Bootcamp Part II

DNA sequence analysis-

Martina Armas BSc

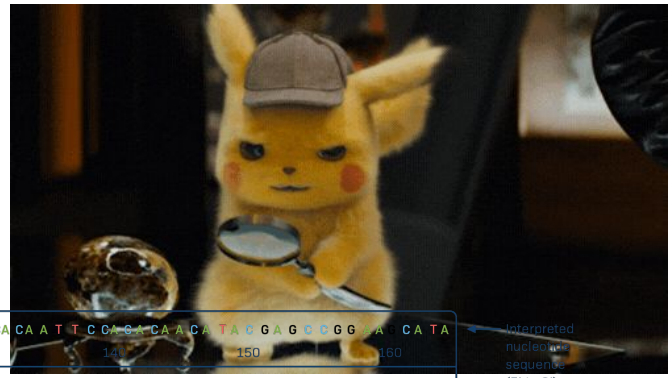
What is DNA sequencing?

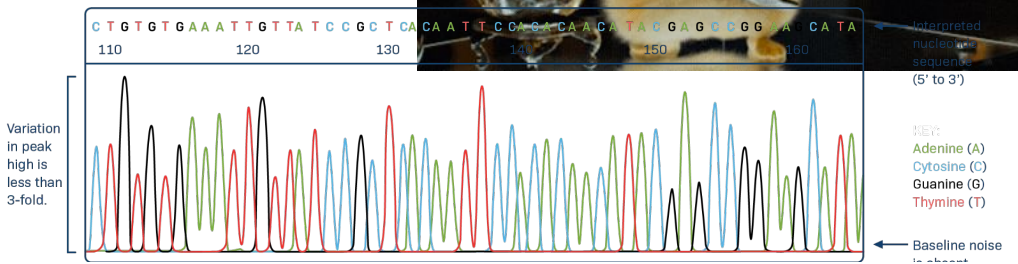
Goal:

Determine the order of nucleotides in a DNA sequence

Basically…

It's decoding a DNA strand, like a detective decodes the order of events from a murder.



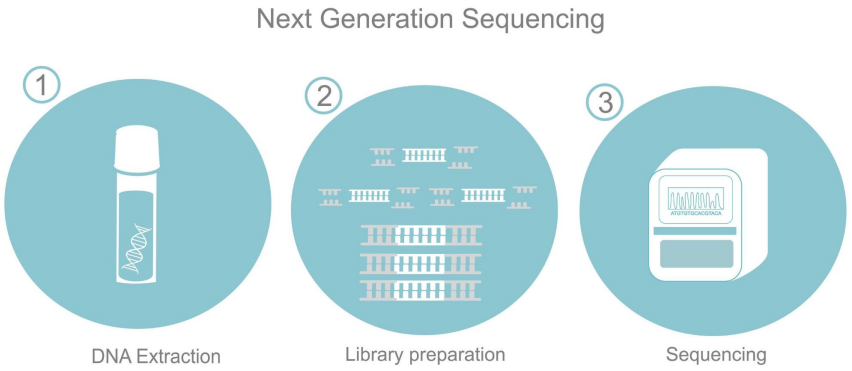


How do you decode the DNA sequence?

What do you need? - OUTLINE

- *DNA*
  - *Good concentration and good quality*
  - *DNA prepared to enter the sequencer*
- *Sequencer*
  - *Different models = use different techniques*
  - *2 Different types of DNA sequencing (long and short reads)*
- *Computer power* *(this gives provides the ability to analysis the massive amounts of data that the sequencer produces)*
  - *Types of archives*
  - *Annotation*

DNA - *for sequencing*



DNA

1. *Good extraction = Good quality/concentration* ✓
2. *Library preparation (using Molecular techniques)*
   - *This allows us to prepare the DNA to ENTER the sequencer*

…3 important steps?

a. *Repair and prepare DNA strand ends for ADAPTERS*
b. *Add BARCODES*
c. *Add ADAPTERS*

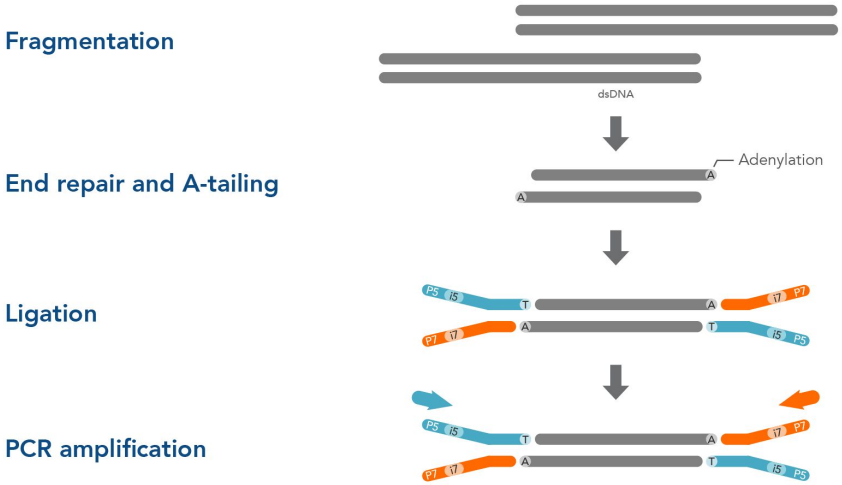
DNA - *for sequencing*

TERMINOLOGY - What does that means?

ADAPTERS: Adapters are short, synthetic DNA sequences ligated to the ends of DNA fragments during library preparation.

Functions:

1. *Compatibility with Sequencers*
2. *Anchoring DNA to the Flow Cell*
3. *Primer Binding Sites*
4. *Indexing (Barcode Integration)*



DNA - *for sequencing*

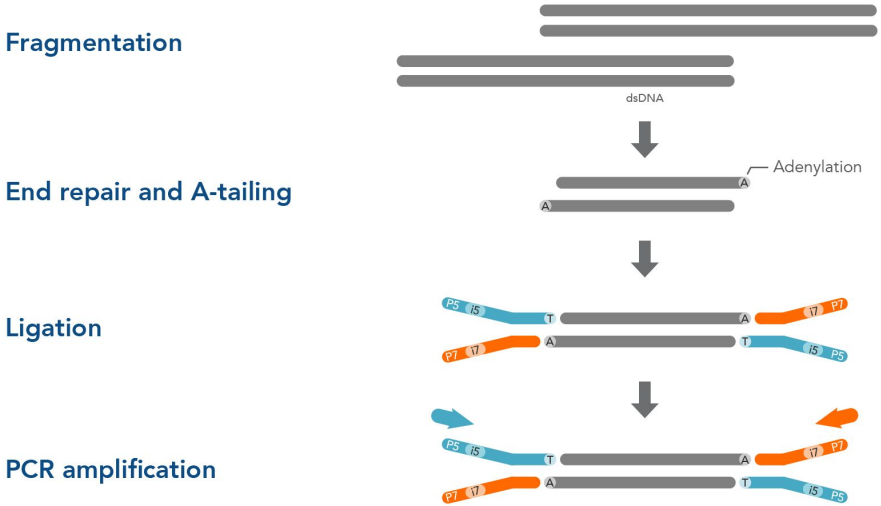
TERMINOLOGY - What does that means?

BARCODES: Barcodes are short, unique DNA sequences included within the adapter or ligated.

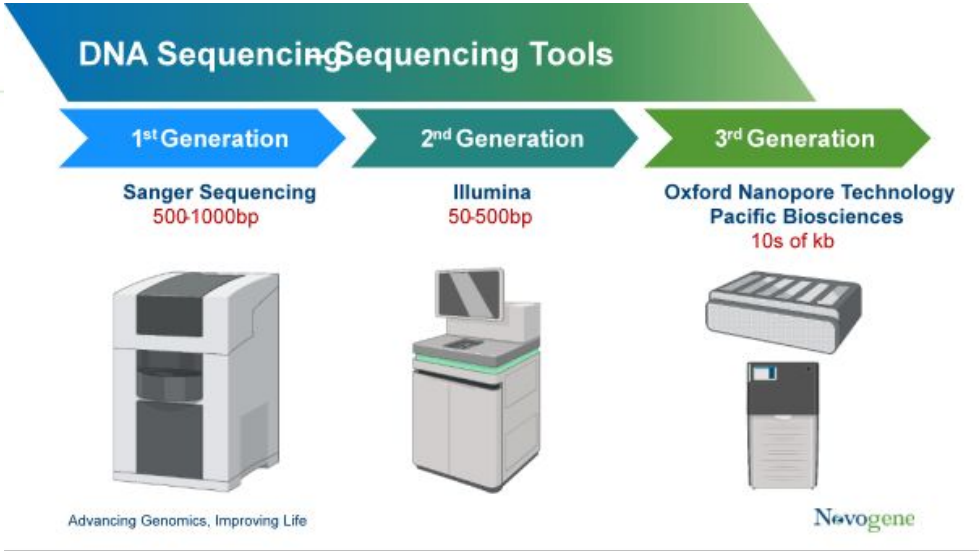
Functions:

Unique Identifiers

a. *They allow for multiplexing, which enables the simultaneous sequencing of multiple samples in a single sequencing run.*



SEQUENCER



1st generation = Sanger (Applied Biosystems 35000

Length: 500-1000 bp

Resolution: High

Cost: $0,50-$1,00 per kilobase

2nd generation = NGS Illumina

Length: 50-300 bp

Resolution: High coverage and precision

Cost: $0,01 per megabase

3rd generation = PacBio SMRT

Length: 10 000-15 000 bp to 100 000bp

Resolution: Moderated to High (according to the number of lectures)

Cost: $0,10 per kilobase

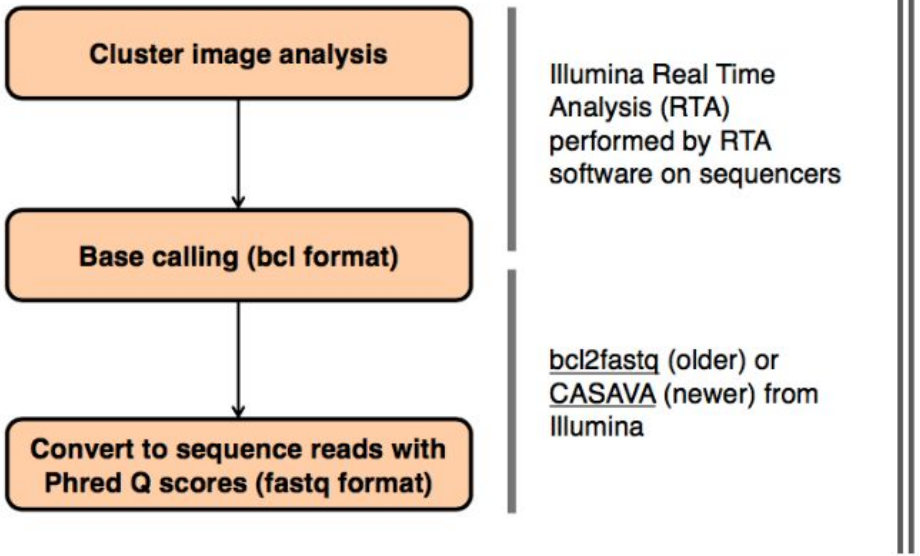
3rd generation = Oxford Nanopore

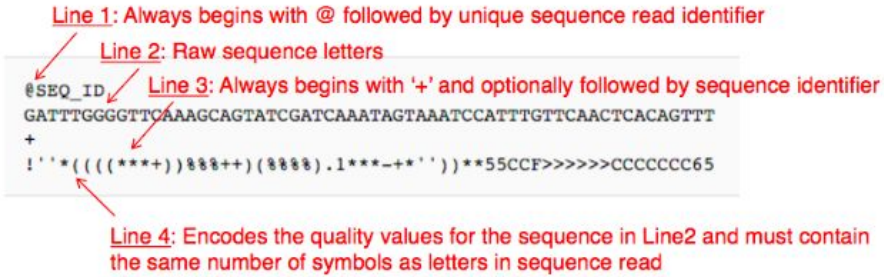
Length: 1000 bp to 100 000 000 bp

Resolution: Moderated

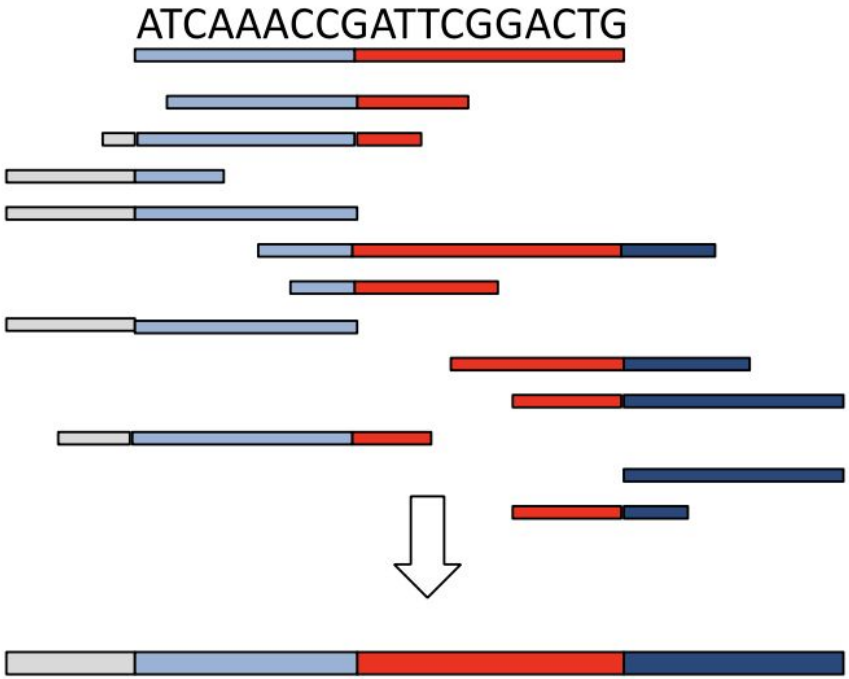
Cost: $0,01 per kilobase

Data Analysis





Data Analysis - *Ensemble "de novo"*

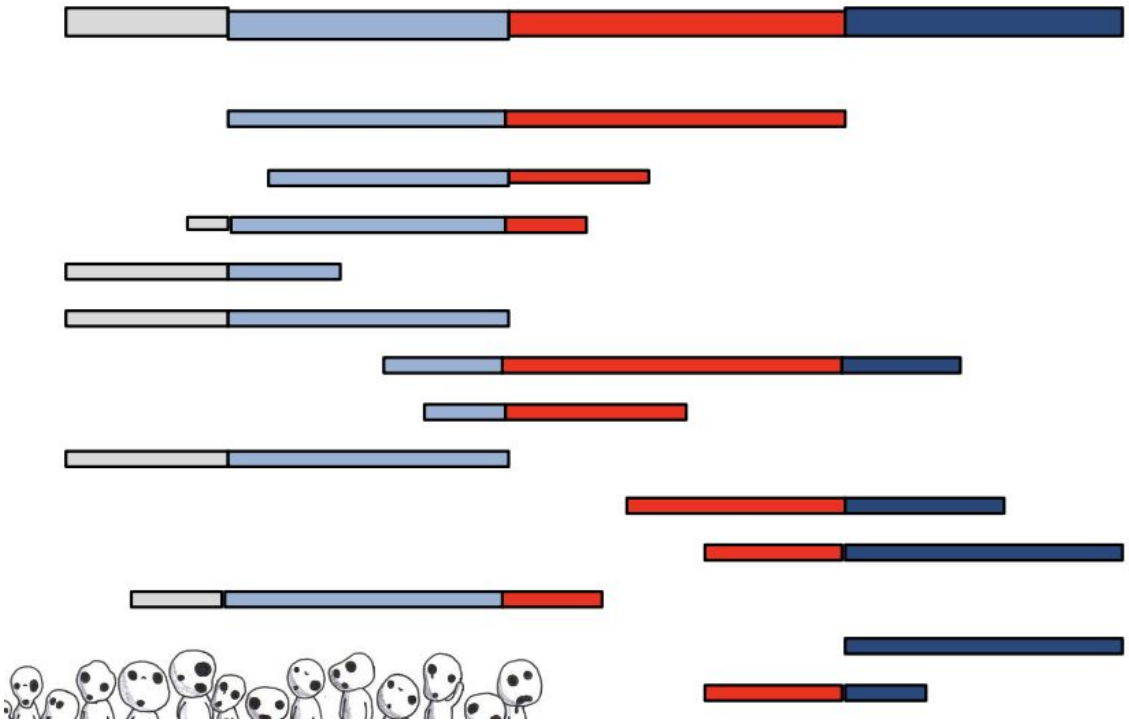


sequence reads

align reads to each other

assemble genome

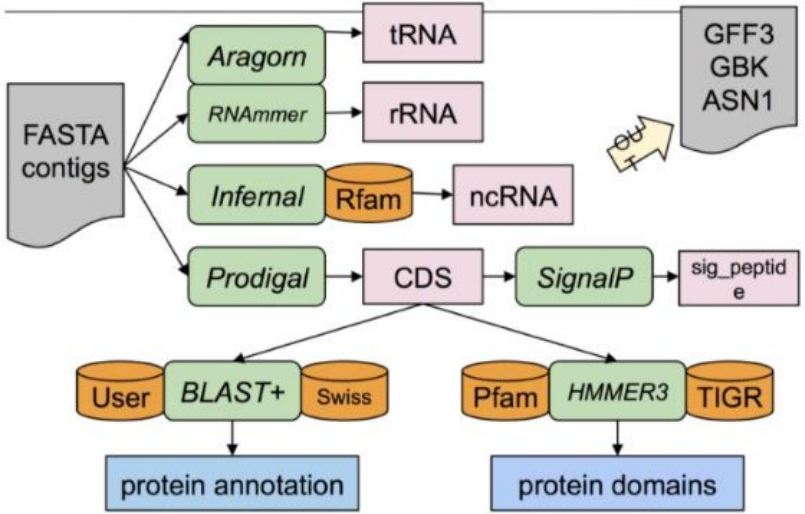
Data Analysis - *Ensemble with "reference" genome*



"reference" genome

align sequence reads to reference

Data Analysis - *Annotation*



Seemann T. *Prokka: rapid prokaryotic genome annotation*, **presentation 2013**
https://scilifelab.github.io/courses/annotation/2017/slides/prokkaLS_08_05_2017_v2.pdf

Programs used to predict genome features

Tool (reference)	Features predicted
Prodigal (Hyatt 2010)	Coding sequence (CDS)
RNAmmer (Lagesen *et al.*, 2007)	Ribosomal RNA genes (rRNA)
Aragorn (Laslett and Canback, 2004)	Transfer RNA genes
SignalP (Petersen *et al.*, 2011)	Signal leader peptides
Infernal (Kolbe and Eddy, 2011)	Non-coding RNA

Seemann, T. *Prokka: rapid prokaryotic genome annotation*. 2014

EXTRA: Free Resources and Platforms to start in BIOINFORMATICS

Software Carpentry Lessons
- https://software-carpentry.org/lessons/

Data Carpentry Lessons
- http://www.datacarpentry.org/lessons/

R Studio
- https://www.rstudio.com/online-learning/

Github
- https://guides.github.com

Libros gratis
- http://www.dsf.unica.it/~fiore/LearningPython.pdf
- http://opencarts.org/sachlaptrinh/pdf/28232.pdf
- http://r4ds.had.co.nz
- Bioinformatics Data Skills PDF Download:
  - Google search "bioinformatics data skills pdf"