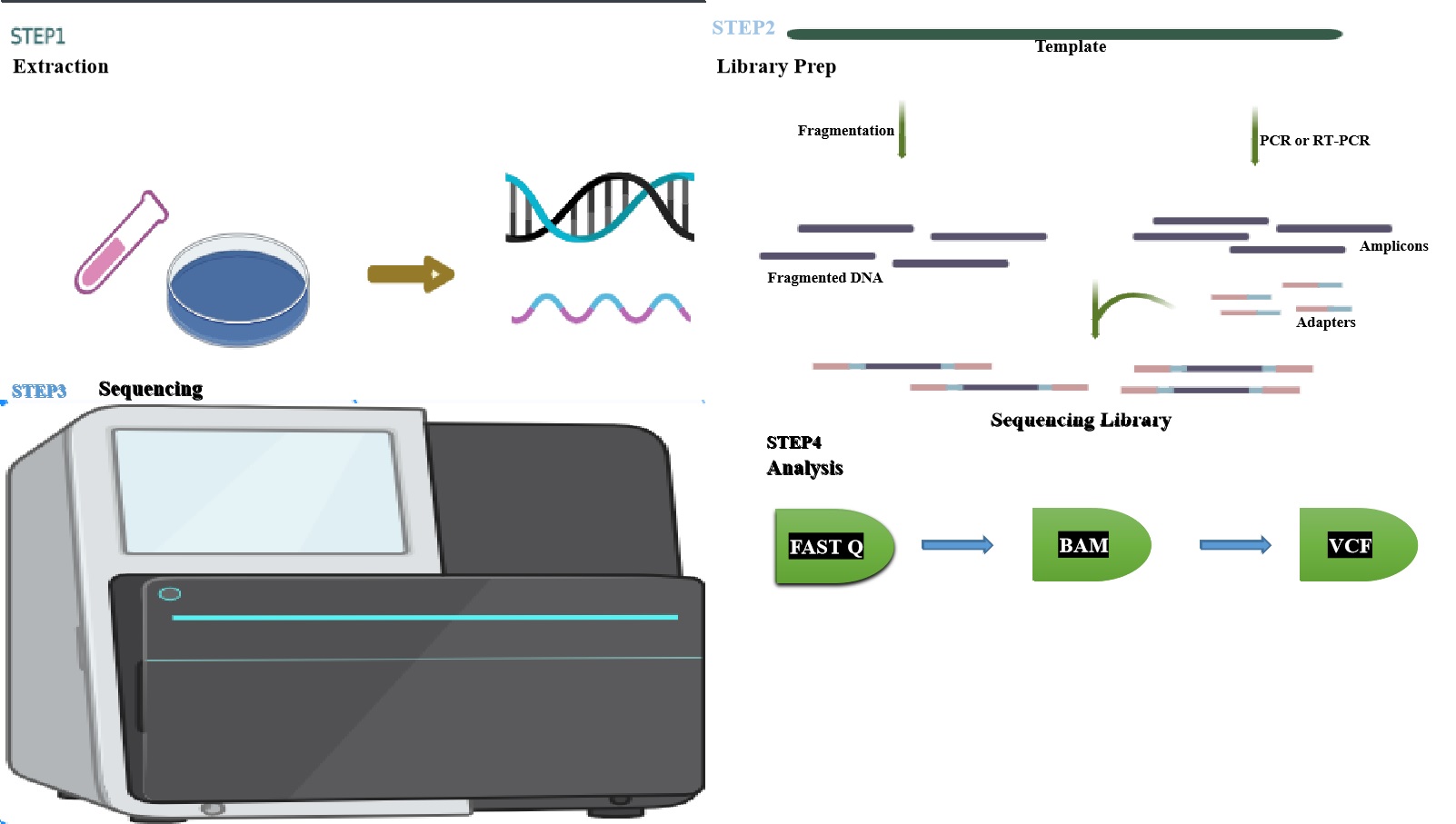
**The Role of Next-Generation Sequencing (NGS) in Biotechnology: Advancements and Applications**

**Introduction**

With the emergence and quick evolution of next sequencing (NGS), also known as massively parallel sequencing, DNA sequencing technologies have advanced significantly over the past several years [1]. High-throughput genomics research has been radically altered by NGS technologies, which have also led to the opening of numerous new fields of study and cutting-edge uses [2]. NGS technologies have proven to have significant potential for scientists studying biology and medicine [3]. Rapid advances in next-generation sequencing (NGS) technologies over the past few years have made it possible for researchers to produce huge numbers of sequence reads at significantly lower costs; this has not only allowed for an unprecedented expansion of the scope of genome-based research projects but has also revolutionized biological and biomedical research, including studies of human disease [4-5]. Moreover, NGS technologies are becoming more affordable and are replacing the micro array-based genotype methods limited to interrogating regions of known sequence variation [5]. This feature opened many contexts of successful applications, the first of which was whole-genome sequencing (WGS), an approach intended for entire genome sequencing. It provides the most complete landscape of genomic information and potential biological consequences. Next generation sequencing technologies are being used more and more in a variety of fields. Their power lies in the possibility to obtain enormous amounts of data and discover novel and essential information about the human genome [6]. The term "NGS" refers to a group of technologies that employ massively parallel sequencing techniques to generate millions of short read sequences more quickly, more affordably, and with higher throughput than Sanger sequencing. The majority of NGS-based approaches that examine genetic variation and its relationship to a certain phenotype use case-control study designs involving unrelated individuals. These study designs are vulnerable to population stratification bias (PSB) because patients and controls have genetically different ancestries [7]. Different methods are used by the NGS platforms that are currently available to achieve high-throughput sequencing. The many approaches of sequencing have an impact on the number, quality, and choice of applications for sequencing. A typical NGS run follows a standard procedure that starts with the extraction of genomic DNA from test samples, library preparation, which entails DNA fragmentation, ligation of adaptors, adaptor sequencing, sample enrichment, and sequencing [8].There are numerous NGS platforms accessible right now [9].

Sanger sequencing, sometimes known as a 'first-generation' sequencing technology, has been partially supplanted in recent years by 'next-generation' sequencing (NGS) approaches. NGS enables the identification of biomarkers for both early diagnosis and tailored treatment. The advent of next-generation sequencing (NGS) has altered clinical research, basic and applied sciences. The NGS enables the production of millions of data points with a lesser expenditure. NGS has not only transformed genome sequencing and personal treatment, but it has also altered earlier genome research. Sequencing is the application of a technique that aids in determining the amount and order of nucleotides present in the DNA of a specific organism. The invaluable knowledge gained through DNA sequencing aids numerous biological research and other sectors such as biotechnology, forensic science, and biological systematics. The unusual arrangement of bases has a substantial impact on health, for example, how we respond to a specific sickness and what type of treatment is required to cure it. NGS is a form of DNA sequencing technique that uses parallel sequencing of several tiny segments of DNA to identify the sequence (Rizzo et al. 2012). High-throughput technology is fast and generates a significant amount of data that is exponentially bigger than the Sanger procedure; it is also less expensive (Voelkerding et al, 2009). Third-generation sequencing is very similar to NGS, but unlike NGS, third-generation sequencing uses single DNA molecules as a template rather than amplified DNA. As a result, third generation sequencing may reduce errors in DNA sequence that arise in the laboratory during the DNA amplification process (Munroe et al. 2010).

This includes concepts, applications, advances, limitations, and the history of technological advances until the emergence of the NGS technique in the era of precision medicine, beginning with a brief history of DNA sequencing and continuing with a comprehensive description of the most commonly used NGS platforms, sequencing chemistries, methodologies, and general workflows (Fig1).



**Step 1: Sample extraction**

NGS can be performed on any sample that yields DNA or RNA (e.g., cell cultures, fresh-frozen tissues, formalin-fixed paraffin-embedded (FFPE) tissues, blood, saliva, and bone marrow). Various extraction protocols based on the starting material are available, and generally each extraction method has been optimized to yield the highest quality and largest amount of nucleic acid from the respective sample type. Following extraction, the amount and quality of DNA or RNA should be determined, as high-quality starting material is critical for successful sequencing.

## Step 2: Library preparation

Preparation of a sequencing library from your RNA or DNA sample involves two basic steps: 1) amplification to yield a pool of appropriately sized target sequences, and 2) the addition of sequencing adapters that will later interact with the NGS platform. If RNA is the starting template, an additional step is needed in which the RNA is first converted to cDNA by reverse transcription. PCR amplification yields a collection of specifically sized DNA fragments – called a library – that are compatible with the sequencing system to be used. The primers used in library preparation are designed based on the sequences of interest, which range from a whole genome to particular RNA transcripts. The adapter ligation step essentially bookends the amplified DNA or cDNA fragments, called amplicons, with specific oligonucleotide sequences that will interact with the surface of a sequencing flow cell. If multiple samples are to be sequenced in a single sequencing run, a unique identifier, or barcode, is additionally ligated to the amplicon. The resulting completed libraries can be pooled into a single sequencing run that is then “demultiplexed” during data analysis.

## Step 3: DNA sequencing

Parallel sequencing is performed using an NGS platform. The library is loaded onto the sequencer which then “reads” the nucleotides one by one. The number of reads produced will vary depending on the sequencing platform and kit used. Several methods of NGS have been developed including pyrosequencing, sequencing by ligation (SOLiD), sequencing by synthesis (SBS – Illumina), and Ion Torrent sequencing.

### Illumina Platforms

While all NGS platforms perform sequencing of millions of small fragments of DNA or cDNA, there are several different sequencing technologies. Some platforms can produce more reads or different read lengths than others. The most prevalent and successful sequencing technology was pioneered by Illumina. Illumina sequencers use a glass flow cell coated with millions of oligonucleotides that are complimentary to the sequencing adaptors. Each library fragment hybridizes with the primers and is further amplified to generate millions to billions of clonal clusters. Then, fluorescently labeled nucleotides are used to synthesize a complementary strand for each fragment. After the addition of each tagged nucleotide, the flow cell is imaged and the emission from each cluster is recorded. The fluorescent emission wavelength and intensity are used to identify the sequence of the templates.

## Step 4. Alignment and Data analysis

After sequencing is complete, specialized software is used to make sense of the large body of data produced. First the reads must be filtered for quality, amplicon size, and agreement between paired ends. The reads are then assembled and aligned to a reference genome. Finally, reads (assembled or raw) can be compared to a reference sequence or to reads from another sample to identify variants based on disease state, etc. If reads are aligned with a reference genome, variant annotation can be used to associate variants with known genes or regulatory sequences.

Other topics will focus on the use of NGS in normal practice, such as variant identification, whole-genome sequencing (WGS), whole-exome sequencing (WES), and multi-gene panels. A unified chapter detailing the main NGS characteristics in the clinic could benefit newcomers, scientists, researchers, and health care professionals, who will be in charge of transforming genetic data into genomic medicine (Schadt et al. 2010).

**Clinical applications**

NGS has recently enabled a better knowledge of genetic illnesses and has become a key technical advancement in the practice of diagnostic and clinical medicine. NGS allows for the study of various sections of the genome in a single reaction and has been proved to be a cost-effective and efficient method in the investigation of genetic illnesses in patients. NGS-generated genomic data has major implications for medical practice, including reliable identification of disease biomarkers, detection of inherited illnesses, and discovery of genetic variables that might assist predict responses to medicines [Jamuar SS and Tan EC. 2015]. However, recommendations on clinical adoption of NGS are still being debated, limiting its usage in the genetic clinic. Sequencing technology is used in a number of molecular diagnostic procedures, including single- and multi-gene panel tests, cell-free DNA for non-invasive prenatal testing, whole-exome sequencing (WES), and whole-genome sequencing (WGS). Given the recent usage of NGS as a diagnostic tool, there are problems such as when to order, who to order from, and how to interpret and convey the results to the patient and family. As a result, understanding the application, strength, and limitations of the various ways is required to determine which one is most suited to your situation [Rabbani et al., 2014].

**Multi-gene panels**

For many illnesses, the classical method is still highly effective. When a patient's clinical symptoms are typical for a certain disorder and the relationship between the disorder and the specific gene is well established with little locus heterogeneity, single-gene testing is advised [Xue et al., 2015]. However, many genetic conditions, such as cardiomyopathies, epilepsy, congenital muscular dystrophy, X-linked intellectual disability, and cancer susceptibility in families with atypical phenotypes, are difficult to diagnose due to clinical variability and genetic locus heterogeneity [LaDuca et al., 2015]. The diagnostic process has been completed, with clinical evaluation followed by sequential laboratory testing, with most tests being negative. The diagnostic rate of undiagnosed genetic illnesses (e.g., developmental delay/ cognitive disability and autism spectrum disorders) can vary substantially, and a multi-gene panel is more appropriate. Tothill and colleagues [Shashi et al., 2014] demonstrate the application of these multi-gene panels in cancer diagnostics by analyzing samples from patients with malignancies of unknown origin (CUP). The lack of a specific location of origin complicates clinical management of CUP patients, and this type of NGS study could assist identify new therapeutic choices. Many genes linked with a specific phenotype are sequenced and analyzed concurrently in multi-gene panel testing, lowering the cost and enhancing the efficiency of genetic diagnostics [Tothill 2013].

**Whole-genome and whole-exome sequencing**

Whole-genome sequencing (also known as WGS, full-genome sequencing, complete genome sequencing, or entire genome sequencing) is the method of determining an organism's complete DNA sequence all at once. The primary advantage of WGS is that it provides entire coverage of the genome, including promoters and regulatory regions. Whole-exome sequencing (WES) sequences all coding regions at a substantially deeper depth. According to Botstein et al. (2003), 85% of disease-causing mutations are found in coding and functional areas of the genome. As a result, exome sequencing has the potential to identify the causes of a huge number of rare, generally monogenic, genetic disorders, as well as predisposing variants in prevalent diseases and cancers [Gonzaga-Jauregui C]. Choi and colleagues demonstrated the utility of WES in clinical practice in 2009 by making genetic diagnoses of congenital chloride diarrhea in patients suspected of having Bartter syndrome, a renal salt-wasting condition. WES was performed on six patients who had no mutations in the traditional genes for Bartter syndrome. All patients had a homozygous deletion in the SLC26A3 gene, which yielded a molecular diagnosis of congenital chloride diarrhea, which was later validated clinically. Long-range PCR or Sanger sequencing. This tailored strategy, in addition to providing more extensive coverage of the 'known' phenotype-specific gene panels, enables for deeper coverage of these genes compared to WES, providing more confidence in the variants discovered. However, all NGS methods are susceptible to sequencing artifacts, and Sanger sequencing is recommended before providing the results to the patient [Botstein et al., 2003]. Furthermore, the patient and their family must be informed of all the intricacies associated with WES and WGS. It is critical to inform them that the test may not produce positive results, and that even positive results can provide diagnoses but do not improve prognosis or therapy.

**RNA-sequencing**

A transcriptome represents the complete set of RNA molecules from any genome at any time or condition and RNA plays essential role in several biological processes, including untranslated RNA species such as microRNAs (miRNAs). RNA-sequencing (RNA-seq) consists of an in-depth RNA analysis through NGS technologies and became the state-of-art technique for transcriptomic [Biesecker et al, 2014]. A typical RNA-seq experiment consists of a good experimental design, sample preparation, library construction, sequencing and data analysis. However, due to several experimental options available, a careful planning and cost estimation is necessary before starting. These include number and type of replicates (technical vs. biological), sequencing platform (e.g. Illumina, Ion Torrent), library preparation method (e.g. rRNA depletion or mRNA enrichment; strand-specific or not; single or paired end), throughput, read length, sequencing depth and coverage. RNA-seq enables detection of novel genes and isoforms, gene fusions, splice and chimeric variants, genomic alterations and gene expression quantification. Although RNA-seq outperforms microarray in transcriptomic analysis [Zhang et al., 2015], its clinical application is still in its infancy and, for instance, will not replace current approaches. RNA-seq is considered a complementary method depending on the needs and resources available, assisting clinicians in making decisions. In clinical practice, RNA measurement has applications across different areas in human health such as therapeutic selection, disease diagnostic and treatment [Byron et al., 2016].

Clinical diagnosis of infectious disease through RNA-seq is still rare, since quantitative PCR (RT-qPCR) assays are still the most common technique used for viral detection and genotyping.

Applications of NGS in virology diagnostic can be used for analysis of patients with unexplained illness, especially during outbreaks and epidemics. It also includes the identification of novel pathogens, viral community characterization, whole viral genome reconstruction, antiviral drug resistance, epidemiology and transcriptomic [Zhang et al., 2016, Sijmons et al., 2014, Chen et al., 2010]. The use of NGS in virology is increasing the knowledge of viral infection dynamics and their correlation with human health and treatment.

NGS can also be applied for circulating tumour RNA (ctRNA) discovery. The analysis of ctRNA in plasma is still in its beginning and presents specific challenges. ctRNA degrades faster than circulating tumour DNA (ctDNA) and needs to be purified rapidly or added in preservative solutions (e.g. TRIzol) and freezed at −80°C, not always an accessible technique to many clinical sites. Despite these challenges, ctRNAs represent good biomarkers of early detection of multiple tumour types, such as breast, lung, prostate and colorectal cancers. NGS is a more powerful tool for ctRNA detection; however, RT-qPCR remains more usable for clinical diagnostic applications [Pimentel et al., 2015].

**Epigenetics**

Epigenetics is a rapidly developing field with enormous implications for health and clinical diagnostics. Conrad Waddington invented the phrase in the 1940s to describe the study of heritable changes in gene activity and expression that do not involve the DNA sequence itself, i.e. a change in phenotype without a change in genotype. Epigenetic mechanisms are another layer of gene control, and NGS enabled researchers to comprehend the status of epigenetics on a wide scale and at a single base-resolution, including primarily DNA methylation, histone modification, and non-coding RNA (ncRNA)-associated silencing [Egger et al., 2004].

DNA methylation was the first epigenetic process discovered, and it is the most well-known and prevalent in human cancer. It entails covalently modifying cytosine by adding a methyl group to cytosines in CpG (cytosine/guanine) islands [Rodenhiser D, Mann, 2004]. DNA methyltransferases (DNMTs) maintain this methylation, which is important for gene transcriptional repression, transposable element silencing, and virus defense. Unmethylated DNA is located in active chromatin regions, while methylated DNA is found in inactive chromatin regions [Skinner et al., 2010].

Post-translational histone modifications, such as acetylation and methylation of conserved lysine residues on the amino-terminal tail domains, are markers for chromatin activity: acetylation is found in active regions of chromatin, whereas hypoacetylation is found in inactive euchromatic or heterochromatic regions. Histone deacetylases (HDACs), histone acetylases, and histone methyltransferases are among the enzymes involved in this process. Because epigenetic mechanisms regulate DNA accessibility, changes in the cell epigenetic pattern alter gene expression and can result in inherited or acquired human disorders [Nicholls et al., 2010].

**References**

1. Metzker ML. Sequencing technologies – the next generation. Nat. Rev. Genet. 11(1), 31–46 (2010).
2. Jacquier A. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. Nat. Rev. Genet. 10(12), 833–844 (2009).
3. Shendure J, Ji H. Next-generation DNA sequencing. Nat. Biotechnol. 26(10), 1135–1145 (2008).
4. Ng SB, Turner EH, Robertson PD, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009;461:272–6.
5. Ng SB, Buckingham KJ, Lee C, et al. Exome sequencing identifies the cause of a Mendelian disorder. Nat Genet. 2010;42:30–5.
6. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole genome sequencing. Nat Rev Genet. 2010;11:415–25.
7. Freedman, M. L., Reich, D., Penney, K. L., McDonald, G. J., Mignault, A. A., Patterson, N., et al. (2004). Assessing the impact of population stratification on genetic association studies. Nat. Genet. 36, 388-393. doi: 10.1038/ng1333
8. Buermans, H. P. J., and den Dunnen, TJ. (2014). Next generation sequencing technology: advances and applications. Biochim. Biophys. Acta 1842, 1932-1941. doi: 10.1016/j.bbadis.2014.06.015
9. Heather, J. M., and Chain, B. (2016). The sequence of sequencers: the history of sequencing DNA. Genomics 107, 1-8. doi: 10.1016/j.ygeno.2015.11.003