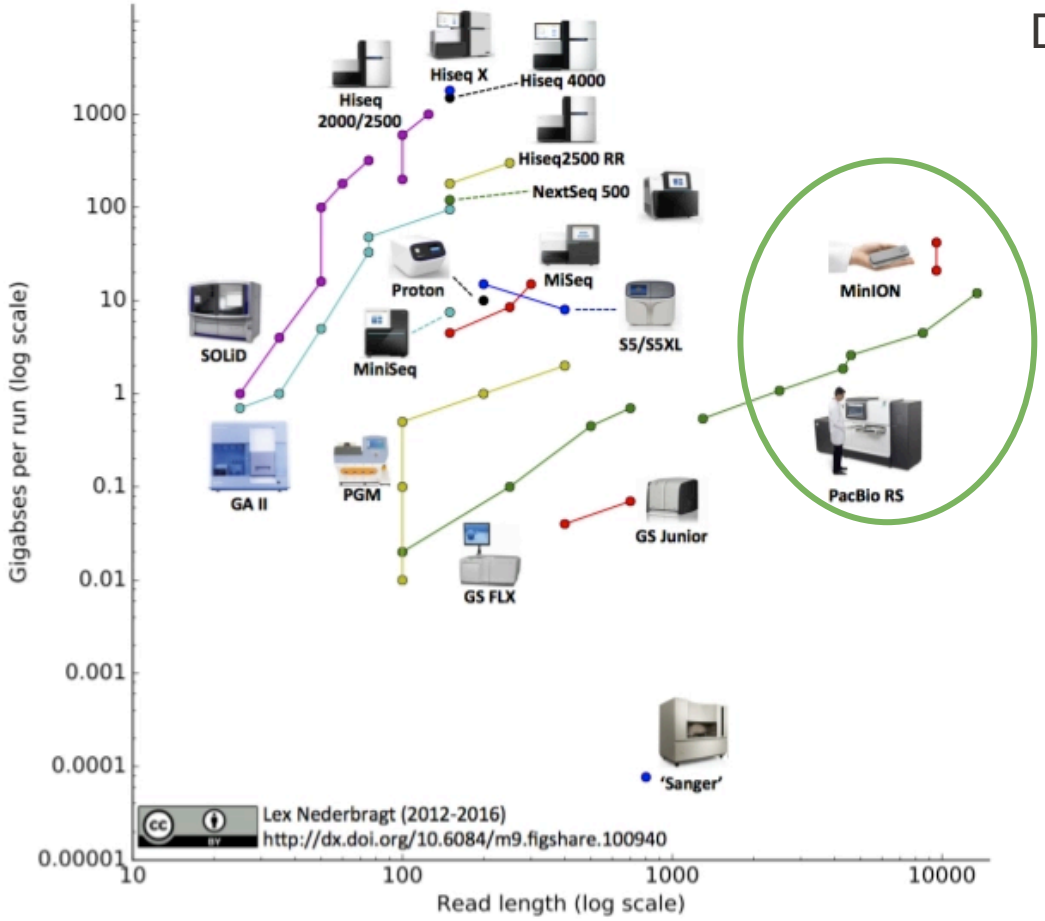
Introduction to Nanopore sequencing

Thomas Haverkamp
@Thomieh

Outline

- The nanopore sequencing method
- Software applications for Nanopore
  - Genome assembly
  - Amplicon Sequencing
- A small NGS comparison

Rapid development in instrumentation



Drastic increase in both

- Read length
- Amount of sequence / run

- Single molecule sequencer
- Long read sequencers

https://figshare.com/articles/developments_in_NGS/100940

A nanopore sequencer

minION sequencer & flowcell

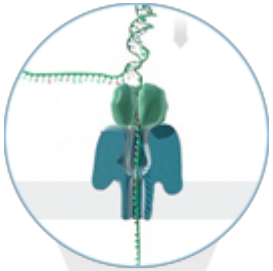
Nanopore minION flowcell

Specifications:

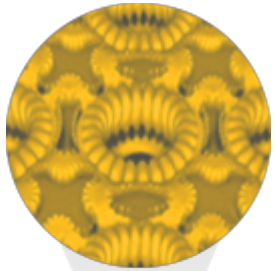
- 512 pores (Guaranteed)
- Needs to be stored at 2-8 ^{O}C
- Pores deteriorate over time - Fresh is best
- Longest single read of a single molecule sequence

  'Record': 2 Mbp
- 'Happy' at about 15 kb
- Up to 450 bases per second / sampling rate 4000 kHz
- May give a near 'realtime sequencing' data for up to 48 hrs
- Current capacity up to 48 hrs/20-40gb

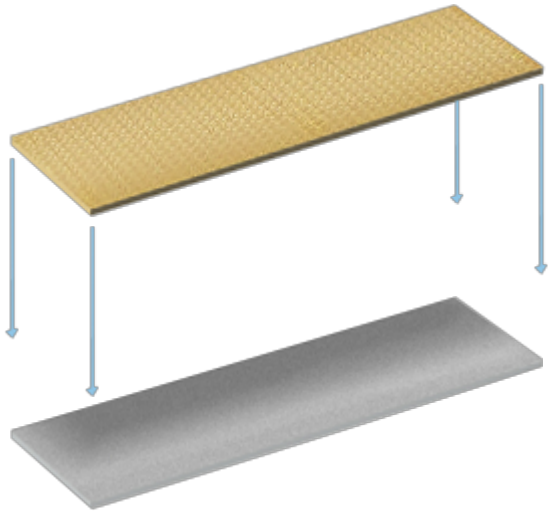
The nanopore sensor chip



Nanopore A protein nanopore is set in an electrically-resistant polymer membrane.



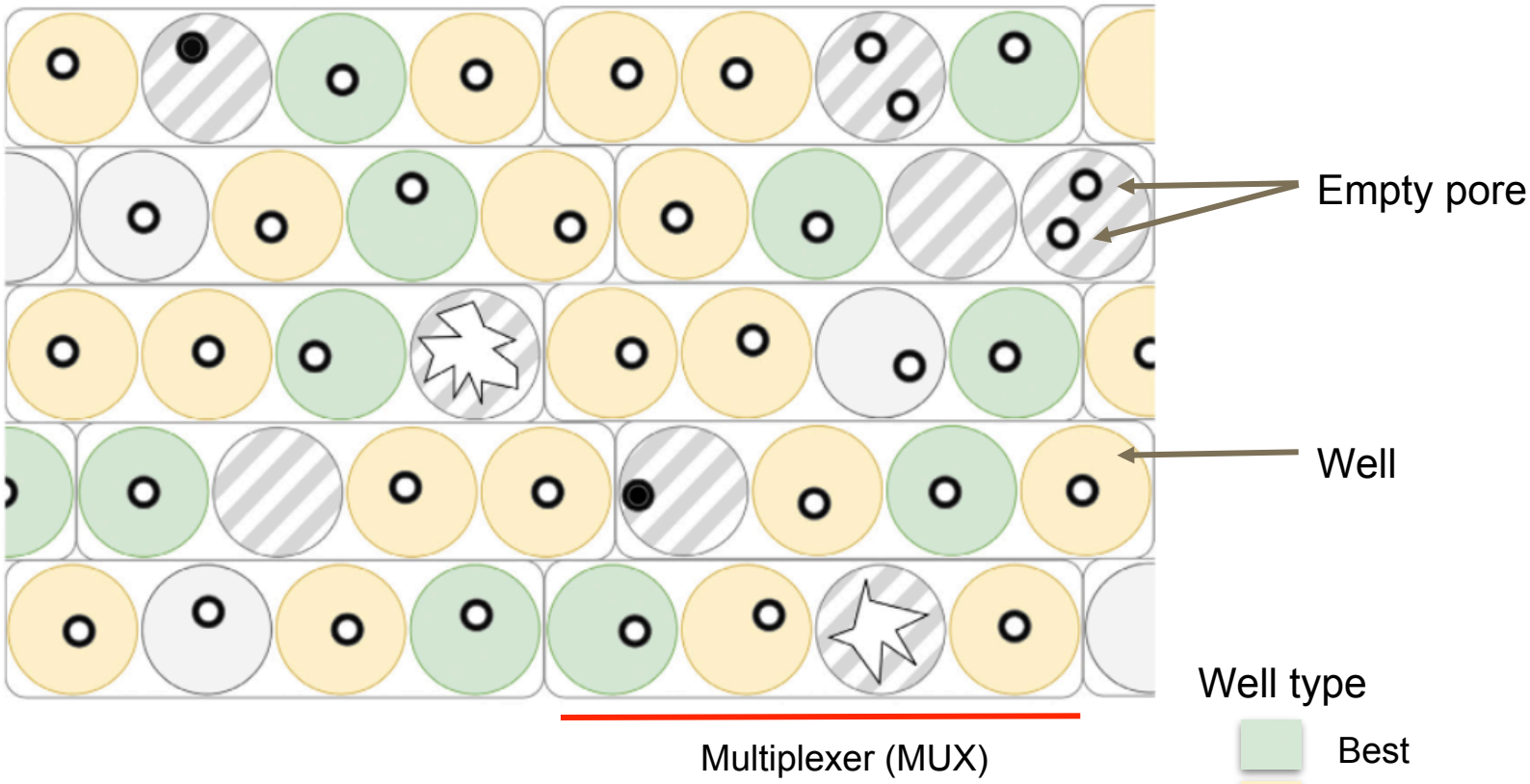
Array of microscaffolds
Each microscaffold supports a membrane and embedded nanopore.



Sensor chip
Each microscaffold corresponds to its own electrode that is connected to a channel in the sensor array chip.

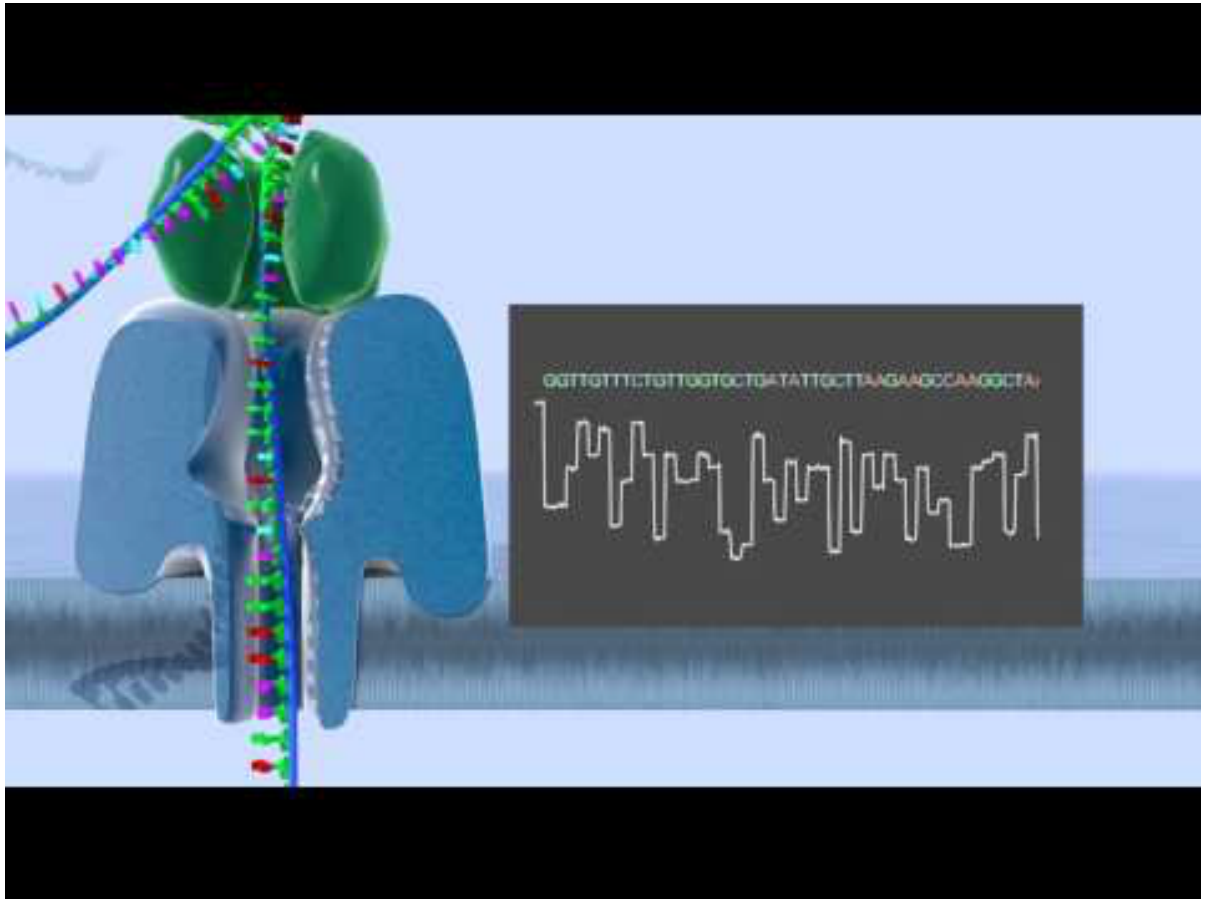
ASIC Application-Specific Integrated Circuit
Each nanopore channel is controlled and measured individually by the bespoke ASIC.

The flow cell layout

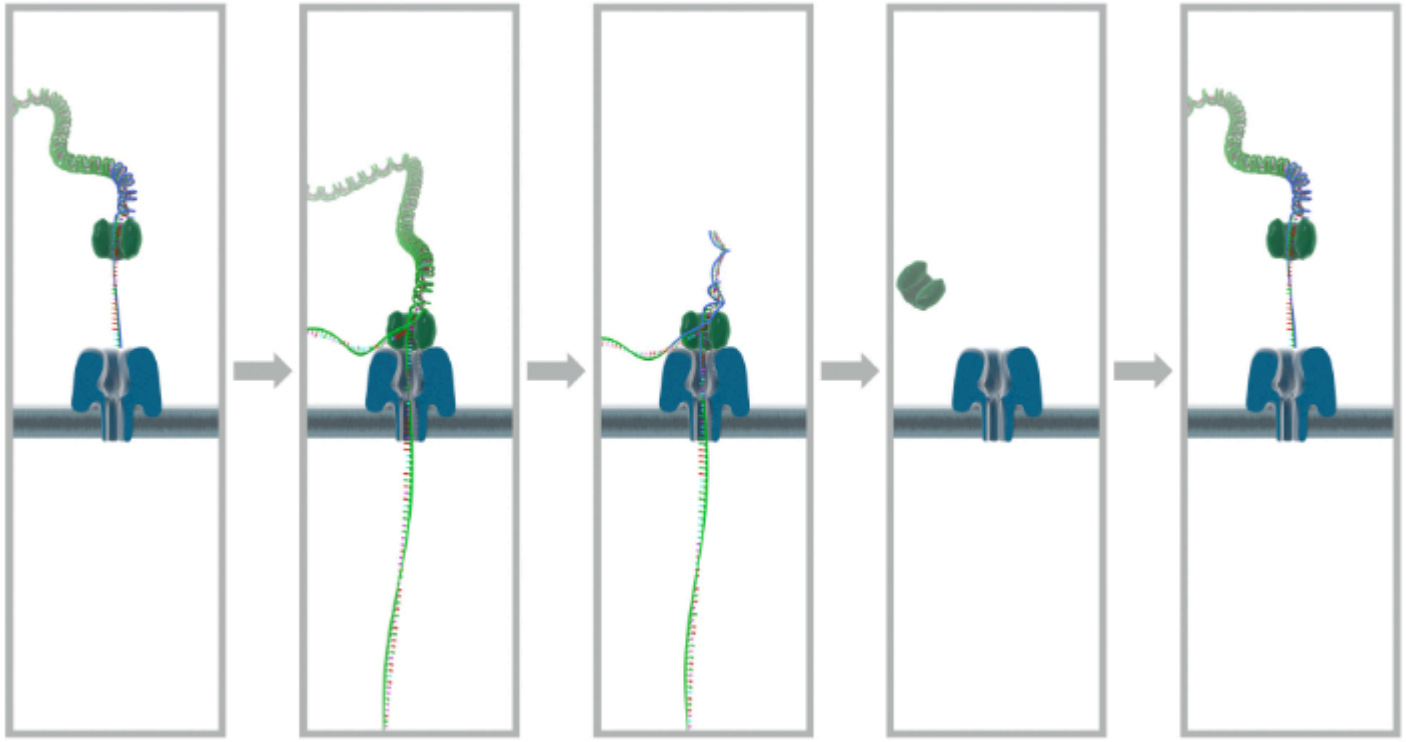


A flow cell has 2048 wells → 512 pores sequenced in parallel

Nanopore sequencing explained



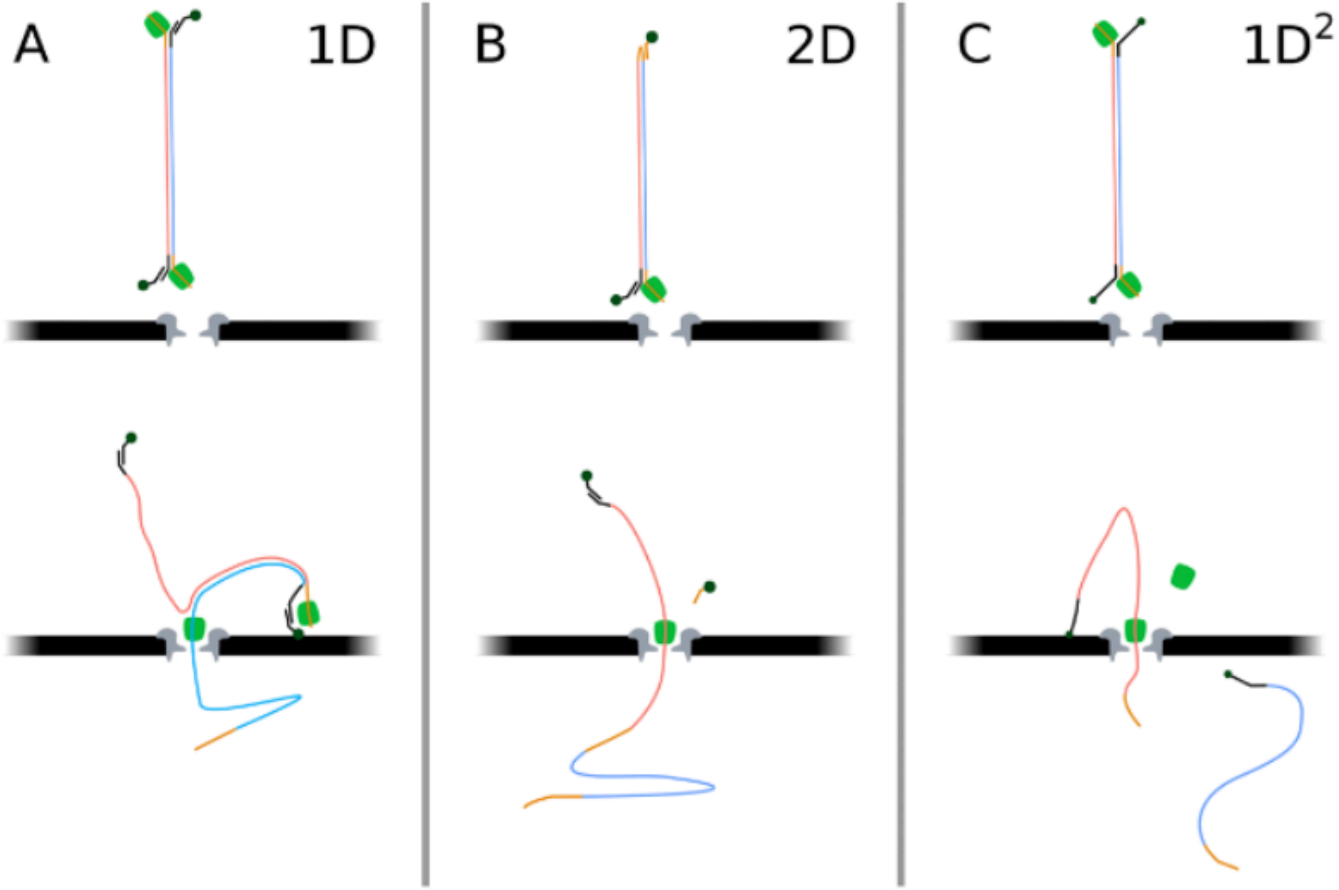
Nanopore sequencing



- The electric potential over the membrane pulls the DNA toward the nanopore.
- The motor protein regulates the speed of sequencing (≈ 450 bases s^{-1}).
- Current changes are measured when a base is pulled through the pore.

De Lannoy et al., 2017 F1000 research (DOI: 10.12688/f1000research.12012.1)

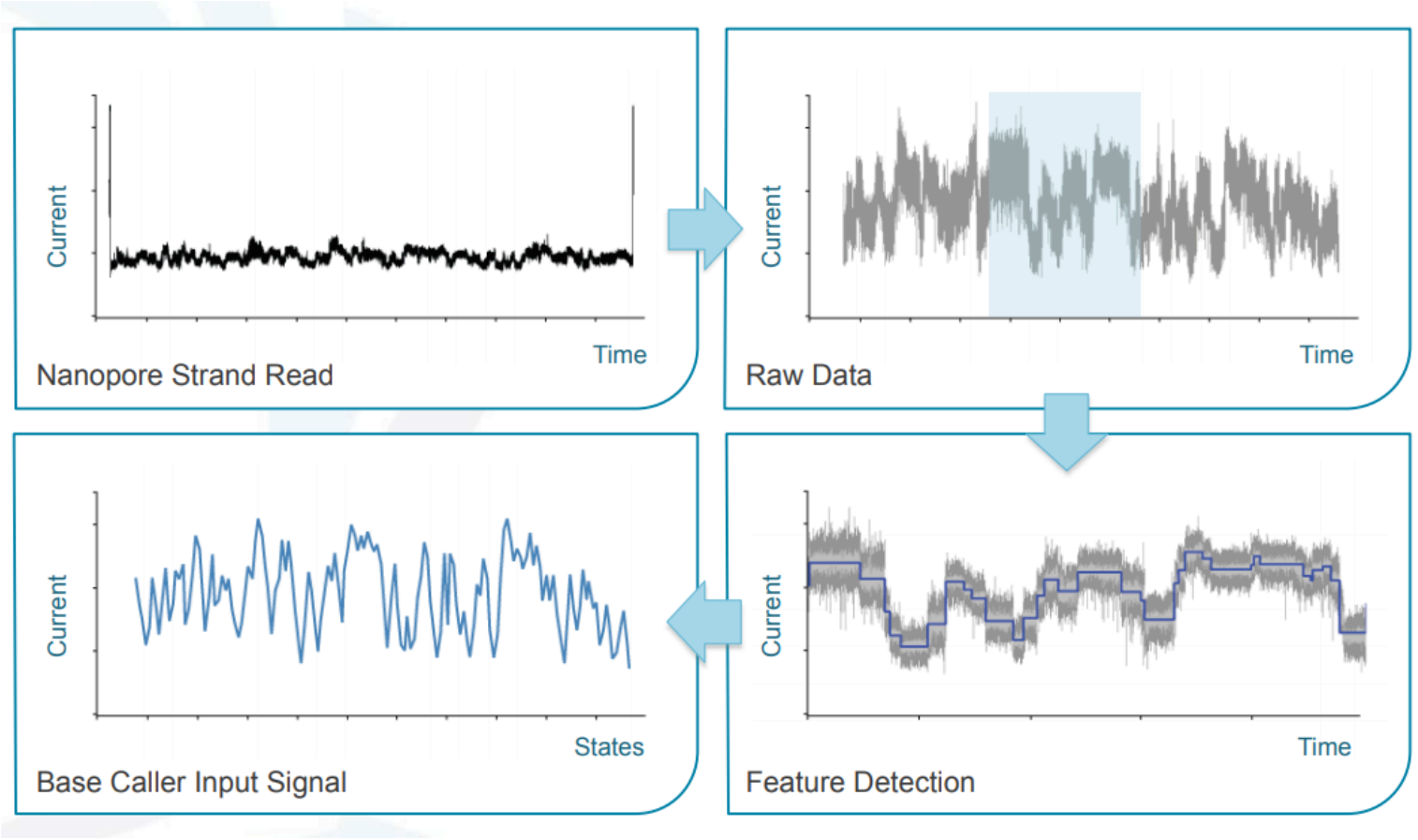
1D vs 2D sequencing



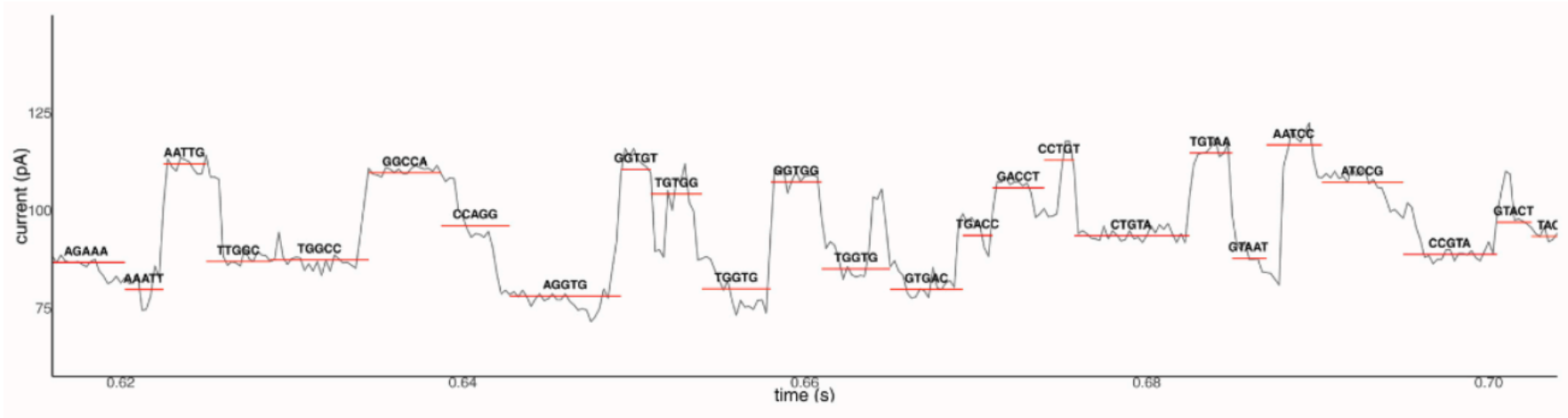
Note: 2D sequencing is no longer available. $1D^2$ is now the standard.

De Lannoy et al., 2017 F1000 research (DOI: 10.12688/f1000research.12012.1)

Nanopore Basecalling



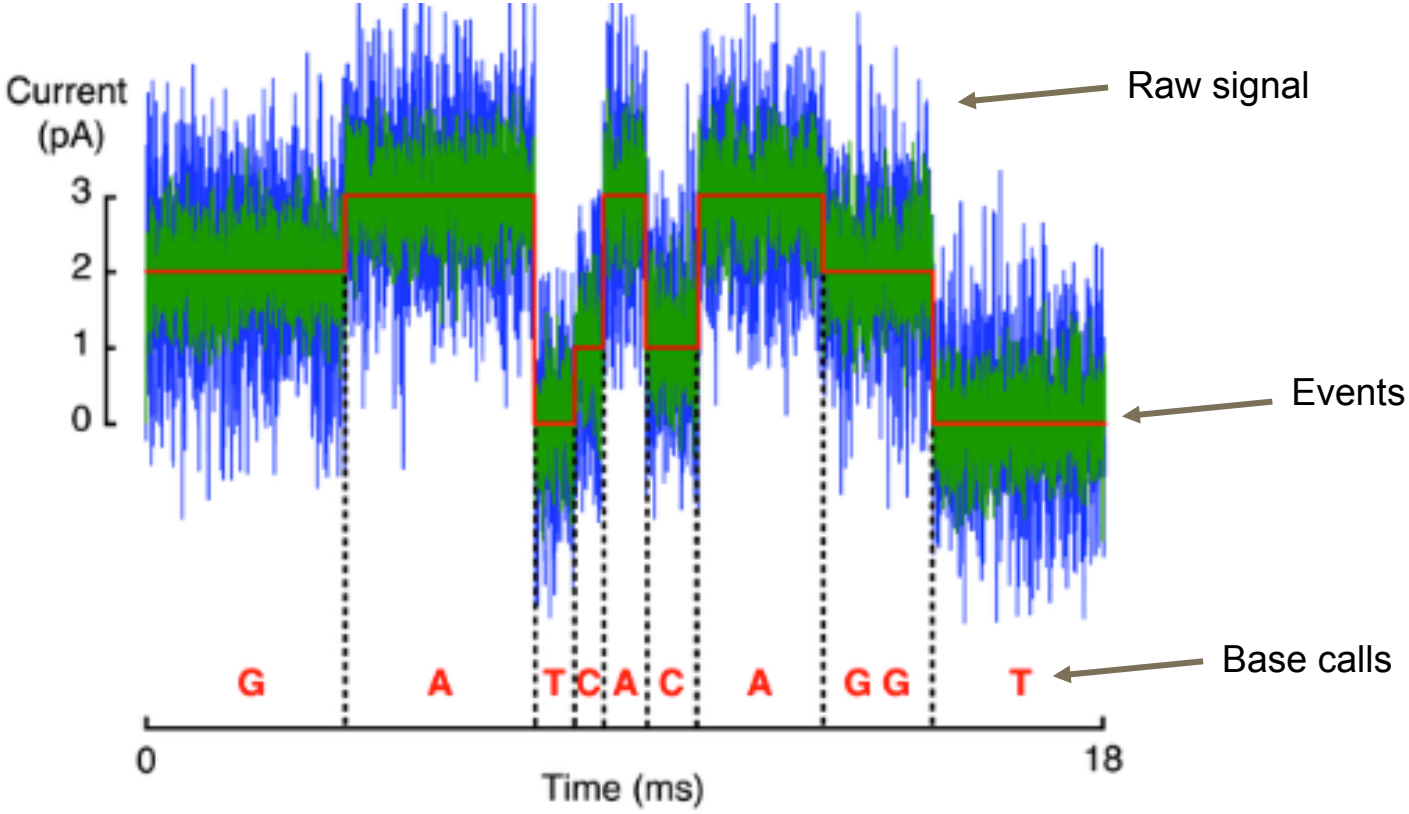
Nanopore basecalling



- The length of the passage (pore) determines the signal
- The assumption was that 5 bases fitted in the pore.
- Newer basecallers dropped assumption and derive basecalls directly from the signal

De Lannoy et al., 2017 F1000 research (DOI: 10.12688/f1000research.12012.1)

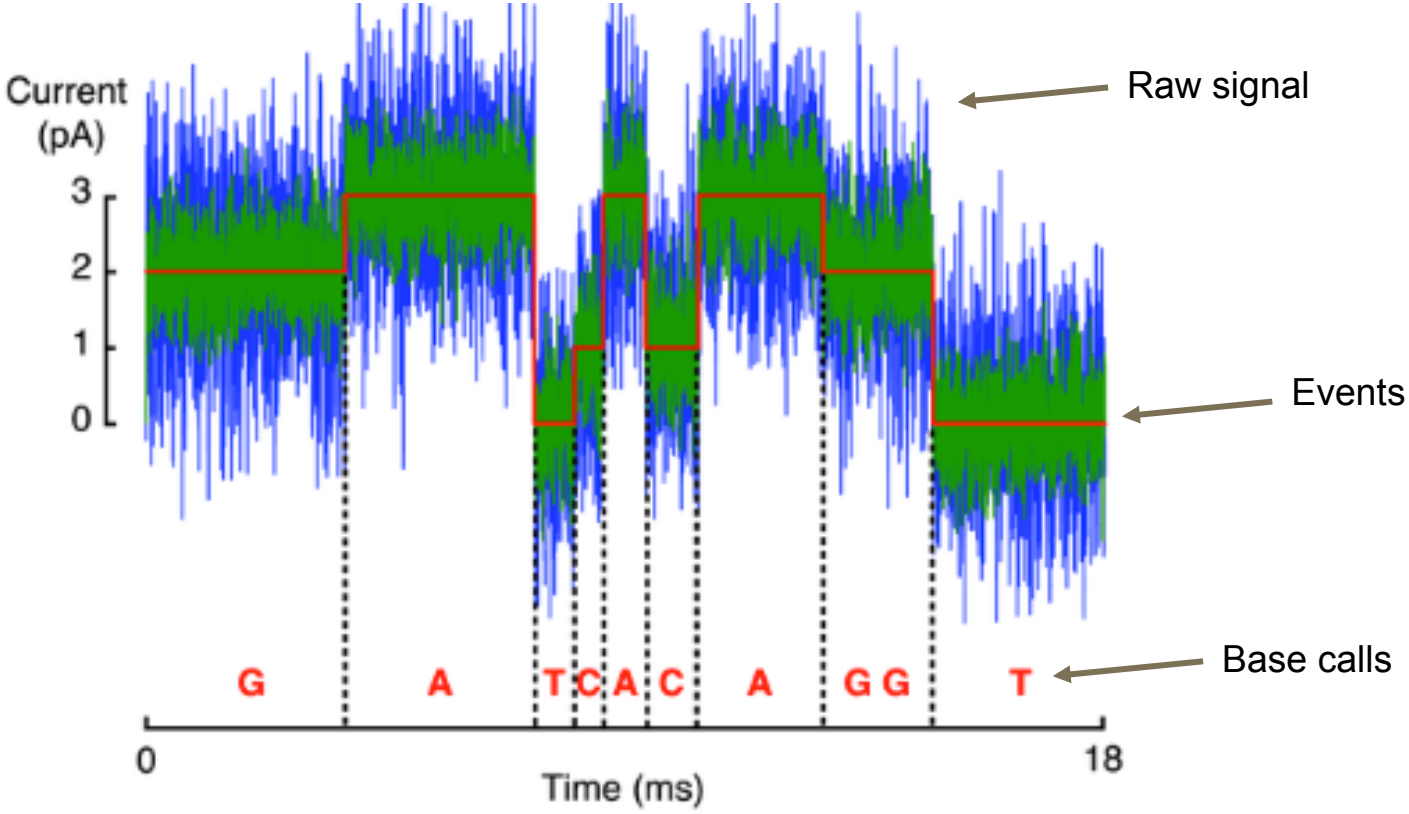
Variation in basecalling



- Translocation time through the pore time is variable
- Depending on the surrounding sequence
- Basecallers need advanced algorithms to deal with this “noisy data”.

O’Donnell et al., 2013 Electrophoresis (DOI: 10.1002/elps.201300174)

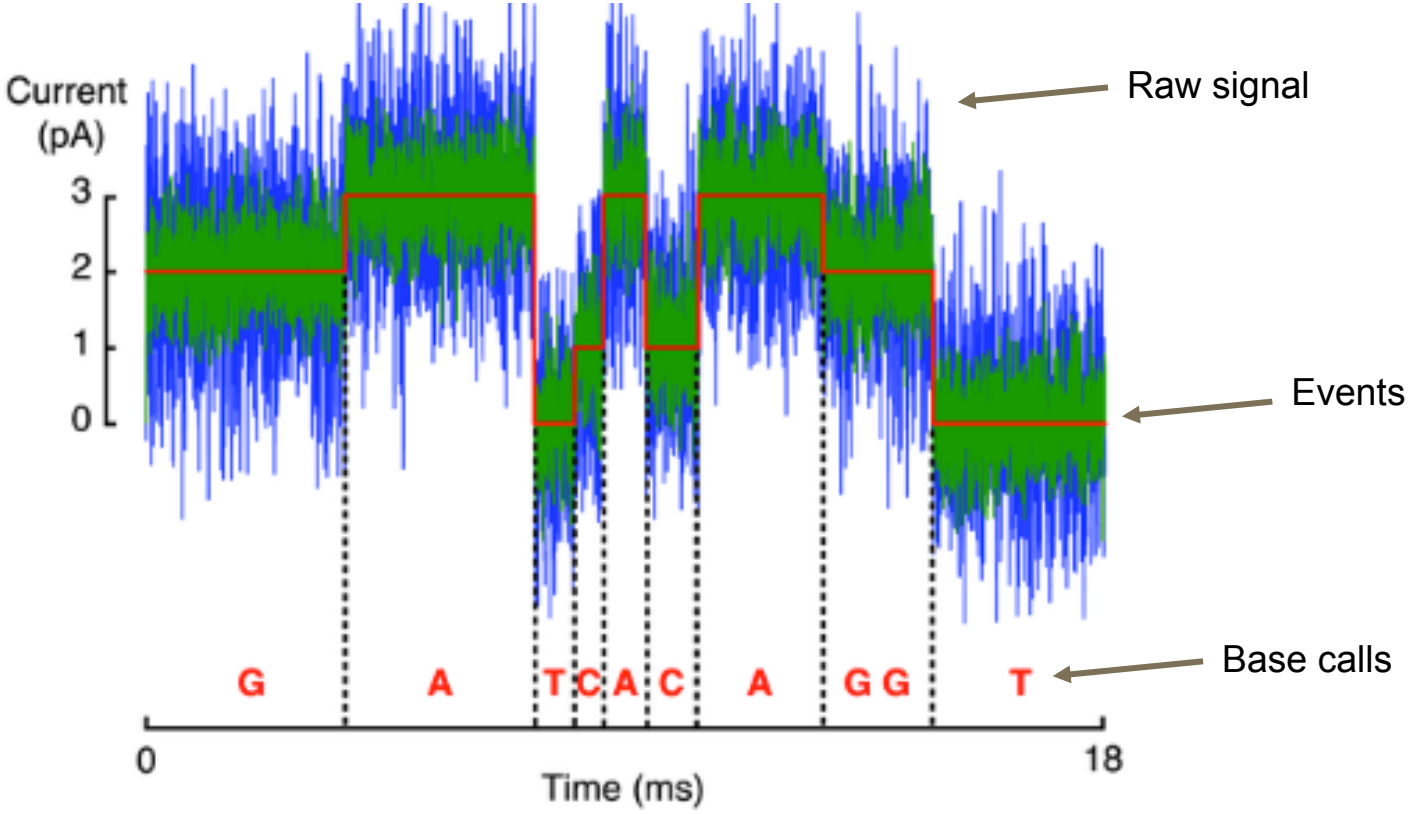
Improving basecalling



- Addition of Lambda DNA might improve basecalling per run.
- But the software needs to be able to use that information

O'Donnell et al., 2013 Electrophoresis (DOI: 10.1002/elps.201300174)

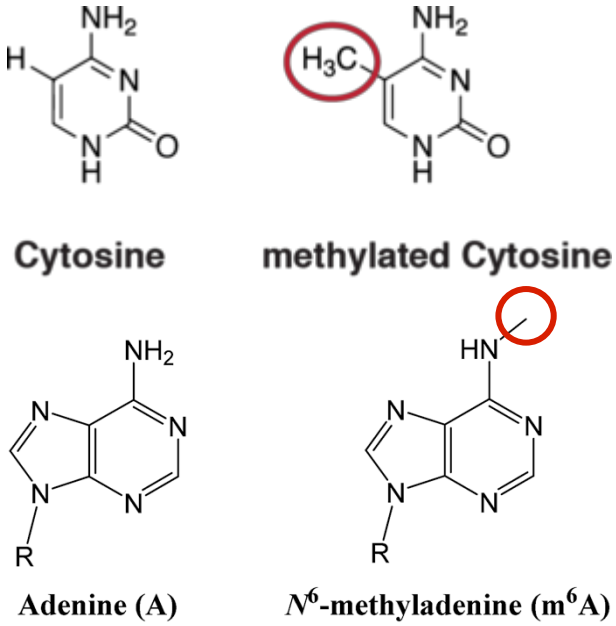
DNA methylation and basecalling



- Basecalling is highly variable.
- Methylated bases have a different signal than non-methylated bases.

O'Donnell et al., 2013 Electrophoresis (DOI: 10.1002/elps.201300174)

DNA methylation



Methylation in Eukaryotes needed for:

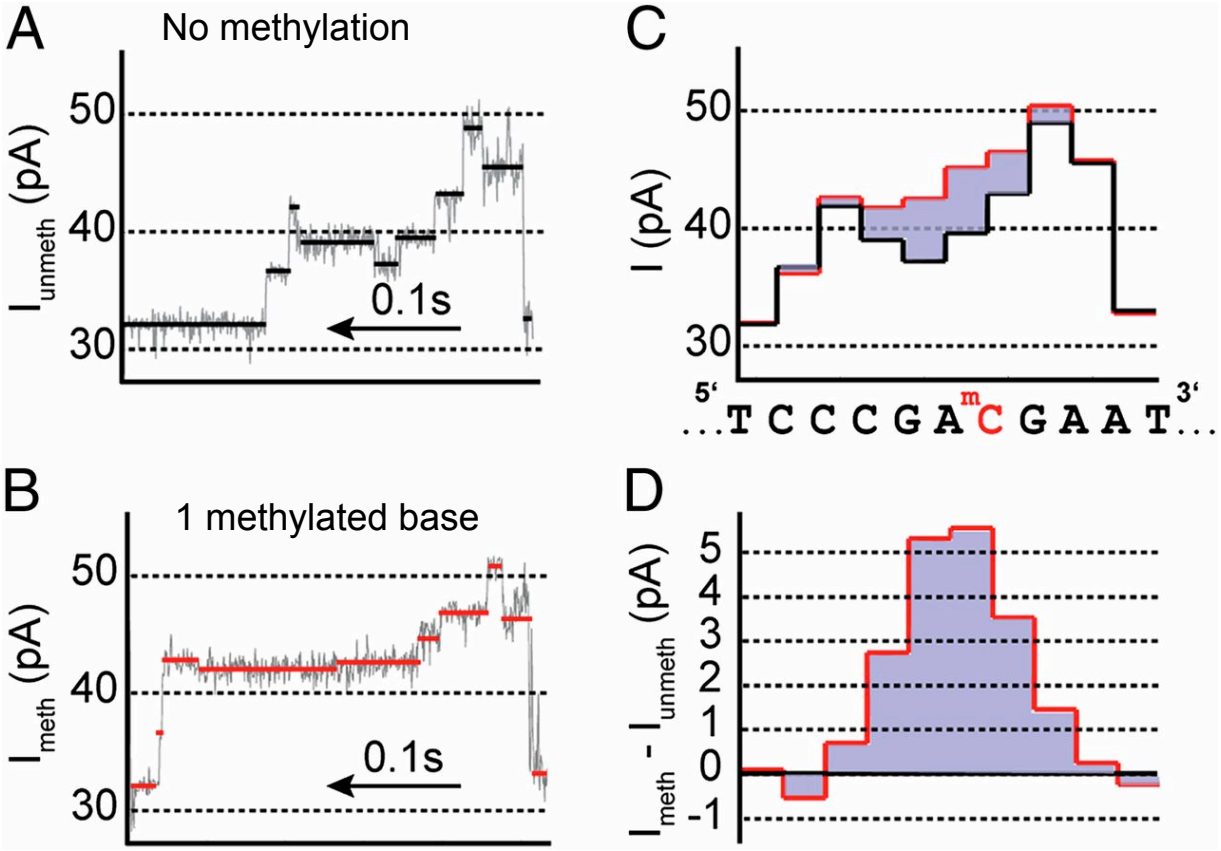
- Gene regulation
- Cell differentiation
- Silencing of mobile elements

Methylation in Prokaryotes:

- Silencing of mobile elements
- Phages recognized
- Gene regulation

Methylated nucleotides.

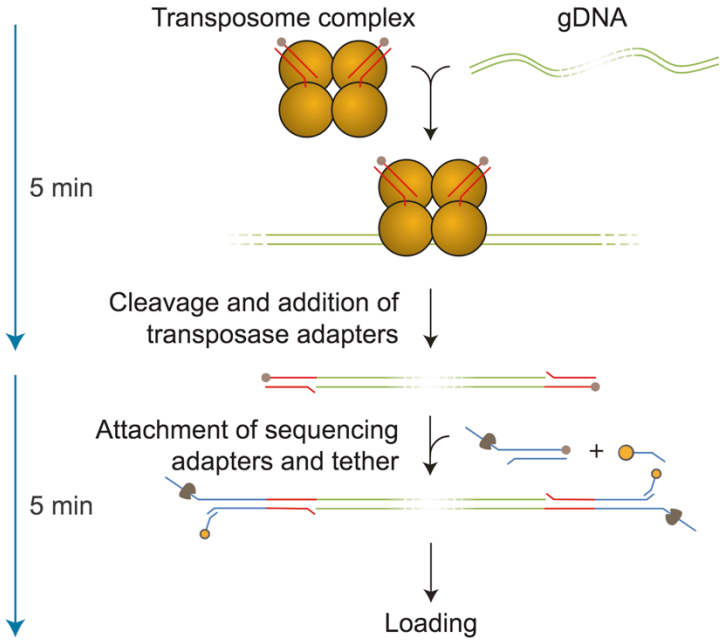
Nanopore DNA methylation detection



Methylation changes the detected current

Lazlo et al., 2013 PNAS (DOI: 10.1073/pnas.1310240110)

Nanopore library preparation – DNA



Rapid Barcoding Kit protocol

- Input: 200ng HMW DNA
- Typical output:
  - 1-2 Gb in 6 hrs
  - 4-8 Gb in 48 hrs
- Enzymatic Shearing of DNA
  → 40-60 % GC required

A very quick library preparation is possible

Nanopore output in “experienced hands”

Sequencing E.coli K-12 MG1655

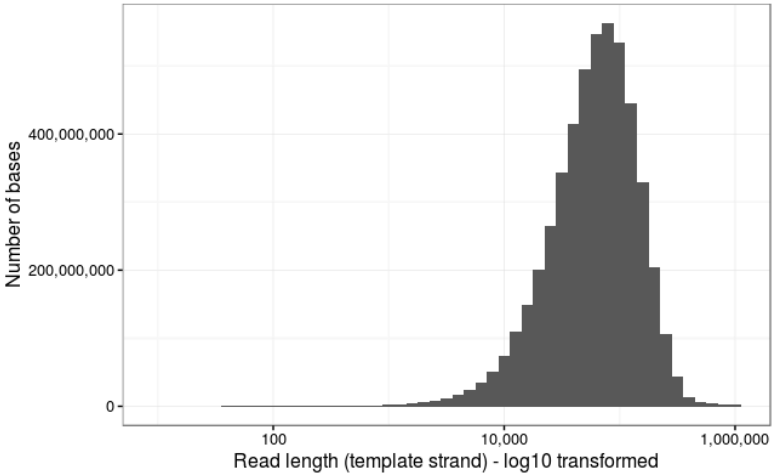
minION output

Total bases: 5.014.576.373 (5Gb)

Number of reads: 150.604

N50: 63.747

Mean: 33.296,44

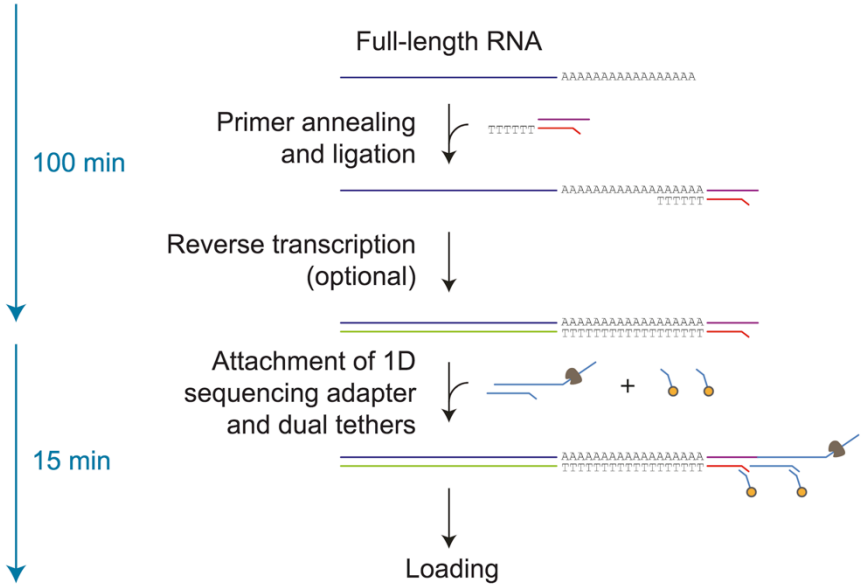


Longest alignable sequence: 778.217 bp (2017)

Possible due to very careful phenol / chloroform extractions with very pure DNA (260/280 ≈ 2.0).!!!

http://lab.loman.net/2017/03/09/ultrareads-for-nanopore/

Nanopore library preparation – RNA



Direct RNA sequencing

- Poly-A tail needed
- Optional reverse transcriptase to make cDNA → improves output
- Input : 500 ng RNA
- Typical output:
  - < 1 Gb in 6 hrs
  - 1-4 Gb in 48 hrs

RNA is very easily degraded.

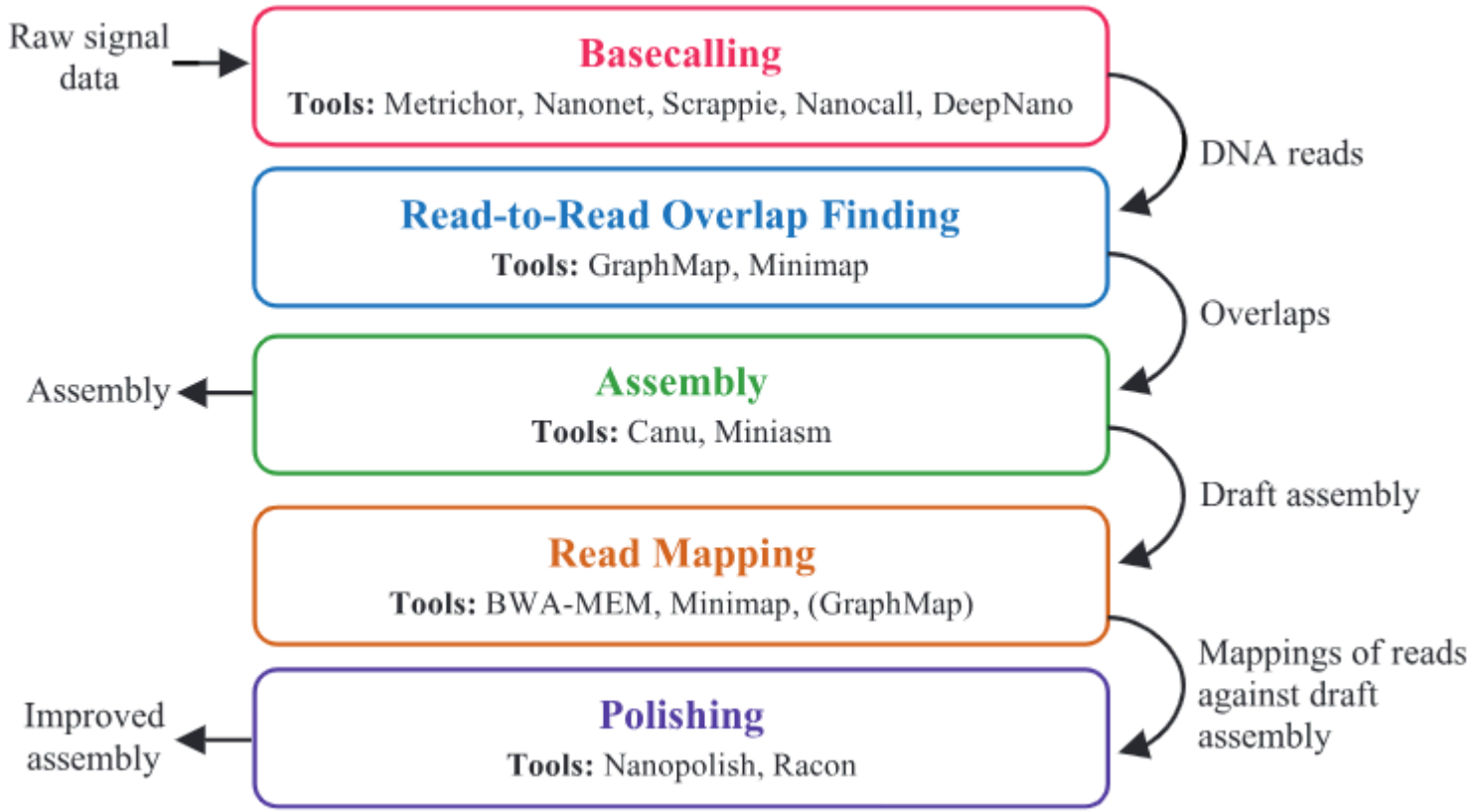
With this “quick” protocol direct sequencing is possible !

MinION Applications

- *De novo* shotgun sequencing (pcr / primer free sequencing)
  - Especially good for repetitive regions
  - Finishing Prokaryote / Eukaryote genomes
  - Detection of structural genome variation (indels)
- Amplicon sequencing
  - Prokaryotes / Eukaryotes: 16S rRNA / 18S rRNA
  - Fungi: ITS-1
  - Animal barcoding: CO1
- Shotgun metagenomics
- Transcriptomics / Direct RNA sequencing

  → Detection of RNA isoforms
- Epigenome (methylation) sequencing

De novo genome assembly



Senol Cali et al., Briefing in Bioinformatics 2018 (DOI: 10.1093/bib/bby017)

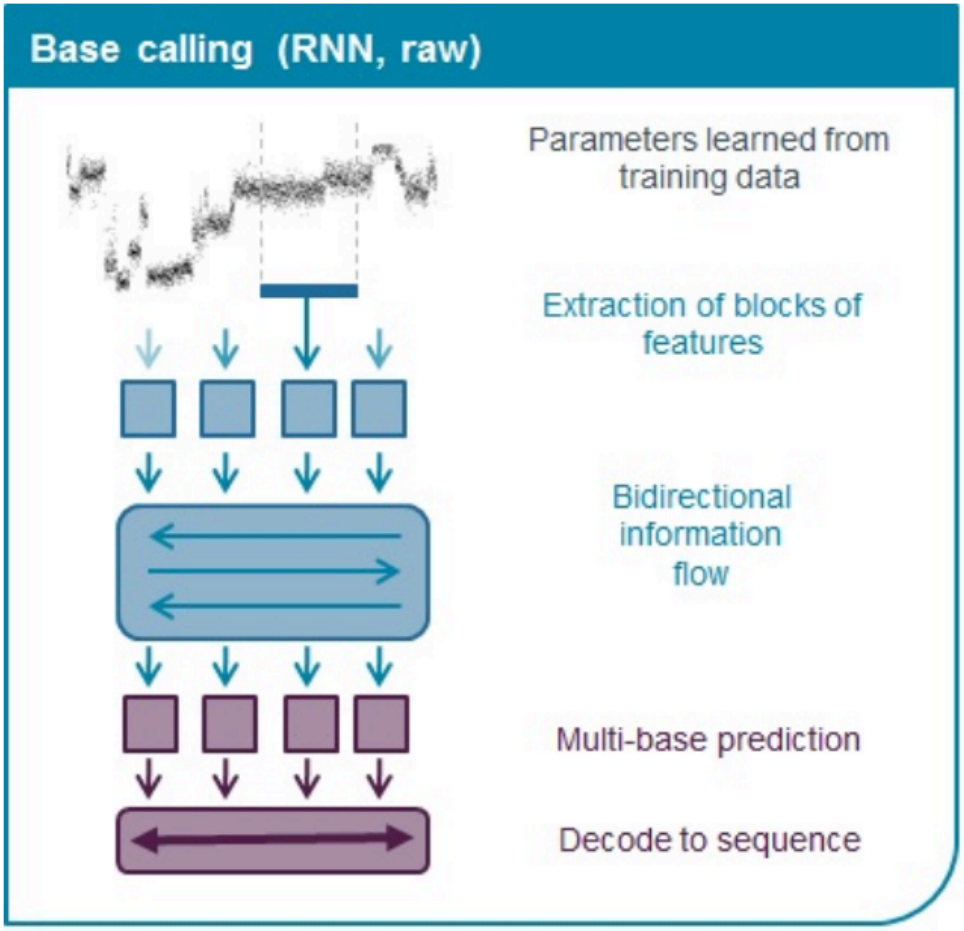
Basecalling software

Many options available:

- Nanopore provides several basecallers
  - MINknow (Included in the sequencing software)
  - Albacore
  - Guppy
  - Scrappie
  - Nanonet
- Other groups have also made basecallers for the nanopore machines:
  - Metrichor (In the cloud basecaller, part of minION workflow)
  - Chiron
  - DeepNano
  - etc

A nice comparison @: https://github.com/rrwick/Basecalling-comparison

Nanopore basecalling

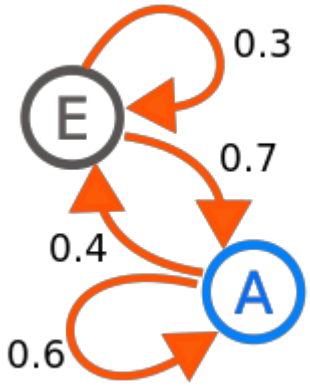


Original basecallers used Hiden Markov Models

Latest basecallers use Recurrent Neural Network (RRN)

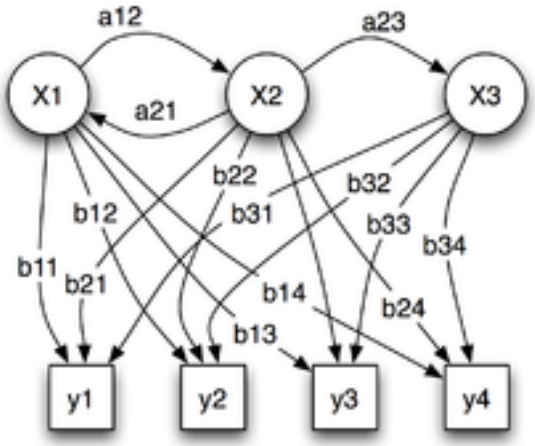
Hiden Markov Models

A **Markov chain** is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event".



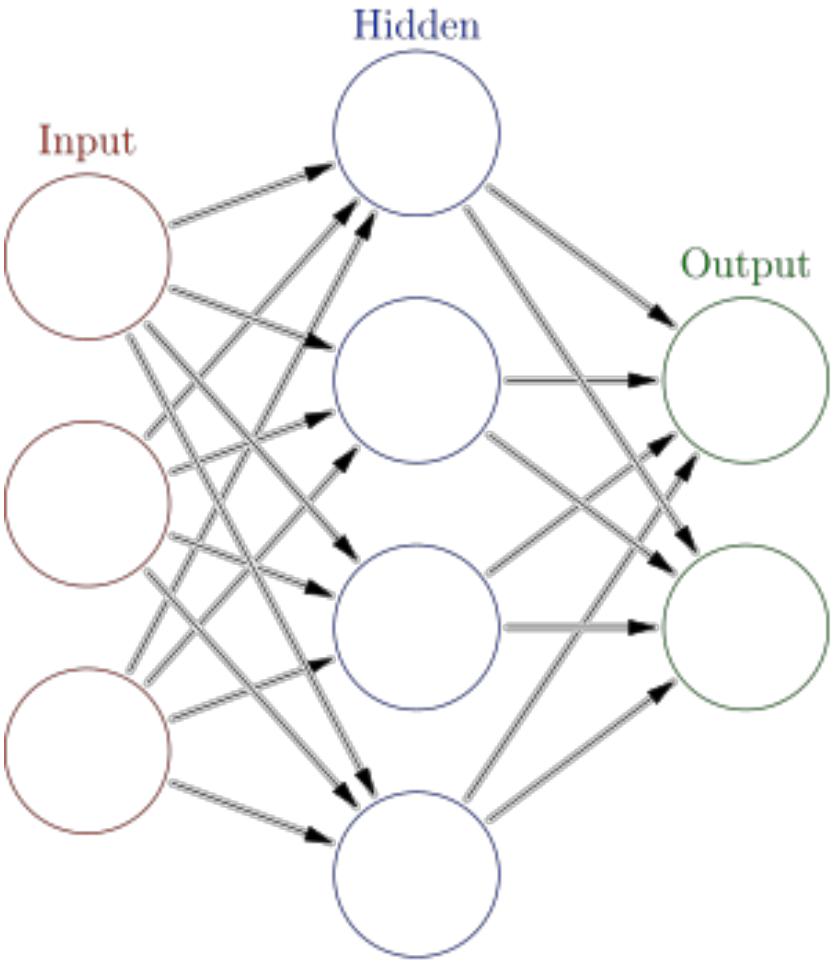
Hidden Markov Model (**HMM**) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (i.e. *hidden*) states.

We assume that the data observed (Y) is not the actual state of the model, but is instead generated by the underlying *hidden* (the H in HMM) states (X).



Wikipedia.org

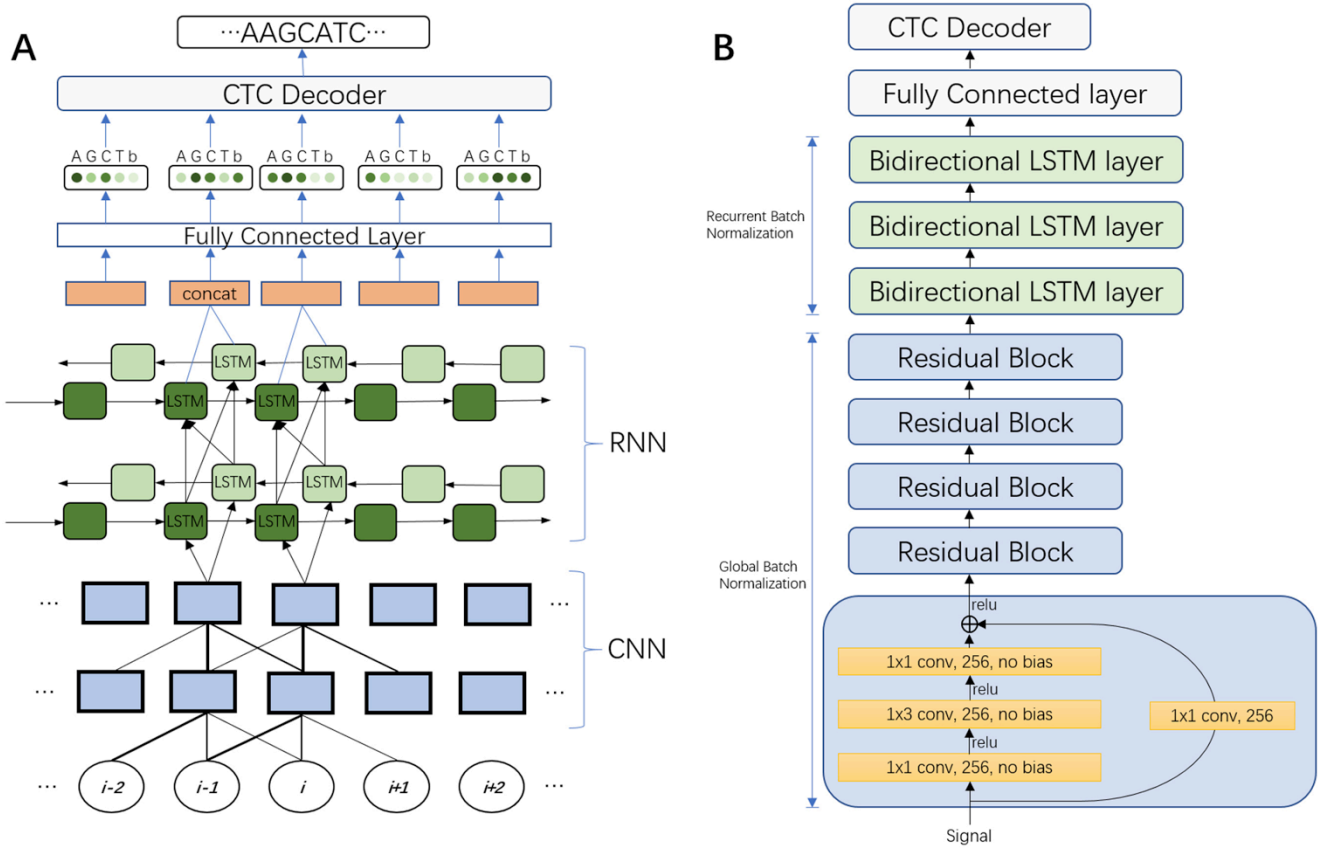
(Artificial) Neural Networks



Learning is achieved by changing the weights from the edges (arrows).

Wikipedia.org

Basecalling software - Chiron



A combined convolutional neural network and a Recurrent Neural Network

Genome assembly with nanopore data

Table 1. Summary of comparisons between long read assemblers. (**A**) Selected metrics for three benchmarking efforts on MinION reads, including chemistries used in the respective studies. Bold values denote the best score per metric. (**B**) Short descriptions and reference papers for all assemblers discussed in this paper. [1]: reads were corrected by Canu prior to assembly.

A	Judge *et al.*[41]			Istace *et al.*[40]			Giordano *et al.*[39]		
	subs/ kbase	indels/ kbase	N50 (Mbase)	subs/ kbase	indels/ kbase	N50 (Mbase)	subs/ kbase	indels/ kbase	N50 (Mbase)
PBcR	1.0	12.2	1.20				0.2	17	0.616
Canu	**0.3**	**7.8**	**2.80**	**0.105**	**10.0**	0.610	**0.1**	17	0.698
SMARTdenovo				0.580	11.1	0.783	0.3	**14**	0.625
Minimap & miniasm	6.7	18.6	6.60	0.207[1]	13.5[1]	0.736[1]	34	67	0.739
ABruijn				0.130	10.1	**0.816**	0.1	15	**0.769**
Chemistry		MAP006			MAP005/MAP006			MAP006/007	
Read type		2D			2D			2D	
Pore		R7.3			R7.3			R7.3/R9	
Basecaller		EPI2ME			EPI2ME			EPI2ME	
Organism		*Enterobacter kobei*			*S. cerevisiae*			*S. cerevisiae*	

B	Description	Ref.
PBcR	Celera OLC assembler adapted for long error-prone reads.	42
Canu	The more accurate successor of PBcR.	43
SMARTdenovo	Fast and reasonably accurate assembler without prior error correction step.	Github
Minimap & miniasm	Fast assembly pipeline without error correction and consensus steps.	44
ABruijn	DBG assembler that fuses unique strings prior to assembly, produces highly contiguous assemblies.	45
TULIP	uses seed extension principle to efficiently assemble large genomes.	25
HINGE	Assesses coverage of low complexity regions prior to assembly and processes them more efficiently.	46

De Lannoy et al., 2017 F1000 research (DOI: 10.12688/f1000research.12012.1)

Canu assembler

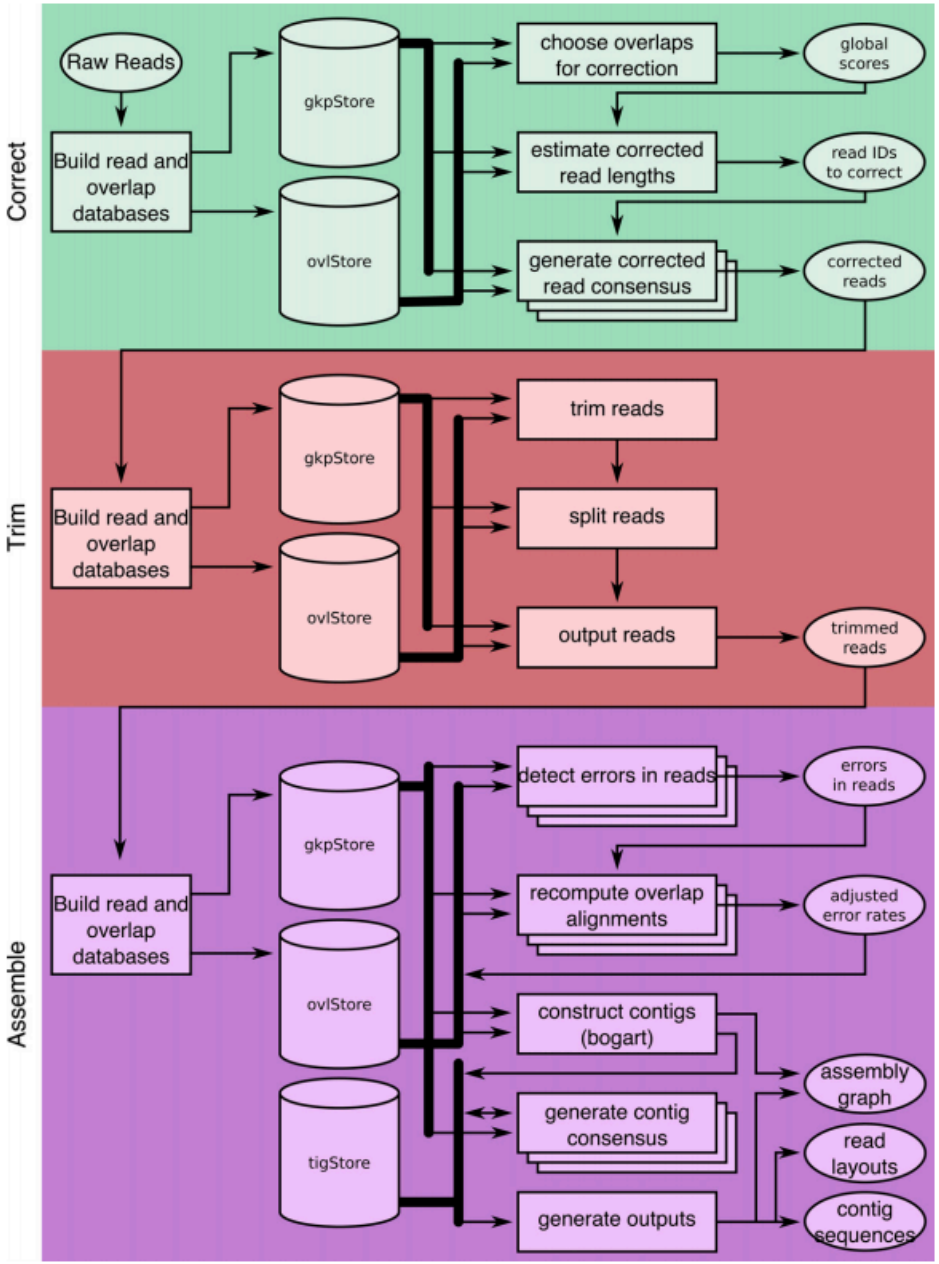
Canu Assembly pipeline

1. Error correction
2. Trimming
3. Assembly

gkpStore: reads database

ovlStora: overlap database

tigStore: contigs database

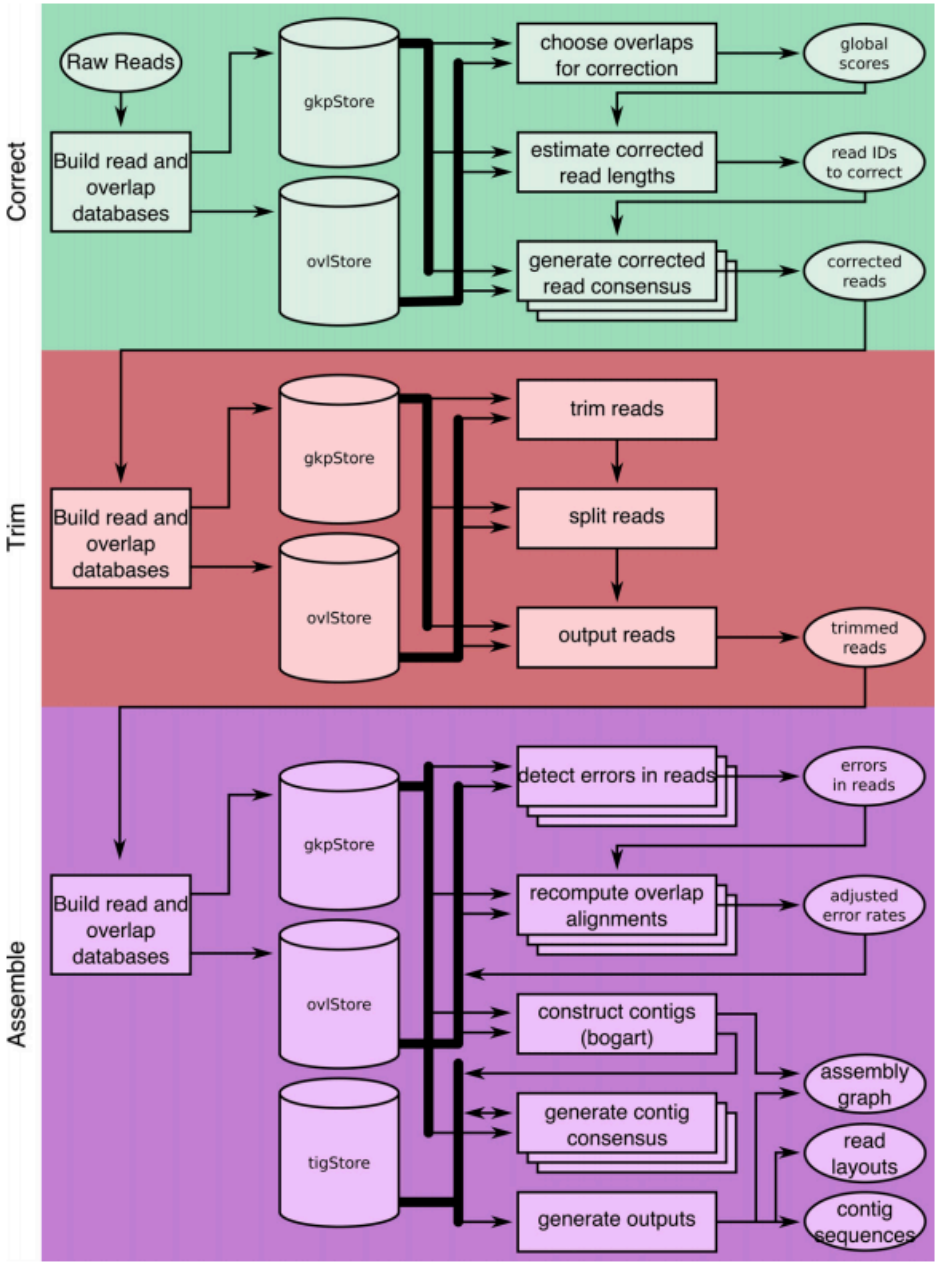


Koren et al., 2017 Genome Research (DOI: 10.1101/gr.215087.116)

Error correction

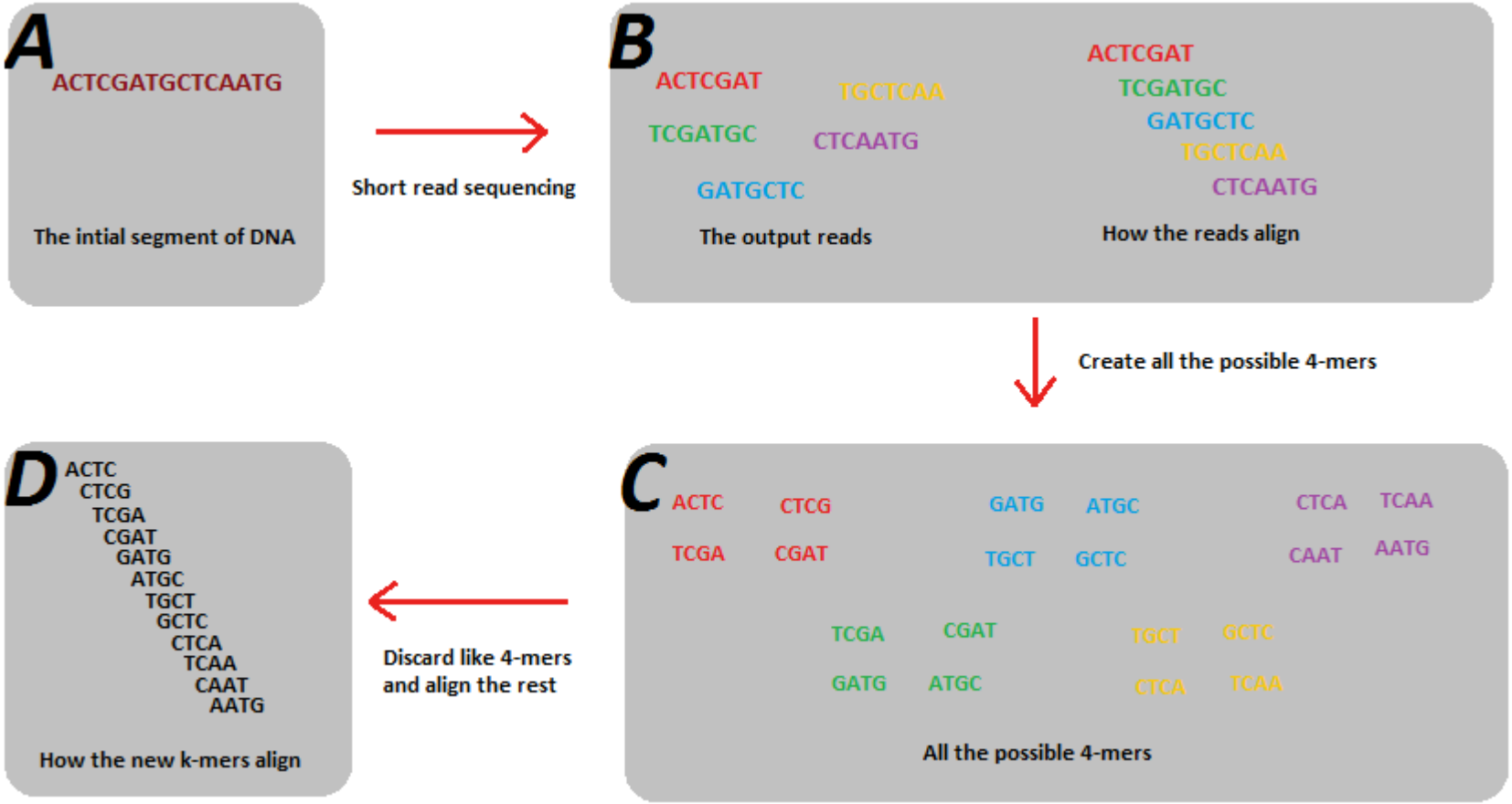
1. Reads split into **kmers**
2. Kmers used to identify overlap
3. Correct reads using overlap

Corrected reads are trimmed



Koren et al., 2017 Genome Research (DOI: 10.1101/gr.215087.116)

Kmers

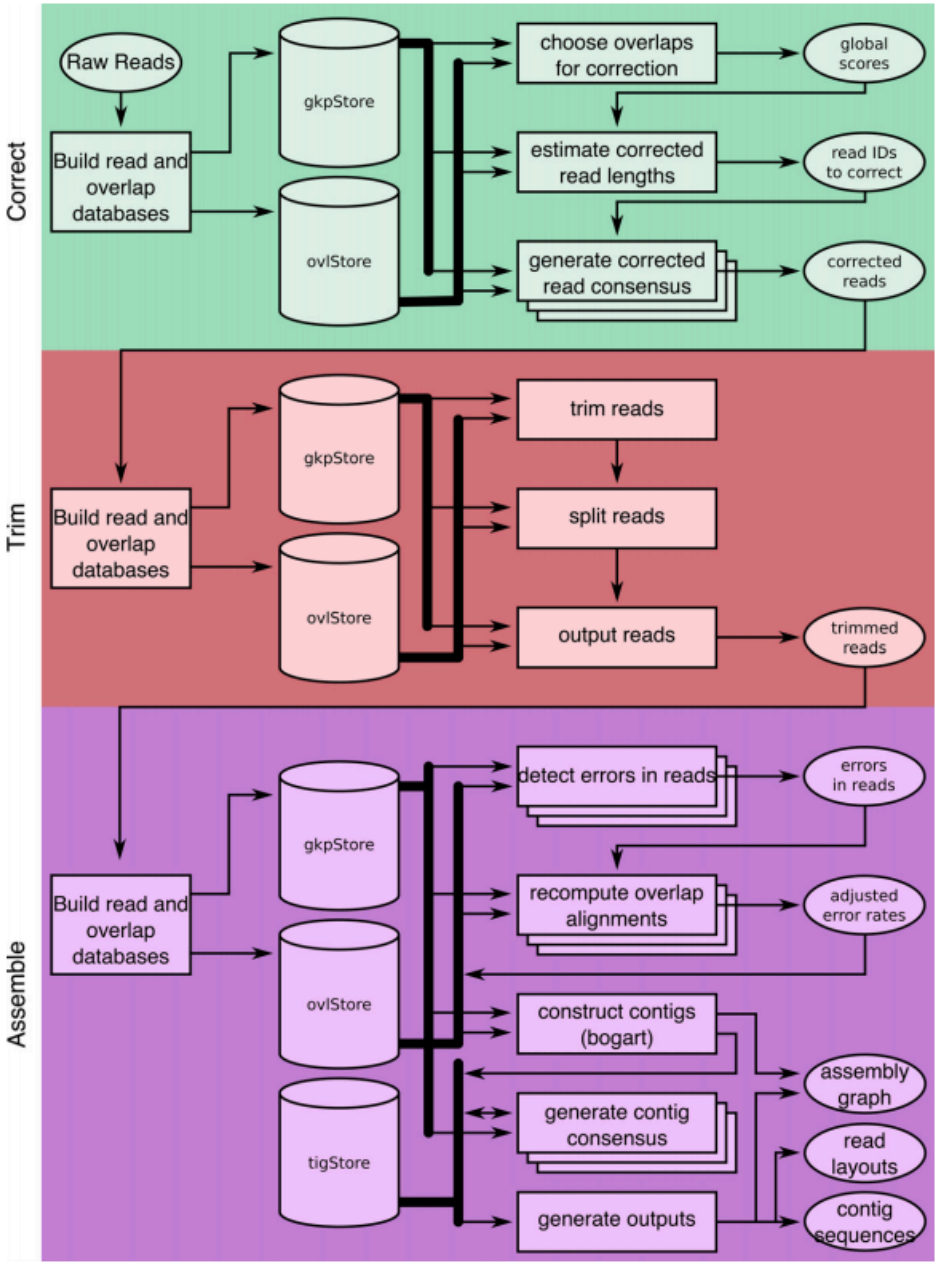


Wikipedia.org

Error correction

1. Reads split into **kmers**
2. Kmers used to identify overlap
3. Correct reads using overlap

Corrected reads are trimmed



Koren et al., 2017 Genome Research (DOI: 10.1101/gr.215087.116)

Canu assembles *E. coli* genome

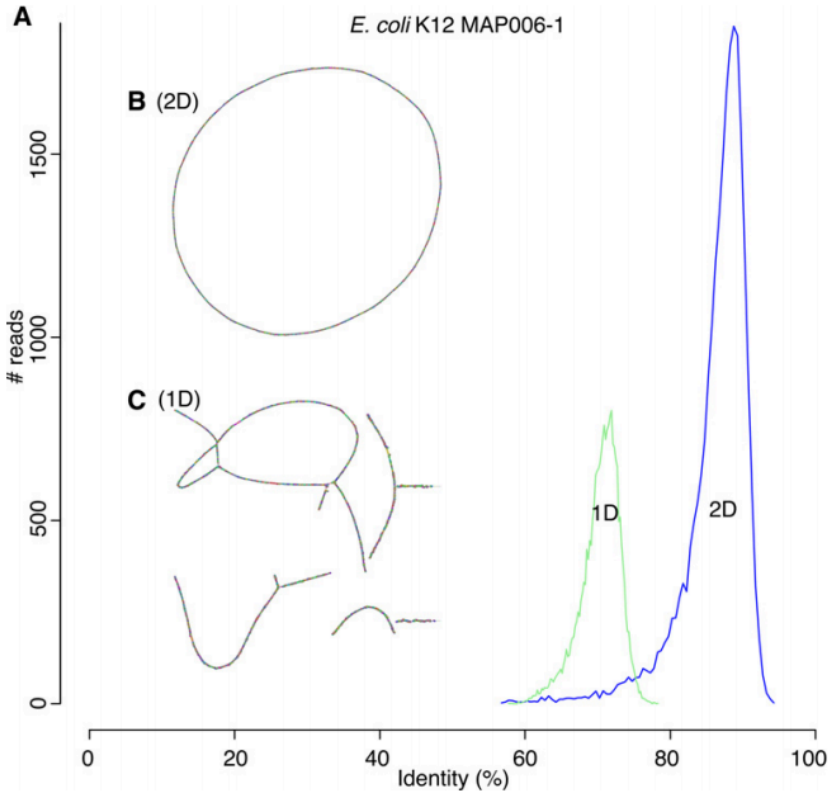


Figure 5. Canu can assemble both 1D and 2D Nanopore *Escherhicia coli* reads. (*A*) A comparison of error rates for 1D and 2D read error rates versus the reference. Template 1D and 2D reads from the MAP006-1 *E. coli* data set were aligned independently to compute an identity for all reads with an alignment >90% of their length (95% of the 2D reads and 86% of the 1D reads had an alignment >90% of their length). The 2D sequences averaged 86% identity, and the 1D reads averaged 70% identity. (*B*) Bandage plot of the Canu BOG for the 2D data. The genome is in a single circle representing the full chromosome. (C) The corresponding plot for 1D data. While highly continuous, there are multiple components due to missed overlaps and unresolved repeats (due to the higher sequencing error rate).

Koren et al., 2017 Genome Research (DOI: 10.1101/gr.215087.116)

Minimap & miniasm

Minimap steps

1. Find overlap between reads using kmers
2. Identify regions of overlapping reads with good coverage (3 - 4 X)
3. Trimming of low coverage areas of reads (< 3 X)

Miniasm steps

1. Each mapped region is classified
2. Overlaps added to assembly graph
3. Clean up of asmembly graph
4. Produce sequence (unitig)

Note: NO error correction of reads.

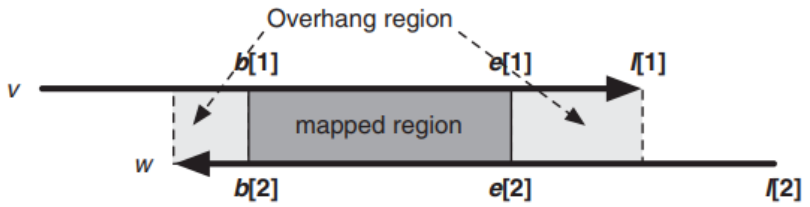


Fig. 1. Mapping between two reads. $b[1]$ and $e[1]$ are the 0-based starting and ending mapping coordinates of the first read v, respectively. $b[2]$ and $e[2]$ are the mapping coordinates of read w. Lightgray areas indicate overhang regions that should be mapped together if the overlap is real. If the overhang regions are small enough, the figure implies an edge $v \to w$ with approximate length $\ell(v \to w) = b[1] - b[2]$ and its complement edge $\bar{w} \to \bar{v}$ with $\ell(\bar{w} \to \bar{v}) = (l[2] - e[2]) - (l[1] - e[1])$

Li. 2016 Bioinformatics (DOI: 10.1093/bioinformatics/btw152)

Nanopolish

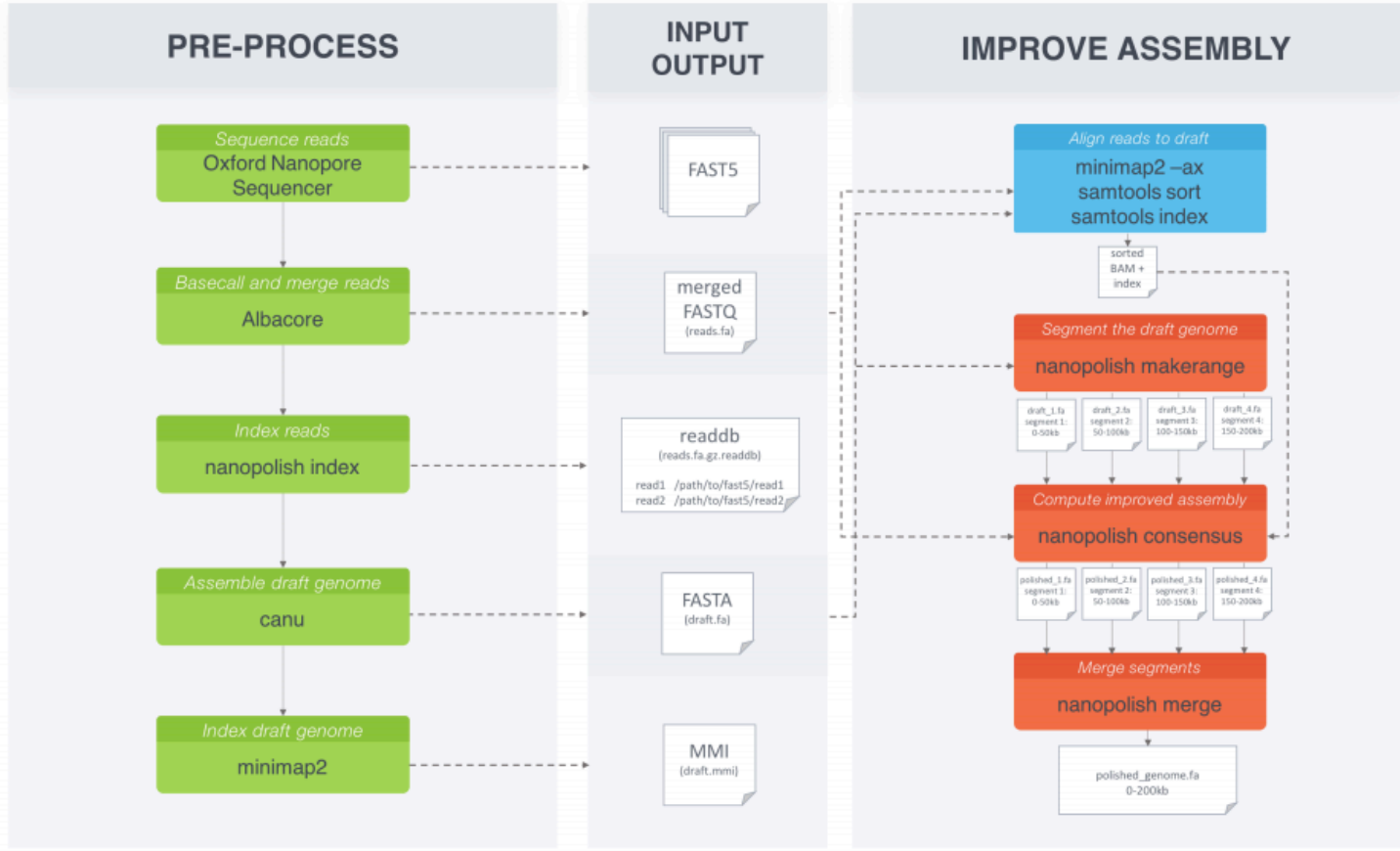
Nanopolish: Improve consensus sequence of assemblies

Options:

- Predict methylated bases
- detect SNPs and indels with respect to a reference genome
- calculate an improved consensus sequence for a draft genome assembly
- align signal-level events to k-mers of a reference genome
    - Align raw sequence data to deal with homopolymers and other hard to analyse sequences

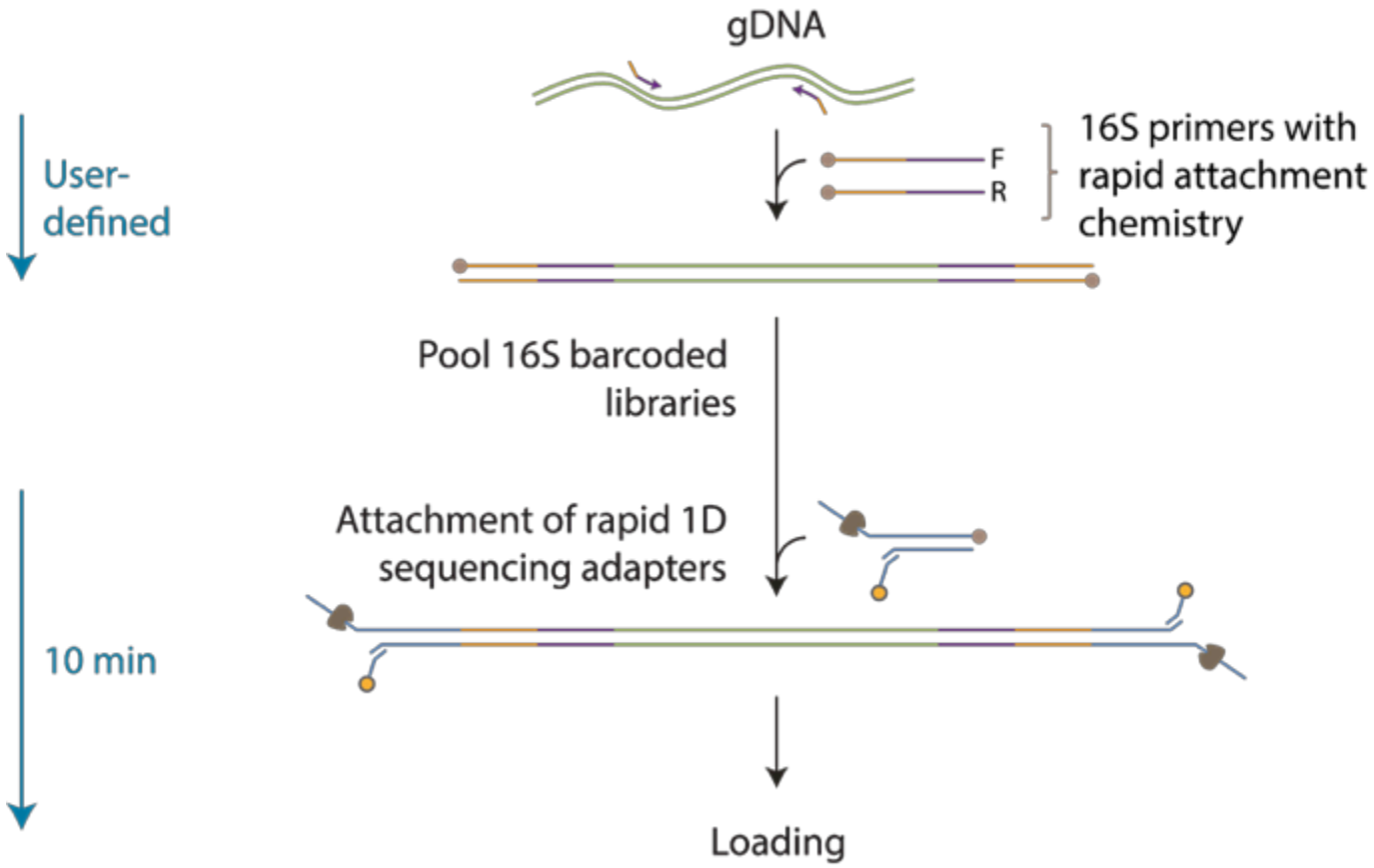
Loman et al., 2015 Nature Methods volume 12, pages 733–735

Nanopolish

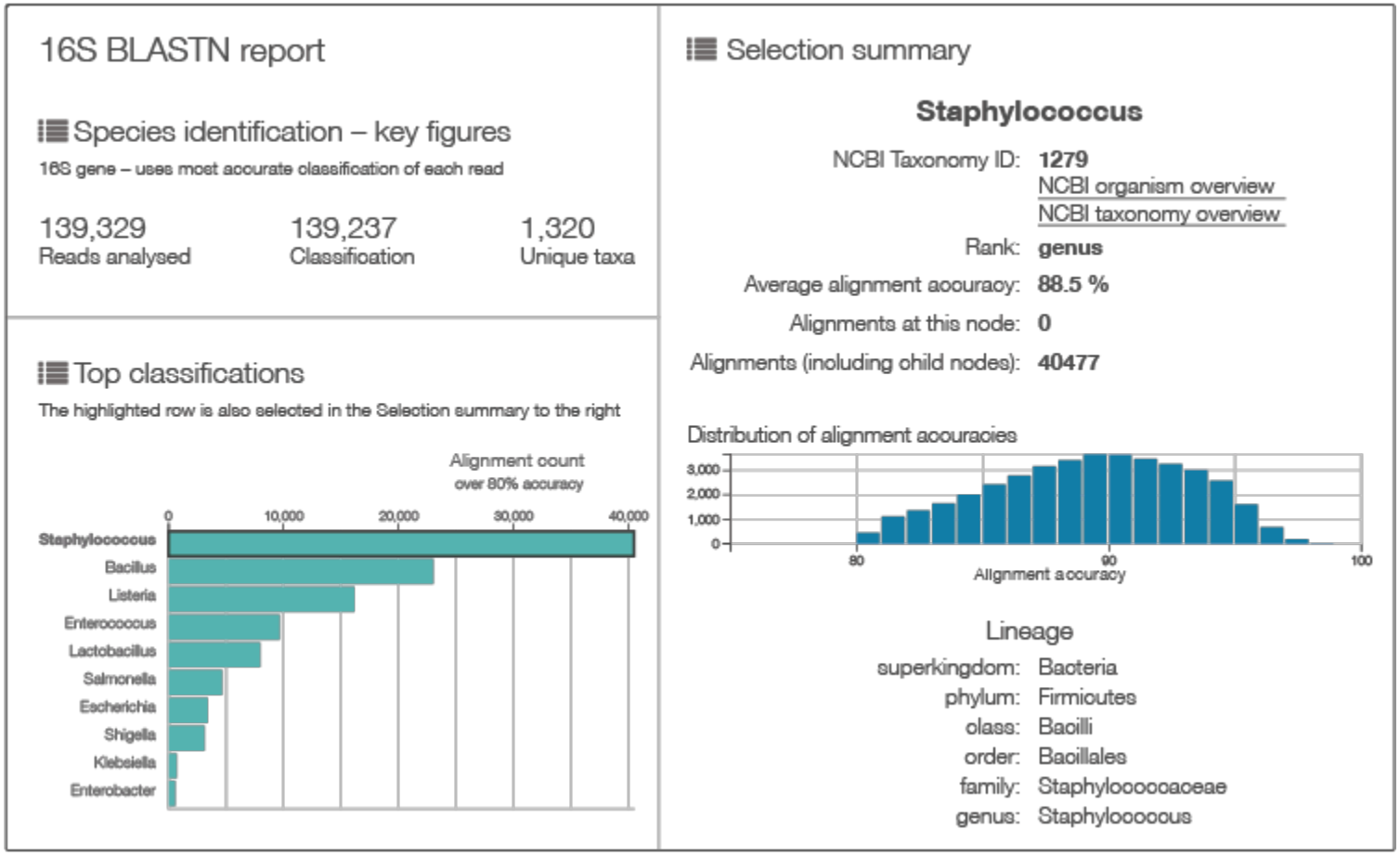


https://media.readthedocs.org/pdf/nanopolish/latest/nanopolish.pdf

Amplicon sequencing

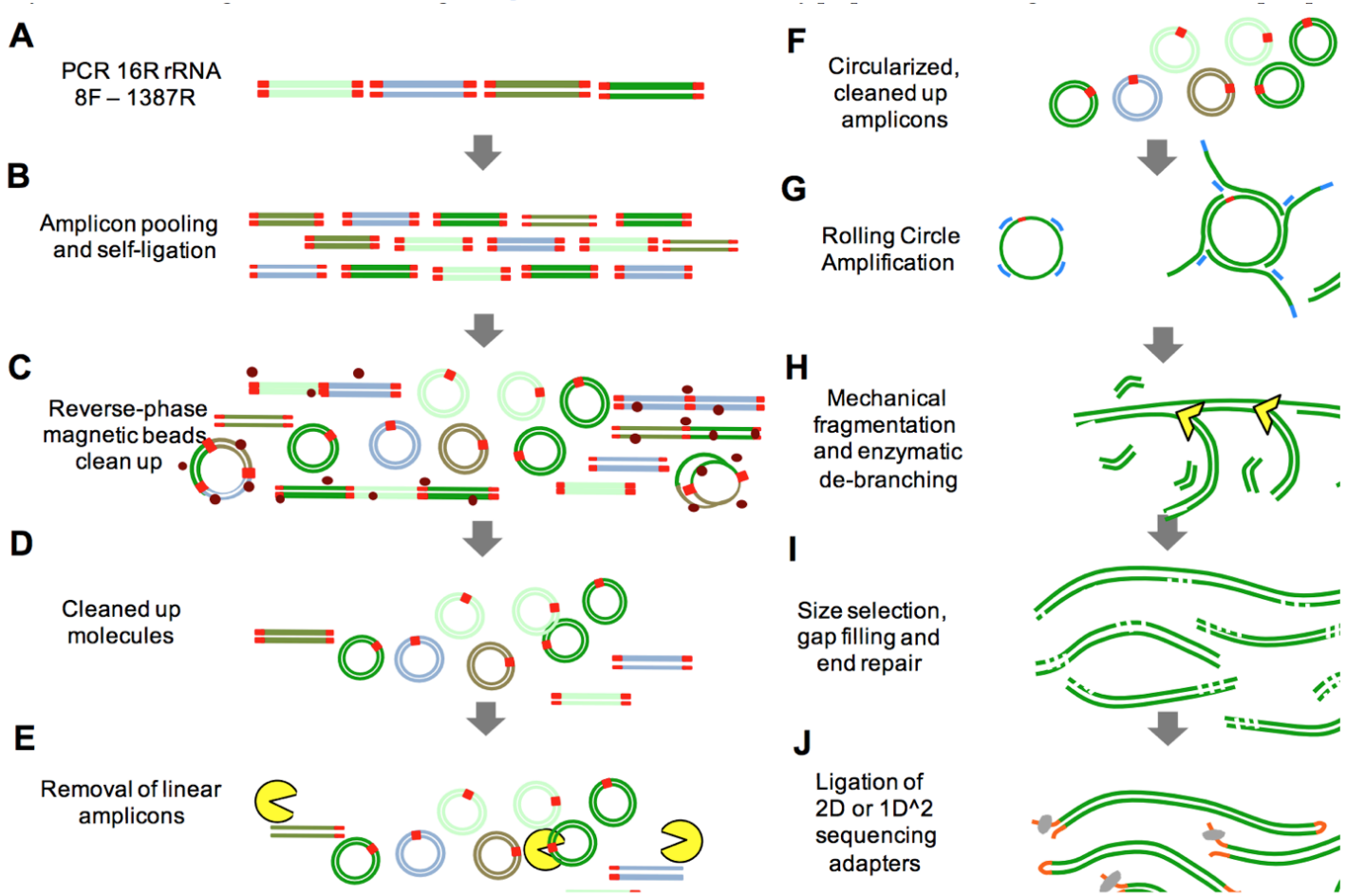


Amplicon sequencing



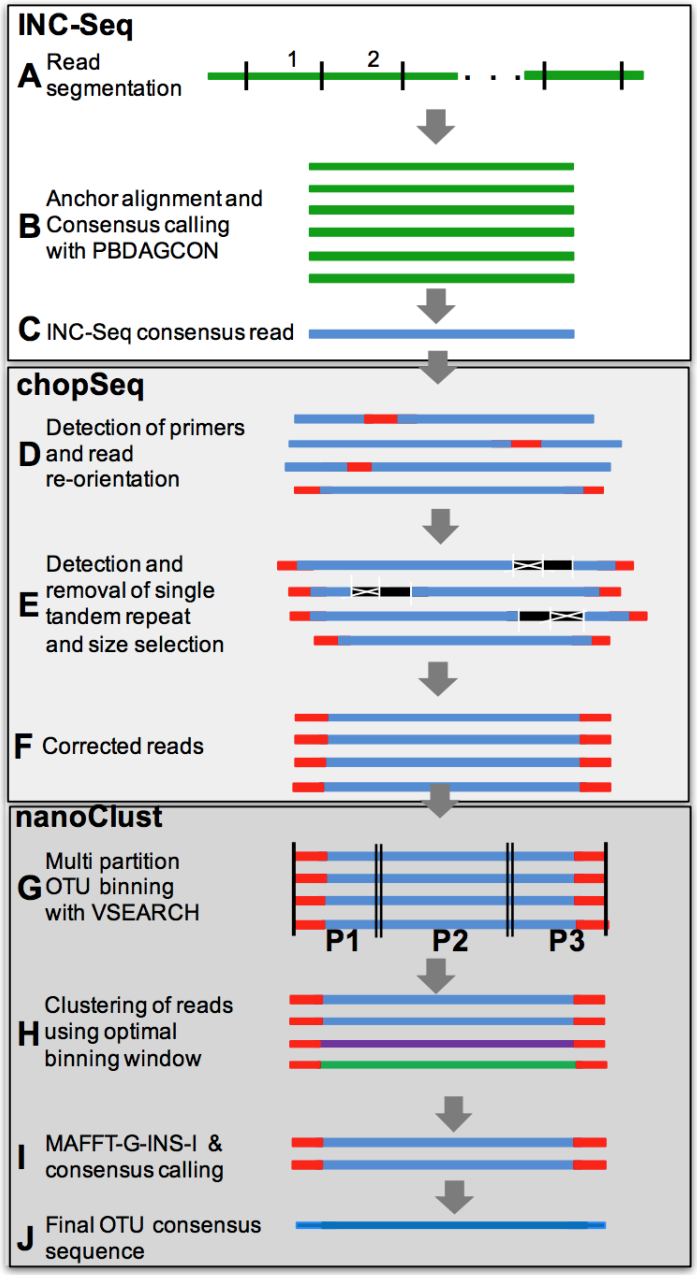
Accuracy is low

Amplicon sequencing



Intramolecular-ligated Nanopore Consensus Sequencing (INC-Seq)

Calus et al., 2018 BioRxiv (DOI: 10.1101/244517)c

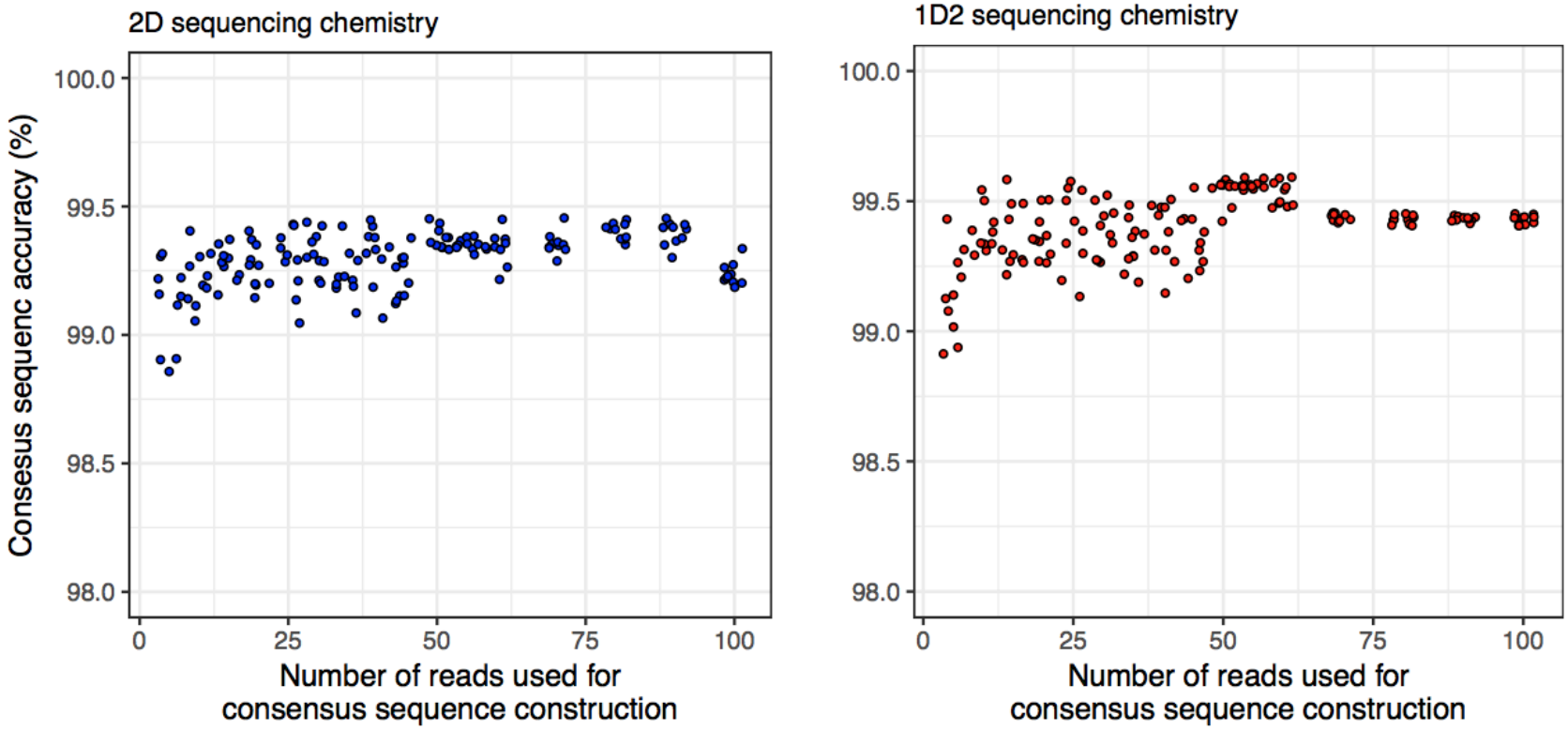


Amplicon sequencing

A pipeline generating consensus amplicons

Calus et al., 2018 BioRxiv (DOI: 10.1101/244517)c

Amplicon sequencing



A minimum of 15 reads is needed to produce high Quality amplicons

Calus et al., 2018 BioRxiv (DOI: 10.1101/244517)c

Nanopore sequencers

Machine	SmidgION	MinION	GridION 5X	promethION
Flowcells	1	1	5	48
Data output	Not yet specified	10-20 Gb	50-100 Gb	> Tb
Pores	Not yet specified	800	5 X 800	3000 (Total 144000)
Application	Field-based	Field / lab based	Sequencing service	Sequencing service

A short comparison

	Illumina	PacBio	minION
Output (Gb)	7.5 – 6000	5-8	10-20
Reads (million)	25 – 20-000	0.15 - 1	≈ 0.15
Read length	150 – 300 bp	0 - 70 Kbp	0 - 800 Kbp
Pros	• Many reads • High quality • Tolerant for poor input material	• Long reads • Improve genome assemblies	• High mobility • Long reads • Improve genome assemblies
Cons	• Fragmented genome assemblies	• High quality input needed • expensive	• High quality input needed • Flowcell has limited shelf life

Experimental design important to decide which platform to use.

The End

A few papers:

The long reads ahead: *de novo* genome assembly using the MinION

- https://f1000research.com/articles/6-1083/v2

Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis

- https://doi.org/10.1186/s13073-015-0220-9

NanoAmpli-Seq: A workflow for amplicon sequencing for mixed microbial communities on the nanopore sequencing platform

- https://www.biorxiv.org/content/early/2018/07/04/244517