

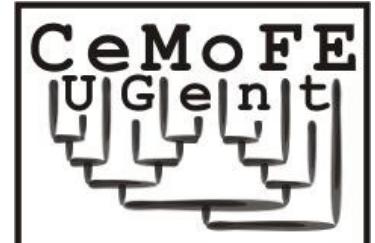
# Next Generation Sequencing for **DUMMIES**

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

3/165

- 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

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. FIERS

*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  $^3\text{H}$ . 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

firmed each other. The results, together with the known specificity of the ribonuclease T<sub>1</sub>, which had released the sequence, establish that MS2 RNA ends in ...GpUpUpApCpCpApCpCpCpA.

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

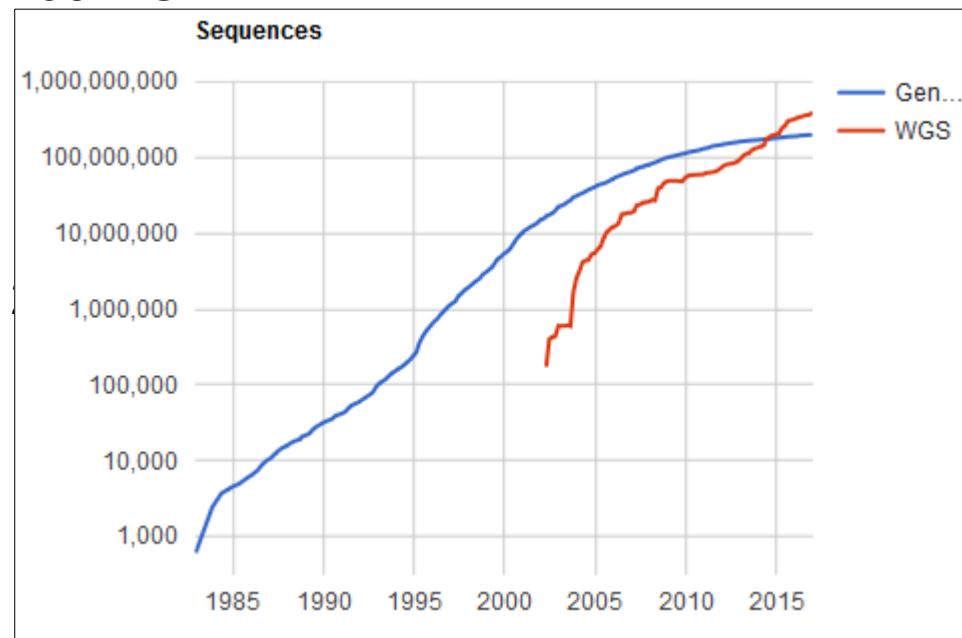
#### 1. Introduction

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

- 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



Release	Date	GenBank		WGS	
		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

- 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: 1<sup>st</sup> 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: 1<sup>st</sup> 454 Life Sciences Next Generation Sequencing system: GS 20 system († mid 2016)
- 2006: 1<sup>st</sup> Solexa Next Generation Sequencer: Genome Analyzer (Illumina)
- 2007: 1<sup>st</sup> Applied Biosystems Next Generation Sequencer: SOLiD († Dec 2017)
- 2009: 1<sup>st</sup> Helicos **single molecule** sequencer: Helicos Genetic Analyser System († Nov 2012)
- 2011: 1<sup>st</sup> Ion Torrent Next Generation Sequencer: PGM
- 1<sup>st</sup> 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: 1<sup>st</sup> BGI Next Generation Sequencer: BGISEQ-500 (sold in China only)
- 2016: 1<sup>st</sup> Oxford Nanopore Technologies sequencer: MinION
- 2017: SeqLL announces tSMS sequencer: **single molecule** (Helicos technology)

## Different platforms

7/165

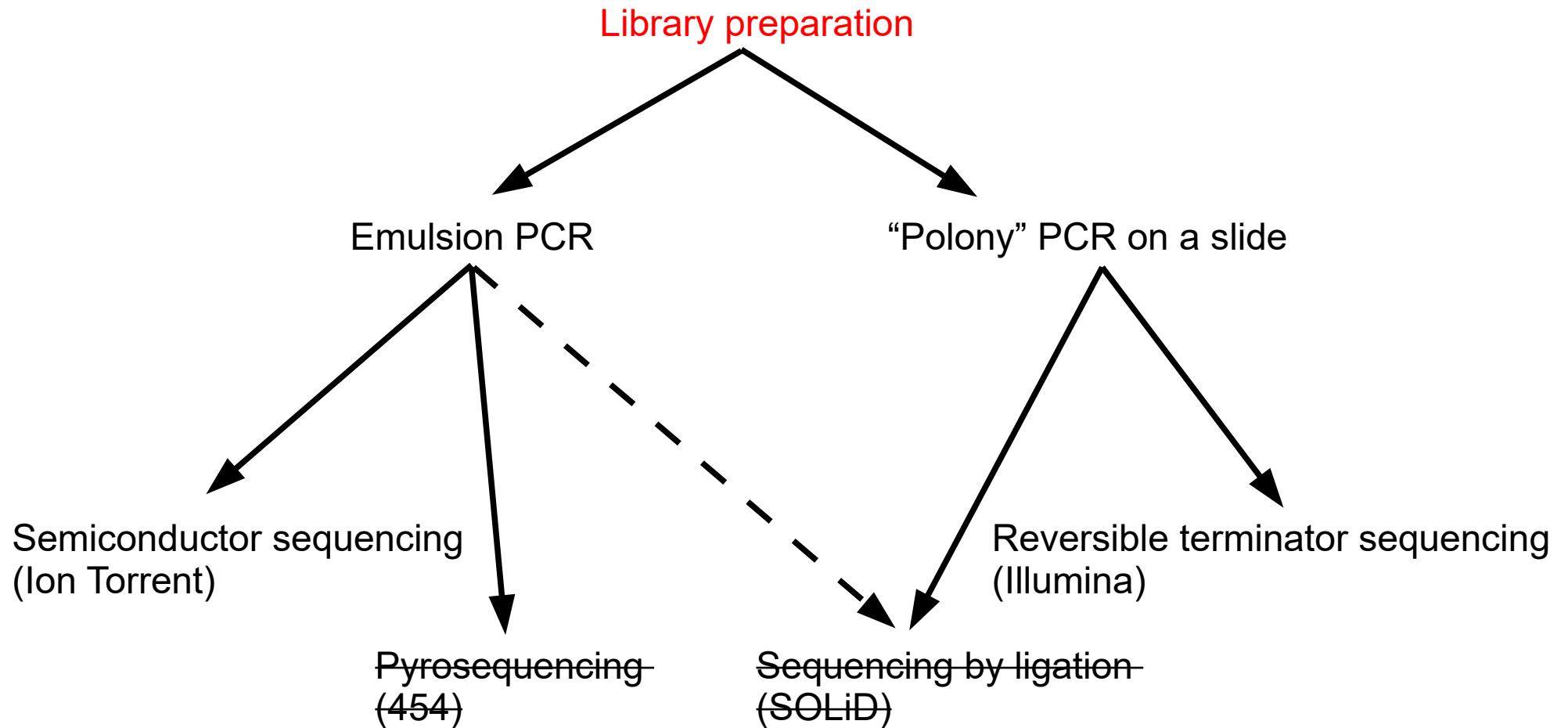
- Illumina (Solexa)
  - iSeq 100
  - MiniSeq
  - MiSeq
  - NextSeq 500 - 550
  - HiSeq 2500 - 3000 – 4000
  - HiSeq X Five – Ten
  - NovaSeq 5000 - 6000
- Thermo Fisher Scientific (Applied Biosystems > Life Technologies)
  - Ion Torrent Personal Genome Machine (PGM)
  - Ion Torrent GeneStudio S5, S5 Plus, S5 Prime
  - Ion Torrent Proton
- Pacific Biosciences
  - Sequel System
  - PacBio RS II
- Oxford Nanopore Technologies
  - SmidgION
  - MinION
  - GridION X5
  - PromethION
- SeqLL
  - tSMS sequencer

Next Generation Sequencing  
Amplified Single Molecule Sequencing

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

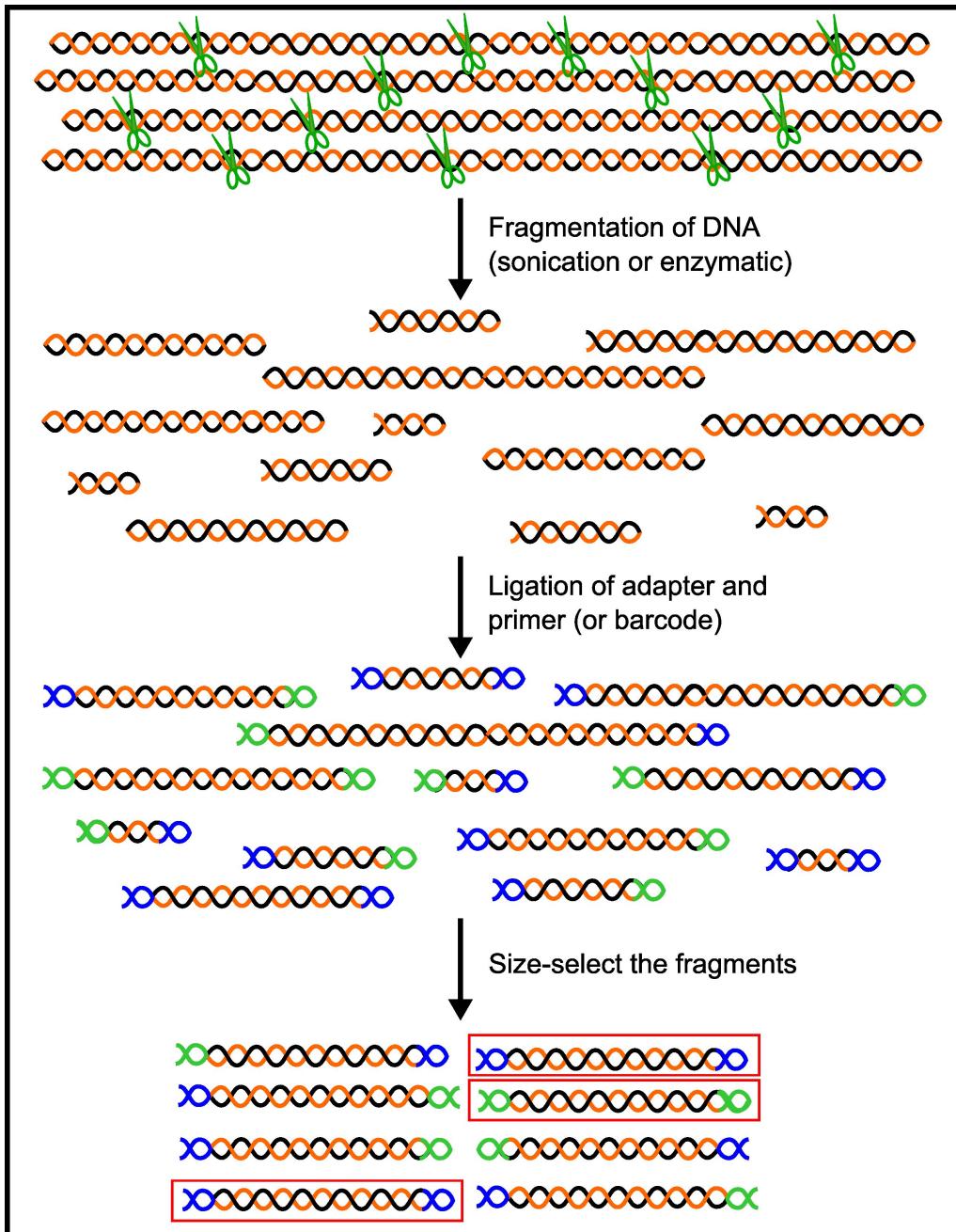
## Workflow

### Next Generation Sequencing: Amplified Single Molecule Sequencing



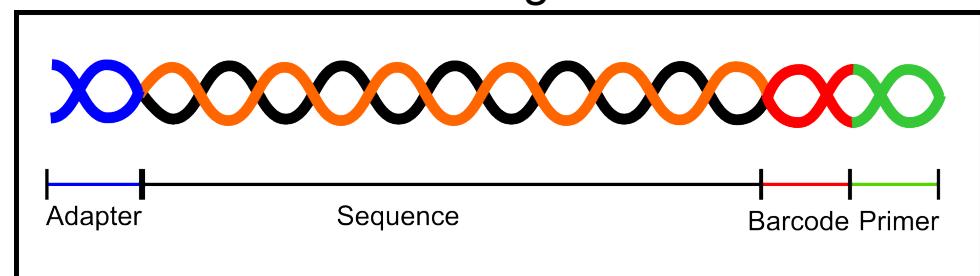
## Workflow

### Next Generation Sequencing: Amplified Single Molecule Sequencing



### Library preparation

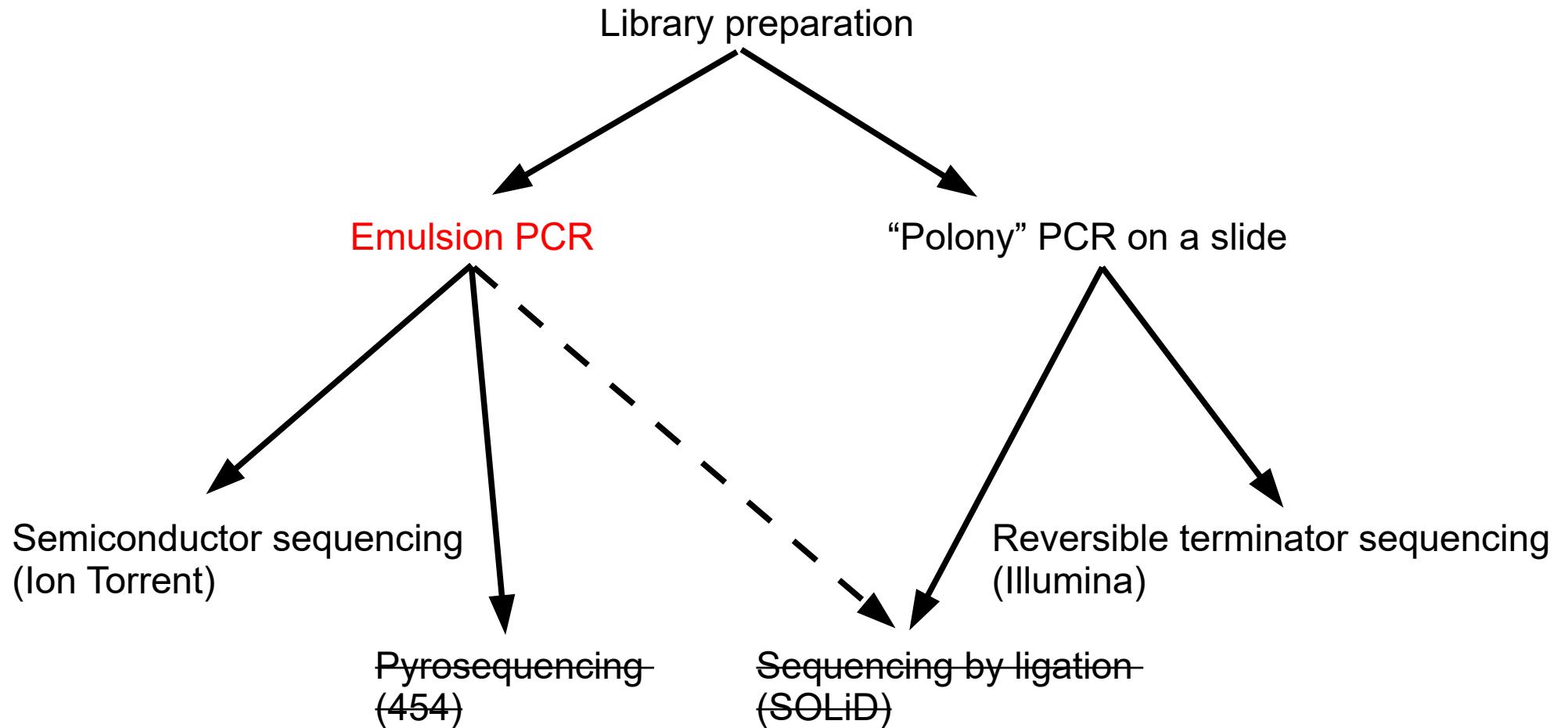
#### Good fragments:



## Workflow

10/165

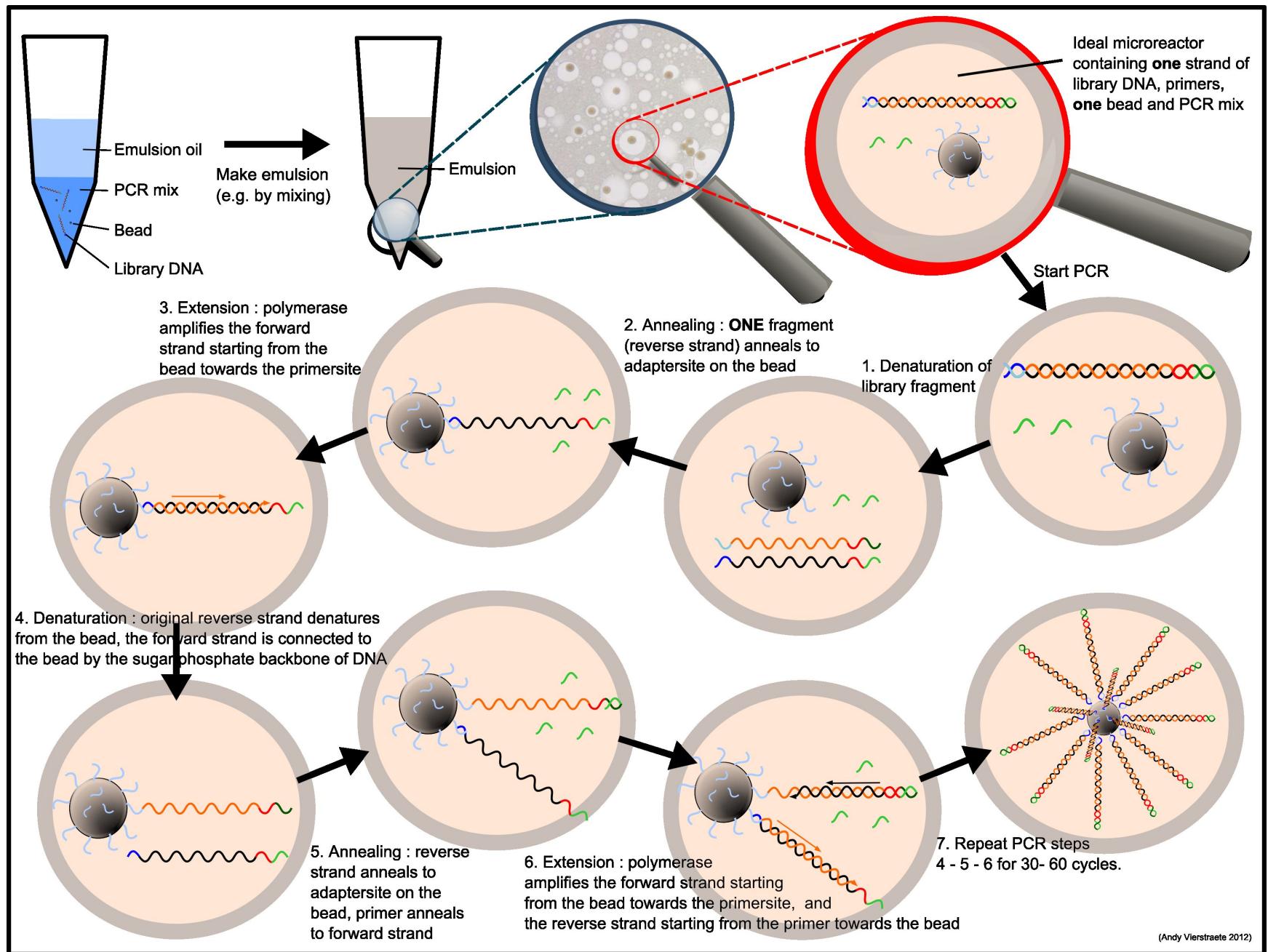
### Next Generation Sequencing: Amplified Single Molecule Sequencing



## Workflow

11/165

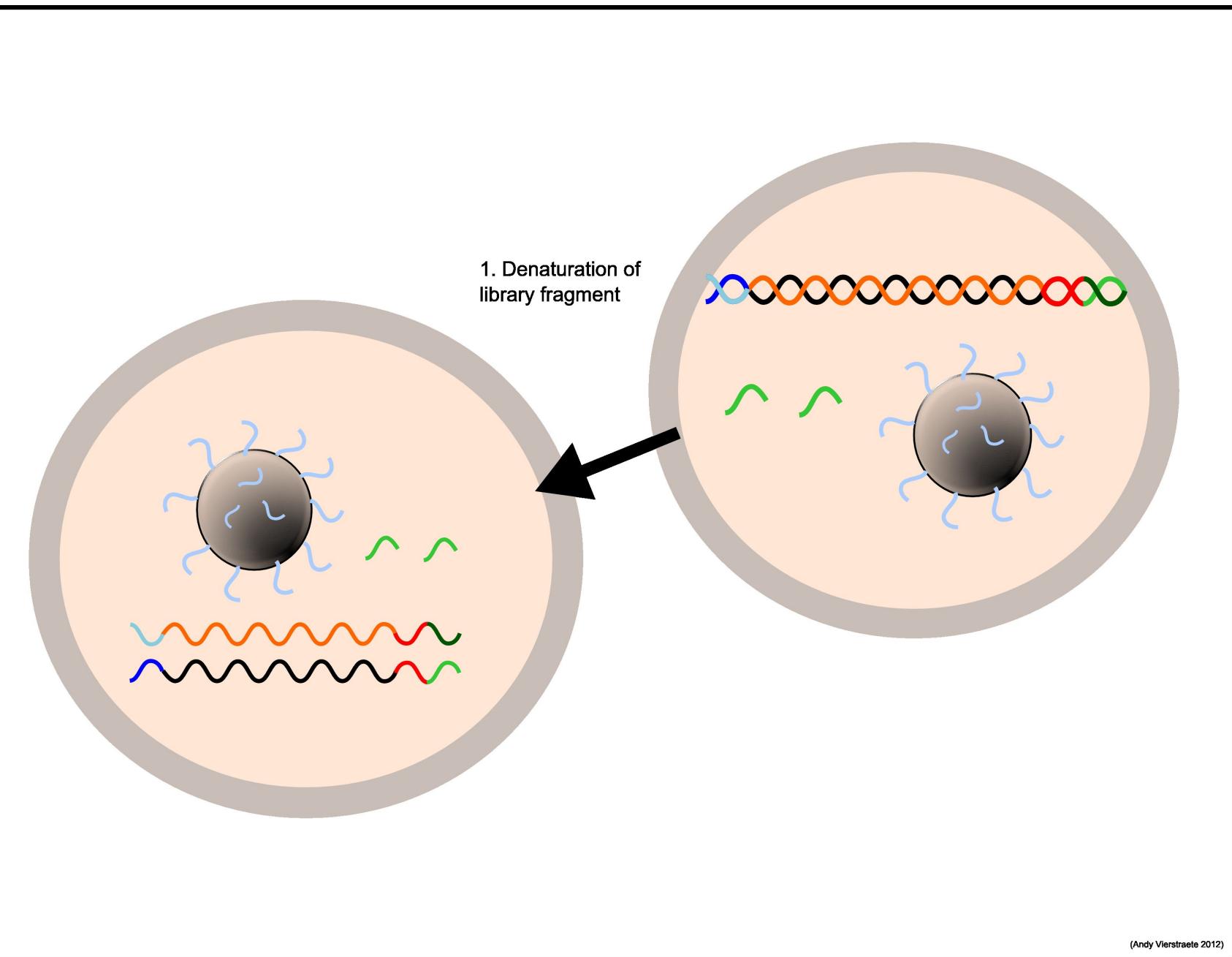
### Next Generation Sequencing: Amplified Single Molecule Sequencing Emulsion PCR



## Workflow

12/165

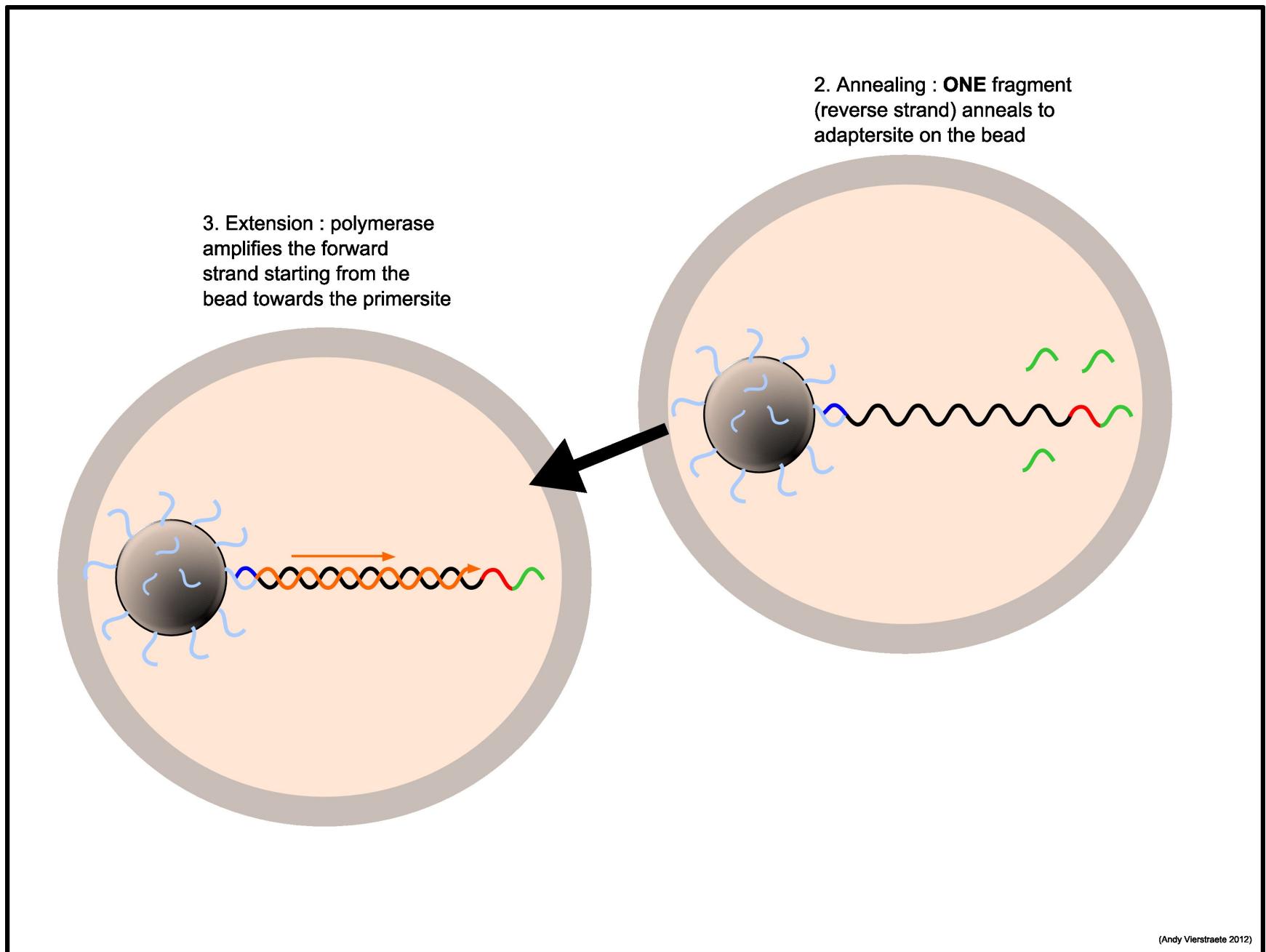
### Next Generation Sequencing: Amplified Single Molecule Sequencing Emulsion PCR



## Workflow

13/165

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

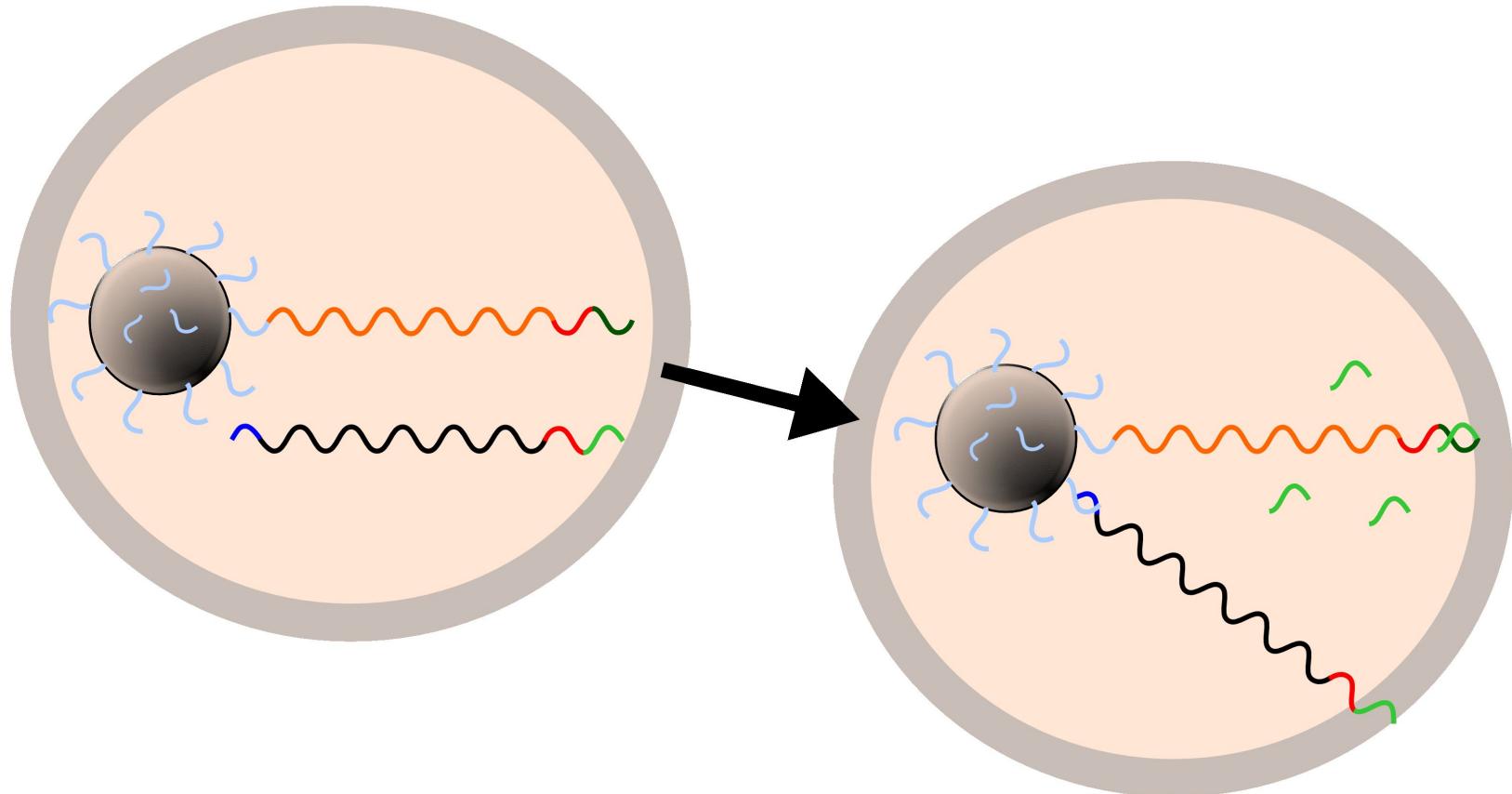


## Workflow

14/165

### 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

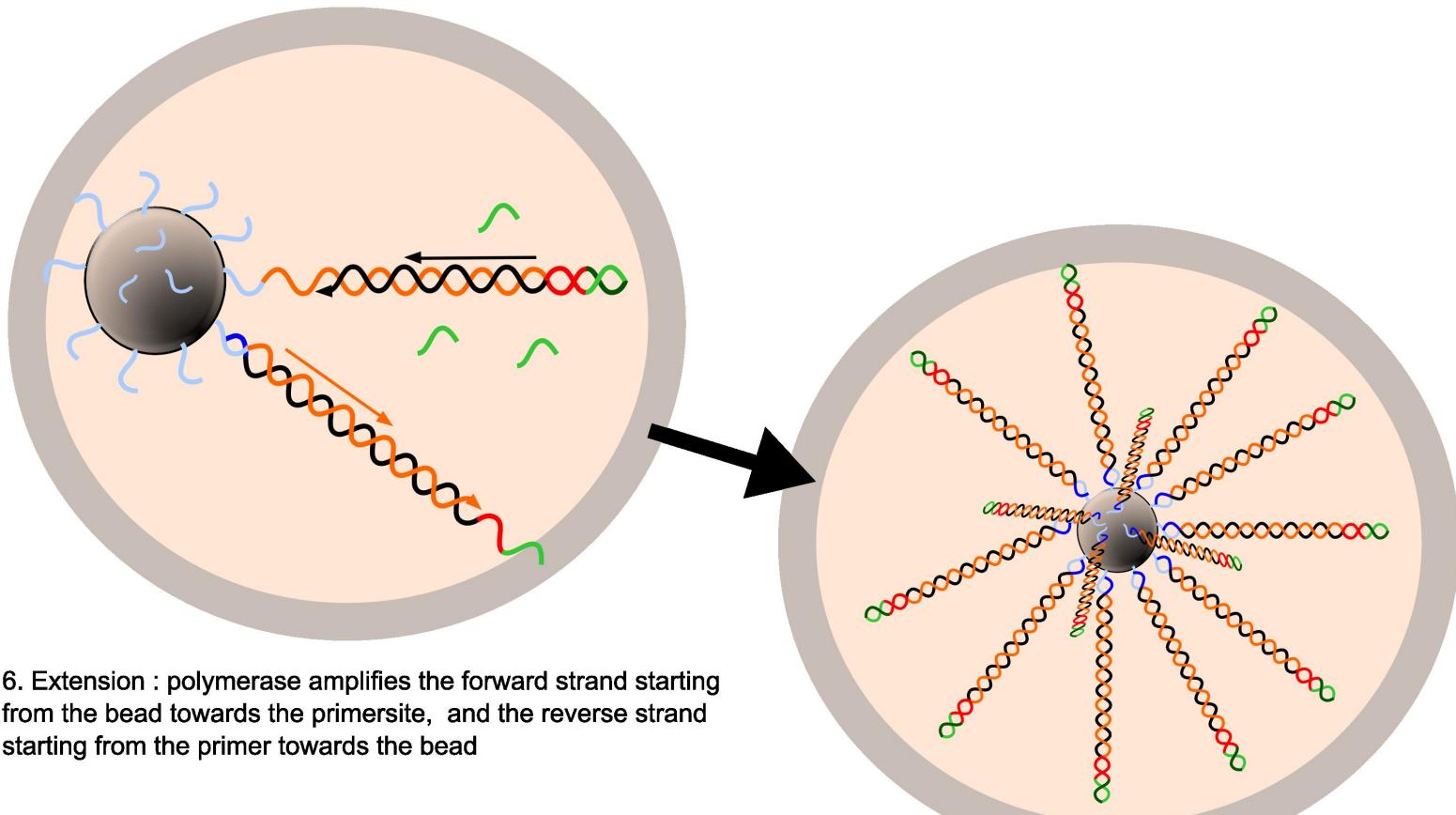


5. Annealing : reverse strand anneals to adaptersite on the bead, primer anneals to forward strand

## Workflow

15/165

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



6. Extension : polymerase amplifies the forward strand starting from the bead towards the primersite, and the reverse strand starting from the primer towards the bead

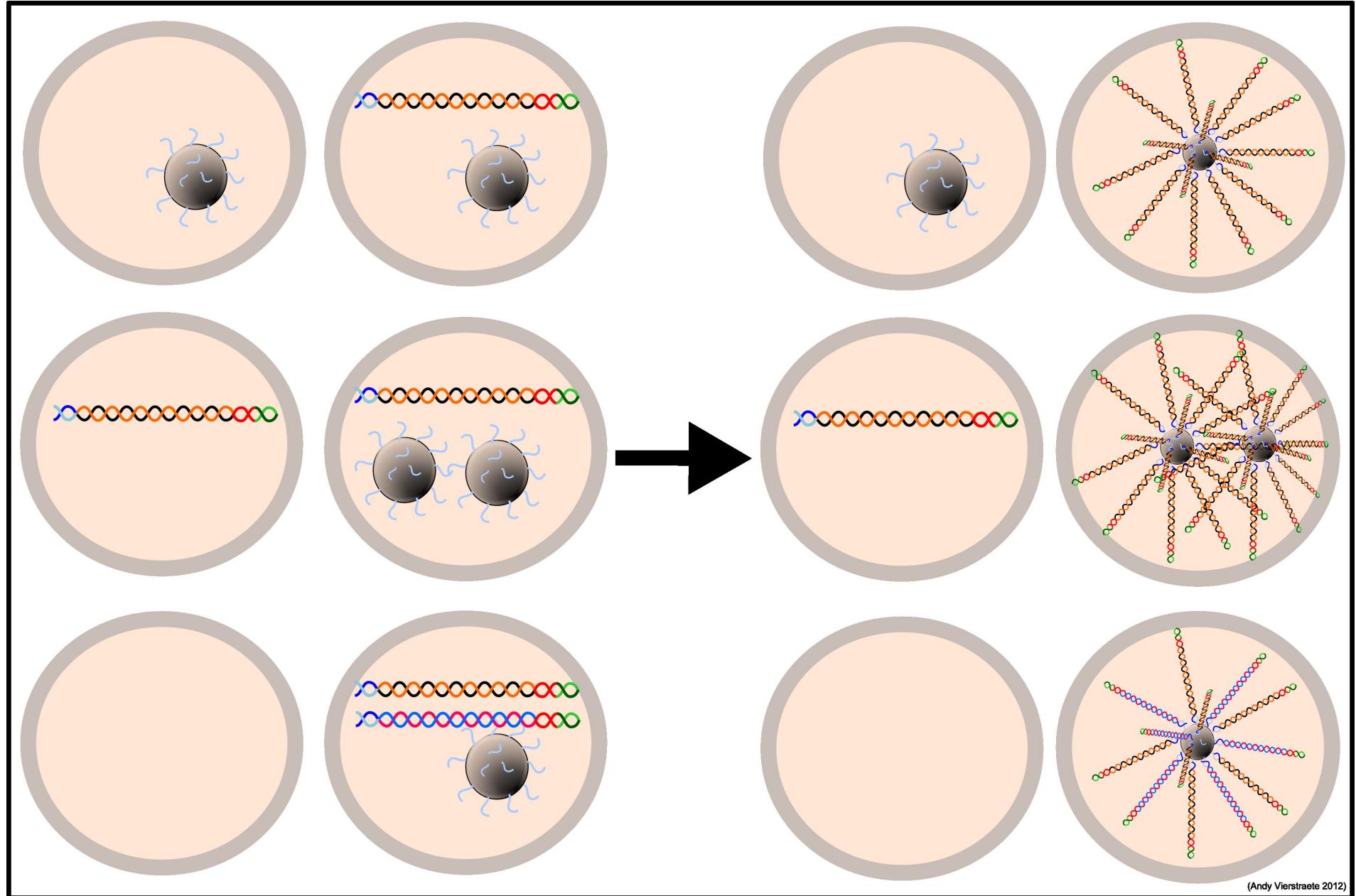
7. Repeat PCR steps  
4 - 5 - 6 for 30- 60 cycles.

# Workflow

16/165

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

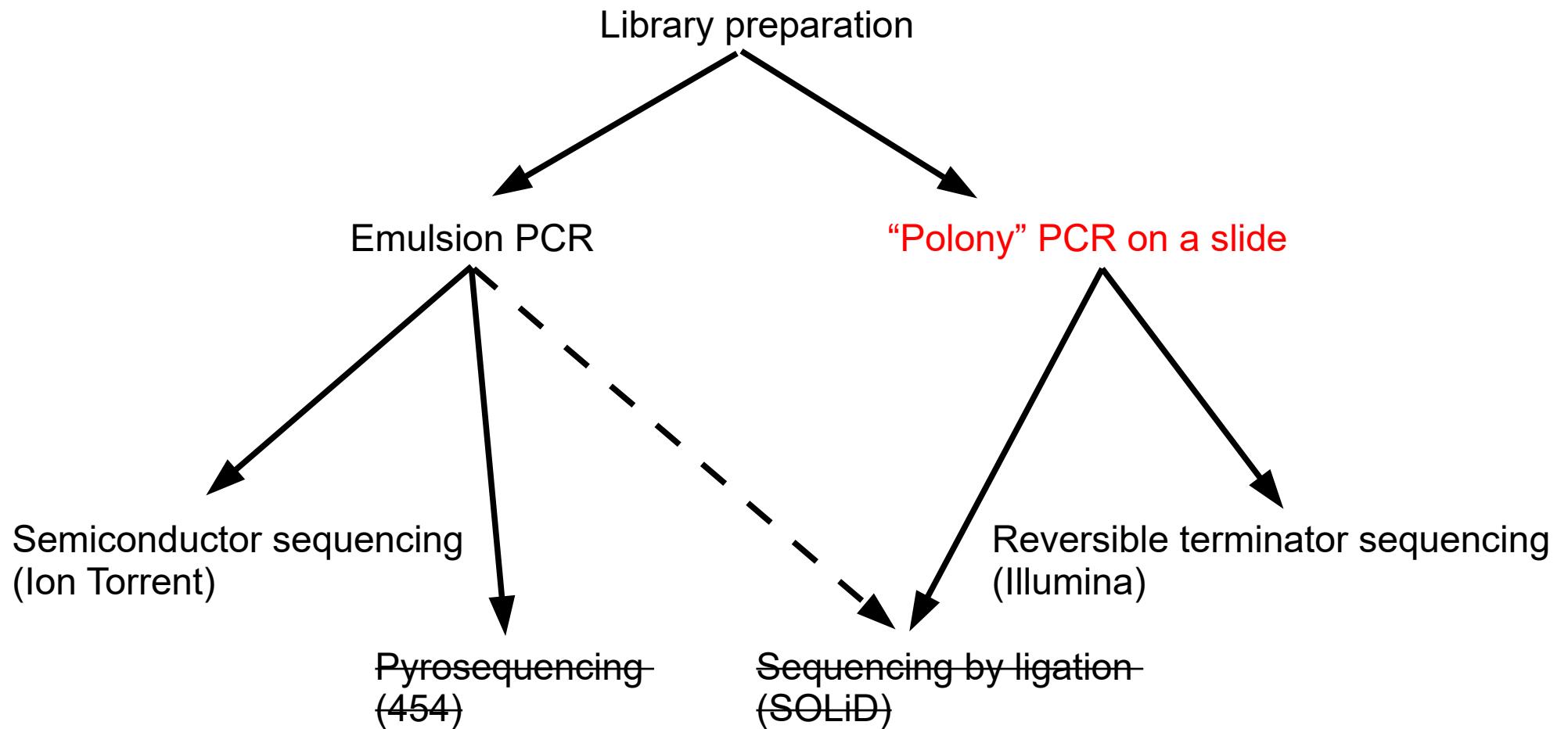
different micro reactors: only 15 % are good ones



## Workflow

17/165

### Next Generation Sequencing: Amplified Single Molecule Sequencing

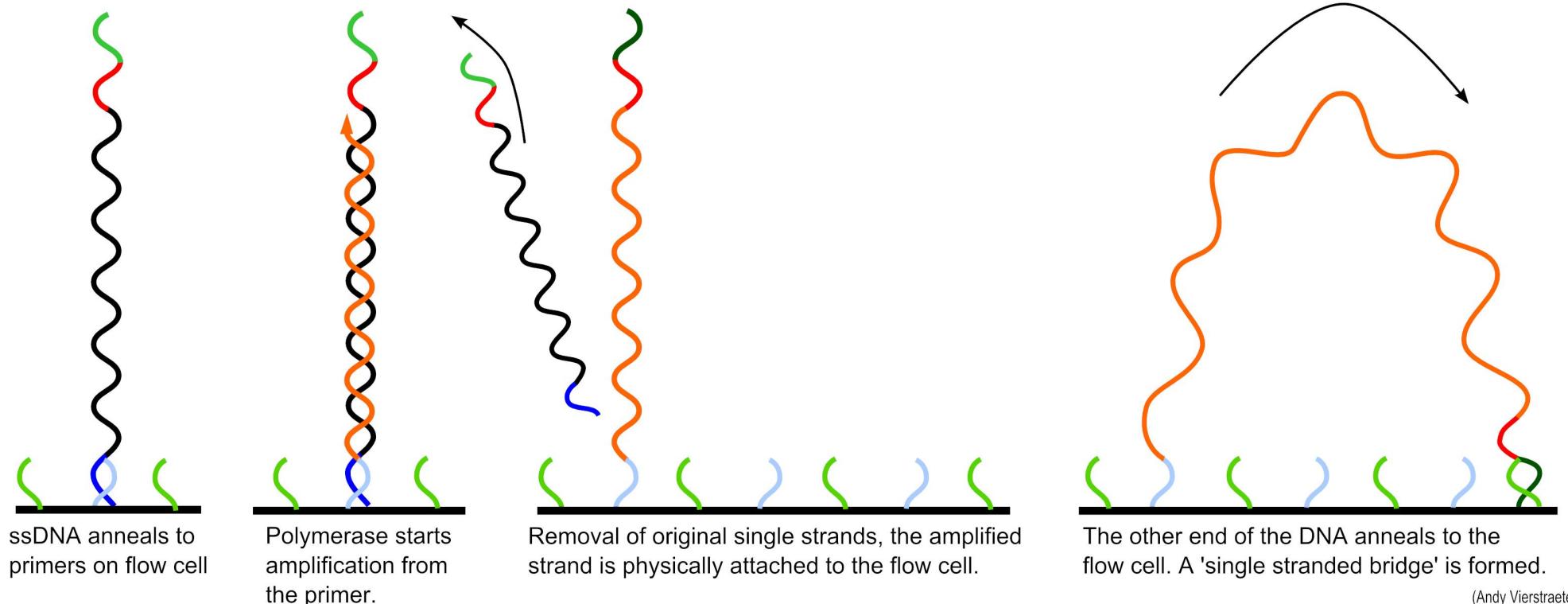


## Workflow

18/165

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

Bridge amplification: Illumina

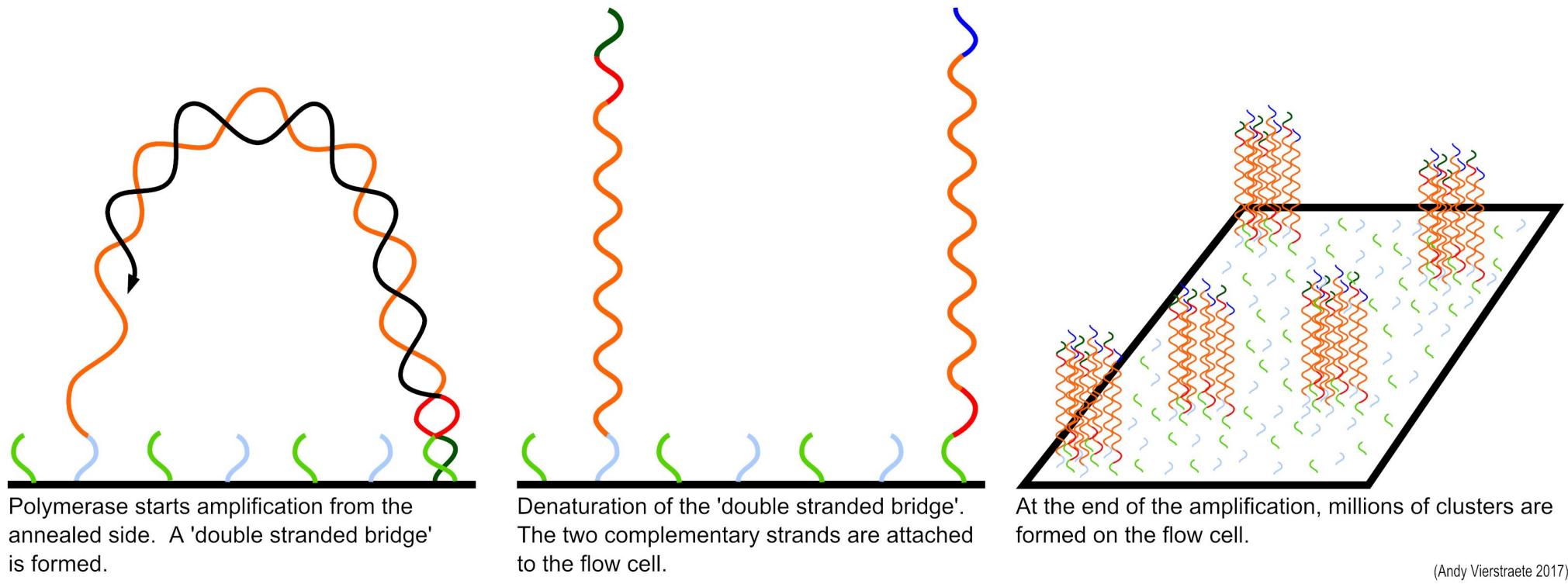


# Workflow

19/165

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

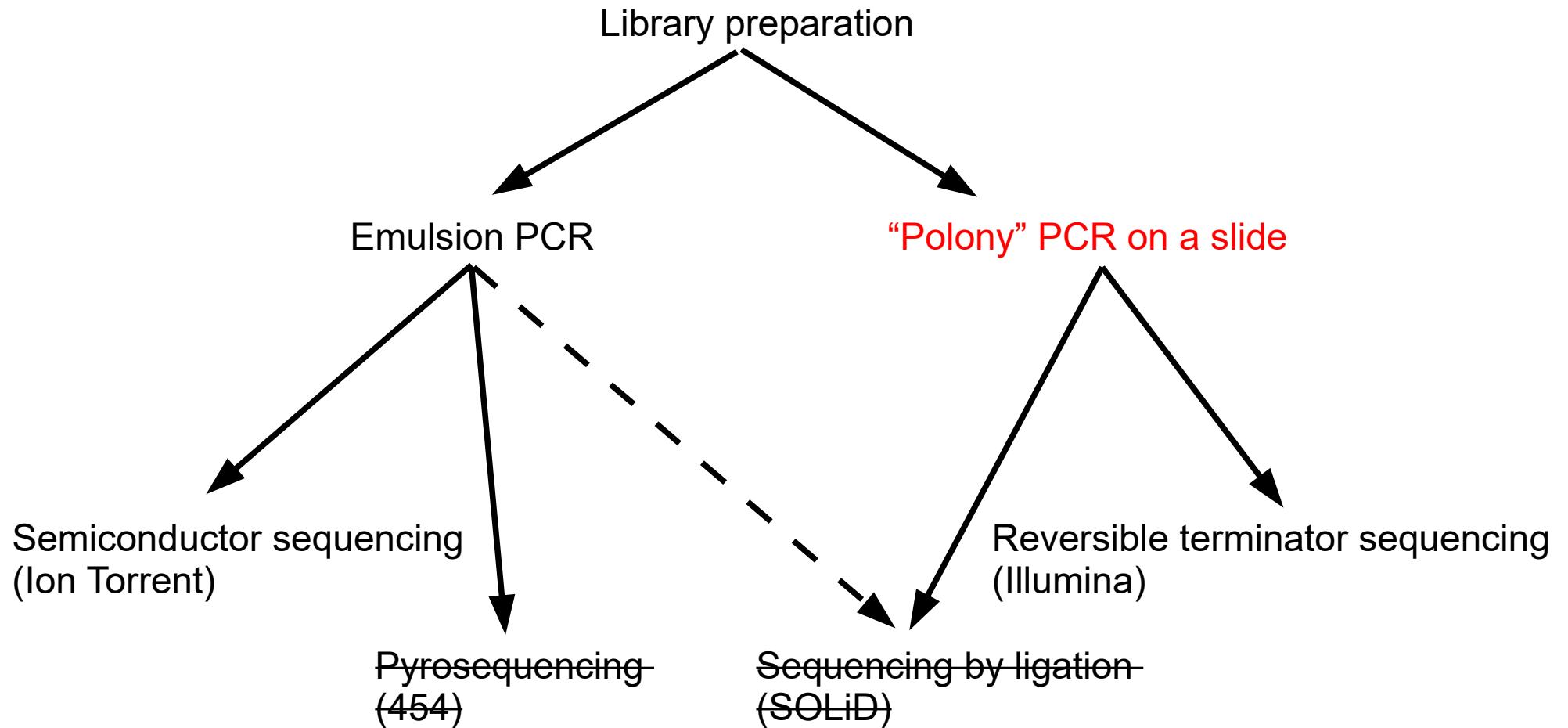
Bridge amplification: Illumina



## Workflow

20/165

### Next Generation Sequencing: Amplified Single Molecule Sequencing

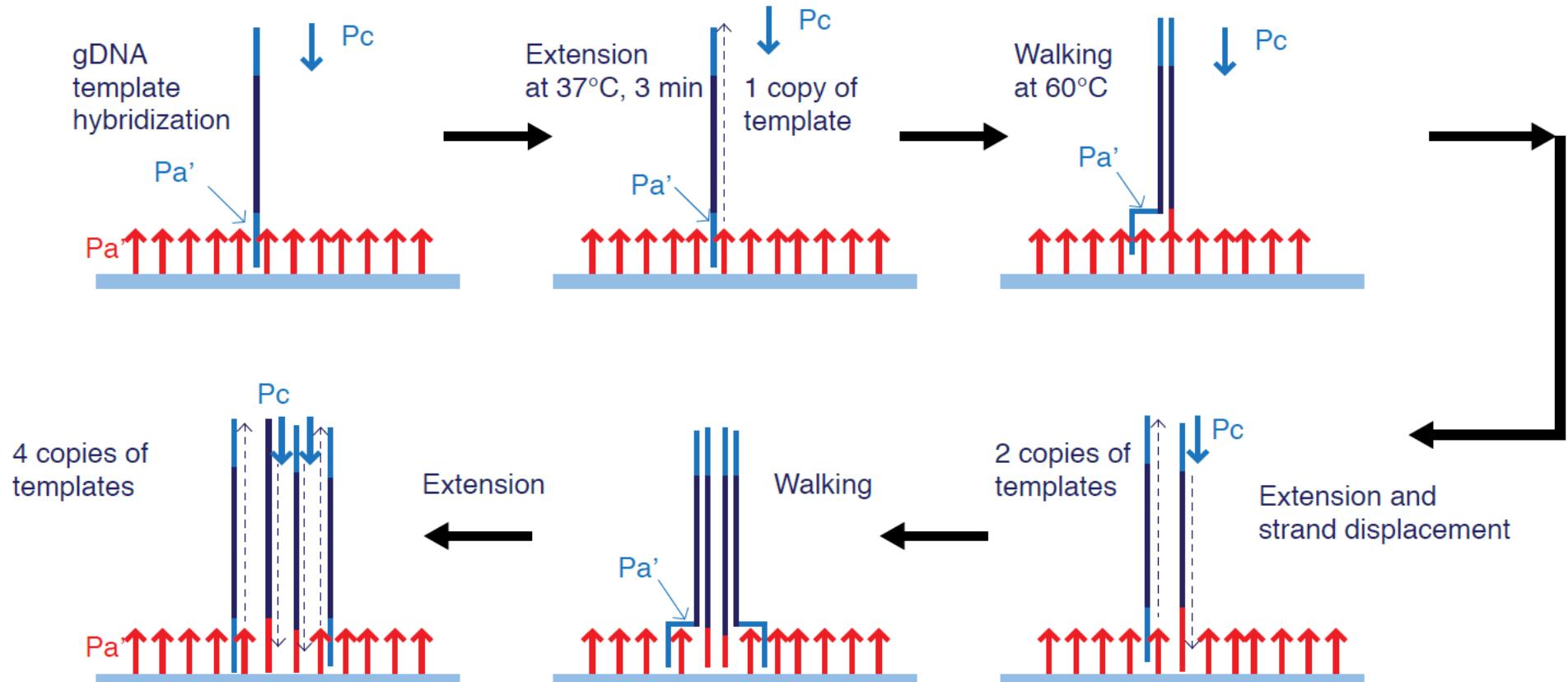


## Workflow

21/165

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

Wildfire amplification: SOLiD

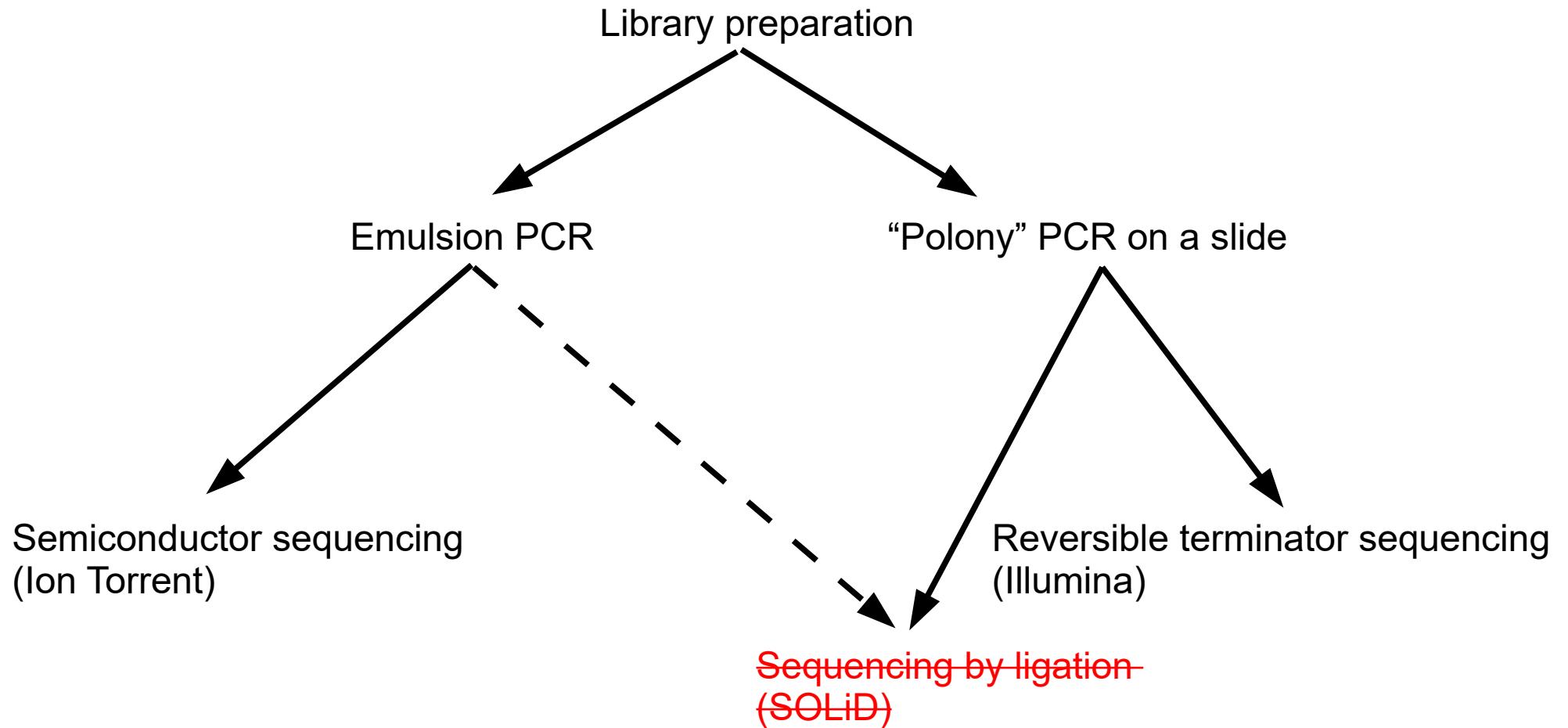


Wildfire chemistry schematic.

## Workflow

22/165

### Next Generation Sequencing: Amplified Single Molecule Sequencing

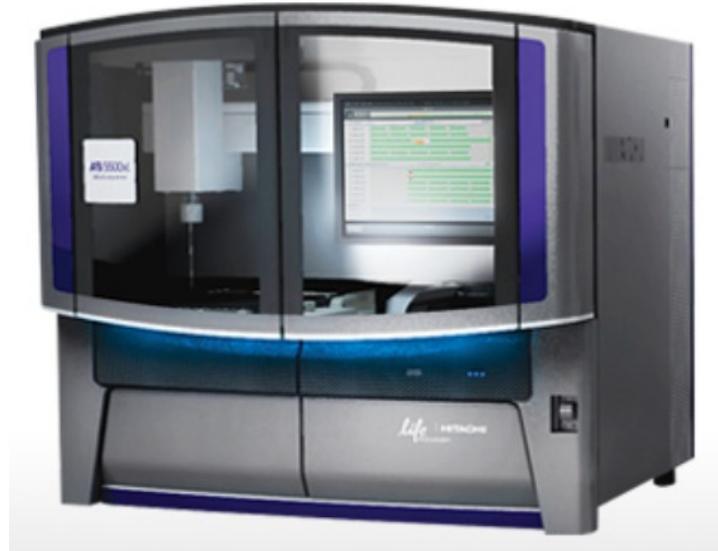


## Different platforms

### Next Generation Sequencing: Amplified Single Molecule Sequencing

#### SOLiD

5500 W SOLiD Sequencer (end of support Dec 2017)



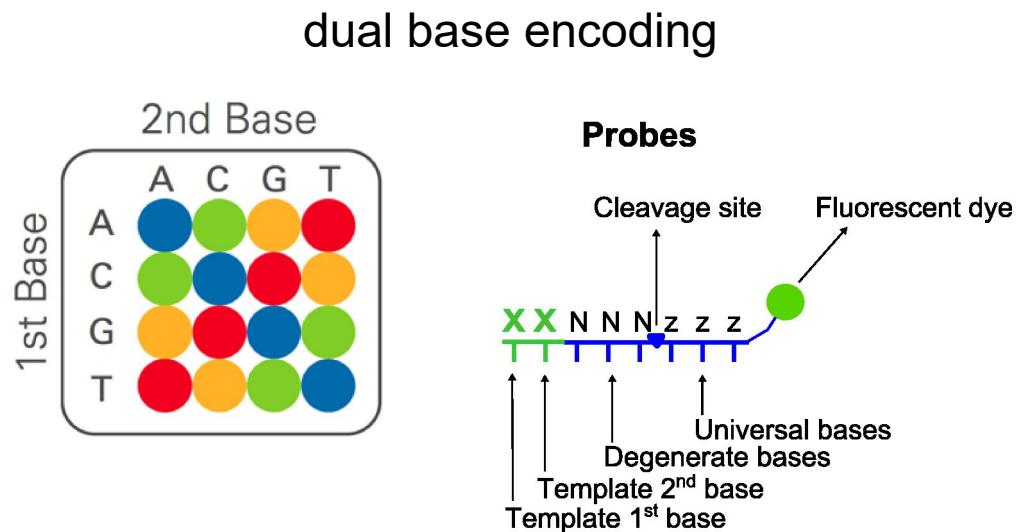
	<b>5500 W</b>	<b>5500 xl W</b>
<b>Read Length</b>	75 bp or 2 x 50 bp	75 bp or 2 x 50 bp
<b>Throughput</b>	120 - 160 Gb	240 - 320 Gb
<b>Reads per run</b>	1,2 Billion	2,4 Billion
<b>Accuracy</b>	99,99 %	99,99 %
<b>Run Time</b>	7 days	7 days

## Different platforms

24/165

### Next Generation Sequencing: Amplified Single Molecule Sequencing SOLiD

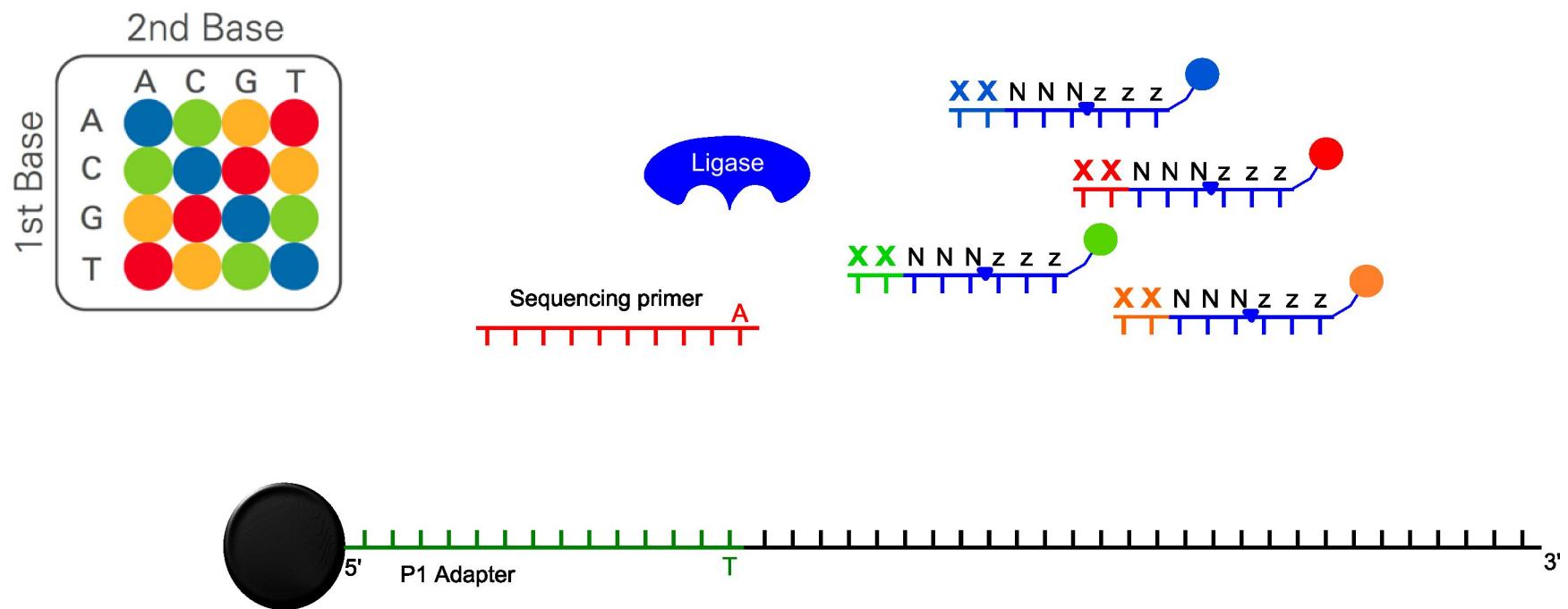
Workflow: Library preparation → Wildfire PCR → Sequencing by Ligation



## Different platforms

25/165

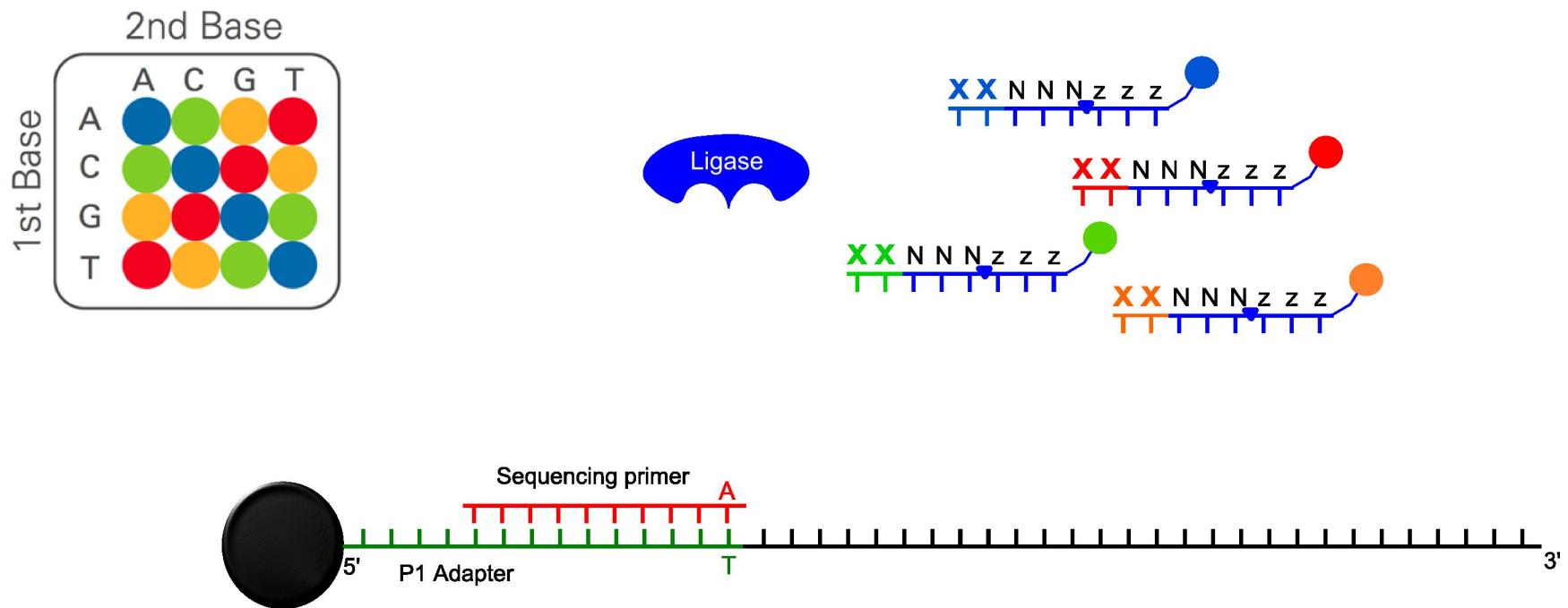
### Next Generation Sequencing: Amplified Single Molecule Sequencing SOLiD



## Different platforms

26/165

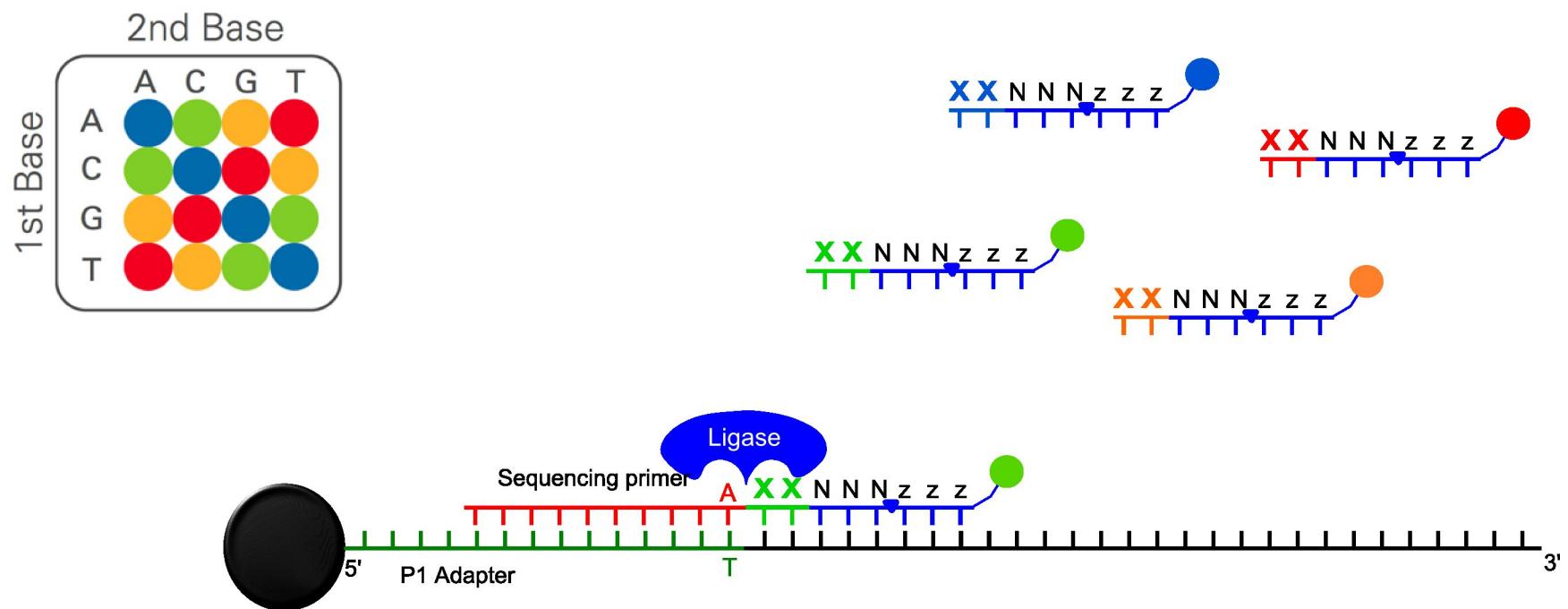
### Next Generation Sequencing: Amplified Single Molecule Sequencing SOLiD



## Different platforms

27/165

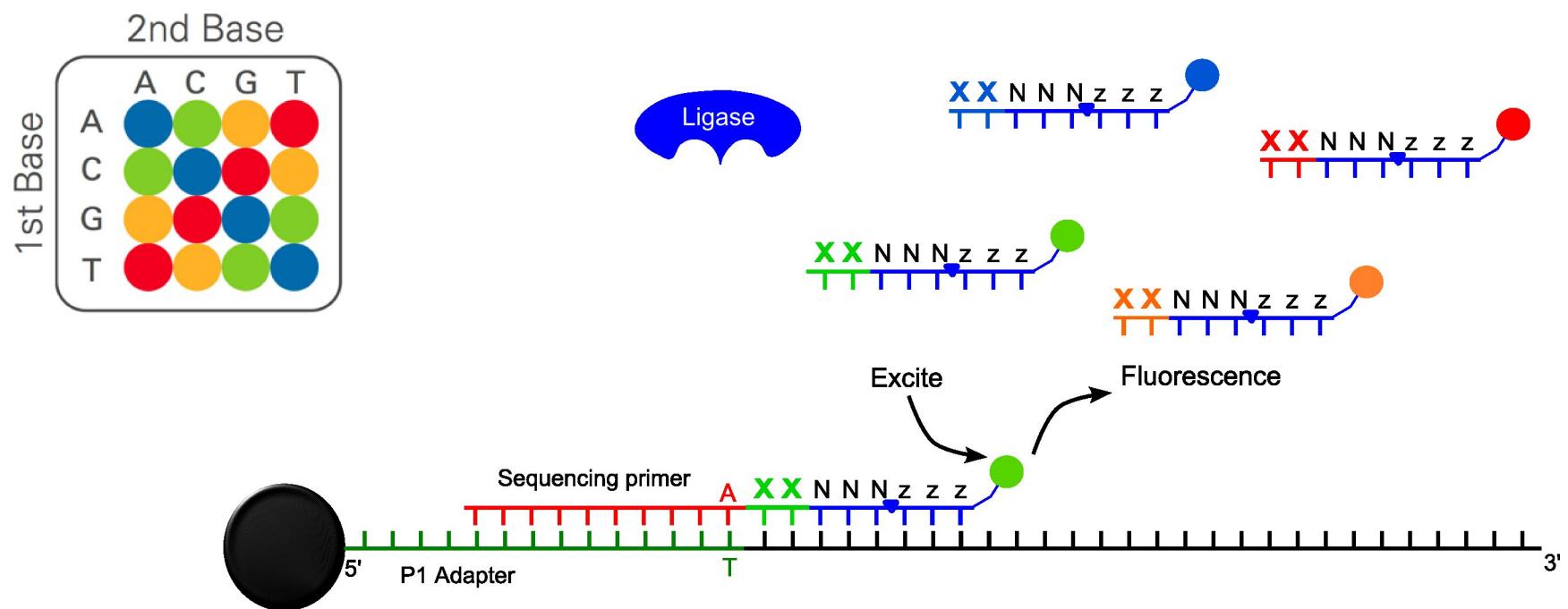
### Next Generation Sequencing: Amplified Single Molecule Sequencing SOLiD



## Different platforms

28/165

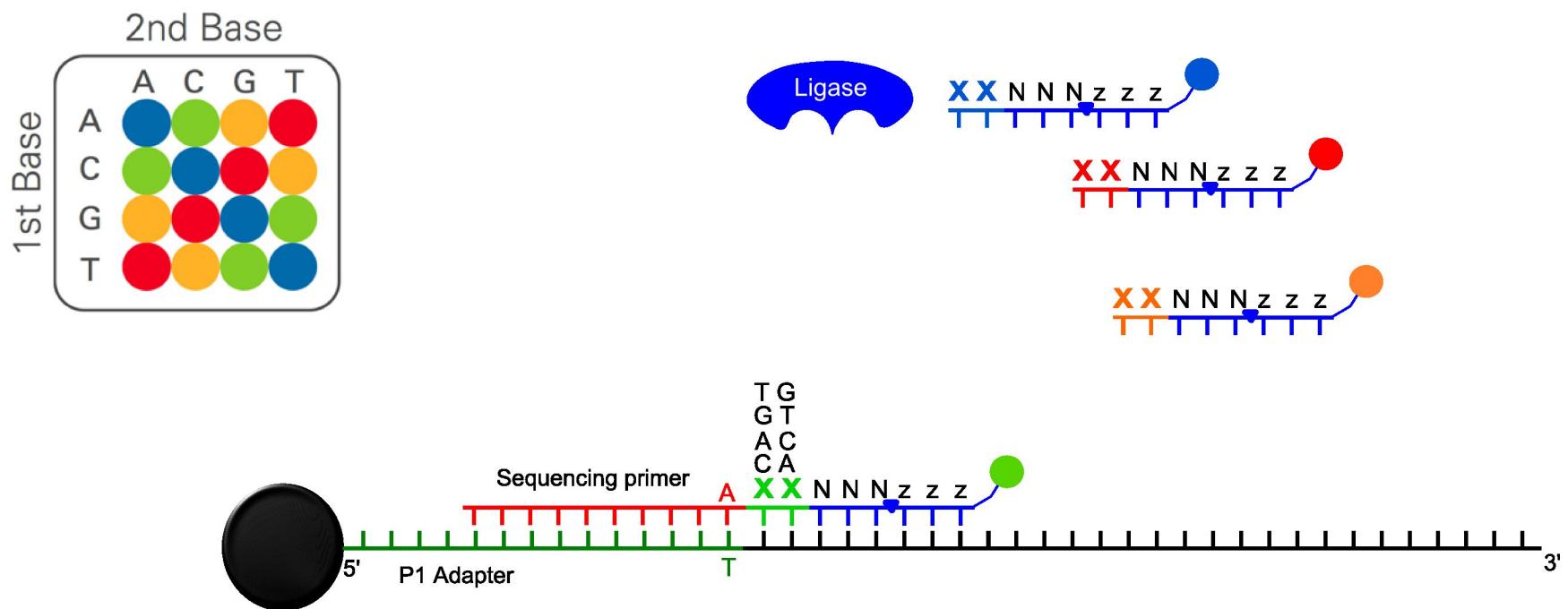
### Next Generation Sequencing: Amplified Single Molecule Sequencing SOLiD



## Different platforms

29/165

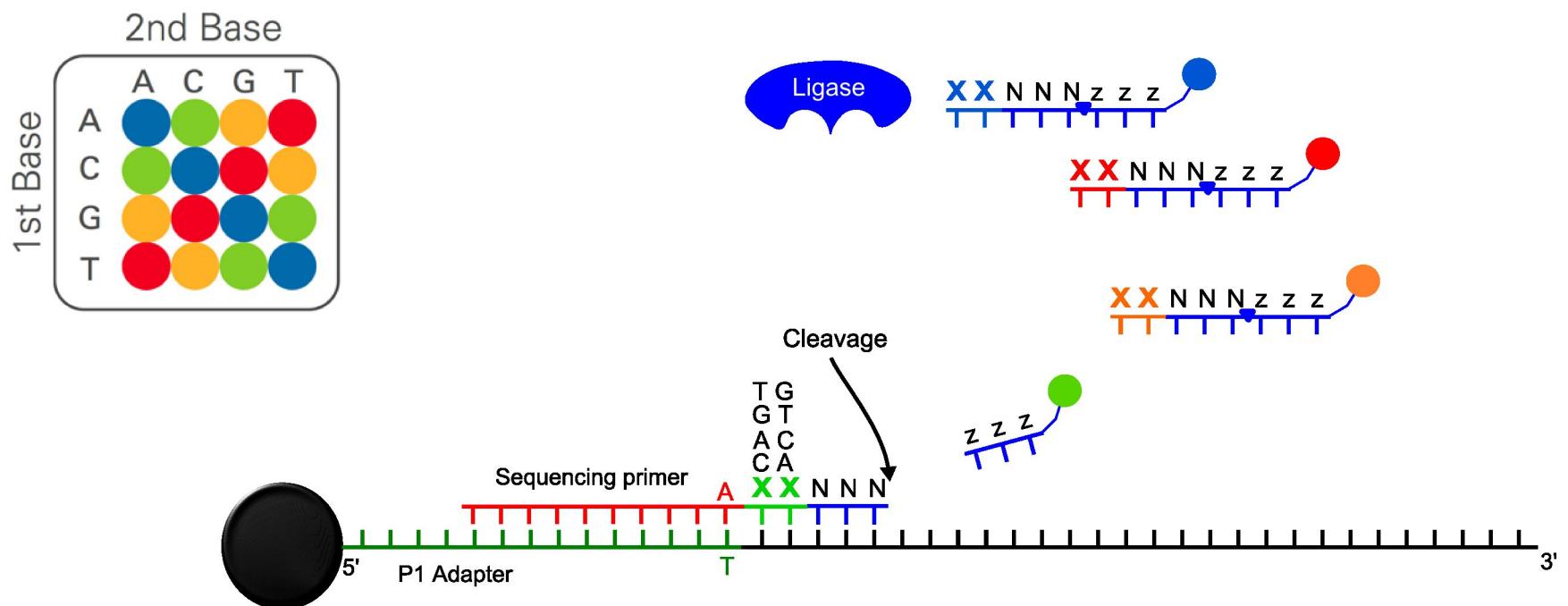
### Next Generation Sequencing: Amplified Single Molecule Sequencing SOLiD



## Different platforms

30/165

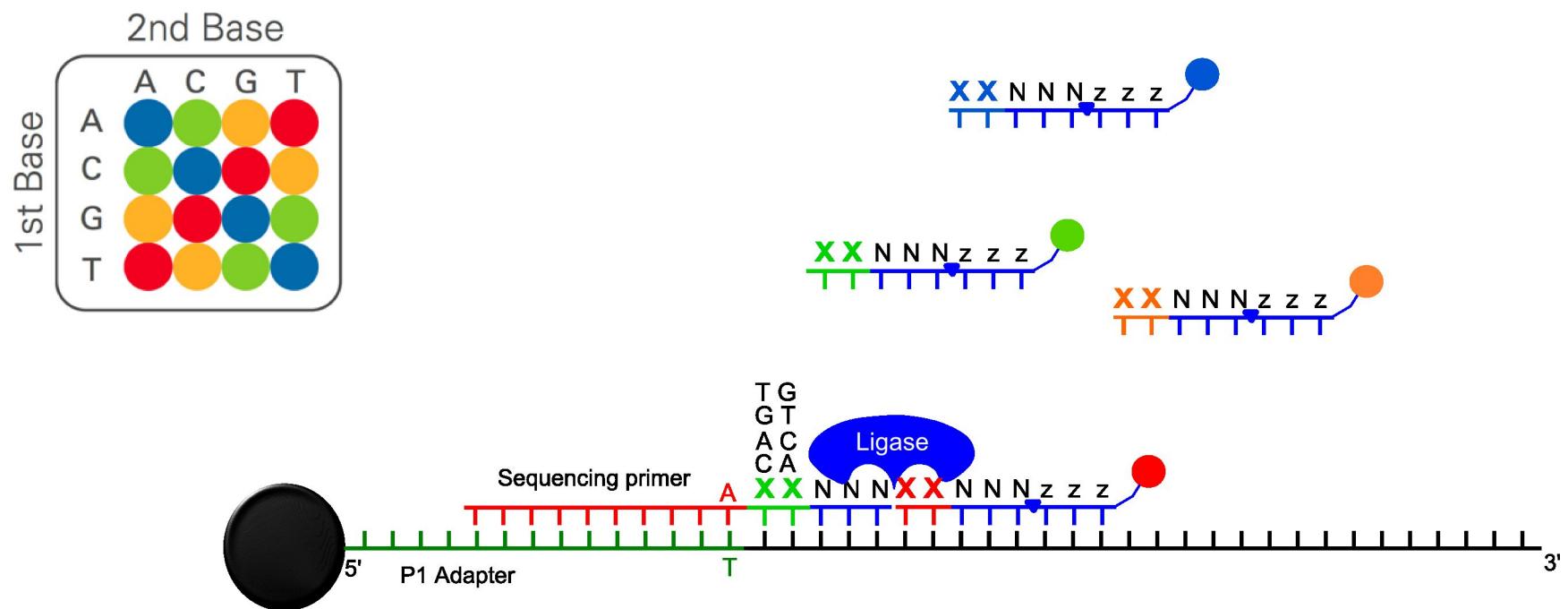
### Next Generation Sequencing: Amplified Single Molecule Sequencing SOLiD



## Different platforms

31/165

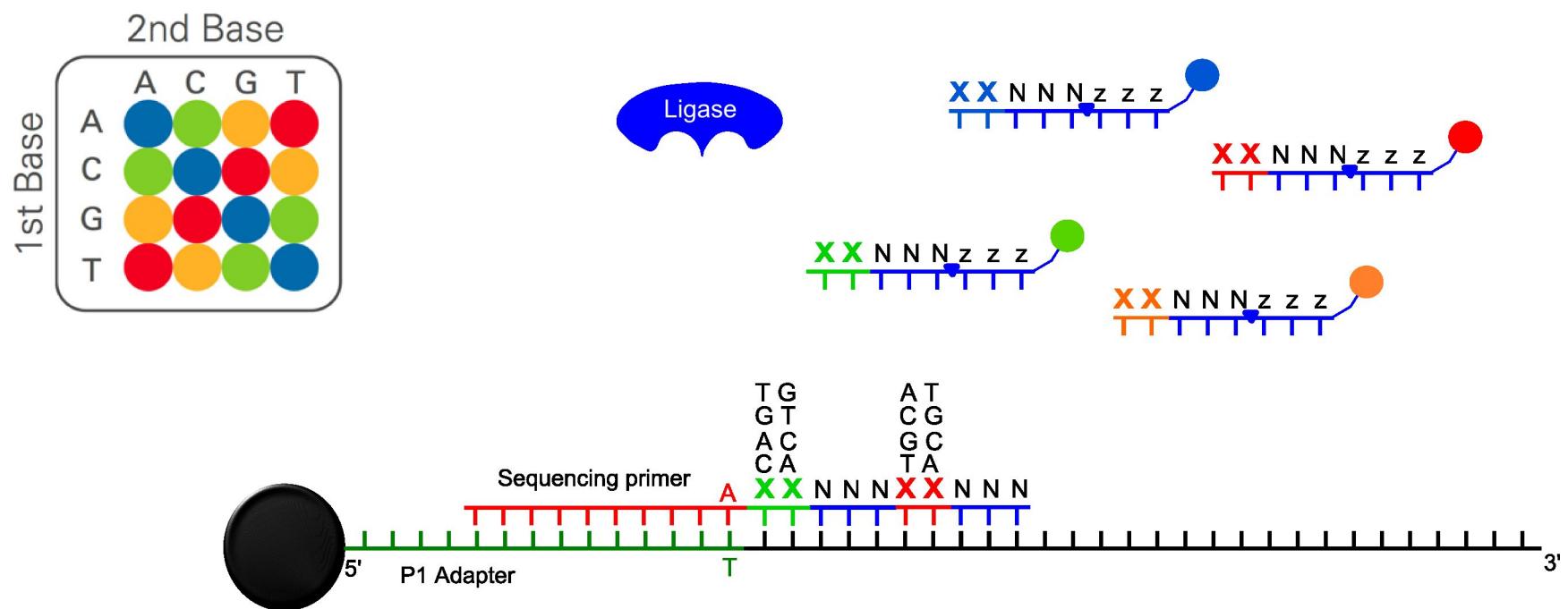
### Next Generation Sequencing: Amplified Single Molecule Sequencing SOLiD



## Different platforms

32/165

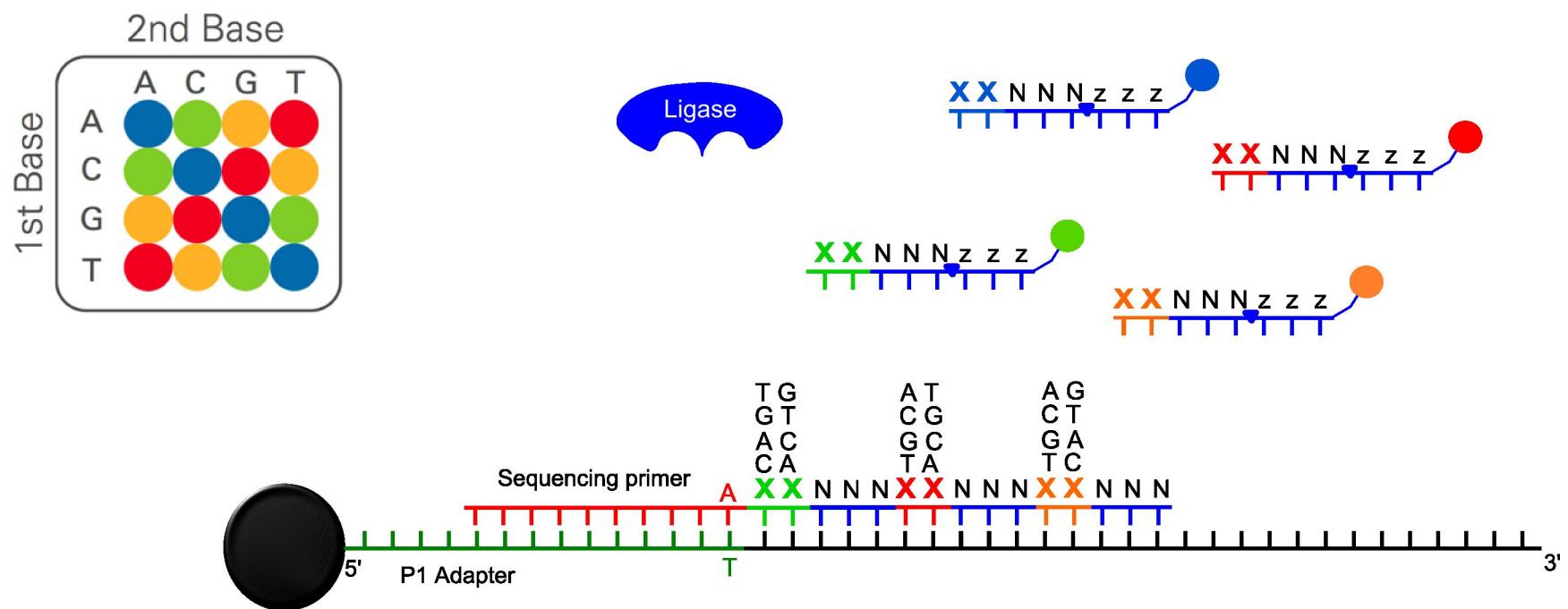
### Next Generation Sequencing: Amplified Single Molecule Sequencing SOLiD



## Different platforms

33/165

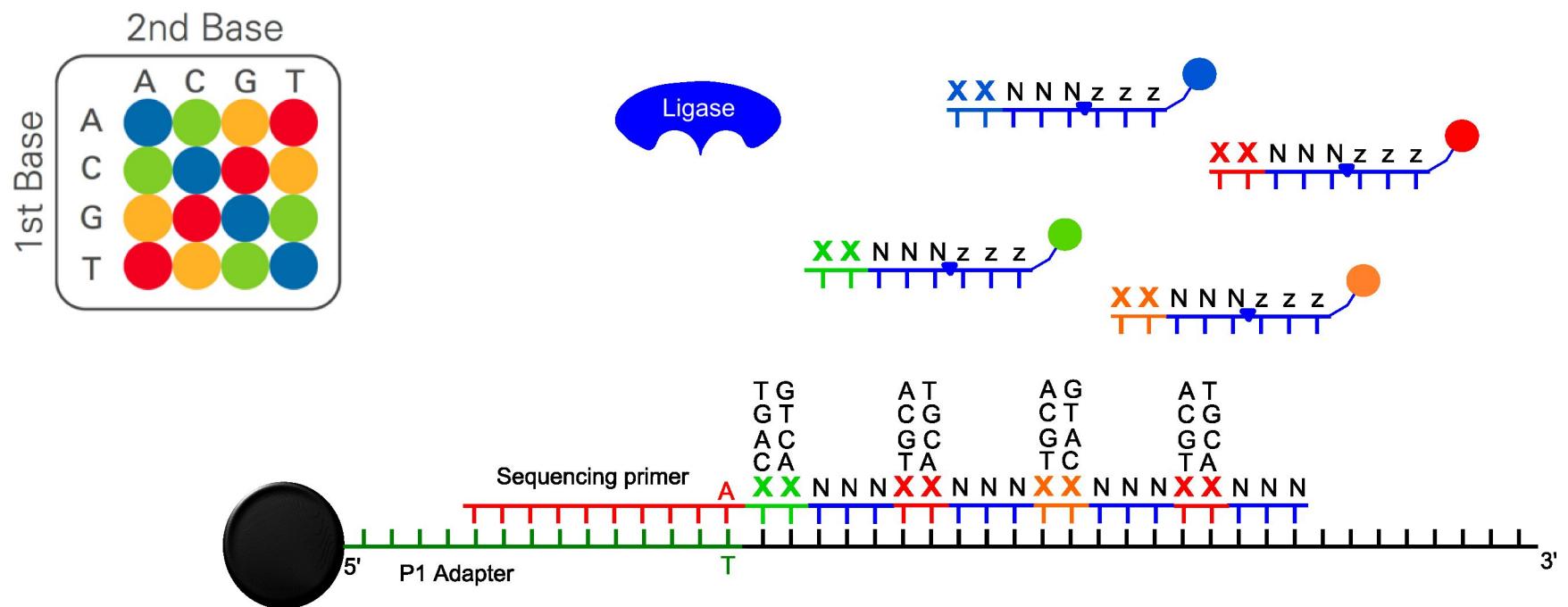
### Next Generation Sequencing: Amplified Single Molecule Sequencing SOLiD



## Different platforms

34/165

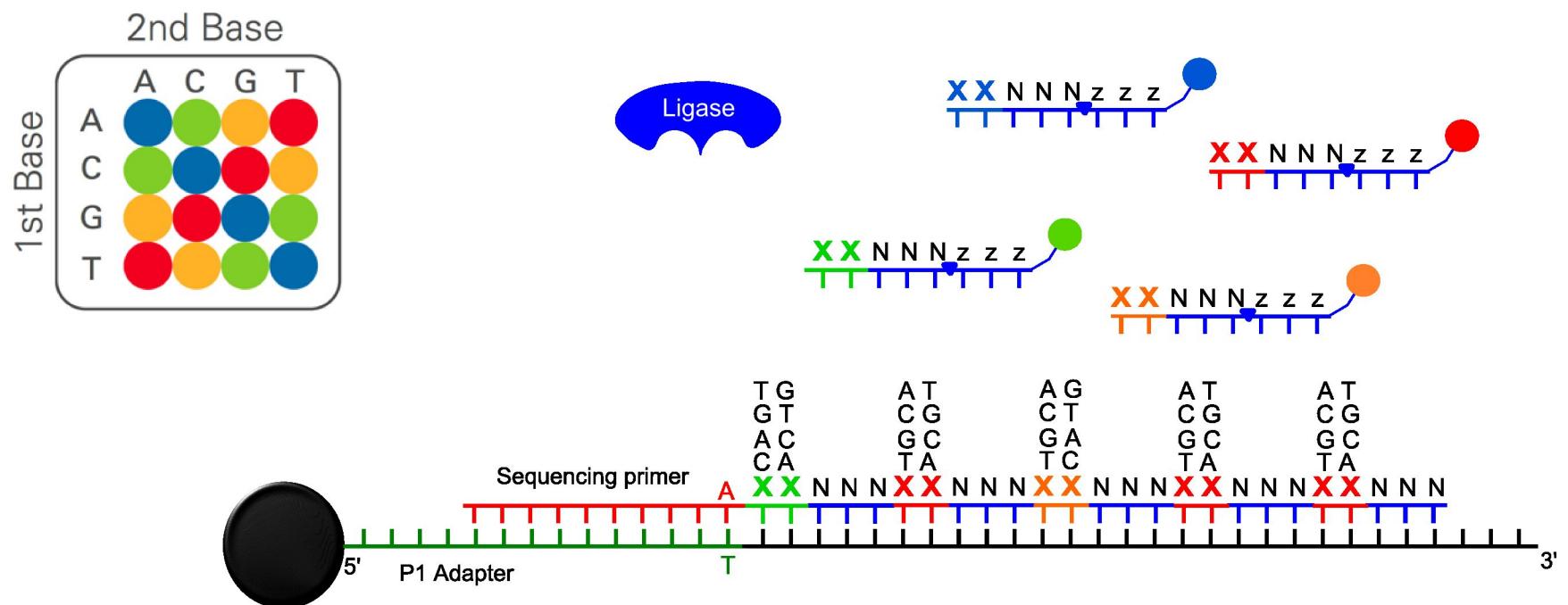
### Next Generation Sequencing: Amplified Single Molecule Sequencing SOLiD



## Different platforms

35/165

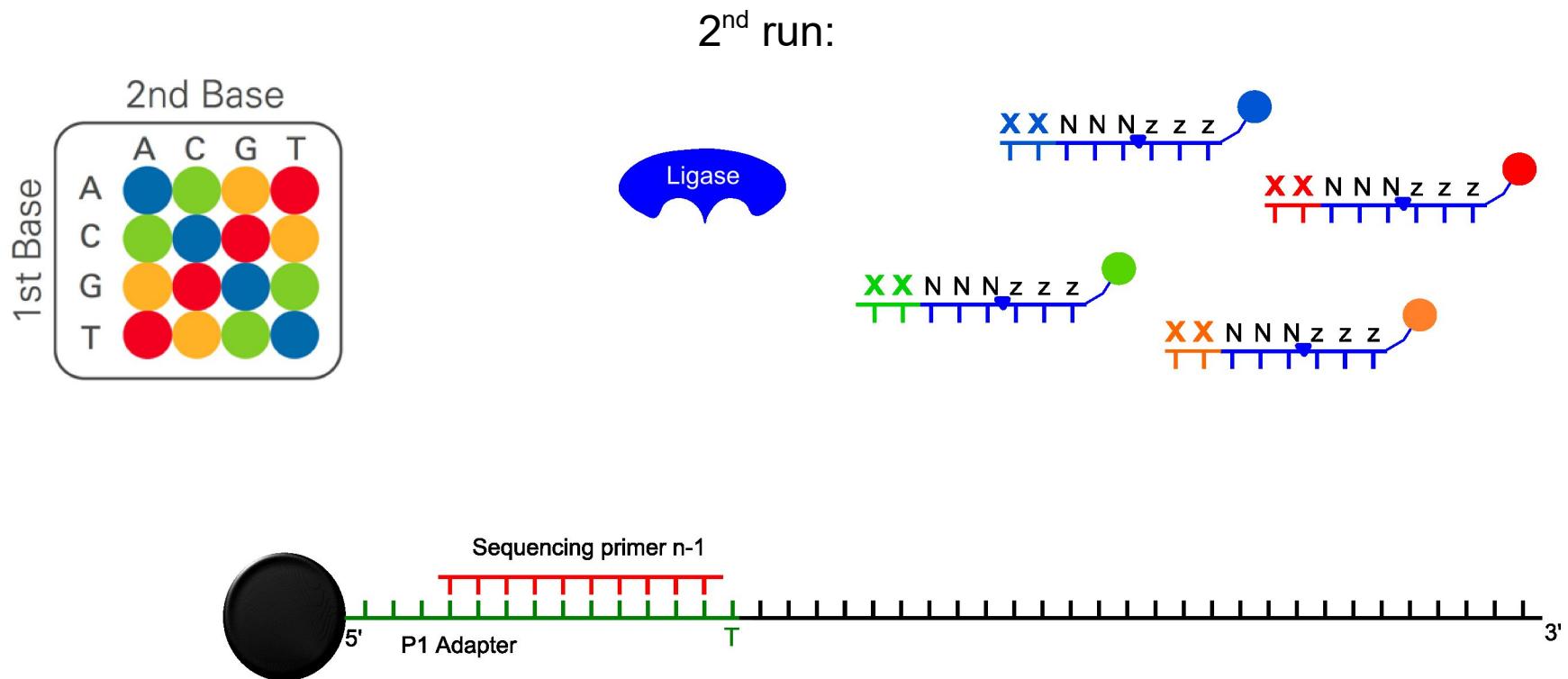
### Next Generation Sequencing: Amplified Single Molecule Sequencing SOLiD



## Different platforms

36/165

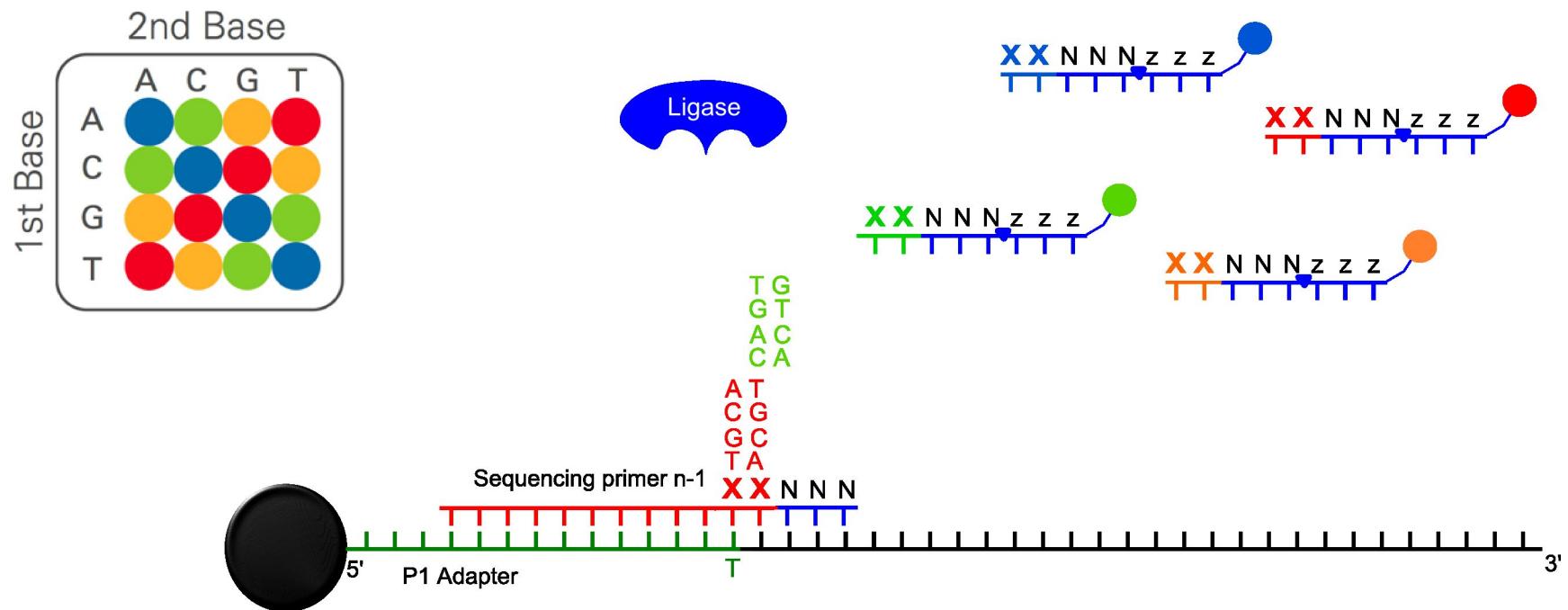
### Next Generation Sequencing: Amplified Single Molecule Sequencing SOLiD



## Different platforms

37/165

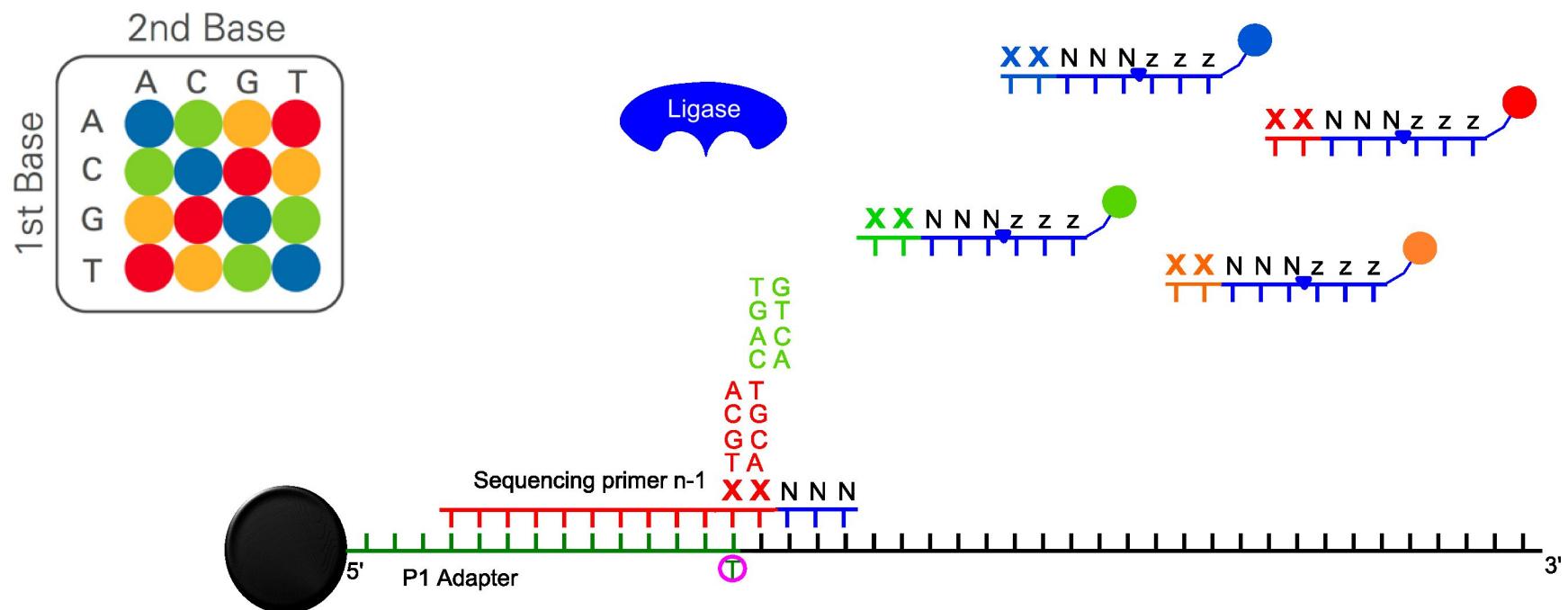
### Next Generation Sequencing: Amplified Single Molecule Sequencing SOLiD



## Different platforms

38/165

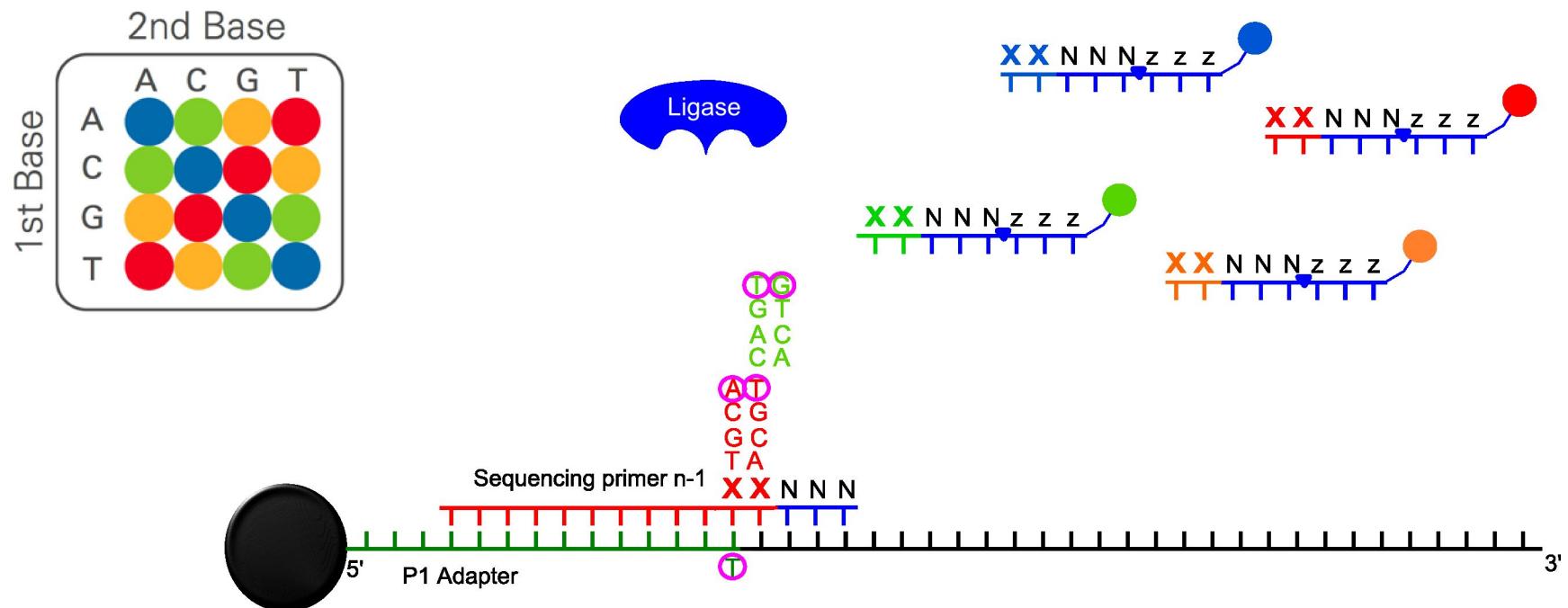
### Next Generation Sequencing: Amplified Single Molecule Sequencing SOLiD



## Different platforms

39/165

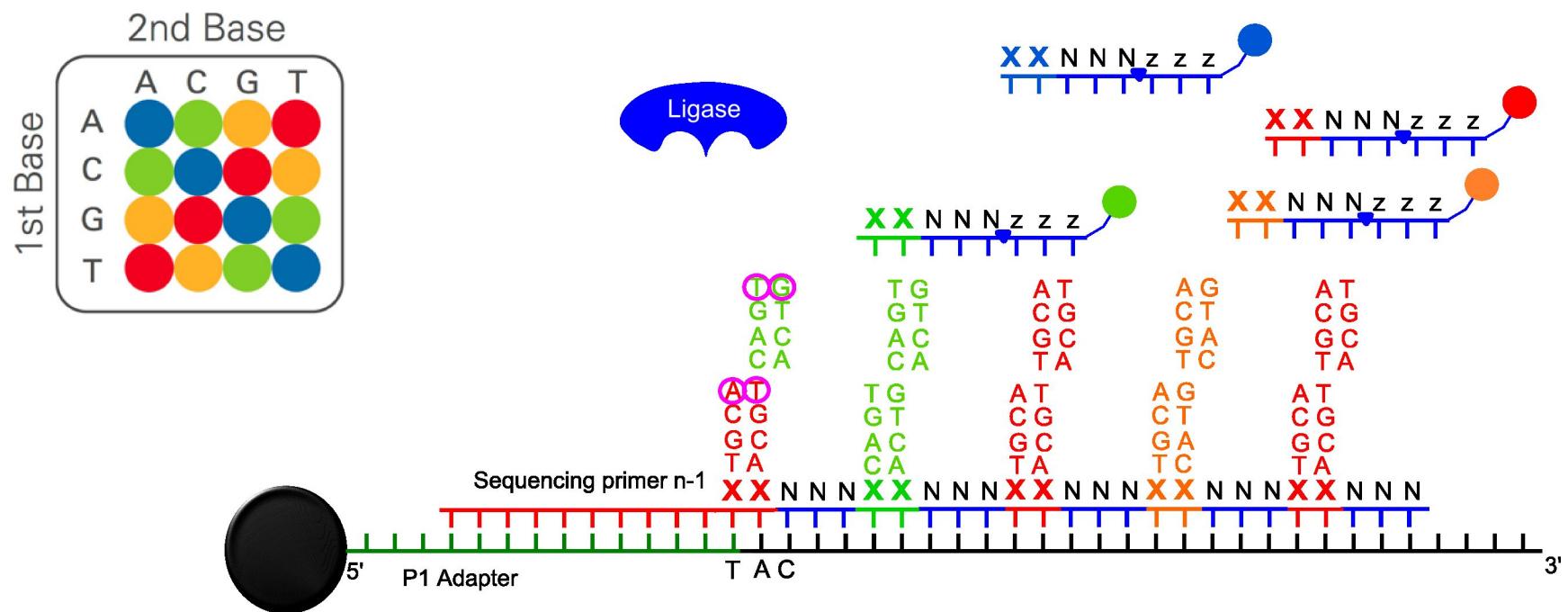
### Next Generation Sequencing: Amplified Single Molecule Sequencing SOLiD



# Different platforms

40/165

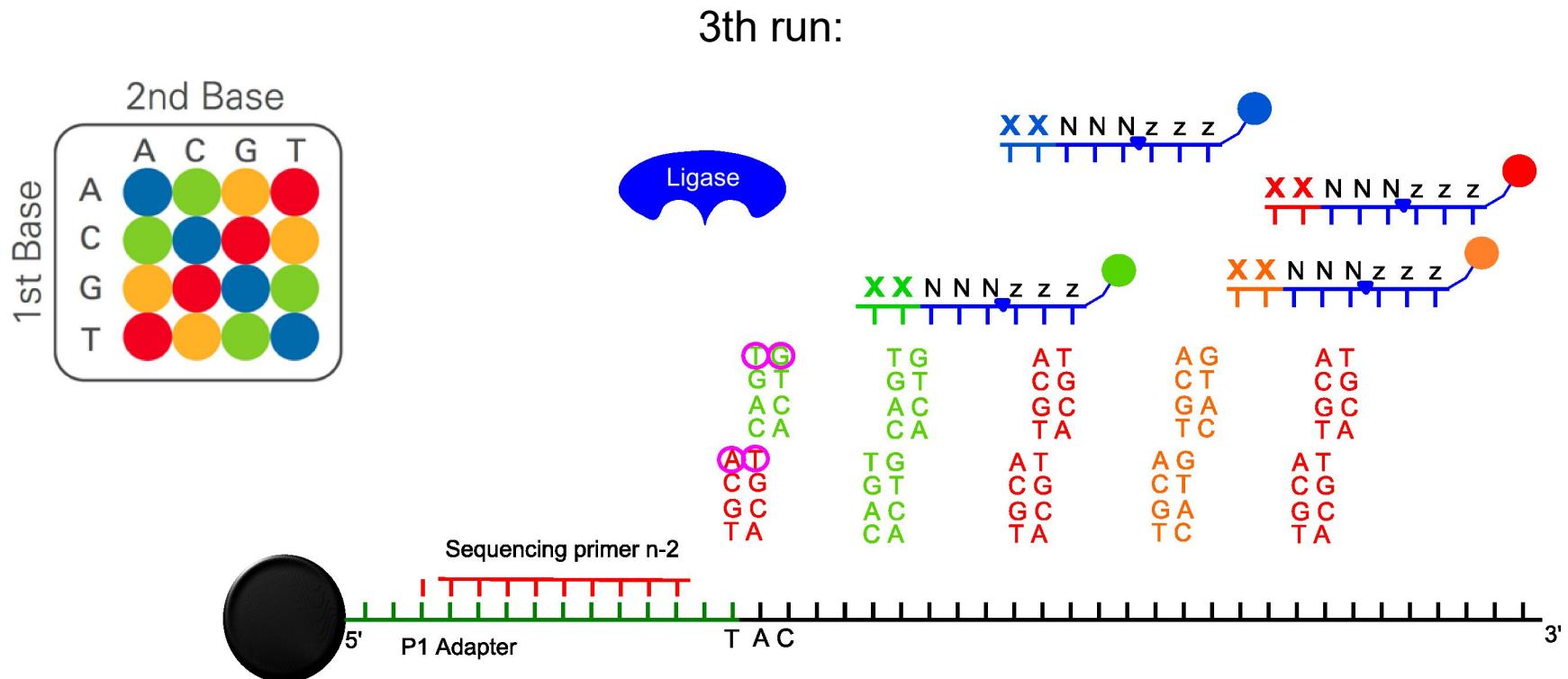
## Next Generation Sequencing: Amplified Single Molecule Sequencing SOLiD



## Different platforms

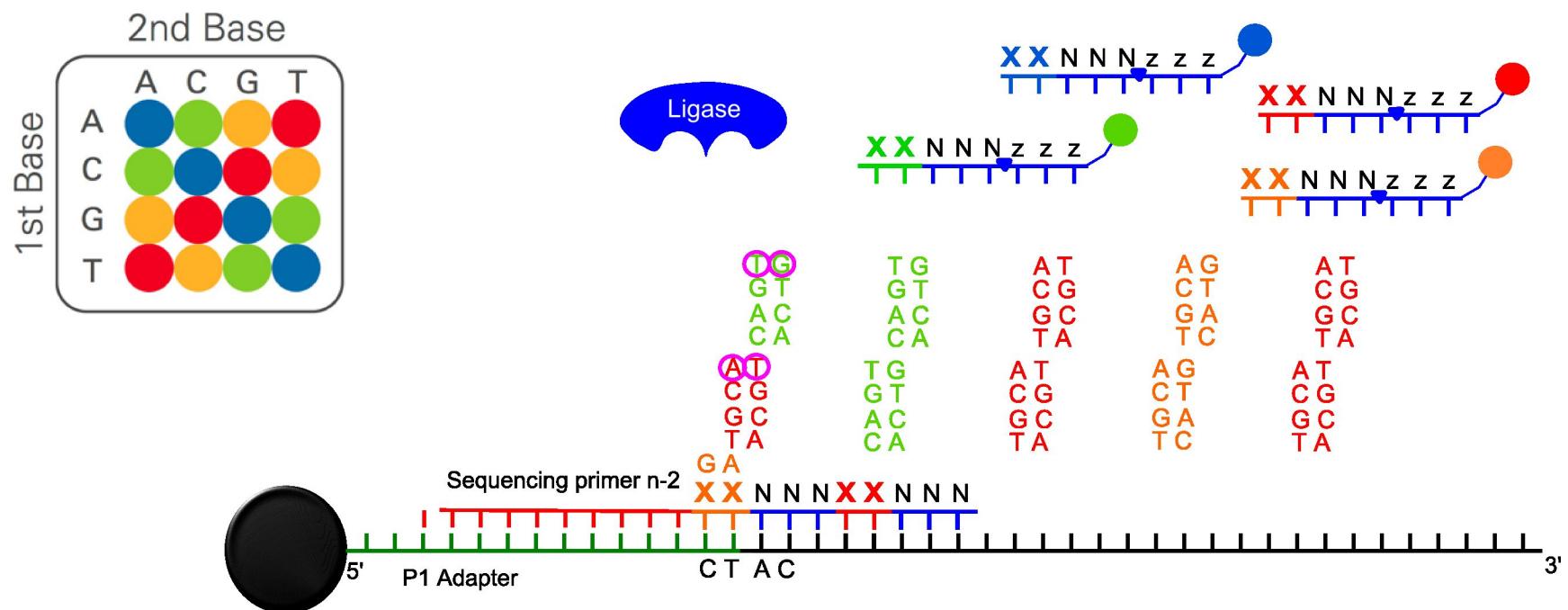
41/165

### Next Generation Sequencing: Amplified Single Molecule Sequencing SOLiD



## Different platforms

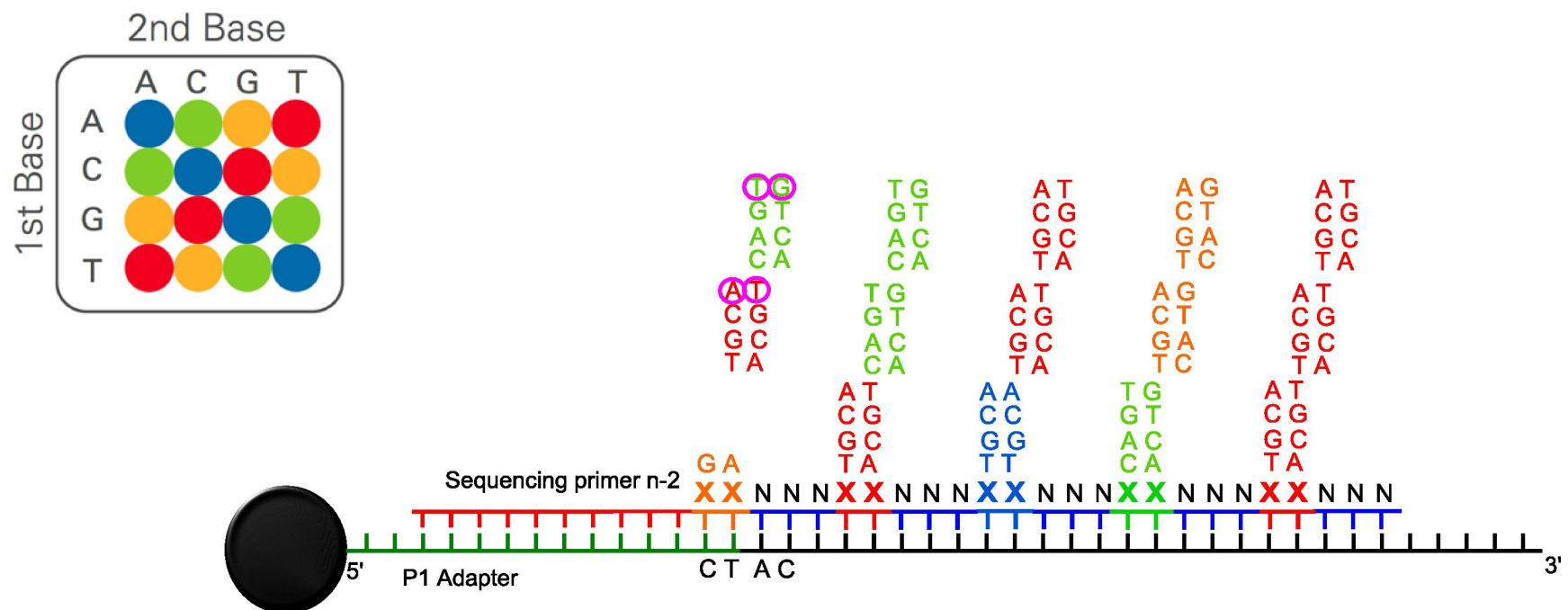
### Next Generation Sequencing: Amplified Single Molecule Sequencing SOLiD



## Different platforms

43/165

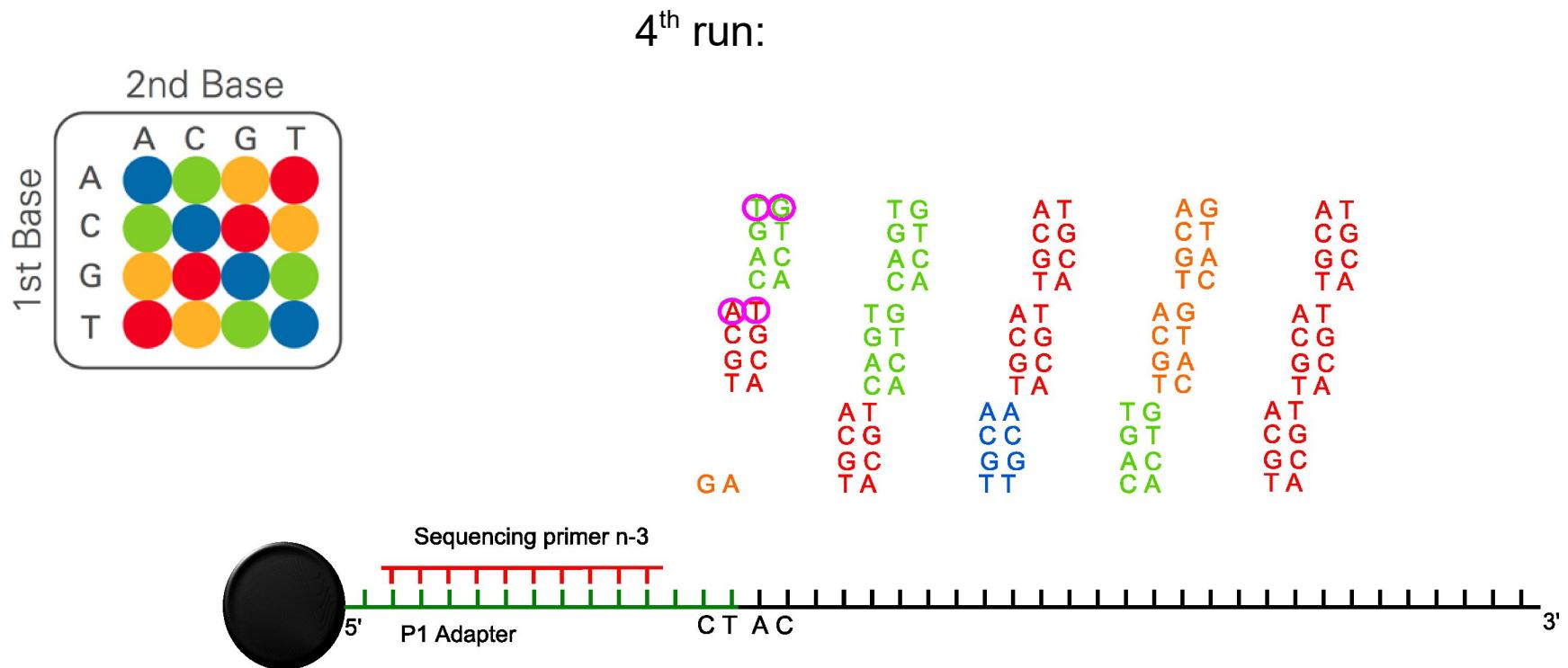
### Next Generation Sequencing: Amplified Single Molecule Sequencing SOLiD



## Different platforms

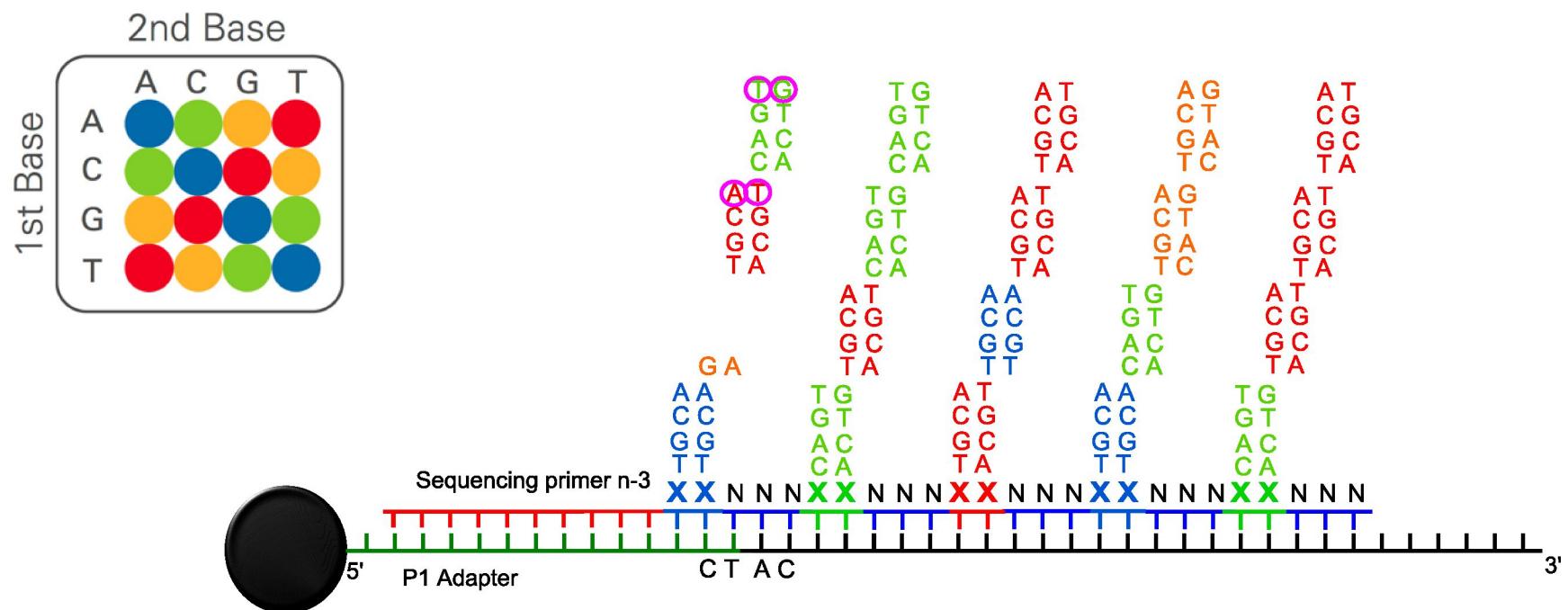
44/165

### Next Generation Sequencing: Amplified Single Molecule Sequencing SOLiD



## Different platforms

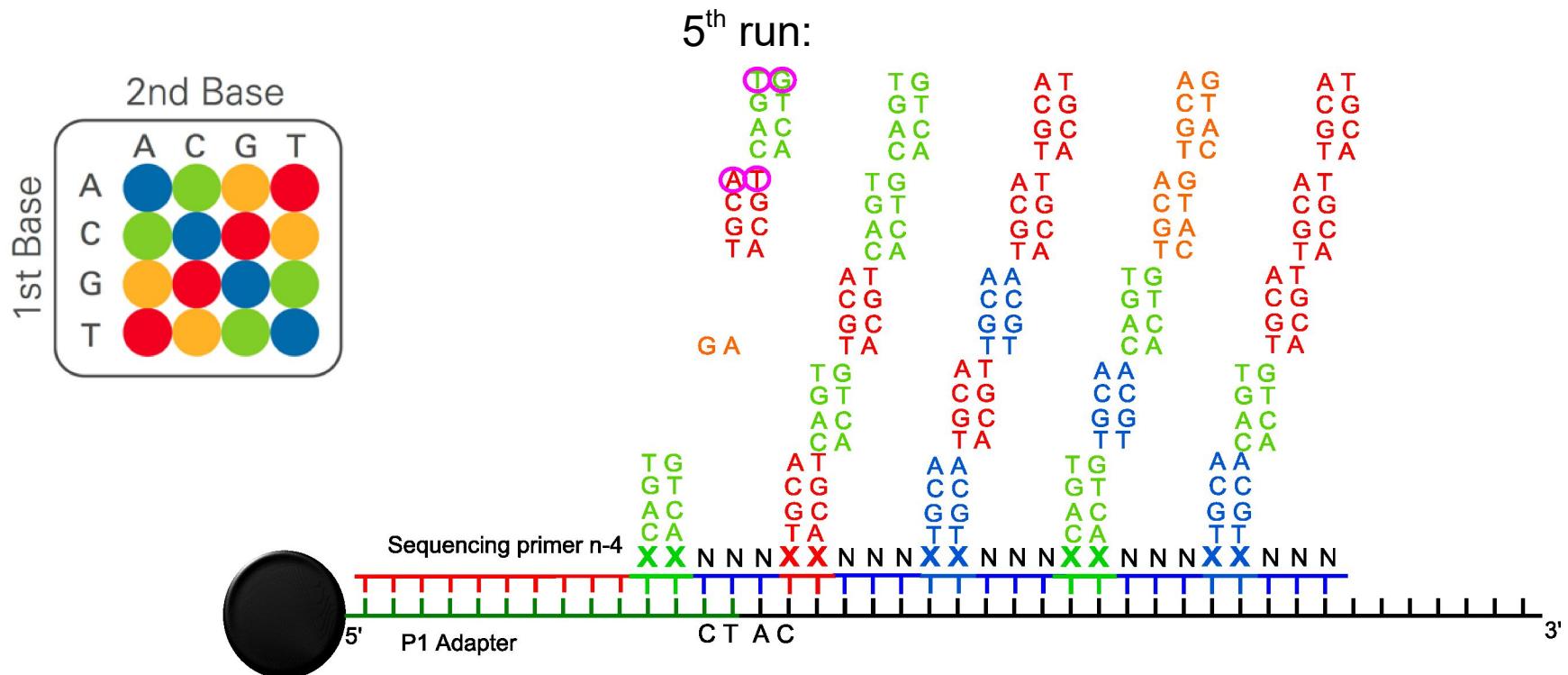
### Next Generation Sequencing: Amplified Single Molecule Sequencing SOLiD



## Different platforms

46/165

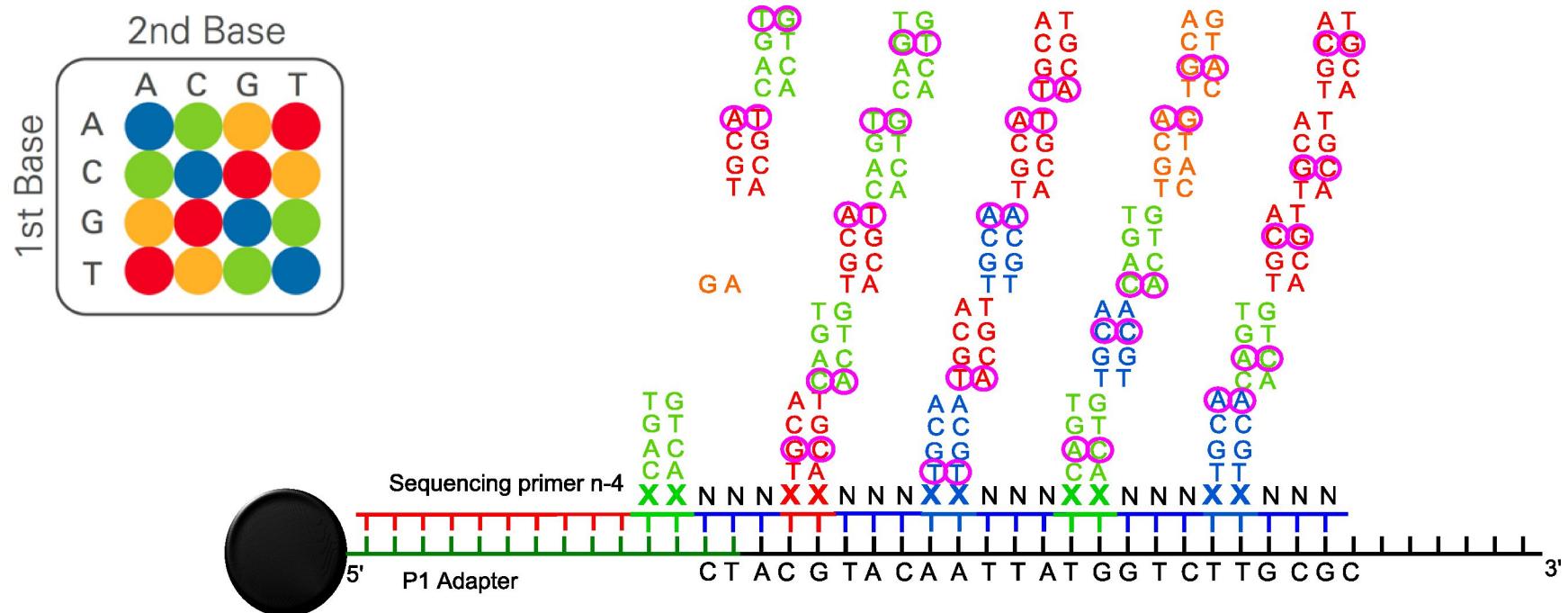
### Next Generation Sequencing: Amplified Single Molecule Sequencing SOLiD



# Different platforms

47/165

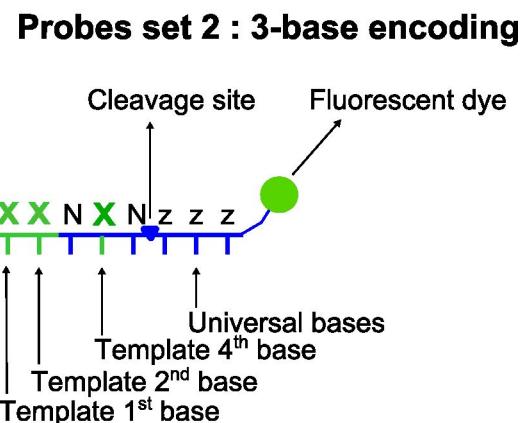
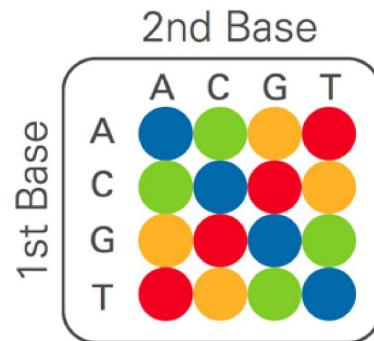
## Next Generation Sequencing: Amplified Single Molecule Sequencing SOLiD



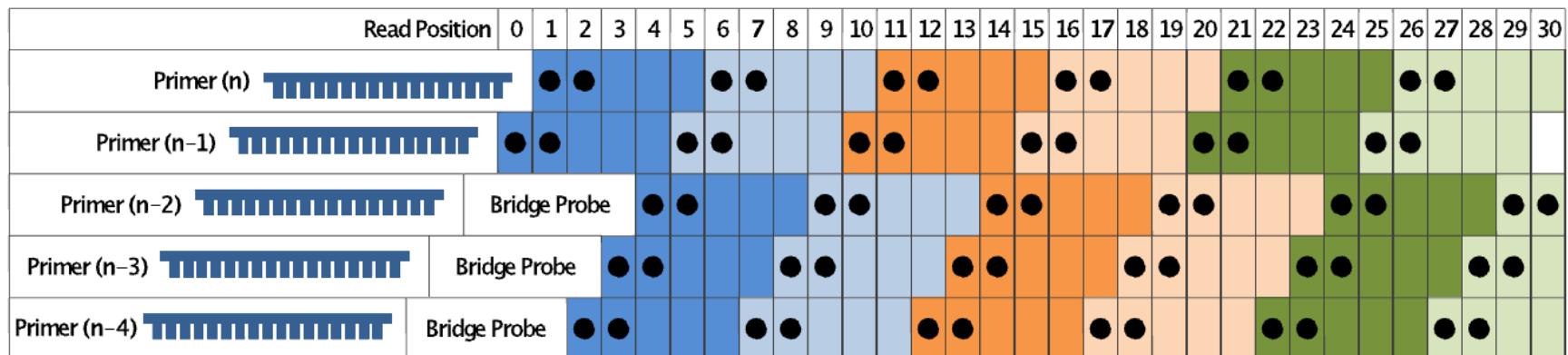
# Different platforms

# Next Generation Sequencing: Amplified Single Molecule Sequencing **SOLiD**

## More accuracy: 3-base encoding probes



**8. Repeat steps 1-7 with Primers n-1, n-2, n-3, and n-4 using Probe Set 1**



#### **9. Repeat steps 1–7 with Primer n-4 using Probe Set 2**

- Interrogated Positions

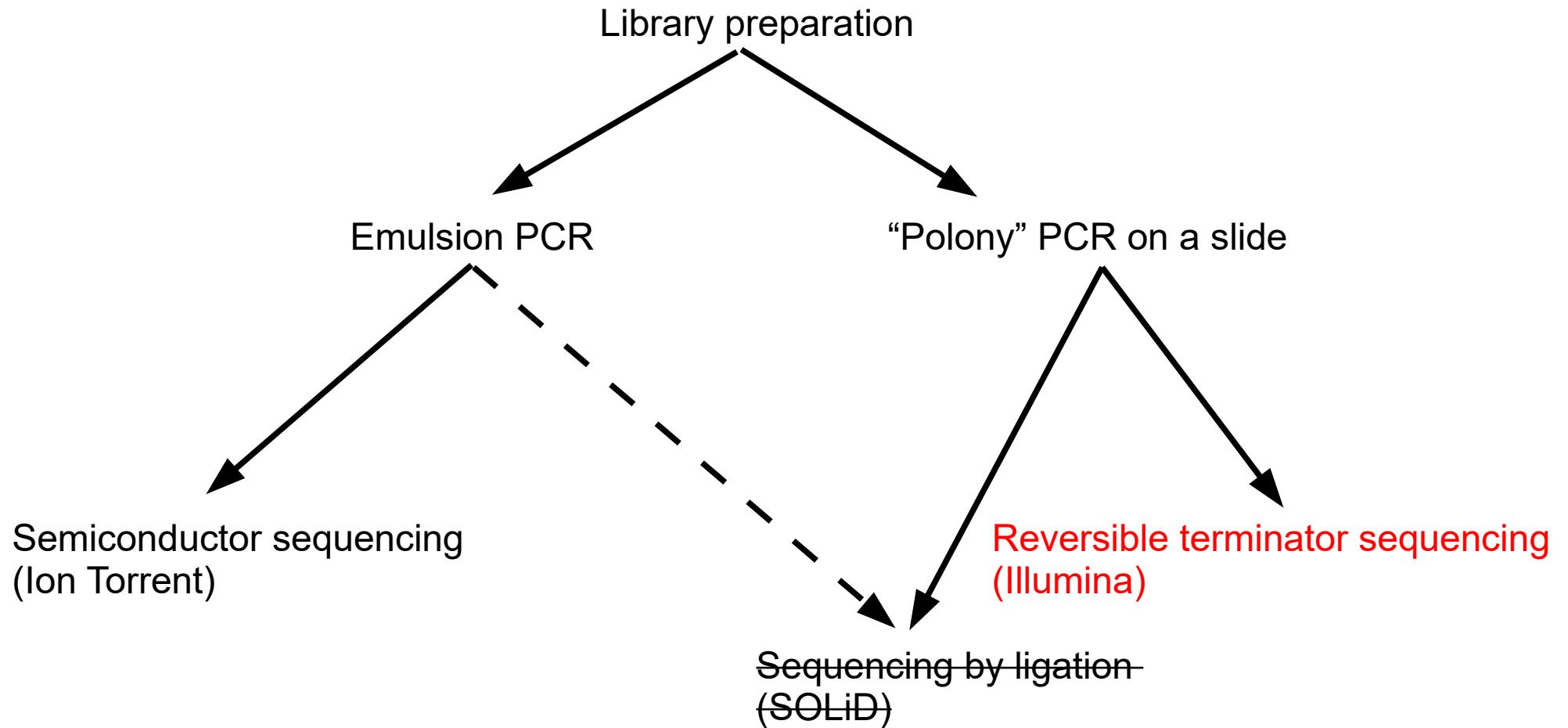
## Ligation Cycle 1 2 3 4 5 6



## Workflow

49/165

### Next Generation Sequencing: Amplified Single Molecule Sequencing



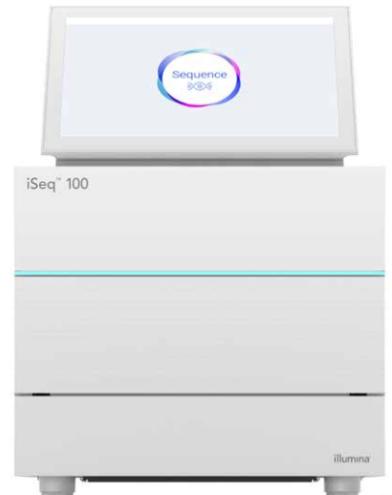
## Different platforms

50/165

### Next Generation Sequencing: Amplified Single Molecule Sequencing

#### Illumina

iSeq 100



MiniSeq



MiSeq



NextSeq 500 / 550



HiSeq 2500 / 3000 / 4000



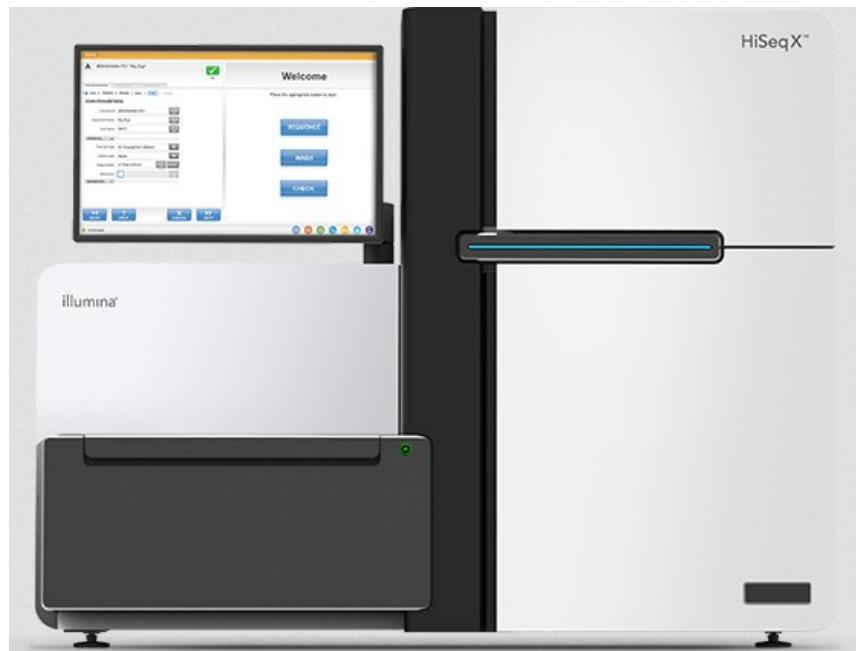
## Different platforms

51/165

### Next Generation Sequencing: Amplified Single Molecule Sequencing

Illumina

HiSeq X



NovaSeq 6000



## Different platforms

### 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

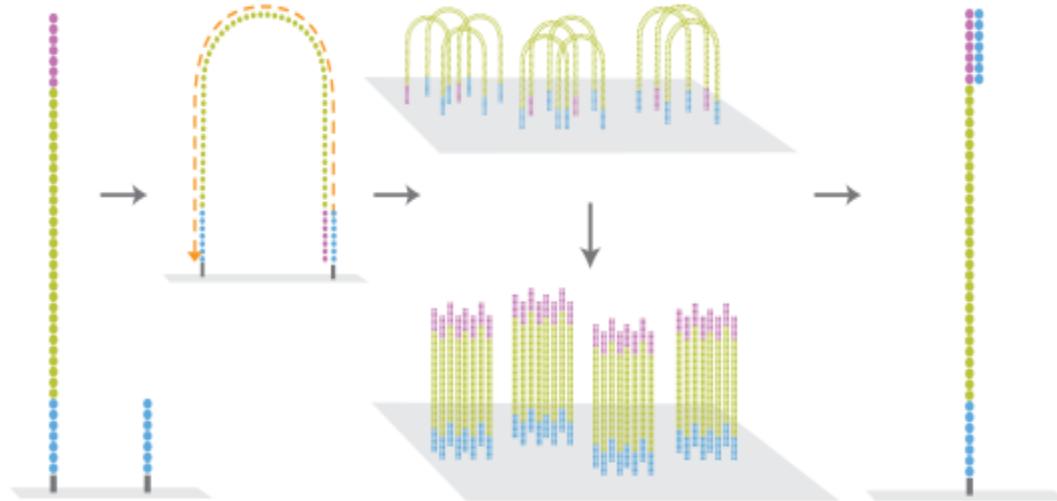
	HiSeq 2500 / 3000 / 4000	HiSeq X	NovaSeq 6000
Read Length	2 x 125 / 2 x 150 / 2 x 150 bp	2 x 150 bp	2 x 150 bp
Throughput	1000 / 750 / 1500 Gb	1800 Gb	850 – 3000 Gb
Reads per run	4 / 2,5 / 5 billion	6 billion	2.8 – 10 billion
Accuracy	99,9 % (>80% of the bases)	99,9 % (>75%)	99,9 % (>75% of the bases)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

Workflow: Library preparation → Bridge amplification → Reversible termination sequencing

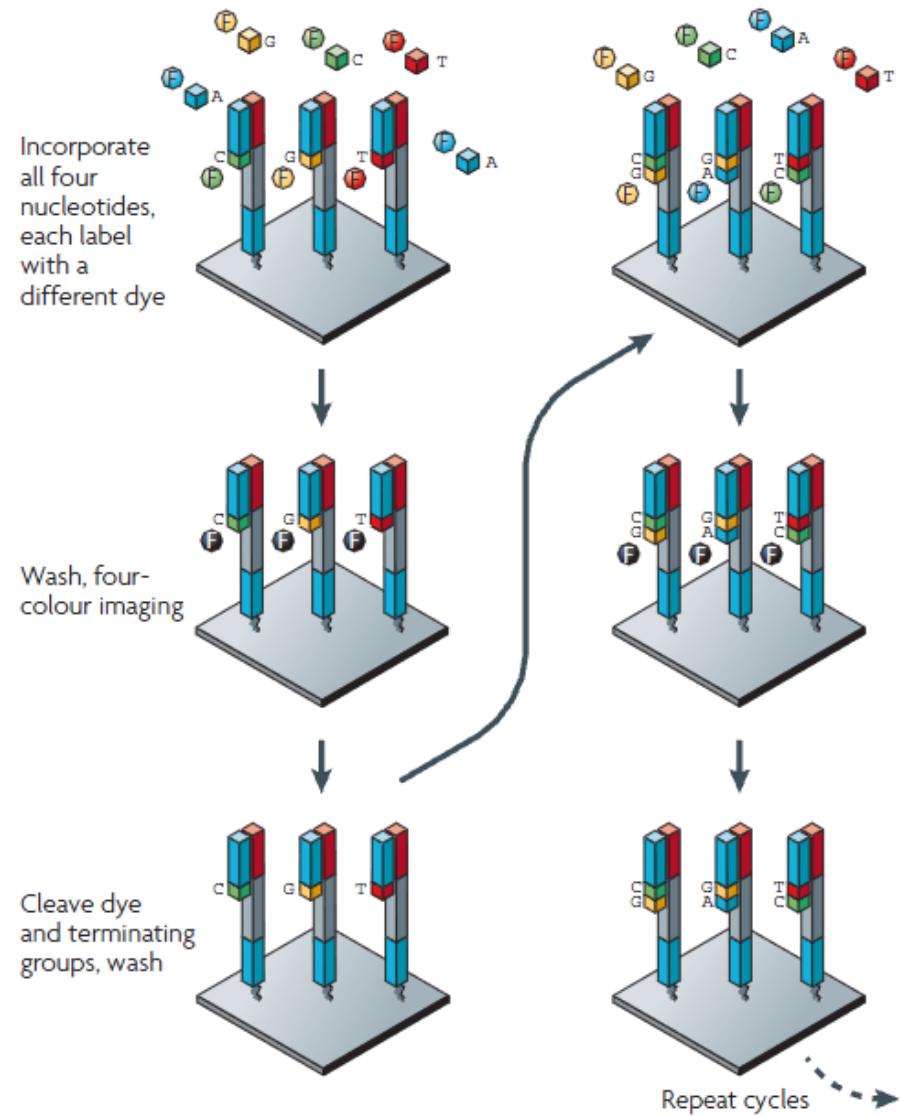
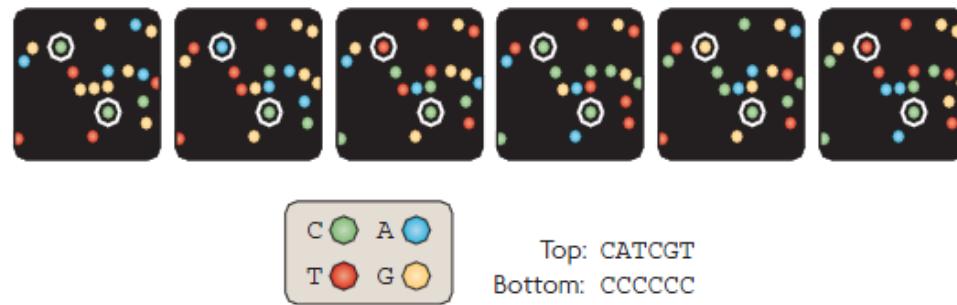
## Different platforms

### Next Generation Sequencing: Amplified Single Molecule Sequencing

Illumina: Reversible termination sequencing



4 nucleotides with different dye flow simultaneous



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

## Different platforms

54/165

### Next Generation Sequencing: Amplified Single Molecule Sequencing

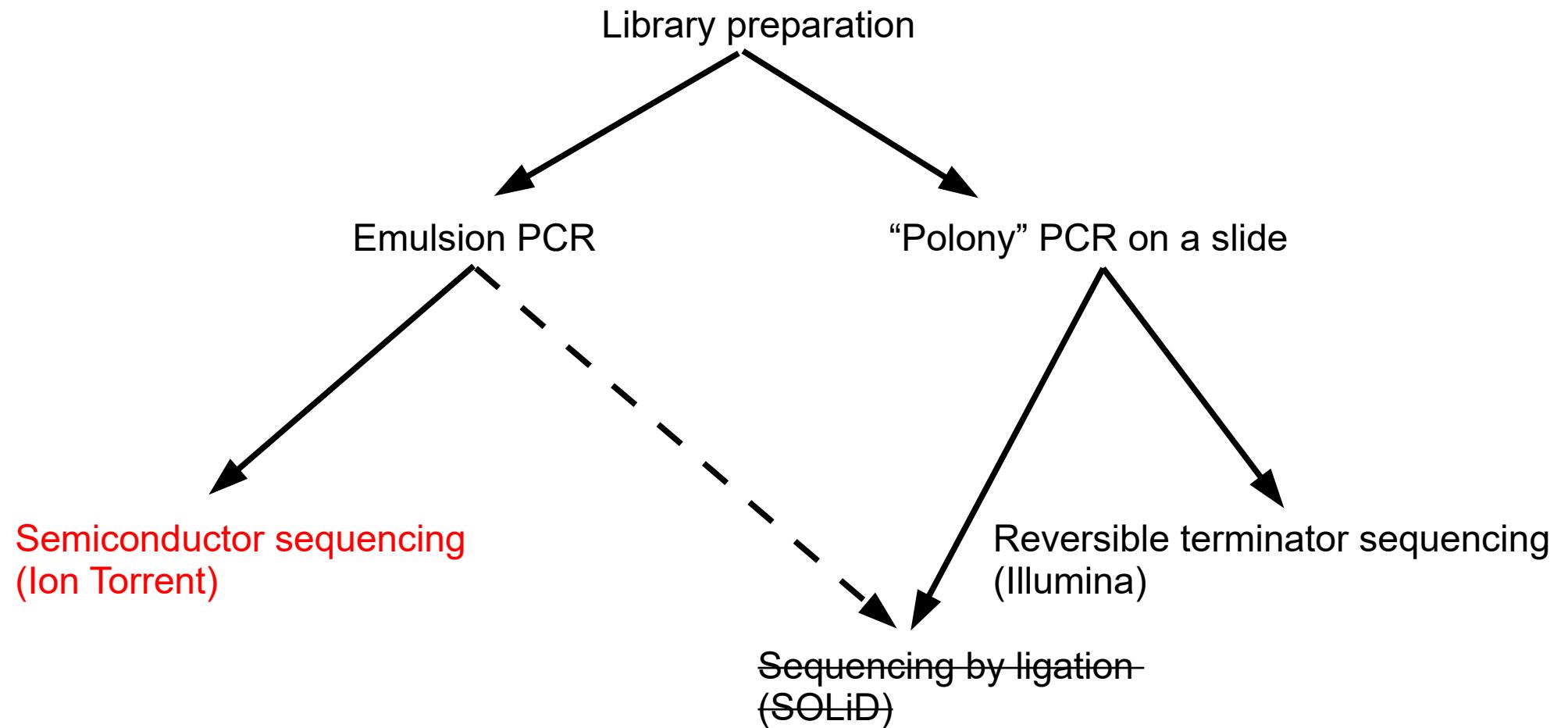
Illumina

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

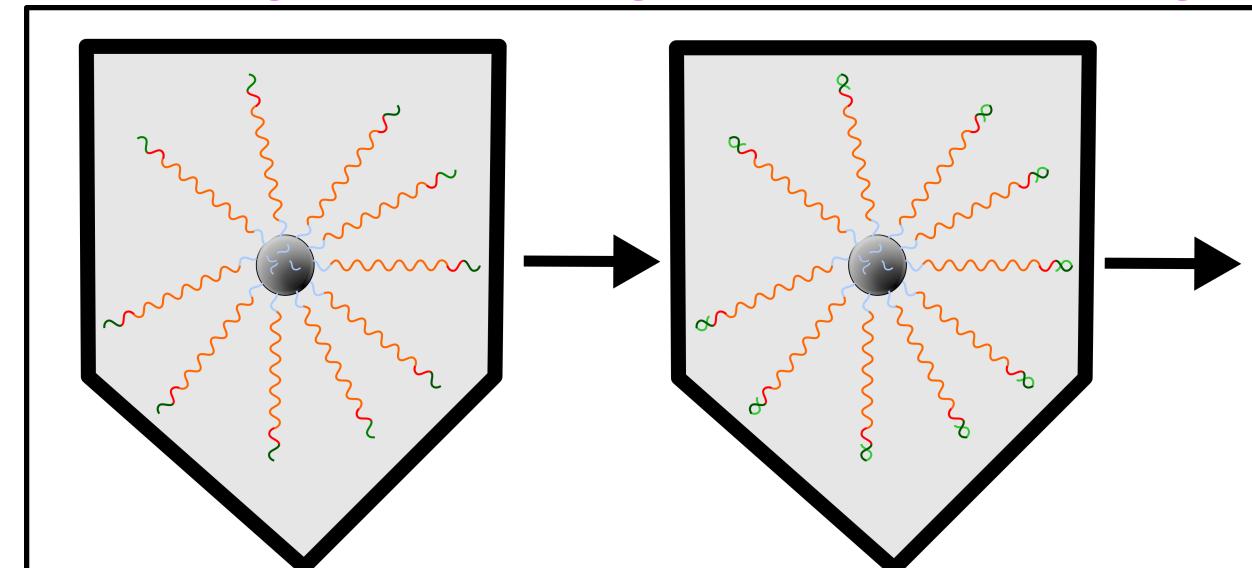


	PGM	GeneStudio S5 / S5 Plus / S5 Prime	Proton
<b>Chip</b>	314 - 316 - 318	510 - 520 - 530 - 540 - 550	PI – PII
<b>Read length</b>	400 bp	400(600) - 400(600) – 400(600) – 200 - 200	200 bp - ?
<b>Throughput</b>	0,1 - 0,6 - 2 Gb	1 - 2 - 8 - 15 - 25 Gb	15 -100 Gb
<b>Reads per run</b>	0,5 - 3 - 6 million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

## Different platforms

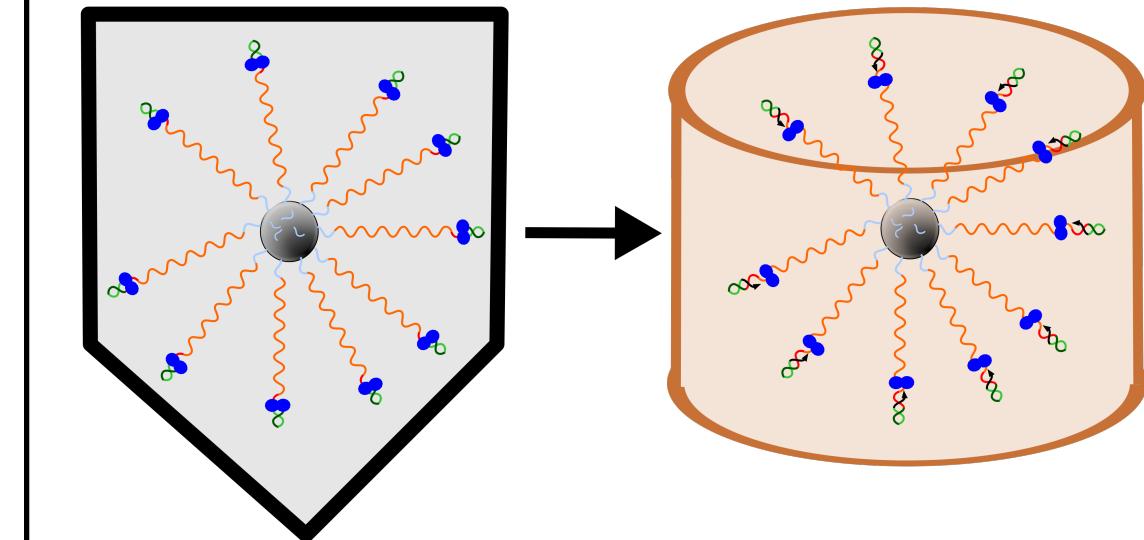
### Next Generation Sequencing: Amplified Single Molecule Sequencing

Ion Torrent



Sequencing

1. Denaturation : the bead contains single stranded DNA.  
2. Annealing : the sequencing primer anneals to the single strands.



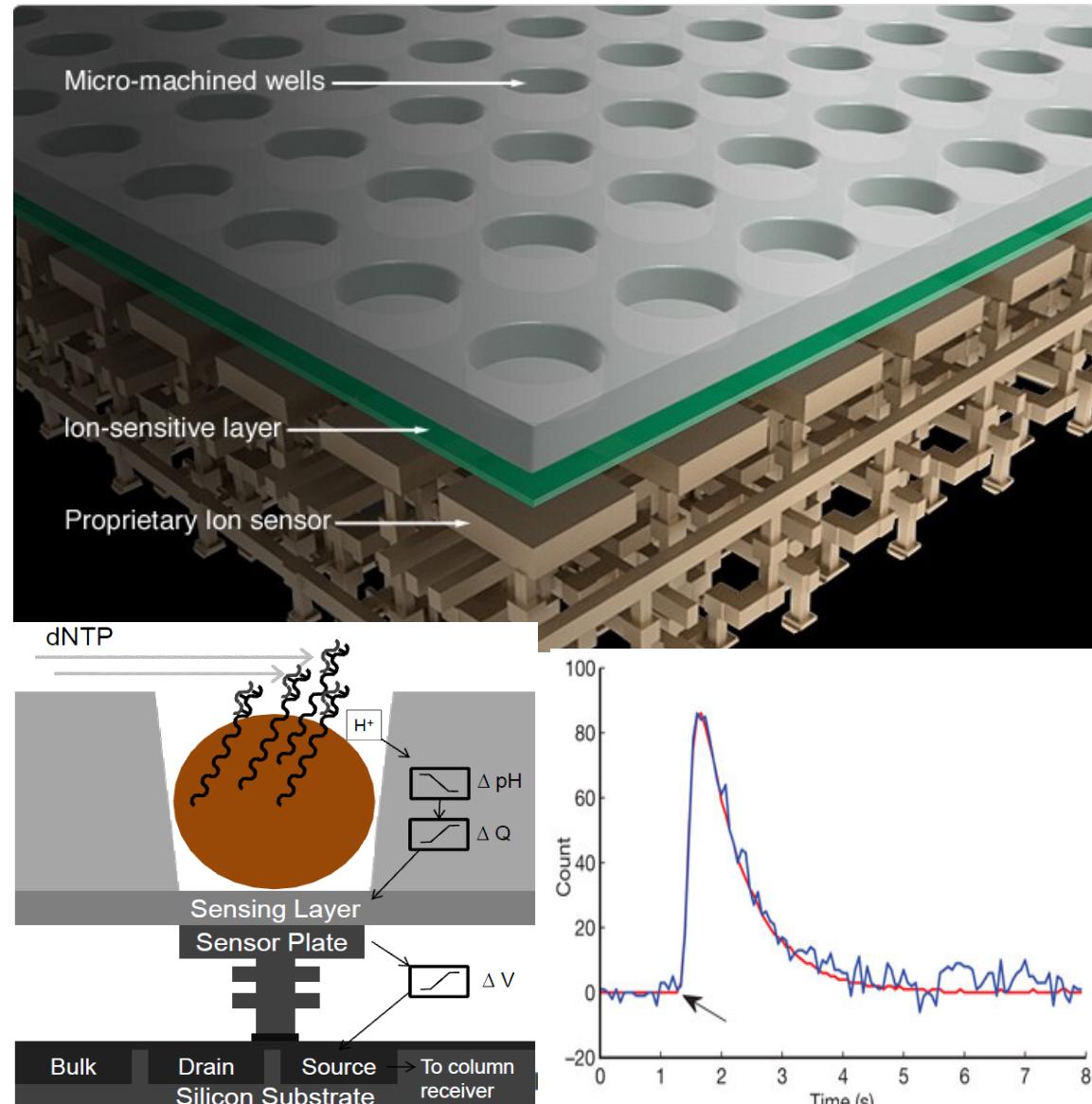
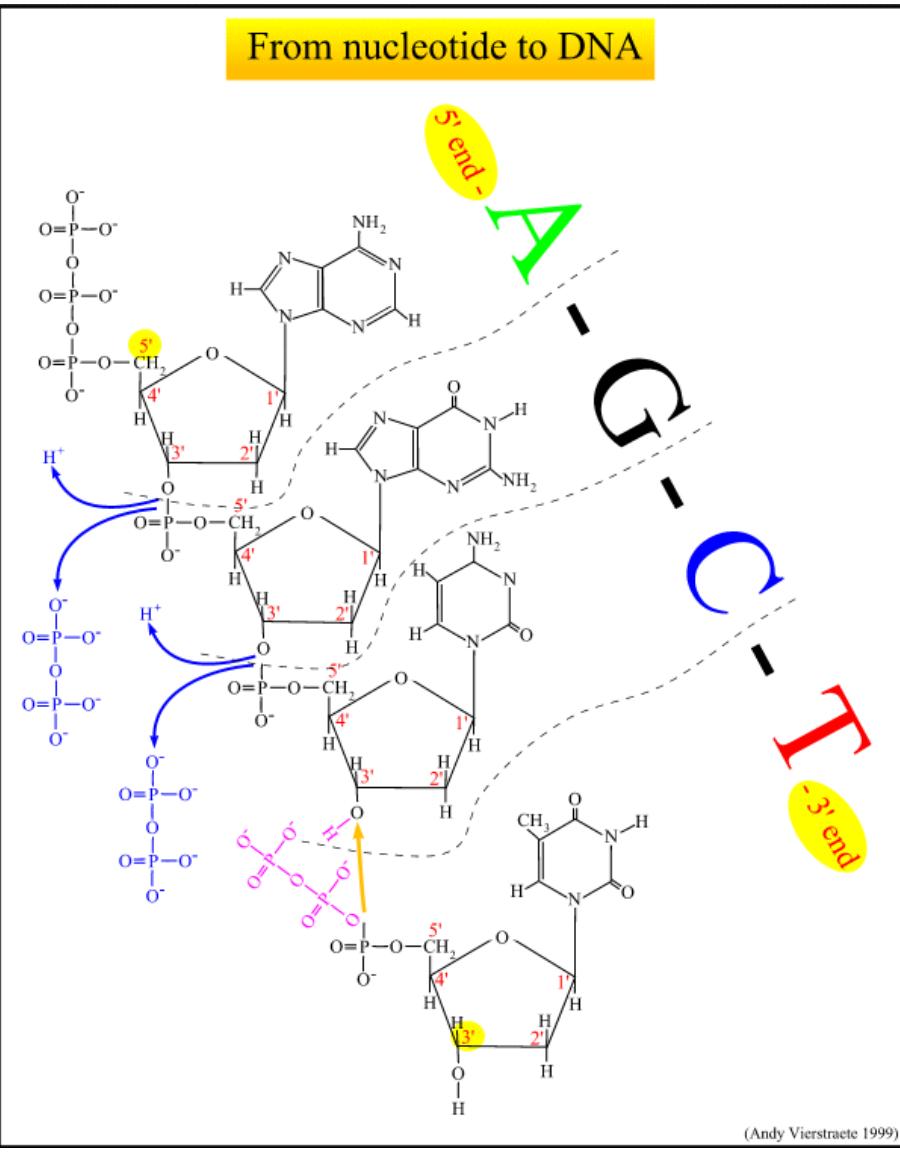
3. Adding polymerase : polymerase attaches to the DNA.

4. Sequencing : the beads with DNA are loaded in the wells, polymerase amplifies the strand starting from the primer towards the bead.  
Nucleotides flow sequentially over the beads.

## Different platforms

### Next Generation Sequencing: Amplified Single Molecule Sequencing Ion Torrent

Workflow: Library preparation → Emulsion PCR → Semiconductor Sequencing

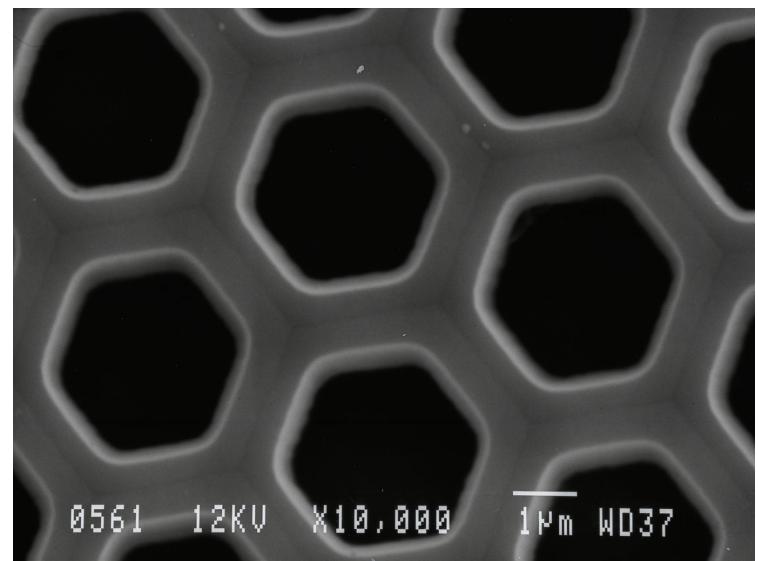
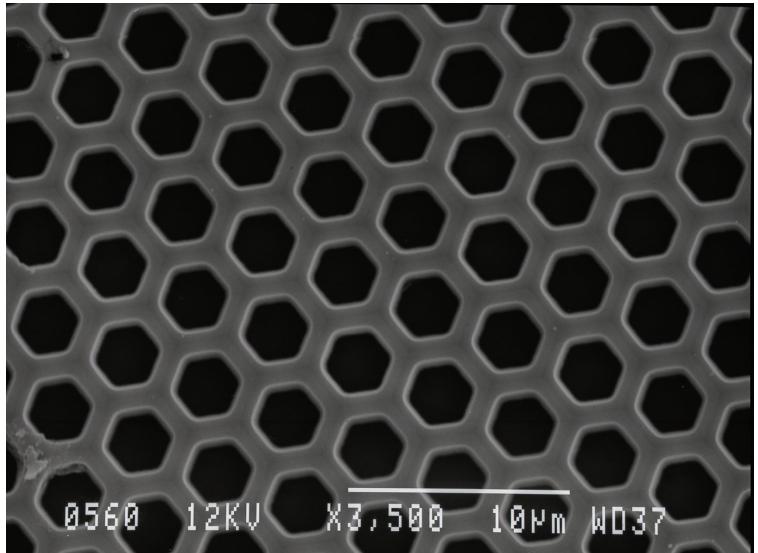


## Different platforms

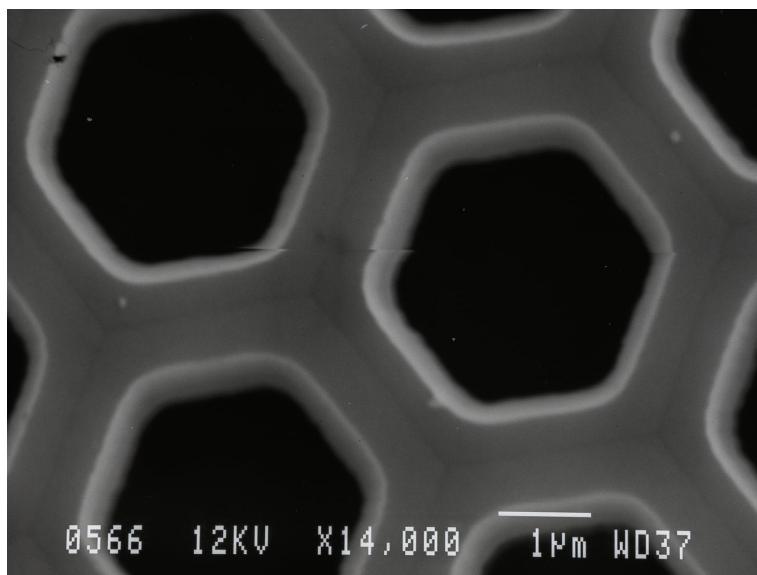
59/165

### Next Generation Sequencing: Amplified Single Molecule Sequencing Ion Torrent

Workflow: Library preparation → Emulsion PCR → Semiconductor Sequencing



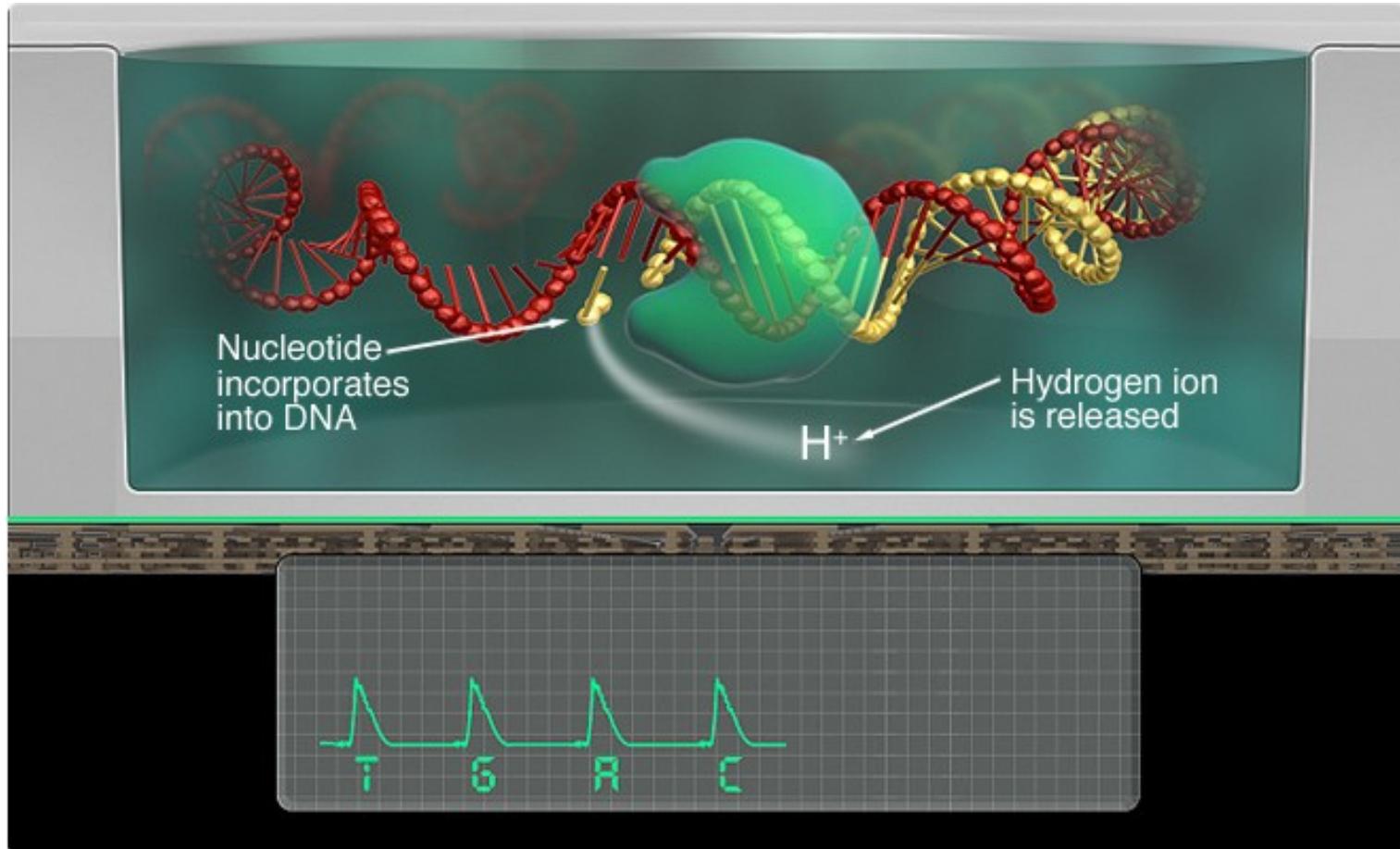
Picture of the wells in a 318 chip



## Different platforms

### Next Generation Sequencing: Amplified Single Molecule Sequencing Ion Torrent

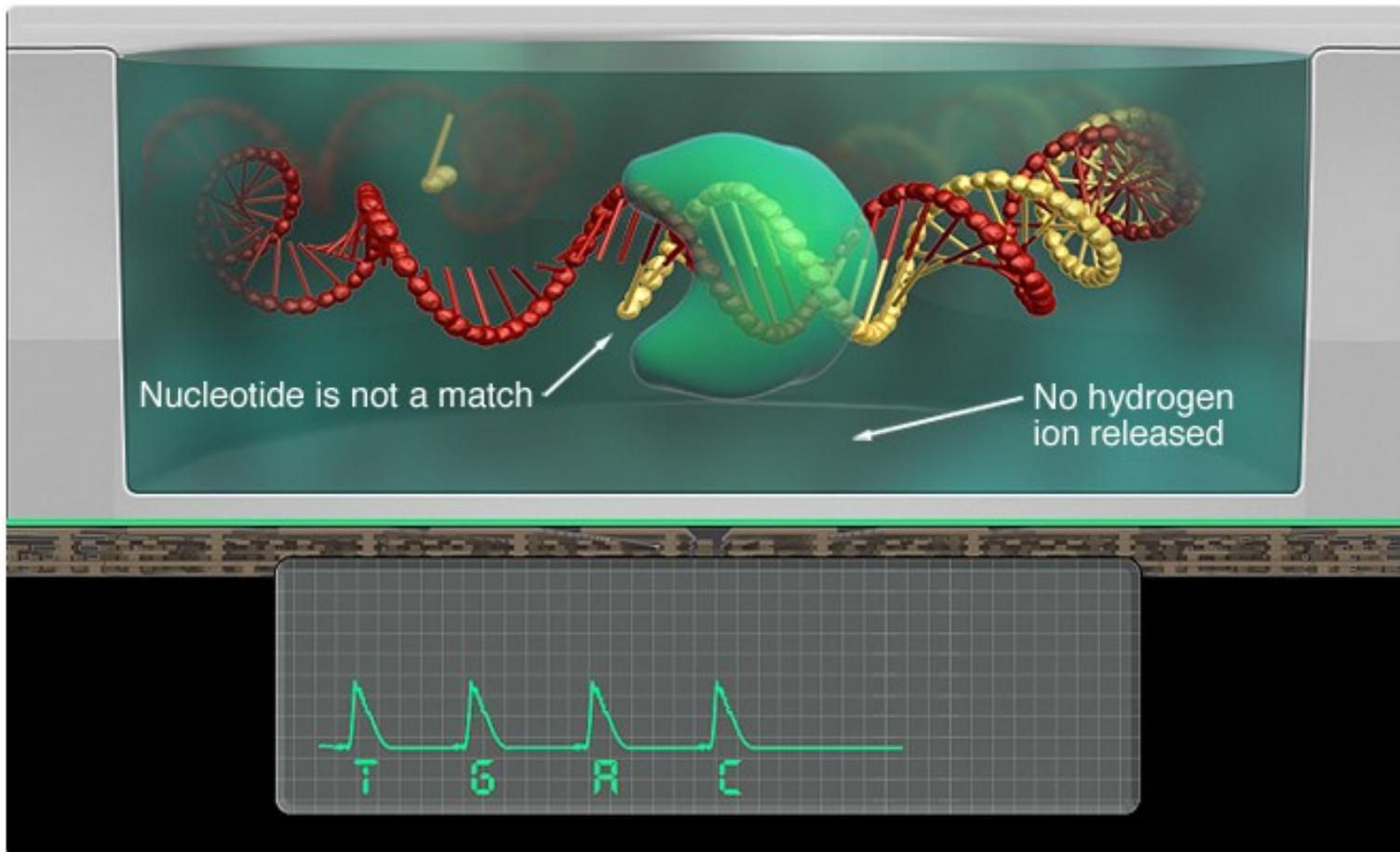
4 nucleotides flow sequentially



No camera, just a pH sensor

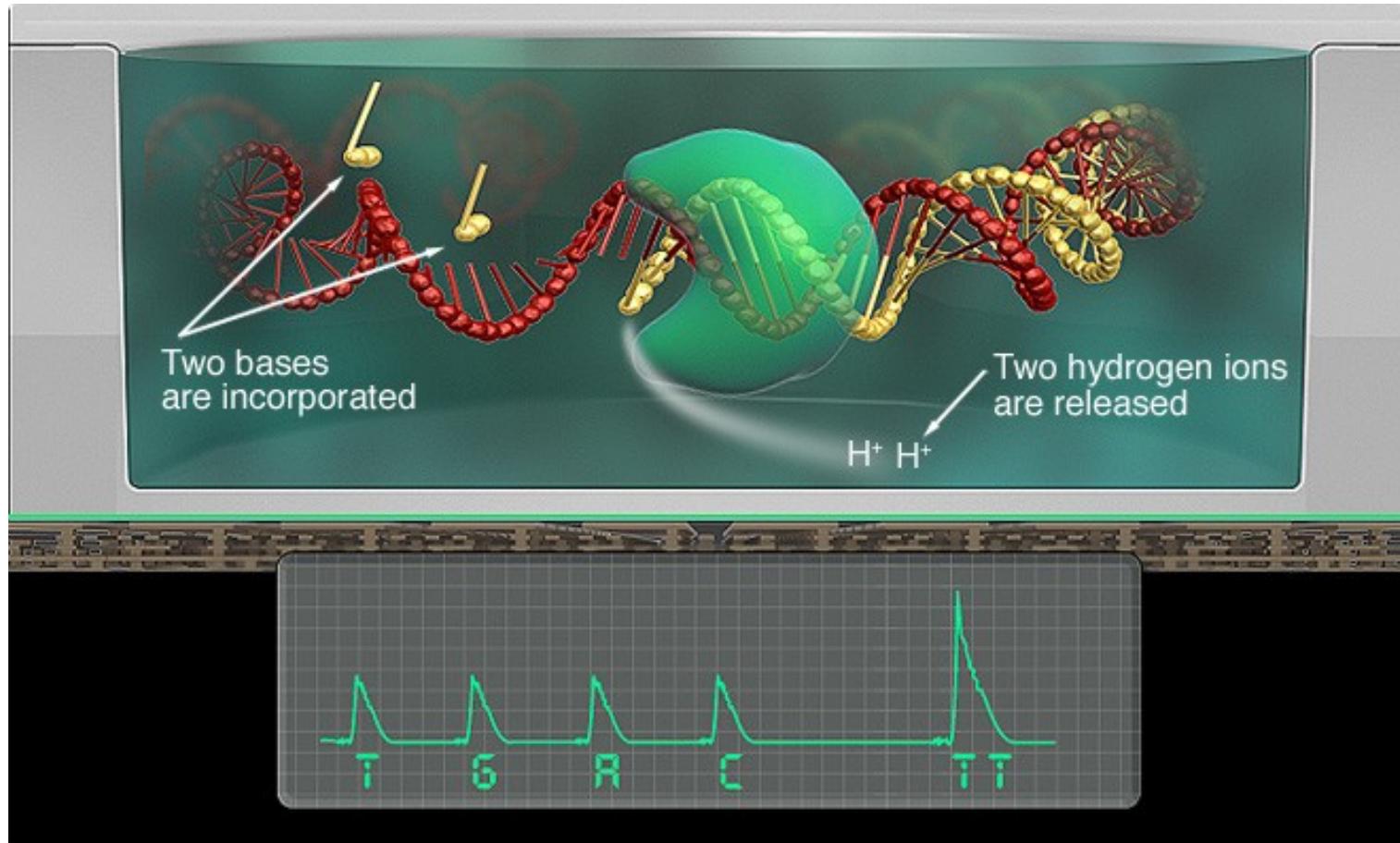
## Different platforms

### Next Generation Sequencing: Amplified Single Molecule Sequencing Ion Torrent



## Different platforms

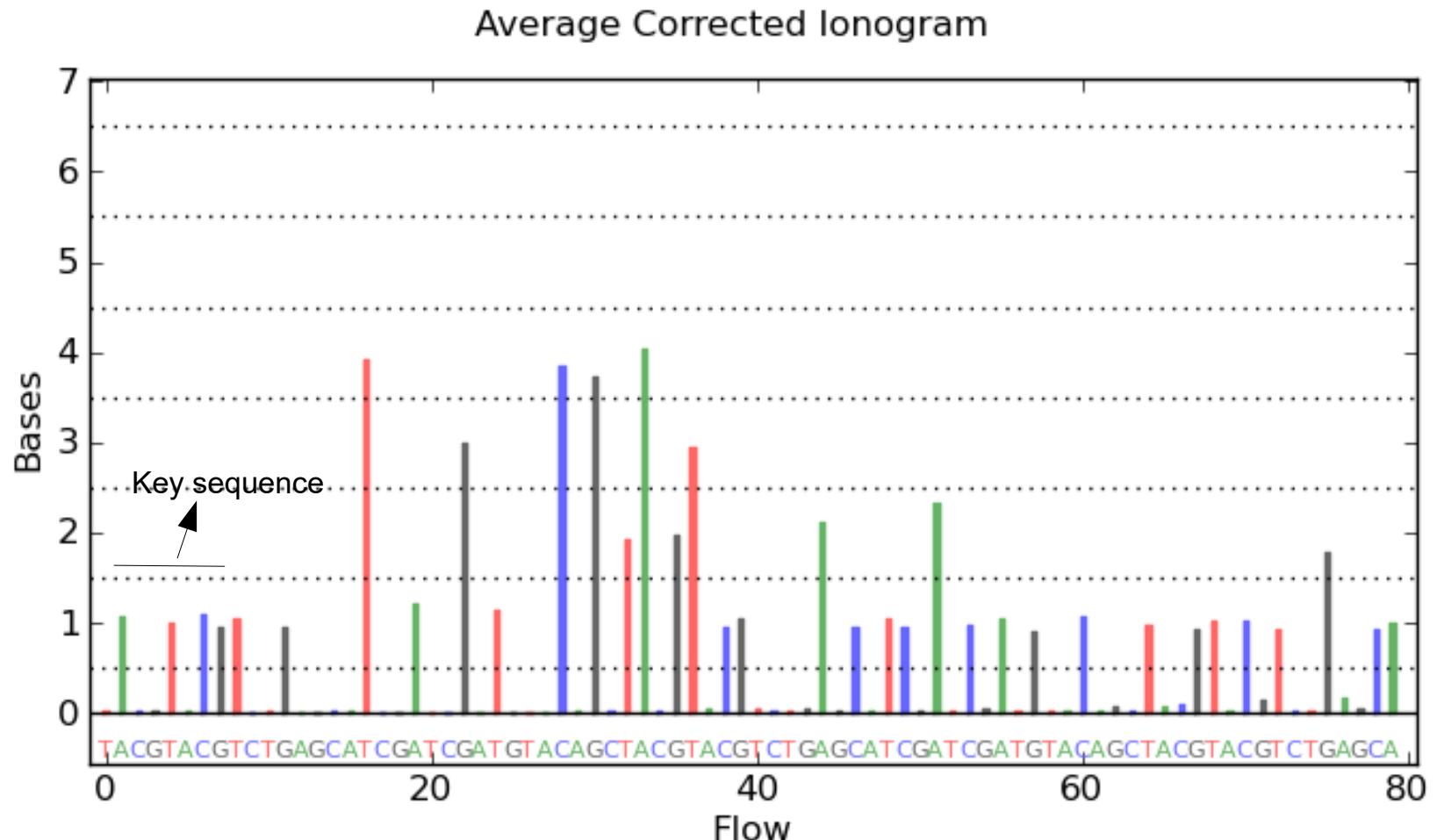
### Next Generation Sequencing: Amplified Single Molecule Sequencing Ion Torrent



## Different platforms

63/165

### Next Generation Sequencing: Amplified Single Molecule Sequencing Ion Torrent



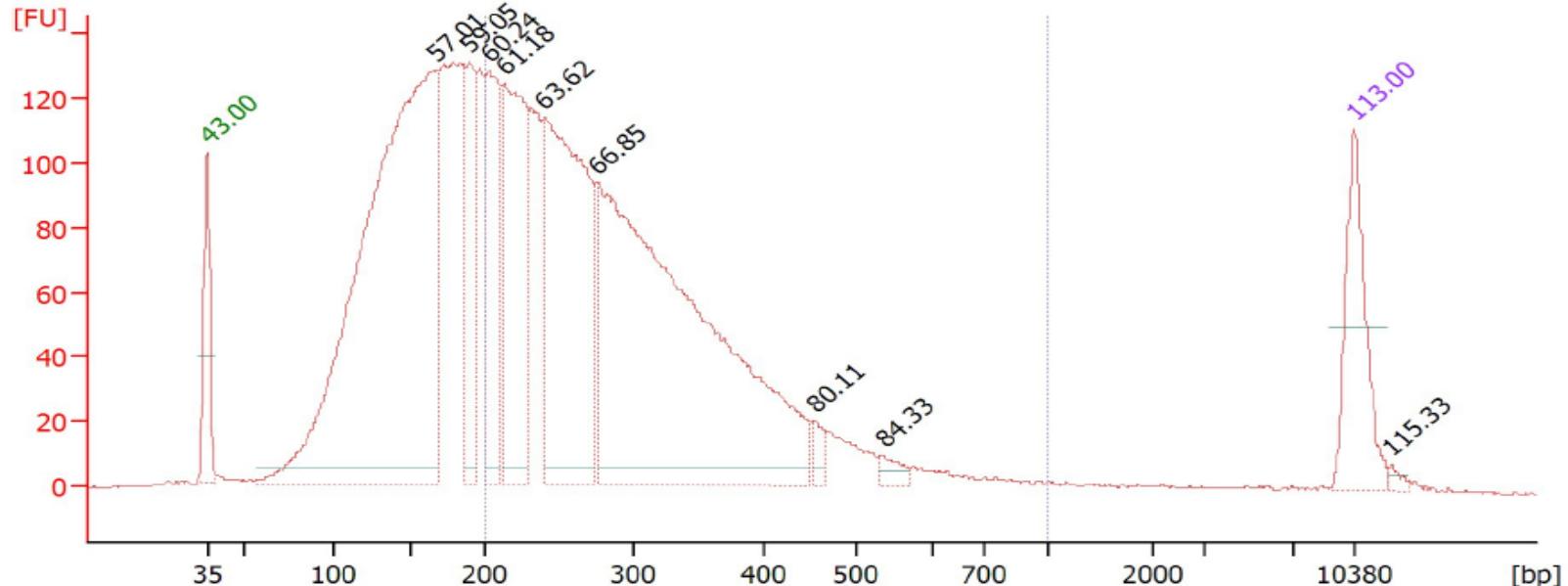
ATCGTGTTCAGGGTCCCCGGGGTT...

## Different platforms

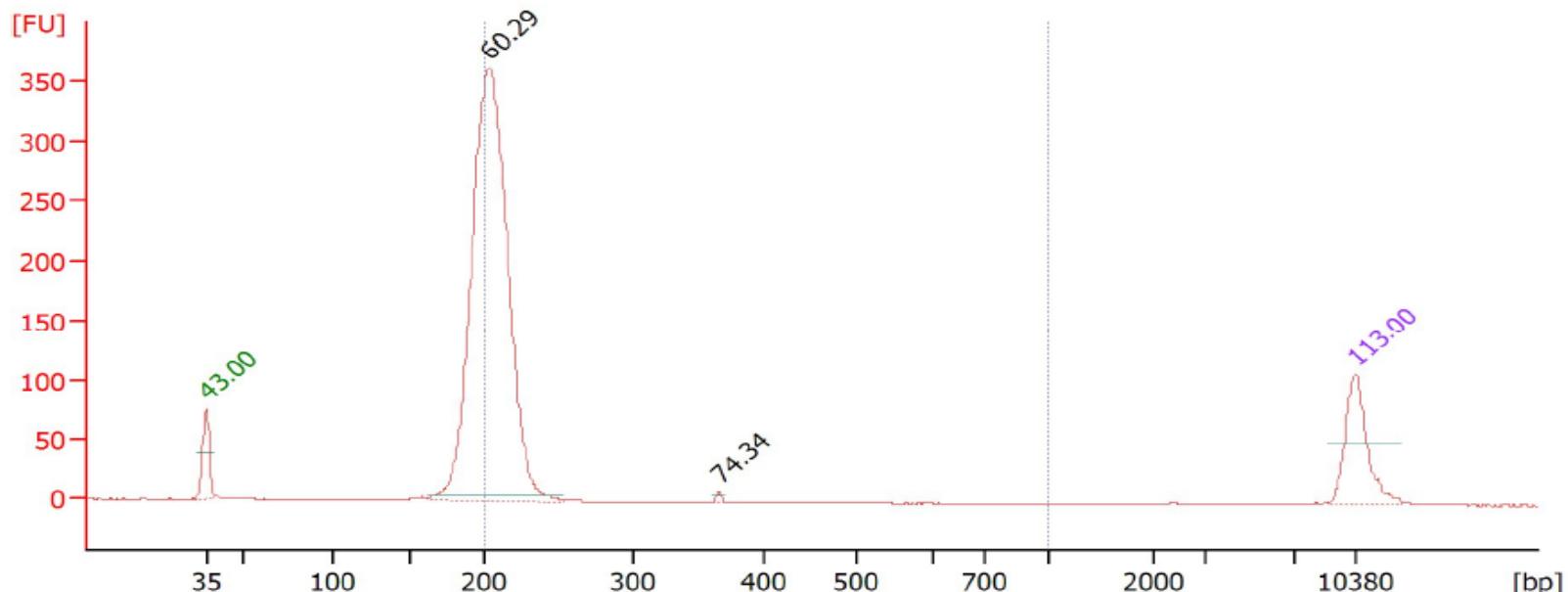
### Next Generation Sequencing: Amplified Single Molecule Sequencing

Ion Torrent

Size selection -> maximizing sequencing yield



green

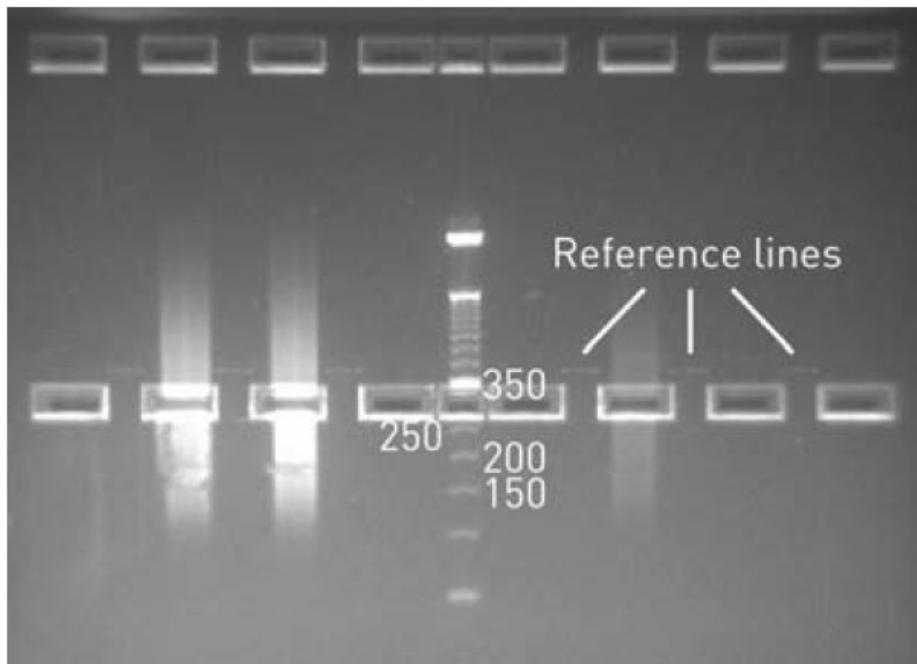


black

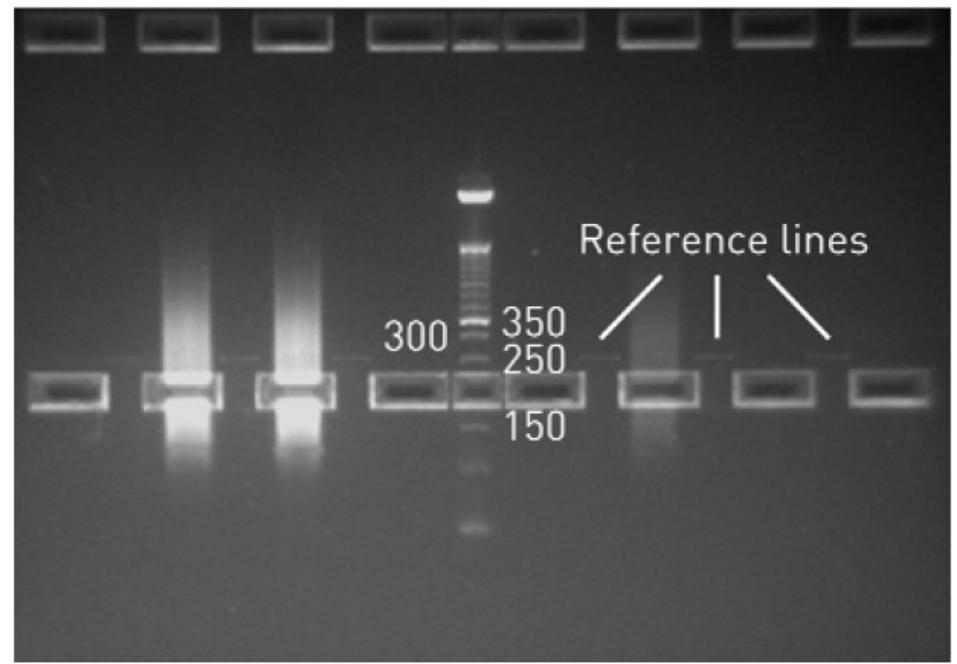
## Different platforms

### Next Generation Sequencing: Amplified Single Molecule Sequencing Ion Torrent

Size selection -> maximizing sequencing yield



200 base-read library gel



100 base-read library gel

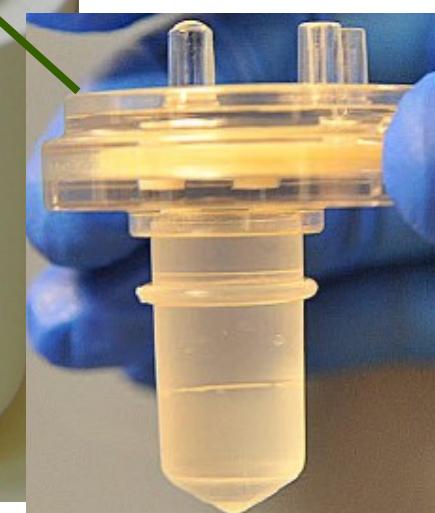
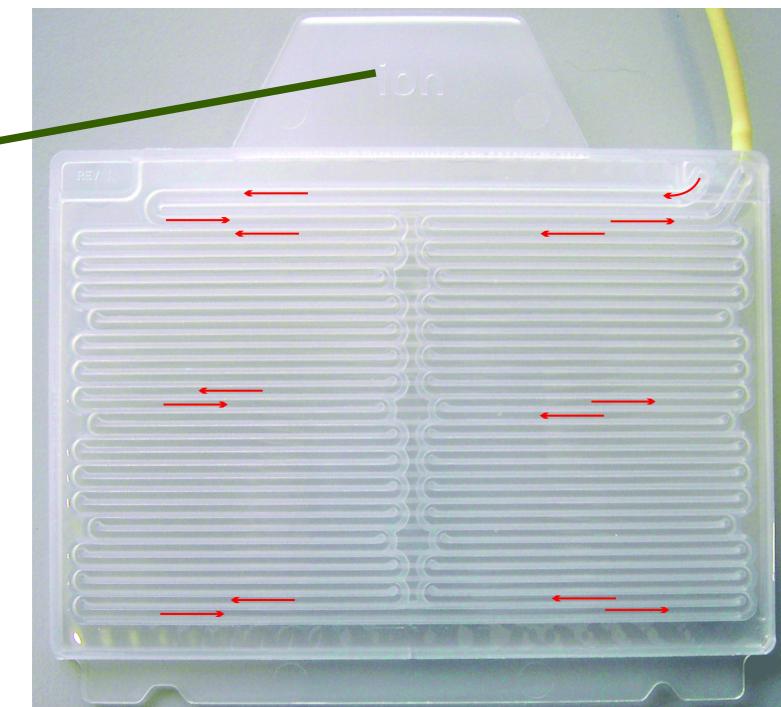
## Different platforms

### Next Generation Sequencing: Amplified Single Molecule Sequencing

Ion Torrent



Emulsion PCR: OneTouch Instrument



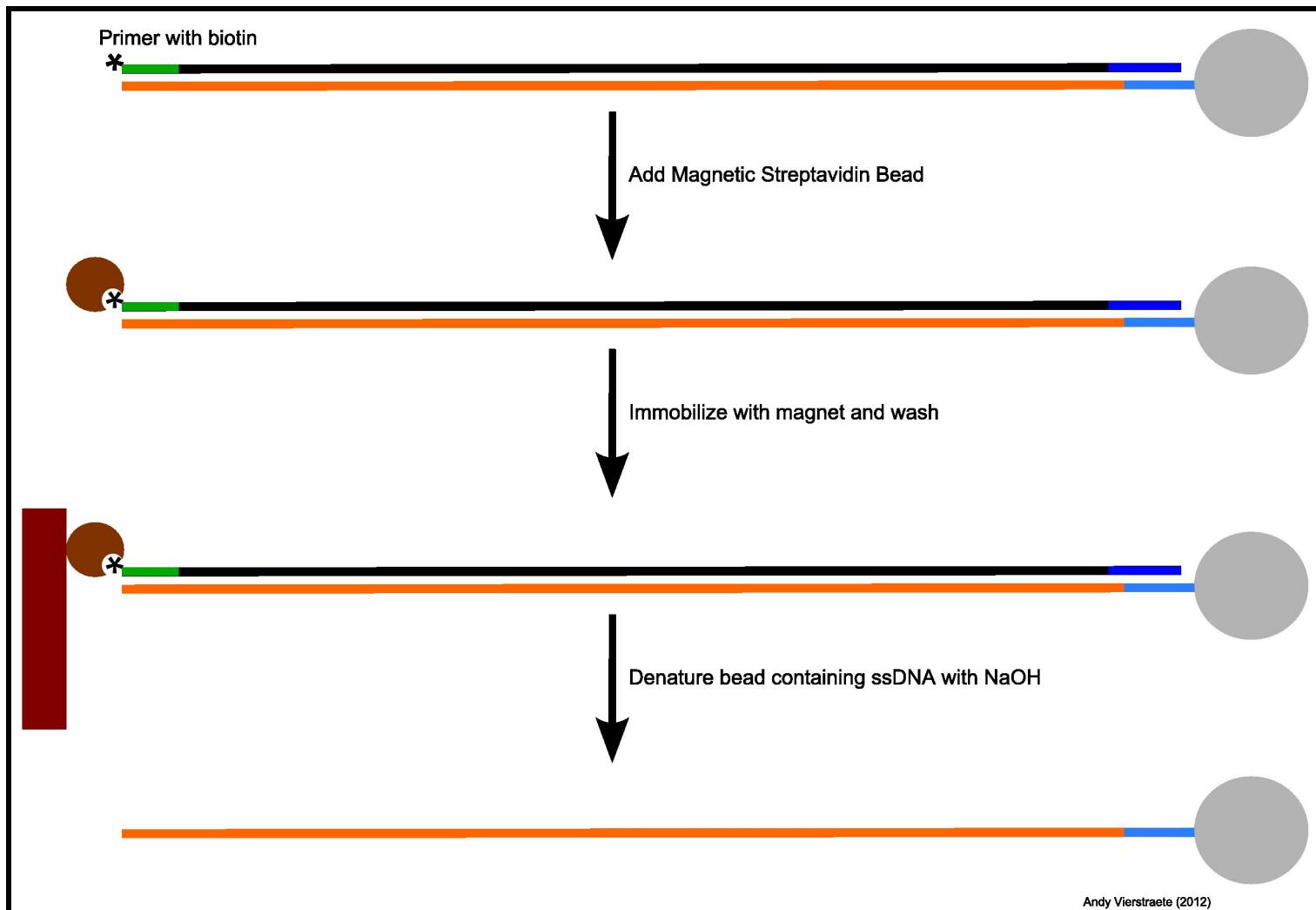
15 min hands-on; 4-8 hours amplification; 35 min enrichment

## Different platforms

67/165

### 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

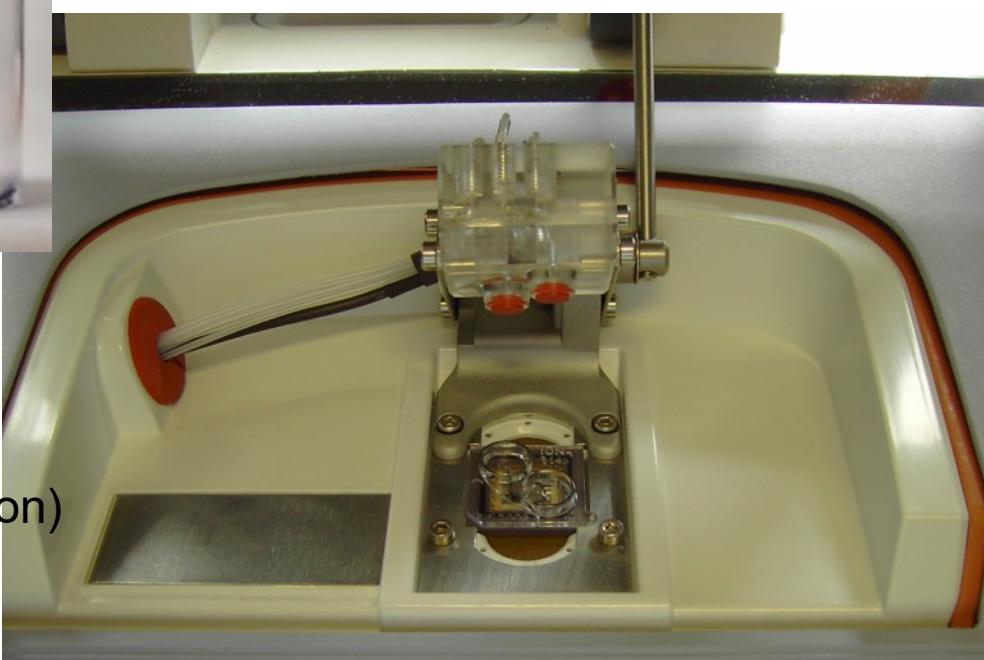
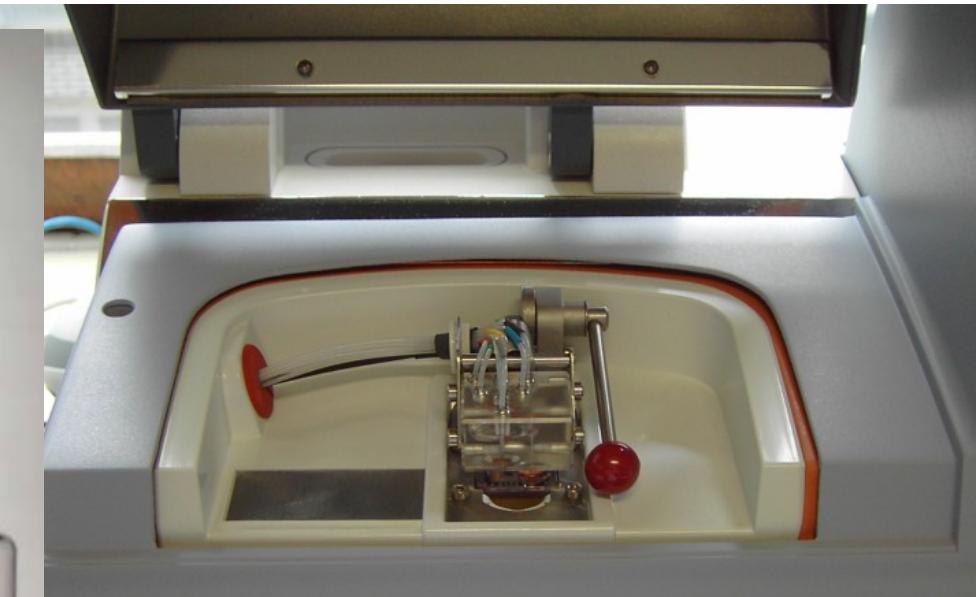


W1 bottle: 350 $\mu$ l 100 mM NaOH

W2 bottle: 2 liter W2 solution (contains PCR solution)

W3 bottle: 50 ml W3 (= buffer with known pH)

4 tubes with 20  $\mu$ l of the different dNTP's



## Different platforms

69/165

### Next Generation Sequencing: Amplified Single Molecule Sequencing

#### Ion Torrent

Movie time

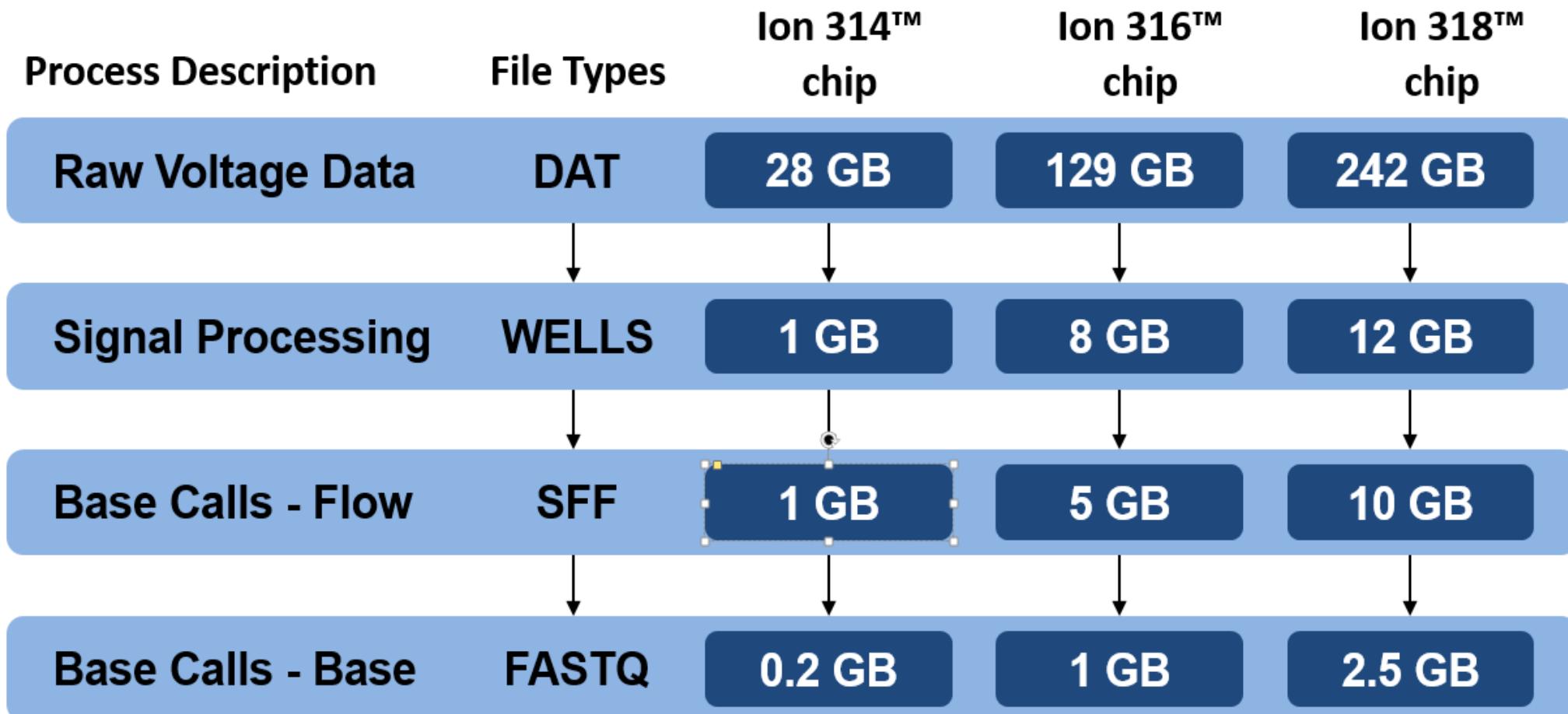
[Ion Torrent Sequencing \(youtube link\)](#)

## Different platforms

### Next Generation Sequencing: Amplified Single Molecule Sequencing

#### Ion Torrent

Torrent Server pipeline



## 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

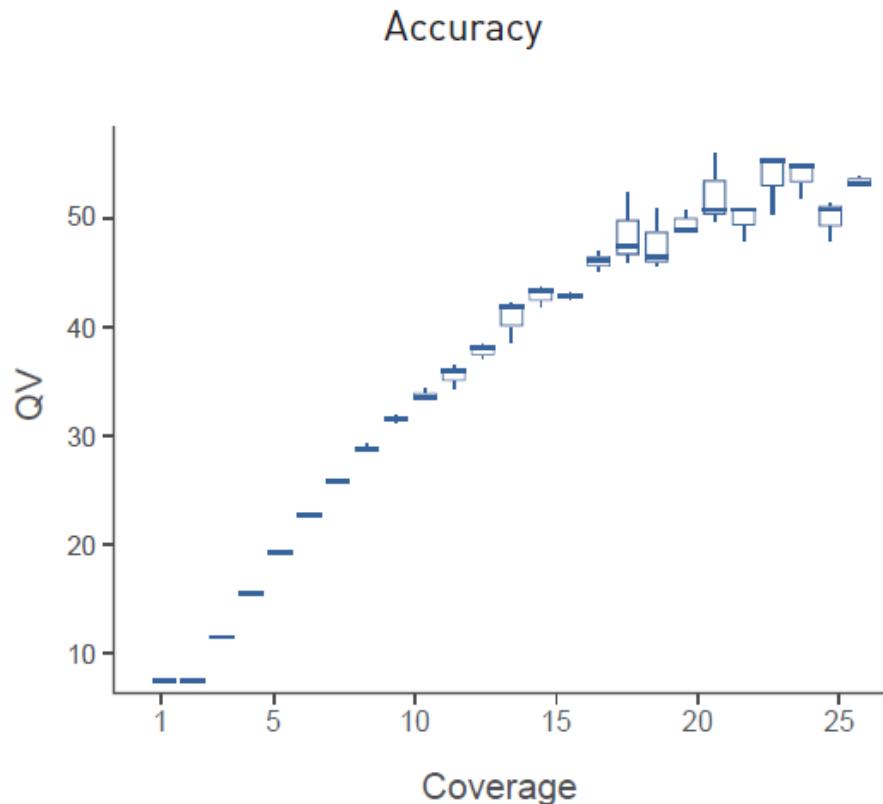


	Pacbio RS	Sequel System
<b>Read Length</b>	50 % > 20 kb (max > 60 kb)	50 % > 20 kb (max > 60 kb)
<b>Throughput</b>	1 Gb/SMRT cell (max 16/run)	5-8 Gb/SMRT cell (max 16/run)
<b>Reads per run</b>	55,000	365,000
<b>Accuracy</b>	86 %	86 %
<b>Run Time</b>	30 minutes – 6 hours/ SMRT cell	30 minutes – 10 hours/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

## Different platforms

### Third Generation Sequencing: Single Molecule Sequencing

Pacific Biosciences



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 - AAACAAACGC - TCC - AC - TATCAAAT - CCGGGTGCG - CTTGTTGTATAACACAAAC - AGG - CGAAAAAAACATA - TCG - AGTT  
TGCAGATCATTACT - AAACAAACGC - TCC - AC - TATCAAAT - CCGGGTGCG - CTTGTTGTATAACACAAAC - AGG - CGAAAAAAACATA - TCGAAGTT  
TGCAGATCATTACT AAAACAAACGC - TCC - AC - TATCAAAT - CCGGGTGCG AACTTGTGTATAACACAAACT - AGG - CGAAAAAAACATA - TCG - AG - T  
TGCAGATCATTACT - AAACAAACGC - TCC - AC - TATCAAAT - CCGGGTGCG - CTTGTTGTATAACACAAAC - AGG - CGAAAAAAACATA - TCG - AGTT  
TGCAGATCATTACT - AAACAAACGC - TCC - AC - TATCAAAT - CCGGGTGCG - CTTGTTGTATAACACAAAC - AGG - CGAAAAAAACATA - TCG - AGTT  
TGCAGATCATTACT - AAACAAACGC - TCC - AC - TATCAAAT - CCGGGTGCG - CTTGTTGTATAACACAAAC - AGG - CGAAAAAAACATA - TCG - AGTT  
TGCAGATCATTACT - AAACAAACGC - TCC - AC - TATCAAAT - CCGGGTGCG - CTTGTTGTATAACACAAAC - AGG - CGAAAAAAACATA - TCG - AGTT  
TGCAGATCATTACT - AAACAAACGC - TCC - AC - TATCAAAT - CCGGGTGCG - CTTGTTGTATAACACAAAC - AGG - CGAAAAAAACATA - TCG - AGTT  
TGCAGATCATTACT - AAACAAACGC - TCC - AC - TATCAAAT - CCGGGTGCG - CTTGTTGTATAACACAAAC - AGG - CGAAAAAAACATA - TCG - AGTT  
TGCAGATCATTACT - AAACAAACGC - TCC - AC - TATCAAAT - CCGGGTGCG - CTTGTTGTATAACACAAAC - AGG - CGAAAAAAACATA - TCG - AGTT  
ACGCGTCCGTAC - TATC - AAT - CCGGGTGCG - C - TGTTGTATAACACAAAC TAGG - CGAAAAAAACATA - TCG - AGTT

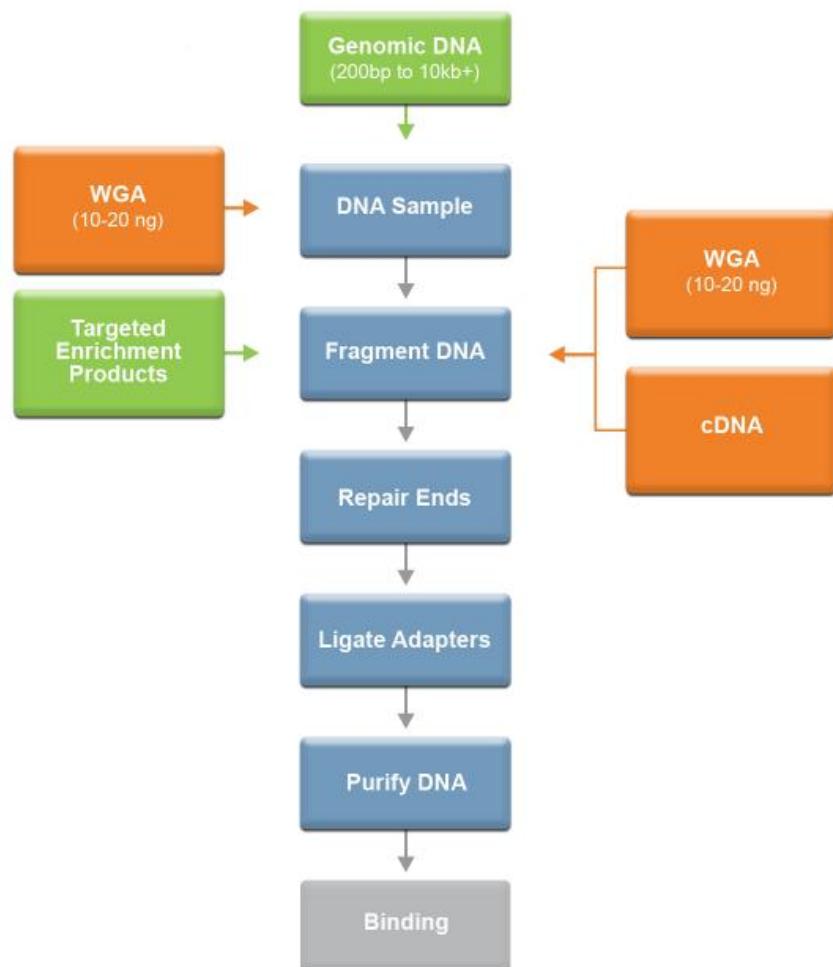
## Different platforms

### Third Generation Sequencing: Single Molecule Sequencing

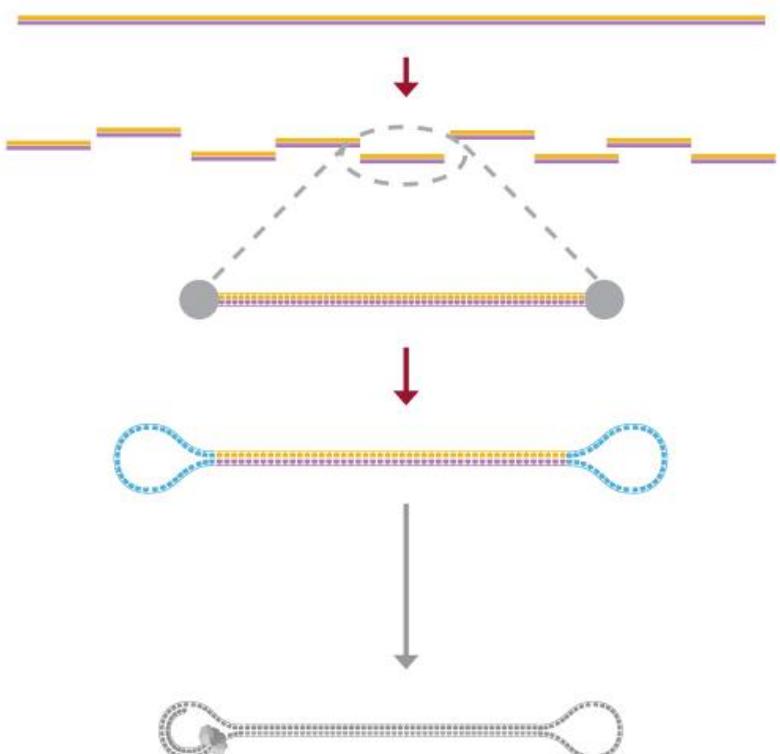
Pacific Biosciences

Workflow: Library preparation → Sequencing

#### Sample Preparation



#### Building of SMRTbell

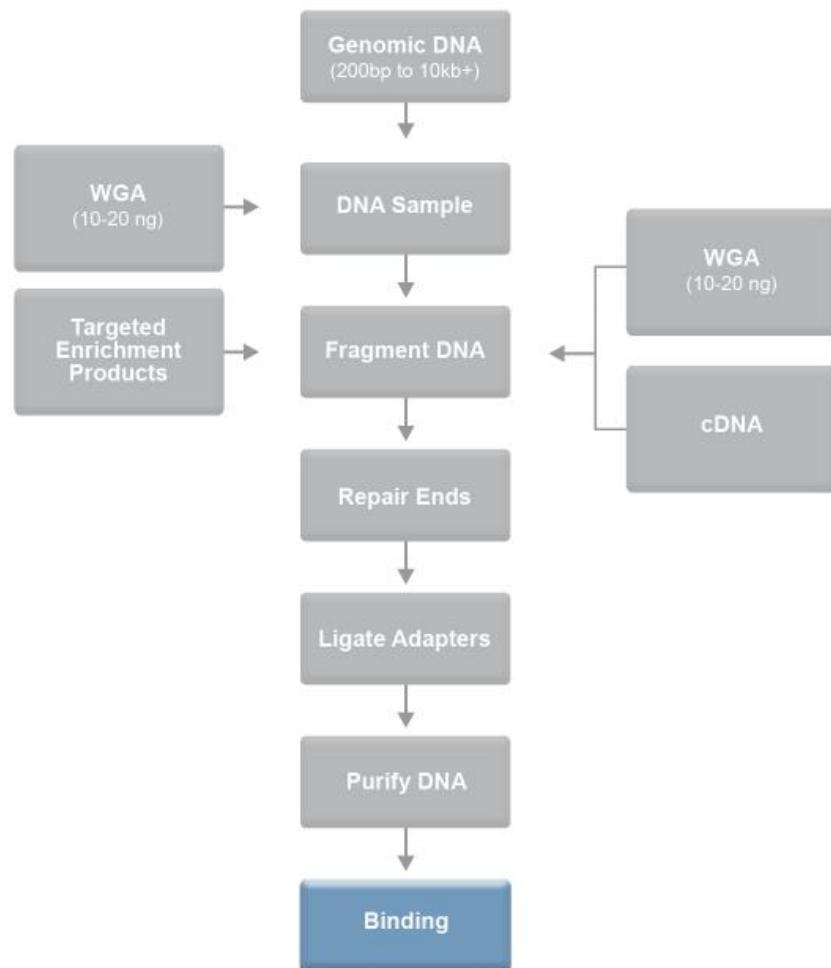


## Different platforms

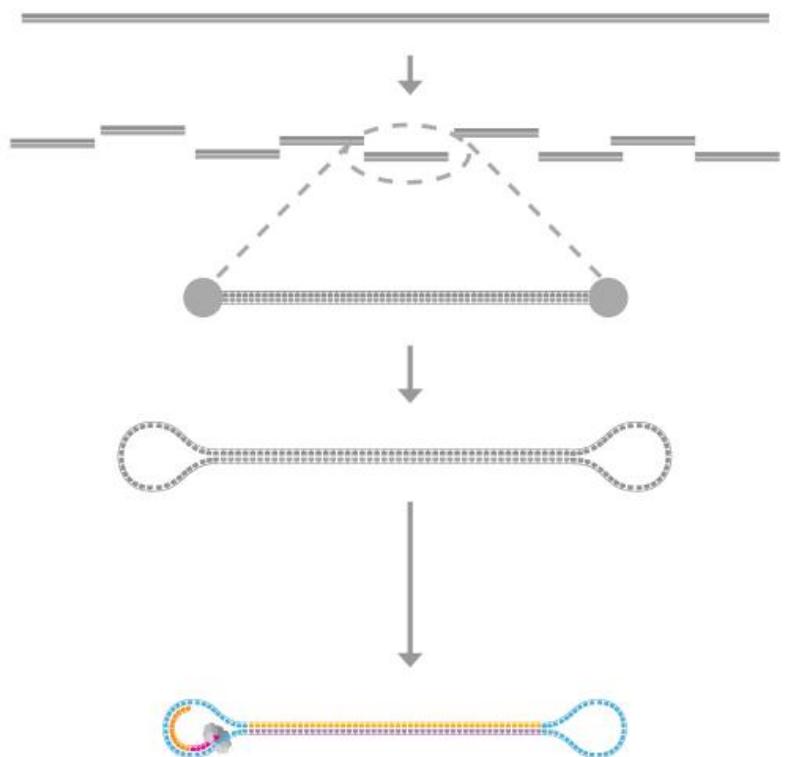
### Third Generation Sequencing: Single Molecule Sequencing

Pacific Biosciences

#### Sample Preparation



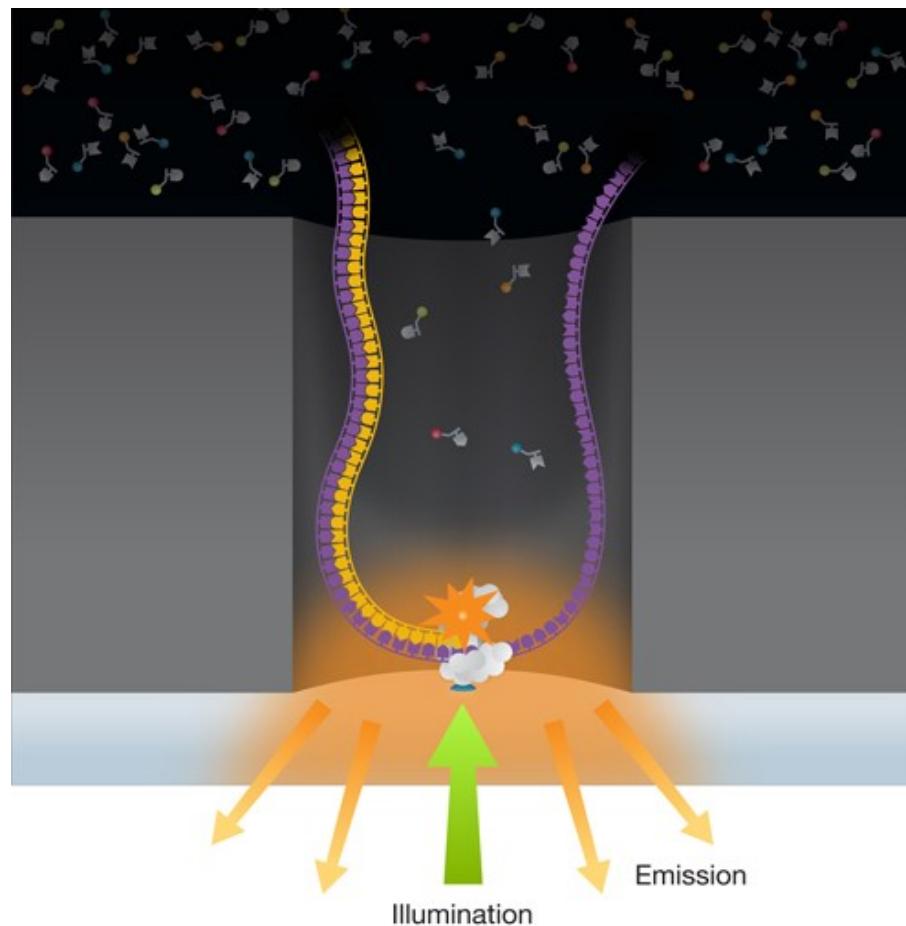
#### Building of SMRTbell



## Different platforms

### Third Generation Sequencing: Single Molecule Sequencing

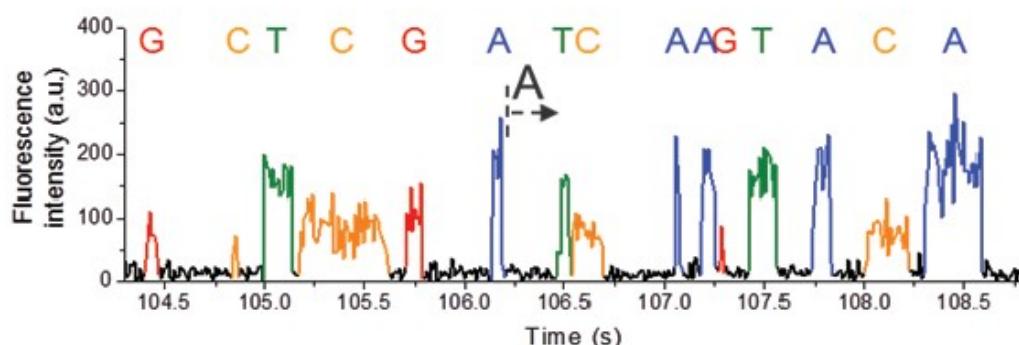
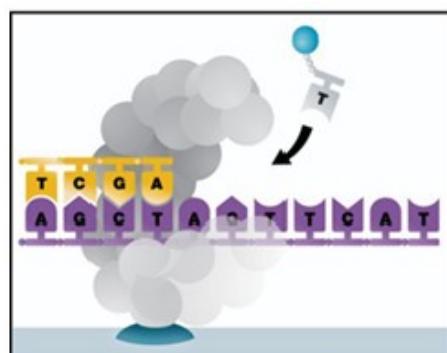
Pacific Biosciences



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

78/165

### 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
<b>Read Length</b>	?	> 200 kb (record: 1,2 Mb)		
<b>Throughput</b>	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
<b>Reads per run</b>	?	10,000 – >300,000		
<b>Accuracy</b>		90 % (1D) – 96 %(1D <sup>2</sup> )		
<b>Run Time</b>	1 – 4 hours	1 - 48 (70) hours	1 – 48 hours	1 - 48 hours

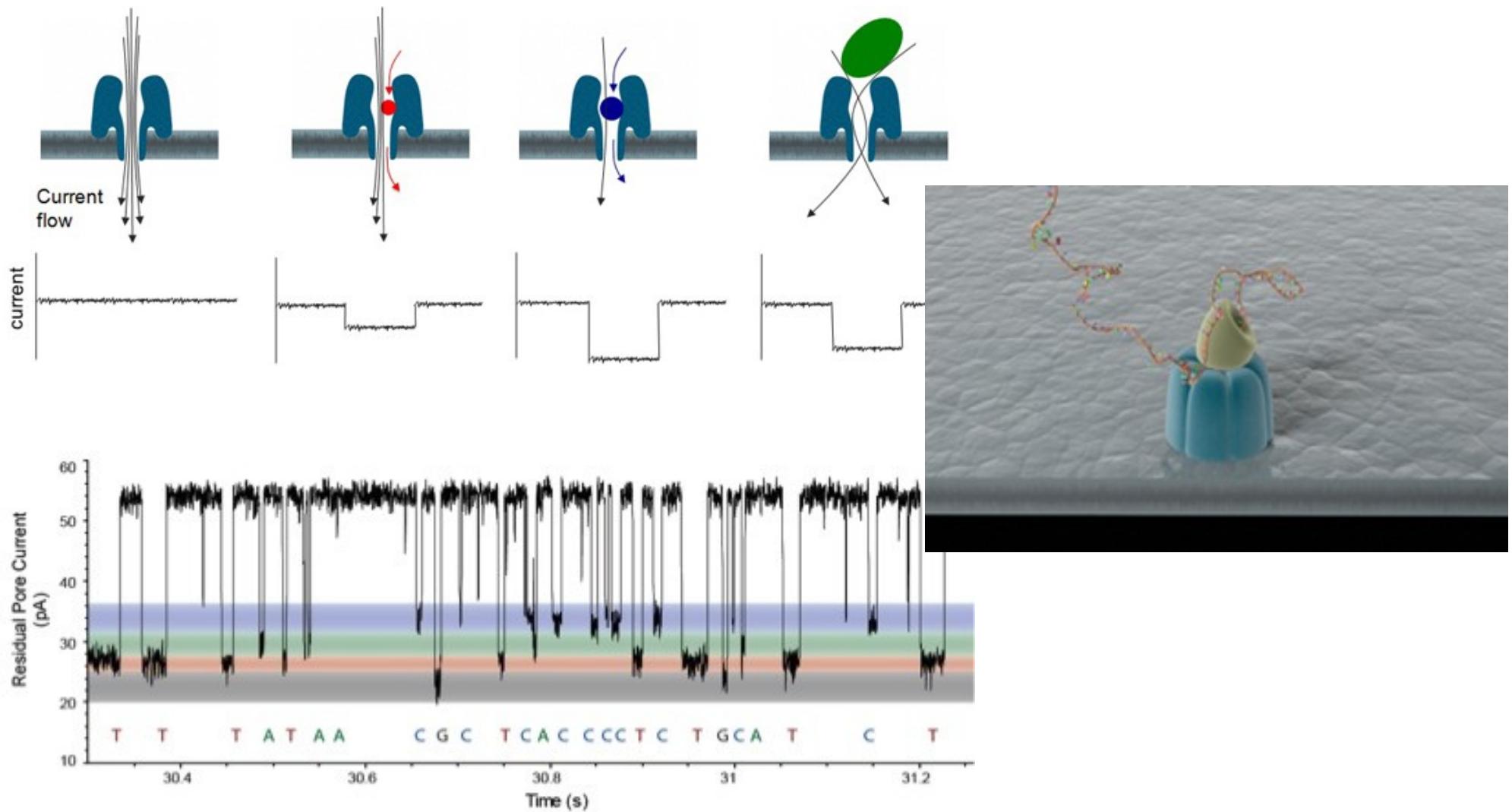
consensus accuracy improved to 99.5% at 30× coverage

## Different platforms

80/165

### Third Generation Sequencing: Single Molecule Sequencing

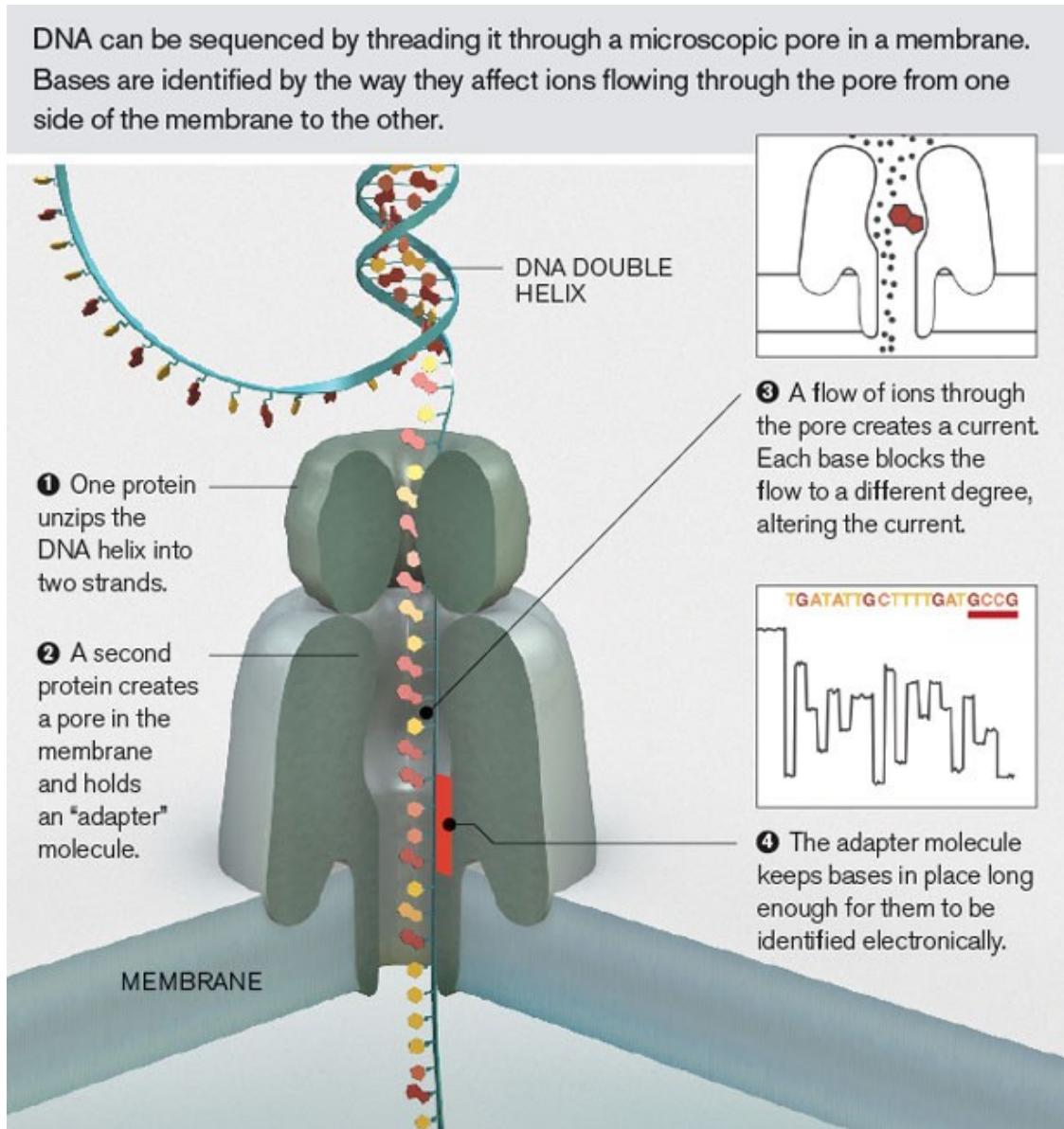
#### Oxford Nanopore



## Different platforms

### Third Generation Sequencing: Single Molecule Sequencing

#### Oxford Nanopore



R9.4: 450 nucleotides/second

## Different platforms

### Third Generation Sequencing: Single Molecule Sequencing

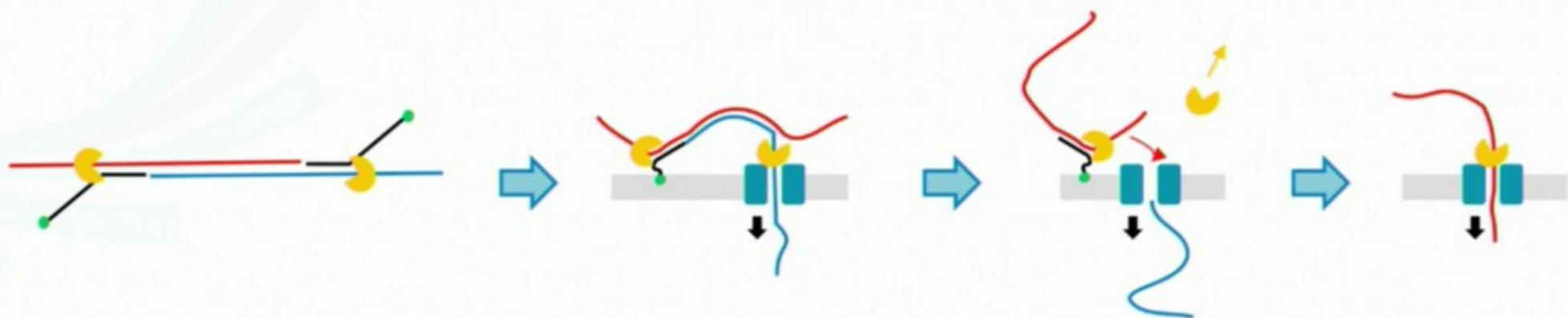
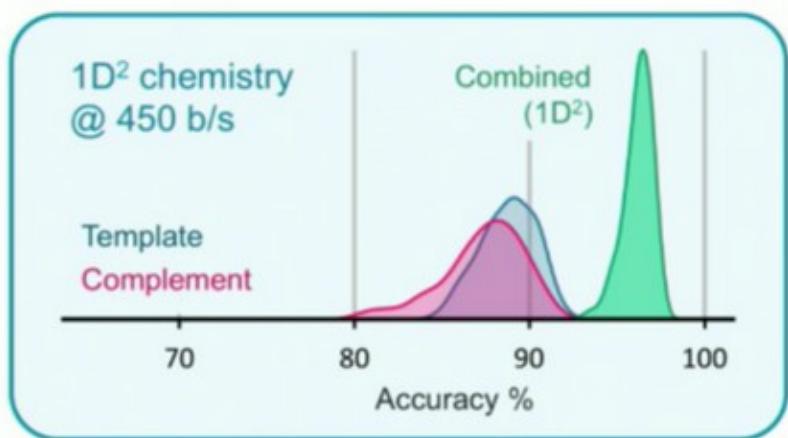
Oxford Nanopore

1D<sup>2</sup>

Improved template – complement data

#### SEQUENCING SCHEME WHERE STRANDS ARE NOT JOINED

- Complement follows template as separate independent strand
  - Each molecule has its own motor-adapter
  - 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



## Different platforms

83/165

### 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



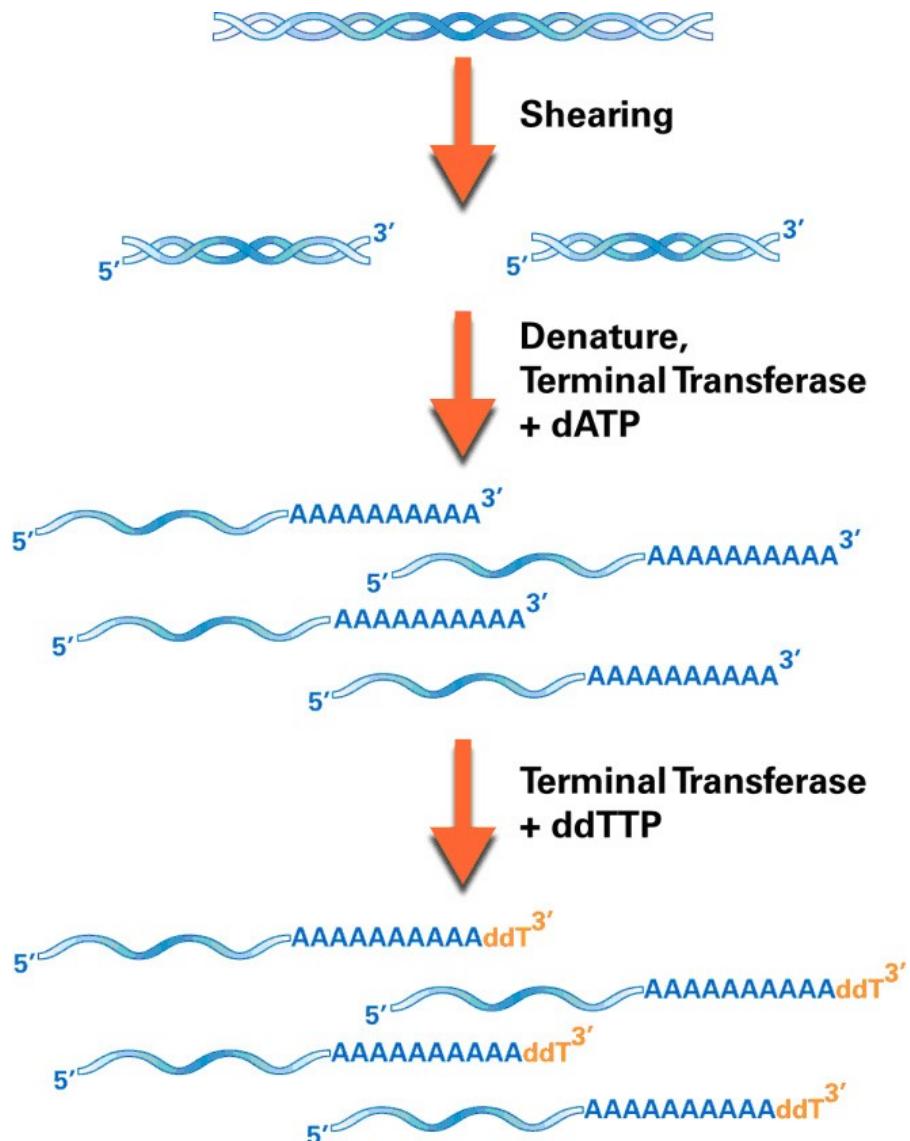
	tSMS Sequencer
<b>Read Length</b>	20-55 bp
<b>Throughput</b>	21-35 Gb/run (2x25 channels/run)
<b>Reads per run</b>	600 – 1000 Million
<b>Accuracy</b>	95-96 % ?
<b>Run Time</b>	30 hours ?

## Different platforms

85/165

### Third Generation Sequencing: Single Molecule Sequencing

SeqLL (Sequence the Lower Limit)

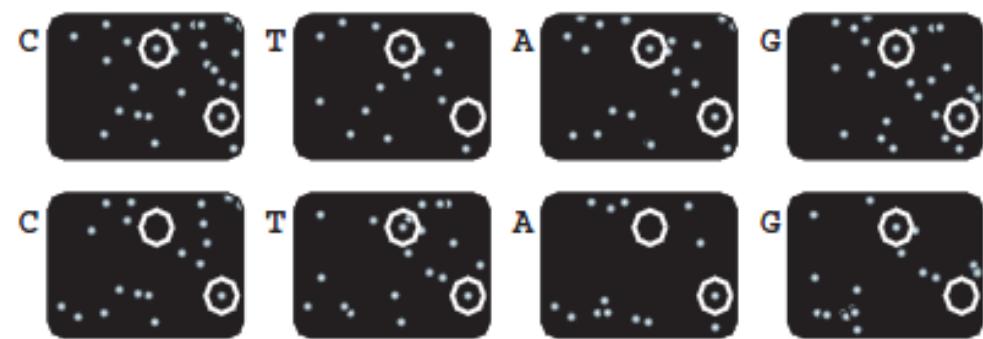
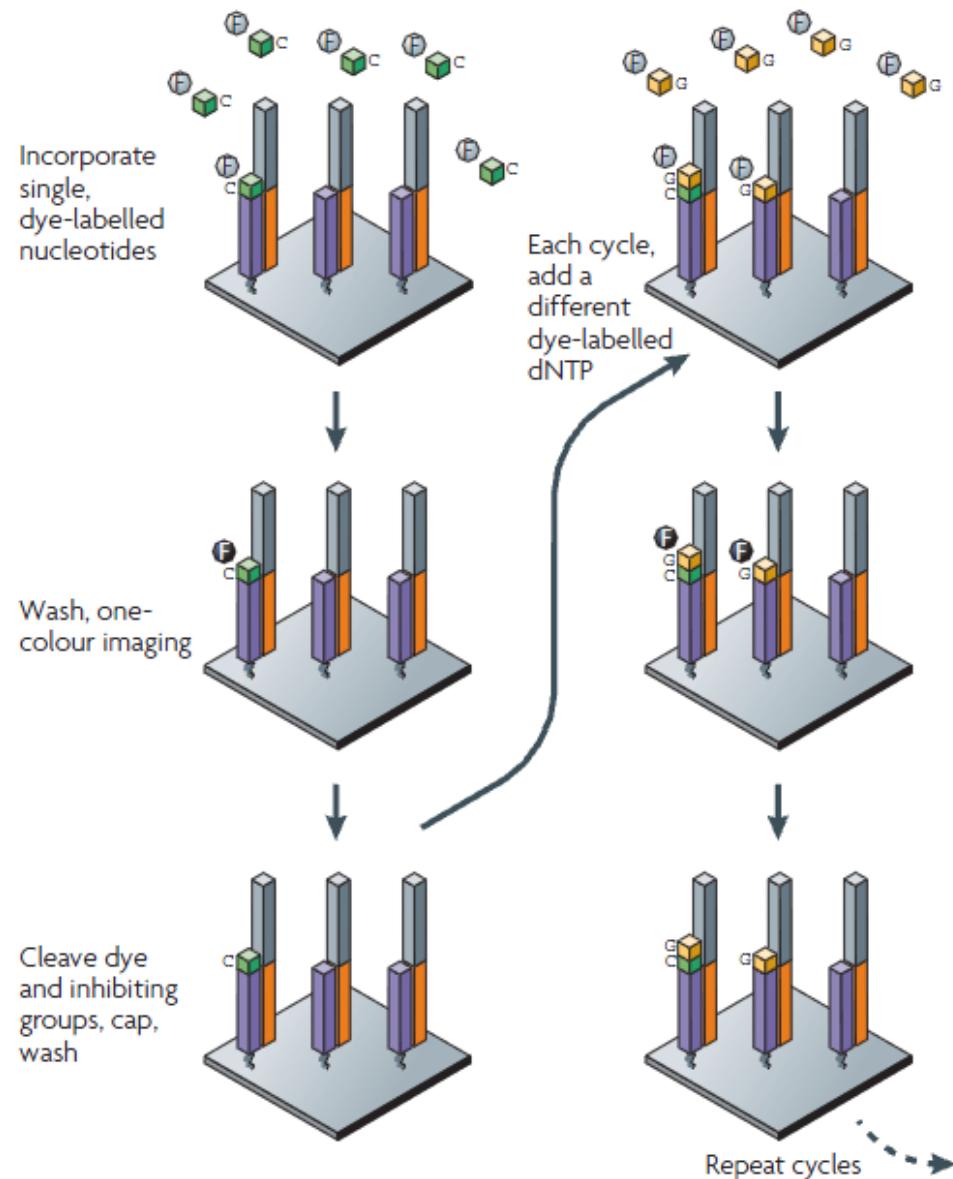


## Different platforms

### Third Generation Sequencing: Single Molecule Sequencing

SeqLL (Sequence the Lower Limit)

Sequencing by synthesis



Nucleotides flown sequentially

## Different platforms

87/165

### Third Generation Sequencing: Single Molecule Sequencing

SeqLL (Sequence the Lower Limit)

Movie time

SeqLL sequencing by synthesis (Youtube link)

## Different platforms

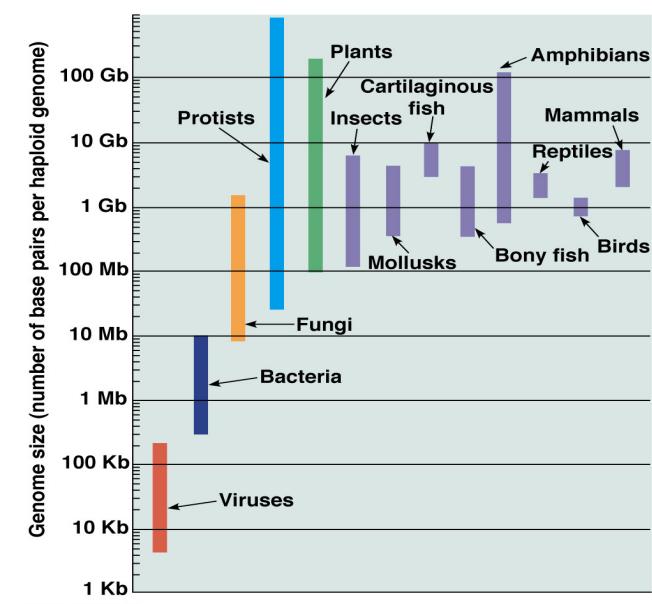
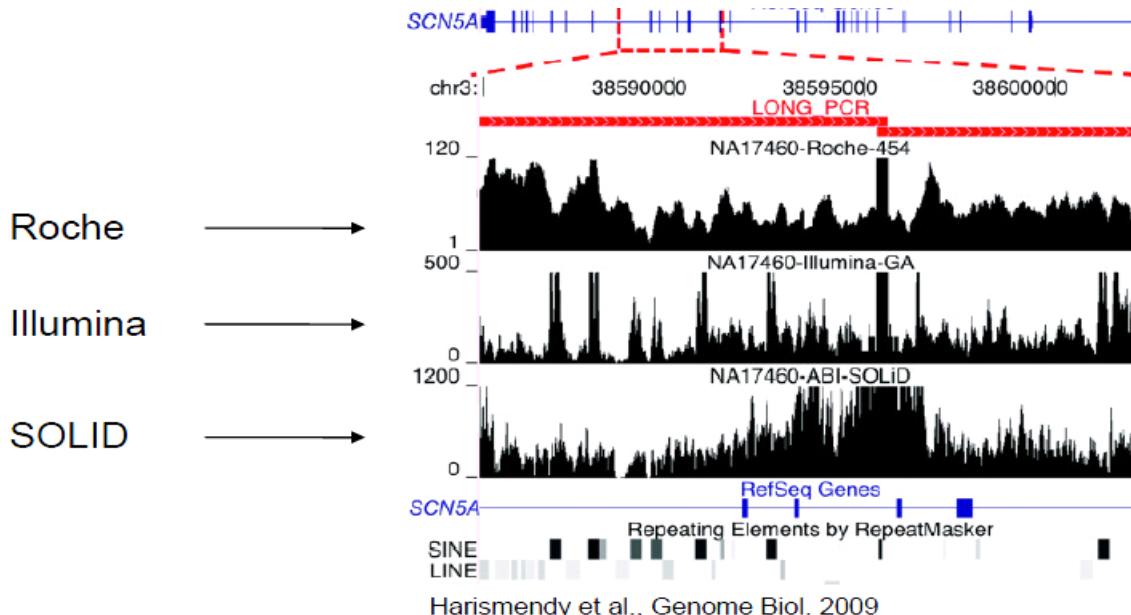
Which Next Generation Sequencer to choose for your project ?

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

90/165

## Sequencing: homopolymer problems

Ion Torrent



# Quality scores in sequencing

## Sequencing: homopolymer problems

Illumina HiSeq



# Quality scores in sequencing

## Sequencing: homopolymer problems

Illumina HiSeq



## Quality scores in sequencing

### Sequencing: homopolymer problems

Illumina MiSeq



# Quality scores in sequencing

94/165

## Sequencing: homopolymer problems

Illumina MiSeq



## Quality scores in sequencing

### 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

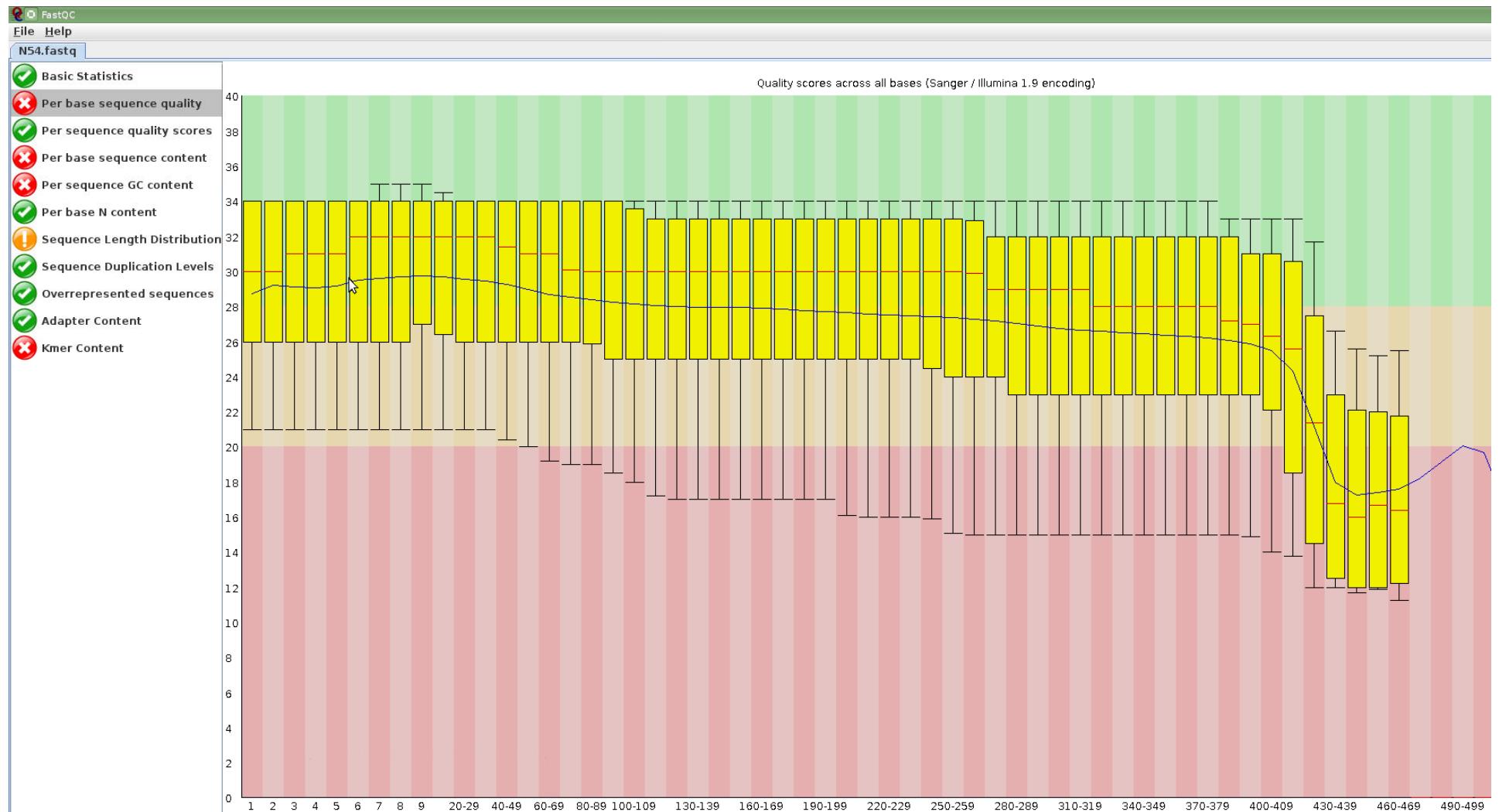
# Quality scores in sequencing

96/165

## Sequencing: quality scores (Phred scores) and accuracy

FastQC : average quality from the sequencing run

Ion Torrent

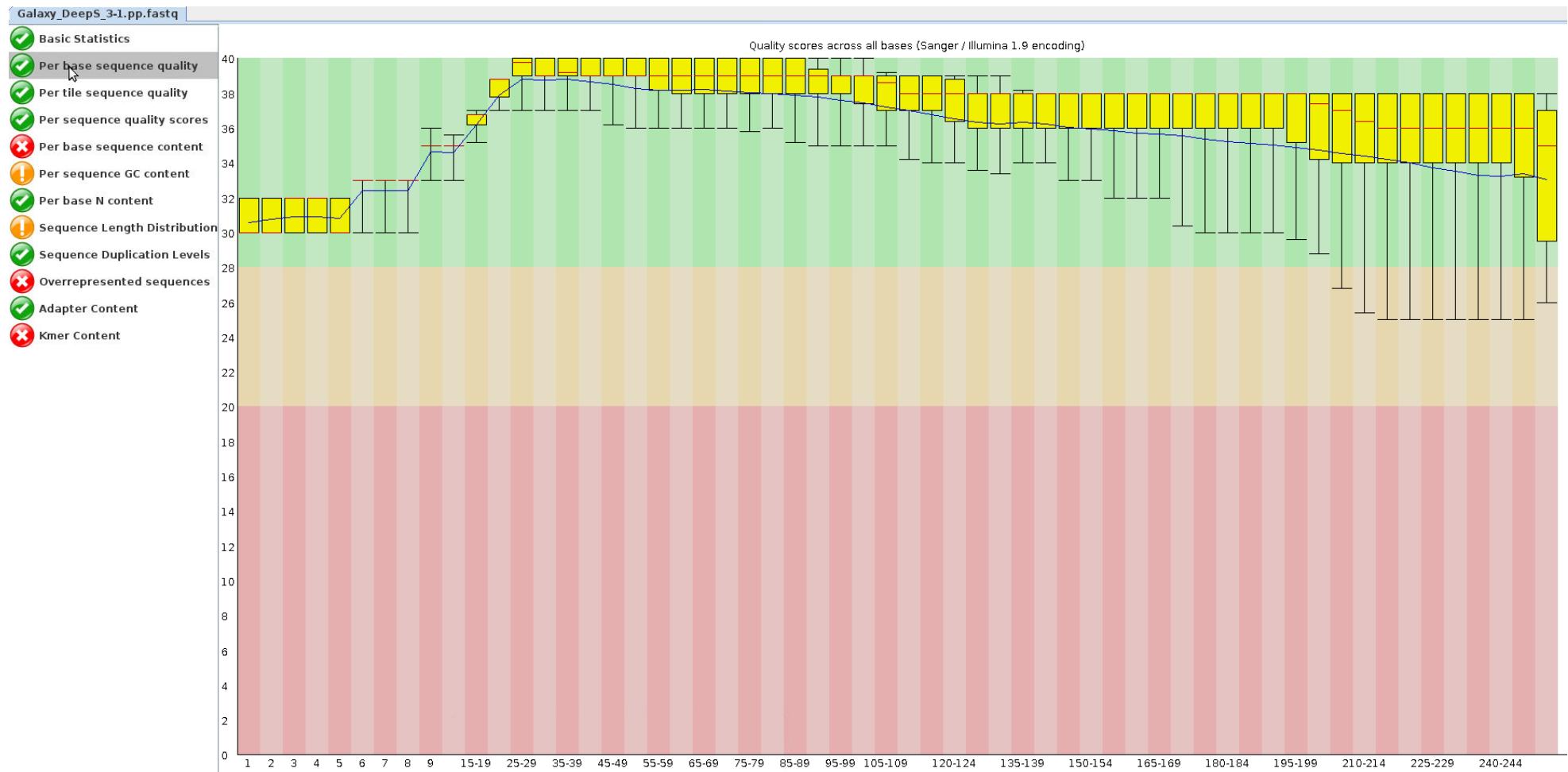


## Quality scores in sequencing

### Sequencing: quality scores (Phred scores) and accuracy

FastQC : average quality from the sequencing run

MiSeq forward sequence

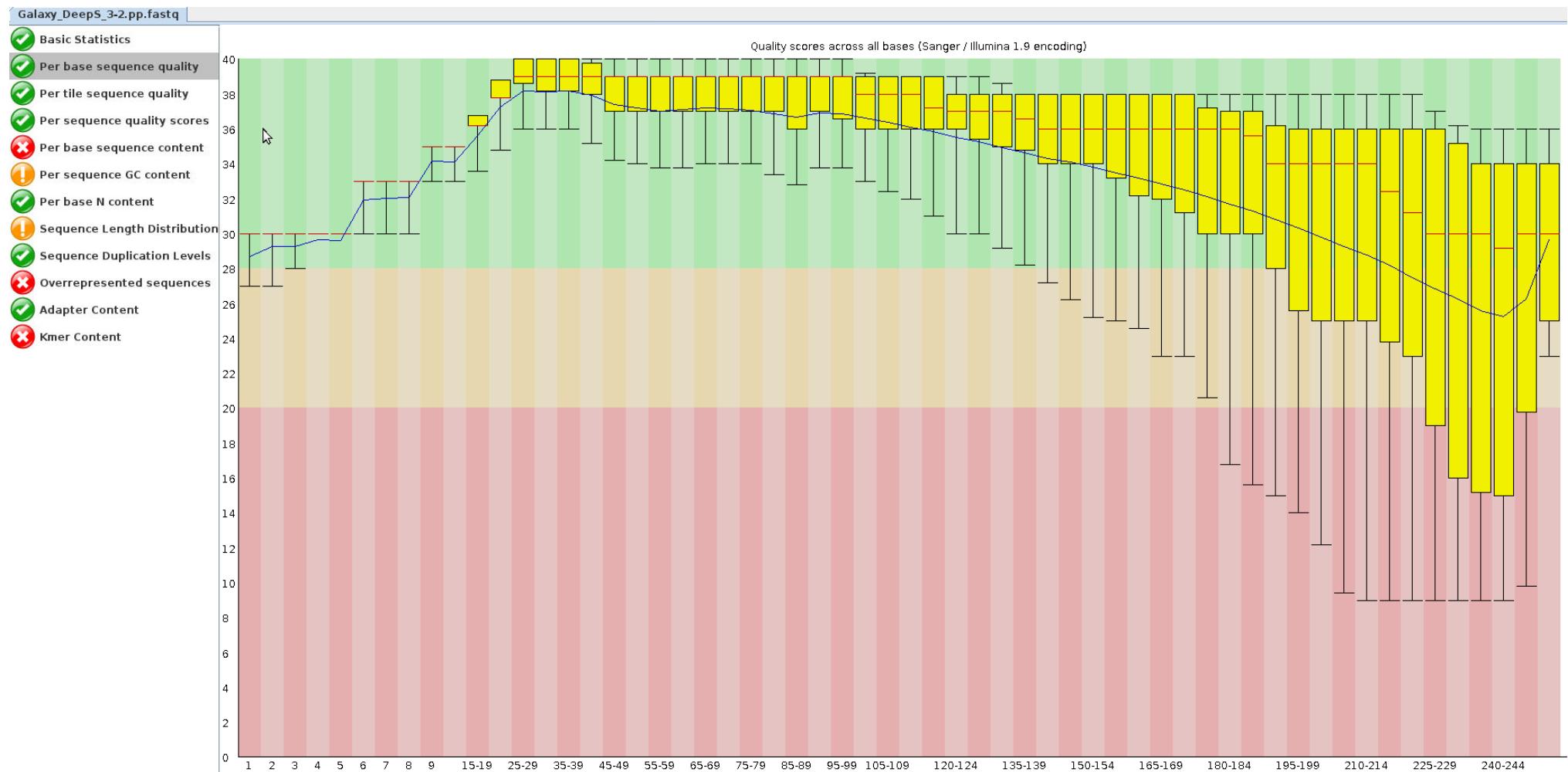


## Quality scores in sequencing

Sequencing: quality scores (Phred scores) and accuracy

FastQC : average quality from the sequencing run

MiSeq reverse sequence

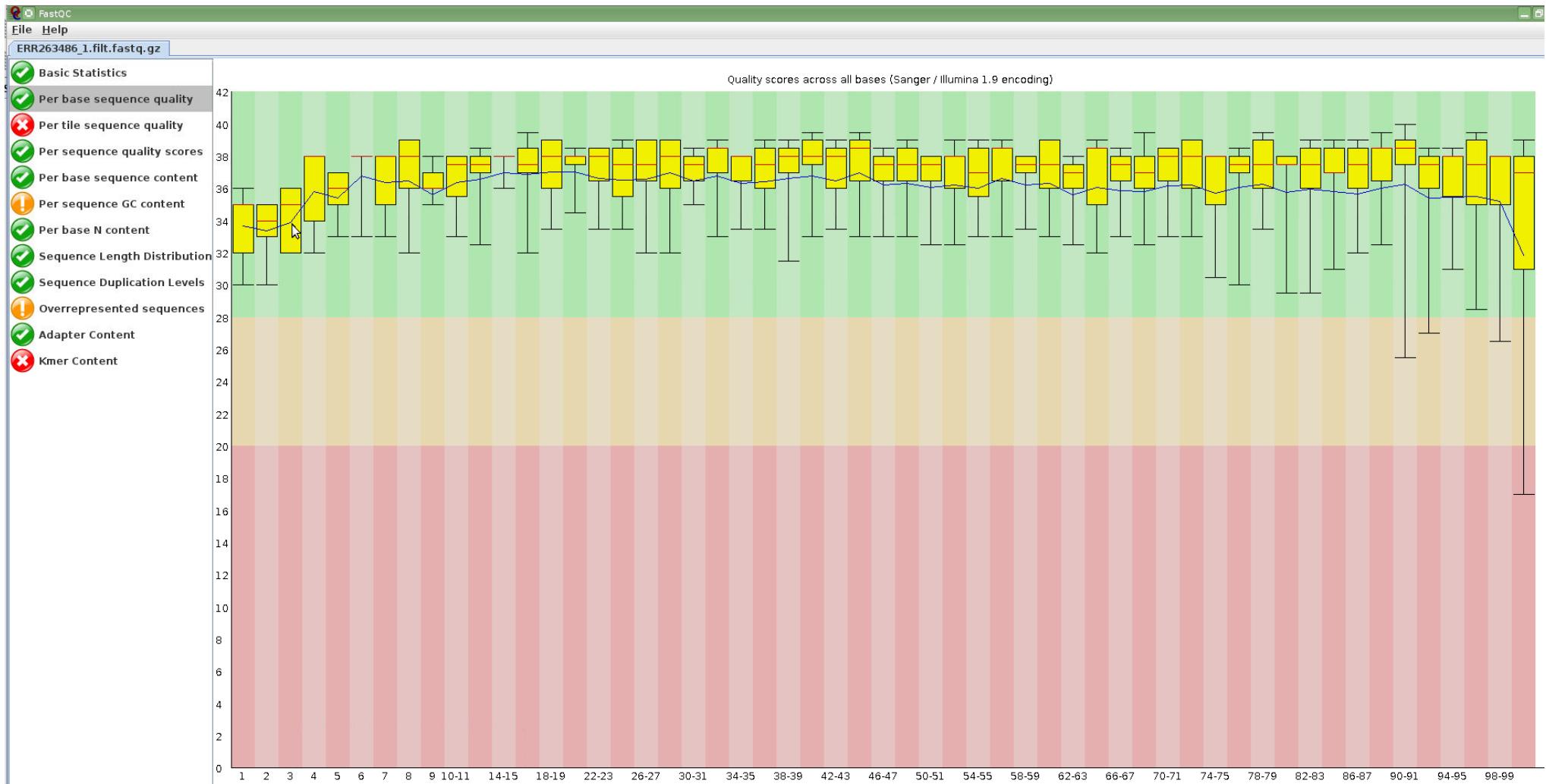


## Quality scores in sequencing

Sequencing: quality scores (Phred scores) and accuracy

FastQC : average quality from the sequencing run

HiSeq



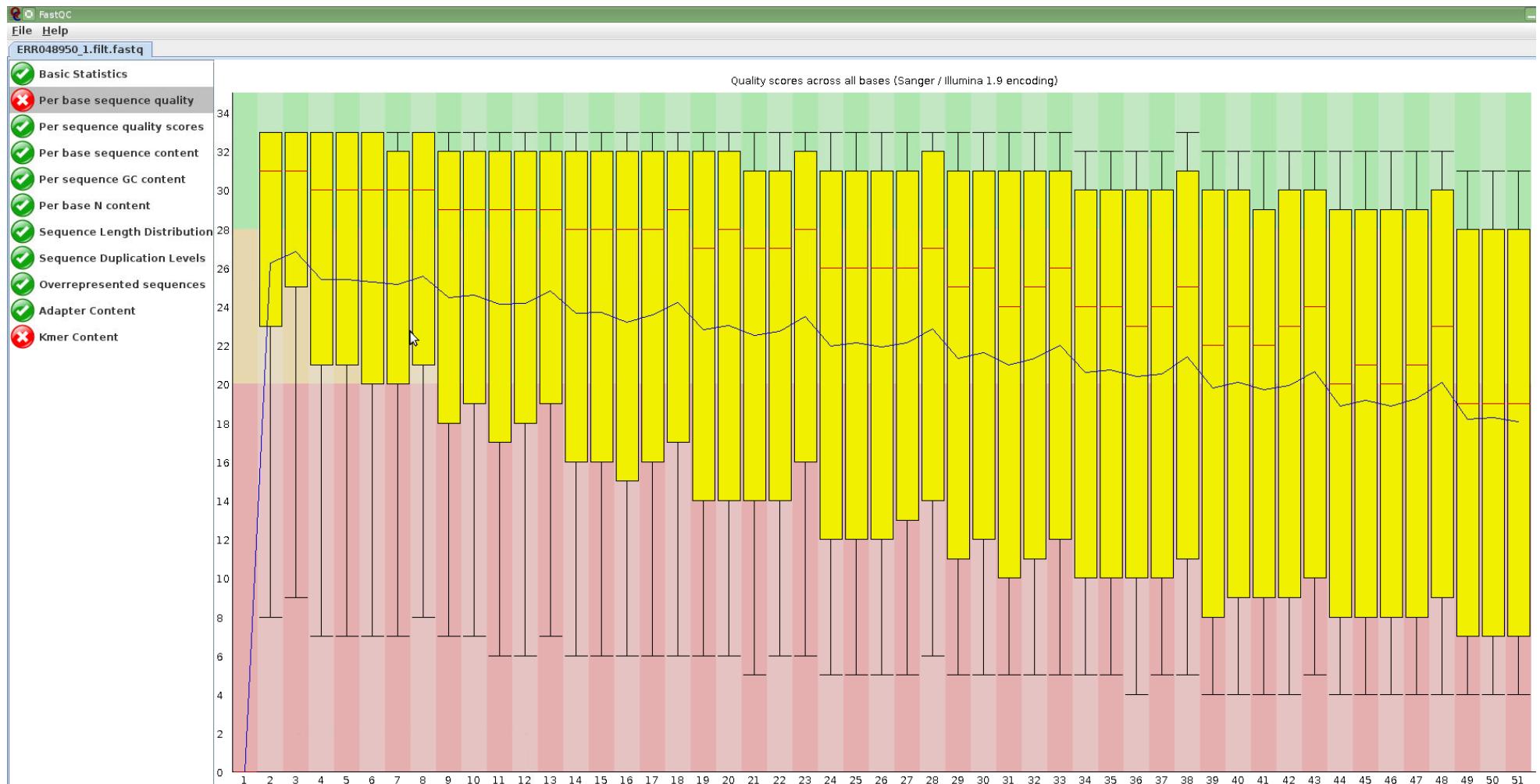
# Quality scores in sequencing

100/165

## Sequencing: quality scores (Phred scores) and accuracy

FastQC : average quality from the sequencing run

SOLiD



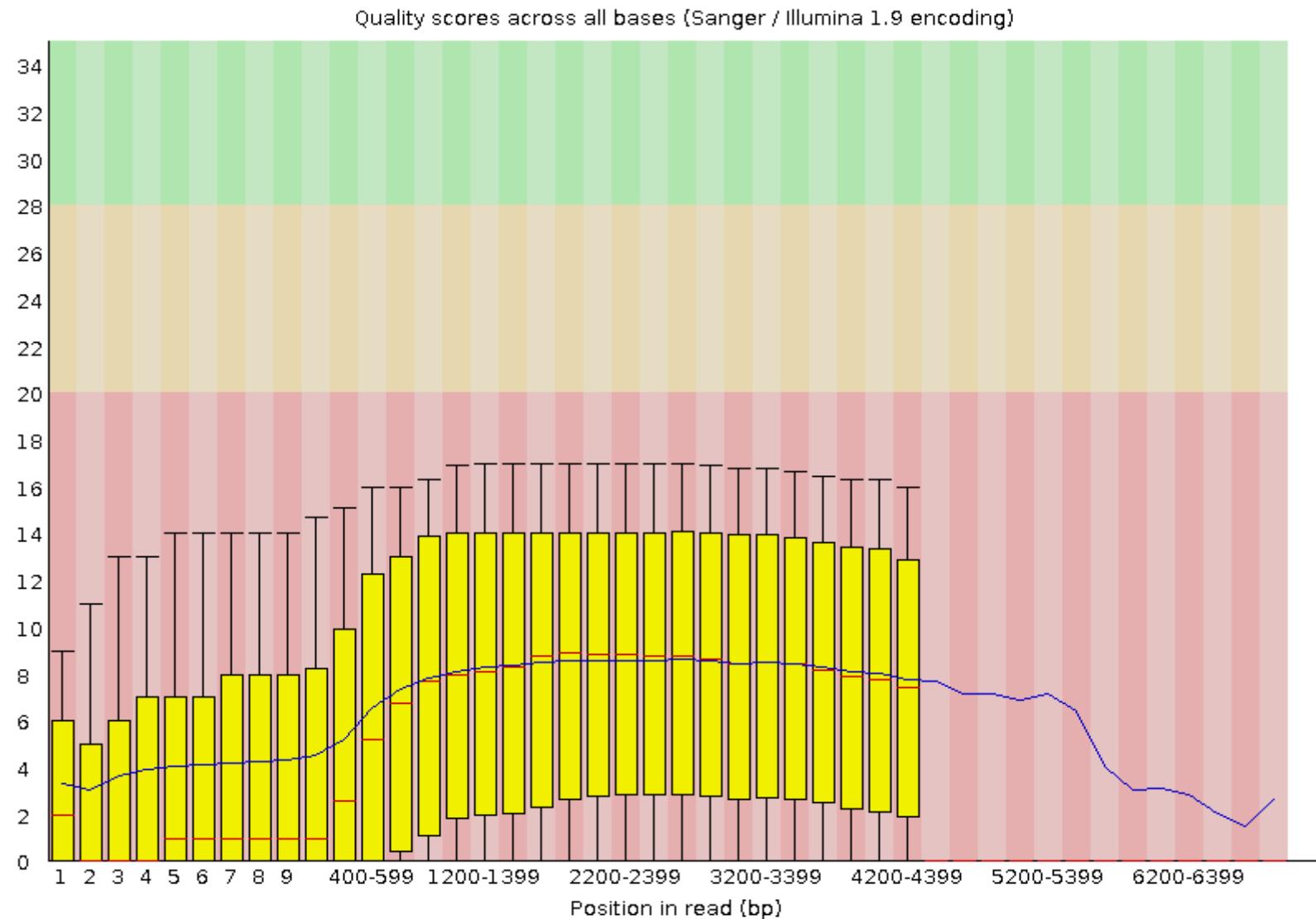
## Quality scores in sequencing

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



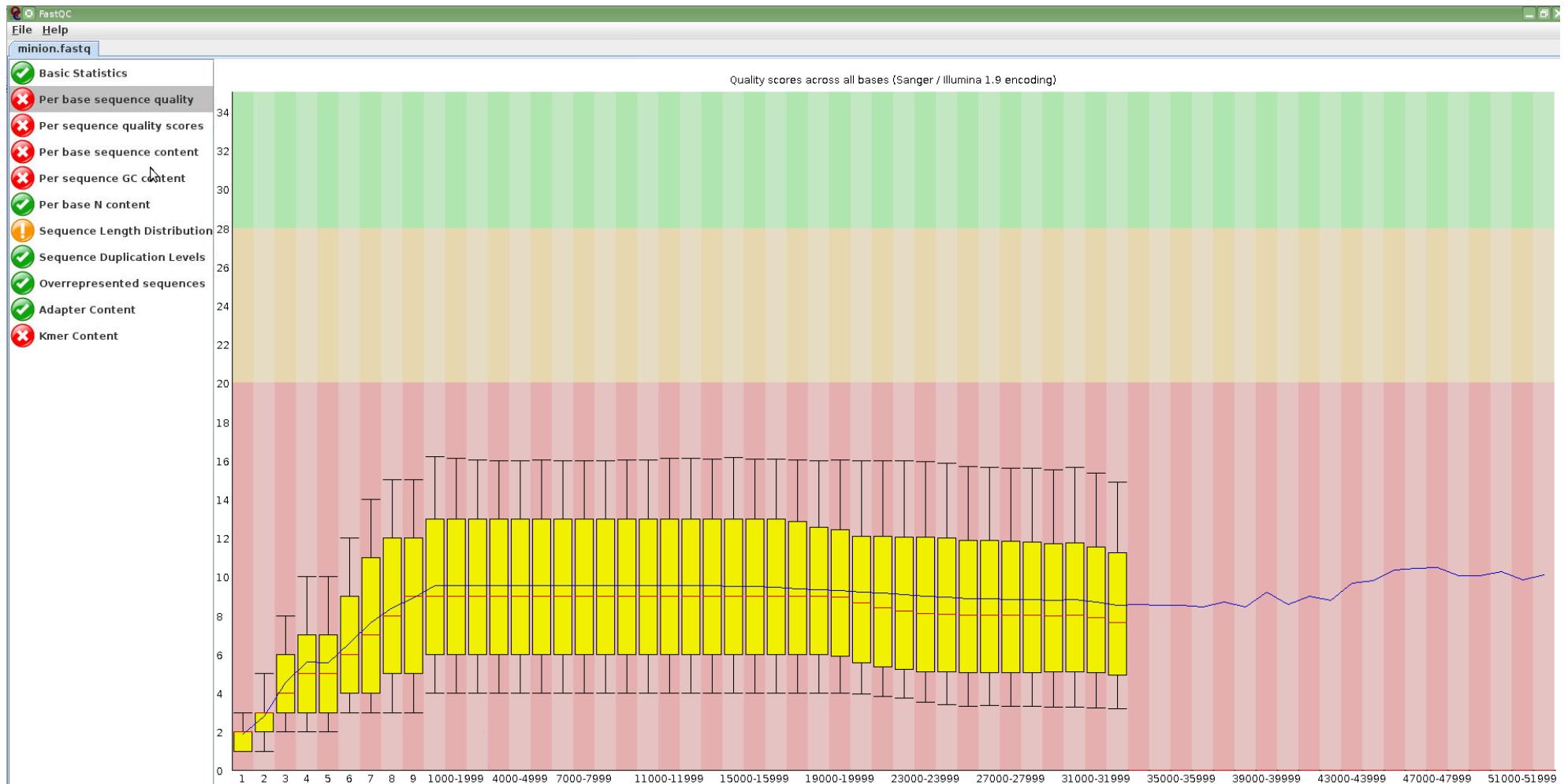
# Quality scores in sequencing

102/165

## Sequencing: quality scores (Phred scores) and accuracy

FastQC : average quality from the sequencing run

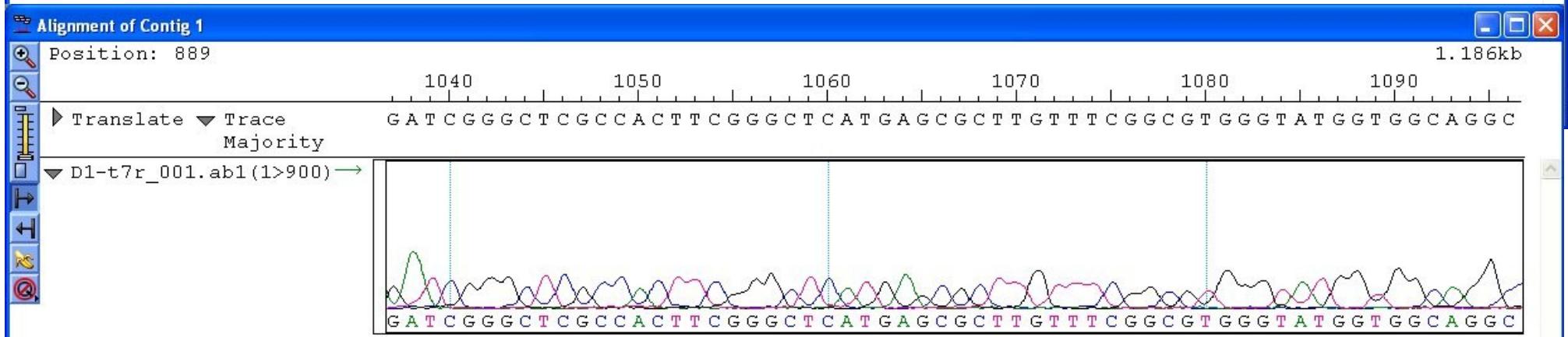
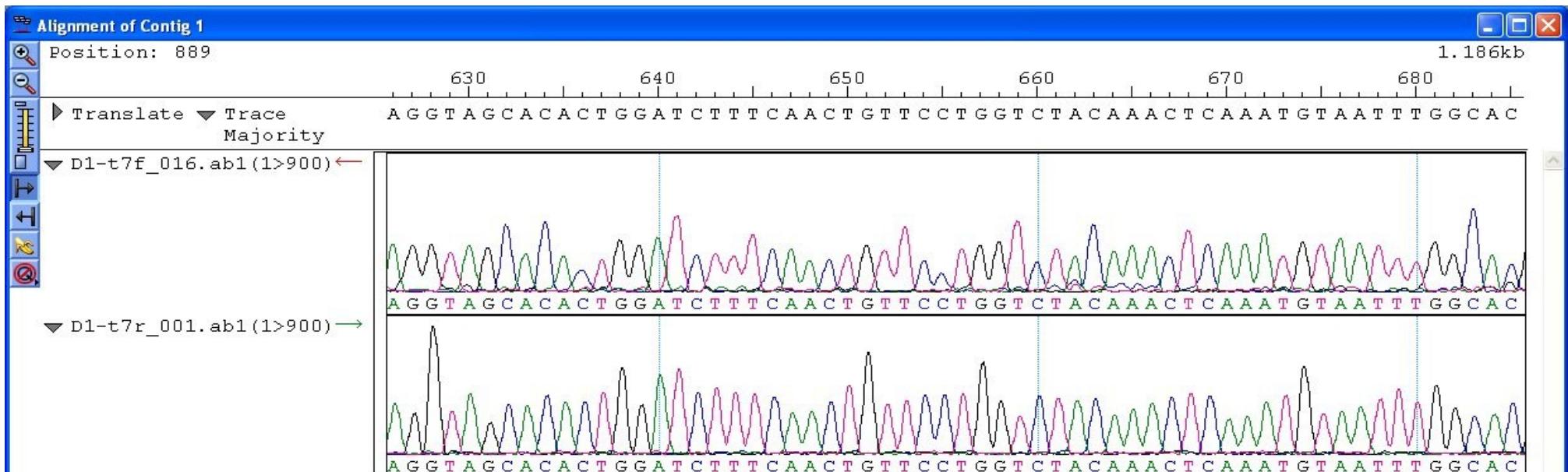
MinION



# Quality scores in sequencing

103/165

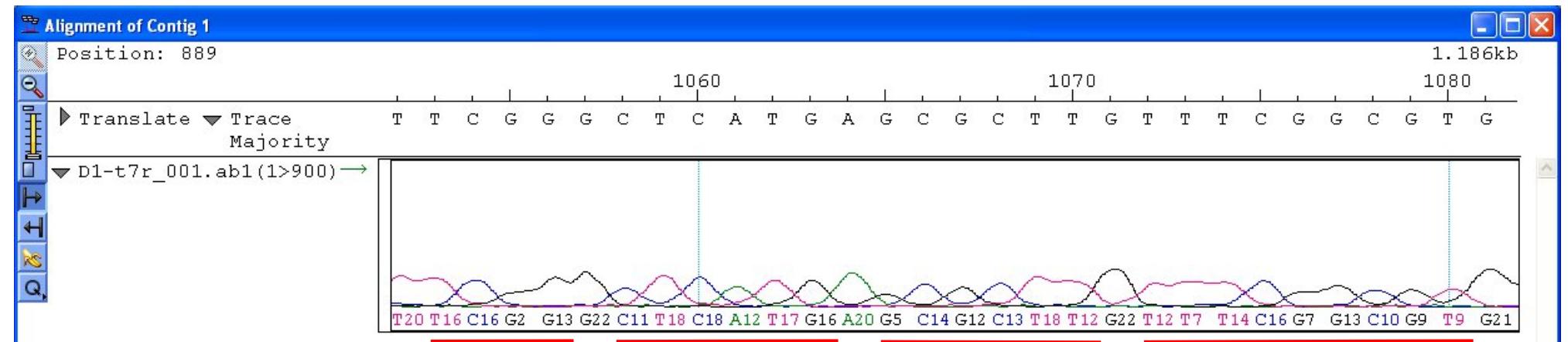
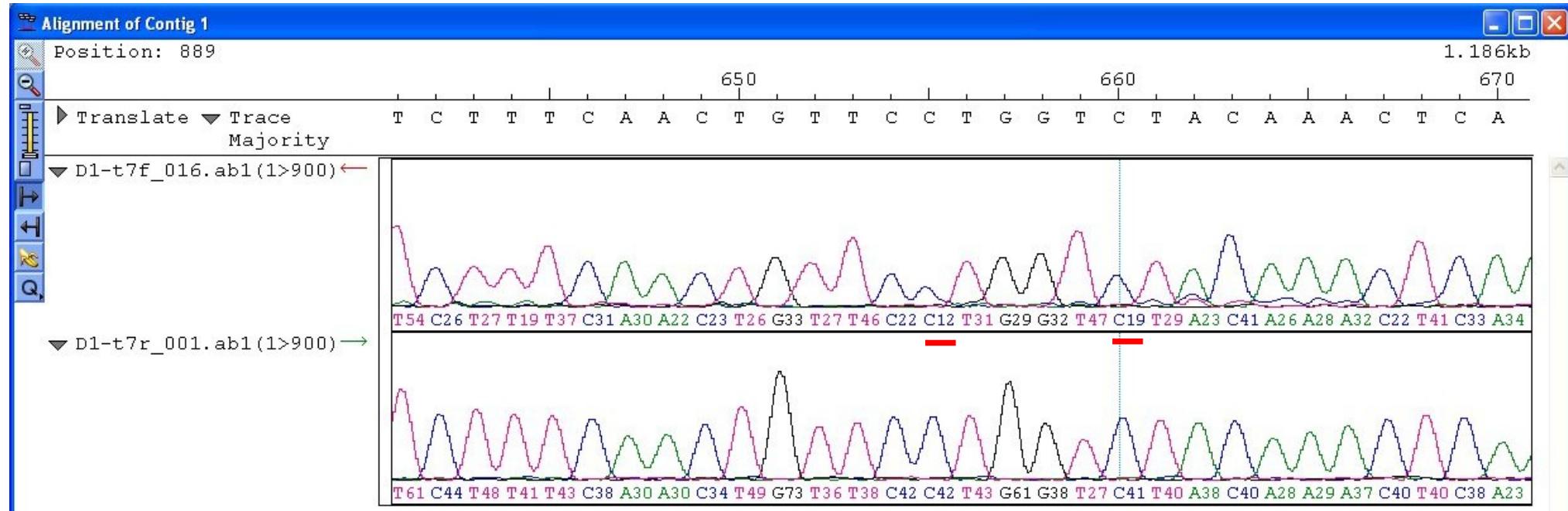
## Sequencing: quality scores (Phred scores) and accuracy



# Quality scores in sequencing

104/165

## Sequencing: quality scores (Phred scores) and accuracy

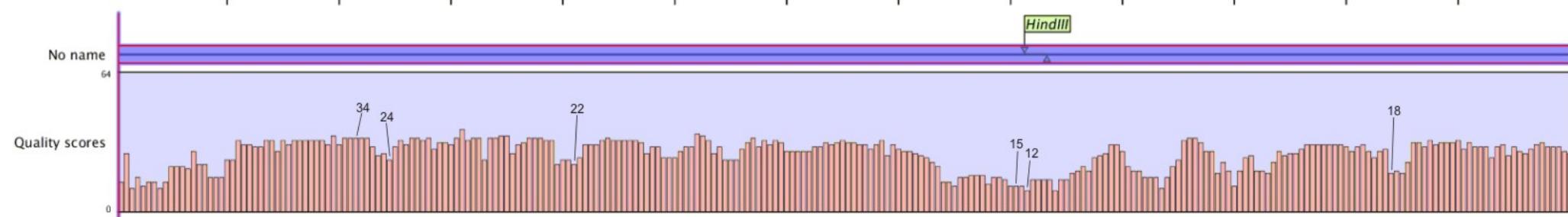


# Quality scores in sequencing

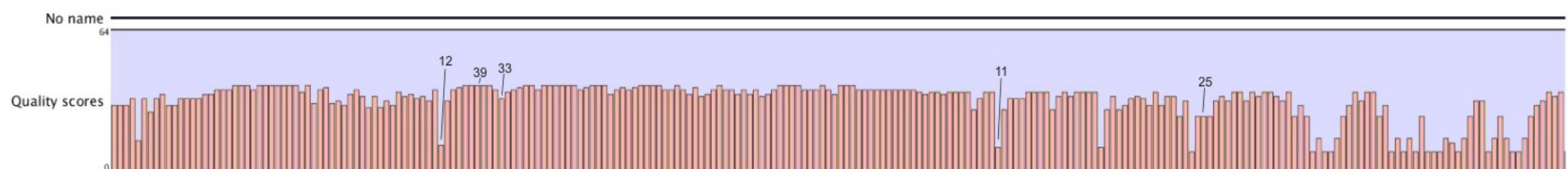
105/165

## Sequencing: quality scores (Phred scores) and accuracy

Ion Torrent PGM

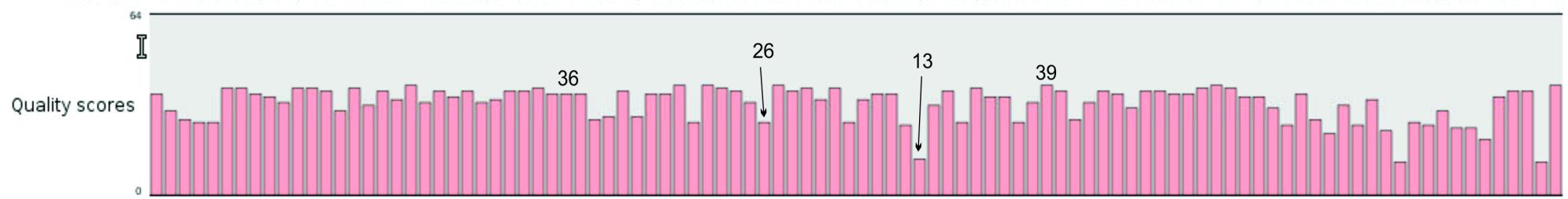


Illumina MiSeq



Illumina HiSeq

B:67869#37/1 GAGCAGAAAGATTAGTAGTCCTGCGATAATCACAGATTATGATTACGCCCTAAGAGTGTTCATCCTTACTTACCCCTGGCATGCCAGGGAAACTAGCCATACT



SOLiD

A\_3\_97\_809/2  
Color space

TCACTTGACTTGTTGTCCTGGTTCAACTGTCAAACAGTC

Quality scores

0

64

20

40

60

80

100

120

140

160

180

200

220

240

260

280

300

320

340

360

380

400

420

440

460

480

500

520

540

560

580

600

620

640

660

680

700

720

740

760

780

800

820

840

860

880

900

920

940

960

980

1000

1020

1040

1060

1080

1100

1120

1140

1160

1180

1200

1220

1240

1260

1280

1300

1320

1340

1360

1380

1400

1420

1440

1460

1480

1500

1520

1540

1560

1580

1600

1620

1640

1660

1680

1700

1720

1740

1760

1780

1800

1820

1840

1860

1880

1900

1920

1940

1960

1980

2000

2020

2040

2060

2080

2100

2120

2140

2160

2180

2200

2220

2240

2260

2280

2300

2320

2340

2360

2380

2400

2420

2440

2460

2480

2500

2520

2540

2560

2580

2600

2620

2640

2660

2680

2700

2720

2740

2760

2780

2800

2820

2840

2860

2880

2900

2920

2940

2960

2980

3000

3020

3040

3060

3080

3100

3120

3140

3160

3180

3200

3220

3240

3260

3280

3300

3320

3340

3360

3380

3400

3420

3440

3460

3480

3500

3520

3540

3560

3580

3600

3620

3640

3660

3680

3700

3720

3740

3760

3780

3800

3820

3840

3860

3880

3900

3920

3940

3960

3980

4000

4020

4040

4060

4080

4100

4120

4140

4160

4180

4200

4220

4240

4260

4280

4300

4320

4340

4360

4380

4400

4420

4440

4460

4480

4500

4520

4540

4560

4580

4600

4620

4640

4660

4680

4700

4720

4740

4760

4780

4800

4820

4840

4860

4880

4900

4920

4940

4960

4980

5000

5020

5040

5060

5080

5100

5120

5140

5160

5180

5200

5220

5240

5260

5280

5300

5320

5340

5360

5380

5400

5420

5440

5460

5480

5500

5520

5540

5560

5580

5600

5620

5640

5660

5680

5700

5720

5740

5760

5780

5800

5820

5840

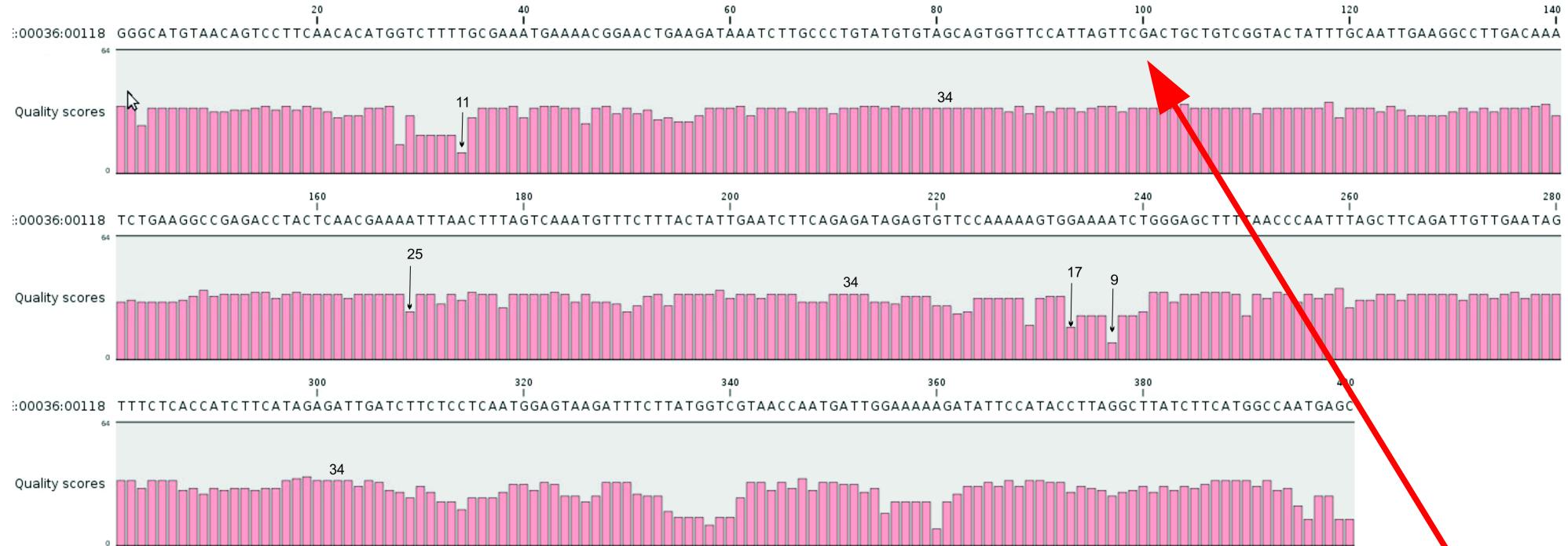
5860

# Quality scores in sequencing

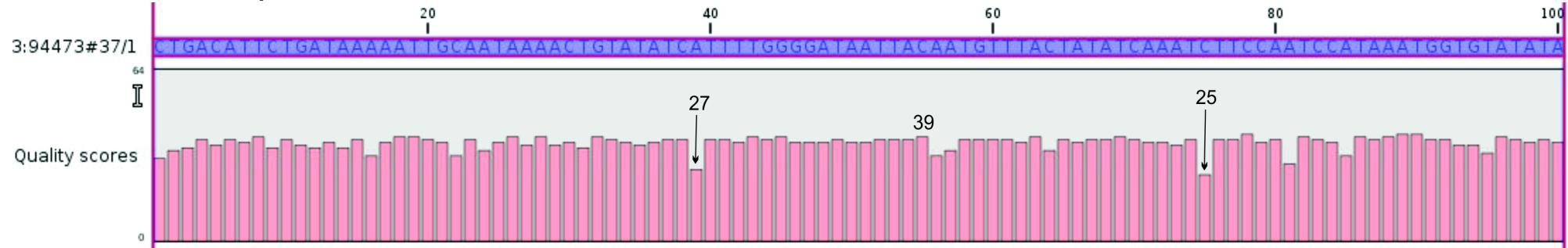
106/165

## Sequencing: quality scores (Phred scores) and accuracy

### Ion Torrent PGM



### Illumina HiSeq

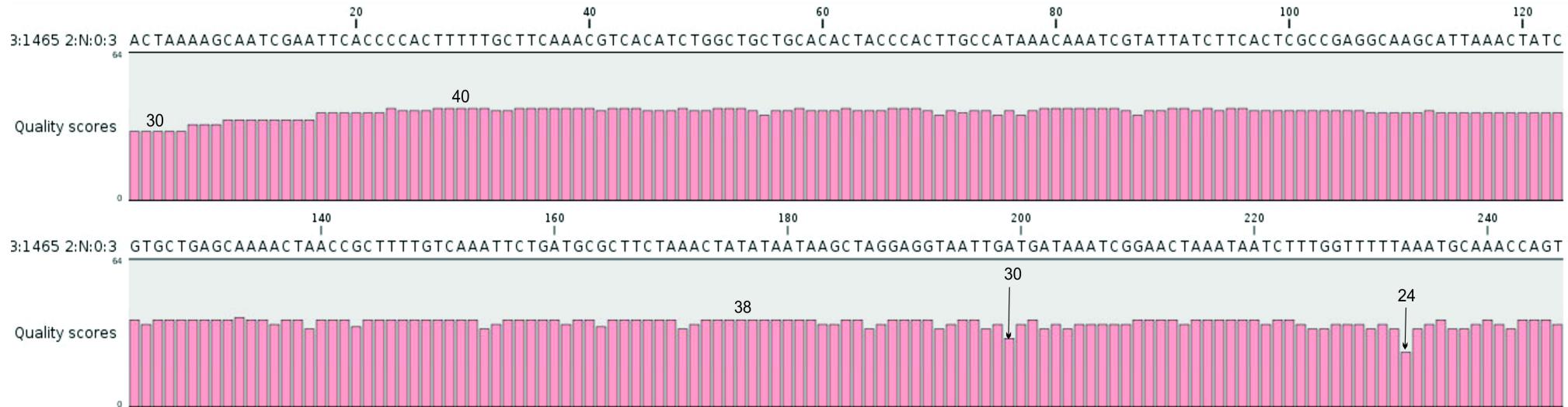


# Quality scores in sequencing

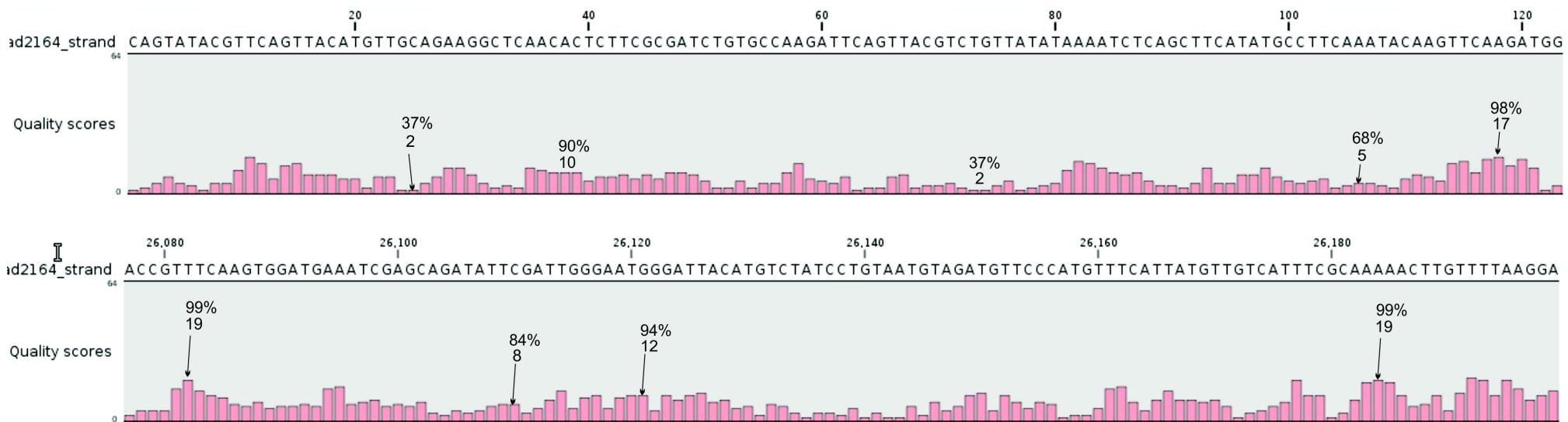
107/165

## Sequencing: quality scores (Phred scores) and accuracy

### Illumina MiSeq



### MinIon

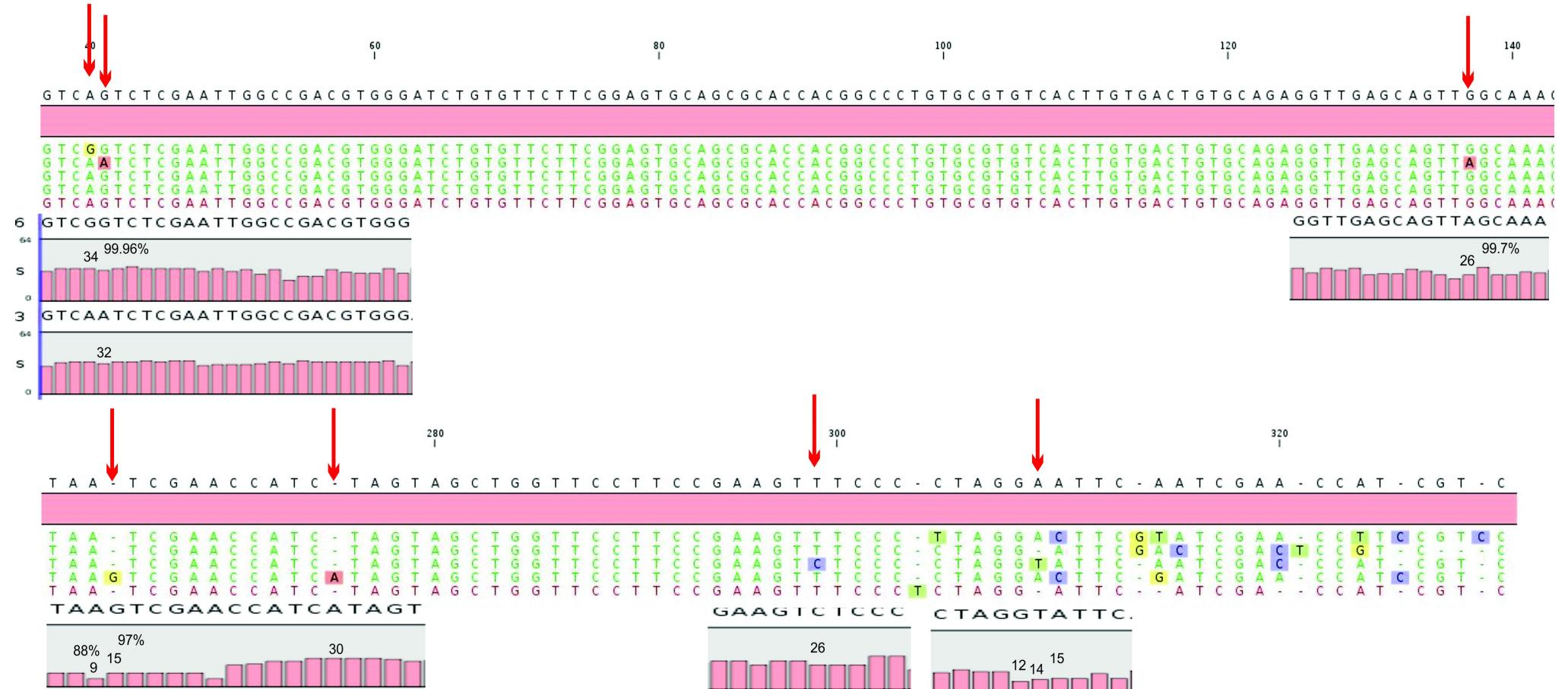


## Quality scores in sequencing

Sequencing: quality scores (Phred scores) and accuracy

IonTorrent Assembly

?? how important are quality scores ??



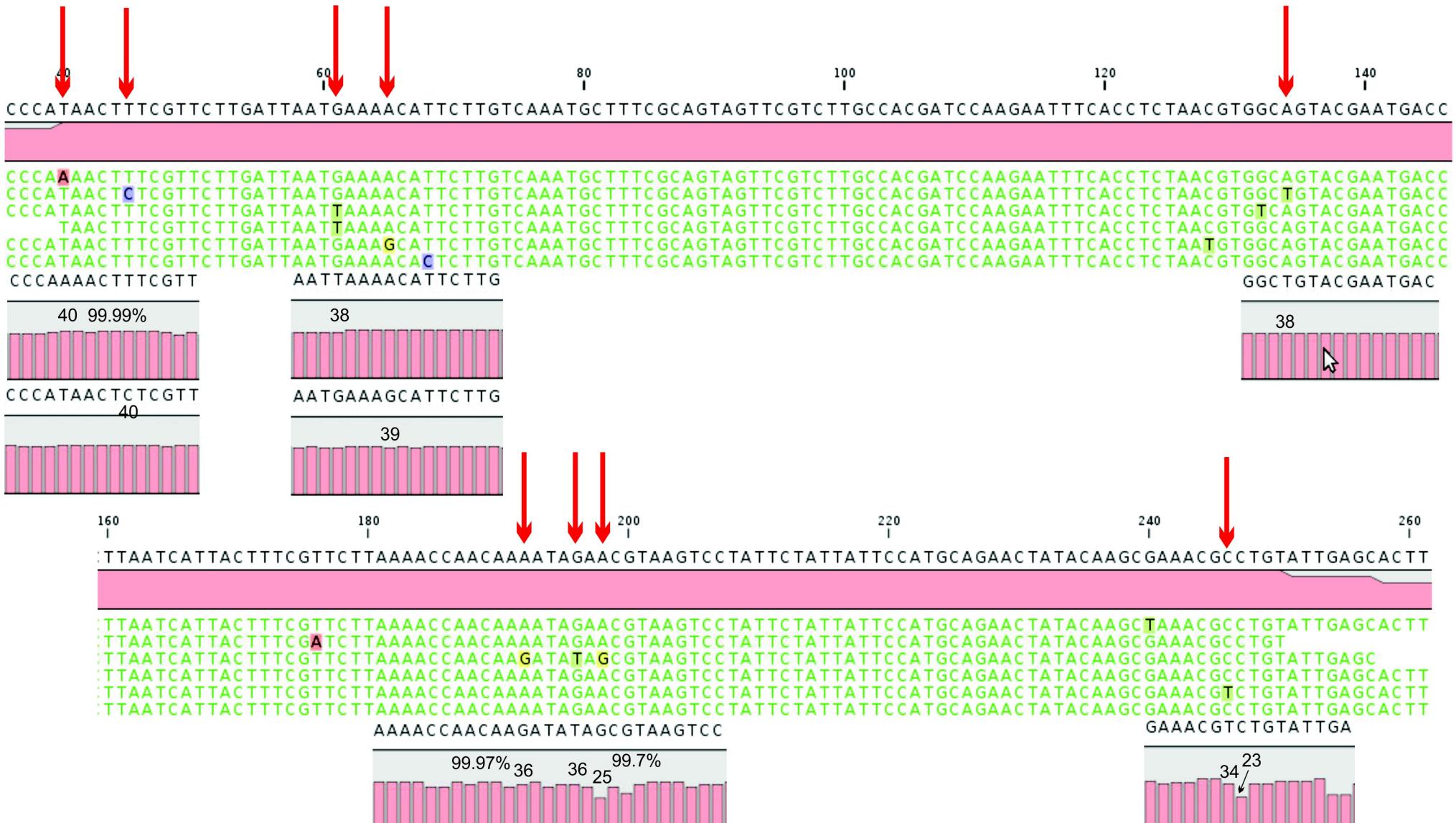
# Quality scores in sequencing

109/165

## Sequencing: quality scores (Phred scores) and accuracy

MiSeq Assembly (Halomonhyphera 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:**  $2,28 \times 10^{-5}$  → 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:**  $4,4 \times 10^{-7}$  → 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.)

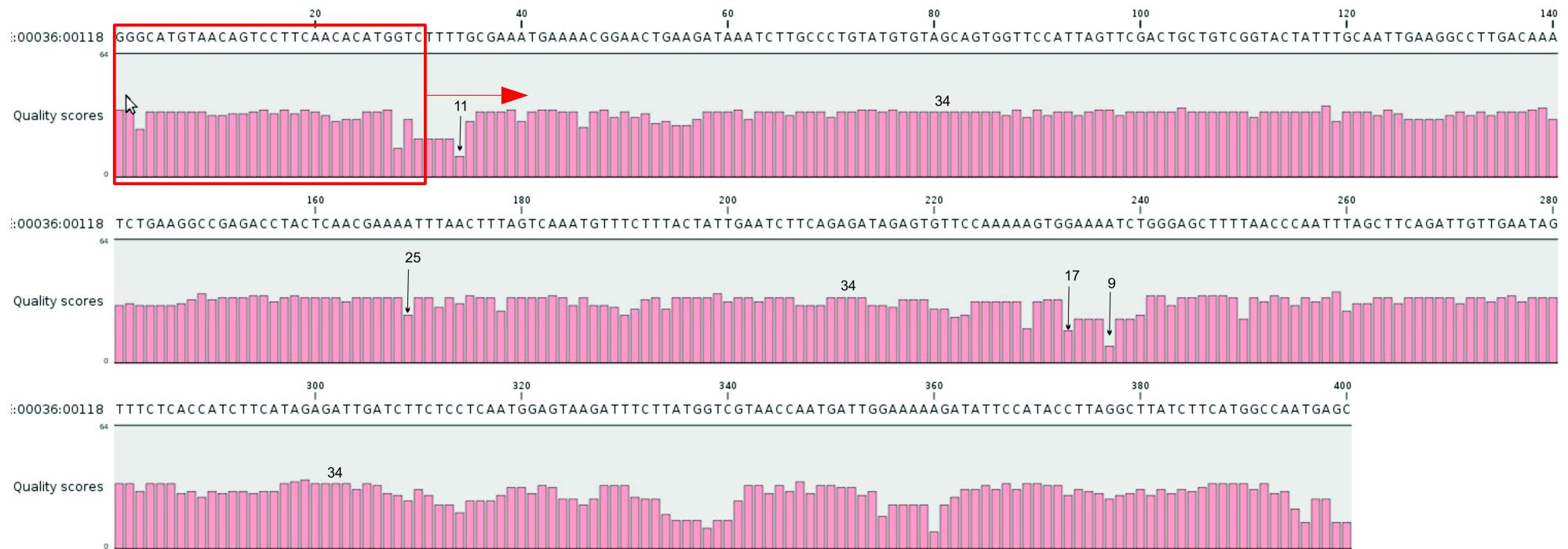
- DNA damage in cells by aging
- ... ?

## Quality scores in sequencing

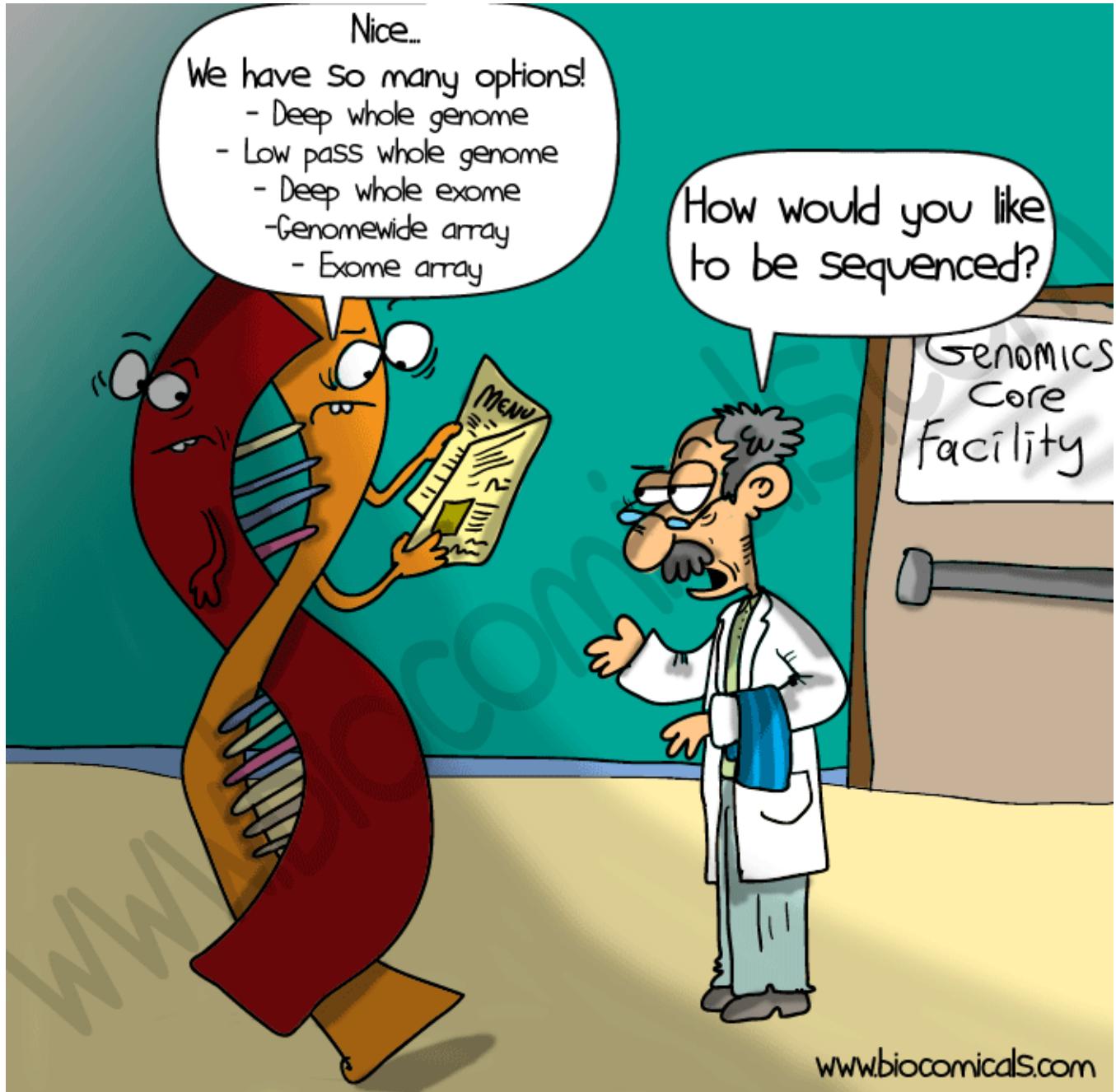
111/165

### Sequencing: quality scores (Phred scores) and accuracy

Quality trimming of reads with sliding window 30, cutoff 15



## Applications

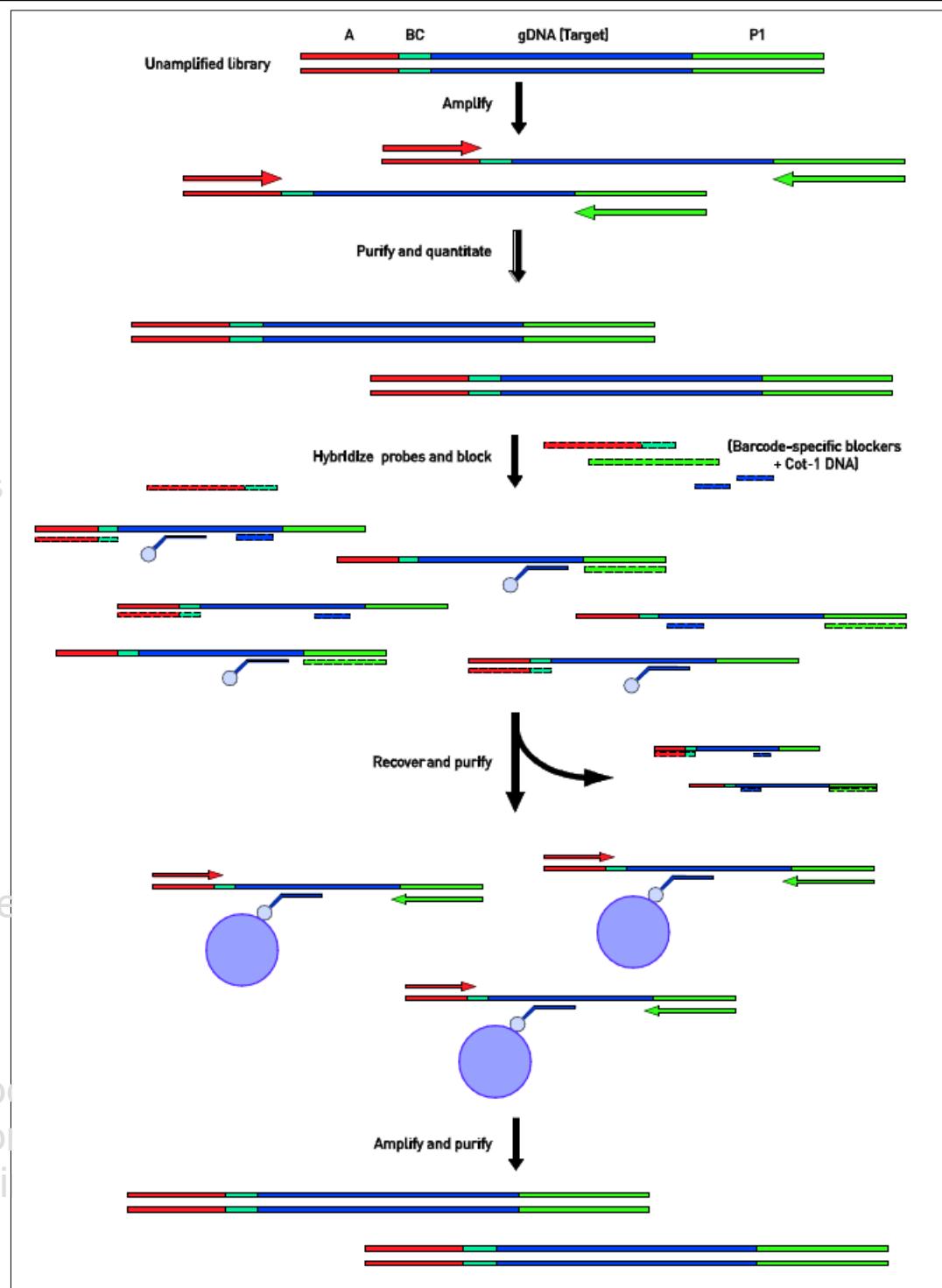


## Applications

- 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)
- ...

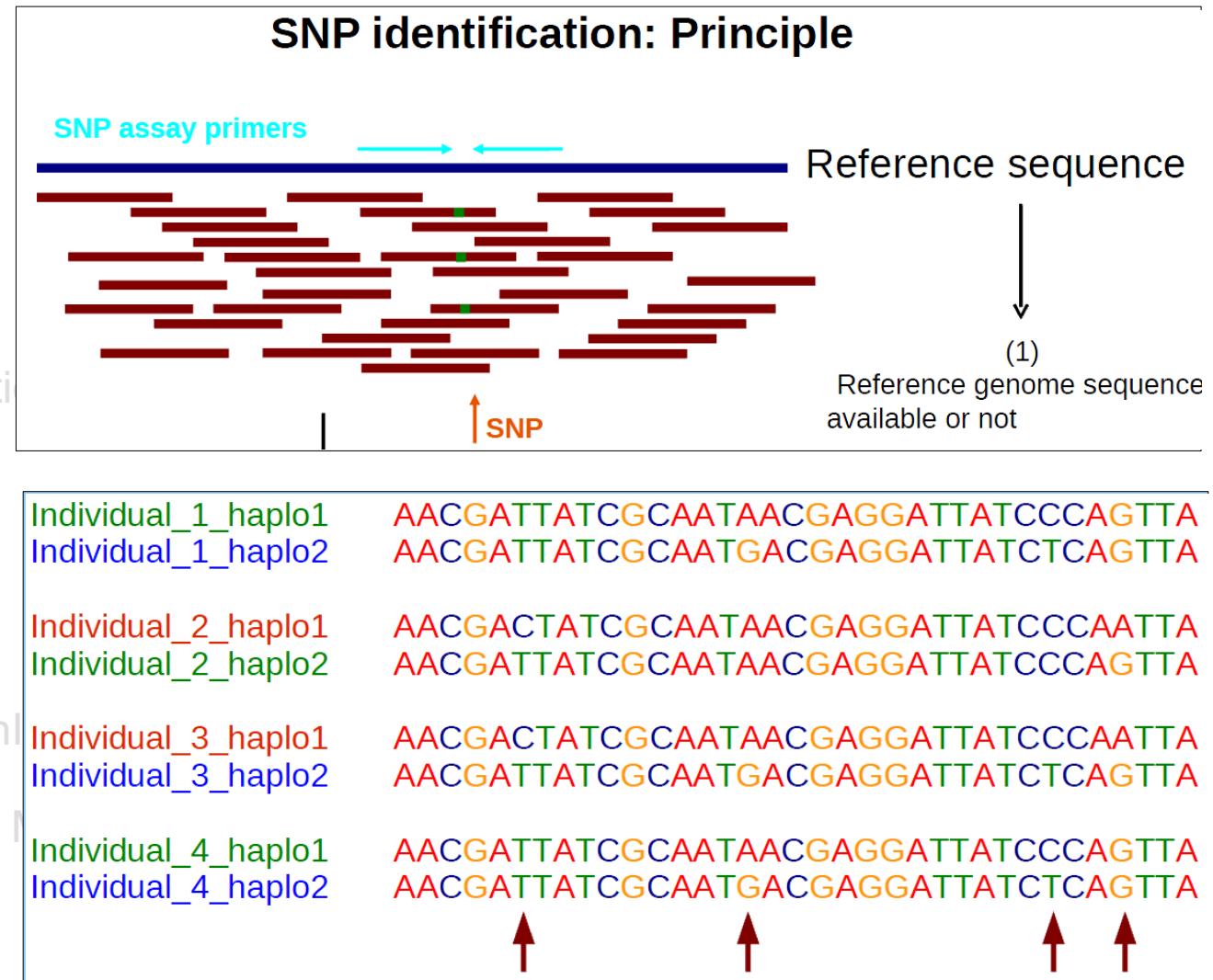
## Applications

- 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-Assoc)
  - RRLs (Reduced-Representation)
  - GBS (Genotyping By Sequenc
- ...



## 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-seq
- 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- Ph

- Transcript Expression Profiling

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

- Transcription factor binding sites

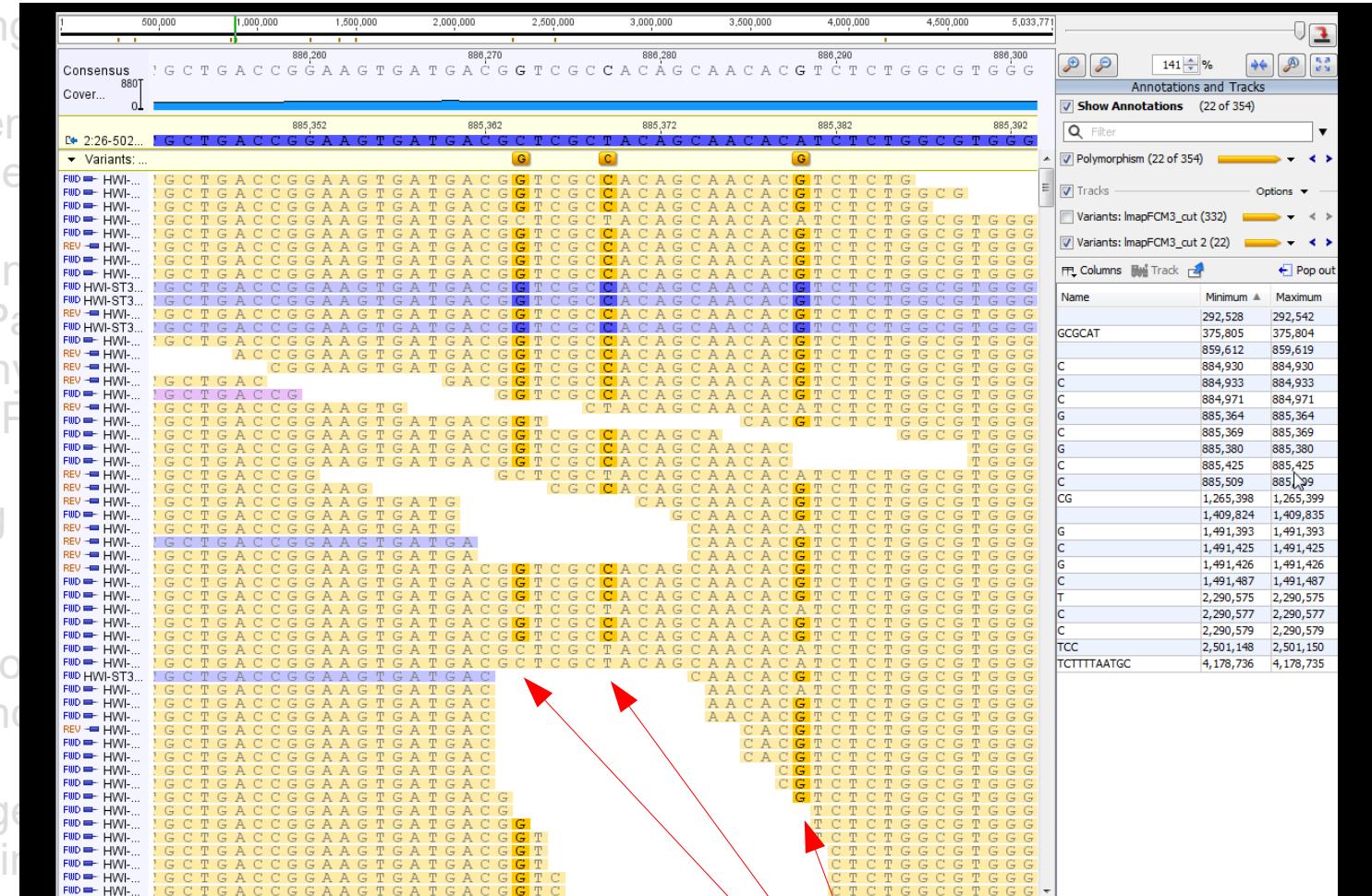
- Structural variation

- Metagenomics / Metabarcoding

- 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

- RNA sequencing

• miRN

• Deep

• Deep

• PAS:

• Transcri

• Structur

• Metagen

• Microsat

• Genetic

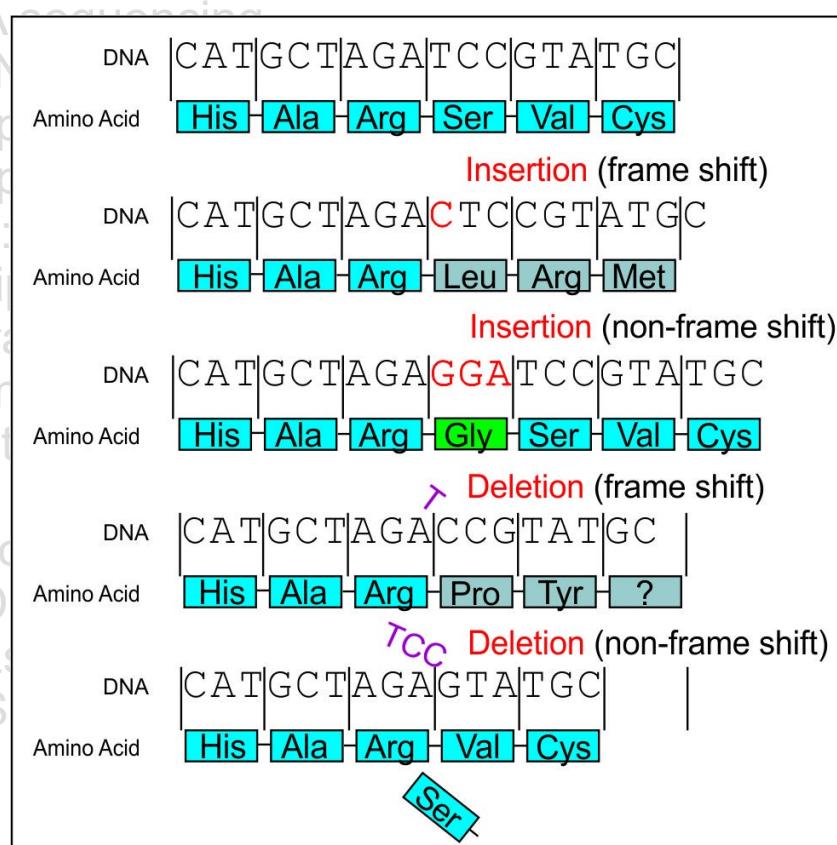
• Micro

• RAD

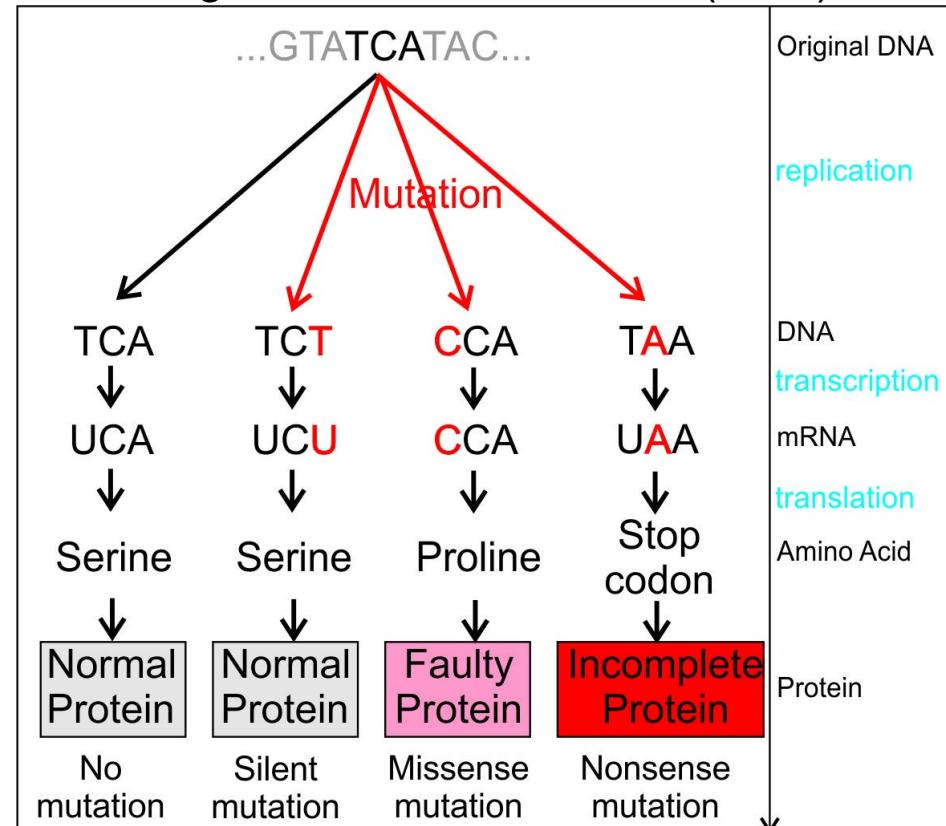
• RRLS

• GBS

• ...



## 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- Phylogenetic

- Transcript Expression Profiling
  - RNA sequencing
  - miRNA sequencing
  - Deep-SAGE
  - Deep-CAGE
  - PAS: polyadenylation site

- Transcription factor binding

### • Structural variation

#### • Metagenomics / Metagenetics

### • Microsatellite sequencing

## Microsatellite sequencing • Genetic marker discovery

### • Genetic marker discovery

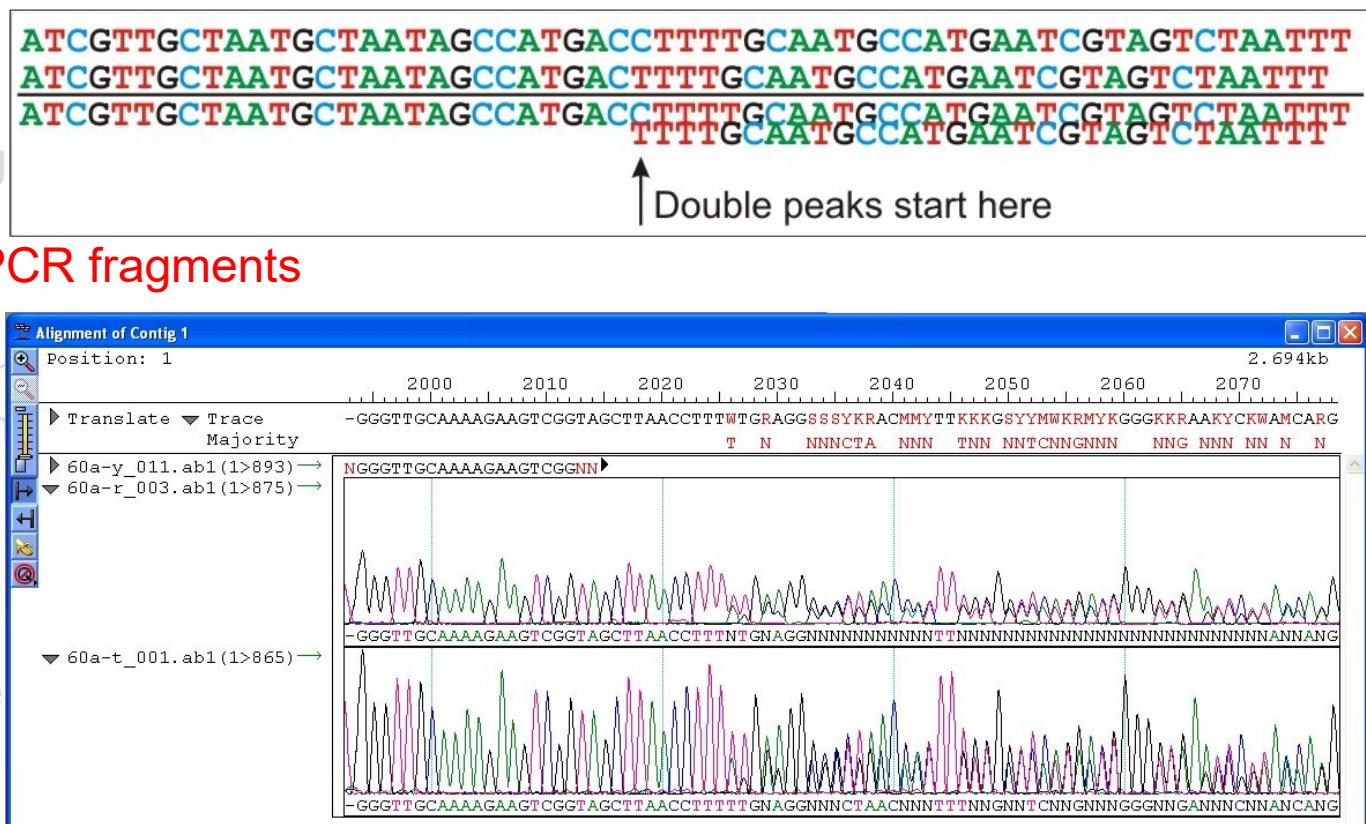
- Microsatellite development  
PARS (Partial

- RADSeq (Restriction-site Associated DNA Sequencing)

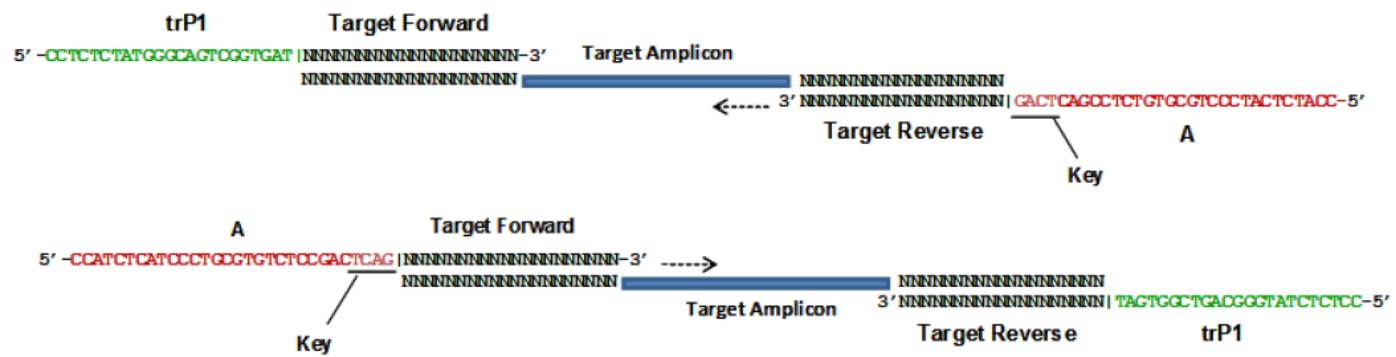
- RRLs (Reduced-Represen

- GBS (Genotyping By S

• • • • •

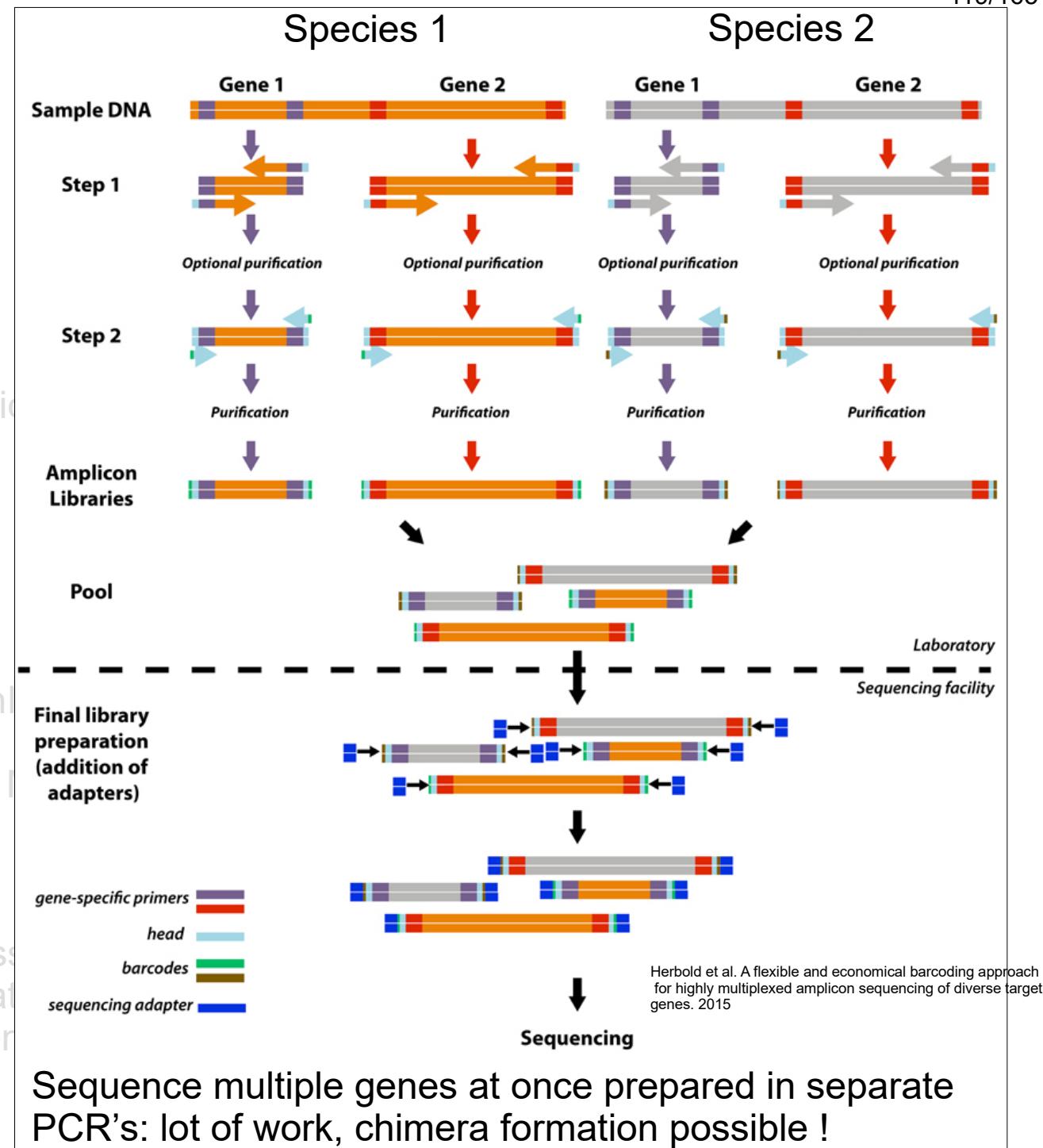


Fusion method: pcr fragments of 400 bp (IonTorrent)



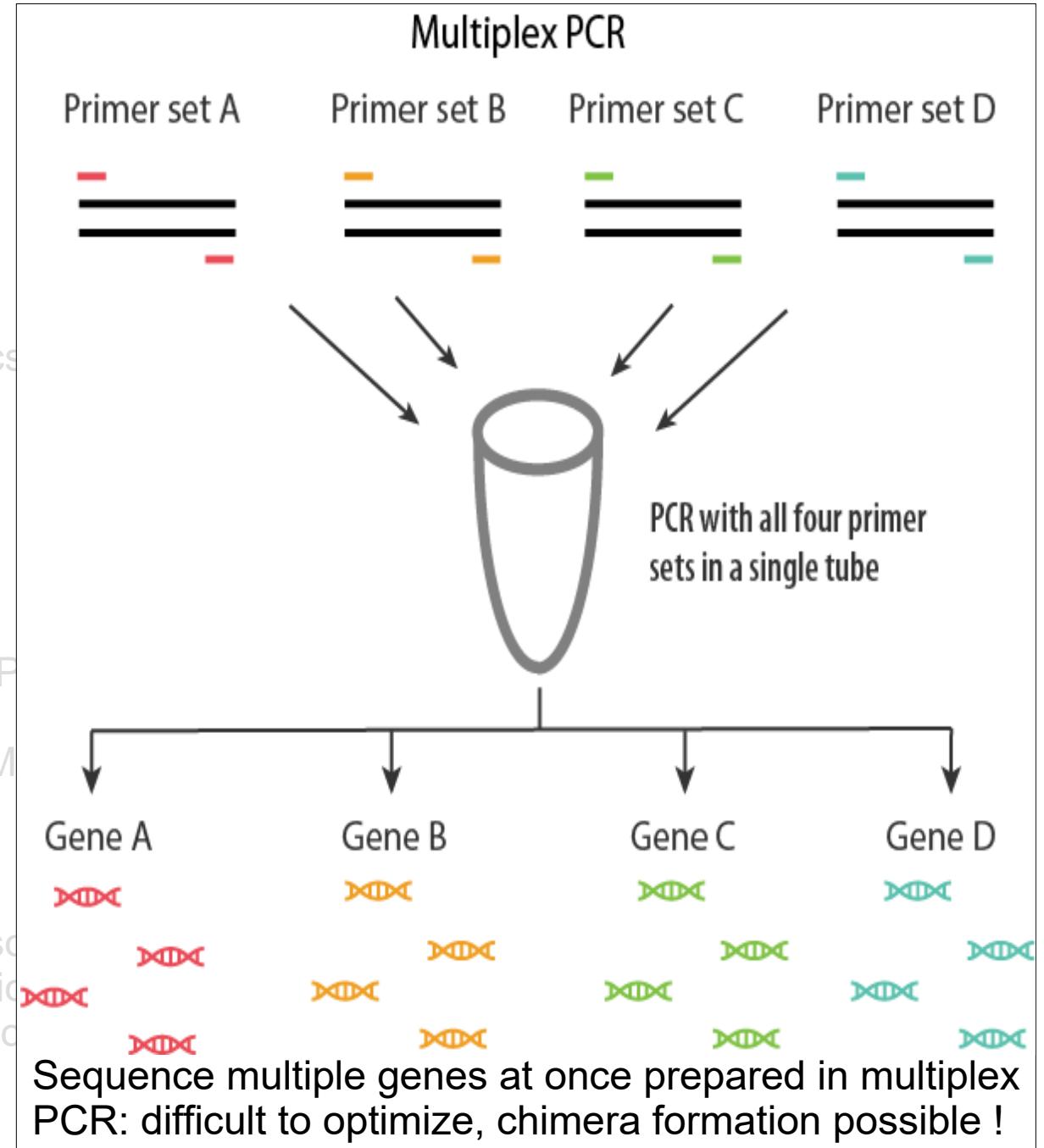
## Applications

- 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-seq
- Structural variation
- Metagenomics / Metagenetics /
- Microsatellite sequencing
- Genetic marker discovery
  - Microsatellite development
  - RADSeq (Restriction-site-Associated DNA Sequencing)
  - RRLs (Reduced-Representation Linkage Analysis)
  - 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 / M
- Microsatellite sequencing
- Genetic marker discovery
  - Microsatellite development
  - RADSeq (Restriction-site-Asso
  - RRLs (Reduced-Representati
  - GBS (Genotyping By Sequenc
- ...



## Applications

- Genome Sequencing
  - De novo sequencing
  - Resequencing
  - Targeted (re)sequencing
  - Mitochondrial sequencing
  - Mutation detection
  - Amplicon sequencing
  - **Amplicon Cancer Panel**
  - Phylogenomics- Phylogenetic analysis
- Transcript Expression Profiling
  - RNA sequencing
  - miRNA sequencing
  - Deep-SAGE
  - Deep-CAGE
  - PAS: polyadenylation site analysis
- Transcription factor binding
- Structural variation
- Metagenomics / Metatranscriptomics
- Microsatellite analysis
- Genetic mapping
  - Microsatellites
  - RADSeq
  - RRLs (Recombinant Random Libraries)
  - GBS (Genome-wide SNP analysis)
  - ...

Ion AmpliSeq Comprehensive Cancer Panel		
Targets	Exons with >400 oncogenes and tumor suppressor genes	
Specification	Observed performance	
Coverage uniformity*	>90%	
On-target bases**	>95%	
Average depth of coverage	NA	

\* 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



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

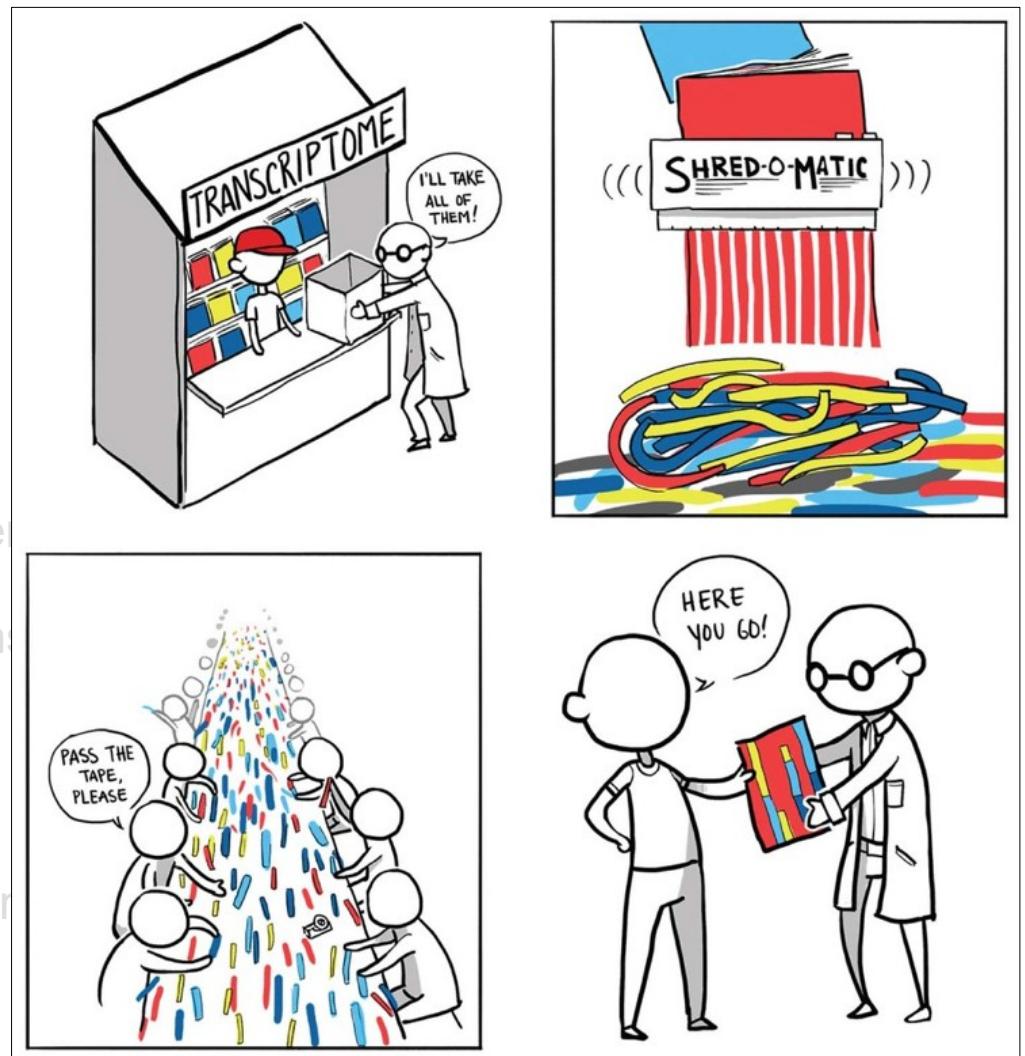
- Microsatellite development
  - RADSeq (Restriction-site-Associated Sequencing)
  - RRLs (Reduced-Representation libraries)
  - GBS (Genotyping By Sequencing)

**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 sequencing
- Structural variation
- Metagenomics / Metagenetics / Metatranscriptomics
- Microsatellite sequencing
- Genetic marker discovery
  - Microsatellite development
  - RADSeq (Restriction-site-Associated DNA sequencing)
  - RRLs (Reduced-Representation library)
  - GBS (Genotyping By Sequencing)
  - ...
- \* Different expression profiles of genes in different conditions (avoid rRNA)
- \* Determine the genes in the transcriptome (also lncRNA)



## Applications

- Genome Seq
  - De novo sequencing
  - Resequencing
  - Targeted sequencing
  - Mitochondria
  - Mutation detection
  - Amplicon sequencing
  - Amplicon Cancer Panel
  - Phylogenomics- Phylogenetics
- Transcript Expression Profiling
  - RNA sequencing
  - miRNA sequencing
  - Deep-Seq
  - Deep-CAGE
  - PAS: polyA sequencing
- Transcription
- Structural variation
- Metagenomics
- Microsatellite analysis
- Genetic mapping
  - Microsatellite analysis
  - 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.

**ThermoFisher**  
SCIENTIFIC

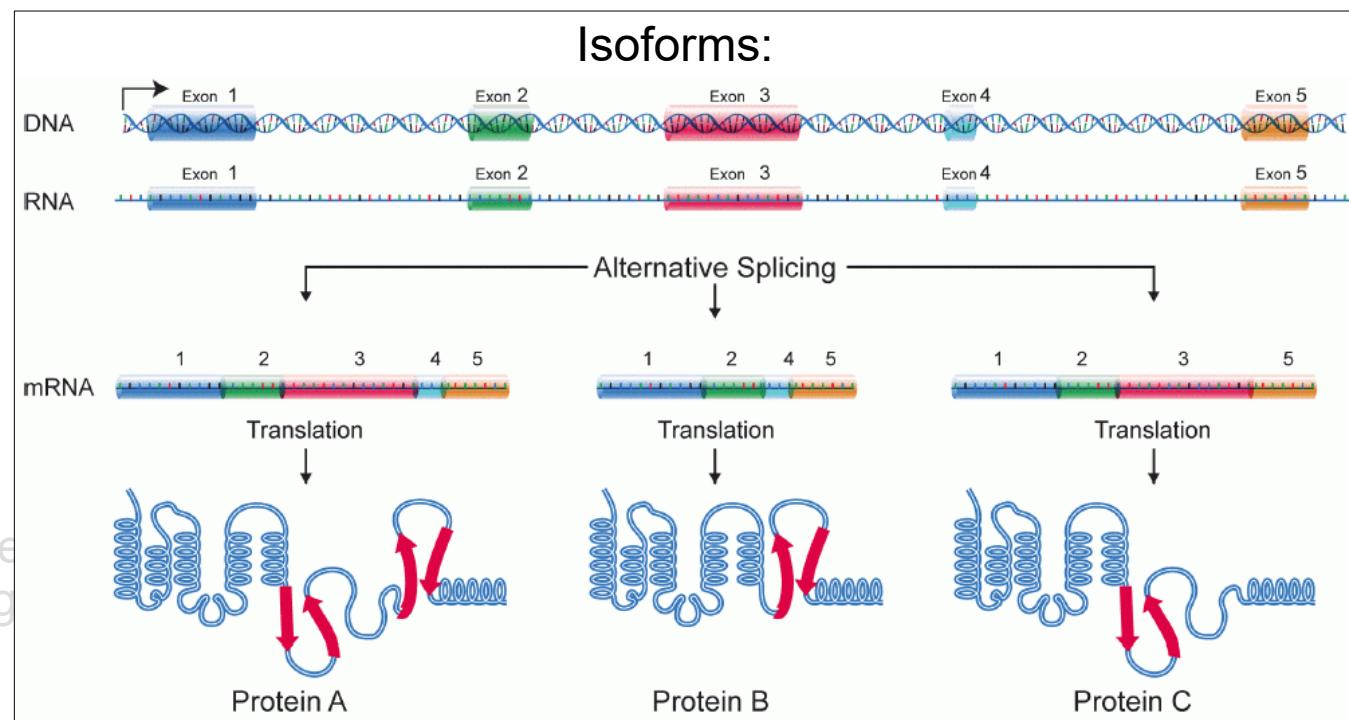
### Ion AmpliSeq Colon and Lung Cancer Research Panel v2 and Ion AmpliSeq RNA Fusion Lung Cancer Research Panel

<b>Application</b>	Somatic mutation detection
<b>Genes</b>	<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>

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-seq
- Structural variation
- Metagenomics / Metagenetics /
- Microsatellite sequencing
- Genetic marker discovery
  - Microsatellite development
  - RADSeq (Restriction-site-Associated DNA Sequencing)
  - RRBS (Reduced Representation Bisulfite Sequencing)



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

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

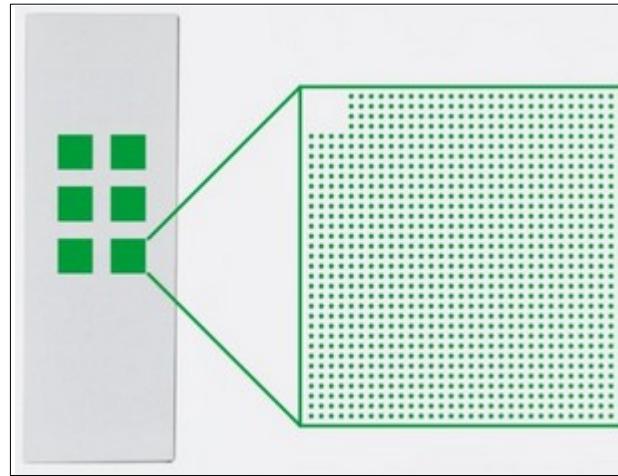
RESEARCH ARTICLE

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

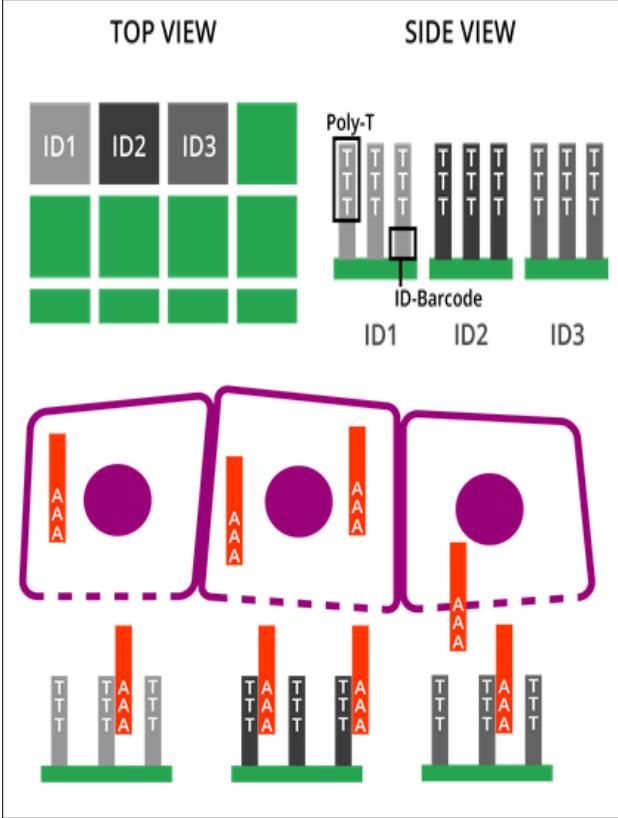
Jason L Weirather<sup>1\*</sup>, Mariateresa de Cesare<sup>2\*</sup>, Yunhao Wang<sup>1,3,4\*</sup>, Paolo Piazza<sup>2</sup>, Vittorio Sebastian<sup>5,6</sup>, Xiu-Jie Wang<sup>3</sup>, David Buck<sup>2</sup>, Kin Fai Au<sup>1,7</sup>

## Applications

- Genome Sequencing

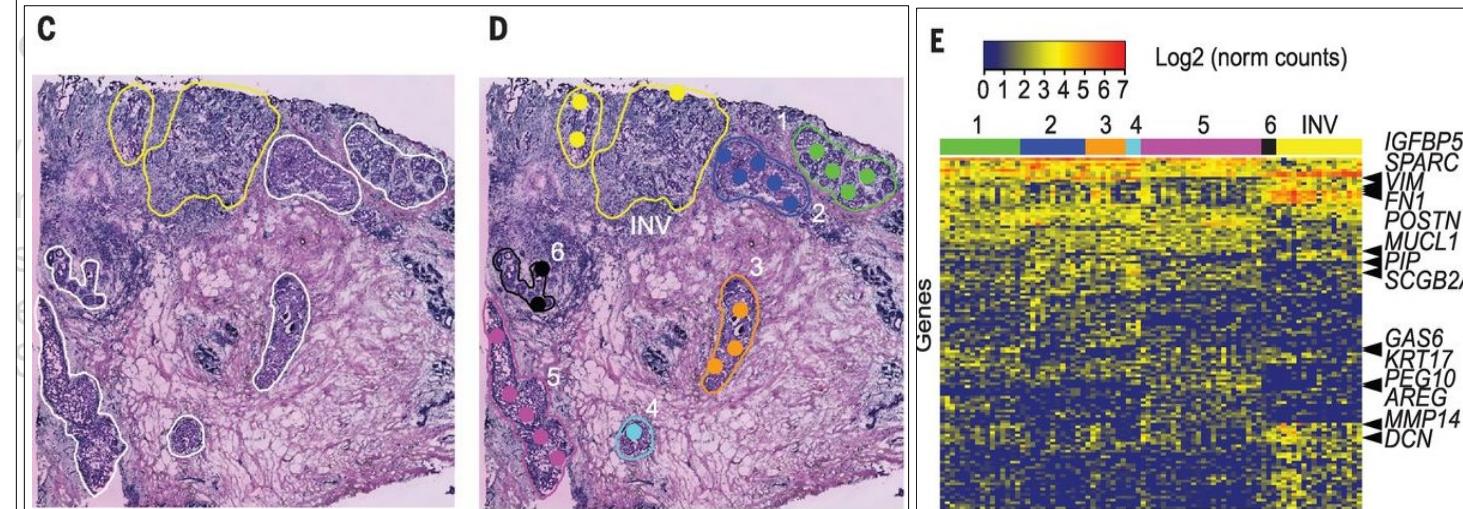
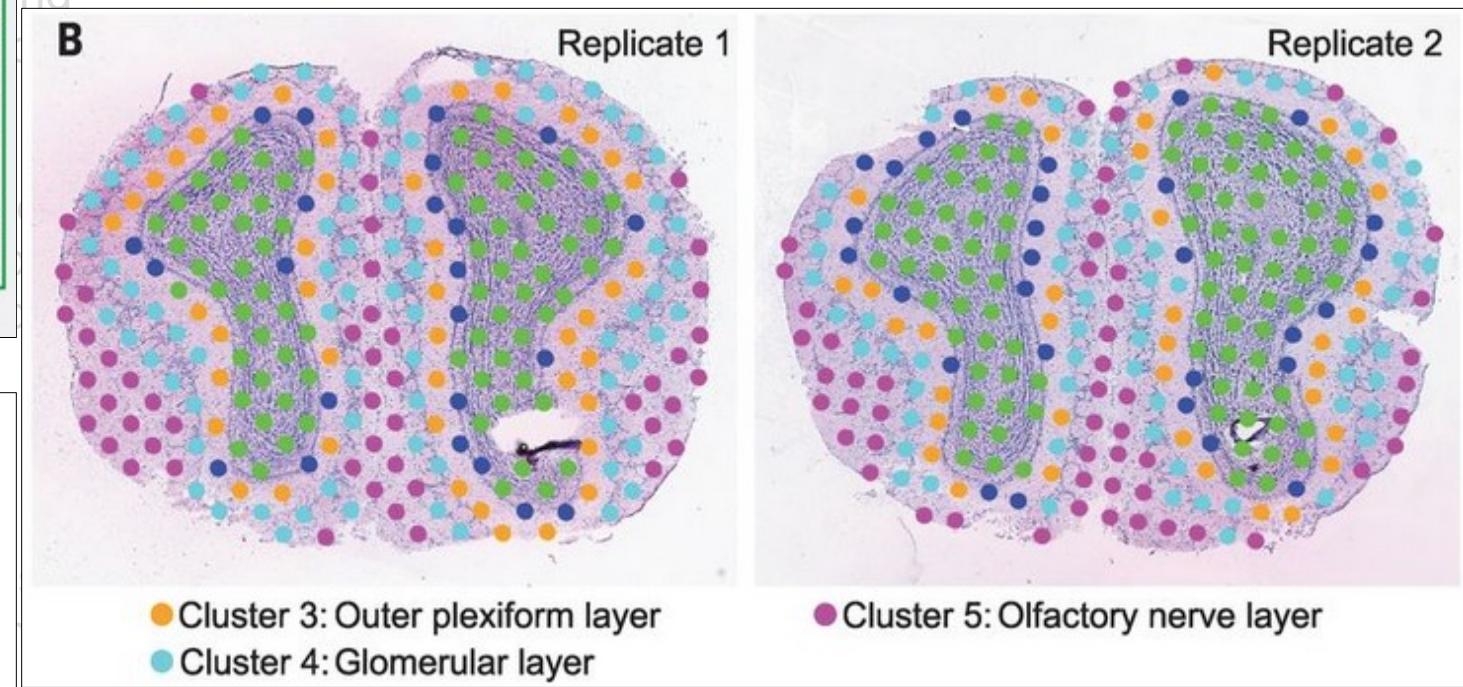


- RNA sequencing



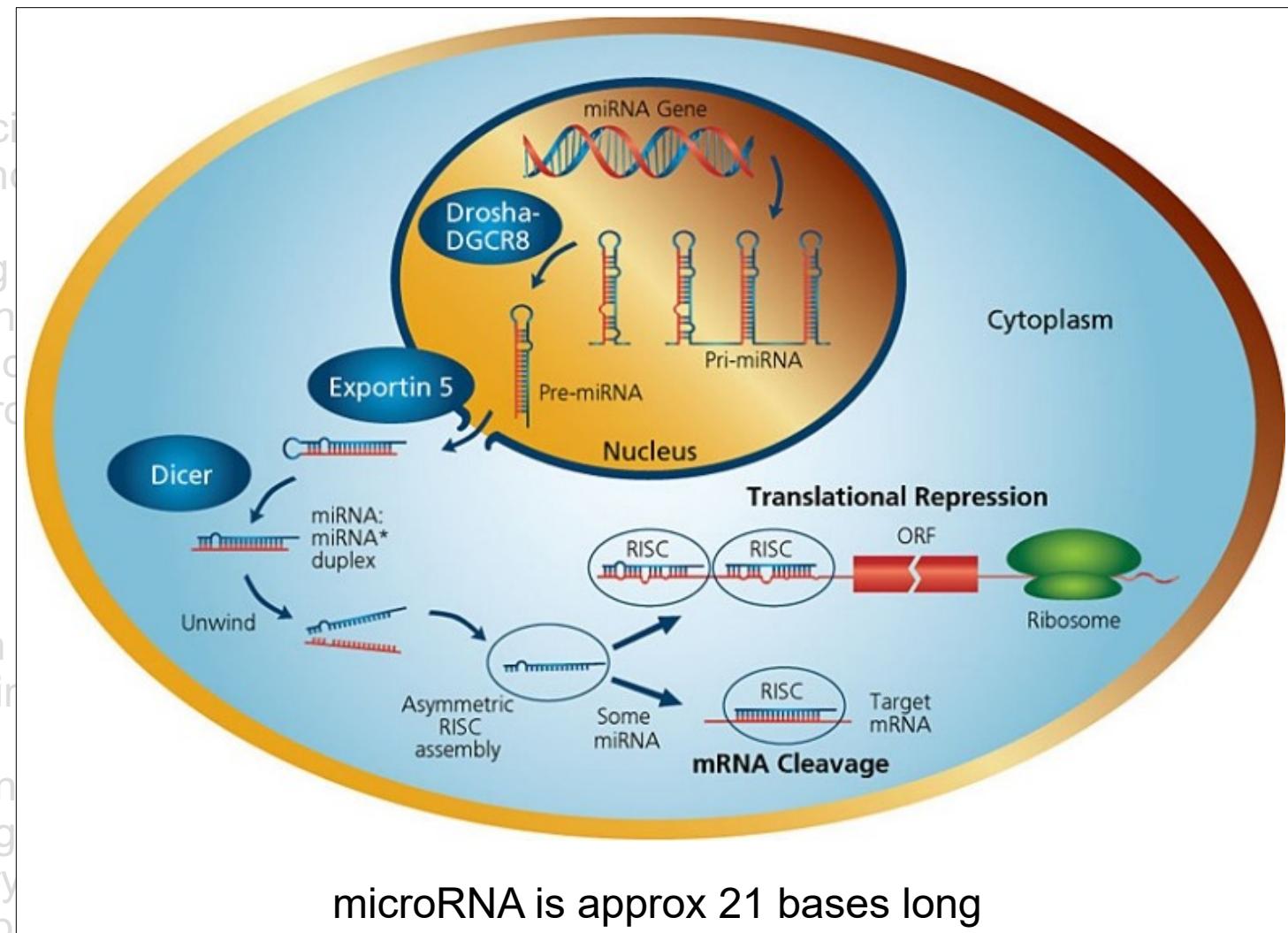
2D transcriptomics: [spatialtranscriptomics.com](http://spatialtranscriptomics.com)

*Visualization and analysis of gene expression in tissue sections by spatial transcriptomics  
Patrik L. Ståhl et all. Science 1 July 2016*



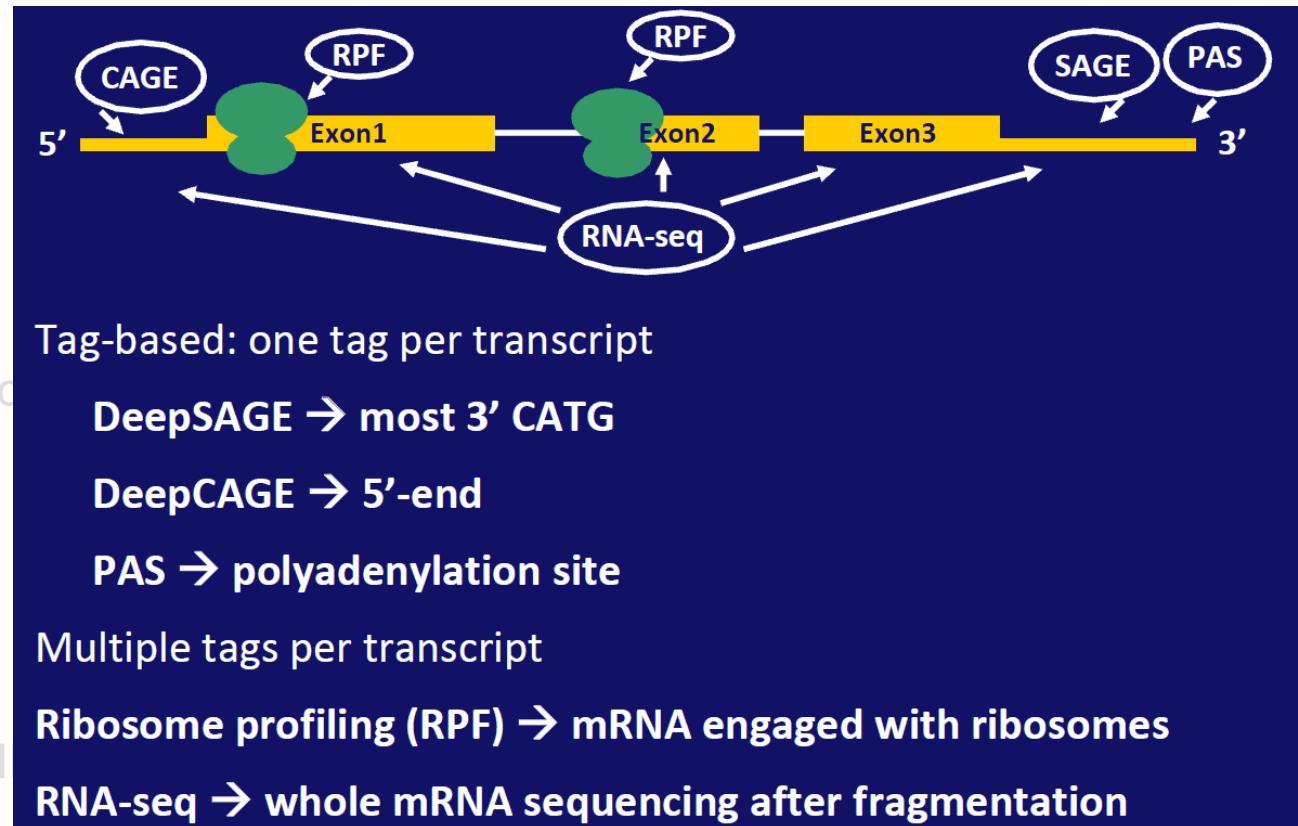
## Applications

- 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
- Transcription factor binding sites
- Structural variation
- Metagenomics / Metagenomic
- 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-seq
- 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)
- ...

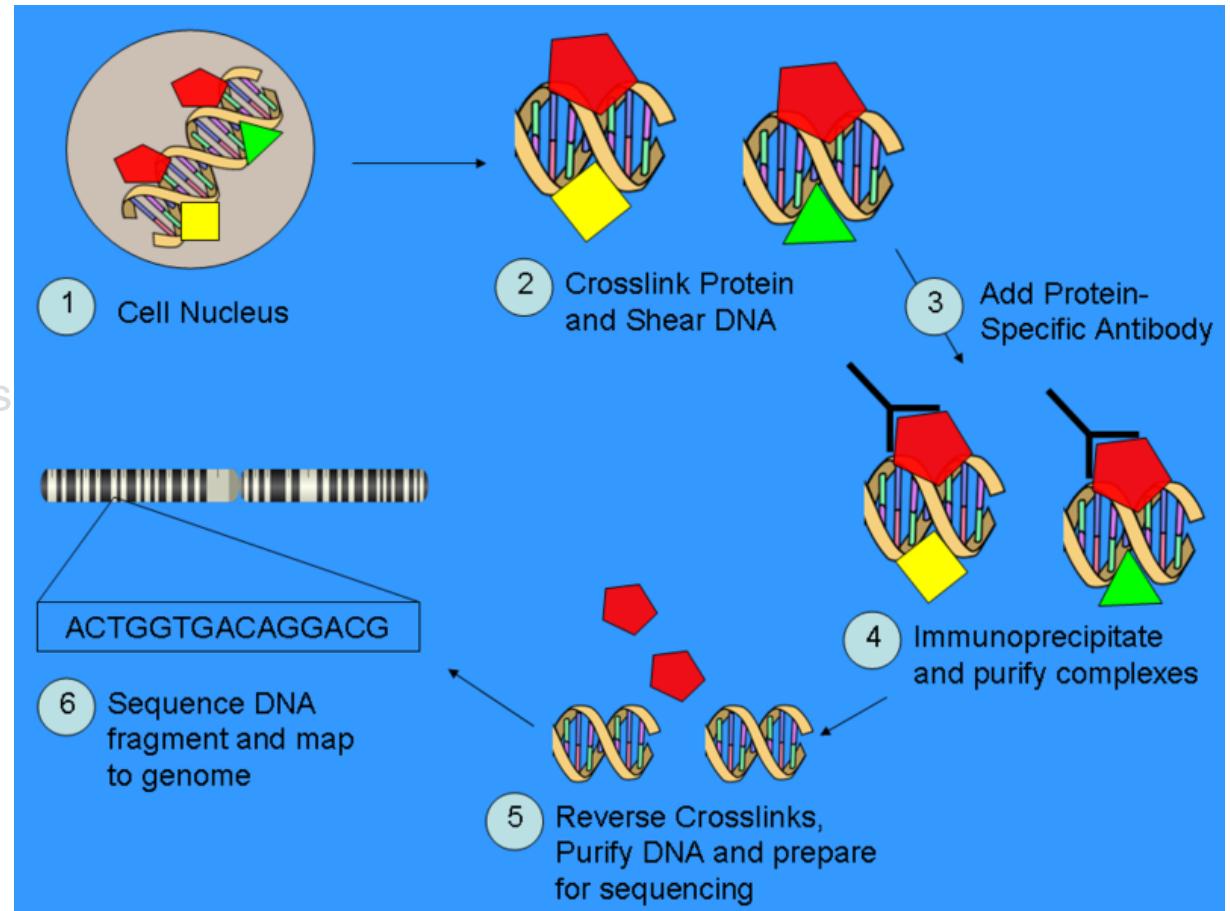


Sequence 17 bp of the transcription start or stop site  
-> discovery of new start and stop sites,  
antisense transcription possibilities

## Applications

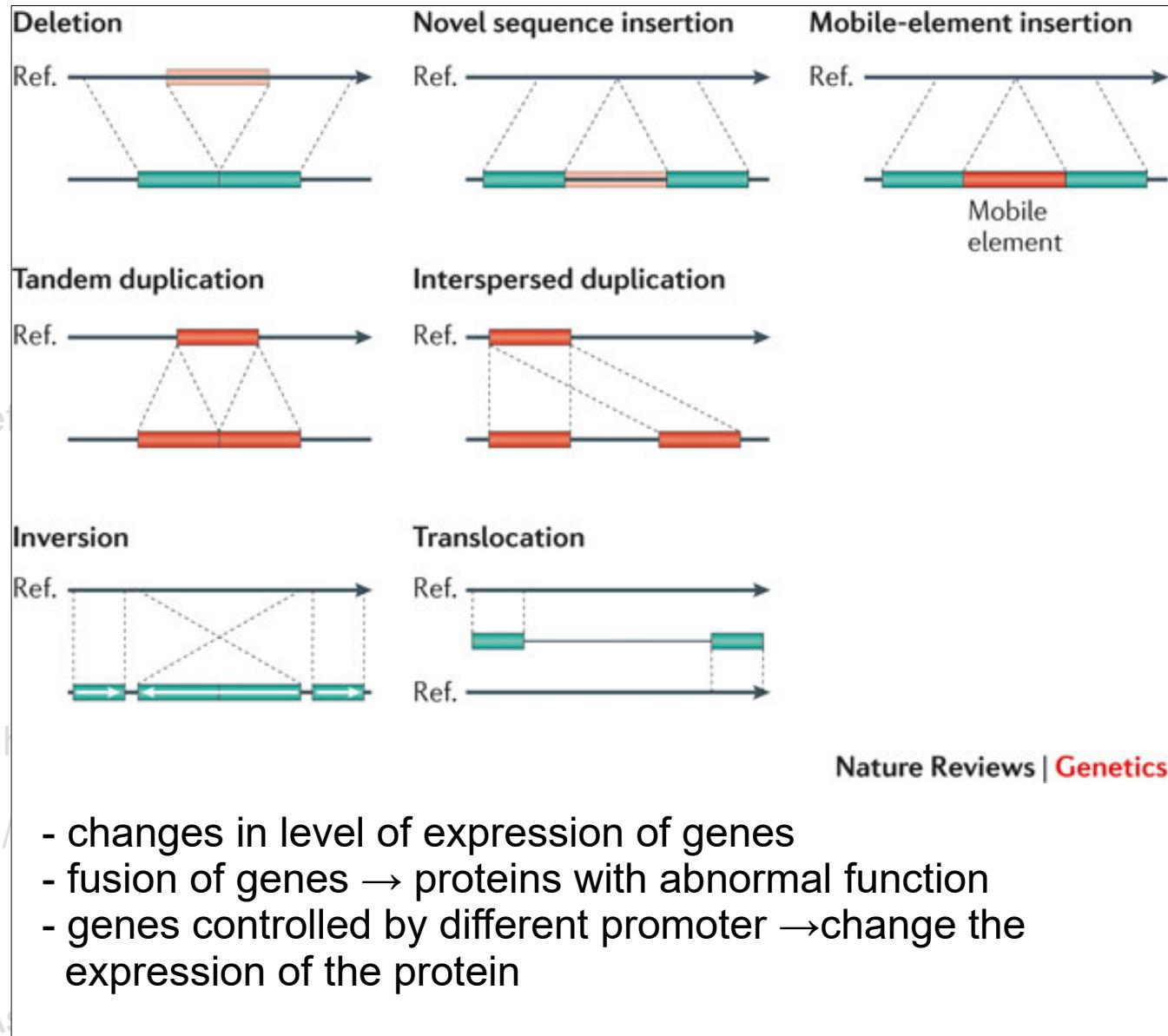
- Genome Sequencing
  - De novo sequencing
  - Resequencing
  - Targeted (re)sequencing
  - Mitochondrial sequencing
  - Mutation detection
  - Amplicon sequencing
  - Amplicon Cancer Panel
  - Phylogenomics- Phylogenetics
- Transcript Expression Profiling
  - RNA sequencing
  - miRNA sequencing
  -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

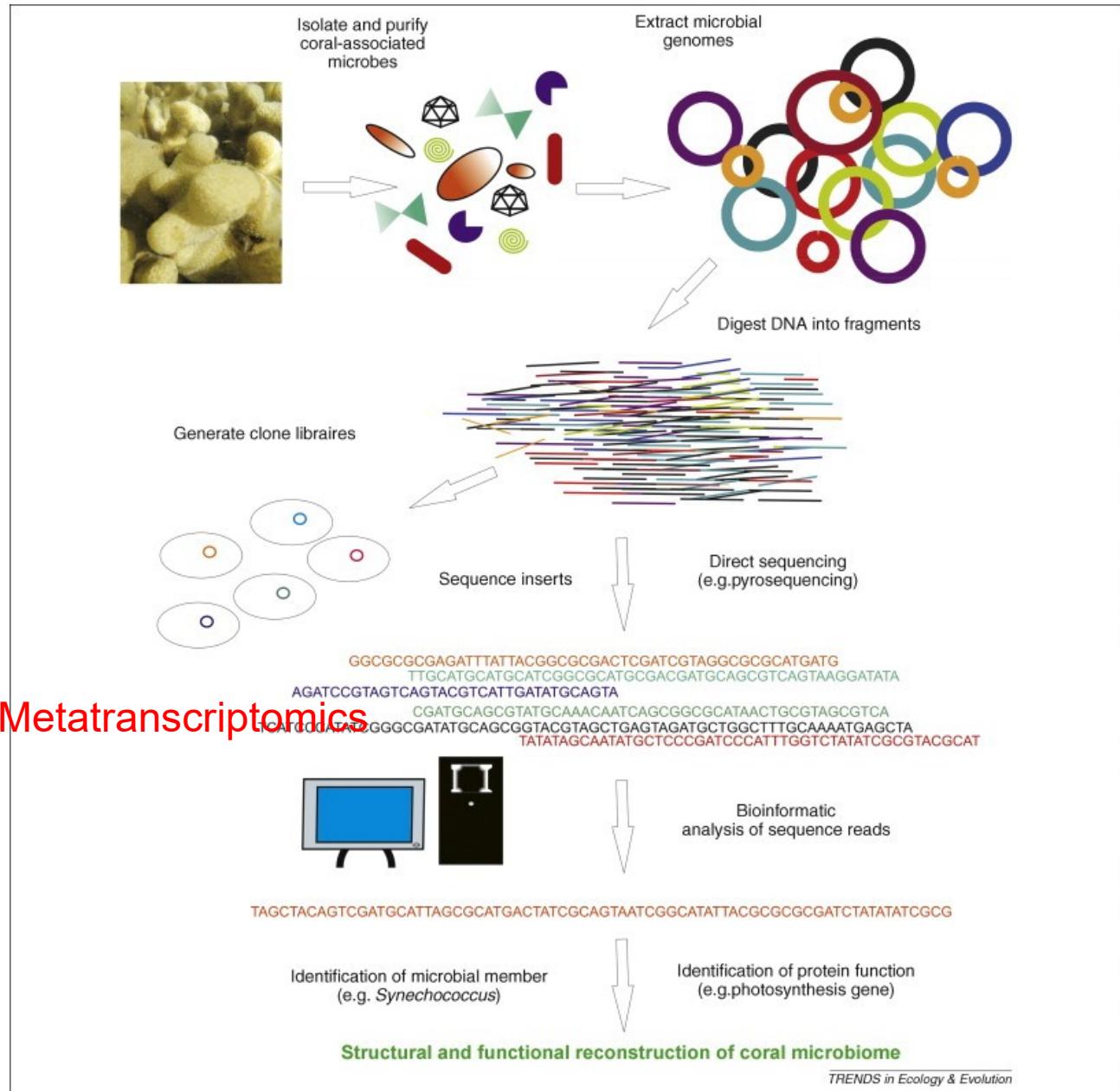
- Genome Sequencing
  - De novo sequencing
  - Resequencing
  - Targeted (re)sequencing
  - Mitochondrial sequencing
  - Mutation detection
  - Amplicon sequencing
  - Amplicon Cancer Panel
  - Phylogenomics- Phylogenetic
- Transcript Expression Profiling
  - RNA sequencing
  - miRNA sequencing
  - Deep-SAGE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

### Study of genomes recovered from environmental samples



# Applications

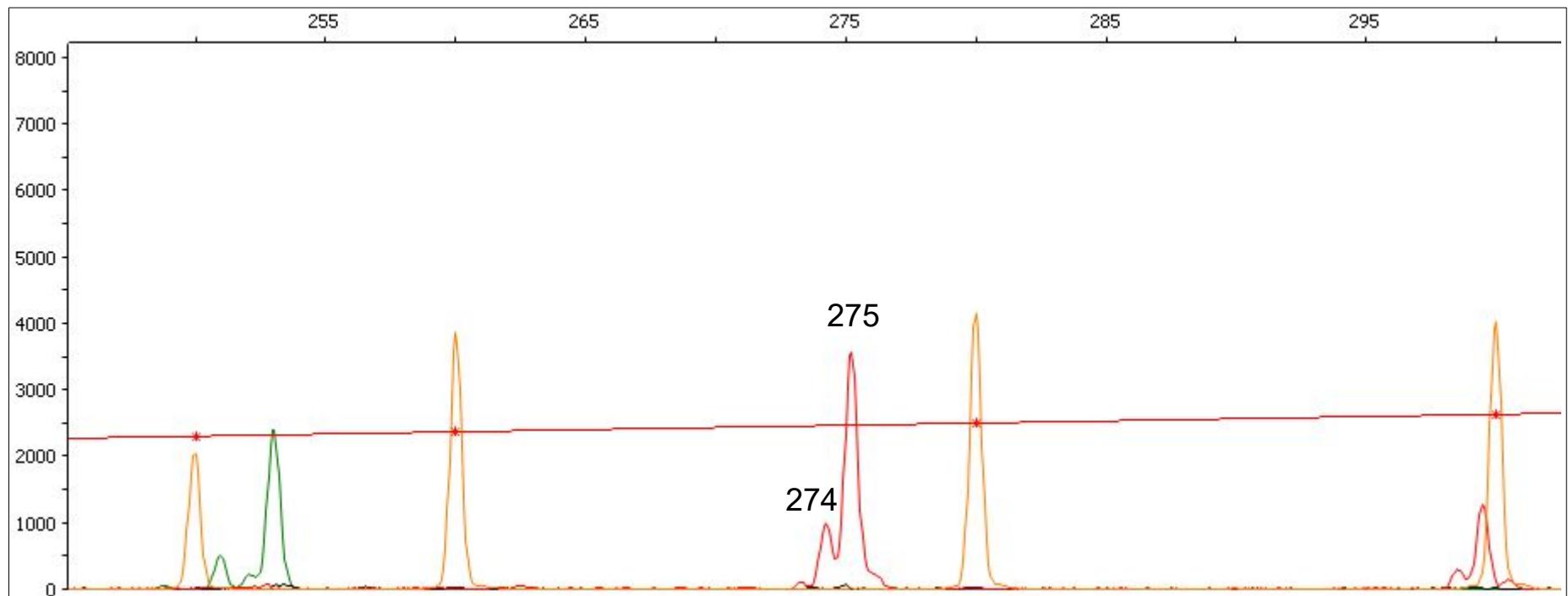
- Genome Sequencing
    - De novo sequencing
    - Resequencing
    - Targeted (re)sequencing
    - Mitochondrial sequencing
    - Mutation detection
    - Amplicon sequencing
    - Amplicon Cancer
    - Phylogenomics
  - Transcript Expression
    - RNA sequencing
    - miRNA sequencing
    - Deep-SAGE
    - Deep-CAGE
    - PAS: polyadenylation
  - Transcription factor binding
  - Structural variation
  - Metagenomics / Metabarcoding
  - **Microsatellite sequencing**
  - Genetic marker discovery
    - Microsatellite development
    - RADSeq (Restriction Site Associated DNA Sequencing)
    - RRLs (Reduced Representation Linkage Analysis)
    - GBS (Genotyping-by-Sequencing)
  - ...

## Applications

### Microsatellite sequencing

Based on sizing on capillary sequencer  
→ one number as result

Small peaks before or after big one ???



# Applications

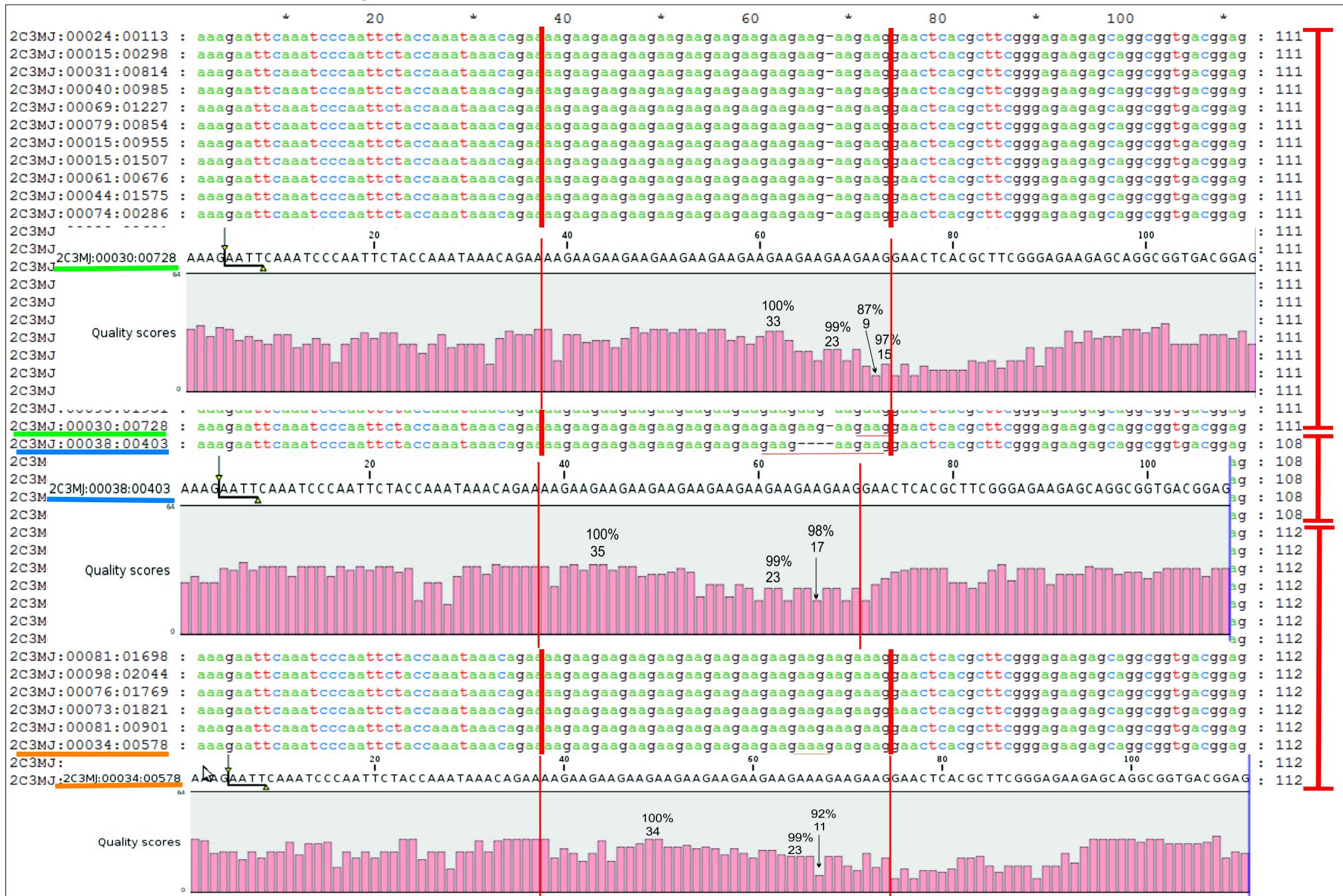
# Microsatellite sequencing

Fragment of 111 bases (all identical), fragment of 108 bases (all identical), fragment of 112 bases (some different)

# Applications

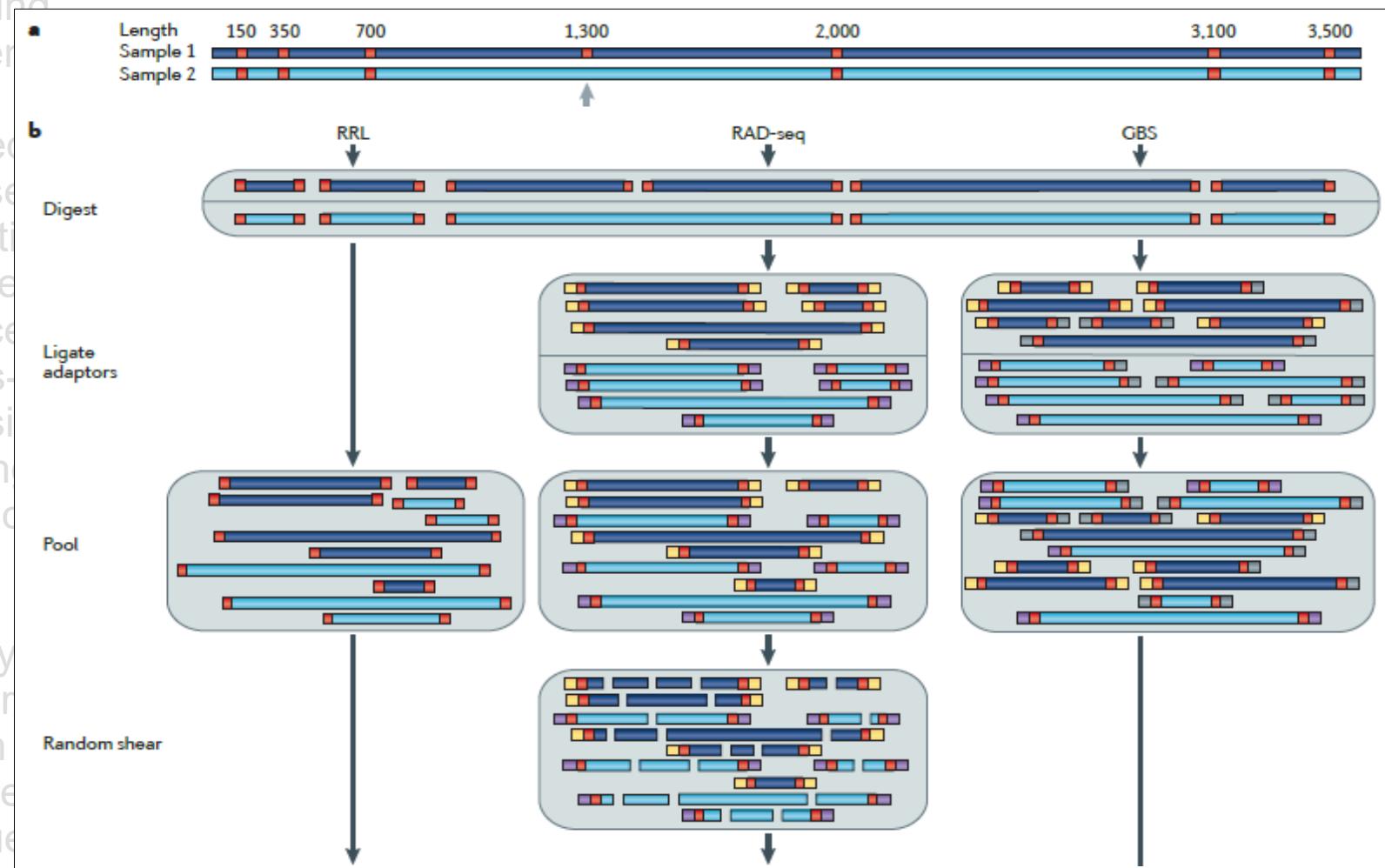
## Microsatellite sequencing

135/165



## Applications

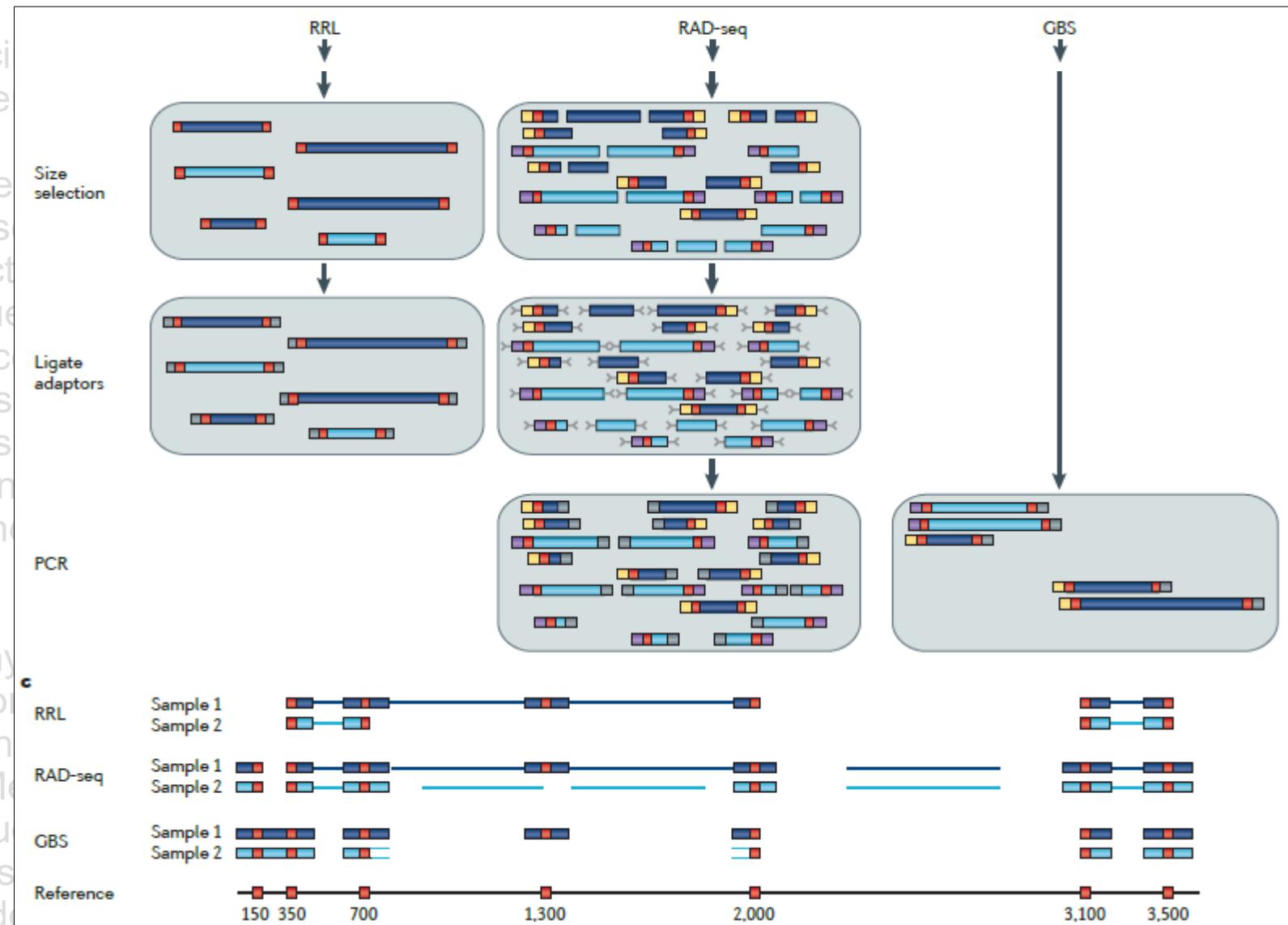
- Genome Sequencing
  - De novo sequencing
  - Resequencing
  - Targeted (re)sequencing
  - Mitochondrial sequencing
  - Mutation detection
  - Amplicon sequencing
  - Amplicon Cancer
  - Phylogenomics
- Transcript Expression
  - RNA sequencing
  - miRNA sequencing
  - Deep-SAGE
  - Deep-CAGE
  - PAS: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



Genome-wide genetic marker discovery and genotyping using next-generation sequencing.  
Davey et al.  
NATURE REVIEWS, GENETICS VOLUME 12, JULY 2011

## 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 factors
- Structural variation
- Metagenomics / Microbiome
- Microsatellite sequencing
- Genetic marker discovery
  - Microsatellite diversity



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

Genome-wide genetic marker discovery and genotyping using next-generation sequencing.  
Davey et al.  
NATURE REVIEWS, GENETICS VOLUME 12, JULY 2011

## Run types

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



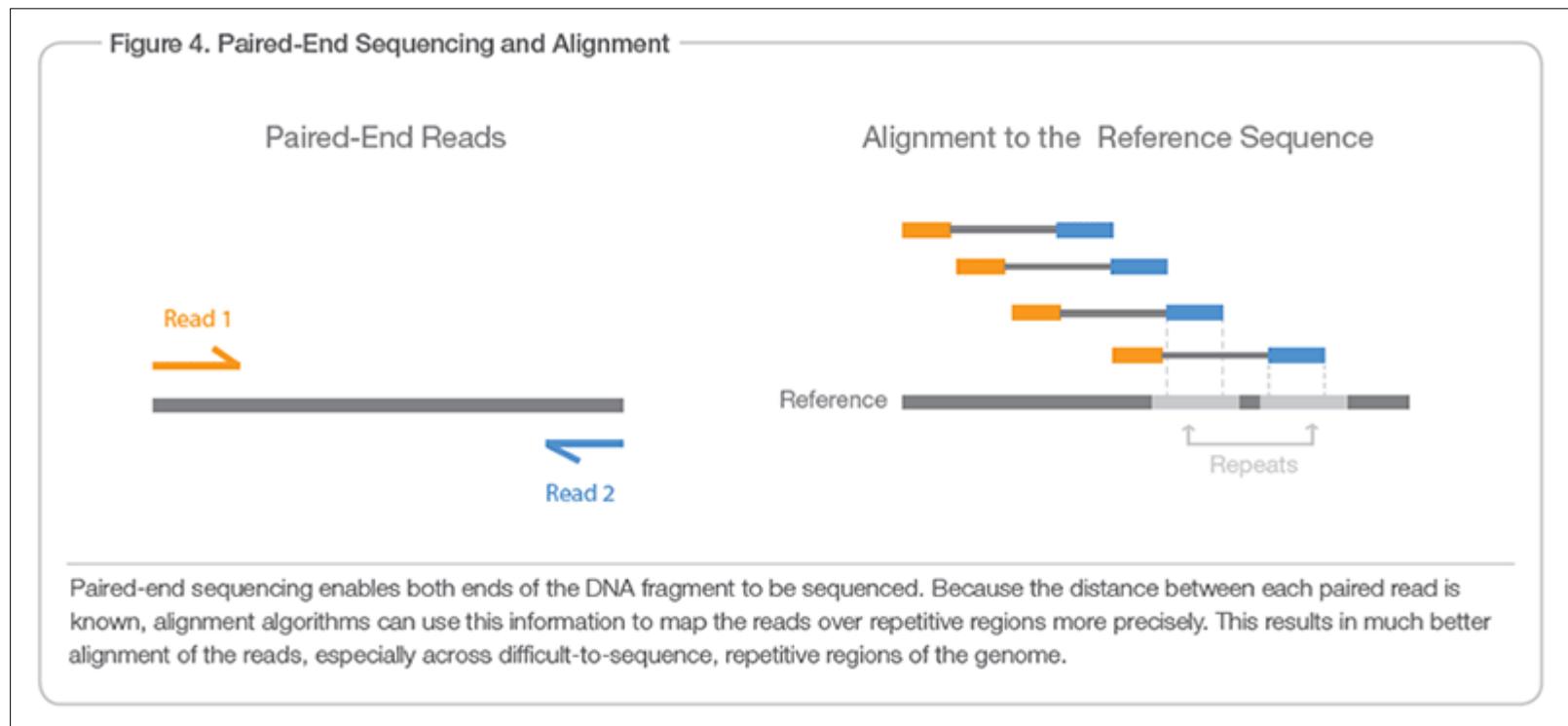
OR



## Run types

139/165

- 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,

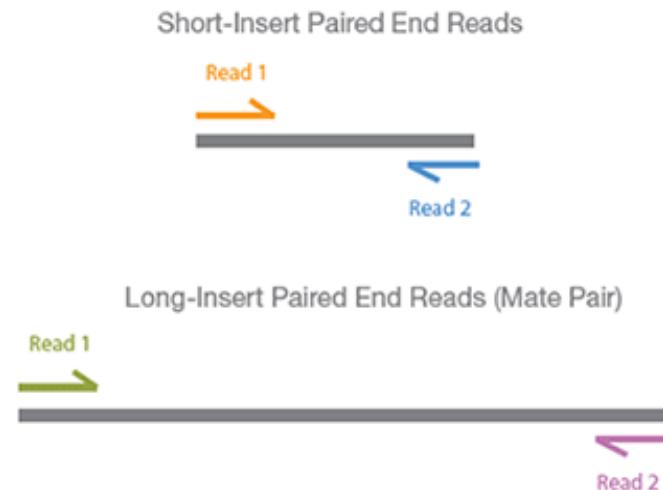
## Run types

140/165

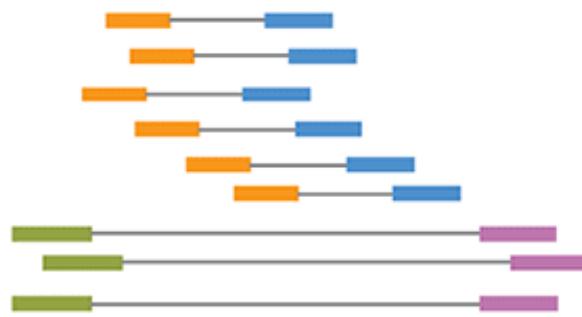
- Single-end sequencing
- Paired-end sequencing
- Mate-pair sequencing

• B

Figure 5. De Novo Assembly with Mate Pairs



De Novo Assembly



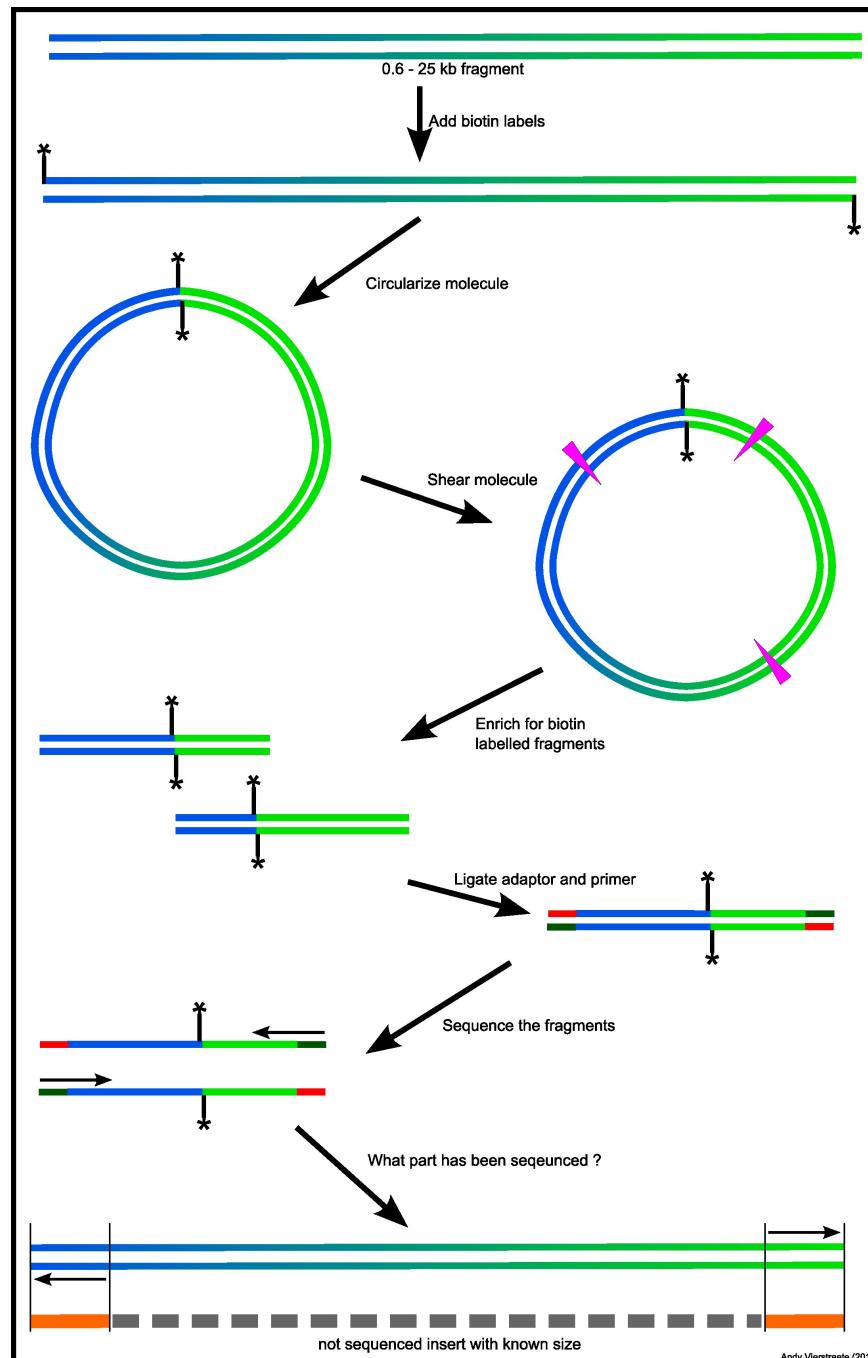
Using a combination of short and long insert sizes with paired-end sequencing results in maximal coverage of the genome for de novo assembly. Because larger inserts can pair reads across greater distances, they provide a better ability to read through highly repetitive sequences and regions where large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

Long inserts possible: 0.6-25 kb

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

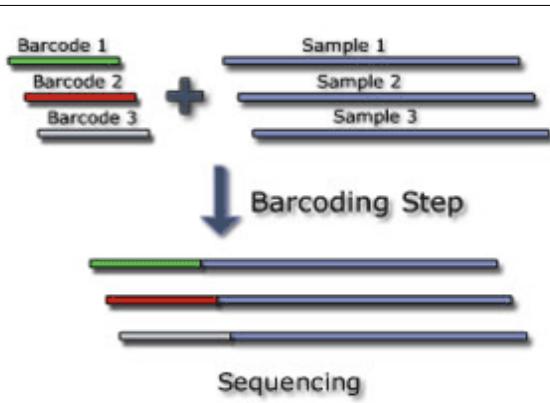
## Run types

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

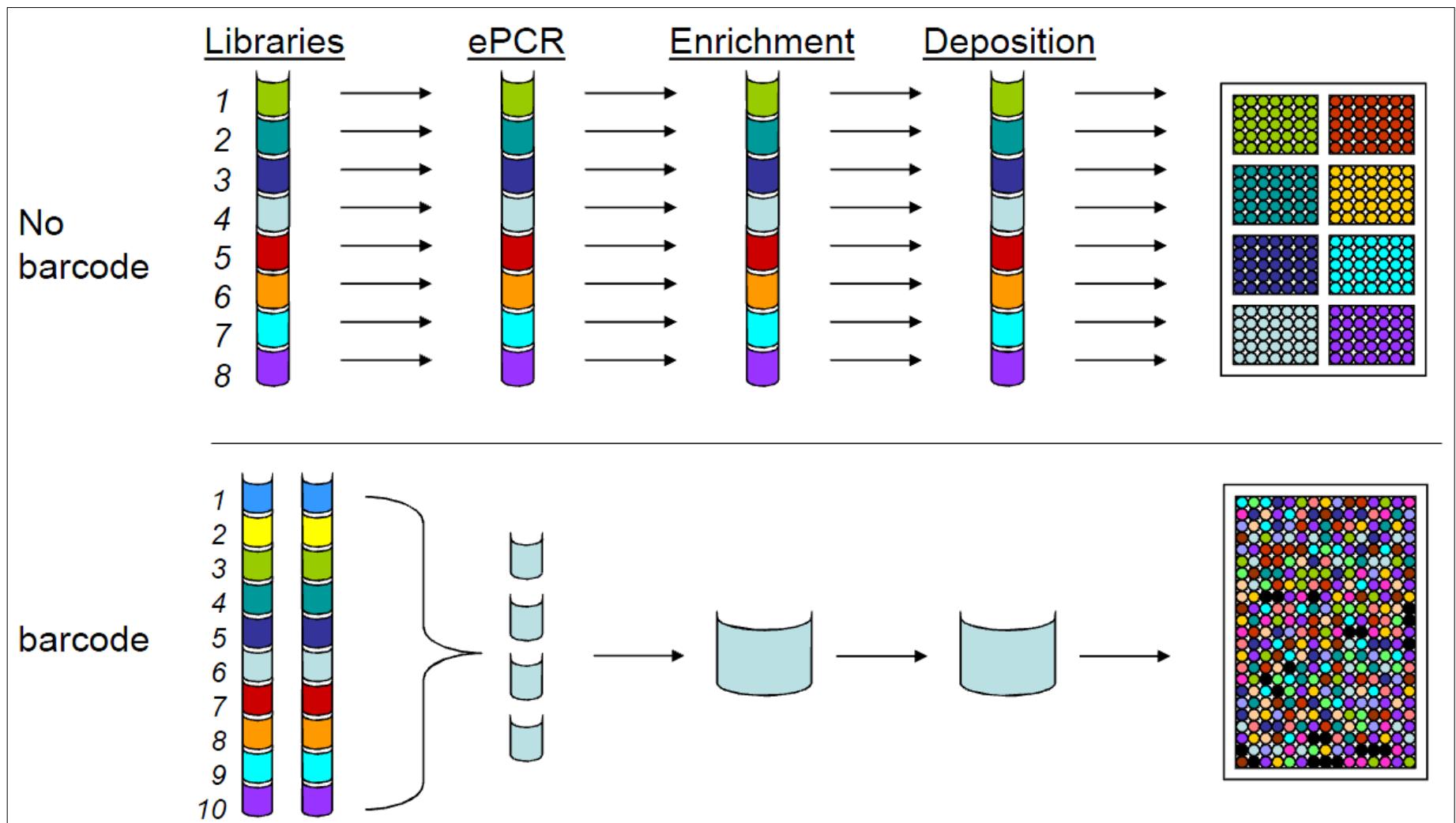


## Run types

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



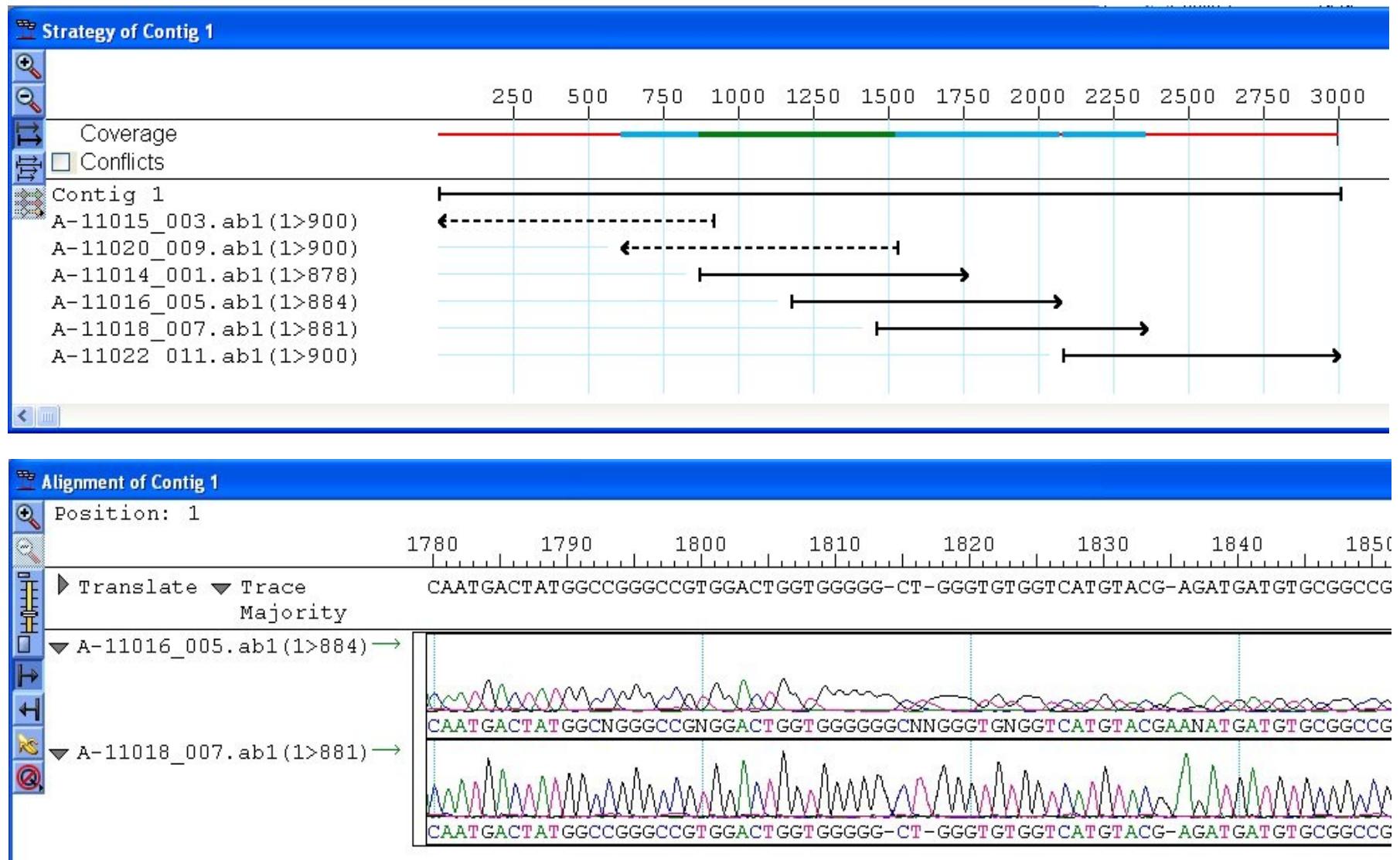
10-12 bp unique sequence is added tot the DNA



## Data analyses

143/165

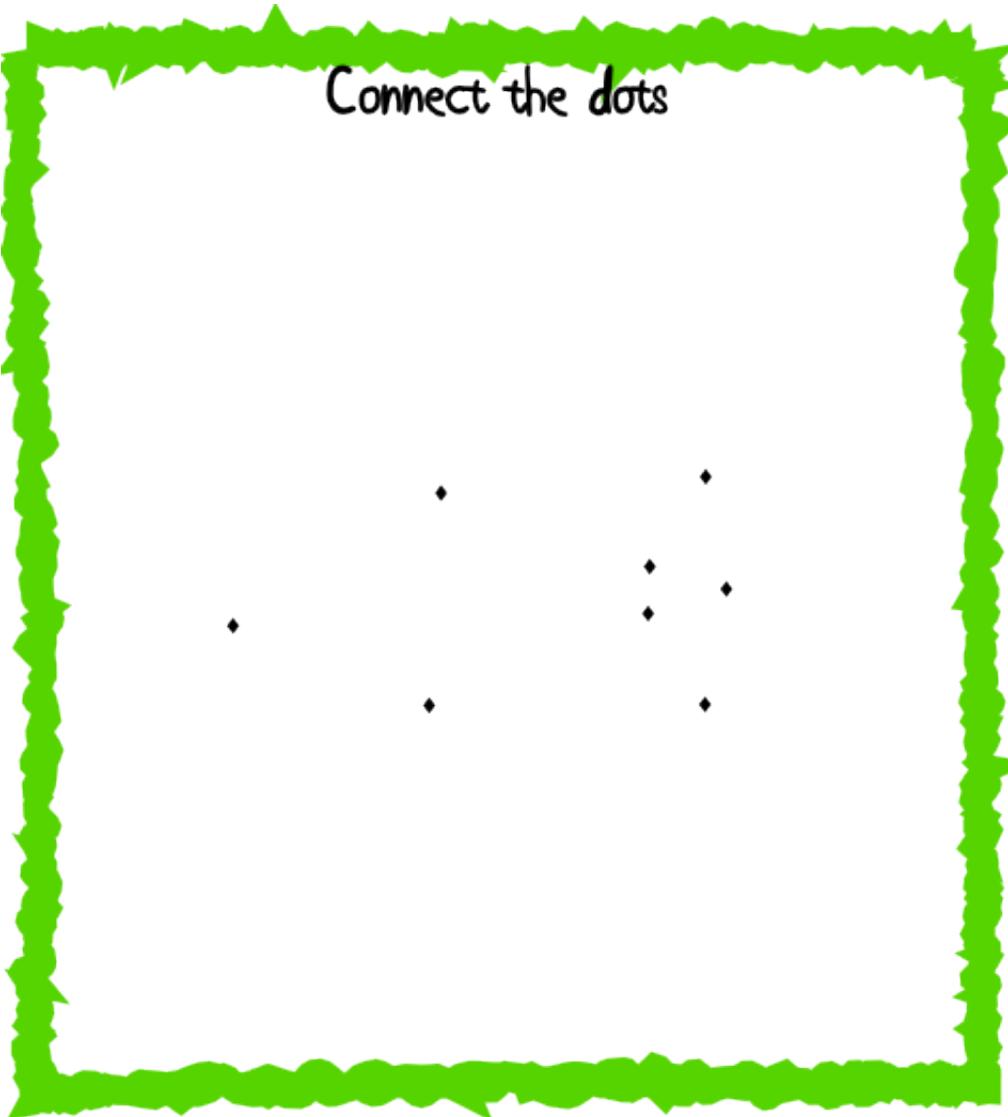
Sanger sequencing: e.g.: one gene sequenced with 6 primers



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

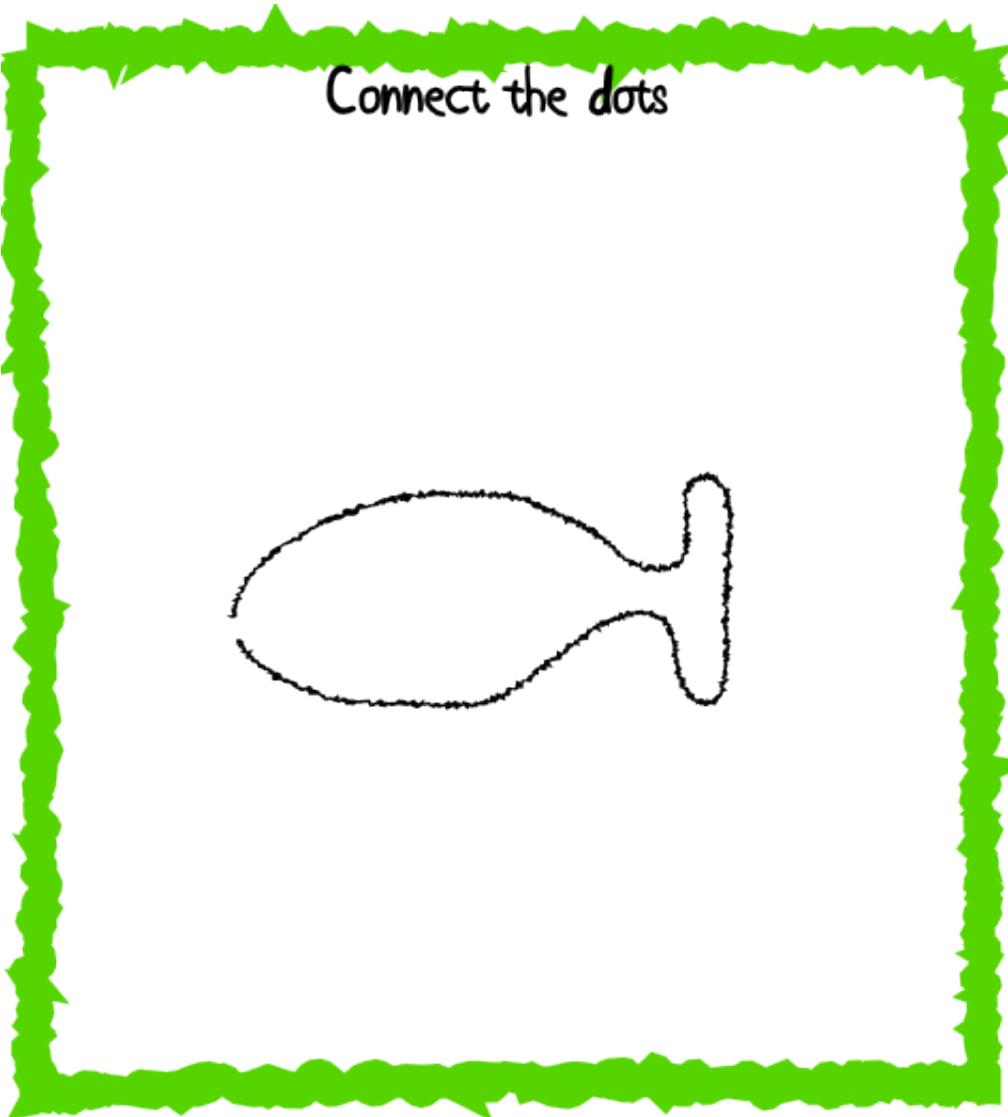
## Data analyses

Sanger sequencing: simplified:



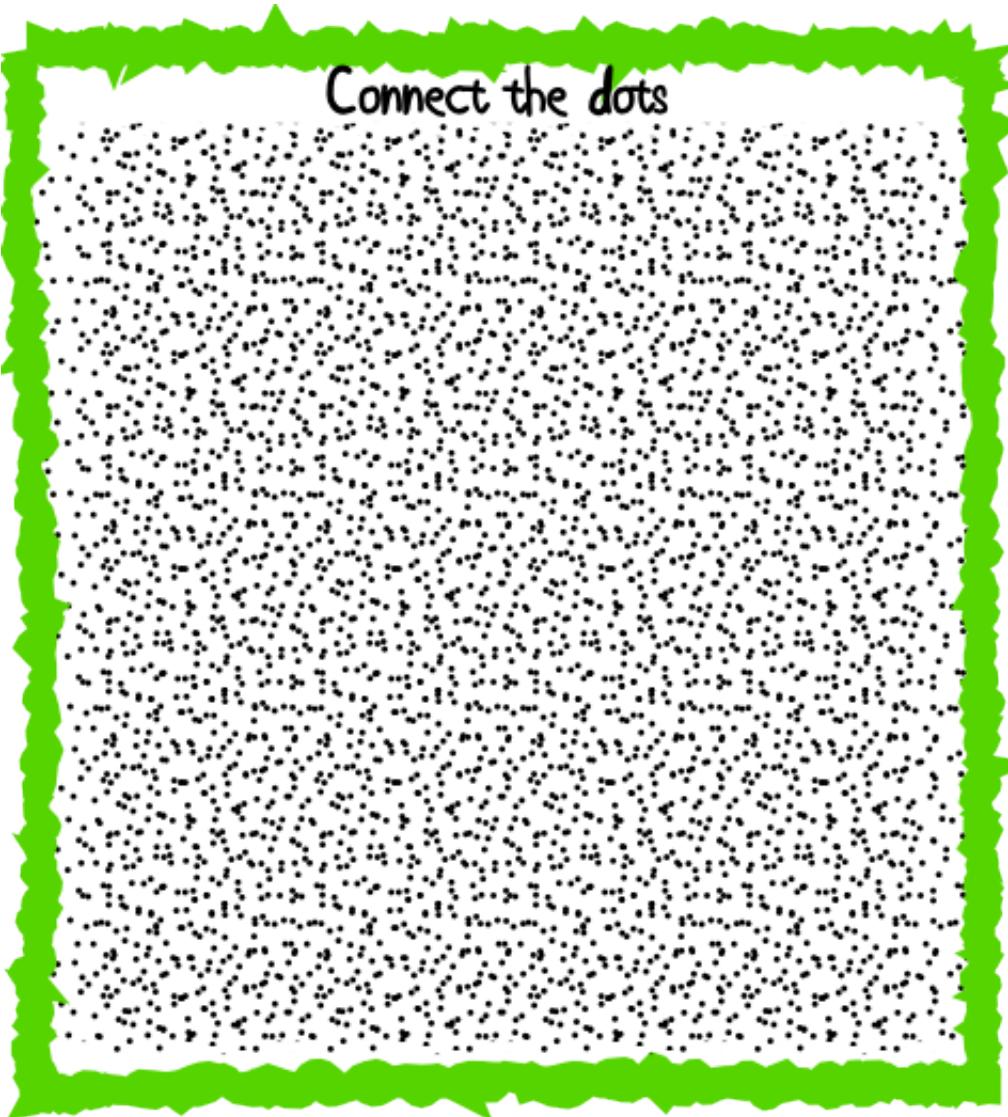
## Data analyses

Sanger sequencing: simplified:



## Data analyses

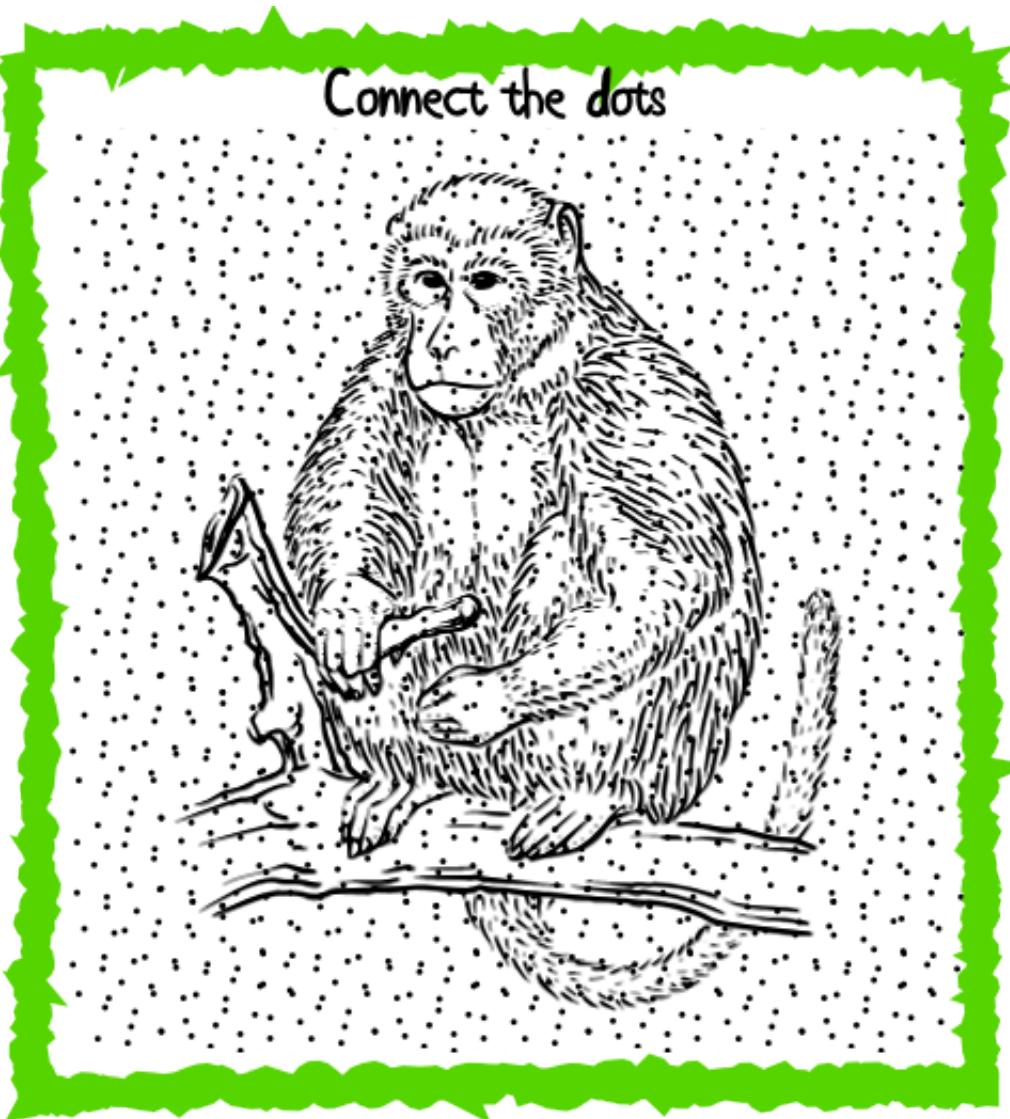
Next Generation sequencing: simplified:



Impossible to assemble manually

## Data analyses

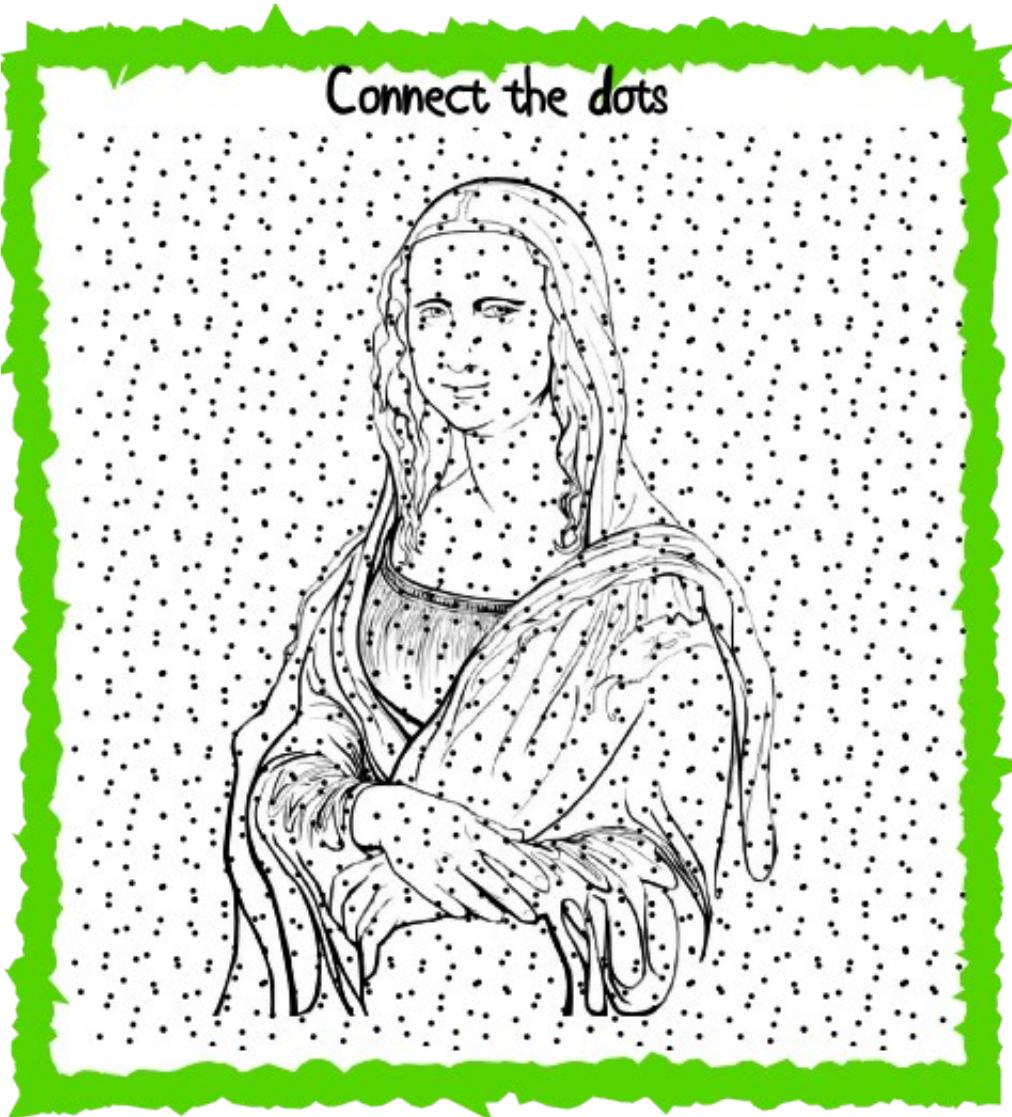
Next Generation sequencing: simplified:



## Data analyses

148/165

Next Generation sequencing: simplified:



Same dataset, different parameters

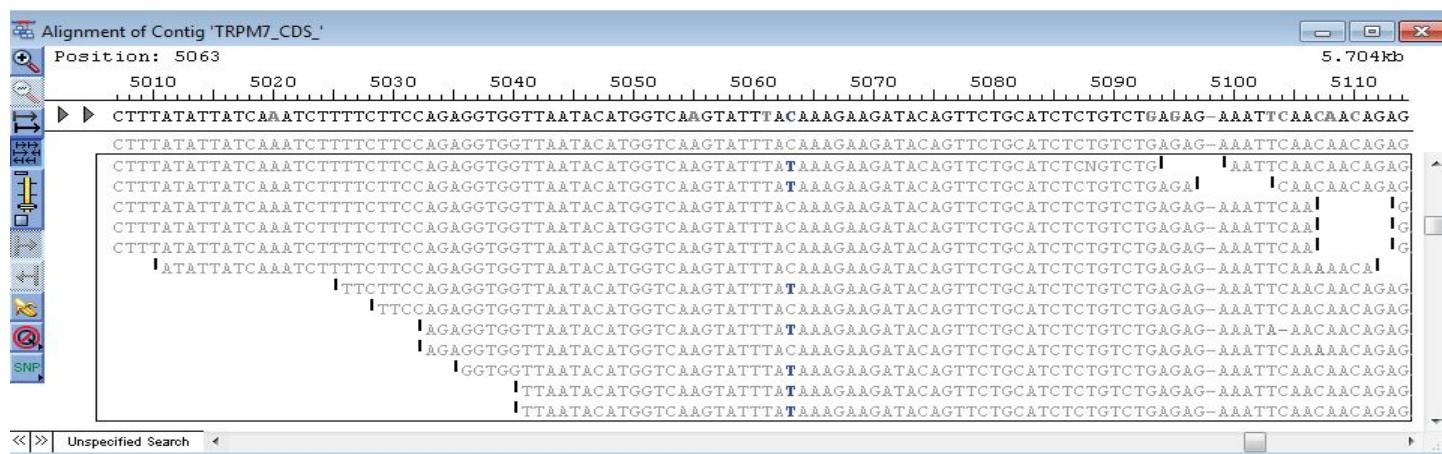
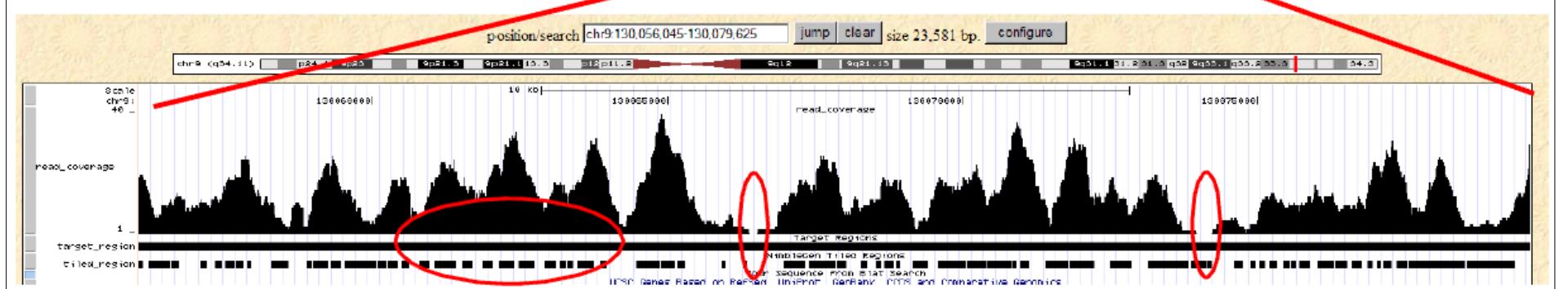
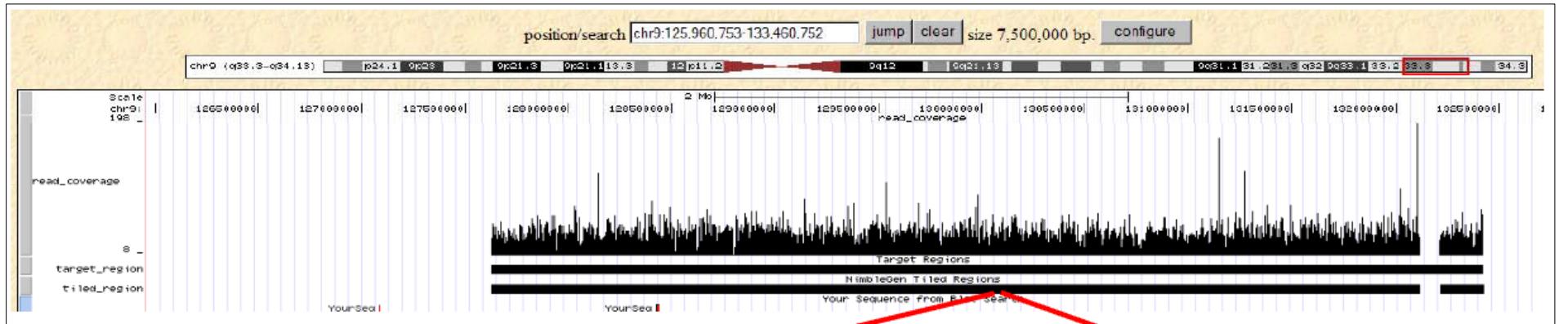
*Determining which assembly is the best is not an easy question. (Monya Baker. Nature Methods Volume 9 No.4, 333-337 (2012))*

## Data analyses

149/165

Next Generation sequencing:

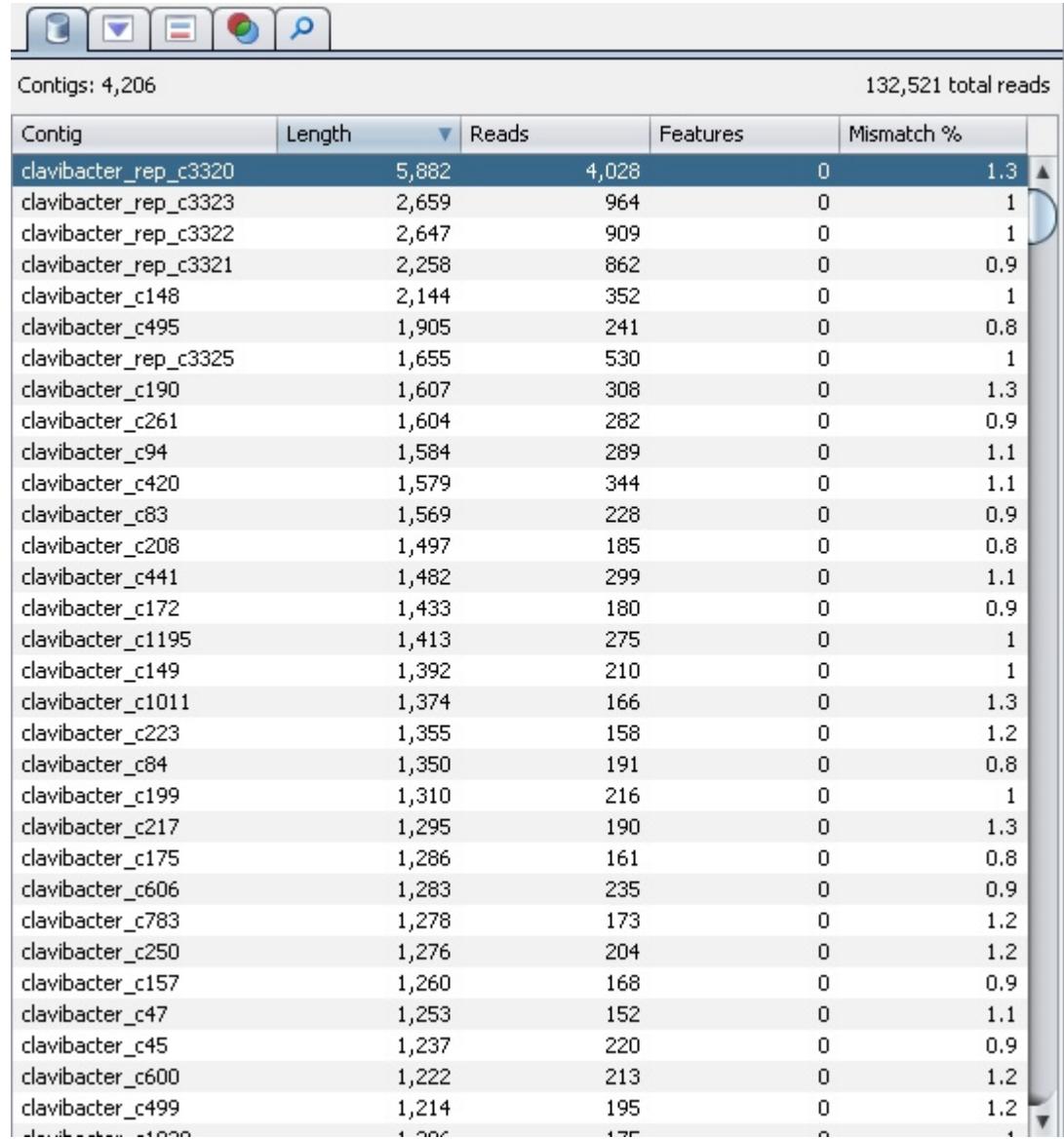
Impossible to check manually



## Data analyses

Next Generation sequencing:

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



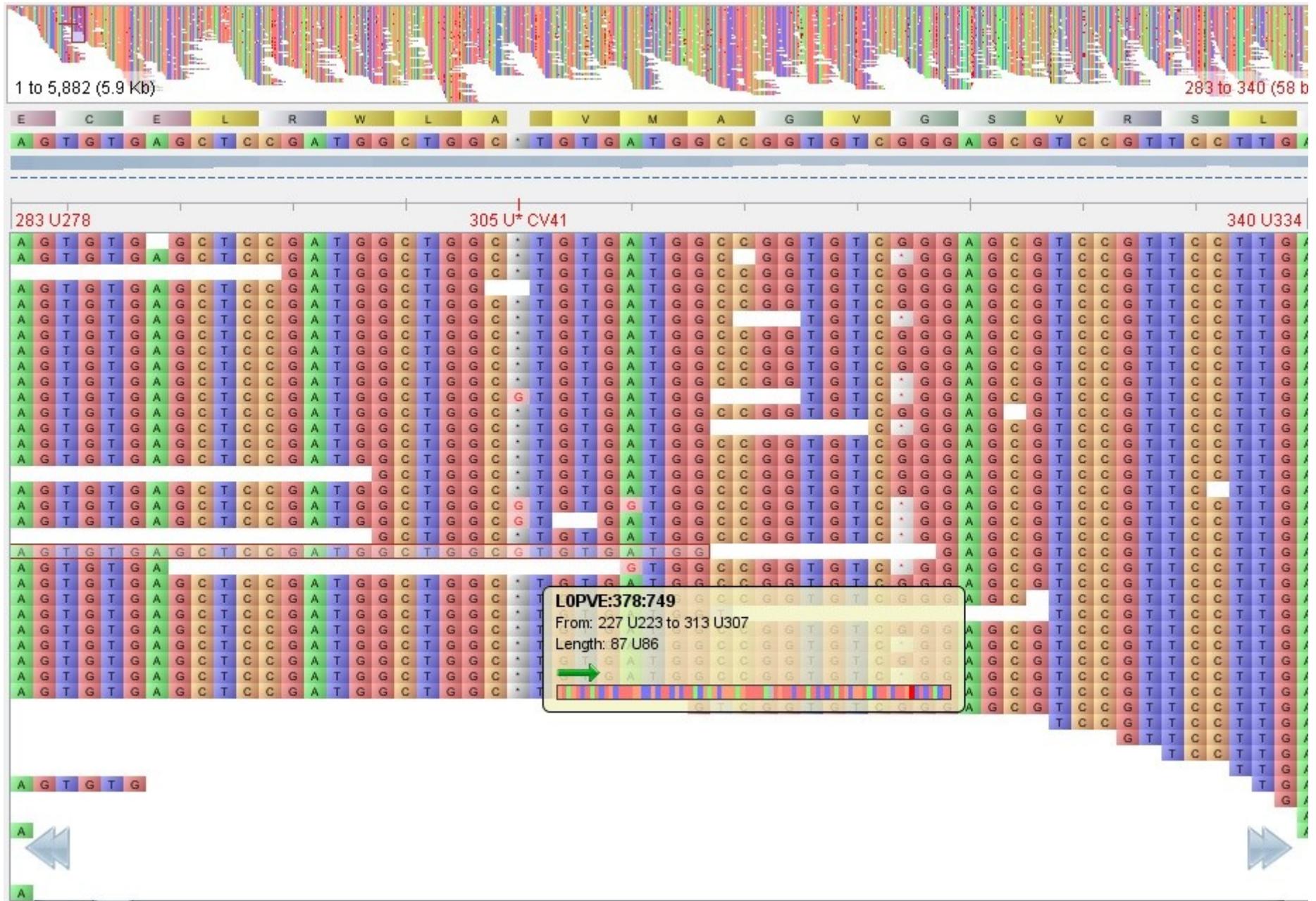
The screenshot shows a software window with a toolbar at the top containing icons for file operations, search, and other functions. Below the toolbar, the text "Contigs: 4,206" and "132,521 total reads" is displayed. A table follows, with columns labeled "Contig", "Length", "Reads", "Features", and "Mismatch %". The data is sorted by length in descending order. The first few rows of the table are as follows:

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
clavibacter_c400	1,205	175	0	1.2

# Data analyses

## Next Generation sequencing:

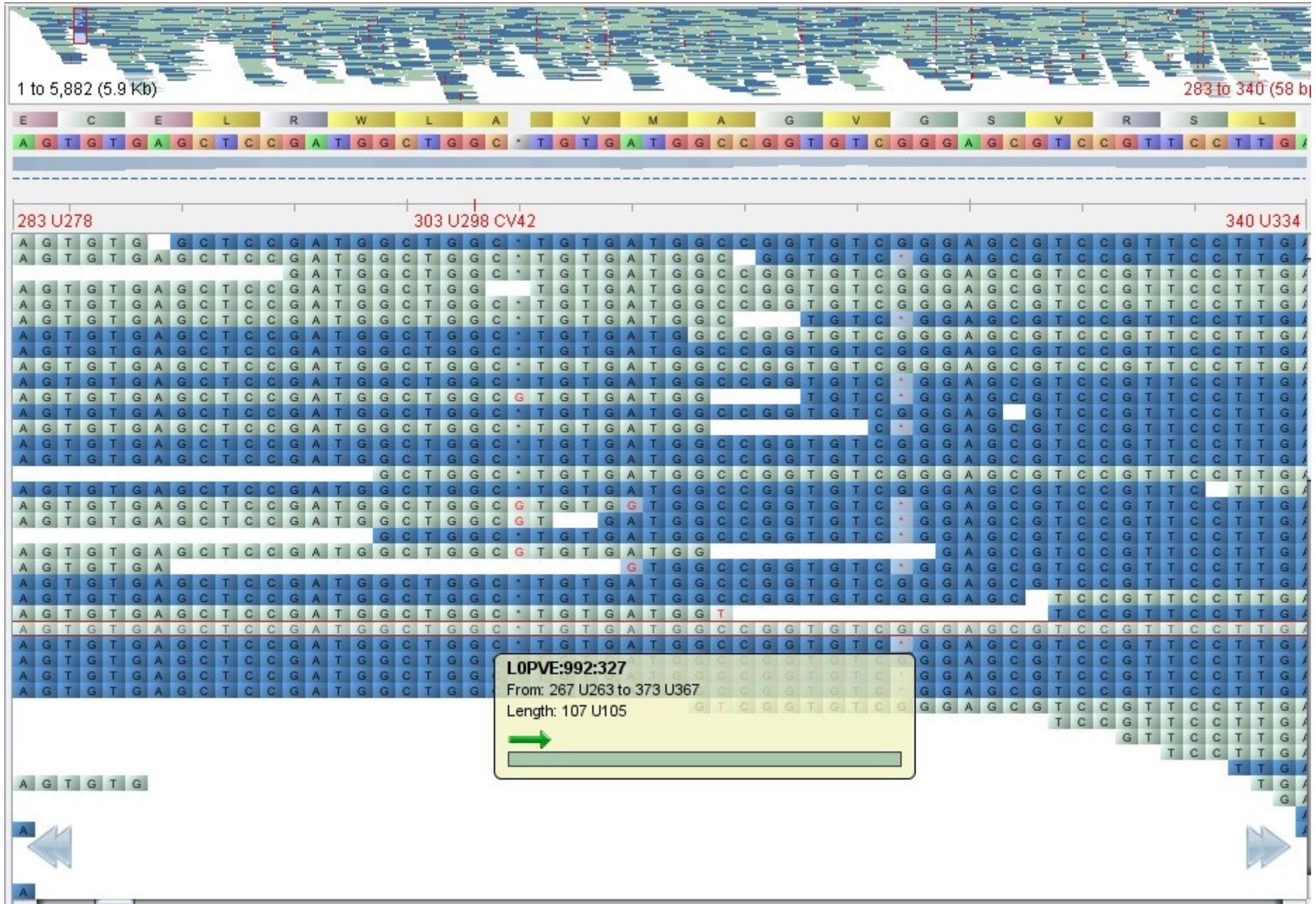
## Assembly of the largest contig



# Data analyses

## Next Generation sequencing:

Light blue: forward  
dark blue: reverse sequence



## Data analyses

Next Generation sequencing:

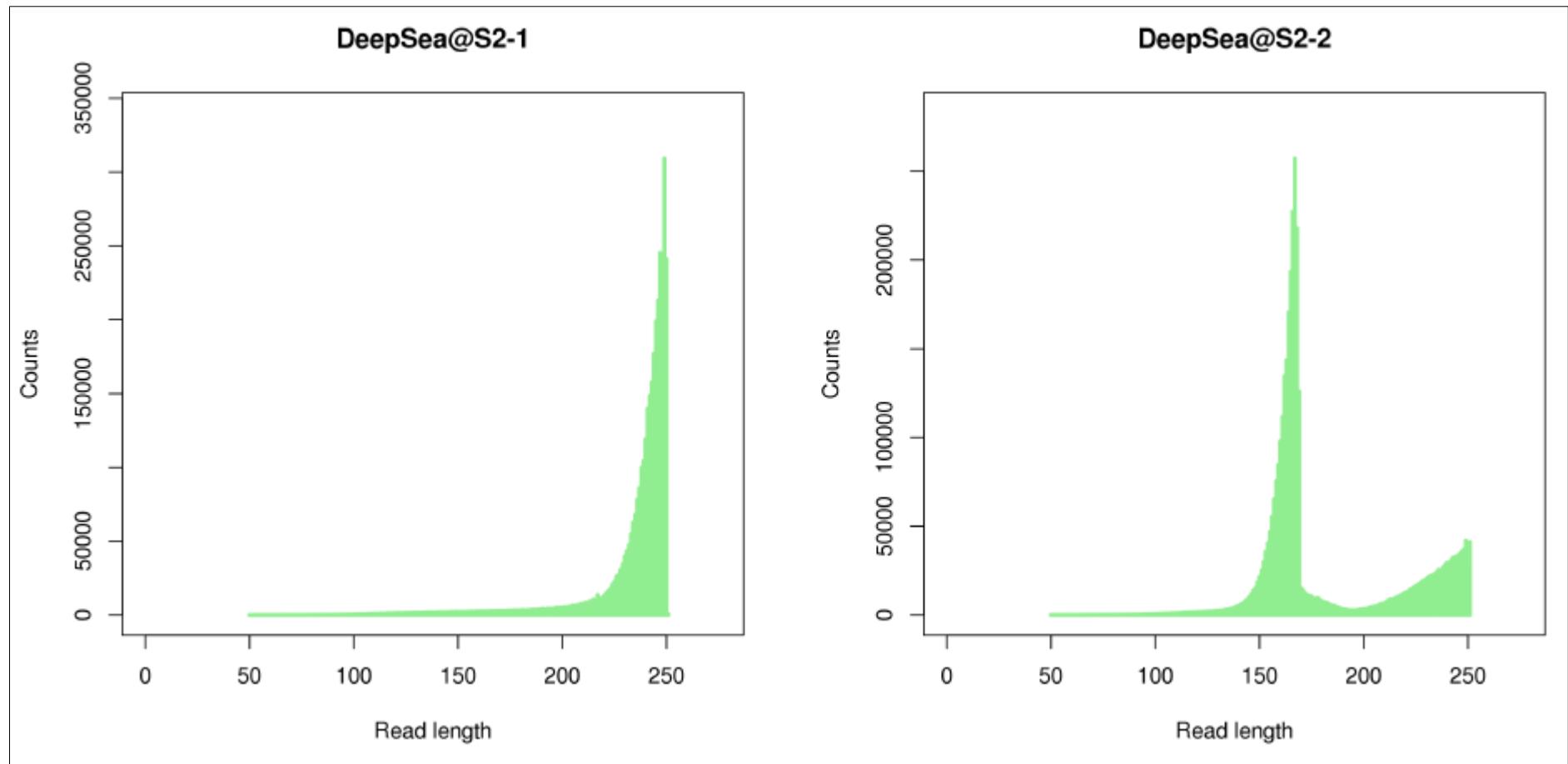
CV: coverage of a specific part



## Data analyses

Next Generation sequencing: expectations from a run.

MiSeq run 2 x 250 bp



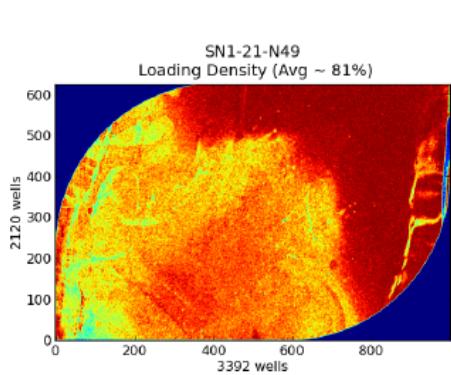
## Data analyses

Next Generation sequencing: expectations from a run.

Ion Torrent PGM run 400 bp

990 M                    58  
Total Bases      Key Signal

81%  
ISP Loading  
**ISP Density**



3,360,432

Total Reads

66%

Usable Reads

### ISP Summary

81% Loading	5,114,378	19% Empty Wells
100% Enrichment	5,112,504	0% No Template
71% Clonal	3,622,186	29% Polyclonal
93% Final Library	3,360,432	0% Test Fragments 3% Adapter Dimer 4% Low Quality

294 bp

346 bp

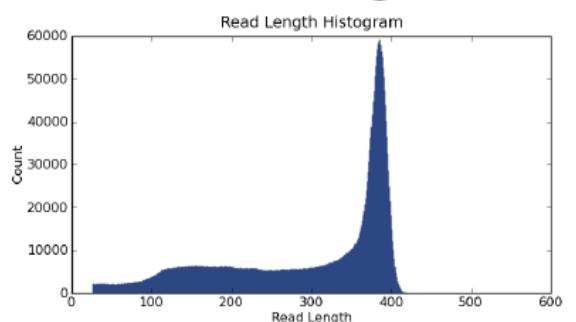
384 bp

Mean

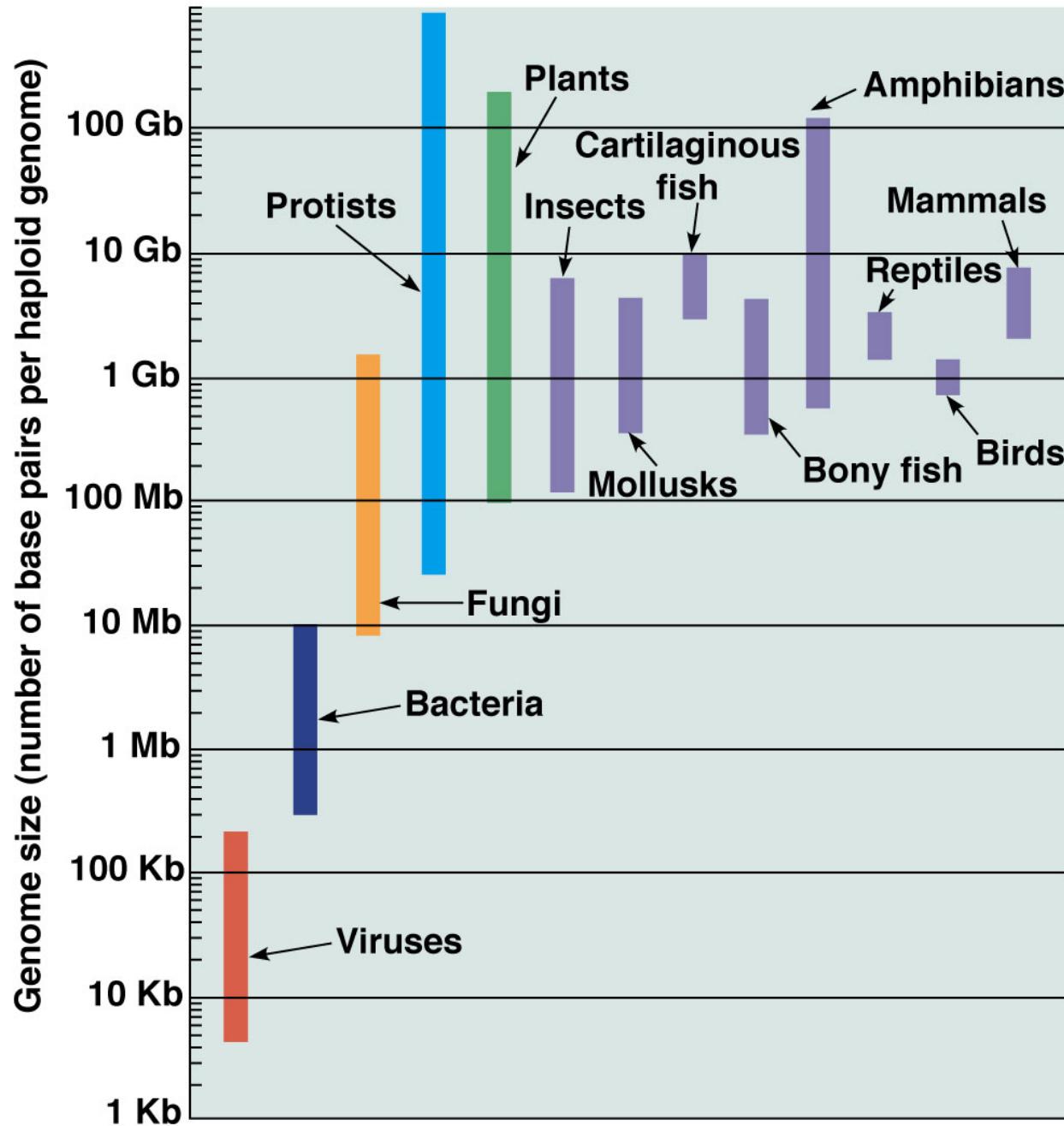
Median

Mode

### Read Length



## 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

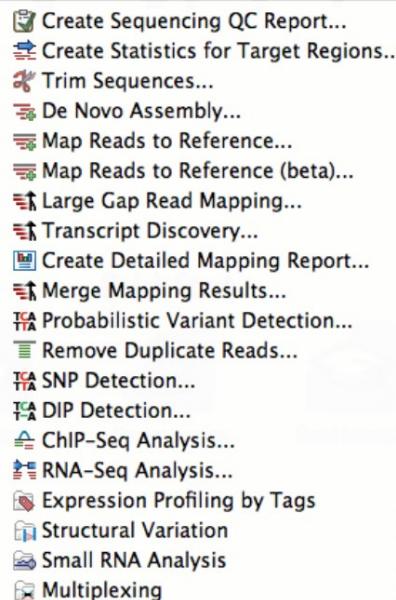
Lots of programs available: commercial or open source

Pro's:

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

Con's:

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

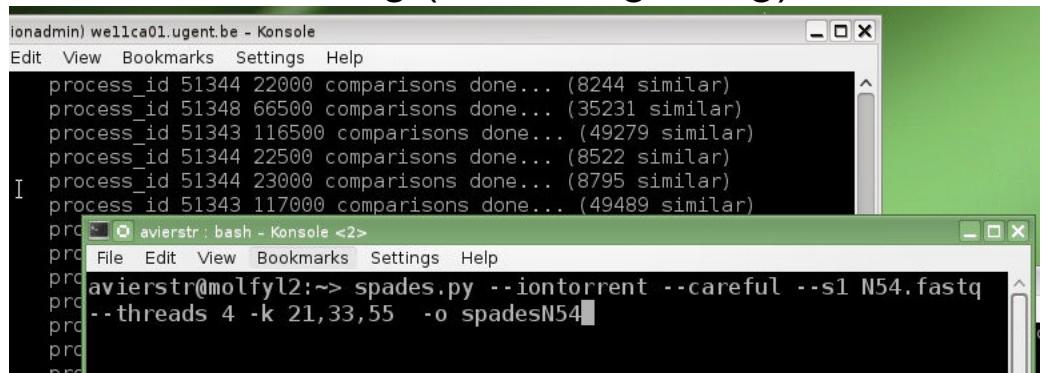


Pro's:

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

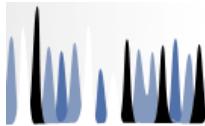
Con's:

- 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)



## Data analyses

158/165

 **SEQanswers**  
the next generation sequencing community

[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 » | ▾

01-24-2008, 08:19 AM #1 Thread Tools ▾ | Search this Thread ▾

**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.

# Data analyses

159/165

- \* [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 Genenetc. 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 Strömborg 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 [Synamatix](#). 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,

## Data analyses

160/165

- \* [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. Linus.
- \* [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](#)

## Data analyses

161/165

\* [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

# Data analyses

162/165

The screenshot shows the Galaxy web interface at [galaxy.nbic.nl/galaxy/](http://galaxy.nbic.nl/galaxy/). The left sidebar lists various bioinformatics tools categorized under 'Tools'. The main panel displays the 'Upload File (version 1.1.3)' tool. It includes fields for 'File Format' (set to 'Auto-detect'), 'File' (with a 'Browse...' button), and 'URL/Text' (with a large input area). A note states: '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.)'. Below this is a section for 'Files uploaded via FTP' with a table header 'File', 'Size', and 'Date'. A note says: 'Please [create](#) or [log in](#) to a Galaxy account to view files uploaded via FTP.' Another note says: 'This Galaxy server allows you to upload files via FTP. To upload some files, log in to the FTP server at [galaxy.nbic.nl](http://galaxy.nbic.nl) using your Galaxy credentials (email address and password).' There is also an option 'Convert spaces to tabs:' with a checkbox 'Yes'. A 'Genome:' dropdown is set to 'Arabidopsis thaliana (tair9)'. A 'Execute' button is present. The right sidebar shows an empty 'History' pane with a message: 'Your history is empty. Click 'Get Data' on the left pane to start'.

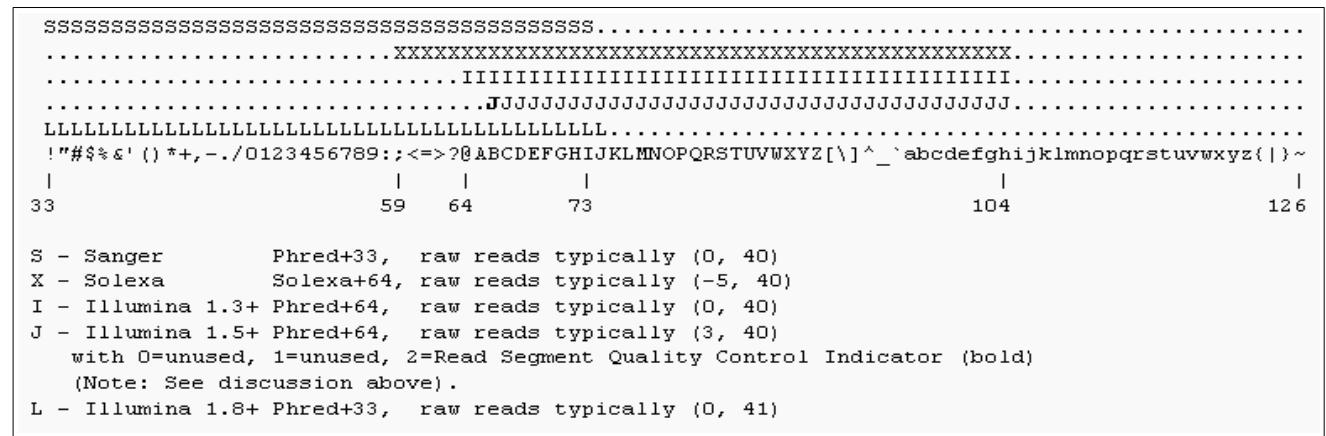
# Data analyses

163/165

## FastQ format:

Header 1 @DBRHHJN1\_0097:1:1:1111:1176#CGATGT/1  
Sequence CCTTCTGGGGCAAGACTTGAGCAAAGACAATGAAAGTAAAGTGCCTGAAAACCGAAAATCGTATGACTTTTTAATGTTGATCAGGACCAT  
Header 2 +DBRHHJN1\_0097:1:1:1111:1176#CGATGT/1  
Quality score gggggggggcgggggggdggggggggdaggdeacedee\ \_eed^dee [ decbgegXdebee^ddddddUd [ceacI \ ^dadbbccc^bbBBBBBBB  
@DBRHHJN1\_0097:1:1:1200:1169#CGATGT/1  
TGGATGTAATTAAATTAAATTAGTAGGTCCCTCTTTGTCTGAAACTAGGGCTCTATTAGAAATATAATAGCACTTTGAATAAAACTGAAAGG  
+DBRHHJN1\_0097:1:1:1200:1169#CGATGT/1  
ffffdeggegggggeegffdgadddeeeeeeggeggefegegggdfggfg\_dd\ddu\_\ \_\_fggf\gdddcccbdb\ffefggdc\_cfdedffffd  
@DBRHHJN1\_0097:1:1:1201:1199#CGATGT/1  
GGTAAAGACTAAATGCTAGACAAAGACACCATCAAAACTCCTAGAAAAGAACATAGGCAAAACACTCTGTATCAACATTACAAATGTTCTCAGGT

## Quality score using the ascii table



Depending on the application:

- Pre-assembly:  
Quality trimming, demultiplexing, adapter-barcode removal, data statistics, ...
  - Assembly:  
denovo, mapping to reference
  - Post-assembly:  
Quality Control, statistics, SNP calling, annotation, blasting,...

## Considerations

- 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 !

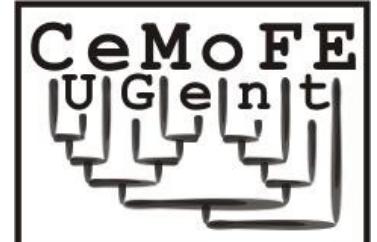
<http://users.ugent.be/~avierstr/>

Andy.Vierstraete@ugent.be



CTAGGTAGCTAGTCG  
GCTLIFECISGATAG  
C4-LETTERWORDT  
GCTATATCGTAGCTG

[WWW.DNA.UGENT.BE](http://WWW.DNA.UGENT.BE)



Center for Molecular Phylogeny and Evolution