Erciyes Med J 2014 36(3): 99-103 • DOI: 10.5152/etd.2014.7803

Next-Generation DNA Sequencing Technologies

REVIEW

Gülsüm Kayman Kürekçi, Pervin Dinçer

ABSTRACT

Recently, the requirement for fast, low-cost, and highly reliable methods has triggered the development of next-generation DNA sequencing technologies. In addition to overcoming the restrictions in Sanger sequencing methods, large-scale genetic data have provided the development of different analysis methods. In this paper, widely used next-generation sequencing platforms, the analyses of the data obtained, and the applications of these technologies are discussed.

Key words: Genome, next-generation sequencing, high-throughput sequencing

INTRODUCTION

Since it was published in 1977, the most widely used DNA sequencing method has been the Sanger sequencing method, which is based on the chain-termination method (1). The human genome sequence, approximately 3 billion bases long, was completed in 2001 with automated Sanger sequencing (2). Despite technical developments that occurred over the years, the restrictions that are encountered in large-genome sequencing have been substantially overcome via the development of next-generation sequencing technologies in recent years. The common feature of next-generation DNA sequencing platforms that have been in ongoing development since 2005 is that they allow large-scale data output through parallelized sequencing of millions of DNA sequences (3). Next-generation sequencing technologies have rapidly had many fields of application in genetic and biological research (4). In this paper, the common features and differences of the Roche/454, Illumina/Solexa, and Life/APG platforms, which are among the most common sequencing technologies, are primarily discussed. The development of the latest technologies, also known as "single DNA molecule," platforms, such as Helicos Biosciences and Pacific Biosciences, are also briefly mentioned. Finally, a considerable amount of data analysis methods that have been obtained and the application fields of these technologies are outlined.

Sanger sequencing and next-generation DNA sequencing technologies

Sanger sequencing

The first human reference genome was completed in 13 years through the chain termination method, also known as the Sanger sequencing method (1, 2). For the approach used for sequencing the entire human genome, more than 20.000 artificial bacteria were cloned to chromosomes for each sequence, which is approximately 100 kb in length. The sequencing operation is carried out with the addition of 2'-deoxynucleotides (dNTPs) and 2 ',3'-dideoxynucleotide (ddNTPs) to the template chain, which will be sequenced by DNA polymerase (Figure 1). During synthesis, whenever a dideoxynucleotide that does not carry a 3'OH group is added, chain extension stops. The resulting DNA molecules of different lengths are eluted by capillary electrophoresis. Consequently, DNA sequences up to 700-800 bp can usually be read without error. The most important reason for using Sanger sequencing to verify genetic variations defined by next-generation sequencing is that the error ratio in sequencing is very low.

Next-generation DNA sequencing technologies

Although various next-generation sequencing platforms that have been developed are based on different biochemical bases, the basic approaches that are used have similarities (5). Instead of a bacterial cloning phase used in Sanger sequencing, the DNA library is formed by means of nebulization, in which it is forced to go through a small hole and fragmented via genomic DNA sonication (with the help of sound waves) or gas pressure and randomly fragmented in similar dimensions. Short adapter sequences that are used for replication and sequencing phases that are complementary to oligonucleotide sequences are attached to all of the ends of the DNA fragments. The DNA library that is created is copied and sequenced via different methods, depending on the platform used.

Department of Medical Biology,
Hacettepe University
Faculty of Medicine,
Ankara, Turkey

Submitted
07.08.2013

Accepted
16.09.2013

Correspondance
Gülsüm Kayman Kürekçi MSc,
Department of Medical Biology,
Hacettepe University
Faculty of Medicine,
Ankara, Turkey
Phone: +90 312 305 25 41
e.mail:
gulsum.kayman@hacettepe.edu.tr

Roche/454 pyrosequencing

The first Next-generation sequencing platform that was created in 2005 and used a few years later in a developed GS FLX Titanium (Roche/454 Sequencing; Branford, USA) device is based on the pyrosequencing method that was presented by Ronaghi et al. (6). The basis of pyrosequencing depends on the detection of pyrophosphate molecules released during chain synthesis. First of all, DNA fragments are connected to beads coated with oligonucleotides that are complementary to adapter sequences at the ends, and amplification operation is performed through emulsion PCR (7). During emulsion PCR, within the micelles containing DNA polymerase, nucleotides, and primers, short DNA sequences are amplified. The beads carrying thousands of copies of a different DNA fragment on their surface are inserted into a plate consisting of millions of wells. Wells also contain enzymes that provide chemiluminescent detection. Sequencing is accomplished, respectively, by the addition of DNA polymerase and a type of nucleotide. When a nucleotide is added to template chains that are fixed to the beads by DNA polymerase, ATP sulfurylase enzyme converts released pyrophosphate to ATP. By converting ATP to a light-generating molecule in direct proportion to the amount of luciferase enzyme, the type and number of base that is added are determined per cycle (Figure 2) (8). The most important advantage of pyrosequencing compared to other next-generation sequencing platforms is that it has the longest readings (up to 1000 base pairs), which facilitates reference genome alignment of sequenced DNA fragments and de novo (without reference sequence) binding. On the other hand, some of the disadvantages of this platform are that it has the lowest total output (700 Mb) and the highest cost per base (9).

Illumina/Solexa Genome Analyzer

In the platform called Genome Analyzer (Illumina/Solexa; San Diego, USA), issued in 2006, a bridge PCR method is used during the amplification phase (Figure 3) (10, 11). The DNA templates that have adapter sequences on both ends are connected to them via a glass surface that is coated with primer complementary sequences on one end. During amplification, the free end is connected to the closest complementary primer and takes the form of a bridge. After each amplification cycle, in which DNA polymerase synthesizes the chain, chains are denatured. The sequencing of the resulting DNA clusters is performed through the cyclic reversible termination method. With the addition of primer, DNA polymerase, and different fluorescently labeled terminator nucleotides, sequencing begins. With the addition of each nucleotide, synthesis pauses, and fluorescence light is recorded. After the removal of the terminator chemical group that is connected to the nucleotide, the next synthesis cycle takes place. Although Genome Analyzer Gb is the

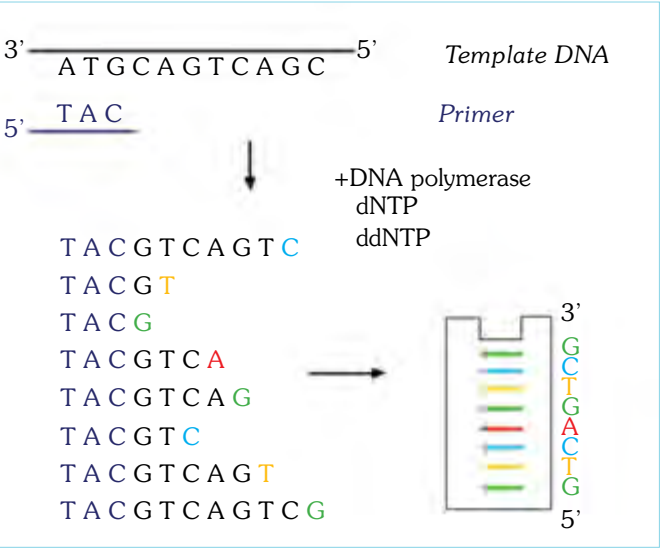

Figure 1. Sanger sequencing

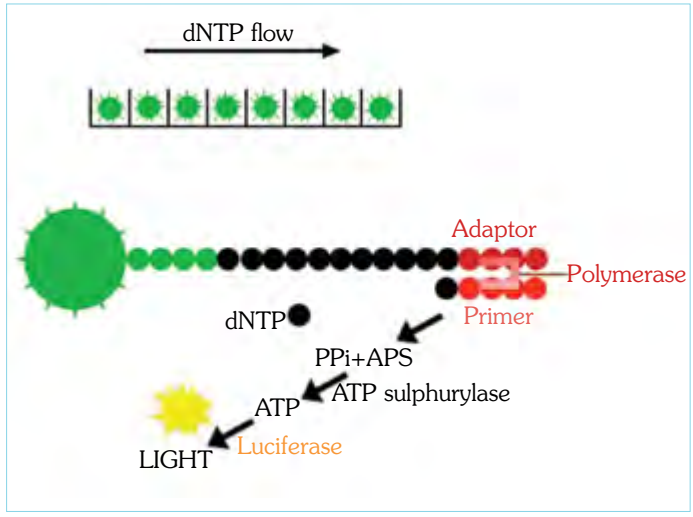

Figure 2. Roche/454 Pyrosequencing. In pyrosequencing, pyrophosphate that is released during DNA template synthesis is converted into a light-generating molecule

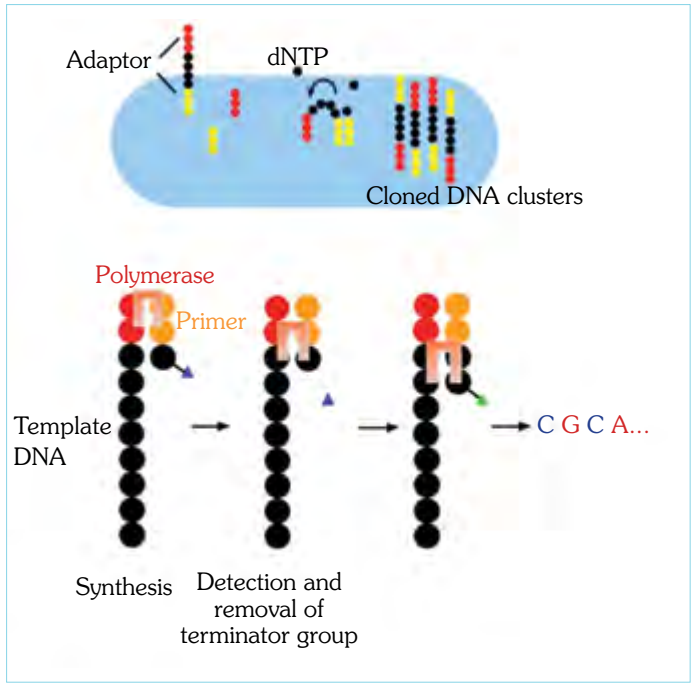

Figure 3. Illumina/Solexa Genome Analyzer. Amplification of templates in a Genome Analyzer device is carried out with bridge PCR. The sequencing operation starts with the attachment of a complementary nucleotide that carries the terminator group. After the nucleotide is added by polymerase, the fluorescent signal is recorded. Before moving on with the next cycle, the terminator chemical group is removed

highest-scale next-generation sequencing platform, compared to 454 technology, the reading length is shorter (2x100 bp).

Applied biosystems/SOLiD and Ion torrent

In the SOLiD (*Supported Oligonucleotide Ligation and Detection*) (Applied Biosystems/Life Technologies; Carlsbad, USA) system, released by Life Technologies in 2006, DNA fragments, amplified by emulsion PCR, are sequenced via ligation (12, 13). In this sequencing approach, instead of a synthesis phase with DNA polymerase, short nucleotide ligation, marked by DNA ligase enzyme, is performed. These short sequences, also known as interrogation probes, consist of 2 bases specific to the probe and 6 variable bases. Furthermore, the 5' end of each probe is marked with one of four different fluorescence molecules. The sequencing reaction blend consists of interrogation probes that include 16 different combinations that might be created by the first interrogated base pair. When an interrogated probe hybridizes with the template sequence, fluorescent light is recorded. After the 5' end that is attached to the fluorescent label is cut and removed, the next interrogation probe is connected, and ligation is performed through connecting the DNA ligase enzyme probe (Figure 4). After 7 ligation cycles, templates are denatured and separated from each other, and the sequencing operation is repeated, starting from one base behind the beginning base. In this way, since each base is interrogated twice, a low error rate is provided. Another DNA sequencing possibility presented by Life Technologies is the Ion Torrent platform (14). In this method, no fluorescent light or chemical modification is used, and the hydrogen ion that is released when each nucleotide is added during DNA synthesis causes a pH change in the solution. The pH change is detected by an ion detector and recorded.

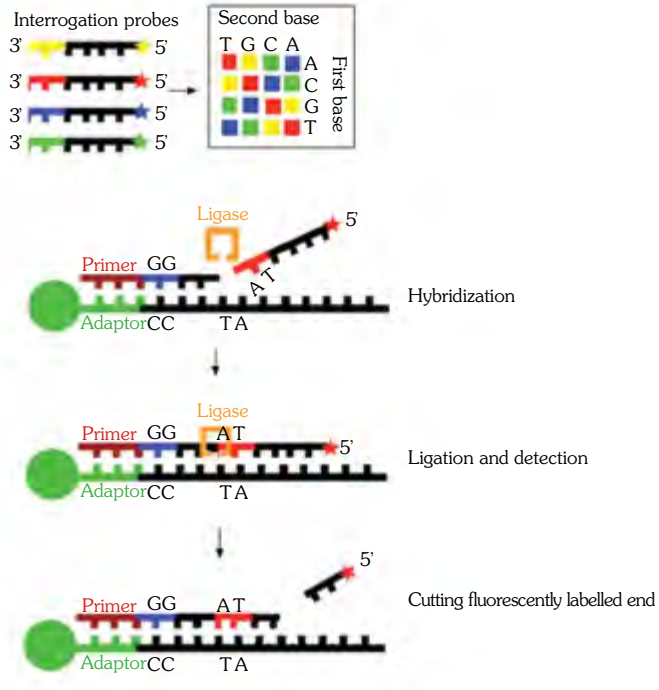

Figure 4. Applied Biosystems/SOLiD. Ligase enzyme combines fluorescently labeled interrogation probes via genetic binding. With the help of interrogation probes specific to 16 different combinations that might be created by the first interrogated base pair, the DNA sequence is detected

Single-molecule DNA template sequencing technologies

Because the fluorescent detection methods that are used in the next-generation sequencing platforms mentioned above are designed to detect amplified signals, they require template chains to be amplified. However, apart from these technologies, devices that do not require amplification phases in which single DNA molecules are used as templates and are more sensitive to low fluorescent signals have been developed. The first of these technologies is the HeliScope platform, developed by Helicos Biosciences (Cambridge, USA) (15). 3' ends of genomic DNA fragments are attached to poly-A oligonucleotides, in which the last adenosine fluorescent light is carried. After the fragments are hybridized to poly-T oligonucleotides serving as the primer and fixed on the surface of a flow cell, complementary chain synthesis occurs with DNA polymerase and labeled dNTPs. The label of each added labeled dNTP is recorded before it is removed. In the SMRT (single-molecule real-time sequencing-by-synthesis) platform developed by Pacific Biosciences (Menlo Park, USA), DNA polymerase enzymes that are fixed on the flow cell are used (16). It is capable of real-time recording for millisecond fluorescent signals that are comprised of the addition of each labeled nucleotide. Another real-time sequencing-based platform is the GridION device, developed by Oxford Nanopore Technologies (Oxford, United Kingdom), but it has not yet been commercialized. Single-chain DNA molecules, without any operation, pass through nanometric-diameter wells where electric current is applied. By using specific current changes for each nucleotide type, real-time detection of a DNA molecule sequence is provided. As a result, the most important advantage of single-molecule DNA technologies is that they do not require an amplification phase, and the initial DNA sample amount (<1 µg) is less than other methods. However, since the sequencing operation can not be repeated, as in other platforms that use cloned DNA, and since a single molecule is sequences, the error ratio in the sequencing may increase (18).

Next-generation sequencing data analysis

The huge amount of data that have been obtained from next-generation sequencing methods needs to be analyzed (19). Next-generation data analysis basically consists of several steps (20). The first step is called "base calling," in which fluorescent and light signal images that are recorded in the experiment phase are transformed into a sequence reading phase. In different algorithms offered by the platforms themselves, for each base that is read, quality scoring is performed, and sequence readings are acquired.

In the second step, the obtained short readings are sorted and combined according to the reference sequence, if available; if not, "de novo" is used. Genetic variations are studied through comparisons with the most current human genome sequence (GRCh37) that is used as the reference. One important point in this step is the coverage ratio of the readings of the reference sequence. The more that the overlapping reading sequence is covered in a genomic region, the more accurate the obtained sequence data become. Despite the high coverage ratio presented in the reading combination of next-generation sequence platforms, there is an important restriction as well. The shortness of the reading sequences makes the combination of the regions consisting of repeat sequences difficult.

It may also cause some difficulties in the sequencing of pseudogenes and related genes that share homologous sequences accurately (21). Another restriction of short reading sequences is that it is difficult to detect structural changes that cover larger areas (at least 1 kilobase) in DNA sequences, such as copy number variations. However, owing to recent changes in analysis methods, the detection of copy number variations is now possible (22).

Application fields of next-generation sequence technologies

Studying genetic variations

The most important application field of sequencing methods is to study genetic variations related to complex or monogenic diseases in which all genomes or a targeted genetic region is sequenced (23, 24). In this way, it is possible to study gene expression differences in the human genome and the effects of polymorphisms on phenotype. In this context, after quality filtering of the obtained data, sequences are compared with reference sequences, and all of the changes that are detected are classified according to the various identified criteria. Generally, these criteria are elimination of polymorphisms, prioritization of the variants in exonic or regulatory regions, and a specific mutation type (missense, nonsense, insertion-deletion) (25). Furthermore, there are programs that provide classification of detected genetic variations according to interspecific immunity and the functional effect on gene products (26).

Genomic analysis and exome sequencing

The high-scale data analysis opportunity offered by next-generation sequencing methods contributes biological information by sequencing all genomes of various organisms, from bacteria to human (27, 28). Among all genome-targeted studies, in recent years, studies on the cancer genome, defined as genetically instable, have accelerated (29, 30). Studying genetically sequential and structural changes that may take place in tumor cell genomes provides a better understanding of factors that affect tumor formation, growth, and spread.

Apart from all genome sequencing approaches, there are methods that allow capturing and sequencing of a specific region in the genome (31). This approach is called targeted sequence capture, and it allows for the classification of a very large amount of data obtained from the entire genome sequencing method and facilitates data assessment. By using this approach, the genome is preferred only in a coding sequence analysis when studying new genes responsible for single-gene diseases. This method is also called the exome sequencing method, which allows the capture and sequencing of exonic regions, which constitute approximately 1% of the genome (25). The NimbleGen (Roche; Branford, USA) and Agilent (Agilent Technologies; Santa Clara, USA) systems are the most widely used exome sequencing platforms (32, 33). The main reasons to prefer only exonic region sequencing to entire-genome sequencing are that it not only provides cost and data load alleviation but also holds 85% of rare variants responsible for single-gene diseases in these regions (24).

Transcriptome analysis

Owing to next-generation sequencing technologies, sequencing, mapping, and determining the amount of transcripts in biological samples have been easier (34). This approach, comprising transcript data analysis, is known as RNA-Seq (35). In this approach, RNA molecules are isolated and transformed into cDNA molecules. After the fragmentation of single-chain cDNA molecules, adapter sequences that suitable for next-generation sequencing platforms are attached to the ends, and the sequencing operation takes place. The RNA-Seq approach is applied for many different reasons, such as gene annotation, gene expression analysis, detection of splice variants, and determination of exon-intron limits. On the other hand, transcriptome analysis also includes non-coding RNAs that do not code proteins. Many RNA molecules, such as transfer RNA, ribosomal RNA, nuclear and nucleolar RNAs, and microRNAs, are in this group. Owing to next-generation sequencing technologies, it is now possible to detect and determine the amount of new short RNAs that regulate gene expression after translation (36).

Chromatin immunoprecipitation analysis

The analysis of proteins (transcription factors, histone proteins) that are attached to DNA molecules and regulate gene expression is performed with the chromatin immunoprecipitation analysis (ChIP) method (37). In this approach, DNA is fragmented after proteins are fixed on binding domains on DNA. Immunoprecipitation is performed on targeted proteins with the help of specific antibodies, and the determined DNA-binding domains are sequenced. When combined with next-generation sequencing methods, the ChIP-Seq method allows the sequencing of a DNA-binding domain library with next-generation sequencing (38).

Genomic methylation mapping

Examination of DNA methylation, which is an important epigenetic mechanism, especially in terms of its relationship with diseases, has gained importance in recent years (39). It is known that there is a relationship between the DNA methylation mechanism, which takes place as a result of addition of a methyl group to cytosines in certain regions of the genome, with important cellular processes, such as gene expression regulation, X-chromosome inactivation, and genomic imprinting. The most common methods for determining DNA methylation in the genome are restriction endonuclease enzyme digestion, which distinguishes methylated and unmethylated DNA; methylated DNA immunoprecipitation; and bisulfite sequencing methods in combination with next-generation sequencing approaches (40). Some of the approaches applied with next-generation sequencing methods are HELP-Seq (Hpa II tiny fragment enrichment by ligation-mediated PCR), MeDIP-Seq (methylated DNA immunoprecipitation), and WGSBS (whole-genome shotgun bisulfite sequencing).

CONCLUSION

High-scale next-generation DNA sequencing methods that allow millions of DNA sequences to be sequenced at one time and at the same time continue to be developed rapidly. The most recent approaches allow single-DNA molecule templates to be sequenced in real time without the requirement of an amplification phase. Today, despite the fact that the high sequencing cost and technical restrictions of large-scale genome sequencing play an important role and that researchers are directed to the target sequencing approach, in the coming years, it is expected that these restrictions will rapidly be overcome, and next-generation sequencing technologies will take place in many research fields, in addition to clinical diagnostic use.

Peer-review: Externally peer-reviewed.

Authors' Contributions: Conceived and designed the experiments or case: GKK, PD. Performed the experiments or case: GKK, PD. Analyzed the data: GKK, PD. Wrote the paper: GKK, PD. All authors have read and approved the final manuscript.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study has received no financial support.

REFERENCES

1. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 1977; 74(12): 5463-7. [CrossRef]
2. International Human Genome Consortium. Finishing the euchromatic sequence of the human genome. Nature 2004; 431(7011): 931-45. [CrossRef]
3. Metzker ML. Sequencing technologies – the next generation. Nature Rev Genet 2010; 11(1): 31-46. [CrossRef]
4. Shendure J, Lieberman Aiden E. The expanding scope of DNA sequencing. Nat Biotechnol 2012; 30(11): 1084-94. [CrossRef]
5. Mardis ER. Next-Generation DNA sequencing methods. Annu Rev Genomics Hum Genet 2008; 9: 387-402. [CrossRef]
6. Ronaghi M, Uhlen M, Nyren P. A sequencing method based on real-time pyrophosphate. Science 1998; 281(5375): 363-5. [CrossRef]
7. Tawfik DS, Griffiths AD. Man-made cell-like compartments for molecular evolution. Nat Biotechnol 1998; 16(7): 652-6. [CrossRef]
8. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al.. Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005; 437(7057): 376-80.
9. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol 2007; 8(7): R143. [CrossRef]
10. Bentley DR. Whole-genome re-sequencing. Curr Opin Genet Dev 2006; 16(6): 545-52. [CrossRef]
11. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 2008; 456(7218): 53-9. [CrossRef]
12. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, et al. Accurate multiplex polony sequencing of an evolved bacterial genome. Science 2005; 309(5741): 1728-32. [CrossRef]
13. McKernan K, Blanchard A, Kotler L, Costa G. Reagents, methods and libraries for bead-based sequencing. US20080003571; 2008.
14. Rothberg JM, Wolfgang H, Rearick TD, Schultz J, Mileski W, Davey M, et al.. An integrated semiconductor device enabling non-optical genome sequencing. Nature 2011; 475(7356):348-52. [CrossRef]
15. Braslavsky I, Hebert B, Kartalov E, Quake R. Sequence information can be obtained from single DNA molecules. Proc Natl Acad Sci U S A 2003; 100(7): 3960-4. [CrossRef]
16. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science 2009. 323(5910); 133-8. [CrossRef]
17. Oxford Nanopore Technologies, Available from: URL: http://www.nanoporetech.com.
18. Pop M, Salzberg SL. Bioinformatics challenges of new sequencing technology. Trends Genet 2008; 24(3): 142-9. [CrossRef]
19. Su Z, Ning B, Fang H, Hong H, Perkins R, Tong W, et al. Next-generation sequencing and its applications in molecular diagnostics. Expert Rev Mol Diagn 2011; 11(3): 333-43.
20. Trapnell C, Salzberg SL. How to map billions of short reads onto genomes. Nature Biotech 2009; 27(5): 455-7. [CrossRef]
21. Voelkerding KV, Dames SA, Durtschi JD. Next-generation sequencing: from basic research to diagnostics. Clin Chem 2009; 55(4): 641-58. [CrossRef]
22. Duan J, Zhang J-G, Deng H-W, Wang Y-P. Comparative studies of copy number variation detection methods for next-generation sequencing technologies. PloS ONE 2013; 8(3): e59128. [CrossRef]
23. Rabbani B, Mahdieh N, Hosomichi K, Nakaoka H, Inoue I, et al. Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders. J Hum Genet 2012; 57(10): 621-32. [CrossRef]
24. Casals F, Idaghdour Y, Hussin J, Awadalla P. Next-generation sequencing approaches for genetic mapping of complex diseases. J Neuroimmunol 2012; 248(1-2): 10-22. [CrossRef]
25. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, et al.. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet 2011; 12(11): 745-55. [CrossRef]
26. Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, et al. A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform 2013; 1-23.
27. Nusbaum C, Ohsumi TK, Gomez, J, Aquadro J, Victor TC, Warren RM, et al. Sensitive, specific polymorphism discovery in bacteria using massively parallel sequencing. Nat Methods 2009; 6(1): 67-9. [CrossRef]
28. Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D. Sequencing of natural strains of Arabidopsis thaliana with short reads. Genome Res 2008; 18(12): 2024-33. [CrossRef]
29. Bell DW. Our changing view of the genomic landscape of cancer. J Pathol 2010; 220(2): 231-43.
30. Walter MJ, Graubert TA, DiPersio JF, Mardis ER, Wilson RK, Ley TJ. Next-generation sequencing of cancer genomes: back to the future. Per Med 2009; 6(6): 653-62. [CrossRef]
31. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, et al. Target-enrichment strategies for next-generation sequencing. Nat Methods 2010; 7(2) 111-8. [CrossRef]
32. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, et al. Genome-wide in situ exon capture for selective resequencing. Nat Genet 2007; 39(12): 1522-7. [CrossRef]
33. Gnirke A, Melkinov A, Maguire J, Rogov P, LeProust EM, Brockman W, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol 2009; 27(2): 182-9. [CrossRef]
34. Mutz KO, Heilkenbrinker A, Lönne M, Walter J-G, Stahl F. Transcriptome analysis using next-generation sequencing. Curr Opin Biotechnol 2013; 24(1): 22-30. [CrossRef]
35. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature Rev Genet 2009; 10(1): 57-63. [CrossRef]
36. Morozova O, Marra MA. Applications of next-generation sequencing technologies in functional genomics. Genomics 2008; 92(5): 255-64. [CrossRef]
37. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, et al. Genome-wide location and function of DNA binding proteins. Science 2000; 290(5500): 2306-9. [CrossRef]
38. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet 2009; 10(10): 669-80. [CrossRef]
39. Suzuki MM, Bird A. DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet 2008; 9(6): 465-76. [CrossRef]
40. Laird PW. Principles and challenges of genomewide DNA methylation analysis. Nat Rev Genet 2010; 11(3): 191-203. [CrossRef]