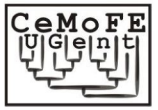
Andy Vierstraete,
Department of Biology,
Ghent University. June 2012





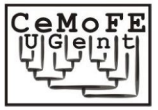
Next Generation Sequencing

for

DUMMIES

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012





- History and future of DNA sequencing
- Workflow
- Different platforms
- Applications
- Run types
- Data analysis
- Considerations
- What do we offer and what does it cost ?

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

In the beginning there was nothing.

God said : 'Let there be light ! '

There was still nothing, but now you could see it.

(Dave Thomas)

History and future of DNA sequencing

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

1953 : Discovery of DNA structure by Watson and Crick

1973 : First sequence of 24 bp published

History and future of DNA sequencing

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

1953 : Discovery of DNA structure by Watson and Crick

1973 : First sequence of 24 bp published

Proc. Nat. Acad. Sci. USA
Vol. 70, No. 12, Part I, pp. 3581–3584, December 1973

The Nucleotide Sequence of the *lac* Operator

(regulation/protein–nucleic acid interaction/DNA–RNA sequencing/oligonucleotide priming)

WALTER GILBERT AND ALLAN MAXAM

Department of Biochemistry and Molecular Biology, Harvard University, Cambridge, Massachusetts 02138

Communicated by J. D. Watson, August 9, 1973

ABSTRACT **The *lac* repressor protects the *lac* operator against digestion with deoxyribonuclease. The protected fragment is double-stranded and about 27 base-pairs long. We determined the sequence of RNA transcription copies of this fragment and present a sequence for 24 base pairs. It is:**

5′--T G G A A T T G T G A G C G G A T A A C A A T T 3′
3′--A C C T T A A C A C T C G C C T A T T G T T A A 5′

The sequence has 2-fold symmetry regions; the two longest are separated by one turn of the DNA double helix.

bind again to the repressor, and is a
Here we shall describe its sequence.

METHODS

Sonicated DNA Fragments. Sonicat
were made by growing a temperatu
λc1857p*lac*5S7 at 34° in a glucose–50
(pH 7.4) medium in 3 mM phosphat
min at a cell density of 4 × 10⁸/ml, t
pending the cells at a density of 8 ×

system

ystem

ems

cule reads

History and future of DNA sequencing

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

- 1953 : Discovery of DNA structure by Watson and Crick
- 1973 : First sequence of 24 bp published
- 1977 : Sanger sequencing method published
- 1980 : Nobel Prize Wally Gilbert and Fred Sanger
- 1982 : Genbank started
- 1983 : Development of PCR
- 1987 : 1st automated sequencer : Applied Biosystems Prism 373
- 1996 : Capillary sequencer : ABI 310
- 1998 : Genome of Caenorhabditis elegans sequenced
- 2000 : Human genome sequenced
- 2005 : 1st 454 Life Sciences Next Generation Sequencing system : GS 20 System
- 2006 : 1st Solexa Next Generation Sequencer : Genome Analyzer
- 2007 : 1st Applied Biosystems Next Generation Sequencer : SOLiD
- 2009 : 1st Helicos single molecule sequencer : Helicos Genetic Analyser System
- 2011 : 1st Ion Torrent Next Generation Sequencer : PGM
- 2011 : 1st Pacific Biosciences single molecule sequencer : PacBio RS Systems
- 2012 : Oxford Nanopore Technologies demonstrates ultra long single molecule reads

Different platforms

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

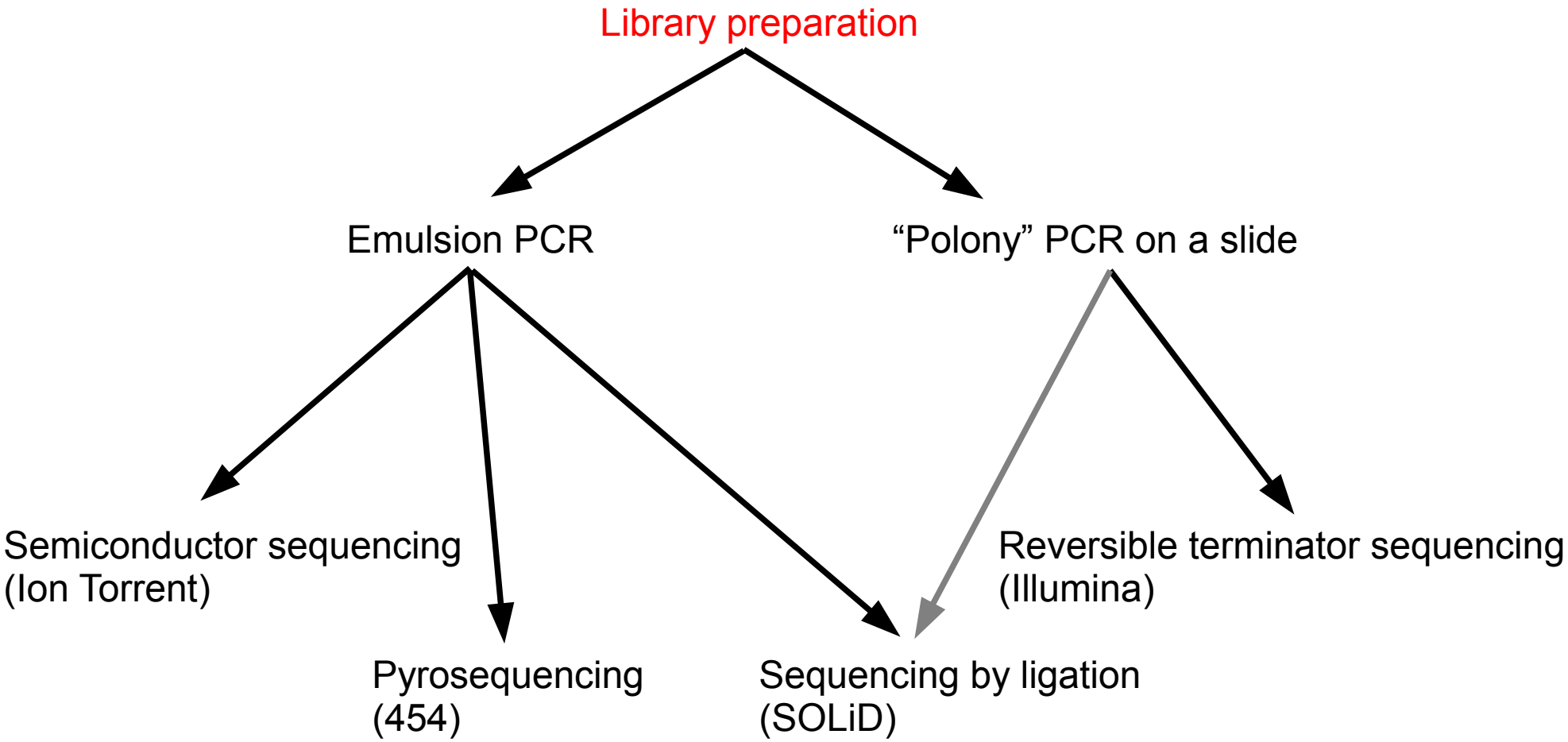
- 454 Sequencing / Roche
  - GS Junior System
  - GS FLX+ System
- Illumina (Solexa)
  - HiSeq System
  - Genome analyzer IIx
  - MySeq
- Applied Biosystems - Life Technologies
  - SOLiD 5500 System
  - SOLiD 5500xl System
- Ion Torrent - Life Technologies
  - Personal Genome Machine (PGM)
  - Proton
- Helicos
  - Helicos Genetic Analysis System
- Pacific Biosciences
  - PacBio RS
- Oxford Nanopore Technologies
  - GridION System
  - MinION

Next Generation Sequencing
Amplified Single Molecule Sequencing

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

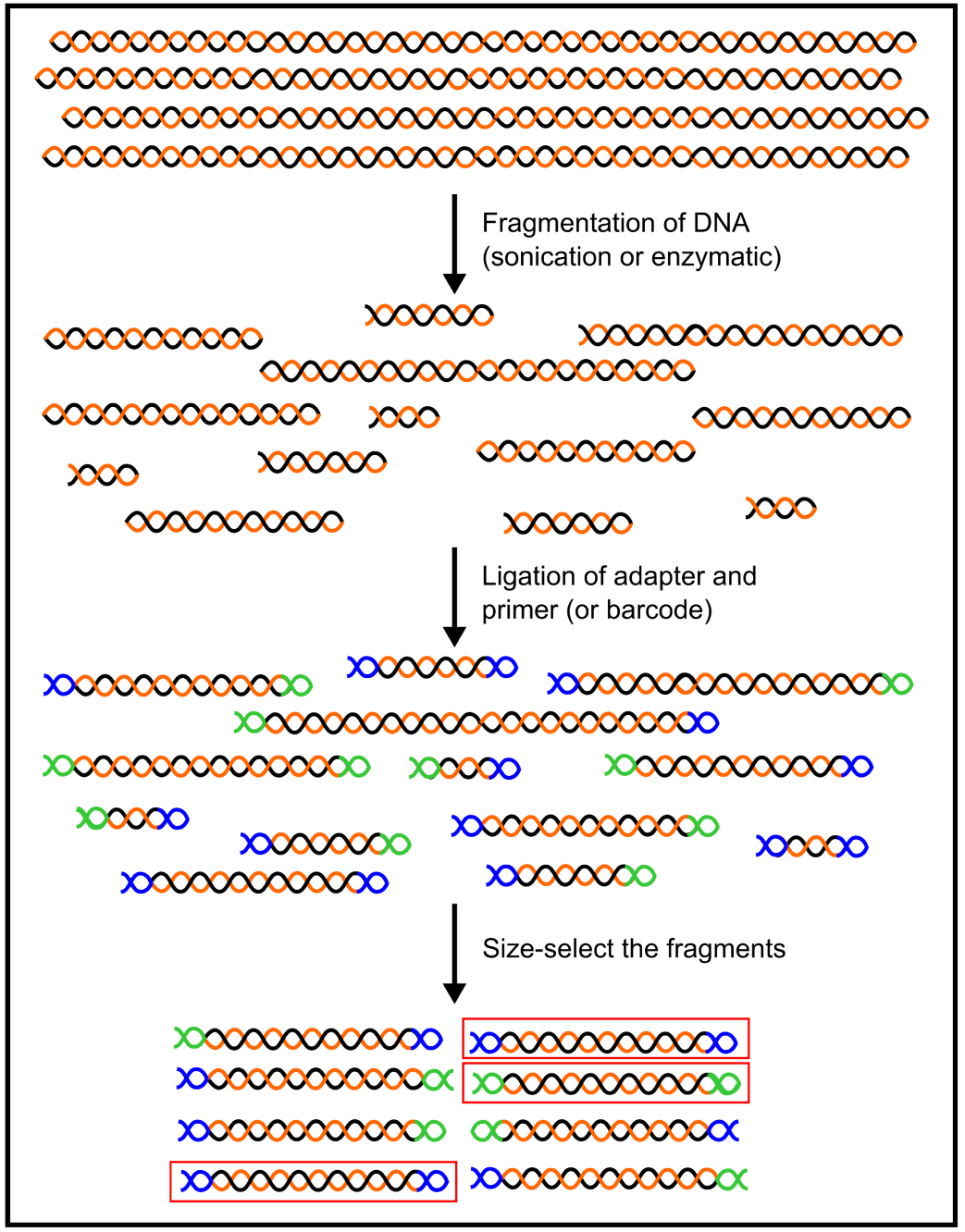
Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing



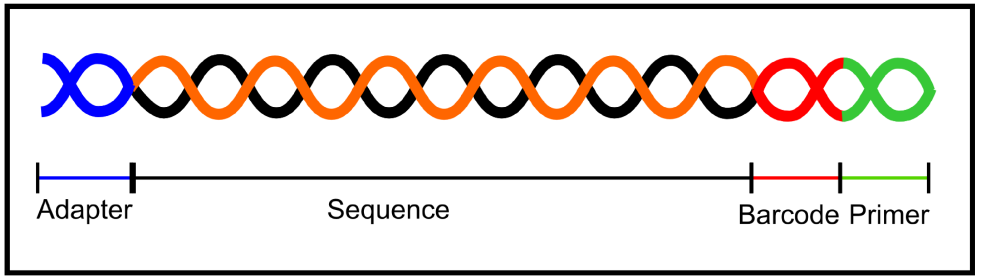
Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing



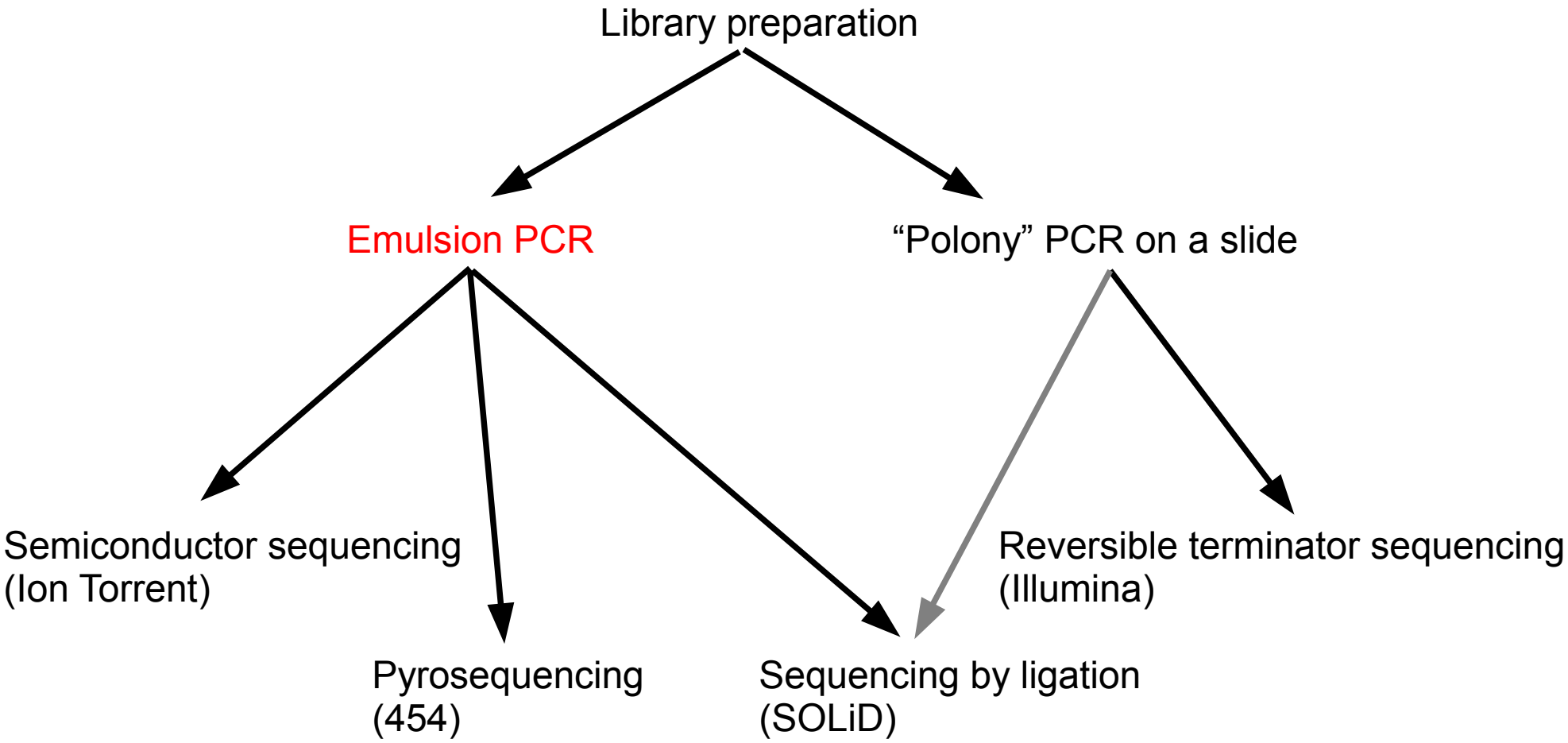
Library preparation

Good fragments :



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

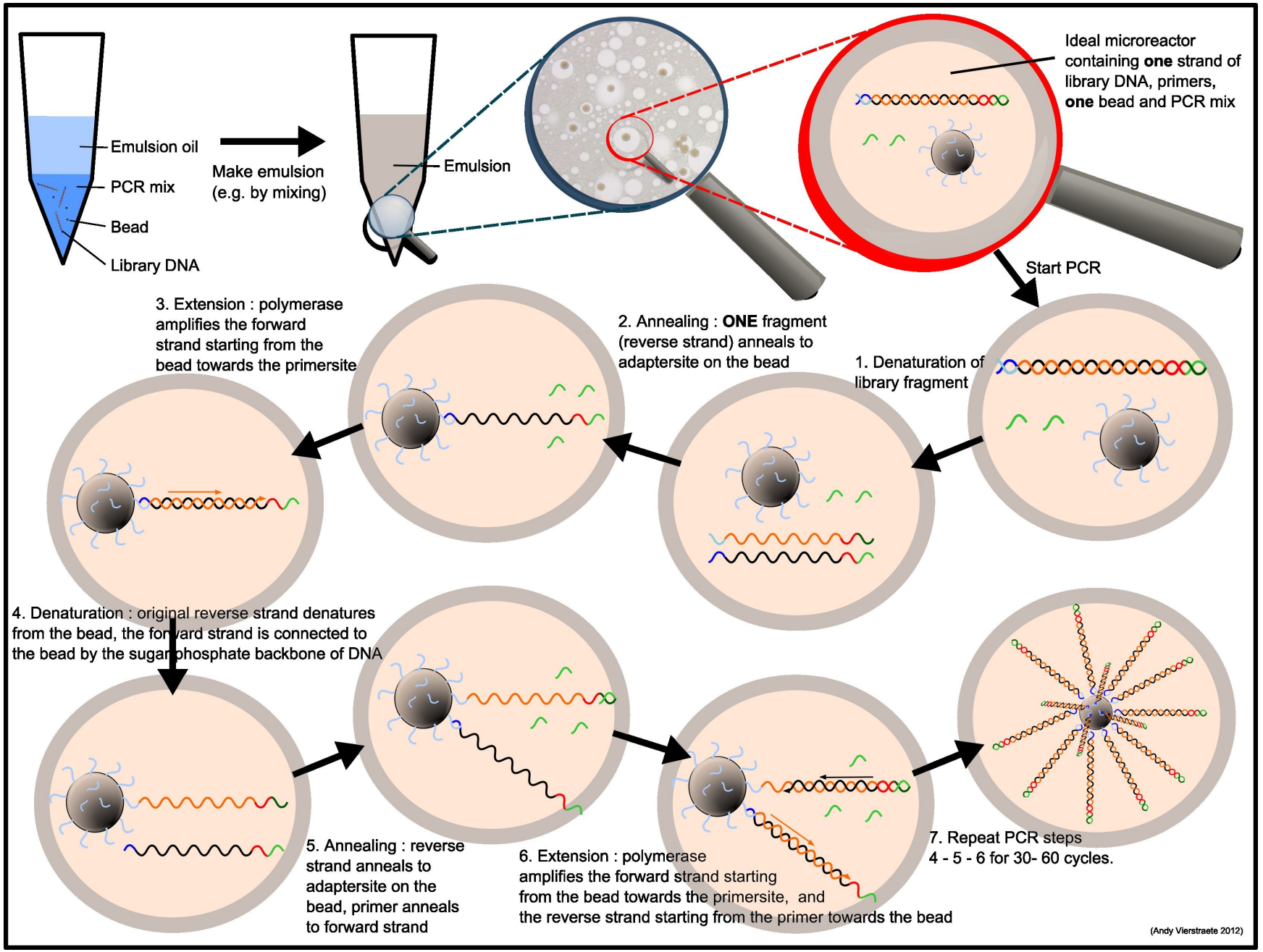
Next Generation Sequencing : Amplified Single Molecule Sequencing



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

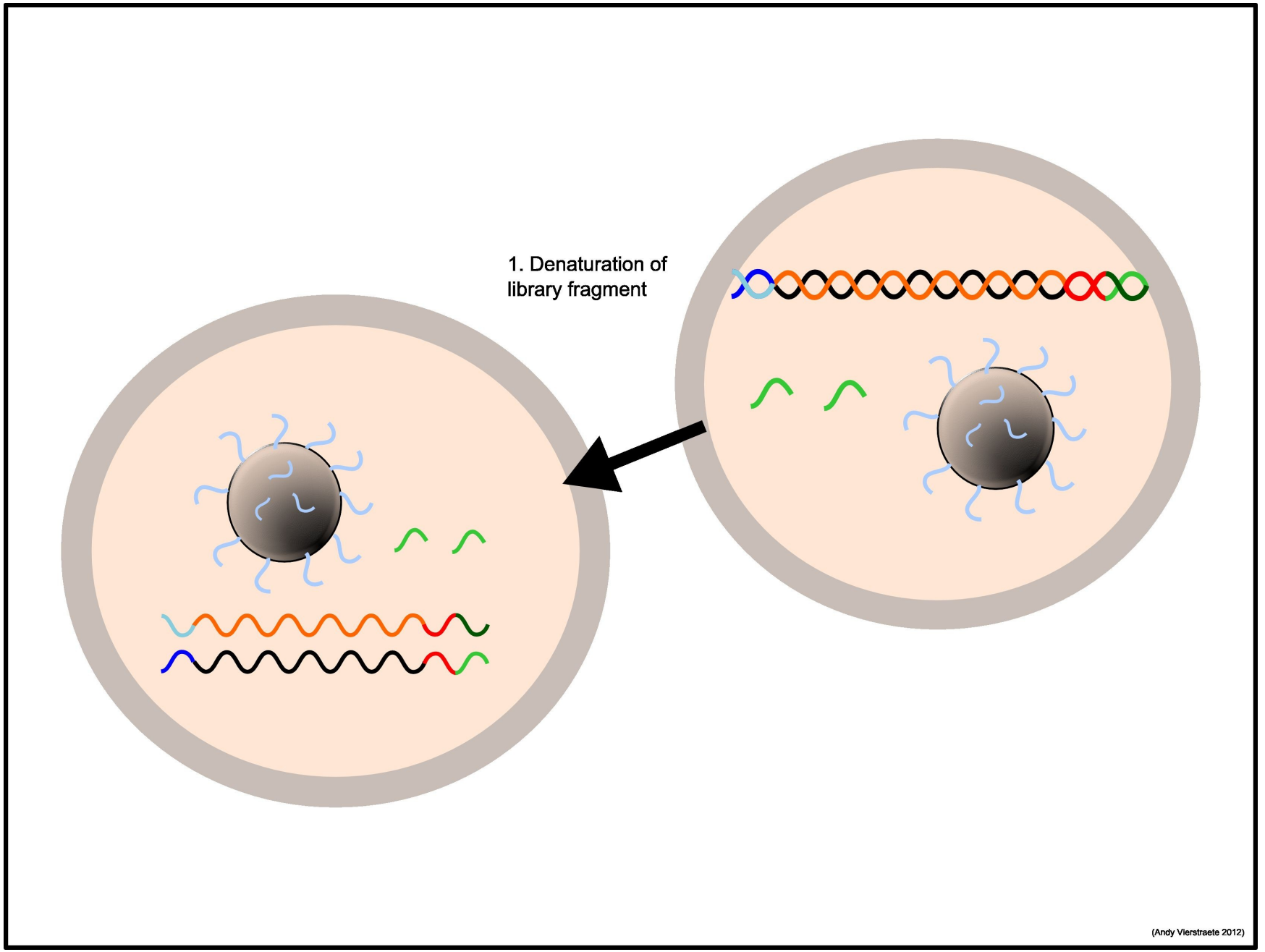
Next Generation Sequencing : Amplified Single Molecule Sequencing

Emulsion PCR



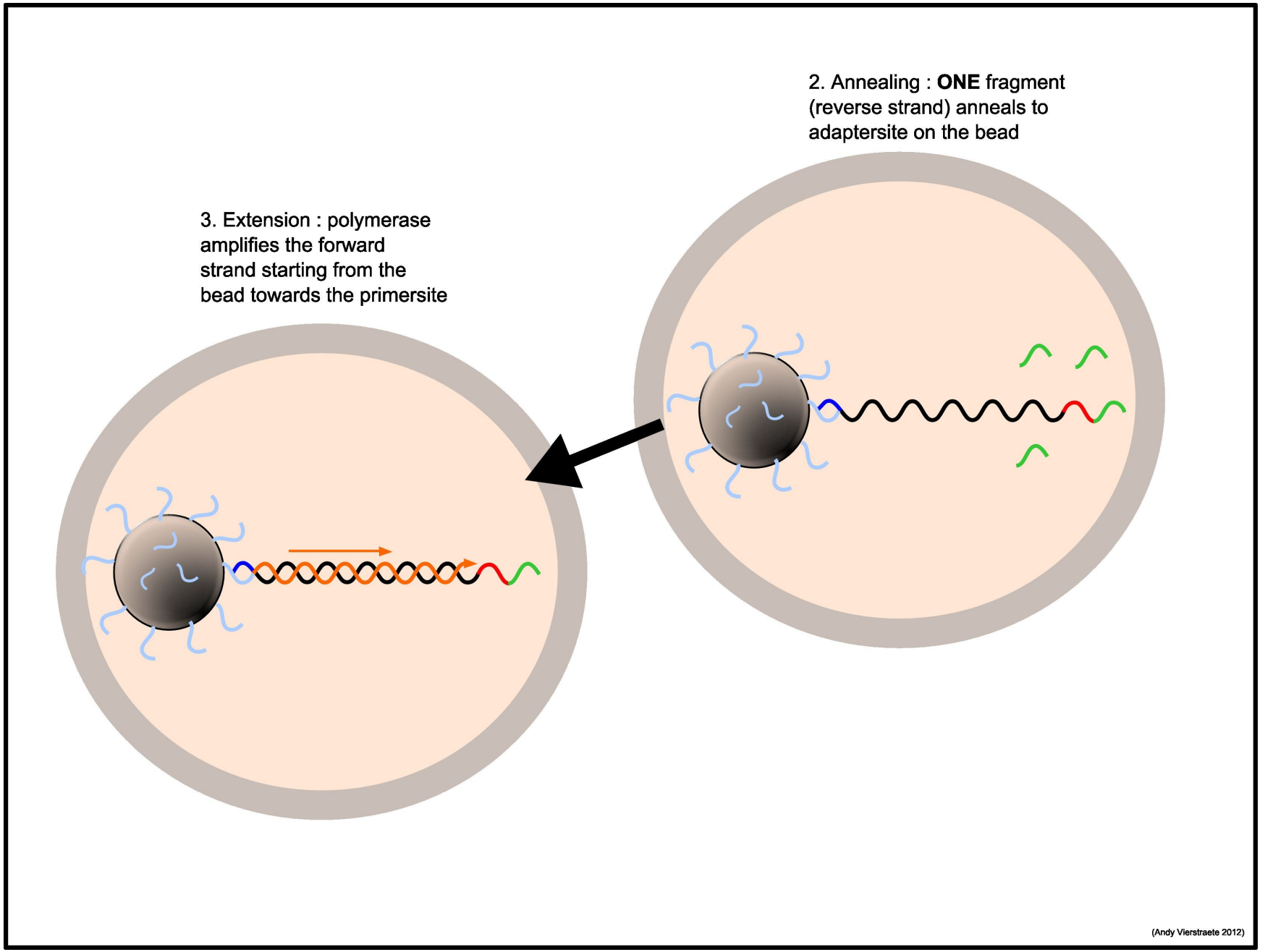
Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing Emulsion PCR



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

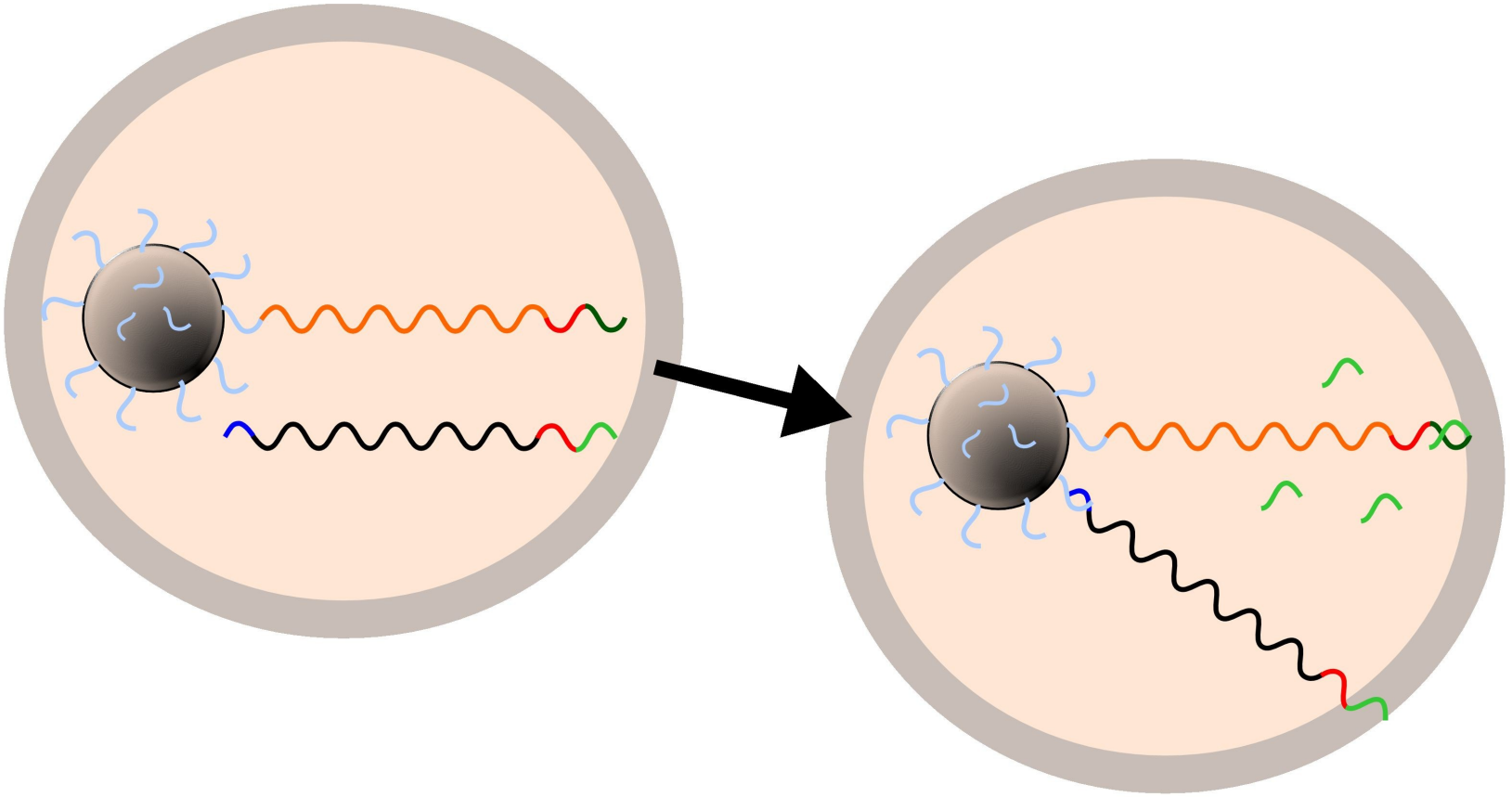
Next Generation Sequencing : Amplified Single Molecule Sequencing Emulsion PCR



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

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

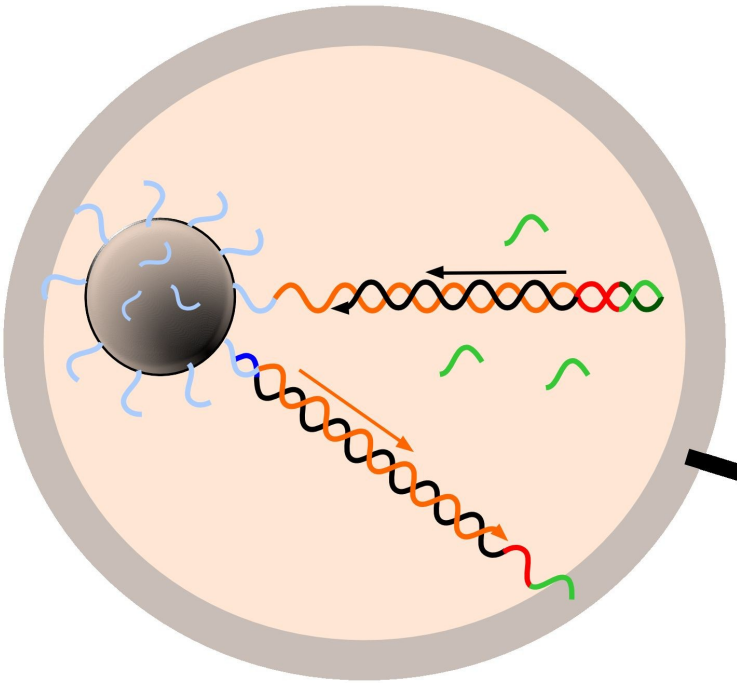


5. Annealing : reverse strand anneals to adaptersite on the bead, primer anneals to forward strand

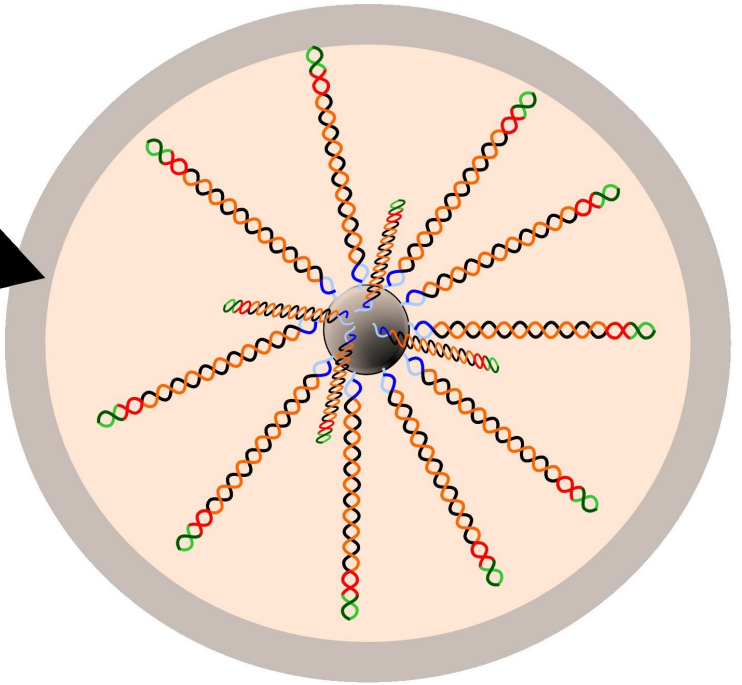
(Andy Vierstraete 2012)

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

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



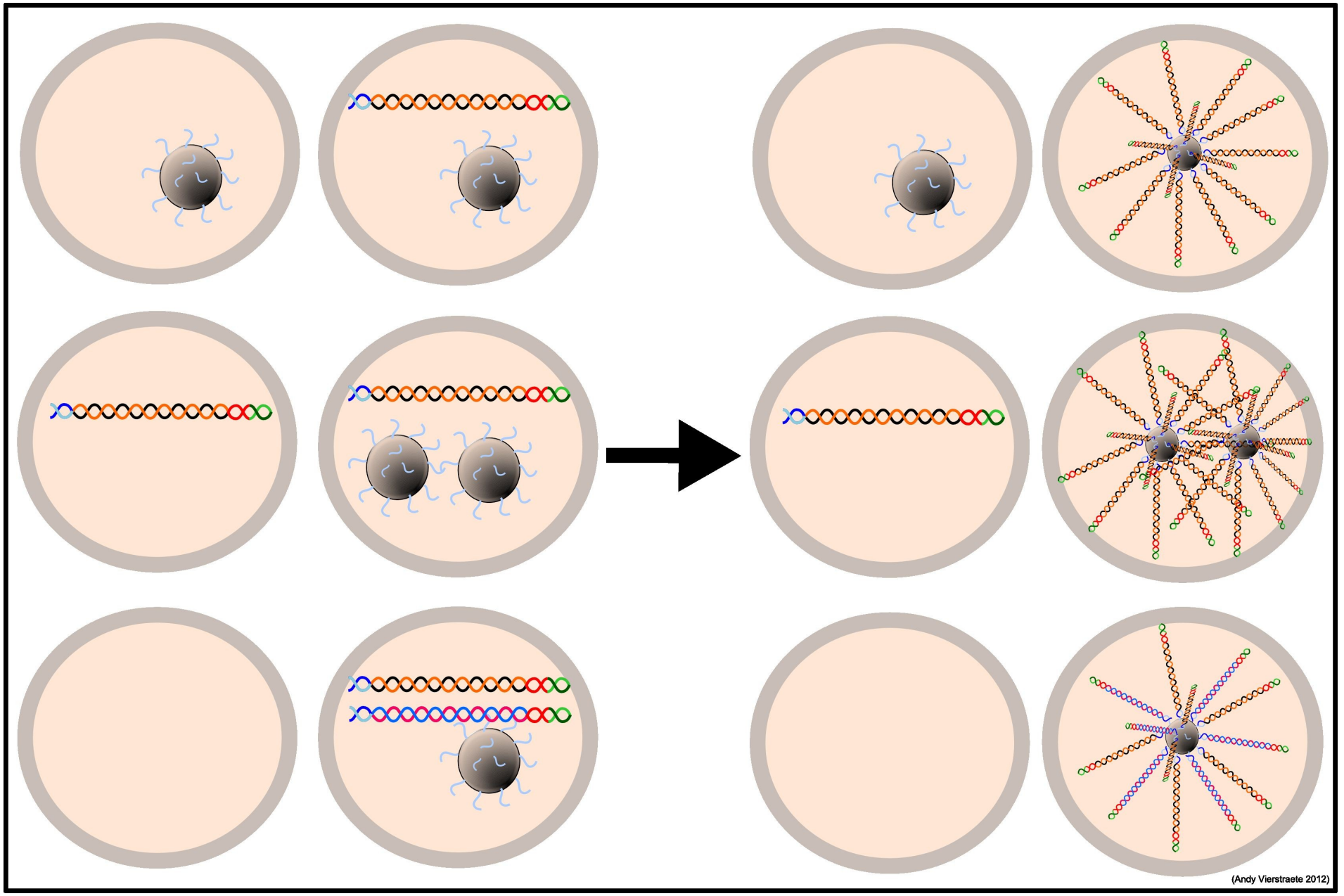
7. Repeat PCR steps
4 - 5 - 6 for 30- 60 cycles.

(Andy Vierstraete 2012)

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

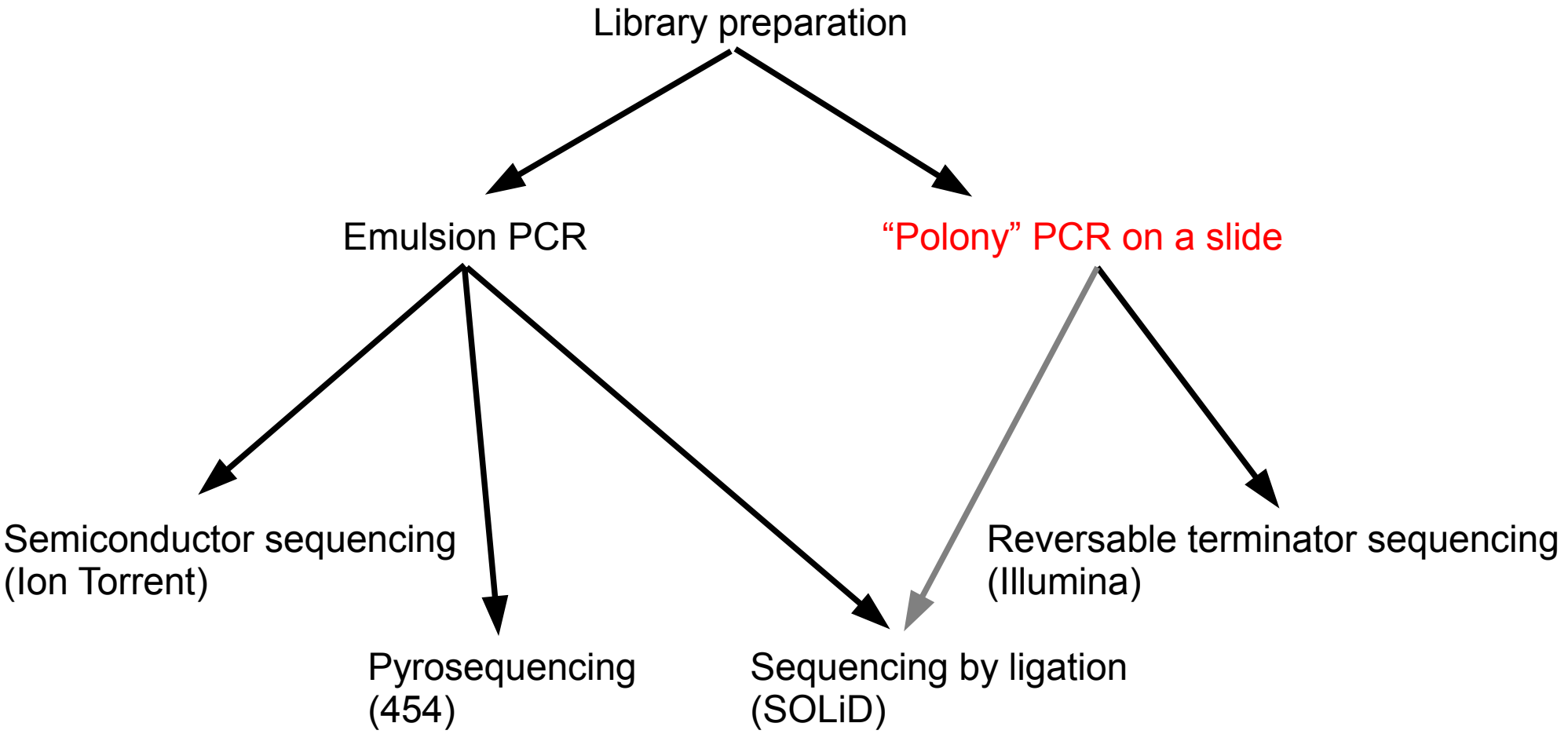
Next Generation Sequencing : Amplified Single Molecule Sequencing Emulsion PCR

different micro reactors : only 15 % are good ones



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

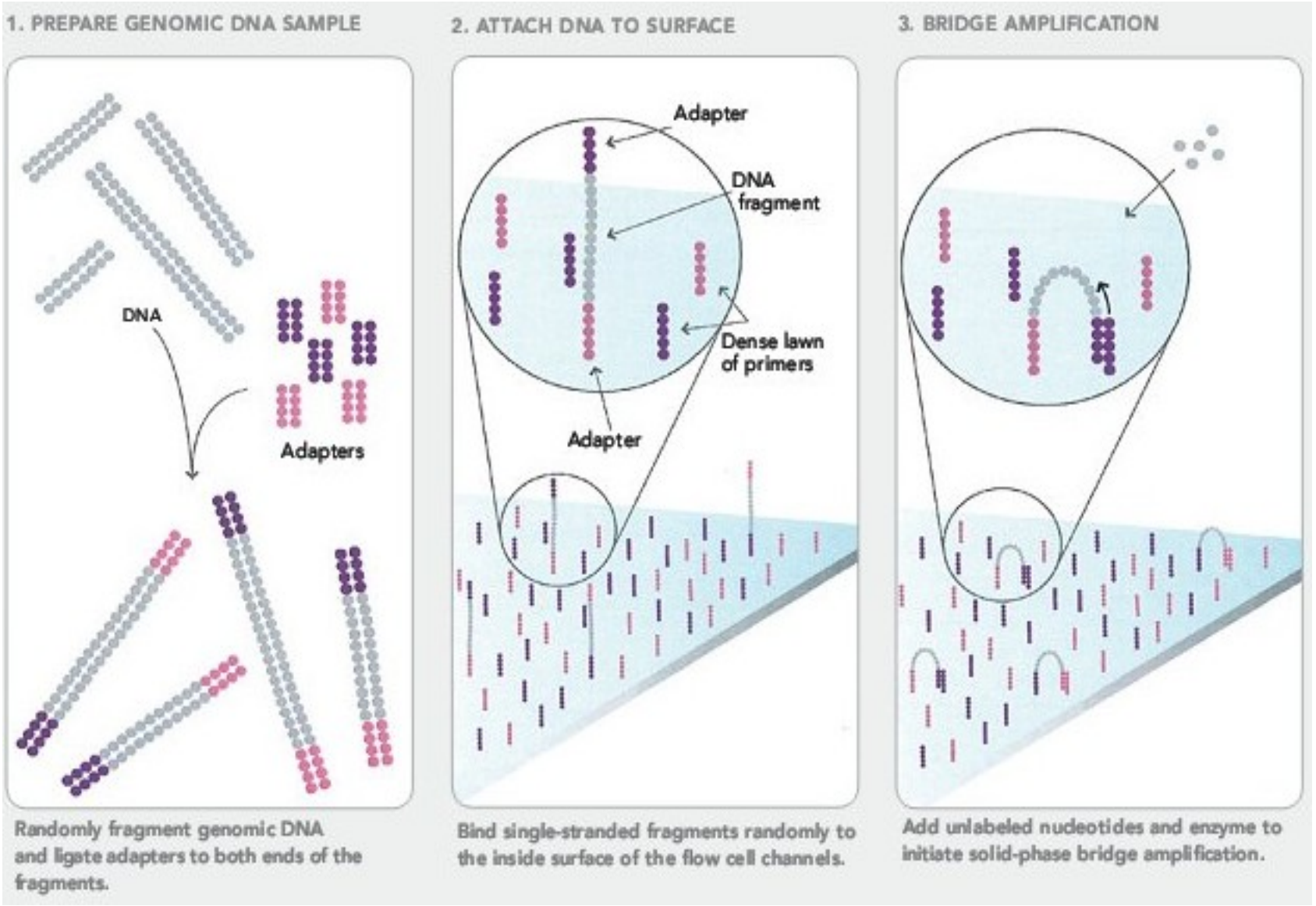
Next Generation Sequencing : Amplified Single Molecule Sequencing



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

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

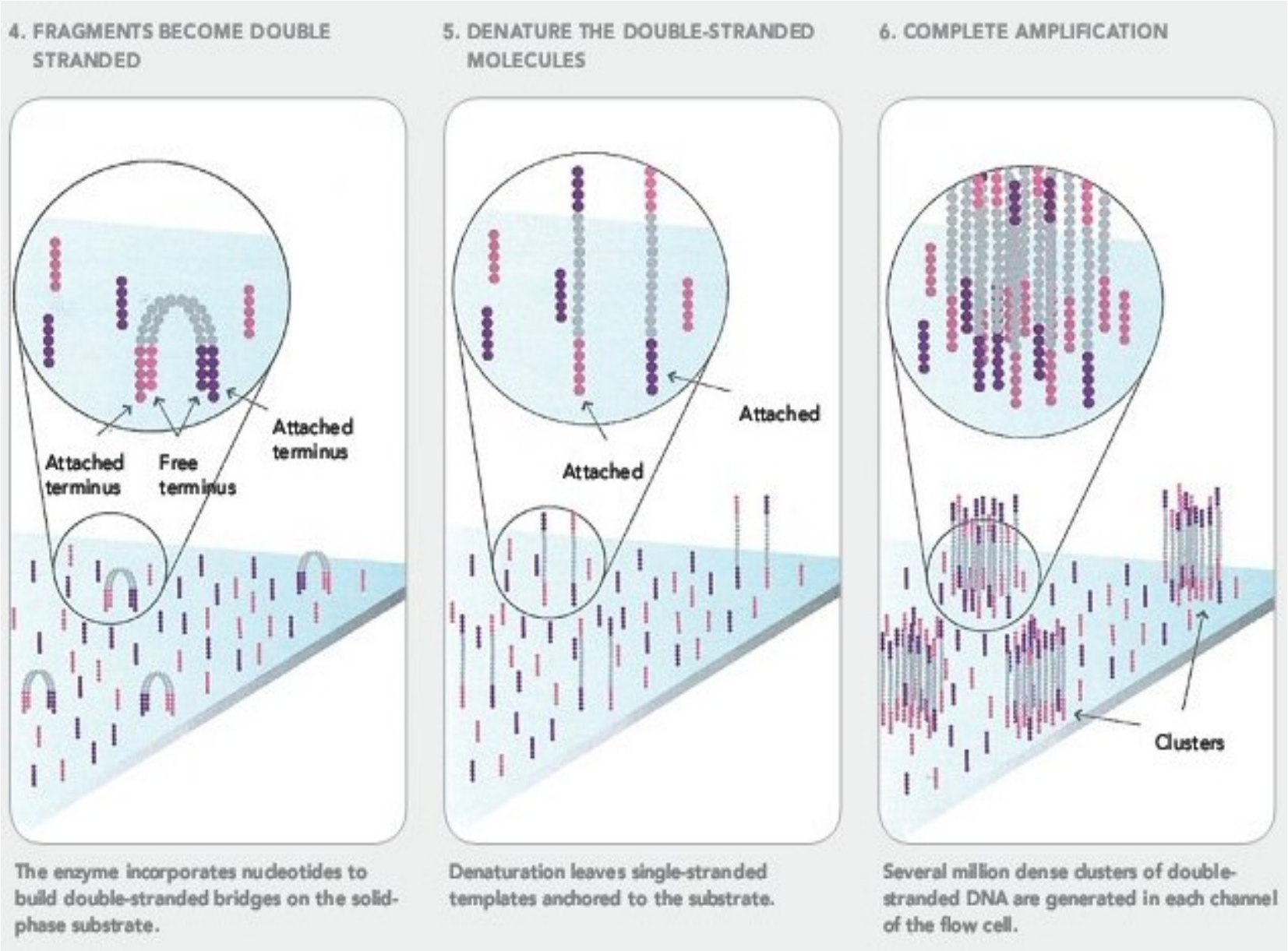
Bridge amplification : Illumina



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

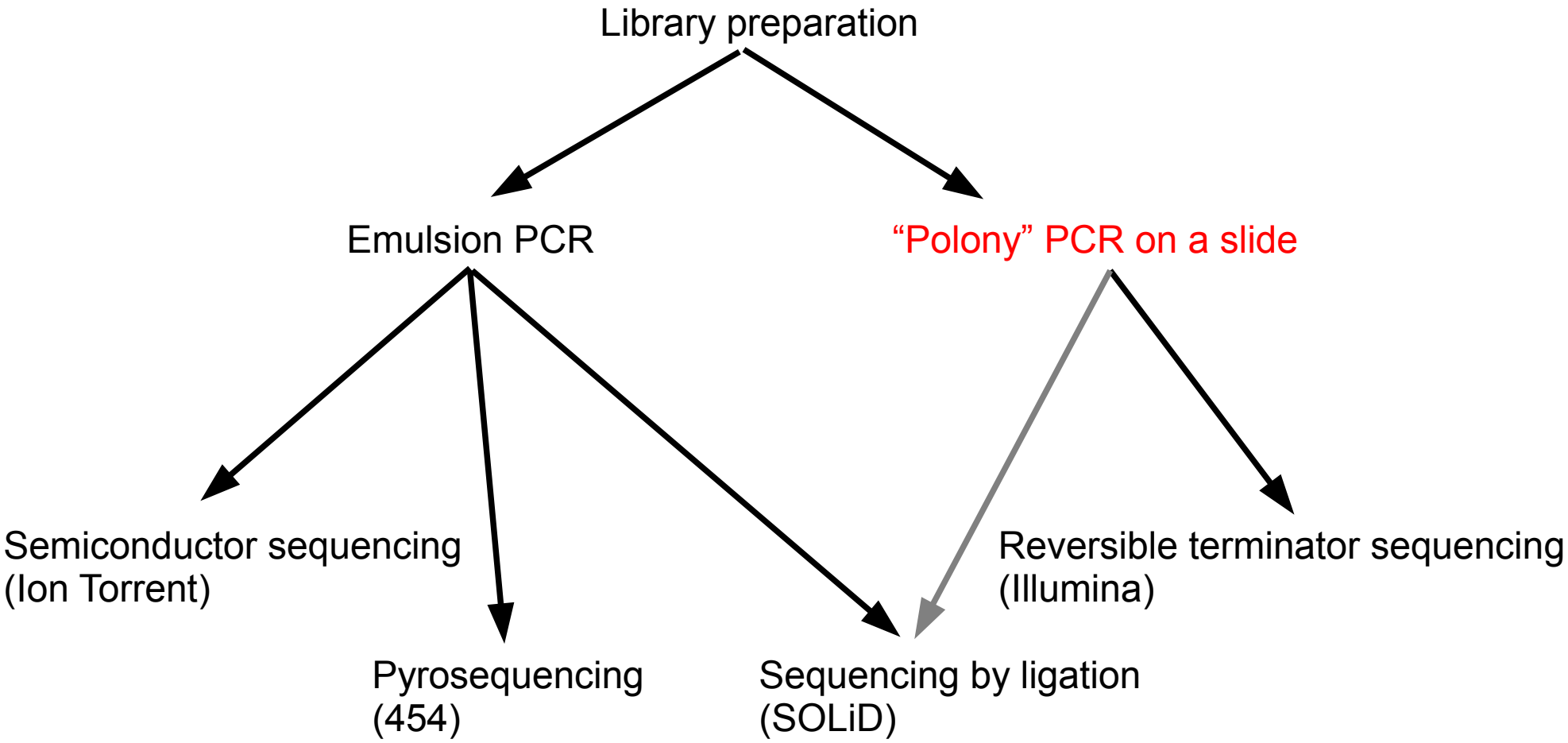
Next Generation Sequencing : Amplified Single Molecule Sequencing "Polony" PCR

Bridge amplification : Illumina



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

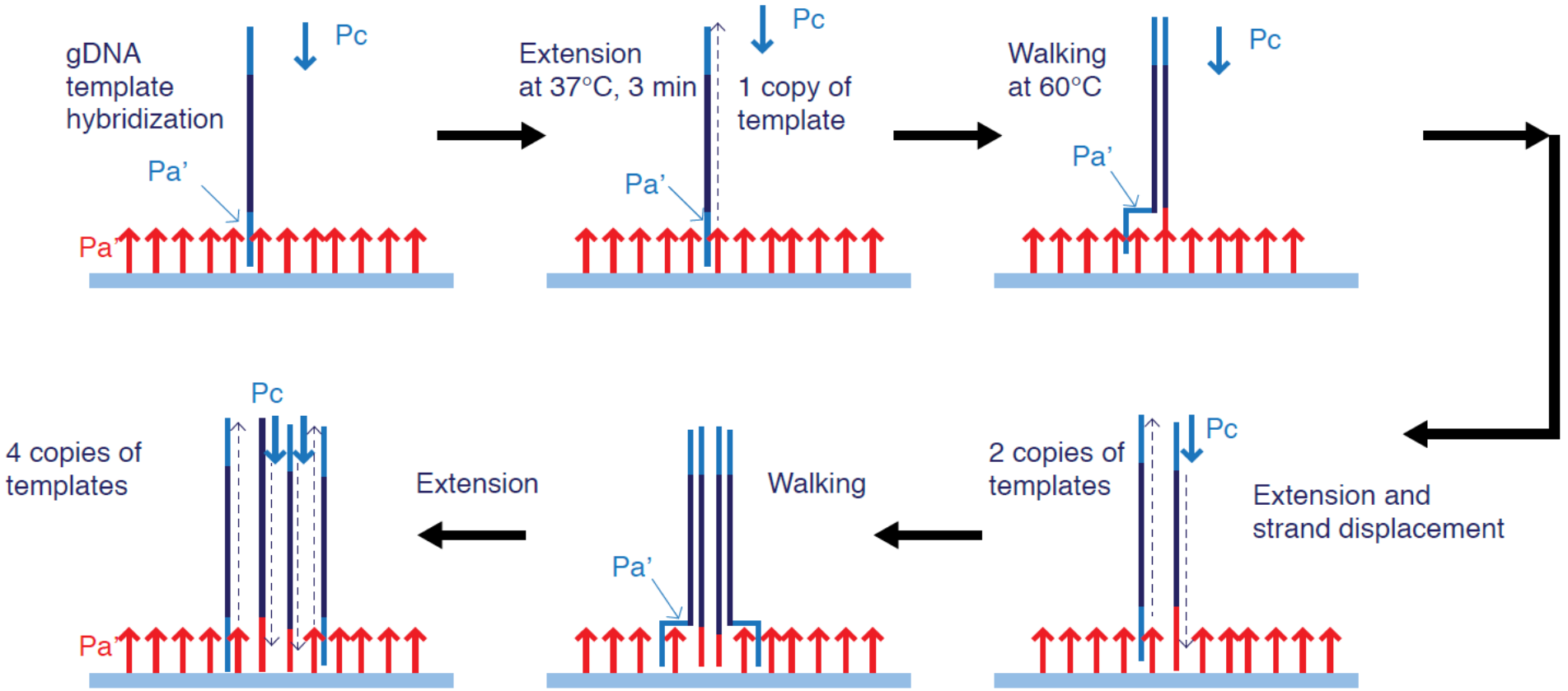
Next Generation Sequencing : Amplified Single Molecule Sequencing



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

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

Wildfire amplification : SOLiD

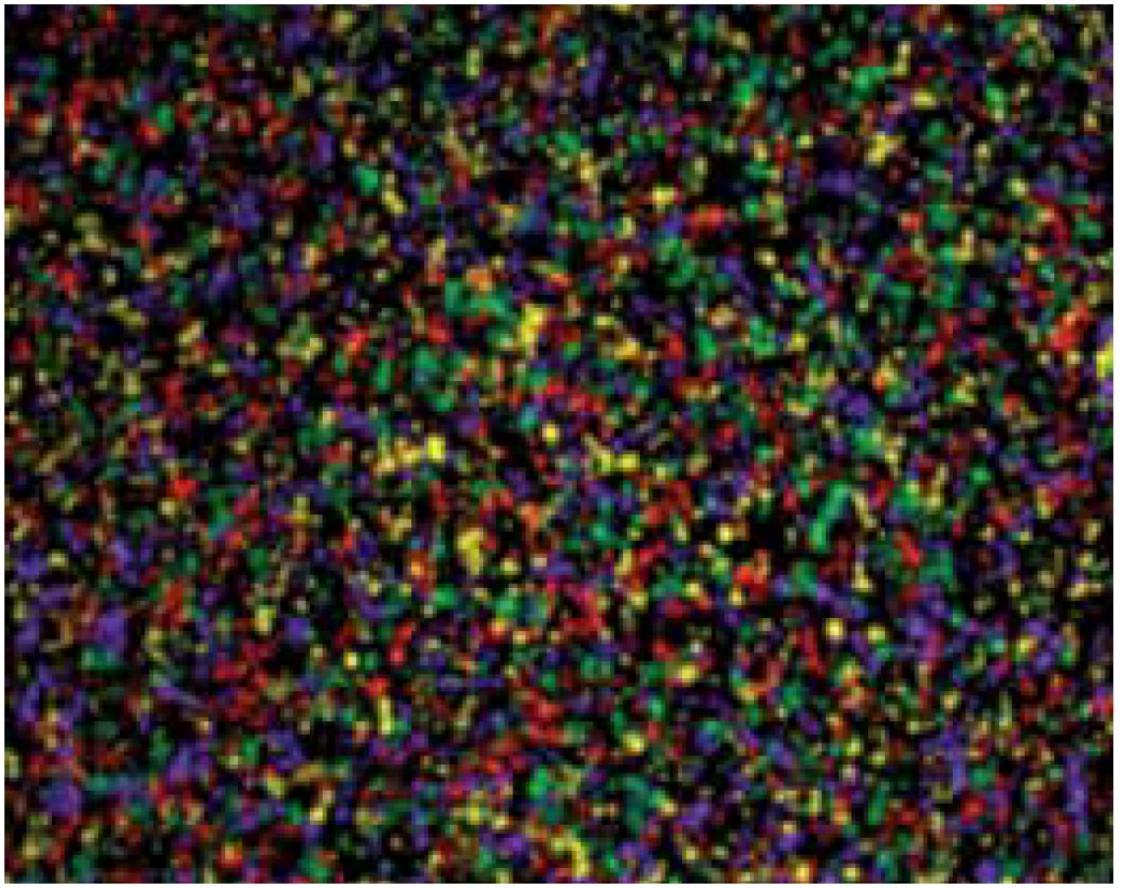


Wildfire chemistry schematic.

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

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

Wildfire amplification : SOLiD



One million colonies/mm² per FlowChip surface.

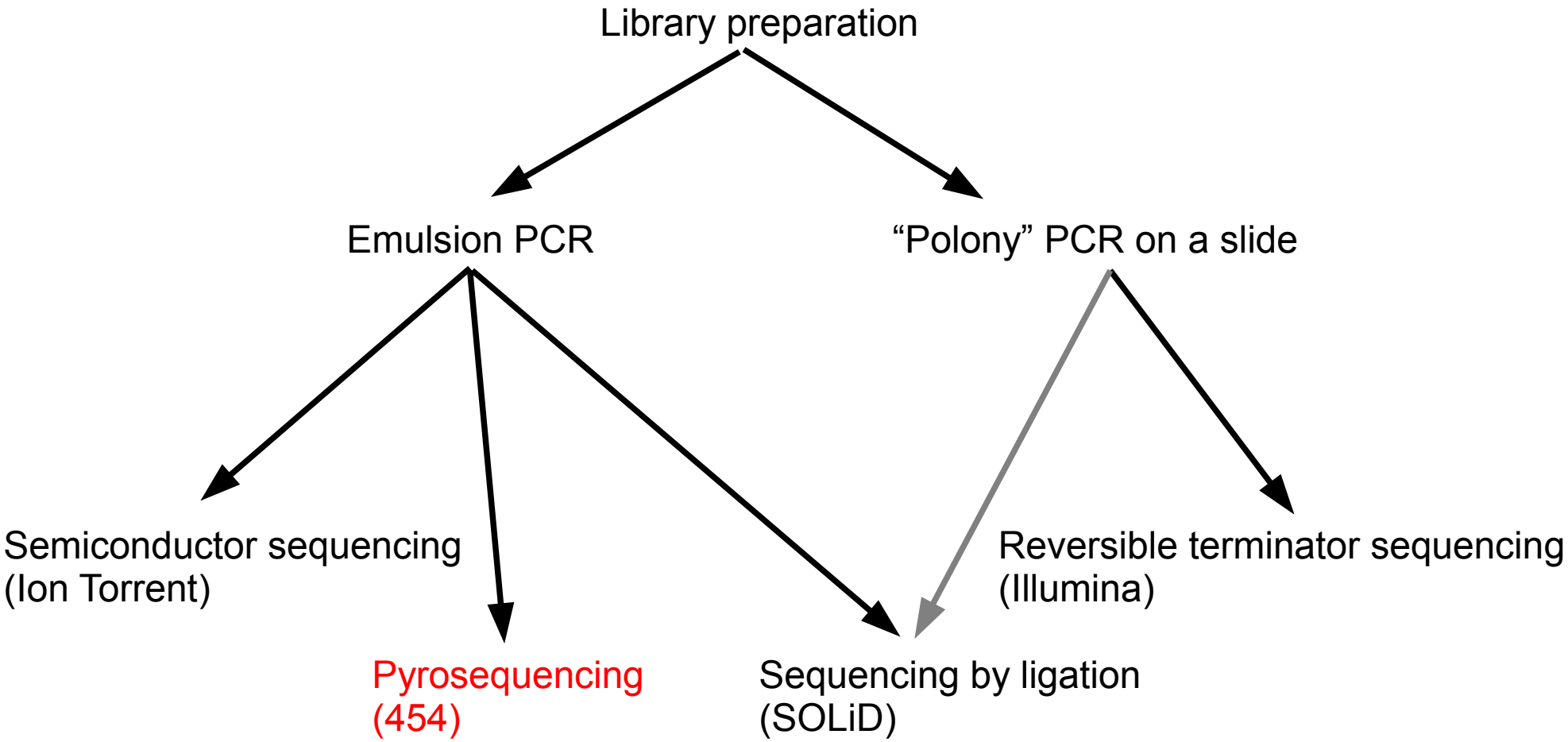
Quality scores in sequencing :
Q17, Q20, Q30, ...

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%

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

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

454 Sequencing / Roche

GS Junior System



GS FLX+ System



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

454 Sequencing / Roche

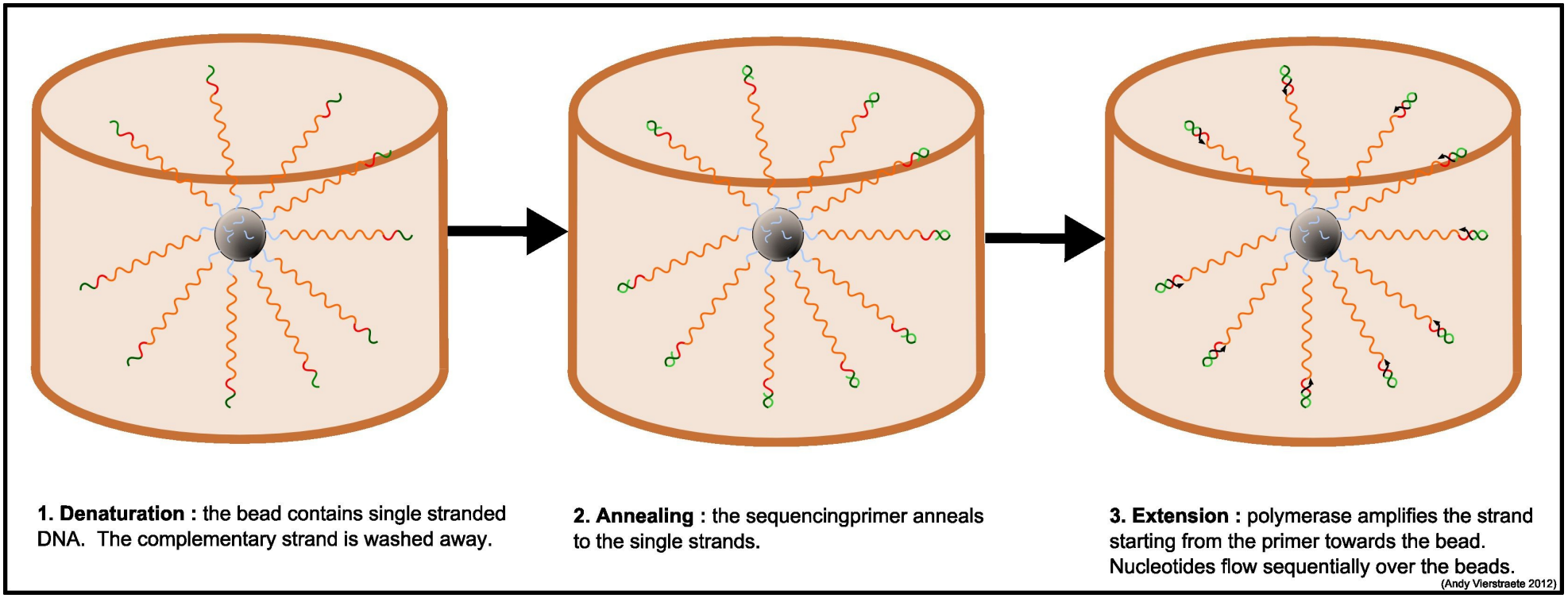
	GS Junior	GS FLX Titanium XL+	GS FLX Titanium XLR70
Read Length	400 bp	700 bp	450 bp
Throughput	35 Mb	700 Mb	450 Mb
Reads per run	100,000	1,000,000	1,000,000
Accuracy	99 %	99,997 %	99,995 %
Run Time	10 hours	23 hours	10 hours

Workflow : Library preparation → Emulsion PCR → Pyrosequencing

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

454 Sequencing / Roche Pyrosequencing



1. Denaturation : the bead contains single stranded DNA. The complementary strand is washed away.

2. Annealing : the sequencingprimer anneals to the single strands.

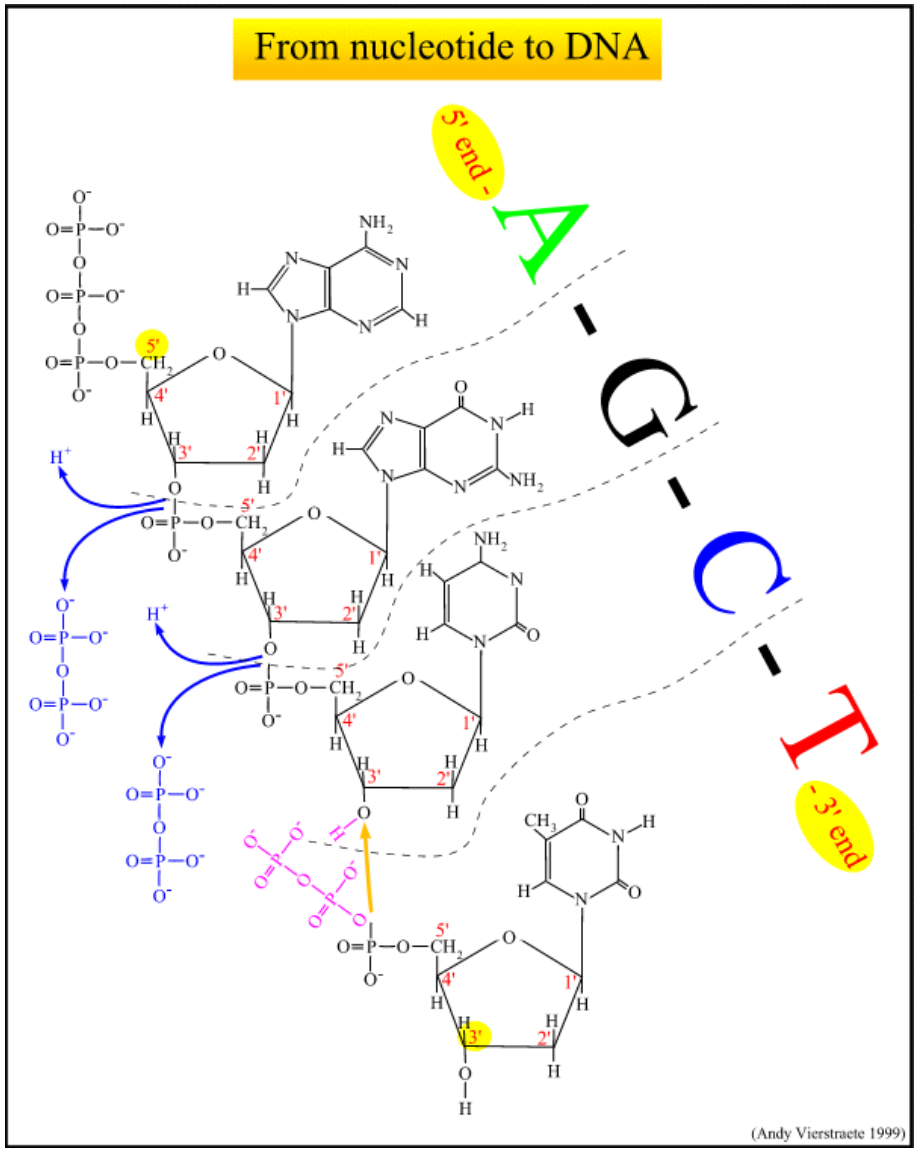
3. Extension : polymerase amplifies the strand starting from the primer towards the bead. Nucleotides flow sequentially over the beads.

(Andy Vierstraete 2012)

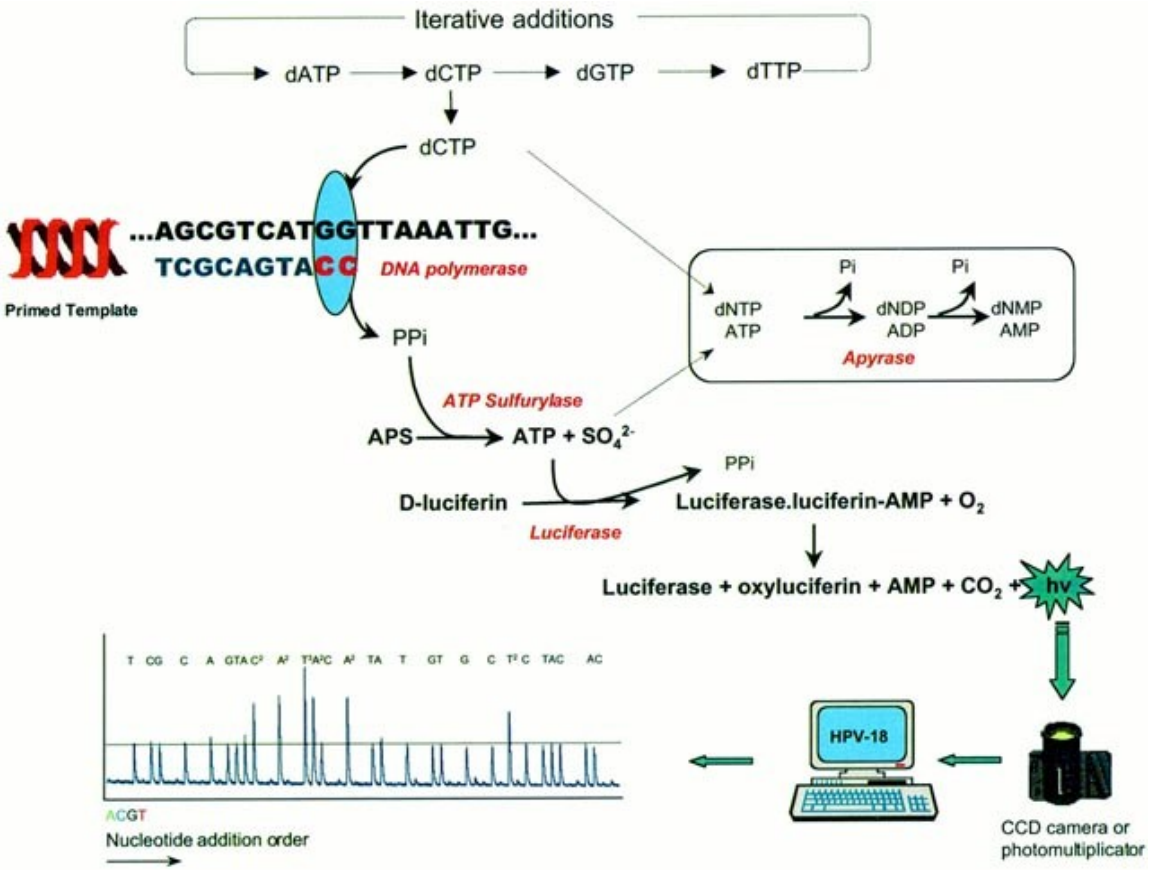
Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

454 Sequencing / Roche Pyrosequencing



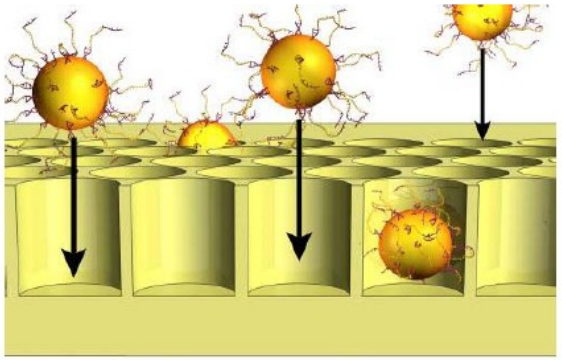
4 nucleotides flow sequentially



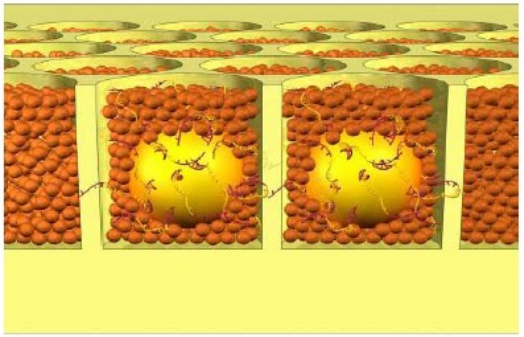
Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

454 Sequencing / Roche Pyrosequencing

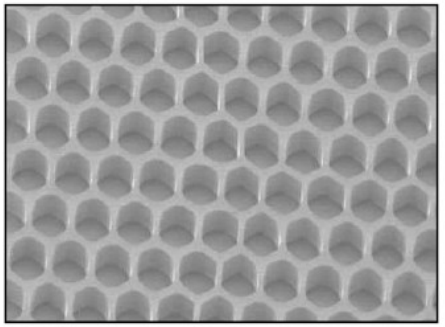


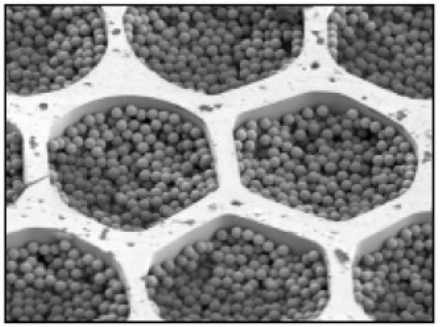
Load enzyme + packing beads

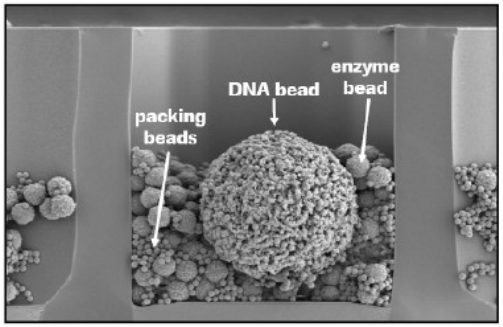


Single clonally amplified DNA beads are deposited in a PicoTiter™Plate (max 4 *10^6 beads)





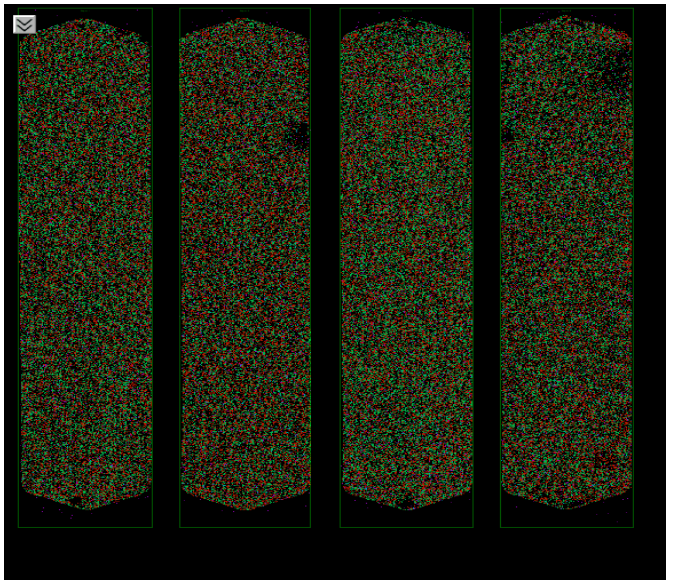


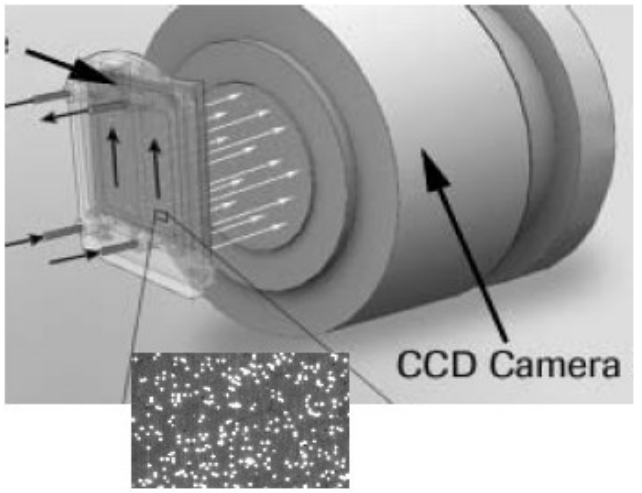


Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

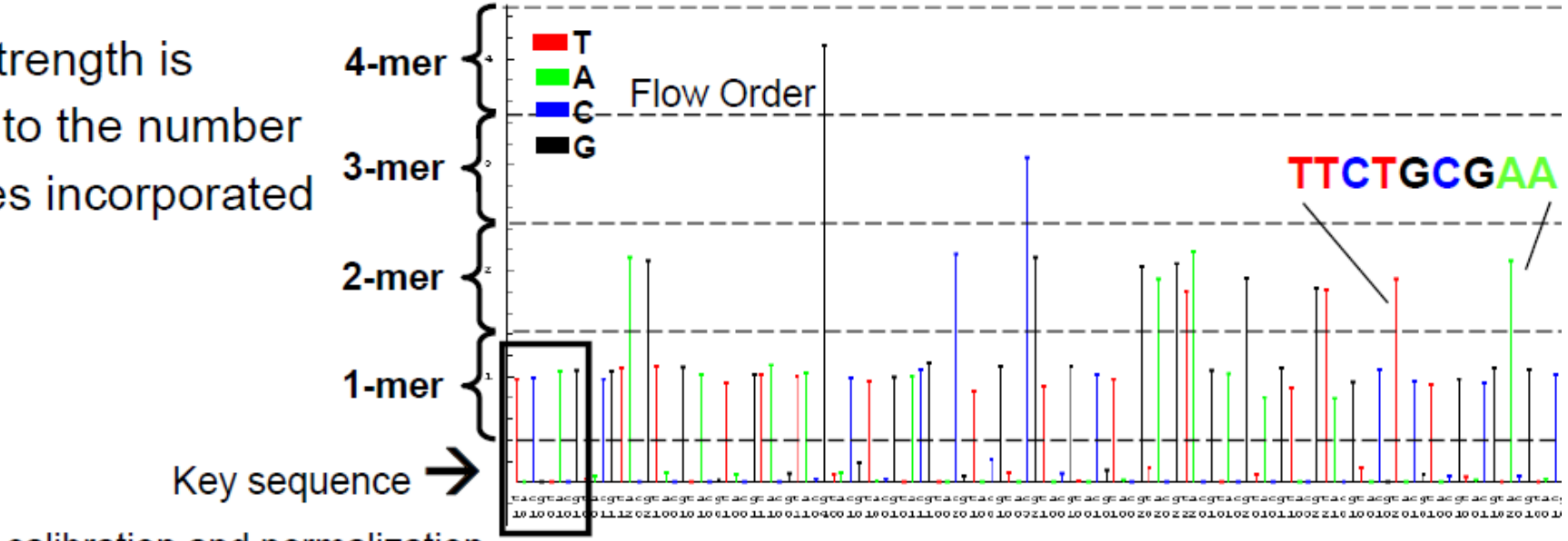
Next Generation Sequencing : Amplified Single Molecule Sequencing

454 Sequencing / Roche Pyrosequencing





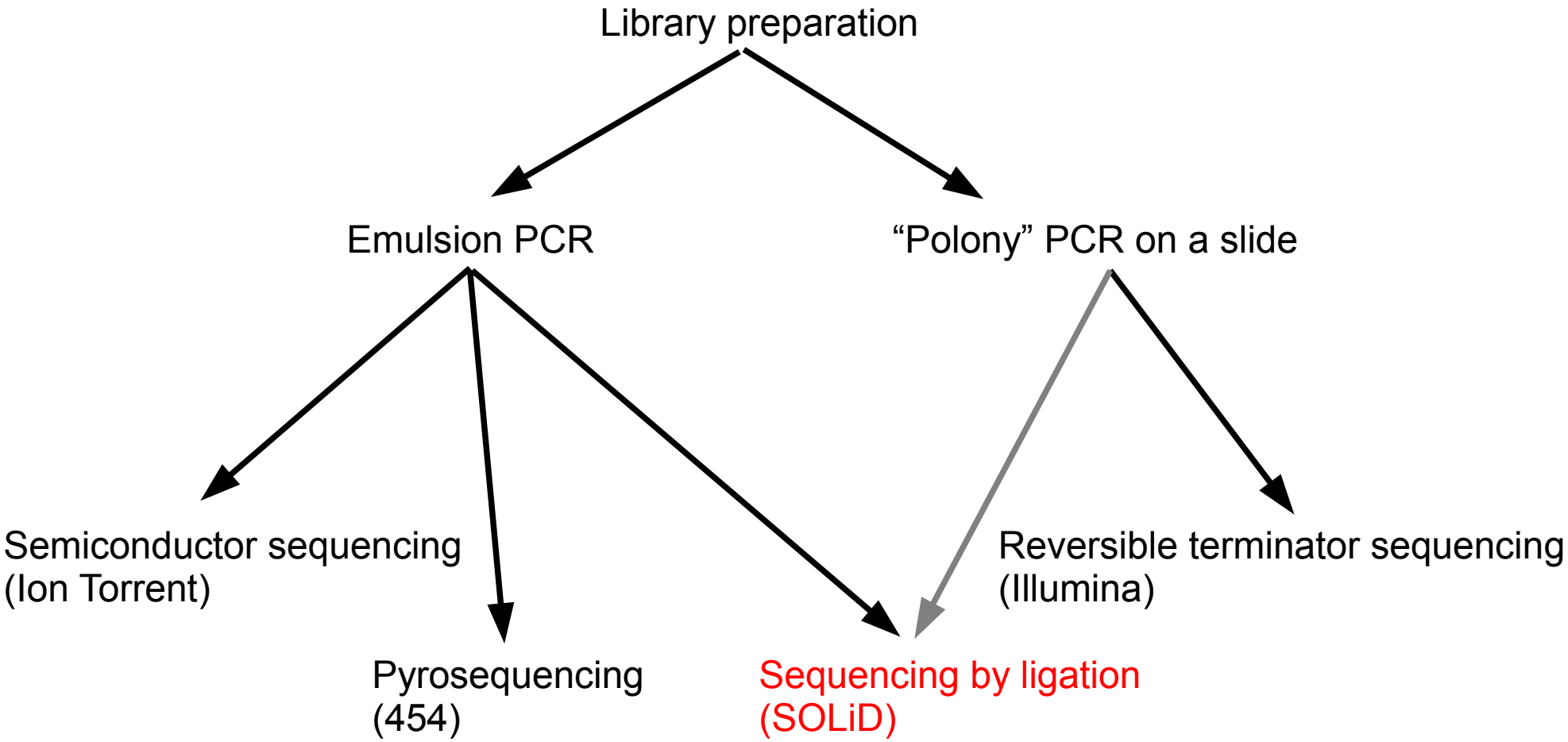
The signal strength is proportional to the number of nucleotides incorporated



TCAG for signal calibration and normalization

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

SOLiD

5500xl SOLiD Sequencer

5500 SOLiD Sequencer



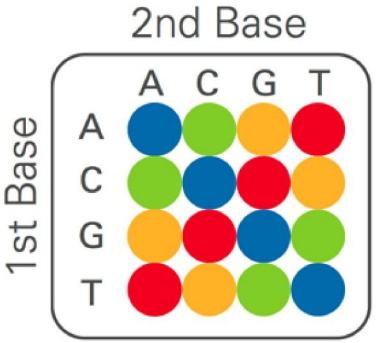
Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

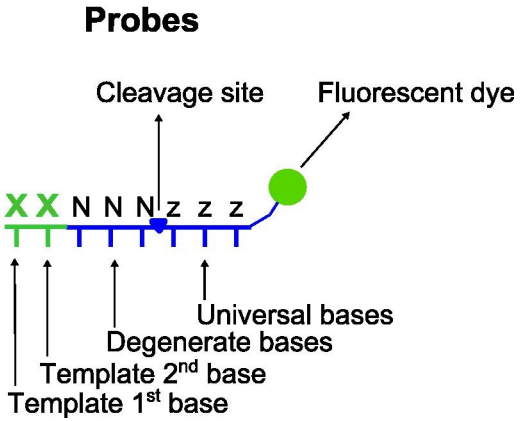
Next Generation Sequencing : Amplified Single Molecule Sequencing

SOLiD

	5500xl	5500	5500 W
Read Length	75 bp	75 bp	75 bp
Throughput	180 Gb	90 Gb	360 Gb ?
Reads per run	3,000,000	1,500,000	6,000,000 ?
Accuracy	99,99 %	99,99 %	99,99 %
Run Time	7-12 days	7-12 days	7-12 days

Workflow : Library preparation → Emulsion PCR → Sequencing by Ligation
Wildfire PCR



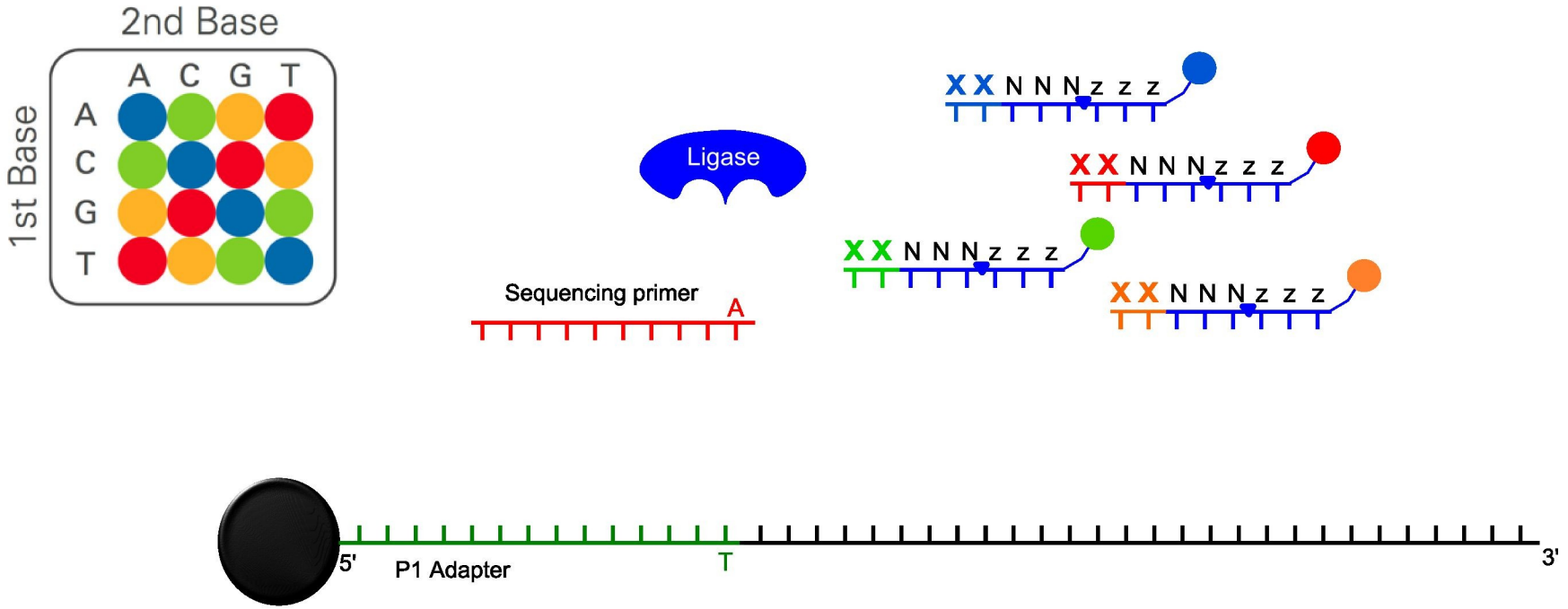


dual base encoding

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

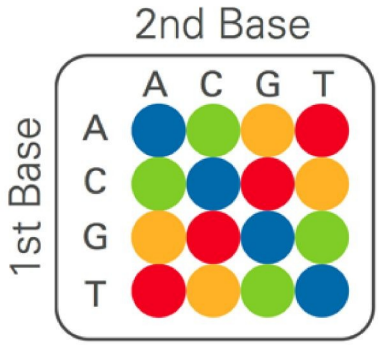
SOLiD

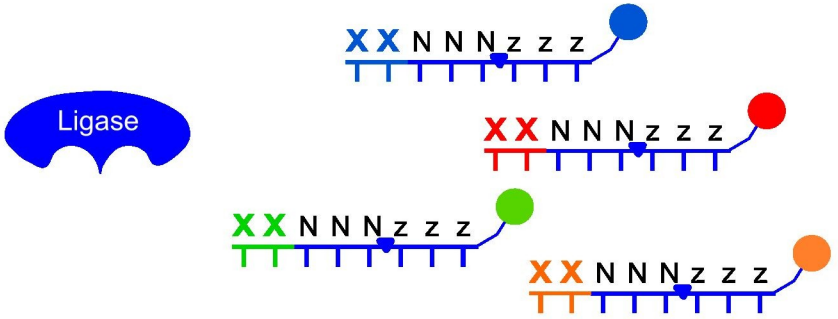


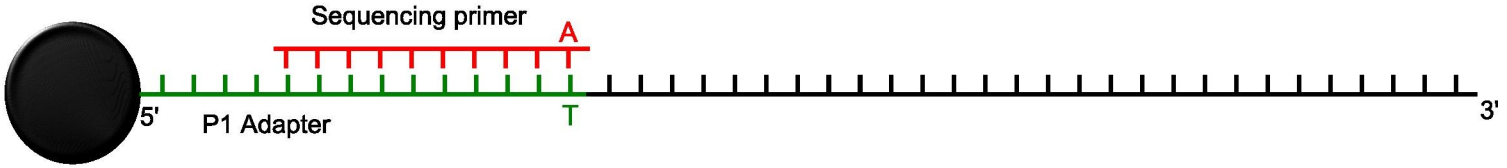
Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

SOLiD



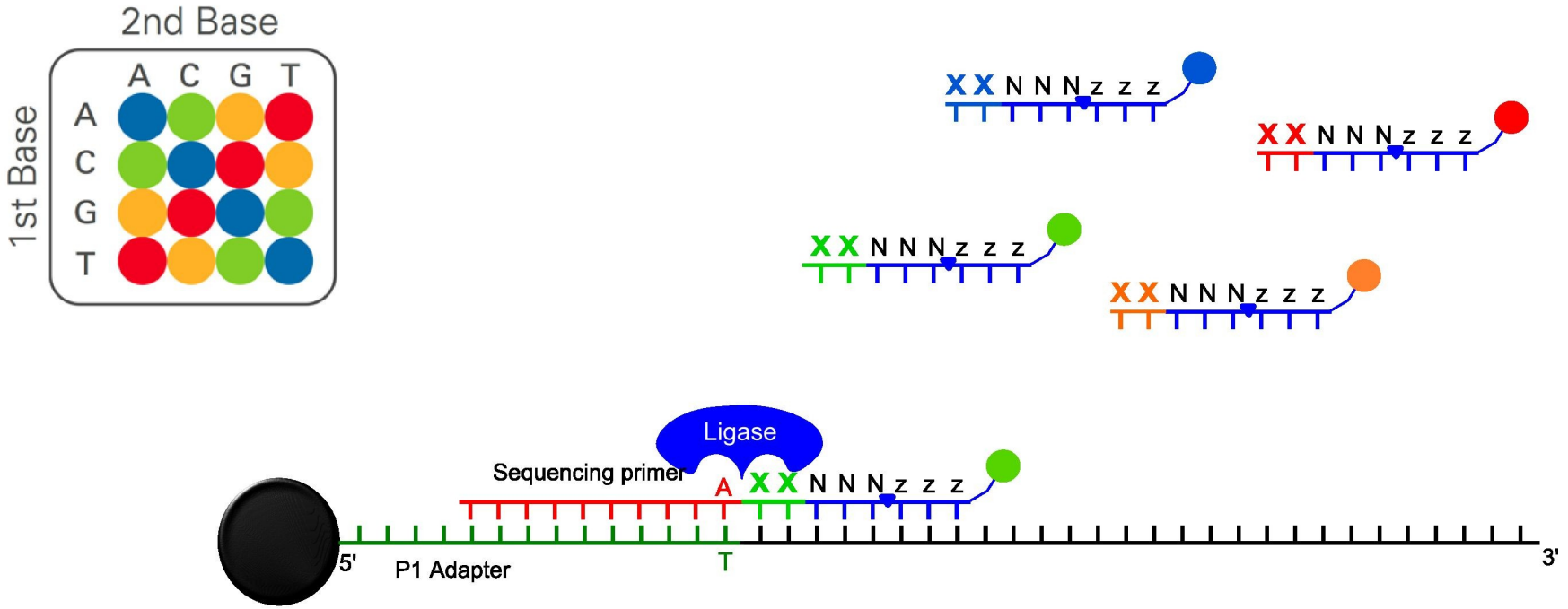




Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

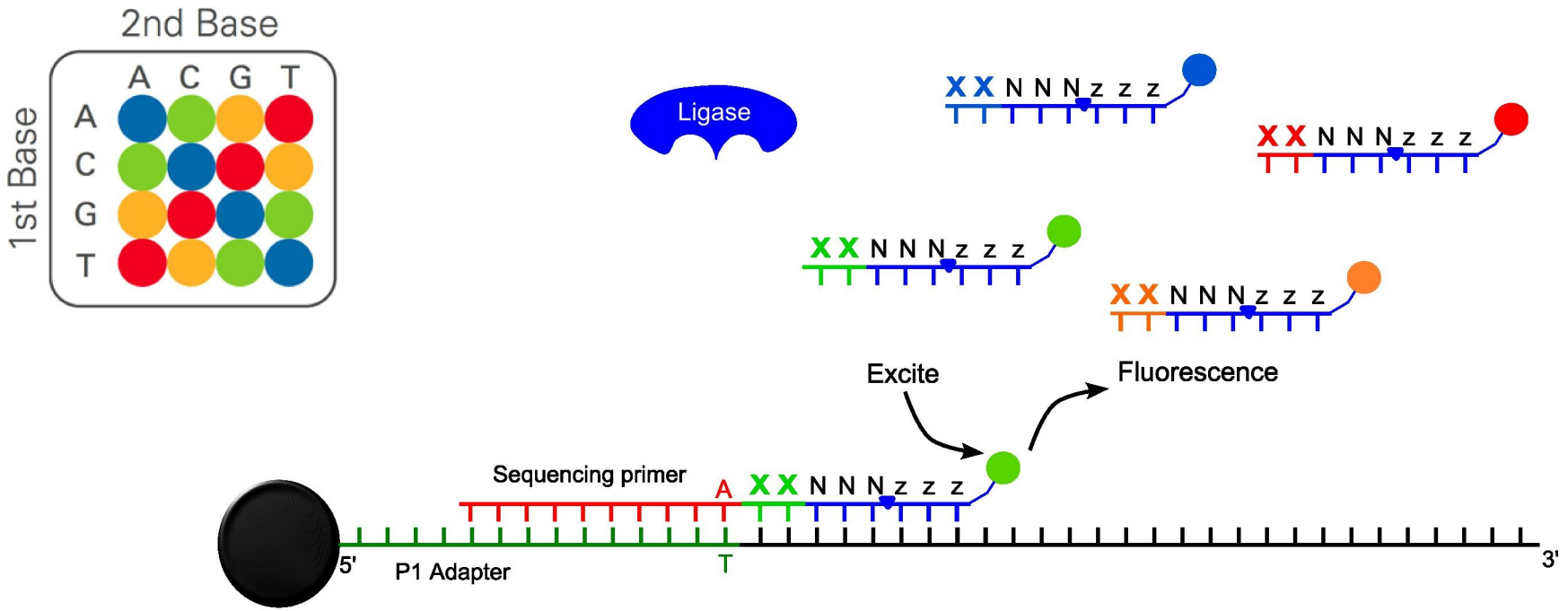
SOLiD



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

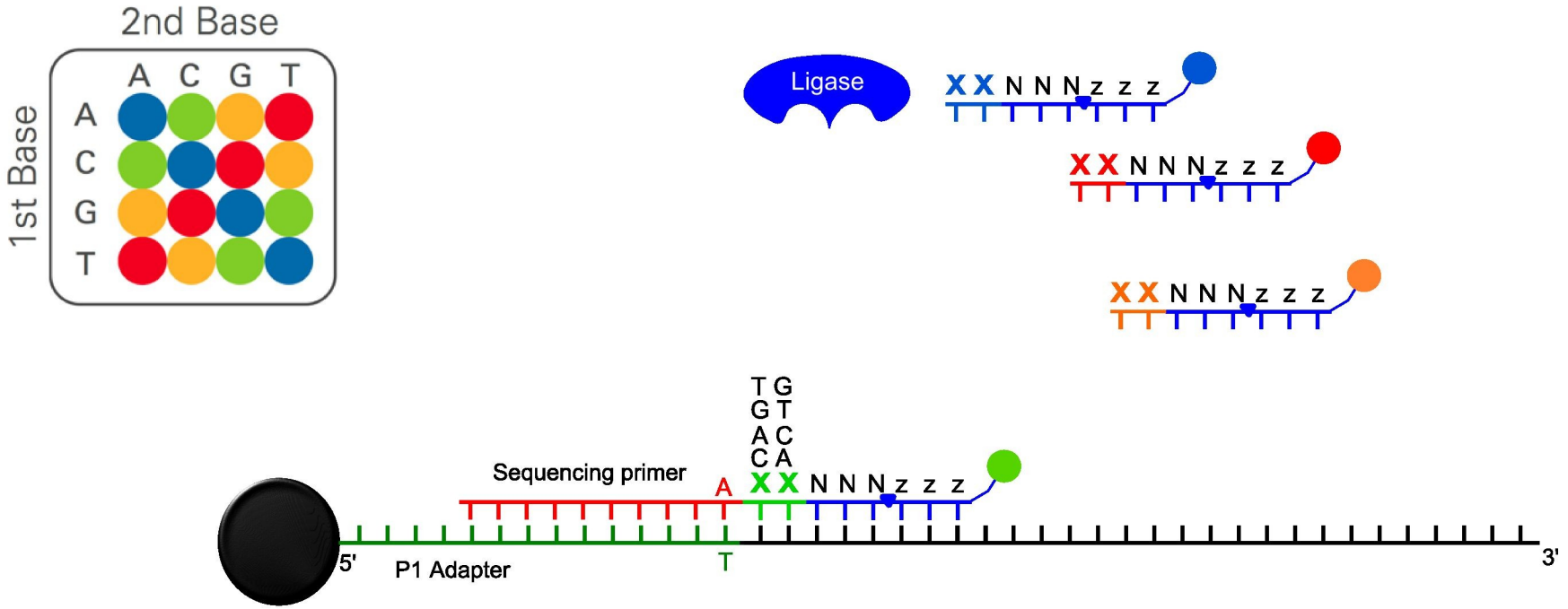
SOLiD



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

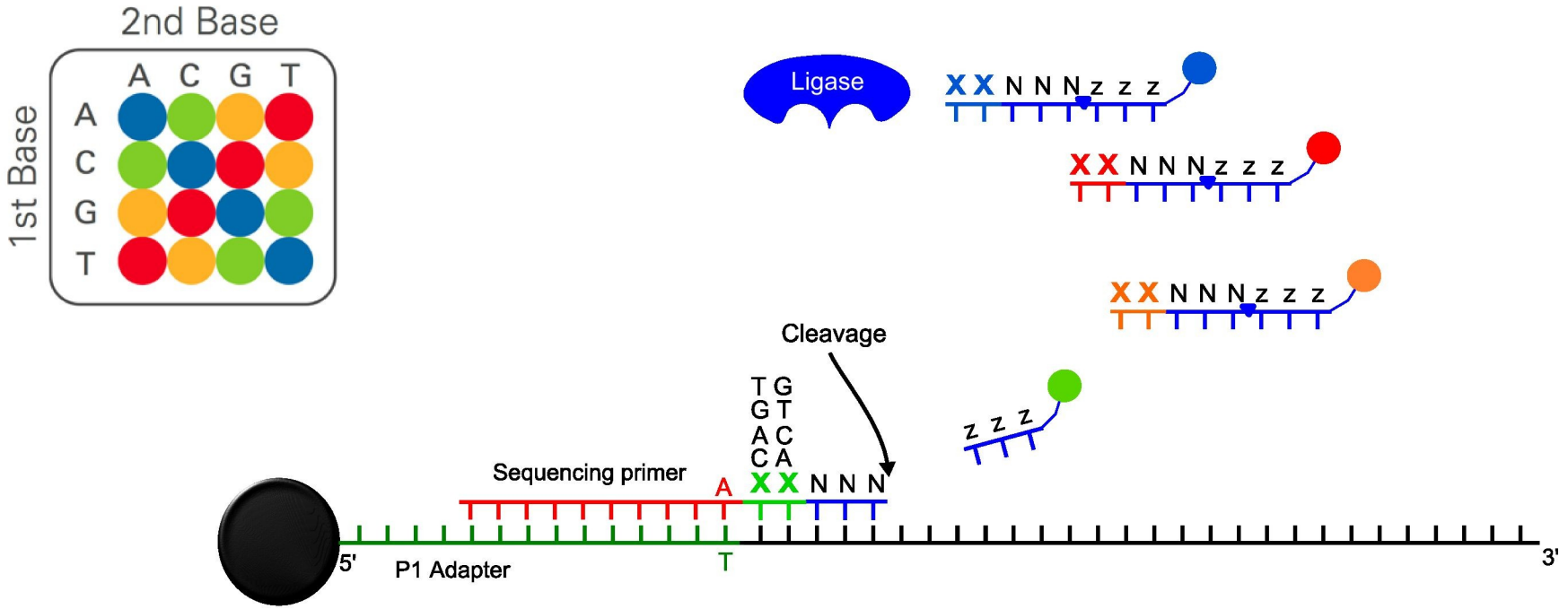
SOLiD



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

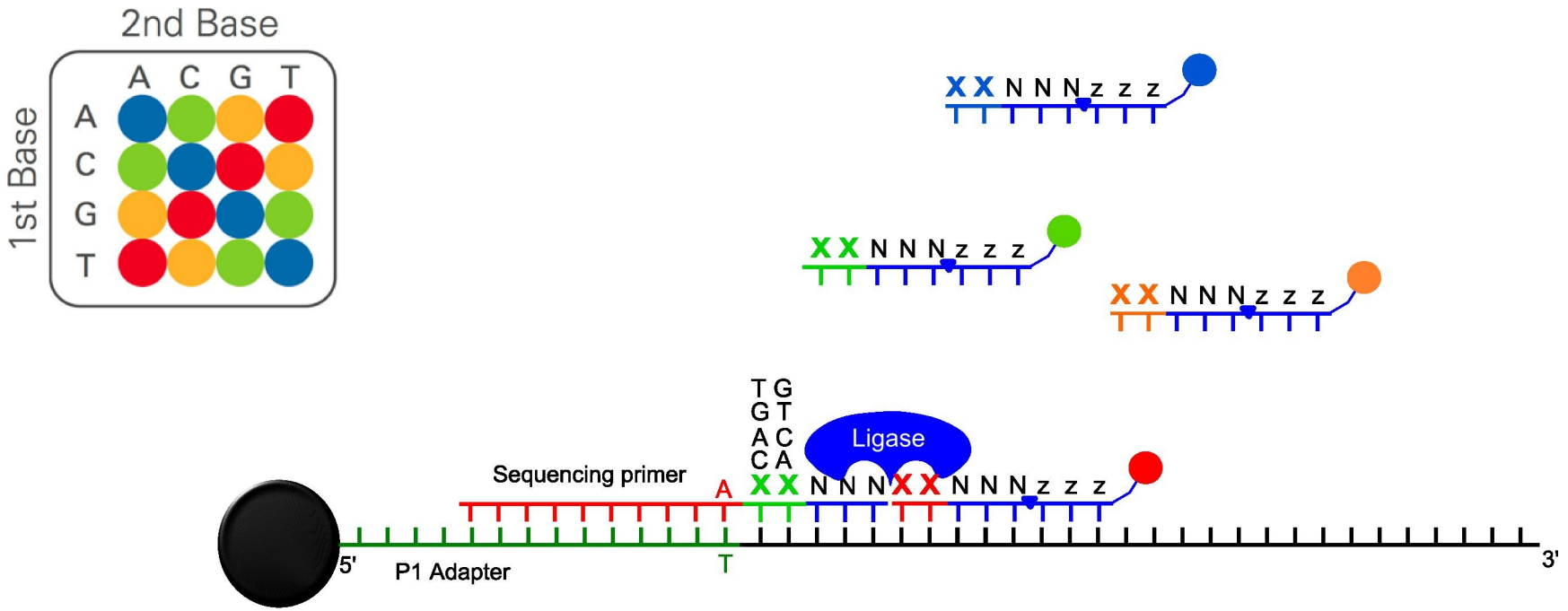
SOLiD



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

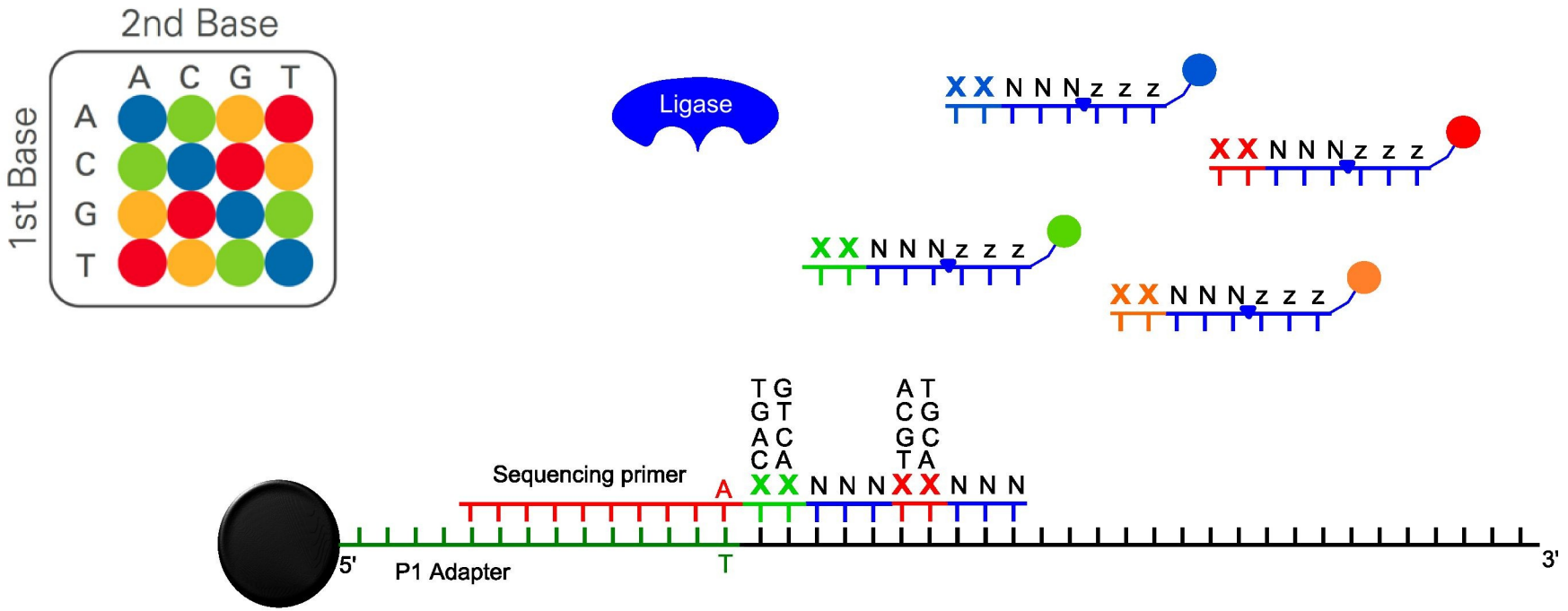
SOLiD



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

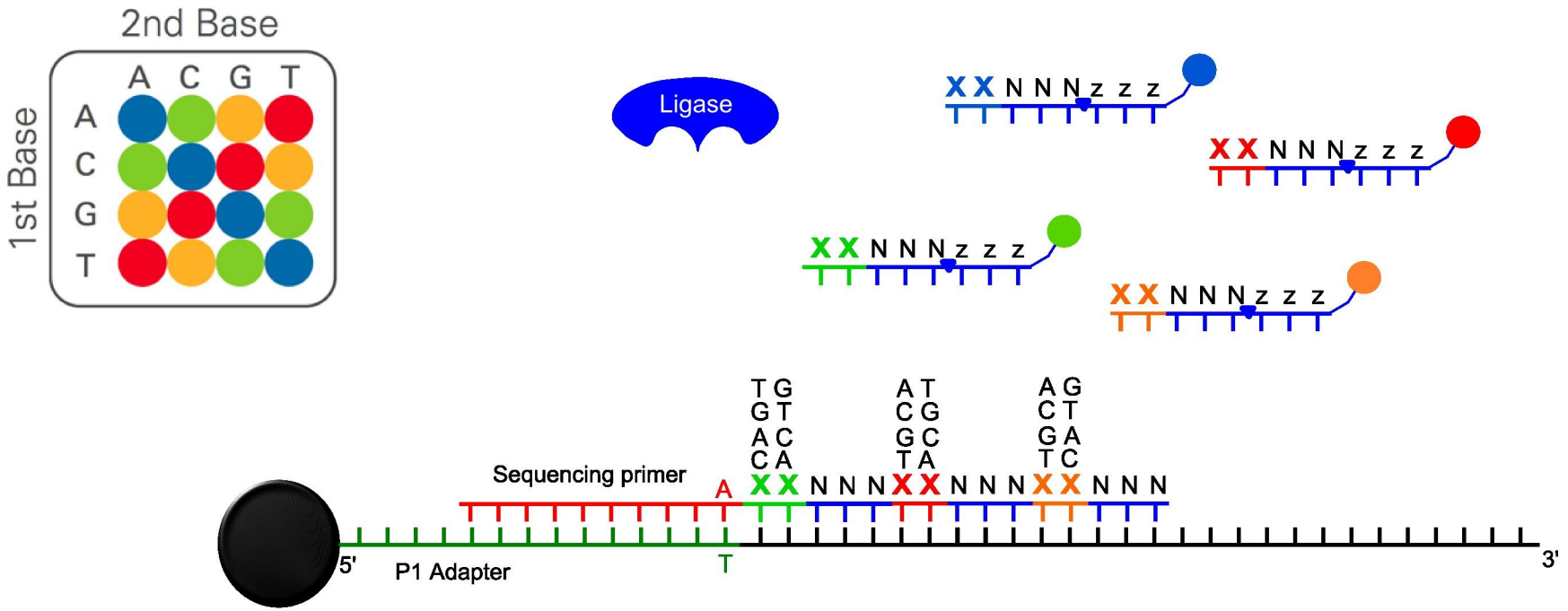
SOLiD



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

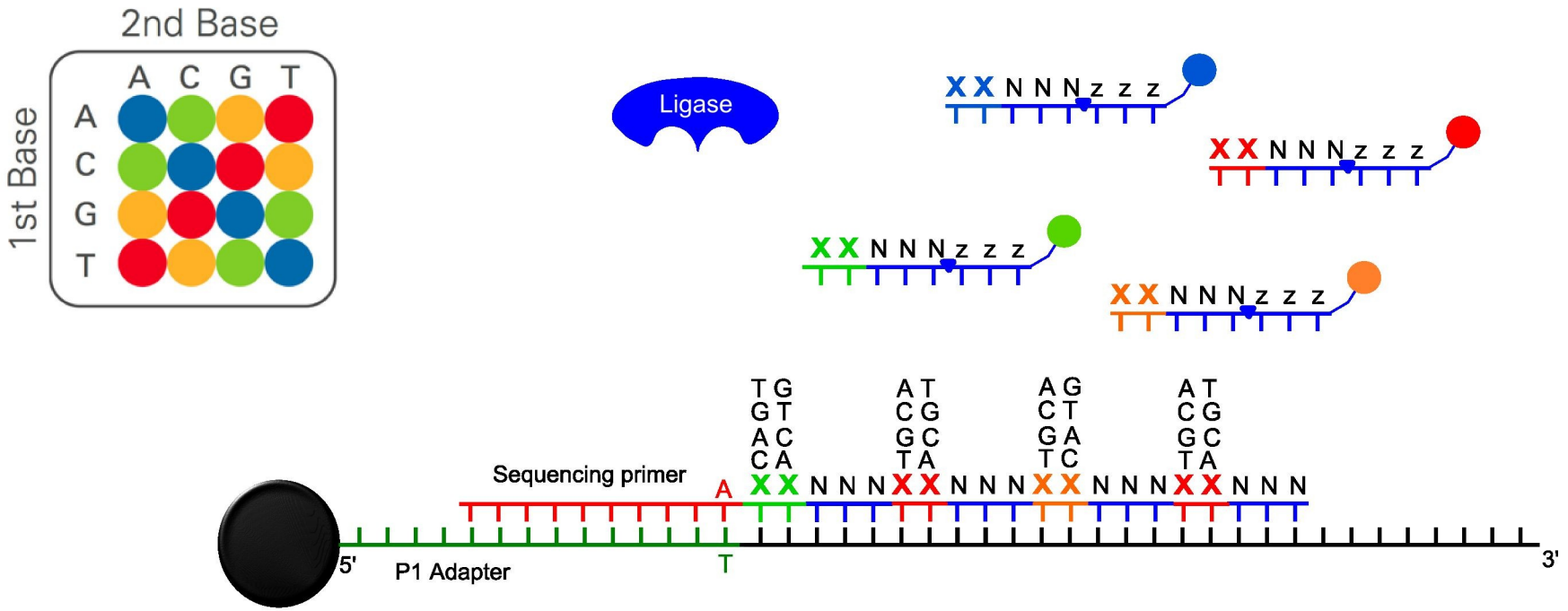
SOLiD



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

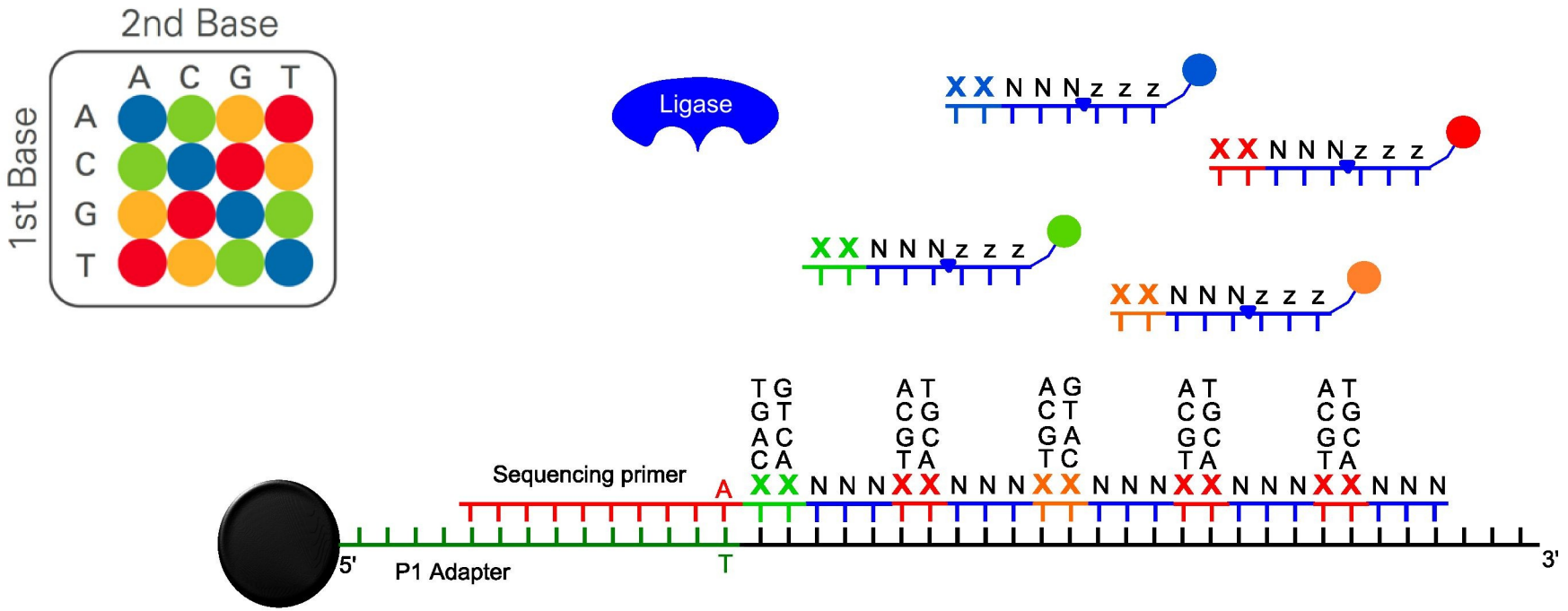
SOLiD



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

SOLiD

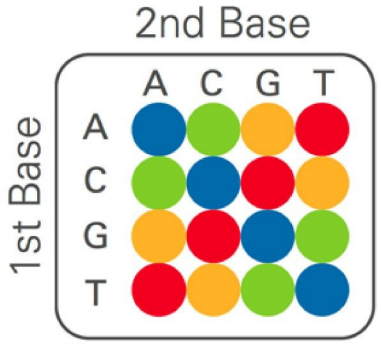


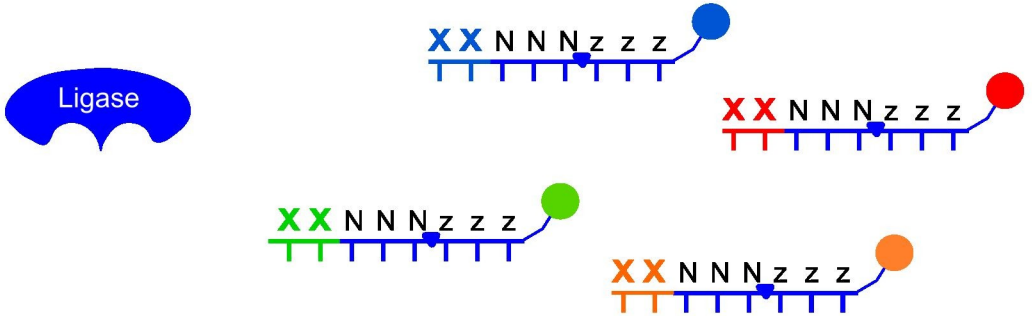
Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

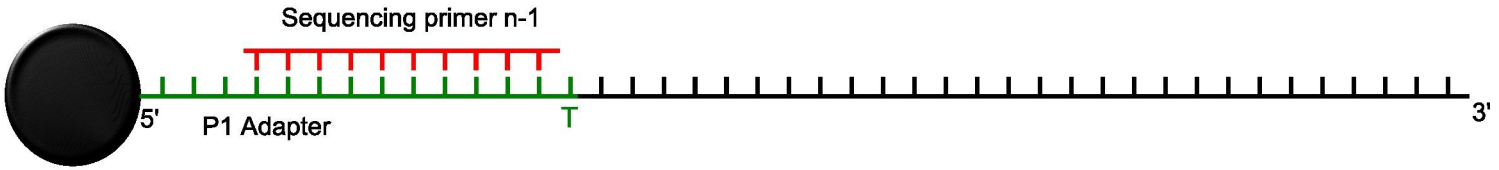
Next Generation Sequencing : Amplified Single Molecule Sequencing

SOLiD

2nd run :



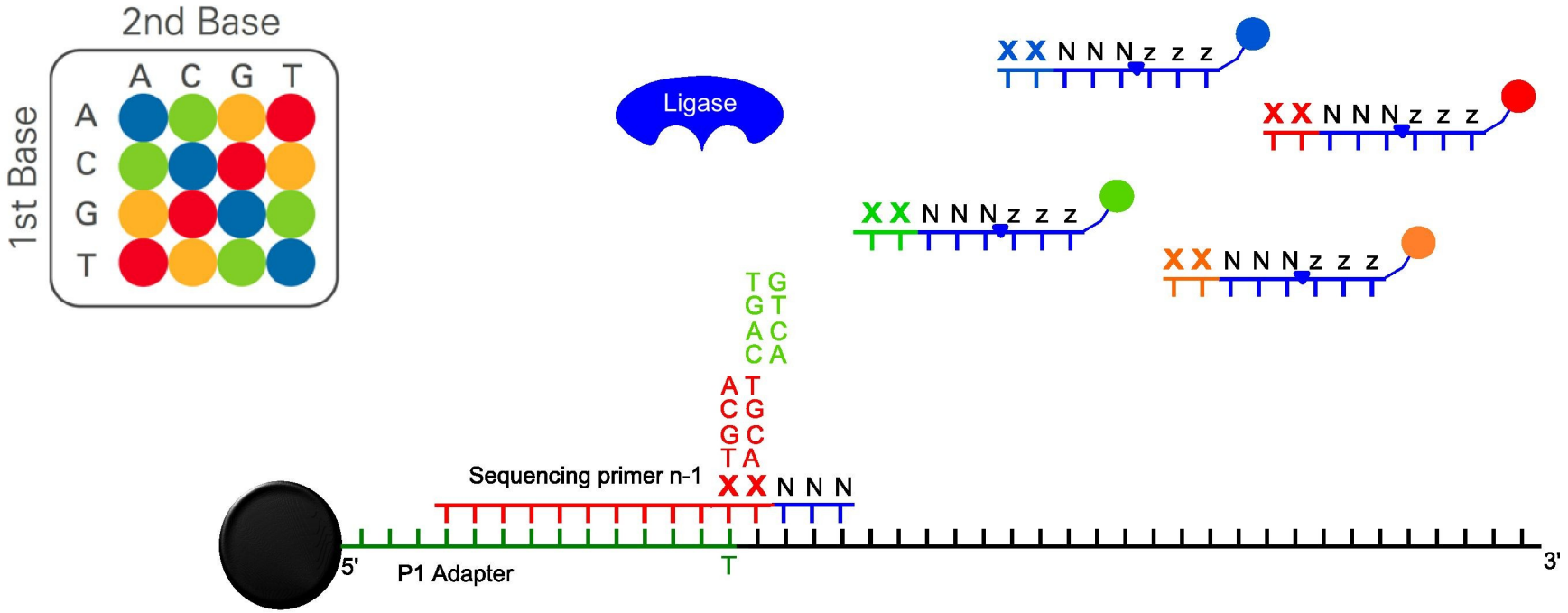




Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

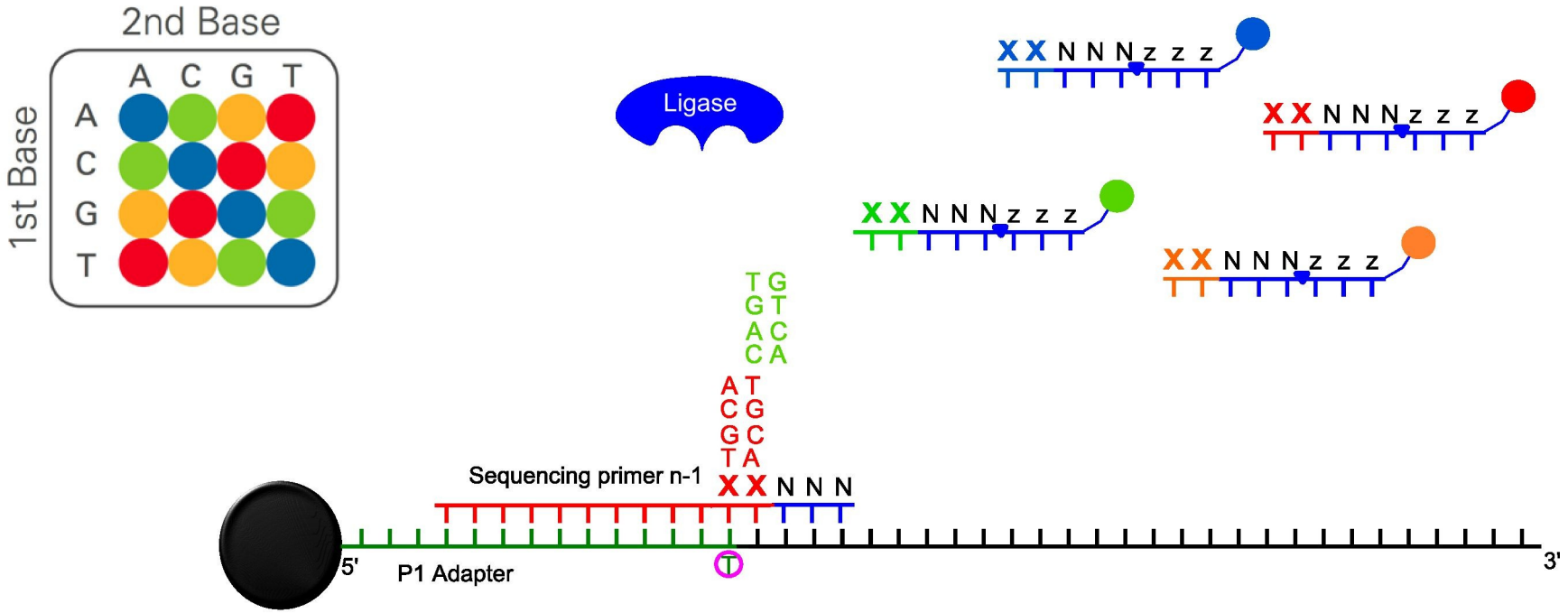
SOLiD



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

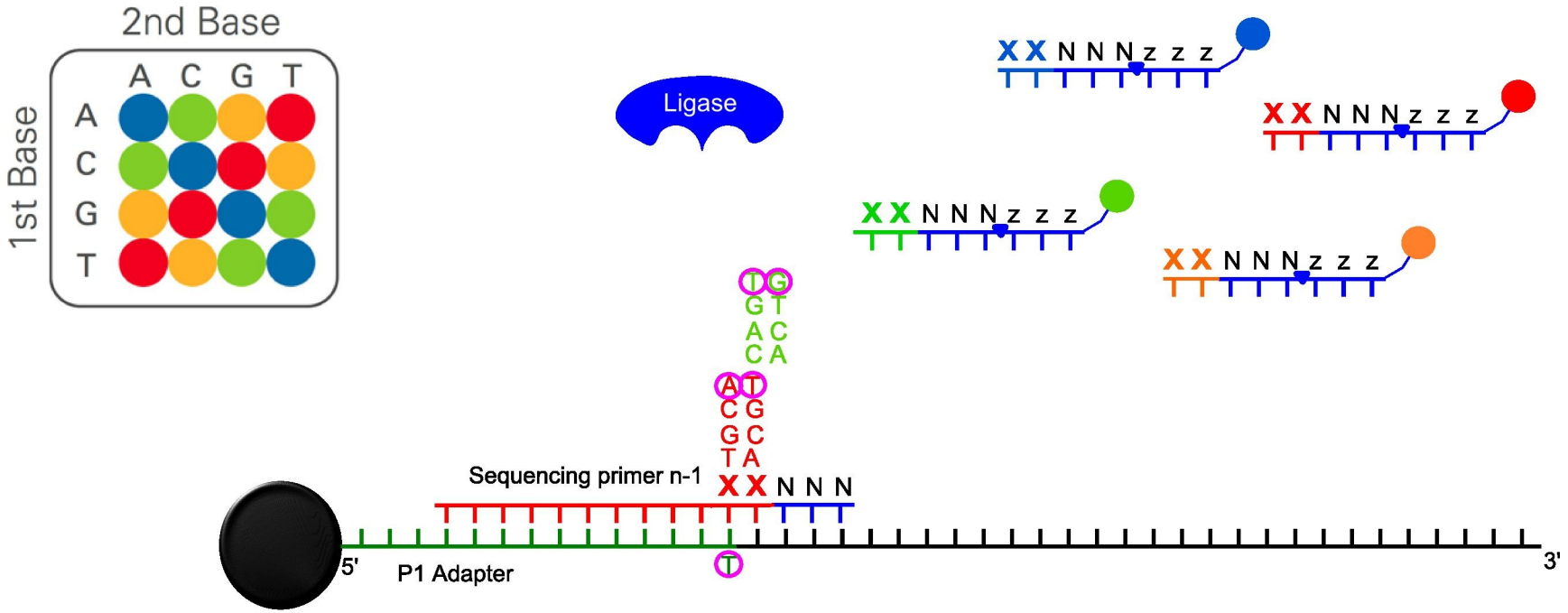
SOLiD



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

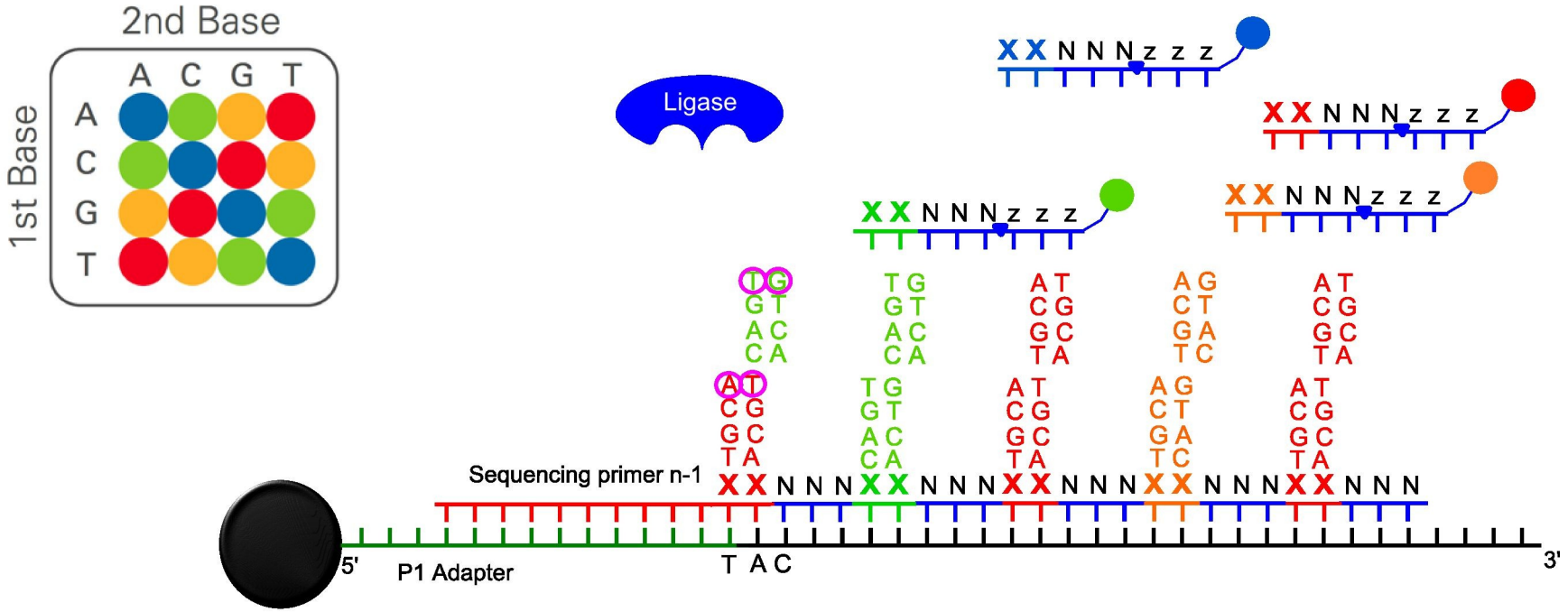
SOLiD



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

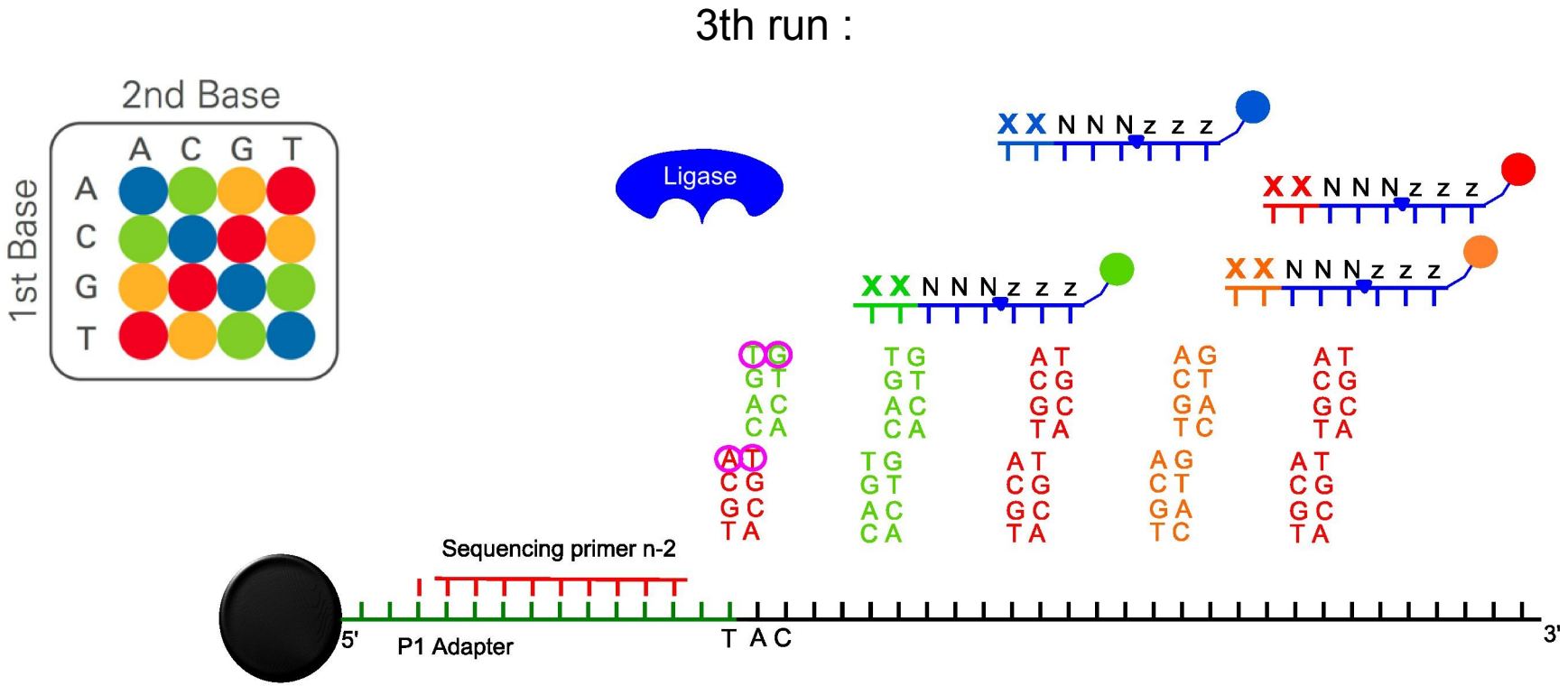
SOLiD



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

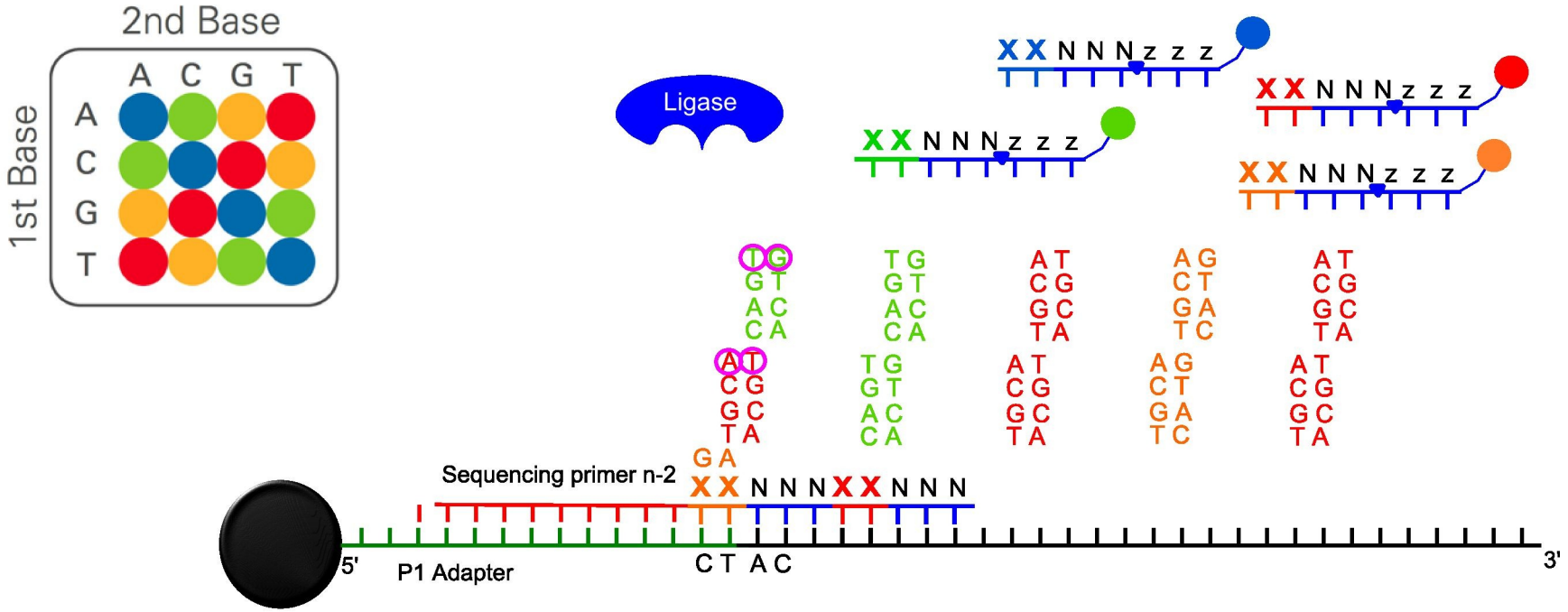
SOLiD



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

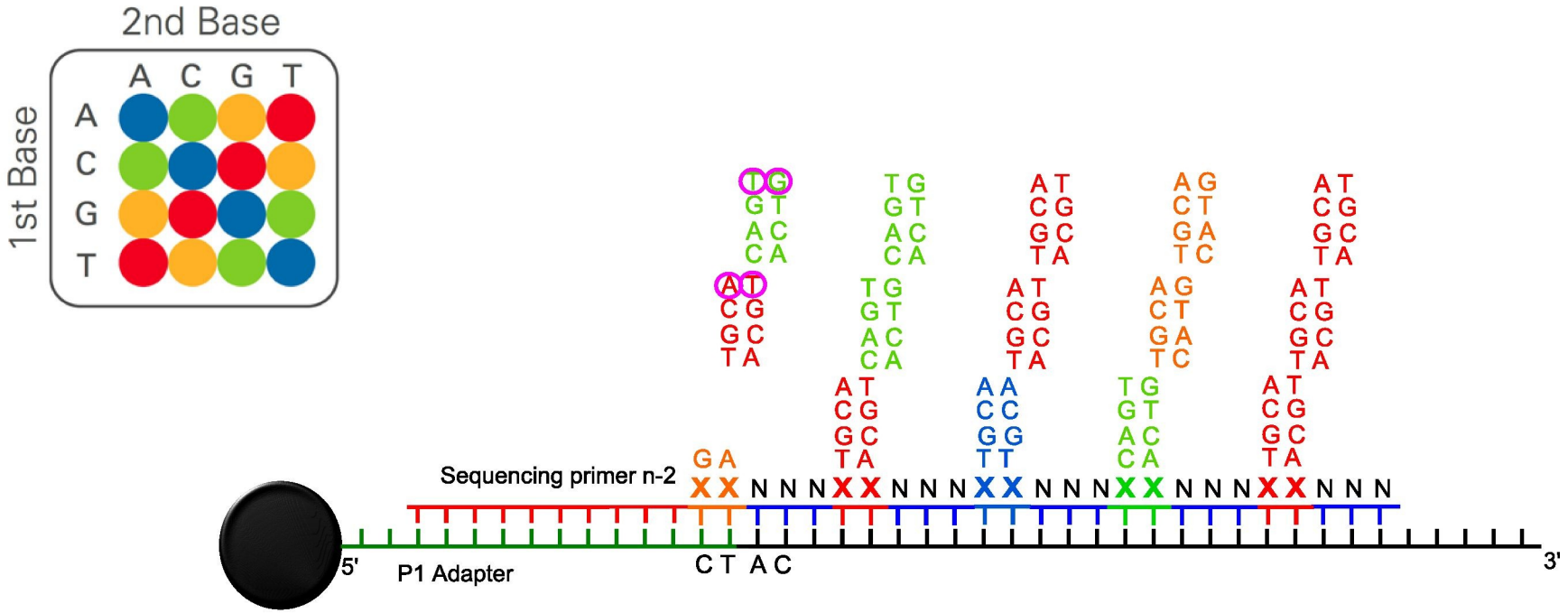
SOLiD



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

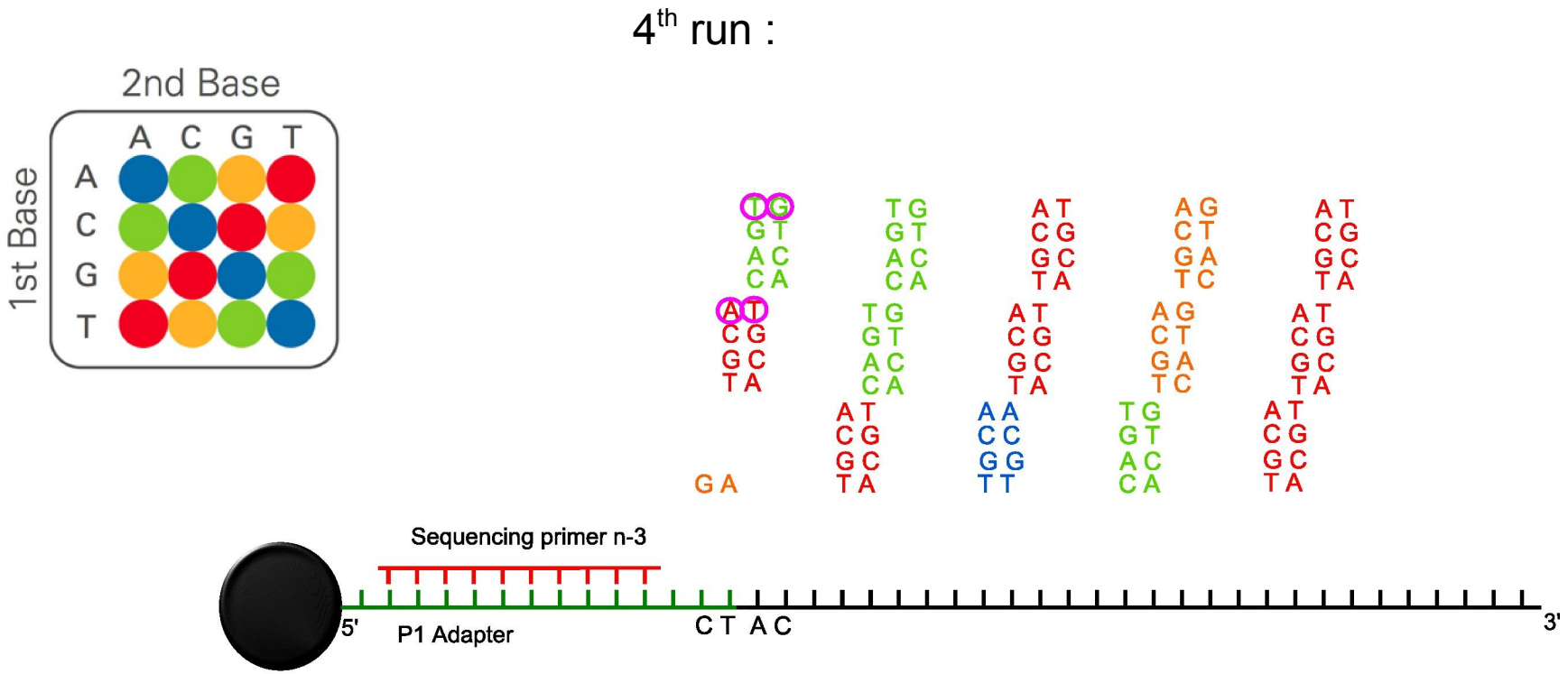
SOLiD



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

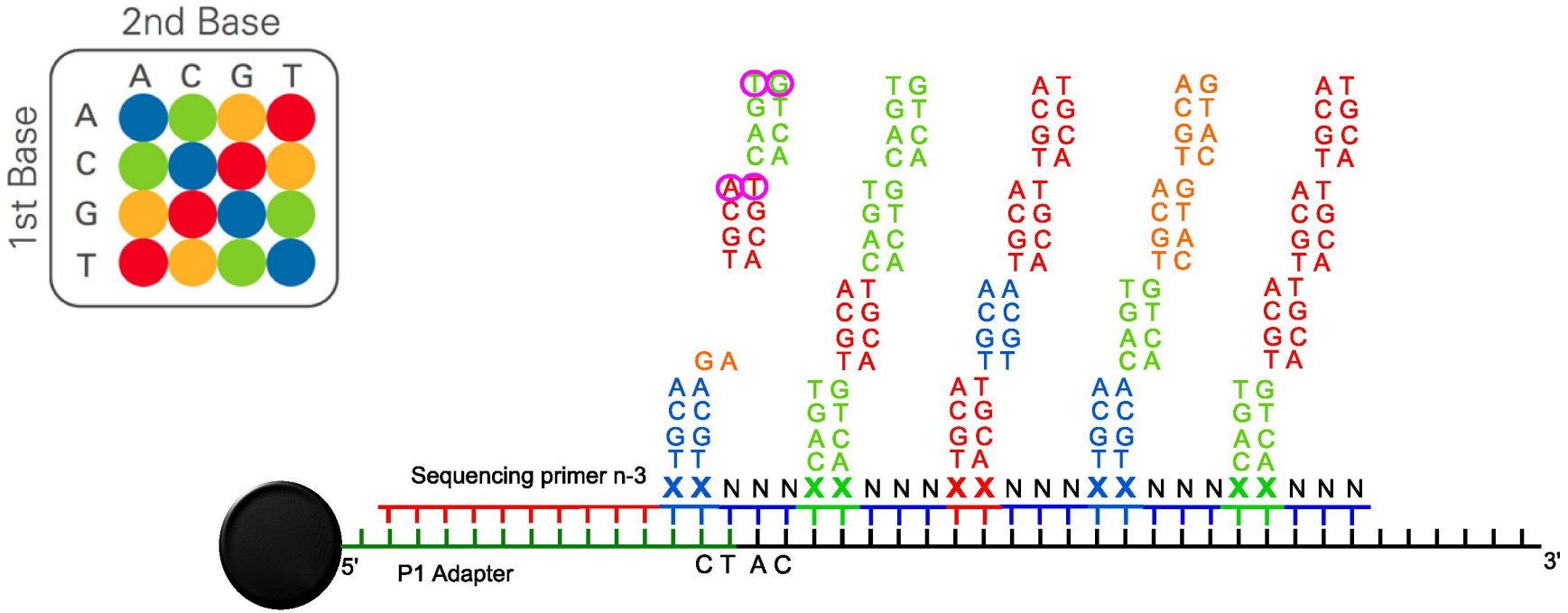
SOLiD



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

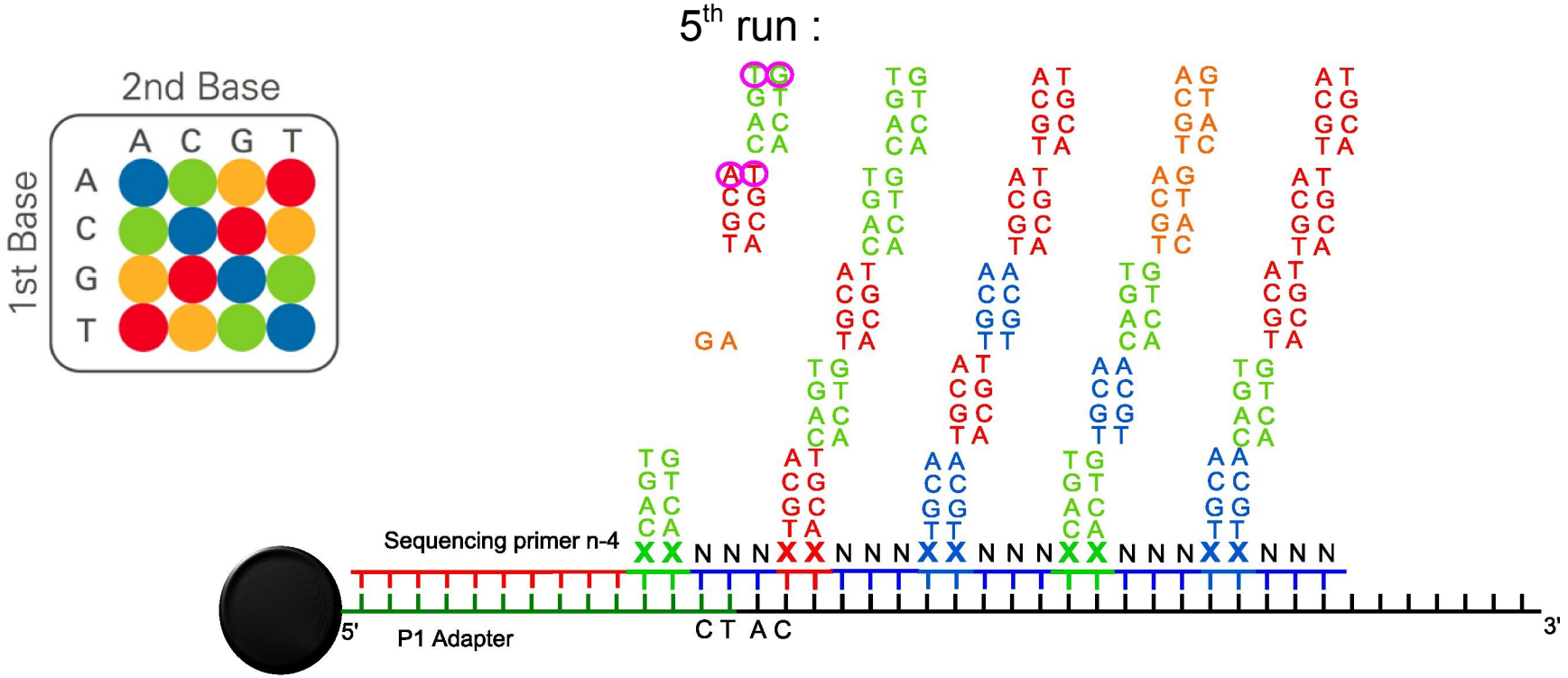
SOLiD



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

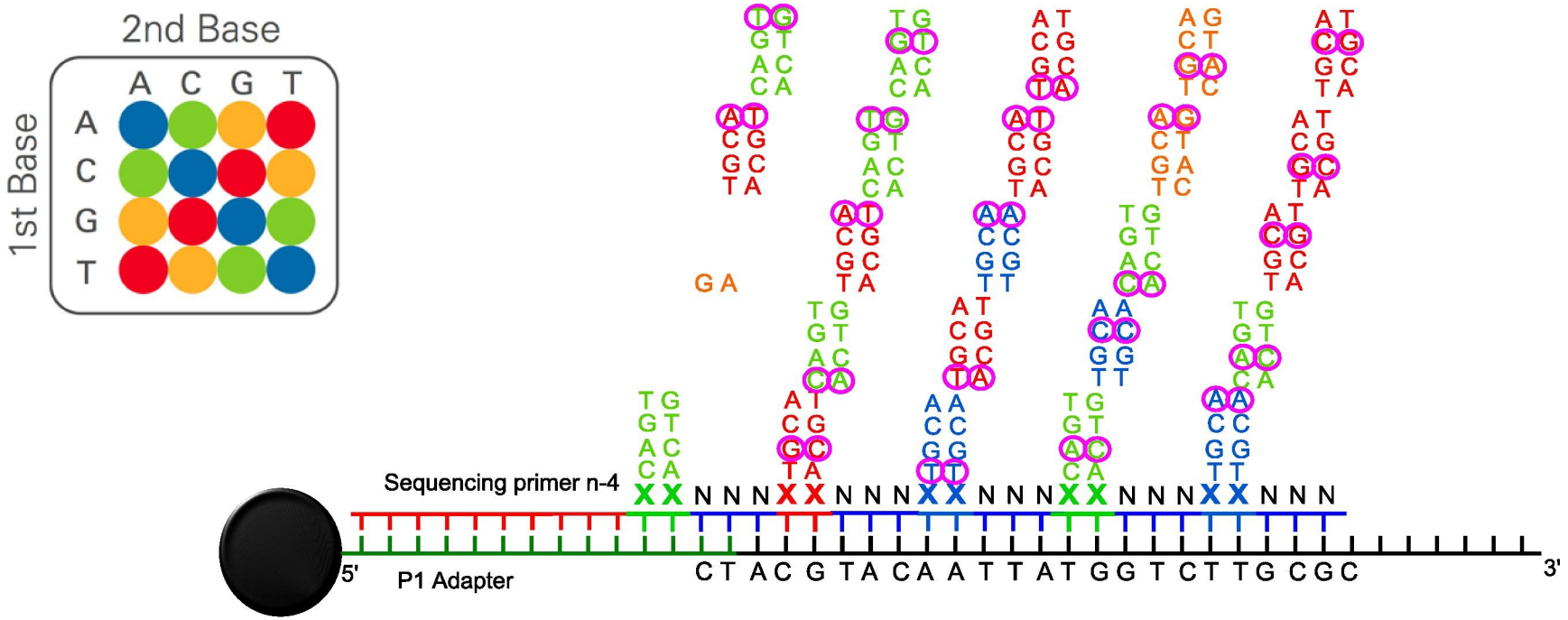
SOLiD



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

SOLiD

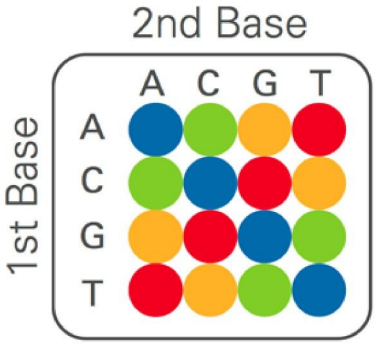


Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

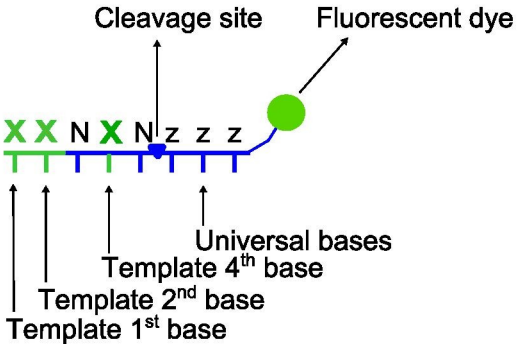
Next Generation Sequencing : Amplified Single Molecule Sequencing

SOLiD

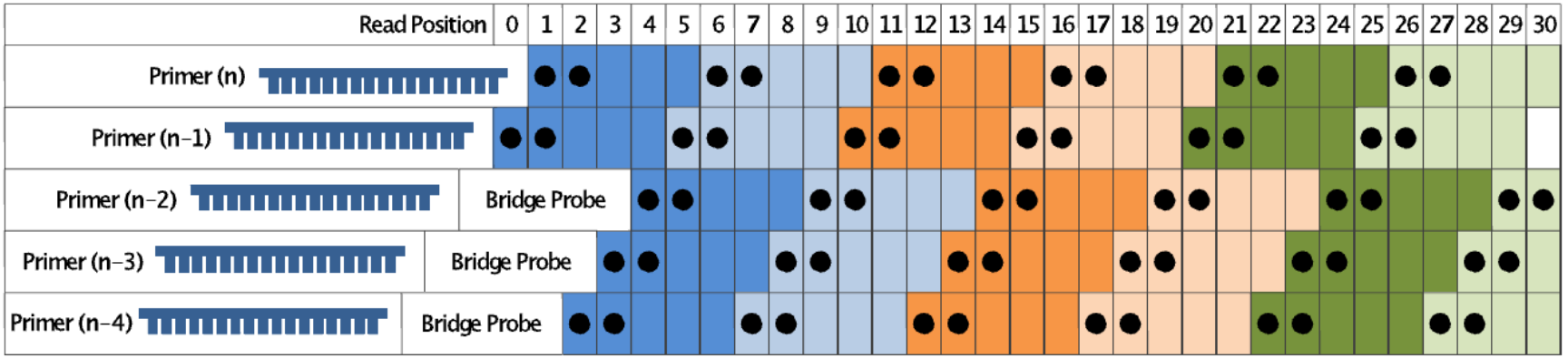
More accuracy : 3-base encoding probes



Probes set 2 : 3-base encoding

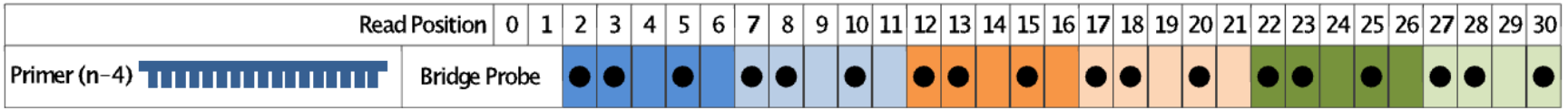


8. Repeat steps 1–7 with Primers n–1, n–2, n–3, and n–4 using Probe Set 1



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

● Interrogated Positions Ligation Cycle 1 2 3 4 5 6

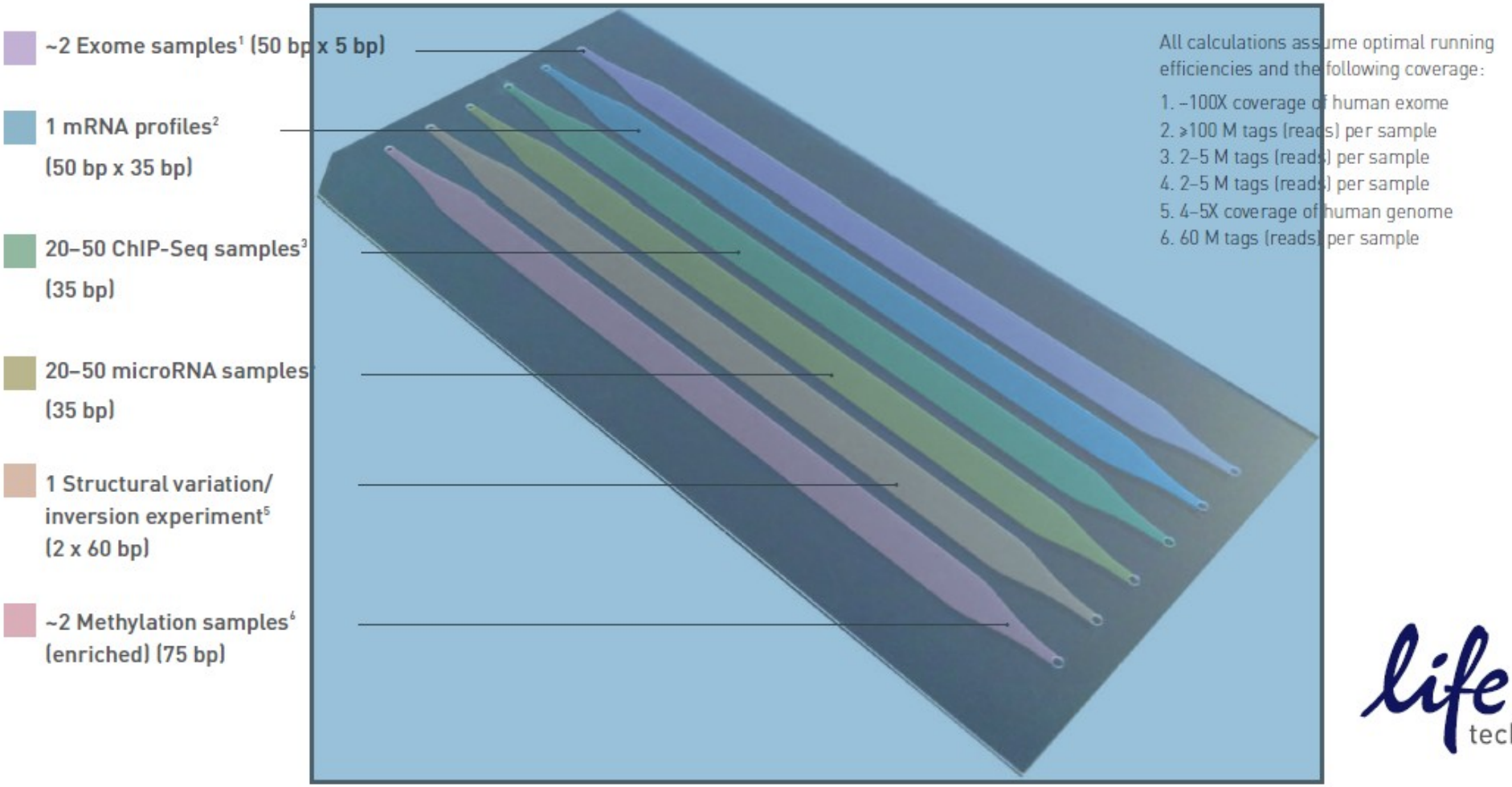


Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

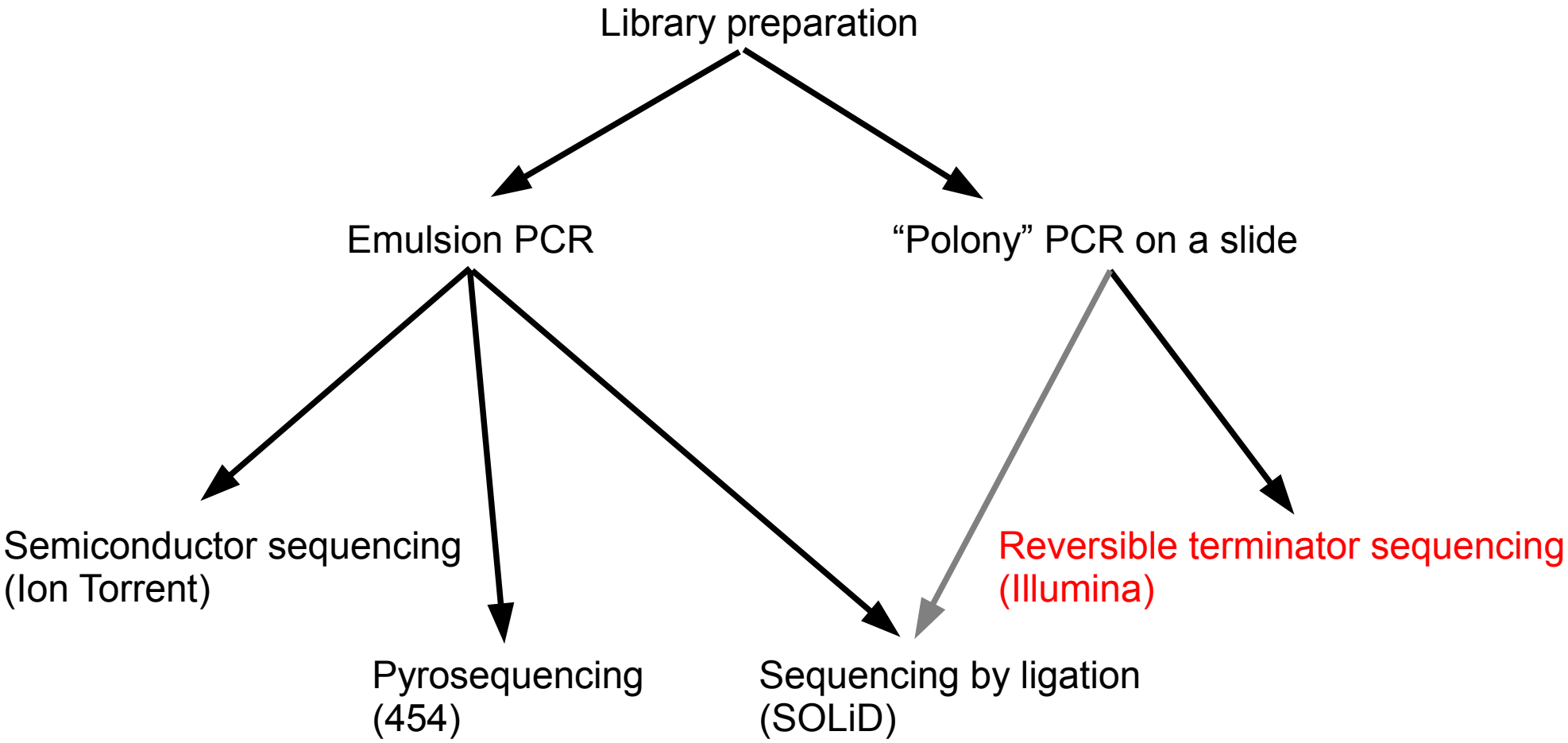
SOLiD

Figure 1. Multiple applications on a single FlowChip with different read lengths and chemistries.



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

Illumina

HiSeq 2500 / 2000 / 1500 / 1000



HiScanSQ



Genome Analyzer IIx



MiSeq



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

Illumina

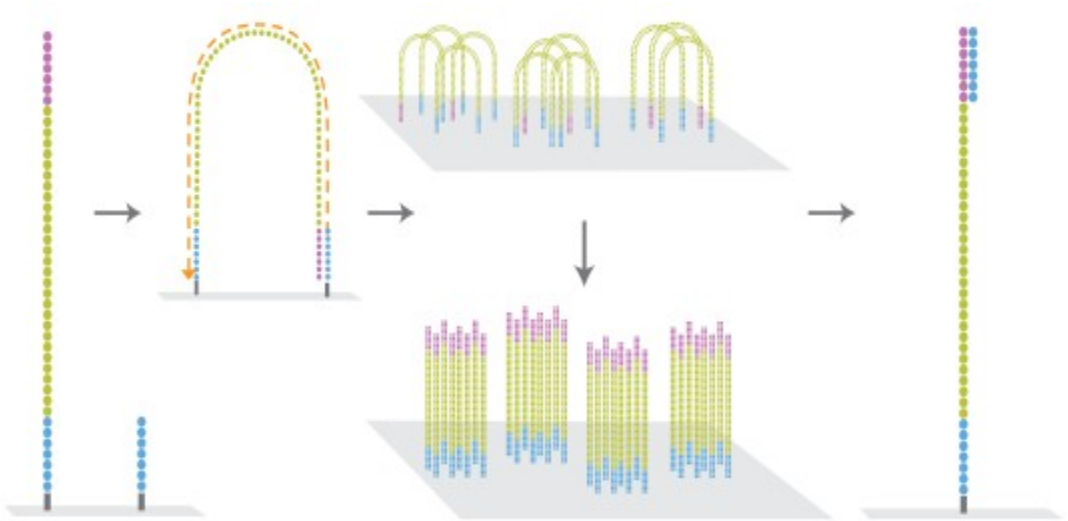
	HiSeq	HiScanSQ	Genome Analyzer IIx	MiSeq
Read Length	100 bp	100 bp	150 bp	250 bp
Throughput	600 Gb	150 Gb	95 Gb	6 Gb
Reads per run	3,000,000,000	750,000,000	320,000,000	12,000,000
Accuracy	99,9 %	99,9 %	99,9 %	99,9 %
Run Time	11 days	8 days	14 days	20-35 hours

Workflow : Library preparation → Bridge amplification → Reversible termination sequencing

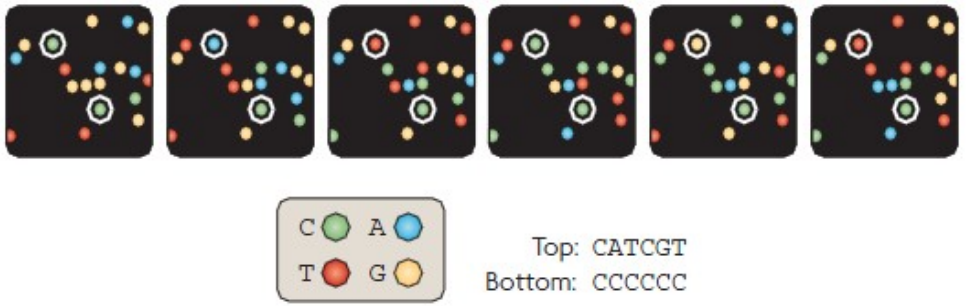
Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

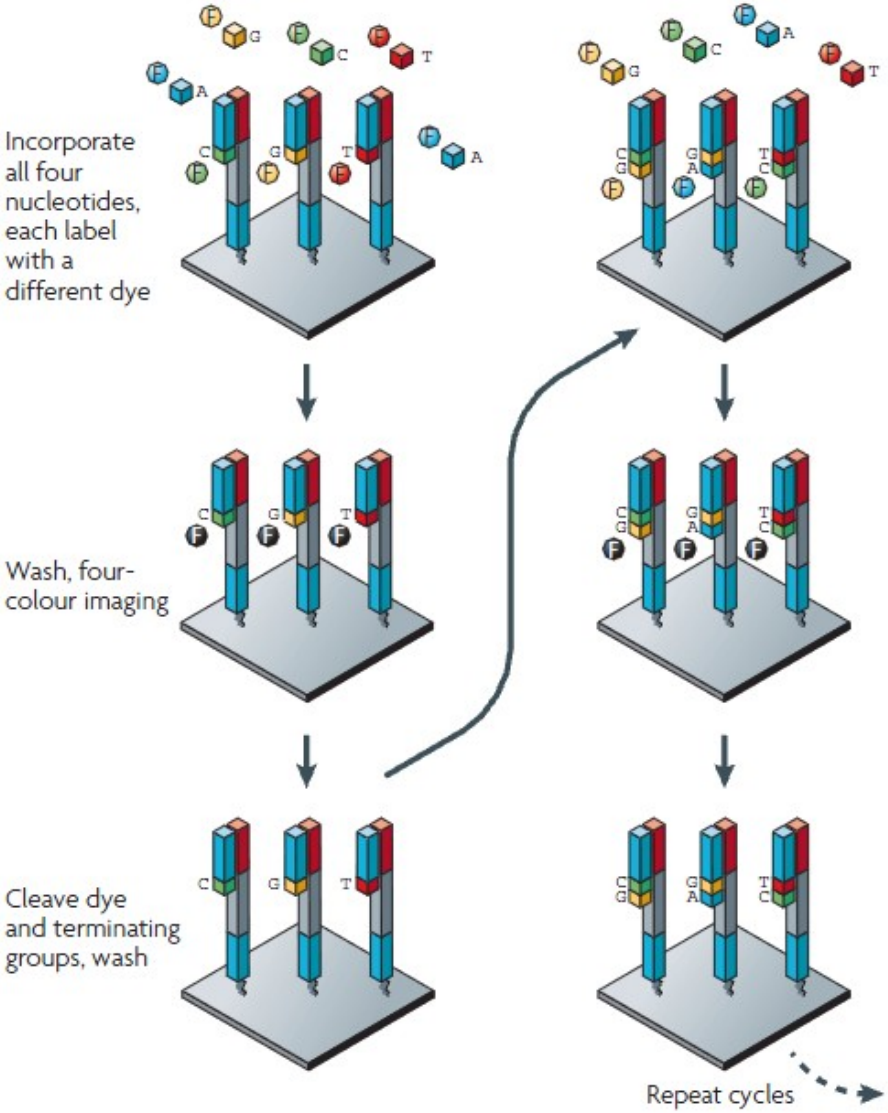
Next Generation Sequencing : Amplified Single Molecule Sequencing

Illumina



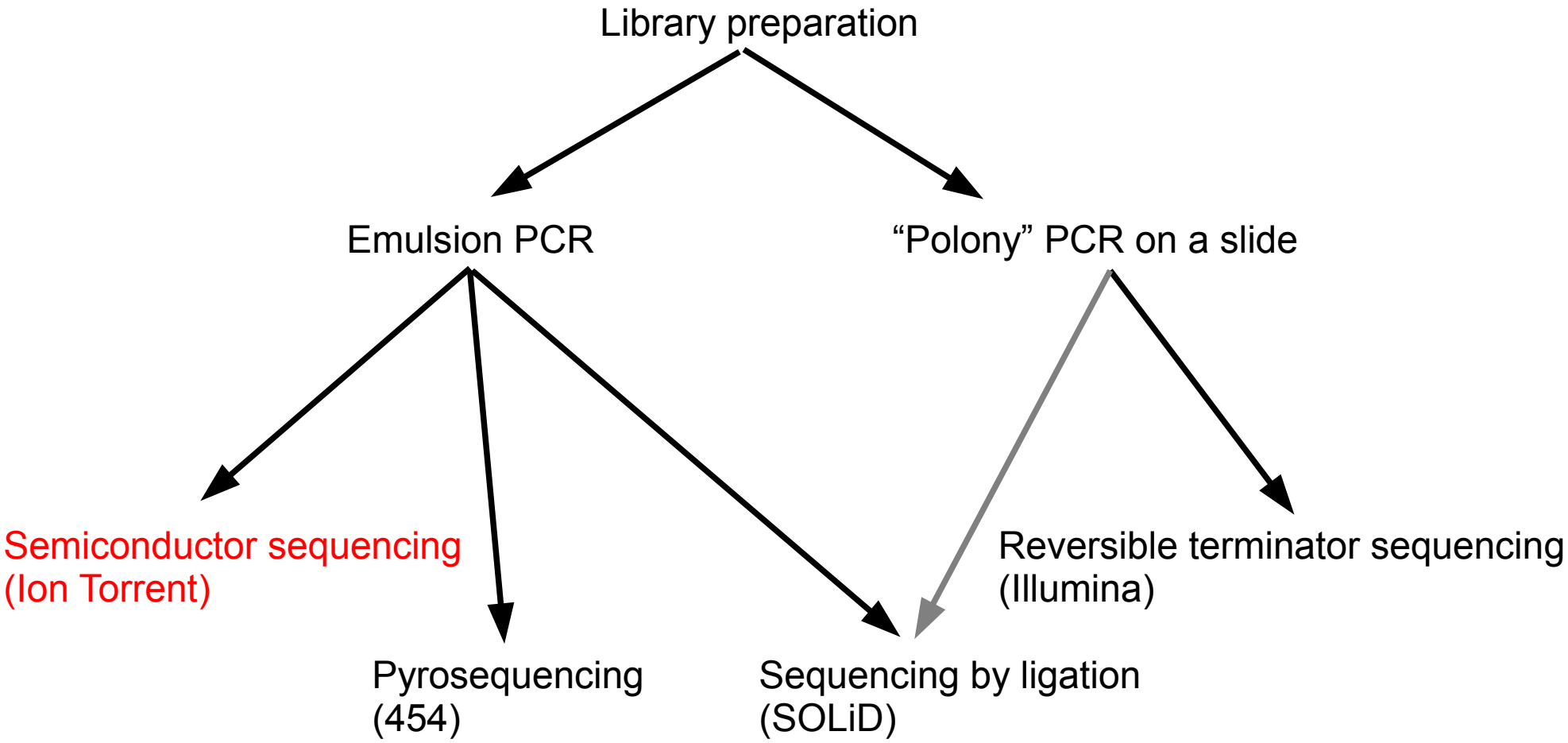
4 nucleotides with different dye
flow simultaneous





Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

Ion Torrent

PGM (Personal Genome Machine)



Proton

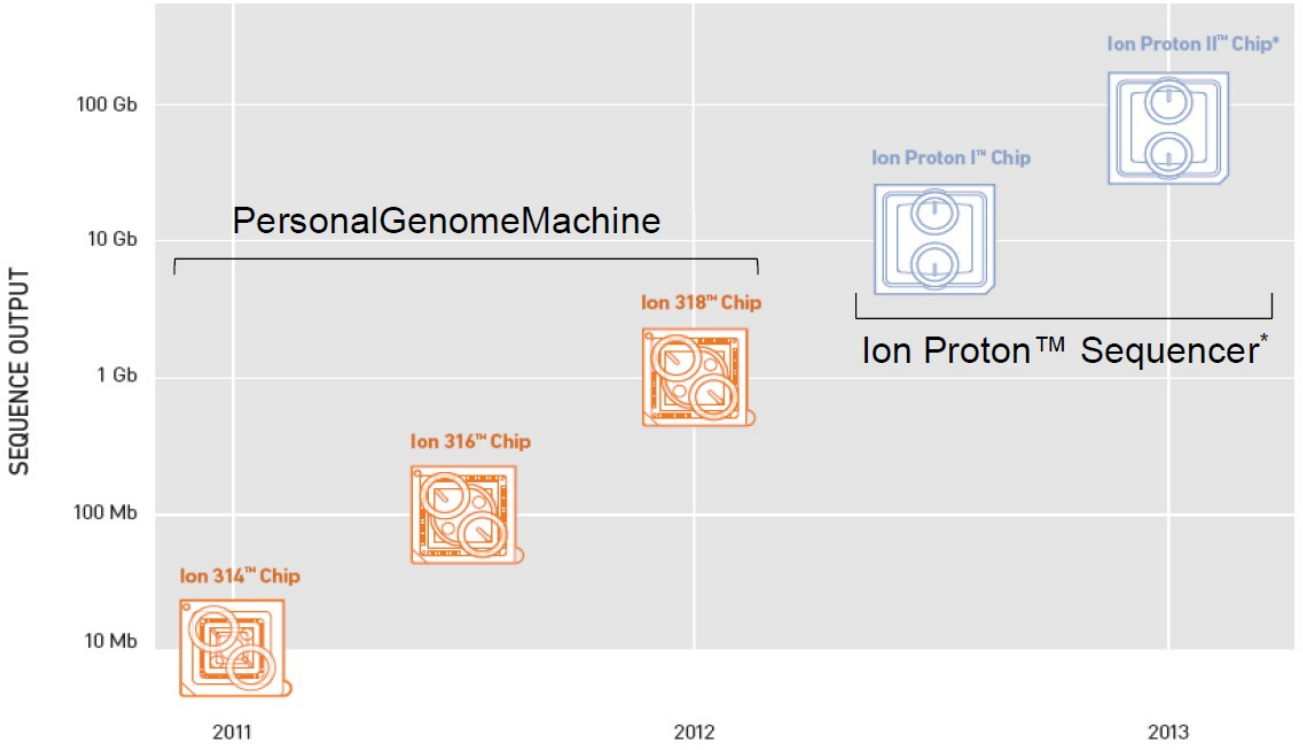


Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

Ion Torrent

	PGM	Proton
Read Length	200 bp	200 bp
Throughput	20 Mb - 1 Gb	10 -100 Gb
Reads per run	11,000,000	660,000,000
Accuracy	99 %	99 %
Run Time	4,5 hours	4,5 hours

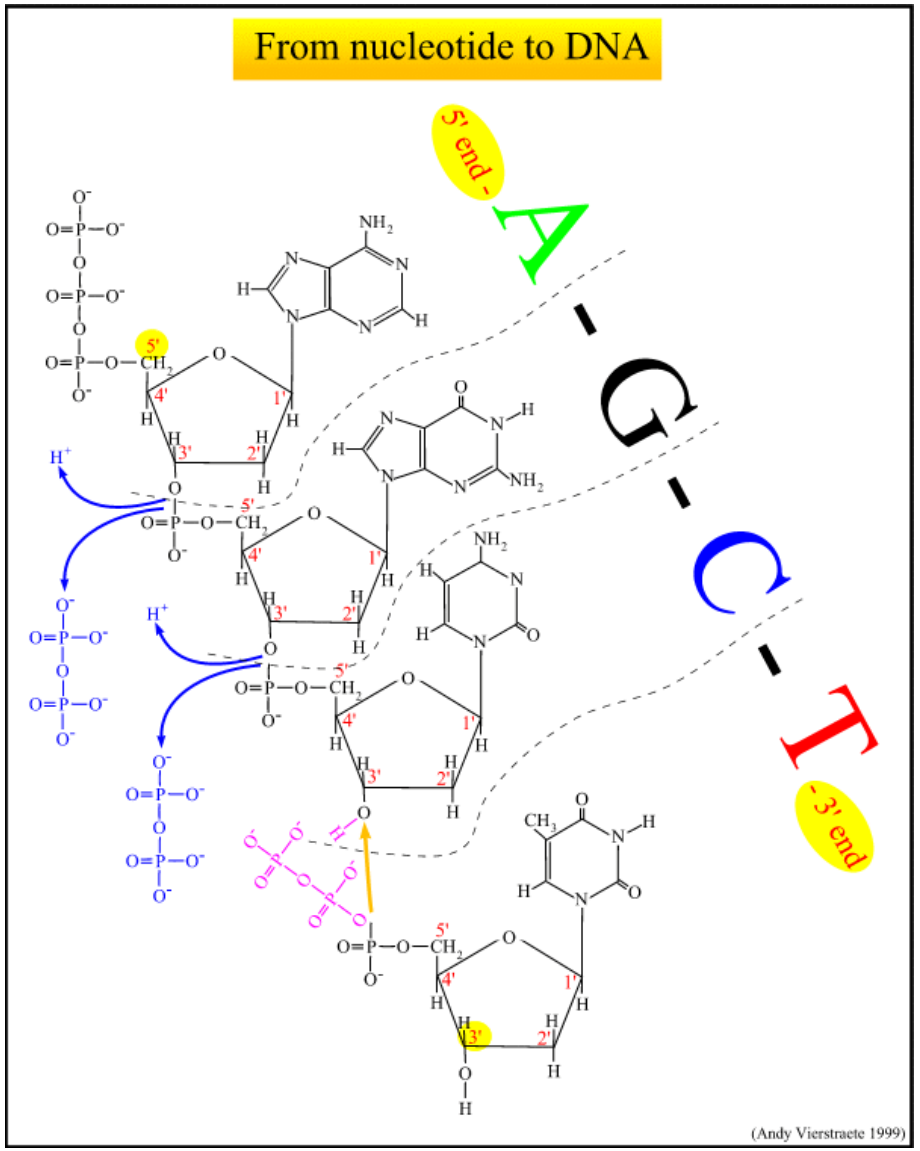


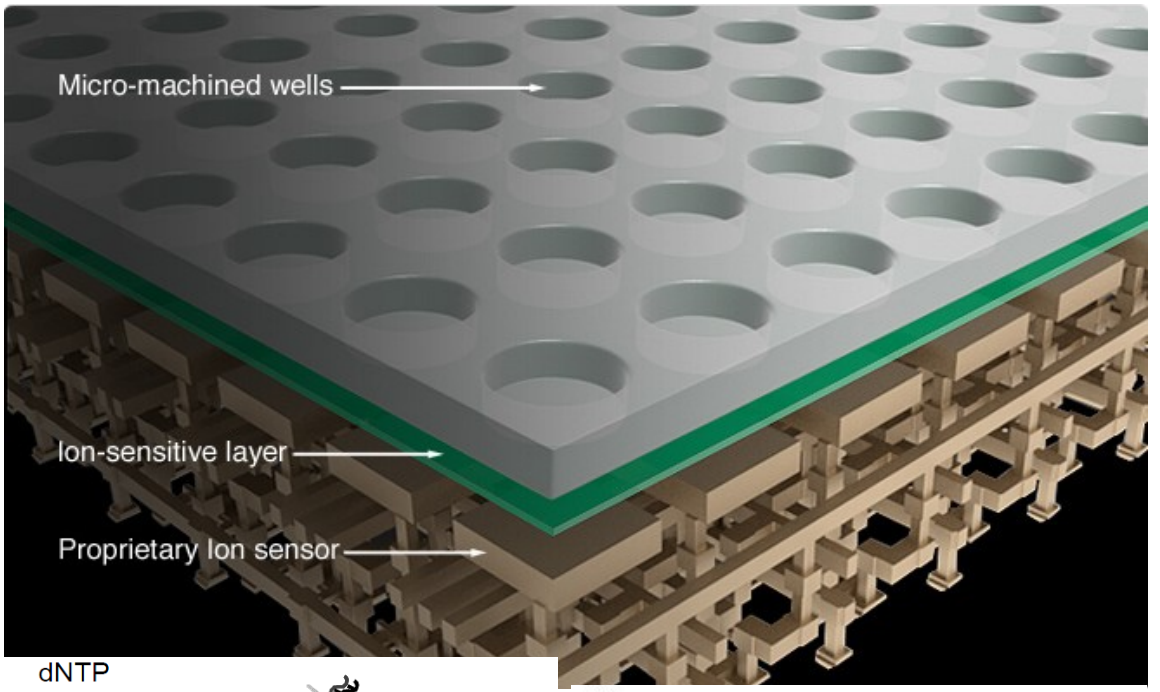
Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

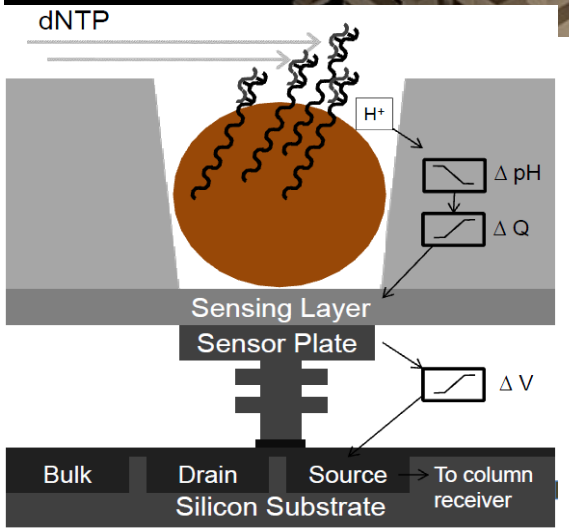
Next Generation Sequencing : Amplified Single Molecule Sequencing

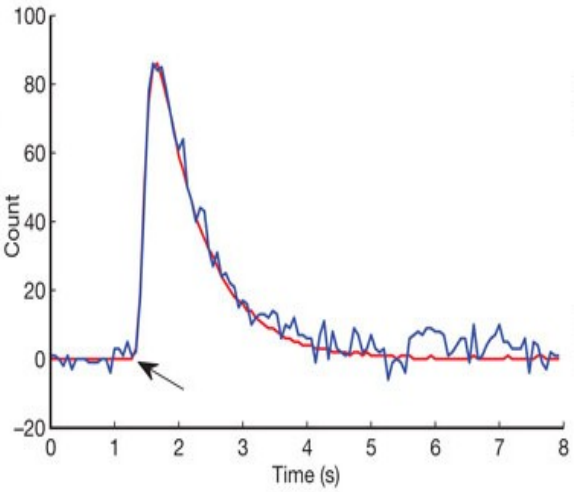
Ion Torrent

Workflow : Library preparation → Emulsion PCR → Semiconductor Sequencing







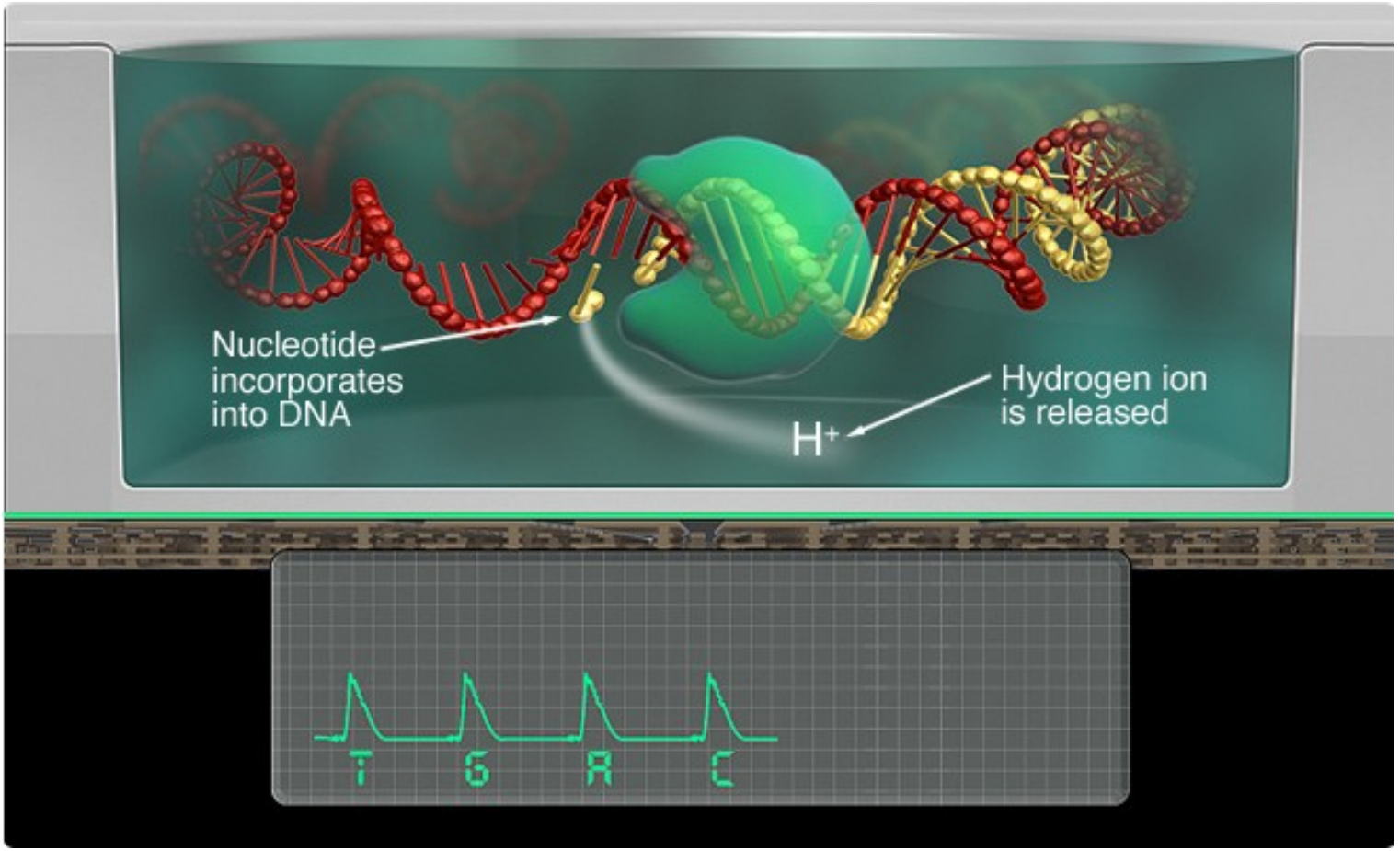


Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

Ion Torrent

4 nucleotides flow sequentially

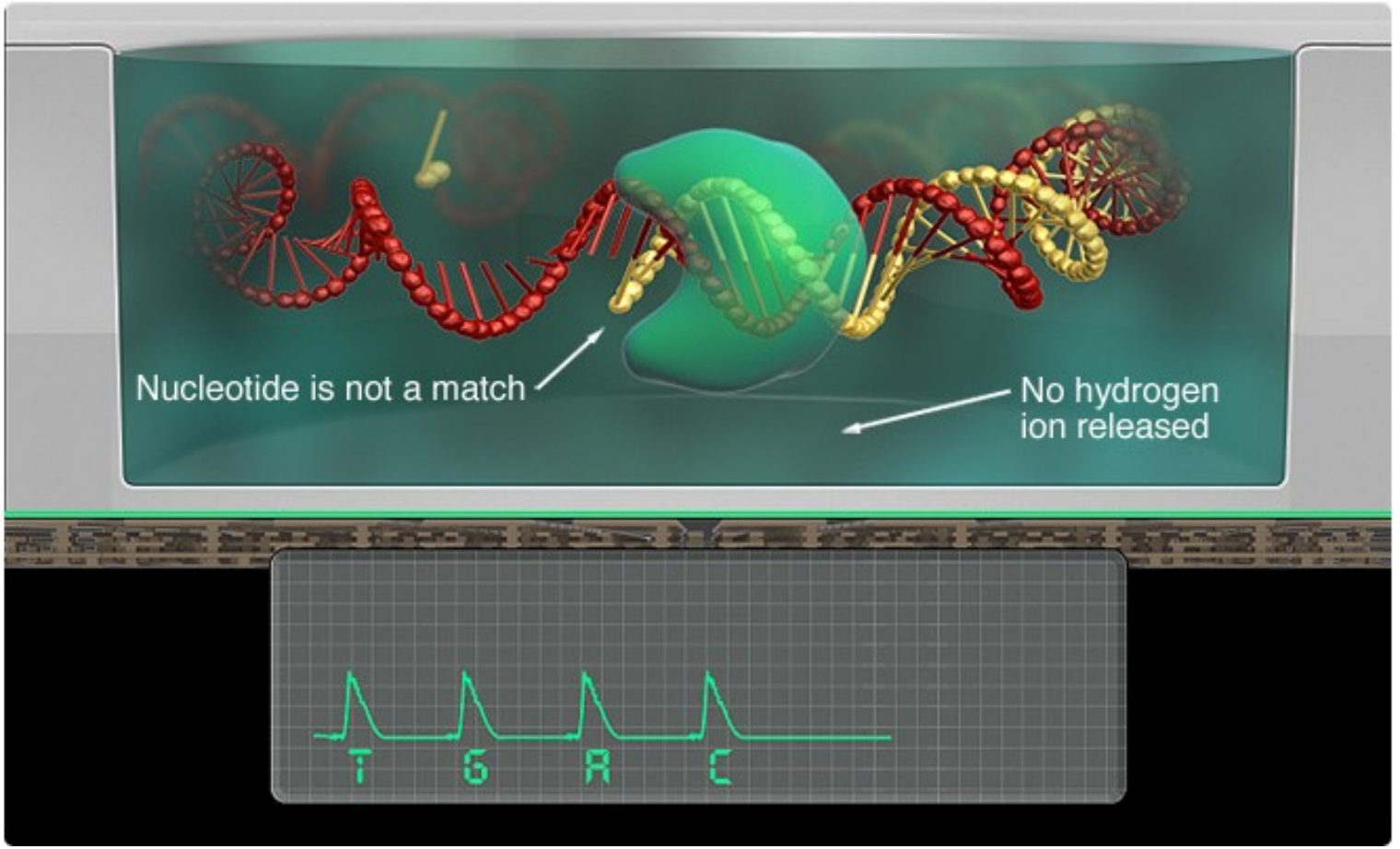


No camera, just a pH sensor

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

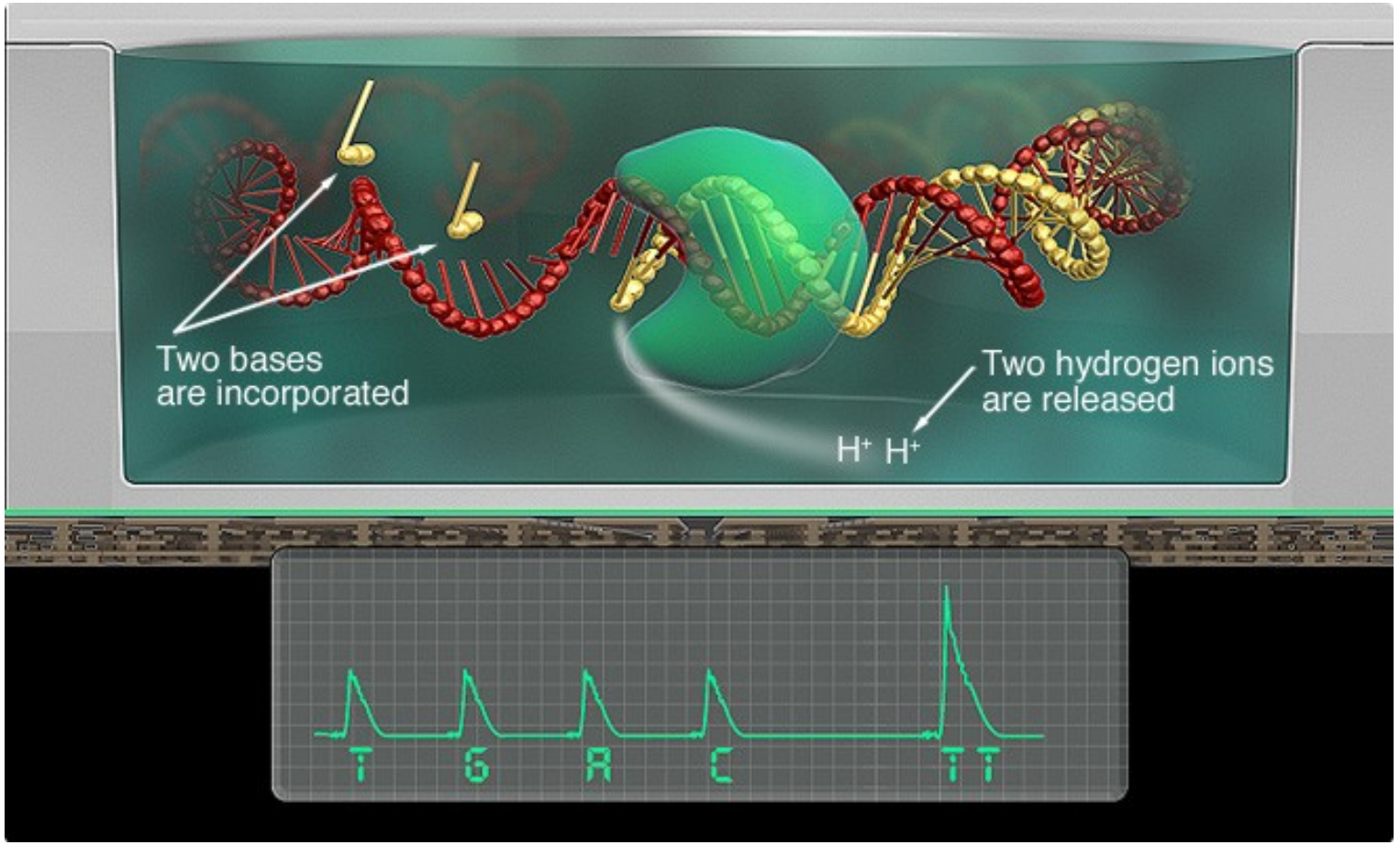
Ion Torrent



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

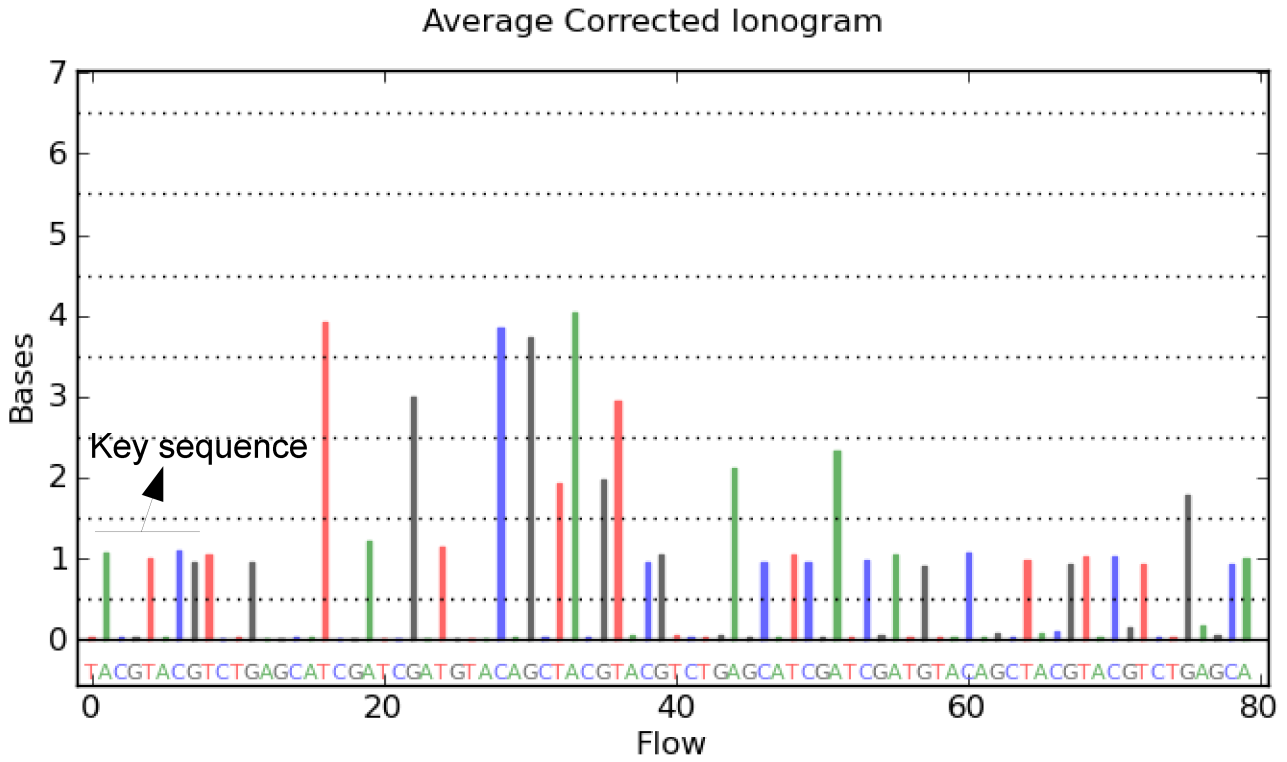
Ion Torrent



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

Ion Torrent



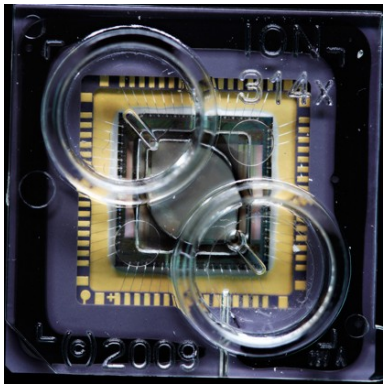
ATCGTGTTTTAGGGTCCCCGGGGTT...

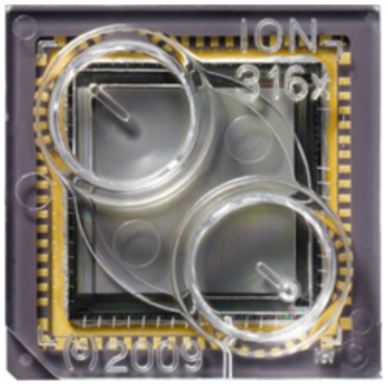
Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

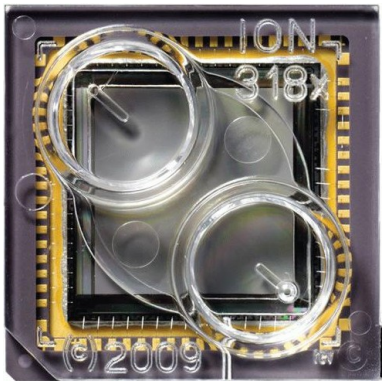
Next Generation Sequencing : Amplified Single Molecule Sequencing

Ion Torrent

400 base reads end of 2012 ?







Chip	Expected Sequencing Run Time			Expected Output		
	35 base reads	100 base reads	200 base reads	35 base reads	100 base reads	200 base reads
 Ion 314™ Chip	0.5 hr	1.5 hr	2.4 hr	3 Mb	10 Mb	20 Mb
 Ion 316™ Chip	0.7 hr	1.7 hr	3.1 hr	30 Mb	100 Mb	200 Mb
 Ion 318™ Chip	0.9 hr	2.4 hr	4.5 hr	300 Mb	500 Mb	up to 1 GB

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

Ion Torrent

Number of wells : 1 million 6 million 11 million
Expected output : 20 Mb 200 Mb 1Gb

Week Ending In	314	316	318
April 22, 2012	jodeeg (101.78Mb)	chris8796 (330.71Mb)	pli (954.53Mb)
April 15, 2012	leemurphy (75.02Mb)	stopkad (401.54Mb)	ymong (676.30Mb)
April 8, 2012	gallagation (56.26Mb)	howedder (488.81Mb)	stopkad (602.72Mb)
April 1, 2012	mhaak (25.99Mb)	choips (461.59Mb)	jjackson (752.64Mb)
March 2012	cover_bio (80.38Mb)	choips (461.59Mb)	jjackson (752.64Mb)
March 25, 2012	OwenLeiser (60.96Mb)	dcarter (374.30Mb)	-
March 18, 2012	cover_bio (80.38Mb)	cover_bio (516.99Mb)	stopkad (291.71Mb)
March 11, 2012	jsheffield (78.65Mb)	jamiezhou (152.39Mb)	pli (723.79Mb)
March 4, 2012	mhaak (57.17Mb)	chris8796 (238.49Mb)	-
February 2012	cavatorm (77.28Mb)	pli (466.46Mb)	doneill (816.26Mb)
February 26, 2012	cavatorm (77.28Mb)	chris8796 (139.88Mb)	-
February 19, 2012	dgallo (50.66Mb)	chris8796 (205.92Mb)	doneill (816.26Mb)
February 12, 2012	rk8052zh (71.74Mb)	pli (466.46Mb)	-
February 5, 2012	klilakos (68.36Mb)	howedder (271.11Mb)	-
January 2012	rk8052zh (96.06Mb)	pli (596.76Mb)	jkron (924.10Mb)
January 30, 2012	sherke (54.83Mb)	sherke (151.81Mb)	jkron (924.10Mb)
January 23, 2012	rk8052zh (96.06Mb)	tstinear (231.38Mb)	-
January 16, 2012	cindau (30.78Mb)	pli (596.76Mb)	Jkron (752.74Mb)
January 9, 2012	tneff (33.48Mb)	RMease (269.20Mb)	jboland (804.45Mb)
January 2, 2012	-	jboland (315.76Mb)	-
December 2011	rtewh (103.30Mb)	pli (589.74Mb)	mholder (989.96Mb)
December 25, 2011	leemurphy (56.37Mb)	cover_bio (470.92Mb)	mholder (988.96Mb)
December 18, 2011	cindau (61.97Mb)	doneill (436.20Mb)	jkron (778.66Mb)
December 11, 2011	rtewh (103.30Mb)	pli (589.74Mb)	-
December 4, 2011	cavatorm (78.47Mb)	mzkg (355.13Mb)	-
November 2011	cavatorm (78.47Mb)	jsheffield (350.93Mb)	-
November 27, 2011	-	-	-
November 20, 2011	-	-	-
November 13, 2011	kgould (65.73Mb)	jsheffield (350.93Mb)	-

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

Ion Torrent

Library Preparation

Fragment Library

1. Fragmentation
2. Adapter ligation
3. Size selection
4. Nick Translation and Amplification
5. Quantification

Total Time ~3.5 hrs

Amplicon Library

1. PCR
2. Clean-Up
3. Quantification

Total Time ~3 hours

RNA Library

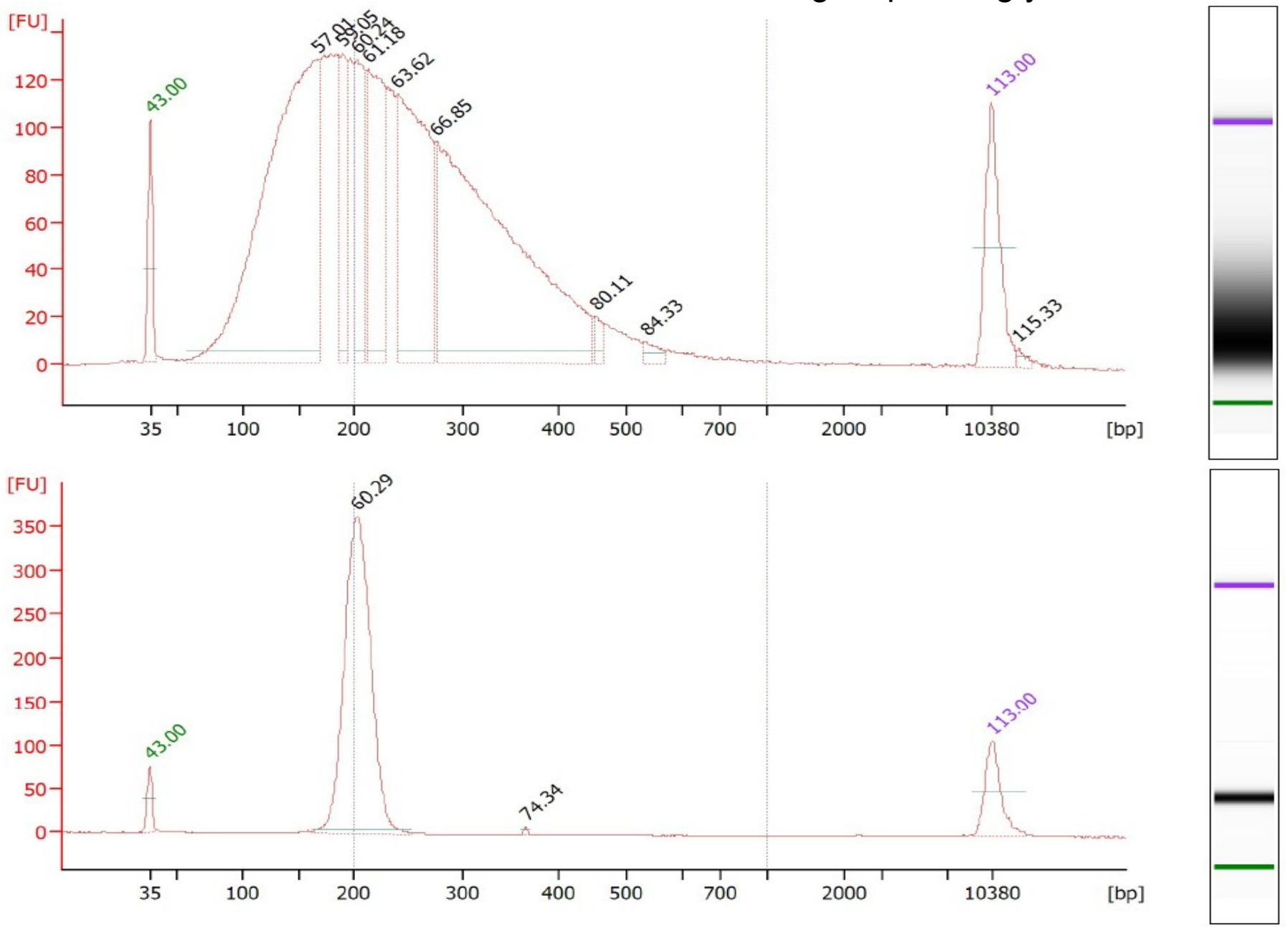
1. Prepare WT or miRNA
2. Adapter Ligation
3. Reverse Transcription
4. Size selection
5. Amplification
6. Quantification

Total Time ~1.5 Days)

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

Ion Torrent Size selection - > maximizing sequencing yield

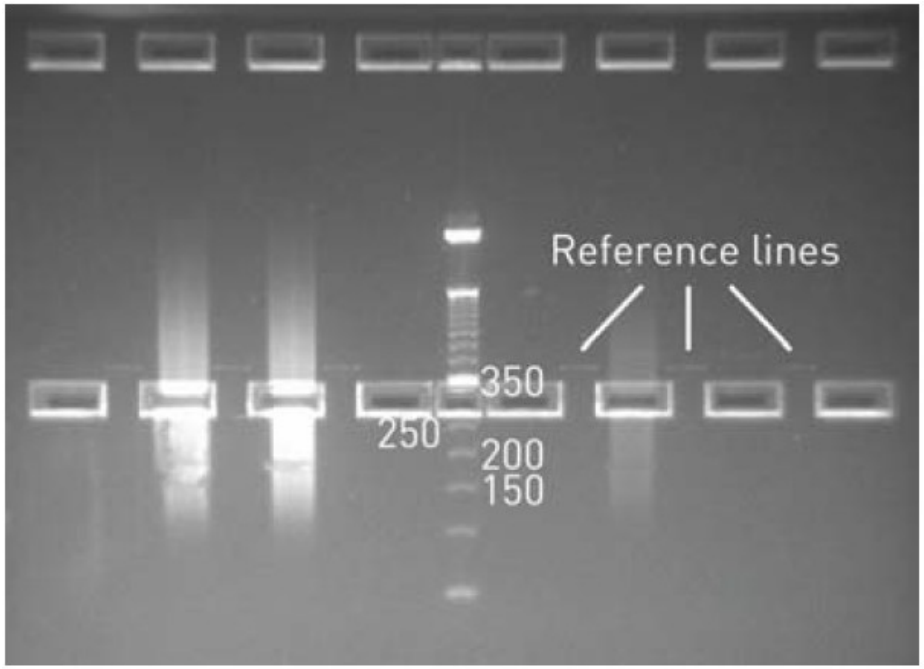


Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

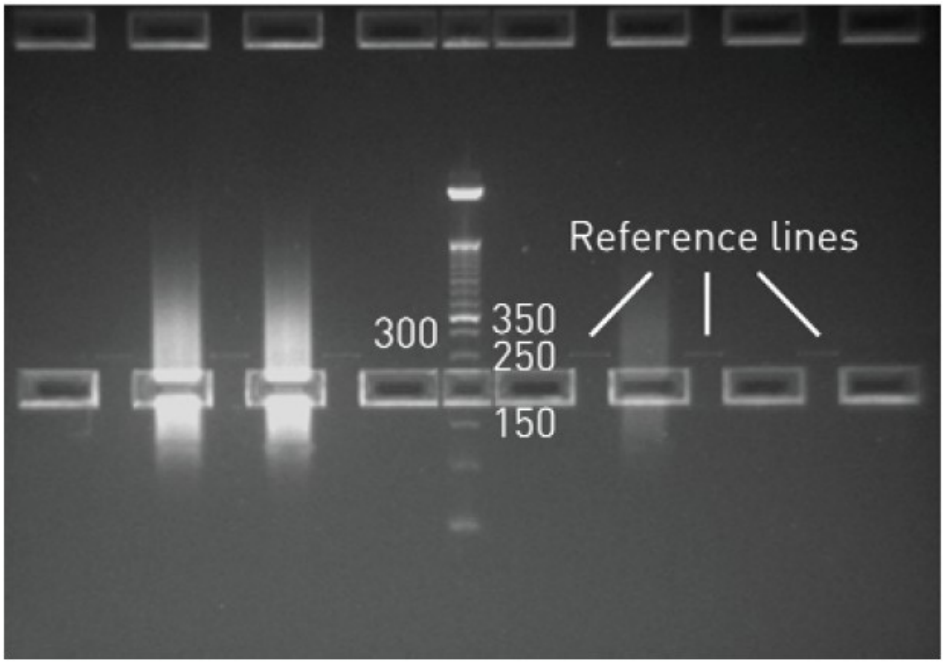
Next Generation Sequencing : Amplified Single Molecule Sequencing

Ion Torrent

Size selection - > maximizing sequencing yield



200 base-read library gel



100 base-read library gel

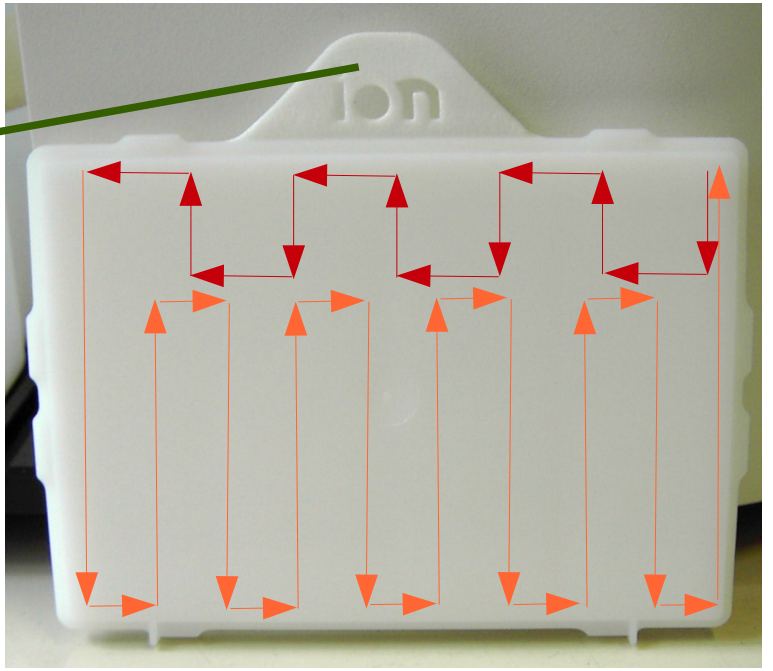
Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

Ion Torrent

Emulsion PCR : OneTouch Instrument









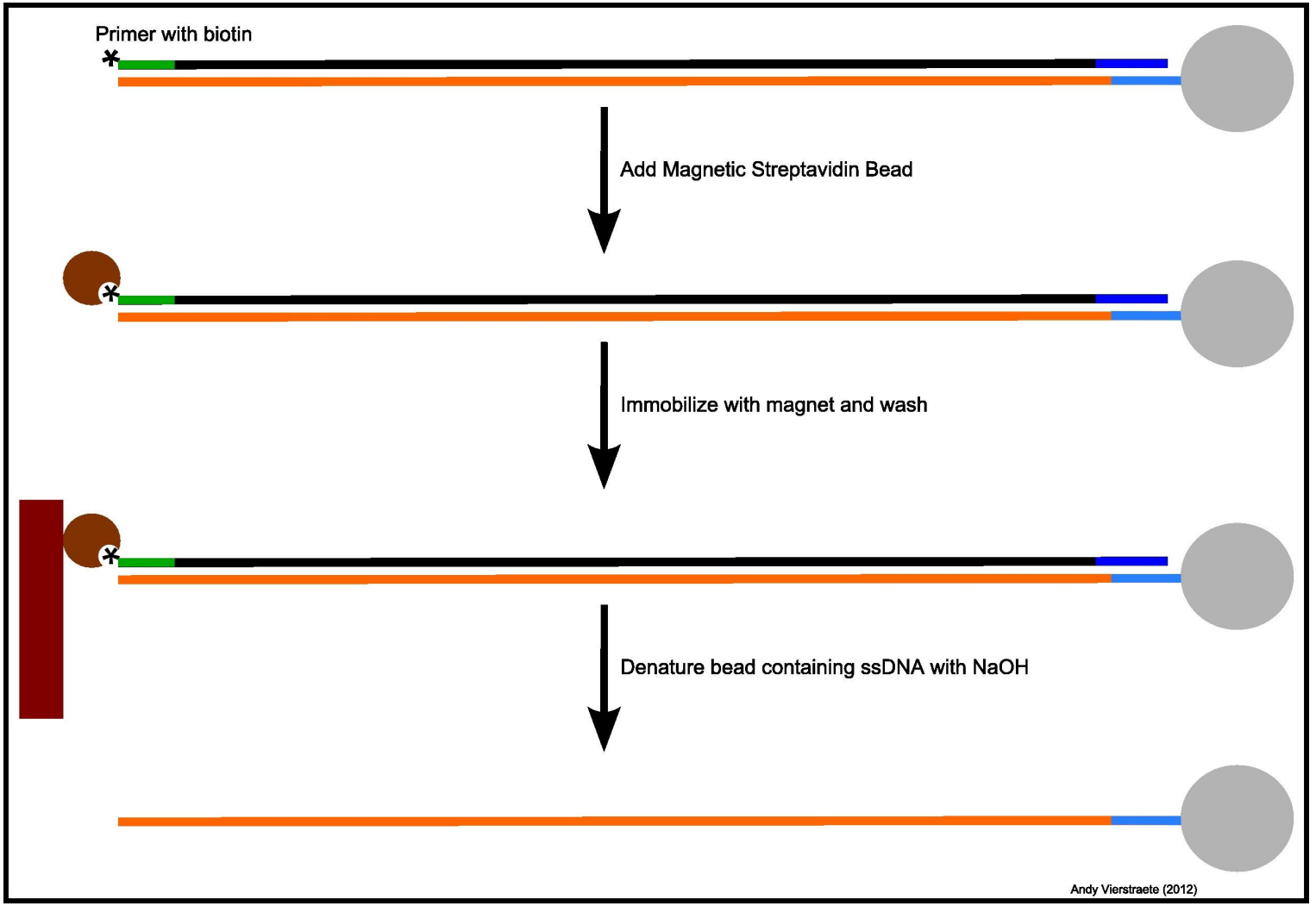
15 min hands-on; 4 hours amplification; 35 min enrichment

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

Ion Torrent

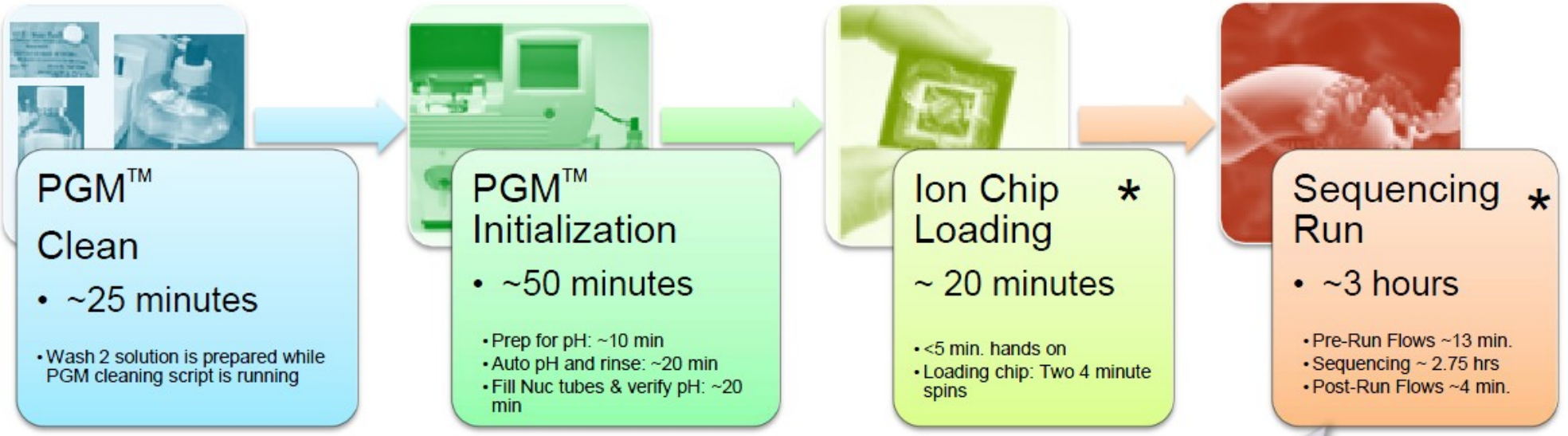
Enrichment : select only the beads that contain DNA
-> maximizing sequencing yield

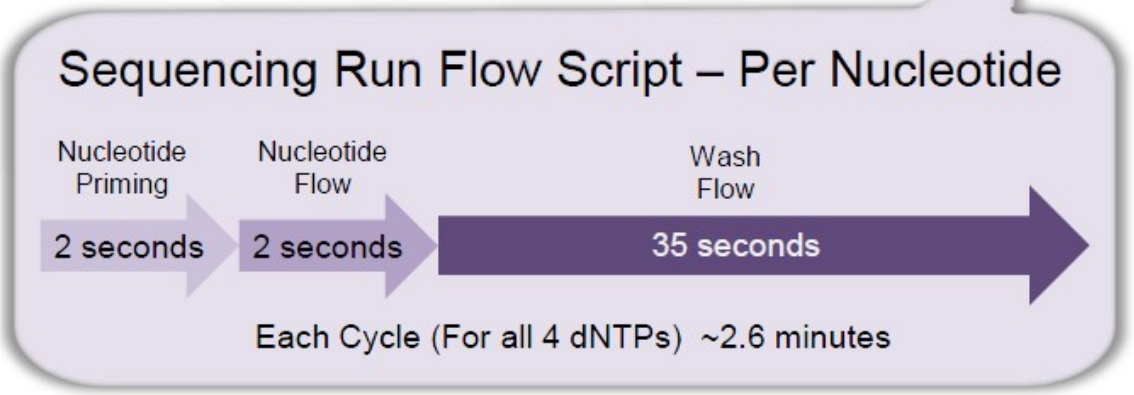


Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

Ion Torrent





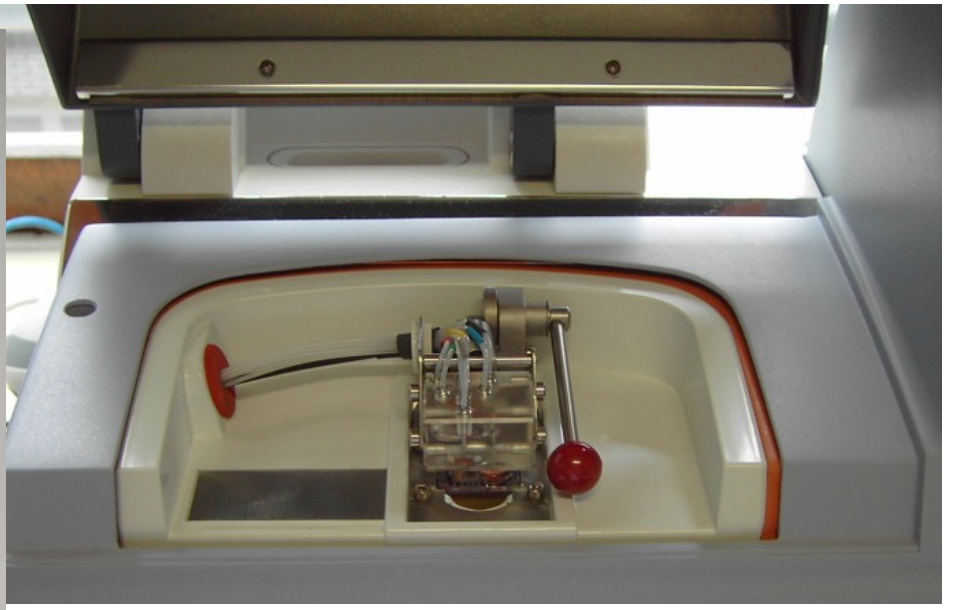
* Each initialized PGM™ Sequencer is capable of more than one sequencing run. The number of experiments possible per "Init" depends on the number of cycles selected (based on desired read length). To conduct subsequent sequencing experiments, repeat the loading and sequencing steps once the initial sequencing run is complete.

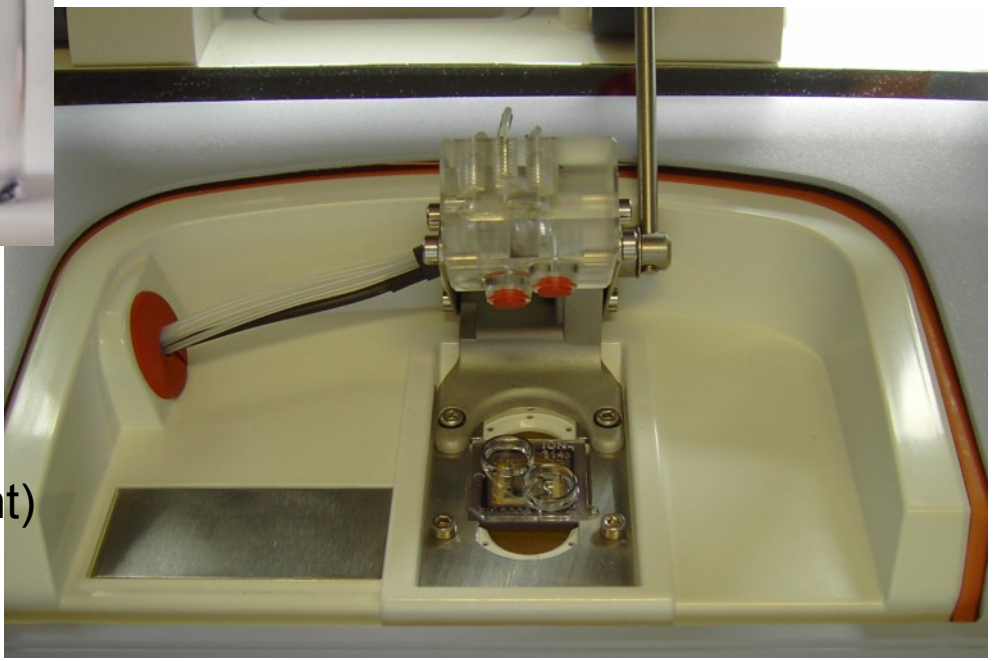
Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

Ion Torrent







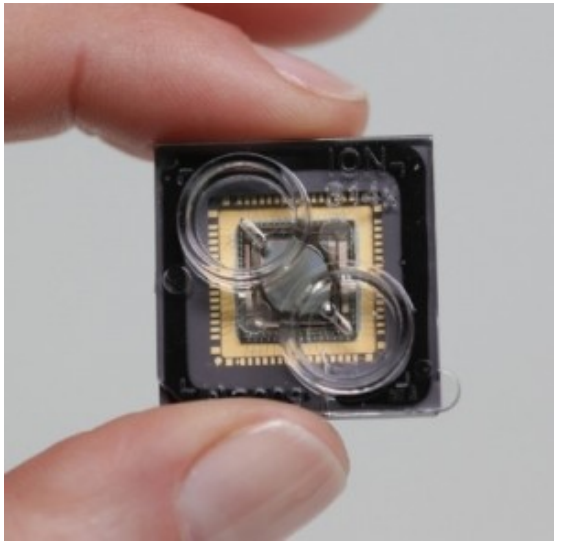
W1 bottle : 350µl 100 mM NaOH
W2 bottle : 2 liter W2 solution (contains detergent)
W3 bottle: 50 ml W3 (= buffer)
4 tubes with 20 µl of the different dNTP's

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

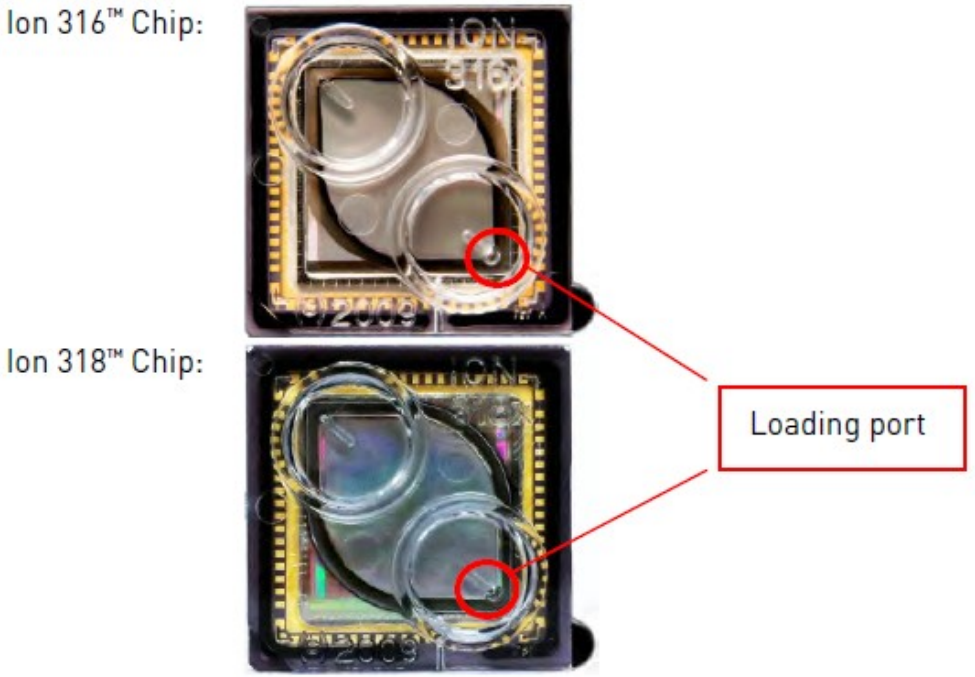
Next Generation Sequencing : Amplified Single Molecule Sequencing

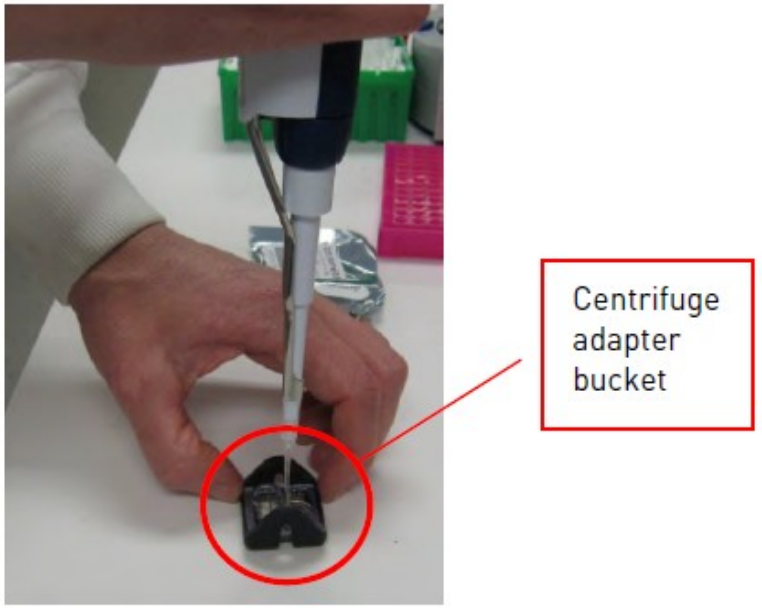
Ion Torrent

Chip loading :



- Add annealing buffer to the beads
- Add primer
- Denaturation : 2 minutes 95 °C
- Annealing : 2 minutes 37°C
- Chip check
- Add polymerase
- Load chip



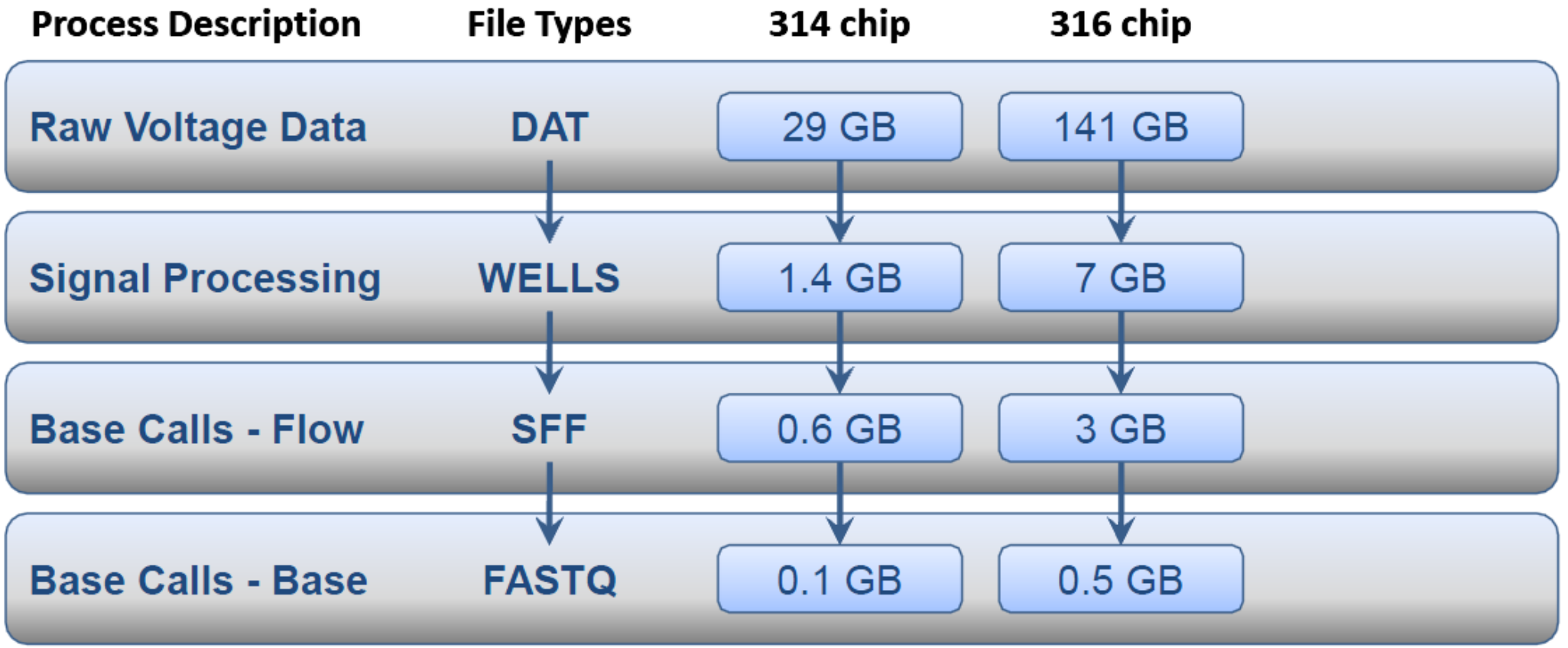


Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

Ion Torrent

Torrent Server pipeline



Different platforms

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

- 454 Sequencing / Roche
  - GS Junior System
  - GS FLX+ System
- Illumina (Solexa)
  - HiSeq System
  - Genome analyzer IIx
  - MySeq
- Applied Biosystems - Life Technologies
  - SOLiD 5500 System
  - SOLiD 5500xl System
- Ion Torrent - Life Technologies
  - Personal Genome Machine (PGM)
  - Proton
- Helicos
  - Helicos Genetic Analysis System
- Pacific Biosciences
  - PacBio RS
- Oxford Nanopore Technologies
  - GridION System
  - MinION

Next Generation Sequencing
Amplified Single Molecule Sequencing

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

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Third Generation Sequencing : Single Molecule Sequencing

Helicos (BioSciences Corporation)

Helicos Genetic Analysis System



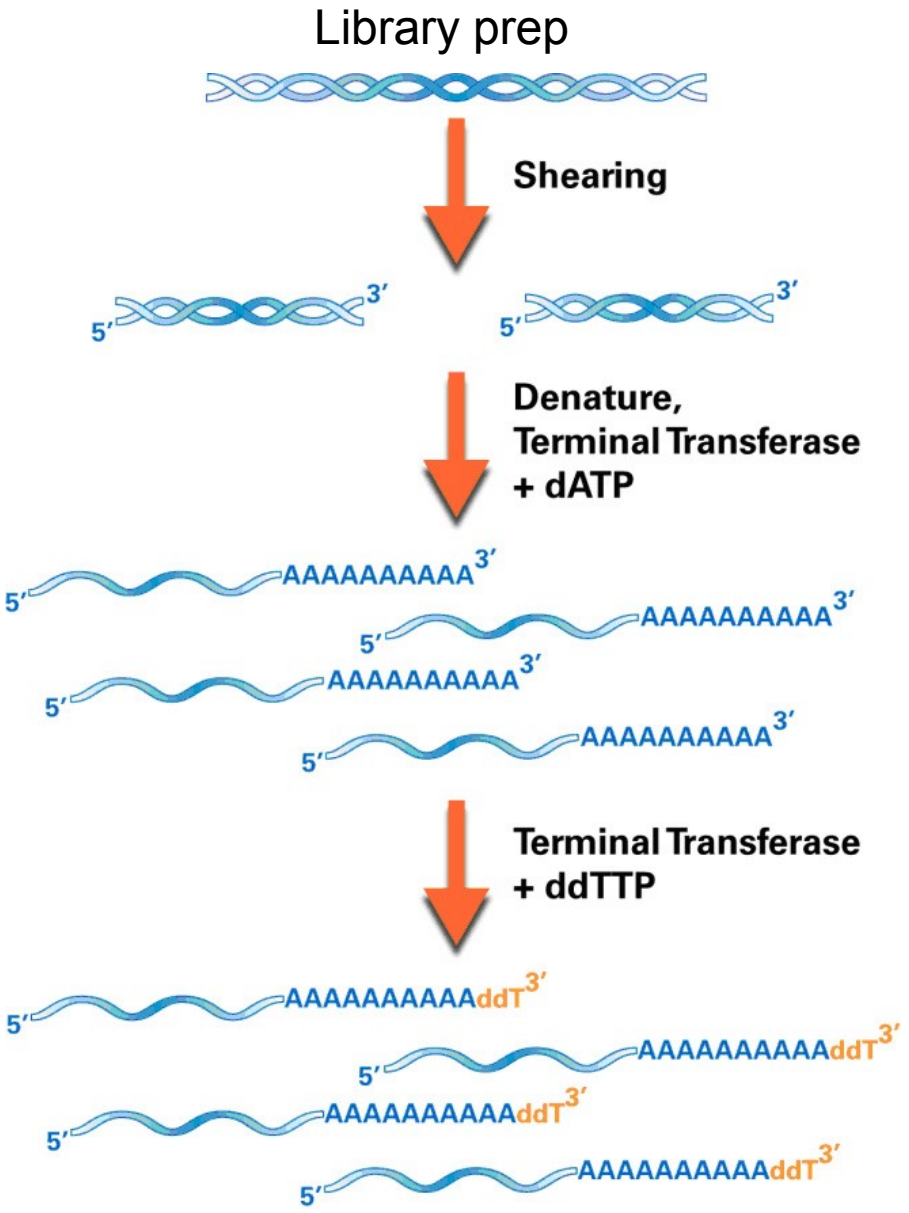
	Helicos
Read Length	35 bp
Throughput	35 Gb
Reads per run	600,000,000 - 1,000,000,000
Accuracy	97 %
Run Time	8 days

Workflow : Library preparation ⟶ Sequencing

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Third Generation Sequencing : Single Molecule Sequencing

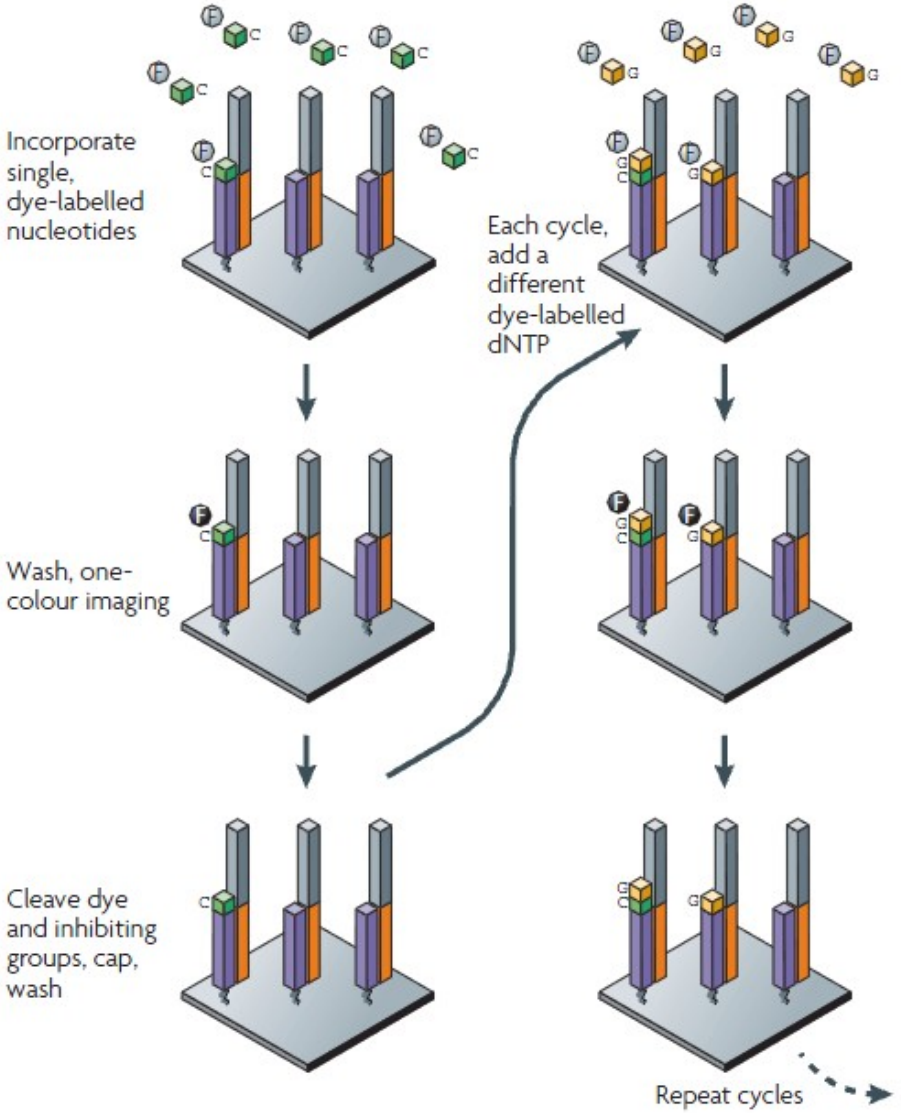
Helicos

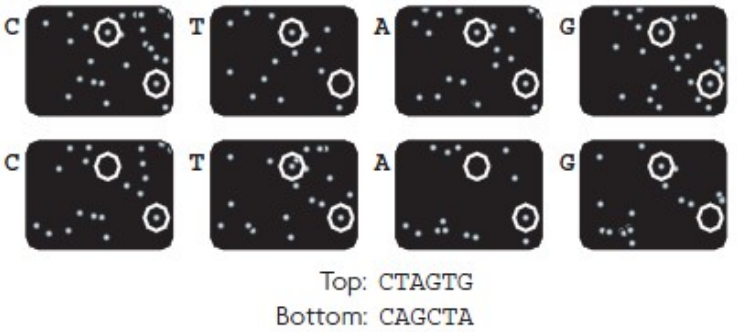


Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Third Generation Sequencing : Single Molecule Sequencing

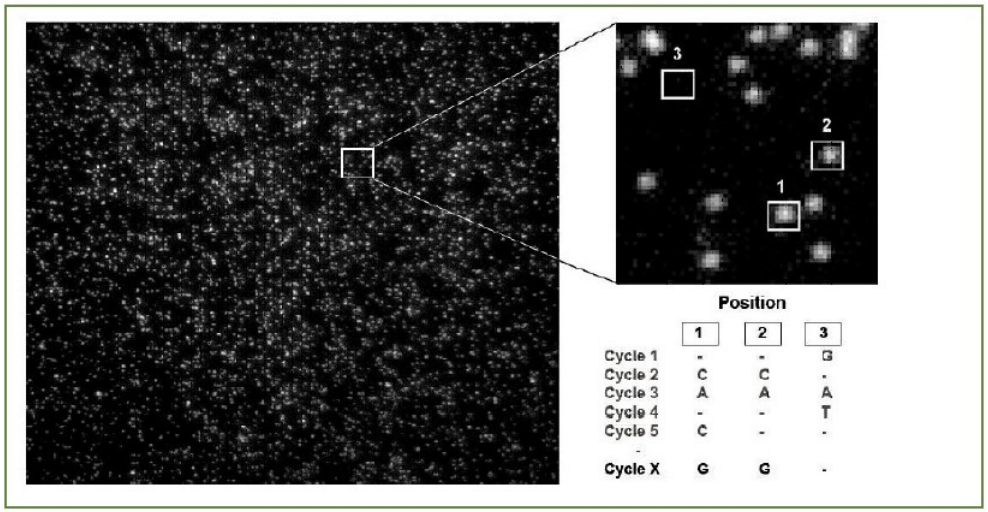
Helicos





Nucleotides flown sequentially

(Dark nucleotides : incorporation not detected)



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Third Generation Sequencing : Single Molecule Sequencing

Helicos

Advantages Single molecule sequencing :

- Less sample preparation (no PCR)
- No amplification
  - -> no PCR errors
  - -> fewer contamination issues
  - -> no GC-bias
  - -> analyze every sample (unPCRable / unclonable)
  - -> analyze low quality DNA (museum, archeological, forensics samples)
- Absolute quantification
- Sequence RNA directly

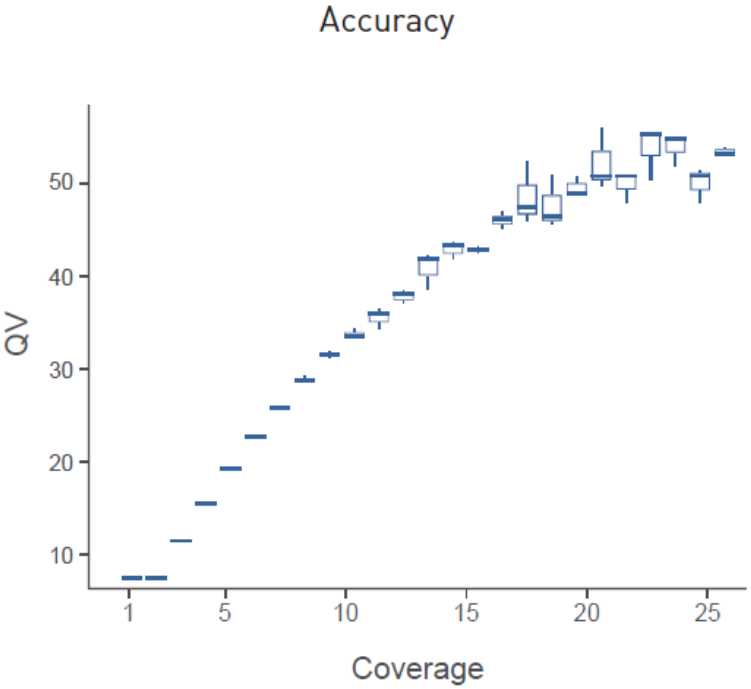
Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Third Generation Sequencing : Single Molecule Sequencing

Pacific Biosciences

Pacbio RS





	Pacbio RS
Read Length	3000 - 15,000 bp
Throughput	1 Gb
Reads per run	70,000
Accuracy	95 %
Run Time	30 minutes

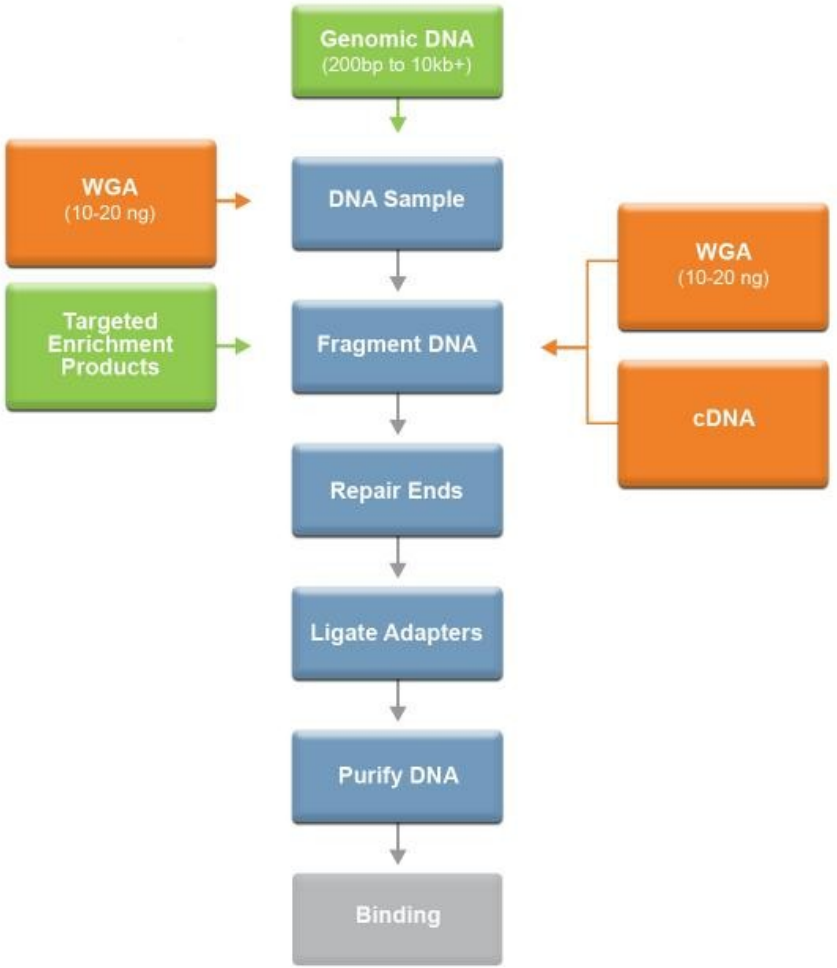
Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Third Generation Sequencing : Single Molecule Sequencing

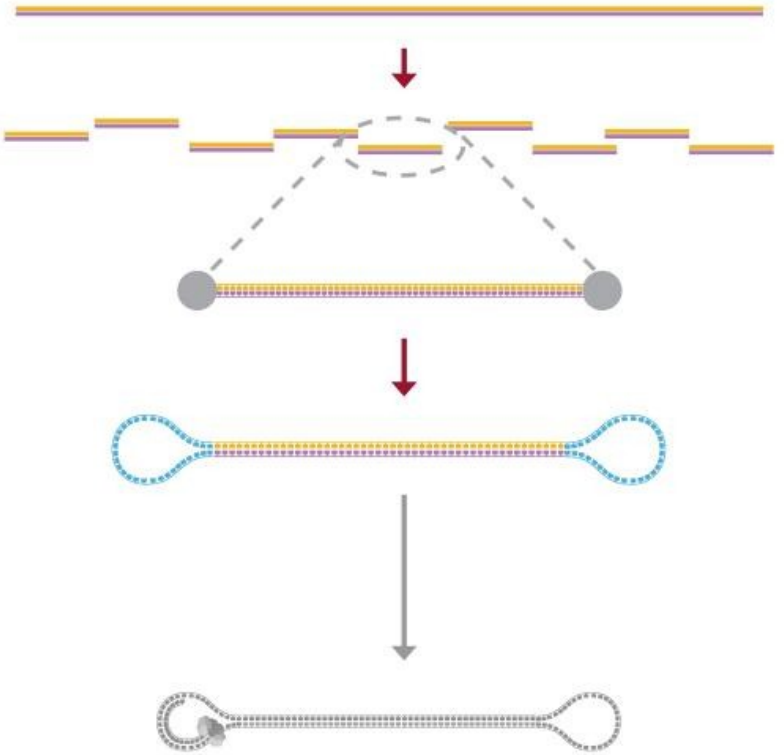
Pacific Biosciences

Workflow : Library preparation → Sequencing

Sample Preparation



Building of SMRTbell

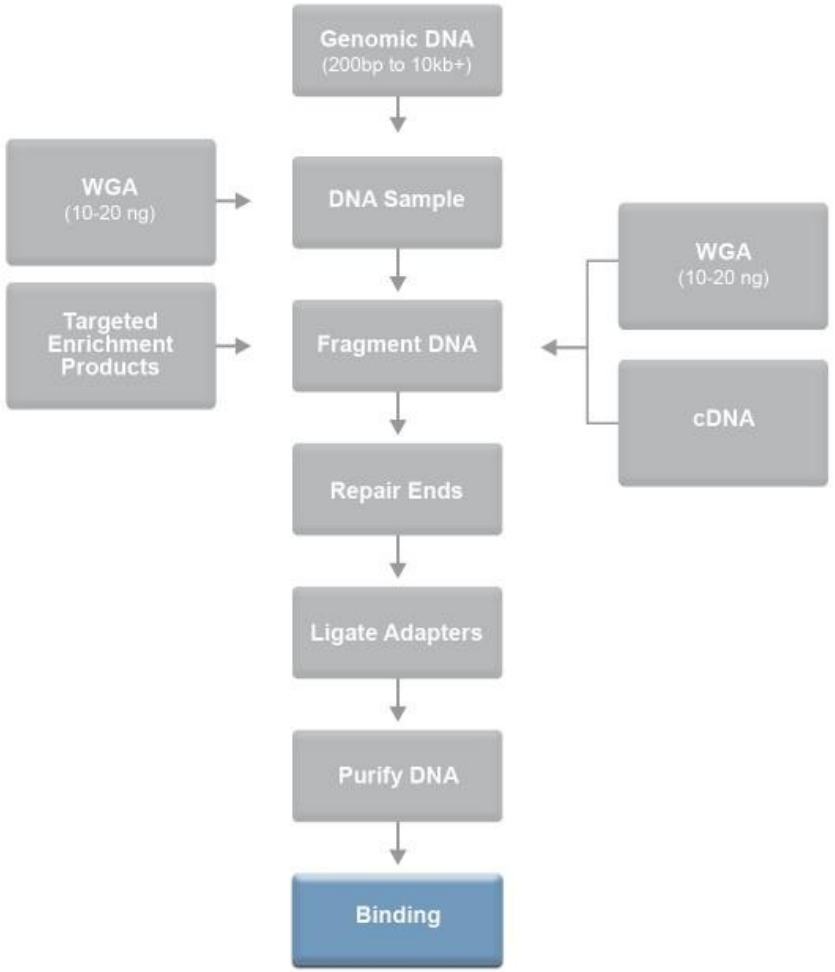


Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

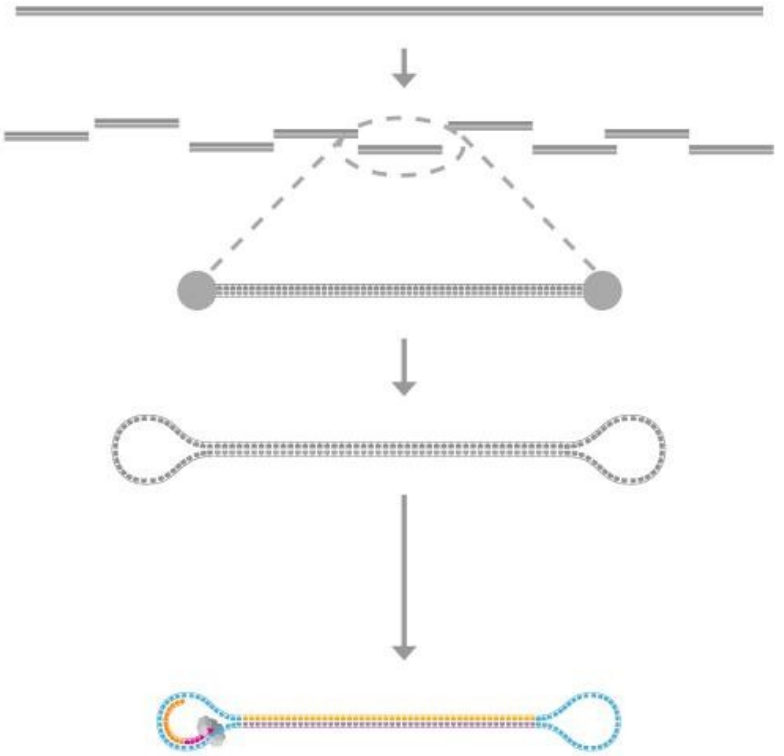
Third Generation Sequencing : Single Molecule Sequencing

Pacific Biosciences

Sample Preparation



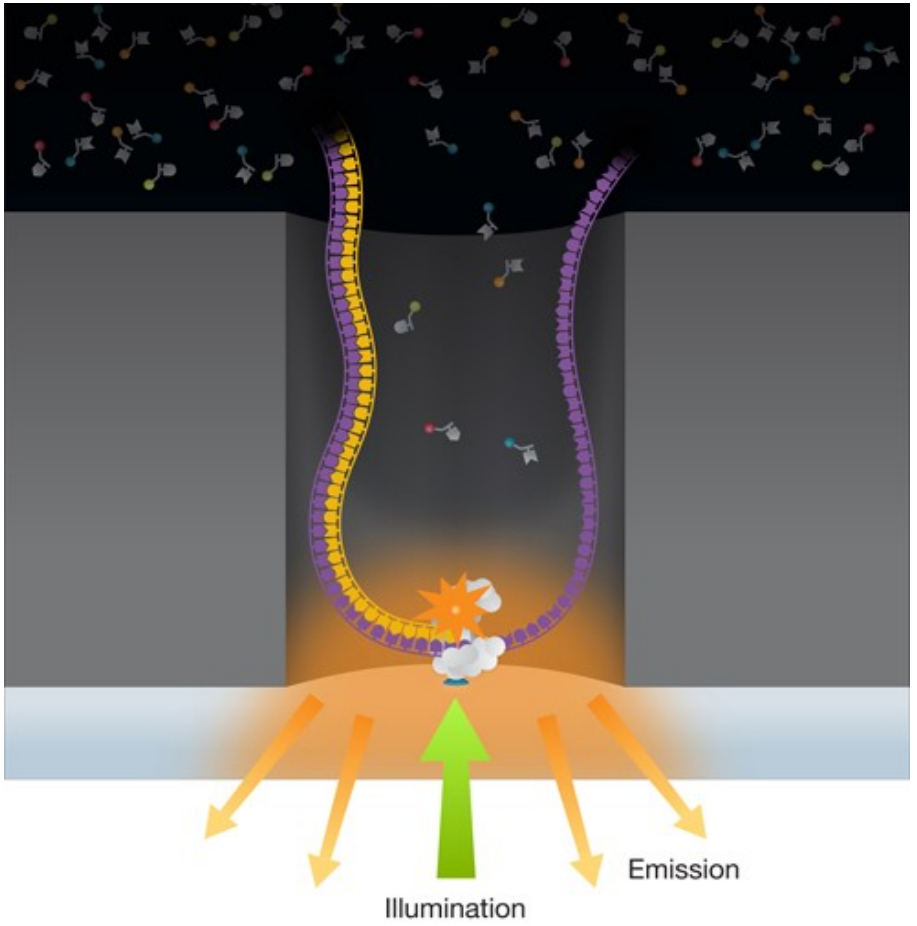
Building of SMRTbell



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Third Generation Sequencing : Single Molecule Sequencing

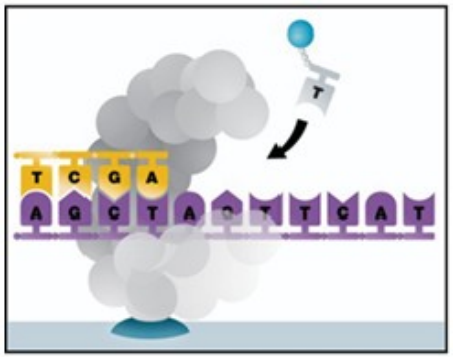
Pacific Biosciences

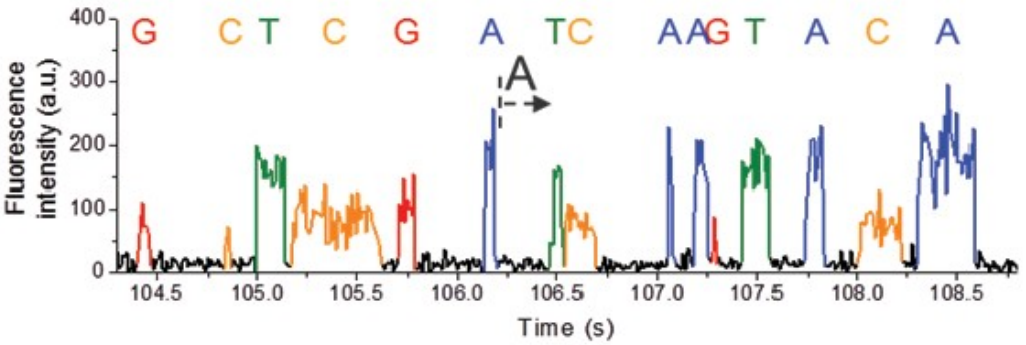


4 nucleotides with different fluorescent dye simultaneous present

2-3 nucleotides/sec
2-3 Kb (up to 50) read length
6 TB data in 30 minutes

laser damages polymerase





Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Third Generation Sequencing : Single Molecule Sequencing

Pacific Biosciences

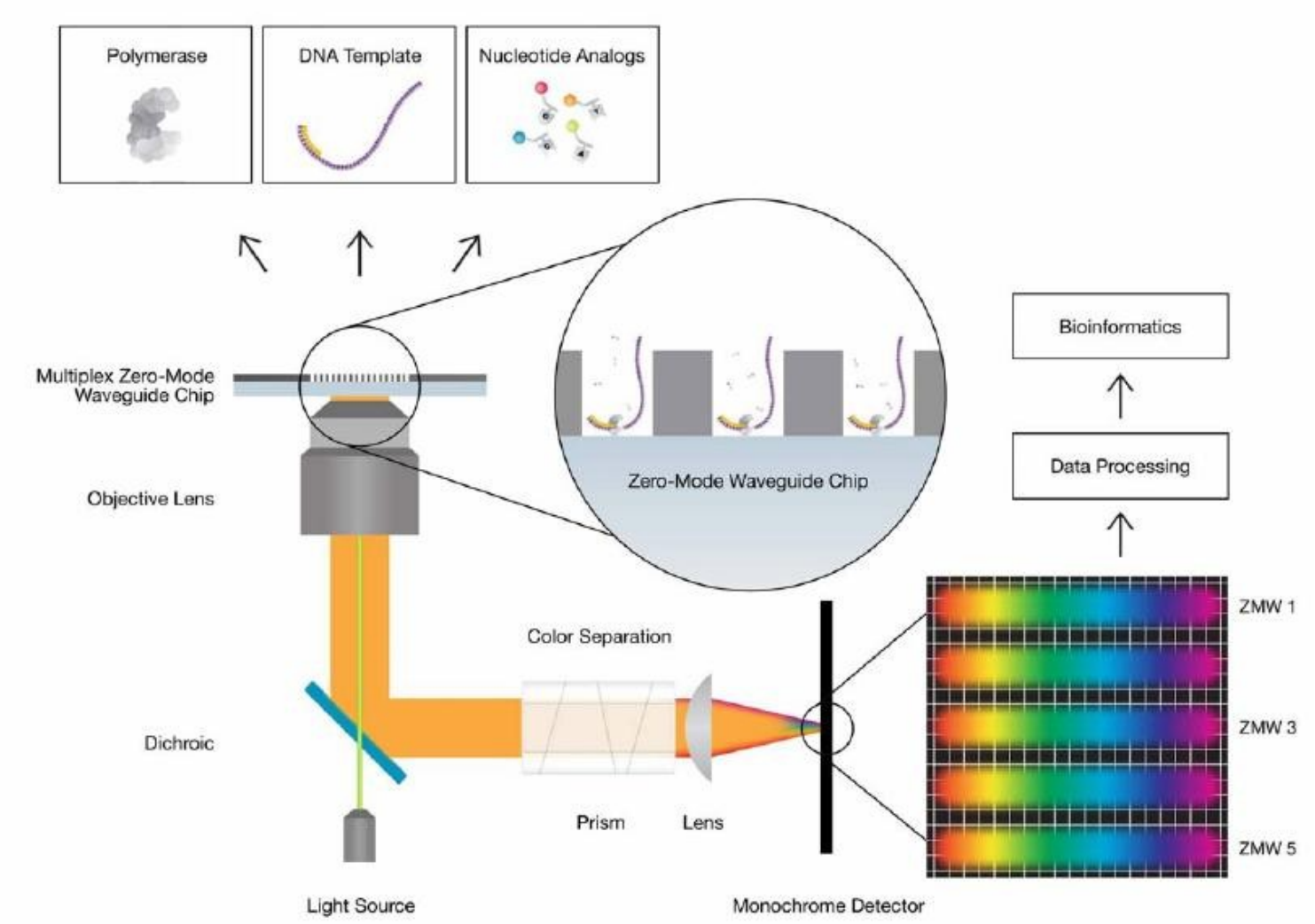


Figure 13. Highly parallel optics system.

The detected flash of light is separated into a spatial array, from which the identity of the incorporated base is determined.

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Third Generation Sequencing : Single Molecule Sequencing

Oxford Nanopore

Single use cartridge



GridION system



MinION

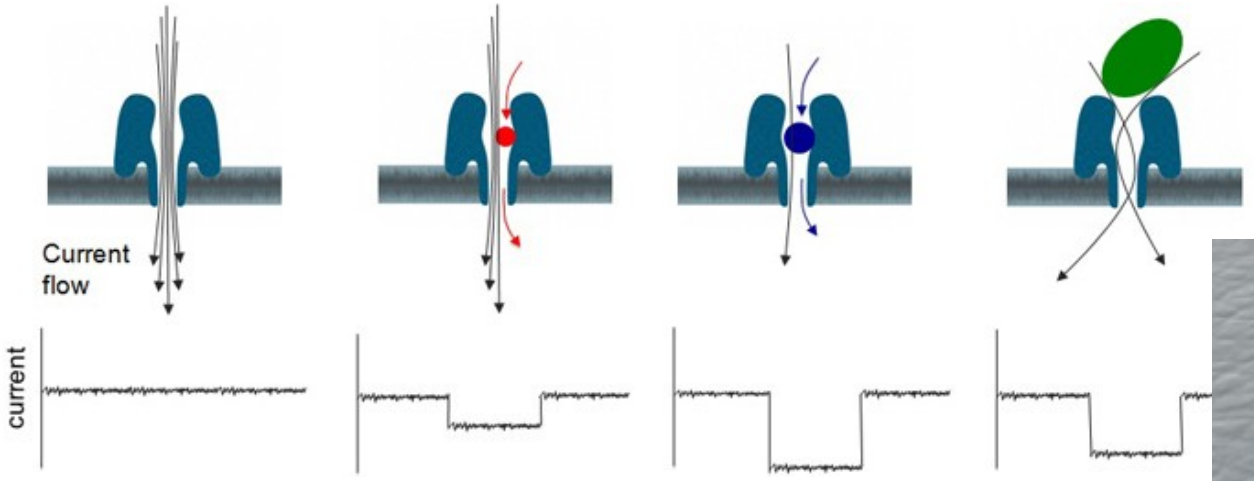


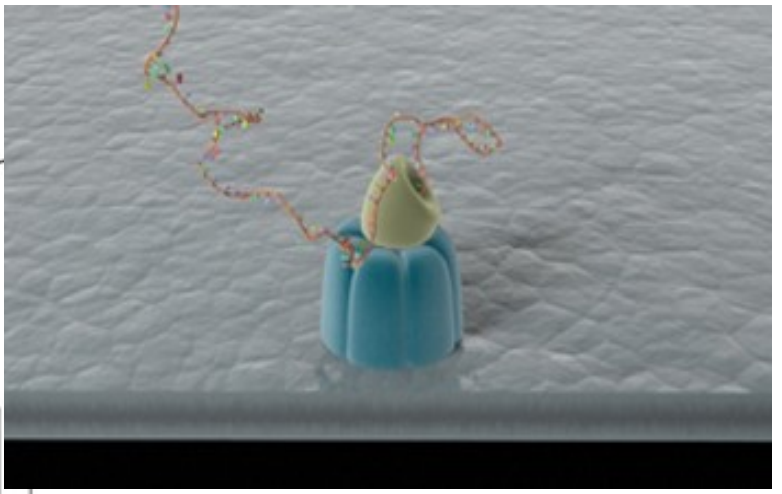
	Nanopore
Read Length	48 kb ?
Throughput	? Gb
Reads per run	2000
Accuracy	75 %
Run Time	? minutes

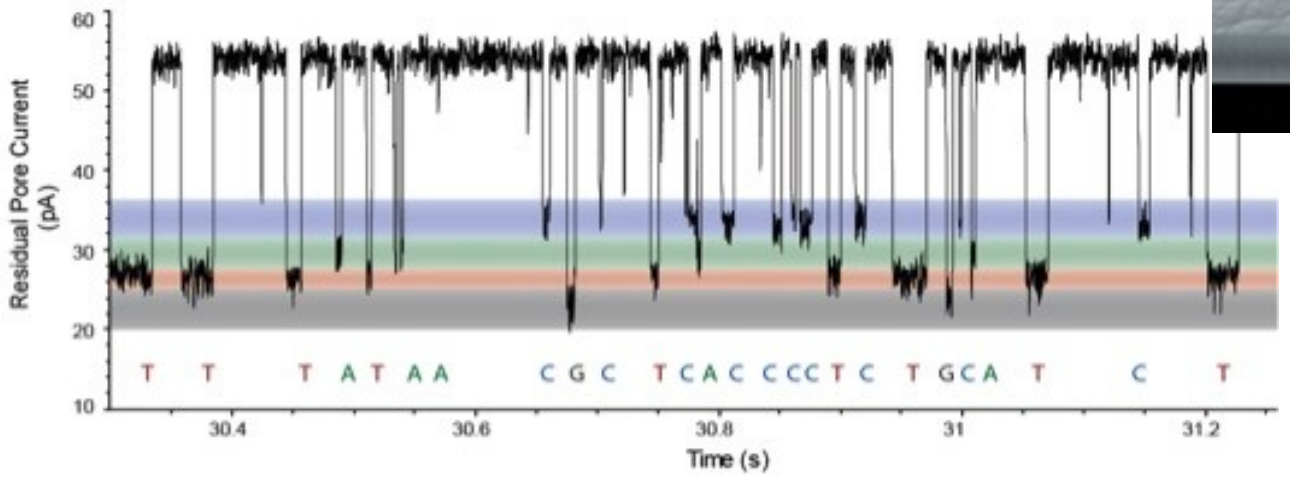
Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Third Generation Sequencing : Single Molecule Sequencing

Oxford Nanopore



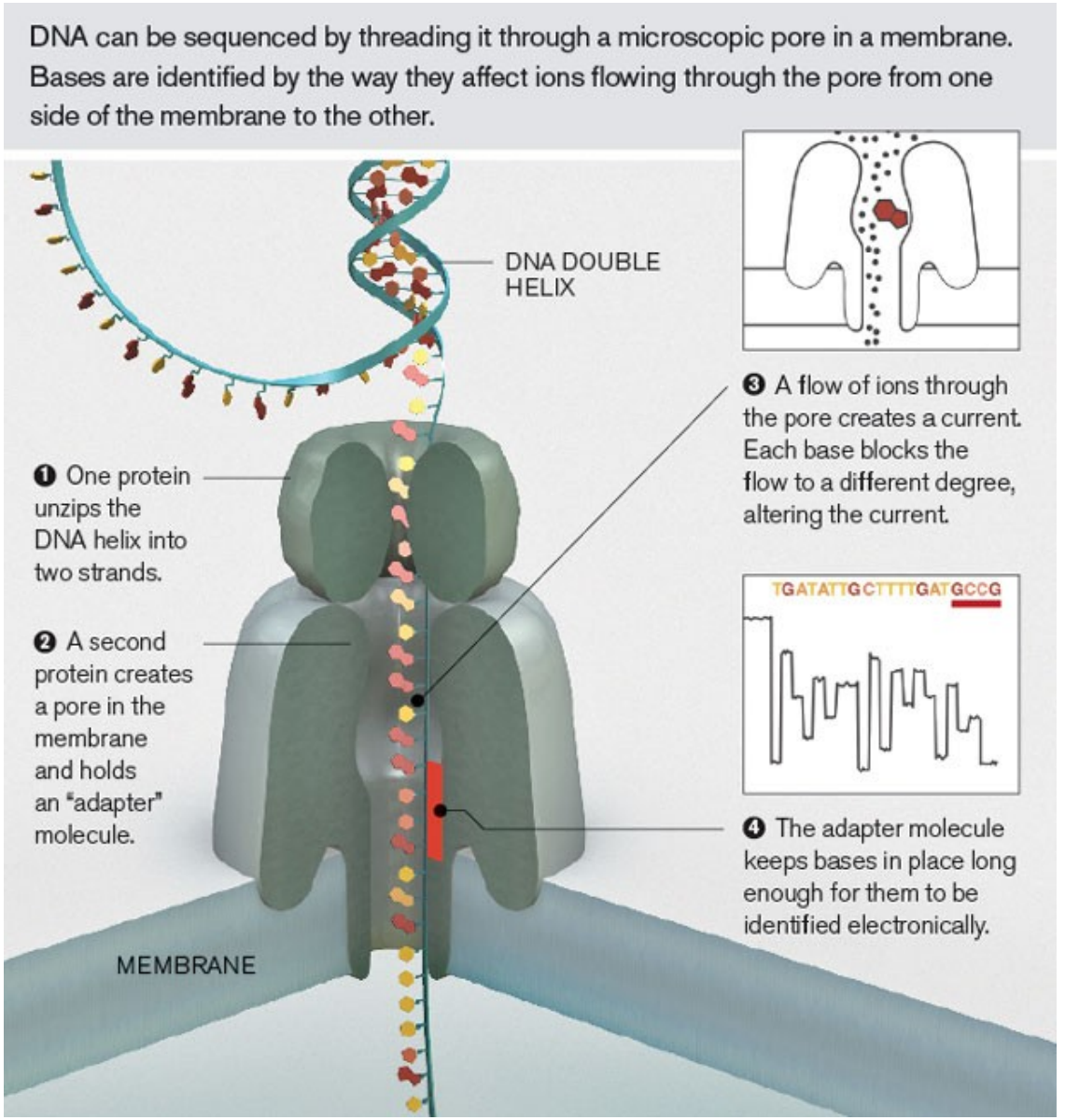




Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Third Generation Sequencing : Single Molecule Sequencing

Oxford Nanopore



Different platforms

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

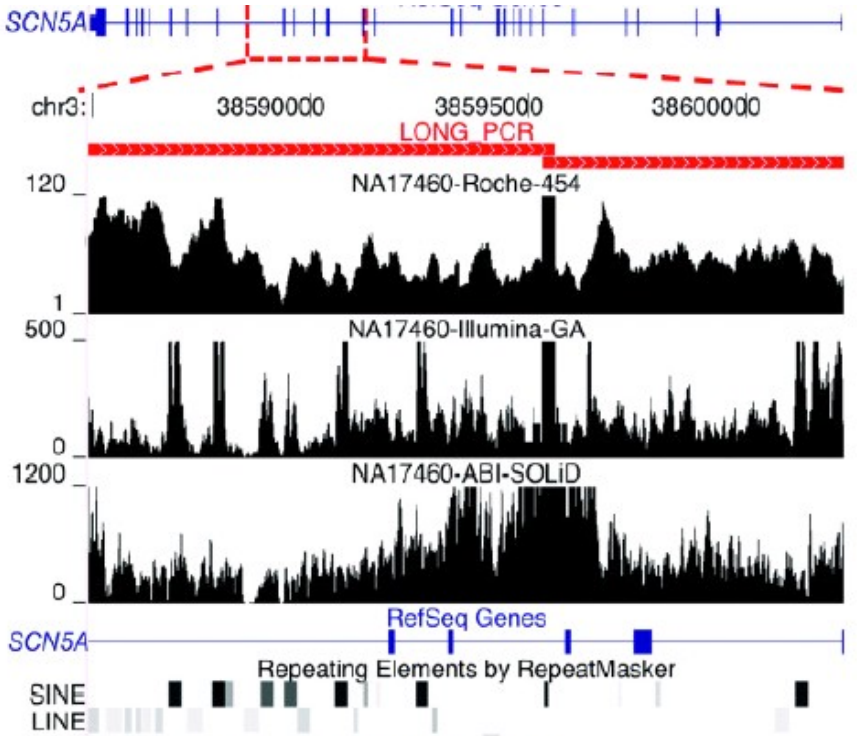
Which Next Generation Sequencer to choose for your project ?

	Capacity	Speed	Read Length	Homopolymers	Cost/run	Amplification
454 Roche	35-700 Mb	10-23 hours	400-700 bp	-	5.000 €	Yes
SOLiD	90-180 Gb	7-12 days	75 bp	+	5.000 €	Yes
Illumina	6-600 Gb	2-14 days	100-250 bp	+	10.000-20.000 €	Yes
Ion Torrent	20 Mb- 1Gb	4,5 hours	200 bp	-	1.000-2.000 €	Yes
Helicos	35 Gb	8 days	35 bp	+	20.000 €	No
PacBio	1Gb	30 minutes	3000 bp	+	600-800 €	No

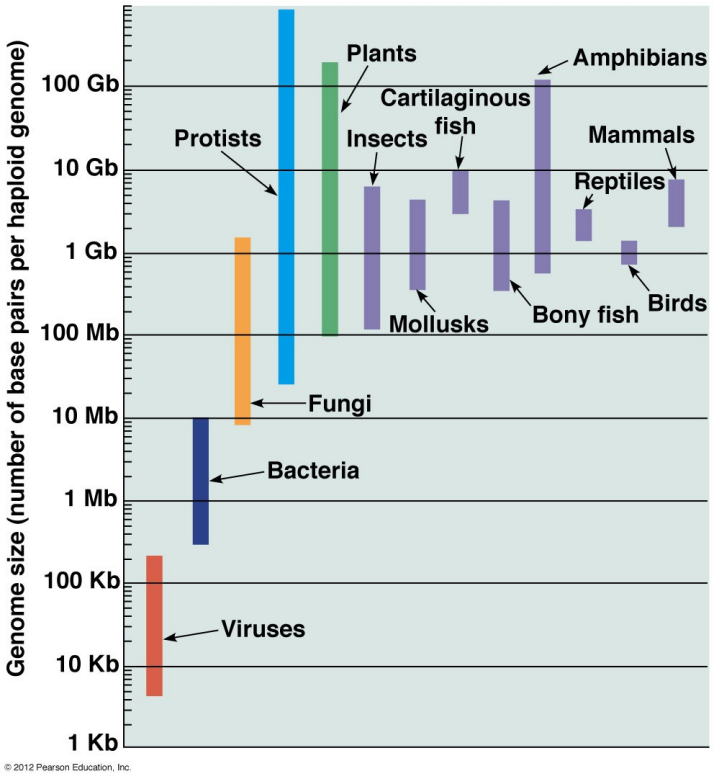
Roche →

Illumina →

SOLID →



Harismendy et al., Genome Biol. 2009



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012



Applications

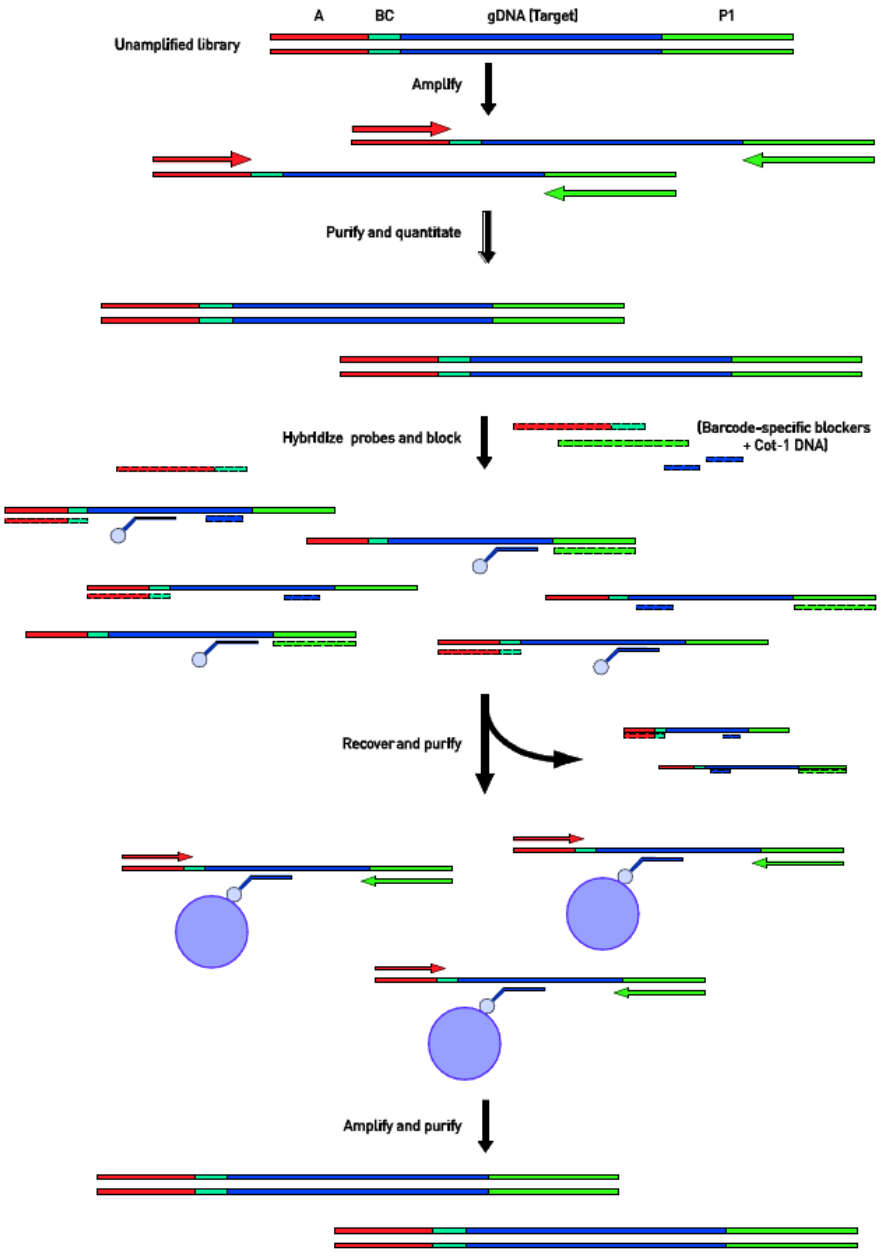
Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

- Genome Sequencing
  - De novo sequencing
  - Resequencing
  - Targeted (re)sequencing
  - Mitochondrial sequencing
  - Mutation detection
  - Amplicon sequencing
  - Amplicon Cancer Panel
  - ...
- Transcript Expression Profiling
  - RNA sequencing
  - miRNA sequencing
  - Deep-SAGE
  - Deep-CAGE
  - PAS : polyadenylation site
  - ...
- Transcription factor binding
  - ChIP sequencing
- Structural variation
- Metagenomics
- ...

Applications

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

- Genome Sequencing
  - De novo sequencing
  - Resequencing
  - Targeted (re)sequencing
  - Mitochondrial sequencing
  - Mutation detection
  - Amplicon sequencing
  - Amplicon Cancer Panel
  - ...
- Transcript Expression Profiling
  - RNA sequencing
  - miRNA sequencing
  - Deep-SAGE
  - Deep-CAGE
  - PAS : polyadenylation site
  - ...
- Transcription factor binding
  - ChIP sequencing
- Structural variation
- Metagenomics
- ...

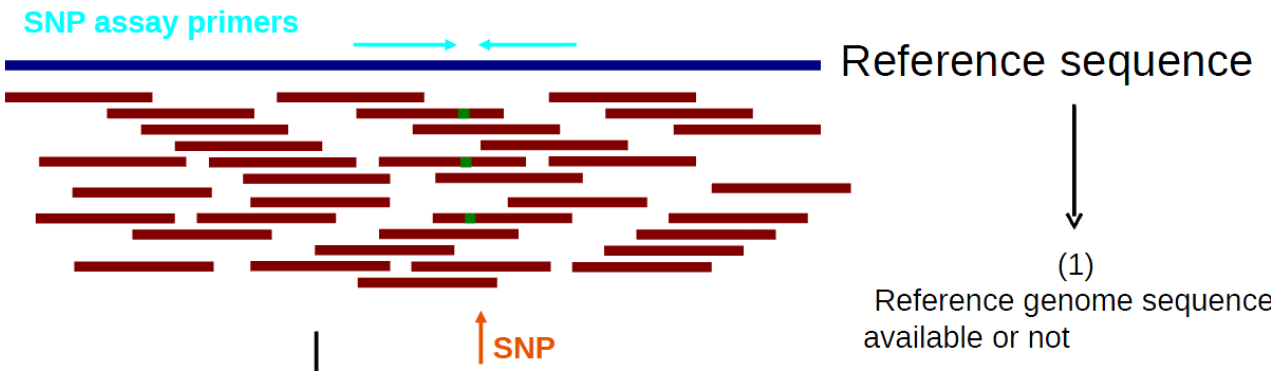


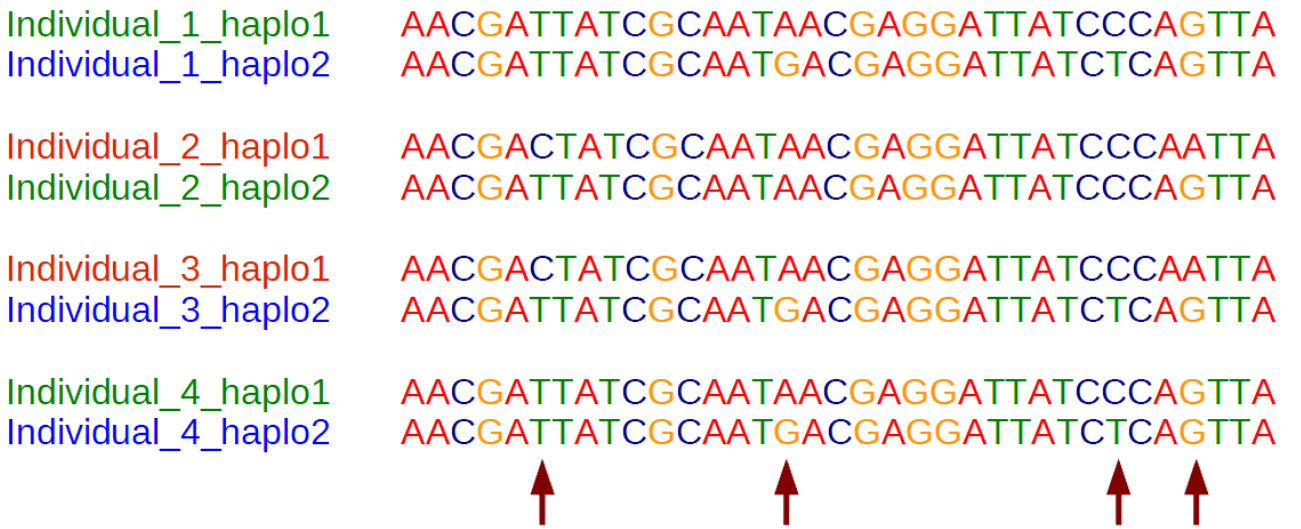
Applications

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

- Genome Sequencing
  - De novo sequencing
  - Resequencing
  - Targeted (re)sequencing
  - Mitochondrial sequencing
  - Mutation detection
  - Amplicon sequencing
  - Amplicon Cancer Panel
  - ...
- Transcript Expression Profiling
  - RNA sequencing
  - miRNA sequencing
  - Deep-SAGE
  - Deep-CAGE
  - PAS : polyadenylation site
  - ...
- Transcription factor binding
  - ChIP sequencing
- Structural variation
- Metagenomics
- ...

SNP identification: Principle

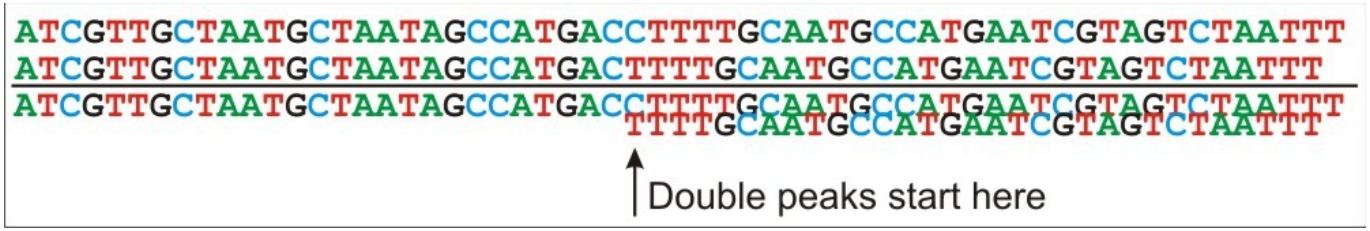




Applications

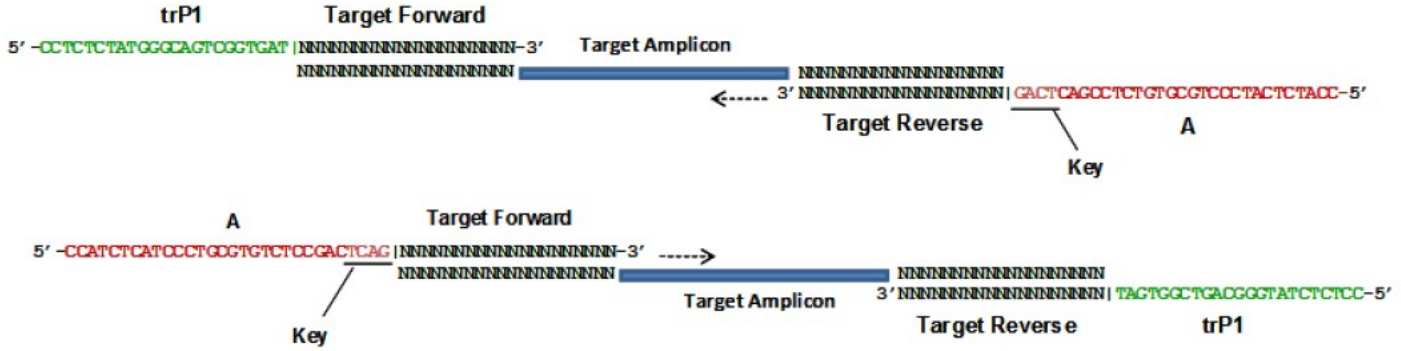
Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

- Genome Sequencing
  - De novo sequencing
  - Resequencing
  - Targeted (re)sequencing
  - Mitochondrial sequencing
  - Mutation detection
  - Amplicon sequencing (eg for heterozygote genes : no cloning necessary)
  - Amplicon Cancer Panel
  - ...
- Transcript Expression Profilin
  - RNA sequencing
  - miRNA sequencing
  - Deep-SAGE
  - Deep-CAGE
  - PAS : polyadenylation site
  - ...
- Transcription factor binding
  - ChIP sequencing
- Structural variation
- Metagenomics
- ...





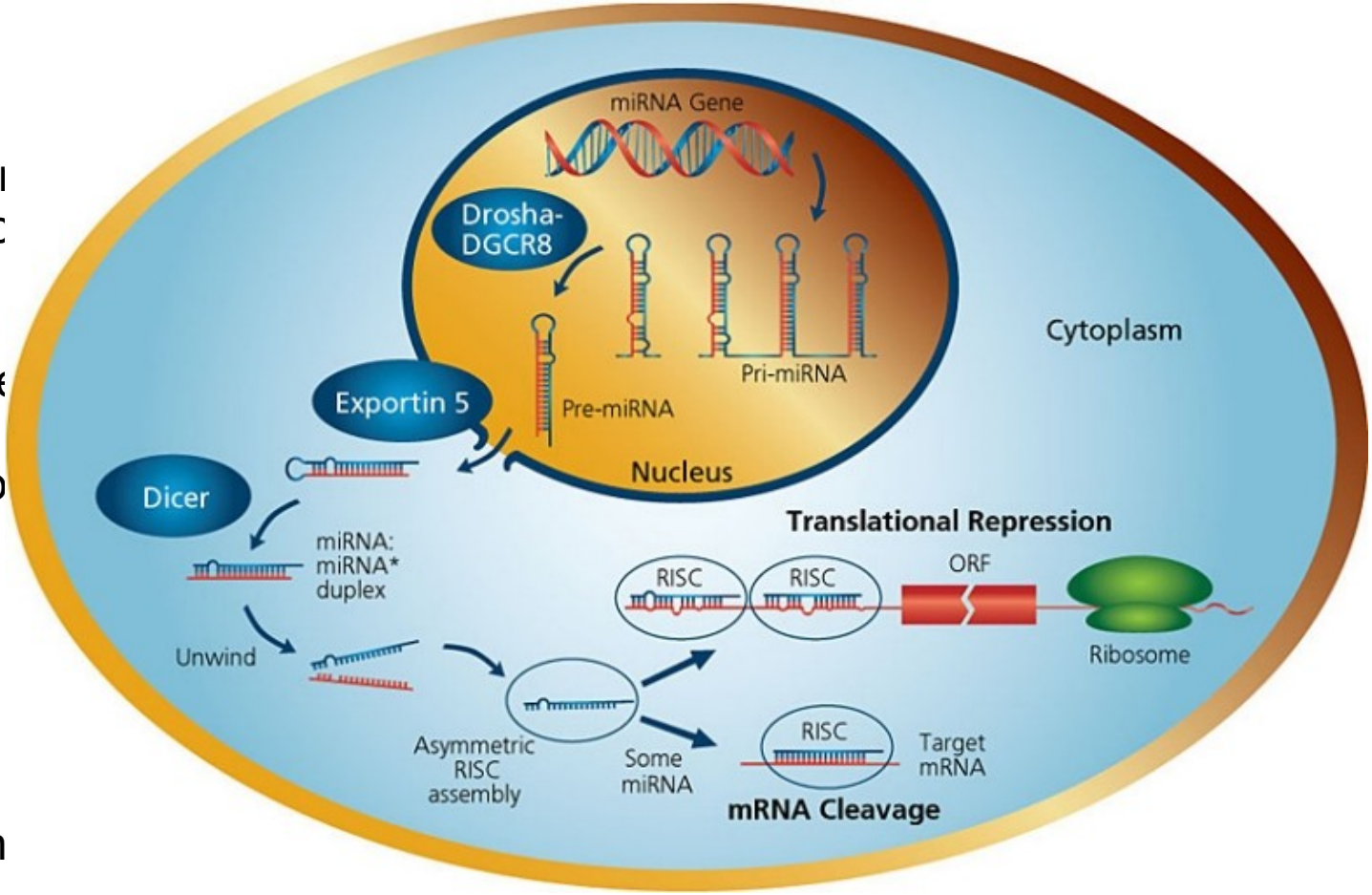
Fusion method : pcr fragments of 250-300 bp (IonTorrent)



Applications

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

- Genome Sequencing
  - De novo sequencing
  - Resequencing
  - Targeted (re)sequenci
  - Mitochondrial sequenc
  - Mutation detection
  - Amplicon sequencing
  - Amplicon Cancer Pane
  - ...
- Transcript Expression Pro
  - RNA sequencing
  - miRNA sequencing
  - Deep-SAGE
  - Deep-CAGE
  - PAS : polyadenylation
  - ...
- Transcription factor bindin
  - ChIP sequencing
- Structural variation
- Metagenomics
- ...

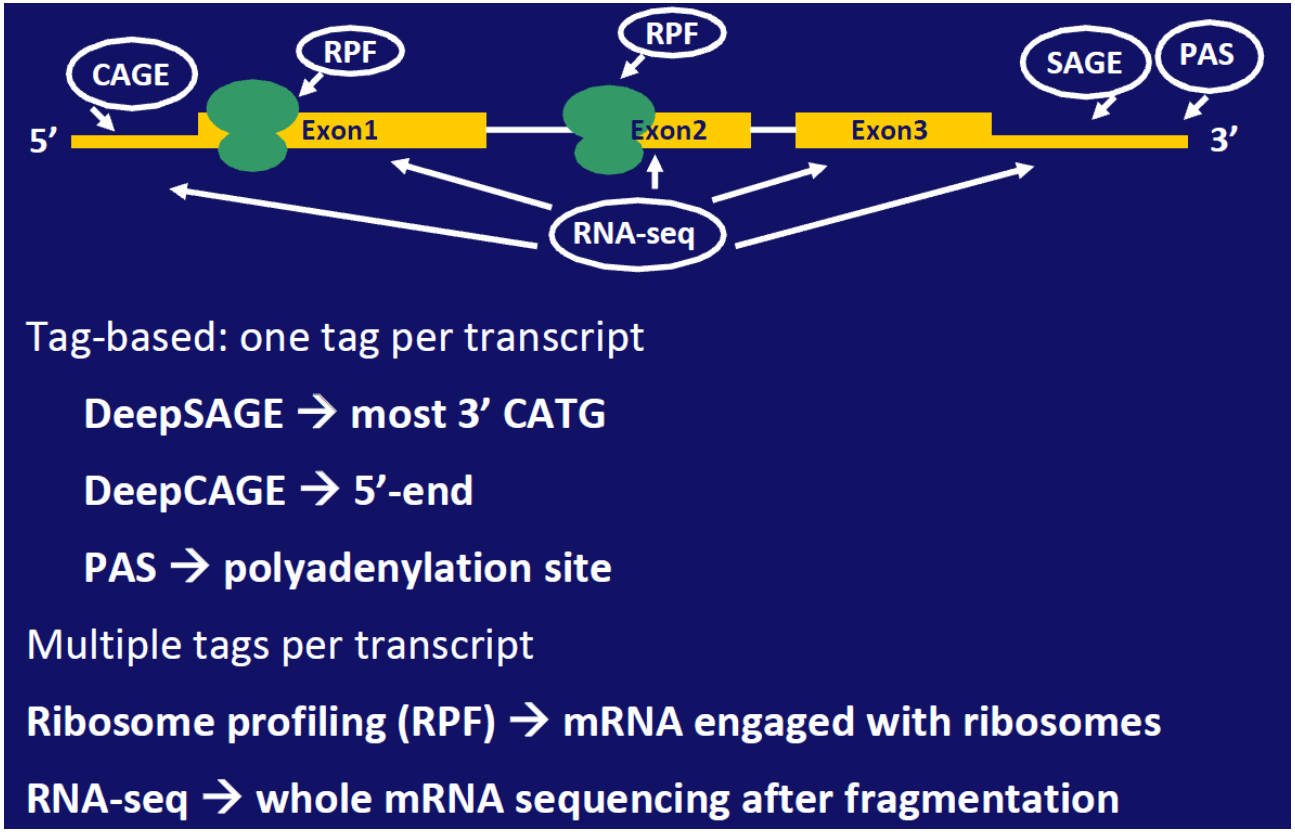


microRNA is approx 21 bases long

Applications

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

- Genome Sequencing
  - De novo sequencing
  - Resequencing
  - Targeted (re)sequencing
  - Mitochondrial sequencing
  - Mutation detection
  - Amplicon sequencing
  - Amplicon Cancer Panel
  - ...
- Transcript Expression Profiling
  - RNA sequencing
  - miRNA sequencing
  - Deep-SAGE
  - Deep-CAGE
  - PAS : polyadenylation site
  - ...
- Transcription factor binding
  - ChIP sequencing
- Structural variation
- Metagenomics
- ...



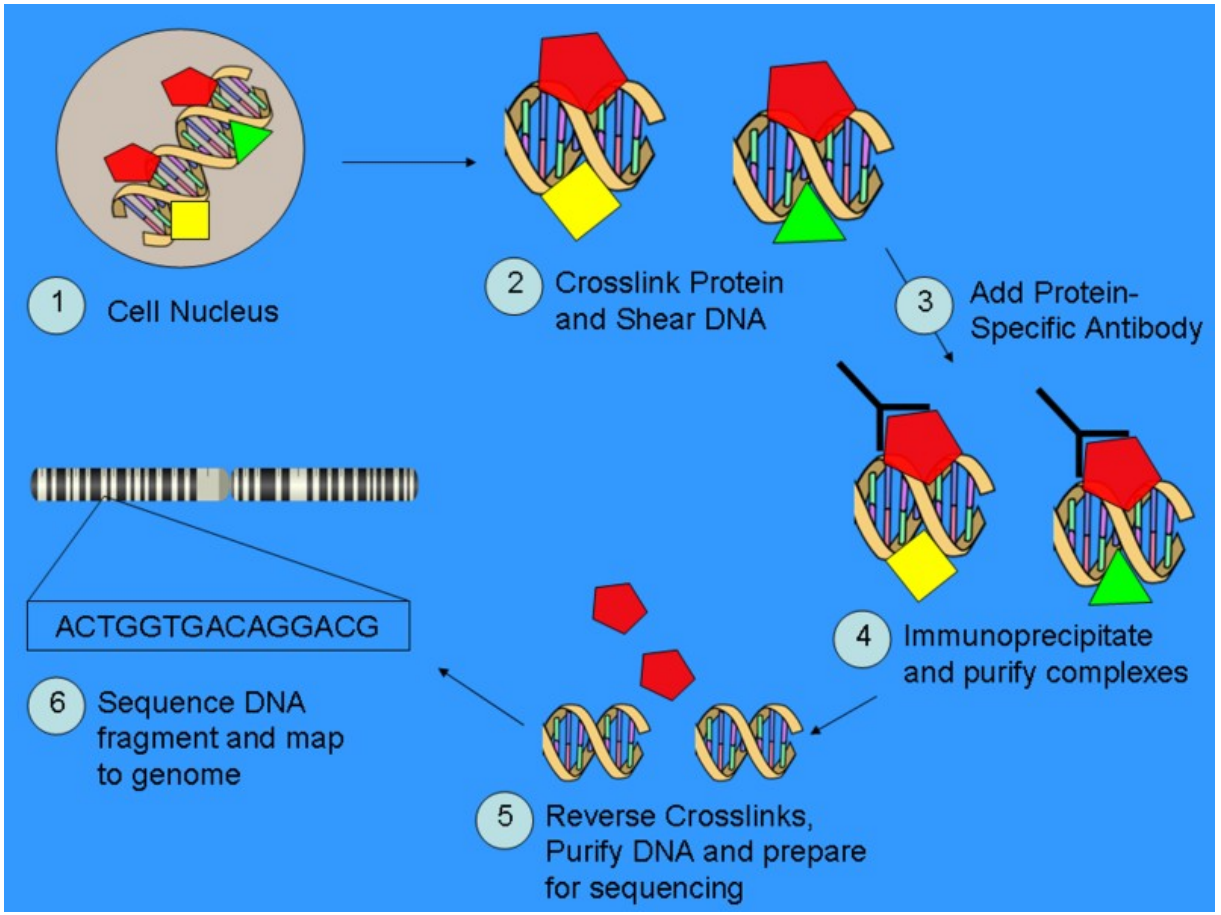
Sequence 17 bp of the transcription start or stop site
-> discovery of new start and stop sites,
antisense transcription possibilities

Applications

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

- Genome Sequencing
  - De novo sequencing
  - Resequencing
  - Targeted (re)sequencing
  - Mitochondrial sequencing
  - Mutation detection
  - Amplicon sequencing
  - Amplicon Cancer Panel
  - ...
- Transcript Expression Profiling
  - RNA sequencing
  - miRNA sequencing
  - Deep-SAGE
  - Deep-CAGE
  - PAS : polyadenylation site
  - ...
- Transcription factor binding
  - ChIP sequencing : Chromatin Immuno Precipitation
- Structural variation
- Metagenomics
- ...

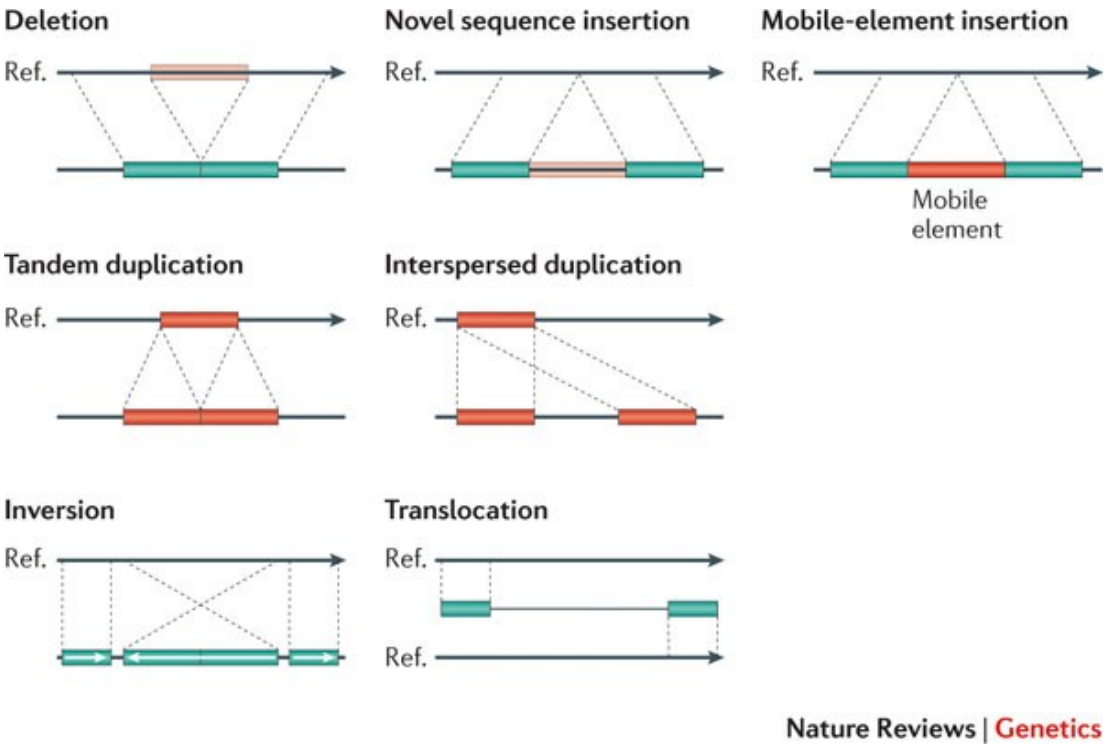
Protein-DNA interactions



Applications

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

- Genome Sequencing
  - De novo sequencing
  - Resequencing
  - Targeted (re)sequencing
  - Mitochondrial sequencing
  - Mutation detection
  - Amplicon sequencing
  - Amplicon Cancer Panel
  - ...
- Transcript Expression Profiling
  - RNA sequencing
  - miRNA sequencing
  - Deep-SAGE
  - Deep-CAGE
  - PAS : polyadenylation site
  - ...
- Transcription factor binding
  - ChIP sequencing
- Structural variation
- Metagenomics
- ...

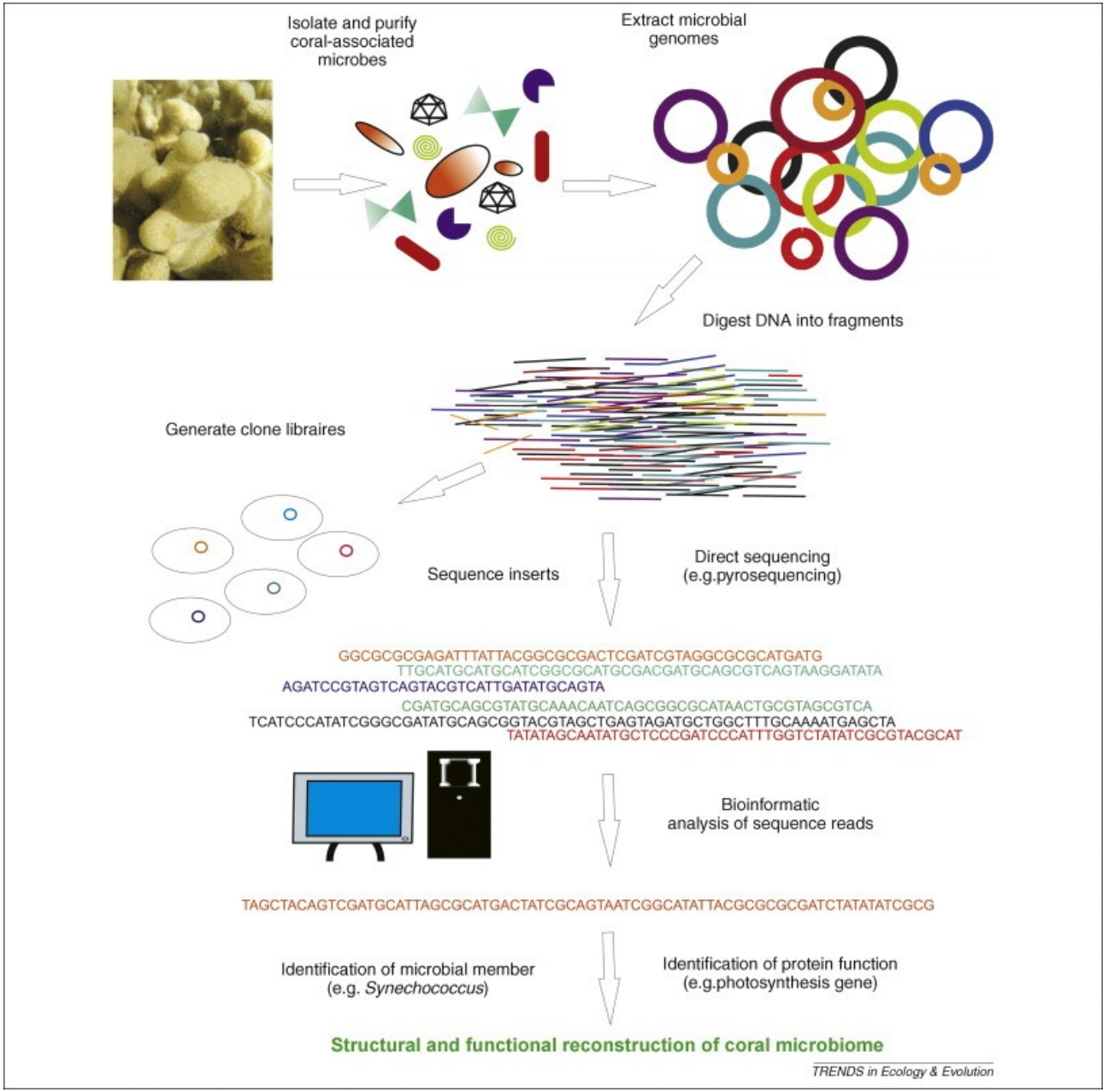


Applications

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

- Genome Sequencing
  - De novo sequencing
  - Resequencing
  - Targeted (re)sequencing
  - Mitochondrial sequencing
  - Mutation detection
  - Amplicon sequencing
  - Amplicon Cancer Panel
  - ...
- Transcript Expression Profiling
  - RNA sequencing
  - miRNA sequencing
  - Deep-SAGE
  - Deep-CAGE
  - PAS : polyadenylation site
  - ...
- Transcription factor binding
  - ChIP sequencing
- Structural variation
- Metagenomics
- ...

Study of genomes recovered from environmental samples



Run types

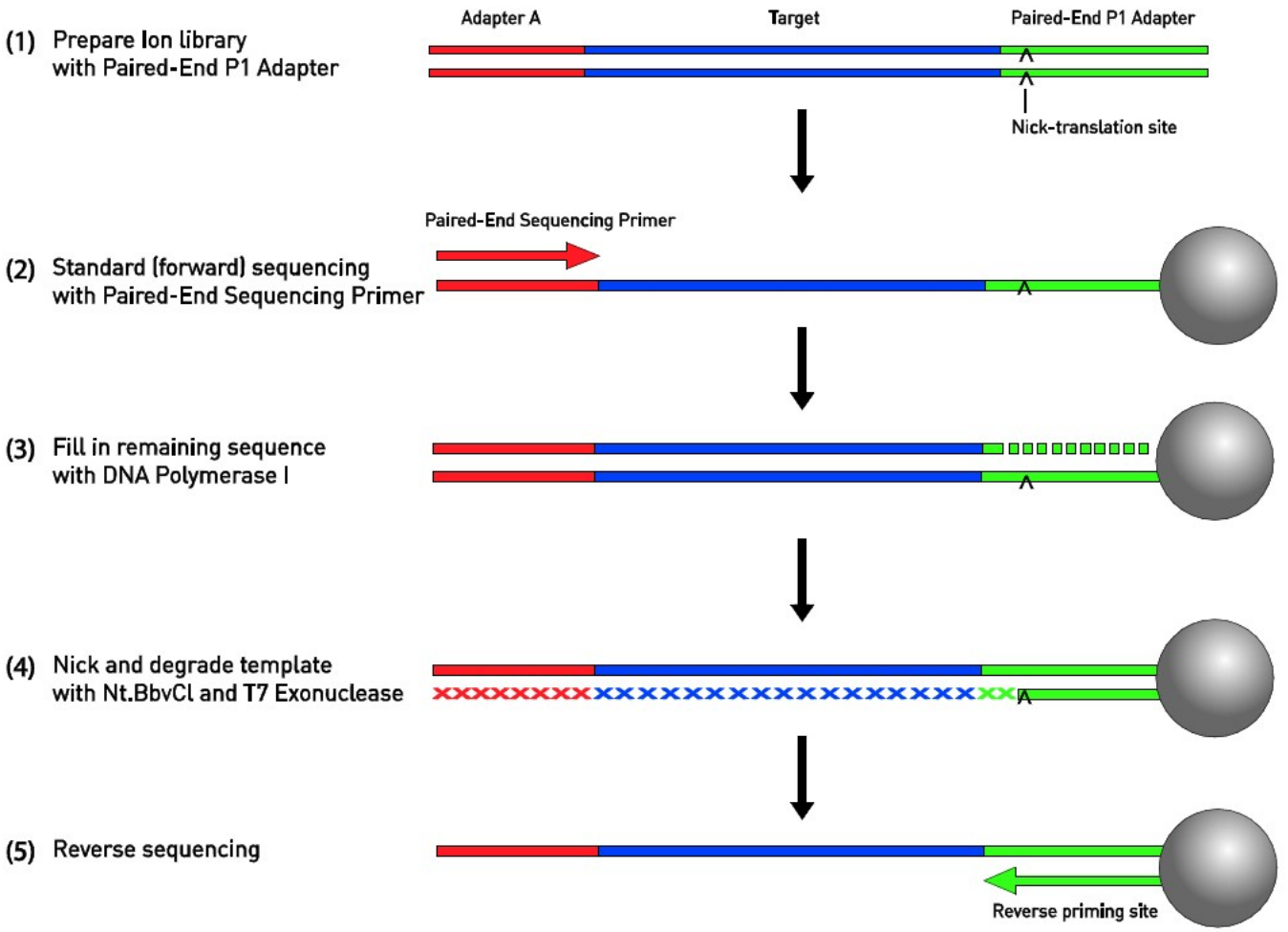
Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

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

Run types

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

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

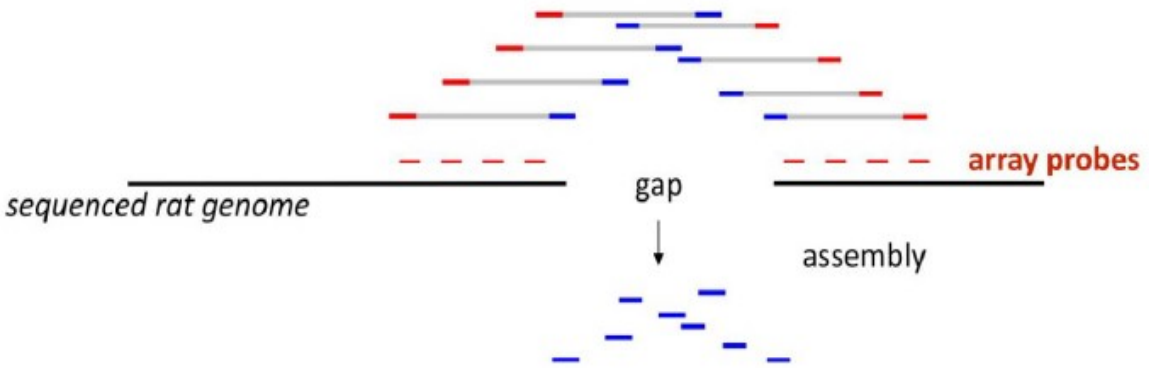


Short fragments (100-400 bp) : forward and reverse reads

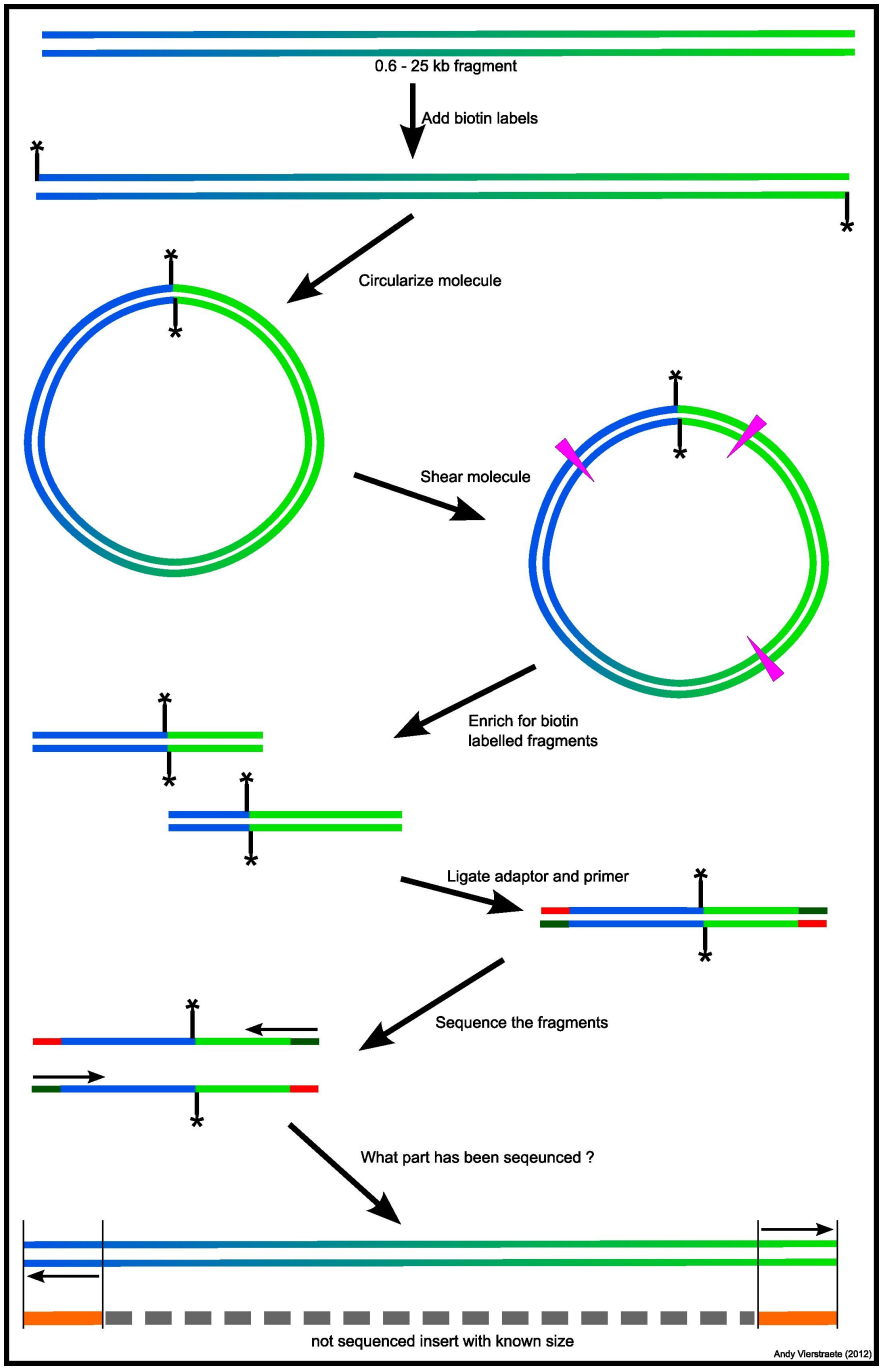
Run types

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

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



Long inserts possible : 0.6-25 kb

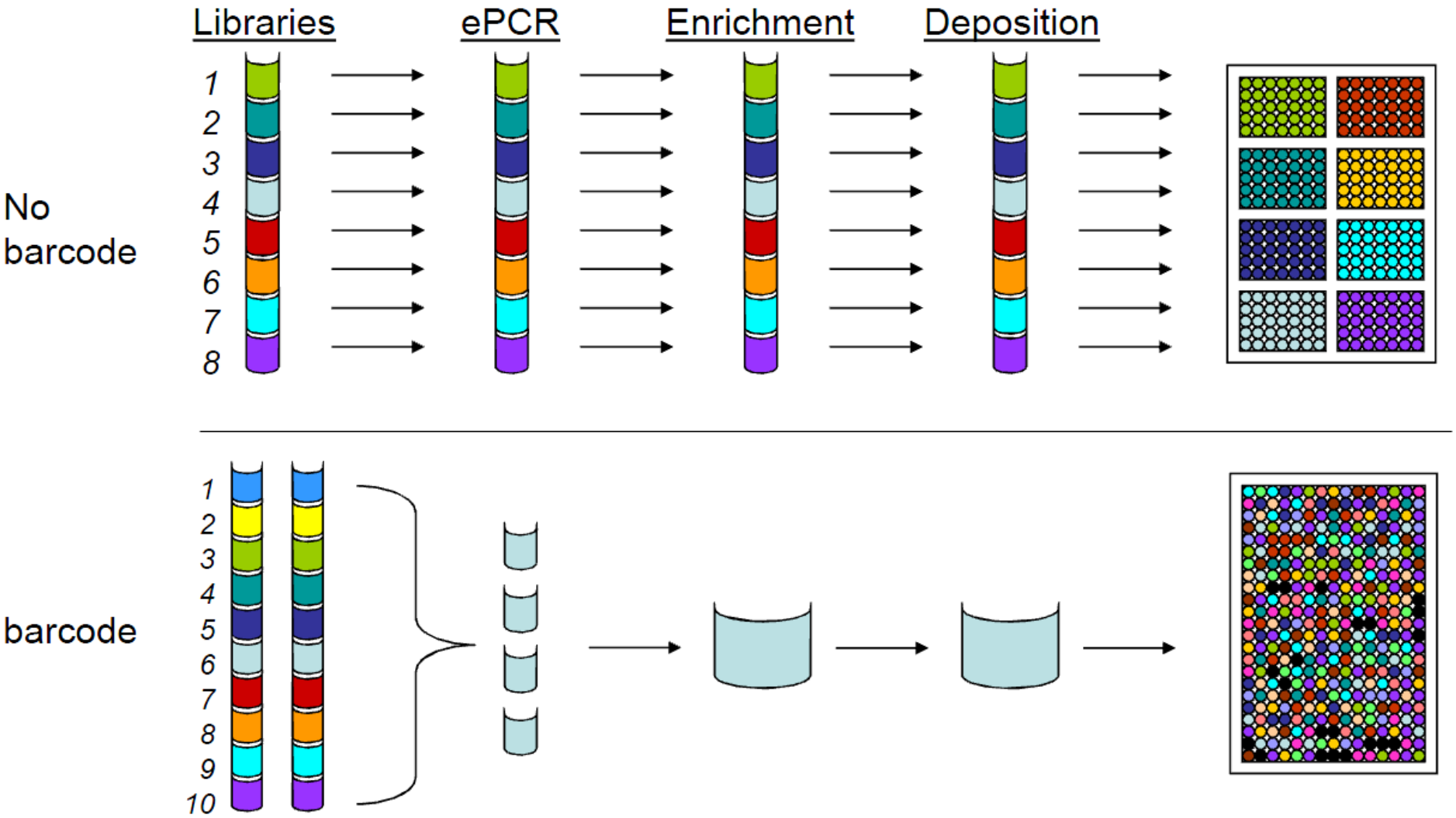


Run types

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

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

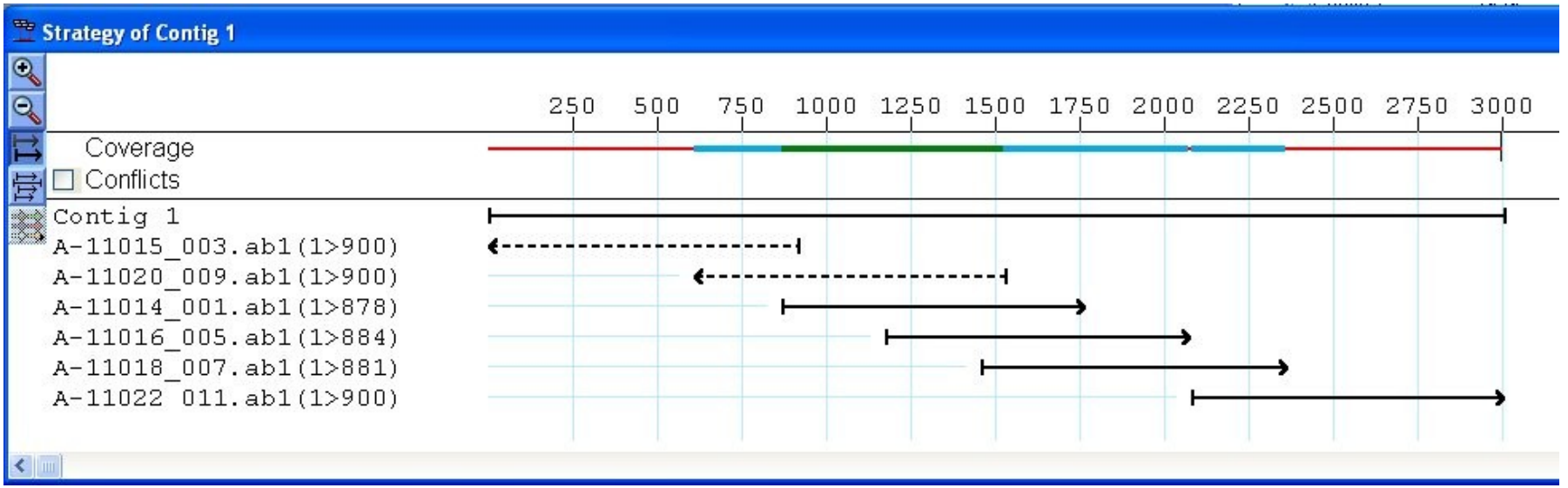
10 bp unique sequence is added tot the DNA

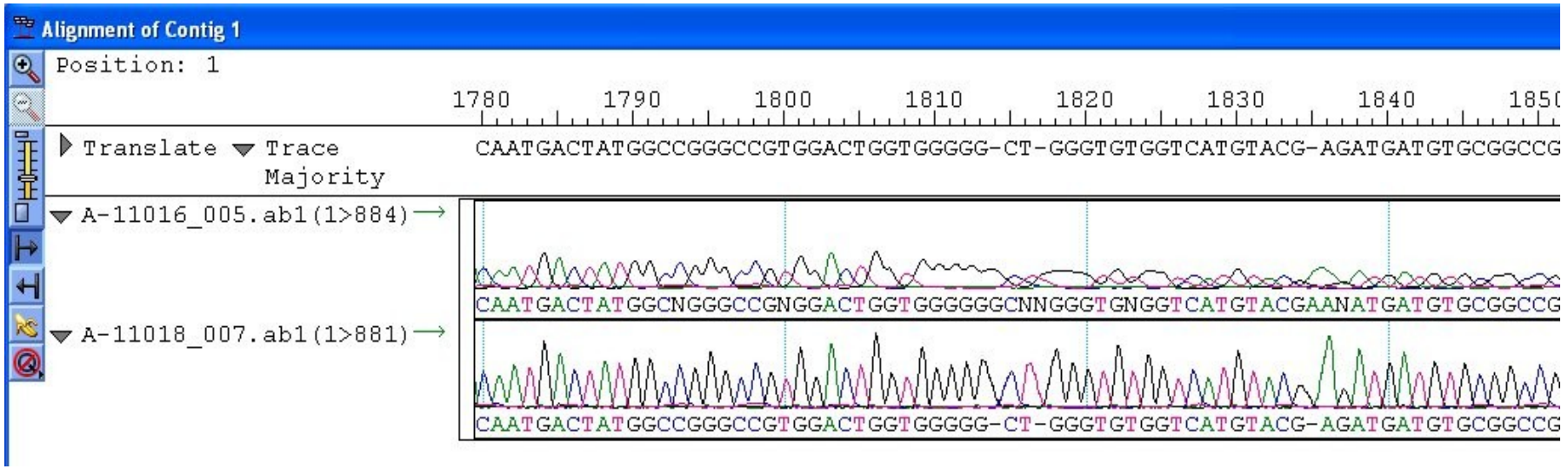


Data analyses

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

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

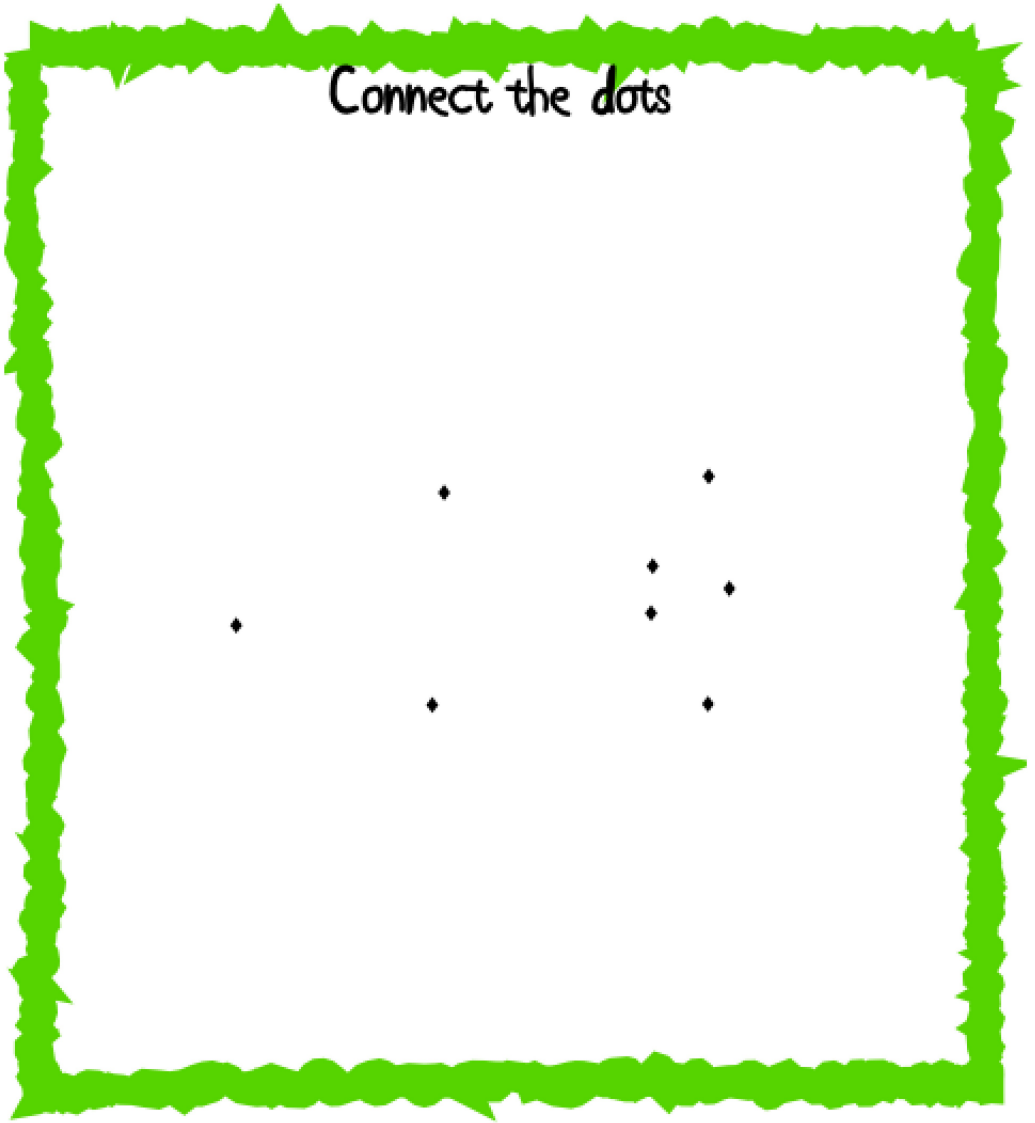




Manually check the assembly and correct errors.
2000 bp takes 5-10 minutes.

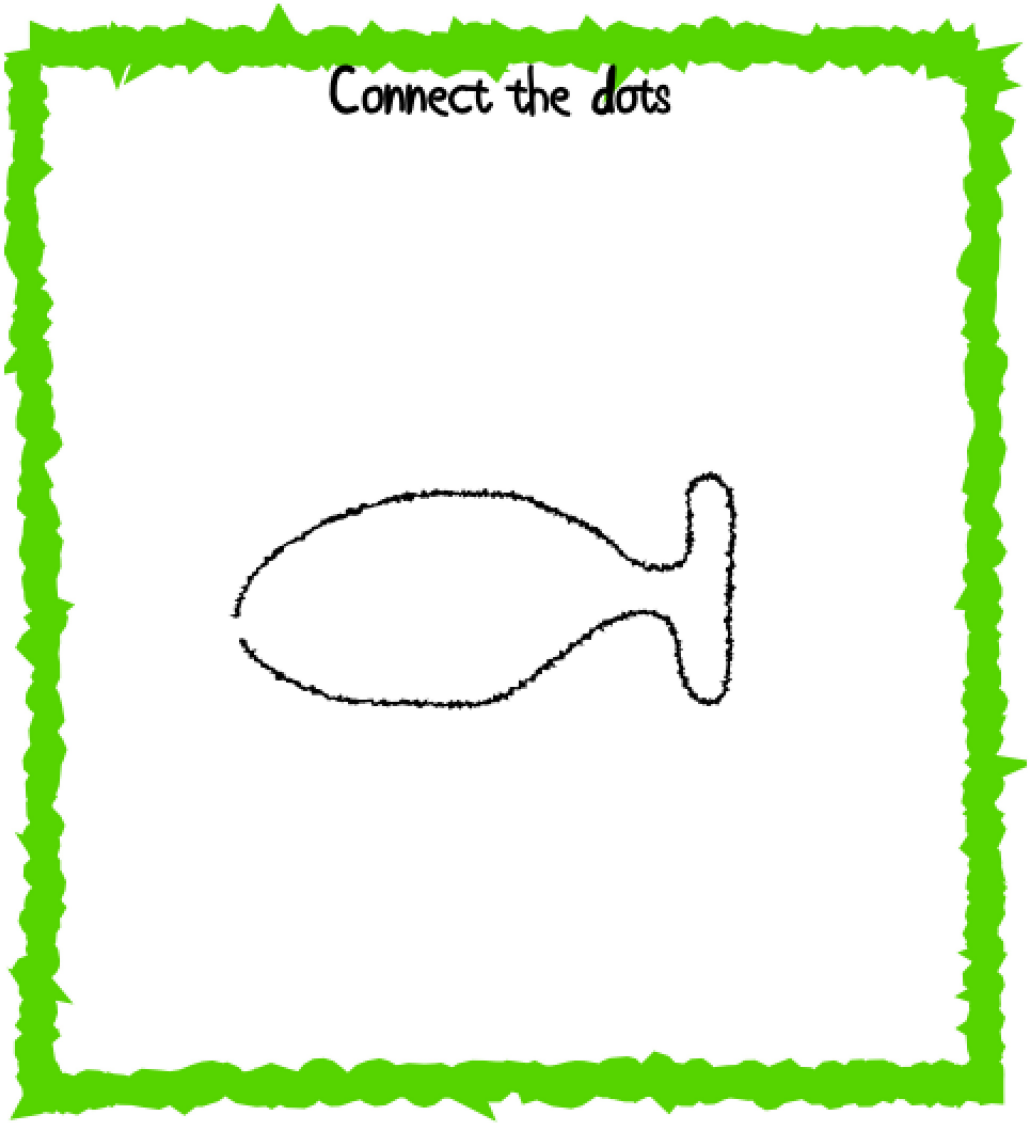
Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Sanger sequencing : simplified :



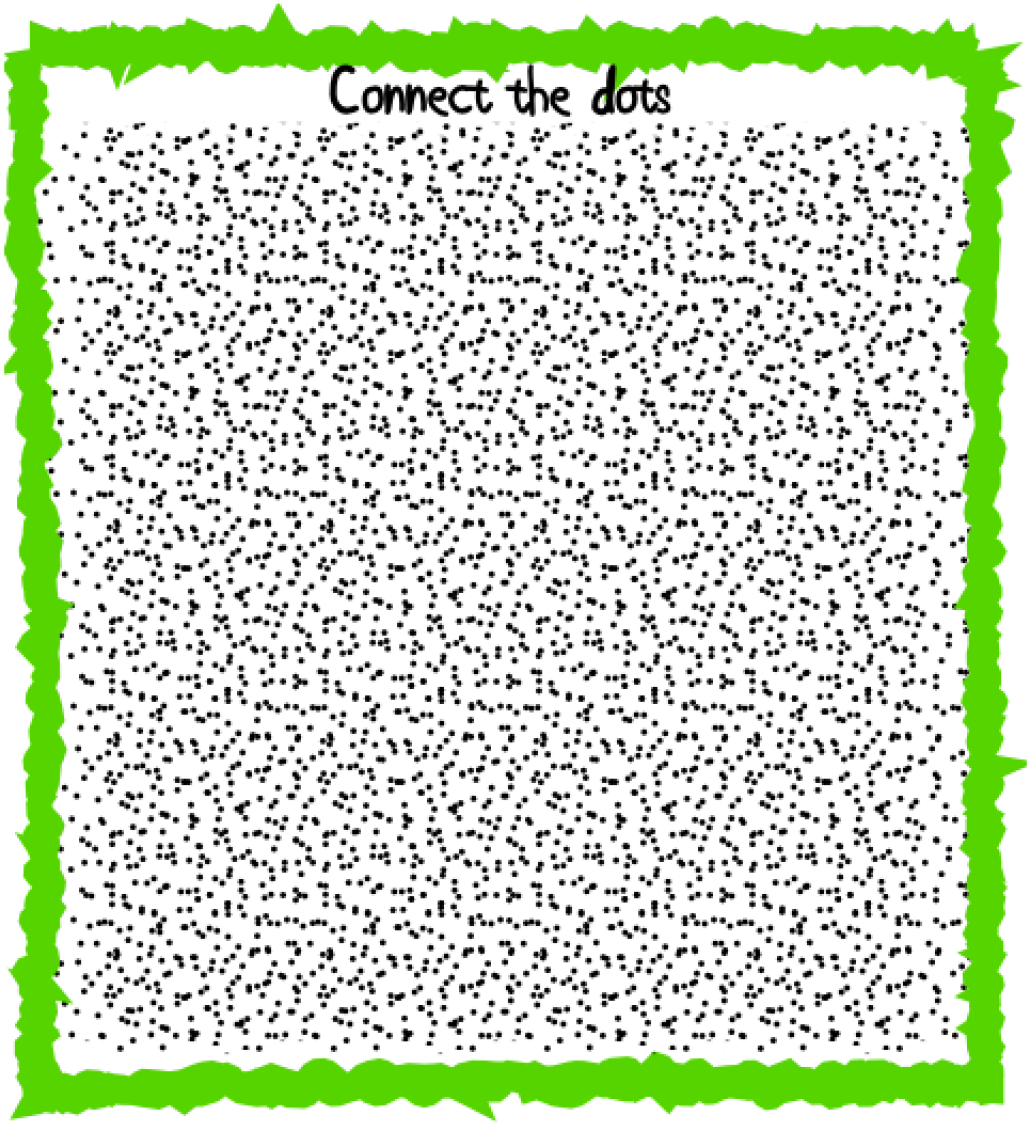
Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Sanger sequencing : simplified :



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

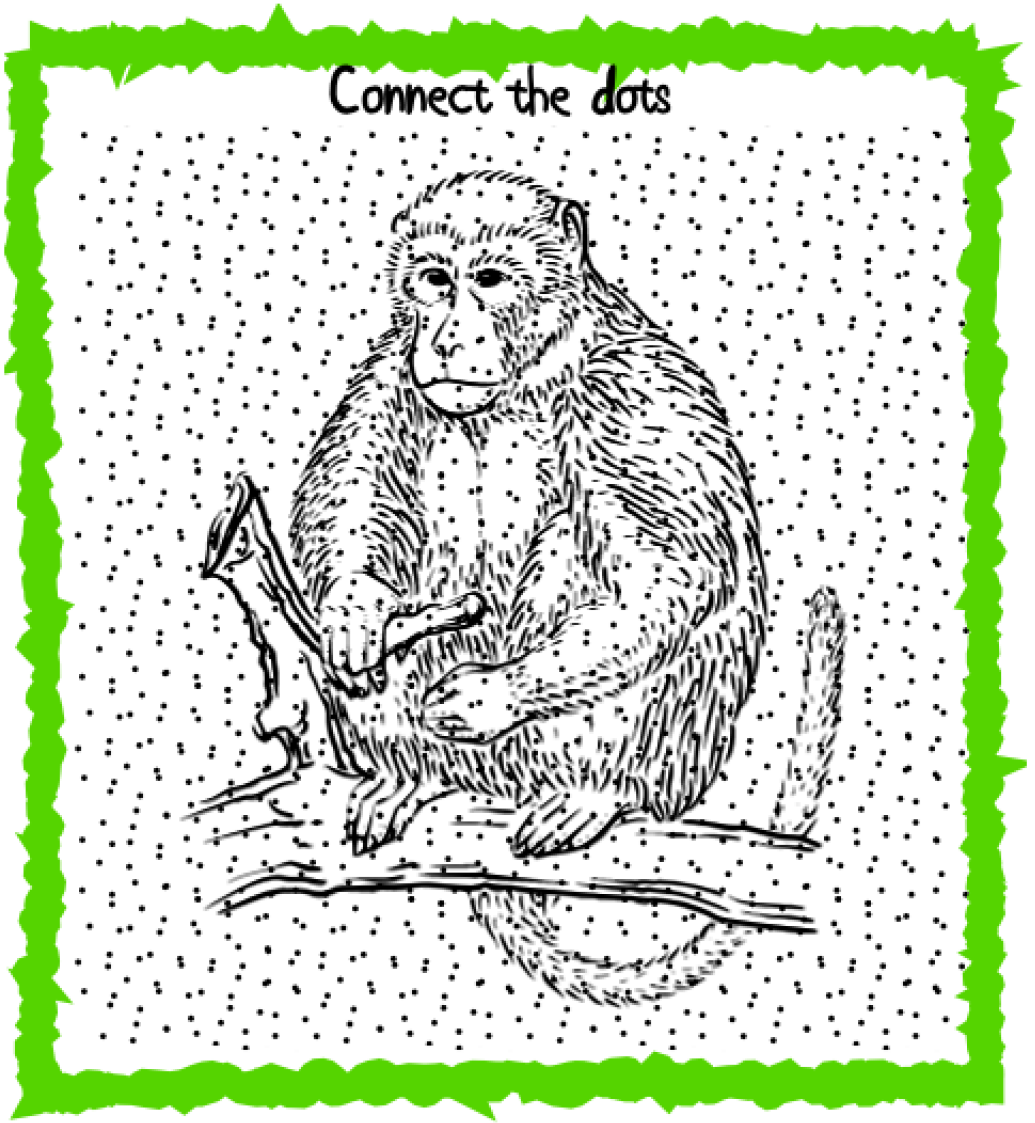
Next Generation sequencing : simplified :



Impossible to assemble manually

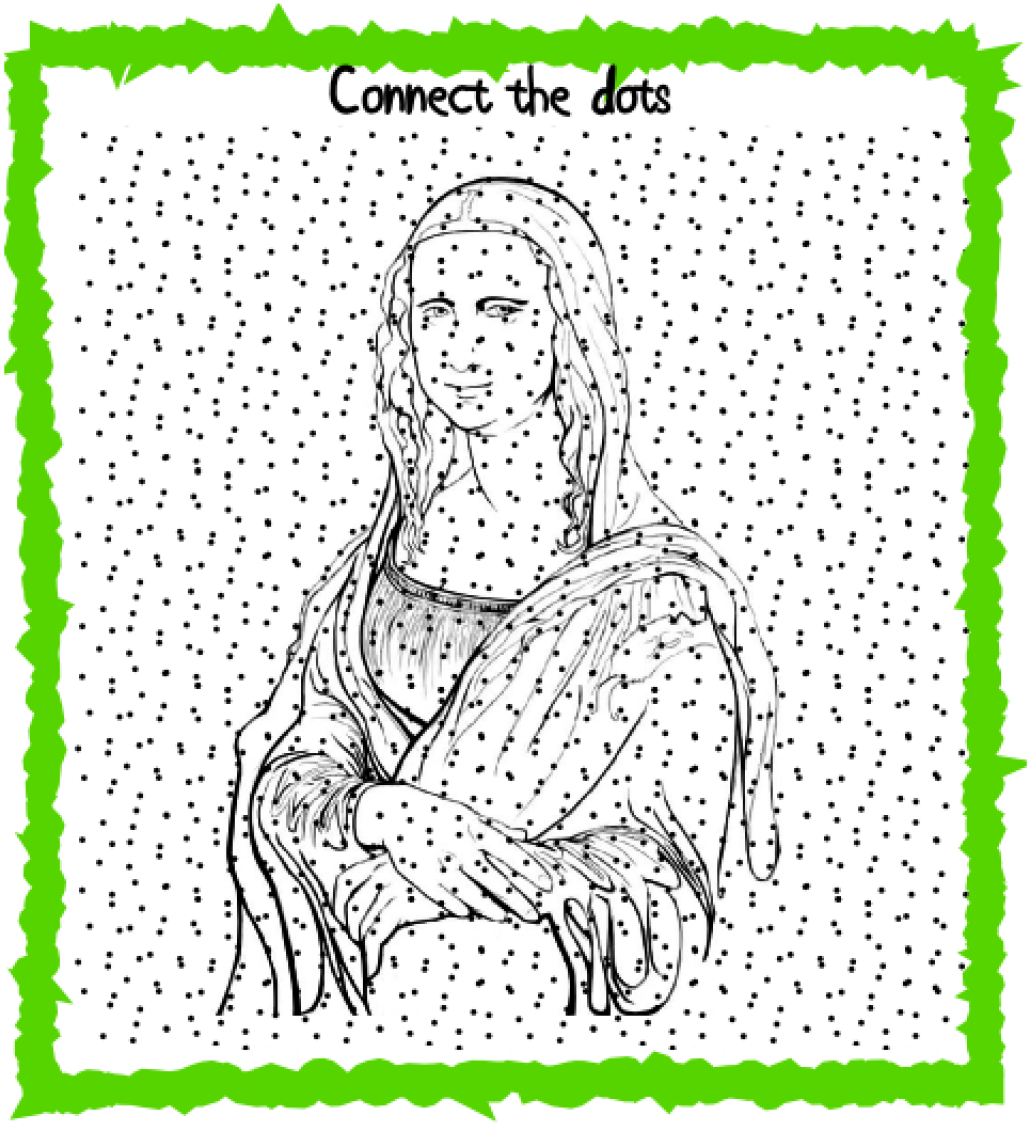
Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation sequencing : simplified :



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

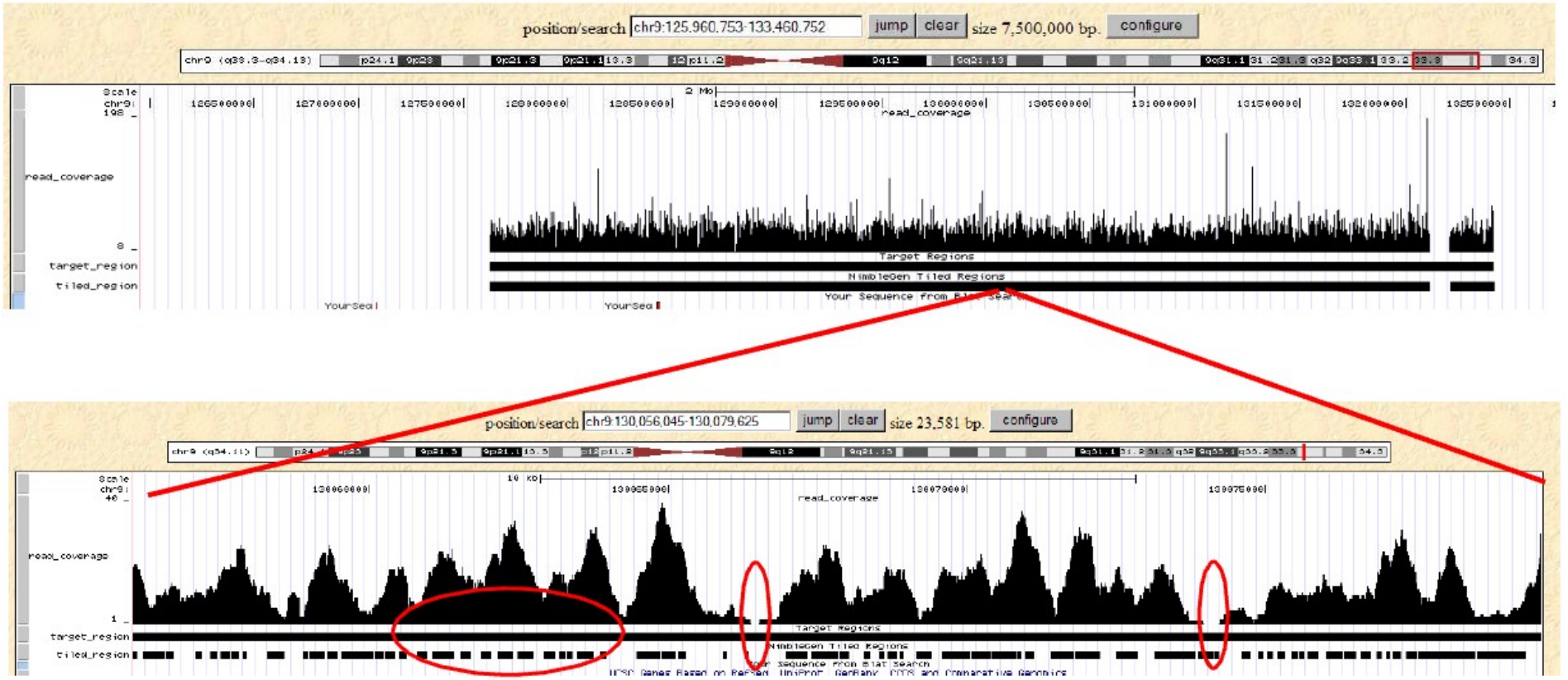
Next Generation sequencing : simplified :



Same dataset, different parameters

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation sequencing : Impossible to check manually

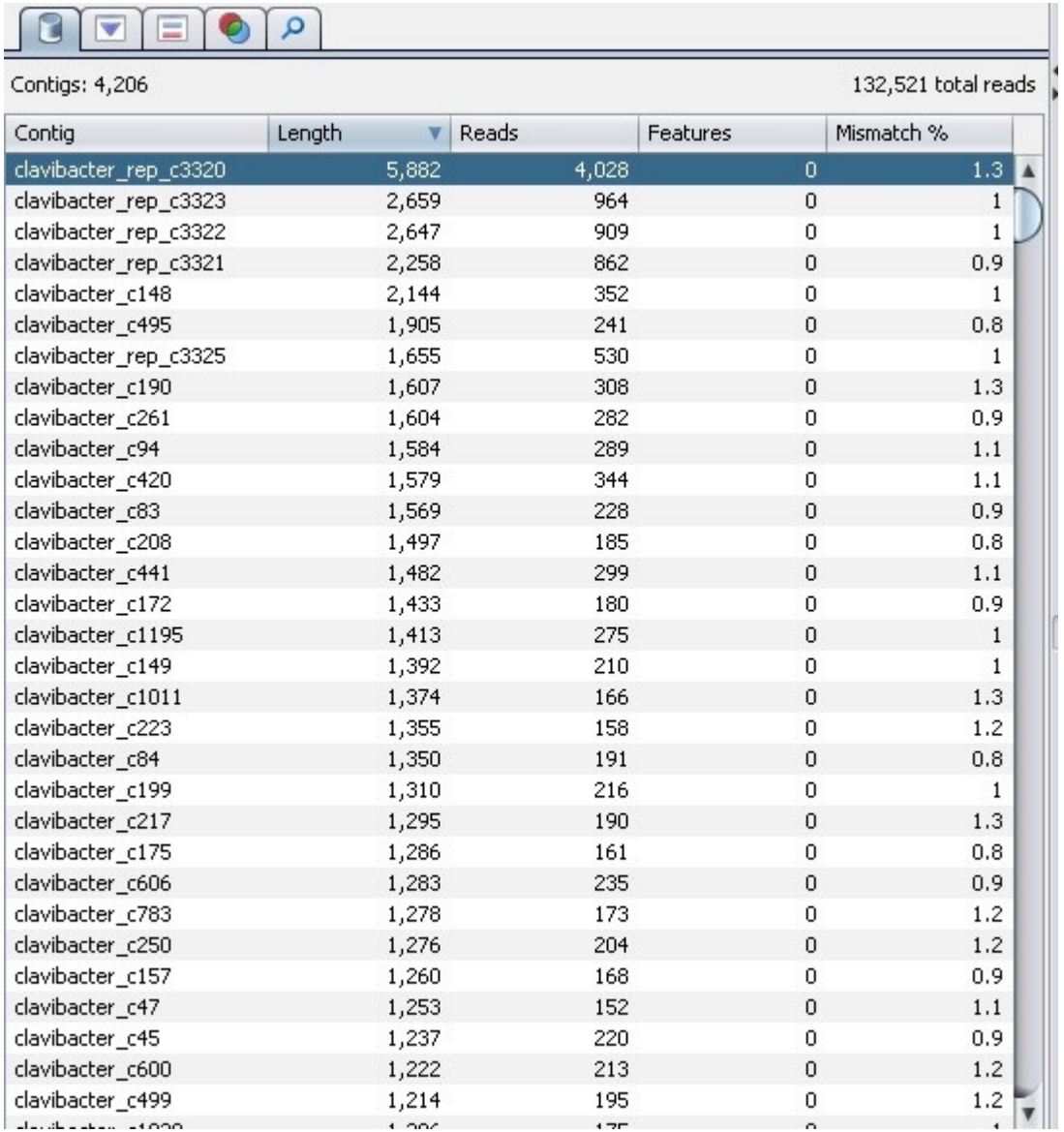




Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation sequencing :

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

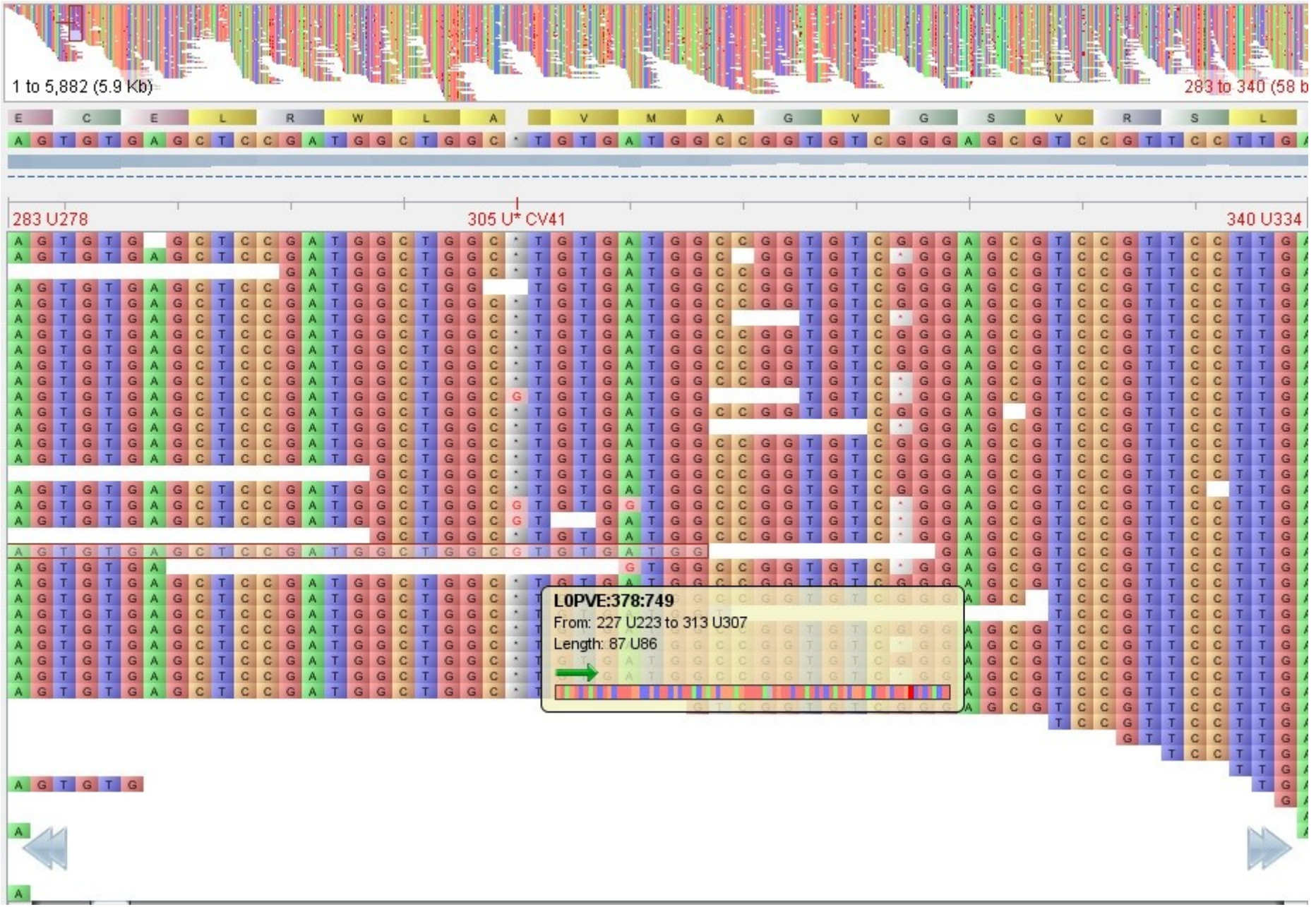


Contigs: 4,206 — 132,521 total reads

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

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

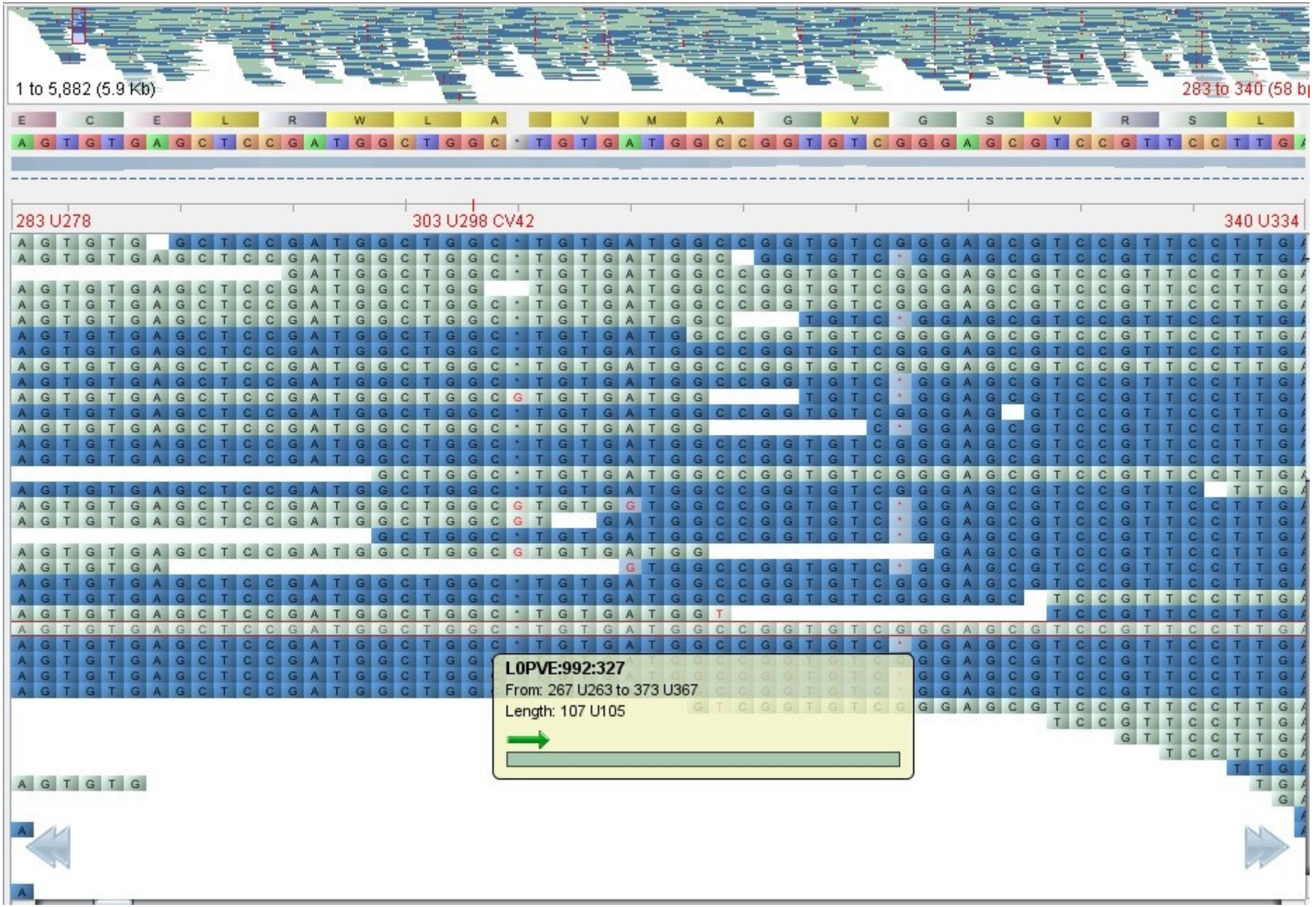
Next Generation sequencing : Assembly of the largest contig



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation sequencing :

Light blue : forward
dark blue : reverse sequence



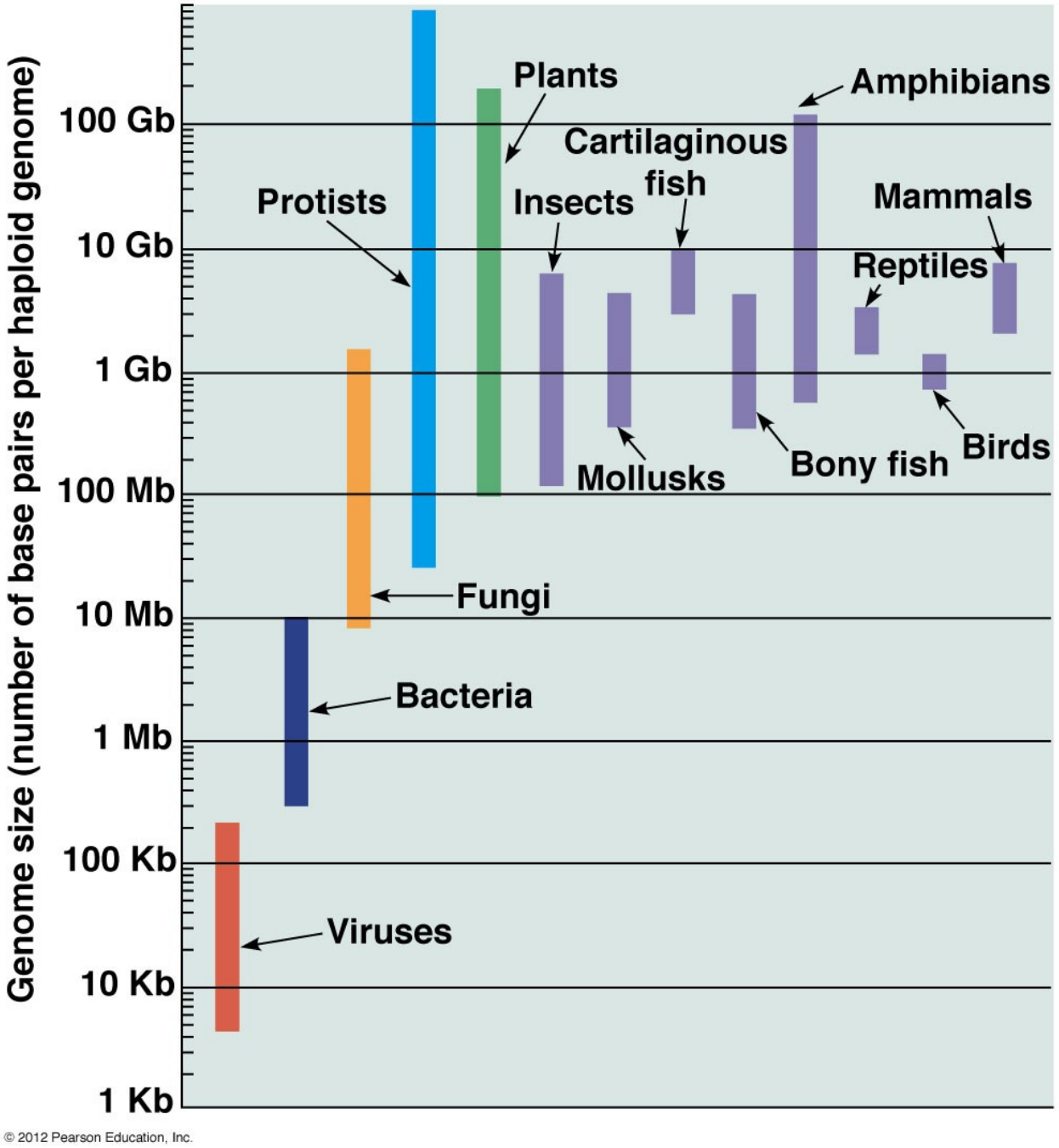
Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation sequencing : CV : coverage of a specific part



Data analyses

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012



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 !

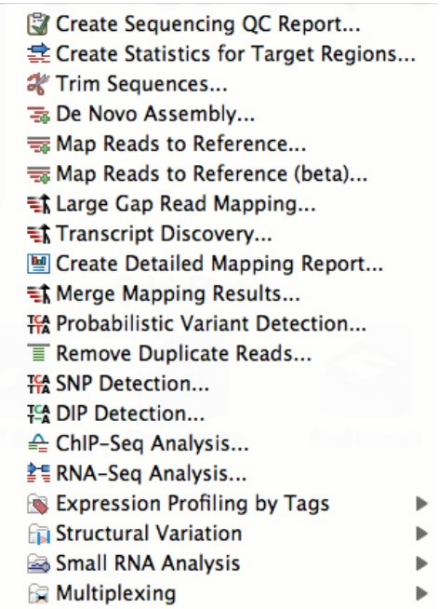
Total DNA sequencing :
1x gDNA
100x mDNA

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

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
- 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

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

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 »

Thread Tools | Search this Thread

01-24-2008, 08:19 AM #1

sci_guy
Member

Location: Sydney

Join Date: Jan 2008
Posts: 80

Software packages for next gen sequence analysis

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

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

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

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

* Exonerate - Various forms of pairwise alignment (including Smith-Waterman-Gotoh) of DNA/protein against a reference. Authors are Guy St C Slater and Ewan Birney from EMBL. C for POSIX.
* GenomeMapper - GenomeMapper is a short read mapping tool designed for accurate read alignments. It quickly aligns millions of reads either with ungapped or gapped alignments. A tool created by the 1001 Genomes project. Source for POSIX.
* GMAP - GMAP (Genomic Mapping and Alignment Program) for mRNA and EST Sequences. Developed by Thomas Wu and Colin Watanabe at Genentec. C/Perl for Unix.
* gnumap - The Genomic Next-generation Universal MAPper (gnumap) is a program designed to accurately map sequence data obtained from next-generation sequencing machines (specifically that of Solexa/Illumina) back to a genome of any size. It seeks to align reads from nonunique repeats using statistics. From authors at Brigham Young University. C source/Unix.
* MAQ - Mapping and Assembly with Qualities (renamed from MAPASS2). Particularly designed for Illumina with preliminary functions to handle ABI SOLiD data. Written by Heng Li from the Sanger Centre. Features extensive supporting tools for DIP/SNP detection, etc. C++ source
* MOSAIK - MOSAIK produces gapped alignments using the Smith-Waterman algorithm. Features a number of support tools. Support for Roche FLX, Illumina, SOLiD, and Helicos. Written by Michael Strömberg at Boston College. Win/Linux/MacOSX
* MrFAST and MrsFAST - mrFAST & mrsFAST are designed to map short reads generated with the Illumina platform to reference genome assemblies; in a fast and memory-efficient manner. Robust to INDELs and MrsFAST has a bisulphite mode. Authors are from the University of Washington. C as source.
* MUMmer - MUMmer is a modular system for the rapid whole genome alignment of finished or draft sequence. Released as a package providing an efficient suffix tree library, seed-and-extend alignment, SNP detection, repeat detection, and visualization tools. Version 3.0 was developed by Stefan Kurtz, Adam Phillippy, Arthur L Delcher, Michael Smoot, Martin Shumway, Corina Antonescu and Steven L Salzberg - most of whom are at The Institute for Genomic Research in Maryland, USA. POSIX OS required.
* Novocraft - Tools for reference alignment of paired-end and single-end Illumina reads. Uses a Needleman-Wunsch algorithm. Can support Bis-Seq. Commercial. Available free for evaluation, educational use and for use on open not-for-profit projects. Requires Linux or Mac OS X.
* PASS - It supports Illumina, SOLiD and Roche-FLX data formats and allows the user to modulate very finely the sensitivity of the alignments. Spaced seed intial 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,

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

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

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

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

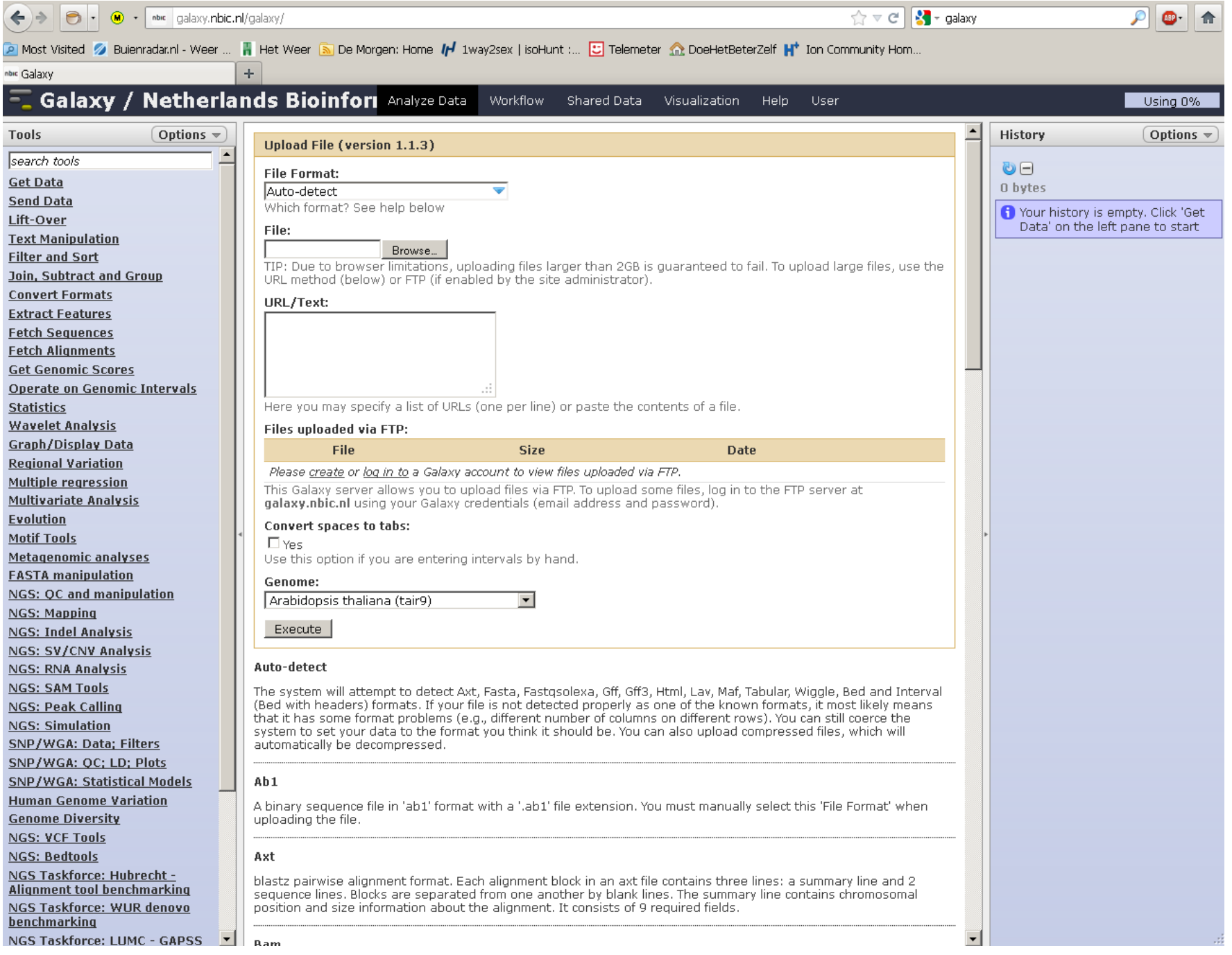
Immunoprecipitation enrichment of sequence fragments, indicating the location of a bound protein of interest. Original algorithm by Matthew Bainbridge, in collaboration with Gordon Robertson. Current code and implementation by Anthony Fejes. Authors are from the Canada's Michael Smith Genome Sciences 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

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012



Data analyses

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

FastQ format :

Barcode

Header 1
Sequence
Header 2
Quality score

```
@DBRHHJN1_0097:1:1:1111:1176#CGATGT/1
CCTTCTGGGGCAAGACTTGAGCAAAGACAAATGAAAGTAAAAGTTGCCCTGAAAACCCGAAAATCGTATGACTCTTTTTTTAATGTTTGATCAGGACCAT
+DBRHHJN1_0097:1:1:1111:1176#CGATGT/1
ggggggggcgggggggdggegggegggggdaggdegacedee\_eed^dee[decbgegXdebee^dddddUd[ceacI\^dadbbccc^bbBBBBBBB
@DBRHHJN1_0097:1:1:1200:1169#CGATGT/1
TGGATGTAAATTTAATTAATTTAGTAGGTCCTTCTCTTTTGTCTGAAACTAGGGCTCTATTAAGAATATAATAGTAGCACTTTTGAATAAAACTGAAAGG
+DBRHHJN1_0097:1:1:1200:1169#CGATGT/1
fgffdeggegggggeegffdgadddeeeeeeegeggefegegggdfgggfg_dd\ddU_`__fgggfggf\gdddddcccbbd\ffefggdc_cfdefdfffd
@DBRHHJN1_0097:1:1:1201:1199#CGATGT/1
GGTTAAAGACTTAAATGCTAGACAAGACACCATCAAACTCCTAGAAAAGAACATAGGCAAAACACTCTCTGATATCAACATTACAAATGTTTTCTCAGGT
```

Quality score using the ascii table

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
 ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII..........................
 .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                          |    |        |                                    |                     |
33                        59   64       73                                  104                   126

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

Depending on the application :

- Pre-alignment :
  - QC, demultiplexing, data statistics, ...
- Alignment :
  - denovo, mapping to reference, ...
- Post-alignment :
  - QC, statistics, SNP calling, annotation, ...

Considerations

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

- Can Next Generation Sequencing solve my problem ?
- What application do I need (de novo, RNA, ...) ?
- What is the best platform to run it on ?
- What is your experiment design ?
- What about bioinformatics ?
- Are your results correct ? (XY - XX chromosome SNP)
- 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.

What do we offer and what does it cost ?

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Price list for Next Generation Sequencing on the IonTorrent PGM
Lab Prof. Bart Braeckman

Price version : 9 May 2012

	Ugent		
	Biology (€)	Other (€)	Extern (€) (3)
1. Choose your library preparation : (1)(5)			
gDNA	300	350	410
total RNA	350	500	590
total polyA RNA (2)	410	560	650
rRNA depleted total RNA (2)	530	680	800
Barcoded	18	18	21
2. Choose your sequencing chip : (4)(6)			
314 (20 Mb)	650	775	900
316 (200 Mb)	950	1075	1250
318 (500 Mb)	1250	1375	1600

Example :

gDNA sequenced on a 316 chip will cost € 1250 (300 + 950) for people in the Biology department.

(1) One library preparation is enough for >50 sequencing reactions. All quality checks with Bioanalyzer included.

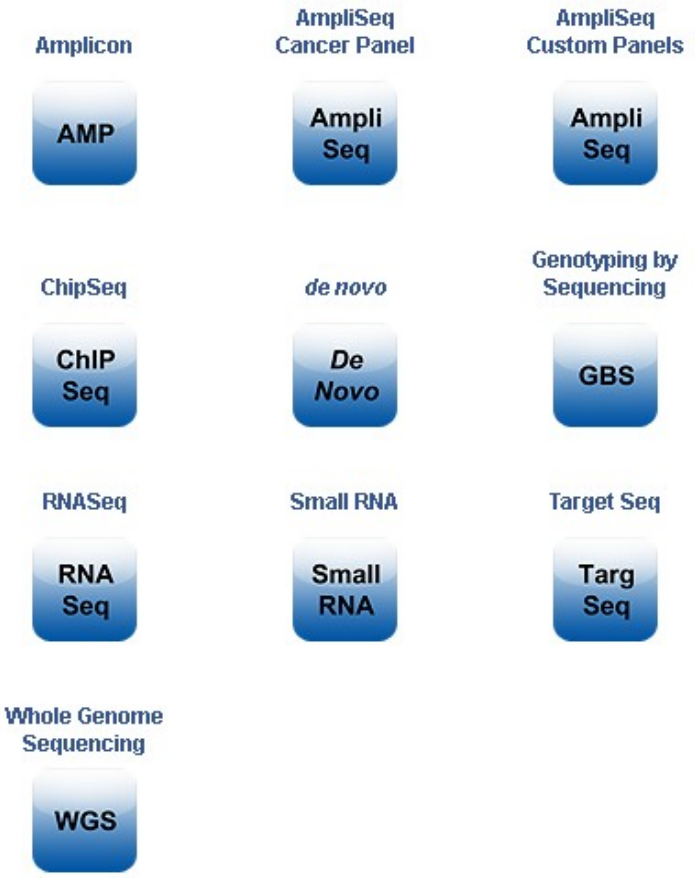
(2) You can prepare your own specific RNA of interest instead of delivering total RNA. In that case, we will treat it as total RNA.

(3) Prices excl. VAT

(4) USB stick with results in FASTQ format included.

(5) Other library preparations : ask for options and prices.

(6) Guaranteed sequencing output is of Q17 quality with the 200 bp kit. If, for any technical reason, sequencing output is lower than the guaranteed minimum, the overall sequencing price will be reduced proportionally. Upon agreement, extra chips of smaller capacity can be run at normal cost to meet the expected sequencing output.



End of 2012 : 400 bp kits ?

314 : 40 Mb
316 : 400 Mb
318 : 1-2 Gb

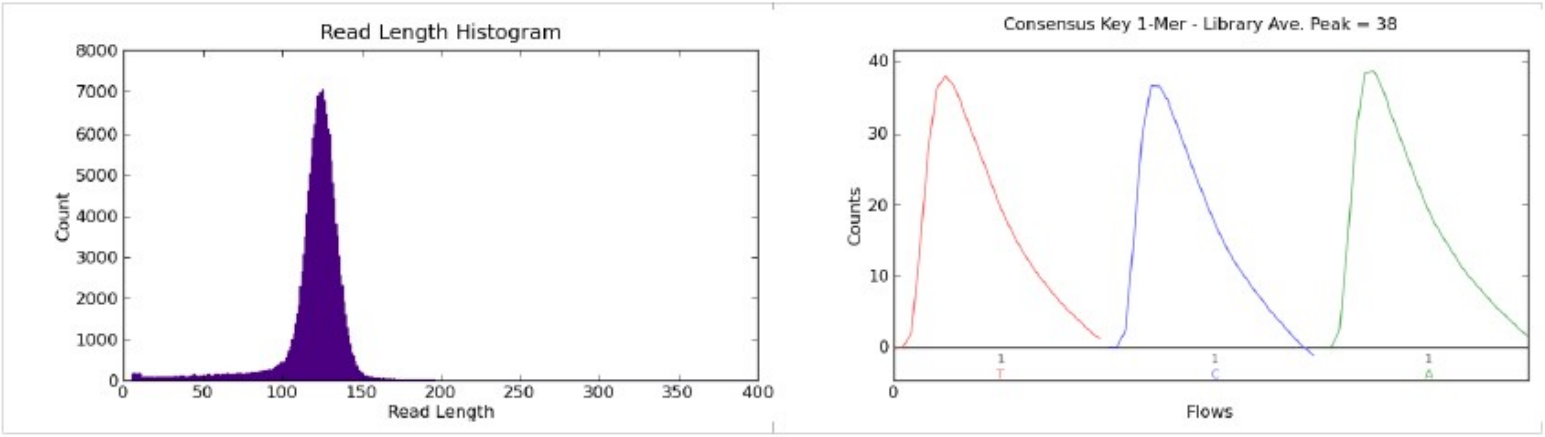
Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Report for clavibacter_no_reference

Library Summary

Based on Predicted Per-Base Quality Scores - Independent of Alignment

Total Number of Bases [Mbp]	20.25
▸ Number of Q17 Bases [Mbp]	13.50
▸ Number of Q20 Bases [Mbp]	11.97
Total Number of Reads	170,247
Mean Length [bp]	119
Longest Read [bp]	202



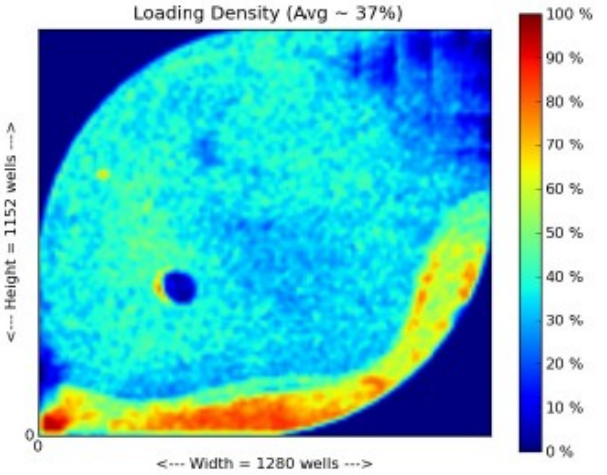
What do we offer and what does it cost ?

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

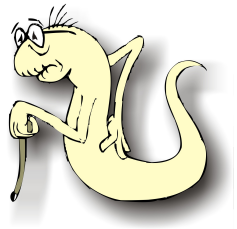
Ion Sphere™ Particle (ISP) Identification Summary

	Count	Percentage
Total Addressable Wells	1,262,519	
▸ Wells with ISPs	479,921	38%
▸ Live ISPs	380,141	79%
▸ Test Fragment ISPs	16,488	4%
▸ Library ISPs	363,653	96%

	Count	Percentage
Library ISPs / Percent Enrichment	363,653	78%
▸ Filtered: Polyclonal	109,448	30%
▸ Filtered: Primer dimer	137	<1%
▸ Filtered: Low quality	83,821	23%
▸ **Final Library Reads**	**170,247**	**47%**



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012



Thanks for your interest !

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

CTAGGTAGCTAGTCG
GCTLIFECISGATAG
C4-LETTERTWORDT
GCTATATCGTAGCTG

WWW.DNA.UGENT.BE