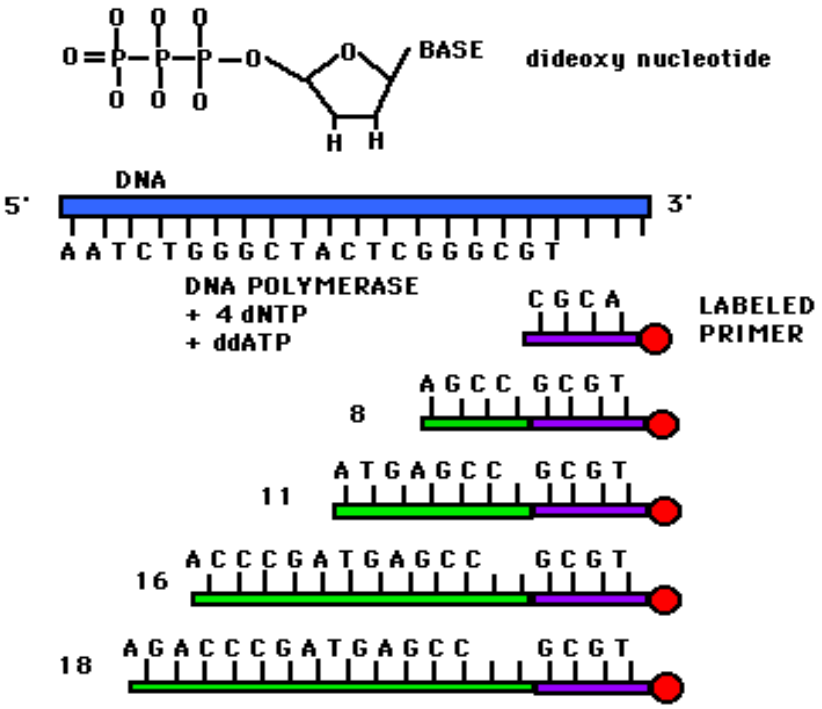
Maastricht University Leading in Learning!

RNA sequencing technologies and data analysis workflow

Florian Caiment (TGX)

December 2023

Sanger sequencing (1977)



DNA Polymerase
+ 4 dNTPs

ddATP ddTTP ddCTP ddGTP

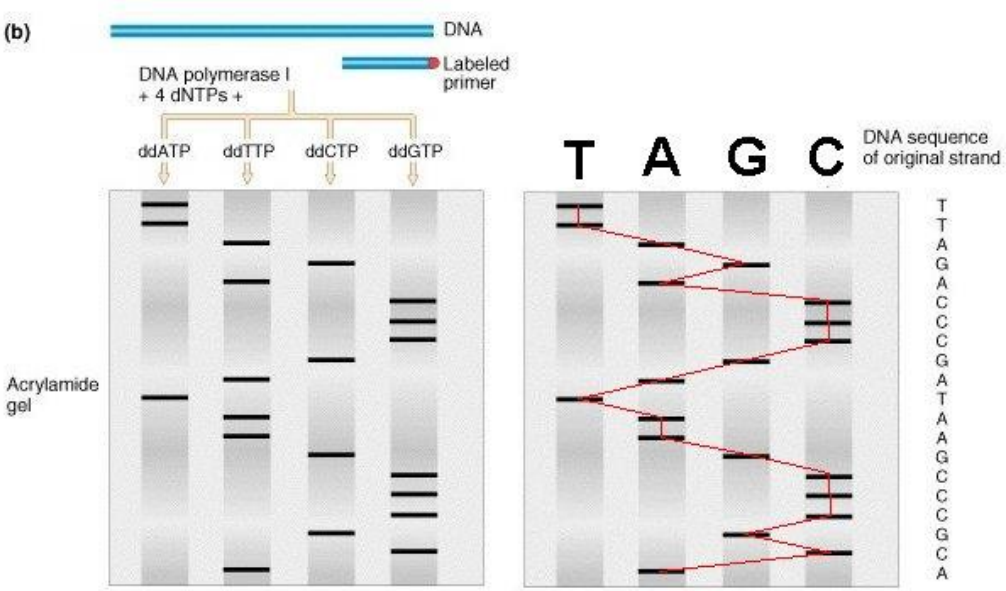
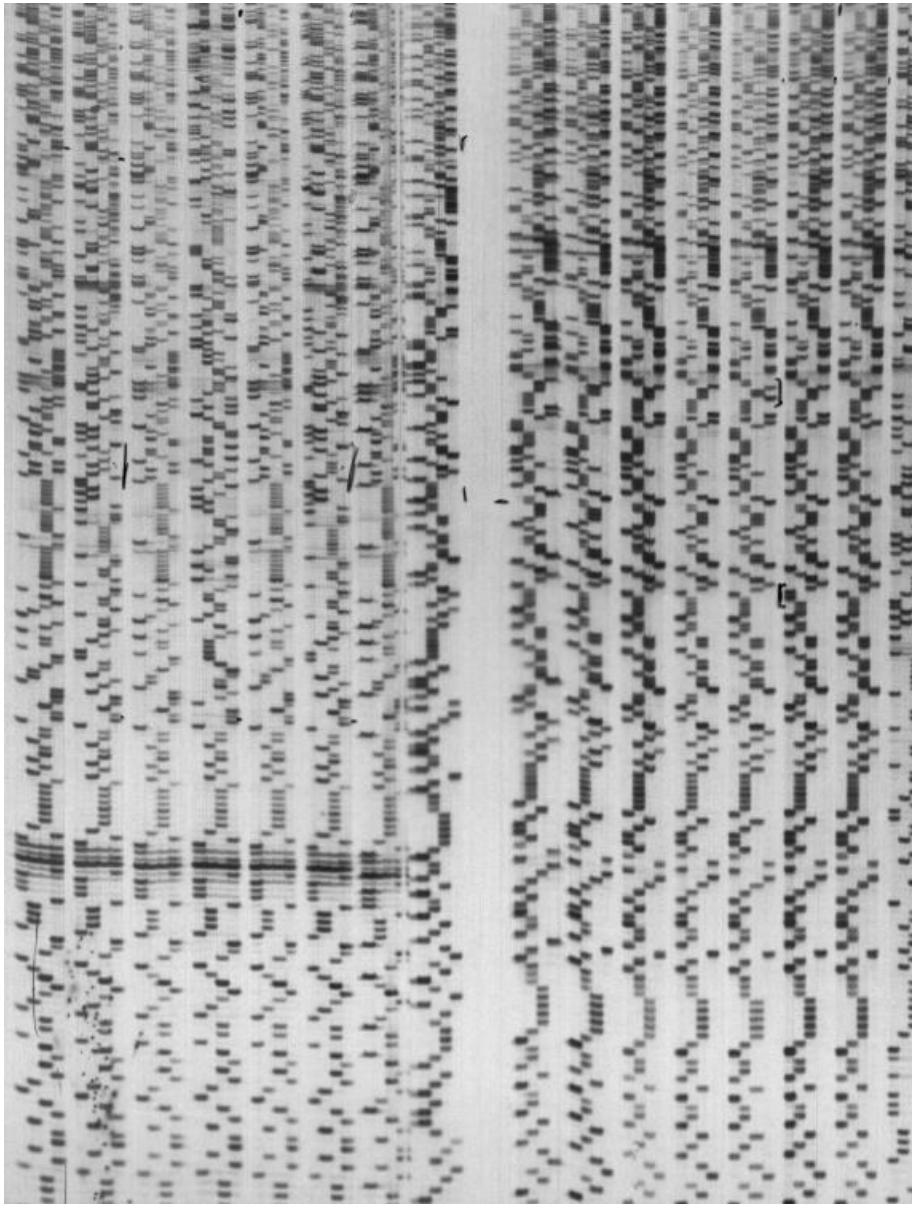


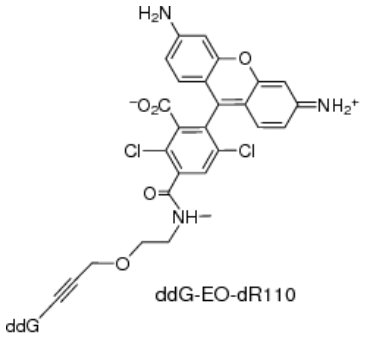
Figure ©2002 by Griffiths *et al.*

Sanger sequencing (1977)

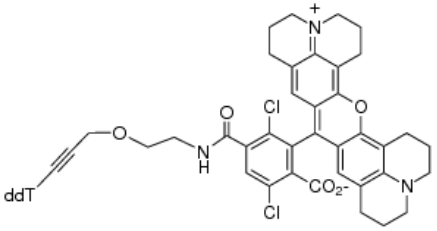


Applied Biosystem (1987)

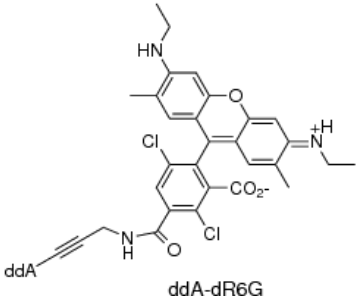
Bigdye terminator: Mix of 4 fluorescent dd-NTP



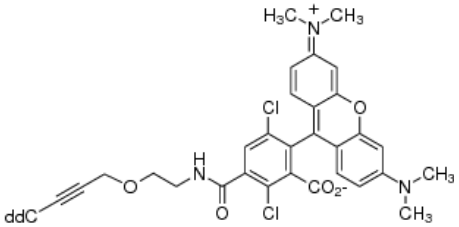
ddG-EO-dR110



ddT-EO-dROX

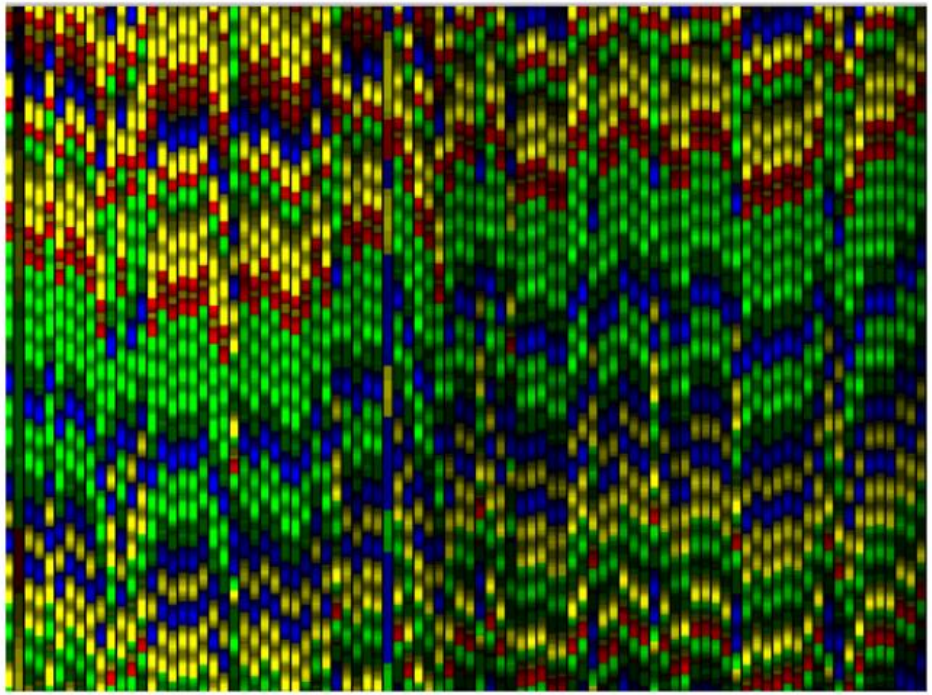


ddA-dR6G

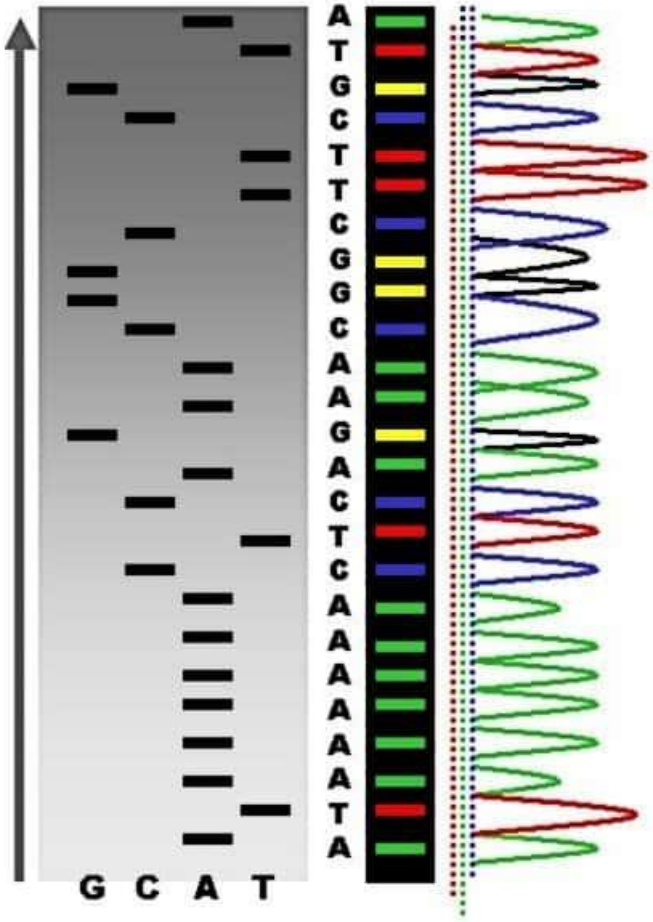


ddC-EO-dTAMRA

Applied Biosystem (1987)



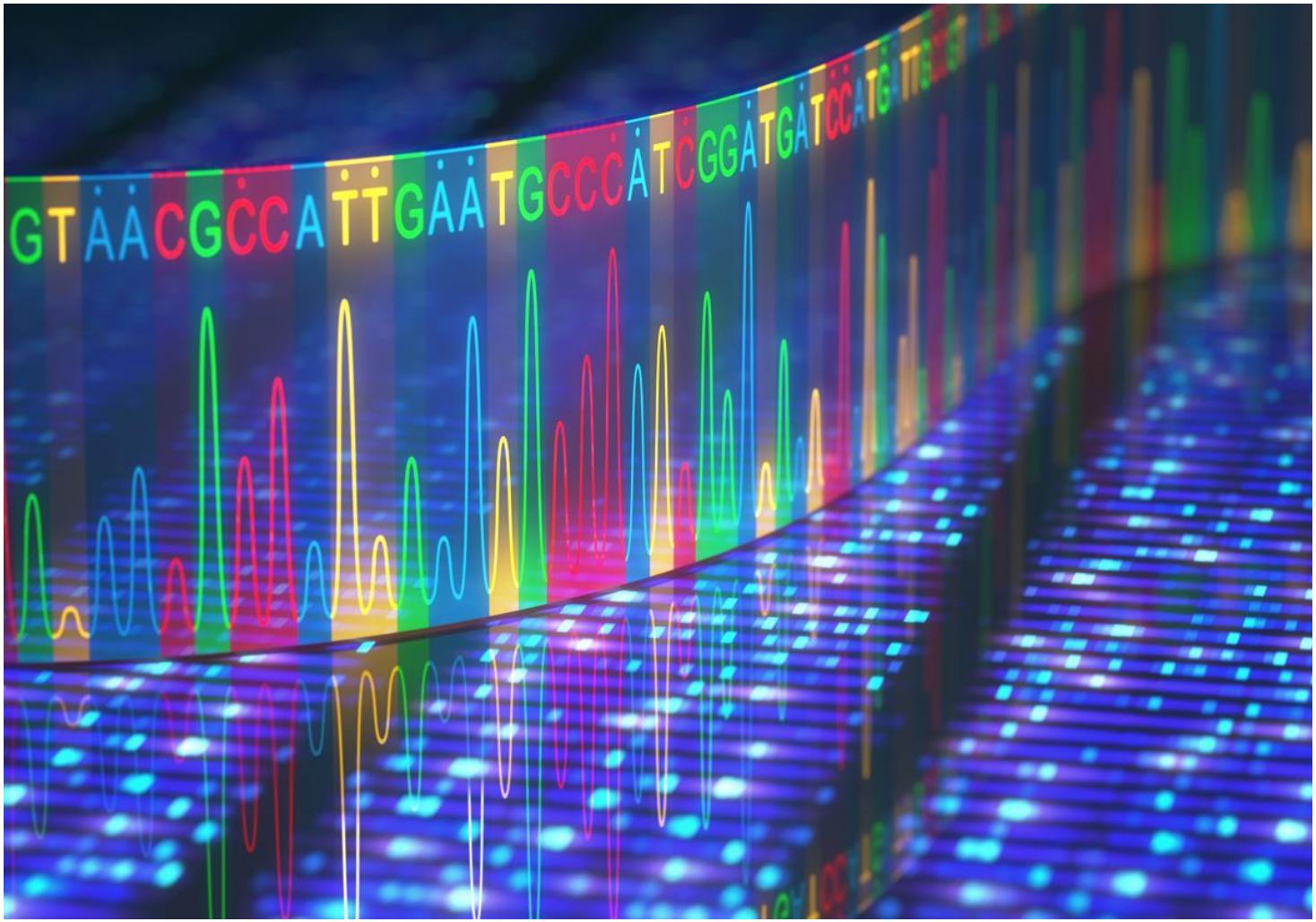
Sequence 800-1000 bp long DNA molecule



Ruled sequencing for 2 decades

1999: Human Chromosome 22 sequenced

2005: The beginning of NGS



Campbell, D., (2020) *Next Generation Sequencing 101*

HTS / NGS / Deep Seq Evolution

454/Roche

Solexa/Illumina

Solid/AB

IonTorrent

Nanopore

Pacific Biosciences

And more…

NGS sequencing techniques

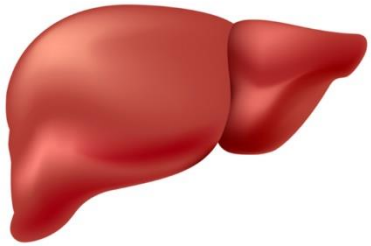
Current Available NGS Platforms

Category	Company	Platform	Read length (bp)	# of reads/run	Sequencing output/run	Run time
PCR-based	Illumina	HiSeq 2500	100–200	1×10^9	~120 Gb	27 h
		HiSeq 2000	100–200	6×10^9	~600 Gb	11 d
		MiSeq	100–150	7×10^6	~2 Gb	1 d
	Thermo Fisher	Ion Torrent	100–200	1×10^7	~2 Gb	5.5 h
		Ion Proton	100-200	7×10^7	~10 Gb	4 h
Single Molecular	Pacific Biosciences	PACBIO RS	~ 18,000	50×10^3	~600 Mb	4 h
	Oxford Nanopore	GridION	$>5 \times 10^6$	2000	~10 Gb	15 min

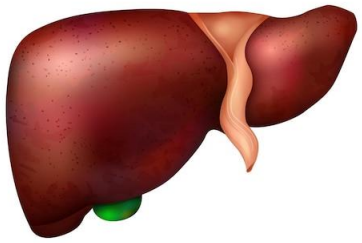
Instrument	Primary Errors	Error Rate (%)	Cost/MB	Market share
3730xl (capillary)	Substitution	0.1-1	$2,308	-
Illumina All Models	Substitution	~0.1	$0.05	63%
Ion Torrent – all chips	Indel	~1	$0.01	12%
PacBio RS	CG deletions	≤15	$7-38	~2%
Oxford Nanopore	Deletions	<20	$0.02	0

For a recent update: https://pubmed.ncbi.nlm.nih.gov/33745759/

An example: From the Wet Lab To the Dry Lab



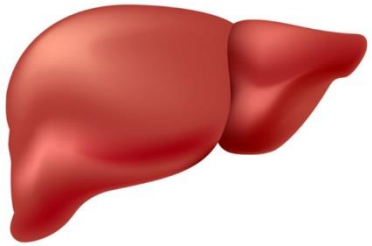
Healthy Liver



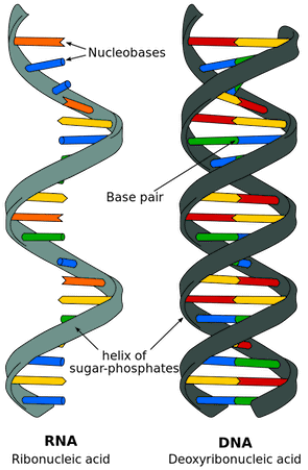
Diseased Liver

Let’s compare the transcriptome of the two samples using NGS

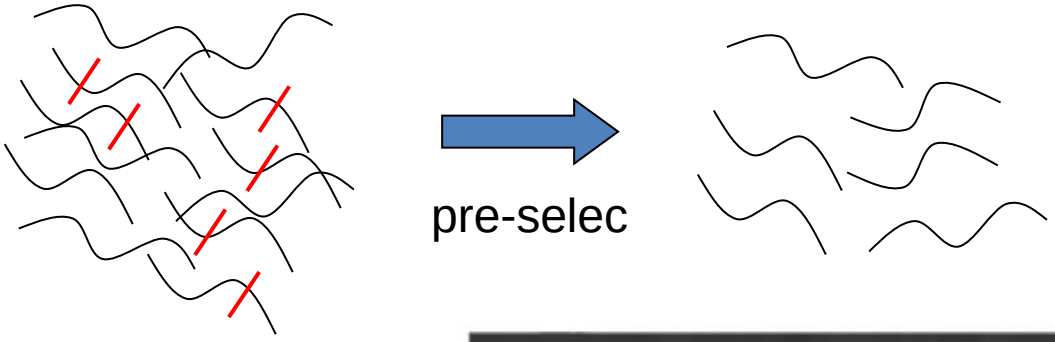
RNA Sequencing

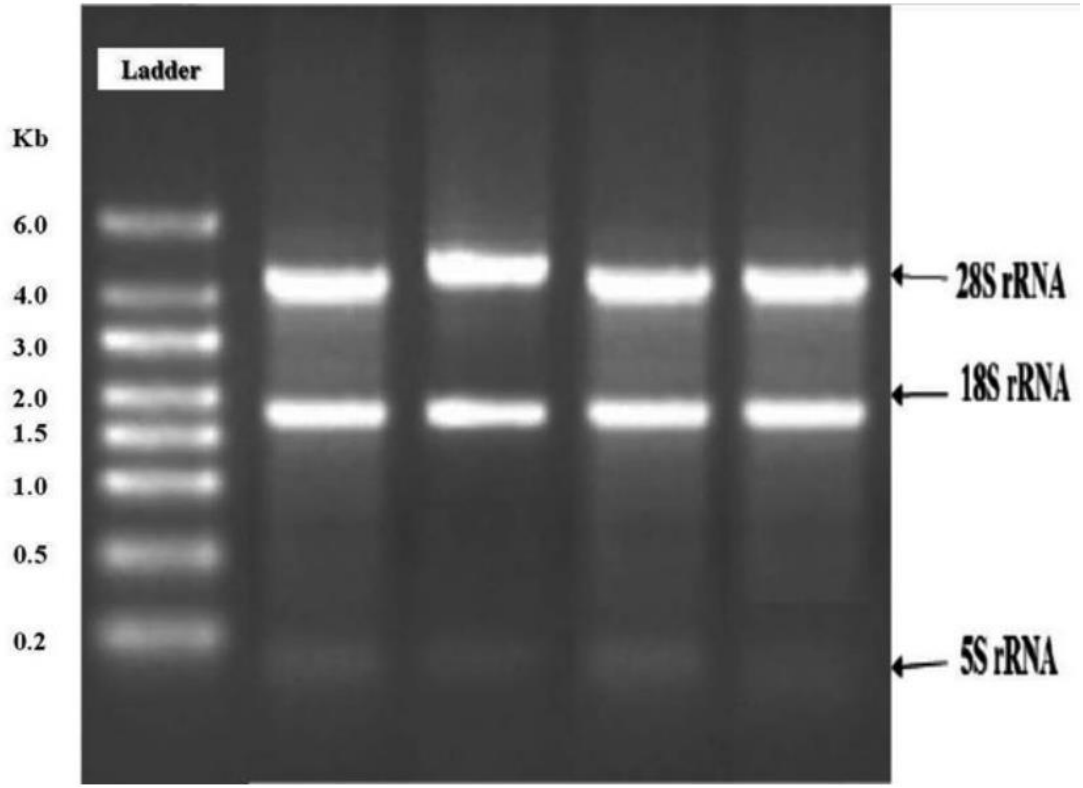


RNA Extraction

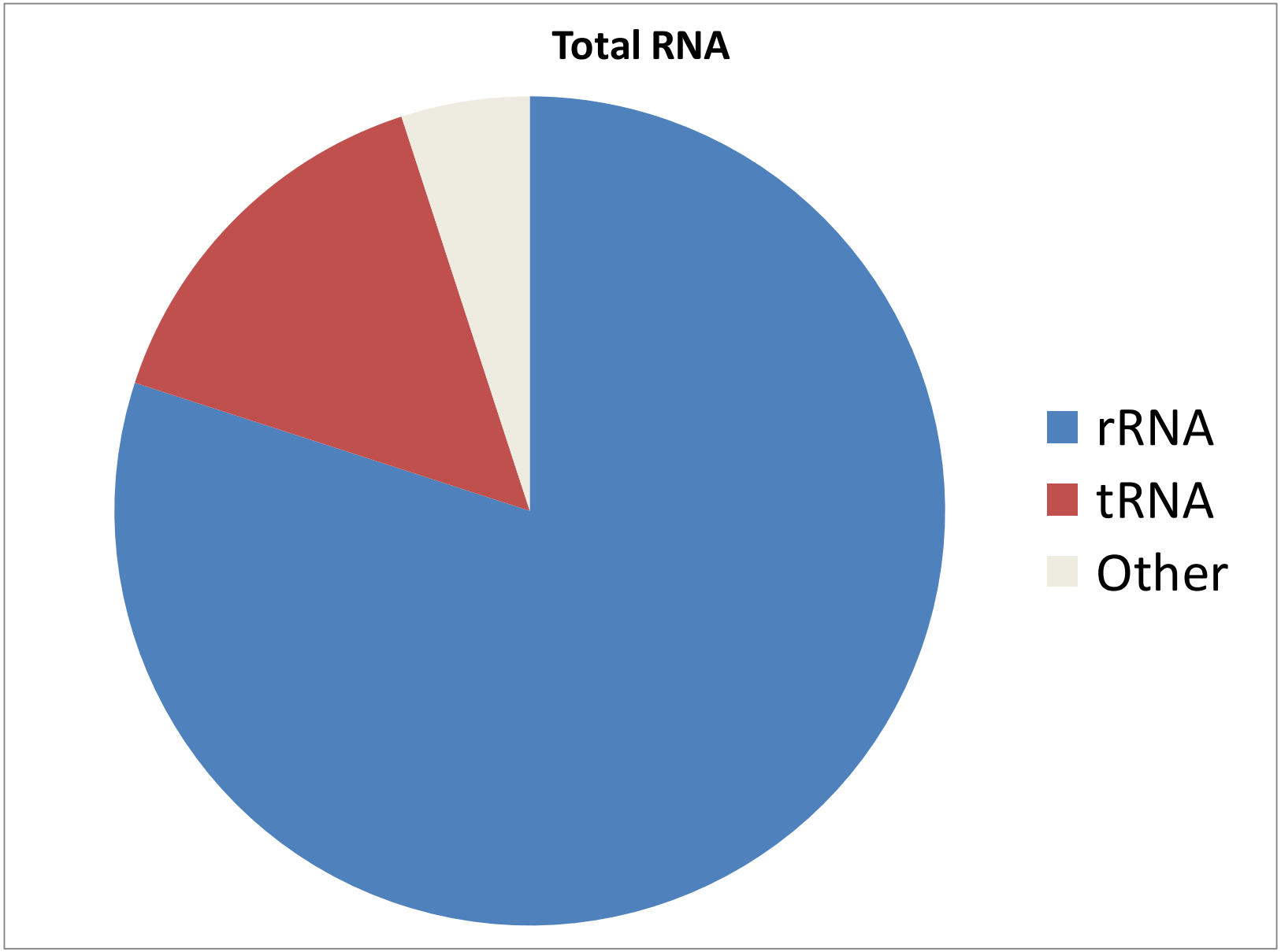


NGS Protocol

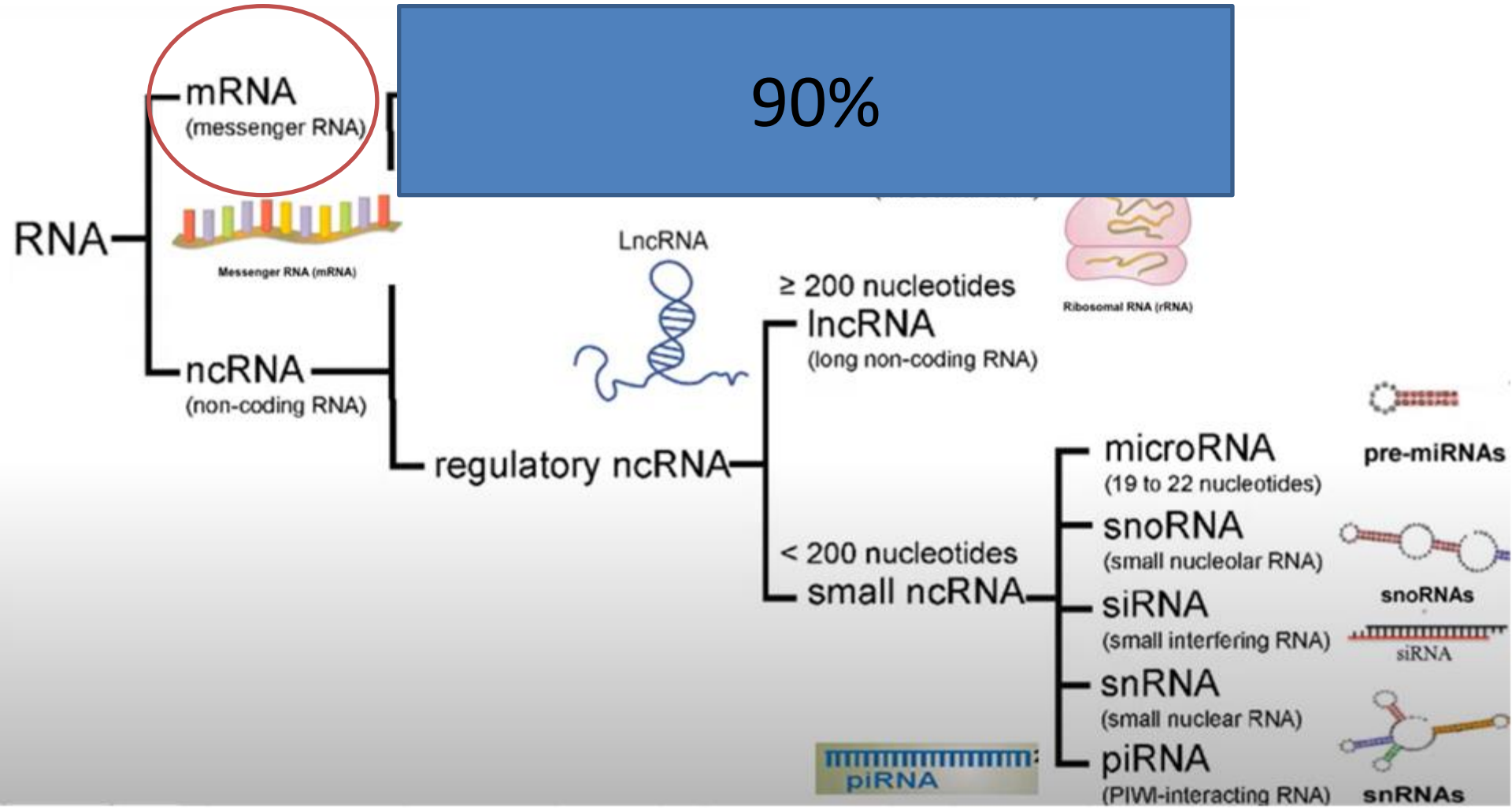




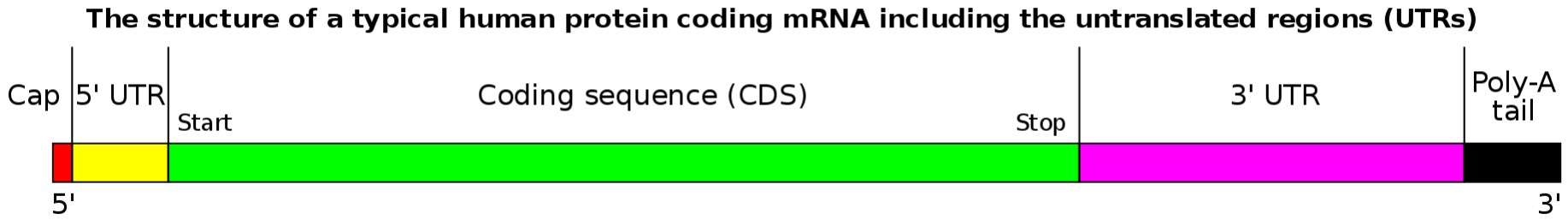
What do you find in total RNA ?

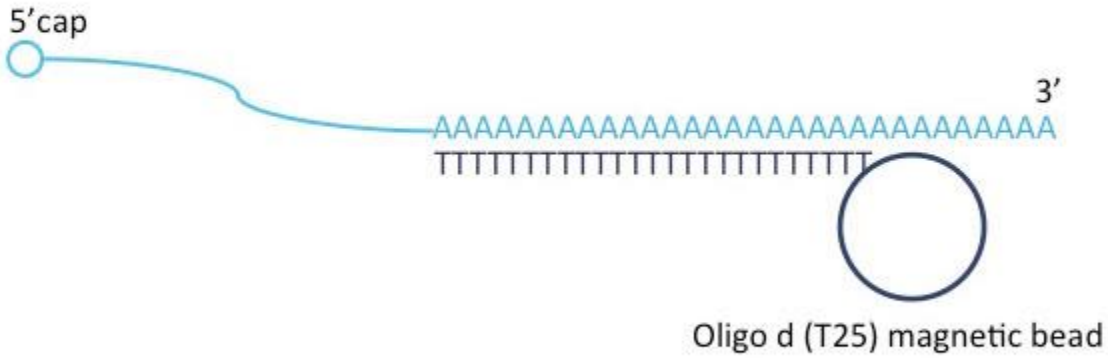

Total RNA
rRNA
tRNA
Other

“Other RNA”


90%
RNA
mRNA
(messenger RNA)
Messenger RNA (mRNA)
ncRNA
(non-coding RNA)
regulatory ncRNA
LncRNA
≥ 200 nucleotides
lncRNA
(long non-coding RNA)
Ribosomal RNA (rRNA)
< 200 nucleotides
small ncRNA
microRNA
(19 to 22 nucleotides)
pre-miRNAs
snoRNA
(small nucleolar RNA)
snoRNAs
siRNA
(small interfering RNA)
siRNA
snRNA
(small nuclear RNA)
piRNA
(PIWI-interacting RNA)
piRNA
snRNAs

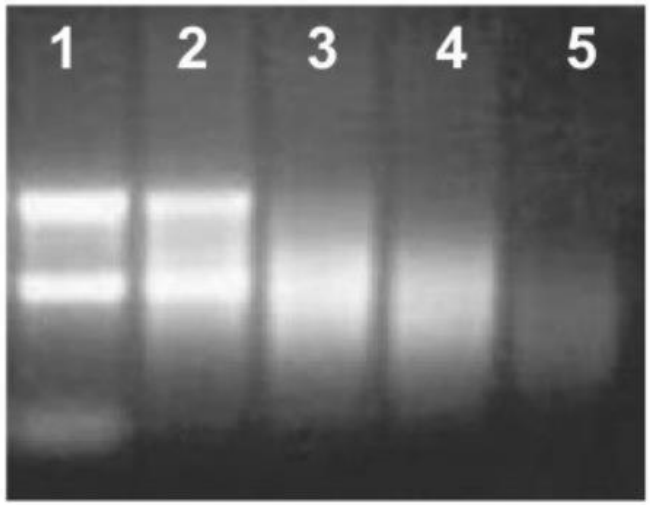
polyA library

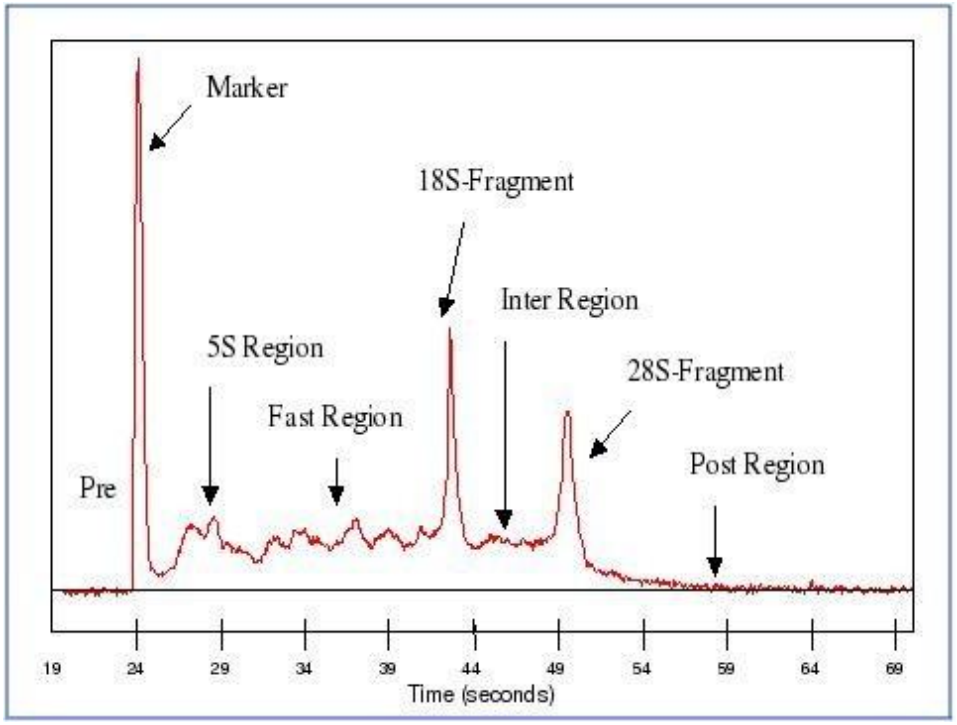




Require intact (good quality) RNA !

RIN (RNA Integrity Number)

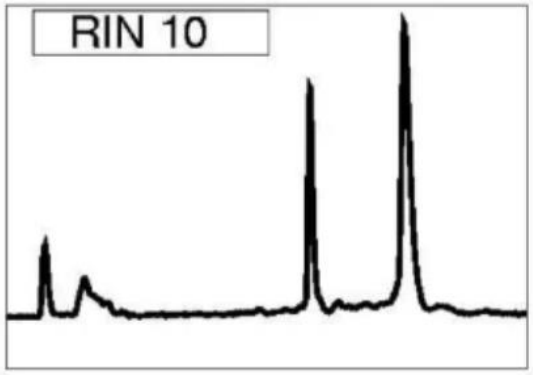


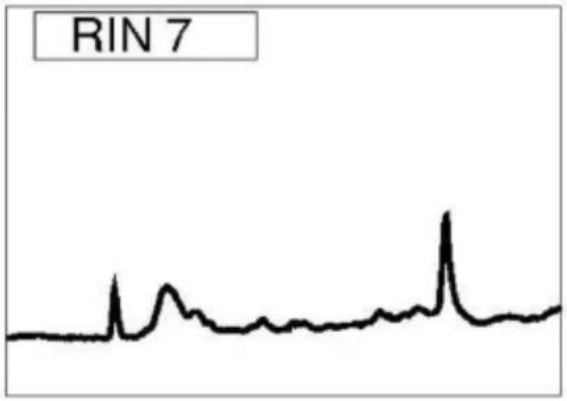


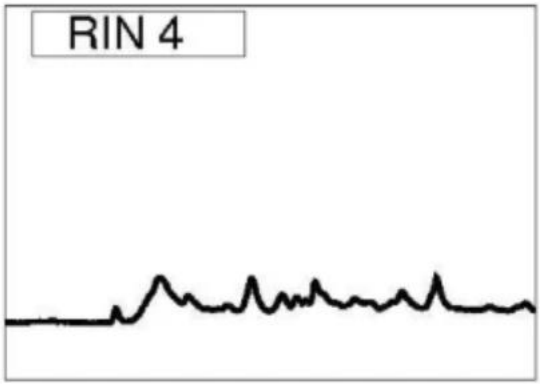
Bioanalyzer Track

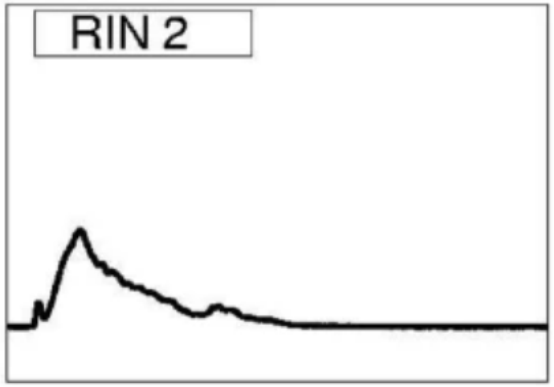
Schroeder A, Mueller O, Stocker S, et al. *BMC Mol Biol*. 2006

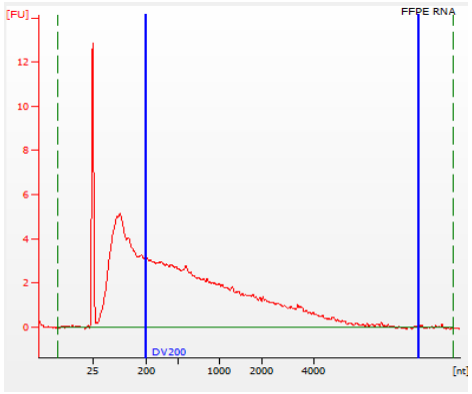
RIN (RNA Integrity Number)







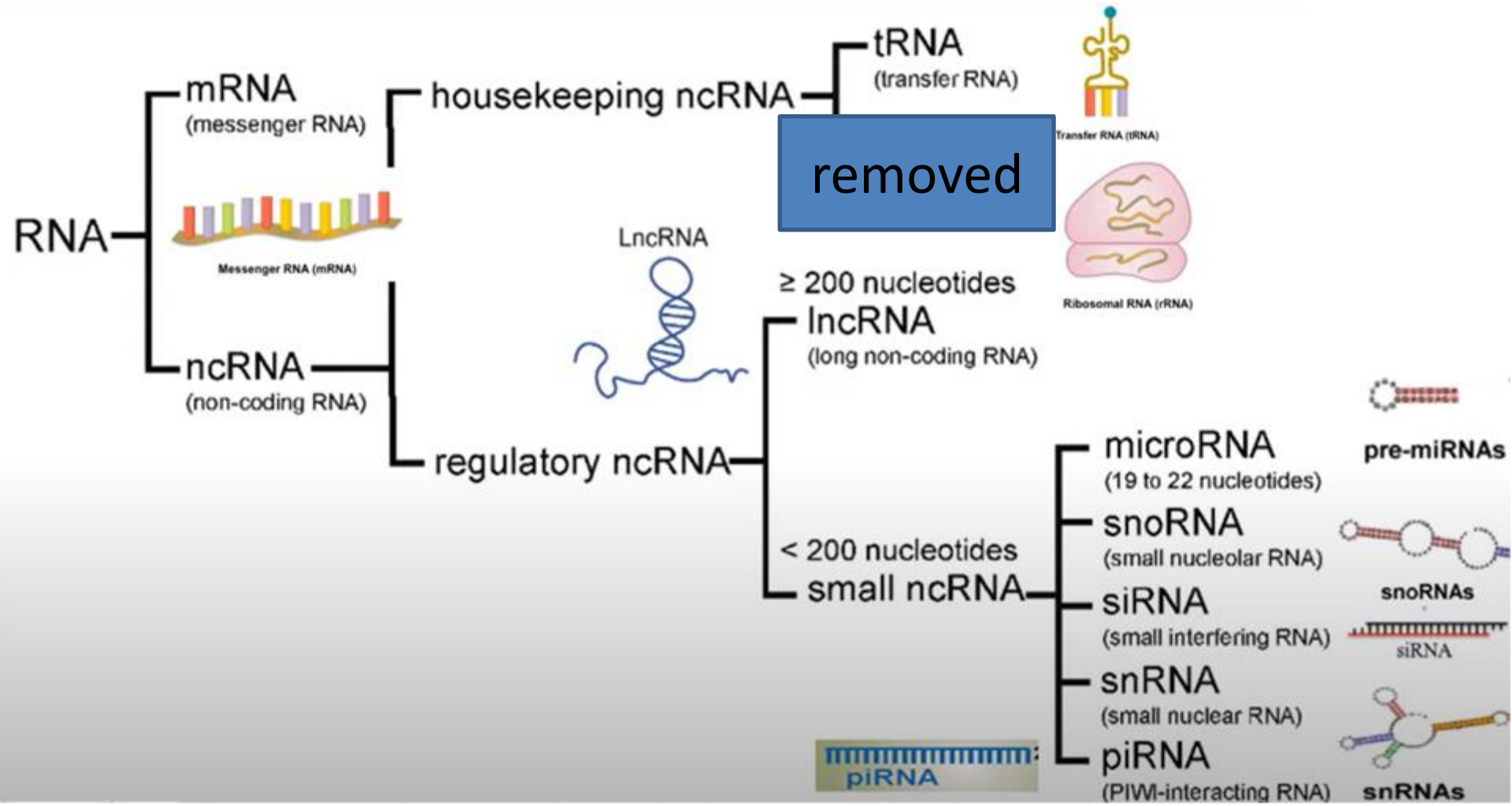




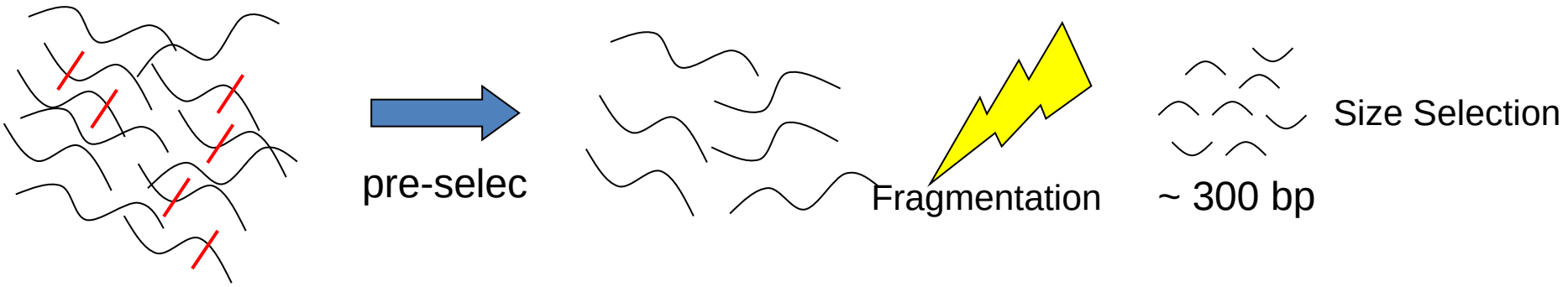
DV200

Schroeder A, Mueller O, Stocker S, et al. *BMC Mol Biol*. 2006

Ribo-Depleted Library



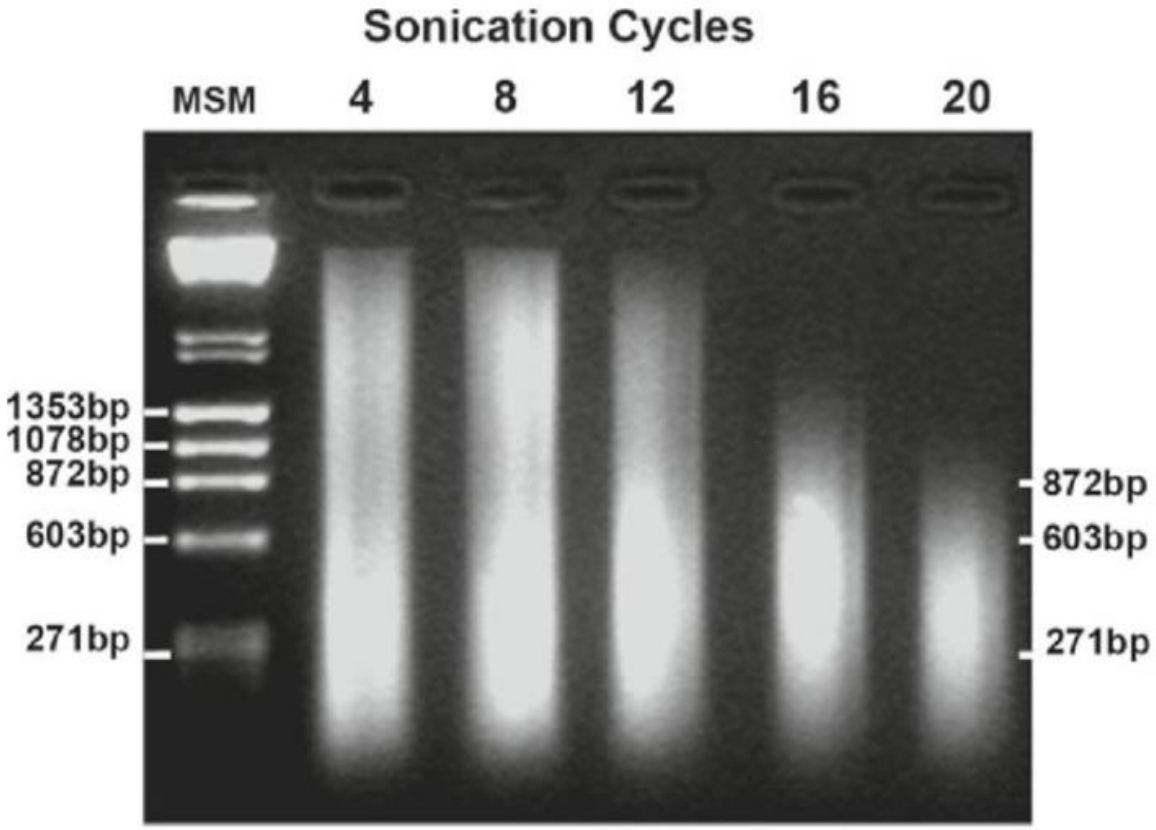
NGS sequencing



Fragmentation methods

Mechanical methods : Sonication

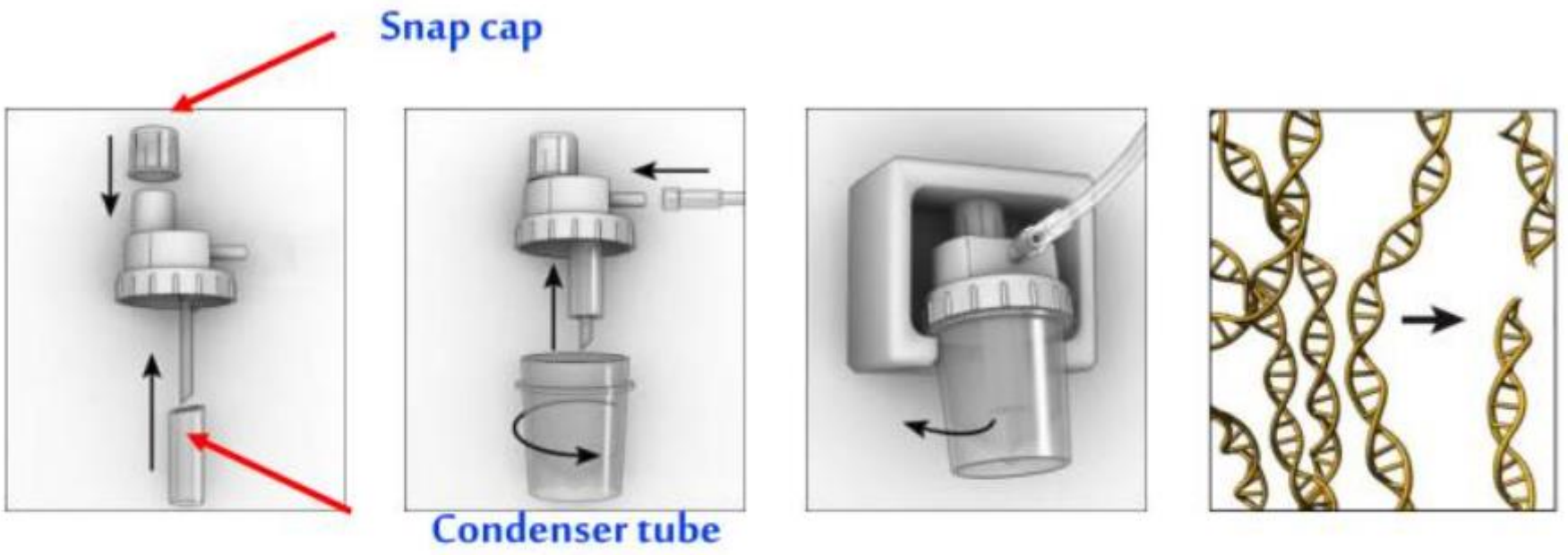
Size of fragment influenced by wavelength, duration and number of cycle



Heliot and Cereghini, 2012

Fragmentation methods

Mechanical methods : Nebulization / Hydrodynamic shear



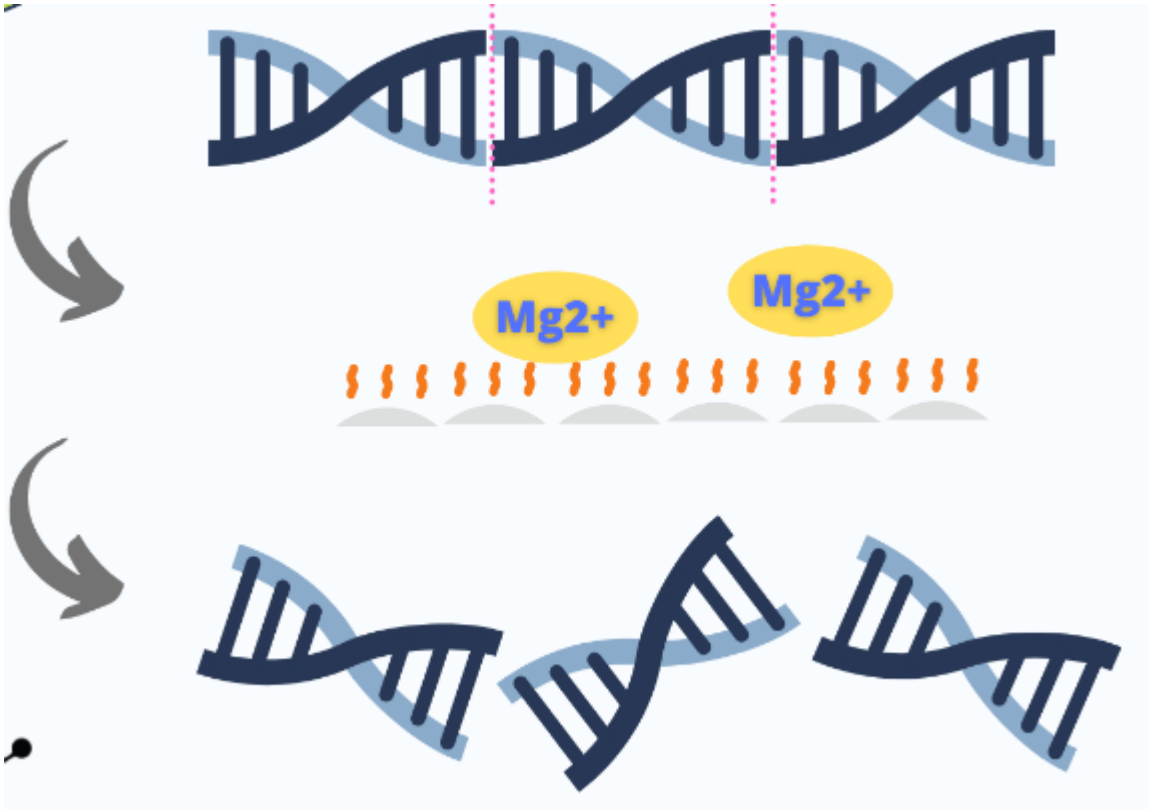
- Nebulization shears double-stranded DNA into fragments ranging from 50 to 900 base pairs.
- High-pressure **nitrogen gas** is used to force the sample into small droplets of liquid which shears the DNA.

Fragmentation methods

Enzymatic methods : DNAse I, Restriction enzyme…

Chemical methods

Heat plus divalent metal cations (Mg2+, Zn2+, etc…)

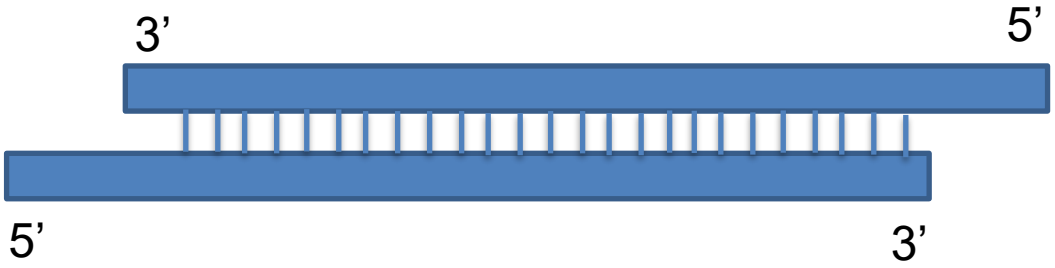


Fragmentation methods

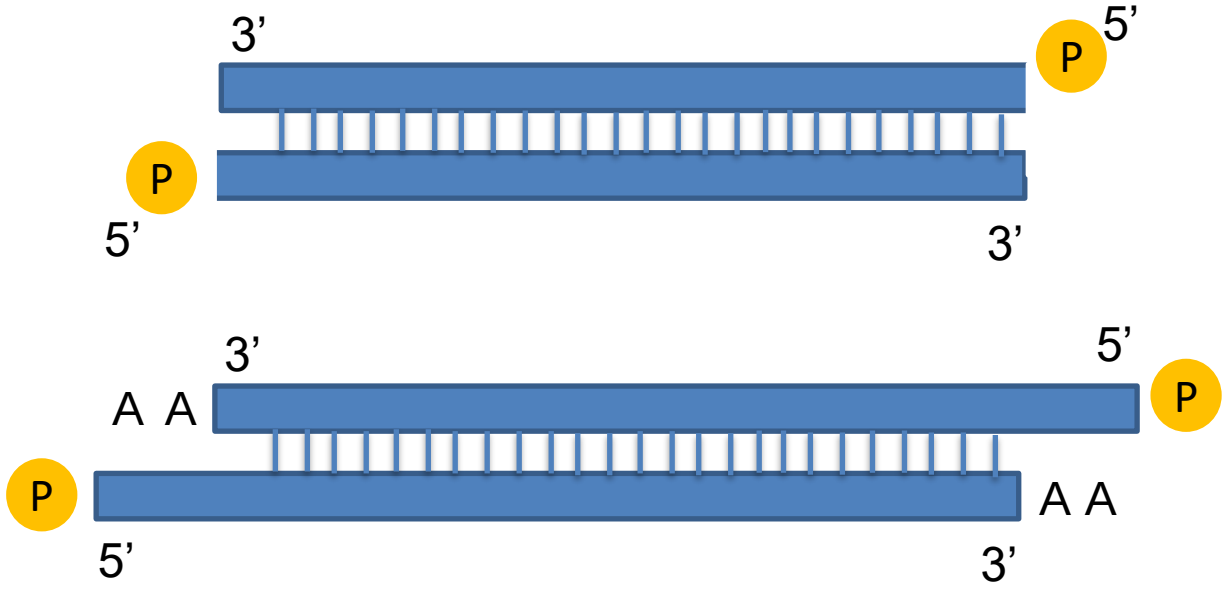
	Physical	Chemical	Enzymatic
Advantages	Tolerant broad range input material Unbiased fragmentation method Less sample-to-sample size variation Creates evenly sized fragments Considered a "clean" method because it does not interfere chemically with the sample Ease of use and implementation	Well-known for breaking RNA	Uses standard lab equipment Highly scalable Requires lower input amounts for PCR-free library generation
Disadvantages	Requires dedicated, expensive equipment Requires large amounts of sample and can result in nucleic acid loss	Requires cations, which can interfere with the sequencing process.	Can cause fragmentation bias Can be sensitive to nucleic acid input amounts because of the relative abundances of enzymes and substrates Can cause sample-to-sample variation

https://www.goldbio.com/articles/article/how-to-fragment-DNA-for-NGS

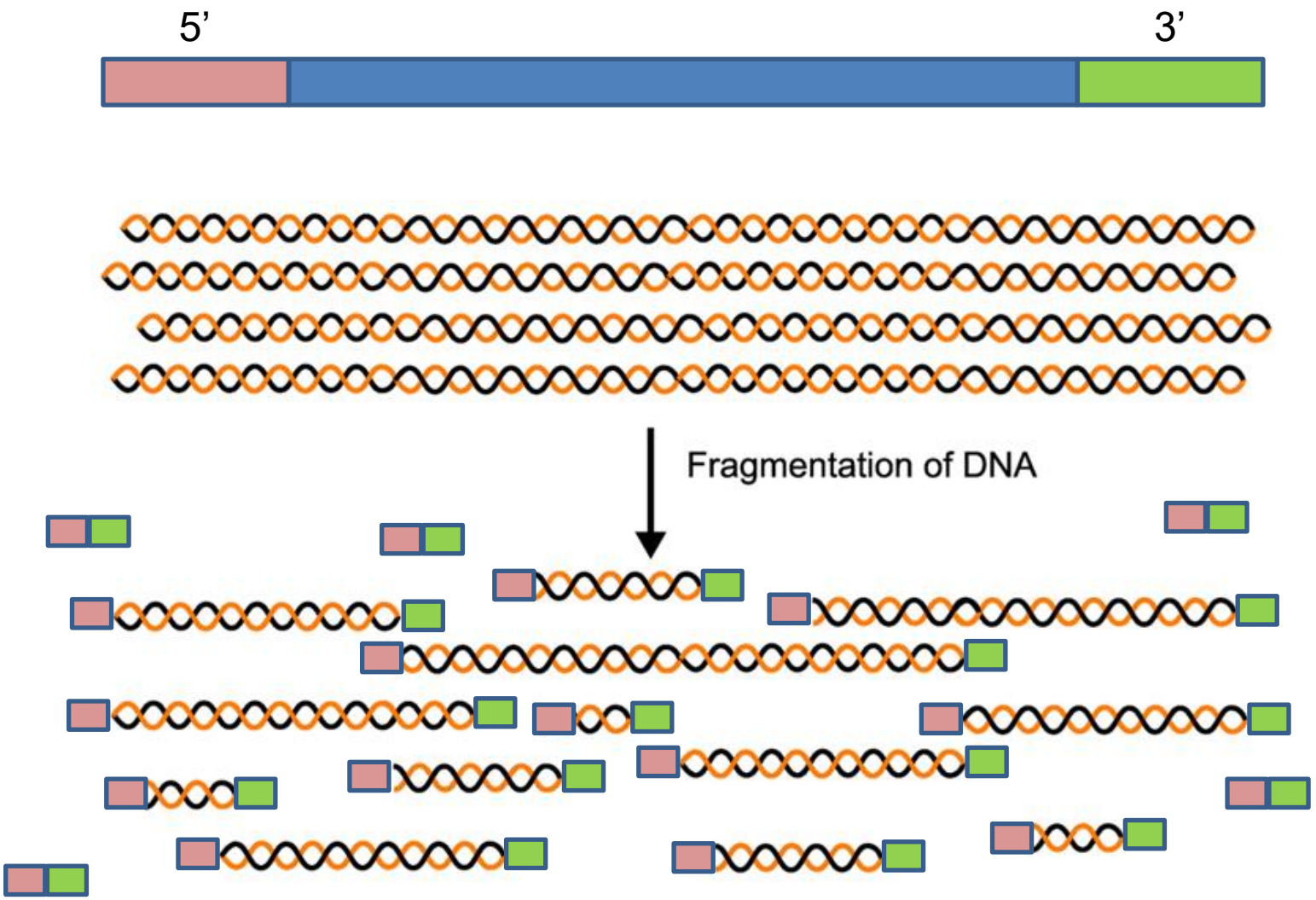
End Repair Reaction



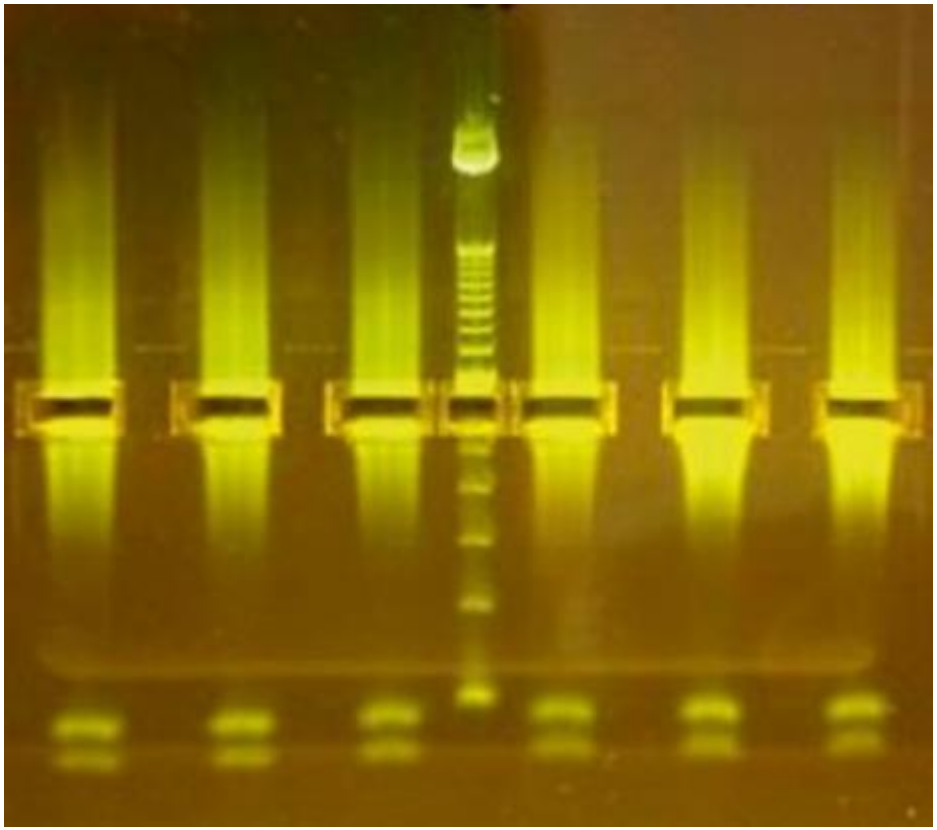
Overhanged ends are either removed (exonuclease) or extended (polymerase)



Adaptor Ligation



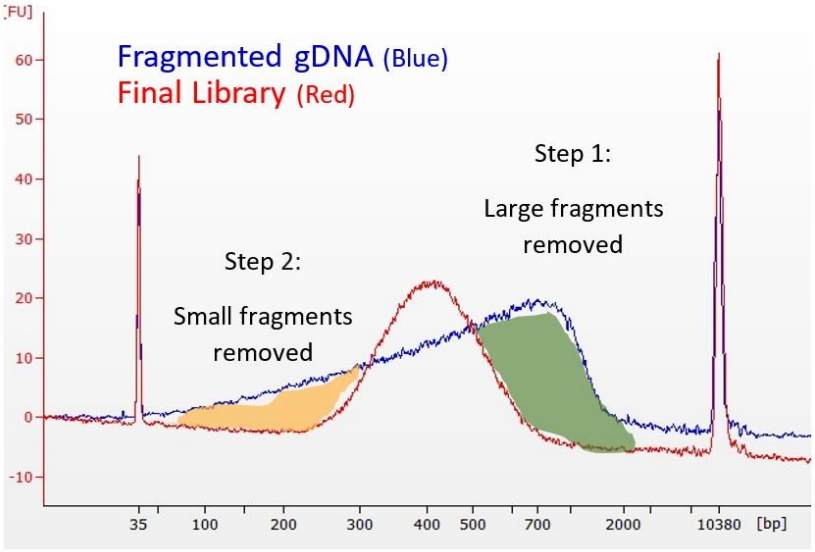
Option : Size Selection



Gel Selection

Efficient, Intuitive,
VERY time consuming,

Option : Size Selection

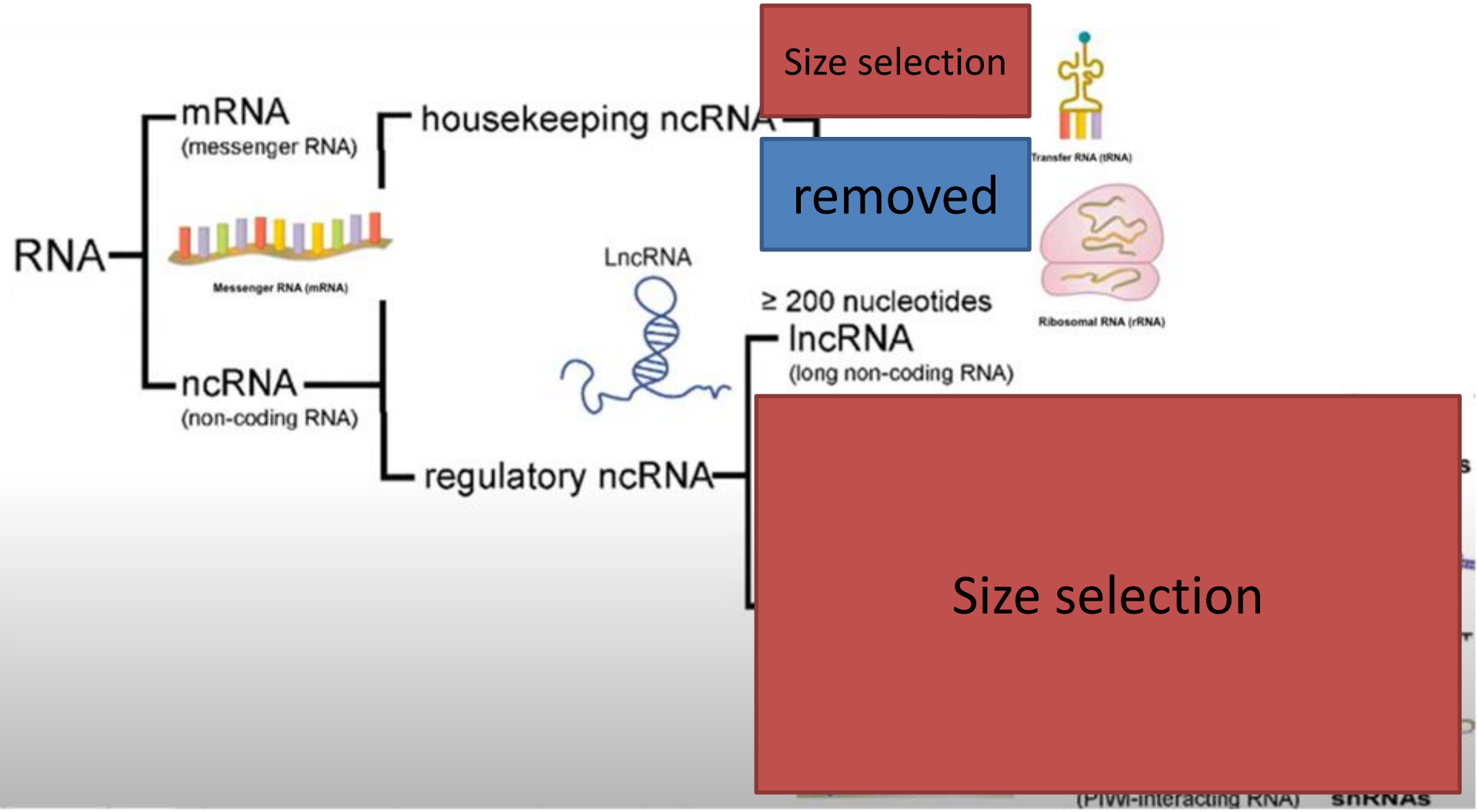


Double sided clean up

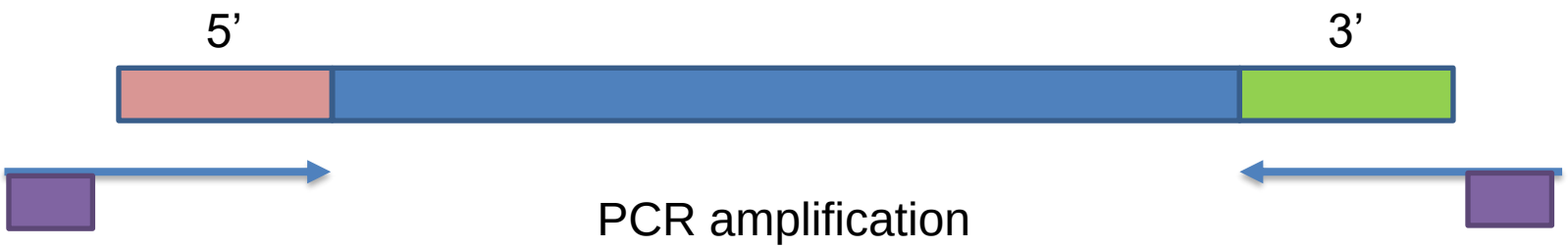
Approximate Insert Peak Size (bp)	150 - 250	250 - 350	300 - 500	400 - 600	500 - 700
Approximate Library Peak Size (bp)	270 - 370	370 - 470	420 - 620	520 - 720	620 - 820
Bead Volume #1	35	32	30	27	24
Bead Volume #2	12	9	8	8	8

https://www.youtube.com/watch?v=WBJTilok2xM&ab_channel=Eppendorf

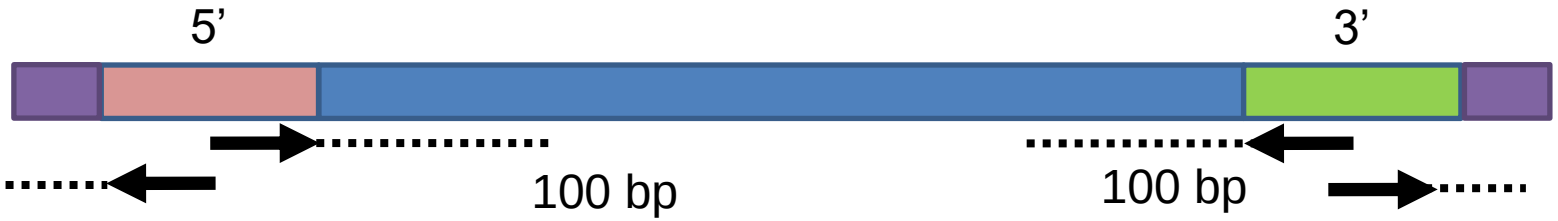
“Other RNA”


Size selection
removed
RNA
mRNA
(messenger RNA)
housekeeping ncRNA
Transfer RNA (tRNA)
Messenger RNA (mRNA)
Ribosomal RNA (rRNA)
LncRNA
≥ 200 nucleotides
lncRNA
(long non-coding RNA)
ncRNA
(non-coding RNA)
regulatory ncRNA
Size selection
(PIWI-interacting RNA)
snRNAs

PCR and Sequencing



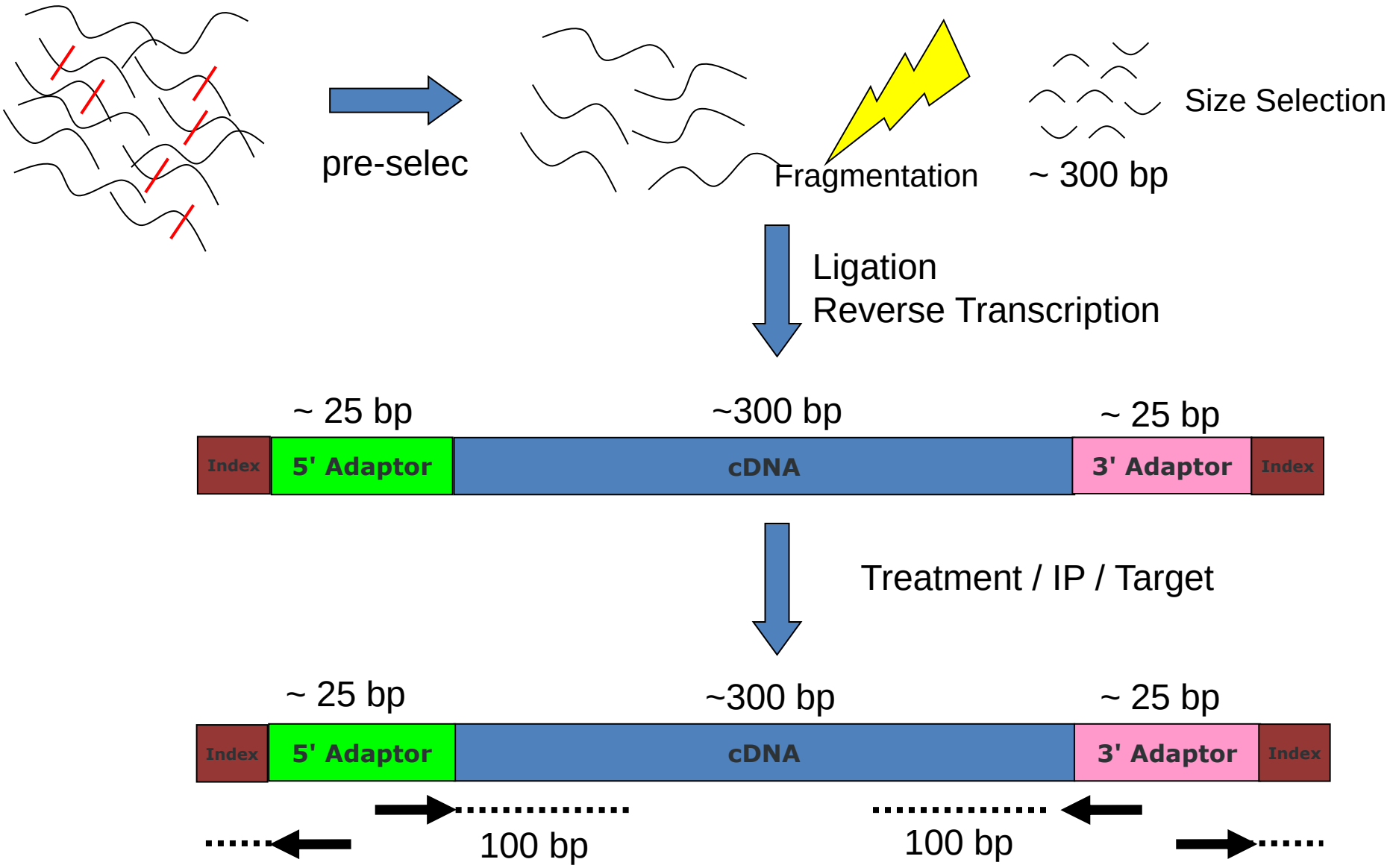
Number of PCR cycle depending on the amount of starting material
PCR-Free library preparation methods exist



Single Read (SR)
Paired-End (PE)

Barcodes (also called Indexes) are now added to multiplex the samples in the same sequencing flowcell

NGS sequencing



MANY different library preparation kits…

… which can impact the data analysis pipeline!

Illumina

NextSeq Series

HiSeq 4000 System

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

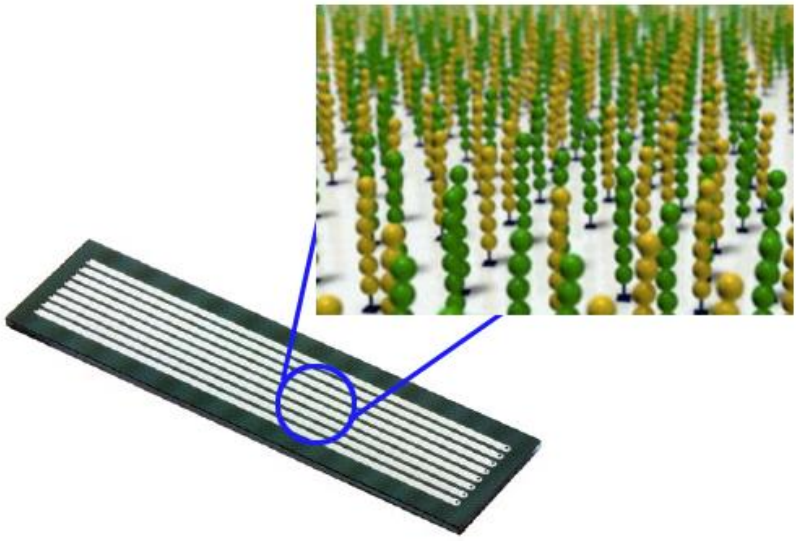
See how a Laboratory Information Management System (LIMS) enabled this large genomics lab to standardize lab procedures and cope with increasing sample volumes from diverse clients.

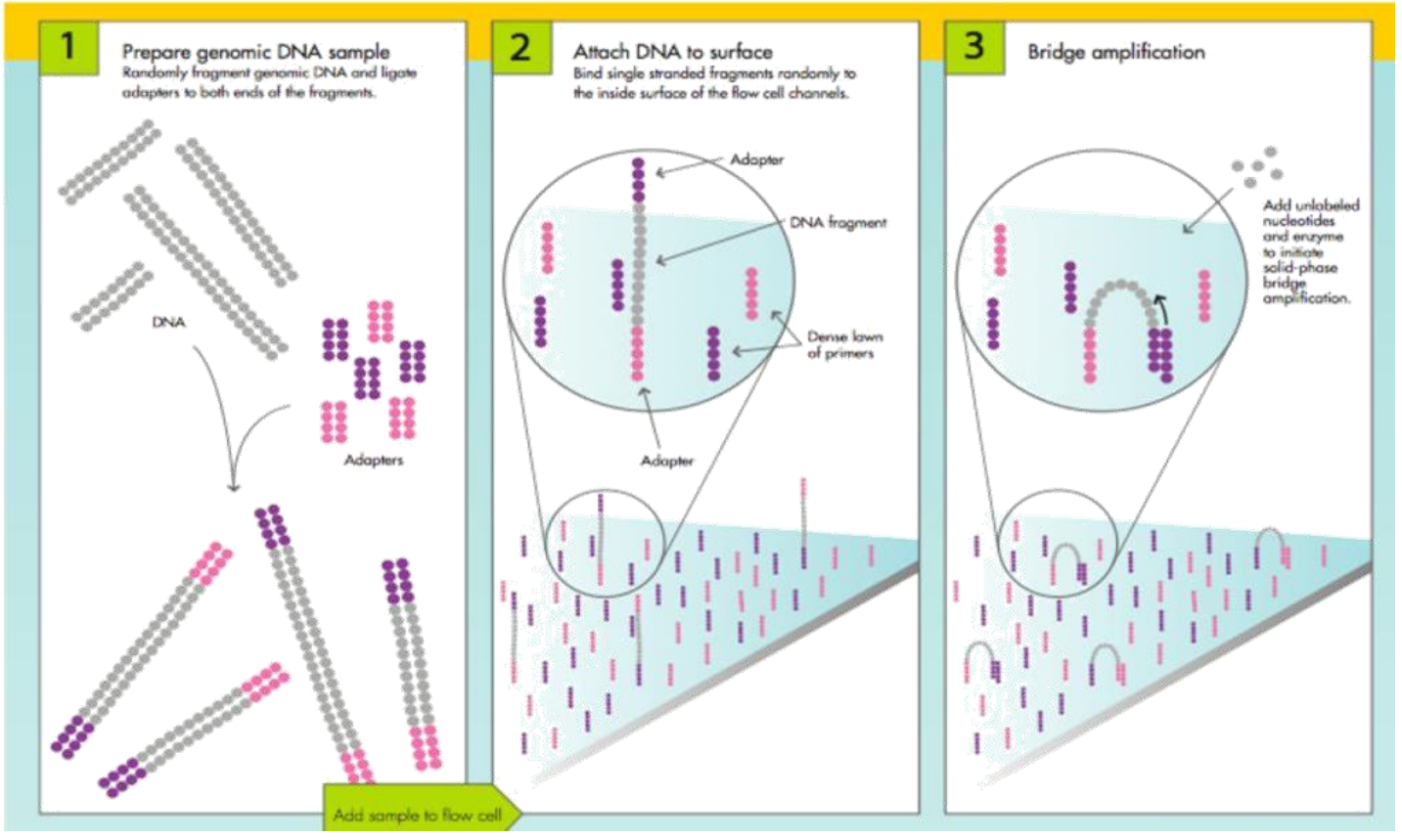
Read Case Study ›

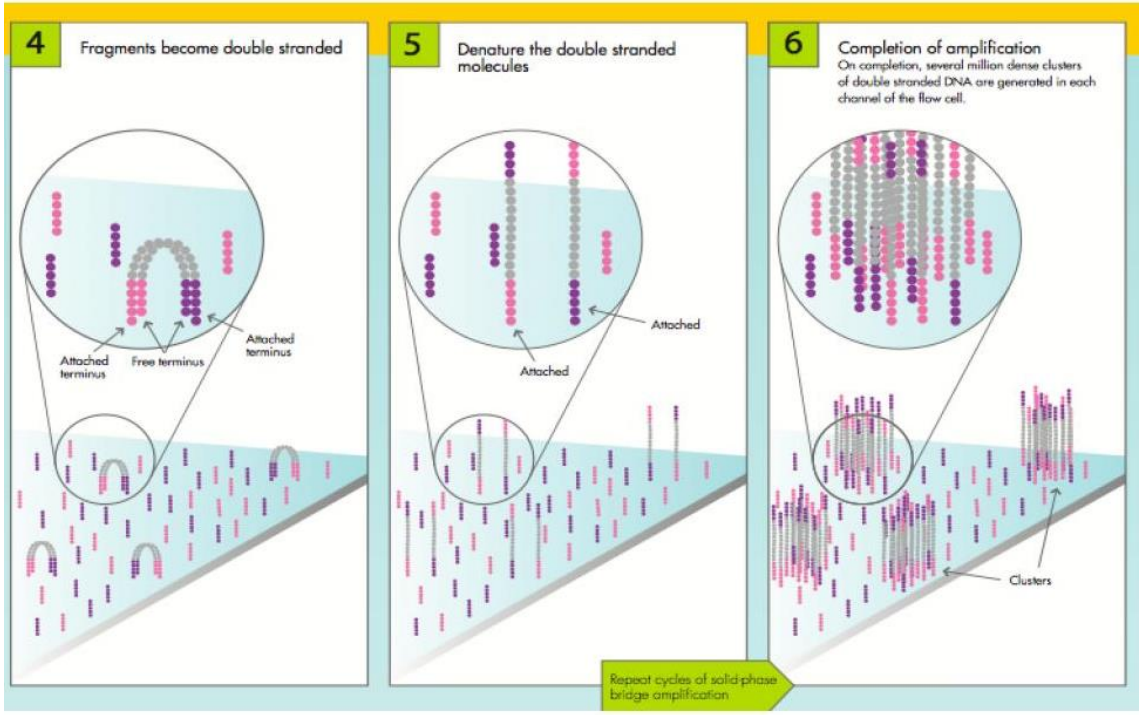
Run Time	12–30 hours	< 1–3.5 days	< 3 days	~13 - 38 hours (dual SP flow cells) ~13–25 hours (dual S1 flow cells) ~16–36 hours (dual S2 flow cells) ~44 hours (dual S4 flow cells)
Maximum Output	120 Gb	1500 Gb	1800 Gb	6000 Gb
Maximum Reads Per Run	400 million	5 billion	6 billion	20 billion
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 150 bp	2 x 250**

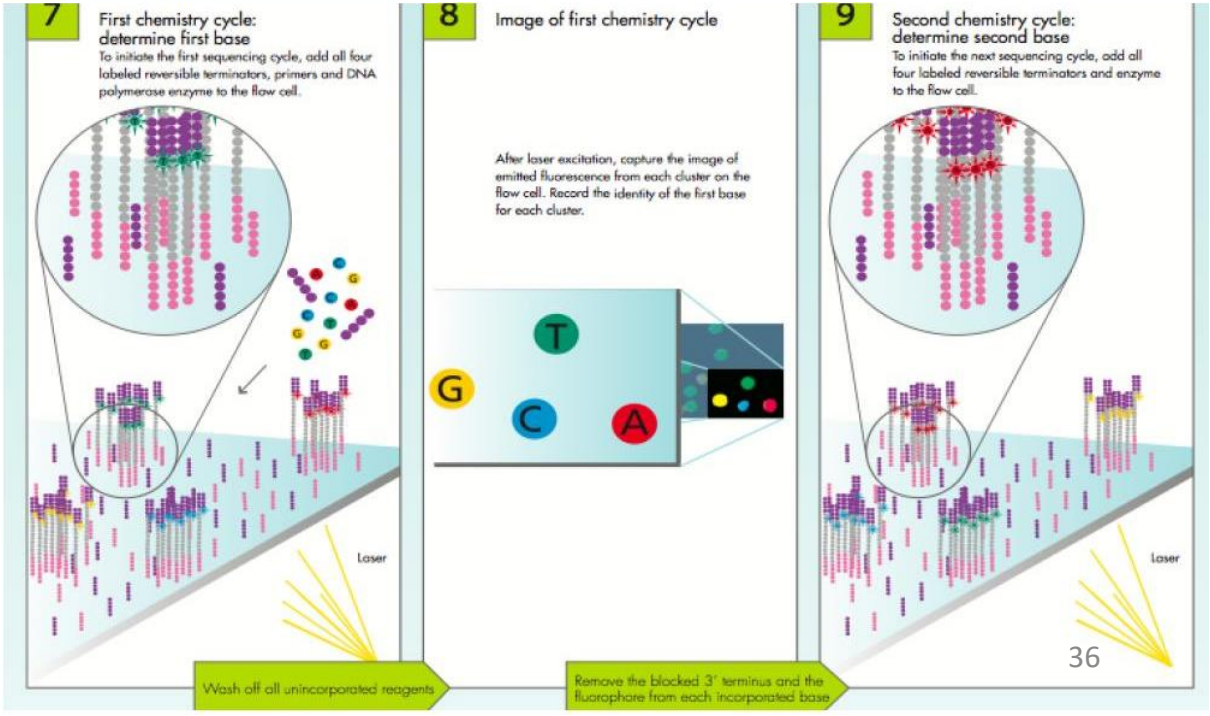
Illumina Novaseq

https://emea.illumina.com/systems/sequencing-platforms/novaseq/system-explorer.html?langsel=/nl/

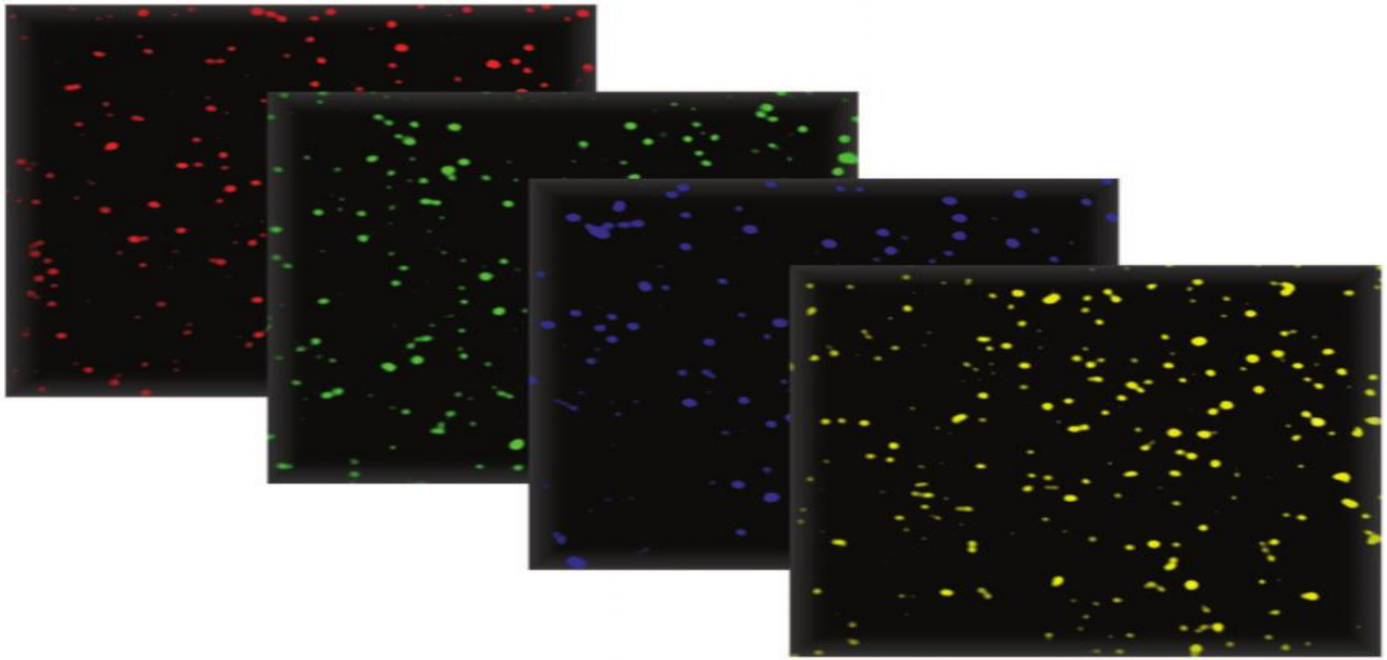


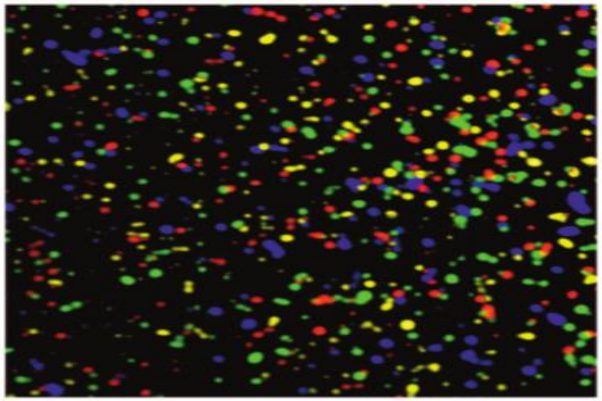

1
Prepare genomic DNA sample
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.
DNA
Adapters
Add sample to flow cell
2
Attach DNA to surface
Bind single stranded fragments randomly to the inside surface of the flow cell channels.
Adapter
DNA fragment
Dense lawn of primers
Adapter
3
Bridge amplification
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

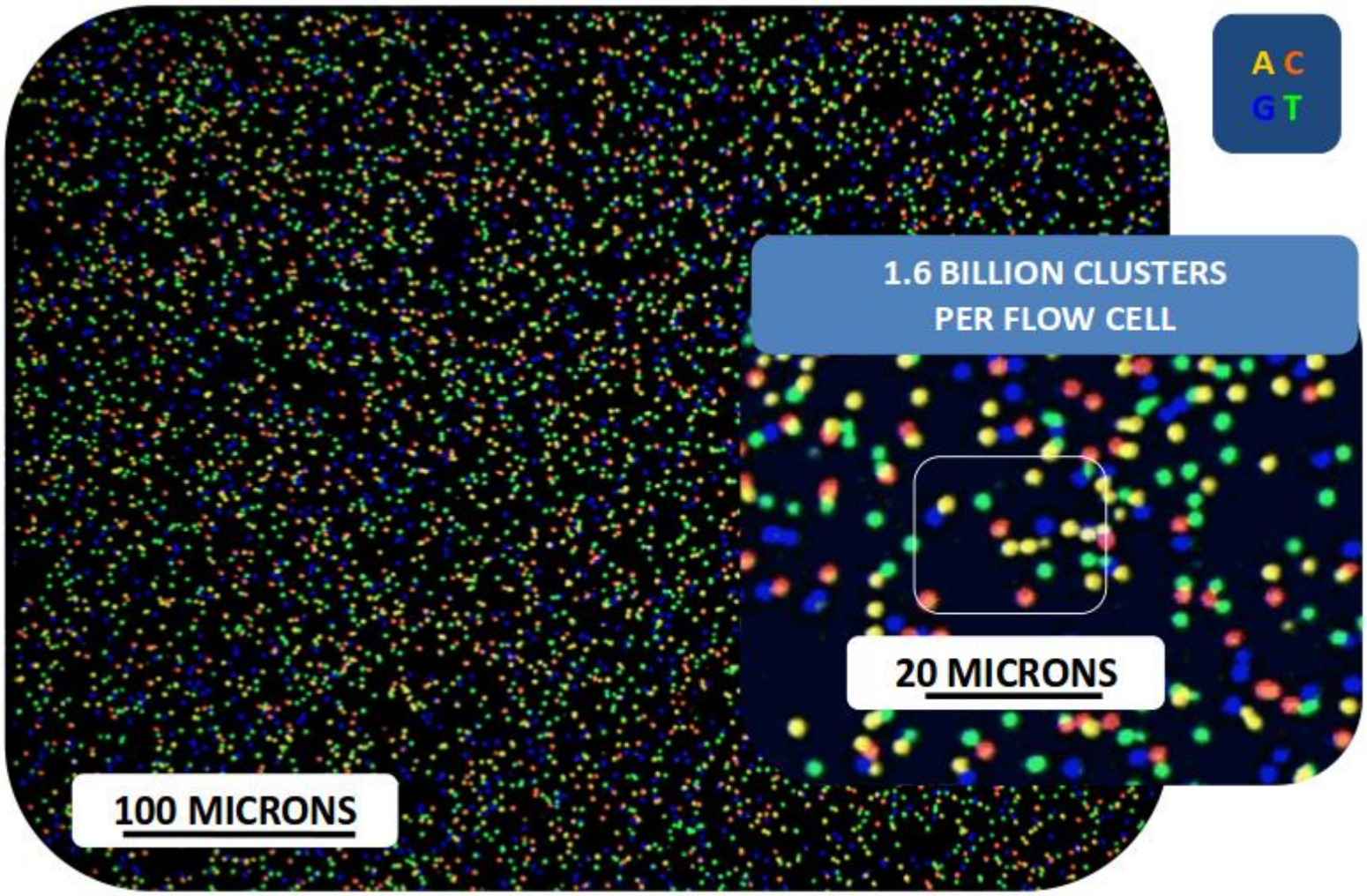

4
Fragments become double stranded
Attached terminus
Free terminus
Attached terminus
5
Denature the double stranded molecules
Attached
Attached
6
Completion of amplification
On completion, several million dense clusters of double stranded DNA are generated in each channel of the flow cell.
Clusters
Repeat cycles of solid-phase bridge amplification


7
First chemistry cycle: determine first base
To initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.
Laser
Wash off all unincorporated reagents
8
Image of first chemistry cycle
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.
T
G
C
A
Remove the blocked 3' terminus and the fluorophore from each incorporated base
9
Second chemistry cycle: determine second base
To initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.
Laser

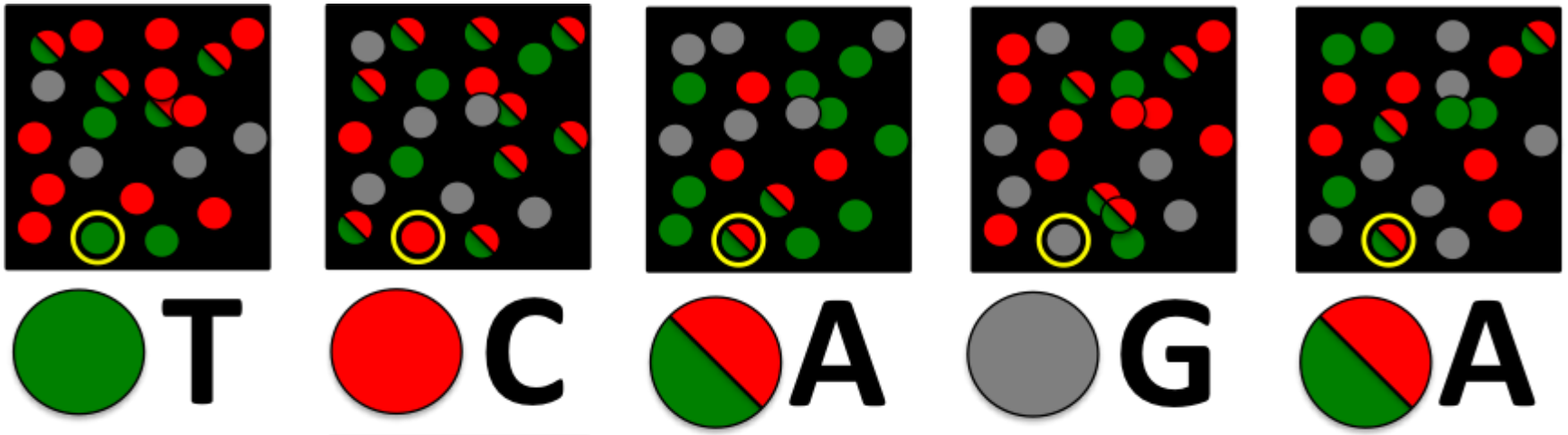
4 channels chemistry



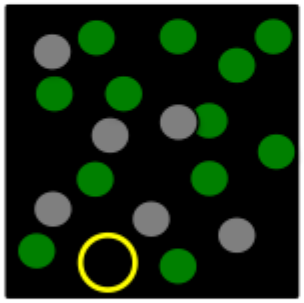



A C
G T
1.6 BILLION CLUSTERS
PER FLOW CELL
20 MICRONS
100 MICRONS

2 channel chemistry

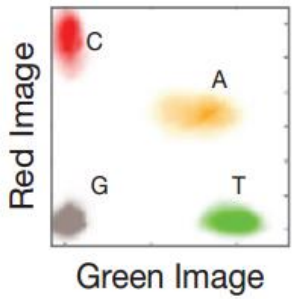


Green channel



Red channel

Figure 2: Two-Channel SBS Imaging



Accelerated detection of all four DNA bases is performed on the NextSeq 500 System using only two images to capture red and green filter wavelength bands. A bases will be present in both images (yellow cluster), C bases in red only, T bases in green only, and G bases in neither.

A common output:

FASTQ files

1 ”read”

```
@read1
AGCTTATCCTCTGCTCACCCCCGGGTTAGCGCACTTGATGTATTCACAGC
+
BA1@CC7CBCCC9C8;B2@>C?B@B@B3=9?@B1:AB7B?B8B?B6B.7.
@read2
TTGGGCGGGATCTCCAGAAGCATATGGATGTGATCCACACAGCATTCTGC
+
?>?B@)<?@,AA7A@C<C?=@@B;+)?B5*@2=@+=BB,=B6C>AB@B24
@read3
TATGCTCAAGAAGGGGCTGATGAGTTGGTGTTTTACGATATCACTGCCTC
+
A3AB:B1:B;9/0BBBCBB<BB@AA0?BB9:BB<A@BB@7@6@<A@@@<3
```

Several billions


Coffee break

A common output:

FASTQ files

1 ”read”

```
@read1
AGCTTATCCTCTGCTCACCCCCGGGTTAGCGCACTTGATGTATTCACAGC
+
BA1@CC7CBCCC9C8;B2@>C?B@B@B3=9?@B1:AB7B?B8B?B6B.7.
@read2
TTGGGCGGGATCTCCAGAAGCATATGGATGTGATCCACACAGCATTCTGC
+
?>?B@)<?@,AA7A@C<C?=@@B;+)?B5*@2=@+=BB,=B6C>AB@B24
@read3
TATGCTCAAGAAGGGGCTGATGAGTTGGTGTTTTACGATATCACTGCCTC
+
A3AB:B1:B;9/0BBBCBB<BB@AA0?BB9:BB<A@BB@7@6@<A@@@<3
```

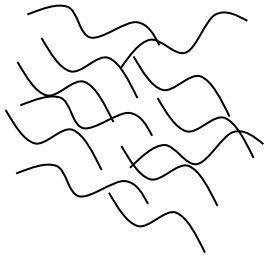
Several billions

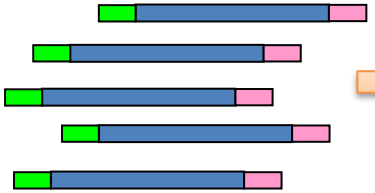
NGS QScore

Table 1: Quality Scores and Base Calling Accuracy

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy	ASCII
10	1 in 10	90%	+
20	1 in 100	99%	5
30	1 in 1,000	99.9%	?
40	1 in 10,000	99.99%	|

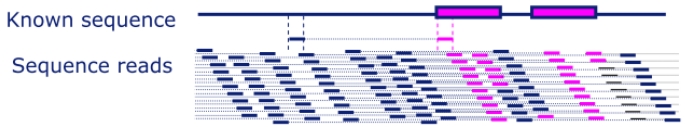
RNA-Sequencing





```
@read1
AGCTTATCCTCTGCTCACCCCCGGGTTAGCGCACTTGATGTATTCACAGC
+
BA1@CC7CBCCC9C8;B2@>C?B@B@B3=9?@B1:AB7B?B8B?B6B.7.
@read2
TTGGGCGGGATCTCCAGAAGCATATGGATGTGATCCACACAGCATTCTGC
+
?>?B@)<?@,AA7A@C<C?=@@B;+)?B5*@2=@+=BB,=B6C>AB@B24
```

Trimming / Mapping / alignment



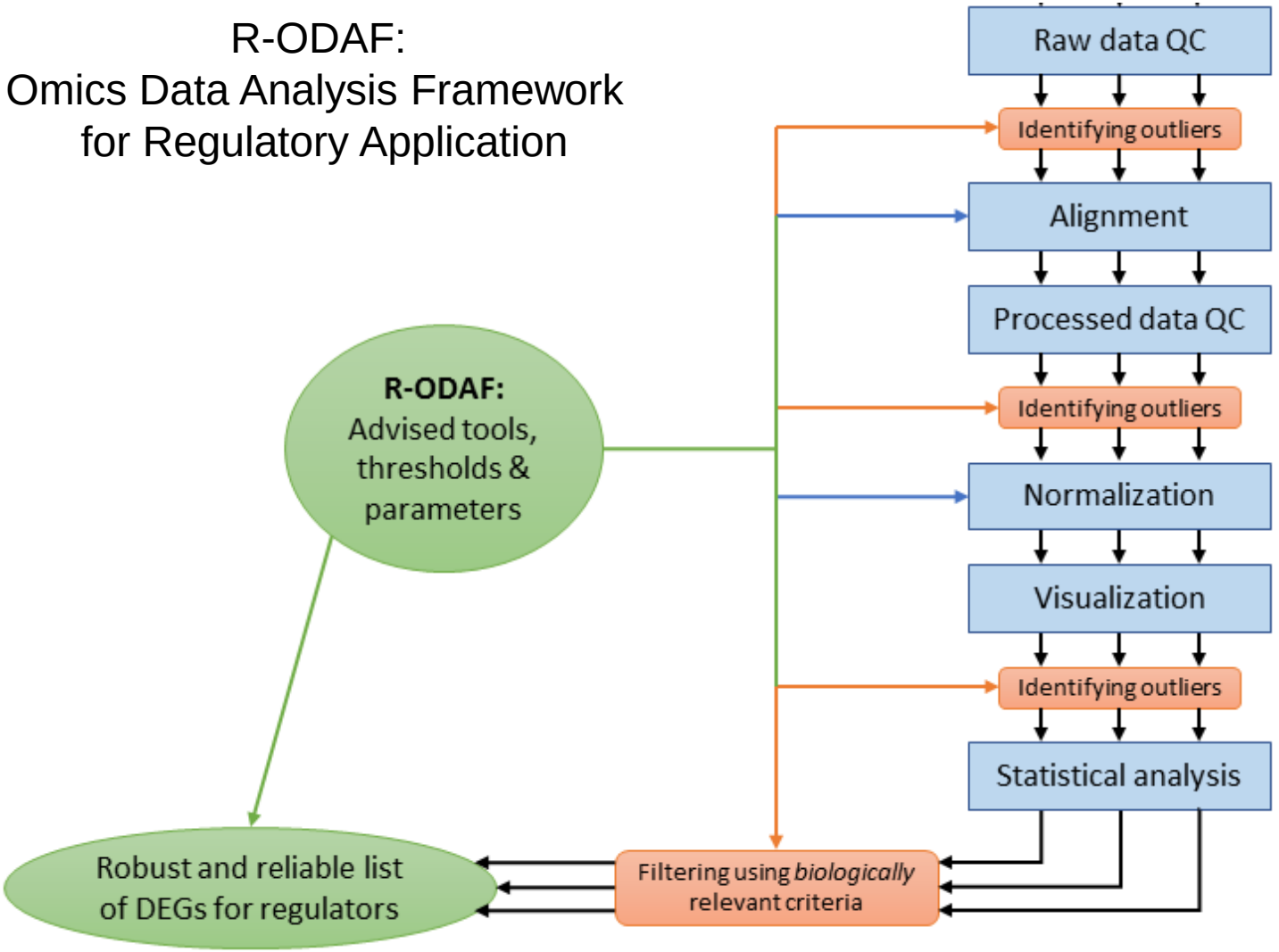
quantification

Raw count table

Gene_id	Sample_1	Sample_2	Sample_3	Sample_4
ENSG00000000003	639	304	616	324
ENSG00000000005	2287	1071	1906	729
ENSG00000000419	534	58	253	298
ENSG00000000457	390	301	29	174
...	...	...	...	...

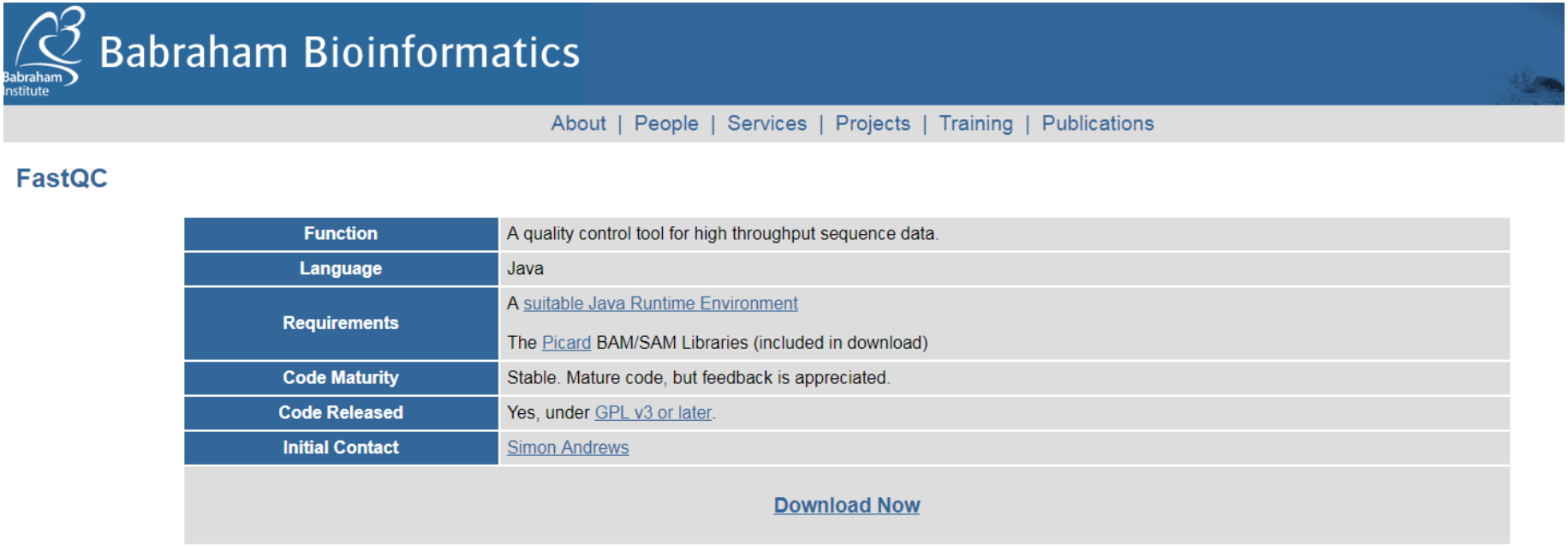
Statistical Analysis

R-ODAF:
Omics Data Analysis Framework for Regulatory Application



https://www.sciencedirect.com/science/article/pii/S0273230022000307

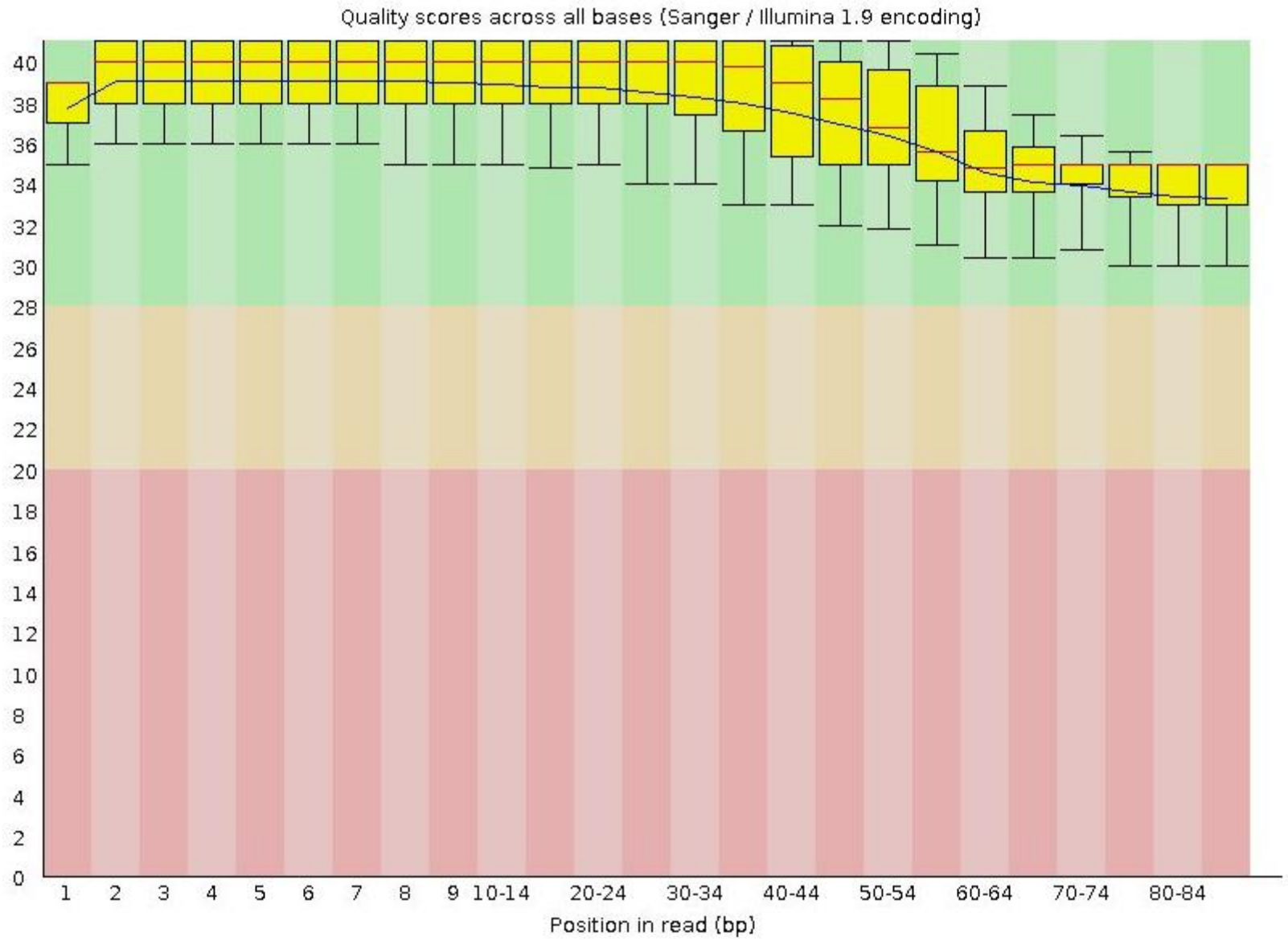
FastQC

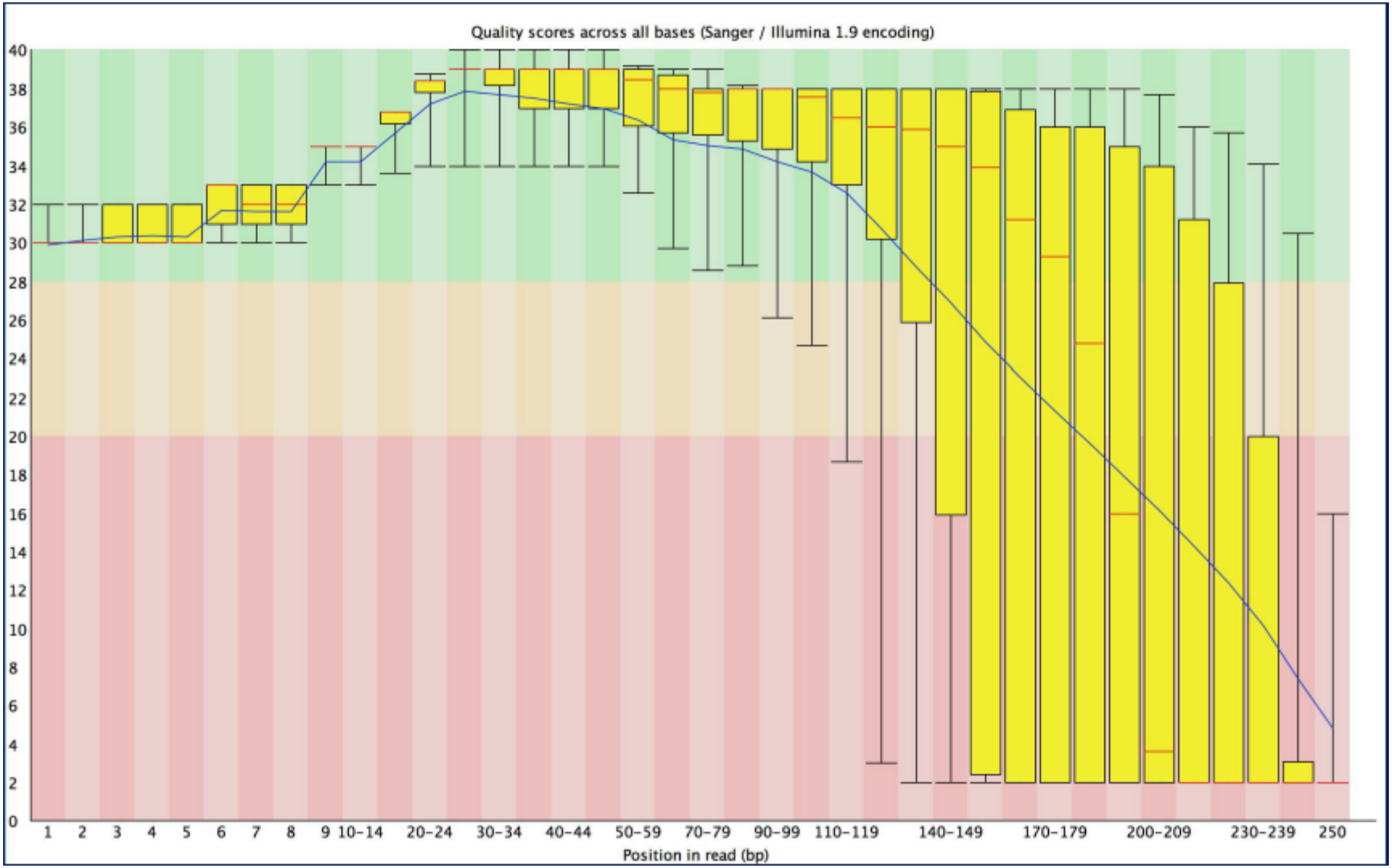



```
@read1
AGCTTATCCTCTGCTCACCCCCGGGTTAGCGCACTTGATGTATTCACAGC
+
BA1@CC7CBCCC9C8;B2@>C?B@B@B3=9?@B1:AB7B?B8B?B6B.7.
```

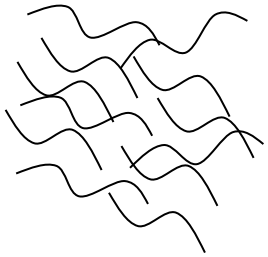
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

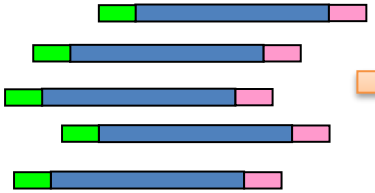
Per base sequence quality




Quality scores across all bases (Sanger / Illumina 1.9 encoding)
40
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1
2
3
4
5
6
7
8
9
10-14
20-24
30-34
40-44
50-59
70-79
90-99
110-119
140-149
170-179
200-209
230-239
250
Position in read (bp)

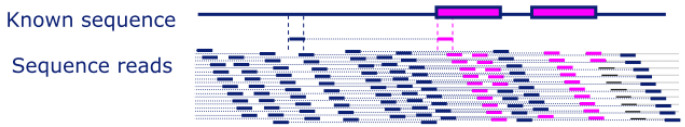
RNA-Sequencing





```
@read1
AGCTTATCCTCTGCTCACCCCCGGGTTAGCGCACTTGATGTATTCACAGC
+
BA1@CC7CBCCC9C8;B2@>C?B@B@B3=9?@B1:AB7B?B8B?B6B.7.
@read2
TTGGGCGGGATCTCCAGAAGCATATGGATGTGATCCACACAGCATTCTGC
+
?>?B@)<?@,AA7A@C<C?=@@B;+)?B5*@2=@+=BB,=B6C>AB@B24
```

Trimming / Mapping - alignment



Raw count table

Gene_id	Sample_1	Sample_2	Sample_3	Sample_4
ENSG00000000003	639	304	616	324
ENSG00000000005	2287	1071	1906	729
ENSG00000000419	534	58	253	298
ENSG00000000457	390	301	29	174
...	...	...	...	...

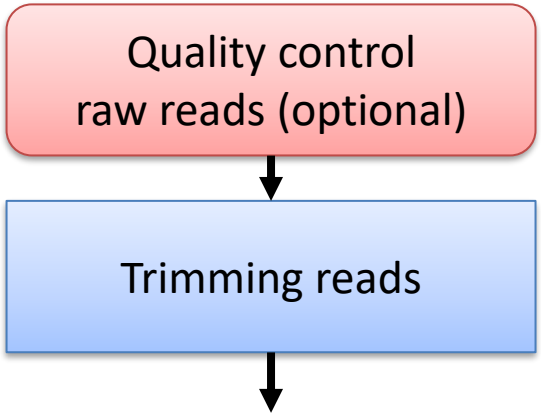
quantification

Statistical Analysis

Dr. Marcha Verheijen

- **List of Alignment bioinformatics tools**

QC/trimming tools	cites	year	cites/year
Trimmomatic	8247	2014	1649.4
HTSeq	4993	2015	1248.3
FastQC	2278	2010	253.1
NGS QC toolkit	1133	2012	161.9
MultiQC	292	2016	97.3
SolexaQA	813	2010	90.3
fastp	45	2018	45



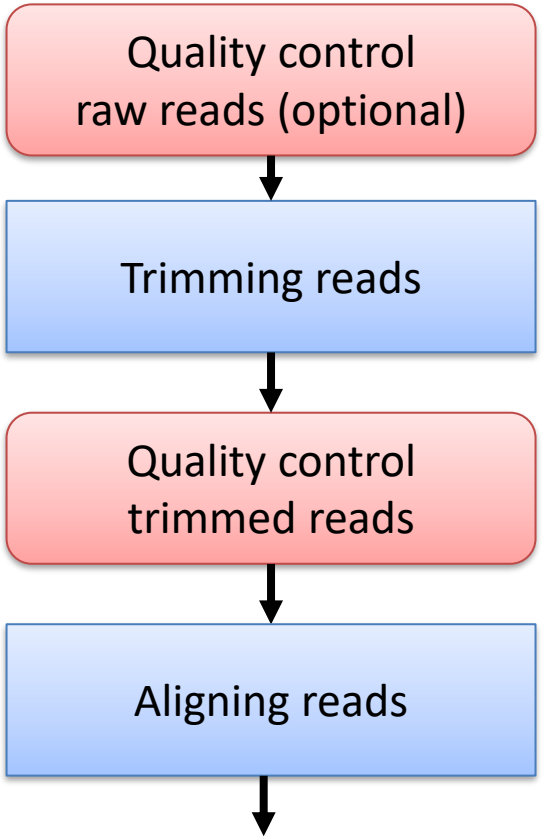
Trimming

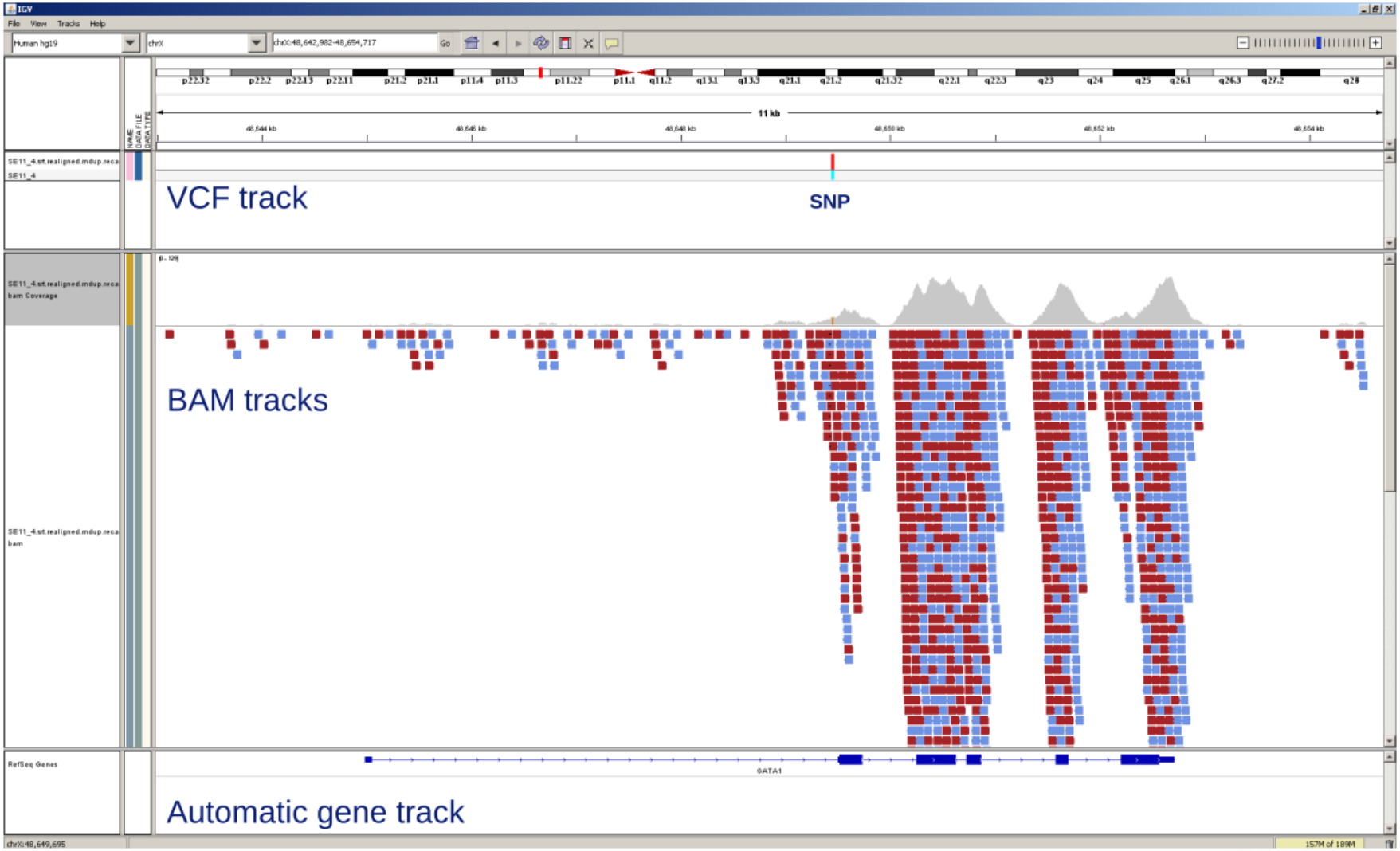
Dr. Marcha Verheijen

- **List of Alignment bioinformatics tools**

QC/trimming tools	cites	year	cites/year
Trimmomatic	8247	2014	1649.4
HTSeq	4993	2015	1248.3
FastQC	2278	2010	253.1
NGS QC toolkit	1133	2012	161.9
MultiQC	292	2016	97.3
SolexaQA	813	2010	90.3
fastp	45	2018	45

Alignment tools	cites	year	cites/year
BWA	18717	2009	1871.7
Bowtie2	12733	2012	1819.0
TopHat2	6049	2012	864.1
STAR	5741	2013	956.8
SOAP2	2816	2009	281.6
HiSat2	2123	2015	530.7




VCF track
SNP
BAM tracks
Automatic gene track

Alignment file : Sam

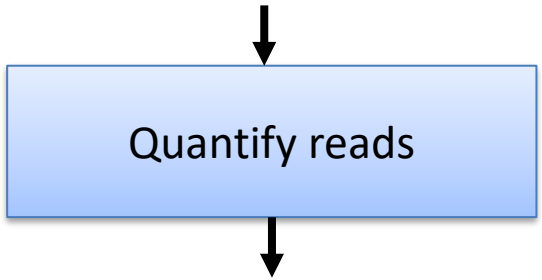
- Read name
- Map: 0 OK, 4 unmapped, 16 mapped reverse strand
- Sequence, quality score, XA (mapper-specific)
- MD: mismatch info: 3 match, then C ref, 30 match, then T ref, 3 match
- NM: number of mismatch
- BAM: binary compressed SAM format.

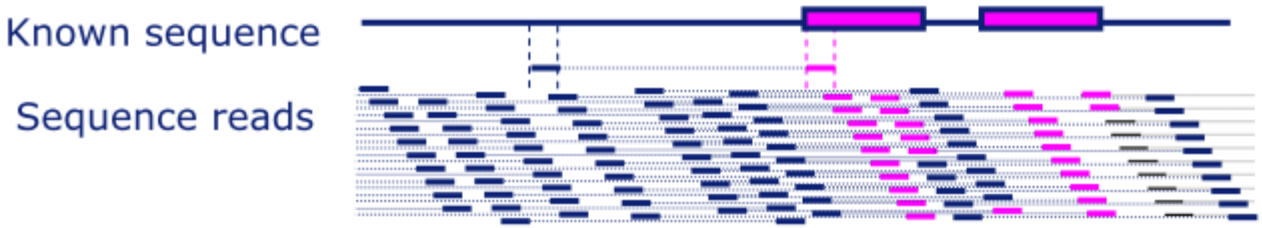
```
HWUSI_EAS366_0112:6:1:1298:18828#0/1    16      chr9    98116600        255     38M     *
     0       0       TACAATATGTCTTTATTTGAGATATGGATTTTAGGCCG  Y\]bc^dab\[_UU`^`LbTUT\ccLbbYaY`
   cWLYW^  XA:i:1  MD:Z:3C30T3     NM:i:2
HWUSI_EAS366_0112:6:1:1257:18819#0/1    4       *       0       0       *       *       0
     0       AGACCACATGAAGCTCAAGAAGAAGGAAGACAAAAGTG  ece^dddT\cT^c`a`ccdK\c^^__]Yb\_cKS^_W\
   XM:i:1
HWUSI_EAS366_0112:6:1:1315:19529#0/1    16      chr9    102610263       255     38M     *
     0       0       GCACTCAAGGGTACAGGAAAAGGGTCAGAAGTGTGGCC  ^c_Yc\Lcb`bbYdTa\dd\`dda`cdd\Y\d
   dd^cT`  XA:i:0  MD:Z:38 NM:i:0
```

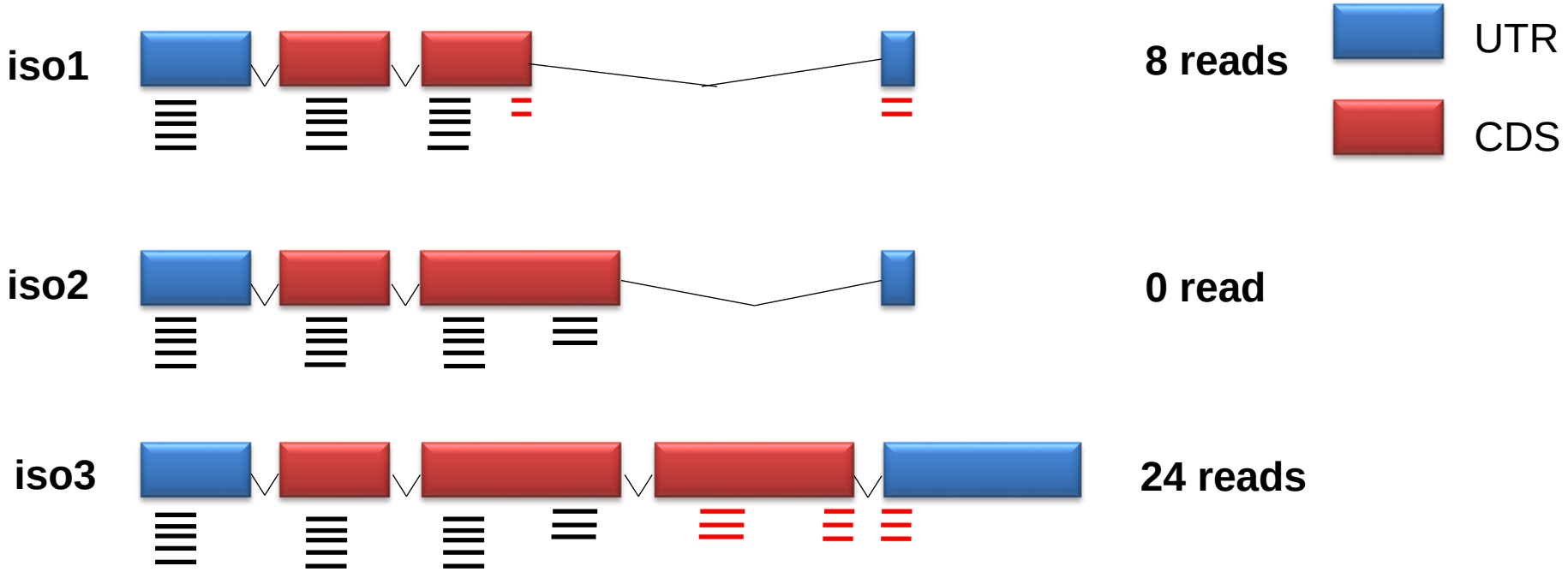
Bam: Binary compression of the Sam file

- **List of quantification tools**

Quantification tools	cites	year	cites/year
Cufflinks	6049	2009	604.9
RSEM	4670	2011	583.8
kallisto	1077	2016	359.0
featurecounts	2047	2013	341.2
Subread	756	2013	126.0







Total Gene: 32 reads

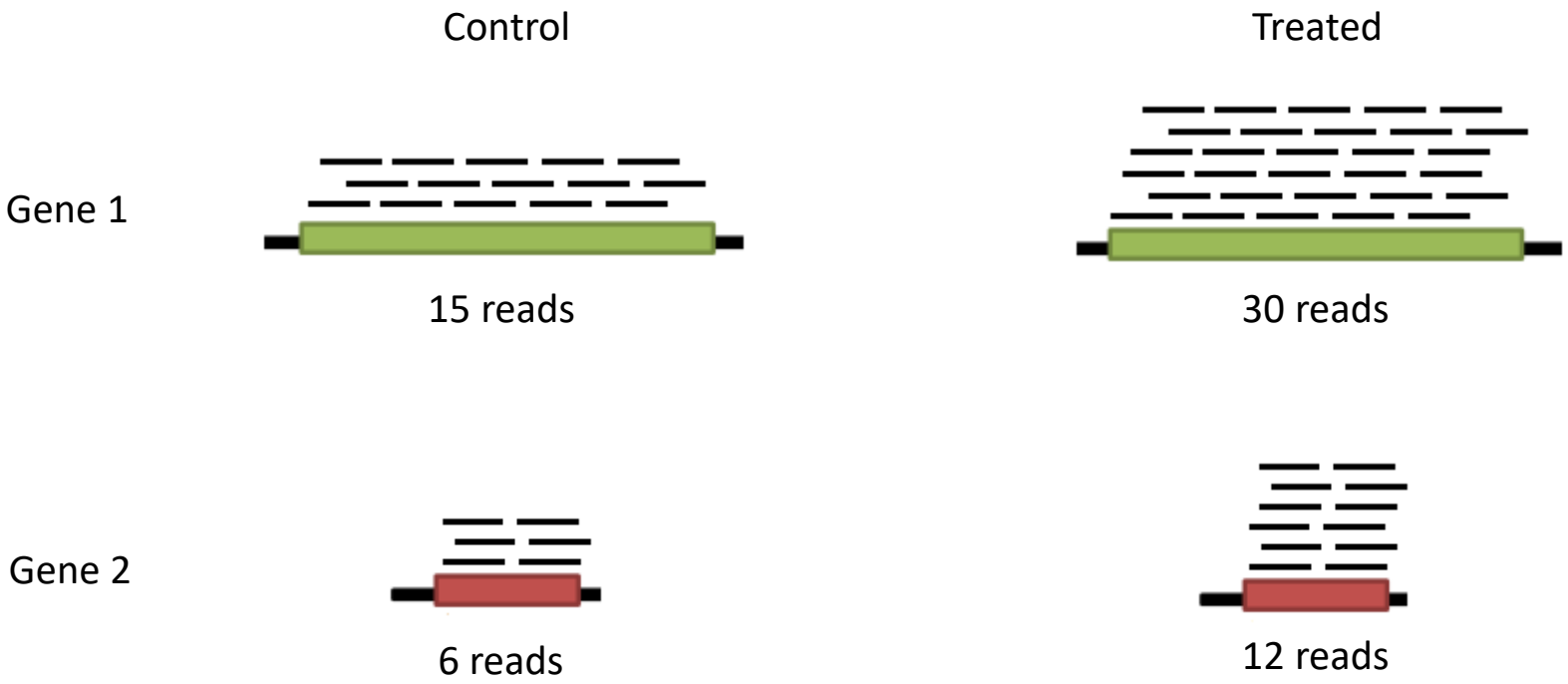
Raw mapped data

Gene_id	Sample_1	Sample_2	Sample_3	Sample_4
ENSG00000000003	639	304	616	324
ENSG00000000005	2287	1071	1906	729
ENSG00000000419	534	58	253	298
ENSG00000000457	390	301	29	174
ENSG00000000460	419	226	144	207
ENSG00000000938	1355	948	1967	886
ENSG00000000971	636	396	691	247
ENSG00000001036	2287	1071	1906	729
ENSG00000001084	534	58	253	298
...	...	...	...	...

Normalization
Statistical Analysis
Biological Interpretation

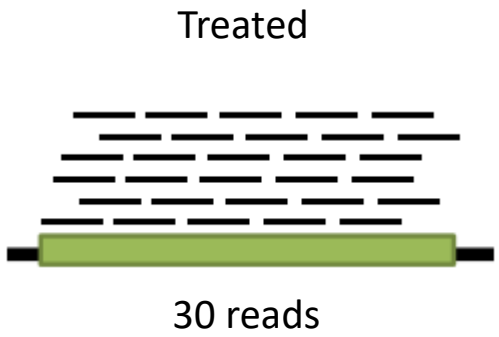
Rest of the week

Relative and Absolute



2 fold change

Relative and Absolute

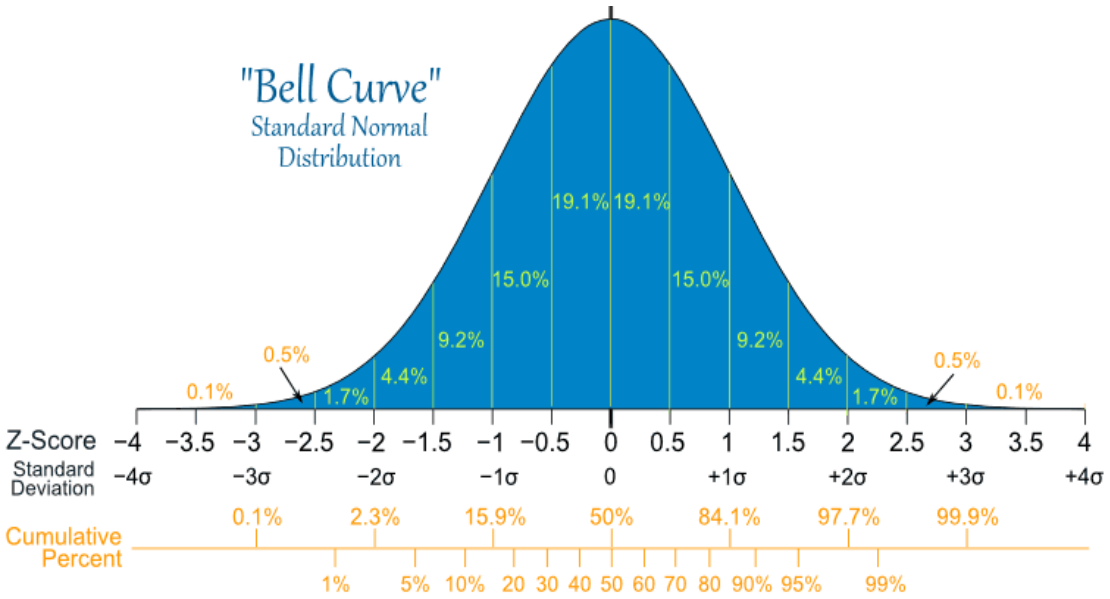


30 / 5 = 6 RPKM

12 / 2 = 6 RPKM

RPKM = Reads per Kilobase per Million mapped reads

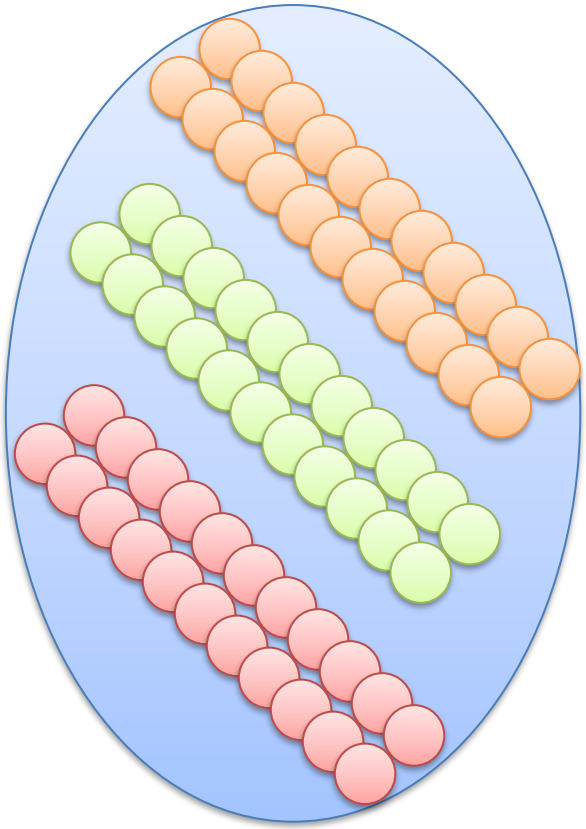
Distribution

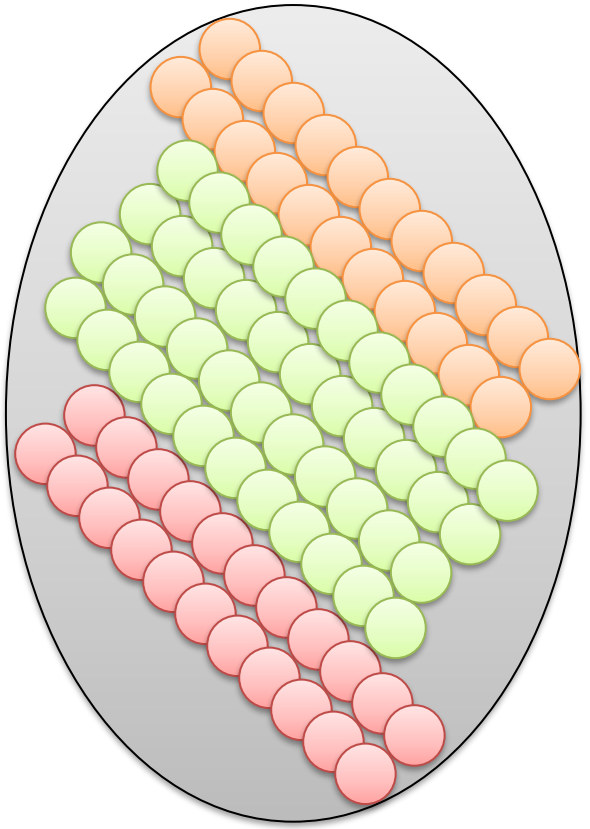


"I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order. The law would have been personified by the Greeks and deified, if they had known of it"

Francis Galton

Everywhere… but not in NGS !

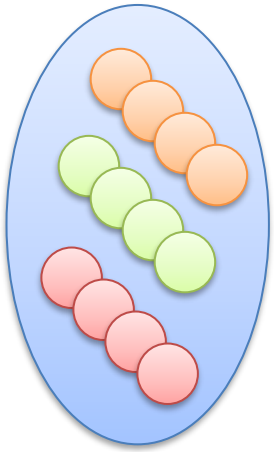




33,3 %

33,3 %

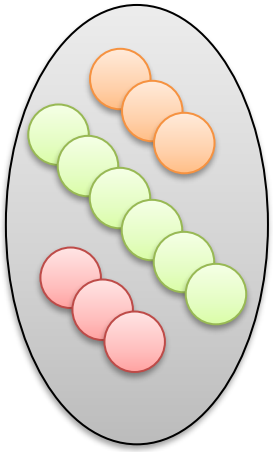
33,3 %

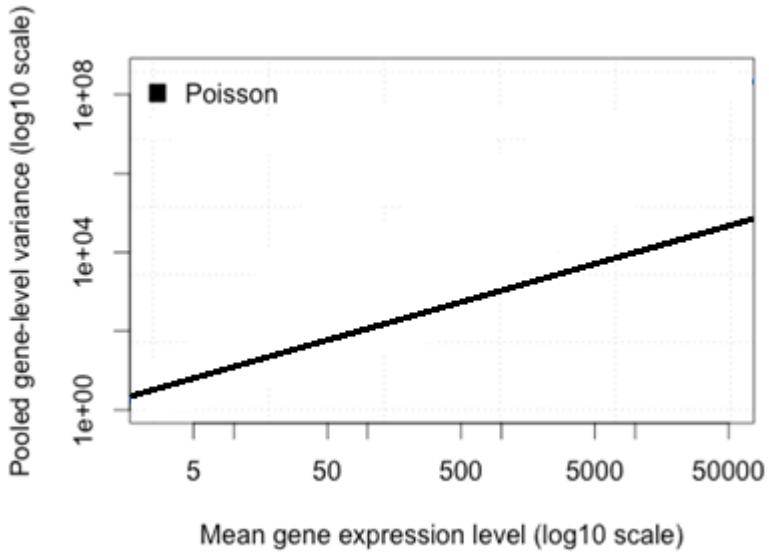


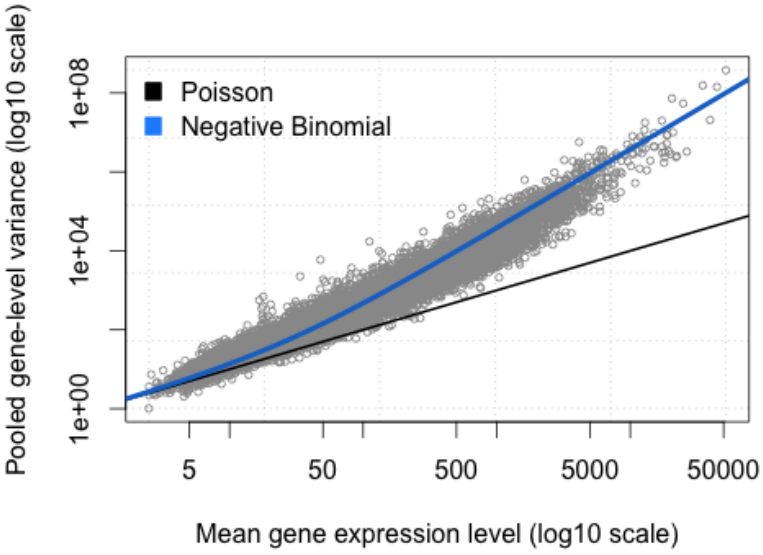
25 %

50 %

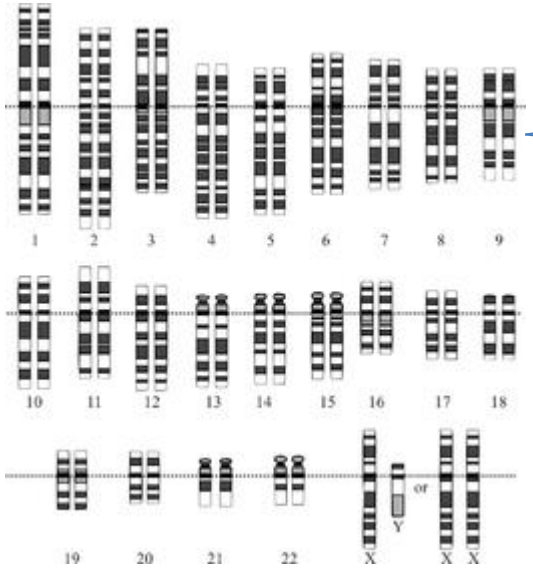
25 %

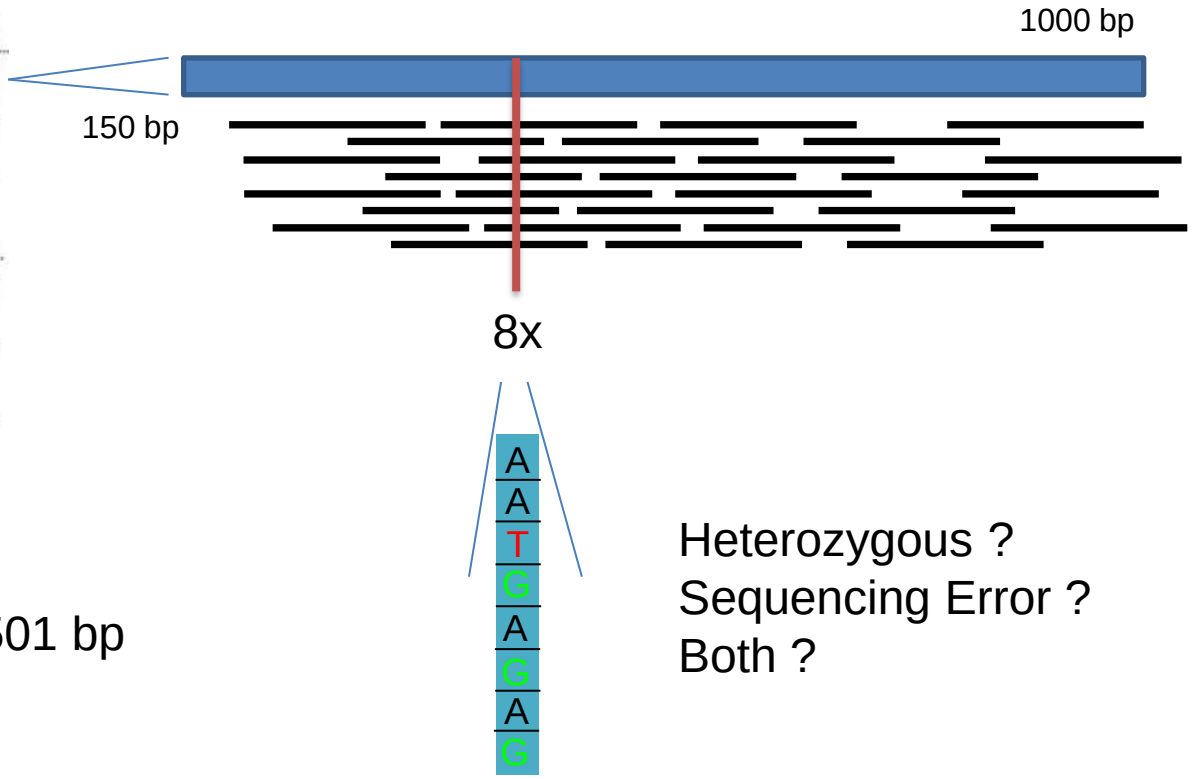



Poisson
Pooled gene-level variance (log10 scale)
1e+00
1e+04
1e+08
5
50
500
5000
50000
Mean gene expression level (log10 scale)


Poisson
Negative Binomial
Pooled gene-level variance (log10 scale)
1e+00
1e+04
1e+08
5
50
500
5000
50000
Mean gene expression level (log10 scale)

Coverage

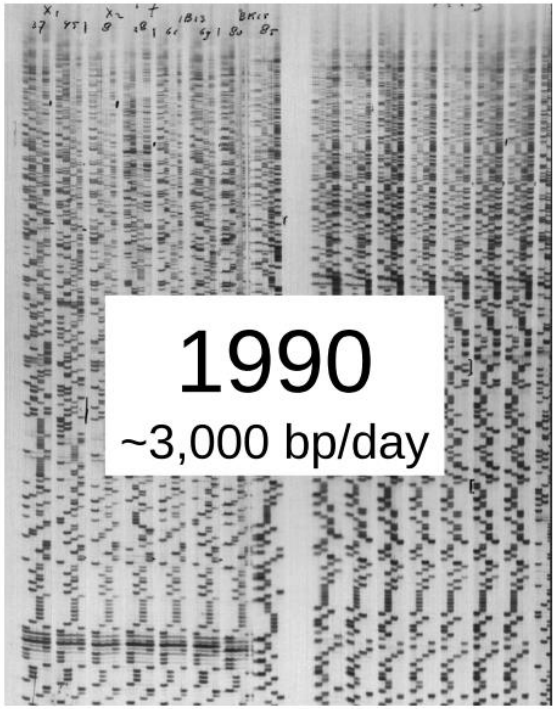


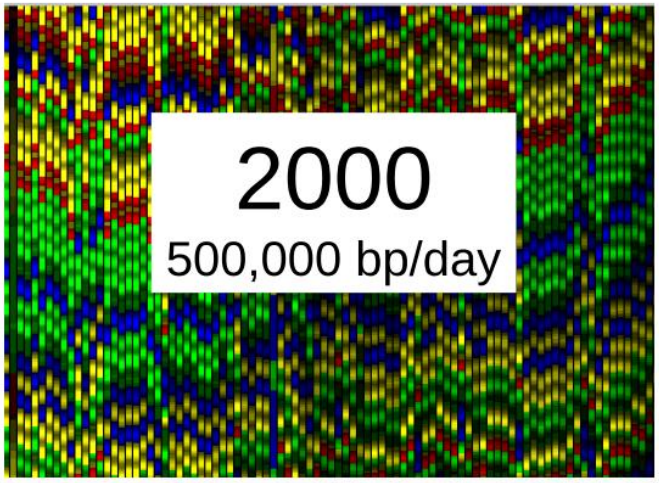


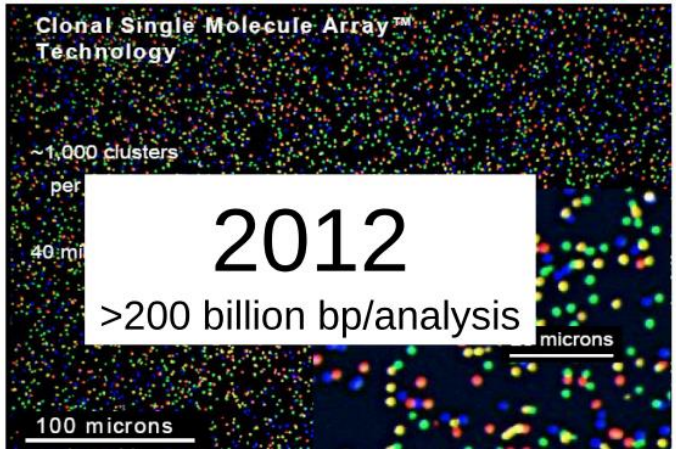
Heterozygous ?
Sequencing Error ?
Both ?

Genome size: 3,117,275,501 bp

To have a good estimation of heterozygosity versus sequencing error : 30x coverage
Need to generate 30x 3,117 billion = 93,5 billion bp per human genome







~19 billion bp/day
~ 5 days

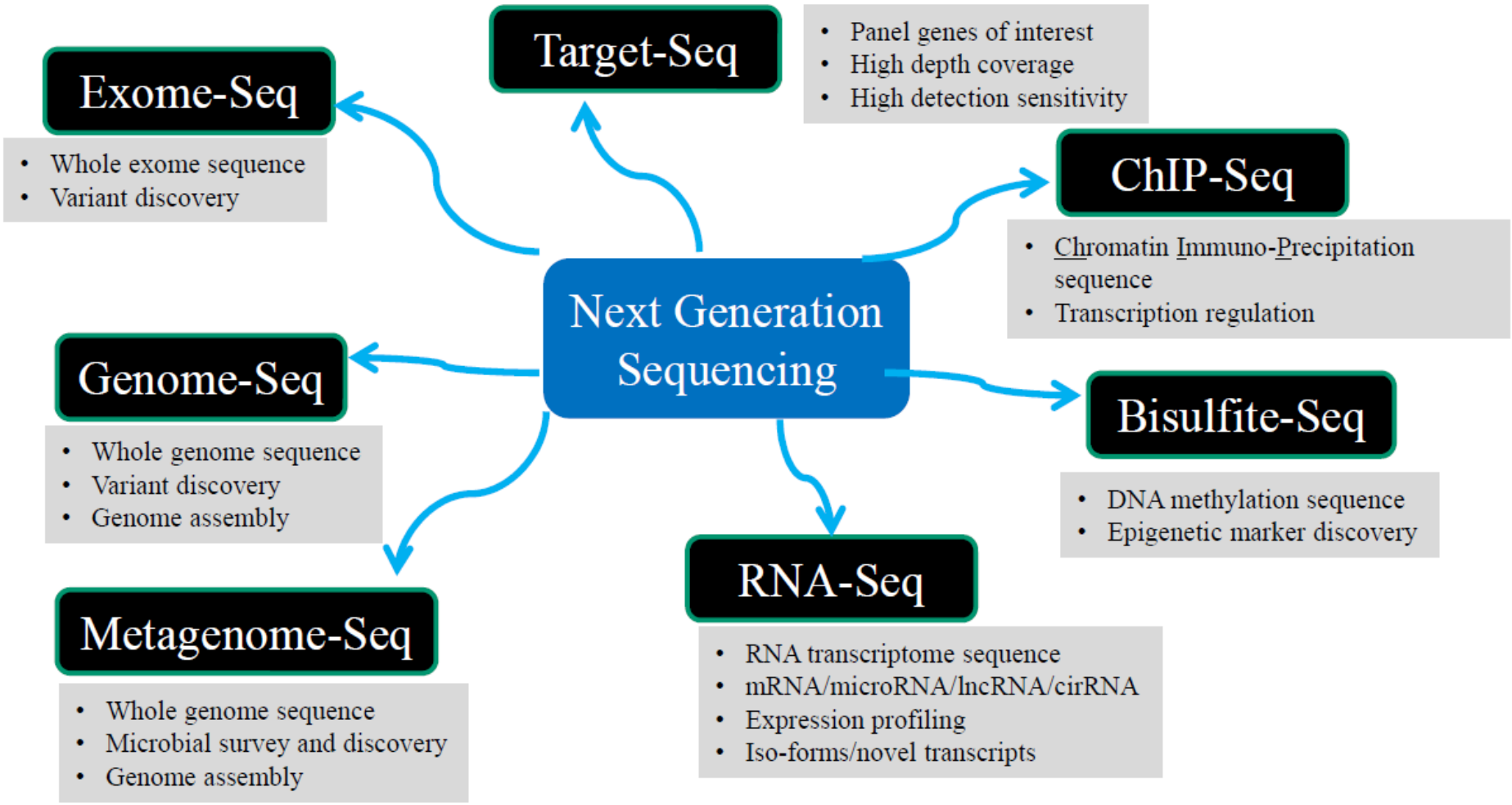
2020

>3000 billion bp/analysis
~1500 billion bp/day
16 genomes/day

2023

NovaSeq X
~7800 billion bp/day
- 80 genomes / day

NGS Application

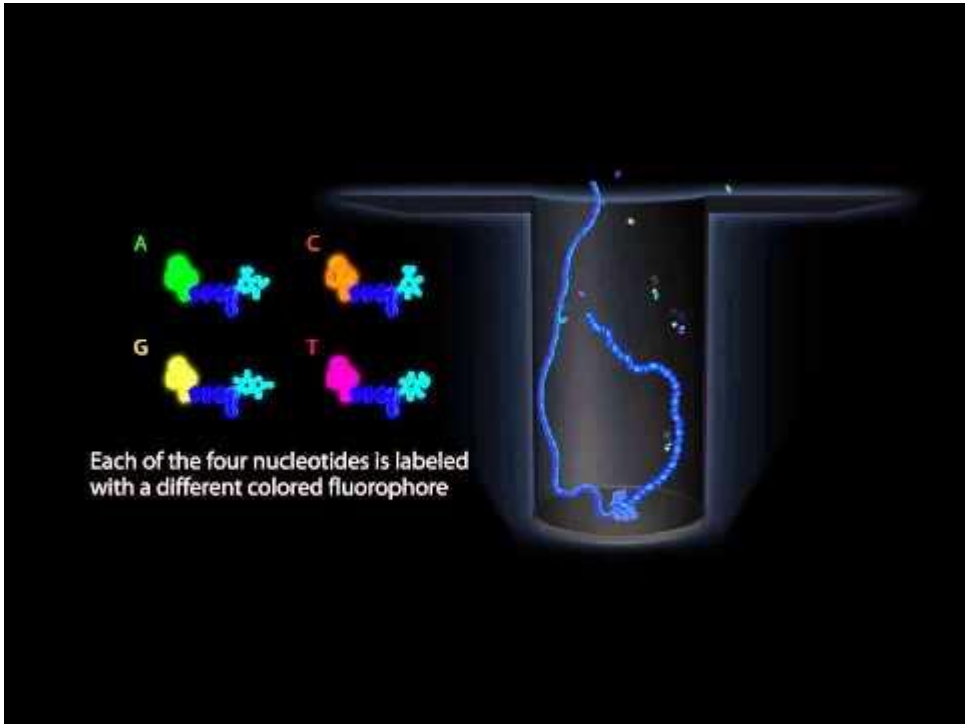


Thanks !
Questions ?

2011: PacBio

(Pacific Biosciences)

SMRT SEQUENCING

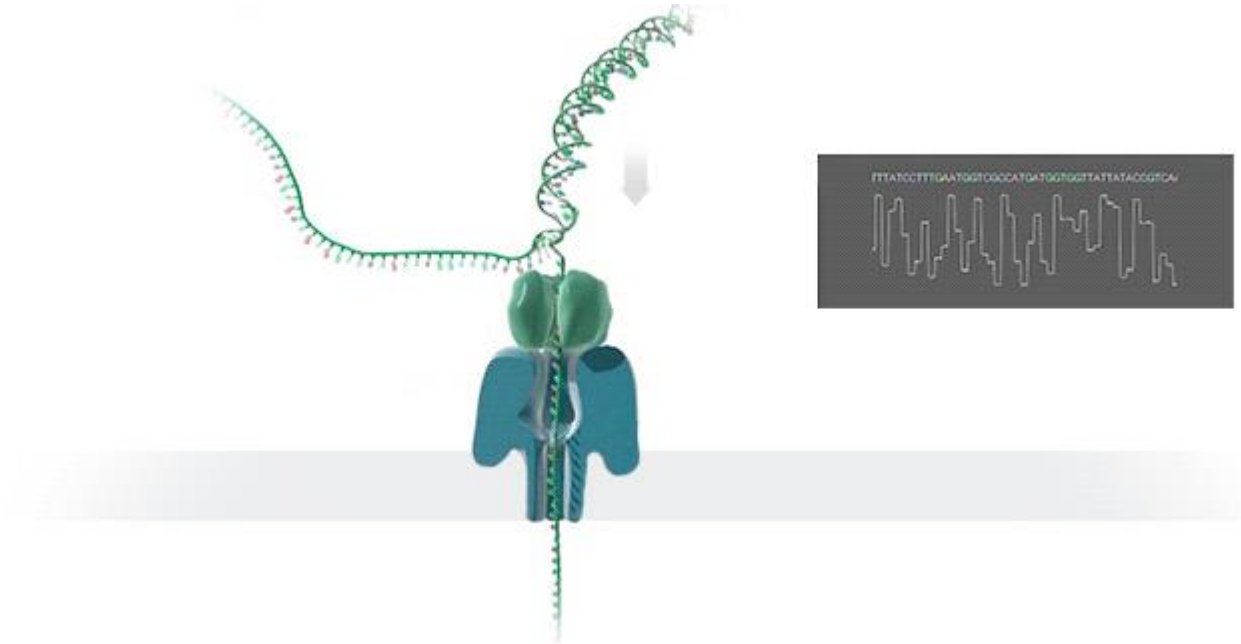


Zero-mode Waveguides (ZMWs)

DNA-Template polymerase complex

https://www.youtube.com/watch?v=NHCJ8PtYCFc

2012: Nanopore MinION



https://www.youtube.com/watch?v=GUb1TZvMWsw