« High-throughput sequencing data analysis» cycle

Unit 5/5: RNA-seq analysis

Oct. 3rd, 4rth, 5th, & 6th, 2023

Pierre Pericard - Sarah Guinchard
Bilille platform - PLBS
pierre.pericard@univ-lille.fr

Unit 5/5: RNA-seq analyses

- **Day 1: Prerequisites for NGS** (Pierre Pericard, Sarah Guinchard)
  - Introduction to Galaxy
  - Next Generation Sequencing (NGS)
  - Quality control, cleaning and preprocessing
  - Mapping on a reference genome
- **Day 2 & 3: RNA-seq Bioinfo** (Camille Marchet, P. Pericard, S. Guinchard)
  - With reference RNA-seq
  - De-novo RNA-seq
  - Introduction to 3rd generation sequencing (long reads)
- **Day 4: RNA-seq Biostats** (Pierre Pericard, Estelle Chatelain)
  - Data exploration (descriptive stats, PCA, clustering, ...)
  - Differentially expressed genes analysis (DEG)
  - Enrichment analyses: over-representation (ORA), ranking tests (GSEA)

Day 1: Bases of NGS analyses for RNA-seq

Morning

- Lectures
  - Introduction to NGS
  - Reads Quality Control + Cleaning
- FastQC + MultiQC + cleaning practical

  https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html

Lunch break

Afternoon

- Lectures
  - Reads mapping on reference
  - Reads duplicates
- Mapping practical

  https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/mapping/tutorial.html

Day 1: Bases of NGS analyses for RNA-seq

- NGS Introduction

➔ What is NGS?

➔ Sequencers

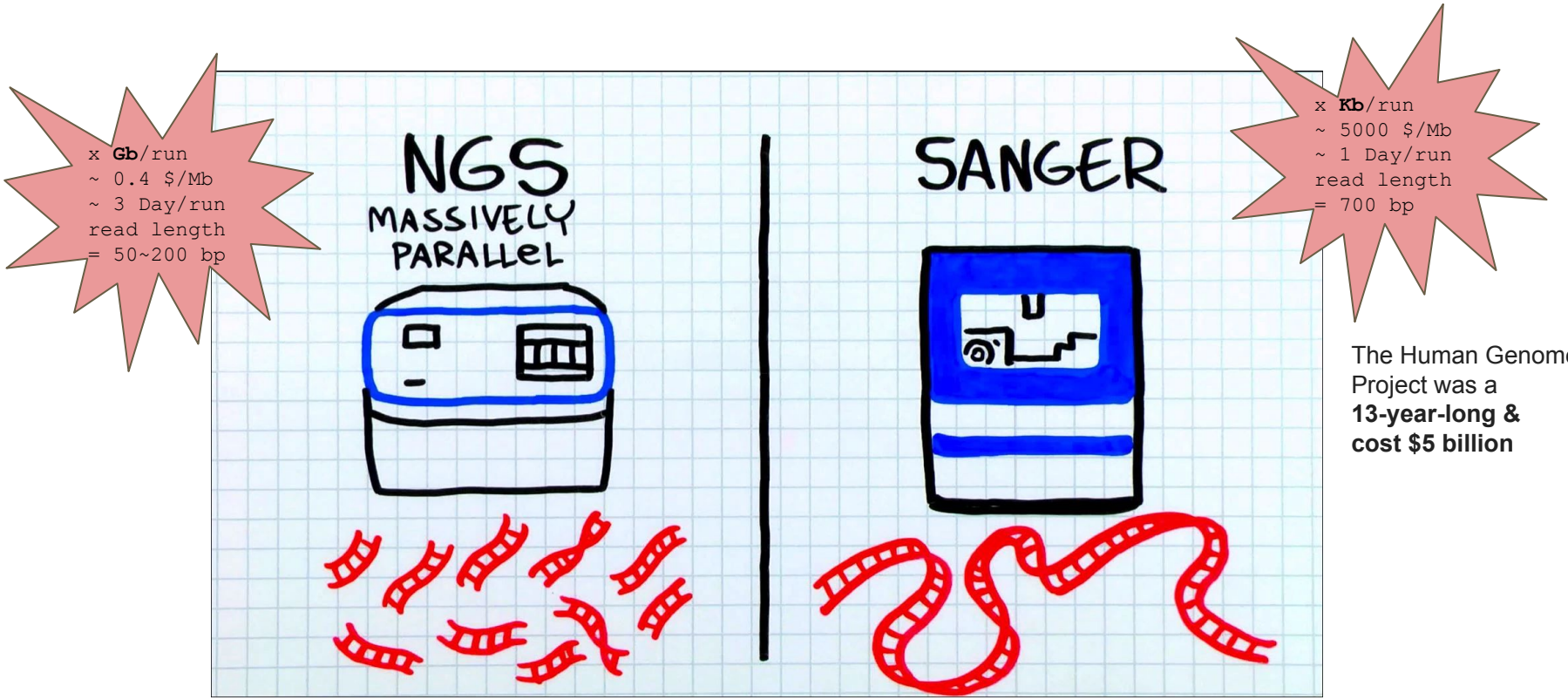
➔ Applications

➔ NGS workflow

➔ Output data

What is Next-Generation Sequencing (NGS)?

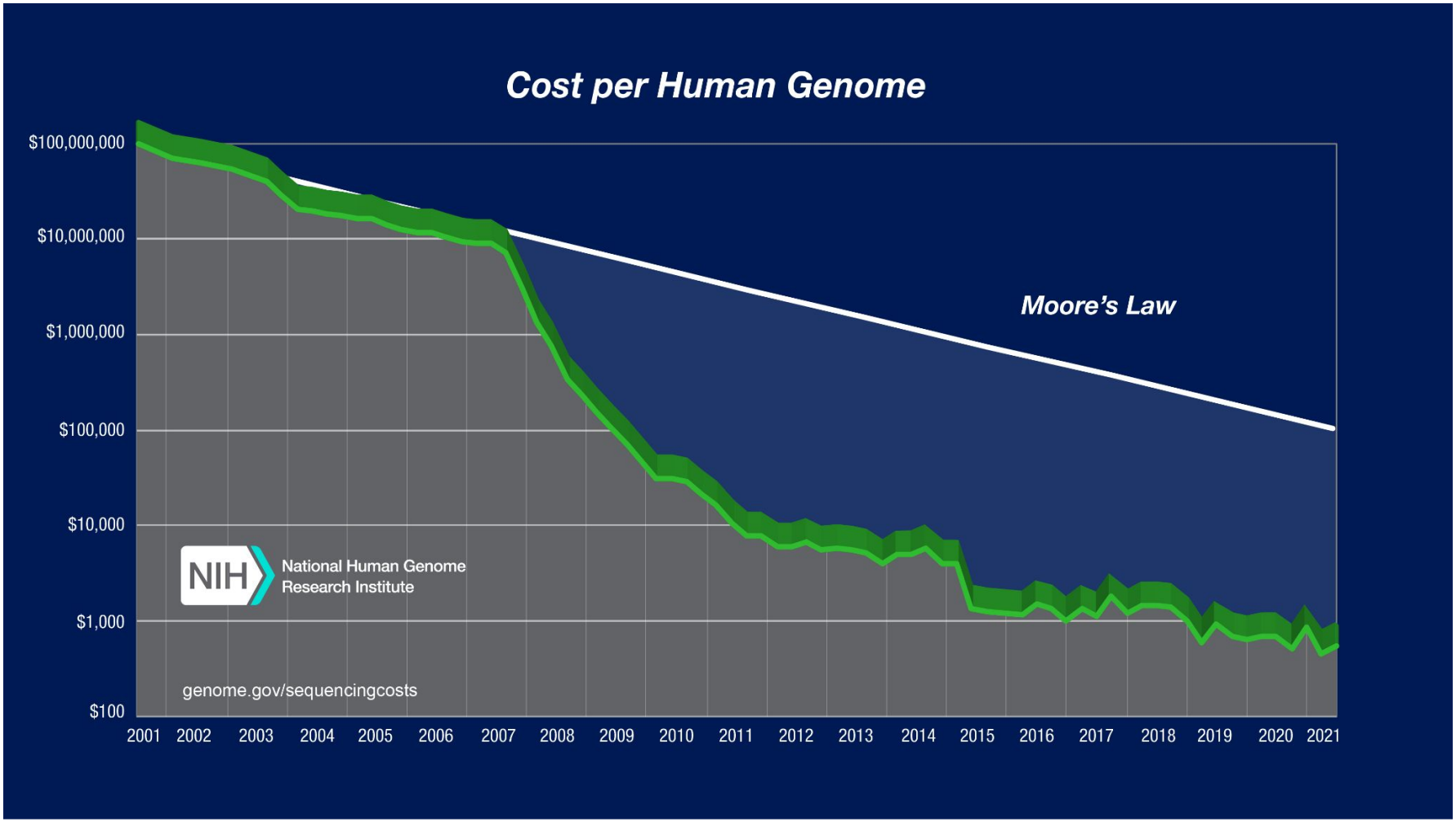
"Next-generation sequencing (NGS), also known as high-throughput sequencing, is the catch-all term used to describe a number of different modern sequencing technologies. These technologies allow for sequencing of DNA and RNA much more quickly and cheaply than the previously used Sanger sequencing, and as such revolutionised the study of genomics and molecular biology"



The Human Genome Project was a **13-year-long & cost $5 billion**

https://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/what-you-will-learn/what-next-generation-dna-
https://www.thermofisher.com/blog/behindthebench/when-do-i-use-sanger-sequencing-vs-ngs-seq-it-out-7/

What is Next-Generation Sequencing (NGS)?



What is Next-Generation Sequencing (NGS)?

Illumina sequencing

Illumina sequencing works by simultaneously identifying DNA bases, as each base emits a unique fluorescent signal, and adding them to a nucleic acid chain

Ion Torrent: Proton / PGM sequencing (thermofisher)

Ion Torrent sequencing measures the direct release of H+ (protons) from the incorporation of individual bases by DNA polymerase and therefore differs from the previous two methods as it does not measure light.

https://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/what-you-will-learn/what-next-generation-dna-

What is Next-Generation Sequencing (NGS)?

Illumina sequencing

Illumina sequencing works by simultaneously identifying DNA bases, as each base emits a unique fluorescent signal, and adding them to a nucleic acid chain

Ion Torrent: Proton / PGM sequencing (thermofisher)

Ion Torrent sequencing measures the direct release of H+ (protons) from the incorporation of individual bases by DNA polymerase and therefore differs from the previous two methods as it does not measure light.

MGI sequencing (BGI group)

https://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/what-you-will-learn/what-next-generation-dna-

Sequencers – Illumina

	Benchtop Sequencers			Production-Scale Sequencers	
	 iSeq 100	 MiniSeq	 MiSeq Series	 NextSeq 550 Series	 NextSeq 1000 & 2000
Popular Applications & Methods	Key Application	Key Application	Key Application	Key Application	Key Application
Large Whole-Genome Sequencing (human, plant, animal)					
Small Whole-Genome Sequencing (microbe, virus)	●	●	●	●	●
Exome & Large Panel Sequencing (enrichment-based)				●	●
Targeted Gene Sequencing (amplicon-based, gene panel)	●	●	●	●	●
Single-Cell Profiling (scRNA-Seq, scDNA-Seq, oligo tagging assays)				●	●
Transcriptome Sequencing (total RNA-Seq, mRNA-Seq, gene expression profiling)				●	●
Targeted Gene Expression Profiling	●	●	●	●	●
miRNA & Small RNA Analysis	●	●	●	●	●
DNA-Protein Interaction Analysis (ChIP-Seq)			●	●	●
Methylation Sequencing				●	●
16S Metagenomic Sequencing		●	●	●	●
Metagenomic Profiling (shotgun metagenomics, metatranscriptomics)				●	●
Cell-Free Sequencing & Liquid Biopsy Analysis				●	●
Run Time	9.5–19 hrs	4–24 hours	4–55 hours	12–30 hours	11–48 hours
Maximum Output	1.2 Gb	7.5 Gb	15 Gb	120 Gb	330 Gb*
Maximum Reads Per Run	4 million	25 million	25 million †	400 million	1.1 billion*
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp

https://www.illumina.com/systems/sequencing-platforms.html

Sequencers – Illumina

	Benchtop Sequencers	Production-Scale Sequencers	
	 NextSeq 1000 & 2000	 NovaSeq 6000	 NovaSeq X Series
Popular Applications & Methods	Key Application	Key Application	Key Application
Large Whole-Genome Sequencing (human, plant, animal)		●	●
Small Whole-Genome Sequencing (microbe, virus)	●	●	●
Exome & Large Panel Sequencing (enrichment-based)	●	●	●
Targeted Gene Sequencing (amplicon-based, gene panel)	●	●	●
Single-Cell Profiling (scRNA-Seq, scDNA-Seq, oligo tagging assays)	●	●	●
Transcriptome Sequencing (total RNA-Seq, mRNA-Seq, gene expression profiling)	●	●	●
Chromatin Analysis (ATAC-Seq, ChIP-Seq)	●	●	●
Methylation Sequencing	●	●	●
Metagenomic Profiling (shotgun metagenomics, metatranscriptomics)	●	●	●
Cell-Free Sequencing & Liquid Biopsy Analysis	●	●	●
Run Time	11-48 hours	~13–38 hours (dual SP flow cells) ~13–25 hours (dual S1 flow cells) ~16–36 hours (dual S2 flow cells) ~44 hours (dual S4 flow cells)	~13–21 hours (1.5B flow cells‡) ~18–24 hours (10B flow cells‡) ~48 hours (25B flow cells‡)
Maximum Output	360 Gb *	6000 Gb	16 Tb
Maximum Reads Per Run	1.2 billion *	20 billion	26 billion (single flow cells) 52 billion (dual flow cells)
Maximum Read Length	2 × 150 bp	2 × 250 bp**	2 × 150 bp

https://www.illumina.com/systems/sequencing-platforms.html

Sequencers – Illumina (pre-2020)

illumina®

NextSeq Series

HiSeq Series

HiSeq X Series‡

NovaSeq 6000 System

Popular Applications & Methods	Key Application	Key Application	Key Application	Key Application
Large Whole-Genome Sequencing (human, plant, animal)	●	●	●	●
Small Whole-Genome Sequencing (microbe, virus)	●	●		●
Exome Sequencing	●	●		●
Targeted Gene Sequencing (amplicon, gene panel)	●	●		●
Whole-Transcriptome Sequencing	●	●		●
Gene Expression Profiling with mRNA-Seq	●	●		●
miRNA & Small RNA Analysis	●	●		●
DNA-Protein Interaction Analysis	●	●		●
Methylation Sequencing	●	●		●
Shotgun Metagenomics	●	●		●

Optimized NGS Sample Tracking and Workflows

See how BaseSpace Clarity LIMS (Laboratory Information Management System) enabled this large genomics lab to standardize lab procedures and cope with increasing sample volumes from diverse clients.

Read Case Study ›

Run Time	12–30 hours	< 1–3.5 days (HiSeq 3000/HiSeq 4000) 7 hours–6 days (HiSeq 2500)	< 3 days	16–36 hours (Dual S2 flow cells) 44 hours (Dual S2 flow cells)
Maximum Output	120 Gb	1500 Gb	1800 Gb	6000 Gb§
Maximum Reads Per Run	400 million	5 billion	6 billion	20 billion**
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp

https://www.illumina.com/systems/sequencing-platforms.html

Sequencers – Thermo Fisher Scientific

Plateformes de séquençage	 Système Ion PGM™ pour le séquençage de nouvelle génération	 Système Ion S5™ pour le séquençage de nouvelle génération	 Système Ion S5™ XL pour le séquençage de nouvelle génération
Avantages	**Évolutivité** : de 30 Mo à 2 Go **Rapidité** : séquençage exécuté en 2 à 7 heures, selon la longueur de lecture et la sortie par la puce	**Simplicité** : solutions de flux de travaux automatisé, de la préparation des échantillons à l'analyse **Évolutivité** : de 600 Mo à 15 Go **Rapidité** : séquençage effectué en 2,5 à 4 heures (quelle que soit la sortie par la puce)	**Simplicité** : solutions de flux de travaux automatisé, de la préparation des échantillons à l'analyse **Évolutivité** : de 600 Mo à 15 Go **Rapidité** : de l'ADN aux données en 24 heures
Applications de séquençage	ARN ciblé ADN ciblé Microbien	ARN ciblé ADN ciblé Microbien Transcriptome Exome Séquençage de l'ARN	ARN ciblé ADN ciblé Microbien Transcriptome Exome Séquençage de l'ARN

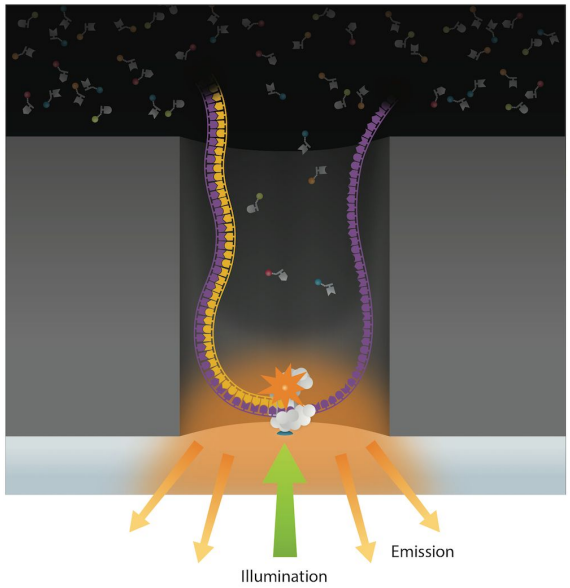
https://www.thermofisher.com/fr/fr/home/life-science/sequencing/sequencing-technology-solutions.html

Sequencers - MGI (BGI group)

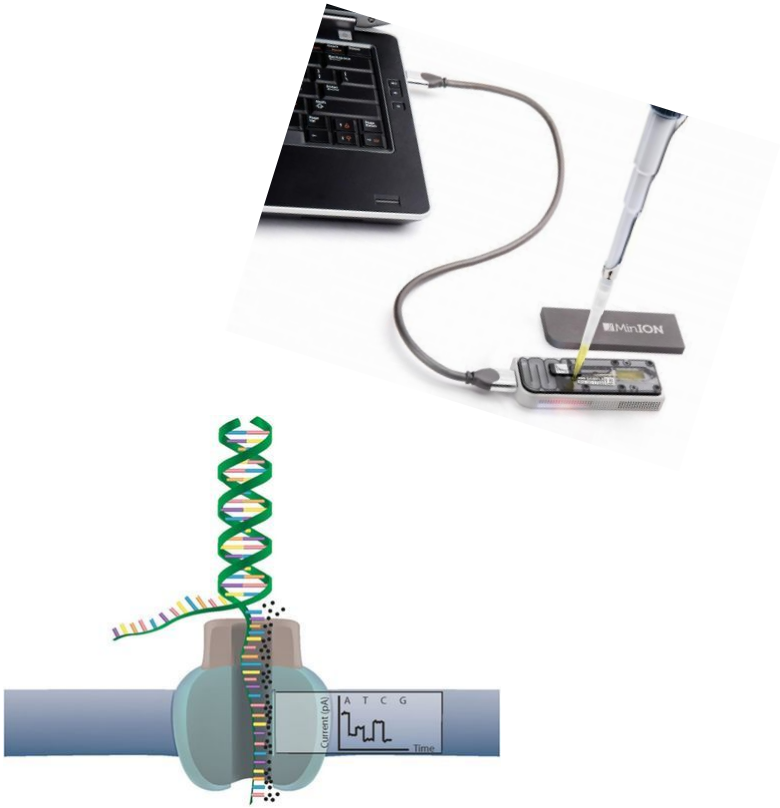
							
	Sequencers	Sequencers	Sequencers	Sequencers	Sequencers	Sequencers	Sequencers
Product Model	DNBSEQ-T7	DNBSEQ-T7* For HotMPS Only	DNBSEQ-G400	DNBSEQ-G400* For HotMPS Only	DNBSEQ-G400C*	DNBSEQ-G99	DNBSEQ-G50
Features	Ultra-high Throughput	Ultra-high Throughput	Adaptive	Adaptive	Adaptive	Fast	Effective
Applications	Whole Genome Sequencing, Deep Exome Sequencing, Transcriptome Sequencing, and Targeted Panel Projects.	Whole Genome Sequencing, Deep Exome Sequencing, Transcriptome Sequencing, and Targeted Panel Projects.	WGS, WES, Transcriptome sequencing, etc.	WGS, WES, Transcriptome sequencing, etc.	Small RNA,Pathogen Fast Identification etc.	Targeted oncology panel sequencing, infectious disease sequencing, oncology methylation sequencing, small whole-genome sequencing	Small whole genome sequencing, targeted DNA/RNA panels, low-pass whole genome sequencing
Flow Cell Type	FC	FC	FCL & FCS	FCL	FCL	FC	FCL & FCS
Lane/Flow Cell++	1 lane	1 lane	2 or 4 lanes	4 lanes	4 lanes	1 lane	1 lane
Operation Mode	Ultra-high Throughput	Ultra-high Throughput	High Throughput	High Throughput	High Throughput	Small and Medium Throughput	Medium Throughput
Max. Throughput / RUN	6Tb	4Tb	1440Gb	720Gb	360G	48Gb	150Gb
Effective Reads / Flow Cell	5000M	5000M	300M/550M/1500-1800M	1500-1800M	1500-1800M	80M	500M / 100M
Average run time	24~30 hours for PE150 sequencing	20~22 hrs for PE100 sequencing	FCS: 13~37 hours FCL: 14~109 hours	15.5-50.5 hours	17/30 hours	12 hours（PE150 ）	9~40 hours
Min. Read Length	PE100	PE100	SE50	SE50	SE50	SE100	SE50
Max. Read Length	PE150	PE100	PE300	PE100	SE100	PE150	PE150

https://en.mgi-tech.com/products/

Third-generation sequencing



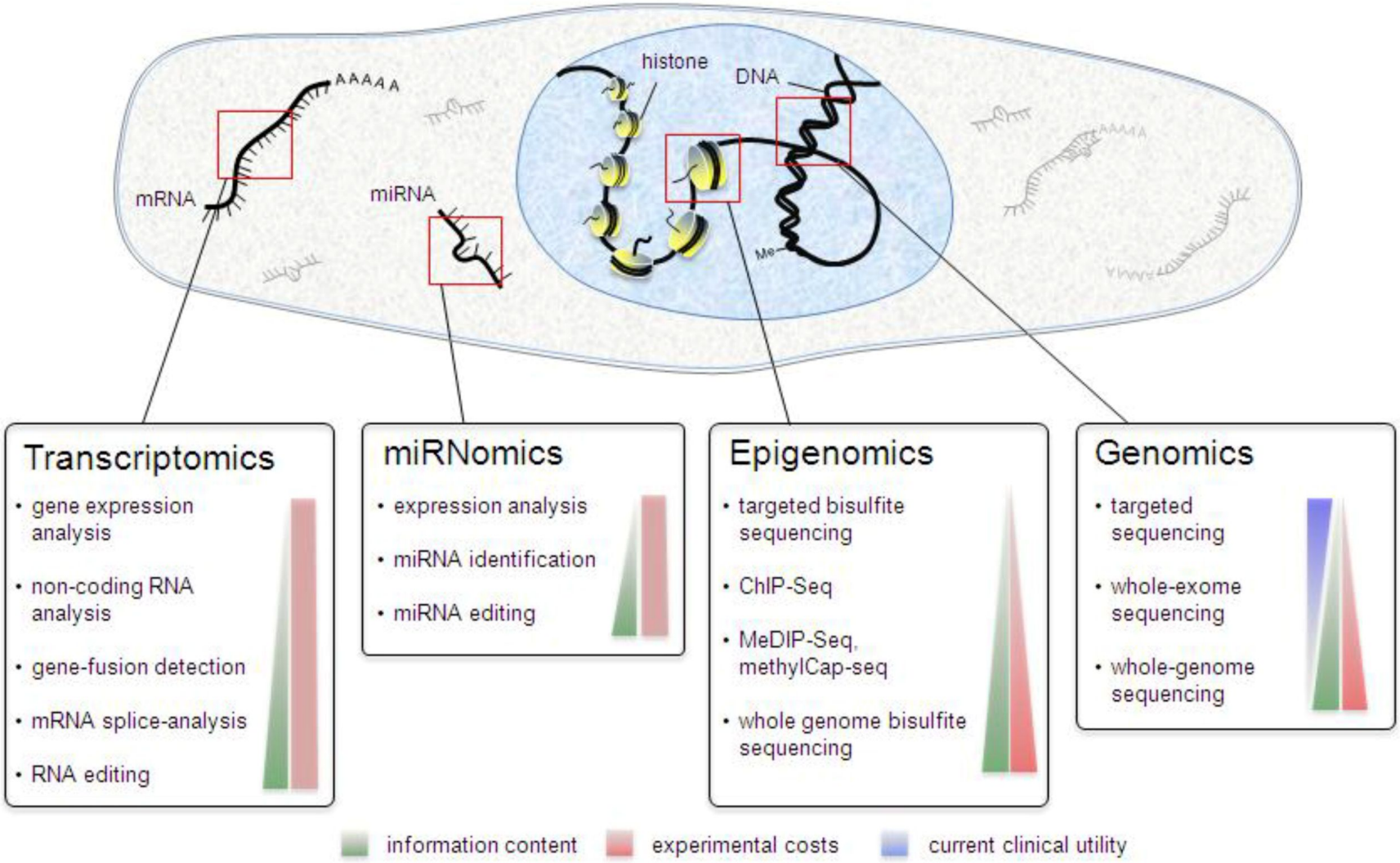
PacBio Sequencing



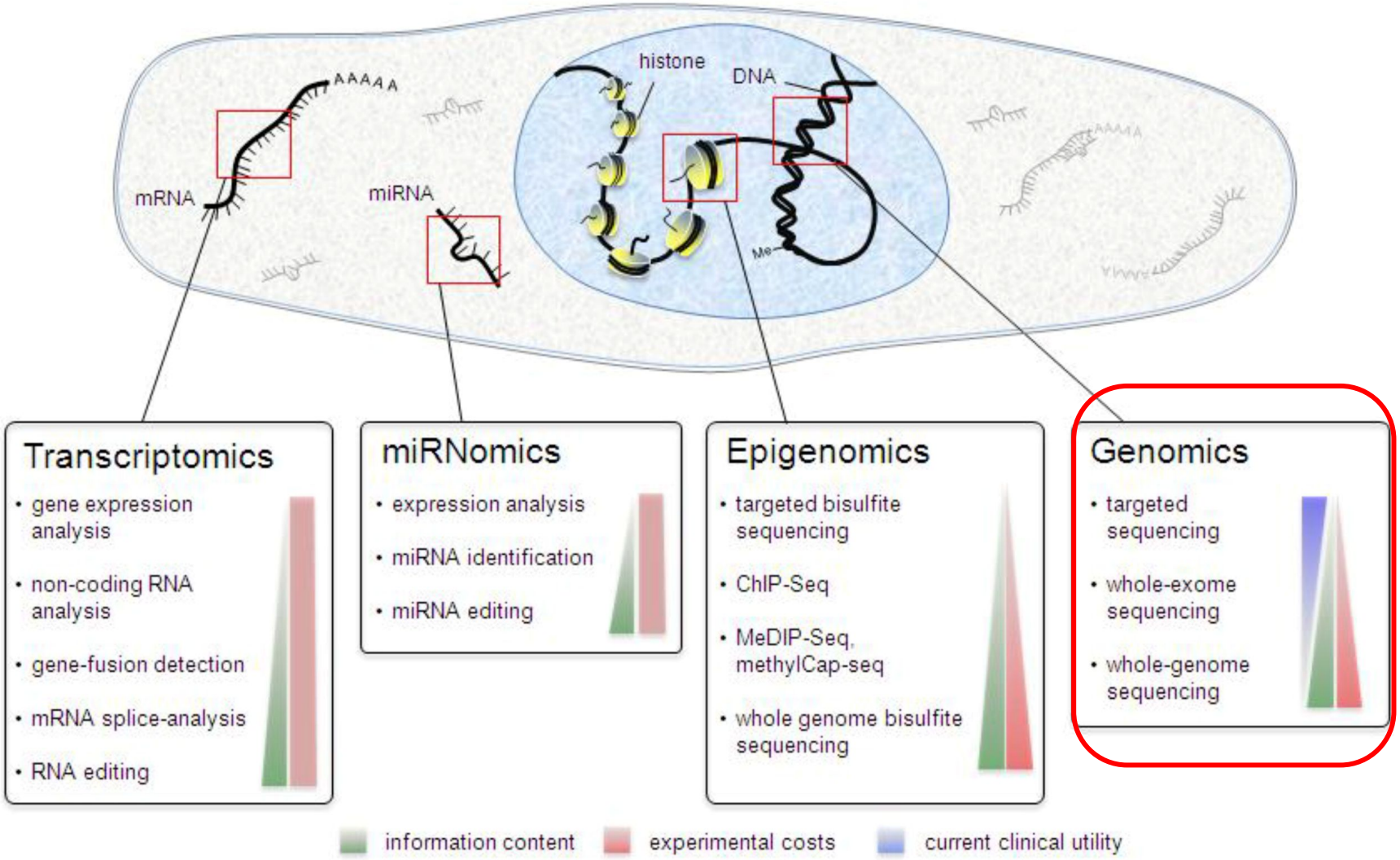
Nanopore technology (ONT)

- much longer reads (> Kb)
- error rate (~ 1-10 %)

Applications



Applications

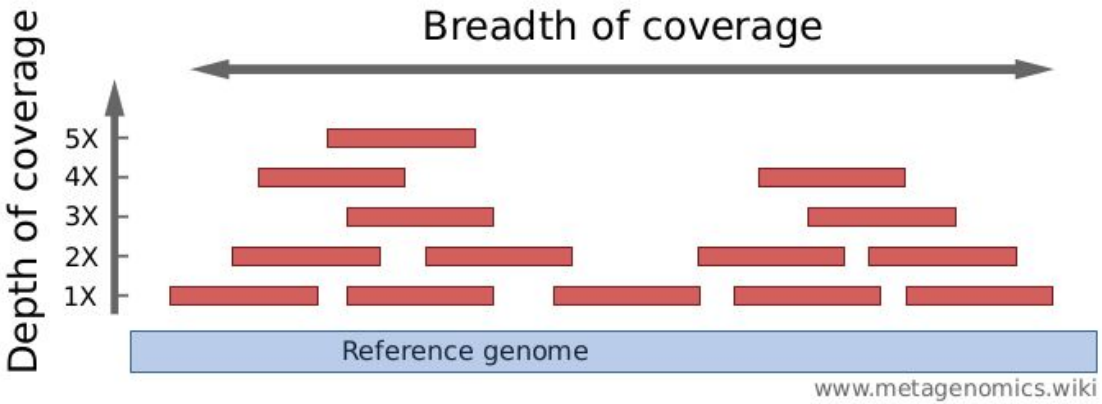


Unit of the sequencing output: the reads

- Result of a sequencing reading
- Short sequences (generally 100-200 bp)
- Includes an identifier that can contain various metadata (fasta, fastq)
- One sequencing run → millions of reads
- In text format, bases written as A, T, C, G or N (unknown)
- Can include quality information for each base (fastq)
- Stay tuned for more details

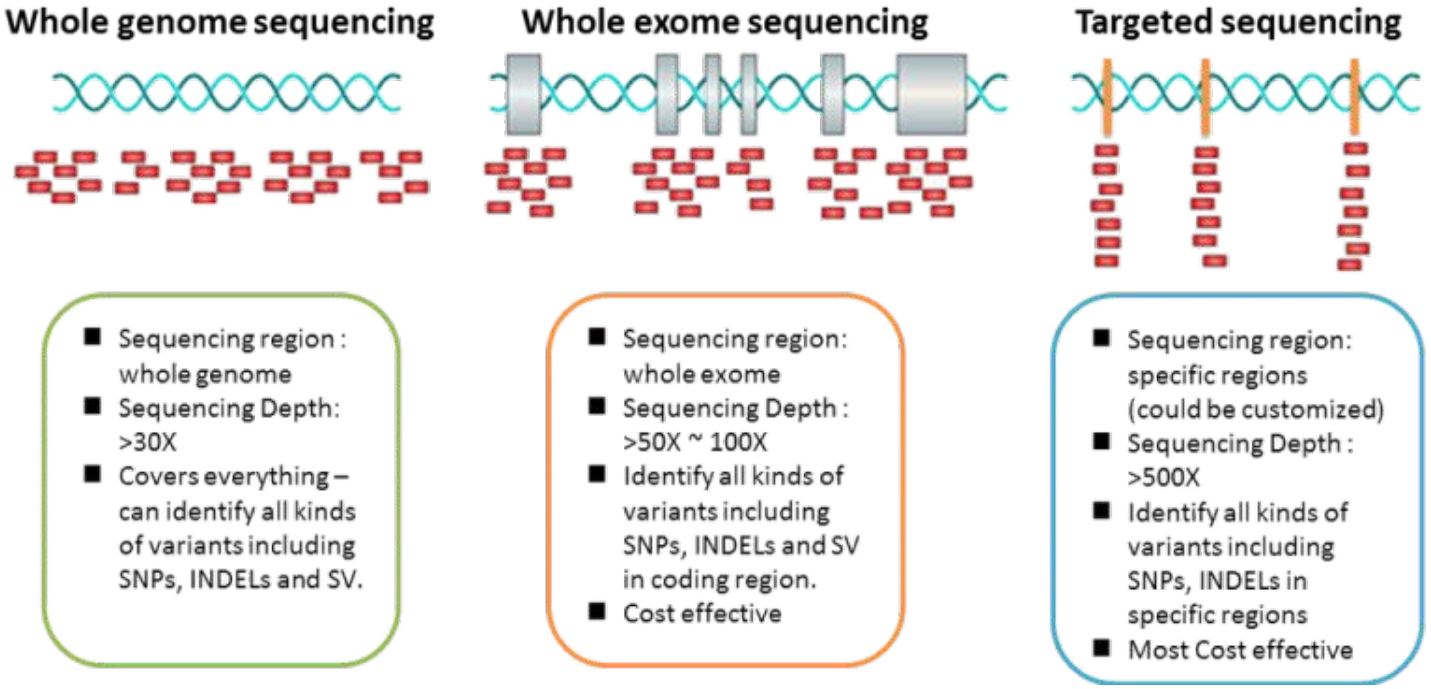
Coverage and depth of coverage

- **Depth of coverage (at one position) =** number of reads covering that position
- **Average depth of coverage =** average number of reads covering a base (X)
  - Example: 30X for normal sample, 100X for tumor sample
- **(Breadth of) Coverage =** percentage of the targeted regions covered by at least X read
  - For example: 90% of a genome is covered at 1X depth; and still 40% is covered at 4X depth.



Source :
- Élodie Girard , 5ème Ecole de bioinformatique AVIESAN-IFB 2016 ,http://www.france-bioinformatique.fr/sites/default/files/V01_ITMO_2016_EG_from_fastq_to_mapping_1.pdf
- http://www.metagenomics.wiki/pdf/definition/coverage-read-depth

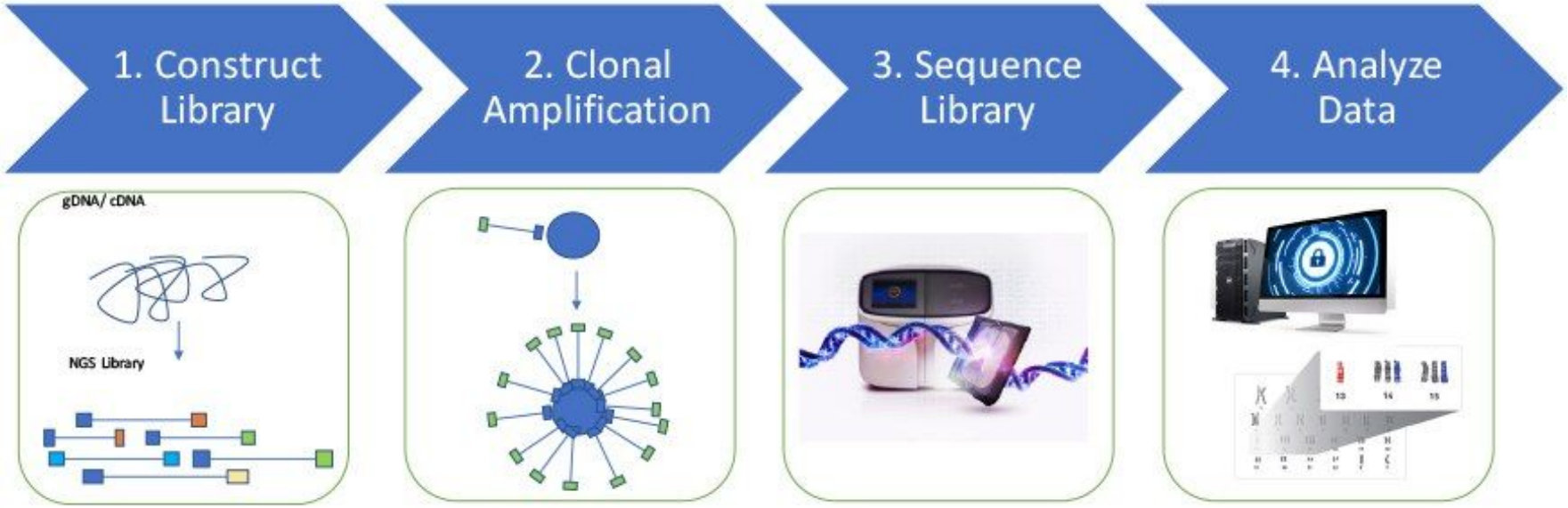
Applications : genomics (DNA-seq)



- Targeted sequencing : rapid and cost-effective way to detect known and novel variants in selected sets of genes or genomic regions
- Whole exome sequencing : sequencing all of the protein-coding regions of genes in a genome (applications : discover rare-variants, adjacent splice-sites,...)
- Whole genome sequencing : alterations in regulatory sequences and non-coding regions, chromosomal rearrangements, ...

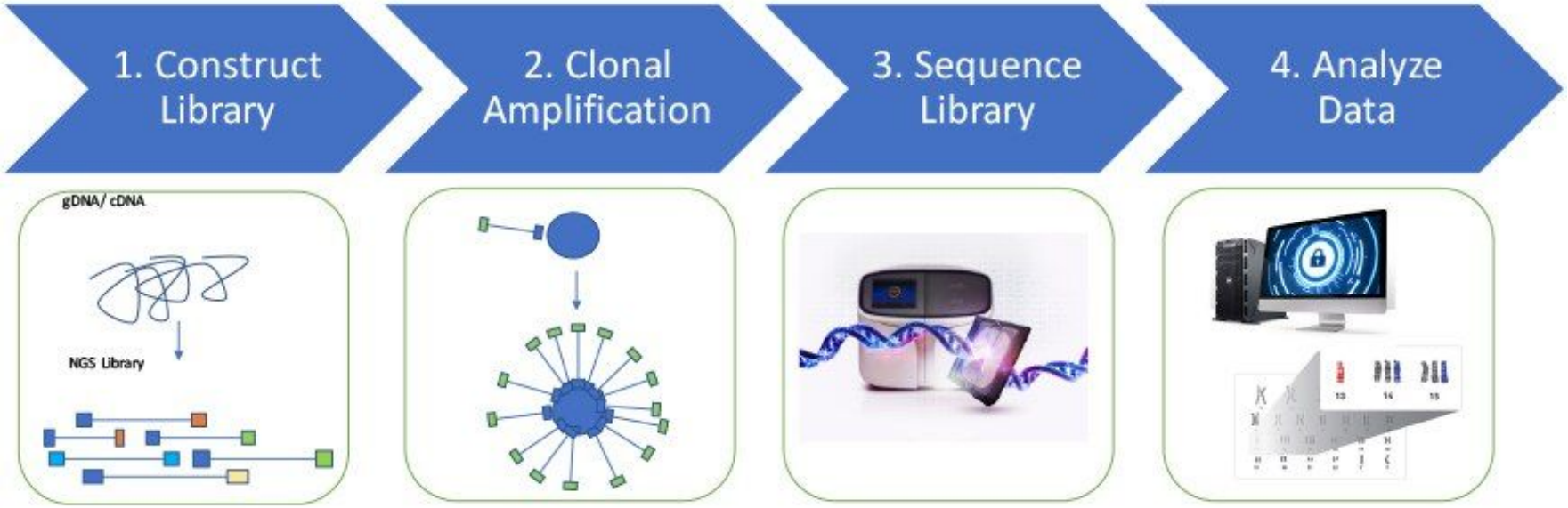
http://www.genomesop.com/somatic-mutations/

NGS workflow



https://www.thermofisher.com/fr/fr/home/life-science/sequencing/sequencing-education/next-generation-sequencing-basics/what-is-next-generation-sequencing.html

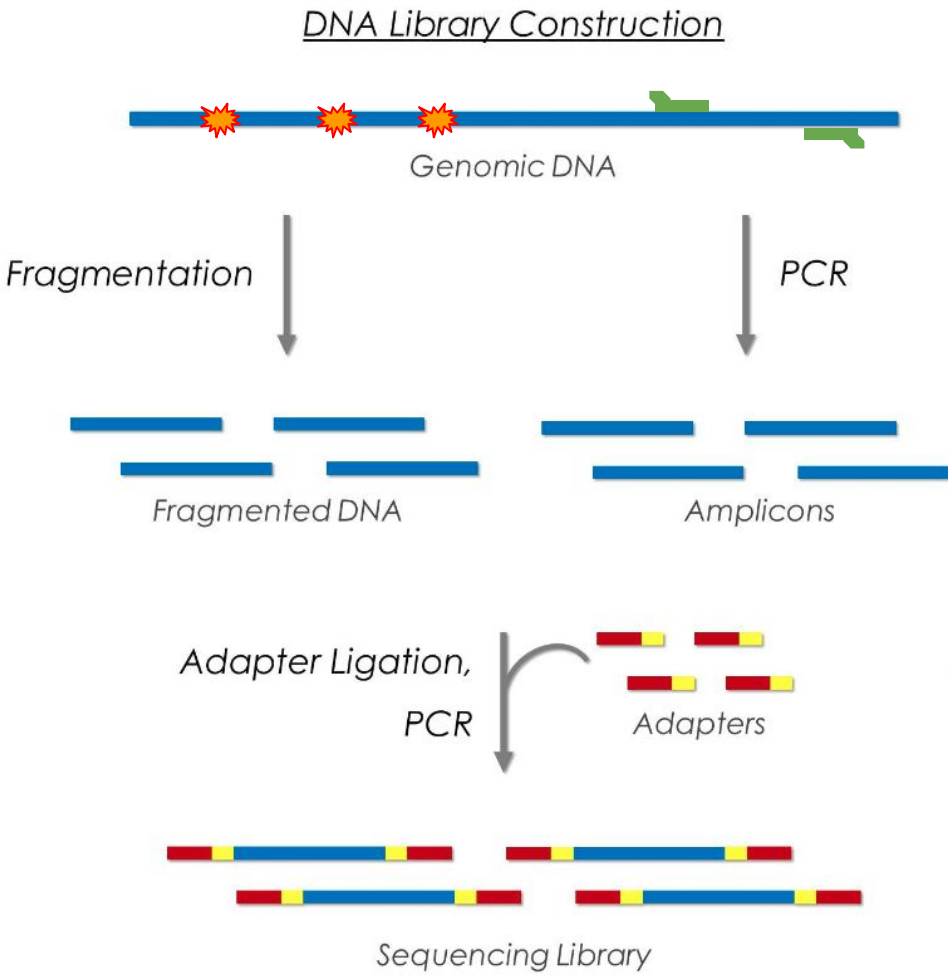
NGS workflow



https://www.thermofisher.com/fr/fr/home/life-science/sequencing/sequencing-education/next-generation-sequencing-basics/what-is-next-generation-sequencing.html

Library construction

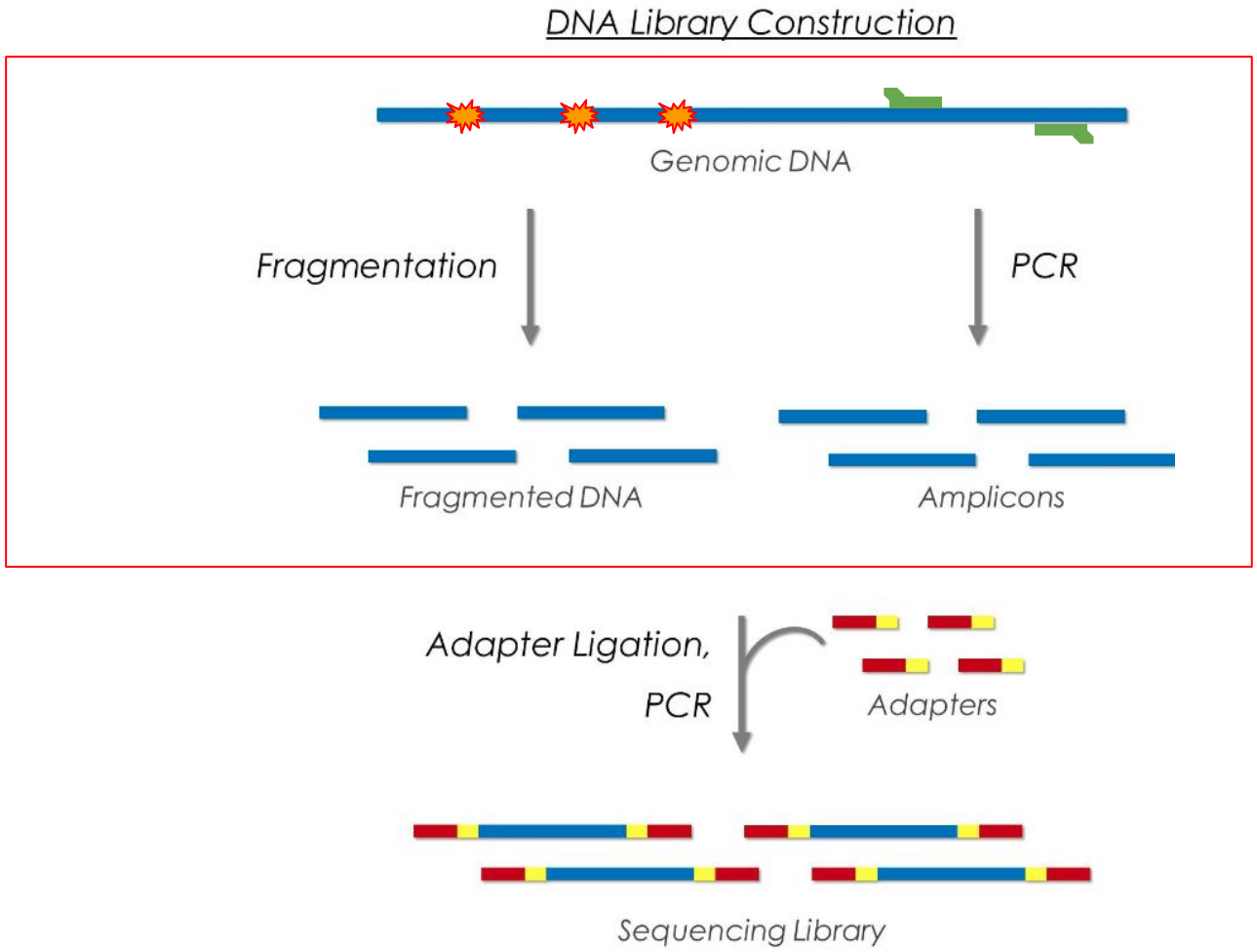
A sequencing “library” must be created from the sample. The DNA (or cDNA) sample is processed into relatively short double-stranded fragments (100–800 bp)



source : https://www.biocompare.com/Molecular-Biology/9187-Next-Generation-Sequencing/

Library construction

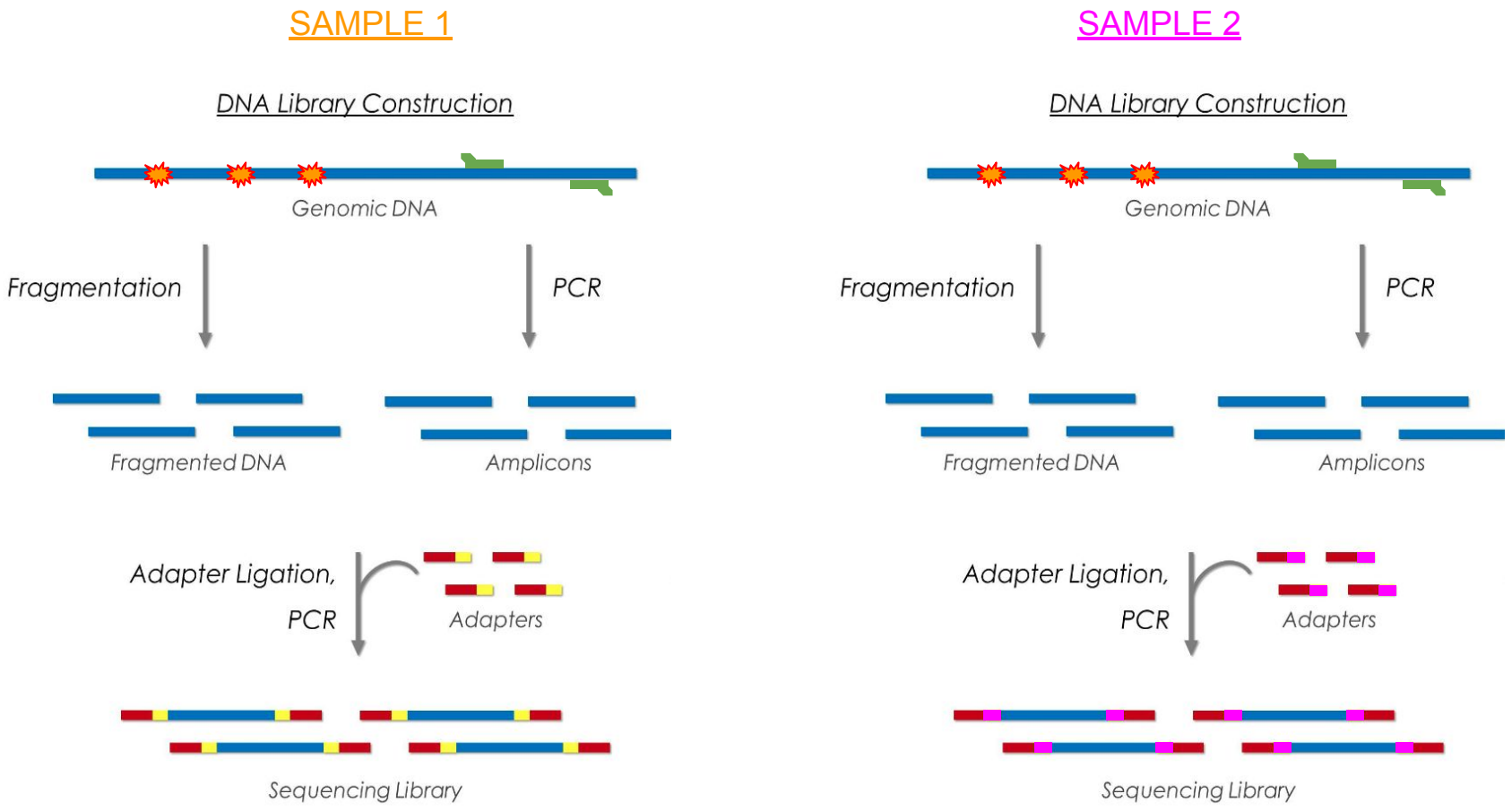
A sequencing “library” must be created from the sample. The DNA (or cDNA) sample is processed into relatively short double-stranded fragments (100–800 bp)



source : https://www.biocompare.com/Molecular-Biology/9187-Next-Generation-Sequencing/

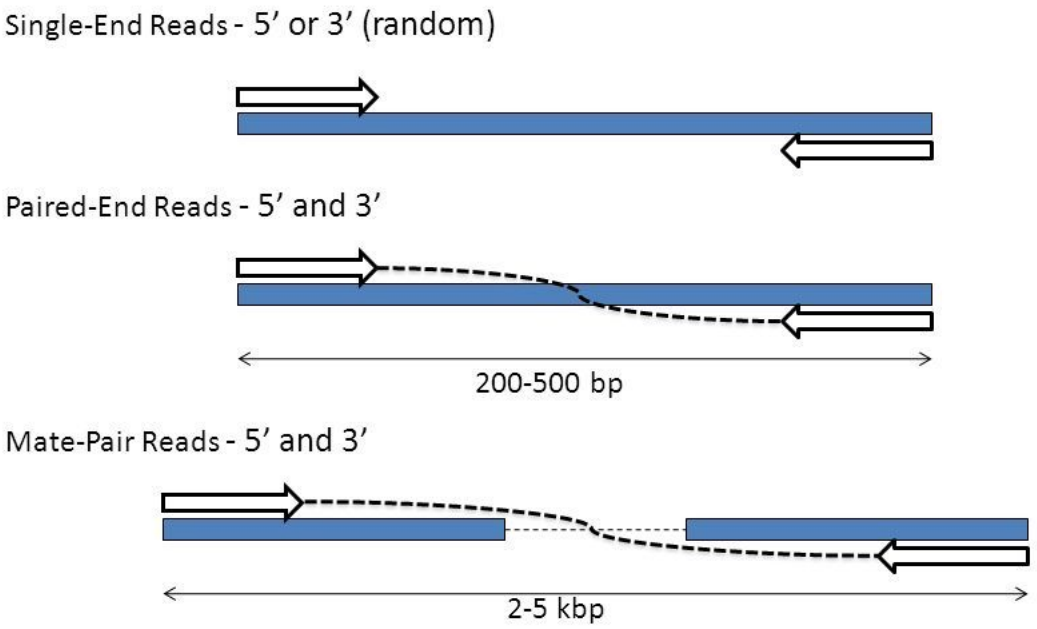
Library construction

Multiplex sequencing using DNA barcoding



Single-end vs paired-end

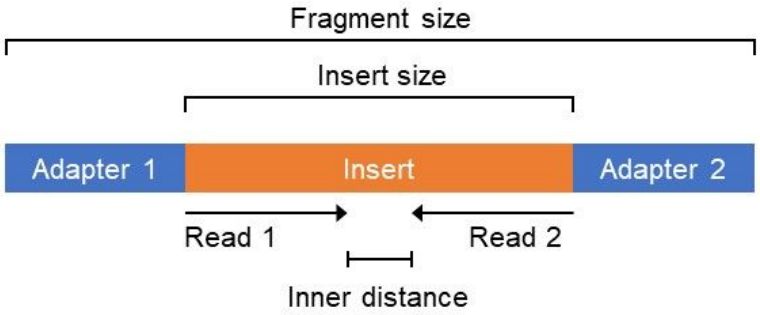
- **Single-End Read:** When sequencing process only occurs in 1 direction
- **Paired-End Read:** When sequencing process occurs in both directions
- **Mate-pair Read:** Short fragments consisting of two segments that originally had a separation of several kilobases in the genome.

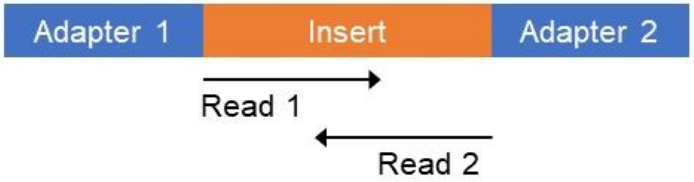


source: http://slideplayer.com/slide/7847747/25/images/7/Types+of+Sequencing+Libraries.jpg

Paired-end

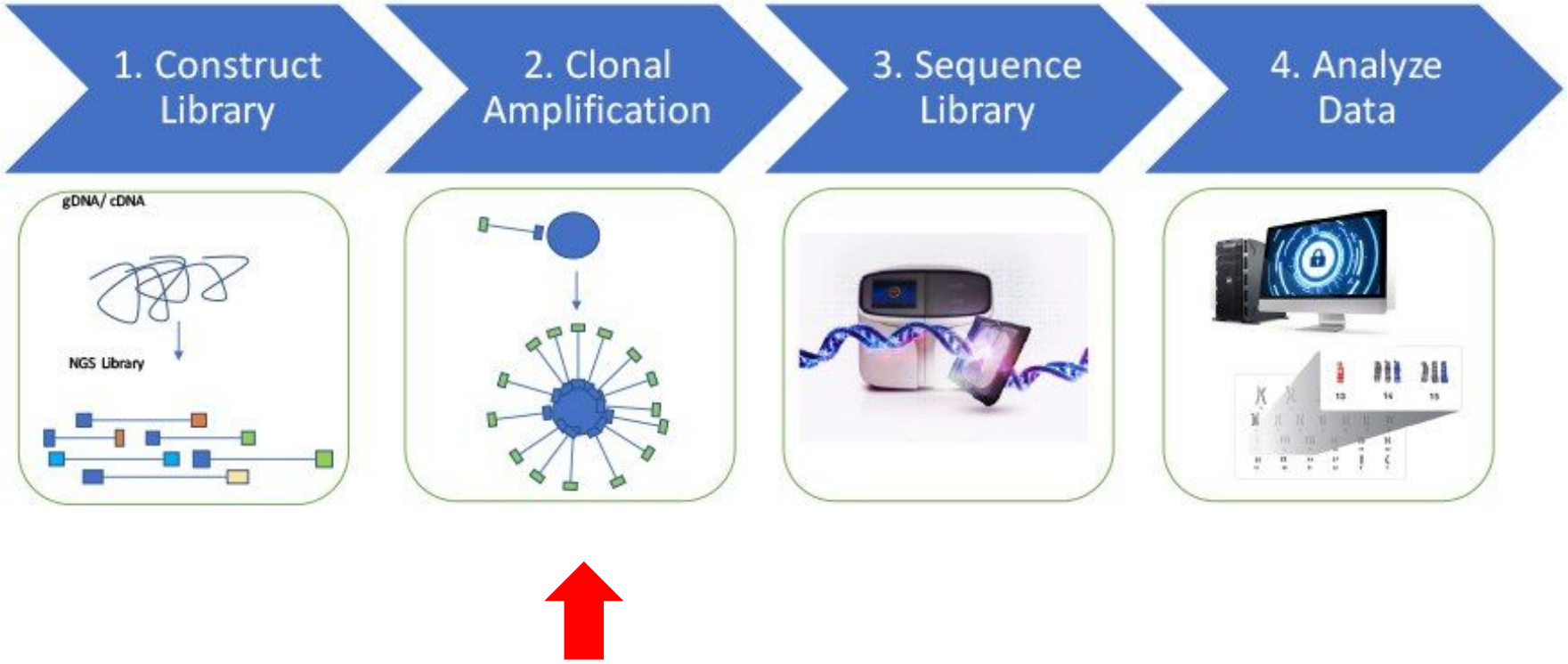
➢ **The insert size** is the size of the piece of DNA of interest, without the adapters.





https://www.badgrammargoodsyntax.com/compbio/2018/3/5/compbio-020-reads-fragments-and-inserts-what-you-need-to-know-for-understanding-your-sequencing-data-1

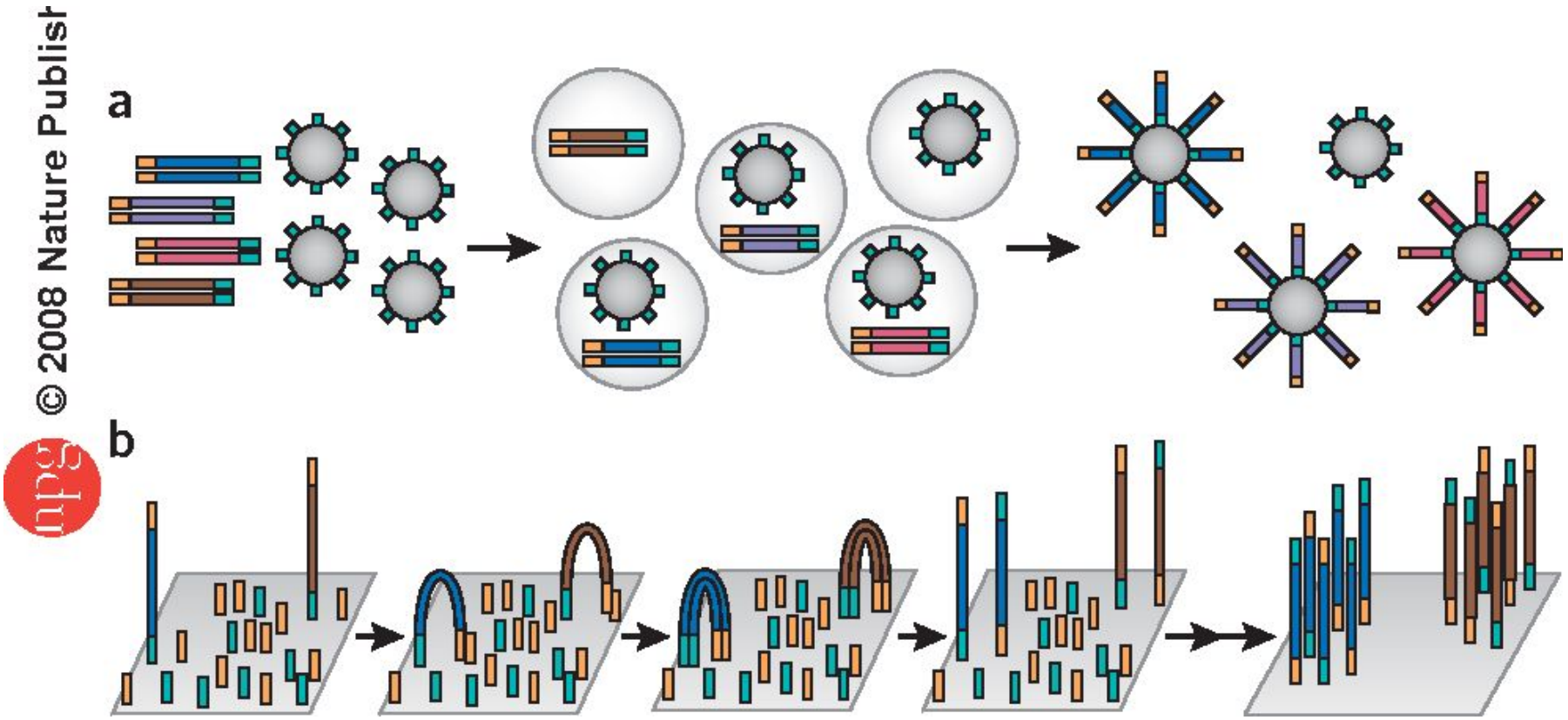
NGS workflow



https://www.thermofisher.com/fr/fr/home/life-science/sequencing/sequencing-education/next-generation-sequencing-basics/what-is-next-generation-sequencing.html

Clonal amplification

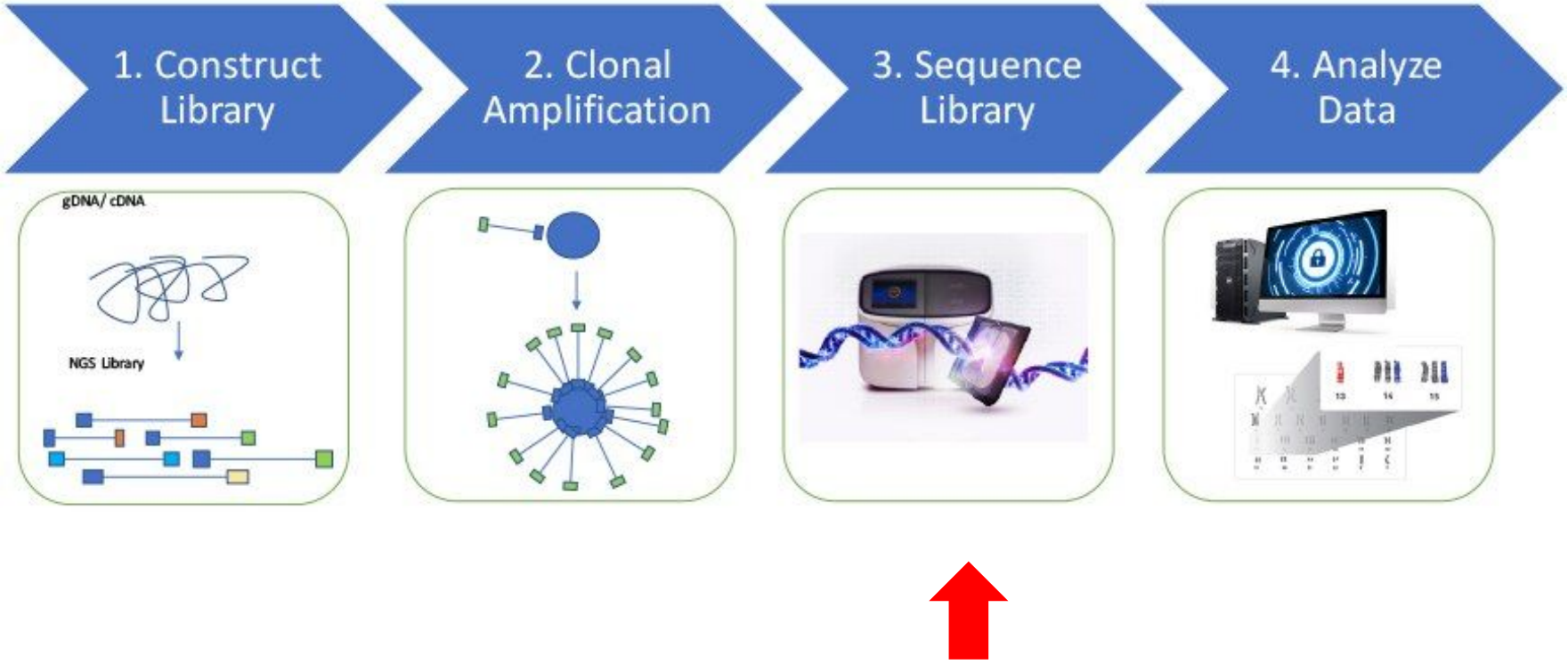
Prior to sequencing, the DNA library must be attached to a solid surface and clonally amplified to increase the signal that can be detected from each target during sequencing.



(a) thermofisher platforms rely on emulsion to amplify clonal sequencing features. (b) The Illumina technology relies on bridge PCR21,22 (aka 'cluster PCR') to amplify clonal sequencing features.

Nature Biotechnology 2008, DOI:10.1038/nbt1486

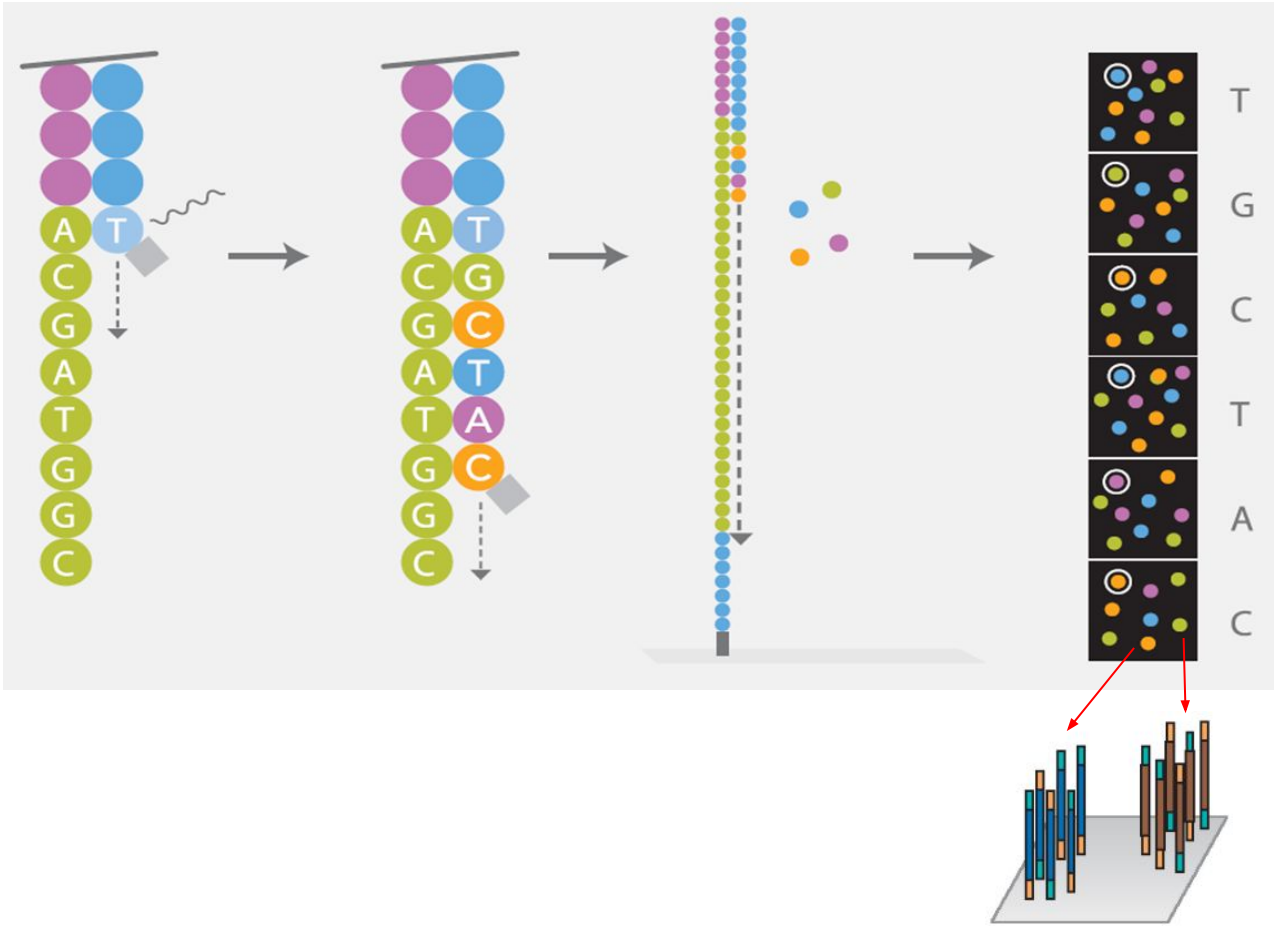
NGS workflow

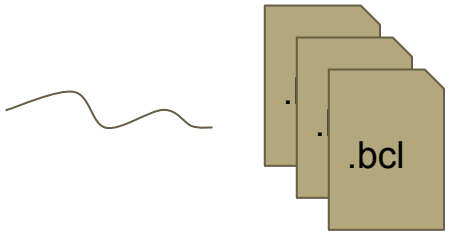


https://www.thermofisher.com/fr/fr/home/life-science/sequencing/sequencing-education/next-generation-sequencing-basics/what-is-next-generation-sequencing.html

Sequencing

Illumina technology

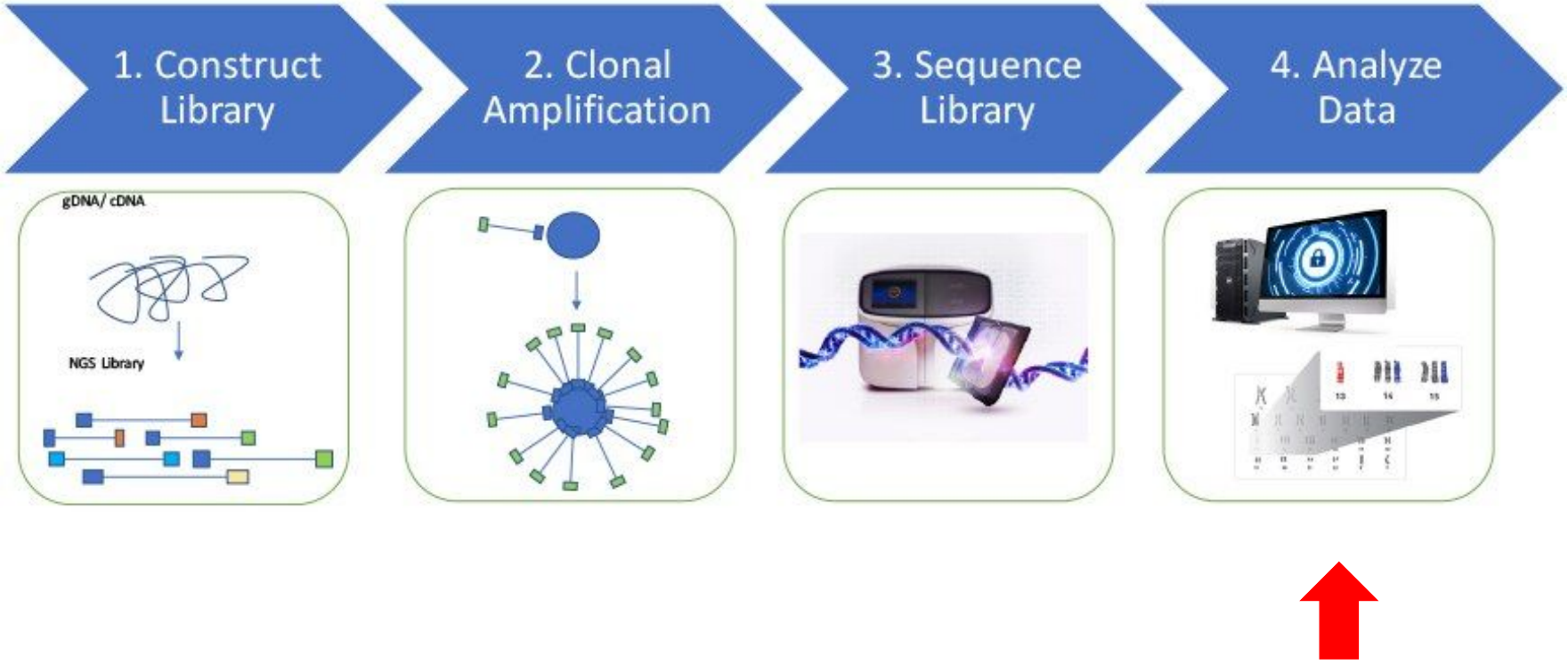




Raw data

https://www.cegat.de/en/services/next-generation-sequencing/

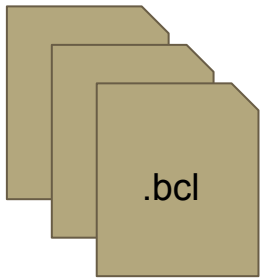
NGS workflow



https://www.thermofisher.com/fr/fr/home/life-science/sequencing/sequencing-education/next-generation-sequencing-basics/what-is-next-generation-sequencing.html

Data analyses

Extracting reads, Demultiplexing



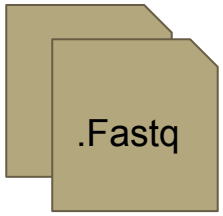
+

Sample Sheet

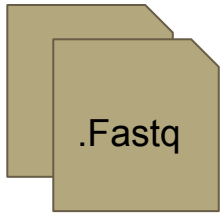
[Header]							
IEMFileVersion	4						
Experiment Name	Project1						
Date	4/16/2016						
Workflow	GenerateFASTQ						
Application	NextSeq FASTQ Only						
Assay	TruSeq LT						
Description							
Chemistry	Default						
[Reads]							
151							
151							
[Settings]							
Adapter	AGATCGGAAGAGCACACGTCTGAACTCCAGTCA						
AdapterRead2	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT						
[Data]							
Sample_ID	Sample_Name	Sample_Plate	Sample_Well	I7_Index_ID	index	Sample_Project	Description
Sample_1				A002	CGATGT		
Sample_2				A004	TGACCA		
Sample_3				A005	ACAGTG		
Sample_4				A006	GCCAAT		

bcl2fastq

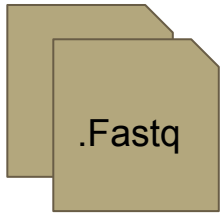
sample 1



sample 2

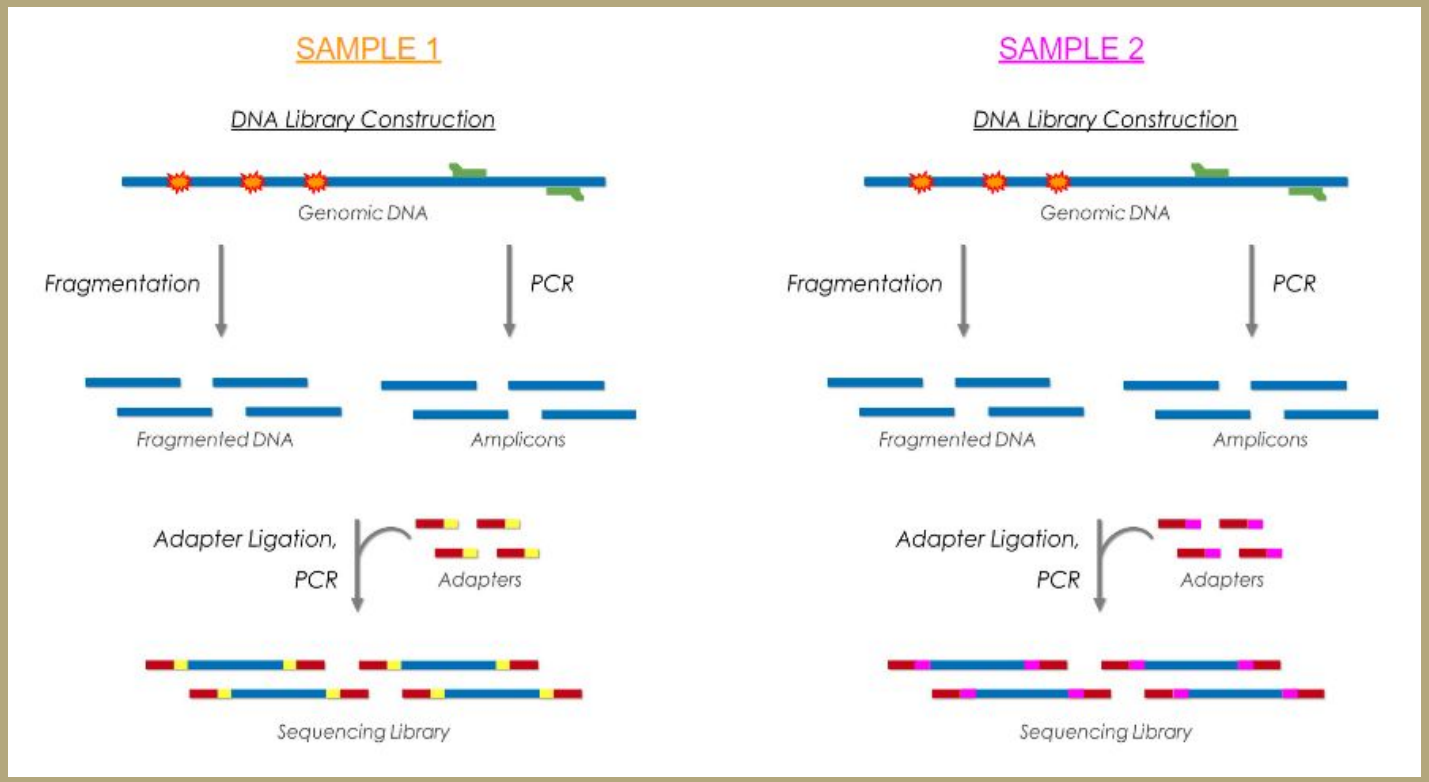


sample 3



Data analyses

Extracting reads, Demultiplexing



sample 1 .Fastq

sample 2 .Fastq

sample 3 .Fastq

Sequencing reads file formats

FastQ

1. Identifier
2. Sequence
4. Quality scores *(as ASCII chars)*

READ

```
@SRR062641.6751359
CGCCCGGCCAATCATTGTGGTTTTAAGTCACTAAGTTTGAGGCTATTTTGTTTTACAGCAAAAGCTAACTGATGCAGACAGGGACAAGTCAGTCTCATCT
+
CBLNPGJQQQJPPQPPQPQRGPPPPRRQQRPSPGRQQQRLRRRMEPQQPMJHQQEHKMMFIIRH?SIIHKNJIKRLJJIKHEABHIFGCGGEFCGDGDCE
```

```
@SRR062634.16249693
CTAAGTTTGAGGCTATTTTGTTTTACAGCAAAAGCTAACTGATGCAGACAGGGACAAGTCAGTCTCATCTCTGTGCACCCAGCATTGCCCAGAACAGGGC
+
ALKMOOOOPPQJQOPPPPPQPPPPPPPRJQRQQQQQRPQPRQQPFQSQQPRLIMHKSNRJQORMFELRPQNQRQJQRRPQQLIRKDMKQJPN8CFDGCCCB

@SRR062634.20060465
CTCCCAGCTTCCAACAGACCCTGTCCCAGCTCCCTCCAAGCTGAGTGTTGGCCTGATACCTACCAGTGGAGCGAGGGGAACCCGAGGACTGCCAAGGGCA
+
D?KMPQEPGCPQQNPQIQIGR@DPERQHEKBED=HCHG8EHFDCD6<329@<:69A<6,;<967>;=C:>AA8BBED######################
```

FastA

```
>SRR062641.6751359
CGCCCGGCCAATCATTGTGGTTTTAAGTCACTAAGTTTGAGGCTATTTTGTTTTACAGCAAAAGCTAACTGATGCAGACAGGGACAAGTCAGTCTCATCT

>SRR062634.16249693
CTAAGTTTGAGGCTATTTTGTTTTACAGCAAAAGCTAACTGATGCAGACAGGGACAAGTCAGTCTCATCTCTGTGCACCCAGCATTGCCCAGAACAGGGC

>SRR062634.20060465
CTCCCAGCTTCCAACAGACCCTGTCCCAGCTCCCTCCAAGCTGAGTGTTGGCCTGATACCTACCAGTGGAGCGAGGGGAACCCGAGGACTGCCAAGGGCA
```

Mais aussi: FAST5, BAM, ...

Fastq files (Paired-end)

2 files : R1, R2

Reads1.fq

```
@ERR229776.100000840
CTAGGAAGCGTAGTCCTGGGGTCATCTCTCCTATTAATACTGTTGGGGAATGTTTAGTA
+
BAEEAGEED96EHFE@BF><>EAAC;EBH<K<6:HJGFFHBC>DDIKG4AIHFFD@0/=
@ERR229776.100020365
CATTATTTCATAGTAGCCAAAAAGTGGAAACAGTCAAAATATCCGTCAGTGAATTGACC
+
1.*/./&((&3=;B@F860C>@51(3:).6GG-68C*:CG)#B4/=HDJ6;79)<@C/
@ERR229776.100104918
TATTTCTGGAATTTTCCATTTAATATTTTCAGACTGCAGTTGACTGCGGGTAACTGAAA
+
CEEEEEFEDAEGGGFDHGFFHGIHHHIIIIGKHBKJJIGHFHKILJKLEJLJJIFJMJK
```

Reads2.fq

```
@ERR229776.100000840
TTCTGGTCAGTAAGACCTCAAAAGGTTAAATACTAGCGATTTACACACCTTAAATGATT
+
CFIEEG@FFFGKFJHJ>HHKLLJIIJILLJIILJHKAKJKKJJJJJJLMKJMKJJJJKJ
@ERR229776.100020365
CCTAAAATGGTGTGTTTTCGTATATTCACAATGCTGTGGAACCATCACCACTATCTGAT
+
4B@EDFF=(/CHBHEHCE6@ED8E@@I6HJB6E:6%@C46FFIBGCIGKD,DN=CBBE@
@ERR229776.100104918
TCTTTCTTTTGTTTTTTTTTCTGAGATGTCTTTTGTTTTTGTTCTGAGGTCTTGTTATG
+
CFIGGGKHHHFHHFIJIIIJKLIIHJIIIKLJKKIJKLLKJFJJMHJJLFJMJIKKJJJ
```

1 interleaved paired file

Reads.fq

```
@SRR531199.1 ILLUMINA 0130:3:1101:1249:1993 length=101
TTTTCAGAGTAGTTGGTACCCAATTGGAAGATGTGACCCACTTCGATACCGCGCTTGAG
+
dfffffffffdffeffdadffffeeefdeffeffeffffffffffddeefeYdfefefe[e
@SRR531199.1 ILLUMINA 0130:3:1101:1249:1993 length=99
ANNNNNNCTTCGGTATNAACTGGGGNNNNGATGTTGAACTGGGTAAAGTCGAAGATCTG
+
BBBBBBBSZTUVWO]YB_[cbabbWBBBBSVVUUgggadcdedbedcddfffdegeggef
@SRR531199.2 ILLUMINA_0130:3:1101:1463:1964 length=101
NTGAGTAGCTCAATGCGCTGACGCCAATAGCTATACCAACGACTGGCCAGATTATGTTT
+
BXSSRU[X[Wcc_ccccccccccc_ccccccccc_ccccccccccccccccccccc
@SRR531199.2 ILLUMINA_0130:3:1101:1463:1964 length=99
AAGTGACCCATCGCGATAAAGTGCTGCGCAGTAAANAGCANCTGTTNGATGCTGGCTTA
+
ggggggggggggggggggfgfgggggggggggg^BbbbaBbbaZ]BZ[ccccfggggg
@SRR531199.3 ILLUMINA_0130:3:1101:1366:1970 length=101
NAAGTCGCGGCGACCCCTATCGTGGCTTTCGGCGTACGCCATTTCAATGCGGCCGCCGC
+
B[[X[YY[YVcc_cccc_cc__________[[[V[^^^^^V[[]SXWUX[\\]]Z^^^B
@SRR531199.3 ILLUMINA_0130:3:1101:1366:1970 length=99
TGGTCAATACAAGCCGCAATACCTGCATCATGCGGNGGAANAATTTGCGCGCCGTTTTC
+
ggfegggggggdeggggfgcgggaggggggggega^Bb`^]B[Y[[[Zffffh_afeefe
```

Day 1: Bases of NGS analyses for RNA-seq

- NGS Introduction
- Reads Quality Control

Reads quality

- Errors when reading bases
- Depends on sequencing technologie
- Error rate tends to increase with read size

⇒ For each position in the read

- One base (A/T/C/G)
- One error probability

Phred Quality Score (for a base)

Phred quality scores ***Q*:** logarithmically related to the base-calling error probabilities P

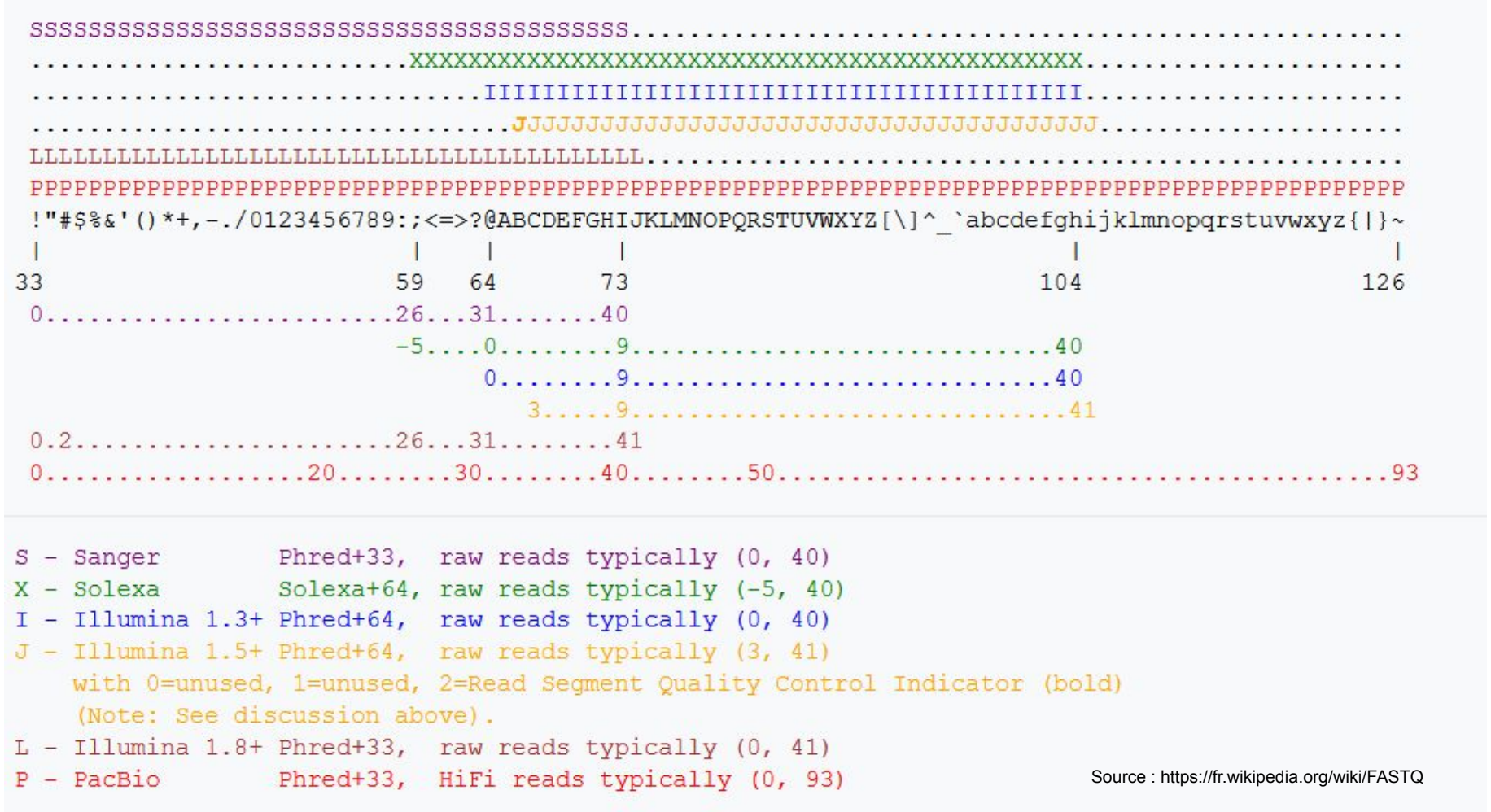
$$Q = -10 \log_{10} P$$

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

source: https://en.wikipedia.org/wiki/Phred_quality_score

Quality score encoding

For history reasons, more than one coding convention



S - Sanger Phred+33, raw reads typically (0, 40)
X - Solexa Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64, raw reads typically (3, 41)
with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
(Note: See discussion above).
L - Illumina 1.8+ Phred+33, raw reads typically (0, 41)
P - PacBio Phred+33, HiFi reads typically (0, 93)

Source : https://fr.wikipedia.org/wiki/FASTQ

ASCII table

Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	*[NULL]*	32	20	*[SPACE]*	64	40	@	96	60	`
1	1	*[START OF HEADING]*	33	21	!	65	41	A	97	61	a
2	2	*[START OF TEXT]*	34	22	"	66	42	B	98	62	b
3	3	*[END OF TEXT]*	35	23	#	67	43	C	99	63	c
4	4	*[END OF TRANSMISSION]*	36	24	$	68	44	D	100	64	d
5	5	*[ENQUIRY]*	37	25	%	69	45	E	101	65	e
6	6	*[ACKNOWLEDGE]*	38	26	&	70	46	F	102	66	f
7	7	*[BELL]*	39	27	'	71	47	G	103	67	g
8	8	*[BACKSPACE]*	40	28	(	72	48	H	104	68	h
9	9	*[HORIZONTAL TAB]*	41	29	)	73	49	I	105	69	i
10	A	*[LINE FEED]*	42	2A	*	74	4A	J	106	6A	j
11	B	*[VERTICAL TAB]*	43	2B	+	75	4B	K	107	6B	k
12	C	*[FORM FEED]*	44	2C	,	76	4C	L	108	6C	l
13	D	*[CARRIAGE RETURN]*	45	2D	-	77	4D	M	109	6D	m
14	E	*[SHIFT OUT]*	46	2E	.	78	4E	N	110	6E	n
15	F	*[SHIFT IN]*	47	2F	/	79	4F	O	111	6F	o
16	10	*[DATA LINK ESCAPE]*	48	30	0	80	50	P	112	70	p
17	11	*[DEVICE CONTROL 1]*	49	31	1	81	51	Q	113	71	q
18	12	*[DEVICE CONTROL 2]*	50	32	2	82	52	R	114	72	r

Example for score interpretation using sanger encoding

```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
  ............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
  ..............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
  LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                         |    |       |                           |                     |
 33                        59   64      73                         104                   126
  0.....................26...31.......40
                           -5....0........9.............................40
                                0........9.............................40
                                   3.....9.............................41
  0.2...................26...31........41

 S - Sanger       Phred+33,  raw reads typically (0, 40)
```

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

S - Sanger Phred+33

```
Bad      :   0-19
Correct  :  20-29
Good     :  30-40
```

```
@SEQ:ID
ACTGTACGATCGATCGCATGCATCAGTACGTCGTACCAGAT
+
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
|                     |         |         |
0.........1.........2.........3.........4
01234567890123456789012345678901234567890
```

Quality Control (QC)

Quality Control (QC) is important to:

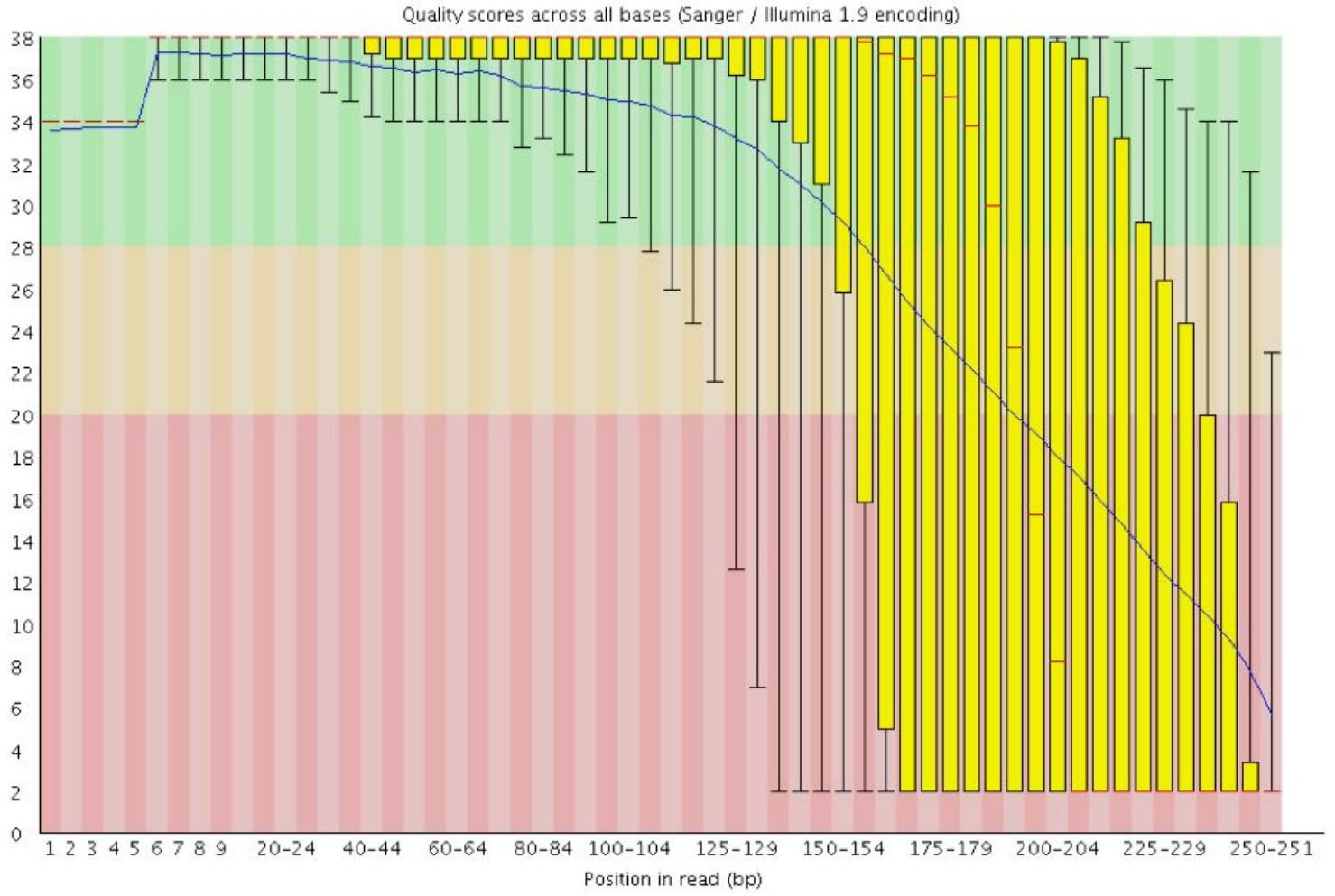
- Check if your sample sequencing went well
- Know when you need to sequence again (sequencing platform QC fail)
- Identify potential problems that can be fixed, or not
- Follow the impact of preprocessing steps

⇒ FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)

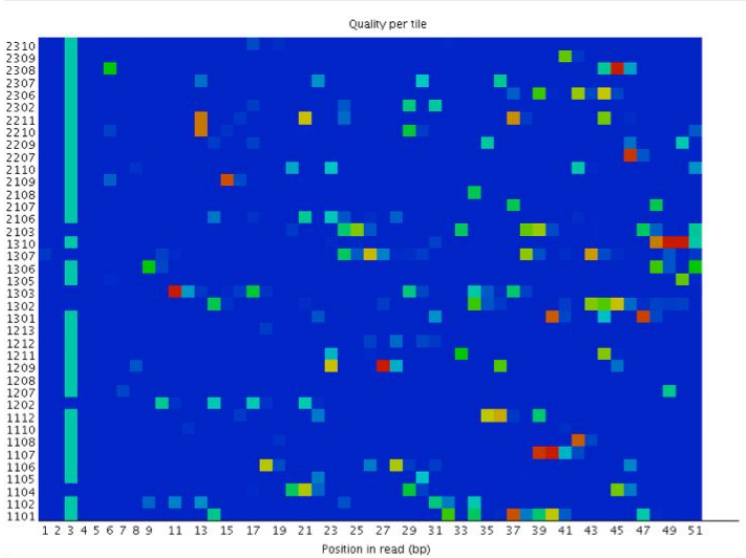
\+ MultiQC (https://multiqc.info/) when comparing multiple datasets

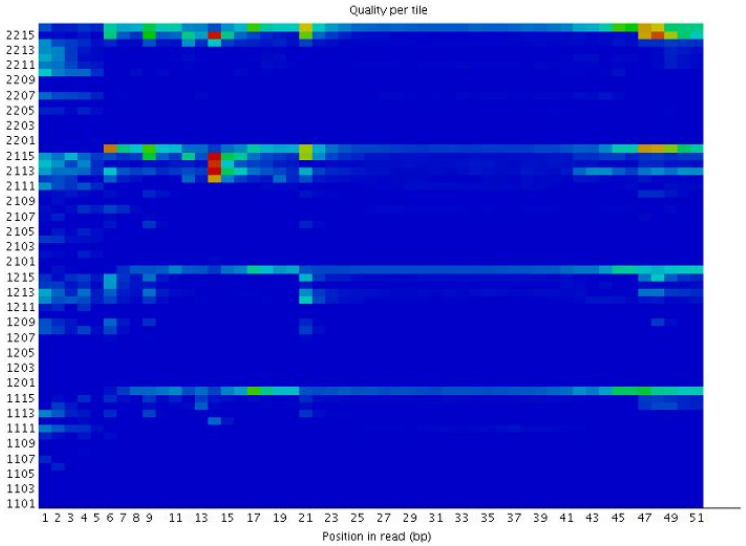
Loss of base call accuracy with increasing sequencing cycles

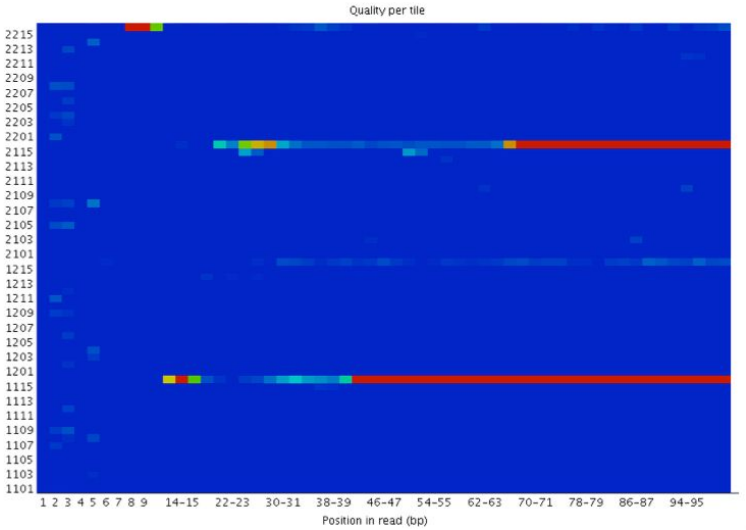
Source: https://sequencing.qcfail.com

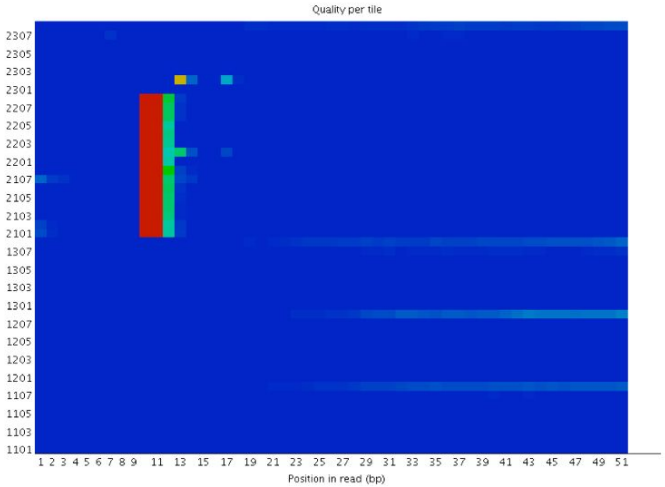


Position specific failures of flowcells





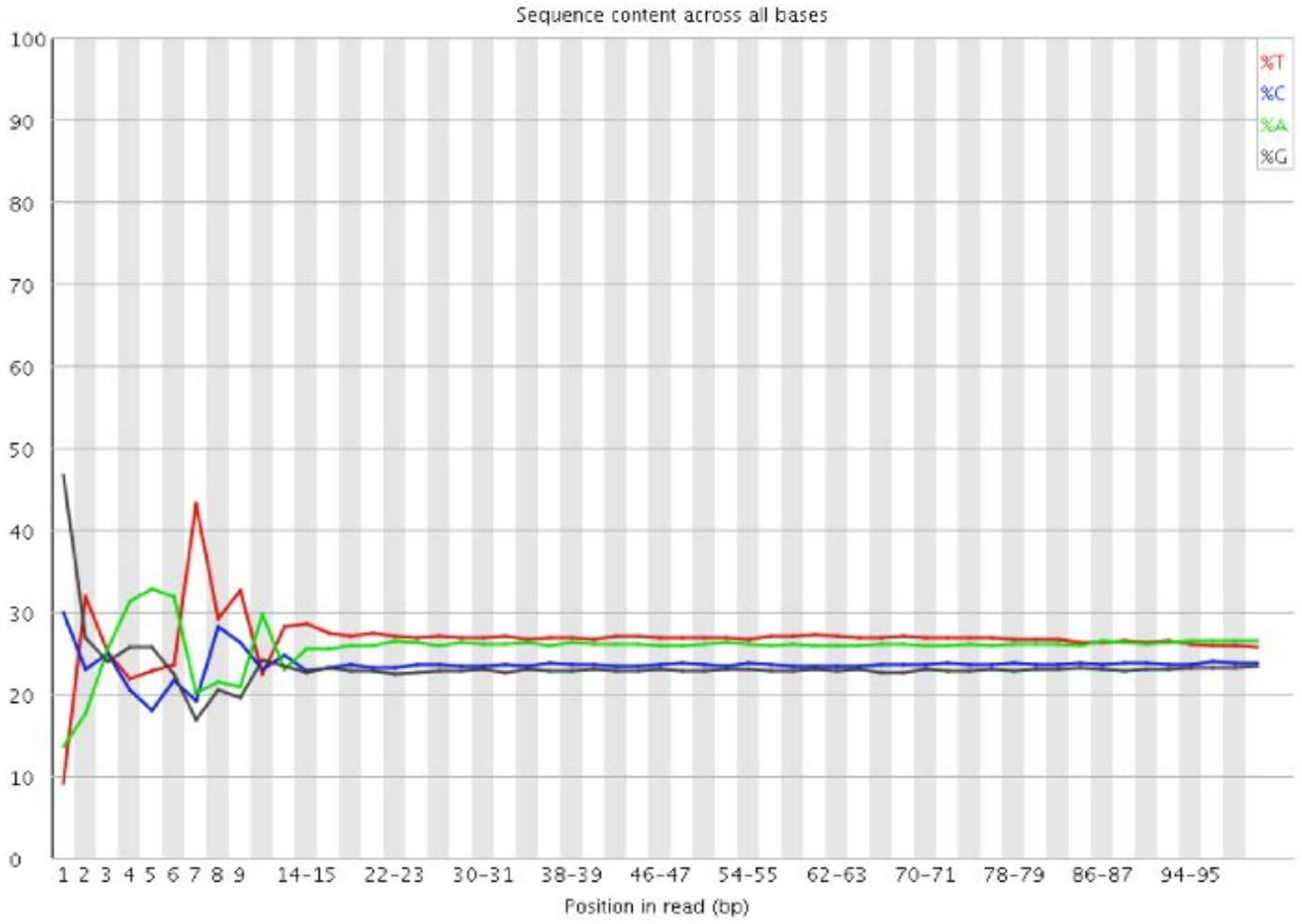




Source: https://sequencing.qcfail.com

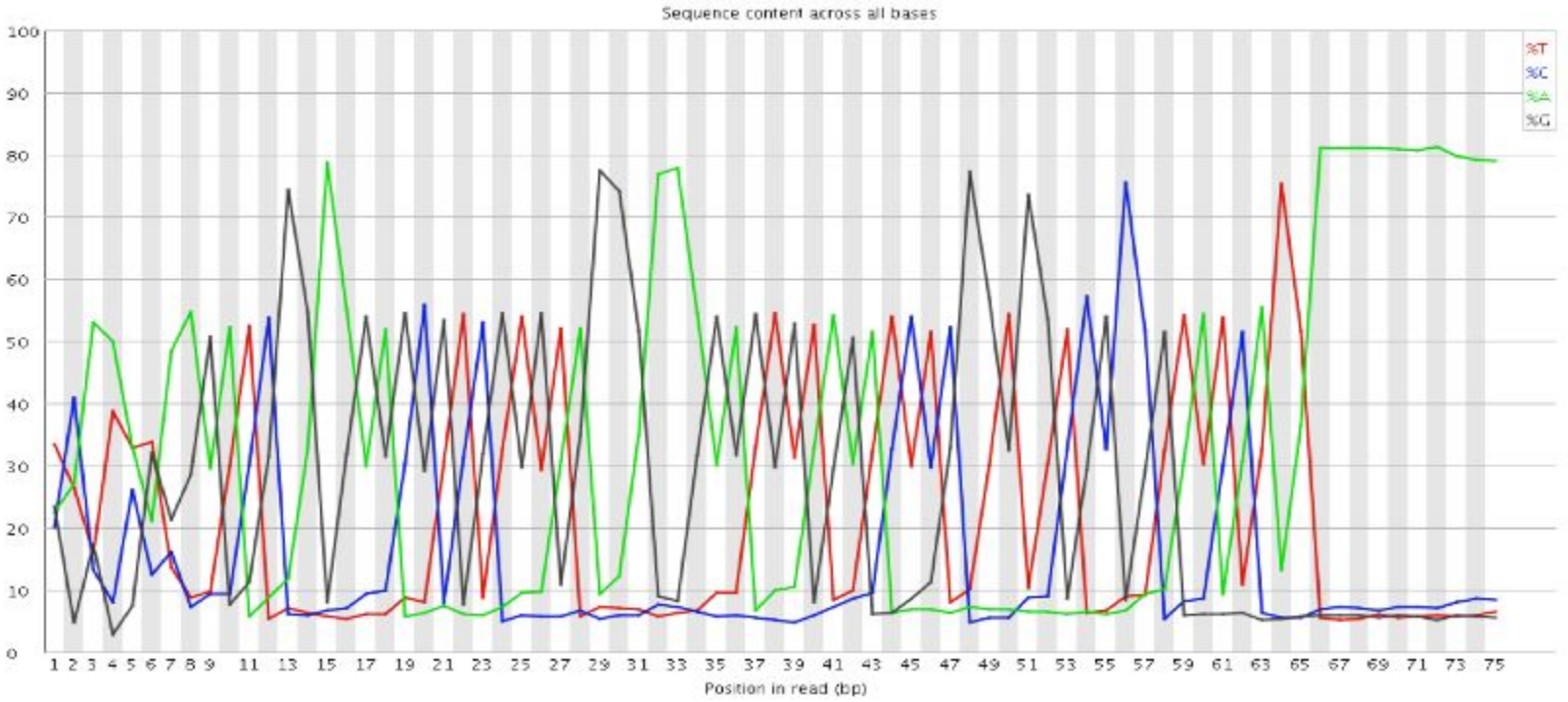
Positional sequence bias in random primed libraries

Source: https://sequencing.qcfail.com



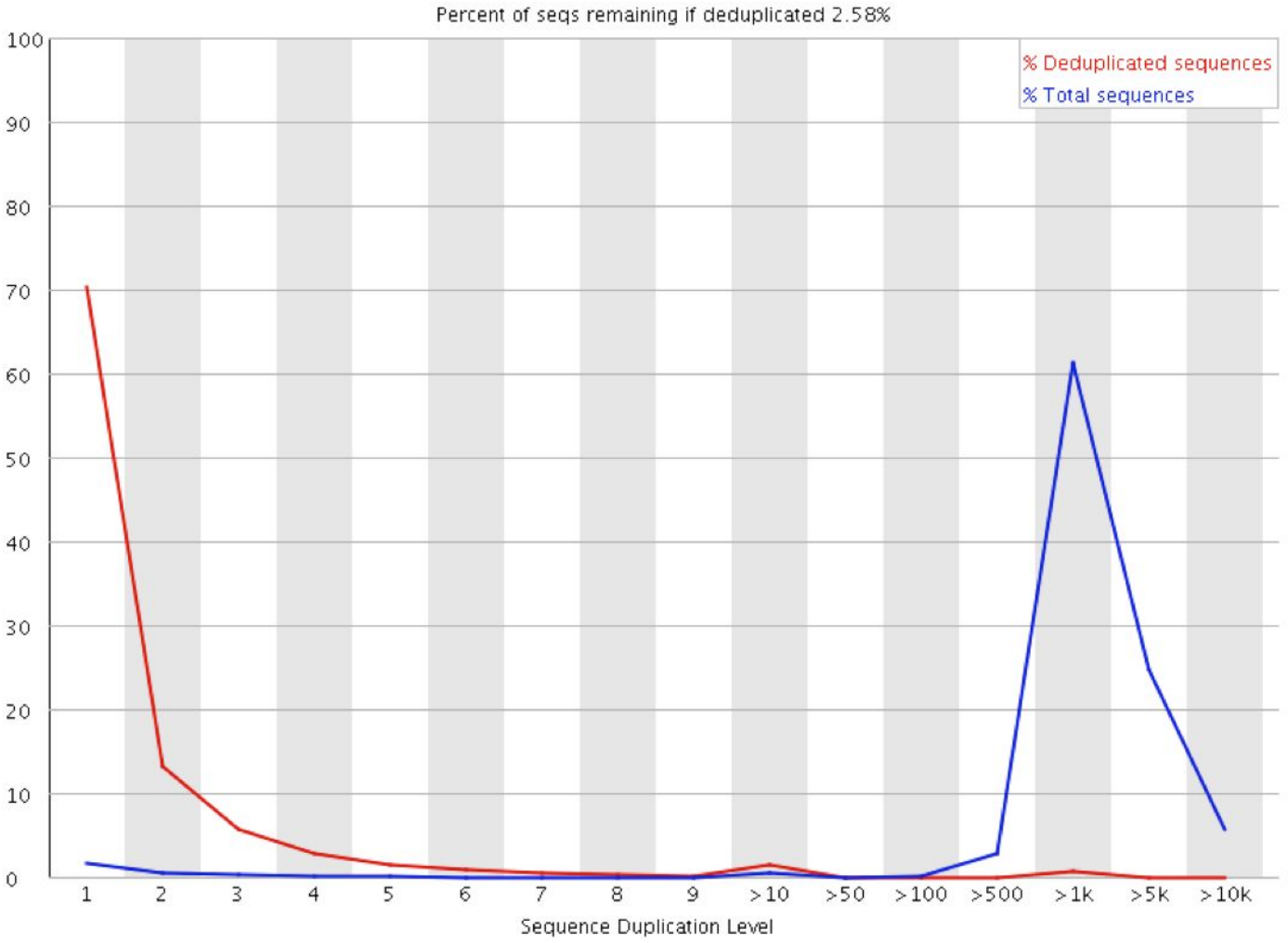
Contamination with adapter dimers

Source: https://sequencing.qcfail.com



Libraries contain technical duplication

Source: https://sequencing.qcfail.com

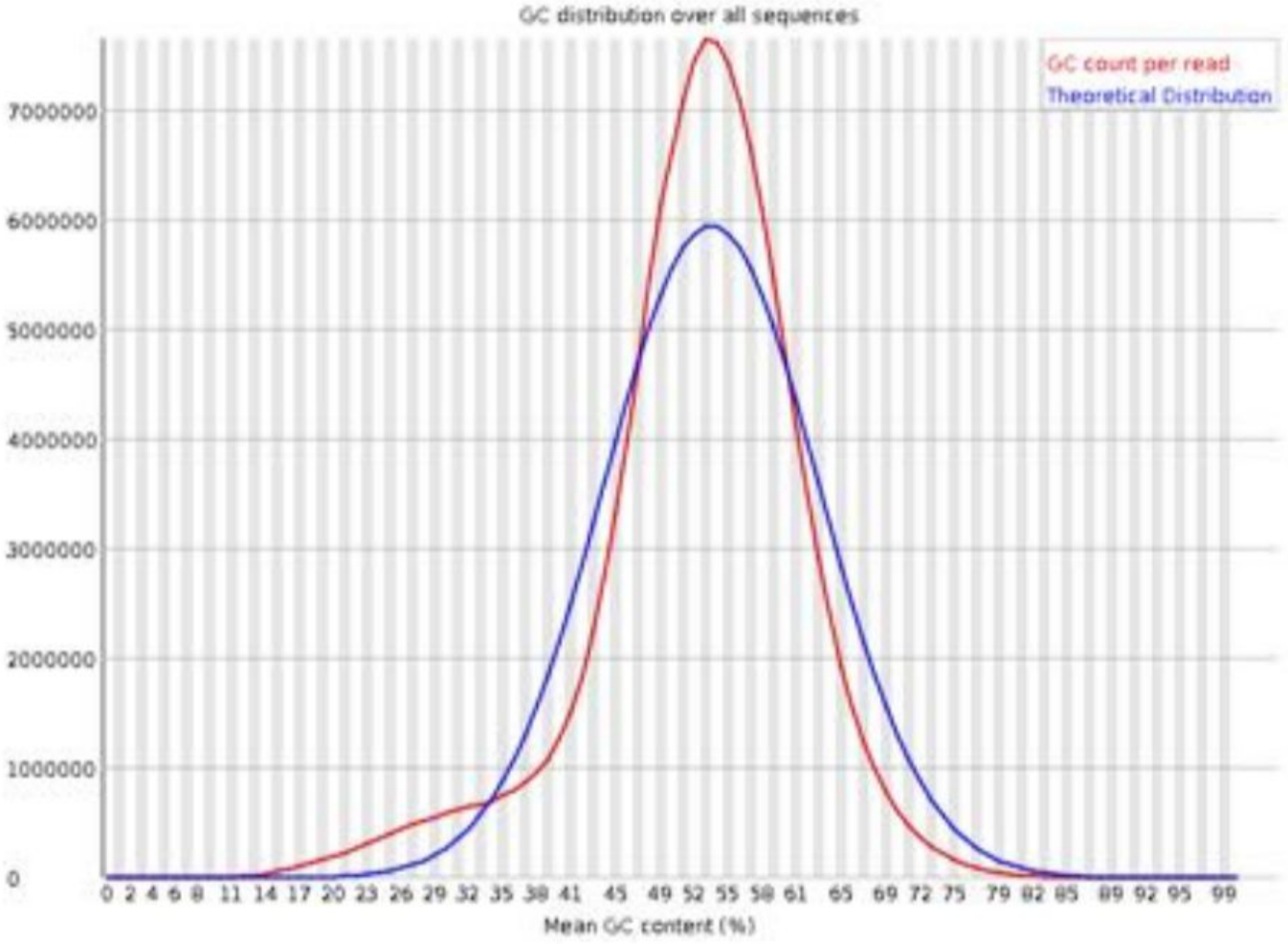


Why do we not expect duplicate reads in DNA-seq?

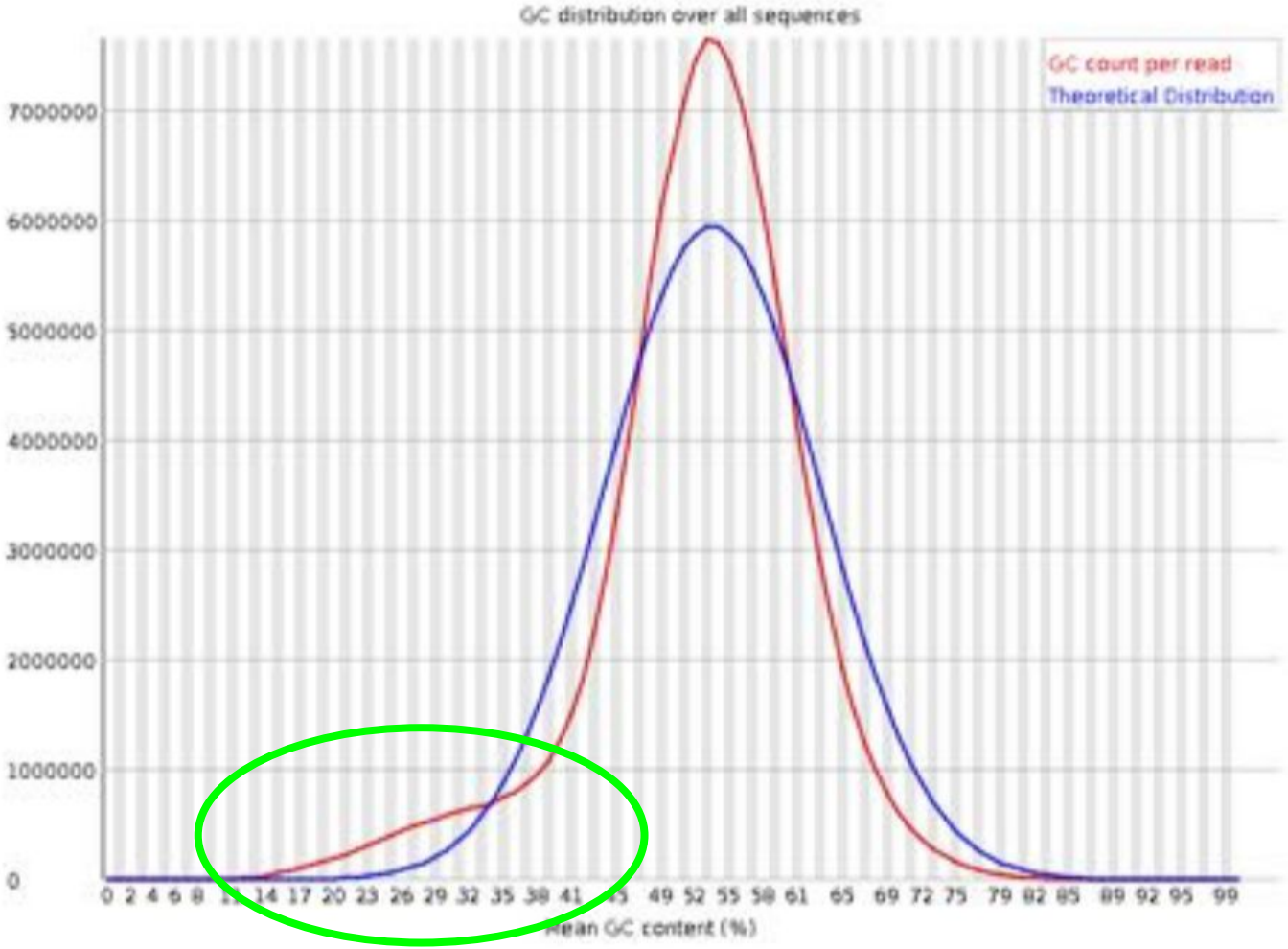
Why do we not expect duplicate reads in DNA-seq?

- Reads are randomly distributed on the genome
- Unless a region has a very high depth of coverage, the probability of two reads being identical is very low

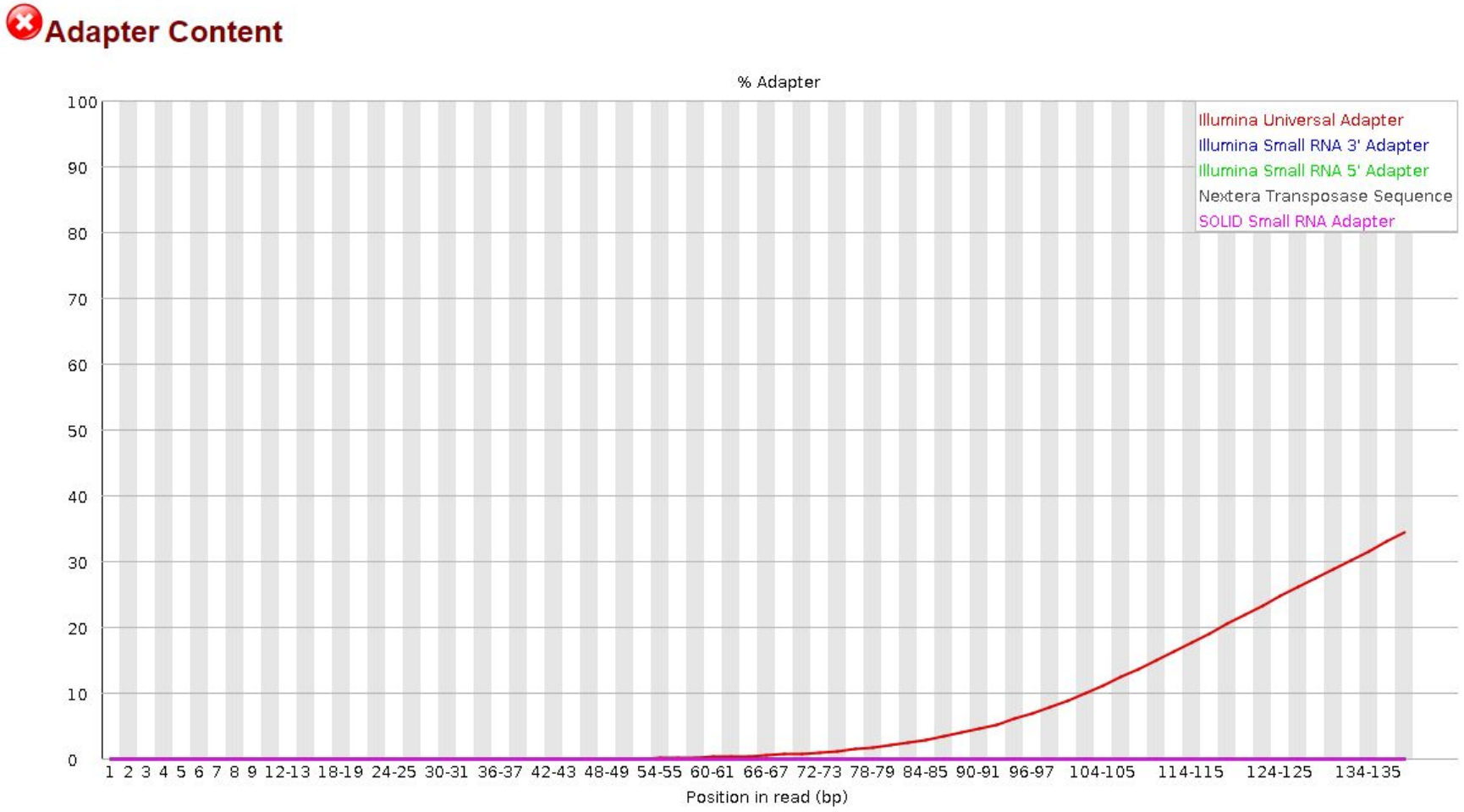
GC content / Contamination ?



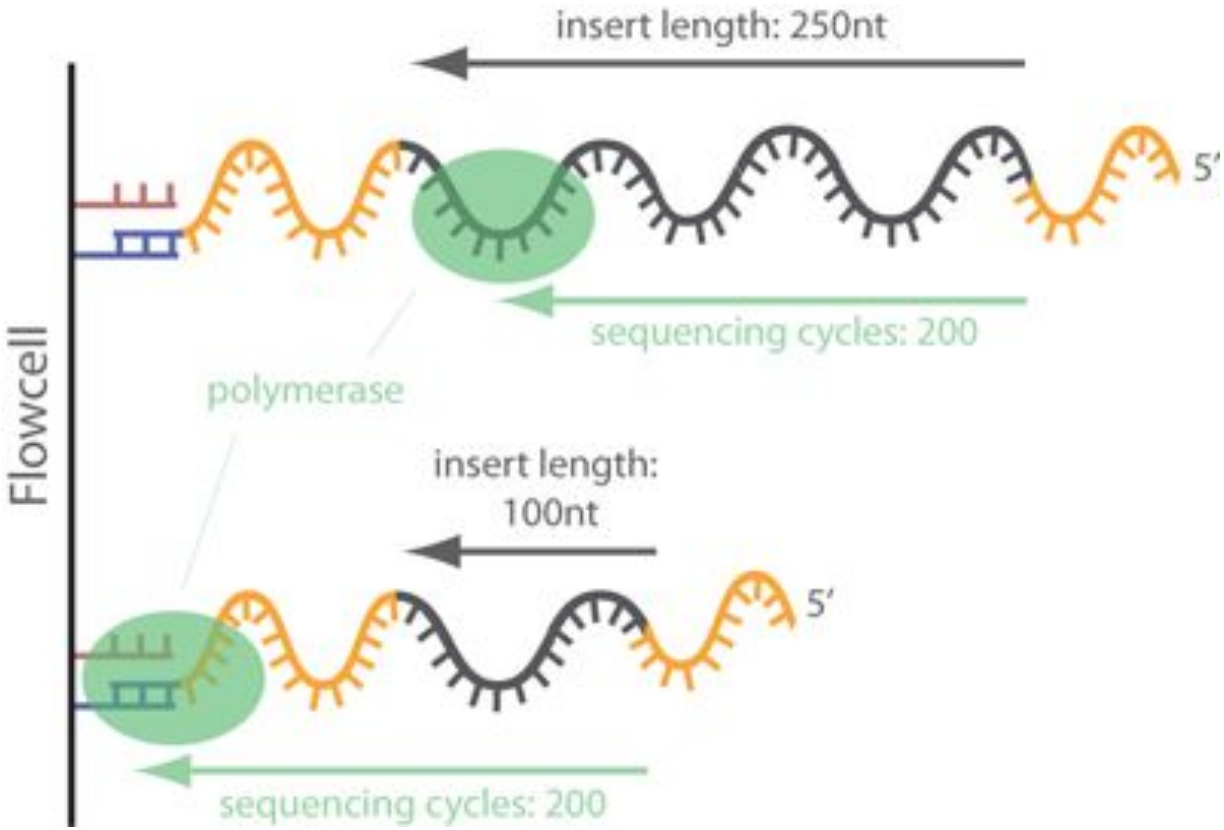
GC content / Contamination ?



Adapter content



Adapter content



https://www.ecseq.com/support/ngs/trimming-adapter-sequences-is-it-necessary

Day 1: Bases of NGS analyses for RNA-seq

- NGS Introduction
- Reads Quality Control
- Reads Cleaning

Goal: read cleaning

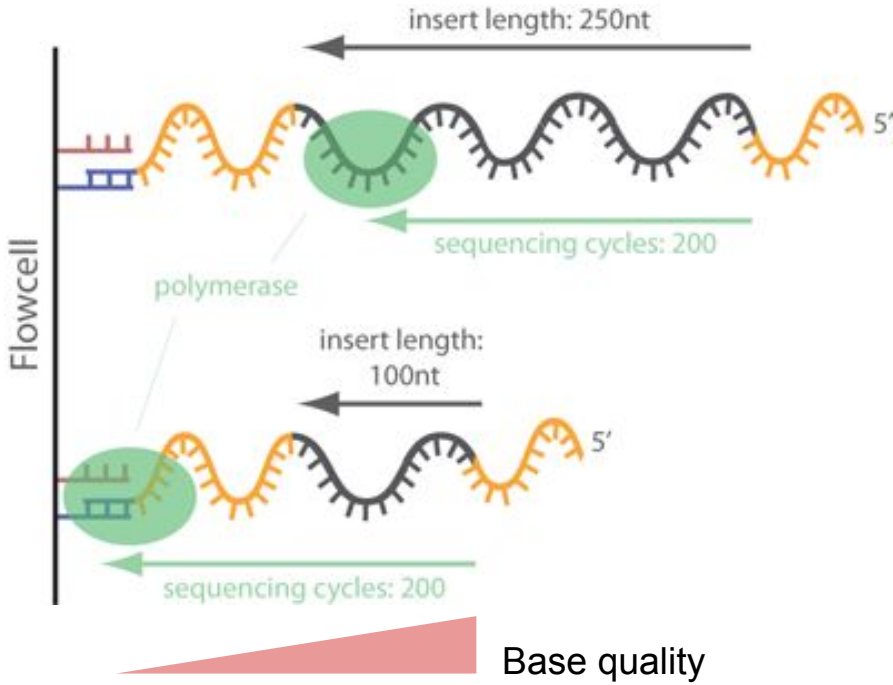
RAW

```
@SRR062641.6751359
CGCCCGGCCAATCATTGTGGTTTTAAGTCACTAAGTTTGAGGCTATTTTGTTTTACAGCAAAAGCTAACTGATGCAGACAGGGACAAGTCAGTCTCATCT
+
CBLNPGJQQQJPPQPPQPQRGPPPPRRQQRPSPGRQQQRLRRRMEPQQPMJHQQEHKMMFIIRH?SIIHKNJIKRLJJIKHEABHIFGCGGEFCGDGDCE
@SRR062634.16249693
CTAAGTTTGAGGCTATTTTGTTTTACAGCAAAAGCTAACTGATGCAGACAGGGACAAGTCAGTCTCATCTCTGTGCACCCAGCATTGCCCAGAACAGGGC
+
ALKMOOOOPPQJQOPPPPPQPPPPPPPRJQRQQQQQRPQPRQQPFQSQQPRLIMHKSNRJQORMFELRPQNQRQJQRRPQQLIRKDMKQJ#KNFDGCCCB
@SRR062634.20060465
CTCCCAGCTTCCAACAGACCCTGTCCCAGCTCCCTCCAAGCTGAGTGTTGGCCTGATACCTACCAGTGGAGCGAGGGGAACCCGAGGACTGCCAAGGGCA
+
D?KMPQEPGCPQQNPQIQIGR@DPERQHEKBE#HCHG8EHFDCD6<329@<:69A<6,;<967>;=C:>AA8BBED######################
@SRR062635.15516129
AAAAAAAAAAAAAAAAAAAAAAAAAAAAGGGGGCCCCCCTTTCCCCCCCGGGGGGGGGACAGGGGGGGTGTTCGGGCCCCGCGCCGCCCTTGACCACGG
+
EKLMPPPPPQQQQQQQQQQQQQQQQH#########################################################################
```

Clean

```
@SRR062641.6751359
CGCCCGGCCAATCATTGTGGTTTTAAGTCACTAAGTTTGAGGCTATTTTGTTTTACAGCAAAAGCTAACTGATGCAGACAGGGACAAGTCAGTCTCATCT
+
CBLNPGJQQQJPPQPPQPQRGPPPPRRQQRPSPGRQQQRLRRRMEPQQPMJHQQEHKMMFIIRH?SIIHKNJIKRLJJIKHEABHIFGCGGEFCGDGDCE
@SRR062634.16249693
CTAAGTTTGAGGCTATTTTGTTTTACAGCAAAAGCTAACTGATGCAGACAGGGACAAGTCAGTCTCATCTCTGTGCACCCAGCATTGCCCAGAACAGGGC
+
ALKMOOOOPPQJQOPPPPPQPPPPPPPRJQRQQQQQRPQPRQQPFQSQQPRLIMHKSNRJQORMFELRPQNQRQJQRRPQQLIRKDMKQJ#KNFDGCCCB
@SRR062634.20060465
CTCCCAGCTTCCAACAGACCCTGTCCCAGCTCCCTCCAAGCTGAG
+
D?KMPQEPGCPQQNPQIQIGR@DPERQHEKBE#HCHG8EHFDCD
```

Reads cleaning



- Cut adaptators at read ends

- Trimming : cut read ends (5' ou 3')
  - Fixed number of bases
  - Individual base quality
  - Mean quality of bases in a sliding window

- Filtering : remove read
  - Size criteria (example < 60bp)
  - Mean base quality for all bases criteria (example < 25)

Reads cleaning example

Tool: Trimmomatic

01
- Clean adaptators

02
- Trimming 5' et 3' on base quality (> 3)

03
- Trimming using sliding window (4 bases, Q < 20)

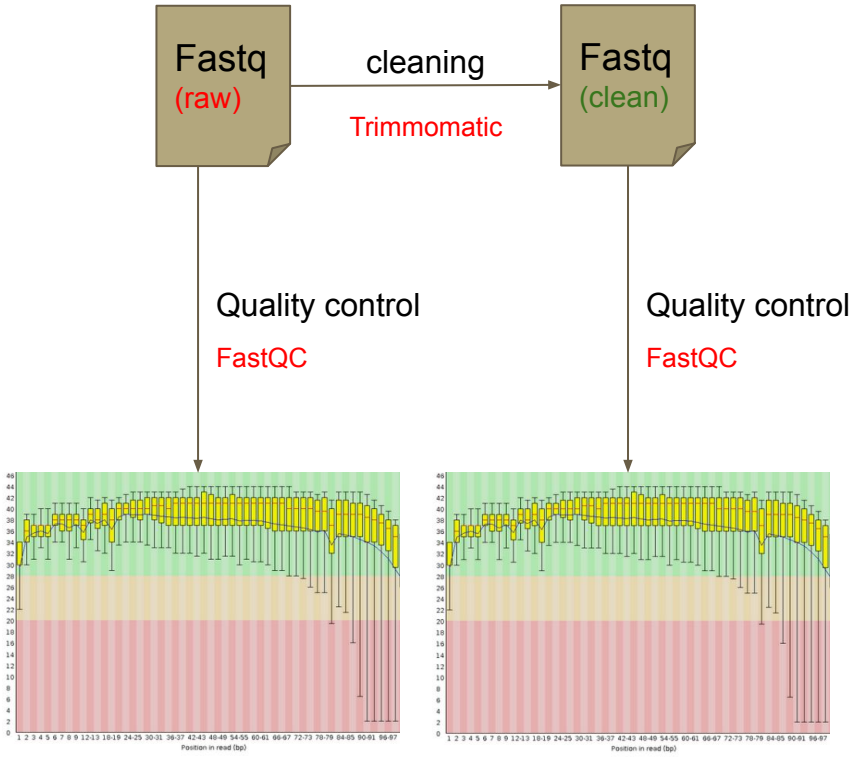
04
- Filtering on mean read quality (Q < 25)

05
- Filtering on read size (taille < 20)

Source : Erwan Core, 5ème Ecole de bioinformatique AVIESAN-IFB 2016 https://www.france-bioinformatique.fr/fr/evenements/EBA2016

Bolger, A. M. and Lohse, M. and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. In Bioinformatics, 30 (15), pp. 2114–2120

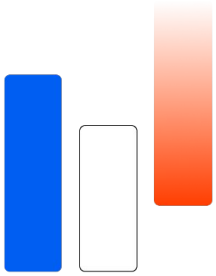
Workflow


Fastq
(raw)
cleaning
Trimmomatic
Fastq
(clean)
Quality control
FastQC
Quality control
FastQC

usegalaxy.fr presentation

Practical: Quality Control (QC) & Cleaning

Open Galaxy



usegalaxy.fr

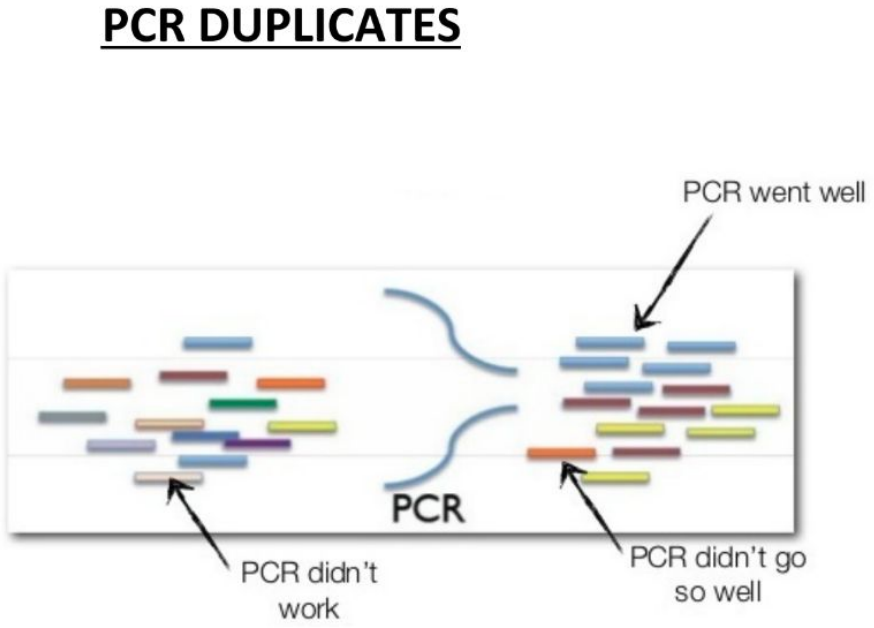
Practical:
https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html

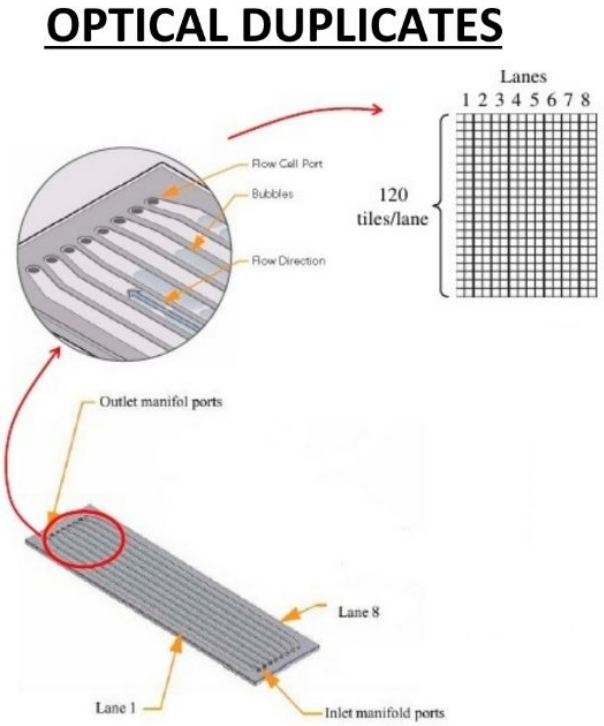
TIAAS: https://usegalaxy.fr/join-training/bilille-rnaseq-2023/

Mapping

Cleaning duplicated reads

How do duplication events arise?





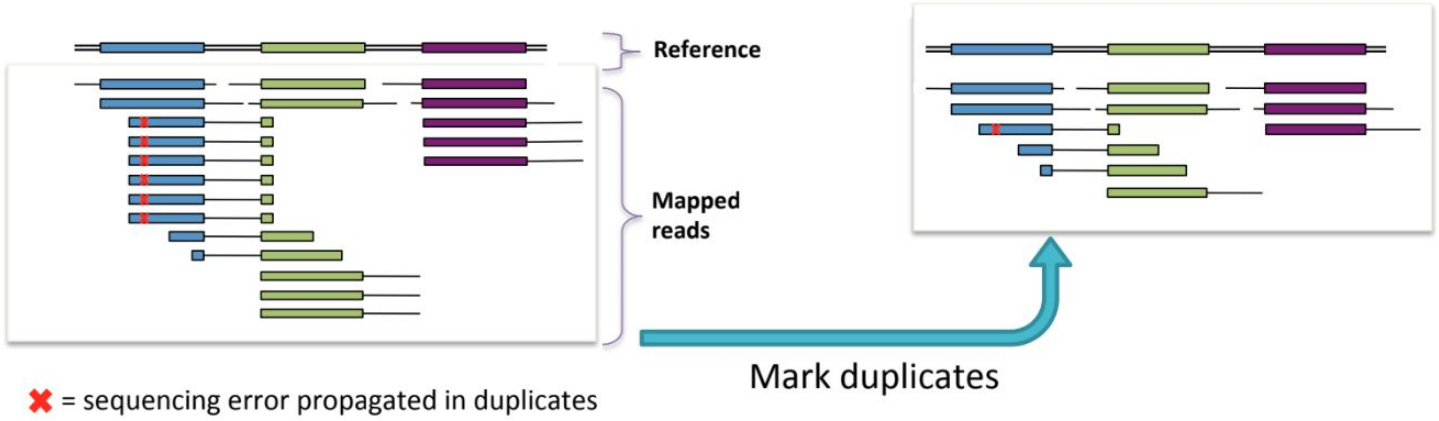
http://www.slideshare.net/jandot/next-generation-sequencing-course-part-2-sequence-mapping
http://www.slideshare.net/cosentia/illumina-gaiix-for-high-throughput-sequencing

source: GATK workshops, https://drive.google.com/drive/folders/0BwTg3aXzGxEDR2JKTUIxVGxCT1k

Cleaning duplicated reads

Why mark duplicates?

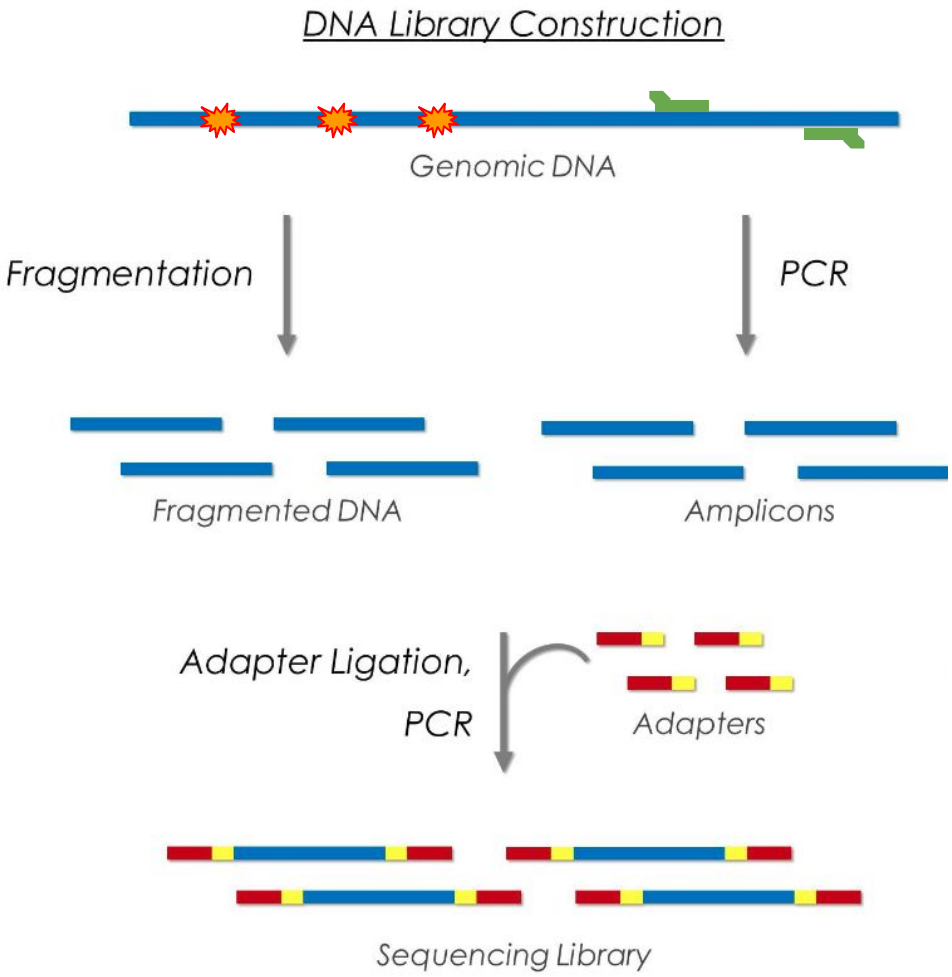
- Duplicates are sets of reads pairs that have the same unclipped alignment start and unclipped alignment end
- They're suspected to be **non-independent measurements** of a sequence
  - Sampled from the exact same template of DNA
  - Violates assumptions of variant calling
- What's more, errors in sample/library prep will get propagated to *all* the duplicates
  - Just pick the "best" copy – mitigates the effects of errors



source: GATK workshops, https://drive.google.com/drive/folders/0BwTg3aXzGxEDR2JKTUIxVGxCT1k

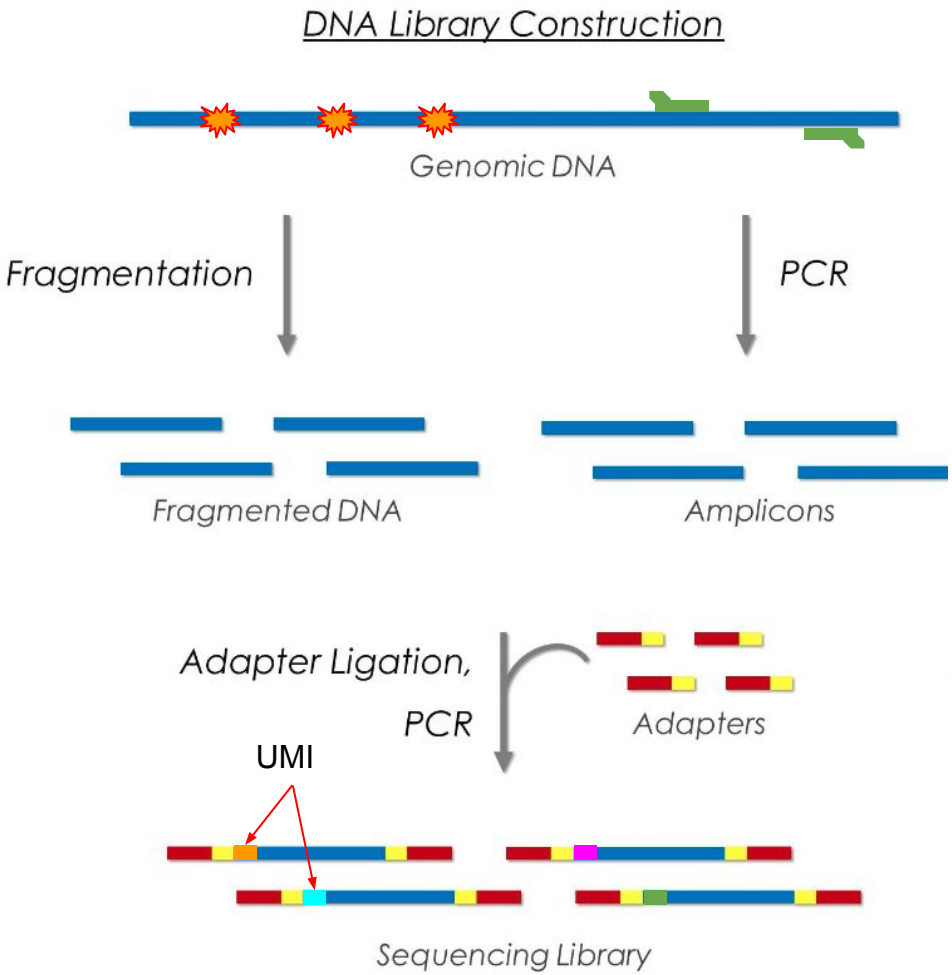
Cleaning duplicated reads

Molecular Barcoding (UMI, *unique molecular identifiers*)



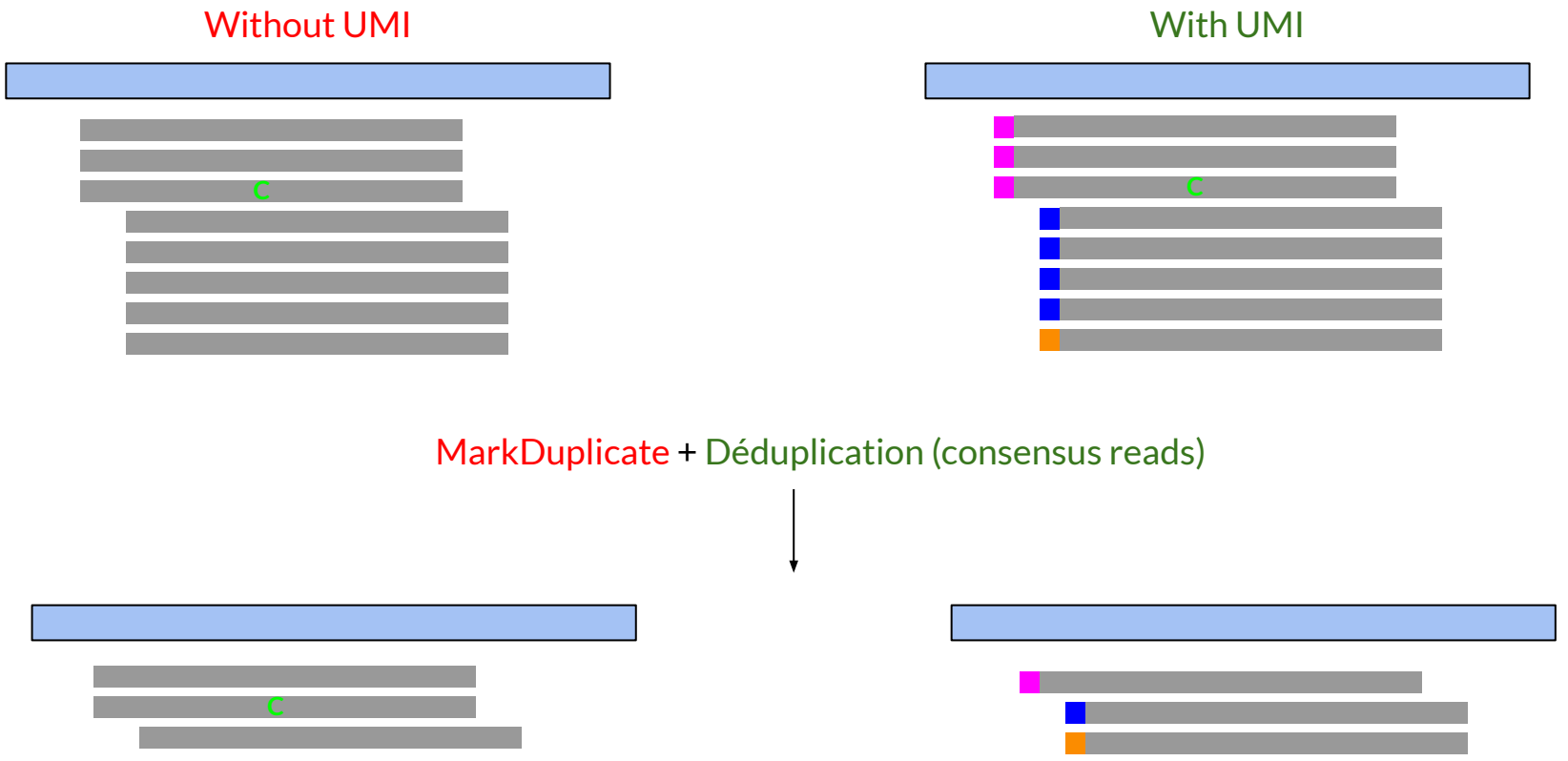
Cleaning duplicated reads

Molecular Barcoding (UMI, *unique molecular identifiers*)

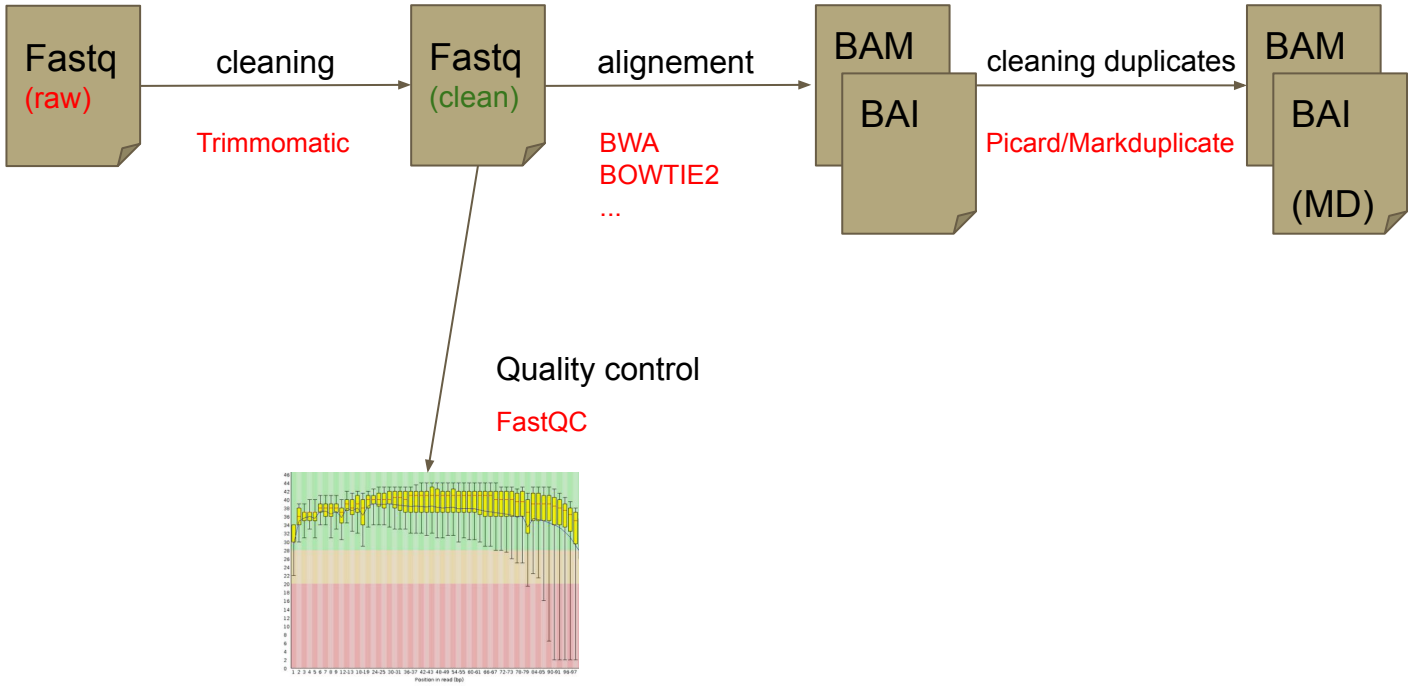


Cleaning duplicated reads

Molecular Barcoding (UMI, *unique molecular identifiers*)

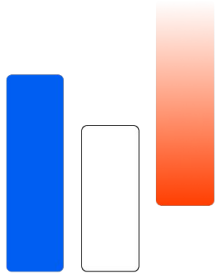


Workflow


Fastq
(raw)
cleaning
Trimmomatic
Fastq
(clean)
alignement
BWA
BOWTIE2
...
BAM
BAI
cleaning duplicates
Picard/Markduplicate
BAM
BAI
(MD)
Quality control
FastQC

Practical: Mapping

Open Galaxy



usegalaxy.fr

Practical:
https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/mapping/tutorial.html

TIAAS: https://usegalaxy.fr/join-training/bilille-rnaseq-2023/