Next Generation Sequencing Technologies, Bioinformatics and Artificial Intelligence: A Shared Timeline

Martina Elena Tarozzi, PhD
Indipendent researcher. Florence, Tuscany, Italy

https://doi.org/10.57098/SciRevs.Biology.3.2.2
Received June 10, 2024. Revised June 25, 2024. Accepted June 26, 2024

Abstract: This review provides a comprehensive overview of the fast-paced and intertwined evolution of three pivotal fields: next-generation sequencing (NGS) technologies, bioinformatics, and artificial intelligence (AI). The paper begins by tracing the development of sequencing technologies and highlights how advancements in genetic sequencing have led to an explosion of biological data, necessitating the rise of bioinformatics for data management and analysis. The review next covers the primary steps and methods used in bioinformatic analysis and concludes by reporting some of the technical and biological challenges in which AI methods have been applied.

Introduction

Determining the order of nucleic acids in polynucleotide molecules and its functional meaning has been a major biological question since the discovery of the molecular structure of DNA in 1953 (1). Only twenty-four years later in 1977, the first method for DNA sequencing was published by Fredrick Sanger and colleagues, who developed the "chain-termination" or dideoxy technique (2). Improvement of this groundbreaking method represented the first generation of DNA sequencing technologies, which produced reads nearly one kilobase (kb) in length. The development of additional techniques such as polymerase chain reaction (PCR) (3) in 1983 and recombinant DNA technologies provided the means for generating high quantities of DNA required by first-generation technologies, triggering the genomic revolution and ultimately the first draft of the human genome in 2001(4). A pivotal turning point was achieved in 2005 with the advent of Next Generation Sequencing (NGS) technologies (5), allowing for the massive and parallel sequencing of whole genomes. In the last decade, sequencing technologies have expanded to include methods for RNA sequencing (6,7), giving rise to the transcriptomic field and to methods for unveiling the structural features and environmental-mediated modifications of chromatin and DNA (8), the epigenomics field, and the single-cell omics technologies starting in 2009 (9).

Artificial intelligence (AI) was born and rapidly improved within this same time frame (Figure 1). Here again, the initial milestones date back to the early 1950s, with the Turing test and the first use of the term "Artificial Intelligence" at the Dartmouth Conference by John McCarthy (10). In the following three decades, the IT sector released groundbreaking innovations, such as the Internet and World Wide Web, as well as algorithms that laid the foundation for deep learning, like Convolutional Neural Networks (CNN), Support Vector Machine (SVM) and Backpropagation (11–14). In less than a decade, the AI Deep Blue was able to outperform human beings in complex tasks such as playing chess. Starting in 2001, the conjunction of increasingly powerful computational resources (storage space and processing speed) (15) and the biotechnological development that led to NGS, increased the potential for tandem use of AI and biology. NGS enabled a massive increase in omics data production, necessitating the development of computational methods able to handle such data. The complexity of biological processes and data provided opportunities and challenges that machine learning techniques are well suited to solve. Consequently, starting from the early 2000s, computational biology became an increasingly relevant field.

As molecular biology becomes more data-intensive and AI algorithms better able to handle biological complexity, the interconnection between these fields is bound to strengthen. In this review, we cover the main technological features of NGS technologies and bioinformatic analysis and provide an overview of current applications of AI on sequencing data.

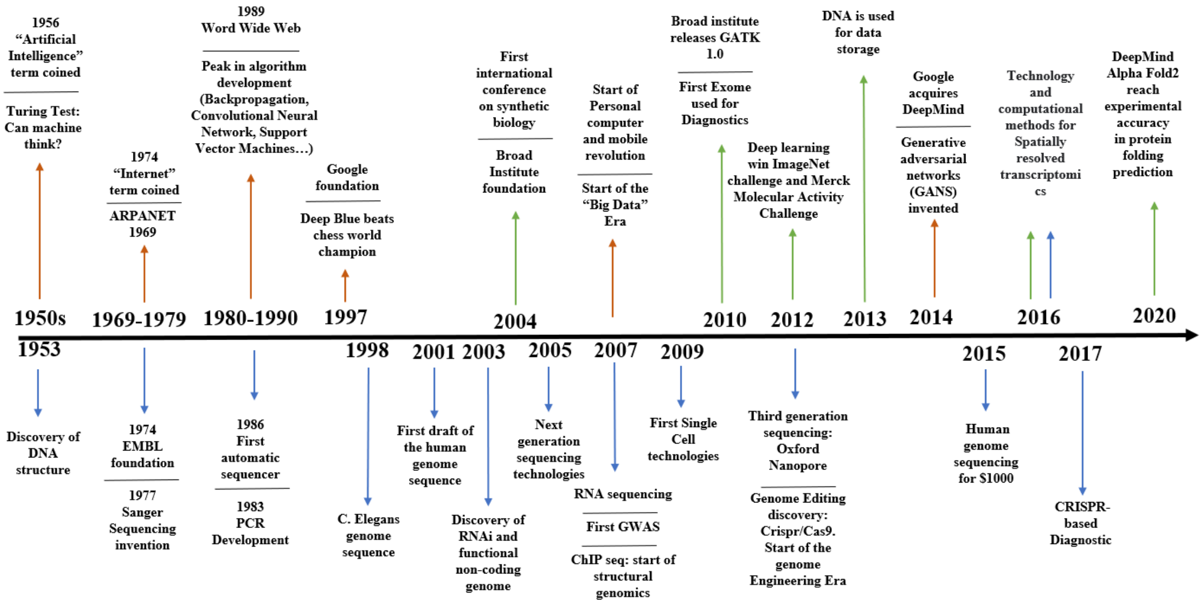

Figure 1: *Timeline of improvement milestones in genomics and sequencing technologies (blue arrows), informatics and artificial intelligence (orange arrows) and computational biology (green arrows).*

Overview of illumina technology and sequencing assays

NGS refers to modern high-throughput sequencing technologies that can be applied to DNA or RNA. Illumina platforms are the NGS technology most frequently used in research and clinical settings, and will therefore be the focus in this paper.

The sequencing workflow starts with library preparation, which varies between different omics. In genomic workflows, DNA is fragmented either mechanically, enzymatically, or with transposons in fragments of appropriate length (typically around 400bp). Next, blunt ends at both ends are repaired: typically, 5′ ends are phosphorylated and 3′ ends are repaired with Adenine residues. Subsequently, adapters are ligated to both ends. Depending on the experimental design, it is then possible to select genome regions of interest via enrichment (such as the exome or a more restricted selection of genes in target sequencing) or retain the entire genome.

RNA sequencing is a versatile high throughput sequencing technique introduced in 2008 that allows for the investigation of gene expression, as well as alternative splicing , allele-specific expression , variation in linear nucleotide sequence, novel transcript expression and gene fusion events . The appropriate library preparation protocol for RNA sequencing must be considered based on the study's questions in order to reduce bias . There are four main categories of possible library preparations: i) total RNAseq: all structural, regulatory, coding and non-coding RNAs are sequenced, ii) RNAseq with ribosomal RNA reduction: only rRNAs useful for phylogenic reconstruction are kept together with regulatory and coding RNAs, iii) cDNA capture: only coding RNAs are enriched using probes targeting exon sequences, iv) polyA selection: only mature mRNAs are isolated through poly-T probes that bind the 3′ poly-A tail of mRNAs.

The third main omic in terms of sequencing of nucleic acids is epigenomics. Gene expression can strongly depend on epigenetic regulation in terms of changes in the accessibility of certain genomic regions through DNA methylation, histone modifications, three-dimensional chromatin organization, and post-transcriptional regulation mechanisms. Each type of epigenetic modification can be measured via an experimental procedure performed before the library preparation described above. A thorough overview of these complex techniques can be found in Mehrmohamadi et al, 2021[24].

There are four main categories of epigenomic assays: i) DNA methylation: normally investigated with bisulfite-conversion-based libraries, where only unmethylated cytosines are converted to uracil (23) ii) histone modifications: mostly studied with chromatin immunoprecipitation assays (ChIP-seq), where crosslinked DNA is treated with antibodies targeting the histone modification of interest and pulled down (24), iii) chromatin accessibility: studied with assays based on the fact that open chromatin is more accessible to fragmentation agents, like digestion enzymes (e.g. DNase-seq) or transposase (e.g. ATAC-seq)(25), iv) 3D organization: mostly investigated with Chromatin Conformation Capture (3C) derived assays (such as 4C-seq, Hi-C, ChIA-Drop), whereby nuclei DNA is crosslinked, chimeric DNA molecules made of genomic regions close to one another are formed, and the proximal genomic regions in the 3D space of the nucleus are measured using pairwise frequencies between genomic loci (26).

Once the library is loaded into the sequencer, it is pumped onto a flow-cell where each single-stranded fragment hybridizes to flow-cell adapters on both ends forming a "bridge" structure that confers the name to this step called "bridge amplification" (27). After cluster generation, sequencing by synthesis with reversible chain terminators and a fluorophore corresponding to each of the four nucleotides (A, C, T, and G), or bases, takes place. The fluorophore wavelength together with its intensity determines the base call. The acquired optical signals for each lane of the flow cell are converted into base call format files (bcl file) that represent the output of sequencing and the first raw input for bioinformatic analysis (Figure 2).

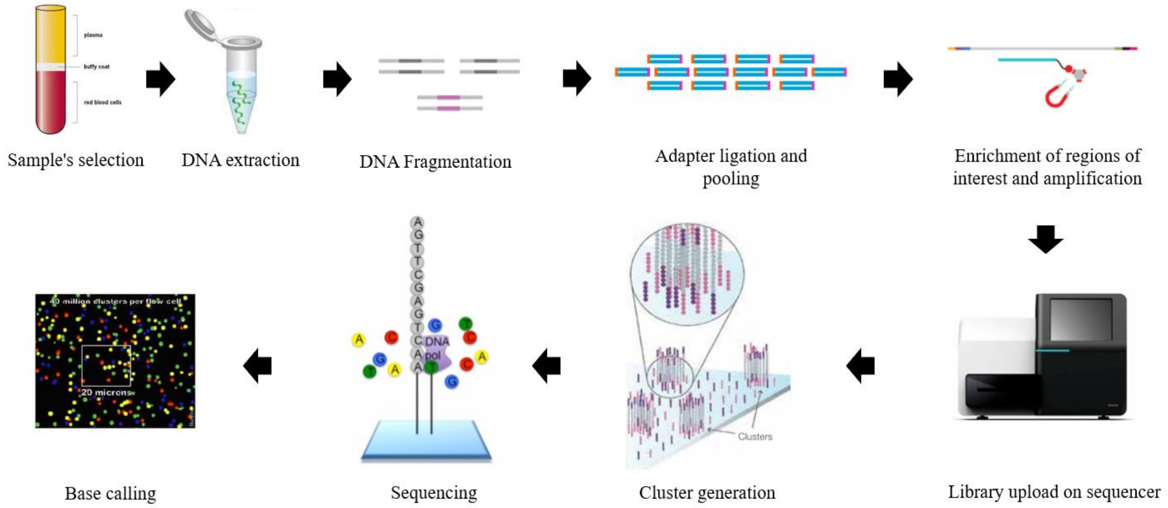

Figure 2: *Schematic representation of a targeted genomic library preparation workflow and Illumina sequencing reaction. Figure adapted from images courtesy of www.illumina.com (https://www.illumina.com/science/technology/next-generation-sequencing.html, https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf).*

Bioinformatic analysis

Bioinformatic analysis involves several steps of computationally intensive data transformations, each requiring specific tools. Substantial research has been done to develop reliable software capable of performing such data transformations and to make the computational analysis of NGS data reproducible. For example, package and environment management systems such as CONDA allow the installation and management of software developed in different languages for different operating systems.

3.1. *Primary Data Analysis*

The primary analysis of Illumina raw data consists in demultiplexing bcl data into FASTQ files. This process transforms the binary raw files obtained from the optic signal acquired during sequencing into a text file in which the nucleotide sequence is matched by a PHRED score defining the quality of the nucleotide, or base, call made by the sequencer. During this step, index sequences contained in each fragment are used to group all calls belonging to the same sample into a single FASTQ file. Depending on the sequencer, this step is performed directly on-instrument by Real-Time Analysis software or afterwards as first step of a command line pipeline (28).

3.2 *Secondary Data Analysis*

Secondary analysis of genomic data is typically a standardized workflow used in diagnostics and research . It generally consists of three main steps: mapping onto the reference genome, post-alignment processing and, if useful, variant calling. Workflow management systems such as Snakemake and Nextflow address the need to concatenate the steps required by bioinformatic pipelines while providing an efficient usage of computational power through process parallelization. The sample's sequence is expressed in FASTQ files containing the nucleotide sequence and quality scores derived from millions of short reads. Quality checks are typically performed first in secondary analysis. One of the most common tools for this task is FASTQC . Adapter reads and base calls with low quality scores are trimmed from the sequence using specific tools such as Trimmomatic , fastp or cutadapt . Mapping algorithms are then used to identify a location in the reference genome that matches the experimental read generated via sequencing. Software settings allow for varying degrees of tolerance towards base mismatches and extra spaces to allow for the detection of possible variants. One commonly used tool for mapping NGS sequences to a reference genome is the Burrows-Wheeler Alignment (BWA). The output of mapping algorithms are stored as sam (sequence alignment map) files, which contain all the information surrounding the mapping procedure in the metadata section, along with data concerning the mapped genomic region and mapping. The binary counterpart to a sam file is a bam file, which represents the type of data that will undergo further post-alignment processing and variant calling. Post-alignment processing consists of sorting, marking duplicates and indexing the bam file. The variant calling step aims to identify single nucleotide variants and small insertions and deletions, reported in Variant Call Format (VCF) files.

3.3 *Tertiary Data Analysis*

Tertiary analysis of genomic data uses the list of variants reported in the VCF file to biologically interpret the data. This component of data analysis is highly adaptable depending on the experimental question. The first and most common step is variant annotation, aimed at obtaining functional information about the type of nucleotide substitution and its predicted effect. A more classic approach is to consider only the variants that affect the primary structure of the coded protein, like missense or truncating variants, especially if in-silico variant predictor tools describe them as likely pathogenic. Functional analysis on groups of significant genes is often performed with network approaches or over-representation methods. In recent years, data science and machine learning methods applied to genomic data have been a resourceful approach for acquiring a more complete understanding of polygenic contributions in complex diseases (39) and in the context of precision medicine (40,41). Some of these applications will be further discussed section 5.

5. *Applications of ai in genomics and transcriptomics*

ML and DL methods have been applied to sequencing data covering various research scopes and topics. Here, we describe some of the most relevant fields of application, focusing first on the use of ML and DL for technical issues associated with NGS data processing and analysis, and then with examples of how these methods are used to explore open biological questions. This section aims at providing common, promising or exemplifying applications of AI methods on NGS data in biology and bioinformatics and should not be considered a complete overview of all its possible applications.

5.1 *Applications on technical problems*

Secondary bioinformatic analysis: variant calling

The accurate discovery of single nucleotide variants from billions of short reads remains a challenging step in bioinformatics because library preparation, sequencing and data processing tools are error-prone procedures. These issues become even more apparent when the object of study are low-frequency somatic mutations or when the input DNA is of lower quality. Most variant callers use statistical methods (such as logistic regression, hidden Markov models, naïve Bayes) to model error sources and to distinguish whether differences between experimental reads and the reference genome are caused by true genetic variants or errors. In recent years, deep learning has been applied to address variant calling on NGS data: a common approach is to address the problem as one of image recognition, where a Deep Neural Network analyzes sequencing data that are transformed as images of read pileups of true genotype calls to compute the genotype likelihoods at each locus. Two of the first and most popular tools of this kind are DeepVariant (42) and DeNovoCNN (43), with the latter specifically used to address the identification of *de novo* mutations. Both tools showed higher accuracy compared to classical methods. An alternative approach is presented in HELLO (44), whereby comparable performances are obtained by designing Deep Neural Networks that examine aligned reads to predict the status (ref or alt) of each candidate allele given the support for that allele in relation to the support for the remaining alleles at the genomic site.

Tertiary bioinformatic analysis: Variant effect prediction

Variant Effect Prediction (VEP) are computational tools that provide a prediction about the functional significance of a single nucleotide variant (SNV). The growing use of NGS technologies for advanced diagnostics has increased the need to better classify variants of uncertain significance. VEPs rely on different types of prior knowledge, such as protein sequence and structural information, evolutionary sequence conservation, functional experiments, epigenomic data and association studies to produce an effect score for the variant. In supervised VEPs, the algorithm is trained on a set of labelled SNVs known as benign or dam2aging according to previous knowledge to perform a classification task. Using this prior knowledge, these methods compute a score expressing the predicted effect of the variant. Examples of well-performing supervised VEPs on human samples are SNP&GO (45), PolyPhen2 (46) and DEOGEN2 (47). Unsupervised methods do not use any labelled data and usually rely exclusively on the evolutionary conservation of the genomic locus. This group also includes deep learning methods, like DeepSequence (48), considered by a recent benchmark study as the top-performing tool among 46 tested in deep mutation scanning data (49). An example of a semi-supervised deep learning method is the Illumina PrimateAI (50), which has performed well in the study of rare diseases.

Visualization of high dimensional datasets

NGS data are highly dimensional because each sample is sequenced simultaneously. The huge amount of information contained in these data can represent an obstacle to the identification of its most meaningful features. Dimensionality reduction techniques such as PCA, t-SNE and UMAP are used to identify latent components in the data that are not easily accessible due to the high number of variables. Data are thus transformed into a lower dimensionality while maintaining the relationships between data points (e.g., samples) as much as possible. These methods are extremely versatile. For example, they can be used in the pre-processing of bulk RNA-seq data to identify possible outliers and relevant covariates(51), to search for recurrent patterns on targeted DNA sequencing data in different classes of samples (52), or to visualize single-cell RNA sequencing data. In this context, dimensionality reduction techniques coupled with clustering algorithms are used for cell-type identification tasks, identifying groups of cells that share similar expression profiles. Another application of these methods is lineage trajectory inference(53), which involves the reconstruction of the position of each individual cell on the lineage trajectory based on scRNA-seq profiles with different time points, allowing for the study of dynamic processes such as the cell cycle, cell differentiation and cell activation.

6. *Future Perspectives*

In this review, we summarized the crucial aspects and timeline of NGS technologies, bioinformatics and AI, highlighted how they are connected in a holistic process, and explained the potential revolutionary insights that can be gained from their concurrent use. Computational and molecular biology have and are continuing to advance at an impressive pace. Machine and deep learning, while relatively recent breakthroughs in the biological and biomedical fields, will almost certainly play an increasing role. Substantial investments are being made by leading technology companies that, together with academic researchers, are implementing innovative methodologies, software and architectures tailored specifically to answer biological questions. Concurrently, we are witnessing the improvement of molecular techniques while sequencing experiments are evolving and becoming more affordable, as evident by the growing interest in long-read sequencing. Taken together, these innovations are a new frontier of research and have the potential to strongly affect our ability understand and interact with genetic information.

References

1. Watson JD, Crick FHC. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nat 1953 1714356 [Internet]. 1953 Apr 25 [cited 2021 Jul 6];171(4356):737–8. Available from: https://www.nature.com/articles/171737a0

2. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A [Internet]. 1977 [cited 2021 Jul 6];74(12):5463. Available from: /pmc/articles/PMC431765/?report=abstract

3. Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H. Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction. Cold Spring Harb Symp Quant Biol. 1986;51(1):263–73.

4. JC V, MD A, EW M, PW L, RJ M, GG S, et al. The sequence of the human genome. Science [Internet]. 2001 Feb 16 [cited 2021 Jul 6];291(5507):1304–51. Available from: https://pubmed.ncbi.nlm.nih.gov/11181995/

5. Heather J, Chain B, Heather JM, Chain B. The Sequence of Sequencers : The History of Sequencing DNA Genomics The sequence of sequencers : The history of sequencing DNA. Genomics [Internet]. 2015;107(1):1–8. Available from: http://dx.doi.org/10.1016/j.ygeno.2015.11.003

6. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. Science [Internet]. 2008 Jun 6 [cited 2021 Jul 19];320(5881):1344. Available from: /pmc/articles/PMC2951732/

7. SJ E, WB B, L L, PS S. Gene discovery and annotation using LCM-454 transcriptome sequencing. Genome Res [Internet]. 2007 Jan [cited 2021 Jul 6];17(1):69–73. Available from: https://pubmed.ncbi.nlm.nih.gov/17095711/

8. JFriedman N, Rando OJ. Epigenomics and the structure of the living genome. Genome Res. 2015;25(10):1482–90.

9. Wang D, Bodovitz S. Single cell analysis: the new frontier in 'Omics.' Trends Biotechnol [Internet]. 2010 Jun [cited 2021 Jul 6];28(6):281. Available from: /pmc/articles/PMC2876223/

10. McCarthy J, Minsky ML, Rochester N, Shannon CE. A proposal for the Dartmouth summer research project on artificial intelligence [Internet]. Vol. 27, AI Magazine. 2006 [cited 2022 Oct 12]. p. 12–4. Available from: http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

11. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, et al. Backpropagation applied to digit recognition [Internet]. Vol. 1, Neural computation. 1989. p. 541–51. Available from: https://www.ics.uci.edu/~welling/teaching/273ASpring09/lecun-89e.pdf

12. Lecun Y, Bottou E, Bengio Y, Haffner P. Gradient-Based Learning Applied to Document Recognition. 1998.

13. Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. Psychol Rev. 1958;65(6):386–408.

14. Boser BE, Guyon IM, Vapnik VN. Training algorithm for optimal margin classifiers. Proc Fifth Annu ACM Work Comput Learn Theory. 1992;144–52.

15. Muir P, Li S, Lou S, Wang D, Spakowicz DJ, Salichos L, et al. The real cost of sequencing: Scaling computation to keep pace with data generation. Genome Biol [Internet]. 2016 Mar 23 [cited 2023 Apr 28];17(1):1–9. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0917-0

16. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, et al. Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. Cell [Internet]. 2008 May 2 [cited 2021 Jul 19];133(3):523. Available from: /pmc/articles/PMC2723732/

17. U N, Z W, K W, C S, D R, M G, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science [Internet]. 2008 Jun 6 [cited 2021 Jul 6];320(5881):1344–9. Available from: https://pubmed.ncbi.nlm.nih.gov/18451266/

18. Mehmood A, Laiho A, Venäläinen MS, McGlinchey AJ, Wang N, Elo LL. Systematic evaluation of differential splicing tools for RNA-seq studies. Brief Bioinform [Internet]. 2020 Dec 1 [cited 2021 Jul 20];21(6):2052–65. Available from: https://academic.oup.com/bib/article/21/6/2052/5648232

19. Castel SE, Levy-Moonshine A, Mohammadi P, Banks E, Lappalainen T. Tools and best practices for data processing in allelic expression analysis. Genome Biol [Internet]. 2015;16(1):1–13. Available from: http://dx.doi.org/10.1186/s13059-015-0762-6

20. Hutchins AP, Poulain S, Fujii H, Miranda-Saavedra D. Discovery and characterization of new transcripts from RNA-seq data in mouse CD4+ T cells. Genomics. 2012 Nov 1;100(5):303–13.

21. Haas BJ, Dobin A, Stransky N, Li B, Yang X, Tickle T, et al. STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. bioRxiv [Internet]. 2017 Mar 24 [cited 2021 Jul 20];120295. Available from: https://www.biorxiv.org/content/10.1101/120295v1

22. Van Dijk EL, Jaszczyszyn Y, Thermes C. Library preparation methods for next-generation sequencing: Tone down the bias. Exp Cell Res. 2014 Mar 10;322(1):12–20.

23. Mehrmohamadi M, Sepehri MH, Nazer N, Norouzi MR. A Comparative Overview of Epigenomic Profiling Methods. Front Cell Dev Biol. 2021;9(July):1–14.

24. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res [Internet]. 2012 Sep 1 [cited 2022 Oct 27];22(9):1813–31. Available from: https://genome.cshlp.org/content/22/9/1813.full

25. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr Protoc Mol Biol [Internet]. 2015 [cited 2022 Oct 27];109:21.29.1. Available from: /pmc/articles/PMC4374986/

26. Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science (80-) [Internet]. 2009 Oct 9 [cited 2021 Jul 5];326(5950):289–93. Available from: /pmc/articles/PMC2858594

27. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate Whole Human Genome Sequencing using Reversible Terminator Chemistry. Nature [Internet]. 2008 Nov 11 [cited 2022 Oct 18];456(7218):53. Available from: /pmc/articles/PMC2581791/

28. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med [Internet]. 2015 May 8 [cited 2021 Mar 31]; 17(5):405–24. Available from: /pmc/articles/PMC4544753/

29. Alonso CM, Llop M, Sargas C, Pedrola L, Panadero J, Hervás D, et al. Clinical Utility of a Next-Generation Sequencing Panel for Acute Myeloid Leukemia Diagnostics. J Mol Diagn [Internet]. 2019 Mar 1 [cited 2022 Oct 13];21(2):228–40. Available from: https://pubmed.ncbi.nlm.nih.gov/30576870/

30. Di Resta C, Galbiati S, Carrera P, Ferrari M. Next-generation sequencing approach for the diagnosis of human diseases: open challenges and new opportunities. EJIFCC [Internet]. 2018 Apr 1 [cited 2022 Oct 13];29(1):4. Available from: /pmc/articles/PMC5949614/

31. Bartoletti-Stella A, Tarozzi M, Mengozzi G, Asirelli F, Brancaleoni L, Mometto N, et al. Dementia-related genetic variants in an Italian population of early-onset Alzheimer's disease. Front Aging Neurosci. 2022;14(September):1–13.

32. Köster J, Rahmann S. Snakemake-a scalable bioinformatics workflow engine. Bioinformatics. 2012.

33. DI Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol 2017 354 [Internet]. 2017 Apr 11 [cited 2023 Mar 2];35(4):316–9. Available from: https://www.nature.com/articles/nbt.3820

34. Andrews, Simon, Krueger, Felix , Segonds-Pichon, Anne , Biggins, Laura , Krueger, Christel , Wingett S. FastQC [Internet]. Babraham, UK. 2010. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

35. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics [Internet]. 2014 Aug 1 [cited 2021 Nov 17];30(15):2114–20. Available from: https://academic.oup.com/bioinformatics/article/30/15/2114/2390096 metanalysis of machine and deep learning-based CRISPR gRNA design tools. RNA Biol. 2020; 17:13

36. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics [Internet]. 2018 Sep 1 [cited 2023 Jan 10];34(17):i884–90. Available from: https://academic.oup.com/bioinformatics/article/34/17/i884/5093234

37. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal [Internet]. 2011 May 2 [cited 2023 Jan 10];17(1):10–2. Available from: https://journal.embnet.org/index.php/embnetjournal/article/view/200/479

38. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;

39. Ho DSW, Schierding W, Wake M, Saffery R, O'Sullivan J. Machine learning SNP based prediction for precision medicine. Front Genet. 2019;10(MAR):267.

40. Dias R, Torkamani A. Artificial intelligence in clinical and genomic diagnostics. Genome Med. 2019;11(1):1–12.

41. Álvarez-Machancoses Ó, Galiana EJD, Cernea A, de la Viña JF, Fernández-Martínez JL. On the role of artificial intelligence in genomics to enhance precision medicine. Pharmgenomics Pers Med. 2020;13:105–19.

42. Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal snp and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36(10):983.

43. Khazeeva G, Sablauskas K, van der Sanden B, Steyaert W, Kwint M, Rots D, et al. DeNovoCNN: a deep learning approach to de novo variant calling in next generation sequencing data. Nucleic Acids Res. 2022;50(17):e97.

44. Ramachandran A, Lumetta SS, Klee EW, Chen D. HELLO: improved neural network architectures and methodologies for small variant calling. BMC Bioinformatics [Internet]. 2021;22(1):1–31. Available from: https://doi.org/10.1186/s12859-021-04311-4

45. Capriotti E, Calabrese R, Fariselli P, Martelli PL, Altman RB, Casadio R. WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation. BMC Genomics [Internet]. 2013;14 Suppl 3(Suppl 3):S6. Available from: http://www.biomedcentral.com/1471-2164/14/S3/S6

46. Adzhubei I, Jordan DM, Sunyaev SR. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. Curr Protoc Hum Genet [Internet]. 2013 [cited 2021

47. Raimondi D, Tanyalcin I, FertCrossed JSD, Gazzo A, Orlando G, Lenaerts T, et al. DEOGEN2: Prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. Nucleic Acids Res. 2017;45(W1):W201–6.

48. Riesselman AJ, Ingraham JB, Marks DS. Deep generative models of genetic variation capture the effects of mutations. Nat Methods [Internet]. 2018;15(10):816–22. Available from: http://dx.doi.org/10.1038/s41592-018-0138-4

49. Livesey BJ, Marsh JA. Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. Mol Syst Biol [Internet]. 2020 Jul 1 [cited 2022 Oct 26];16(7):e9380. Available from: https://onlinelibrary.wiley.com/doi/full/10.15252/msb.20199380

50. Sundaram L, Gao H, Padigepati SR, McRae JF, Li Y, Kosmicki JA, et al. Predicting the clinical impact of human mutation with deep neural networks. Nat Genet 2018 508 [Internet]. 2018 Jul 23 [cited 2022 Oct 27];50(8):1161–70. Available from: https://www.nature.com/articles/s41588-018-0167-z

51. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17(1):1–19.

52. Tarozzi M, Bartoletti-Stella A, Dall'Olio D, Matteuzzi T, Baiardi S, Parchi P, et al. Identification of recurrent genetic patterns from targeted sequencing panels with advanced data science: a case-study on sporadic and genetic neurodegenerative diseases. BMC Med Genomics 2022 151 [Internet]. 2022 Feb 10 [cited 2022 Feb 25];15(1):1–12. Available from: https://bmcmedgenomics.biomedcentral.com/articles/10.1186/s12920-022-01173-4

53. Saelens W, Cannoodt R, Todorov H, Saeys Y. A comparison of single-cell trajectory inference methods. Nat Biotechnol [Internet]. 2019;37(5):547–54. Available from: http://dx.doi.org/10.1038/s41587-019-0071-9