Next Generation DNA Sequencing Approaches and Genomic Data Analysis

Kshatrapal Singh[1*[0000-0002-5965-0783]], Raja Sarath Kumar Boddu[2]

Primary affiliation
[1, 2,]School of Computer Engineering, Lincoln University College, Petaling Jaya, 47810, Malaysia
pdf.kshatrapal@lincoln.edu.my
Secondary affiliation
[1]Department of CSE, KCC Institute of Technology and Management, Greater Noida, 201308, India,
mekpsingh1@gmail.com
[2]Department of AI&ML, Raghu Engineering College, Vishakhapatnam, 531162, India
iamrajaboddu@gmail.com

Abstract

By enabling quick, high-throughput analysis of DNA and RNA, next-generation sequencing (NGS) has revolutionized genetics and significantly advanced a number of sectors, including personalized medicine, rare disease diagnosis, and cancer research. Sequence alignment, variant calling, and data quality control are important elements in the NGS workflow that are provided by both open-source and commercial applications. The massive datasets produced by NGS technology may now be stored, managed, and processed more easily thanks to cloud-based platforms. Ethical issues, particularly those pertaining to informed consent and the privacy of genomic data, continue to be crucial factors as NGS develops. In the future, single-cell sequencing and the integration of multi-omics data could improve our comprehension of intricate biological processes. Important databases for analyzing NGS results and their clinical implications include dbSNP, COSMIC, and The Cancer Genome Atlas. Researchers can produce new genetic insights with broad significance for improving human health and comprehending disease mechanisms by skillfully utilizing these techniques and databases.

Keywords

Bioinformatics, Genomics, Next-generation sequencing, Data analysis, NGS methodology.

1. Introduction

The majority of biological and medical fields are starting to change as a result of the Human Genome Project's (HGP) exponential acceleration of technological advancements in genomics and genetics. Personalized medicine and systems biology are now accessible [1]. The Precision Medicine Initiative has emerged as the 21st-century strategy for the prevention and treatment of human disease, and whole genome sequencing (WGS) is not restricted to a select few. The potency of genomic technology has also been accepted by anthropology and evolution [2]. Less than 40 years ago, chain-terminating inhibitors were employed to efficiently and accurately sequence DNA, setting us on the current road in genetics and genomics. The velocity of discovery since the middle of the last century is remarkable. Human gene mapping and gene discovery saw tremendous conceptual and technological advancements in the ensuing decades, culminating in the HGP and the initial draft of the 3 billion-letter human genome [3]. By comparing a disease sample with a standard reference genome, the HG represents a significant advancement in biomedicine by making it possible to identify potential genetic abnormalities.

Sydney Brenner's team published the first massively parallel sequencing method, which produced gene expression profiles from yeast and a human cell line using microbeads [4]. The demand for low-cost, repeatable, high-throughput nucleic acid sequence data creation has been met by more popular next-generation sequencing (NGS) techniques, which have sparked a revolution in the biomedical sciences [5]. NGS has given biologists and clinicians access to the benefits and challenges of "big data science" for the genome wide assessment of genetic variants, expression of unique RNA species, and epigenetic modifications linked to development, aging, and illness. Nowadays, high throughput cellular molecule cataloging and/or analysis is referred to as "omics." We are making progress in identifying every functional genomic element and comprehending the function of non-coding variations in tissue-specific settings [6]. Systems-level methods for quantitative study of the dynamic interaction of molecules within a particular cell or tissue have been made possible by a significant increase in genomic, transcriptomic, and epigenomic data. Additionally, NGS-based methods have rapidly become widely applicable in medicine, ranging from pharmacogenomics and drug development to genetic diagnostics and disease networks. Our distinct identities are defined by the millions of genetic variants that each of us possesses [7]. Some of these polymorphisms are linked to human characteristics or clinically recognized disorders, but the majority do not appear to have a clear pathogenic consequence. In Section 2, we will go over the fundamentals of NGS technology and go into more detail on the Illumina platform, which is a popular tool among genomic scientists.

2. Bioinformatic Techniques for Next Generation Sequencing Data Analysis

Large volumes of DNA or RNA sequences are produced by NGS, and handling, analyzing, and interpreting this data requires computational techniques. To obtain biological insights, the raw sequencing data generated by NGS equipment must be processed, examined, and evaluated. Bioinformatic methods are useful in this situation. Preprocessing, alignment, variant calling, gene expression quantification, differential expression analysis, and other specific studies are all included in this broad category of computational techniques, algorithms, and tools. Once processed, several computational techniques, like as de novo assembly, reference-based mapping, and transcriptome analysis, are applied to extract useful biological information [8]. Additionally, the detection of genetic variations, such as single-nucleotide polymorphisms (SNPs), copy number variations (CNVs), and structural variants, is made easier by sophisticated bioinformatic techniques. Integrative analyses, combining NGS data with other genomic and functional data sources, enabling the analysis of gene expression and regulatory networks [9].

NGS has become a popular genomic technology in less than ten years due to increased sequencing throughput and significant cost savings. NGS technology has advanced significantly since the first commercially accessible device with a throughput of 20 megabase pairs (Mbp) per run was released [10]. The current Illumina HiSeq X system can generate 1.8 terabase pairs (Tbp) of sequencing data each run, which is almost a 100,000-fold increase in just ten years [11]. Single molecule sequencers, such as Pacific Biosciences' RS and Nanopore's minION, are among the emerging NGS systems that can offer high read lengths and resolution of DNA alterations. The juggernaut of the industry are the sequence-by-synthesis (SBS) systems from Illumina that tout a wide range of applications, relative ease of use, numerous levels of throughput, flexibility in setup, and comparatively low sequencing cost [12]. The next-generation sequencing workflow includes the following steps: Sample preparation, library preparation, sequencing, and data analysis (Figure 1).

Figure 1: NGS workflow steps

3. Applications and Benefits of Next Generation Sequencing

Next-generation sequencing technology has fundamentally changed the kinds of questions scientists can ask and answer. Innovative sample preparation and data analysis options enable a broad range of applications. For example, NGS allows us to:

- Rapidly sequence whole genomes
- Deeply sequence target regions
- Utilize RNA sequencing (RNA-Seq) to discover novel RNA variants and splice sites, or quantify mRNAs for gene expression analysis
- Analyze epigenetic factors such as genome-wide DNA methylation and DNA-protein interactions
- Sequence cancer samples to study rare somatic variants, tumor subclones, and more
- Study the human microbiome
- Identify novel pathogens

NGS-based research currently focuses on the transcriptome, epi-transcriptome, and epigenome in addition to genomic research. Human genome-based research using WGS and WES has yielded new insights into biological processes and has been applied in wellness research, agriculture and food research, genome-wide association research studies revealing a wide range of population genetic variants, their molecular basis and genetic linkage to various diseases, including cancer, and the study of new pathogenic/emerging variants like SARS-CoV-2 variants in human diseases. The ever-expanding list of targeted big gene panels, including the 261 gene panel, the 400 gene panel, and the TSO 500 panel from Illumina, IDT, Agilent, and Thermo Fisher, has made it possible to translate the redefining of the mutational landscapes in cancers into clinical research [13]. In addition to SNVs, these panels evaluate clinically significant CNVs, RNA fusion transcripts, TMB, and microsatellite instability (MSI) for lung, breast, and colorectal cancers as well as challenging malignancies such ovarian, pancreatic, renal, and urothelial cancers.

The method of syndromic testing/multiple pathogen testing assays like BioFire or multiplex PCRs has been motivated by the need for precision medicine to identify the precise etiological agent in microbial diseases. Nevertheless, NGS panels that can identify any pathogen utilizing a shotgun or focused approach from different sick specimens or clinical isolates are being developed in response to the

constraints of multiplexing. These panels can detect drug-resistant mutations, such as antimicrobial and antiviral drug-resistant mutations, in addition to identifying causal organisms. As demonstrated during the COVID-19 outbreak, the valuable information produced by NGS on microbial identification and drug resistance genotyping, such as in MTB, HIV, and SARS-CoV-2, has proven crucial for disease surveillance, disease containment, public health epidemiological studies, policy making, and quick therapeutic interventions.

DNA analysis has become a key investigative tool in forensic science since Sir Alec Jeffreys initially suggested using DNA profiling to differentiate between various samples at a crime scene. In order to answer a variety of criminal problems, NGS has supplanted earlier DNA fingerprinting techniques like restriction fragment length polymorphism (RFLP), mitochondrial DNA, variable number of tandem repeats (VNTR) profiling, and short tandem repeat (STR) typing [14]. Because NGS can produce very accurate, repeatable, and highly sensitive results from severely contaminated and degraded sample qualities encountered in forensic labs, it has rapidly gained prominence in this field.

Following are the benefits of DNA sequencing with NGS:

- Sequences large stretches of DNA in a massively parallel fashion, offering advantages in throughput and scale compared to capillary electrophoresis–based Sanger sequencing
- Provides high resolution to obtain a base-by-base view of a gene, exome, or genome
- Delivers quantitative measurements based on signal intensity
- Detects virtually all types of genomic DNA alterations, including single nucleotide variants, insertions and deletions, copy number changes, and chromosomal aberrations
- Offers flexibility to discover novel DNA variants, scale studies, and sequence multiple samples simultaneously

4. Future Scope and Conclusion

NGS has enormous promise for future developments and applications across a wide range of industries. The advances in bioinformatics, robotics, liquid handling, and nucleic acid preparation will transform NGS sequencing procedures, making them faster and more exact. These next sequencing technologies will scale down to zeptoliters and even a few molecules, requiring less input DNA and reagents. They will also become more portable, which will allow them to be used in diagnostic applications in a variety of sectors, including ecology, agriculture, medicine, and other field-based contexts. Taken together, NGS presents great potential for transformational breakthroughs across numerous fields. NGS has already changed domains such as clinical diagnostics, cancer genomics, and microbial genomics, delivering new insights into the genetic foundations of diseases and driving customized therapeutics. As technology improves, NGS is projected to play a crucial role in areas such as single-cell genomics, long-read sequencing, epigenomics, and multi-omics integration, enabling a deeper understanding of cellular processes, disease mechanisms, and tailored treatment options. NGS's reach will be further expanded by the creation of real-time sequencing and point-of-care applications, enabling quick monitoring and diagnosis in a variety of contexts. Furthermore, the extraction of significant insights from the enormous volume of NGS data created will depend on developments in bioinformatics and data analysis. Advances in robotics, liquid handling, and sample processing will assist the higher order multiplexing, which will allow more samples to be processed in less time and at a lower cost. Improvements in data transfer and storage, as well as quicker and more precise bioinformatic data analysis, will be as significant. NGS will become more widely available and accessible with continued technology advancements and cost reductions, making it easier to incorporate it into everyday clinical practice, research, agribusiness, and environmental investigations. NGS has a bright future ahead of it, one that promises to open up new avenues for

research and spur developments that will significantly affect environmental preservation, agriculture, human health, and other areas.

References

1. Satam, H. et al. Next-generation sequencing technology: current trends and advancements. *Biology* **12**, 997 (2023).
2. Zhang, C. et al. Terminal deoxynucleotidyl transferase: properties and applications. *Eng. Microbiol.* **5**, 100179 (2025).
3. Shanker, V. R. et al. Unsupervised evolution of protein and antibody complexes with a structure-informed language model. *Science* **385**, 46–53 (2024).
4. Subramanian, A. M., Thomson, M. Unexplored regions of the protein sequence-structure map revealed at scale by a library of foldtuned language models. Preprint at *bioRxiv* https://doi.org/10.1101/2023.12.22.573145 (2023).
5. Kshatrapal Singh, Laxman Singh, Vijay Shukla, Yogesh Kumar Sharma, Arun Kumar Rai, "An Advanced Approach for DNA Sequencing and Similarities Analysis on the basis of Groupings of Nucleotide Bases", International Journal of Data Mining and Bioinformatics, ISSN 1748-5673, vol. 29(1/2), pp. 133-149, 2025. https://doi.org/10.1504/IJDMB.2025.143005
6. Hoose, A. et al. DNA synthesis technologies to close the gene writing gap. *Nat. Rev. Chem.* **7**, 144–161 (2023).
7. Fornace, M. E. et al. NUPACK: analysis and design of nucleic acid structures, devices, and systems. Preprint at https://doi.org/10.26434/chemrxiv-2022-xv98l (2022).
8. Costa VG, C.E.J.A.I R, Pedreira. Recent advances in decision trees: An updated survey 2023. 56(5): p. 4765–4800.
9. Robertson, W. E. et al. Creating custom synthetic genomes in *Escherichia coli* with REXER and GENESIS. *Nat. Protoc.* **16**, 2345–2380 (2021).
10. Kshatrapal Singh, Ashish Kumar, Manoj Kumar Gupta, Modified k-string in composition vector method for DNA sequence comparison based on maximum entropy principle, Journal of Interdisciplinary mathematics, ISSN 0972-0502, vol. 23 (1), pp. 31-41, 2020.
11. Nguyen, E. et al. Sequence modeling and design from molecular to genome scale with Evo. *Science* **386**, eado9336 (2024).
12. Wolfsberg, E. et al. Machine-guided dual-objective protein engineering for deimmunization and therapeutic functions. *Cell Syst.* **16**, 101299 (2025).
13. Holston, A. S. et al. Degenerate DropSynth for simultaneous assembly of diverse gene libraries and local designed mutants. Preprint at *bioRxiv* https://doi.org/10.1101/2023.12.11.569291 (2023).
14. Sakamoto, Y.; Sereewattanawoot, S.; Suzuki, A. A new era of long-read sequencing for cancer genomics. *J. Hum. Genet.* **2020**, *65*, 3–10.