Long-read sequencing

Method

Genomes have many long **repetitive elements, copy number alterations** relevant to evolution, adaptation and disease

many of these complex elements are so long that **short-read** paired-end technologies are **insufficient** to resolve them.

Long-read sequencing
-> several kilobases, span complex or repetitive regions with a single continuous read
-> transcriptomic research, as they are capable of spanning entire mRNA transcripts

2 main types:
single-molecule real-time sequencing approaches
synthetic approaches that rely on existing shortread technologies to construct long reads *in silico*.

Single-molecule real-time sequencing (SMRT)

PacBio (Pacific Biosciences***) and ONT (****Oxford Nanopore Technologies)*

PacBio:

1. specialized **flow cell** with many thousands of individual picolitre **wells with transparent bottoms** = zero-mode waveguides (ZMW)

2. **fixes the polymerase to the bottom** of the well and allows the DNA strand to progress through the ZMW.

-> constant location of incorporation -> the system can focus on a single molecule.

3. continuously visualized with a **laser and camera system** that records the colour and duration of emitted light as the labelled nucleotide momentarily pauses

during incorporation

4. polymerase **cleaves** the dNTP-bound fluorophore during incorporation, allowing it to d**if- fuse away from the sensor** area before the next labelled dNTP is incorporated.

5. unique **circular template** that allows each template to be sequenced multiple times as the polymerase repeatedly traverses the circular molecule. -> multiple passes are used to generate a consensus read of insert, known as a circular consensus sequence

https://youtu.be/NHCJ8PtYCFc

https://youtu.be/NHCJ8PtYCFc

ONT:

consumer prototype: MinION
don't monitor incorporations of nucleotides guided by a template DNA strand.
directly detect the DNA composition of a native **ssDNA** molecule

1. DNA is **passed through a protein pore** as **current** is passed through the pore

2. As the DNA translocates through the action of a secondary motor protein, a **voltage blockade** occurs that modulates the current passing through the pore

3. shifts in voltage are **characeristic of the particular DNA sequence** in the pore, which can then be interpreted as a **k-mer.** (*k-mers* are unique subsequences of a sequence of length k)

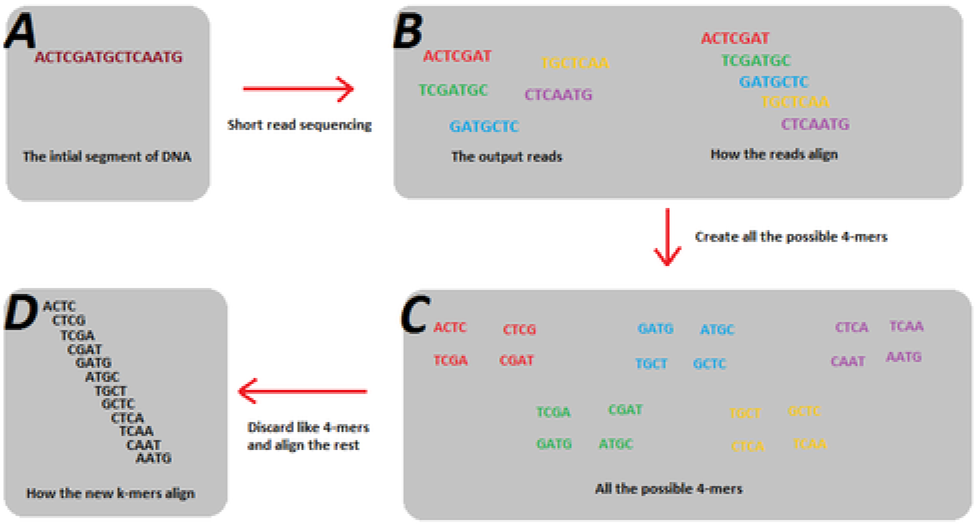
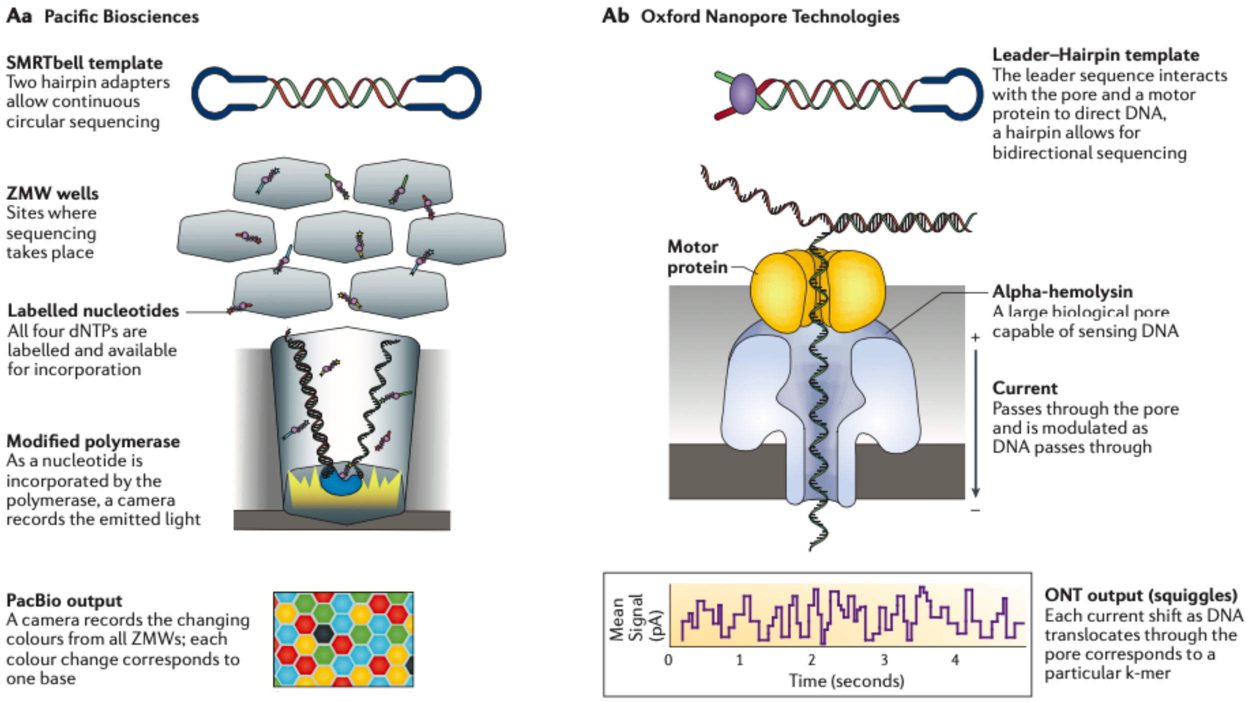


Image result for k-mer"

-> not 1–4 possible signals, the instrument has more than 1,000 — one for each possible k-mer


Aa Pacific Biosciences
SMRTbell template
Two hairpin adapters allow continuous circular sequencing
ZMW wells
Sites where sequencing takes place
Labelled nucleotides
All four dNTPs are labelled and available for incorporation
Modified polymerase
As a nucleotide is incorporated by the polymerase, a camera records the emitted light
PacBio output
A camera records the changing colours from all ZMWs; each colour change corresponds to one base
Ab Oxford Nanopore Technologies
Leader–Hairpin template
The leader sequence interacts with the pore and a motor protein to direct DNA, a hairpin allows for bidirectional sequencing
Motor protein
Alpha-hemolysin
A large biological pore capable of sensing DNA
+
Current
Passes through the pore and is modulated as DNA passes through
–
Mean Signal (pA)
0
1
2
3
4
Time (seconds)
ONT output (squiggles)
Each current shift as DNA translocates through the pore corresponds to a particular k-mer

Existing shortread technologies to construct long reads *in silico*

relies on a system of barcoding to associate fragments that are sequenced on existing short-read sequencers

1. par- tition large DNA fragments into either **microtitre wells or an emulsion** such that very few molecules exist in each partition.

2. template frag- ments are **broken off and barcoded**. -> existing short-read instrumentation

3. data are split by barcode and reassembled with the knowledge that fragments sharing **barcodes** are derived from the **same original large fragment**

Illumina and 10X Genomics

Illumina:
microtitre plate, no futher special instrument

10X Genomics:
use **emulsion** to partition DNA
require the use of a microfluidic instrument to perform pre-sequencing reactions.

Throughput, accuracy, cost, application

2016 data:

Throughput:

Illumina HiSeq X	800–900 Gb per flow cell*
Pacific BioSciences RS II	500 Mb–1 Gb*
Oxford Nanopore MK 1 MinION	Up to 1.5 Gb
However, new PromethION	3.6 Tb has been achieved!

Error profile:

Short-read: usually 0.1% - 1%

Pacific BioSciences RS II	13% single pass, ≤1% circular consensus read
Oxford Nanopore MK 1 MinION	~12%

A major limitation of nanopore sequencing is its high error rate, which despite recent improvements to the nanopore chemistry and computational tools still ranges between 5% and 15%.

error rate of 30% during the early phase of its release around 2014.[22] With the latest R9 release in 2016 raw error rates have been reduced to between 2-13% for various types of DNA sequencing

Cost:

Illumina HiSeq X	1000$ instrument + 7$ cost per GB
Pacific BioSciences RS II	695$ instrument + 1,000$ per GB
Oxford Nanopore MK 1 MinION	1,000$ instrument + 750$ per GB

Overall:

lower throughput, higher error rate and higher cost per base relative to short read sequencing

Long-read technologies are improving rapidly, and may become the mainstay of sequencing

Adventages:
High resolution genome assemblies
close gaps in genomes by spanning the low complexity regions

Tree of Life initiative, a collaboration across multiple centres is in the process of developing high resolution **reference sequences for >50 vertebrate species** using a combination of long read, short read and linked-read approaches

Another leading project is the large **bacterial sequencing project** NCTC 3000 at the **Wellcome Sanger Institute**, which is using PacBio sequencing to sequence complete bacterial genomes

A recent example of this was a study where SMRT sequencing was used to identify a **reservoir of antibiotic resistant plasmids within hospitals**