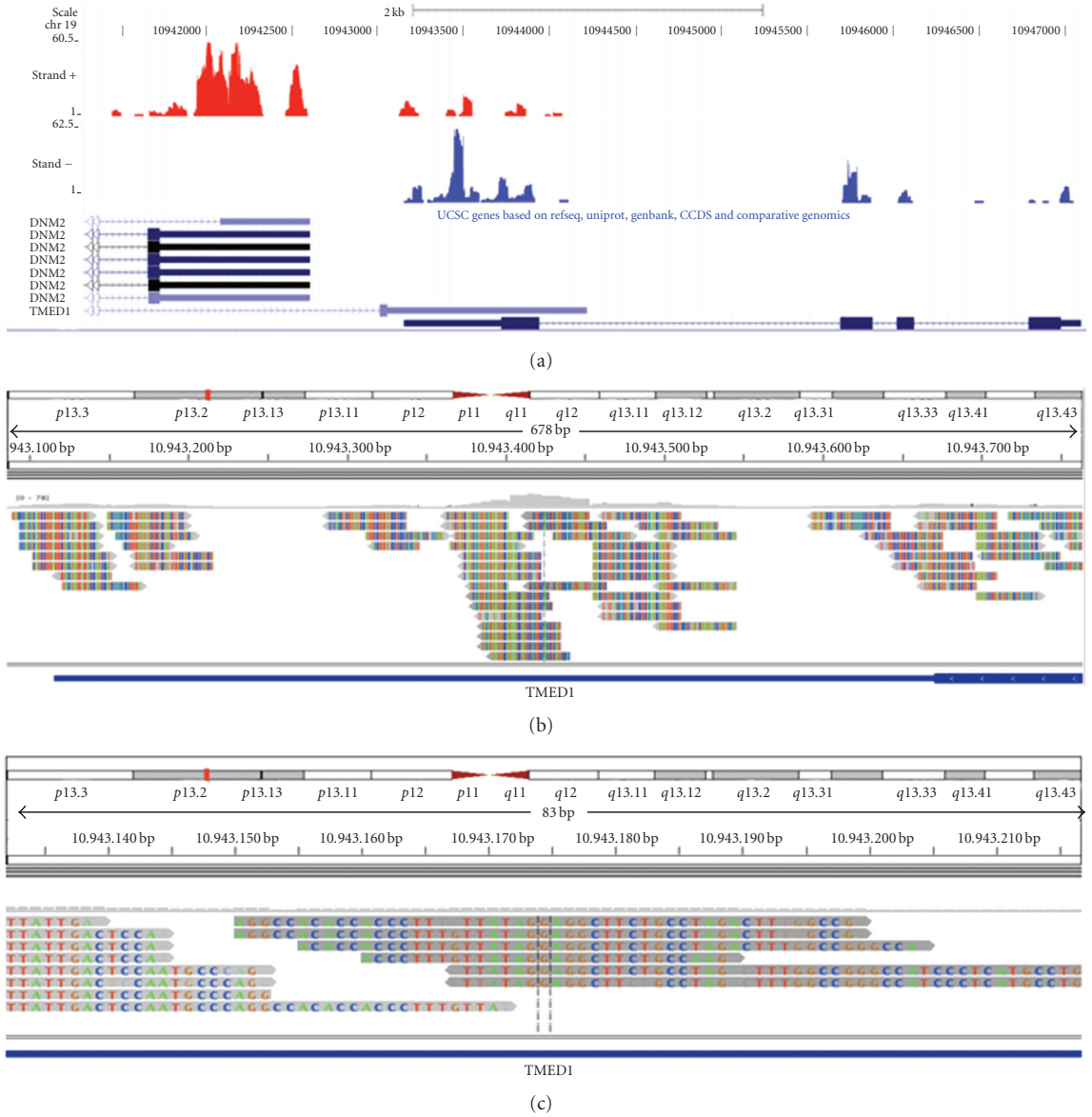

Figure 4: *Strand-Specific Read Distribution in UCSC Genome Browser and IGV.* (a) UCSC Genome Browser showing an example of stranded sequences generated by RNA-Seq experiment on NGS platform. In particular, the screenshot—of a characteristic "tail to tail" orientation of two human genes—clearly shows the specific expression in both strands where these two genes overlap, indicating that the strandedness of reads is preserved. (b) The same genomic location in the IGV browser, showing the reads (coloured blocks) distribution along TMED1 gene. The grey arrows indicate the sense of transcription. The specific expression in both strands where the genes overlap, indicates that the strandedness of reads is preserved. In (c) a greater magnification of the reads mapping to the same region at nucleotide level, useful to SNP analysis. The chromosome positions are shown at the top and genomic loci of the genes are shown at the bottom of each panel.

3.5. Quantifying Gene Expression and Isoforms' Abundance. Browser-driven analyses are very important for visualizing the quality of the data and to interpret specific events on the basis of the available annotations and mapped reads. However they only provide a qualitative picture of the phenomenon under investigation and the enormous amount of data does not allow to easily focus on the most relevant details. Hence, the second phase of most of the RNA-Seq pipeline consists of the automatic quantification of the transcriptional events across the entire genome (see Figure 4). From this point of view the interest is both quantifying known elements (i.e., genes or exons already annotated) and detecting new transcribed regions, defined as transcribed segments of DNA not yet annotated as exons in databases. The ability to detect these unannotated regions, even though biologically relevant, is one of the main advantages of the RNA-Seq over microarray technology. Usually, the quantification step is preliminary to any differential expression approach, see Figure 5.

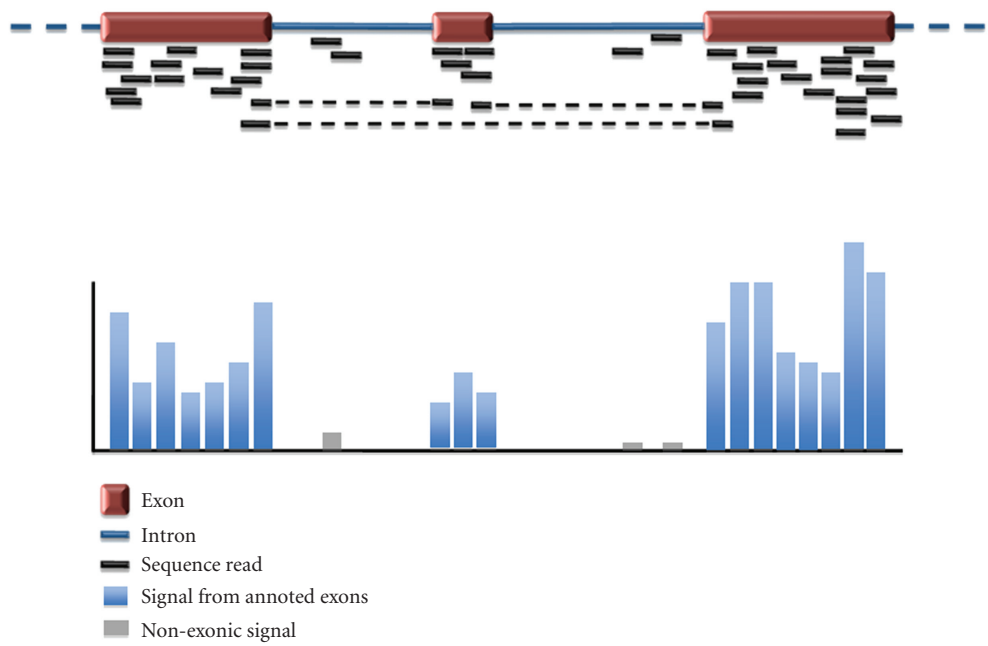

Figure 5: *Mapping and quantification of the signal.* RNA-seq experiments produce short reads sequenced from processed mRNAs. When a reference genome is available the reads can be mapped on it using efficient alignment software. Classical alignment tools will accurately map reads that fall within an exon, but they will fail to map spliced reads. To handle such problem suitable mappers, based either on junctions library or on more sophisticated approaches, need to be considered. After the mapping step annotated features can be quantified.

In order to derive a quantitative expression for annotated elements (such as exons or genes) within a genome, the simplest approach is to provide the expression as the total number of reads mapping to the coordinates of each annotated element. In the classical form, such method weights all the reads equally, even though they map the genome with different stringency. Alternatively, gene expression can be calculated as the sum of the number of reads covering each base position of the annotated element; in this way the expression is provided in terms of base coverage. In both cases, the results depend on the accuracy of the used gene models and the quantitative measures are a function of the number of mapped reads, the length of the region of interest and the molar concentration of the specific transcript. A straightforward solution to account for the sample size effect is to normalize the observed counts for the length of the element and the number of mapped reads. In [37], the authors proposed the *Reads Per Kilobase per Million* of mapped reads (RPKM) as a quantitative normalized measure for comparing both different genes within the same sample and differences of expression across biological conditions. In [84], the authors considered two alternative measures of relative expression: the fraction of transcripts and the fraction of nucleotides of the transcriptome made up by a given gene or isoform.

Although apparently easy to obtain, RPKM values can have several differences between software packages, hidden at first sight, due to the lack of a clear documentation of the analysis algorithms used. For example ERANGE [37] uses a union of known and new exon models to aggregate reads and determines a value for each region that includes spliced reads and assigned multireads too, whereas [30, 40, 81, 90] are restricted to known or prespecified exons/gene models. However, as noticed in [91], several experimental issues influence the RPKM quantification, including the integrity of the input RNA, the extent of ribosomal RNA remaining in the sample, the size selection steps and the accuracy of the gene models used.

In principle, RPKMs should reflect the true RNA concentration; this is true when samples have relatively uniform sequence coverage across the entire gene model. The problem is that all protocols currently fall short of providing the desired uniformity, see for example [37], where the Kolmogorov-Smirnov statistics is used to compare the observed reads distribution on each selected exon model with the theoretical uniform one. Similar conclusions are also illustrated in [57, 58], among others.

Additionally, it should be noted that RPKM measure should not be considered as the panacea for all RNA-Seq experiments. Despite the importance of the issue, the expression quantification did not receive the necessary attention from the community and in most of the cases the choice has been done regardless of the fact that the main question is the detection of differentially expressed elements. Regarding this point in [92] it is illustrated the inherent bias in transcript length that affect RNA-Seq experiments. In fact the total number of reads for a given transcript is roughly proportional to both the expression level and the length of the transcript. In other words, a long transcript will have more reads mapping to it compared to a short gene of similar expression. Since the power of an experiment is proportional to the sampling size, there will be more statistical power

to detect differential expression for longer genes. Therefore, short transcripts will always be at a statistical disadvantage relative to long transcripts in the same sample. RPKM-type measures provide an expression level normalized by the length of the gene and this only apparently solves the problem; it gives an unbiased measure of the expression level, but also changes the variance of the data in a length dependent manner, resulting in the same bias to differential expression estimation. In order to account for such an inherent bias, in [92] the authors proposed to use a fixed length window approach, with a window size smaller than the smallest gene. This method can calculate aggregated tag counts for each window and consequently assess them for differential expression. However, since the analysis is performed at the window level some proportion of the data will be discarded; moreover such an approach suffers for a reduced power and highly expressed genes are more likely to be detected due to the fact that the sample variance decreases with the expression level. Indeed, it should be noticed that the sample variance depends on both the transcript length and the expression level.

Finally, we observe that annotation files are often inaccurate; boundaries are not always mapped precisely, ambiguities and overlaps among transcripts often occur and are not yet completely solved. Concerning this issue in [93] the authors proposed a method based on the definition of "union-intersection genes" to define the genomic region of interest and normalized absolute and relative expression measures within. Also, in this case we observe that all strategies work much better with data that preserve RNA strandedness, which is an extremely valuable information for transcriptome annotation, especially for regions with overlapping transcription from opposite directions.

The quantification methods described above do not account for new transcribed region. Although several studies have already demonstrated that RNA-Seq experiments, with their high resolution and sensitivity have great potentiality in revealing many new transcribed regions, unidentifiable by microarrays, the detection of new transcribed regions is mainly obtained by means of a sliding window and heuristic approaches. In [94] stretches of contiguous expression in intergenic regions are identified after removing all UTRs from the intergenic search space by using a combination of information arising from tiling-chip and sequence data and visual inspection and manual curation. The procedure is quite complex and is mainly due to the lack of strandedness information in their experiment. On the contrary, the hybridization data are less affected by these issues because they distinguish transcriptional direction and do not show any 5′ bias (see [94] for further details). Then, new transcribed regions are required to have a length of at least 70 bp and an average sequence coverage of 5 reads per bp. A similar approach, with different choices of the threshold and the window, was proposed in [40], where the authors investigated either intergenic and intronic regions. The choices of the parameters are assessed by estimating noise levels by means of a Poisson model of the noncoding part of the genome. In [45] the whole genome is split into 50 bp windows (non-overlapping). A genomic region is defined as a new transcribed region if it results from the union of two consecutive windows, with at least two sequence reads mapped per window. Additionally, the gap between each new transcribed regions should be at least 50 bp, and the gap between a new transcribed region and an annotated gene (with the same strand) at least 100 bp. A slightly more sophisticated approach is used in ERANGE [37]. Reads that do not fall within known exons are aggregated into candidate exons by requiring regions with at least 15 reads, whose starts are not separated by more than 30 bp. Most of the candidate exons are assigned to neighboring gene models when they are within a specifiable distance of the model.

These studies, among others, reveal many of these new transcribed regions. Unfortunately, most of them do not seem to encode any protein, and hence their functions remain often to be determined. In any case, these new transcribed regions, combined with many undiscovered new splicing variants, suggest that there is considerably more transcript complexity than previously appreciated. Consequently further RNA-Seq experiments and more sophisticated analysis methods can disclose it.

The complexity of mammalian transcriptomes is also compounded by alternative splicing which allows one gene to produce multiple transcript isoforms. Alternative splicing includes events such as exon skipping, alternative 5′ or 3′ splicing, mutually exclusive exons, intron retention, and "cryptic" splice sites (see Figure 6). The frequency of occurrence of alternative splicing events is still underestimated. However it is well known that multiple transcript isoforms produced from a single gene can lead to protein isoforms with distinct functions, and that alternative splicing is widely involved in different physiological and pathological processes. One of the most important advantages of the RNA-Seq experiments is the possibility of understanding and comparing the transcriptome at the isoform level (see [95, 96]). In this context, two computational problems need to be solved: the detection of different isoforms and their quantification in terms of transcript abundance.

Initial proposals for solving these problems were essentially based on a gene-by-gene manual inspection usually focusing the attention to the detection of the presence of alternative splicing forms rather than to their quantification. For example, the knowledge of exon-exon junction reads and of junctions that fall into some isoform-specific regions can provide useful information for identifying different isoforms. The reliability of a splicing junction is usually assessed by counting features like the number of reads mapping to the junction, the number of mismatches on each mapped read, the mapping position on the junction and the mismatches location in a sort of heuristic approach. Unfortunately, these techniques cannot be scaled to the genome level and they are affected by a high false positive and false negative rate.

Following the above mentioned ideas, in [40] the authors detected junctions by computing the probability of a random hits for a read of length R on the splice junctions of length J with at most a certain number of mismatches. In [95], the authors used several information similar to those described above to train classifiers based on logistic regression for splicing junction detection. In [97], the authors introduced

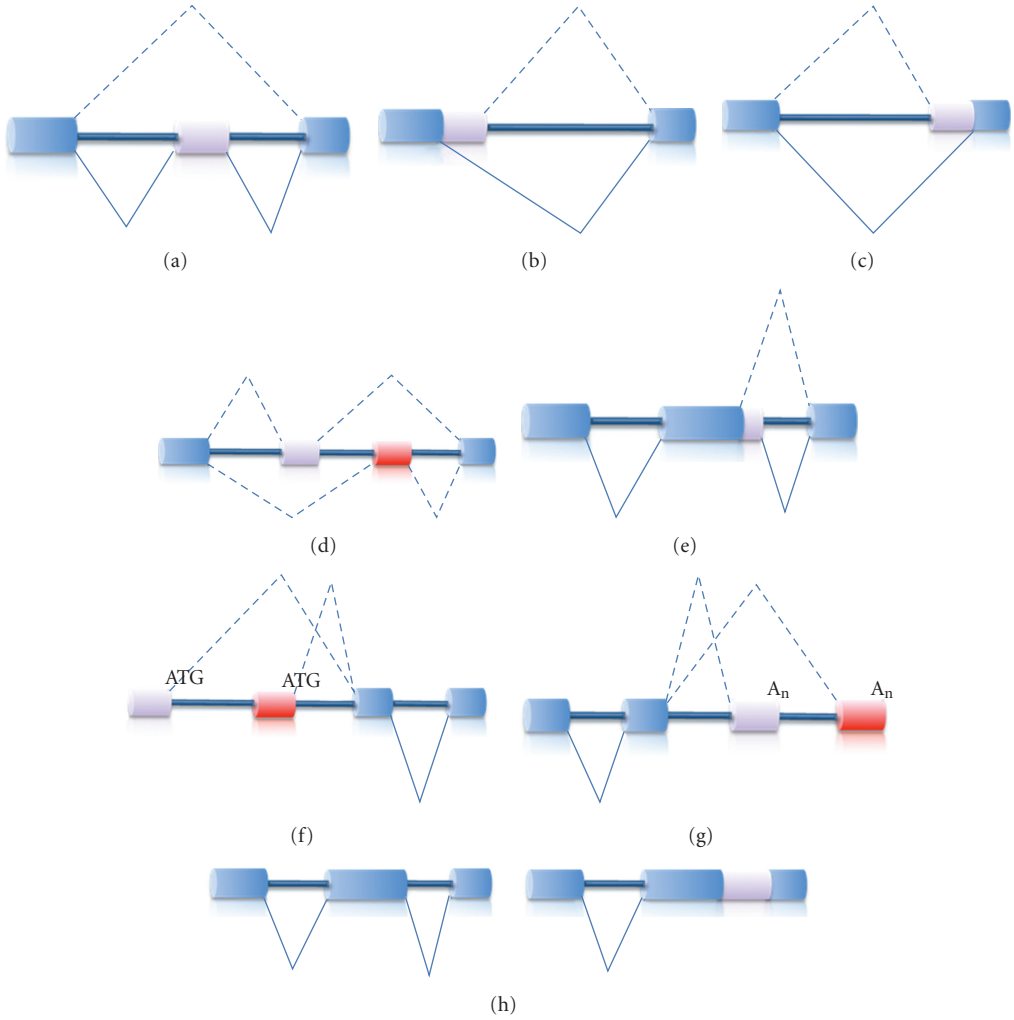

Figure 6: *Alternative splicing*. Schematic representation of the possible patterns of alternative splicing of a gene. Boxes are discrete exons that can be independently included or excluded from the mRNA transcript. Light blue boxes represent constitutive exons, violet and red boxes are alternatively spliced exons. Dashed lines represent alternative splicing events. (a) Canonical exon skipping; (b) 5′ or (c) 3′ alternative splicing; (d) Mutually exclusive splicing event involving the selection of only one from two or more exon variants; (e) Intra-exonic "cryptic" splice site causing the exclusion of a portion of the exon from the transcript; (f) Usage of new alternative 5′ or (g) 3′ exons; (h) Intron retention.

a new metric to measure the quality of each junction read. Then they estimated the distribution of such metric either with respect to known exon splice junctions and random splice junctions, and implemented an empirical statistical model to detect exon junctions evaluating the probability that an observed alignment distribution comes from a true junction.

The simple detection of specific isoforms does not provide useful information about their quantitative abundance. In principle, the quantification methods described above are equally applicable to quantify isoform expression. In practice, however, it is difficult to compute isoform-specific expression because most reads that are mapped to the genes are shared by more than one isoform and then it becomes difficult to assign each read only to a specific isoform. As a consequence, the assignment should rely on inferential methods that consider all data mapping to a certain region.

Several proposed methods for inferring isoforms' abundance are based on the preliminary knowledge of precise isoforms' annotation, on the assumption of uniform distribution of the reads across the transcript, on Poisson model for the reads' counts and equal weight for each read, regardless the quality of the match. The methods are often limited to handle only the cases where there is a relative small number of isoforms without confounding effects due to the overlap between genes. In particular in [98], the authors showed that the complexity of some isoform sets may still render the estimation problem nonidentifiable based on current RNA-Seq protocols and derived a mathematical characterization of identifiable isoform set. The main reason for such an effect is that current protocols with short single-end reads RNA-Seq are only able to asses local properties of a transcript. It is possible that the combination of short-read data with longer reads or paired-end reads will be able to go further in addressing such challenges.

Recently, in [90] the authors proposed a statistical method where, similar to [34], the count of reads falling into an annotated gene with multiple isoforms is modeled as a Poisson variable. They inferred the expression of each individual isoform using maximum likelihood approach, whose solution has been obtained by solving a convex optimization problem. In order to quantify the degree of uncertainty of the estimates, they carried out statistical inferences about the parameters from the posterior distribution by importance sampling. Interestingly, they showed that their method can be viewed as an extension of the RPKM concept and reduces to the RPKM index when there is only one isoform. An attempt to relax the assumption of uniform reads sampling is proposed in [84]. In this paper, the authors unified the notions of reads that map to multiple locations, that is, that could be potentially assigned to several genes, with those of reads that map to multiple isoforms through the introduction of latent random variables representing the true mappings. Then, they estimated the isoforms' abundance as the maximum likelihood expression levels using the EM algorithm. The Poisson distribution is also the main assumption in [99], where a comprehensive approach to the problem of alternative isoforms prediction is presented. In particular, the presence of alternative splicing event within the same sample is assessed by using Pearson's chi-square test on the parameter of a multinomial distribution and the EM algorithm is used to estimate the abundance of each isoform.

3.6. Differential Expression. The final goal in the majority of transcriptome studies is to quantify differences in expression across multiple samples in order to capture differential gene expression, to identify sample-specific alternative splicing isoforms and their differential abundance.

Mimicking the methods used for microarray analysis, researchers started to approach such crucial question using statistical hypothesis' tests combined with multiple comparisons error procedures on the observed counts (or on the RPKM values) at the gene, isoform or exon level. Indeed, in [30] the authors applied the empirical Bayes moderated t-test proposed in [100] to the normalized RPKM. However in microarray experiments, the abundance of a particular transcript is measured as a fluorescence intensity, that can be effectively modeled as a continuous response, whereas for RNA-Seq data the abundance is usually a count. Therefore, procedures that are successful for microarrays do not seem to be appropriate for dealing with such type of data.

One of the pioneering works to handle such difference is [34], where the authors modeled the aggregated reads count for each gene using Poisson distribution. One can prove that the number of reads observed from a gene (or transcript isoform) follows a binomial distribution that can be approximated by a Poisson distribution, under the assumption that RNA-Seq reads follow a random sampling process, in which each read is sampled independently and uniformly from every possible nucleotide in the sample. In this set-up, in [34] the authors used a likelihood ratio test to test for significant differences between the two conditions. The Poisson model was also employed by [40], where the authors used the method proposed in [101] to determine the significance of differential expression. On the contrary, in [83], the authors simply estimated the difference in expression of a gene between two conditions through the difference of the count proportions p_1 and p_2 computed using a classical Z-test statistics. In [18], the authors employed the Fishers exact test to better weigh the genes with relatively small counts. Similarly in [99] the authors used Poisson model and Fishers exact test to detect alternative exon usage between conditions.

Recently, more sophisticated approaches have been proposed in [102, 103]. In [102], the authors proposed an empirical Bayesian approach, based on the negative binomial distribution; it results very flexible and reduces to the Poisson model for a particular choice of the hyperparameter. They carried out differential expression testing using a moderated Bayes approach similar in the spirit to the one described in [100], but adapted for data that are counts. We observed that the method is designed for finding changes between two or more groups when at least one of the groups has replicated measurements. In [103], the observed counts of reads mapped to a specific gene obtained from a certain sample was modeled using Binomial distribution. Under such assumption, it can be proved that the log ratio between the two samples conditioned to the intensity signal (i.e., the average of the two logs counts) follows an approximate normal distribution, that is used for assessing the significance of the test. All the above-mentioned methods assume that the quantification of the features of interest under the experimental conditions has been already done and each read has been assigned to only one elements, hence the methods are directly applicable to detect genes or exons differences provided that overlapping elements are properly filtered out. By contrast the above described methods are not directly suited for detecting isoforms' differences unless the quantification of the isoform abundance has been carried out using specific approaches. To handle such difficulties, in [104], the authors proposed a hierarchical Bayesian model to directly infer the differential expression level of each transcript isoform in response to two conditions. The difference in expression of each isoform is modeled by means of an inverse gamma model and a latent variable is introduced for guiding the isoform's selection. The model can handle the heteroskedasticity of the sequence read coverage and inference is carried out using Gibbs sampler.

It should be noticed that although these techniques already provide interesting biological insights, they have not been sufficiently validated on several real data-sets where different type of replicates are available, neither sufficiently compared each others in terms of advantages and disadvantages. As with any new biotechnology it is important to carefully study the different sources of variation that can affect measure of the biological effects of interest and to statistically asses the reproducibility of the biological findings in a rigorous way, and to date this has been often omitted. Indeed, it should be considered that there are a variety of experimental effects that could possibly increase the variability, the bias, or be confounded with sequencing-based measures, causing miss-understanding of the results.

Unfortunately, such problems have received little of attention until now. In order to fill this gap, in [93] the authors presented a statistical inference framework for transcriptome analysis using RNA-Seq mapped read data. In particular, they proposed a new statistical method based on log-linear regression for investigating relationships between read counts and biological and experimental variables describing input samples as well as genomic regions of interest. The main advantage of the log-linear regression approach is that it allows to account both for biological effect and a variety of experimental effects. Their paper represents one of the few attempts of looking at the analysis of RNA-Seq data from a general point of view.

4. Challenges and Perspective for NGS

From the development of the Sanger method to the completion of the HGP, genetics has made significant advances towards the understanding of gene content and function. Even though significant achievements were reached by Human Genome, HapMap and ENCODE Projects [7, 105, 106], we are far from an exhaustive comprehension of the genomic diversity among humans and across the species, and from understanding gene expression variations and its regulation in both physio and pathological conditions. Since the appearance of first NGS platforms in the 2004, it was clear that understanding this diversity at a cost of around $5–10 million per genome sequence [107], placed it outside the real possibilities of most research laboratories, and very far from single individual economical potential. To date, we are in the "$1,000 genome" era, and, although this important barrier has not yet been broken, its a current assumption that this target is going to be reached within the end of 2010. It is likely that the rapid evolution of DNA sequencing technology, able to provide researchers with the ability to generate data about genetic variation and patterns of gene expression at an unprecedented scale, will become a routine tool for researchers and clinicians within just a few years.

As we can see, the number of applications and the great amount of biological questions that can be addressed by "Seq" experiments on NGS platforms is leading a revolution in the landscape of molecular biology, but the imbalance between the pace at which technology innovations are introduced in the platforms and the biological discoveries derivable from them is growing up. The risk is the creation of a glut of "under-used" information that in few months becomes of no use because the new one is produced. It is necessary to invest in an equivalent development of new computational strategies and expertise to deal with the volumes of data created by the current generation of new sequencing instruments, to maximize their potential benefit.

These platforms are creating a new world to explore, not only in the definition of experimental/technical procedures of large-scale analyses, but also in the downstream computational analysis and in the bioinformatics infrastructures support required for high-quality data generation and for their correct biological interpretation. In practice, they have shifted the bottleneck from the generation of experimental data to their management and to their statistical and computational analysis. There are few key points to consider. The first one is the data management: downstream computational analysis becomes difficult without appropriate Information Technology (IT) infrastructure. The terabytes of data produced by each sequencing run requires conspicuous storage and backup capacity, which increases considerably the experimental costs. The second one regards the protocols used for the production of raw data: each platform has its peculiarity in both sample preparation and type and volume of raw data produced, hence they require individualized laboratory expertise and data processing pipelines. Third, beside vendor specific and commercial software, several other open-source analysis tools are continuously appearing. Unfortunately, there is often an incomplete documentation and it is easy to spend more time in evaluating software suites than in analyzing the output data. Whichever software is used, the most important question is to understand its limitations and assumptions. Community adoption of input/output data standards is also essential to efficiently handle the data management problem. Till now the effort has been mainly devoted to the technological development rather than to the methodological counterpart. The choice of a careful experimental design has been also not always adequately considered.

As regards the RNA-Seq, we have still to face several critical issues either from a biological and computational point of view. RNA-seq protocols are extremely sensitive and need a very careful quality control for each wet laboratory step. For instance, the contamination of reagents with RNAse and the degradation of RNA, even partial, must be avoided during all the technical procedures. The quality of total isolated RNA is the first, and probably the most crucial point for an RNA-Seq experiment. Poor yield of polyA enrichment or low efficiency of total RNA ribodepletion are also critical issues for preparing high-quality RNA towards the library construction. It is clear that, independently on the library construction procedure, particular care should be taken to avoid complete degradation of RNA during the controlled RNA fragmentation step. Furthermore, in order to correctly determine the directionality of gene transcription and to facilitate the detection of opposing and overlapping transcripts within gene-dense genomic regions, particular care should be taken to preserve the strandedness of RNA fragments during the library preparation. In addition, to provide a more uniform coverage throughout the transcript length, random priming for reverse transcription protocols, rather than oligo dT priming (with the bias of low coverage at the 5′ ends), should be done after removal of rRNA. Finally, it should be considered that for the platforms based on CRT and SBL, substitutions and under representation of AT-rich and GC-rich regions, probably due to amplification bias during template preparation, are the most common error type. In contrast, for pyrosequencing platforms, insertions and deletions represent a common drawback.

For what concern the data analysis, to the above-mentioned points, we should note that most of the available software for read alignment are designed for genomic

mapping hence they are not fully capable to discover exon junctions. The classical extension for handling RNA-Seq data involves the preconstruction of junction libraries reducing the possibility of discovering new junctions. It would be desirable to develop new methods that allow either new junction detection and also the use of paired-end reads, that are particularly promising for more accurate study. Additionally further developments are required to assess the significance of new transcribed regions, the construction of new putative genes and the precise quantification of each isoform, for which there is still a lack of statistical methodologies. For what concerns the detection of differential expression, existing techniques were not sufficiently validated on biological data and compared in terms of specificity and sensitivity. Moreover, of potentially great impact, is the lack of biological replicates which precludes gauging the magnitude of individual effects in relation to technical effects. Biological replicates is essential in a RNA-Seq experiment to draw generalized conclusions about the "real" differences observed between two or more biological groups.

Facing such multidisciplinary challenges will be the key point for a fruitful transfer from laboratory studies to clinical applications. Indeed, the availability of low-cost, efficient and accurate technologies for gene expression and genome sequencing will be useful in providing pathological gene expression profiles in a wide number of common genetic disorders including type II diabetes, cardiovascular disease, Parkinson disease and Downs syndrome. Moreover, the application of NGS to the emerging disciplines of pharmacogenomics and nutrigenomics will allow to understand drug response and nutrient-gene interactions on the basis of individual patient's genetic make-up, leading in turn to the development of targeted therapies for many human diseases or tailored nutrient supplementation [108].

Acknowledgment

We are grateful to the anonymous referees whose valuable comments helped to substantially improve the paper. This work was supported by the CNR-Bioinformatics Project.

References

[1] D. D. Licatalosi and R. B. Darnell, "RNA processing and its regulation: global insights into biological networks," *Nature Reviews Genetics*, vol. 11, no. 1, pp. 75–87, 2010.

[2] V. E. Velculescu, L. Zhang, W. Zhou, et al., "Characterization of the yeast transcriptome," *Cell*, vol. 88, no. 2, pp. 243–251, 1997.

[3] J. Lindberg and J. Lundeberg, "The plasticity of the mammalian transcriptome," *Genomics*, vol. 95, no. 1, pp. 1–6, 2010.

[4] W. F. Doolittle and C. Sapienza, "Selfish genes, the phenotype paradigm and genome evolution," *Nature*, vol. 284, no. 5757, pp. 601–603, 1980.

[5] R. J. Taft, M. Pheasant, and J. S. Mattick, "The relationship between non-protein-coding DNA and eukaryotic complexity," *BioEssays*, vol. 29, no. 3, pp. 288–299, 2007.

[6] T. Cavalier-Smith, "Cell volume and the evolution of eukaryote genome size," in *The Evolution of Genome Size*, T. Cavalier-Smith, Ed., pp. 105–184, John Wiley & Sons, Chichester, UK, 1985.

[7] E. Birney, J. A. Stamatoyannopoulos, A. Dutta, et al., "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project," *Nature*, vol. 447, no. 7146, pp. 799–816, 2007.

[8] A. Jacquier, "The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs," *Nature Reviews Genetics*, vol. 10, no. 12, pp. 833–844, 2009.

[9] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[10] R. W. Holley, "Alanine transfer RNA," in *Nobel Lectures in Molecular Biology 1933–1975*, pp. 285–300, Elsevier North Holland, New York, NY, USA, 1977.

[11] A. M. Maxam and W. Gilbert, "A new method for sequencing DNA," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, no. 2, pp. 560–564, 1977.

[12] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, no. 12, pp. 5463–5467, 1977.

[13] E. R. Mardis, "Next-generation DNA sequencing methods," *Annual Review of Genomics and Human Genetics*, vol. 9, pp. 387–402, 2008.

[14] J. Shendure and H. Ji, "Next-generation DNA sequencing," *Nature Biotechnology*, vol. 26, no. 10, pp. 1135–1145, 2008.

[15] M. L. Metzker, "Sequencing technologies the next generation," *Nature Reviews Genetics*, vol. 11, no. 1, pp. 31–46, 2010.

[16] R. A. Irizarry, D. Warren, F. Spencer, et al., "Multiple-laboratory comparison of microarray platforms," *Nature Methods*, vol. 2, no. 5, pp. 345–349, 2005.

[17] P. A. C. 't Hoen, Y. Ariyurek, H. H. Thygesen, et al., "Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms," *Nucleic Acids Research*, vol. 36, no. 21, article e141, 2008.

[18] J. S. Bloom, Z. Khan, L. Kruglyak, M. Singh, and A. A. Caudy, "Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays," *BMC Genomics*, vol. 10, article 221, 2009.

[19] M. Harbers and P. Carninci, "Tag-based approaches for transcriptome research and genome annotation," *Nature Methods*, vol. 2, no. 7, pp. 495–502, 2005.

[20] M. P. Horan, "Application of serial analysis of gene expression to the study of human genetic disease," *Human Genetics*, vol. 126, no. 5, pp. 605–614, 2009.

[21] H. Misu, T. Takamura, N. Matsuzawa, et al., "Genes involved in oxidative phosphorylation are coordinately upregulated with fasting hyperglycaemia in livers of patients with type 2 diabetes," *Diabetologia*, vol. 50, no. 2, pp. 268–277, 2007.

[22] T. Takamura, H. Misu, T. Yamashita, and S. Kaneko, "SAGE application in the study of diabetes," *Current Pharmaceutical Biotechnology*, vol. 9, no. 5, pp. 392–399, 2008.

[23] D. V. Gnatenko, J. J. Dunn, S. R. McCorkle, D. Weissmann, P. L. Perrotta, and W. F. Bahou, "Transcript profiling of human platelets using microarray and serial analysis of gene expression," *Blood*, vol. 101, no. 6, pp. 2285–2293, 2003.

[24] C. A. Sommer, E. C. Pavarino-Bertelli, E. M. Goloni-Bertollo, and F. Henrique-Silva, "Identification of dysregulated genes in lymphocytes from children with Down syndrome," *Genome*, vol. 51, no. 1, pp. 19–29, 2008.

[25] W. Malagó Jr., C. A. Sommer, C. Del Cistia Andrade, et al., "Gene expression profile of human Down syndrome leukocytes," *Croatian Medical Journal*, vol. 46, no. 4, pp. 647–656, 2005.

[26] B. T. Wilhelm and J.-R. Landry, "RNA-Seq-quantitative measurement of expression through massively parallel RNA-Sequencing," *Methods*, vol. 48, no. 3, pp. 249–257, 2009.

[27] M. N. Bainbridge, R. L. Warren, M. Hirst, et al., "Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach," *BMC Genomics*, vol. 7, article 246, 2006.

[28] U. Nagalakshmi, Z. Wang, K. Waern, et al., "The transcriptional landscape of the yeast genome defined by RNA sequencing," *Science*, vol. 320, no. 5881, pp. 1344–1349, 2008.

[29] T. T. Torres, M. Metta, B. Ottenwälder, and C. Schlötterer, "Gene expression profiling by massively parallel sequencing," *Genome Research*, vol. 18, no. 1, pp. 172–177, 2008.

[30] N. Cloonan, A. R. R. Forrest, G. Kolle, et al., "Stem cell transcriptome profiling via massive-scale mRNA sequencing," *Nature Methods*, vol. 5, no. 7, pp. 613–619, 2008.

[31] L. J. Core, J. J. Waterfall, and J. T. Lis, "Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters," *Science*, vol. 322, no. 5909, pp. 1845–1848, 2008.

[32] S.-I. Hashimoto, W. Qu, B. Ahsan, et al., "High-resolution analysis of the 5′-end transcriptome using a next generation DNA sequencer," *PLoS ONE*, vol. 4, no. 1, article e4108, 2009.

[33] H. Li, M. T. Lovci, Y.-S. Kwon, M. G. Rosenfeld, X.-D. Fu, and G. W. Yeo, "Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 51, pp. 20179–20184, 2008.

[34] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, "RNA-Seq: an assessment of technical reproducibility and comparison with gene expression arrays," *Genome Research*, vol. 18, no. 9, pp. 1509–1517, 2008.

[35] R. D. Morin, M. D. O'Connor, M. Griffith, et al., "Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells," *Genome Research*, vol. 18, no. 4, pp. 610–621, 2008.

[36] R. D. Morin, M. Bainbridge, A. Fejes, et al., "Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing," *BioTechniques*, vol. 45, no. 1, pp. 81–94, 2008.

[37] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature Methods*, vol. 5, no. 7, pp. 621–628, 2008.

[38] R. Rosenkranz, T. Borodina, H. Lehrach, and H. Himmelbauer, "Characterizing the mouse ES cell transcriptome with Illumina sequencing," *Genomics*, vol. 92, no. 4, pp. 187–194, 2008.

[39] D. J. Sugarbaker, W. G. Richards, G. J. Gordon, et al., "Transcriptome sequencing of malignant pleural mesothelioma tumors," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 9, pp. 3521–3526, 2008.

[40] M. Sultan, M. H. Schulz, H. Richard, et al., "A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome," *Science*, vol. 321, no. 5891, pp. 956–960, 2008.

[41] Y. W. Asmann, E. W. Klee, E. A. Thompson, et al., "3′ tag digital gene expression profiling of human brain and universal reference RNA using Illumina Genome Analyzer," *BMC Genomics*, vol. 10, article 531, 2009.

[42] I. Chepelev, G. Wei, Q. Tang, and K. Zhao, "Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq," *Nucleic Acids Research*, vol. 37, no. 16, article e106, 2009.

[43] J. Z. Levin, M. F. Berger, X. Adiconis, et al., "Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts," *Genome Biology*, vol. 10, no. 10, article R115, 2009.

[44] C. A. Maher, N. Palanisamy, J. C. Brenner, et al., "Chimeric transcript discovery by paired-end transcriptome sequencing," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 30, pp. 12353–12358, 2009.

[45] D. Parkhomchuk, T. Borodina, V. Amstislavskiy, et al., "Transcriptome analysis by strand-specific sequencing of complementary DNA," *Nucleic Acids Research*, vol. 37, no. 18, article e123, 2009.

[46] T. E. Reddy, F. Pauli, R. O. Sprouse, et al., "Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation," *Genome Research*, vol. 19, no. 12, pp. 2163–2171, 2009.

[47] F. Tang, C. Barbacioru, Y. Wang, et al., "mRNA-Seq whole-transcriptome analysis of a single cell," *Nature Methods*, vol. 6, no. 5, pp. 377–382, 2009.

[48] R. Blekhman, J. C. Marioni, P. Zumbo, M. Stephens, and Y. Gilad, "Sex-specific and lineage-specific alternative splicing in primates," *Genome Research*, vol. 20, no. 2, pp. 180–189, 2010.

[49] G. A. Heap, J. H. M. Yang, K. Downes, et al., "Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing," *Human Molecular Genetics*, vol. 19, no. 1, pp. 122–134, 2010.

[50] D. Raha, Z. Wang, Z. Moqtaderi, et al., "Close association of RNA polymerase II and many transcription factors with Pol III genes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 8, pp. 3639–3644, 2010.

[51] S. Marguerat and J. Bahler, "RNA-Seq: from technology to biology," *Cellular and Molecular Life Sciences*, vol. 67, no. 4, pp. 569–579, 2010.

[52] Y. He, B. Vogelstein, V. E. Velculescu, N. Papadopoulos, and K. W. Kinzler, "The antisense transcriptomes of human cells," *Science*, vol. 322, no. 5909, pp. 1855–1857, 2008.

[53] R. Lister, R. C. O'Malley, J. Tonti-Filippini, et al., "Highly integrated single-base resolution maps of the epigenome in Arabidopsis," *Cell*, vol. 133, no. 3, pp. 523–536, 2008.

[54] B. T. Wilhelm, S. Marguerat, I. Goodhead, and J. Bahler, "Defining transcribed regions using RNA-Seq," *Nature Protocols*, vol. 5, no. 2, pp. 255–266, 2010.

[55] N. T. Ingolia, S. Ghaemmaghami, J. R. S. Newman, and J. S. Weissman, "Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling," *Science*, vol. 324, no. 5924, pp. 218–223, 2009.

[56] T. D. Harris, P. R. Buzby, H. Babcock, et al., "Single-molecule DNA sequencing of a viral genome," *Science*, vol. 320, no. 5872, pp. 106–109, 2008.

[57] J. C. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer, "Substantial biases in ultra-short read data sets from high-throughput DNA sequencing," *Nucleic Acids Research*, vol. 36, no. 16, article e105, 2008.

[58] O. Harismendy, P. C. Ng, R. L. Strausberg, et al., "Evaluation of next generation sequencing platforms for population targeted sequencing studies," *Genome Biology*, vol. 10, no. 3, article R32, 2009.

[59] L. W. Hillier, G. T. Marth, A. R. Quinlan, et al., "Whole-genome sequencing and variant discovery in *C. elegans*," *Nature Methods*, vol. 5, no. 2, pp. 183–188, 2008.

[60] J. D. McPherson, "Next-generation gap," *Nature Methods*, vol. 6, no. 11S, pp. S2–S5, 2009.

[61] D. R. Zerbino and E. Birney, "Velvet: algorithms for de novo short read assembly using de Bruijn graphs," *Genome Research*, vol. 18, no. 5, pp. 821–829, 2008.

[62] I. Birol, S. D. Jackman, C. B. Nielsen, et al., "De novo transcriptome assembly with ABySS," *Bioinformatics*, vol. 25, no. 21, pp. 2872–2877, 2009.

[63] F. Denoeud, J.-M. Aury, C. Da Silva, et al., "Annotating genomes with massive-scale RNA sequencing," *Genome Biology*, vol. 9, no. 12, article R175, 2008.

[64] M. Yassoura, T. Kaplana, H. B. Fraser, et al., "Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 9, pp. 3264–3269, 2009.

[65] C. Trapnell and S. L. Salzberg, "How to map billions of short reads onto genomes," *Nature Biotechnology*, vol. 27, no. 5, pp. 455–457, 2009.

[66] P. Flicek and E. Birney, "Sense from sequence reads: methods for alignment and assembly," *Nature Methods*, vol. 6, supplement 11, pp. S6–S12, 2009.

[67] D. S. Horner, G. Pavesi, T. Castrignanò, et al., "Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing," *Briefings in Bioinformatics*, vol. 11, no. 2, pp. 181–197, 2009.

[68] A. Cox, "ELAND: efficient local alignment of nucleotide data," unpublished, http://bioit.dbi.udel.edu/howto/eland.

[69] "Applied Biosystems mappread and whole transcriptome software tools," http://www.solidsoftwaretools.com/.

[70] H. Li, J. Ruan, and R. Durbin, "Mapping short DNA sequencing reads and calling variants using mapping quality scores," *Genome Research*, vol. 18, no. 11, pp. 1851–1858, 2008.

[71] A. D. Smith, Z. Xuan, and M. Q. Zhang, "Using quality scores and longer reads improves accuracy of Solexa read mapping," *BMC Bioinformatics*, vol. 9, article 128, 2008.

[72] R. Li, Y. Li, K. Kristiansen, and J. Wang, "SOAP: short oligonucleotide alignment program," *Bioinformatics*, vol. 24, no. 5, pp. 713–714, 2008.

[73] R. Li, C. Yu, Y. Li, et al., "SOAP2: an improved ultrafast tool for short read alignment," *Bioinformatics*, vol. 25, no. 15, pp. 1966–1967, 2009.

[74] B. D. Ondov, A. Varadarajan, K. D. Passalacqua, and N. H. Bergman, "Efficient mapping of Applied Biosystems SOLiD sequence data to a reference genome for functional genomic applications," *Bioinformatics*, vol. 24, no. 23, pp. 2776–2777, 2008.

[75] H. Jiang and W. H. Wong, "SeqMap: mapping massive amount of oligonucleotides to the genome," *Bioinformatics*, vol. 24, no. 20, pp. 2395–2396, 2008.

[76] H. Lin, Z. Zhang, M. Q. Zhang, B. Ma, and M. Li, "ZOOM! Zillions of oligos mapped," *Bioinformatics*, vol. 24, no. 21, pp. 2431–2437, 2008.

[77] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biology*, vol. 10, no. 3, article R25, 2009.

[78] D. Campagna, A. Albiero, A. Bilardi, et al., "PASS: a program to align short sequences," *Bioinformatics*, vol. 25, no. 7, pp. 967–968, 2009.

[79] N. Cloonan, Q. Xu, G. J. Faulkner, et al., "RNA-MATE: a recursive mapping strategy for high-throughput RNA-sequencing data," *Bioinformatics*, vol. 25, no. 19, pp. 2615–2616, 2009.

[80] F. De Bona, S. Ossowski, K. Schneeberger, and G. Rätsch, "Optimal spliced alignments of short sequence reads," *Bioinformatics*, vol. 24, no. 16, pp. i174–i180, 2008.

[81] C. Trapnell, L. Pachter, and S. L. Salzberg, "TopHat: discovering splice junctions with RNA-Seq," *Bioinformatics*, vol. 25, no. 9, pp. 1105–1111, 2009.

[82] G. J. Faulkner, A. R. R. Forrest, A. M. Chalk, et al., "A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE," *Genomics*, vol. 91, no. 3, pp. 281–288, 2008.

[83] T. Hashimoto, M. J. L. de Hoon, S. M. Grimmond, C. O. Daub, Y. Hayashizaki, and G. J. Faulkner, "Probabilistic resolution of multi-mapping reads in massively parallel sequencing data using MuMRescueLite," *Bioinformatics*, vol. 25, no. 19, pp. 2613–2614, 2009.

[84] B. Li, V. Ruotti, R. M. Stewart, J. A. Thomson, and C. N. Dewey, "RNA-Seq gene expression estimation with read mapping uncertainty," *Bioinformatics*, vol. 26, no. 4, pp. 493–500, 2009.

[85] W. J. Kent, C. W. Sugnet, T. S. Furey, et al., "The human genome browser at UCSC," *Genome Research*, vol. 12, no. 6, pp. 996–1006, 2002.

[86] W. Huang and G. Marth, "EagleView: a genome assembly viewer for next-generation sequencing technologies," *Genome Research*, vol. 18, no. 9, pp. 1538–1543, 2008.

[87] H. Bao, H. Guo, J. Wang, R. Zhou, X. Lu, and S. Shi, "MapView: visualization of short reads alignment on a desktop computer," *Bioinformatics*, vol. 25, no. 12, pp. 1554–1555, 2009.

[88] I. Milne, M. Bayer, L. Cardle, et al., "Tablet-next generation sequence assembly visualization," *Bioinformatics*, vol. 26, no. 3, pp. 401–402, 2010.

[89] H. Li, B. Handsaker, A. Wysoker, et al., "The sequence alignment/map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[90] H. Jiang and W. H. Wong, "Statistical inferences for isoform expression in RNA-Seq," *Bioinformatics*, vol. 25, no. 8, pp. 1026–1032, 2009.

[91] S. Pepke, B. Wold, and A. Mortazavi, "Computation for ChIP-Seq and RNA-Seq studies," *Nature Methods*, vol. 6, no. 11S, pp. S22–S32, 2009.

[92] A. Oshlack and M. J. Wakefield, "Transcript length bias in RNA-Seq data confounds systems biology," *Biology Direct*, vol. 4, article 14, 2009.

[93] J. H. Bullard, E. A. Purdom, K. D. Hansen, S. Durinck, and S. Dudoit, "Statistical inference in mRNA-Seq: exploratory data analysis and differential expression," Tech. Rep. 247/2009, University of California, Berkeley, 2009.
[94] B. T. Wilhelm, S. Marguerat, S. Watt, et al., "Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution," *Nature*, vol. 453, no. 7199, pp. 1239–1243, 2008.
[95] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe, "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing," *Nature Genetics*, vol. 40, no. 12, pp. 1413–1415, 2008.
[96] E. T. Wang, R. Sandberg, S. Luo, et al., "Alternative isoform regulation in human tissue transcriptomes," *Nature*, vol. 456, no. 7221, pp. 470–476, 2008.
[97] L. Wang, Y. Xi, J. Yu, L. Dong, L. Yen, and W. Li, "A statistical method for the detection of alternative splicing using RNA-Seq," *PLoS ONE*, vol. 5, no. 1, article e8529, 2010.
[98] D. Hiller, H. Jiang, W. Xu, and W. H. Wong, "Identifiability of isoform deconvolution from junction arrays and RNA-Seq," *Bioinformatics*, vol. 25, no. 23, pp. 3056–3059, 2009.
[99] H. Richard, M. H. Schulz, M. Sultan, et al., "Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments," *Nucleic Acids Research*, vol. 38, no. 10, p. e112, 2010.
[100] G. K. Smyth, "Linear models and empirical Bayes methods for assessing differential expression in microarray experiments," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, article 3, 2004.
[101] S. Audic and J.-M. Claverie, "The significance of digital gene expression profiles," *Genome Research*, vol. 7, no. 10, pp. 986–995, 1997.
[102] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.
[103] L. Wang, Z. Feng, X. Wang, X. Wang, and X. Zhang, "DEGseq: an R package for identifying differentially expressed genes from RNA-Seq data," *Bioinformatics*, vol. 26, no. 1, pp. 136–138, 2009.
[104] S. Zheng and L. Chen, "A hierarchical Bayesian model for comparing transcriptomes at the individual transcript isoform level," *Nucleic Acids Research*, vol. 37, no. 10, article e75, 2009.
[105] F. S. Collins, E. S. Lander, J. Rogers, and R. H. Waterson, "Finishing the euchromatic sequence of the human genome," *Nature*, vol. 431, no. 7011, pp. 931–945, 2004.
[106] International Human Genome Sequencing Consortium, "A haplotype map of the human genome," *Nature*, vol. 437, no. 7063, pp. 1299–1320, 2005.
[107] E. R. Mardis, "Anticipating the 1,000 dollar genome," *Genome Biology*, vol. 7, no. 7, article 112, 2006.
[108] V. Costa, A. Casamassimi, and A. Ciccodicola, "Nutritional genomics era: opportunities toward a genome-tailored nutritional regimen," *The Journal of Nutritional Biochemistry*, vol. 21, no. 6, pp. 457–467, 2010.