Metagenomics 101

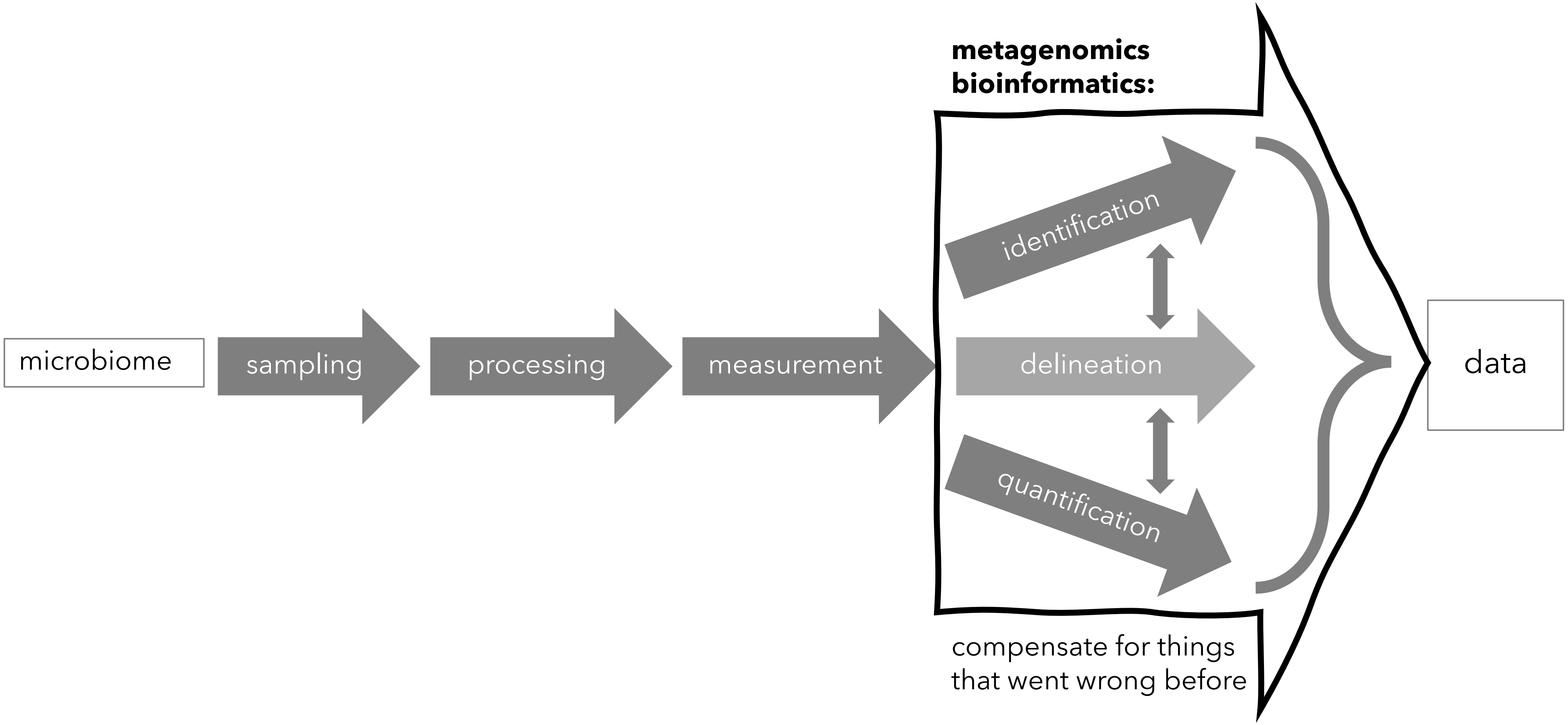
Session 2: Raw data & QC

Anna Heintz-Buschart

February 2022

UNIVERSITY OF AMSTERDAM
Life Sciences

Omics paradigm



metagenomics bioinformatics:

microbiome

sampling

processing

measurement

identification

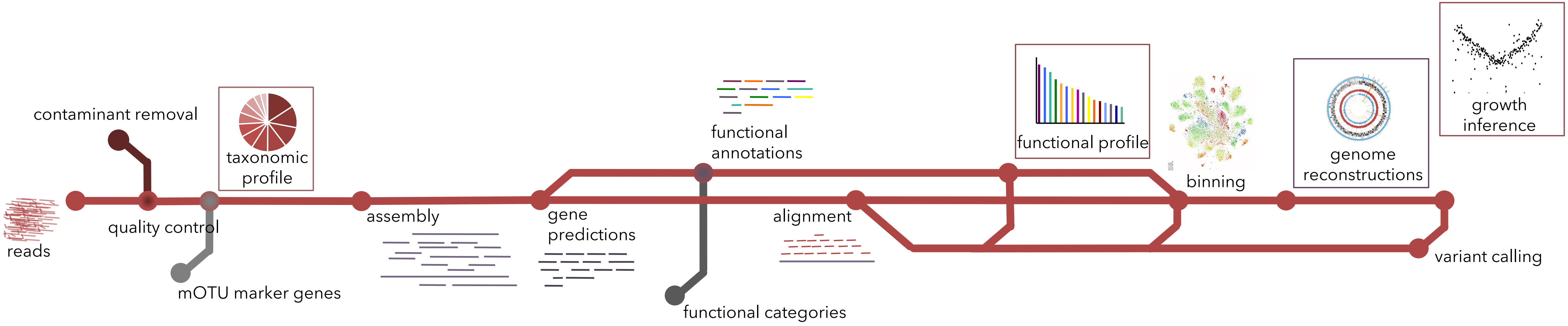
delineation

quantification

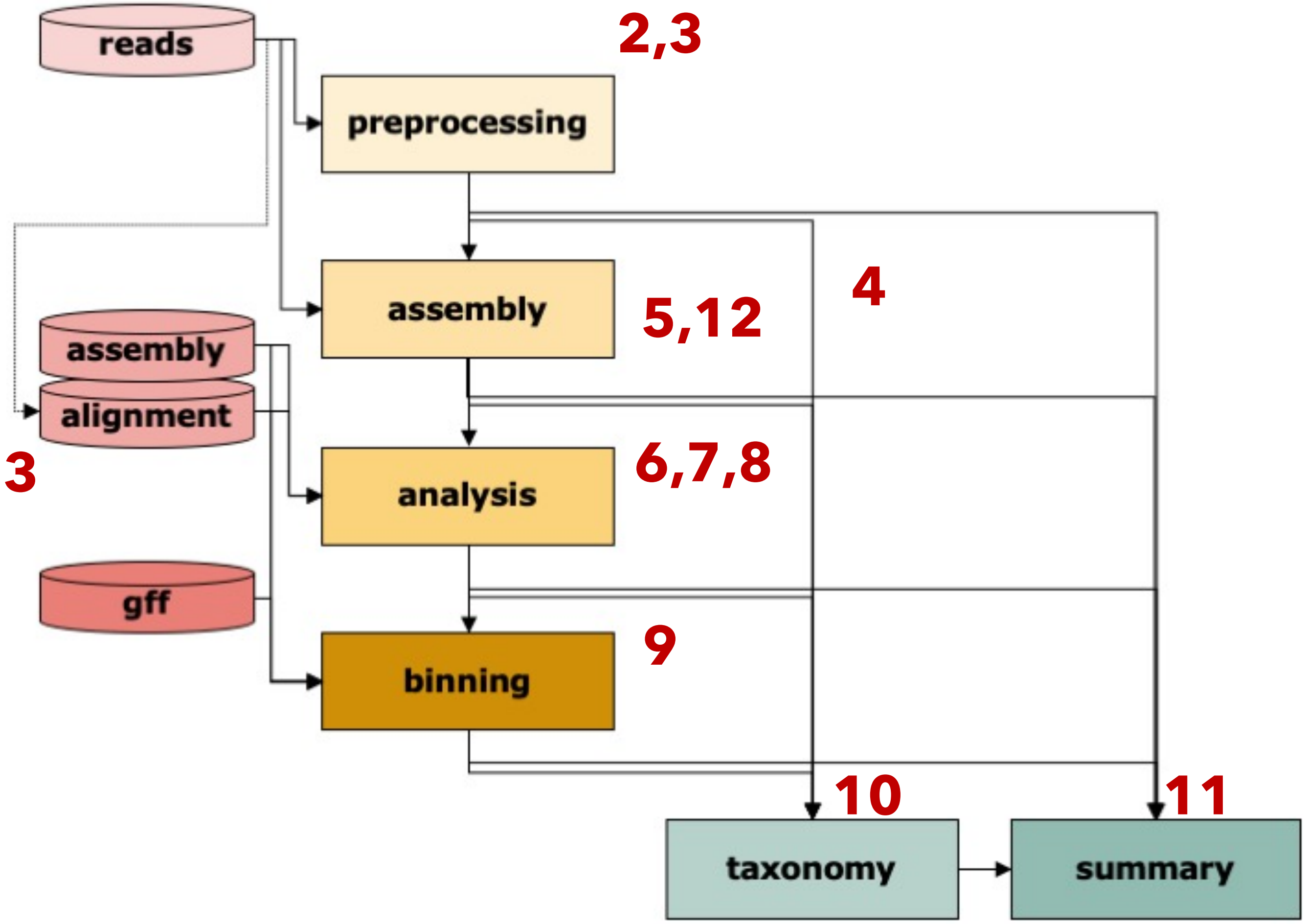
data

compensate for things that went wrong before

Metagenomics (+ other omics) pipeline



Metagenomics (+ other omics) pipeline



reads

preprocessing

2,3

assembly

alignment

3

assembly

5,12

4

analysis

6,7,8

gff

binning

9

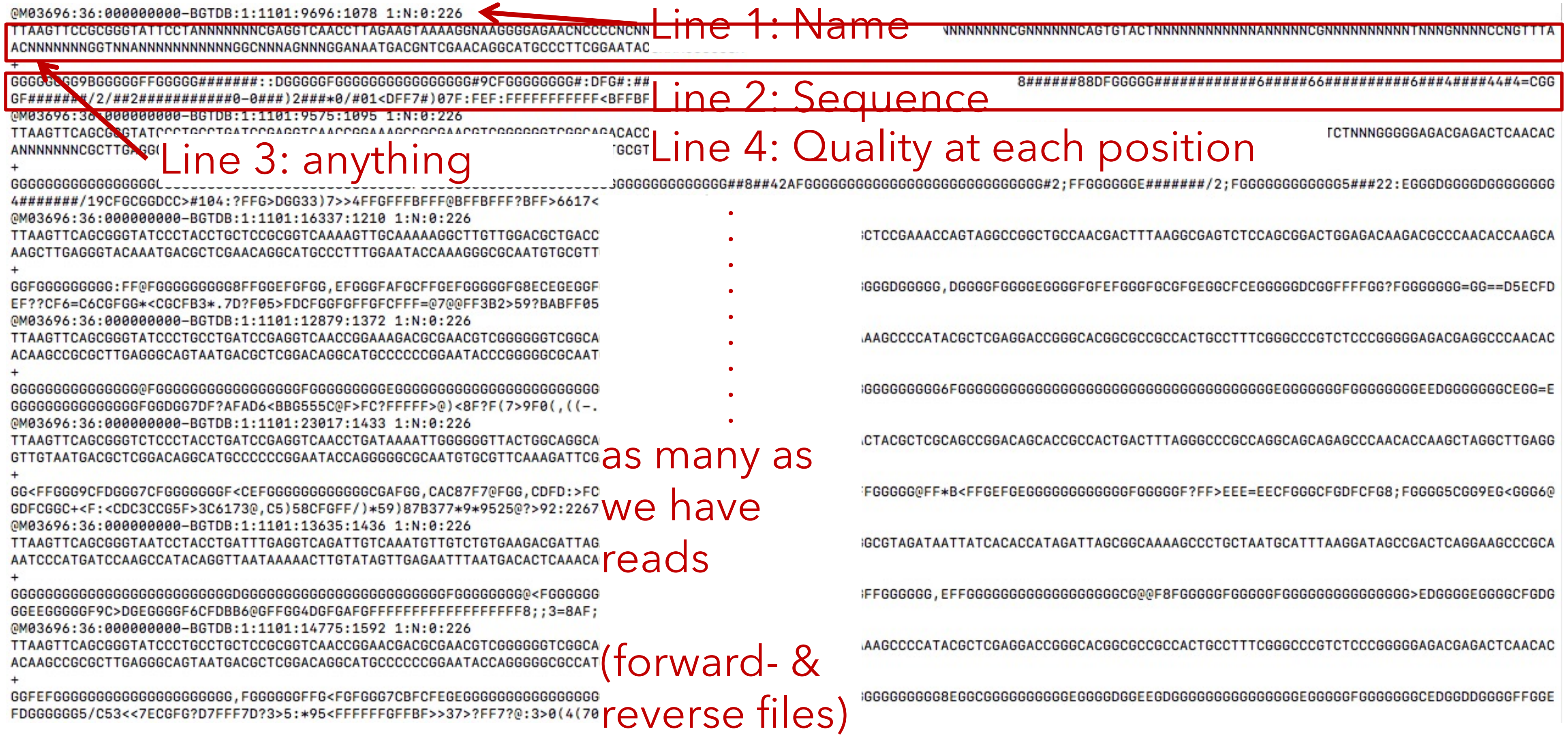
taxonomy

10

summary

11

What does sequencing data look like?



```
@M03696:36:000000000-BGTDB:1:1101:9696:1078 1:N:0:226
TTAAGTTCCGCGGGTATTCCTANNNNNNNNCGAGGTCAACCTTAGAAGTAAAAGGNAAGGGGAGAACNCCCCNCNN      NNNNNNNNCGNNNNNNCAGTGTACTNNNNNNNNNNNNNANNNNNCGNNNNNNNNNNTNNNGNNNNCCNGTTTA
ACNNNNNNNGGTNNANNNNNNNNNNNNGGCNNNAGNNNGGANAATGACGNTCGAACAGGCATGCCCTTCGGAATAC
+
GGGGGGG9BGGGGGFFGGGGG#######::DGGGGGFGGGGGGGGGGGGGGGG#9CFGGGGGGGG#:DFG#:##      8######88DFGGGGG#############6#####66##########6###4#####44#4=CGG
GF#######/2/##2###########0-0###)2###*0/#01<DFF7#)07F:FEF:FFFFFFFFFFF<BFFBF
@M03696:36:000000000-BGTDB:1:1101:9575:1095 1:N:0:226
TTAAGTTCAGCGGGTATCCCTGCCTGATCCGAGGTCAACCGGAAAGCCGCGAACGTCGGGGGGTCGGCAGACACC      TCTNNNGGGGGAGACGAGACTCAACAC
ANNNNNNNCGCTTGAGG      TGCGT
+
GGGGGGGGGGGGGGGG      GGGGGGGGGGGGG##8##42AFGGGGGGGGGGGGGGGGGGGGGGGGGG#2;FFGGGGGGE#######/2;FGGGGGGGGGGGG5###22:EGGGDGGGGDGGGGGGG
4#######/19CFGCGGDCC>#104:?FFG>DGG33)7>>4FFGFFFBFFF@BFFBFFF?BFF>6617<
@M03696:36:000000000-BGTDB:1:1101:16337:1210 1:N:0:226
TTAAGTTCAGCGGGTATCCCTACCTGCTCCGCGGTCAAAAGTTGCAAAAAGGCTTGTTGGACGCTGACC      CTCCGAAACCAGTAGGCCGGCTGCCAACGACTTTAAGGCGAGTCTCCAGCGGACTGGAGACAAGACGCCCAACACCAAGCA
AAGCTTGAGGGTACAAATGACGCTCGAACAGGCATGCCCTTTGGAATACCAAAGGGCGCAATGTGCGTT
+
GGFGGGGGGGGG:FF@FGGGGGGGGG8FFGGEFGFGG,EFGGGFAFGCFFGEFGGGGGFG8ECEGEGGF      GGGDGGGGG,DGGGGFGGGGEGGGGFGFEFGGGFGCGFGEGGCFCEGGGGGDCGGFFFFGG?FGGGGGGG=GG==D5ECFD
EF??CF6=C6CGFGG*<CGCFB3*.7D?F05>FDCFGGFGFFGFCFFF=@7@@FF3B2>59?BABFF05
@M03696:36:000000000-BGTDB:1:1101:12879:1372 1:N:0:226
TTAAGTTCAGCGGGTATCCCTGCCTGATCCGAGGTCAACCGGAAAGACGCGAACGTCGGGGGTCGGCA      AAGCCCCATACGCTCGAGGACCGGGCACGGCGCCGCCACTGCCTTTCGGGCCCGTCTCCCGGGGGAGACGAGGCCCAACAC
ACAAGCCGCGCTTGAGGGCAGTAATGACGCTCGGACAGGCATGCCCCCCGGAATACCCGGGGGCGCAAT
+
GGGGGGGGGGGGGGG@FGGGGGGGGGGGGGGGGFGGGGGGGGGEGGGGGGGGGGGGGGGGGGGGGGG      GGGGGGGGG6FGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGEGGGGGGGFGGGGGGGGEEDGGGGGGGCEGG=E
GGGGGGGGGGGGGGGFGGDGG7DF?AFAD6<BBG555C@F>FC?FFFFF>@)<8F?F(7>9F0(,((-.
@M03696:36:000000000-BGTDB:1:1101:23017:1433 1:N:0:226
TTAAGTTCAGCGGGTCTCCCTACCTGATCCGAGGTCAACCTGATAAAATTGGGGGGTTACTGGCAGGCA      CTACGCTCGCAGCCGGACAGCACCGCCACTGACTTTAGGGCCCGCCAGGCAGCAGAGCCCAACACCAAGCTAGGCTTGAGG
GTTGTAATGACGCTCGGACAGGCATGCCCCCCGGAATACCAGGGGGCGCAATGTGCGTTCAAAGATTCG
+
GG<FFGGG9CFDGGG7CFGGGGGGGF<CEFGGGGGGGGGGGGCGAFGG,CAC87F7@FGG,CDFD:>FC      FGGGGG@FF*B<FFGEFGEGGGGGGGGGGGGGFGGGGGF?FF>EEE=EECFGGGCFGDFCFG8;FGGGG5CGG9EG<GGG6@
GDFCGGC+<F:<CDC3CCG5F>3C6173@,C5)58CFGFF/)*59)87B377*9*9525@?>92:2267
@M03696:36:000000000-BGTDB:1:1101:13635:1436 1:N:0:226
TTAAGTTCAGCGGGTAATCCTACCTGATTTGAGGTCAGATTGTCAAATGTTGTCTGTGAAGACGATTAG      GCGTAGATAATTATCACACCATAGATTAGCGGCAAAAGCCCTGCTAATGCATTTAAGGATAGCCGACTCAGGAAGCCCGCA
AATCCCATGATCCAAGCCATACAGGTTAATAAAAACTTGTATAGTTGAGAATTTAATGACACTCAAACA
+
GGGGGGGGGGGGGGGGGGGGGGGGGGDGGGGGGGGGGGGGGGGGGGGGGGGGGFGGGGGGGG@<FGGGGGG      FFGGGGGG,EFFGGGGGGGGGGGGGGGGGCG@@F8FGGGGGFGGGGGFGGGGGGGGGGGGGG>EDGGGGEGGGGCFGDG
GGEEGGGGGF9C>DGEGGGGF6CFDBB6@GFFGG4DGFGAFGFFFFFFFFFFFFFFFFFF8;;3=8AF;
@M03696:36:000000000-BGTDB:1:1101:14775:1592 1:N:0:226
TTAAGTTCAGCGGGTATCCCTGCCTGCTCCGCGGTCAACCGGAACGACGCGAACGTCGGGGGGTCGGCA      AAGCCCCATACGCTCGAGGACCGGGCACGGCGCCGCCACTGCCTTTCGGGCCCGTCTCCCGGGGGAGACGAGACTCAACAC
ACAAGCCGCGCTTGAGGGCAGTAATGACGCTCGGACAGGCATGCCCCCCGGAATACCAGGGGGCGCCAT
+
GGFEFGGGGGGGGGGGGGGGGGGGGG,FGGGGGGFFG<FGFGGG7CBFCFEGEGGGGGGGGGGGGGGG      GGGGGGGGG8EGGCGGGGGGGGGGEGGGGDGGEEGDGGGGGGGGGGGGGGEGGGGGFGGGGGGGCEDGGDDGGGGFFGGE
FDGGGGGG5/C53<<7ECGFG?D7FFF7D?3>5:*95<FFFFFFGFFBF>>37>?FF7?@:3>0(4(70
```

Line 1: Name

Line 2: Sequence

Line 3: anything

Line 4: Quality at each position

as many as we have reads

(forward- & reverse files)

Where do errors come from?

Sequencing

short read sequencing

illumina®

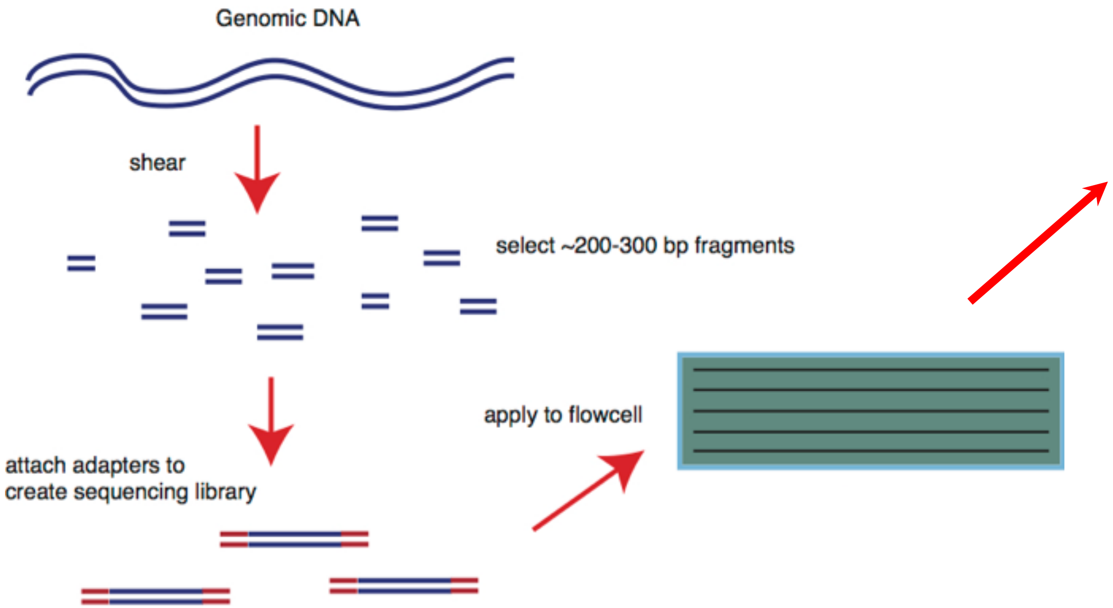
MGI

long read sequencing

PACBIO®

Oxford NANOPORE Technologies

Short read sequencing



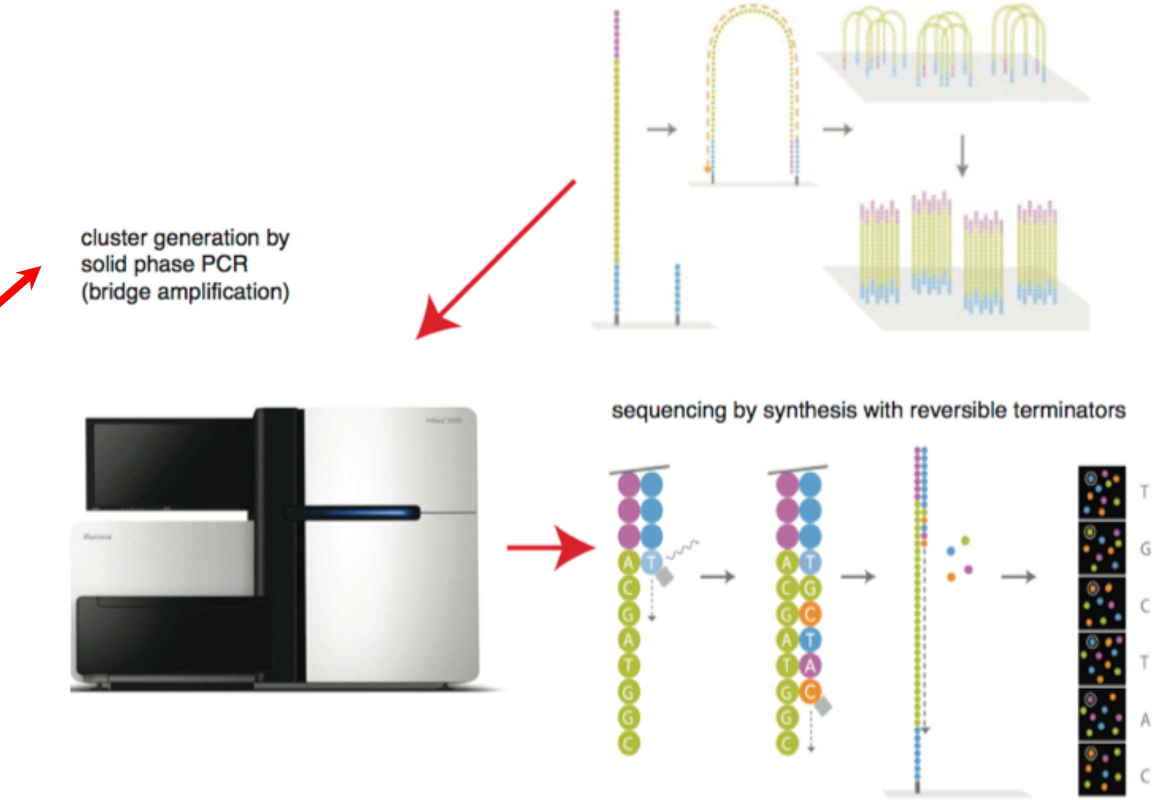
Genomic DNA

shear

select ~200-300 bp fragments

attach adapters to create sequencing library

apply to flowcell

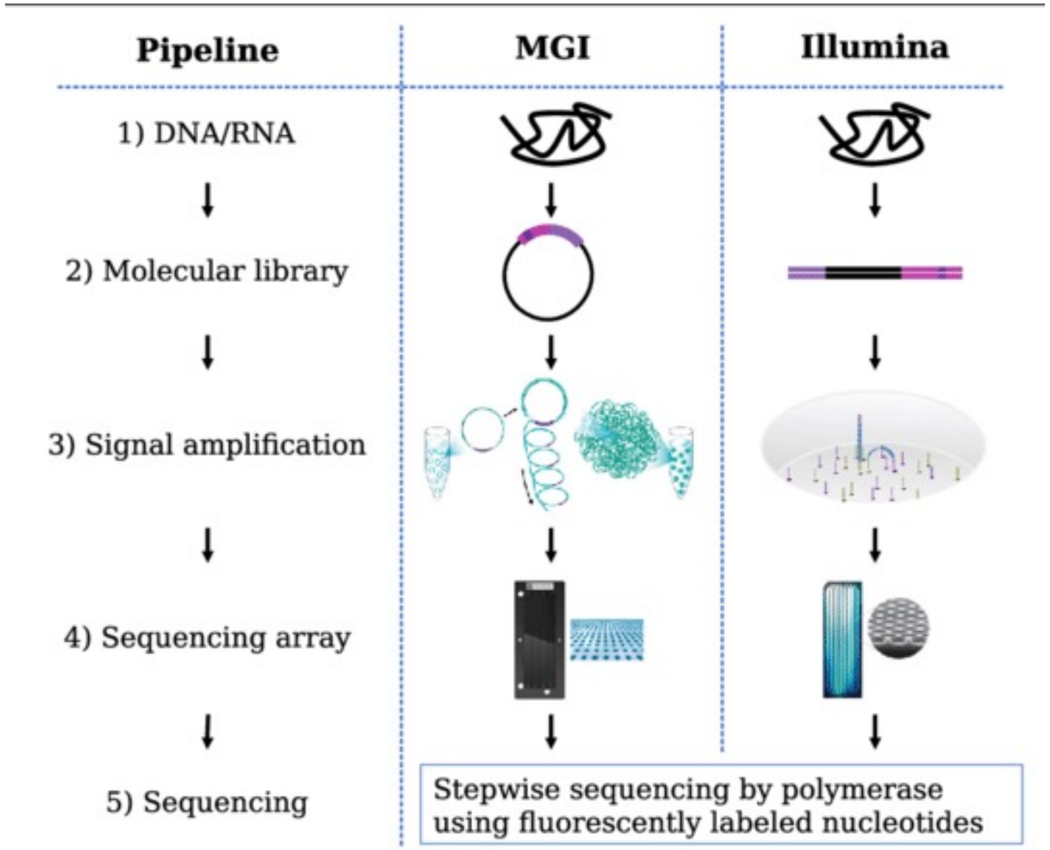


cluster generation by solid phase PCR (bridge amplification)

sequencing by synthesis with reversible terminators

https://bitesizebio.com/13546/sequencing-by-synthesis-explaining-the-illumina-sequencing-technology/

Short read sequencing



Pipeline | MGI | Illumina

1) DNA/RNA

2) Molecular library

3) Signal amplification

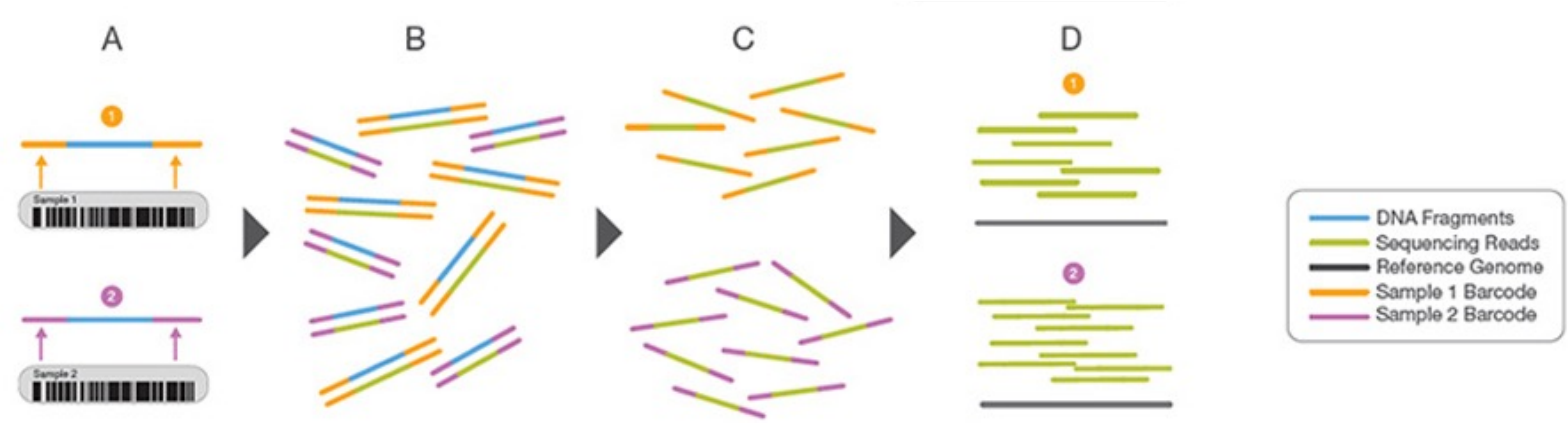
4) Sequencing array

5) Sequencing

Stepwise sequencing by polymerase using fluorescently labeled nucleotides

Patterson, J., Carpenter, E.J., Zhu, Z. et al. BMC Genomics 20, 604 (2019)

Short read sequencing



A. Two representative DNA fragments from two unique samples, each attached to a specific barcode sequence that identifies the sample from which it originated.

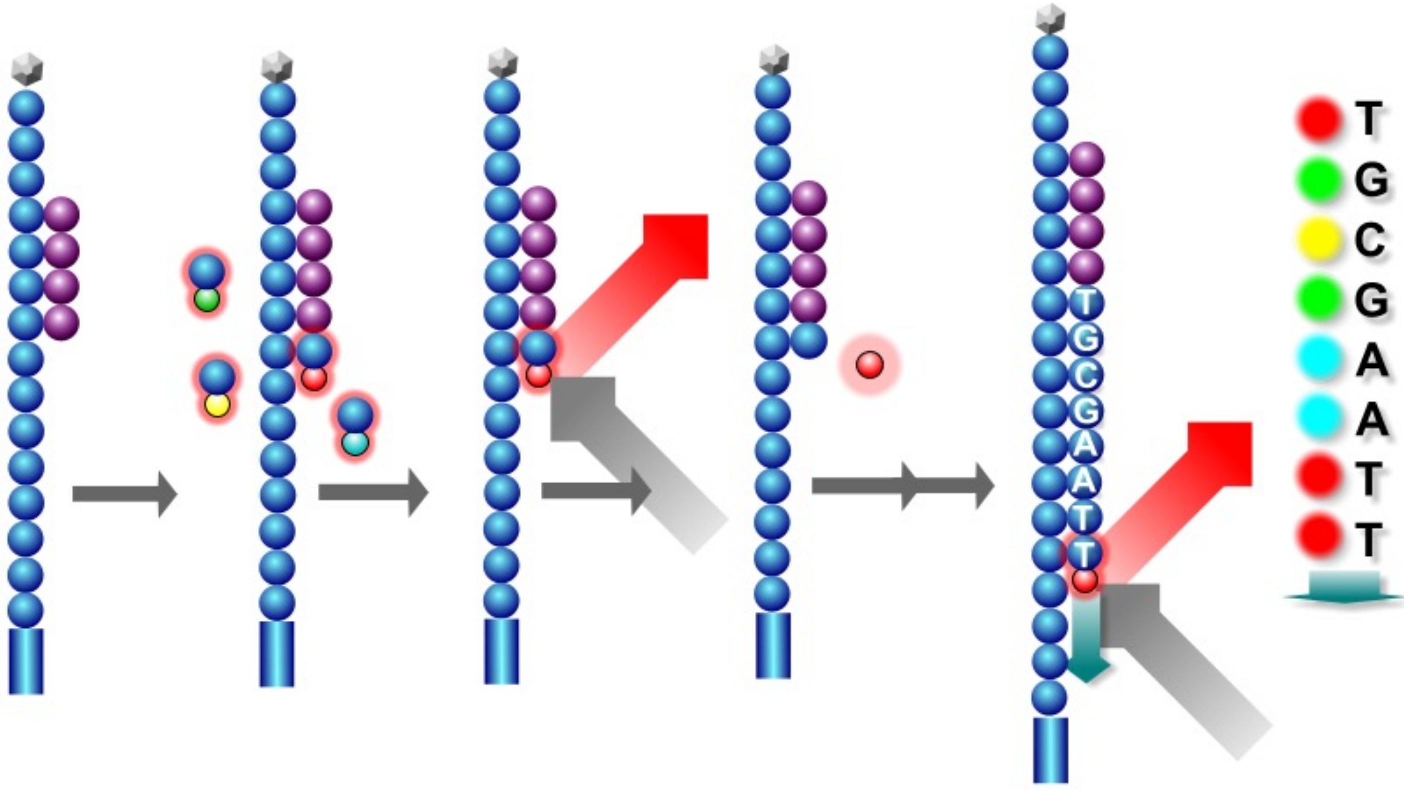
B. Libraries for each sample are pooled and sequenced in parallel. Each new read contains both the fragment sequence and its sample-identifying barcode.

C. Barcode sequences are used to de-multiplex, or differentiate reads from each sample.

D. Each set of reads is aligned to the reference sequence.

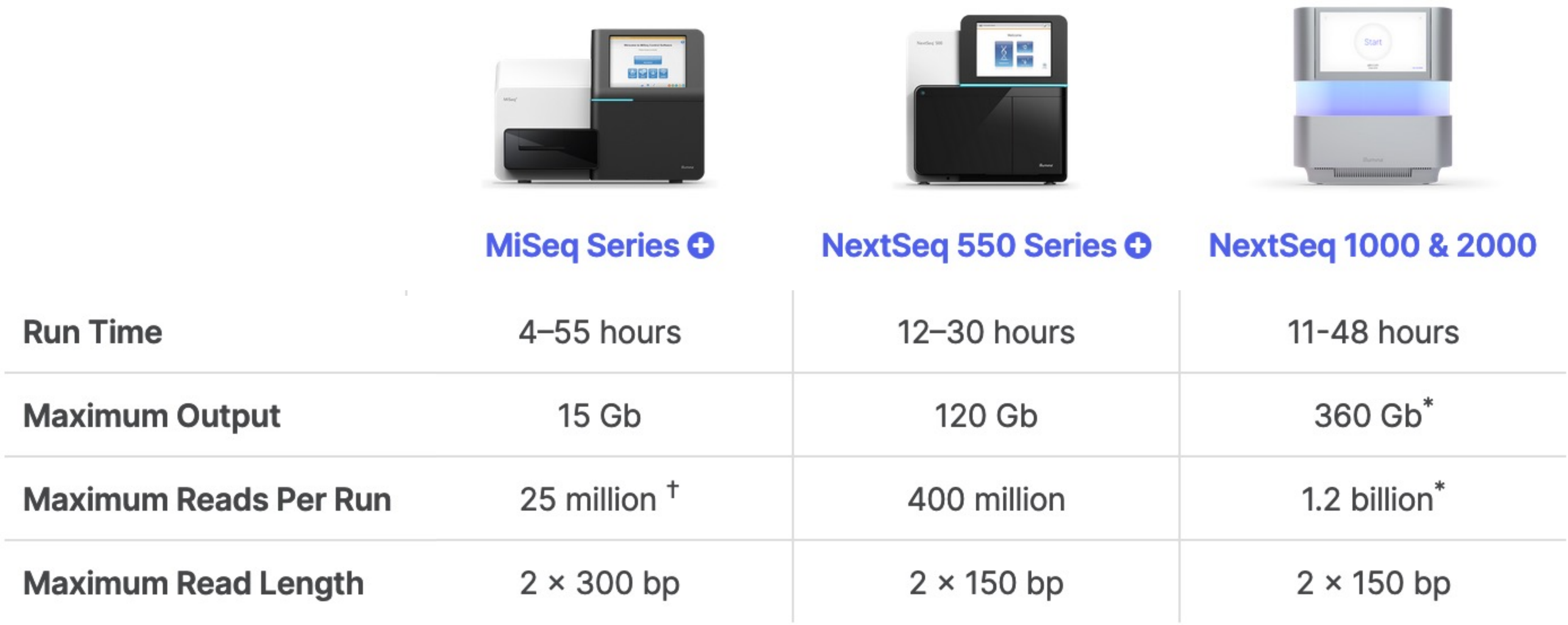
https://www.illumina.com/techniques/sequencing/ngs-library-prep/multiplexing.html

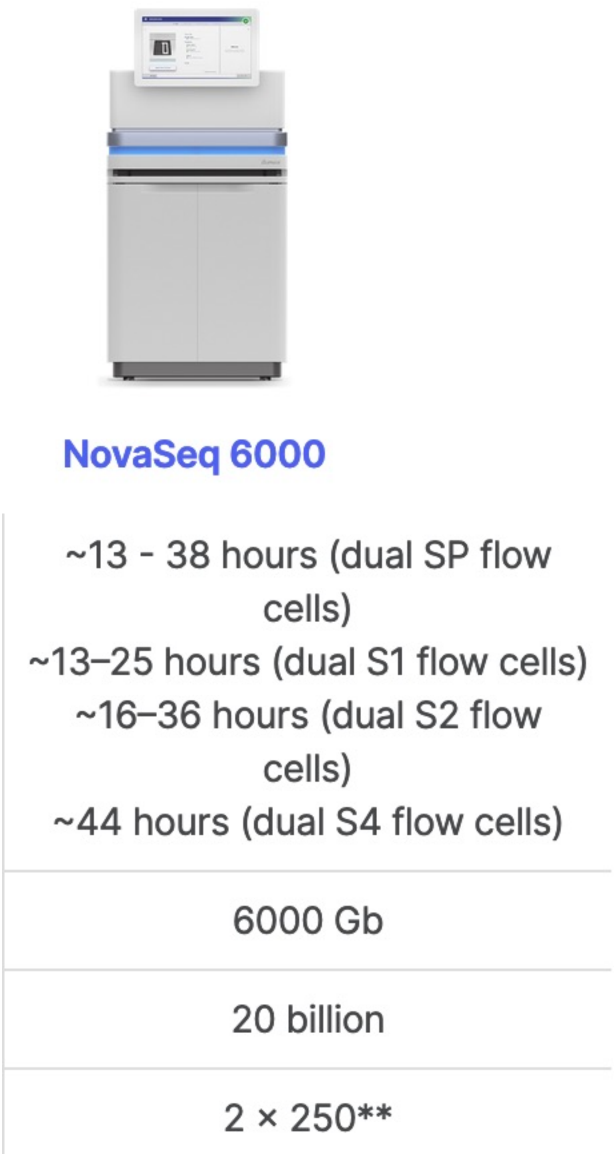
Short read sequencing



https://en.wikipedia.org/wiki/File:Sequencing_by_synthesis_Reversible_terminators.png

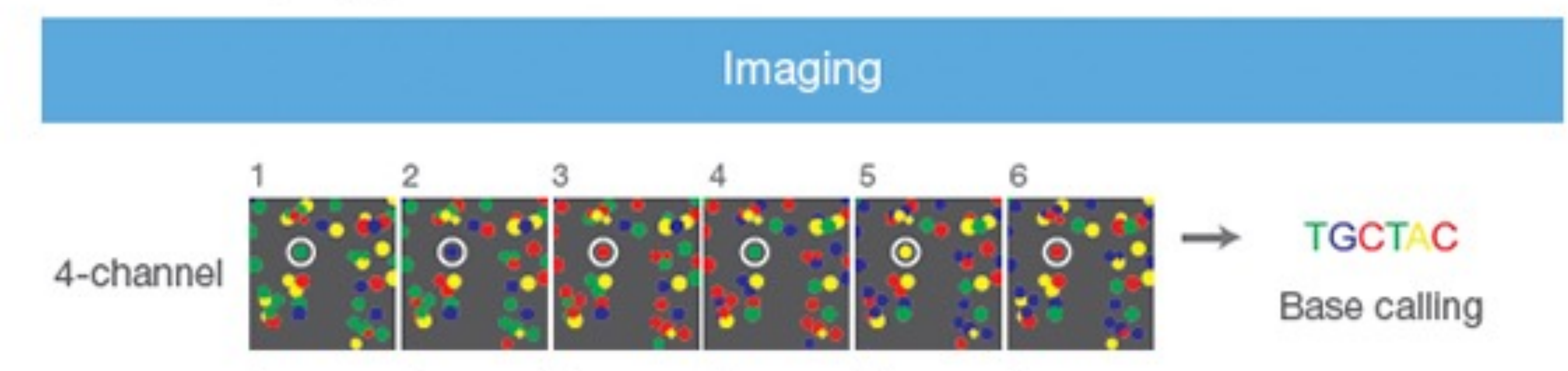
Short read sequencing





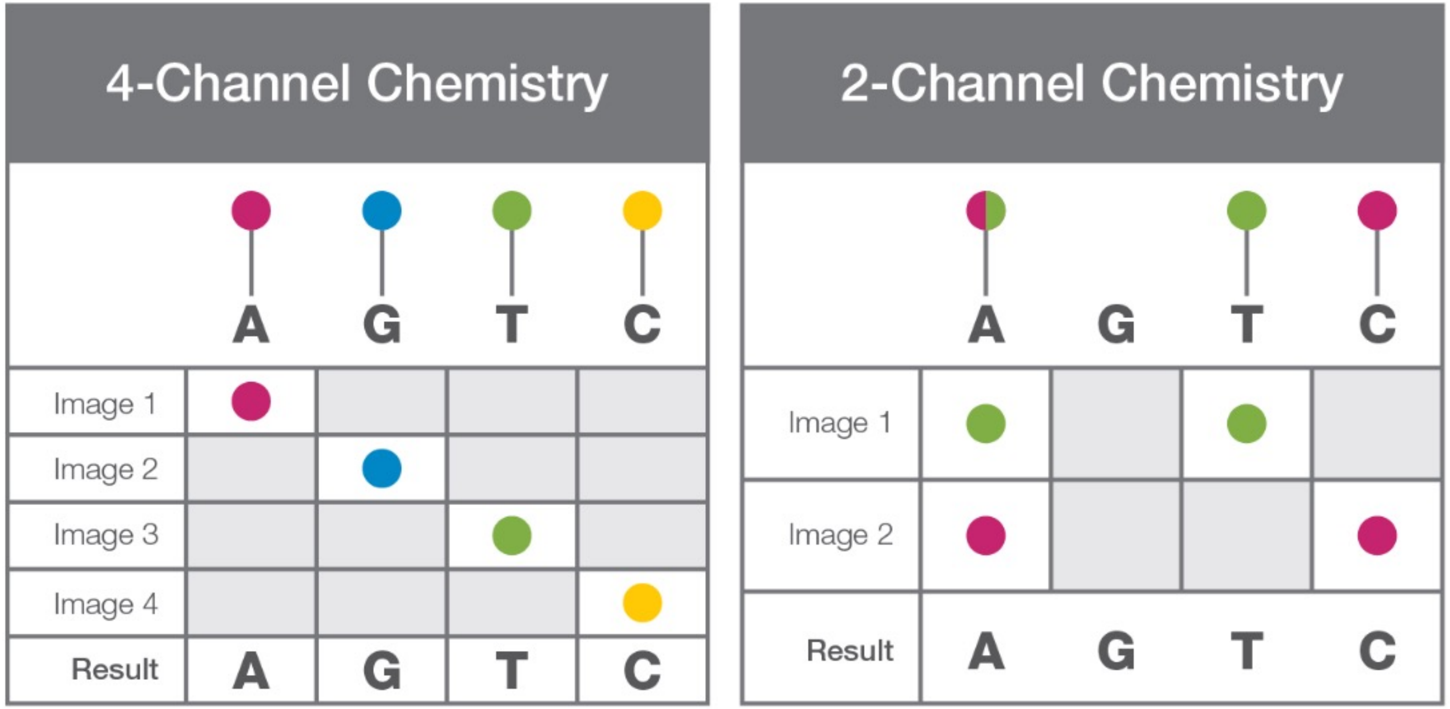
	MiSeq Series ⊕	NextSeq 550 Series ⊕	NextSeq 1000 & 2000	NovaSeq 6000
Run Time	4–55 hours	12–30 hours	11-48 hours	~13 - 38 hours (dual SP flow cells) ~13–25 hours (dual S1 flow cells) ~16–36 hours (dual S2 flow cells) ~44 hours (dual S4 flow cells)
Maximum Output	15 Gb	120 Gb	360 Gb*	6000 Gb
Maximum Reads Per Run	25 million †	400 million	1.2 billion*	20 billion
Maximum Read Length	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 × 250**

Short read sequencing



https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology/2-channel-sbs.html

Short read sequencing



4-Channel Chemistry

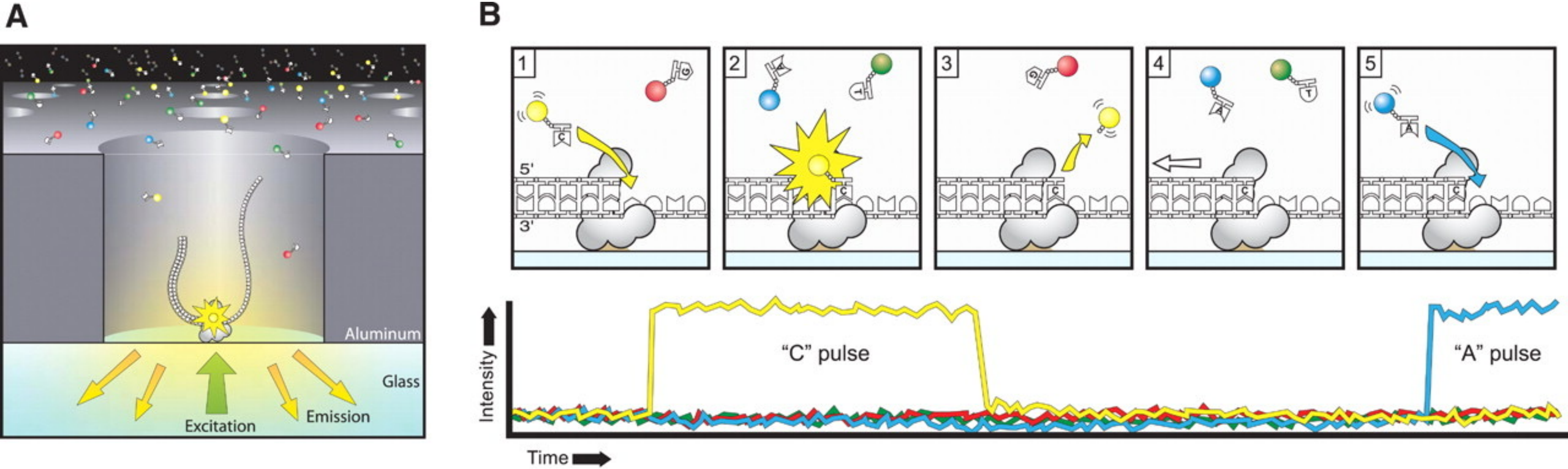
	🔴 A	🔵 G	🟢 T	🟡 C
Image 1	🔴			
Image 2		🔵		
Image 3			🟢	
Image 4				🟡
Result	A	G	T	C

2-Channel Chemistry

	🔴🟢 A	G	🟢 T	🔴 C
Image 1	🟢		🟢	
Image 2	🔴			🔴
Result	A	G	T	C

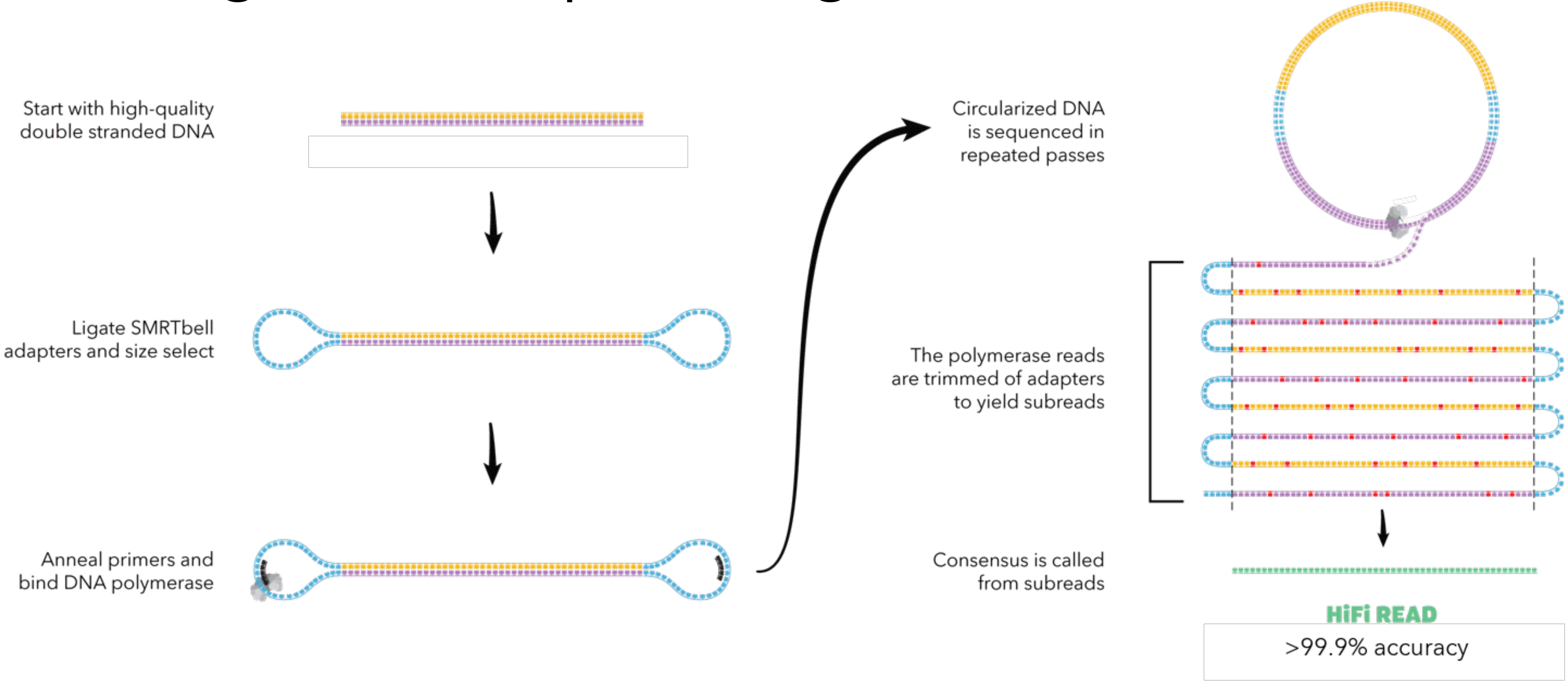
https://www.illumina.com/content/dam/illumina-marketing/documents/products/techspotlights/cmos-tech-note-770-2013-054.pdf

Long read sequencing: Pacbio SMRT



Eid, J., et al. (2009) Science, 323(5910), 133–138.

Long read sequencing: Pacbio SMRT



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

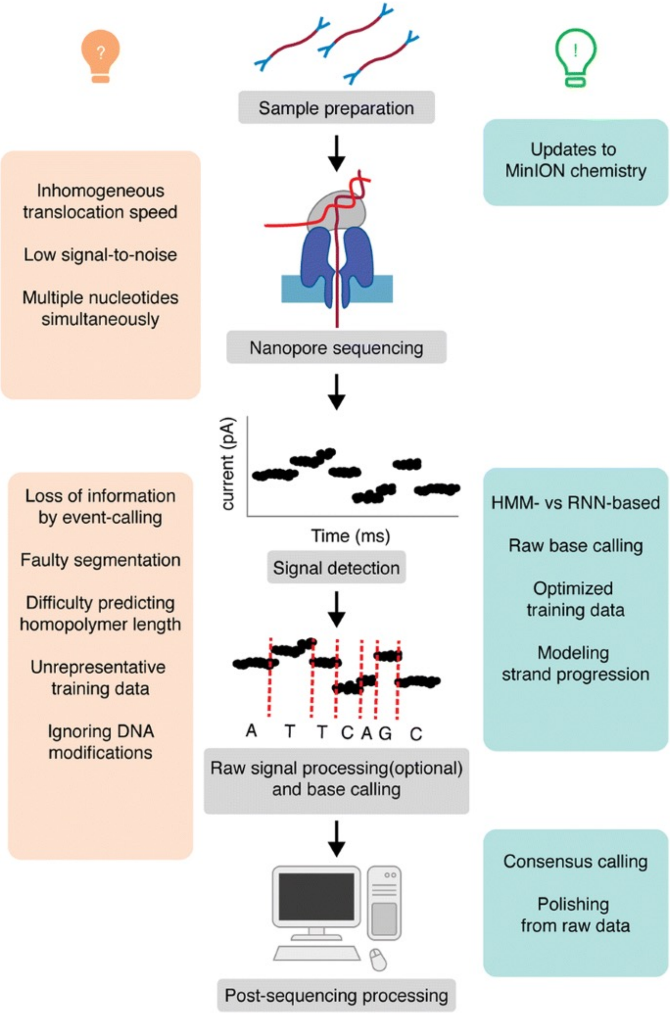
Consensus is called from subreads

HiFi READ

>99.9% accuracy

https://www.pacb.com/smrt-science/smrt-sequencing/hifi-reads-for-highly-accurate-long-read-sequencing/

Long read sequencing: ONT



Sample preparation

Nanopore sequencing

Signal detection

Raw signal processing(optional) and base calling

Post-sequencing processing

Inhomogeneous translocation speed

Low signal-to-noise

Multiple nucleotides simultaneously

Loss of information by event-calling

Faulty segmentation

Difficulty predicting homopolymer length

Unrepresentative training data

Ignoring DNA modifications

Updates to MinION chemistry

HMM- vs RNN-based

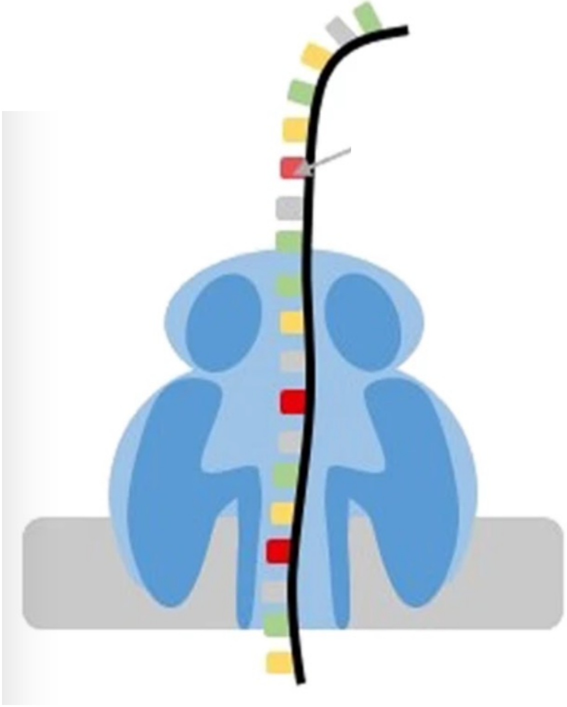
Raw base calling

Optimized training data

Modeling strand progression

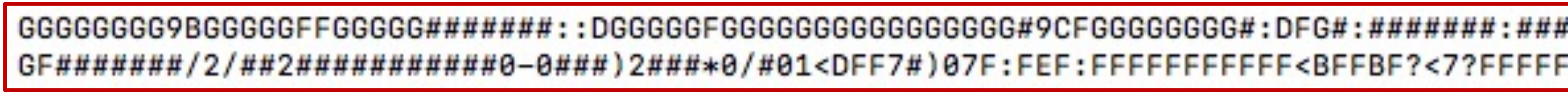
Consensus calling

Polishing from raw data



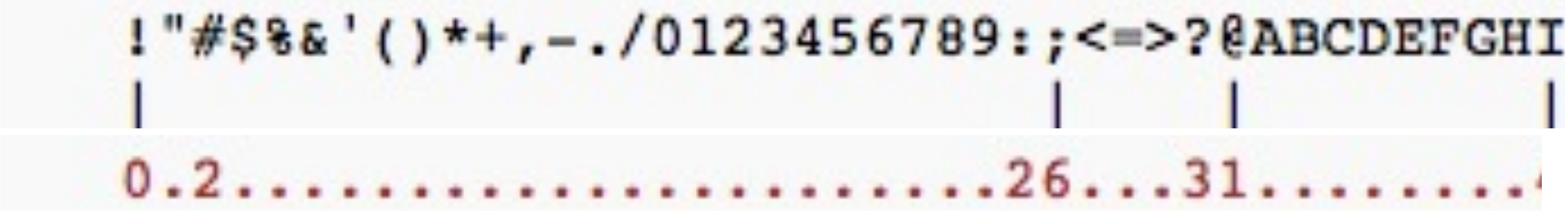
Rang, F.J., Kloosterman, W.P. & de Ridder, J. Genome Biol 19, 90 (2018)
Xu, L., Seki, M. J Hum Genet 65, 25–33 (2020)

Quality?



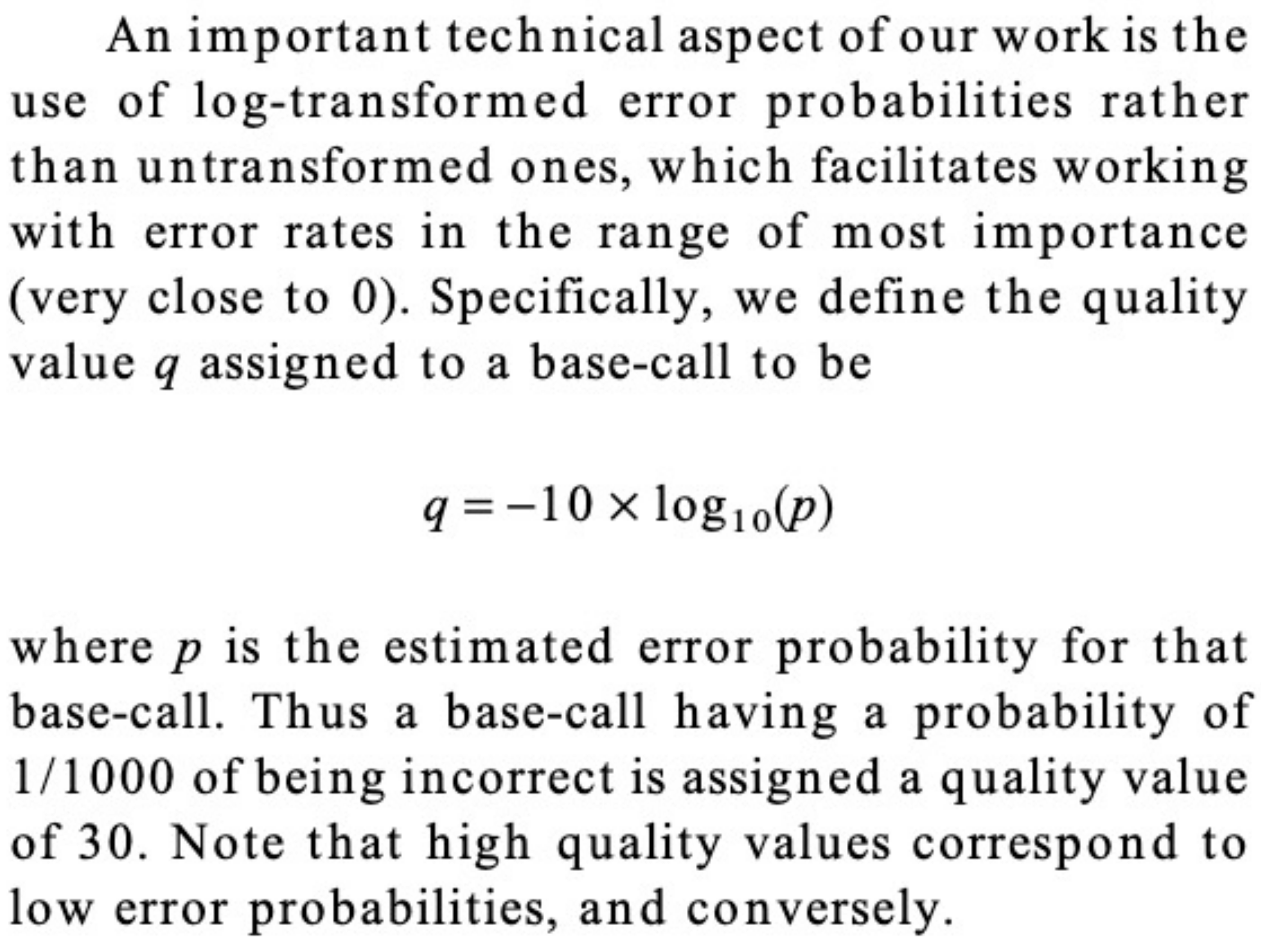
```
GGGGGGGG9BGGGGGFFGGGGG#######::DGGGGGFGGGGGGGGGGGGGGGG#9CFGGGGGGGG#:DFG#:#######:###
GF#######/2/##2###########0-0###)2###*0/#01<DFF7#)07F:FEF:FFFFFFFFFFF<BFFBF?<7?FFFFF
```

Phred+33 score:



```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
|                         |    |        |
0.2......................26...31........
```

Quality scoring



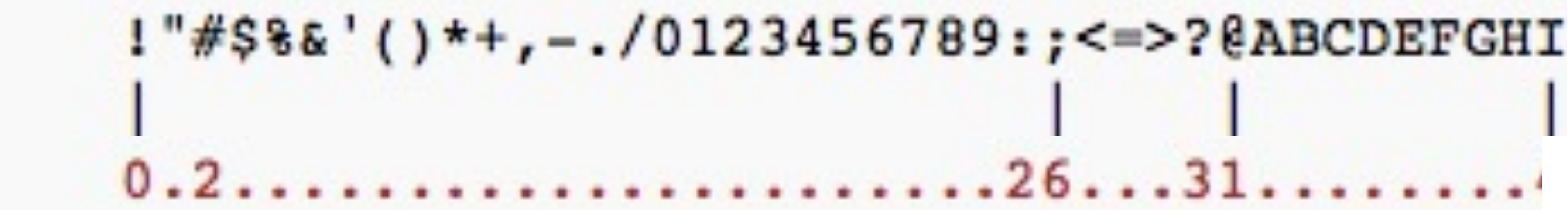
An important technical aspect of our work is the use of log-transformed error probabilities rather than untransformed ones, which facilitates working with error rates in the range of most importance (very close to 0). Specifically, we define the quality value q assigned to a base-call to be

$$q = -10 \times \log_{10}(p)$$

where p is the estimated error probability for that base-call. Thus a base-call having a probability of 1/1000 of being incorrect is assigned a quality value of 30. Note that high quality values correspond to low error probabilities, and conversely.

Ewing B, Green P. Genome Research. 1998 8(3):186-94

Quality scoring

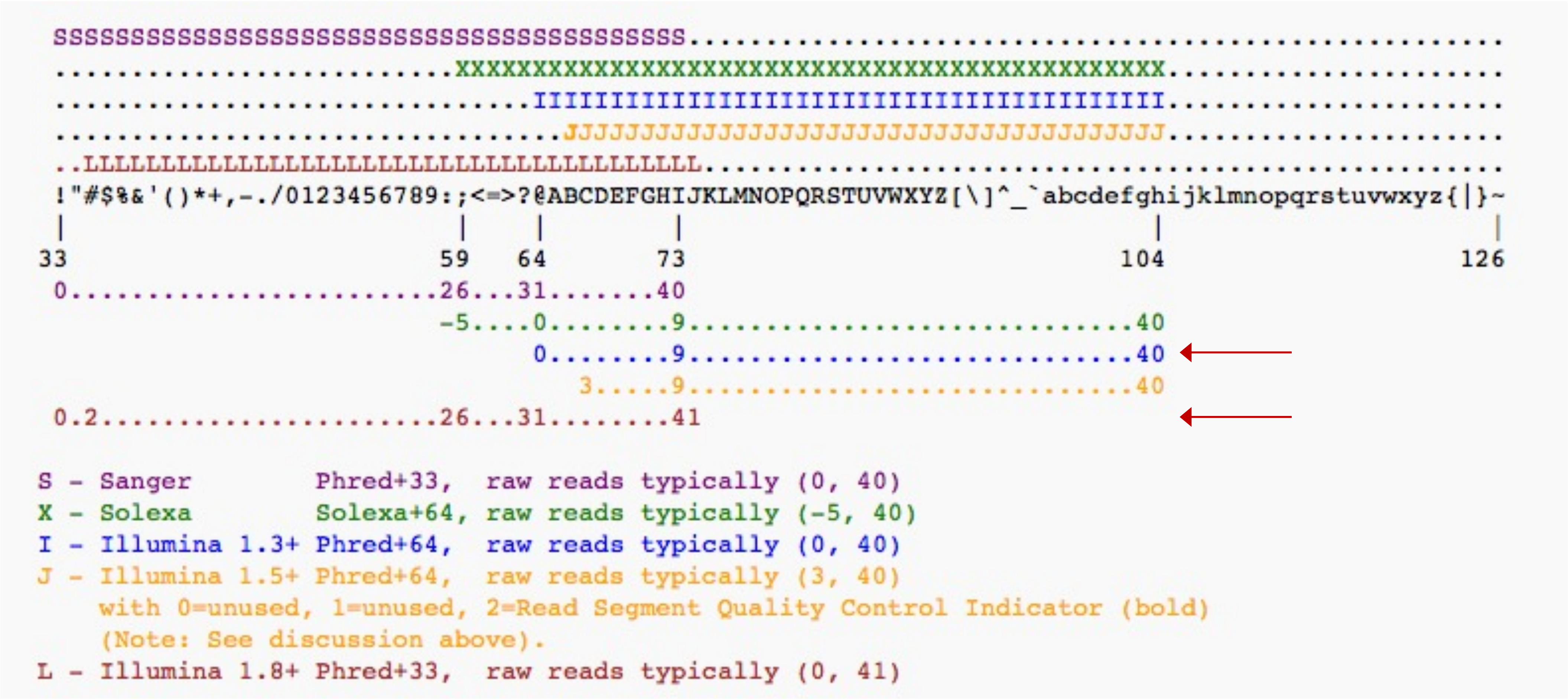


```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
|                         |    |        |
0.2......................26...31........
```

```
error        1  0.63           0.1                    0.01    0.0025    0.00079                0.0001
probability
```

$$q = -10 \times \log_{10}(p)$$

Quality scoring



```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
 ..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0........................26...31.......40
                          -5....0........9.............................40
                                0........9.............................40
                                   3.....9.............................40
 0.2......................26...31........41

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)
```

Quality scoring

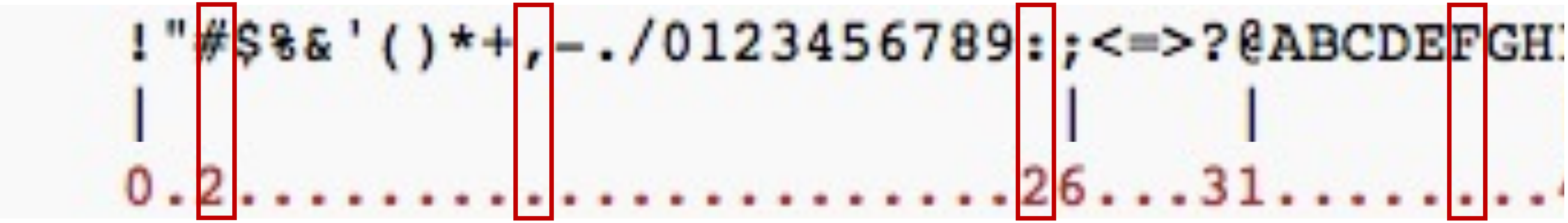
novaseq data:

```
@SRR15010442.1
CCTGTTTGCTCCCCACGCTTTCGCGCCTCAGCGGCAGTTACAGACCAAAAAGCCGCCTTCGCCACTGGTGTTC
CTCCACATCTCTACGCATTTCACCGCTACACGTGGAATTCTACCCCCC
+
F:FFFF,FF:FFFF:FFFFFFFFFFFFFF:FFF,FFFFF,FFFFFFFF:FFFFFFFFFFFFFF:FFFFFF:FF
:FFFFFFFFFF,FF:F:FFFFFFFF:F:F:FFFFFF,F,FF,FFFFFF
```

Quality scoring

novaseq data:

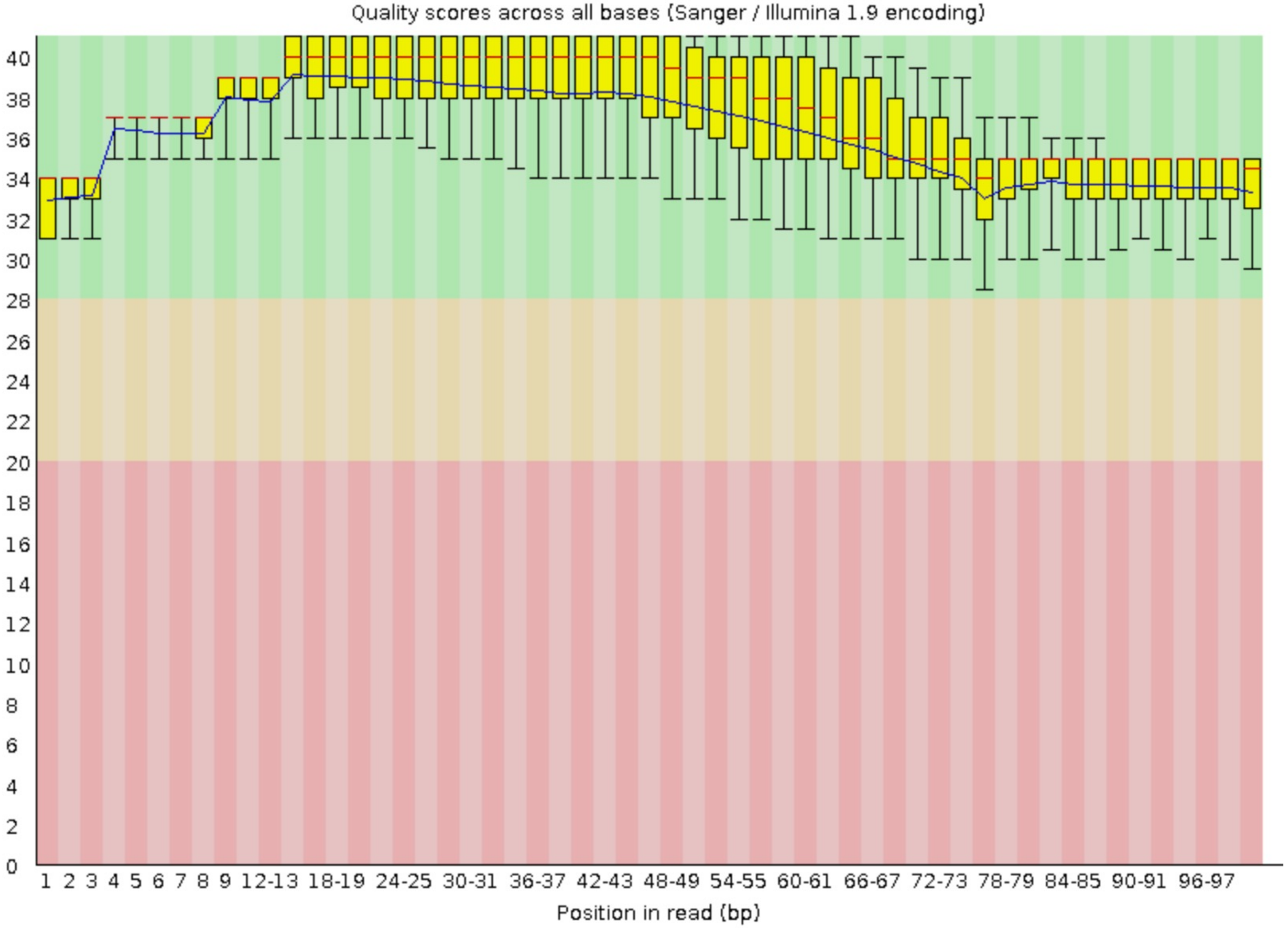
```
@SRR15010442.1
CCTGTTTGCTCCCCACGCTTTCGCGCCTCAGCGGCAGTTACAGACCAAAAAGCCGCCTTCGCCACTGGTGTTC
CTCCACATCTCTACGCATTTCACCGCTACACGTGGAATTCTACCCCCC
+
F:FFFF,FF:FFFF:FFFFFFFFFFFFFF:FFF,FFFFF,FFFFFFFF:FFFFFFFFFFFFFF:FFFFFF:FF
:FFFFFFFFFF,FF:F:FFFFFFFF:F:F:FFFFFF,F,FF,FFFFFF
```



error probability 0.63 0.079 0.0032 0.0002

Quality reports

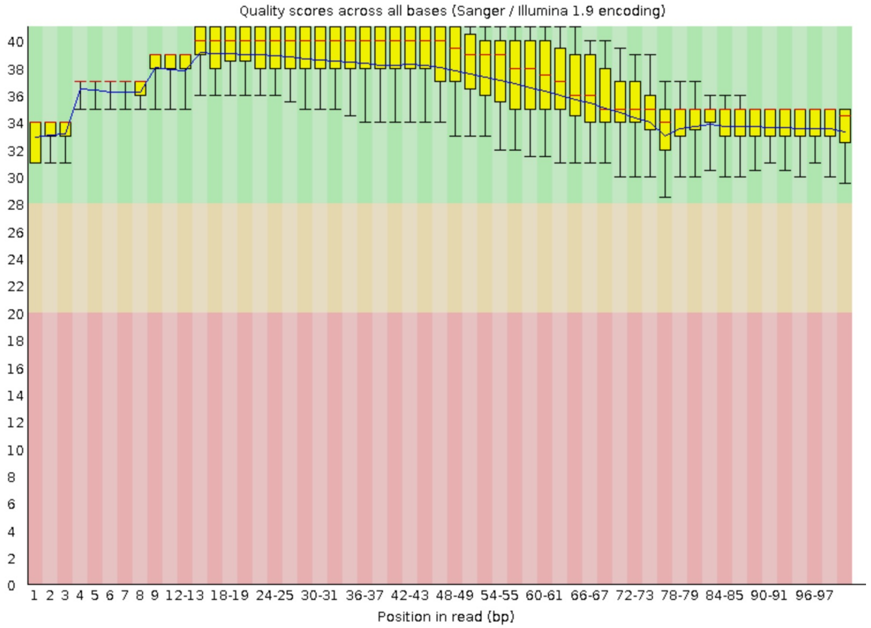
summary of 1 dataset:



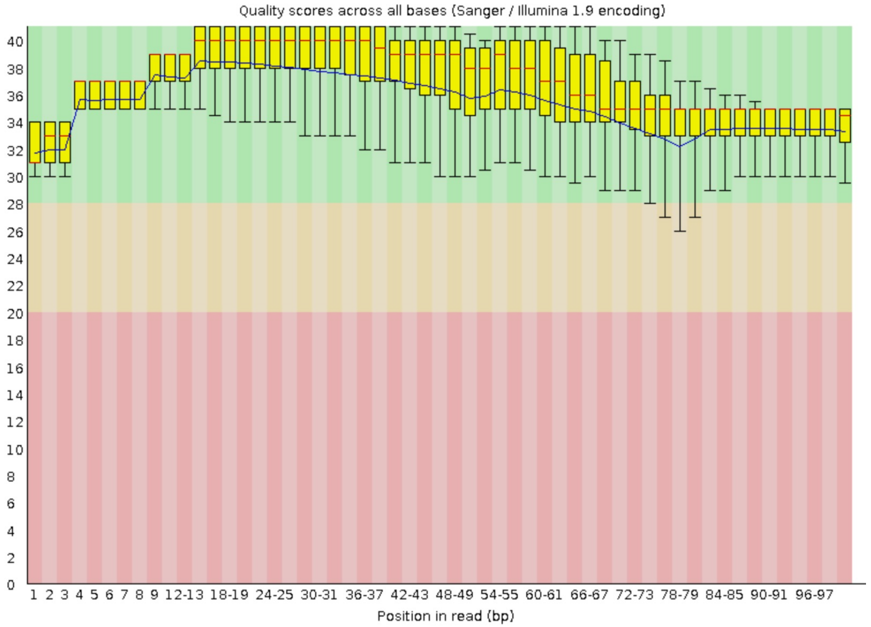
Quality reports

summary of 1 dataset:

forward reads:

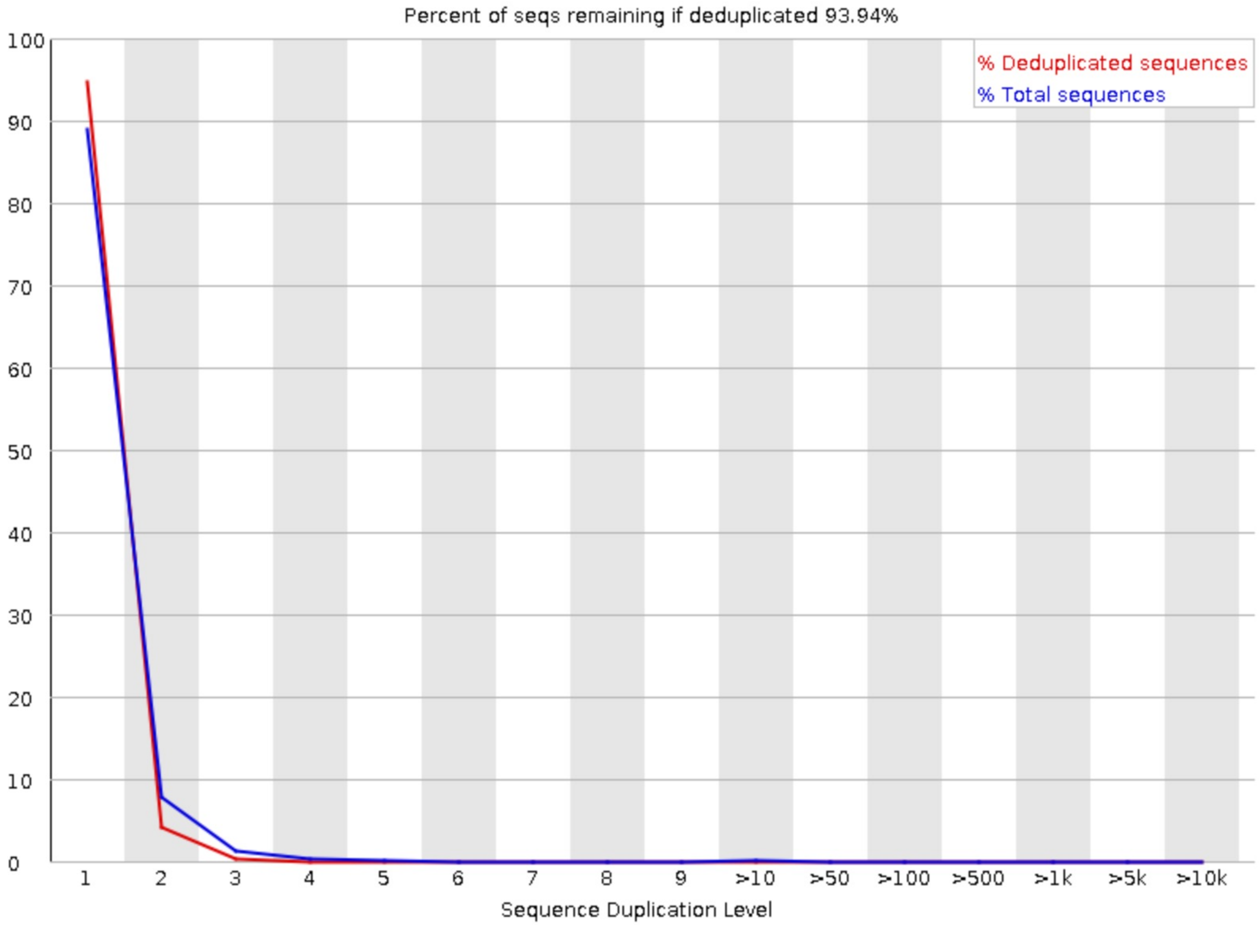


reverse reads:



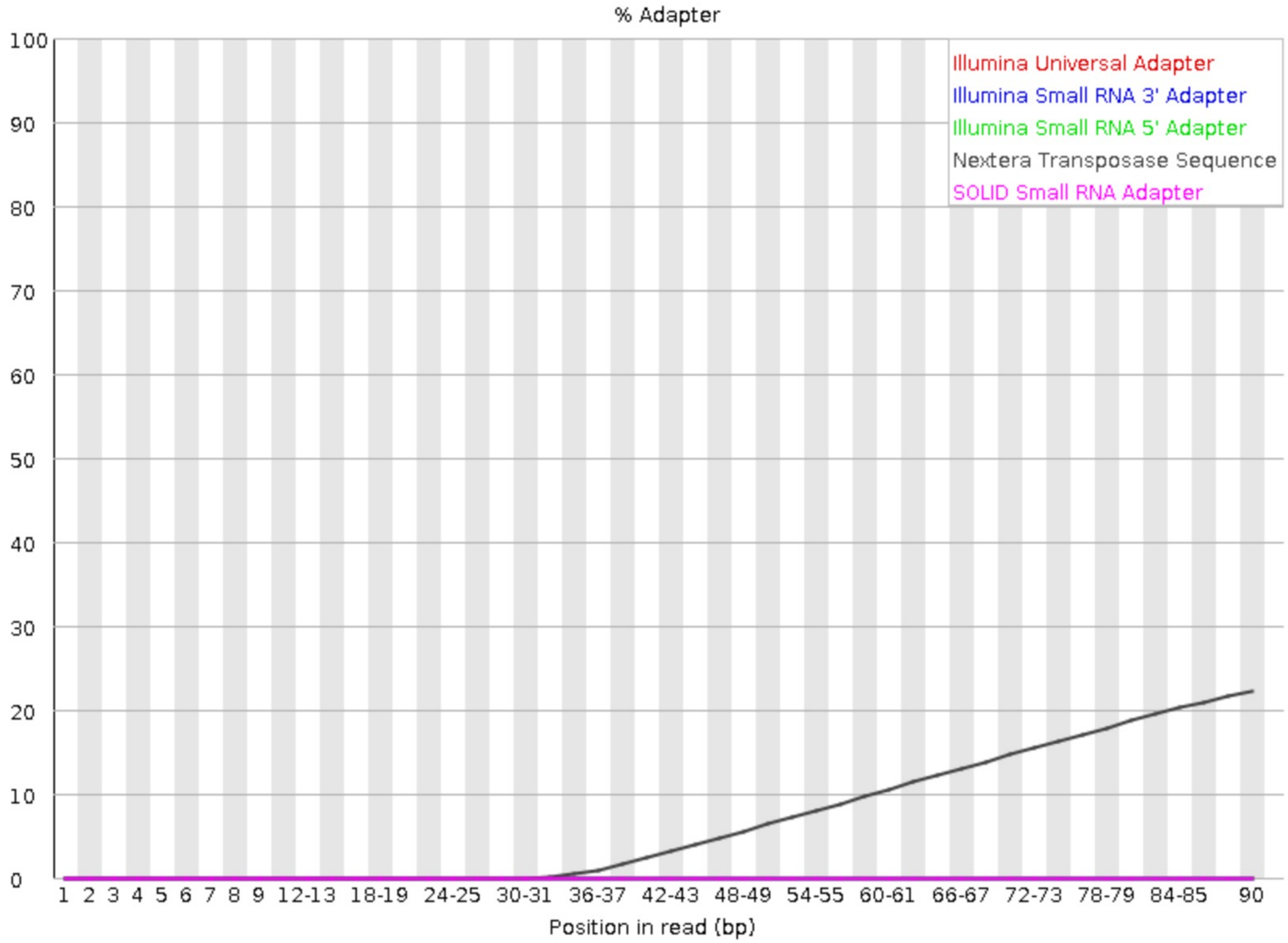
Quality reports

summary of 1 dataset:



Quality reports

summary of 1 dataset:



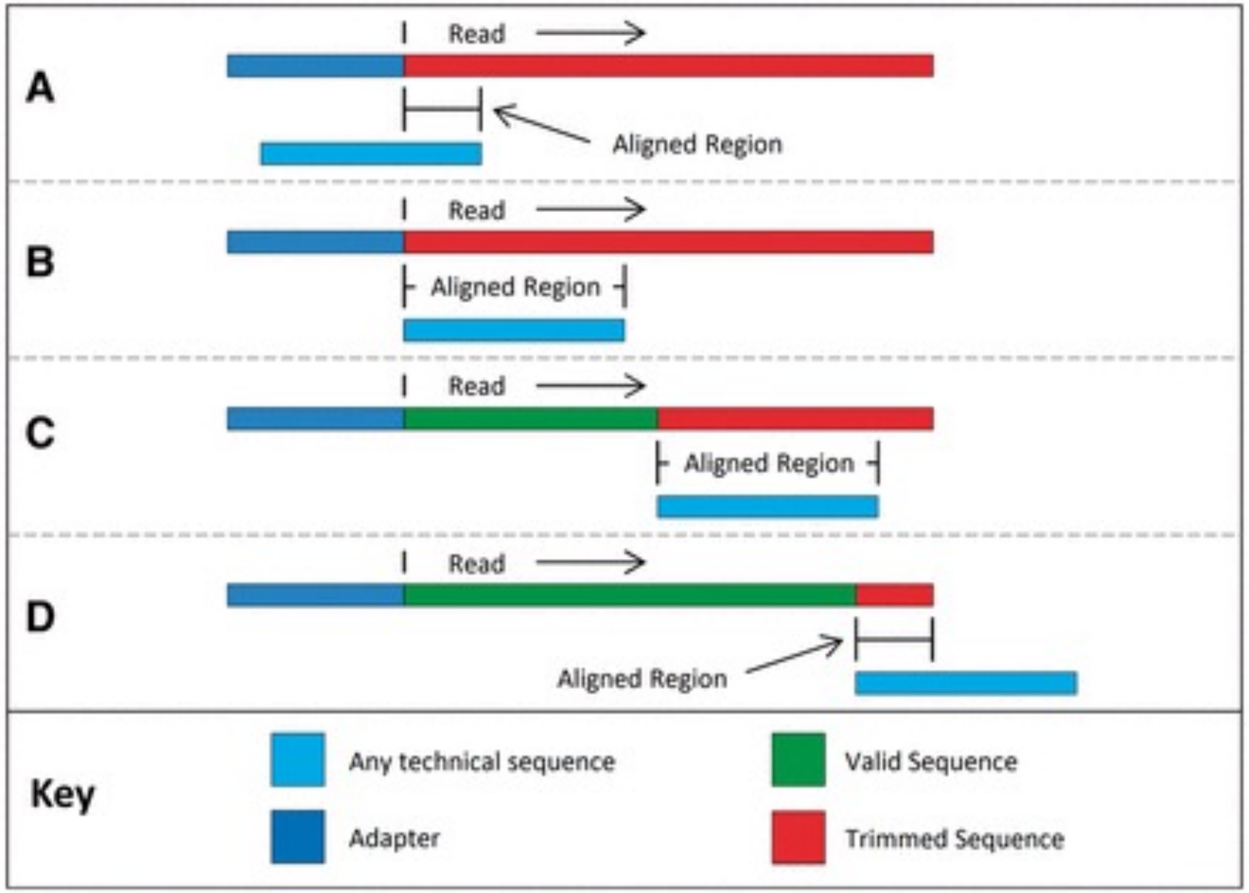
Data preprocessing: filtering & trimming

- remove adapter sequences

- remove low-quality ends

- remove dark-cycle poly-G ends

Data preprocessing : filtering & trimming

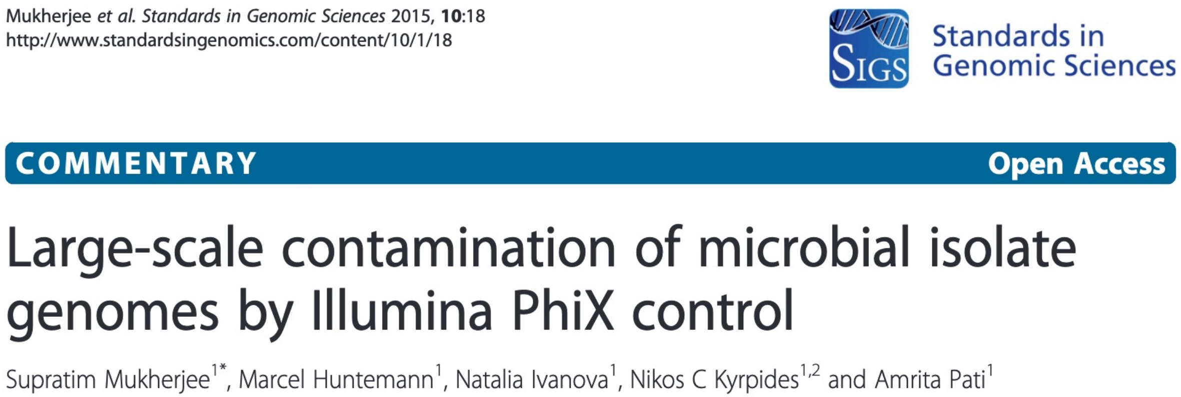


Anthony M. Bolger, Marc Lohse, Bjoern Usadel, Bioinformatics, Volume 30, Issue 15, 2014, Pages 2114–2120

Data preprocessing – remove contaminants!

- remove uninformative sequences:
- phiX spike-in
- host genome
- for rRNA-depleted RNAseq: remove rRNA

Data preprocessing – remove contaminants!



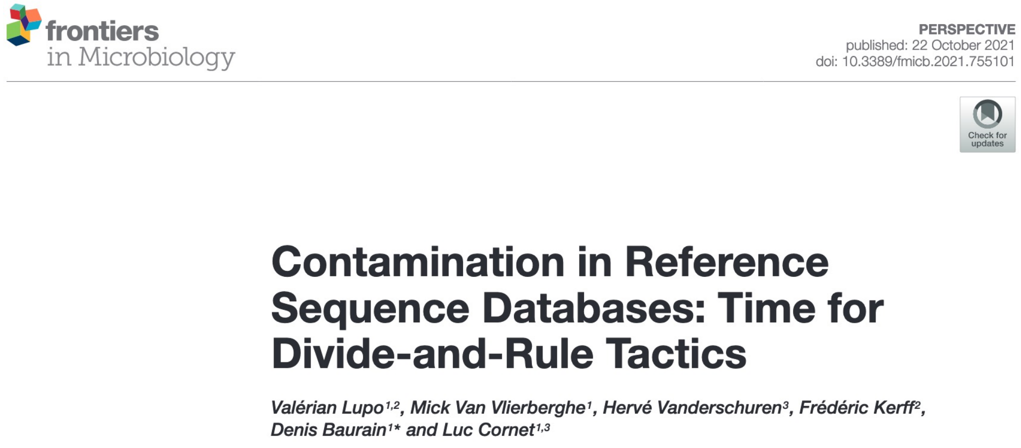
Mukherjee *et al. Standards in Genomic Sciences* 2015, **10**:18
http://www.standardsingenomics.com/content/10/1/18

Standards in Genomic Sciences

COMMENTARY **Open Access**

Large-scale contamination of microbial isolate genomes by Illumina PhiX control

Supratim Mukherjee[1*], Marcel Huntemann[1], Natalia Ivanova[1], Nikos C Kyrpides[1,2] and Amrita Pati[1]



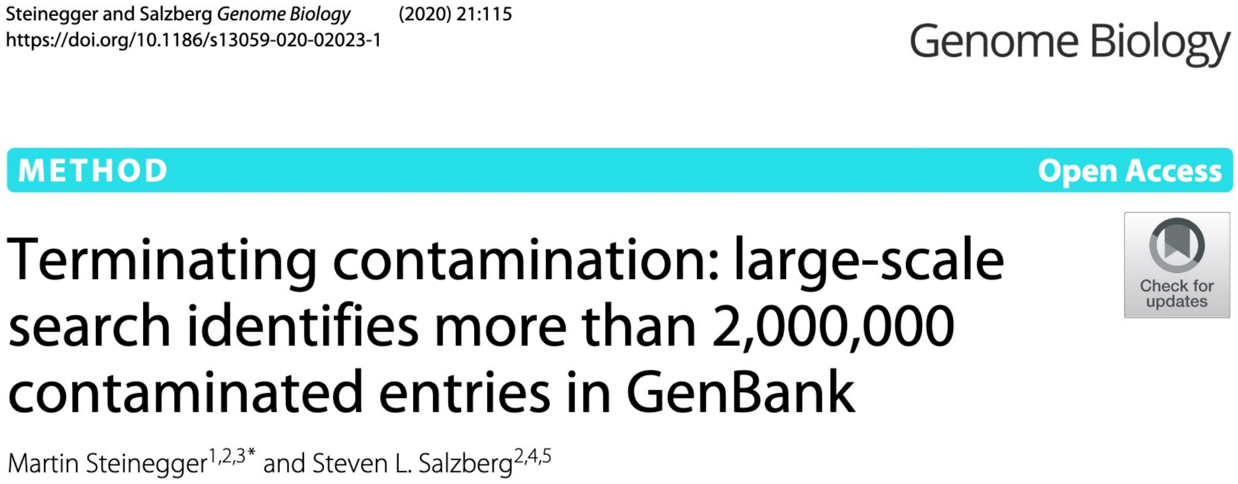
frontiers in Microbiology

PERSPECTIVE
published: 22 October 2021
doi: 10.3389/fmicb.2021.755101

Check for updates

Contamination in Reference Sequence Databases: Time for Divide-and-Rule Tactics

Valérian Lupo[1,2], Mick Van Vlierberghe[1], Hervé Vanderschuren[3], Frédéric Kerff[2], Denis Baurain[1] and Luc Cornet[1,3]*



Steinegger and Salzberg *Genome Biology* (2020) 21:115
https://doi.org/10.1186/s13059-020-02023-1

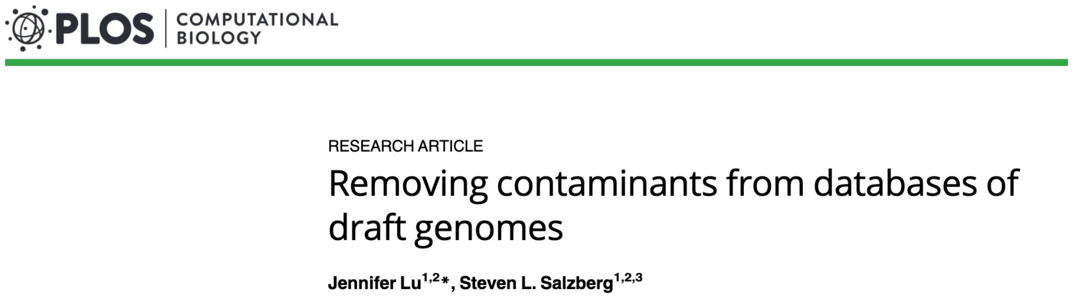
Genome Biology

METHOD **Open Access**

Check for updates

Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank

Martin Steinegger[1,2,3*] and Steven L. Salzberg[2,4,5]



PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Removing contaminants from databases of draft genomes

Jennifer Lu[1,2*], Steven L. Salzberg[1,2,3]

Thanks for your attention!

a.u.s.heintzbuschart@uva.nl

SP C2.205

github.com/a-h-b

twitter.com/_a_h_b_