Introduction to Bioinformatics and Computational Genomics

Serghei Mangul, Ph.D
Assistant Professor of Clinical Pharmacy and Biological Sciences,
University of Southern California

Objectives

By the end of this module, you will be able to:

- Understand the basics of bioinformatics analysis.
- Understand the basic workflow of RNA-seq and its analyses
- Calculate genome length, coverage and throughput
- Explain the differences between short and long read sequencing technologies
- Understand the purpose of quantification of transcripts and differential gene expression

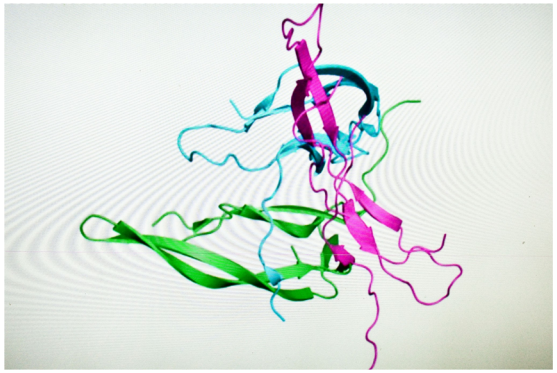
What is bioinformatics?

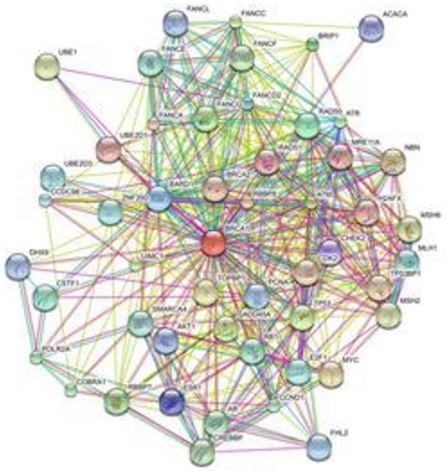
- Bioinformatics is the branch of biology that is concerned with the acquisition, storage, and analysis of the **information found in nucleic acid and protein sequence data**.
- By selecting an appropriate computer program, scientists can use sequence data to look for genes, get clues to gene functions, examine genetic variation, and explore evolutionary relationships. Bioinformatics is a young and dynamic science.
- New bioinformatic software is being developed while existing software is continually updated.

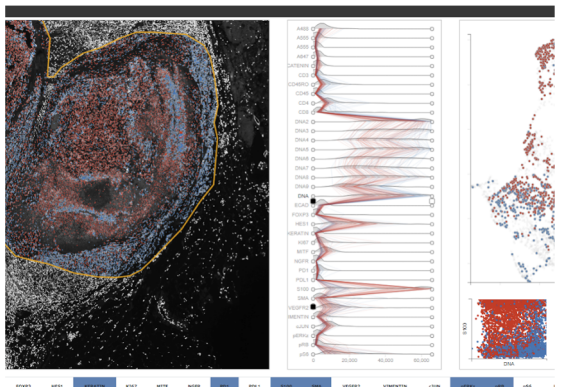
https://www.genome.gov/25020000/online-education-kit-bioinformatics-introduction

Where can we apply bioinformatics?

- Genomics
  - Genomic feature predictions
  - Sequencing data analysis
- Proteomics
  - Protein 3D structure modeling
  - Drug design
- Systems Biology
  - Gene set enrichment
  - Pathway analysis
- Phenotype
  - Image analysis
  - Integration

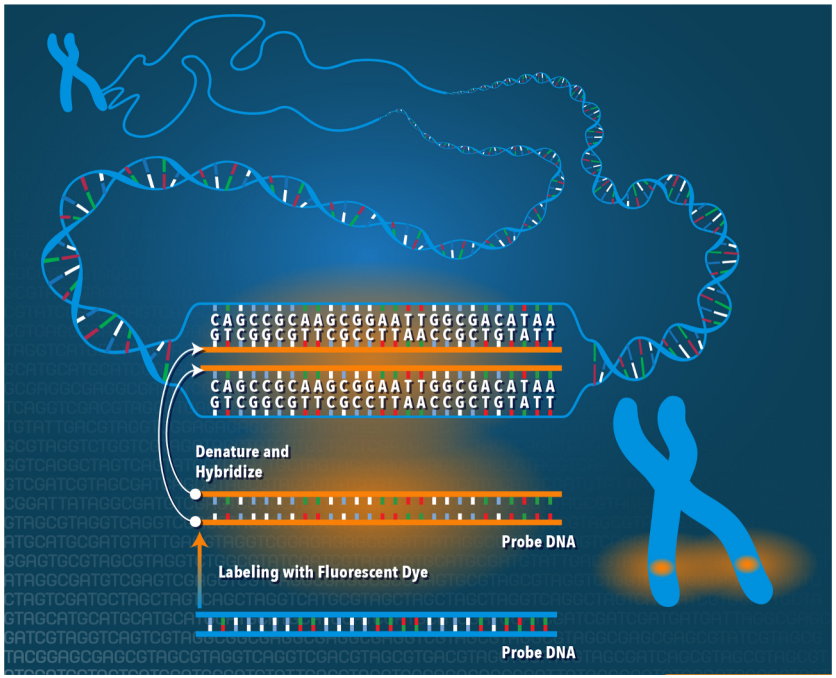






What is genomics?

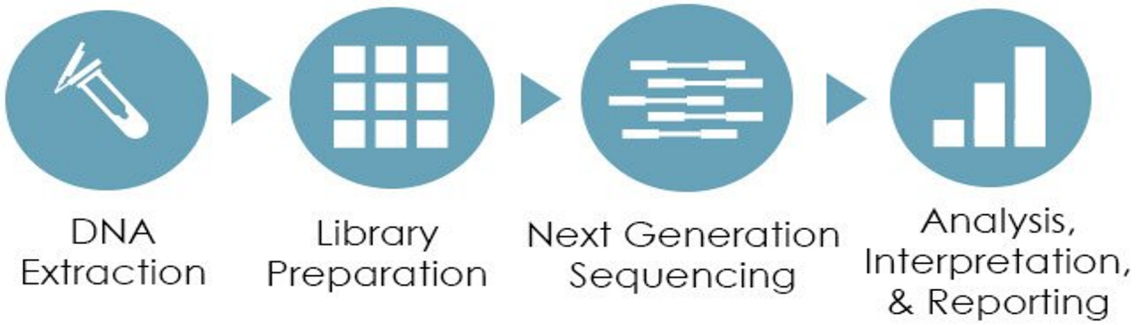
- Genomics is the study of **whole genomes of organisms**, and incorporates elements from genetics.
- Genomics uses a combination of recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble, and analyze the structure and function of genomes
- Genomics harnesses the availability of complete DNA sequences for entire organisms and was made possible by both the pioneering work of Fred Sanger and the more recent next-generation sequencing technology.



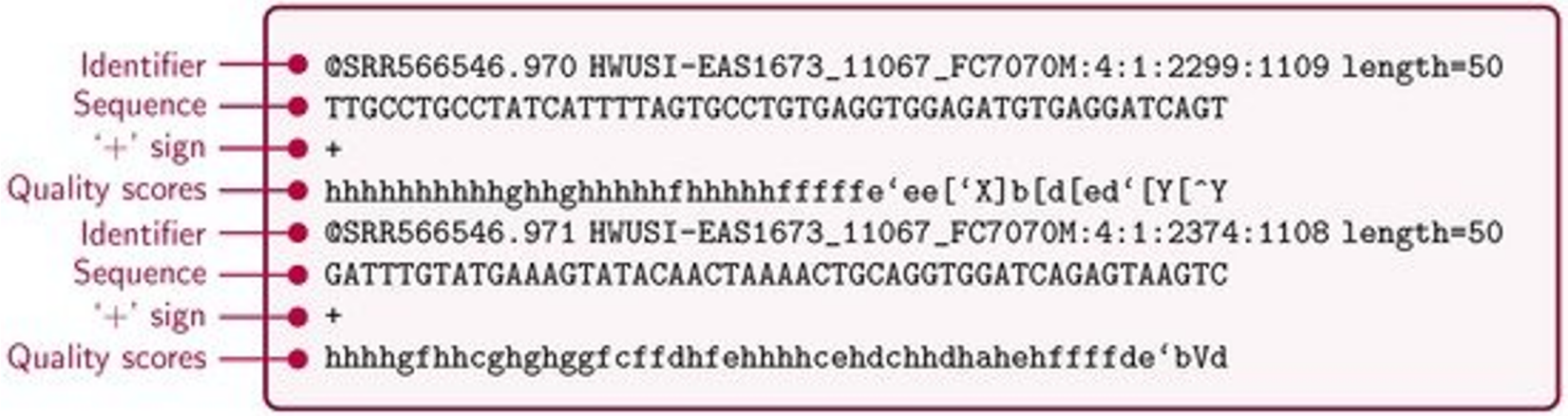
https://www.ebi.ac.uk/training/online/course/genomics-introduction-ebi-resources/what-genomics

Next generation sequencing

- A high-throughput method used to determine a portion of the nucleotide sequence of an individual's genome. This technique utilizes DNA sequencing technologies that are capable of processing multiple DNA sequences in parallel.
- Next-generation sequencing (NGS), also known as high-throughput sequencing, is the catch-all term used to describe a number of different modern sequencing technologies. Also called ‘Massive parallel sequencing’ or ‘Second generation sequencing”
- These technologies allow for sequencing of DNA and RNA much more quickly and cheaply than the previously used Sanger sequencing.

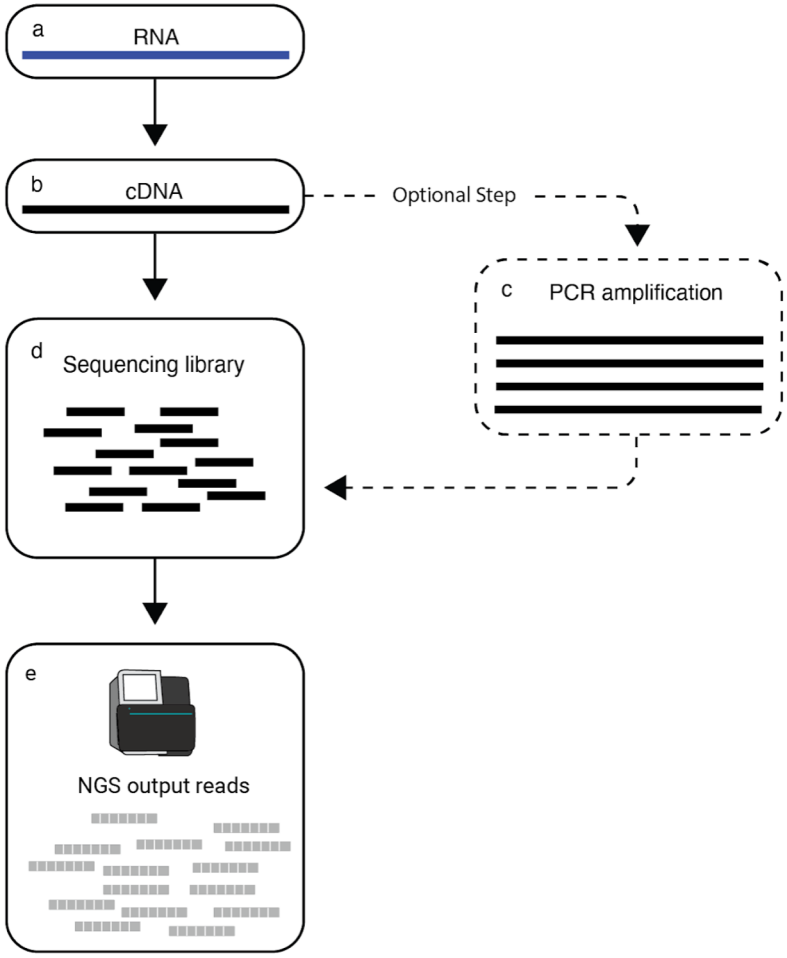


Reads

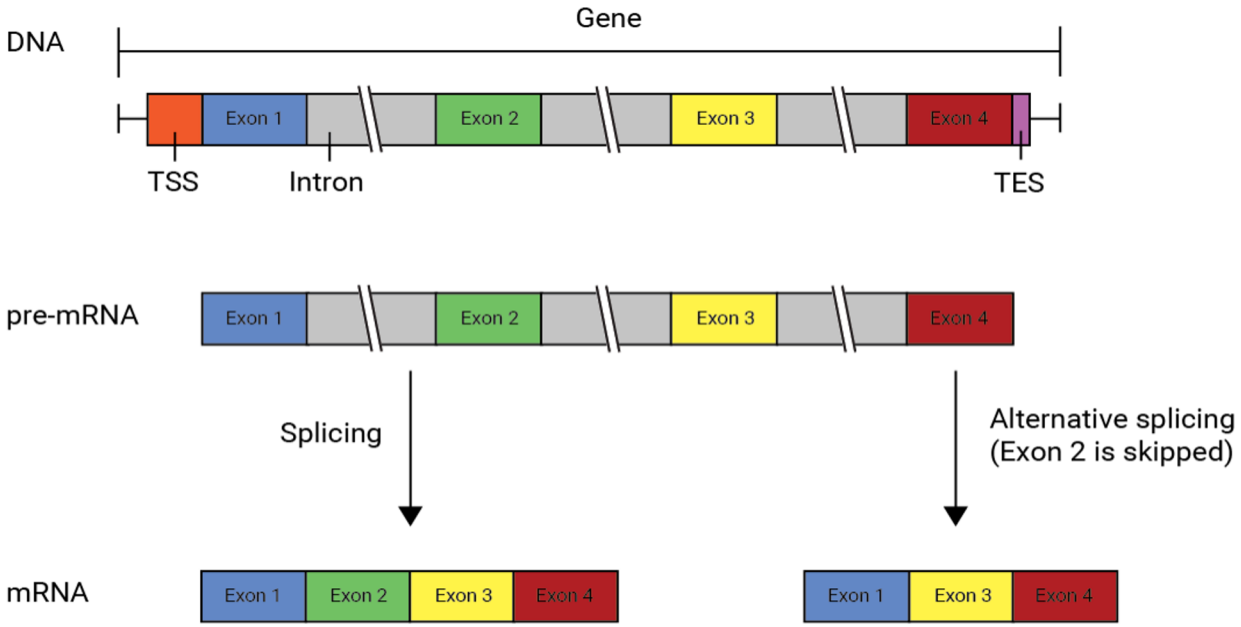

Identifier
Sequence
'+' sign
Quality scores
Identifier
Sequence
'+' sign
Quality scores
@SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
+
hhhhhhhhhhghhghhhhhfhhhhhfffffe'ee['X]b[d[ed'[Y[^Y
@SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50
GATTTGTATGAAAGTATACAACTAAAACTGCAGGTGGATCAGAGTAAGTC
+
hhhhgfhhcghghggfcffdhfehhhhcehdchhdhahehffffde'bVd

RNA-seq

- Preparation of an RNA-seq library starts with extracting and **isolating RNA** from a biological sample, such as a cell line or a frozen tissue sample.
- The RNA then undergoes **reverse transcription** and is converted into **cDNA,** which is then amplified by **polymerase chain reaction (PCR)** and fragmented into short sequences.
- The constructed **sequencing library** becomes the input for the **NGS machine** after the RNA molecules are processed.
- The machine's main function is to output a file containing sequences of the fragments in the library.
- The reads are usually delivered as a FASTQ file, which contains the read sequences.



- During transcription, DNA is used as a template to create a single strand of RNA. First, the DNA molecule is copied by RNA polymerase into a form known as pre-mRNA, which contains all the bases of the original DNA strand.
- Portions of the pre-mRNA, called introns, are removed because these portions do not code for proteins. The remaining portions, called exons, are then combined together. This process is known as alternative splicing.
- The final mRNA contains just the combined exons and carries the genetic code, which is expressed as a protein and referred to as a transcript.

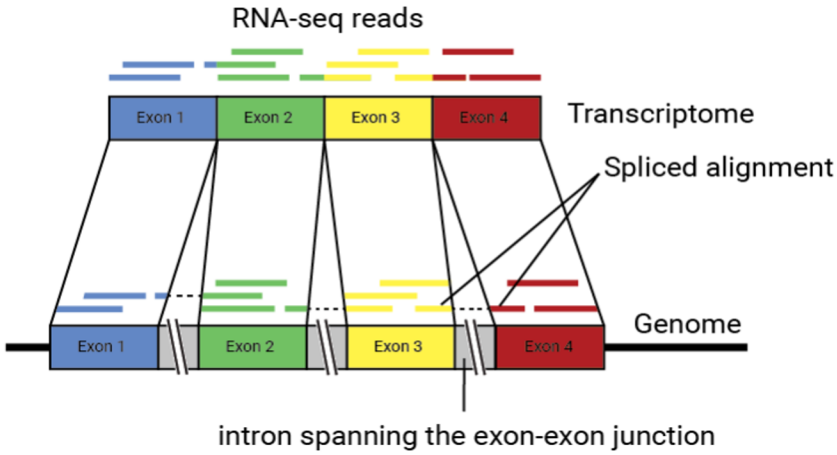


Read alignment

- Read alignment compares the reads derived from an individual genome, which is unknown and also referred to as the donor genome, to a known reference genome.
- The goal of read alignment is to find a section of the genome that matches or nearly matches the read sequence
- Algorithms pre-process the reference sequence, the short reads, or both of these genomic elements, into an indexed form that facilitates rapid searching for matches.
- Alignment tools reconstruct the individual genome by replacing the nucleotide bases in the reference genome which sometimes differ from the nucleotide bases present in the reads obtained from sequencing in the corresponding loci.

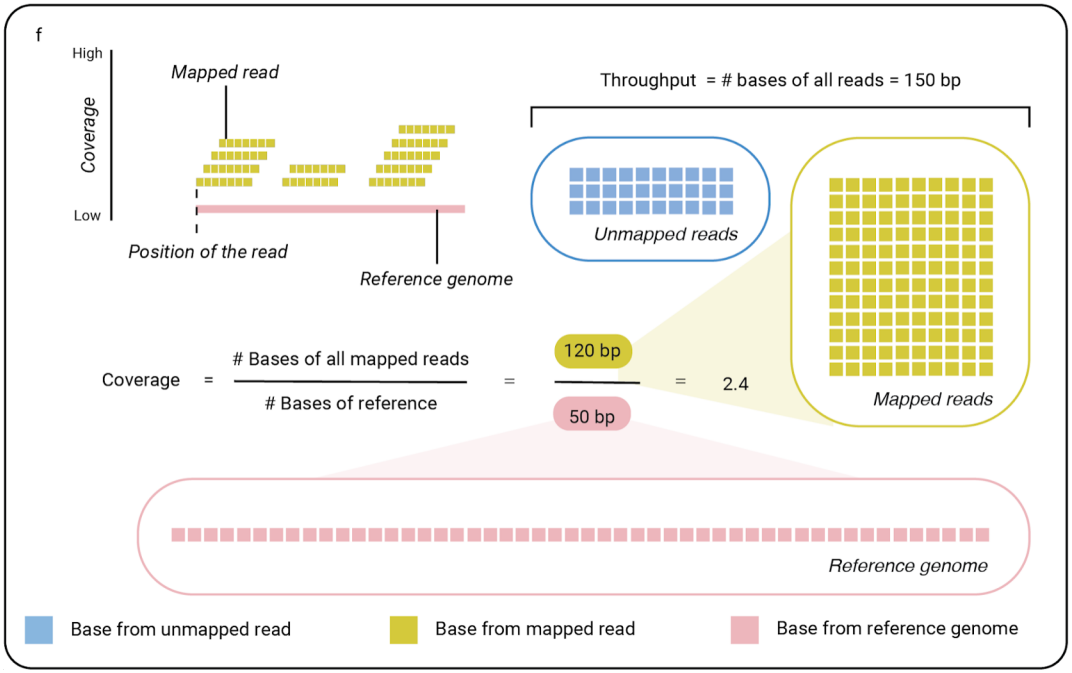
Alignment of RNA-seq reads

- Aligning RNA-seq reads is complicated because, while reads originate from the transcripts, they are often aligned to the reference genome which does not contain the junction regions where two exons were combined.
- The reads could also be aligned to reference transcripts prior to aligning to the genome. However, the presence of novel junctions—junctions not present in the reference transcripts—may require the RNA-seq reads, also called spliced reads, to be aligned to the reference genome.



Multi-mapped reads, coverage, and throughput

- A typical genome contains regions with repeat sequences - reads are shorter than the repetitive regions
- Reads may align to multiple regions of the reference genome equally well - multi-mapped reads - produce false matches
- To overcome the problem of false matches produced by multi-mapped reads, coverage rate (sequencing depth) can be increased which can aid in increasing the accuracy of matched reads



File formats

- **Fastq**
- **Fasta**
- **Sam**
- **Bam**
- **Quant.sf (main salmon output)**

Fasta

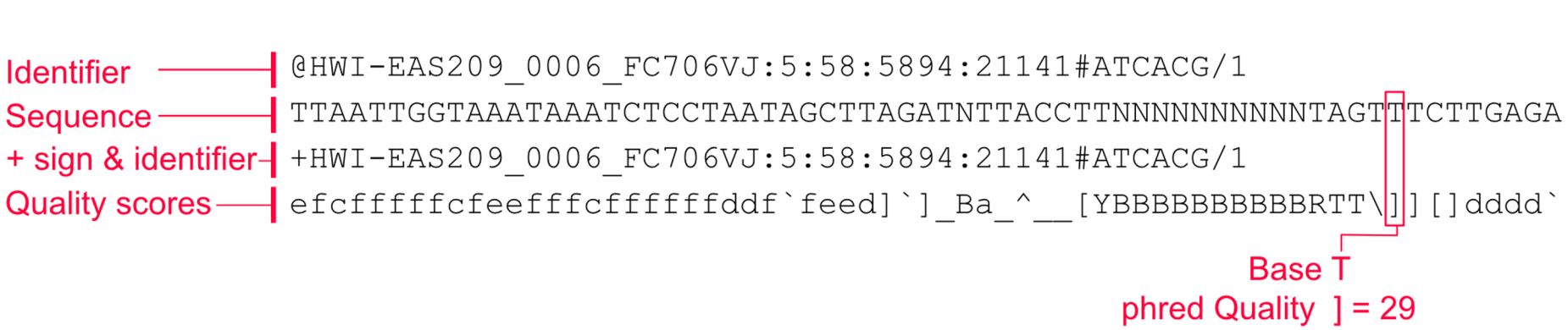
The first line in a FASTA file usually starts with a “>” (greater-than) symbol. This first line is called the “description line”, and can contain descriptive information about the sequence in the subsequent lines. The description can be id or name of the sequence such as gene names. However, very infrequently you may see lines starting with a “;” (semicolon). These lines will be taken as a comment, and can hold additional descriptive information about the sequence in subsequent lines.

Ex:

```
>Human:Chromosome1
ATGCATCGGCTAGGTCTCTTGATCGATCGATGCTAGCTACGTACGTAC
>Human:Chromosome2
ATGCATCGGCTAGGTCTCTTGATCGATCGATGCTAGCTACGTACGTAC
>Human:Chromosome3
ATGCATCGGCTAGGTCTCTTGATCGATCGATGCTAGCTACGTACGTAC
```

Fastq

An extension of the FASTA format is FASTQ format. This format is designed to handle base quality metrics output from sequencing machines. In this format, both the sequence and quality scores are represented as single ASCII characters. The format uses for lines for each sequence, and these four lines are stacked on top of each other in text files output by sequencing workflows. Each of the 4 lines will represent a read.



https://compgenomr.github.io/book/fasta-and-fastq-formats.html

Sam

- A Sam (sequence alignment mapping))file contains an alignment of reads to a reference sequence in a human readable format.

Bam

- A Bam file contains the exact same information about an alignment as a Sam file.
- The key difference is that a Bam file is encoded in binary, which is not human readable. The reason for this is to compress the alignment data for faster computation time in downstream bioinformatics analysis pipelines.
- Tools such as samtools are capable of converting the binary format to a human readable sam file format on the fly.

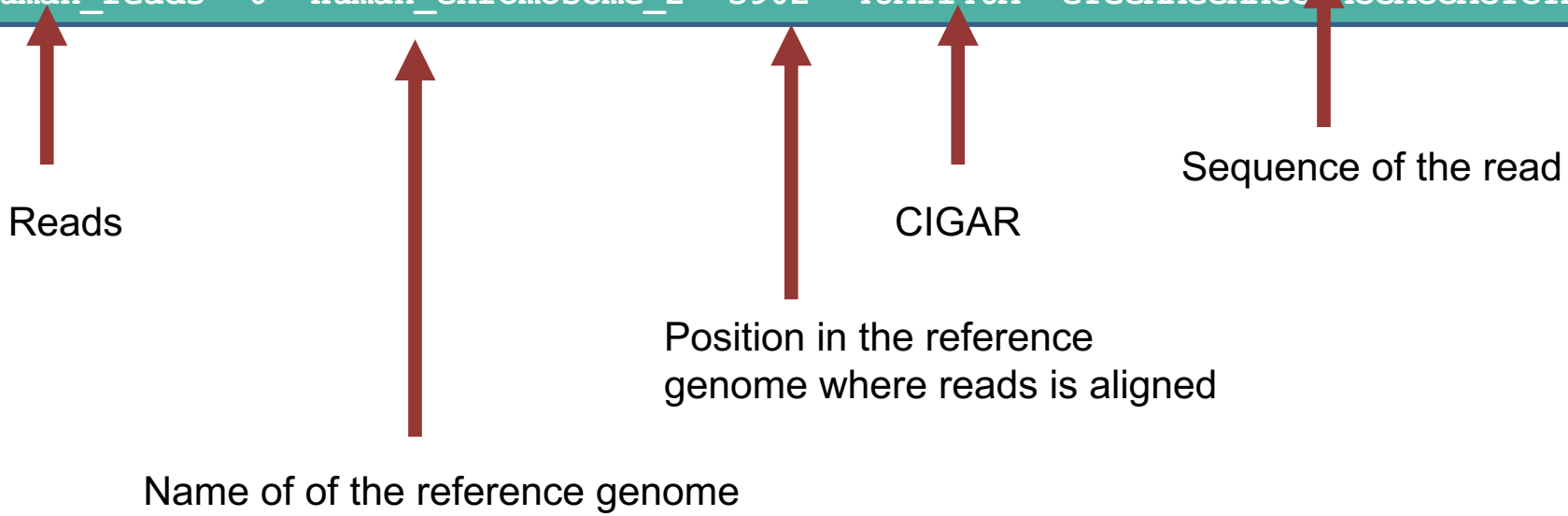
Interpreting SAM/BAM alignment file

```
Run this command to view aligned reads:
~> less reads.sam

human_read1   0   Human_chromosome_1   150    18M1I52M    MCCCTTTTAGTCAGTGTGGAAAAATCT....

human_read2   0   Human_chromosome_2   1702   3M1D68M     CAGCCCACCAGAAGAGAGCGGCAGGTCT....

human_read3   0   Human_chromosome_2   5902   48M1I46M    GTGGAAGGAAGC[illegible]ACCACCACTCTATTT....
```



Understanding the CIGAR

CIGAR (Concise Idiosyncratic Gapped Alignment Report) is a way to represent additional bases of the read that are not in the reference or may be missing bases of the read that are in the reference.

Interpreting alignment of read1

Line from reads.sam file:

```
read1   0   Human_chromosome_1   150    18M1I52M …
```

Interpretation:

Read1 was aligned to HIV_complete_genome as follows:

18 matches (18M) then 1 insertion (1I) and then 52 matches (52M)

Interpreting alignment of read2

Line from reads.sam file:

```
read2   0   Human_chromosome_2   1702   3M1D68M
```

Interpretation:

Read2 was aligned to HIV_complete_genome as follows:

```
3 matches, then 1 deletion then 68 matches
```

Quant.sf (Main salmon output)

Salmon's main output is its quantification file and contains the gene transcripts and the number of reads supporting them within the RNA seq data. This file is a plain-text, tab-separated file with a single header line (which names all of the columns). This file is named `quant.sf` and appears at the top-level of Salmon's output directory. The columns appear in the following order:

Name	Length	EffectiveLength	TPM	NumReads

- **Name** — This is the name of the target transcript provided in the input transcript database (FASTA file).
- **Length** — This is the length of the target transcript in nucleotides.
- **NumReads** — This is salmon's estimate of the number of reads mapping to each transcript that was quantified. It is an "estimate" insofar as it is the expected number of reads that have originated from each transcript given the structure of the uniquely mapping and multi-mapping reads and the relative abundance estimates for each transcript.

How do you calculate genome length in UNIX?

1. Download the human chromosome 1:

```
~:> !wget ftp://ftp.ensembl.org/pub/release-88/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.chromosome.1.fa.gz
```

1. Gunzip the gz file:

```
~:> !gunzip Homo_sapiens.GRCh38.dna.chromosome.1.fa.gz
```

1. For convenience, Change the name of the file to chr1.fasta.

```
~:> !mv Homo_sapiens.GRCh38.dna.chromosome.1.fa.gz chr1.fasta
```

1. Calculate genome length

```
~:> !genome_length=$(grep -v ">" chr1.fasta | wc | awk '{print $3-$1}')
```

```
~:> !echo "$genome_length"
```

Genome length Calculation: code walkthrough

Given a reference sequence in a fasta file, calculate its length by summing the number of bases after the headers which are marked by the '>' character.

1) **Subtract all lines in the fasta file that contain the '>' key character.**
2) **Pipe the output to the wc command, which will return word/line/character counts.**
3) **Parse the wc output by piping to the awk command and subtract the number of characters from the number of lines beginning with '>'.**
4) **Print the genome_length variable containing the result.**

```
                                        1                2              3
~:> !genome_length=$(grep -v ">" chr1.fasta | wc | awk '{print $3-$1}')

~:> !echo "$genome_length"
          4
```

How do you calculate coverage?

Before an accurate calculation of coverage can be determined, it is preferred to align the reads to a reference sequence. This ensures that the reads used for calculating coverage actually overlap on the reference sequence. Accordingly, this alignment will produce a bam file from which we can easily determine the coverage using samtools.

To run this command, samtools requires a sorted bam file and its index. The Bam file provided is already sorted so we simply generate the index before calculating coverage. Samtools automatically checks the current directory for the associated index file. (extension .bai)

Download bam file:
https://drive.google.com/file/d/1l7tVSzvuGQVBNpe3n_anvyjDxjCvCPlW/view?usp=sharing

~:> samtools index sample.sorted.bam

~:> samtools coverage sample.sorted.bam

#rname	startpos	endpos	numreads	covbases	coverage	meandepth	meanbaseq	meanmapq
1	1	248956422	9972	1478152	0.593739	0.00599355	17	36

In this example, we look at the average coverage over the entire chromosome 1. This simulated bam file has extremely low coverage. True sequencing will yield higher coverage. Requires samtools

Throughput = number of reads multiplied by read length

1) Download the simulated fastq.gz file

https://drive.google.com/file/d/1-PgMpn0EgqtdceOyf_XnZcTqGV6O0k6Q/view?usp=sharing

2) Count the number of reads in the fastq file by summing the number of lines and dividing the total by four. Store result in variable Num_Reads.

```
~:> !gunzip sample_name.fastq.gz
~:> !line_count=$(cat sample_name.fastq | wc -l )
~:> !Num_Reads=$(((line_count + 1) / 4))
```

3) Count the number of bases per read. Store Result in variable Bases_Per_Read.

```
~:> !Bases_Per_Read=$(cat sample_name.fastq | awk 'NR==2' | wc -c)
```

4) Count the number of bases per read. Store result in variable Throughput.

```
~:> !Throughput=$((Bases_Per_Read * Num_Reads))
```

5) Print the contents of the throughput variable to the terminal to see the final result.

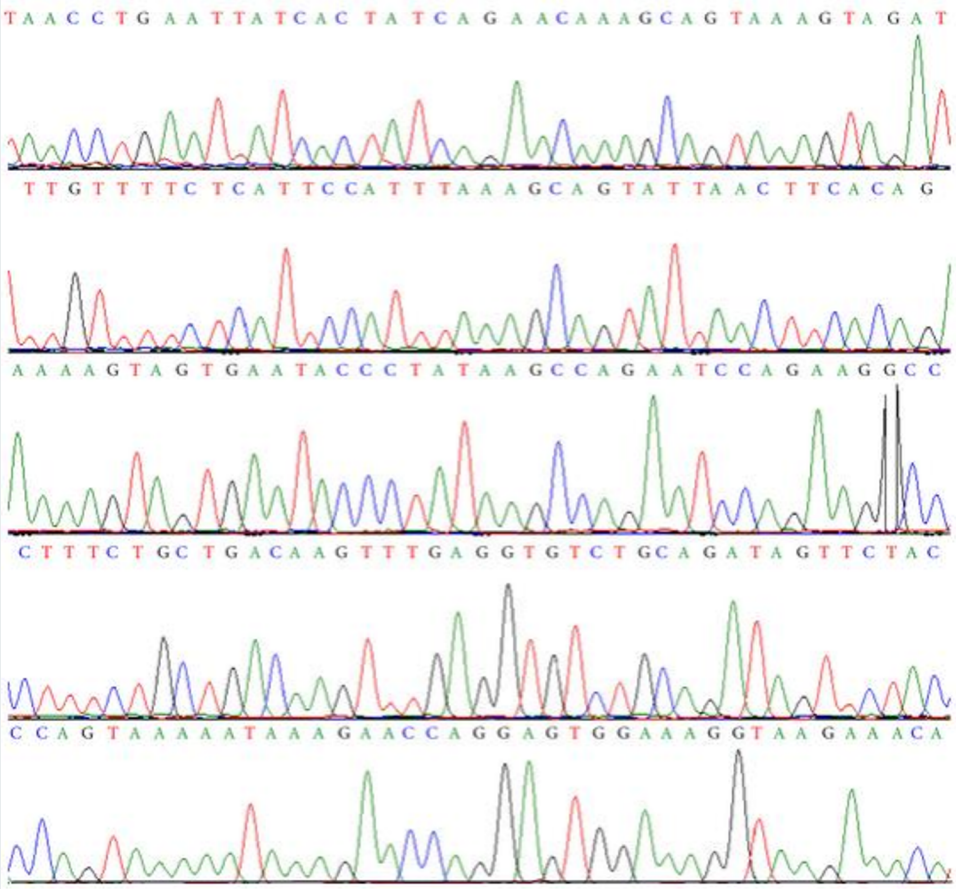
```
~:> !echo "$Throughput"
```

*notes: (See fastq file format for details)

1) Every 4.0 lines in a fastq file is a read.
2) The 2 line contains the first read.

Sequencing Technologies

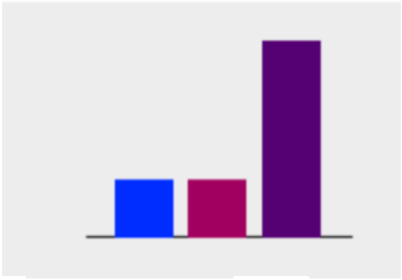
1. Short Read Sequencing Technologies
    a. Illumina
2. Long Read Sequencing Technologies
    a. PacBio
    b. Nanopore



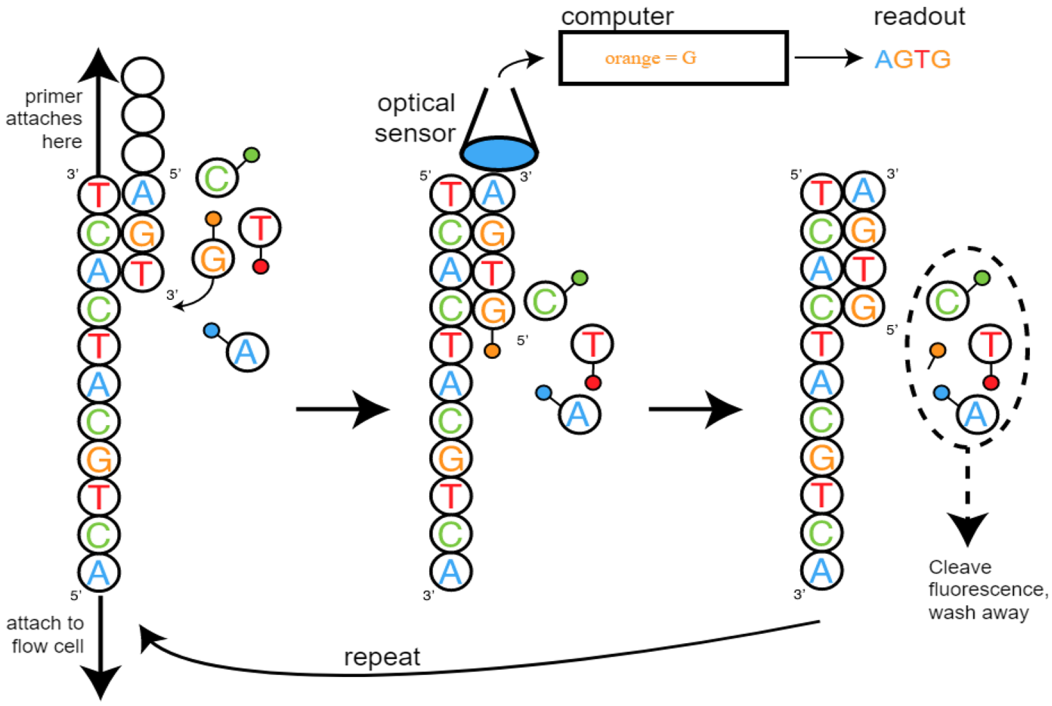
Illumina Sequencing

- Short read sequencing technology
- Sequencing is sequencing-by-synthesis chemistry
- Uses novel reversible terminator nucleotides for the four bases, a different fluorescent dye and a special polymerase enzyme
- Requires PCR amplification

NextSeq 550 by Illumina 2006

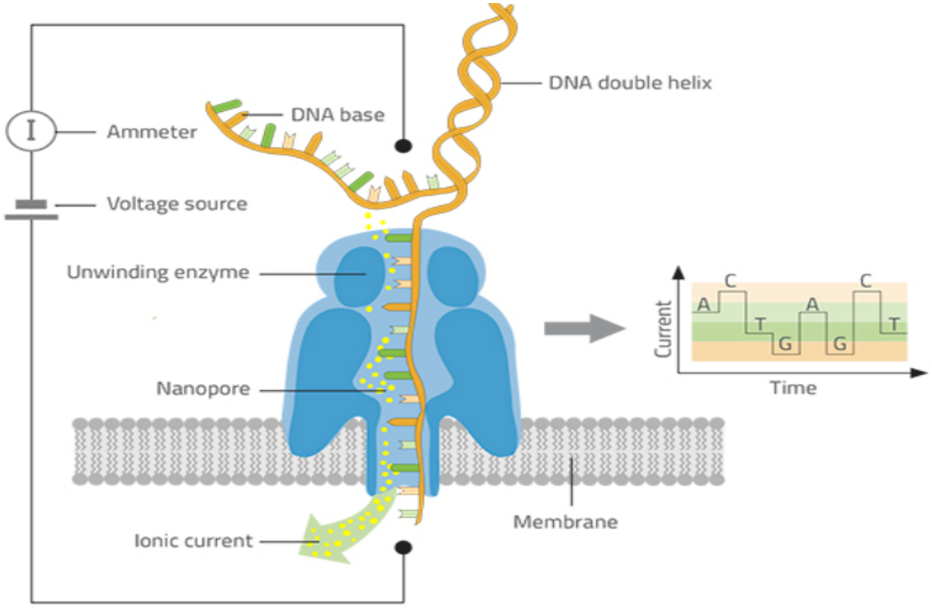


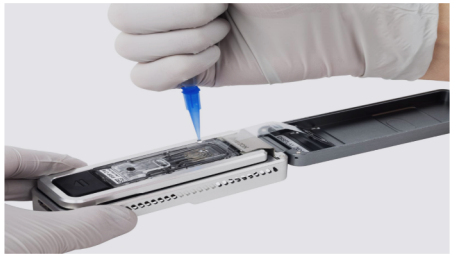
Read Length Error Rate Throughput



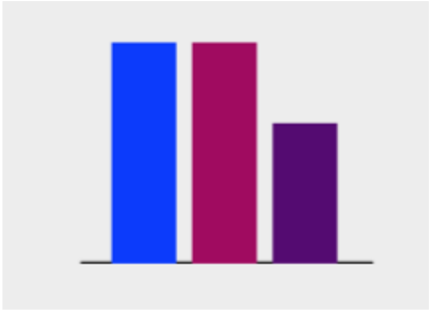
Nanopore

- Sequencing technology utilizes a **nanopore:** Pores of nanometer size located in electrically insulating environment.
- Can sequence any length of fragments of native DNA and RNA resulting in short or long reads.
- No PCR amplification step is needed





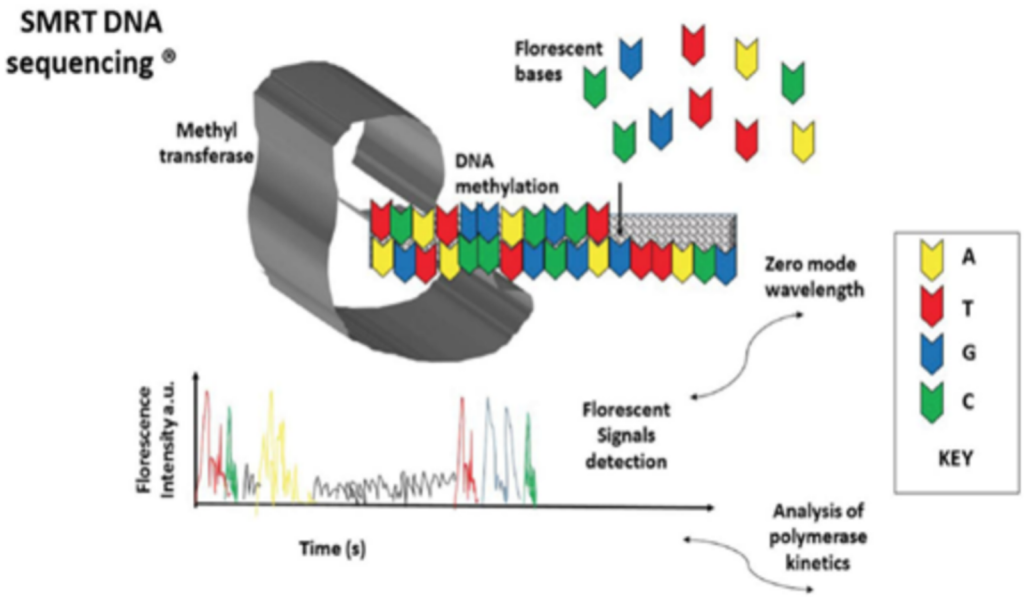
Nanopore sequencing by MinION: Oxford Nanopore Technologies (2014)



Read Length Error Rate Throughput

PacBio

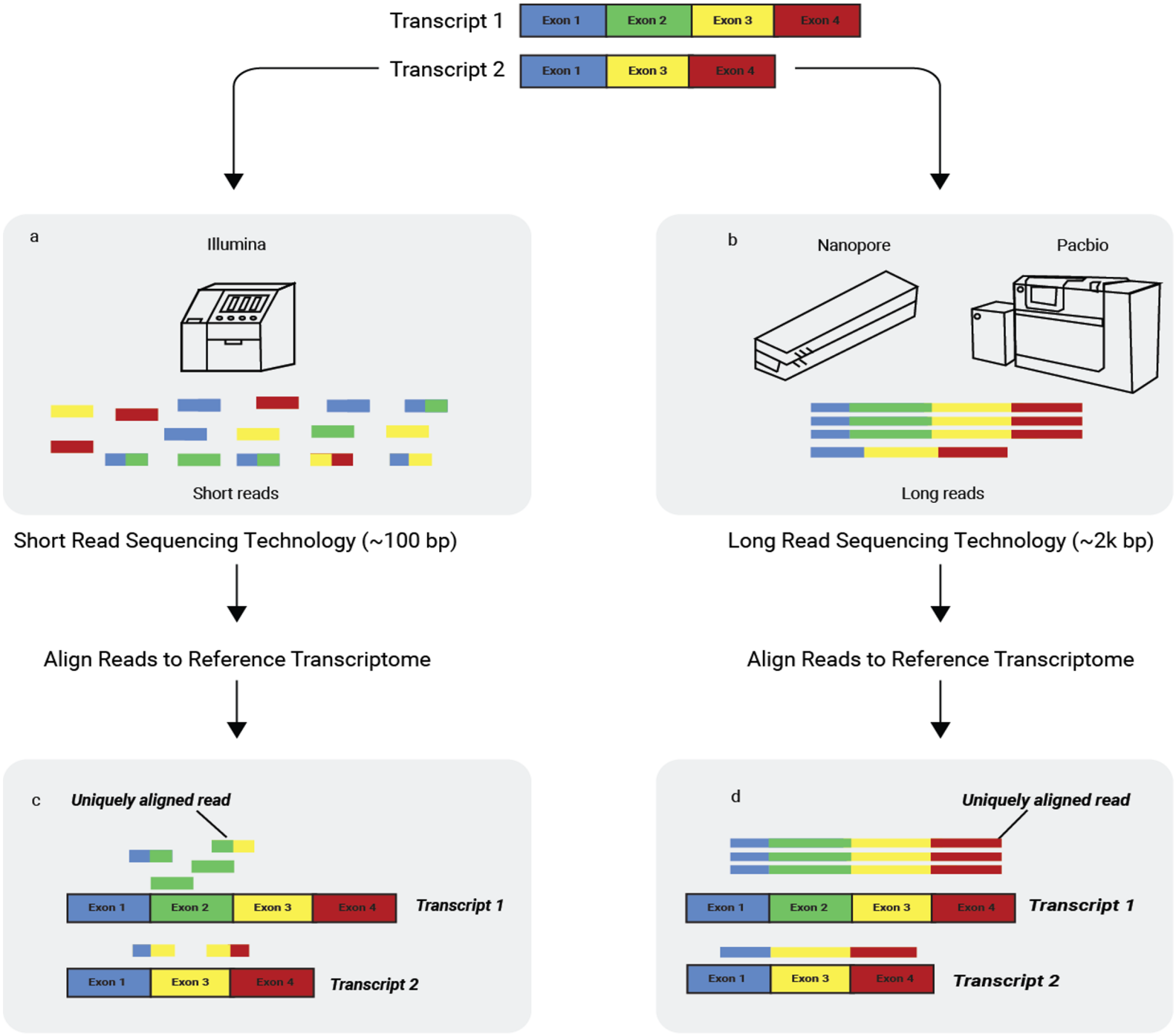
- PacBio (SMRT) sequencing technology
  - No PCR amplification step
  - Sequencing is based on real time detection of a fluorescent signal, emitted when a nucleotide incorporation occurs.



“Single Molecule, Real-Time (SMRT) PacBio by Pacific BioScience (2010)”

Comparison of Sequencing Technologies

- **Throughput:** Long read sequencing technologies < short read sequencing technologies
- **Error rate:** Short read sequencing technologies < long read sequencing technologies
- **Alignment:**
  - Short reads are usually aligned or mapped to multiple locations on the transcriptome and/or genome.
  - Longer reads are more likely to be aligned to a single location.
- **Assemble novel transcripts:**
  - Longer reads are preferred for *de novo* assembly → assembly step more clear
  - Short reads do not span the shared region or shared exon junction → ambiguous assembly .
- **Estimate transcripts and gene expression**:
  - Shorter reads preferred for quantification → higher throughput.
  - Longer reads → lower throughput → determination of the transcript for each read is a straightforward process.


Transcript 1
Exon 1
Exon 2
Exon 3
Exon 4
Transcript 2
Exon 1
Exon 3
Exon 4
a
Illumina
Short reads
b
Nanopore
Pacbio
Long reads
Short Read Sequencing Technology (~100 bp)
Long Read Sequencing Technology (~2k bp)
Align Reads to Reference Transcriptome
Align Reads to Reference Transcriptome
c
Uniquely aligned read
Exon 1
Exon 2
Exon 3
Exon 4
Transcript 1
Exon 1
Exon 3
Exon 4
Transcript 2
d
Uniquely aligned read
Exon 1
Exon 2
Exon 3
Exon 4
Transcript 1
Exon 1
Exon 3
Exon 4
Transcript 2

Quantification of Transcript and Gene Expression

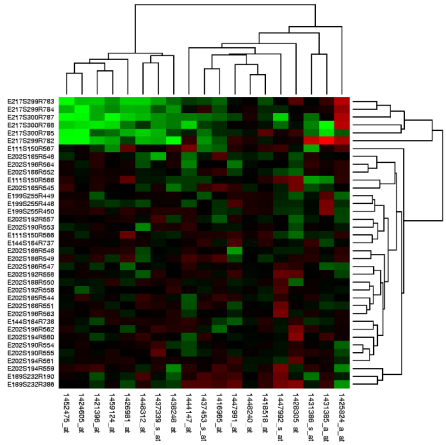
- **Simplest approach for quantification - Extracting gene counts**
  - HT-seq-count, R Count, featureCounts
  - Useful in bioinformatics analysis pipelines - cell type decomposition, quantify differential expression of genes
- Isoforms of transcripts coming from the same gene - highly similar in sequence - share multiple overlapping regions - due to alternative splicing
  - Multi-mapped reads pose a challenge in identifying the true origin of the reads
  - In order to effectively used this method, only the reads mapped to a single transcript can be considered.

- **Transcript level quantification**
  - Recommended for RNA-seq data analysis - more accuracy over traditional methods
  - Probabilistic methods - Cufflinks - estimates isoform abundance from short read information
  - EM algorithm - RSEM - assigns reads to the isoforms from which they originated
- **Emerging approaches - Kallisto and Salmon**
  - Do not require alignment - use raw reads
  - Computationally fast, use comparatively lesser memory
  - Called pseudoalignment
  - These methods leverage the idea that only required information when quantifying reads is ‘which’ transcripts have generated the read and not ‘where’ the transcripts align.

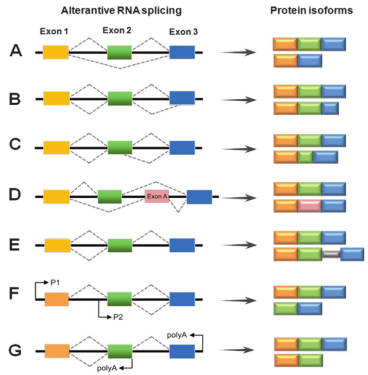
Differential Gene Expression Analysis

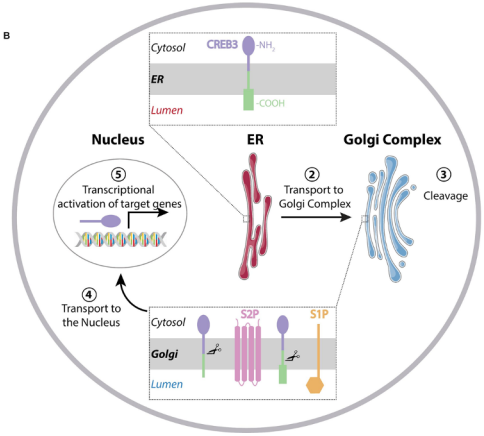
- Differential gene expression can be analysed by normalising, statistically analysing the data or using eQTL analyses.
- Normalisation is controlling the length of transcripts while comparing transcripts of different lengths. This can also be due to batch effects.
- Variations in data obtained can lead to differential expression analyses (DE) which aims to find expression levels of genomic features which are significantly different in distinct samples.
    - Statistical testing - P-value
- Several tools are available for DE analysis - DESeq2, EdgeR

Repurposing of RNA-Seq Tools



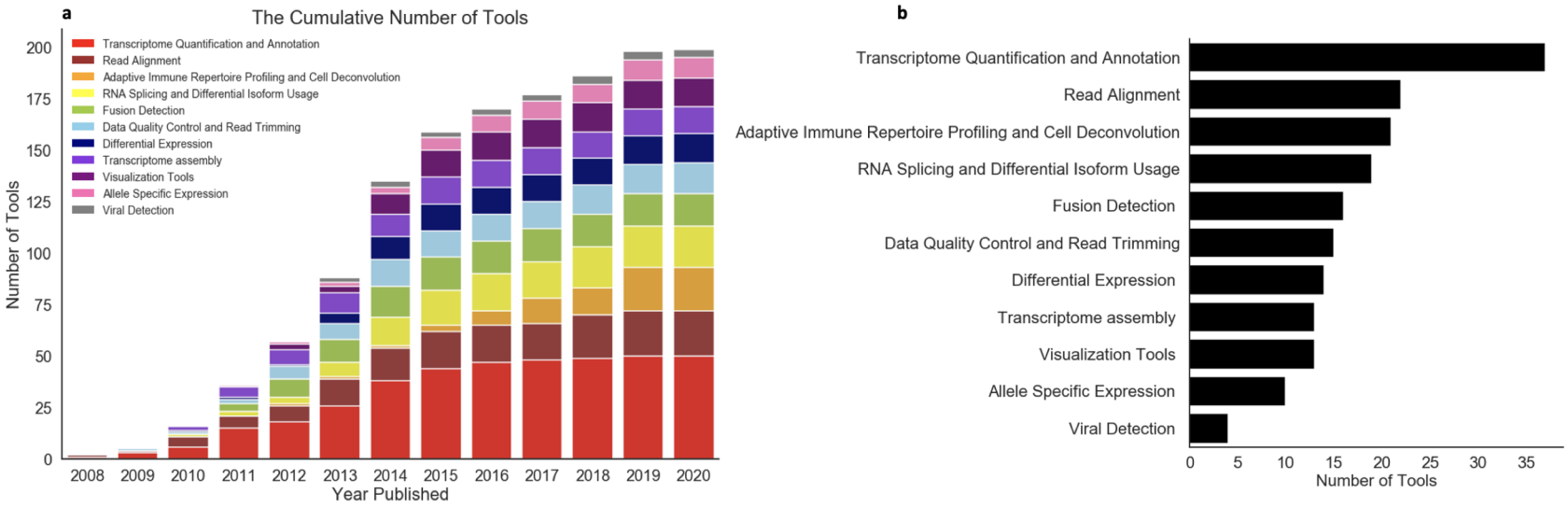
- RNA-seq is capable of addressing biological problems:
  - gene expression profiles across various phenotypes and conditions,
  - detecting novel alternative splicing on specific exons
  - detecting changes in concentration, function, or localization of transcription factors that affect splicing which can lead to downstream neurodegenerative diseases and multiple cancers
  - RNA-seq technology was designed for a new set of computational tools which are capable of **repurposing** RNA-seq data such as individual adaptive immune repertoire and microbial communities of a sample





Survey of Computational Methods for RNA-Seq Analysis

- We surveyed 184 computational tools that are used for RNA-seq analysis from 2008 to 2020
  - Divided the tools into different categories representing different stages of the RNA-seq analysis workflow
- The rate of new tool development slowed after 2015 and the average annual growth rate in available tools from 2015 to 2020 was 6.81%
- Of the 184 tools currently available, the largest category of RNA-seq tools (n=43) are designed for Transcriptome Quantification Annotation, and the smallest category of tools (n=4) is Viral Detection



- The percentage of required computational expertise of current computational tools that are used for RNA-seq analysis in different categories based on three-level classification scheme.
- **Most tools require medium to high level computational expertise.**
  - → **Low level of computational expertise:** RNA-seq tools with a GUI or web-based interface; in other words, the user is not required to use the command line.
  - → **Medium level of computational expertise:** RNA-seq tools that require the user to implement R commands and navigate package managers
  - → **High level of computational expertise:** RNA-seq tools that require the user to be fluent in programming languages such as C, C++, Java, Python, Perl, etc.
- The percentage of availability of package managers (Anaconda, Bioconductor or CRAN, or no package manager) for tools in different categories.
- **Most tools had no availability of package manager.**