H3ABioNet

معهد باستور تونس
Institut Pasteur de Tunis

H3ABioNet

Pan African Bioinformatics Network for H3Africa

Introduction to Bioinformatics Online Course: IBT

Genomics

Sequencing technologies and NGS Overview

Learning Objectives

Session 1:
Sequencing technologies and NGS Overview

- **Part 1:** Introduction to DNA Sequencing
- **Part 2:** DNA Sequencing in the NGS era
- **Part 3:** Overview of NGS Technologies
- **Part 4:** DNA-Seq Protocol : Overview
- **Part 5:** DNA-Seq Analysis Pipeline and File Formats

Learning Outcomes

Session 1:
Sequencing technologies and NGS Overview

- Understand basics of NGS technologies
- Understand different NGS file formats
- Navigate through database repositories to retrieve NGS datasets

Part 1

Introduction to DNA Sequencing

What is DNA Sequencing ?

- DNA Sequencing is the process of reading the nucleotides present in DNA : **determining the precise order of nucleotides within a DNA molecule**.

- DNA-Seq generally refers today to any NGS method or technology that is used to determine the order of the four bases (A, T, C, G) in a strand of DNA.

What is DNA Sequencing ?

- In fact, there are 2 main types of DNA sequencing technologies that are used today: **Sanger sequencing** and **Next-Generation Sequencing (NGS)**.

- Each of these technologies has utility in today's genetic analysis environment.

Main DNA sequencing technologies

1953
(Watson & Crick) DNA Structure as a double helix

1972-1976
(Min Jou et al., 1972; Fiers et al., 1976) sequence of the 1st complete gene and the complete genome of Bacteriophage MS2 (viral RNA 3,569 nt)

1977
Maxam & Gilbert
+ Sanger sequencing

1977
(Sanger et al., Feb 1977)
The first full DNA genome to be sequenced : bacteriophage ϕX174

1987
ABI markets ABI370, the 1st fully automated sequencing machine

1995
First published use of WGS sequencing

1995
(C.Venter) 1st complete genome of a free-living organism, the bacterium *H. influenzae* (circular chr >1,8Mb)

1996
First NGS technology : Pyrosequencing

2001
Draft sequence of the human genome (shotgun sequencing methods). Public & Private (Lander et al., Feb 2001 Nature; Venter et al., Feb 2001 Science)

The Human Genome Project: the objectives

- The Human Genome Project (**HGP**)= a 13-years (**1990-April 14, 2003**) international effort to sequence of the 3 billion "letters" of human DNA.
- $300 million project, led by the U.S. DoE and the NIH.
- International Human Genome Sequencing Consortium (**IHGSC**)= group of publicly funded researchers
- At any given time, ≈ 200 labs in the United States supported these efforts + > 18 different countries from across the globe had contributed to the HGP.

The Human Genome Project: the objectives

- 2 groups competing for sequencing:
  - Public
  - Private (Celera Genomics)

- Opposing philosophies :
  - **HGP** Bermuda Agreement (1996)

  → all information from the project would be made freely available to all within 24h.
  - **Private**

  → access restricted to paying customers !

- In February 2001, drafts of the human genome sequence were published simultaneously by both public-private groups in separate articles (Lander et al (IHGSC)., Feb 2001 Nature; Venter et al., Feb 2001 Science).

Chial, 2008
http://www.genome.gov/sequencingcosts/
http://www.yourgenome.org/
www.sanger.ac.uk

The Human Genome Project: the method (WGS)

- **Hierarchical genome shotgun sequencing**

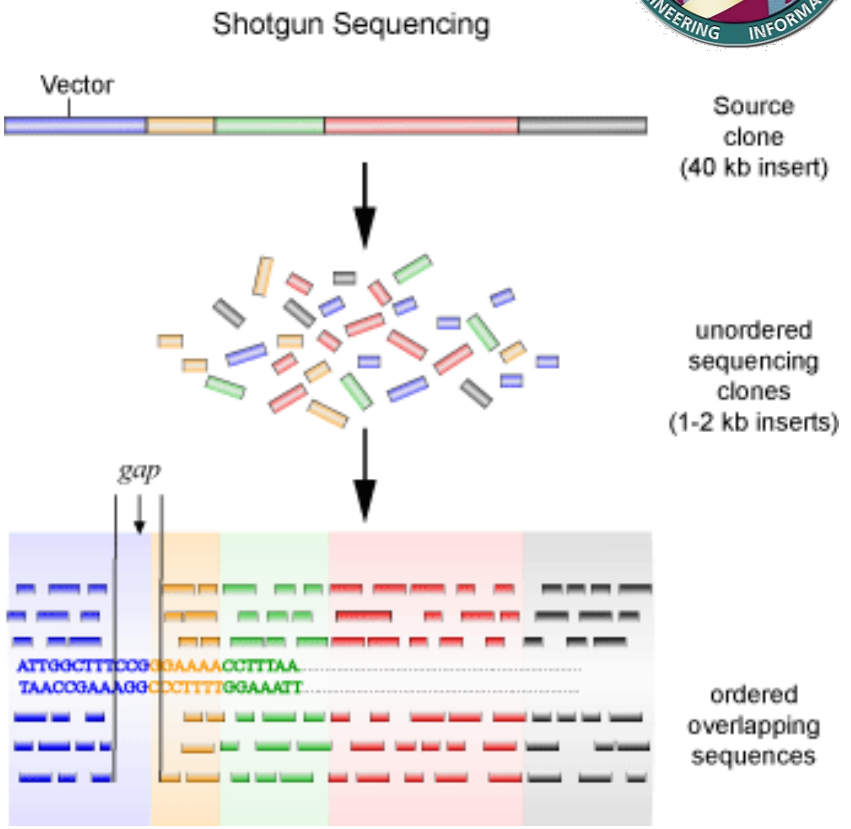
- Shotgun phase

→ Genome fragmented into larger segments
→ cloning into vectors
→ clones sequencing
→ shotgun sequences assembly
→ relied on the physical map of the human genome established earlier.

- Finishing phase : filling in gaps and resolving DNA sequences in ambiguous areas

- **Whole-genome shotgun sequencing** (Celera genomics)

→ Genome sheared randomly into small fragments (appropriately sized for sequencing)
→ Reassembly.



The IHGSC and Celera used the same general method of termination chain for the DNA sequencing (Hood & Galas, 2003).

Sequencing Quality

- Sequencing quality depend upon the average number of times each base in the genome is 'read' during the sequencing process.

- For the Human Genome Project (HGP) :
  - **'draft sequence'** (covering ~90% of the genome at ~99.9% accuracy)
  - **'finished sequence'** (covering >95% of the genome at ~99.99% accuracy).

  Producing truly high-quality 'finished' sequence by this definition is very expensive and labor-intensive.

- **Several releases of the human genome sequences**

http://www.genome.gov/sequencingcosts/

Finished Genome vs Draft Genome

- Variable degrees of completion of published genomes

• Draft Sequencing

- high-throughput or shotgun phase (whole genome or clone-based approach)
- Assembly using specific algorithms (whole-genome or single-clone assembly)

→ lower accuracy than finished sequence; some segments are missing or in the wrong order or orientation.

• Finishing

- Accuracy in bases identification + Quality Check + few if any gaps.
- Contiguous segments of sequence are ordered and linked to one another
- No ambiguities or discrepancies about segments order and orientation

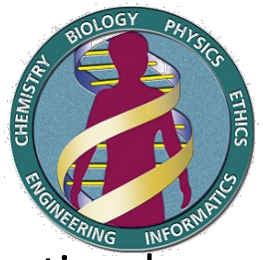
• Complete Genome

A Genome represented by a single contiguous sequence with no ambiguities

- → The sequences available are ***finished to a certain high quality***.

(Mardis et al., 2002)

The Human Genome Project: the heritage



- The HGP project required that all human genome sequence information be **freely and publicly available**. The existing **DNA sequences** have been stored in databases available to anyone willing to exploit and analyze them.

- Dedicated databases house various data for model organisms such as sequences of known and hypothetical genes and proteins (**GenBank**, **NCBI**). Other databases (**Ensembl** http://www.ensembl.org) present additional data and annotation as well as powerful tools for visualizing and searching it.

- Community efforts for non-model organisms like Eukaryotic Pathogens : **EuPathDB** (http://eupathdb.org/eupathdb/).

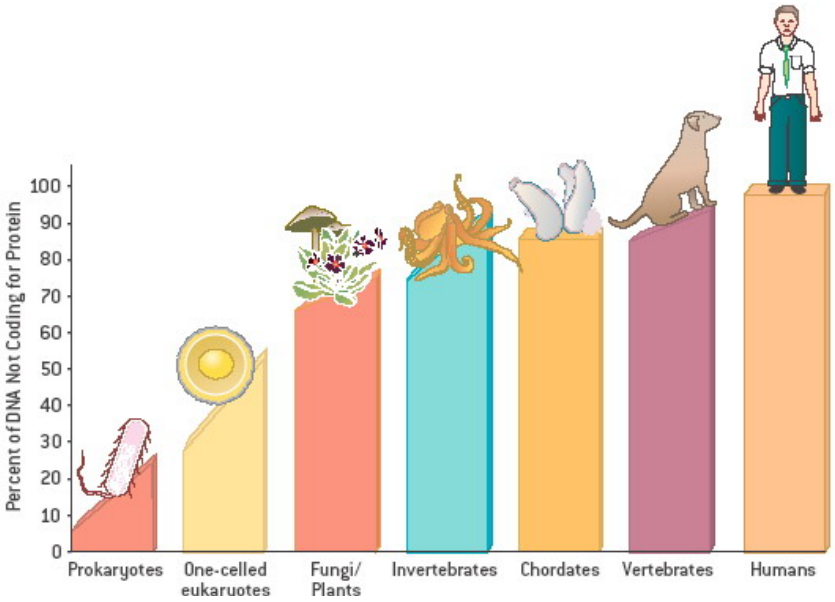
- **Computer programs** have been developed to analyze and interpret the data.

The Human Genome Project: the heritage

- The human genome contains only about 20,000 protein-coding genes, similar in number and with largely orthologous functions as those in nematodes that have only 1,000 somatic cells.

- The **extent of non-protein-coding DNA increases with increasing complexity**, reaching > **98% in humans**.

→ **Encode**
→ **Gencode**



(Mattick, 2011)

The Human Genome Project: the heritage

- **Encode** (https://www.encodeproject.org)
  The Encyclopedia of DNA Elements (ENCODE) Project aims to provide "a list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active. "
  "**the generation of such a catalogue is crucial for understanding genome function**."

- **Gencode** (http://www.gencodegenes.org)
  The human genome has been the focus of intensive manual annotation:
  The GENCODE Consortium aims to identify all gene features in the human and mouse genomes using a combination of computational analysis, manual annotation, and experimental validation.

(Djebali et al., 2012)
(Harrow et al., 2012)
(Birney et al., 2007)

The Human Genome Project: the heritage

- Genetic differences in individual bases (SNPs) of a genome are by far the most common type of genetic variation.
- Goal: develop a **Haplotype Map** of the Human Genome
  = identification and cataloging of most of the millions of SNPs estimated to occur commonly in the human genome.
- **Described variants are, their location, their distribution among people within populations and among populations** in different parts of the world. → designed to **provide information** to **link genetic variants to the risk for specific diseases**
- **1000 Genomes Project**: has become more complete and reliable as many novel variants have been discovered !!

The Human Genome Project: the heritage

- 1000 Genomes (www.1000genomes.org/)

  - identify most genetic variants with frequencies of at least 1%.
  - freely accessible ressource of human genetic variation.
  - final data set = data for 2,504 individuals from 26 populations. (low coverage sequencing and exome sequence data for all)
  - International Genome Sample Resource (IGSR) for ongoing usability of data generated by the 1000 Genomes Project.

- UK10K (www.uk10k.org/)

  - identification of rare genetic variants through the study of the DNA of 4,000 individuals and their comparison to the protein-coding areas of 6,000 people with documented diseases.
  - link between genetic variants and rare diseases.

The Human Genome Project: the heritage

- Development of novel technologies to help increase the depth of sequencing: Next-Generation Sequencing **(NGS) technologies**

- Since their development, NGS technologies have gained increasing attention with a considerable potential application in both diagnostic and public health microbiology.

- Revolutionized the sequencing process: **from Sanger to HT sequencing**

(Salipante *et al.*, 2013)

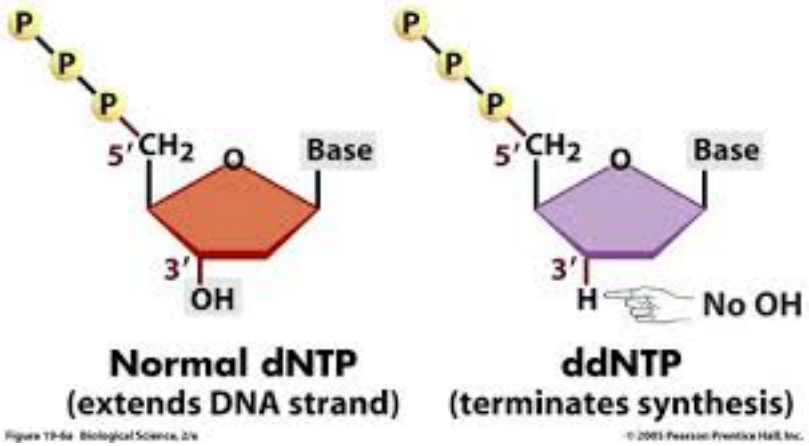
Part 2

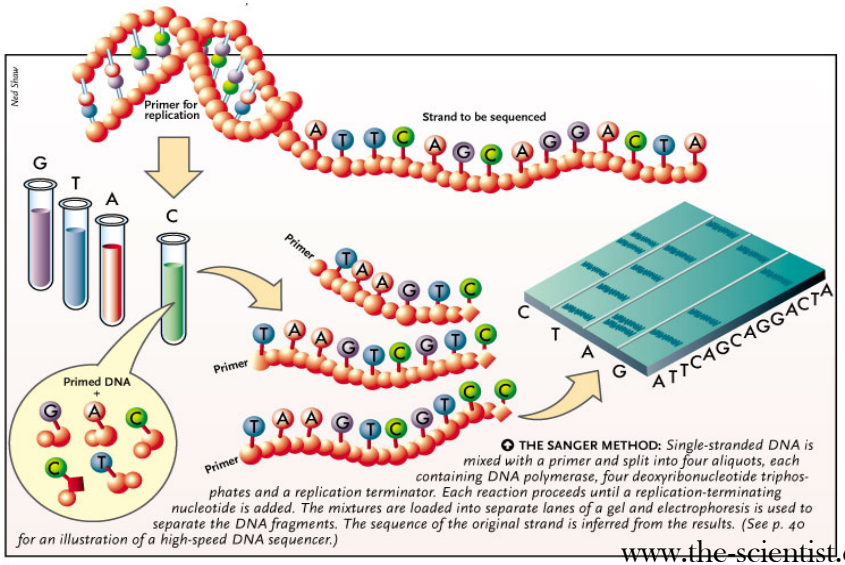
DNA Sequencing in the NGS era

DNA Sequencing method: from Sanger to NGS

- **Sanger sequencing** is the method developed by Frederick Sanger in 1977. This method involves copying single-stranded DNA with chemically altered bases called dideoxynucleotides (**ddNTPs**).

- ddNTPs when incorporated at the 3' end of the growing chain, terminate the chain selectively at A, C, G, or T. The terminated chains are then resolved by capillary electrophoresis.





DNA Sequencing method: from Sanger to NGS

- Applied Biosystems (Life Technologies), manufactured the automated capillary sequencers utilized by both Celera Genomics and The Human Genome Project.
- While capillary sequencing was the first approach to successfully sequence a full human genome, it was still too expensive and took too long for commercial purposes !!!
- Because of this, sequencing using Sanger technology has been displaced by technologies like pyrosequencing, or SMRT sequencing…

DNA-Seq

- DNA-Seq is nowadays used as an effective sequencing strategy after the advent of rapid DNA sequencing methods that has greatly accelerated biological and medical research and discovery : *de novo*…

- **DNA-Seq may be used to determine the sequence of individual genes, larger genetic regions, full chromosomes, or entire genomes.**

- 'DNA-Seq' and other related 'seq' technologies allow to cover genome complexity : genomic DNA-Seq, Methyl-Seq, ChIP-Seq, exome sequencing…

NGS

- Next-generation sequencing (NGS), or high-throughput (HT) sequencing = catch-all term describing different modern sequencing technologies used by different platforms:

  - Illumina (Solexa) sequencing
  - Roche 454 sequencing
  - Ion torrent: Proton / PGM sequencing
  - …

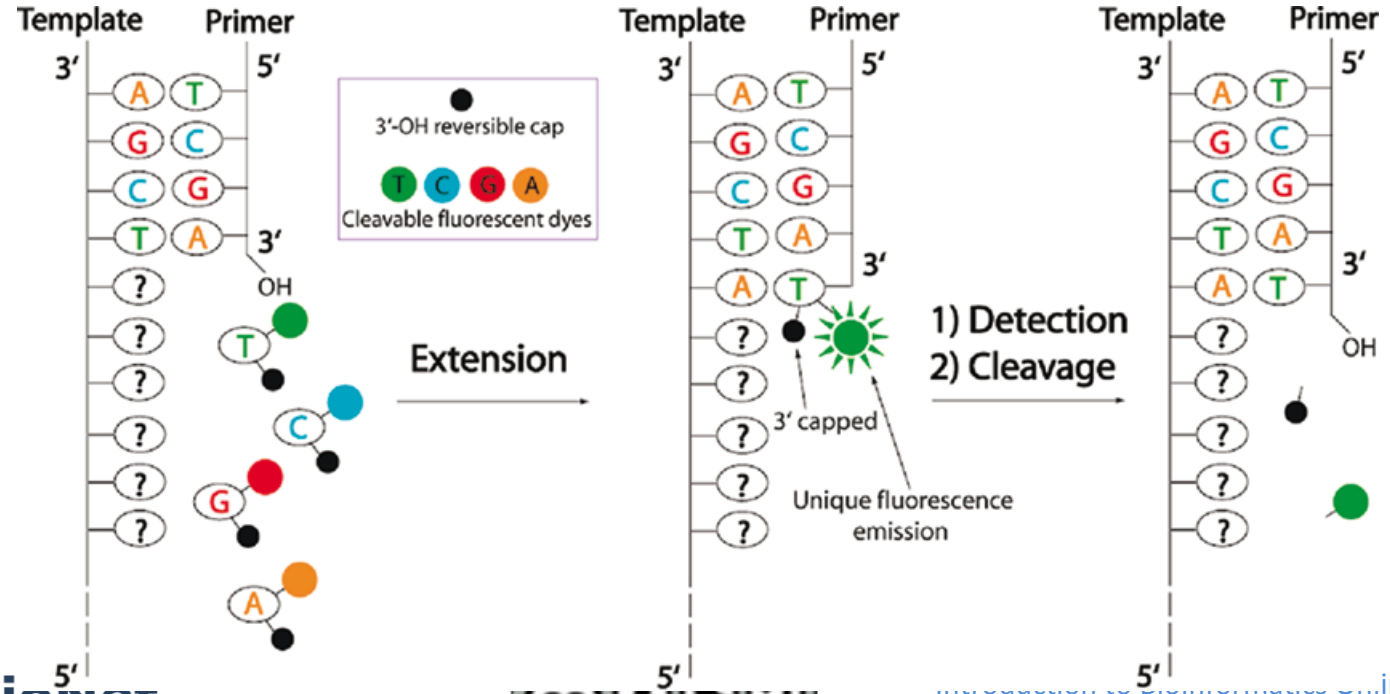
- DNA sequence faster and cheaper than the **Sanger sequencing** = revolution for genomics and molecular biology.

NGS:
Sequencing Technologies and Platforms

Technology	Company	Support	Chemistry
Massively Parallel Sequencing			
Solexa	Illumina	**Bridge PCR** on flowcell	**Seq-By-Synthesis**
454	Roche Applied Science	**emPCR** on beads	**Pyrosequencing**
SOLiD	AB / Life Technologies	emPCR on beads	Seq-By-Ligation
Ion Torrent	Life Technologies	emPCR on beads	Proton detection
Single Molecule Sequencing			
PacBio SMRT	Pacific Biosciences	Pol performance	Real-time-Seq
Nanopore	Oxford Nanopore Tech/McNally	Translocation	NA

NGS:
Principles of DNA Sequencing: SBS

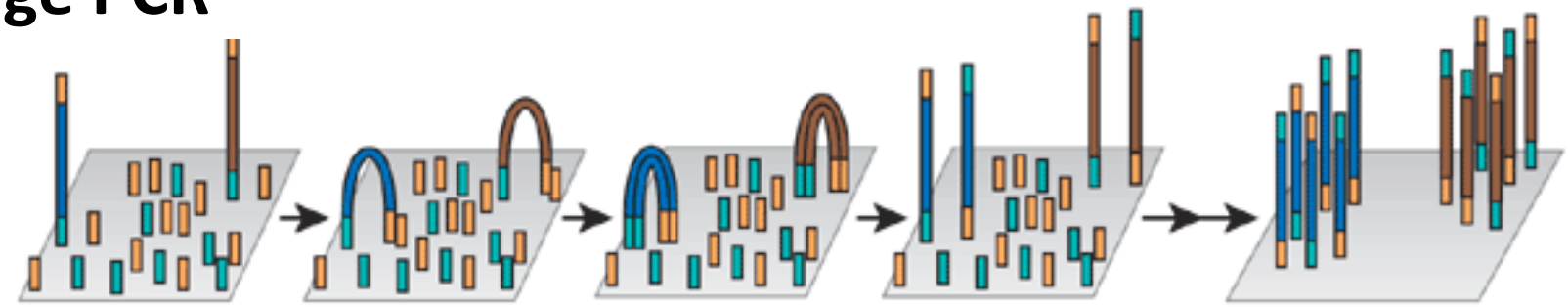
- **Tracking the addition of labeled nucleotides as the DNA chain is copied**
- The **DNA template** is **immobilized.**
- Solutions of A, nucleotides C , G and T sequentially added and removed.
- **Light** is generated when a nucleotide complements the first unpaired base.
- Chemiluminescent signal detected to determine the sequence.



NGS:

In brief : emPCR (454/SOLiD/Ion Torrent) vs Bridge PCR (Illumina)

Bridge PCR



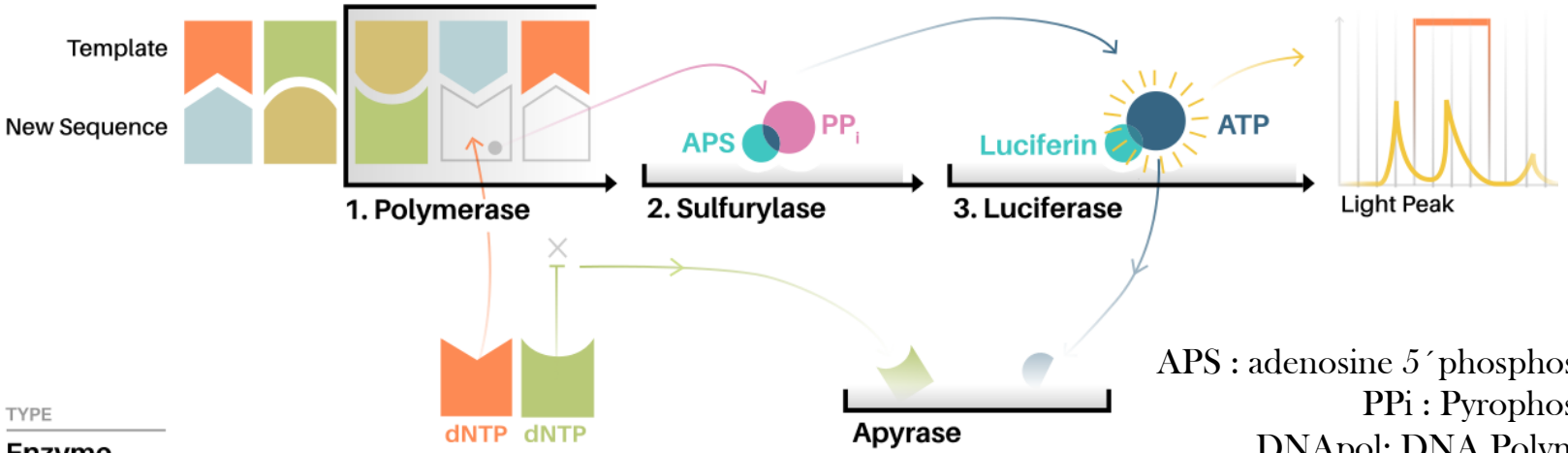
- **The adaptor-flanked shotgun library is PCR amplified on a flow cell**
- **both primers coat the surface of a solid substrate**
- **Amplification products** from any given member of the library **remain locally fixed near the point of origin = cluster**
- The PCR produces **clonal clusters contains copies** of a single DNA.

(Shendure & Ji, 2008)

NGS:
Principles of DNA Sequencing: Pyrosequencing

- **Incorporation of dNTPs by DNApol releases pyrophosphate (PPi).**
- ATP sulfurylase converts PPi to ATP in the presence of APS.
- ATP = substrate for the luciferase-mediated conversion of luciferin to oxyluciferin
- This conversion generates light in amounts proportional to the amount of ATP detected by a camera.

Unincorporated nucleotides and ATP are degraded by the apyrase, and the reaction can restart with another nucleotide



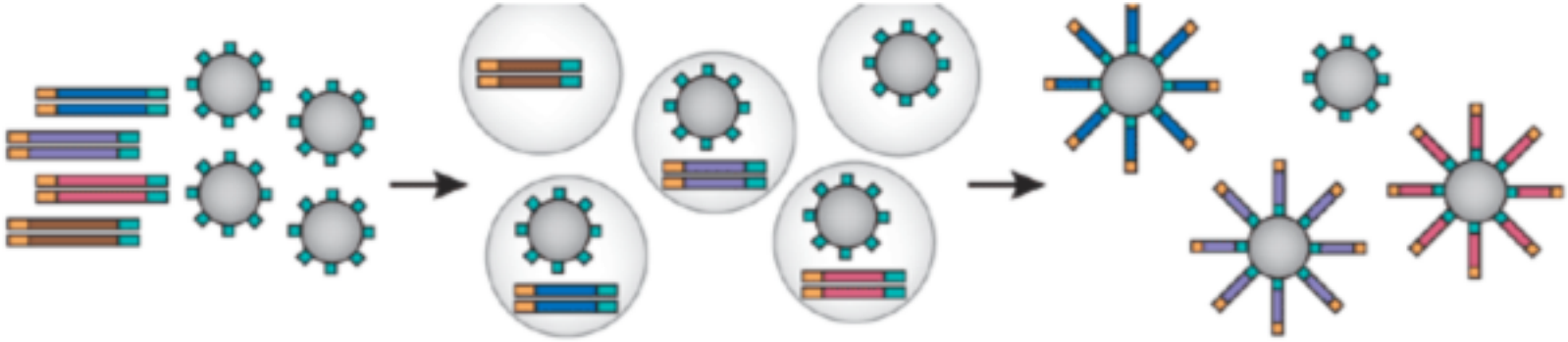
APS : adenosine 5´phosphosulfate
PPi : Pyrophosphate
DNApol: DNA Polymerase

(https://en.wikipedia.org)

NGS:

In brief : emPCR (454/SOLiD/Ion Torrent) vs Bridge PCR (Illumina)

emPCR



- **The adaptor-flanked shotgun library is PCR amplified** in the context of a **water-in-oil emulsion**.
- **PCR primer is 5'-attached** on micron-scale beads.
- 1 bead-containing compartments = 0 or 1 template DNA.
- **PCR amplicons are captured to the surface of the bead**.
- **1 clonally amplified bead** = PCR products corresponding to **amplification of a single molecule** from the library.

(Shendure & Ji, 2008)

Part 3

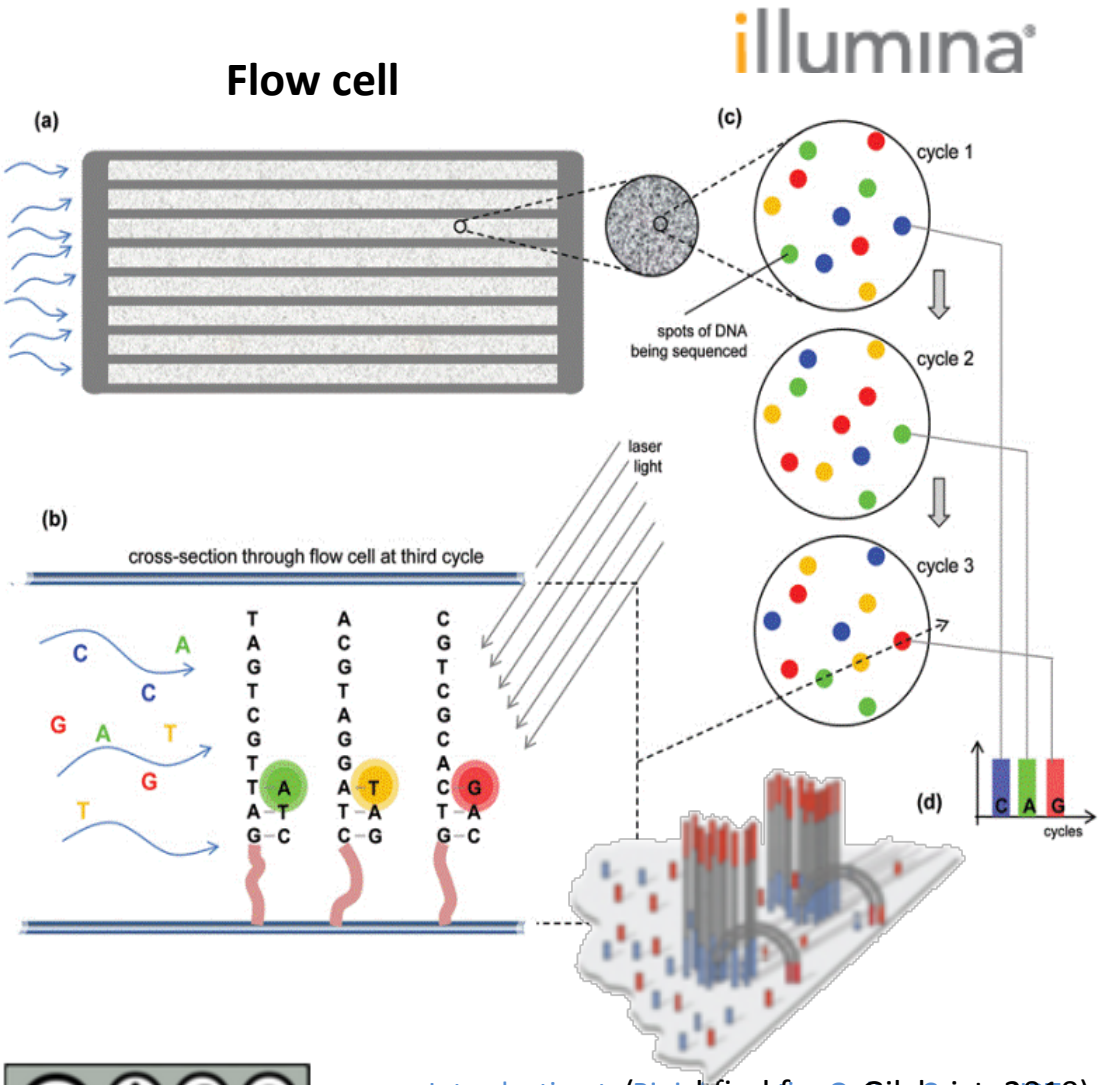
Overview of NGS Technologies

NGS:
Sequencing Technologies and Platforms

Technology	Company	Support	Chemistry
Massively Parallel Sequencing			
Solexa	Illumina	Bridge PCR on flowcell	Seq-By-Synthesis
454	Roche Applied Science	emPCR on beads	Pyrosequencing
SOLiD	AB / Life Technologies	emPCR on beads	Seq-By-Ligation
Ion Torrent	Life Technologies	emPCR on beads	Proton detection
Single Molecule Sequencing			
PacBio SMRT	Pacific Biosciences	Pol performance	Real-time-Seq
Nanopore	Oxford Nanopore Tech/McNally	Translocation	NA

Solexa

- The input sample must be cleaved into short sections.
- Fragments are ligated to adaptors and annealed to the slide using the adaptors.
- Fragments are separated into single strands to be sequenced.
- Nucleotides are modified so that each emits a different coloured light when excited by a laser.
  + they have a terminator, so that only one base is added at a time.
- PCR, process repeated in cycles, images analyzed.



(modified from Gilchrist, 2010)

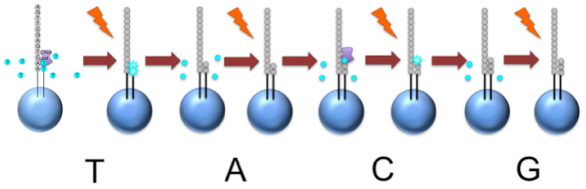
Solexa
Intra-Platform Comparison

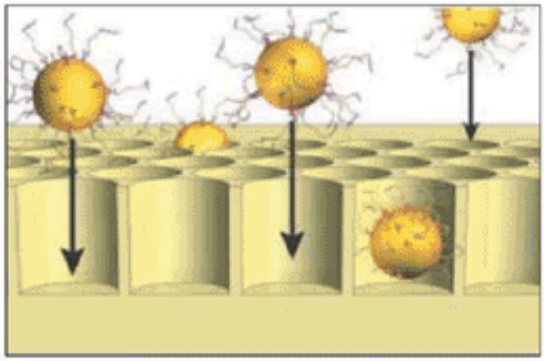
	 MiSeq **Focused power.** Speed and simplicity for targeted and small genome sequencing.	 NextSeq 500 **Flexible power.** Speed and simplicity for everyday genomics.		 HiSeq 2500 **Production power.** Power and efficiency for large-scale genomics.		 HiSeq X* **Population power.** $1,000 human genome and extreme throughput for population-scale sequencing.
Key applications	Small genome, amplicon, and targeted gene panel sequencing.	Everyday genome, exome, transcriptome sequencing, and more.		Production-scale genome, exome, transcriptome sequencing, and more.		Population-scale human whole-genome sequencing.
Run mode	N/A	Mid-Output	High-Output	Rapid Run	High-Output	N/A
Flow cells processed per run	1	1	1	1 or 2	1 or 2	1 or 2
Output range	0.3-15 Gb	20-39 Gb	30-120 Gb	10-180 Gb	50-1000 Gb	1.6-1.8 Tb
Run time	5-65 hours	15-26 hours	12-30 hours	7-40 hours	< 1 day - 6 days	< 3 days
Reads per flow cell†	25 Million‡	130 Million	400 Million	300 Million	2 Billion	3 Billion
Maximum read length	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp	2 × 125 bp	2 × 150 bp

http://www.illumina.com

454 Technology

- **As in Illumina, the DNA is fragmented**.
- **Adaptors** added, end annealed to beads.
  1 DNA fragment = 1 bead.
- Fragments amplified by **PCR using adaptor-specific primers**.
- The sequence can then be determined computationally.
- **Longer reads than Illumina**, different lengths.





(http://www.ebi.ac.uk)

454 Technology
Intra-Platform Comparison

Roche

GS Junior System

brings the power of 454 Sequencing Systems directly to the laboratory benchtop

Read lengths:
GS Junior (up to 400bp)
GS Junior + (700-800bp)

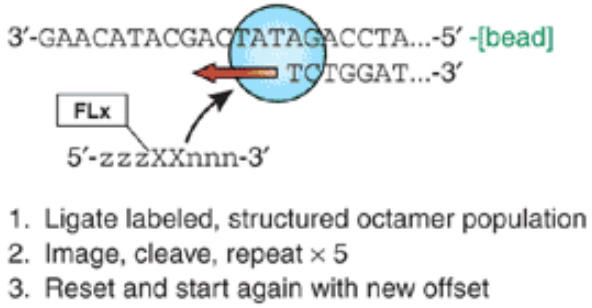
GS FLX+ System

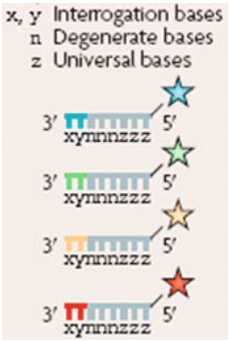
Features combination of long reads, accuracy and high-throughput, making the system well suited for larger genomic projects.

Read lengths:
GS FLX Titanium XL+ (up to 1,000bp)
GS FLX Titanium XLR70 (up to 600bp)

SOLID

- Sequencing is performed with a **ligase**, rather than a polymerase.
- Each sequencing cycle introduces a **partially degenerated population of fluorescently labeled octamers.** The population is structured such that **the label correlates with the identity of the central 2 bp in the octamer.**
- **After ligation and imaging** in four channels, the **labeled portion of the octamer (that is, 'zzz') is cleaved** leaving a free end for another cycle of ligation.



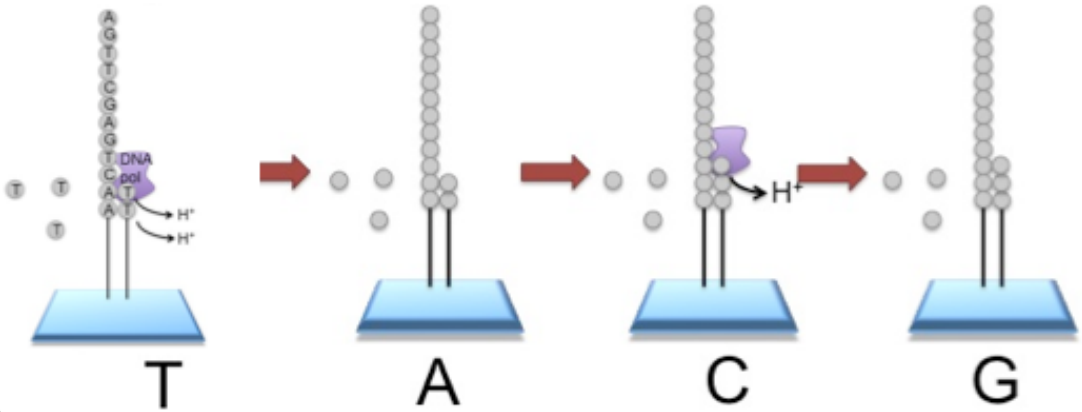


*In the **SOLiD sequencing** method, unlike other NGS technologies (which base detection of DNA fragments is performed through polymerase reaction) sequencing is achieved by **Sequencing-By-Ligation**.*

Adapted from (Shendure & Ji, 2008) and (http://www.springer.com/978-1-4614-7725-9)

Ion Torrent

- **As in other kinds of NGS, the input DNA is fragmented.**
- **Adaptors** are added and **one molecule is placed onto a bead**.
- Amplification on the bead by **emulsion PCR**. Each bead is placed into 1 well of a slide.
- **The pH is detected is each of the wells, as each H+ ion released will decrease the pH**. The changes in pH allow us to determine if that base, and how many thereof, was added to the sequence read.
- The dNTPs are washed away, and the process is **repeated in cycles.**



Adapted from (http://www.ebi.ac.uk) and (http://www.springer.com/978-1-4614-7725-9)

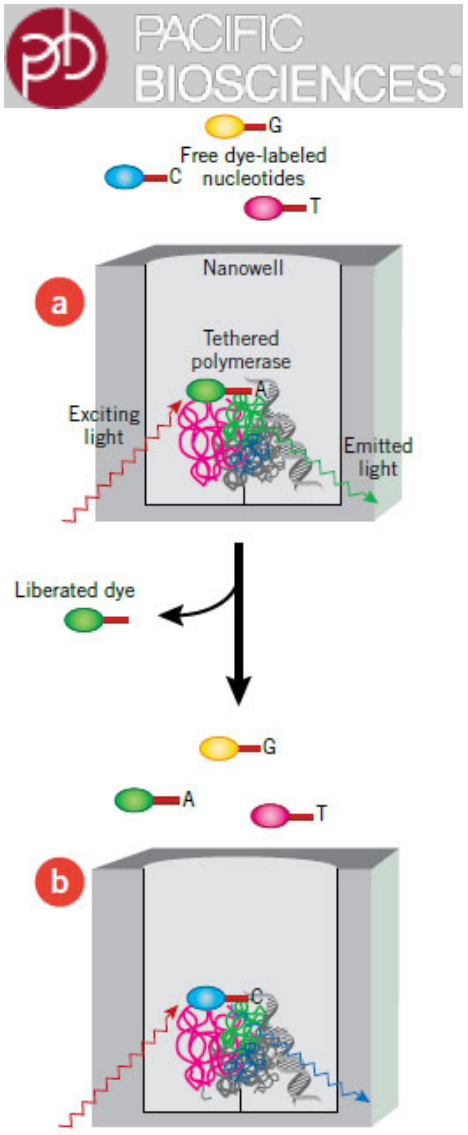
PacBio

- **Single Molecule Real Time (SMRT) DNA Sequencing technology**

- **a →** DNA polymerase molecule is tethered to the bottom of a nanowell → ZMW design ensures only one nucleotide-linked dye can be directly excited at a time.

- **b →** Each incorporated phospholinked nucleotide will reside on the enzyme's active site for a few milliseconds, which is enough time for a fluorescent signal to be recorded.

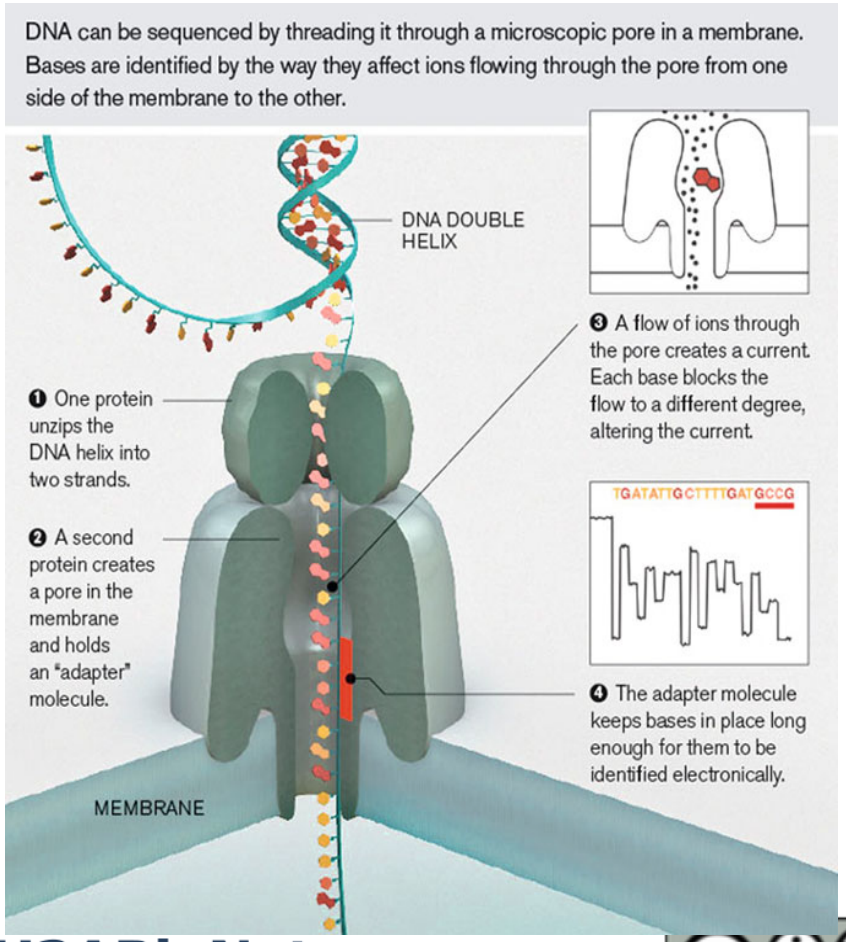
NB: In other systems, the fluorescent label is attached to the base in nucleotides. In SMRT technology, the fluorescent label is attached to the phosphate chain → The released labled pentaphosphates will diffuse quickly.



(Osherovich et al., 2010)

Nanopore

Single Molecule Real Time (SMRT) DNA Sequencing technology: How it works



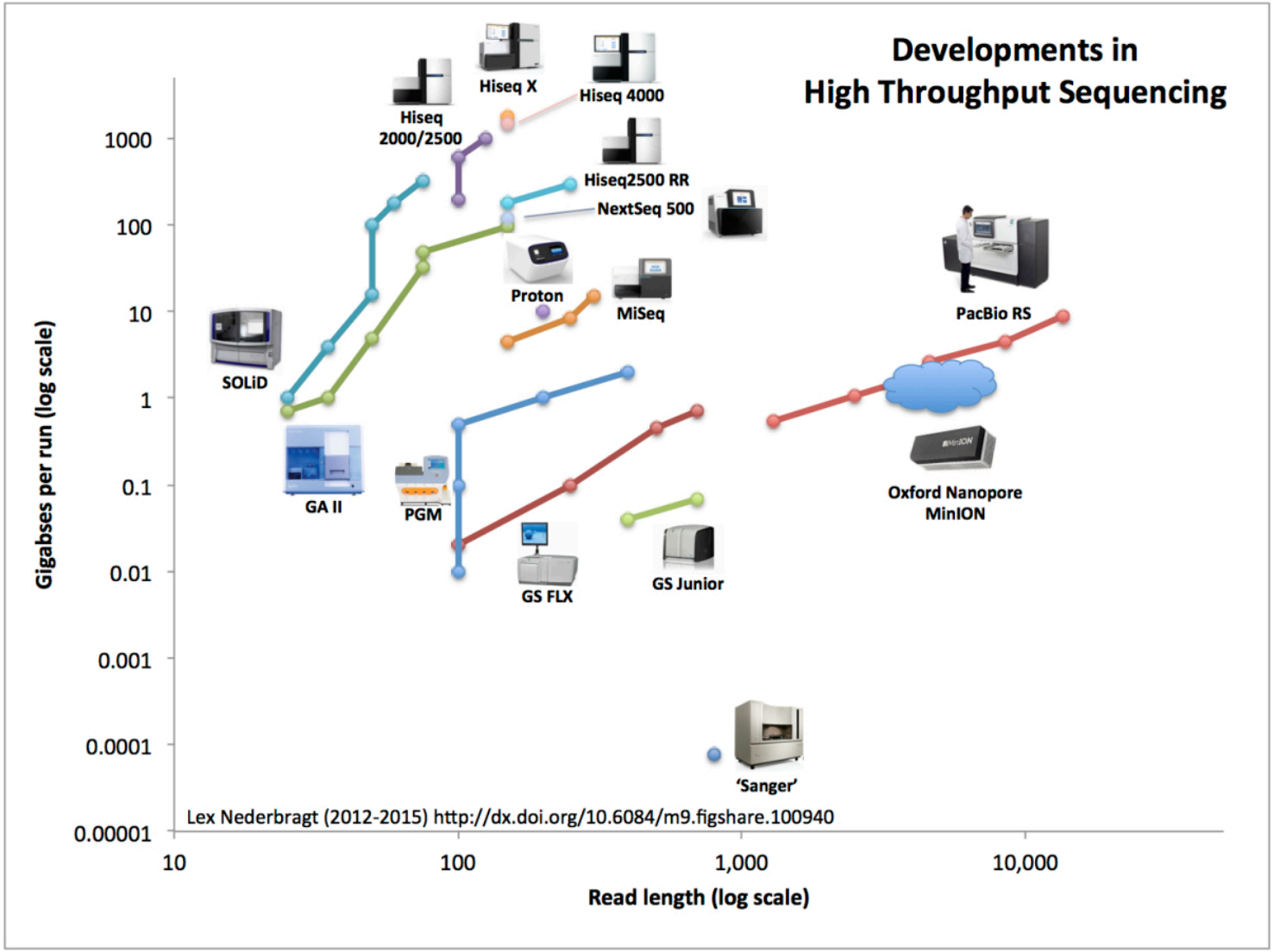
Schematic representation of nanopore sequencing system.

- Upper protein → ssDNA.
- 2nd protein
- forms a nanopore in a membrane.
- contains an adaptor molecule reduce the speed of passing DNA through the pore.
- Each base obstructs the flow to a different degree.
- PromethION…

(http://www.springer.com/978-1-4614-7725-9)

(picture from MIT's Technology Review)

Inter-Platform Comparison



(https://flxlexblog.wordpress.com)

The four main advantages of NGS over classical Sanger sequencing

- **Speed**
  NGS is quicker than Sanger sequencing in two ways.
  - Chemical reaction may be combined with the signal detection, whereas in Sanger sequencing these are two separate processes.
  - 1 read can be taken at a time in Sanger sequencing, whereas NGS is massively parallel.

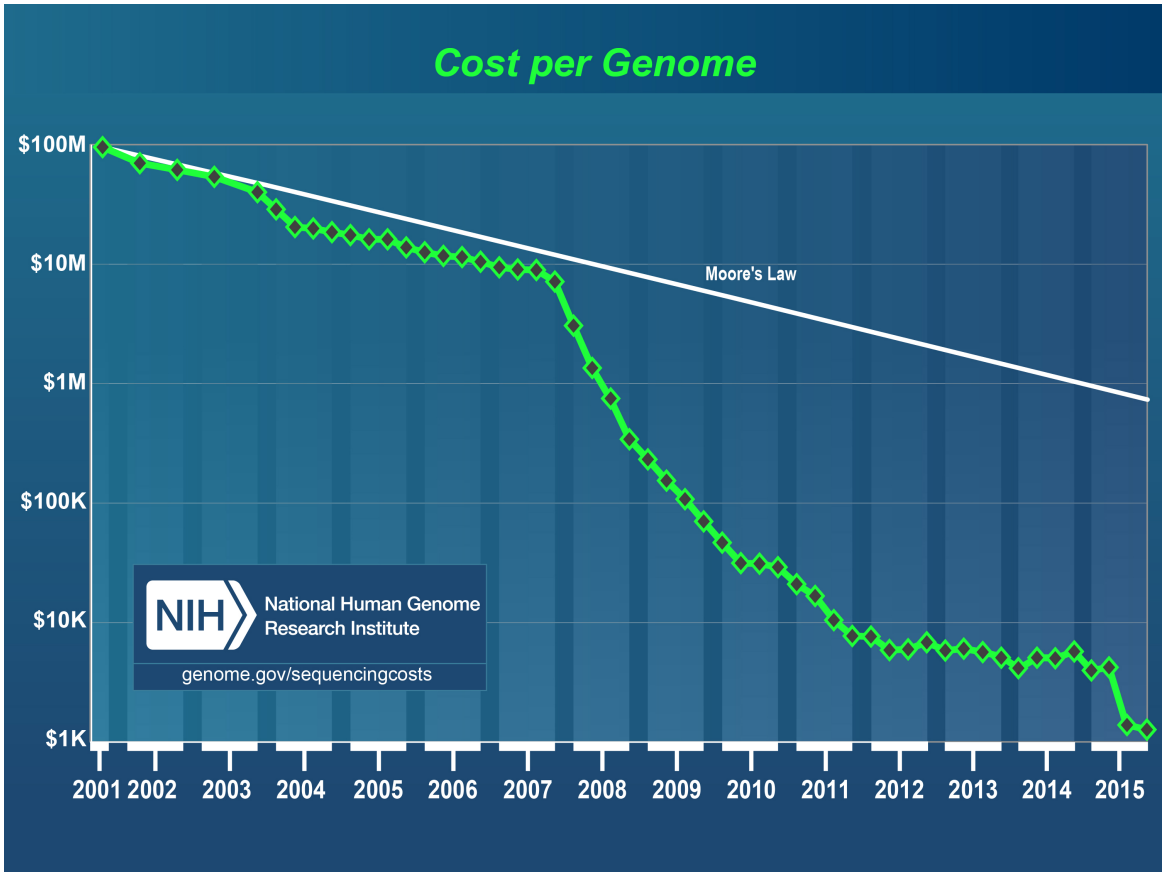
- **Cost**
  The human genome sequence cost $300M.
  Sequencing a human genome with Illumina allows to approach the $1,000 expected.

- **Sample size**
  needs significantly less starting amount of DNA/RNA

- **Accuracy**
  More repeats than with Sanger sequencing → greater coverage, higher accuracy and sequence reliability (individual reads less accurate for NGS).

DNA Sequencing costs

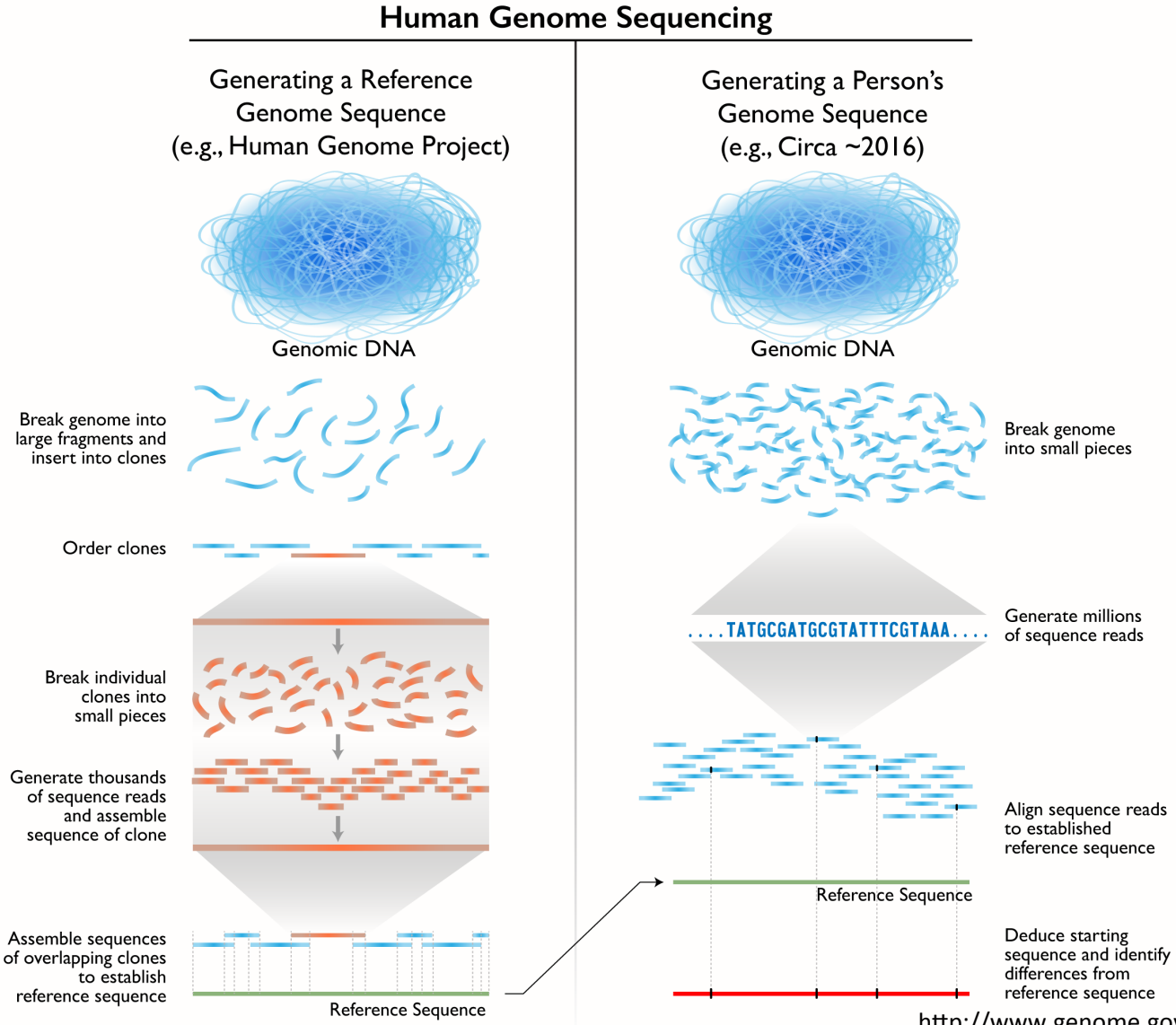
- (Data from the NHGRI Genome Sequencing Program (GSP)



Accurately determining the cost for sequencing a given genome (e.g., a human genome) is not simple.

http://www.genome.gov/sequencingcosts/

Comparison of human genome sequencing methods HGP vs. ~ 2016



http://www.genome.gov/sequencingcosts/

Part 4

DNA-Seq Protocol : Overview

DNA-Seq

Protocol for Library Construction

STEP 01 Genomic DNA Purification

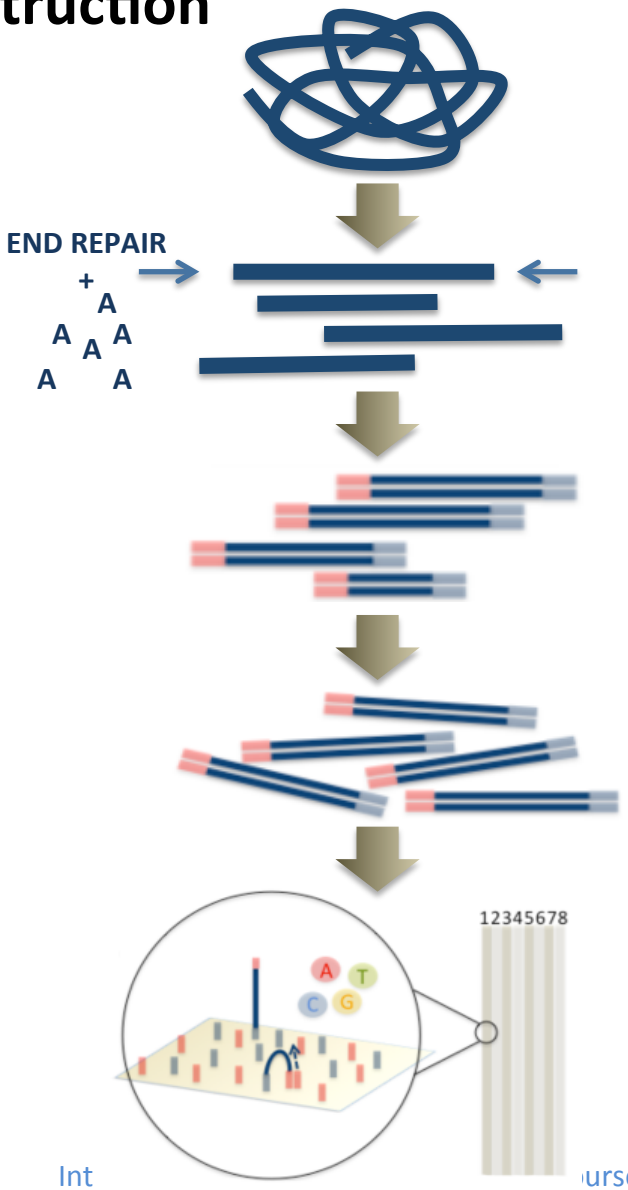
STEP 02 Genomic DNA Fragmentation

STEP 03 End repair and A-tailing

STEP 04 Adapter Ligation

STEP 05 Size Selection & PCR

STEP 06 Sequencing



DNA-Seq

Kits for DNAseq Library prep

- **Illumina-compatible DNA-Seq Library Prep Kits**

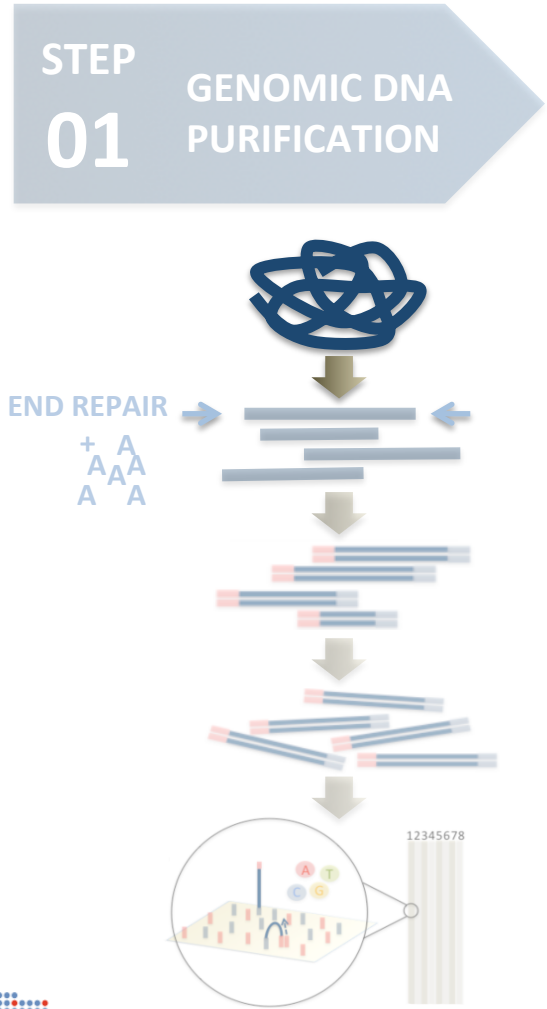
NEXTflex™ Rapid DNA-Seq Kit - DNA-Seq library prep kit, 1 ng - 1 µg input DNA
NEXTflex mtDNA-Seq Kit - mtDNA libraries
NEXTflex™ DNA Sequencing Kits - DNA-Seq library prep kit, 1 µg of input DNA
NEXTflex™ PCR-Free DNA Sequencing Kit - Amplification-free DNA-Seq library prep kit for sequencing 0.5 µg – 3 µg of input DNA
NEXTflex™ PCR-Free Barcodes - Up to 48 barcodes for use with the NEXTflex™ PCR-Free DNA-Seq Kits and other DNA-Seq protocols

KAPA HyperPlus Kits - input DNA from 1 ng – 1 µg
KAPA Hyper Prep Kits - 250 ng FFPE DNA or less + fewer cycles of amplification with KAPA HiFi DNA Polymerase (duplication rates + coverage)

(http://www.biooscientific.com/Next-Gen-Sequencing/Illumina-DNA-Library-Prep-Kits/à (https://www.kapabiosystems.com/product-applications/products/next-generation-sequencing-2/dna-library-preparation/

DNA-Seq

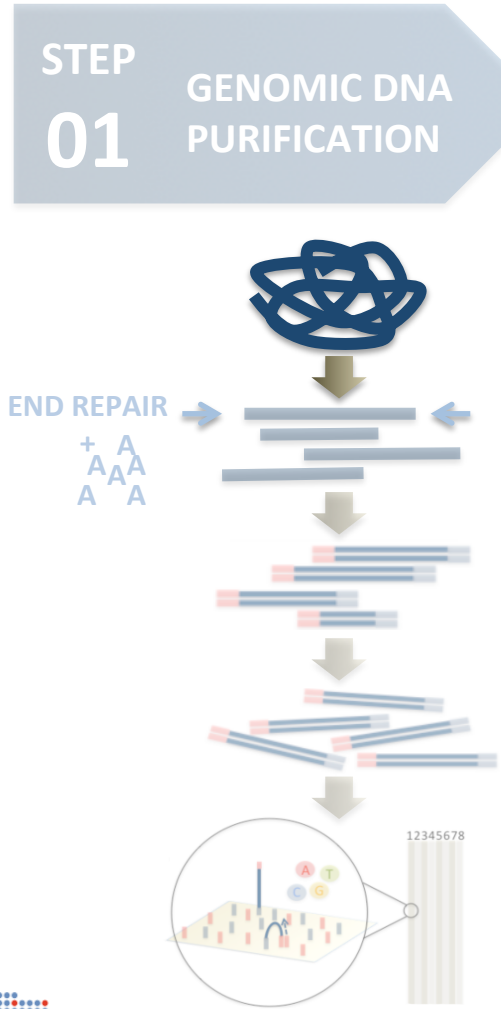
Critical steps in DNAseq Library prep



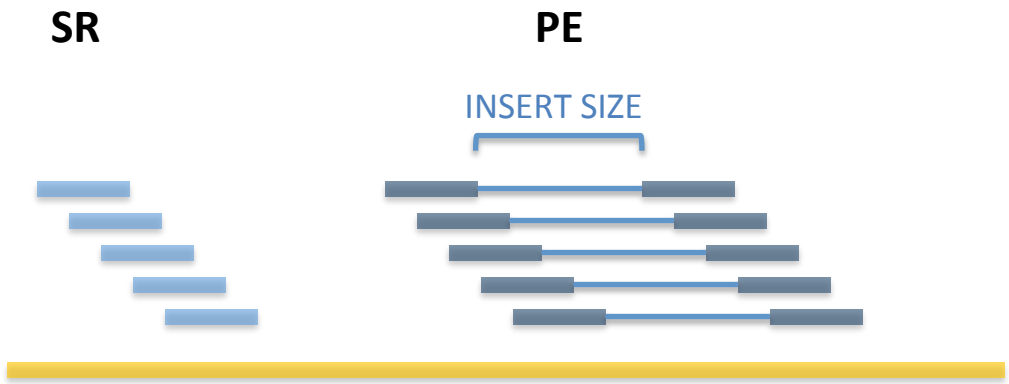
- **Starting material: QC**
  - ***Quality Control***
    - gel visualization, Bioanalyzer (Agilent, Bio-rad)
  - ***Quantity Control***
    - Nanodrop, Qubit…
- **Experimental design**
  - SR (single read) or PE (paired-end)
  - Multiplexing or not
  - *de novo* or not
  - …

DNA-Seq

Critical steps in DNAseq Library prep



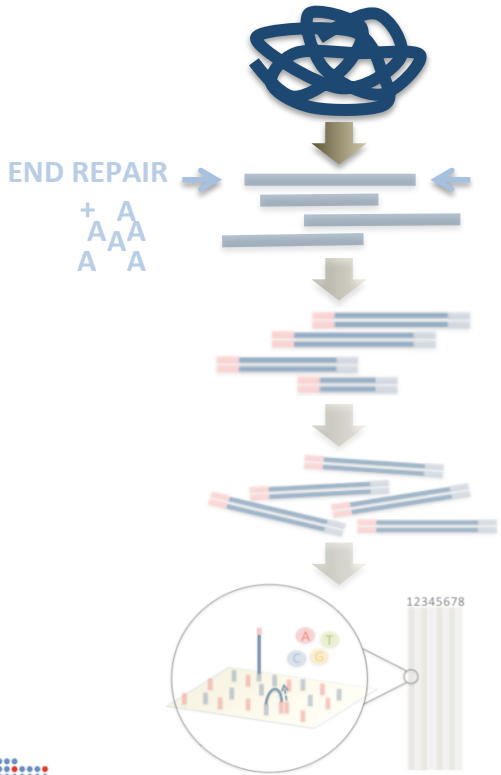
- **Experimental design**
  - SR (single-end reads) or PE (paired-end reads)



DNA-Seq

Critical steps in DNAseq Library prep

STEP 01 GENOMIC DNA PURIFICATION



Experimental design

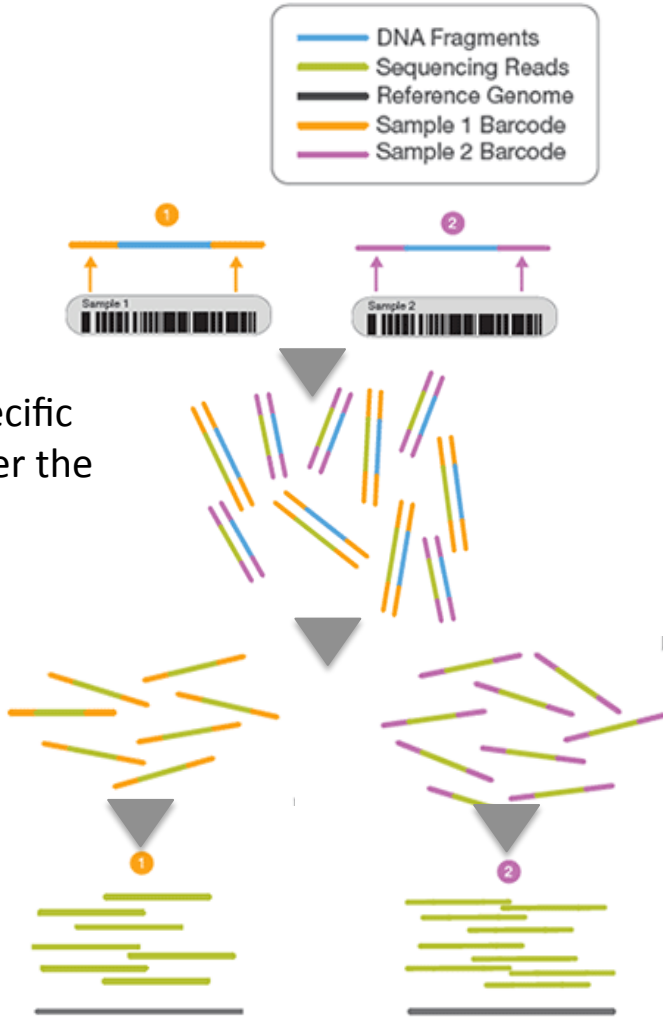
- Multiplexing or not ?

- **multiplexing** = attach to a specific **barcode** sequence to identify later the sample from which it originates.

- Libraries pooled and sequenced in parallel.

- Reads from each library are differenciated by using barcode to de-multiplex

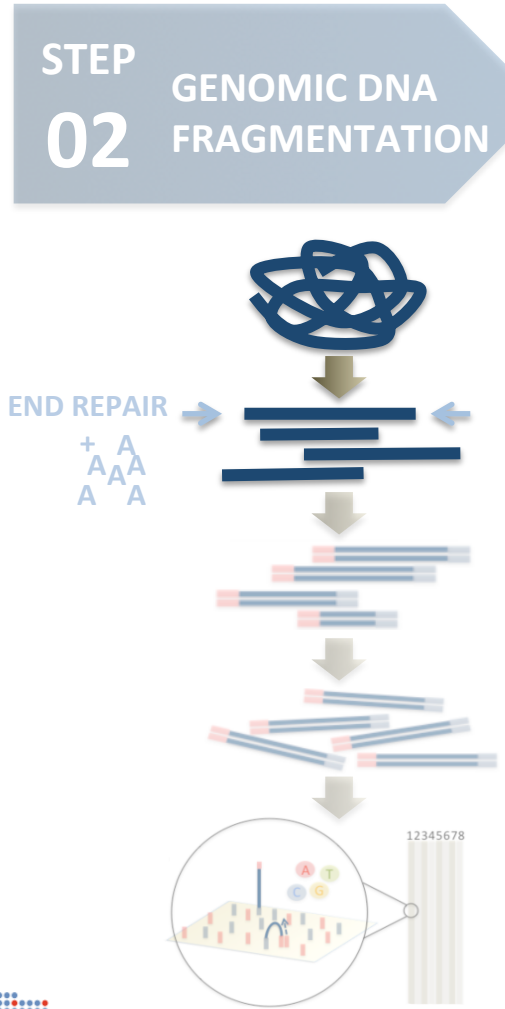
- Each set is aligned to the reference genome



www.illumina.com

DNA-Seq

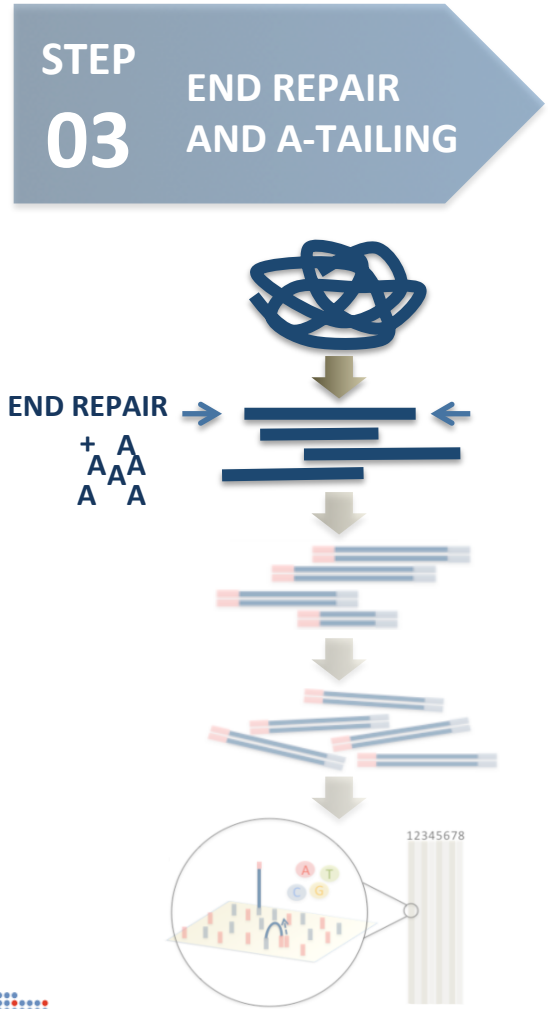
Critical steps in DNAseq Library prep



- **Fragmentation**
  - ***Can be included in the kit***
    - Optimization of fragmentation parameters
  - ***Several methods***
    - Enzymatic, Nebulization, acoustic shearing…
- **Starting material: input**
  - ***Low Quality DNA***
    - Caution in size selection
  - ***High Quality DNA***
    - Size selection

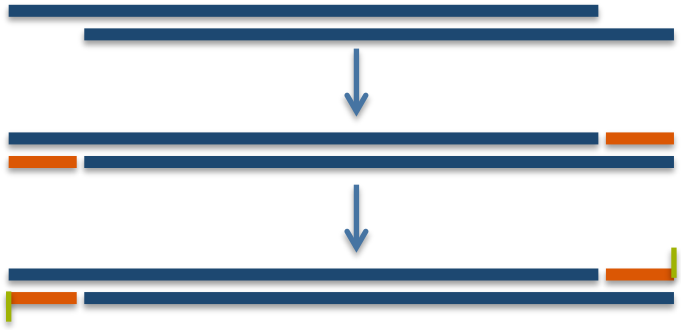
DNA-Seq

Critical steps in DNAseq Library prep



- **Repair ends**

  - Converts overhangs:
    Blunt ends + Phosphorylates 5’-end
  - Reagents:
    dNTP, T4 DNA pol, Klenow – Kinase/ATP (T4 PNK)
  - Simple enzymatic reaction

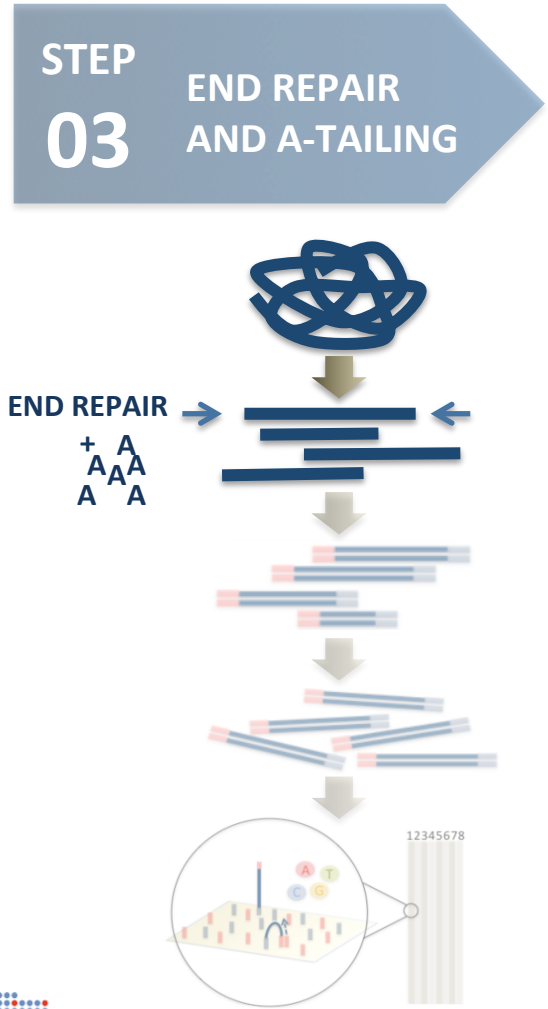


BLUNT ENDING BY EXONUCLEASE

5’-END PHOSPHORYLATION

DNA-Seq

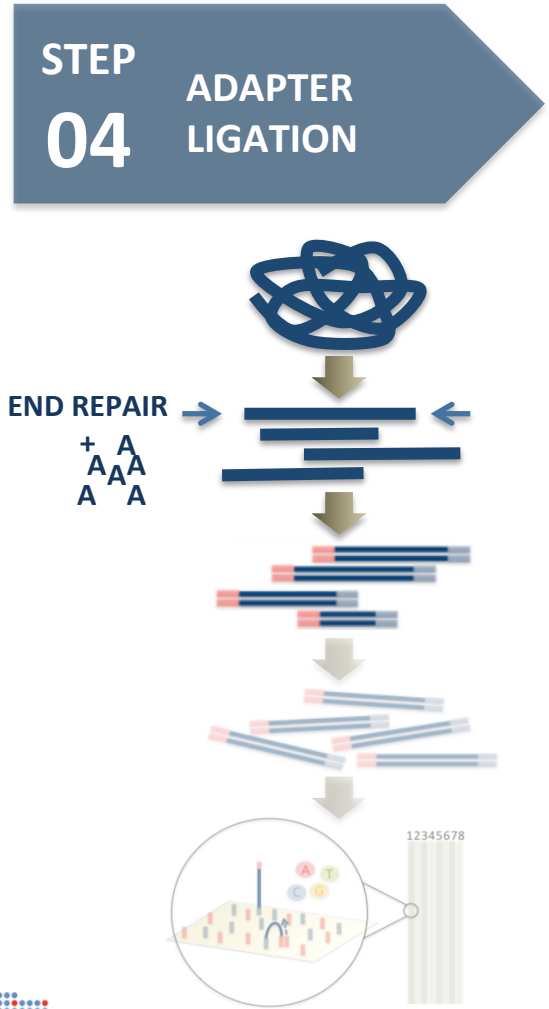
Critical steps in DNAseq Library prep



- **A-tailing (Adenylation)**
  - Adds ‘A’ base to the 3' end of the blunt phosphorylated DNA fragments
  - Prevents
    - Formation of adapters dimers
    - Concatemers
  - Reagents
    1 mM dATP, Klenow exo (3' to 5' exo minus)

DNA-Seq

Critical steps in DNAseq Library prep

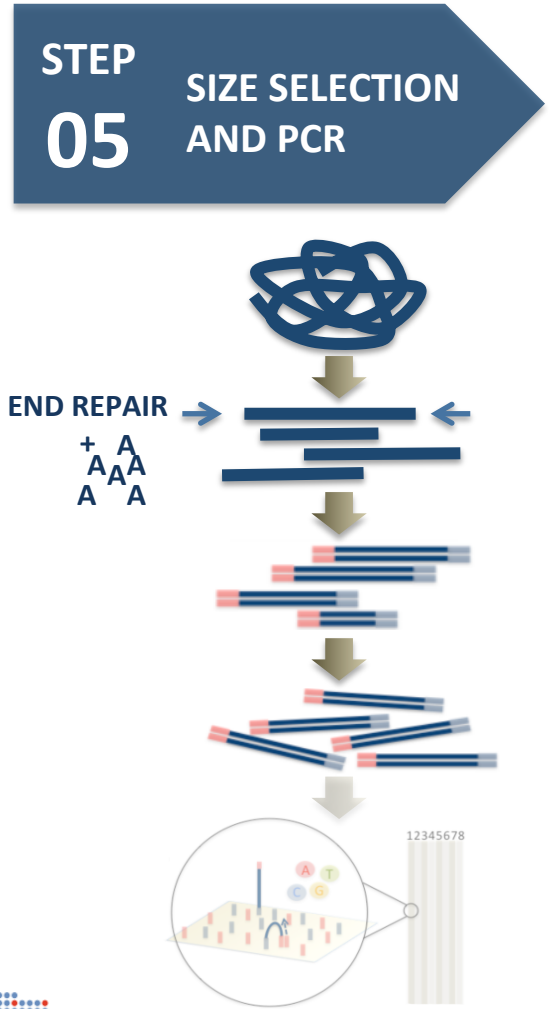


- **Adapter Ligation**

- Provided or custom-designed
- Adapter concentration affects ligation, adapter and adapter-dimer carryover
- Robust Ligation efficiency for adapter:insert molar ratios between 10:1 and >200:1
- Adapter ratio >200:1 for low-input applications.
- Adapter quality
- Post-Ligation cleanup

DNA-Seq

Critical steps in DNAseq Library prep



- **Size selection : Read length considerations**

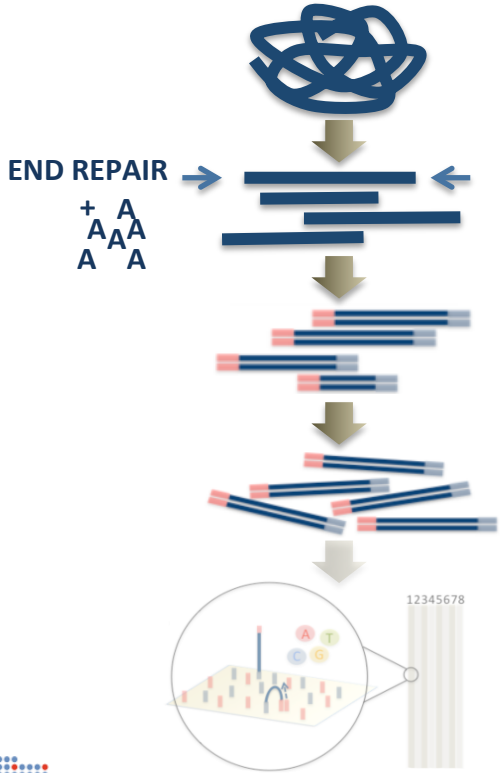
  - Size select 300 – 400 bp or 350 – 500 bp, post-ligation
  - Ensures maximum coverage of most inserts
  - Problem of non-uniform genome coverage
  - Problem of material loss

→ Strategy to focus read lengths during sample and library preparation

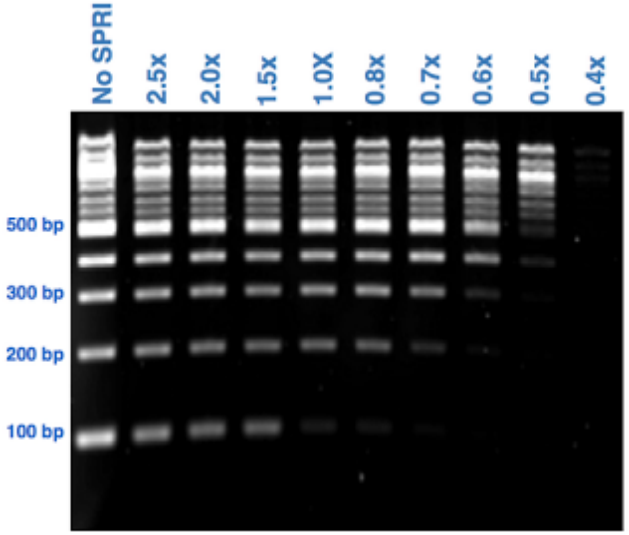
DNA-Seq

Critical steps in DNAseq Library prep

STEP 05 SIZE SELECTION AND PCR



- **Size selection : Read length considerations**
  - Double solid-phase reverse immobilization (SPRI) selection methods allow to reshape the input fragment distribution into well-defined ranges.
  - SPRI + Reverse-SPRI



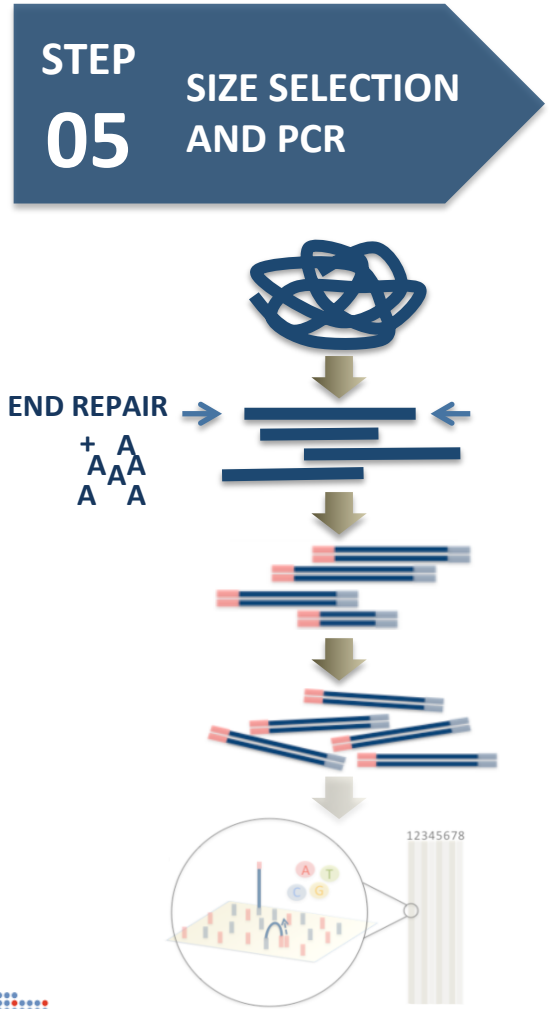
FRAGMENTS ON BEADS

FRAGMENTS IN SUPERNATANT

https://www.kapabiosystems.com/

DNA-Seq

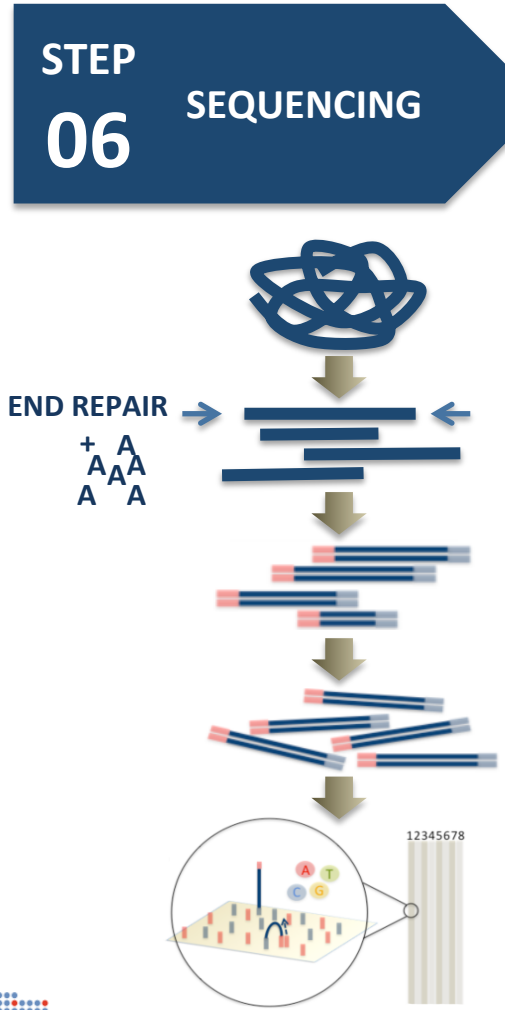
Critical steps in DNAseq Library prep



- **Library Amplification (PCR)**
  - Amplifies the amount of DNA in the library
  - Selectively enrich DNA fragments with adapter molecules on both ends
  - Post-amplification cleanup
- **QC**
  - Quality & Quantity & size check

DNA-Seq

Critical steps in DNAseq Library prep



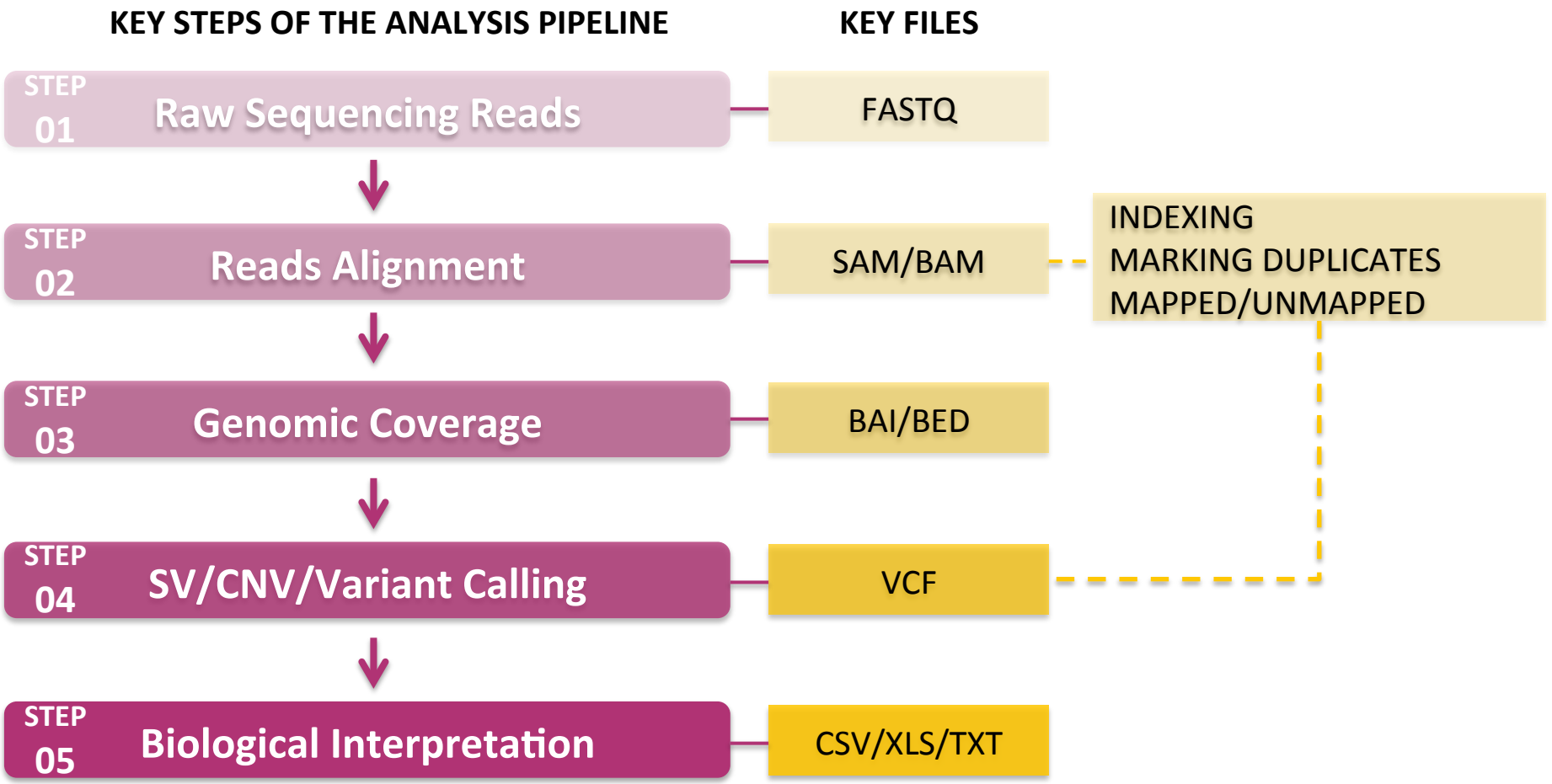
- **DNA Sequencing**

  - input : Library constructed
    - Whole-genome
    - Whole-exome
    - Target region
  - Cluster amplification + sequencing + base calling
  - QC (run report)
  - output : sequenced « reads » (fastq files)

Part 5

DNA-Seq Analysis Pipeline and File Formats

DNA-Seq Analysis Pipeline and associated files



DNA-Seq Analysis Pipeline and associated files

STEP 01 RAW SEQUENCING READS

FASTQ

SEQUENCING RUN/QC

```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
@EAS54_6_R1_2_1_540_792
TTGGCAGGCCAAGGCCGATGGATCA
+
;;;;;;;;;;;7;;;;;-;;;3;83
@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
;;;;;;;;;;;9;7;;.7;393333
```

Sequencing output

- **FASTQ** (text) format.
- Potentially **SRA** (binary), but rather used for public data online.

Fastq file

- Improvement of the Sanger breakthrough (associating each nucleotide to a quality score)
- Hundreds of millions of lignes/rows
- Blocks of 4 lignes (@)
- Example

DNA-Seq Analysis Pipeline and associated files

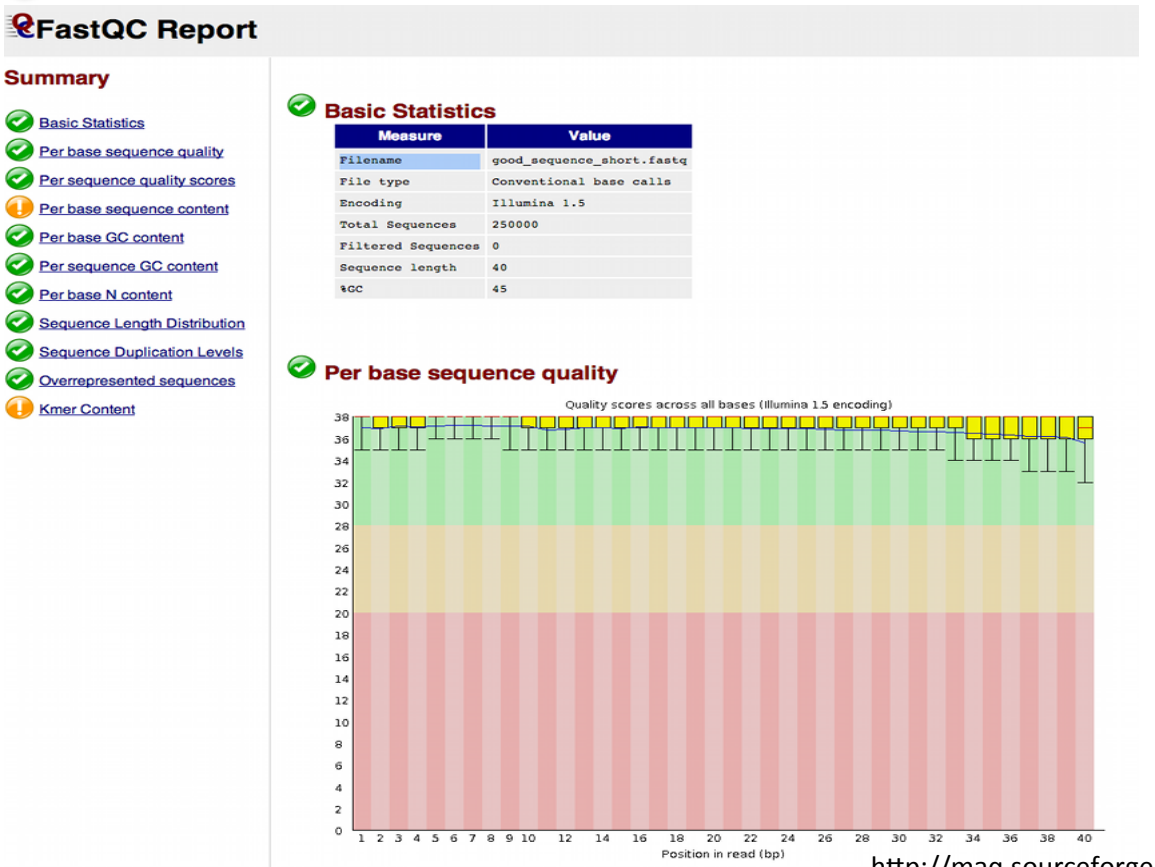
STEP 01 RAW SEQUENCING READS

FASTQ

SEQUENCING RUN/QC

```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
@EAS54_6_R1_2_1_540_792
TTGGCAGGCCAAGGCCGATGGATCA
+
;;;;;;;;;;;7;;;;;-;;;3;83
@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
;;;;;;;;;;;9;7;;.7;393333
```

FastQC



http://maq.sourceforge.net
https://wiki.hpcc.msu.edu

DNA-Seq Analysis Pipeline and associated files

STEP 01 RAW SEQUENCING READS

FASTQ

SEQUENCING RUN/QC

```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
@EAS54_6_R1_2_1_540_792
TTGGCAGGCCAAGGCCGATGGATCA
+
;;;;;;;;;;;7;;;;;-;;;3;83
@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
;;;;;;;;;;;9;7;;.7;393333
```

FastQC + Trimming + bad reads removal

FastQC Report

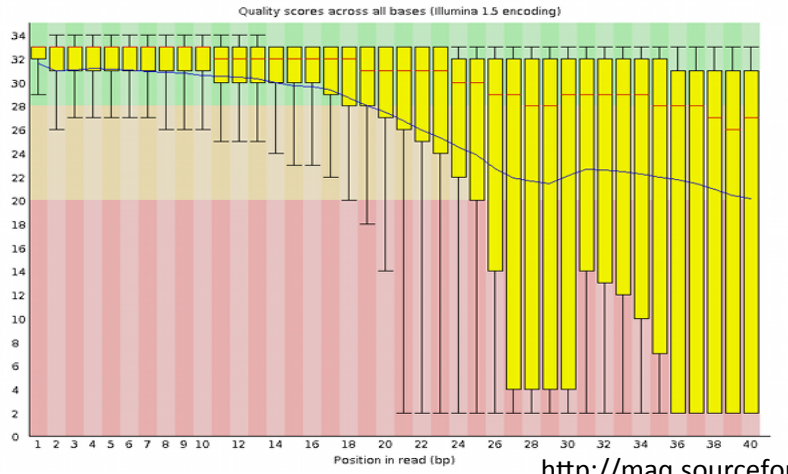
Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

Basic Statistics

Measure	Value
Filename	bad_sequence_short.fastq
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	395288
Filtered Sequences	0
Sequence length	40
%GC	47

Per base sequence quality



http://maq.sourceforge.net
https://wiki.hpcc.msu.edu

DNA-Seq Analysis Pipeline and associated files

STEP 02 READS MAPPING/ ALIGNMENT

SAM/BAM

BWA/BOWTIE…/SAMTOOLS

Input files

- .fastq
- Reference Genome (.fasta, .fa, .fai)
- GFF/GTF, GFF3

Reads alignment

- BWA / BOWTIE (Burrows-Wheeler transform)
- *de novo*: NEWBLER (454)…

Output files

- .sam / .bam

DNA-Seq Analysis Pipeline and associated files

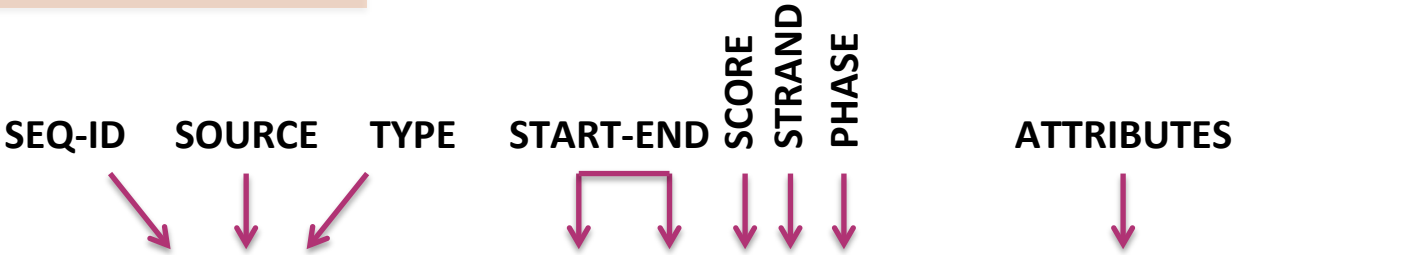
STEP 02 READS MAPPING/ ALIGNMENT

SAM/BAM

BWA/BOWTIE…/SAMTOOLS

GFF3

- 1 line for 1 feature
- tab-separated columns
- 9 columns + optional additional information




```
##gff-version 3
ctg123 . mRNA            1300  9000  .  +  .  ID=mrna0001;Name=sonichedgehog
ctg123 . exon            1300  1500  .  +  .  ID=exon00001;Parent=mrna0001
ctg123 . exon            1050  1500  .  +  .  ID=exon00002;Parent=mrna0001
ctg123 . exon            3000  3902  .  +  .  ID=exon00003;Parent=mrna0001
ctg123 . exon            5000  5500  .  +  .  ID=exon00004;Parent=mrna0001
ctg123 . exon            7000  9000  .  +  .  ID=exon00005;Parent=mrna0001
```

DNA-Seq Analysis Pipeline and associated files

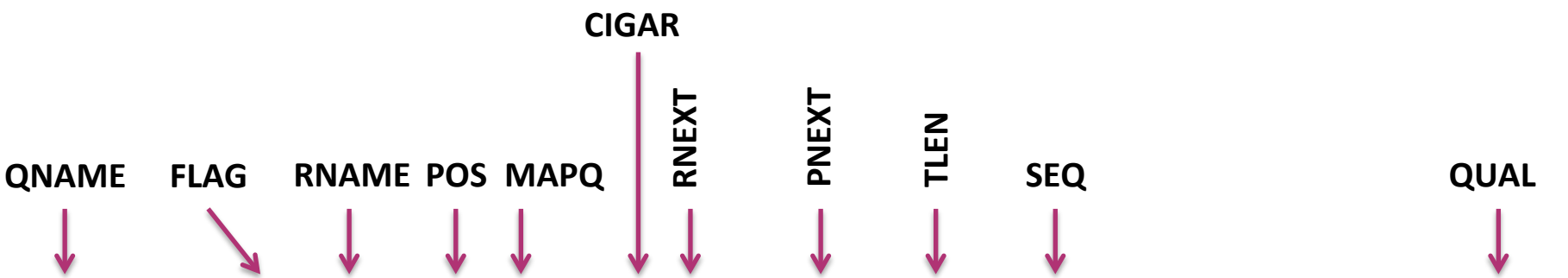
STEP 02 READS ALIGNMENT

SAM/BAM

BWA/BOWTIE…/SAMTOOLS

SAM / BAM

- Header lines + Alignments sections
- tab-separated columns
- 11 columns
- Samtools (view, sort, index…)
- Removal of duplicated reads that affects Variant Calling




```
1:497:R:-272+13M17D24M      113     1       497     37      37M     15      100338662       0       CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG   0;==-==9;>>>
19:20389:F:275+18M2D19M     99      1       17644   0       37M     =       17919   314     TATGACTGCTAATAATACCTACACATGTTAGAACCAT    >>>>>>>>>>>>>>>>>>>>
19:20389:F:275+18M2D19M     147     1       17919   0       18M2D19M        =       17644   -314    GTAGTACCAACTGTAAGTCCTTATCTTCATACTTTGT    ;44999;499<8
9:21597+10M2I25M:R:-209     83      1       21678   0       8M2I27M =       21469   -244    CACCACATCACATATACCAAGCCTGGCTGTGTCTTCT    <;9<<5><<<<><<<><><<>
```

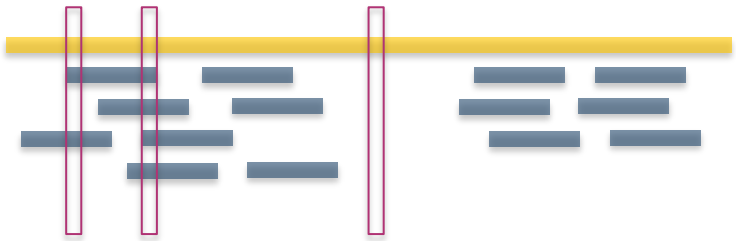
http://samtools.sourceforge.net/SAM1.pdf
http://genome.sph.umich.edu/wiki/SAM

DNA-Seq Analysis Pipeline and associated files

STEP 03 GENOMIC COVERAGE

BAI/BED

SAMTOOLS/BEDTOOLS



Coverage

- IGV = Integrative Genomics Viewer (https://www.broadinstitute.org/igv/)

- other useful file formats
BED = (Browser Extensible Data)
(http://genome.ucsc.edu/FAQ/FAQformat)

- Coverage associated to 3 different concepts:
  - Fold Coverage (number + X)
  - Breadth of Coverage
  - Depth of Coverage

DNA-Seq Analysis Pipeline and associated files

STEP 04 SV/CNV/VARIANT CALLING

VCF

SV/CNV/Variant Calling

- **Structural variations (SV)**

Deletions, duplications, copy-number variations, insertions, inversions, translocations…

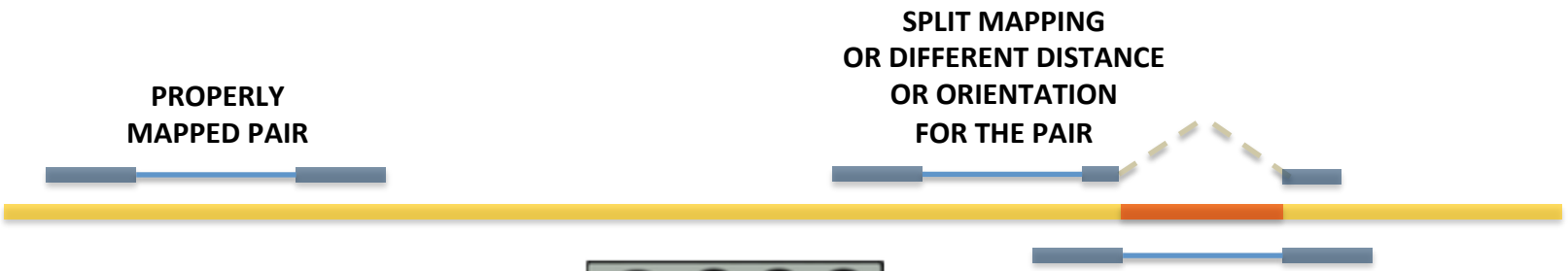
- **Copy number Variations (CNV)**

Deletions or duplications of genes or relatively large regions of the genome that affect chromosomes▸

Variant Calling (SNPs and small InDels)

- SNPs: affects only 1 nucleotide

- InDels: affects 1 or several nucleotides

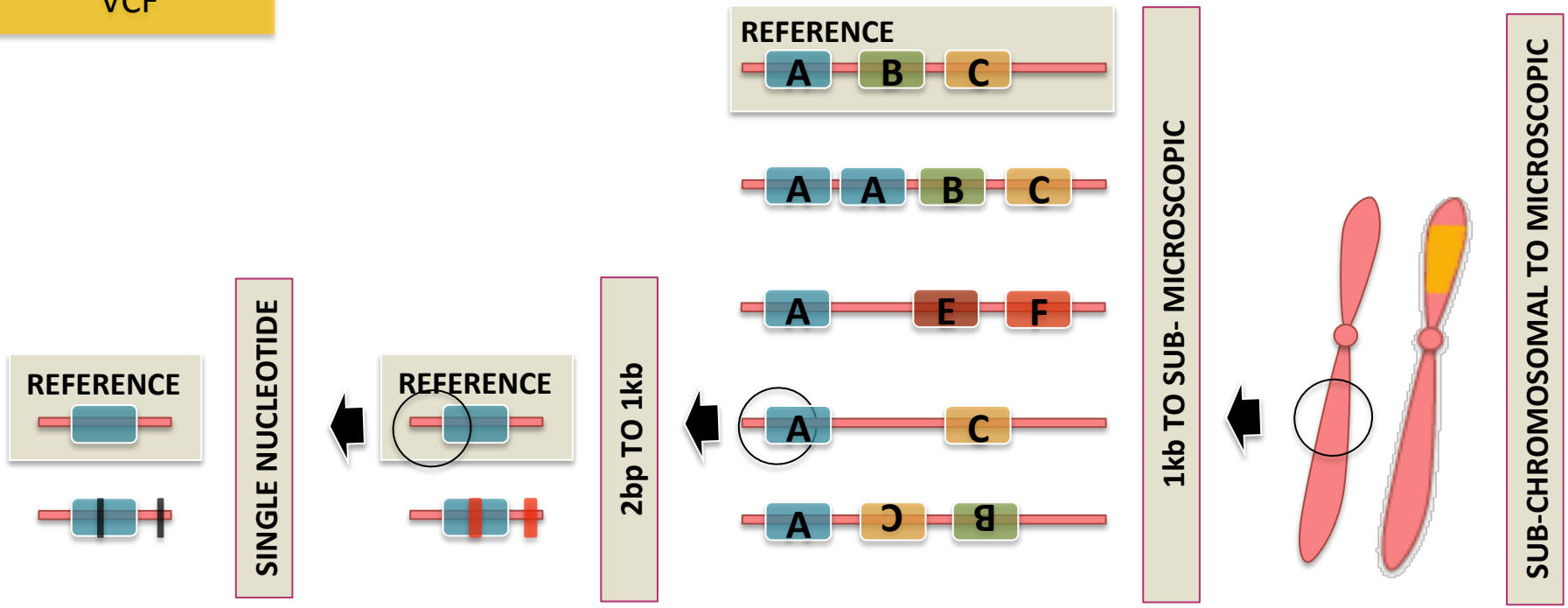


DNA-Seq Analysis Pipeline and associated files

STEP 04 SV/CNV/VARIANT CALLING

- SV/CNV/Variant Calling

VCF



(definitions adapted from Scherer et al., 2007)

DNA-Seq Analysis Pipeline and associated files

STEP 04 SV/CNV/VARIANT CALLING

VCF

SV/CNV/Variant Calling

- **VCF** (Variant Call Format)
  - Text file format storing SNPs and InDels information
  - http://www.1000genomes.org/node/101
  - Obtaining variants listed in this format is a multistep procedure involving different tools but standardized
  - Headers (meta-information) + data lines
  - 8 required fields, tab-delimited

```
#CHROM POS     ID        REF ALT    QUAL FILTER INFO
FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G       A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2
GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T       A       3    q10    NS=3;DP=11;AF=0.017
GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A       G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667
;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T       .       47   PASS   NS=3;DP=13;AA=T
GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTCT    G,GTACT 50   PASS   NS=3;DP=9;AA=G
GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

DNA-Seq Analysis Pipeline and associated files

STEP 05 BIOLOGICAL INTERPRETATION

CSV/XLS/TXT

From Variant annotation to data mining

- web-based
- available packages

Aim

- Functional impact of variants (synonymous or not…)
- Gene Ontology Annotation (BP, MF, CC)
- Pathway / Network information
- Predictions of pathogenicity / severity

NB: DAVID (**D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery) to switch between databases https://**david**.ncifcrf.gov/

NIH NIAID

H3ABioNet
Pan African Bioinformatics Network for H3Africa

DNA-Seq Analysis
Take-home messages

- **Biological question**
  - need to be clearly defined first, so that the design of the experiment, the library construction and the pipeline of analysis could be prepared accordingly

- **Platfom**
  - Each one has its own specificities that needs to be understood before choosing one
  - Different technologies, short reads (Illumina…) vs long reads (PacBio…)
  - Rapidly evolving, several limitaions (PCR bias for GC rich regions…)
  - Combination of different platforms possible (*de novo*…)

- **Input / Output files**
  - Companion indexed files needed (.fa & .fai / .bam & .bai / .vcf & .vcf.idx…)
  - text based (FASTA, FASTQ, SAM, GTF/GFF, BED, VCF, WIG) or binary (BAM, BCF, SFF)
  - 1-based (GFF/GFT, SAM/BAM, WIG) or (0-based : BED)

DNA-Seq Analysis

The command-line environment

- Understand see how these files are generated practically (no demo)
- Give you an idea about how to deal with these files easily once they are generated and given to you by your sequencing plateform.
- Make you work a bit on one specific file (a vcf file), using the command line interface (assignment).

DNA-Seq Analysis

The command-line environment

- Reminder of the command line syntax and some basic commands to manipulate files
  → the same syntax is used for NGS analysis…but…using other specific commands (algorithms, tools…)
- Examples of command lines to generate or retrieve data from NGS data files using these specific commands
- Basic linux command lines can be useful to parse files: How to interrogate a vcf file using basic linux command lines?

DNA-Seq Analysis

The command-line environment

- **The NGS datasets: reminder**
  - Outputs large files
  - Output files contains various kinds of informations that you need to parse
  - fastq: quality associated to each read…
  - sam/bam: quality of the mapping…
  - vcf: variants, annotation of the effects that these variants can have…

- **The command-line environment**
  - UNIX Operating System: able to deal with multi-task & multi-user needs
  - i.e. can even handle multiple files at a time (useful if multiple samples)
  - Brings flexibility to handle large files
  - Allows to easily parse the content of a (big) file

DNA-Seq Analysis

The command-line environment

- **The basic commands you have seen are useful for NGS**

→Remember that many kind of files are generated through the NGS analysis pipeline

▸ So you should know at this stage the file system basics that allows you to work with many files and classify them

DNA-Seq Analysis

The command-line environment

- **What you know about the command-line environment should allow you to:**

▸ Create directories and move through file system

▸ At this stage you should be able to handle easily queries to parse large files

- able to search for a particular pattern
- able to select specific information columns

→ Easily **interrogate** the large amount of information in the output files

DNA-Seq Analysis

The command-line environment

Viewing & Manipulating

cat

- view the content of a short file

```
$ cat    file1
```

more

- view the content of a long file, step by step

```
$ more    file1
```

less

- view the content of a long file, by portions

```
$ less    file1
```

DNA-Seq Analysis

The command-line environment

Viewing & Manipulating

head

- view the first lines of a (long) file

```
$ head    file1
```

NB: By default (without options), displays the 10 first lines

tail

- view the last lines of a (long) file

```
$ tail    file1
```

NB: By default (without options), displays the 10 last lines

DNA-Seq Analysis

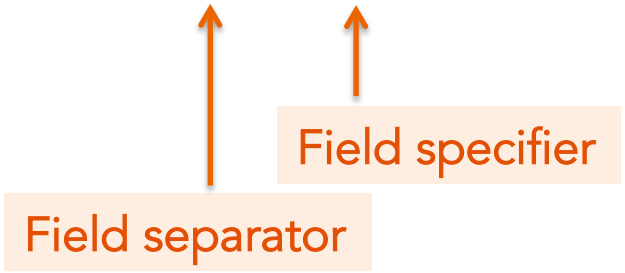
The command-line environment

Viewing & Manipulating

cut

- Extract specific fields from a file

```
$ cut   -d' '   -f1,2     file1
```



DNA-Seq Analysis

The command-line environment

Viewing & Manipulating

grep

- search for the occurrence of a specific pattern in a file (regular expression using the wildcards…)

Careful : grep displays the whole line containing this specific pattern XXX

```
$ grep   XXX   file1
```

Could be used to display all lines that DO NOT contain a specific pattern

```
$ grep   -v   XXX   file1
```

DNA-Seq Analysis

The command-line environment

Viewing & Manipulating

WC

- Prints different kind of counts for a file

```
$ wc    -l    file1
```

Prints line counts

DNA-Seq Analysis

The command-line environment

Redirecting characters

|

The "|" character allows to combine several commands, by sending the result of one command to another

```
$ grep   XXX   file1   |    wc    -l
```

Prints line counts instead of displaying the result on the screen

DNA-Seq Analysis

The command-line environment

Redirecting characters

>

The “>” character allows to redirect the result of a command to a new file

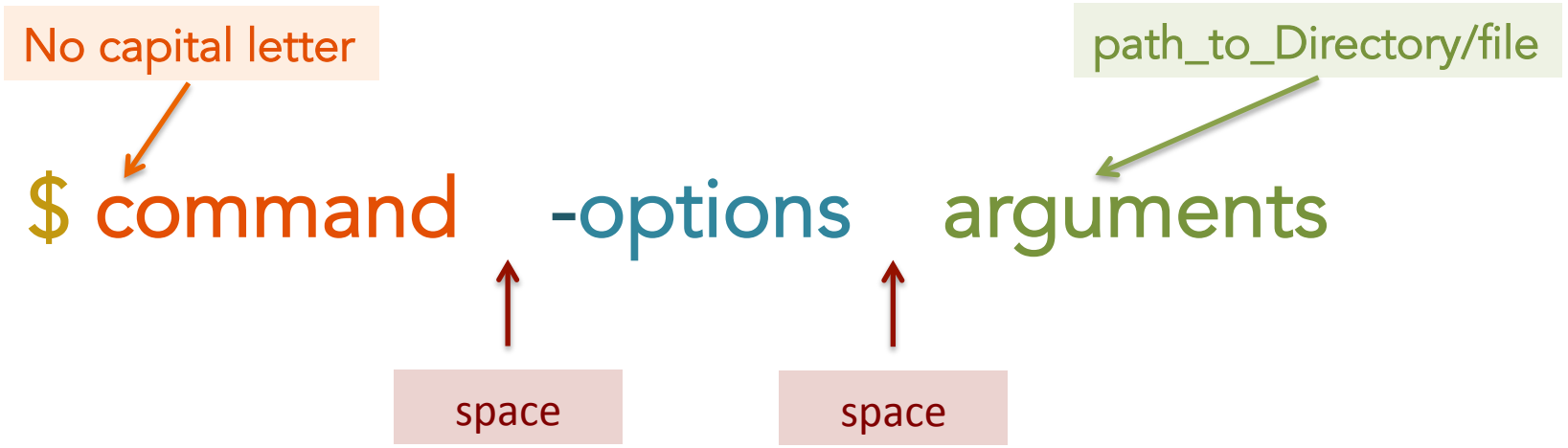
```
$ grep   XXX   file1   >    file2
```

Prints line counts instead of displaying the result on the screen

DNA-Seq Analysis

The command-line environment

- **Command-line syntax**



Useful commands:
$ man NameOfTheCommand
$ pwd

DNA-Seq Analysis

The command-line environment

FASTQ	FASTQC/…
SAM/BAM	BWA/BOWTIE…/SAMtools
VCF	VCFtools/jannovar

DNA-Seq Analysis

The command-line environment

FASTQ | FASTQC/…

To run FastQC (provided it is installed):

- specify the files you want to process on the command line
- FastQC will generate an HTML report for each file (embeded graphs)

```
fastqc seqfile1 seqfile2 .. seqfileN
```

```
fastqc   --help
```

```
fastqc [-o output dir] [--(no)extract]
```

-o –outdir

Create all output files in the specified output directory (dir must already exist).

--extract

Uncompress the zipped output file in the same dir after being created.

--noextract

Do not uncompress the output file after creating it.

DNA-Seq Analysis

The command-line environment

FASTQ | FASTQC/…

Trimming

→ Trimming low Quality Bases

Low quality base reads from the sequencer can cause an otherwise mappable sequence not to align → different open source tools can trim off 3' bases → produce a FASTQ file of the trimmed reads to use as input to the alignment program.

e.g.: FASTX-Toolkit

```
gunzip -c Sample_R1.cat.fastq.gz | fastx_trimmer -l 50 -Q 33 > trimmed.fq
```

Trim down to 50 bases (last base is 50)

option that specifies how base qualities on the 4th line of each fastq entry are encoded

DNA-Seq Analysis

The command-line environment

FASTQ | FASTQC/…

Trimming

→ Trimming Adapters

A 3' adapter contamination can cause the insert sequence not to align (adapter sequence ≠ bases at the 3' end of the reference genome sequence).
Unlike general fixed-length trimming, adapter trimming removes different numbers of 3' bases depending on where the adapter sequence is found.

e.g.: cutadapt

```
cutadapt -a   AAAT   test.fastq   -o  test_new.fastq
cutadapt -m 22 -O 10 -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
```

-m22 = discard any sequence that is smaller than 22 bases after trimming
-O10 = says not to trim 3' adapter sequences unless at least the first 10 bases of the adapter are seen at the 3' end of the read

DNA-Seq Analysis

The command-line environment

SAM/BAM | BWA/BOWTIE…/SAMtools

Mapped and unmapped reads are imported and stored into SAM/BAM format

samtools : A suite of useful commands to visualize or get informations from .sam/.bam

```
#from SAM to BAM conversion
samtools view  test.sam > test.bam

# for sorting and indexing alignment
samtools sort    file.bam     -o     file.sorted.bam
samtools index   file.sorted.bam   file.sorted.bam.bai

#all reads mapping on a certain portion of chr1 or all the chr1 in another bam
samtools index test.bam
samtools view   test.bam     chr1:200000-500000
samtools view -b   test.bam   chr1   >    test_chr1.bam
```

DNA-Seq Analysis

The command-line environment

SAM/BAM | BWA/BOWTIE…/SAMtools

Mapped and unmapped reads are imported and stored into SAM/BAM format

samtools

The flagstat command provides simple statistics on a BAM file

```
#from SAM to BAM conversion
samtools flagstat  file.bam
```

1	6874858 + 0 in total (QC-passed reads + QC-failed reads)
2	90281 + 0 duplicates
3	6683299 + 0 mapped (97.21%)
4	6816083 + 0 paired in sequencing
5	3408650 + 0 read1
6	3407433 + 0 read2
7	6348470 + 0 properly paired (93.14NaV)
8	6432965 + 0 with itself and mate mapped
9	191559 + 0 singletons (2.81NaV)
10	57057 + 0 with mate mapped to a different chr
11	45762 + 0 with mate mapped to a different chr (mapQ>=5)

DNA-Seq Analysis

The command-line environment

VCF | VCFtools/SNPEff

VCF files contains information about variants
VCF can be used as input and output file for many tools

Variant calling can be done using many available tools and methods (GATK, samtools…) and the output used by many others (VCFtools, VCFminer, snpeff…)

When a mapped read shows a mismatch from the reference genome
→ is the mismatch due to a real SNP???

e.g. How does samtools detect SNPs?
Samtools computes statistics to incorporate different types of information such as:

- number of different reads that share a mismatch from the reference
- the sequence quality data
- the expected sequencing error rates

http://samtools.sourceforge.net/mpileup.shtml

DNA-Seq Analysis

The command-line environment

VCF | VCFtools/jannovar

VCF files contains information about variants
VCF can be used as input and output file for many tools

e.g.samtools & bcftools: 2 steps are required using these commands :

1.samtools

- collect summary information in the input BAMs
- compute the likelihood of data given each possible genotype
- and store the likelihoods in the BCF format (see below). It does not call variants at this stage.

2. Bcftools

- applies the prior and does the actual calling
- can also concatenate BCF files, index BCFs for fast random access and convert BCF to VCF.

http://samtools.sourceforge.net/mpileup.shtml

DNA-Seq Analysis

The command-line environment

VCF | VCFtools/jannovar

Suppose we have :

- a reference sequence in **genome.fasta**, indexed by **samtools faidx**
- position sorted alignment files aln1.sorted.bam and aln2.sorted.bam.

→ you can call SNPs and short INDELs using:

```
#1.Generate a BCF file (binary data format : information about sequence variants (SNPs…)
samtools mpileup -uD -f genome.fasta aln1.sorted.bam aln2.sorted.bam |
bcftools view -bvcg - > file.bcf
```

-u output into an uncompressed bcf file
-D keep read depth for each sample
-f next argument is reference genome file

-b output to BCF format
-v only output potential variant sites (i.e., exclude monomorphic ones)
-c do SNP calling
-g call genotypes for each sample in addition to just calling SNPs

http://samtools.sourceforge.net/mpileup.shtml

DNA-Seq Analysis

The command-line environment

VCF | VCFtools/jannovar

Suppose we have :

- a reference sequence in **genome.fasta**, indexed by **samtools faidx**
- position sorted alignment files aln1.sorted.bam and aln2.sorted.bam.

→ you can call SNPs and short INDELs using:

```
#2.Convert BCF into VCF (flat text file rather than a binary = easier to view)
bcftools view   file.bcf  |   vcfutils.pl   varFilter  -D100   >   file.filt.vcf
```

-D100 filters out SNPs that had read depth higher than 100

http://samtools.sourceforge.net/mpileup.shtml

DNA-Seq Analysis

The command-line environment

VCF | VCFtools/jannovar

Many ways to nnotate the VCF file (vcftools, jannovar…)

e.g. jannovar

Jannovar identifies all transcripts affected by a given variant, and provides HGVS-compliant annotations for diiferent types of variants.

```
#Download the RefSeq transcript database for the release hg19/GRCh37.
$ java -jar jannovar-cli-0.23-SNAPSHOT.jar download   -d   hg19/refseq
#Annotate the test.vcf file
$ java -jar jannovar-cli-0.23-SNAPSHOT.jar annotate-vcf \ -d   hg19_refseq.ser  -i
test.vcf  -o  test.jv.vcf
```

```
1 866511   rs60722469      C       CCCCT    258.62  PASS    EFFECT=INTRONIC;HGVS=SAMD11:NM_152486.2:c.
1 879317   rs7523549       C       T        150.77  PASS    EFFECT=MISSENSE;HGVS=SAMD11:XM_005244727.1
1 879482   .       G       C       484.52  PASS    EFFECT=MISSENSE;HGVS=SAMD11:XM_005244727.1:exon9:c
```

http://charite.github.io/jannovar/

DNA-Seq Analysis

The command-line environment

- all these commands can be run as part of an analysis pipeline
- all files generated can be parsed using specific tools (samtools…)
- text-based files generated can be parsed using basic linux commands
- At this stage you should be able to handle easily queries to parse large files
  - able to search for a particular pattern
  - able to select specific information columns

→ Assignment !