Chapter 7

New Advances in NGS Technologies

Edo D'Agaro

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/66924

Abstract

In the next-generation sequencing (NGS) methods, a DNA molecule of an individual is broken down into many small fragments to make up the so-called sequencing library. These small fragments serve as a template for the synthesis of numerous complementary fragments (called reads). Every small piece of the original DNA is copied many times in a variable number of reads. Depending on the desired accuracy level, it is possible to set the system to achieve a certain level of coverage, i.e., a number of reads per fragment. A level of 30X coverage is already sufficient for the routine diagnosis of most of the Mendelian diseases. All the sequences are then transferred into a computer and aligned with a reference sequence available in the international databases. By this way, all sequences of reads can be recomposed as a fine puzzle to obtain the sequence of a single gene or whole genome. The NGS machines, available today, are very flexible devices. In fact, an NGS sequencer can be used for different types of applications: (1) whole-genome sequencing (WGS): analysis of the entire genome of an individual; (2) whole exome sequencing (WES): analysis of the entire coding genes of an individual; (3) targeted sequencing: analysis of a set of genes or a single gene; (4) transcriptome analysis: analysis of all the RNA produced by specific cells.

Keywords: genomics, NGS, genetic screening, animal diseases, bioinfomatics

1. Introduction

The next-generation sequencing (NGS) technologies, with the capability to produce, at a low price, millions of DNA sequences per analysis, have changed the way to make the genetic analysis in the veterinary and animal fields. In the past few years, the observed reduction of cost and time required to obtain the genomic information (whole genome, whole exome or targeted gene sequencing) have had several repercussions in various disciplines such as animal selection, animal research, genetic disease diagnosis and control, cancer research, and metagenomic studies. The focus of this chapter is to give an updated overview of the next-generation

techniques, analyzing the general framework, the bioinformatic analysis pipeline, present applications in the genetic disease detection, problems and future possible developments.

2. Patterns of transmission of genetic diseases

2.1. The hereditary transmission

The hereditary traits are transmitted from the DNA (deoxyribonucleic acid) contained in the chromosomes and in the mitochondria. The gene is formed by regions that encode proteins, the exons, and segments that are not translated, the introns. Genetic diseases are due to inherited mutations, which are changes in DNA sequence and transmitted in familiars in different ways. Several diseases are inherited in a simple way and follow the Mendel's law. These are monogenic or Mendelian diseases. Monogenic diseases can be classified according to their mechanisms of inheritance: autosomal dominant and recessive, X-linked and mitochondrial diseases.

2.2. Autosomal dominant

In the case of dominant disease, the most common genotype is the heterozygous genotype. The disease is transmitted to an average of 50% of the offspring regardless of their sex. The most common cross is between a heterozygous affected (Aa) and a homozygous unaffected (aa). A typical pedigree of an autosomal dominant disease has the following distinguish features:

(a) vertical segregation, with affected individuals in all generations, both males and females;

(b) an equal number of affected males and females;

(c) each affected individual have at least one affected parent.

An autosomal dominant disease may also arise from a new mutation. A risk factor of a new mutation is the parental age. The frequency of new mutations increases proportionally with the age of the parents. One feature of the dominant autosomal diseases is to show a variable expression among the affected individuals, even in the same family. Some autosomal dominant diseases exhibit the phenomenon of anticipation, which consists of an earlier onset and an aggravation of symptoms with the passing generations. Another characteristic of the autosomal dominant diseases is the penetrance, which is the percentage of animals expressing the phenotype over the total number of individuals which carry the mutation. Other genes or the environment can interfere with an occurrence of a disease resulting in a lack of penetrance. In this case, the subject that carries the dominance mutation will not be affected but, 50% of the offspring will manifest the disease with varying degrees.

2.3. Autosomal recessive diseases

The presence of two recessive alleles at the same locus causes the disorders. The most frequent cross is between heterozygous individuals for the mutated allele (healthy carriers of an autosomal recessive disease). A typical pedigree of an autosomal recessive disease has the following distinguish features:

(a) unaffected parents could have affected offspring;

(b) equal number of males and females;

(c) all offspring are affected when parents are affected.

The heterozygous parents will show a risk of 25% to get affected offspring. Inbreeding is the most important risk factor for the occurrence of an autosomal recessive disease.

2.4. Disease linked to the chromosome X

The X chromosome contains hundreds of genes. The dominant or recessive effect is determined in the females because they have two chromosomes. Males have only one X chromosome and therefore they express the recessive mutations (called hemizygous). In this case, the recessive trait is always expressed. If the alleles are different, a female is heterozygous, i.e., a carrier. A typical pedigree of an x-linked dominant disease has the following distinguish features:

(a) each generation usually has an affected individual;

(b) all daughters of an affected male are affected;

(c) all males and females of a heterozygous are affected.

The family tree study shows a zig-zag or diagonal pattern, with a variable number of affected males in the following generations.

A typical pedigree of an x-linked recessive disease has the following distinguish features:

(a) all males of an affected female are affected;.

(b) affected males never transmit the trait to their male offspring;

(c) unaffected parents may have affected offspring.

2.5. Multifactorial diseases

Some genetic diseases show a polygenic and/or multifactorial inheritance. The multifactorial traits can be continuous and measurable, such as the weight, height, blood pressure, or discontinuous, such as congenital heart disease, diabetes, osteoporosis, and cardiovascular disease.

2.6. Mitochondrial disease

The defects in the mitochondrial DNA show a maternal inheritance because mitochondria are transmitted to the offspring by the egg cell. Characteristics of mitochondrial diseases are incomplete penetrance, variable expressivity, and pleiotropy. The phenotypic expression of mitochondrial disorders depends on the ratio of normal and mutated mtDNA present in the cells of various tissues. The organ impairment and its severity depend on the percentage of mutated mitochondria, which is variable in different tissues. Mitochondria are distributed in cells throughout the body and are therefore responsible for pathologies with impairment of different organs and apparatuses.

3. Mutations

A mutation is defined as a stable and heritable modification of the genetic material. The mutation can occur at different levels and change the attitude of the entire genome showing a variation in the number of chromosomes (genomic mutation), a variation in a single chromosome (mutation chromosome) or, it may involve a single gene (gene mutation). Mutations can be spontaneous or induced. The most common causes of mutation can be physical, chemical, or biological factors (called mutagens). The gene mutation can be lethal to the cell or, when the gene is expressed, give rise to senseless, inactive, or less active proteins or, although much rarely, resulted in products with a greater or different activity. Most of the mutations are harmful and only a small percentage appears to be advantageous. The carrier of an unfavorable gene may die before to reproduce (lethal disease). Natural selection tends to reduce mutated genes. Sometimes, however, due to the changing environment, insignificant mutations can be expressed or even turn out to be advantageous and to be favored by natural selection. Genomic and chromosomal mutations, when they are compatible with life, give rise to a more complex phenotypic pattern, because they involve a large number of genes. On the contrary, gene mutations affect the sequence of a single gene and therefore they have an effect only on the functionality of a single protein. A *transition mutation* is one substitution of a purine with a purine or pyrimidine with a pyrimidine while a *transversion mutation* is a substitution between a purine and a pyrimidine. At phenotypic level, effects of mutations can vary from the complete neutrality to a range of events whose gravity depends on the interaction with other genes and with the environment. At the molecular level, a gene mutation consists in a variation of the sequence of a nucleotide of the gene. The normal DNA sequence may change as a result of insertion or deletion of one or more bases. A gene mutation, if it concerns a single point of DNA, i.e., one or a few bases, it is called a point mutation or SNP. The first element to be considered, in assessing the effects of a gene mutation, is what part of the gene has changed. The sequence of a gene is made, in addition to a promoter, of a number of exons interspersed with introns. As is known, the entire sequence of the gene is transcribed, but the intronic fragments are removed during the process of splicing. The exon region of a gene is characterized by the presence of a coding region, which contains the information for the synthesis of the coded protein, and two noncoding regions that are located at the beginning and end of the mRNA molecule (5'UTR and 3'UTR that are transcribed but not translated). Mutations that modify the coding region may have important effects on the translation of the proteins, mutations at the junction between intron and exon can interfere with the recognition of splice sites causing the production of modified transcripts while mutations in the untranslated regions (UTRs) and the introns are often neutral or in some cases may alter the regulatory functions of these regions. The mutations that show (most frequently) phenotypic effects are those affecting the coding regions. Therefore, in this chapter, we will focus on them. In the case of a *silent mutation*, the replacement of a base causes a change in the codon without a change of the amino acid (the genetic code is degenerated and the same amino acid can be coded by more than one codon). For example, if a mutation changes the codon CUA in CUG, as both codons code for the same amino acid leucine, the amino acid does not change. For this reason, a silent mutation has no effect at the level of the protein and thus on the phenotype. In the case of a *missense mutation*, the substitution of a base of a codon with a

http://dx.doi.org/10.5772/66924

different base causes the change in the codon function. For example, if the codon AGC, which codes for the amino acid serine, becomes AGA, encoding for the amino acid arginine, during the protein synthesis, an arginine will be inserted instead of a serine. The substitution of an amino acid may therefore change the protein function. Based on the amino acid substitution, the missense mutations are divided into: (a) conservative substitutions: the new amino acid has similar characteristics to the replaced one (neutral mutation); (b) nonconservative substitutions: the new amino acid has different characteristics [e.g. GAG (Glu) changing to GTG (Val)]. The consequences of a missense mutation are more or less serious depending on the physical-chemical characteristics of the two amino acids involved. Characteristics of the amino acids present in the primary structure determine the secondary and tertiary structures of the proteins by means of ionic and hydrophobic interactions. In turn, the correct folding of the protein chain determines its biological functionality. The maintenance of the protein functions depends on the amino acid position and its structural role. In the case of sickle cell disease, due to a missense mutation in the gene of the beta-globin, a polar amino acid (water-soluble, glutamic acid) is replaced by an insoluble apolar amino acid valine. This causes the production of a beta-globin chain resulting in an altered hemoglobin solubility, which in turn causes the characteristic shape of the sickle red blood cells. A gene mutation can change a codon sense in a noncodon sense (*nonsense mutation*). For example, if the codon AAG, coding for lysine, becomes UAG, which is a stop codon, then, the protein synthesis will end prematurely. The synthesized protein is therefore incomplete and, in most cases, it will not be functional. An elongation mutation arises when, a nonsense triplet present in the wild gene, following a base substitution, becomes a triplet "sense," encoding for an amino acid. In this case, the stop triplet is eliminated and the mRNA translation will continue with the formation of a longer protein. Deletions and insertions of bases in the coding region of a gene cause a slippage of the reading code system (*frameshift mutations*). The inclusion or loss of one or more bases causes a slip, resulting in loss of sense of the whole protein. All codons downstream of the mutation change and from that point they are then incorporated into the protein, giving rise to an abnormal protein and almost always nonfunctional. A deletion of an entire portion of a gene can severely alter the three-dimensional protein structure, the chemical properties and functionality. Even in this case, the deletion may lead to a slip of reading (frameshift), determining the loss of a part of the protein. A mutation in a gene can also determine a quantitative variation (not qualitative) of the protein. This occurs when the mutation falls in a regulatory region (i.e., the sequence of the promoter) and modifies the transcription of the gene. A point mutation (by substitution or by insertion/deletion of bases), both in coding and noncoding regions, can give rise to an RFLP (polymorphism of length of the restriction fragments) and to an SNP (single nucleotide polymorphism).

4. DNA sequencing

Over the last 60 years, it has observed a significant increase in the knowledge of animal genomes and the genetic code, starting from the discovery of the structure of DNA in 1953 until the publication of the first draft of the human genome in 2001 [1, 2]. The Sanger sequencing [3], also known as sequencing of the first generation, was the method used to sequence

the genome within the project "Human Genome' bringing the entire genome sequence in 2003, after 13 years from the beginning of the project with a cost of $3 billion and the contribution of six different nations. In the recent years, several animal and plant genome sequencing projects have been carried out thanks to the implementation of both molecular biology and computer science innovations [4]. The evolution in parallel of these two sectors has enabled the advent of several next-generation sequencing (NGS) platforms. Nowadays, by means of this new technology, which is more efficient and economic compared to the previous methods, it is possible to sequence a new genome at the cost of less than one thousand euros in a very short time (1–2 days). Furthermore, the cost and time can be greatly reduced by analyzing only the coding regions (exomes). Despite the fact that the exomes are only 1% of the entire genome, by mean of this method, it is possible to identify the 85% of the monogenic diseases.

4.1. Last-generation sequencing (NGS) possible

All NGS methods are characterized by two important phases: a molecular biology step, which goes from the preparation of the sample to the sequencing phase, which allows to carry out, at the same time, more than one reaction and with less dexterity than the Sanger method, and a computerized stage for the analysis of the data. The first part of the process is divided into three steps: *sample preparation*, *amplification*, and *sequencing*. Sample preparation, which is the common passage for all platforms, is the fragmentation of DNA (dimensions vary from 100 to 800 base pairs (bp) in relation to the platform and the adapters used). The amplification is based on two methods: emulsion or solid PCR amplifications. In the emulsion PCR method, described for the first time by Tawfik Griffiths, the individual molecules of DNA are clonally amplified in micro-compartments consisting of mixtures of water and oil [5]. The adapters bound to individual DNA molecules, hybridize to complementary sequences coating the surface. Cycles of upcoming amplification allow to the formation of "clusters" of folded fragments clonally amplified. All the NGS platforms are characterized by the ability to sequence and massively parallel amplifying the DNA molecules in a clonal or single way. Unlike the Sanger method, where the fragments of different sizes from individual sequencing reactions were separated by electrophoresis, in the NGS technologies, sequencing is accomplished through the repetition of nucleotide extension cycles or oligonucleotide ligations. The principles of sequencing and image acquisition are the peculiar steps that characterize the different platforms on the market. To date, different methods of sequencing are known. The cyclic reversible termination (CRT) sequencing method uses reversible labeled nucleotides. Each sequencing cycle comprises: the incorporation of the nucleotide, the acquisition of the fluorescence, and the cutting of the nucleotide. The subsequent washing step allows the elimination of all nonincorporated nucleotides. At this point, the image for identifying the incorporated nucleotide is captured, followed by a cleavage step that removes the terminator group and the fluorophore. After the elimination, the polymerase can continue the reaction of extension and tie the second nucleotide. This process is used by two types of trading platforms: Illumina and Helicos, which differ in the templates of sequencing. While Illumina uses clonally amplified fragments on a solid surface, Helicos is currently the only commercial platform able to use single DNA molecules. In addition, the acquisition platform of Illumina uses a four-color (the four reversible nucleotides are labeled with a different fluorophore and

are dispensed at the same time in the sequencer), while in the Helicos platform, all nucleotides they are labeled with the same fluorophore and are dispensed into the sequencer in a determined hierarchical order.

4.2. Commercial systems for NGS

We describe here the main platforms used for the last-generation sequencing:

Roche 454 system was the first genome sequencer to be marketed in 2004. This system uses the sequencing by synthesis technology known as pyrosequencing. Initially, the 454 sequencing method used reads of 100–150 bp, producing about 200,000 reads, with an output of 20 Mb for each run. In 2008, a new sequencer the 454 GS FLX titanium was produced (reads of 700 bp, with an accuracy after filtering of 99.9%, with an output of 0.7 Gb for run in 24 hours). In 2009, the GS Junior and 454 GS FLX platforms were used to set up a new sequencer with an output of 14 Gb for run. Further developments have led to the production of the GS FLX +, which is able to sequence reads up to 1 kb. The high speed of analysis combined with the long reads is the positive characteristics of this platform. However, the cost of reagents remains a problem to be solved.

Ion Torrent. The method used by Ion Torrent Genome Machine (PGM) is very similar to the Roche 454 system. This platform, instead of using images to capture the incorporated nucleotide, detects the change of pH. The output per run is of 270 Mb with reads of 100–200 bp.

AB SOLiD system. The Applied Biosystems method was first marketed in 2006. This system uses the ligation sequencing method (in both directions) to ligate fluorescently labeled octomers to the DNA fragment. Initially, the length of the reads was only 35 bp and the output of 3 Gb for run. In 2010, the 5500x1 SOLiD platform was released (reads with length of 85 bp, precision of 99.99% and output 30 Gb per run). The main problem of this method is the low length of the reads.

Illumina GA/HiSeq system. In 2006, the Solexa company released the Genomic Analyzer (GA) and in 2007, the company was bought by Illumina. The system uses the sequencing by synthesis method (SBS) and the amplification bridge which is an alternative method of the PCR. At the beginning, the analyzer's output was of 1 Gb for run, after upgraded to 50 Gb. In 2010, the HiSeq 2000 was released with an output of 600 Gb for run and reads of 200 bp. The basic method is to generate a great number of polonies (generated DNA colonies) that are simultaneously sequenced in parallel reactions occurring on the surface of a flow cell. The cost of sequencing is lower compared to the competitors. **Table 1** presents the main NGS platforms.

4.2.1. Third-generation sequencing technologies.

Helicos single-molecule sequencing device: Heliscope method was presented for the first time in 2007. This platform uses a technique that individually analyzes the molecules achieving a greater accuracy. Using this system, you can obtain an output in the order of 28 Gb. However, the main disadvantage of this method remains the low capacity to correctly identify indels,

Platform	Type
Illumina	Sequencing by synthesis
454 Roche	Pyrosequencing
Ion Torrent	Ion semiconductor
ABI Solid	Ligation based sequencing (color spaced)
Oxford nanopore	Nanopore sequencing
Pacific biosciences	Single molecule real time (SMRT)

Table 1. List of sequencing platforms.

with a consequent increase in errors. Another problem is the length of the reads, which has never exceeded 50 bp.

Oxford Nanopore Miniontechnology. Most of the nanopore sequencing technologies are based on the transit of DNA molecules, after the application of an electric potential, through an array of protein nanopores located in a high-salt buffer. Single bases are identified according to the electric change. Several methods have been proposed based on the nanopore technique. Among these, one is produced by Oxford Nanopore (exploits the combination of three molecules), one based on MSpa (a protein). This method can be very useful for identifying infectious diseases and to analyze environmental samples (metagenomic analysis of bacteria, viruses and fungi) without a previous knowledge of the sample composition.

An additional advantage of the third-generation sequencing methods is the production of long reads (>than 10,000 pb and up to 100,000 bp) enabling to analyze insertions, deletions, and translocations [6–9]. Using these technologies and mapping methods, it is possible to study entire chromosome arms.

4.3. NGS data analysis

Although the various sequencing technologies use different methods, they all provide as output the FASTQ sequences. The FASTQ string is the type of information used in molecular biology to store genetic sequences and the related quality scores, namely the score that the algorithm assigns to the string and which is then used to choose the best match against the reference genome. At this point, the bioinformatics analysis is divided into three steps: *alignment*, which search of correspondences between the reads and the reference genome and the *variant calling* that attempts to separate the differences due to genetic mutations and instrumental errors made during the analysis and *filtration and annotation*, which attempt to align the reads to the reference genome.

4.3.1. Alignment

The alignment is the process by which you map short reads to a reference genome. It is a complex task, since the software must compare each reads in all of the reference DNA positions

[8, 10, 11]. It is a computationally challenging passage and wasteful in terms of time. The SAM (sequence alignment map) and BAM (binary alignment map) are the standard file formats for storing the data obtained using the next-generation technologies (NGS). There are many commercially available software, free or on sale, to perform this task. Most software use a method based on indices, which are very fast in the search for all alignment positions (without gaps) in the reference genome. Other algorithms, instead, allow the search for alignments with gaps. The various methods to solve the problem include the use of hash tables (e.g., MAQ, ELAND), algorithms based on Burrows-Wheeler transformation (e.g., BWA, Bowtie, SOAP2), an algorithm that uses a reversible compression method commuting the order of the characters without changing the values; genome-based hash (e.g., Novoalign, SOAP). Some software can take into account of gaps (e.g., BWA, Bowtie2) while others do not (e.g., MAQ, Bowtie).

4.3.2. Variant calling

After the alignment, the DNA sequence can be compared to the reference genome, identifying the possible changes. These variations can be due to genetic diseases or they can be only noise. The complexity of this subject lies in the difficulty to distinguish between the true variations and the sequencing errors [12]. The continuous development and improvement of NGS technologies has brought continuous advantages in this area, improving the quality of analysis [13, 14]. The main difficulty in this kind of analysis is the presence of indels, i.e., phenomena of inserts (insertion) or cancelation (deletion) of DNA segments. In fact, indels are the major cause of false-positives. The number of false-positives increases when using algorithms that do not take it into account of gaps. Another source of errors arises during the preparation of sequences and the PCR analysis. In order to reduce the errors, it is advisable to increase the sensitivity of the PCR analysis, use updated alignment software and a big reference database for getting wider comparisons.

4.3.3. Filtering and annotation

After the alignment and variant calling steps, a list of thousands of potential differences is generated between the genome under study and the reference genome. The next step is then to determine which of these variations are due to sequencing errors. The use of specific filters allows to remove variants that do not follow the models under study and to make annotations. The comparison of reads against a genetic reference tree will find all elements of known function. In addition to the filtration step, the annotation process provides another tool to select and restrict the test sample, applying specific functional models [15]. The low cost of these new instruments is leading to the discovery of a great number of genetic variants and identifying various diseases [16]. For example, it has been observed that about 1300 positions of a gene or a sequence within a chromosome are associated with 200 diseases [17].

4.3.4. Definitions of DNA sequence variants

In general, there are three principal categories of variants:

(a) Causative variants

Variants showing a clear pathogenetic role and associated with a phenotype disease.

(b) Unrelated variants

The most commonly used definition to define this class of variants is "incidental findings."

(c) Variants with undefined functional and clinical effects (variants of uncertain significance, VUS).

Table 2 reports some useful Internet sites for the genetic variant discovery analysis.

SAMTools	http:// mamtools.sourceforce.net
	http//htslib.org
GATK	https://www.broadinstitute.org/gatk
Platypus	http://www.well.eox.ac/platypus
Freebayes	http://github.com/ekg/freebayes
BreakDancer	http://breakdancer.souceforge.net
Dindel	https://sanger.ac.uk/resources/software/dindel

Table 2. List of variant discovery resources (Internet sites).

4.3.5. Example of a variant calling analysis using the galaxy platform (www.galaxy.org)

The protocol in a nutshell:

- Inspect the reads;
- Raw data cleanup/quality trimming;
- Align reads to a reference genome;
- Mark duplicates;
- Combine samples into a single file;
- Realign reads around insertions and deletions (indels);
- Correct inaccurate base qualities (optional);
- Make variant calls, SNVs, indels;
- Annotate the results.

(a) Read quality control

Steps involved and suggested tools:

NGS: QC and manipulation → FastQC

Command line: fastqc

http://dx.doi.org/10.5772/66924

Some of the important outputs of FastQC for our purposes are:

- Quality encoding type: important for quality trimming software;
- Total number of reads: gives you an idea of coverage;
- Presence of highly recurring k-mers;

(b) Quality trimming/cleanup of read files

NGS: QC and manipulation → trimmomatic

Suggested trimmomatic functions:

- Adapter trimming;
- Sliding window trimming;
- Trailing bases quality trimming;
- Leading bases quality trimming;
- Minimum read length.

(c) Genome alignment

Suggested tools:

NGS: Mapping → BWA

NGS: Picard → MarkDuplicates

NGS: SAM Tools → Merge BAM Files

NGS: GATK Tools → Indel Realigner

NGS: GATK Tools → Count Covariates

NGS: GATK Tools → Table Recalibration

(d) Variant calling

NGS: GATK Tools → Unified Genotyper

Possible alternative software:

Varscan http://varscan.sourceforge.net/

(e) Annotation

NGS: GATK Tools → Variant Annotation

Figure 1 shows an example of a variant analysis workflow.

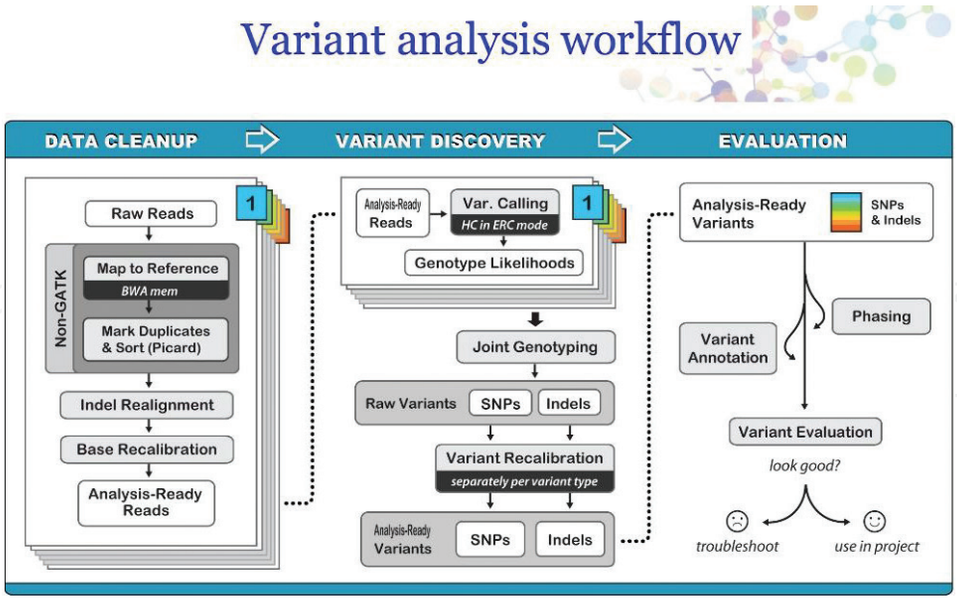

Figure 1. Example of a variant analysis workflow.

4.3.6. Bioinformatic analysis of the NGS data

Genetic testing based on NGS analysis can be divided depending on the width of the portion of the analyzed genome:

(a) sequencing the entire genome (whole genome sequencing; WGS);

(b) sequencing the entire exome (whole exome sequencing; WES) [18];

(c) targeted gene sequencing.

The choice of the different method will take into account several factors:

- *Costs*: since the introduction of the NGS practice, the use of specific panels has been privileged for economic reasons. Nowadays, the cost of the different methods has gradually narrowed and today this aspect is not, in most of the cases, the primary choice element. WES may be, in fact, less expensive of sequencing a panel of genes and, similarly, the cost of WGS is getting closer to that of WES;
- *Purpose*: in general, the NGS techniques find the best applications in the case of high heterogeneity genetic diseases or for the Mendelian diseases in which the gene is not yet been identified;
- *Diagnosis:* direct clinical applications are more and more frequent, especially for the congenital diseases;
- *Sensitivity*: the higher the number of reads of a specific region, the greater will be the sensibility for that particular stretch of DNA. In principle, in the analysis of specific panels, the smaller is the portion of the genome analyzed, the greater is the coverage and then the

sensitivity. This is the reason why, for a mosaic alteration, the analysis of known genes shows a greater sensitivity than the WES analysis;

- *Data storage*: in relation to the increasing number of sequenced genomes and data to be analyzed, new bioinformatic platforms for storage and data analysis are needed.

The amount of data generated by NGS platforms is on the order of terabytes (Tb). Software used for data analysis differ according to the NGS technology but, all follow a "pipeline" system of data analysis converting the luminescence images or fluorescence data to "reads" sequences. In this process, defined, "base calling," a quality score value is set for each nucleotide, indicating the probability of an error associated with it. The "quality score" is an important value for selecting the reads during the analysis process. If they fall below a specific level, they are eliminated improving the accuracy of the alignment process. In order to get an adequate efficiency of the alignment, this value should not be below 30. Another limitation for the alignment process is represented by the repeated sequences. The error rates associated with the NGS technologies seem to be higher than the traditional Sanger method. However, the accuracy of the NGS sequencing is increased by means of a repeated and massive reading of each gene fragment, which determines the "coverage" of the genome. The latter parameter represents an essential value in the NGS analysis (value ranging between 20 and 50 times, in relation to the platform used) determining the presence of false negatives (for heterozygous individuals) in the detection of nucleotide variants. At the end of the variant annotation process, the number of identified variants varies in relation to the application of various filters, used to reduce the number of candidates. The filters most commonly used are based on:

- inheritance pattern (autosomal/X-linked, dominant/recessive);
- sharing equal variations in well-characterized families;
- removal of variants already known through the use of public databases (dbSNP);
- based on the potential effect of the change (no sense, missense, changes in splicing sites or insertions and deletions that modify the reading frame);
- prediction of the functional effects of changes through the use of *in silico* bioinformatic tools (SIFT, POLYPHEN, and ANNOVAR software).

4.3.6.1. Advantages of using the NGS technology

Although, in previous decades, the development of the Sanger method has brought many findings in chemistry, automation and miniaturization of the process, only the advent of NGS technologies has led to a significant improvement of quality and speed of the genome analysis [19]. The release of new sequencing platforms and the reduction of costs associated with the NGS technology is the consequence of three factors:

(1) Thousands/millions of sequencing reactions can be conducted in parallel exceeding the limit of 1–96 possible reactions obtainable with a traditional sequencer;

(2) Cloning or amplification of the DNA fragment is, in the new technologies, not necessary or completely automated within the platforms;

(3) Ability to detect the minor allele with high accuracy. This in term allows a better identification of a variant in mosaic sample or heterozygous deletions. The number of times that a DNA fragment is amplified and sequenced is proportional to the abundance of such segment in the original sample.

4.3.6.2. Limits of NGS technology

The NGS technology has, however, some limitations related mostly to the magnitude of data products. In fact, in the NGS results we can observe false positives and false negatives.

False positives can result from:

- an incorrect alignment with the reference sequence. You can overcome this problem using different alignment software;
- systematic sequencing errors. This type of error can be identified in all the samples and removed from the final list [20];
- technical errors of the sequencer. For example, with the pyrosequencing method, a common error was observed using homopolymers longer using the pyrosequencing method, was observed when reading homopolymers longer than 5–6 bases;

False negatives instead come from:

- low coverage;
- low coverage in regions of interest;
- alignment of repeated regions.

The reduction of errors can be obtained increasing the coverage and the quality of DNA fragmentation (fragments of greater size). The implementation of the "paired-end" sequencing, that is able to sequence fragments of a greater length by both ends, allows the analysis of fragments of 5–10 kb. The use of the Sanger method is, however, required at the end of the analysis because the results, obtained using the NGS methods, require to be validated.

4.3.7. Applications of NGS technology

The genome can be assessed globally, only in the coding regions or in target regions. The main interest of the veterinary genetic sector is to study the coding regions as the greater number of diseases is caused by mutations or exonic splicing which alter the correct amino acid sequence of the protein. Indeed, exomes, while constituting only 1% of the genome, are the location of 85% of pathogenetic mutations. The number of mutations known to be associated with a genetic disease exceeds 110,000 variants, in most of 3700 different genes. It was estimated that only half of the Mendelian diseases have a known genetic basis. For these reasons, the scientific research focused mostly on exomes for the identification of new genes [21]. The approach, commonly used in the past, was to identify the transmitted loci associated with the phenotype by means of the segregation or linkage analysis, identifying shared genes of affected individuals in large size families. The advent of WES technology has

triggered a rapid growth of the sector (identification of new disease), having the advantage of requiring a limited number of samples. The first successful application of this method has led to the identification of the gene DHODH as causative of Miller syndrome. Since 2010, a large amount of studies have identified new disease-causing mutations, through the exome sequencing method using a reduced number of affected individuals, in various types of diseases including cases of neuropathy, poikiloderma associated with neutropenia, familial exudative vitreoretinopathy, immune disorders, and tumor predisposition. The use of WES allowed to identify some causative mutations also in diseases with high phenotypic heterogeneity where the traditional linkage analysis approach is more difficult. A new area of the development of NGS technology is linked to the identification of new biomarkers or pharmagenomics with the development of personalized therapies. The impact of next-generation sequencing technology on genomics has in turn led to a new revolution in the genetic field changing the nature of genetic trials. The production of a large number of low cost NGS platforms is useful for many applications. These include: new variants discovered by targeted resequencing regions of interest or whole genomes; *de novo* assemblies of bacterial genomes and lower eukaryotes; cataloging of cell, tissues and organism transcriptomes (RNA-Seq); profiling of genome-wide epigenetic marks and chromatin structure using seq-based methods (Chip-seq and Methyl-seq); classification of species and/or discovery of the gene using metagenomic studies. NGS technologies can also accelerate exploration of the natural world [22, 23]. Despite a dramatic increase in the number of complete genome sequences available in public databases, most of the biological diversity, in our world, remains to be explored. Nowadays, *de novo* assembly of NGS data requires the development of new software tools that can overcome the technical limitations of these technologies. In fact, the main limitation is a rapid deterioration in the quality of assembly as the length of reading decreases.

4.3.8. De novo assembly

De novo genome assembly is often compared to solving a great puzzle without knowing the picture that we are trying to rebuild [7, 21, 24]. Mathematically, the issue *de novo* assembly is difficult regardless of the method of sequencing. During the *de novo* assembly process, a high number of repeated segments in the genome may cause several errors. The assembler tool should guess the right genome, starting from a great number of alternative options (the number of attempts increases exponentially with a high pattern of repetitions in the genome) [25]. As the technology has evolved, new methods for assembling genomes have continuously changed [26]. Genome sequencers have never been able to read more than a relatively short stretch of DNA at once, with read lengths gradually increasing over time.

Assembly quality: High coverage is necessary to sequence polymorphic alleles within diploid or polyploidy genomes. However, using shorter reads, the coverage should be increased in order to balance the low connectivity of the system and to obtain an optimal assembly. However, some times, a poor assembly process cannot be improved by a higher coverage. In fact, in the case of a high number of long repeated sequences in the genome, a high coverage will never fully compensate for the increasing errors and gaps which are produced during the editing phase. Using paired end technology gaps can be crossed and eliminated [5].

Assembly methods aim to create the most comprehensive reconstruction as possible without introducing errors. The central challenge of assembling the genome is to solve the problem of repetitive sequences. If the DNA reads are random, than the expected number of cases of each sequence will decrease exponentially and the number of repeats in the genome is reduced. However, several genomes may share highly repeated structures which do not allow an easy assembly process [27].

Scaffolding: The phase of scaffolding focuses on repeats to fix linking contig initials in data-driven scaffolds. A scaffold is a collection of contig linked by pair partners, where gaps between contigs between contig might be repeats, in which the gap can in theory be filled or outright gaps that original sequencing project does not capture [6]. If the distances are long enough, the assembler is able to connect contig in nearly all repetitions. Assembler tools may vary in the method how they call the contigs. The great majority of them are based on a combination of two factors: contig length and a number of reads. A contig containing too many reads is called as a repetition. High-copy-number patterns are easy to be identified. On the contrary, the identification of two copies is more difficult) [25]. If the contigs are overlapping in a scaffold, the assembler can merge at this point. Otherwise, the assembler will record a gap inside the scaffold.

Example of de novo assembly using the galaxy platform (www.galaxy.org)

(a) Read quality control

Steps involved and suggested tools:

NGS: QC and manipulation → FastQC: comprehensive QC

Command line: fastqc

Some of the important outputs of FastQC for our purposes are:

- Read length: Will be important in setting maximum k-mer size value for assembly;
- Quality encoding type: Important for quality trimming software;
- % GC: High GC organisms don't tend to assemble well and may have an uneven read coverage distribution;
- Total number of reads: Gives you an idea of coverage;
- Dips in quality near the beginning, middle, or end of the reads: Determines possible trimming/cleanup methods and parameters and may indicate technical problems with the sequencing process/machine run;
- Presence of highly recurring k-mers: May point to contamination of reads with barcodes, adapter sequences, etc.;
- Presence of large numbers of Ns in reads: May point to poor quality sequencing run. You need to trim these reads to remove Ns.

http://dx.doi.org/10.5772/66924

(b) Quality trimming/cleanup of read files

Steps involved and suggested tools:

NGS: QC and manipulation > trimmomatic

Command lines:

- Adapter trimming
- Sliding window trimming
- Trailing bases quality trimming
- Minimum read length

(c) Assembly

Steps involved and suggested tools:

NGS-Assembly → Velvet Optimizer

Possible alternative software:

Spades

SOAP-denovo

MIRA

ALLPATHS

Possible tools for improving your assemblies:

QUAST – http://bioinf.spbau.ru/quast

Mauve

InGAP-SV – https://sites.google.com/site/nextgengenomics/ingap

Semi-automated gap fillers:

Gap filler – http://www.baseclear.com/landingpages/basetools-a-wide-range-of-bioinformatics-solutions/gapfiller/

IMAGE – http://sourceforge.net/apps/mediawiki/image2/

Genome visualizers and editors:

Artemis – http://www.sanger.ac.uk/resources/software/artemis/

IGV – http://www.broadinstitute.org/igv/

Geneious – http://www.geneious.com/

CLC BioWorkbench – http://www.clcbio.com/products/clc-genomics-workbench/

Automated and semi-automated annotation tools:

Prokka – https://github.com/tseemann/prokka

RAST – http://www.nmpdr.org/FIG/wiki/view.cgi/FIG/RapidAnnotationServer

JCVI – http://www.jcvi.org/cms/research/projects/annotation-service/.

Figure 2 shows an example of a genome assembly workflow.

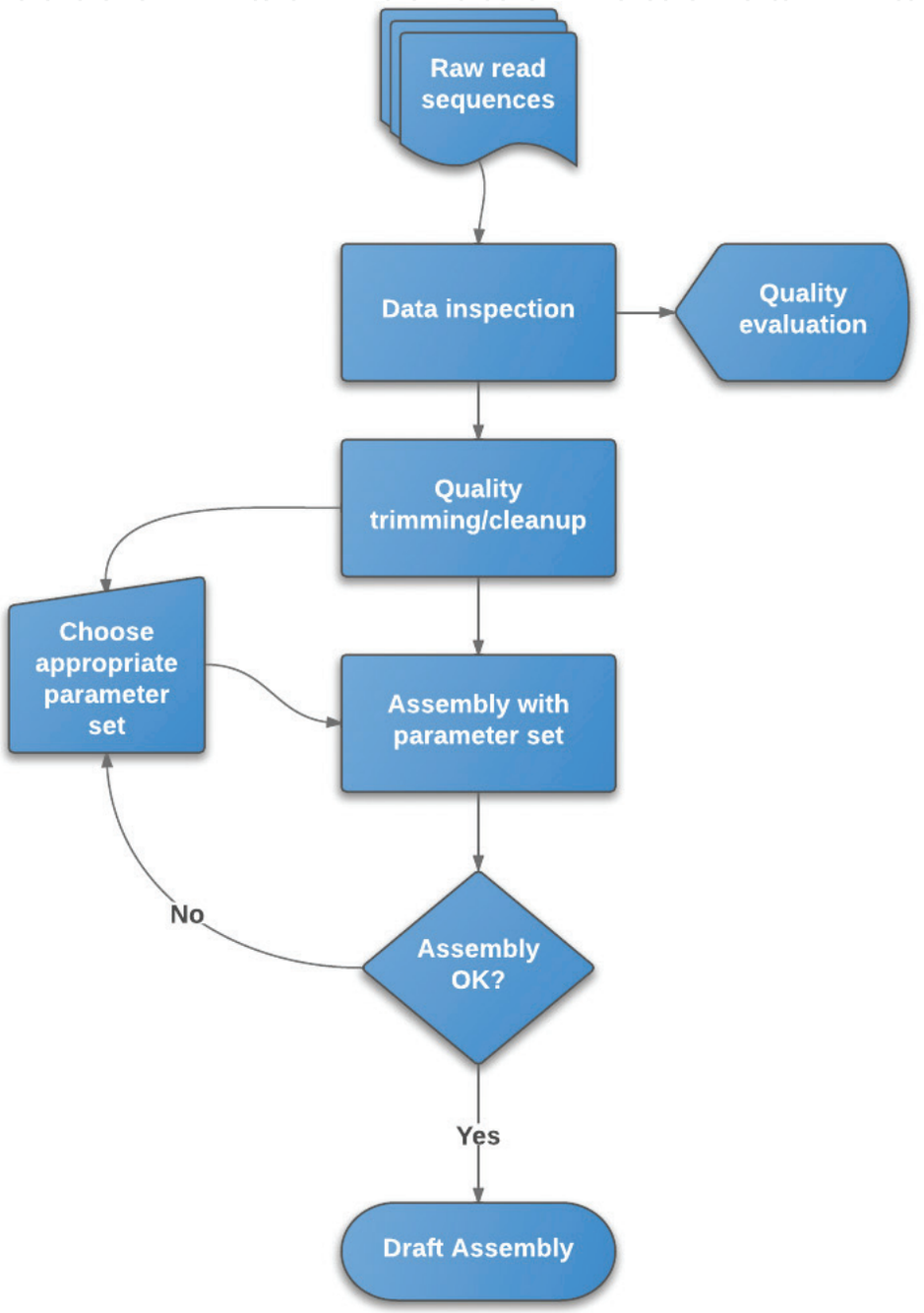

Figure 2. Example of a genome assembly workflow.

4.3.9. RNA-seq method

The transcriptome is the totality of the transcripts present in a cell, each with their abundance and varies depending on the age stage and physiological conditions of the animals. The study of the transcriptome allow to interpret the functional genome elements identifying the different types of transcripts (mRNA, coding RNA, and small RNAs), to determine their level of expression and their structure (terminations to the 5' and 3', exon-intron structure, alternative splicing). The introduction of NGS methods has revolutionized the characterization and quantification of transcriptomes overcoming limitations of previous methods (for example, microarrays or RT-PCR), as the need for a previous knowledge of the genomic sequence, the background noise (due to cross-hybridization), and the limited dynamic range. Starting from the mRNA, a cDNA library is constructed and specific adapters to both ends of each sequence are used. Each molecule is then sequenced after the PCR amplification. The sequencing produces a series of short reads, which can be aligned to the genome or the reference transcriptome. These are used to construct an expression profile for each gene, with a resolution that can reach up to a single base. If the reads are short and very numerous (up to several million per run) you can identify transcripts with a low level of expression. The RNA-seq is highly reproducible and requires less RNA for the synthesis library, since it lacks the step of cloning. Nc RNAs (noncoding RNA) are small molecules which are not translated into proteins. This class of RNAs includes several RNAs as the transfer RNA (tRNA), ribosomal RNA (rRNA), small nuclear and small nucleolar RNA, micro RNA and small interfering RNA (miRNA and siRNA). Micro RNAs (21-nucleotide-long RNAs) are very important posttranscriptional regulators of genes in animals. **Table 3** reports the most important Internet reference databases.

NCBI–RefSeq	htttp://www.ncbi.nlm.nih.gov/refseq
UCSC	http://genome.ucsc.edu
Ensembl	http://www.ensembl.org/index.html
dbSNP	htpp://www.ncbi,nlm.nih.gov/SNP/index.html
dbVar	http://www.ncbi.nlm.nih.gov/dbvar http://ncbi.nlm.gov/dbvar/content/overview
dbGaP	www.ncbi.nlm.nih.gov/gap
DNA Data of Japan	http://www.dbbj.nig.ac.jp
ExPASy	http://ca.expasy.org
GenAtlas	http://citi2.fr/GENATLAS/
GenBamk	http://wwww.ncbi.nlm.nih.gov/
SNP consortium	http://snp.shl.org
Stanford microarray	http://genome.stanford.edu/microarray
Swiss-Protein	http://www.expasy/sprot

Table 3. List of reference databases (Internet sites).

5. Other methods for screening the genetic diseases

5.1. Mendelian disorders

Karyotyping and *in situ* fluorescent hybridization technique (FISH) were traditionally used to detect the large genetic disorders. In the last decade, microarrays have been used to improve the resolution. Nowadays, NGS technologies can be used to sequence all the exomes or targeted exome regions.

5.2. Microarrays or SNP chip

SNP array is a specific type of chip array which is used to identify the single point mutations. After the DNA extraction, DNA is fragmented and processed by means of biochemical methods and labeled with a fluorescent dye. The DNA spots (20–100 bp) are attached to a solid surface (glass, plastic, or silicone) and the hybridization is performed using the samples of DNA to be tested. For each SNP to be genotyped, several DNA probes are designed with similar sequences but variable at one position corresponding to a polymorphism. The measure of the fluorescence, at each spot, allows us to detect the presence/absence of a specific allele. Two individual are identical at 99.9% of the genome and the wide linkage disequilibrium (LD) observed throughout the genome reduces the number of needed SNPs. If many SNPs are linked together in one region, only one maker is needed to be accurately measured. The ability to detect SNPs using array-based approaches is often limited by the density of the array.

5.3. Epigenetics

Methylation effects and histone modifications explain most of the epigenetic and posttranscriptional modifications. Chip-seq (chromatin-immuno-precipitation and direct sequencing) is used at the whole-genome level [28].

5.4. Analysis of a nucleotide sequence

The operations to be carried out in the molecular laboratory are as follows:

- extracting DNA from a tissue, usually the blood;
- DNA amplification. It consists in amplifying selectively the DNA;
- DNA sequencing. Once you obtained the various DNA fragments amplified by PCR, you run the "sequencing reaction";
- Analysis of the nucleotide sequences and search for gene mutations.

Specific software is used to perform an alignment between the nucleotide sequences of an individual showing a monogenic disease (from which you want to search for the mutation) and those of a control healthy individual (i.e., which does not have one specific disease). From this comparison, if the animal has a monogenic disease, may emerge one or more differences in the nucleotide sequence. These mutations (silent mutations, missense, not sense, frameshift, and splicing) shall be individually analyzed and evaluate the effects. The prediction of the effects of a missense mutation in the protein structure level is a

complex process, in which we must take into account all the characteristics of the amino acids protein chain and their mutual interactions. Currently, some programs are available (such as SwissPdbViewer, for free on the websitehttp://www.expasy.org/spdbv/text/download.htm), able to process all of these data and provide a prediction of how the mutation will affect the three-dimensional structure of the protein. The automated sequencer provides the data files that can be displayed in the form of electropherograms with specific programs such as MT-Navigator (Applied Biosystems). The succession of peaks that make up the electropherogram is the result of the detection the fluorescence emitted by the fluorophores linked to the four ddNTPs used in the reactions of sequence. For each fluorochrome, and then to each nucleotide, it is conventionally associated a different color (A = green, T = red, C = blue, and G = black), thus facilitating the reading of the sequence. The reference sequences for the comparison may be obtained from the National Database Center for Biotechnology Information (NCBI, website: http://www.ncbi.nlm.nih.gov). The individual mutations are recognizable on the basis of several peak patterns. The substitutions of a single nucleotide or more nucleotides can be recognized because of its presence in a specific point of different color peaks corresponding to the replaced bases. One individual homozygous shows only one peak while a heterozygous shows two peaks corresponding to two alleles. A mutation can be localized in correspondence of the first base of the junction of splicing (GT becomes TT). In fact in most eukaryotic genes, introns begin with the "GT" bases and end with "AG."

6. Clinical applications in domestic animals

Animal geneticist, whose role consisted only once in calculating the probability of transmission hereditary of a given genetic disease in a given individual, is now an active figure throughout the clinical process, from diagnosis to treatment. The collection of family history is a central point in clinical genetics. It, in fact, can help to arrive at a correct diagnosis. In addition, through early diagnosis, it is possible to prevent or delay the onset of the disease through a suitable therapy. This possibility is often about relatives of an individual and can take to extend the analysis beyond the individual subject. Even for a correct formulation of the prognosis, it is important to be able to refer to the specific form assumed by a given disease in a certain family. For example, in some cases, analysis of family history has allowed to recognize some individuals with less aggressive forms of a disease, a less severe clinical picture, and then improved life expectancy.

7. Genetic counseling

Genetic counseling is a communication process that aims to provide to individual pet owners or farmers the risk of genetic diseases in order to help these people to analyze the consequences and to make responsible decisions. The first task of the genetic counselor is to identify the individual's needs and provide information with an appropriate manner to the degree of its culture and its beliefs. In addition to information about the disease, it must

present the possible available genetic tests, highlighting their reliability. Finally, it should discuss the implications of the various choices.

Author details

Edo D'Agaro

Address all correspondence to: edo.dagaro@uniud

Department of Agricultural, Food, Environment and Animal Sciences, University of Udine, Udine, Italy

References

[1] Mardis, E R. The impact of next-generation sequencing technology on genetics. Trends Genet. 2008; **24**: 133–141.

[2] Schatz M C, Delcher A L, Salzberg, S L. Assembly of large genomes using second – generation sequencing. Genome Res. 2010; **20**: 1165–1173.

[3] Dong Y. Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (*Capra hircus*). Nat. Biotechnol. 2013; **31**: 135–141.

[4] Narzisi G, Mishra B, Schatz MC. Algorithms for computational biology. Lecture Notes in Computer Science. 2014; **21**: 183–195.

[5] Gnerre S. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. USA. 2011; **108**: 1513–1518.

[6] Berlin K. Assembling large genomes with single-molecule sequencing and locality – sensitive hashing. Nat. Biotechnol. 2015; **33**: 623–633.

[7] Chaisson, M J, et al. Resolving the complexity of the human genome using single-molecule sequencing. Nature. 2015; **517**: 608–611.

[8] Pendleton M. Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat. Methods. 2015; **12**: 780–786.

[9] Sharon, D., Tilgner, H., Grubert, F., Snyder, M. A single-molecule long-read survey of the human transcriptome. Nat Biotechnol. 2013; **31**: 1009–1014.

[10] Gordon D. Long-read sequence assembly of the gorilla genome. Science. 2016; **352**: aae0344.

[11] Koren S. Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biol. 2013; **14**: R101.

[12] Sebat J. Large-scale copy number polymorphism in the human genome.Science. 2004; **305**: 525–528.

[13] Browning S R, Browning B L. Haplotype phasing: existing methods and new developments. Nat. Rev. Genet. 2011; **12**: 703–714.

[14] Zheng G X. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nat. Biotechnol. 2016; **34**: 303–311 (2016).

[15] Phillippy A M, Schatz M C, Pop M. Genome assembly forensics: finding the elusive misassembly. Genome Biol. 2008; **9**: R55.

[16] Baker M. Structural variation: the genome's hidden architecture. Nat. Methods. 2012; **9**: 133–137.

[17] Hosomichi K, Shiina T, Tajima A, Inoue I. The impact of next-generation sequencing technologies on HLA research. J. Hum. Genet. 2015; **60**: 665–673.

[18] Kuleshov V. Whole-genome haplotyping using long reads and statistical methods. Nat. Biotechnol. 2014; **32**: 261–266.

[19] Chen X. The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development. Cell. 2014; **158**: 1187–1198.

[20] Ross M G. Characterizing and measuring bias in sequence data. Genome Biol. 2014; **14**: R51.

[21] Cao H. Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. GigaScience. 2014; **3**: 34.

[22] Oulas A. Metagenomics: tools and insights for analyzing next–generation sequencing data derived from biodiversity studies. Bioinform. Biol. Insights. 2015; **9**: 75–88.

[23] Voskoboynik A. The genome sequence of the colonial chordate, *Botryllus* schlosseri. eLife. 2013; **2**: e00569.

[24] Li R. The sequence and de novo assembly of the giant panda genome. Nature. 2010; **63**: 311–317.

[25] Church D M. Extending reference assembly models. Genome Biol. 2016; **16**: 13.

[26] Putnam NH. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res. 2016; **26**: 342–350.

[27] Paszkiewicz K, Studholme DJ. *De novo* assembly of short sequence reads. Brief Bioinform. 2010; **11**(5):457–72.

[28] Fang G. Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia coli* using single-molecule real-time sequencing. Nat. Biotechnol. 2012; **30**: 1232–1239.

IntechOpen