Genome & Transcriptome studies using Next Generation Sequencing (NGS) Technologies : Past and Present

Fatma Guerfali, PhD
Institut Pasteur de Tunis
fatma.guerfali@gmail.com

OVERVIEW

PARTS 1 - 3

Part 1 Genomes and Transcriptomes : From central dogmas to ongoing discoveries

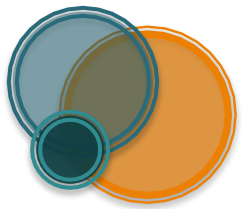
(Central dogma of molecular Biology, Human Genome Project, GENCODE, ENCODE…)

Part 2 Advances in Sequencing Technologies

(Examples of technologies developed for Closed/Open systems for gene expression analysis, Next-Generation Sequencing platforms and technologies …)

Part 3 Overview of NGS (DNA / RNA Seq) Protocols and related file formats

(Overview of protocols for Genomic and Transcriptomic analysis of standard samples using Next-Generation Sequencing and overview of the different formats generated at each step…)



OVERVIEW

PARTS 1 - 3

Part 1 Genomes and Transcriptomes : From central dogmas to ongoing discoveries

(Central dogma of molecular Biology, Human Genome Project, GENCODE, ENCODE…)

Part 2 Advances in Sequencing Technologies

(Examples of technologies developed for Closed/Open systems for gene expression analysis, Next-Generation Sequencing platforms and technologies …)

Part 3 Overview of NGS (DNA / RNA Seq) Protocols and related file formats

(Overview of protocols for Genomic and Transcriptomic analysis of standard samples using Next-Generation Sequencing and overview of the different formats generated at each step…)

NGS PROTOCOLS

CRITICAL STEPS BEFORE LIBRARY PREP

A Library Preparation for DNA-seq

B Library Preparation for RNA-seq

C File Formats

PART 3

NGS PROTOCOLS

CRITICAL STEPS BEFORE LIBRARY PREP

What Biologists think Bioinformaticians can do

- Biologist with an external hard drive in hands, looking completely lost.
- He wants you to analyze his NGS data. These COST A LOT and he hopes that *for once he hasn't missed his experiments…*
- He is counting on you, as he has no idea of what he has in his hands, but he wants "candidate genes, and small p-values, please…"

Adapted from (Delafontaine, 2013)

NGS PROTOCOLS

CRITICAL STEPS BEFORE LIBRARY PREP

What Bioinformaticians need to understand …. Is much more !!!!

- What species you are working on and what are the characteristics of this/ these species
- The short-term and long-term goal in performing the experiment
- Experiments details ("single-end" or "paired-end")
- Statistical considerations (replicates…)
- …

Adapted from (Delafontaine, 2013)

NGS PROTOCOLS

CRITICAL STEPS BEFORE LIBRARY PREP

- Robust library preparation methods are of crucial importance.

- Nevertheless: clear that NGS libraries (all types of applications) contain biases that compromise the quality of NGS datasets and can lead to their erroneous interpretation.

- A detailed knowledge of each step in the protocol and of the nature of these biases will be essential for
  - a careful interpretation of NGS data
  - finding ways to improve library quality
  - developing bioinformatics tools to compensate for the biases

Important for both biologists, statisticians and bioinformaticians !

NGS PROTOCOLS

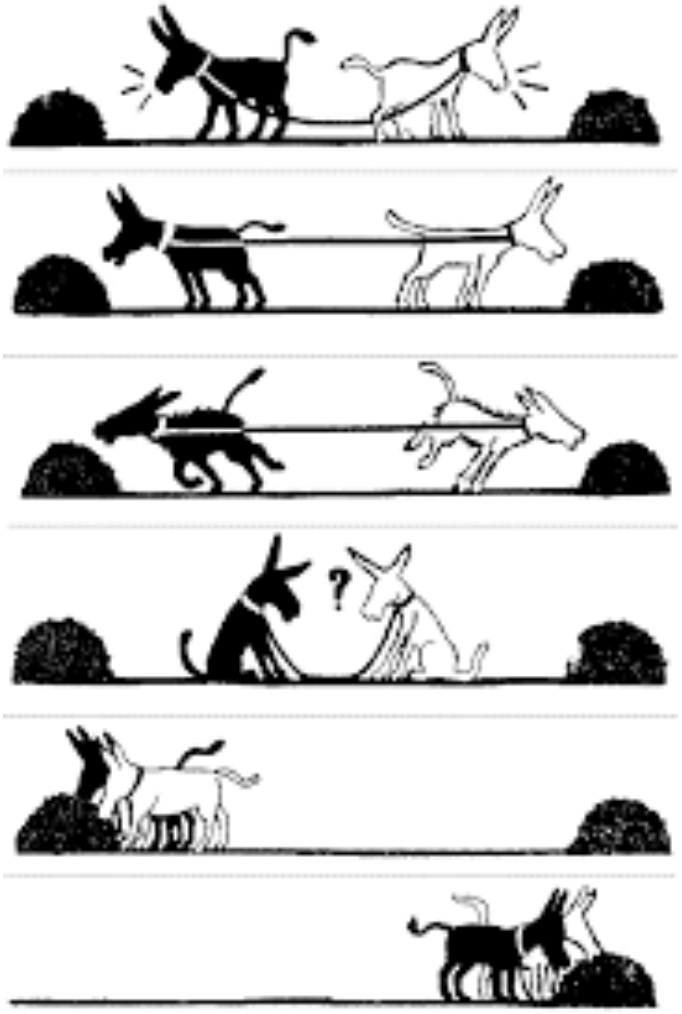
CRITICAL STEPS BEFORE LIBRARY PREP

Need to Collaborate… But also to understand each other !!!!

The Biologist need to know how the experiments were performed and what is his exact biological question

The Statistician need to know the details of your experiment and the question you are willing to answer

The Bioinformatician need to perform a "biologically relevant" analysis based on your experiment and question



PART 3

NGS PROTOCOLS

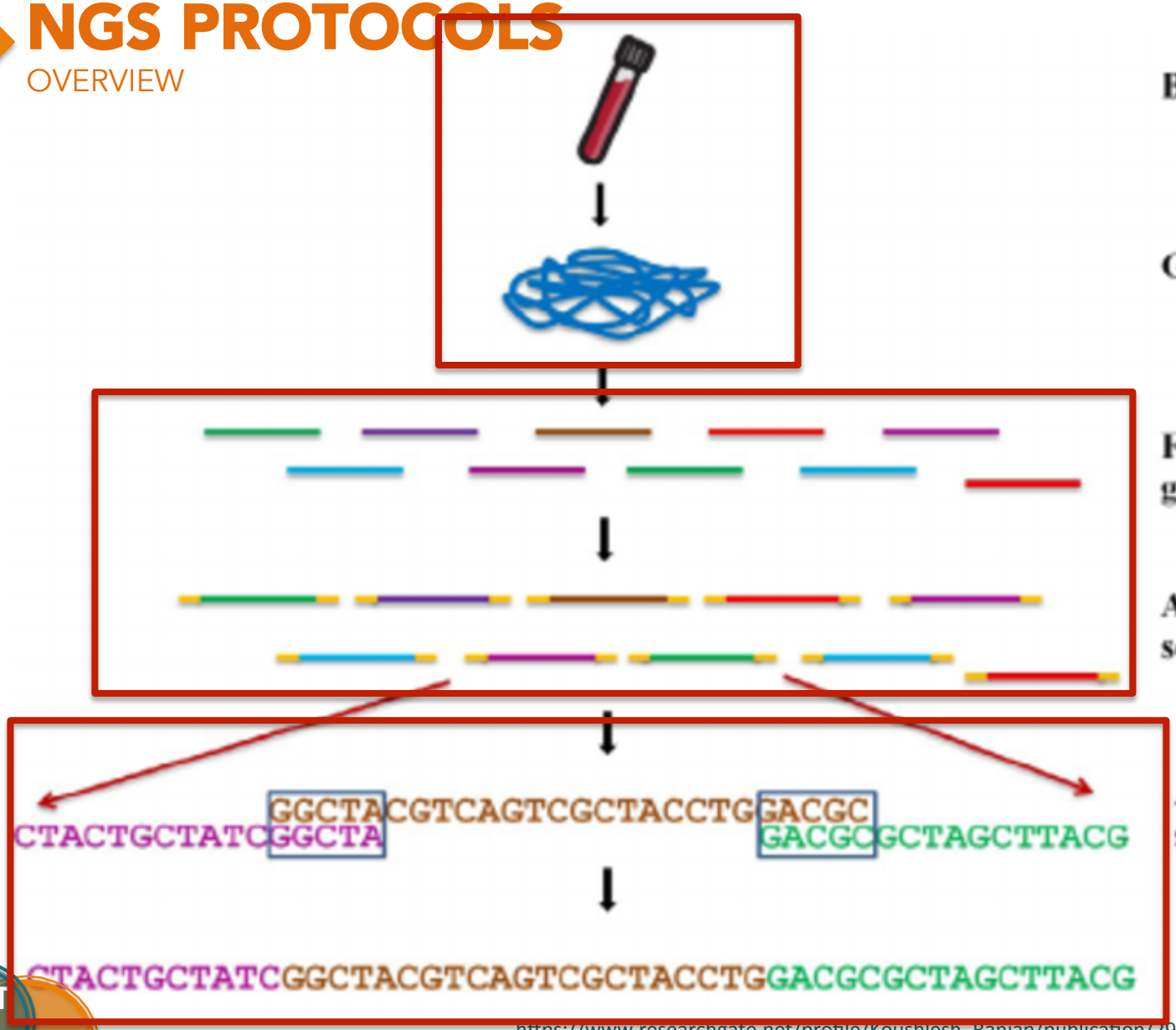
CRITICAL STEPS BEFORE LIBRARY PREP

...At least don't be this kind of biologist !

PART 3

NGS PROTOCOLS

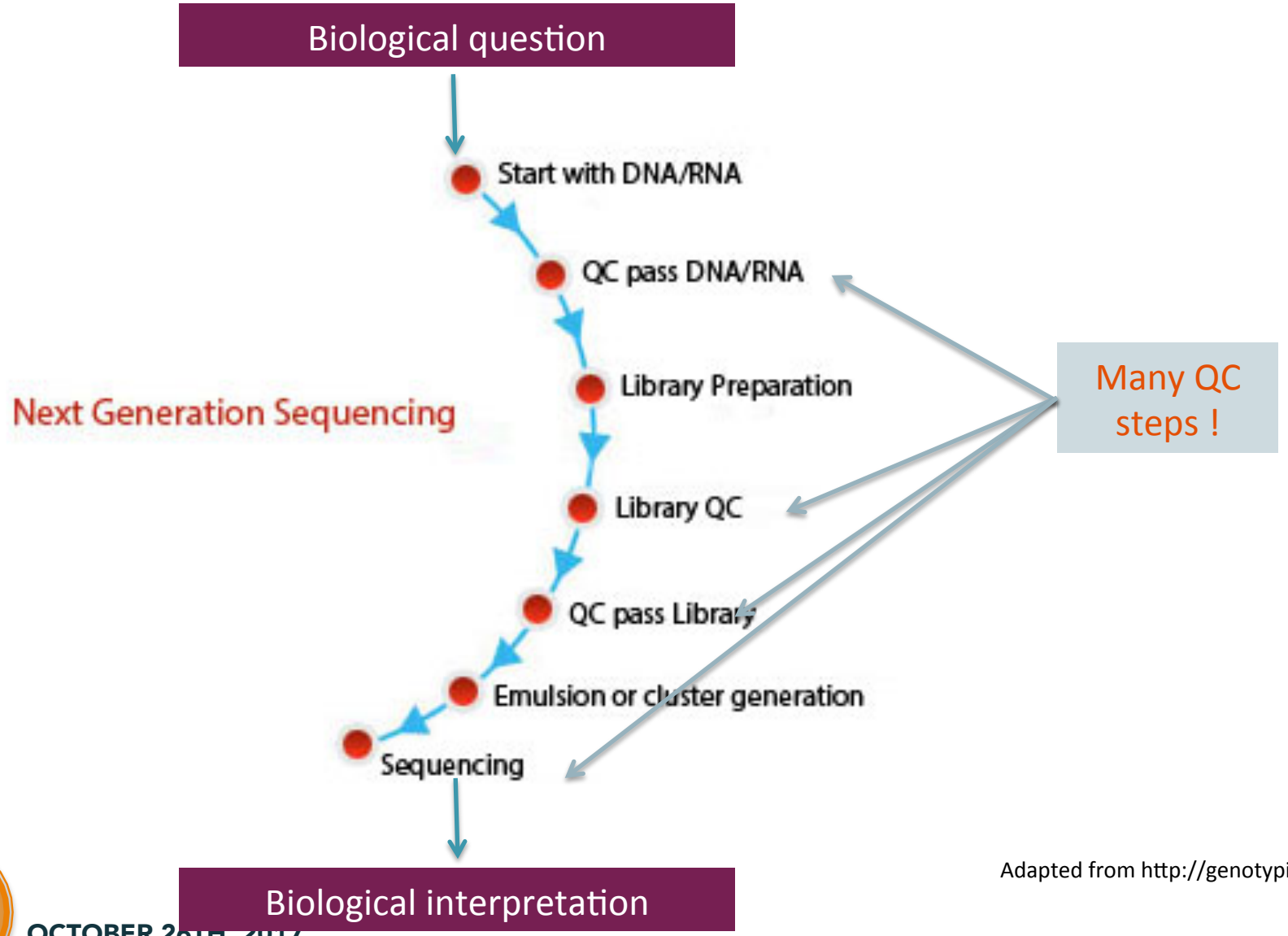
OVERVIEW



https://www.researchgate.net/profile/Koushiesh_Ranjan/publication/291171327/figure/fig4/AS:319876653699075@1453275945801/Figure-5-Basic-principle-of-next-generation-sequencing-technologies.png

NGS PROTOCOLS

OVERVIEW



Adapted from http://genotypic.co.in

Library Preparation (DNA-Seq / RNA-Seq)

Overview

NGS PROTOCOLS

DNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 01 Genomic DNA Purification

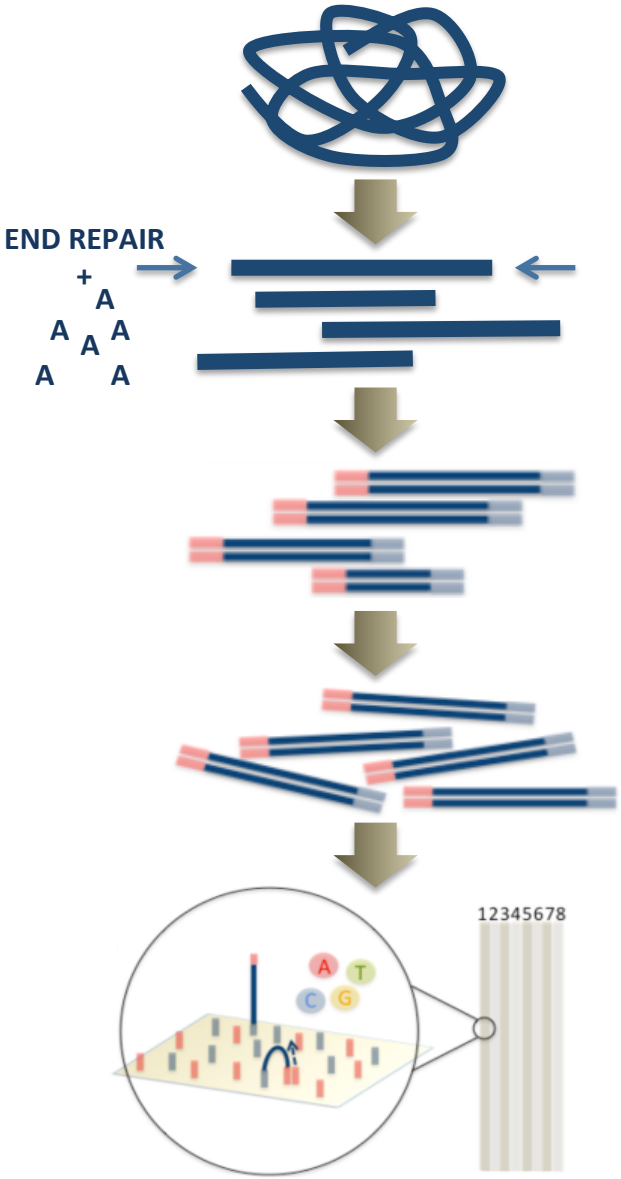
STEP 02 Genomic DNA Fragmentation

STEP 03 End repair and A-tailing

STEP 04 Adapter Ligation

STEP 05 Size Selection & PCR

STEP 06 Sequencing



NGS PROTOCOLS

DNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

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)

BIOO SCIENTIFIC

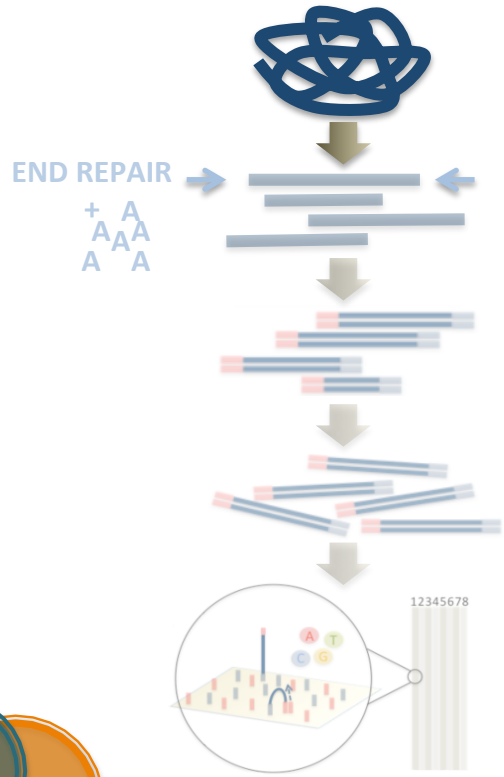
KAPABIOSYSTEMS evolving better science

(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/)

NGS PROTOCOLS

DNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 01 GENOMIC DNA PURIFICATION

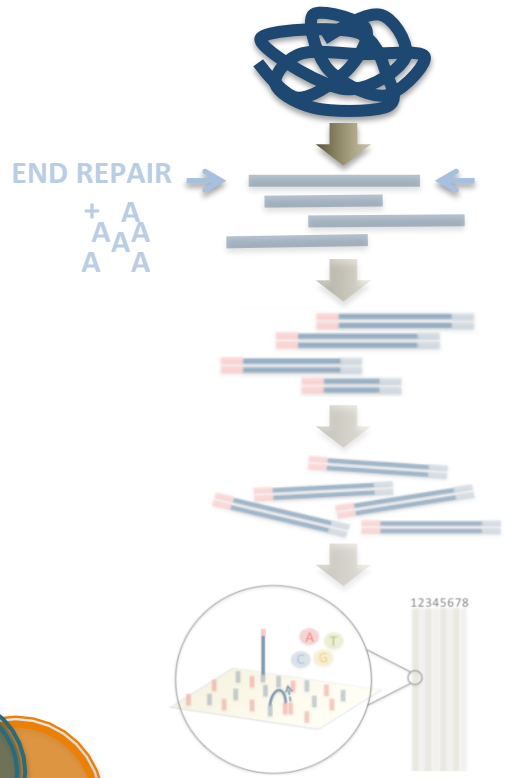


- **Starting material: QC**
  - ***Quality Control***
    - gel visualization, Bioanalyzer (Agilent, Bio-rad)
  - ***Quantity Control***
    - Nanodrop, Qubit…

NGS PROTOCOLS

DNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 01 GENOMIC DNA PURIFICATION

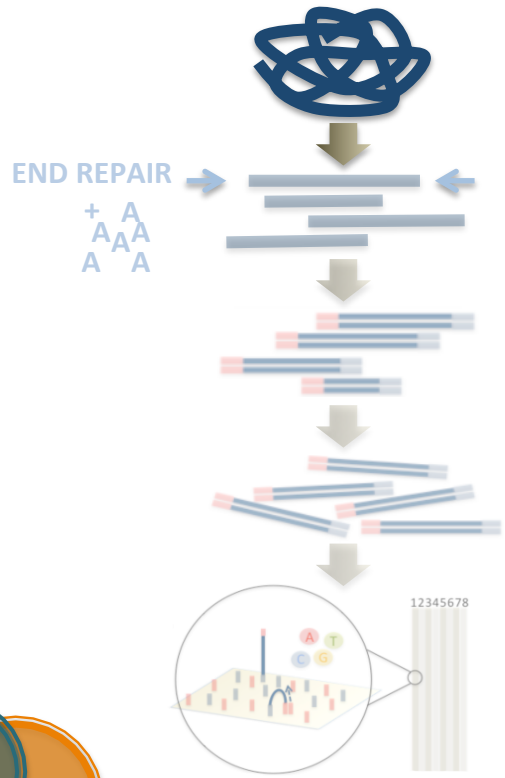


- **Experimental design**
  - SR (single read) or PE (paired-end)
  - Multiplexing or not
  - *de novo* or not
  - Statistics…

NGS PROTOCOLS

DNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

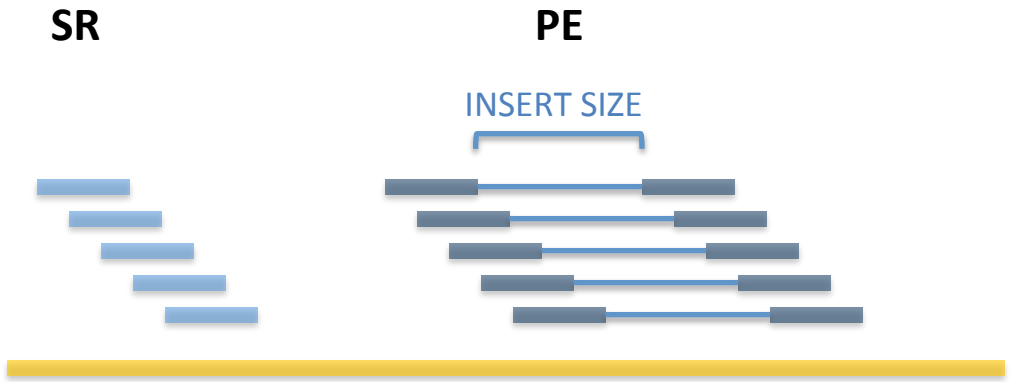
STEP 01 GENOMIC DNA PURIFICATION



- **Experimental design**

  - SR (single-end reads) or PE (paired-end reads)

PE involves sequencing both ends of the DNA fragments and aligning the forward and reverse reads as read pairs

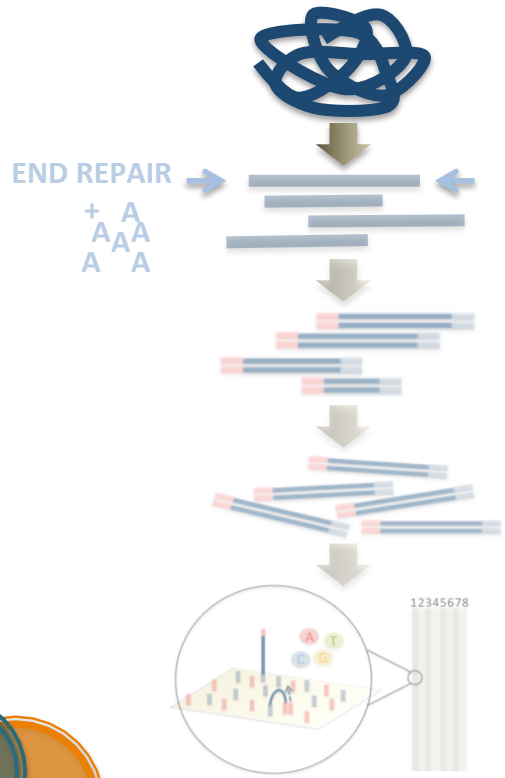


Adapted from https://www.biostars.org/p/162806/

NGS PROTOCOLS

DNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 01 GENOMIC DNA PURIFICATION

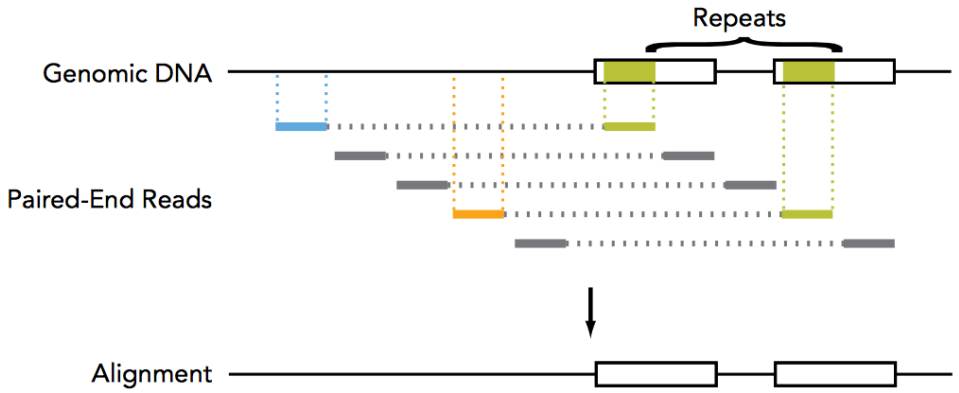


Experimental design

- SR (single-end reads) or PE (paired-end reads)

Advantage of PE:

- more accurate read alignment
- Unambiguous mapping of repeats



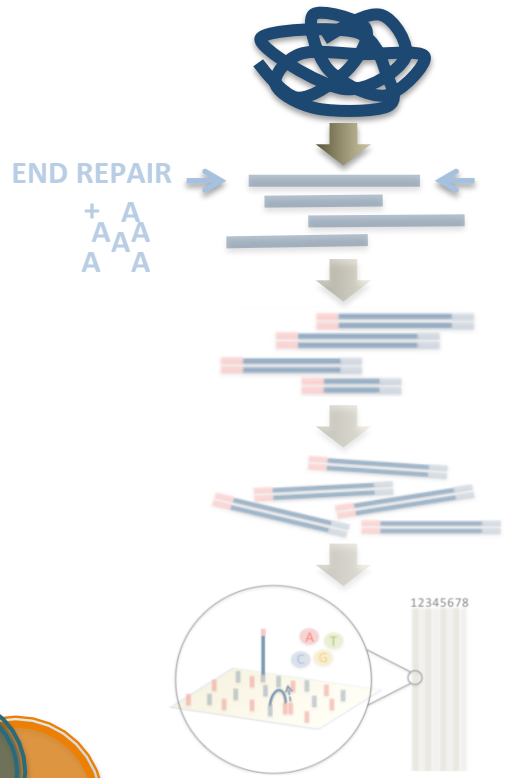
Reads in repeats (green) can be unambiguously aligned in complex genomes. Each read is associated with a paired read (blue or orange) and the separation between read pairs is known from the fragment size of the input DNA.

http://www.illumina.com/

NGS PROTOCOLS

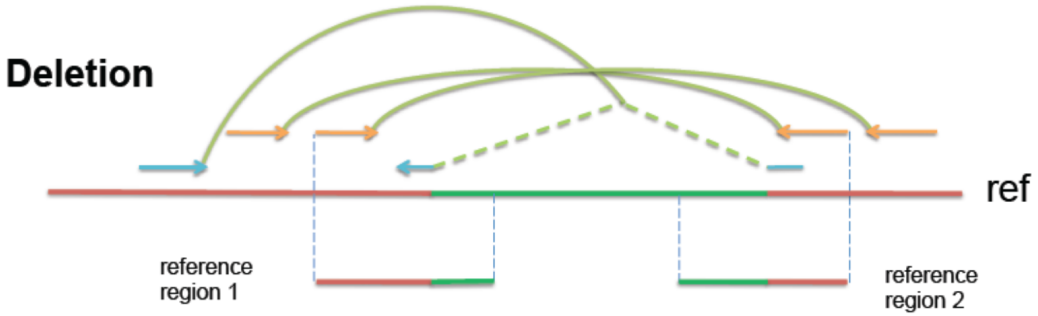
DNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 01 GENOMIC DNA PURIFICATION



- **Experimental design**

  - SR (single-end reads) or PE (paired-end reads)
  - Detection of even small deletions
  - Estimation of InDels sizes
  - allows removal of PCR duplicates (common artifact resulting from PCR amplification during library preparation: via Analysis of differential read-pair spacing)



NGS PROTOCOLS

DNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 01 GENOMIC DNA PURIFICATION

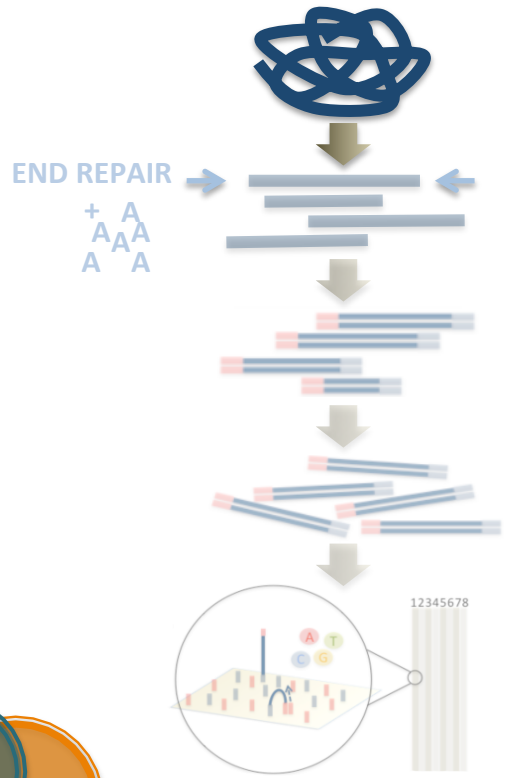


Table 1: Flexible Paired Sequencing Provides Optimal Detection Of Any Variant

Variant	Single Read	Short Insert Paired-Ends (200–500 bp)	Long Insert Mate Pairs (2–5 kb)	Paired-End And Mate Pair Combined
SNP	++	++++	++	++++
Small indels	++	++++	++	++++
Insertion	+	+++	+++	++++
Amplification	++	+++	+++	++++
Deletion	+	+++	++	++++
Inversion	+	+++	++	++++
Complex rearrangement	+	+++	++	++++
Large rearrangement	+	++	+++	++++

Only by combing short and long inserts can researchers be certain to find all different sizes and types of variants. In particular, short inserts are essential to identifying small indels and mate pairs are essential for identifying the largest rearrangements.

NGS PROTOCOLS

DNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

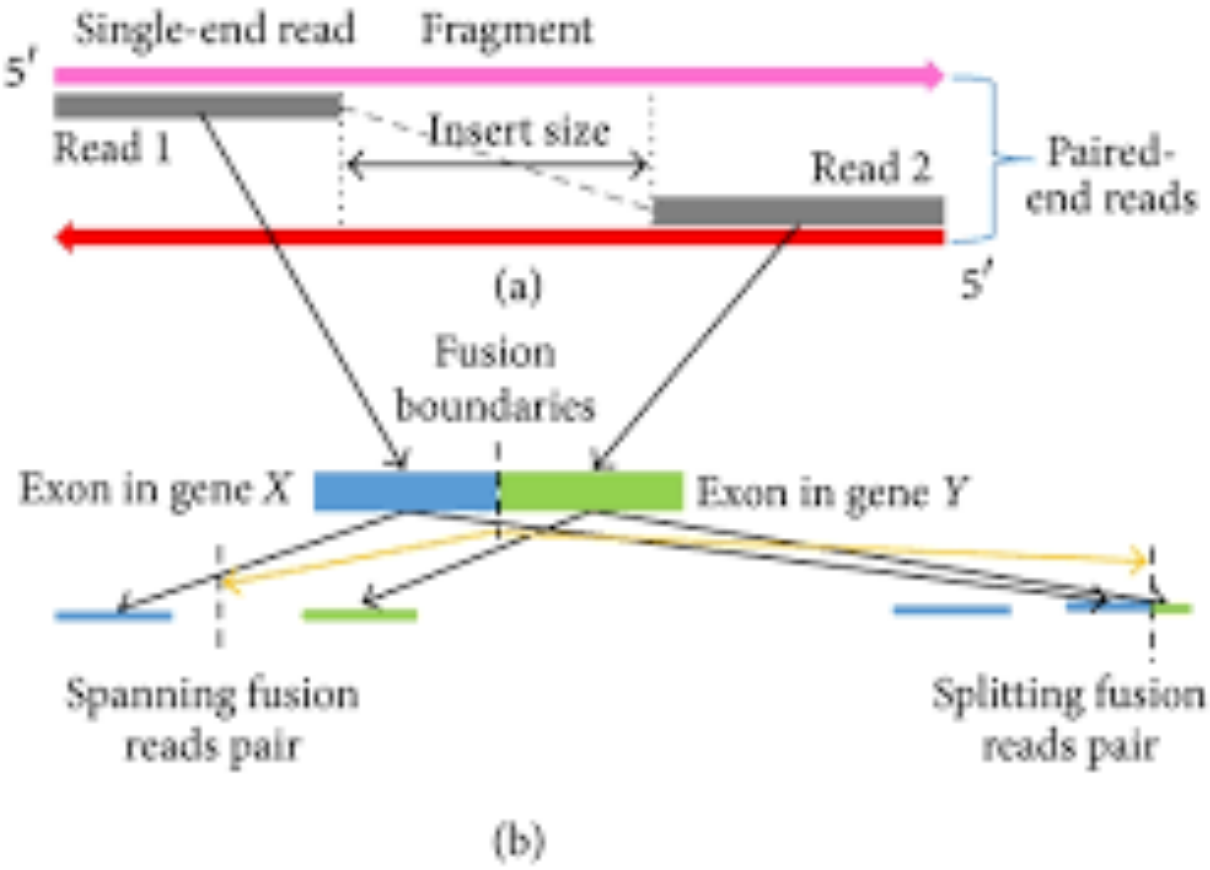
Figure 5. *De Novo* Assembly with Mate Pairs

Using a combination of short and long insert sizes with paired-end sequencing results in maximal coverage of the genome for de novo assembly. Because larger inserts can pair reads across greater distances, they provide a better ability to read through highly repetitive sequences and regions where large structural rearrangements have occurred. Shorter inserts sequenced at higher depths can fill in gaps missed by larger inserts sequenced at lower depths. Thus a diverse library of short and long inserts results in better de novo assembly, leading to fewer gaps, larger contigs, and greater accuracy of the final consensus sequence.

SEQ TECHNOLOGIES

NEXT-GENERATION SEQUENCING

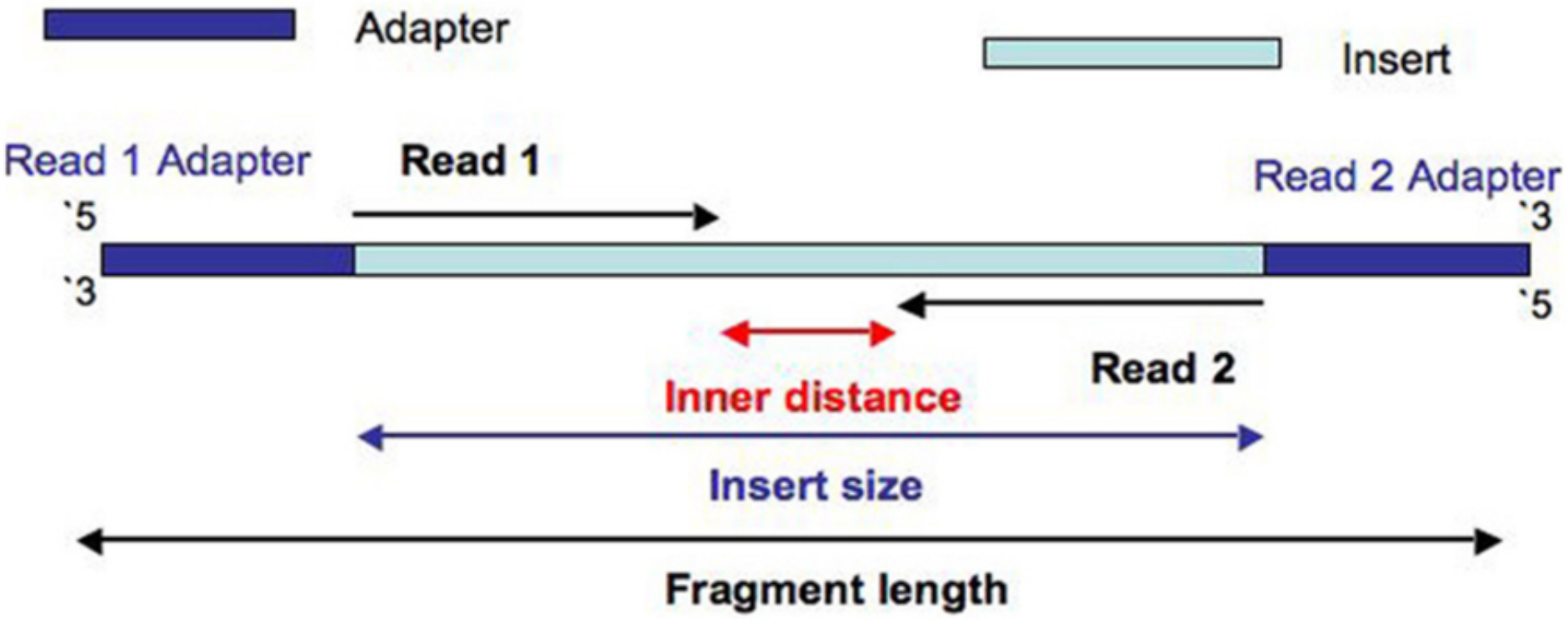
- Lexical considerations



SEQ TECHNOLOGIES

NEXT-GENERATION SEQUENCING

- Lexical considerations

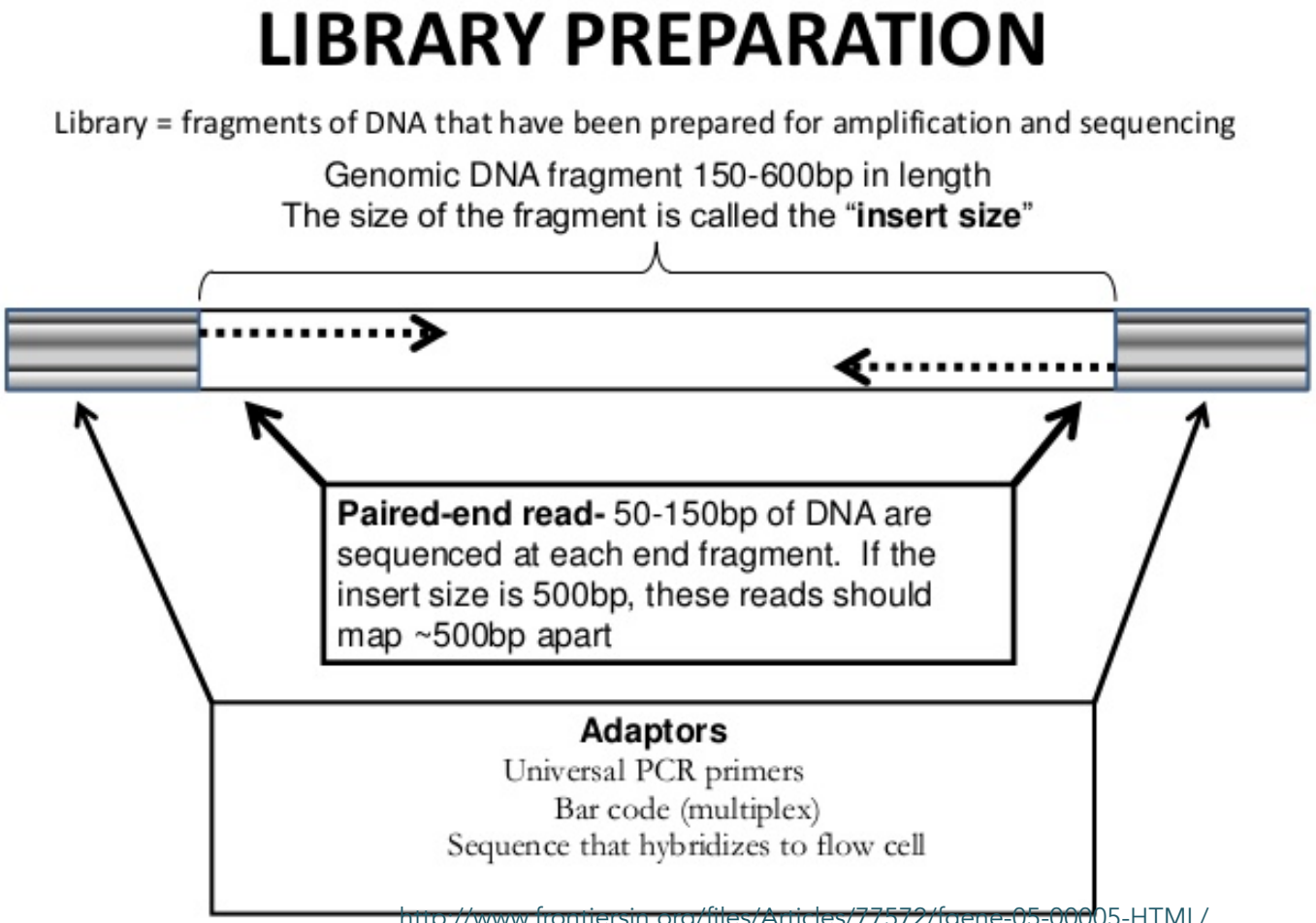


http://www.frontiersin.org/files/Articles/77572/fgene-05-00005-HTML/image_m/fgene-05-00005-g001.jpg

SEQ TECHNOLOGIES

NEXT-GENERATION SEQUENCING

- Lexical considerations

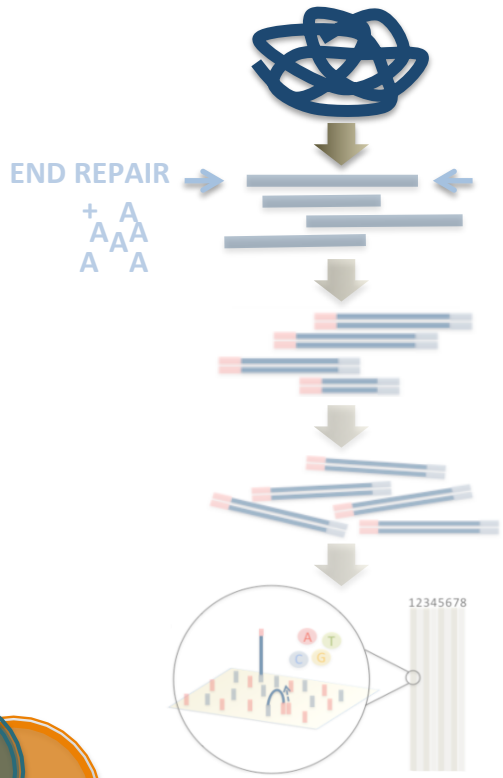


http://www.frontiersin.org/files/Articles/77572/fgene-05-00005-HTML/image_m/fgene-05-00005-g001.jpg

NGS PROTOCOLS

DNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 01 GENOMIC DNA PURIFICATION



Experimental design

- Multiplexing or not ?

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

- Libraries pooled and sequenced in parallel.

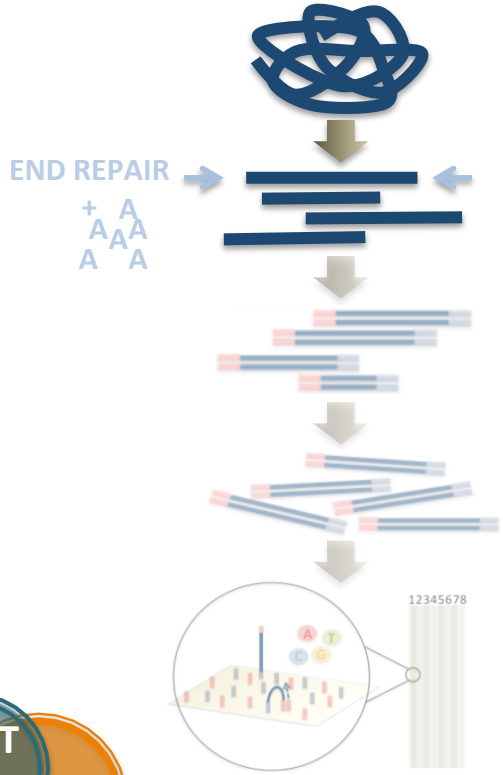
- Reads from each library are differenciated by using barcode to de-multiplex

- Each set is aligned to the reference genome

NGS PROTOCOLS

DNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 02 GENOMIC DNA FRAGMENTATION

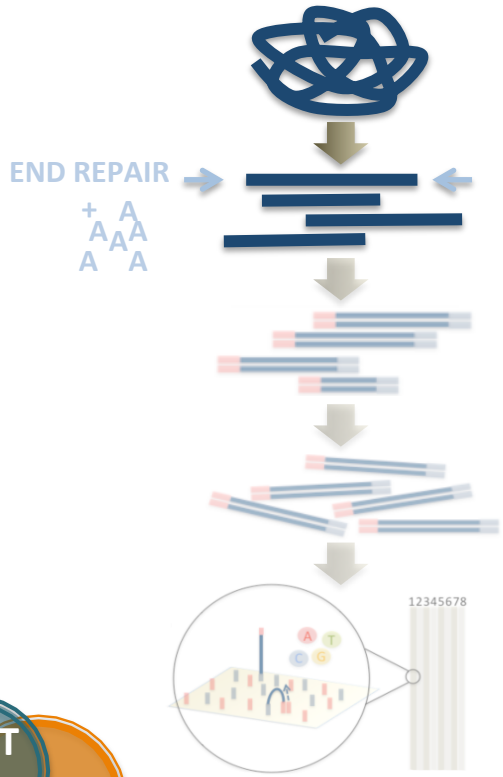


- **Fragmentation**
  - ***Can be included in the kit***
    - Optimization of fragmentation parameters
  - ***Several methods***
    - Enzymatic, Nebulization, acoustic shearing…

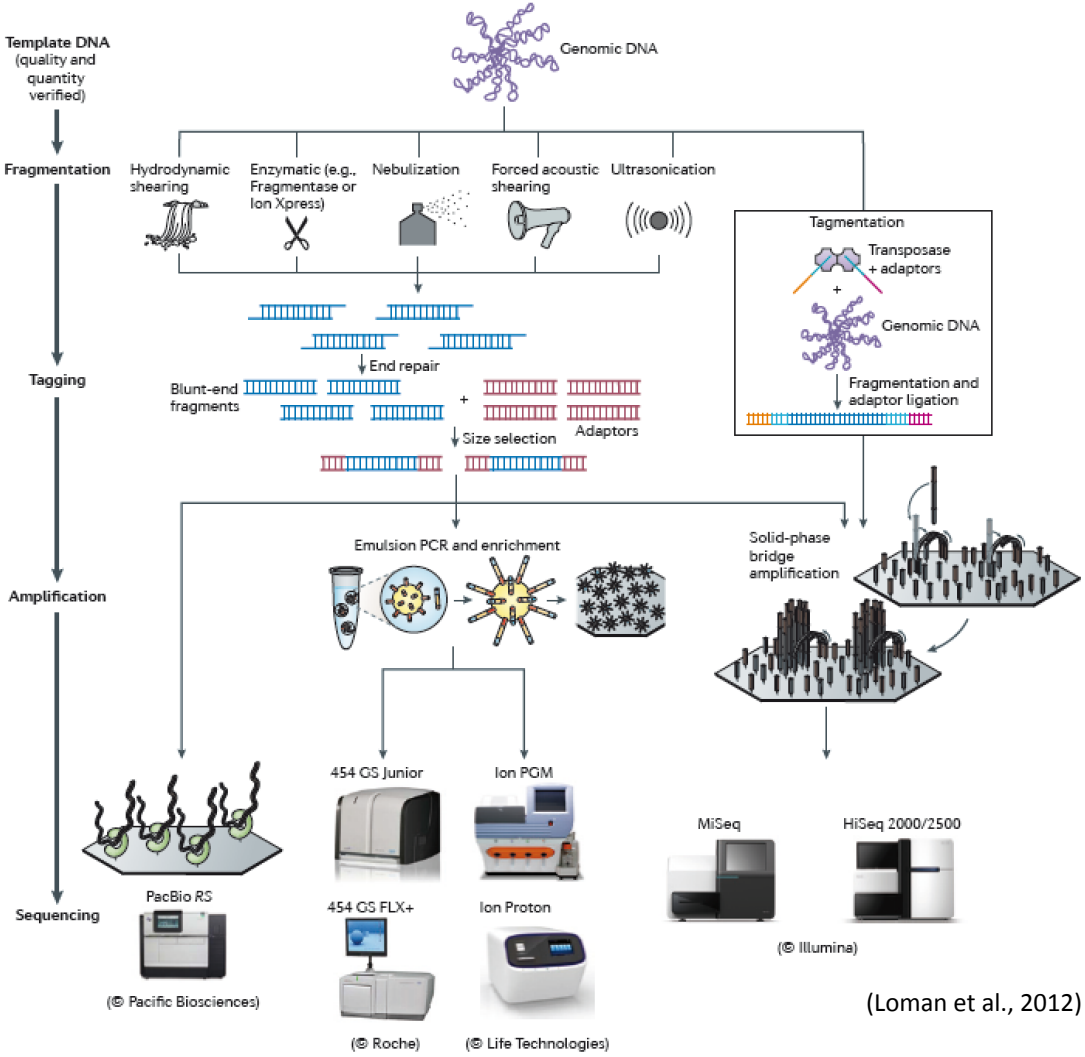
NGS PROTOCOLS

DNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 02 GENOMIC DNA FRAGMENTATION



Different fragmentation procedures according to High-throughput sequencing platforms.



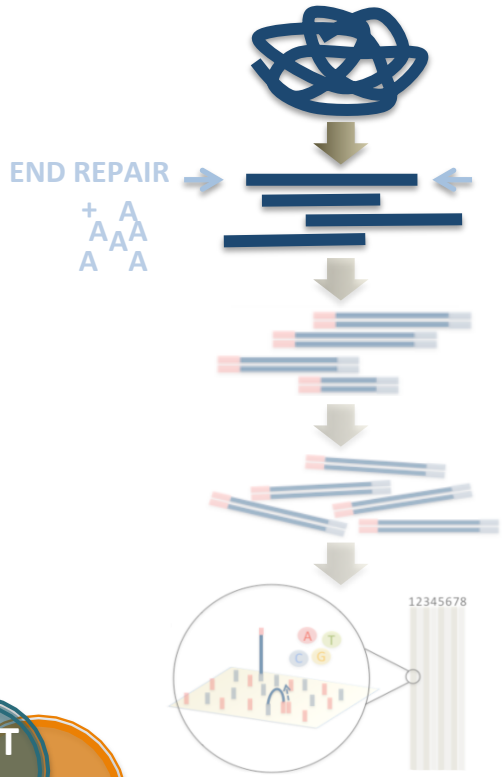
(Loman et al., 2012)

PART 3 A

OCTOBER 26TH, 2017

IPT COURSE, TUNIS, TUNISIA

STEP 02 GENOMIC DNA FRAGMENTATION



Starting material: input

- ***Low vs High Quality DNA***
- Caution in size selection

Low Quality DNA

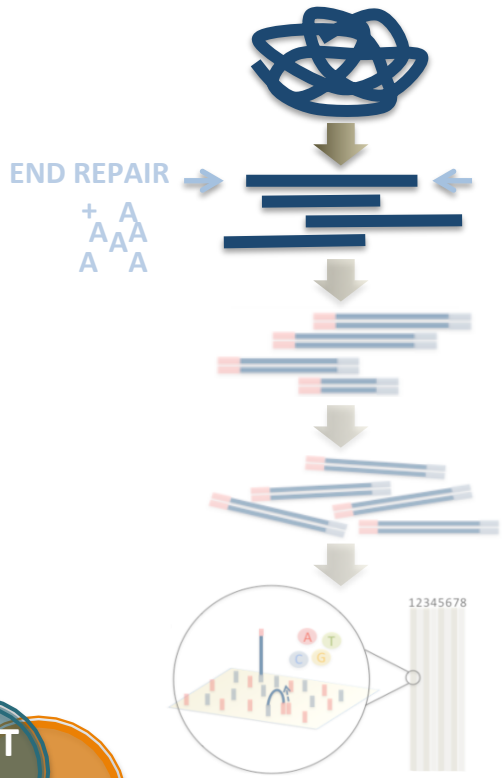
Nucleic acids can be highly degraded and fragmented, a consequence of the nature of preservation.

If starting with **sub-nanogram quantities** of **low quality** DNA: **size selection not advised** (limited number of amplifiable DNA molecules going into PCR = result in greatly reduced library yield).

NGS PROTOCOLS

DNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 02 GENOMIC DNA FRAGMENTATION



- **Starting material: input**

  - ***Low vs High Quality DNA***
  - Caution in size selection

Sufficient Quantity of High Quality DNA

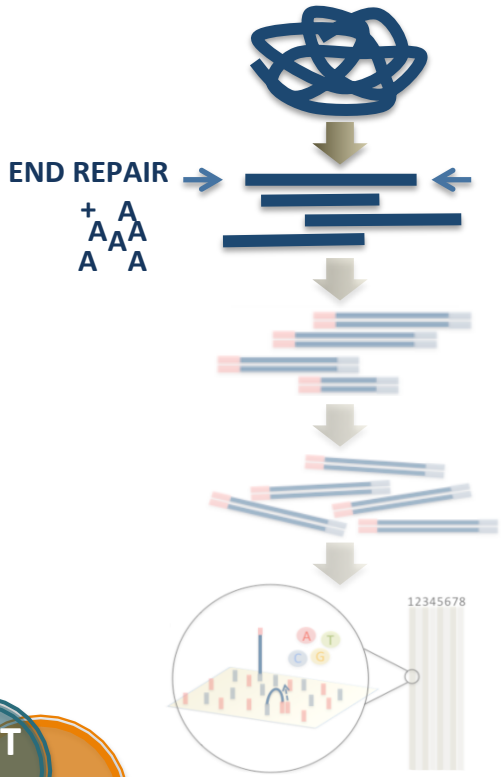
Size selecting a specific region of a broad range shear is advisable if starting with **≥ 10 ng of DNA**.
→ if DNA is not a limiting factor + many barcoded samples are being processed in parallel → Highly recommended

Why? Serves as an internal control to ensure each library gets similar reads or coverage

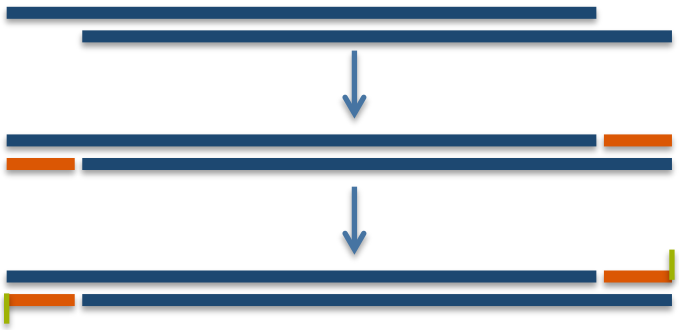
NGS PROTOCOLS

DNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 03 END REPAIR AND A-TAILING



- **Repair ends**
  - Converts overhangs:
    Blunt ends + Phosphorylates 5'-end
  - Reagents:
    dNTP, T4 DNApol, Klenow - Kinase/ATP (T4 PNK)
  - Simple enzymatic reaction



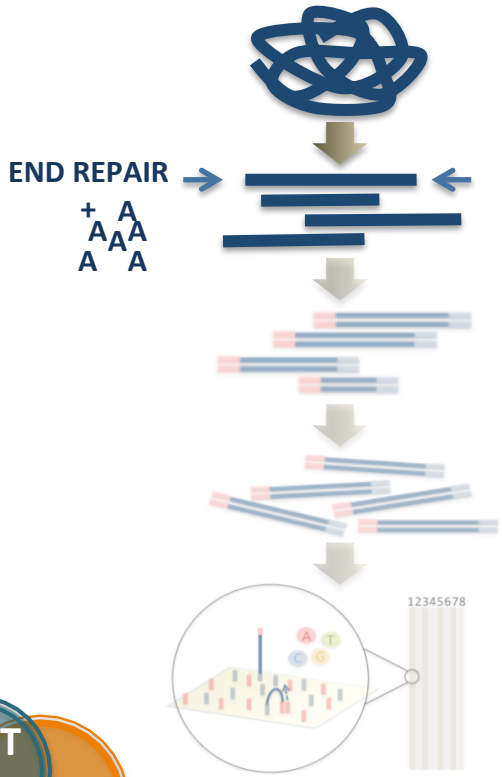
BLUNT ENDING BY EXONUCLEASE

5'-END PHOSPHORYLATION

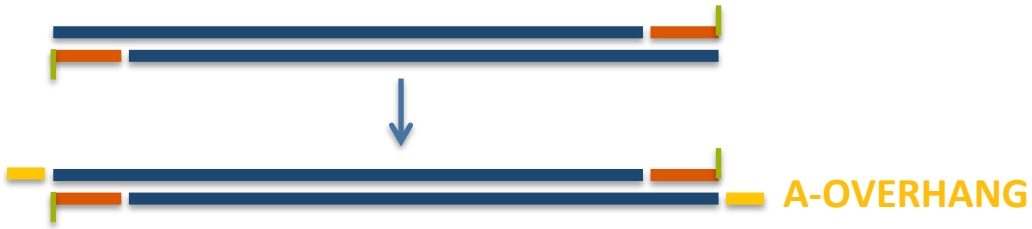
NGS PROTOCOLS

DNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 03 — END REPAIR AND A-TAILING



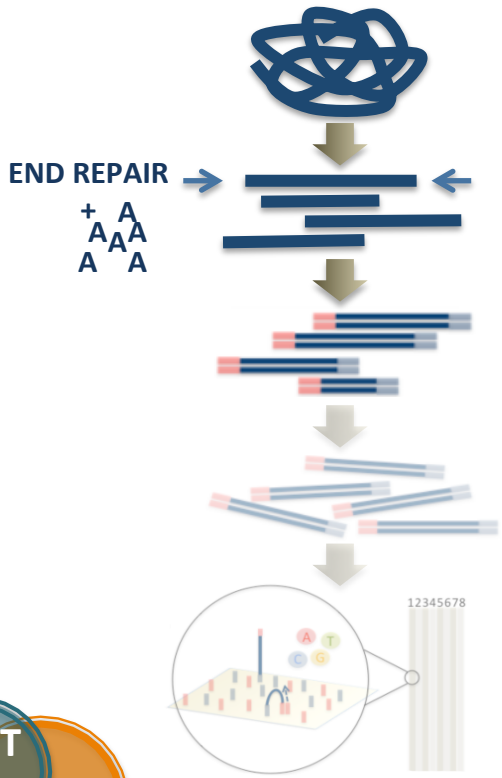
- **A-tailing (Adenylation)**
  - Adds ‘A’ base to the 3' end of the blunt phosphorylated DNA fragments
  - Prevents
    - Formation of adapters dimers
    - Concatemers
  - Reagents
    1 mM dATP, Klenow exo (3' to 5' exo minus)



NGS PROTOCOLS

DNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 04 ADAPTER LIGATION



Adapter Ligation

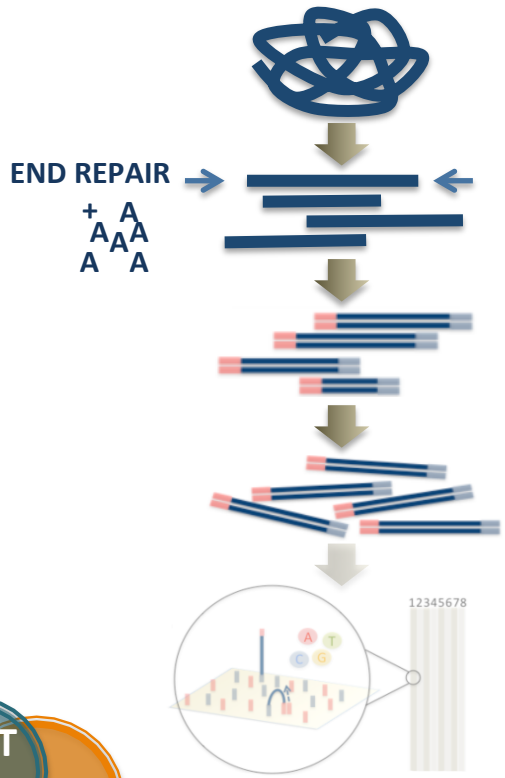
- Provided or custom-designed
- Adapter concentration affects ligation, adapter and adapter-dimer carryover
- Robust Ligation efficiency for adapter:insert molar ratios between 10:1 and >200:1
- Adapter ratio >200:1 for low-input applications.
- Adapter quality
- Post-Ligation cleanup

https://www.kapabiosystems.com/

NGS PROTOCOLS

DNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 05 SIZE SELECTION AND PCR



- **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

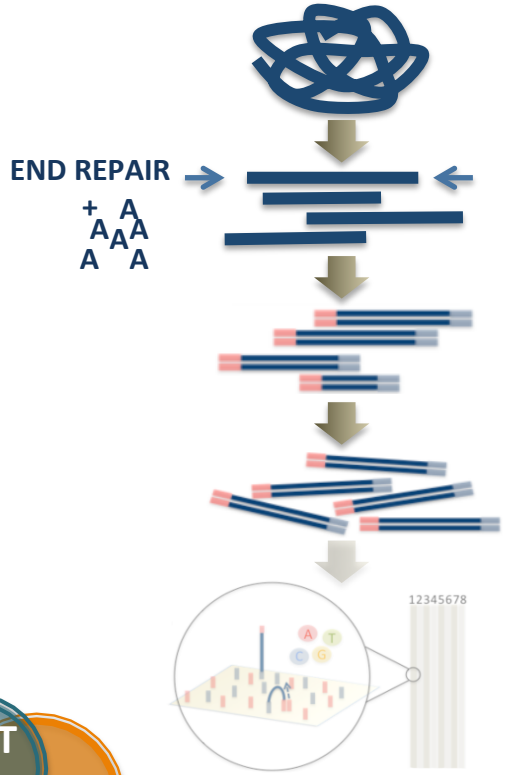
= ensures maximum and uniform coverage

https://www.kapabiosystems.com/

NGS PROTOCOLS

DNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 05 SIZE SELECTION AND PCR



Size selection: Read length considerations

Strategies to focus read lengths during sample and library preparation:

- *Proper parameters for shearing genomic DNA important*
- *but all have limitations in obtaining tight focused bands.*

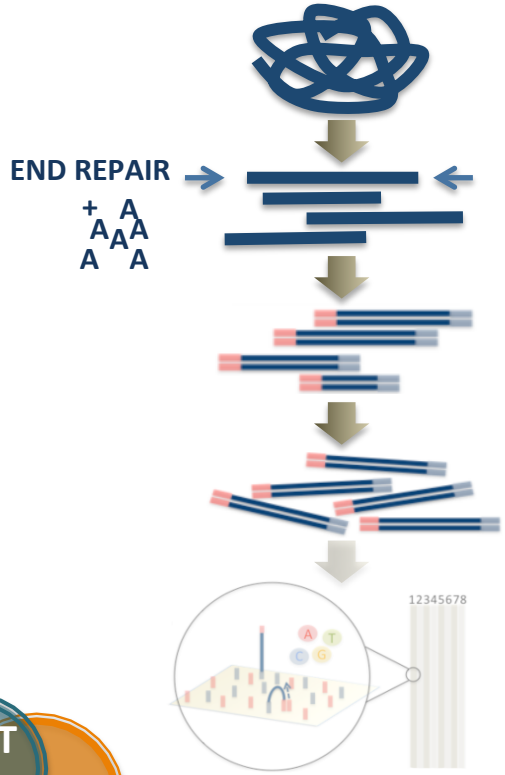
- Double solid-phase reverse immobilization (SPRI) selection methods allow for reshaping the input fragment distribution into well-defined ranges.

https://www.kapabiosystems.com/

NGS PROTOCOLS

DNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 05 SIZE SELECTION AND PCR



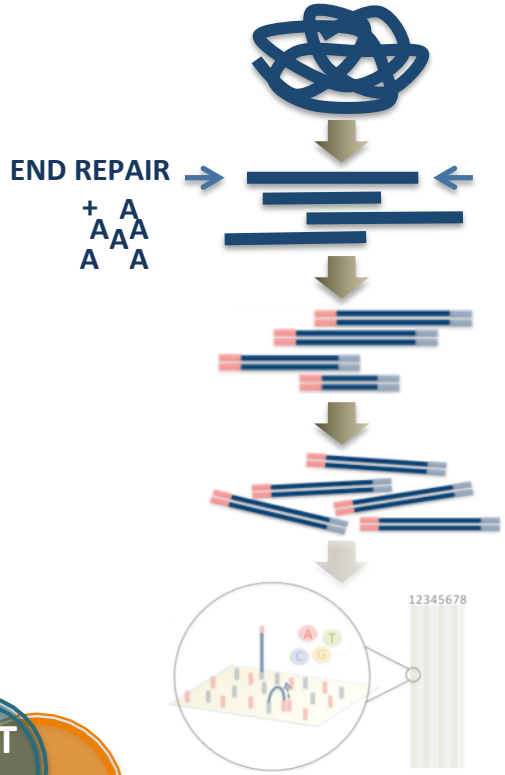
- **Size selection: Read length considerations**
  - SPRI beads are paramagnetic (magnetic only in a magnetic field) and this prevents them from clumping and falling out of solution.
  - *Popular and present in many kits because:*
    - *The binding capacity of SPRI beads is huge. 1µl of AmpureXP will bind over 3µg DNA.*
    - *SPRI is great for low concentration DNA cleanup*

https://www.kapabiosystems.com/

NGS PROTOCOLS

DNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 05 SIZE SELECTION AND PCR



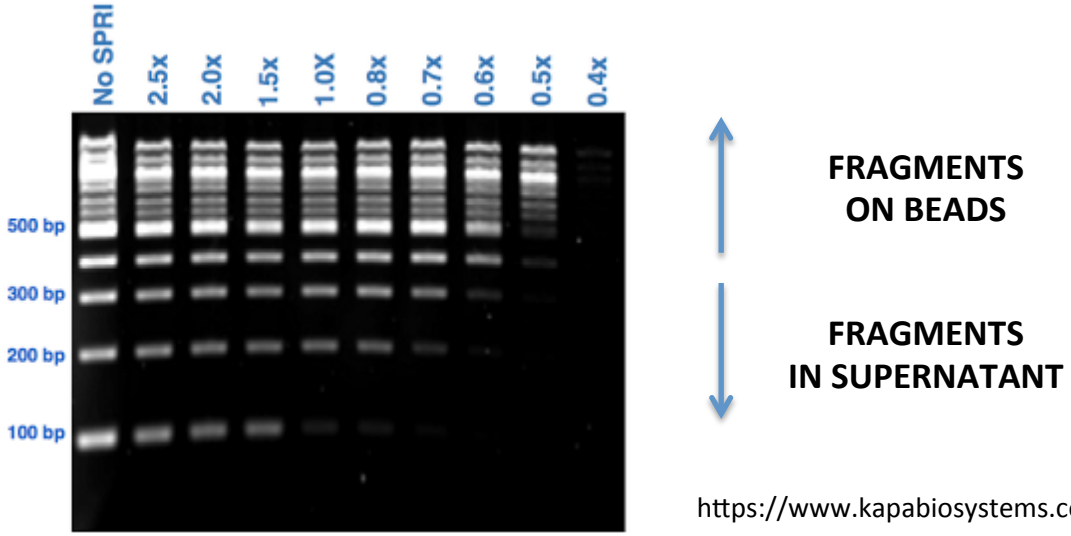
Size selection: Read length considerations

- SPRI + Reverse-SPRI

By using a combination of shearing using SPRI + Rev. SPRI , one can detect a quite tight size range with no gel.

The "X" of SPRI refers to the volume ratio of SPRI to DNA

- 1X SPRI is a 1:1 vol of SPRI:DNA

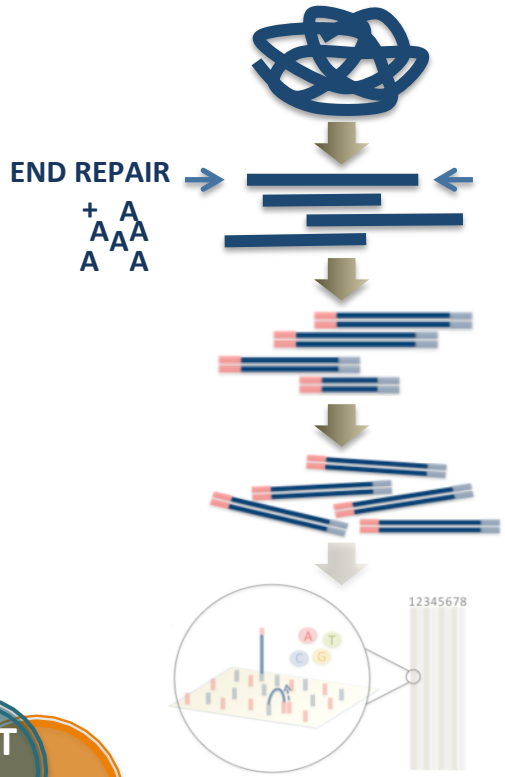


https://www.kapabiosystems.com/

NGS PROTOCOLS

DNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 05 SIZE SELECTION AND PCR



- **Library Amplification (PCR)**
  - Amplifies the amount of DNA in the library
  - Selectively enriches DNA fragments with adapter molecules on both ends
  - Post-amplification cleanup
  - PCR-free kits

  → reduce bias: ***PCR-free library preparation kits*** (Illumina) (Korazewa et al., 2009): proposed to ligate adapters that contain all necessary elements for bridge amplification on Illumina flow cells, eliminating the need for PCR.

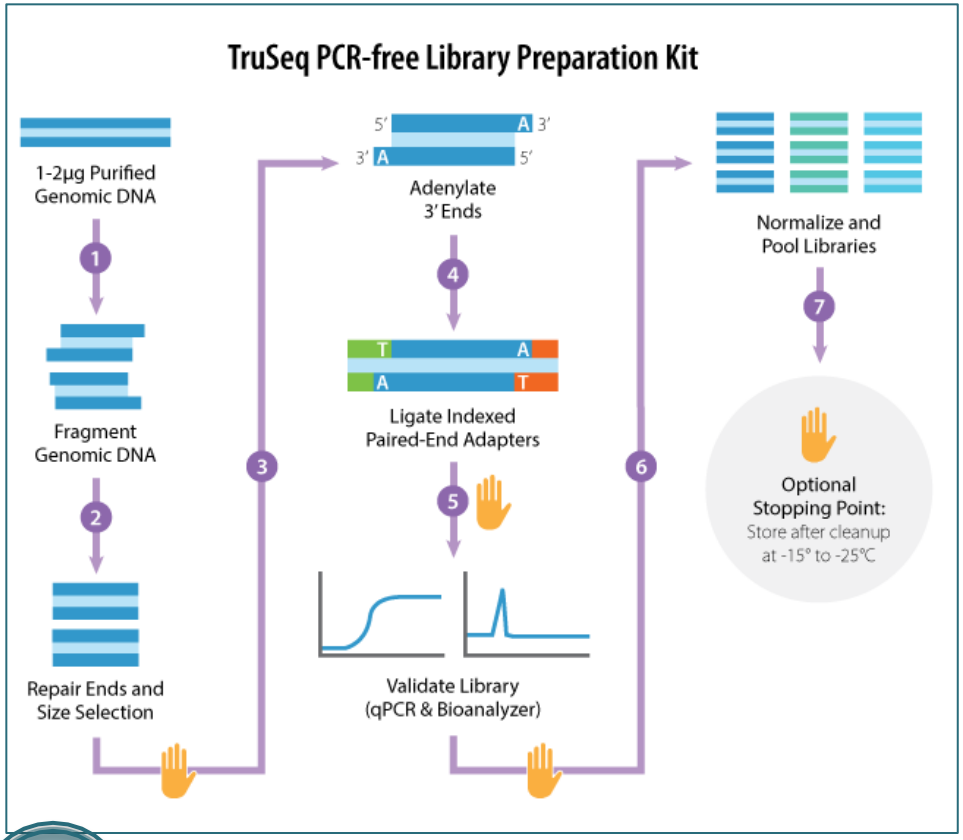
- **QC**
  - Quality & Quantity & size check

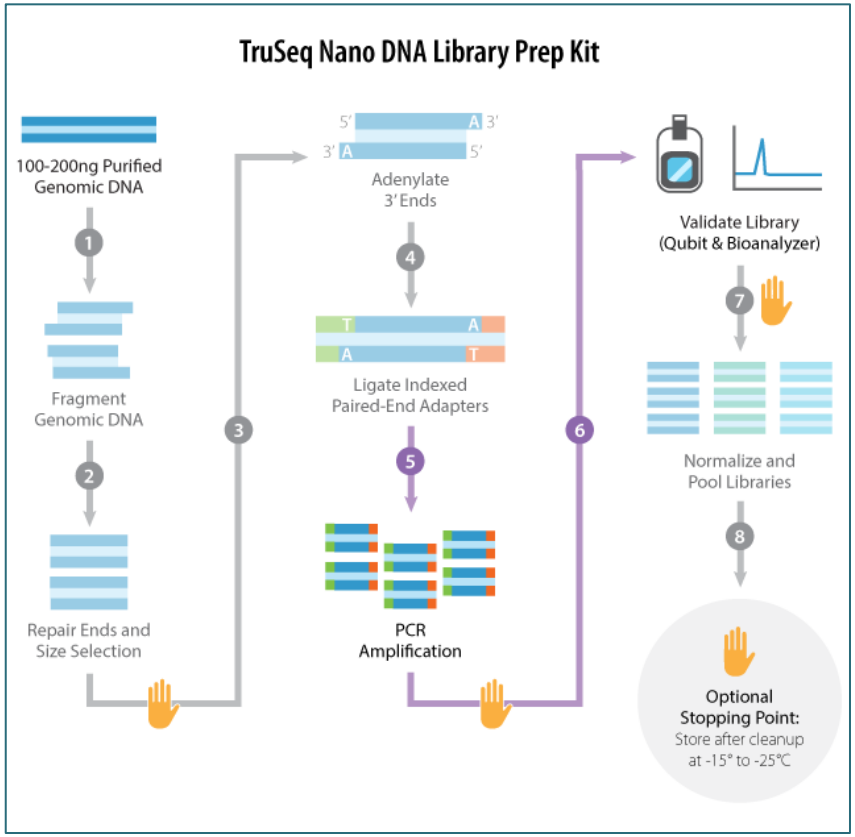
https://www.kapabiosystems.com/

NGS PROTOCOLS

DNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

Continuous improvement in kit design

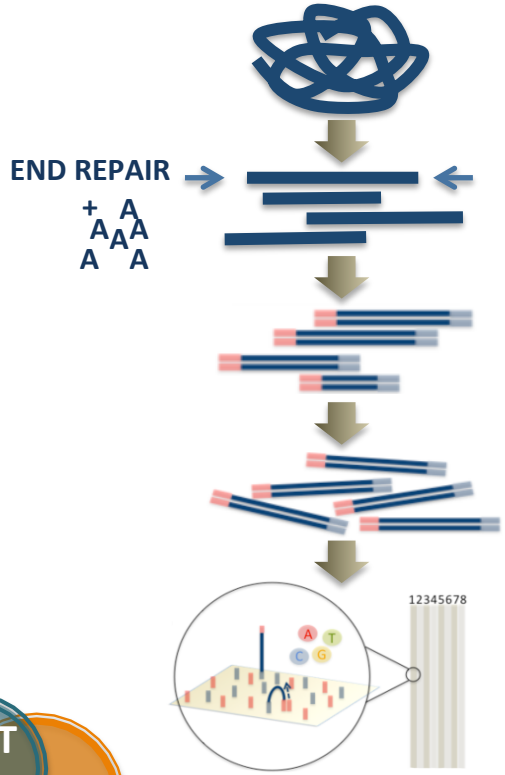




NGS PROTOCOLS

DNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 06 SEQUENCING



- **DNA Sequencing**

  - input : Library constructed
    - Whole-genome
    - Whole-exome
    - Target region
    - …

  - Cluster amplification + sequencing + base calling

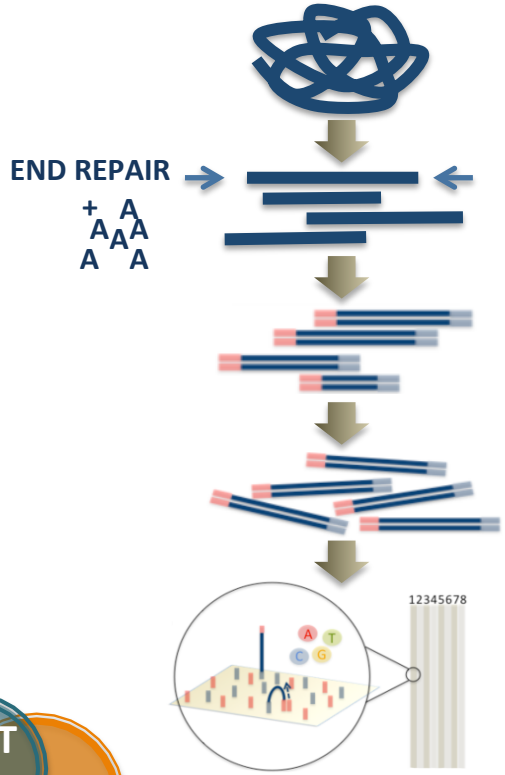
  - QC (run report)

https://www.kapabiosystems.com/

NGS PROTOCOLS

DNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 06 SEQUENCING



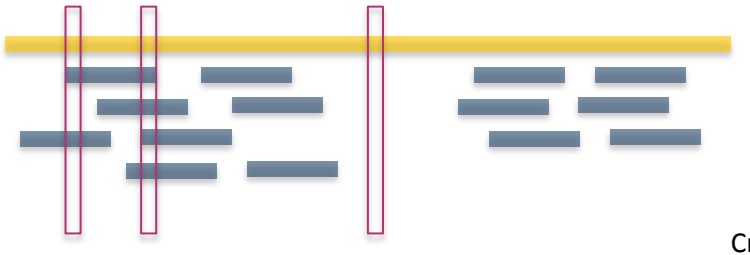
DNA Sequencing

- output : sequenced « reads » (fastq files)

NB: the number of sequenced reads can be approximated knowing the overall genome size and the coverage required

- Coverage

The number of times each nucleotide is « read »
→ Fold Coverage (number + X)



Cronn & al., 2012
https://www.kapabiosystems.com/

NGS PROTOCOLS

DNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 06 **SEQUENCING**

Coverage

- By increasing or decreasing the number of sequencing reads, researchers can tune the sensitivity of an experiment to accommodate various study objectives.

- Sequencing runs can be tailored to zoom in with high resolution on particular regions of the genome, or provide a more expansive view with lower resolution. This offers several experimental design advantages. Examples:

- Detection of low frequency mutations within a mixed cell population: somatic mutations may only exist within a small proportion of cells in a given tissue sample

→ region of DNA having the mutation must be sequenced at extremely high coverage, often >1000×

- Genome-wide variant discovery: study design involves sequencing many samples (hundreds to thousands) at lower coverage → allows to achieve greater statistical power within a given population.

- **Sequencing technologies all include error-prone steps**

Errors and biases may be introduced at any step of NGS

▸ **Shearing DNA**

DNA fragmentation is not trivial and is not a random process
DNA shearing may introduce AT or GC biases:

- acoustic shearing can produce C → A/G → T artifacts in CCG sequences

- sonication preferentially cleaves GC-rich fragments

(Costello et al., 2013)
(Poptsova et al., 2014)

NGS PROTOCOLS

DNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

- **Sequencing technologies all include error-prone steps**

Errors and biases may be introduced at any step of NGS

- **Adapter ligation**

- **Amplification procedure**

- templates with balanced GC-AT content are preferred, thus resulting in an underrepresentation of both GC-rich and AT-rich regions (Benjamini & Speed, 2012)

- differential denaturation and annealing kinetics of different target regions according to the GC content (Van Dijk & al., 2014)
→ Review on optimization (Van Dijk & al., 2014)

- **kits of different manufacturers show substantial differences** in terms of proportion of PCR duplicates, GC bias, target coverage, number of off-target reads… (Bodi & al., 2013)

(Seguin-Orlando et al., 2013)
(Poptsova et al., 2014)

- **Sequencing technologies all include error-prone steps**

▸ **Coverage bias**

Coverage most extremely biased towards GC-rich sequences:
- **Ion torrent:** AT-rich organellar DNA fragments underrepresented
- **PacBio:** slight unevenness of coverage and bias towards GC

Coverage biased toward AT-rich sequences:
- **SOLID:** AT bias

(Quail et al., 2012)

NGS PROTOCOLS

DNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

- **Sequencing technologies all include error-prone steps**

  - **Error Frequency (EF)**
  - **Error Type (ET)**

	EF			ET	
	10^{-1}	10^{-2}	10^{-3}	SNPs	Indels
Sanger capillary seq	X			X	
Illumina MiSeq & HiSeq			X	X	
454 GS/GS FLX		X			X
PacBio RS		X*			X
Ion Torrent		X**			X
SOLID		X ($2x10^{-2}$)***			

- *EF of CG deletions
  **EF of short deletions
  ***AT bias

(Ross et al. 2013)
(Fox et al. 2014)

Library Preparation (DNA-Seq / RNA-Seq)

Overview

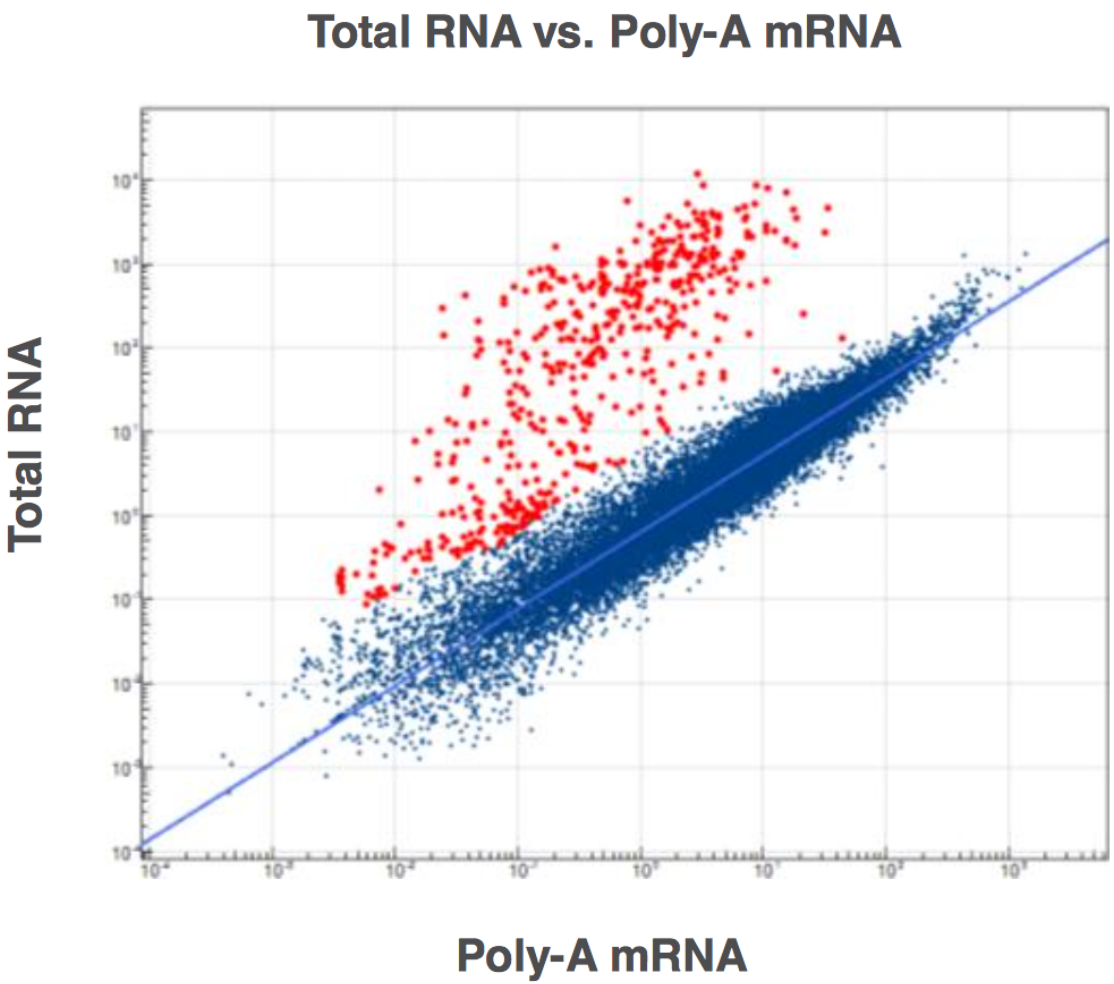
NGS PROTOCOLS

RNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

Method	Principle	Kit
Total RNA-seq	detects coding plus multiple forms of noncoding RNA (Species-specific rRNA probes (Ribo-Zero chemistry)	TruSeq Stranded Total RNA with Ribo-Zero Kits (whole transcriptome with precise strand orientation measurement)
mRNA-seq	•measurement of strand orientation •low quality/FFPE samples	•TruSeq Stranded mRNA Library Prep Kit •TruSeq RNA Access Library Prep Kit
Targeted RNA-seq	sequences specific transcripts of interest via either enrichment or amplicon-based approaches	•Fixed Panels: TruSeq Targeted RNA Expression Kits (Apoptosis, Cardiotoxicity, NFkB Pathways…) •Custom Panels: DesignStudio…
Small RNA-seq	generates small RNA libraries directly from total RNA. Supports customizable size selection (any small RNA: 17-35 nt).	TruSeq Small RNA Library Prep Kit
Single-cell RNA-seq	•single-cell isolation (process tens of thousands of single cells/day). •Transcriptome profiling (hundreds to tens of thousands of single cells/exp). •Synthesize full-length cDNA from only 1–1000 whole cells or 10 pg–10 ng high-quality total RNA.	•ddSEQ™ Single-Cell Isolator •Illumina Bio-Rad® SureCell™ WTA 3' Library Prep Kit for the ddSEQ™ System •SMART-Seq® Ultra® Low Input RNA Kit
Ribosome profiling	ribosome-protected mRNA fragments	ARTseq/TruSeq Ribo Profile Library Preparation Kit

GENOMES & TRANSCRIPTOMES

HGP: THE HERITAGE



Plot shows gene level count data

for all RefSeq Genes

illumina

NGS PROTOCOLS

RNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

Library Prep Kit Selector

Determine the best kit for your needs based on project type, starting material, and method or application.

Compare Total RNA-Seq Library Prep Kits

Determine the best kit for your needs.

Compare mRNA-Seq Library Prep Kits

Determine the best kit for your needs.

MiniSeq System
Targeted Power. Access proven Illumina quality with this small, robust NGS system. From 1-12 small RNA samples per run.

MiSeq Series
Focused power. Speed and simplicity for targeted and small genome sequencing. From 1-12 small RNA samples per run.

NextSeq Series
Flexible power. Speed and simplicity for everyday genomics. Up to 48 small RNA samples per run.

HiSeq Series
Production power. Max throughput and lowest cost for production-scale genomics. Up to 96 small RNA samples per run.

Platform Comparison Tool
Compare sequencing platforms and identify the best system for your lab and applications.

Sequencing Reagents
Find kits that include sequencing reagents, flow cells, and buffers tailored to each Illumina sequencing system.

NGS PROTOCOLS

RNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

How to prepare the library: critical steps

• Strikingly, **almost all steps of the various protocols have been reported to introduce biases**, especially in the case of RNASeq.

• **RNASeq protocols are technically more challenging than DNASeq** protocols and often include biased procedures.

• Common types of biases include :

- Library generation artifacts
- low complexity (many reads with the same starting point)
- uneven coverage across different regions of transcription units
- antisense artifacts in the case of stranded libraries.

Koboldt et al., 2010
Linnarsson, 2010

NGS PROTOCOLS

RNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

How to prepare the library: critical steps

- Standard methods for RNA library preparation do not always retain **information on the DNA strand** from which the RNA strand was transcribed. However this could be a very useful information for many reasons among which:

- identification of **antisense transcripts**
- determination of the transcribed strand of **noncoding RNAs**
- determination of expression levels of **coding or non-coding** overlapping transcripts.

→ multiple **published modifications of the original method for strand-specific RNASeq** (Levin et al., 2010) with differences in strand specificity, library complexity, evenness and continuity of coverage (…): ***dUTP second-strand marking***

(Linnarsson, 2010)
(van Dijk, Jaszczyszyn &Thermes, 2014)
https://www.neb.com/tools-and-resources/usage-guidelines/getting-started-with-rna-seq

NGS PROTOCOLS

RNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 01 RNA Purification & Selection

↓

STEP 02 cDNA Synthesis

↓

STEP 03 RNA/cDNA Fragmentation

↓

STEP 04 End repair and A-tailing

↓

STEP 05 Adapter Ligation

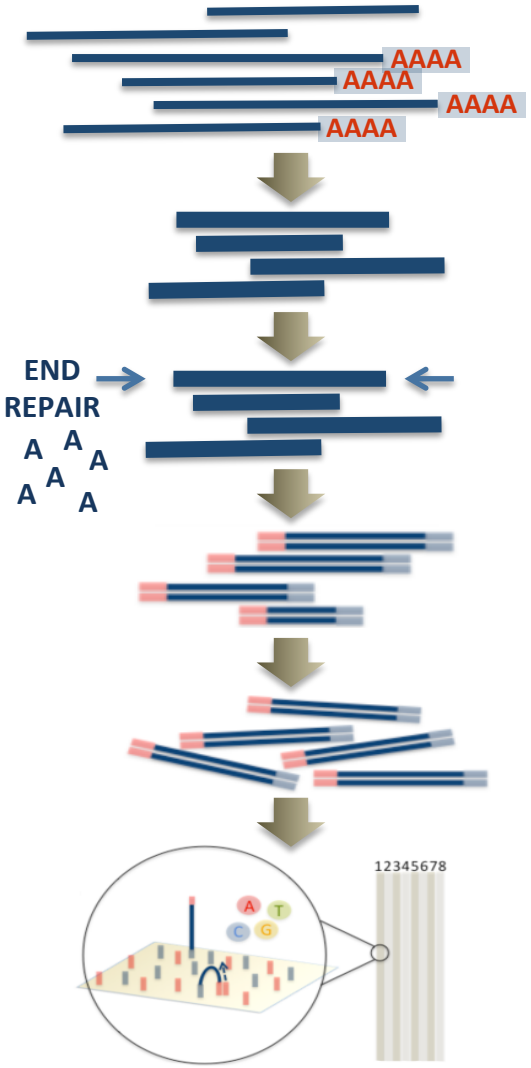
↓

STEP 06 Size Selection & PCR

↓

STEP 07 Sequencing

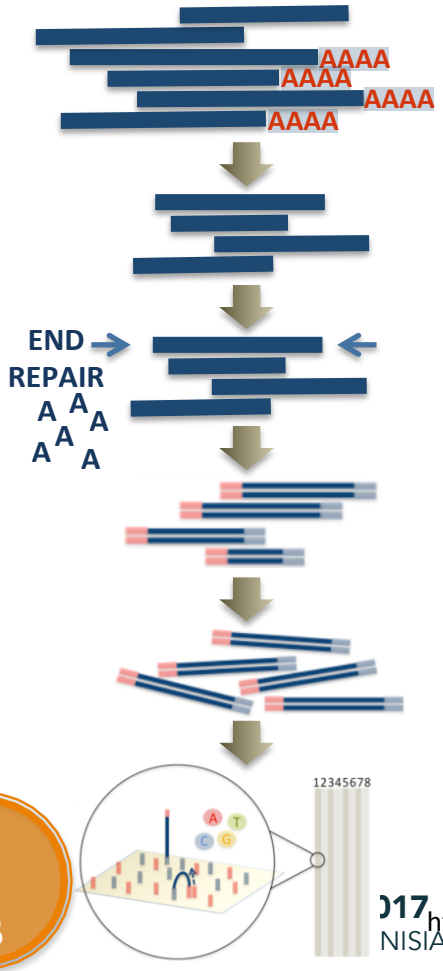
PART 3



NGS PROTOCOLS

RNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 01 RNA PURIFICATION



Starting material: QC

- ***Quality Control***
  - ***Important to start with high quality RNA*** (Bioanalyzer (Agilent, Bio-rad), RIN)
  - Integrity and size distribution (denaturing agarose gel visualization. Ex: for eukaryotes sharp bands of 28S:18S with a 2:1 ratio of intensity)

- ***Quantity Control***
  - Nanodrop, Qubit…(organic compunds or free nucleotides could result in overestimation)

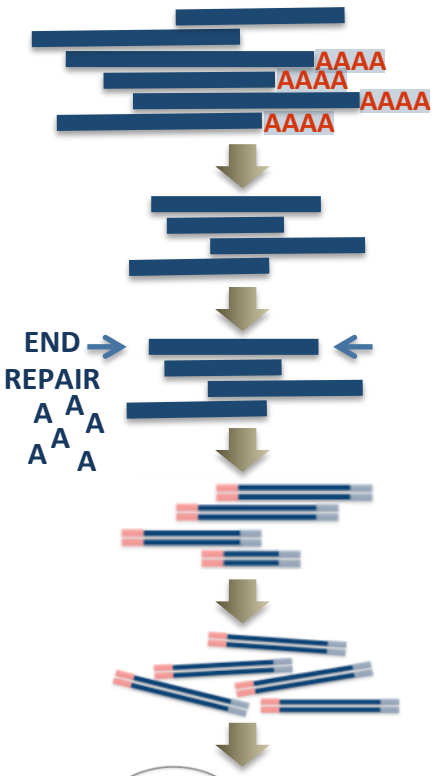
The quality and accurate quantitation of input RNA is critical to ensure successful cDNA synthesis + libraries.

https://www.neb.com/tools-and-resources/usage-guidelines/getting-started-with-rna-seq

NGS PROTOCOLS

RNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 01 RNA PURIFICATION



Starting material: isolate your RNA of interest

- A single mammalian cell may contain about 10 pg of RNA, or about 10 million molecules, of which ≈90% represent ribosomal RNA.

- **Isolate the RNA fraction of interest**
- **directly** (mRNA, miRNA…)
- or **indirectly** by depleting **any other RNA undesired fraction**

Example: interest in the non-ribosomal fraction (mRNA, micro-RNA, tRNA…) → **Ribo-Zero rRNA Removal Kit**

NB: it may be desirable to suppress other highly expressed transcripts, e.g. α- and β-globin in blood.

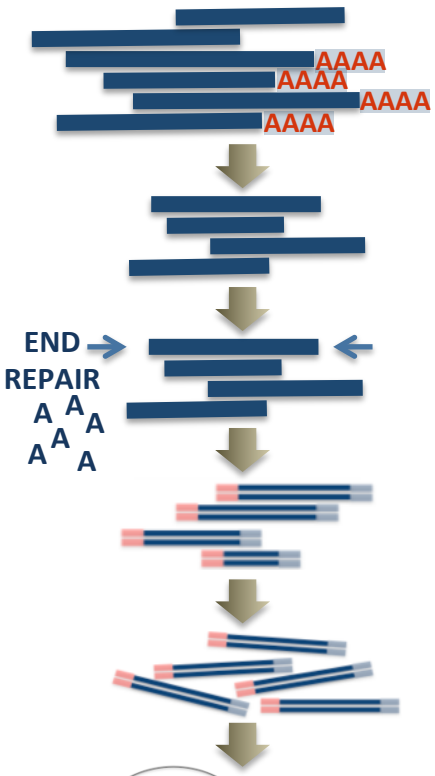
- **Remove any contaminant DNA** using a DNase

http://www.illumina.com/products/by-type/molecular-biology-reagents/ribo-zero-rrna-removal-human-mouse-rat.html

NGS PROTOCOLS

RNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 01 RNA PURIFICATION



Starting material: isolate your RNA of interest

Avoid wrong extraction method

• At low RNA concentrations, specific miRNAs (GC poor or highly structured) are lost during classical RNA extractions using Trizol reagent.
= important take home message : “total” RNA preparations may actually be biased due to unequal precipitation efficiencies.

→Could lead to artifacts in interpreting the results (Kim et al., 2012)

• If miRNA levels are to be compared among samples, it is recommended to :

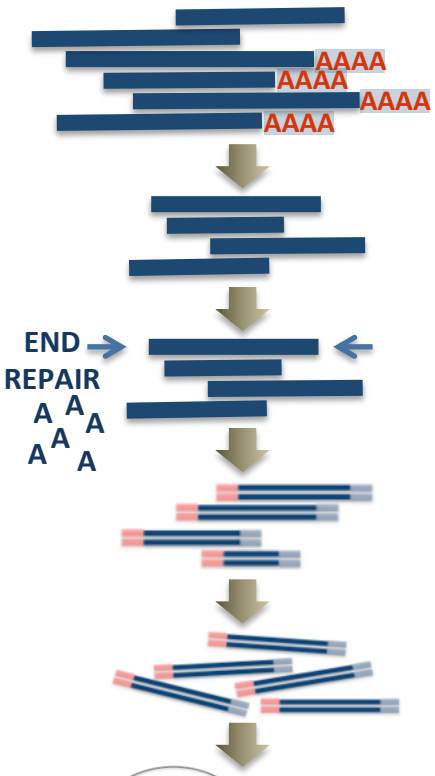
- use similar amounts of material of each sample for RNA extraction
- Or avoid Trizol extraction and use dedicated methods to extract miRNAs (MirVana kit (Ambion – Life Technologies)).

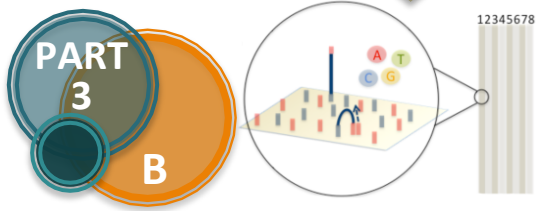
(Linnarsson, 2010)

NGS PROTOCOLS

RNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 02 RNA FRACTION SELECTION





Starting material: isolate your RNA of interest

rRNA depletion

Two main categories of rRNA depletion methods exist:

- **Subtractive hybridization** with probes specific for rRNA to deplete them, method introducing only poor bias in relative transcript abundance (He et al., 2010)
- **Selective degradation** of rRNAs and other 5'monophosphate RNAs by exonucleases, while mRNAs are protected by their 5'cap structures, or triphosphates in prokaryotes.

However :

- highly expressed genes with short half-lives and partially degraded mRNAs preferentially lost after exonuclease treatment.
- + exonuclease treatment tends to be less efficient in rRNA depletion when stable secondary structures block its progression.

(Linnarsson, 2010)

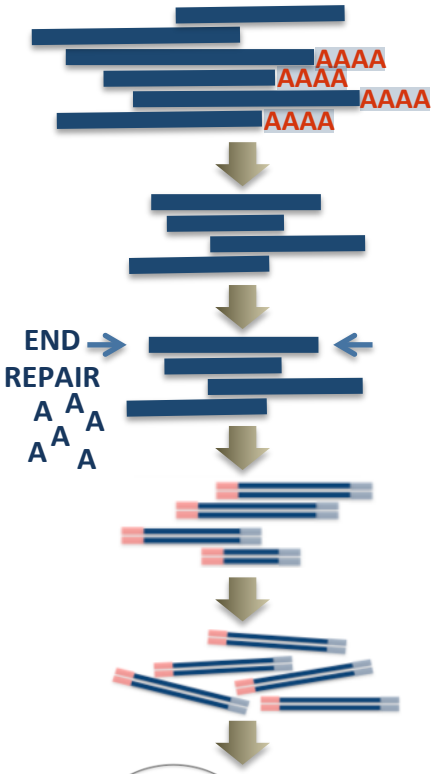
(van Dijk, Jaszczyszyn &Thermes, 2014)

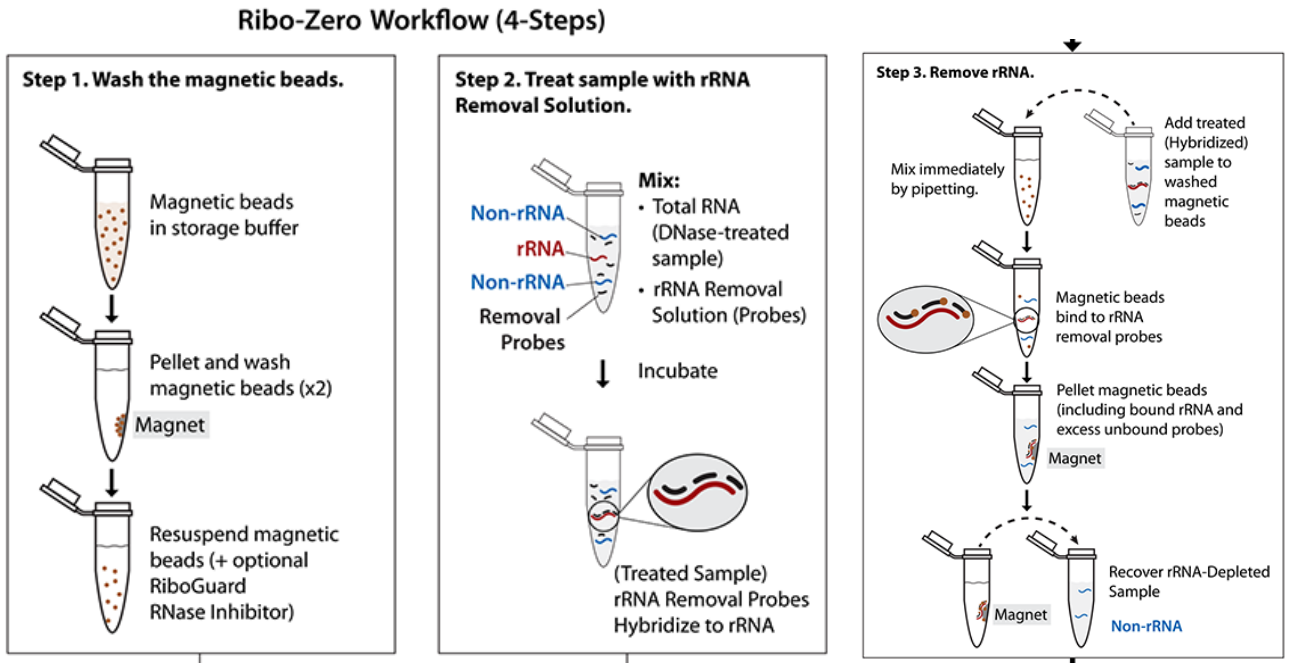
NGS PROTOCOLS

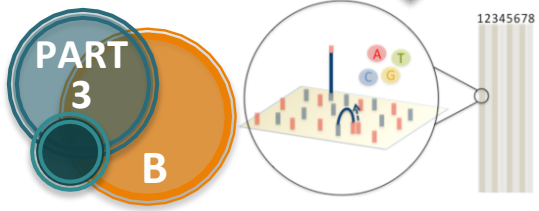
RNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 02 RNA FRACTION SELECTION

- **Starting material: isolate your RNA of interest**





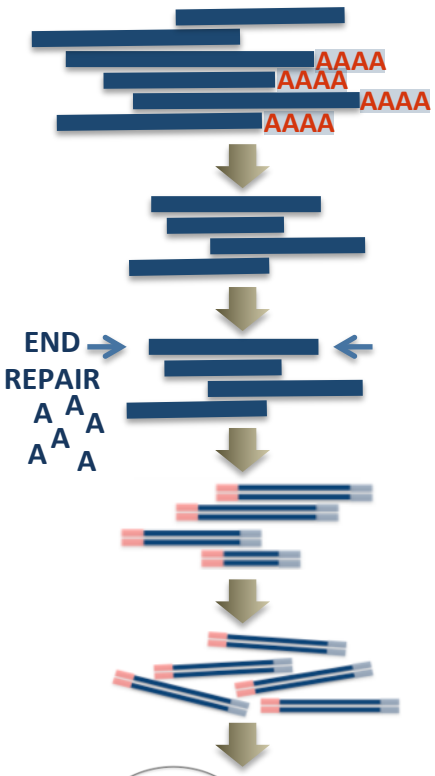


(Linnarsson, 2010)
(van Dijk, Jaszczyszyn &Thermes, 2014)

NGS PROTOCOLS

RNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 02 RNA FRACTION SELECTION



- **Starting material: isolate your RNA of interest**

Avoid loss of starting material

This could result from:

- Unsufficient starting material
- Non cautious purification
- RNAses contamination
- unproper storage
- cross-contamination (very difficult to detect)
- whenever working with low concentration of nucleic acid, it is necessary to use low-adsorbing plasticware (use 'non-stick' tubes).

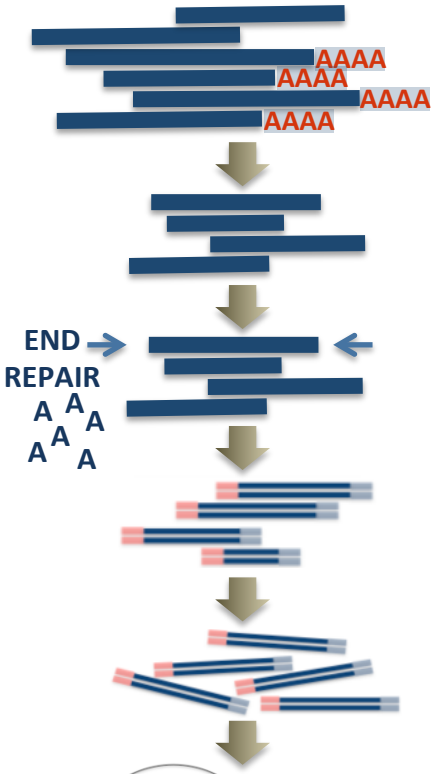
PART 3 B

(Linnarsson, 2010)
(van Dijk, Jaszczyszyn &Thermes, 2014)

NGS PROTOCOLS

RNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 03 RNA/cDNA FRAGMENTATION



Fragmentation

Examples of widely used fragmentation methods

- physical methods

official protocols provided by the manufacturers call for fragmentation by nebulization driven by pressurized air (GA, FLX), or by the Covaris AFA ultrasound device (SOLiD). Other shearing instruments exist…

- enzymatic fragmentation kits

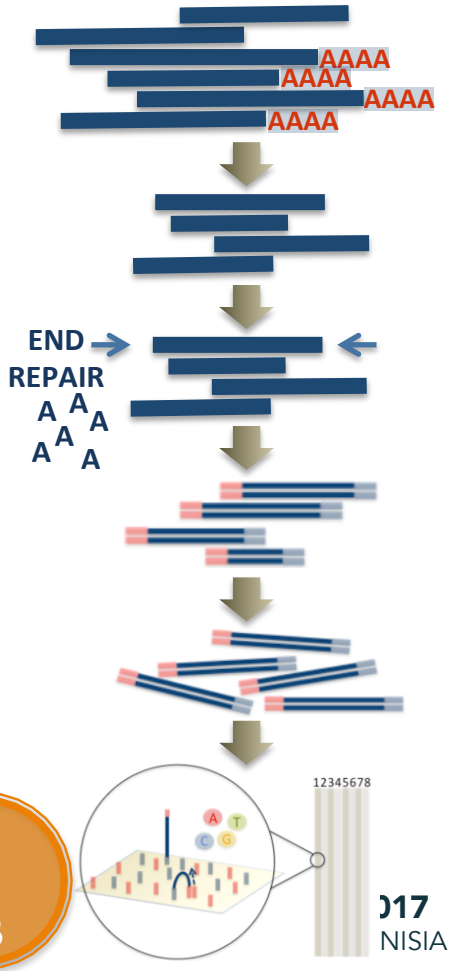
Fragmentase (New England Biolabs) based on bacterial nuclease, Nextera (Epicentre) based on random transposon insertion

-…

NGS PROTOCOLS

RNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 03 RNA/cDNA FRAGMENTATION



2 strategies

***RNA fragmentation prior to cDNA synthesis*:** (Illumina 'TruSeq Stranded mRNA Sample PrepKit', standard Illumina protocol, RNA ligation method). Different techniques of RNA fragmentation have been described. Some protocols use chemical zinc-mediated cleavage, while other protocols are based on RNase III.

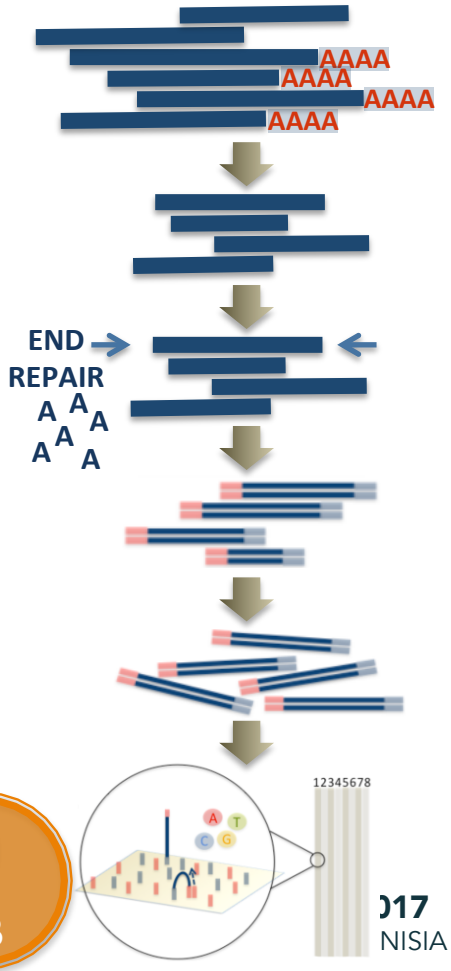
cDNA fragmentation: in the original dUTP protocol double-stranded cDNAs generated from intact RNA are fragmented.

PART 3 B

NGS PROTOCOLS

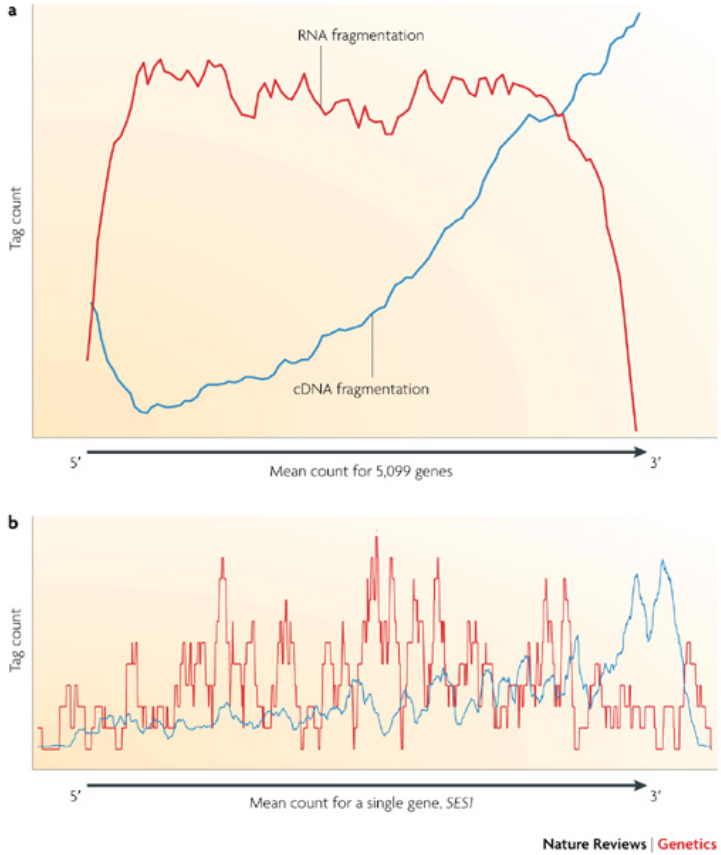
RNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 03 RNA/cDNA FRAGMENTATION



- **2 strategies**

Biases during library construction



Fragmentation of oligo-dT primed cDNA (blue line) is more biased towards the 3' end of the transcript. RNA fragmentation (red line) provides more even coverage along the gene body, but is relatively depleted for both the 5' and 3' ends.

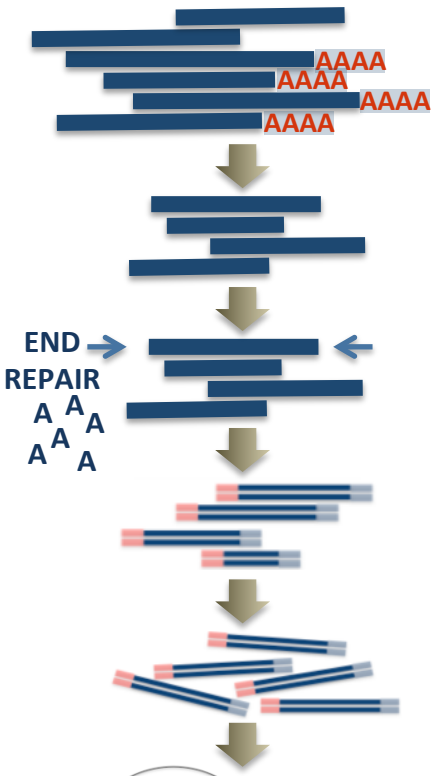
A specific yeast gene, SES1 (seryl-tRNA synthetase)

PART 3 B

NGS PROTOCOLS

RNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 03 RNA/cDNA FRAGMENTATION



Reverse Transcription and 2nd strand DNA Synthesis

- **Reverse Transcriptase**: Complementary DNA (cDNA) is generated from the RNA template.
  → 1st strand cDNA synthesis
- **DNA Polymerase**: RNA/cDNA hybride is then made + ds cDNA after RNA degradation.
  → 2nd strand cDNA synthesis
- Biases described:
  - Enzyme (ligase)-dependant (incubation T°C…)
  - RNA sequence and structure-dependance

PART 3 B

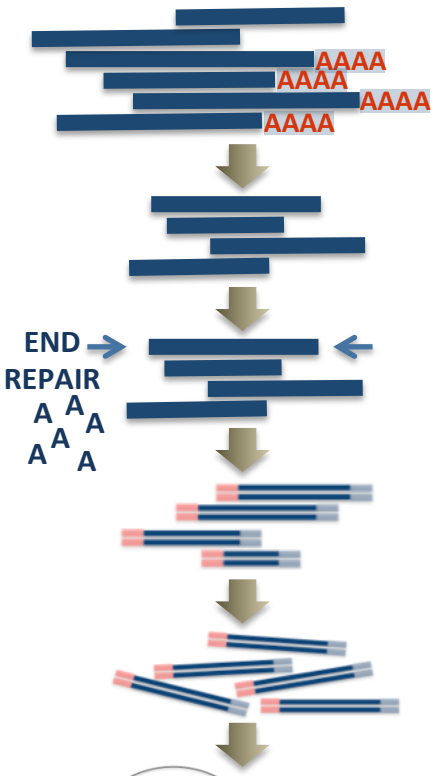
Hafner et al., 2011

https://www.neb.com/tools-and-resources/usage-guidelines/getting-started-with-rna-seq

NGS PROTOCOLS

RNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 04 END REPAIR AND A-TAILING



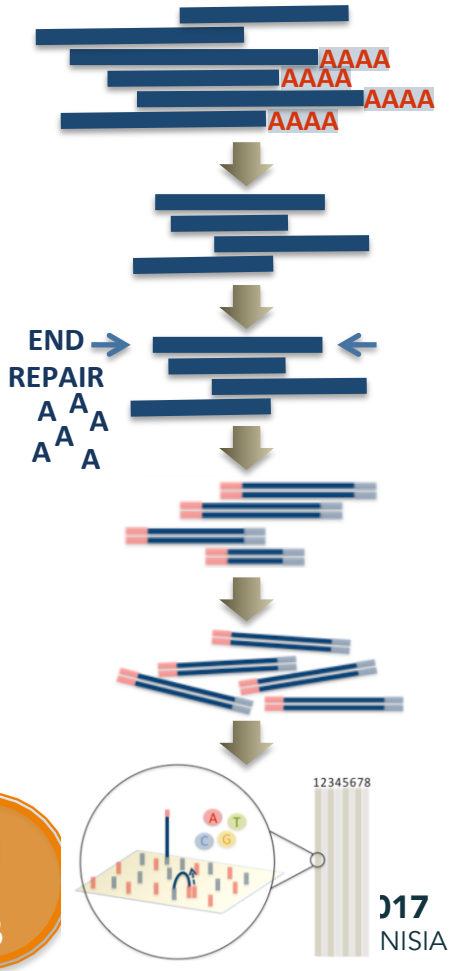
- **After the fragmented library is repaired** (that is, protruding 3’ and 5’ ends are removed or filled in), adapters must be ligated.
- The library is then ready for amplification and sequencing.

https://www.neb.com/tools-and-resources/usage-guidelines/getting-started-with-rna-seq

NGS PROTOCOLS

RNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 05 **ADAPTER LIGATION**



- In the Illumina RNA ligation method, **RNA fragments are ligated with a pre-adenylated 3'adapter, followed by ligation of a 5'adapter**, using Rnl1 and Rnl2, 2 different families of **RNA end-joining enzymes**.

Systems require each fragment to have distinct upstream and downstream adapters (A and B).

Risks:

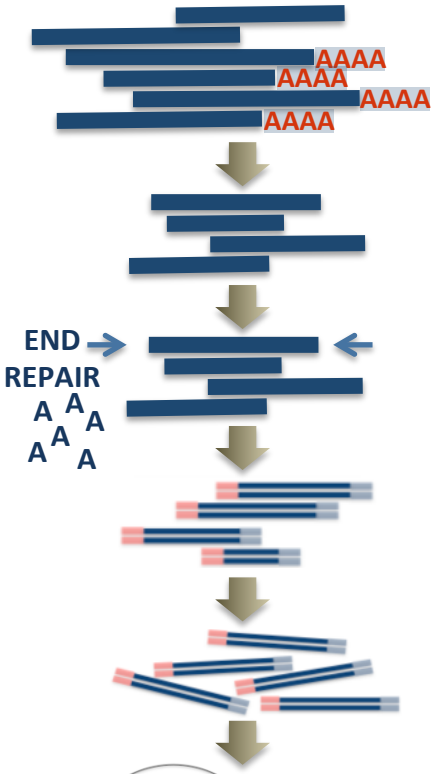
- Rnl1 and Rnl2 have been shown to have **different substrate specificities**
- 2 fragments can be accidentally joined together, forming a **chimera which generates misleading reads**.

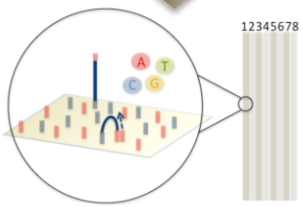
(Linnarsson, 2010)
(van Dijk, Jaszczyszyn &Thermes, 2014)

NGS PROTOCOLS

RNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 05 ADAPTER LIGATION





- **Ligation bias** can have dramatic effects on both:

  - the fidelity of expression profiles
  - reproducibility across samples.

(3'-adapter ligation reactions critical steps in introducing biases in cDNA sequence read representation)

Ex: Small RNAs characterized by unstable secondary structures underperformed in 3'- and 5'-adapter ligation, especially when using Rnl1

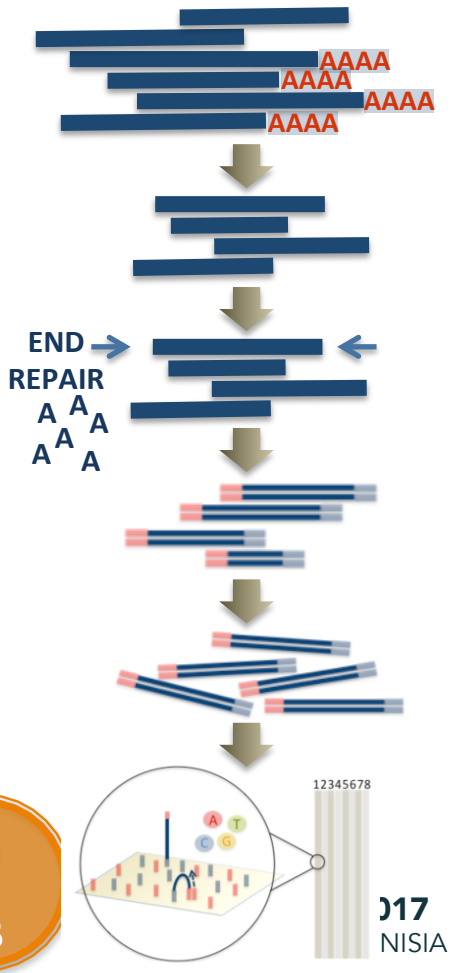
→ Perform adapter ligations with

- an excess of adapter
- reaction conditions including time and temperature are held constant

NGS PROTOCOLS

RNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 06 **SIZE SELECTION AND PCR**



- **PCR Amplification**

  - The **ligation products are subsequently reverse transcribed and amplified by PCR**.

  - The **most problematic step** is the PCR amplification, which results in loss of specific regions of the template DNA as other regions are more efficiently amplified

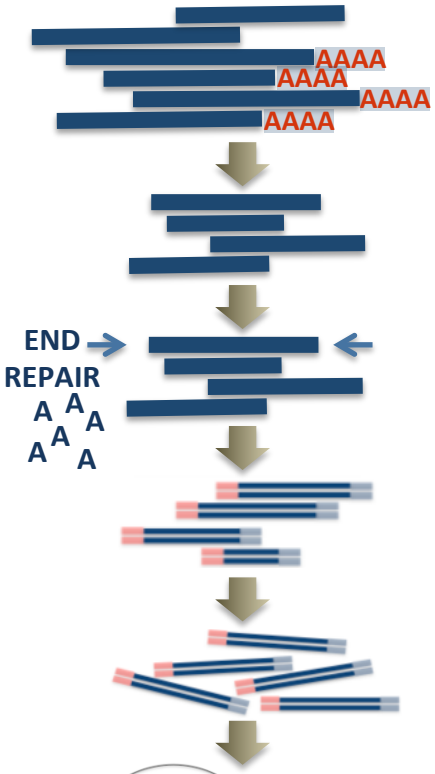
GC-rich or AT-rich fragments may be under represented

(Linnarsson, 2010)
(van Dijk, Jaszczyszyn &Thermes, 2014)

NGS PROTOCOLS

RNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 06 **SIZE SELECTION AND PCR**



- **Size selection**

Size selection may still be performed by agarose gel electrophoresis or on-column + bioanalyzer check possible.

Risk: the standard protocols for gel extraction include a **heating step** that **may denature some AT-rich sequences**.

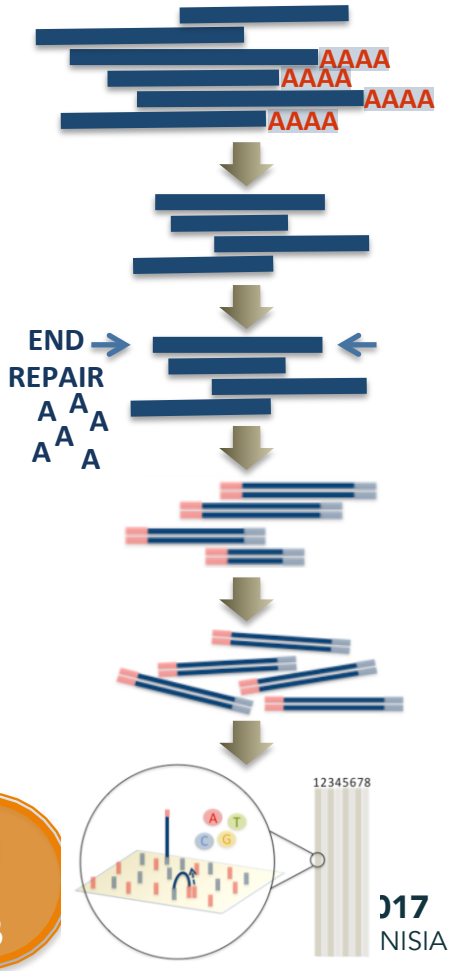
NB: **This effect severely complicates the sequencing of extremely AT-rich genomes** (*P. falciparum*)

(Linnarsson, 2010)
(van Dijk, Jaszczyszyn &Thermes, 2014)

NGS PROTOCOLS

RNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 06 SIZE SELECTION AND PCR



- **Quality Control**

An often overlooked but crucial sample preparation step before running the sequencing is this QC

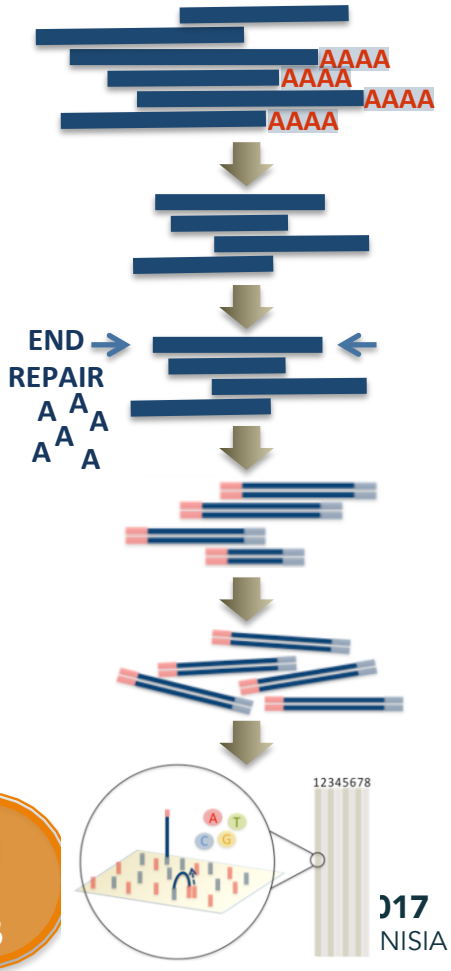
- **Quality control** (using capillary electrophoresis BioAnalyzer or Experion)
- **Quantification** (Nanodrop)

(Linnarsson, 2010)
(van Dijk, Jaszczyszyn &Thermes, 2014)

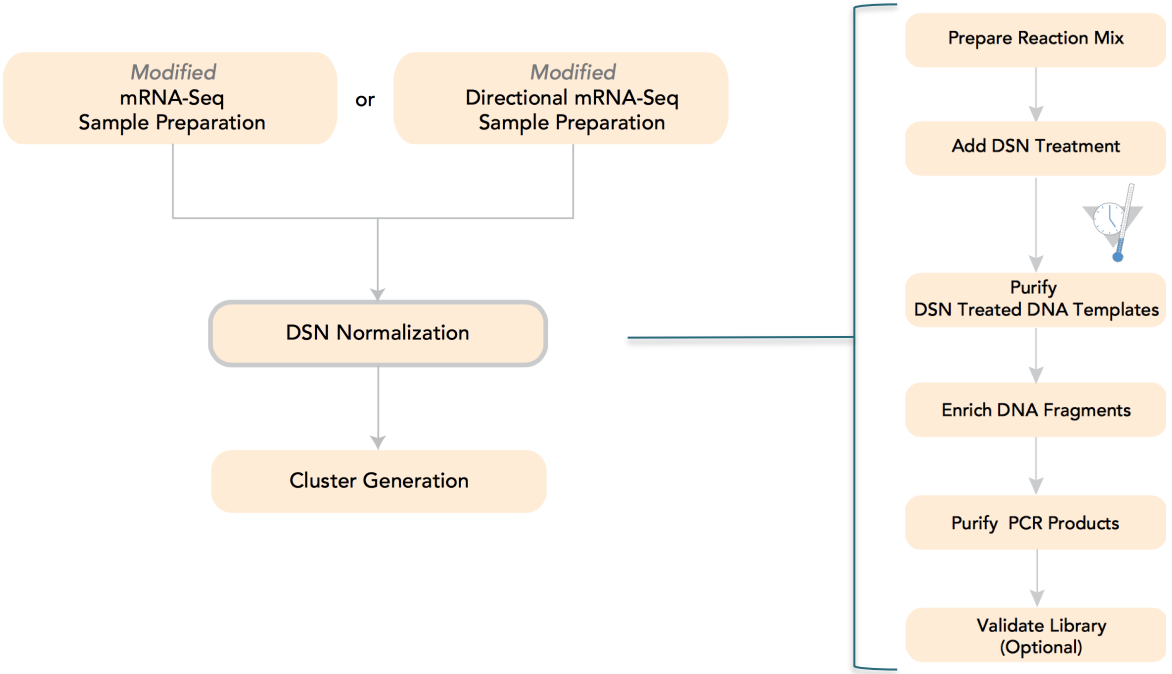
NGS PROTOCOLS

RNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 06 **SIZE SELECTION AND PCR**



- **Optional: DSN treatment**
  - normalize Illumina® RNA-seq sample preparation
  - Uses the Duplex-Specific thermostable nuclease (DSN) enzyme (www.evrogen.com).
  - Degradation of abundant molecules (rRNA, tRNA, housekeeping genes…) while preserving molecules derived from less abundant transcripts.

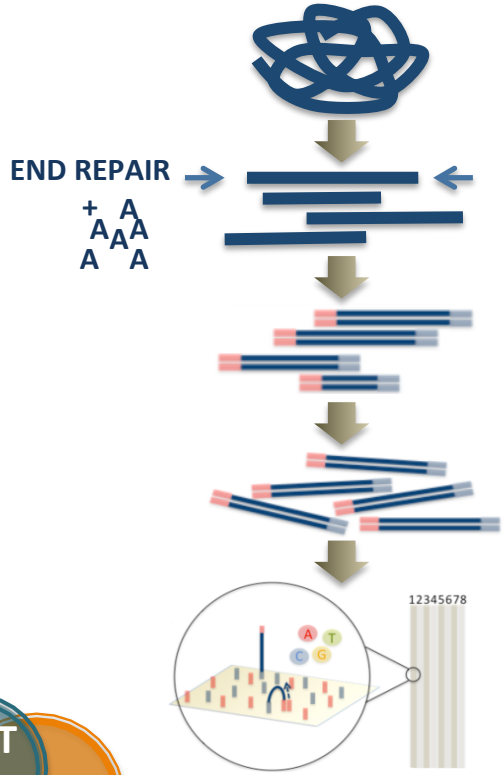


PART 3 B

NGS PROTOCOLS

RNA-seq: PROTOCOL FOR LIBRARY CONSTRUCTION

STEP 06 SEQUENCING



RNA Sequencing

- output : sequenced « reads » (fastq files)

NB: the number of sequenced reads can be approximated knowing the overall genome size and the coverage required

NGS PROTOCOLS

TAKE-HOME MESSAGES

- **How much sequence data is enough?**

Often this question is answered by the **practical considerations** of funding, instrument access, and/or the availability of sample material.

- **Choose your indicators of performance**

Certain **performance metrics** can indicate the expected quality and completeness of a sequenced genome.

number of reads, average read length, alignment rate, and inferred error rate are the most obvious indicators of success or failure.

All depends on your question !!

NGS PROTOCOLS

TAKE-HOME MESSAGES

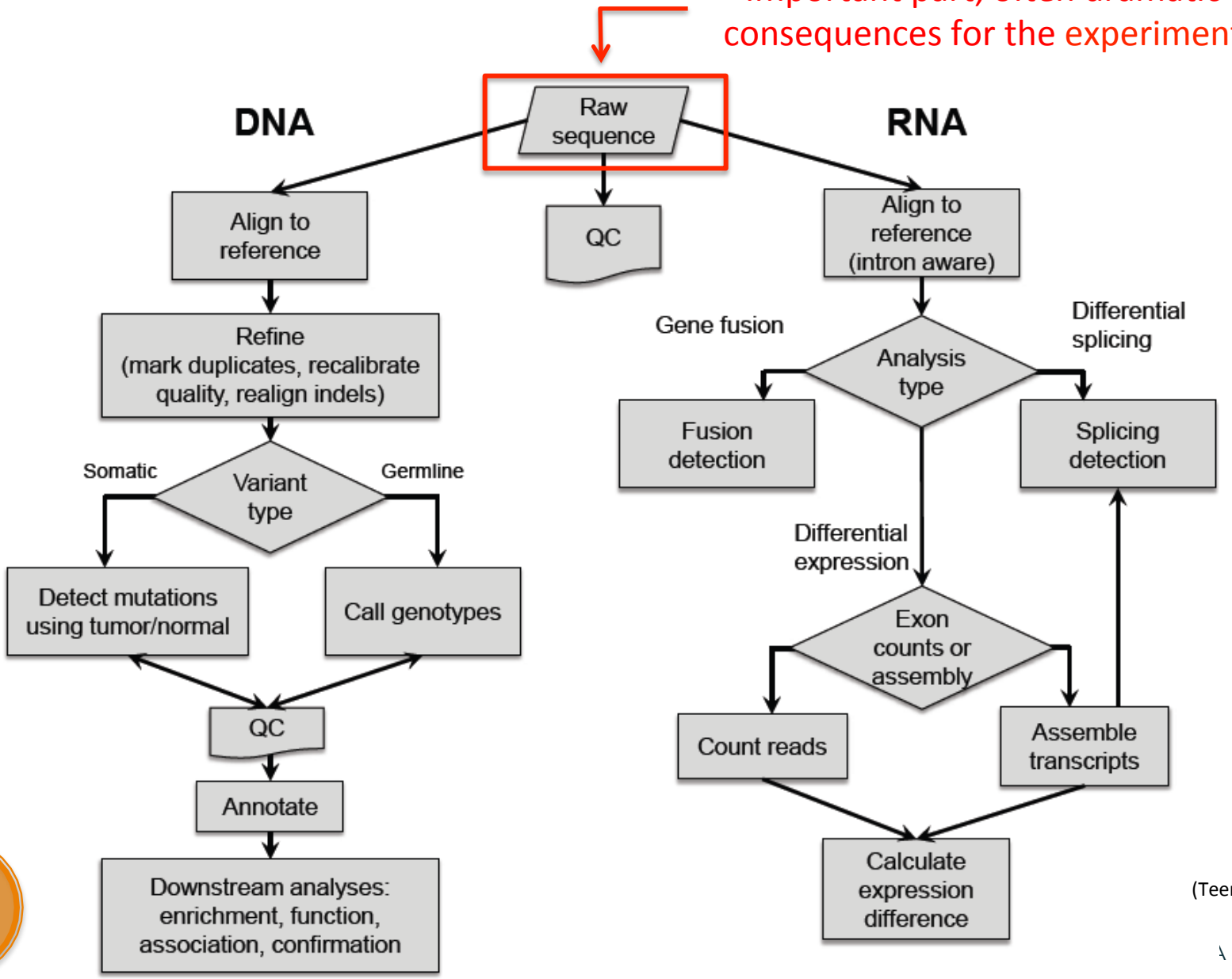
- **Biological question**
  - need to be clearly defined first, so that the design of the experiment, the library construction and the pipeline of analysis could be prepared accordingly
- **Platfom**
  - Each one has its own specificities that needs to be understood before choosing one
  - Each one has its own errors
  - Different technologies, short reads (Illumina…) vs long reads (PacBio…)
  - Rapidly evolving, several limitations (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)

File Formats
(DNA-seq)

NGS PROTOCOLS

OVERVIEW

Important part, often dramatic consequences for the experiment



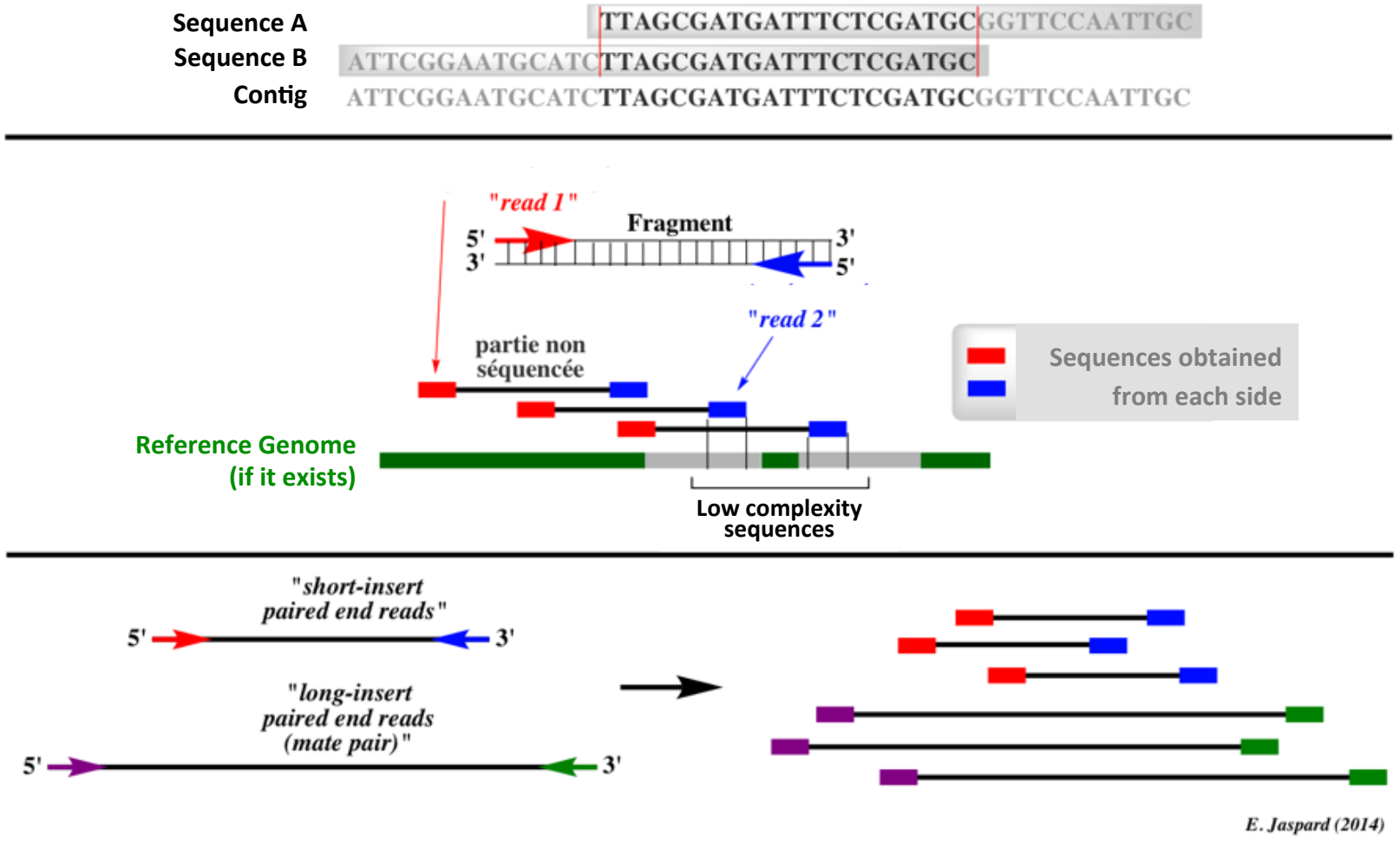
(Teer et al., 2014)

PART 3 C

SEQ TECHNOLOGIES

MOST COMMON FILE FORMATS

- NGS

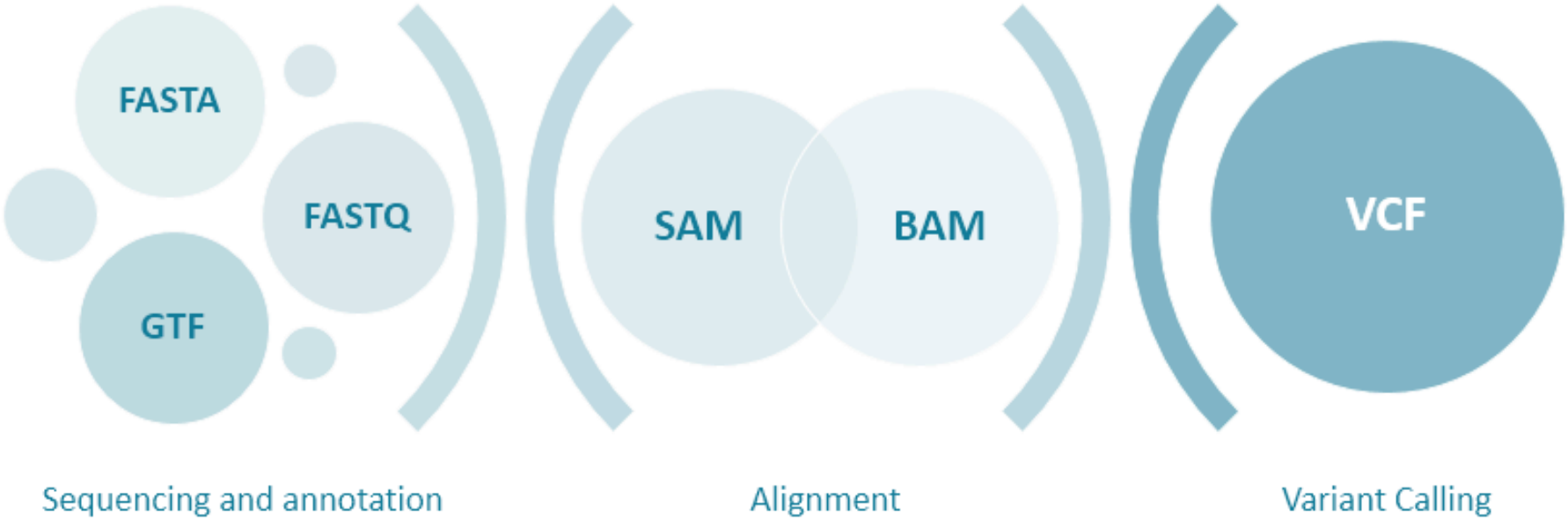


(Jaspar, 2014)

NGS PROTOCOLS

MOST COMMON FILE FORMATS

- At each step of the analysis process, different files are generated.
- Each file contains an additional information related to the analysis performed

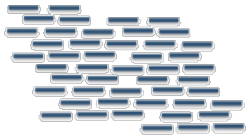


NGS PROTOCOLS

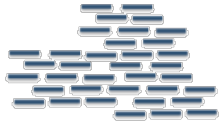
FILE FORMATS

FASTQ FILES 1

FASTQC
QUALITY CONTROL OF READS



TRIMMING
FILTERING BAD QUALITY READS



SEQUENCE ALIGNMENT 2

ASSEMBLY (*DE NOVO*)
RECONSTRUCTION OF A GENOME

MAPPING
OF READS TO A REFERENCE GENOME

FASTA FILE | **GFF FILE**

3

SAM FILES

BAM FILES

4

READ DEPTH

STRUTURAL VARIATIONS

GENE / CHR CNV

VARIANT CALLING

5

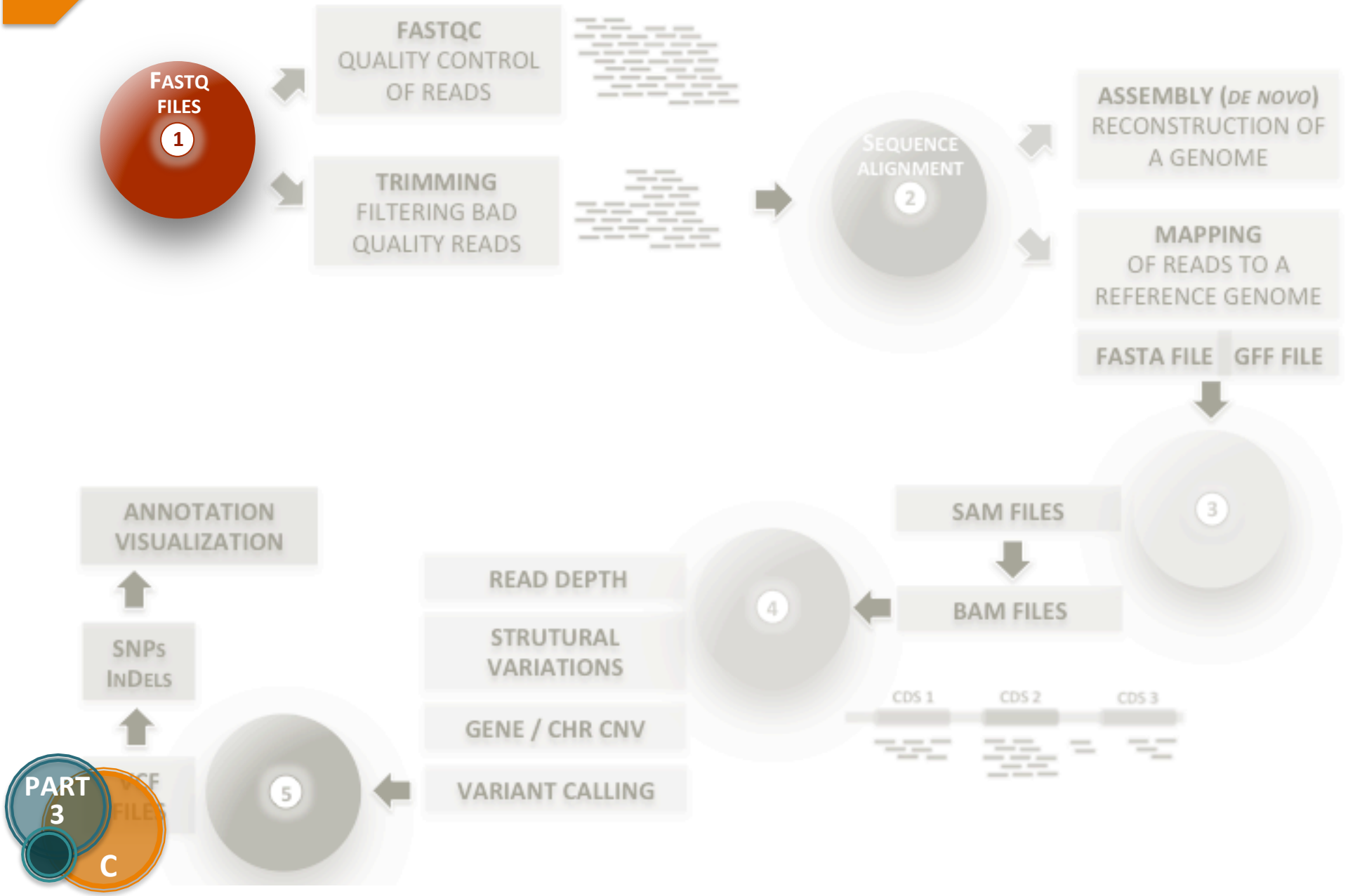
VCF FILES

SNPs InDels

ANNOTATION VISUALIZATION

PART 3 C

NGS PROTOCOLS



FASTQ

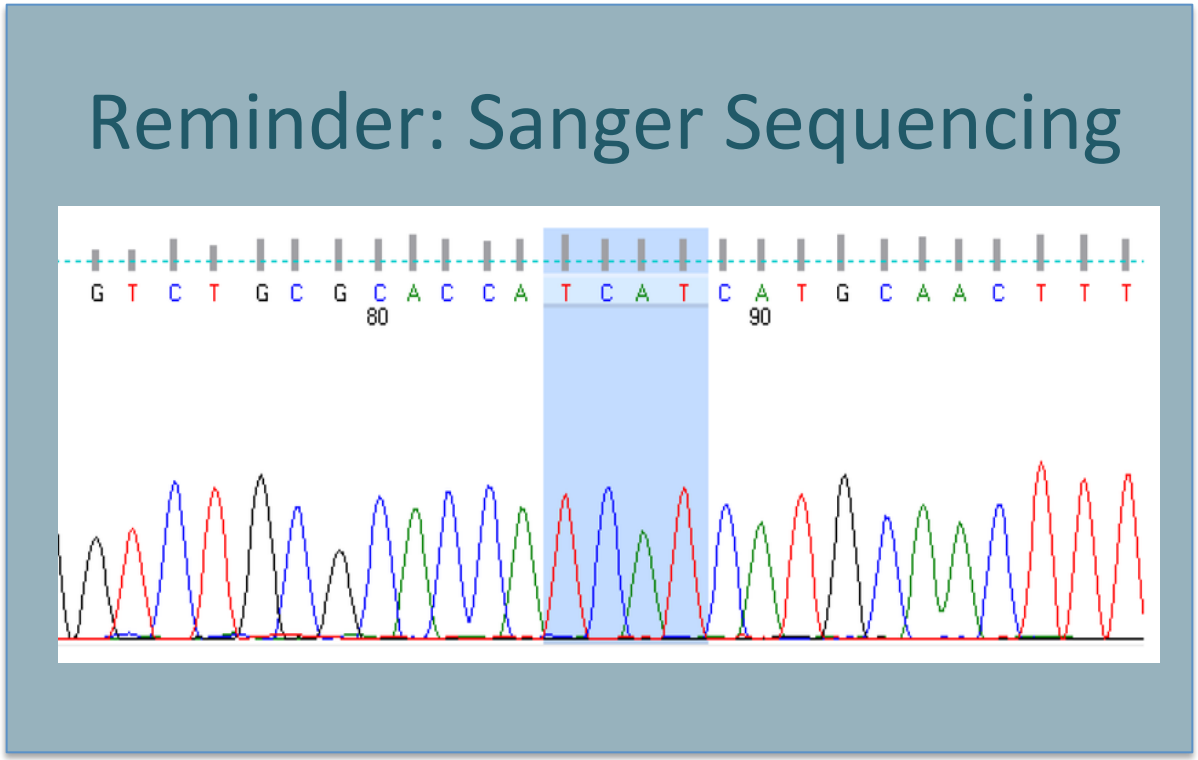
a FASTQ file is the most common output provided by sequencing platforms.

NB: Additional files exist depending on the technology used. Example: for a PacBio sequencing project, there might even be a difference between the different sequencing devices:

PacBio RS Platform	Data Files Delivered
PacBio RS	1. xxxx.metadata.xml (optional but desirable) 2. xxxx.bas.h5
PacBio RS II	1. xxxx.metadata.xml (optional but desirable) 2. xxxx.bas.h5 3. xxxx.1.bax.h5 4. xxxx.2.bax.h5 5. xxxx.3.bax.h5

https://www.ncbi.nlm.nih.gov/sra/docs/submitformats/
https://www.ncbi.nlm.nih.gov/core/assets/sra/files/bas.h5ReferenceGuide.pdf

FASTQ



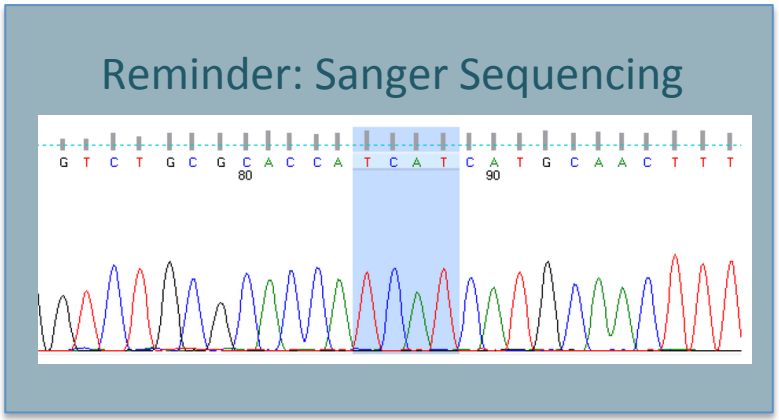
→a FASTQ file is a file containing :

- Reads sequences
- a Quality score associated to each Read

NGS PROTOCOLS

FILE FORMATS

FASTQ



a FASTQ file is a file containing 4 lines / Read:

- Read Sequence identifier (encoded descriptions of instrument, lane…)
- Reads sequences
- « + » sign (optional: « + » followed by seq identifier) separating line2 and line4
- a Quality score associated to each Read

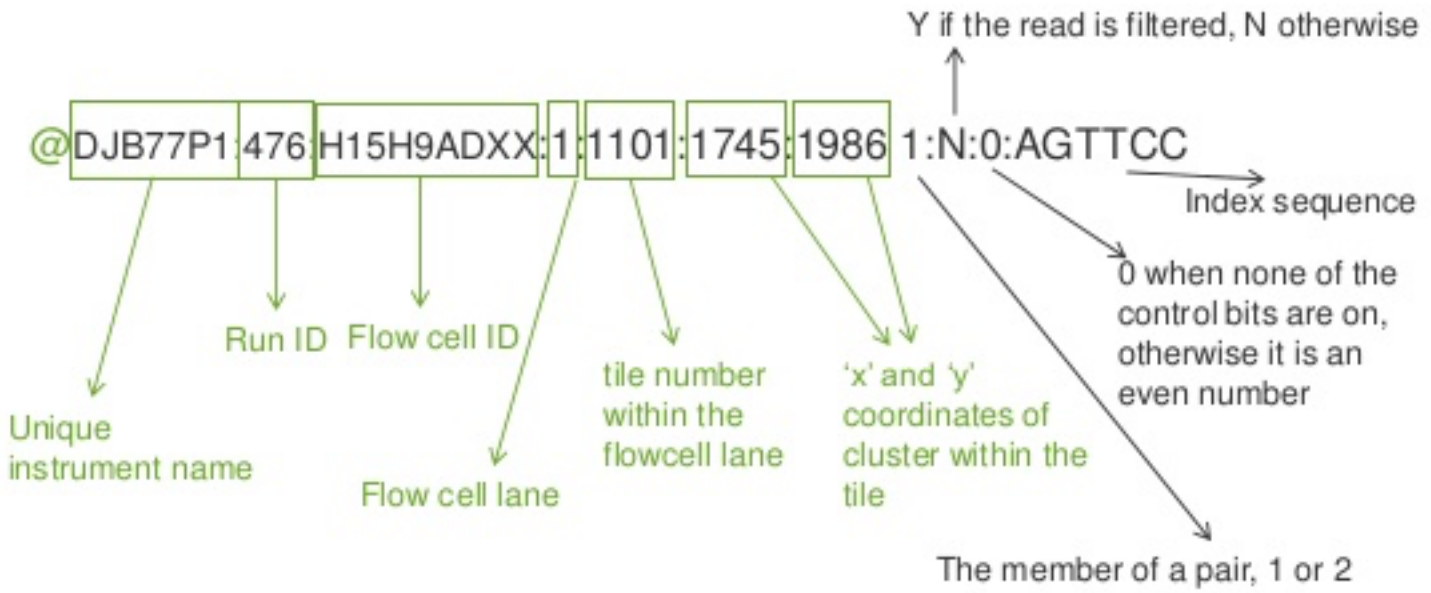
http://drive5.com/usearch/manual/fastq_fig.jpg

PART 3 C

FASTQ

a FASTQ file is a file containing :

- Reads sequences
- a Quality score associated to each Read



https://image.slidesharecdn.com/03dmqb20151-150504081828-conversion-gate02/95/data-management-for-quantitative-biology-data-sources-next-generation-technologies-apr-30-2015-dr-stefan-czemmel-27-638.jpg?cb=1430727633

FASTQ

a FASTQ file is a file containing :

- Reads sequences
- a Quality score associated to each Read

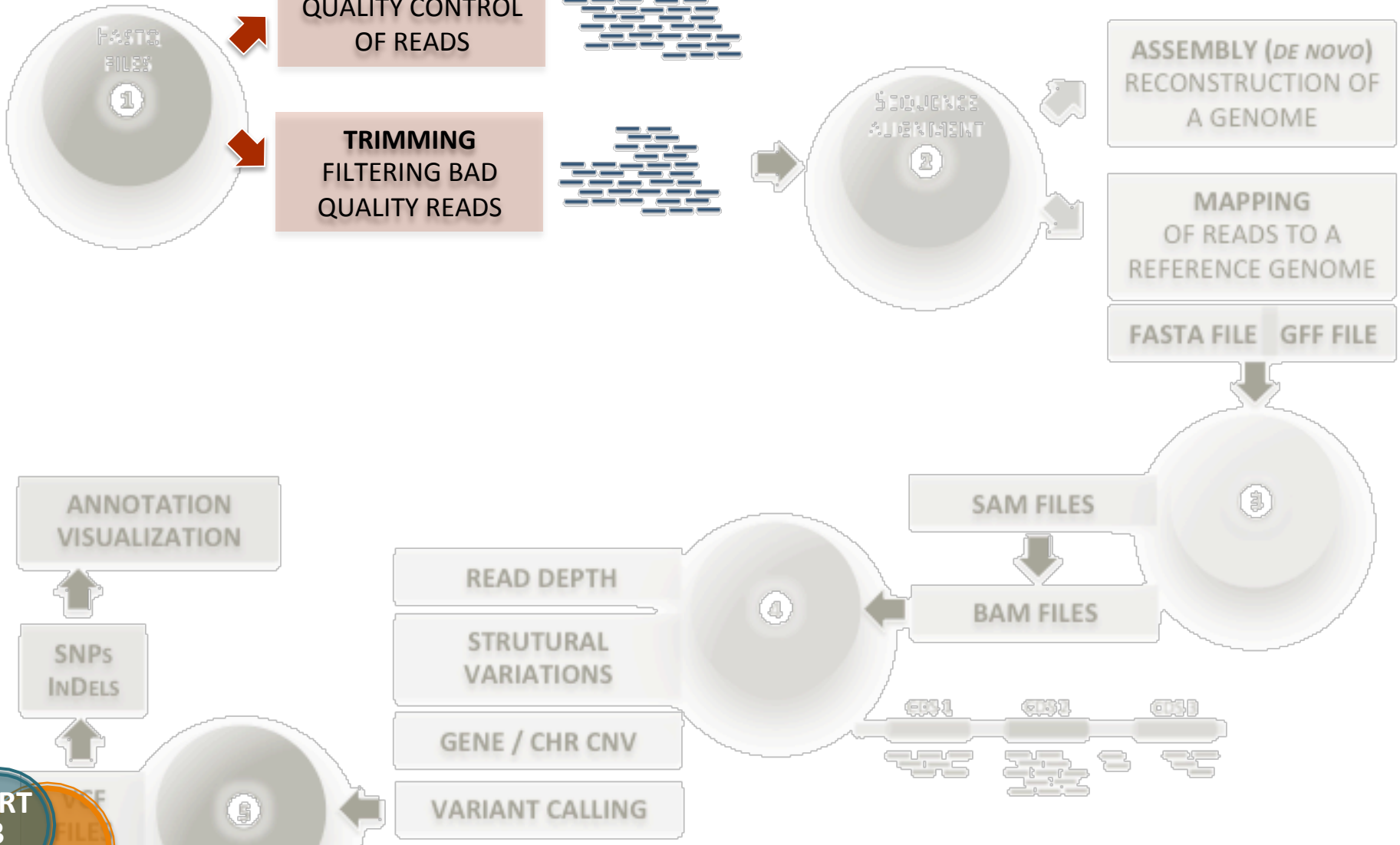
$$Q = -10\log_{10}P$$

Q = Quality Value
P = Error Probability

Prob.of incorrect base call	Phred quality Score	Base call accuracy
1 in 10	10	90%
1 in 100	20	99%
1 in 1000	30	99.9%
1 in 10000	40	99.99%
1 in 100000	50	99.999%

NGS PROTOCOLS

FILE FORMATS



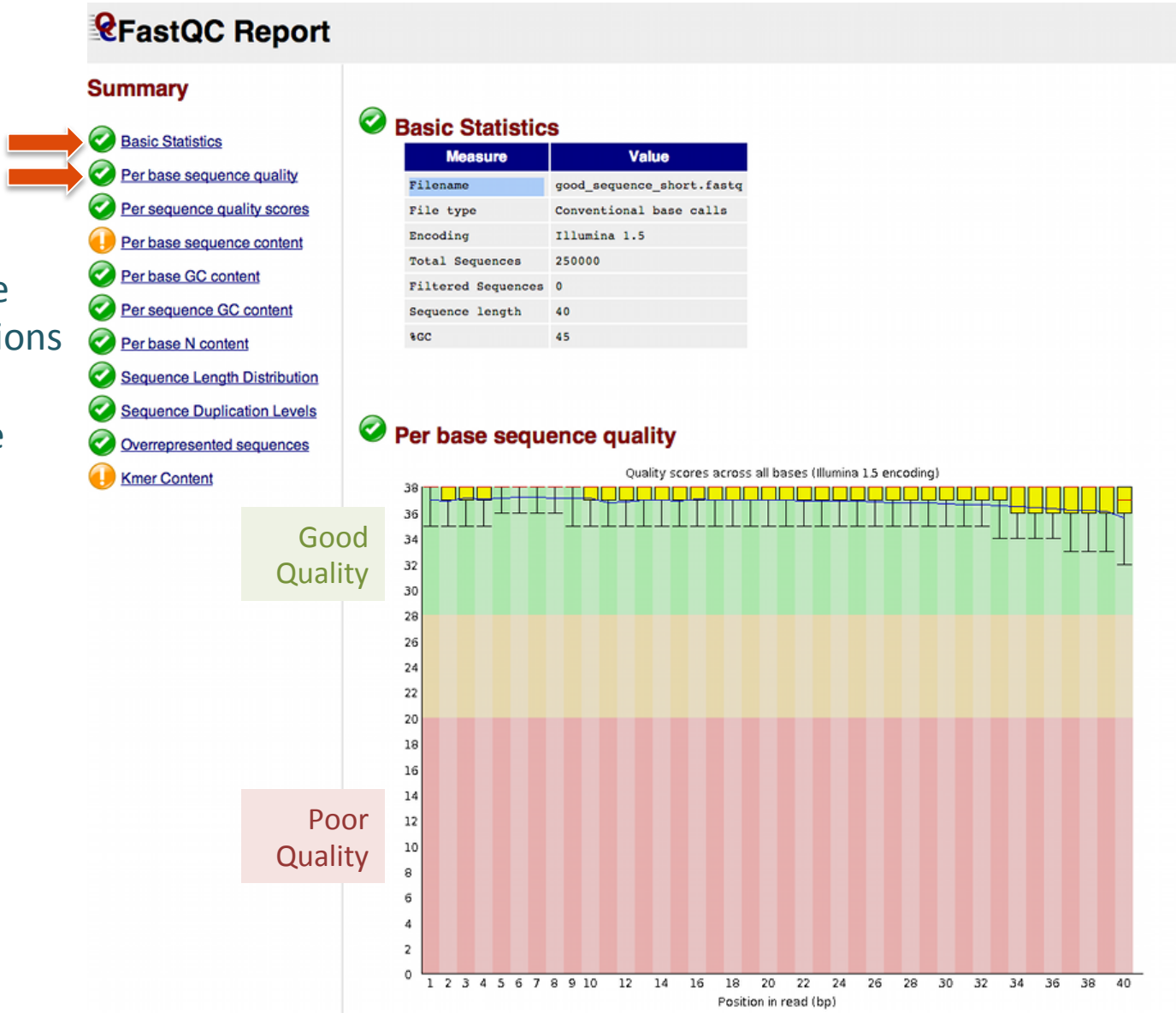
PART 3 C

NGS PROTOCOLS

FILE FORMATS

FASTQC

a FASTQC file is a file containing informations about the quality control made on the FASTQ file



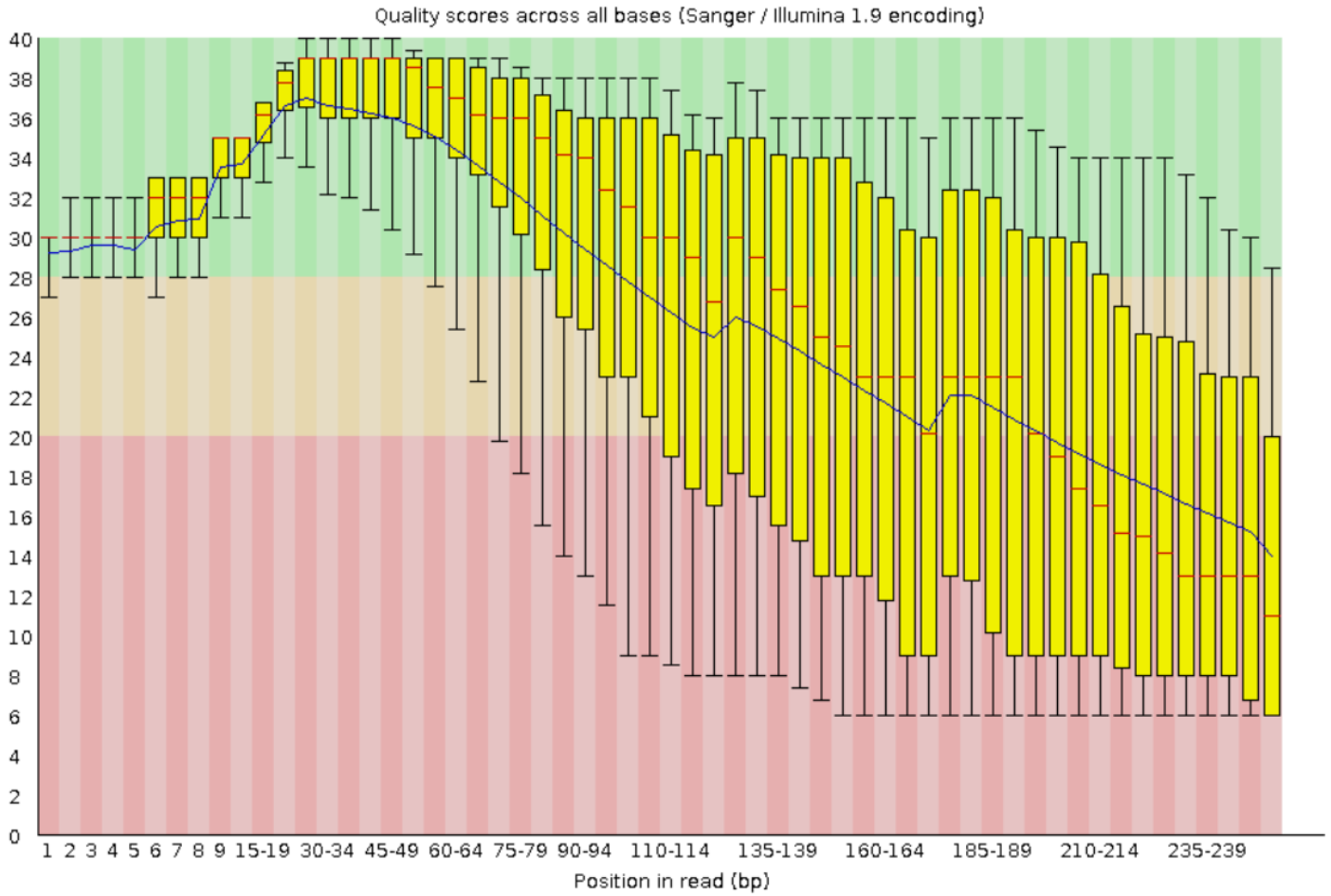
https://wiki.hpcc.msu.edu/download/attachments/15434467/fastqc-1.png?version=1&modificationDate=1365623346000&api=v2

NGS PROTOCOLS

FILE FORMATS

FASTQC

Per base sequence quality



- Bad quality data
- High variability
- High decrease in terms of quality towards the end of the read

→ Important to remove bad quality data

TRIMMING

- **Trimming low Quality Bases:**
  - pre-processing step : removes the low quality bases, identified by the probability that they are called incorrectly.

  - widely but heterogeneously applied

  - However the impact of trimming on subsequent alignment to a genome could influence downstream analyses (gene expression estimation…)

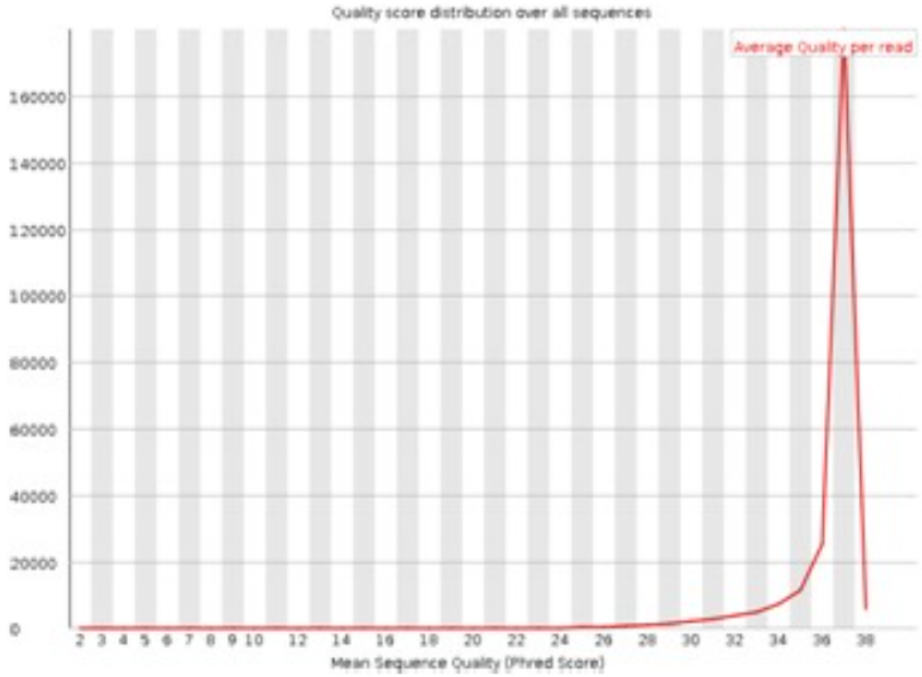
- **Adapter Trimming**

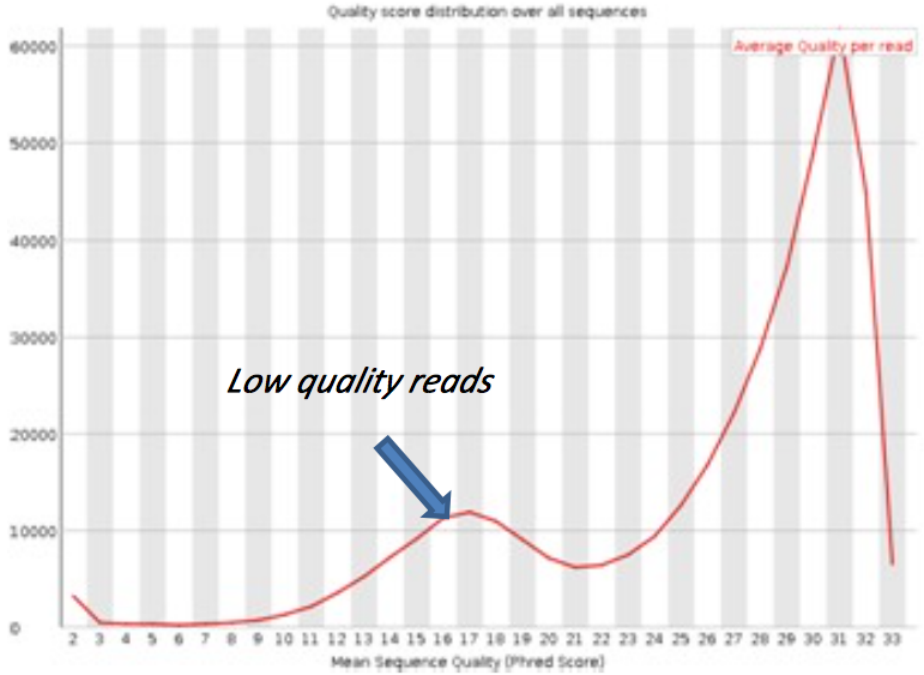
NGS PROTOCOLS

FILE FORMATS

Per Sequence Quality Scores

- 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



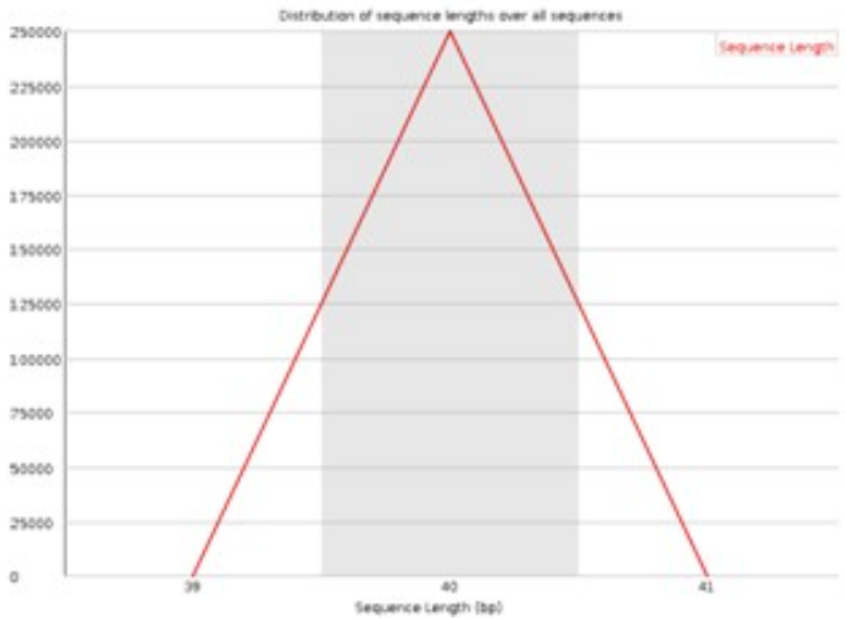


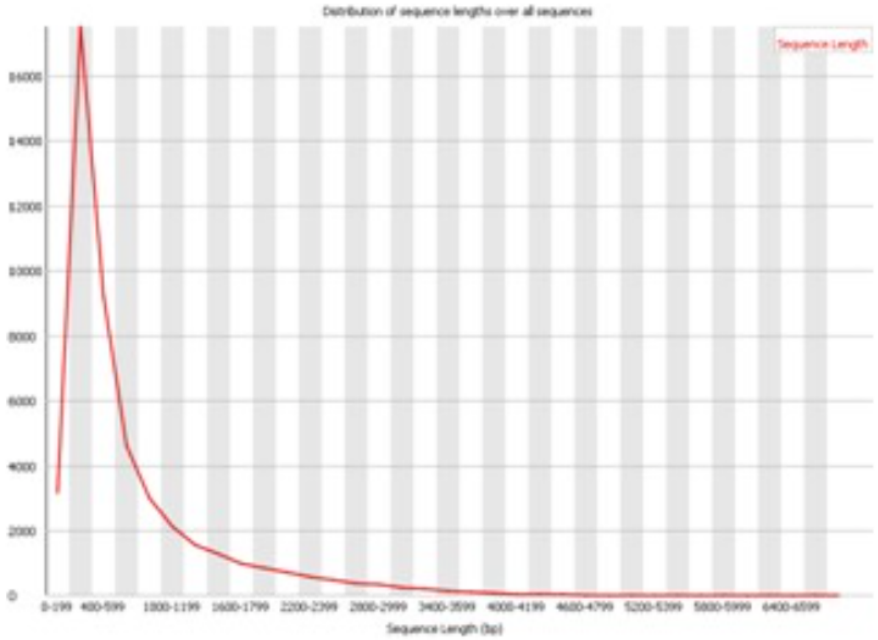
NGS PROTOCOLS

FILE FORMATS

Sequence Length Distribution

- 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





TRIMMING

- Trimming low Quality Bases

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

Manipulating tools

FASTX-Toolkit provides a set of command line tools for manipulating fasta and fastq files. The available modules include a fastx_trimmer utility for trimming fastq sequences (and quality score strings) before alignment.

```
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

TRIMMING

- Adapters Trimming

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 differing numbers of 3' bases depending on where the adapter sequence is found.

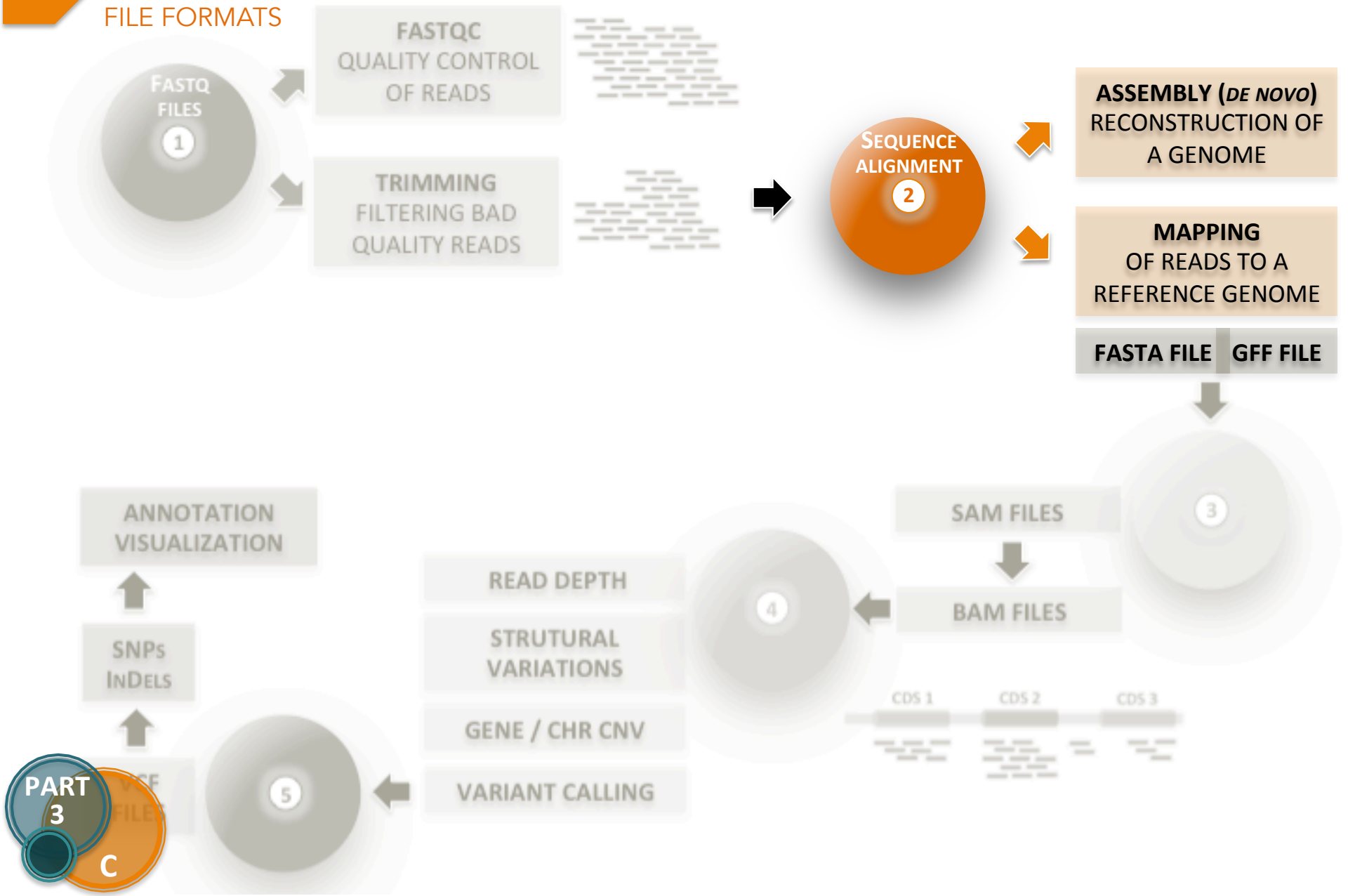
Manipulating tools
Cutadapt program is an excellent tool for removing adapter contamination.
Ex cutadapt on small RNA-seq library data (give it different sequences to trim for R1 and R2 reads). Example for one:

```
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

NGS PROTOCOLS

FILE FORMATS



NGS PROTOCOLS

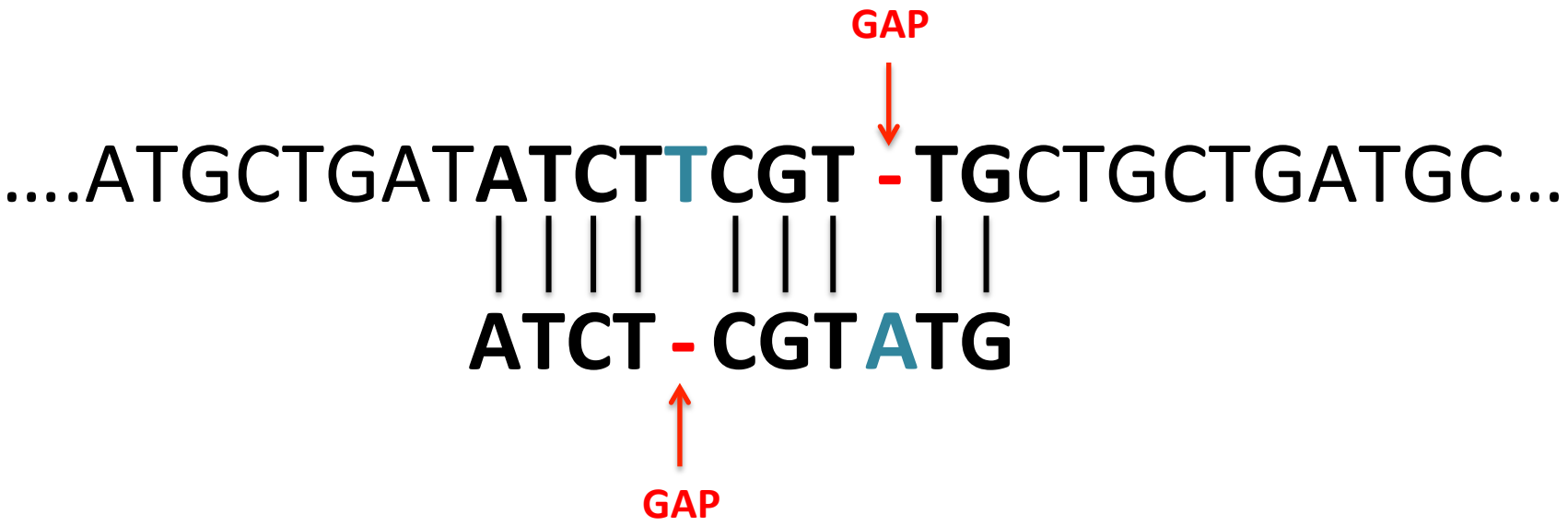
FILE FORMATS

- Alignment is the process by which we discover how and where one **queried sequence** is similar to a **reference sequence**.

- More specifically: an alignment is a way of "lining up" **some or all of the characters in a sequence** with **some or all characters** from the reference in a way that reveals how similar they are.

Compare sequences (BLAST…)

Reference ….ATGCTGAT**ATCTTCGTTG**CTGCTGATGC…
and **ATCTCGTATG**



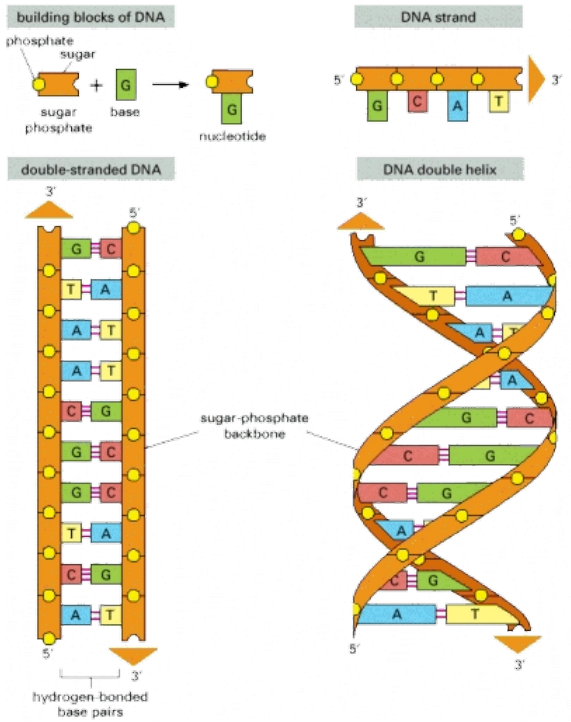
NGS PROTOCOLS

FILE FORMATS

- **Important questions in Biology**:

How to find:

- the origin of a sequence
- The relationship between different sequences/organisms ?
- ….



QUERY

ATCTCGTATG

DATABASE

NGS PROTOCOLS

FILE FORMATS

1 query sequence

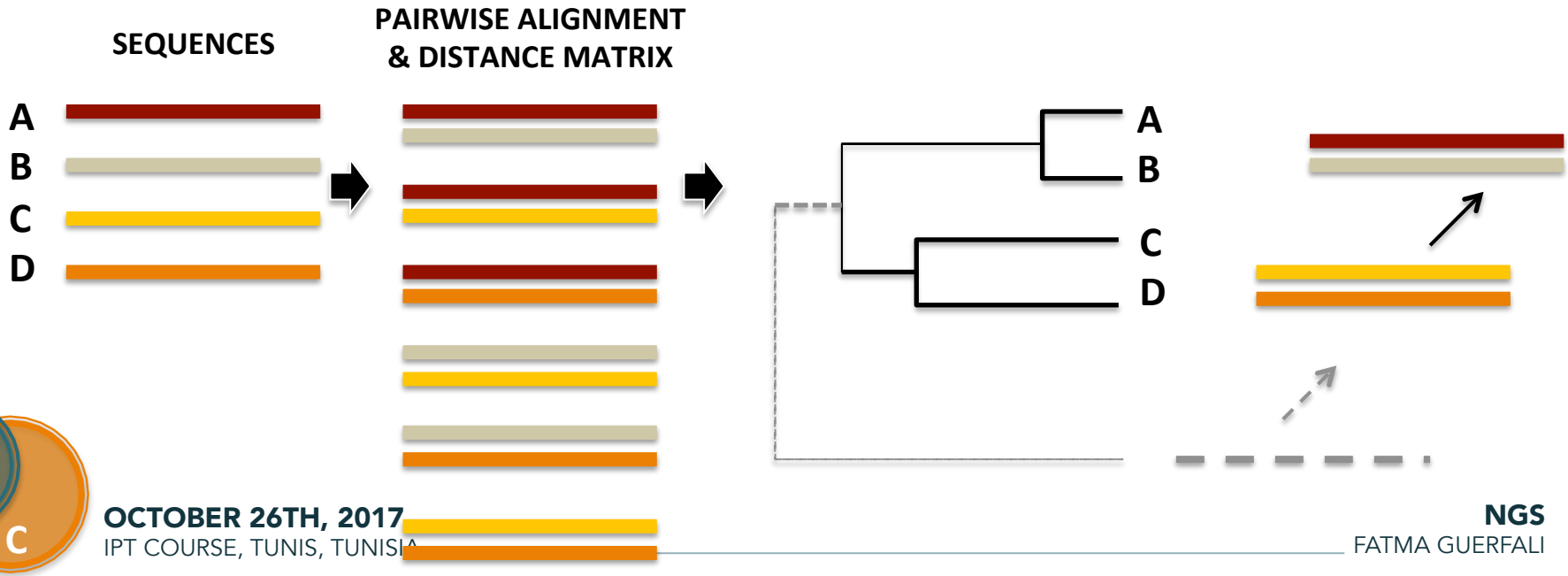
QUERY

ATCTCGTATG

DATABASE

ATCATTCGTTG
ATTCGTTGACT
TCGTTGACTGT
CGTTGACTGTG

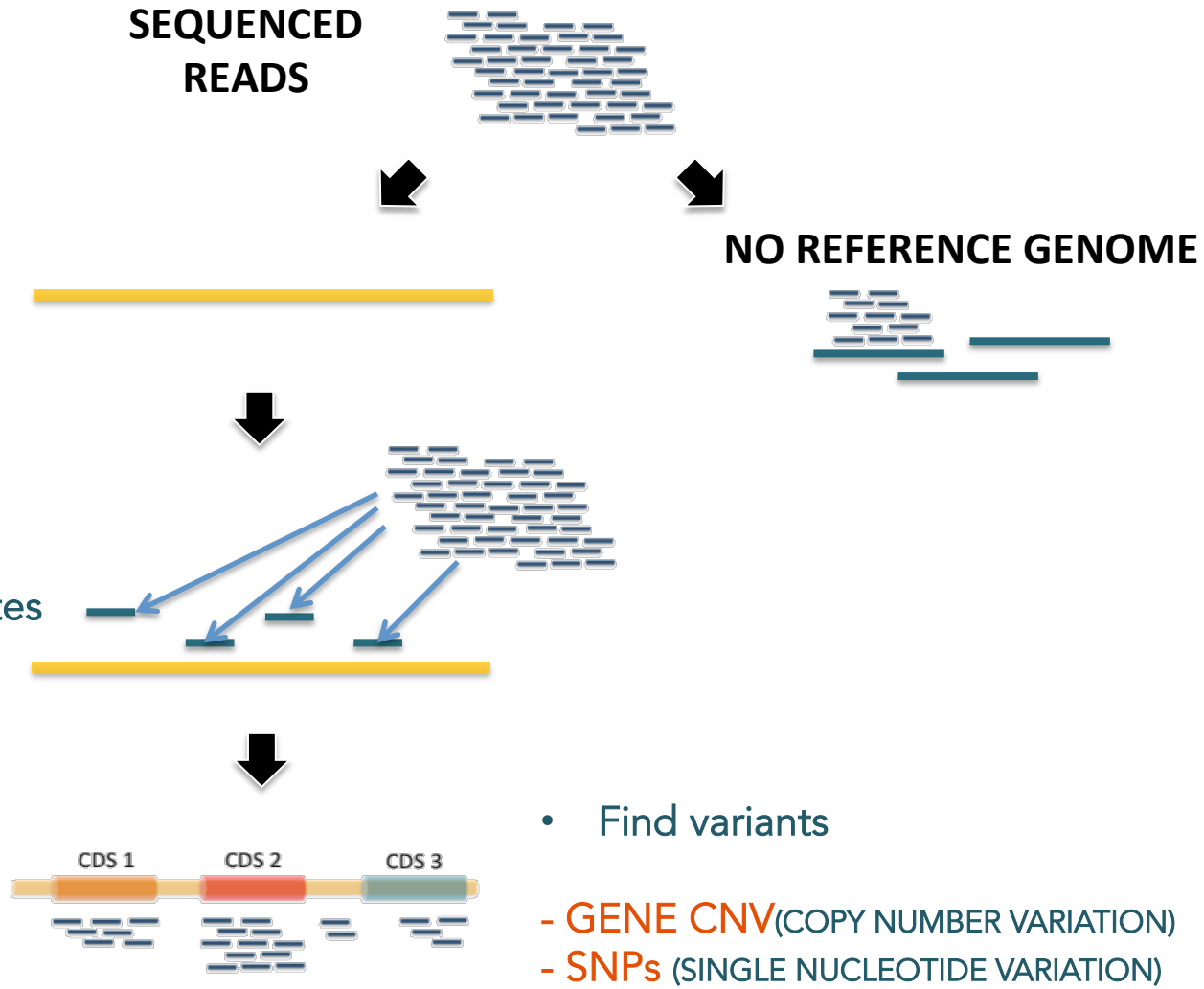
Multiple query sequences



Now how to do that with millions of sequences ?

NGS PROTOCOLS

FILE FORMATS



- Find best match
- Find genomic coordinates
- Find chromosome

→ FASTA file

→ GFF file

- Find variants

- GENE CNV (COPY NUMBER VARIATION)
- SNPs (SINGLE NUCLEOTIDE VARIATION)
- InDels (INSERTION / DELETION)

…

FASTA / GFF Files

FASTA

- « > » sequence

GFF (Generic Feature Format)

- standard file format for storing genomic features in a text file.
- 9 column, tab-delimited files.
- Versions exist. GFF3 files can include sequence in FASTA format at the end of the file. GFF4 adds a reference to any feature using an integer (unique ID).

```
## gff-version 4
## sub-version 1.02
## generated: 2015-02-01
## sequence-region   chr1 1 2097228
chrX    Coding_transcript   intron 14192   14266   .   -
gene=Gene00071  FeatureID=125731789
```

NGS PROTOCOLS

FILE FORMATS

READS « MAPPING » & « MAPPABILITY »

Comparing the DNA of the sequenced sample to its reference sequence:

→ we need to find the corresponding part of that sequence for each read in our sequencing data.

This is called aligning or mapping the reads against the reference sequence.

Once this is done, we can look for variation (e.g. SNPs) within the sample.

https://en.wikibooks.org/wiki/Next_Generation_Sequencing_(NGS)/Alignment

NATURE REVIEWS | GENETICS

(Sims et al., 2014)

READS « MAPPING » & « MAPPABILITY »

This poses a number of problems:

- The short reads do not come with position information.

- The reference sequence can be quite long (~3 billion bases for human), making it a daunting task to find a matching region.

- Reads are short → may be multiple matches (especially true for repetitive regions).

- Impossible to look only for perfect matches to the reference otherwise we would never see any variation → need to allow some mismatches and small structural variation (InDels) in our reads.

https://en.wikibooks.org/wiki/Next_Generation_Sequencing_(NGS)/Alignment

NATURE REVIEWS | GENETICS

(Sims et al., 2014)

READS « MAPPING » & « MAPPABILITY »

Even the best mapping algorithms cannot align all reads to a reference genome !
→ Any sequencing technology produces errors. Similar to the "real" variation, we need to tolerate a low level of sequencing errors in our reads, and separate them from the "real" variation later.

Errors can be due to :

- PCR artifacts (*PCR duplicates…)*
- Sequencing errors (*often random, can be filtered out as singleton reads…*)
- Mapping errors (around repeats (homopolymers…) or other low-complexity regions.
- Structural rearrangements or insertions in the query genome, or deletions in the reference

→ it is not possible to unambiguously assign reads to all genomic regions.

https://en.wikibooks.org/wiki/Next_Generation_Sequencing_(NGS)/Alignment

NATURE REVIEWS | GENETICS

(Sims et al., 2014)

NGS PROTOCOLS

FILE FORMATS

READS « MAPPING » & « MAPPABILITY »

→ Important to assess the uniformity of coverage

Genomic alignment and assembly can result in :

- regions of the genome that lack coverage
  - Gaps
  - GC-rich regions
- regions with much higher coverage than theoretically expected.

→ Analyses of unmapped reads are often used for the identification of structural variants and non-reference insertions.

→ *de novo* Sequencing: critical ! (hybrid approaches)

(Sims et al., 2014)

READS « MAPPING » & « MAPPABILITY »

The NGS reads (RNA/genomic DNA) resulting from the high-throughput sequencing are mapped to a reference sequence.

- If a “reference” genome exists for the organism you are sequencing, reads can be “aligned” to the reference !
- This involves finding the place in the reference genome that each read matches to!
- Due to high sequence similarity within members of the same species, most reads should map to the reference.

Efficient mapping of short reads to a large reference sequence has remained a considerable computational challenge, spurring the development of dozens of alignment algorithms

NGS PROTOCOLS

DEFINITIONS

READS « MAPPING » & « MAPPABILITY »

Bowtie (Langmead et al., 2009), BWA (Li & Durbin, 2009) , SOAP2 (Li et al., 2009) : have leveraged the Burrows–Wheeler transformation (BWT) algorithm to dramatically decrease alignment time (~20 million reads) in hours ≠ Maq or Novoalign (several days).

Aligner	Description	URL
Illumina platform		
ELAND	Vendor-provided aligner for Illumina data	http://www.illumina.com
Bowtie	Ultrafast, memory-efficient short-read aligner for Illumina data	http://bowtie-bio.sourceforge.net
Novoalign	A sensitive aligner for Illumina data that uses the Needleman–Wunsch algorithm	http://www.novocraft.com
SOAP	Short oligo analysis package for alignment of Illumina data	http://soap.genomics.org.cn/
MrFAST	A mapper that allows alignments to multiple locations for CNV detection	http://mrfast.sourceforge.net/
SOLiD platform		
Corona-lite	Vendor-provided aligner for SOLiD data	http://solidsoftwaretools.com
SHRiMP	Efficient Smith–Waterman mapper with colorspace correction	http://compbio.cs.toronto.edu/shrimp/
454 Platform		
Newbler	Vendor-provided aligner and assembler for 454 data	http://www.454.com
SSAHA2	SAM-friendly sequence search and alignment by hashing program	http://www.sanger.ac.uk/resources/software
BWA-SW	SAM-friendly Smith–Waterman implementation of BWA for long reads	http://bio-bwa.sourceforge.net
Multi-platform		
BFAST	BLAT-like fast aligner for Illumina and SOLiD data	http://bfast.sourceforge.net
BWA	Burrows-Wheeler aligner for Illumina, SOLiD, and 454 data	http://bio-bwa.sourceforge.net
Maq	A widely used mapping tool for Illumina and SOLiD; now deprecated by BWA	http://maq.sourceforge.net

(Koboldt et al., 2010)

NGS PROTOCOLS

DEFINITIONS

READS « MAPPING » & « MAPPABILITY »

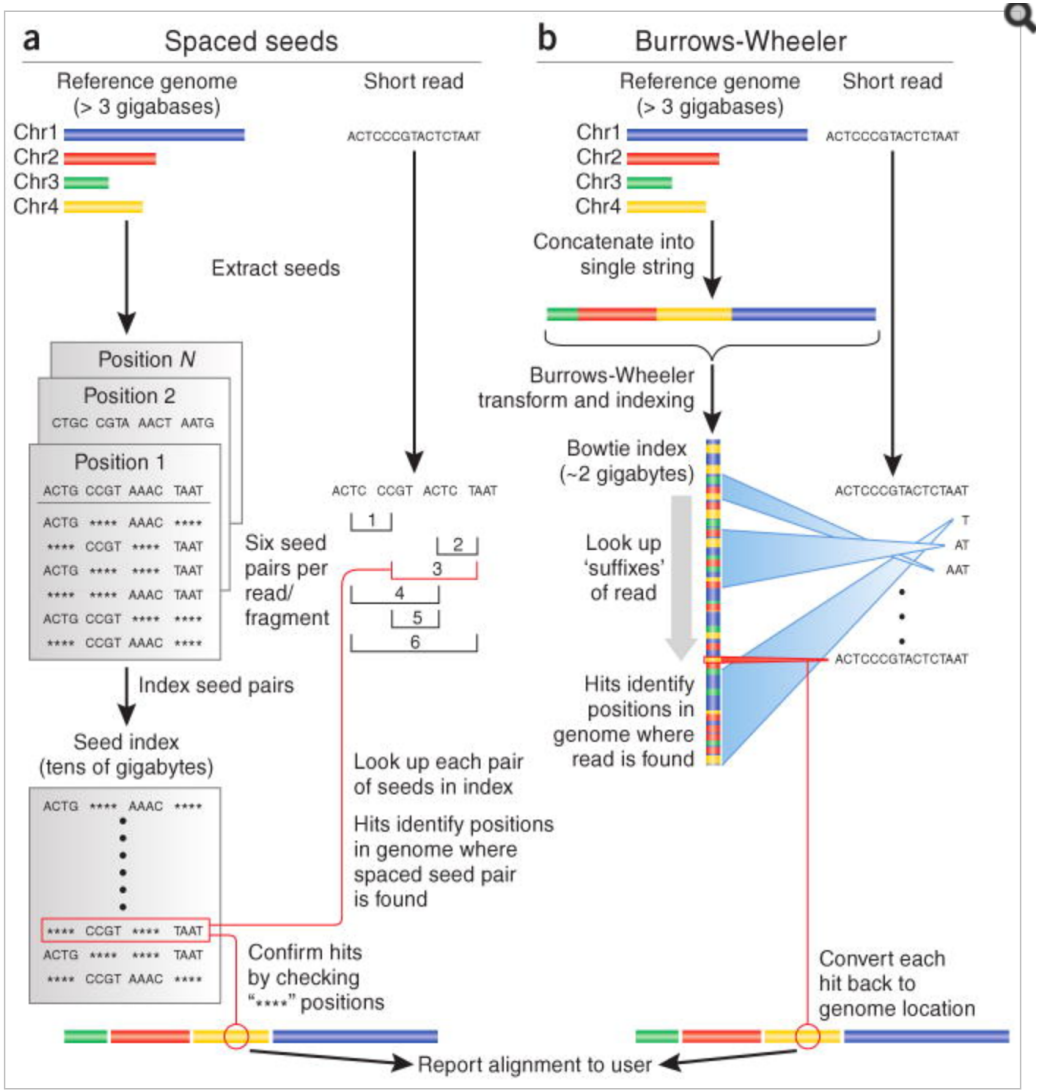
2 different algorithmic approaches for aligning short (20–200-bp) sequencing reads.

Spaced seeds (Maq):

- each position in the reference is cut into equal-sized pieces, called 'seeds'
- seeds are paired and stored in a lookup table.
- Each read is also cut up according to this scheme
- pairs of seeds are used as keys to look up matching positions in the reference.

Burrows-Wheeler (Bowtie)

- Store a memory-efficient representation of the reference genome.
- Align Reads character by character from right to left against the transformed string.
- When all characters in the read have been processed, alignments are represented by any positions within the interval.
- Faster algorithms mainly due to the memory efficiency of Burrows-Wheeler search.



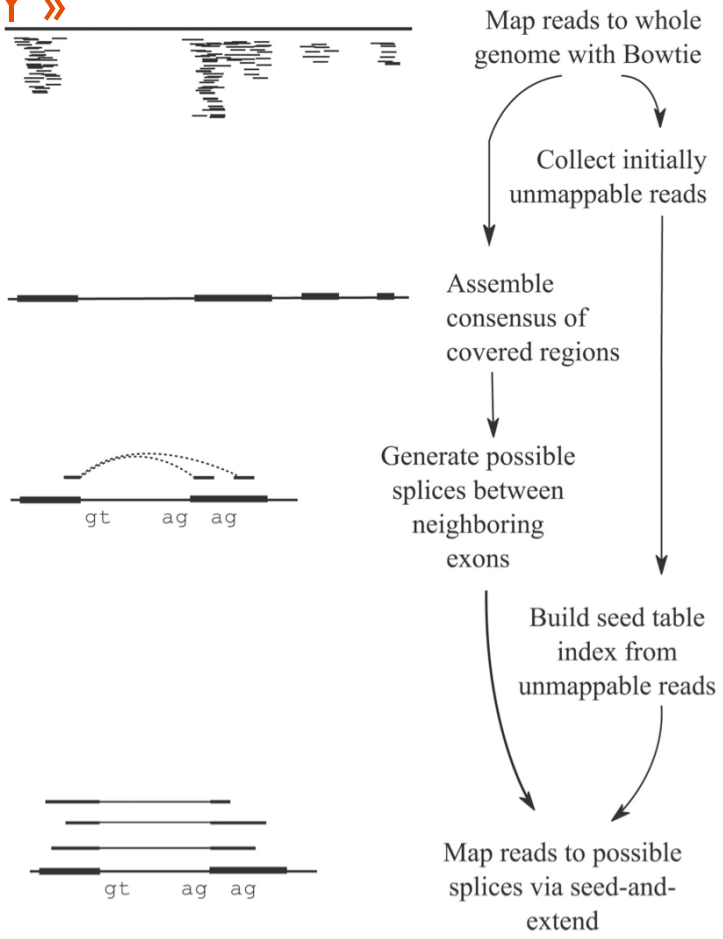
(Trapnell & Salzberg, 2009)

READS « MAPPING » & « MAPPABILITY »

Mapping to the genome achieves two major objectives of RNA-Seq experiments:

- Identification of novel transcripts from these regions covered in the mapping.
- Estimation of the abundance of the transcripts from their depth of coverage in the mapping.

Because RNA-Seq reads are short, the first task is challenging



(Trapnell, Pachter and Salzberg, 2009)

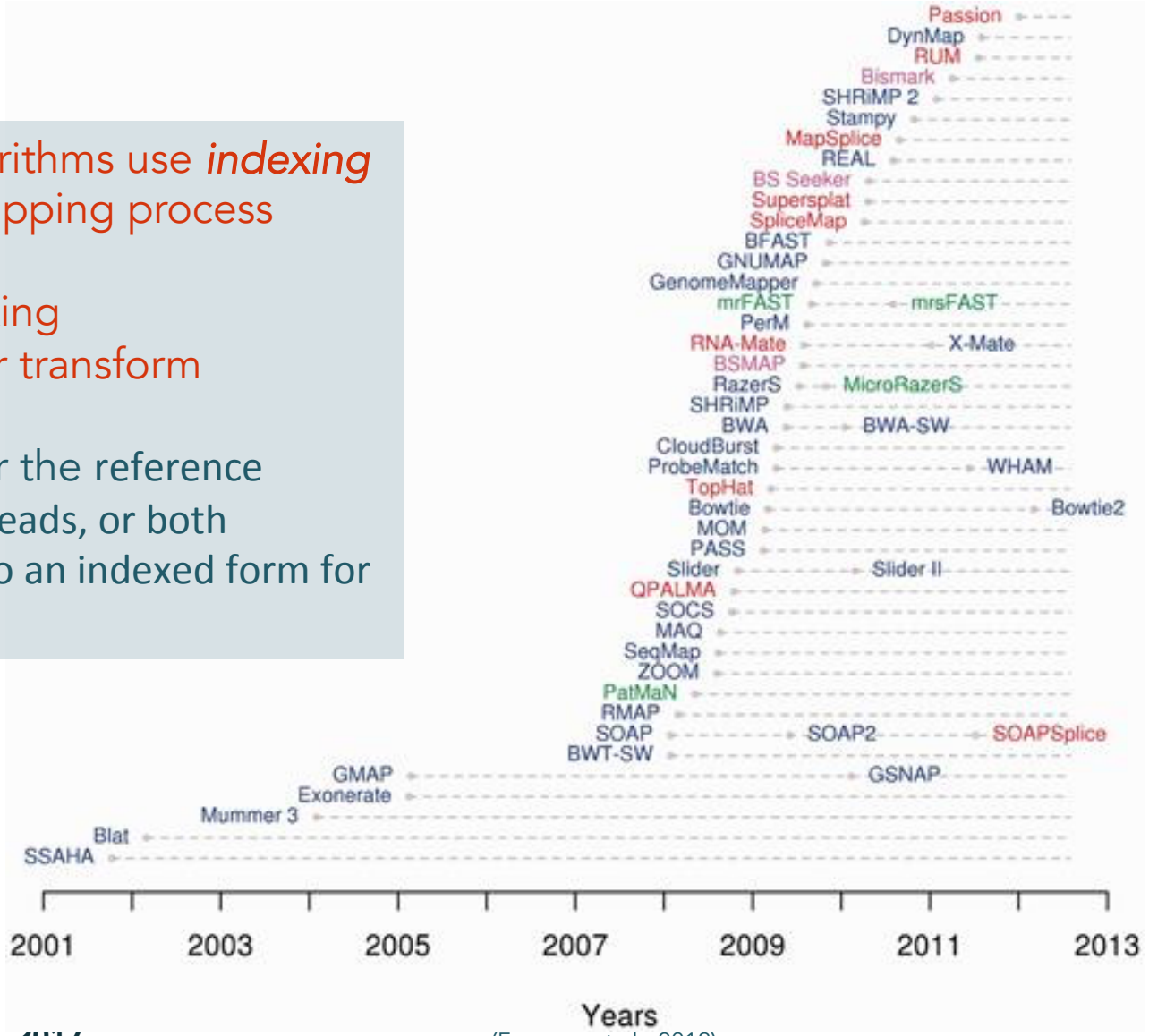
NGS PROTOCOLS

FILE FORMATS

Improvement: Algorithms use *indexing* to speed up the mapping process

- Hash table indexing
- Burrows-Wheeler transform

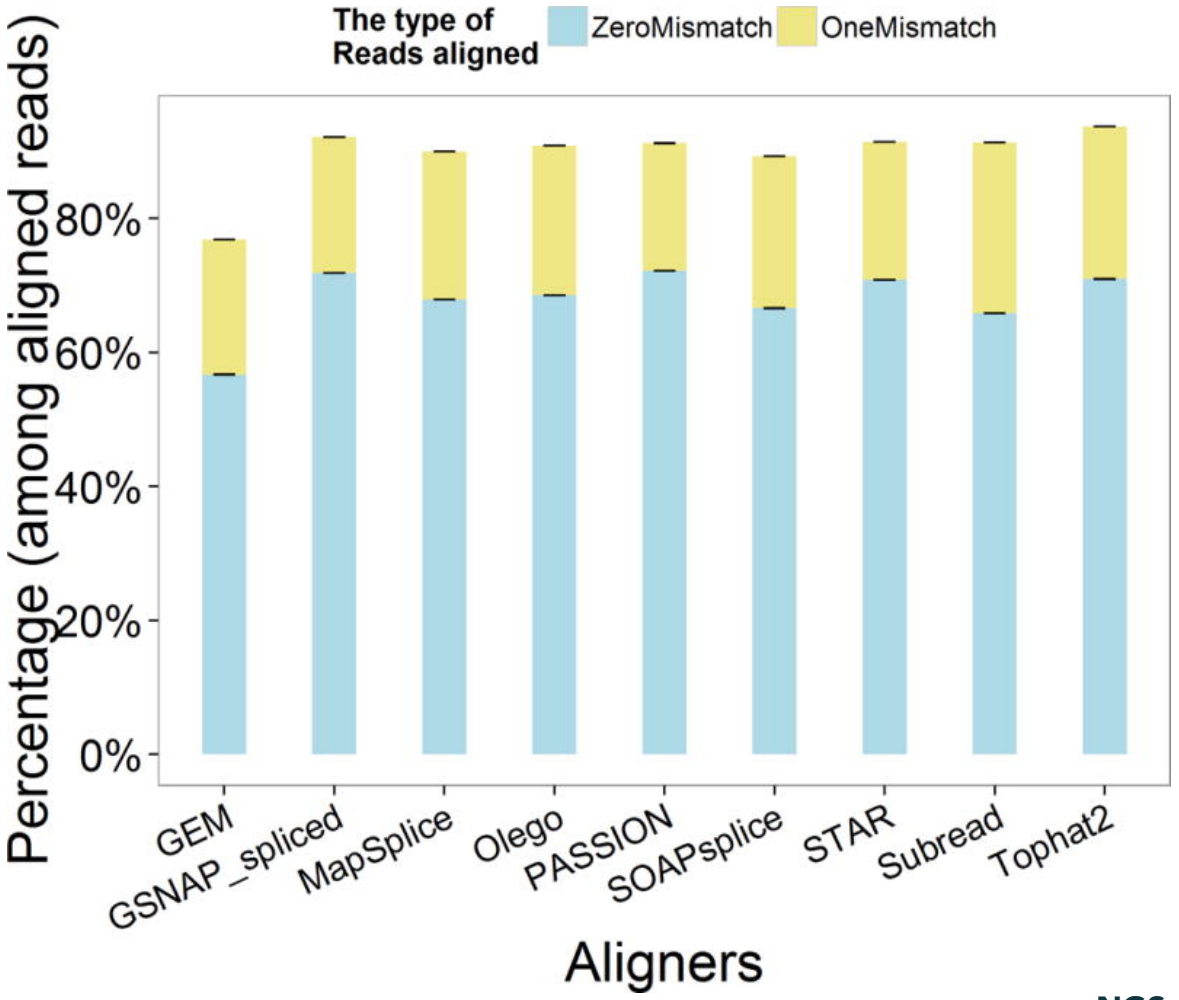
Indexing is done for the reference sequence, the short reads, or both
→ pre-processing into an indexed form for faster search



(Fonseca et al., 2012)

Impact of RNA-seq aligners on gene expression estimation

« While numerous RNA-seq data analysis pipelines are available, research has shown that the choice of pipeline influences the results of differentially expressed gene detection and gene expression estimation. »



(Yang et al., 2015)

NGS PROTOCOLS

FILE FORMATS

Choose the correct mapper:

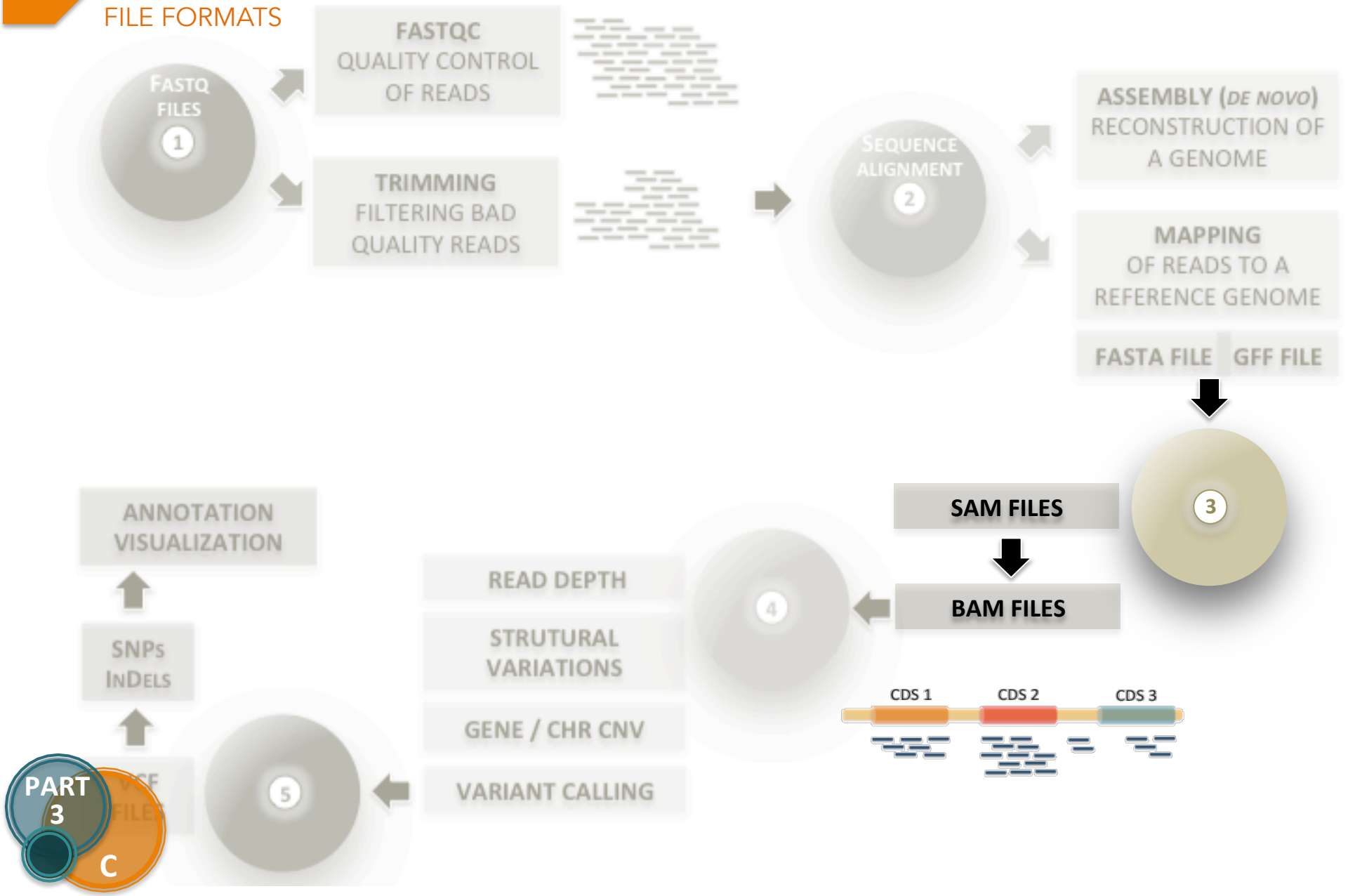
- Does it support my sequencing platform? (read length…)
- Does it handle PE data?
- Does it allow mismatches? (SNPs)
- Does it allow gapped alignment? (InDels)
- How does it deal with multiple matches? (user's choice)
- How does it deal with spliced reads? (reads spanning exon-exon boundaries)

"… there is no tool that outperforms all of the others in all the tests. Therefore, the end user should clearly specify his needs in order to choose the tool that provides the best results."

*(Hatem et al BMC Bioinformatics 2013, **14**:184)*

NGS PROTOCOLS

FILE FORMATS



PART 3 C

SAM / BAM FILES

SAM file (.sam) = "Sequence Alignment/Map format"

Tab-delimited text file format that can store information about alignment (mapped, unmapped) and even QC-failed reads…All this information can be easily retrieved (samtools).

But SAM takes to much space…

BAM file (.bam) is the binary version of a SAM file

NB: Due to the large data volumes being generated and stored, the EBI has designed CRAM files as alignment files (like BAM files, but represent a compressed version of the alignment, driven by the reference the sequence data is aligned to).

NGS PROTOCOLS

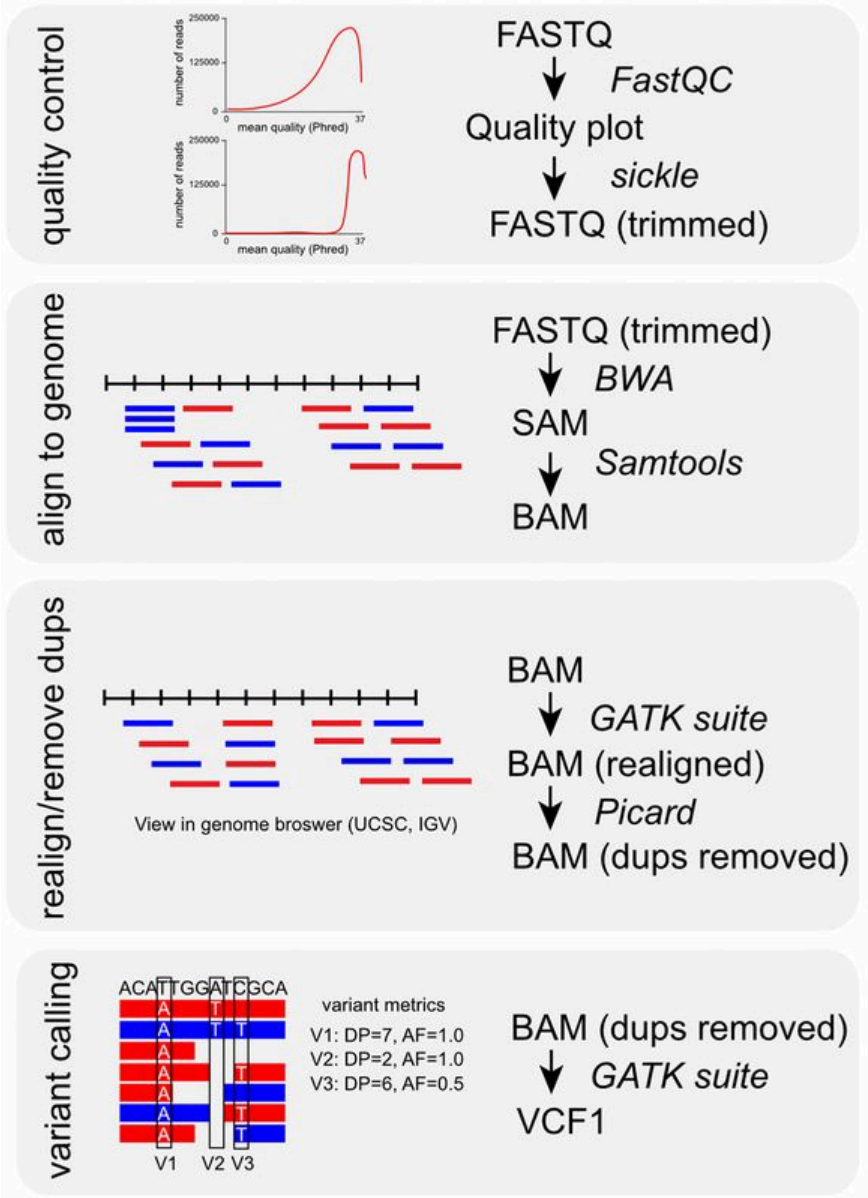
FILE FORMATS

The mapping tools calculate a probability for the correctness of the alignment for the whole read.

```
coor      12345678901234    567890123456789012345 6
ref       aggttttataaaac----aattaagtctacagagcaacta
sample    aggttttataaaacAAATaattaagtctacagagcaacta
read1     aggttttataaaac****aaAtaa
read2      ggttttataaaac****aaAtaaTt
read3          ttataaaacAAATaattaagtctaca
read4              CaaaT****aattaagtctacagagcaac
read5                aaT****aattaagtctacagagcaact
read6                  T****aattaagtctacagagcaacta
```

Some bases can be misaligned !
Approaches to this problem :

- GATK ralignment = solve it by realigning the problematic regions
- Samtools = detect it and mark it with alignment qualities per base and not only per read. Resulting qualities are known as BAQ (Base Alignment Quality) (see mpileup).



INDEXES

Some programs require that for faster access we need a companion file (often called index) for the different formats

FASTA (.fa & .fai)

BAM and BAI formats (suffixes .bam & .bai)

VCF (.vcf & .vcf.idx)

NGS PROTOCOLS

FILE FORMATS

Sequence ID	Flag	Chr	Position	Map Qual	Cigar	Paired end info		
HWI-ST1136:196:HS113:4:1101:4333:28021	163	chr2	217279469	255	100M	=	217279487	117
HWI-ST1136:196:HS113:4:1101:4333:28021	83	chr2	217279487	255	99M1S	=	217279469	-117
HWI-ST1136:196:HS113:4:1101:4320:28039	163	chr11	65271253	255	100M	=	65271335	182
HWI-ST1136:196:HS113:4:1101:4320:28039	83	chr11	65271335	255	100M	=	65271253	-182
HWI-ST1136:196:HS113:4:1101:4274:28047	99	chr4	763497	255	100M	=	763607	210
HWI-ST1136:196:HS113:4:1101:4274:28047	147	chr4	763607	255	100M	=	763497	-210
HWI-ST1136:196:HS113:4:1101:4333:28054	99	chr17	74433086	255	100M	=	74433100	114
HWI-ST1136:196:HS113:4:1101:4333:28054	147	chr17	74433100	255	100M	=	74433086	-114
HWI-ST1136:196:HS113:4:1101:4353:28065	99	chr11	62293812	255	100M	=	62293909	197
HWI-ST1136:196:HS113:4:1101:4353:28065	147	chr11	62293909	255	100M	=	62293812	-197

...

The following table gives an overview of the mandatory fields in the SAM format:

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
2	FLAG	Int	$[0, 2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0, 2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0, 2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	$[0, 2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1, 2^{31}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

SAM file format

SAM= TEXT file, BAM= SAM file compressed and indexed (binary) format

```
@SQ     SN:chr9_random  LN:449403
@SQ     SN:chrM LN:16299
@SQ     SN:chrUn_random LN:5900358
@SQ     SN:chrX LN:166650296
@SQ     SN:chrX_random  LN:1785075
@SQ     SN:chrY LN:15902555
@SQ     SN:chrY_random  LN:58682461
HWI-EAS038:6:1:23:122#0 4       *       0       0       *       *       0       0       TAGCCTTGATGTTTACCTATTGTATCAAAGGGC        OJYMKLTPKOPOXYBBBBBBBBBBBBBBBBBBB
B
HWI-EAS038:6:1:25:283#0 0       chr14   27002726        0       33M     *       0       0       AGAGACCCAGGAAATTGAAGTCAGAGCAGTTAG        abaa_Z_`X]PW^BBBBBBBBBB
BBBBBBBBB       XT:A:R  NM:i:1  X0:i:3  X1:i:0  XM:i:1  XO:i:0  XG:i:0  MD:Z:18T22
HWI-EAS038:6:1:26:649#0 0       chr9    27884899        37      33M     *       0       0       CCTTTCTTTTGTCTACTCCTTTCCTCTTGGTAT        abbaabbbbbb```aZ\`a\aa[]
_QWaa`YX5       XT:A:U  NM:i:0  X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:33
HWI-EAS038:6:1:30:918#0 16      chr17   95265601        0       33M     *       0       0       GTGTTTATCAGTCCCAAGGCCACTAGAGGCTTG        BBBBBBBBBBBBBBB[[`\aaZaa
__aaaa`a`       XT:A:R  NM:i:2  X0:i:3  X1:i:0  XM:i:2  XO:i:0  XG:i:0  MD:Z:3G8T20
HWI-EAS038:6:1:32:1507#0        16      chr13   57505480        37      33M     *       0       0       CGGAGCTGGTGGTAGACATTGTGTGCTGCCTAG        \Z]W_[`]Z^[]^^Z^T
`bbab_[W\_]bb_W_b       XT:A:U  NM:i:0  X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:33
HWI-EAS038:6:1:32:298#0 4       *       0       0       *       *       0       0       TATAATAAAAATGACATTTTATTAAATACGCCT        `^aaa__\]^[5BBBBBBBBBBBBBBBBBBBB
B
HWI-EAS038:6:1:32:1938#0        0       chr7    65636851        37      33M     *       0       0       TTTATATTTCTCCCCTTATCATTCCATTTTTTT        ]aa^^X\`YQ\Y[^UY
ZHMHXWZEVFO][8888       XT:A:U  NM:i:1  X0:i:1  X1:i:0  XM:i:1  XO:i:0  XG:i:0  MD:Z:31G1
HWI-EAS038:6:1:32:861#0 4       *       0       0       *       *       0       0       TGCATTCTAAGTTGGTTTAATATAAATCAACAT        ]bUSJGKHWaK_\BBBBBBBBBBBBBBBBBBB
B
HWI-EAS038:6:1:32:1814#0        0       chr2    98586740        0       33M     *       0       0       CCACTTGACGACTTCAAAAATGACGAAATCACT        W^R^X`]Z]a]XZ]aZ
W]PYVV\YRW[SUZSST       XT:A:R  NM:i:1  X0:i:12 X1:i:44 XM:i:1  XO:i:0  XG:i:0  MD:Z:14G18
HWI-EAS038:6:1:34:2002#0        0       chr10   97252488        37      33M     *       0       0       CCTAGATTCCTTAGGGTATAAAAGGAGGAGAGC        _a`_ba]_Oa]aV[`a
OHDT^_BBBBBBBBBB        XT:A:U  NM:i:1  X0:i:1  X1:i:0  XM:i:1  XO:i:0  XG:i:0  MD:Z:29T3
HWI-EAS038:6:1:37:667#0 0       chrX    90652654        37      33M     *       0       0       CAAGTCCAAAAATTCCTTGAAAAATTTCACAAT        Y`_TOMPT^^_[PUWOJQLQQYW]
BBBBBBBBB       XT:A:U  NM:i:1  X0:i:1  X1:i:0  XM:i:1  XO:i:0  XG:i:0  MD:Z:19C13
HWI-EAS038:6:1:37:1236#0        4       *       0       0       *       *       0       0       ATGATTTCTTGTTGTGTATCACTATTCTAGGGG        _Q\LYBBBBBBBBBBBBBBBBBBB
BBBBBBBBB
HWI-EAS038:6:1:37:262#0 16      chr2    3386587 23      33M     *       0       0       TCTAGTACCCACATGGTGCAAGGAGAGAACCAA        BB]Z[LFTXX]TZYQRXHJUOISL^X]_[UO]
a       XT:A:U  NM:i:1  X0:i:1  X1:i:1  XM:i:1  XO:i:0  XG:i:0  MD:Z:6C26
HWI-EAS038:6:1:38:385#0 0       chr9    35113013        25      33M     *       0       0       AAAAAACGTGAAAAATAAGAAATGCCAACTGAA        [aa``_]PTUUZY[_[R]BBBBBB
BBBBBBBBB       XT:A:U  NM:i:2  X0:i:1  X1:i:0  XM:i:2  XO:i:0  XG:i:0  MD:Z:16G9C6
HWI-EAS038:6:1:38:37#0  16      chr16   49998240        37      33M     *       0       0       ATTTGTCTGTGATGATTTTCTGTTCTTTCAATG        B[_XHJJJTMPWWNR__`]__Wa^
`^`R`]a_a       XT:A:U  NM:i:0  X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:33
HWI-EAS038:6:1:40:991#0 16      chr13   75619559        0       33M     *       0       0       TTTAATATTCTATCTTTATTTAGTGCATTTGTT        a_ZQPX`__[`RY[]\]PVT]\\]
WXU[]`V_^       XT:A:R  NM:i:0  X0:i:6619       XM:i:0  XO:i:0  XG:i:0  MD:Z:33
HWI-EAS038:6:1:40:767#0 0       chr11   34713793        25      33M     *       0       0       TAACTTATTCCTTTAGGTCCTGTGTTTTCTATT        aaa0`]aQYBBBBBBBBBBBBBBBB
```


https://image.slidesharecdn.com/day1s2barte-130308084536-phpapp02/95/chipseq-data-analysis-35-638.jpg?cb=1362733593

Mapped and unmapped reads are imported into SAM/BAM format

The standard CIGAR description of pairwise alignment defines three operations:
'M' for **match/mismatch**, 'I' for insertion compared with the reference and 'D' for deletion.

(NB: The POS indicates that the read aligns starting at position 5 on the reference)

The CIGAR :
3M = 3 bases in the read sequence align with the reference.
1I = The next base in the read does not exist in the reference.
1D = The reference base does not exist in the read sequence

```
RefPos:      1  2  3  4  5  6  7  8  9 10 11 12 13
Reference:   C  C  A  T  A  C  T  G  A  A  C  T  G
Read: ACTAGAATG
```

```
RefPos:      1  2  3  4  5  6  7     8  9 10 11 12 13
Reference:   C  C  A  T  A  C  T     G  A  A  C  T  G
Read:                  A  C  T  A  G  A  A     T  G
```

POS: 5
CIGAR: 3M1I3M1D2M

samtools index test.bam
samtools view test.bam chr1:200000-500000

Mapped and unmapped reads are imported into SAM/BAM format

SAMTools

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

```
#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
```

NGS PROTOCOLS

FILE FORMATS

Mapped and unmapped reads are imported into SAM/BAM format

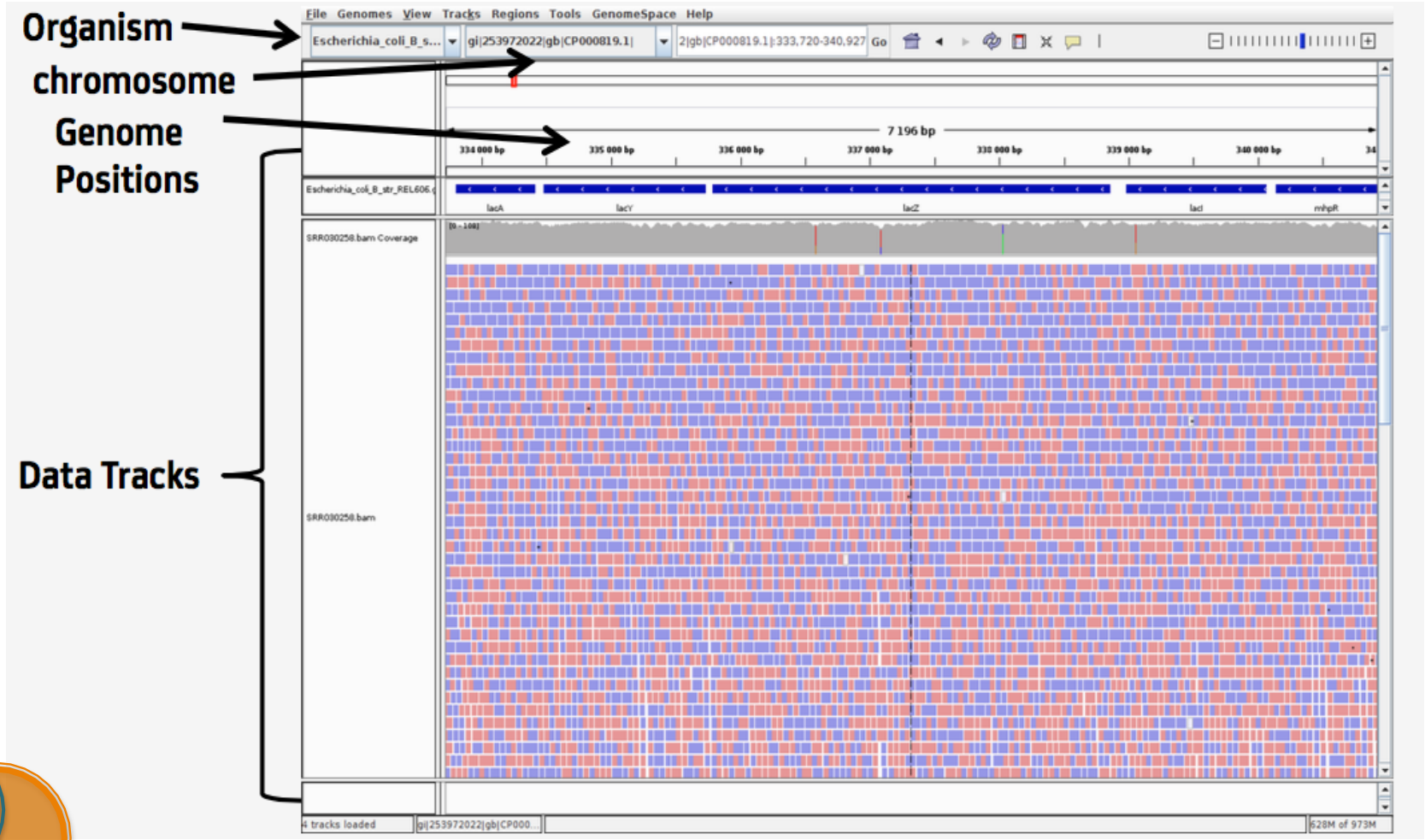
Samtools

The flagstat command provides simple statistics on a BAM file

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

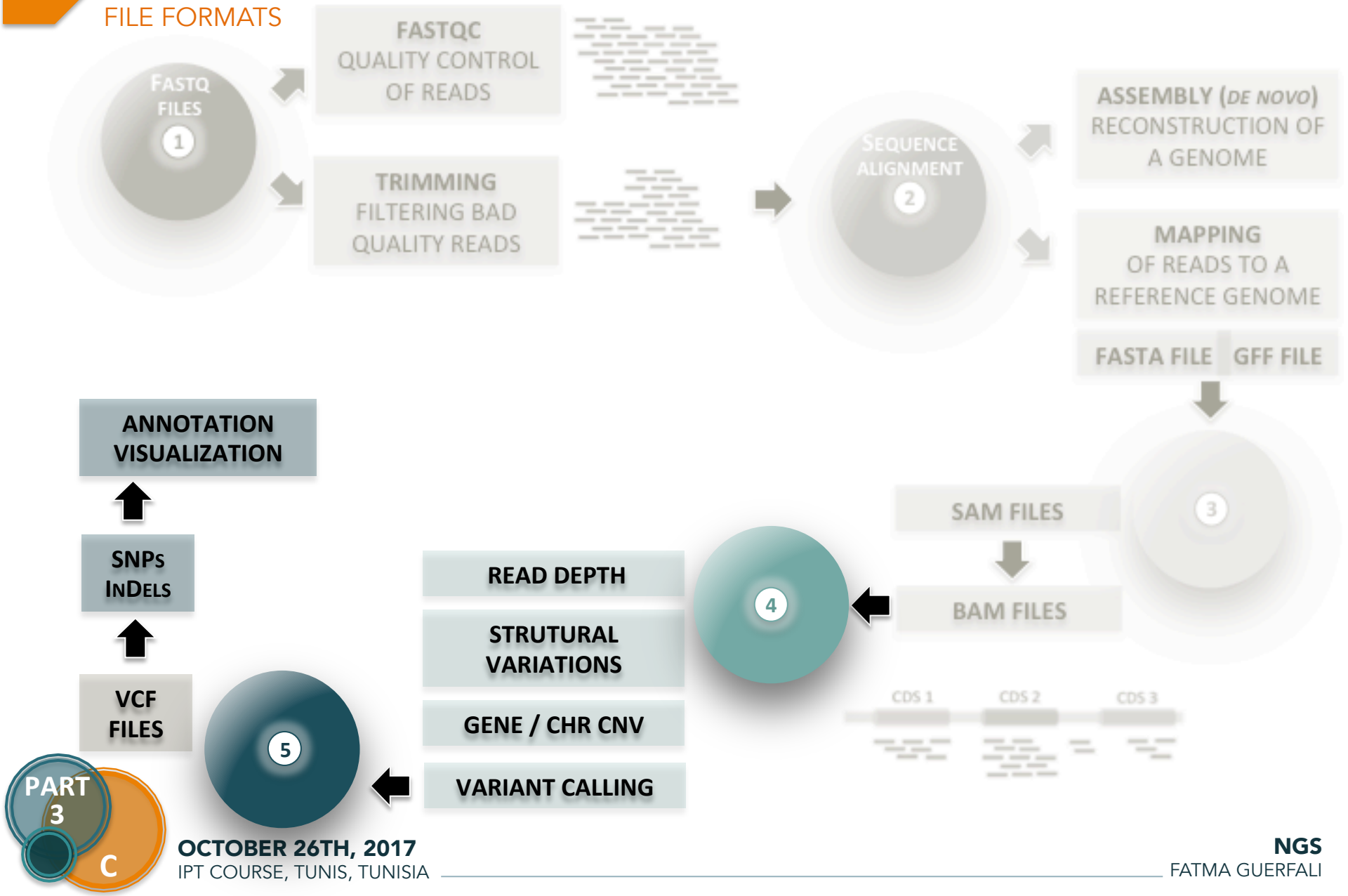
1	6874858 + 0 in total (QC-passed reads + QC-failed reads)
2	90281 + 0 duplicates
3	6683299 + 0 mapped (97.21%)
4	6816083 + 0 paired in sequencing
5	3408650 + 0 read1
6	3407433 + 0 read2
7	6348470 + 0 properly paired (93.14NaV)
8	6432965 + 0 with itself and mate mapped
9	191559 + 0 singletons (2.81NaV)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

IGV : Integrated Genome Viewer



NGS PROTOCOLS

FILE FORMATS



PART 3 C

COVERAGE

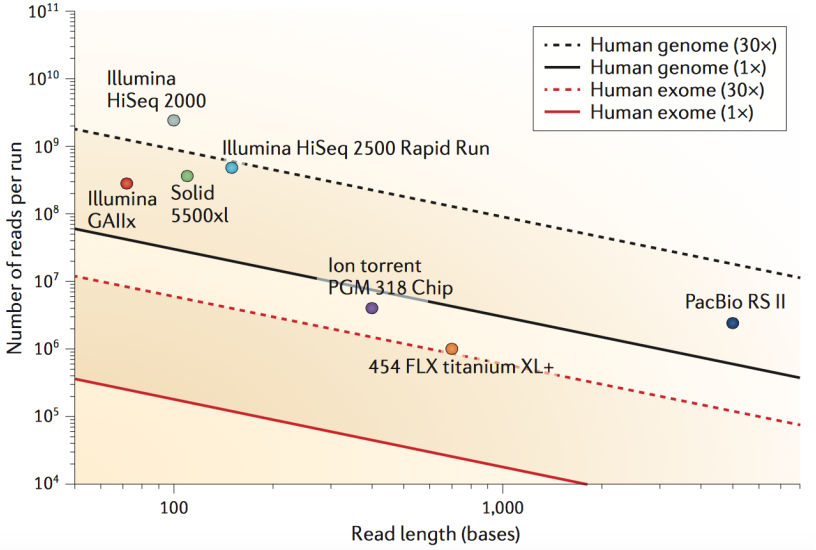
Theoretical redundancy of coverage (Lander & Waterman, 1988)

the average number of times that:

- each nucleotide is expected to be sequenced
- given a certain number of reads of a given length
- and the assumption that reads are randomly distributed across an idealized genome

$$c = LN/G$$

L : the read length
N : the number of reads
G : the haploid genome length



Theoretical coverage (shown as diagonal lines; c = 1× or 30×) according to the Lander–Waterman formula for human genome or exome sequencing.

(Sims et al., 2014)

NATURE REVIEWS | GENETICS

COVERAGE

Redundancy of coverage
= depth or depth of coverage = The number of times a nucleotide is sequenced

Coverage ≈ Depth
Coverage ≈ Breadth of coverage = the percentage of target bases that are sequenced a given number of times

Example of output for a target genome sequencing study:
- Genome sequenced to 30X average coverage
- 95% breadth of coverage of the reference genome
- at a minimum depth of 10 reads

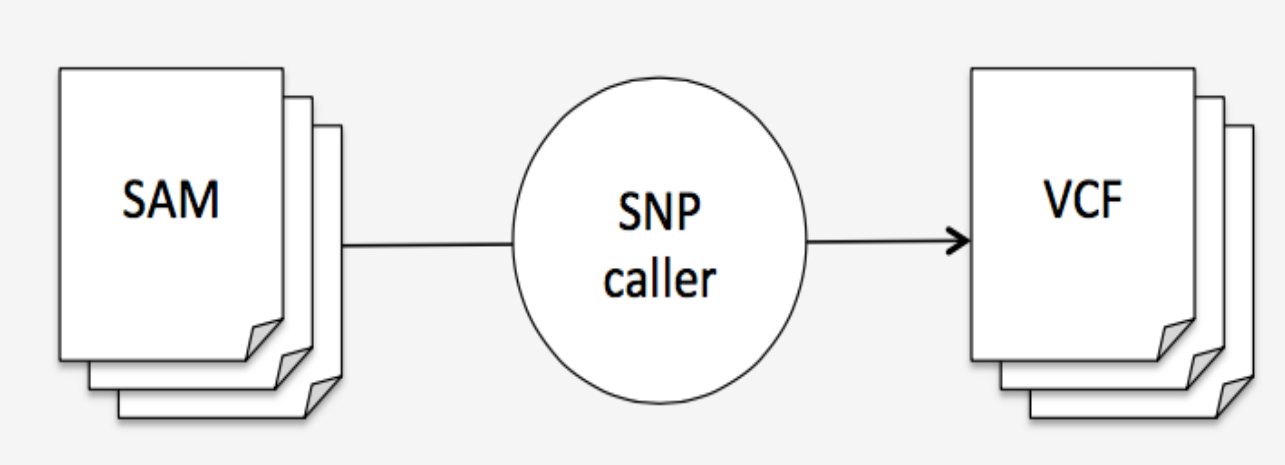
COVERAGE

Regardless of average read depth, depth-of-coverage methods are vulnerable to false positives that are being called owing to local variations in coverage even after correction for both GC bias and 'mappability'.

→ cross-sample calling is required to reduce this effect.

(Sims et al., 2014)

VARIANT CALLING



Many tools for variants detection: GATK, Samtools (mpileup), FreeBayes, PICARD, etc.

VARIANT CALLING

SNP callers are based on:

- Simple method (VarScan SNP caller + most of the commercial SNP callers): counting the number of reads for each alleles once appropriate thresholds for the sequencing and mapping qualities have been applied.

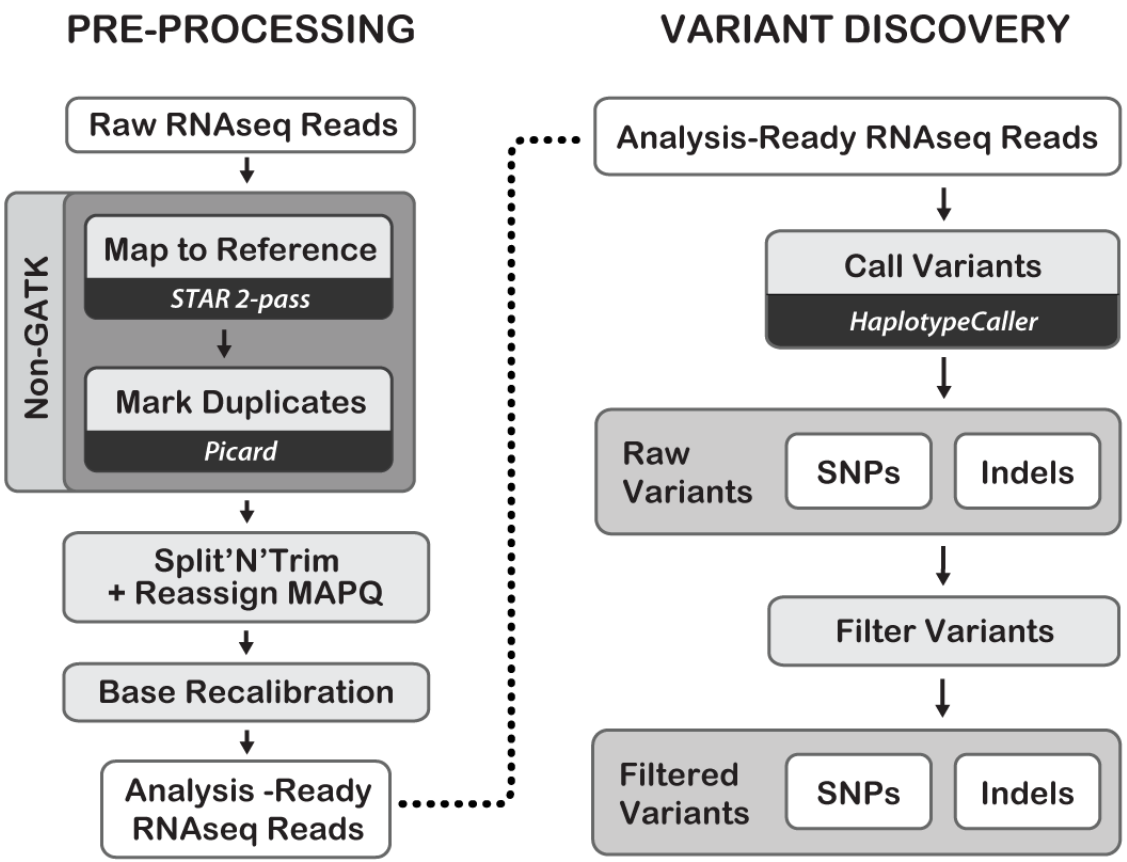
- Methods based on more advanced statistics:
  - Often perform better, specially with low coverages
  - Do assumptions to create bayesian models
  - Most assume diploid individuals
  - Some take into account the Hardy-Weinberg equilibrium and Linkage Disequilibrium information as well as previous information about the SNPs present in the species and their allele frequencies.

VARIANT CALLING

GATK best practices.

Several different Best Practices workflows tailored to particular applications depending on the type of variation of interest and the technology employed.

example : for Germline SNP and Indels Discovery in RNA-seq



Best Practices for Germline SNPs and Indels in RNAseq

VCF FILE

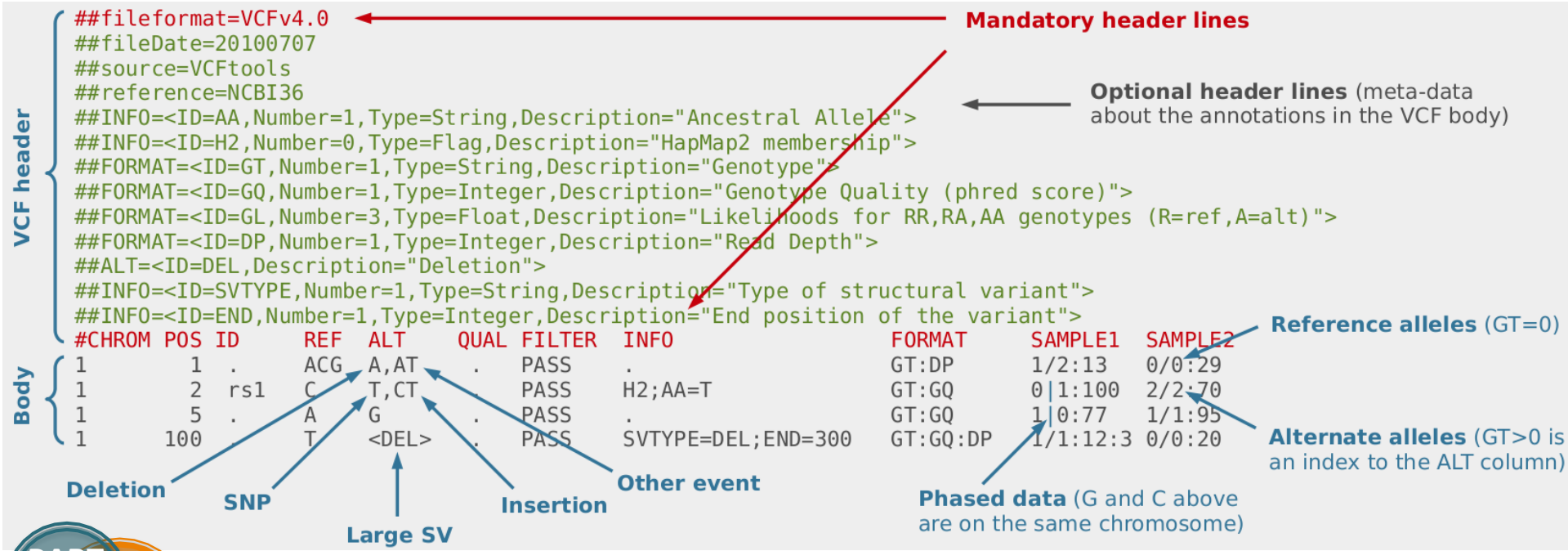
VCF file (.vcf) = "Variant Calling format"

Tab-delimited text file format that can store information about variants…
All this information can be easily queried (VCFtools).

- Format validation.
- SNV annotation.
- VCF comparison.
- Statistics.
- …

VCF FILES

- header lines starting with # describing the rest of the file
- 8 separated tab columns = chromosome, position, the REF base, the ALT base, the QUAL score …



VARIANT CALLING FILTERING:
Variant annotation programs (SnpEff…)

```
$ java -jar snpEff.jar
SnpEff version SnpEff 4.1 (build 2015-01-07), by Pablo Cingolani
Usage: snpEff [command] [options] [files]

Run 'java -jar snpEff.jar command' for help on each specific command

Available commands:
        [eff|ann]                       : Annotate variants / calculate effects (you can use either 'ann' or 'eff
        build                           : Build a SnpEff database.
        buildNextProt                   : Build a SnpEff for NextProt (using NextProt's XML files).
        cds                             : Compare CDS sequences calculated form a SnpEff database to the one in a
        closest                         : Annotate the closest genomic region.
        count                           : Count how many intervals (from a BAM, BED or VCF file) overlap with eac
        databases                       : Show currently available databases (from local config file).
        download                        : Download a SnpEff database.
        dump                            : Dump to STDOUT a SnpEff database (mostly used for debugging).
        genes2bed                       : Create a bed file from a genes list.
        len                             : Calculate total genomic length for each marker type.
        protein                         : Compare protein sequences calculated form a SnpEff database to the one
        spliceAnalysis                  : Perform an analysis of splice sites. Experimental feature.
```

VARIANT CALLING FILTERING:
Variant annotation programs (SnpEff…)

Here is an example of a file before and after being annotated using SnpEff:

VCF file before annotations

```
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO
1      889455  .       G       A       100.0   PASS    AF=0.0005
1      897062  .       C       T       100.0   PASS    AF=0.0005
```

VCF file after being annotated using SnpEff

```
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO
1      889455  .       G       A       100.0   PASS    AF=0.0005;EFF=STOP_GAINED(HIGH|NONSENSE|Cag/Tag|Q236*|
1      897062  .       C       T       100.0   PASS    AF=0.0005;EFF=STOP_GAINED(HIGH|NONSENSE|Cag/Tag|Q141*|
```

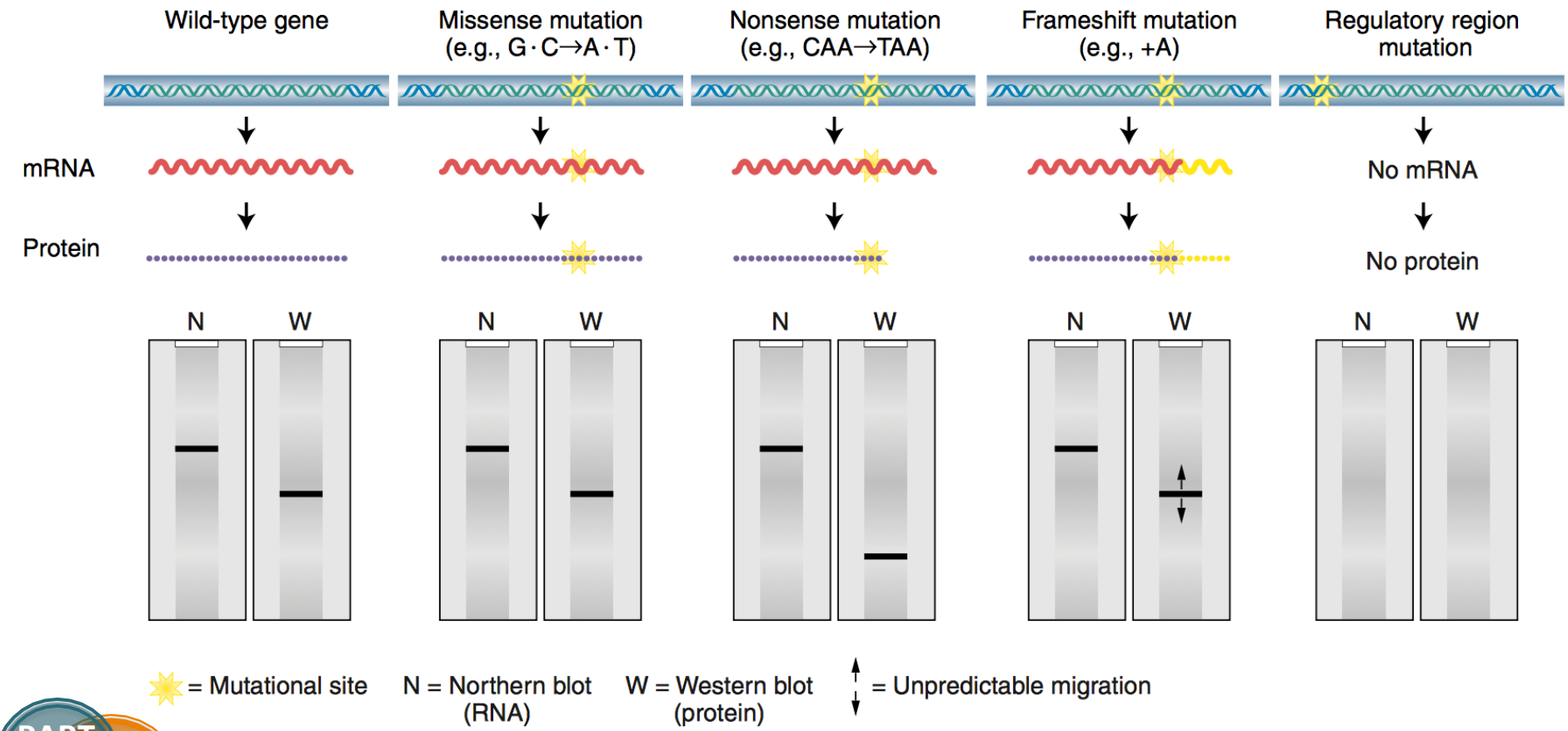
A you can see, SnpEff added an 'EFF' tag to the INFO field (eigth column).

```
|749|NOC2L||CODING|NM_015658|)
|642|KLHL17||CODING|NM_198317|)
```

NGS PROTOCOLS

FILE FORMATS

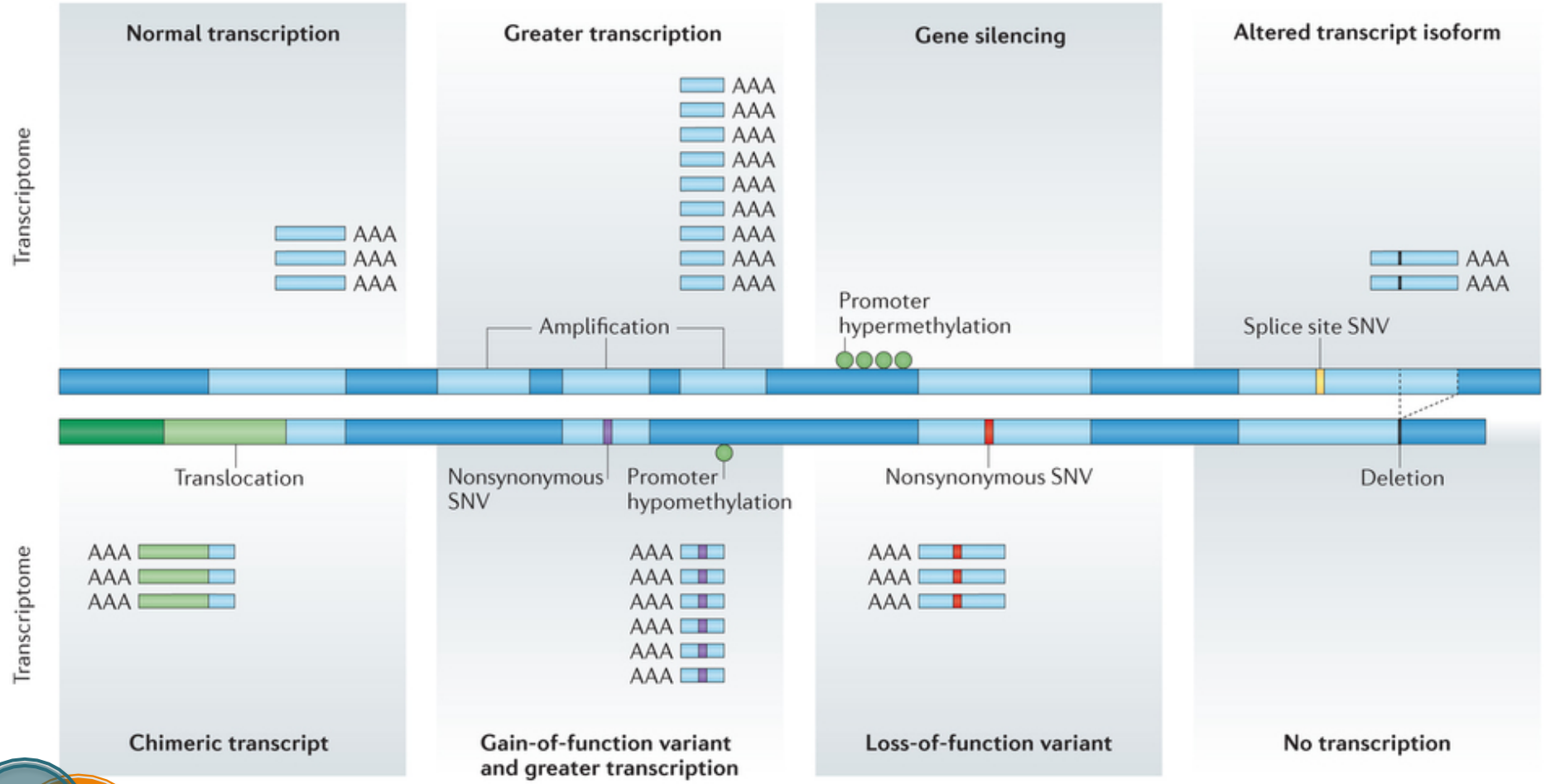
VCF FILES



NGS PROTOCOLS

FILE FORMATS

Integrate transcriptome (± epigenome) with WGS



(Mwenifumbo & Marra, 2013)

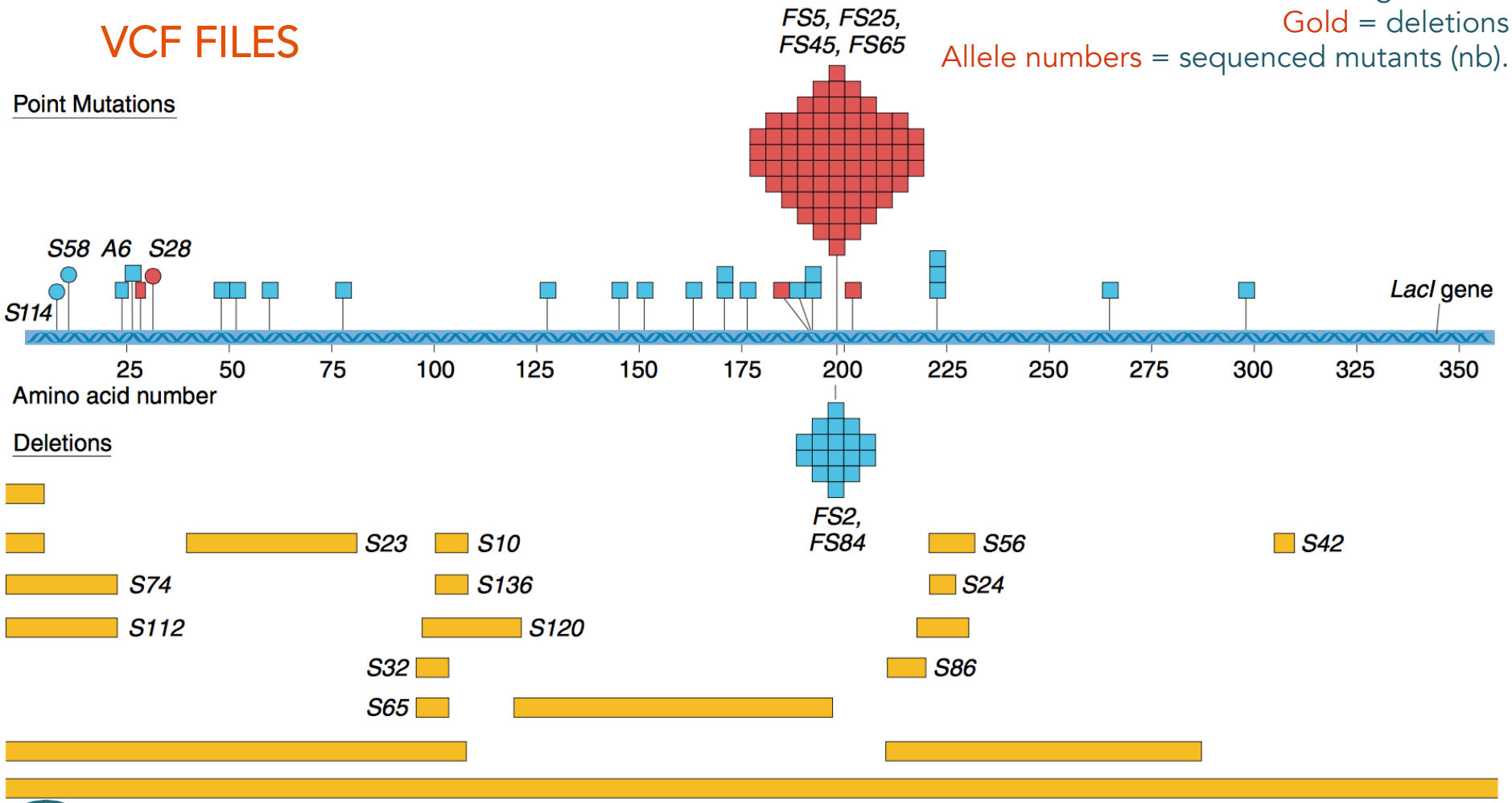
Nature Reviews | Genetics

NGS PROTOCOLS

FILE FORMATS

VCF FILES

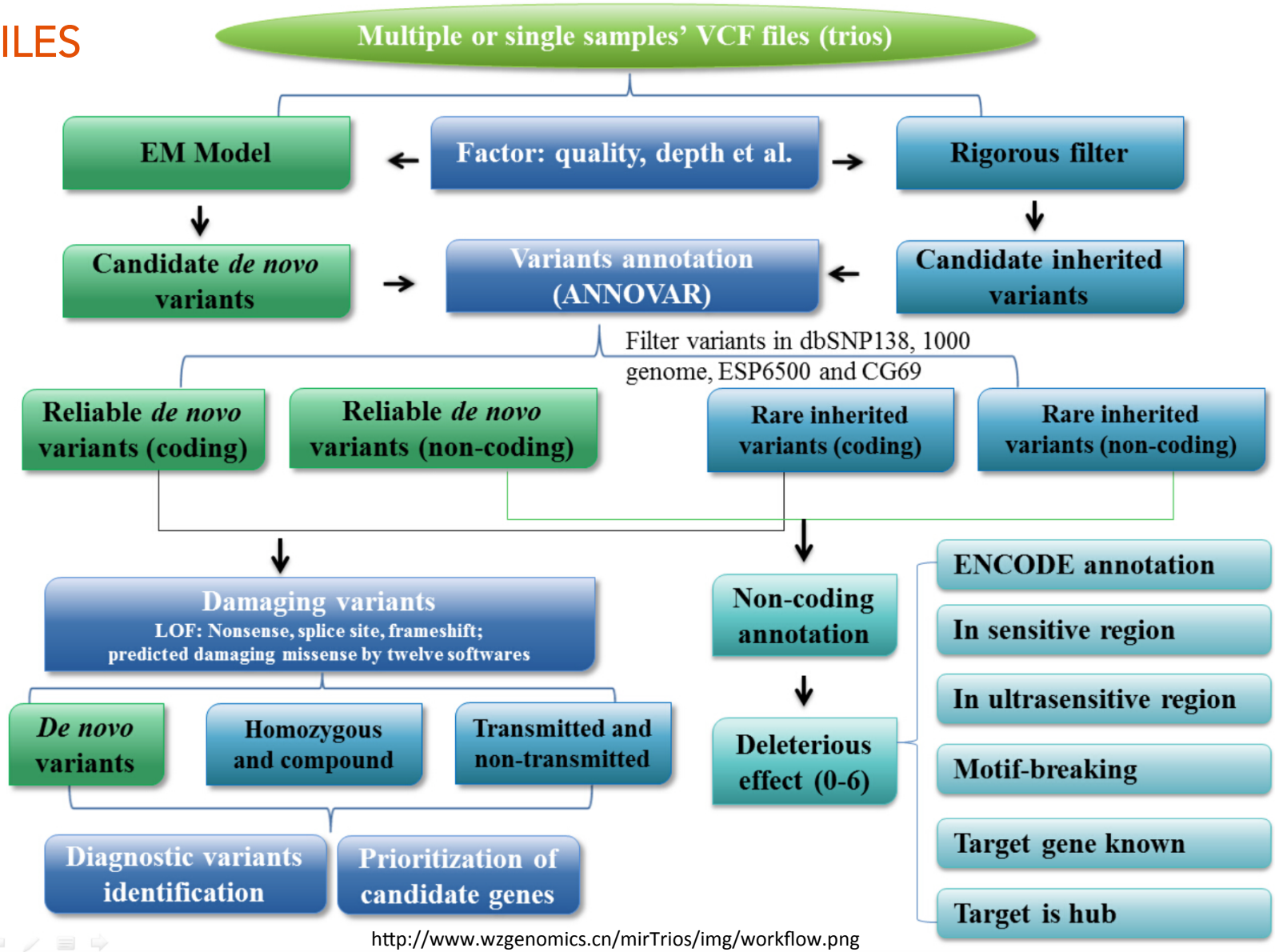
The distribution of 140 spontaneous mutations in *lacI*.
Boxes = position of point mutations
Circles = larger InDels mutants.
red = fast-reverting mutants
Gold = deletions
Allele numbers = sequenced mutants (nb).



NGS PROTOCOLS

FILE FORMATS

VCF FILES

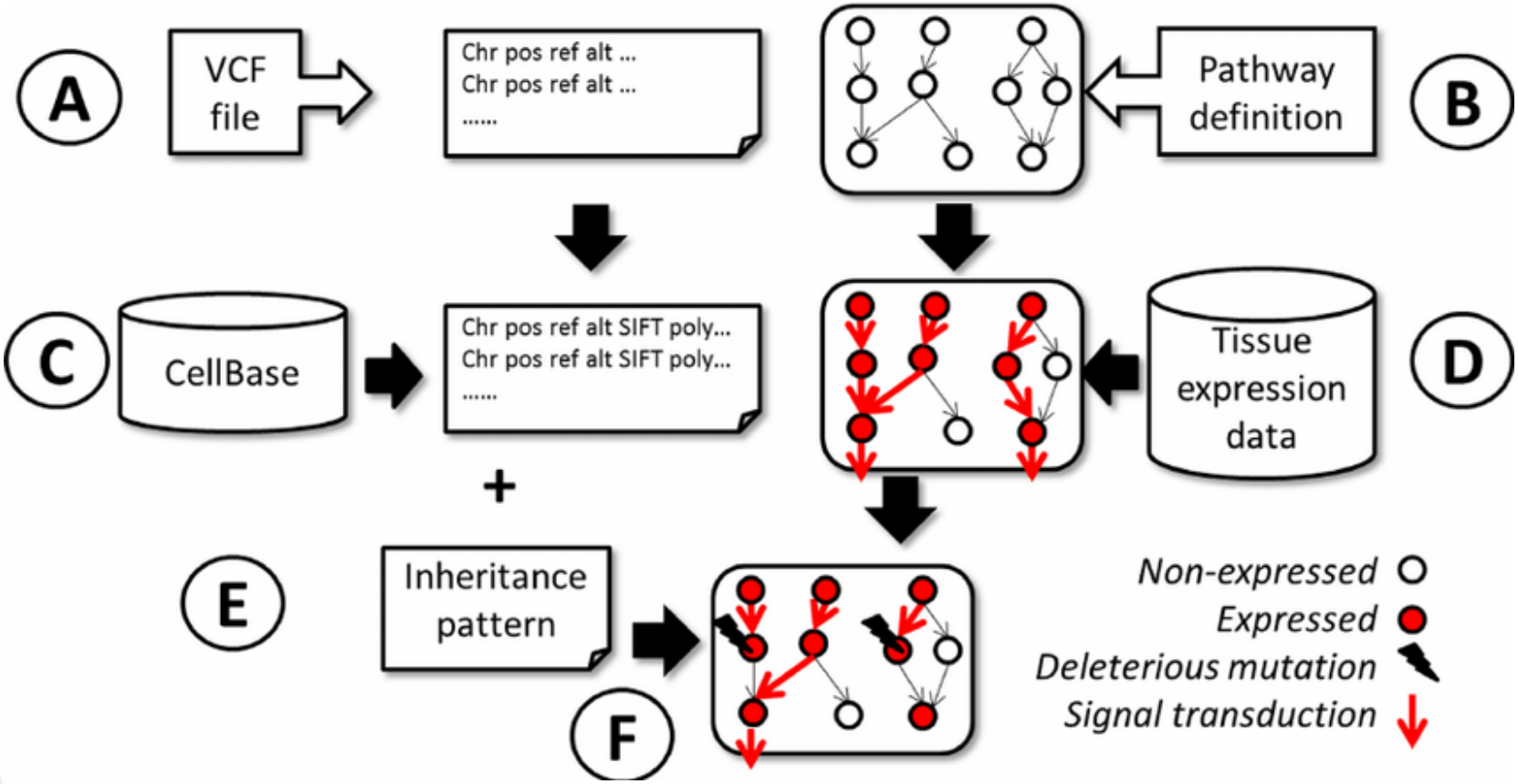


http://www.wzgenomics.cn/mirTrios/img/workflow.png

PART 3 C

VCF FILES

Output example: impact of mutations found in next generation sequencing data over human signaling pathways



(Hernansaiz-Ballesteros et al., 2015)

NGS PROTOCOLS

FILE FORMATS

VCF FILES

Problem:
How can you make sure what you observe are real variants?
When considered alone, an error is indistinguishable from a sequence variant !!

- Errors?
- **Sequencing** errors (Polymerase error (except SOLID), Ligase error (SOLID)…)
- **Mapping** errors
- **Variant calling** errors
- **Human** error (sample preparation bias: e.g. extraction day effect…)
- Reliability of reference genome

- or real heterozygous variant?

VARIANT DETECTION / CALLING

The power to detect variants is reduced by:

- low base quality
- by non-uniformity of coverage.

The power to detect true variants is increased by:

→ Ensuring uniformity of coverage
  - affected by sample preparation (GC-bias during PCR amplification → Limitate PCR amplification if GC rich)
  - influenced by repetitive or low-complexity sequences

→ increasing the sequencing depth/reads = Increase the depth of coverage !
(but cannot solve gaps from repetitive regions of sizes exceeding reads length…)

→ Use PE reads (could solve repetitive regions if smaller than the inner distance

→ DNA resequencing: targeted sequencing if relevant

DNA resequencing

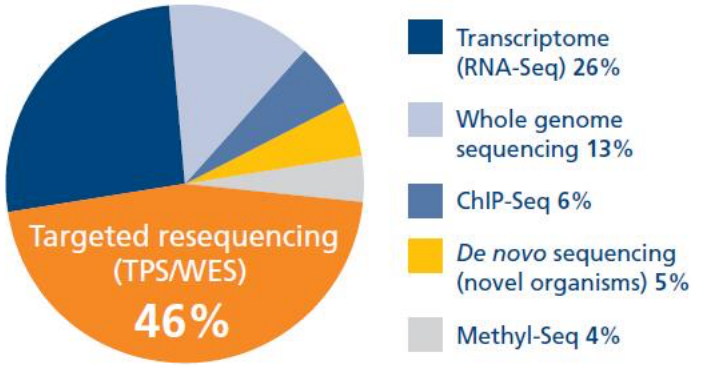
- DNA resequencing explores genetic variation in individuals, families and populations, particularly with respect to human genetic disease.

- Resequencing can better reveal:
- single-nucleotide variants (SNVs)
- small insertions and deletions (indels)
- larger structural variants (such as inversions and translocations)
- copy number variants (CNVs)

NGS PROTOCOLS

FILE FORMATS

DNA resequencing



- Targeted resequencing:

WGS
HIGH VOLUME OF DATA GENERATED
+ GENOMIC REGION OF INTEREST KNOWN FOR MANY DISEASES

→ Targeted resequencing:
- selects chosen regions of the genome for analysis
- offers a highly flexible and economical alternative to WGS.

→ for the same cost, more samples can be sequenced to the same depth but over a smaller genomic region.

(Sims et al., 2014)

https://www.biosciencetechnology.com/white-papers/2013/

NGS PROTOCOLS

FILE FORMATS

DNA resequencing

- Targeted resequencing comprises:

- whole exome sequencing (WES) to target just the gene-encoding regions
- custom panel sequencing, which includes smaller panels focussing on specific regions of the genome.

	Targeted resequencing		Targeted panel
	Whole genome	Whole exome	500 kb
Target size (bp)	3×10^9	5×10^7	500,000
Sample/sequencing run*	6	120	1536
Depth of coverage*	x30	x100	≥500
Data analysis time/sample	>48 hr	4 hr	<1 hr

(Sims et al., 2014)

https://www.biosciencetechnology.com/white-papers/2013/

NGS PROTOCOLS

FILE FORMATS

DNA resequencing

- Targeted resequencing:

WES

- exome accounts for only 1.5% of the human genome
- yet includes 85% of all disease-causing mutations

→ WES methods are highly popular for detecting causal variants without compromising the chance of discovering de novo mutations.

→ NB: Prone to reference bias due to designed capture probes (preferential enrichment of ref allele at heterozygous site : affect SNV calling)

(Sims et al., 2014)
https://www.biosciencetechnology.com/white-papers/2013/

NATURE REVIEWS | GENETICS

NGS PROTOCOLS

FILE FORMATS

DNA resequencing

- Targeted resequencing:

Custom panel sequencing

- more suitable option if the biological question is more focused.
- can provide information on both intronic and exonic regions (association with a particular disease, linked to therapeutic intervention…)
- Can enhance the detection of rare variants or variants present in highly heterogeneous samples : depth of coverage can be further increased when focussing on a smaller target.

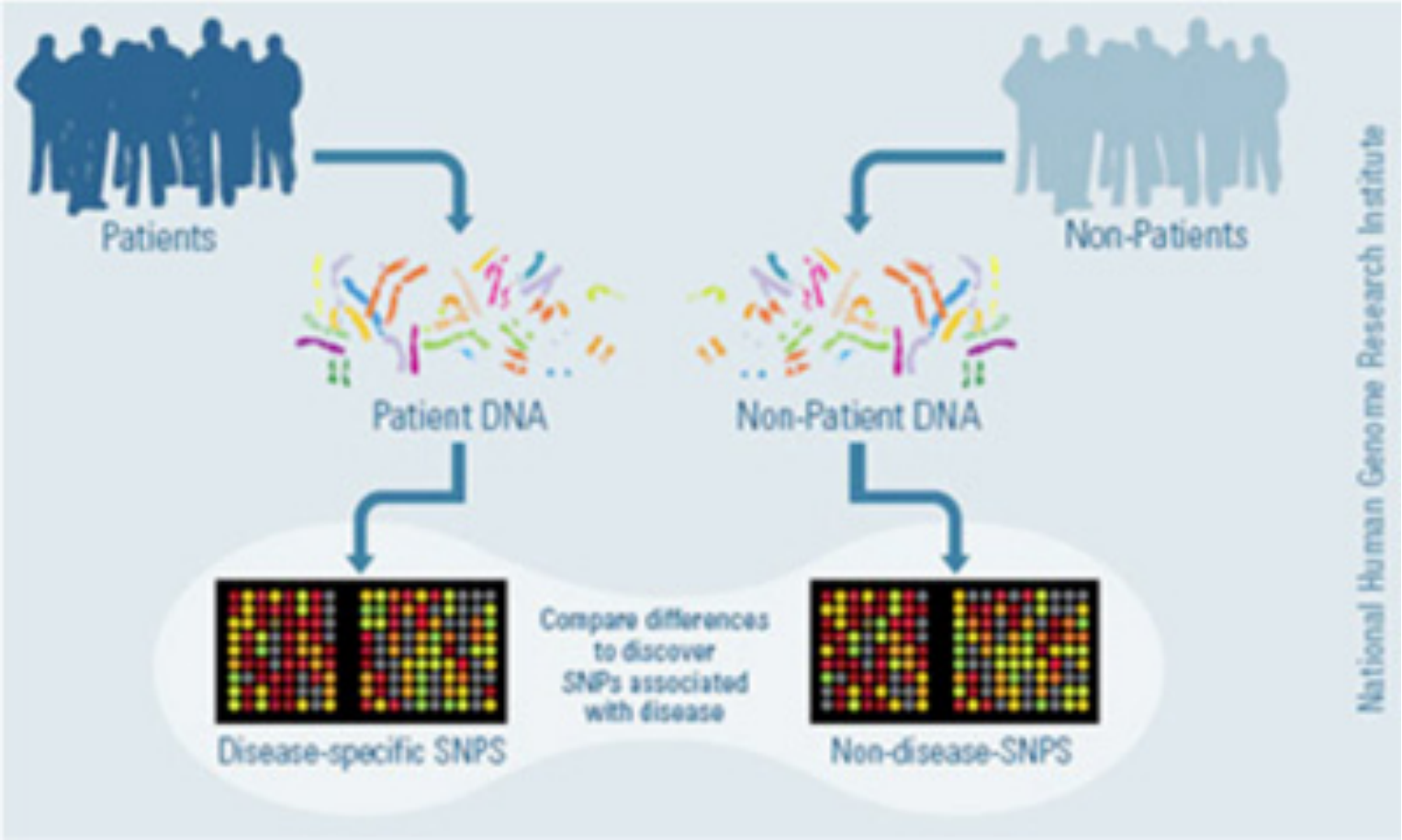
→ This approach is ideal for cancer research, as such variants may be missed or under-represented in WES.

(Sims et al., 2014)
https://www.biosciencetechnology.com/white-papers/2013/

NGS PROTOCOLS

FILE FORMATS

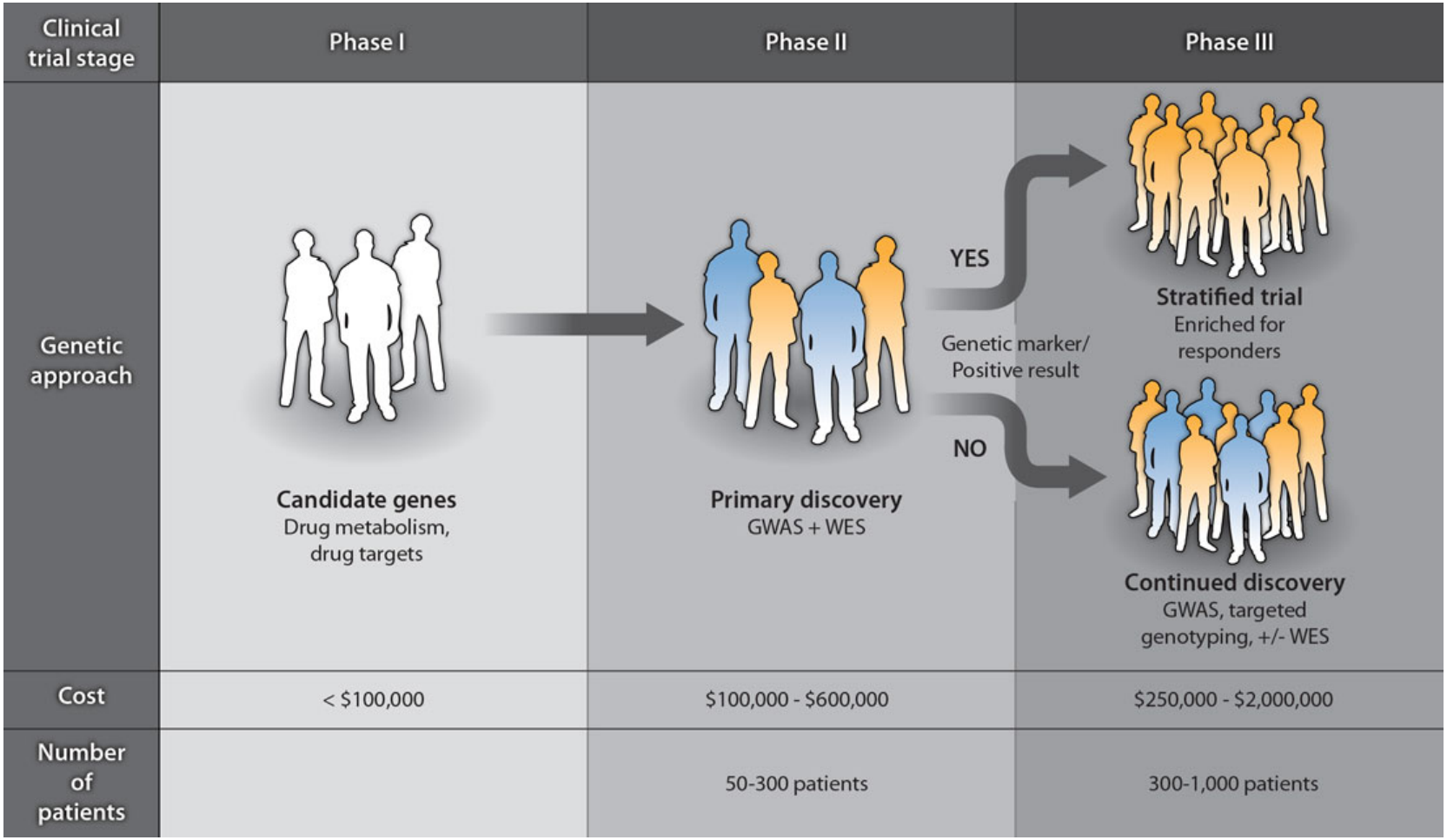
GWAS



NGS PROTOCOLS

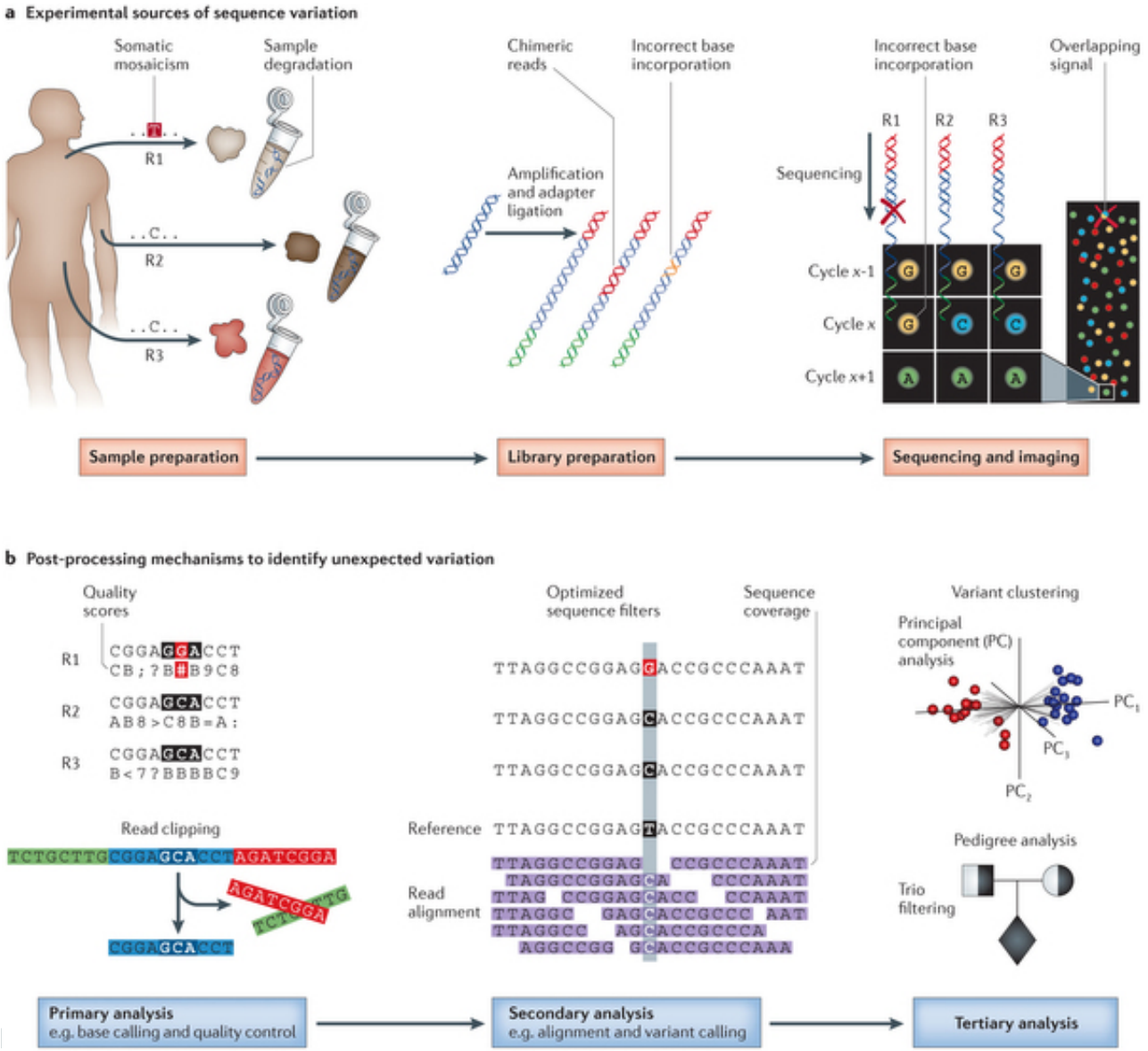
FILE FORMATS

GWAS +/- Targeted +/- WES



NGS PROTOCOLS

SUMMARY



Nature Reviews | Genetics

(Robasky, Lewis & Church, 2014)

NGS PROTOCOLS

SUMMARY

- Many file formats are generated at each stage of the analysis process (.fastq, .sam, .bam, .vcf…)

- Many other formats are available

- *BED (Browser Extensible Data) : flexible way to define the data lines displayed in an annotation track.*
- *bedGraph displays continuous-valued data in track format (useful for probability scores, transcriptome data).*
- *MAF (Multiple Alignment Format) stores multiple alignments at the DNA level between entire genomes.*
- *HAL : graph-based structure, efficiently stores and indexes multiple genome alignments and ancestral reconstructions. HAL files are represented in HDF5 format, an open standard for storing and indexing large, compressed scientific data sets. Genomes within HAL are organized according to the phylogenetic tree that relate them.*
- *….*

- Others are platform specific

The ChIP-Seq programs use SGA (Simple Genome Annotation) files as INPUT and OUTPUT. SGA files are used to represent ChIP-Seq data as well as other genome annotations such as the location of TSSs or matches to consensus sequences.

NGS PROTOCOLS

SUMMARY

- By increasing or decreasing the number of sequencing reads, researchers can tune the sensitivity of an experiment to accommodate various study objectives.

- Sequencing runs can be tailored to zoom in with high resolution on particular regions of the genome, or provide a more expansive view with lower resolution. This offers several experimental design advantages. Examples:

- Detection of low frequency mutations within a mixed cell population: somatic mutations may only exist within a small proportion of cells in a given tissue sample

→ region of DNA having the mutation must be sequenced at extremely high coverage, often >1000×

- Genome-wide variant discovery: study design involves sequencing many samples (hundreds to thousands) at lower coverage → allows to achieve greater statistical power within a given population.

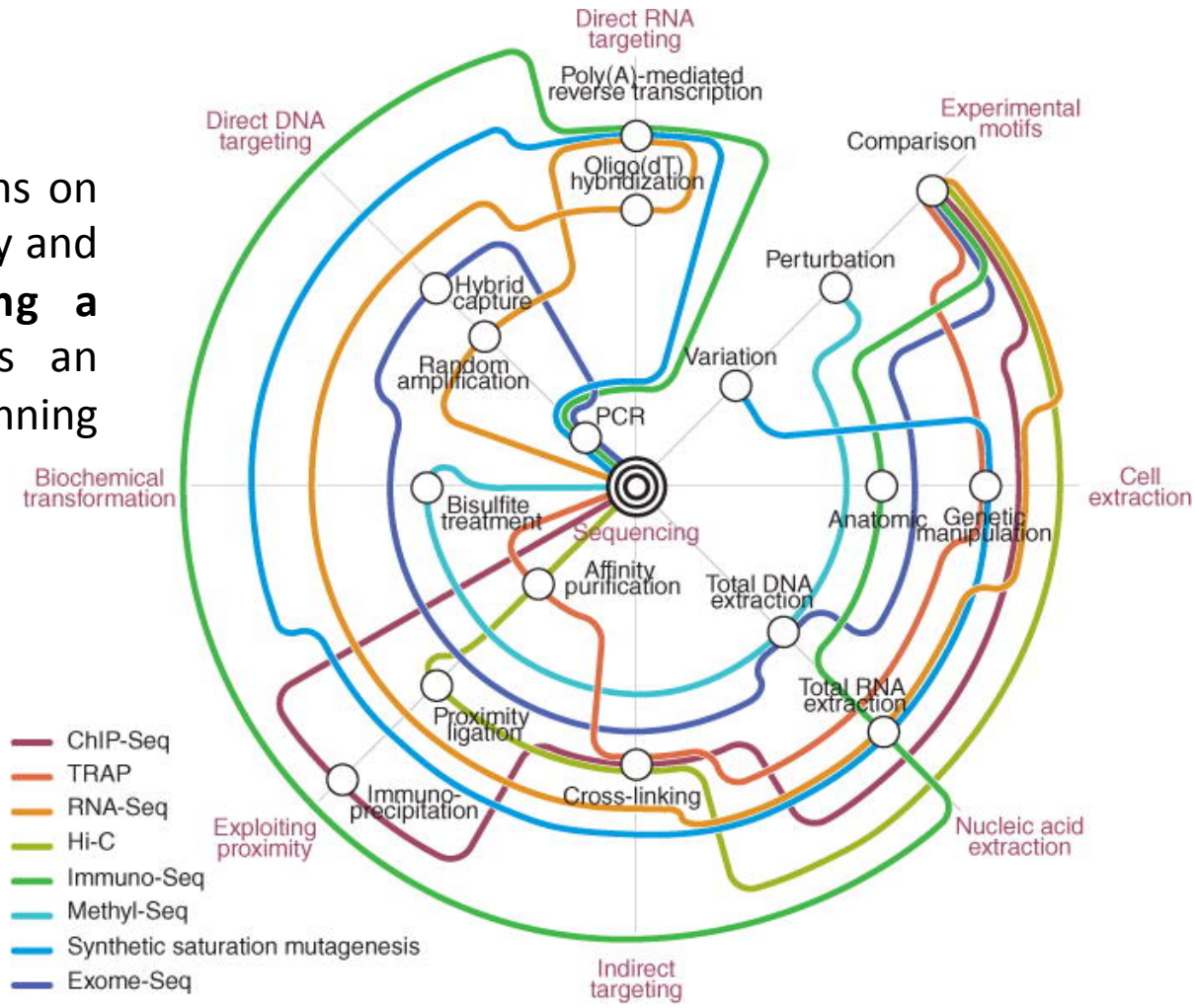
NGS PROTOCOLS

TAKE-HOME MESSAGES

- **Choose your protocol**

Given that sequencing runs on NGS instruments are costly and time-consuming, **defining a data generation goal** is an important step of the planning process.

"How we are getting there: a subway map of sequencing technology. *Despite the disparate goals of different sequencing experiments, the great variety of sequencing experiments is a result of distinct combinations of a relatively small set of core techniques, which are represented as open circles or 'stations'. "*



(Shendure and Aiden, 2012)