

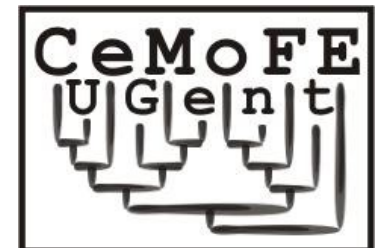
Next Generation Sequencing *for* **DUMMIES**

Andy Vierstraete,
Department of Biology,
Ghent University.
Version March 14th 2018

Andy.Vierstraete@ugent.be



<http://users.ugent.be/~avierstr/>



Center for Molecular Phylogeny and Evolution

Next Generation Sequencing **=** ***High-throughput Sequencing***

- History and future of DNA sequencing
- Workflow
- Different platforms
- Quality scores in sequencing
- Applications
- Run types
- Data analysis
- Considerations

History and future of DNA sequencing



— 1953: Discovery of DNA structure by Watson and Crick

— 1967: First DNA sequence of 11 bp published (20 pages)

History and future of DNA sequencing

4/165

1953: Discovery of DNA structure by Watson and Crick

1967: First DNA sequence of 11 bp published (20 pages) *J. Mol. Biol.* (1967) **30**, 507–527

Studies on the Bacteriophage MS2

IV†. The 3'-OH Terminal Undecanucleotide Sequence of the Viral RNA Chain

R. DE WACHTER AND W. FLIERS

Laboratory of Physiological Chemistry, State University of Ghent, Belgium

(Received 1 May 1967, and in revised form 29 July 1967)

The 3'-OH terminus of bacteriophage MS2 RNA was selectively labelled with ^3H . This was achieved by oxidation of the free 2', 3'-diol group with sodium periodate to a dialdehyde, and reduction of the latter with tritiated sodium borohydride. Treatment of this RNA with alkali and separation of the hydrolysis products

firmly each other. The results, together with the known specificity of the ribonuclease T_1 , which had released the sequence, establish that MS2 RNA ends in ...GpUpUpApCpCpApCpCpA.

It is suggested that the termination signal for the translation into polypeptides

1. Introduction

Apart from several transfer RNA's, little is known about the primary structure of macromolecular RNA's. Particularly, one would like to gain information on the beginning and on the ending of a messenger RNA, as this might possibly be related to genetic signals for polypeptide chain initiation and termination. Viral RNA, although not a typical messenger *sensu stricto*, behaves nevertheless in many respects as a simple, polycistronic message. Sugiyama & Fraenkel-Conrat (1961) identified the 3'-OH terminal nucleoside of tobacco mosaic virus RNA as adenosine. Subsequently,

History and future of DNA sequencing

5/165

1953: Discovery of DNA structure by Watson and Crick

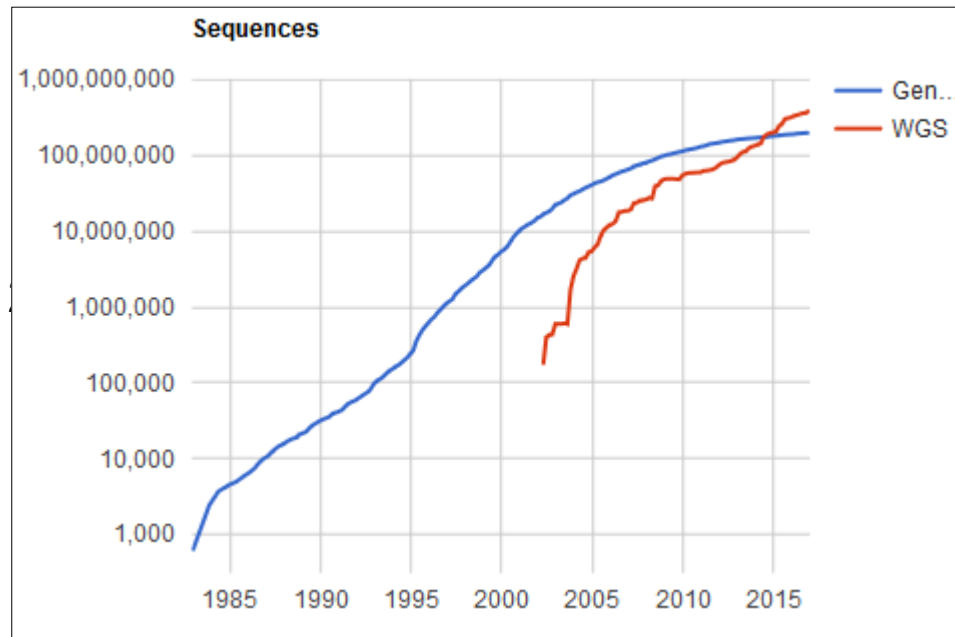
1967: First DNA sequence of 11 bp published (20 pages)

1976: First genome sequenced: Bacteriophage MS2 (3569 bp) by Walter Fiers

1977: Sanger sequencing method published

1980: Nobel Prize Wally Gilbert and Fred Sanger


1982: Genbank started



		GenBank		WGS	
Release	Date	Bases	Sequences	Bases	Sequences
3	Dec 1982	680338	606		
14	Nov 1983	2274029	2427		
20	May 1984	3002088	3665		
216	Oct 2016	220731315250	197390691	1676238489250	363213315
217	Dec 2016	224973060433	198565475	1817189565845	395301176

History and future of DNA sequencing

6/165

- 
- 1953: Discovery of DNA structure by Watson and Crick
 - 1967: First DNA sequence of 11 bp published (20 pages)
 - 1976: First genome sequenced: Bacteriophage MS2 (3569 bp) by Walter Fiers
 - 1977: Sanger sequencing method published
 - 1980: Nobel Prize Wally Gilbert and Fred Sanger
 - 1982: Genbank started
 - 1983: development of PCR
 - 1987: 1st automated sequencer: Applied Biosystems Prism 373
 - 1996: Capillary sequencer: ABI 310
 - 1998: Genome of *Caenorhabditis elegans* sequenced (100 million bp)
 - 2001: Human genome sequenced (3,2 billion bp)
 - 2005: 1st 454 Life Sciences Next Generation Sequencing system: GS 20 system^(† mid 2016)
 - 2006: 1st Solexa Next Generation Sequencer: Genome Analyzer (Illumina)
 - 2007: 1st Applied Biosystems Next Generation Sequencer: SOLiD^(† Dec 2017)
 - 2009: 1st Helicos **single molecule** sequencer: Helicos Genetic Analyser System^(† Nov 2012)
 - 2011: 1st Ion Torrent Next Generation Sequencer: PGM
1st Pacific Biosciences **single molecule** sequencer: PacBio RS Systems
 - 2012: Oxford Nanopore Technologies demonstrates ultra long **single molecule** reads
 - 2014: Roche acquires Genia: development of NanoTag **single molecule** sequencing
 - 2015: 1st BGI Next Generation Sequencer: BGISEQ-500 (sold in China only)
 - 2016: 1st Oxford Nanopore Technologies sequencer: MinION
 - 2017: SeqLL announces tSMS sequencer: **single molecule** (Helicos technology)

Different platforms

7/165

- Illumina (Solexa)
 - iSeq 100
 - MiniSeq
 - MiSeq
 - NextSeq 500 - 550
 - HiSeq 2500 - 3000 – 4000
 - HiSeq X Five – Ten
 - NovaSeq 5000 - 6000
- Thermo Fisher Scientific (Applied Biosystems -> Life Technologies)
 - Ion Torrent Personal Genome Machine (PGM)
 - Ion Torrent GeneStudio S5, S5 Plus, S5 Prime
 - Ion Torrent Proton

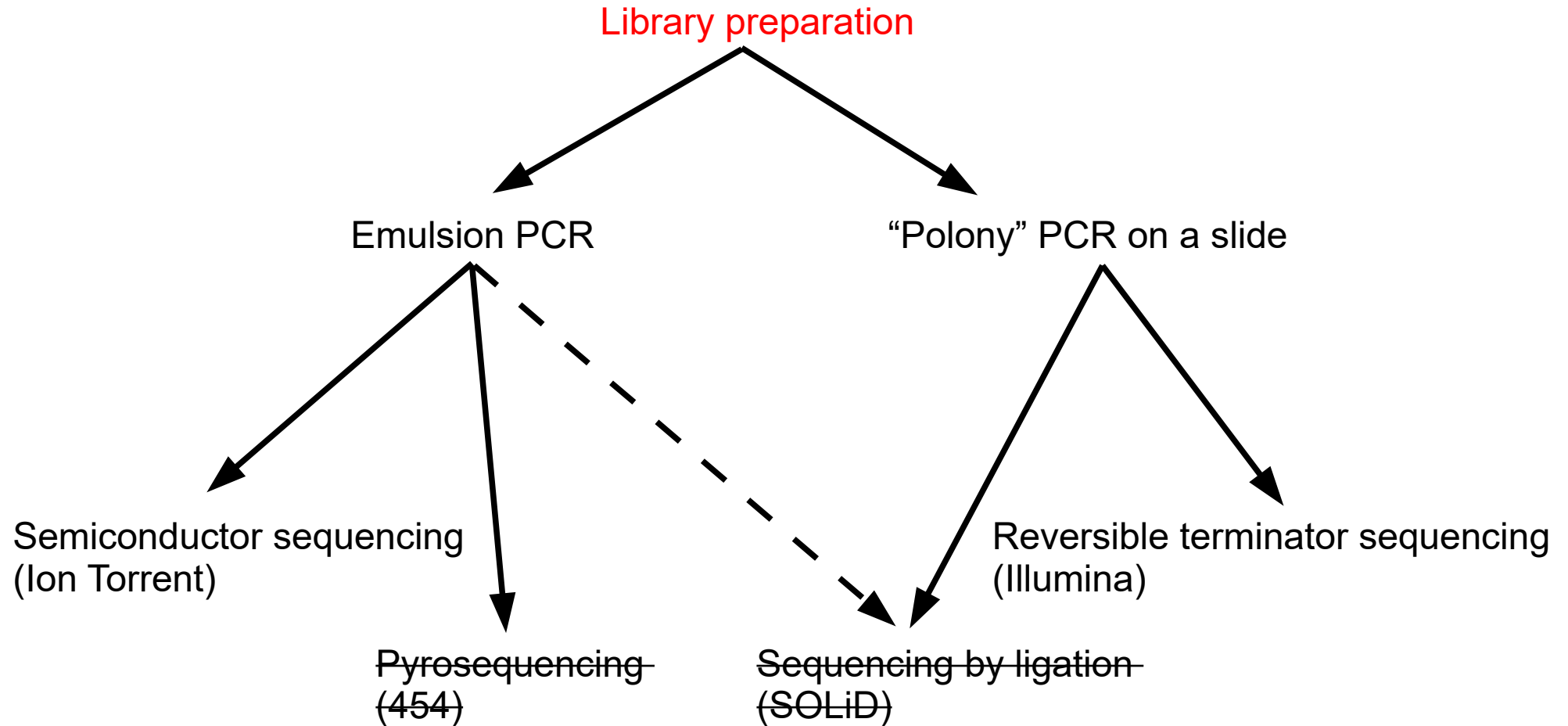
Next Generation Sequencing
Amplified Single Molecule Sequencing

- Pacific Biosciences
 - Sequel System
 - PacBio RS II
- Oxford Nanopore Technologies
 - SmidgION
 - MinION
 - GridION X5
 - PromethION
- SeqLL
 - tSMS sequencer

Third Generation Sequencing,
Next Next Generation Sequencing,
Single Molecule Sequencing

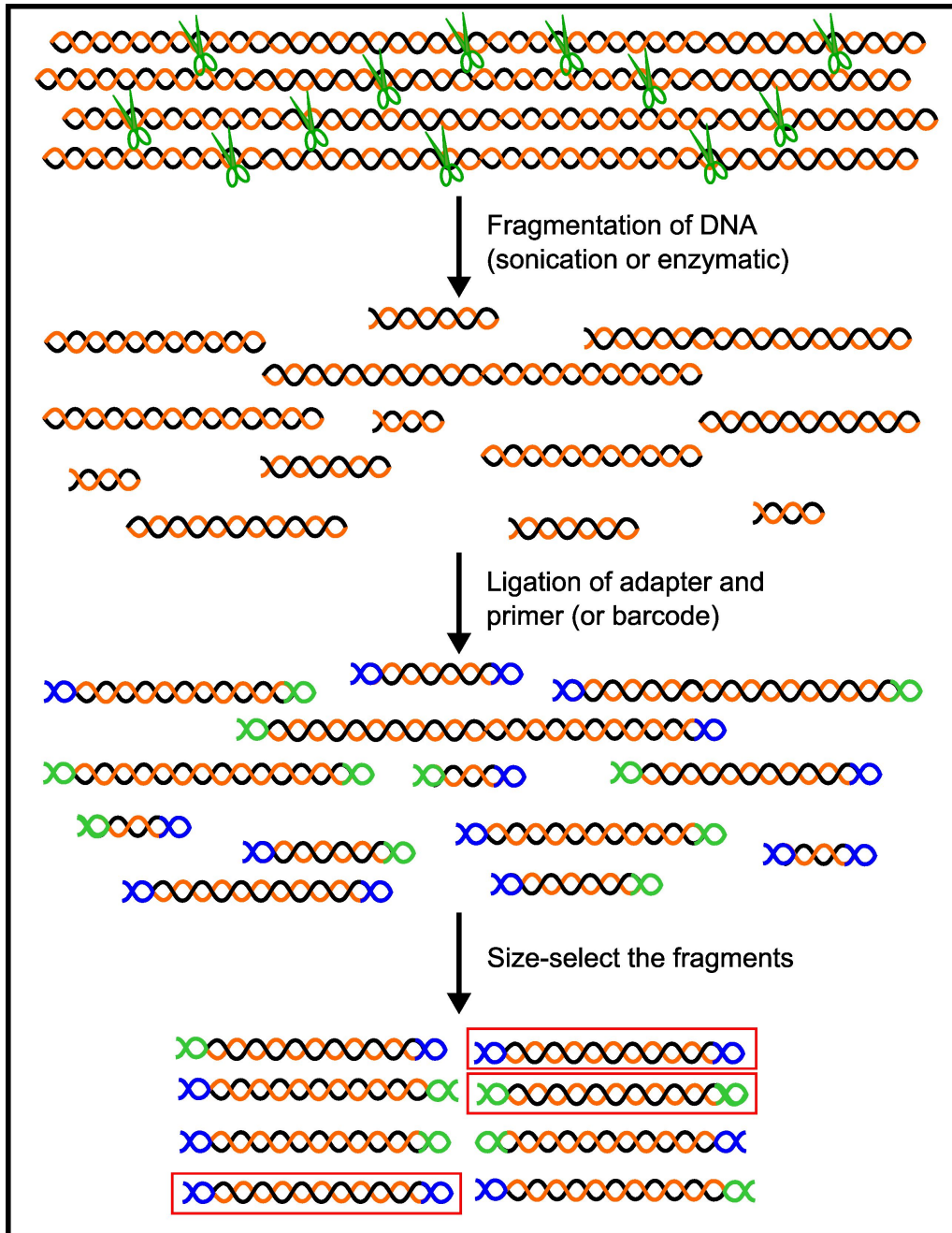
Workflow

Next Generation Sequencing: Amplified Single Molecule Sequencing



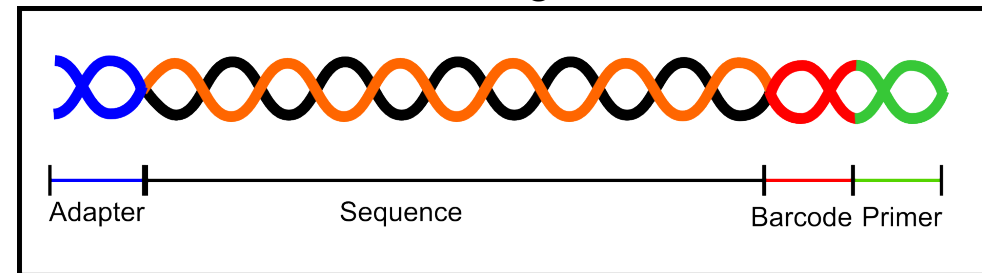
Workflow

Next Generation Sequencing: Amplified Single Molecule Sequencing



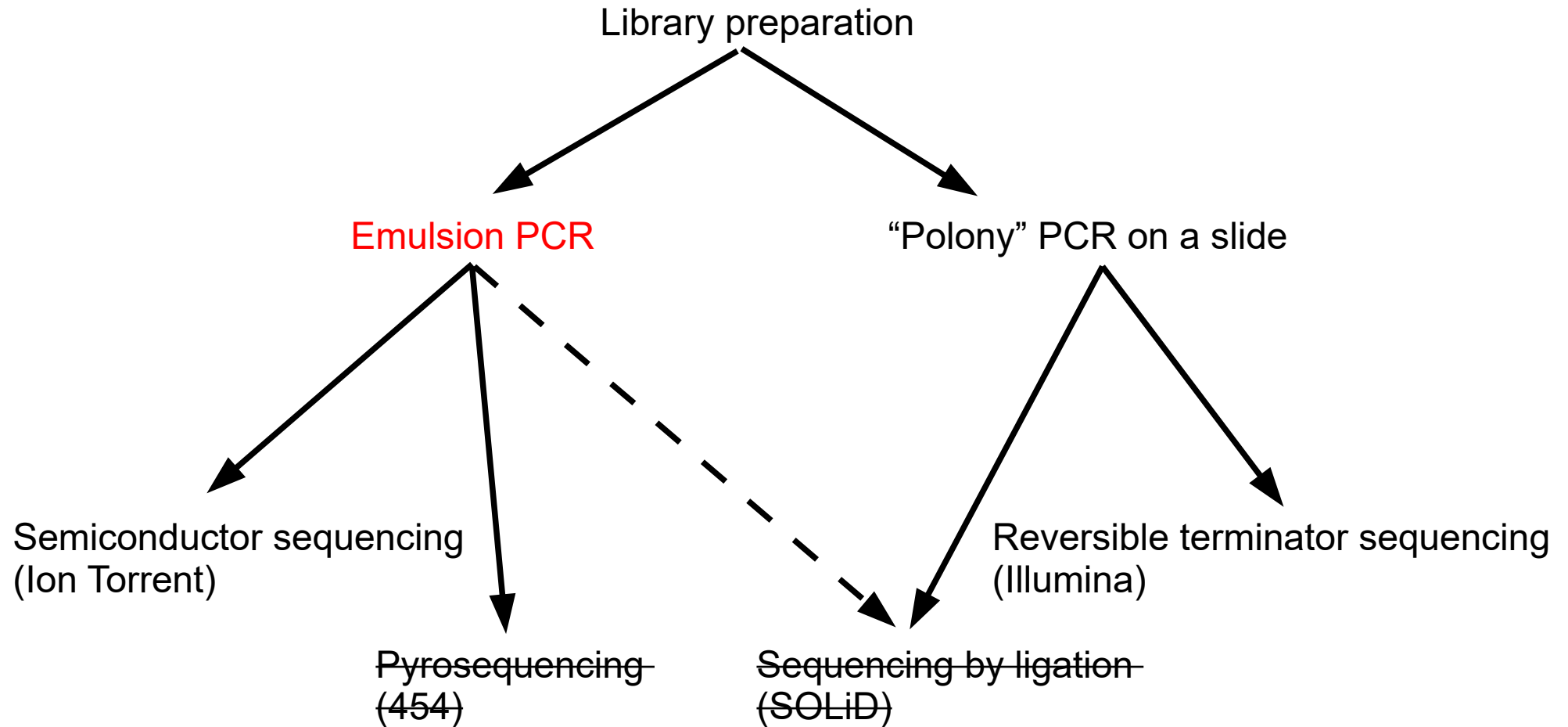
Library preparation

Good fragments:



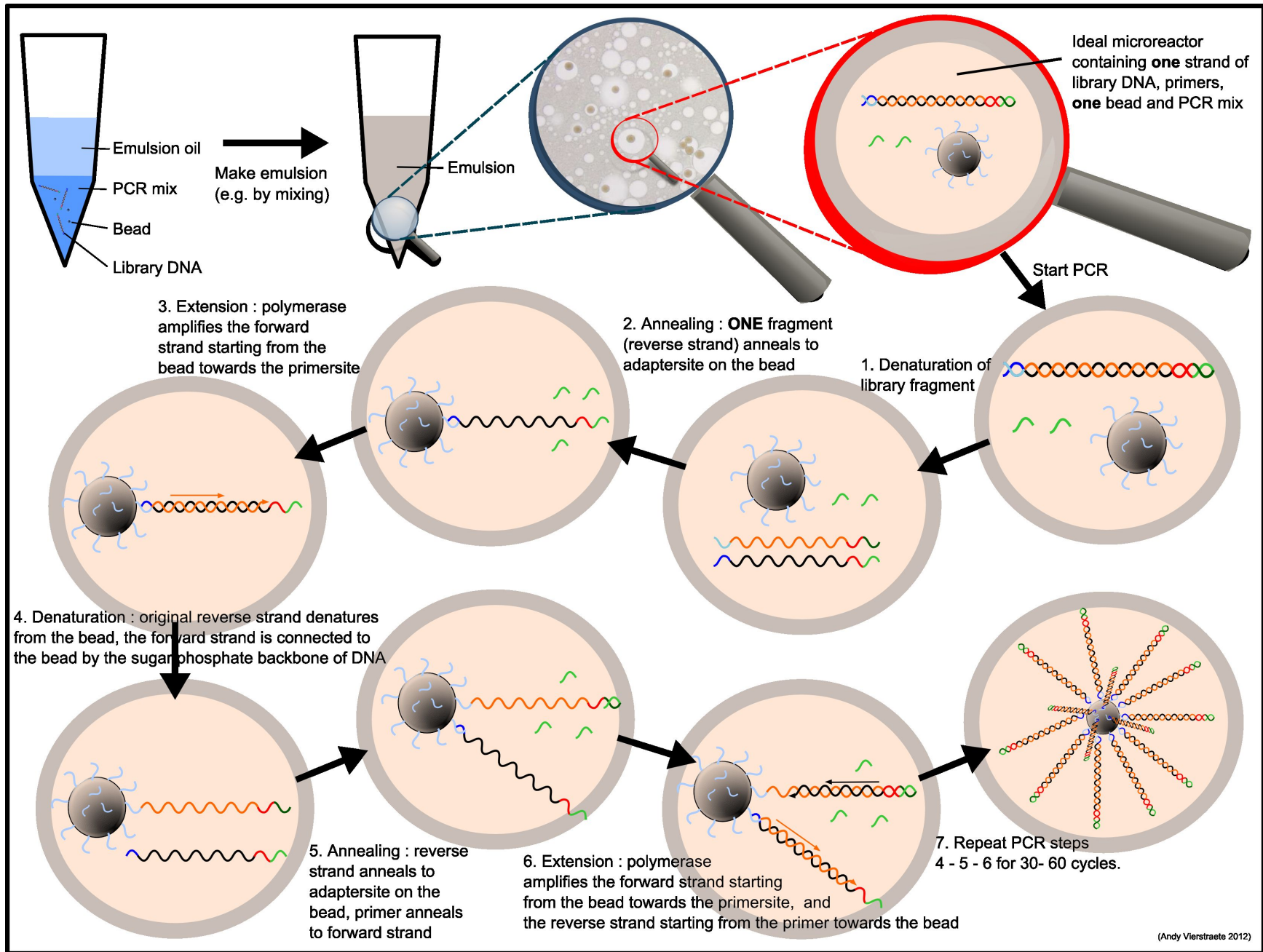
Workflow

Next Generation Sequencing: Amplified Single Molecule Sequencing



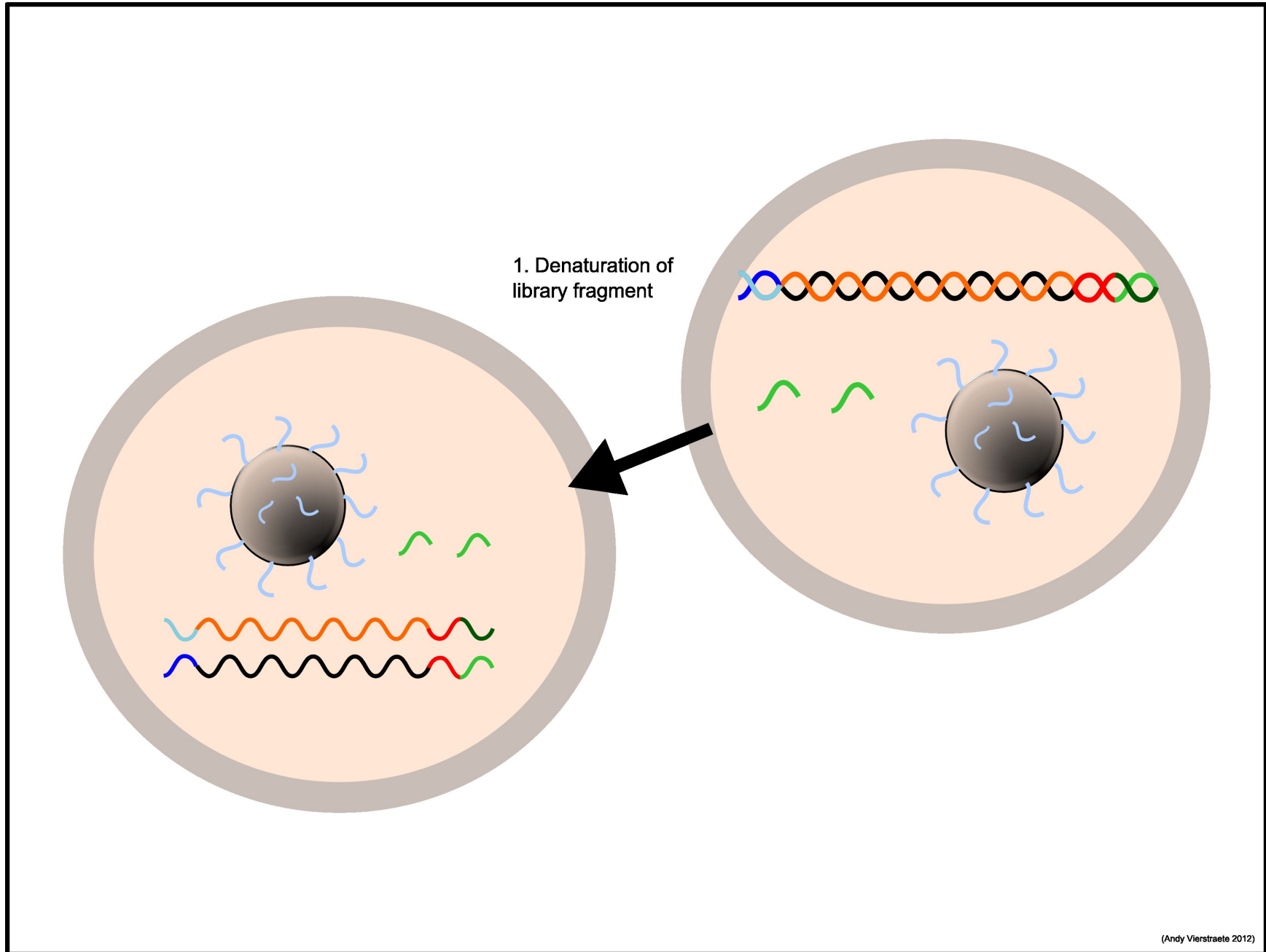
Workflow

Next Generation Sequencing: Amplified Single Molecule Sequencing Emulsion PCR



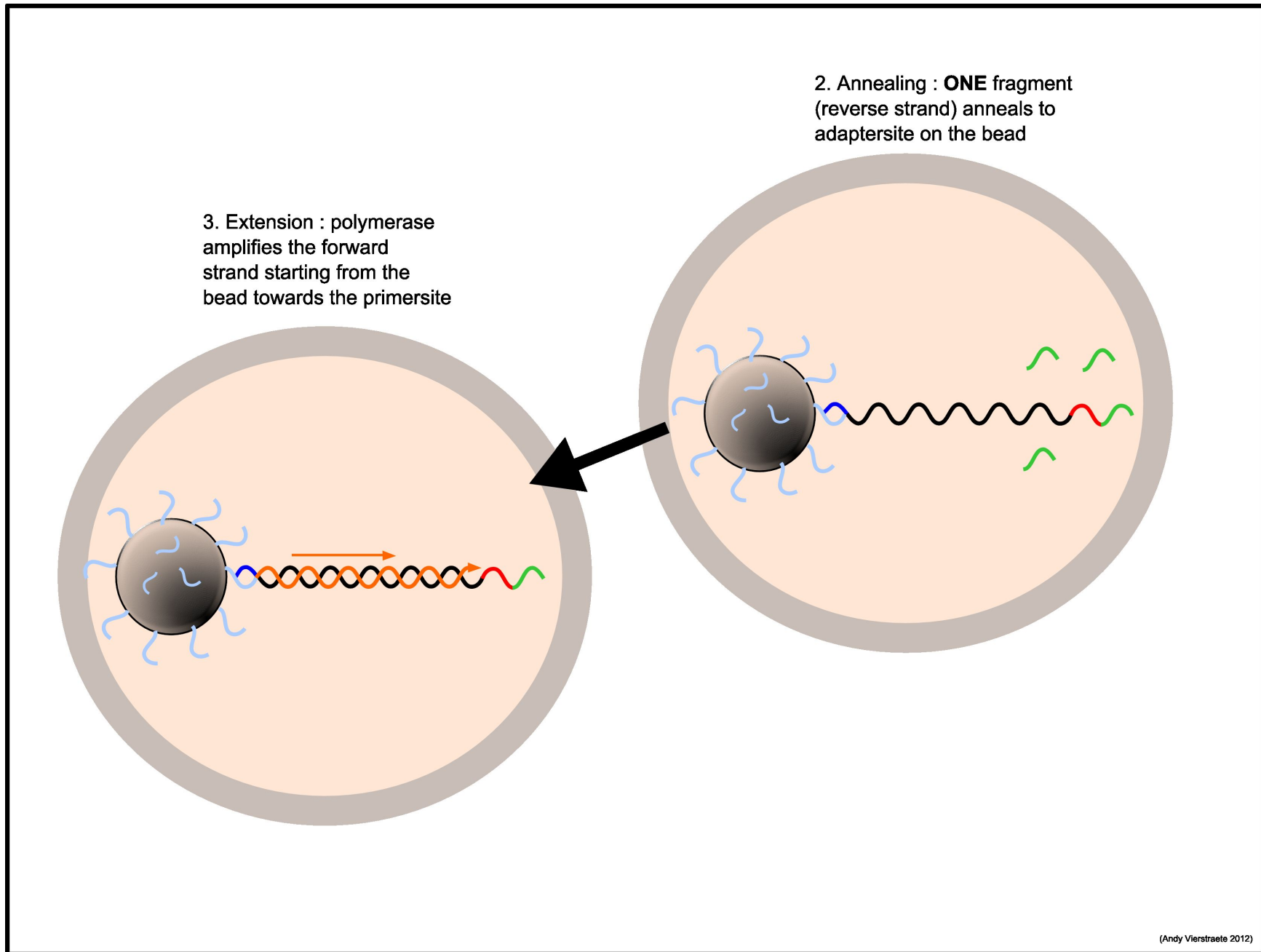
Workflow

Next Generation Sequencing: Amplified Single Molecule Sequencing Emulsion PCR



Workflow

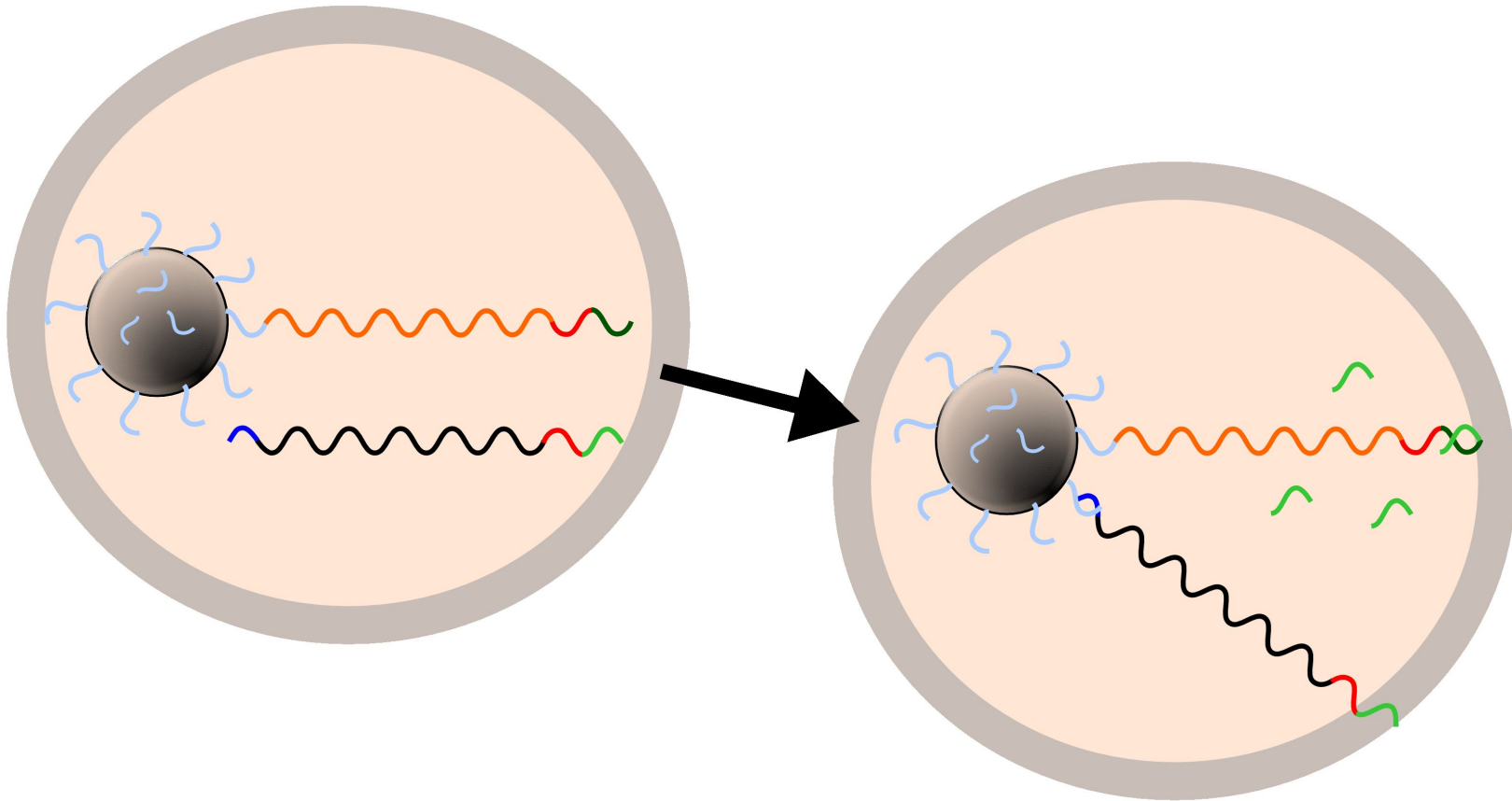
Next Generation Sequencing: Amplified Single Molecule Sequencing Emulsion PCR



Workflow

Next Generation Sequencing: Amplified Single Molecule Sequencing Emulsion PCR

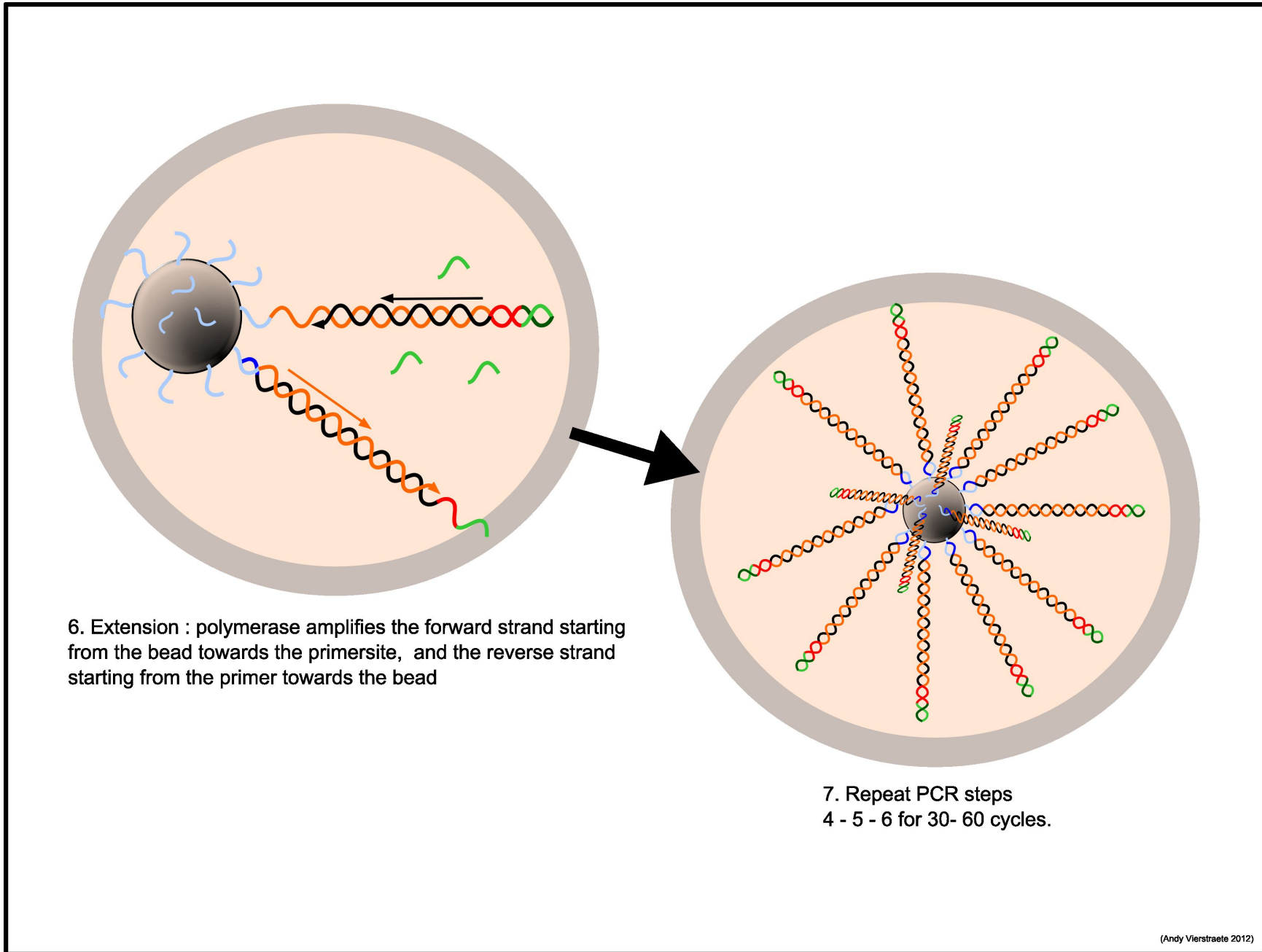
4. Denaturation : original reverse strand denatures from the bead, the forward strand is connected to the bead by the sugar phosphate backbone of DNA



5. Annealing : reverse strand anneals to adaptersite on the bead, primer anneals to forward strand

Workflow

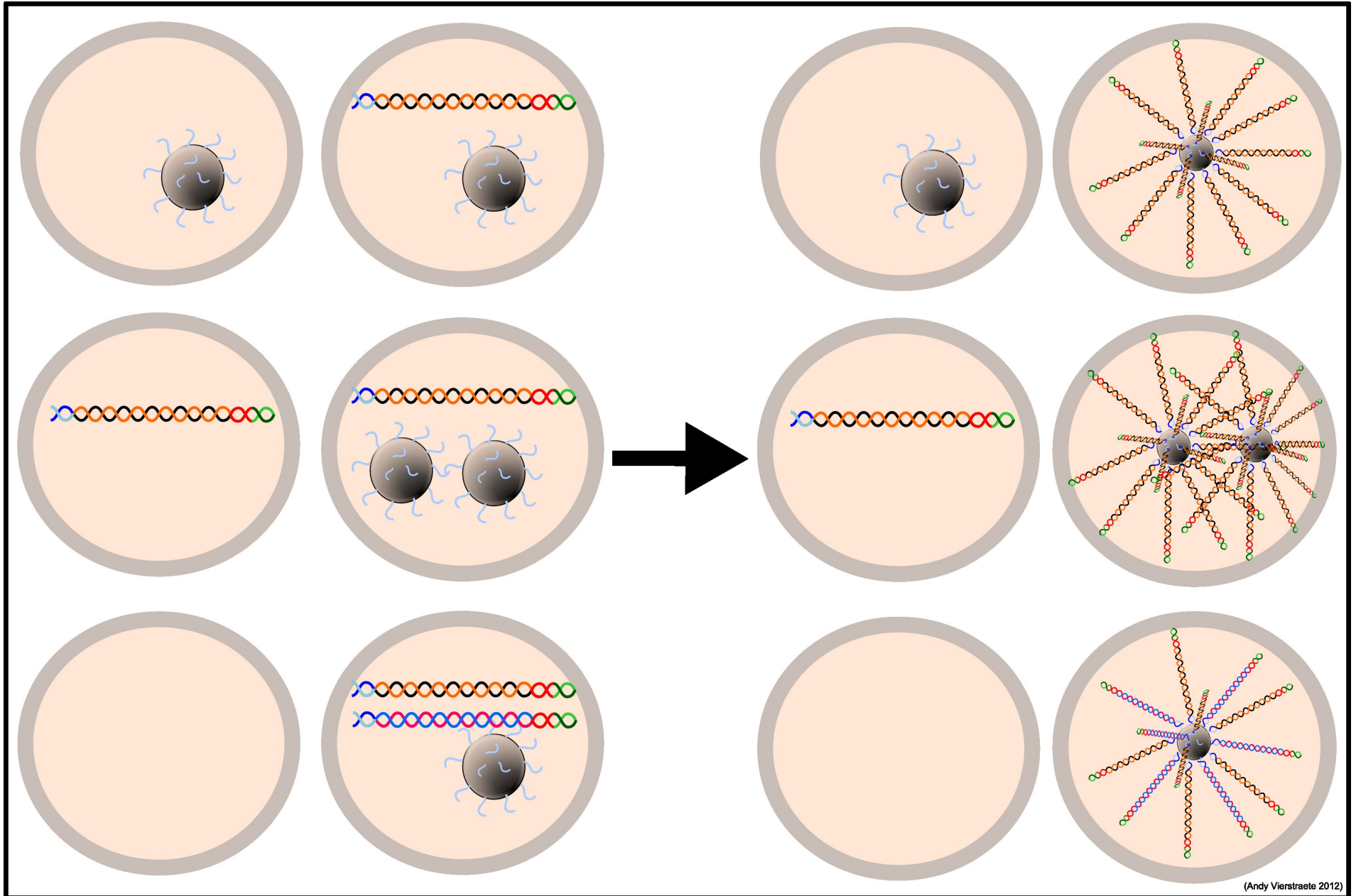
Next Generation Sequencing: Amplified Single Molecule Sequencing Emulsion PCR



Workflow

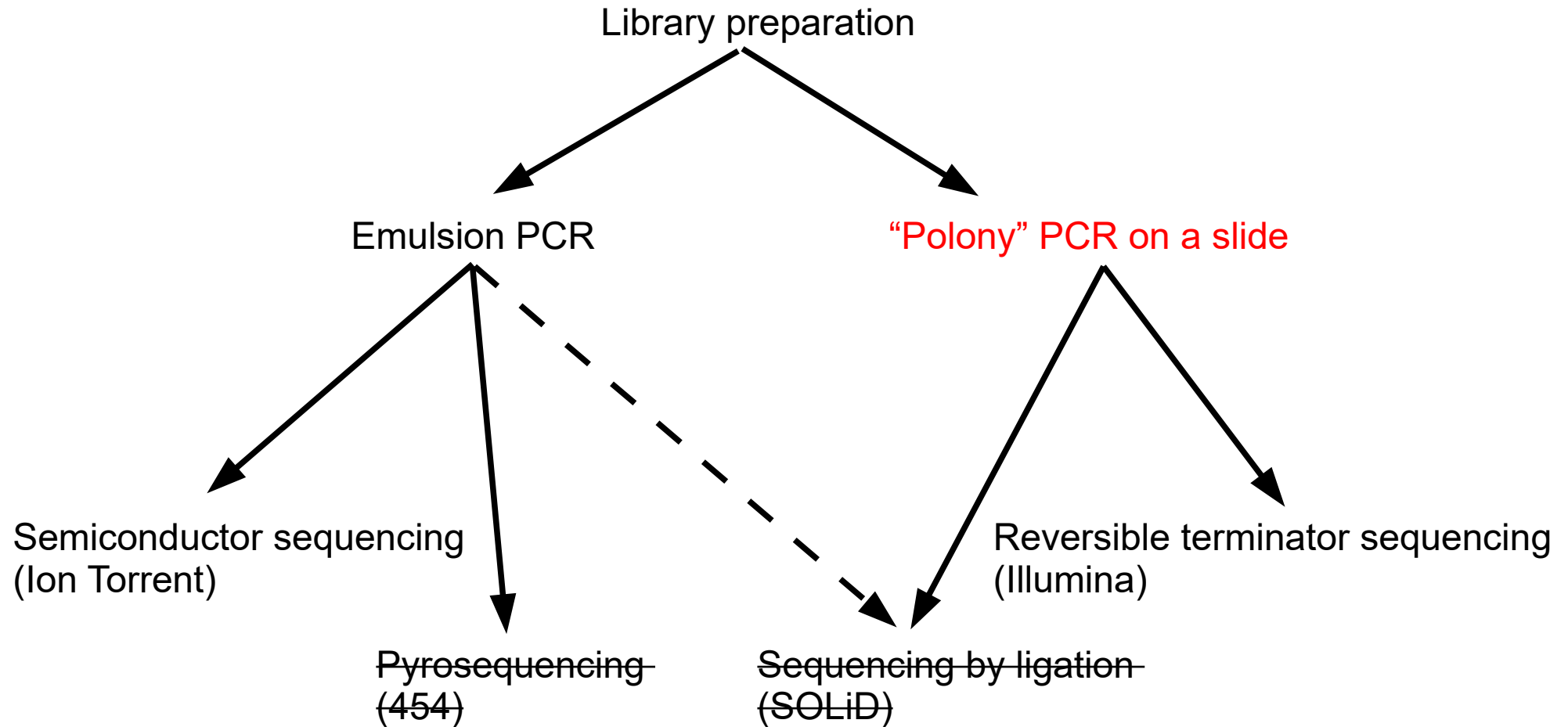
Next Generation Sequencing: Amplified Single Molecule Sequencing **Emulsion PCR**

different micro reactors: only 15 % are good ones



Workflow

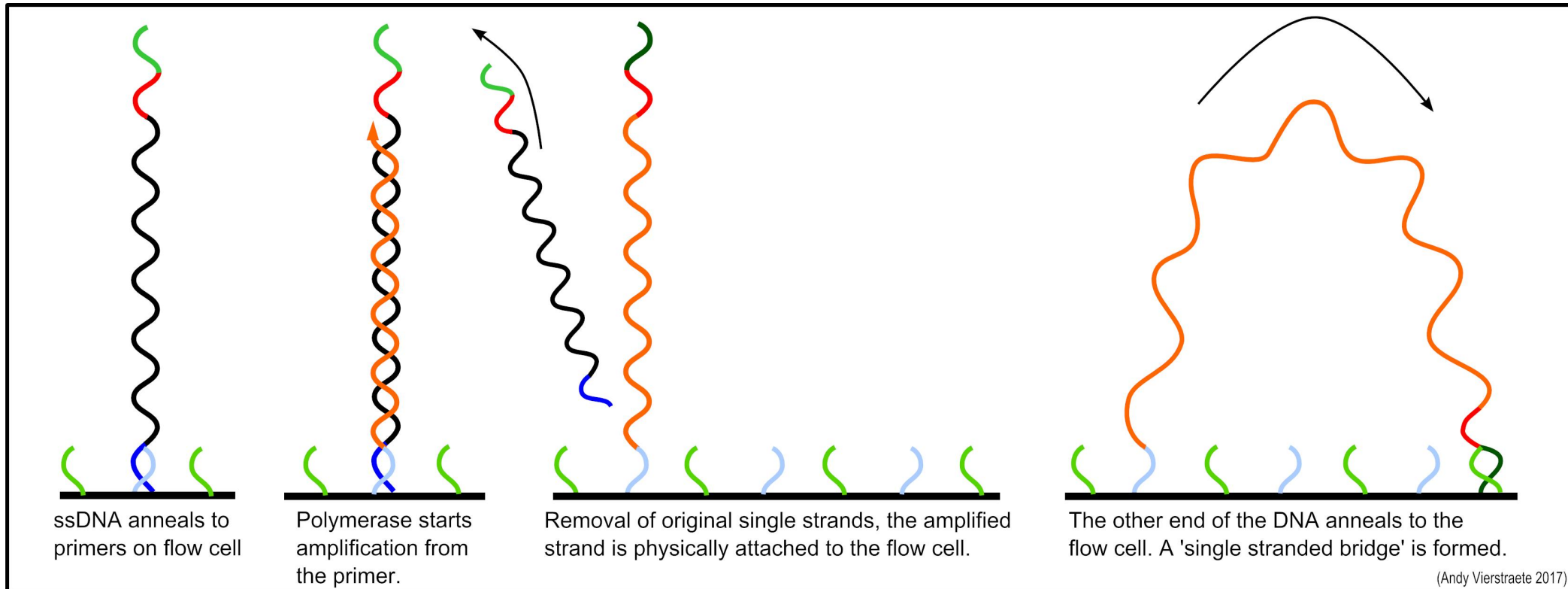
Next Generation Sequencing: Amplified Single Molecule Sequencing



Workflow

Next Generation Sequencing: Amplified Single Molecule Sequencing “Polony” PCR

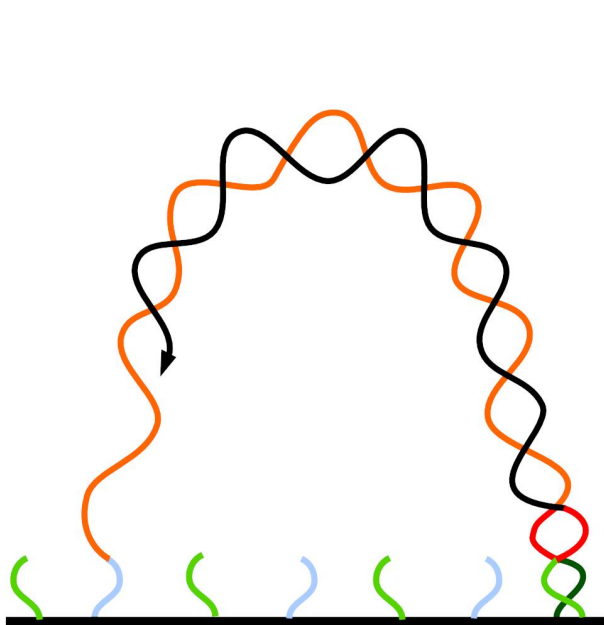
Bridge amplification: Illumina



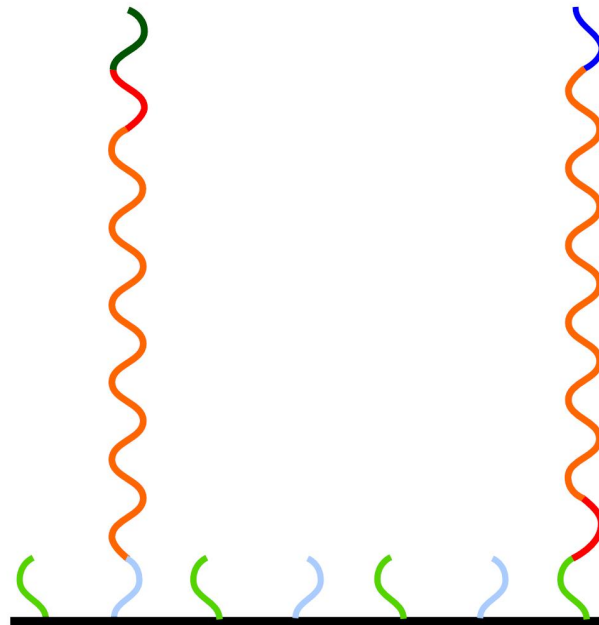
Workflow

Next Generation Sequencing: Amplified Single Molecule Sequencing "Polony" PCR

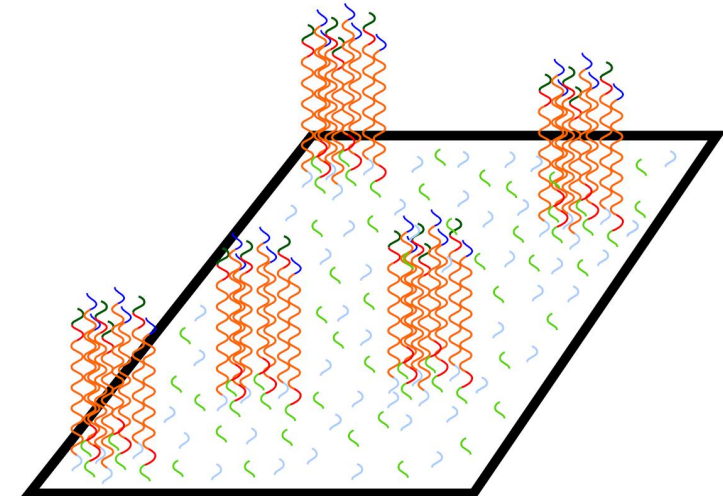
Bridge amplification: Illumina



Polymerase starts amplification from the annealed side. A 'double stranded bridge' is formed.



Denaturation of the 'double stranded bridge'. The two complementary strands are attached to the flow cell.

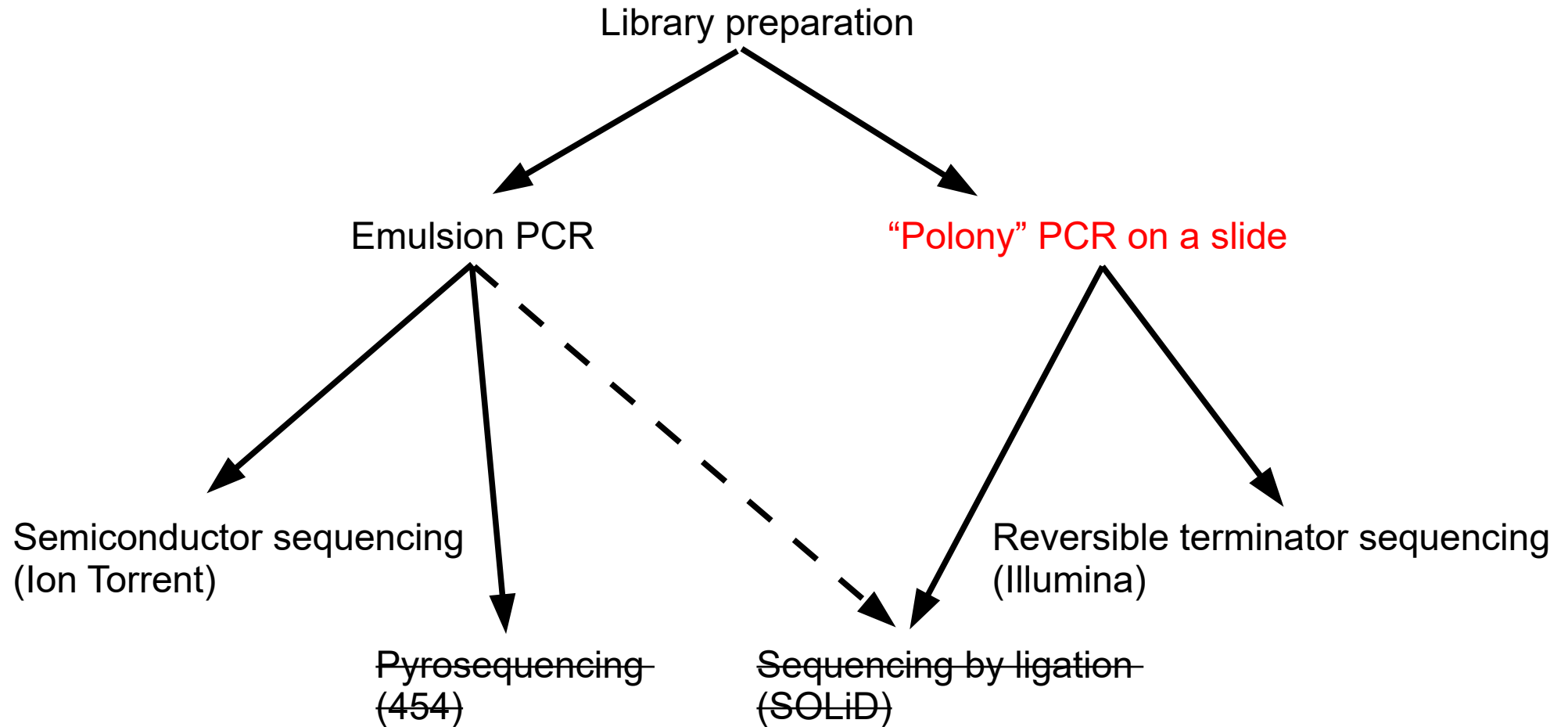


At the end of the amplification, millions of clusters are formed on the flow cell.

(Andy Vierstraete 2017)

Workflow

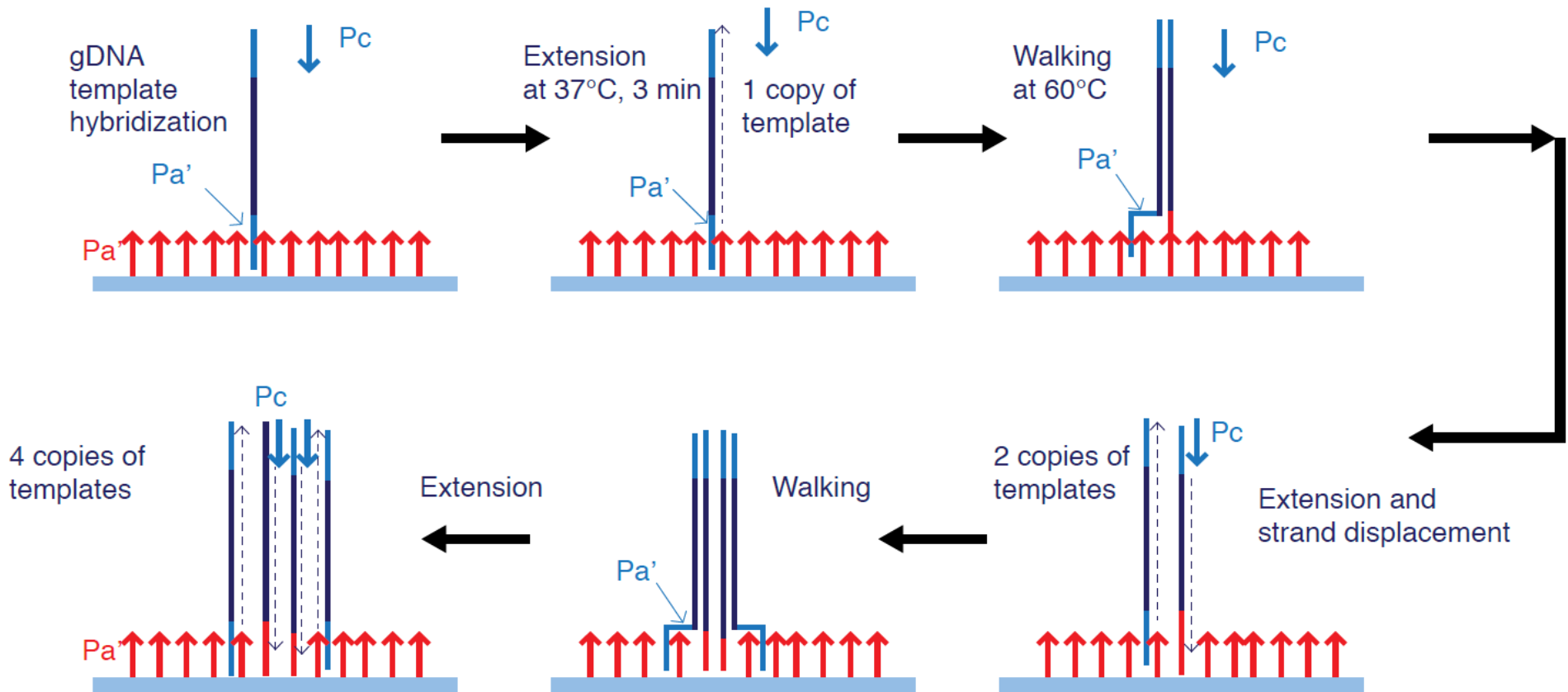
Next Generation Sequencing: Amplified Single Molecule Sequencing



Workflow

Next Generation Sequencing: Amplified Single Molecule Sequencing “Polony” PCR

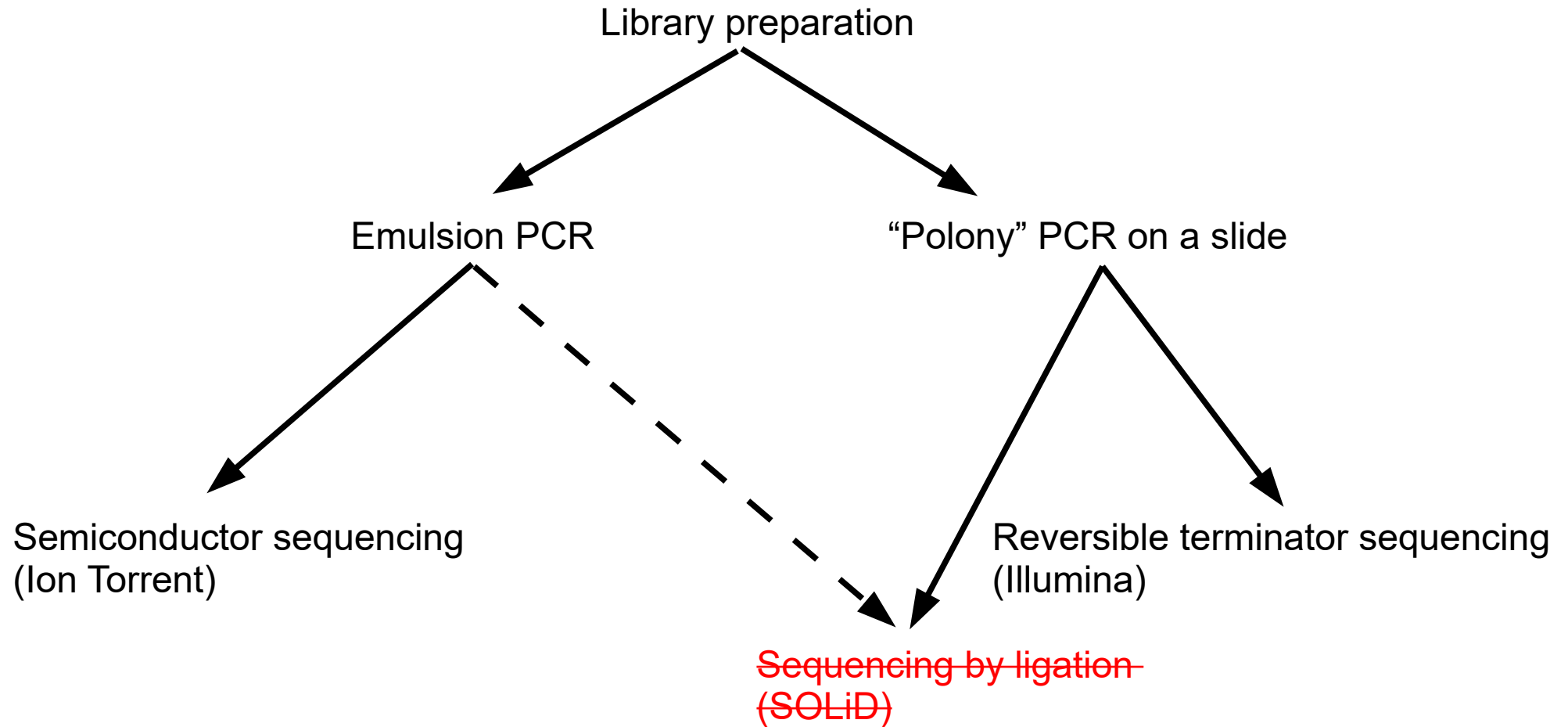
Wildfire amplification: SOLiD



Wildfire chemistry schematic.

Workflow

Next Generation Sequencing: Amplified Single Molecule Sequencing

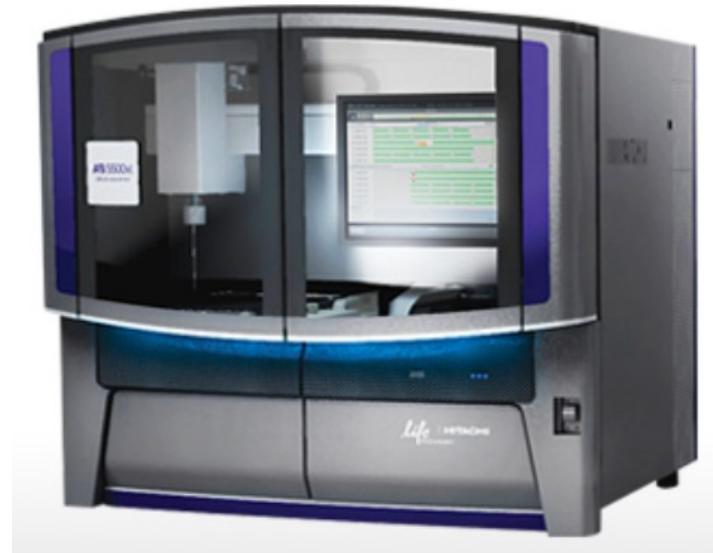


Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing

SOLiD

5500 W SOLiD Sequencer (end of support Dec 2017)



	5500 W	5500 xl W
Read Length	75 bp or 2 x 50 bp	75 bp or 2 x 50 bp
Throughput	120 - 160 Gb	240 - 320 Gb
Reads per run	1,2 Billion	2,4 Billion
Accuracy	99,99 %	99,99 %
Run Time	7 days	7 days

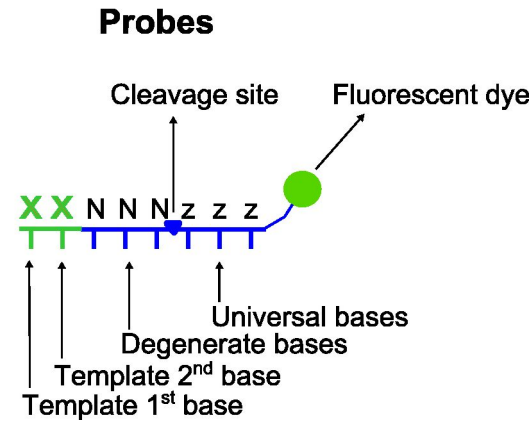
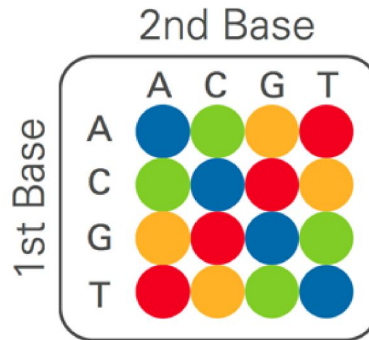
Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing

SOLiD

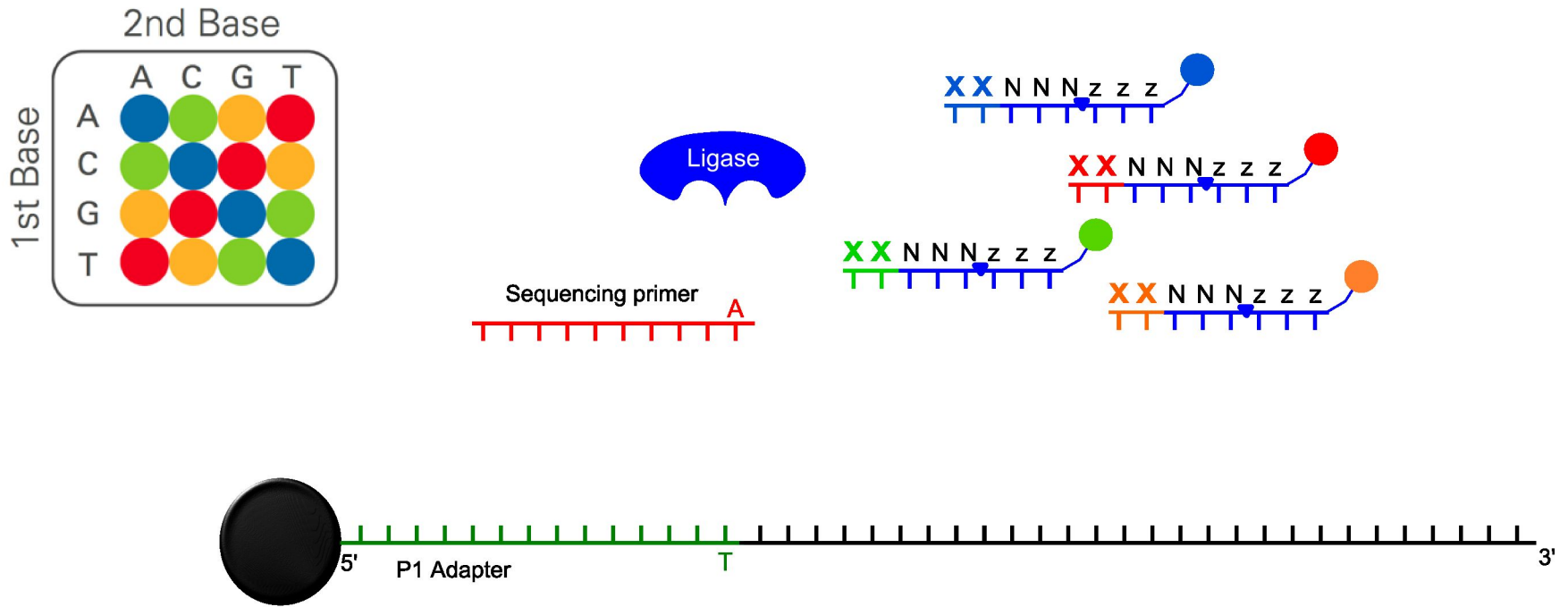
Workflow: Library preparation → Wildfire PCR → Sequencing by Ligation

dual base encoding



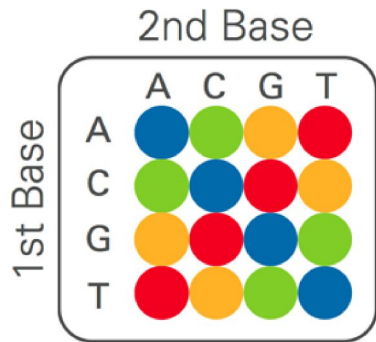
Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing SOLiD

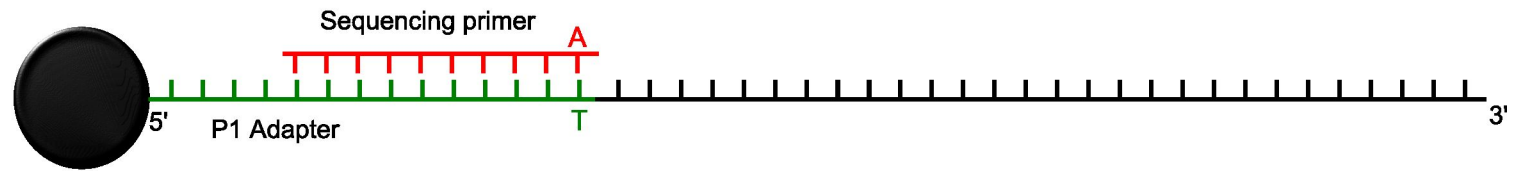
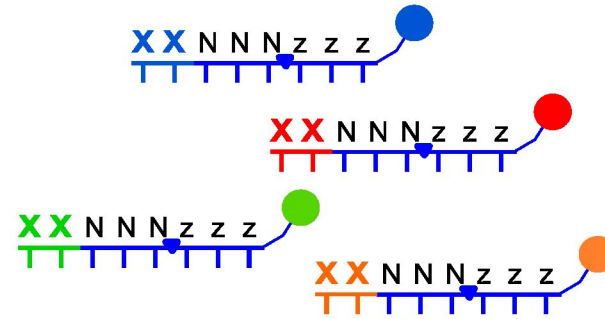


Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing SOLiD



Ligase



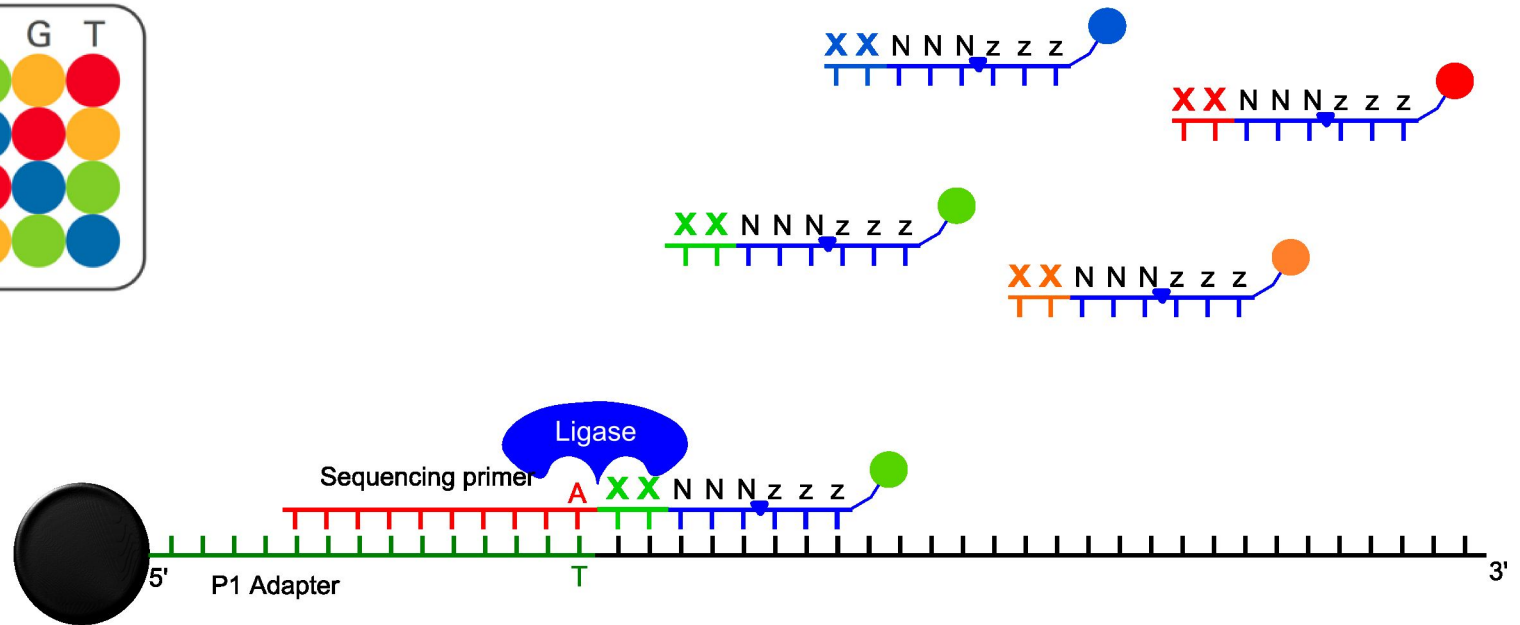
Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing

SOLiD

2nd Base

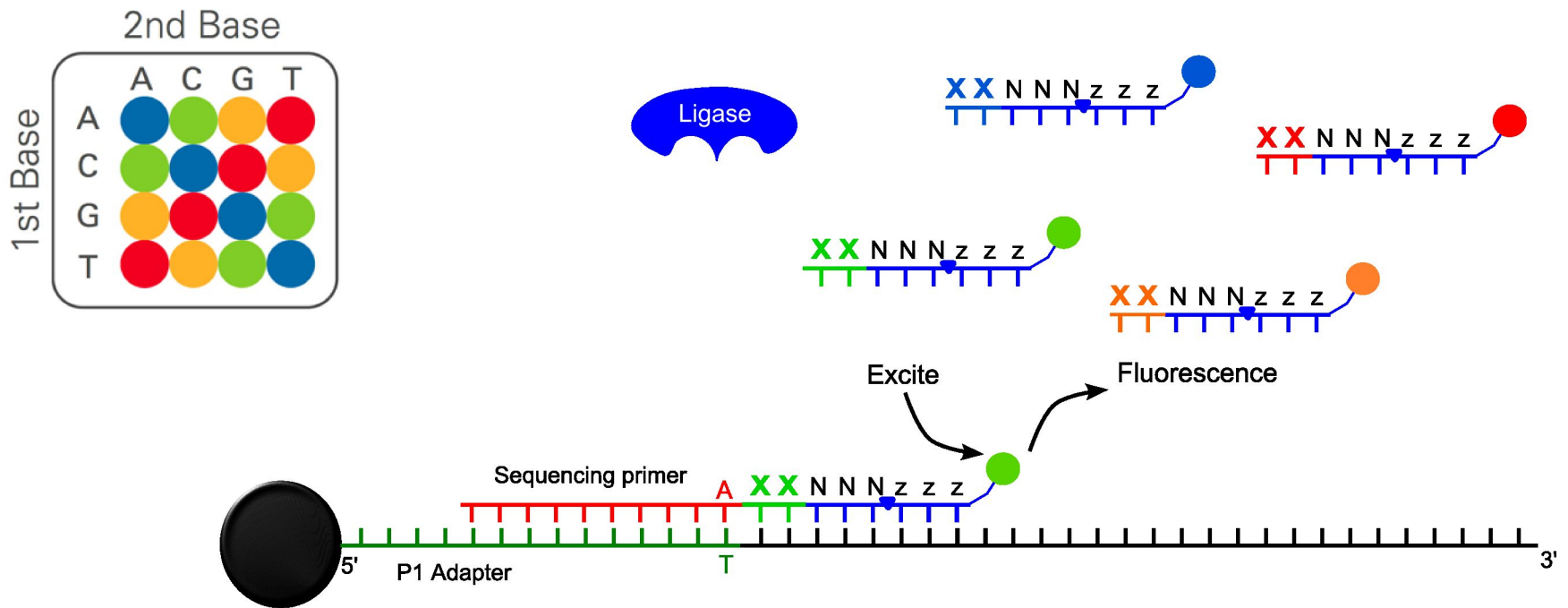
	A	C	G	T
1st Base	A	C	G	T
	C	A	T	C
	G	T	C	A
	T	C	A	G



Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing

SOLiD

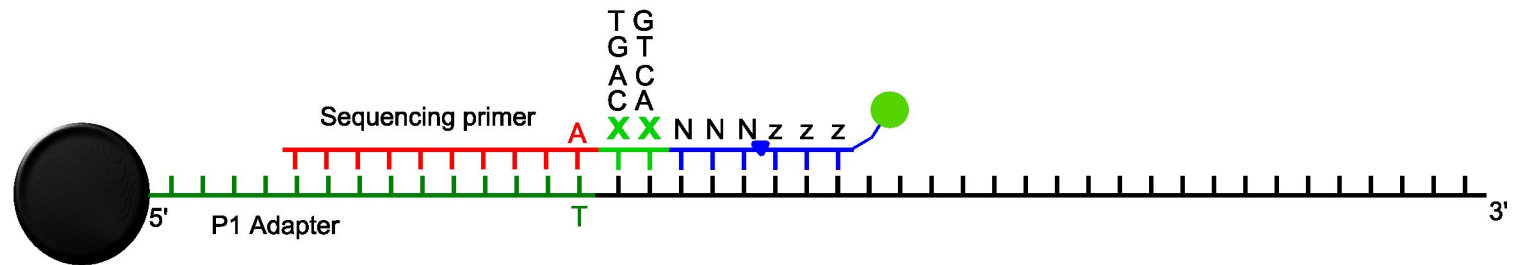
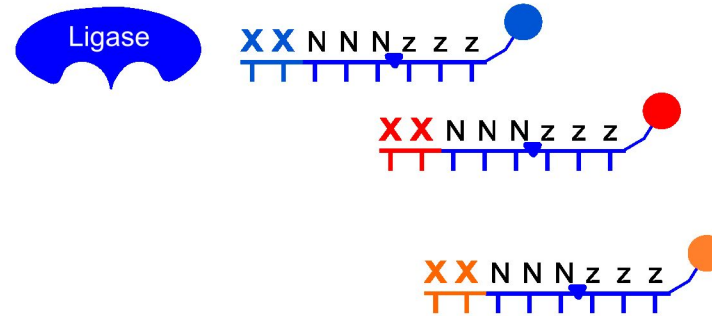


Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing

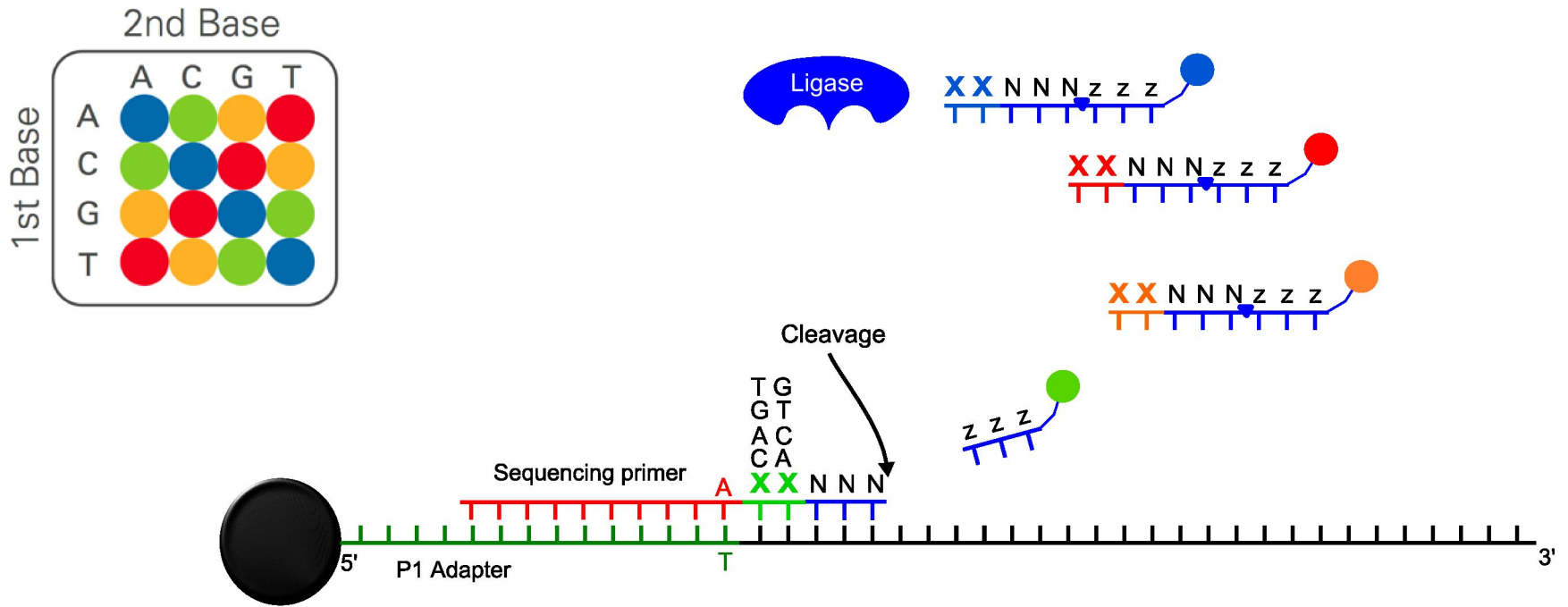
SOLiD

		2nd Base			
		A	C	G	T
1st Base	A	●	●	●	●
	C	●	●	●	●
	G	●	●	●	●
	T	●	●	●	●



Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing SOLiD



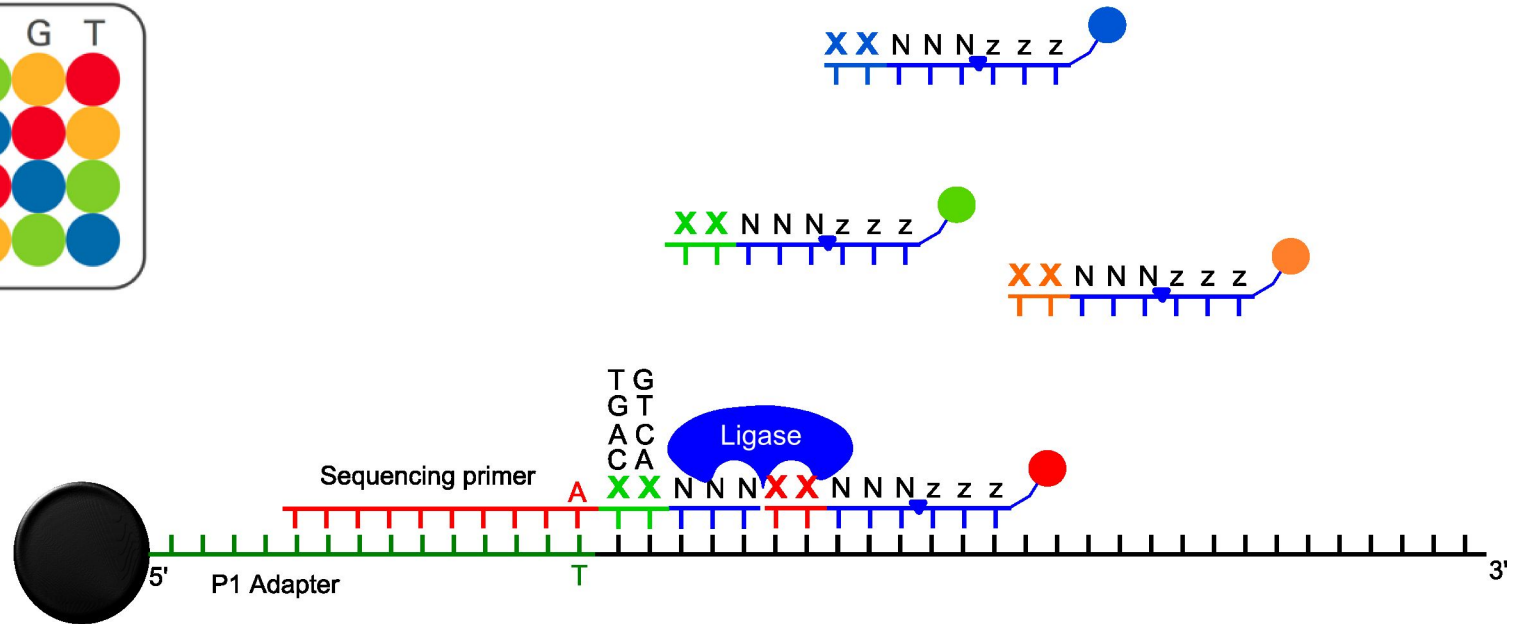
Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing

SOLiD

2nd Base

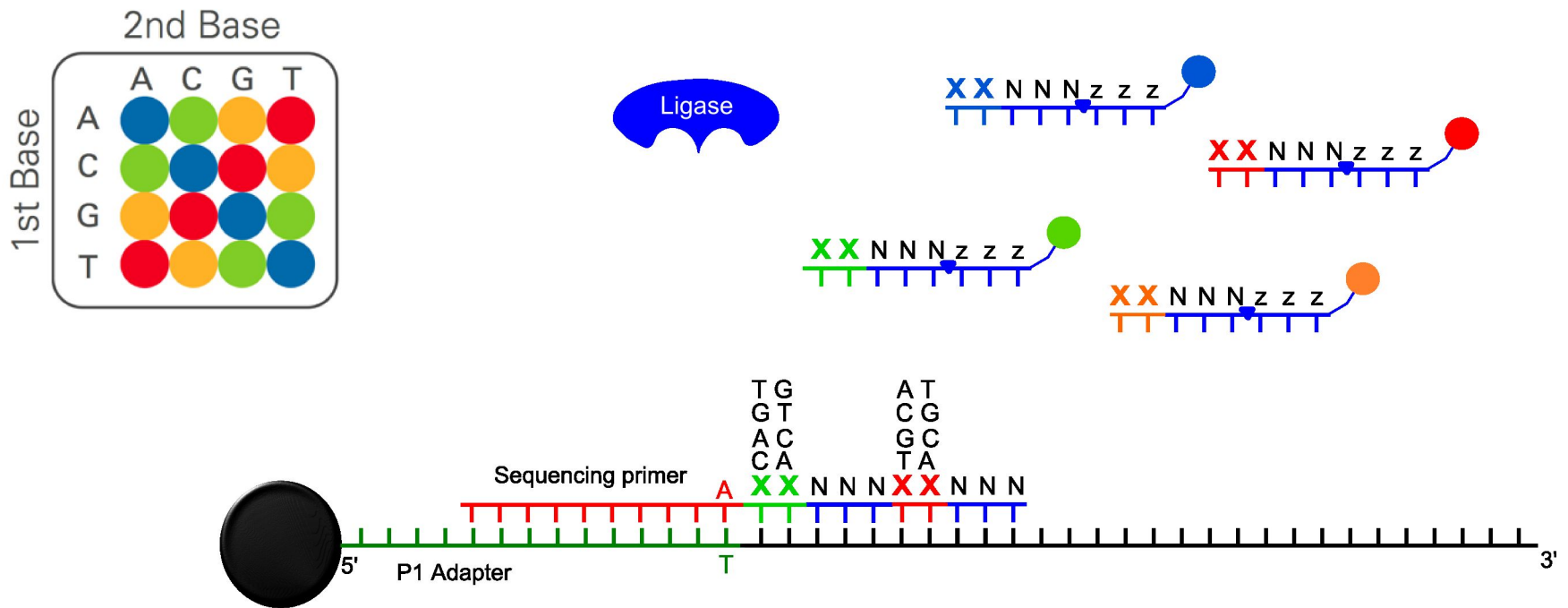
	A	C	G	T
1st Base	A	C	G	T
	C	A	T	G
	G	T	C	A
	T	G	A	C



Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing

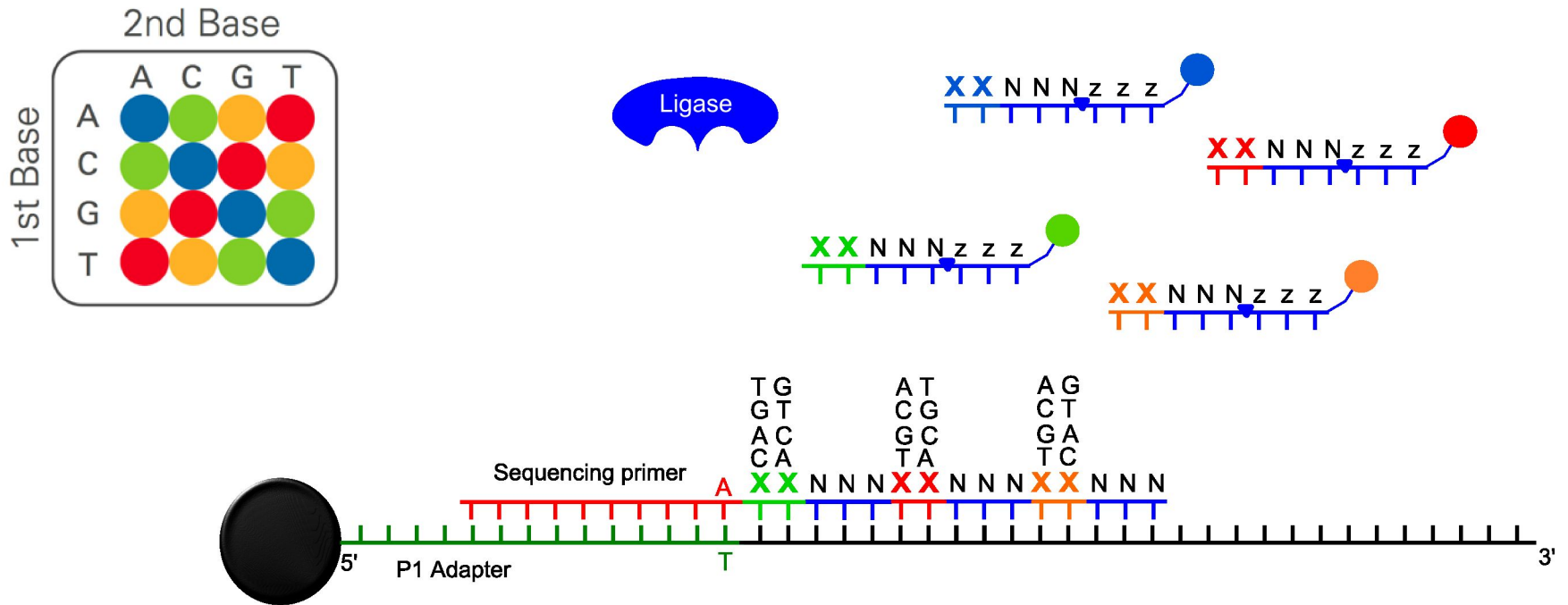
SOLiD



Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing

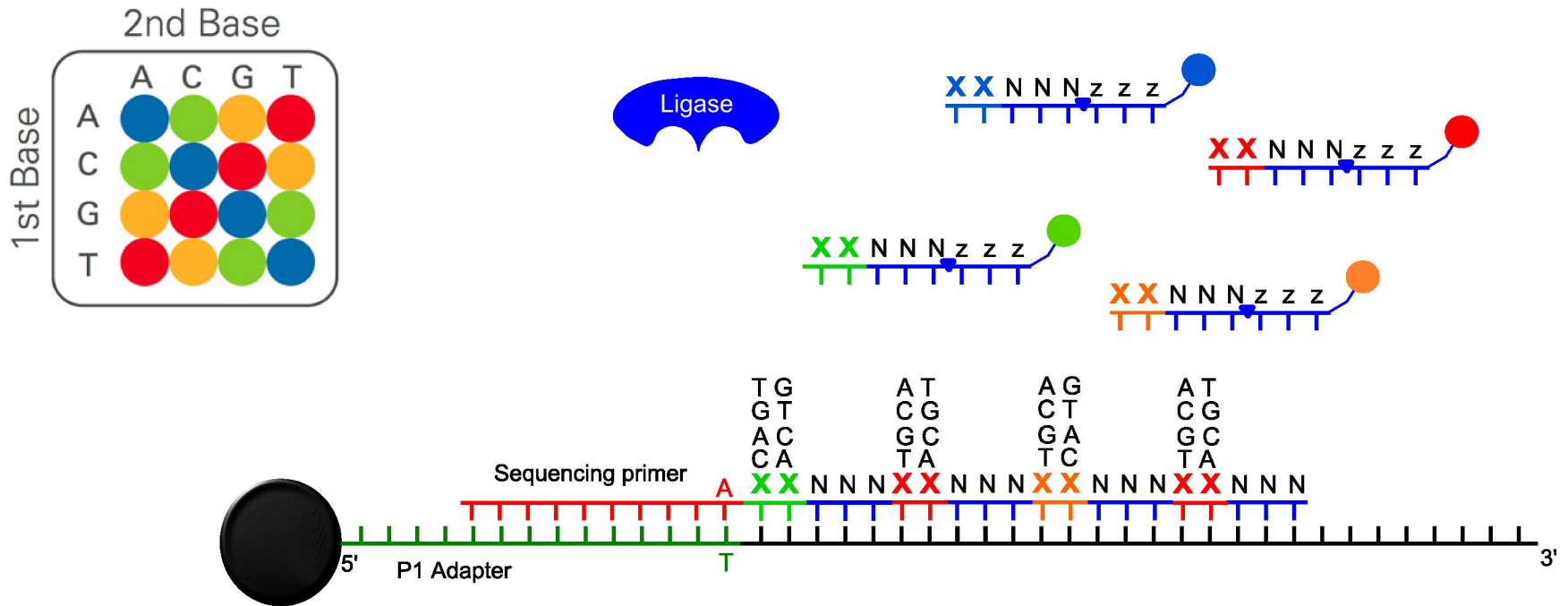
SOLiD



Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing

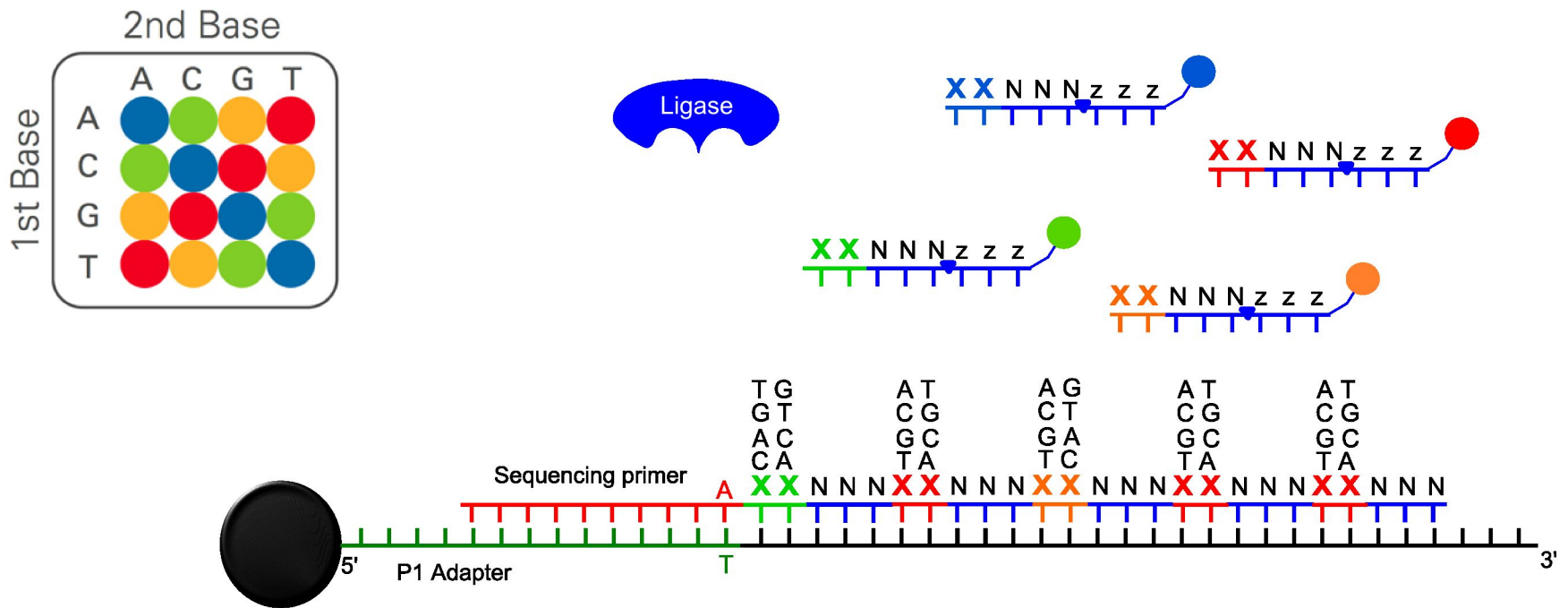
SOLiD



Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing

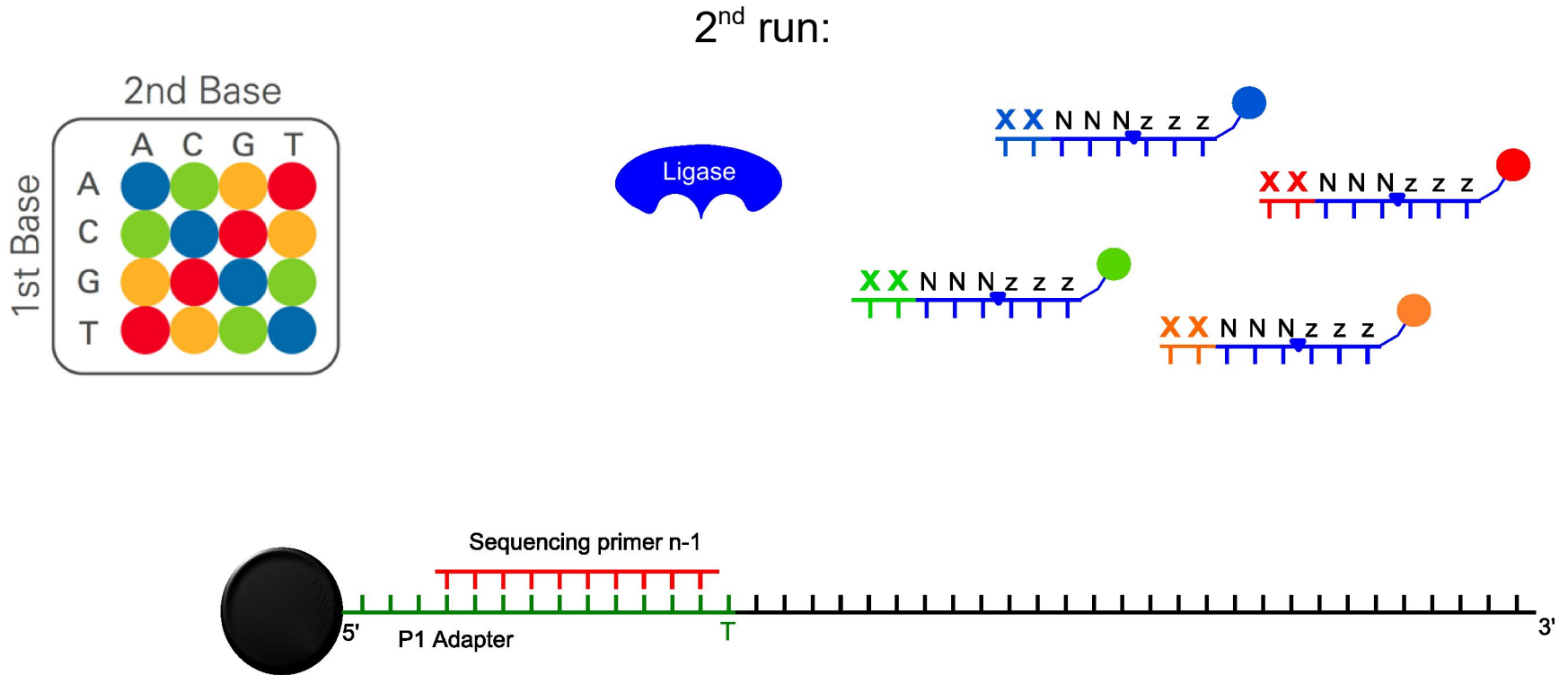
SOLiD



Different platforms

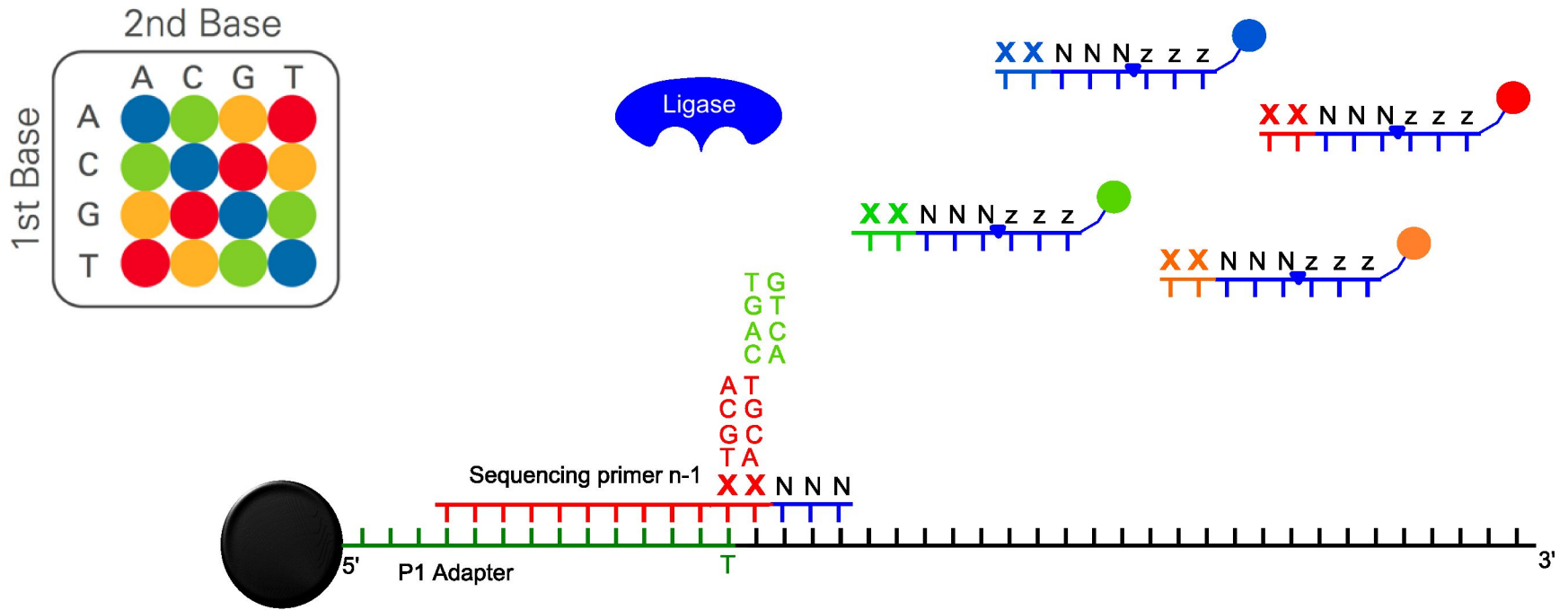
Next Generation Sequencing: Amplified Single Molecule Sequencing

SOLiD



Different platforms

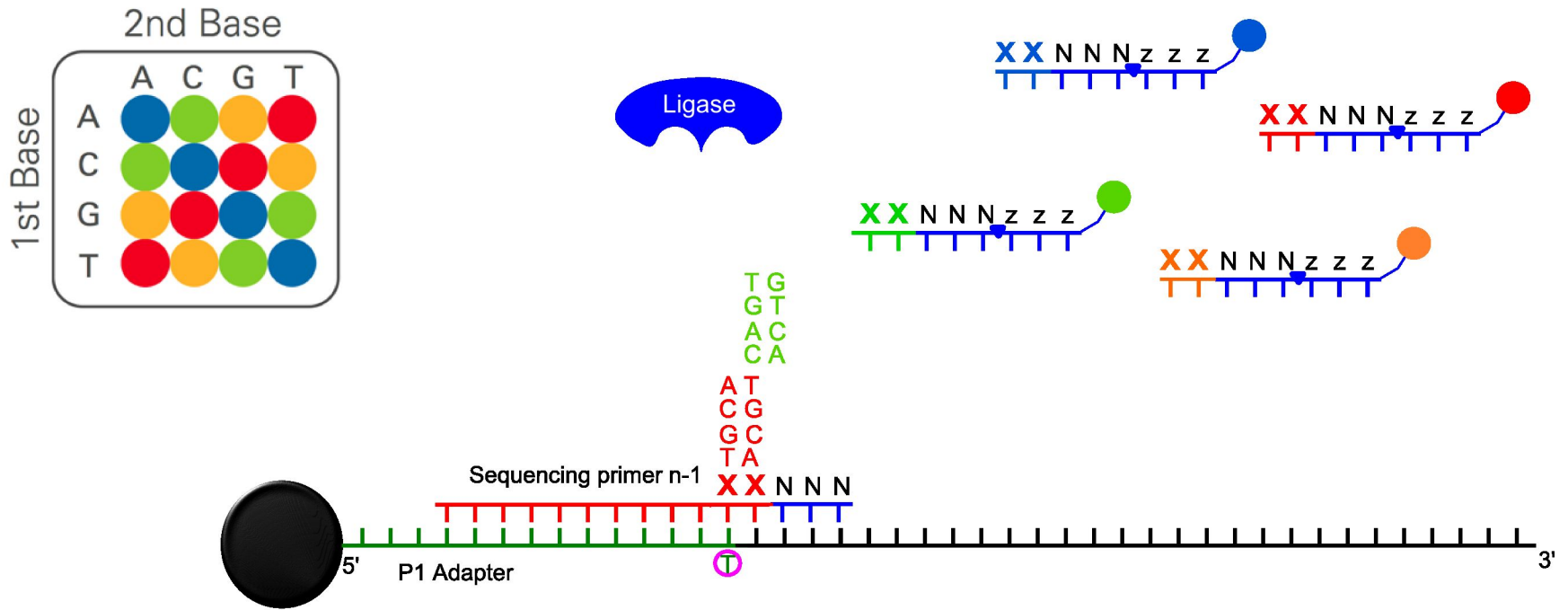
Next Generation Sequencing: Amplified Single Molecule Sequencing SOLiD



Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing

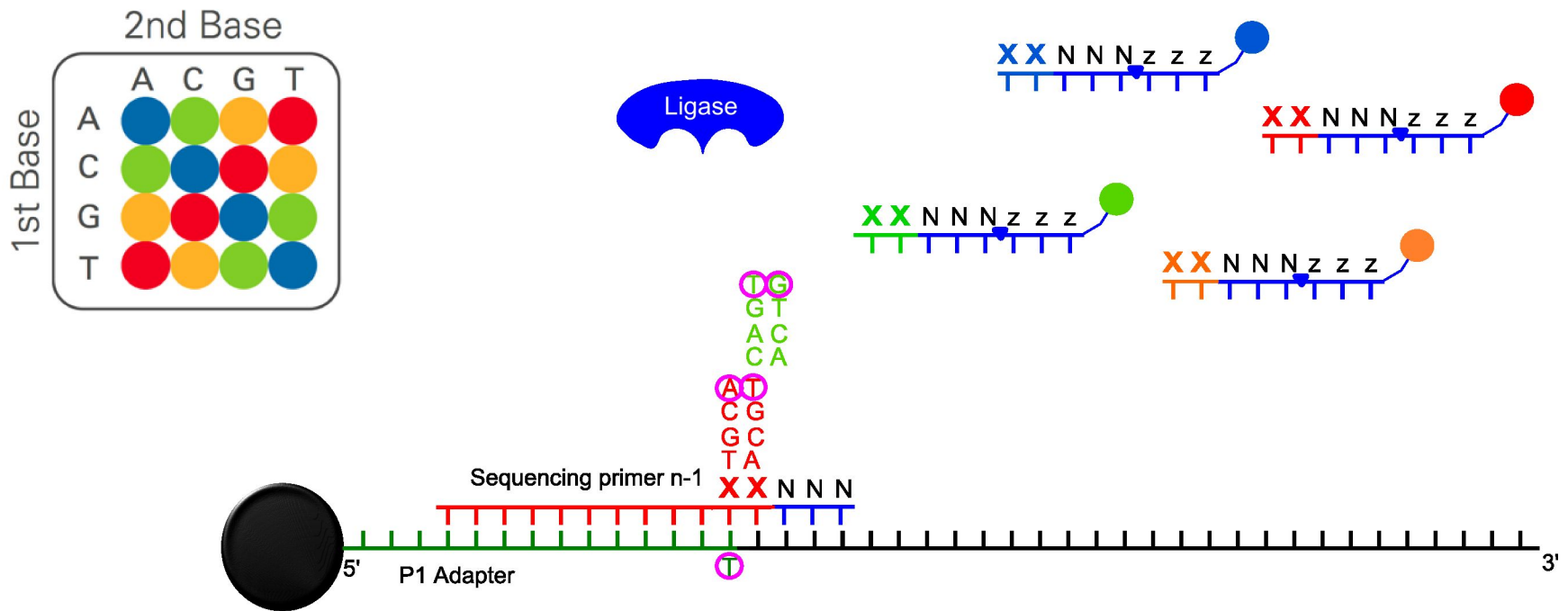
SOLiD



Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing

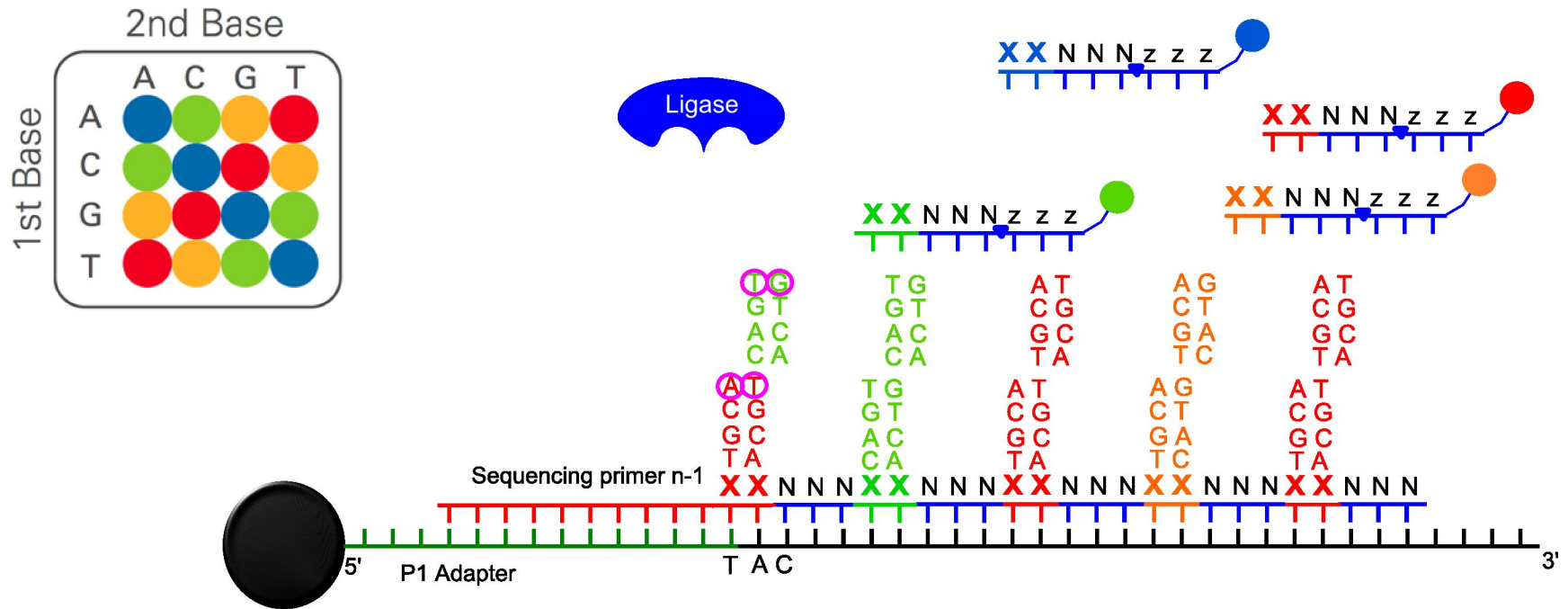
SOLiD



Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing

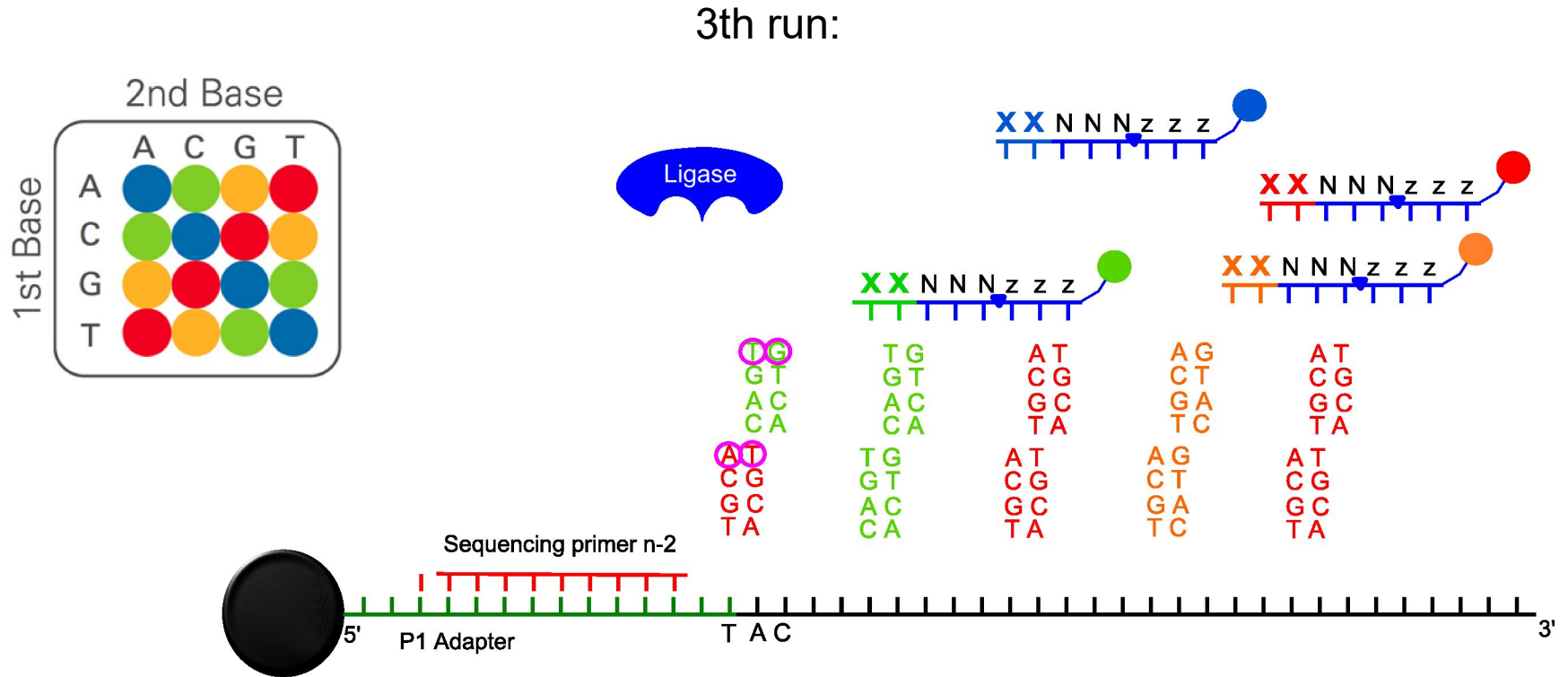
SOLiD



Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing

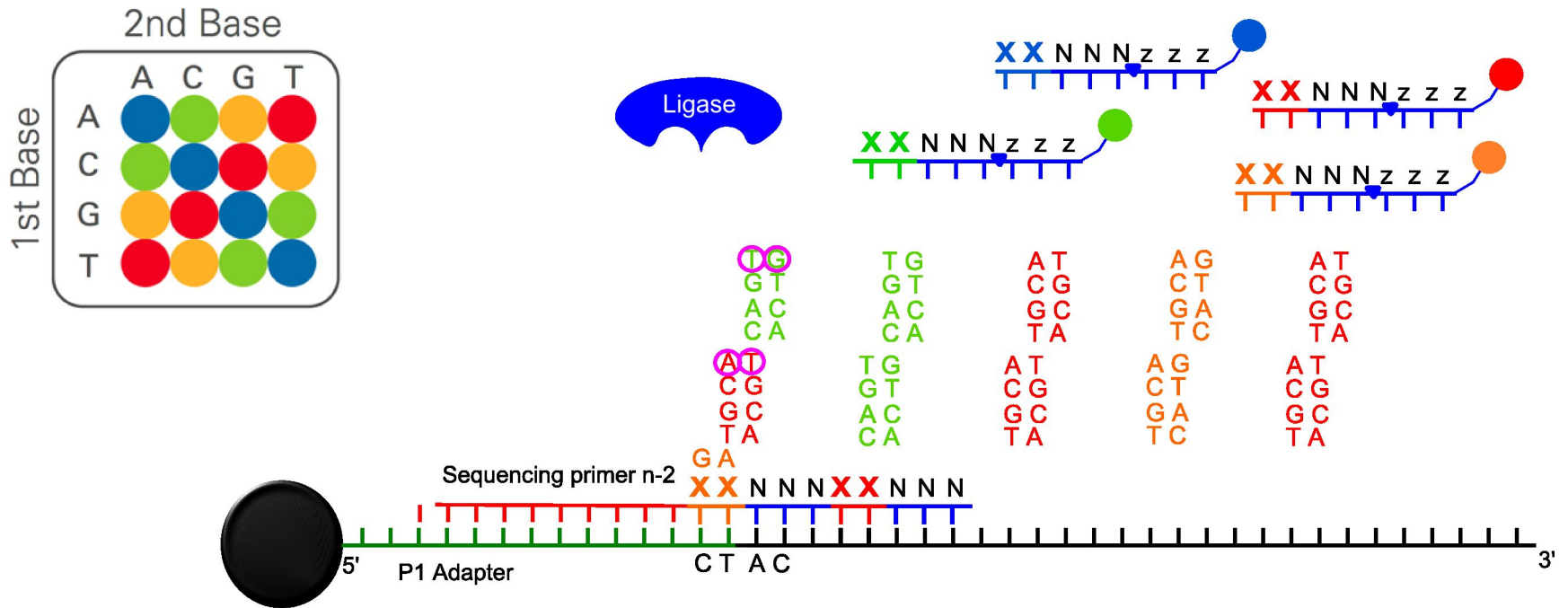
SOLiD



Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing

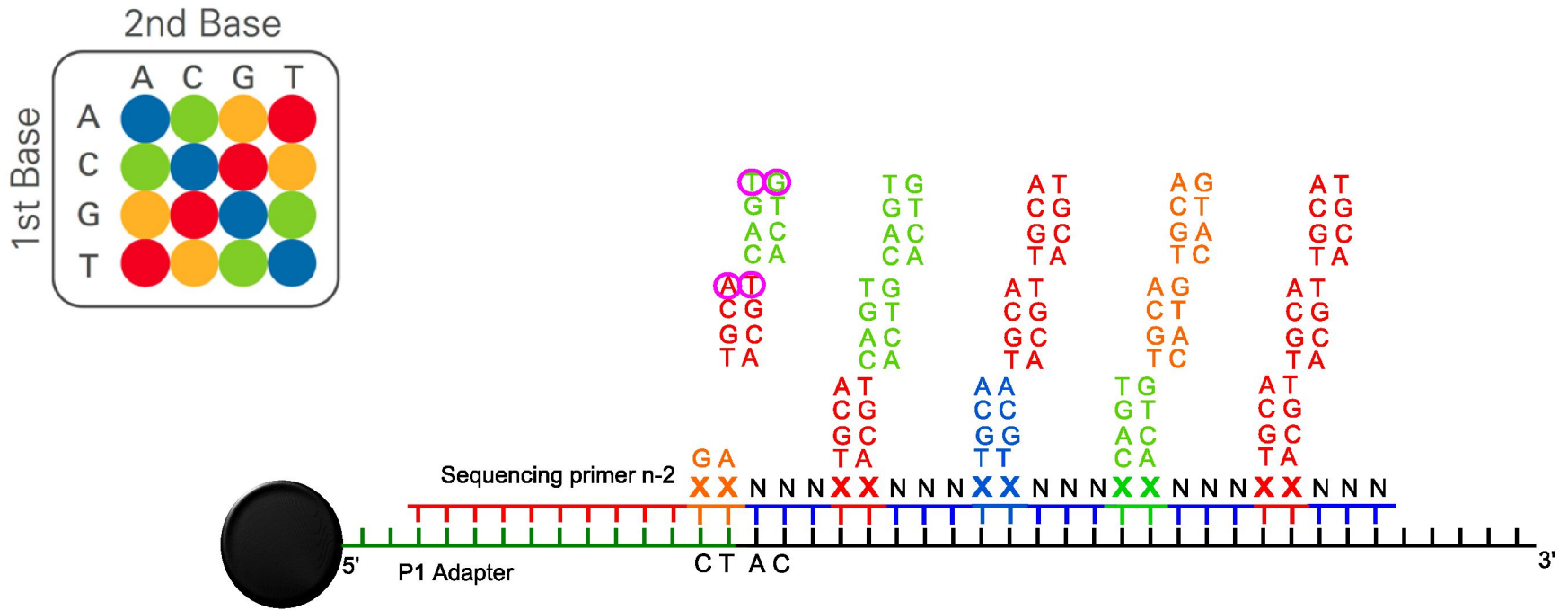
SOLiD



Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing

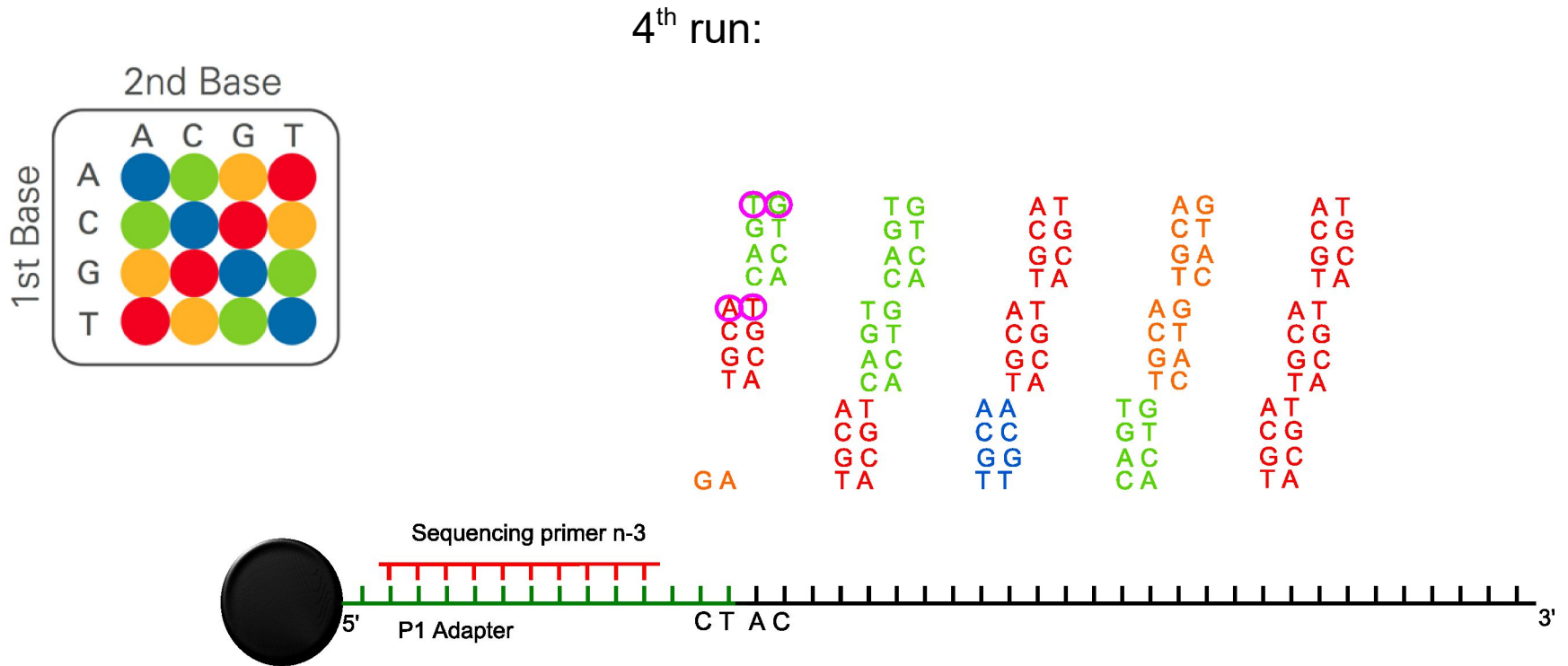
SOLiD



Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing

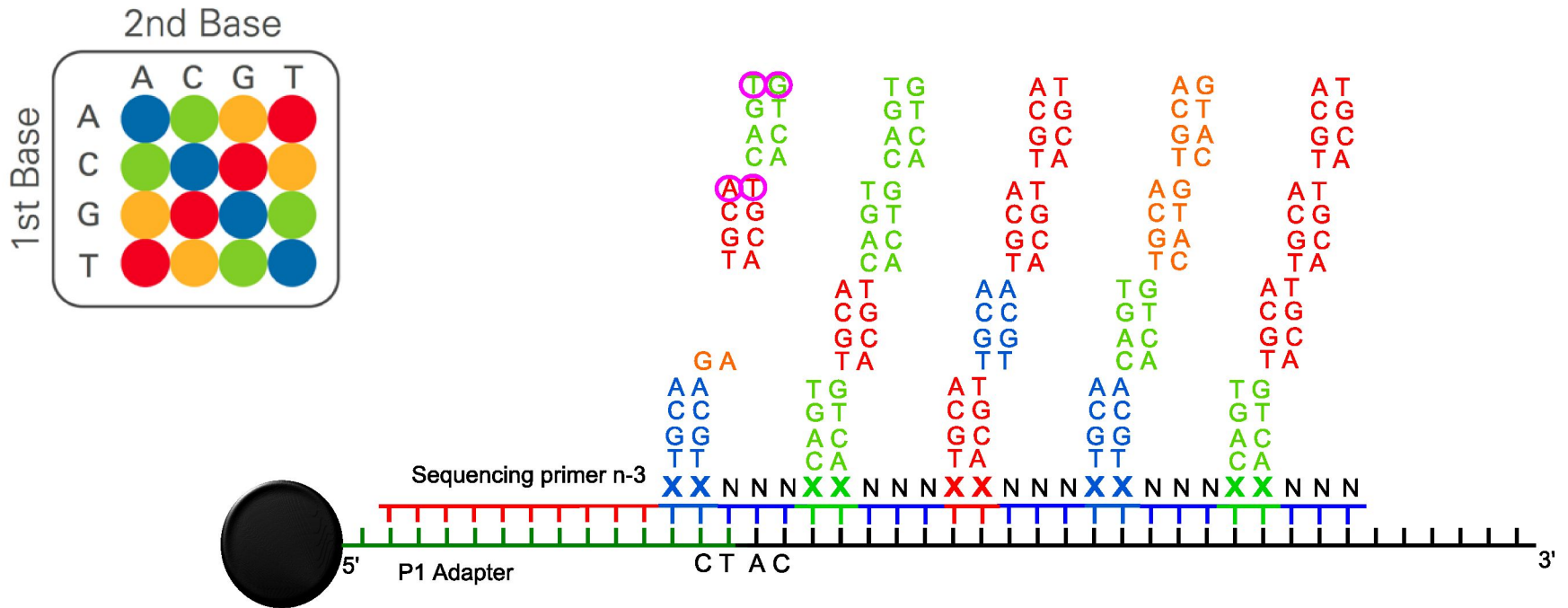
SOLiD



Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing

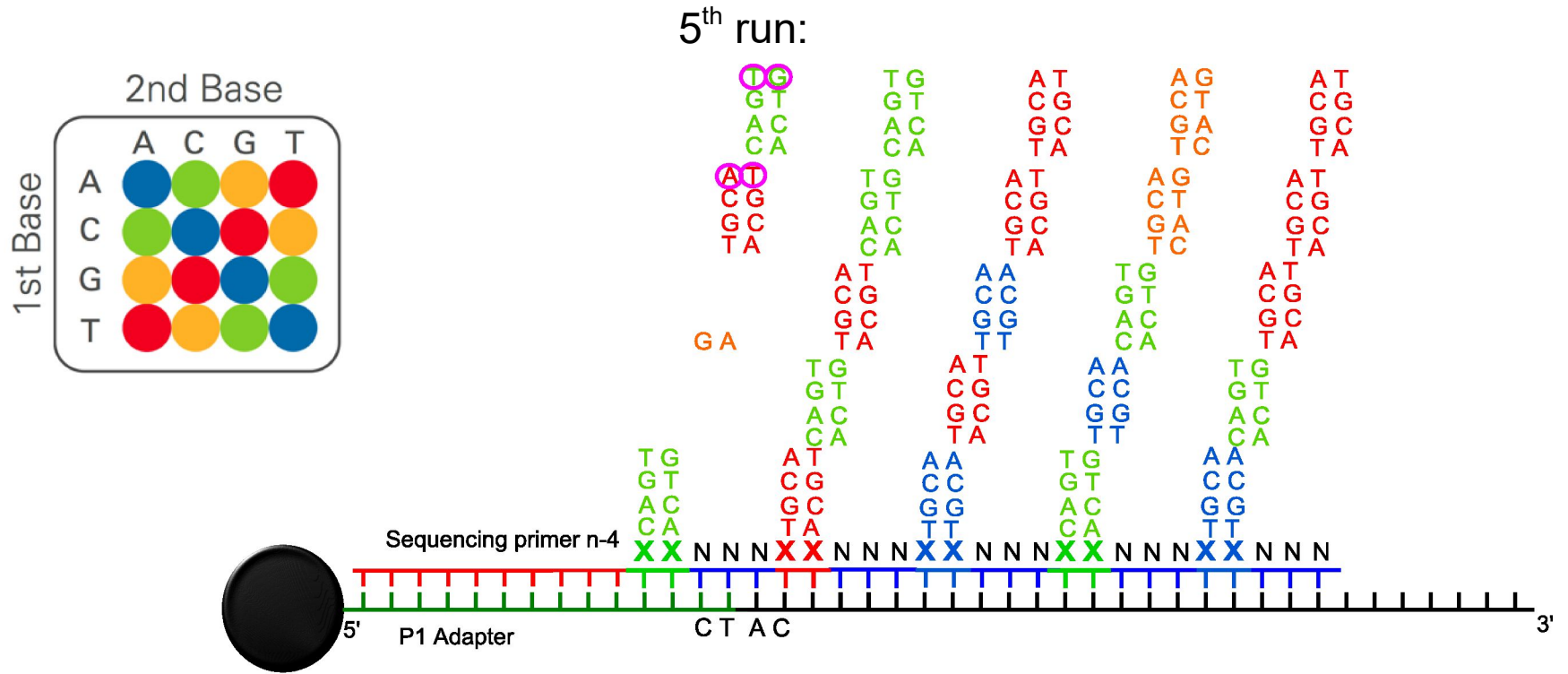
SOLiD



Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing

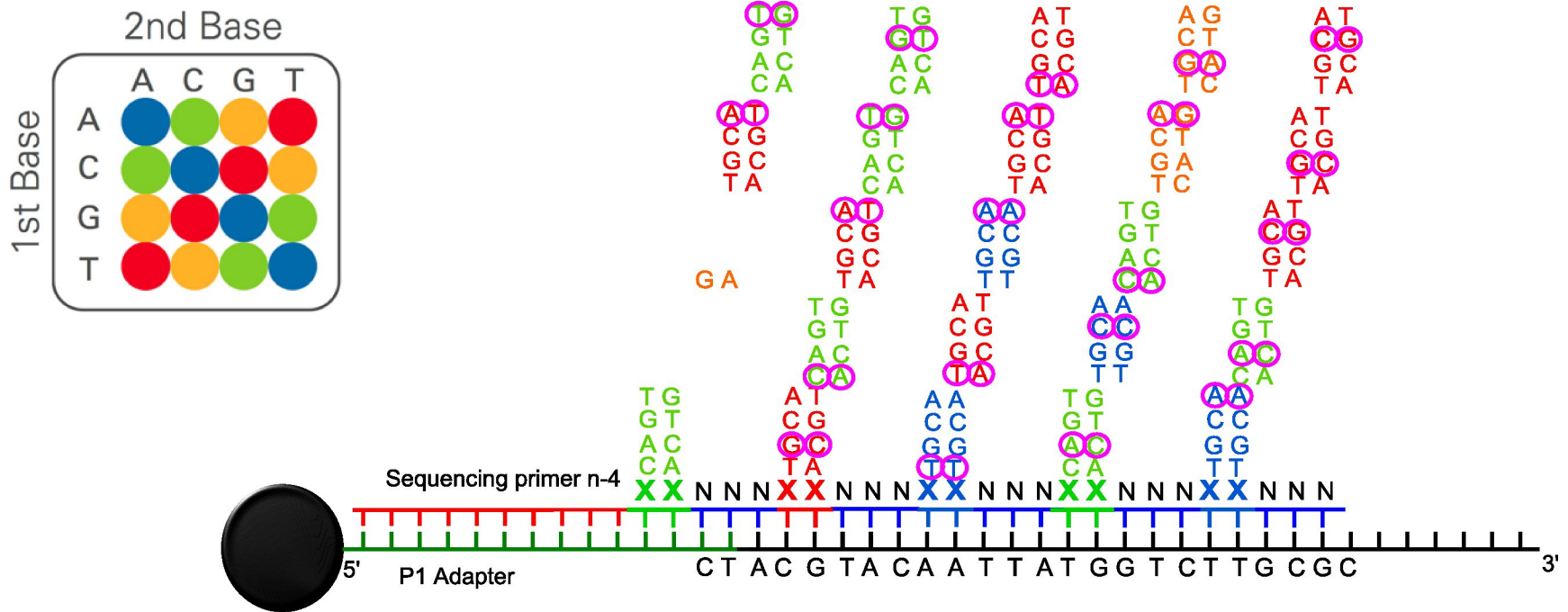
SOLiD



Different platforms

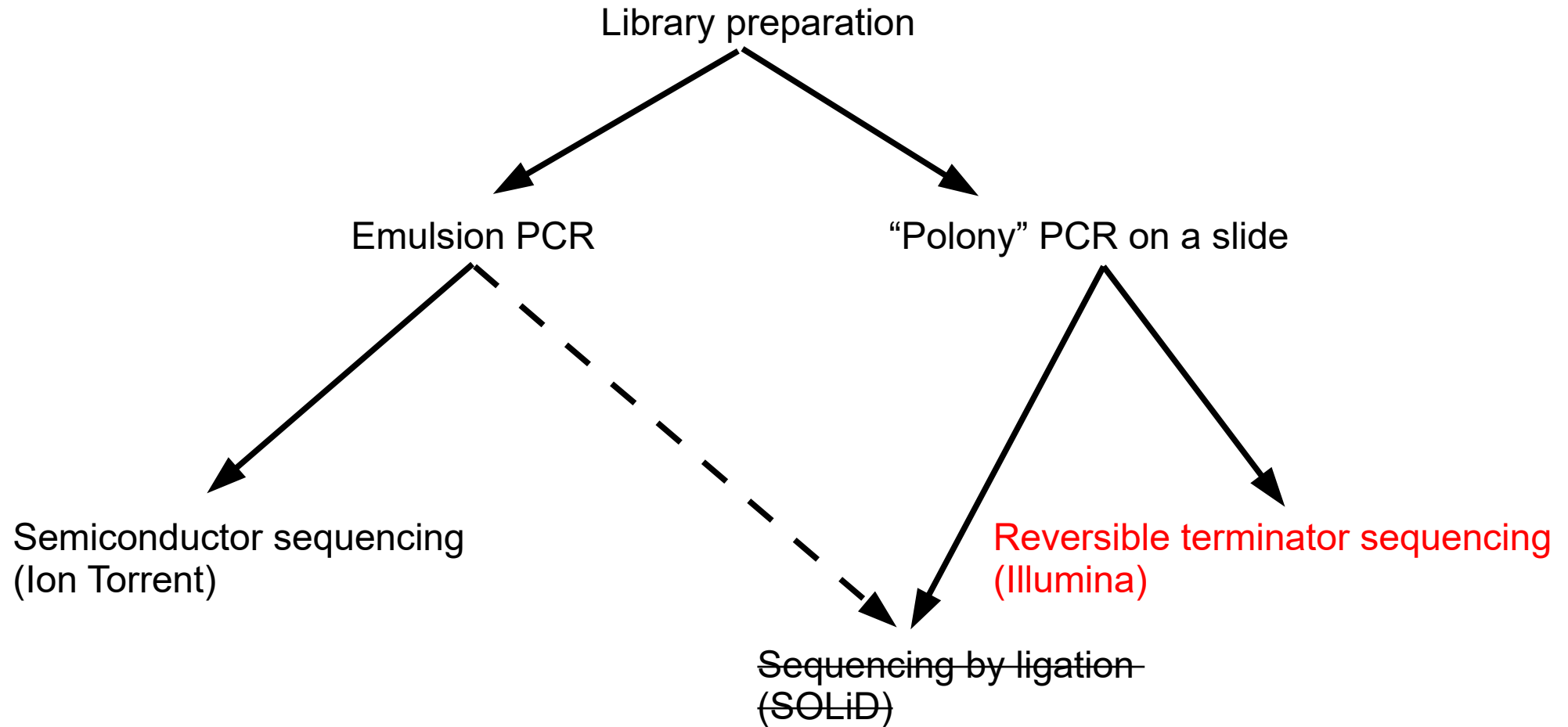
Next Generation Sequencing: Amplified Single Molecule Sequencing

SOLiD



Workflow

Next Generation Sequencing: Amplified Single Molecule Sequencing



Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing

Illumina

iSeq 100



MiniSeq



MiSeq



NextSeq 500 / 550



HiSeq 2500 / 3000 / 4000

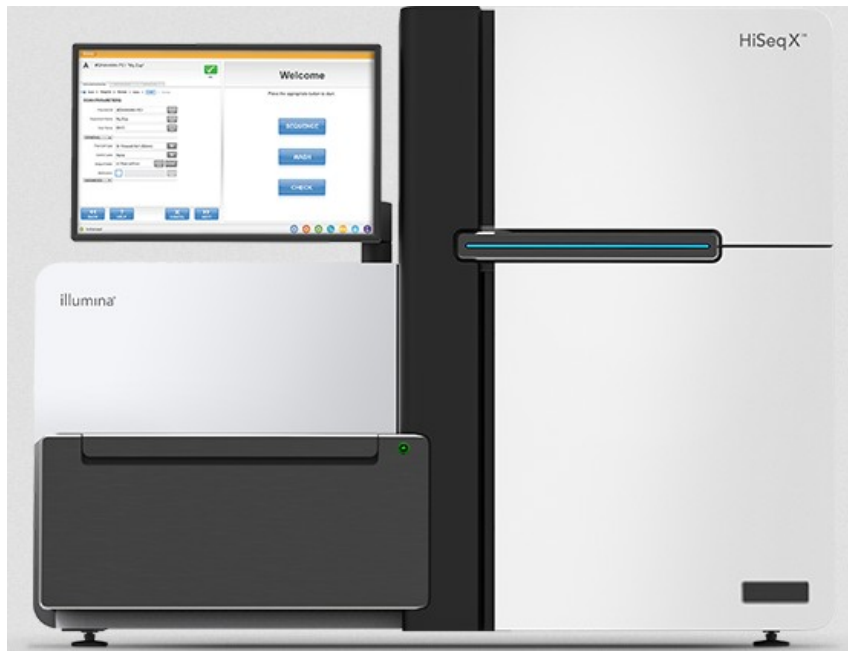


Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing

illumina

HiSeq X



NovaSeq 6000



Different platforms

52/165

Next Generation Sequencing: Amplified Single Molecule Sequencing

Illumina

	iSeq 100	MiniSeq	MiSeq	NextSeq 550
Read Length	2 x 150 bp	2 x 150 bp	2 x 300 bp	2 x 150 bp
Throughput	1.2 Gb	7.5 Gb	15 Gb	120 Gb
Reads per run	4 million	50 million	50 million	800 million
Accuracy	99,9 % (>80%)	99,9 % (>80%)	99,9 % (>70%)	99,9 % (>80% of the bases)
Run Time	17.5 hours	24 hours	55 hours	29 hours

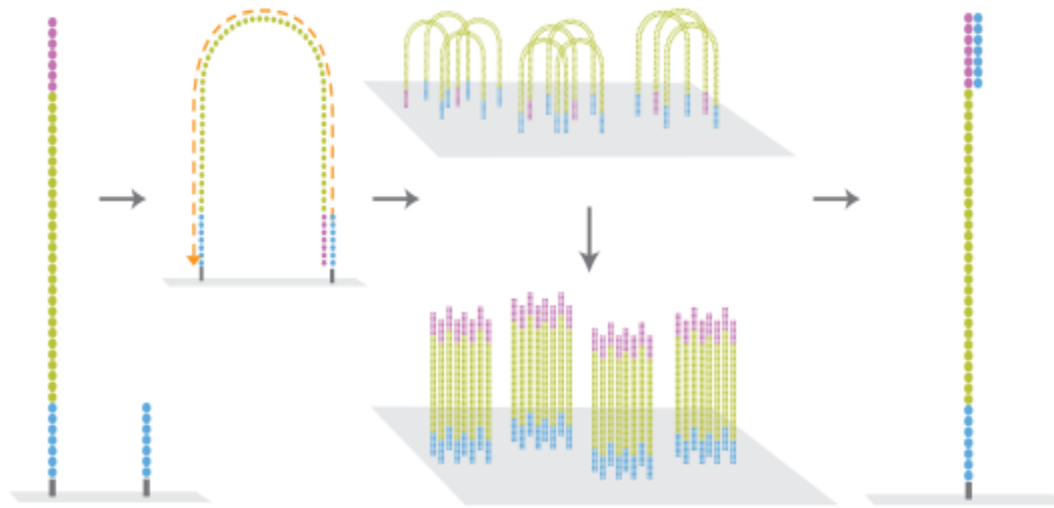
	HiSeq 2500 / 3000 / 4000	HiSeq X	NovaSeq 6000
Read Length	2 x 125 / 2 x 150 / 2 x 150 bp	2 x 150 bp	2 x 150 bp
Throughput	1000 / 750 / 1500 Gb	1800 Gb	850 – 3000 Gb
Reads per run	4 / 2,5 / 5 billion	6 billion	2.8 – 10 billion
Accuracy	99,9 % (>80% of the bases)	99,9 % (>75%)	99,9 % (>75% of the bases)
Run Time	6 / 3,5 / 3,5 days	< 3 days	36 – 44 hours

Workflow: Library preparation \longrightarrow Bridge amplification \longrightarrow Reversible termination sequencing

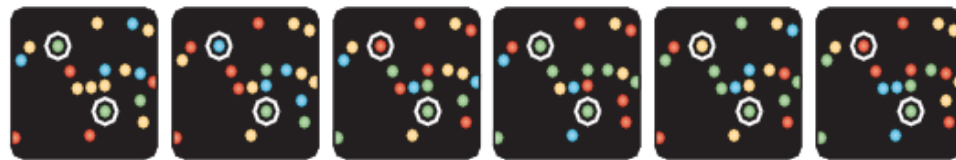
Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing

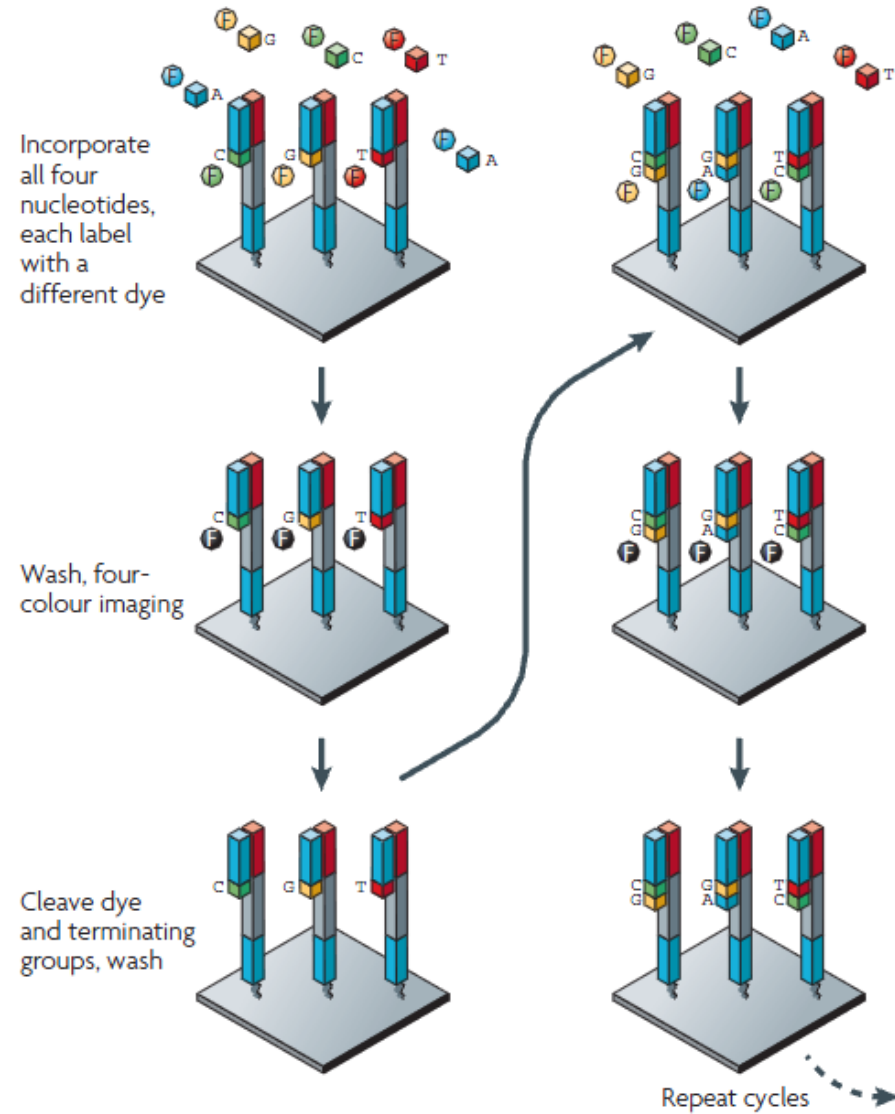
Illumina: Reversible termination sequencing



4 nucleotides with different dye flow simultaneous



Top: CATCGT
Bottom: CCCCCC



Illumina 2- and 4-channel SBS (sequencing by synthesis) sequencing technology

Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing

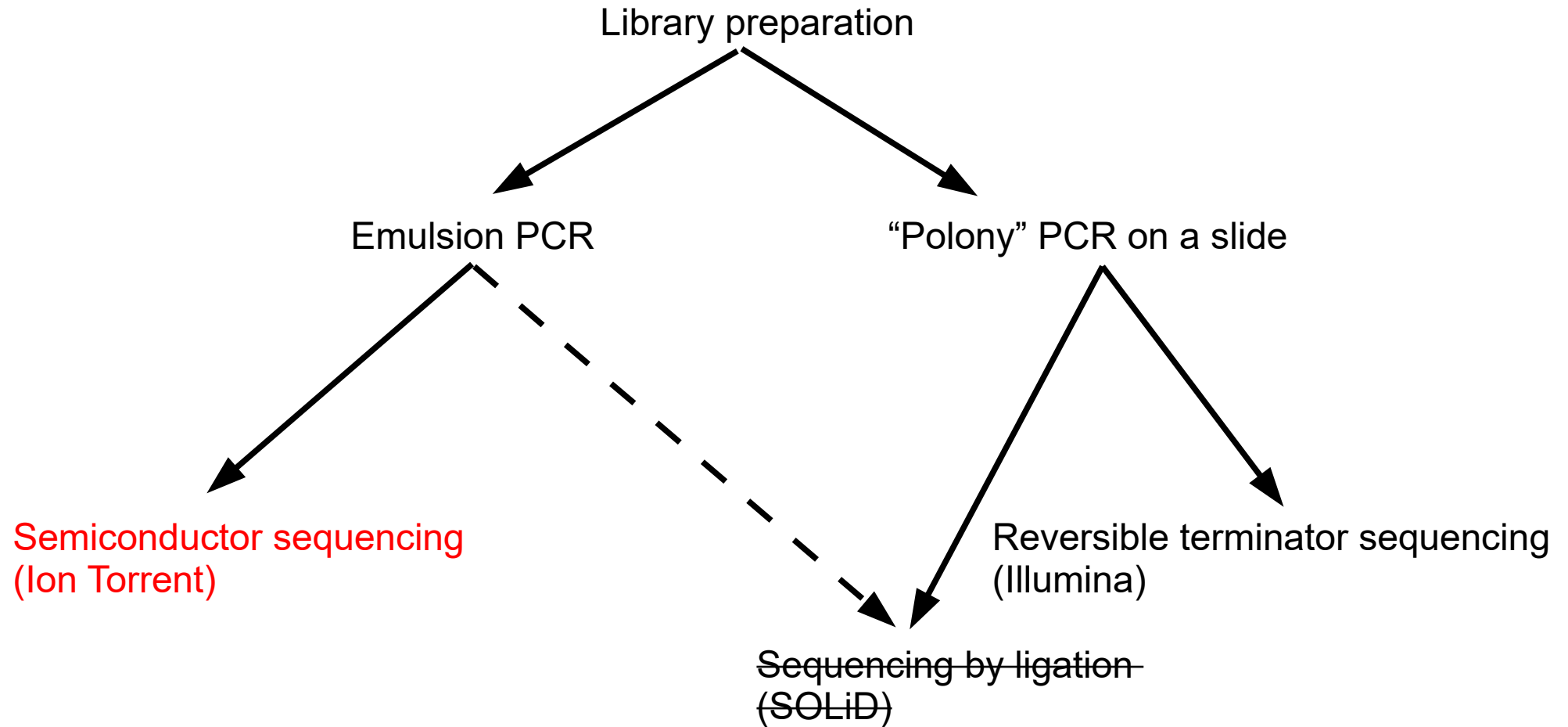
Illumina

Movie time

[Illumina sequencing \(youtube\)](#)

Workflow

Next Generation Sequencing: Amplified Single Molecule Sequencing



Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing

Ion Torrent

PGM
(Personal Genome Machine)



GeneStudio S5 / S5 Plus / S5 Prime



Proton

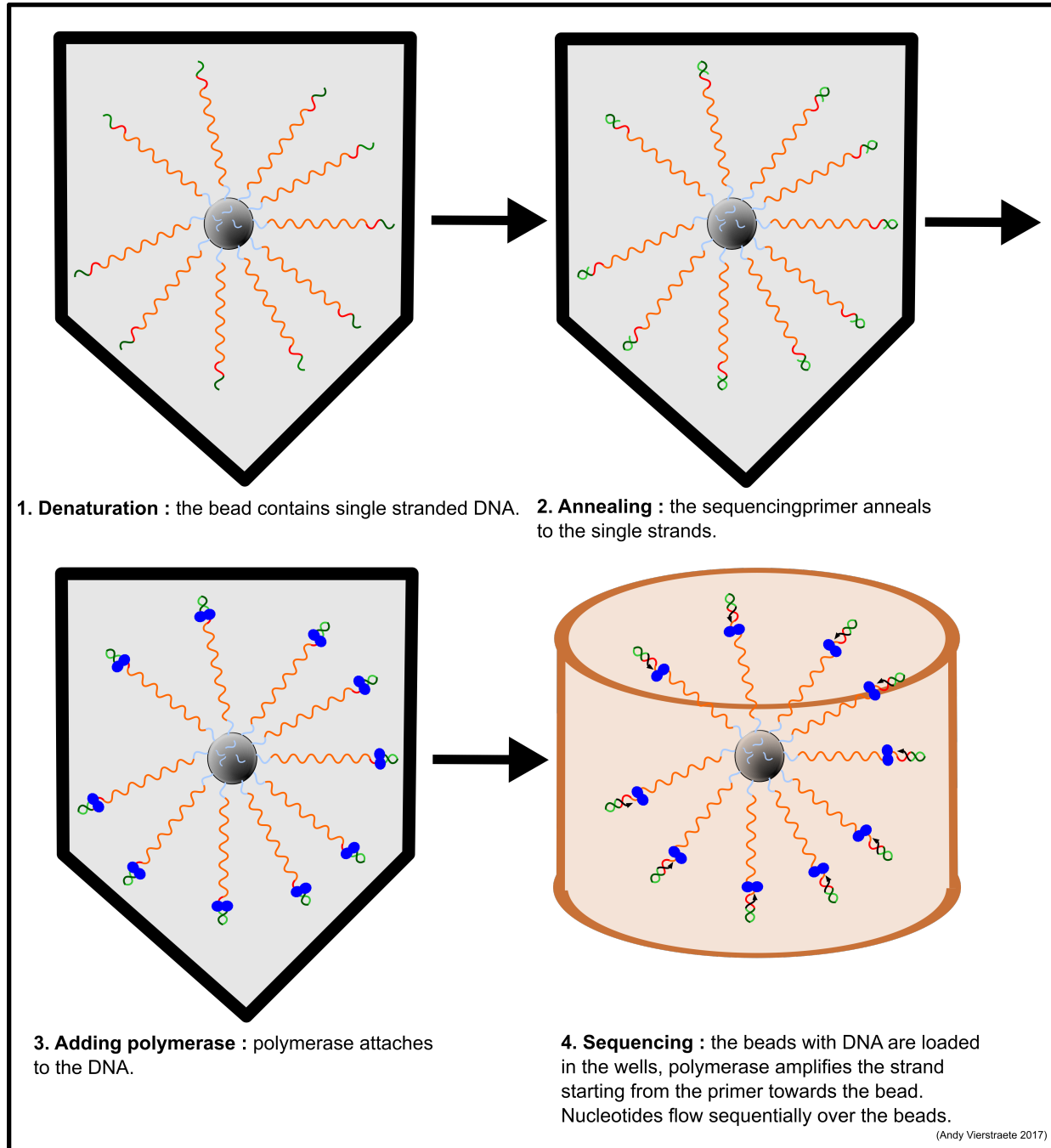


	PGM	GeneStudio S5 / S5 Plus / S5 Prime	Proton
Chip	314 - 316 - 318	510 - 520 - 530 - 540 - 550	PI – PII
Read length	400 bp	400(600) - 400(600) – 400(600) – 200 - 200	200 bp - ?
Throughput	0,1 - 0,6 - 2 Gb	1 - 2 - 8 - 15 - 25 Gb	15 -100 Gb
Reads per run	0,5 - 3 - 6 million	3 - 5 - 20 - 80 - 130 million	80 - 250 million
Accuracy	99 % (raw read)	99 % (raw read)	99 % (raw read)
Run Time	4 - 5 - 7 hours	4 - 4 - 4 - 2,5 hours	2,5 hours
Data processing	2 - 4 - 6 hours	6,5 - 8 - 17,5 - 16,5 hours /(up to 4x faster for Plus and Prime)	2,5 hours

Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing

Ion Torrent



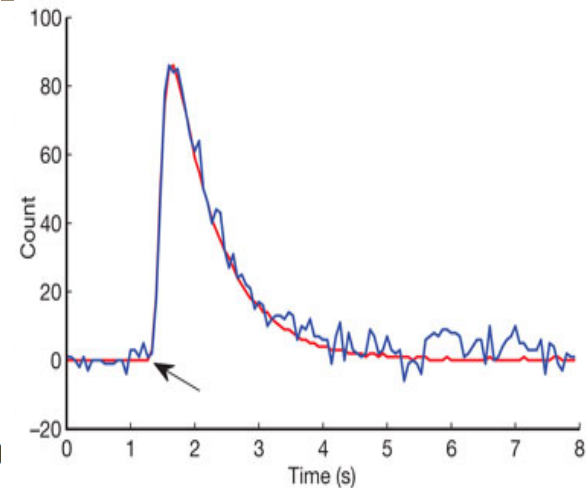
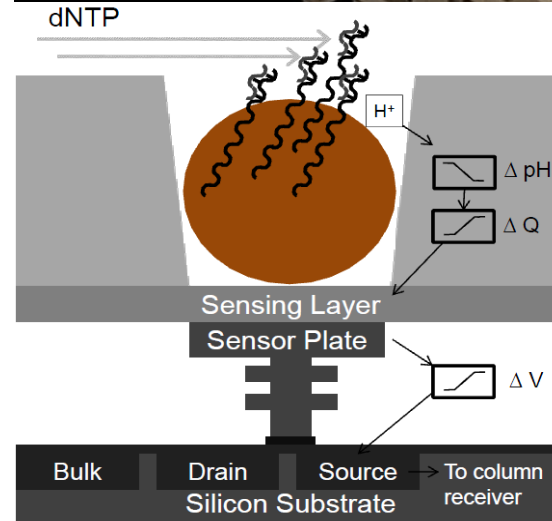
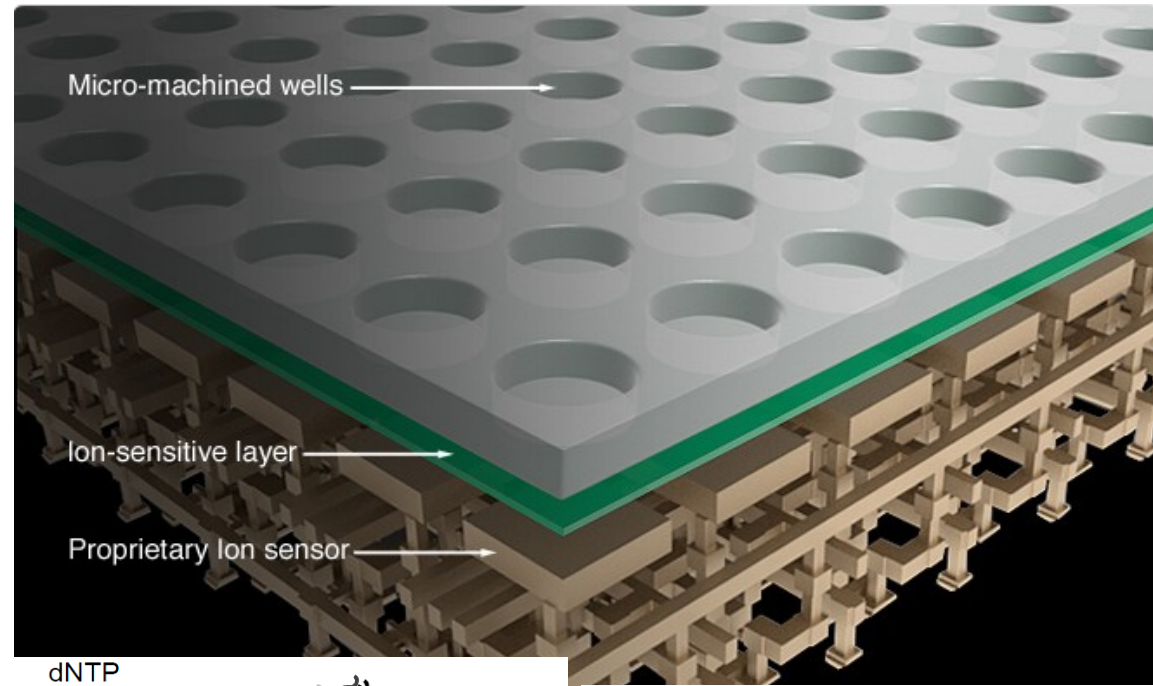
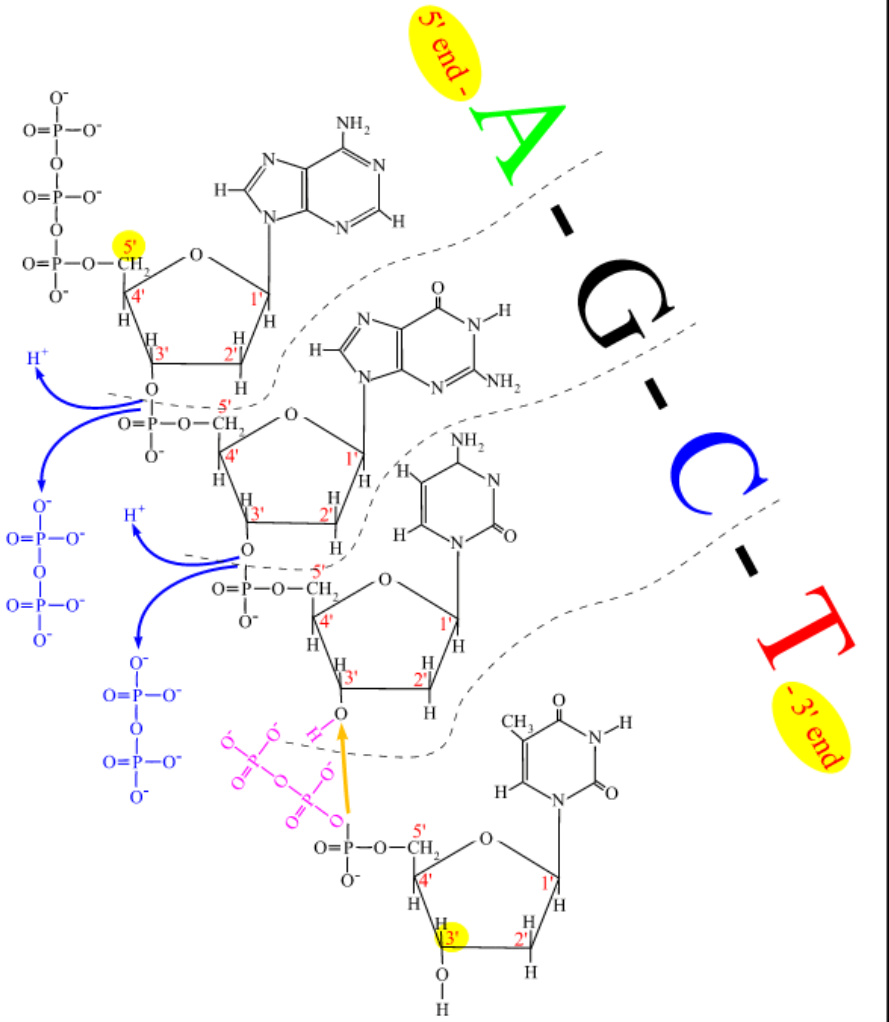
Sequencing

Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing Ion Torrent

Workflow: Library preparation → Emulsion PCR → Semiconductor Sequencing

From nucleotide to DNA

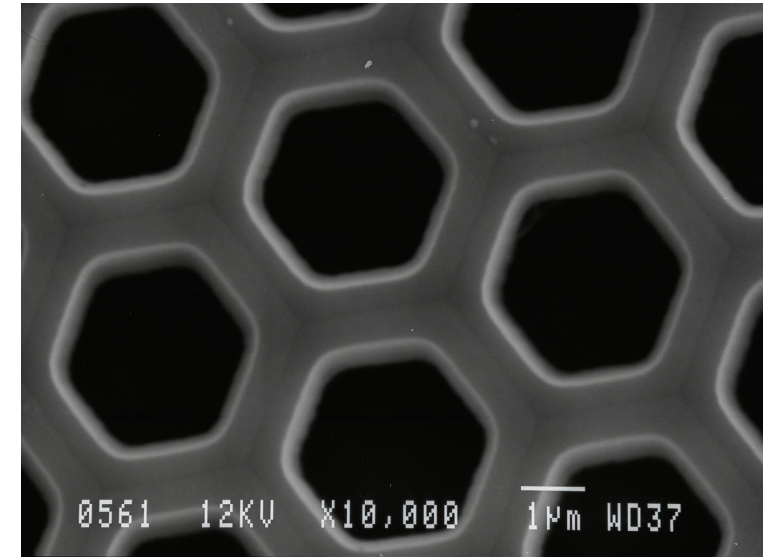
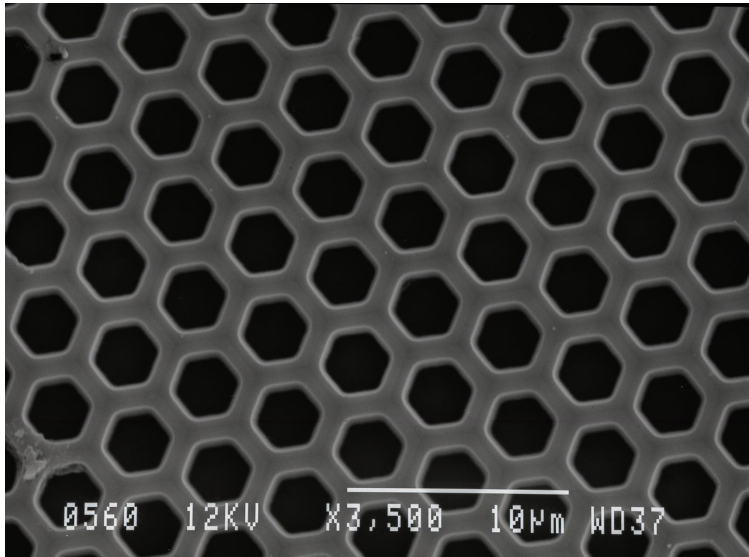


Different platforms

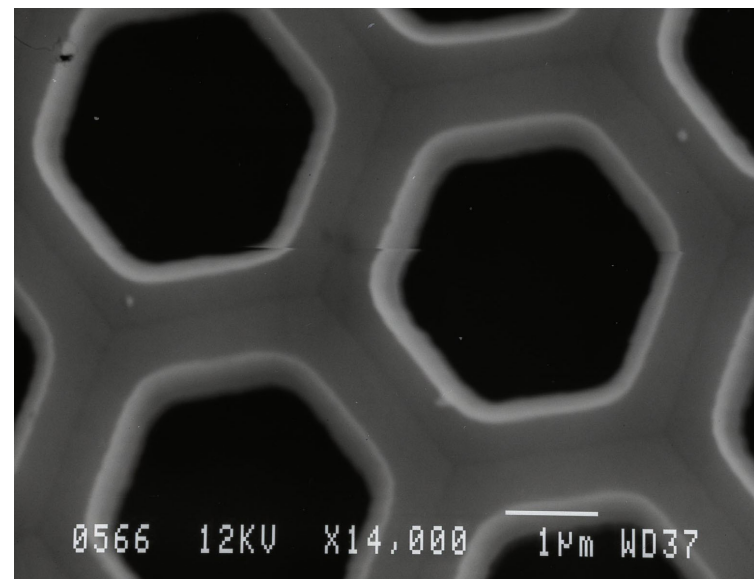
Next Generation Sequencing: Amplified Single Molecule Sequencing

Ion Torrent

Workflow: Library preparation → Emulsion PCR → Semiconductor Sequencing



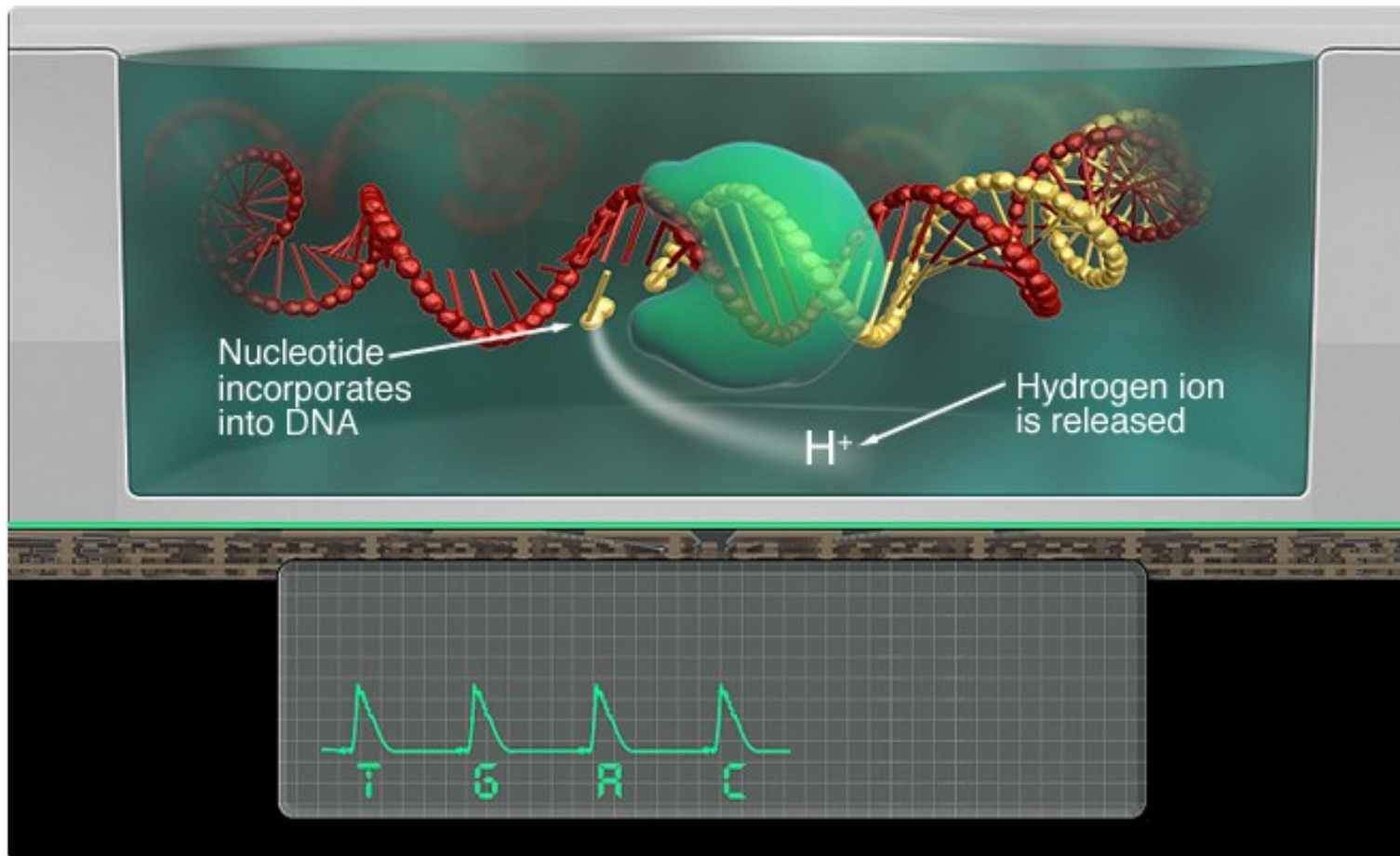
Picture of the wells in a 318 chip



Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing Ion Torrent

4 nucleotides flow sequentially

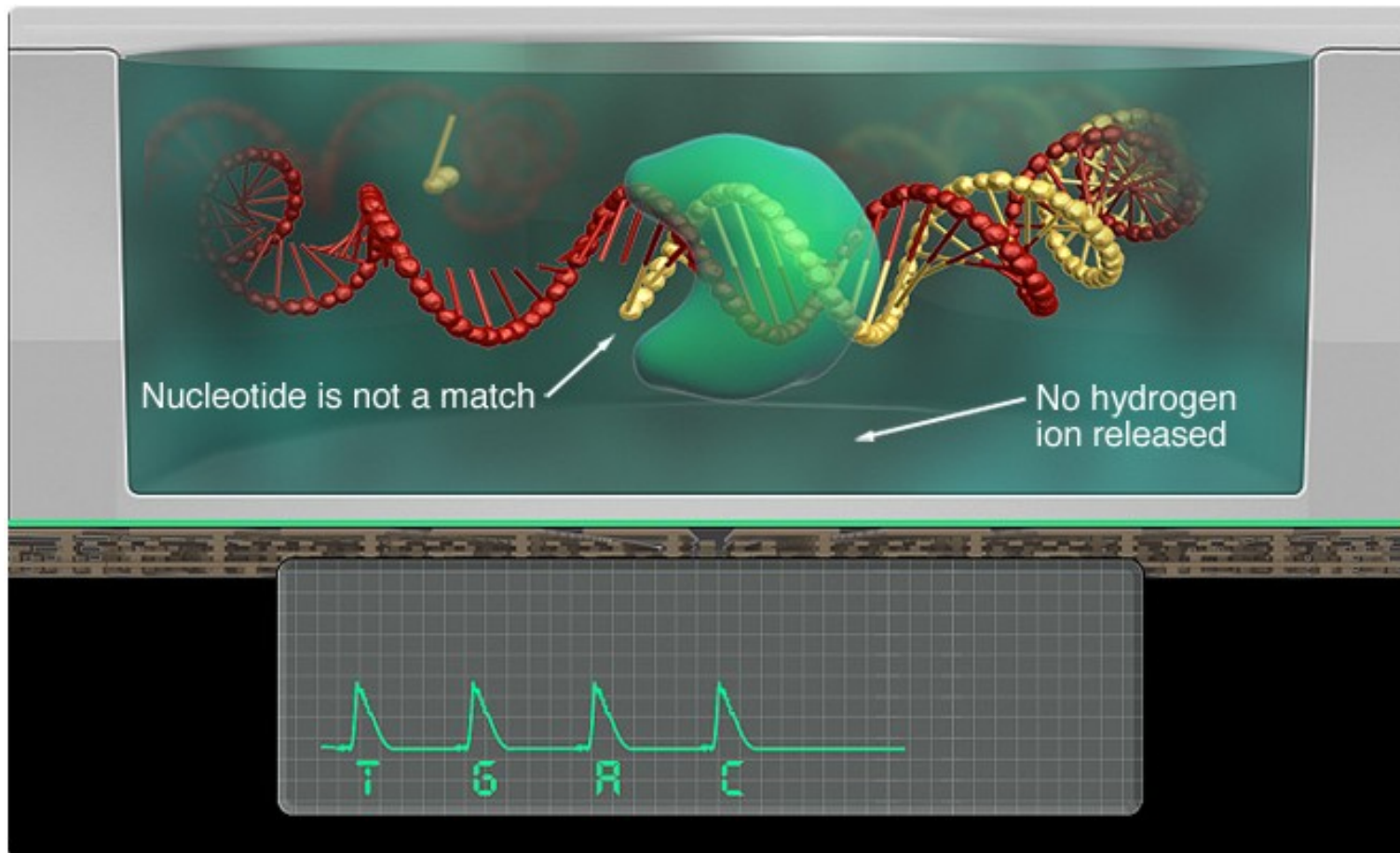


No camera, just a pH sensor

Different platforms

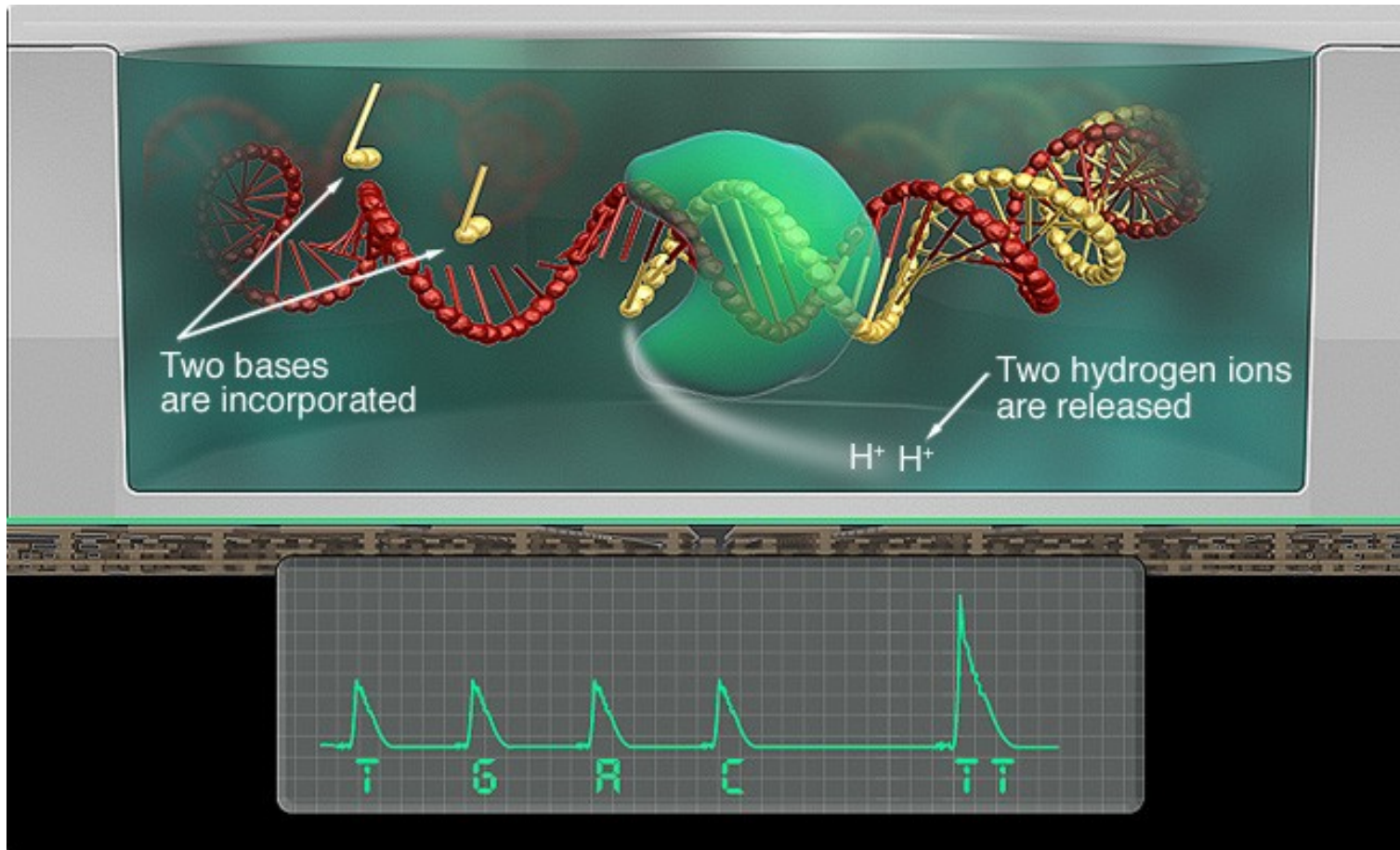
Next Generation Sequencing: Amplified Single Molecule Sequencing

Ion Torrent



Different platforms

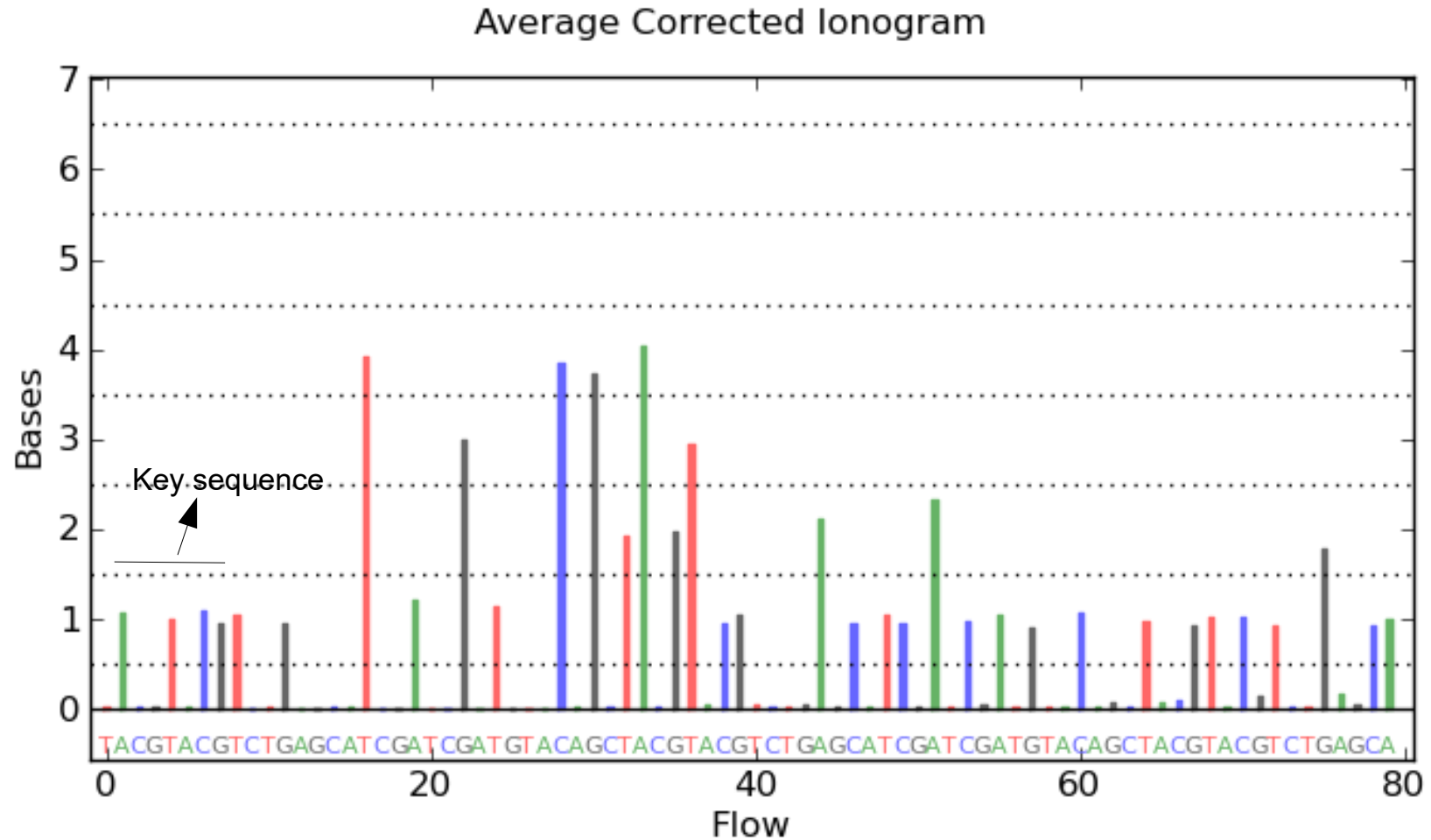
Next Generation Sequencing: Amplified Single Molecule Sequencing Ion Torrent



Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing

Ion Torrent



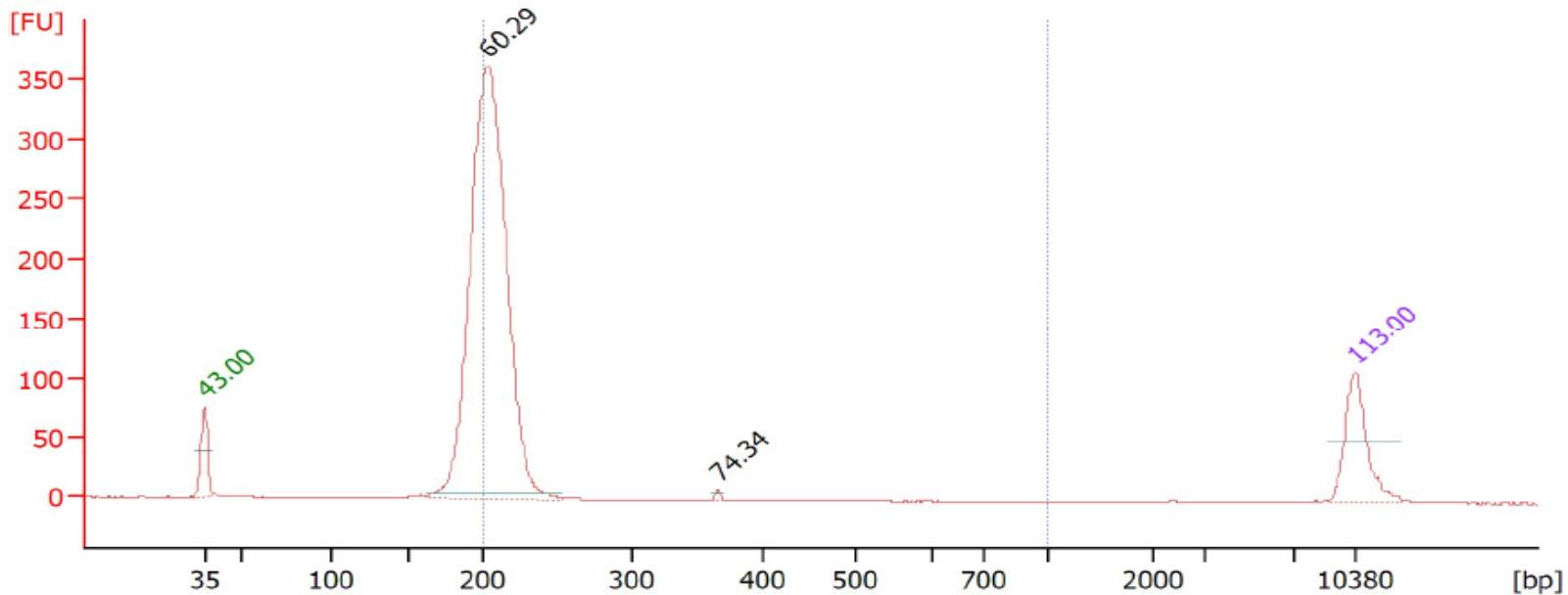
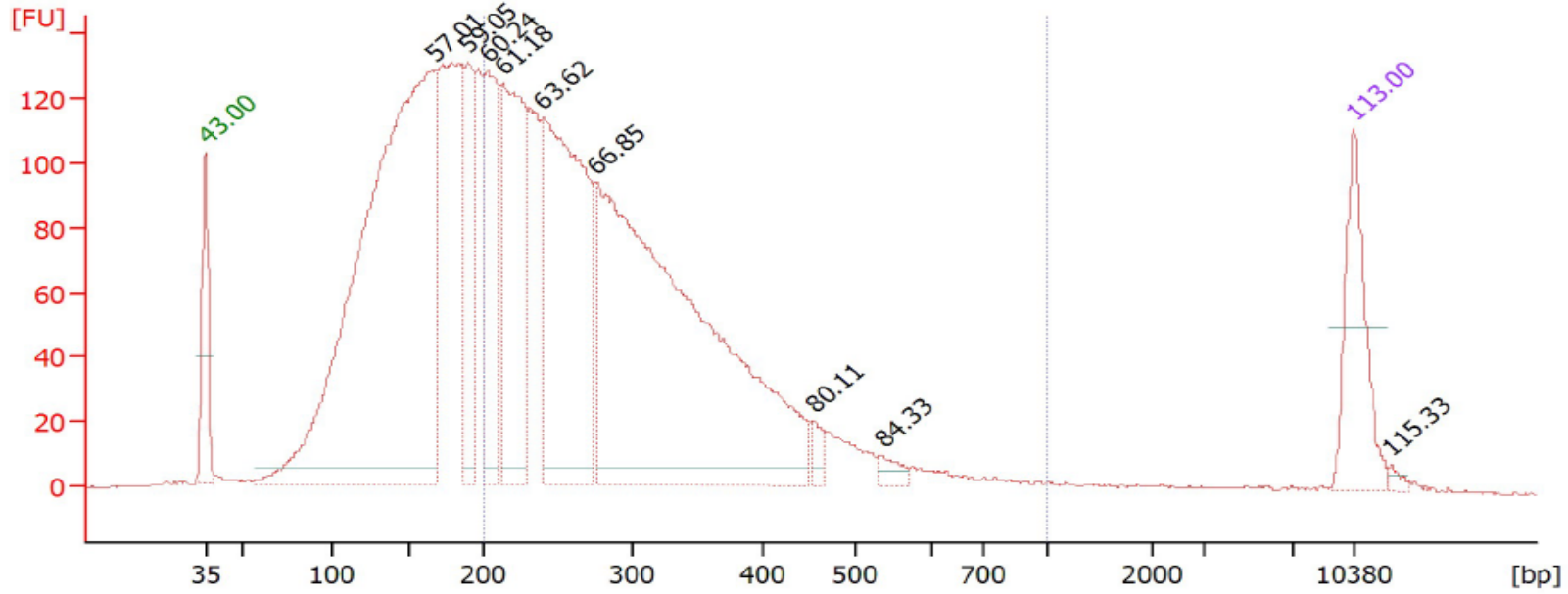
ATCGTGT TTTAGGGTCCCCGGGGTT...

Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing

Ion Torrent

Size selection - > maximizing sequencing yield

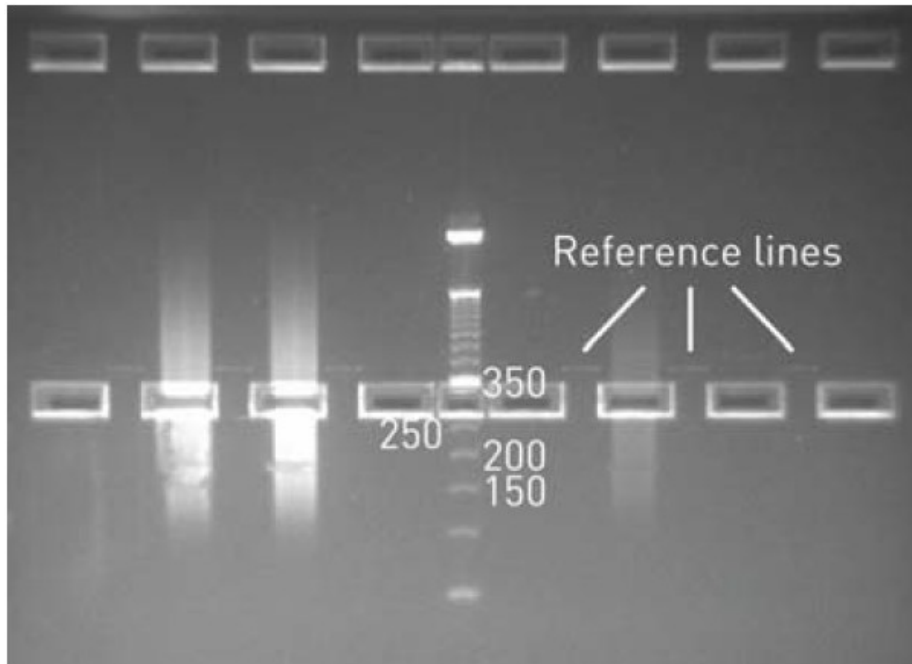


Different platforms

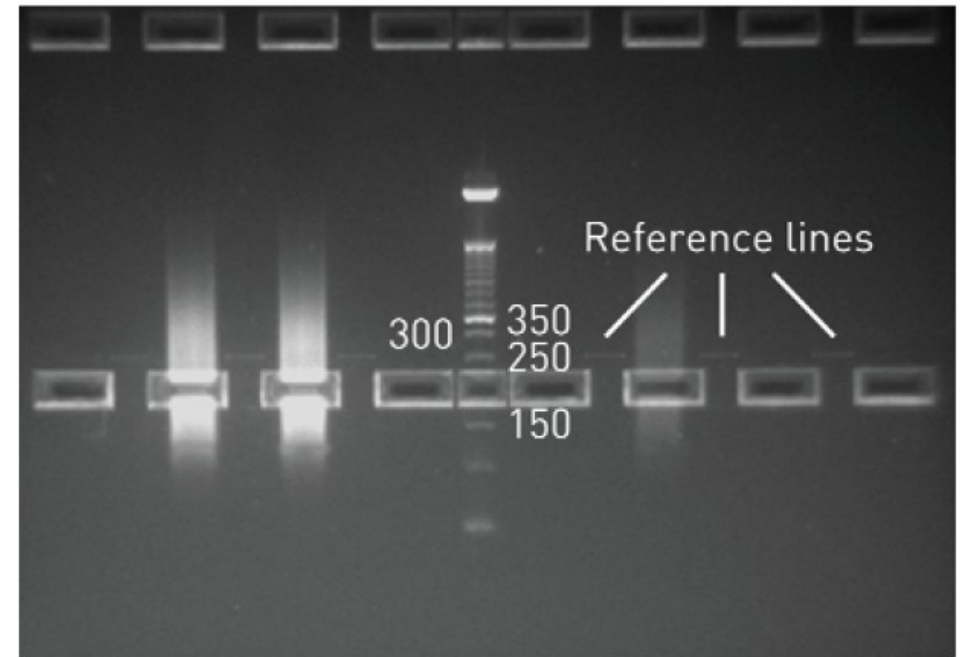
Next Generation Sequencing: Amplified Single Molecule Sequencing

Ion Torrent

Size selection -> maximizing sequencing yield



200 base-read library gel



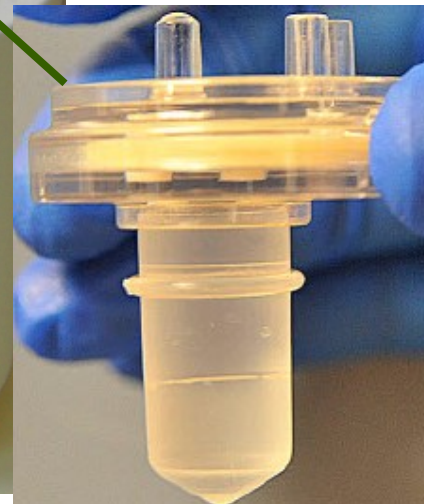
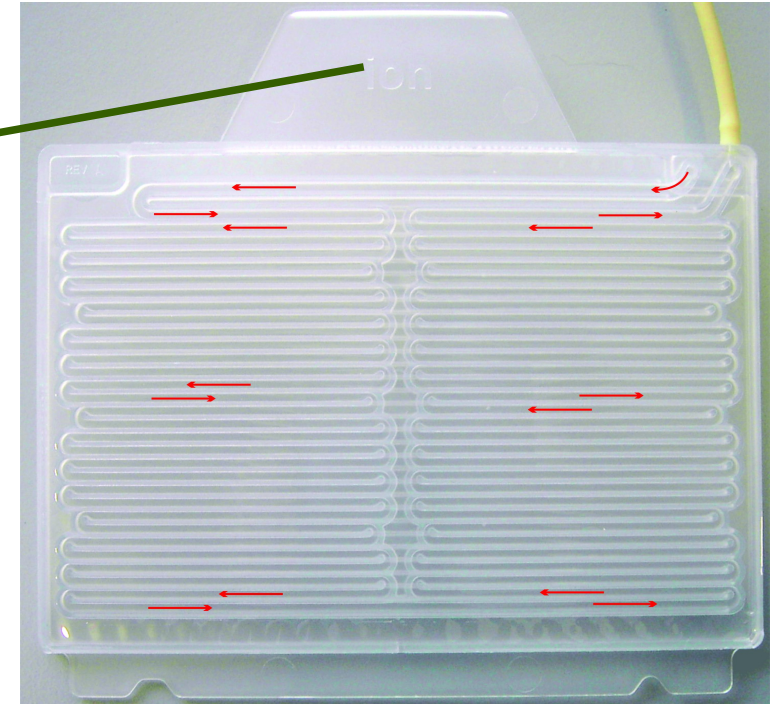
100 base-read library gel

Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing

Ion Torrent

Emulsion PCR: OneTouch Instrument



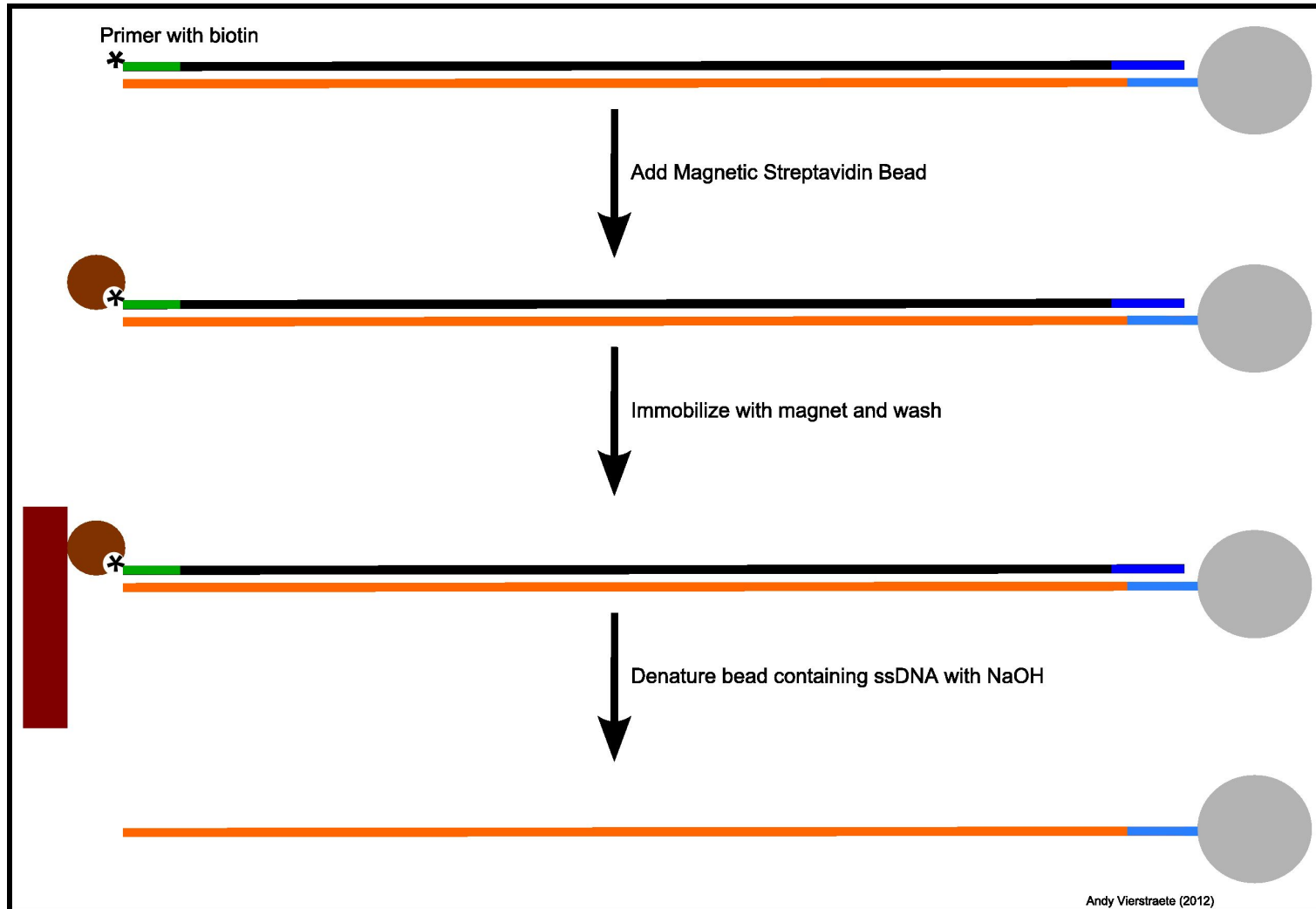
15 min hands-on; 4-8 hours amplification; 35 min enrichment

Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing

Ion Torrent

Enrichment: select only the beads that contain DNA
-> maximizing sequencing yield



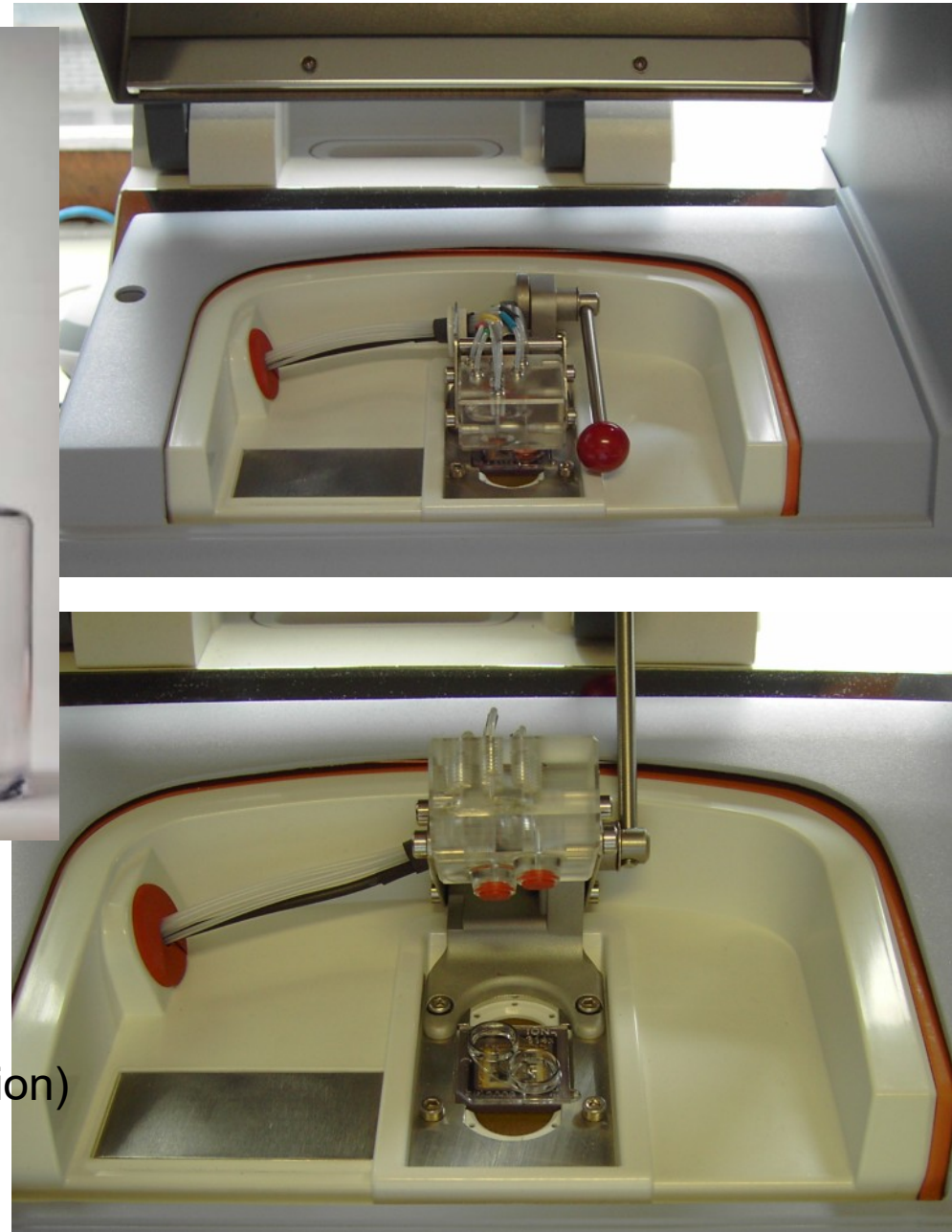
Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing

Ion Torrent



W1 bottle: 350 μ l 100 mM NaOH
 W2 bottle: 2 liter W2 solution (contains pcr solution)
 W3 bottle: 50 ml W3 (= buffer with known pH)
 4 tubes with 20 μ l of the different dNTP's



Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing

Ion Torrent

Movie time

[Ion Torrent Sequencing \(youtube link\)](#)

Different platforms

Next Generation Sequencing: Amplified Single Molecule Sequencing

Ion Torrent

Torrent Server pipeline

Process Description	File Types	Ion 314™ chip	Ion 316™ chip	Ion 318™ chip
Raw Voltage Data	DAT	28 GB	129 GB	242 GB
Signal Processing	WELLS	1 GB	8 GB	12 GB
Base Calls - Flow	SFF	1 GB	5 GB	10 GB
Base Calls - Base	FASTQ	0.2 GB	1 GB	2.5 GB

Different platforms

- Illumina (Solexa)
 - MiniSeq
 - MiSeq
 - NextSeq 500 - 550
 - HiSeq 2500 - 3000 – 4000
 - NovaSeq 5000 - 6000
 - HiSeq X Five - Ten
- Thermo Fisher Scientific (Applied Biosystems -> Life Technologies)
 - Ion Torrent Personal Genome Machine (PGM)
 - Ion Torrent S5 and S5XL
 - Ion Torrent Proton

- Pacific Biosciences
 - Sequel System
 - PacBio RS II
- Oxford Nanopore Technologies
 - SmidgION
 - MinION
 - GridIONx5
 - PromethION
- SeqLL
 - tSMS sequencer

Next Generation Sequencing
Amplified Single Molecule Sequencing

Third Generation Sequencing,
Next Next Generation Sequencing,
Single Molecule Sequencing

Different platforms

Third Generation Sequencing: Single Molecule Sequencing

Pro's:

- Less sample preparation (no PCR, no loss in sequences)
- Longer read lengths (PacBio and Oxford Nanopore)
- No amplification
 - > all reads are unique
 - > no PCR errors or PCR bias
 - > fewer contamination issues
 - > no GC-bias
 - > analyze every sample (un-PCR-able / unclonable)
 - > analyze low quality DNA (museum, archaeological, forensic samples)
- Absolute quantification (low abundance transcripts)
- Sequence RNA directly → easy detection of isoforms
- Detection of base modifications (methylation)
- Detection of structural variants (copy number variants, gene duplications, deletions, insertions, inversions, and translocations)
- start and stop sequencing as required (PacBio and Ox. Nanopore)
- data available in real time (Ox. Nanopore)
- Possibility for real-time targeted sequencing (Ox. Nanopore) (if no match with target, DNA strand is ejected and a new is captured and sequenced)
- Possibility to flatten coverage variation: “read until”: stop reading if gene has enough coverage, load an other strand and sequence. Less sequencing to cover all variants.

Cons:

- Lower read quality

Different platforms

Third Generation Sequencing: Single Molecule Sequencing

Pacific Biosciences

Pacbio RS



Sequel System



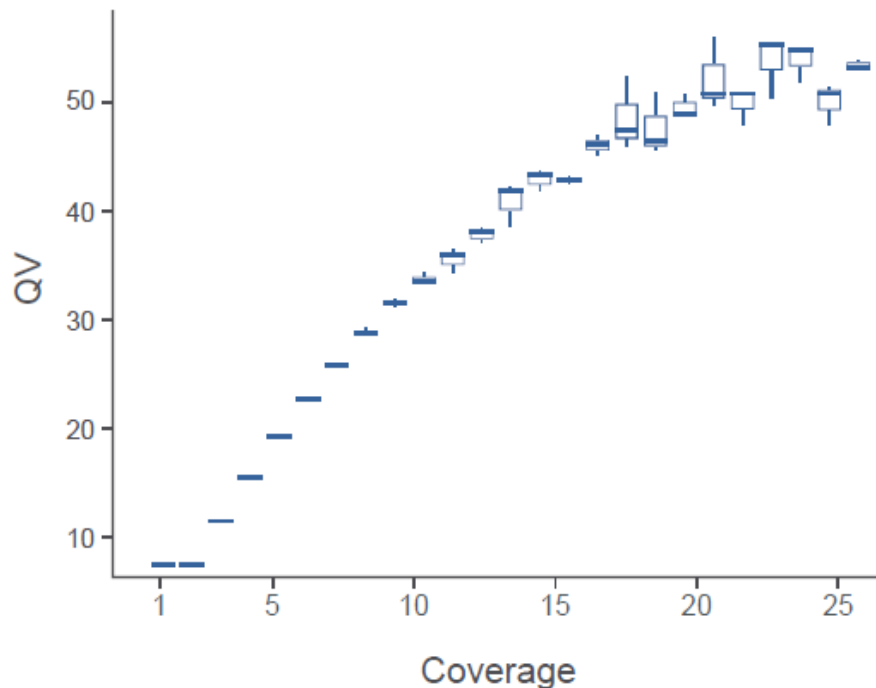
	Pacbio RS	Sequel System
Read Length	50 % > 20 kb (max > 60 kb)	50 % > 20 kb (max > 60 kb)
Throughput	1 Gb/SMRT cell (max 16/run)	5-8 Gb/SMRT cell (max 16/run)
Reads per run	55,000	365,000
Accuracy	86 %	86 %
Run Time	30 minutes – 6 hours/ SMRT cell	30 minutes – 10 hours/ SMRT cell

Different platforms

Third Generation Sequencing: Single Molecule Sequencing

Pacific Biosciences

Accuracy



Quality scores in sequencing:
Q17, Q20, Q30, ...

Quality score	Probability of incorrect bases	Base call accuracy
10	1 in 10	90 %
17	1 in 50	98 %
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%

- Circular Consensus Sequencing (CCS reads)
- Consensus by sequencing many reads

```
TGCAGATCATTACT - AAACAACGC - TCC - AC - TATCAAAT - CCGGGTGCG - CTTGTTGTATAACACAAAC - AGG - CGAAAAAACATA - TCG - AGTT
TGCAGATCATTACT - AAACAACGC - TCC - AC - TATCAAAT - CCGGGTGCG - CTTGTTGTATAACACAAAC - AGG - CGAAAAAACATA - TCG - AGTT
TGCAGATCATTACT - AAACAACGC - TCC - AC - TATCAAAT - CCGGGTGCG - CTTGTTGTATAACACAAAC - AGG - CGAAAAAACATA - TCG - AG - T
TGCAGATCATTACT - AAACAACGC - TCC - AC - TATCAAAT - CCGGGTGCG - CTTGTTGTATAAC
TGCAGATCATTACT - AAACAACGC - TCC - AC - TATCAAAT - CCGGGTGCG - CTTGTTGTATAACACAAAC - AGG - CGAAAAAACATA - TCG - AGTT
TGCAGATCATTACT - AAACAACGC - TCC - AC - TATCAAAT - CCGGGTGCG - CTTGTTGT
TGCAGATCATTACT - AAACAACGC - TCC - AC - TATCAAAT - CCGGGTGCG - CTTGTTGTATA
TGCAGATCATTACT - AAACAACGC - TCC - AC - TATCAAAT - CCGGGTGCG - CTTGTTGTATAACACAAAC - AGG - CG - AAAAACATA - TCG - AGTT
TGCAGATCATTACT - AAACAACGC - TCC - AC - TATCAAAT - CCGGGTGCG - CTTGTTGTATAACACAAAC - AGG - TAGG - CGAAAAAACATA - TCG - AGTT
```

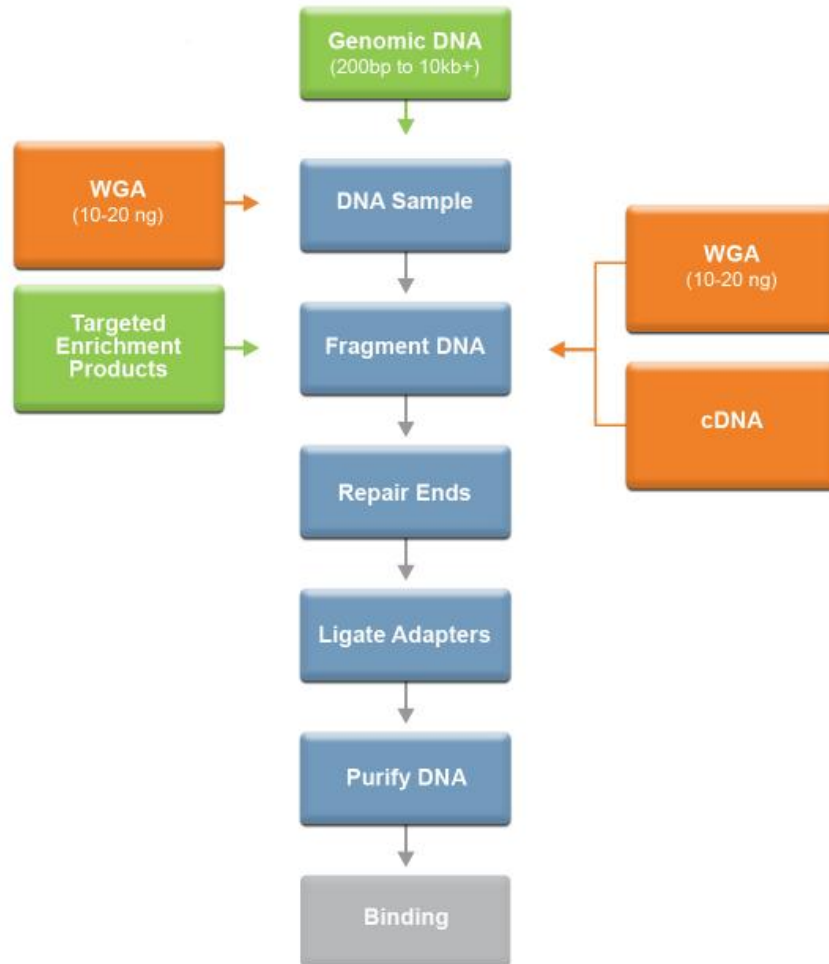
Different platforms

Third Generation Sequencing: Single Molecule Sequencing

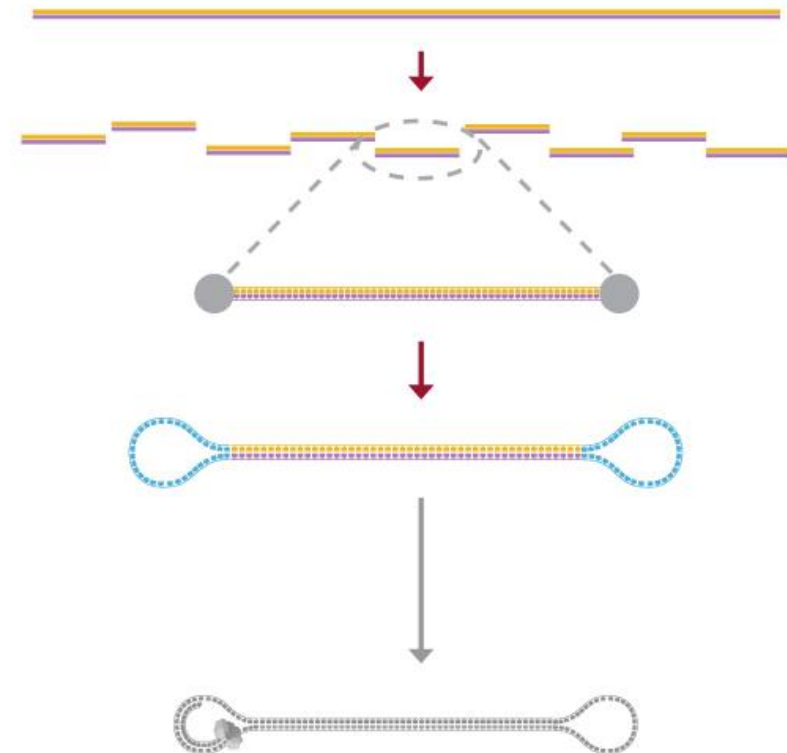
Pacific Biosciences

Workflow: Library preparation \longrightarrow Sequencing

Sample Preparation



Building of SMRTbell

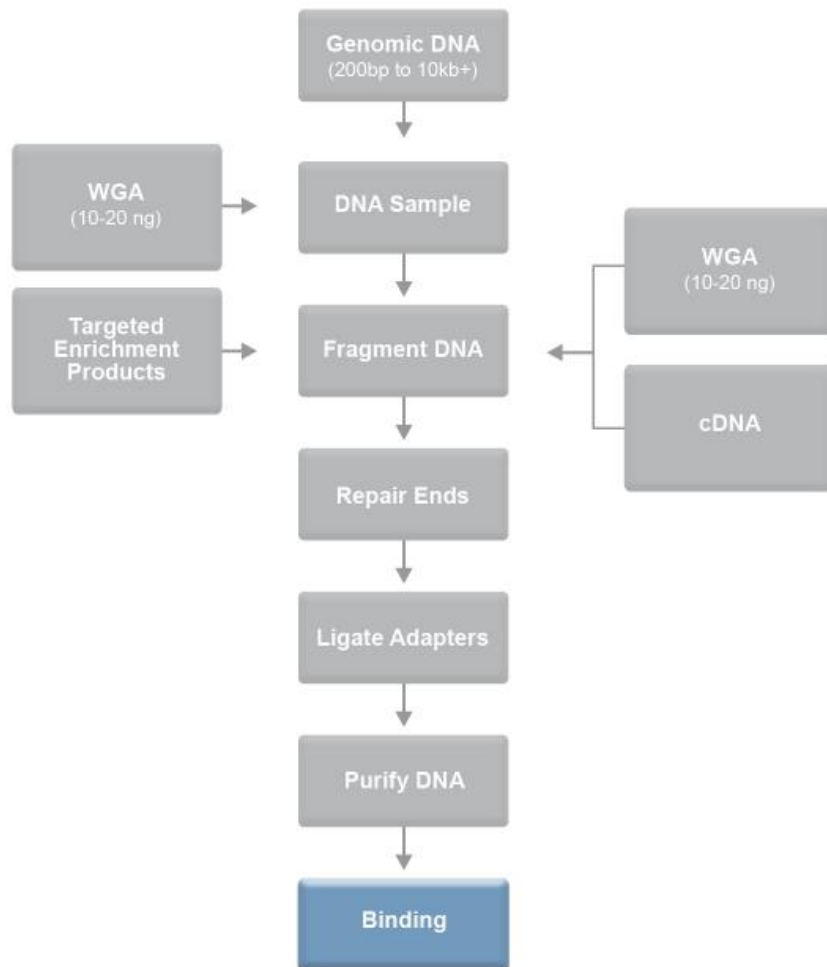


Different platforms

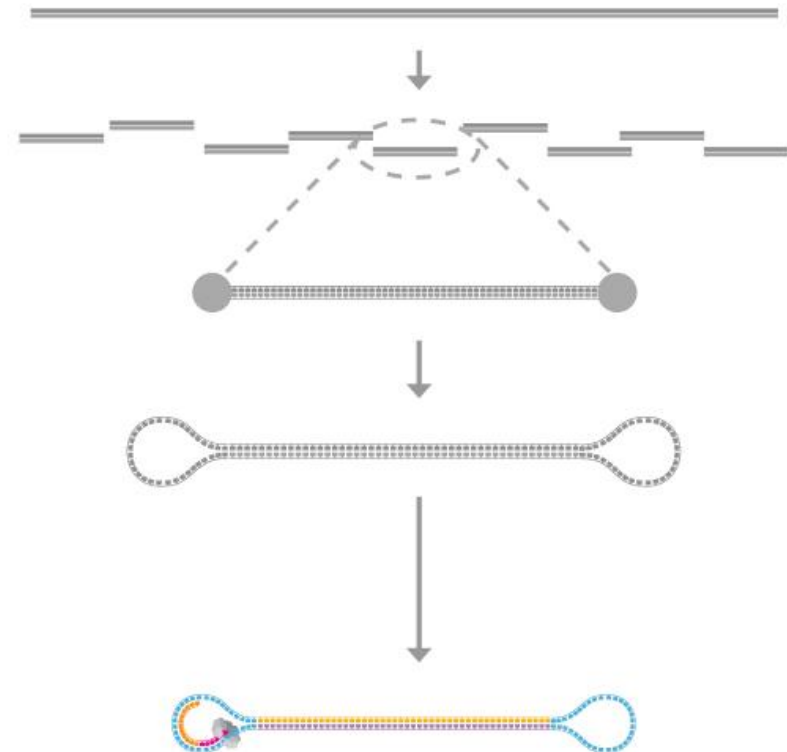
Third Generation Sequencing: Single Molecule Sequencing

Pacific Biosciences

Sample Preparation



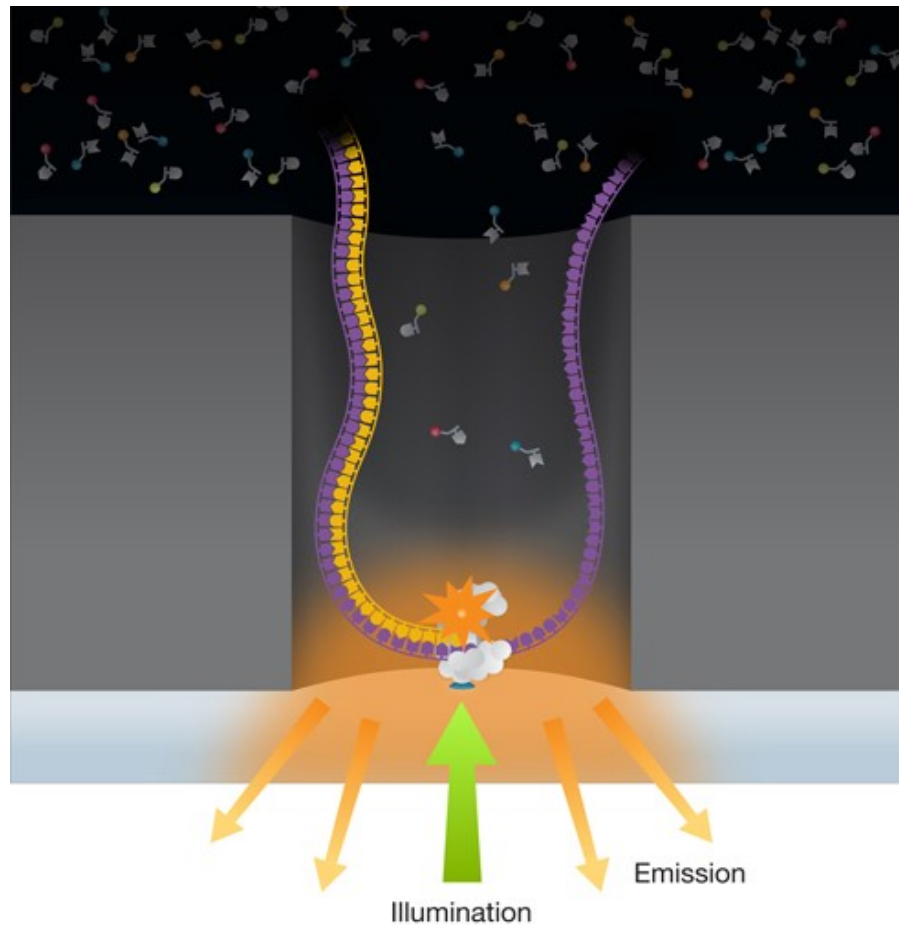
Building of SMRTbell



Different platforms

Third Generation Sequencing: Single Molecule Sequencing

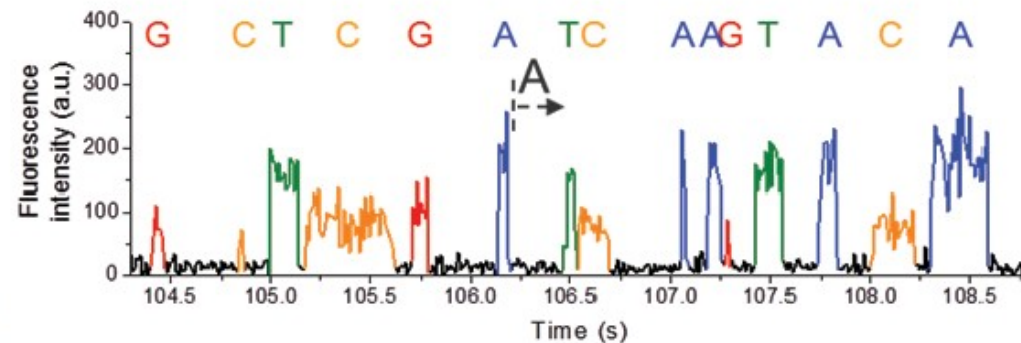
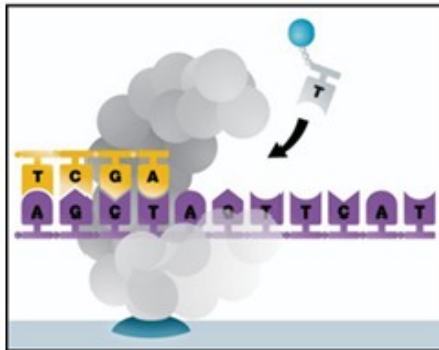
Pacific Biosciences



4 nucleotides with different fluorescent dye simultaneous present

2-4 nucleotides/sec
2-20 Kb read length
6 TB raw data in 30 minutes

laser damages polymerase



Different platforms

Third Generation Sequencing: Single Molecule Sequencing

Pacific Biosciences

Movie time

Pacific Biosciences (YouTube)

Different platforms

Third Generation Sequencing: Single Molecule Sequencing

Oxford Nanopore

SmidgION



MinION



GridION X5



PromethION



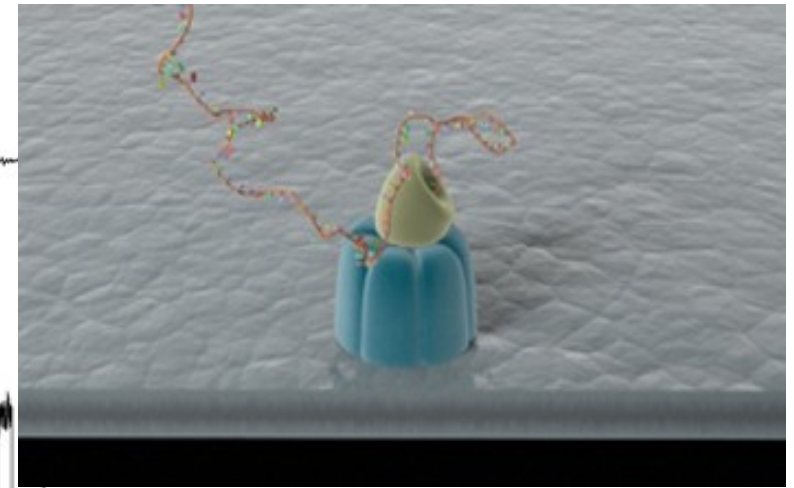
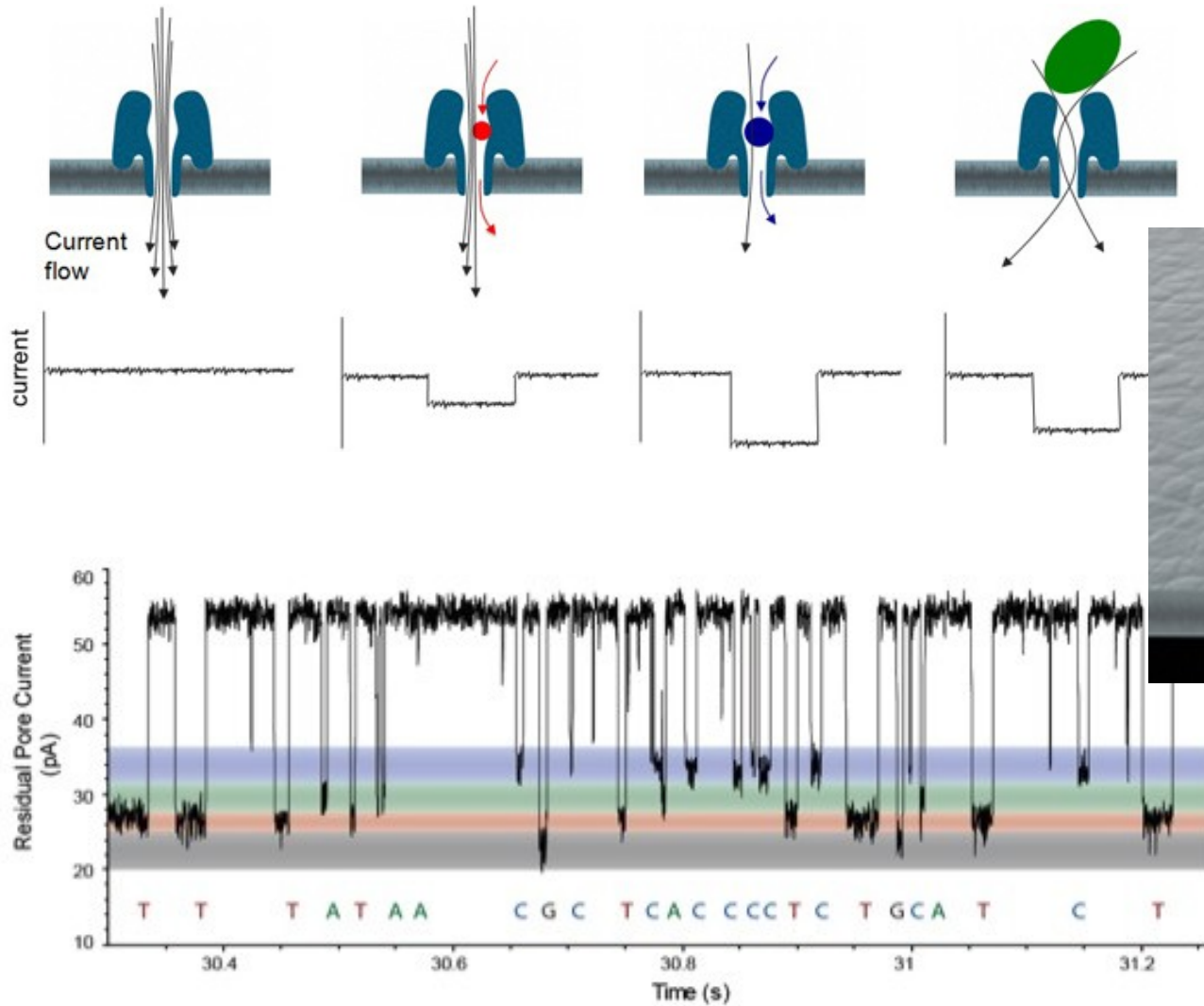
	SmidgION	MinION	GridION X5	PromethION
Read Length	?	> 200 kb (record: 1,2 Mb)		
Throughput	1 Gb (1 flow cell with 256 pores) ?	10-20 Gb (1 flow cell with 512 pores)	100 Gb (5 flow cells with 512 pores/cell) 2560 pores	50 – 250 Gb per flow cell/48hours ? (48 flow cells, 3000 pores/cell) 144,000 pores
Reads per run	?	10,000 – >300,000		
Accuracy		90 % (1D) – 96 % (1D ²)		
Run Time	1 – 4 hours	1 - 48 (70) hours	1 – 48 hours	1 - 48 hours

consensus accuracy improved to 99.5% at 30× coverage

Different platforms

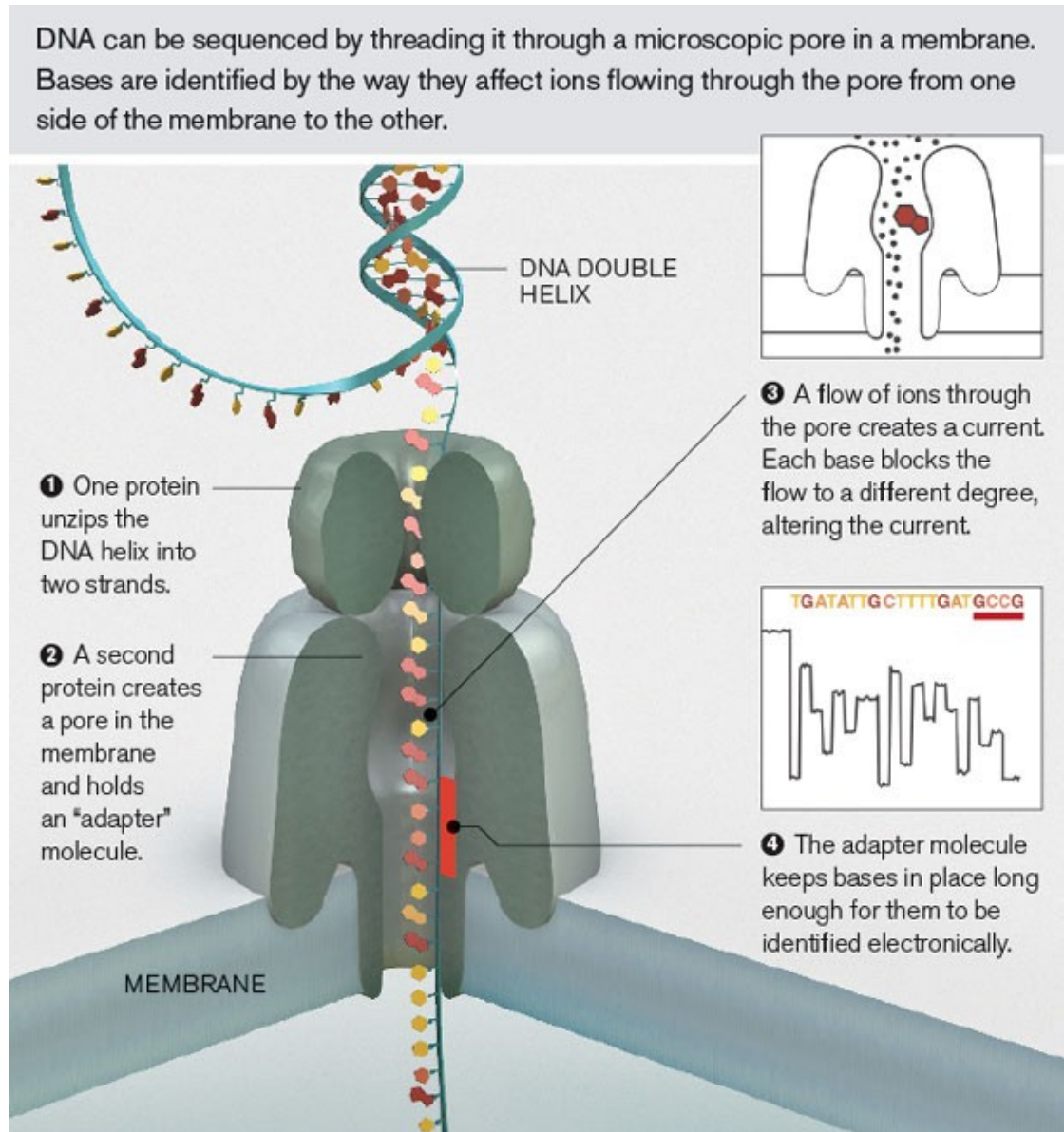
Third Generation Sequencing: Single Molecule Sequencing

Oxford Nanopore



Different platforms

Third Generation Sequencing: Single Molecule Sequencing Oxford Nanopore



R9.4: 450 nucleotides/second

Different platforms

Third Generation Sequencing: Single Molecule Sequencing

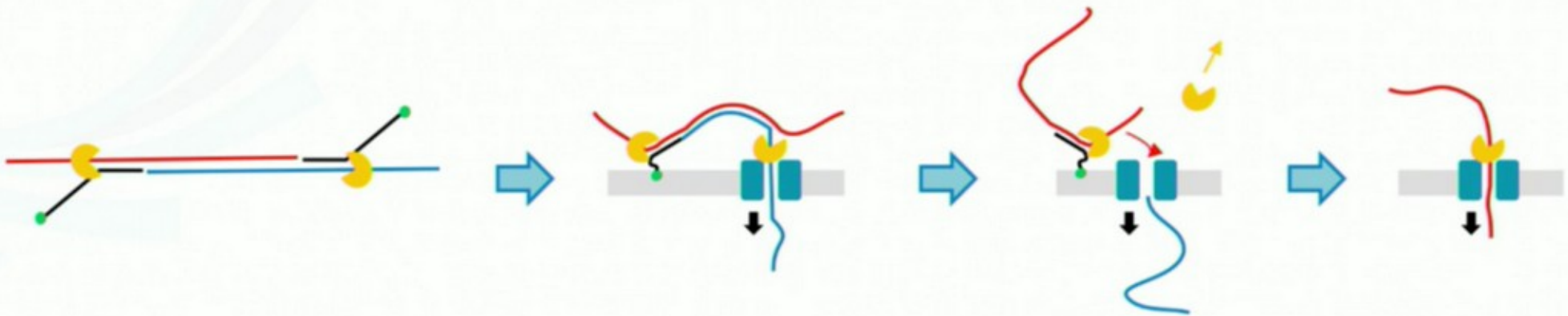
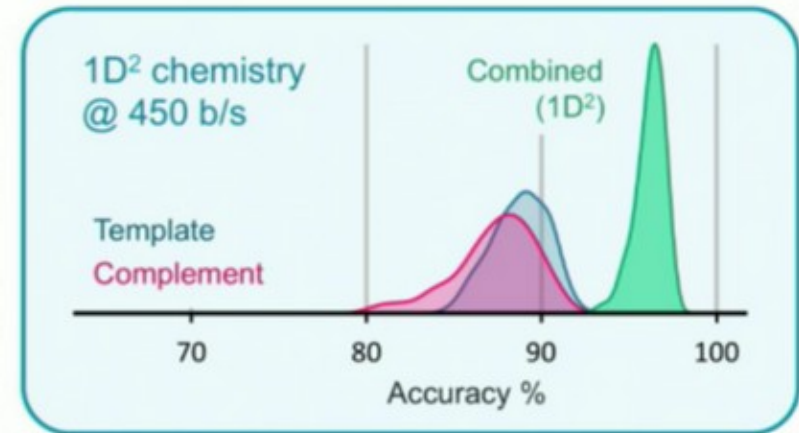
Oxford Nanopore

1D²

Improved template – complement data

SEQUENCING SCHEME WHERE STRANDS ARE NOT JOINED

- Complement follows template as separate independent strand
 - Each molecule has its own motor-adaptor
 - Each individual strand has high 1D accuracy
 - No secondary structure problems
- Simple library preparation, compatible addition to 1D methods
 - Compatible with E8 and 450 bps



Accuracy will improve to 99% with new basecaller
Homopolymer reads will improve with Scrappie basecaller

Different platforms

Third Generation Sequencing: Single Molecule Sequencing

Oxford Nanopore

Movie time

[Oxford Nanopore Sequencing \(YouTube\)](#)

Different platforms

Third Generation Sequencing: Single Molecule Sequencing

SeqLL (Sequence the Lower Limit)

tSMS sequencer

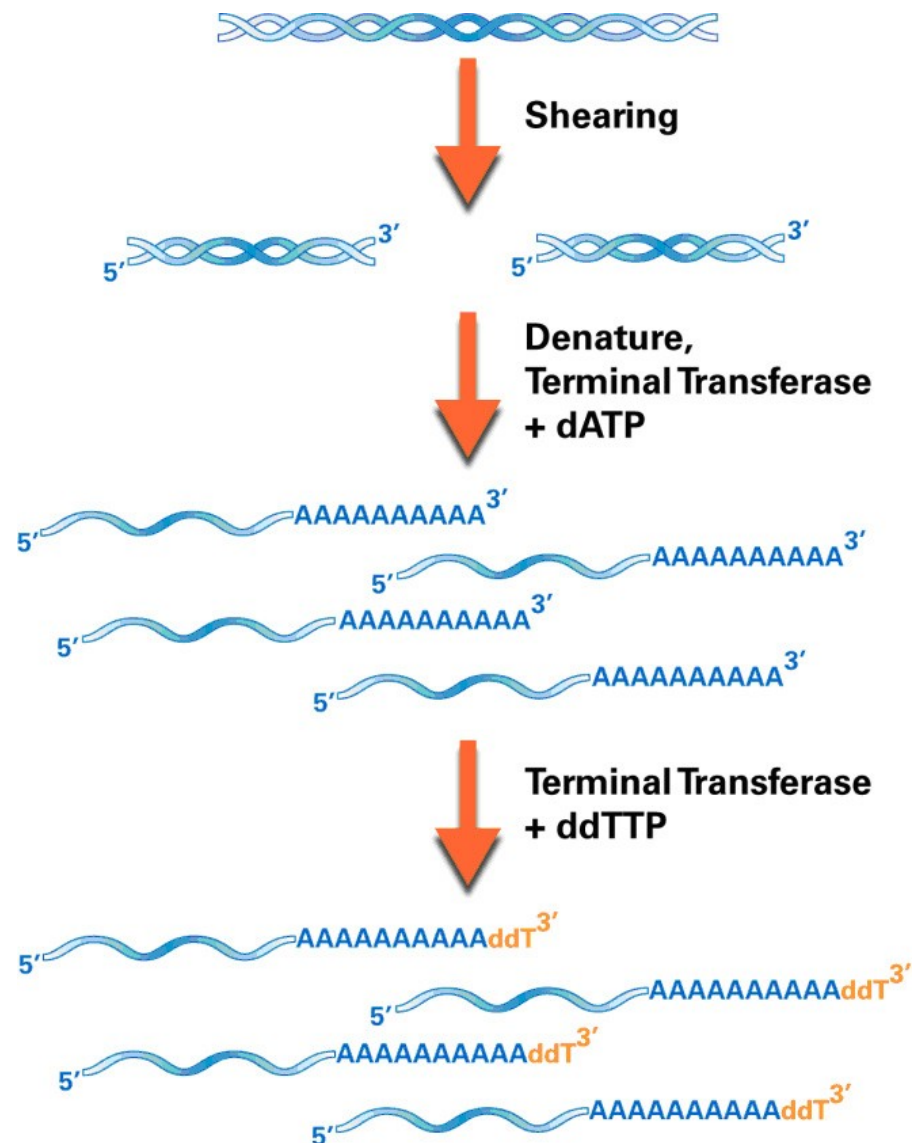


	tSMS Sequencer
Read Length	20-55 bp
Throughput	21-35 Gb/run (2x25 channels/run)
Reads per run	600 – 1000 Million
Accuracy	95-96 % ?
Run Time	30 hours ?

Different platforms

Third Generation Sequencing: Single Molecule Sequencing

SeqLL (Sequence the Lower Limit)

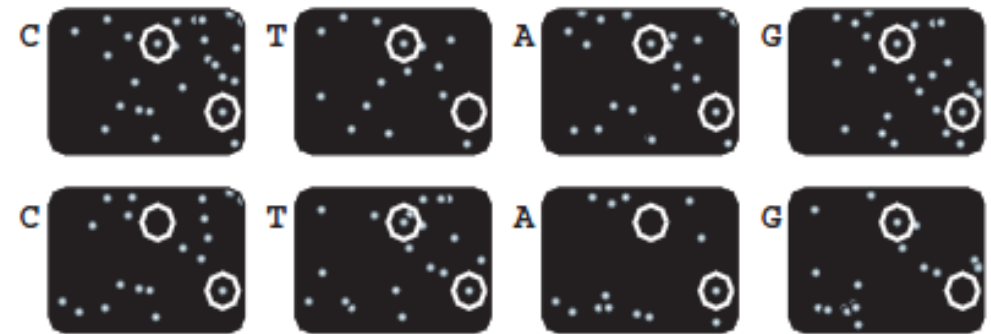
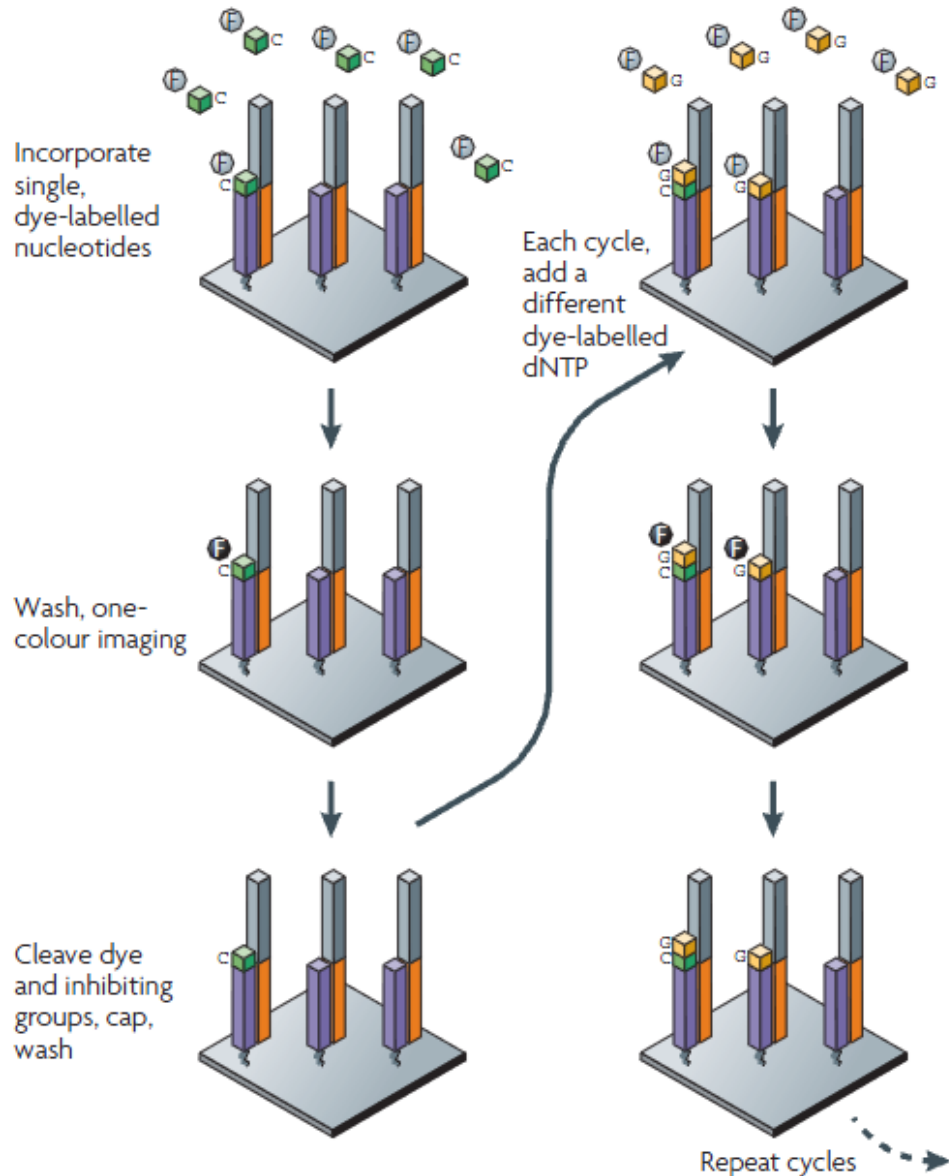


Different platforms

Third Generation Sequencing: Single Molecule Sequencing

SeqLL (Sequence the Lower Limit)

Sequencing by synthesis



Top: CTAGTG
Bottom: CAGCTA

Nucleotides flow sequentially

Different platforms

Third Generation Sequencing: Single Molecule Sequencing

SeqLL (Sequence the Lower Limit)

Movie time

[SeqLL sequencing by synthesis \(Youtube link\)](#)

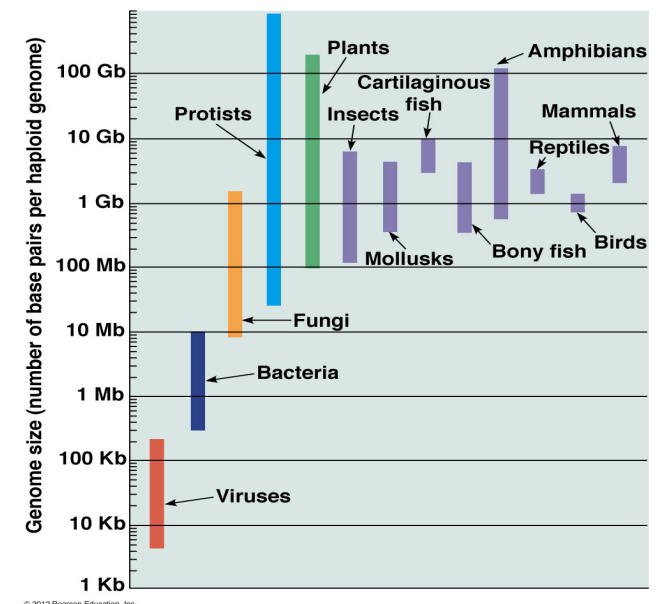
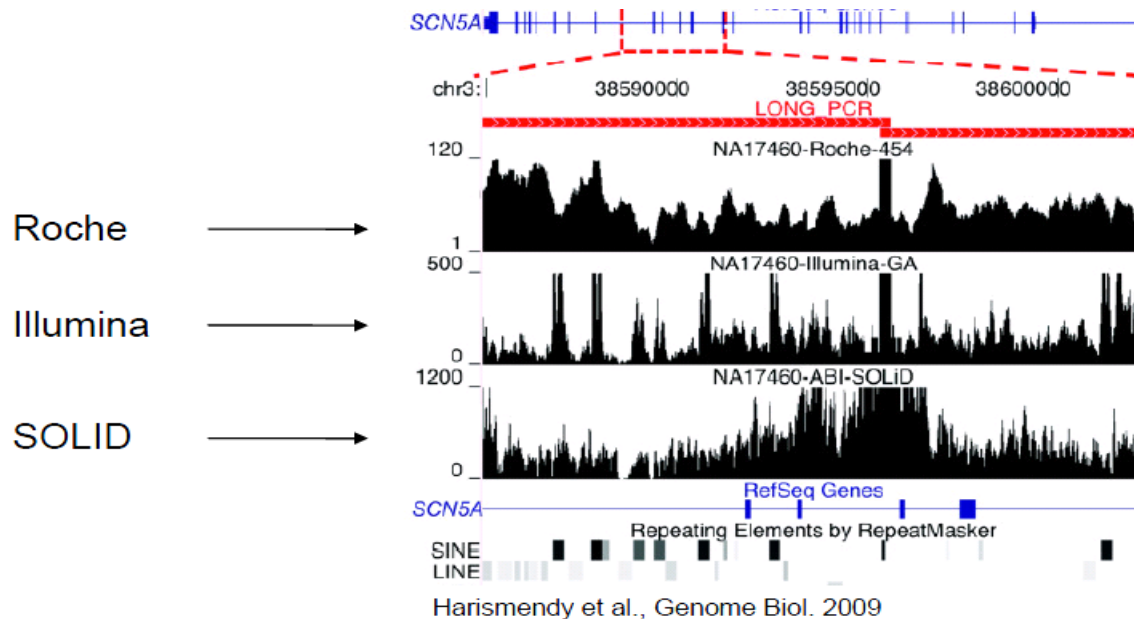
Different platforms

Which Next Generation Sequencer to choose for your project ?

	Capacity	Speed	Read Length	Read Homopolymers	Cost/run (*) 10,000 reads	Amplification
SOLiD	120 - 320 Gb	7 days	75 bp	+	5.000 € ?	Yes
Illumina	7,5 - 3000 Gb	1 - 6 days	125 - 2x300	+	3.000-5.000 €	Yes
Ion Torrent	20 Mb - 15Gb	4 - 17 hours	200 - 400 bp	-	800-3.000 €	Yes
PacBio	1 – 8 Gb	0,5 – 10 h	> 20,000 bp	+	600-800 € ?	No
Oxford nanopore	10 – 20 Gb	1 – 48 h	>20,000 bp	+/-	< 1000 €	No
SeqLL	21-35 Gb	30 h ?	20-55 bp	+	?	No

(*) only sequencing cost, for a minimum of 10,000 reads (more reads → becomes cheaper)

<https://genohub.com/> <https://www.scienceexchange.com/browse?category=ngs>



Quality scores in sequencing

Sequencing: homopolymer problems

Ion Torrent



Quality scores in sequencing

Sequencing: homopolymer problems

Illumina HiSeq



Quality scores in sequencing

95/165

Sequencing: quality scores (Phred scores) and accuracy

Quality scores in sequencing: Q17, Q20, Q30, ... is a probability

Quality score	Probability of incorrect bases	Base call accuracy
8	1 in 6	84 %
10	1 in 10	90 %
15	1 in 30	97%
17	1 in 50	98 %
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%

Q10: 90,0% chance that the base is correct
Q30: 99,9% chance that the base is correct

1 Gb genome: 1 time coverage:
Q20: possible 10.000.000 errors
Q30: possible 1.000.000 errors
More coverage reduce the errors

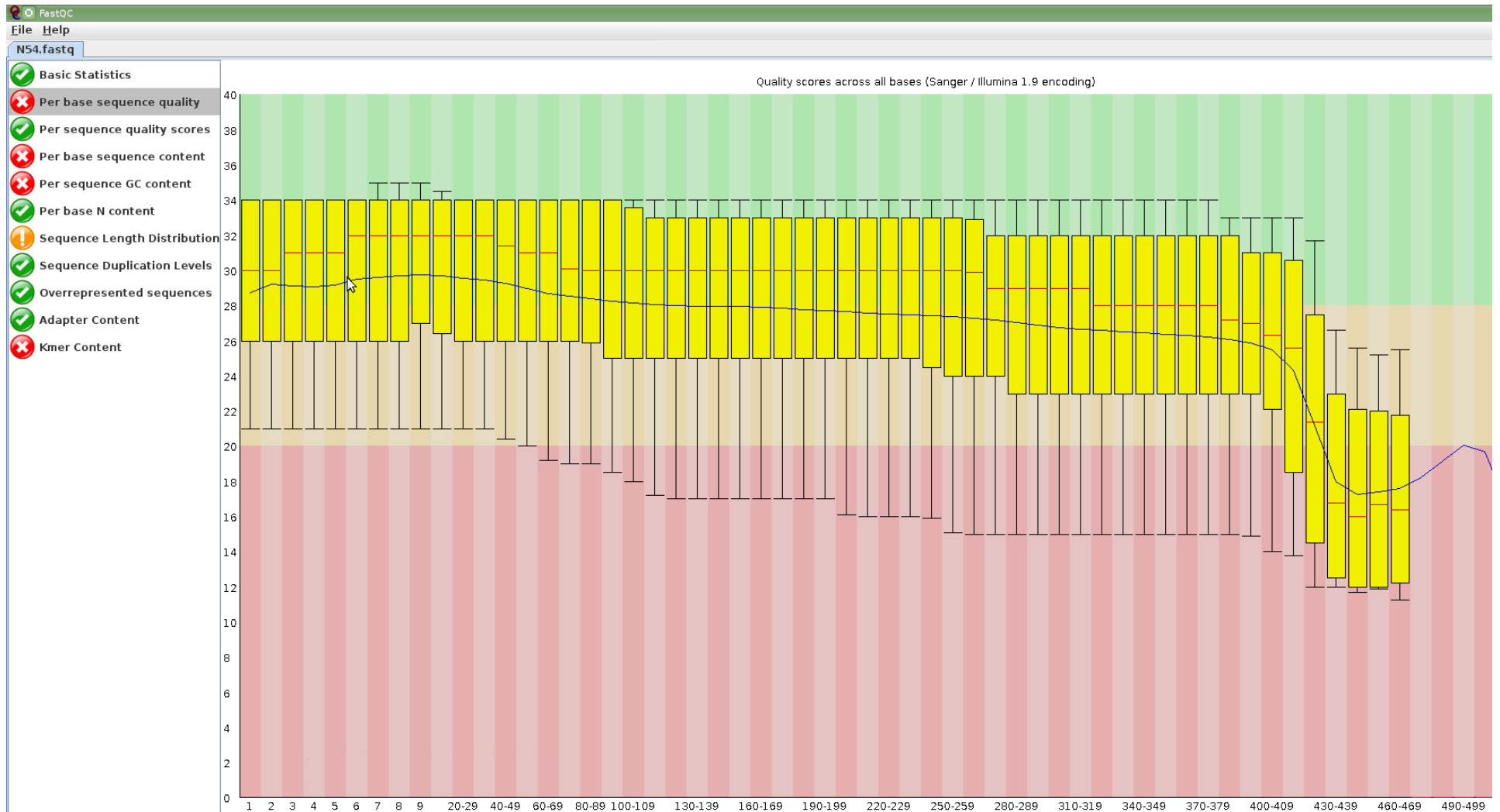
Quality scores in sequencing

96/165

Sequencing: quality scores (Phred scores) and accuracy

FastQC : average quality from the sequencing run

Ion Torrent

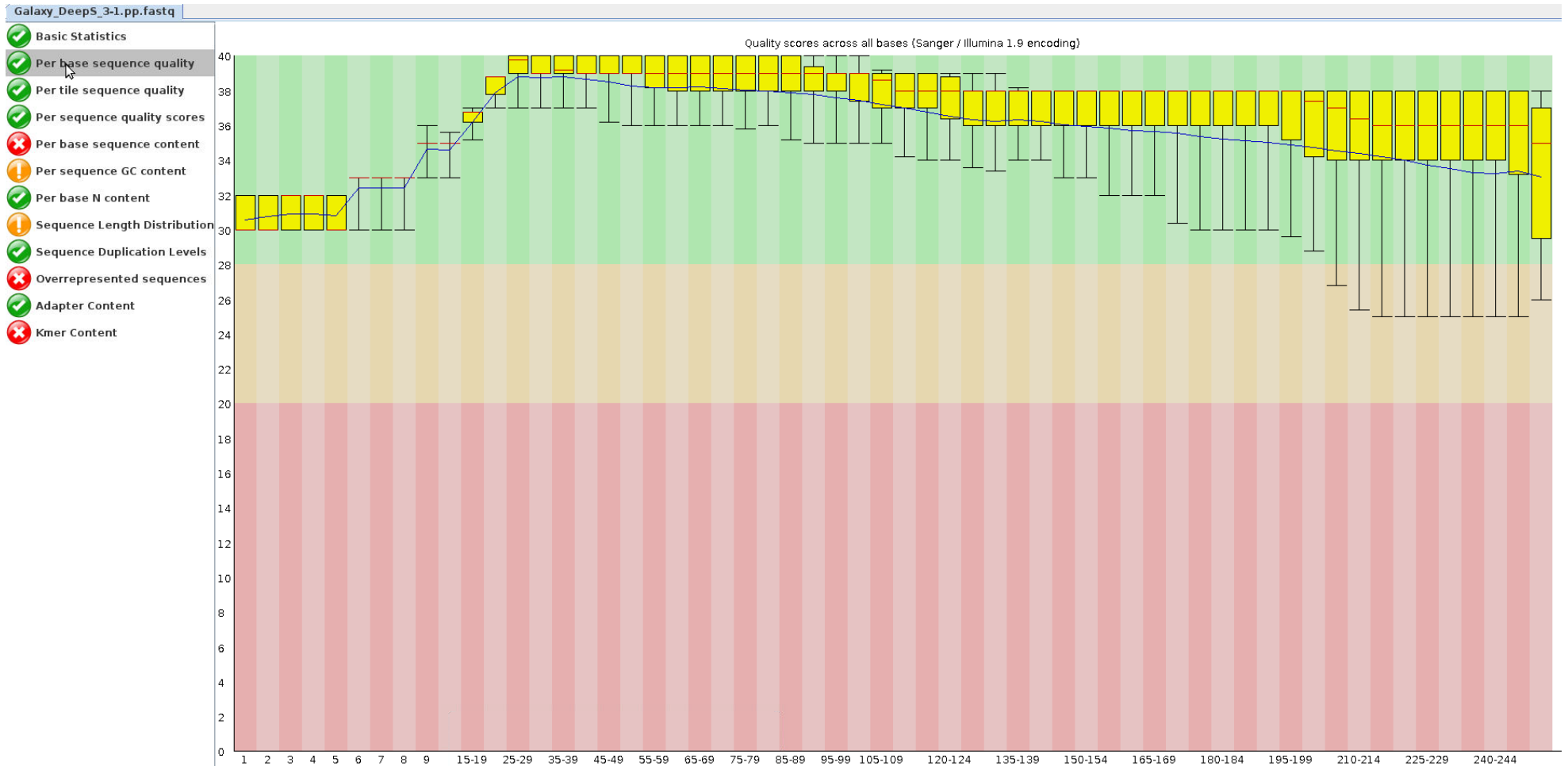


Quality scores in sequencing

Sequencing: quality scores (Phred scores) and accuracy

FastQC : average quality from the sequencing run

MiSeq forward sequence

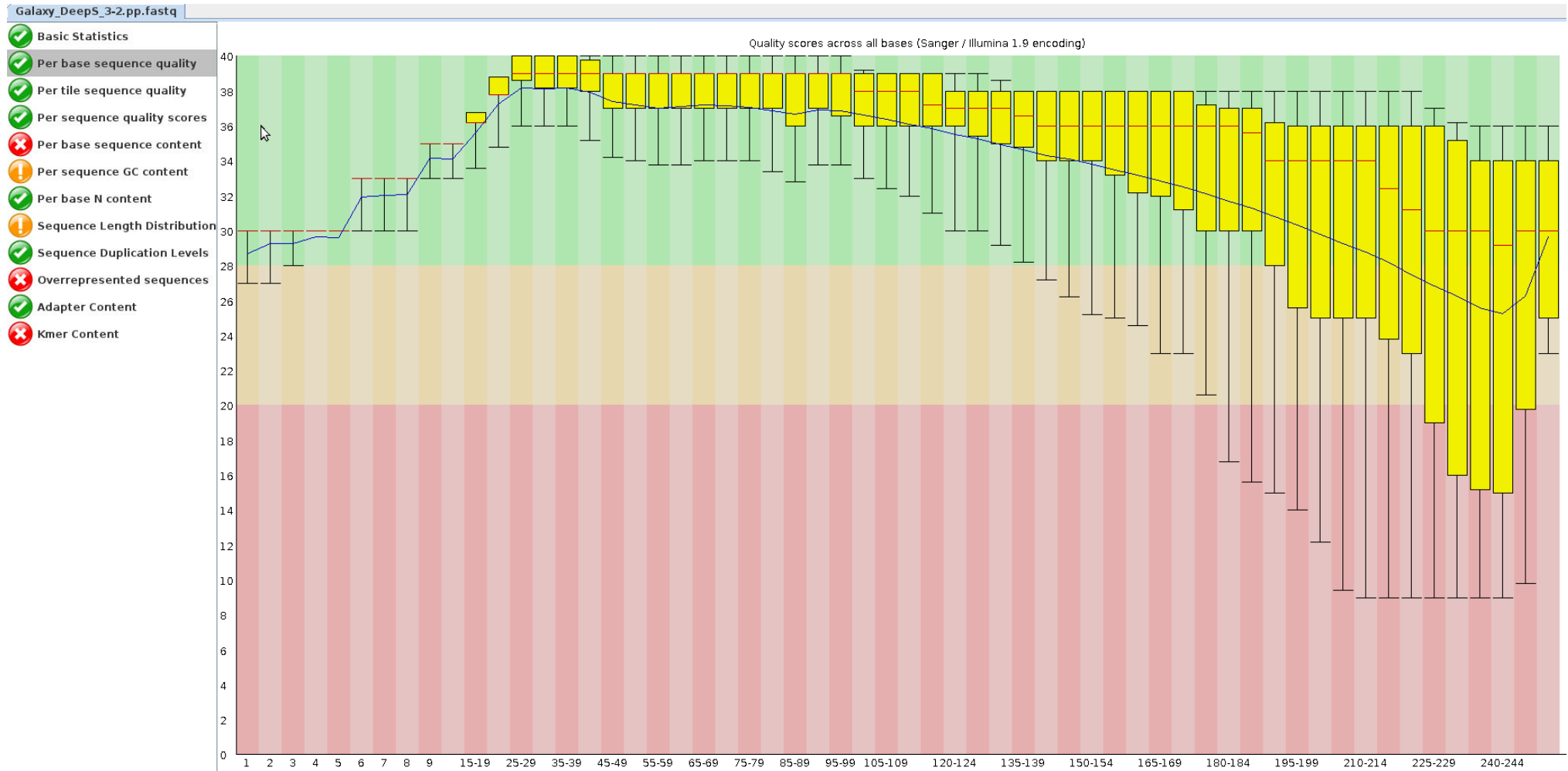


Quality scores in sequencing

Sequencing: quality scores (Phred scores) and accuracy

FastQC : average quality from the sequencing run

MiSeq reverse sequence

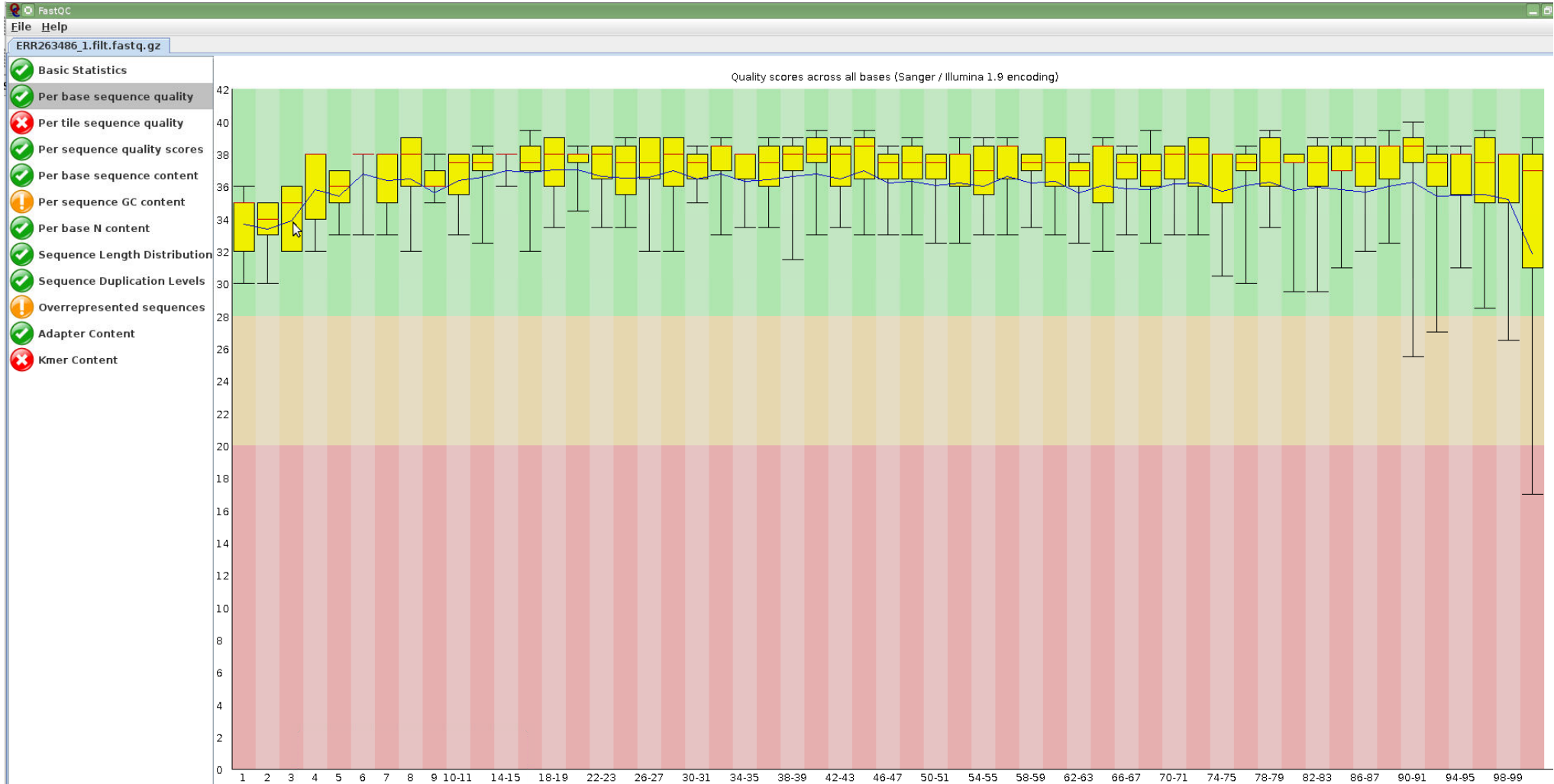


Quality scores in sequencing

Sequencing: quality scores (Phred scores) and accuracy

FastQC : average quality from the sequencing run

HiSeq



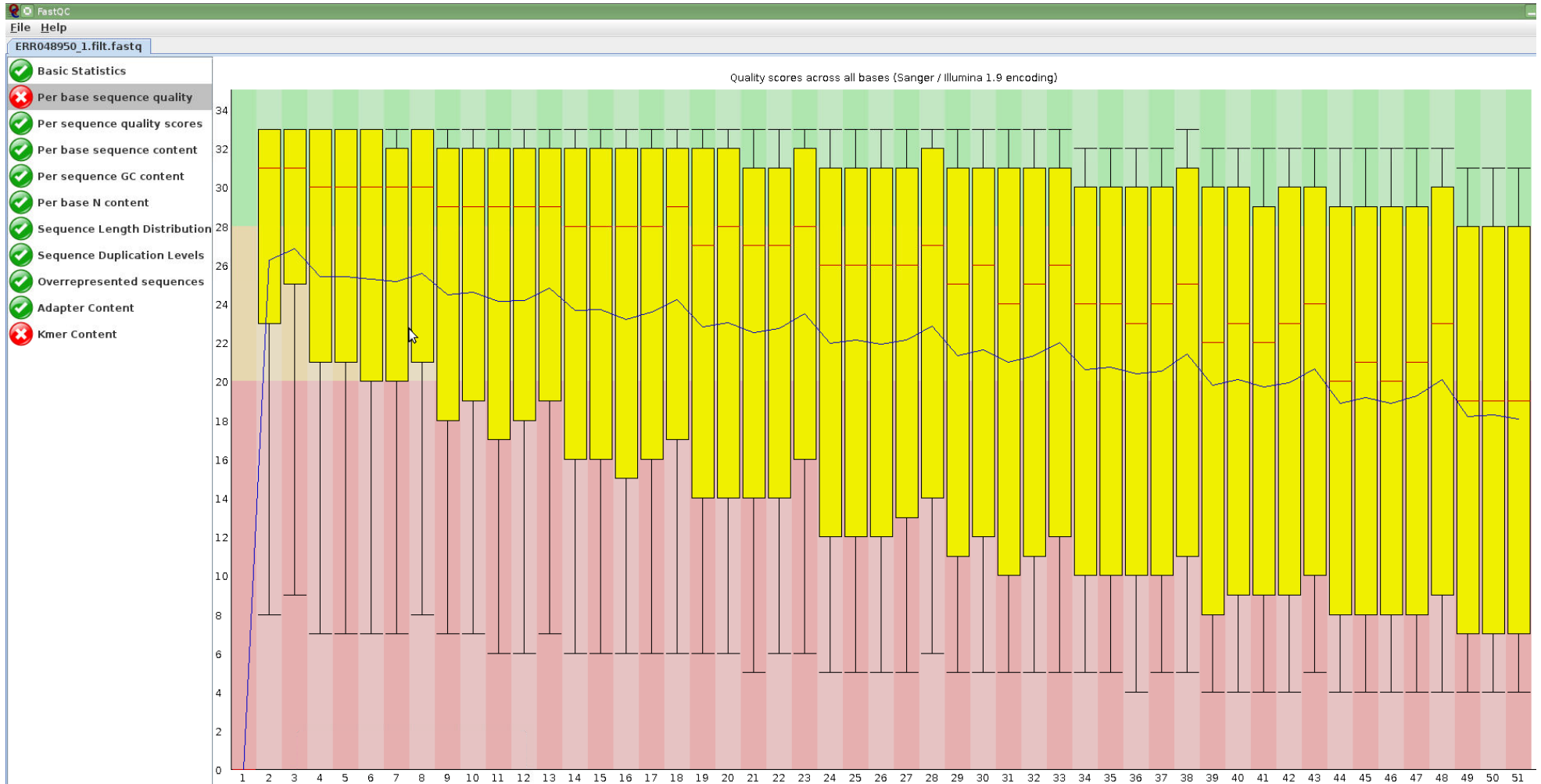
Quality scores in sequencing

100/165

Sequencing: quality scores (Phred scores) and accuracy

FastQC : average quality from the sequencing run

SOLiD



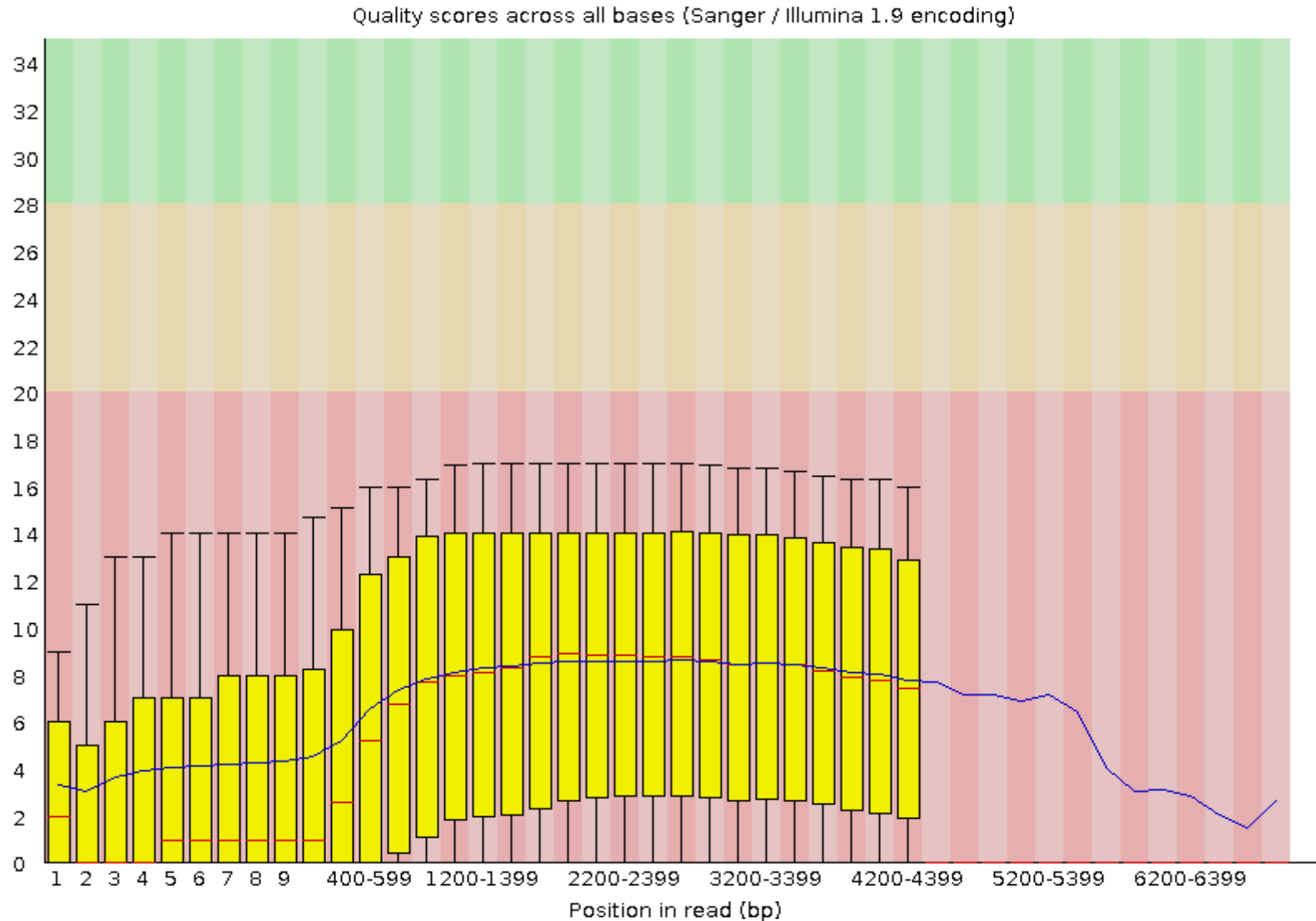
Quality scores in sequencing

101/165

Sequencing: quality scores (Phred scores) and accuracy

FastQC : average quality from the sequencing run

PacBio (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/pacbio_srr075104_fastqc.html)



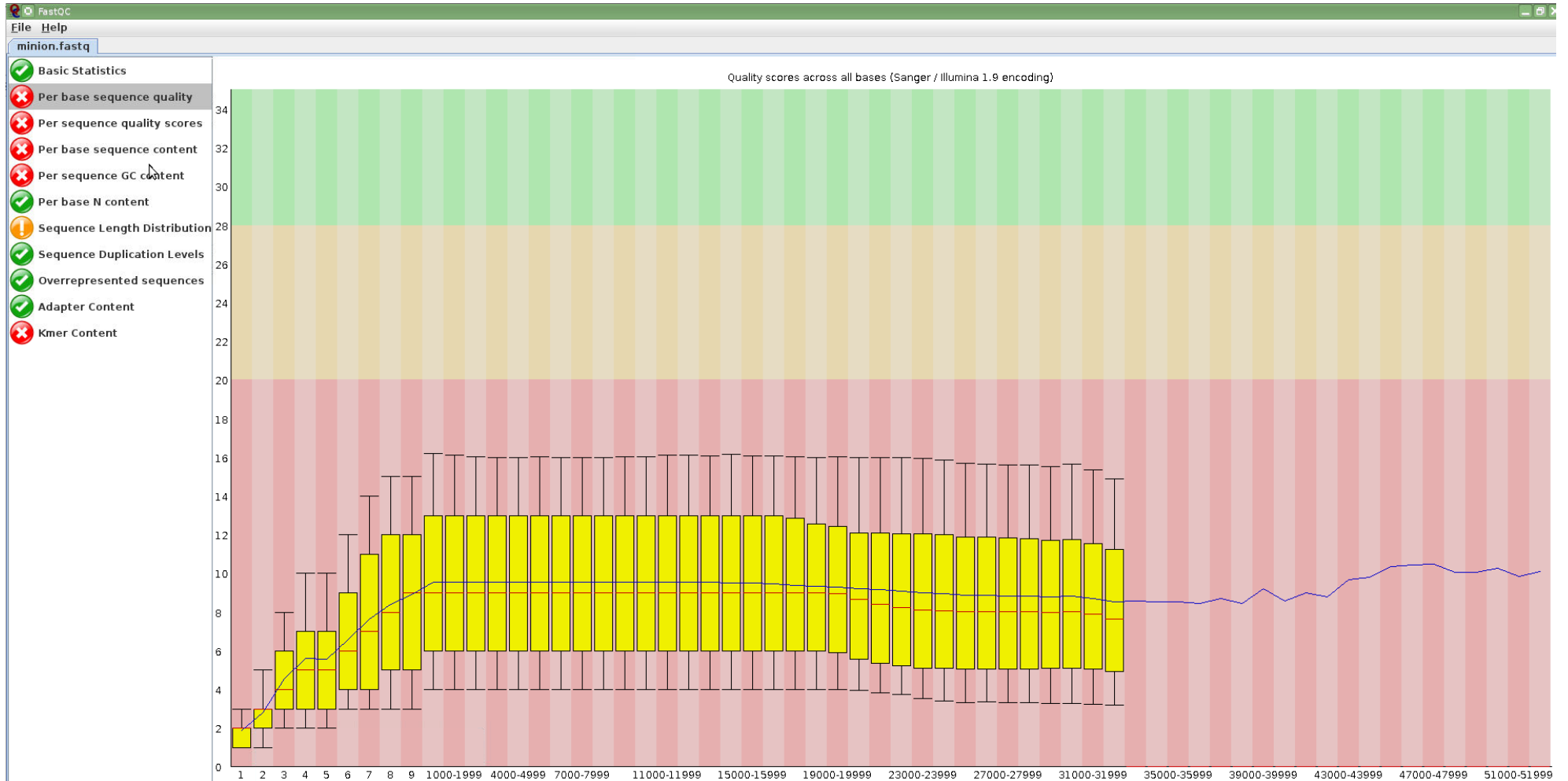
Quality scores in sequencing

102/165

Sequencing: quality scores (Phred scores) and accuracy

FastQC : average quality from the sequencing run

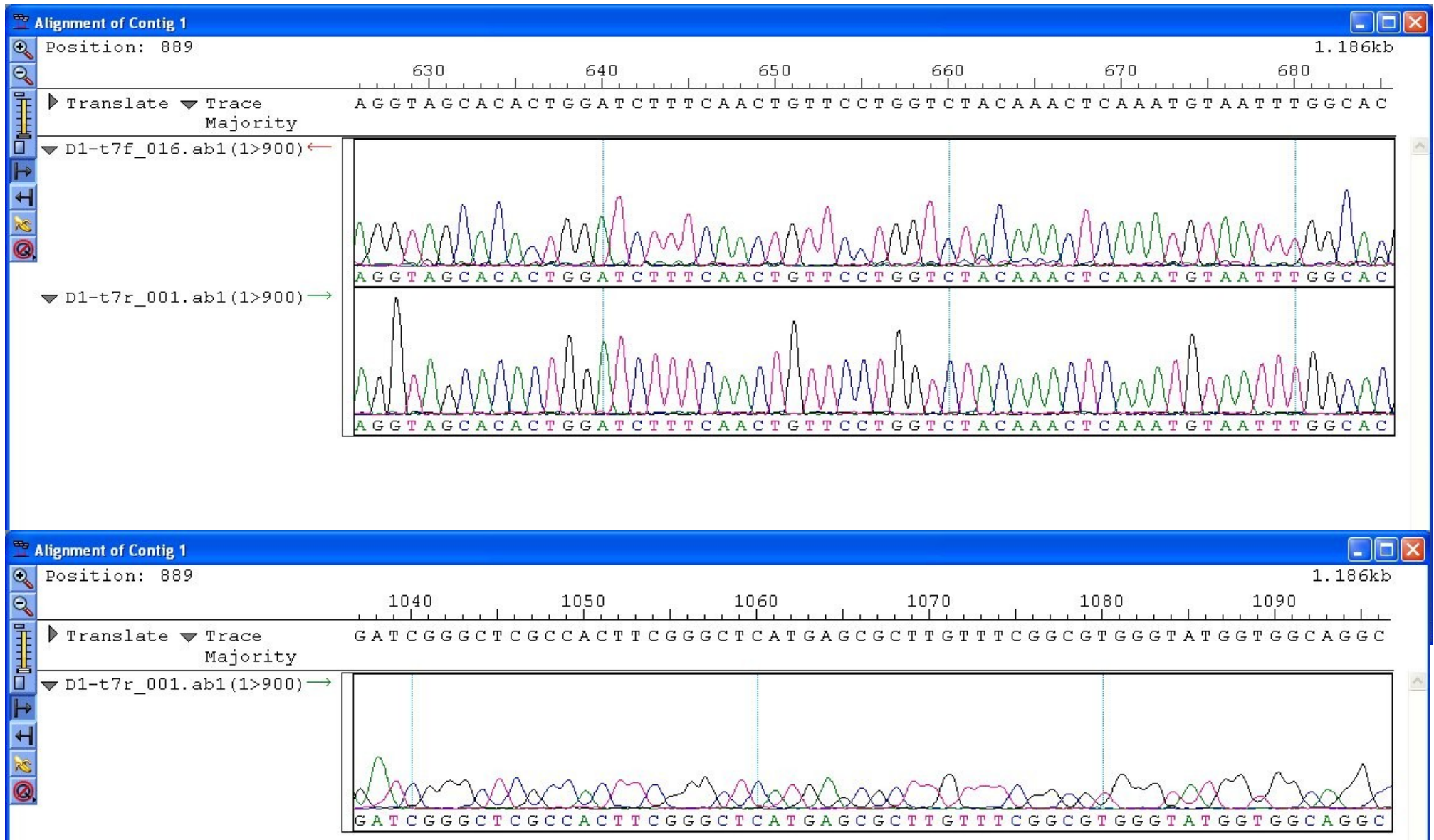
MinION



Quality scores in sequencing

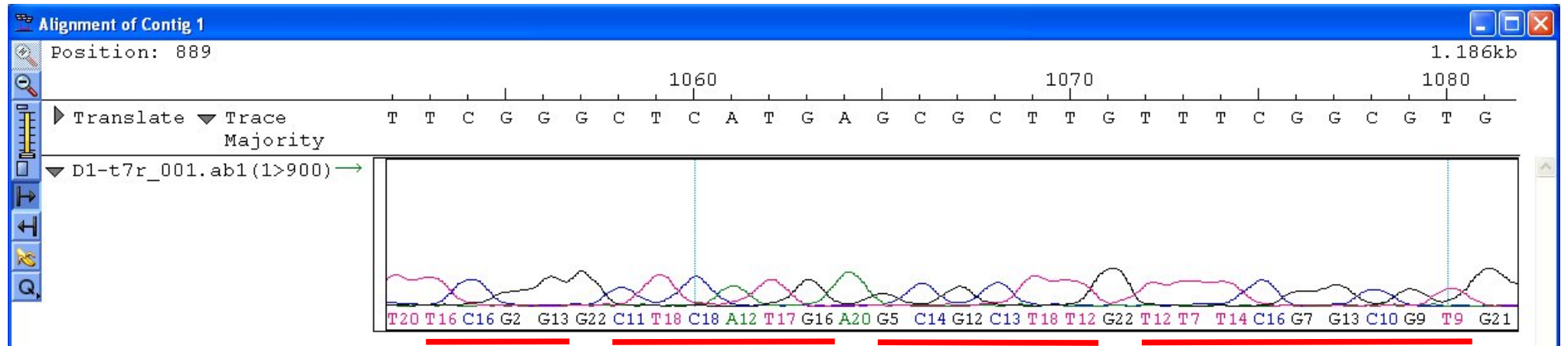
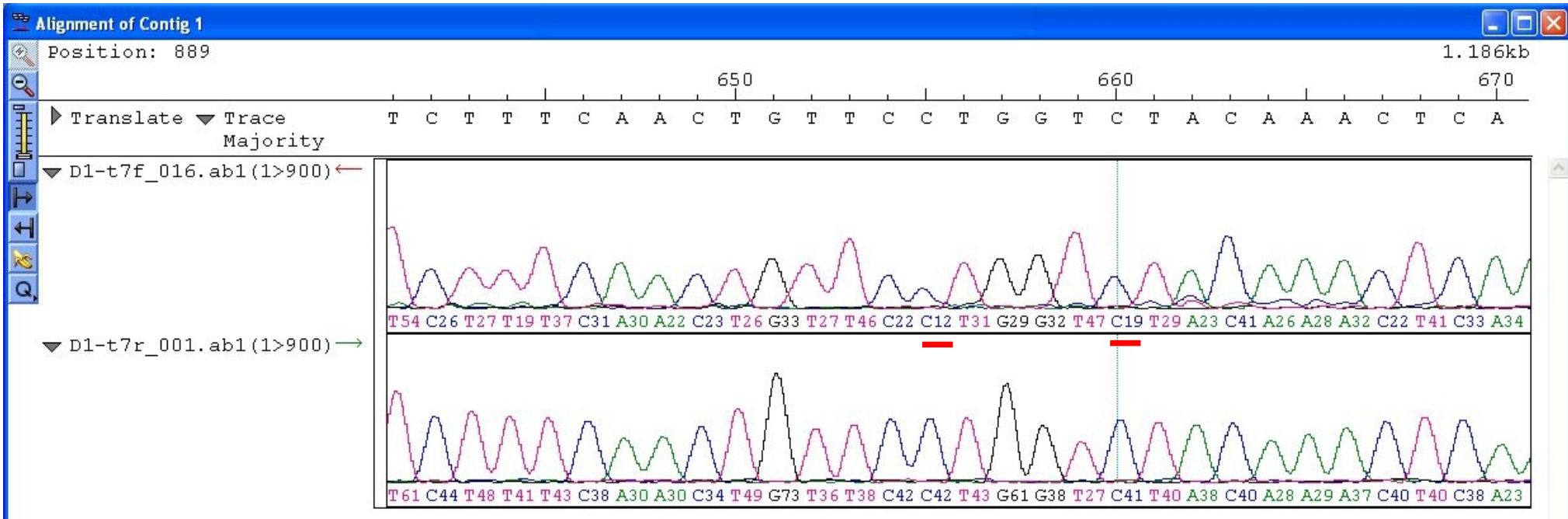
103/165

Sequencing: quality scores (Phred scores) and accuracy



Quality scores in sequencing

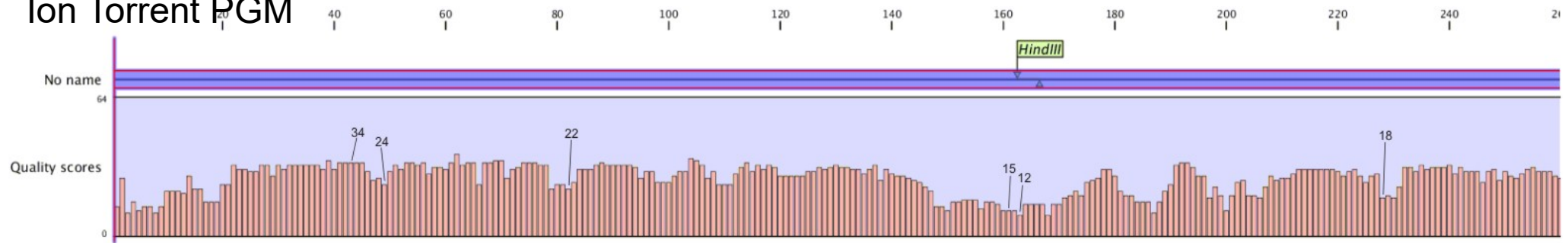
Sequencing: quality scores (Phred scores) and accuracy



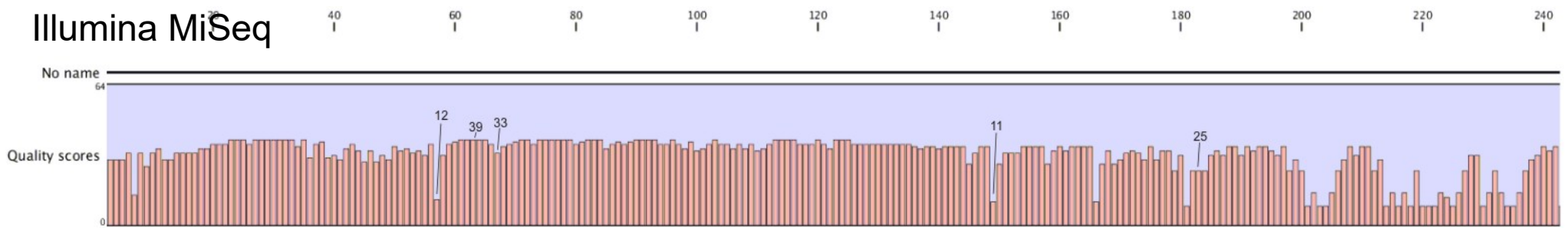
Quality scores in sequencing

Sequencing: quality scores (Phred scores) and accuracy

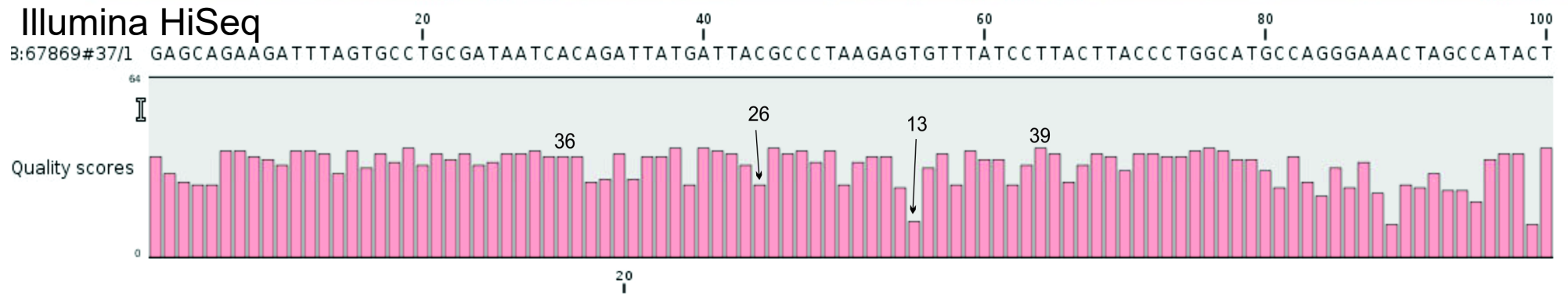
Ion Torrent PGM



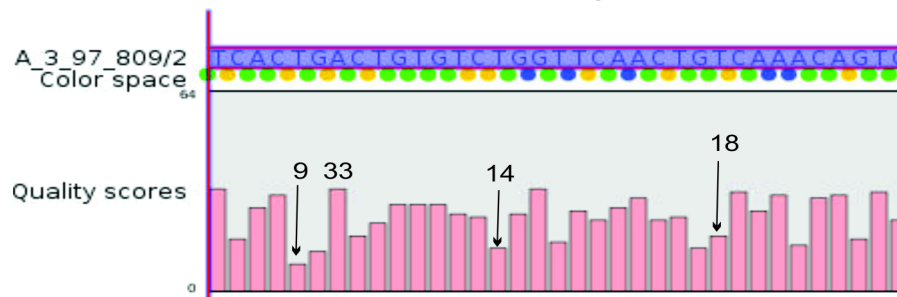
Illumina MiSeq



Illumina HiSeq



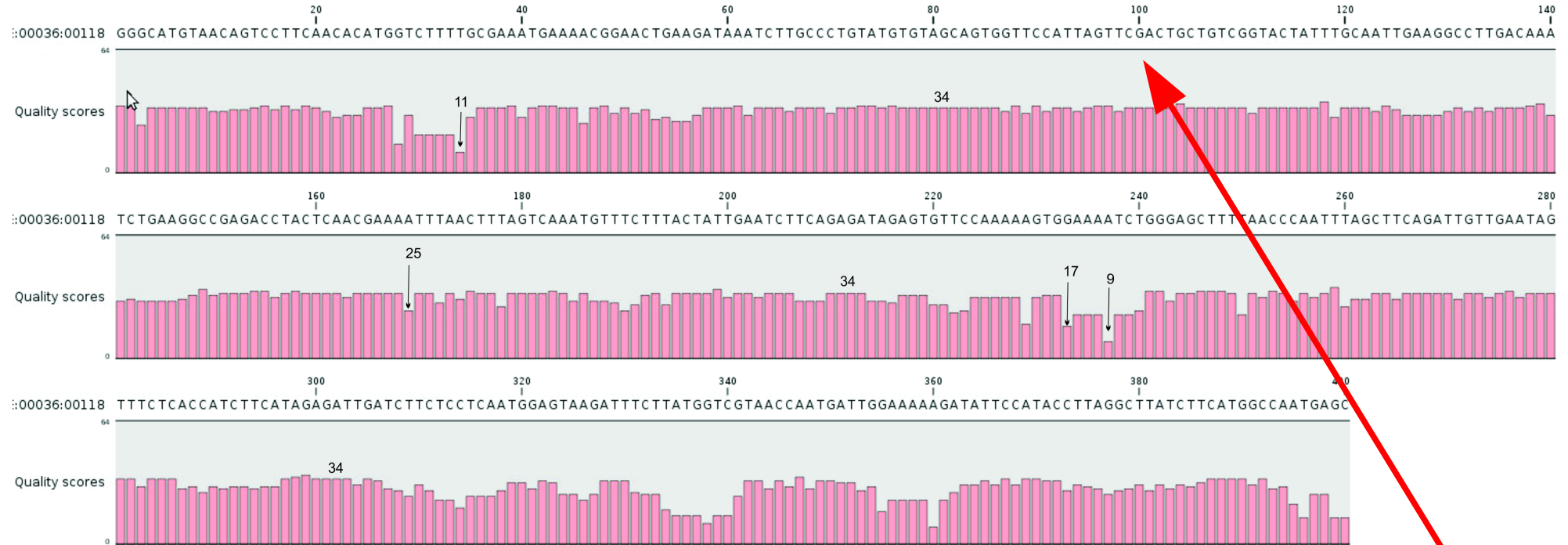
SOLiD



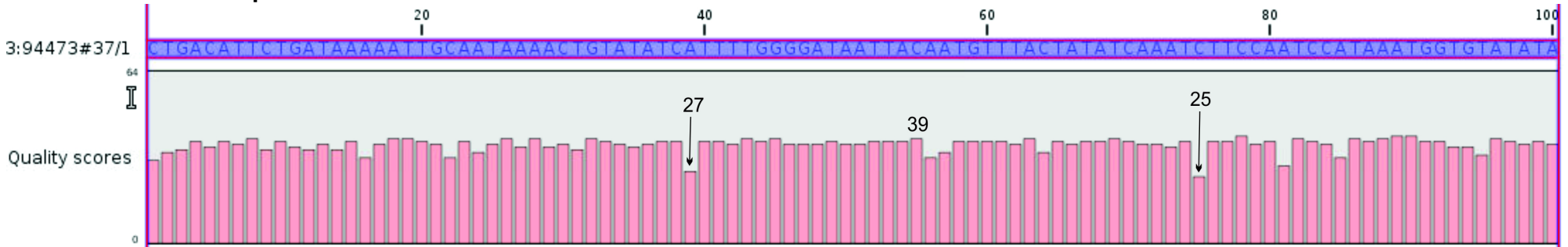
Quality scores in sequencing

Sequencing: quality scores (Phred scores) and accuracy

Ion Torrent PGM



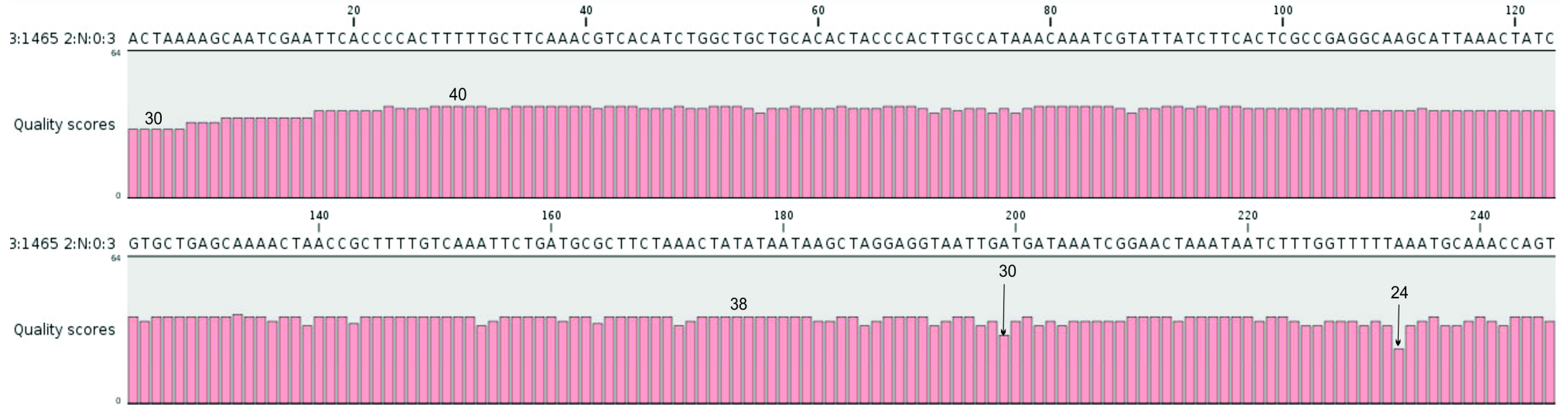
Illumina HiSeq



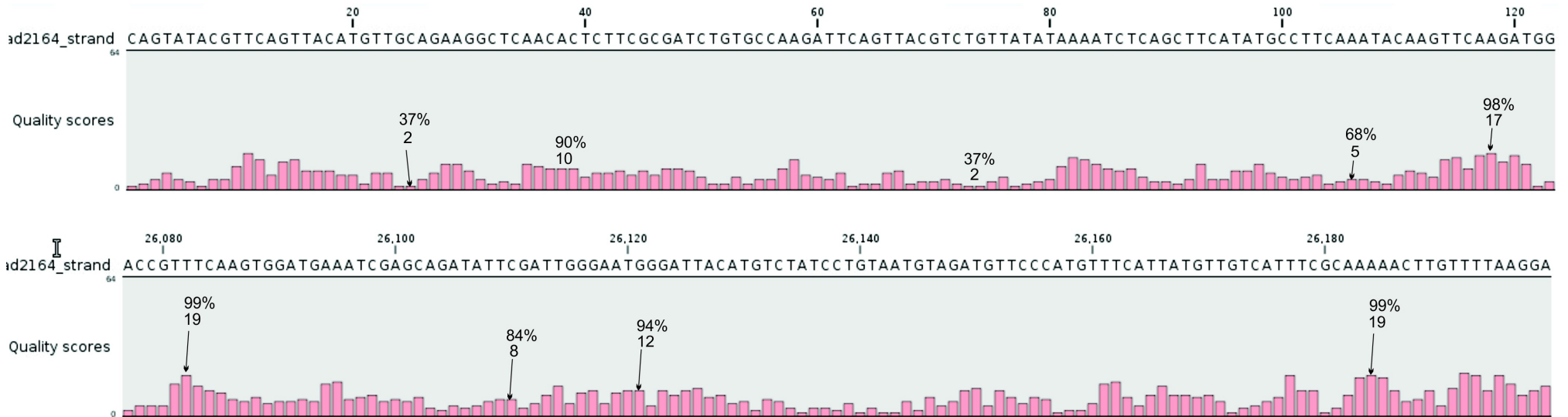
Quality scores in sequencing

Sequencing: quality scores (Phred scores) and accuracy

Illumina MiSeq



Minlon

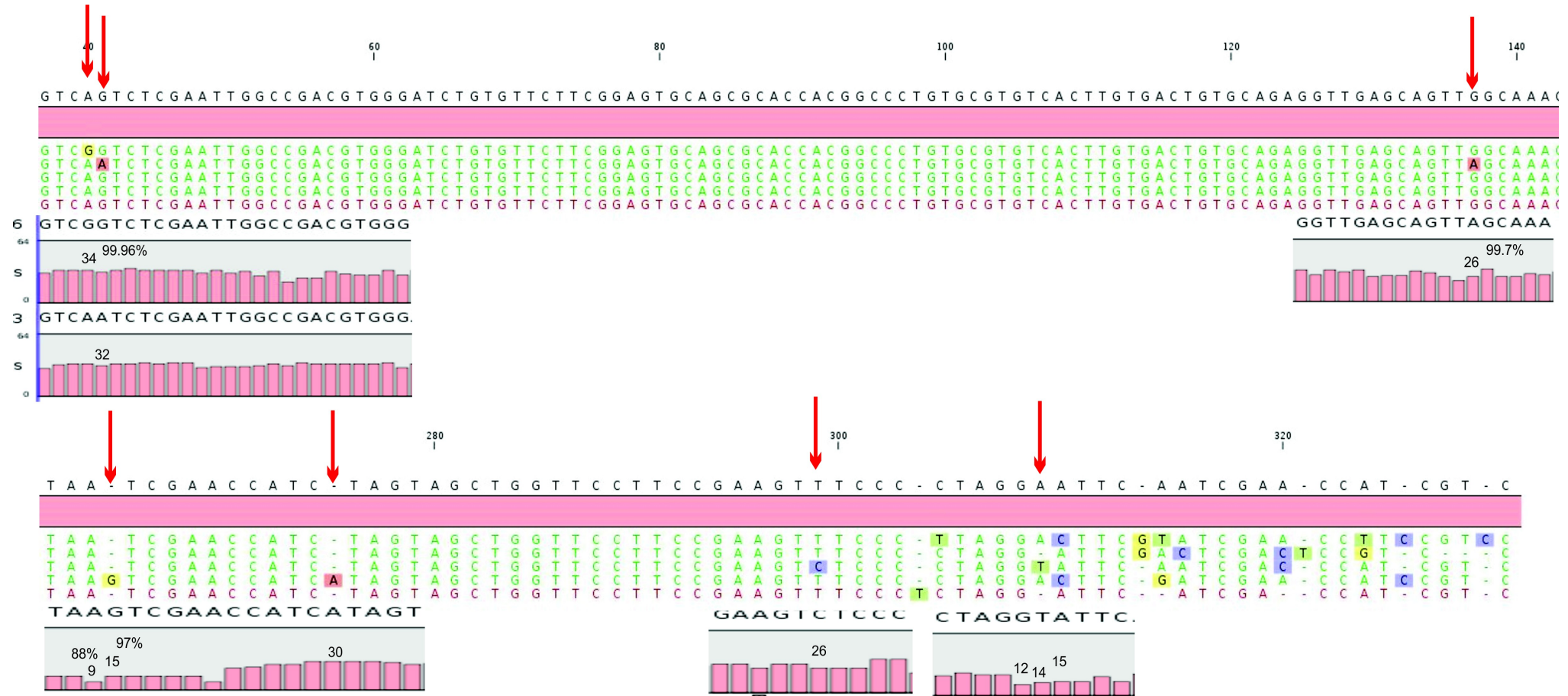


Quality scores in sequencing

Sequencing: quality scores (Phred scores) and accuracy

IonTorrent Assembly

?? how important are quality scores ??

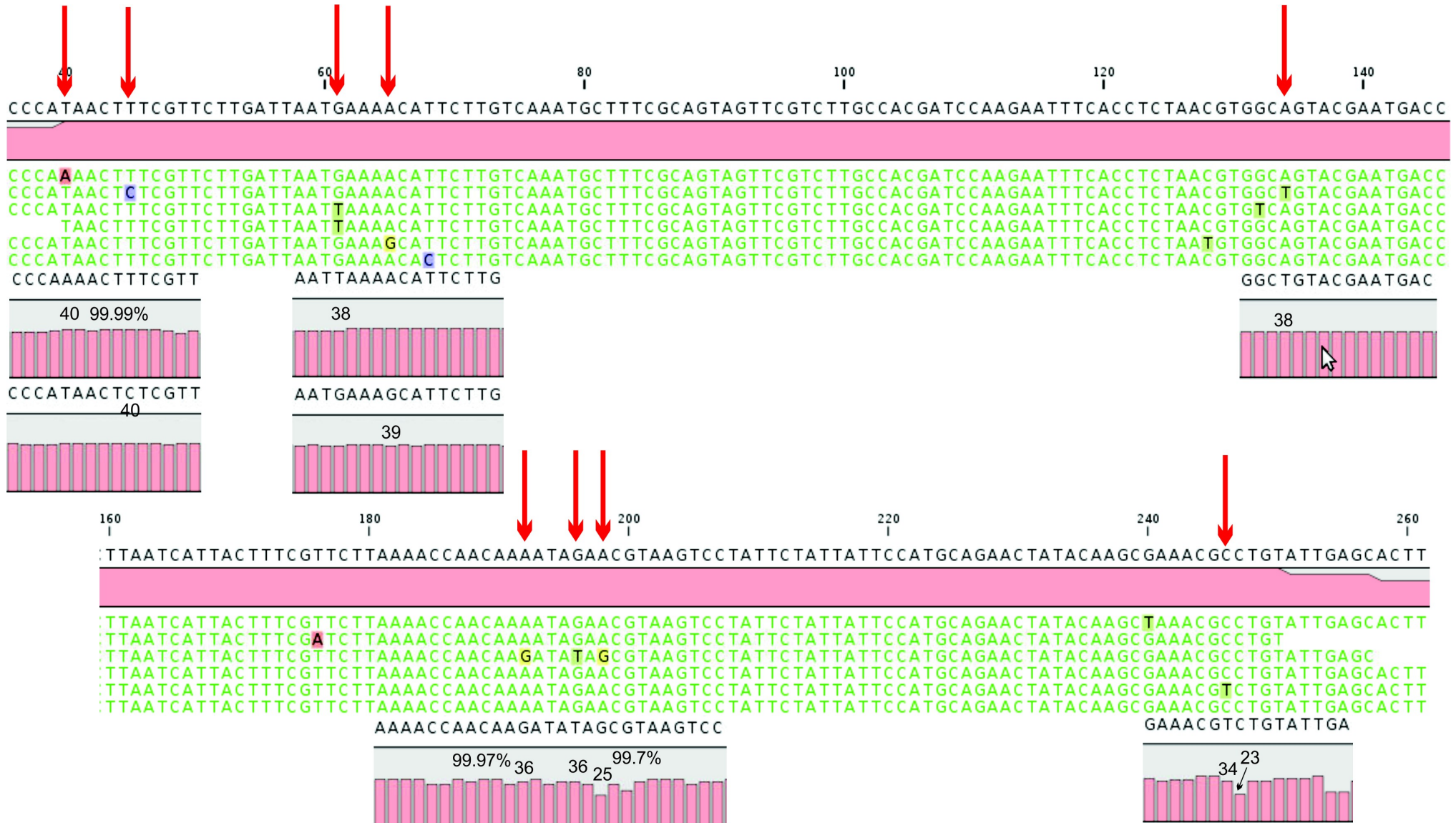


Quality scores in sequencing

Sequencing: quality scores (Phred scores) and accuracy

MiSeq Assembly (Halomonhystera sp.)

?? how important are quality scores ??



→ consensus sequences is ok

Quality scores in sequencing

110/165

Sequencing: quality scores (Phred scores) and accuracy

What can cause sequencing errors ?

- Oxidation artifact induced by acoustic shearing of DNA in library prep.

(Costello, 1013. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation)

- PCR errors in library prep
- PCR errors in emulsion PCR or Polony PCR

Taq: $2,28 \times 10^{-5} \rightarrow 2,28$ errors in 100,000 bases polymerized

600 bp template: $100,000/600=167$ fragments

2,28 fragments per 167 fragments contain 1 wrong base (1,36%) after 1 pcr cycle

50M reads (MiSeq) => 680,000 reads with 1 error.

High Fidelity polymerase: $4,4 \times 10^{-7} \rightarrow 4,4$ errors in 10,000,000 bases polymerized

600 bp template: $10,000,000/600=16,667$ fragments

4,4 fragments per 16,667 fragments contain 1 wrong base (0,026%) after 1 pcr cycle

50M reads (MiSeq) => 13,000 reads with 1 error.

Polymerase	600 bp template (% PCR products with an error)		MiSeq 50 M reads (# reads with an error)	
	5 cycles	20 cycles	5 cycles	20 cycles
Taq (error rate $2,28 \times 10^{-5}$)	6,84%	27,36%	3,42M	13,68M
High Fidelity Taq (error rate $4,4 \times 10^{-7}$)	0,132%	0,528%	66,000	264,000

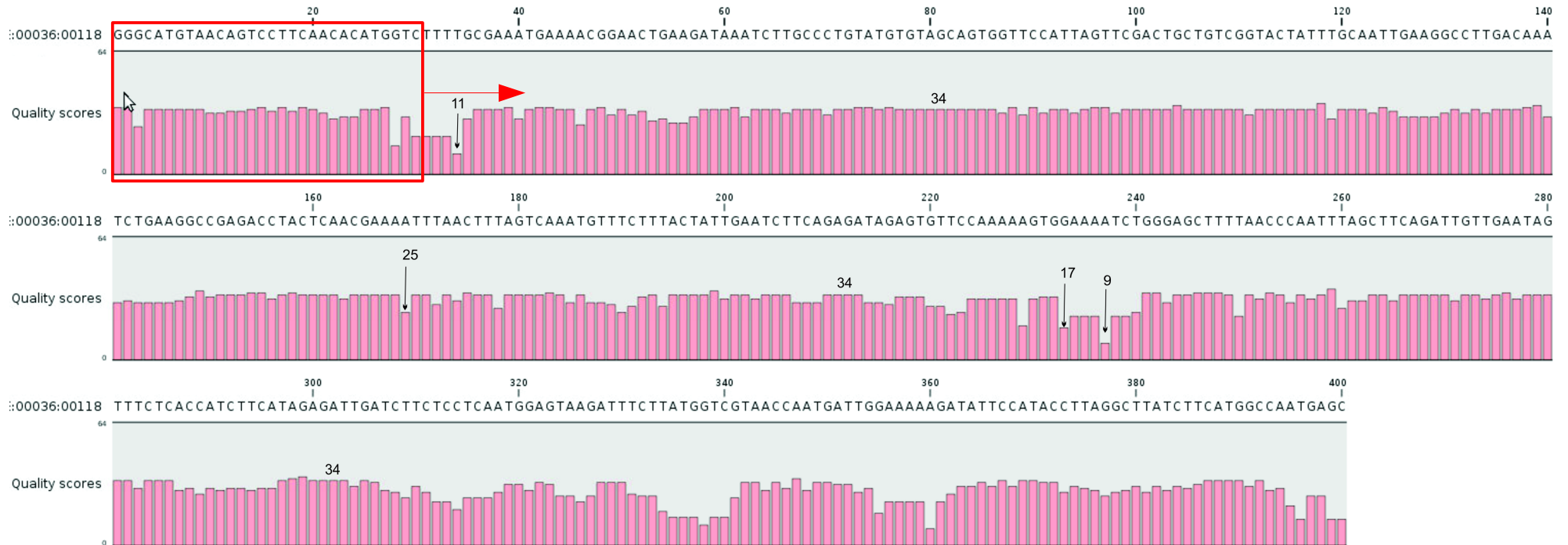
- indel sequencing errors (Illumina 0,005% - Ion Torrent 1%)
 - substitution sequencing errors (Illumina 0,1% - Ion Torrent 0,08%)
 - DNA damage: oxidated G to T transversion during amplification step
- (Chen, 2016. DNA damage is a major cause of sequencing errors, directly confounding variant identification.)
- DNA damage in cells by aging
 - ... ?

Quality scores in sequencing

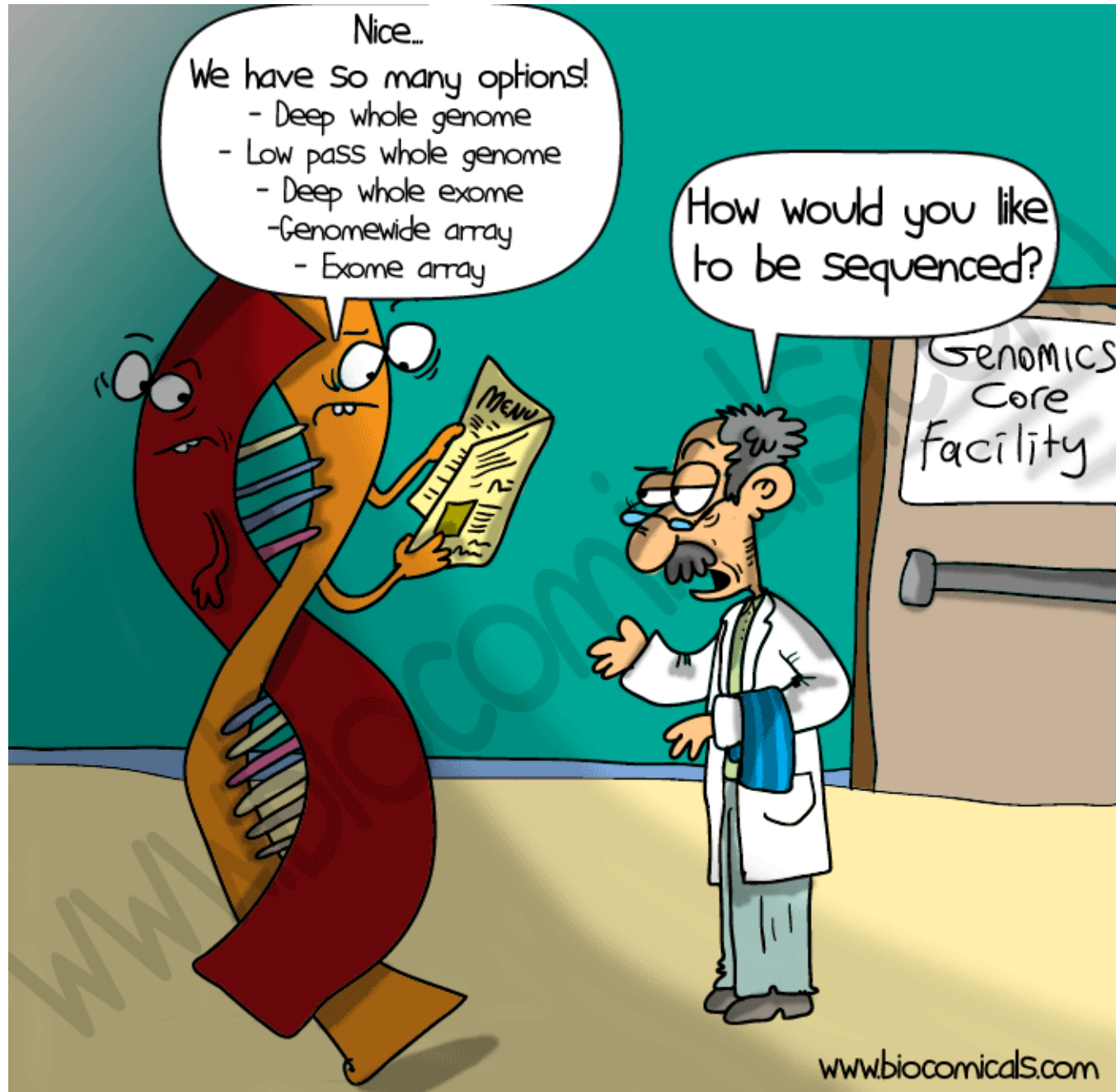
111/165

Sequencing: quality scores (Phred scores) and accuracy

Quality trimming of reads with sliding window 30, cutoff 15



Applications

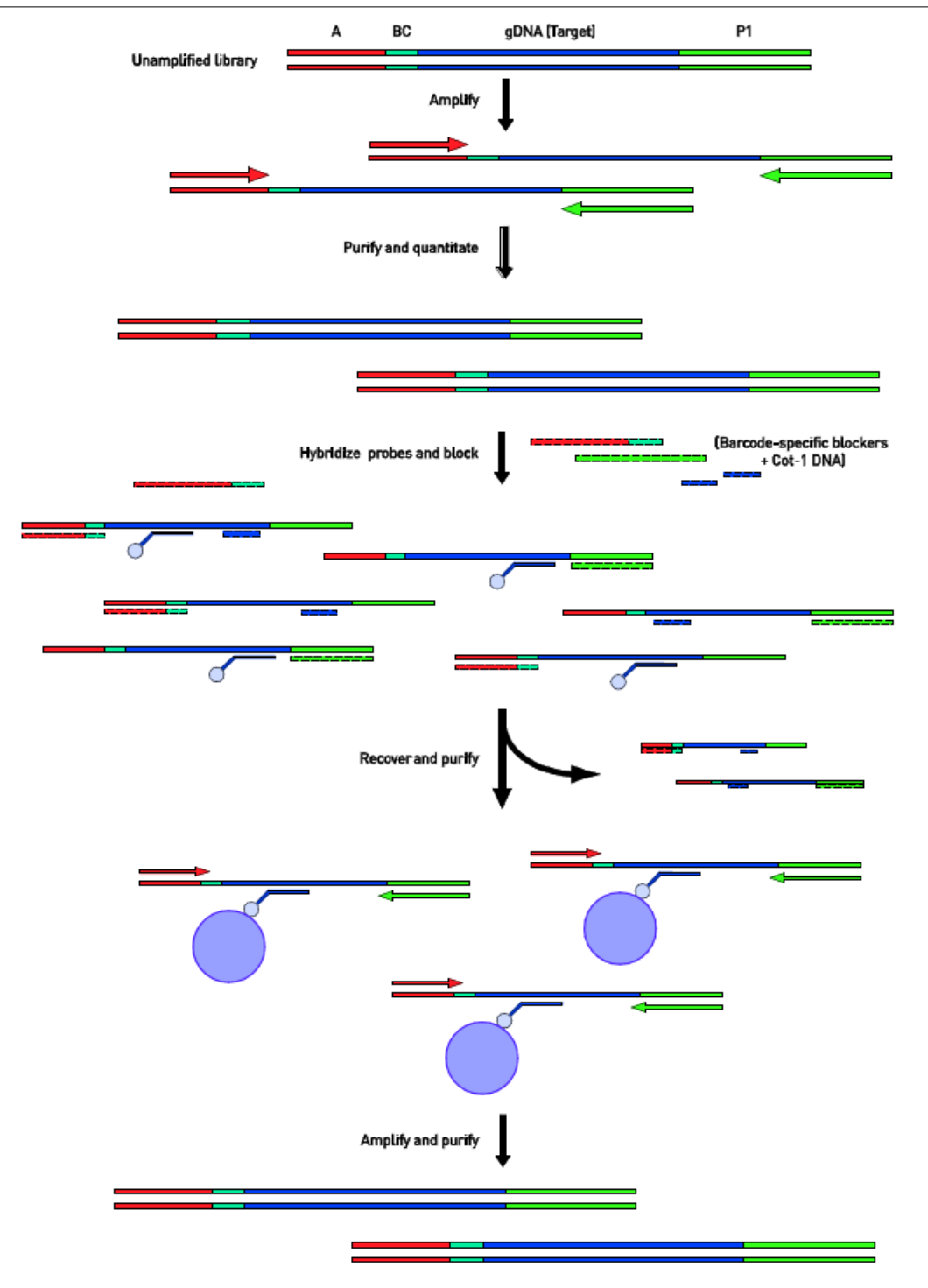


Applications

- Genome Sequencing
 - De novo sequencing
 - Resequencing
 - Targeted (re)sequencing
 - Mitochondrial sequencing
 - Mutation detection
 - Amplicon sequencing
 - Amplicon Cancer Panel
 - Phylogenomics- Phylogenetics
- Transcript Expression Profiling
 - RNA sequencing
 - miRNA sequencing
 - Deep-SAGE
 - Deep-CAGE
 - PAS: polyadenylation site
- Transcription factor binding : ChIP sequencing
- Structural variation
- Metagenomics / Metagenetics / Metatranscriptomics
- Microsatellite sequencing
- Genetic marker discovery
 - Microsatellite development
 - RADSeq (Restriction-site-Associated Sequencing)
 - RRLs (Reduced-Representation libraries)
 - GBS (Genotyping By Sequencing)
- ...

Applications

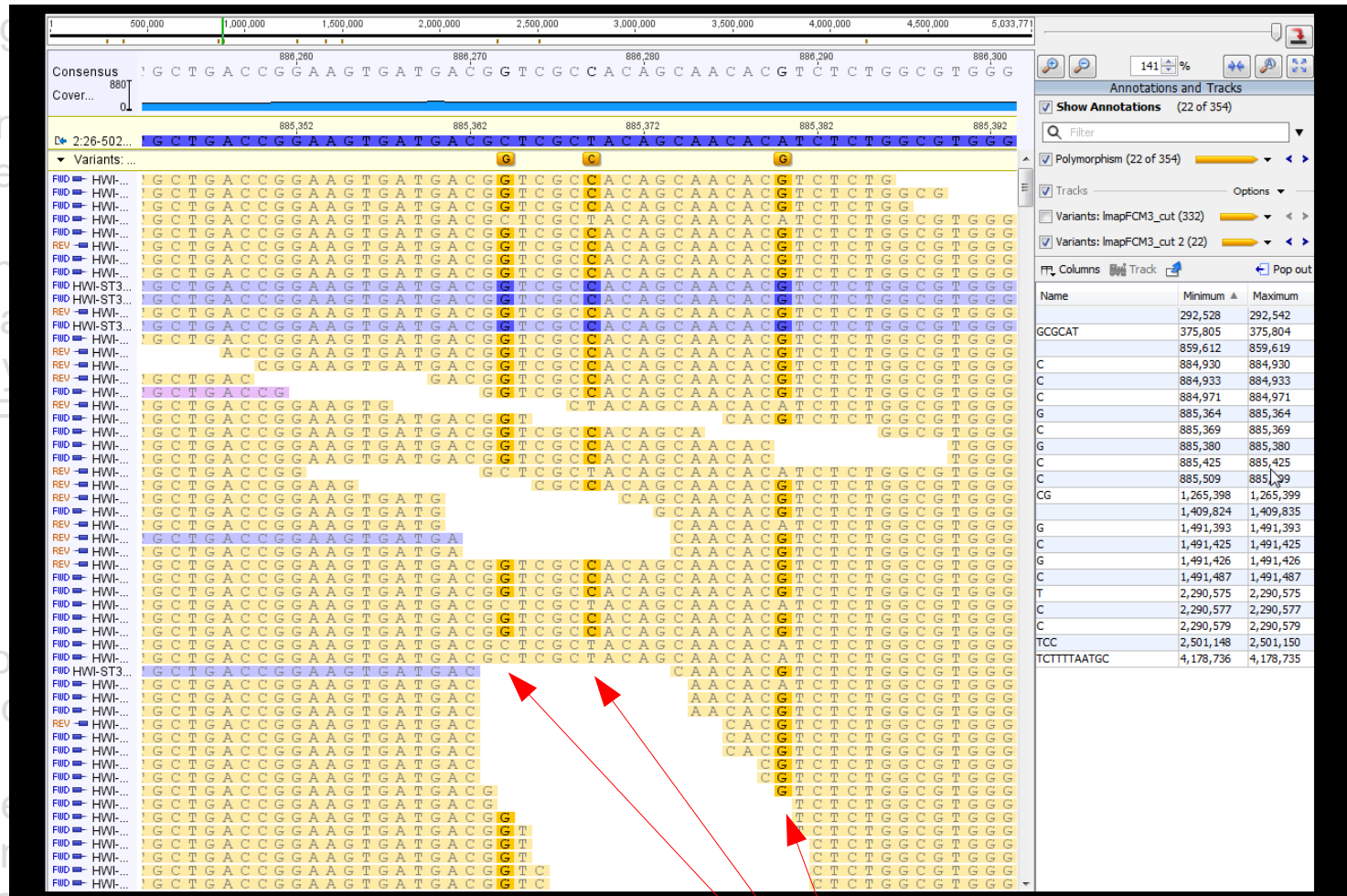
- Genome Sequencing
 - De novo sequencing
 - Resequencing
 - **Targeted (re)sequencing**
 - Mitochondrial sequencing
 - Mutation detection
 - Amplicon sequencing
 - Amplicon Cancer Panel
 - Phylogenomics- Phylogenetics
- Transcript Expression Profiling
 - RNA sequencing
 - miRNA sequencing
 - Deep-SAGE
 - Deep-CAGE
 - PAS: polyadenylation site
- Transcription factor binding : ChIP
- Structural variation
- Metagenomics / Metagenetics / Me
- Microsatellite sequencing
- Genetic marker discovery
 - Microsatellite development
 - RADSeq (Restriction-site-Asso
 - RRLs (Reduced-Representatio
 - GBS (Genotyping By Sequenci



• ...

Applications

- Genome Sequencing
 - De novo sequencing
 - Resequencing
 - Targeted (re)sequencing
 - Mitochondrial sequencing
 - **Mutation detection**
 - Amplicon sequencing
 - Amplicon Cancer Panel
 - Phylogenomics- Phylogenetic
- Transcript Expression Profiling
 - RNA sequencing
 - miRNA sequencing
 - Deep-SAGE
 - Deep-CAGE
 - PAS: polyadenylation site
- Transcription factor binding
- Structural variation
- Metagenomics / Metagenome
- Microsatellite sequencing
- Genetic marker discovery
 - Microsatellite development
 - RADSeq (Restriction-site-Associated Sequencing)
 - RRLs (Reduced-Representation libraries)
 - GBS (Genotyping By Sequencing)
 - ...

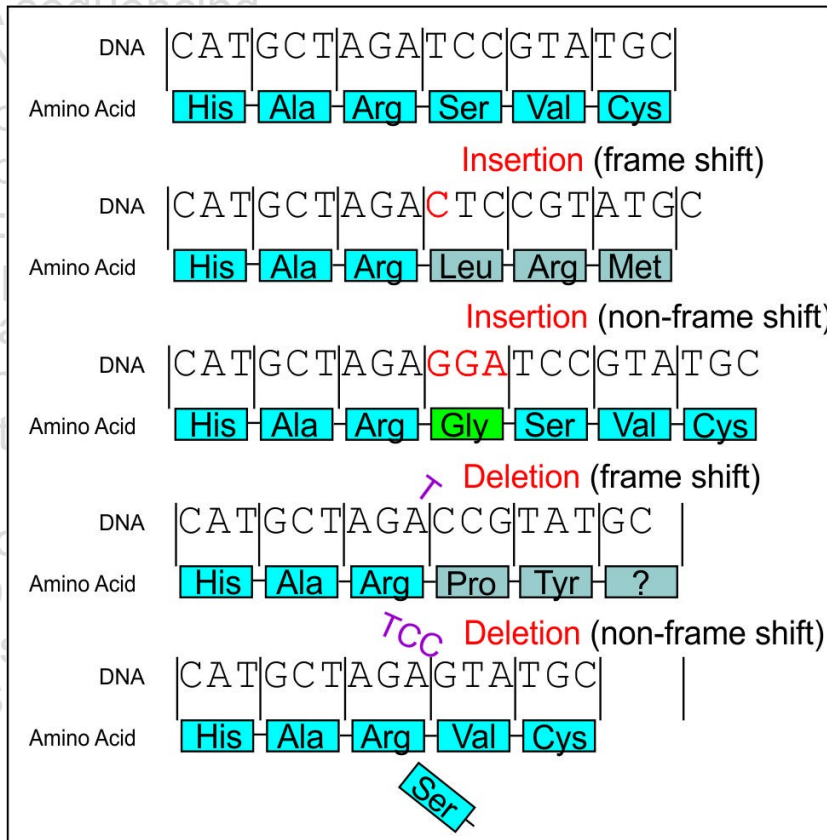
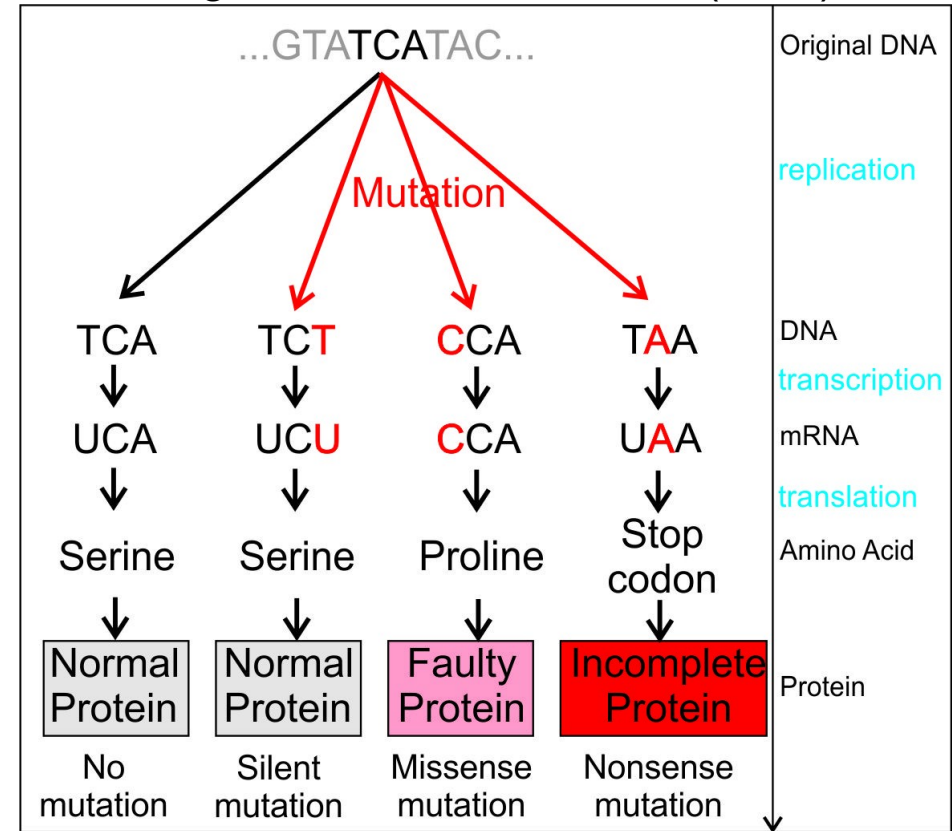


Not random errors, but 2 possible bases at one position.

Applications

- Genome Sequencing
 - De novo sequencing
 - Resequencing
 - Targeted (re)sequencing
 - Mitochondrial sequencing
 - **Mutation detection**
 - Amplicon sequencing
 - Amplicon Cancer Panel
 - Phylogenomics- Phylogenetics
- Transcript Expression Profiling

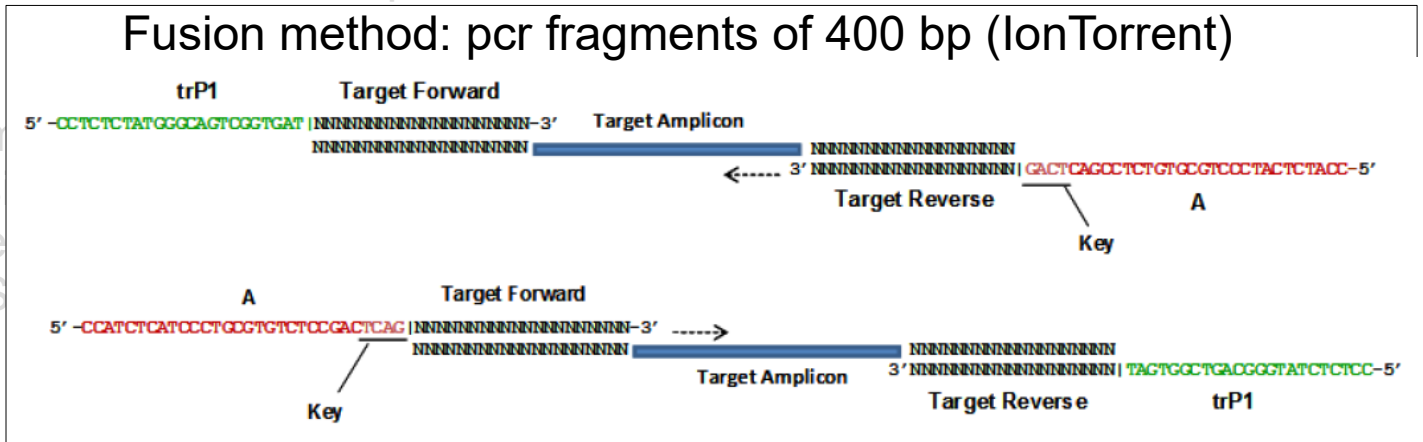
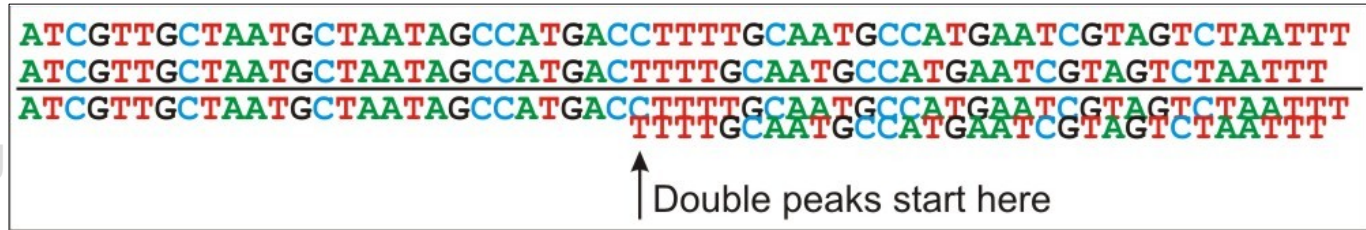
Single nucleotide variation (SNV)



Insertions and deletions

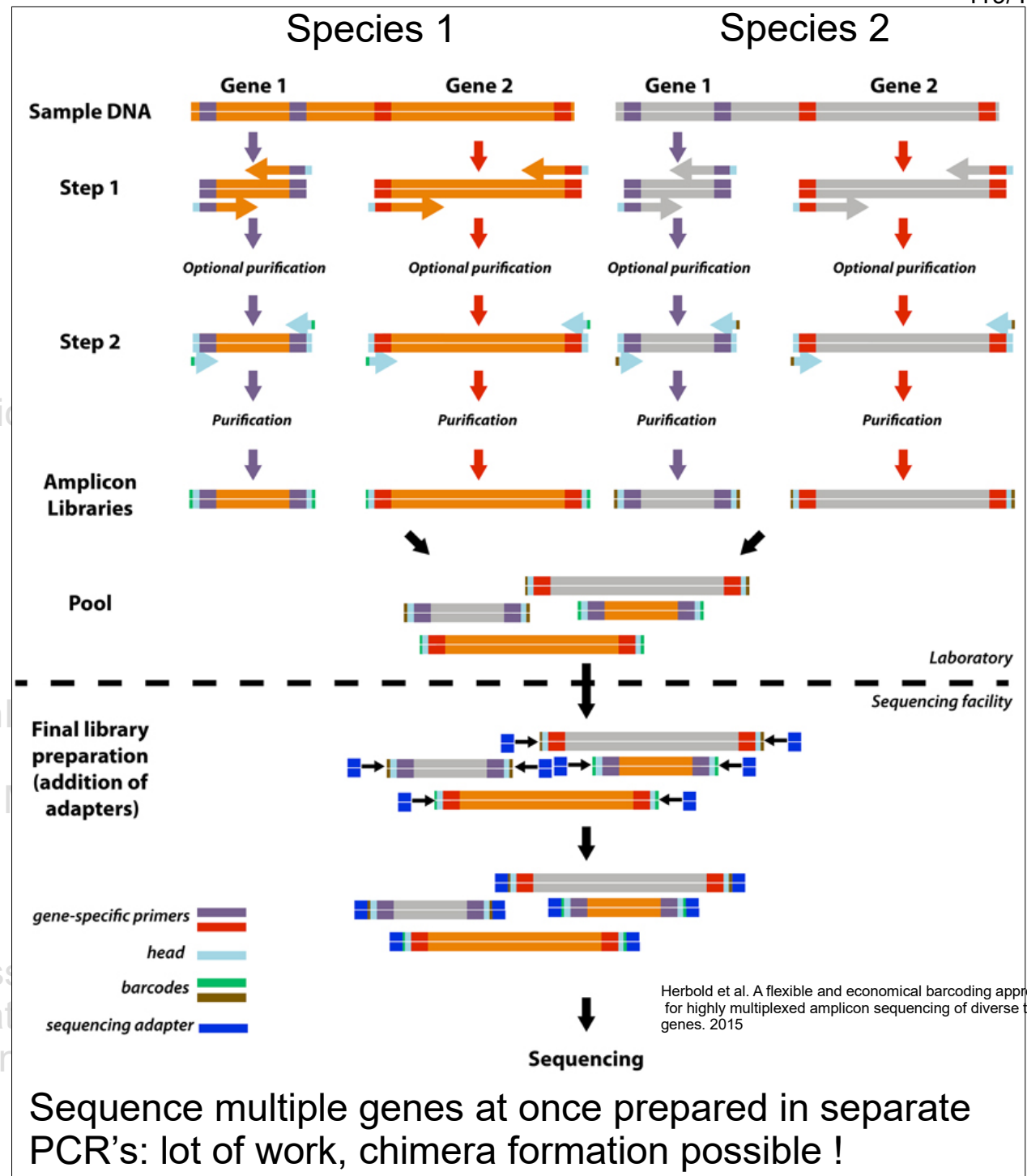
Applications

- Genome Sequencing
 - De novo sequencing
 - Resequencing
 - Targeted (re)sequencing
 - Mitochondrial sequencing
 - Mutation detection
 - Amplicon sequencing = PCR fragments**
 - Amplicon Cancer Panel
 - Phylogenomics- Phylogenetics
- Transcript Expression Profiling
 - RNA sequencing
 - miRNA sequencing
 - Deep-SAGE
 - Deep-CAGE
 - PAS: polyadenylation site
- Transcription factor binding
- Structural variation
- Metagenomics / Metagenetics / Metatranscriptomics
- Microsatellite sequencing
- Genetic marker discovery
 - Microsatellite development
 - RADSeq (Restriction-site associated DNA sequencing)
 - RRLs (Reduced-Representation Libraries)
 - GBS (Genotyping By Sequencing)
- ...



Applications

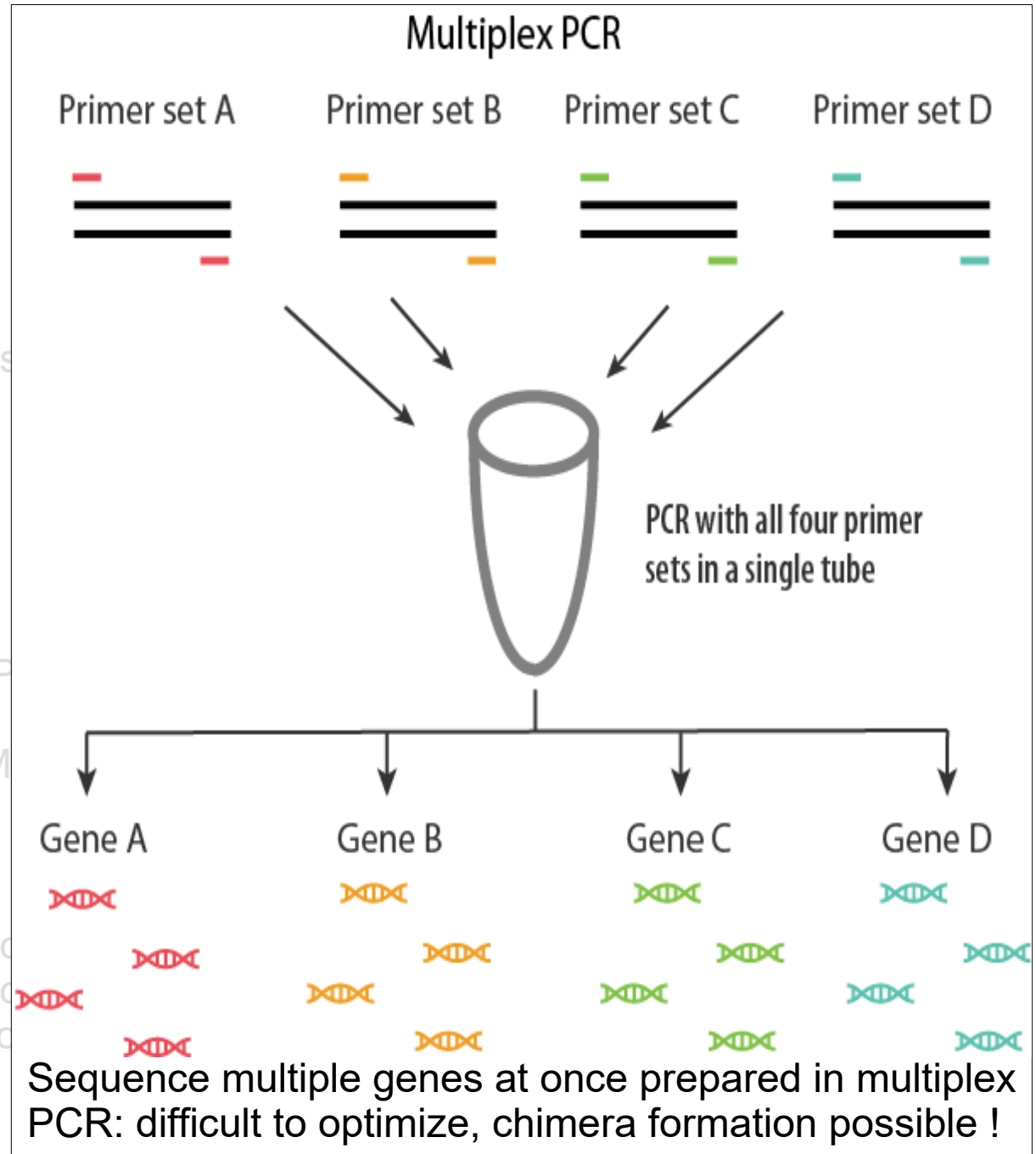
- Genome Sequencing
 - De novo sequencing
 - Resequencing
 - Targeted (re)sequencing
 - Mitochondrial sequencing
 - Mutation detection
 - **Amplicon sequencing**
 - Amplicon Cancer Panel
 - Phylogenomics- Phylogenetic
- Transcript Expression Profiling
 - RNA sequencing
 - miRNA sequencing
 - Deep-SAGE
 - Deep-CAGE
 - PAS: polyadenylation site
- Transcription factor binding : ChIP
- Structural variation
- Metagenomics / Metagenetics /
- Microsatellite sequencing
- Genetic marker discovery
 - Microsatellite development
 - RADSeq (Restriction-site-Association)
 - RRLs (Reduced-Representation)
 - GBS (Genotyping By Sequencing)
- ...



Sequence multiple genes at once prepared in separate PCR's: lot of work, chimera formation possible !

Applications

- Genome Sequencing
 - De novo sequencing
 - Resequencing
 - Targeted (re)sequencing
 - Mitochondrial sequencing
 - Mutation detection
 - **Amplicon sequencing**
 - Amplicon Cancer Panel
 - Phylogenomics- Phylogenetics
- Transcript Expression Profiling
 - RNA sequencing
 - miRNA sequencing
 - Deep-SAGE
 - Deep-CAGE
 - PAS: polyadenylation site
- Transcription factor binding : ChIP
- Structural variation
- Metagenomics / Metagenetics / M
- Microsatellite sequencing
- Genetic marker discovery
 - Microsatellite development
 - RADSeq (Restriction-site-Asso
 - RRLs (Reduced-Representatio
 - GBS (Genotyping By Sequenc
- ...



Applications

- Genome Sequencing
 - De novo sequencing
 - Resequencing
 - Targeted (re)sequencing
 - Mitochondrial sequencing
 - Mutation detection
 - Amplicon sequencing
 - **Amplicon Cancer Panel**
 - Phylogenomics- Phylog
- Transcript Expression Profile
 - RNA sequencing
 - miRNA sequencing
 - Deep-SAGE
 - Deep-CAGE
 - PAS: polyadenylation si
- Transcription factor binding
- Structural variation
- Metagenomics / Metagenetics / Metatranscriptomics
- Microsatellite
- Genetic markers
 - Microsat
 - RADSeq
 - RRLs (R
 - GBS (Ge
- ...

Ion AmpliSeq Comprehensive Cancer Panel		
Targets	Exons with >400 oncogenes and tumor suppressor genes	
Amplicon length	125–175 bp (average 155 bp)	
Primer pool size	~16,000 primers in 4 tubes	
Input DNA required	10 ng per pool, 40 ng per DNA sample	
Time to results	Single-day workflow from DNA to annotated variants (run time varies by read length and chip type)	
Sample multiplexing	Ion PI™ or Ion 540™ Chip: 4 samples, ~1,000x average coverage	
	Specification	Observed performance
Coverage uniformity*	>90%	94%
On-target bases**	>95%	97%
Average depth of coverage	NA	350x

* Coverage uniformity = bases covered at >20% of the mean coverage.
 ** On-target bases = bases mapped to target regions, out of total mapped bases per run.

Technical Note: DNA Sequencing

illumina®

Sequencing Panel for 4813 Genes with Known Associated Clinical Phenotypes

TruSight™ One Sequencing Panel provides high depth of coverage for accurate variant calling

Applications

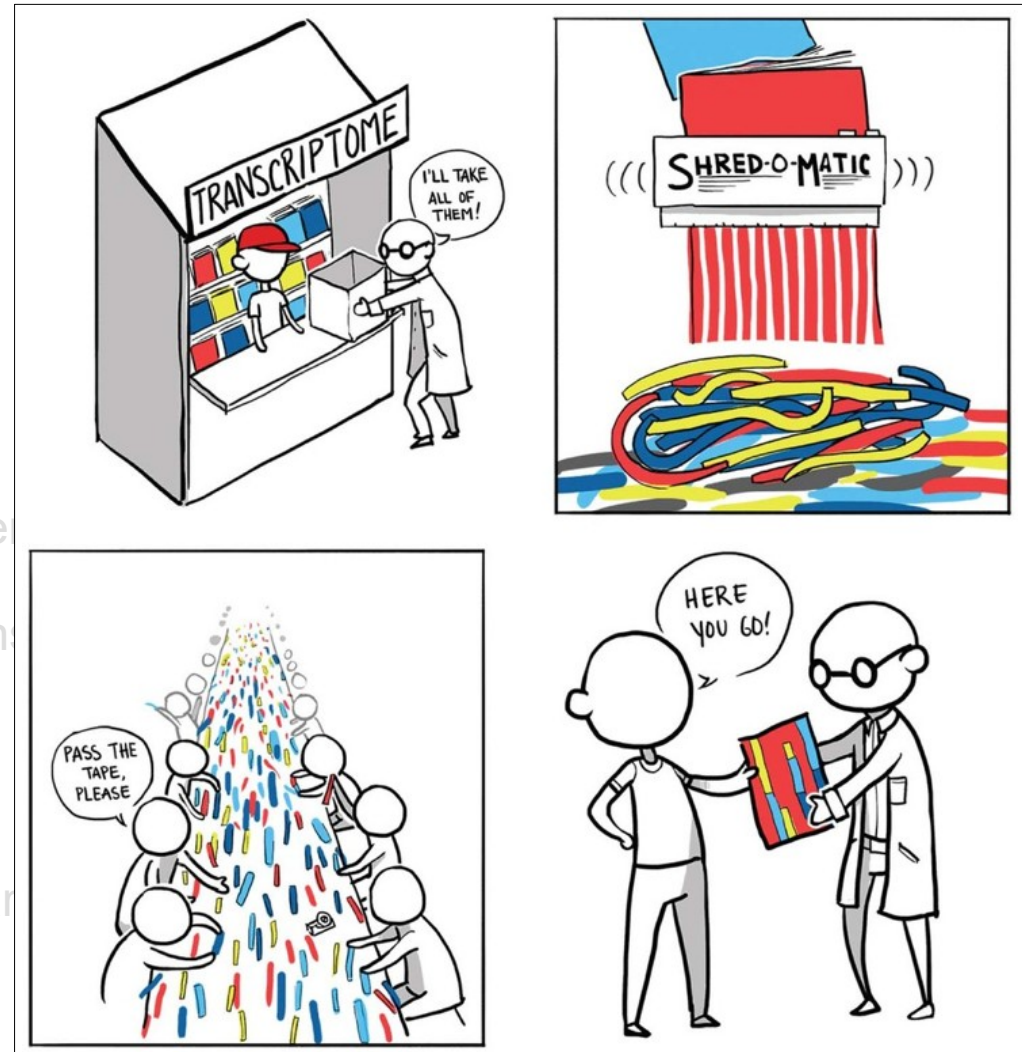
- Genome Sequencing
 - De novo sequencing
 - Resequencing
 - Targeted (re)sequencing
 - Mitochondrial sequencing
 - Mutation detection
 - Amplicon sequencing
 - Amplicon Cancer Panel
 - **Phylogenomics- Phylogenetics**
- Transcript Expression Profiling
 - RNA sequencing
 - miRNA sequencing
 - Deep-SAGE
 - Deep-CAGE
 - PAS: polyadenylation site
- Transcription factor binding : ChIP sequencing
- Structural variation
- Metagenomics / Metagenetics / Metatranscriptomics
- Microsatellite sequencing
- Genetic marker discovery
 - Microsatellite development
 - RADSeq (Restriction-site-Associated Sequencing)
 - RRLs (Reduced-Representation libraries)
 - GBS (Genotyping By Sequencing)
- ...

Phylogenomics draws information by comparing entire genomes, or at least large portions of genomes.

Phylogenetics compares and analyzes the sequences of single genes, or a small number of genes, as well as many other types of data

Applications

- Genome Sequencing
 - De novo sequencing
 - Resequencing
 - Targeted (re)sequencing
 - Mitochondrial sequencing
 - Mutation detection
 - Amplicon sequencing
 - Amplicon Cancer Panel
 - Phylogenomics- Phylogenetics
 - Transcript Expression Profiling
 - **RNA sequencing**
 - miRNA sequencing
 - Deep-SAGE
 - Deep-CAGE
 - PAS: polyadenylation site
 - Transcription factor binding : ChIP seque
 - Structural variation
 - Metagenomics / Metagenetics / Metatran
 - Microsatellite sequencing
 - Genetic marker discovery
 - Microsatellite development
 - RADSeq (Restriction-site-Associated
 - RRLs (Reduced-Representation librar
 - GBS (Genotyping By Sequencing)
 - ...
- * Different expression profiles of genes in different conditions (avoid rRNA)
- * Determine the genes in the transcriptome (also lncRNA)



Applications

- Genome Sequencing
 - De novo sequencing
 - Resequencing
 - Targeted (amplicon) sequencing
 - Mitochondrial DNA sequencing
 - Mutation detection
 - Amplicon sequencing
 - Amplicon Cancer Panel
 - Phylogenomics- Phylogenetics
- Transcript Expression Profiling
 - RNA sequencing
 - miRNA sequencing
 - Deep-SEA
 - Deep-CA
 - PAS: poly(A) sequencing
- Transcription start site (TSS) analysis
- Structural variant detection
- Metagenomics
- Microsatellite genotyping
- Genetic marker genotyping
 - Microsatellite genotyping
 - RADSeq
 - RRLs (Reduced-Representation libraries)
 - GBS (Genotyping By Sequencing)
- ...

Prepare Library | Sequence | Analyze Data



TruSight[®] RNA Pan-Cancer Panel

Comprehensive assessment of cancer-related RNA transcripts and fusion detection in FFPE tissues and other oncology samples.



Ion AmpliSeq Colon and Lung Cancer Research Panel v2 and Ion AmpliSeq RNA Fusion Lung Cancer Research Panel

Application	Somatic mutation detection
Genes	<i>KRAS, EGFR, BRAF, PIK3CA, AKT1, ERBB2, PTEN, NRAS, STK11, MAP2K1, ALK, DDR2, CTNNB1, MET, TP53, SMAD4, FBX7, FGFR3, NOTCH1, ERBB4, FGFR1, and FGFR2</i>

Applications

- Genome Sequencing
 - De novo sequencing
 - Resequencing
 - Targeted (re)sequencing
 - Mitochondrial sequencing
 - Mutation detection
 - Amplicon sequencing
 - Amplicon Cancer Panel
 - Phylogenomics- Phylogenetics
- Transcript Expression Profiling
 - **RNA sequencing**
 - miRNA sequencing
 - Deep-SAGE
 - Deep-CAGE
 - PAS: polyadenylation site
- Transcription factor binding : ChIP-seq
- Structural variation
- Metagenomics / Metagenetics /
- Microsatellite sequencing
- Genetic marker discovery
 - Microsatellite development
 - RADSeq (Restriction-site-Associated DNA Sequencing)
 - RPLs (Reduced Representation Libraries)

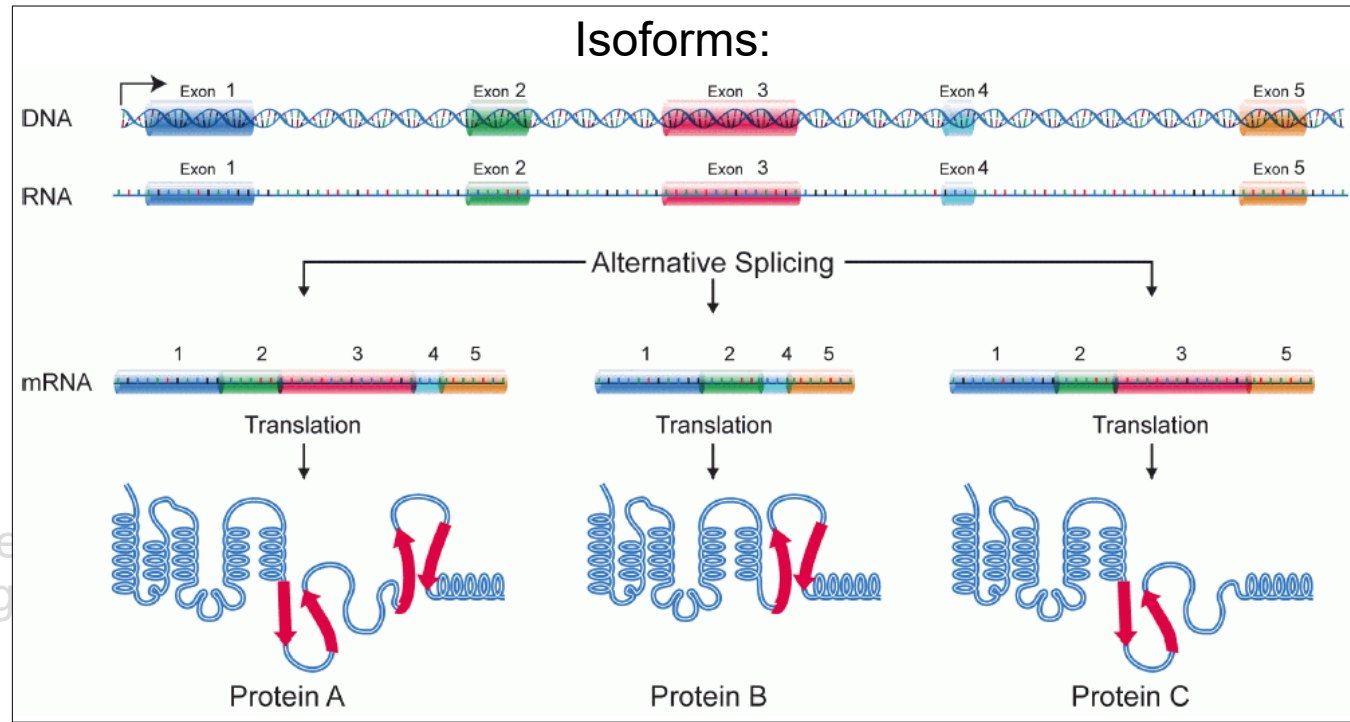


Table 2. Performance of Illumina, PacBio and ONT on isoform identification in the gold standard SIRVs.

Strategy+Library	True positive <u>68</u>		False positive	
	Over-annotated library (100)			
	Correct library (68)			
	Insufficient library (43)			
Illumina+Insufficient	39	5	-	33
Illumina+Correct	<u>63</u>	-	-	<u>27</u>
Illumina+Over-annotated	62	15	-	24
PacBio+Correct	<u>67</u>	-	-	-
ONT+Correct	<u>68</u>	-	-	-

RESEARCH ARTICLE

Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis [version 1; referees: awaiting peer review]

Jason L Weirather^{1*}, Mariateresa de Cesare^{2*}, Yunhao Wang^{1,3,4*}, Paolo Piazza², Vittorio Sebastiano^{5,6}, Xiu-Jie Wang³, David Buck², Kin Fai Au ^{1,7}

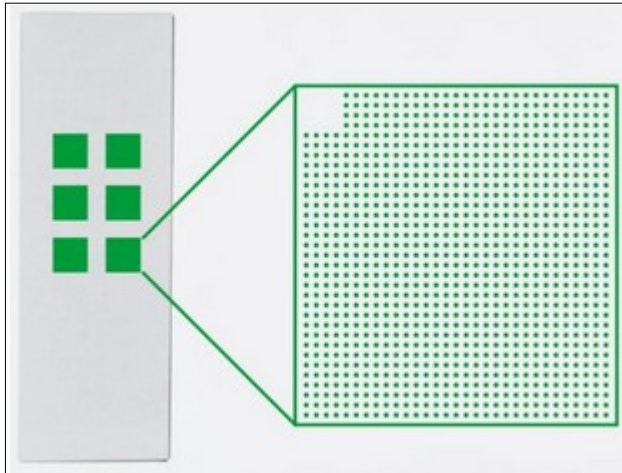
Applications

- Genome Sequencing

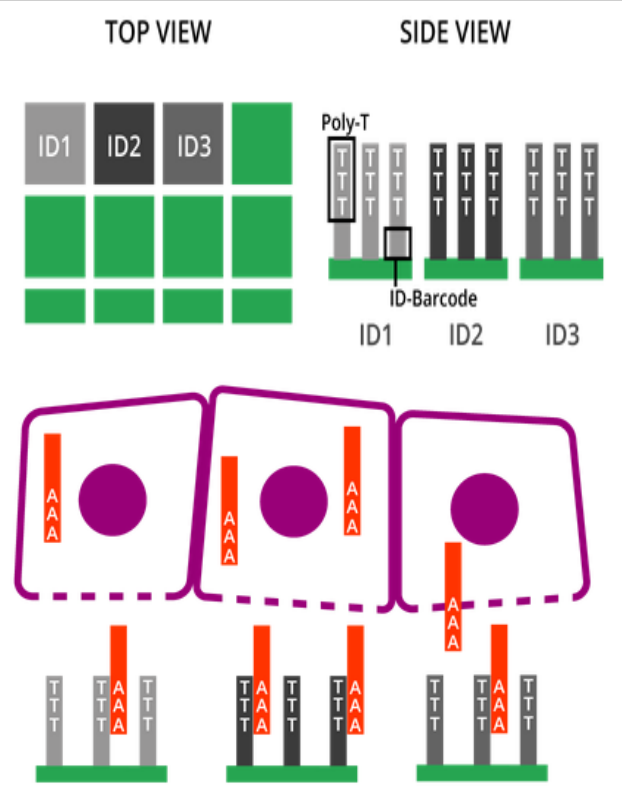
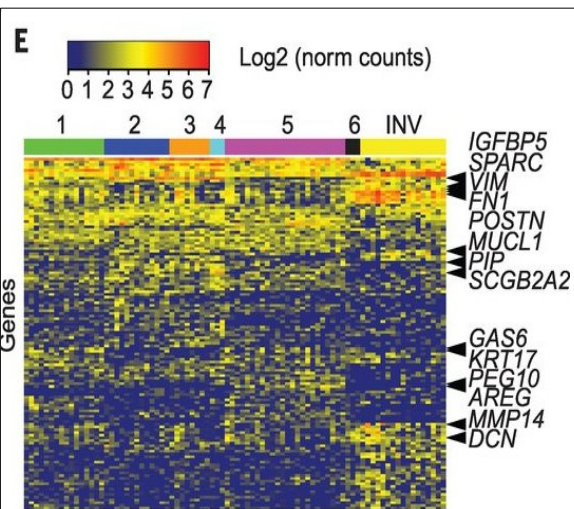
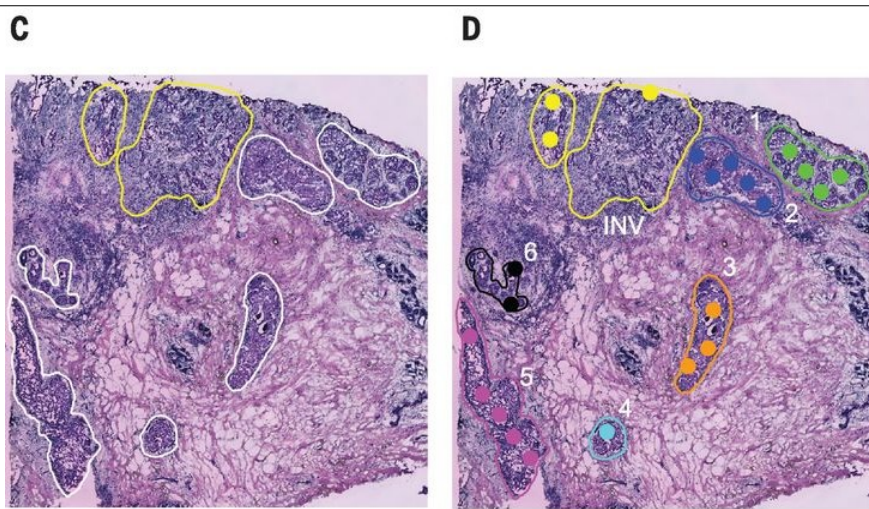
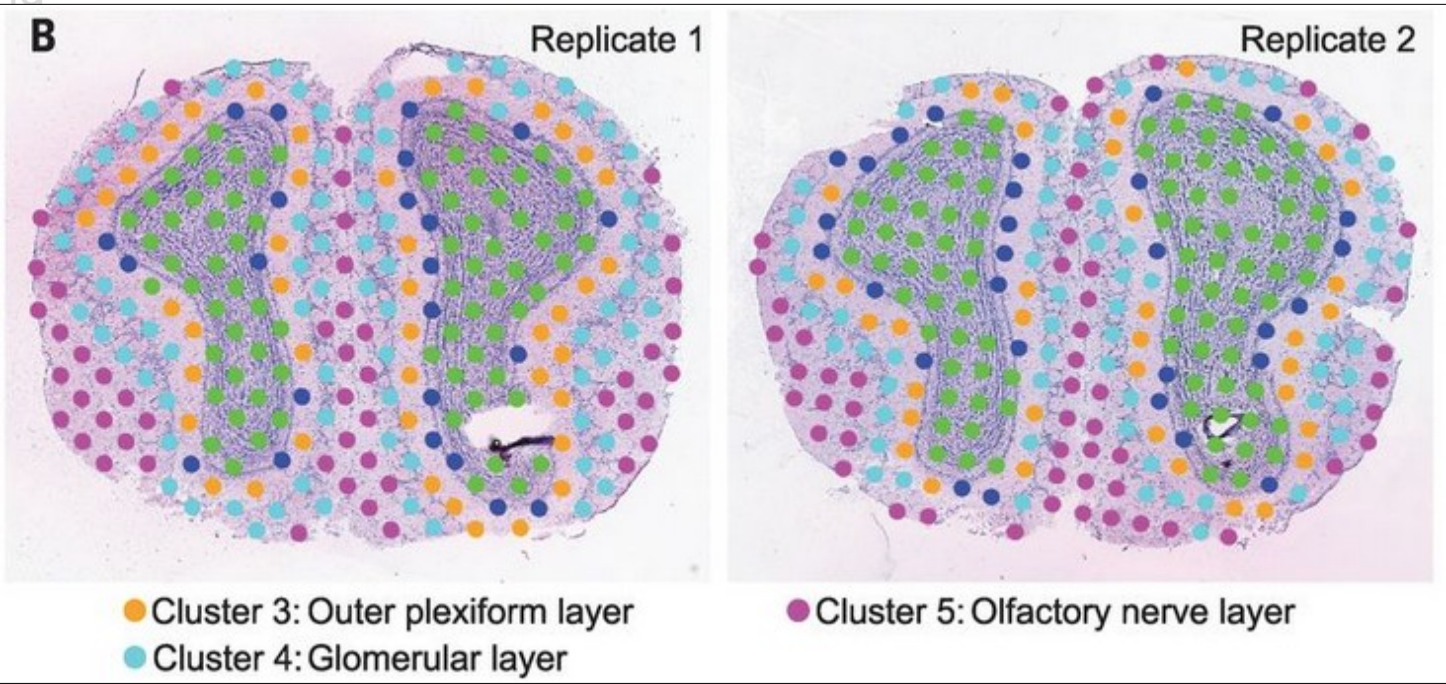
2D transcriptomics: spatialtranscriptomics.com

Visualization and analysis of gene expression in tissue sections by spatial transcriptomics

Patrik L. Ståhl et al. Science 1 July 2016

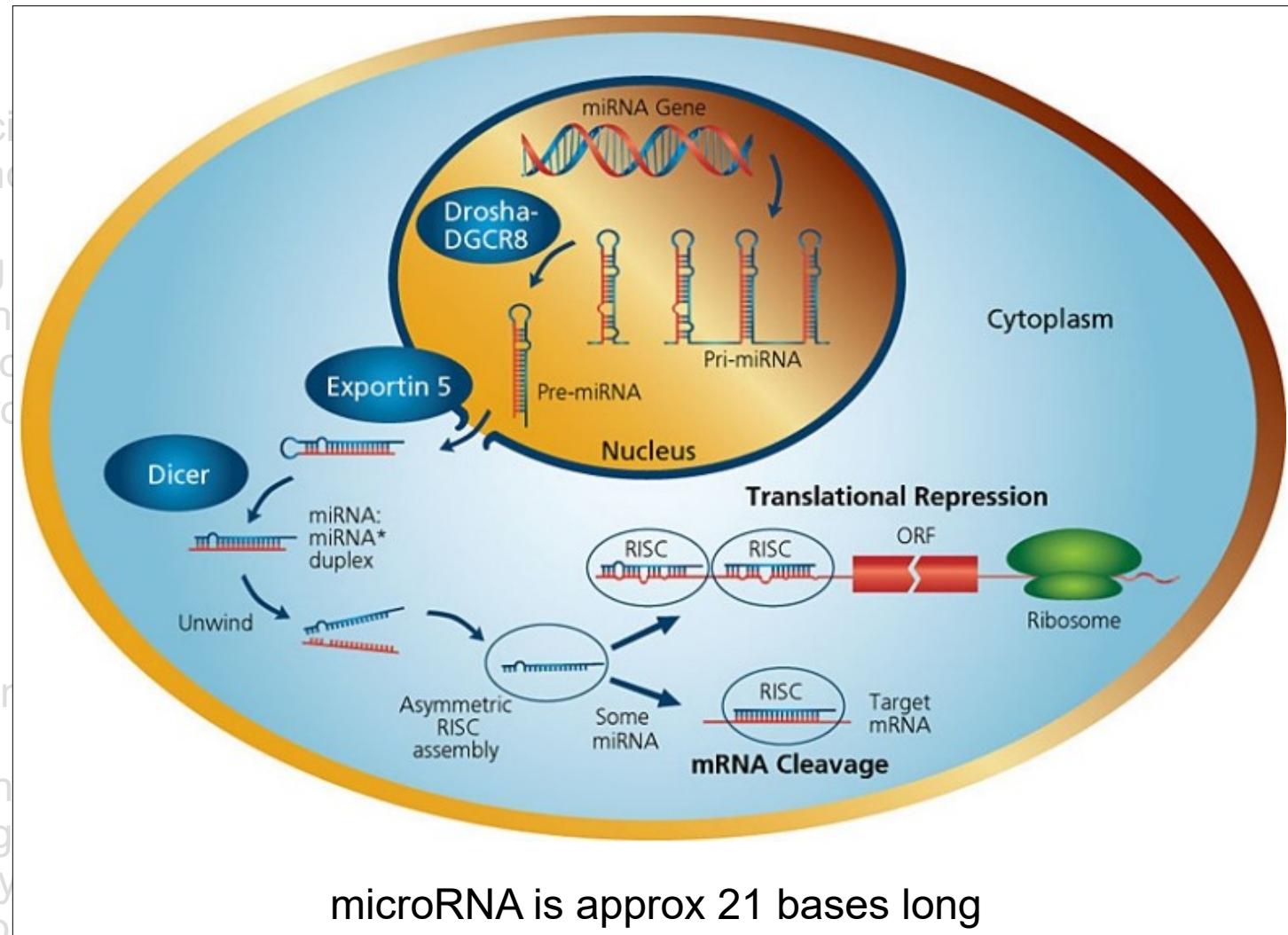


- RNA sequencing



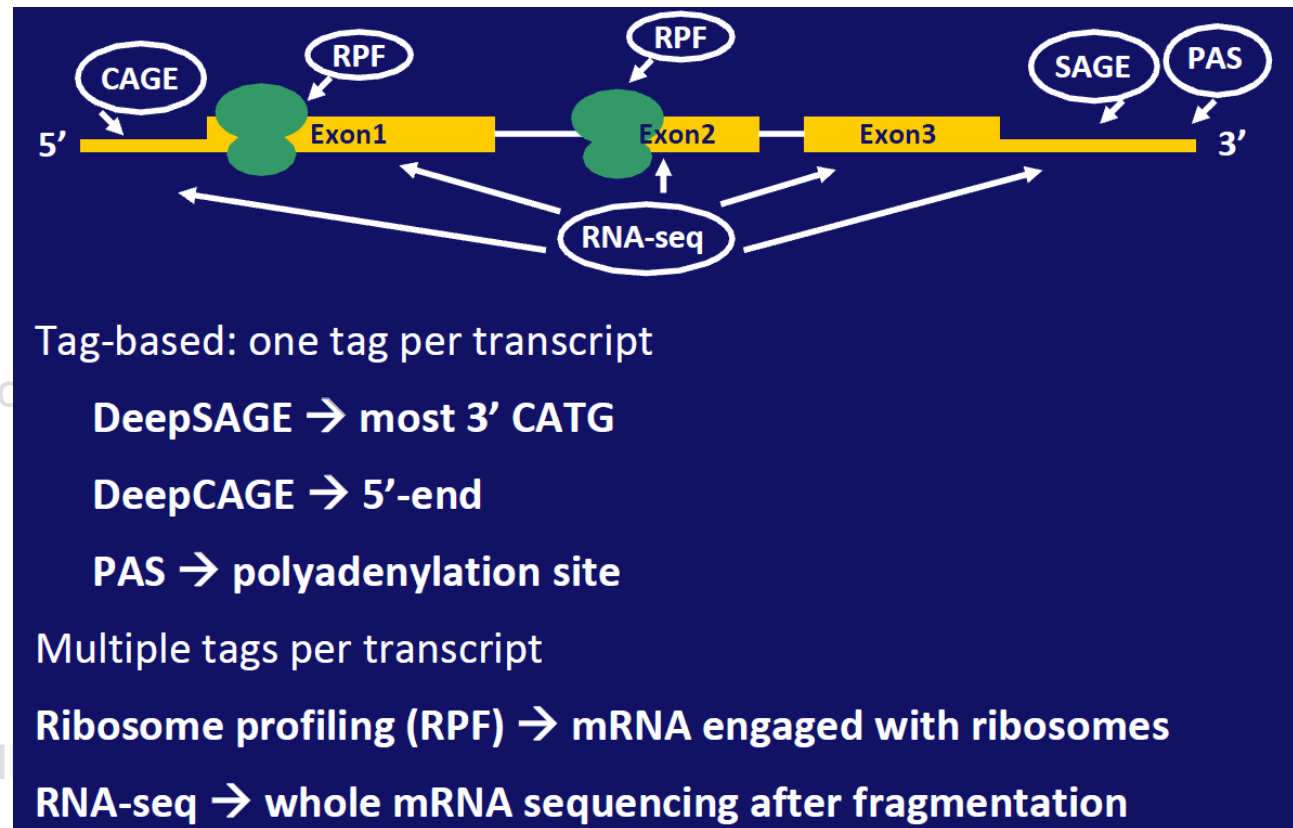
Applications

- Genome Sequencing
 - De novo sequencing
 - Resequencing
 - Targeted (re)sequencing
 - Mitochondrial sequencing
 - Mutation detection
 - Amplicon sequencing
 - Amplicon Cancer Panel
 - Phylogenomics- Phylogenetics
- Transcript Expression Profiling
 - RNA sequencing
 - **miRNA sequencing**
 - Deep-SAGE
 - Deep-CAGE
 - PAS: polyadenylation
- Transcription factor binding
- Structural variation
- Metagenomics / Metagenetics
- Microsatellite sequencing
- Genetic marker discovery
 - Microsatellite development
 - RADSeq (Restriction-site-Associated Sequencing)
 - RRLs (Reduced-Representation libraries)
 - GBS (Genotyping By Sequencing)
- ...



Applications

- Genome Sequencing
 - De novo sequencing
 - Resequencing
 - Targeted (re)sequencing
 - Mitochondrial sequencing
 - Mutation detection
 - Amplicon sequencing
 - Amplicon Cancer Panel
 - Phylogenomics- Phylogenetic
- Transcript Expression Profiling
 - RNA sequencing
 - miRNA sequencing
 - **Deep-SAGE**
 - **Deep-CAGE**
 - **PAS: polyadenylation site**
- Transcription factor binding : ChIP
- Structural variation
- Metagenomics / Metagenetics / Metatranscriptomics
- Microsatellite sequencing
- Genetic marker discovery
 - Microsatellite development
 - RADSeq (Restriction-site-Associated Sequencing)
 - RRLs (Reduced-Representation libraries)
 - GBS (Genotyping By Sequencing)
- ...

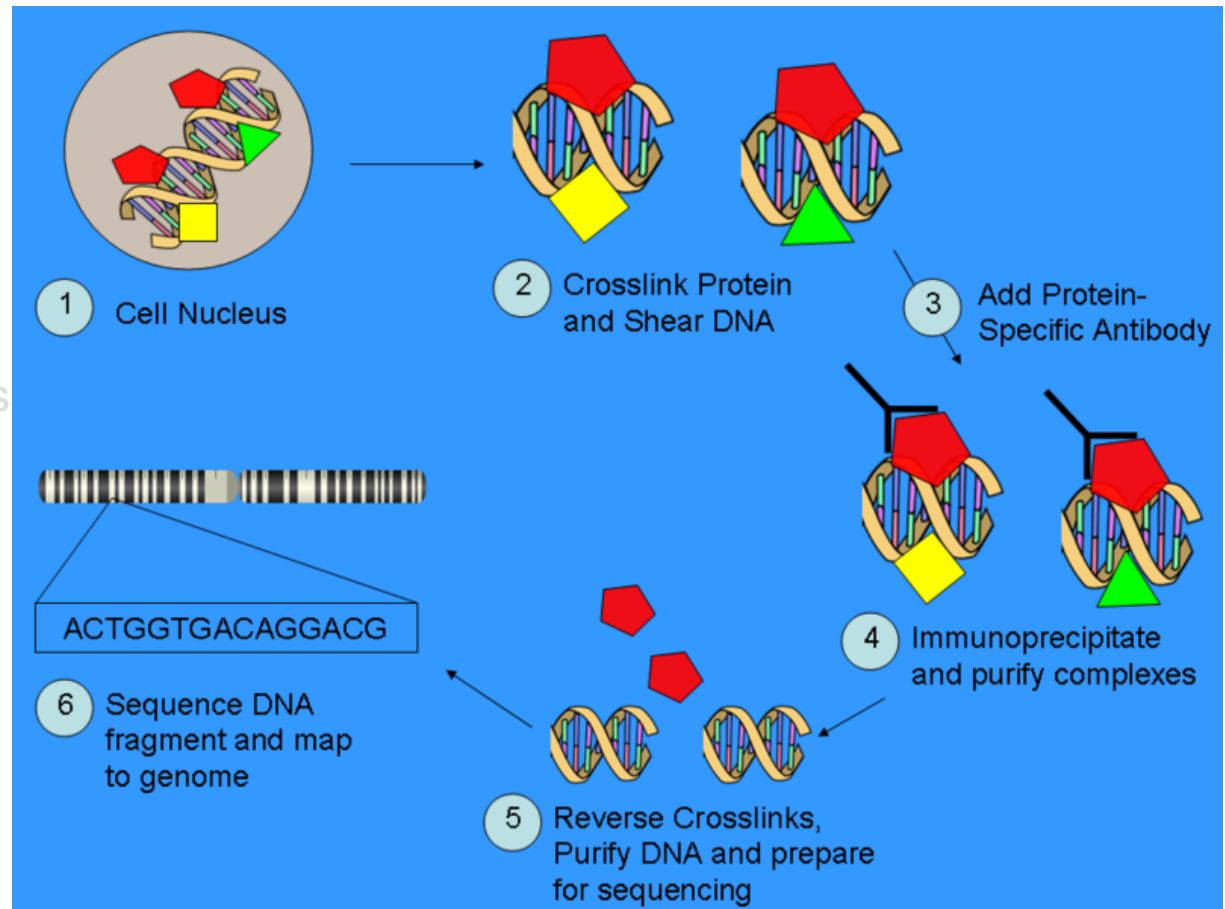


Sequence 17 bp of the transcription start or stop site
 -> discovery of new start and stop sites,
 antisense transcription possibilities

Applications

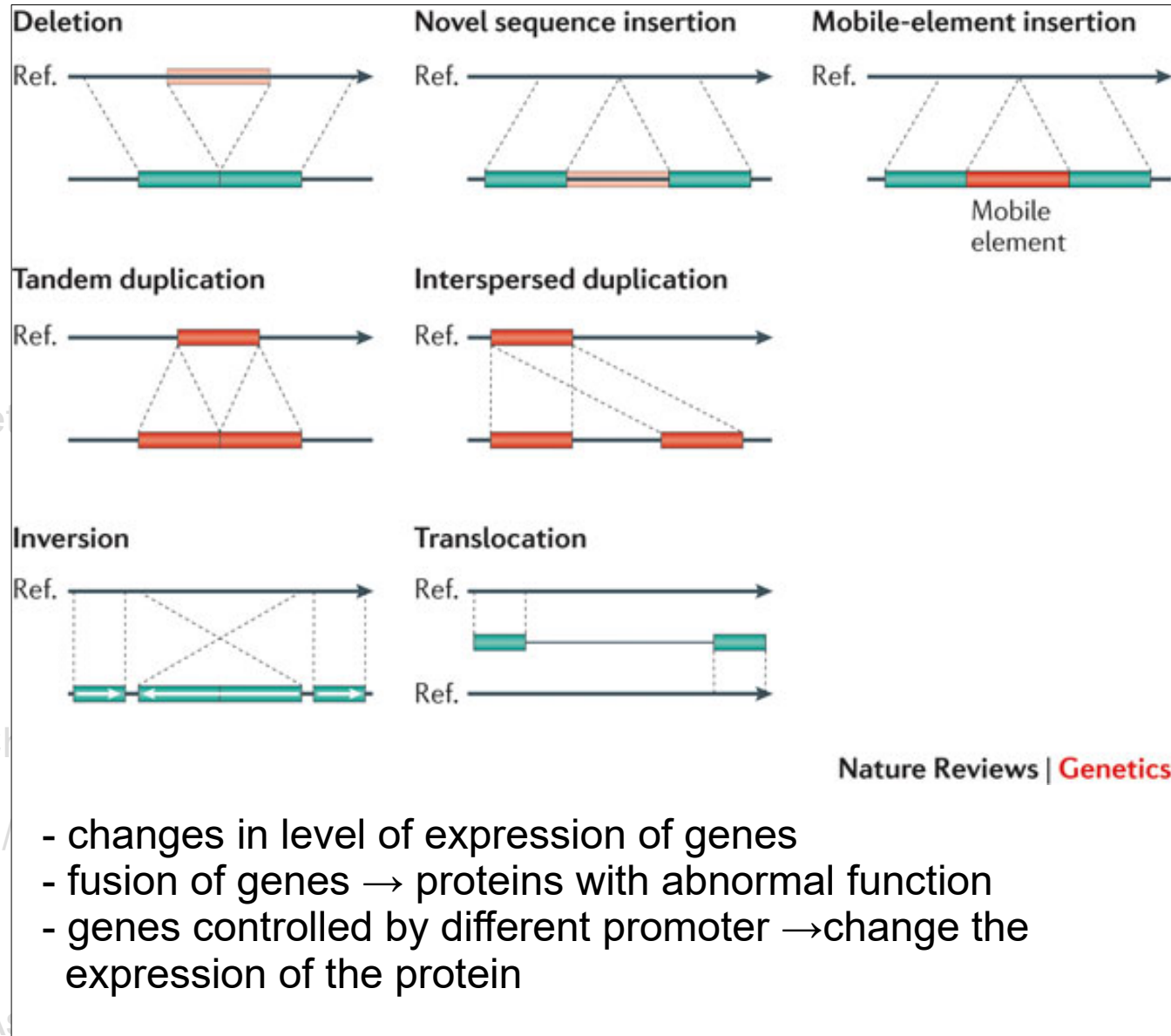
- Genome Sequencing
 - De novo sequencing
 - Resequencing
 - Targeted (re)sequencing
 - Mitochondrial sequencing
 - Mutation detection
 - Amplicon sequencing
 - Amplicon Cancer Panel
 - Phylogenomics- Phylogenetics
- Transcript Expression Profiling
 - RNA sequencing
 - miRNA sequencing
 - Deep-SAGE
 - Deep-CAGE
 - PAS: polyadenylation site
- **Transcription factor binding :**
 - **ChIP sequencing**
- Structural variation
- Metagenomics / Metagenetics / Metatranscriptomics
- Microsatellite sequencing
- Genetic marker discovery
 - Microsatellite development
 - RADSeq (Restriction-site-Associated Sequencing)
 - RRLs (Reduced-Representation libraries)
 - GBS (Genotyping By Sequencing)
- ...

Protein-DNA interactions



Applications

- Genome Sequencing
 - De novo sequencing
 - Resequencing
 - Targeted (re)sequencing
 - Mitochondrial sequencing
 - Mutation detection
 - Amplicon sequencing
 - Amplicon Cancer Panel
 - Phylogenomics- Phylogenetics
- Transcript Expression Profiling
 - RNA sequencing
 - miRNA sequencing
 - Deep-SAGE
 - Deep-CAGE
 - PAS: polyadenylation site
- Transcription factor binding : ChIP-seq
- **Structural variation**
- Metagenomics / Metagenetics / Metatranscriptomics
- Microsatellite sequencing
- Genetic marker discovery
 - Microsatellite development
 - RADSeq (Restriction-site-Associated DNA Sequencing)
 - RRLs (Reduced-Representation libraries)
 - GBS (Genotyping By Sequencing)
- ...

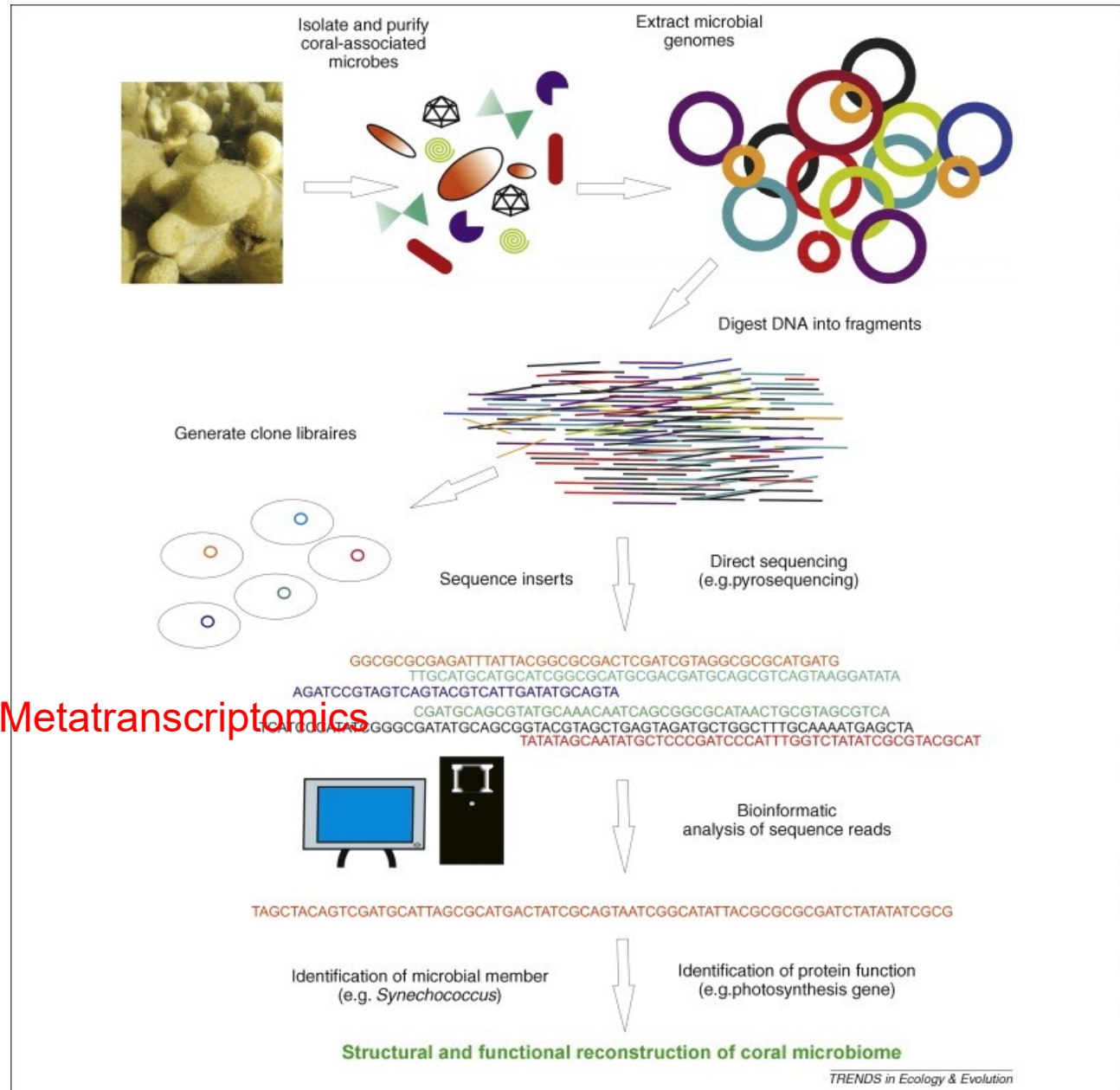


- changes in level of expression of genes
- fusion of genes → proteins with abnormal function
- genes controlled by different promoter → change the expression of the protein

Applications

- Genome Sequencing
 - De novo sequencing
 - Resequencing
 - Targeted (re)sequencing
 - Mitochondrial sequencing
 - Mutation detection
 - Amplicon sequencing
 - Amplicon Cancer Panel
 - Phylogenomics- Phylogeneti
- Transcript Expression Profiling
 - RNA sequencing
 - miRNA sequencing
 - Deep-SAGE
 - Deep-CAGE
 - PAS: polyadenylation site
- Transcription factor binding
- Structural variation
- **Metagenomics / Metagenetics / Metatranscriptomics**
- Microsatellite sequencing
- Genetic marker discovery
 - Microsatellite development
 - RADSeq
 - RRLs
 - GBS
- ...

Study of genomes recovered from environmental samples

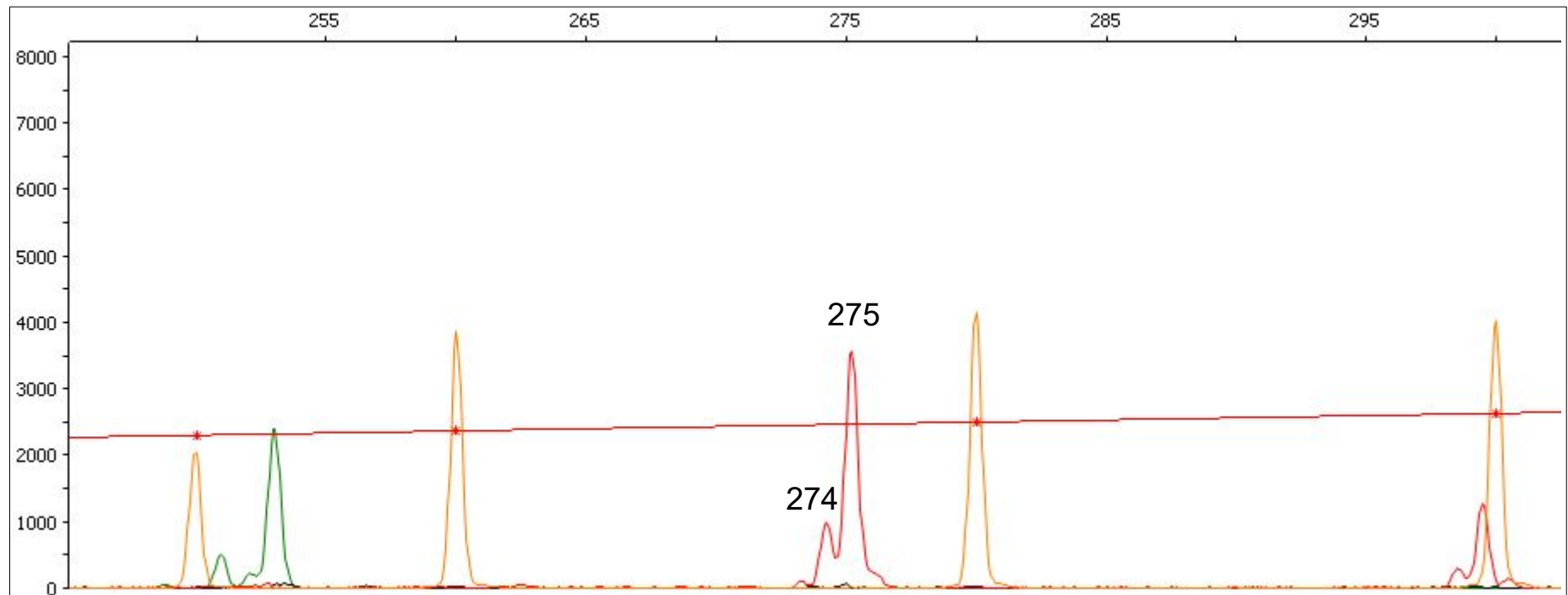


Applications

Microsatellite sequencing

Based on sizing on capillary sequencer
→ one number as result

Small peaks before or after big one ???



Applications

• Genome Sequencing

- De novo sequencing
- Resequencing
- Targeted (re)sequencing
- Mitochondrial sequencing
- Mutation detection
- Amplicon sequencing
- Amplicon Cancer
- Phylogenomics

• Transcript Expression

- RNA sequencing
- miRNA sequencing
- Deep-SAGE
- Deep-CAGE
- PAS: polyadenylation

• Transcription factor

• Structural variation

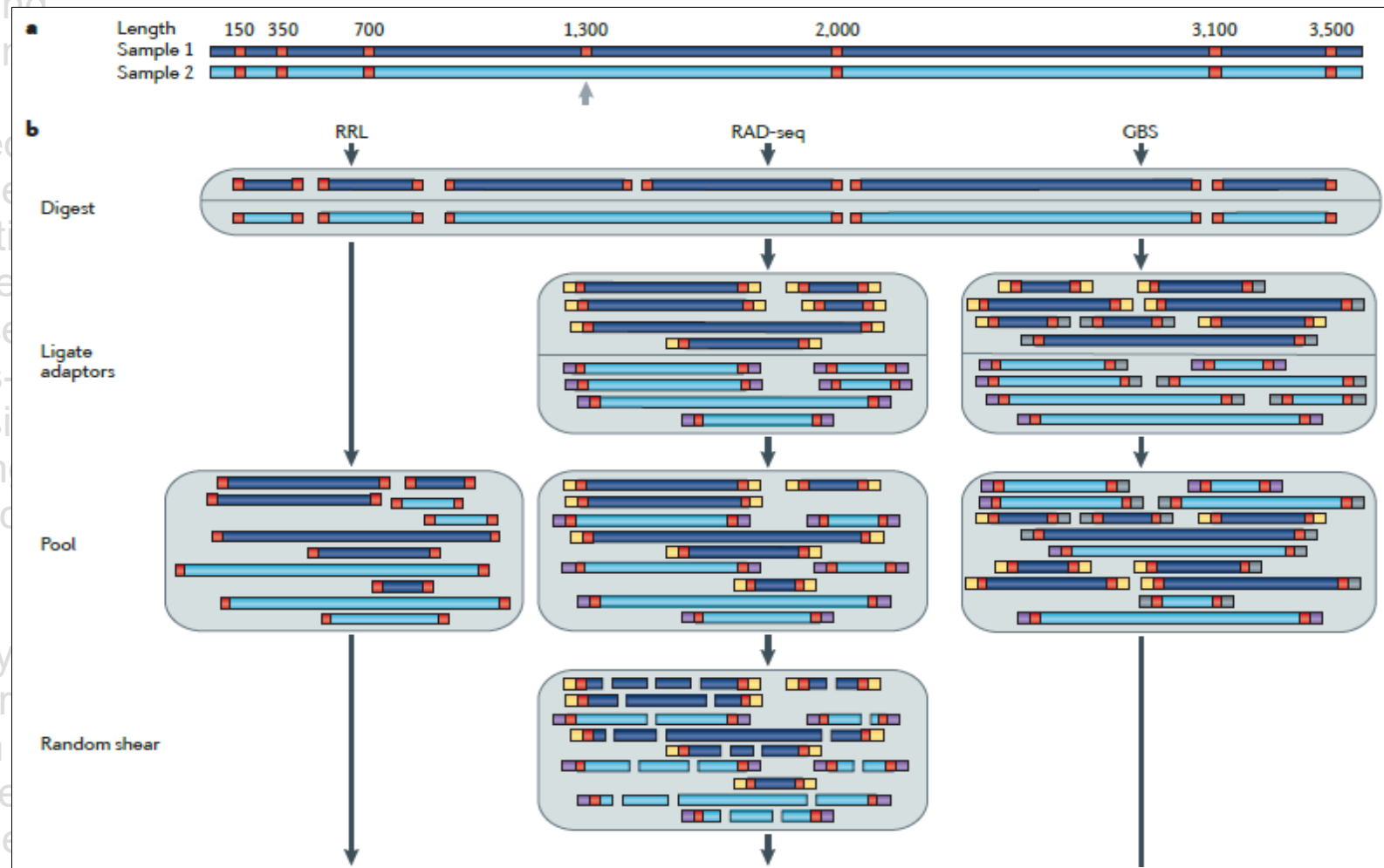
• Metagenomics / Met

• Microsatellite sequen

• Genetic marker discovery

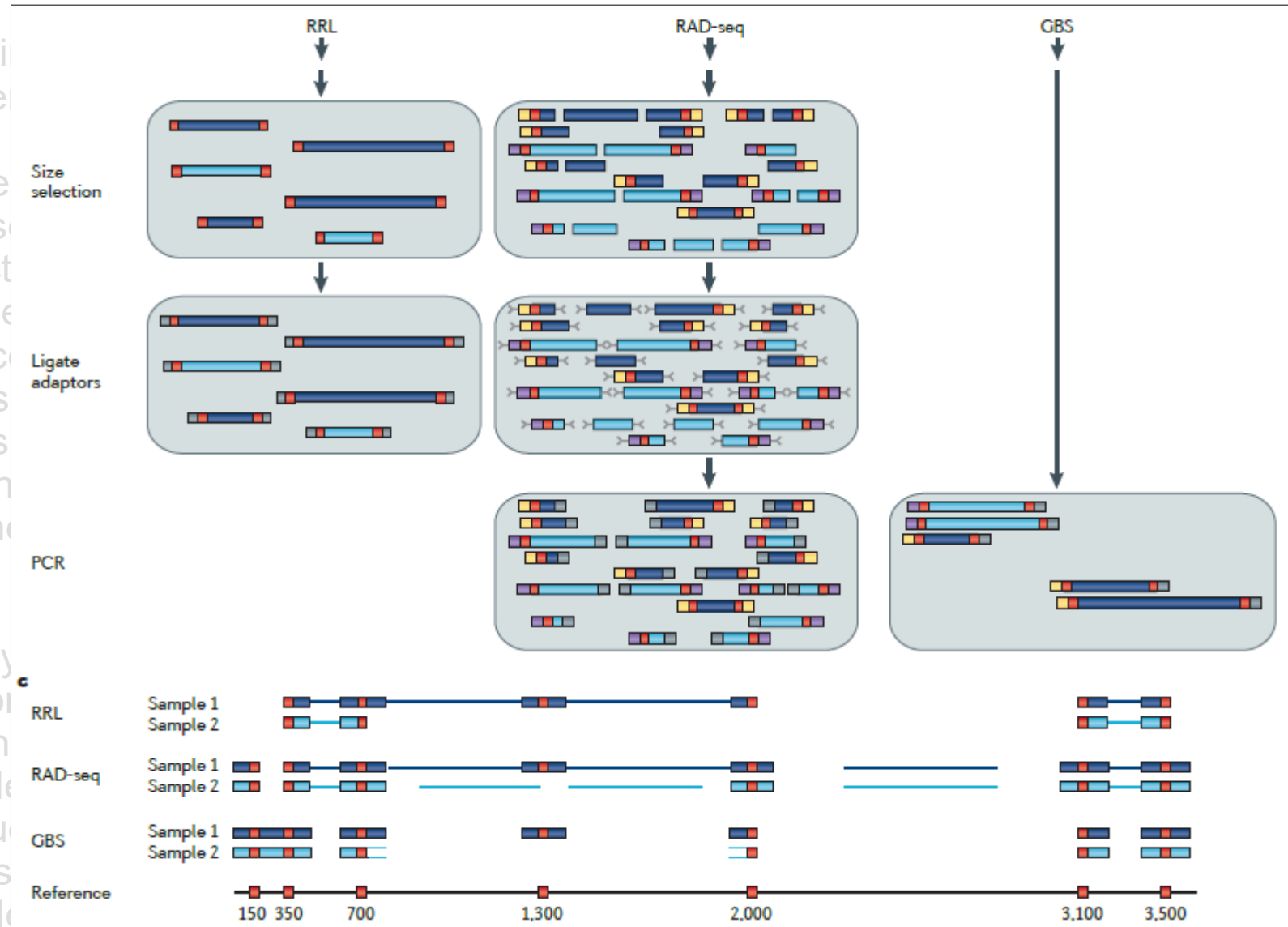
- Microsatellite development
- **RADSeq (Restriction-site-Associated Sequencing)**
- **RRLs (Reduced-Representation libraries)**
- **GBS (Genotyping By Sequencing)**

• ...



Applications

- Genome Sequencing
 - De novo sequencing
 - Resequencing
 - Targeted (re)sequencing
 - Mitochondrial sequencing
 - Mutation detection
 - Amplicon sequencing
 - Amplicon Cancer
 - Phylogenomics
- Transcript Expression
 - RNA sequencing
 - miRNA sequencing
 - Deep-SAGE
 - Deep-CAGE
 - PAS: polyadenylation
- Transcription factor
- Structural variation
- Metagenomics / Microbiome
- Microsatellite sequencing
- Genetic marker discovery
 - Microsatellite discovery
- ...



- **RADSeq (Restriction-site-Associated Sequencing)**
- **RRLs (Reduced-Representation libraries)**
- **GBS (Genotyping By Sequencing)**

Run types

- Single-end sequencing
- Paired-end sequencing
- Mate-pair sequencing
- Barcoding samples

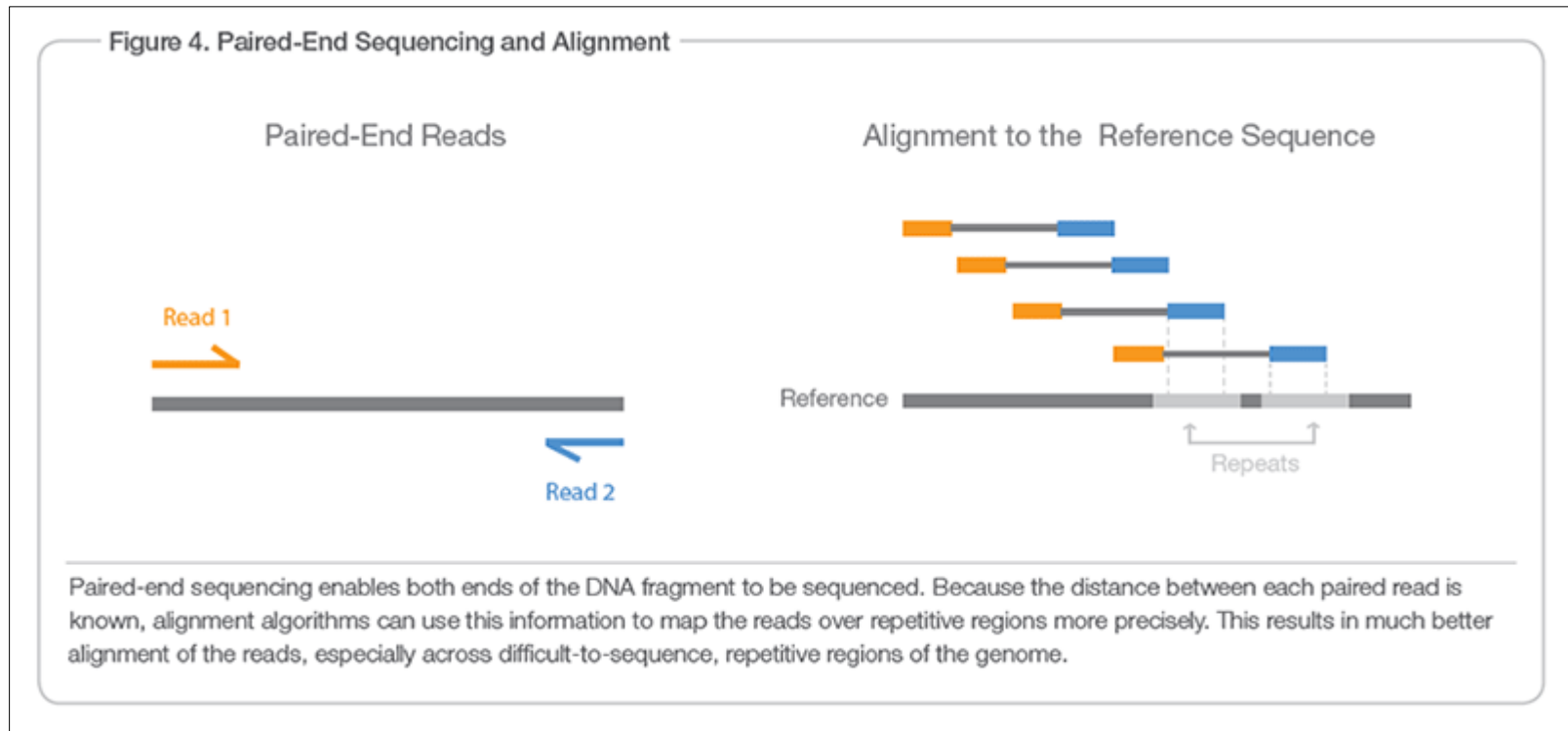


OR



Run types

- Single-end sequencing
- Paired-end sequencing
- Mate-pair sequencing
- Barcoding samples

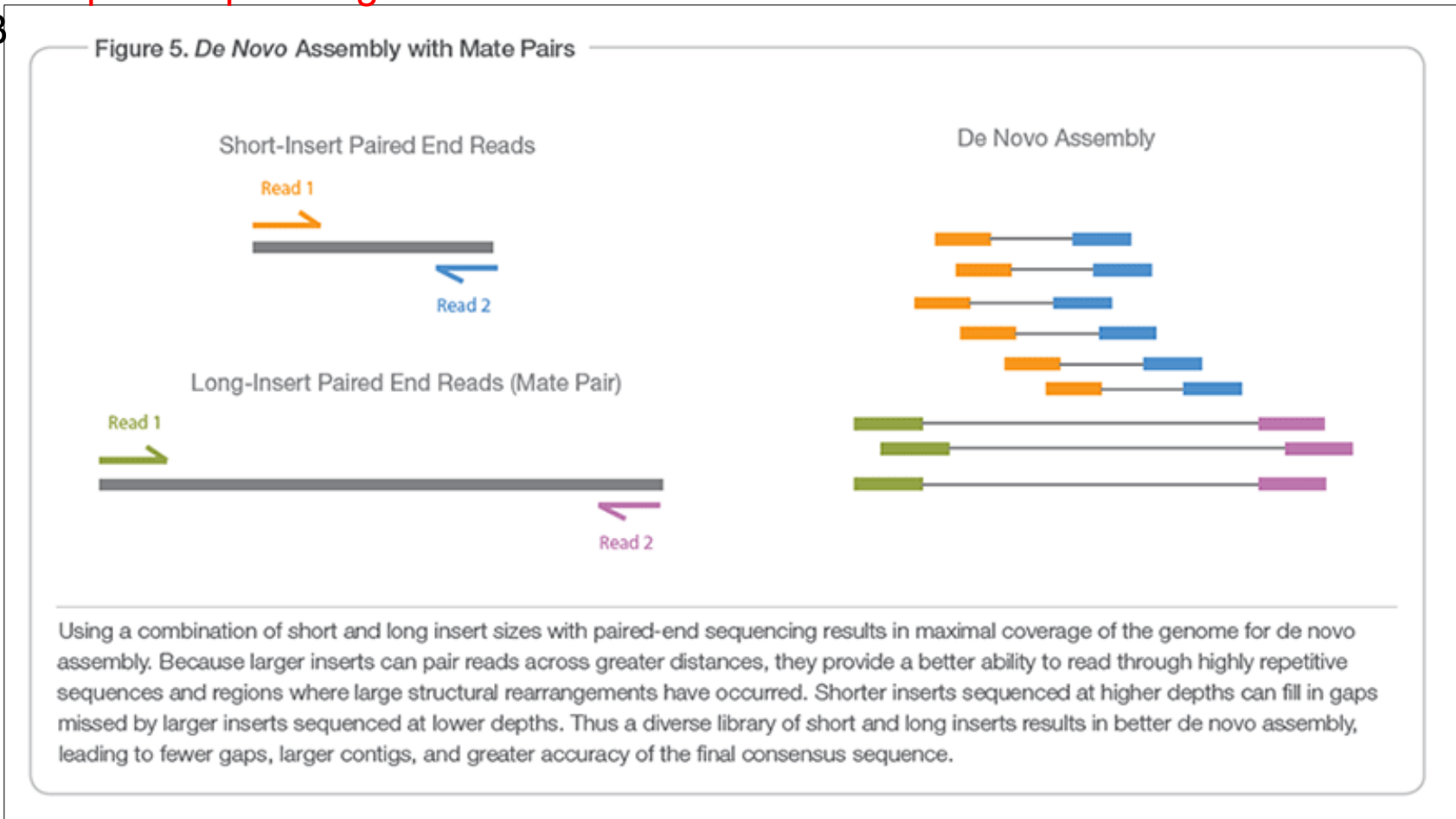


- Short fragments: overlap possible between forward and reverse reads
 - Longer fragments: no overlap (insert between 200 -1200 bp long)
- Better for detecting rearrangements, repetitive sequences, gene fusions, novel transcripts,

Run types

- Single-end sequencing
- Paired-end sequencing
- **Mate-pair sequencing**

• B

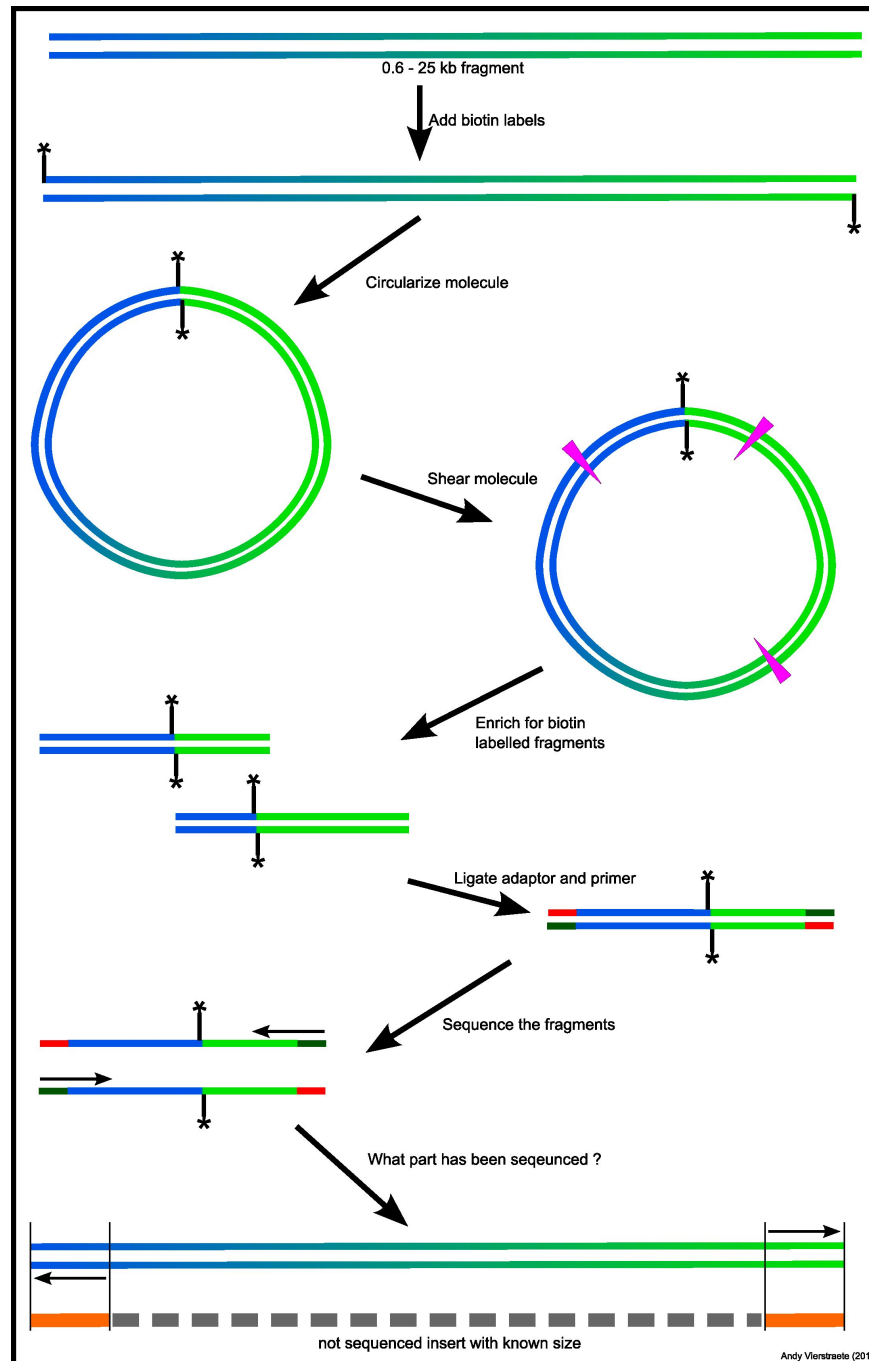


Long inserts possible: 0.6-25 kb

Better for detecting complex rearrangements, denovo sequencing, genome finishing, structural variant detection.

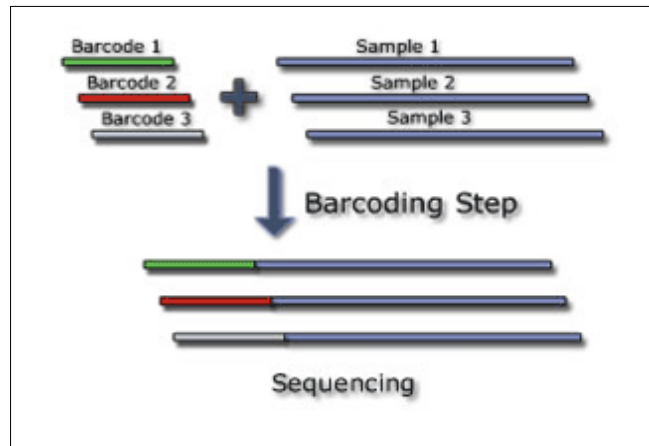
Run types

- Single-end sequencing
- Paired-end sequencing
- **Mate-pair sequencing**
- Barcoding samples

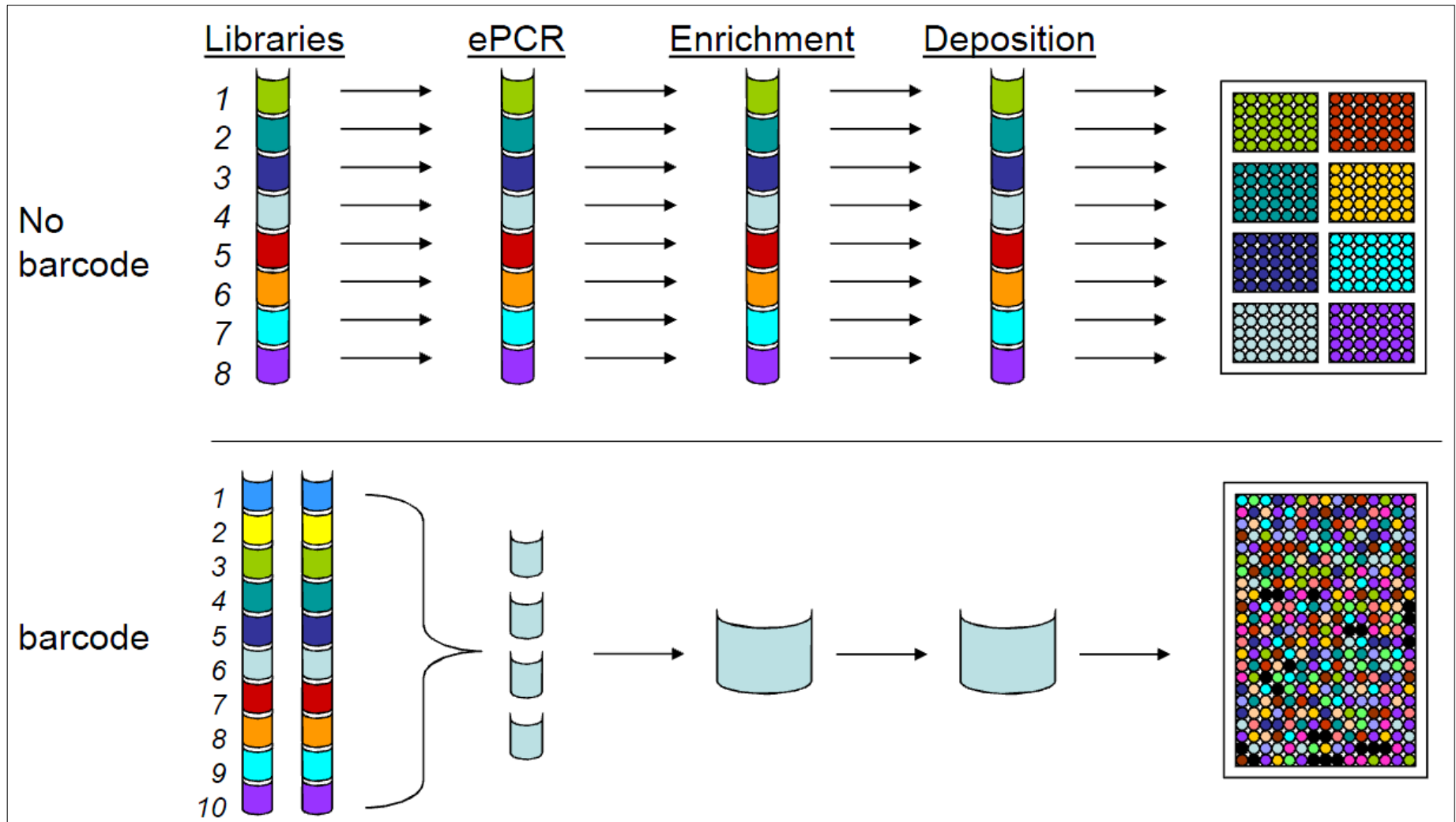


Run types

- Single-end sequencing
- Paired-end sequencing
- Mate-pair sequencing
- **Barcoding samples**



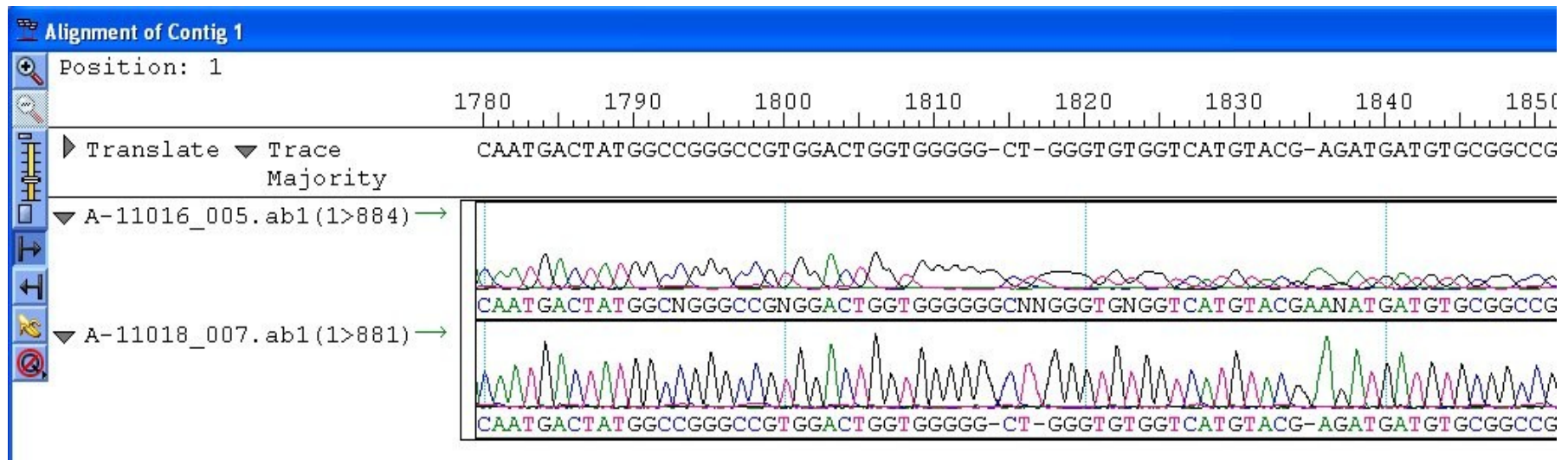
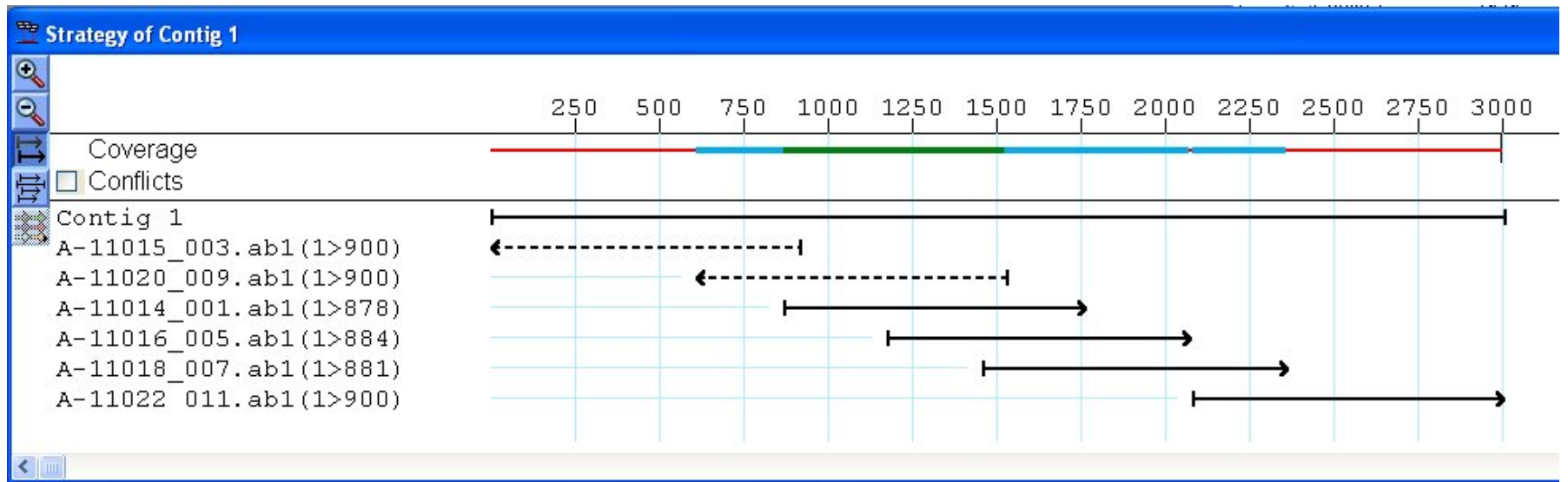
10-12 bp unique sequence is added to the DNA



Data analyses

143/165

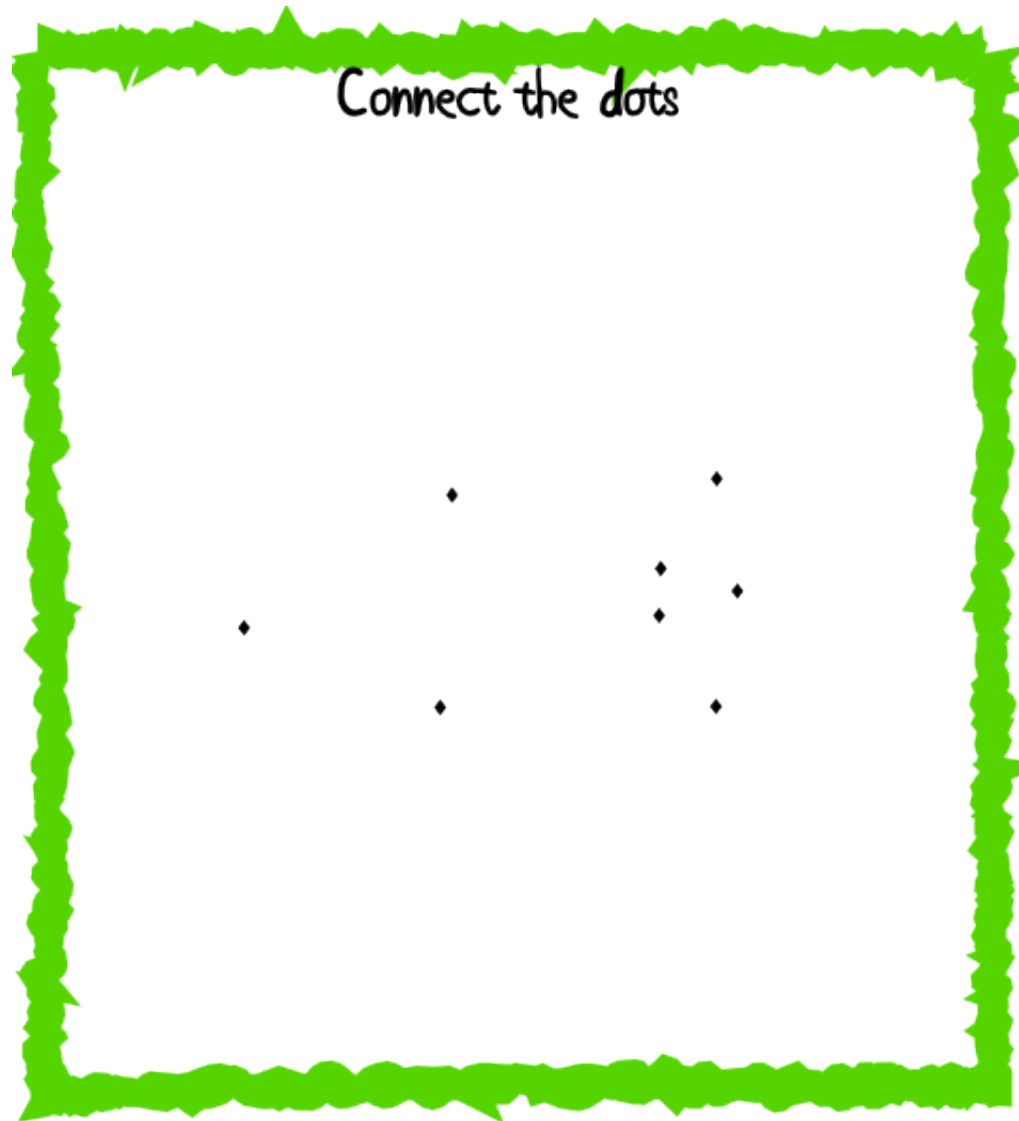
Sanger sequencing: e.g.: one gene sequenced with 6 primers



Manually check the assembly and correct errors.
2000 bp takes 5-10 minutes.

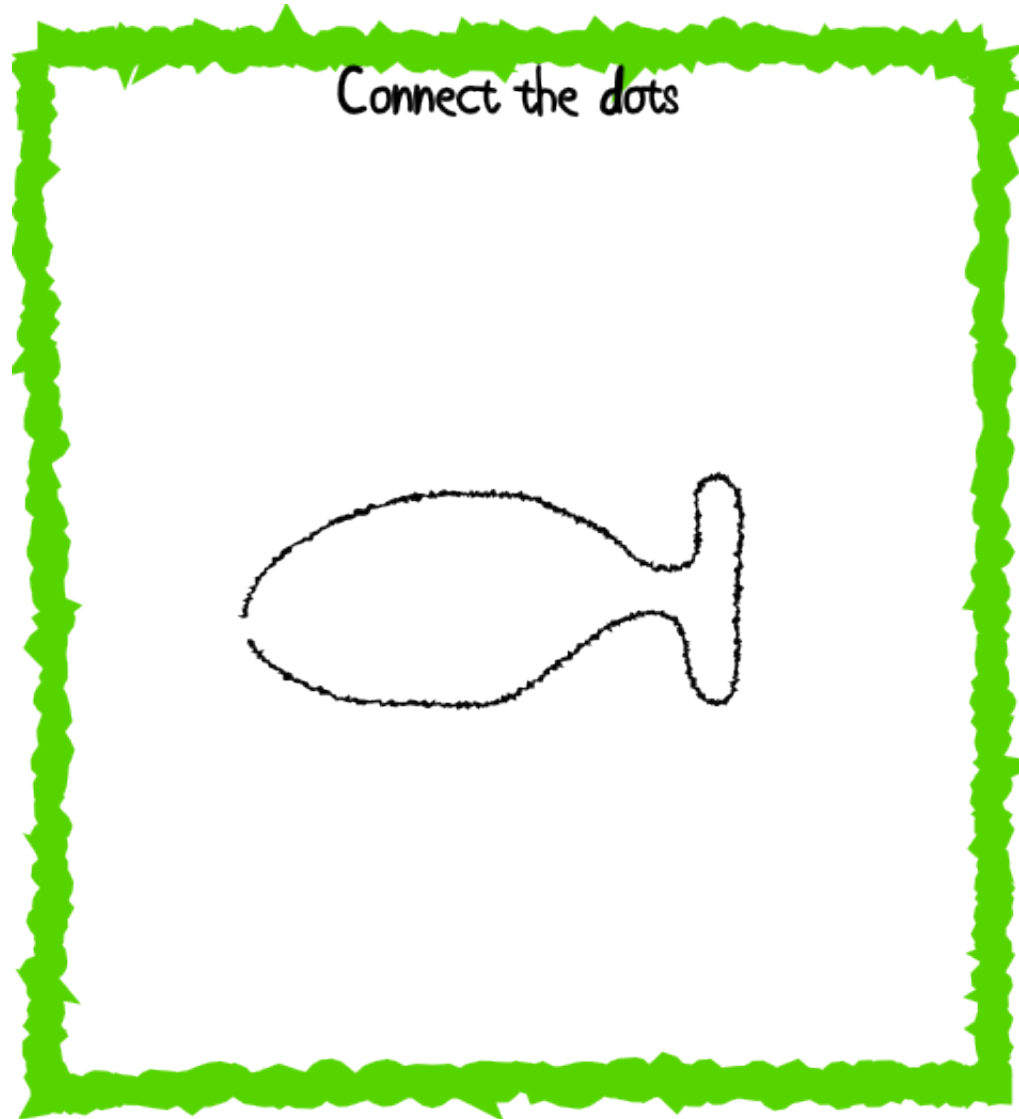
Data analyses

Sanger sequencing: simplified:



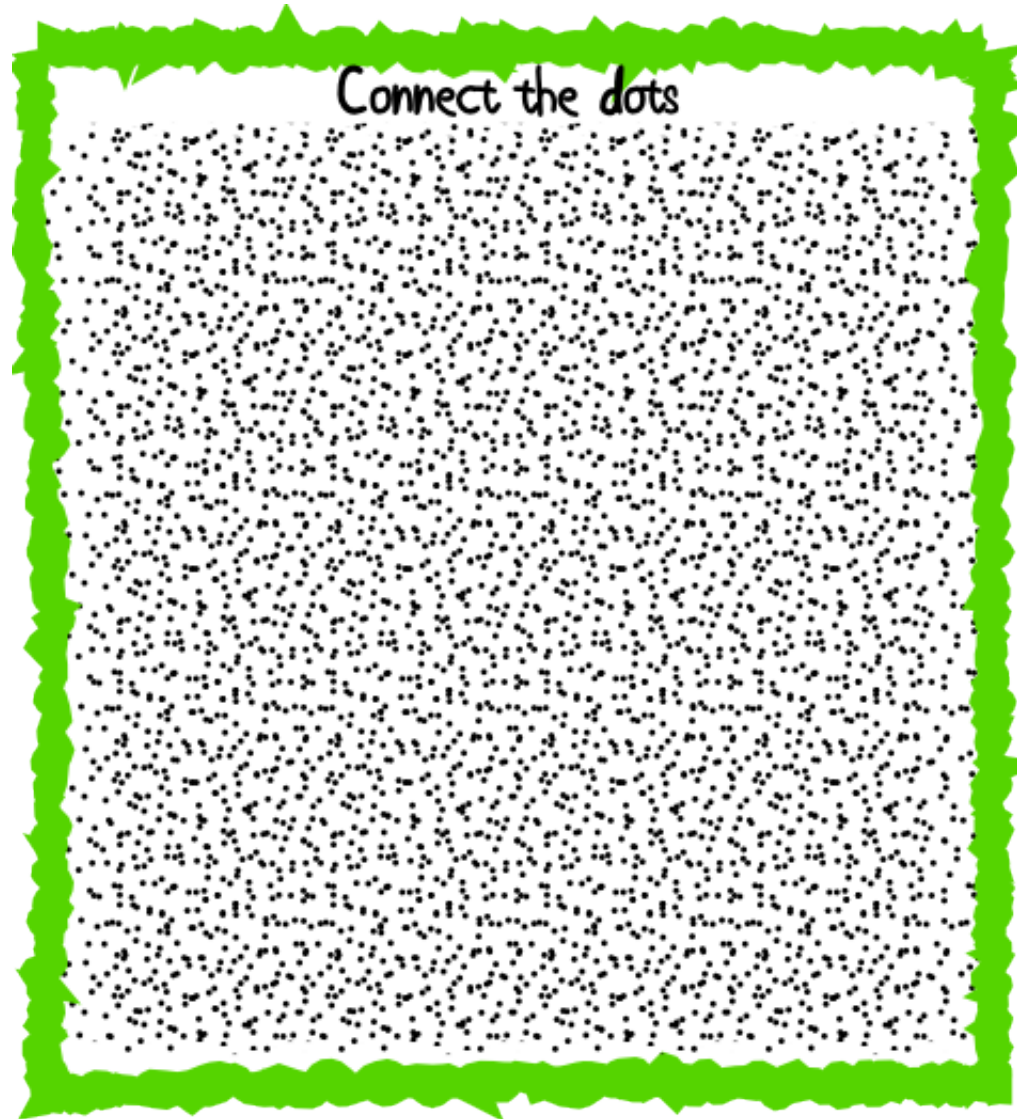
Data analyses

Sanger sequencing: simplified:



Data analyses

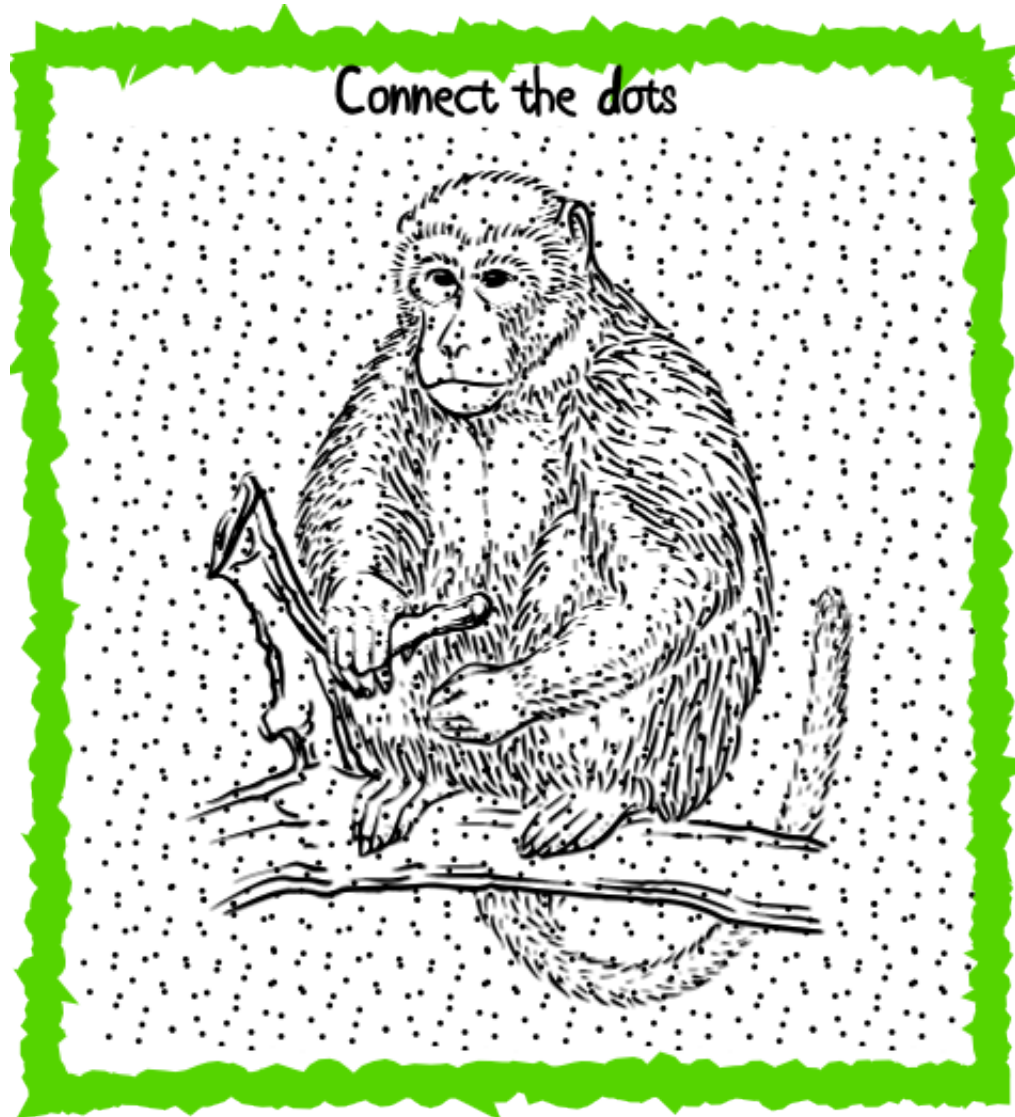
Next Generation sequencing: simplified:



Impossible to assemble manually

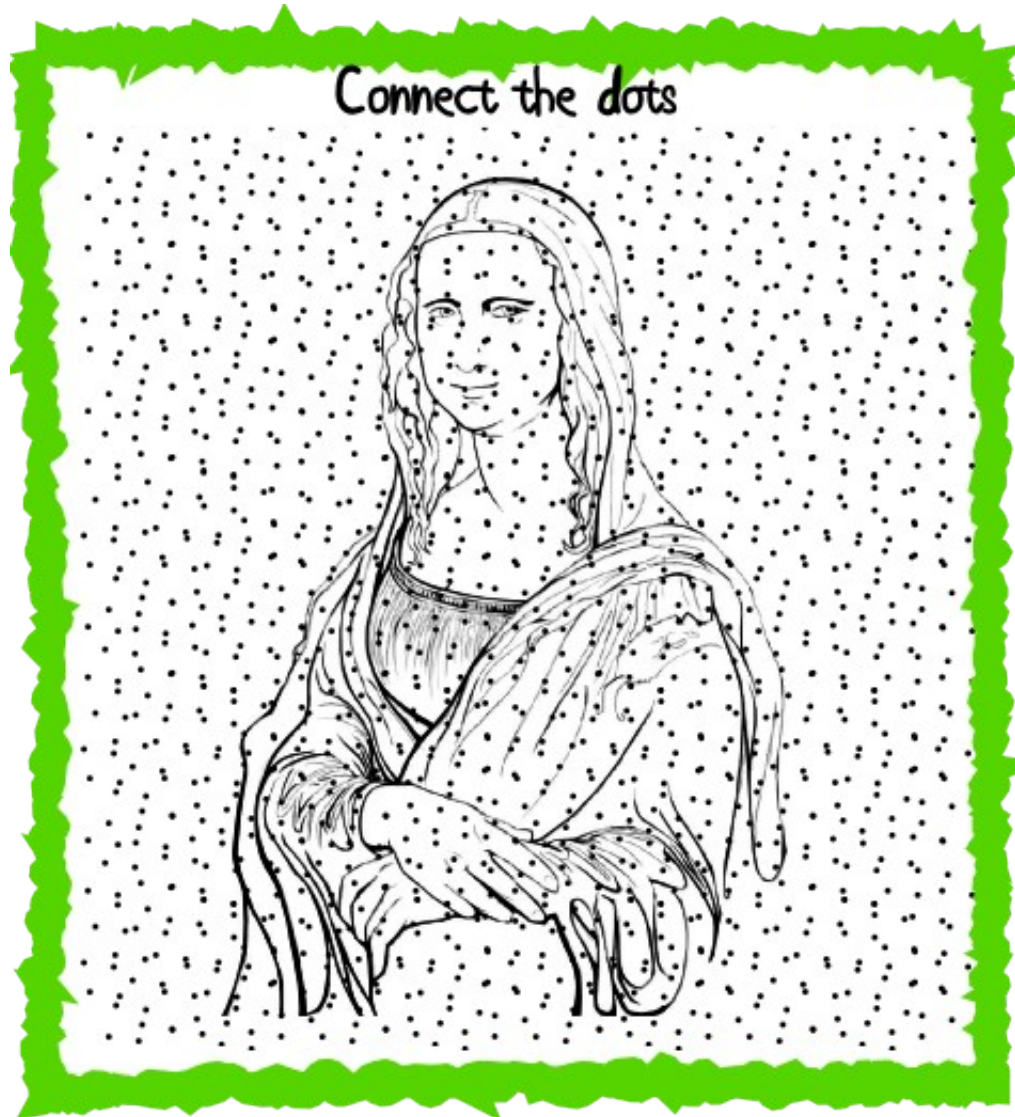
Data analyses

Next Generation sequencing: simplified:



Data analyses

Next Generation sequencing: simplified:

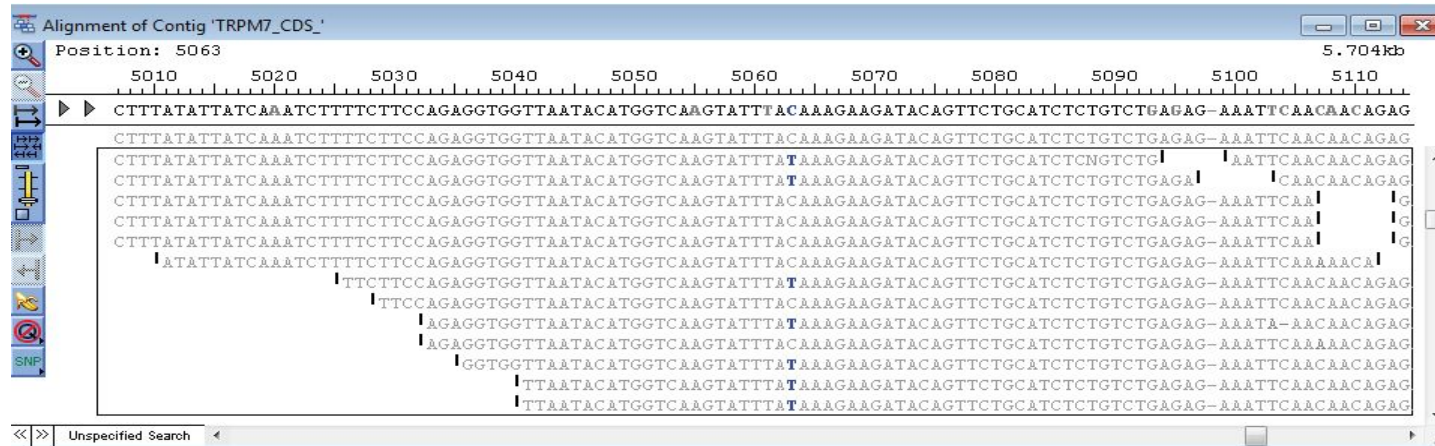
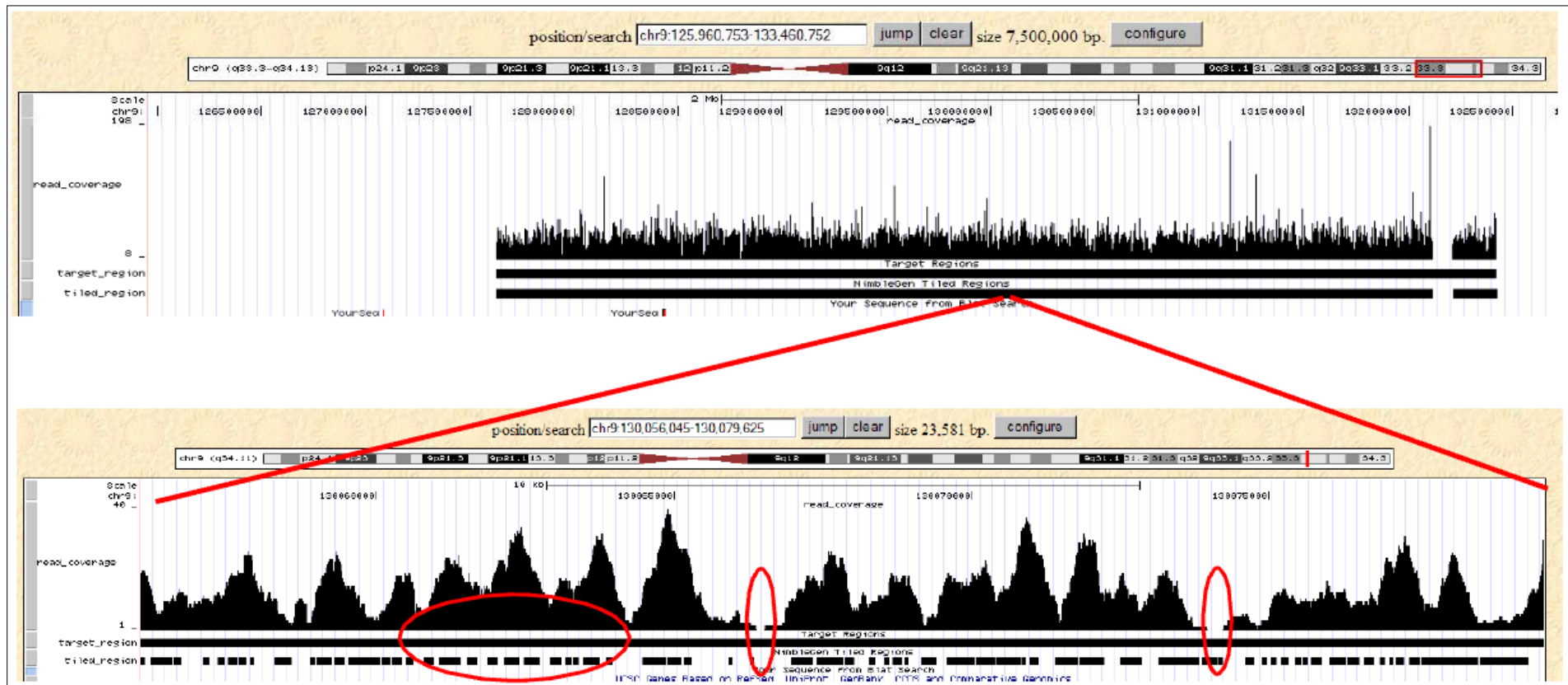


Same dataset, different parameters

Determining which assembly is the best is not an easy question. (Monya Baker. Nature Methods Volume 9 No.4, 333-337 (2012))

Data analyses

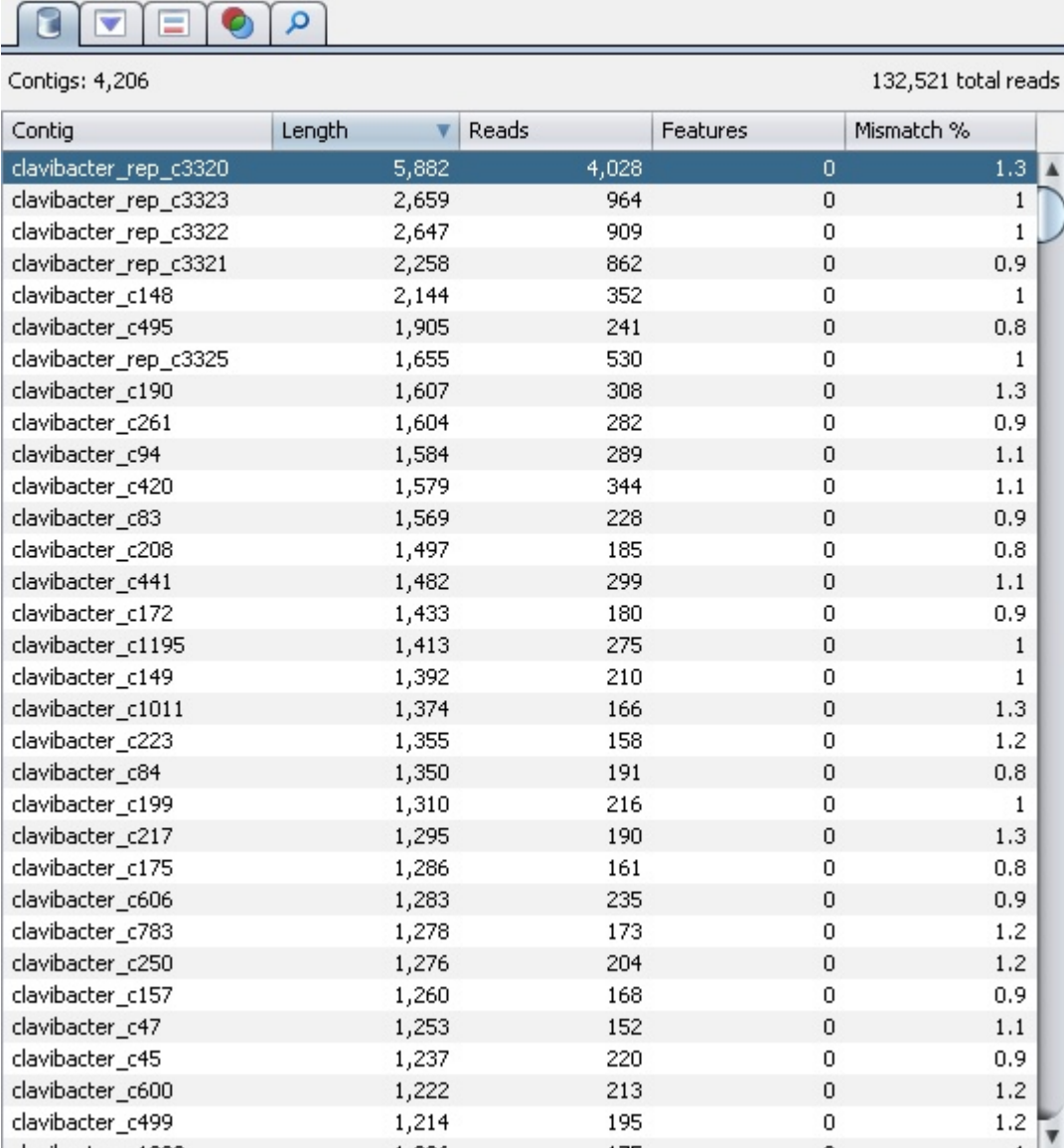
Next Generation sequencing: Impossible to check manually



Data analyses

150/165

Next Generation sequencing: Example of a bacteria sequenced on a 314 chip
 100 bp reads: 11,97 Mb: not enough to assemble complete genome.



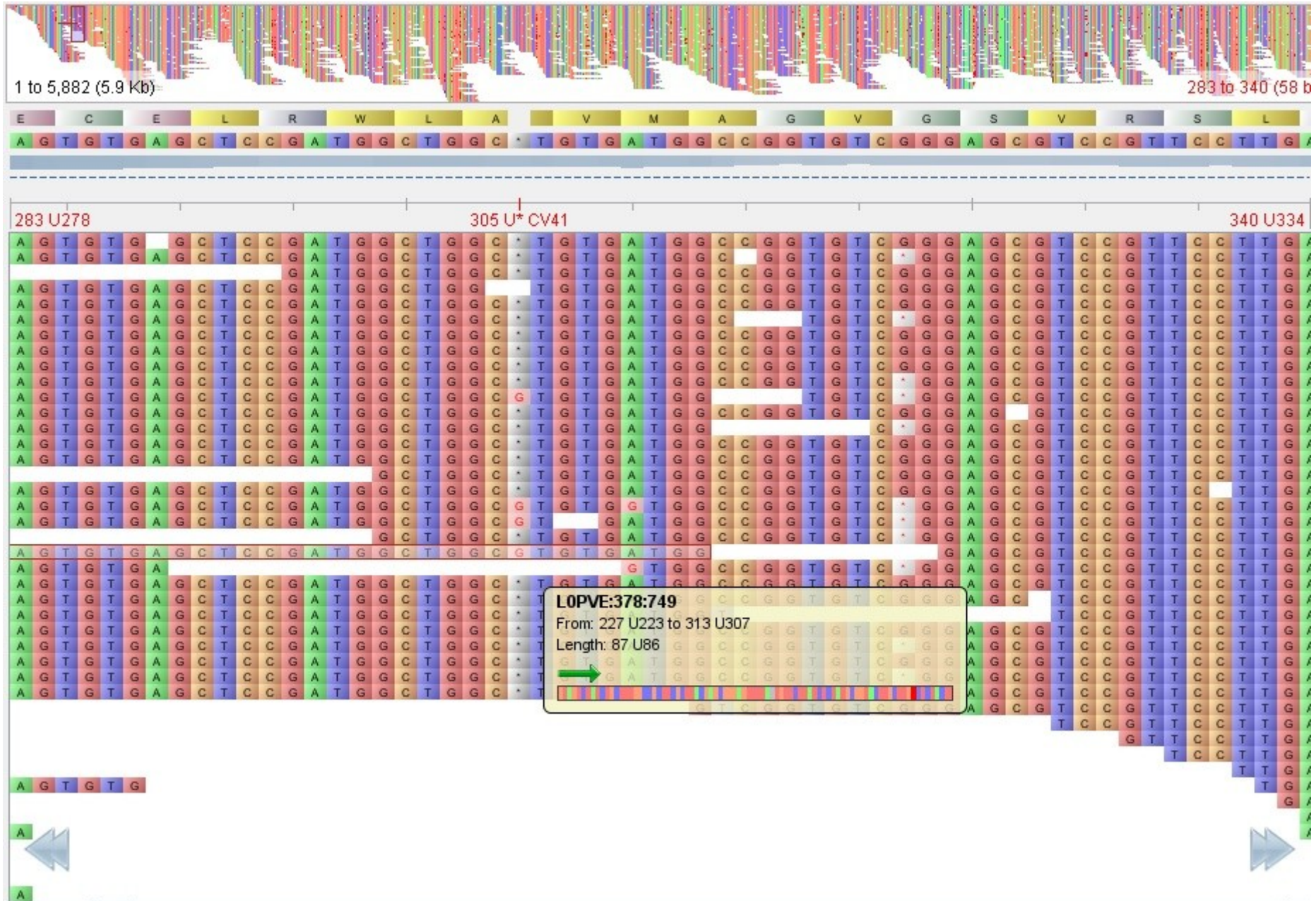
Contigs: 4,206 132,521 total reads

Contig	Length	Reads	Features	Mismatch %
clavibacter_rep_c3320	5,882	4,028	0	1.3
clavibacter_rep_c3323	2,659	964	0	1
clavibacter_rep_c3322	2,647	909	0	1
clavibacter_rep_c3321	2,258	862	0	0.9
clavibacter_c148	2,144	352	0	1
clavibacter_c495	1,905	241	0	0.8
clavibacter_rep_c3325	1,655	530	0	1
clavibacter_c190	1,607	308	0	1.3
clavibacter_c261	1,604	282	0	0.9
clavibacter_c94	1,584	289	0	1.1
clavibacter_c420	1,579	344	0	1.1
clavibacter_c83	1,569	228	0	0.9
clavibacter_c208	1,497	185	0	0.8
clavibacter_c441	1,482	299	0	1.1
clavibacter_c172	1,433	180	0	0.9
clavibacter_c1195	1,413	275	0	1
clavibacter_c149	1,392	210	0	1
clavibacter_c1011	1,374	166	0	1.3
clavibacter_c223	1,355	158	0	1.2
clavibacter_c84	1,350	191	0	0.8
clavibacter_c199	1,310	216	0	1
clavibacter_c217	1,295	190	0	1.3
clavibacter_c175	1,286	161	0	0.8
clavibacter_c606	1,283	235	0	0.9
clavibacter_c783	1,278	173	0	1.2
clavibacter_c250	1,276	204	0	1.2
clavibacter_c157	1,260	168	0	0.9
clavibacter_c47	1,253	152	0	1.1
clavibacter_c45	1,237	220	0	0.9
clavibacter_c600	1,222	213	0	1.2
clavibacter_c499	1,214	195	0	1.2

Data analyses

Next Generation sequencing:

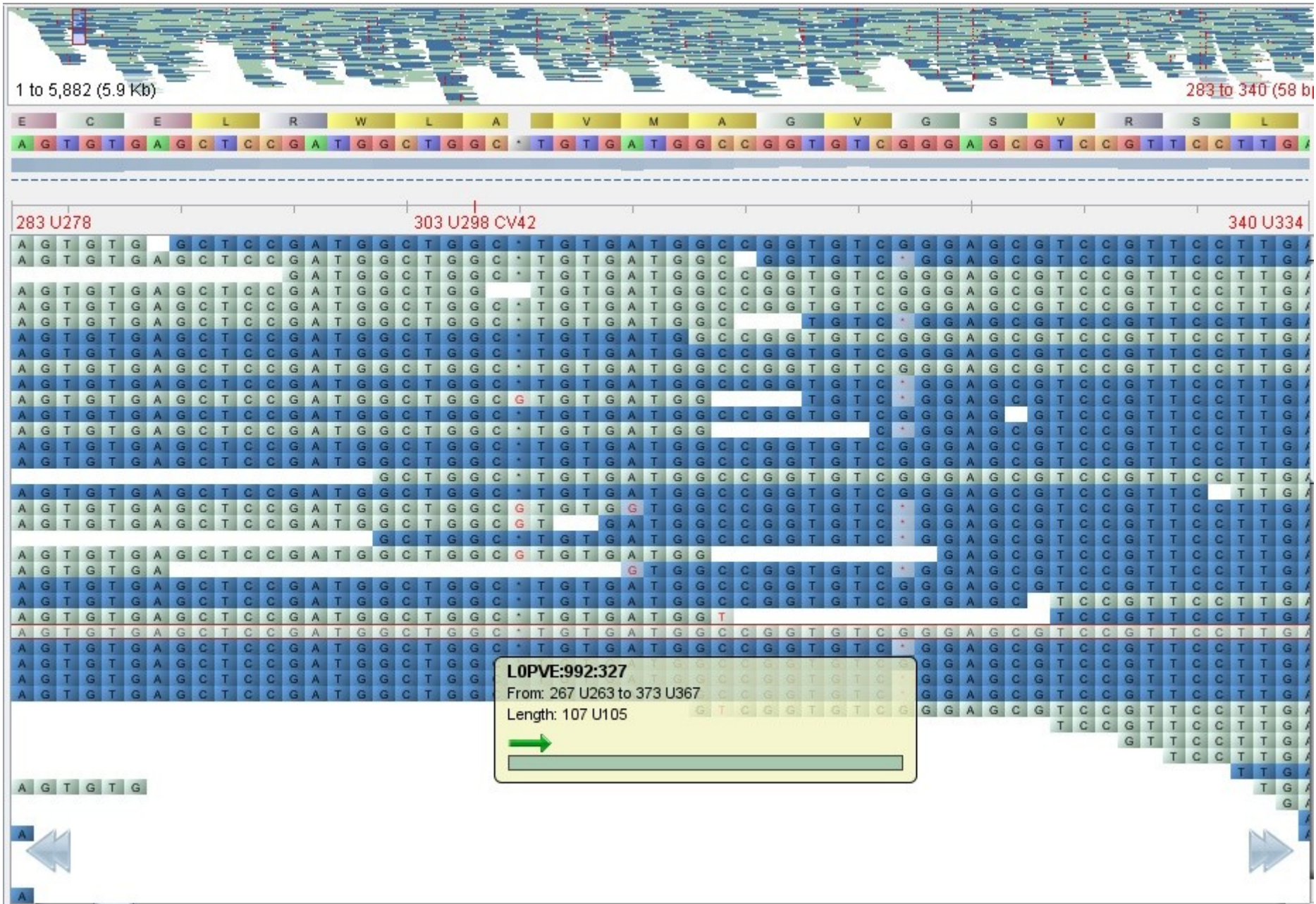
Assembly of the largest contig



Data analyses

Next Generation sequencing:

Light blue: forward
dark blue: reverse sequence



Data analyses

Next Generation sequencing:

CV: coverage of a specific part

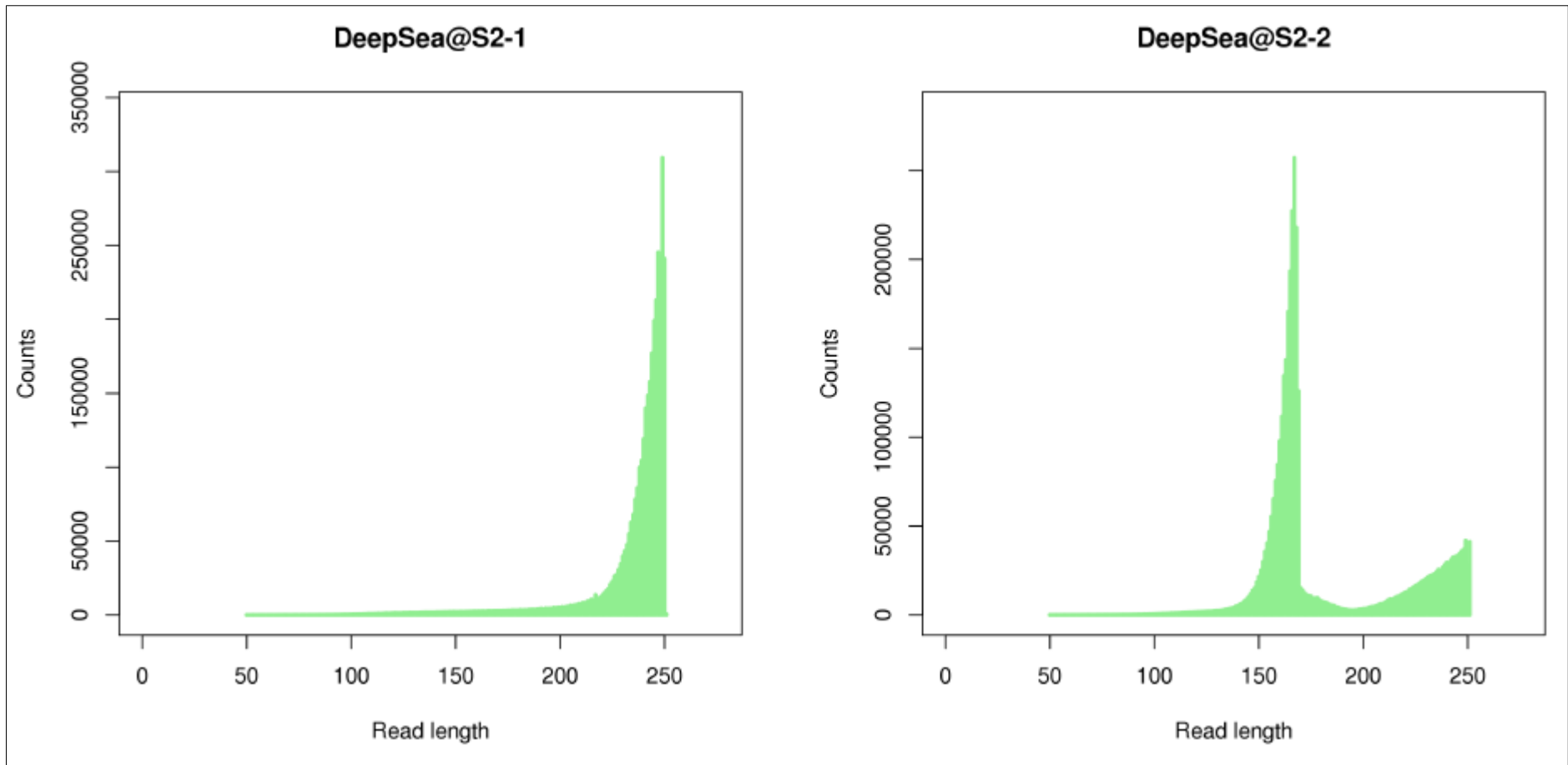


Data analyses

154/165

Next Generation sequencing: expectations from a run.

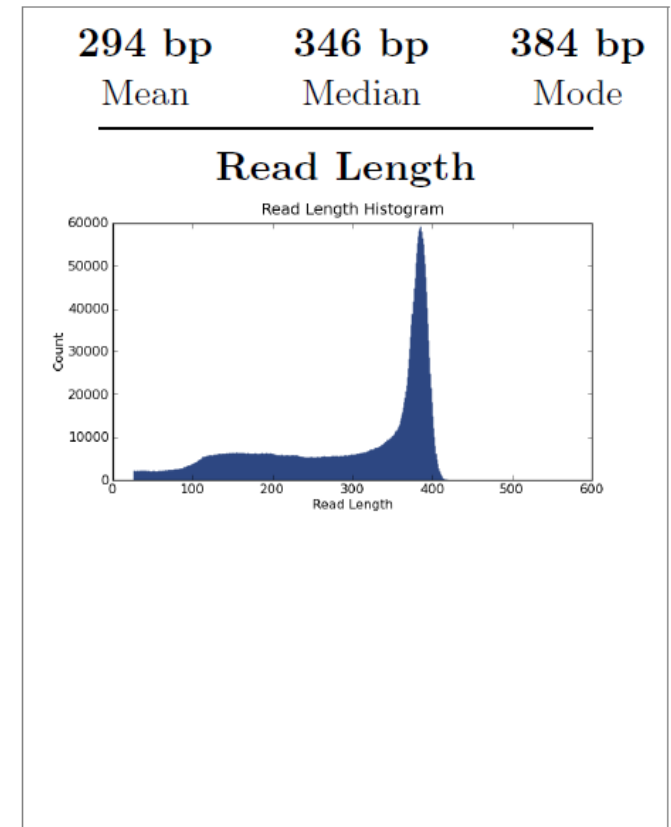
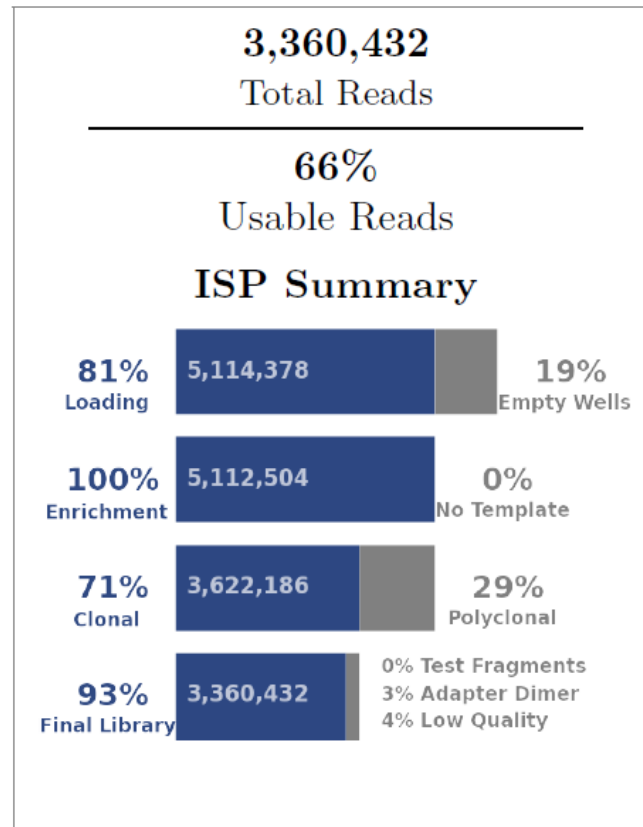
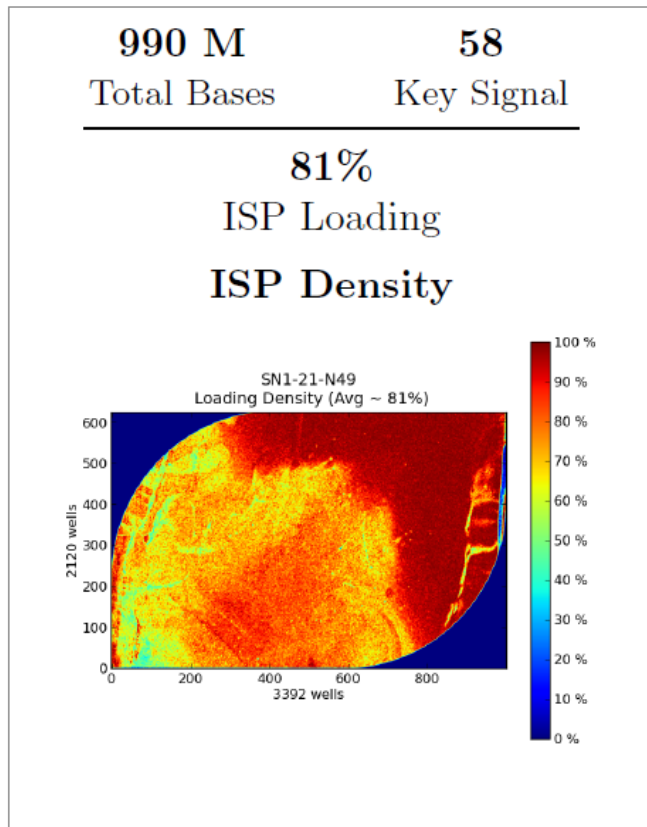
MiSeq run 2 x 250 bp



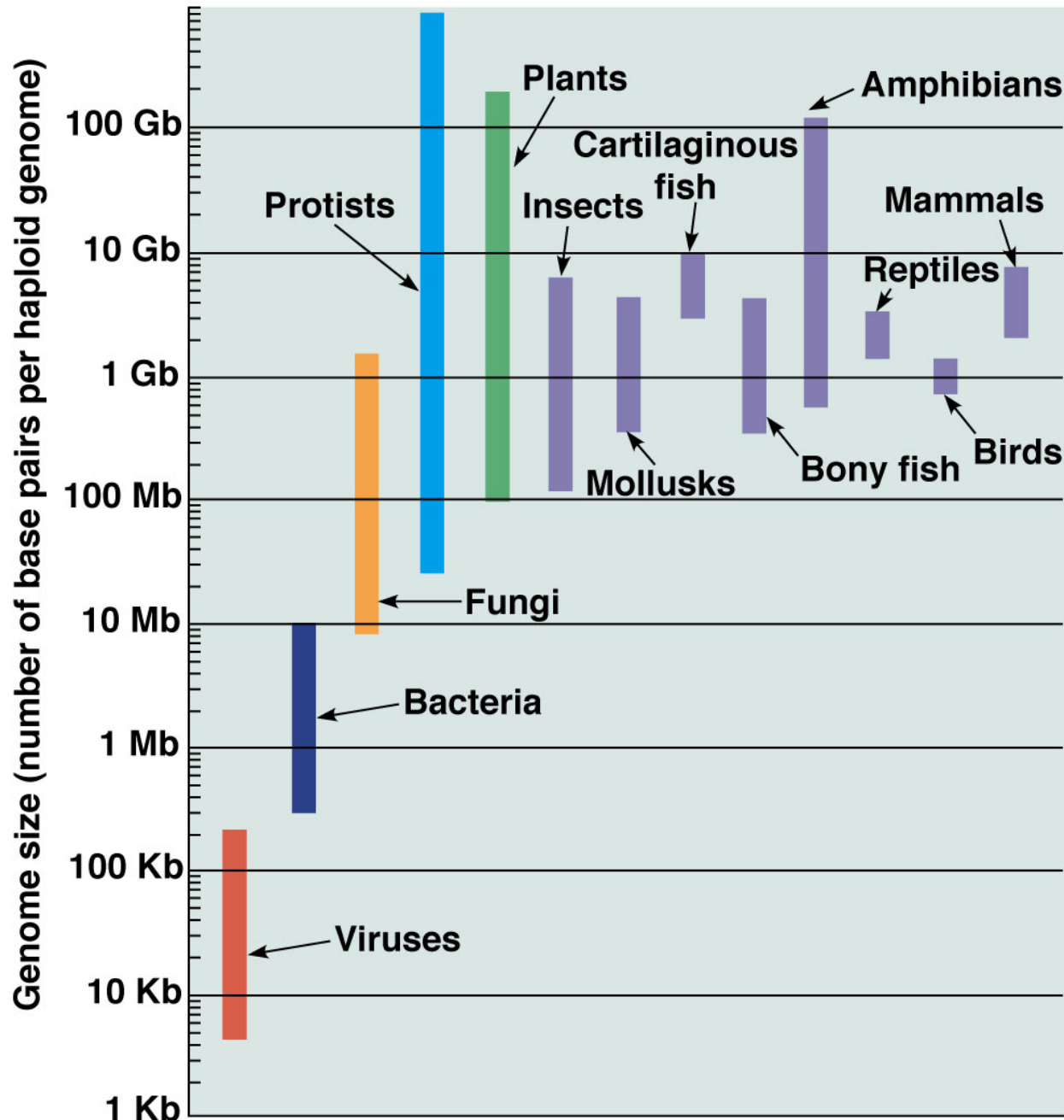
Data analyses

Next Generation sequencing: expectations from a run.

Ion Torrent PGM run 400 bp



Data analyses



One SOLiD run: 90 Gb (gigabases)
 -> 200 GB (gigabytes) of raw data
 -> mapping to reference:
 4 h on 250 cores server

1 Gb genome, 15 x coverage =
 15 Gbases to assemble or
 to map to a reference !

Total DNA sequencing:
 1x gDNA
 100x mDNA

Data analyses

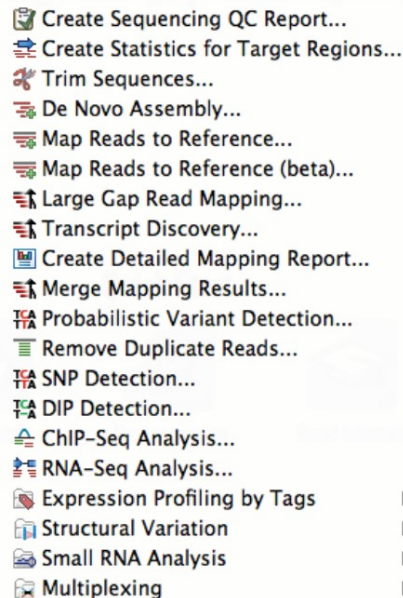
Lots of programs available: commercial or open source

Pro's:

- user friendly
- plug and play
- automated processing available
- integrated packages
- support

Con's:

- "limited" set of options
- expensive
- black box



Pro's:

- free
- most run on Linux platforms (stable)
- endless possibilities
- you can make your own pipelines
- (sometimes) clear methods

Con's:

- most run on Linux platforms (requires bioinformatics skills in Linux operating system and shell scripting)
- lots of small programs that only do one specific job
- maintenance
- time consuming (in the beginning)

A screenshot of a terminal window showing the output of a command. The output consists of several lines of text, each representing a comparison result:

```

process_id 51344 22000 comparisons done... (8244 similar)
process_id 51348 66500 comparisons done... (35231 similar)
process_id 51343 116500 comparisons done... (49279 similar)
process_id 51344 22500 comparisons done... (8522 similar)
process_id 51344 23000 comparisons done... (8795 similar)
process_id 51343 117000 comparisons done... (49489 similar)

```

Below the terminal output, a command is shown being executed in a shell:

```

avierstr@molfyl2:~$ spades.py --iontorrent --careful --s1 N54.fastq
--threads 4 -k 21,33,55 -o spadesN54

```



SEQanswers > Bioinformatics > Bioinformatics
 Software packages for next gen sequence analysis

Welcome, **avierstr.**

You last visited: Yesterday at 10:51 PM
 Private Messages: Unread 0, Total 0.

User CP

FAQ

Community ▾

Calendar

New Posts

Search ▾

Quick Links ▾

Log Out

Similar Threads

Closed

Page 1 of 12 1 2 3 11 > Last >> ▾

Thread Tools ▾ Search this Thread ▾

01-24-2008, 08:19 AM

#1

[sci_guy](#)

Member

Location: Sydney

Join Date: Jan 2008
 Posts: 80

Software packages for next gen sequence analysis

28 Dec 2009: This thread has been closed. Please see our [wiki software portal](#) for information about each of these packages.

A reasonably thorough table of next-gen-seq software available in the commercial and public domain

Integrated solutions

- * [CLCbio Genomics Workbench](#) - *de novo* and reference assembly of Sanger, Roche FLX, Illumina, Helicos, and SOLiD data. Commercial next-gen-seq software that extends the CLCbio Main Workbench software. Includes SNP detection, CHIP-seq, browser and other features. Commercial. Windows, Mac OS X and Linux.
- * [Galaxy](#) - Galaxy = interactive and reproducible genomics. A job webportal.
- * [Genomatix](#) - Integrated Solutions for Next Generation Sequencing data analysis.
- * [JMP Genomics](#) - Next gen visualization and statistics tool from SAS. They are [working with NCGR](#) to refine this tool and produce others.
- * [NextGENe](#) - *de novo* and reference assembly of Illumina, SOLiD and Roche FLX data. Uses a novel Condensation Assembly Tool approach where reads are joined via "anchors" into mini-contigs before assembly. Includes SNP detection, CHIP-seq, browser and other features. Commercial. Win or MacOS.
- * [SeqMan Genome Analyser](#) - Software for Next Generation sequence assembly of Illumina, Roche FLX and Sanger data integrating with Lasergene Sequence Analysis software for additional analysis and visualization capabilities. Can use a hybrid templated/*de novo* approach. Commercial. Win or Mac OS X.
- * [SHORE](#) - SHORE, for Short Read, is a mapping and analysis pipeline for short DNA sequences produced on a Illumina Genome Analyzer. A suite created by the 1001 Genomes project. Source for POSIX.
- * [SlimSearch](#) - Fledgling commercial product.

Align/Assemble to a reference

- * [BFAST](#) - Blat-like Fast Accurate Search Tool. Written by Nils Homer, Stanley F. Nelson and Barry Merriman at UCLA.
- * [Bowtie](#) - Ultrafast, memory-efficient short read aligner. It aligns short DNA sequences (reads) to the human genome at a rate of 25 million reads per hour on a typical workstation with 2 gigabytes of memory. Uses a Burrows-Wheeler-Transformed (BWT) index. [Link to discussion thread here](#). Written by Ben Langmead and Cole Trapnell. Linux, Windows, and Mac OS X.
- * [BWA](#) - Heng Lee's BWT Alignment program - a progression from Maq. BWA is a fast light-weighted tool that aligns short sequences to a sequence database, such as the human reference genome. By default, BWA finds an alignment within edit distance 2 to the query sequence. C++ source.
- * [ELAND](#) - Efficient Large-Scale Alignment of Nucleotide Databases. Whole genome alignments to a reference genome. Written by Illumina author Anthony J. Cox for the Solexa 1G machine.
- * [Exonerate](#) - Various forms of pairwise alignment (including Smith-Waterman-Gotoh) of DNA/protein against a reference. Authors are Guy St C Slater and Ewan Birney from EMBL. C for POSIX.

- * [Exonerate](#) - Various forms of pairwise alignment (including Smith-Waterman-Gotoh) of DNA/protein against a reference. Authors are Guy St C Slater and Ewan Birney from EMBL. C for POSIX.
- * [GenomeMapper](#) - GenomeMapper is a short read mapping tool designed for accurate read alignments. It quickly aligns millions of reads either with ungapped or gapped alignments. A tool created by the 1001 Genomes project. Source for POSIX.
- * [GMAP](#) - GMAP (Genomic Mapping and Alignment Program) for mRNA and EST Sequences. Developed by Thomas Wu and Colin Watanabe at Genentec. C/Perl for Unix.
- * [gnumap](#) - The Genomic Next-generation Universal MAPper (gnumap) is a program designed to accurately map sequence data obtained from next-generation sequencing machines (specifically that of Solexa/Illumina) back to a genome of any size. It seeks to align reads from nonunique repeats using statistics. From authors at Brigham Young University. C source/Unix.
- * [MAQ](#) - Mapping and Assembly with Qualities (renamed from MAPASS2). Particularly designed for Illumina with preliminary functions to handle ABI SOLID data. Written by Heng Li from the Sanger Centre. Features extensive supporting tools for DIP/SNP detection, etc. C++ source
- * [MOSAIK](#) - MOSAIK produces gapped alignments using the Smith-Waterman algorithm. Features a number of support tools. Support for Roche FLX, Illumina, SOLID, and Helicos. Written by Michael Strömberg at Boston College. Win/Linux/MacOSX
- * [MrFAST and MrsFAST](#) - mrFAST & mrsFAST are designed to map short reads generated with the Illumina platform to reference genome assemblies; in a fast and memory-efficient manner. Robust to INDELS and MrsFAST has a bisulphite mode. Authors are from the University of Washington. C as source.
- * [MUMmer](#) - MUMmer is a modular system for the rapid whole genome alignment of finished or draft sequence. Released as a package providing an efficient suffix tree library, seed-and-extend alignment, SNP detection, repeat detection, and visualization tools. Version 3.0 was developed by Stefan Kurtz, Adam Phillippy, Arthur L Delcher, Michael Smoot, Martin Shumway, Corina Antonescu and Steven L Salzberg - most of whom are at The Institute for Genomic Research in Maryland, USA. POSIX OS required.
- * [Novocraft](#) - Tools for reference alignment of paired-end and single-end Illumina reads. Uses a Needleman-Wunsch algorithm. Can support Bis-Seq. Commercial. Available free for evaluation, educational use and for use on open not-for-profit projects. Requires Linux or Mac OS X.
- * [PASS](#) - It supports Illumina, SOLID and Roche-FLX data formats and allows the user to modulate very finely the sensitivity of the alignments. Spaced seed initial filter, then NW dynamic algorithm to a SW(like) local alignment. Authors are from CRIBI in Italy. Win/Linux.
- * [RMAP](#) - Assembles 20 - 64 bp Illumina reads to a FASTA reference genome. By Andrew D. Smith and Zhenyu Xuan at CSHL. (published in BMC Bioinformatics). POSIX OS required.
- * [SeqMap](#) - Supports up to 5 or more bp mismatches/INDELS. Highly tunable. Written by Hui Jiang from the Wong lab at Stanford. Builds available for most OS's.
- * [SHRIMP](#) - Assembles to a reference sequence. Developed with Applied Biosystem's colourspace genomic representation in mind. Authors are Michael Brudno and Stephen Rumble at the University of Toronto. POSIX.
- * [Slider](#) - An application for the Illumina Sequence Analyzer output that uses the probability files instead of the sequence files as an input for alignment to a reference sequence or a set of reference sequences. Authors are from BCGSC. Paper is [here](#).
- * [SOAP](#) - SOAP (Short Oligonucleotide Alignment Program). A program for efficient gapped and ungapped alignment of short oligonucleotides onto reference sequences. The updated version uses a BWT. Can call SNPs and INDELS. Author is Ruiqiang Li at the Beijing Genomics Institute. C++, POSIX.
- * [SSAHA](#) - SSAHA (Sequence Search and Alignment by Hashing Algorithm) is a tool for rapidly finding near exact matches in DNA or protein databases using a hash table. Developed at the Sanger Centre by Zemin Ning, Anthony Cox and James Mullikin. C++ for Linux/Alpha.
- * [SOCS](#) - Aligns SOLID data. SOCS is built on an iterative variation of the Rabin-Karp string search algorithm, which uses hashing to reduce the set of possible matches, drastically increasing search speed. Authors are Ondov B, Varadarajan A, Passalacqua KD and Bergman NH.
- * [SWIFT](#) - The SWIFT suit is a software collection for fast index-based sequence comparison. It contains: SWIFT - fast local alignment search, guaranteeing to find epsilon-matches between two sequences. SWIFT BALSAM - a very fast program to find semiglobal non-gapped alignments based on k-mer seeds. Authors are Kim Rasmussen (SWIFT) and Wolfgang Gerlach (SWIFT BALSAM)
- * [SXOligoSearch](#) - SXOligoSearch is a commercial platform offered by the Malaysian based [Synamatrix](#). Will align Illumina reads against a range of Refseq RNA or NCBI genome builds for a number of organisms. Web Portal. OS independent.
- * [Vmatch](#) - A versatile software tool for efficiently solving large scale sequence matching tasks. Vmatch subsumes the software tool REPuter, but is much more general, with a very flexible user interface, and improved space and time requirements. Essentially a large string matching toolbox. POSIX.
- * [Zoom](#) - ZOOM (Zillions Of Oligos Mapped) is designed to map millions of short reads, emerged by next-generation sequencing technology, back to the reference genomes, and carry out post-analysis. ZOOM is developed to be highly accurate, flexible, and user-friendly with speed being a critical priority. Commercial. Supports Illumina and SOLID data.

De novo Align/Assemble

- * [ABYSS](#) - Assembly By Short Sequences. ABYSS is a de novo sequence assembler that is designed for very short reads. The single-processor version is useful for assembling genomes up to 40-50 Mbases in size. The parallel version is implemented using MPI and is capable of assembling larger genomes. By Simpson JT and others at the Canada's Michael Smith Genome Sciences Centre. C++ as source.
- * [ALLPATHS](#) - ALLPATHS: De novo assembly of whole-genome shotgun microreads. ALLPATHS is a whole genome shotgun assembler that can generate high quality assemblies from short reads. Assemblies are presented in a graph form that retains ambiguities, such as those arising from polymorphism,

- * [ALLPATHS](#) - ALLPATHS: De novo assembly of whole-genome shotgun microreads. ALLPATHS is a whole genome shotgun assembler that can generate high quality assemblies from short reads. Assemblies are presented in a graph form that retains ambiguities, such as those arising from polymorphism, thereby providing information that has been absent from previous genome assemblies. Broad Institute.
- * [Edena](#) - Edena (Exact DE Novo Assembler) is an assembler dedicated to process the millions of very short reads produced by the Illumina Genome Analyzer. Edena is based on the traditional overlap layout paradigm. By D. Hernandez, P. François, L. Farinelli, M. Osteras, and J. Schrenzel. Linux/Win.
- * [EULER-SR](#) - Short read *de novo* assembly. By Mark J. Chaisson and Pavel A. Pevzner from UCSD (published in Genome Research). Uses a de Bruijn graph approach.
- * [MIRA2](#) - MIRA (Mimicking Intelligent Read Assembly) is able to perform true hybrid de-novo assemblies using reads gathered through 454 sequencing technology (GS20 or GS FLX). Compatible with 454, Solexa and Sanger data. Linux OS required.
- * [SEQAN](#) - A Consistency-based Consensus Algorithm for De Novo and Reference-guided Sequence Assembly of Short Reads. By Tobias Rausch and others. C++, Linux/Win.
- * [SHARCGS](#) - De novo assembly of short reads. Authors are Dohm JC, Lottaz C, Borodina T and Himmelbauer H. from the Max-Planck-Institute for Molecular Genetics.
- * [SSAKE](#) - The Short Sequence Assembly by K-mer search and 3' read Extension (SSAKE) is a genomics application for aggressively assembling millions of short nucleotide sequences by progressively searching for perfect 3'-most k-mers using a DNA prefix tree. Authors are René Warren, Granger Sutton, Steven Jones and Robert Holt from the Canada's Michael Smith Genome Sciences Centre. Perl/Linux.
- * [SOAPdenovo](#) - Part of the SOAP suite. See above.
- * [VCAKE](#) - De novo assembly of short reads with robust error correction. An improvement on early versions of SSAKE.
- * [Velvet](#) - Velvet is a de novo genomic assembler specially designed for short read sequencing technologies, such as Solexa or 454. Need about 20-25X coverage and paired reads. Developed by Daniel Zerbino and Ewan Birney at the European Bioinformatics Institute (EMBL-EBI).

SNP/Indel Discovery

- * [ssahaSNP](#) - ssahaSNP is a polymorphism detection tool. It detects homozygous SNPs and indels by aligning shotgun reads to the finished genome sequence. Highly repetitive elements are filtered out by ignoring those kmer words with high occurrence numbers. More tuned for ABI Sanger reads. Developers are Adam Spargo and Zemin Ning from the Sanger Centre. Compaq Alpha, Linux-64, Linux-32, Solaris and Mac
- * [PolyBayesShort](#) - A re-incarnation of the PolyBayes SNP discovery tool developed by Gabor Marth at Washington University. This version is specifically optimized for the analysis of large numbers (millions) of high-throughput next-generation sequencer reads, aligned to whole chromosomes of model organism or mammalian genomes. Developers at Boston College. Linux-64 and Linux-32.
- * [PyroBayes](#) - PyroBayes is a novel base caller for pyrosequences from the 454 Life Sciences sequencing machines. It was designed to assign more accurate base quality estimates to the 454 pyrosequences. Developers at Boston College.

Genome Annotation/Genome Browser/Alignment Viewer/Assembly Database

- * [EagleView](#) - An information-rich genome assembler viewer. EagleView can display a dozen different types of information including base quality and flowgram signal. Developers at Boston College.
- * [LookSeq](#) - LookSeq is a web-based application for alignment visualization, browsing and analysis of genome sequence data. LookSeq supports multiple sequencing technologies, alignment sources, and viewing modes; low or high-depth read pileups; and easy visualization of putative single nucleotide and structural variation. From the Sanger Centre.
- * [MapView](#) - MapView: visualization of short reads alignment on desktop computer. From the Evolutionary Genomics Lab at Sun-Yat Sen University, China. Linux.
- * [SAM](#) - Sequence Assembly Manager. Whole Genome Assembly (WGA) Management and Visualization Tool. It provides a generic platform for manipulating, analyzing and viewing WGA data, regardless of input type. Developers are Rene Warren, Yaron Butterfield, Asim Siddiqui and Steven Jones at Canada's Michael Smith Genome Sciences Centre. MySQL backend and Perl-CGI web-based frontend/Linux.
- * [STADEN](#) - Includes GAP4. GAP5 once completed will handle next-gen sequencing data. A partially implemented test version is available [here](#)
- * [XMatchView](#) - A visual tool for analyzing cross_match alignments. Developed by Rene Warren and Steven Jones at Canada's Michael Smith Genome Sciences Centre. Python/Win or Linux.

Counting e.g. ChIP-Seq, Bis-Seq, CNV-Seq

- * [BS-Seq](#) - The source code and data for the "Shotgun Bisulphite Sequencing of the Arabidopsis Genome Reveals DNA Methylation Patterning" Nature paper by [Cokus et al.](#) (Steve Jacobsen's lab at UCLA). POSIX.
- * [ChIPSeq](#) - Program used by Johnson et al. (2007) in their Science publication
- * [CNV-Seq](#) - CNV-seq, a new method to detect copy number variation using high-throughput sequencing. Chao Xie and Martti T Tammi at the National University of Singapore. Perl/R.
- * [FindPeaks](#) - perform analysis of ChIP-Seq experiments. It uses a naive algorithm for identifying regions of high coverage, which represent Chromatin Immunoprecipitation enrichment of sequence fragments, indicating the location of a bound protein of interest. Original algorithm by Matthew Bainbridge, in collaboration with Gordon Robertson. Current code and implementation by Anthony Fejes. Authors are from the Canada's Michael Smith Genome Sciences Centre. JAVA/OS independent. Latest versions available as part of the [Vancouver Short Read Analysis Package](#)

[ChIP-seq](#) - perform analysis of ChIP-Seq experiments. It uses a naive algorithm for identifying regions of high coverage, which represent enrichment. Immunoprecipitation enrichment of sequence fragments, indicating the location of a bound protein of interest. Original algorithm by Matthew Bainbridge, in collaboration with Gordon Robertson. Current code and implementation by Anthony Fejes. Authors are from the Canada's Michael Smith Genome Sciences Centre. JAVA/OS independent. Latest versions available as part of the [Vancouver Short Read Analysis Package](#)

- * [MACS](#) - Model-based Analysis for ChIP-Seq. MACS empirically models the length of the sequenced ChIP fragments, which tends to be shorter than sonication or library construction size estimates, and uses it to improve the spatial resolution of predicted binding sites. MACS also uses a dynamic Poisson distribution to effectively capture local biases in the genome sequence, allowing for more sensitive and robust prediction. Written by Yong Zhang and Tao Liu from Xiaole Shirley Liu's Lab.
- * [PeakSeq](#) - PeakSeq: Systematic Scoring of ChIP-Seq Experiments Relative to Controls. a two-pass approach for scoring ChIP-Seq data relative to controls. The first pass identifies putative binding sites and compensates for variation in the mappability of sequences across the genome. The second pass filters out sites that are not significantly enriched compared to the normalized input DNA and computes a precise enrichment and significance. By Rozowsky J et al. C/Perl.
- * [QuEST](#) - Quantitative Enrichment of Sequence Tags. Sidow and Myers Labs at Stanford. From the 2008 publication [Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data](#). (C++)
- * [SISSRs](#) - Site Identification from Short Sequence Reads. BED file input. Raja Jothi @ NIH. Perl.

**See also [this thread](#) for ChIP-Seq, until I get time to update this list.

Alternate Base Calling

- * [Rolexa](#) - R-based framework for base calling of Solexa data. Project [publication](#)
- * [Alta-cyclic](#) - "a novel Illumina Genome-Analyzer (Solexa) base caller"

Transcriptomics

- * [ERANGE](#) - Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq. Supports Bowtie, BLAT and ELAND. From the Wold lab.
- * [G-Mo.R-Se](#) - G-Mo.R-Se is a method aimed at using RNA-Seq short reads to build de novo gene models. First, candidate exons are built directly from the positions of the reads mapped on the genome (without any ab initio assembly of the reads), and all the possible splice junctions between those exons are tested against unmapped reads. From CNS in France.
- * [MapNext](#) - MapNext: A software tool for spliced and unspliced alignments and SNP detection of short sequence reads. From the Evolutionary Genomics Lab at Sun-Yat Sen University, China.
- * [QPAlma](#) - Optimal Spliced Alignments of Short Sequence Reads. Authors are Fabio De Bona, Stephan Ossowski, Korbinian Schneeberger, and Gunnar Rätsch. A paper is [available](#).
- * [RSAT](#) - RSAT: RNA-Seq Analysis Tools. RNASAT is developed and maintained by Hui Jiang at Stanford University.
- * [TopHat](#) - TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons. TopHat is a collaborative effort between the University of Maryland and the University of California, Berkeley

last edited 16 Jun 09, Sci_guy

Last edited by ECO; 12-28-2009 at 05:45 PM. Reason: Add link to wiki

The screenshot shows the Galaxy web interface for the Netherlands Bioinformatics Center. The main content area displays the 'Upload File (version 1.1.3)' tool configuration page. The 'File Format' dropdown is set to 'Auto-detect'. The 'File' field is empty, with a 'Browse...' button. A tip indicates that files larger than 2GB may fail due to browser limitations. The 'URL/Text' field is also empty. Below this, there is a table for 'Files uploaded via FTP' which is currently empty. The 'Convert spaces to tabs' option is unchecked. The 'Genome' dropdown is set to 'Arabidopsis thaliana (tair9)'. An 'Execute' button is visible at the bottom of the tool configuration area.

Tools (left sidebar):

- Get Data
- Send Data
- Lift-Over
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- Statistics
- Wavelet Analysis
- Graph/Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Evolution
- Motif Tools
- Metagenomic analyses
- FASTA manipulation
- NGS: QC and manipulation
- NGS: Mapping
- NGS: Indel Analysis
- NGS: SV/CNV Analysis
- NGS: RNA Analysis
- NGS: SAM Tools
- NGS: Peak Calling
- NGS: Simulation
- SNP/WGA: Data; Filters
- SNP/WGA: QC; LD; Plots
- SNP/WGA: Statistical Models
- Human Genome Variation
- Genome Diversity
- NGS: VCF Tools
- NGS: Bedtools
- NGS Taskforce: Hubrecht - Alignment tool benchmarking
- NGS Taskforce: WUR denovo benchmarking
- NGS Taskforce: LUMC - GAPSS

History (right sidebar):

0 bytes

Your history is empty. Click 'Get Data' on the left pane to start.

Upload File (version 1.1.3)

File Format:
Auto-detect
Which format? See help below

File:
[Empty field] [Browse...]

TIP: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail. To upload large files, use the URL method (below) or FTP (if enabled by the site administrator).

URL/Text:
[Empty text area]

Here you may specify a list of URLs (one per line) or paste the contents of a file.

Files uploaded via FTP:

File	Size	Date
<i>Please create or log in to a Galaxy account to view files uploaded via FTP.</i>		

This Galaxy server allows you to upload files via FTP. To upload some files, log in to the FTP server at galaxy.nbic.nl using your Galaxy credentials (email address and password).

Convert spaces to tabs:
 Yes
Use this option if you are entering intervals by hand.

Genome:
Arabidopsis thaliana (tair9)

[Execute]

Auto-detect
The system will attempt to detect Axt, Fasta, Fastqsolexa, Gff, Gff3, Html, Lav, Maf, Tabular, Wiggle, Bed and Interval (Bed with headers) formats. If your file is not detected properly as one of the known formats, it most likely means that it has some format problems (e.g., different number of columns on different rows). You can still coerce the system to set your data to the format you think it should be. You can also upload compressed files, which will automatically be decompressed.

Ab1
A binary sequence file in 'ab1' format with a '.ab1' file extension. You must manually select this 'File Format' when uploading the file.

Axt
blastz pairwise alignment format. Each alignment block in an axt file contains three lines: a summary line and 2 sequence lines. Blocks are separated from one another by blank lines. The summary line contains chromosomal position and size information about the alignment. It consists of 9 required fields.

Ram

Considerations

164/165

- Can Next Generation Sequencing solve my problem ?
- What application do I need (de novo, RNA, amplicon, ...) ?
- What is the best platform to run it on ? (capacity, price, speed, accuracy, read length...)
- What is your experimental design ?
- What about bioinformatics ?
- Are your results correct ? (XY - XX chromosome for SNP)
- In cancer research: mutation in gene increase “probability” for developing cancer
- What about statistics ?
- 15x coverage is probably not over the whole genome

- Rubbish in = rubbish out (contamination, sample degradation, mixed samples)

- If you don't know, ask and discuss with others.

Thanks for your interest !

<http://users.ugent.be/~avierstr/>

Andy.Vierstraete@ugent.be

CTAGGTAGCTAGTCG
GCTLIFECISGATAG
C4-LETTERWORDT
GCTATATCGTAGCTG

WWW.DNA.UGENT.BE

