Sequencing technologies for metagenomics

Espen Mikal Robertsen
(Espen Åberg)

Applied Metagenomics, AM21

2021- Portugal

www.elixir-europe.org

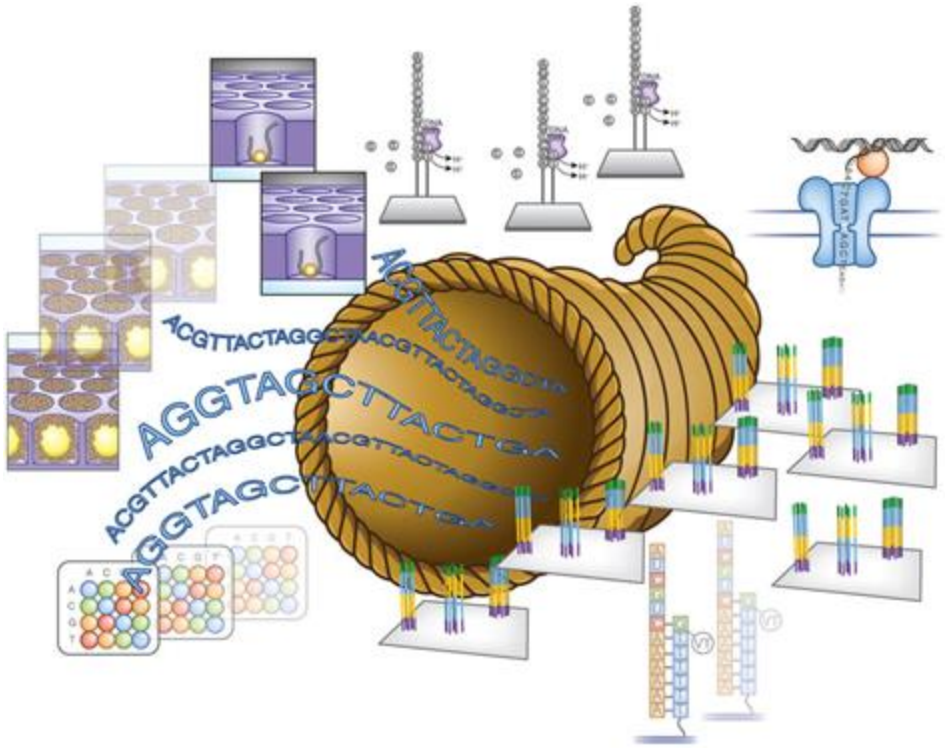
Early metagenomic sequencing

- Pioneering metagenomic studies used the Sanger platform
  - i.e Venter, J.C. et al. Environmental genome shotgun sequencing of the Sargasso Sea. Science 304, 66–74 (2004).
    - 1800 genomic species , 148 novel bacterial phylotypes
    - High-quality DNA sequence
    - Relatively long (500-1000 bp)

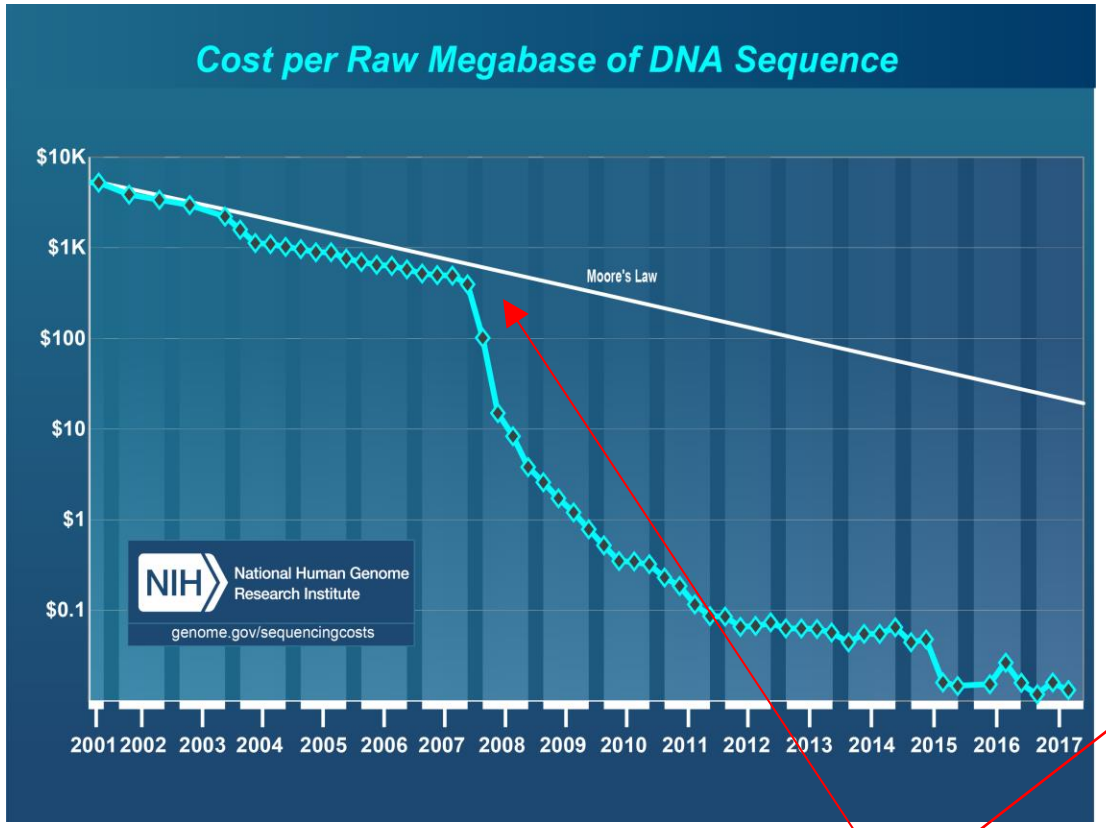
- This technology can not provide sufficient read depth to saturate moderately diverse communities
  - Sanger-based metagenomic projects are often limited to:
    - Fosmid or bacterial artificial chromosome libraries
    - low-diversity microbial communities.

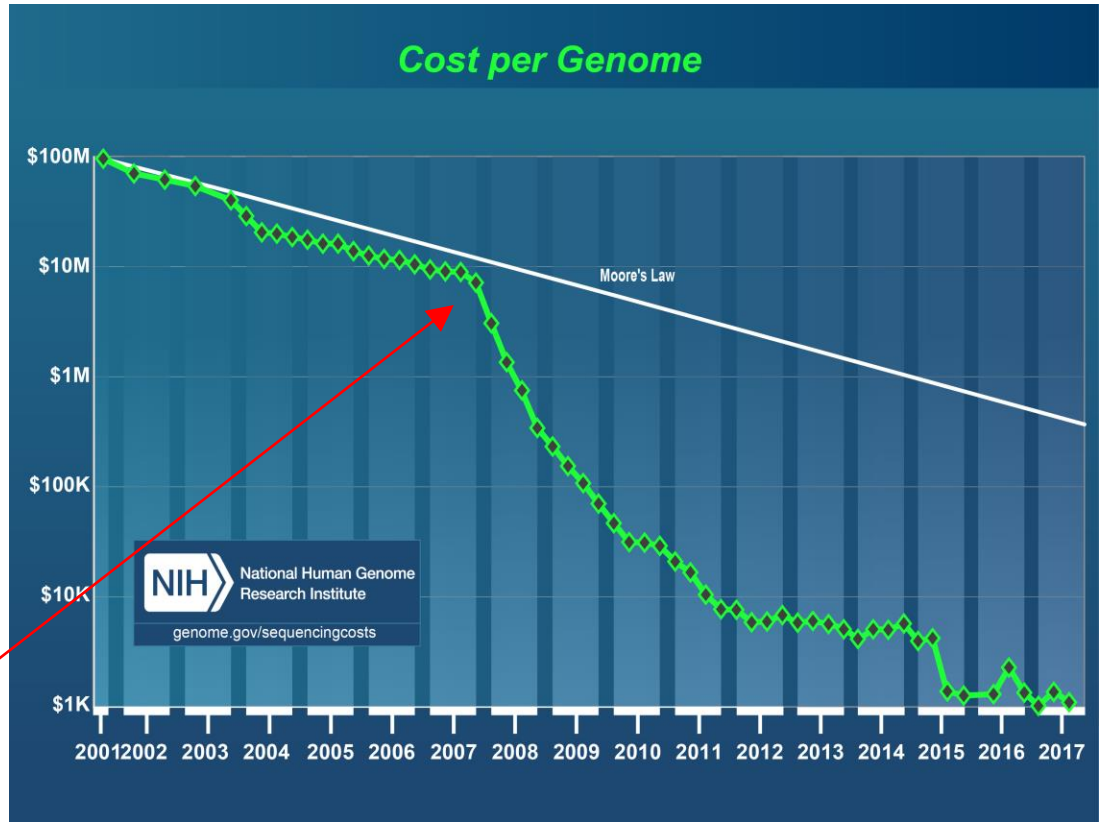
Next-generation sequencing (NGS)

- Overcomes several of the disadvantages of Sanger sequencing
  1. Substantially higher throughput
  2. Cheaper cost per base sequencing
  3. Simpler library preparation
  4. No cloning step
  5. Real time



DNA Sequencing Costs over time





NGS enters the scene here

the cost of sequencing a human-sized genome

Technology improvements that 'keep up' with Moore's Law are widely regarded to be doing exceedingly well

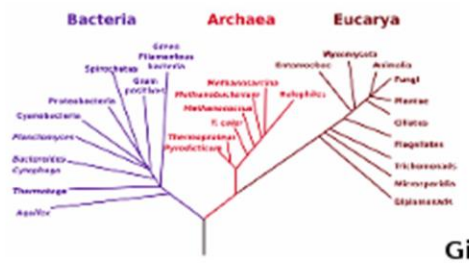
Sequence data analysis is changing rapidly

- Few methods are completely static
- Software is still under active development
- New methods and tools are reported every month
- Staying on the learning curve is essential

Metagenomics: a dominant contributor to sequence databases

Robert Koch isolates microorganisms using solid cultures

1888



Carl Woese propose rRNA as marker for taxonomy

1977

Giovannoni et al., perform the first microbial community study by 16S rRNA libraries

1990

1676

Leeuwenhoek reports his observations about oral microbiota

1931

Winogradsky microbial ecology experiments

Modified from https://www.hhmi.org/biointeractive/

1980

Kary Mullis develops PCR

1998

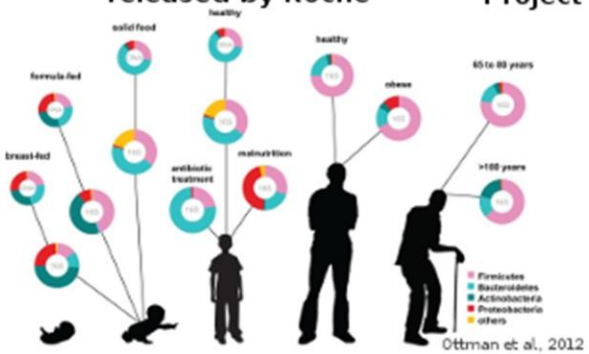
Handelsman et al., propose the term 'metagenomics'

2005

First NGS machine released by Roche

2006

GA sequencer from Solexa is released



2008

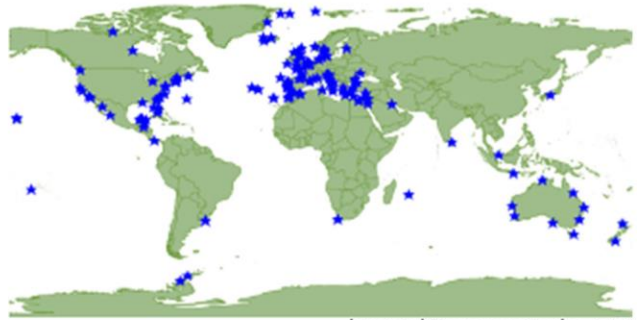
Human Microbiome Project publication

2011

PacBio RS sequencer is released

2015

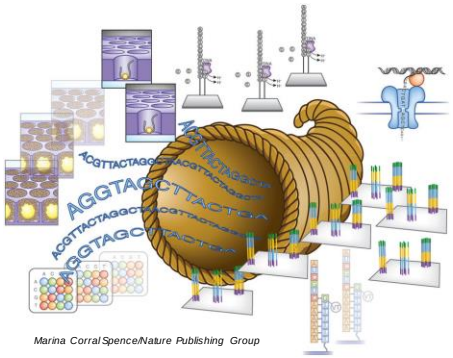
Ocean Samplig Day



Next-generation sequencing (NGS)

- Not without new challenges…
  - Each new technology has a different error model and biases that need to be considered during experimental design and sequence analysis
  - Errors that occur in the output sequence on NGS
    - Indels (insertion/deletion) = bases inserted (In) or absent (del)
    - Base substitutions
  - Increased coverage can overcome errors but absolute number of sequencing errors will increase with coverage

NGS?



- Next Generation Sequencing
  - **Second** Generation Sequencing
    - Sequencing By Synthesis
      - Pyrosequencing
        - 454
      - Sequencing by reversible terminator
        - Illumina
      - Sequencing by Detection of Hydrogen Ion
        - Ion Torrent
    - Sequencing By Ligation
      - Sequencing by hybridisation and ligation
        - SOLiD
  - **Third** Generation Sequencing
    - Single Molecule real-time long reads
      - Pacific Biosciences Single molecule real time (SMRT) sequencing
      - Oxford Nanopore
    - Synthetic long reads
      - Illumina Synthetic Long-Read
      - 10X Genomics

Illumina

- Market leader
  - Latest addition Novaseq 6000
    - Output: 80 - 6000 Gb
    - Paired end reads: 1.6 - 40B
  - 100$ genomes?
  - iSeq 100 (benchtop sequencer)

- Long-read sequencing market?

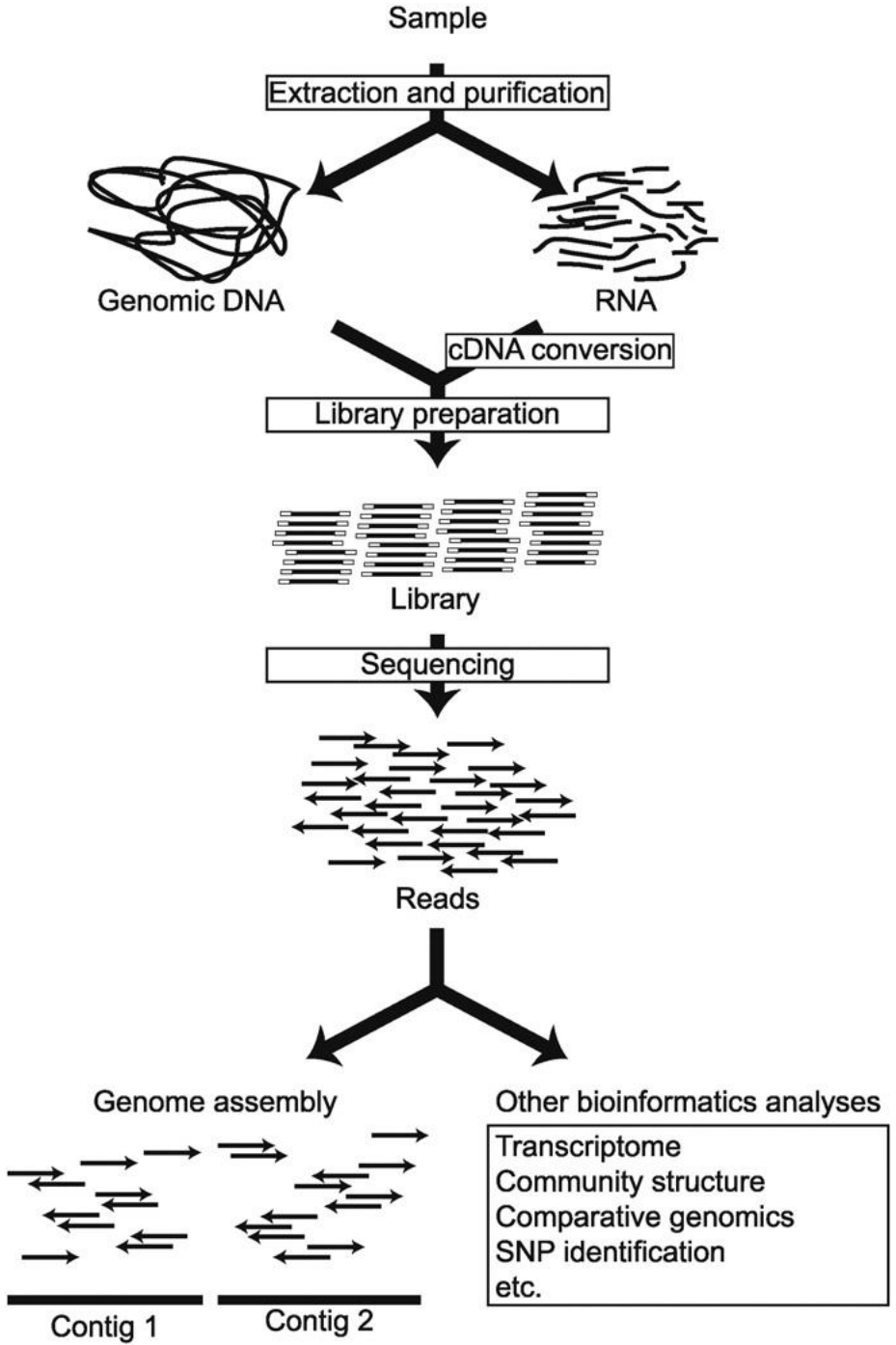
Bejing Genomic Institute (BGI)

- Biggest sequencing centre on earth.
- Short-read sequencing platform, the **BGISEQ-500,MGI-200, MGI-2000**
- An initial study suggests it may produce data of a comparable quality to Illumina (Mak *et al.* 2017).

General NGS Principle

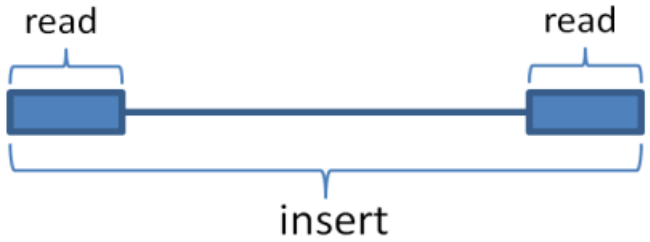
- Sequence a large number of DNA fragments (thousands to millions) in parallel in a single machine run

- Possible downstream analyses depends on:
  - Choice of the sequencing instrument and associated technology
  - The way libraries are prepared



Vincent AT, et al; J Microbiol Methods; vol138:p60-71 (2017)

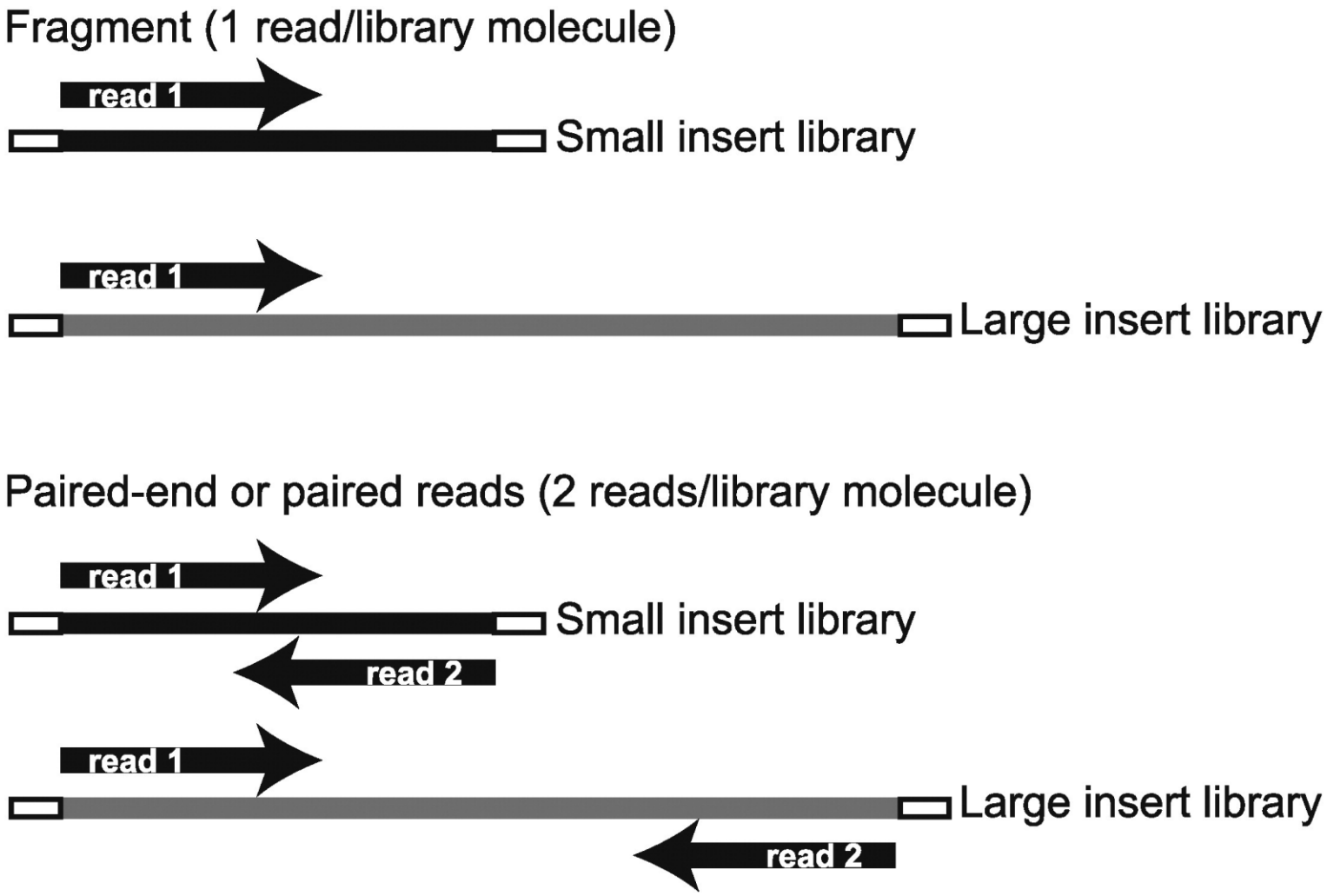
Basic concepts



Insert: The DNA fragment that is used for sequencing.

Read: The part of the insert that is sequenced.

Single-end or Paired-end reads…

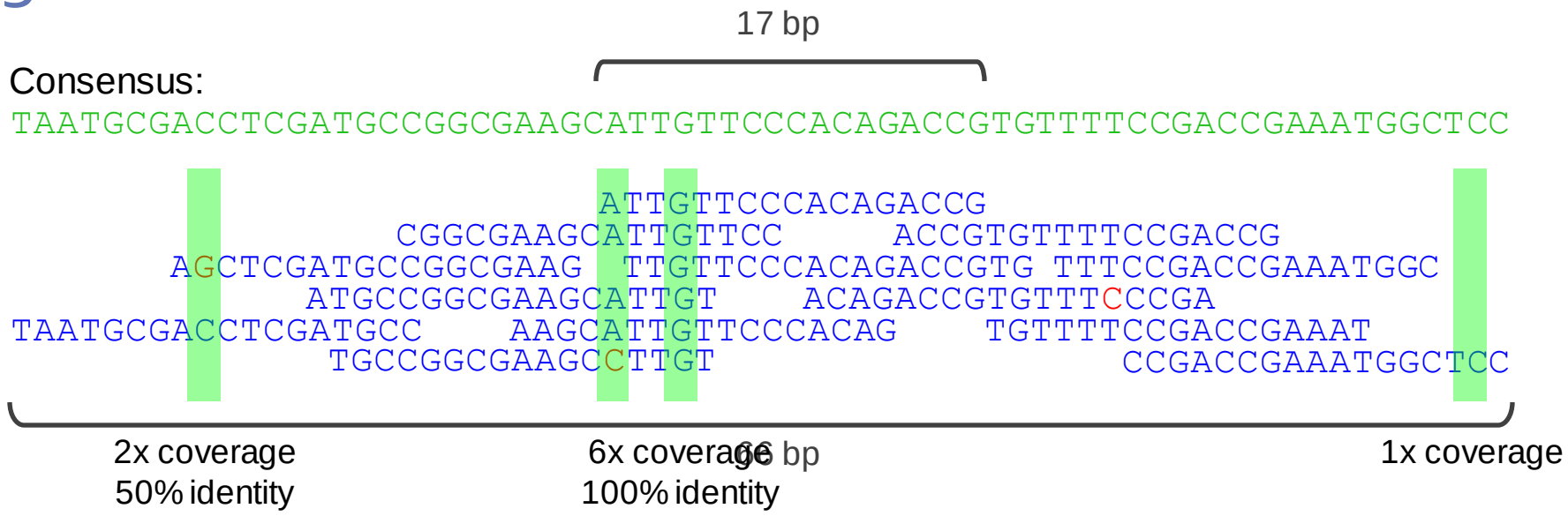


…Determines:

- Sequencer choice
- How the libraries are produced

Vincent AT, et al; J Microbiol Methods; vol138:p60-71 (2017)

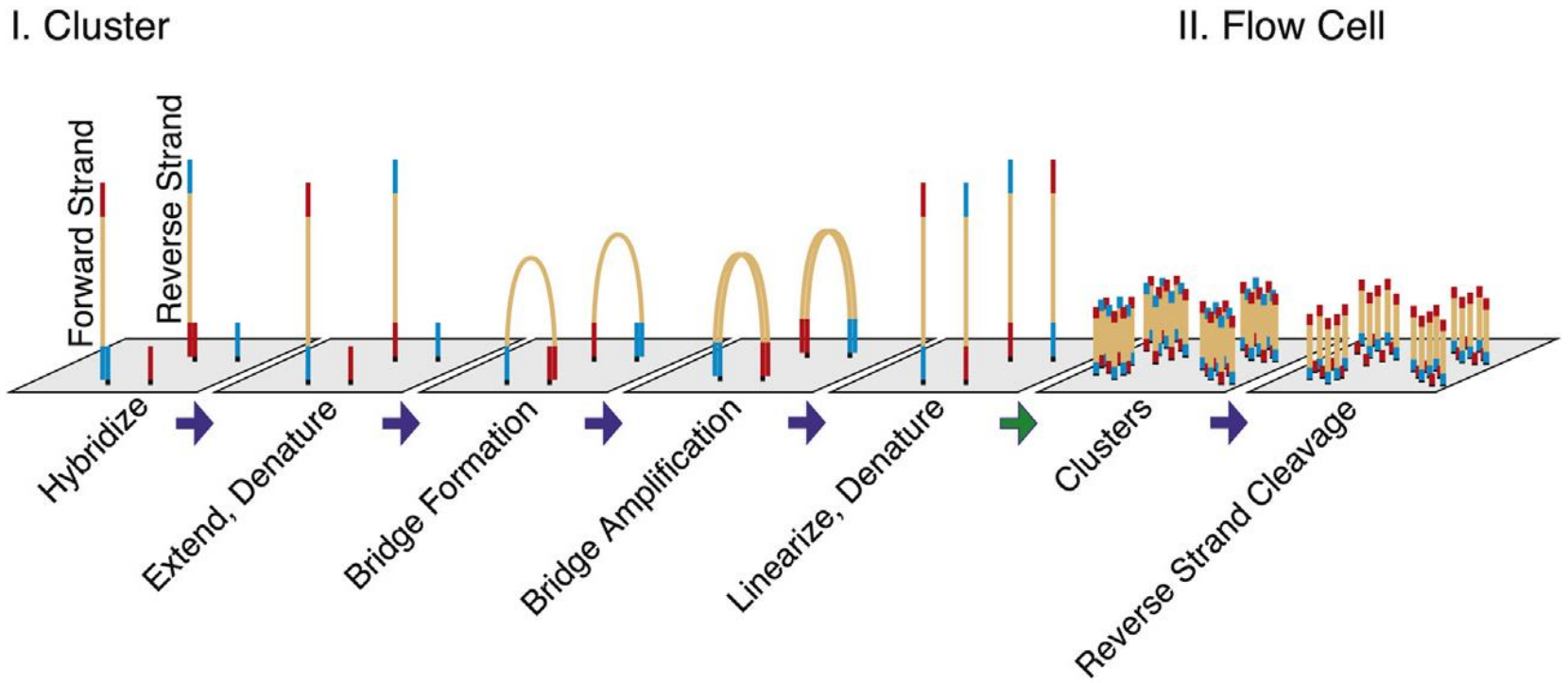
Coverage



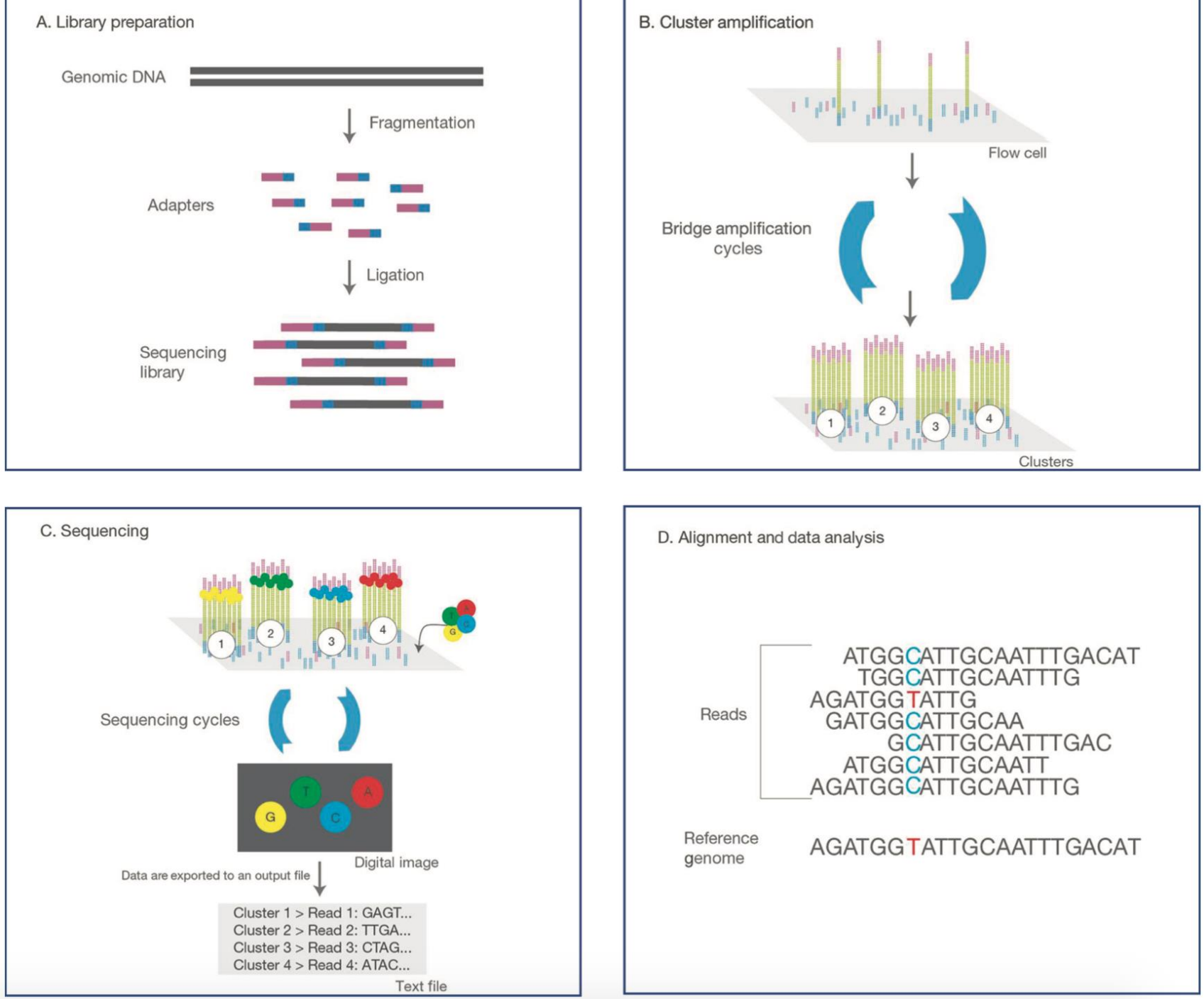
Coverage: # of reads underlying the consensus

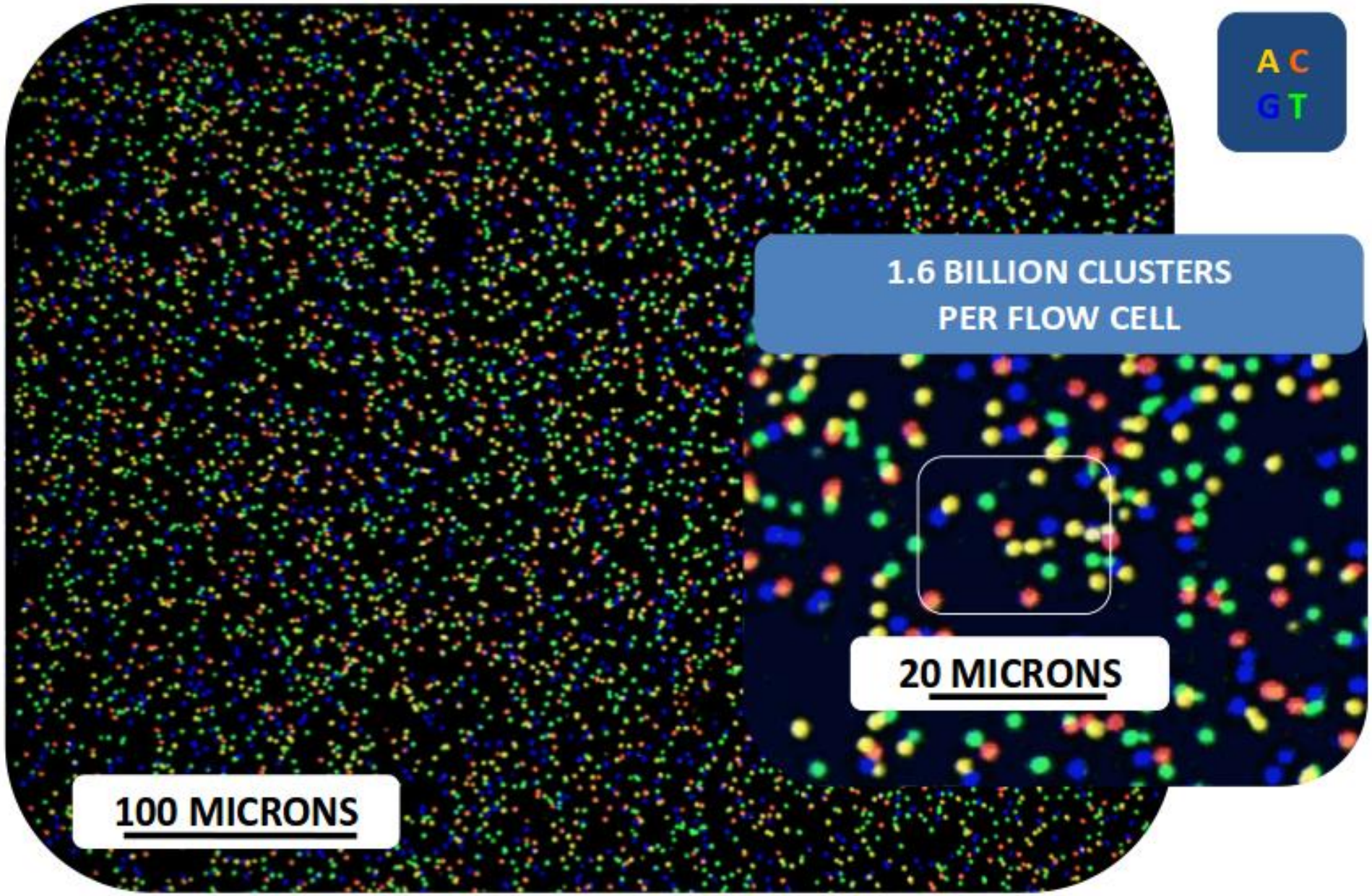
Overview - Illumina

Clustering

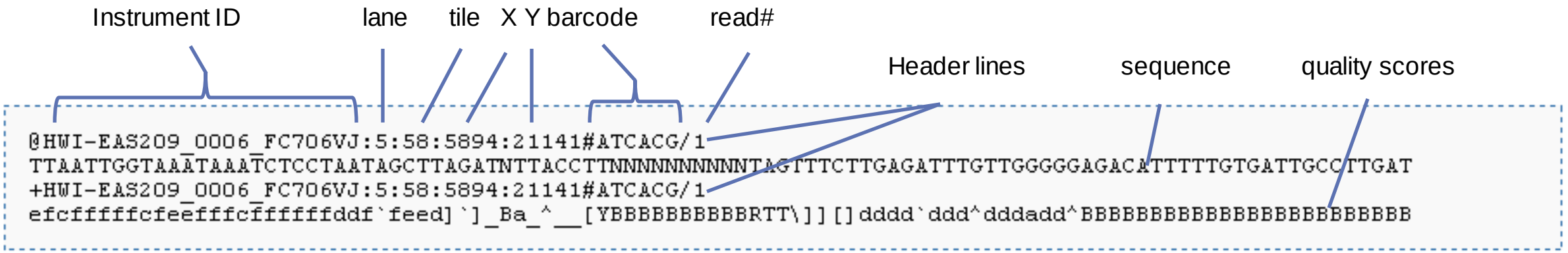


Next generation sequencing technology and genomewide data analysis: Perspectives for retinal research


A. Library preparation
Genomic DNA
Fragmentation
Adapters
Ligation
Sequencing library
B. Cluster amplification
Flow cell
Bridge amplification cycles
1
2
3
4
Clusters
C. Sequencing
1
2
3
4
Sequencing cycles
T
A
G
C
Digital image
Data are exported to an output file
Cluster 1 > Read 1: GAGT...
Cluster 2 > Read 2: TTGA...
Cluster 3 > Read 3: CTAG...
Cluster 4 > Read 4: ATAC...
Text file
D. Alignment and data analysis
Reads
ATGGCATTGCAATTTGACAT
TGGCATTGCAATTTG
AGATGGTATTG
GATGGCATTGCAA
GCATTGCAATTTGAC
ATGGCATTGCAATT
AGATGGCATTGCAATTTG
Reference genome
AGATGGTATTGCAATTTGACAT


A C
G T
1.6 BILLION CLUSTERS
PER FLOW CELL
20 MICRONS
100 MICRONS

Sequence data output format - fastq



- An important aspect of data analysis is knowing what you have.
  - At least four different ways to report quality scores
  - Header line formats differ with technology

```
                    2   3    1
CL100025298L1C002R050_244547
              XxxxYyyy FOV(~Tile)
```

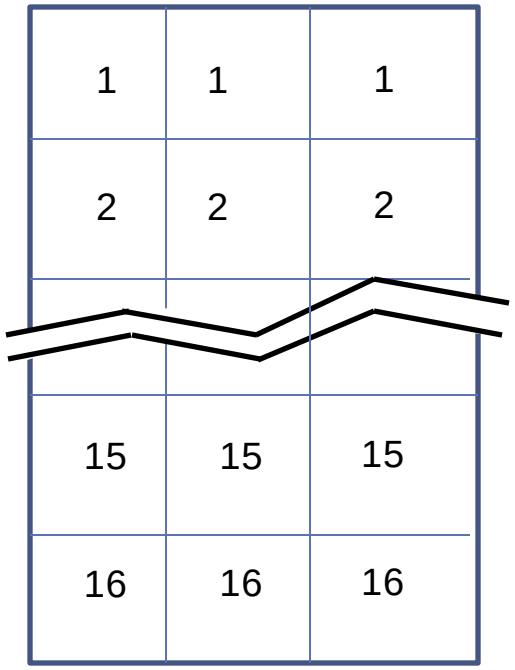
Image: BIT 815: Analysis of Deep Sequencing Data

See http://en.wikipedia.org/wiki/FASTQ_format for more details

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

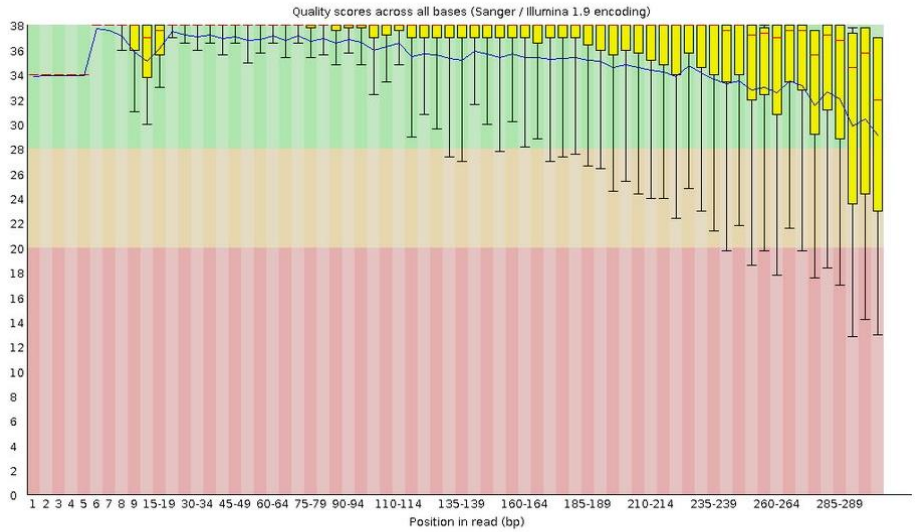
(Illumina v1.8 header version)

EAS139	the unique instrument name
136	the run id
FC706VJ	the flowcell id
2	flowcell lane
2104	tile number within the flowcell lane 4-digit number: 2 = over- (1) or underside (2) of flowcell 1 = 1 number of Swaths 04 = tile (image) number from 1-16 (or more depending on technology)
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 is filtered, N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	index sequence

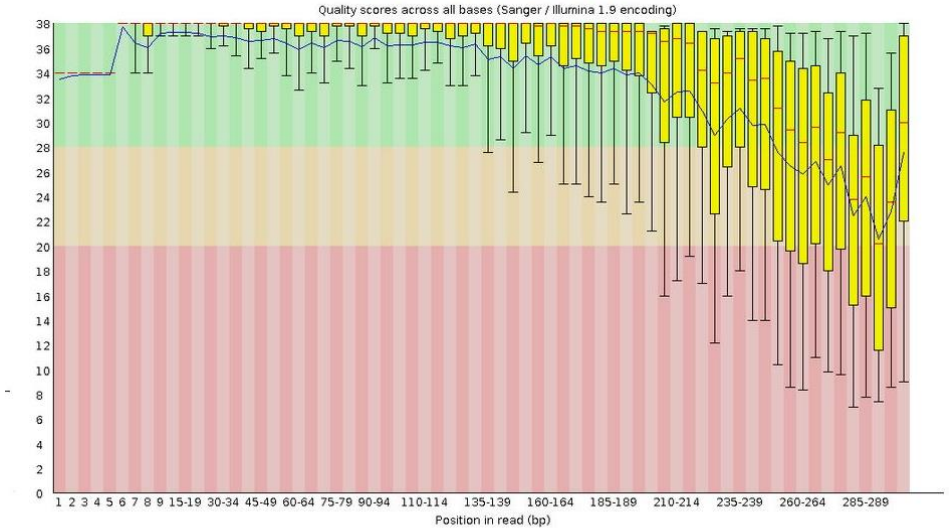


Challenges/Limitations with Illumina. R1 R2 variations

Forward reads



Reverse reads



https://rachaellappan.github.io/16S-analysis/pre-processing-reads.html

Why are repeats a problem?

- The law of repeats
  - It is impossible to resolve repeats of length L unless you have reads longer than L
  - It is impossible to resolve repeats of length L unless you have reads longer than L

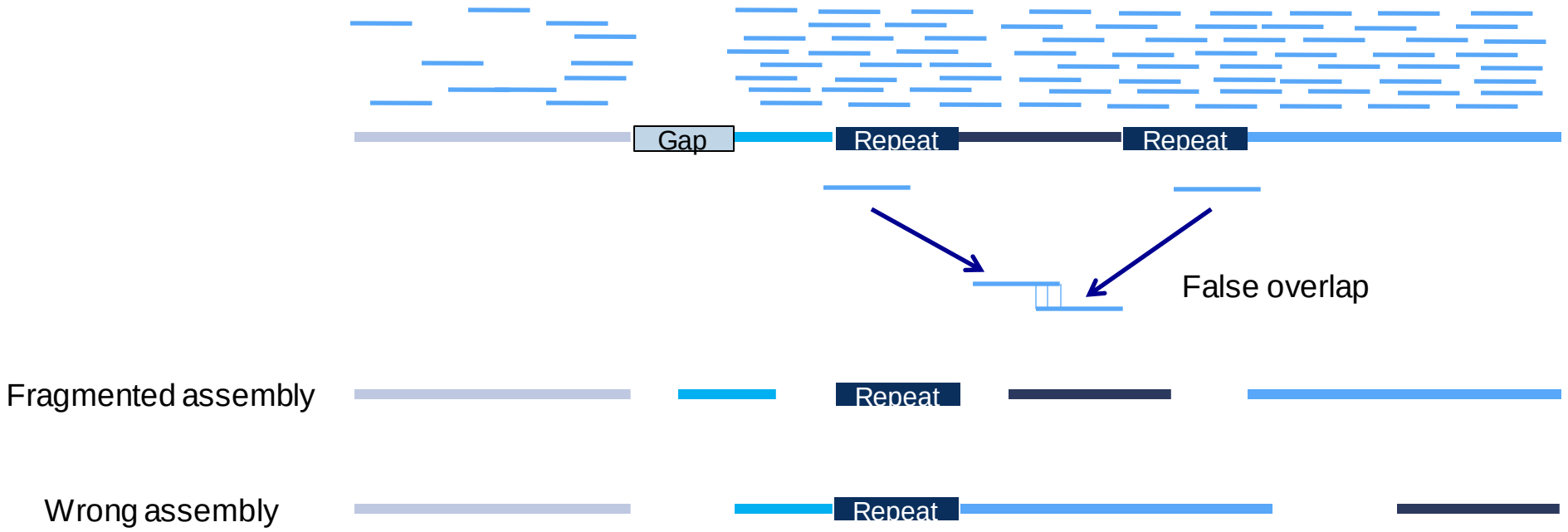
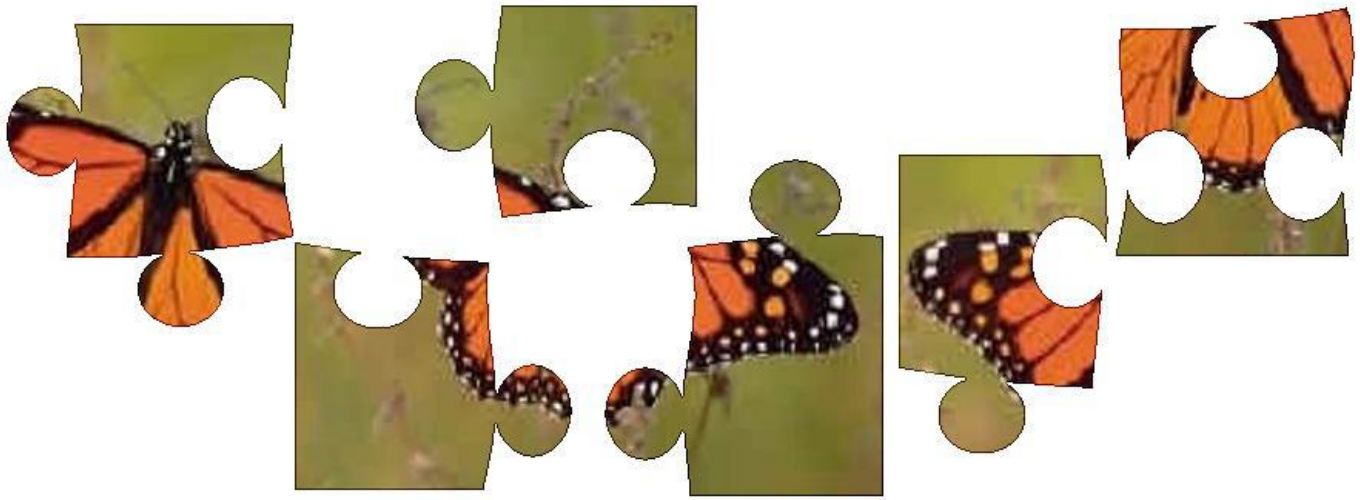


Image: Erik Hjerde

Long vs short reads



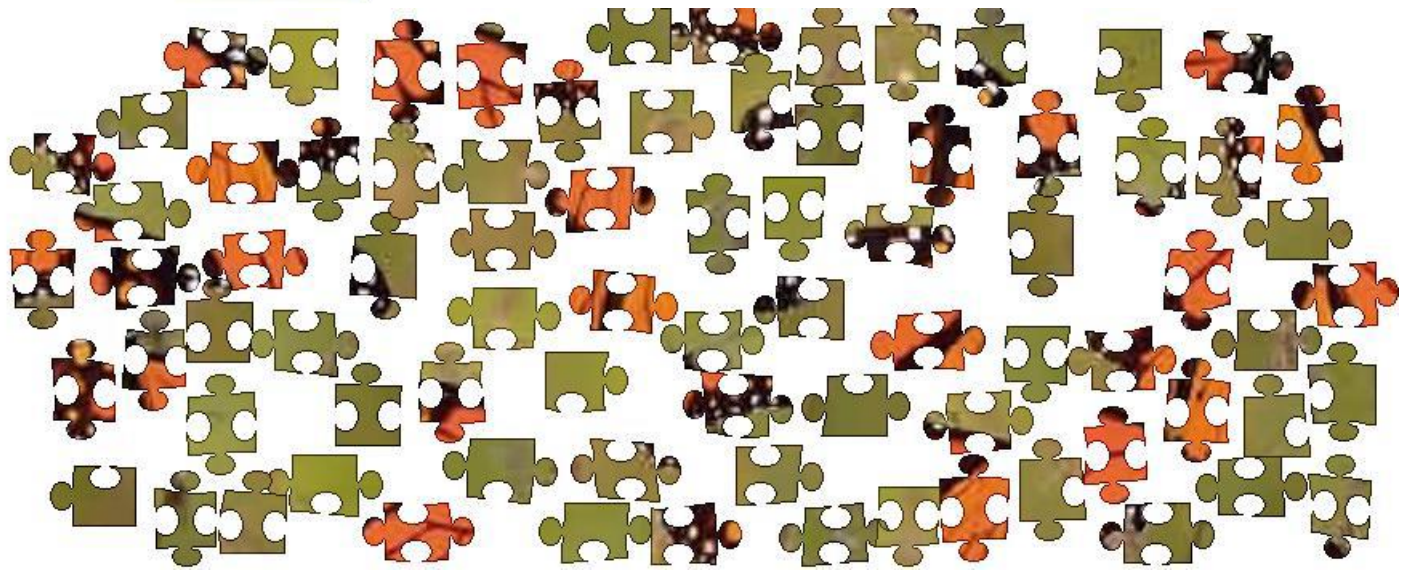
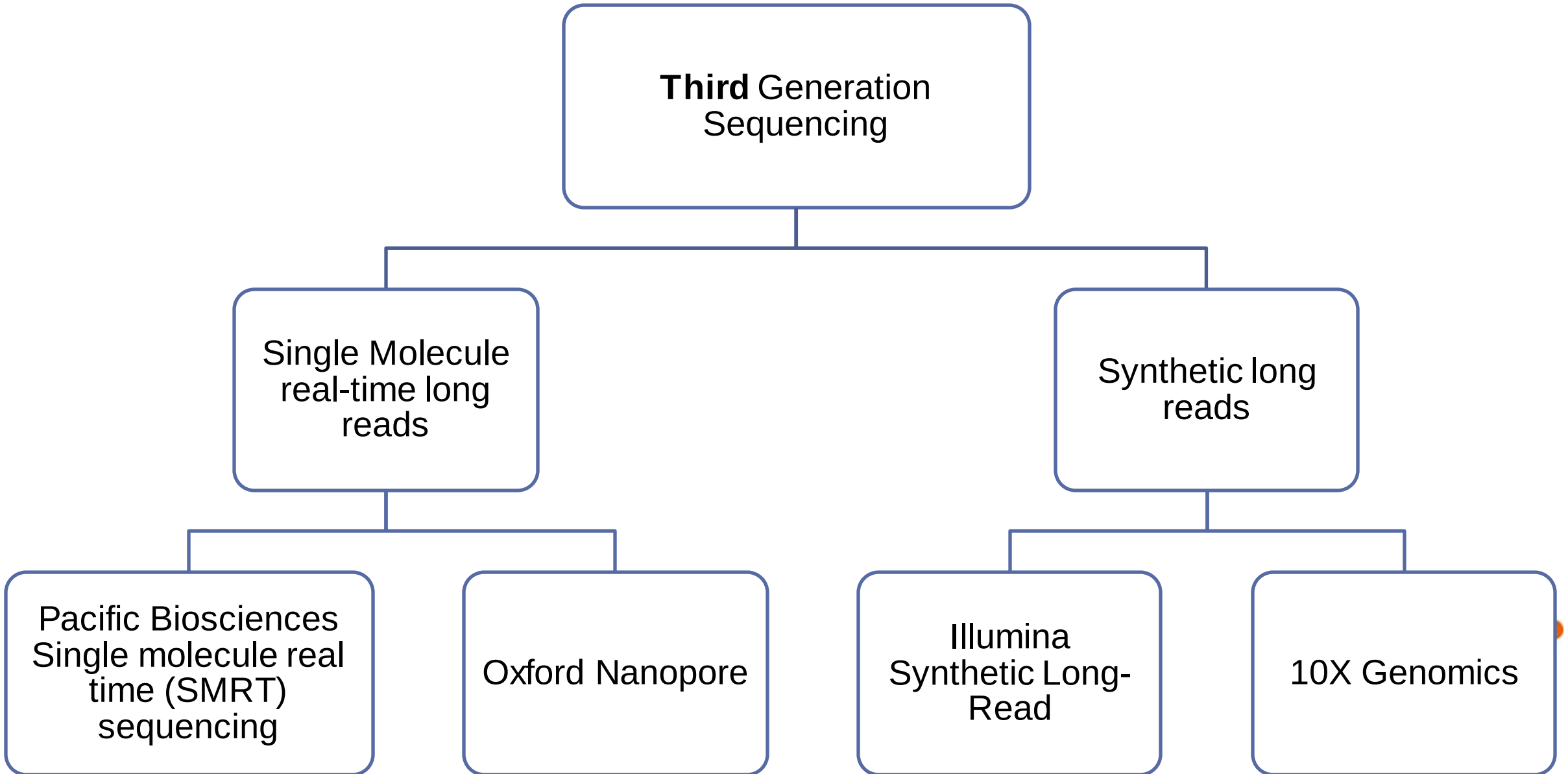
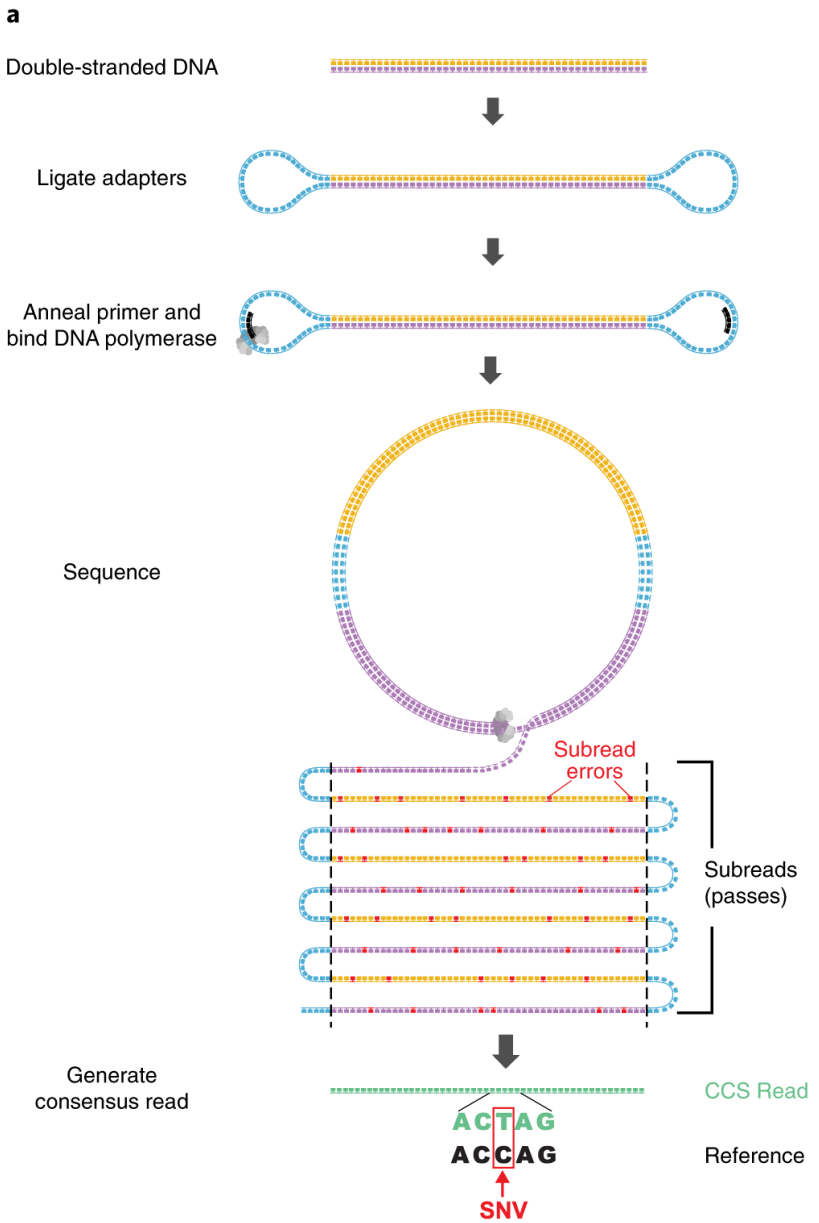


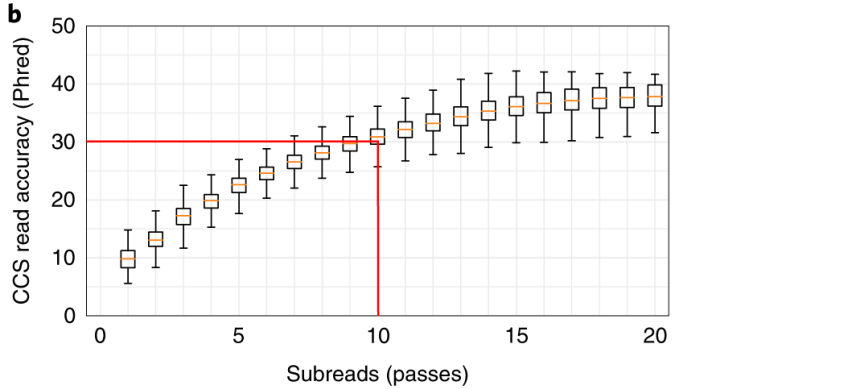
Image: Petri Auvinen

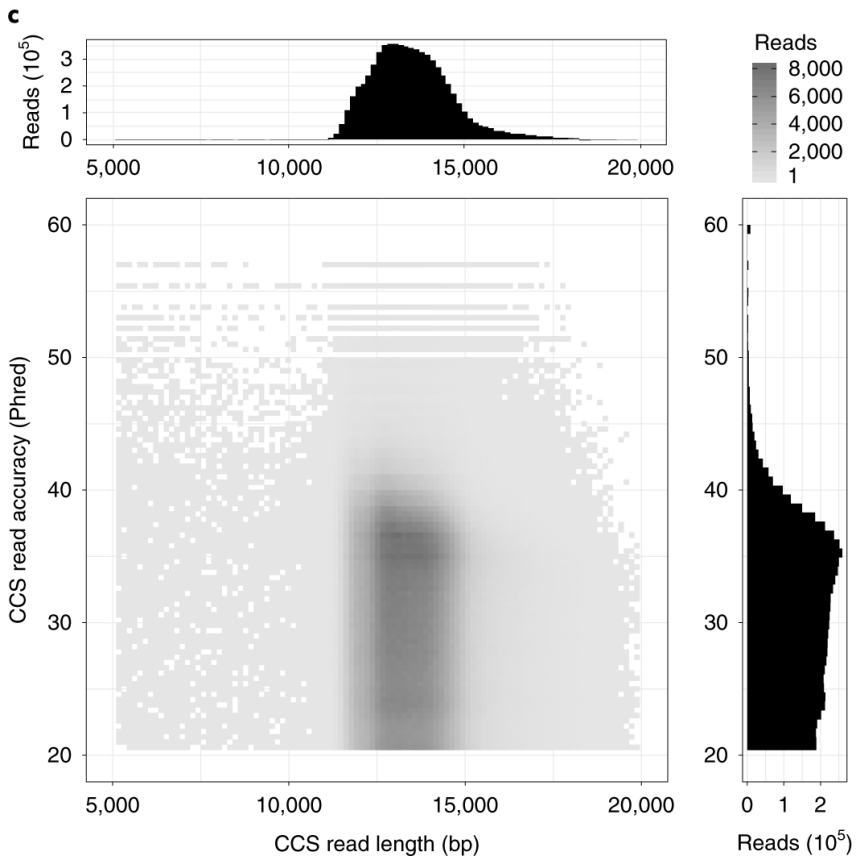
The solution?



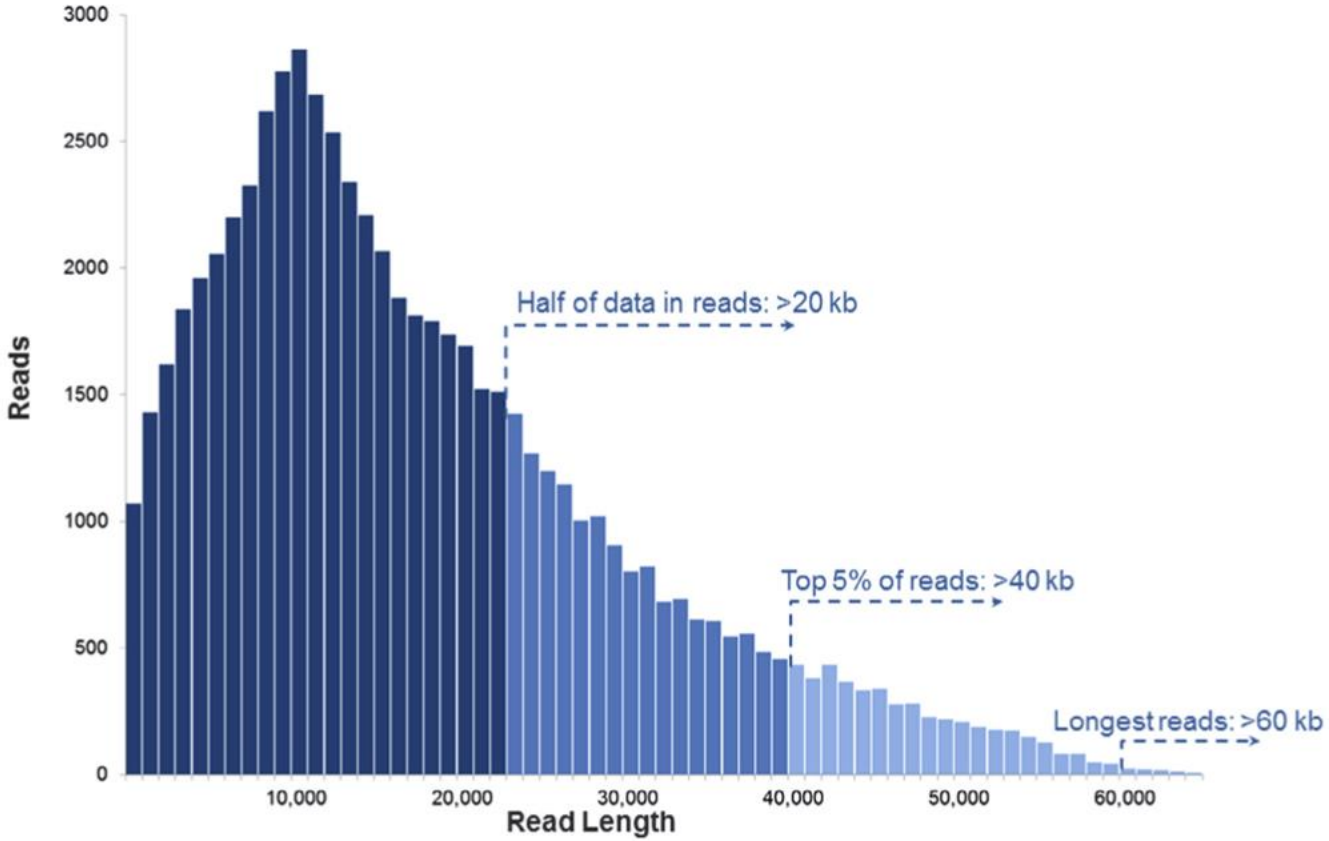
Pacific Biosciences: Sequencing DNA with highly accurate long reads



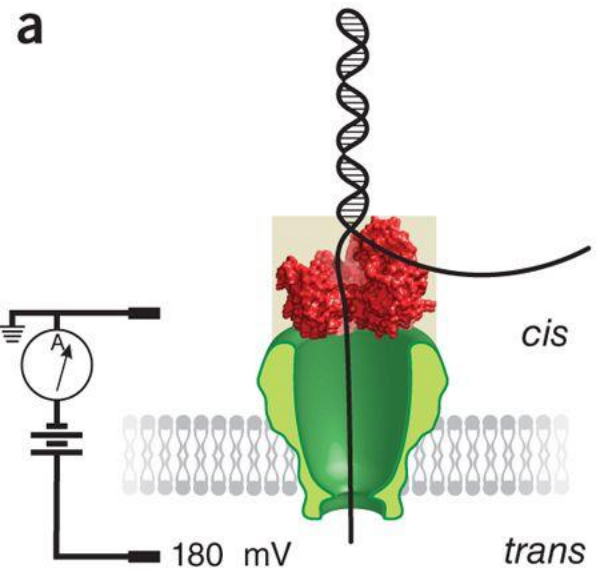


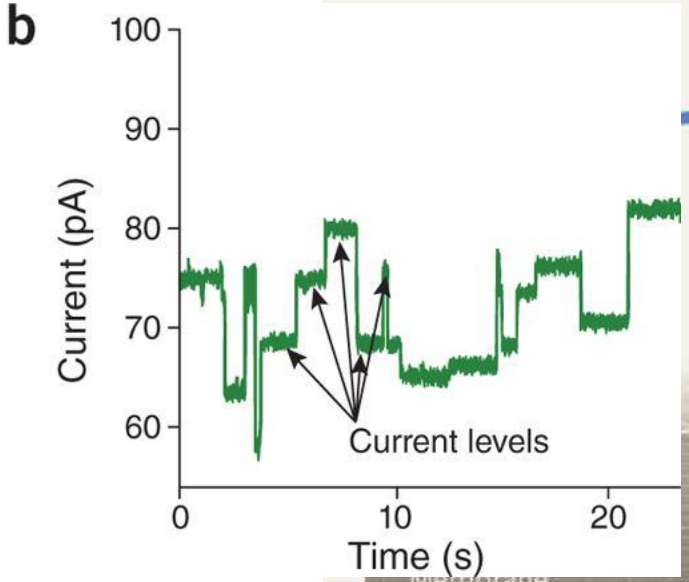


Pacific Biosciences: Sequencing DNA with highly accurate long reads



Genomics Proteomics Bioinformatics 13 (2015) 278–289


a
cis
trans
180 mV

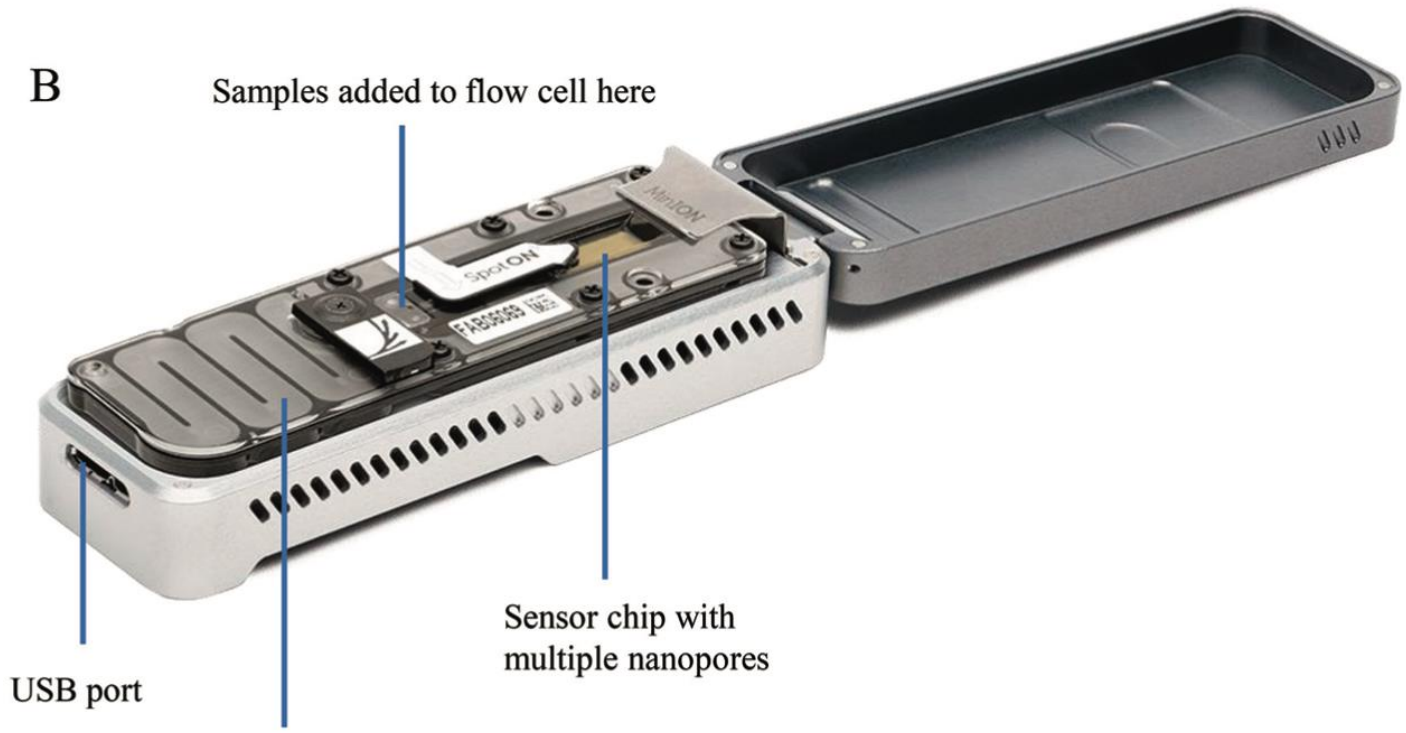

b
Current (pA)
100
90
80
70
60
0
10
20
Time (s)
Current levels

NANOPORE SEQUENCING
At the heart of the MinION device, an enzyme unwinds DNA, feeding one strand through a protein pore. The unique shape of each DNA base causes a characteristic disruption in electrical current, providing a readout of the underlying sequence.
DNA double helix
DNA base
Unwinding enzyme
Protein pore
Membrane
Ion
Current
Current
Sequence A A C T C G T

MinION (Oxford Nanopore)

- Very portable
- No special equipment to run
- Simple run
  - 10 minute prep
- Very cheap to run
  - $500-900 per (reusable flow-cell)
- Very long (100kb is not unusual) (record 2272580 bases)
- Max throughput 10-30 Gb pr single flowcell
- Reads appear in real-time (pull the USB plug when you have enough data)

B


Samples added to flow cell here
MinION
SpotON
Sensor chip with
multiple nanopores
USB port
Flow cells containing
sensing chemistry,
nanopore, and electronics

Futuromics: SmidgION and the Flongle (Oxford Nanopore)

Hybrid / long read assemblies in metagenomics are getting more commonplace

nature communications

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature communications › articles › article

Article | Open Access | Published: 11 July 2019

Strain-level metagenomic assignment and compositional estimation for long reads with MetaMaps

Alexander T. Dilthey ✉, Chirag Jain, Sergey Koren & Adam M. Phillippy

Nature Communications **10**, Article number: 3066 (2019) | Cite this article

12k Accesses | **24** Citations | **66** Altmetric | Metrics

BMC Genomics

Home About Articles Submission Guidelines

Research | Open Access | Published: 06 May 2021

Long-read metagenomics retrieves complete single-contig bacterial genomes from canine feces

Anna Cuscó ✉, Daniel Pérez, Joaquim Viñes, Norma Fàbregas & Olga Francino

BMC Genomics **22**, Article number: 330 (2021) | Cite this article

1866 Accesses | **2** Citations | **36** Altmetric | Metrics

PacBio Long Reads Improve Metagenomic Assemblies, Gene Catalogs, and Genome Binning

Haiying Xie[1,2†], Caiyun Yang[1†], Yamin Sun[3], Yasuo Igarashi[1], Tao Jin[4*] and Feng Luo[1,2*]

[1]Research Center of Bioenergy and Bioremediation, College of Resources and Environment, Southwest University, Chongqing, China
[2]PUROTON Gene Medical Institute Co., Ltd., Chongqing, China
[3]Research Center for Functional Genomics and Biochip, Tianjin Biochip Co., Ltd., Tianjin, China
[4]The Beijing Genomics Institute (BGI)-Shenzhen, Shenzhen, China

nature communications

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature communications › articles › article

Article | Open Access | Published: 04 January 2021

Long-read metagenomics using PromethION uncovers oral bacteriophages and their interaction with host bacteria

Koji Yahara ✉, Masato Suzuki, Aki Hirabayashi, Wataru Suda, Masahira Hattori, Yutaka Suzuki & Yusuke Okazaki

Nature Communications **12**, Article number: 27 (2021) | Cite this article

5538 Accesses | **4** Citations | **66** Altmetric | Metrics

	Sequence in contigs	Contig N50 (bp)	Number of contigs
Short read only	7.91 Gb	12,036	2,507,175
Short read + nanopore long reads	8.12 Gb	359,531	49,676

Questions?