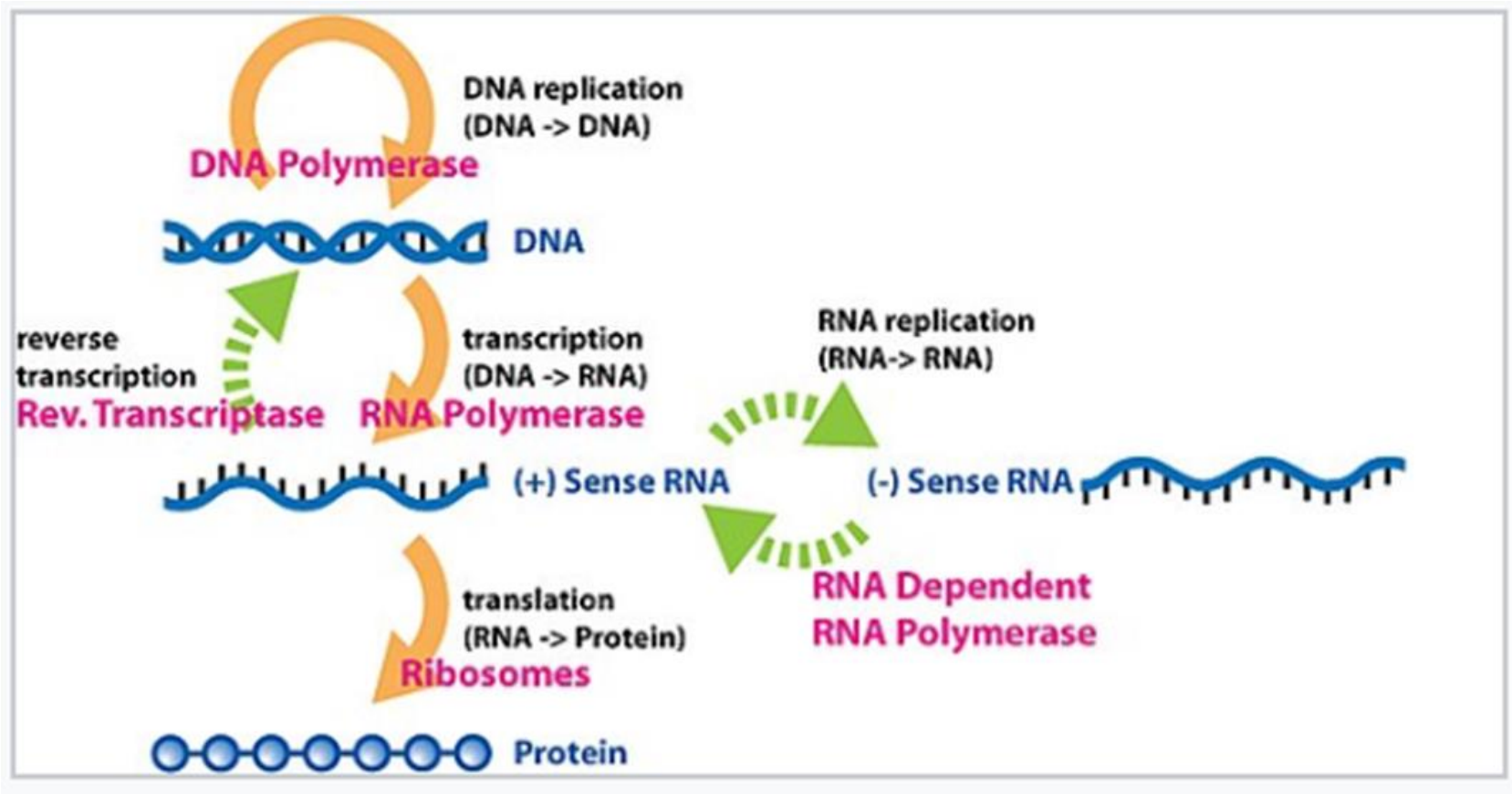
DNA sequencing Informatics in MATLAB

Central Dogma of Biology



Genomes and genomics

Genomes: the collection of all DNA of a species
Genetics: the study of the inheritance of phenotype
Genomics: the study of genomes

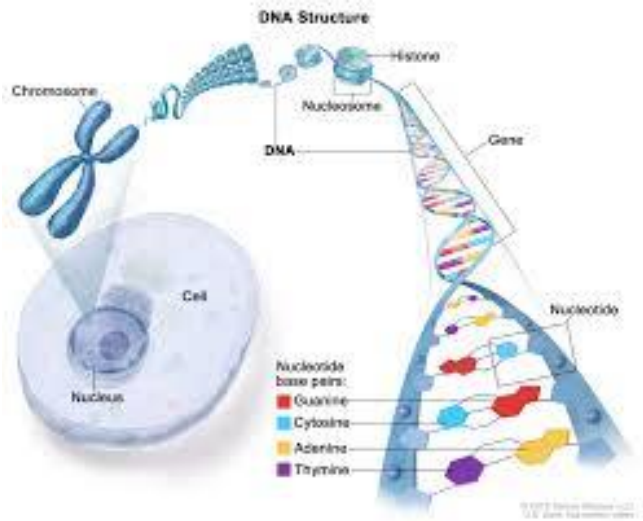
Genome size: the number of A C G T in a genome (bp)

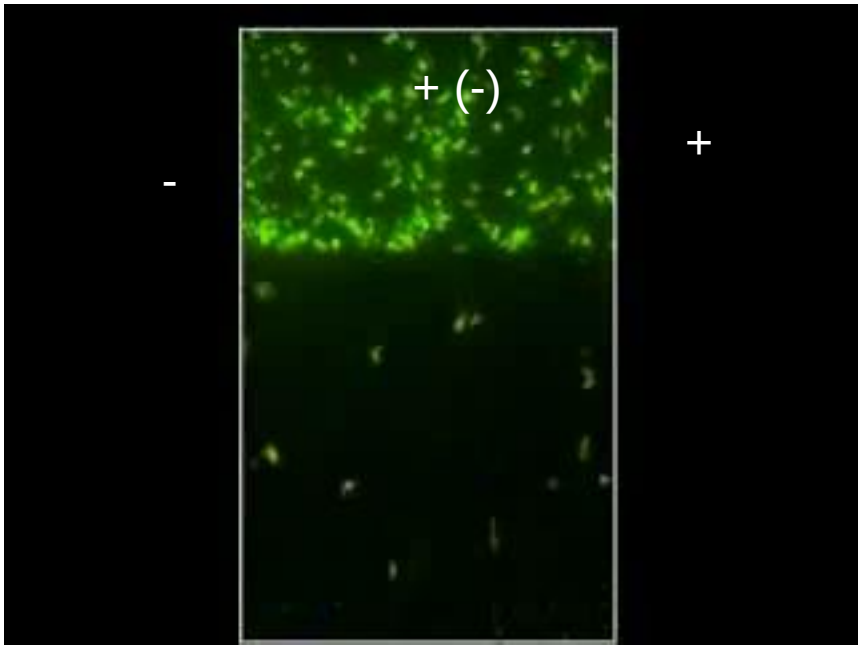
Species	*T2 phage*	*Escherichia coli*	*Drosophila melanogaster*	*Homo sapiens*	*Paris japonica*
Genome Size	170,000 bp	4.6 million bp	130 million bp	3.2 billion bp	150 billion bp
Common Name	 Virus	 Bacteria	 Fruit fly	 Human	 Canopy Plant

Largest genome

http://ib.bioninja.com.au/standard-level/topic-3-genetics/32-chromosomes/genome-size.html

DNA molecules inside nano-channels





Human Genome Overview

Mitochondrial genome

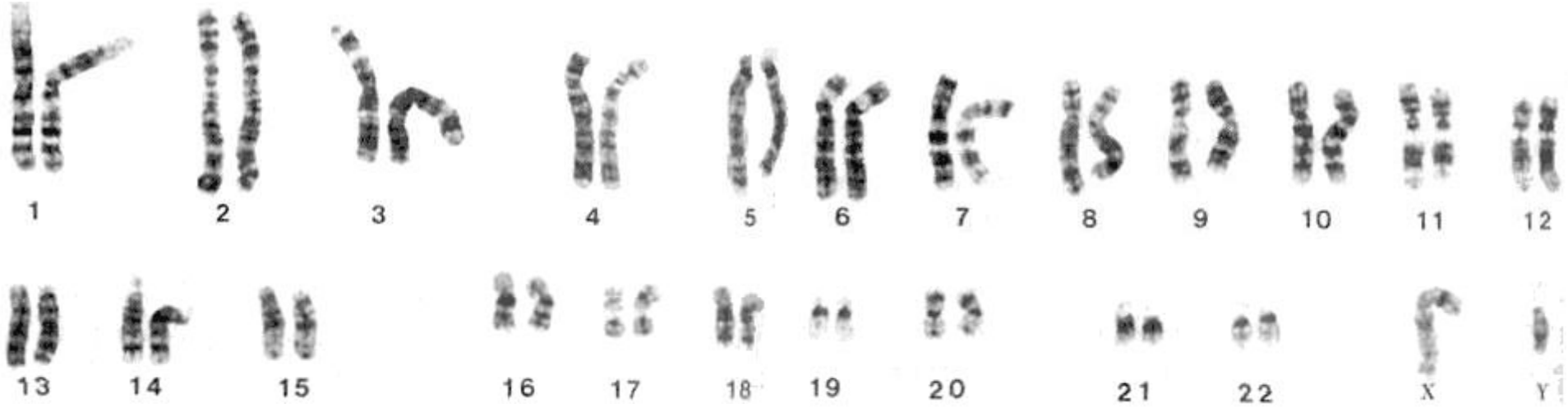
Nuclear genome (human genome)

Mitochondrial genome

- 16,569 bp, 37 genes, 44% (G+C)
- 93% of mt DNA is for protein coding, all genes lack introns.
- Human cells vary in the number of mt DNA molecules (typically thousands of copies/cell).
- Higher mutation rate in the mitochondrial genome
- Human sperms have very few mt molecules (<5), and degradation of mt molecules are frequent.
- mtDNA is inherited from the mother (maternally inherited).

Human Genome

- 30 to 40 trillion human cells? The majority are somatic cells (few germline cells)
- Each somatic cell contains two nearly identical sets of DNA molecules. Germline cells have only one set
- They come from your parents and are organized in 23 pairs of chromosomes
- Maternal and paternal genomes are slightly different but are the same across all the cells in your body.
- 22 autosomal chromosomes numbered according to the size of the chromosomes, plus two sex chromosomes x and y



Genetic Variation

- Describes differences between DNA sequences of individual genomes.
- As each individual has two nuclear genomes (a paternal genome and a maternal genome), genetic variation occurs within the individual as well as between individuals.

Why we study human genome?
Genetic Variations in Disease Studies

case

```
….GCCGTTGAC….
….GCCGTTGAC….
```

control

```
….GCCATTGAC….
….GCCATTGAC….
```

Whole genome association study

Single Nucleotide Polymorphisms (SNPs)

Individual 1	paternal	A C T T A G C T **T** A C	heterozygous
	maternal	A C T T A G C T **C** A C	
Individual 2	paternal	A C T T A G C T **T** A C	homozygous
	maternal	A C T T A G C T **T** A C	
Individual 3	paternal	A C T T A G C T **C** A C	
	maternal	A C T T A G C T **C** A C	

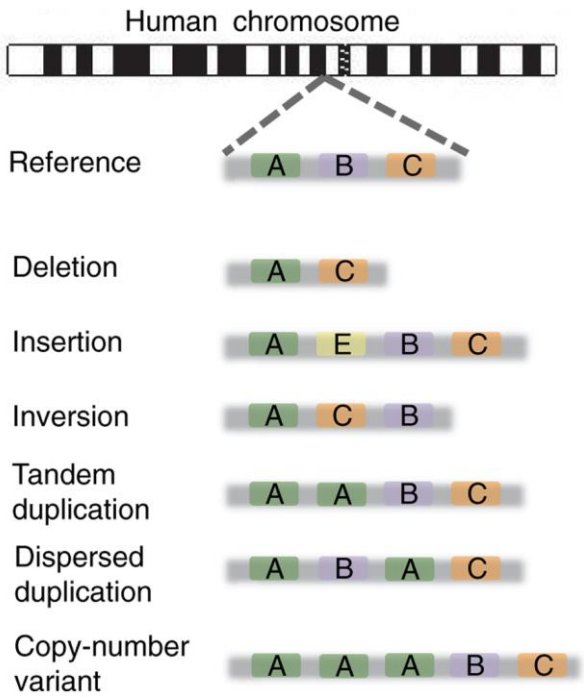
The nucleotide on a SNP locus is called

- a major allele (if allele frequency > 50%), or
- a minor allele (if allele frequency < 50%).

94% → A C T T A G C T **T** ← **T: Major allele**

6% → A C T T A G C T **C** ← **C: Minor allele**

Large structural variations



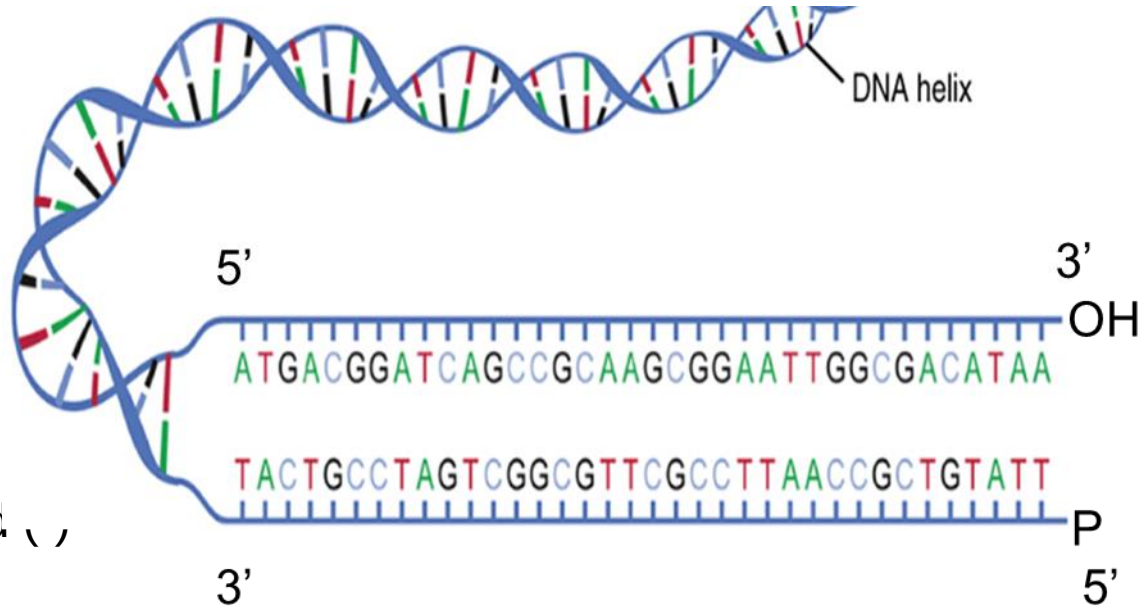
1kb to hundreds of kb in size

Associated with many genetic diseases

Hard to detect by most genetic analysis tools

Other reasons/applications for studying Genomes/Genetics (any genomes)

Convention of writing DNA sequences



Forward strand
Sequence (+)

Complementary strand (-)

Always write from 5'->3'

```
Sequence (+):                          5′ ATGACGGATCAGCC 3′
complementary strand (-):              3′ TACTGCCTAGTCGG 5′
Reverse complementary strand           5′ GGCTGATCCGTCAT 3'
```

Informatics: FASTA Format

```
>hg38_dna range=chr1:11102837-11267747 5'pad=0 3'pad=0 strand=+
repeatMasking=none
TGGACAACATGGCAAGAATCAGTCTCTACAGAAAATACAAAAATTAGCCG
AGTGTGATGGCATGCACCTGTAGTCCCAGCTACTCAGGAGGCTGAGGTGG
GAGGATAACTCGTGCCCGGGAGGTGGAGGTTGCAGTAAGCTGAGATTGCA
CCACTGCCCCCCAGCCCGGGTGATAGTGCCAGACCTTGTCTCCAAAAAAA
AAAAAAAAAAAAAAAAAAAAAGAAAAAAATAGTTAACCTGTTAACAACAT
CTGGTGATTAGGGGCTCCCATCCCTTTGAGCAGTCGAAAATCCACATATA
GCTTTTAACTCCCCCAAAACTTAACTAATAGCCTCCTGTTGACCAAAGGC
CTTACCGATAACAAACGGTCAACTGACACATATTGTGTGTATGTATTGTA
TTCTTACAATAAATTACGAAAATTTTAAGGAAAATATATTTACTATTCAT
TAAGTGGATCATCATAAAGATCTTCATCCTTGTTTATGTATTTATTTTTA
GAGATAGGGTCTCTGTCATCCAGGCTGGAGTGCAGTGGCACAATCATAGC
CCACTGCCCCCCAGCCCGGGTGATAGTGCCAGACCTTGTCTCCAAAAAAA
AAAAAAAAAAAAAAAAAAAAAGAAAAAAATAGTTAACCTGTTAACAACAT
CTGGTGATTAGGGGCTCCCATCCCTTTGAGCAGTCGAAAATCCACATATA
CCACTGCCCCCCAGCCCGGGTGATAGTGCCAGACCTTGTCTCCAAAAAAA
AAAAAAAAAAAAAAAAAAAAAGAAAAAAATAGTTAACCTGTTAACAACAT
CTGGTGATTAGGGGCTCCCATCCCTTTGAGCAGTCGAAAATCCACATATA
```

Large data size of a human genome

- The last page contains about 850 characters
  Need 5,000,000 pages to print the human genome
- Some large projects study the DNA of 100,000 people
- Some large projects sequence single cells from different tissue types
- Some large projects sequence the same individual at different ages
- There are many species on this planet to be sequenced

DNA sequencing

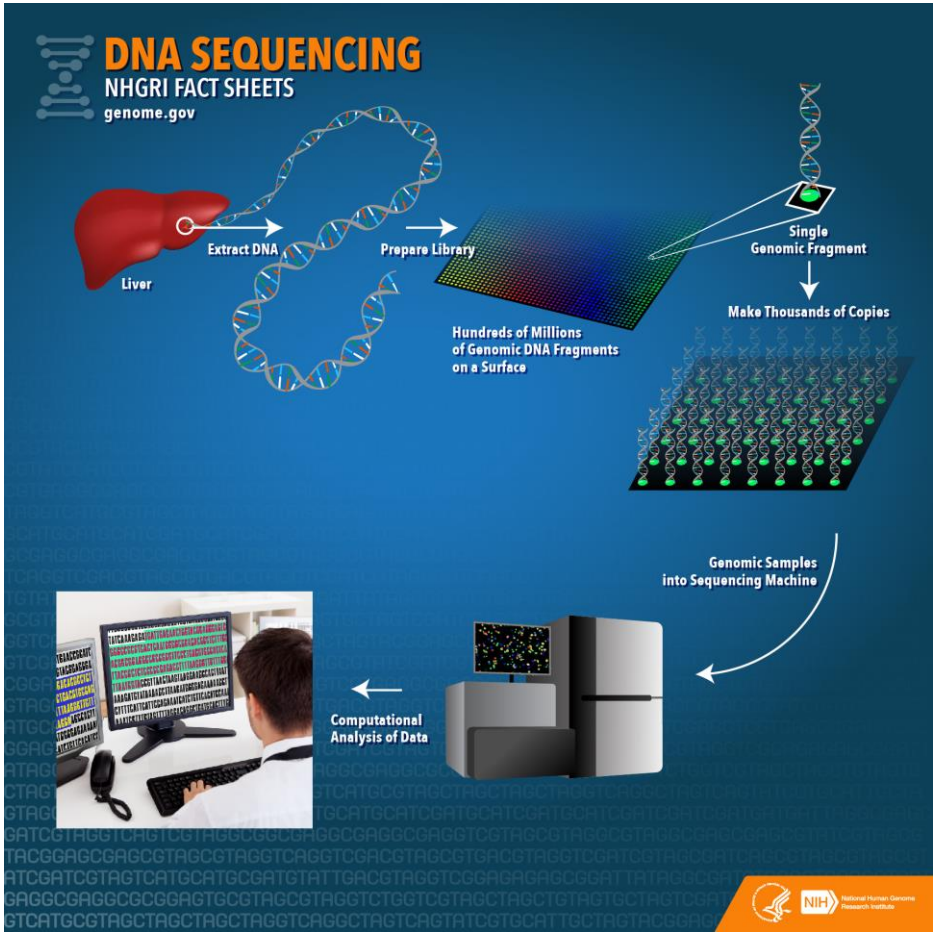
The process of determining the order of nucleotides in a DNA sequence

Sample input:

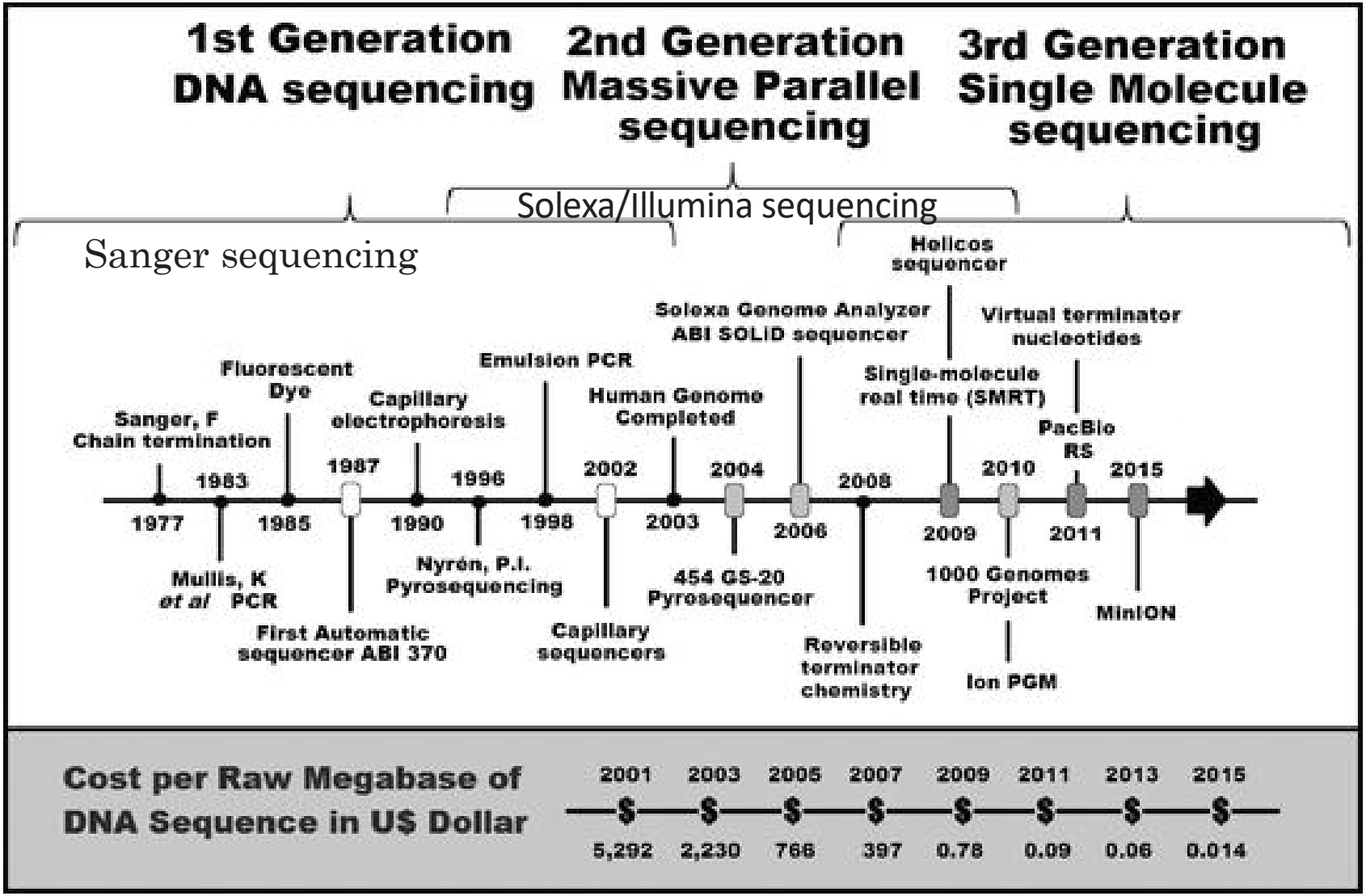
- Purified the DNA molecules from blood, tissue, cell….
- Sequencing library using purified DNA sample

Output:

- The DNA sequences in the form of strings of `A`, `C`, `G` and `T`
- `ACGTGGTGGCCCAAATGGGAGC TGGGAAACTTTCCCCAATGGCG TAGTGTACTGCAGTCAGTCGGG CCTTAAACCCGGTAGAAACTGA CGGGGCCATGGCCCAAAGATGT GTCAAA`



https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Fact-Sheet



Pereira, M.A., Malta, F.S.V., Freire, M.C.M. and Couto, P.G.P. (2017) In Marchi, F. A., Cirillo, P. D. R. and Mateo, E. C. (eds.), *Applications of RNA-Seq and Omics Strategies - From Microorganisms to Human Health*. InTech, Rijeka, pp. Ch. 13.

DNA sequencing technologies

40 year Perspectives; https://www-nature-com.ezproxy2.library.drexel.edu/collections/gljmxjkqgv

Sanger sequencing

sequence-by- synthesis' (SBS) techniques,

Pyrosequencing was later licensed to 454 Life Sciences,

Illumina sequencing

ABI PRISM range developed from Leroy Hood's research, produced by Applied Biosystems

(SOLiD) system from Applied Biosystems (which became Life Technologies-Thermosfisher

Complete Genomic's 'DNA nanoballs' technique,

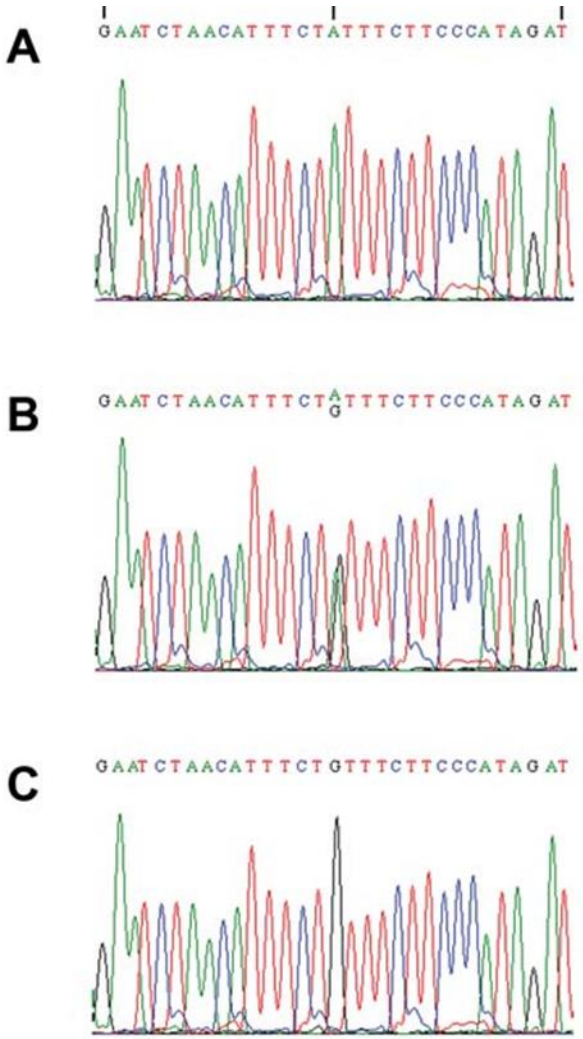
Ion Torrent (another Life Technologies product)

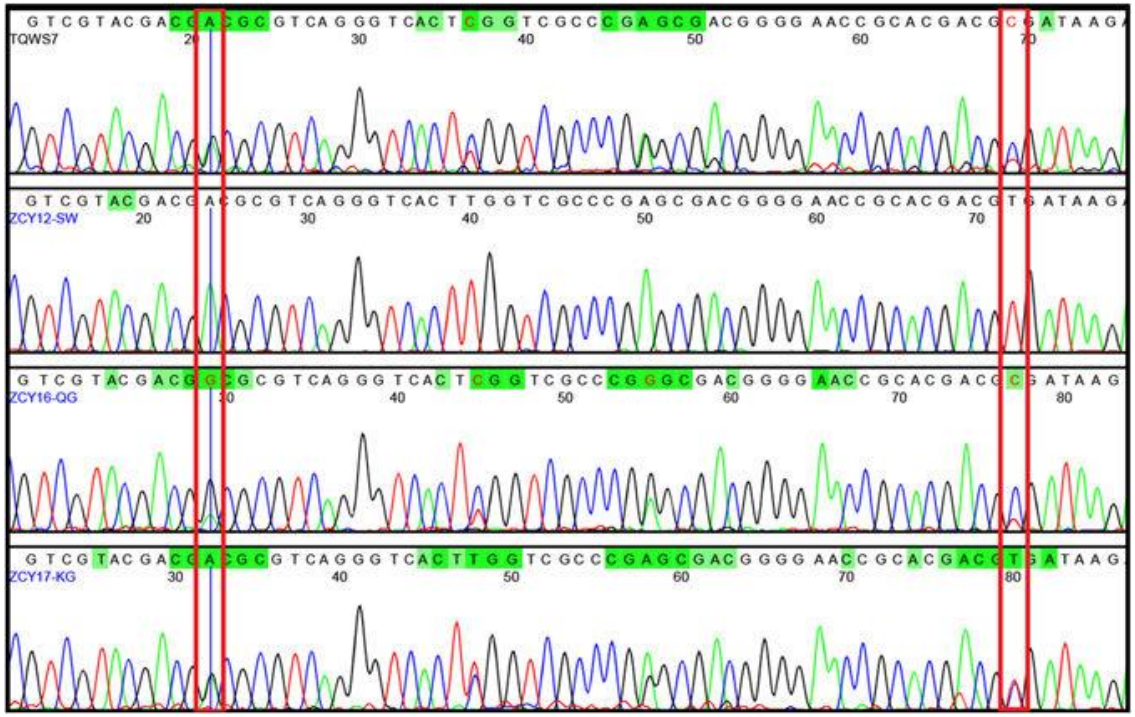
Helicos BioSciences

Pacific Biosciences

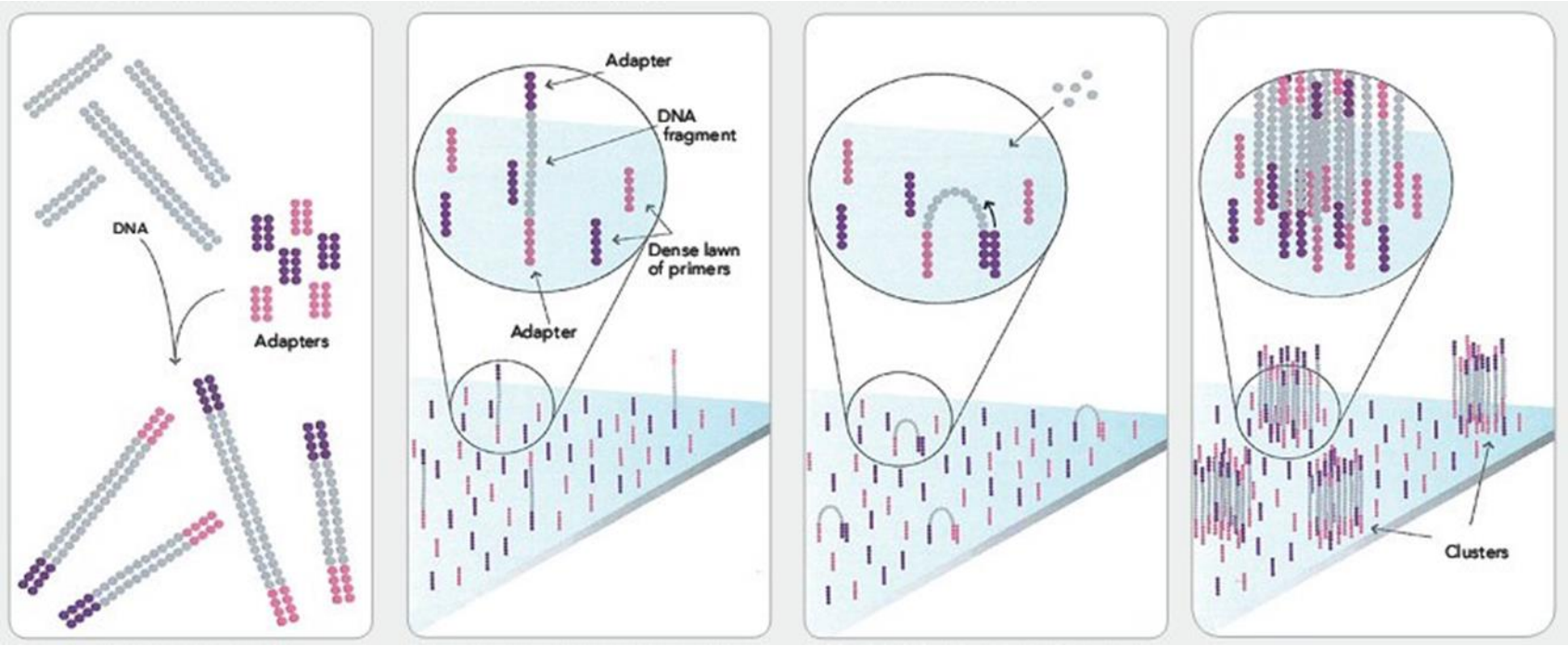
Oxford Nanopore Technologies (ONT)

Sanger sequencing chromatogram



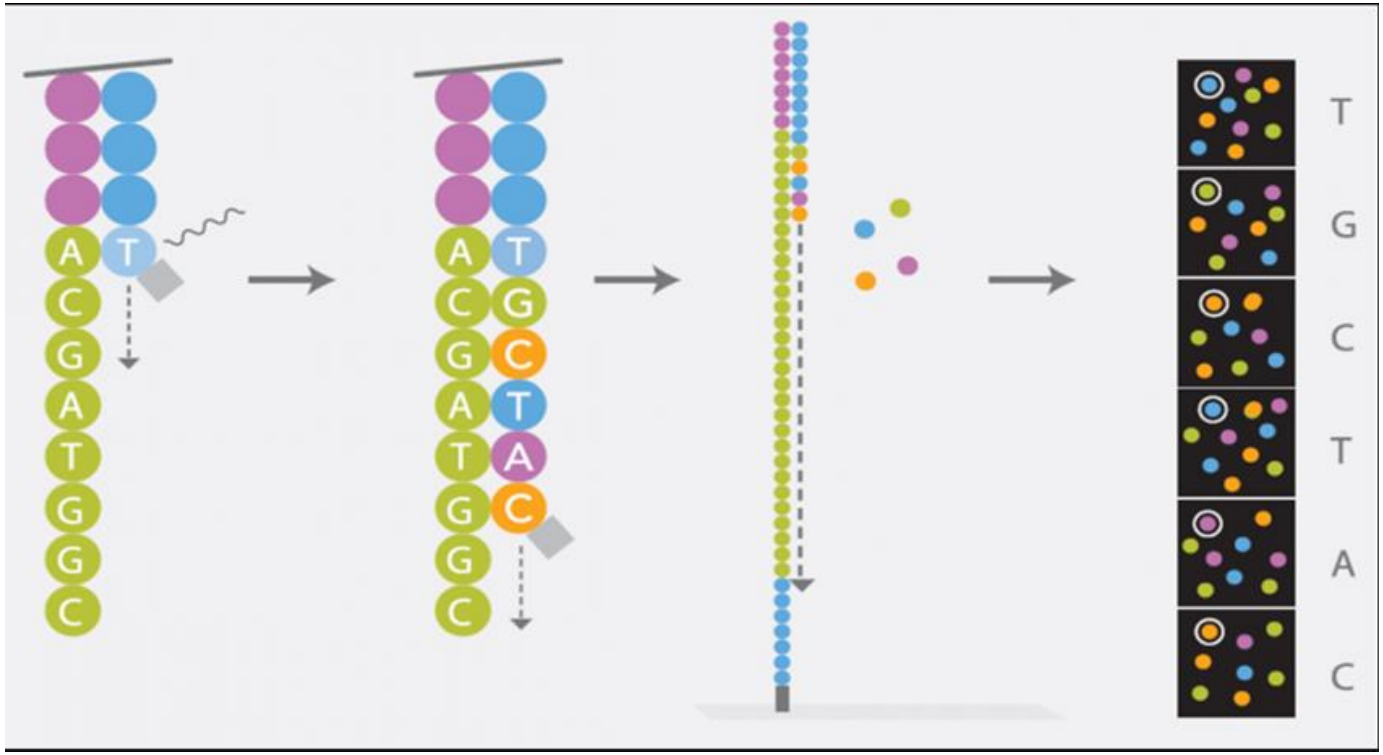


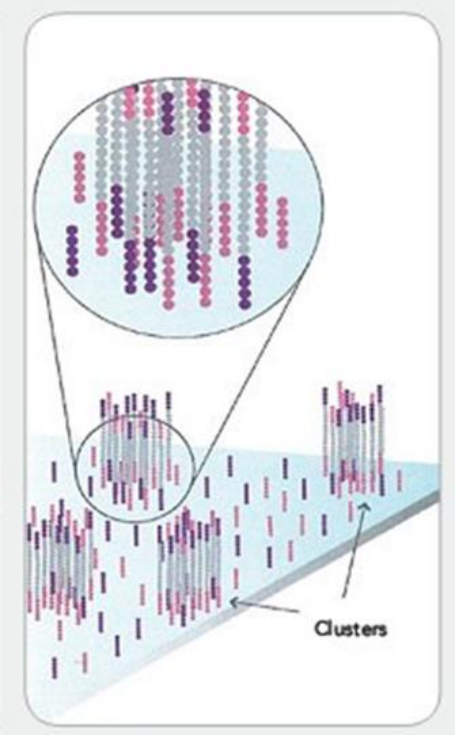
Illumina sequencing (sequencing by synthesis)



https://www.cegat.de/web/wp-content/uploads/2015/10/sequencing-1-1030x558.png

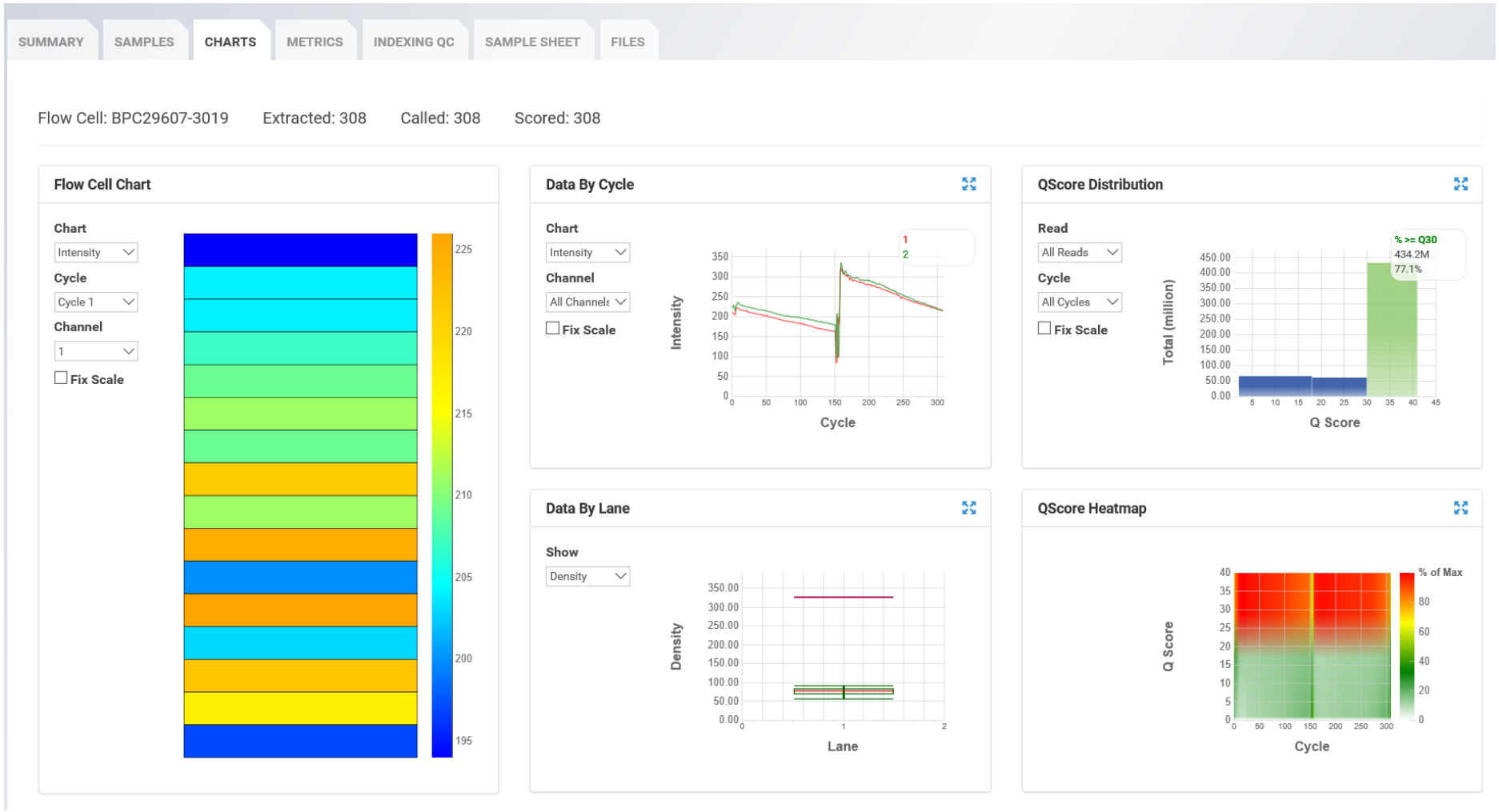
Illumina sequencing (sequencing by synthesis)



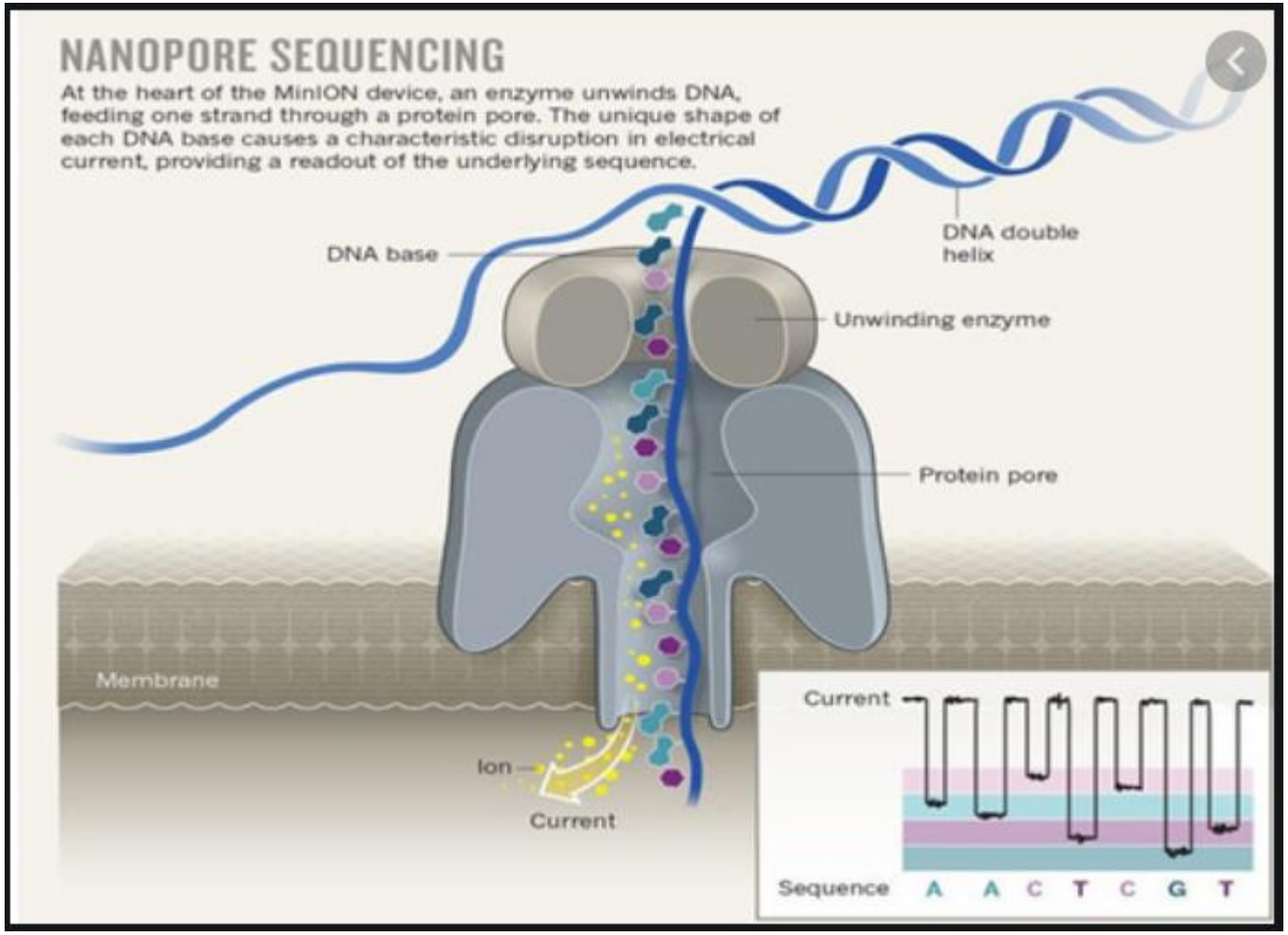


Illumina Desktop Iseq

Illumina sequencing (sequencing by synthesis)



Oxford nanopore sequencing



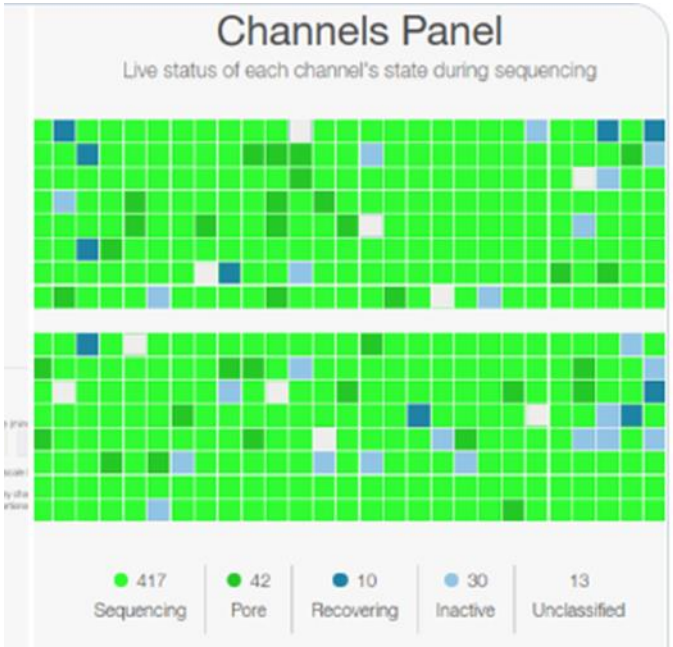
MinION

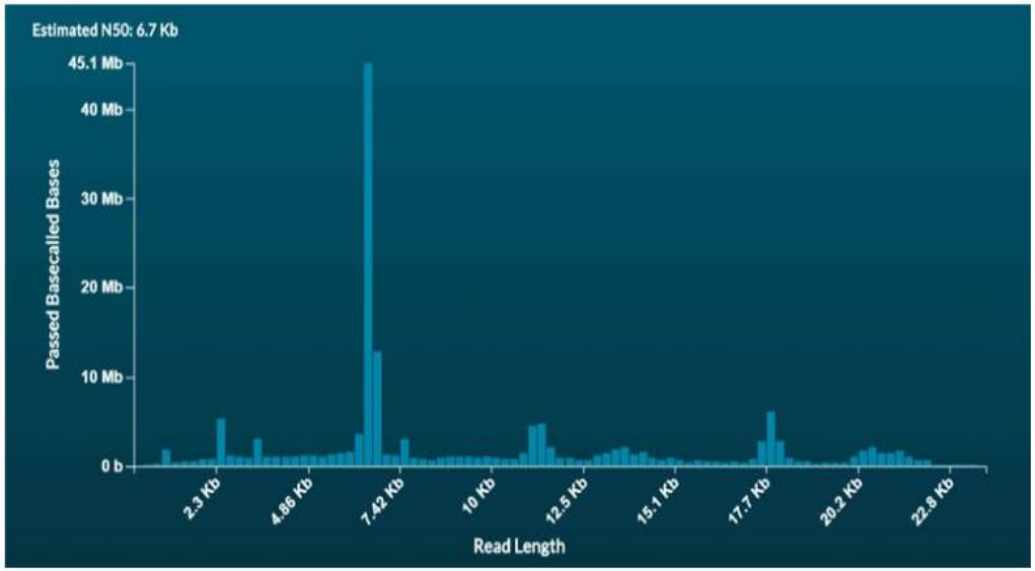
MinION

MinION Mk1C

The only portable, real-time devices for DNA and RNA sequencing

Oxford nanopore sequencing





Informatics: Basic sequencing data analysis

- Raw data analysis (image processing and base calling)
- **Fastq file stores all the sequencing information** (including the quality scores)
- Quality control and read cleaning
- Alignment to the reference genome (SAM/BAM formats)
- *De novo* assembly (pairwise comparison without the reference). Contigs aligned to reference.
- Variant call (VCF formats)
- Bed file, a tab-delimited text file that defines a track

TTGGCAGGCCAAGGCCGATGGATCA +

Understanding Fastq

```
@K00188:208:HFLNGBBXX:3:1101:1428:1508 2:N:0:CTTGTA
ATAATAGGATCCCTTTTCCTGGAGCTGCCTTTAGGTAATGTAGTATCTNATNGACTGNCNCCANAN
+
AAAFFJJJJJJJJJJJJJJJJJFJJFJJJJJFJJJJJJJJJJJJJJJJ#FJ#JJJJF#F#FJJ#F#
```

```
@EAS54_6_R1_2_1_540_792
TTGGCAGGCCAAGGCCGATGGATCA
+
;;;;;;;;;;7;;;;;-;;;3;83
```

4 or more lines

- A line starting with @, showing the sequence ID.
- One or more lines that contain the sequence called by the machine. A,C,G,T,N
- A new line starting with the character +, empty or repeat the sequence ID.
- One or more lines that contain the quality scores (phred score) using ASCII code.

http://maq.sourceforge.net/fastq.shtml

Understanding Quality scores

$$Q = -10 \ \log_{10} P$$

Where P is the probability that the corresponding base call is incorrect.

Phred quality score	Probability that the base is called wrong (P)	Accuracy of the base call	
10	1 in 10	90%	-10*log(0.1)
20	1 in 100	99%	
30	1 in 1000	99.90%	-10*log(0.001)
40	1 in 10000	99.99%	
50	1 in 100000	100.00%	

ASCII characters representing the phred score

ASCII (American Standard Code for Information Interchange) encoding a typical computer key board starting from !

33	!	65	A	97	a
34	"	66	B	98	b
35	#	67	C	99	c
…	…	…	…	…	…

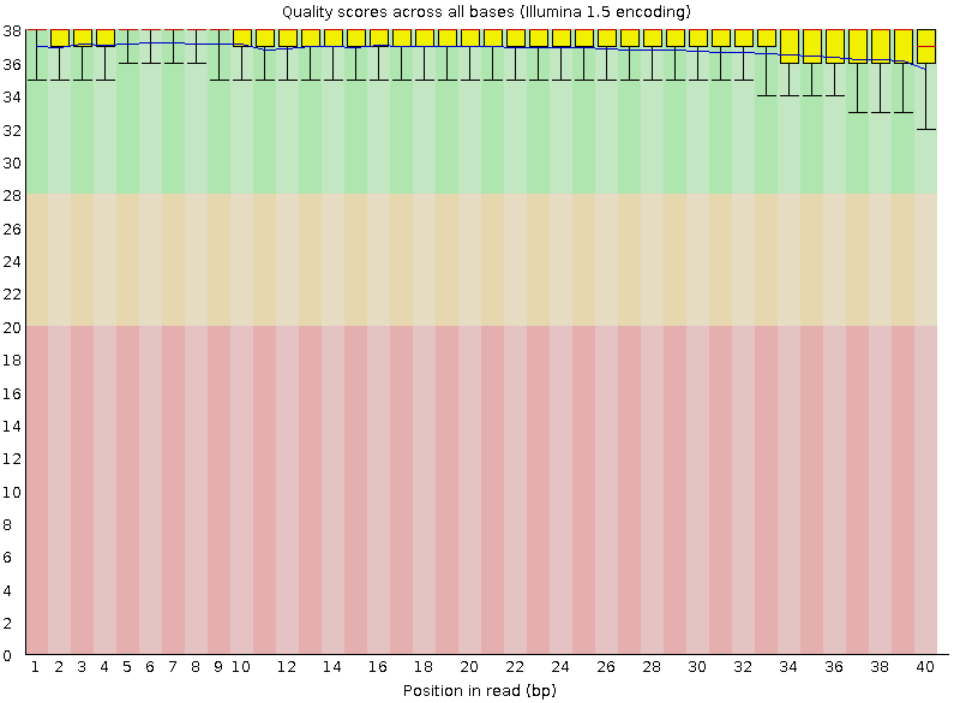
- Standard: <score> = <ASCII number of character> – 33;
  - for example, ‘5’ has an ASCII number of 53, which means the last base has a quality score of 53 – 33 = 20, i.e., error probability is $10^{-(20/10)} = 0.01$
- Illumina has a different standard (with different versions)

Informatics: Basic sequencing data analysis

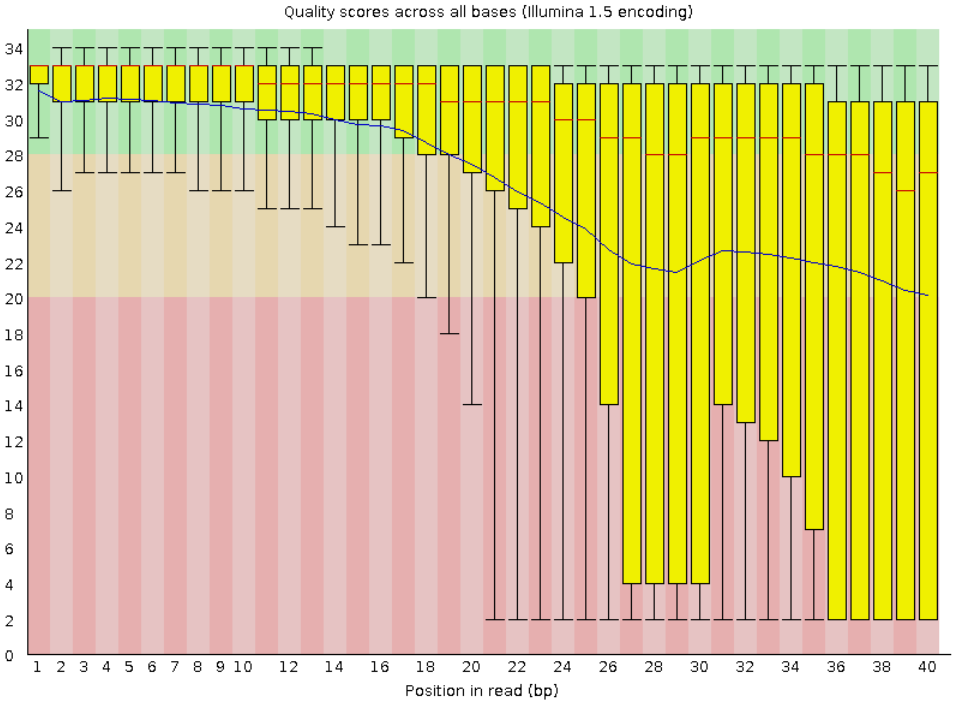
- Raw data analysis (image processing and base calling)
- Fastq file stores all the sequencing information (including the quality scores)
- **Quality control and read cleaning**
- Alignment to the reference genome (SAM/BAM formats)
- *De novo* assembly (pairwise comparison without the reference)
- Variant call (VCF formats)
- Bed file, a tab-delimited text file that defines a track

Quality control and read cleaning

A good case



A bad case

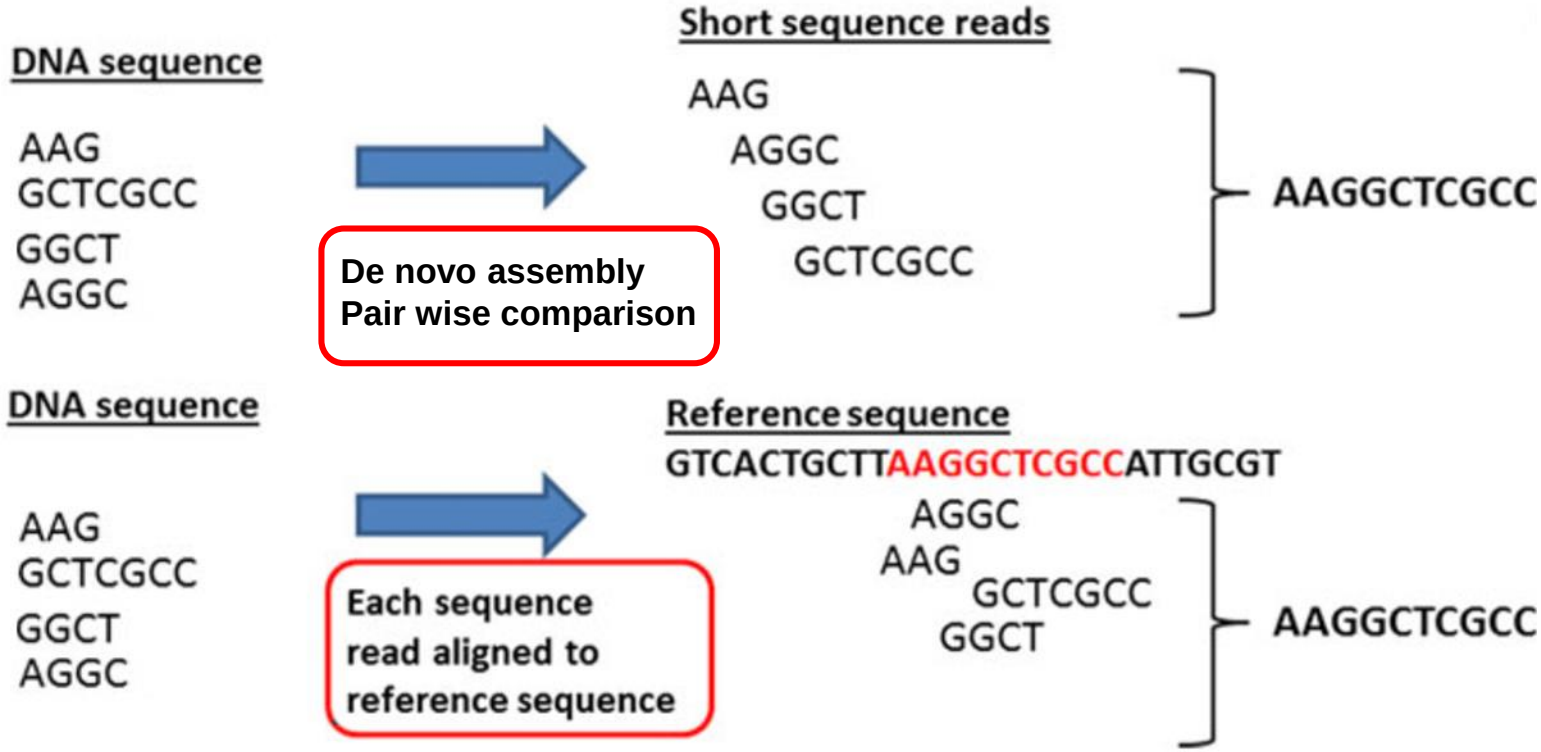


https://blog.horizondiscovery.com/diagnostics/the-5-ngs-qc-metrics-you-should-know

Informatics: Basic sequencing data analysis

- Raw data analysis (image processing and base calling)
- Fastq file stores all the sequencing information (including the quality scores)
- Quality control and read cleaning
- **Alignment to the reference genome (SAM/BAM formats)**
- ***De novo* assembly (pairwise comparison without the reference)**
- Variant call (VCF formats)
- Bed file, a tab-delimited text file that defines a track

De novo assembly vs. alignment to reference sequence



https://pubmed.ncbi.nlm.nih.gov/25225743/

Alignment to the reference genome

Query
Reference

```
   18 tttcgacctgttattcgcatactattaaagaattctttggtagctctcaa    67
      .||||||||||||||| |||| |||||||||||||||||||||||||||||
 1348 attcgacctgttatt-gcat-ctattaaagaattctttggtagctctcaa  1395

   68 ttatcacaattcatggaccaagcaaacccattagctgagttaacgaataa   117
      ||||||||||||||||||||||||||||||||||||||||||||.||||
 1396 ttatcacaattcatggaccaagcaaacccattagctgagttaacgcataa  1445

  118 acgtcgtctatcagcattaggacctggtggtttaacacgtgaacgtgctc   167
      ||||||||||||||||||||||||||||||||||||||||||||||||||
 1446 acgtcgtctatcagcattaggacctggtggtttaacacgtgaacgtgctc  1495

  168 aaatggaagtacgtgacgttcactactctcactatggccgtatgtgtcca   217
      ||||||||||||||||||||||||||||||||||||||||||||||||||
 1496 aaatggaagtacgtgacgttcactactctcactatggccgtatgtgtcca  1545

  218 attgaaacacctgagggaccaaacattggattgattaacttattatcaag   267
      ||||||||||||||||||||||||||||||||||||.|||||||||||
 1546 attgaaacacctgagggaccaaacattggattgattaactcattatcaag  1595

  268 ttatgcacgtgtaaatgaattcggctttattgaaacaccatatcgtaaag   317
      ||||||||||||||||||||||||||||||||||||||||||||||||||
 1596 ttatgcacgtgtaaatgaattcggctttattgaaacaccatatcgtaaag  1645

  318 ttgatttagatacacatgctatcactga--aaa----------aac----   351
      |||||||||||||||||||||||||||  |||          |||
 1646 ttgatttagatacacatgctatcactgatcaaattgactatttaacagct  1695
```

Alignment to the reference genome

Local alignment and global alignment

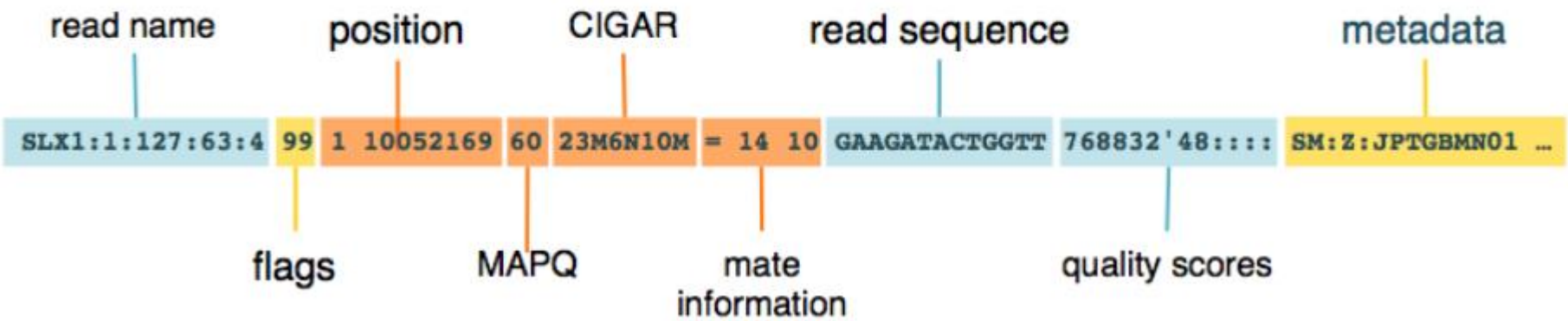
There are many sequence aligners and many short reads aligners

- Bowtie
- BWA
- CLC bio
- Karma
- MAQ
- SOAP/SOPA2
- ZOOM
- …….
- Over 100

SAM/BAM file format

- The Sequence Alignment/Map (SAM) format:
- Format for the storage of sequence alignments and their mapping coordinates
- Supports different sequencing technologies
- Flexible in style, compact in size.
- BAM is the binary version of the SAM format

HEADER containing metadata (sequence dictionary, read group definitions etc)
RECORDS containing structured read information (1 line per read record)



https://gatkforums.broadinstitute.org/gatk/discussion/11014/sam-bam-cram-mapped-sequence-data-formats

SAM/BAM file imported in the MATLAB

```
ans = struct with fields:
            QueryName: '003b4613-7c0f-4461-a2fc-e238bc1f2627'
                 Flag: 0
             Position: 1
        MappingQuality: 60
           CigarString: '28S18M1D84M1D9M1I11M1I11M1D3M2I35M1I18M1I10M1D18M2I1
          MatePosition: 0
            InsertSize: 0
              Sequence: 'CGTGCCTTCGTTCAGTTACGTATTGCTGGGGCGGCGACCTCGCAGGTTTCGC
               Quality: '#%&%#%%&28D*?,)).,+.5668@@D6/651)-?@='*?<(''*12@?443
                  Tags: [1×1 struct]
        ReferenceIndex: 1
    MateReferenceIndex: 0
```

CIGAR string

```
ref                 GGGCGGCGACCTCGCGGG TTTTCGCTATTTATGAAAATTTTCCGGTTT AAGGCGTTTCCGTTCTTCTTCGTCATAACTTAATGTTTTTAT
seq1  CGGTAGTATTGCTGGGCGGCGACCTCGCGGGTTTTTCGCTATTTATGAAAATTTTCCGGTTTAAAGGCGTTTCCGTTCTTCTTCGTCATAACTTAATG  TTTAT
```

13S 18M 1I 30M 1I 35M 2D

cigar1 13S18M1I30M1I35M2D20M2D4M4D11M2D7M

```
ref
seq2                GGGCGGCGACCTCGCGGGTTTT  CGCTATTTATGAAAATTTTCCGGTTTAAGGCG TTTCCGTTCTTCTTCGTCATAACTTAATGTTTTTAT
GCTTCGTTCAGTTACGTATTGCTGGGCGGCGACCTCGCGGGTTTTACCGCTATTTGAAAAAATTACCCGGTTCAAGGCGCTTT  GTTCTTCTTCGTCATAACTTAATGTTTT A
```

26S 22M 2I 32M 1I 3M 2D 24M 1D

cigar2 26S22M2I32M1I3M2D24M1D8M1D5M1D10M5D

M doesn't mean the exact match, just the best score for the whole sequence
S: soft clipped (still in the sequence, the position counts start after S.
H: hard clipped (not in the sequence anymore, the position counts start from 1)

Informatics: Basic sequencing data analysis

- Raw data analysis (image processing and base calling)
- Fastq file stores all the sequencing information (including the quality scores)
- Quality control and read cleaning
- Alignment to the reference genome (SAM/BAM formats)
- *De novo* assembly (pairwise comparison without the reference)
- **Variant call (VCF formats)**
- Bed file, a tab-delimited text file that defines a track

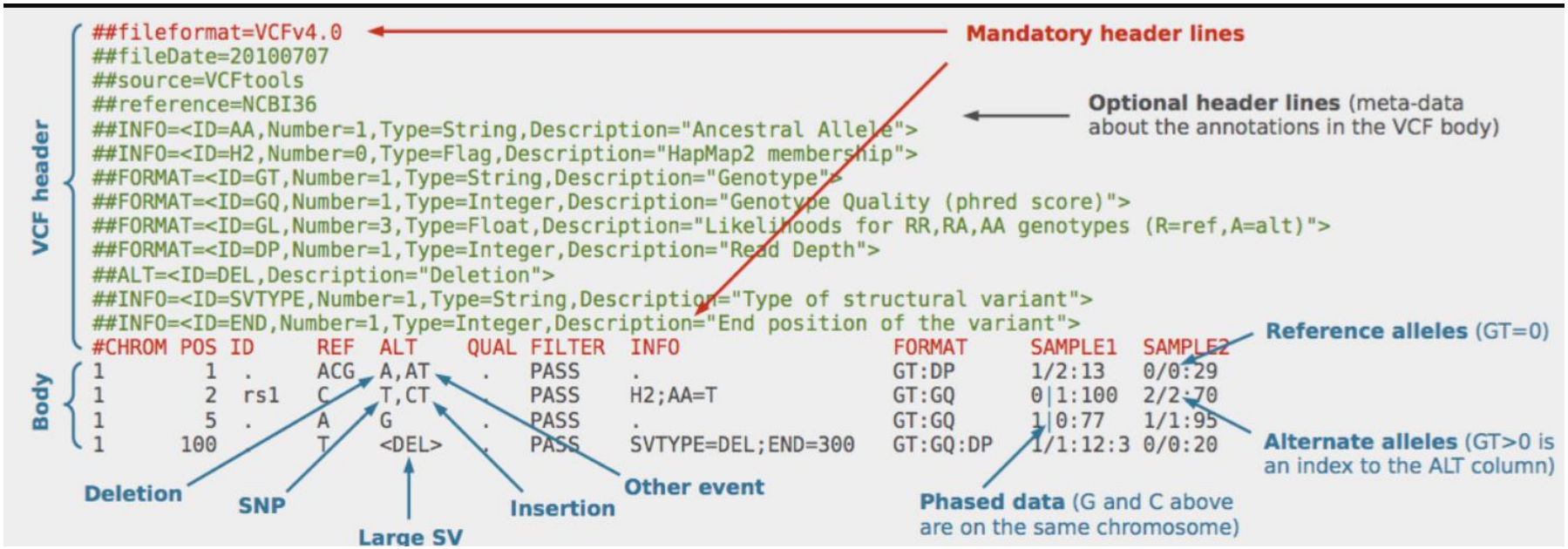
Alignment to the reference genome

Query
Reference

```
   18 tttcgacctgttattcgcatactattaaagaattctttggtagctctcaa    67
      .||||||||||||||| |||| |||||||||||||||||||||||||||||
 1348 attcgacctgttatt-gcat-ctattaaagaattctttggtagctctcaa  1395

   68 ttatcacaattcatggaccaagcaaacccattagctgagttaacgaataa   117
      ||||||||||||||||||||||||||||||||||||||||||||.||||
 1396 ttatcacaattcatggaccaagcaaacccattagctgagttaacgcataa  1445

  118 acgtcgtctatcagcattaggacctggtggtttaacacgtgaacgtgctc   167
      ||||||||||||||||||||||||||||||||||||||||||||||||||
 1446 acgtcgtctatcagcattaggacctggtggtttaacacgtgaacgtgctc  1495

  168 aaatggaagtacgtgacgttcactactctcactatggccgtatgtgtcca   217
      ||||||||||||||||||||||||||||||||||||||||||||||||||
 1496 aaatggaagtacgtgacgttcactactctcactatggccgtatgtgtcca  1545

  218 attgaaacacctgagggaccaaacattggattgattaacttattatcaag   267
      ||||||||||||||||||||||||||||||||||||.|||||||||||
 1546 attgaaacacctgagggaccaaacattggattgattaactcattatcaag  1595

  268 ttatgcacgtgtaaatgaattcggctttattgaaacaccatatcgtaaag   317
      ||||||||||||||||||||||||||||||||||||||||||||||||||
 1596 ttatgcacgtgtaaatgaattcggctttattgaaacaccatatcgtaaag  1645

  318 ttgatttagatacacatgctatcactga--aaa----------aac----   351
      |||||||||||||||||||||||||||  |||          |||
 1646 ttgatttagatacacatgctatcactgatcaaattgactatttaacagct  1695
```

Variant Call Format (VCF) file format

VCF is a text file format (most likely stored in a compressed manner). It contains meta-information lines, a header line, and then data lines each containing information about a position in the genome. The format also has the ability to contain genotype information on samples for each position.



https://vcftools.github.io/specs.html

The types of variants that can be stored in a VCF file are:

```
SNPs
Alignment  VCF representation
ACGT       POS REF ALT
AtGT       2   C   T

Insertions
Alignment  VCF representation
AC-GT       POS REF ALT
ACtGT       2   C   CT

Deletions
Alignment  VCF representation
ACGT       POS REF ALT
A--T       1   ACG A

Complex events
Alignment  VCF representation
ACGT       POS REF ALT
A-tT       1   ACG AT

Large structural variants
VCF representation
POS REF ALT   INFO
100 T   <DEL> SVTYPE=DEL;END=300
```