TABLE 4.4
Software for detecting imprinting and maternal effects using SNP and binary trait data

Software	Data Type	Trait	Ref
Mendel	General pedigrees	Binary	[55]
MC-PDT	General pedigrees	Binary	[520, 542]
LIME	CaPT/CoPT + CaMP/CoMP	Binary	[505]
MCPDTI	General pedigrees	Quantitative	[539]

4.8 Concluding remarks

In most of the methods discussed in this chapter, HWE is not assumed, although some may assume mating symmetry, a condition less stringent than HWE. Such assumptions are typically in place to reduce the number of parameters that need to be estimated. Hidden in some of the methods are rare disease assumptions, which have driven other assumptions about population frequency relationships, hidden factors that may contribute to estimation bias and inflated type I errors [502].

This chapter discusses methods for detecting imprinting and/or maternal genotype effect of a single SNP marker. To increase detection power, multiple SNPs can be analyzed jointly through consideration of haplotype effects [538]. On the other hand, haplotypes may be used to help infer missing genotypes or resolve ambiguity of parental origin to increase detection power with single SNP testing methods [49, 234].

We focus on binary traits in this chapter, but quantitative traits are also of great interest and limited methods are available [165, 398, 452]. Environmental factors and other covariates may also play important roles in causal mechanism of a trait. Log-linear and logistic model can accommodate for such effects [522]. For non-parametric approaches, covariates may be accounted for by adjusting quantitative trait values [165]. How to best incorporate covariates with simple non-parametric tests for binary traits remains an open problem. Information from genetic markers tightly linked to the test locus may be utilized to facilitate detectability of effects or to increase detection power [521, 522].

5

Modeling and Analysis of Next-Generation Sequencing Data

CONTENTS

For more than two decades, high-throughput genomic technologies, starting with the microarrays, have made it possible to simultaneously analyze tens of thousands of genes in an organism's genome. In the last few years, the advent of next-generation sequencing (or deep sequencing) has opened a whole new avenue in high-throughput genomics by increasing the coverage, the resolution and the statistical power of such analyses. The RNA-seq technology offers unprecedented information about the transcriptome, but harnessing this information and extracting knowledge from it using bioinformatics tools remain fraught with challenges and present a bottleneck. A great deal of statistical research is now being devoted to this new, interdisciplinary area, resulting in novel methods to extract signals from noisy data and compare signals across multiple experimental conditions. This chapter is intended to provide a brief yet informative overview of the next-generation sequencing technology, the nature of the data it produces, the quantitative issues involved and the statistical challenges/methodologies (including designs and inferential techniques).

5.1 Isolation, quality control and library preparation

For mRNA molecules, which are the primary target of RNA-seq, stability is a major issue. DNA molecules are inherently stable, but mRNA is not. In general, prokaryotic mRNA molecules are unstable and quickly degraded by enzymes called endoribonucleases and exoribonucleases after transcription. Consequently, most of them are short-lived and the average prokaryotic mRNA half-life is less than 10 minutes. This high turnover rate enables prokaryotic cells to respond to environmental changes promptly by making quick changes to the transcription (i.e. mRNA production) process. Eukaryotic mRNA molecules, however, are in general more stable with a half-life of 7–10 hours. The half-lives of mRNA are subject to regulation, based on the developmental stage of the organism or environmental factors. The regulation of eukaryotic mRNA degradation has been known to involve interactions between some sequence elements on the mRNA molecule and proteins or micro-RNAs. Most eukaryotic mRNA decay starts with de-adenylation at the $3'$ end, which is the removal of the poly-adenine tail by the enzyme deadenylase. Then the decay continues either through de-capping of mRNA at the $5'$ end and subsequent degradation by $5'-3'$ exoribonuclease or through direct $3'-5'$ decay from the tail end by a multi-protein complex called exosome.

RNA is usually isolated from freshly dissected or frozen tissue-samples using commercially available and user-friendly kits such as TRIZOL (Life Technologies, Carlsbad, California), RiboPure (Ambion, Austin, Texas) or

RNAEasy (Qiagen, Hilden, Germany). On the other hand, high-throughput RNA isolation systems are primarily based on RNA attached to magnetic particles that facilitate their washing and isolation. Isolation of RNA from formaldehyde-fixed, paraffin-embedded tissues is sometimes carried out, but not recommended due to the degradation problem discussed earlier. Degradation can be prevented by immersing the tissue-sample in RNA storage reagents such as RNAlater (Ambion) or by processing partially and storing as a phenolic emulsion (Trizol). In any case, often RNA samples are separated into size-specific classes (such as mRNA, miRNA, etc.) at this isolation stage using the miRVana column-system by Ambion or something similar. Some researchers prefer isolating RNA initially as total RNA and then separating it into various sizes by means of polyacrylamide gel electrophoresis (PAGE). However, in this latter case, the total RNA sample is typically contaminated with genomic DNA and has to be treated with DNase to digest the contaminating DNA before library preparation, followed by inactivating the excess DNase using reagents.

After isolation, it is important to check the quality of the isolated RNA in terms of purity, quantity and the extent of degradation. Some researchers use devices such as Nanodrop that are user-friendly, require nano-liter amounts of starting material, produce quick readings and have some parallel-processing capability. However, such devices can neither tell the difference between DNA and RNA (thereby failing to provide any information regarding DNA contamination) nor distinguish between degraded and intact RNA. An alternative is to use a system such as QubitFluorometer (by Life Technologies) that employs a more direct method for measuring the RNA and the DNA in the sample, resulting in more specific, accurate measurements of RNA with a wider dynamic range. These systems provide information regarding DNA contamination, but are still unable to measure RNA degradation. To overcome this stumbling block, Agilent Bioanalyzer uses a method different from the fluorescence-based techniques of Nanodrop and QubitFluorometer. Agilent Bioanalyzer is a micro-fluidics capillary electrophoresis-based system for measuring nucleic acids. In addition to being highly sensitive and requiring tiny amounts of starting material, it contains a microchip that is programmed for size control and has space for up to 12 tissue-samples at a time. Samples are mixed with polymers and a fluorescent dye, which are then loaded and measured through capillary electrophoretic movement.

The final step before sequencing is the conversion of the RNA into a cDNA library representing all of the RNA molecules in the tissue-sample. For RNA-seq library preparation, standard library protocols require 0.1–10 microgram of total RNA, although high-sensitivity protocols can work with as little as 10 picogram of RNA. This conversion of RNA to cDNA has two advantages – the chemical stability of DNA as well as the fact that DNA is more amenable to the sequencing chemistry and protocols of various sequencing platforms. Each commercial RNA-sequencing platform has its own library preparation protocol that is made available by the manufacturer. The major steps in library preparation are as follows. First, starting with the pure, intact and

quality-checked total RNA obtained from the tissue-sample, mRNA has to be purified out of it (the commercial platform ILLUMINA does it in two steps, first by exposing total RNA to magnetic beads and then by washing away the unwanted, non-specifically bound rRNA and other kinds of RNA from the magnetic beads). Secondly, the purified mRNA strands have to be fragmented into smaller pieces by incubation with a fragmentation reagent and the pieces have to be primed with random hexamer primers. Next, the primed mRNA fragments are reverse-transcribed into cDNA fragments using the enzyme reverse transcriptase. For each cDNA fragment thus obtained, the second (i.e. the opposite polarity) strand is then synthesized and the original mRNA fragment is removed, leaving behind only double-stranded cDNA. This double-stranded cDNA is further purified by attaching them to paramagnetic beads and washing those beads to get rid of unwanted stuff such as enzymes, buffers, free nucleotides and residual RNA. Subsequently, the double-stranded cDNA is eluted from the beads and subjected to a process called end-repairing. The 3′ ends of the end-repaired double-stranded cDNA fragments are then adenylated (i.e. the nucleotide adenine or A is added to those ends). They are now ready for the next step, which is adaptor ligation. Adaptors (which are short sequences of nucleotides themselves) are ligated to both ends of the end-repaired and adenylated cDNA fragments in such a way that there is a 6-nucleotide difference between the adaptor-sequences of cDNA fragments to be used for two different library reactions. The idea behind using a different adaptor-index for each library reaction is that it allows for pooling libraries later for sequencing and still keeps a way of tracing a sequence-read back to its original library based on its adaptor-index. A new alternative technique called tagmentation is more efficient as it fragments cDNA and incorporates adaptor sequence-tags simultaneously with the help of transposase enzymes. Nextera (under the Illumina platform) is the only commercially available system employing this technique. Thereafter the adaptor-ligated, end-repaired double-stranded cDNA fragments are purified once more (using paramagnetic beads again) and the library is enriched via several (typically 12–16) cycles of polymerase chain reaction (PCR) amplification, with the nucleotide sequences of the adaptors serving as primer-binding sites. Certain types of nucleotides containing triphosphates (deoxynucleoside triphosphate or dNTP) play the role of substrates for this chain reaction. The output of this process, after undergoing another round of purification using paramagnetic beads, is the final library representing the original mRNA in the tissue-sample under study.

5.2 Validation, pooling and normalization

Before being sent to the sequencing machine, the libraries still have to pass through a few quality-control steps. First, the quality of a library has to be validated by either quantifying the yield of double-stranded cDNA in it or visualizing the abundance and size-distribution of the library via PAGE (capillary electrophoresis in the case of Agilent Bioanalyzer). Once validated, several (half a dozen to two dozens) libraries are normalized and pooled together. This is because modern sequencing platforms can handle many libraries simultaneously and normalization evens out the amounts of double-stranded cDNA in different libraries (by, for example, diluting them differently), ensuring that all libraries are equally represented. At last, the normalized libraries are pooled together and passed on to the sequencing machine.

5.3 Sequencing

5.3.1 Single-end vs. paired-end

The sequencing machine reads the base pairs from the ends of the cDNA fragments. Sequencing both ends of each fragment is now routinely done in many platforms and is known as paired-end sequencing. Some platforms carry out single-end sequencing. At a later stage, when the end-sequences are mapped to locations in the genome with the help of mapping algorithms, the sequences obtained by single-end reading often fail to be mapped uniquely. This ambiguity results in decreased efficiency and increased loss of information, as these ambiguous sequences are usually discarded. Paired-end sequencing is a way around this problem, because if one of the two ends has an ambiguous sequence, the other end can first be mapped to a single location in the genome and that will most likely determine the location of the other end uniquely.

5.3.2 Generations of sequencing technology

In the next subsection, we briefly describe the various commercially available second-generation (or next-generation) sequencing platforms. To put things in perspective, first-generation high-throughput sequencing basically referred to Sanger dideoxy sequencing that used capillary electrophoresis for resolving nucleic acid fragment lengths. A standard run of this system would typically involve 96 capillaries and produce a sequence of length 600–1000 bases per capillary, which amounts to about 60,000 – 100,000 bases of sequence. Second-generation sequencing platforms mostly use a similar sequencing method by synthesis chemistry of individual nucleotides, but do so in a massively parallel way involving numbers of sequencing reactions in a single run that are several orders of magnitude more than first-generation systems (tens or hundreds of billions of bases). This incredibly high throughput of these more advanced systems gives them unprecedented sensitivity and the ability of discovering novel transcripts, small non-coding RNAs and transcription factor binding

sites. However, these systems tend to produce short reads (read-lengths in the hundreds of nucleotides). More advanced sequencing technologies (third generation) have recently begun to be introduced that use individual molecules of DNA or RNA as the starting template and produce longer reads per sequencing reaction (upwards of 1500 nucleotides), albeit carrying out fewer sequencing reactions per run.

5.3.3 Various next-generation sequencing platforms

In the last decade, a number of manufacturers have introduced next-generation sequencing platforms some of which are still in use and others have been discontinued. The following is a brief description of each major platform. See, for example [236] and [470] in this context.

5.3.3.1 Illumina

First introduced by Solexa about a decade ago and later marketed under Illumina's brand name, it is one of the most widely used systems today. It includes several versions of the Genome Analyzer as well as the more recent MiSeq and HiSeq. It uses fluorescently labeled nucleotides passing through a flow cell with many micro-fluidic channels or lanes. That is where the sequencing reaction takes place and detection signals are obtained via optical scanning. For simultaneous detection of nucleotide incorporation in millions of sequencing reactions, the four nucleotides are labeled with four different fluorescent labels. It is based on the Sanger "sequencing by synthesis" principle (mentioned earlier) but differs from the first-generation Sanger dideoxy sequencing in a number of significant ways. The top and bottom surfaces of each micro-fluidic channel are covered with oligonucleotide sequences that are complementary to the anchor sequences in the adaptors ligated to the cDNA fragments (see library preparation). When the libraries are passed through these lanes, the ligated adapters bind to these oligonucleotide sequences, thereby fixing or immobilizing the associated cDNA fragments onto the lane surfaces. Subsequently, each of these cDNA fragments serves as a template and is clonally amplified through a process called bridge amplification, whereby up to a thousand identical copies of each template are generated in close proximity, forming a cluster. These clusters (and not the individual cDNA fragments in the library) are used as the basic detection units, because otherwise the fluorescent signal intensity would be too low for the optical scanner. After cluster generation and the removal of one strand from the double-stranded cDNA fragments in the cluster, reagents are passed through the flow cell to carry out sequencing by synthesis. In each synthesis round, the addition of a single nucleotide (A, C, G or T) takes place and the corresponding fluorescent signal is imaged. The specific nucleotide is identified by this stored signal – a process known as *base calling*. A reconstruction of the sequence of additions at a particular location on the flow-cell surface (that corresponds to a particular bridge-amplified

cluster) produces the sequencing read for an original double-stranded cDNA fragment in the library. This reconstruction process can be performed at one end or both ends of the cDNA fragment, producing single-end or paired-end reads. Depending on the instrument model and the sequencing length, a single run of the Genome Analyzer can take anywhere between 70–280 hours, that of MiSeq takes 5–55 hours (producing up to 30 million paired-end reads) and that of HiSeq 2500 takes 7–265 hours (producing up to 6 billion paired-end reads).

Ideally, the simultaneous addition of nucleotides to the many identical copies (or bridge-clones) of a cDNA template in a cluster should be in perfect synchronization from one step to the next (i.e. should be in phase). In practice, a non-zero percentage of templates lose sync with the majority of templates in a cluster, that is, they either fall behind by a few bases or are a few bases ahead (known respectively as phasing and prephasing). This causes more background noise in the resulting dataset and decreasing accuracy (i.e. decreasing quality of base-calling) as more and more sequencing cycles are completed. Despite this, the Illumina systems have one of the lowest error rates (smaller than 1%) among all the commercially available systems, the most common type of errors being single nucleotide substitution.

The majority of the run time of an Illumina sequencing machine goes to imaging the clusters on the flow-cell surface tiles. For imaging, the fluorescent labels on the nucleotides are illuminated with red and green lasers and scanned through four different filters, producing four images on each tile of the flow-cell surface after every cycle. The raw images captured after each cycle are analyzed (using the proprietary software that comes with the instrument) to record the location coordinates of each cluster, its signal intensity and noise level. Next, this information is used in the base-calling phase by the Real Time Analysis software that makes the base calls and computes a quality score for each call, filtering out low-quality reads. Finally, the base-call files (bcl files) generated in the previous step are converted to FASTQ files by the proprietary software CASAVA.

5.3.3.2 SOLiD

It stands for Sequencing by Oligonucleotide Ligation and Detection and is commercially made available by Applied Biosystems of Carlsbad, California. Instead of the bridge amplification used by Illumina, this system uses a special process (emulsion polymerase chain reaction or emPCR) to amplify the numbers of copies of the DNA templates. Also, the sequencing chemistry is based on ligation instead of synthesis. Another important feature that distinguishes this platform from its competitors is a unique double-interrogation strategy that results in potentially greater sequencing accuracy and makes it the instrument of choice for single nucleotide polymorphism (SNP) detection. In this platform, a library of cDNA fragments is attached to magnetic beads (one molecule per bead). The DNA on each bead is then amplified by PCR

in an emulsion, with the amplified products still remaining attached to the bead. The amplified products are then covalently bound to a glass slide. A sequencing round begins with the addition of a universal primer to all cDNA fragments attached to the same magnetic bead. This renders the starting sequence of all those fragments known and identical. Using several primers that hybridize to that universal primer, di-base probes with fluorescent labels are competitively ligated to the primer. That is the start of a circle. If the bases at the first and second positions of the di-base probe (starting from the 3′ end) are complementary to the DNA that is being sequenced, then the ligation reaction takes place and the fluorescent label generates an optical signal. The remaining unbound probes are washed out. Subsequently, the primer and the probes are all reset for the next round. Now the 5′ end of the new universal primer will match to the base just preceding the earlier base. In this way, primers are reset by a single nucleotide five times so that at the end of the cycle, at least four nucleotides will have been interrogated twice (due to the di-base or dinucleotide probes) and the fifth nucleotide at least once. Ligation of subsequent di-base probes ensures a second interrogation of even that fifth nucleotide. The entire sequencing step consists of five rounds and each round consists of five cycles. This double interrogation method causes a single nucleotide polymorphism (SNP) to produce a two-color change, whereas a measurement error results in a single color change. This is a reliable way of distinguishing between a true SNP and an error in sequencing. However, there is a price to pay for this. Decoding the raw data from SOLiD has an added layer of difficulty as it encodes by two bases instead of one, which necessitates an alternative representation of the nucleotide sequence (called the "color space"). Instead of each single nucleotide being represented by one specific color, as was the case with Sanger-type sequencing chemistry, here each color stands for four potential two-base combinations. This makes any attempt to directly translate color-reads to base reads more error-prone. The best solution to this problem is converting the base reference sequence itself into the color space and then converting it back to the nucleotide space once the sequence has been aligned to a reference genome encoded in color space. All this is, however, easier said than done and results in a substantially higher error rate for SOLiD than Sanger-type sequencing platforms, despite the potential for increased accuracy coming from the double interrogation strategy. The latest instruments (e.g. the 5500 W) have gotten rid of magnetic bead amplification and used flow-chips in the place of amplifying templates. A machine using two flow-chips can produce up to 320 gigabytes of data in a single run.

5.3.3.3 Ion Torrent semiconductor sequencing

It is the first next-generation sequencing platform that does not involve chemically modified nucleotides, fluorescence labeling and image scanning. This increases the speed, decreases the cost and leaves a smaller equipment footprint on the final output. This platform uses the adaptor-ligated library preparation step (involving clonal amplification by emulsion PCR) followed by the synthesis-based sequencing chemistry of other platforms. But instead of detecting optical signals or photons from fluorescently labeled nucleotides, it detects changes in the pH (a measure of proton or hydrogen ion concentration) in a well when a nucleotide is added and protons are released. When a nucleotide is newly incorporated into a DNA strand, the chemical reaction catalyzed by DNA polymerase releases a pyrophosphate group and a proton, with the latter causing a pH change in the vicinity of the reaction. However, the extent of change is not specific to any particular nucleotide, so to determine the DNA sequence, each of the four substrate nucleotides is added to the reaction in sequential order. If a pH change is detected after the introduction of a nucleotide, it strongly indicates that the template DNA strand contains its complementary base at the latest position. In order to detect these very small pH changes, this platform resorts to semiconductor technology.

The sequencing takes place on a set of ion semiconductor chips, each of which contains an array of micro-wells. Each micro-well has in it one single-stranded template DNA and one molecule of DNA polymerase. Underneath each well is an ion sensitive layer. The micro-wells are flooded with a particular type of deoxyribonucleotide triphosphate or dNTP (N = A or T or C or G) sequentially, one after another in order. In each round, one of these dNTPs gets affixed to the template strand, depending on whose complementary base there is at the latest position on the template. The associated chemical reaction releases a proton that triggers the ion-sensitive material underneath. The electric pulses from those sensors are directly transmitted to a computer which immediately translates them into a DNA sequence. The final steps of signal processing and DNA sequence assembly are accomplished via embedded software.

For all the advantages of the above-mentioned technique, there are drawbacks too. The overall error rate of this platform is higher than that of the Illumina platform. The primary reason behind that is indels caused by homopolymers. When the DNA template contains a homo-polymeric region (i.e. a stretch of identical nucleotides), the pH change is stronger and proportional to the number of repeats of that nucleotide. Consequently, as the number of repeats (n) increases, there is a gradual decrease in the signal-strength ratio between n repeats and n-1 (or n+1) repeats. This limits the ability of the system to detect the total number of repeats correctly. In the current version of the system, the error rate for detecting a 5-base homo-polymer (i.e. when n = 5) is about 3.5%. Another significant drawback is the short size of the reads (35–400 base pairs per run)

5.3.3.4 Pacific biosciences single molecule real-time sequencing

The Pacific Biosciences single molecule real-time (SMRT) sequencing platform is considered as a third-generation technology, due to its higher sensitivity than the second-generation platforms described above and the resulting

capability to sequence single DNA molecules. Not only does it render any form of amplification unnecessary, but also it generates much longer reads (median length 8–10 kilobases, longest around 30 kilobases) than most other platforms. Like the previously mentioned platforms, it is also based on sequencing by synthesis, but a crucial difference is that it uses nucleotides carrying fluorescent labels linked to their end phosphate group but no terminator group (as would have been the case with Illumina). When a new nucleotide gets added to an elongating DNA strand, with the cleavage of the end pyrophosphate group, the fluorescent label is released at the same time. This makes real-time signal detection possible. The sequence-detecting signal is continuously recorded as a movie at a speed of 75 frames per second instead of using separate scanner images.

SMRT uses what are called *zero-mode waveguides* (ZMW). These are space-restricted chambers that allow light energy and reagents to be guided into extremely small volumes (of the order of zeptoliters or 10^{-21} liters). To be more precise, a ZMW is a hole only tens of nanometers in diameter that is microfabricated on an ultra-thin metal film (100 nanometers in thickness). The metal film is deposited onto a glass substrate. This provides a single chamber that contains a single molecule of DNA polymerase and a single DNA molecule that is to be sequenced in real time. The diameter of a ZMW being smaller than the wavelength of light, only the bottom 30 nanometers of the ZMW is illuminated by the light coming through the glass substrate. The resulting ultra-miniscule detection volume leads to a substantial reduction in background noise and makes it possible to detect nucleotide incorporation into a single DNA molecule. Using specific flurescent nucleotide triphosphates, the addition of an A, C, G or T to an elongating nucleotide chain can be detected while it is being synthesized. Because of this speed advantage, the runtime can be very short (i.e. of the order of 1–2 hours). The current version of the instrument (PacBio RS II) with its current movie-length of 3 hours can produce up to 375 megabases of sequence in a single run.

Since it does direct sequencing of single DNA molecules, it has another advantage over some of the earlier platforms. It has the capability of detecting nucleic acid modifications (such as DNA methylation, that is, 5-methyl cytosine formation). While the presence of such modifications causes consistent delays in the kinetics of the DNA polymerase used in sequencing, this has been exploited by this platform for the detection of DNA modifications (currently claiming to detect up to 25 base modifications in a single run). Two notable disadvantages of this platform are its high error rate (10–15%, with the most common error-type being indels) and high run cost. Paired-end sequencing is not possible on this platform. Also, the amount of DNA sample required at the beginning is on the high side (1000 nanograms). The library preparation steps for SMRT are similar to some other platforms and involves shotgun fragmentation of DNA into appropriate sizes, fragment end repair and adapter ligation, annealing to sequencing primers and binding of DNA polymerases with it.

5.3.3.5 Nanopore technologies

Nanopore sequencing is a third-generation, single-molecule technology that uses a single enzyme to separate a DNA strand and guides it through a protein pore embedded in a membrane. The simultaneous passage of ions through the pore generates an electric current which is measured. The specific nucleotides passing through the pore (that is, A, G, T or C) impede this current flow differently. This makes the current sensitive to the specific nucleotides and produces a signal that is detected in the pore. This, although conceptually simple, is technologically quite challenging as it involves measuring very small changes in electric current at the single-molecule scale. However, once this technical barrier is overcome and this platform is successfully commercialized, one of the greatest advantages of it will be a small device size (like a cellular phone or even a USB stick). At least that is the claim by the companies that are currently developing it. Overall, this approach is quite promising and will likely have a greater impact on genomics in the future.

5.3.3.6 Choosing a platform

The key factors that go into the choice of a platform are accuracy, the number and lengths of reads, the preference for paired-end over single-end, the amount of sample material needed, cost and run time. If the goal is to detect single-nucleotide polymorphisms (SNP), then choosing a platform with a very low error rate (and hence very high accuracy, such as SOLiD) is essential because of the very low frequency of SNP occurrence and the need for distinguishing the true SNPs from sequencing errors. However, it is important to remember in this context that one can compensate for low accuracy by generating more reads, that is, by repeatedly sequencing the same piece of RNA. On the other hand, if one is trying to quantify gene expression to detect differential expression, identify known protein-coding genes or discover new genes, the need for accuracy is much less stringent. In that case, most of the platforms described above (Illumina, SOLiD, Ion Torrent) can be used. One can see, for example, [236] for more details on this issue.

Regarding lengths of reads, longer reads are needed to reduce the percentage of reads mapping to multiple locations of the reference genome. A relatively long read (e.g. 50 nucleotides) will bring down that percentage below 0.01%, which is good enough for detecting differential expression. Reads longer than that will be essential for purposes such as annotating novel genes in a species for which neither a reference genome nor any other sequence data-source is available. For this, the newer generation models produced by Pacific Biosciences (PacBio RS II) is suitable. The question of how many reads are needed should be put in perspective by considering the overall genome size of the organism under study and the proportion of its genome that protein-coding genes occupy. For humans, those two numbers are about 3 billion nucleotides and 3.3%. This means that 1 million paired-end reads of length 50 nucleotides each (or single-end reads of length 100 nucleotides each), which amount to 100

million nucleotides of sequence data, will cover all the protein-coding genes once. So, if a particular platform has an output of 20 million reads, it will provide 20 times the coverage which means a high to decent amount of coverage for a vast majority of genes and possible omission of a few rarely expressed or low-expression genes. These days the typical output from the common technology platforms is about 30 million, which is good enough to capture a vast majority of (but probably not all of) the genes expressed in a sample. When better coverage is crucial for some reason, one should choose the platforms that yield a large number of reads more easily. Two things to keep in mind are that no platform may ever be able to provide enough coverage to obtain every single transcript from each locus, and there is no consensus on the question of how many reads are necessary to confirm the existence of a transcript.

The issue of single-end vs. paired-end reads has been discussed earlier, so next we comment on the amount of sample material needed. Sequencing platforms that use amplified, double-stranded cDNA basically have no lower limit on the amount of material needed because of the amplification. Also, platforms are available that can sequence the total mount of RNA from a single cell. So any discussion on the amount of sample material may seem irrelevant, except for one reason – the fact that providing more than the minimum required amount of tissue sample to a platform increases the representation of the RNA species in that sample. So in general it is not recommended to supply only the bare minimum needed.

The cost aspect, similarly, has lost some of its relevance over the years as the cost of sequencing has gone down significantly during the past decade. Still, it remains an important issue as not all research projects are equally well-funded and the standards for data quality have also gone up over the years. Purchasing a laboratory sequencer for oneś personal use is now more feasible than ever (e.g. the Personal Genome Machine by Ion Torrent, MiSeq by Illumina, etc.), but outsourcing the sequencing job by sending tissue samples or RNA-seq libraries to commercial NGS facilities still remains an effective way of cost mitigation. Regarding the downward movement of sequencing cost over time, it is widely believed that the bottom has not been reached yet. This belief is reinforced by the increasing competition among commercial and non-profit core NGS facilities.

In a fast-moving field such as genomics, it is neither desirable nor ideal that a sequencing run gets delayed. The unfortunate reality, however, is that a number of platforms experience delays. It is even more frustrating to know that the delay is not caused by the sequencing machine running slowly, but by the insufficiency of libraries to fill a flow cell for a single run and the resulting wait for more libraries to be submitted. If this can somehow be avoided by careful planning, such delays will not occur. Also, further downstream, the huge amount of data generated from an NGS platform often take a really long time to be preprocessed and analyzed. Compared to it, any delay at the sequencing stage may seem insignificant.

5.4 Factors affecting NGS data accuracy

In addition to any errors in base calling (i.e. the final identification of an A, T, C or G done by an NGS platform), there may be biases creeping into the earlier steps of the whole process. Biases can affect both the library preparation step and the sequencing step. It may not be possible to completely eliminate these biases and other factors contributing to inaccurate signals, but we can still be mindful of them while choosing the experimental design, the statistical model, the analysis method and algorithm so as to minimize their adverse effects on the final outcome. See, for example, [470] in this context.

5.4.1 At the library preparation stage

At the very beginning of the library preparation stage, DNA fragmentation by sonication and nebulization tends to break the DNA strands more often than expected after a cytosine (C), compared to the other three. This violates the complete randomness assumption that we make about the fragmentation process. Subsequently, the size selection process of the DNA fragments also introduces bias in its own way. For example, the use of a high gel-melting temperature in the gel extraction method is biased towards recovering DNA fragments with higher GC content.

Next, the ligation step introduces some bias that affects the sequencing of both long-stranded RNA species and short-stranded ones, though in different ways. Size-selected DNA fragments (double stranded) are first subjected to repair and adenylation (i.e. creation of A tails) at both ends, followed by ligation of adapters carrying 5′ T overhangs. This adapter ligation process has been found to be biased against DNA fragments starting with a thymine. This bias affects the cDNA molecules obtained via reverse transcription from messenger RNA or long non-coding RNA, but it does not affect short-stranded RNA molecules (e.g. siRNA) because adapter ligation for them precedes reverse transcription to cDNA. However, the small RNA adapter ligation process brings in another type of sequence-specific bias for some small RNA species, depending on their secondary and tertiary structures (which, in turn, are affected by the temperature and chemical composition of the ligation reaction mixture.

Afterwards, the adapter-ligated DNA fragments are amplified using polymerase chain reaction (PCR) that involves DNA polymerases. This process in known to be biased against highly GC-rich or AT-rich DNA fragments, resulting in an under-representation of the genomic regions that are of this type. One way this problem can be partly dealt with is by optimizing PCR conditions for GC-rich or AT-rich regions, but the only way to completely get rid of it is to use a library preparation process that does not involve PCR.

5.4.2 At the sequencing stage

As was described earlier, many NGS platforms are based on the sequencing by synthesis principle, which uses DNA polymerases. Consequently, the coverage bias against highly GC-rich or AT-rich genomic regions that was mentioned above is present at the sequencing stage too, and it is difficult to completely get rid of. In addition, other activities involved in the sequencing process often introduce their own biases and artifacts. For example, misalignment during scanning or unintended light reflections can lead to inaccuracies in imaging. The presence of lint, dust particles, crystals and air bubbles in the buffers could generate artificial signals. Fortunately, these problems can be avoided to a large extent by being sufficiently careful.

At the signal processing and base-calling steps, some platforms such as the Illumina Genome Analyzer suffer from problems that their proprietary software packages are effective in dealing with, but other commercial or open-source software often used by researchers have different algorithms for these tasks than the proprietary software and they produce varying results (due to making different assumptions on the signal distribution and other reasons). For example, some of them make the assumption that the signals from the four detection channels for A, C, G and T are independent. Some assume that signals from different cycles are independent. However, in reality, the A and C signal channels have some dependence due to the overlap between the emission spectra of their fluorescent labels. So do the G and T signal channels. Also, due to phasing and prephasing, signals from a cycle are dependent on those from cycles preceding and succeeding it.

5.5 Applications of RNA-Seq

The primary objectives of RNA-seq are the determination of the nucleotide sequence (i.e. the particular order of the A, C, G and U residues), learning of the gene structure (i.e. locations of promoters and enhancers, 5′ and 3′ untranslated regions, exon-intron junctions, poly-adenylation sites, etc.) and quantification of the abundance of RNA molecules (i.e. the absolute and normalized numerical amount of each specific sequence) in a tissue sample. Each of these, in turn, opens up other avenues of important applications. For example, knowledge of the sequence enables the identification of known protein-coding genes as well as the discovery of new genes or long non-coding RNA species. It also provides valuable information about the secondary structure (e.g. hairpin bends, bulges, etc) and tertiary structure (i.e. the three-dimensional shape of the molecule) which are crucial in determining the class and function of it (e.g. whether it is a transfer RNA or micro-RNA, etc.). Quantification of abundance enables us to detect differential gene expression between two tissue

samples or two experimental conditions or two organisms. Below we elaborate a little more on some of these.

The sequence reads from an RNA-seq platform are usually mapped at first to known protein-coding genes archived in existing databases. This not only helps us confirm known intron-exon boundaries but also discover completely new exons. In addition, this enables precise identification of a gene's important structural features such as the 3′ untranslated region (UTR), the 5′ transcription start site (TSS) or polyadenylation sites. Based on RNA-seq read counts, it is possible to compare the usage of one axon to that of another (i.e. which one was busier in the transcription process and how much busier). Due to the massively parallel nature of the RNA-seq platforms, it is possible to do all of these in a genomewide manner.

Prior to RNA-seq, annotations of protein-coding genes used to depend on computational predictions based on genomic sequences. A high throughput technology such as RNA-seq not only allows us to verify many of those previous predictions but also to discover novel protein-coding genes with no previous prediction. This is crucial when there is no genome sequence database available for an organism and we are trying to build its transcriptome solely on the basis of its RNA-seq reads.

It has already been mentioned that RNA-seq makes the detection of differential gene expression between two tissue-types (or experimental conditions or organisms) possible. Another important purpose that RNA-seq serves is the study of quantitative traits. Just as genome-wide association studies (GWAS) link single-nucleotide polymorphisms (SNP) to various physical or physiological traits that are quantifiable, eQTL is the study of association between gene expression changes and SNPs. As is now well-known, such association can be a direct or indirect causal relation (direct or local if, for example, the SNP is located in the enhancer region of a gene, thereby changing the gene expression; indirect or distal if, for example, the SNP is located far away from the gene but structurally changes a transcription factor that was needed for the gene's expression and renders it non-functional). As RNA-seq quantifies gene expression levels, it can be used to find these associations. A newer branch of association studies known as sQTL tries to find the association between SNPs and the location and usage of gene-splicing sites.

With the advancement of the RNA-seq technology and its ability to yield more and more reads that are longer, it has now become possible to identify rare transcripts that are potentially important. One example is the transcripts produced by "fusion genes." Such genes come into existence when two previously separate genes contribute parts of their own structures (e.g. protein-coding region, 3′ poly-A region or 5′ UTR) to a fused structure. This is quite common in cancer tissues. However, recently this phenomenon has been detected in normal tissue via RNA-seq studies, implying that it is not necessarily indicative of a disease condition.

The advent of RNA-seq has greatly enhanced our capability of finding long and short non-coding RNAs that are abbreviated as lncRNA and miRNA

respectively. Their existence was known before the RNA-seq era, but this technology has opened our eyes to their plentifulness and ubiquity. An lncRNA is a transcript that is longer than 200 nucleotides, produced by a region that is not a part of (or overlapping with) a protein-coding exon, and does not belong to other known non-coding RNA species such as transfer RNAs and ribosomal RNAs. They are now known to play some role in epigenomics by controlling transcription as enhancers and by binding with histone proteins to change their functions. Short non-coding RNAs can actually be of several different types, the most well-known of which is micro-RNAs or miRNA. The approach developed for sequencing miRNAs (called miRNA-seq) can be implemented on most common sequencing platforms, once the miRNAs are converted to double-stranded cDNAs. However, the steps needed to convert the starting material (either size-selected and fragmented small RNAs, or the total RNA which is all RNA species combined) to double-stranded cDNAs differ from the pre-sequencing protocols of the RNA-seq platforms described earlier in this chapter. The miRNA-seq method can also be used to sequence other species of short non-coding RNA species. Micro-RNAs are known to play a role in regulating gene expression by binding with and degrading messenger RNAs and preventing their translation.

Recently, RNA-seq has been used for "exome sequencing" or "exome capture" which is not RNA-seq in the strictest sense of the word. The purpose is to identify variations in the protein-coding gene sequences from genomic DNA samples. The idea is to sequence fragmented genomic DNA enriched for exons by means of hybridization to exonic sequences. Since the primary motivation behind this has been studying diseases in humans, and SNPs (as well as other variations) need to be identified from large human cohorts, sequencing only the exonic sequences of an individual is a convenient way to keep the cost down. Another advantage is that, focusing on just the exons automatically means focusing mostly on protein-coding genes, so the variations (SNPs, etc.) found in this way are directly relevant to protein structure modification.

Interacting with a variety of proteins (e.g. transcription factors) is a crucial aspect of the day-to-day normal functioning of a genome. Many of these interactions take place in a region-specific manner. Figuring out which specific regions of the genome the interacting proteins bind to (known as *transcription factor binding sites* (TFBS)) starts with the capturing of the protein-bound regions via a mechanism called *Chromatin ImmunoPrecipitation* (ChIP) and then sequenced using an NGS platform. This is one of the most important uses of the NGS technology.

We conclude this section with a few more applications of next-generation sequencing that are proving to be increasingly important. As was mentioned earlier in the context of Pacific Biosciences Single Molecule Real-Time sequencing, some NGS platforms are useful in epigenomics because of their ability to detect DNA methylation (i.e. conversion of the cytosine residue to methylcytosine). Lately, RNA-seq has started replacing Sanger sequencing (once the gold standard) as the preferred technology platform for *de novo* genome assembly of an organism – even for those with large and complex genomes. And last but not least, recently RNA-seq has started being used in metagenomics (the study of all genomes present in a community of organisms) as it is able to quickly sequence everything that there is in a metagenome and give us an overall profile of the composition and functional state of a microbial community.

5.6 RNA-Seq data preprocessing and analysis

Generally speaking, RNA-seq data analysis involves three major tasks. The first one is base calling, which is based on a deconvolution of the optical or physicochemical signals produced by the sequencing mechanism. Almost all sequencing platforms store these base-call results in the FASTQ file format, each FASTQ file containing a huge number of reads (i.e. the A-T-G-C orderings of DNA fragments sampled from a library). Next comes the data quality check or quality control step where the reads in a FASTQ file are checked for their quality and preprocessed before being mapped to a reference genome. Depending on the results of a number of quality metrics that are examined, the FASTQ files are preprocessed to weed out low-quality reads, eliminate portions of reads containing low-quality base calls and get rid of any unwanted items (such as PCR primers and adapter sequences). This is followed by the mapping or alignment of the preprocessed reads to a reference genome in an attempt to find the genome locations where the reads most likely came from. Once this is done, application-specific downstream analysis can begin.

5.6.1 Base calling

Base calling is carried out using algorithms in the proprietary software packages that are platform-specific and come with the instruments (e.g. Illumina's *Bustard* algorithm).The output of such an algorithm is a base call (i.e. identification of a nucleotide) for each sequencing cycle along with a confidence score for that call. These are stored in a file format such as FASTA, CFASTA, QUAL or FASTQ, with the last one being the most widely used. Conversion tools are available for the other file formats to be converted to FASTQ which is a text-based format. The confidence score or quality score (Q-score) that comes with a base call is obtained as $Q = (-10)log_{10}P_E$ where P_E is the probability of an erroneous base call. Typically a Q-score of 20 (which corresponds to $P_E = 0.01$) is the minimum requirement for a base call to be deemed reliable, although the value of Q can go up to 60. The reporting of this Q-score in a FASTQ file, however, is not done by numbers. They are usually encoded with ASCII characters, the most commonly used encoding scheme being the one introduced by Sanger sequencing.

Here are some more details about Illumina's base calling mechanism and the quality scores that come with it from [293]. During each cycle of Illumina's "sequencing by synthesis" process, images produced by a charged coupled device (CCD) record fluorescence intensities in each of the four nucleotide channels. These are stored in an intensity matrix whose columns correspond to the cycles and rows correspond to the channels. Bustard converts these observed intensities into concentrations by multiplying them with the inverse of an estimated "crosstalk" matrix to adjust for the correlation among the four channels. As more and more cycles go on, loss of fragment copies in the library results in reduced intensities which, in turn, leads to reduced concentrations. As a way arond this, Bustard rescales the concentrations in each cycle by a factor proportional to the reciprocal of the average concentration for the cycle. So all cycles end up having the same average concentration. Next, Bustard uses a Markov chain model and its transition probabilities to estimate the probability of one base being correctly synthesized during a cycle, that of no new base being synthesized during a cycle (known as *lagging* or *phasing*), and that of two bases being synthesized during a cycle (known as *leading* or *pre-phasing*). The algorithm adjusts the rescaled concentrations using these estimated probabilities. Subsequently, these adjusted concentrations are used to make base-calls and assign quality scores for the called bases. By assigning such quality scores, it is possible to assess the performance of a base-calling algorithm and compare one algorithm with another (some other algorithms are briefly mentioned below). The conventional quality-scoring algorithm that most people use is *Phred*. It involves a four-phase procedure to estimate a number of parameters regarding peak shape and peak resolution. Then it takes these parameter estimates and searches for a quality score that corresponds to those estimates in some known table of quality scores. The *Phred* quality scores have been found to be quite accurate for several different types of sequencing platforms.

Among the recent alternatives to Bustard is *Alta-Cyclic* [100] that takes a support vector machine (SVM) approach to build a classifier which classifies each newly synthesized nucleotide as one of the four possible bases. Like every supervised machine-learning algorithm, this SVM must be trained first, which is done using a known reference genome in one of the flow cell lines. Another alternative is *Rolexa* [370] that is used in the Solexa platform. Its probabilistic algorithm corrects for positional bias, phasing, rephasing and crosstalk. Next, it estimates the conditional probability of each base given a quadruple of intensities (the quadruple being modeled as a mixture of four multivariate normal densities). The one with the highest conditional probability is called. This algorithm also involves a procedure for the identification and removal of ambiguous base-calls based on entropy calculations. Other alternatives are *BayesCall* [220] that uses a full Bayesian model for the four bases, concentrations of active templates and the observed fluorescence intensities with cycle-dependent parameters and estimation using variations of the EM algorithm (Monte Carlo Expectation Maximization and Expectation

Conditional Maximization), *BING* [237] that uses pixel-based base-calling as opposed to cluster-based base-calling, [33] who probabilistically model the log intensities in such a way that it includes a read effect and a base-cycle effect as well as latent indicator variables for each possible base in each read and cycle, *Ibis* which uses a multi-class SVM classifier and makes phasing in a given cycle dependent on the intensity values from the two adjacent cycles (before it and after it), *AYB* or All Your Base [285] that fits a cluster-specific multivariate regression model for the intensity matrix via the iteratively reweighted least squares (IRLS) algorithm, *freeIbis* [357] which is claimed to be a more efficient version of *Ibis* with calibrated quality scores for the Illumina platform.

Now, back to the discussion on Q-scores. Because each sequencing platform has its own calibration of the Q-score, if the scores obtained from different platforms are to be compared or combined, it is necessary to have an idea about each of their calibration methods and re-calibrate all of them so they become comparable. The platforms use either a control lane or a precomputed calibration table to come up with their own P_E. To make the resulting Q-scores comparable, they are re-calibrated in the following way. A subcollection of the reads is used that map to regions of the reference genome containing no SNPs, so any mismatch between the reference sequence and the reads can be attributed to a sequencing error. Depending on the rate of such mismatches at each base position of the reads, a new calibration table is constructed, which is then used for recalibration.

5.6.2 Quality control and preprocessing of reads

The first step is a general quality control analysis which examines the overall quality of the millions of reads. The overall quality assessment includes scanning the reads for low-confidence bases, biased nucleotide composition, duplicates, adapters, etc. The output of this step are the number of reads and some quality metrics, which are also the input for the second step (i.e. preprocessing). The primary goal of preprocessing is not only the removal of low-quality bases but also that of various artifacts from the individual reads (e.g. adapters, poly-adenine tails, microbiome, etc.). It may also involve trimming and filtering. Once all this is done, the data will be ready for the subsequent step of read alignment to a reference genome.

5.6.2.1 Quality control

The three main data quality metrics are Q-scores, the read length distribution and the percentage of each base across base positions. Examining Q-scores can be done across all base positions of all reads from the first to the last sequenced base. Or it can be done by plotting the average Q-score of each read and looking at their distribution pattern. Either way, the goal is to have a vast majority of reads with an average Q-score > 30 and minimize the percentage of reads with an average score < 20. For platforms based on sequencing by

synthesis, due to reasons mentioned earlier, usually the base positions that are in the early phases of a sequencing run tend to have higher Q-scores than those coming later in the process. It is necessary to have a median Q-score of at least 20 even for these late-phase base positions. If the median Q-score drops significantly below that threshold, the affected base positions must be scrutinized and low-quality bases have to be trimmed from the reads concerned. Also, an increasing number of the ambiguous call "N" (which happens when none of the four bases can be called with enough confidence) is indicative of diminishing base-call quality.

Regarding the read length distribution, it is less of a worry for platforms that produce reads with a high degree of homogeneity in their lengths. For some platforms such as the Pacific Biosciences, however, there is considerable heterogeneity in the read lengths, and so the read length distribution is something to watch carefully. In addition to indicating the total volume of useful data generated in a sequencing run, it also shows the relative amounts of shorter and longer reads. If the cumulative distribution function (CDF) of the read lengths from a platform is stochastically larger than that from another platform (both having comparable base-quality and equal data volume output), then the first platform is preferable because longer reads have an advantage in the upcoming steps of this process. A CDF being stochastically larger than another means that the graph of it is always below that of the competing CDF.

If DNA fragmentation is truly random during the library preparation process, the probability of observing each of the four bases at each base position should be the same. As a result, if one plots the percentage of each base across all base positions, the plots for adenine, guanine, thymine and cytosine should be approximately parallel to each other. Also, the overall percentage visible in each plot should be proportional to the overall frequency of each base in the initial library. Any significant departure from a parallel configuration points to anomalies in the library preparation process, such as nonrandom DNA fragmentation or undesirable over-representation of some RNA species in the library.

The quality control steps are usually carried out using software packages such as *NGS QC Toolkit* [335], *FASTX-Toolkit* or *FastQC*.

5.6.2.2 Preprocessing

If the quality control step detects low-quality reads and/or unwanted artifacts, the preprocessing step gets rid of them. Low-quality base calls at the 3′ end of the sequence are trimmed away, along with artifacts such as adapters and poly-A tails as well as duplicated portions. Some platforms routinely do sequence filtering as a part of their FASTQ file generation process, while others don't. In either case, if the quality scores are found to be below the acceptable threshold 20 in the quality control step, further filtering and/or trimming will be required using software tools such as *Trimmomatic* [30], *ngsShoRT* [52] or

Sickle which is a sliding-window, adaptive, quality-based trimming tool for FASTQ files. In addition to these specialized tools, preprocessing can also be performed using options available in the quality control software packages mentioned earlier.

5.6.3 Read alignment

This step is aimed at finding the point of origin for each individual read by mapping them to a reference genome. If a reference genome is not yet available, reads can be mapped to a transcriptome (created from the reads themselves via a *de novo* assembly method). Mapping a read to a reference genome involves a sequence alignment. Mapping is computationally intensive due to several reasons – the huge number of reads, the large size of a reference genome, the need to map spliced reads non-contiguously, etc. Because of this, often the genome sequence is transformed and compressed into an index to expedite the mapping process. For example, the commonly used Burrows-Wheeler transform – an idea borrowed from string matching theory. Nevertheless, simultaneous mapping of millions of reads (often short in length) to a big reference genome remains a challenging task, unlike mapping a single or a few sequence(s) of moderate length(s) using BLAST or other similar tools. A major source of difficulty is figuring out whether the deviation of a read from the reference genome is due to sequencing errors or it is a true sequence deviation (i.e. meaning that the genome sequence of the tissue sample under study truly differs from the reference genome as a result of indel mutations and/or polymorphisms). Another major challenge is the identification of novel splice junctions in RNA-seq data.

Several alignment algorithms have been put forward in the last one-and-a-half decades. Older algorithms such as BLAST used hash tables and seed-and-extend methods. Examples of newer algorithms based on the optimization of and improvement upon BLAST's original approach are Efficient Larg-scale Alignment of Nucleotide Databases (ELAND), Novoalign and Short Oligonucleotide Alignment Program (SOAP). As explained in [257], SOAP is based on reference genome indexing. So is Novoalign. However, ELAND and another one called Mapping and Assembly with Qualities or MAQ [256] are based on indexing the reads from the sample. It uses an ELAND-like hashing method for alignment, followed by a Bayesian statistical model to produce Phred-scaled quality scores for the resulting alignments (which is actually $10log_{10}(P_E)$ where P_E is the posterior probability of incorrect assignment). MAQ is quite efficient in combining the mapping quality information with the Q-scores and utilizing mate-pair information for paired-end read alignment in diploid samples. To understand the basic difference between the approach used by algorithms such as SOAP and Novoalign and that used by algorithms such as ELAND and MAQ, one needs to understand the original seed-and-extend approach of BLAST. In it, if a match happens between short nucleotide sequences (or "words") in the query sequence and the reference genome, that

matched region is used as a "seed" to extend the alignment to adjacent regions. BLAST only uses seeds that are consecutive sequences of nucleotides and exactly match portions of the reference genome. Since this seed selection criterion is a bit too stringent, resulting in lower alignment sensitivity (as it is not sensitive to sequence variations), some newer algorithms started using non-consecutive or spaced seeds, thereby enhancing the probability of finding a match. All four algorithms mentioned above do this, but some of them implement it differently than the others. SOAP and Novoalign start out by chopping the reference genome into small, equal-sized pieces and saving them in a giant hash table which is subsequently used to search for near-matches with similarly chopped pieces of the reads from the tissue sample. This is called reference genome indexing. ELAND and MAQ, on the other hand, construct the hash table from the reads and extract short subsequences from the reference genome to look for near-matches with the ones in the hash table.

In order to achieve more efficient indexing and faster searching, the algorithms that use the Burrows-Wheeler transform with an efficient backward search are *Bowtie* [242] and *BWA* [255]. Although *Bowtie* is a sub-optimal greedy algorithm, it does produce high-quality alignments with double indexing to prevent excessive backtracking. It also has built-in options for the users to choose their own balance between efficiency and accuracy. *BWA* allows for inexact matching and gapped alignment.

The software called *Tophat* [440] was designed for the discovery of novel splice junctions ab initio. This two-step algorithm starts by mapping all reads to the reference genome using Bowtie. The reads that do not map to the genome are designated "initially unmapped" or IUM. The next step is to assemble the mapped reads using the MAQ algorithm mentioned above and construct an initial consensus. Once this is done, the sequences that flank potential donor/acceptor splice sites within neighboring regions are joined to construct potential splice junctions. Finally, the IUM reads are indexed and aligned to these splice junction sequences. A more recent version of this algorithm that can handle variable-length reads and variable-length indels with respect to the reference genome is called *TopHat2*. Another relatively new and fast-running spliced alignment algorithm is *STAR* or Spliced Transcripts Alignment to a Reference. While it runs faster than TopHat, its memory-space requirement is considerably larger than that of TopHat. In addition to speed, the other advantages of STAR include its ability to perform an unbiased search for splice junctions as it does not need any *a priori* information regarding their locations, sequence signals or introl lengths. It can align a read having any number of splice junctions, mismatches and indels as well as those having poor-quality ends. It uses the so-called "maximum mappable length" approach which chops a read into 50-base long pieces and finds the best portion that can be mapped for each piece. Next, it maps the remaining portion of a piece (which may end up being mapped far away in the case of a splice junction). This sequential "maximum mappable seed" search looks for exact matches and uses the genome in the form of uncompressed suffix arrays. Subsequently,

it stitches the seeds together within a given genomic window, allowing for indels, mismatches and splice junctions. While doing so, it deals with seeds from paired reads concurrently to increase sensitivity.

The combined effect of all these new developments is that the run time for aligning tens of millions of reads to a big and complex genome is now minutes instead of hours. Clearly, the two central issues in sequence alignment are sensitivity and run time. With the advent of newer sequencing platforms capable of producing longer reads, another important issue is efficient alignment of longer reads. One more thing to keep in mind is the reference bias that creeps in if we stick to the same reference genome for all our alignment jobs. Some algorithms such as SOAP2 (a newer version of SOAP) and Bowtie are preferable if lowering run-time is the primary concern. On the other hand, a greater emphasis on sensitivity will dictate that we choose algorithms such as *SHRiMP2* [67] or *Stampy* [277]. The algorithm *GenomeMapper* [380] is capable of reducing reference bias by using multiple reference genomes simultaneously. Algorithms such as BWA-MEM, which is a modified version of BWA [254], LAST [228] and Basic Local Alignment with Successive Refinement or BLASR [45] are designed for efficiently aligning longer reads.

The output from this step is an alignment file listing the mapped reads and their mapping positions in the reference sequence. These are usually stored in a SAM (Sequence Alignment/ Map) or BAM (Binary Alignment/Map) file. The former is in a tab-delimited text format and the latter is a compressed binary version of it. A SAM/BAM file has a header section followed by an alignment section containing the fields "Query sequence read name (or Query template name)," "Bitwise flag," "Reference sequence name," "Leftmost mapping position (on the reference sequence)," "Mapping quality," "CIGAR string," "Reference name of the next read (for paired-end reads," "Position of the next read (for paired-end reads)," "Observed template length," "Segment sequence" and "ASCII of Phred-scaled base quality."

The contents of the SAM/BAM files are carefully checked to see the percentage of aligned (in particular, uniquely aligned) reads, detecting and filtering out multi-reads (i.e. reads that map to multiple locations in the reference genome) and doing the same to duplicate reads. Currently, even the top-of-the-line alignment algorithms are able to match only 70–80% of the reads to unique positions in the reference genome due to a variety of reasons – biological, technological and computational. Increasingly longer reads produced by the latest platforms and increasingly efficient algorithm development may eventually do away with the technological and computational reasons, but the biological ones (e.g. DNA polymorphism, mutation, presence of repetitive sequences in genomes, etc.) will still remain. Regarding multi-reads, which cause problems in the downstream analysis, there are two possible avenues. They can either be filtered out, leading to the removal of a substantial percentage of the reads and information loss, or recycled by algorithms such as *BM-MAP* [515] that probabilistically allocate a multi-read to one of the candidate positions in the reference genome that match it. Duplicate reads, whose existence is

usually a low-probability event unless there is PCR over-amplification, should also be detected after the mapping step and removed for the sake of performance enhancement in the later steps. However, their removal carries a risk. It is impossible to distinguish between biological duplicates (that exist naturally in a genome) and technical duplicates (created by PCR over-amplification). We actually want to remove the latter, but in the process, end up removing the former too and thereby lose true biological information. All three of the tasks mentioned above can be performed using software packages such as *Picard* or *SAMtools*, along with other necessary tasks (e.g. SAM-to-BAM conversion, indexing of those two file-types, merging of multiple BAM files to one, alignment visualization, etc.). Alignment visualization can be either using a text-based alignment viewer or a direct graphical visualization by the superimposition of mapped reads on top of the reference genome.

5.6.4 Genome-guided transcriptome assembly and isoform finding

Once the alignment is done, one important purpose for which it can be used is the discovery of novel genes and splice variants. Most sequencing platforms produce short reads which are much smaller than the length of an mRNA produced by a gene (with the exception of a few platforms such as PAC BIO II that can produce reads as long as a mature mRNA). So a single read does not shed much light on the detailed structure of a gene, such as its intron-exon organization, poly-adenylation sites and transcription start sites. Most exons are quite short (shorter than a couple of hundred base-pairs), so the order of alternative exons in a gene and their use must be reconstructed via mapping to the reference genome and then linking alignments from one region to another. This is called a mapping-based assembly of the transcriptome, as opposed to *de novo* assembly. Both assembly approaches involve constructing a graph for each gene locus based on RNA-seq reads. The graph serves as a starting point for resolving isoforms. However, construction of these graphs is tricky because it involves splitting the data in such a way that a single graph represents only a single locus.

Here we explain the graphical approach in the context of mapping-based assembly. First, any mapping algorithm that allows split (i.e. spaced or non-consecutive) reads can be used to align the RNA-seq reads to the reference genome. If gene models (i.e. structural details of genes) were available for the reference genome, this alignment itself would provide information about which exons belong to which genes. If no gene model is available for the reference genome, mapped reads must first be segmented in order to represent gene loci. Subsequently, a *splicing graph* (also often called an *exon graph*) is constructed for each locus. A *graph* is nothing but a collection of nodes (or vertices) and a collection of edges connecting some pairs of nodes. If you imagine trying to travel from one node to another in a graph, it will only be possible if those two nodes are connected via a sequence of consecutive edges (called a path). In a

splicing graph, each node is an exon, each edge represents an exon junction and a path represents an isoform. By applying a path-finding algorithm on such a graph, one can find out some or all of the paths that exist in it. The number of such paths depends on the edge-set of the graph. If the edge-set is so rich that the graph is fully connected (i.e. every pair of nodes is linked via a path), then all conceivable isoforms are possible. On the other hand, a sparse edge-set will make only some isoforms possible.

So, what kind of a topology (i.e. edge-set) should a particular splicing graph have? That is where the RNA-seq reads provide the crucial data-based guidance. The task is to choose a topology which best corresponds to the data. Those possible splice junctions (or edges) for which there is no support from the RNA-seq reads are removed from the graph, and only the edges with significant support are kept. "Support" here means split-reads and paired-end information. In the case of a split-read, if the beginning of the read is mapped to one exon and the end of the read to another exon, this provides evidence for these two exons to be adjacent in an mRNA sequence. In a paired-end case, the same applies to the two ends of the read-pair (i.e. one end mapped to one exon and the other end to another). This latter case is considered somewhat weaker evidence for the existence of an exon junction than the split-read case.

As [288] points out, accurate estimation of isoform abundance is very challenging if not all isoforms are known, since the read-pairs from unknown isoforms can affect the accuracy of the abundance estimation of the known ones. So, isoform finding is a crucial step before what we describe in the next subsection – abundance estimation and gene expression quantification.

The output of this transcriptome assembly step is gene models and transcript models. Assembled transcripts from different samples are merged and combined with reference annotation in order to produce more complete gene models. These can subsequently be used for gene expression quantification.

5.6.5 Quantification and comparison of expression levels

In this step, each single read is associated with a gene based on its mapping location. Expression of novel or familiar genes and their transcript production can be quantified using the gene models and transcript models from the previous step. When studying an organism whose genome is already well-annotated, obtaining gene models and transcript models via the method described in the previous section is not necessary. Instead, the reference annotation for that well-studied organism can be directly used. However, this restricts the quantification of expression levels to only known genes and transcripts. Expressions of novel genes for such organisms will still have to be quantified with the help of the previous section's output.

Abundance estimates can be reported in the form of raw read-counts or in normalized units such as FPKM (fragments per thousand nucleotides per million mapped reads) or RPKM (reads per thousand nucleotides per million reads). At this stage, the sequencing data simply take the form of a table of

genes and their read counts or FPKM or RPKM values. One of the important purposes that it serves is the comparison of abundance between two different (or among several different) tissue samples or organisms or experimental conditions. This is known as *differential expression analysis* and it involves various types of statistical methods. Some kind of *normalization* of the raw counts is necessary in order to create a "level playing field" for comparison before such statistical methods are applied, because of possible differences in read numbers between two libraries and/or between two sequencing runs on different days using different batches of reagents and/or between two organisms due to the inherent differences in their transcriptome compositions. That is why the concept of RPKM [301] was originally introduced. Based on transcript lengths and the sequencing depth, this RPKM can compare the expression levels across different genes and samples. The main challenge in using RPKM to estimate transcript abundance is the fact that mapped reads are frequently shared by multiple isoforms. The abundance models that used raw counts before the introduction of the RPKM concept had to assume that each transcript had a single isoform and reads were uniquely mappable to transcripts. As it became clear that this naiive assumption was far from reality, some *ad hoc* approaches were put forward, such as the "rescue" method (i.e. allocating fractions of the reads in proportion to the coverage of uniquely mapped reads) or that of allocating fractions of multi-reads (i.e. reads that map to several places in the reference genome) equally to the target transcript isoforms.

Some of the first departures from such simplistic, *ad hoc* approaches are seen in [497], [204], etc. [497] proposed an expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. [204] took a model-based approach assuming that the number of reads coming from an exon of a certain length followed a Poisson distribution with its mean being a normalized function of the exon length. Maximum likelihood estimates of the relative abundances of different transcripts were found using a concave optimization algorithm. The next type of statistical models was necessitated by the advent of paired-end sequencing and are generally known as *insert length* models. The lengths of the fragments used in the sequencing process as well as transcript lengths play an important role in them. In paired-end sequencing, reads correspond to both ends of the sequenced fragments. In this case, information about insert-size must be utilized to ensure that the two exons truly form a junction, as opposed to the possibility that the two exons are merely in the same transcript but something else is between them. In other words, when the reads in a pair map upstream and downstream of an alternatively spliced exon, isoforms will be indicative of different intervening insert lengths and this information should be utilized to maker the abundance estimates more accurate. The insert-size distribution depends on the RNA-seq library. Typically, the average insert-size is used for each read-pair and if the fragment-size variability is large in the library, the estimates of the insert-size for any particular read-pair will also have high variability (i.e. cannot be very

accurate). [440] introduced the earliest insert-length model by extending [204] approach to paired-end reads and made their algorithm available through the widely used software package *Cufflinks*. This was followed by other such models. For example, [225] came up with a Bayesian algorithm (*MISO*) for RNA-seq data and in it, they used an insert-length distribution that was based on the implied lengths of read-pairs that mapped to large intron-less regions (e.g. the 3′ untranslated regions). [373] is an excellent review of this approach and these models. An insert-length model tacitly assumes that the filtering of reads is based on their lengths and independent of their nucleotide sequences. Conditional on insert lengths, transcripts are assumed to be sampled from a uniform distribution. Using the relative probability of observing the given insert-length as a weight, the read-pairs are assigned (using a weighted probabilistic assignment scheme) to isoforms consistent with both individual reads. We conclude this subsection by pointing the reader to the algorithm developed by [108], called *IsoInfer/IsoLasso*, to deal with the problems of alignment, isoform finding and abundance estimation. First it computes a large set of possible isoforms and then weeds out many of them using the Lasso method to select a best subset.

5.6.6 Normalization methods

As was mentioned in the previous subsection, we often observe many artifacts and biases that adversely affect the quantification of expression levels via abundance estimation. So normalization is a very important step in RNA-seq data analysis. The purpose of normalization is to identify and remove systematic technical differences (i.e. technical biases) among samples that are often there. Among the plethora of normalization algorithms developed for RNA-seq data are total read count normalization (*RC*), upper quartile normalization *UQ*), relative log expression (*RLE*), trimmed mean of M-value normalization (*TMM*), median normalization (*med*), quantile normalization (*Q*), *DESeq*, *RPKM* and *FPKM*, *RSEM* and *Sailfish*. Some of these are global normalization procedures while others are not. In a global normalization method, only a single factor is used to scale the read-counts for all the genes from each sample.

Let n_{gj} be the observed read-count for gene g in the j^{th} sample, N be the number of samples and G be the number of genes. Let $D_j = \sum_{g=1}^{G} n_{gj}$ denote the total number of read-counts for the j^{th} sample. Also, let C_j be the normalization factor used for the j^{th} sample. In *total read-count normalization*, the assumption is that the read-counts are proportional to the expression levels of genes and to the sequencing depth. Since we want to ensure that the scaled total number of reads in each sample is the same (call that number K), the C_j's must be chosen in such a way that $C_j D_j = K$ for $j = 1, \ldots, N$. So it is clear that C_j must be K/D_j. As a result, the normalized read-count for gene g in sample j is $C_j n_{gj} = K n_{gj}/D_j$. Typically, $K = 10^6$ is used.

Clearly, the above method does not take into account the distribution of the read-counts for each sample. Bullard et al. (2010) [37] proposed a

normalization method that tries to match the read-count distributions across samples. Let $Q_{j(p)}$ be the upper $100p^{th}$ percentile of the read-counts in the j^{th} sample. The *upper quantile normalization* method seeks to ensure that $C_j D_j Q_{j(p)} = (\Pi_{i=1}^{N} D_i Q_{i(p)})^{1/N}$ for $j = 1, \ldots, N$. From this, it is easy to find the normalization factor C_j for the j^{th} sample. A commonly used value of p is 0.75.

In a differential expression study, a common experience is to find that a vast majority of genes are not differentially expressed. If a gene g is not differentially expressed between the j^{th} and the j'^{th} samples, then the ratio $n_{gj}/n_{gj'}$ of the counts of that particular gene g in those two samples would be expected to be equal to $C_j/C_{j'}$. Often it is observed that a few highly differentially expressed genes dominate in their influence on the total read-counts. When that is the case, using the total read-count D_j of the j^{th} sample to compute that sample's normalization factor C_j (as was done in the two methods described above) runs the risk of inducing bias into the estimation of read-counts for other genes in that sample. It is better to artificially create a *pseudo-reference sample* and assume that the expression level for each gene g in this artificial sample is equal to the geometric mean $(\Pi_{i=1}^{N} n_{gi})^{1/N}$. We will then have, for a particular gene g,

$$\frac{n_{gj}}{(\Pi_{i=1}^{N} n_{gi})^{1/N}} = \frac{C_j}{C_{Pse}}$$

where C_{Pse} denotes the normalization factor for the pseudo-reference sample. If we take C_{Pse} to be equal to 1, then C_j ends up being equal to the left-hand side of the above equation. Now, this was for a particular gene g. If we take the median of the left-hand side of the above equation over all genes (call it M_j), and define C_j^* as $C_j^* = (\Pi_{i=1}^{N} M_i)^{1/N}/M_j$, we can use C_j^* as the normalization factor for the j^{th} sample (clearly the C_j^*'s multiply to 1). This is called *relative log expression normalization*, because $(\Pi_{i=1}^{N} M_i)^{1/N} = \frac{1}{N}\sum_{i=1}^{N} log(M_i)$ and the geometric mean appearing in the expression level for the pseudo-reference sample can also similarly be expressed as an average of log-values.

It has been mentioned above that the total read-count of a sample is often dominated by a few highly expressed genes. In the RLE method, we dealt with this reality by creating a pseudo-reference sample. Another way out would be to remove genes from the upper and the lower ends of the expression levels. This idea leads to the *trimmed mean of M-value* normalization method. First, we define two quantities – the log fold-change and the absolute intensity. The log fold-change for gene g in the j^{th} sample compared to a reference sample *ref* is defined as

$$M_{g,ref}(j) = log_2 \left\{ \left(\frac{n_{gj}}{D_j} \right) / \left(\frac{n_{g,ref}}{D_{ref}} \right) \right\}$$

where $n_{g,ref}$ and D_{ref} denote the read-count for gene g in the reference sample and the total read-count for the reference sample respectively. The absolute intensity of gene g in the j^{th} sample is defined as

$$A_{g,ref}(j) = \frac{1}{2} \left\{ log_2 \left(\frac{n_{gj} n_{g,ref}}{D_j D_{ref}} \right) \right\}$$

Suppose we trim off a certain percentage (say, $100q\%$) of the ordered $M_{g,ref}$ values from the top end and the bottom end, and also trim off a certain percentage (say, $100q^*\%$) of the ordered $A_{g,ref}$ values from the two ends. Let G^* be the set of genes that still remain after the trimming. Using those, the normalization statistic will be defined as

$$TMM_{j,ref} = \frac{\sum_{g \in G^*} W_{g,ref}(j) M_{g,ref}(j)}{\sum_{g \in G^*} W_{g,ref}(j)}$$

where the weight $W_{g,ref}(j)$ is the inverse of the variance of the $M_{g,ref}(j)$ values for $g \in G^*$. Since the variance of the $M_{g,ref}(j)$ values can be shown to be approximately equal to $\{(D_j - n_{gj})/D_j n_{gj} + (D_{ref} - n_{g,ref})/D_{ref} n_{g,ref}\}$, the reciprocal of this expression is the weight used in the formula for $TMM_{j,ref}$. Finally, letting $B_j = 2^{TMM_{j,ref}}$, we define the normalization factor for the j^{th} sample as

$$C_j = \frac{exp\left(\frac{1}{N}\sum_{i=1}^{N} log_e(B_j)\right)}{B_j}$$

and clearly the C_j's multiply to 1.

We conclude this subsection with a discussion of RPKM. As its name suggests, the RPKM approach normalizes for the total transcript length and the number of reads. Let N be the total number of mappable reads in an experiment, L be the sum of the numbers of base-pairs in the exons of the genes and C_{ex} be the number of mappable reads that were mapped to those exons. Then RPKM is defined as

$$RPKM = C_{ex}/(\frac{N}{10^6})(\frac{L}{10^3}).$$

5.6.7 Differential expression analysis

An enormous volume of research has been conducted in the last decade to devise suitable statistical methods for detecting differential expression of genes between tissues, organisms or experimental conditions. Here we provide a glimpse of some of it. Lorenz et al. (2014) is an excellent reference for further details.

5.6.7.1 Binomial and Poisson-based approaches

Suppose that we are quantifying the expressions of G genes in L populations, and X_{ijkg} denotes the number of reads that mapped to gene g in replicate