Introduction to NGS

Dr. Umar Ahmad

September 1, 2024

#### Learning Objectives

After this section you should be able to:

* List the main high-throughput sequencing technologies in use.
* Describe the main differences between Illumina and Oxford Nanopore platforms, including their advantages and disadvantages.
* Recognise the structure of common file formats in bioinformatics, in particular FASTQ, FASTA, GFF3 and CSV/TSV.

## Next Generation Sequencing

The sequencing of genomes has become more routine due to the [rapid drop in DNA sequencing costs](https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data) seen since the development of Next Generation Sequencing (NGS) technologies in 2007. One main feature of these technologies is that they are *high-throughput*, allowing one to more fully characterise the genetic material in a sample of interest.

There are three main technologies in use nowadays, often referred to as 2nd and 3rd generation sequencing:

* Illumina’s sequencing by synthesis (2nd generation)
* Oxford Nanopore Technologies, shortened ONT (3rd generation)
* Pacific Biosciences, shortened PacBio (3rd generation)

The video below from the iBiology team gives a great overview of these technologies.

### Illumina Sequencing

Illumina’s technology has become a widely popular method, with many applications to study transcriptomes (RNA-seq), epigenomes (ATAC-seq, BS-seq), DNA-protein interactions (ChIP-seq), chromatin conformation (Hi-C/3C-Seq), population and quantitative genetics (variant detection, GWAS), de-novo genome assembly, amongst [many others](https://emea.illumina.com/content/dam/illumina-marketing/documents/products/research_reviews/sequencing-methods-review.pdf).

An overview of the sequencing procedure is shown in the animation video below. Generally, samples are processed to generate so-called sequencing libraries, where the genetic material (DNA or RNA) is processed to generate fragments of DNA with attached oligo adapters necessary for the sequencing procedure (if the starting material is RNA, it can be converted to DNA by a step of reverse transcription). Each of these DNA molecule is then sequenced from both ends, generating pairs of sequences from each molecule, i.e. **paired-end sequencing** (single-end sequencing, where the molecule is only sequenced from one end is also possible, although much less common nowadays).

This technology is a type of **short-read sequencing**, because we only obtain short sequences from the original DNA molecules. Typical protocols will generate 2x50bp to 2x250bp sequences (the 2x denotes that we sequence from each end of the molecule).

The main advantage of Illumina sequencing is that it produces very **high-quality sequence reads** (error rate <1%) at a low cost. However, we only get **very short sequences**, which is a limitations when it comes to resolving problems such as long sequence repeats (e.g. around centromeres or transposon-rich areas of the genome), distinguishing gene isoforms (in RNA-seq), or resolving haplotypes (combinations of variants in each copy of an individual’s diploid genome).

**In summary, Illumina:**

* Utilizes sequencing-by-synthesis chemistry.
* Offers short read lengths.
* Known for high accuracy with low error rates (<1%).
* Well-suited for applications like DNA resequencing and variant detection.
* Scalable and cost-effective for large-scale projects.
* Limited in sequencing long DNA fragments.
* Expensive to set up.

### Nanopore Sequencing

Nanopore sequencing, a form of **long-read sequencing technology**, has distinct advantages in the field of genomics. It excels in its ability to sequence exceptionally lengthy DNA molecules, including those reaching megabase sizes, thus effectively addressing that limitation of short-read sequencing. The **portability** of some nanopore sequencing devices is another advantageous feature; some are designed to operate via a simple USB connection to a standard laptop, making it exceptionally adaptable for on-the-go applications, including fieldwork.

|  |
| --- |
| Overview of Nanopore sequencing showing the highly-portable MinION device. The device contains thousands of nanopores embedded in a membrane where current is applied. As individual DNA molecules pass through these nanopores they cause changes in this current, which is detected by sensors and read by a dedicated computer program. Each DNA base causes different changes in the current, allowing the software to convert this signal into base calls. Source: Fig. 1 in [Wang et al. 2021](https://doi.org/10.1038/s41587-021-01108-x) |

Overview of Nanopore sequencing showing the highly-portable MinION device. The device contains thousands of nanopores embedded in a membrane where current is applied. As individual DNA molecules pass through these nanopores they cause changes in this current, which is detected by sensors and read by a dedicated computer program. Each DNA base causes different changes in the current, allowing the software to convert this signal into base calls. Source: Fig. 1 in [Wang et al. 2021](https://doi.org/10.1038/s41587-021-01108-x)

However, optimising this technology presents some challenges, notably in the production of sequencing libraries containing high molecular weight and intact DNA. It’s important to note that nanopore sequencing historically exhibited **higher error rates**, approximately 5% for older chemistries, compared to Illumina sequencing. However, significant advancements have emerged, enhancing the accuracy of nanopore sequencing technology, now achieving [accuracy rates exceeding 99%](https://nanoporetech.com/accuracy).

**In summary, ONT:**

* Operates on the principle of nanopore technology.
* Provides long read lengths, ranging from thousands to tens of thousands of base pairs.
* Ideal for applications requiring long-range information, such as *de novo* genome assembly and structural variant analysis.
* Portable, enabling fieldwork and real-time sequencing.
* Exhibits higher error rates (around 5%), with improvements in recent versions.
* Costs can be higher per base, compared to Illumina for certain projects.

#### Which technology to choose?

Both of these platforms have been widely popular for bacterial sequencing. They can both generate data with high-enough quality for the assembly and analysis for most of the pathogen genomic surveillance. Mostly, which one you use will depend on what sequencing facilities you have access to.

While Illumina provides the cheapest option per sample of the two, it has a higher setup cost, requiring access to the expensive sequencing machines. On the other hand, Nanopore is a very flexible platform, especially its portable MinION devices. They require less up-front cost allowing getting started with sequencing very quickly in a standard molecular biology lab.

## Bioinformatics file formats

Bioinformatics relies on various standard file formats for storing diverse types of data. In this section, we’ll discuss some of the key ones we’ll encounter, although there are numerous others. You can refer to the “[Common file formats](../appendices/01-file_formats.md)” appendix for a more comprehensive list.

### FAST5

FAST5 is a proprietary format developed by ONT and serves as the standard format generated by its sequencing devices. It is based on the hierarchical data format HDF5, designed for storing extensive and intricate data. Unlike text-based formats like FASTA and FASTQ, FAST5 files are binary, necessitating specialized software for opening and reading.

Within these files, you’ll find a Raw/ field containing the original raw current signal measurements. Additionally, tools like basecallers can add Analyses/ fields, converting signals into standard FASTQ data (e.g., Guppy basecaller).

Typically, manual inspection of these files is unnecessary, as specialized software is used for processing them. For more in-depth information about this format, you can refer to [this resource](https://static-content.springer.com/esm/art%3A10.1038%2Fs41587-021-01147-4/MediaObjects/41587_2021_1147_MOESM1_ESM.pdf).

### FASTQ

FASTQ files are used to store **nucleotide sequences along with a quality score** for each nucleotide of the sequence. These files are the typical format **obtained from NGS sequencing** platforms such as Illumina and Nanopore (after basecalling). Common file extensions used for this format include .fastq and .fq.

The file format is as follows:

@SEQ\_ID <-- SEQUENCE NAME  
AGCGTGTACTGTGCATGTCGATG <-- SEQUENCE  
+ <-- SEPARATOR  
%%).1\*\*\*-+\*''))\*\*55CCFF <-- QUALITY SCORES

In FASTQ files each sequence is always represented across 4 lines. The quality scores are encoded in a compact form, using a single character. They represent a score that can vary between 0 and 40 (see [Illumina’s Quality Score Encoding](https://support.illumina.com/help/BaseSpace_OLH_009008/Content/Source/Informatics/BS/QualityScoreEncoding_swBS.htm)). The reason single characters are used to encode the quality scores is that it saves space when storing these large files. Software that work on FASTQ files automatically convert these characters into their score, so we don’t have to worry about doing this conversion ourselves.

The quality value in common use is called a **Phred score** and it represents the **probability that the base is an error**. For example, a base with quality 20 has a probability of being an error. A base with quality 30 has chance of being an error. Typically, a Phred score threshold of >20 or >30 is used when applying quality filters to sequencing reads.

Because FASTQ files tend to be quite large, they are **often compressed** to save space. The most common compression format is called *gzip* and uses the extension .gz. To look at a *gzip* file, we can use the command zcat, which decompresses the file and prints the output as text.

For example, we can use the following command to count the number of lines in a compressed FASTQ file:

zcat sequences.fq.gz | wc -l

If we want to know how many sequences there are in the file, we can divide the result by 4 (since each sequence is always represented across four lines).

### FASTA

FASTA files are used to store **nucleotide or amino acid sequences**. Common file extensions used for this format include .fasta, .fa, .fas and .fna.

The general structure of a FASTA file is illustrated below:

>sample01 <-- NAME OF THE SEQUENCE  
AGCGTGTACTGTGCATGTCGATG <-- SEQUENCE ITSELF

Each sequence is represented by a name, which always starts with the character >, followed by the actual sequence.

A FASTA file can contain several sequences, for example:

>sample01  
AGCGTGTACTGTGCATGTCGATG  
>sample02  
AGCGTGTACTGTGCATGTCGATG

Each sequence can sometimes span multiple lines, and separate sequences can always be identified by the > character. For example, this contains the same sequences as above:

>sample01 <-- FIRST SEQUENCE STARTS HERE  
AGCGTGTACTGT  
GCATGTCGATG  
>sample02 <-- SECOND SEQUENCE STARTS HERE  
AGCGTGTACTGT  
GCATGTCGATG

To count how many sequences there are in a FASTA file, we can use the following command:

grep ">" sequences.fa | wc -l

In two steps:

* find the lines containing the character “>”, and then
* count the number of lines of the result.

FASTA files are commonly used to **store genome sequences**, after they have been assembled. We will see FASTA files several times throughout these materials, so it’s important to be familiar with them.

### GFF3

The **GFF3 (Generic Feature Format version 3)** is a standardized file format used in bioinformatics to describe genomic features and annotations. It primarily serves as a structured and human-readable way to represent information about genes, transcripts, and other biological elements within a genome. Common file extensions used for this format include .gff and .gff3.

Key characteristics of the GFF3 format include:

* **Tab-delimited columns:** GFF3 files consist of tab-delimited columns, making them easy to read and parse.
* **Hierarchical structure:** the format supports a hierarchical structure, allowing the description of complex relationships between features. For instance, it can represent genes containing multiple transcripts, exons, and other elements.
* **Nine standard columns:** this includes information such as the sequence identifier (e.g. chromosome), feature type (e.g. gene, exon), start and end coordinates, strand and several attributes.
* **Attributes field:** the ninth column, known as the “attributes” field, contains additional information in a key-value format. This field is often used to store details like gene names, IDs, and functional annotations.
* **Comments:** GFF3 files can include comment lines starting with a “#” symbol to provide context or documentation.

GFF3 is widely supported by various bioinformatics tools and databases, making it a versatile format for storing and sharing genomic annotations.

### CSV/TSV

**Comma-separated values** (CSV) and **tab-separated values** (TSV) files are text-based formats commonly used to store **tabular data**. While strictly not specific to bioinformatics, they are commonly used as the output of bioinformatic software. CSV files usually have .csv extension, while TSV files often have .tsv or the more generic .txt extension.

In both cases, the data is organized into rows and columns. Rows are represented across different lines of the file, while the columns are separated using a **delimiting character**: a command , in the case of CSV files and a tab space (tab ↹) for TSV files.

For example, for this table:

| sample | date | strain |
| --- | --- | --- |
| VCH001 | 2023-08-01 | O1 El Tor |
| VCH002 | 2023-08-02 | O1 Classical |
| VCH003 | 2023-08-03 | O139 |
| VCH004 | 2023-08-04 | Non-O1 Non-O139 |

This would be its representation as a CSV file:

sample,date,strain  
VCH001,2023-08-01,O1 El Tor  
VCH002,2023-08-02,O1 Classical  
VCH003,2023-08-03,O139  
VCH004,2023-08-04,Non-O1 Non-O139

And this is its representation as a TSV file (the space between columns is a tab ↹):

sample date strain  
VCH001 2023-08-01 O1 El Tor  
VCH002 2023-08-02 O1 Classical  
VCH003 2023-08-03 O139  
VCH004 2023-08-04 Non-O1 Non-O139

CSV and TSV files are human-readable and can be opened and edited using **basic text editors** or **spreadsheet software** like *Microsoft Excel*.

## Summary

#### Key Points

* High-throughput sequencing technologies, often called next-generation sequencing (NGS), enable rapid and cost-effective genome sequencing.
* Prominent NGS platforms include Illumina, Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio).
* Each platform employs distinct mechanisms for DNA sequencing, leading to variations in read length, error rates, and applications.
* Illumina sequencing:
  + Uses sequencing-by-synthesis chemistry, produces short read lenghts and has high accuracy with low error rates (<1%).
  + While it is scalable and cost-effective for large-scale projects, it is expensive to set up and limited in sequencing long DNA fragments.
* Nanopore sequencing:
  + Uses nanopore technology, provides long read lengths, making it ideal for applications such as *de novo* genome assembly.
  + Although the costs can be higher per base, it is cheaper to set up.
  + Exhibits higher error rates (around 5%), but with significant improvements in recent versions (1%).
* Common file formats in bioinformatics include FASTQ, FASTA and GFF. These are all text-based formats.
* FASTQ format (.fastq or .fq):
  + Designed to store sequences along with quality scores.
  + Contains a sequence identifier, sequence data, a separator line and quality scores.
  + Widely used for storing **sequence reads generated by NGS platforms**.
* FASTA format (.fasta, .fa, .fas, .fna):
  + Is used for storing biological sequences, including DNA, RNA, and protein.
  + It Comprises a sequence identifier (often preceded by “>”) and the sequence data.
  + Commonly used for sequence storage and exchange of **genome sequences**.
* GFF format (.gff or .gff3):
  + A structured, tab-delimited format for describing genomic features and annotations.
  + Consists of nine standard columns, including sequence identifier, feature type, start and end coordinates, strand information, and attributes.
  + Facilitates the representation of genes, transcripts, and other genomic elements, supporting hierarchical structures and metadata.
  + Commonly used for storing and sharing **genomic annotation** data in bioinformatics.
* CSV (.csv) and TSV (.tsv):
  + Plain text formats to store tables.
  + The columns in the CSV format are delimited by comma, whereas in the TSV format by a tab.
  + These files can be opened in standard spreadsheet software such as *Excel*.