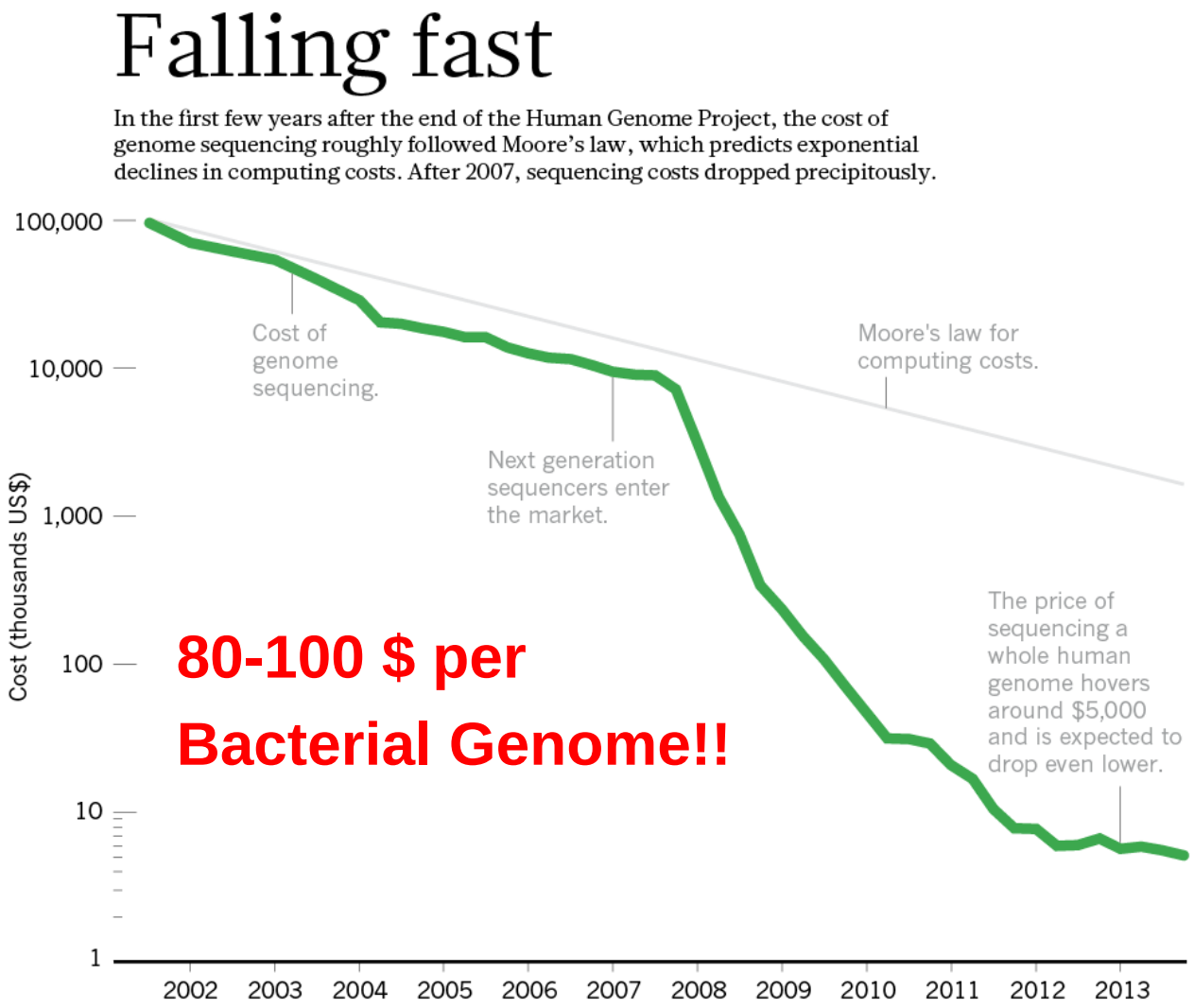
Next Generation Sequencers: from the bacterial culture to raw data

Valeria Michelacci

NGS course, June 2015

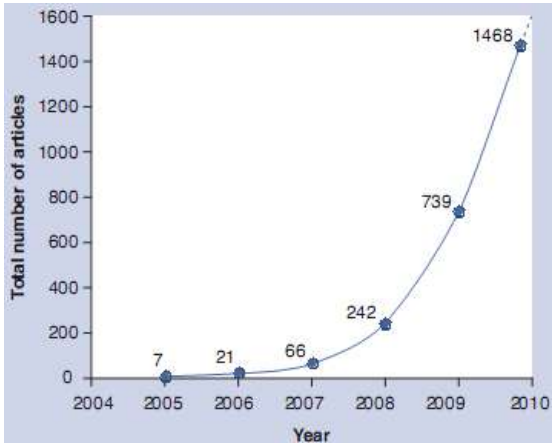
COSTS ASSOCIATED WITH DNA SEQUENCING



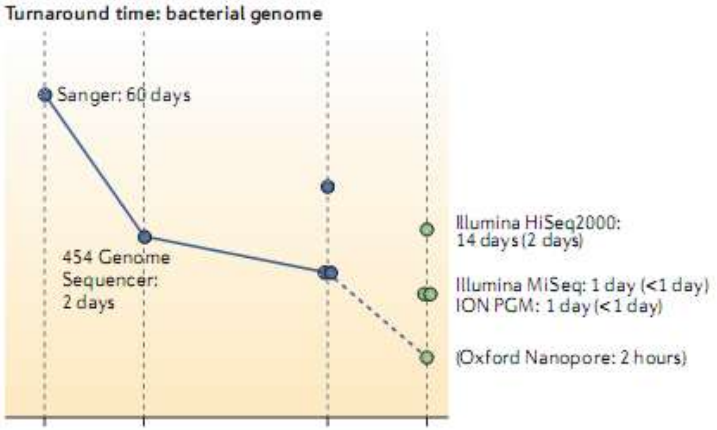
80-100 $ per Bacterial Genome!!

Benefits from NGS

Massive sequence output & low cost/base



Su Z. et al Expert Rev Mol Diagn. 2011 Apr;11(3):333-43.



Didelot X et al. Nat Rev Genet. 2012 Aug 14;13(9):601-12.

NGS in Microbiology

Particularly attractive:

Ability to generate large quantity of **starting material**

Small microbial genomes

Application in microbiology:

Multiple resistance determinants

Epidemiological markers

Virulence factors

Typing based on the SNPs in the WGS

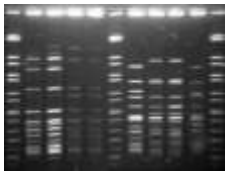
Classical methods and NGS

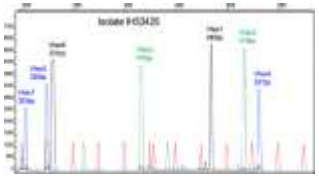
Isolation of the bacterial pathogen

Species identification

Test suscebtibility to antimicrobial drugs

Determining of pathogenic potential





Relating the bacterial pathogen to other strains of the same species

Multistep process

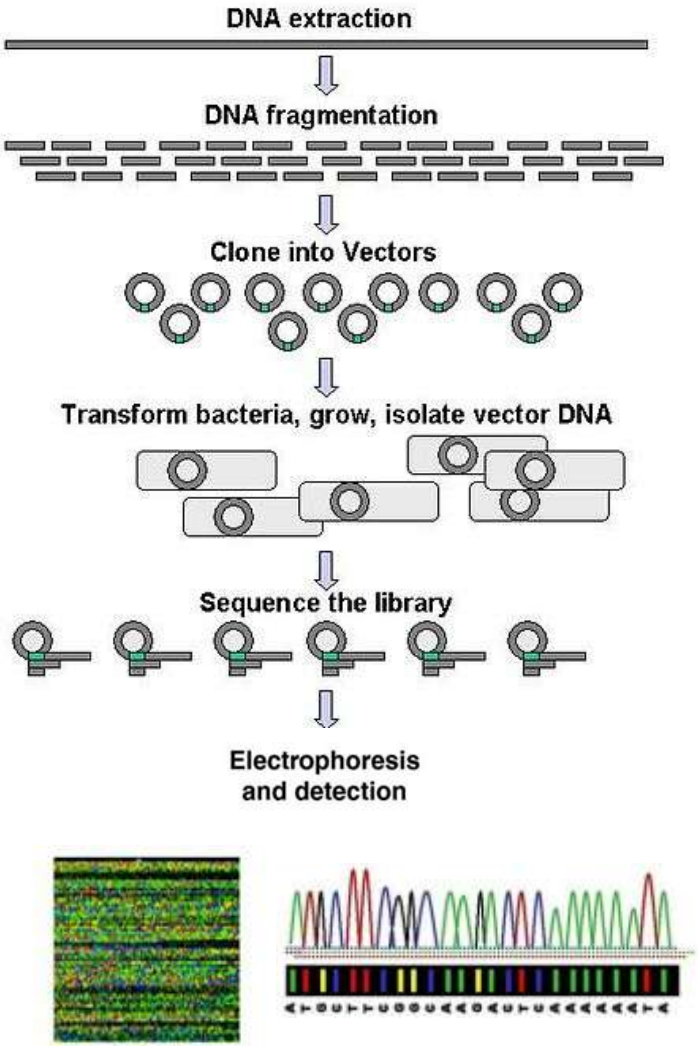
days
---
months

All the results derived from sequencing

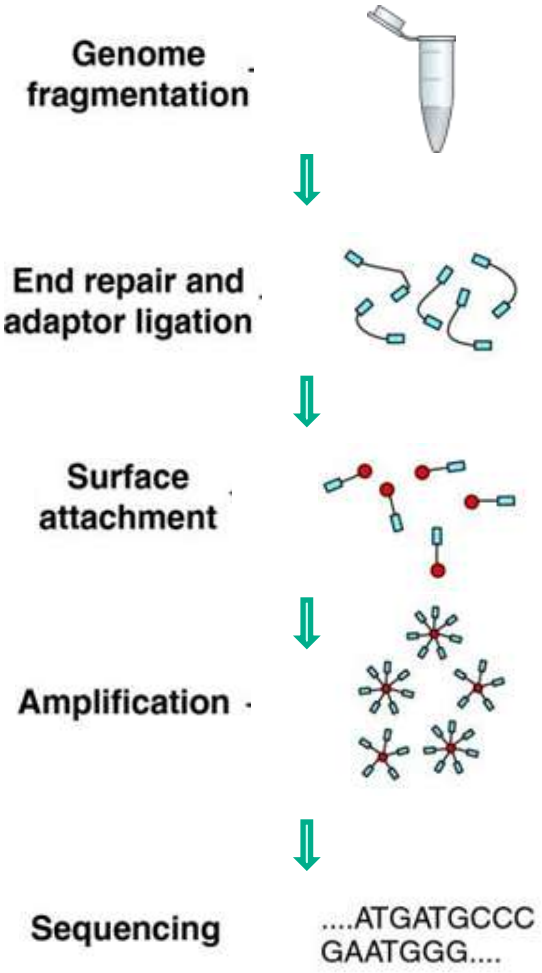
Need for huge preliminary work and of intense data analysis

Conventional sequencing vs NGS

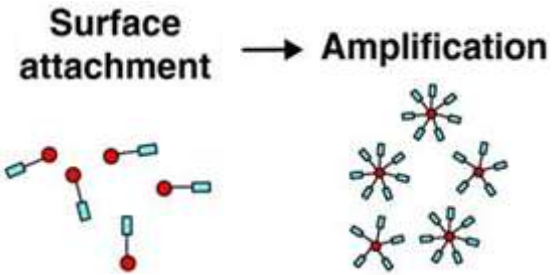
Conventional



NGS Pipeline



Next generation sequencing



...

...

illumina MiSeq

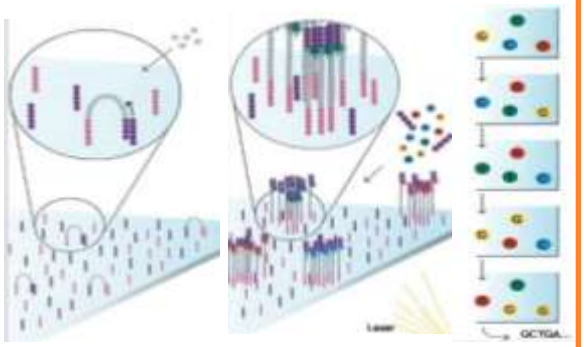
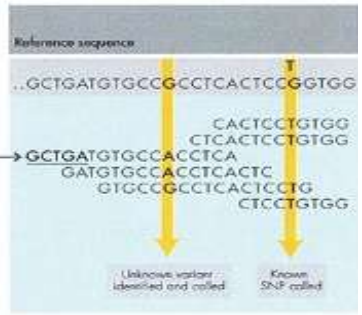
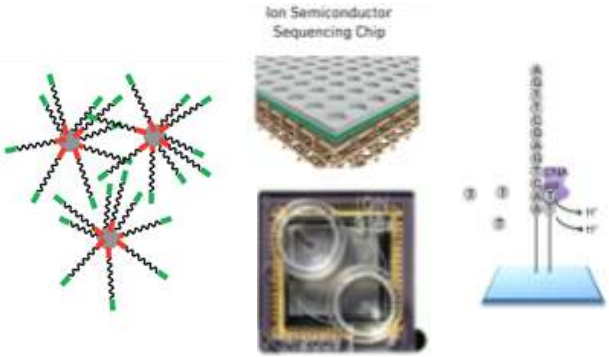


Image capture
Fluoresence detectioin

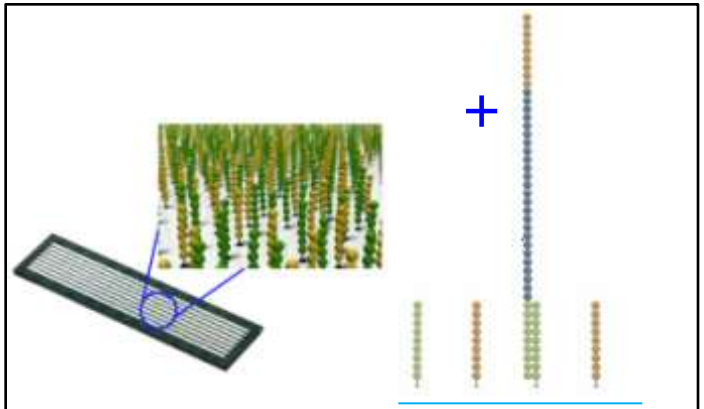


200bp-400bp short reads

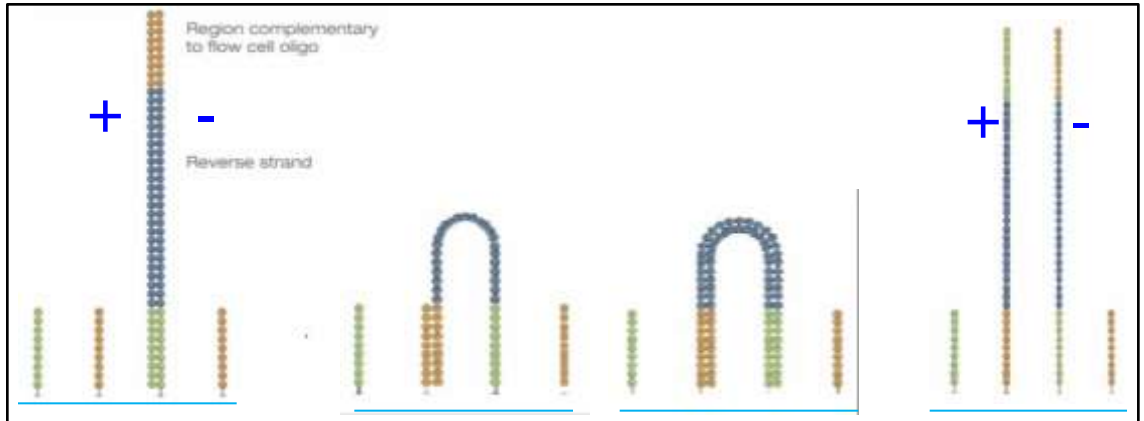


pH variation when incorporating nucleotides in the growing strand

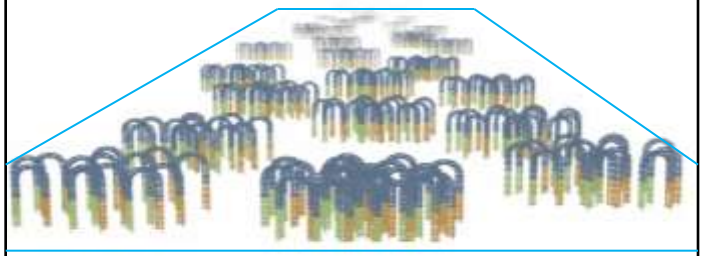
Illumina Sequencing by Synthesis/1



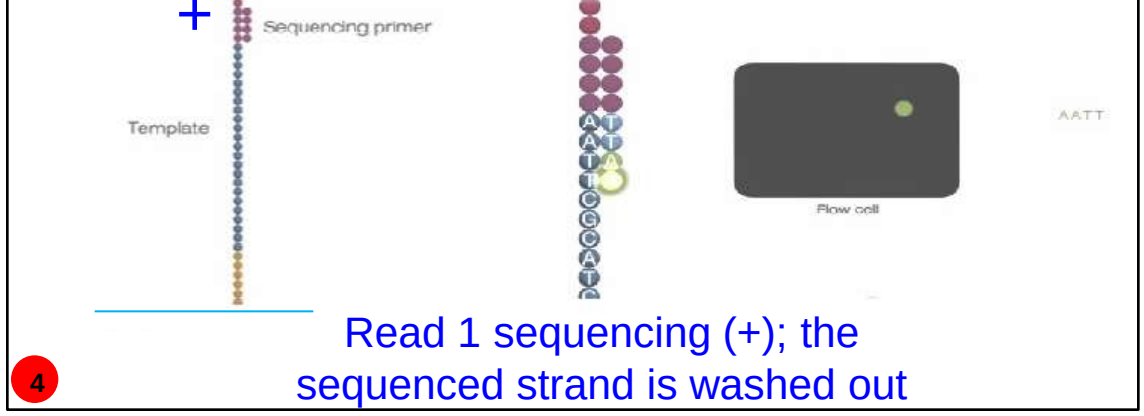
1 Flow cell ibridization



2 Cluster amplification in one position

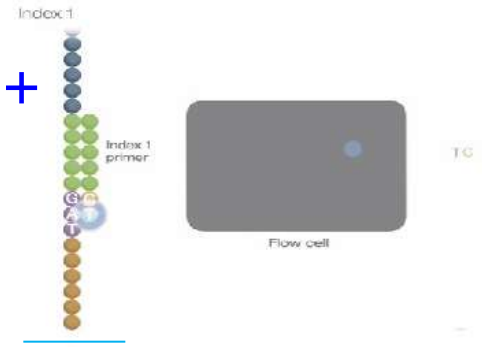


3 Cluster amplification throughout the flow cell



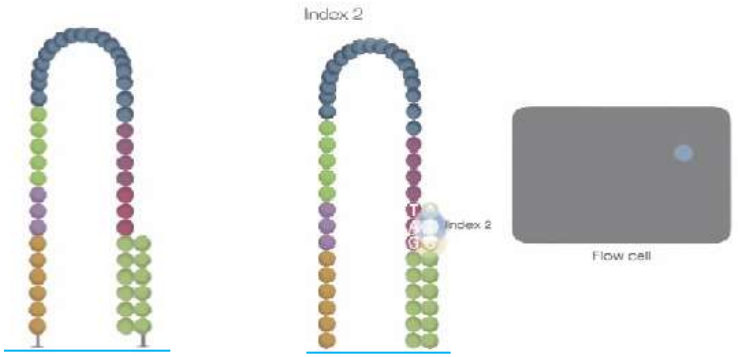
4 Read 1 sequencing (+); the sequenced strand is washed out

Illumina Sequencing by Synthesis/2



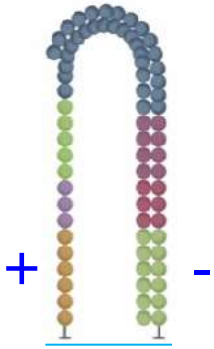
Sequencing of Index1; the sequenced strand is washed out

5

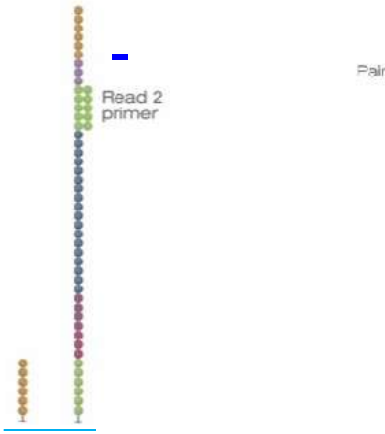


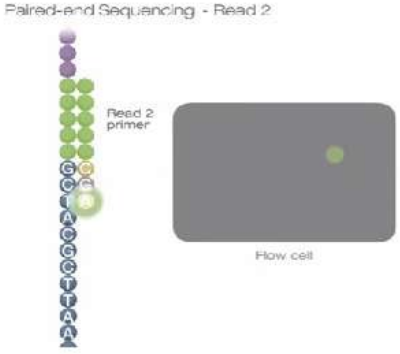
Bridging and sequencing of Index2; the sequenced strand is washed out

6



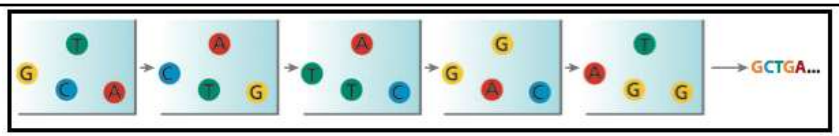
Synthesis of strand -. The strand + is cleaved and washed out





Read 2 sequencing (-)

7



strainname_R1.fastq

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

strainname_R2.fastq

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

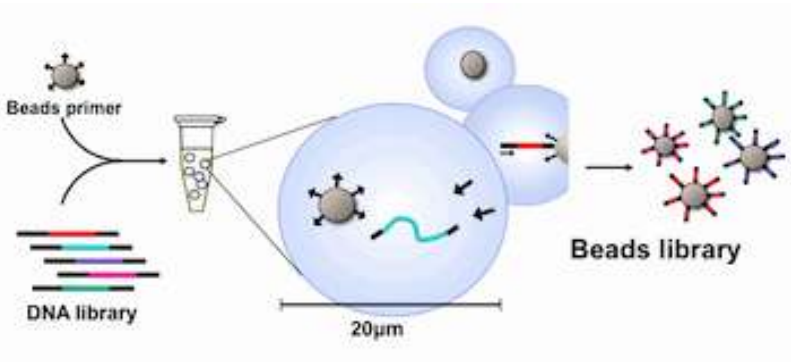
FastQ files compiling

8

Ion Torrent semiconductor sequencing/1

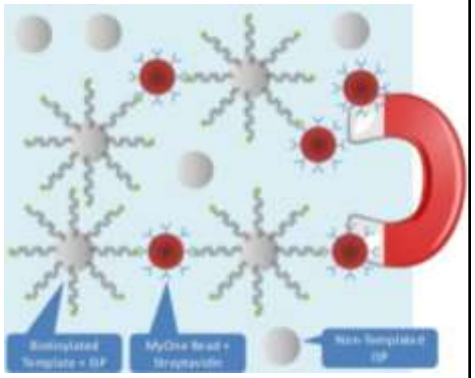
Shearing and adapter ligation

1



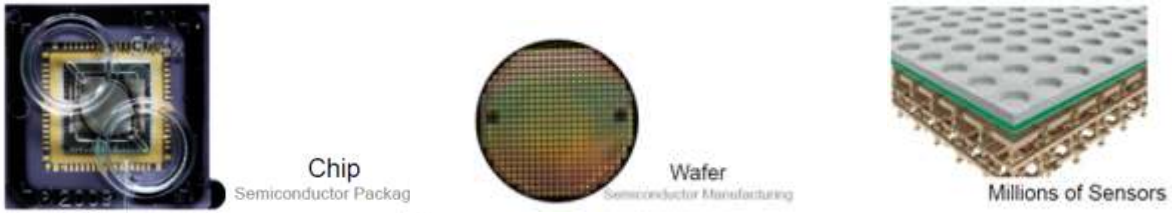
Equimolar emulsion PCR

2



Magnetic enrichment step for loaded spheres

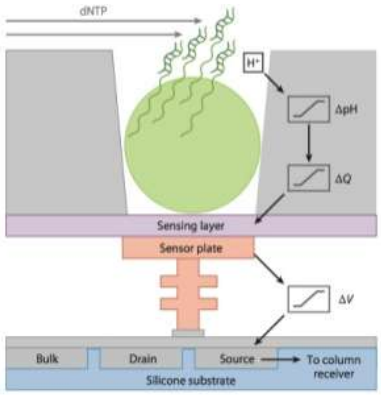
3

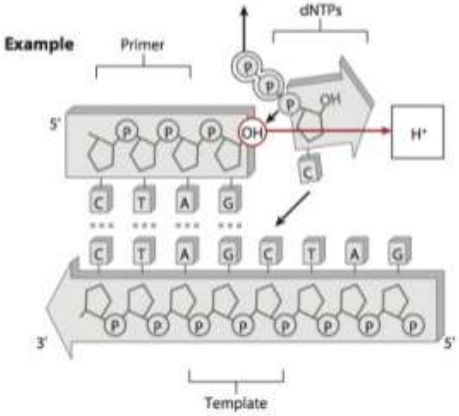


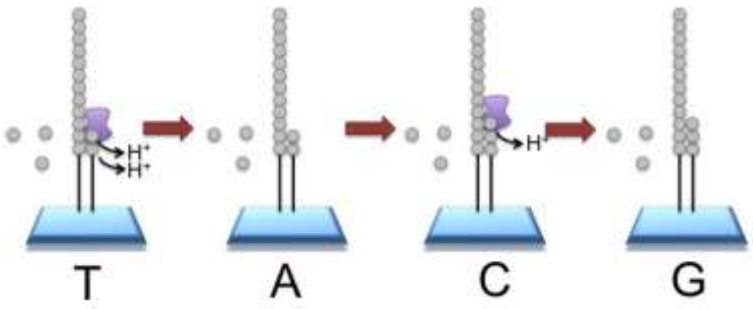
Chip loading

4

Ion Torrent semiconductor sequencing/2

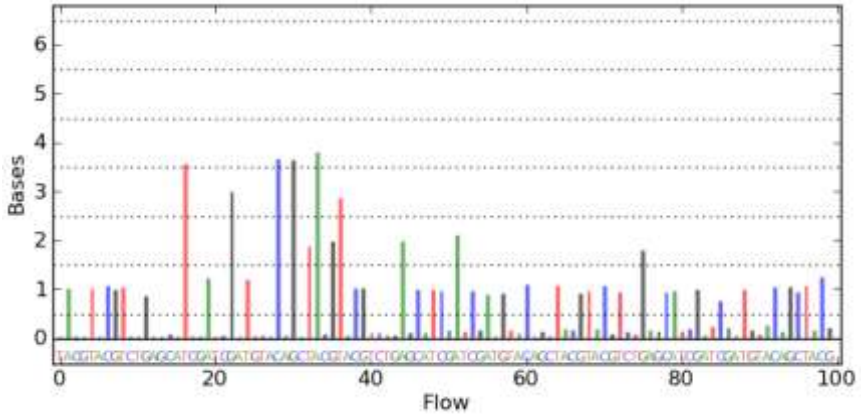






5

dNTPs flow and pH monitoring

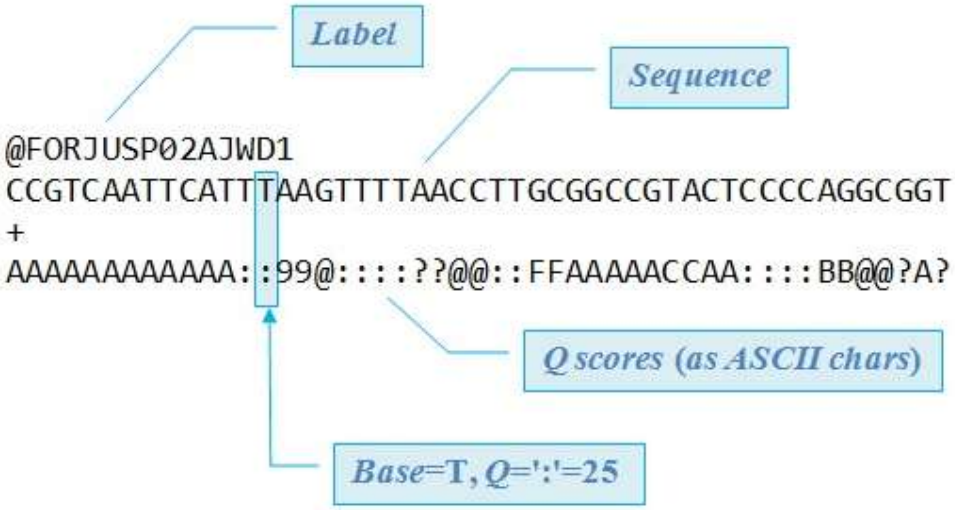



```
@read1
AGCTTATCCTCTGCTCACCCCCGGGTTAGCGCACTTGATGTATTCACAGC
+
BA1@CC7CBCCC9C8;B2@>C?B@B@B3=9?@B1:AB7B?B8B?B6B.7.
@read2
TTGGGCGGGATCTCCAGAAGCATATGGATGTGATCCACACAGCATTCTGC
+
?>?B@)<?@,AA7A@C<C?=@@B;+)?B5*@2=@+=BB,=B6C>AB@B24
@read3
TATGCTCAAGAAGGGGCTGATGAGTTGGTGTTTTACGATATCACTGCCTC
+
A3AB:B1:B;9/0BBBCBB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

6

Flow diagram interpretation and fastQ file compiling

.fastq files



Each .fastq file covering a 5 Mb genome at 30X weights about **300 MB**

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Phred quality score

$$Q = -10 \log_{10} P$$

from 0 to 93 using ASCII characters 33 to 126

.fastq files

@

```
@X1L6C:01561:00672
AAATATCACCAAATAAAAAACGCCTTAGTAAGTATTTTTCAGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTG
GATTAAAAAAAGAGTGTCTGATAGCAGCTTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGGTCAC
TAAATACTTTAACCAATATAGGCATAGCGCACAGACAGATAAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCA
CCACCATTACCACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGACGCGTACAGGAAACACAGAAAAAAAGCCCGCA
CCTGACAGTGCGGGCTTTTTTCGACCAAAGGTAACGAGGTAACAACCATGCG
+
CC:9::FBC<CD7:88888(:>><C<CCCC<CCBBAAB/A@A8888,;<@;AABBB=?;B98992:B<
CGBBCGDCC??BCC;BB<ADEEED*CCCAAACCCBCABBDDBB>B??A;999;@8=>199A7>9::CBCH:B:>>>)999)
77037;<7==5=@@BBCC:C@BBB9B<E<D9>?><<6ADCBCBAABB@@@DDCCBA@@==+=.//?B<??AEB::6;DCD>
C:;;;-:9:BC<BBCCC9??<AA;AG<CB>GD@B;;;A<AE;AA<B??@9@C<BB<???BB;BBBAAAA:::BAB099/9>
@========(<<?)99997>>CCEBA>>=>2373333&3:99-33(3--717---43606704/47761
```

@

```
@X1L6C:01104:03031
AGAAGCTGCTATCAGACACTCTTTTTTTAATCCACACAGAGACATATTGCCCGTTGCAGTCAGAATGAAAAGCTGAAAAATA
CTTACTAAGGCGTTTTTTATTTGGTGATATTTTTTTCAATATCATGCAGCAAACGGTGCAACATTGCCGTGTCTCGTTGCTC
TAAAAGCCCCAGGCG
+
@AC=BCCC???B?@@CBB@???>>>>>*?8??>DAABEBCBABCAAA:@@>+9:8>;<;//.
98283988*44449;;9/88:?29:>>5;78333333&399298:6/./DCDDCC';>:ACBDAABB??9::+9<
1444@:?77-3<03368:8755888;:9833)3777'--'--
```

@

```
@X1L6C:03659:02717
GCTTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAATATAGGCATA
GCGCACAGACAGATAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACCATTACCACCACCATCACCATTA
CCACAGGTAACGGTGCGGGCTGACGCGTACAGGAAACACAGAAAAAAGACCCGCCACTGACCAGTGCG
+
???9?BB@<CAA;A8@?:?@@5::BCCCEC;C=CCC8CEJ8DE;AACF>CC?DDCCCBB:B@???9?;B=B=CAA@?;?BCG
CCCCCCBABBBBCCDDAA2:4;@???CAB@AAA9@@AB?C:;;C;CDCCC>ECCAA<AC<CB>DC<AB=CD=C9::A4::>
CC;@@@A?CI@DDAFKDDD:A@CBCDC::::99199+8;4746@CA?)<444/3:4934333-3888//
```

@

```
@X1L6C:02011:02071
TTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAATATAGGCATAGCGCACAGACAGATAAAAATTACAGAGTACA
CAACATCCATGAAACGCATTAGCACCACCATTACCACCACCATCACCATTACCACAGGTAACGGTGCGGGTGACGCGTACAG
GAAACACAGAAAAAAGCCCGCACCTGACAGTGCGGGCTTTTTTTTCGACCAAAGTAACG
+
=0>>>19;;,;;7=CCDADC;?::::::,5;;==4>273:<@BBCF=CDH;@;MMFEED@?>>>:::::*5/55<
;::@:;:BC=BCBB<B@@@D<@@B:;3:::9@<BB=BD=AC;@B:??3::CAC=CD;;;=BBAB>CC;AA;BAAAA9AD@>>
>>?955>4?9499985555558&4<>2:;661499888...88/56666666$;6/.5:8(..+'++
```

@

```
@X1L6C:01333:03005
GCAATGCCAGGCAGGGCATGTACAGCTACGTACGTCTGAGCATCGATCGATGTACAGCTACGTACGTCTGAGCATCGATCGA
TGTACAGCTACGTACGTCTGAGCATCGATCGATGTACAGCTACG
+
555/55/(///(////(/8:9:<=>?<?@:98A??676<:;;@:5555555554444;=4443333;383338<68>>
68=333118311111111113933644588?==<76992---2+++0/
```

…and so on

Coverage

Reads mapped on a reference genome

Ref seq

COVERAGE

Mapped reads

Throughput of benchtop sequencers

	Read lenght	Total time	Total reads	Output
Illumina MiSeq	2 x 300bp	~56 hrs	44-50 M	13.2-15 Gb
Ion Torrent PGM	400bp	~19 hrs	4-5.5 M	1.6-2.2 Gb

e.g. *E. coli* multiplex genomes sequencing

Average length 5 Mb

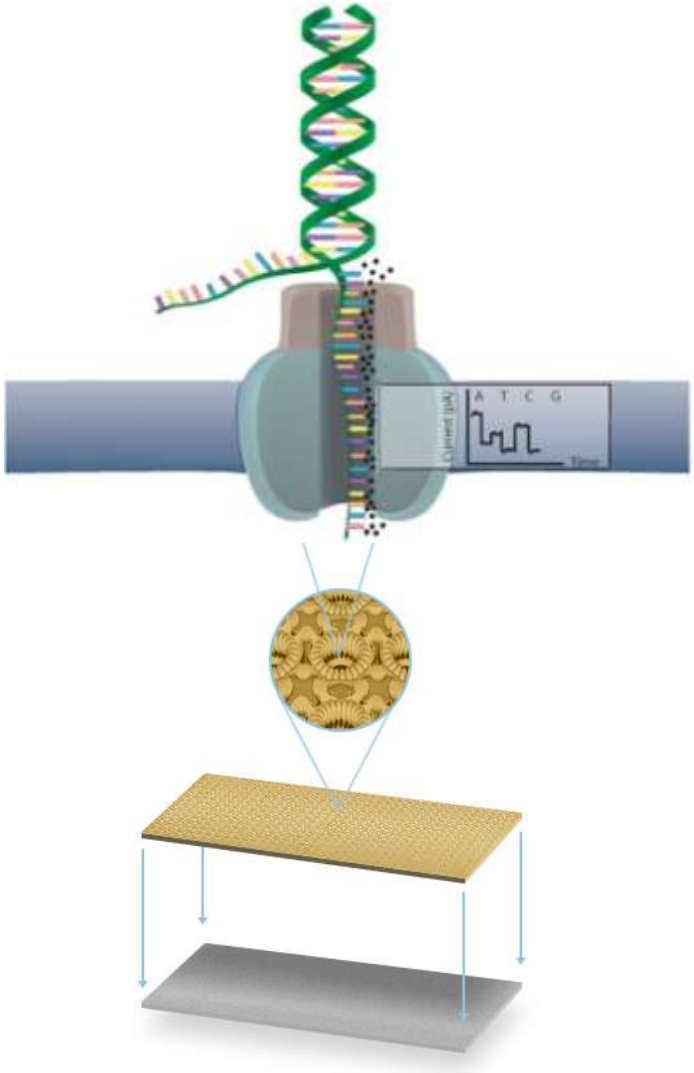
Desired coverage 30X

150 Mb/strain

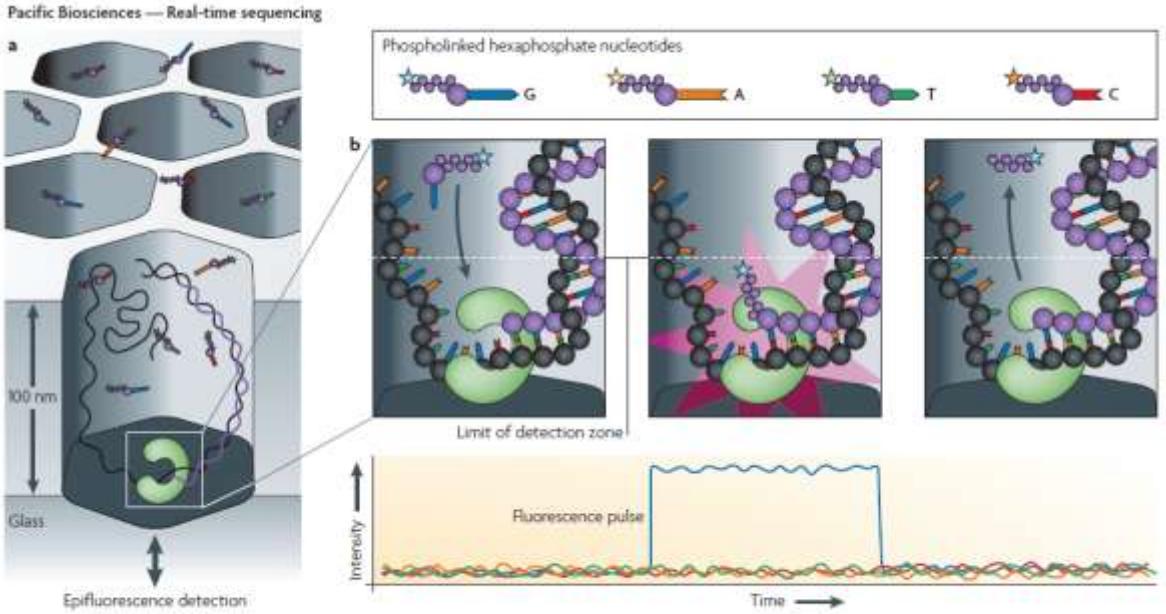
Different systems, different problems

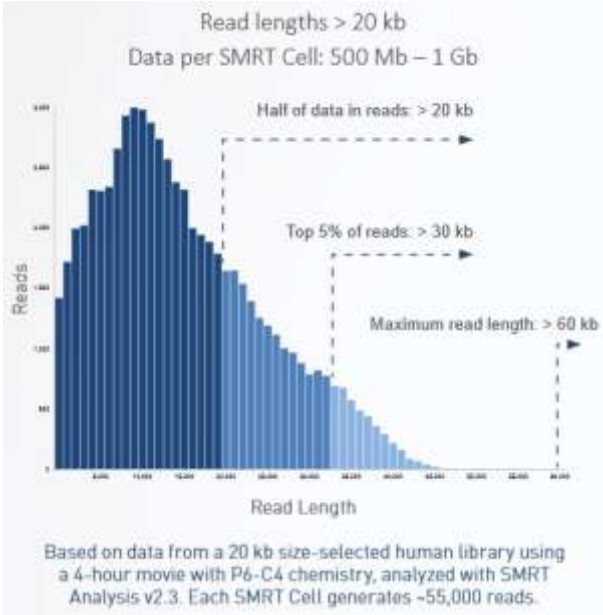
	Illumina	Ion Torrent
Base calling	Blocked fluorescent-labeled dNTPs	pH detection at dNTP incorporation
Critical events	Interference of waves	Enumeration of events of incorporation of identical dNTPs molecules
Errors generated	Substitutions	Length pf homopolymers

MinION - Oxford Nanopore Technologies



Pacific Biosciences





Issues to be addressed

- **Data production still needs to be streamlined**

  Reference laboratories only actively produce data as of today (6 NRLs in our network)

- **Cross-platform compatibility**

  Different platforms = different errors rates and types

- **Intrinsic quality of the sequence reads at the nucleotidic level**

  Filtering algorythms to be developed and harmonized

- **Refinement of existing tools for data analysis and development of new ones**

  Need for new approaches to typing

- **Need for education in bioinformatics**

- **Computationally intense data analysis**

  Accessibility of bioinformatic tools via open-source servers

- **Massive data storage and transfer**

  What data should be stored? Cloud storage?