Bio-technological background for „Covid_alignment“ hands-on exercise

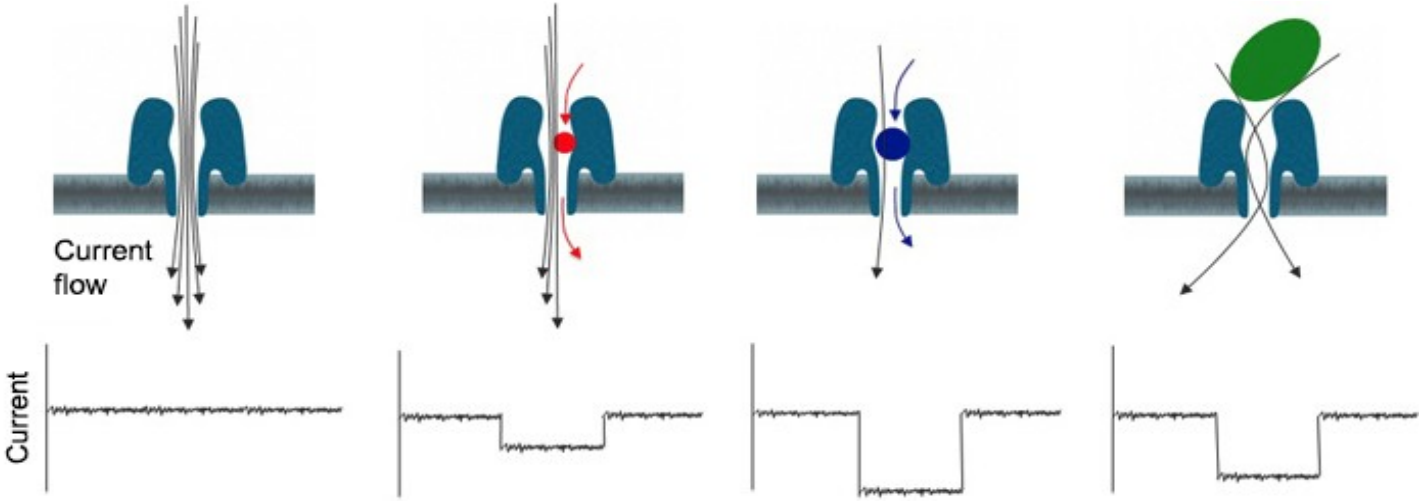
Sophia Derdak

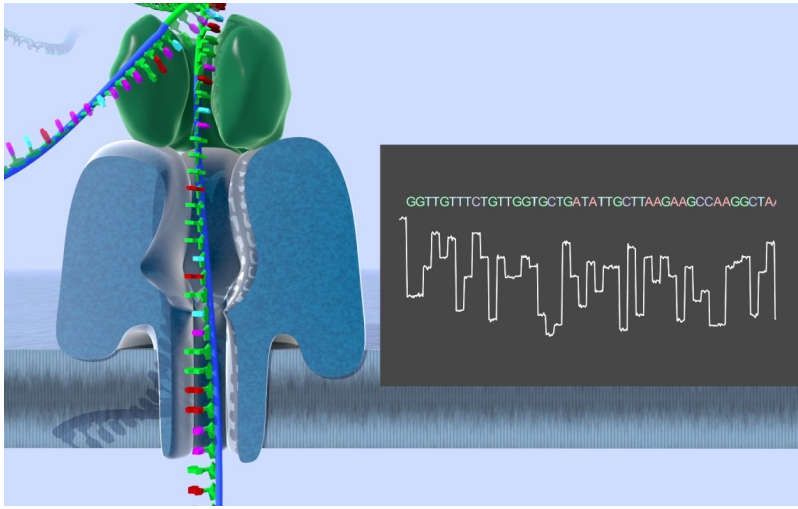
Core Facilities MedUni Vienna

sophia.derdak@meduniwien.ac.at

Oxford Nanopore Technologies long read sequencing

Nanopore sequencing principle



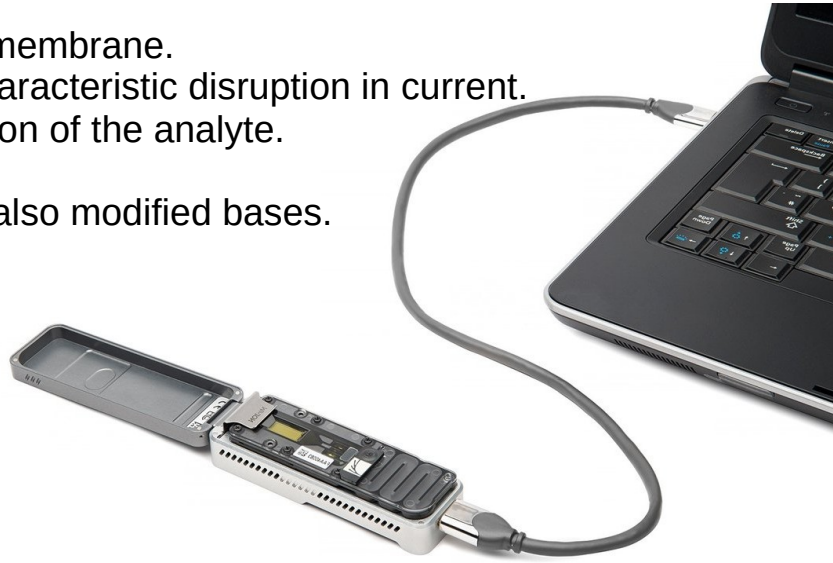


- protein nanopore set in an electrically resistant membrane bilayer

An ionic current is passed through the nanopore by setting a voltage across this membrane.
If an analyte passes through the pore or near its aperture, this event creates a characteristic disruption in current.
Measurement of that current makes it possible to identify the molecular composition of the analyte.

- distinguish between the four standard DNA bases G, A, T and C, U in RNA and also modified bases.

View it in motion: https://www.youtube.com/watch?v=E9-Rm5AoZGw



Nanopore sequencing: features and applications

+

- read long DNA molecules directly (no amplification, no bias)
- read mRNA molecules directly (isoforms)
- read and identify modified (e.g. methylated) nucleotides
- results in real time while sequencing
- portable sequencing device: facilitates field sequencing

−

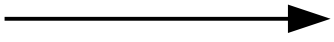
- high input quantities required
- base resolution is poor, error rate of 5-10%, not ideal for point mutation analysis

Alignment of sequencing reads to reference genome

Sequence alignment – the old-fashioned way: BLAST

TGGACCATCTGGTTGAGCATGTGGGGGTCAACTCCCA

query



blastn

BLAST
Nucleotide
Sequence
database

Schistocephalus solidus genome assembly S_solidus_NST_G2 ,scaffold SSLN_scaffold0000849

Sequence ID: emb|LL901051.1| **Length:** 73717 **Number of Matches:** 1

Range 1: 6146 to 6165 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
40.1 bits(20)	1.2	20/20(100%)	0/20(0%)	Plus/Minus

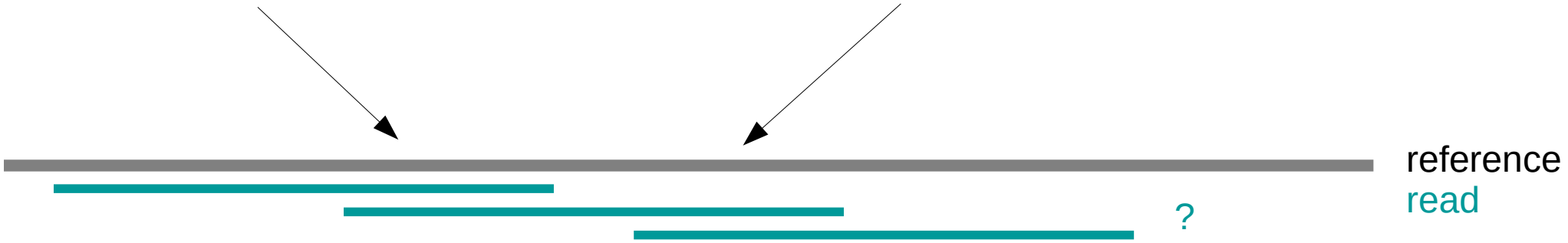
```
Query  8     TCTGGTTGAGCATGTGGGGG  27
             ||||||||||||||||||||
Sbjct  6165  TCTGGTTGAGCATGTGGGGG  6146
```

https://blast.ncbi.nlm.nih.gov/Blast.cgi

Sequence alignment from whole genome re-sequencing – BLAST x 100.000.000?

100.000.000 x reads in fastq format
(~ 100 bases)

1 x reference sequence in fasta format
~3.000.000.000 bases



How long does this take? Hours, days?

What are the computational resources needed? Cpus, memory ...

Do we allow for mismatches? How many?

Gapped alignments? How long a gap do we accept?

Exhaustive? Are all possible matches output?

The most commonly used tools still perform a “query” of each read (short sequence) against the reference “database” (long sequence).

They differ in their algorithmic (Burrows-Wheeler indexing a.o.) and computational (parallelization a.o.) setup, and many of them require computational clusters.

Their names are: e.g. bowtie, bwa, STAR, novoalign, gem, minimap2 and most of them are open source.

CNAG cluster

Alignments in “difficult” genomic regions: The problem of mappability and ambiguous alignments

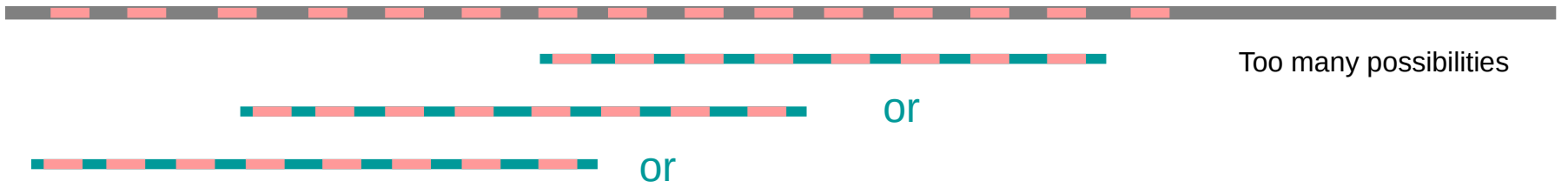
0. aligning without mismatches: the most common case

1. aligning with mismatches: essential for variant calling

2. aligning sequences of highly homologous genes

3. aligning in highly polymorphic regions

4. aligning in repetitive regions



The coverage

represents the number of times a base of the sample genome (or target region) is read during sequencing.

A higher coverage provides higher power for data analysis.

How to get a higher coverage:

- mainly by loading more sequencing units (indexes, lanes, entire flowcells) with the same library preparation

Typical coverage numbers:

- whole genome: 30x
- exome: 50-100x
- custom gene panel capture: >1000x
- RNA-Seq: depending on gene expression

Thank you!

Sophia Derdak

Core Facilities MedUni Vienna

sophia.derdak@meduniwien.ac.at