Article title: Next-generation sequencing with emphasis on Illumina and Ion torrent platforms.
Authors: Eman Hagar[1], Ahmed Hassan Hagar[2]
Affiliations: Department of Microbiology, Faculty of science, Alexandria University, Egypt.[1], Faculty of medicine, El-Qalam College for science and technology, Sudan.[2]
Orcid ids: 0000-0003-3698-338X[1], 0000-0002-0368-1388[2]
Contact e-mail: eman_hagger@yahoo.com
License information: This work has been published open access under Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Conditions, terms of use and publishing policy can be found at https://www.scienceopen.com/.
Preprint statement: This article is a preprint and has not been peer-reviewed, under consideration and submitted to ScienceOpen Preprints for open peer review.
DOI: 10.14293/S2199-1006.1.SOR-.PPA9N9O.v1
Preprint first posted online: 24 August 2022

Next-generation sequencing with emphasis on Illumina and Ion torrent platforms.

- **Corresponding author**: Eman Hassan Hagar

E-mail: eman_hagger@yahoo.com

ORCID ID: 0000-0003-3698-338X

Department of Microbiology, Faculty of Science, Alexandria University

Egypt.

- **Author:** Ahmed Hassan Hagar

E-mail: ahmedhagar717@gmail.com

Faculty of Medicine, El-Qalam College for science and technology

Sudan.

Abstract

Background: Next-generation sequencing is a type of deep sequencing. In comparison to the previously used Sanger's method, Next generation sequencing allowing the sequencing of an entire genome in a single day. Next-generation sequencing (NGS) has revolutionized genomics and molecular biology. NGS has a wide range of medical applications, including tumors and inherited disease diagnosis. It is also used to find genetic variants across the genome. There are several NGS platforms available. Illumina and Iontorrent are the most prevalent sequencing platforms. These NGS platforms reduced the cost and time required to sequence a full genome.

The main body of the abstract: The review paper covered a brief history of next-generation sequencing technology (NGS), followed by the benefits of employing this revolutionary technology and The general approach of NGS, beginning with fragmentation and ending with data analysis, was explained, with a focus on the Illumina and Ion torrent platforms. Finally, the data analysis step was thoroughly covered, beginning with data quality control and ending with data visualization.

Conclusion: According to the review article, Next generation sequencing (NGS) is a promising technology that has revolutionized genome sequencing. The NGS platform has resulted in softwares that can perform the vast majority of NGS steps such as sequencing, variant annotation, and quality checks. The Iontorrent and Illumina platforms have grown in popularity and are frequently used. NGS has gained traction in clinical applications.

Keywords:

Next generation sequencing (NGS), Illumina, Iontorrent, sequencing FASTQC, variant calling.

1. Background:

1.1. Next-generation sequencing

Next-generation sequencing (NGS) is a deep sequencing method that enables quick, accurate, and low-cost determination of the nucleic acid sequence of the entire genome (1). Unlike the earlier sequencing method developed by Frederick Sanger, NGS has revolutionized the field of genome sequencing by sequencing millions of fragments per run, as opposed to the sanger method, which only sequences a single DNA fragment(2). Due to the rapid development of NGS, researchers have been able to identify genetic abnormalities associated with diseases, such as variations in genes associated with tumors and those associated with inherited diseases. There are numerous NGS platforms, but we will concentrate on Illumina and Ion torrent in this paper because most clinical sequencing is performed by both of these two platforms.

In 2000, NGS first technology was developed by Massively Parallel Signature Sequencing (MPSS) Lynx Therapeutics (USA) Company. The company was later purchased by Illumina (3).

In 2004, 454 Life Sciences company produced a paralleled version of pyrosequencing which was considered the second of a new generation of sequencing technologies (3).

In 2005–2006, 454 GS 20 Roche sequencing platform was developed which could produce 20 million bases and revolutionized DNA sequencing (20 Mbp) (3).

In 2007, the GS FLX model replaced the 454 GS 20 Roche sequencing platform and was capable of producing 100 million base pairs in 4 hours, which was increased to 400 million base pairs in 2008(3).

In 2008, the first published paper was about the production of human genome sequences using NGS (Watson personal genome sequence) (4).

In 2014, Illumina produced The HiSeq X Ten which is considered the world's first platform capable of providing full coverage of human genomes for less than $1,000(5).

In 2019, the National Human Genome Research Institute reported that the cost of sequencing an entire human genome is 942 USD (6).

2. Main text

2.1. Next-generation sequencing (NGS) has the following advantages

•Using parallel sequencing، NGS can analyze tens to hundreds of thousands of reads per sample.

•Because NGS has high sensitivity and a low detection limit, it can detect low-frequency variants.

•For large data volumes, NGS can provide rapid large-scale sequencing results.

• Next-generation sequencing (NGS) analyses hundreds of genes in a single assay to provide comprehensive genomic profiling.

2.2. The general workflow of NGS:

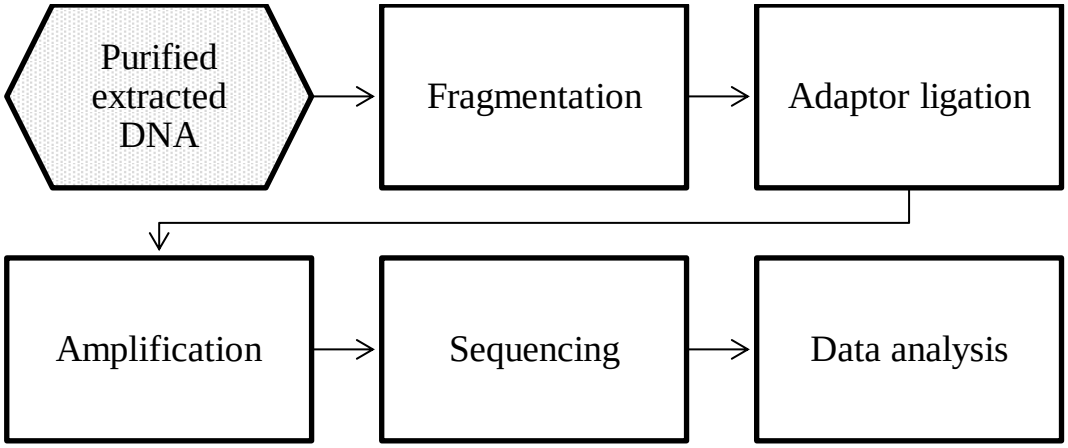

Figure 1: General workflow of NGS.

2.2.1. Fragmentation step

DNA as a complete double strand is too complex and more prone to breakage. To overcome this obstacle, DNA is cut into small fragments, which are more stable and less likely to break. DNA fragmentation can be accomplished in one of two ways: (7):

1. Mechanical techniques such as ultrasonication, shearing, and nebulization
2. Enzyme digestion

2.2.2. Adaptor ligation

Adaptors are oligonucleotides that are added to the blunt end of a DNA fragment. The importance of these adaptors is that they hold the barcoding

sequence as well as the primers, and they are regarded as essential binding sites for immobilization of these DNA fragments to the flowcell, allowing bridge amplification, as in the particular example of the Illumina platform(7).

The barcoding sequence held by the adaptor is known as the unique molecular identifiers (UMIs), which are small DNA sequences. The significance of UMIs is that it is critical to distinguish between real biological DNA duplicates and duplicates caused by PCR clones after the amplification step using PCR. If we have reads that start and end at the same position and have the same UMIs, it means that these reads are not biological duplicates and must be removed and as a result reduce the false negative variants(8). Following these modifications, the DNA fragments are ready to be loaded into the sequencer's flow cell. These flowcells contain oligonucleotides that are complementary to the adaptor attached to the DNA fragment, allowing the DNA fragment to be immobilized within the flowcell.

The flow cell in Illumina is a glass slide with 8 channels known as lanes, each lane with 3 columns, and each column with 100 tiles. Each lane is coated with oligonucleotides which are complementary to the ligation adaptor. As a result, when sequencing libraries are loaded into the lanes, they become immobilized on the lane surface (9). Beads are found in ion torrent flowcells. These beads are coated with oligonucleotides that are complementary to the barcoding sequences, allowing the sequence libraries to be attached to the beads in Ion cells.

2.2.3.Amplification step

Amplification, also known as Cluster generation, is a process that duplicates millions of copies of single-stranded DNA by amplifying clusters of DNA fragments (10). The significance of this step is that it will produce several reads of the same sequence with varying qualities, so we can improve the

quality of our reads by excluding low-quality reads and keeping high-quality reads of the same DNA sequence.

Amplification in Illumina is based on bridge PCR, in which a reverse strand complementary to the original strand is created by PCR and primed by the oligonucleotide coated on the flowcell. After that, the original strand is denatured and washed away. After the reverse strand bridges over to interact with a complementary oligonucleotide coated on the flowcell, the adapter attaches to the complementary oligonucleotide and a new forward strand begins to form up complementary to the reverse strand. When the temperature drops, both the forward and reverse strands become unbridged. Then when the temperature rises, both DNA strands form bridges to adjacent oligonucleotides, and the cycle continues repeatedly, producing millions of DNA pieces (11).

Emulsion PCR is used for amplification in Iontorrent. The flow cell serves as a micro-reactor for PCR because it contains beads, DNA fragments, PCR mix, and emulsion oils. PCR cycling takes place, resulting in the amplification of DNA fragments on each bead (12).

2.2.4. Sequencing

Is the process used to identify the exact nucleotide base sequence (As, Ts, Cs, and Gs) in a fragment of DNA. There are different types of sequencing methods mostly important is sequencing by synthesis which accounts for more than 90% of the total sequencing data worldwide(13).

The Illumina platform is based on sequencing by synthesis (SBS), whereas the IonTorrent platform is based on ion semiconductor sequencing, which is also based on SBS but differs from other sequencing technologies in the fact that it doesn't require chemically modified nucleotides and does not require

fluorescent labeling. Therefore as a side benefit, no fluorescence detection or image scanning is required (14).

2.2.4.1. Sequencing principle in Illumina

In the Illumina platform, Reversible terminators, which are chemically modified nucleotides by the addition of a fluorophore, are used (15). Polymerase enzymes expand sequencing primers by adding reversible terminators. Reversible terminators cause incorporation to stop immediately after the first nucleotide. The polymerases and unbound nucleotides are rinsed away, and the label of the bases integrated for each sequence is read using 4 images taken through different filters, and fluorophores are illuminated with two separate lasers (red: A, C and green: G, T). The fluorophores and terminators are then removed, and the sequencing is resumed with the inclusion of the next base. (16).

Steps of Illumina sequence data generation process:

1. Raw recorded images are analyzed to determine clusters, signal intensity, and noise level for each cluster.
2. Real-time analysis software (RTA) is used for base calling normally, there is no human intervention in the base calling process; everything is left to the software, and researchers focus primarily on the analysis of base calling data.
3. Base calling results are organized in base call files. Afterward, these files are transformed into FASTQ files.

2.2.4.2. Sequencing Principle in Ion torrent

In Ion torrent semi-conductor sequencing, when a nucleotide is integrated into a new DNA strand in the sequencing by synthesis process, a pyrophosphate group and a proton (H+) are released as by-products. The

release of a proton (H+) causes a change in pH. This pH change can be observed and utilized to determine the nucleotide incorporated. It is stated that PH changes are not nucleotide specific. so, to determine the sequence of the DNA nucleotides (dATP, dGTP, dCTP, and dTTP) are introduced into the reaction one at a time during the run. A detected pH change after the introduction of a specific nucleotide indicates that the template strand contains its complementary base at the last position (17).

Steps of Ion torrent sequence data generation process

Unlike Illumina, no images are collected, so there is no need to analyze these images and generate data from them because simply only the change in PH is recorded.

1. Torrent Suite Software is utilized for base calling (18).
2. Base calling results are saved in base call files. These files are then converted into FASTQ files.

There are various NGS file formats, including as FASTAQ, FASTA, CFASTA, SFF, and QUAL. FASTAQ is considered the standard and most widely used format. All other file formats can be converted to FASTQ format. This process is known as this process is called BCF to FASTQ conversion.

FASTQ FILE is a text-based file format that contains each read's sequence as well as the confidence score for each base (19). Compressed FASTQ files are multi-gigabyte in size and contain 200 million or more reads.

FASTQ file consists of 4 lines (19):

1. A sequence identifier containing information about the sequencing run and the cluster.

2. The sequence of bases
3. A separator, which is a plus (+) sign.
4. The base call quality scores (Q-scores). Q-scores are Phred +33 encoded, using ASCII characters to represent the numerical quality scores. The goal of utilizing ASCII characters rather than numbers to convey the Q- score is to reduce the file size by recording a single character rather than recording a number consisting of 2 digits.

2.2.5. Data analysis

Data are analyzed to assess reads quality, align reads against a reference genome, identify variants, and determine the impact of the variant on protein structure and function. Figure (2) depicts 5 main steps of data analysis.

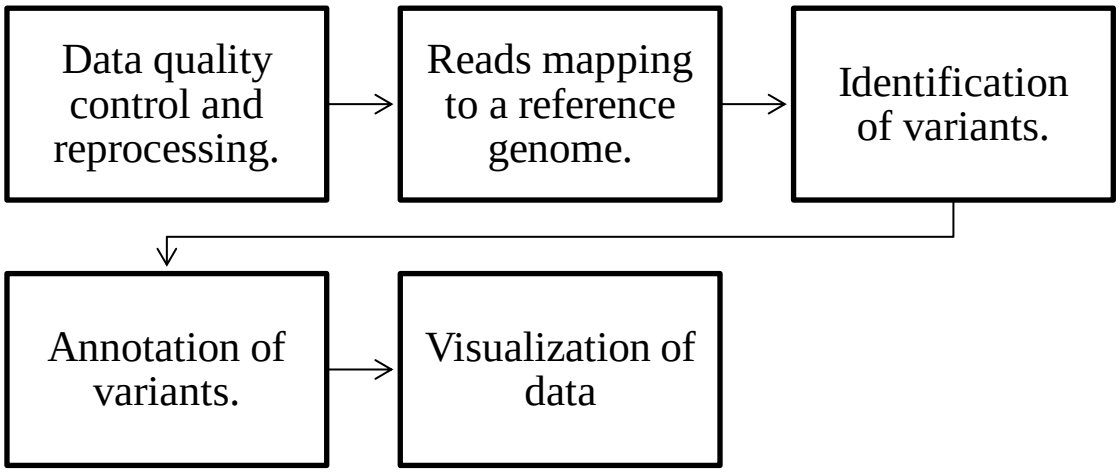

Figure 2: General approach for NGS data analysis.

2.2.5.1. Quality control and reprocessing

The purpose of this phase is checking quality of data to filter and delete low-quality results, so reducing the waste of computational resources and time in subsequent steps.

There are various NGS data quality control softwares available, including NGS QCToolkit, FASTXToolkit, and FASTQC. For quality checks, Illumina uses FASTQC (20), whereas IonTorrent employs the Ion Sphere Quality Control Kit (21). If the Q-scores following quality checks are unsatisfying, reprocessing software can be used to filter and trim unfavorable fragments. These softwares include Sickle, Trimmomatic, ngsShoRT, Skewer, Cutadapt, and NGS QC Toolkit.

2.2.5.2. Reads mapping to a reference genome

The process of allocating reads to a specific place versus a reference genome is known as mapping. This stage could be performed using aligners based on different algorithms (22).

Aligners have varying speed and sensitivity. Bowtie and SOAP2 aligners are faster, novoalgin and stampy are more sensitive, and BWA has a good balance of sensitivity and speed.

The mapping result can be saved in one of two file formats:

1. SAM is a text format with aligned reads that is human readable but takes longer time to parse (23).
2. BAM: is a binary format that is a compressed version of SAM. it is substantially smaller than sam. it saves storage space and is faster to parse, but it is not human readable (23)

2.2.5.3. Identification of variants (Variant calling).

Reads are mapped against a reference genome, and any deviation from the reference genome's sequence is identified as a variant. Single nucleotide polymorphism (SNP) accounts for 90% of the DNA variants of the human genome (24). SNPs can also be detected by identifying single nucleotide mismatches between aligned data and the reference genome (25). Softwares like Samtools, SOAPsnp, GATK, SHORE, SNVer, and MaCH can carry out this process (26). The results of the final calling are saved into variant calling format (VCF) (27).

2.2.5.4. Annotation of variants.

A functional analysis aims to link the alteration in variant sequence with changes in phenotype as well as to investigate the impact of variations on protein structure and function (28).

The evaluated variant can be compared to a known variant. Variants that have previously been sequenced are included in datasets such as dbSNP, which contains diverse genetic molecular variations and can be used as a reference for comparison with the tested variant (28).

In Functional annotation, each variant is classified according to its relationship to coding sequences in the genome and how it may change the coding sequence and affect the gene product (28).

2.2.5.5. Visualization of data.

Visualization is essential for observing trends and outliers in large datasets and linking findings to others.

This process has 2 options:

1. Integrative Genomics Viewer (IGV): an offline application that can be downloaded on personal computers. There is no need to upload data (29).

2. UCSC Genome Browser: it is an online tool, no software must be downloaded. Data must be uploaded to the cloud (29).

3. Conclusion

Next-generation sequencing (NGS) is a very promising technology that has revolutionized genome sequencing by cutting the cost and time necessary to sequence a whole genome. It also made it much easier to detect mutations throughout the genome, making it possible to identify the genes responsible for tumors and genetic disorders. It also allow detecting the effect of variations on the structure and functions of the protein. NGS platform has resulted in software capable of performing the majority of NGS steps. The Illumina and Ion torrent platforms have gained significant popularity and are widely used. Throughout the NGS process, the highest quality reads are achieved by filtering and trimming low-quality data. NGS has gained a growing reputation in clinical applications. Recently, some governments have launched projects to sequence large populations.

Abbreviations

NGS: Next-generation sequencing, , UMIs: Unique molecular identifiers, DNA: Deoxyribonucleic acid, PCR: Polymerase chain reaction, SBS: Sequencing by synthesis, A: Adenine, T: Thiamine, C: Cytosines, G: Guanines, RTA: Real-time analysis software, dATP: Deoxyadenosine triphosphate, dGTP: Deoxyguanosine triphosphate, dCTP: Deoxycytidine Triphosphate, dTTP: Deoxythymidine triphosphate, BCF: Base Calling files, ASCII: American Standard Code for Information Interchange, SAM:

Sequence Alignment Map, BAM: Binary Alignment Map, SNP: Single nucleotide polymorphism, VCF: Variant calling format, dbSNP: Database of single nucleotide polymorphisms, IGV: Integrative Genomics Viewer.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

No datasets were generated during the study.

Competing interests

The authors declare that they have no competing interests.

Funding

No funding was received for the purpose of the study.

Authors' contributions

Both authors contributed equally to the manuscript.

Acknowledgment

None

References:

1. Illumina. (2009). *Next-Generation Sequencing (NGS) | Explore the technology*. Illumina.com.

https://www.illumina.com/science/technology/next-generation-sequencing.html

2. NGS vs. Sanger Sequencing. (n.d.). Www.illumina.com. https://www.illumina.com/science/technology/next-generation-sequencing/ngs-vs-sanger-sequencing.html#:~:text=The%20critical%20difference%20between%20Sanger
3. Barba, M., Czosnek, H., & Hadidi, A. (2014). Historical Perspective, Development, and Applications of Next-Generation Sequencing in Plant Virology. Viruses, 6(1), 106–136. https://doi.org/10.3390/v6010106
4. Robertson, S. (2016, March 6). History of Next Generation Sequencing. News-Medical.net. https://www.news-medical.net/health/History-of-Next-Generation-Sequencing.aspx
5. Illumina Introduces the HiSeq XTM Ten Sequencing System. (2014, January 14). Investor.illumina.com. https://investor.illumina.com/news/press-release-details/2014/Illumina-Introduces-the-HiSeq-X-Ten-Sequencing-System/default.aspx
6. Adewale, B. A. (2020). Will long-read sequencing technologies replace short-read sequencing technologies in the next 10 years? African Journal of Laboratory Medicine, 9(1). https://doi.org/10.4102/ajlm.v9i1.1340
7. Poptsova, M. S., Il'icheva, I. A., Nechipurenko, D. Yu., Panchenko, L. A., Khodikov, M. V., Oparina, N. Y., Polozov, R. V., Nechipurenko, Y. D., & Grokhovsky, S. L. (2014). Non-random DNA fragmentation in next-generation sequencing. Scientific Reports, 4(1). https://doi.org/10.1038/srep04532
8. Unique Molecular Identifiers (UMIs) | For sequencing accuracy. (n.d.). Www.illumina.com. Retrieved August 16, 2022, from

https://www.illumina.com/techniques/sequencing/ngs-library-prep/multiplexing/unique-molecular-identifiers.html

9. admin. (2022, May 13). Principle and Workflow of Illumina Next-generation Sequencing | CD Genomics Blog. CD Genomics Blog. https://www.cd-genomics.com/blog/principle-and-workflow-of-illumina-next-generation-sequencing/
10. NGS Workflow Steps | Illumina sequencing workflow. (n.d.). Www.illumina.com. https://www.illumina.com/science/technology/next-generation-sequencing/beginners/ngs-workflow.html
11. Anasagasti, A., Irigoyen, C., Barandika, O., López de Munain, A., & Ruiz-Ederra, J. (2012). Current mutation discovery approaches in Retinitis Pigmentosa. Vision Research, 75, 117–129. https://doi.org/10.1016/j.visres.2012.09.012
12. http://www.facebook.com/andreweditor, & http://www.facebook.com/andreweditor. (2016, July 9). All in the Chip: Ion Torrent Sequencers - Bitesize Bio. Bitesize Bio. https://bitesizebio.com/27399/all-in-the-chip-ion-torrent-sequencers/
13. Illumina. (2016b). An introduction to next-generation sequencing technology. https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf
14. Next Generation Sequencing. (n.d.). Medicine.yale.edu. Retrieved August 16, 2022, from https://medicine.yale.edu/keck/dna/nextgen/iontorrentpgm/
15. Chen, F., Dong, M., Ge, M., Zhu, L., Ren, L., Liu, G., & Mu, R. (2013). The History and Advances of Reversible Terminators Used in New Generations of Sequencing Technology. Genomics, Proteomics & Bioinformatics, 11(1), 34–40. https://doi.org/10.1016/j.gpb.2013.01.003

16. Kircher, M., & Kelso, J. (2010). High-throughput DNA sequencing - concepts and limitations. BioEssays, 32(6), 524–536. https://doi.org/10.1002/bies.200900181
17. Gupta, A. K., & Gupta, U. D. (2014, January 1). Chapter 19 - Next Generation Sequencing and Its Applications (A. S. Verma & A. Singh, Eds.). ScienceDirect; Academic Press. https://www.sciencedirect.com/science/article/pii/B9780124160026000195
18. Torrent Suite Software - US. (n.d.). Www.thermofisher.com. https://www.thermofisher.com/eg/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-workflow/ion-torrent-next-generation-sequencing-data-analysis-workflow/ion-torrent-suite-software.html
19. FASTQ files explained. (2021, October 26). Support.illumina.com. https://support.illumina.com/bulletins/2016/04/fastq-files-explained.html
20. FastQC. (2019). Illumina.com. https://www.illumina.com/products/by-type/informatics-products/basespace-sequence-hub/apps/fastqc.html
21. Ion SphereTM Quality Control Kit. (n.d.). Www.thermofisher.com. Retrieved August 16, 2022, from https://www.thermofisher.com/order/catalog/product/4468656
22. Schbath, S., Martin, V., Zytnicki, M., Fayolle, J., Loux, V., & Gibrat, J.-F. (2012). Mapping Reads on a Genomic Sequence: An Algorithmic Overview and a Practical Comparative Analysis. Journal of Computational Biology, 19(6), 796–813. https://doi.org/10.1089/cmb.2012.0022
23. What are SAM & BAM Files? (2021, November 23). ZYMO RESEARCH. https://www.zymoresearch.com/blogs/blog/what-are-sam-and-bam-files

24. Ye, S., Dhillon, S., Ke, X., Collins, A. R., & Day, I. N. M. (2001). An efficient procedure for genotyping single nucleotide polymorphisms. Nucleic Acids Research, 29(17), e88. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC55900/

25. Shen, Y., Wan, Z., Coarfa, C., Drabek, R., Chen, L., Ostrowski, E. A., Liu, Y., Weinstock, G. M., Wheeler, D. A., Gibbs, R. A., & Yu, F. (2009). A SNP discovery method to assess variant allele probability from next-generation resequencing data. Genome Research, 20(2), 273–280. https://doi.org/10.1101/gr.096388.109

26. Kumar, S., Banks, T. W., & Cloutier, S. (2012). SNP Discovery through Next-Generation Sequencing and Its Applications. International Journal of Plant Genomics, 2012, 1–15. https://doi.org/10.1155/2012/831460

27. Maurer, I. (2020, April 9). What is a Variant Call Format (VCF) file? Precision Oncology Solutions | GenomOncology. https://www.genomoncology.com/blog/what-is-a-variant-call-format-vcf-file

28. Piper, M. M., Mary. (2016, April 28). Variant annotation with snpEff. In-Depth-NGS-Data-Analysis-Course. https://hbctraining.github.io/In-depth-NGS-Data-Analysis-Course/sessionVI/lessons/03_annotation-snpeff.html

29. NGS - Data Analysis | ABM Inc. (n.d.). Old.abmgood.com. Retrieved August 16, 2022, from https://old.abmgood.com/marketing/knowledge_base/next_generation_sequencing_data_analysis.php#13

Figure legend

Figure 1: shows the general workflow of NGS described in 5 steps: Fragmentation, Adaptor ligation, Amplification, Sequencing, Data analysis

Figure 2: shows the general approach for NGS data analysis described in 5 steps: Data quality control and reprocessing, Reads mapping to a reference genome, Identification of variants, Annotation of variants, Visualization of data.