Next-gen Sequencing Technologies

Bioinformatics Applications (PLPTH813)

Sanzhen Liu

1/31/2017

R

- Data structure

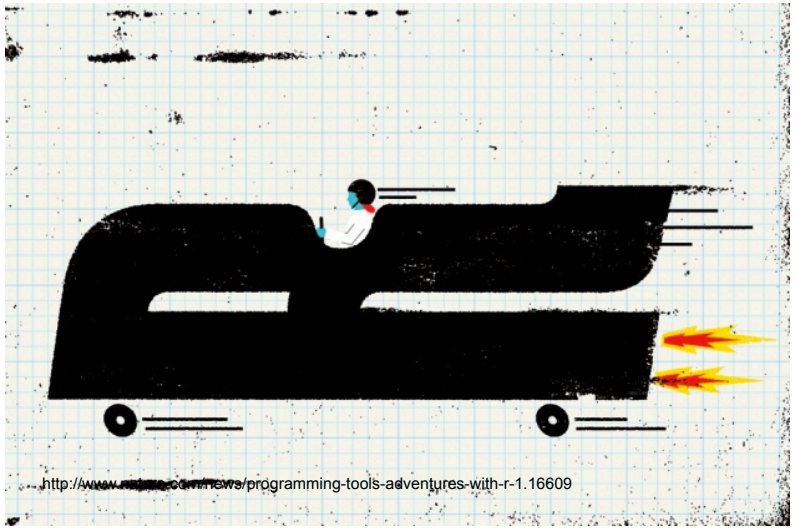
Vector: numeric, character, logical, …
Data frame: 2-dimension table
List: flexible for different types of data

- Data importing and exporting

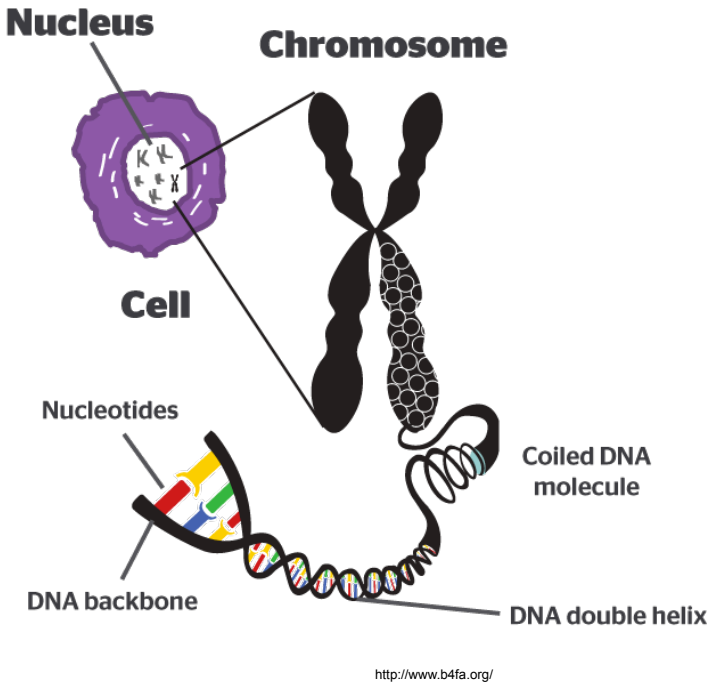
read.table, write.table

- Plotting:

plot, points, lines, abline



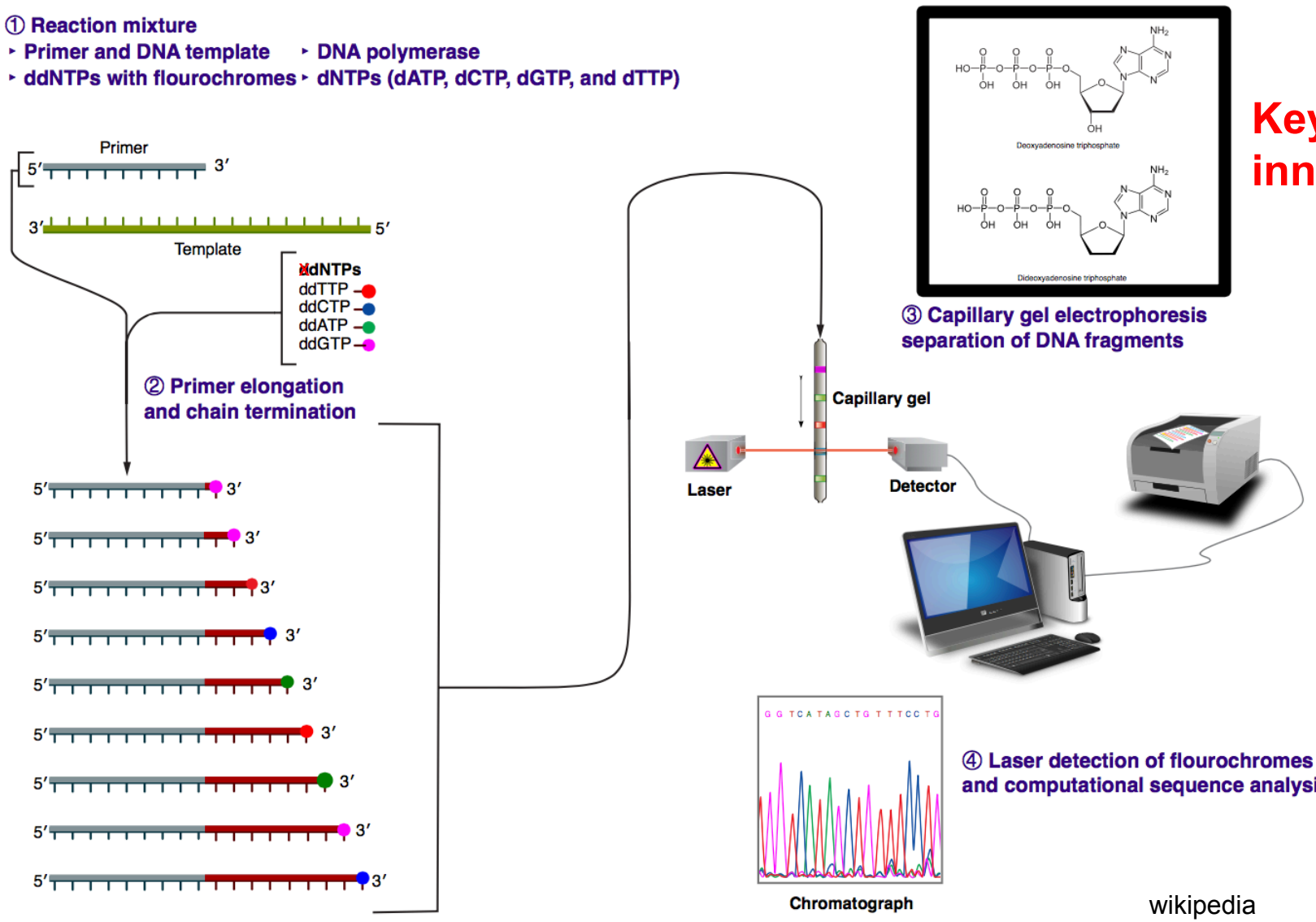
Genome sequencing



DNA sequence

```
   1 ccctgaggct tttcggagcg agctcctcaa atcgcatcca gattttcggg tccgagggaa
  61 ggaggaccct gcgaaagctg cgacgactat cttcccctgg ggccatggac tcggacgcca
 121 gcctggtgtc cagccgcccg tcgtcgccag agcccgatga ccttttctg ccggcccgga
 181 gtaagggcag cagcggcagc gccttcactg ggggcaccgt gtcctcgtcc accccgagtg
 241 actgcccgcc ggagctgagc gccgagctgc gcggcgctat gggctctgcg ggcgcgcatc
 301 ctggggacaa gctaggaggc agtggcttca agtcgtcctc gtccagcacc tcgtcgtcta
 361 cgtcgtcggc ggctgcgtcg tccaccaaga aggacaagaa gcaaatgaca gagccggagc
 421 tgcagcagct gcgtctcaag atcaacagcc gcgagcgcaa gcgcatgcac gacctcaaca
 481 tcgccatgga tggcctccgc gaggtcatgc cgtacgcaca cggcccttcg gtgcgcaagc
 541 tttccaagat cgccacgctg ctgctggcgc gcaactacat cctcatgctc accaactcgc
 601 tggaggagat gaagcgactg gtgagcgaga tctacggggg ccaccacgct ggcttccacc
 661 cgtcggcctg cggcggcctg gcgcactccg cgcccctgcc cgccgccacc gcgcacccgg
 721 cagcagcagc gcacgccgca catcaccccg cggtgcacca ccccatcctg ccgcccgccg
 781 ccgcagcggc tgctgccgcc gctgcagccg cggctgtgtc cagcgcctct ctgcccggat
 841 ccgggctgcc gtcggtcggc tccatccgtc caccgcacgg cctactcaag tctccgtctg
 901 ctgccgcggc cgccccgctg gggggcgggg gcggcggcag tggggcgagc gggggcttcc
 961 agcactgggg cggcatgccc tgcccctgca gcatgtgcca ggtgccgccg ccgcaccacc
1021 acgtgtcggc tatgggcgcc ggcagcctgc cgcgcctcac ctccgacgcc aagtgagccg
1081 actggcgccg gcgcgttctg gcgacagggg agccaggggc cgcggggaag cgaggactgg
1141 cctgcgctgg gctcgggagc tctgtcgcga ggaggggcgc aggaccatgg actgggggtg
1201 gggcatggtg gggattccag catctgcgaa cccaagcaat gggggcgccc acagagcagt
1261 ggggagtgag gggatgttct ctccgggacc tgatcgagcg ctgtctggct ttaacctgag
1321 ctggtccagt agacatcgtt ttatgaaaag gtaccgctgt gtgcattcct cactagaact
```

Sanger sequencing technology



Key innovation

wikipedia

DNA sequencing technology

1st-gen sequence (Sanger)

1980 Nobel Prize

next-gen sequence (NGS)

>sample
TGCCGCCTACTAACCGGTTCTGAGAGTTCTGAGATG
AGAGAATGCCTACTAACCGGTTCTGAGAATGCCTA
CTAACCGATGCCTACTAACCGGTTCTGAGAGTTCTG
AGCTGAGATGCCTACTAACCGGTTCTGAGAATGCCT
ACTAACCGATGCCTACTAACCGGTTCTGAGAGTTCT
GAGATGAGAGAATGCCTACTAACCGGTTCTGAGAA
TGCCTACTAACCGATGCCTACTAACCGGTTCTGAGA
GTTCTGAGCTATGCCTACTAACCGGTTCTGAGAATG
CCTACTAACCGATGCCTACTAACCGGTTCTGAGAGT
TCTGAGATGAGAGAATGCCTACTAACCGGTTCTGA
GAATGCCTACTAAATGCCTACTAACCGGTTCTGAGA
ATGCCTACTAACCGATGCCTACTAACCGGTTCTGAG
AGTTCTGAGATGAGAGAATGCCTACTAACCGGTTCT
GAGAATGCCTACTAACCGATGCCTACTAACCGGTTC
TGAGAGTTCTGAGCTCCGATGCCTACTAACCGGTTC
TGAGAGTTCTGAGCTAATACTAACCGGTTATGCCTA
CTAACCGGTTCTGAGAATGCCTACTAACCGGTTCTG
AGAATGCCTACTAACCGATGCCTACTAACCGGTTCT
GAGAGTTCTGAGATGAGAGAATGCCTACTAACCGG
TTCTGAGAATGCCTACTAACCGATGCCTACTAACCG
GTTCTGAGAGTTCTGAGCTGAGAA

800 letters

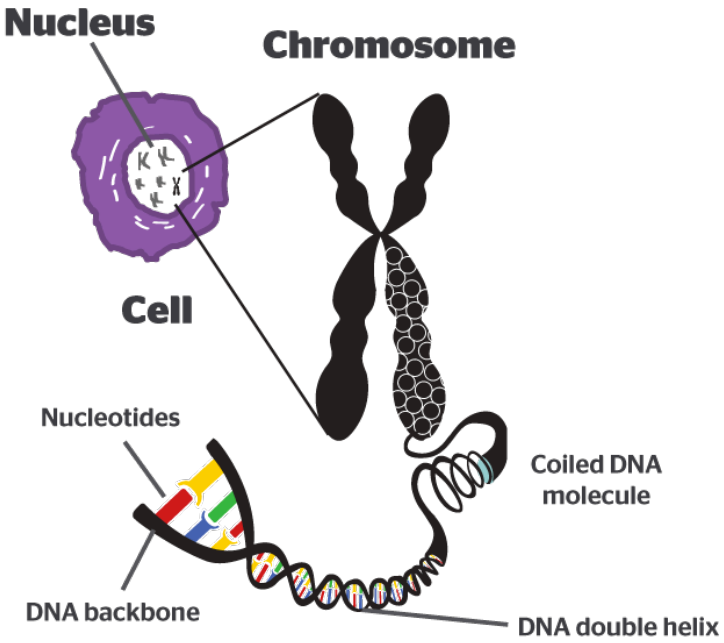
billions of letters

Major NGS technologies in market

Sequencing cost

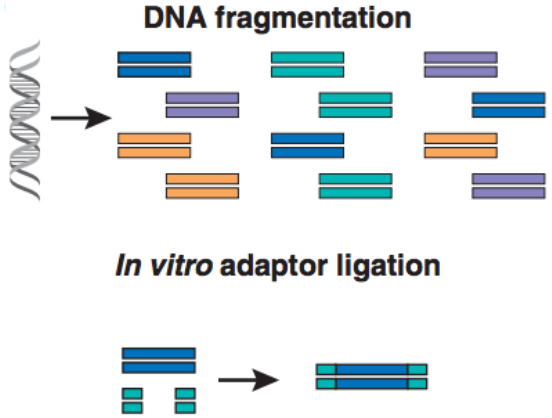
1970's Sanger sequencing

Data source: genome.gov/sequencingcosts



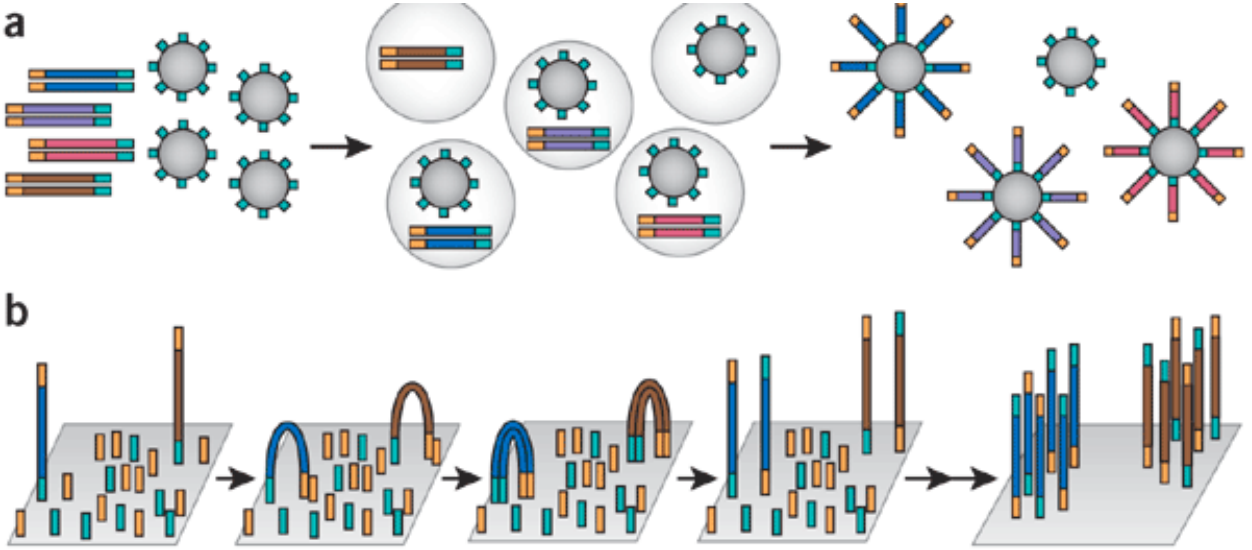
Before single molecular sequencing technology, **amplification/cloning** of a single nucleotide molecule is needed for sequencing.

DNA amplification/cloning



Adaptor ligation

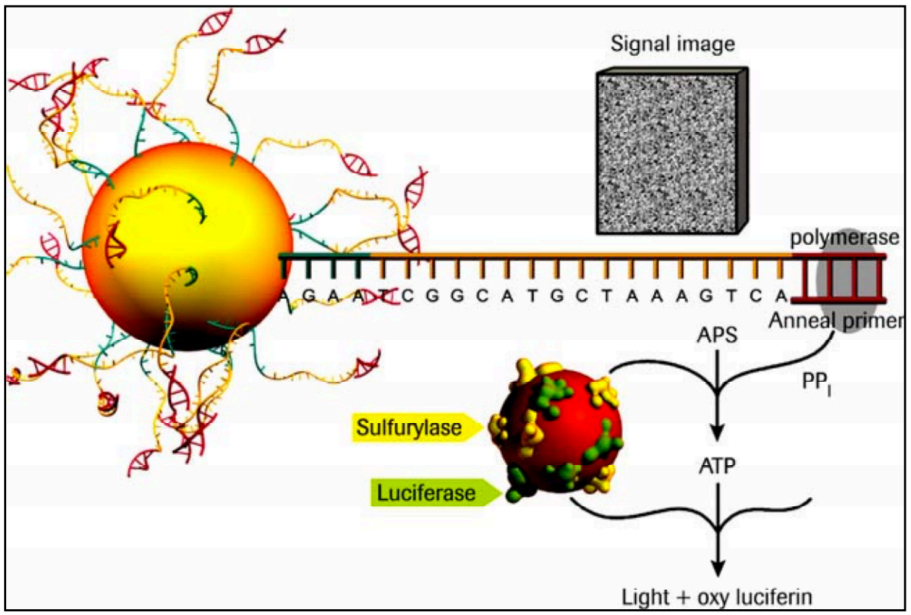
Water-in-oil emulsion PCR
(454 and Ion Torrent)



Bridge PCR on slides
(Illumina)

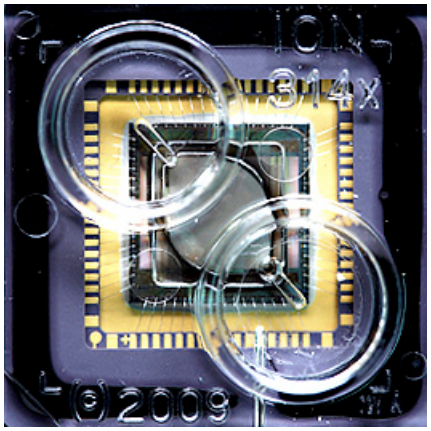
Nature Biotechnology, 2008, 26: 1135-45

454 and Ion Torrent



454 technology, Nature 2005, 437: 376-380

1. Sequencing by synthesis
2. Pyrosequencing (454)



Ion Torrent video

1. Ion Torrent technology is similar to 454 technology
2. The signal is H^+ rather than pyrophosphate

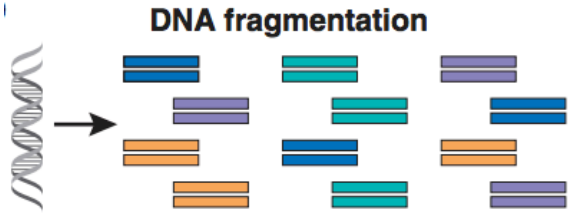
Ion Torrent & 454

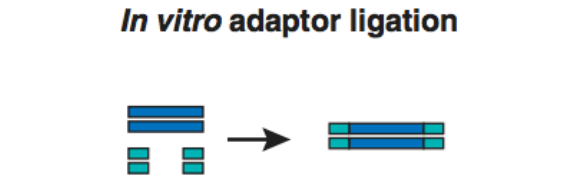
Life technology Ion Torrent & Roche 454:
Record signal per **nucleotide type**:

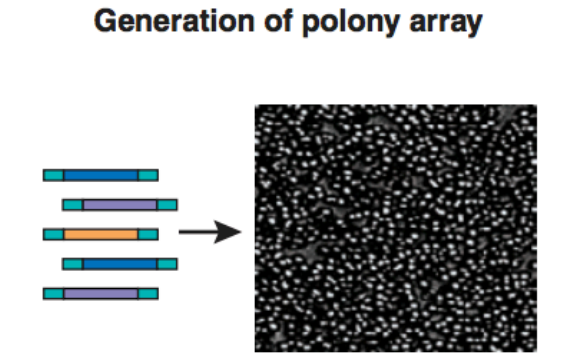
A T G C A A A A

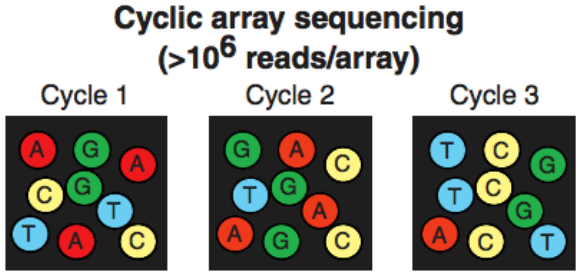
A T G C A A A A

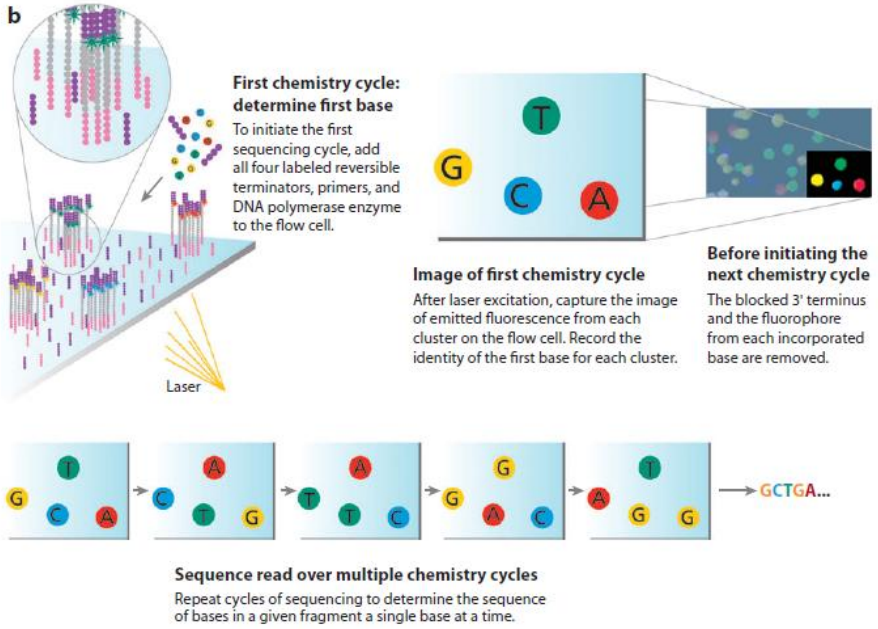
Illumina sequencing











philos.biol.mun.ca

Two key technologies:

1. Bridge PCR
2. Reversible terminator chemistry

Illumina *versus* Ion Torrent & 454

Illumina
Record signal per **nucleotide position**:

A T G C A A A A

A T G C A A A A

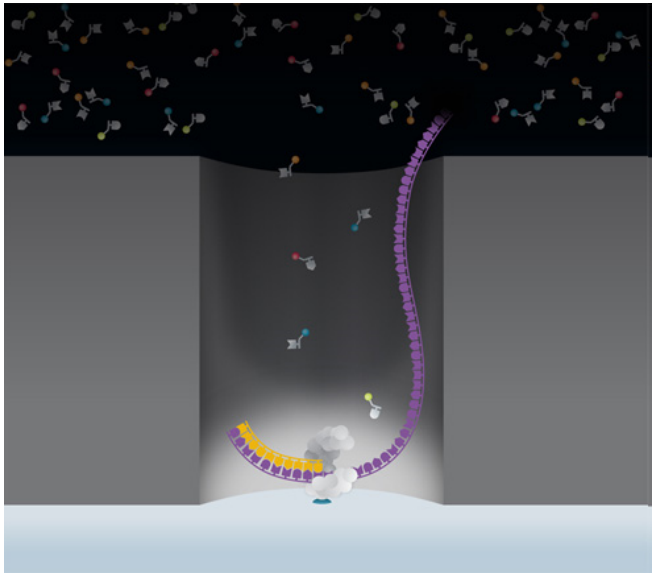
Life technology Ion Torrent & Roche 454:
Record signal per **nucleotide type**:

A T G C A A A A

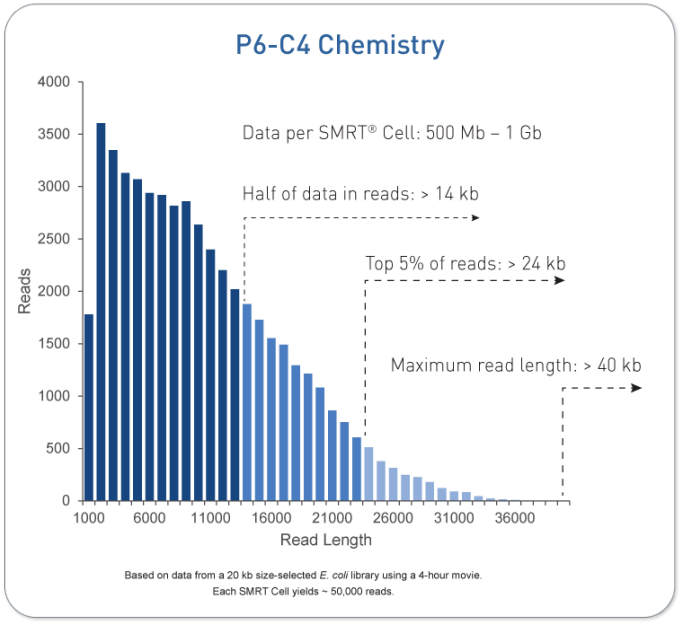
Sequencing errors at homopolymers

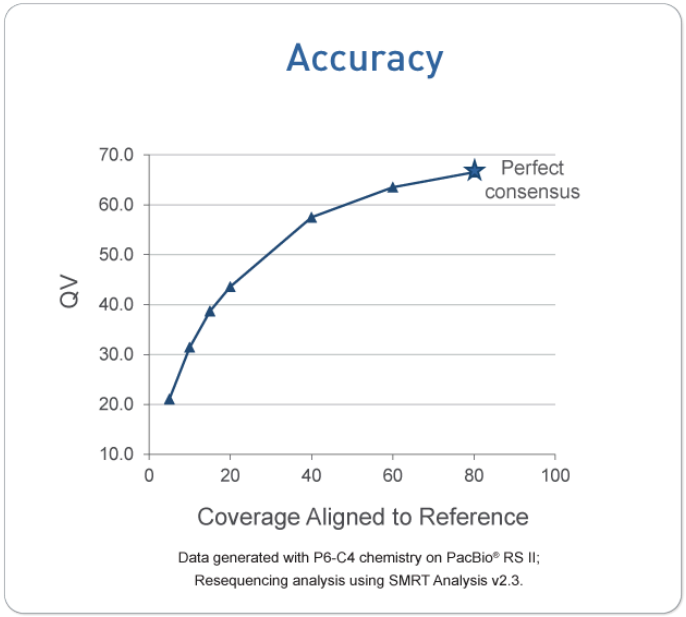
When the single molecular sequencing technology is ready, **amplification or cloning** is not necessary.

PacBio – Single Molecule Real Time (SMRT)



[PacBio tech video]





- Single molecule sequencing
- no amplifications required
- up to 70+ kbp sequencing
- Moderate sequencing throughput
- high sequencing error rate (~15%, random, no-context-specific errors)

pacbio.com

PacBio Sequencing procedure

- Complete template preparation in as few as 6 hrs

- Accepts a variety of sample types and insert sizes
- Low template options available from 10 - 100 ng

- Size selection options to enrich for longest inserts
- Multiplexing and barcoding solutions available

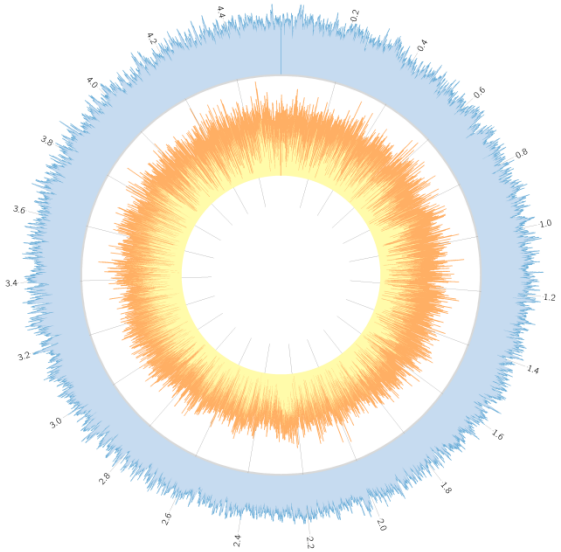
PacBio RS II

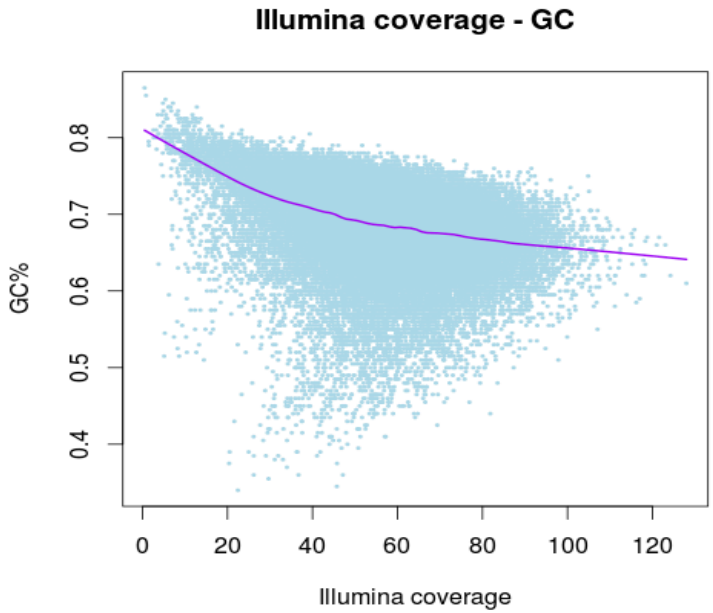
Sequel System

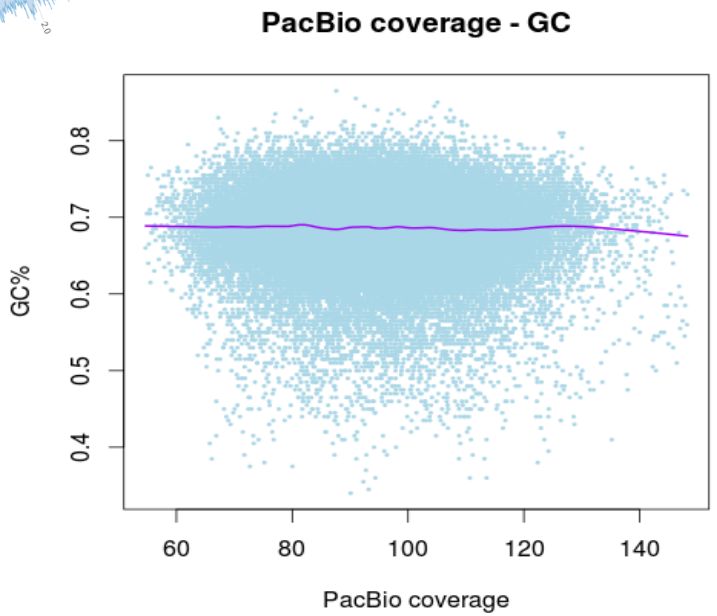
- Rapid sequence time (0.5 to 6 hrs)
- Serially process up to 16 SMRT Cells in a single run with walkaway automation

http://www.pacb.com/wp-content/uploads/2015/09/Revolutionize-Genomics-with-SMRT-Sequencing.pdf

Less GC-related biases

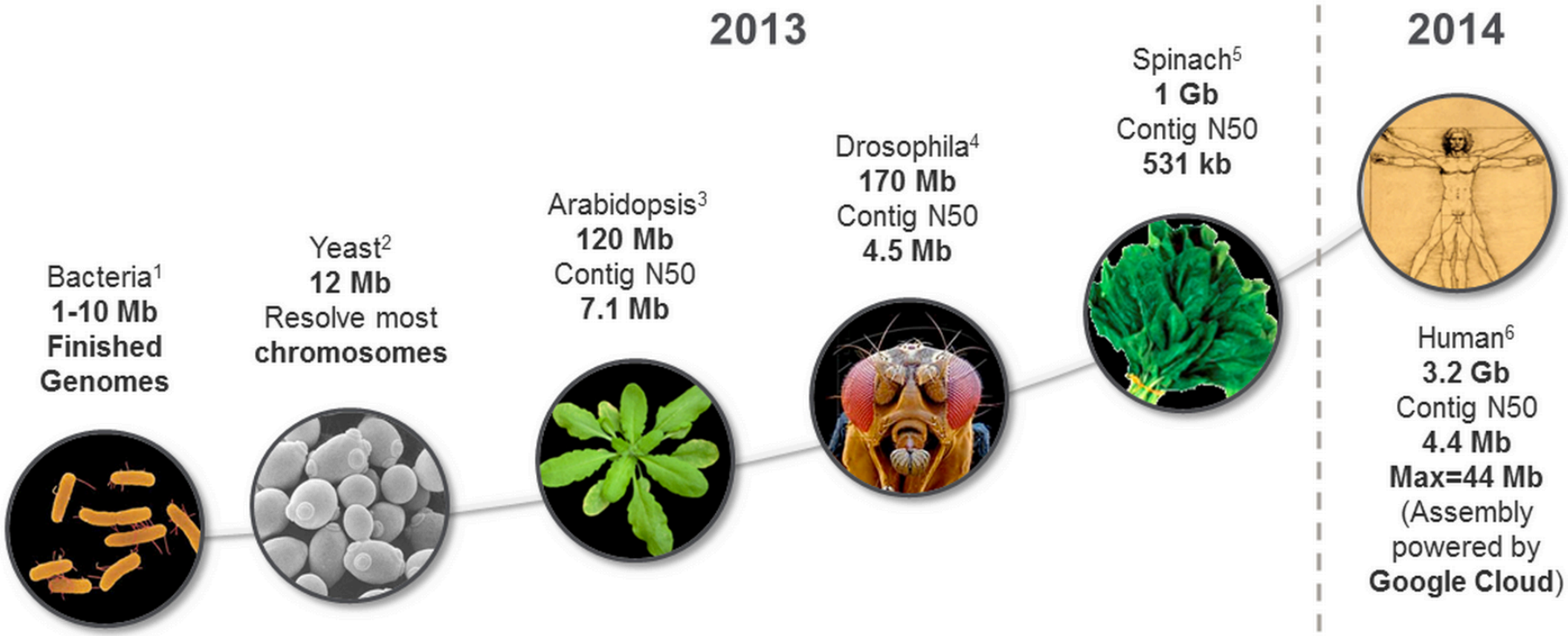






Peng *et al*., BMC Genomics, 2016, 17:21

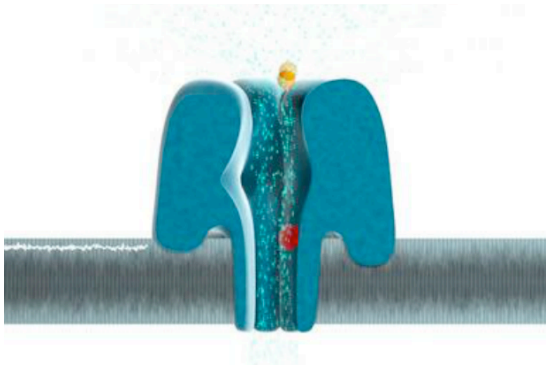
PacBio for genome assembly



PacBio has solved *de novo* assemblies of most bacterial genomes and it will solve assemblies of small “simple” genomes (e.g., <500 Mbp) with increasing read length and improved sequencing quality.

New Platform – **Oxford Nanopore**

A promising technology



As each nucleobase passes through the pore the current is affected and this change allows sequence to be read out.

- Single molecular sequencing
- No amplifications
- **Long reads (kbp)**
- **Error rate is high (~30%)**

MinION

1. USB disposable sequencer
2. Hundreds of Mb in several hours

PromethION (NEW)

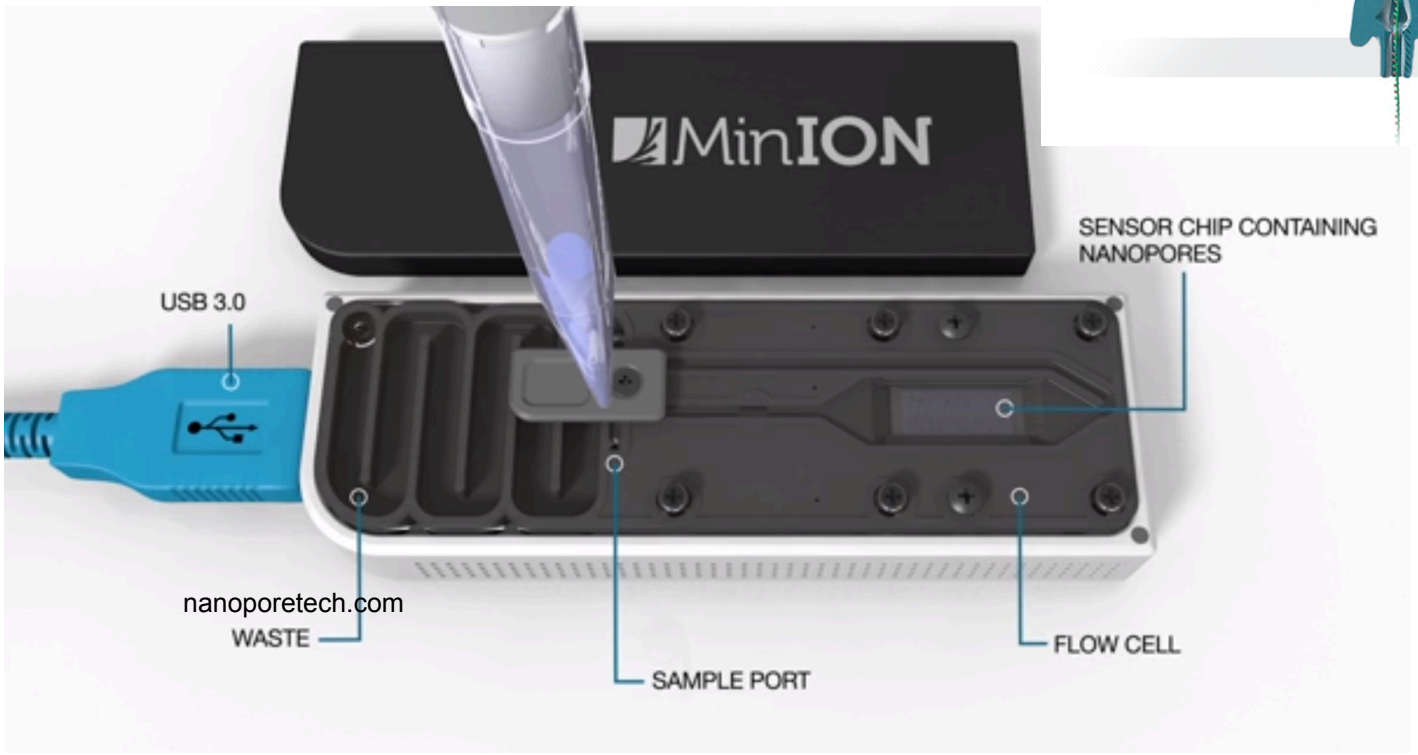
1. "MinION cluster"?
2. High-throughput (1Tb output)

nanoporetech.com

MinION specification (1/31/2017)

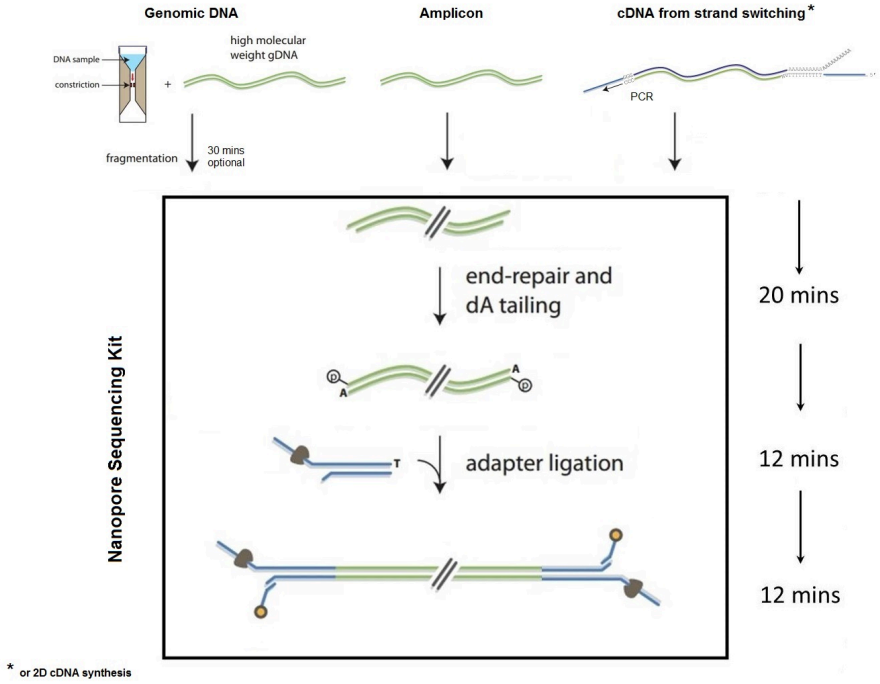
Item	**MinION**
Number of channels available for sequencing	Up to 512
Sample input Requirement PCR Free	10pg - 1µg
Sample preparation time 1D	10 minutes
Sample preparation time 2D	90 minutes
Run time	1 minute - 48 hours
Number of reads at 10Kb at standard speed (250bps)	Up to 2.2M
Read length	up to hundreds of kb
Flow Cell Cost (depending on order type and volume)	$500 - $900

Potentials for Nanopore

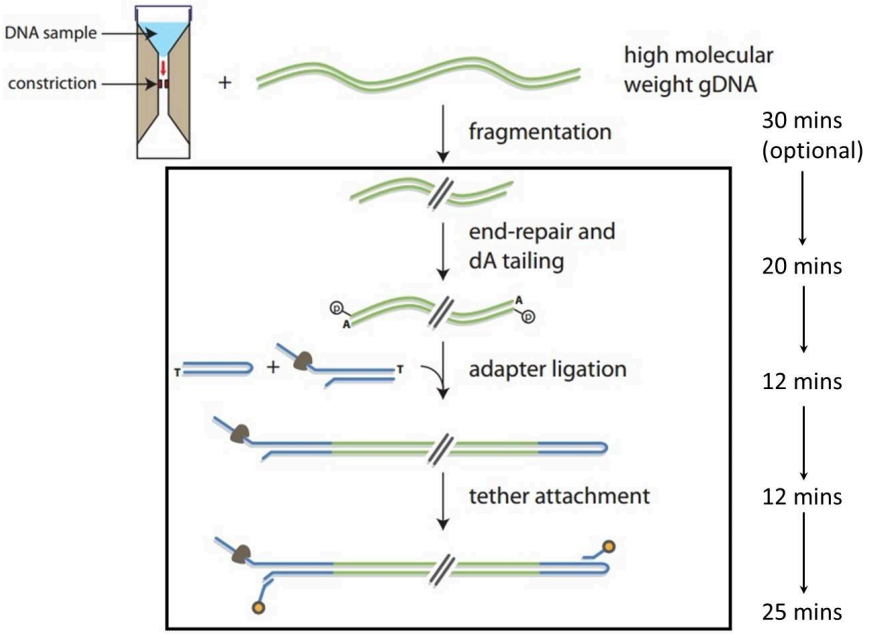


smartphone
+
Nanopore
(SmidgION)

Nanopore library preparation



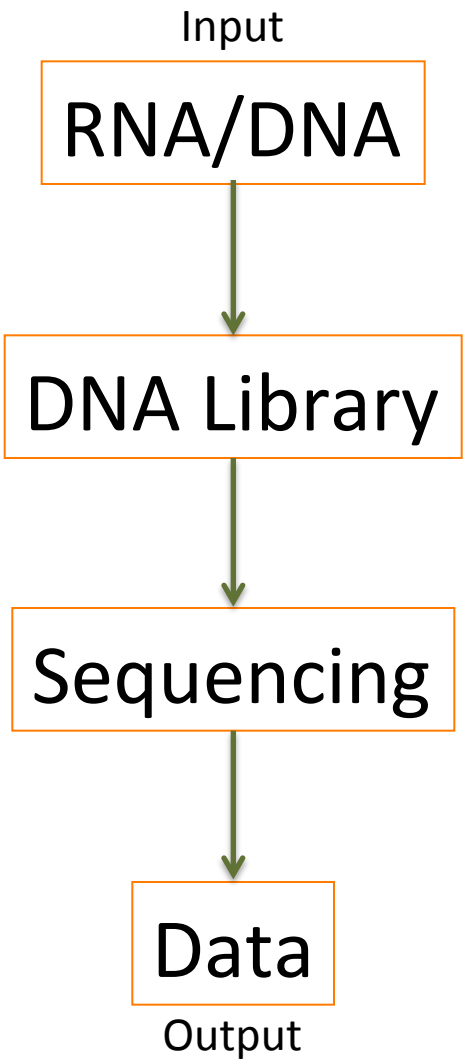
1D



2D

nanopore.com

COMMON in all NGS platforms

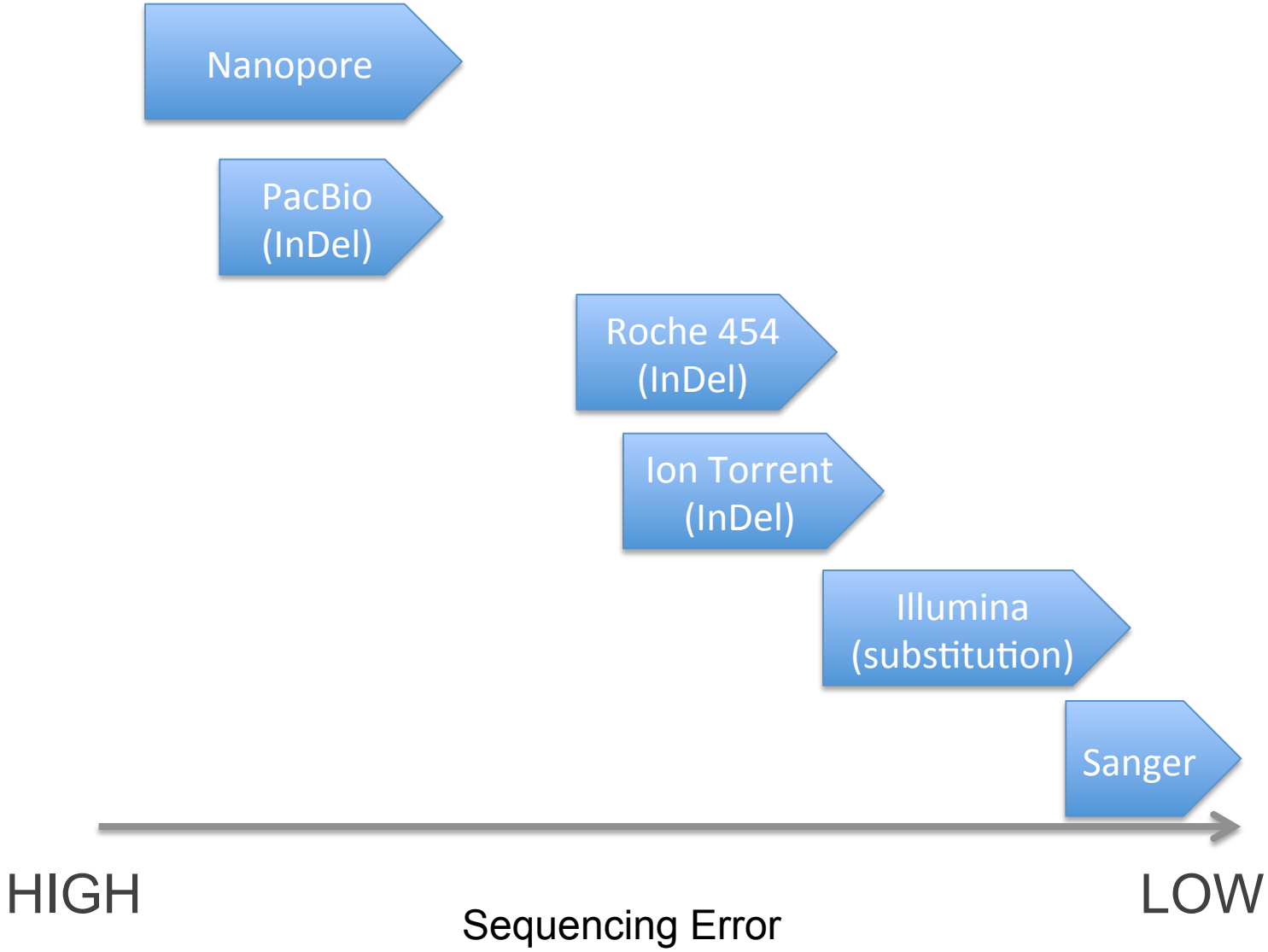


Adaptor is required for library preparation

Thousands to hundreds of millions of fragments are sequenced in parallel

Standard data (FASTQ) are output

Sequencing error rates



Applications of NGS

1. Whole-genome sequencing/re-sequencing / target-region sequencing (Assembly, Variant discovery)
2. Genome-reduction sequencing (GBS, RAD-Seq)
3. RNA-Seq: differential expression, alternative splicing and variant discovery
4. Small RNA-Seq
5. ChIP-Seq: Elucidate DNA-protein interaction
6. Metagenomics
7. Others

Case study – Discussion (3 min)

1. *De novo* assembly of a strain of *E.coli*

2. Human whole genome sequencing for SNP discovery

Which platform(s)?

Sequencing depth?

Sequence platforms

Illumina (MiSeq, NextSeq, HiSeq)

very high throughput, up to 2x300 bp, and high accuracy (<1%)

Proton (Ion Torrent)

high throughput, up to 300-500 bp, but high errors at homopolymer regions

PacBio

Moderate sequencing throughput, very long (up to 70kb+), but high errors (15%)

Nanopore

Moderate sequencing throughput, very long (up to 300kb), but high errors (10-30%)

@anne_churchland (twitter)

Experimental design

- Goal
- Platform
- Read length
- Rate and type of sequence errors
- Sequencing depth
- Replication
- Control
- Budget

Platform	Templates	Signal	Read length	Run time	reads per run	Error type	Error rate
Illumina Miseq	PCR or PCR-free	fluorescent	up to 2x300	1-2 days	Up to 10 Gb	substitutions	~0.1-1%
Illumina Hiseq	PCR or PCR-free	fluorescent	up to 2x250	days	Hundreds of Gb	substitutions	~0.1-1%
Ion Torrent	PCR	H+	300-500	2 hours	10 Gb?	InDel	>1%
PacBio	Amplification not required	fluorescent	Average >5,000	30min	500 Mb – 1 Gb	InDel	~15%
Nanopore	Amplification not required	Electronic flow change	>1,000	hours	? Mb per MinION	Del?	~30%

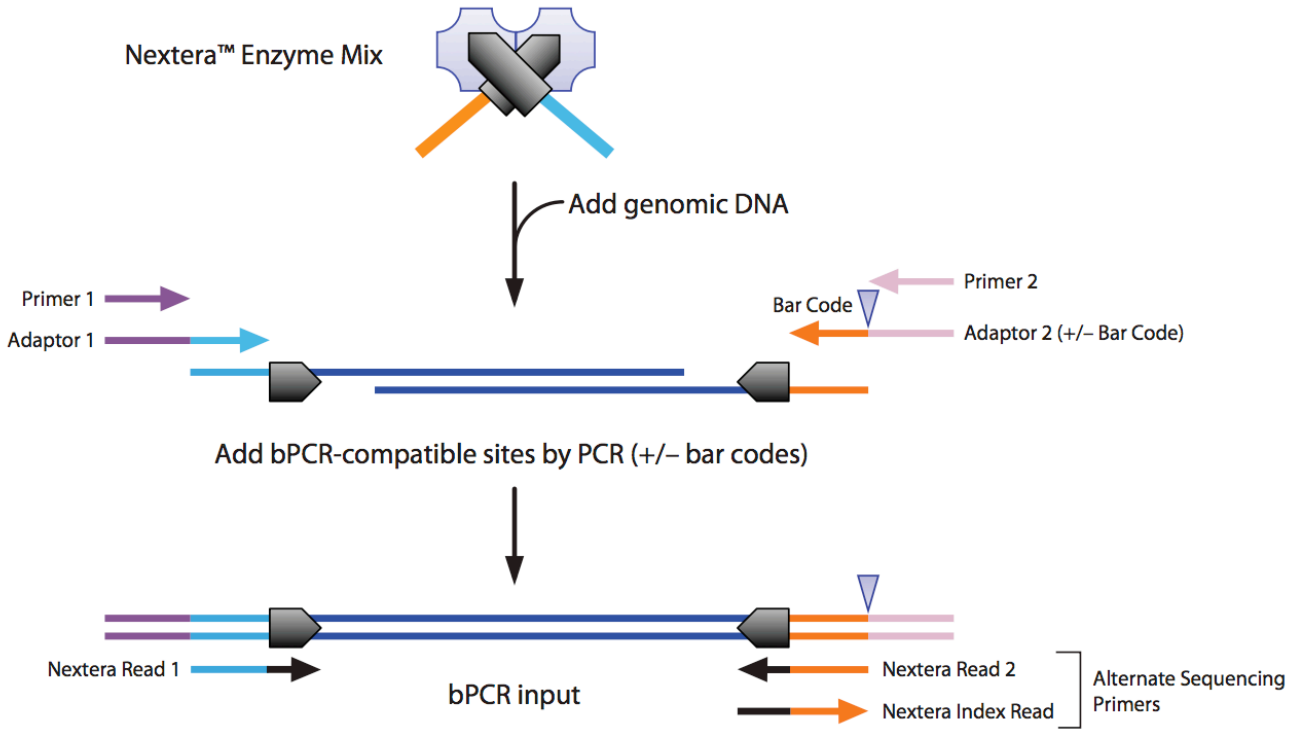
Illumina platforms and terminologies

Illumina video

1. Library preparation
2. Single-ends and paired ends
3. Reads
4. Instruments

Library preparation - Nextera

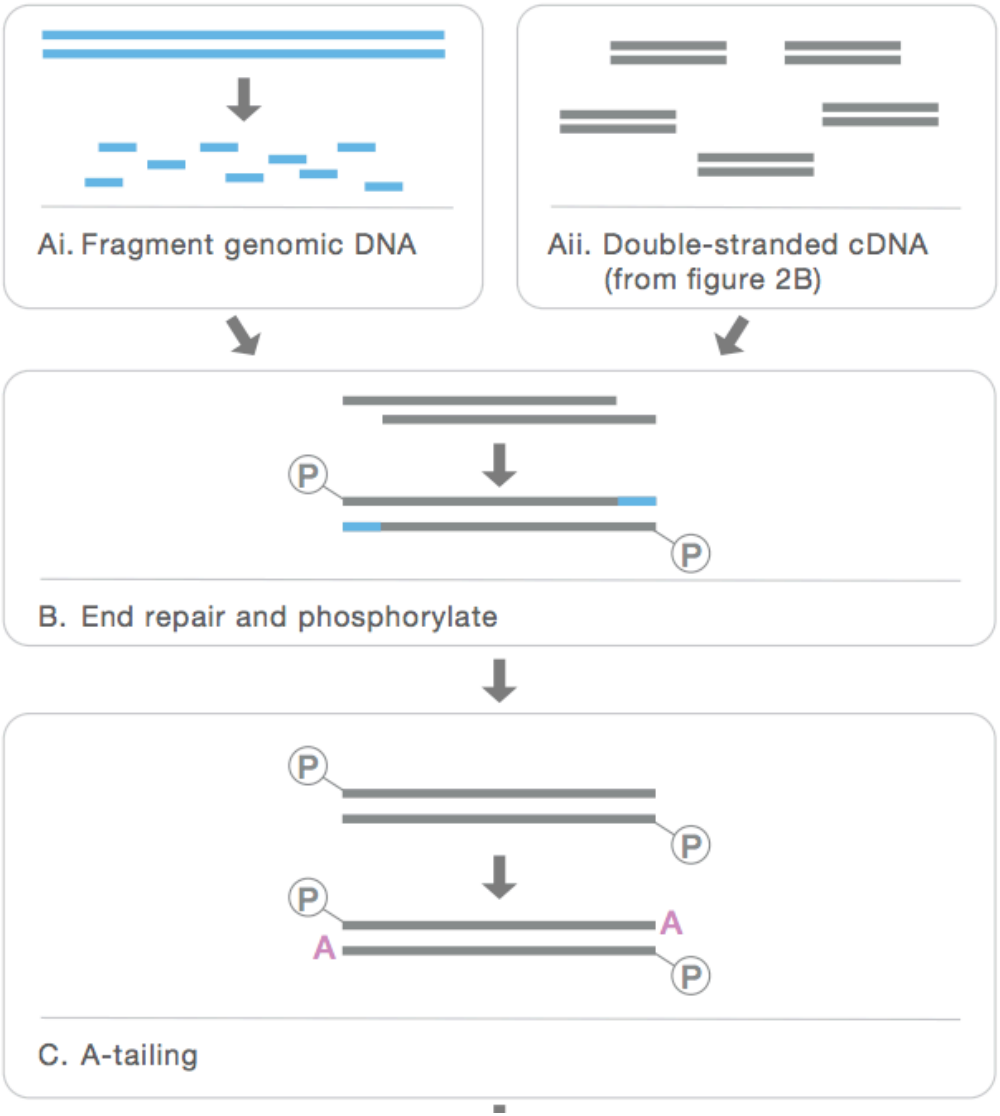
Nextera technology employs in vitro transposition to simultaneously fragment and tag DNA in a single-tube reaction, and prepare sequencer-ready libraries in under 2 hours.



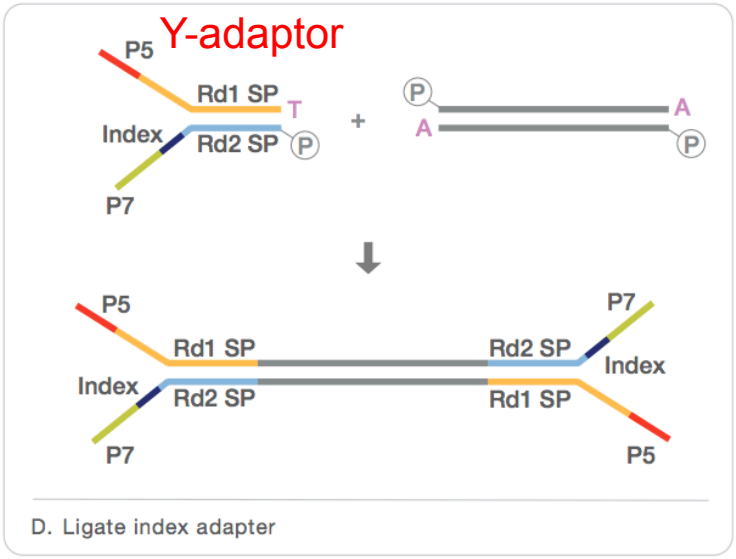
http://www.epibio.com

Library preparation – Y-adaptor method

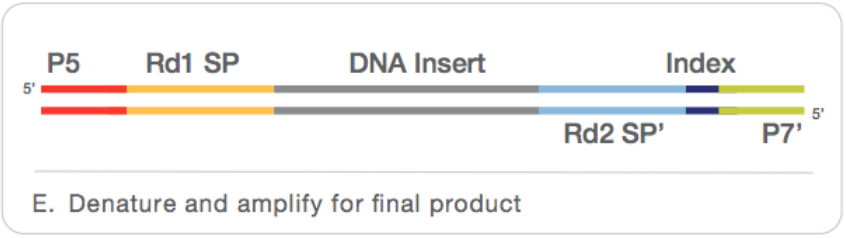
a.



b.



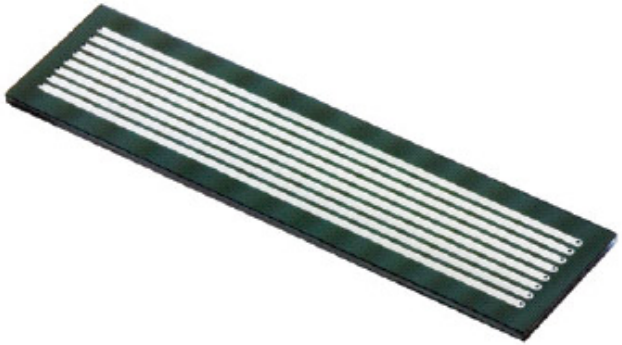
Final product



From TruSeq Manual

Multiplexing (DNA barcode/Index)

Flowcell lane

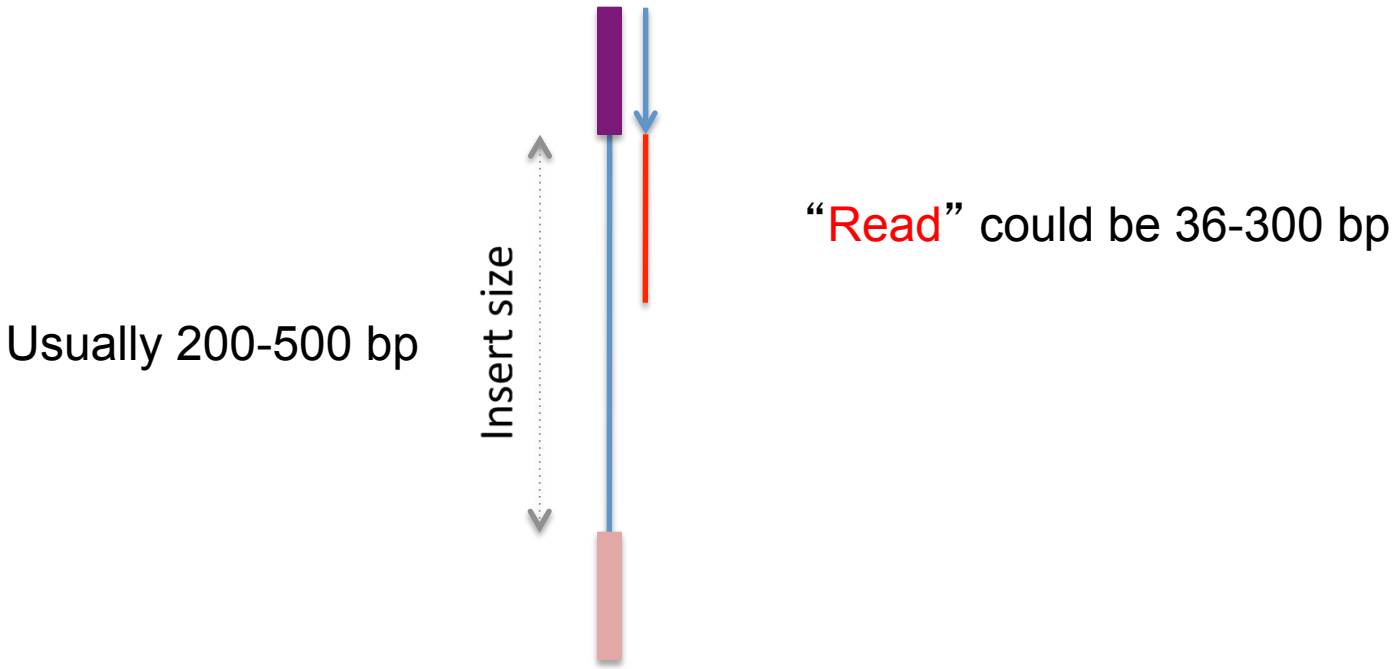


- per lane's data are more than needed in many cases

- Multiplexing: To put multiple samples in a lane via using **DNA barcodes** to distinguish samples

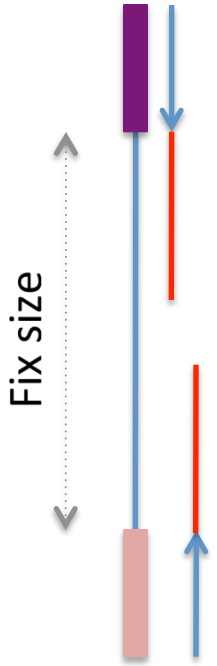
Single-end sequencing

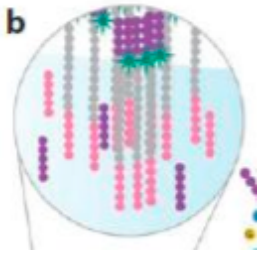
A single read is generated for each template/cluster



Paired-end sequencing

Two reads are generated for each template/ cluster; the 1st is from one end with one primer, the 2nd is for the other end with the other primer.





Illumina Sequencers

NovaSeq

NovaSeq5000 & 600 output (as of 1/31/2017)

Sequencing Output per Flow Cell

	NovaSeq 5000 and 6000 Systems		NovaSeq 6000 System	
Flow Cell Type	S1*	S2	S3*	S4*
2 x 50 bp	up to 167 Gb	280–333 Gb	NA**	NA**
2 × 100 bp	up to 333 Gb	560–667 Gb	NA**	NA**
2 × 150 bp	up to 500 Gb	850–1000 Gb	up to 2000 Gb	up to 3000 Gb

Specifications based on Illumina PhiX control library at supported cluster densities.

* The NovaSeq 5000 System, NovaSeq 5000 System Upgrade, and NovaSeq Reagent Kits with S1, S3, or S4 flow cells will be available later in 2017.

** NA: not applicable

Reads Passing Filter

	NovaSeq 5000 and 6000 Systems		NovaSeq 6000 System	
Flow Cell Type	S1*	S2	S3*	S4*
	up to 1.6 B	2.8–3.3 B	up to 6.6 B	up to 10 B

Data - FASTQ

Standard data format - FASTQ

@HWI-EAS225:3:1:2:854#0/1
GGGGGGAAGTCGGCAAAATAGATCCGTAACTTCGGG
+HWI-EAS225:3:1:2:854#0/1
a`abbbbabaabbbababb^`[aaa`_N]b^ab^``a

http://en.wikipedia.org/wiki/FASTQ_format

```
@HWUSI-EAS100R:6:73:941:1973#0/1
```

HWUSI-EAS100R	the unique instrument name
6	flowcell lane
73	tile number within the flowcell lane
941	'x'-coordinate of the cluster within the tile
1973	'y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
/1	the member of a pair, /1 or /2 *(paired-end or mate-pair reads only)*

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

EAS139	the unique instrument name
136	the run id
FC706VJ	the flowcell id
2	flowcell lane
2104	tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)*
Y	Y if the read fails filter (read is bad), N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	index sequence