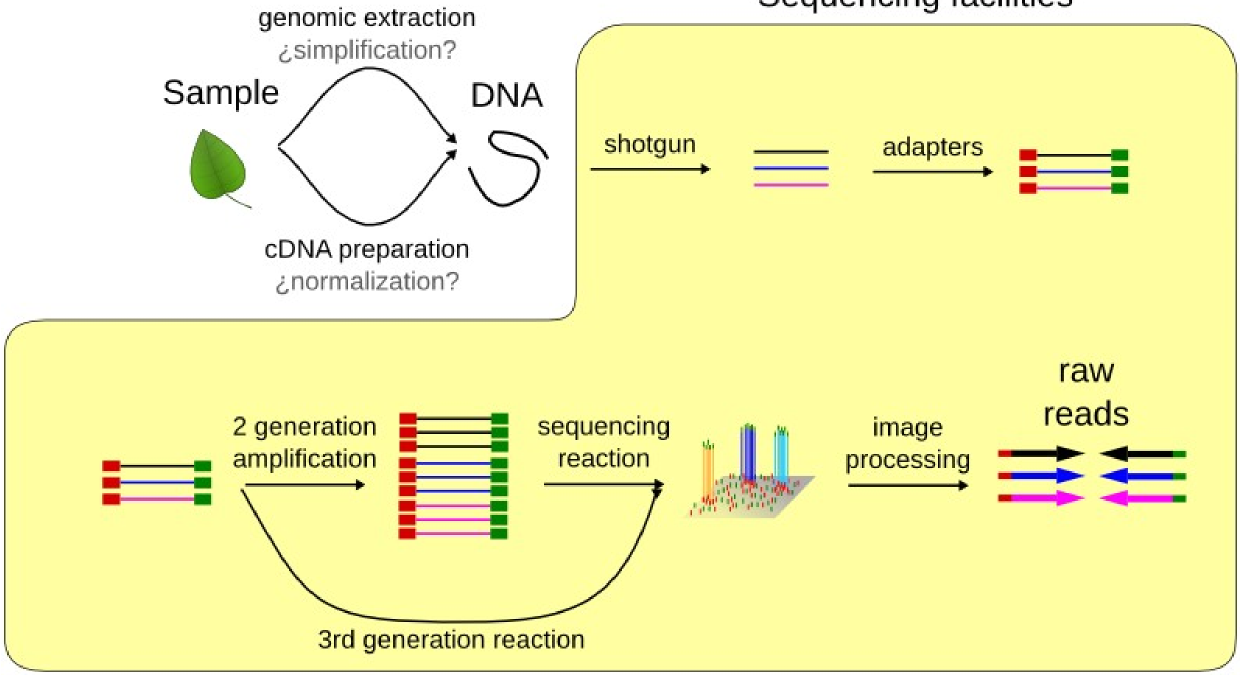
# Sequencing technologies

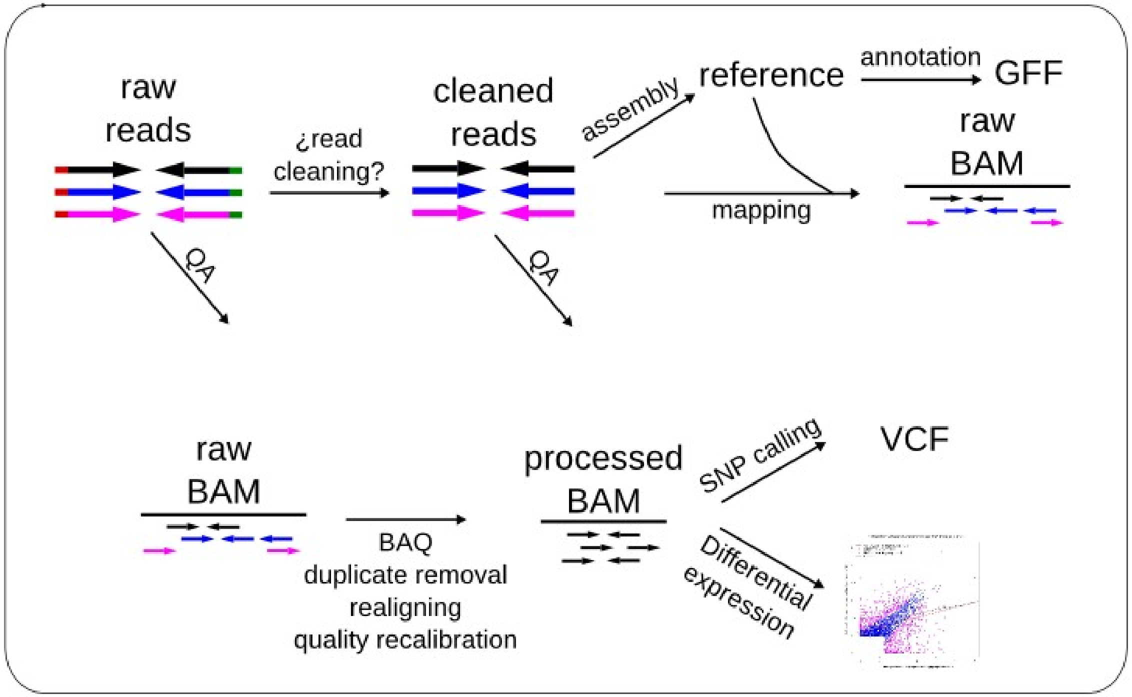
Jose Blanca
COMAV institute
bioinf.comav.upv.es

## Sequencing facilities



genomic extraction
¿simplification?
Sample
DNA
cDNA preparation
¿normalization?
shotgun
adapters
2 generation amplification
sequencing reaction
image processing
raw reads
3rd generation reaction

## Sequence analysis



raw reads
¿read cleaning?
cleaned reads
QA
QA
assembly
reference
annotation
GFF
mapping
raw BAM
raw BAM
BAQ
duplicate removal
realigning
quality recalibration
processed BAM
SNP calling
VCF
Differential expression

# Outline

Sequencing technologies:

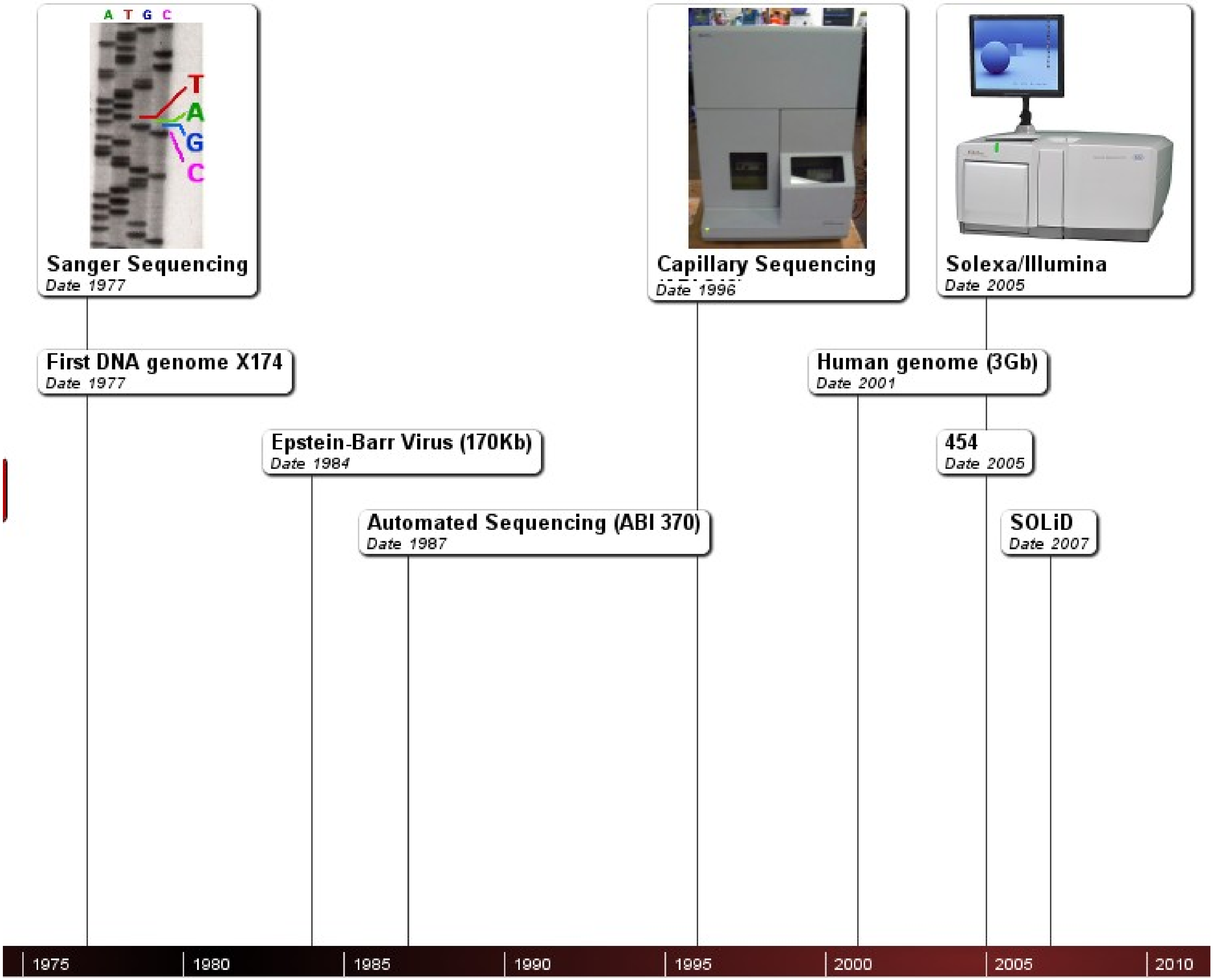
- Sanger

- 2nd generation sequencing:
  - 454
  - Illumina
  - SOLiD
  - Ion Torrent

- 3er generation sequencing:
  - PacBio
  - Nanopore

- General considerations

Reducing the complexity



**Sanger Sequencing**
*Date 1977*

**First DNA genome X174**
*Date 1977*

**Epstein-Barr Virus (170Kb)**
*Date 1984*

**Automated Sequencing (ABI 370)**
*Date 1987*

**Capillary Sequencing**
*Date 1996*

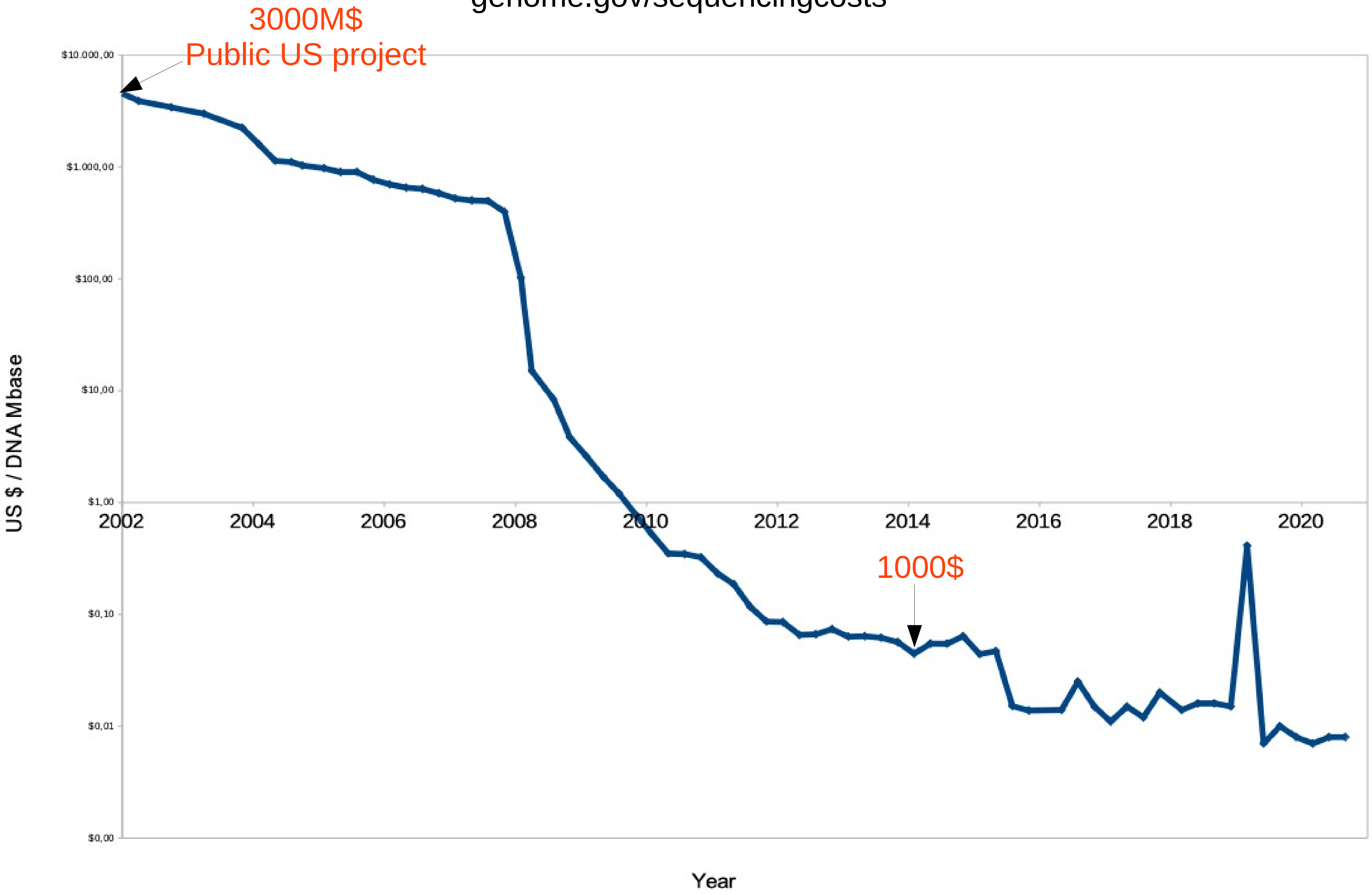
**Human genome (3Gb)**
*Date 2001*

**Solexa/Illumina**
*Date 2005*

**454**
*Date 2005*

**SOLiD**
*Date 2007*

genome.gov/sequencingcosts



3000M$
Public US project

1000$

# Sanger sequencing

# Sanger sequencing

Traditional DNA sequencing method

Ideal for small sequencing projects

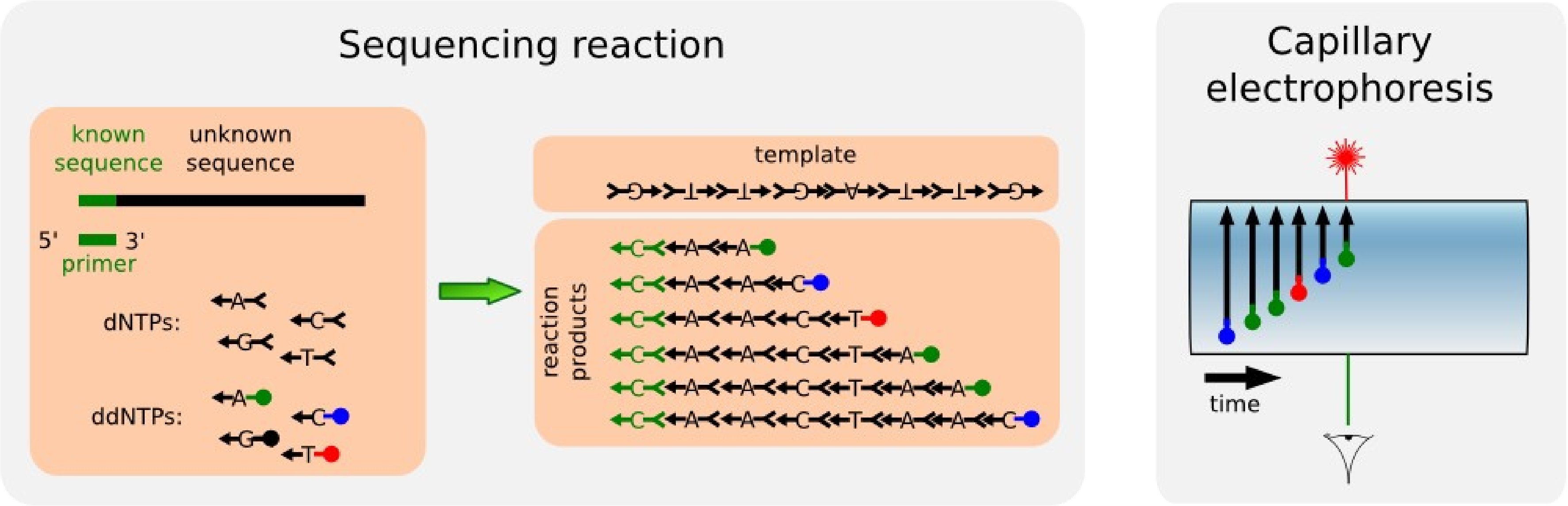
Read length around 600-800 bp

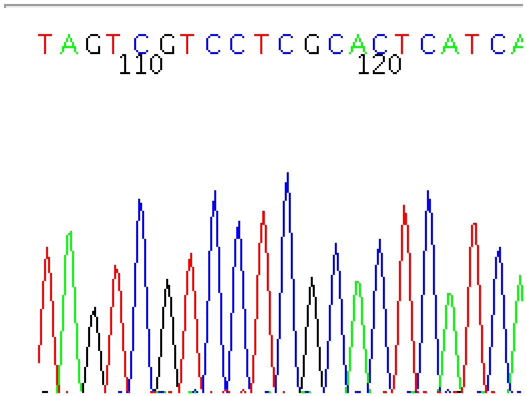
Around 5-10$ per reaction

384 reactions in parallel at most

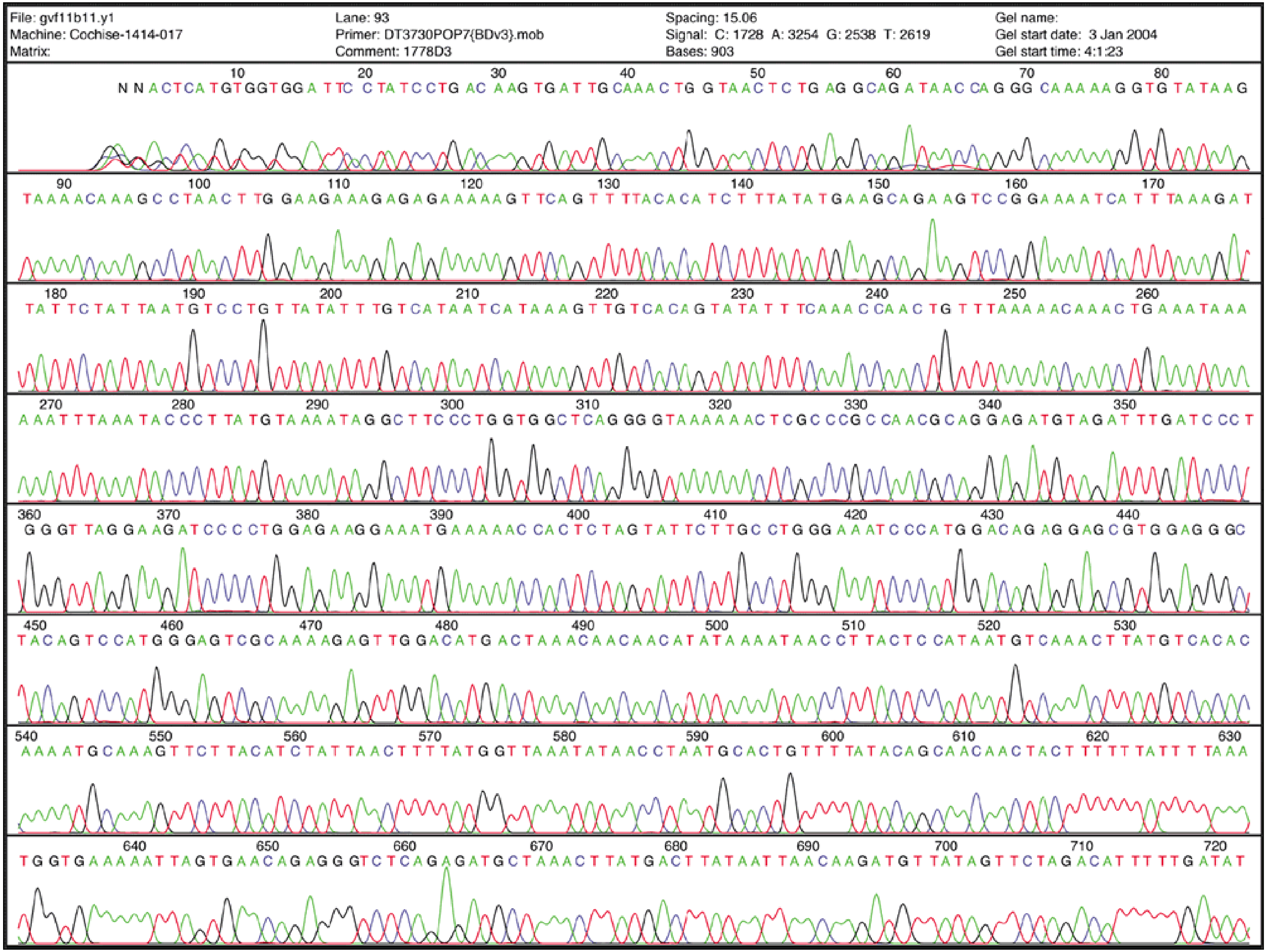
Applied Biosystems is the main technological provider

# Sanger sequencing





# Sanger sequencing



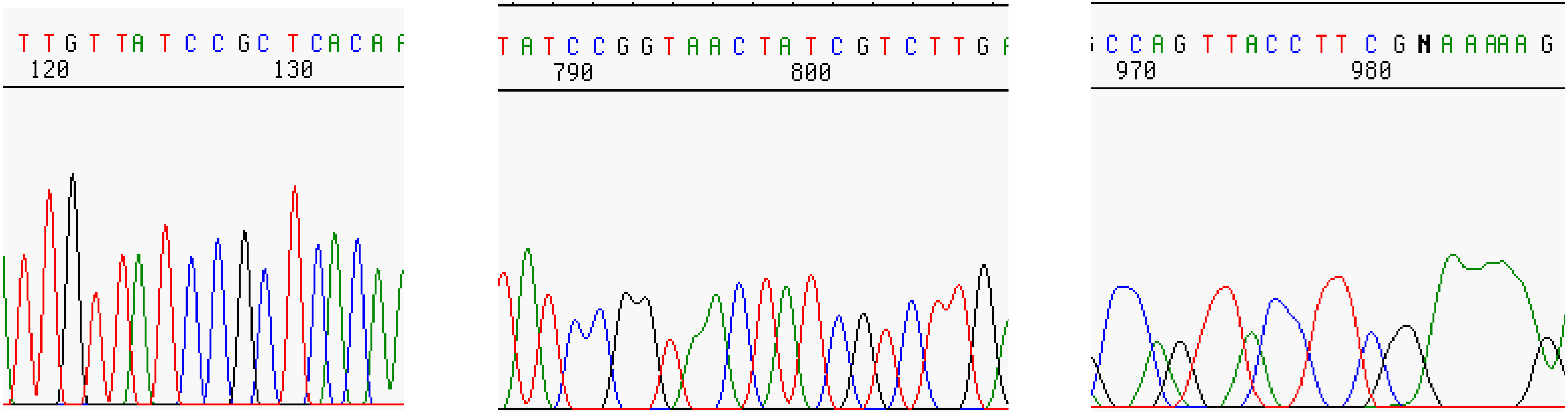
File: gvf11b11.y1
Machine: Cochise-1414-017
Matrix:

Lane: 93
Primer: DT3730POP7{BDv3}.mob
Comment: 1778D3

Spacing: 15.06
Signal: C: 1728 A: 3254 G: 2538 T: 2619
Bases: 903

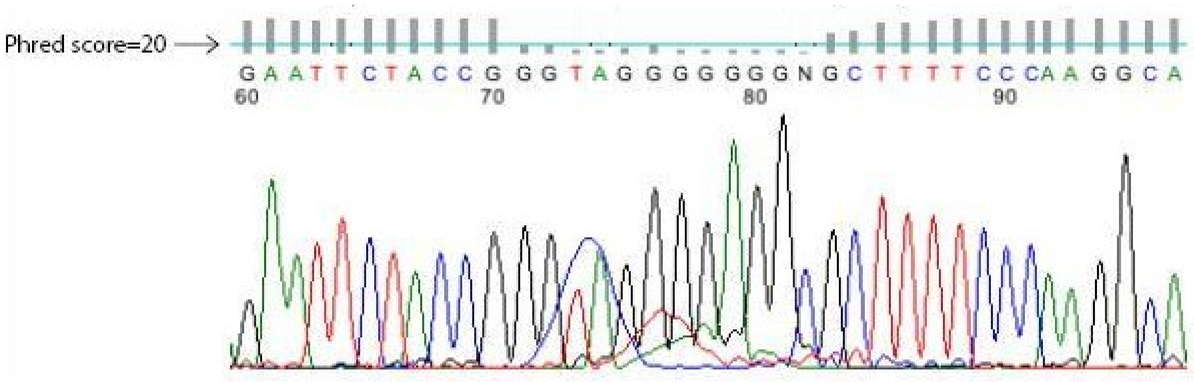
Gel name:
Gel start date: 3 Jan 2004
Gel start time: 4:1:23

# Sequence and quality

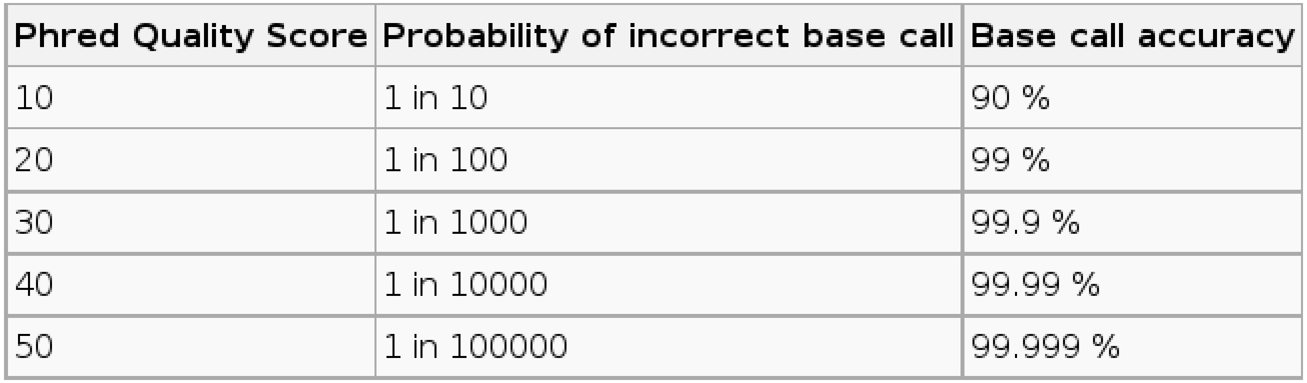


Due to technical limitations different technologies have different errors patterns.

# Sequence and quality



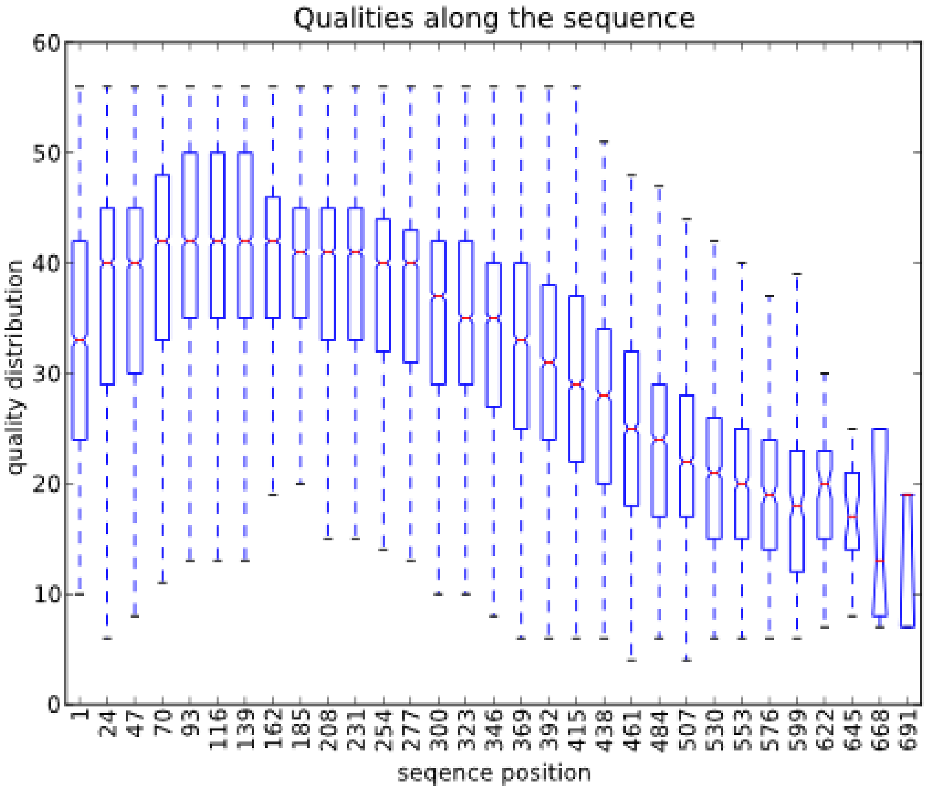
Phred score = - 10 log (prob error)



| 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 10000 | 99.99 % |
| 50 | 1 in 100000 | 99.999 % |

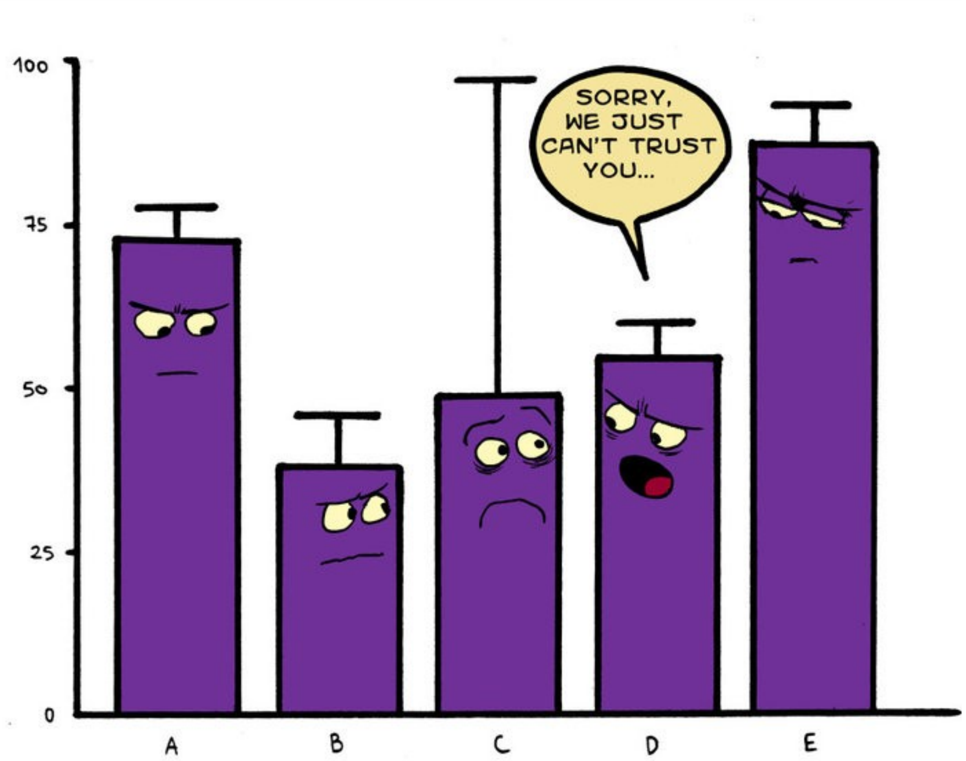
# Sanger sequencing

In Sanger quality is worst at the beginning and at the end.



Any evidence has error bars

Any conclusion has error bars



# Other sources of error

Pre-sequencing:

- PCR mutation-like errors
- Polymerase slippage (low complexity regions)
- PCR primers (e.g. hexamers in random priming)
- Cloning artifacts, chimeric molecules
- Sample contamination
- Index/flag assignment errors

Post-sequencing:

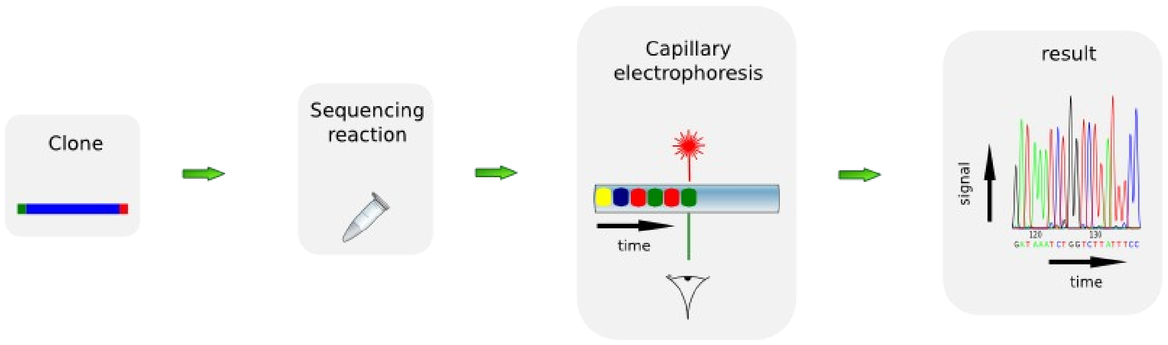
- Assembly artifacts
- Alignment errors due to:
  - Reference
  - Alignment algorithms
- SNV calling software

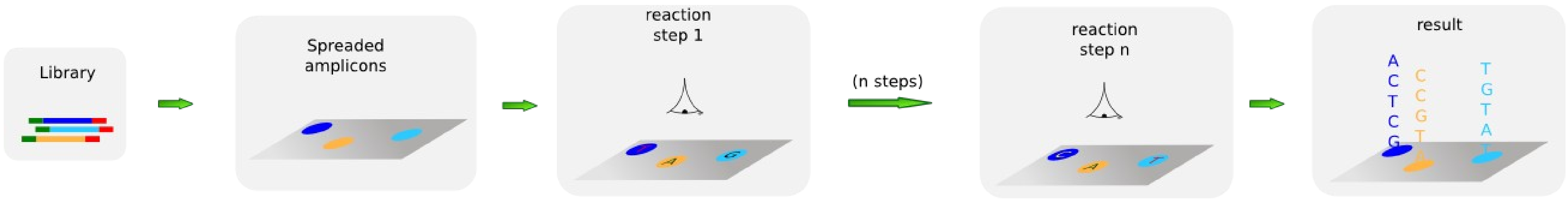
# 2nd generation sequencing

# Sanger vs NGS sequencing

| | Sanger | NGS |
|---|---|---|
| Num. sequences per reaction | 1 clone | Millions of molecules |
| Max. parallelization | 384 | Several millions |
| Sequence quality | High | Low |
| Sequence length | 600-800 bp | 35-20000 (depends on the platform) |
| Throughtput | Low | High |

# Sanger vs NGS





# Library preparation

Fragmentation

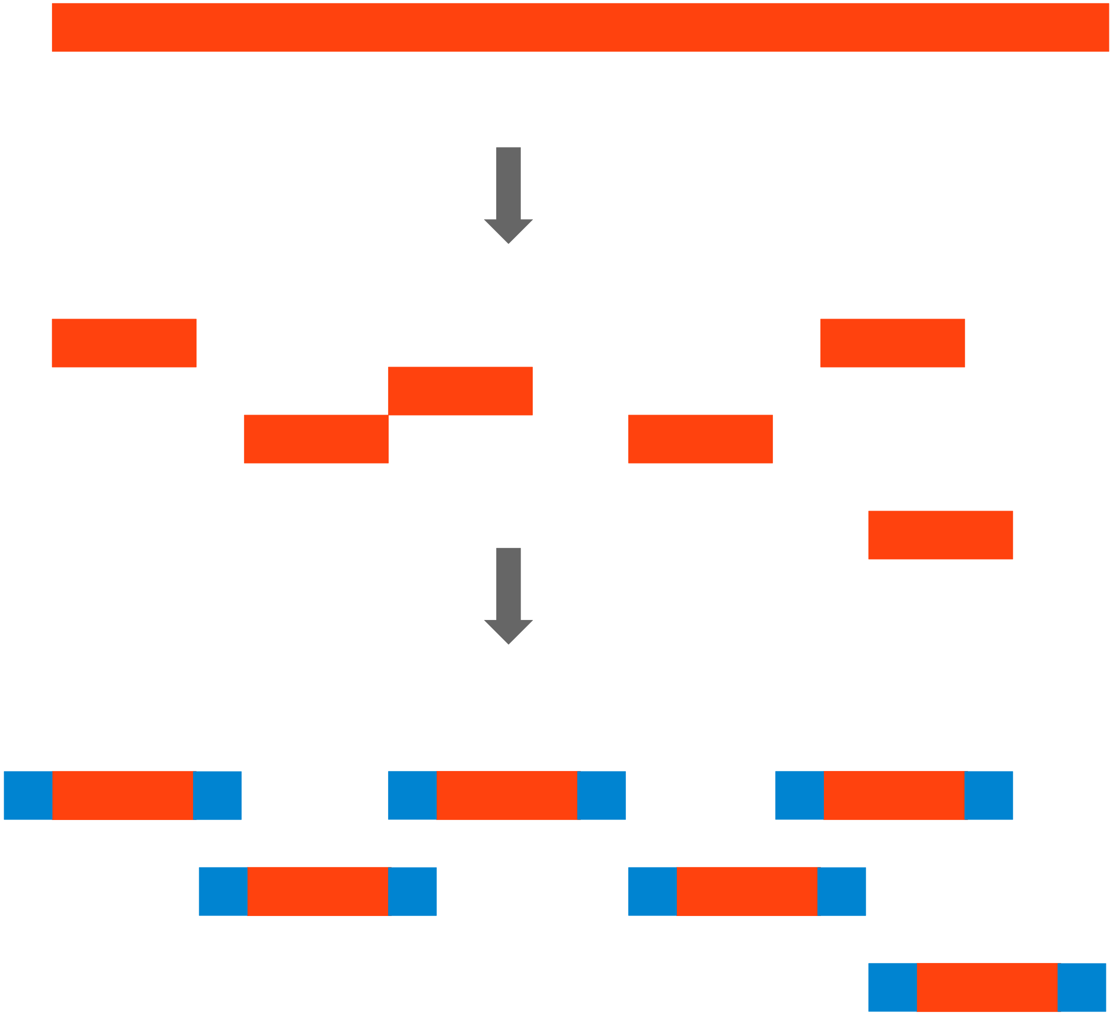
- Sonication
- Nebulization
- Shearing

Size selection

End repair

Sequencing adaptor ligation

Purification



# 454

First NGS platform (and first to be phased out)

Pirosequencing based chemistry

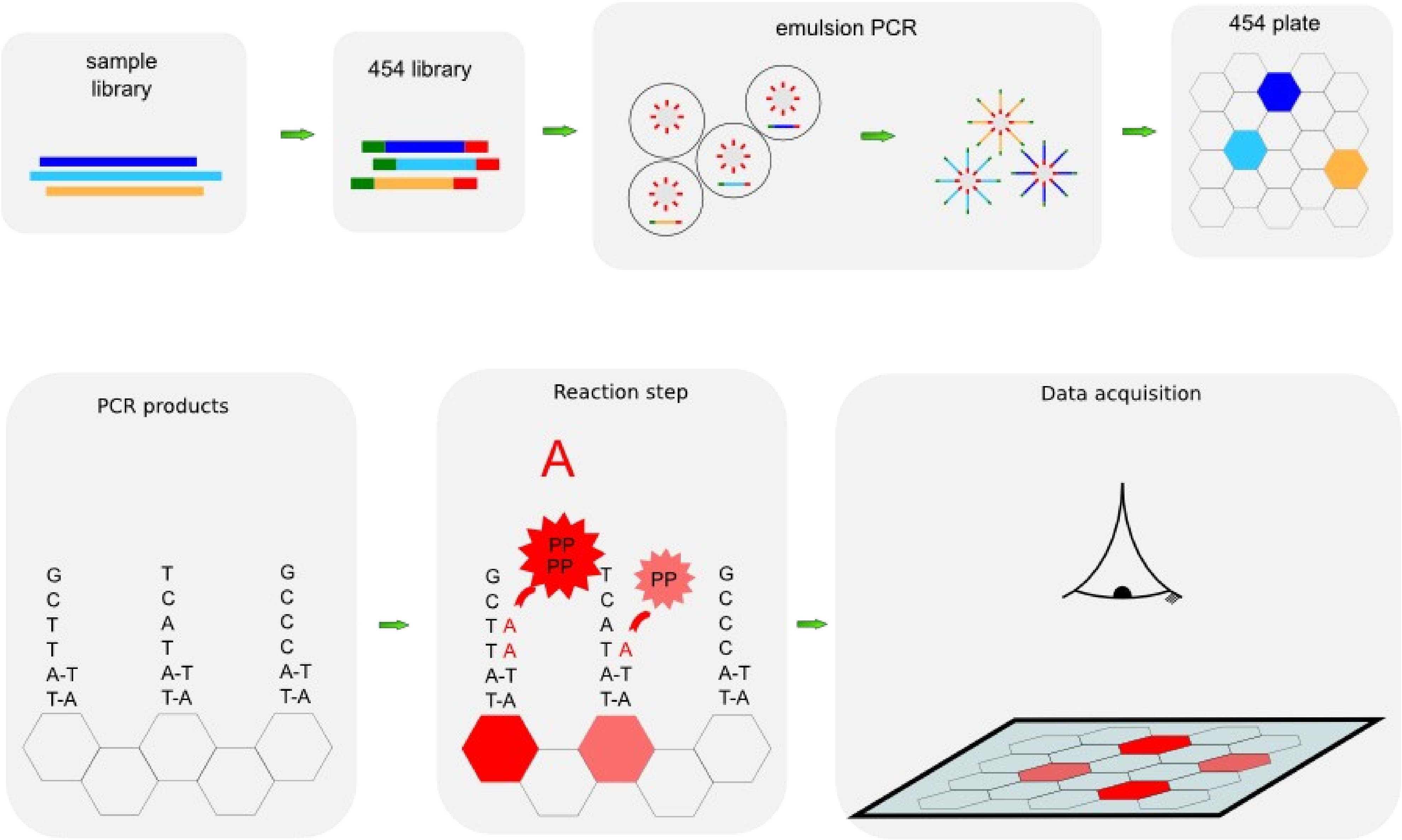
Long reads (400-700bp)

Owned by Roche

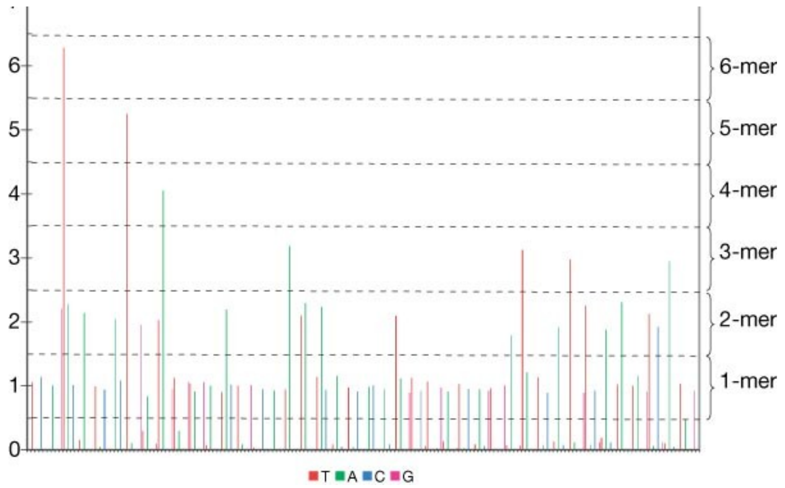
>1 million reads

Obsolete

# 454


sample library
454 library
emulsion PCR
454 plate
PCR products
G
C
T
T
A-T
T-A
T
C
A
T
A-T
T-A
G
C
C
C
A-T
T-A
Reaction step
A
PP
PP
PP
G
C
T A
T A
A-T
T-A
T
C
A
T A
A-T
T-A
G
C
C
C
A-T
T-A
Data acquisition

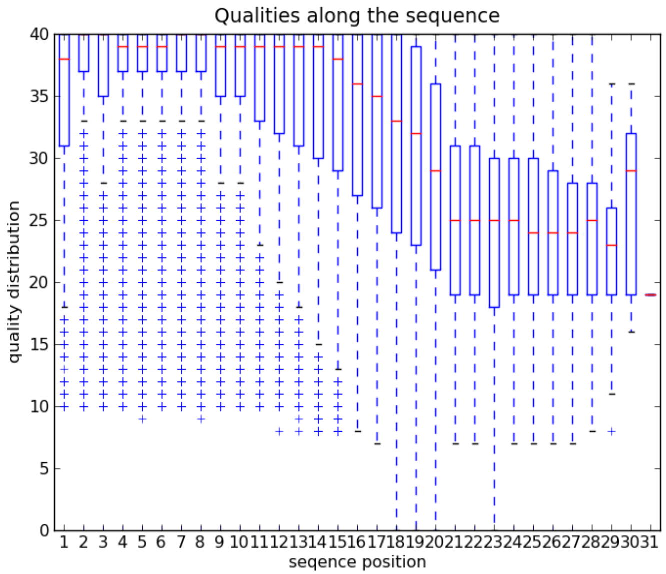
# 454 quality



The lengthiest the homopolymer the less quality.

It is very difficult to differentiate AAAAAA from AAAAA.

Qua
leng



# Illumina

Previously known as Solexa

Reversible terminators based sequencing technique

Short reads (50 or 250bp depending on the version)

Lowest cost per base

Ideal for resequencing projects

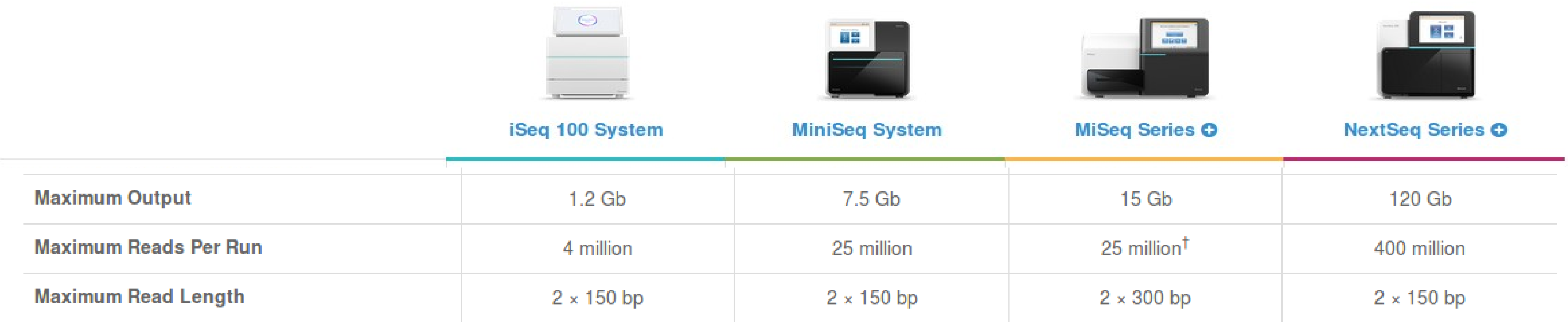
Highest throughput

Runs divided in 8 lanes

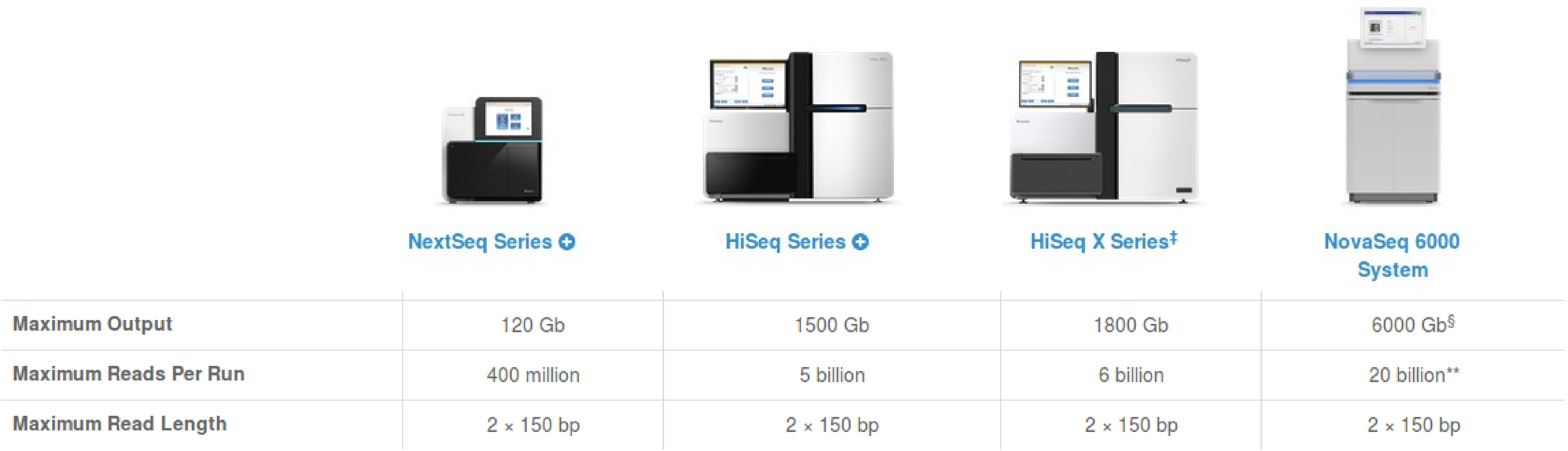
Up to 4000 million reads

Can sequence both ends of the molecules (paired ends)

# Illumina instruments

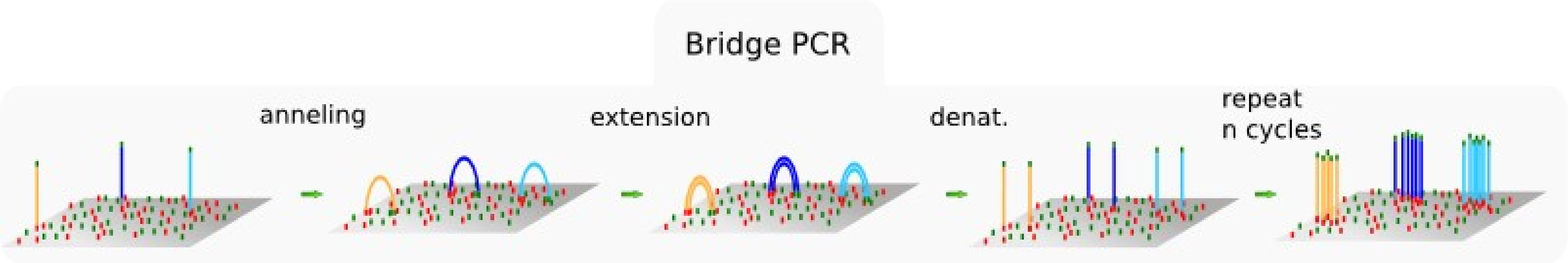


| | iSeq 100 System | MiniSeq System | MiSeq Series ⊕ | NextSeq Series ⊕ |
|---|---|---|---|---|
| **Maximum Output** | 1.2 Gb | 7.5 Gb | 15 Gb | 120 Gb |
| **Maximum Reads Per Run** | 4 million | 25 million | 25 million[†] | 400 million |
| **Maximum Read Length** | 2 × 150 bp | 2 × 150 bp | 2 × 300 bp | 2 × 150 bp |



| | NextSeq Series ⊕ | HiSeq Series ⊕ | HiSeq X Series[‡] | NovaSeq 6000 System |
|---|---|---|---|---|
| **Maximum Output** | 120 Gb | 1500 Gb | 1800 Gb | 6000 Gb[§] |
| **Maximum Reads Per Run** | 400 million | 5 billion | 6 billion | 20 billion** |
| **Maximum Read Length** | 2 × 150 bp | 2 × 150 bp | 2 × 150 bp | 2 × 150 bp |

# Illumina



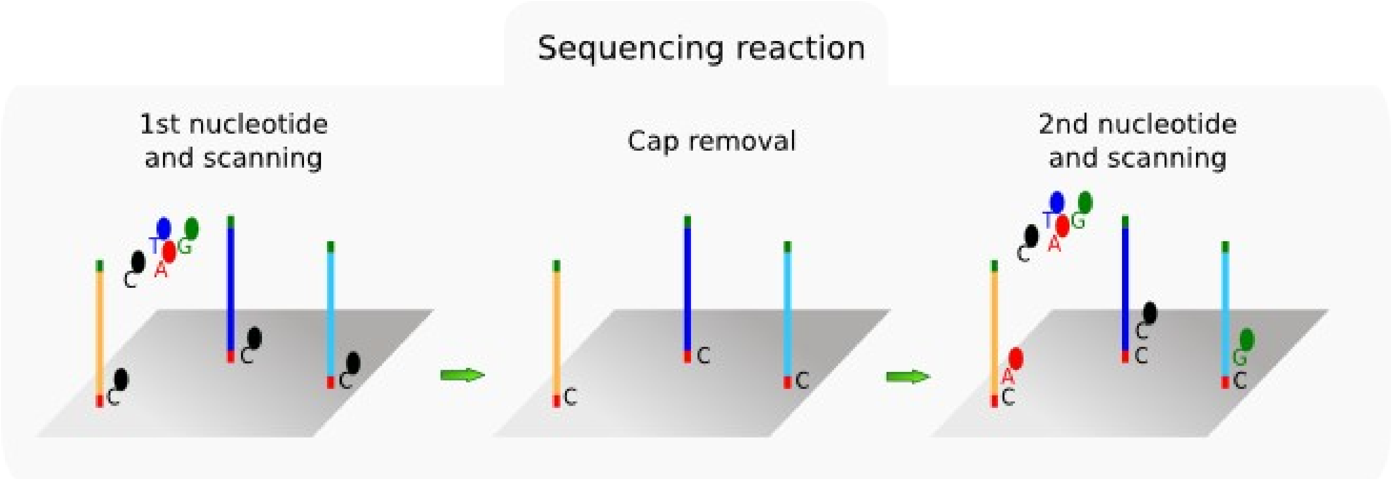
Bridge PCR

anneling

extension

denat.

repeat
n cycles

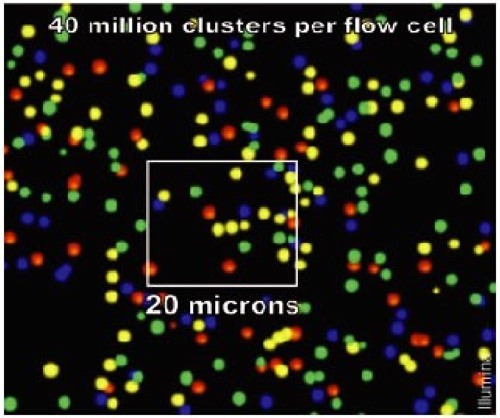


Sequencing reaction

1st nucleotide
and scanning

Cap removal

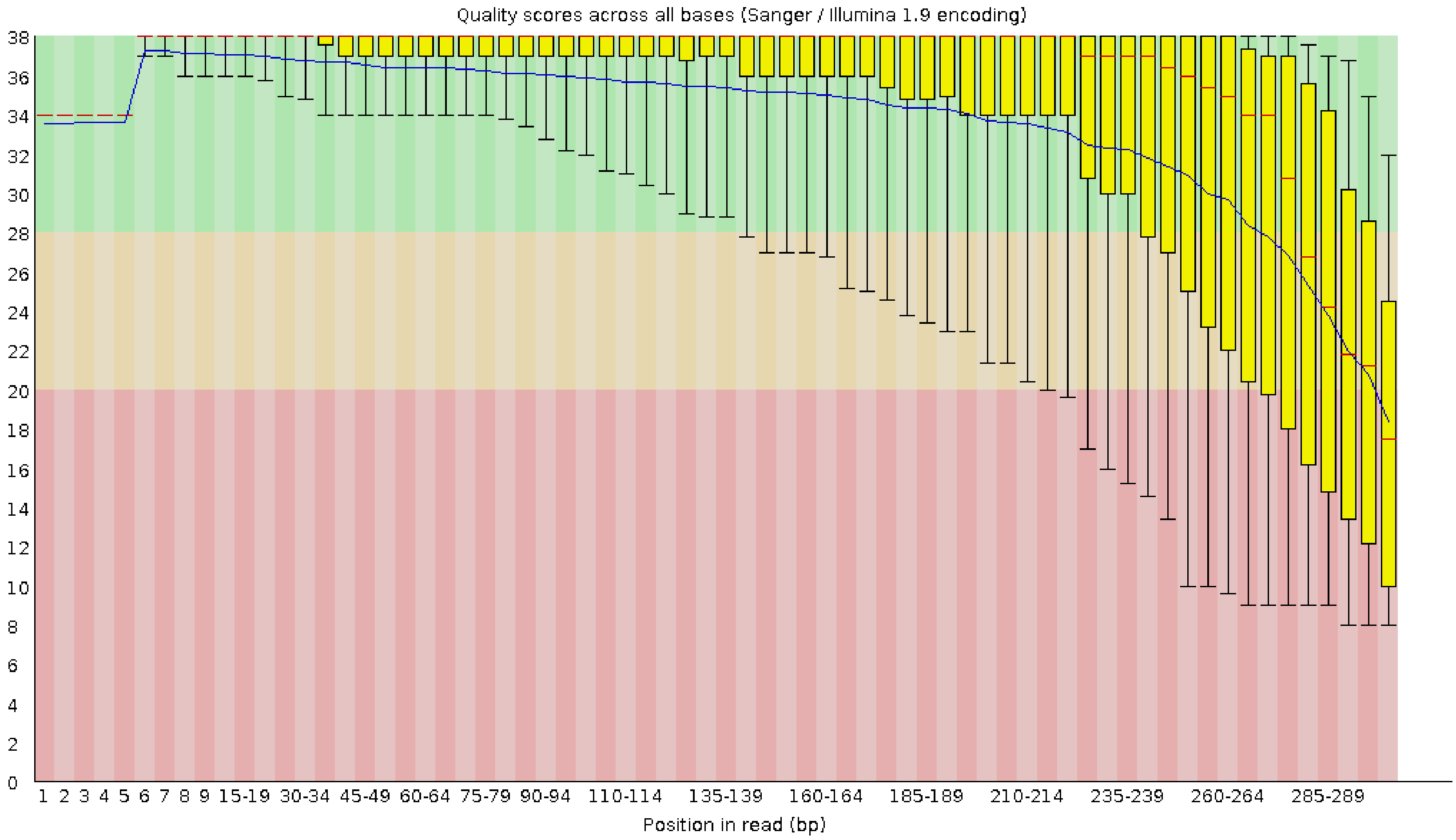
2nd nucleotide
and scanning



# Illumina

Quality diminishes with sequence length.

No homopolymer problem, mainly substitution errors.



# SOLiD

Ligation based sequencing chemistry

Short reads (35 - 75bp depending on the version)

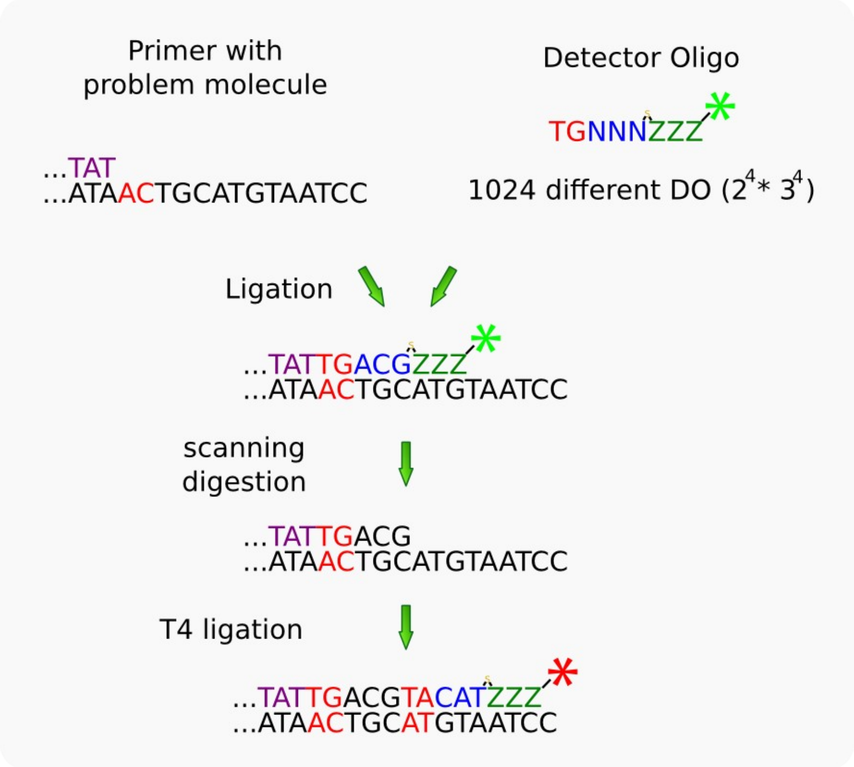
Only for resequencing projects

It used to produce color sequences, not nucleotides

Color sequences have poor quality, but nucleotide sequences have high quality

115 or 320 million reads

# SOLiD



Primer with problem molecule

...TAT
...ATAACTGCATGTAATCC

Detector Oligo

TGNNNZZZ*

1024 different DO ($2^4$* $3^4$)

Ligation

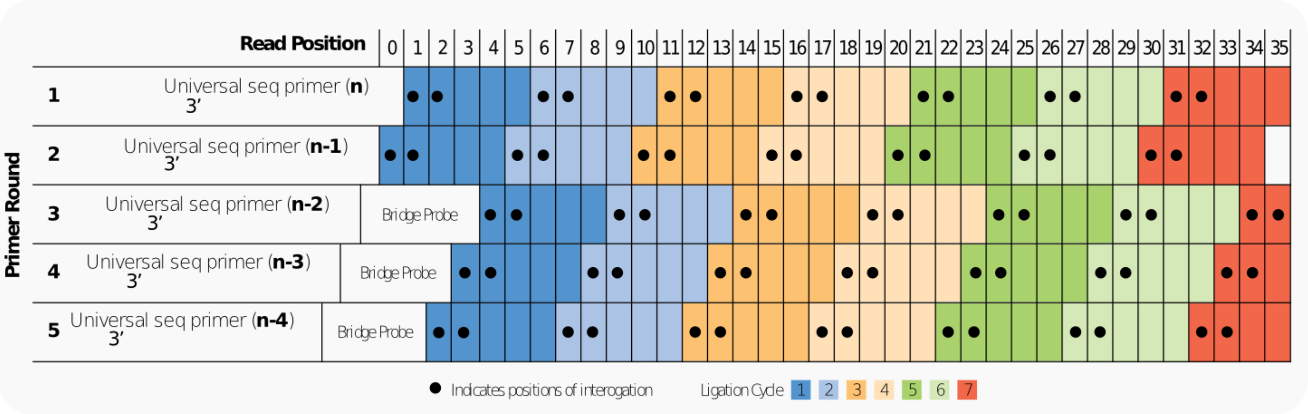
...TATTGACGZZZ*
...ATAACTGCATGTAATCC

scanning digestion

...TATTGACG
...ATAACTGCATGTAATCC

T4 ligation

...TATTGACGTACATZZZ*
...ATAACTGCATGTAATCC

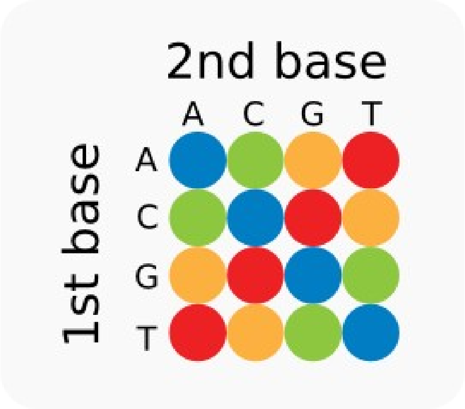


| Primer Round | Read Position | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Universal seq primer (**n**) 3′ | | 1 ● | 1 ● | 1 | 1 | 1 | 2 ● | 2 ● | 2 | 2 | 2 | 3 ● | 3 ● | 3 | 3 | 3 | 4 ● | 4 ● | 4 | 4 | 4 | 5 ● | 5 ● | 5 | 5 | 5 | 6 ● | 6 ● | 6 | 6 | 6 | 7 ● | 7 ● | 7 | 7 | 7 |
| 2 | Universal seq primer (**n-1**) 3′ | 1 ● | 1 ● | 1 | 1 | 1 | 2 ● | 2 ● | 2 | 2 | 2 | 3 ● | 3 ● | 3 | 3 | 3 | 4 ● | 4 ● | 4 | 4 | 4 | 5 ● | 5 ● | 5 | 5 | 5 | 6 ● | 6 ● | 6 | 6 | 6 | 7 ● | 7 ● | 7 | 7 | 7 | |
| 3 | Universal seq primer (**n-2**) 3′ | Bridge Probe | | | | 1 ● | 1 ● | 1 | 1 | 1 | 2 ● | 2 ● | 2 | 2 | 2 | 3 ● | 3 ● | 3 | 3 | 3 | 4 ● | 4 ● | 4 | 4 | 4 | 5 ● | 5 ● | 5 | 5 | 5 | 6 ● | 6 ● | 6 | 6 | 6 | 7 ● | 7 ● |
| 4 | Universal seq primer (**n-3**) 3′ | Bridge Probe | | | 1 ● | 1 ● | 1 | 1 | 1 | 2 ● | 2 ● | 2 | 2 | 2 | 3 ● | 3 ● | 3 | 3 | 3 | 4 ● | 4 ● | 4 | 4 | 4 | 5 ● | 5 ● | 5 | 5 | 5 | 6 ● | 6 ● | 6 | 6 | 6 | 7 ● | 7 ● | 7 |
| 5 | Universal seq primer (**n-4**) 3′ | Bridge Probe | | 1 ● | 1 ● | 1 | 1 | 1 | 2 ● | 2 ● | 2 | 2 | 2 | 3 ● | 3 ● | 3 | 3 | 3 | 4 ● | 4 ● | 4 | 4 | 4 | 5 ● | 5 ● | 5 | 5 | 5 | 6 ● | 6 ● | 6 | 6 | 6 | 7 ● | 7 ● | 7 | 7 |

● Indicates positions of interrogation

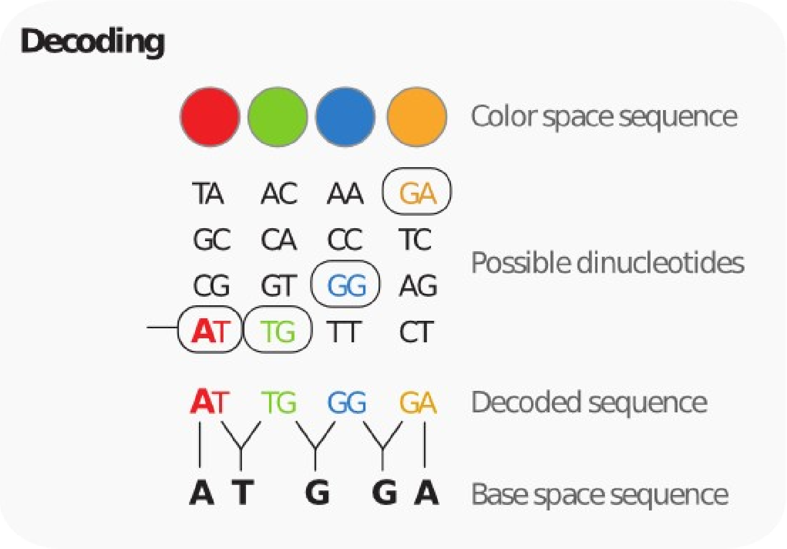
Ligation Cycle 1 2 3 4 5 6 7

# SOLiD



2nd base

| 1st base | A | C | G | T |
|---|---|---|---|---|
| A | 🟦 | 🟩 | 🟧 | 🟥 |
| C | 🟩 | 🟦 | 🟥 | 🟧 |
| G | 🟧 | 🟥 | 🟦 | 🟩 |
| T | 🟥 | 🟧 | 🟩 | 🟦 |



**Decoding**

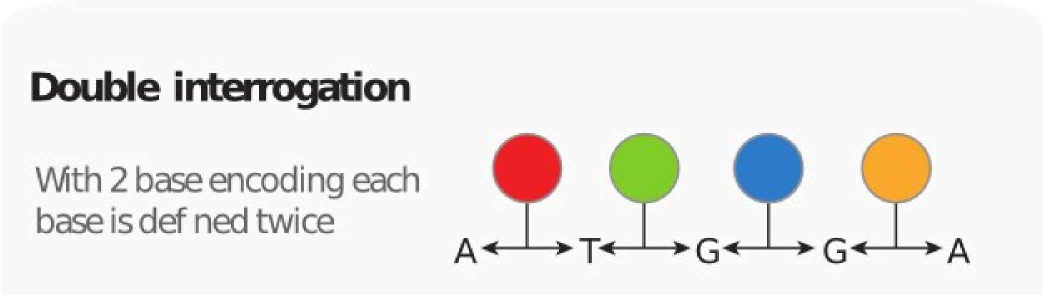
🟥 🟩 🟦 🟧 Color space sequence

| | | | |
|---|---|---|---|
| TA | AC | AA | GA |
| GC | CA | CC | TC |
| CG | GT | GG | AG |
| AT | TG | TT | CT |

Possible dinucleotides

AT TG GG GA Decoded sequence

A T G G A Base space sequence



**Double interrogation**

With 2 base encoding each base is defined twice

Really?

# Ion Torrent

Around 60-80 M reads.

200 pb length.

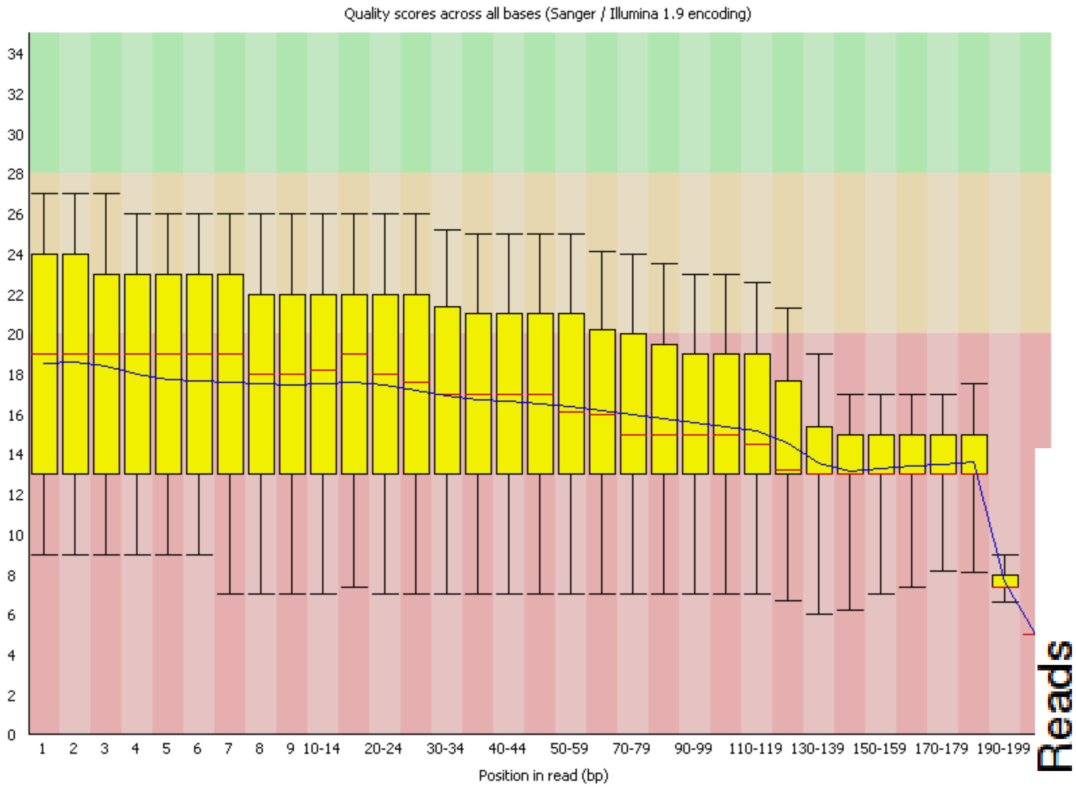
Sequences based on H+ production

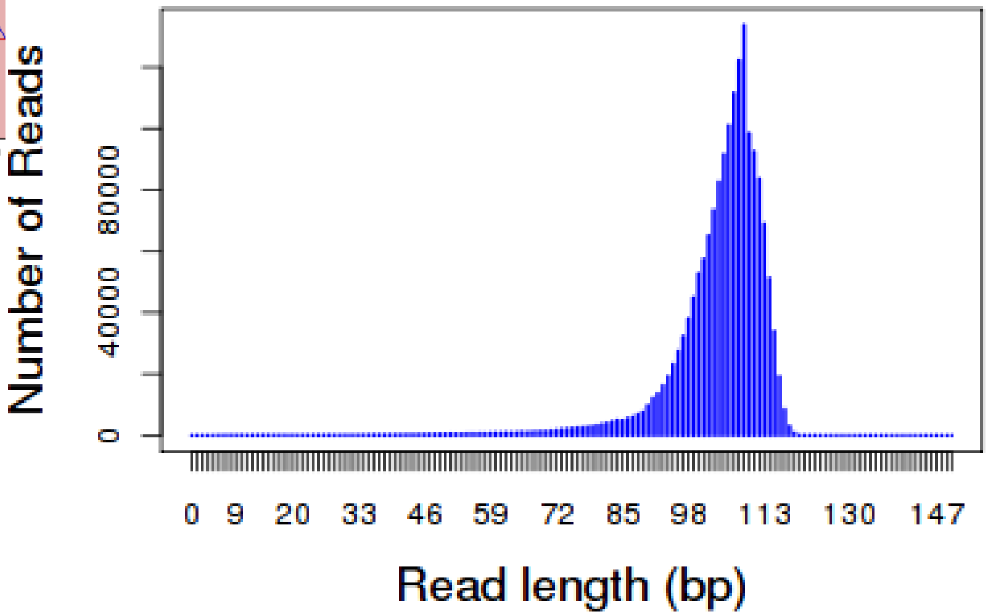
Error rates higher than other 2nd generation

Error pattern similar to 454, with homopolymer problem.

Belongs to Life technologies (Applied Biosystems)

# Ion Torrent


Quality scores across all bases (Sanger / Illumina 1.9 encoding)
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1
2
3
4
5
6
7
8
9
10-14
20-24
30-34
40-44
50-59
70-79
90-99
110-119
130-139
150-159
170-179
190-199
Position in read (bp)


Number of Reads
0
40000
80000
0
9
20
33
46
59
72
85
98
113
130
147
Read length (bp)

# 3rd generation sequencing

# PacBio

3rd generation platform (single molecule)

Polymerase based chemistry (SMRT)

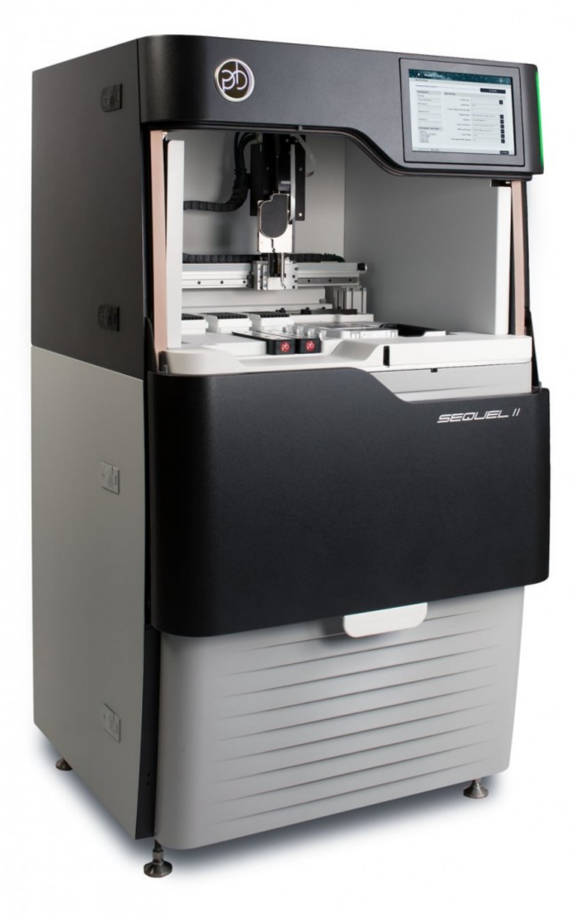
Long reads (typically 5 to 60 kb)

Very high error rate for the standard mode

- It has a HiFi platform with low error rate

Ideal for de novo sequencing projects

Not many reads

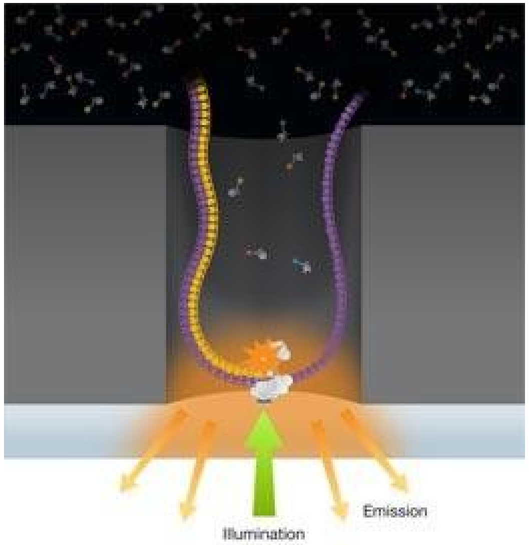


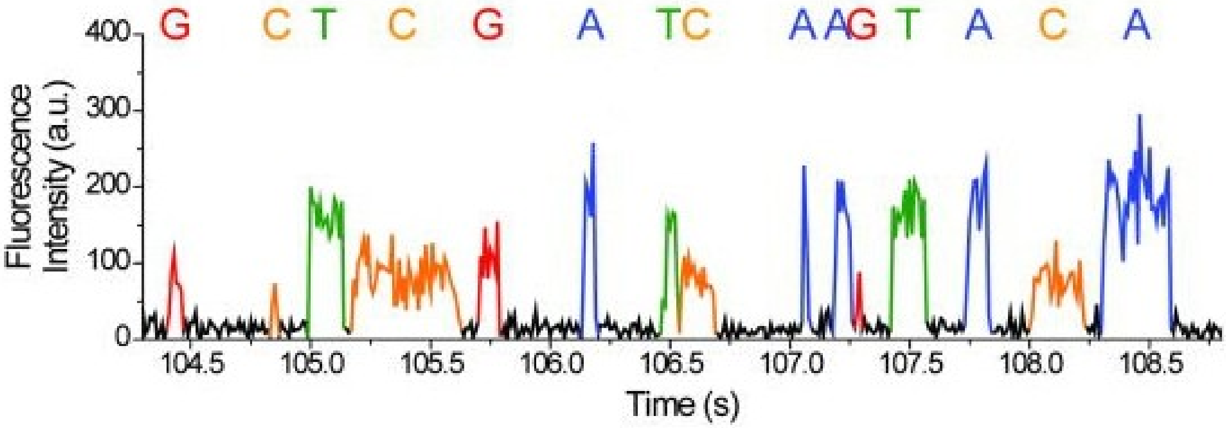
# PacBio

3rd generation, single molecule detection. No amplification step required.

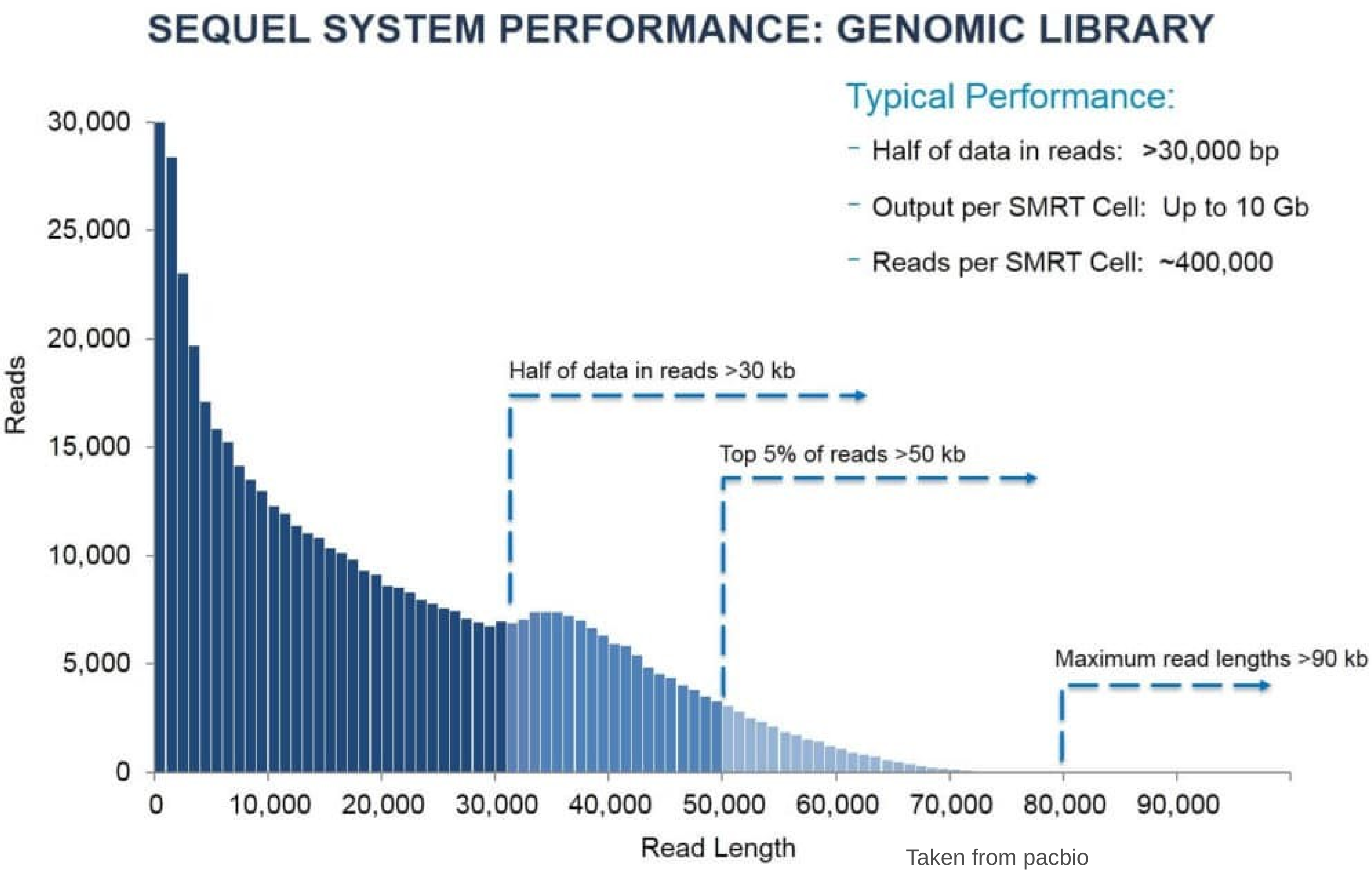
Nucleotides labeled on the phosphate removed during the polymerization.

Sequencing based on the time required by the polymerase to incorporate a nucleotide (Polymerase requires milliseconds versus microseconds for the stochastic diffusion)





# PacBio length distribution



SEQUEL SYSTEM PERFORMANCE: GENOMIC LIBRARY

Typical Performance:

- Half of data in reads: >30,000 bp
- Output per SMRT Cell: Up to 10 Gb
- Reads per SMRT Cell: ~400,000

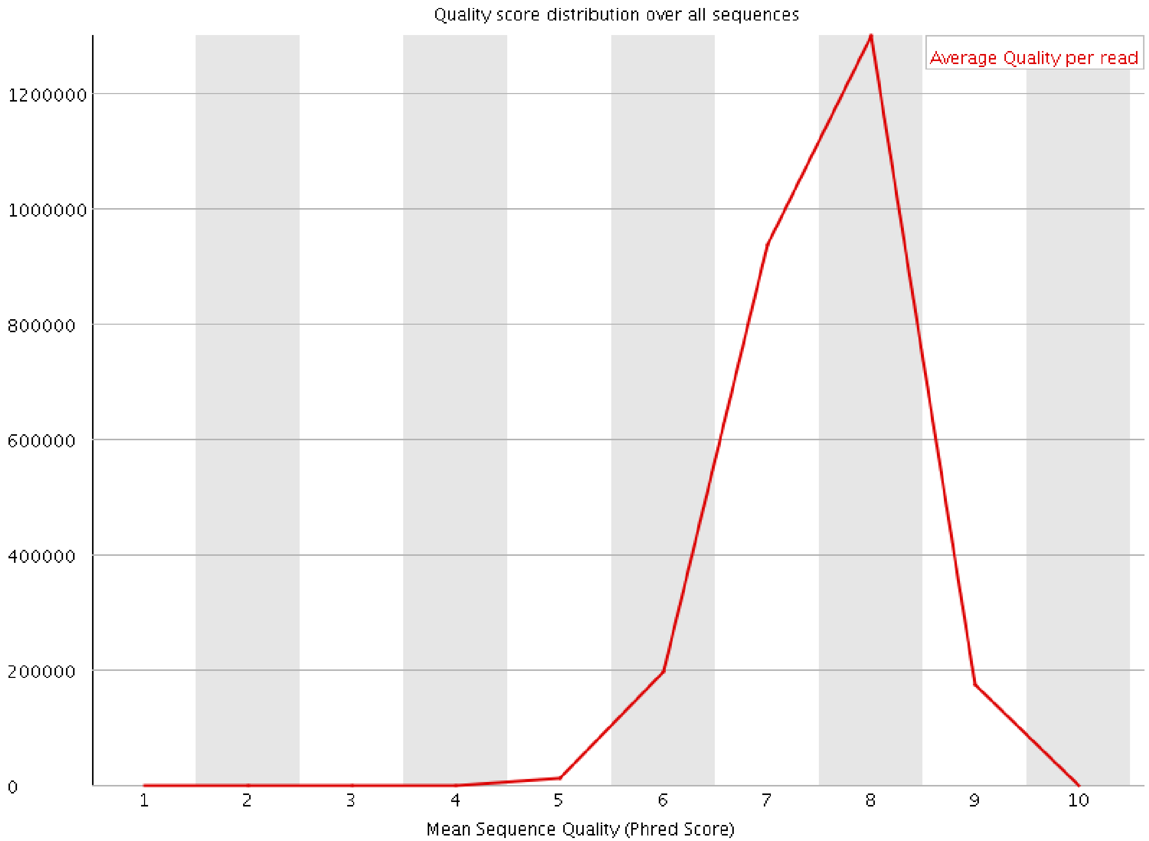
Half of data in reads >30 kb

Top 5% of reads >50 kb

Maximum read lengths >90 kb

Taken from pacbio

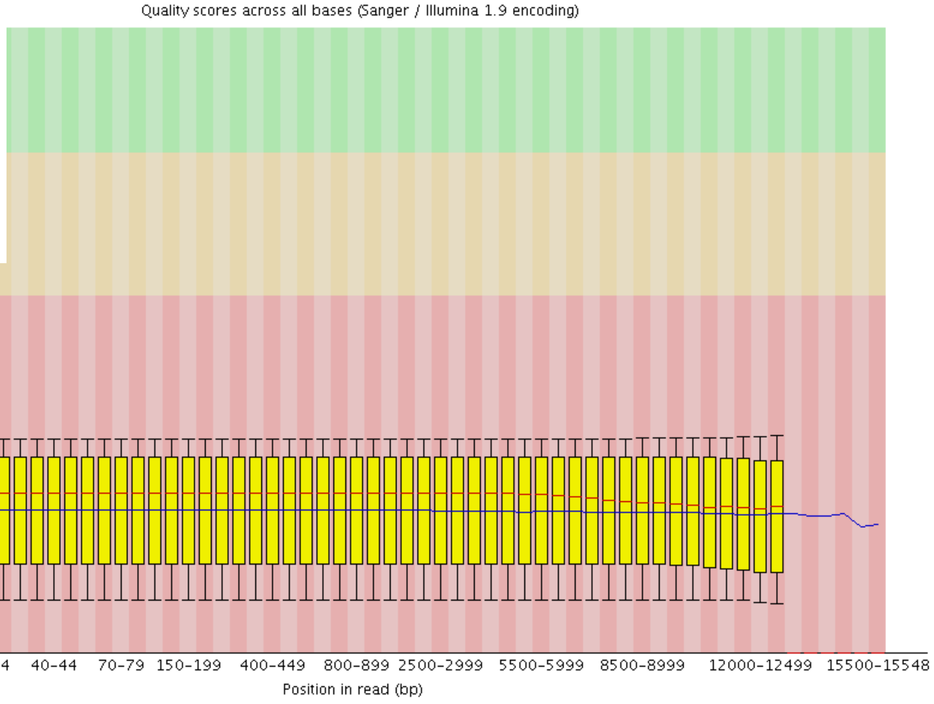
# PacBio quality distribution



Taken from flxlexblog.wordpress.com

Distributions for standard sequencing.

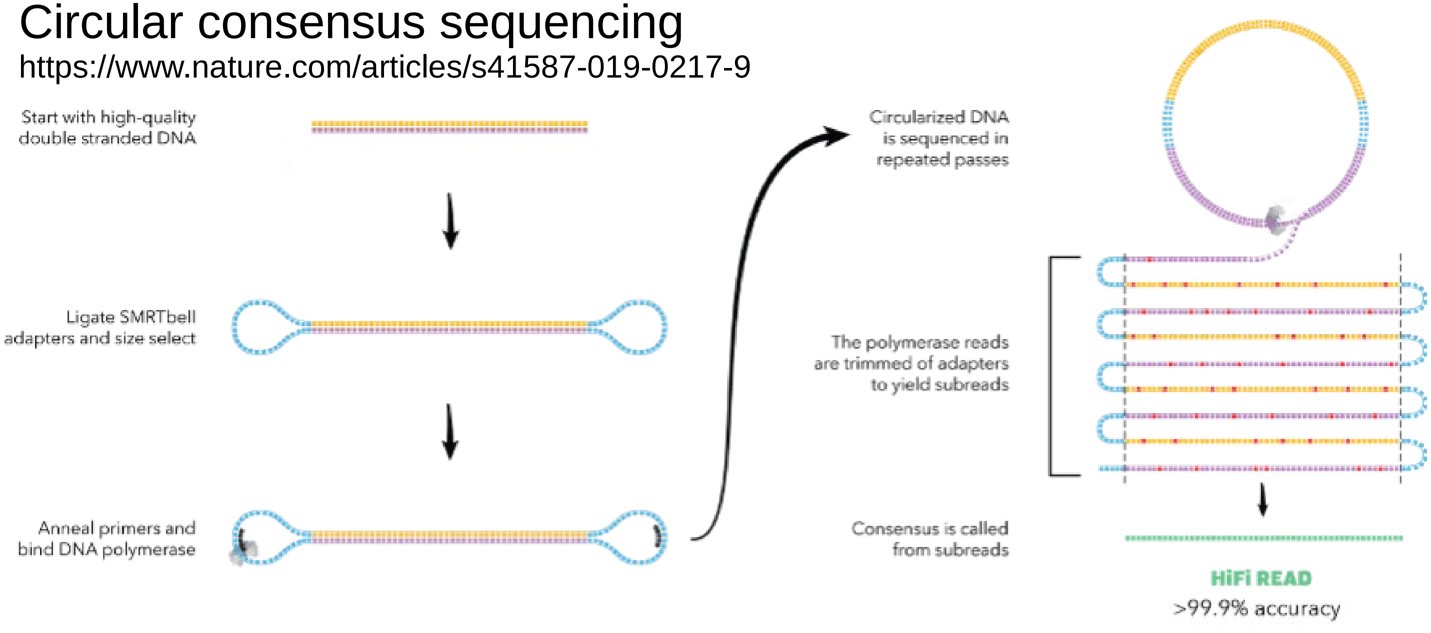
85-90 % error rate



# Pacbio High Fidelity (HiFi)

Circular consensus sequencing

https://www.nature.com/articles/s41587-019-0217-9


Start with high-quality double stranded DNA
Ligate SMRTbell adapters and size select
Anneal primers and bind DNA polymerase
Circularized DNA is sequenced in repeated passes
The polymerase reads are trimmed of adapters to yield subreads
Consensus is called from subreads
HiFi READ
>99.9% accuracy

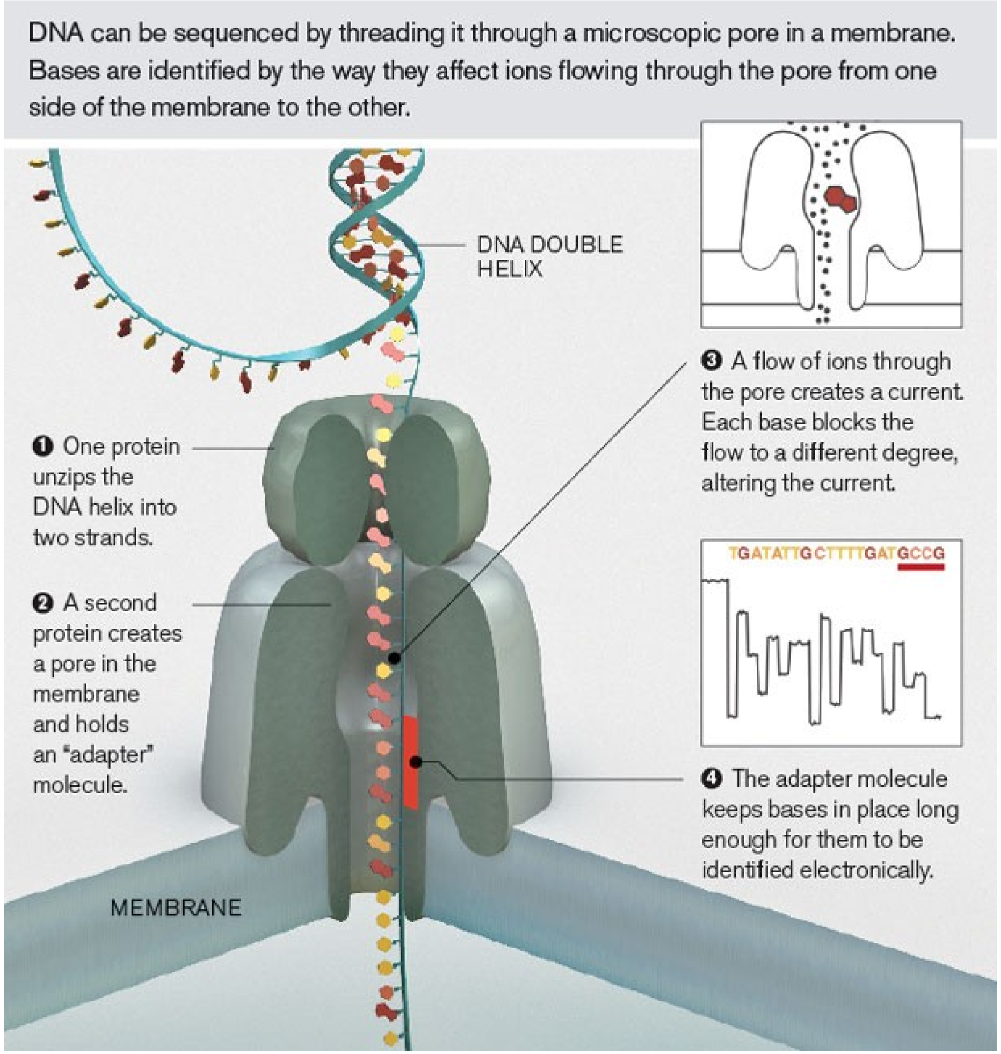
Long reads and high quality: 99% error rate

Compared with standard mode:

- smaller read lengths: 10-30 kb
- Lower yields

# Nanopore

Senses differences in ion flow



DNA can be sequenced by threading it through a microscopic pore in a membrane. Bases are identified by the way they affect ions flowing through the pore from one side of the membrane to the other.

❶ One protein unzips the DNA helix into two strands.

❷ A second protein creates a pore in the membrane and holds an "adapter" molecule.

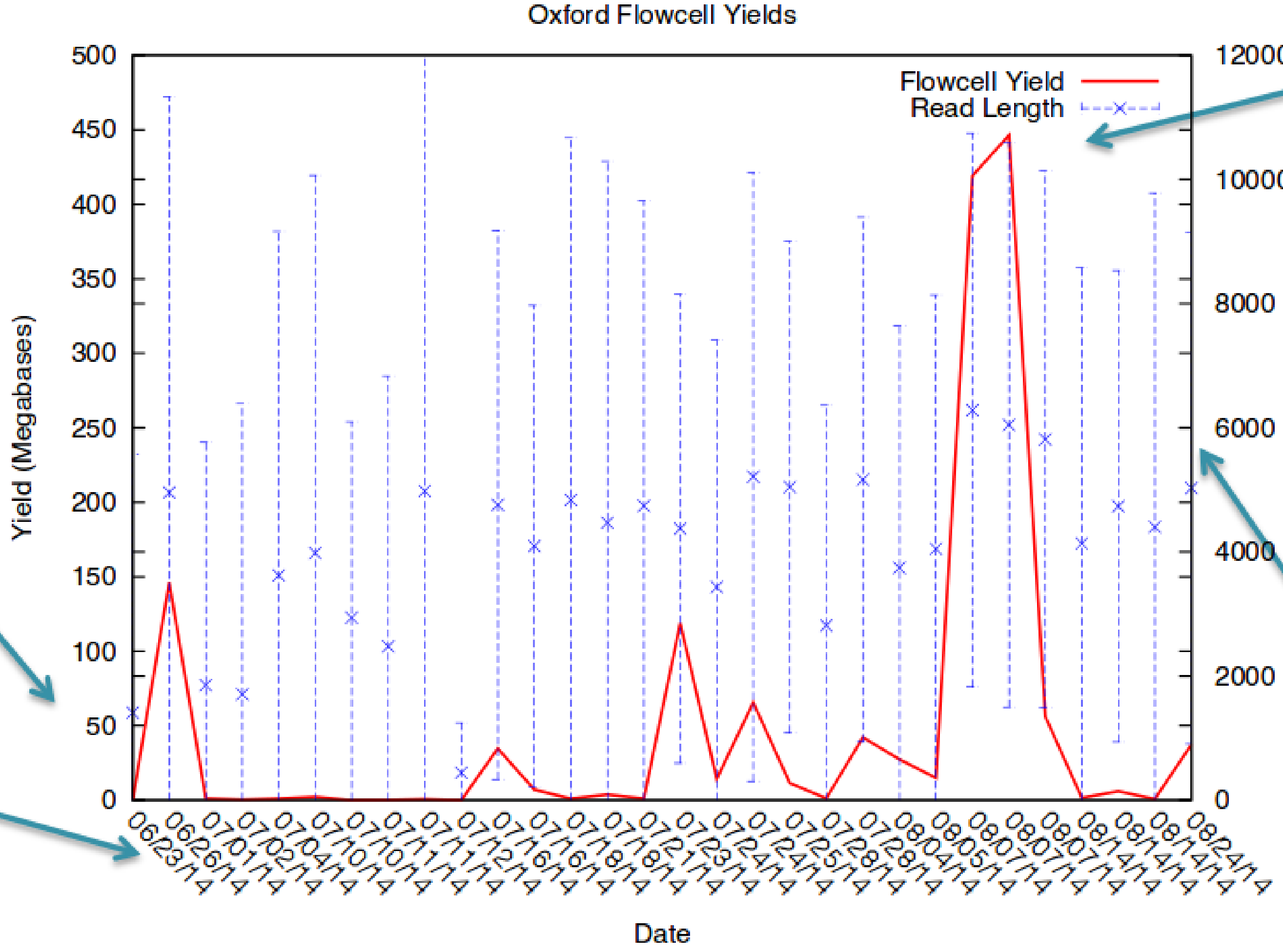
❸ A flow of ions through the pore creates a current. Each base blocks the flow to a different degree, altering the current.

❹ The adapter molecule keeps bases in place long enough for them to be identified electronically.

miniION

# Nanopore-first data

Not very reliable ¿yet?



Best: 446Mb

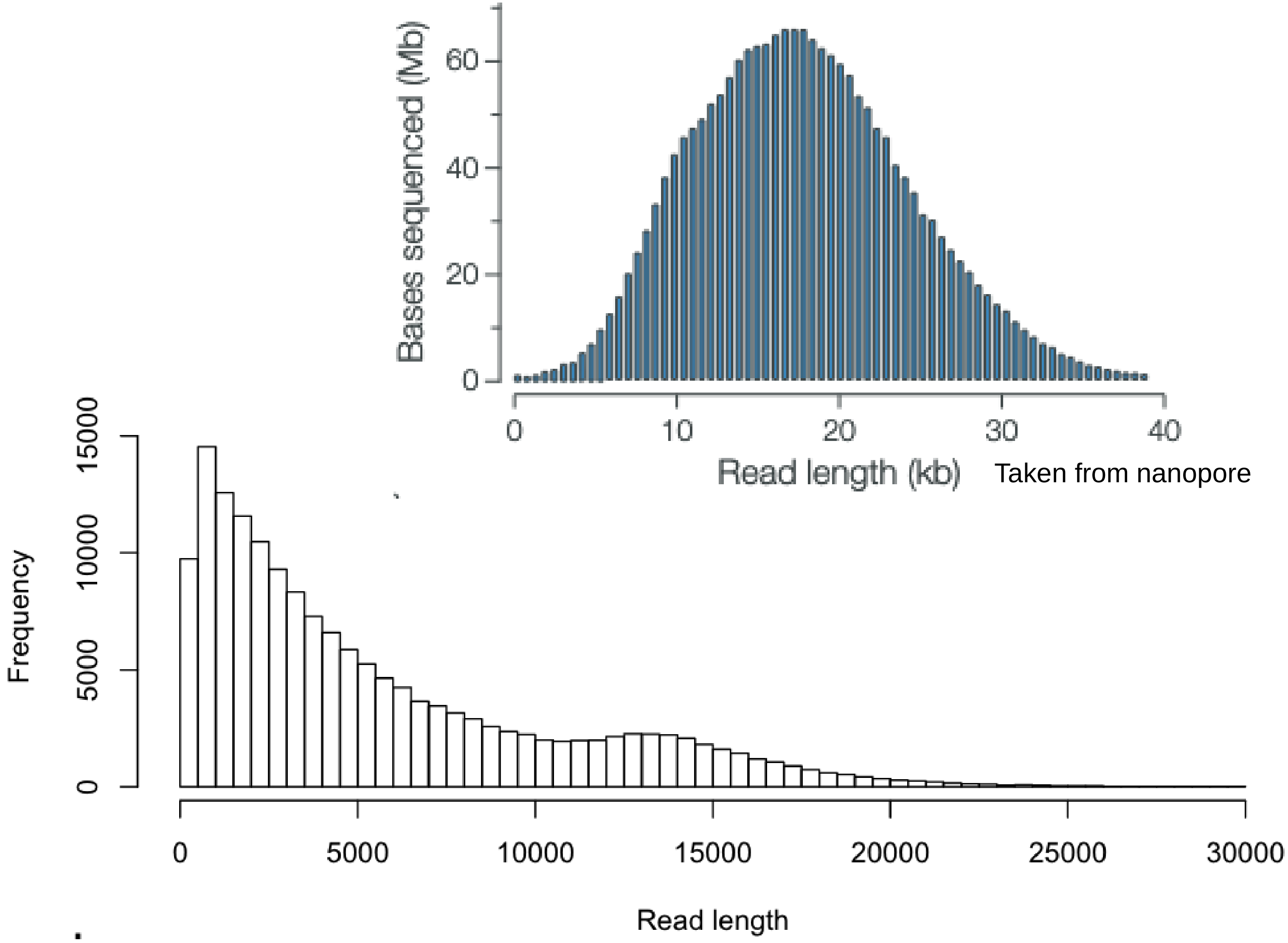
Mean: 50Mb

Mean Read Length: ~6kb (10kb shear)

30 Flowcells

# Nanopore read length and quality

Reads are typically 10–100 kb in length and 87–98% accurate



Taken from nanopore

https://melbourne.figshare.com/articles/figure/Nanopore_read_length_distribution/12652034/2

# Nanopore ultra long reads

Typically longer than 100Kb

- Reads one order of magnitude longer than Pacbio reads
- Main limiting factor is DNA extraction

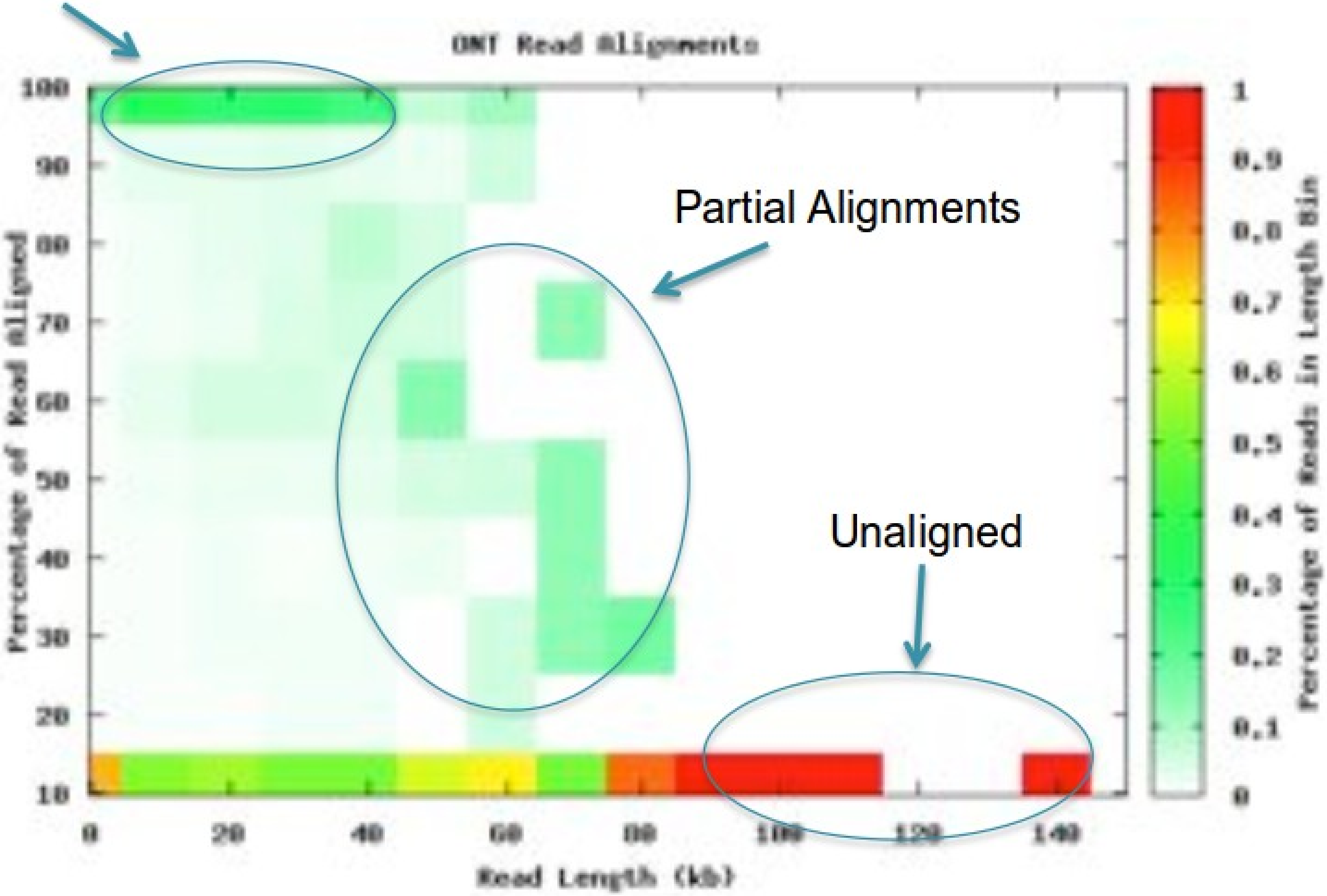
Low accuracy: 87–98%

Lower yield than standard nanopore reads

# Nanopore alignments

32% of the data map using BLASTN

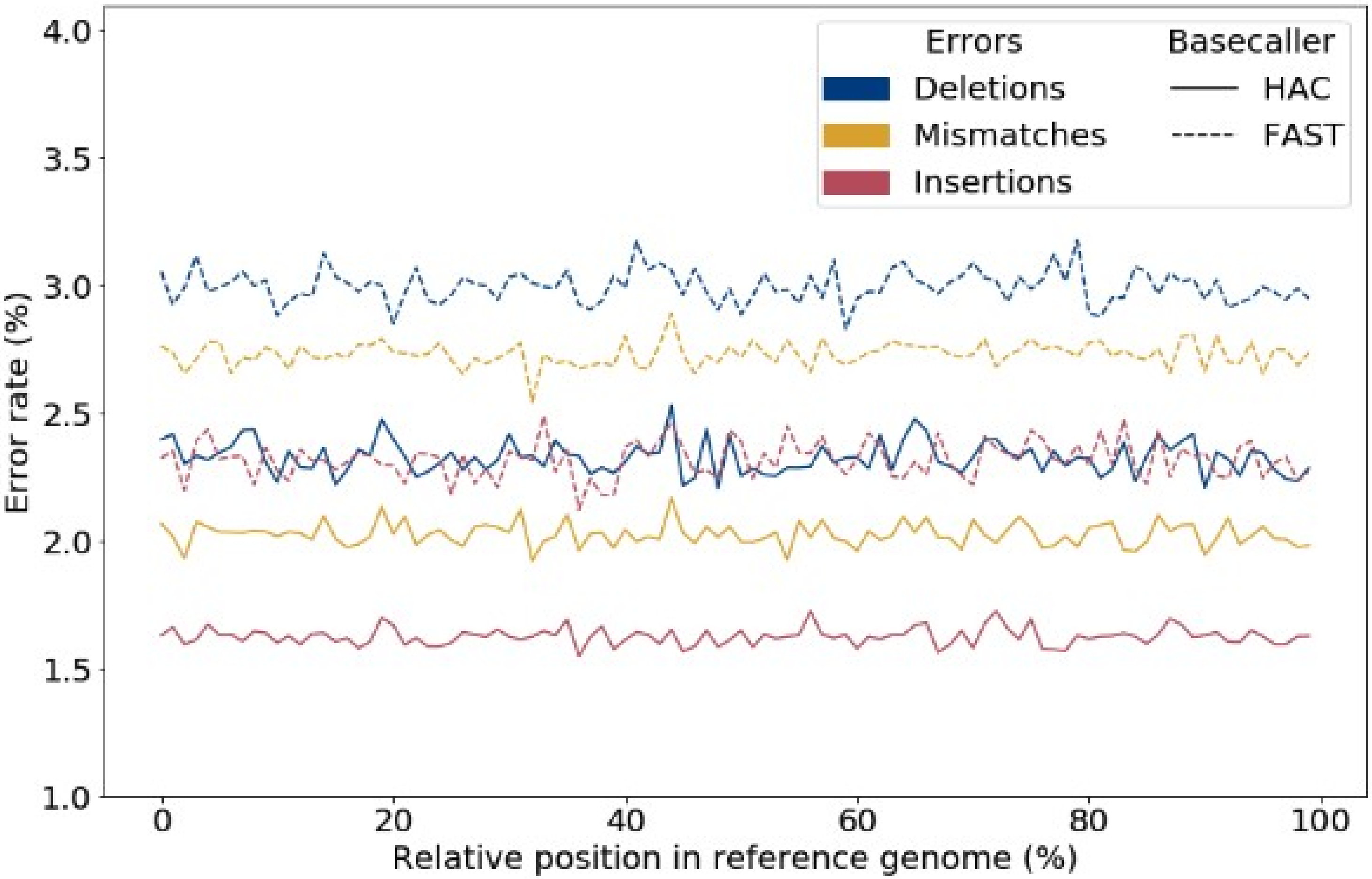
Full Length Alignments



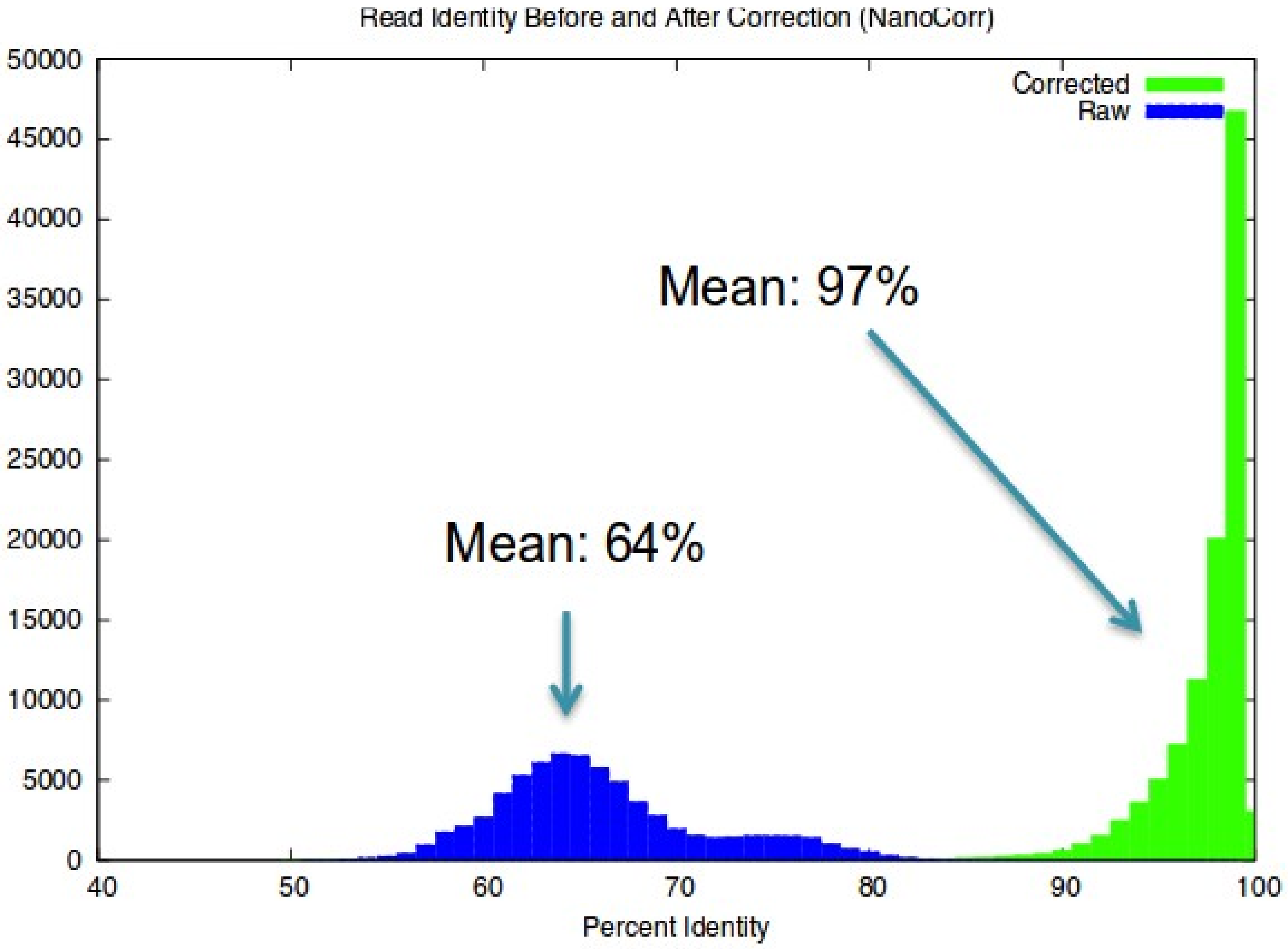
Partial Alignments

Unaligned

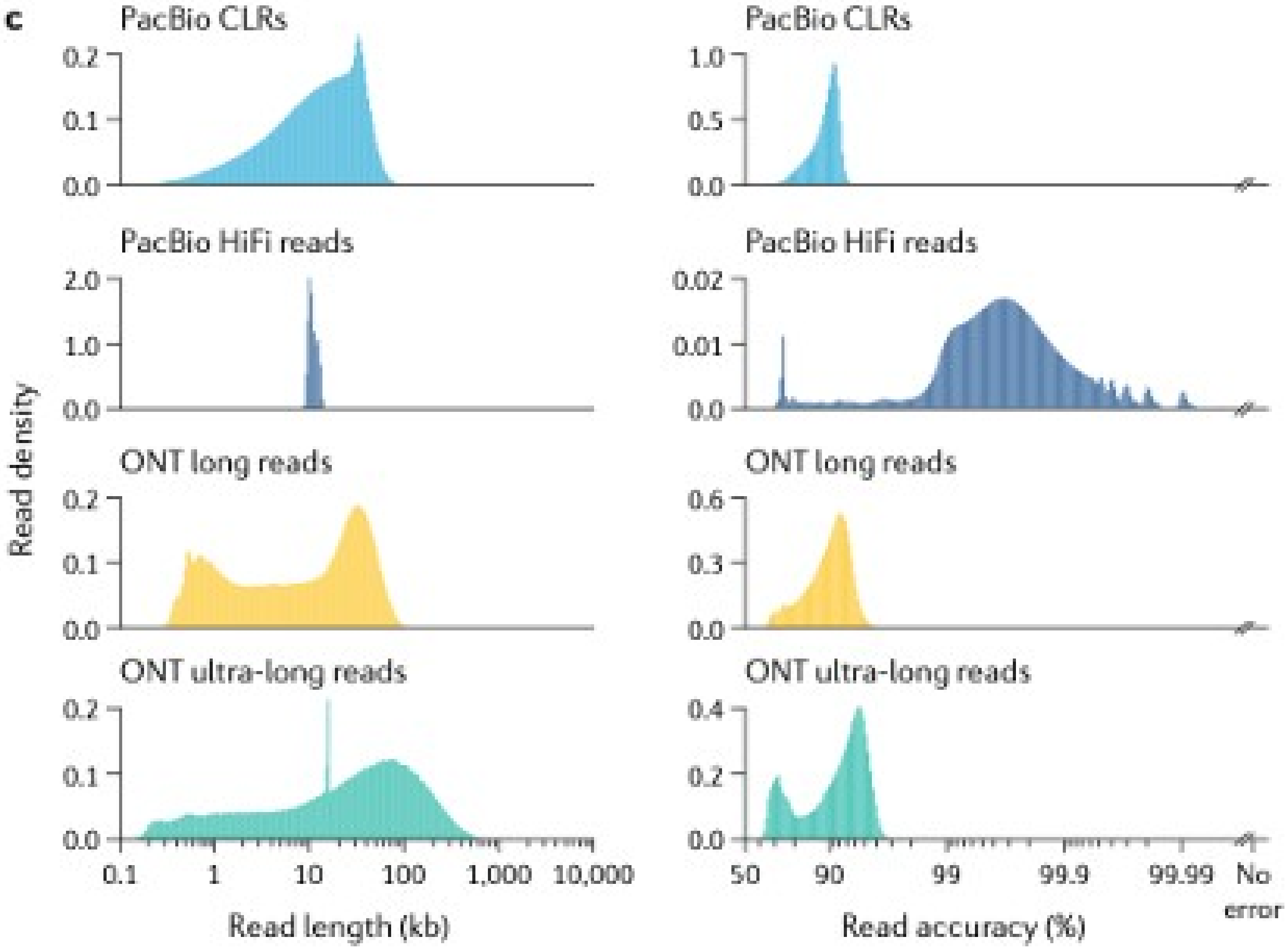
# Nanopore accuracy

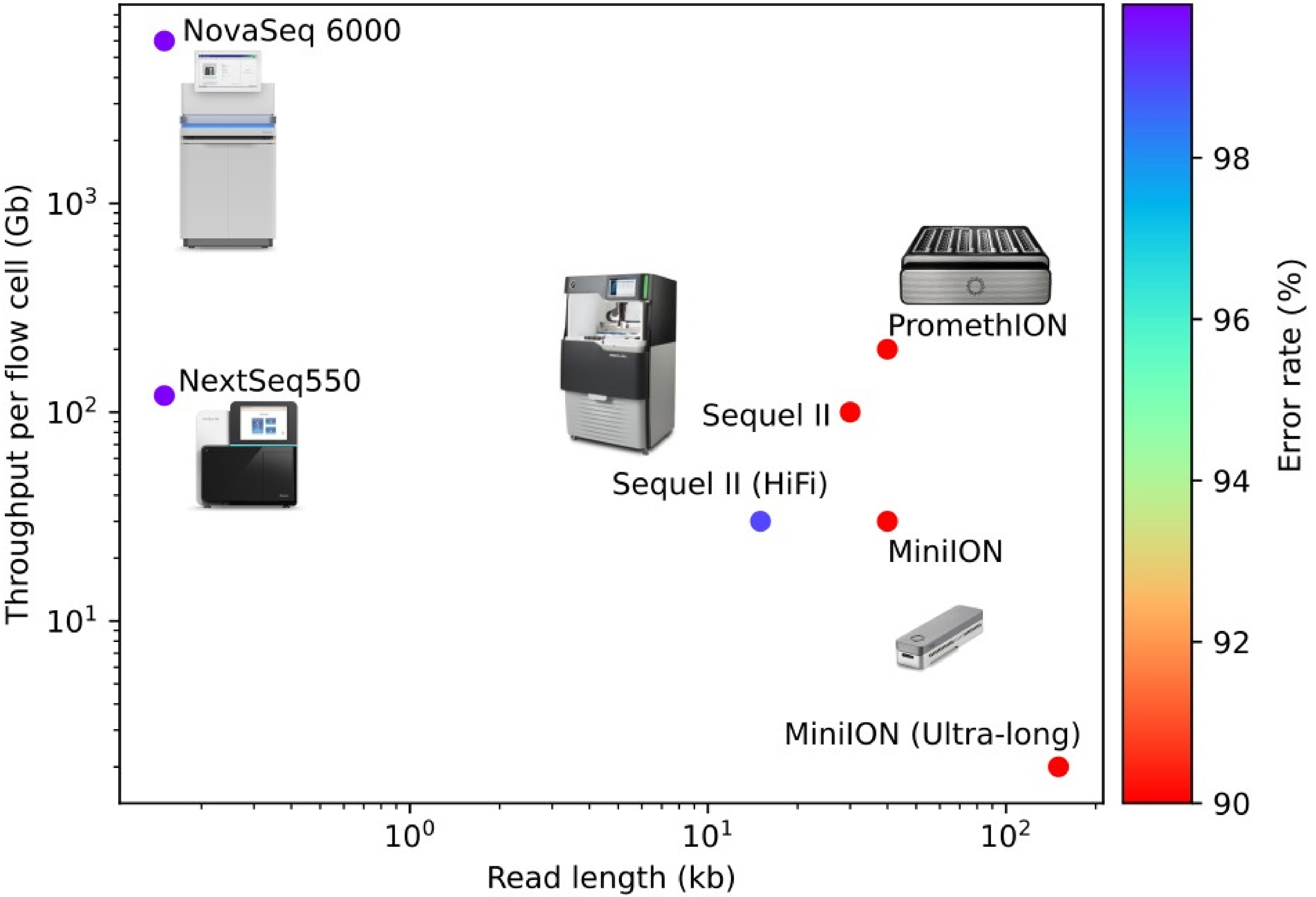


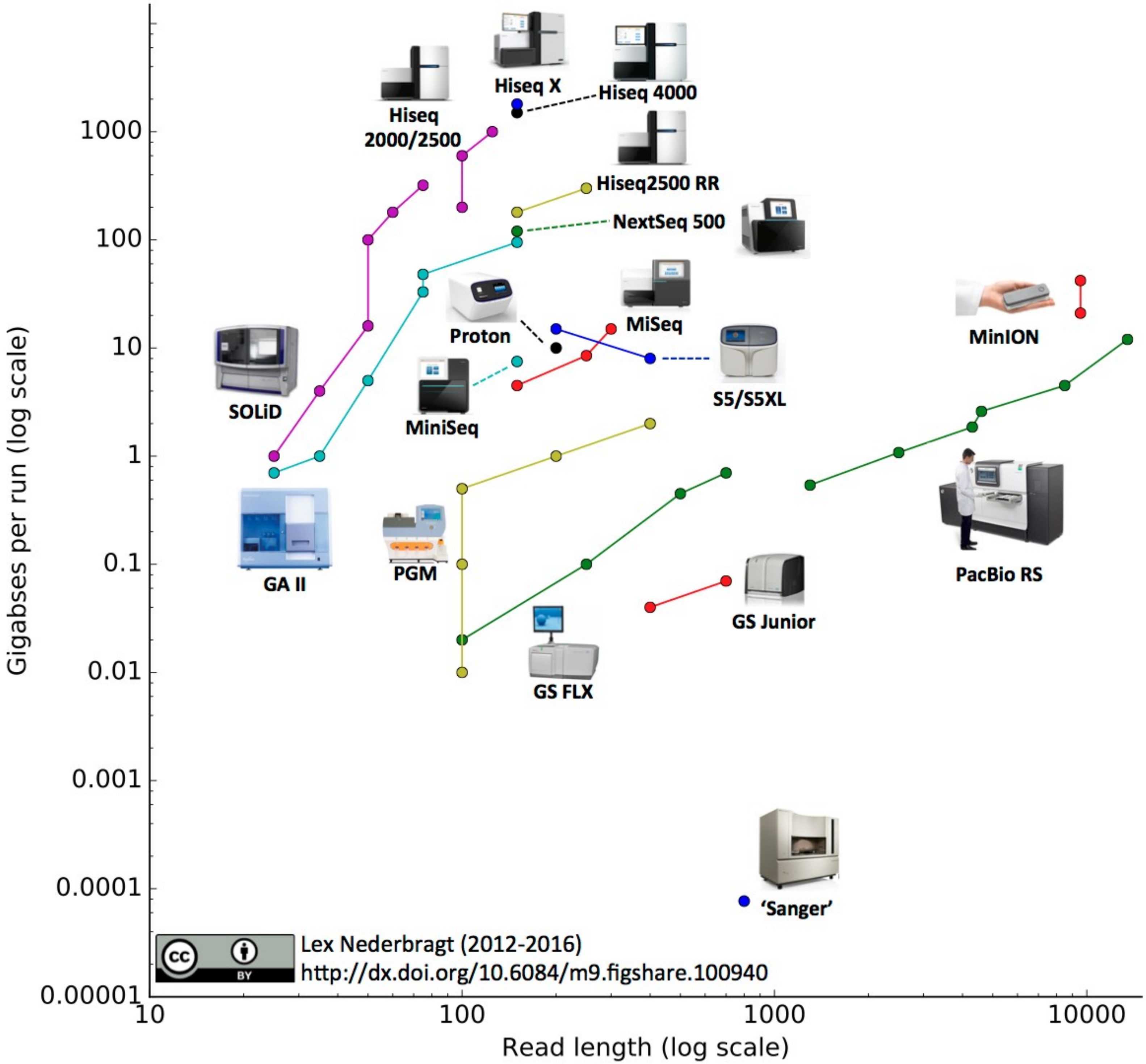
# Nanopore-Illumina hybrid error correction



# Long NGS reads comparison




NovaSeq 6000
NextSeq550
PromethION
Sequel II
Sequel II (HiFi)
MiniION
MiniION (Ultra-long)
Throughput per flow cell (Gb)
$10^3$
$10^2$
$10^1$
$10^0$
$10^1$
$10^2$
Read length (kb)
Error rate (%)
98
96
94
92
90



Lex Nederbragt (2012-2016)
http://dx.doi.org/10.6084/m9.figshare.100940

# Bioinformatic challenges

Huge data files handling.

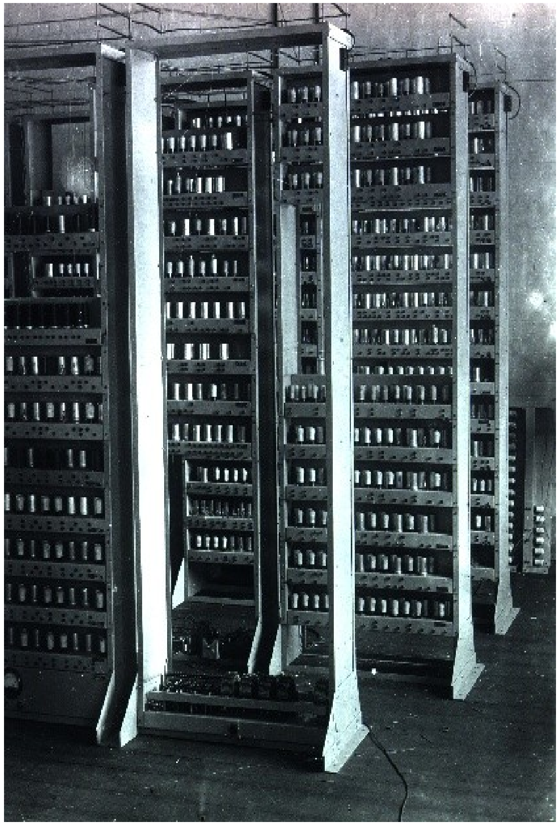
Beefy computers required.

Software still being developed or missing.

Ad-hoc software required during the analysis.

Existing software tailored to experienced bioinformaticians.

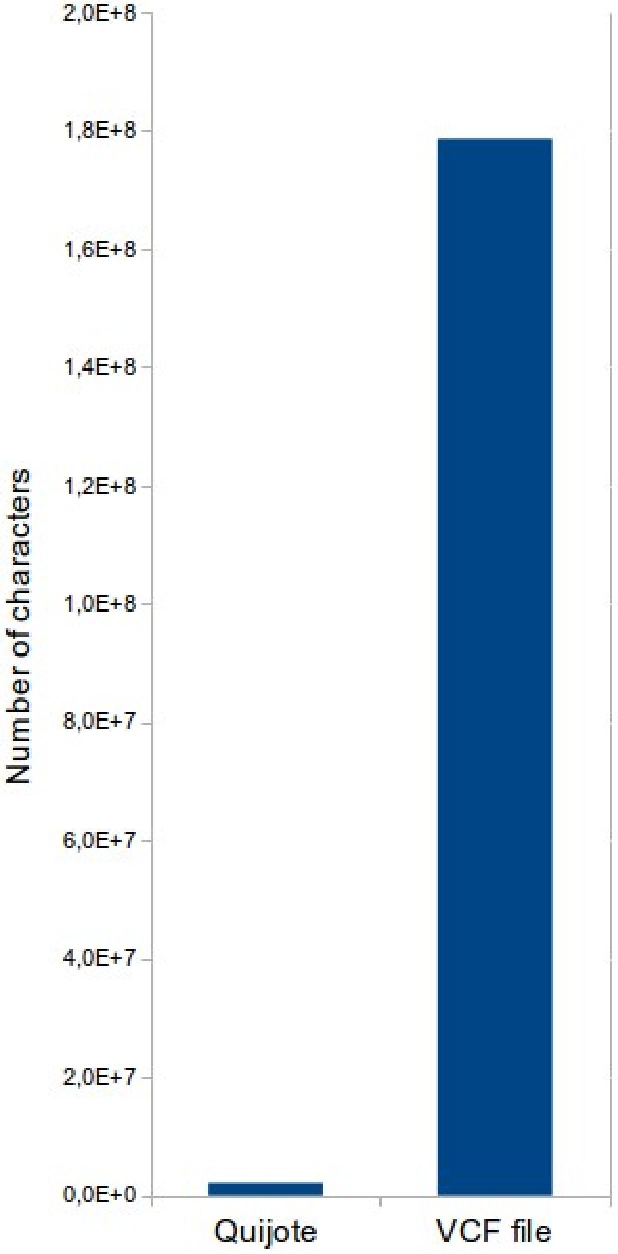
Dollar for dollar rule proposed

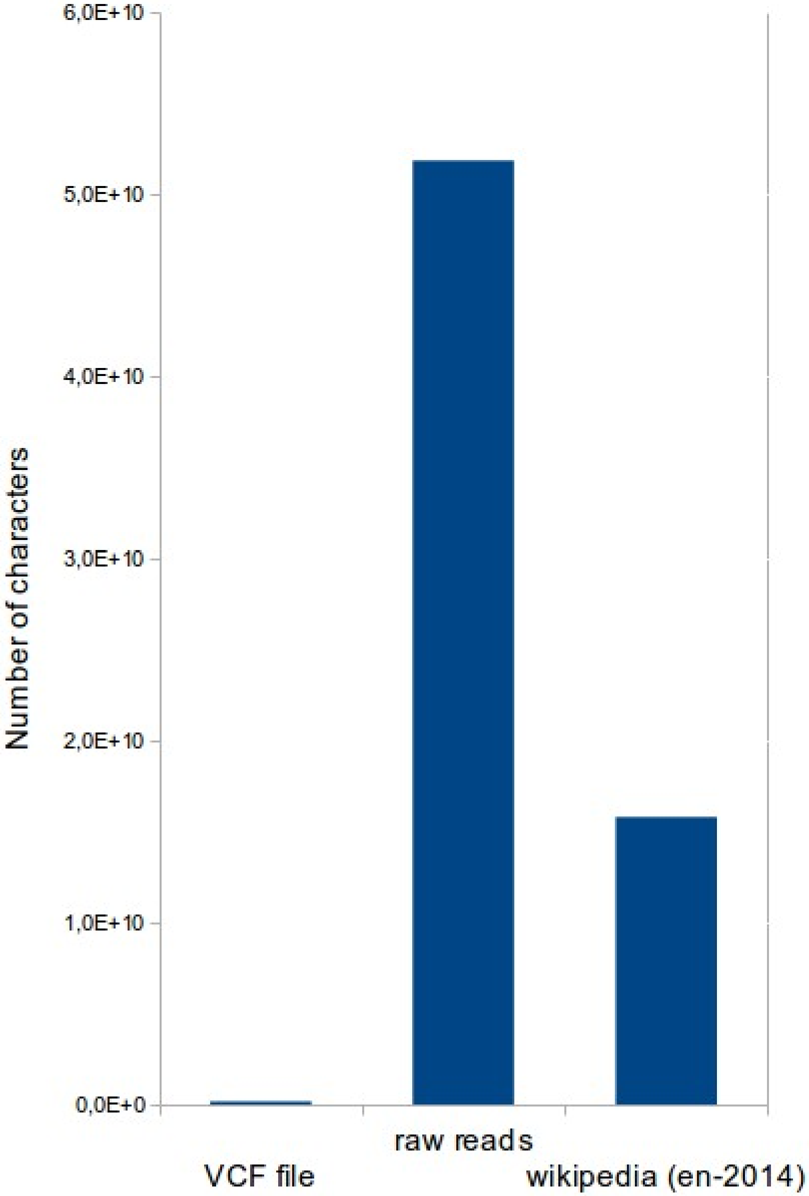


EDSAC by Computer Laboratory Cambridge

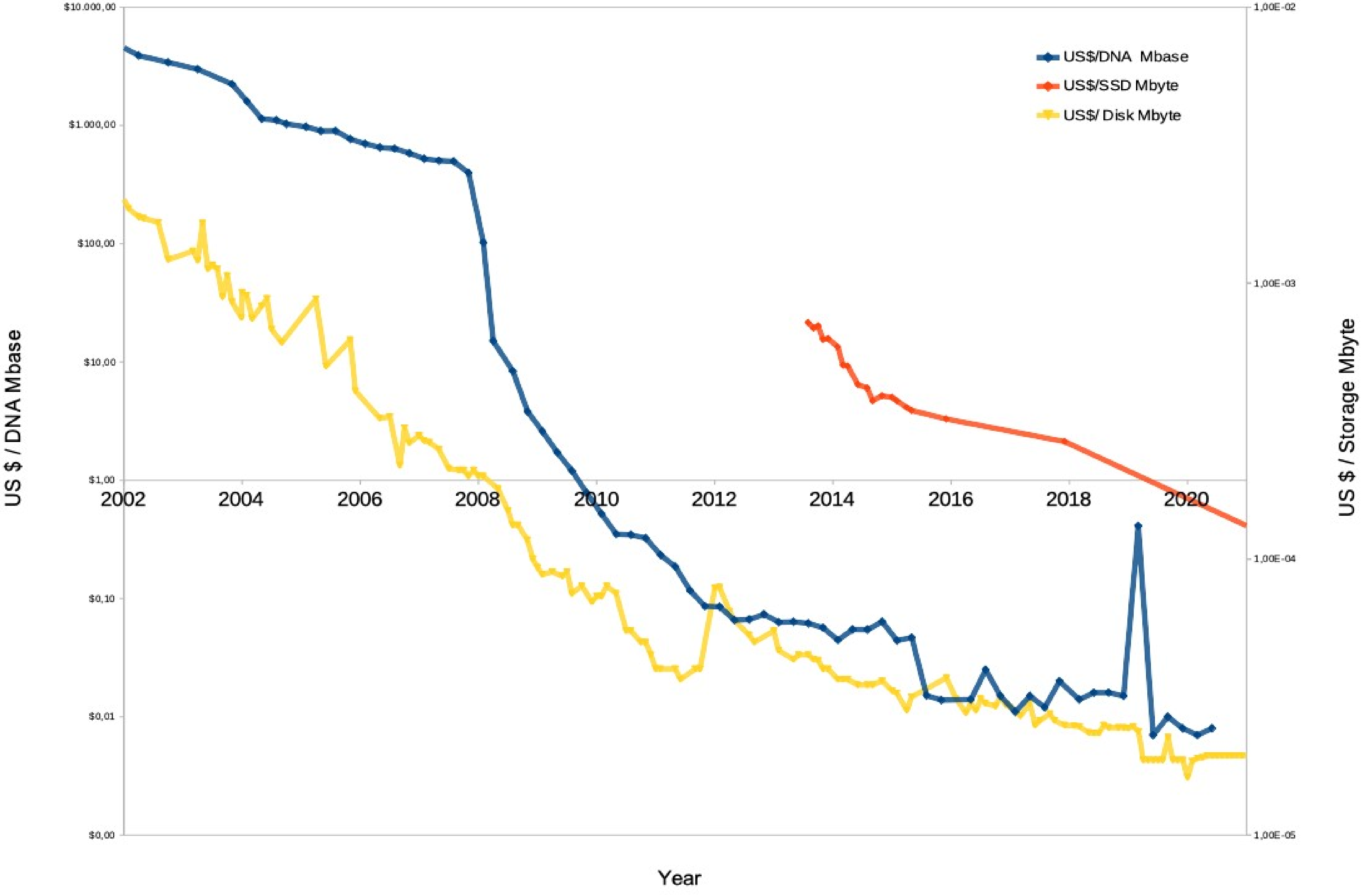
# Bioinformatic challenges

Amount of data managed on a small transcriptome assembly





# Sequencing and storage costs



# Reducing the complexity

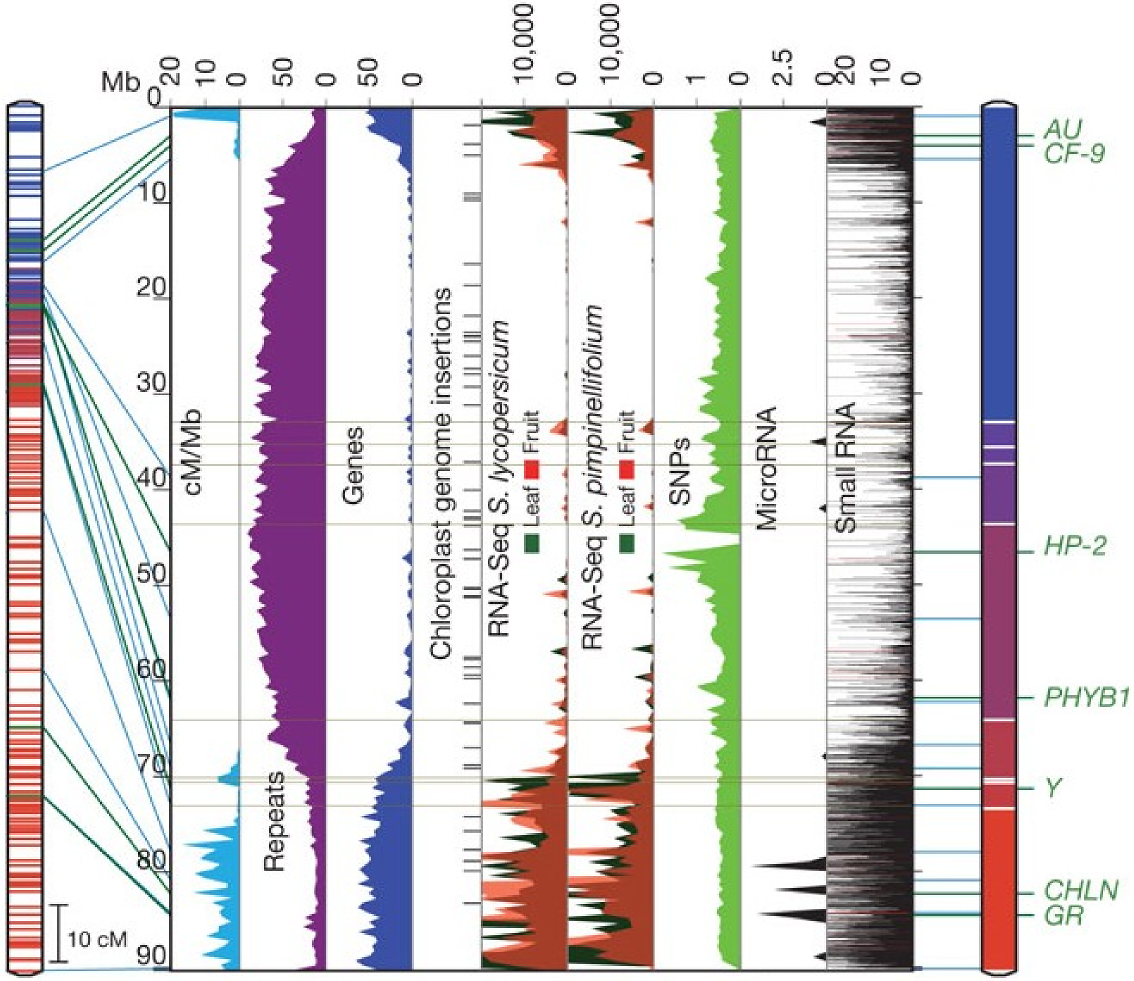
# Genome

Pros:

- Finest resolution
- Reproducible

Cons:

- Expensive ($600 per sample)
- Lots of information will be lost if no reference is available, especially in the repetitive regions.



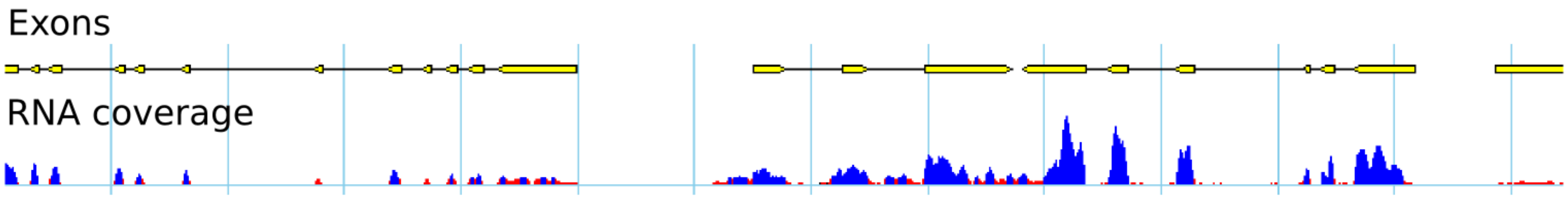
# RNASeq

Pros:

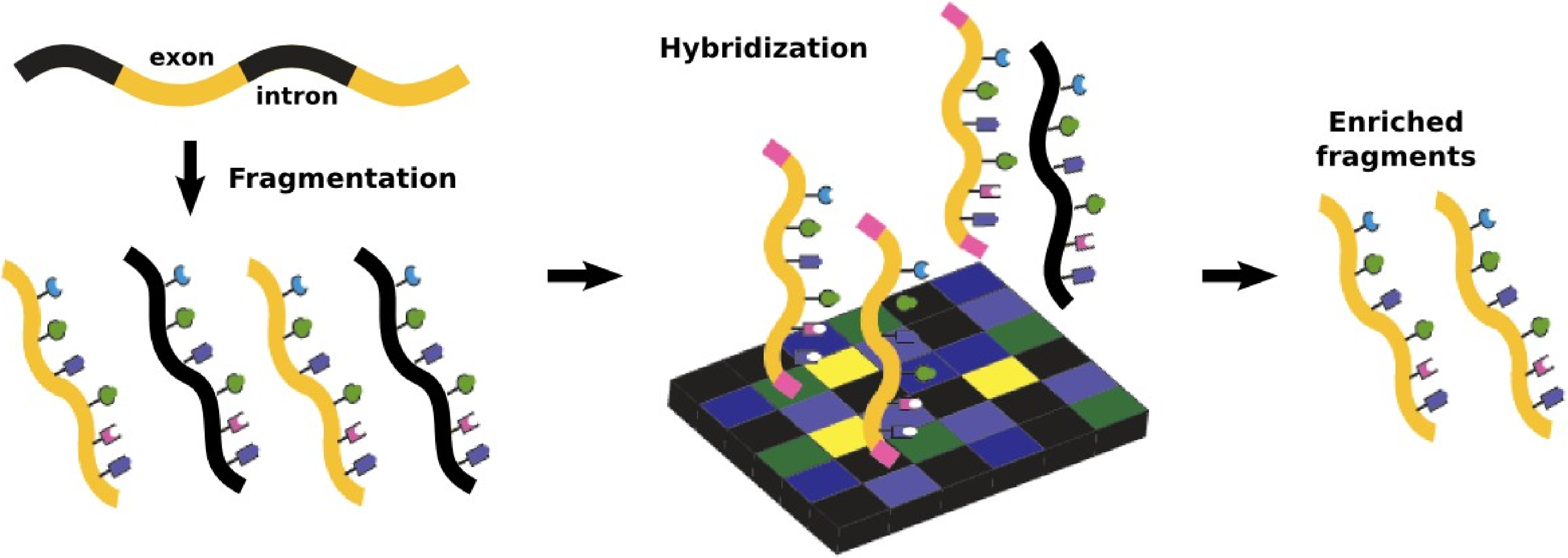
- Cheaper than whole genome sequencing ($300)
- Well proven methodologies
- Reproducible
- Follows gene density

Cons:

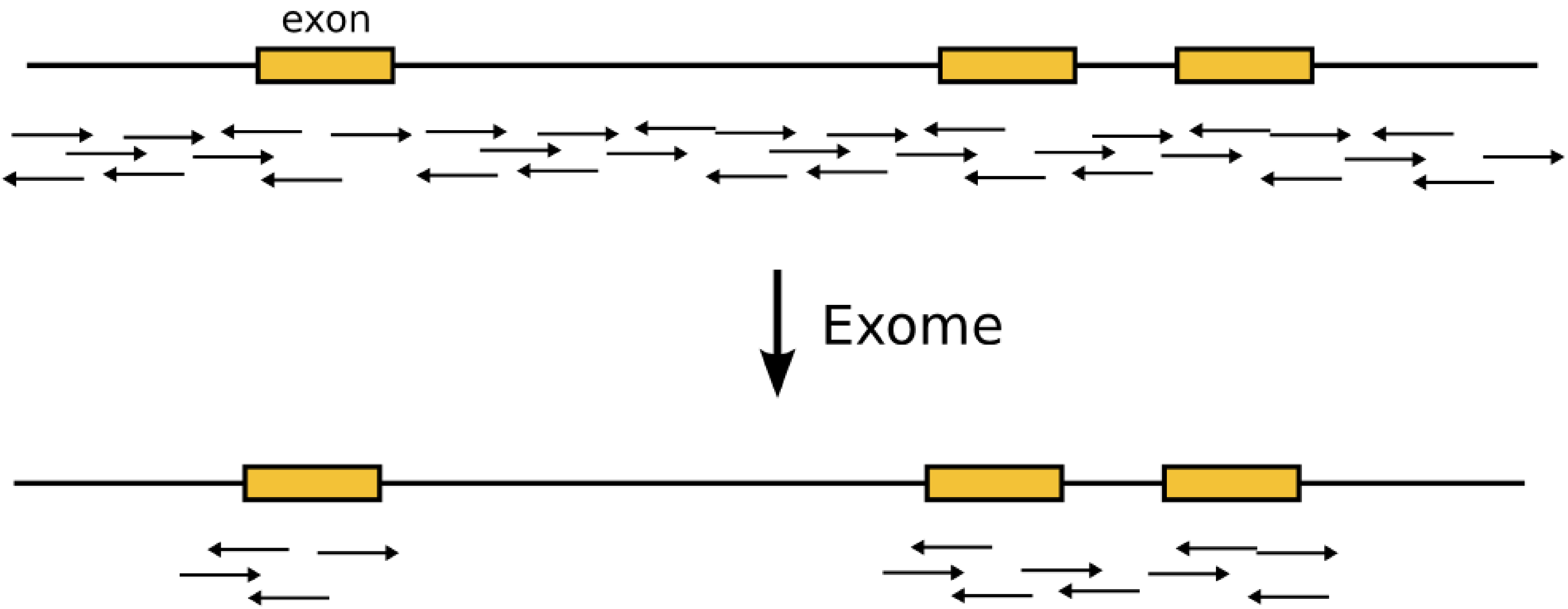
- RNA handling
- For many samples is pricier than GBS
- Follows gene density



# Exome


exon
intron
Fragmentation
Hybridization
Enriched fragments

# Exome



# Exome

Pros:

- More complete representation than RNASeq
- More reproducible than RNASeq

Cons:

- Exome capture platforms only available in model species
- Pricier than RNASeq

# Sequence capture

Targeted sequence capture as a powerful tool for evolutionary analysis
Am. J. Bot  doi: 10.3732/ajb.1100323

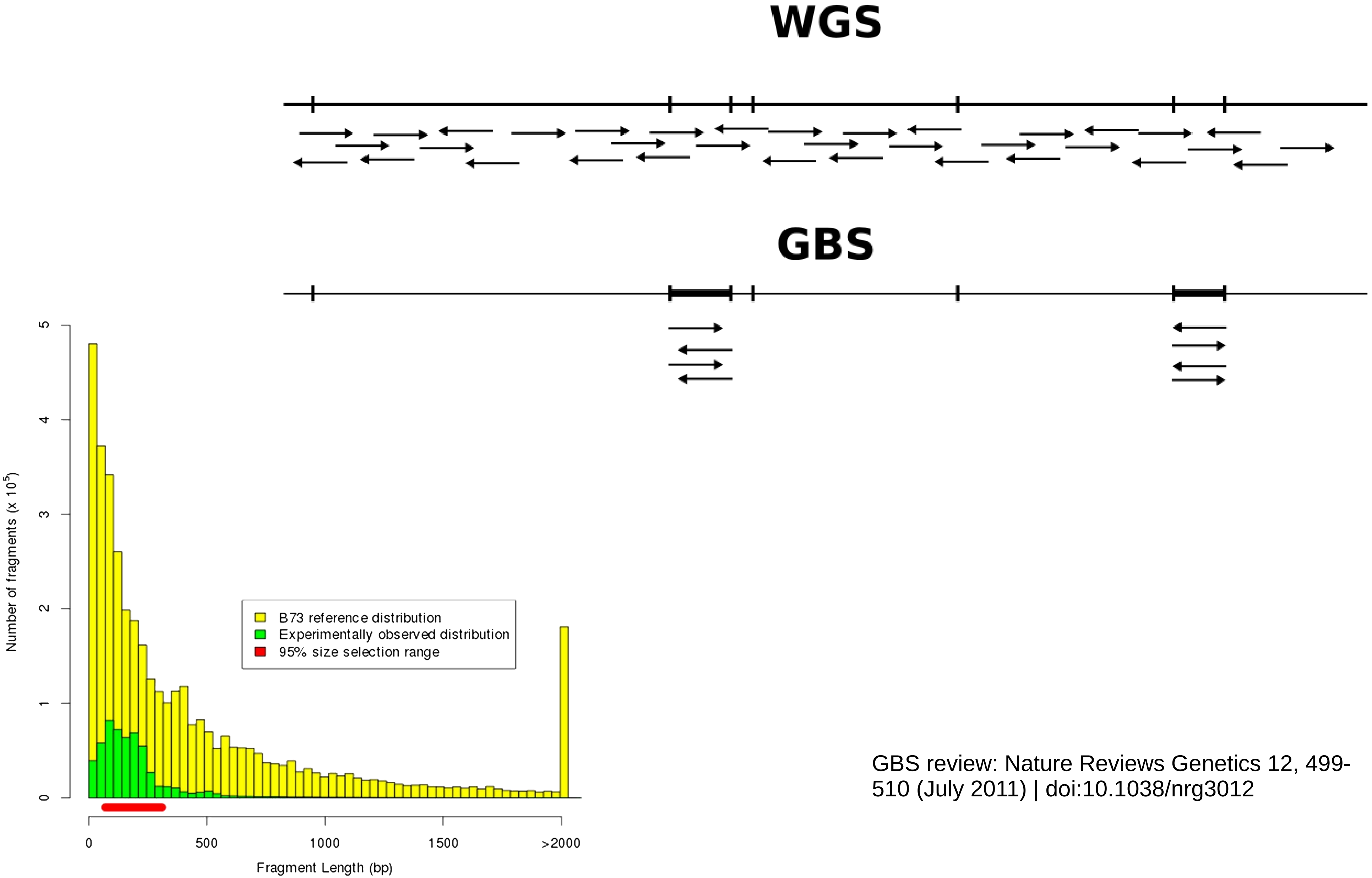
Hibridization against designed probes

From several targeted loci to over a million target regions

It is expensive to design and create the probe set

Costs per sample will depend on the number of probes

# Genotyping by Sequencing (GBS)



doi: 10.1534/genetics.112.147710

GBS review: Nature Reviews Genetics 12, 499-510 (July 2011) | doi:10.1038/nrg3012

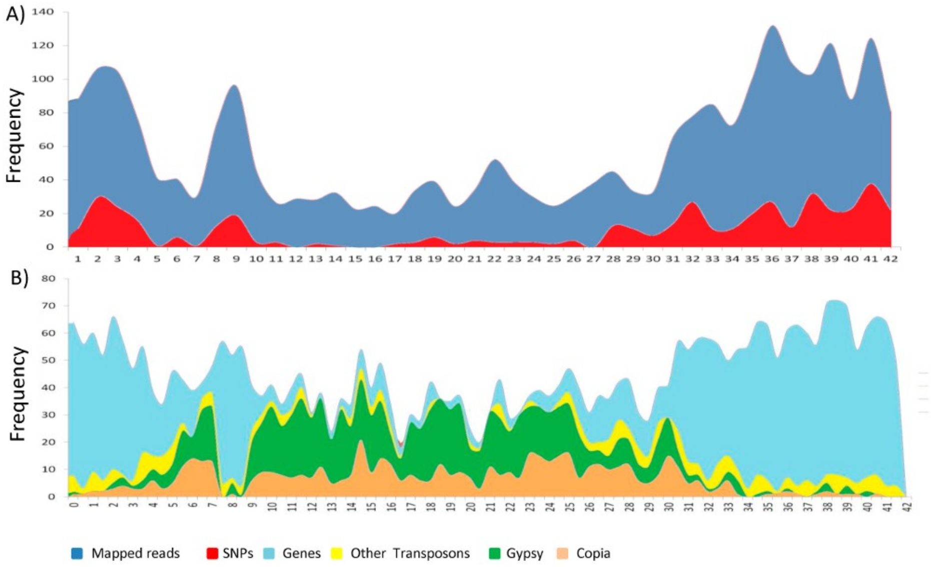
# Genotyping by Sequencing (GBS)

Pros:

- Cheap ($50 per sample)
- Lots of variation

Cons:

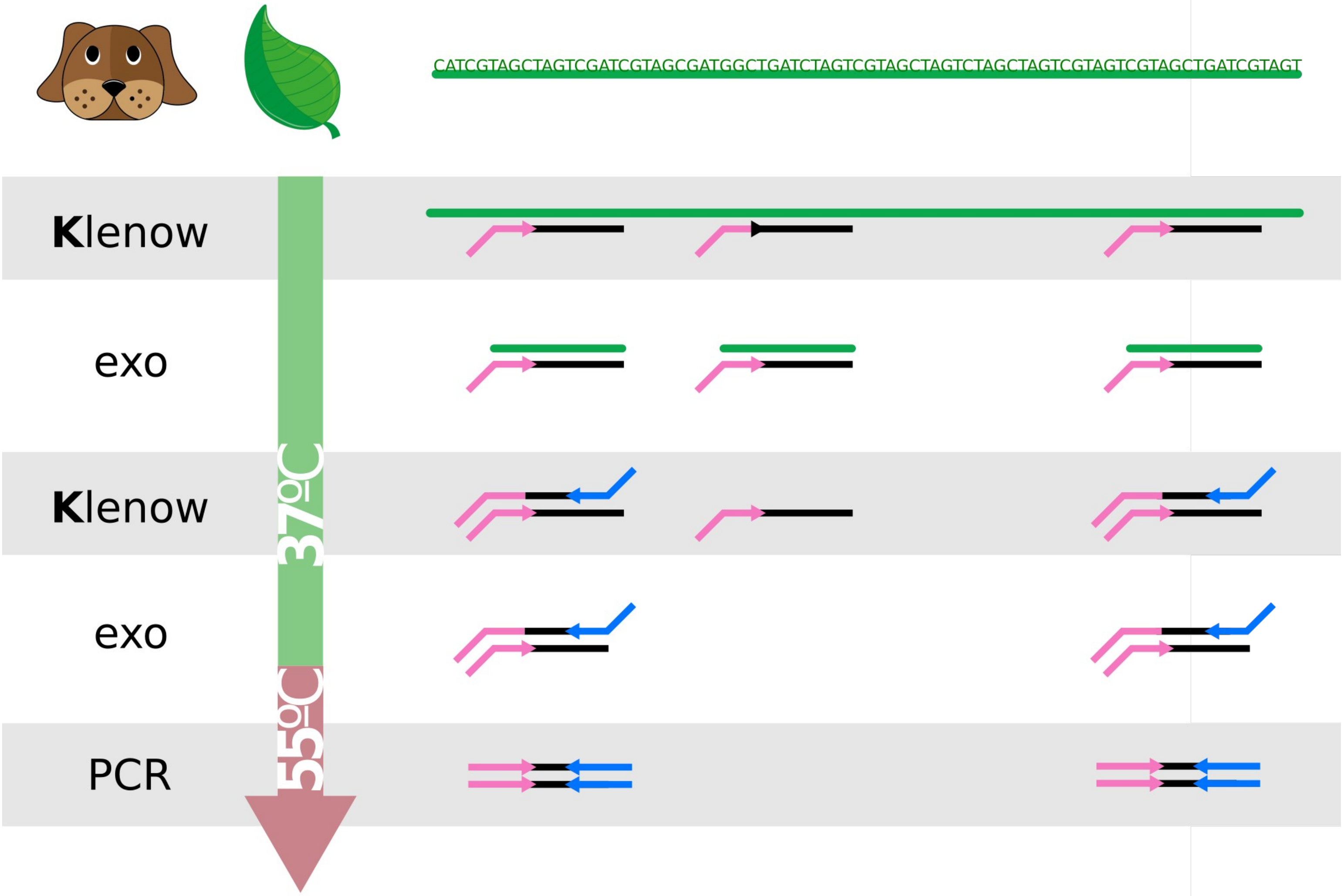
- Prone to artifacts (e.g. false SNPs due to repetitive DNA) if no reference genome is available.
- Degree of coverage along the genome depends on the Restriction Enzyme chosen
- How reproducible is it?
- Patent trolls



GBS based SNPs in Soy

doi: 10.1534/genetics.112.147710

# K-seq



CATCGTAGCTAGTCGATCGTAGCGATGGCTGATCTAGTCGTAGCTAGTCTAGCTAGTCGTAGTCGTAGCTGATCGTAGT

**K**lenow

exo

**K**lenow

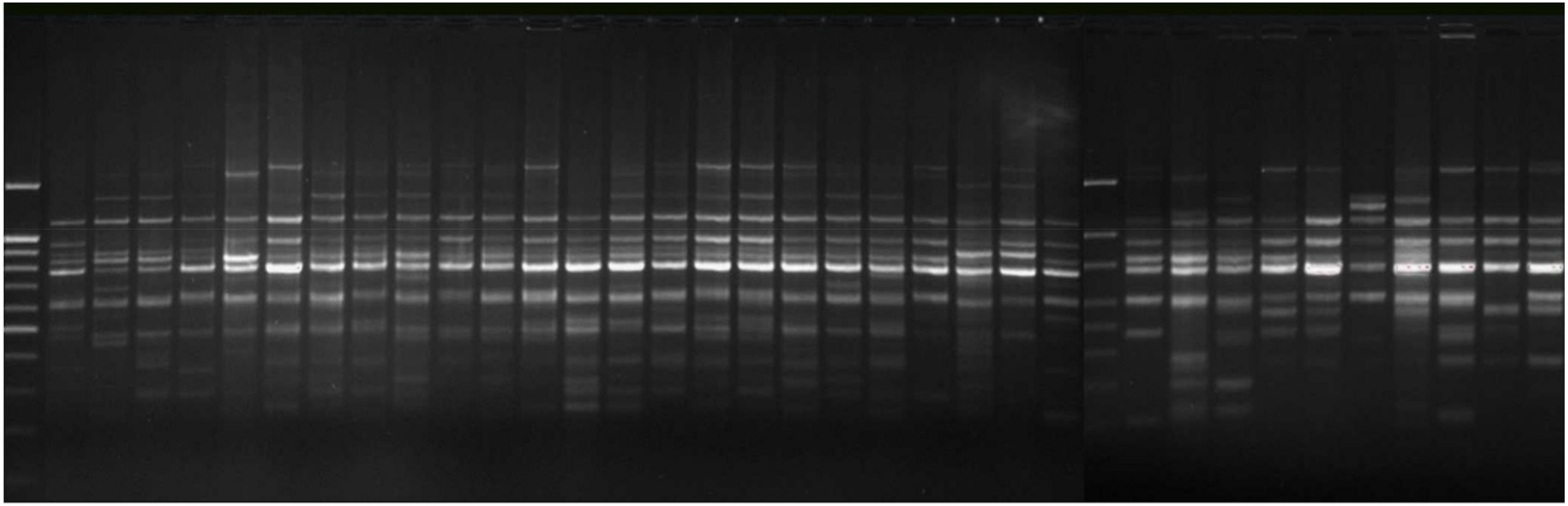
exo

PCR

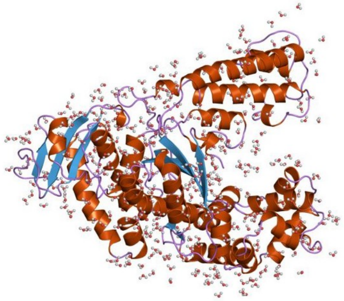
37ºC

55ºC

# El fracaso RAPD



# Inestabilidad térmica de los cebadores cortos



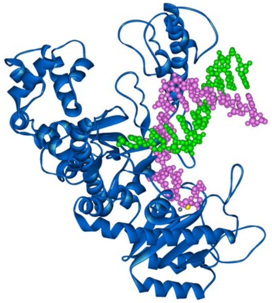
Polimerasa Taq

*Thermus aquaticus*

Funciona entre 50-72ºC

```
CTAGCTAGCTGACGTAGCTGATGCTATCTAGCTACGTAGCTACTACGAGTCGATGCTAGTCATGTCGTA
                            ||||||||
                            TAGCTACG
```

# Klenow



Polimerasa klenow

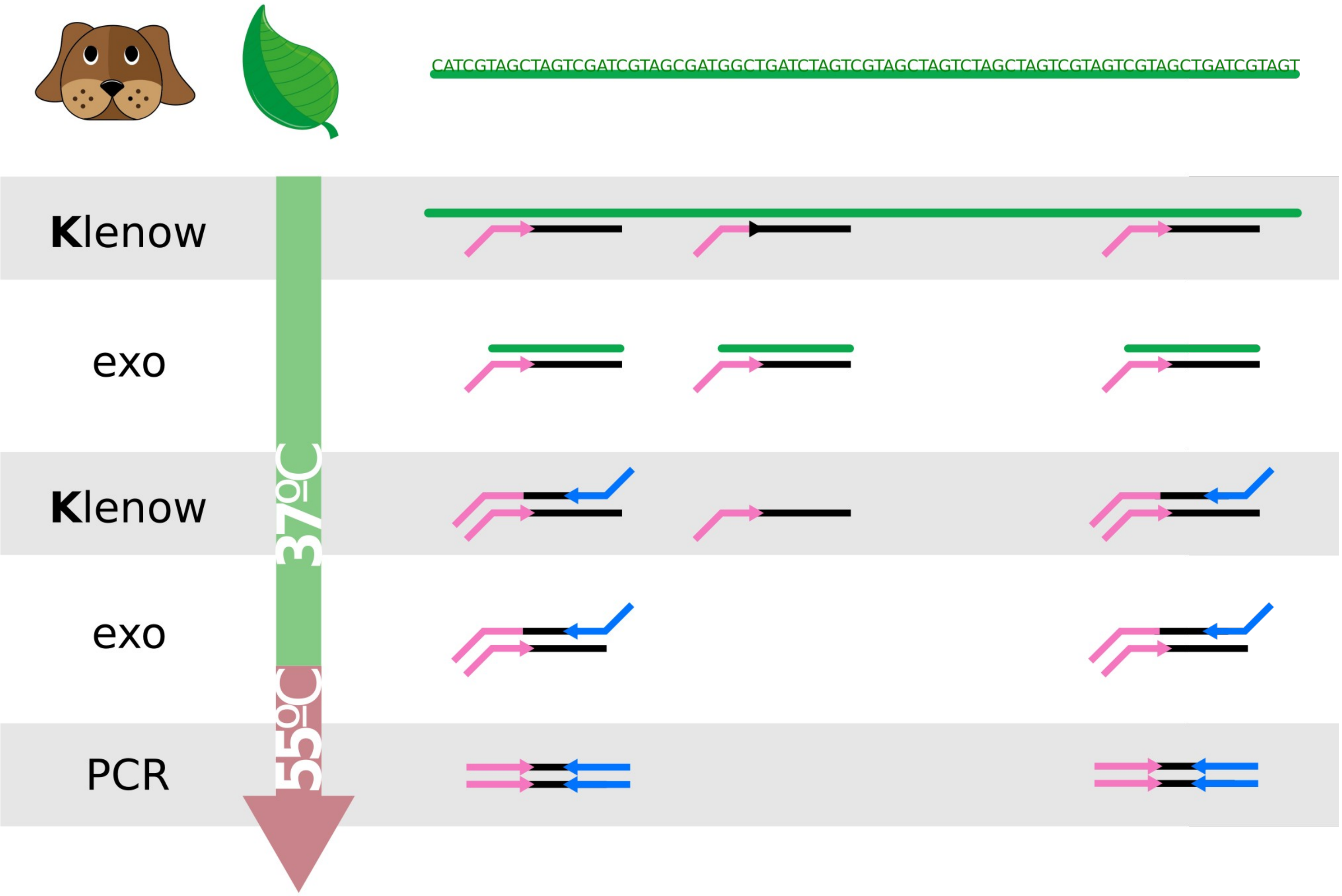
*E. coli*

Funciona a 37ºC

Se destruye a 95ºC

```
CTAGCTAGCTGACGTAGCTGATGCTATCTAGCTACGTAGCTACTACGAGTCGATGCTAGTCATGTCGTA
                            ||||||||
                            TAGCTACG
```

# K-seq



CATCGTAGCTAGTCGATCGTAGCGATGGCTGATCTAGTCGTAGCTAGTCTAGCTAGTCGTAGTCGTAGCTGATCGTAGT

**K**lenow

exo

**K**lenow

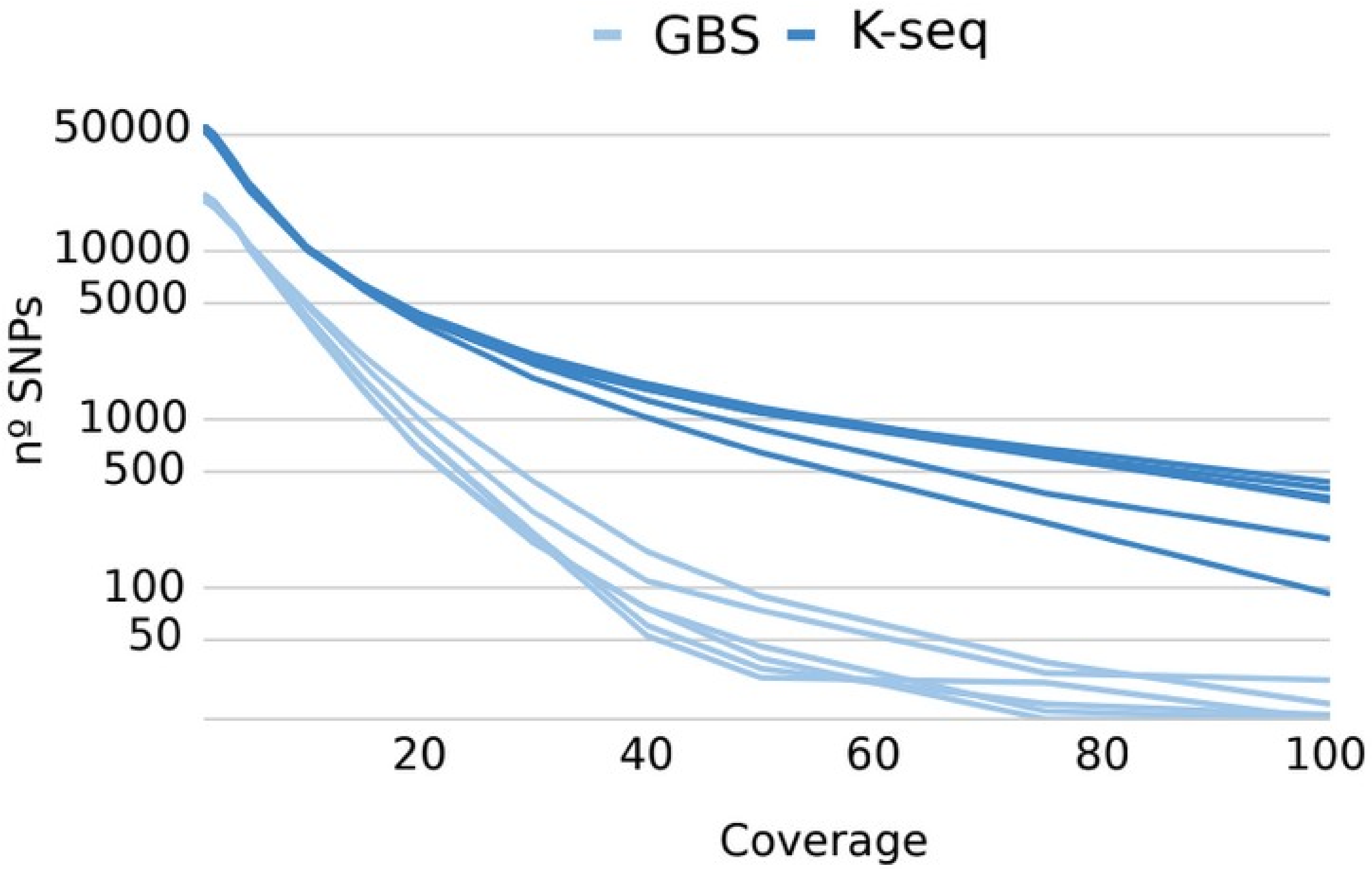
exo

PCR

37ºC

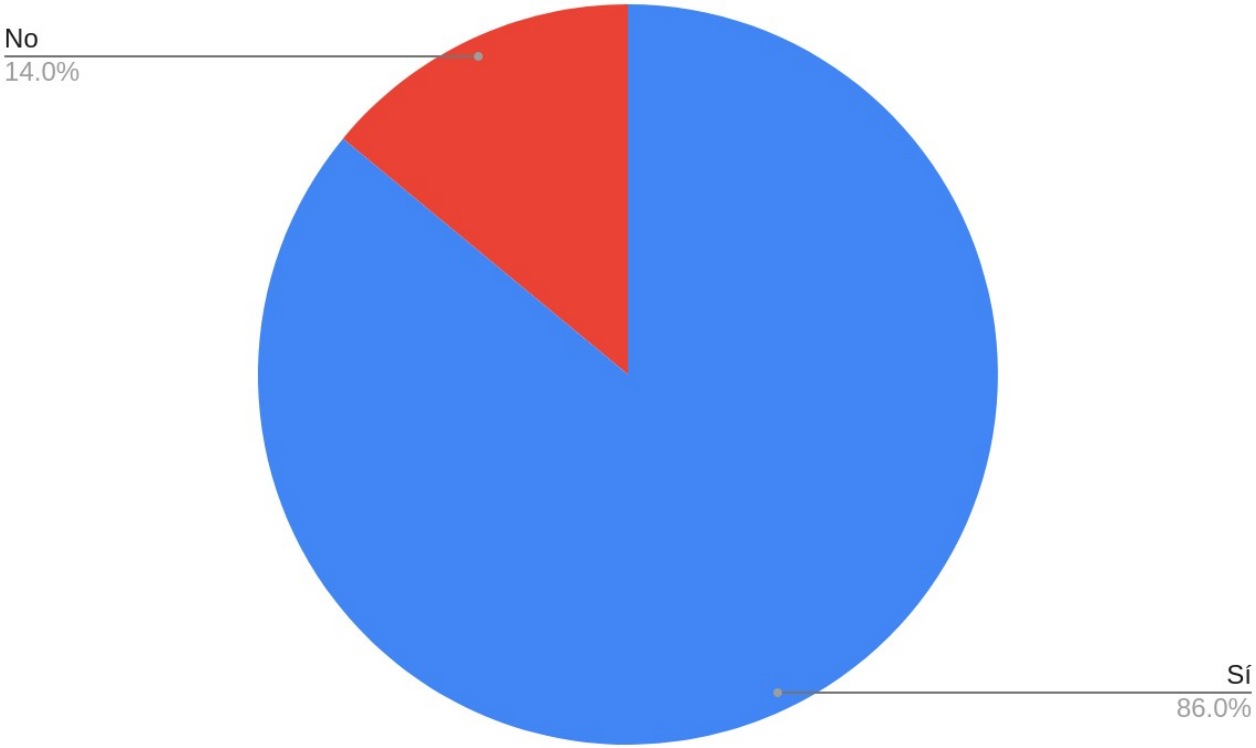
55ºC

# GBS vs K-seq

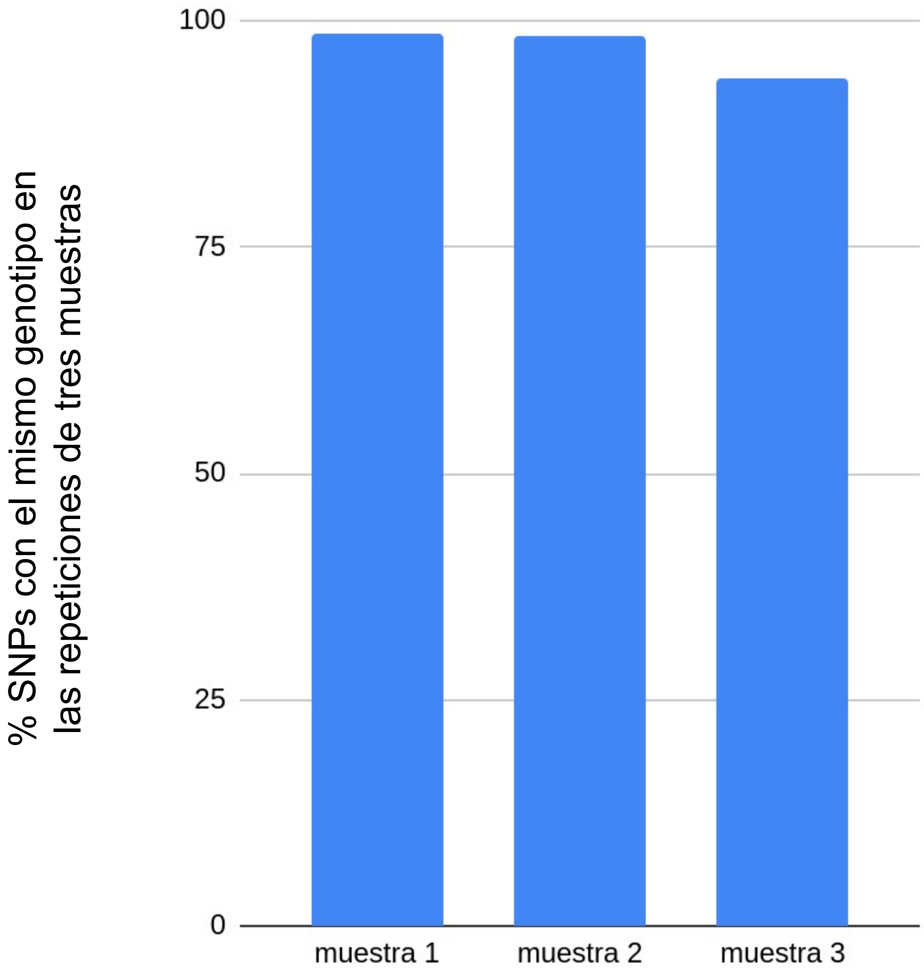

GBS
K-seq
50000
10000
5000
1000
500
100
50
nº SNPs
20
40
60
80
100
Coverage

# Reproducibilidad

% SNPs genotipados en tres muestras independientes



# Reproducibilidad



# Funciona en especies cercanas con los mismos cebadores

Cebadores solanaceas: tomate, patata, pimiento, berenjena, petunia

# Amplicons

Pros:

- Cheap for few genes
- Amplicon sets can be ordered, but the design is expensive

Cons:

- Not scalable for lots of genes
- Previous sequence information is required

Jose Blanca
COMAV institute
bioinf.comav.upv.es