Introduction to Nanopore sequencing

Thomas Haverkamp (Norwegian veterinary institute)

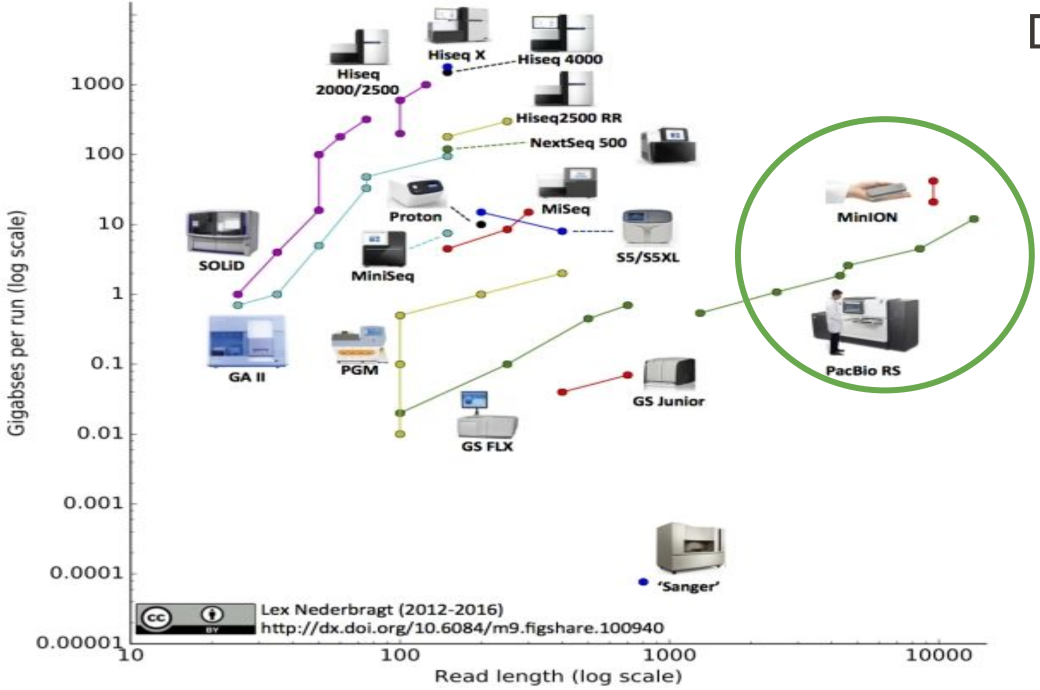
@Thomieh

Veterinærinstituttet
Norwegian Veterinary Institute

Outline

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

Rapid development in instrumentation



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

Drastic increase in both

- Read length
- Amount of sequence / run

- Single molecule sequencer
- Long read sequencers

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

Oxford Nanopore sequencers

Machine	SmidgION	MinION	GridION 5X	promethION
Flowcells	1	1	5	48
Data output	Not yet specified	10-50 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

The minION nanopore sequencer

minION sequencer & flowcell

The minION nanopore flowcell

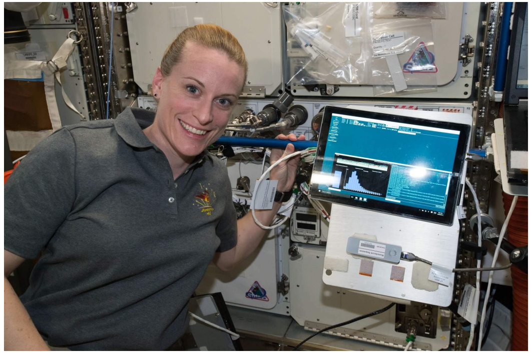
minION sequencer & flowcell

Flongle flowcell - 2.8 Gb

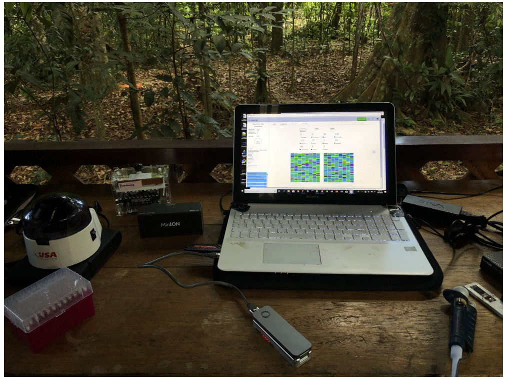
Normal flowcell - 50 Gb

Out of the lab usage...

Antarctica



ISS spacestation



The Jungle

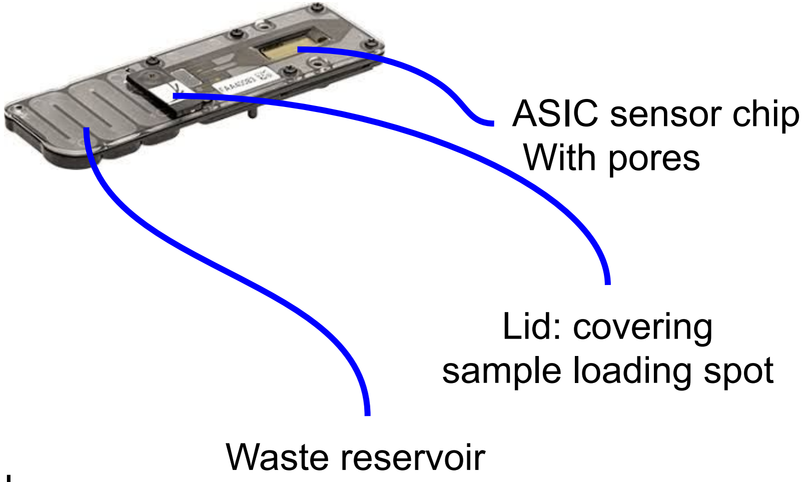
You do need lab equipment to process your samples !!!

The minION flowcell

Specifications:

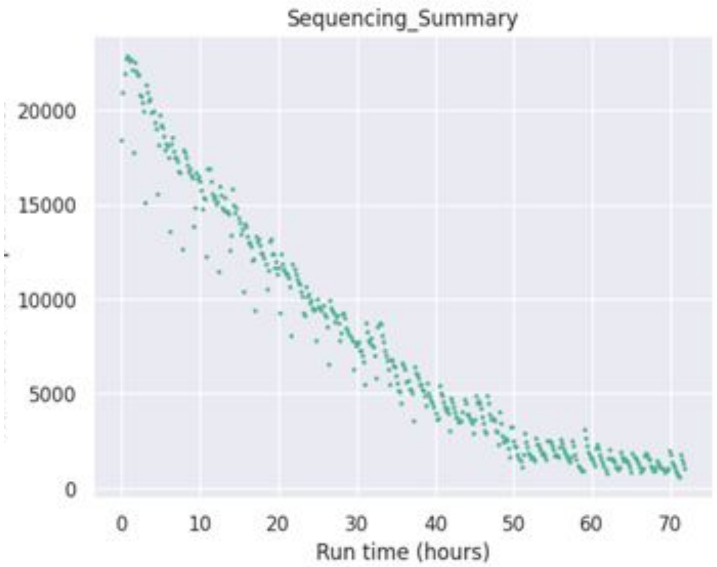
- 512 pores (Guaranteed)
- Needs to be stored at 2-8 ^{O}C
- Pores deteriorate over time - Fresh is best
- Longest single read of a single molecule sequence

  '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

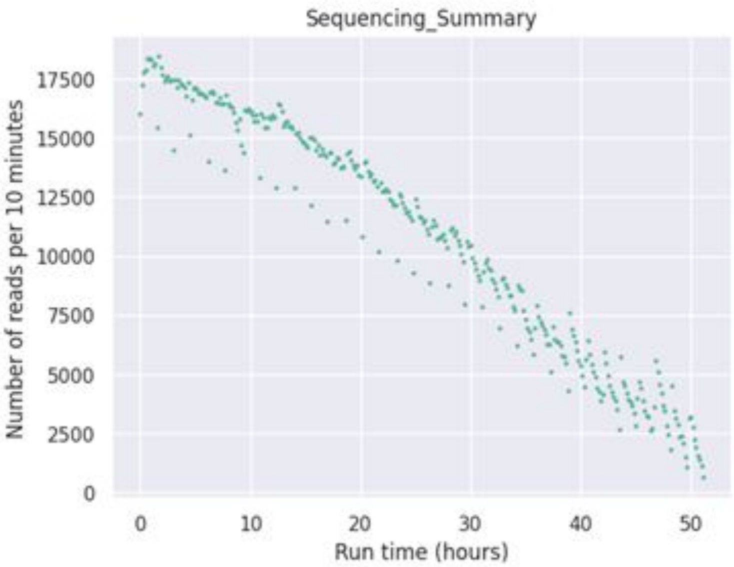


The minION output - reads

run_20220831

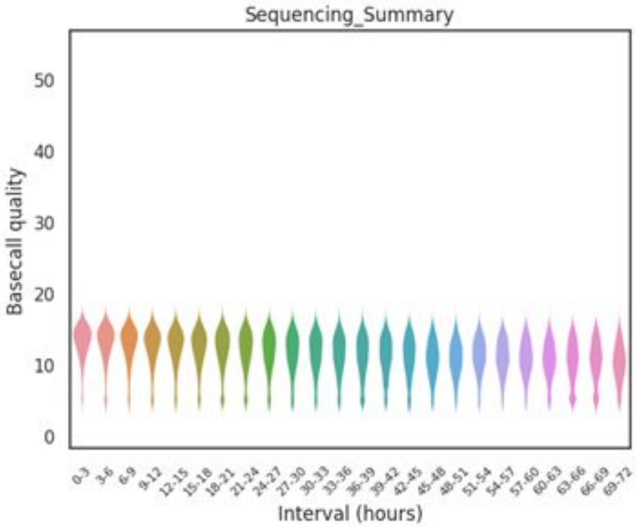


run_20220920



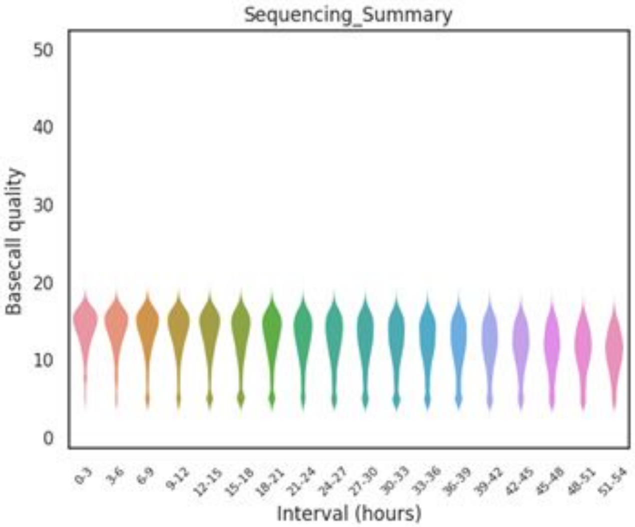
The minION output - quality

run_20220831



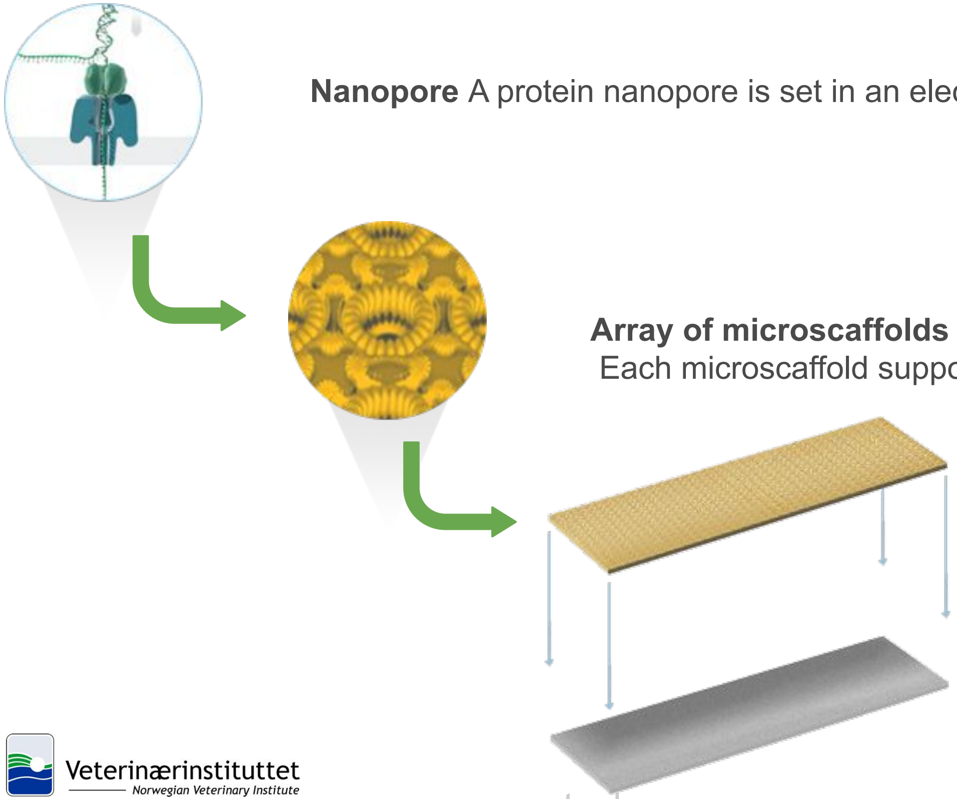
Highest quality
54.2 (696)

run_20220920



Highest quality
49.9 (230)

The nanopore sensor chip



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

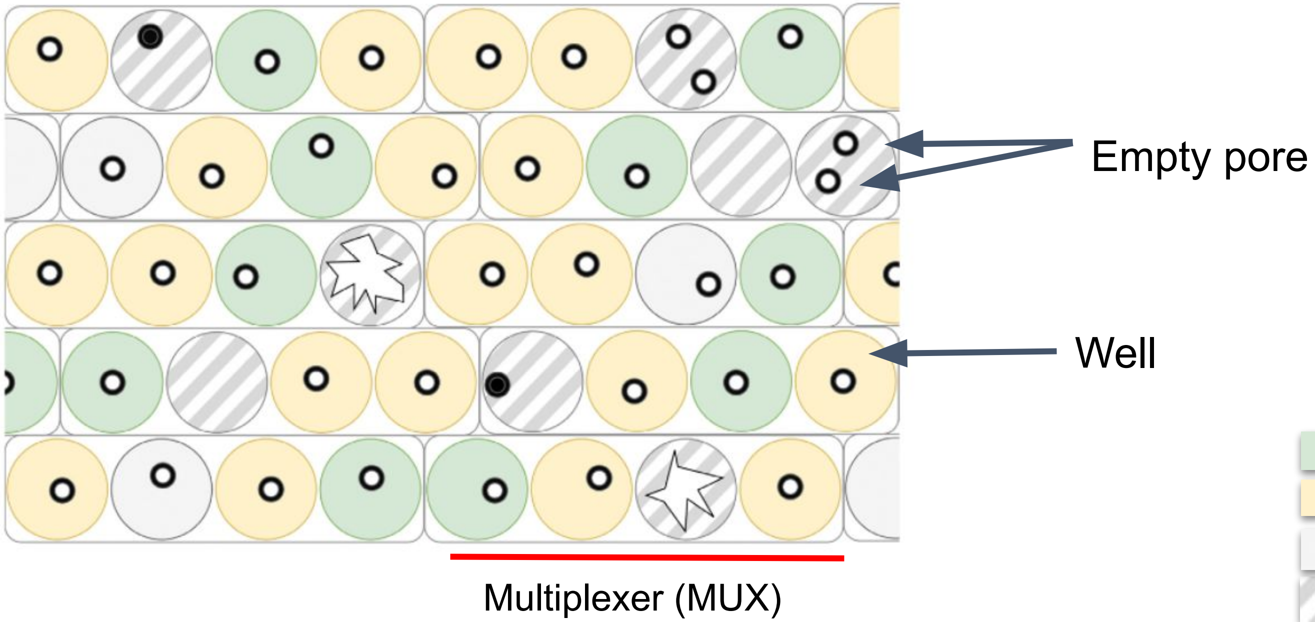
Array of microscaffolds
Each microscaffold supports a membrane and embedded nanopore.

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.

www.nanoporetech.com

The flowcell layout



Well type

🟩 Best

🟨 2 and 3rd place

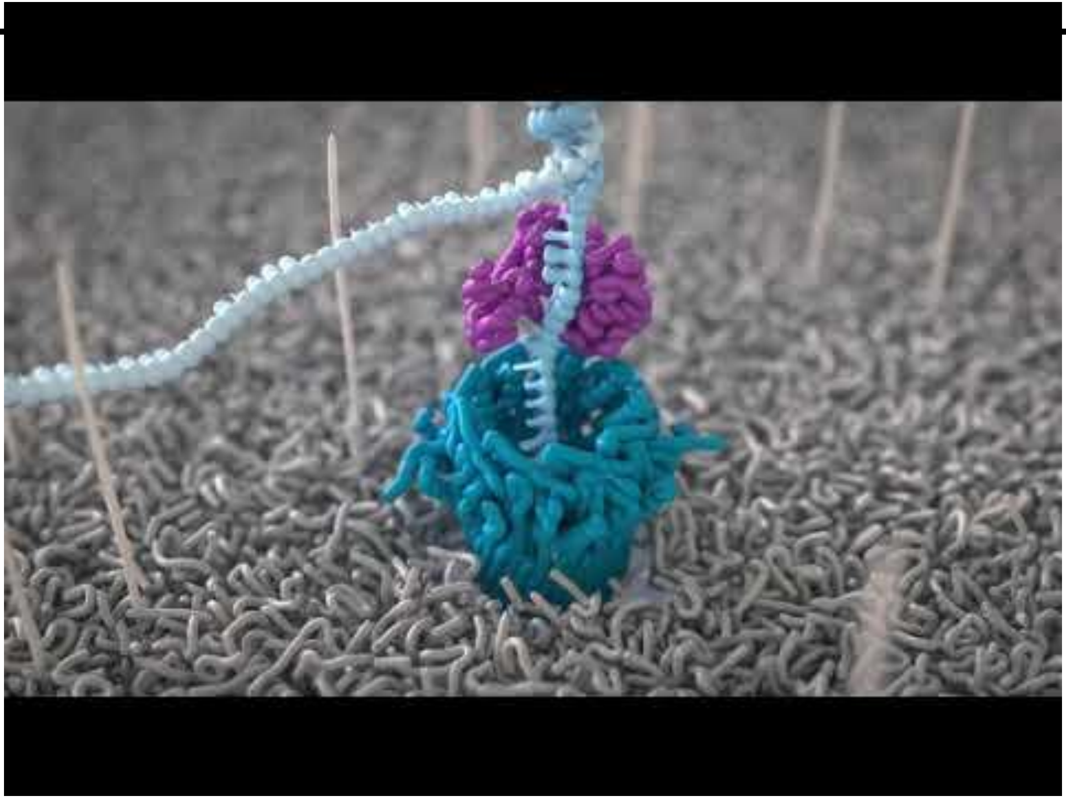
⬜ 4rd = not used

▨ Not usable

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

www.nanoporetech.com

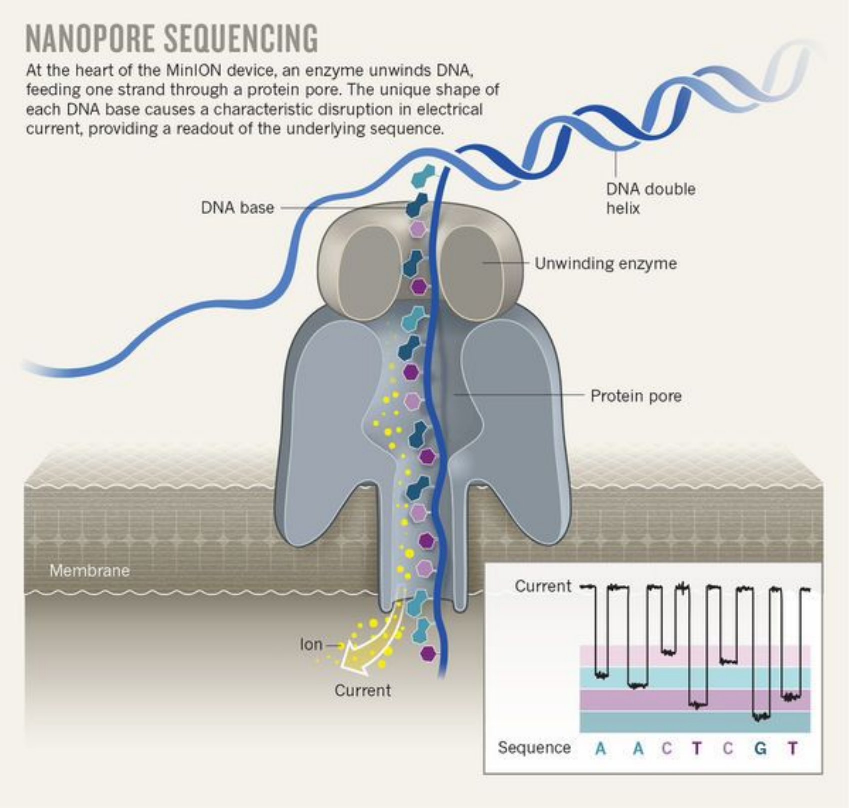
Nanopore sequencing explained



www.nanoporetech.com

Veterinærinstituttet
Norwegian Veterinary Institute

Nanopore sequencing explained



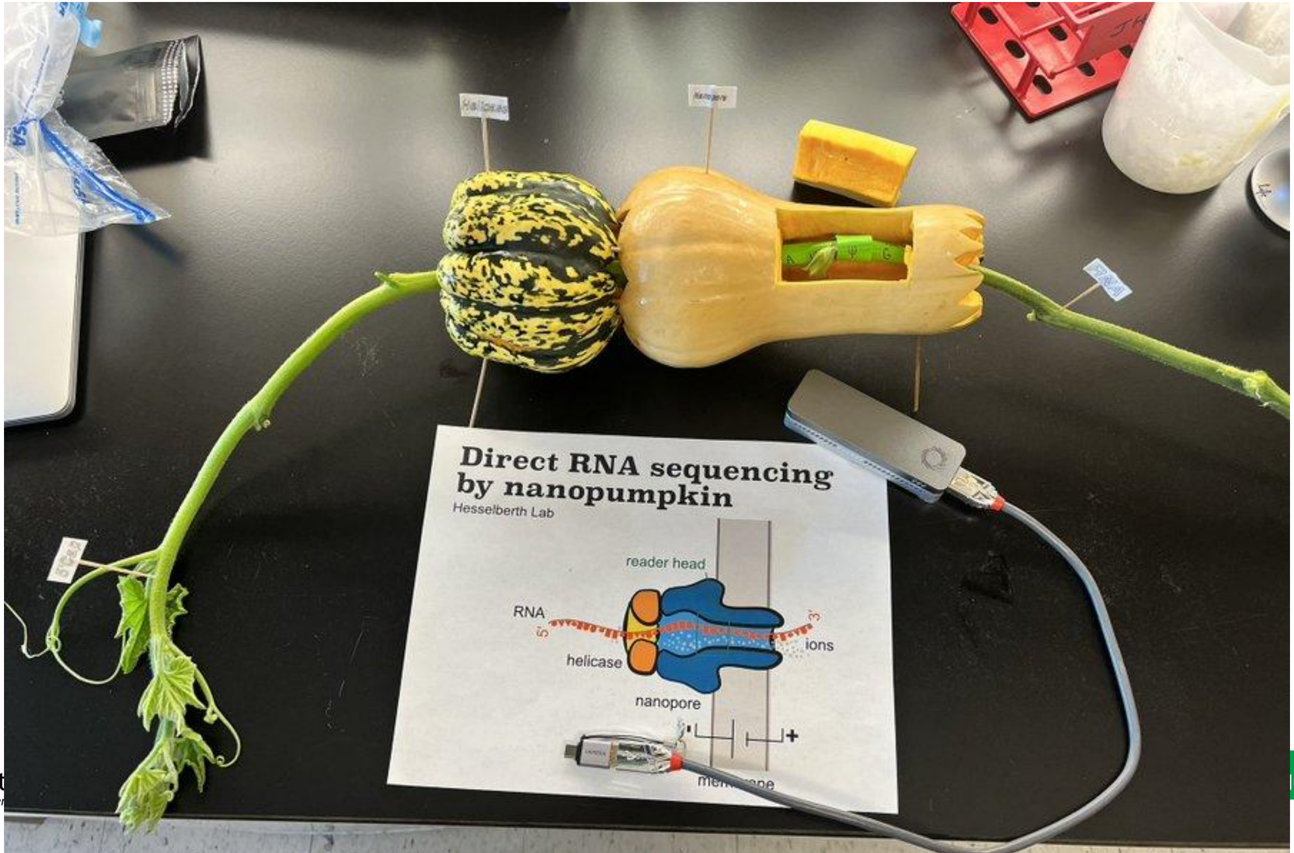
NANOPORE SEQUENCING

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.

www.nanoporetech.com

DNA

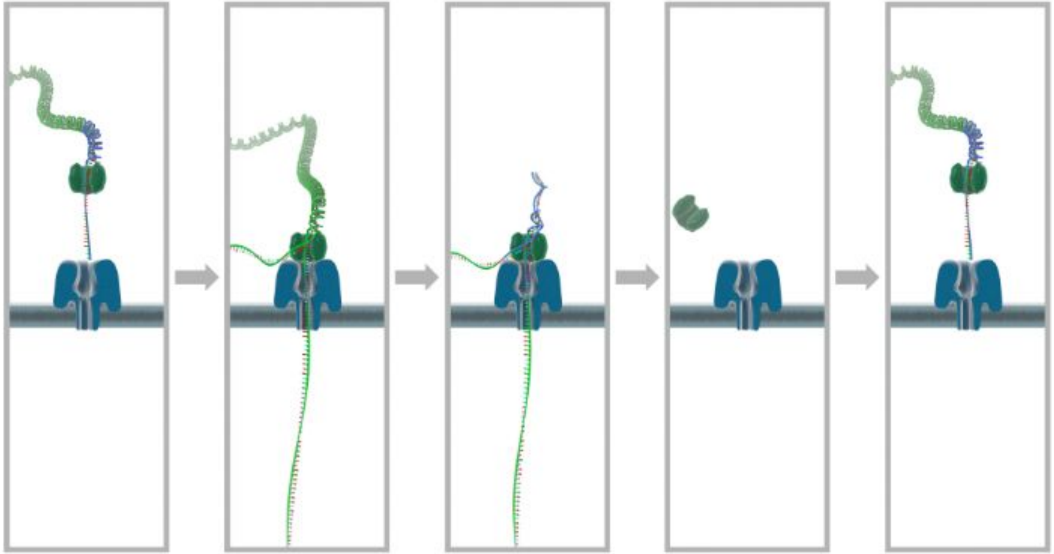
Nanopore sequencing explained



RNA

by
Jay Hesselberth

Nanopore sequencing

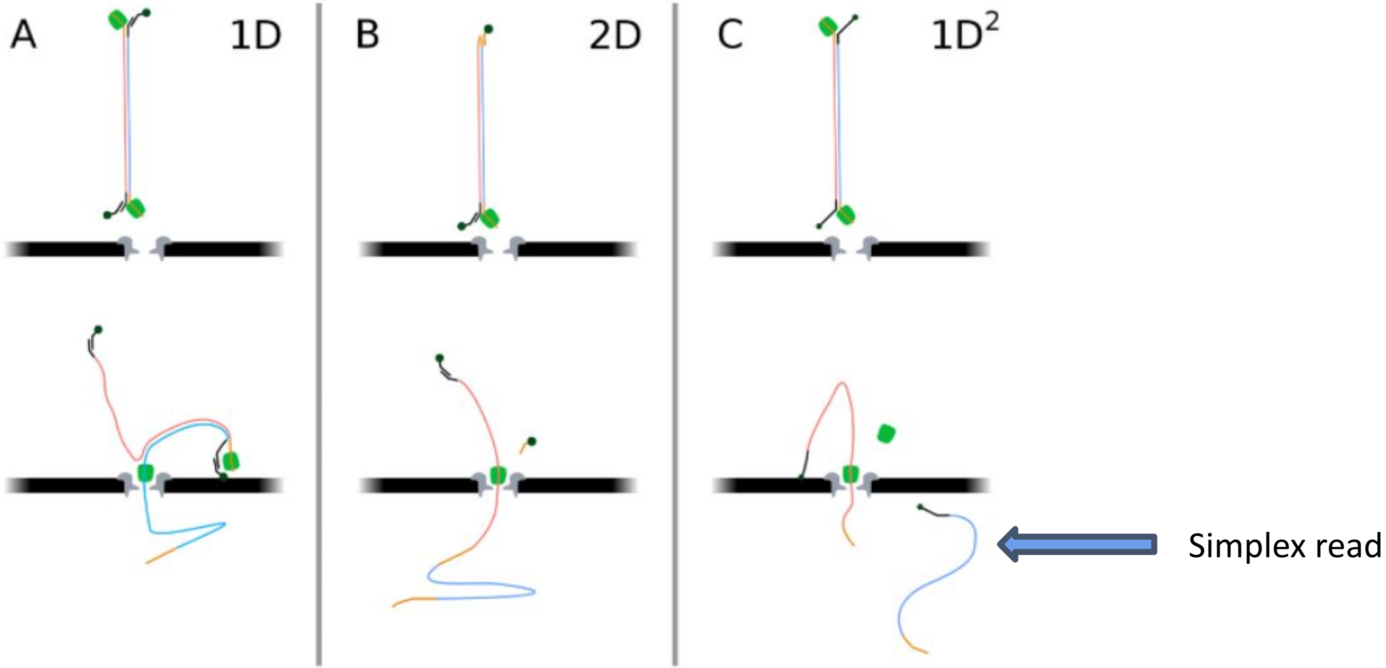


- The electric potential over the membrane pulls the DNA toward the nanopore.
- The motor protein regulates the speed of sequencing (≈ 450 bases s^{-1}).
- Current changes are measured when a base is pulled through the pore.

Veterinærinstituttet
Norwegian Veterinary Institute

De Lannoy et al. - DOI: 10.12688/f1000research.12012.1

1D vs 2D sequencing



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

De Lannoy et al. - DOI: 10.12688/f1000research.12012.1

Simplex vs Duplex reads

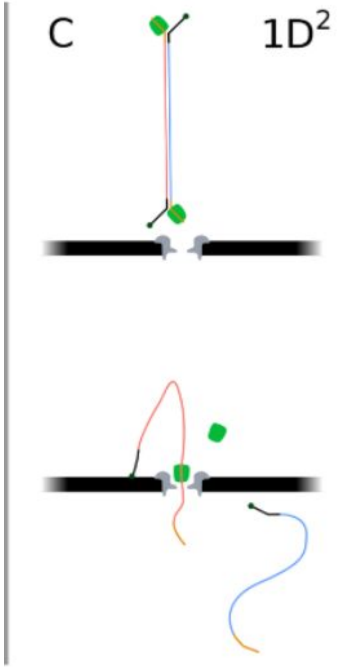
Simplex reads: Single strand of a DNA molecule

Duplex reads: two strands of the “same” DNA molecule

Algorithm:
Sequencing of two single strands
—> at the same pore
—> with little time between them

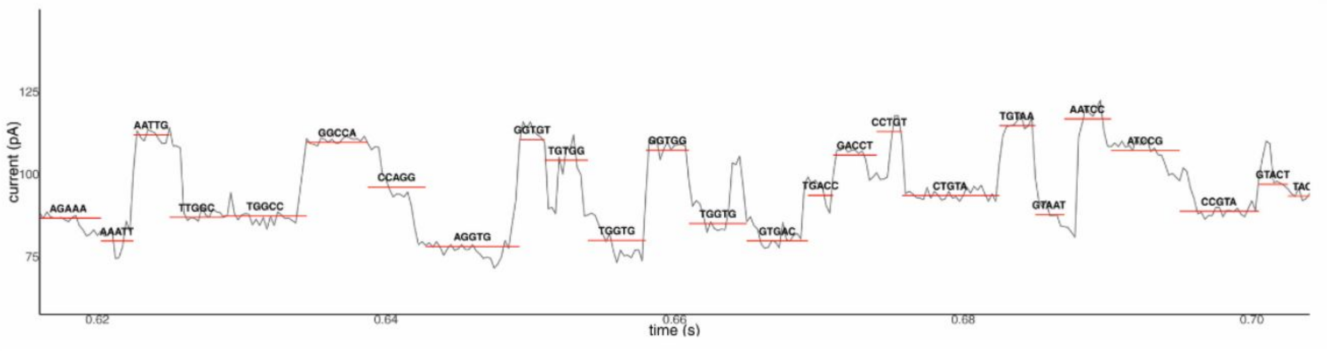
—> Create a consensus sequence of both reads, with higher quality scores.

About 5 -15 % of all single reads can be used for this.



Veterinærinstituttet
Norwegian Veterinary Institute

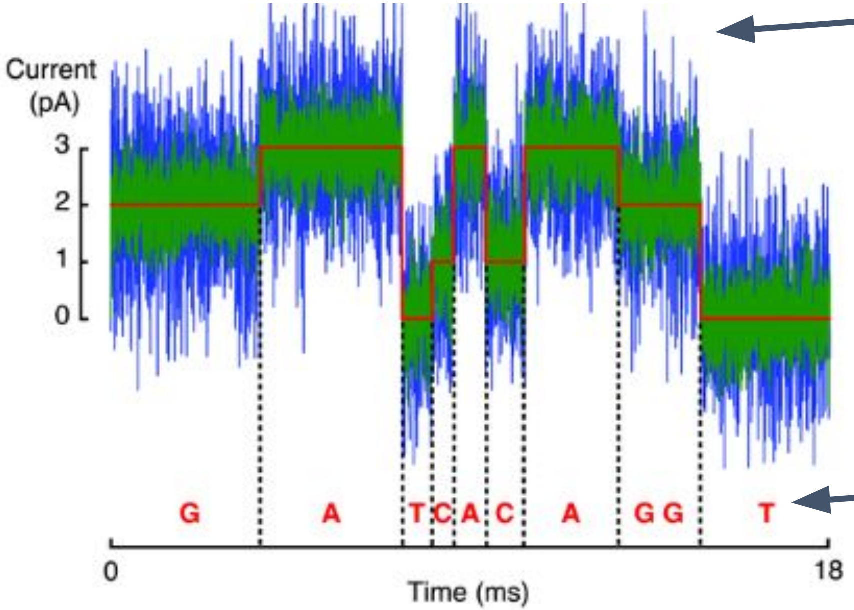
Nanopore basecalling



- 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. - DOI: 10.12688/f1000research.12012.1

Variation in basecalling



Raw signal

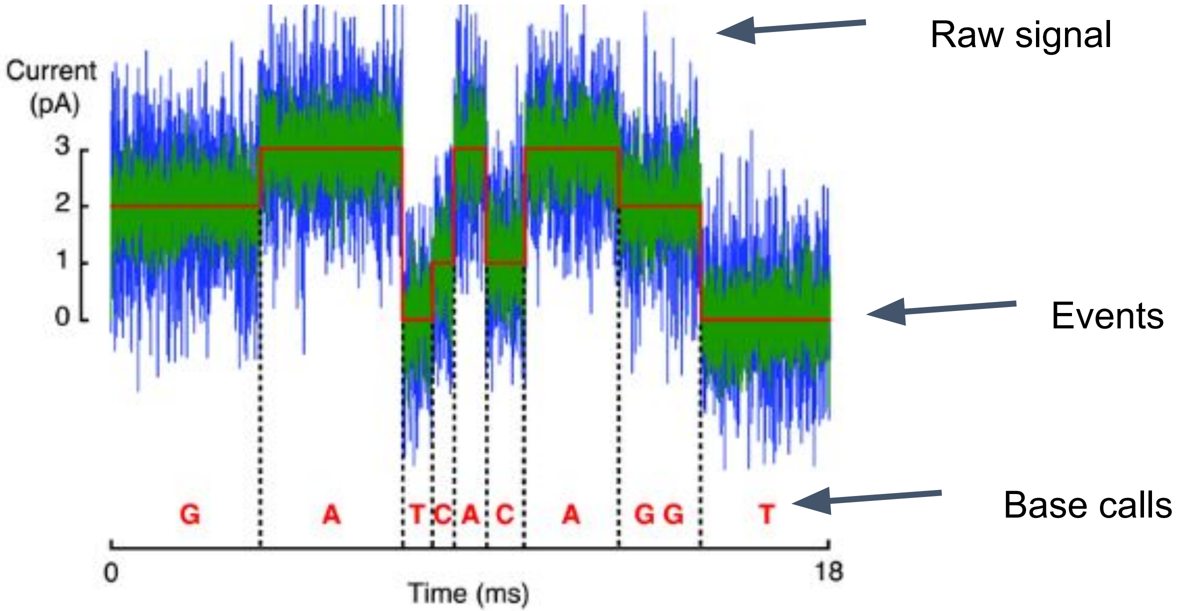
Events

Base calls

- 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 - DOI: 10.1002/elps.201300174

Improving basecalling



Raw signal

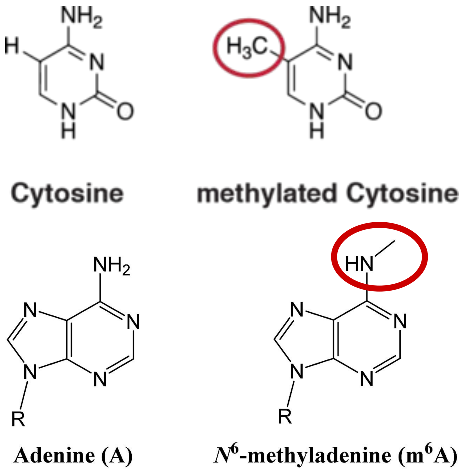
Events

Base calls

- 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 - DOI: 10.1002/elps.201300174

DNA methylation



Cytosine

methylated Cytosine

Adenine (A)

N^6-methyladenine (m^{6}A)

Methylated nucleotides.

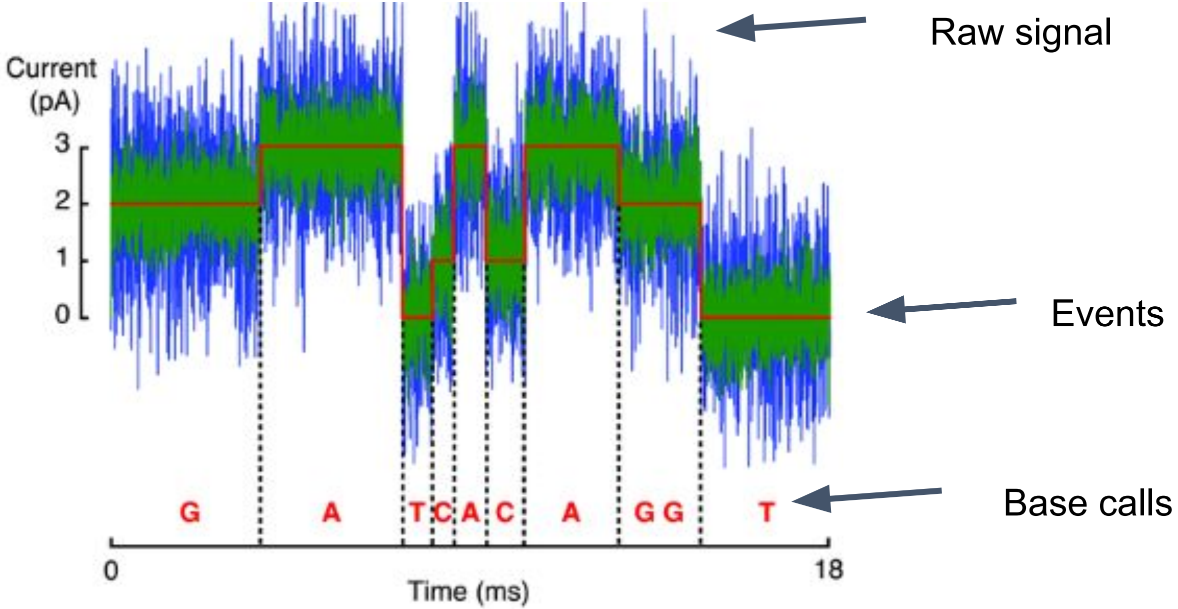
Methylation in Eukaryotes needed for:

- Gene regulation
- Cell differentiation
- Silencing of mobile elements

Methylation in Prokaryotes:

- Silencing of mobile elements
- Phages recognized
- Gene regulation

DNA methylation ?



Raw signal

Events

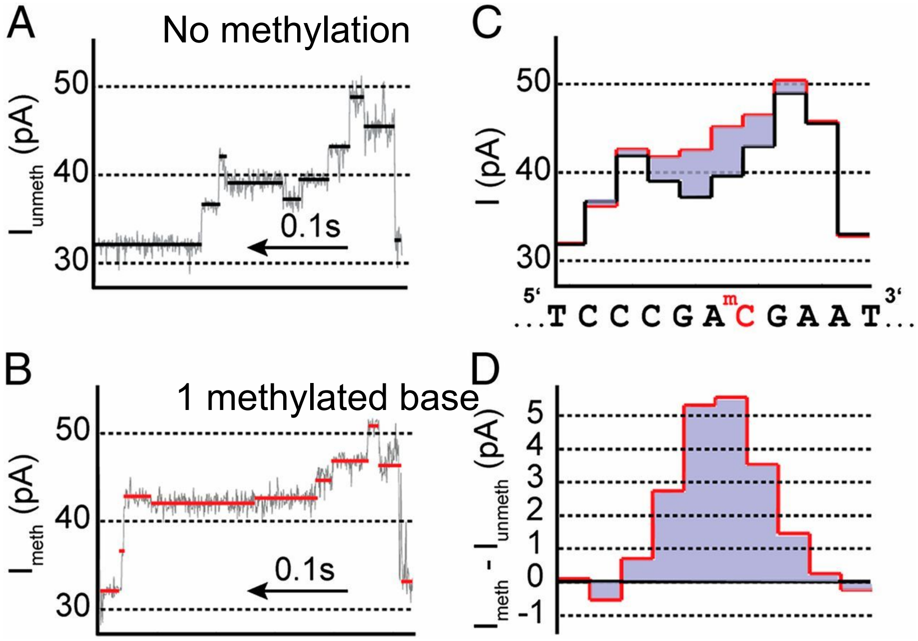
Base calls

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

Veterinærinstituttet
Norwegian Veterinary Institute

O’Donnell et al., 2013 - DOI: 10.1002/elps.201300174

Detecting methylation



Methylation changes the detected current

Lazlo et al., - DOI: 10.1073/pnas.1310240110

The nanopore data

FAST5 files contain:

- multiple raw sequences per file
- not demultiplexed
- Good run can produces > 700 files (compressed tar file: > 200 Gb)

FAST5 files are used for Nanopore basecalling.

- Guppy version 6 (only downloadable when you are registered at Nanopore in a team).
- Output from Guppy can be analyzed with Nanoplot to check the quality.
- Guppy can do demultiplexing, previously done by qcat.

POD5 files are recently introduced as a replacement for FAST5 files.

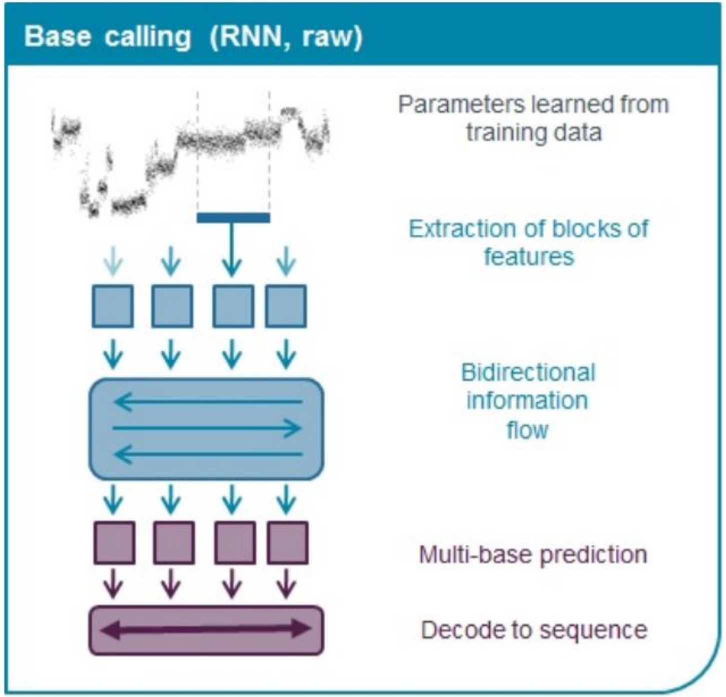
Basecalling software

Many options available:

- Nanopore provides several basecallers
  - MINknow (Included in the sequencing software)
  - **Guppy**(standard)
  - Bonito
  - Dorade
- 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



Base calling (RNN, raw)

Parameters learned from training data

Extraction of blocks of features

Bidirectional information flow

Multi-base prediction

Decode to sequence

www.nanoporetech.com

Original basecallers used Hiden Markov Models

Latest basecallers use Recurrent Neural Network (RRN)

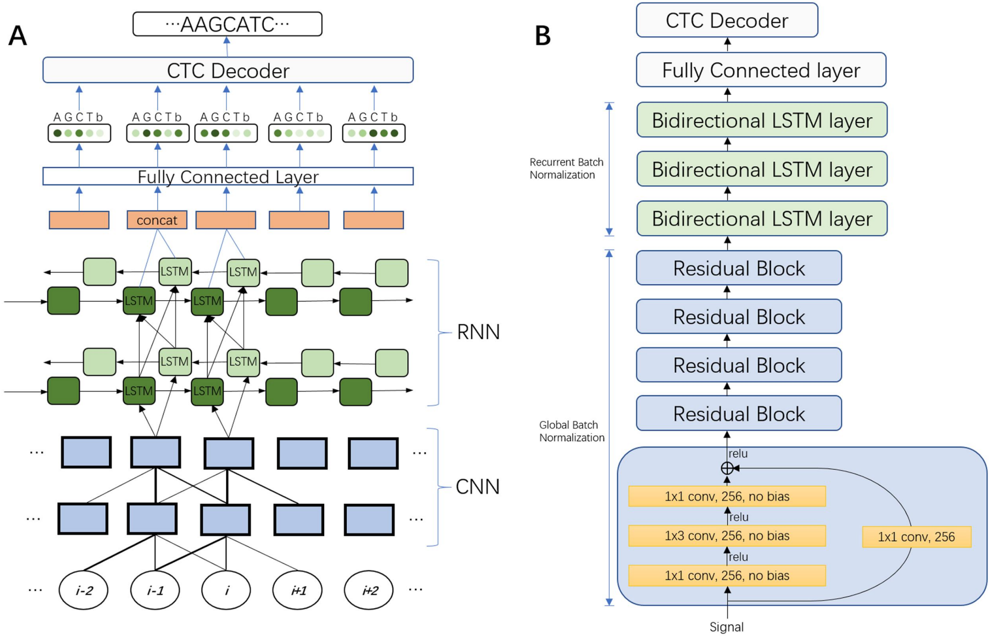
To run basecalling software:

Use Graphical Processor Units (GPU),

instead of

Central Processor Units (CPU).

Basecalling software - Chiron



A combined convolutional neural network and a Recurrent Neural Network

www.nanoporetech.com

Basecalling software - Guppy

A good run gives:

- > 800 Fast5 files (raw signal)
- > 170 Gb data

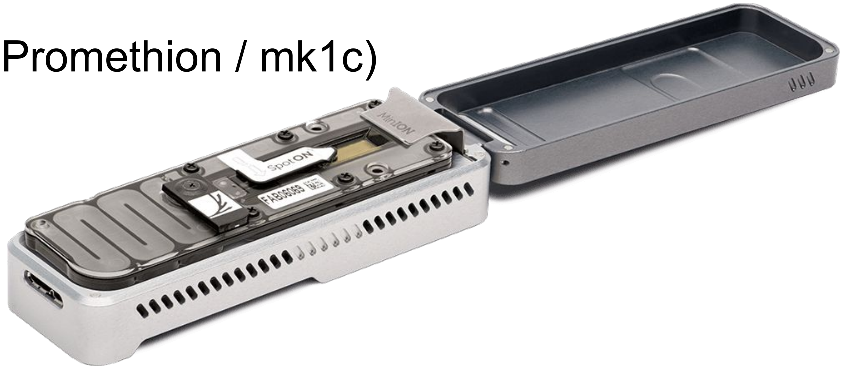
Basecalling with Guppy (V6) using a GPU.

- < 20 Gb fastq data (quality >= 7)
- time needed: about 48 hrs

Basecalling software - Guppy

Basecalling is dependent on:

- Sequencing machine (e.g. minION/Gridion / Promethion / mk1c)
- Flowcell type (9.4.1 or 10.3 / 10.4)
- Library prep
- DNA / RNA (Proteins) as input
- Interest in DNA modifications (Cytosine methylation)



The models for Basecalling can be found in the folder of each guppy installation:

the latest: /cluster/projects/nn9305k/src/miniconda/envs/guppy_gpu_v6.3/data

Guppy - which model to use?

Flowcell r9 : 55 models

Flowcell r10 : 67 models

Mk1C

dna_r9.4.1_450bps_fast.cfg > Fast basecalling

dna_r9.4.1_450bps_fast_mk1c.cfg > Fast basecalling for the mk1c

dna_r9.4.1_450bps_hac.cfg > High accuracy basecalling

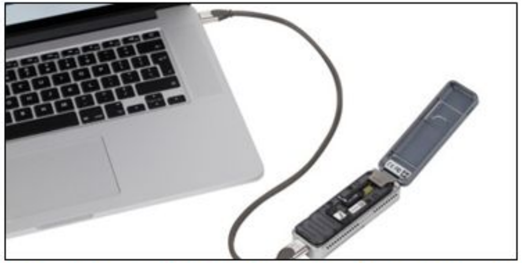
dna_r9.4.1_450bps_sup.cfg > Super High accuracy basecalling

dna_r10.4.1_e8.2_260bps_sup.cfg >

Super High accuracy with newest flowcell 10.4.1 and the kit E8.2 Kit 14

More accurate basecalling takes longer. !!!

Working with the minION



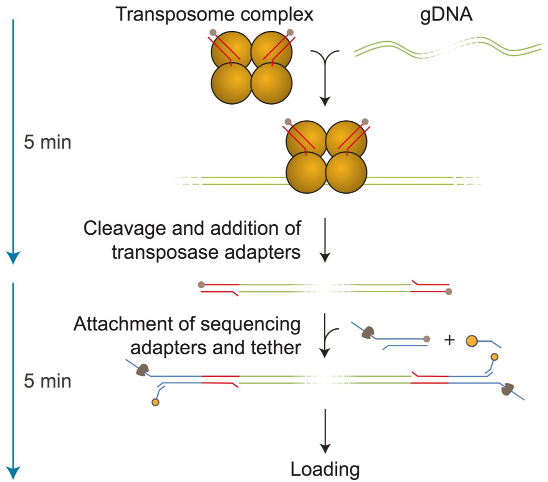
WHAT THE COMPANY SAYS IT LOOKS LIKE

WHAT YOUR PI THINKS IT LOOKS LIKE

HOW THE ACTUAL DATA LOOKS LIKE

HOW REALITY LOOKS LIKE

Sequencing library preparation - DNA


Transposome complex
gDNA
5 min
Cleavage and addition of transposase adapters
Attachment of sequencing adapters and tether
+
5 min
Loading

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

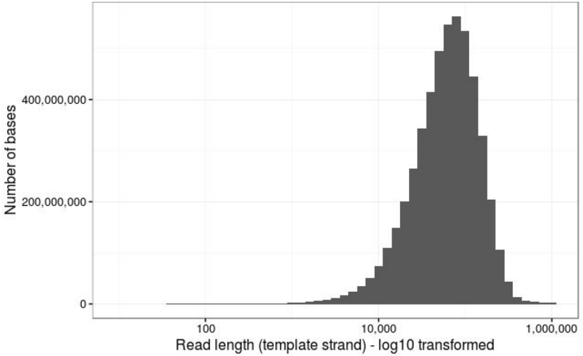
A very quick library preparation is possible

Veterinærinstituttet
Norwegian Veterinary Institute

www.nanoporetech.com

Sequencing output

Sequencing E.coli K-12 MG1655



minION output

Total bases: 5.014.576.373 (5Gb)

Number of reads: 150.604

N50: 63.747

Mean lenght: 33.296,44

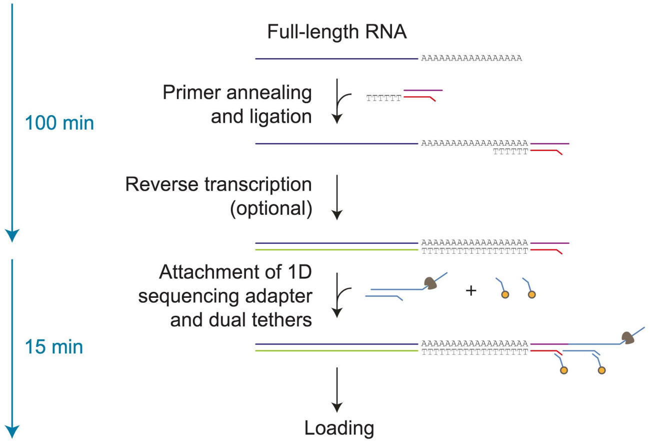
Longest alignable sequence: 2,272,580 bp (2018)

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

Veterinærinstituttet
Norwegian Veterinary Institute

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

Sequencing library preparation - RNA



Full-length RNA
AAAAAAAAAAAAAAAAA
100 min
Primer annealing and ligation
TTTTTT
Reverse transcription (optional)
Attachment of 1D sequencing adapter and dual tethers
+
15 min
Loading

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 !

Veterinærinstituttet
Norwegian Veterinary Institute

www.nanoporetech.com / Garalde et al., 2018 Nature Methods volume 15, pages 201–206

MinION applications

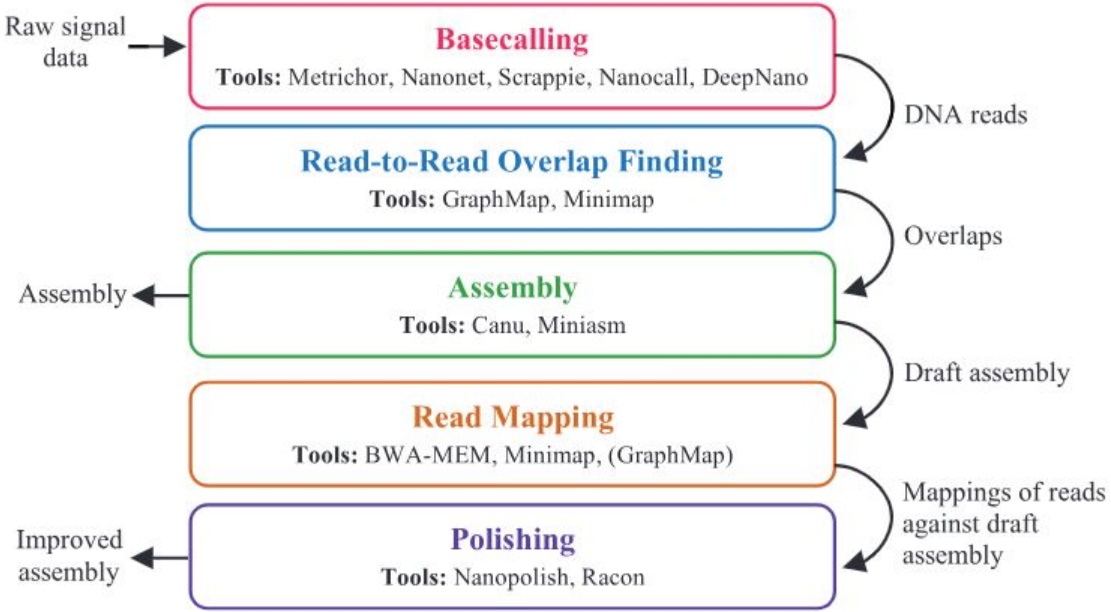
- *De novo* shotgun sequencing (pcr / primer free sequencing)
  - Especially good for repetitive regions
  - Finishing Prokaryote / Eukaryote genomes
  - Detection of structural genome variation (indels)

MinION applications

- 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



Raw signal data

Basecalling

Tools: Metrichor, Nanonet, Scrappie, Nanocall, DeepNano

DNA reads

Read-to-Read Overlap Finding

Tools: GraphMap, Minimap

Overlaps

Assembly

Assembly

Tools: Canu, Miniasm

Draft assembly

Read Mapping

Tools: BWA-MEM, Minimap, (GraphMap)

Mappings of reads against draft assembly

Improved assembly

Polishing

Tools: Nanopolish, Racon

Senol Cali et al., - DOI: 10.1093/bib/bby017

Genome assembly

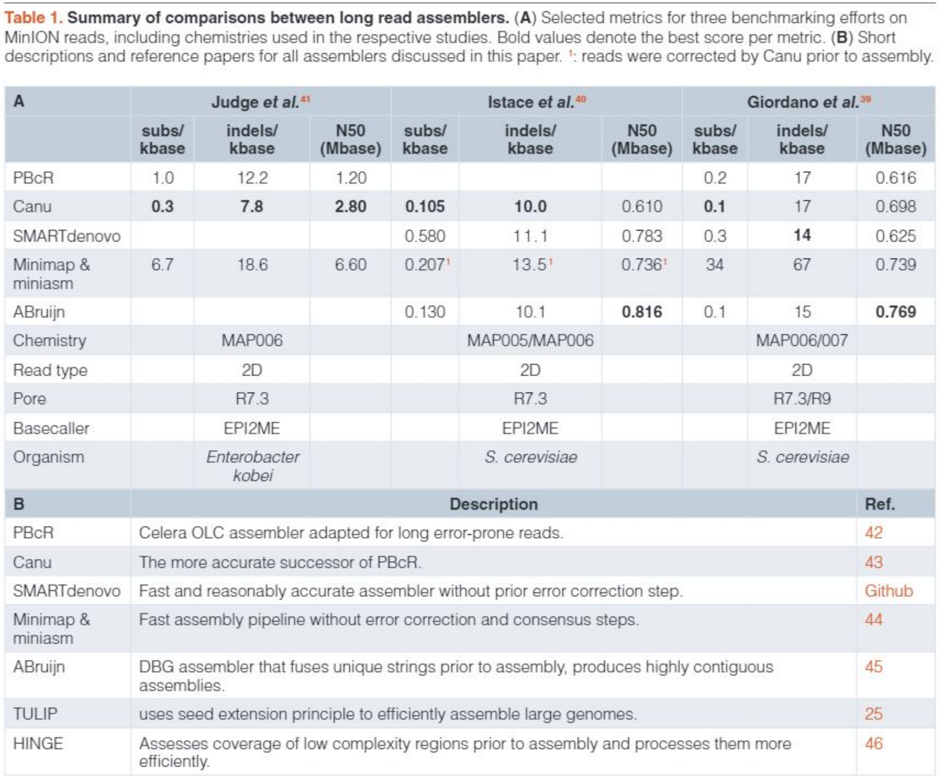


Table 1. **Summary of comparisons between long read assemblers.** (**A**) Selected metrics for three benchmarking efforts on MinION reads, including chemistries used in the respective studies. Bold values denote the best score per metric. (**B**) Short descriptions and reference papers for all assemblers 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., - DOI: 10.12688/f1000research.12012.1

Canu assembler

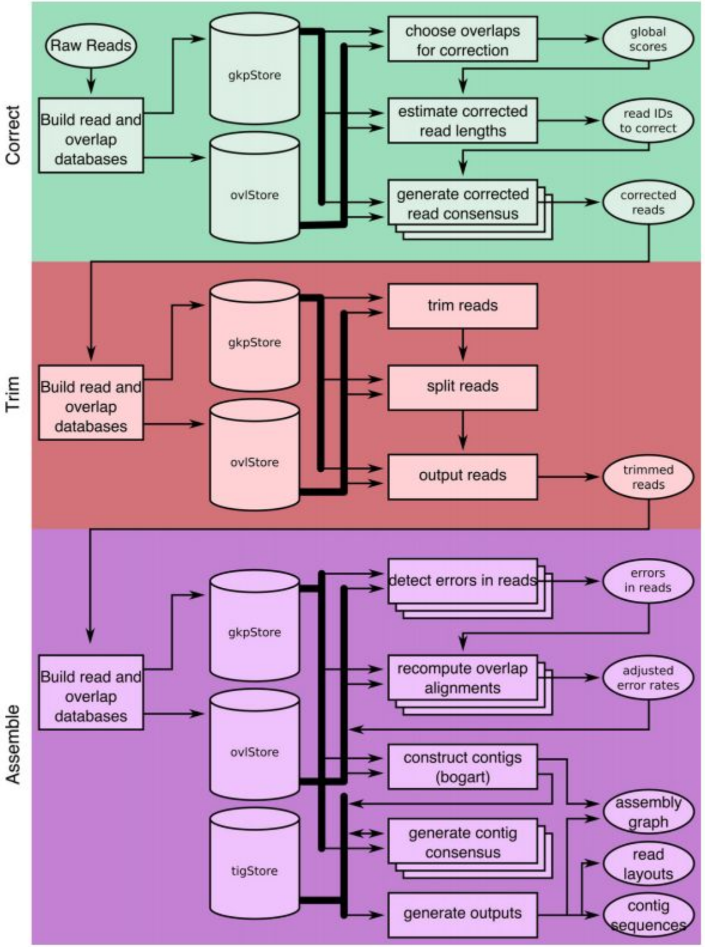
Canu Assembly pipeline

1. Error correction
2. Trimming
3. Assembly

gkpStore: reads database

ovlStora: overlap database

tigStore: contigs database

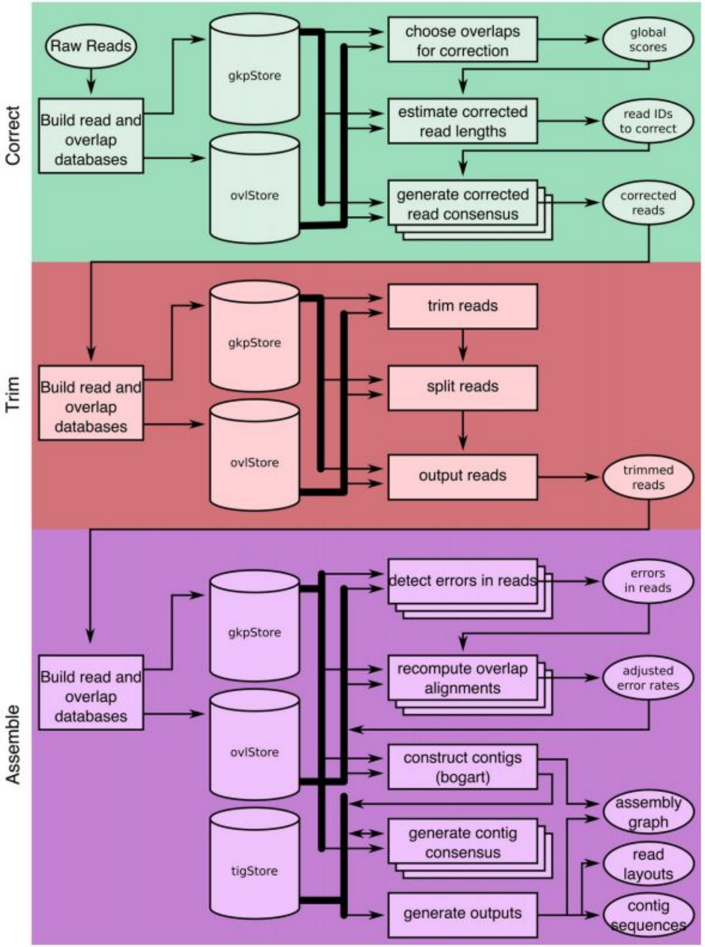


Koren et al., 2017 - DOI: 10.1101/gr.215087.116

Error correction

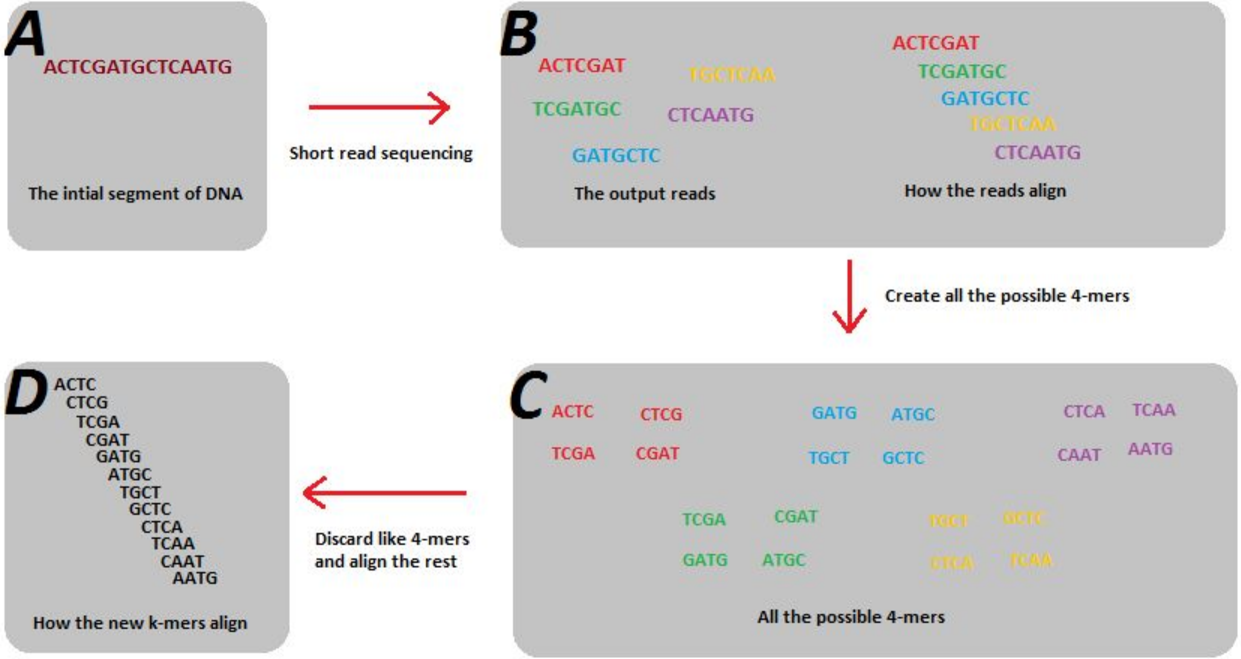
1. Reads split into **kmers**
2. Kmers used to identify overlap
3. Correct reads using overlap

Corrected reads are trimmed



Koren et al., 2017 - DOI: 10.1101/gr.215087.116

Kmers



A

ACTCGATGCTCAATG

The intial segment of DNA

Short read sequencing

B

ACTCGAT

TGCTCAA

TCGATGC

CTCAATG

GATGCTC

The output reads

ACTCGAT
TCGATGC
GATGCTC
TGCTCAA
CTCAATG

How the reads align

Create all the possible 4-mers

C

ACTC CTCG GATG ATGC CTCA TCAA

TCGA CGAT TGCT GCTC CAAT AATG

TCGA CGAT TGCT GCTC

GATG ATGC CTCA TCAA

All the possible 4-mers

Discard like 4-mers and align the rest

D

ACTC
CTCG
TCGA
CGAT
GATG
ATGC
TGCT
GCTC
CTCA
TCAA
CAAT
AATG

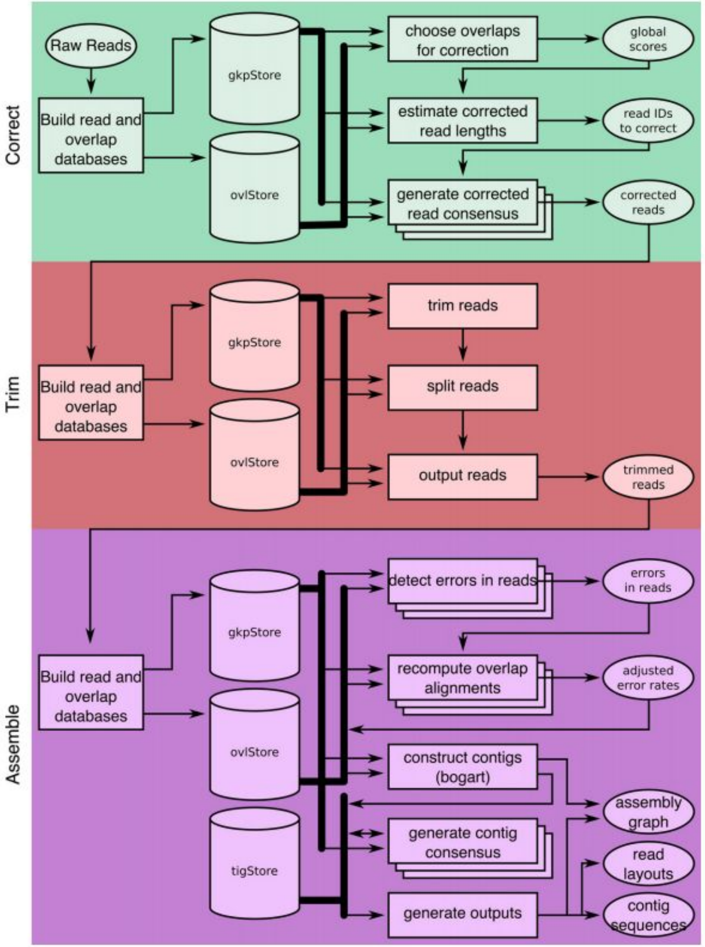
How the new k-mers align

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 - DOI: 10.1101/gr.215087.116

Canu assembly *E. coli* genome

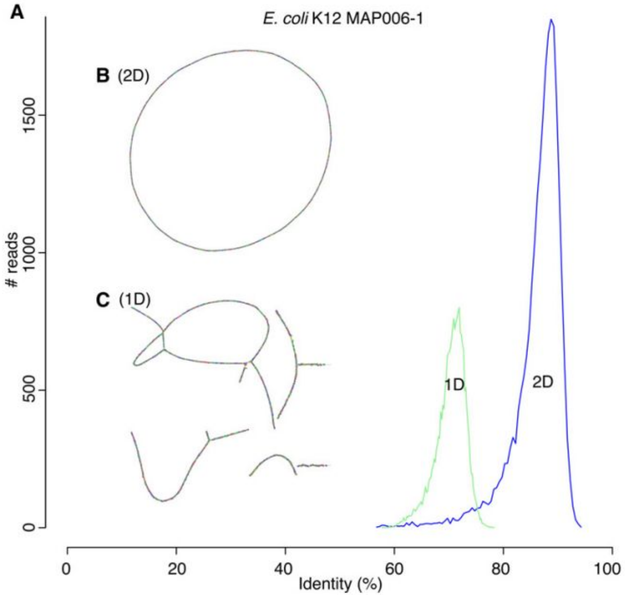


Figure 5. Canu can assemble both 1D and 2D Nanopore *Escherhicia coli* reads. (*A*) A comparison of error rates for 1D and 2D read error rates versus the reference. Template 1D and 2D reads from the MAP006-1 *E. coli* data set were aligned independently to compute an identity for all reads with an alignment >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 - DOI: 10.1101/gr.215087.116

Polishing

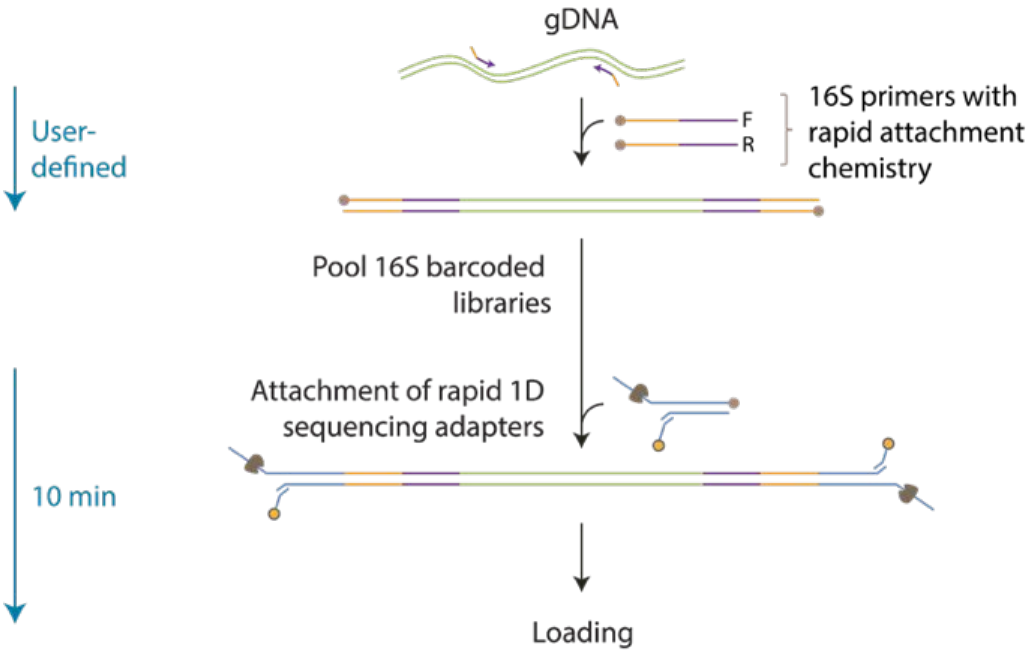
Nanopolish: Improve consensus sequence of assemblies

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

Amplicon sequencing



gDNA

16S primers with rapid attachment chemistry

User-defined

Pool 16S barcoded libraries

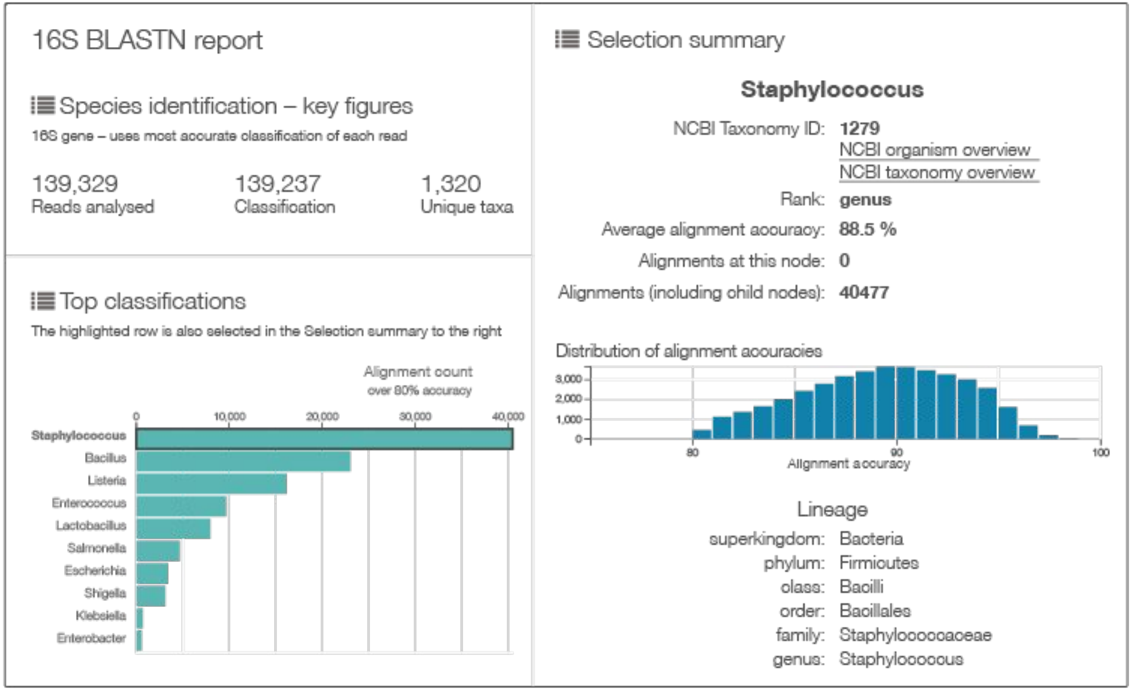
Attachment of rapid 1D sequencing adapters

10 min

Loading

www.nanoporetech.com

Amplicon sequencing



16S BLASTN report

Species identification – key figures

16S gene – uses most accurate classification of each read

139,329	139,237	1,320
Reads analysed	Classification	Unique taxa

Top classifications

The highlighted row is also selected in the Selection summary to the right

Selection summary

Staphylococcus

NCBI Taxonomy ID:	**1279** NCBI organism overview NCBI taxonomy overview
Rank:	**genus**
Average alignment accuracy:	**88.5 %**
Alignments at this node:	**0**
Alignments (including child nodes):	**40477**

Distribution of alignment accuracies

Lineage

superkingdom:	Bacteria
phylum:	Firmicutes
class:	Bacilli
order:	Bacillales
family:	Staphylococcaceae
genus:	Staphylococcus

Accuracy is low

Several pipelines available

- NanoClust (16s)
- NanoAmpli-Seq (16S)
- Amplicon_sorter
- NGSpeciesID (COI)
- …

How to generate a consensus?

A short comparison

	Illumina	PacBio	minION
Output (Gb)	7.5 – 6000	5-8	10-50
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

Nanopore sequencing technology, bioinformatics and applications (2021)

- https://pubmed.ncbi.nlm.nih.gov/34750572/

The End

Contact details:

Thomas Haverkamp
Thomas.haverkamp@vetinst.no

twitter: @Thomieh