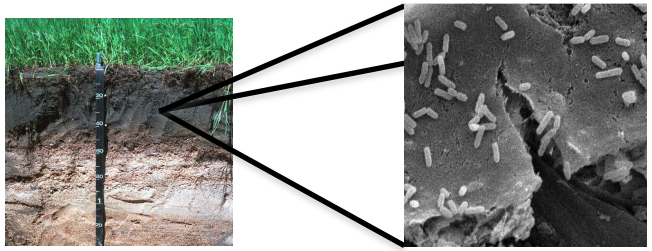
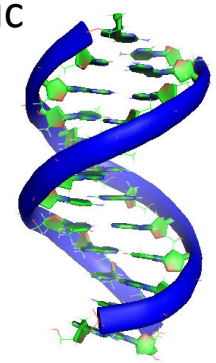
DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

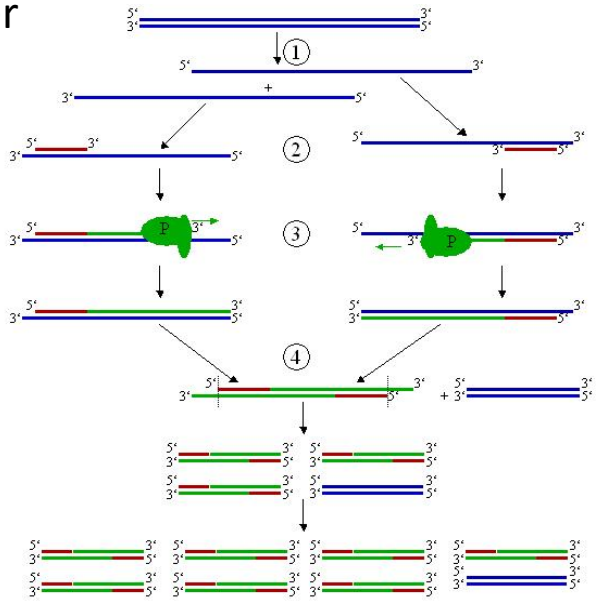
Microbial diversity with NGS-PMGA at a glance

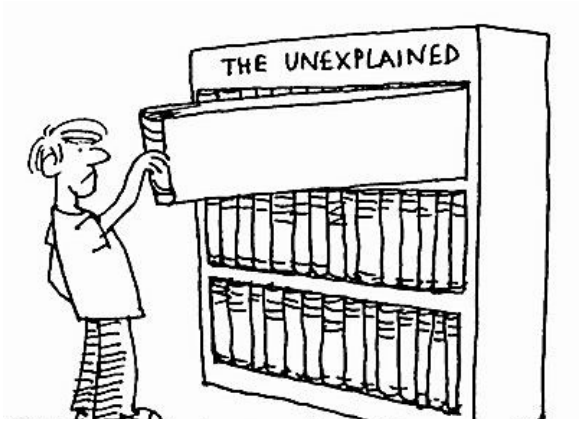


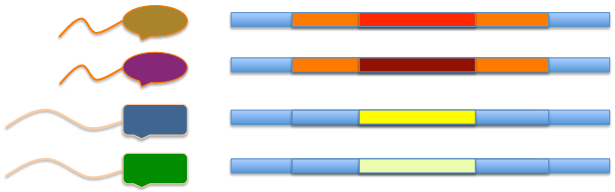
1. Sampling and nucleic acids extraction



2. PCR amplification of phylogenetic marker







4. Data analysis

3. Sequencing

DEPARTMENT OF
Biochemistry &
Biotechnology
UNIVERSITY OF THESSALY

Microbial diversity with NGS-PMGA

1. Sampling and nucleic acids extraction

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

1. Sampling and nucleic acids extraction

- Sampling
  - Representative
  - Number of replicates (be aware of power tests etc, or consult your statistician!!! E.g. PERMANOVA requires ≥4 reps for obtaining *P*-values lower than 0.05 in pairwise cmps)
  - Controls!!!
  - Selective extraction/sampling (in case of environments with same gene carrying off-target organisms; e.g. sampling of epiphytes)
  - …

- Nucleic acids extraction… steps:
  - Cell wall/membrane disruption
  - Nuclease inactivation denaturation
  - Removal of proteins and PCR inhibitors

DEPARTMENT OF
Biochemistry &
Biotechnology
UNIVERSITY OF THESSALY

Microbial diversity with NGS-PMGA

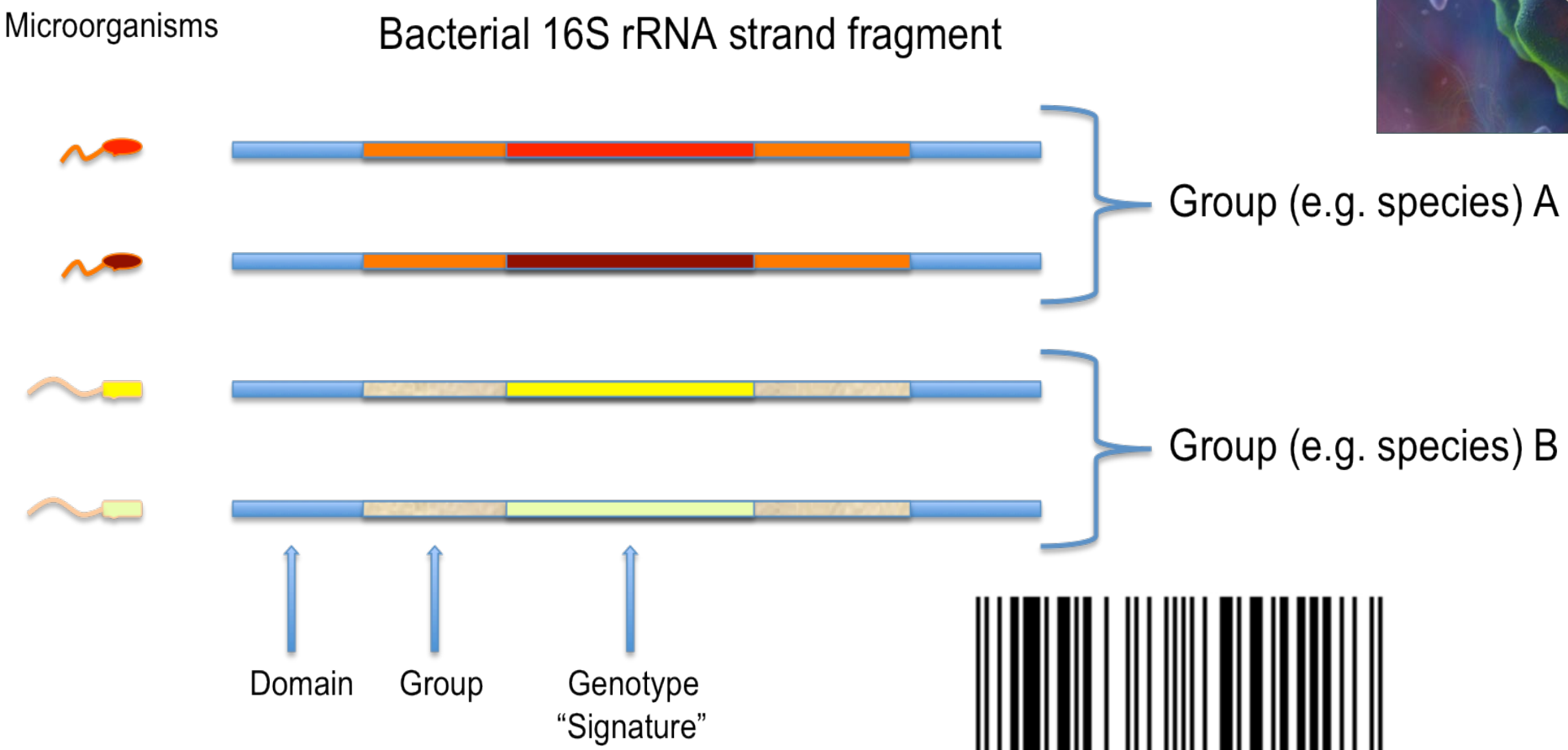
2a. Phylogenetic marker choice

2a. Phylogenetic marker choice

- Has to exist in the target organisms:
  - 16S rRNA gene: prokaryotes
  - ITS: fungi
  - 18S rRNA gene: protists, mycorhizae, eukaryotes
  - amoA: ammonia oxidizers
  - …
- Suitable conserved sequences (exist in all target organisms) that can facilitate primer designing around hypervariable regions (differ between target organisms like barcodes)
- PCR products compatible with sequencing technologies

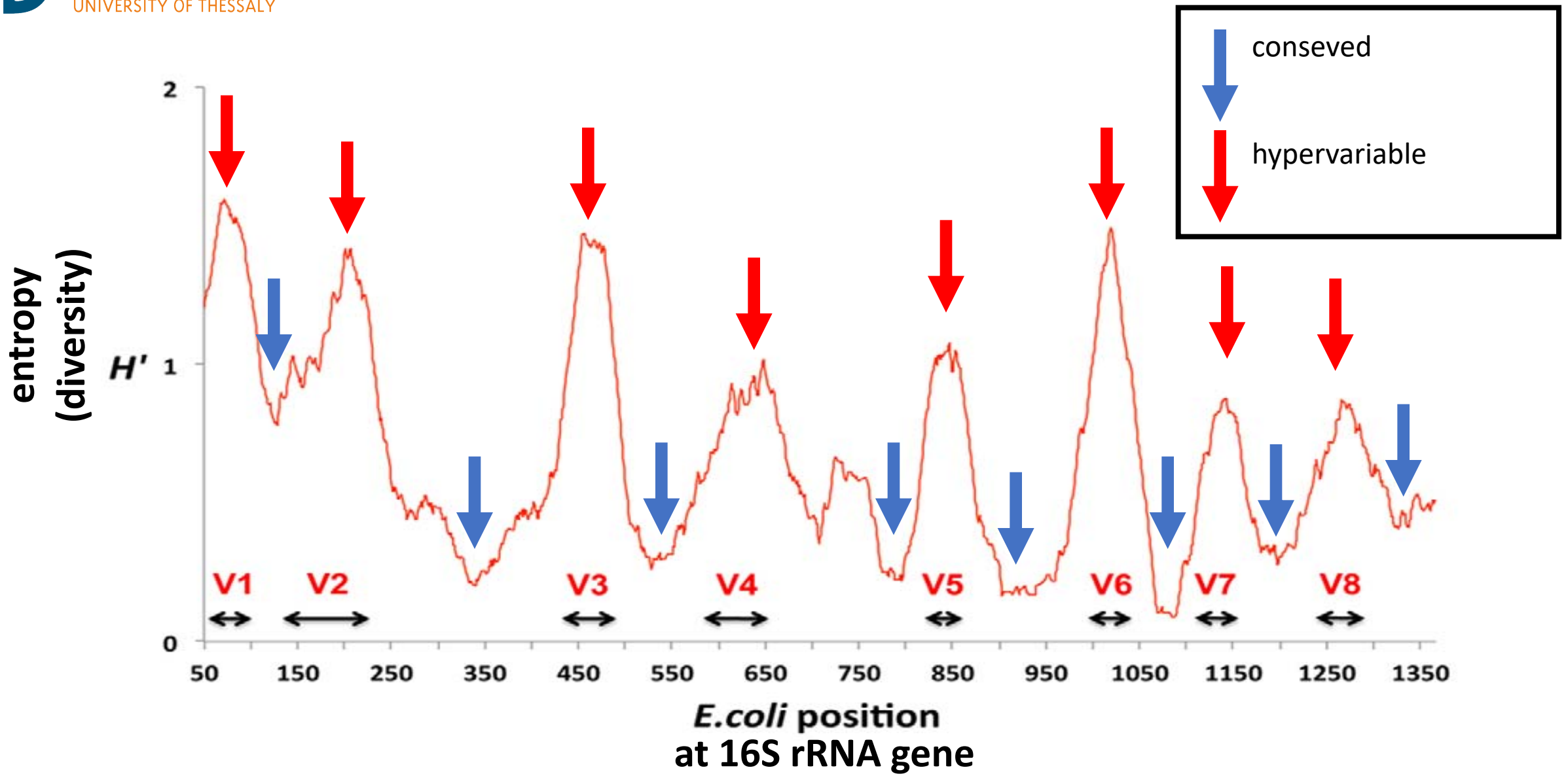
DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

2a. E.g. 16S rRNA gene



DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

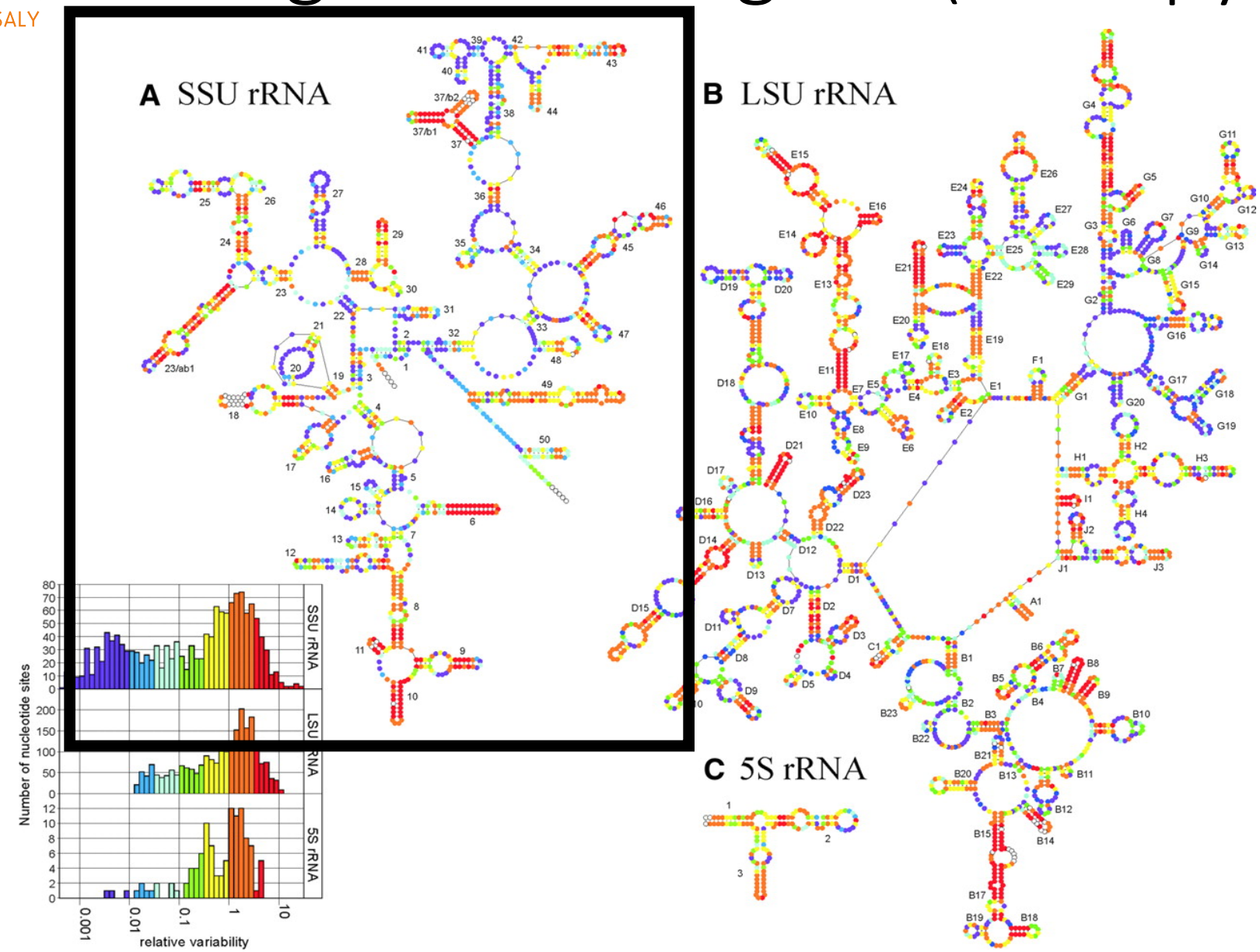
2a. E.g. 16S rRNA gene



Vasileiadis, S., et al. (2012). Plos One 7, 11, doi: 10.1371/journal.pone.0042671

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

2a. E.g. 16S rRNA gene (entropy maps)

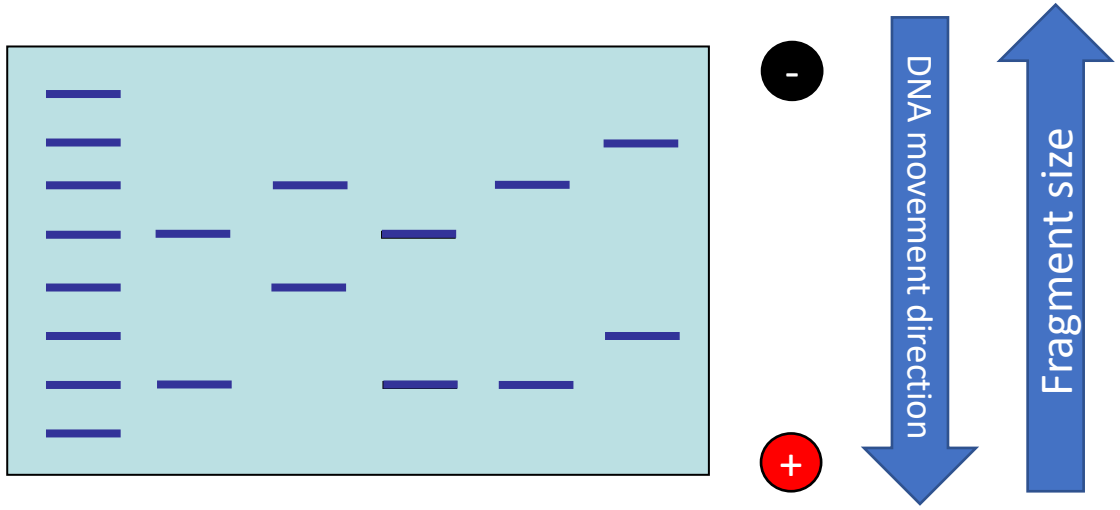


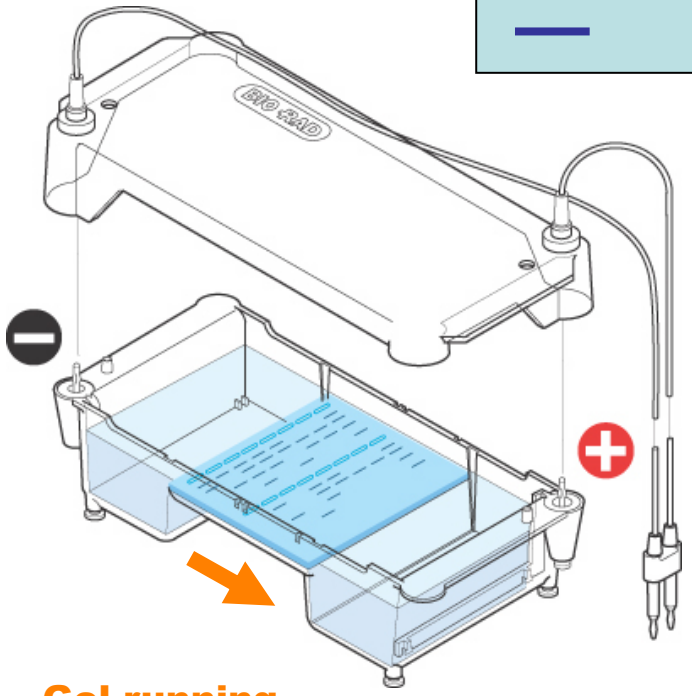
Wuyts J et al. Nucl. Acids Res. 2001;29:5017-5028

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

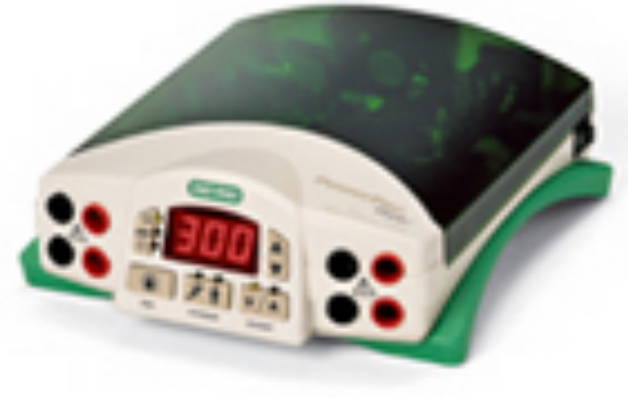
2b. PCR errors & prevention/correction

PCR Demo





Gel running



BioRad: A brief history of PCR

DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

2b. PCR biases & prevention/correction

- Prevention/correction:

❖mispriming -> stringency ↑: temperature/hybridization conditions & *

❖chimeras -> minimization of PCR cycles, high fidelity polymerases, relieving agents & *

❖amplification of homologous non-target genes (e.g. for 16S rRNA gene, mitochondrial or chloroplast homologue) -> selective extraction and selective environment sampling & *

*overall sequencing depth outweighs the errors (the higher the depth the more the sequences we can spare)

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

2b. PCR biases (mispriming)

Dealing with mispriming in data analysis:

- Contrasting each amplicon sequence with curated databases.

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

2b. PCR errors & prevention/correction

- Prevention/correction:

❖mispriming -> temperature/hybridization conditions & *

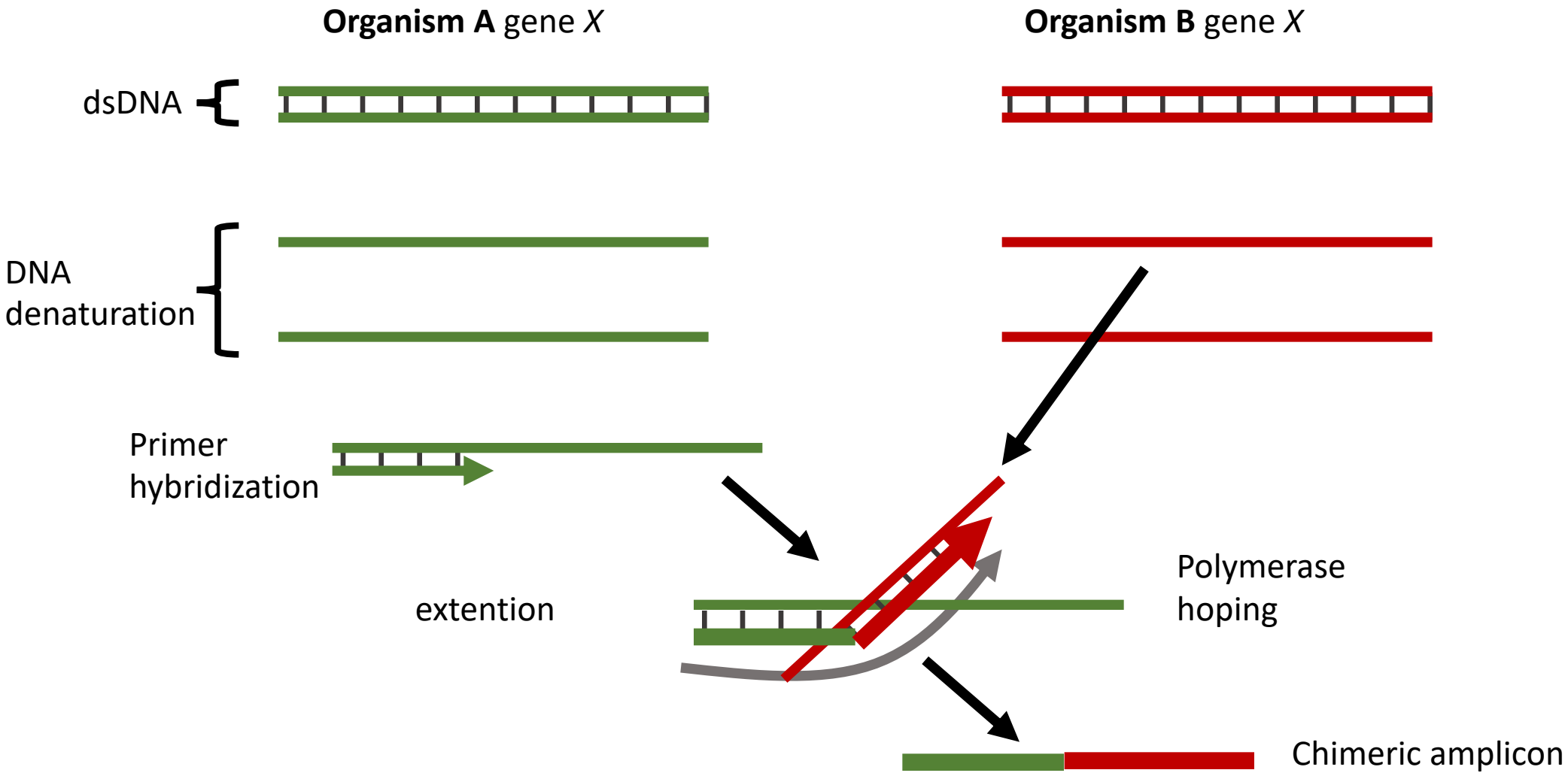
❖chimeras -> minimization of PCR cycles, high fidelity polymerases, relieving agents & *

❖amplification of homologous non-target genes (e.g. for 16S rRNA gene, mitochondrial or chloroplast homologue) -> selective extraction and selective environment sampling & *

*overall sequencing depth outweighs the errors (the higher the depth the more the sequences we can spare)

DEPARTMENT OF
Biochemistry &
Biotechnology
UNIVERSITY OF THESSALY

2b. PCR errors (Chimeras)



DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

2b. PCR errors & prevention/correction

- Prevention/correction:

❖mispriming -> temperature/hybridization conditions & *

❖chimeras -> minimization of PCR cycles, high fidelity polymerases, relieving agents & *

❖amplification of homologous non-target genes (e.g. for 16S rRNA gene, mitochondrial or chloroplast homologue) -> selective extraction and selective environment sampling & *

*overall sequencing depth outweighs the errors (the higher the depth the more the sequences we can spare)

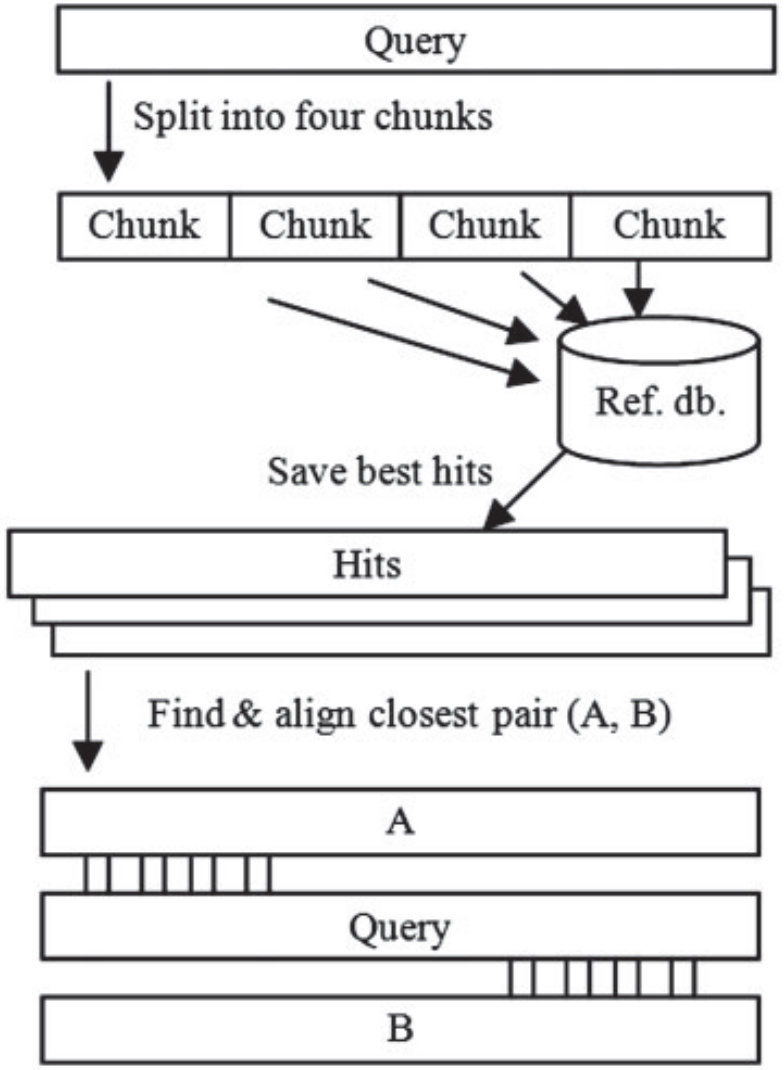
DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

2b. PCR errors (chimeras)

Dealing with chimeras in data analysis:

- Contrasting amplicon sequence portions with:
  - curated database
  - *de novo* built database using the exp. data (the most abundant sequences are the correct)

- A sequence containing portions of distant others in the database is chimeric



E.g. Uchime ----> Edgar et al. 2011, *Bioinformatics*, https://doi.org/10.1093/bioinformatics/btr381

DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

2b. PCR errors & prevention/correction

- Prevention/correction:

❖mispriming -> temperature/hybridization conditions & *

❖chimeras -> minimization of PCR cycles, high fidelity polymerases, relieving agents & *

❖amplification of homologous non-target genes (e.g. for 16S rRNA gene, mitochondrial or chloroplast homologue) -> selective extraction and selective environment sampling & *

*overall sequencing depth outweighs the errors (the higher the depth the more the sequences we can spare)

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

2b. PCR biases (same gene, off-target org.)

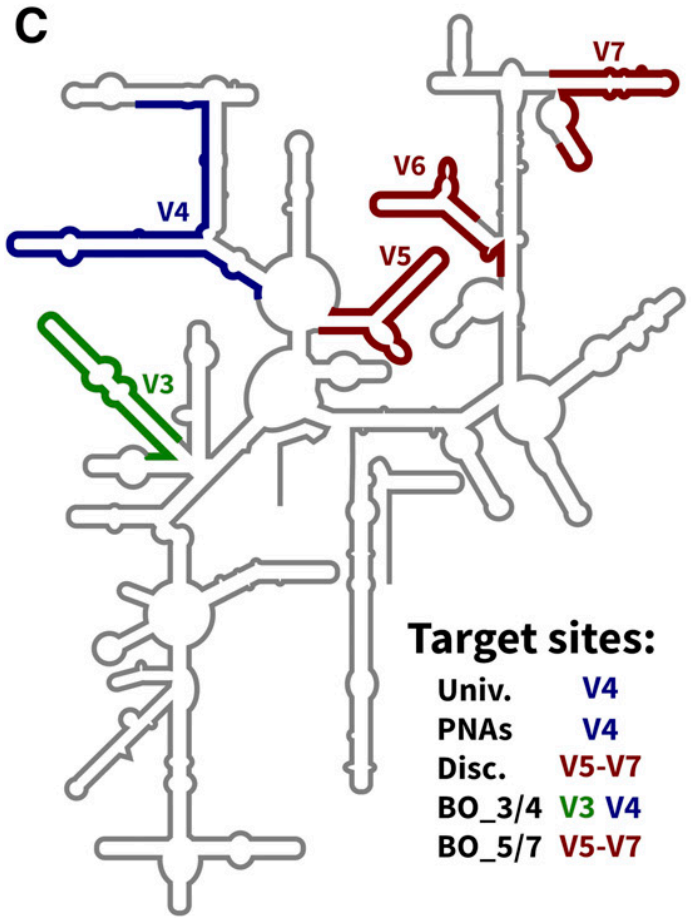
- Strategies for endophyte community analysis:

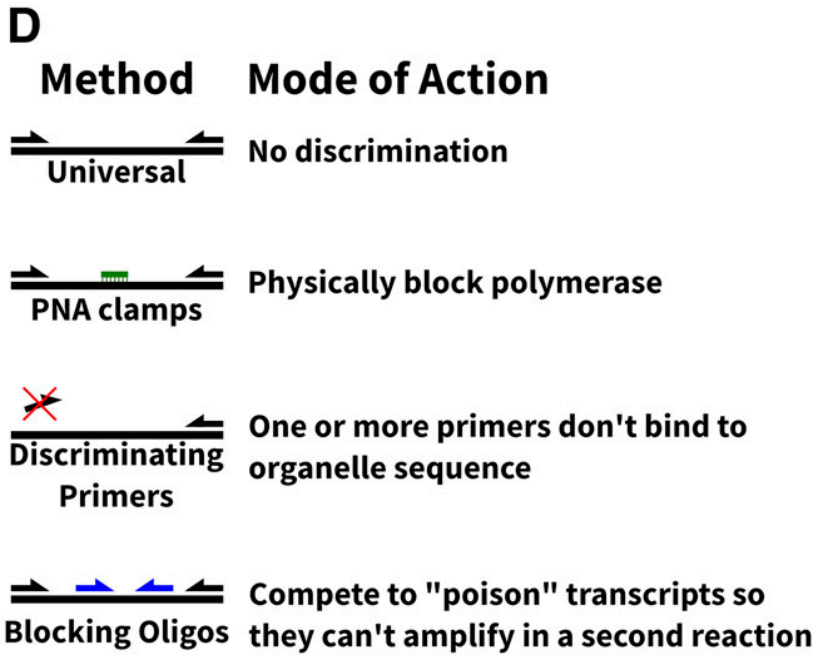
❖Be selective in the bioinformatics

❖Use peptide nucleic acid (PNA) or lock nucleic acid (LNA) blockers in the lab

❖Use selective primers in the lab

❖Use blocking oligos in the lab





Giangacomo, C., et al. (2021). Phytobiomes Journal, https://doi.org0.1094/pbiomes-07-20-0055-r

DEPARTMENT OF
Biochemistry &
Biotechnology
UNIVERSITY OF THESSALY

Microbial diversity with NGS-PMGA

3. Sequencing

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

The 3 generations of sequencing technologies

(size *vs* analyzed sequence variant numbers)

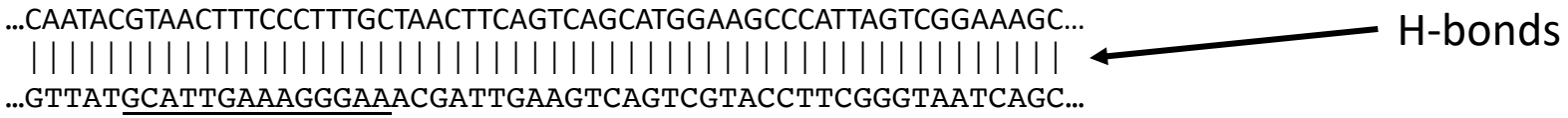
- 1st generation sequencing (single DNA fragment per reaction, ~700-1000 bp long)
  - Dideoxy termination method (**Sanger**)
- 2nd generation sequencing (massively parallel, max of ~300-600 bp total)
  - Sequencing by ligation (e.g. ABI-SOLiD)
  - Sequencing by synthesis (e.g. *Pyrosequencing*, **Illumina**, *Ion-torrent*)
- 3rd generation sequencing (massively parallel and very long reads, max of ~ 10kbp-250kbp)
  - Sequencing by synthesis (e.g. single molecule real time – **PacBio**, **Nanopore**)
  - Synthetic long reads (e.g. Moleculo, 10X)

DEPARTMENT OF
Biochemistry &
Biotechnology
UNIVERSITY OF THESSALY

The 3 generations of sequencing technologies

All three generations have been proposed for PMGA sequencing

1st generation: Dideoxy termination (Sanger) method



Ingredients:

1) dsDNA

https://en.wikipedia.org/wiki/Sanger_sequencing

1st generation: Dideoxy termination (Sanger) method

…CAATACGTAACTTTCCCTTTGCTAACTTCAGTCAGCATGGAAGCCCATTAGTCGGAAAGC…

…GTTATGCATTGAAAGGGAAACGATTGAAGTCAGTCGTACCTTCGGGTAATCAGC…

Ingredients:

1) dsDNA

Process:

1) Denature dsDNA (95˚C)

https://en.wikipedia.org/wiki/Sanger_sequencing

1st generation: Dideoxy termination (Sanger) method

…GTTATGCATTGAAAGGGAAACGATTGAAGTCAGTCGTACCTTCGGGTAATCAGC…

Ingredients:
1) dsDNA

Process:
1) Denature dsDNA (95˚C)

https://en.wikipedia.org/wiki/Sanger_sequencing

1st generation: Dideoxy termination (Sanger) method

```
    5'CGTAACTTTCCCTT
      ||||||||||||||
...GTTATGCATTGAAAGGGAAACGATTGAAGTCAGTCGTACCTTCGGGTAATCAGC...
```

Ingredients:

1) dsDNA
2) Primer

Process:

1) Denature dsDNA (95˚C)
2) Primer hybridization (e.g. 50 ˚C)

https://en.wikipedia.org/wiki/Sanger_sequencing

1st generation: Dideoxy termination (Sanger) method

Ingredients:

1) dsDNA
2) Primer
3) Polymerase

Process:

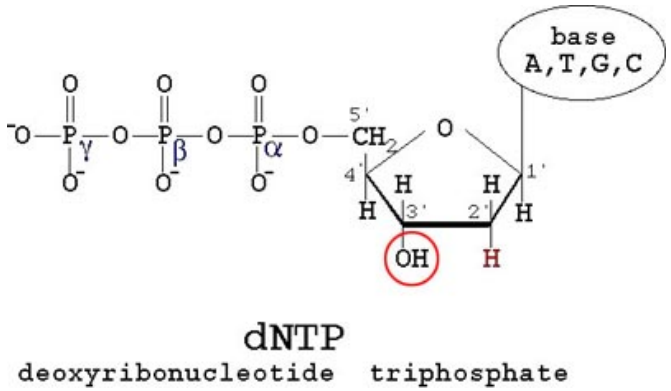
1) Denature dsDNA (95˚C)
2) Primer hybridization (e.g. 50 ˚C)

https://en.wikipedia.org/wiki/Sanger_sequencing

1st generation: Dideoxy termination (Sanger) method

Ingredients:

1) dsDNA
2) Primer
3) Polymerase
4) dNTPs

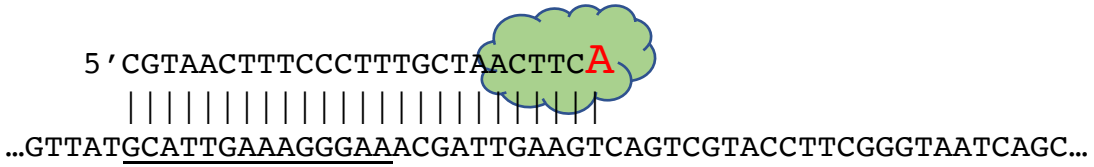


Process:

1) Denature dsDNA (95˚C)
2) Primer hybridization (e.g. 50 ˚C)
3) Extension/elongation (e.g. 72 ˚C)

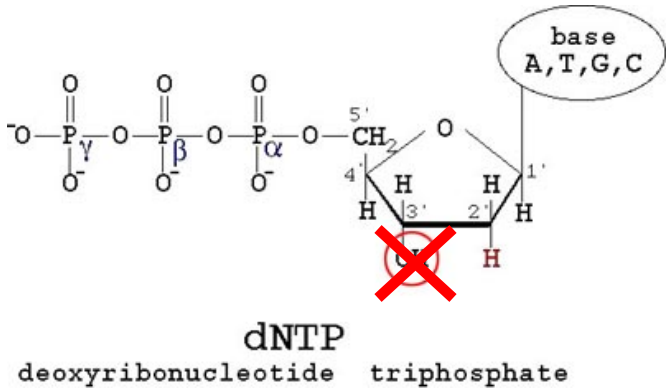
https://en.wikipedia.org/wiki/Sanger_sequencing

1st generation: Dideoxy termination (Sanger) method



Ingredients:

1) dsDNA
2) Primer
3) Polymerase
4) dNTPs
5) ddNTPs (e.g. ddATP)



Process:

1) Denature dsDNA (95˚C)
2) Primer hybridization (e.g. 50 ˚C)
3) Extension/elongation (e.g. 72 ˚C)
4) Termination

https://en.wikipedia.org/wiki/Sanger_sequencing

1st generation: Dideoxy termination (Sanger) method

5′CGTAACTTTCCCTTTGCTAACTTCA

…GTTATGCATTGAAAGGGAAACGATTGAAGTCAGTCGTACCTTCGGGTAATCAGC…

Ingredients:

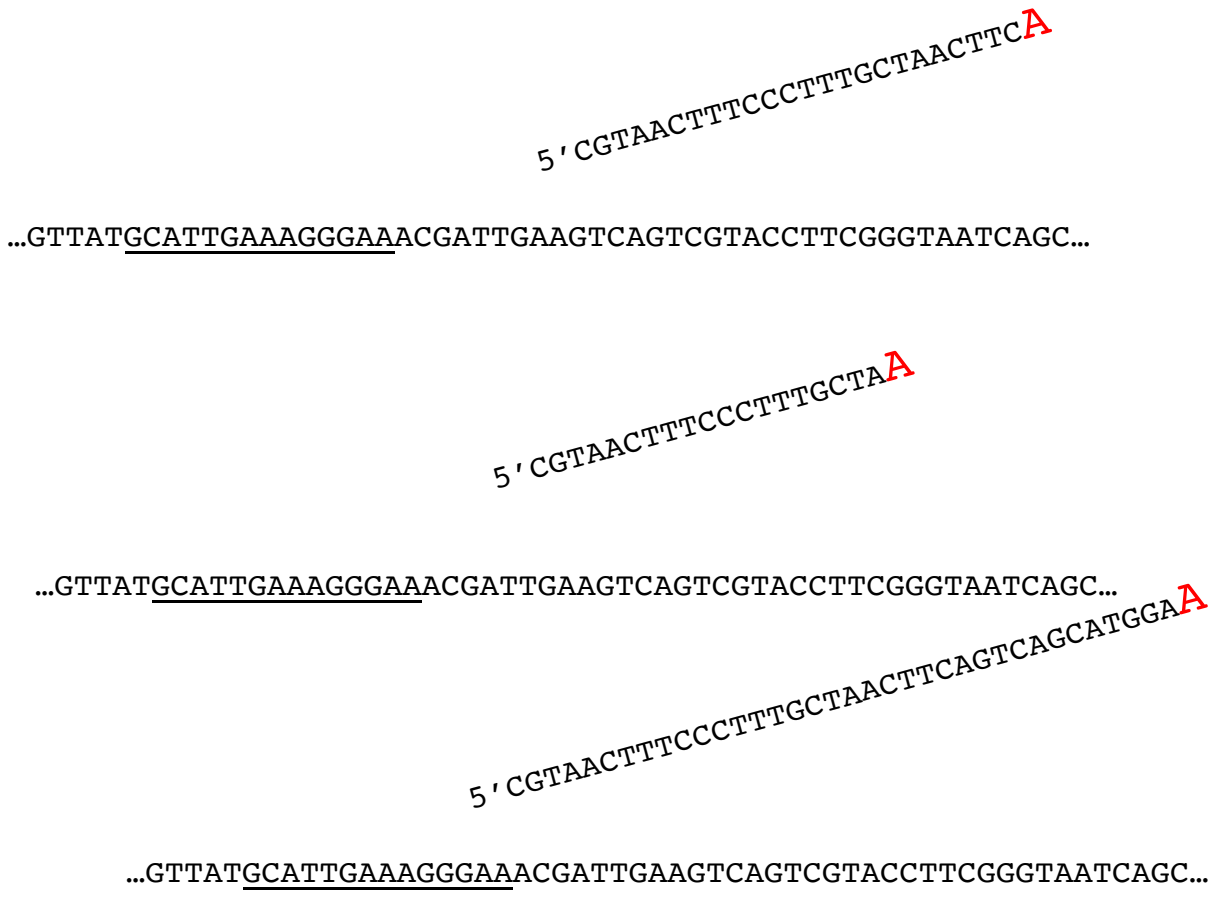
1) dsDNA
2) Primer
3) Polymerase
4) dNTPs
5) ddNTPs (e.g. ddATP)

Process:

1) Denature dsDNA (95˚C)
2) Primer hybridization (e.g. 50 ˚C)
3) Extension/elongation (e.g. 72 ˚C)
4) Termination
5) Denature dsDNA (95˚C)

https://en.wikipedia.org/wiki/Sanger_sequencing

1st generation: Dideoxy termination (Sanger) method

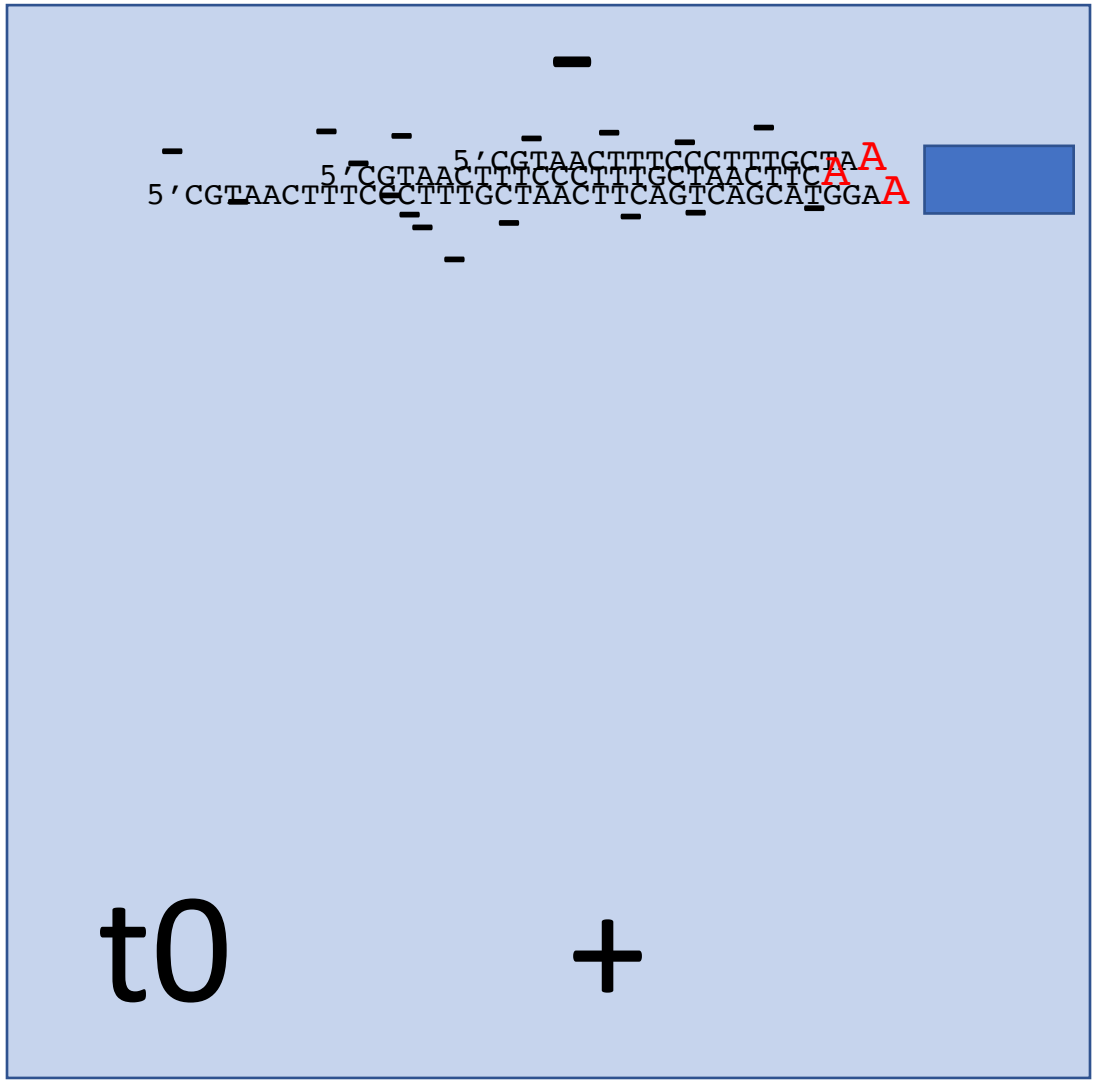


https://en.wikipedia.org/wiki/Sanger_sequencing

1st generation: Dideoxy termination (Sanger) method

Diagnostics process:

1) Add fragments to a matrix that separates them according to size (e.g. polyacrylamide gel)
2) Apply current (negatively charged DNA moves toward positive electrode)
3) Identify terminator base according to location on gel



https://en.wikipedia.org/wiki/Sanger_sequencing

1st generation: Dideoxy termination (Sanger) method

Diagnostics process:

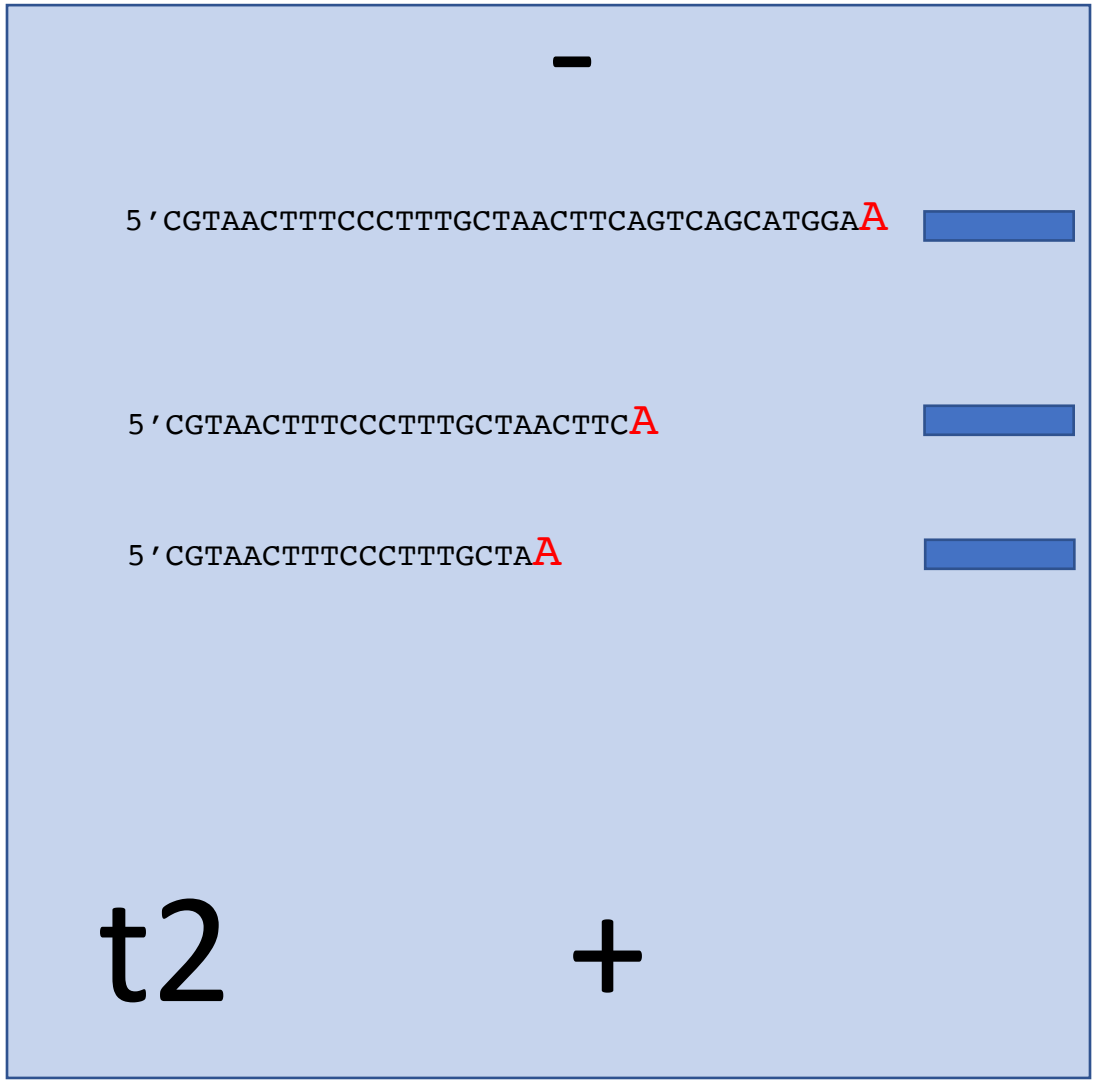
1) Add fragments to a matrix that separates them according to size (e.g. polyacrylamide gel)
2) Apply current (negatively charged DNA moves toward positive electrode)
3) Identify terminator base according to location on gel

https://en.wikipedia.org/wiki/Sanger_sequencing

1st generation: Dideoxy termination (Sanger) method

Diagnostics process:

1) Add fragments to a matrix that separates them according to size (e.g. polyacrylamide gel)
2) Apply current (negatively charged DNA moves toward positive electrode)
3) Identify terminator base according to location on gel

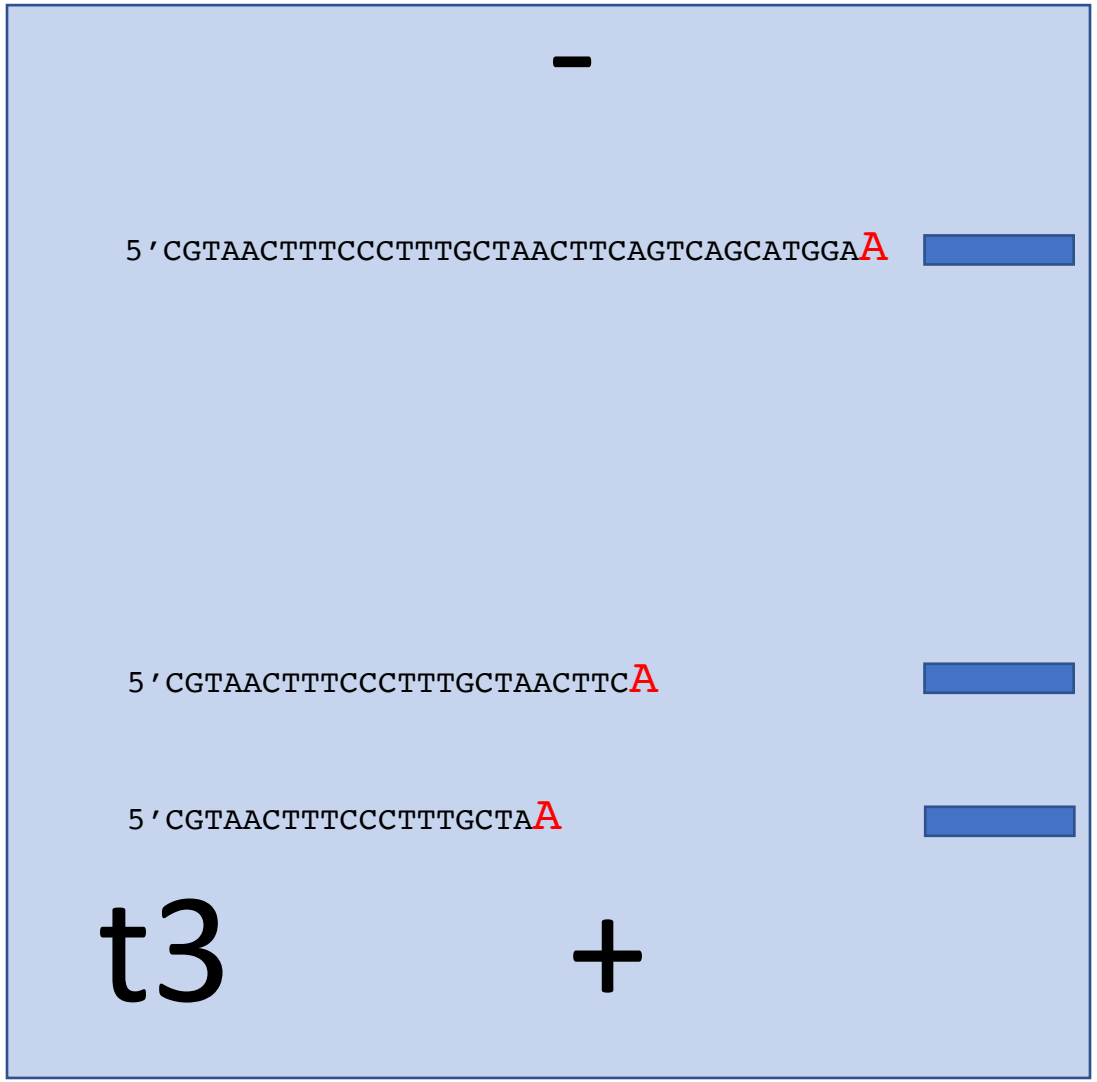


https://en.wikipedia.org/wiki/Sanger_sequencing

1st generation: Dideoxy termination (Sanger) method

Diagnostics process:

1) Add fragments to a matrix that separates them according to size (e.g. polyacrylamide gel)
2) Apply current (negatively charged DNA moves toward positive electrode)
3) Identify terminator base according to location on gel



https://en.wikipedia.org/wiki/Sanger_sequencing

1st generation: Dideoxy termination (Sanger) method

Diagnostics process:

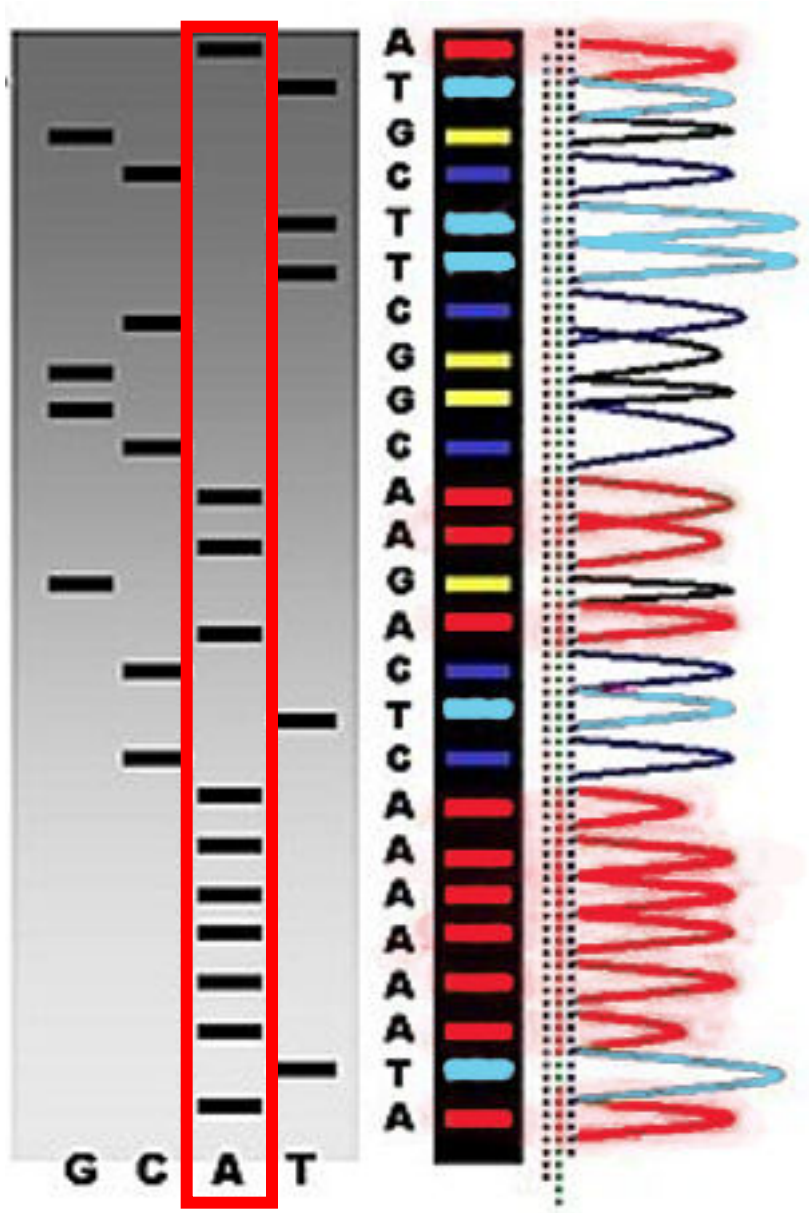
1) Add fragments to a matrix that separates them according to size (e.g. polyacrylamide gel)
2) Apply current (negatively charged DNA moves toward positive electrode)
3) Identify terminator base according to location on gel

https://en.wikipedia.org/wiki/Sanger_sequencing

1st generation: Dideoxy termination (Sanger) method

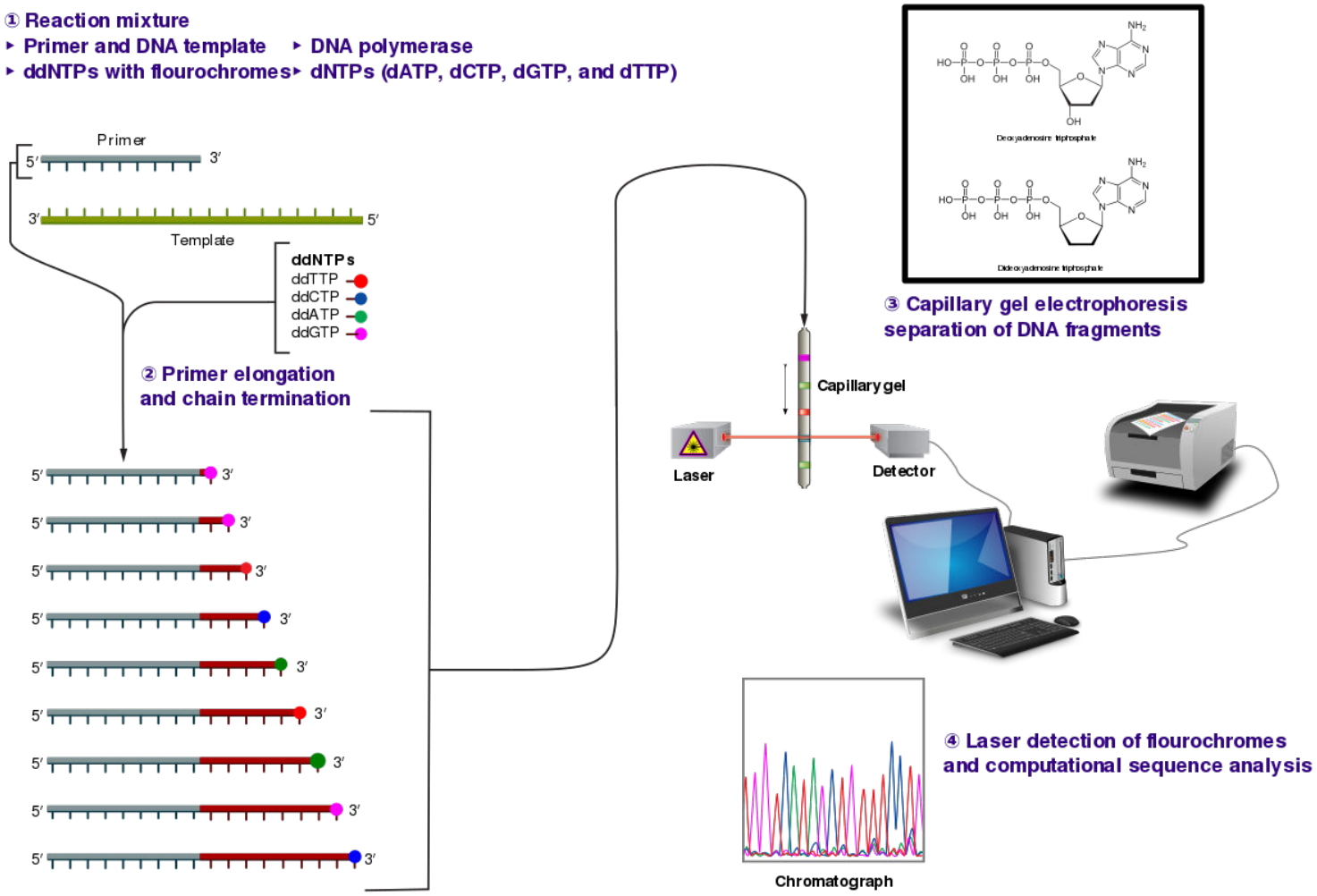
Diagnostics process:

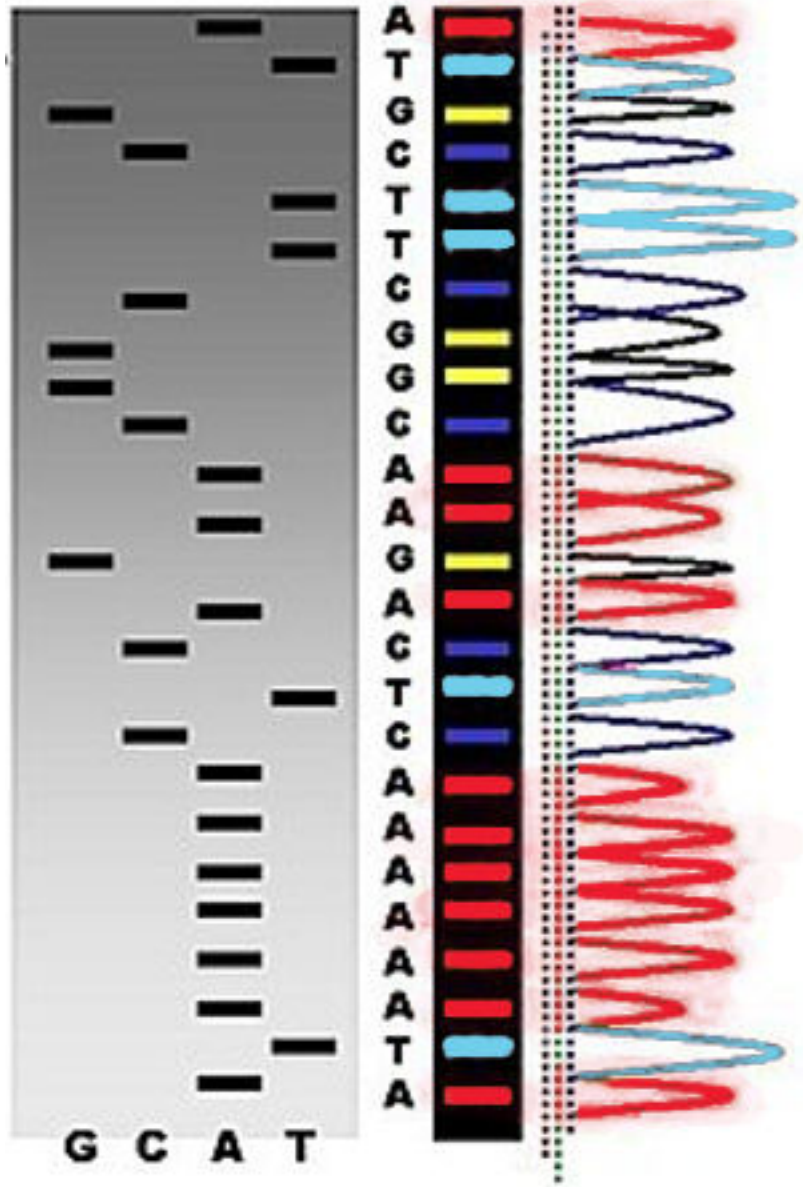
1) Add fragments to a matrix that separates them according to size (e.g. polyacrylamide gel)
2) Apply current (negatively charged DNA moves toward positive electrode)
3) Identify terminator base according to location on gel



https://en.wikipedia.org/wiki/Sanger_sequencing

1st generation: Dideoxy termination (Sanger) method (Summary slide)

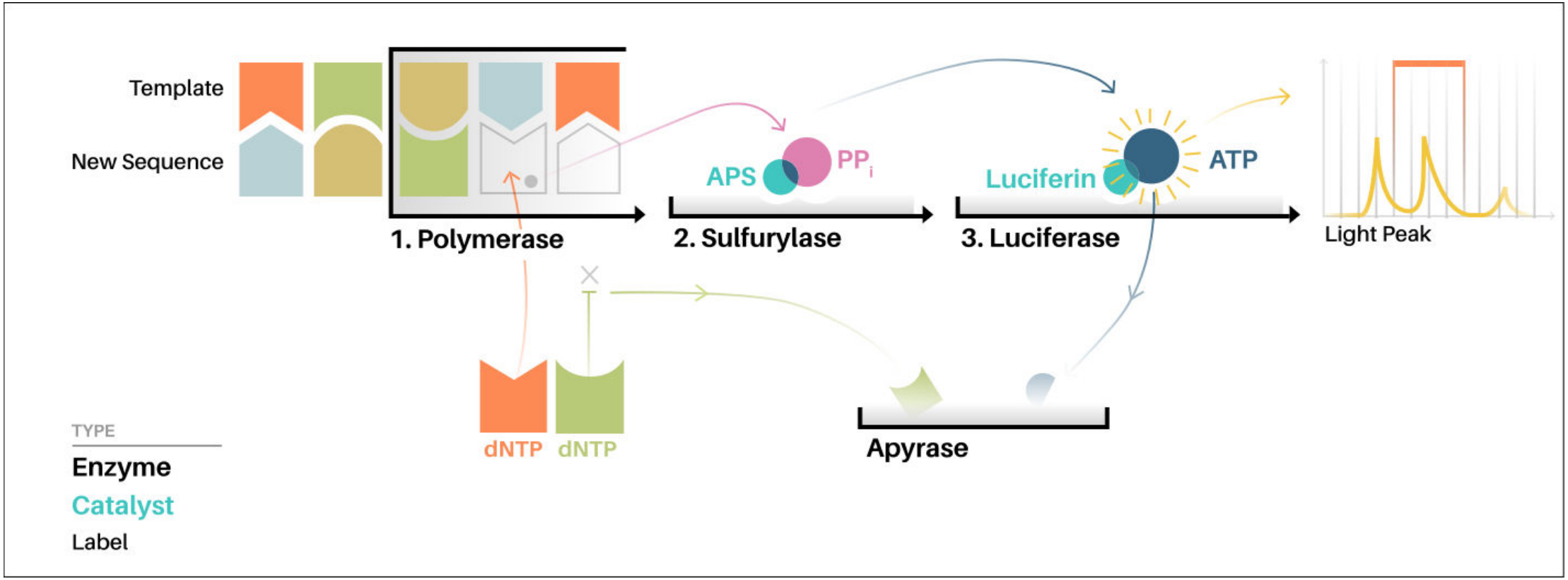




https://en.wikipedia.org/wiki/Sanger_sequencing

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

3. 2nd generation Pyrosequencing

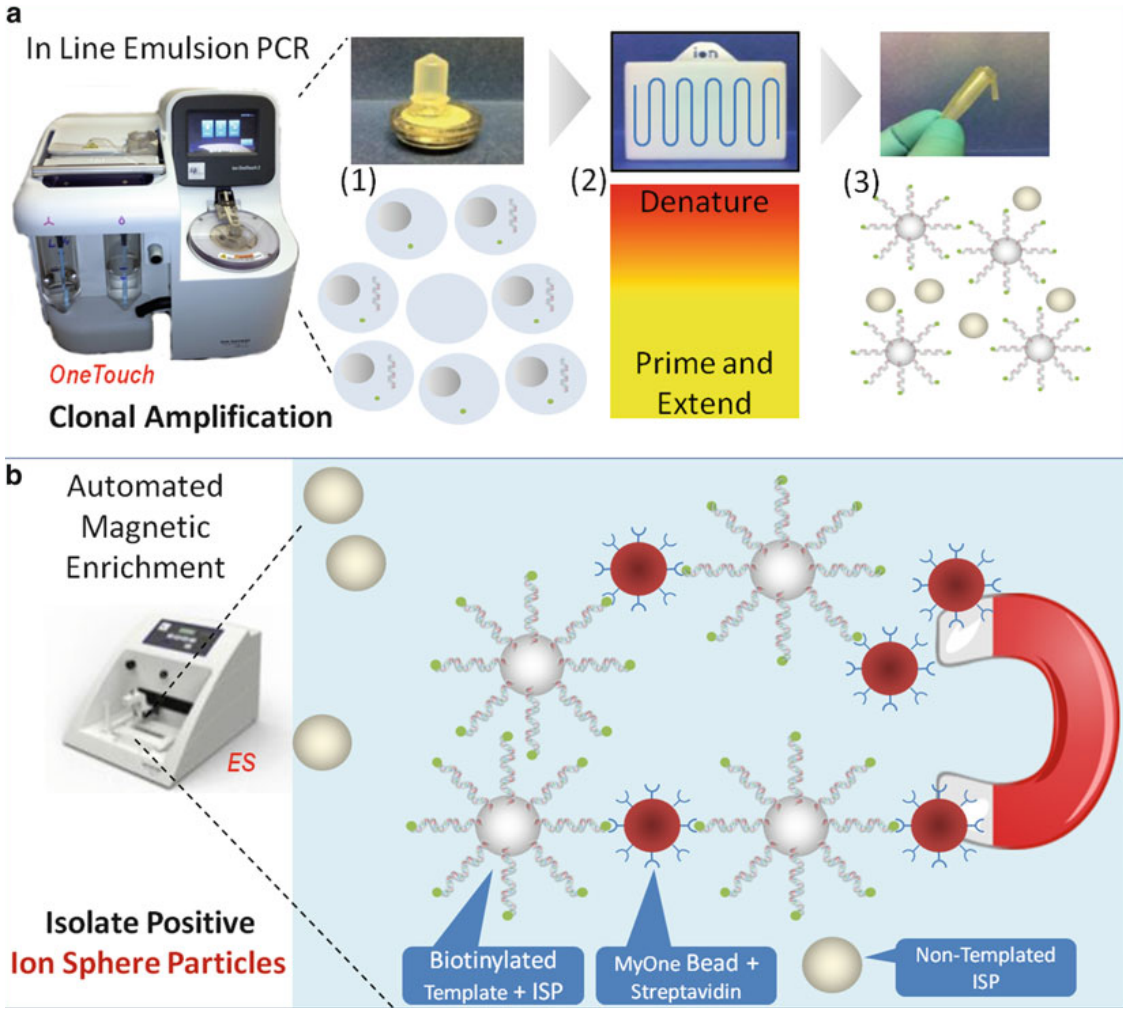


https://en.wikipedia.org/wiki/Pyrosequencing
https://www.youtube.com/watch?v=KzdWZ5ryBlA

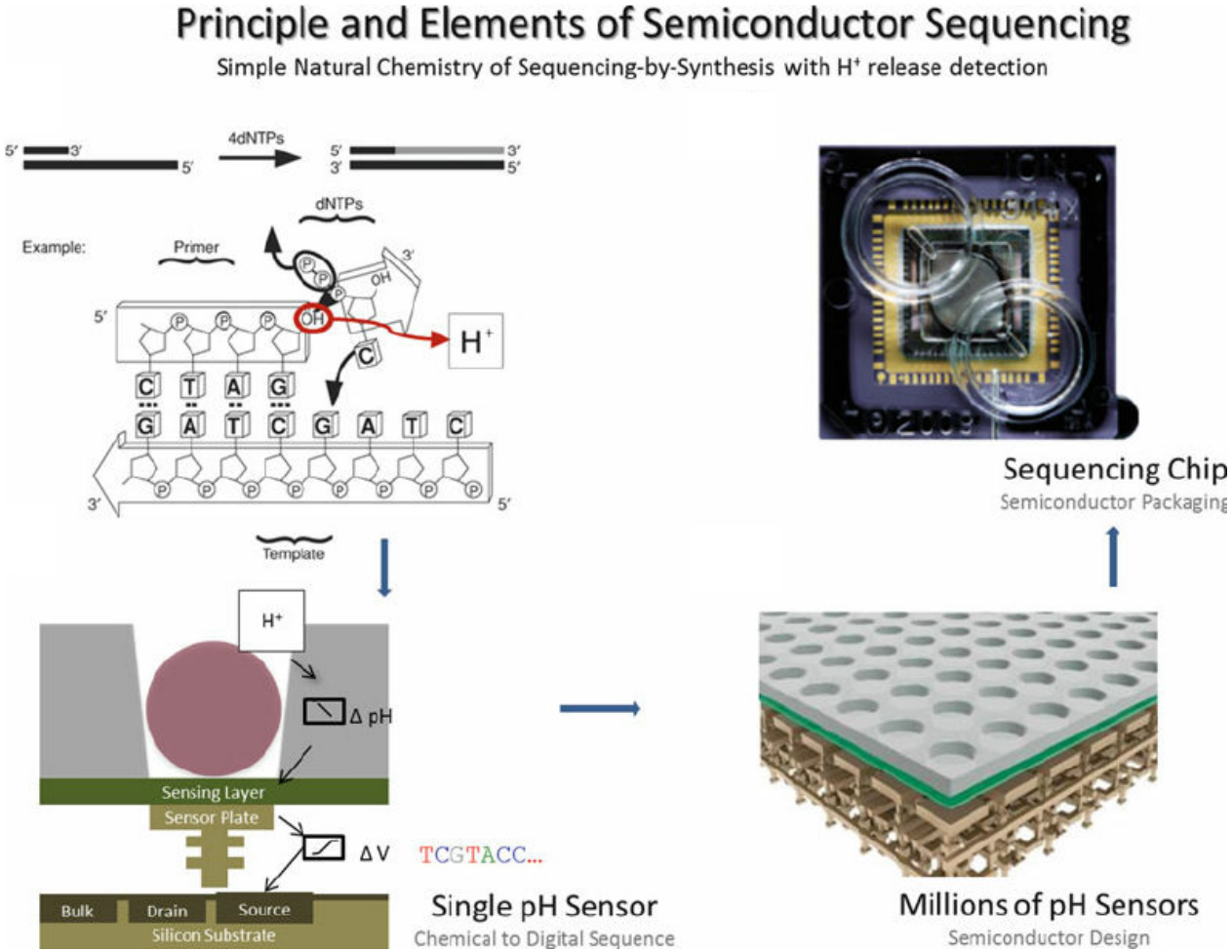
DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

3. 2nd generation Ion Torrent/Proton

Sample prep



Sequencing

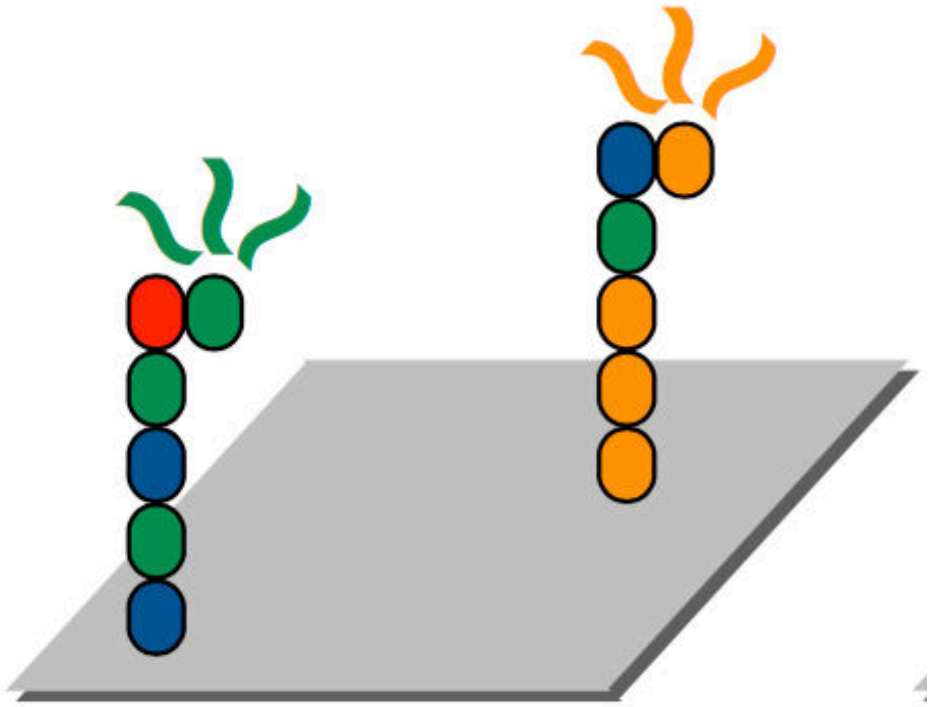


Kohn, A.B., *et al.* (2013). Single-Cell Semiconductor Sequencing. Methods in molecular biology (Clifton, NJ) 1048, 247-284 (sightly modified)

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

3. 2nd generation Illumina (sequencing)

Besides bases, polymerase and primers

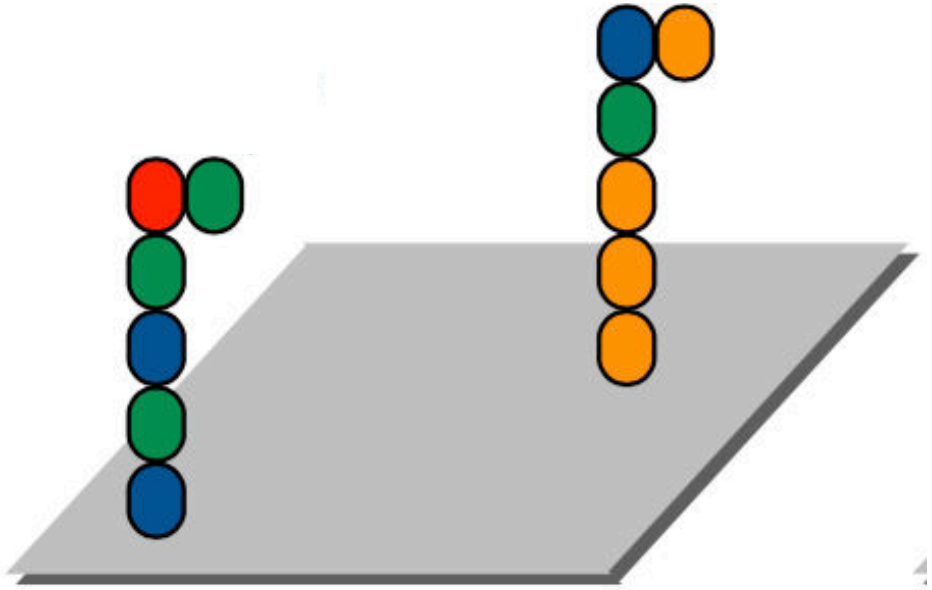


http://data-science-sequencing.github.io/Win2018/lectures/lecture2/#illumina

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

3. 2nd generation Illumina (sequencing)

Important ingredients are

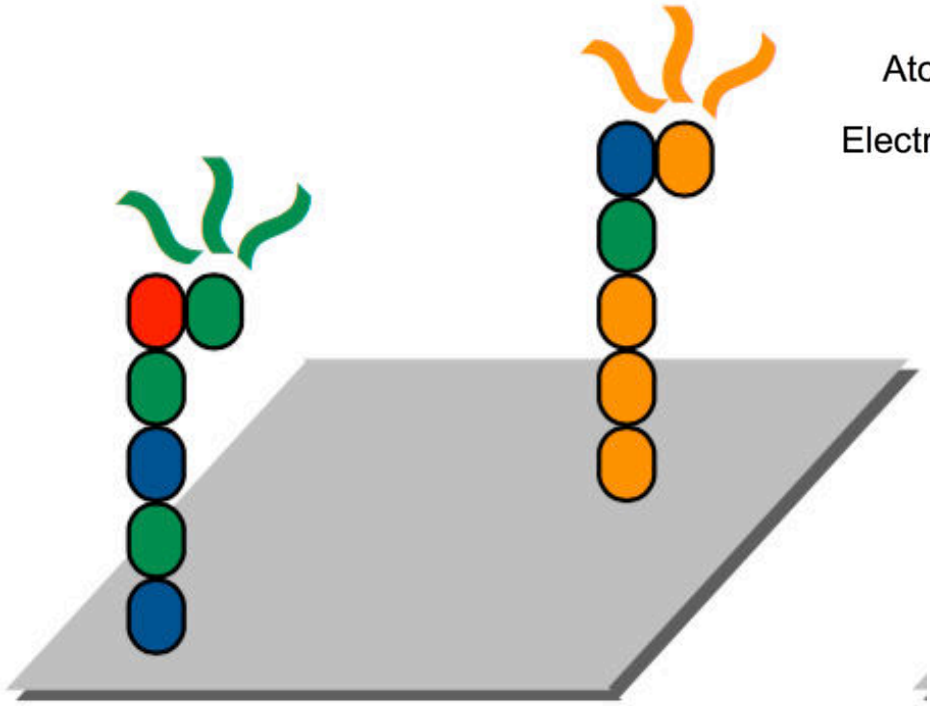


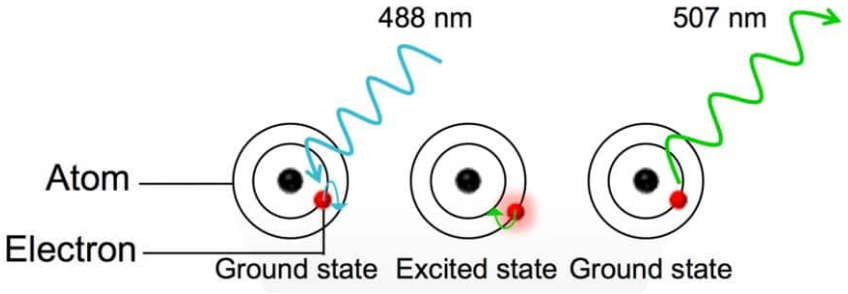
http://data-science-sequencing.github.io/Win2018/lectures/lecture2/#illumina

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

3. 2nd generation Illumina (sequencing)

Important ingredients are the fluorophore



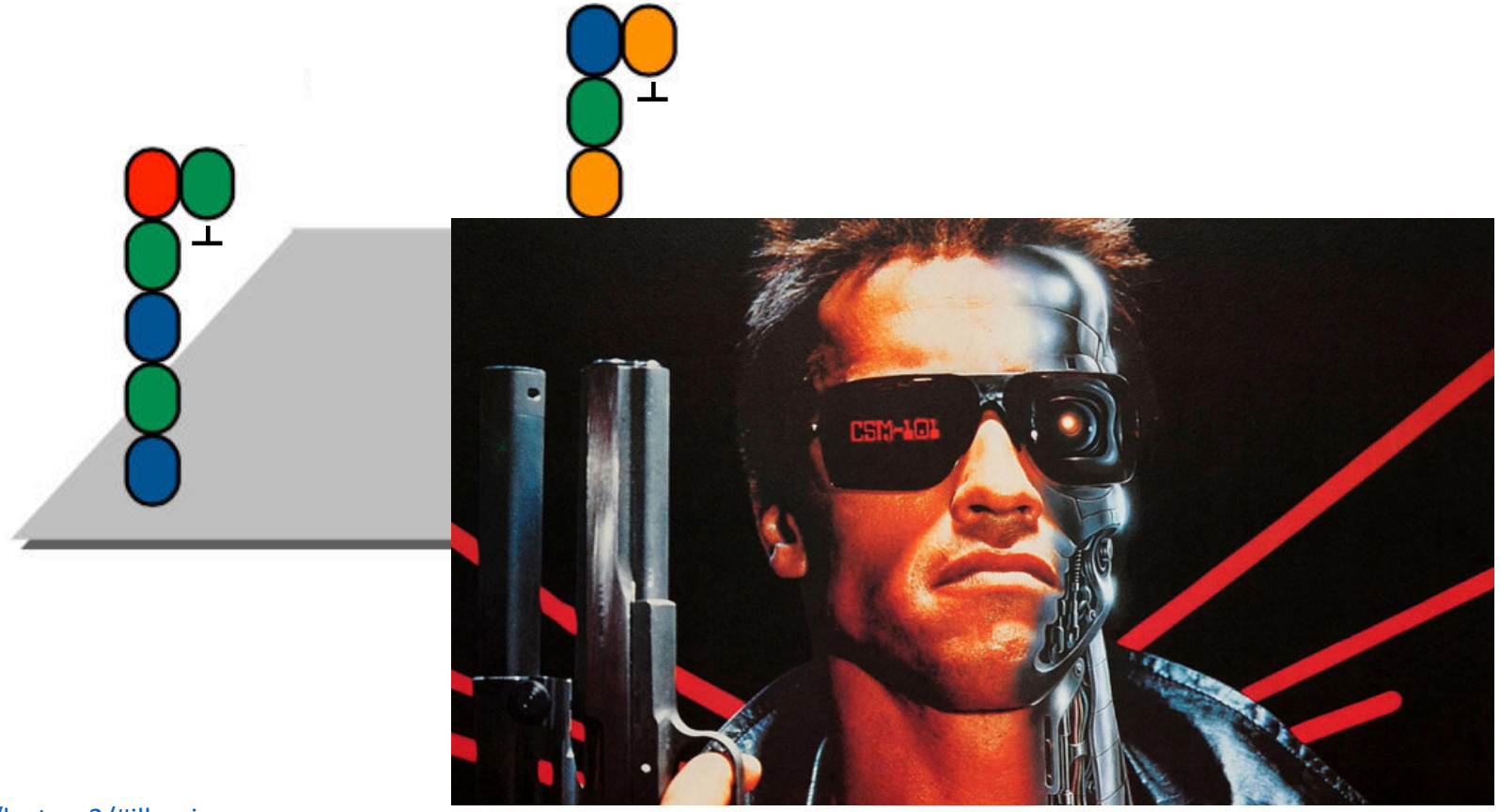


http://data-science-sequencing.github.io/Win2018/lectures/lecture2/#illumina

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

3. 2nd generation Illumina (sequencing)

Important ingredients are the fluorophore & terminator!



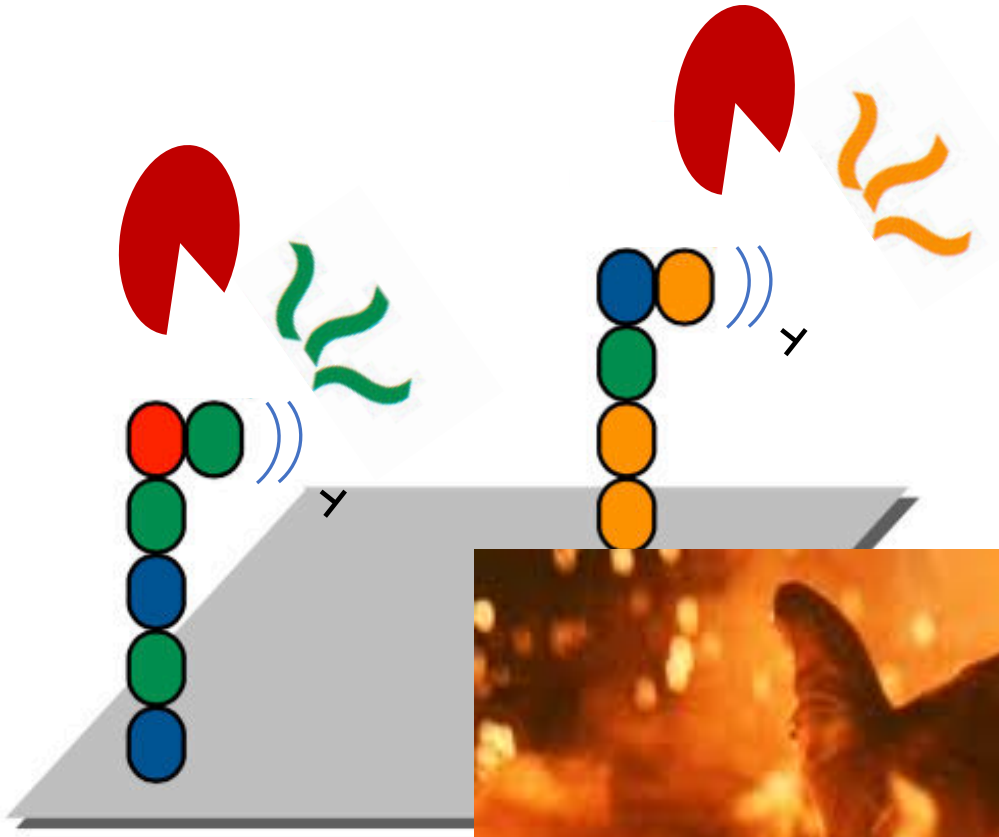
http://data-science-sequencing.github.io/Win2018/lectures/lecture2/#illumina

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

3. 2nd generation Illumina (sequencing)

Important ingredients are the fluorophore & terminator!

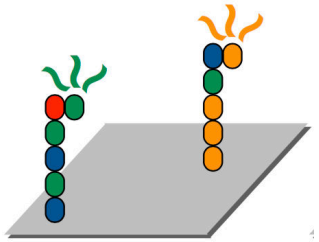
… a cleaving enzyme …
AND then, the terminator is…
NO MORE!



http://data-science-sequencing.github.io/Win2018/lectures/lecture2/#illumina

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

3. 2nd generation Illumina (sequencing)

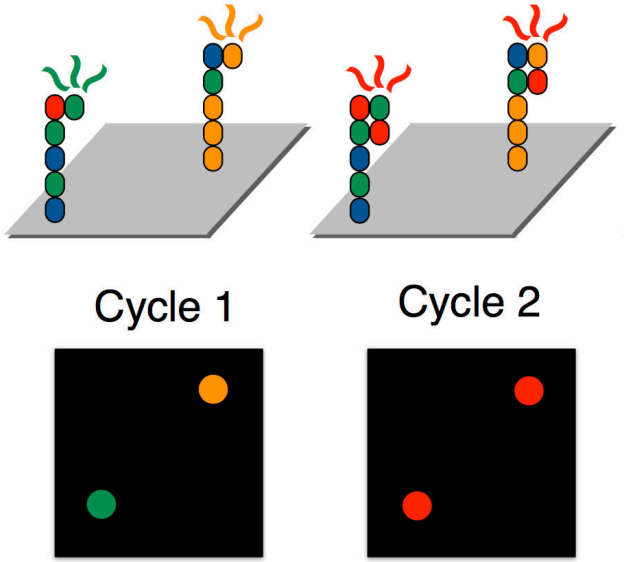


Cycle 1

http://data-science-sequencing.github.io/Win2018/lectures/lecture2/#illumina
http://tinyurl.com/6zbzh4

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

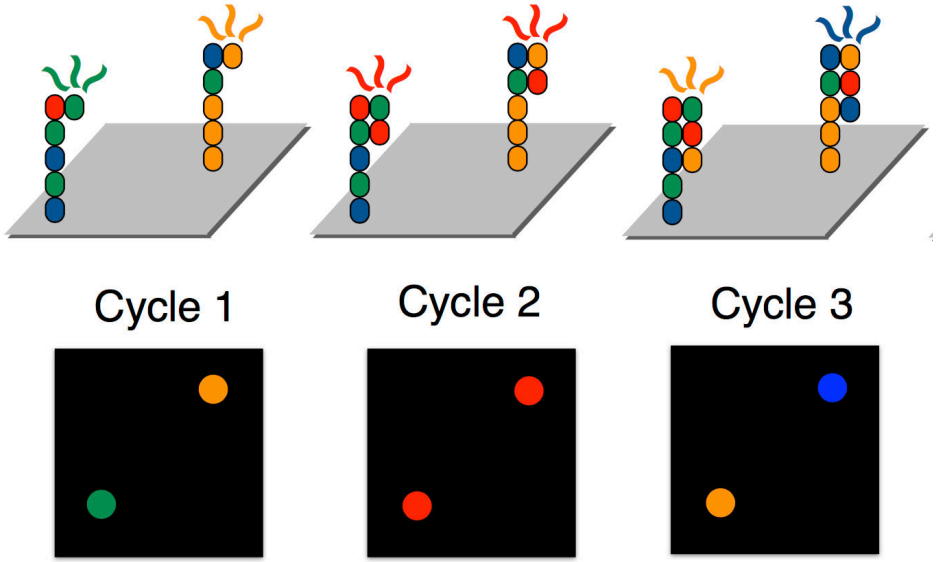
3. 2nd generation Illumina (sequencing)



http://data-science-sequencing.github.io/Win2018/lectures/lecture2/#illumina
http://tinyurl.com/6zbzh4

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

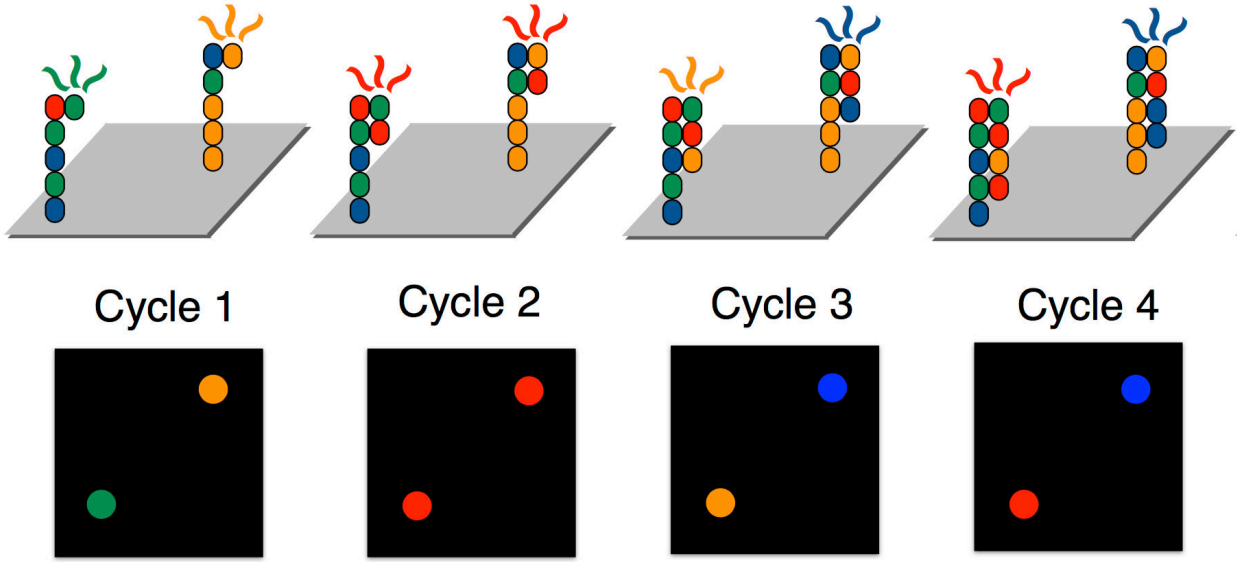
3. 2nd generation Illumina (sequencing)



http://data-science-sequencing.github.io/Win2018/lectures/lecture2/#illumina
http://tinyurl.com/6zbzh4

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

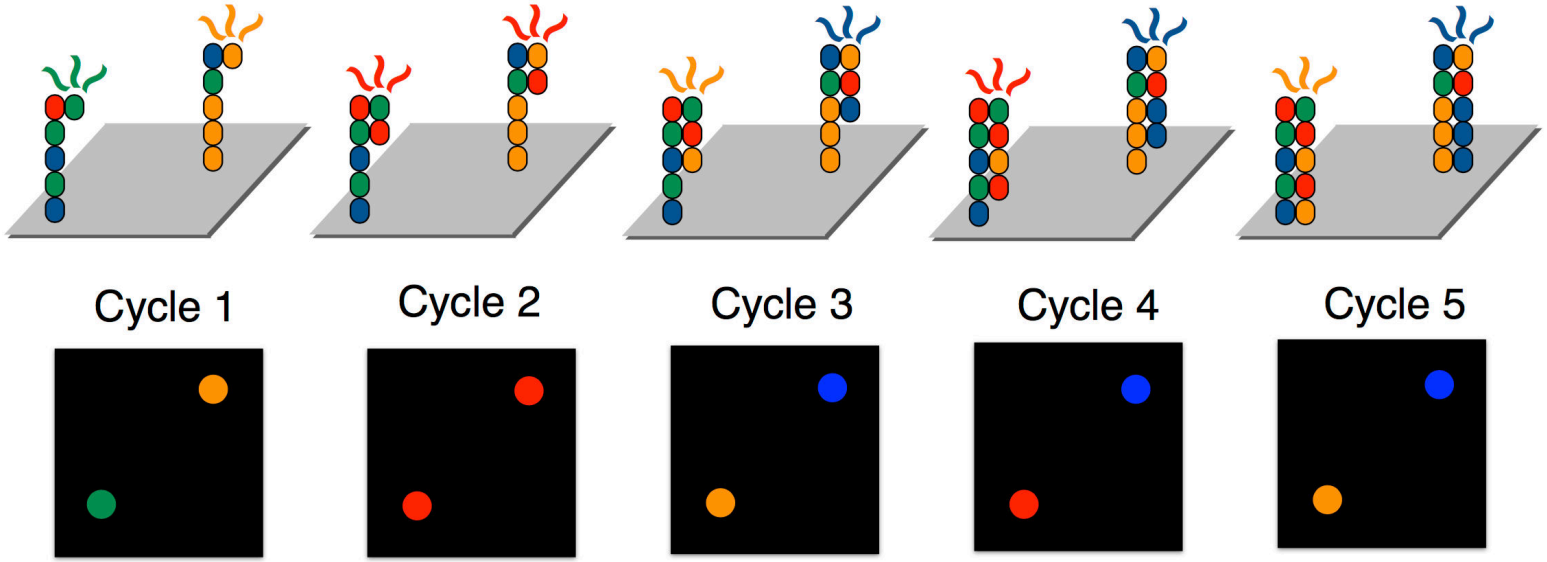
3. 2nd generation Illumina (sequencing)

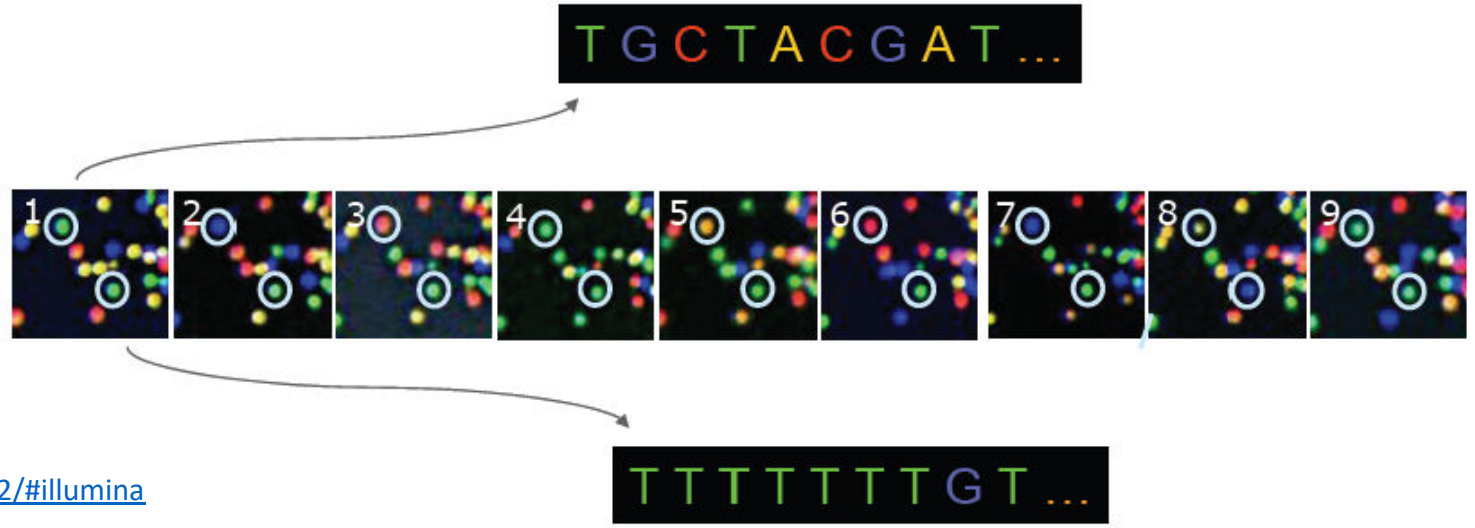


http://data-science-sequencing.github.io/Win2018/lectures/lecture2/#illumina
http://tinyurl.com/6zbzh4

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

3. 2nd generation Illumina (sequencing)



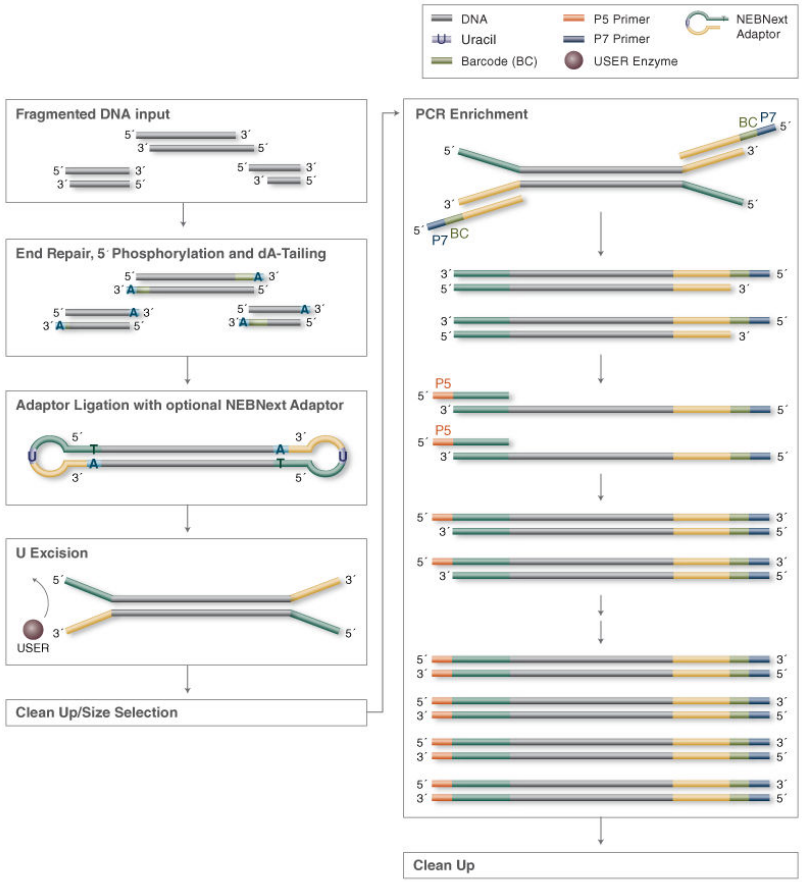


http://data-science-sequencing.github.io/Win2018/lectures/lecture2/#illumina
http://tinyurl.com/6zbzh4

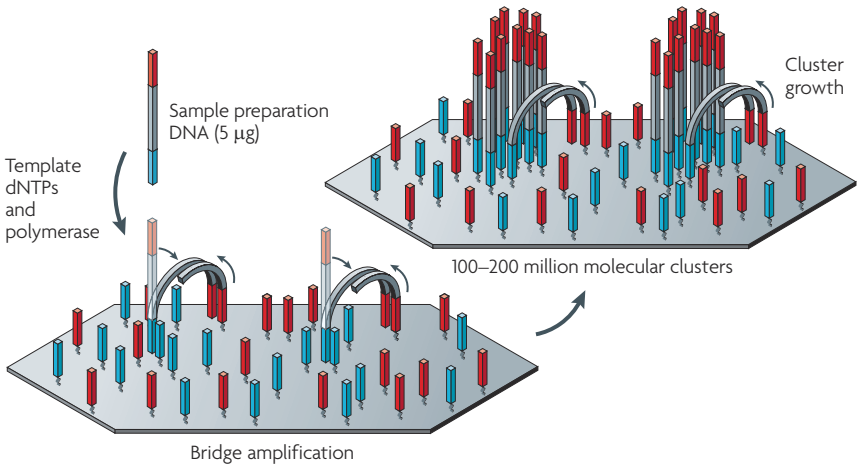
DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

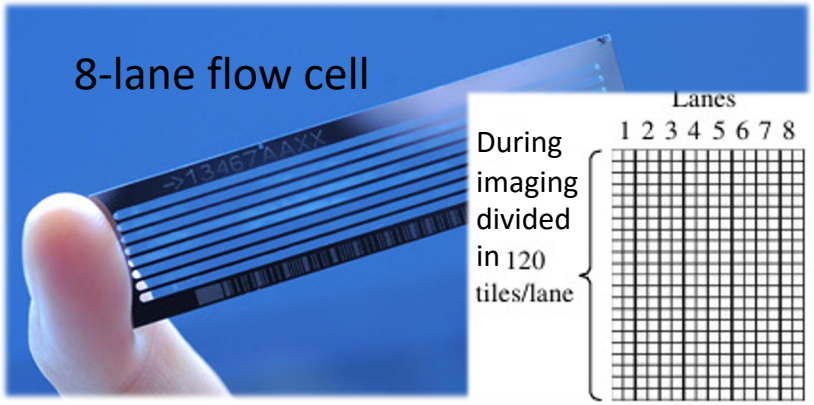
3. 2nd generation Illumina

Library prep

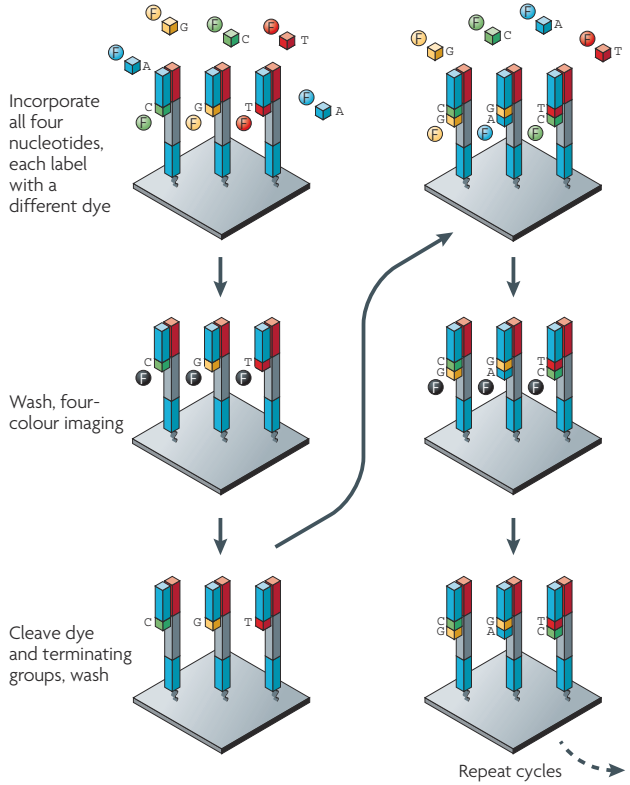


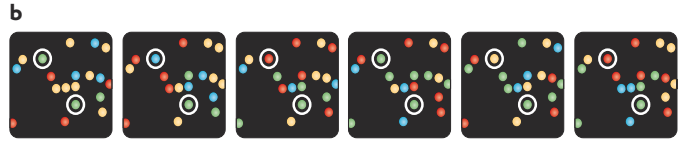
Flow cell attachment and bridge amplification

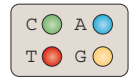




Sequencing data generation







Top: CATCGT
Bottom: CCCCCC

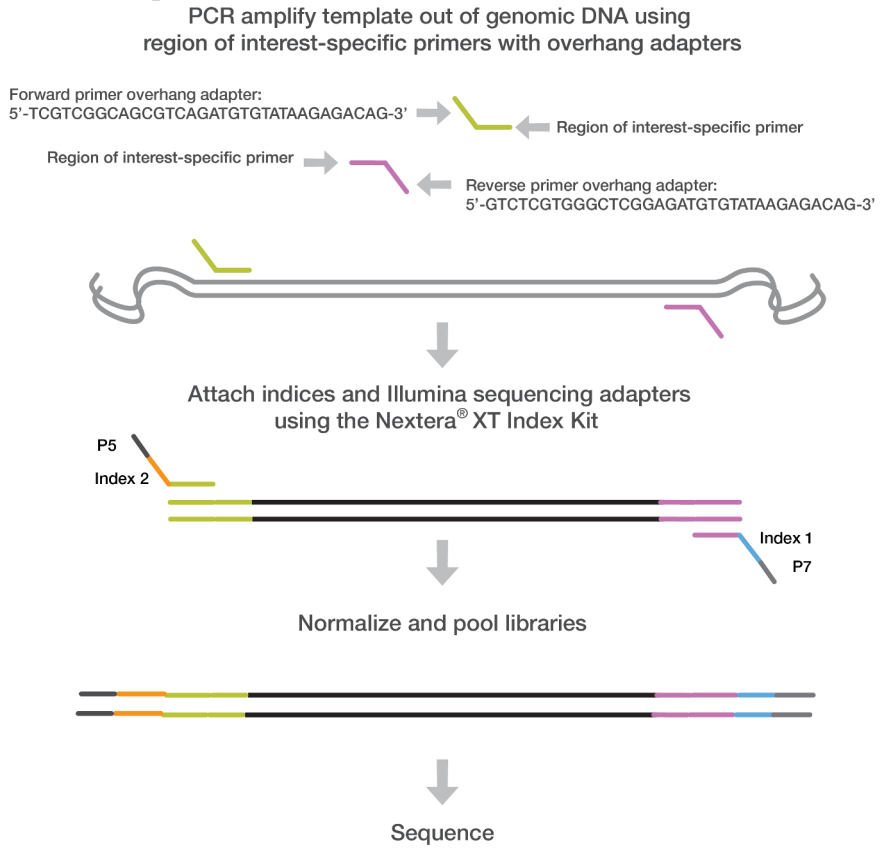
https://www.youtube.com/watch?v=fCd6B5HRaZ8
https://www.neb.com/applications/library-preparation-for-next-generation-sequencing/illumina-library-preparation

Kawashima, E., *et al.* (1998). Method of nucleic acid amplification, Google Patents
Metzker, M.L. (2009). Sequencing technologies — the next generation. Nat Rev Genet 11, 31 (modified)

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

3. 2nd generation Illumina (library prep)

Figure 1 16S V3 and V4 Amplicon Workflow

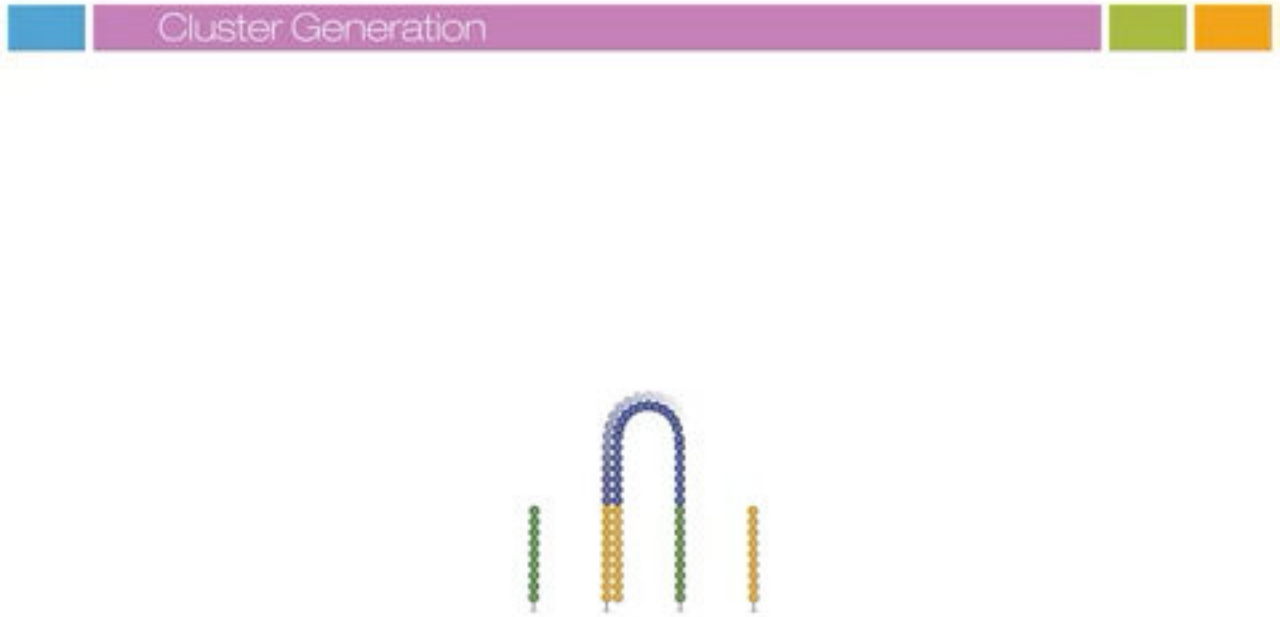


User-defined forward and reverse primers that are complementary upstream and downstream of the region of interest are designed with overhang adapters, and used to amplify templates from genomic DNA. A subsequent limited-cycle amplification step is performed to add multiplexing indices and Illumina sequencing adapters. Libraries are normalized and pooled, and sequenced on the MiSeq system using v3 reagents.

Illumina documentation for amplicon sequencing library prep

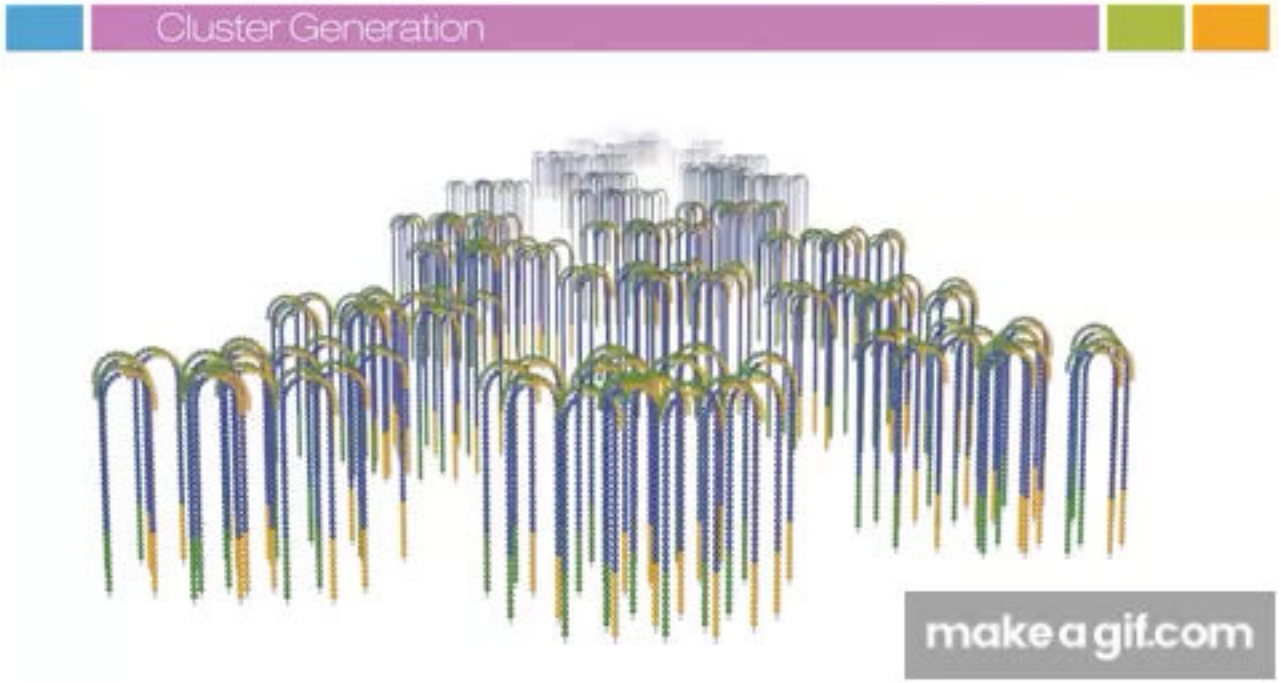
DEPARTMENT OF
Biochemistry &
Biotechnology
UNIVERSITY OF THESSALY

3. 2nd generation Illumina (cluster prep)



DEPARTMENT OF
Biochemistry &
Biotechnology
UNIVERSITY OF THESSALY

3. 2nd generation Illumina (cluster prep)

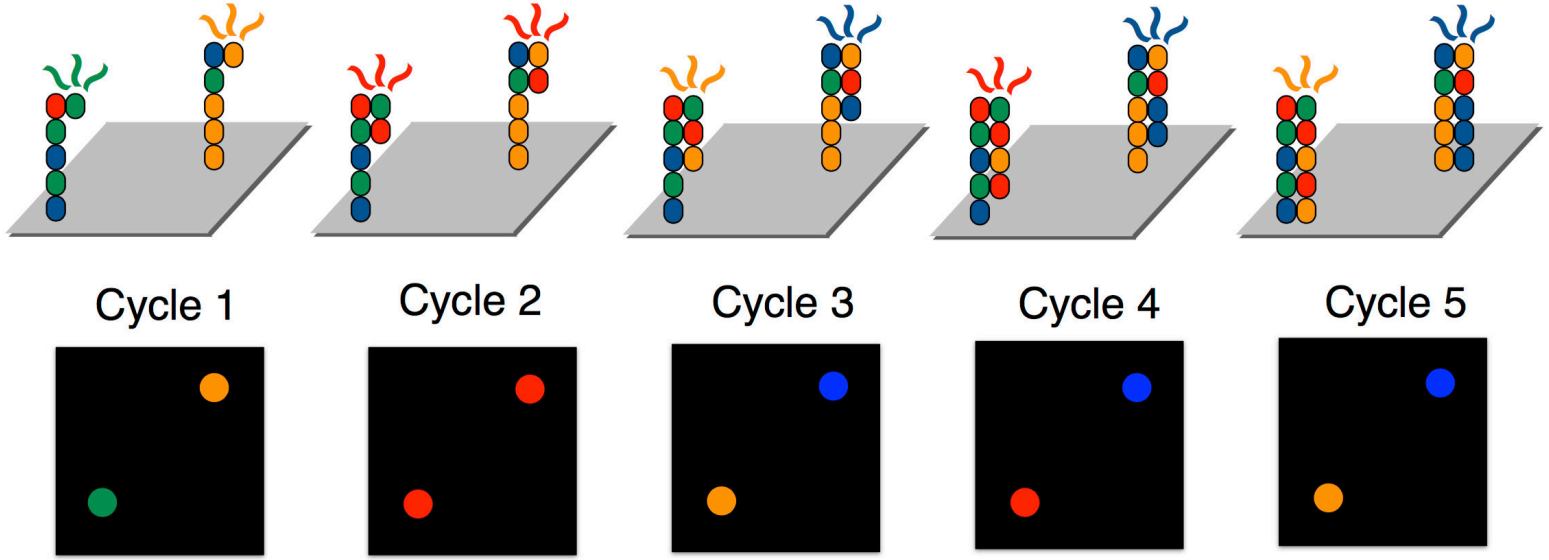


DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

3. 2nd generation Illumina (sequencing)

Basic ingredients:

1) Single stranded template DNA attached to a slide (flow-cell lane)
2) Four nucleic acid bases with reversible terminators and cleavable dyes (each having its own colour: e.g. Adenine emits at yellow when excited, G at blue, T at green, C at red)
3) DNA polymerase



Process:

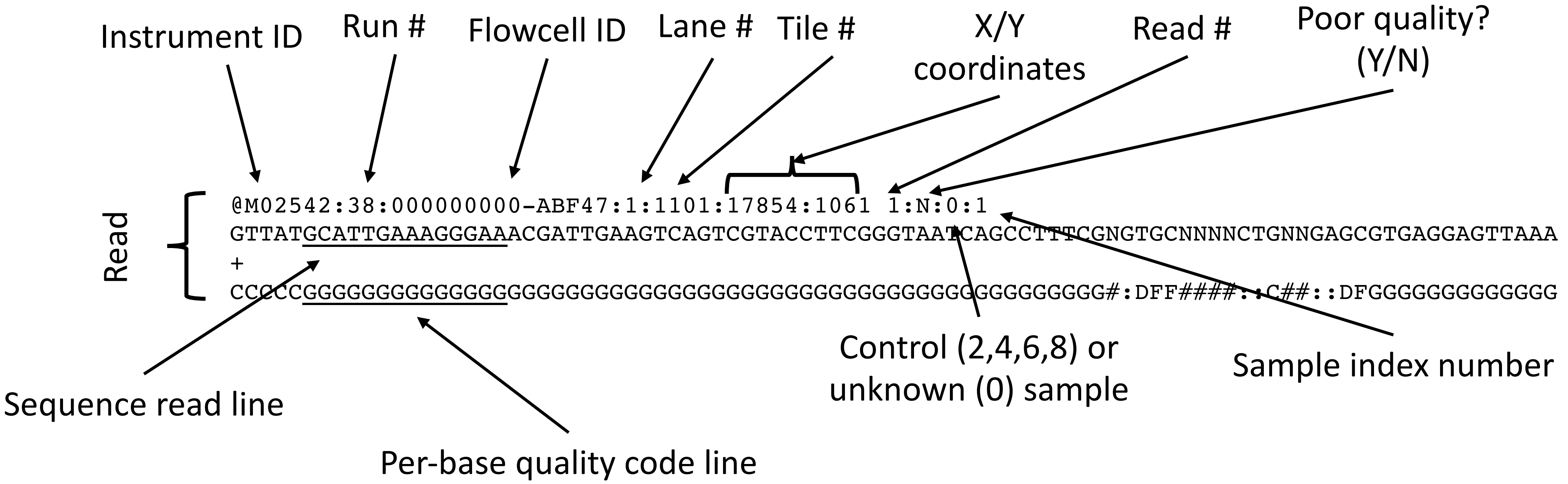
1) The 4 bases are added together with the polymerase on the slide with the attached template ssDNA fragments
2) A base with its dye and terminator is added in a complementary fashion against each ssDNA (5' -> 3') to synthesize dsDNA and the reaction is terminated
3) The dye is excited and an image is acquired
4) The dye and the terminator are enzymatically removed
5) The cycle is repeated (go to step 1)

http://data-science-sequencing.github.io/Win2018/lectures/lecture2/#illumina

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

3. 2nd generation Illumina (sequencing)

3. Illumina reads output (the fastq format)




```
@M02542:38:000000000-ABF47:1:1101:17854:1061 1:N:0:1
GTTATGCATTGAAAGGGAAACGATTGAAGTCAGTCGTACCTTCGGGTAATCAGCCTTTCGNGTGCNNNNCTGNNGAGCGTGAGGAGTTAAA
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG#:DFF####::C##::DFGGGGGGGGGGGGG
@M02542:38:000000000-ABF47:1:1101:18832:1061 1:N:0:1
TACTGCGGTTGAAAGGGAAACGATTGAAGTCAGTCGTACCTGCGGGTAATCAGCCTTTTGNGTGTNNNNATGNNAGGCGTGAGGAGTTAAT
+
CCCCCGGGFGGFGGGGGGGGGGEGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG#:DGF####::D##::@FGGGGGGGGGGGG
@M02542:38:000000000-ABF47:1:1101:17854:1061 2:N:0:1
TACGTCAACATCCTTAACGNATTTANCNNGTNCTAACCTTTTGAACGNNANNTNTGATNACCGTANGCAAGCTTTCGGNACCAGAGCAACT
+
CCCCCGGGGGGGGGGGGG#=CFFG#:##::#:CDFGGGGGGGGGGG##:##:#::FG#:CFGFF#:CDGGGGGGGGE#:AFGGGFGGGGG
@M02542:38:000000000-ABF47:1:1101:18832:1061 2:N:0:1
TACGTCAACATCCTTAACGNATTTANCNNGTNCTAACCTATTGAACGNNANNTNTGATNCCATAANCAAGCTTTCAGCNCCAGAGCAACTT
+
CCCCCGGGGGGFFGGGGGG#=CFFF#:##::#:CFGFGGFGGEFGDF##:##:#::,C#:CFFFG#:DFGGGGGGGGG#:AFGCFGGDGGG
......
```

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

3. Error probability (Phred Q values)

$$Q = -10 \times \log_{10}(P_{err}) \quad <=> \quad P_{err} = 10^{Q/-10}$$

Representation of quality scores

Table 1 ASCII Characters Encoding Q-scores 0–40

Symbol	ASCII Code	Q-Score	Symbol	ASCII Code	Q-Score	Symbol	ASCII Code	Q-Score
!	33	0	/	47	14	=	61	28
"	34	1	0	48	15	>	62	29
#	35	2	1	49	16	?	63	30
$	36	3	2	50	17	@	64	31
%	37	4	3	51	18	A	65	32
&	38	5	4	52	19	B	66	33
'	39	6	5	53	20	C	67	34
(	40	7	6	54	21	D	68	35
)	41	8	7	55	22	E	69	36
*	42	9	8	56	23	F	70	37
+	43	10	9	57	24	G	71	38
,	44	11	:	58	25	H	72	39
-	45	12	;	59	26	I	73	40
.	46	13	<	60	27			

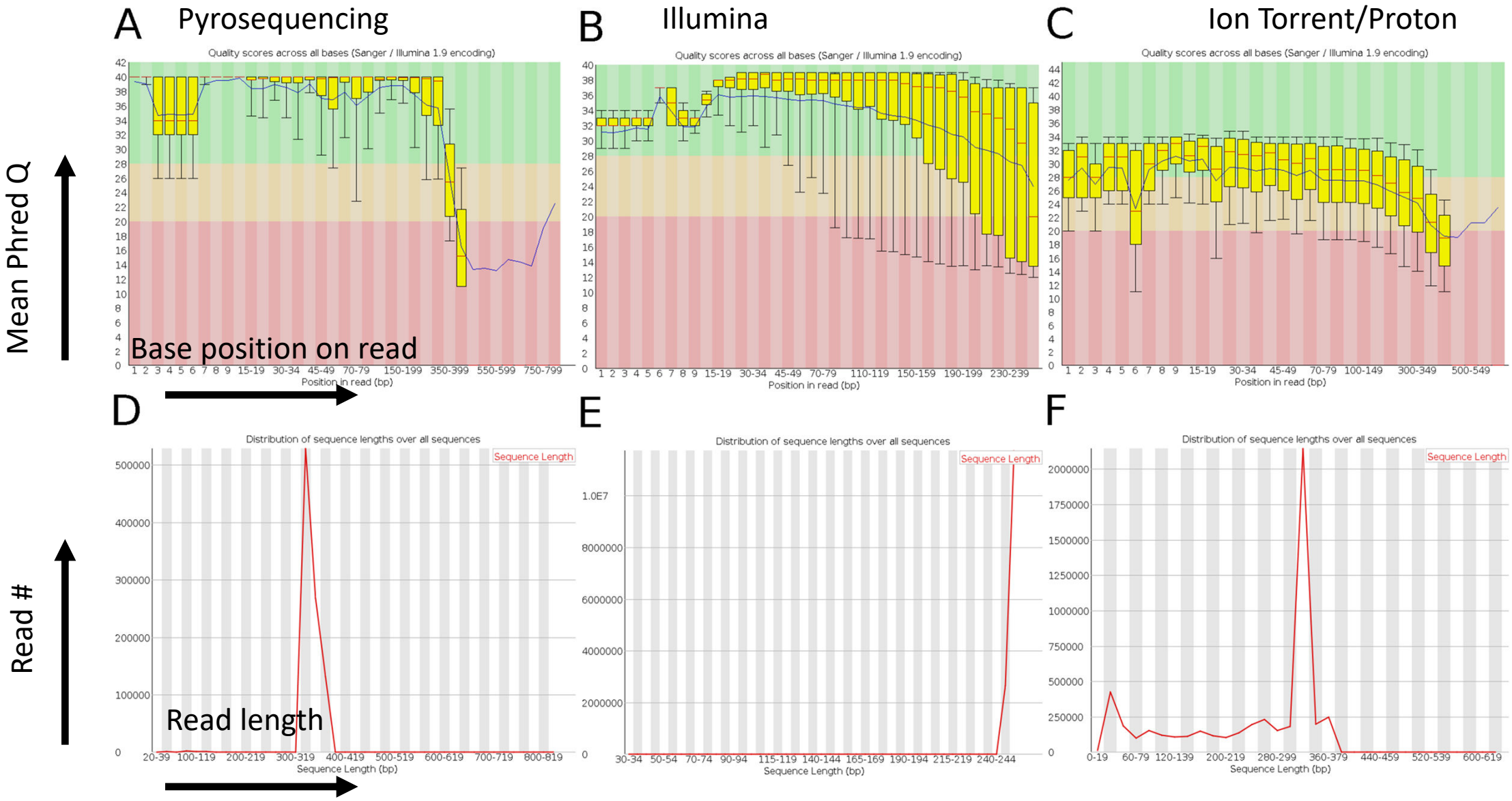
Phred quality scores are logarithmically linked to error probabilities

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%

https://en.wikipedia.org/wiki/Phred_quality_score
https://www.omixon.com/bioinformatics-for-beginners-file-formats-part-2-short-reads/

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

3. 2nd generation error probability



Fuellgrabe, M.W., et al. (2015). PLOS ONE 10, e0129195, doi: 10.1371/journal.pone.0129195

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

3. 2nd gen Inherent sequencing error types

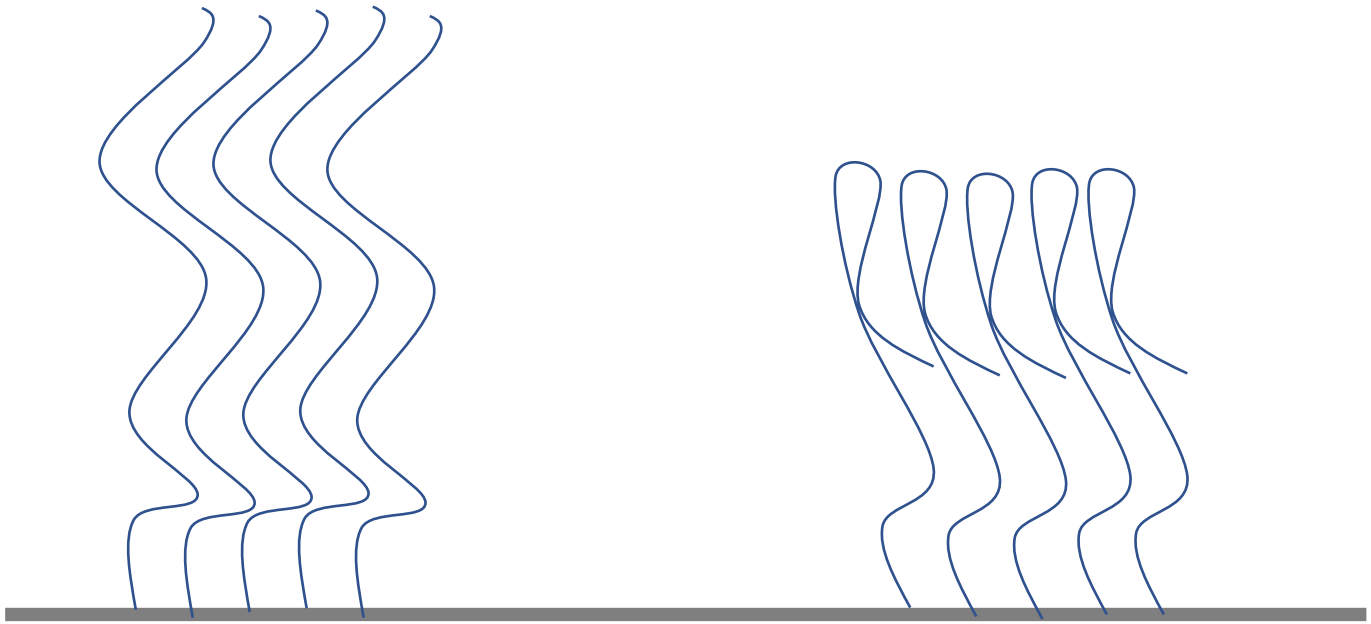
- Substitutions in Illumina read-ends (due to phasing): can be dealt with using paired-end sequencing overlaps

- Homopolymers in pyrosequencing and Ion Torrent/Proton reads

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

Error sources (inverted repeats)

Difficulties with template replication due to secondary structures (e.g. palindromic repeats).

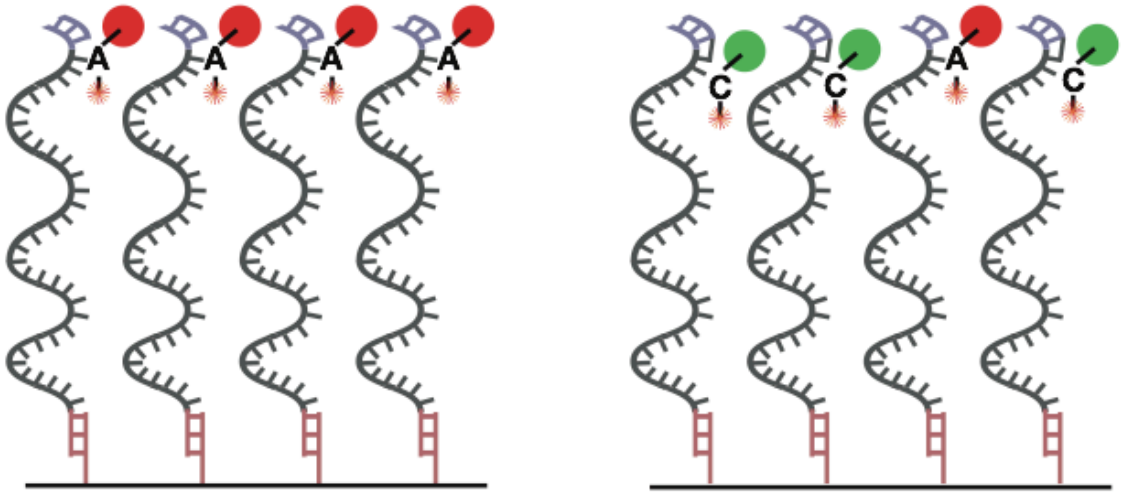


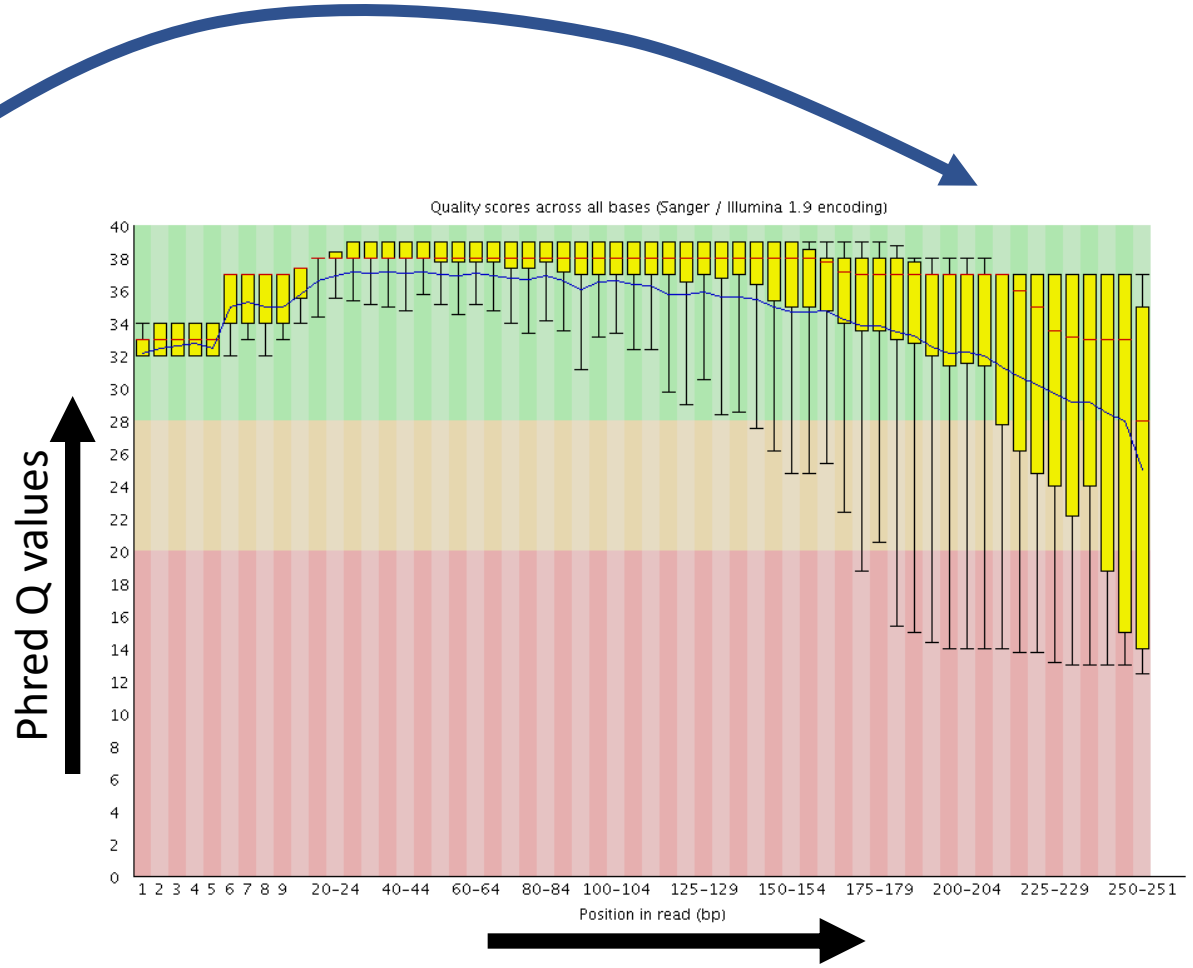
Schrimer et al., 2015; https://doi.org/10.1093/nar/gku1341
https://www.ecseq.com/support/ngs/why-does-the-sequence-quality-decrease-over-the-read-in-illumina

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

3. Error source in Illumina (phasing)

Fluorescence phasing due to imperfect function of enzymes. Occasional lack of terminator/fluorophore removal accumulates toward the read-end and reduces Phred Q (increases error probability).

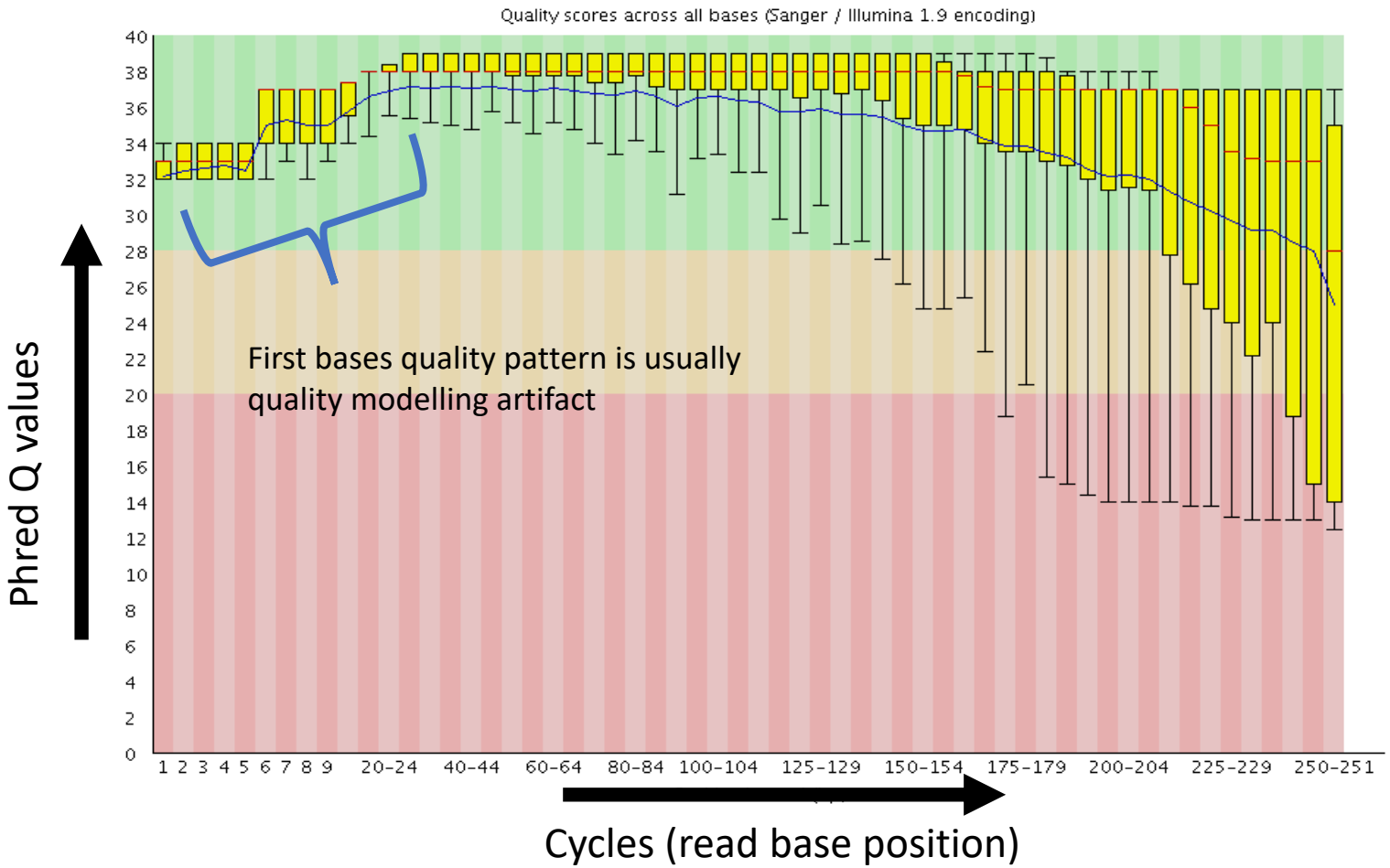




https://www.ecseq.com/support/ngs/why-does-the-sequence-quality-decrease-over-the-read-in-illumina

DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

3. Error source in Illumina (initial; basecalling algorithmic)



https://www.ecseq.com/support/ngs/why-does-the-sequence-quality-decrease-over-the-read-in-illumina

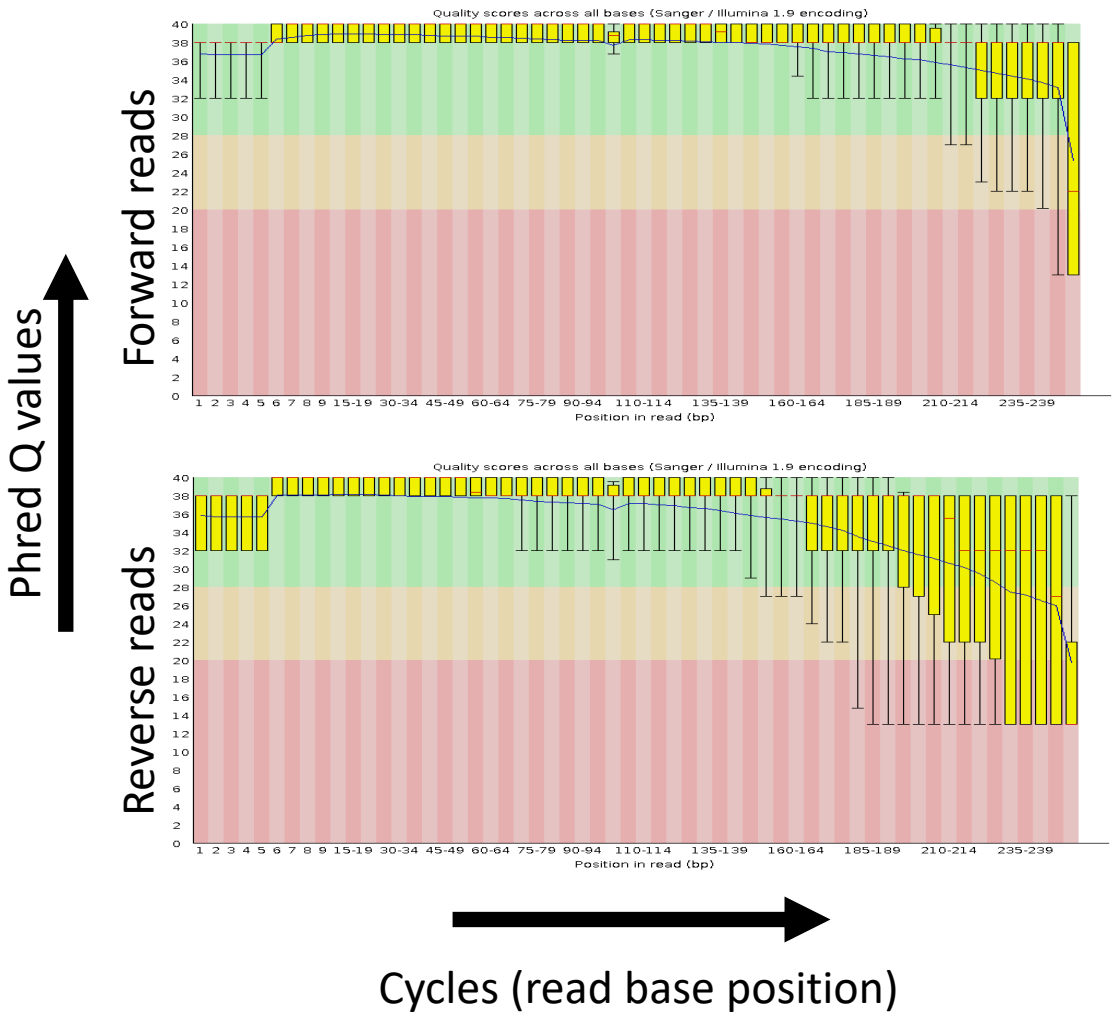
DEPARTMENT OF
Biochemistry &
Biotechnology
UNIVERSITY OF THESSALY

3. Error source in Illumina (for *vs* rev)

Besides their similar per-base error patterns, forward reads tend to have slightly higher per base quality score statistics than the reverse reads.

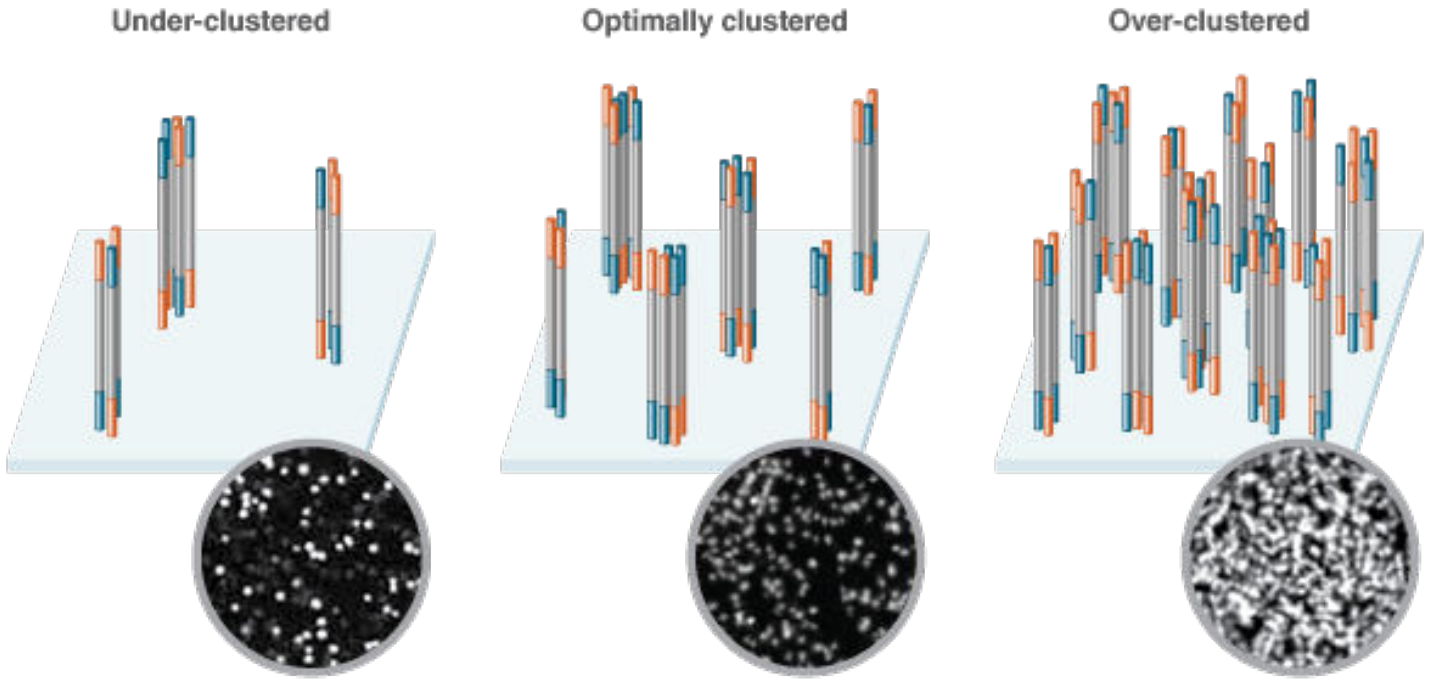
Possibly attributed to

- overclustering (extra bridge amplification for second strand)
- Possible DNA wear-out



DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

3. Error source in Illumina (over-clustering)



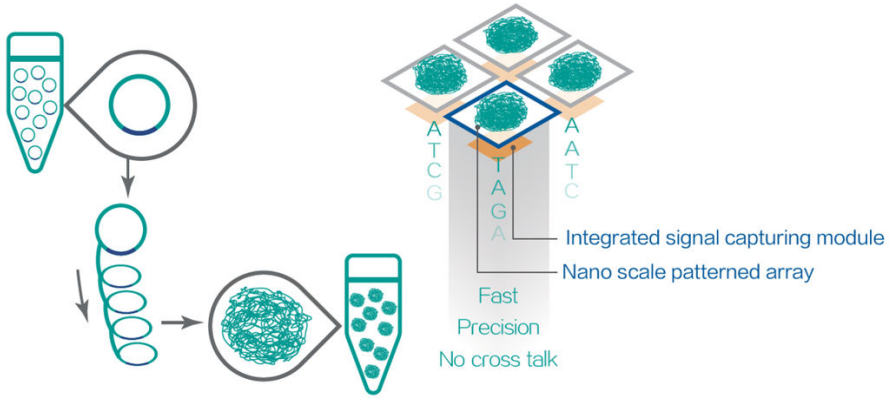
https://www.neb.com/tools-and-resources/feature-articles/the-quantitation-question-how-does-accurate-library-quantitation-influence-sequencing

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

DNA nanoball

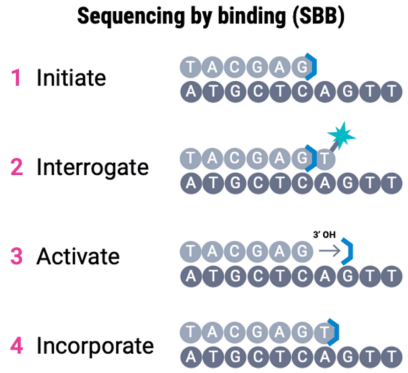
DNA nanoball

- Basic difference: rolling cycle amplification during library prep -> no secondary errors



DNA sequencing by binding similar with Illumina

- Enzymatic template interrogation with fluorescent base matching prior polymerization.

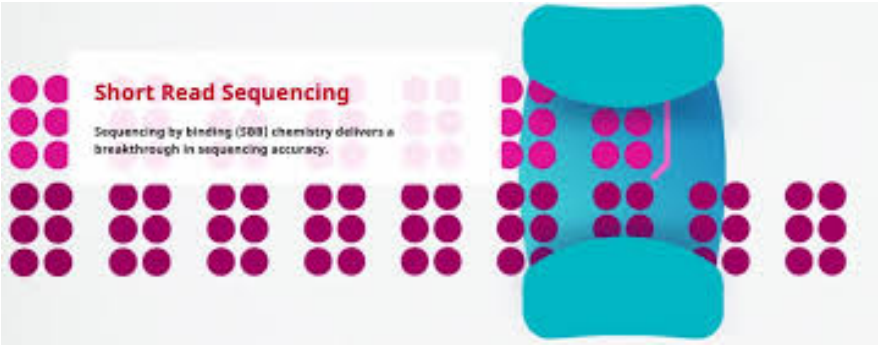


https://www.pacb.com/blog/sbb-sequencing/
https://en.mgitech.cn/Home/Products/instruments_info/id/40.html

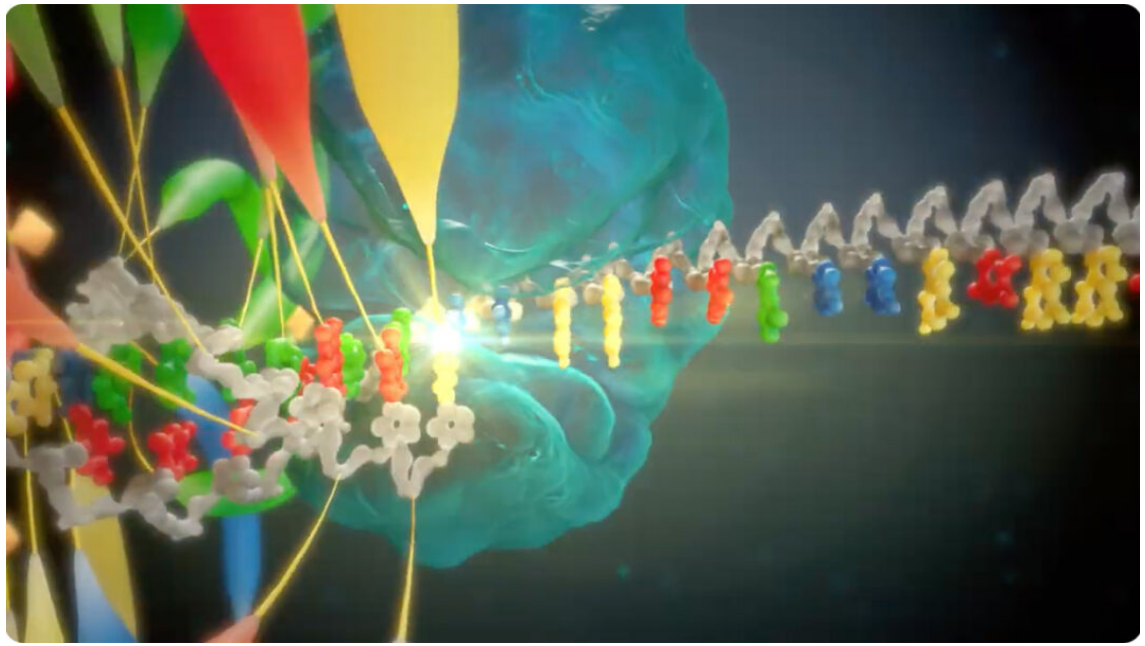
DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

Several others

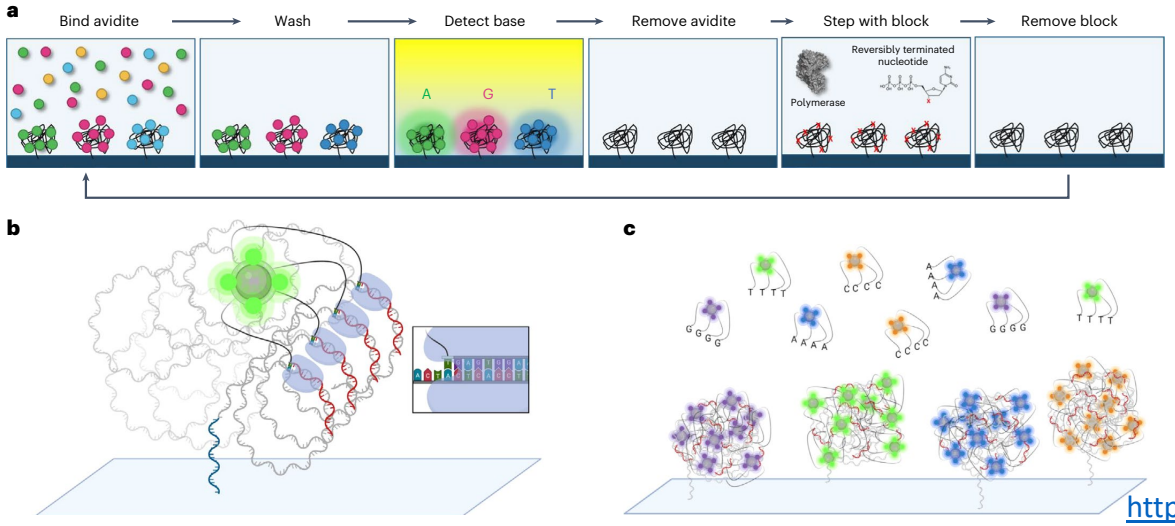
Pacific Biosciencies… sequencing by binding (Onso instrument)



Roche…Sequencing by expansion (SBX)



Element Biosciencies… rolling cycle amplification and Avidites



https://www.genengnews.com/topics/omics/the-best-of-ngs-instrument-companies-to-watch-in-2025/

Arslan, S., et al. (2024). Nat Biotechnol 42, 132-138, doi: 10.1038/s41587-023-01750-7
https://www.pacb.com/technology/sequencing-by-binding/

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

3. Synthetic long reads

Barcode and Amplify

Ligation or direct amplification add LoopSeq-specific sequences. Amplification then adds one unique barcode to each molecule and creates thousands of copies of each molecule and its unique barcode

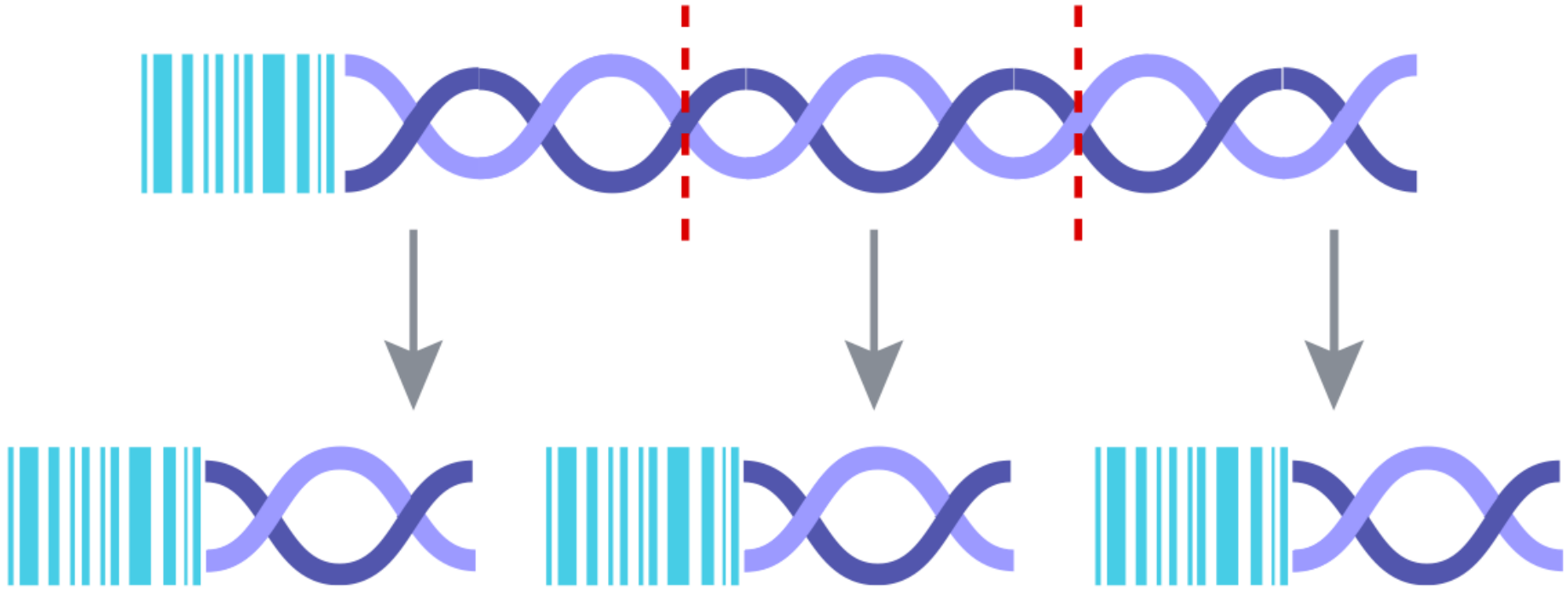
https://www.elementbiosciences.com/products/loopseq
For sFL16S see Callahan et al 2021, Microbiome https://doi.org/10.1186/s40168-021-01072-3
Or Jeong et al (2021). Scientific Reports 11, 1727, 10.1038/s41598-020-80826-9
https://hpst.cz/sites/default/files/oldfiles/loopseq-16s-18s-microbiome-24-plex-user-manual_0.pdf

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

3. Synthetic long reads

Distribute and Prep

Enzymes digest each molecule into fragments and distribute copies of the barcode throughout the fragments, so each fragment has the same barcode as the source molecule.



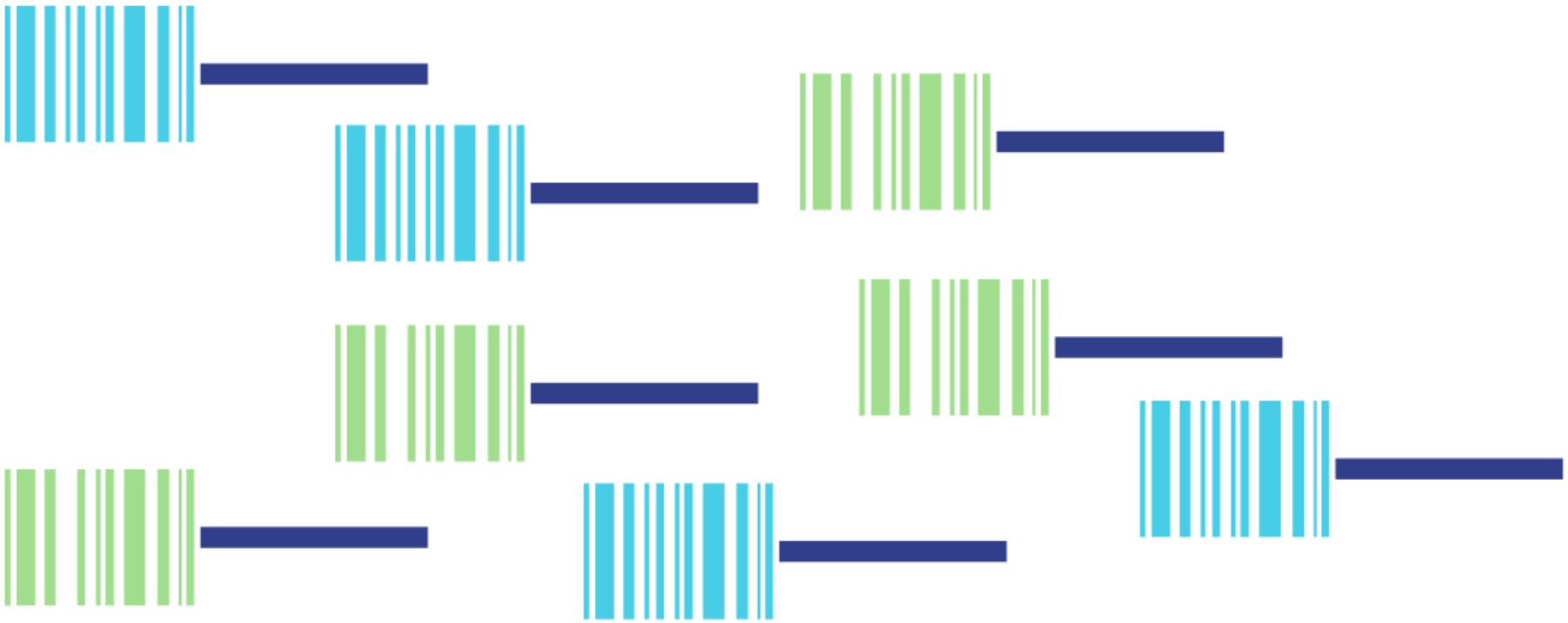
https://www.elementbiosciences.com/products/loopseq
For sFL16S see Callahan et al 2021, Microbiome https://doi.org/10.1186/s40168-021-01072-3
Or Jeong et al (2021). Scientific Reports 11, 1727, 10.1038/s41598-020-80826-9
https://hpst.cz/sites/default/files/oldfiles/loopseq-16s-18s-microbiome-24-plex-user-manual_0.pdf

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

3. Synthetic long reads

Sequence

The AVITI System sequences the fragments as short reads. Subsequent demultiplexing groups the short reads by barcode.



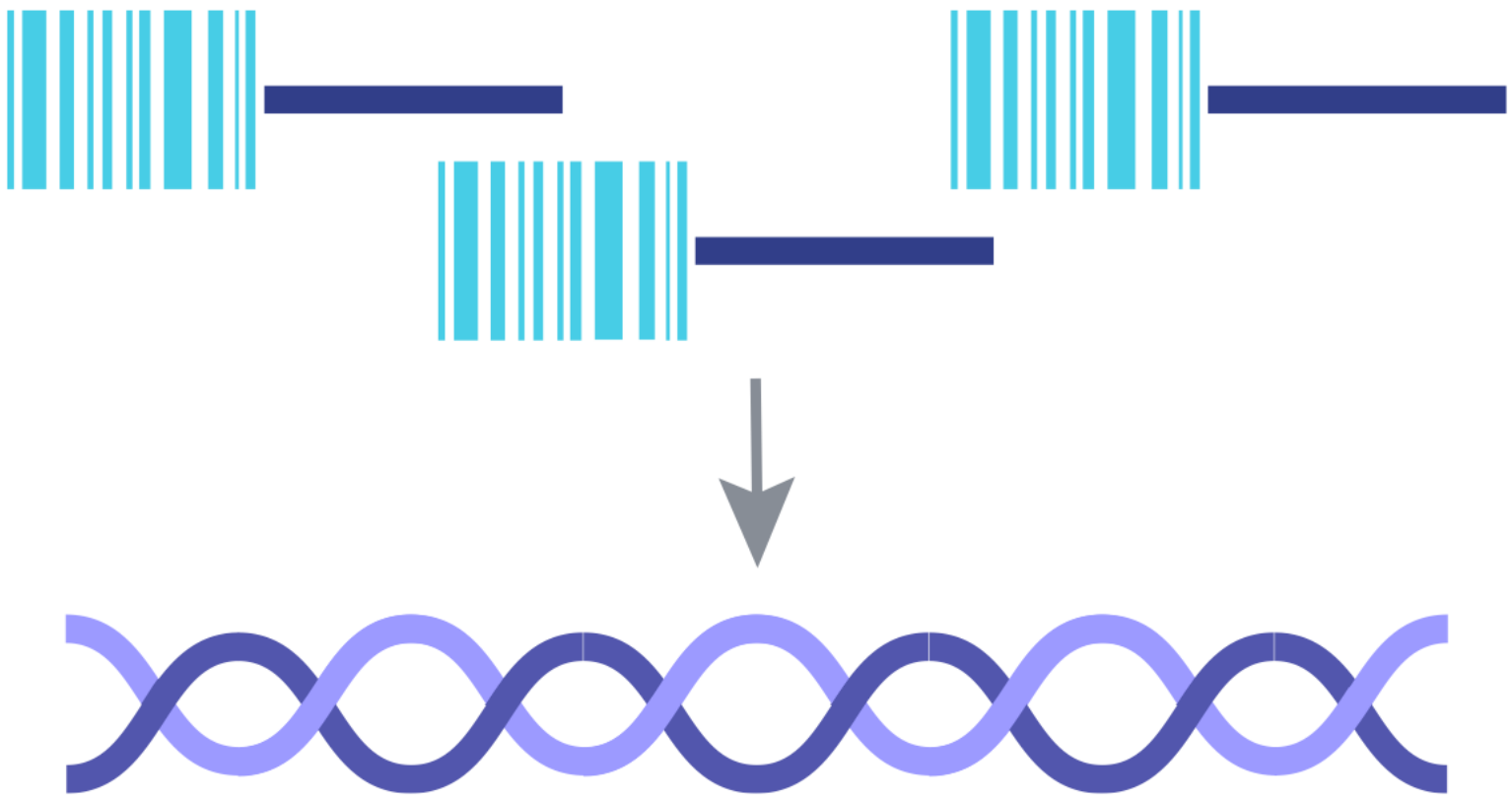
https://www.elementbiosciences.com/products/loopseq
For sFL16S see Callahan et al 2021, Microbiome https://doi.org/10.1186/s40168-021-01072-3
Or Jeong et al (2021). Scientific Reports 11, 1727, 10.1038/s41598-020-80826-9
https://hpst.cz/sites/default/files/oldfiles/loopseq-16s-18s-microbiome-24-plex-user-manual_0.pdf

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

3. Synthetic long reads

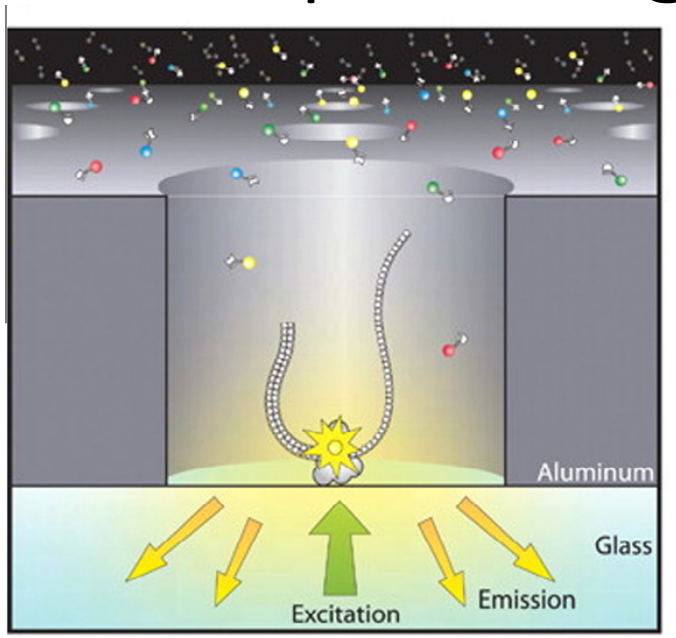
Reassemble

De novo assembly reassembles overlapping short reads into the original full-length molecule.

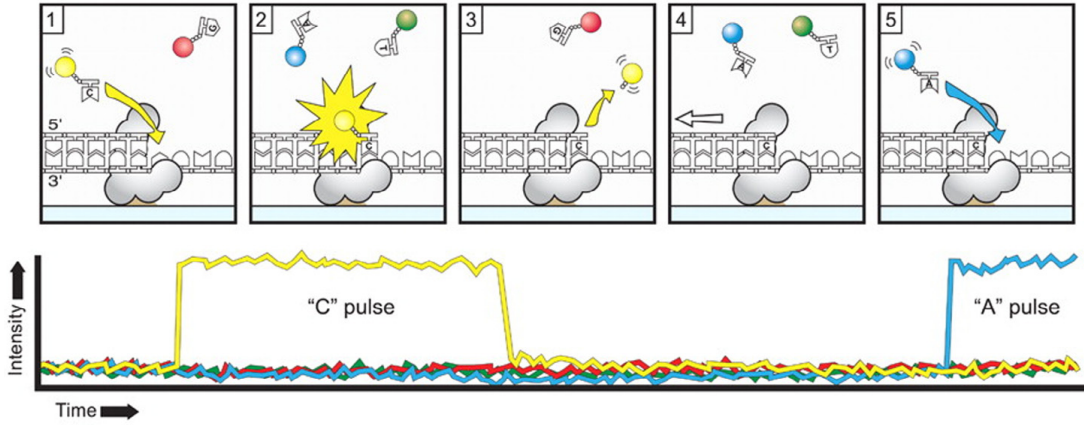


https://www.elementbiosciences.com/products/loopseq
For sFL16S see Callahan et al 2021, Microbiome https://doi.org/10.1186/s40168-021-01072-3
Or Jeong et al (2021). Scientific Reports 11, 1727, 10.1038/s41598-020-80826-9
https://hpst.cz/sites/default/files/oldfiles/loopseq-16s-18s-microbiome-24-plex-user-manual_0.pdf

3rd generation: PacBio (single molecule real-time – SMRT – sequencing)

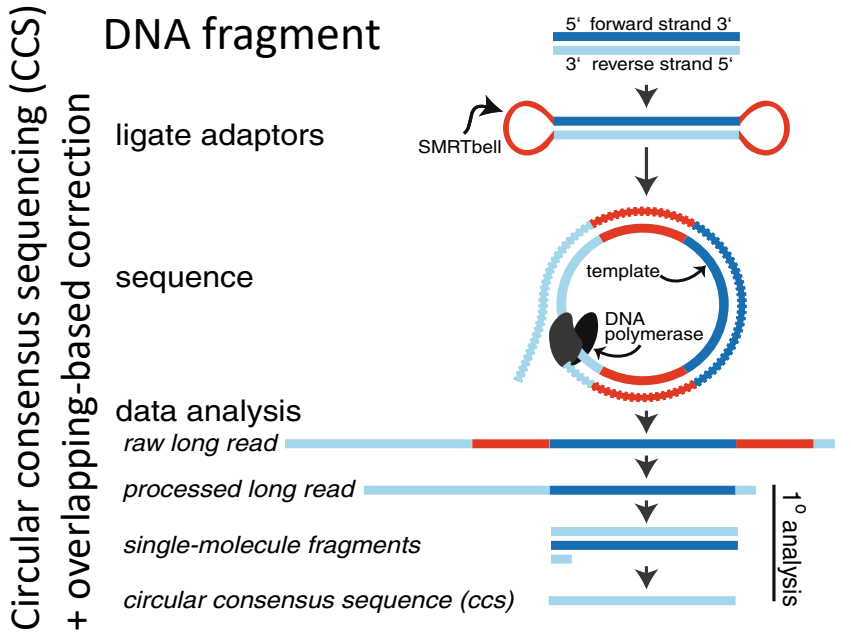


Zero-Mode Waveguide (ZMW)



Very long reads (currently of mean length of > 10 kbp)
but…
Random errors up to 10 % (mostly InDels)…

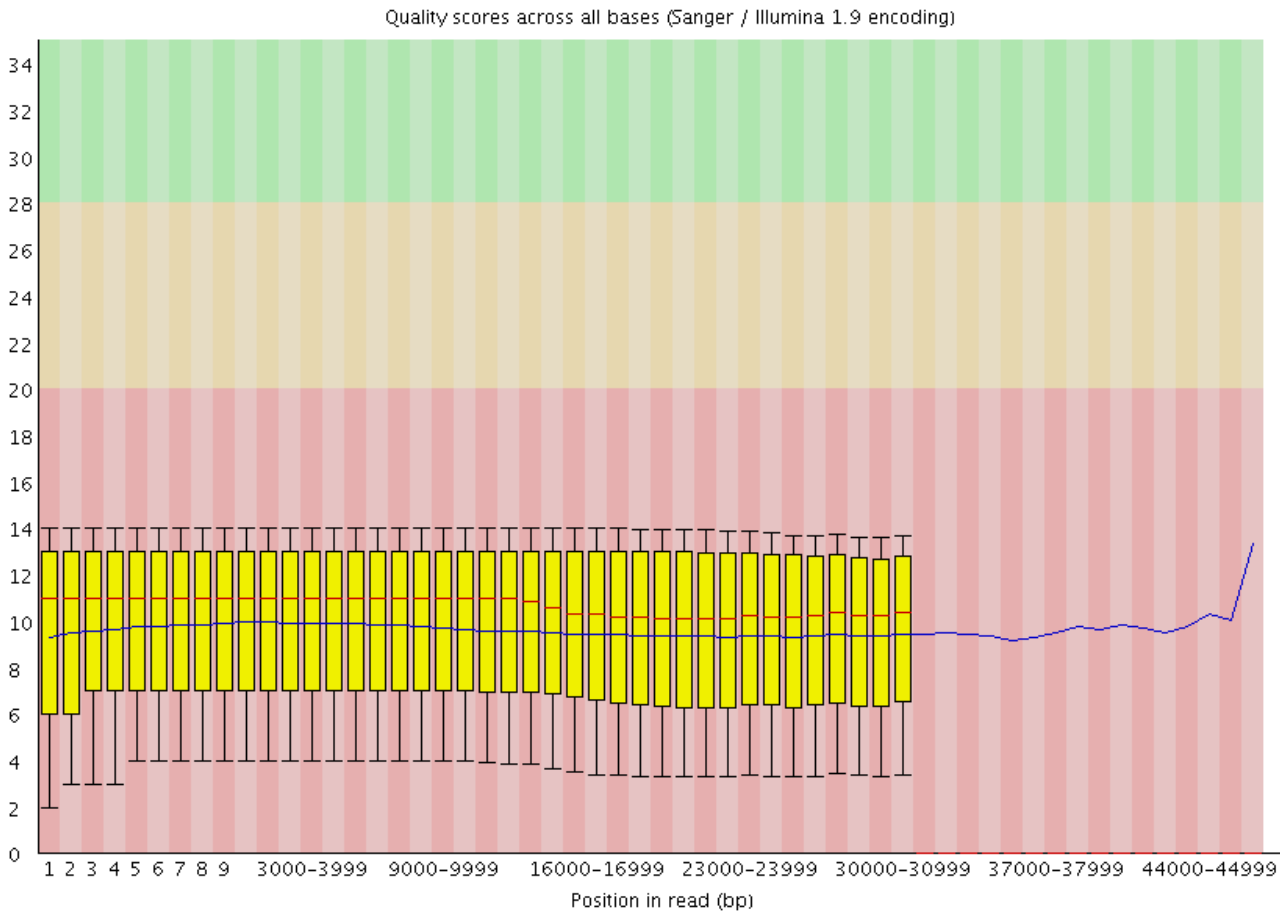
HiFi-reads!!!
Solution

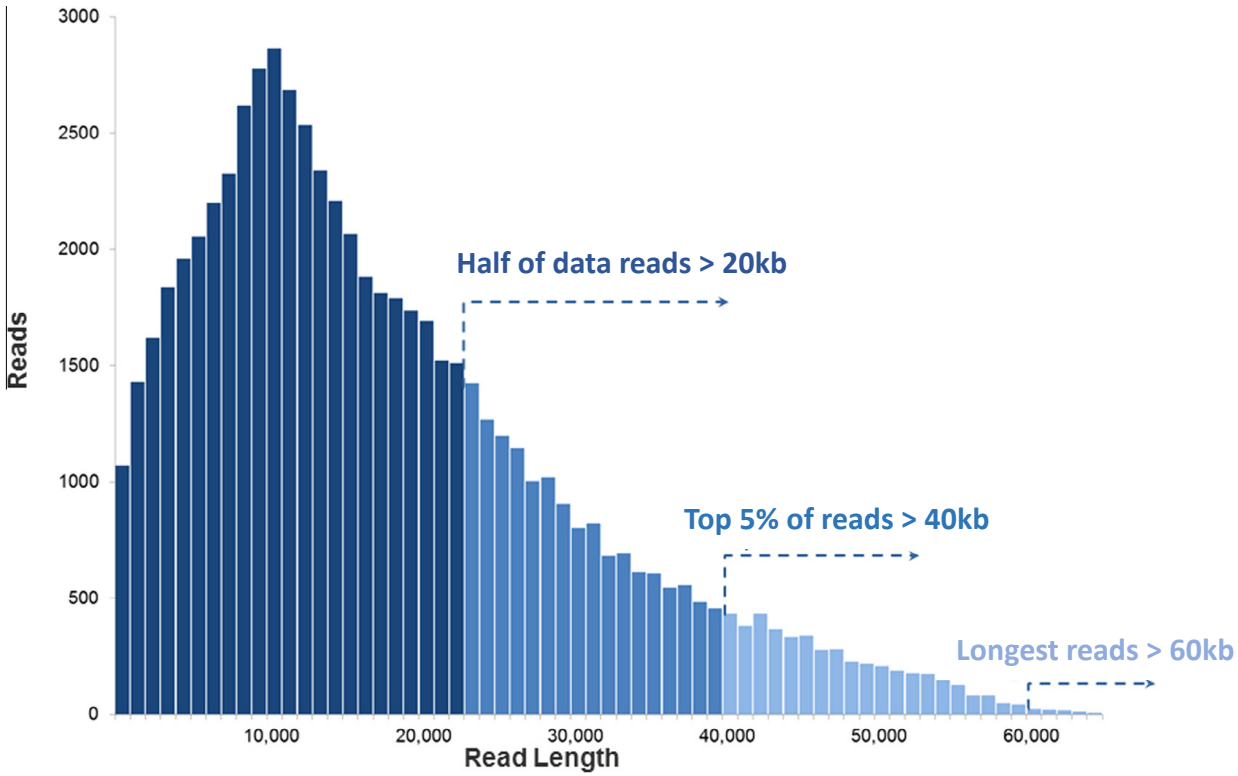


https://www.youtube.com/watch?v=v8p4ph2MAvI

Rhoads, A. and Au, K.F. (2015). PacBio Sequencing and Its Applications. Genomics, Proteomics & Bioinformatics 13, 278-289
Metzker, M.L. (2009). Sequencing technologies — the next generation. Nat Rev Genet 11, 31

3rd generation: PacBio per-base error probabilities



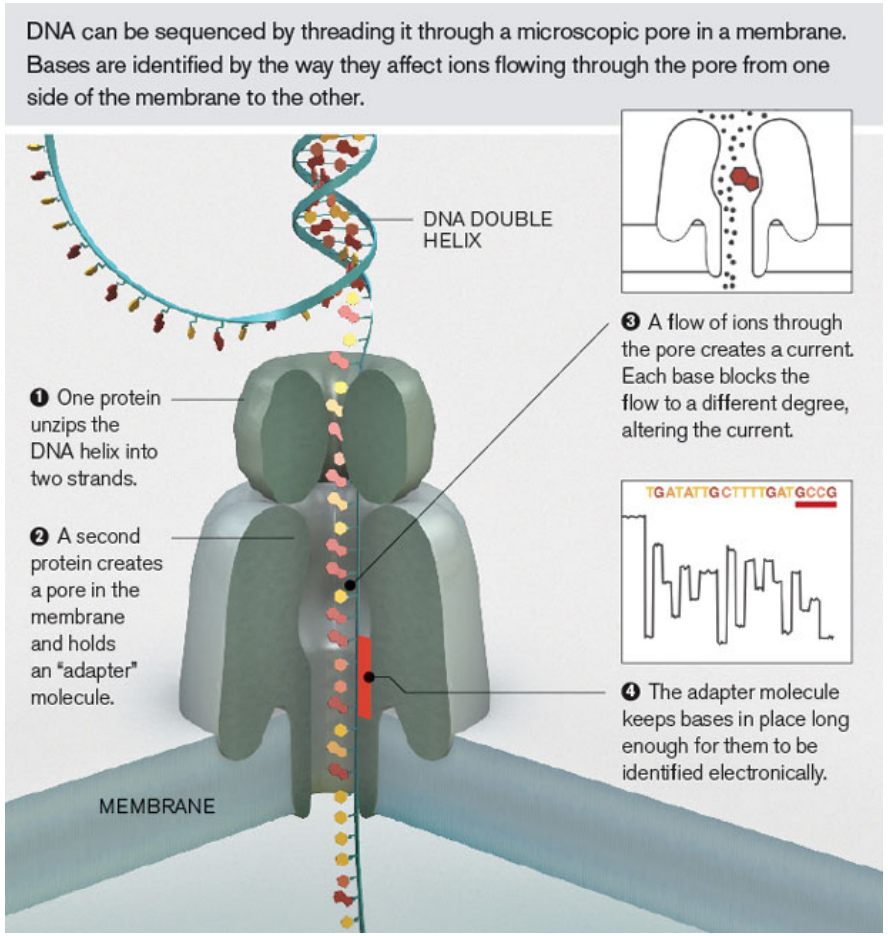


3rd generation: PacBio read error types

- Mainly Insertion/deletion
- Due to:
  - Polymerase monitoring sync issues (e.g. polymerization speed variation – usually deletion due to polymerase high speed)
  - Unincorporated nucleotide pass-bys (insertions)
- Corrected via the circular consensus sequencing process (multiple passes of the same sequences) and also computationally (overlapping-based correction)

Koren, S., *et al.* (2013). Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biol 14, R101

3rd generation: Nanopore



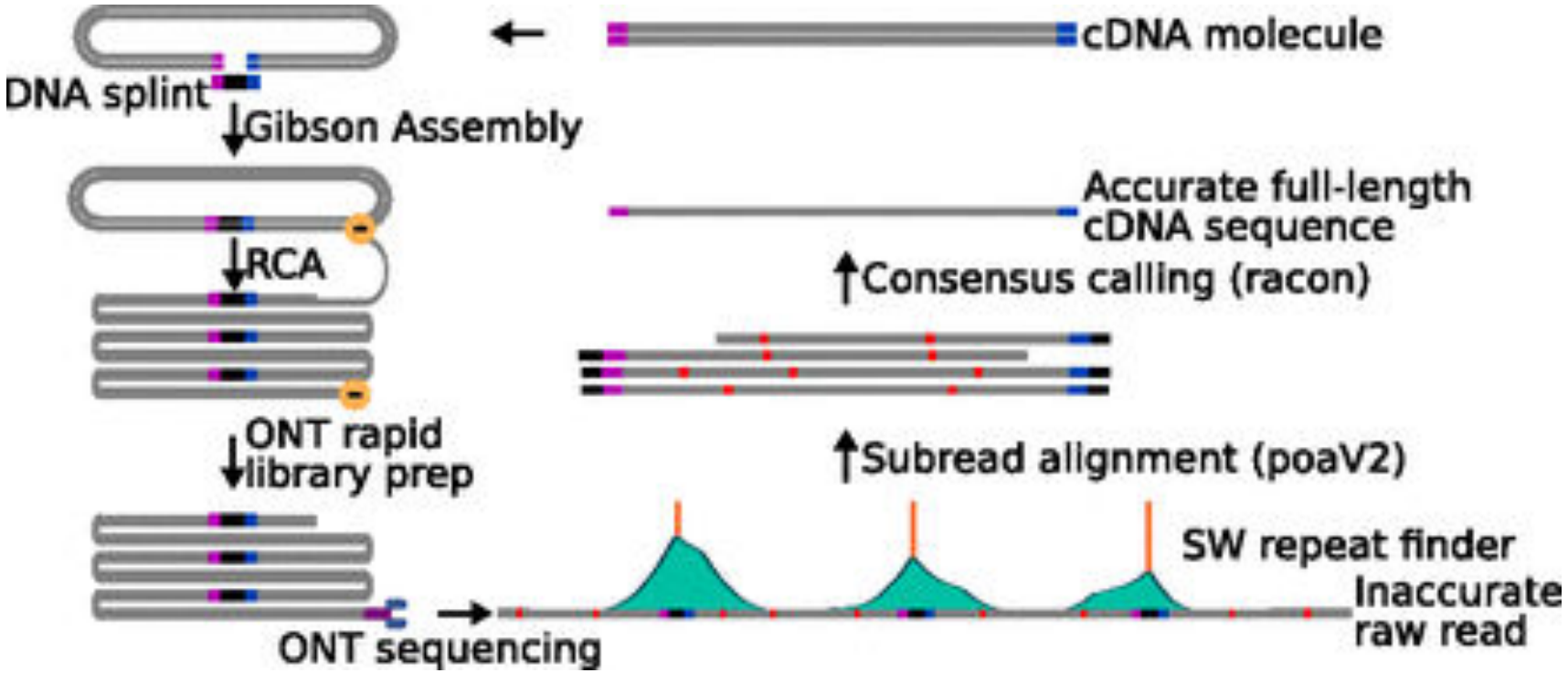
- Higher error rates than PacBio
- Error types are usually InDels
- Minimal sample prep and benchtop convenience

https://nanoporetech.com/products/minion

Lu, H., *et al.* (2016). Oxford Nanopore MinION Sequencing and Genome Assembly. Genomics, Proteomics & Bioinformatics 14, 265-279
O'Donnell, C.R., *et al.* (2013). Error analysis of idealized nanopore sequencing. Electrophoresis 34, 2137-2144

3rd generation: Nanopore

- Dramatic error rate reduction approach Rolling Circle Amplification (RCA)… similarly to PacBio



Volden, R., *et al.* (2018). PNAS 115, 9726-9731; doi 10.1073/pnas.1806447115
Zee, A, *et al.* (2021). BioRxiv 2021.10.30.466545; doi 10.1101/2021.10.30.466545

3rd generation: high quality amplicon sequencing with Nanopore/PacBio

Karst et al 2021, Nat Methods https://doi.org/10.1038/s41592-020-01041-y

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

The 3 generations of sequencing technologies

(size *vs* analyzed sequence variant numbers)

- 1st generation sequencing (single DNA fragment per reaction, ~700-1000 bp long)
  - Dideoxy termination method (**Sanger**)
- 2nd generation sequencing (massively parallel, max of ~300-600 bp total)
  - Sequencing by ligation (e.g. ABI-SOLiD)
  - Sequencing by synthesis (e.g. *Pyrosequencing*, **Illumina**, *Ion-torrent*)
- 3rd generation sequencing (massively parallel and very long reads, max of ~ 10kbp-250kbp)
  - Sequencing by synthesis (e.g. single molecule real time – **PacBio**, **Nanopore**)
  - Synthetic long reads (e.g. Moleculo, 10X… e.g. see sFL16S)

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

The 3 generations of sequencing technologies

(size *vs* analyzed sequence variant numbers)

- 1st generation sequencing (single DNA fragment per reaction, ~700-1000 bp long)
  - Dideoxy termination method (**Sanger**)
- 2nd generation sequencing (massively parallel, max of ~300-600 bp total)
  - Sequencing by ligation (e.g. ABI-SOLiD)
  - Sequencing by synthesis (e.g. *Pyrosequencing*, **Illumina**, *Ion-torrent*)
- 3rd generation sequencing (massively parallel and very long reads, max of ~ 10kbp-250kbp)
  - Sequencing by synthesis (e.g. single molecule real time – **PacBio**, **Nanopore**)
  - Synthetic long reads (e.g. Moleculo, 10X… e.g. see sFL16S)

Which method to select ?

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

The 3 generations of sequencing technologies
(size *vs* analyzed sequence variant numbers)

Which method to select?

- Amplicon size (150-450bp most established cases; complete marker genes – 1000-1500bp – are always desirable): 1st, 2nd, 3rd generation
- 10s of thousands of reads per sample required: 2nd generation is more cost-effective
- Paired reads desirable: Illumina (2x250bp, 2x300bp)
- Lowest error rates desirable: Illumina, synthetic, highQ-long

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

The 3 generations of sequencing technologies
(size *vs* analyzed sequence variant numbers)

Which method to select?

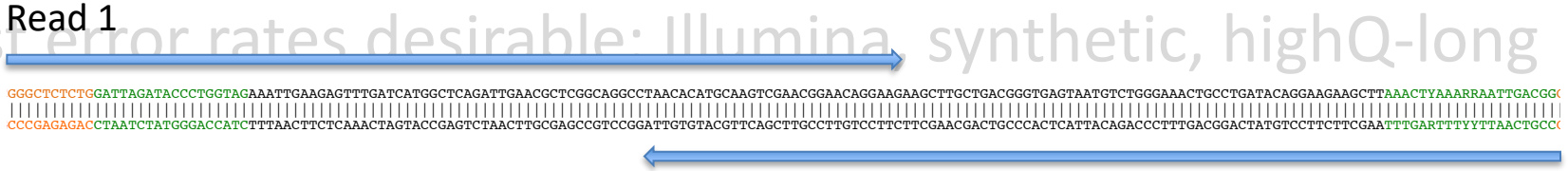
- Amplicon size (150-450bp most established cases; complete marker genes – 1000-1500bp – are always desirable): 1st, 2nd, 3rd generation
- 10s of thousands of reads per sample required: 2nd generation is more cost-effective
- Paired reads desirable: Illumina (2x250bp, 2x300bp)
- Lowest error rates desirable: Illumina, synthetic, highQ-long

DEPARTMENT OF
Biochemistry &
Biotechnology
UNIVERSITY OF THESSALY

The 3 generations of sequencing technologies
(size *vs* analyzed sequence variant numbers)

Which method to select?

- Amplicon size (150-450bp most established cases; complete marker genes – 1000-1500bp – are always desirable): 1st, 2nd, 3rd generation
- 10s of thousands of reads per sample required: 2nd generation is more cost-effective
- Paired reads desirable: Illumina (2x250bp, 2x300bp)
- Lowest error rates desirable: Illumina, synthetic, highQ-long



DEPARTMENT OF
Biochemistry &
Biotechnology
UNIVERSITY OF THESSALY

The 3 generations of sequencing technologies
(size *vs* analyzed sequence variant numbers)

Which method to select?

- Amplicon size (150-450bp most established cases; complete marker genes – 1000-1500bp – are always desirable): 1st, 2nd, 3rd generation
- 10s of thousands of reads per sample required: 2nd generation is more cost-effective
- Paired reads desirable: Illumina (2x250bp, 2x300bp)
- Lowest error rates desirable: Illumina, synthetic, highQ-long

DEPARTMENT OF
Biochemistry &
Biotechnology
UNIVERSITY OF THESSALY

The 3 generations of sequencing technologies

(size *vs* analyzed sequence variant numbers)

Which method to select?

- Amplicon size (150-450bp most established cases; complete marker genes – 1000-1500bp – are always desirable): 1st, 2nd, 3rd generation (3rd generation is advantageous here)
- 10s of thousands of reads per sample required: 2nd generation is more cost-effective still
- Paired reads desirable: Illumina (2x250bp, 2x300bp)
- Lowest error rates desirable: Illumina, synthetic, highQ-long

Main traits of sequencing technologies

Three generations of sequencing (read size, analyzed fragment numbers)

- 1st: single fragment per reaction, ~700-1000 bp long (Sanger – dideoxy termination method)
- 2nd: millions of fragments per reaction up to 2x300 bp reads (Illumina – sequencing by synthesis)
- 3rd: 10's-100's of thousands of reads of up to 250kbp long (PacBio, Nanopore; reports for reads of 4Mbp for Nanopore)

Sequencing error types and positional bias per technology

- Sanger: low quality read beginnings/ends
- Illumina: substitutions, error prone read ends (phasing)
- PacBio, Nanopore: indels, randomly distributed (high error rates which can be corrected with improved base-calling algorithms, CCS)

Current main applications (practically every method is used for several tasks but mostly for…)

- Sanger: single locus analysis
- Illumina (extremely high throughput): genomics, **amplicon diversity**, ChipSeq, RNAseq…
- PacBio, Nanopore: genomics, meta-gnomics, RNAseq

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

The 3 generations of sequencing technologies
(size *vs* analyzed sequence variant numbers)

Which method to select?

Illumina? (low-cost/high-quality/high-depth)

synthetic? (high-cost/high-quality/high-depth)

PacBio/Nanopore? (high-cost/intermediate-to-high-quality/low-depth)

Interesting Nanopore approach https://nanoporetech.com/sites/default/files/s3/posters/lc2022/LukaszKrych%20LC22.pdf

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

The 3 generations of sequencing technologies
(size *vs* analyzed sequence variant numbers)

How much can the read length of a single marker gene affect our study outcome?

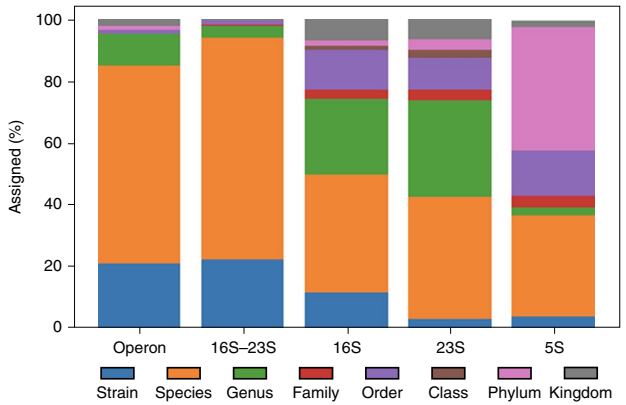
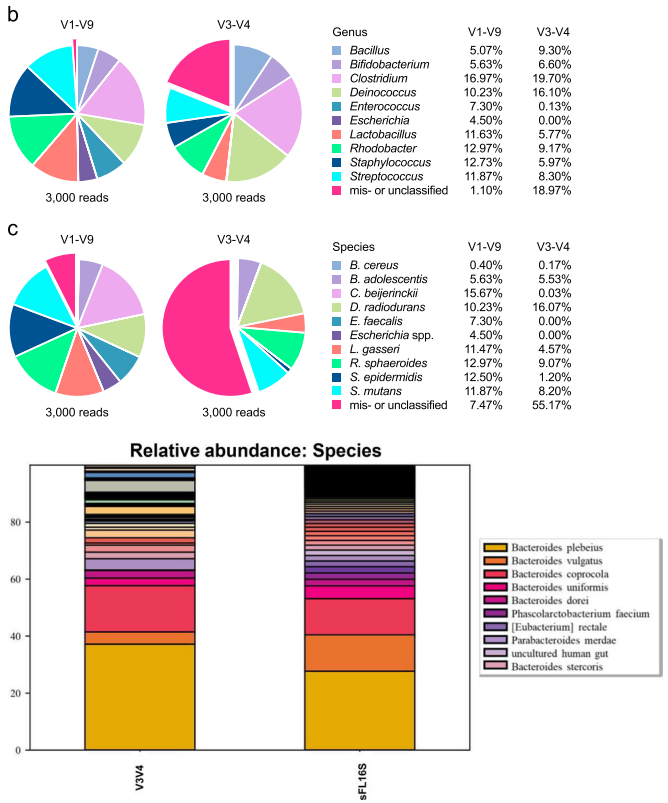
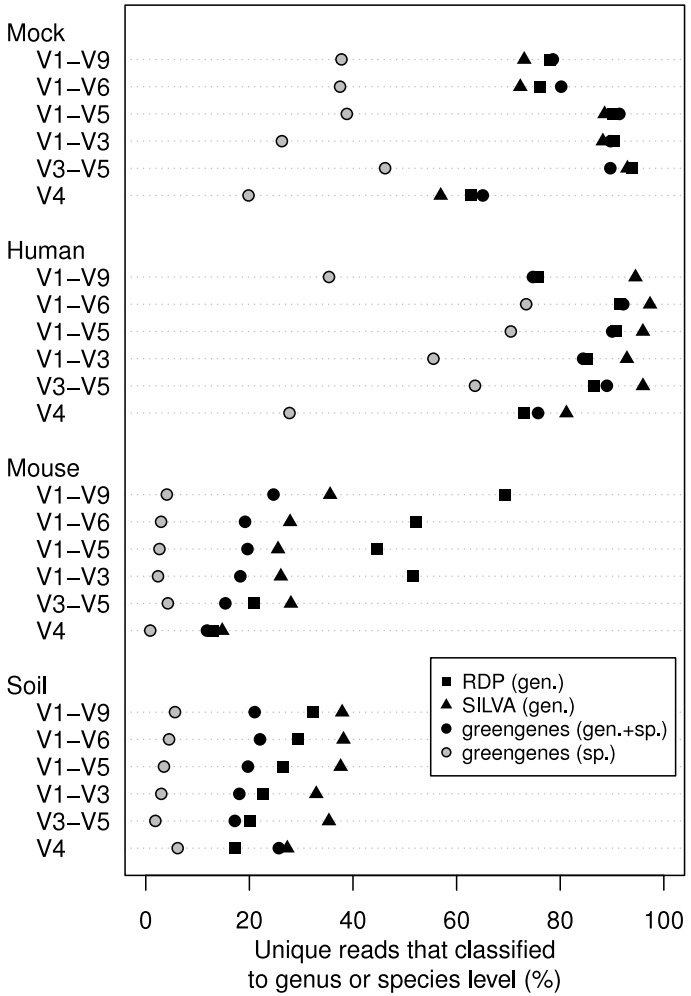


Fig. 3 | BLAST-based consensus taxonomic assignment against the Web of Life database for whole rRNA operons, using the combination of 16S and 23S rRNAs and individual rRNA genes. In the dataset, 253,089 operons were available and used for assignment. Of these, $n = 253,087$ had an annotatable 23S rRNA gene, $n = 253,088$ had an annotatable 16S rRNA gene and $n = 50,560$ had an annotatable 5S rRNA gene.





Karst, et al 2021. Nat Methods https://doi.org/10.1038/s41592-020-01041-y

Matsuo, et al 2021, BMC Microbiol, https://doi.org/10.1186/s12866-021-02094-5

Jeong et al 2021, Scientific Reports https://doi.org/10.1038/s41598-020-80826-9

Schloss et al 2016, Peer J, https://doi.org/10.7717/peerj.1869

Johnson et al 2019, Nat Commun https://doi.org/10.1038/s41467-019-13036-1

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

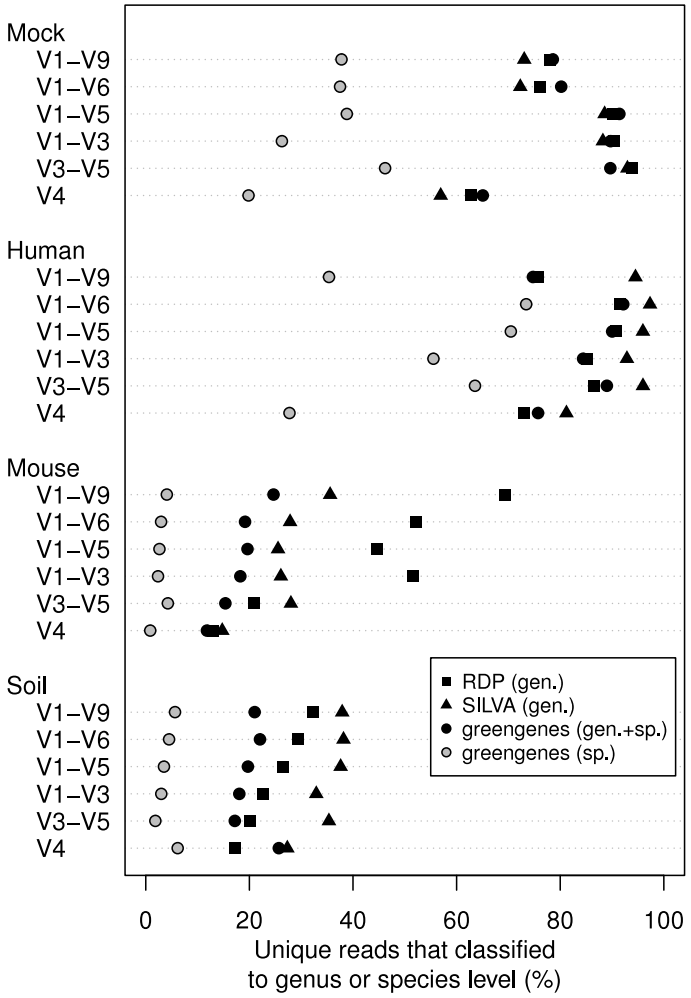
The 3 generations of sequencing technologies
(size *vs* analyzed sequence variant numbers)

How much can the read length of a single marker gene affect our study outcome?

- Prokaryotes (genomic plasticity)?
- Environment?
- Reference database?
- Eukaryotes (makes sense)?

Can we resort to cost-effective hybrid solutions? E.g.:

- Generate habitat/experiment specific (pooled DNA template) long-read reference databases for our short-read sample data (AutoTax)



Schloss et al 2016, Peer J, https://doi.org/10.7717/peerj.1869
Dueholm et al 2020, mBio, https://doi.org/10.1128/mbio.01557-20