ELEC6205
ELEC3223
DNA sequencing
What is a genome?

The first human genome sequence was completed in 2003 through the Human Genome Project, a 13-year effort with an estimated cost of $2.7 billion.

In 2008 a human genome was sequenced over a 5-month period for approximately $1.5 million.

In the rapidly evolving field of “next-generation” sequencing (NGS) many commercial technologies have emerged during the past decade.

The cost has now dropped to <$1000 and sequencing will become a routinely used clinical diagnostic tool during the next decade.

see Goodwin et al  Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics 17 2016
Sequencing approaches

• “Classical” chain termination - Sanger method

• Next Generation Sequencing (NGS):
  – Illumina – market lead but many others
  – Semiconductor (Ion Torrent)

• 3\textsuperscript{rd} Generation (single molecule)
  – Nanopore (Oxford Nanopore)
DNA Sequencing

• DNA sequencing determines the sequence of nucleotides in a fragment of DNA.
• Sequencing an entire genome is complex:
  • DNA of the genome is broken into many small pieces,
  • Pieces are sequenced
  • Reassembled into a single long "consensus."

• Sanger sequencing: target DNA is copied many times, making many fragments of different lengths.
• Fluorescent “chain terminator” nucleotides mark the ends of the fragments and allow the sequence to be determined.
Sanger sequencing: Chain termination method

- Developed by British biochemist Fred Sanger and his colleagues in 1977.
- **Human Genome Project**: Sanger sequencing used to determine the sequences of many small fragments of human DNA (from many individuals). Fragments aligned based on overlapping portions to assemble the sequences of larger regions of DNA and, eventually, entire chromosomes.
- Although genomes are now typically sequenced using NGS, Sanger sequencing is still widely used for individual pieces of DNA (fragments for cloning or generated through PCR).
The details of Sanger sequencing

- Start by making **many copies** of a **target DNA region** (usually around 900bp). Similar to DNA replication in a cell, or PCR.
- We need:
  - A DNA **polymerase** enzyme
  - A **primer** - short piece of single-stranded DNA that binds to the template DNA and acts as a "starter" for the polymerase
  - Four **regular** DNA nucleotides (dATP, dCTP, dGTP, dTTP)
  - The **template** DNA to be sequenced
ONE BIG DIFFERENCE

• Sanger sequencing also contains a unique ingredient:

• **Dideoxy**, or **chain-terminating**, versions of all four nucleotides (ddATP, ddTTP, ddCTP, ddGTP), each labelled with a different colour fluorescent dye.
Dideoxy nucleotides

Once a dideoxy nucleotide has been added to the chain, there is no hydroxyl available and no further nucleotides can be added. The chain ends with the dideoxy nucleotide, which is marked with a particular color of dye depending on the base (A, T, C or G) that it carries.

Regular nucleotide: the 3’ hydroxyl group acts as a “hook,” where a new nucleotide is added to an existing chain (recall first lecture). For short it’s called **dNTP**

Dideoxy nucleotides have one key difference: they lack a hydroxyl group on the 3’ carbon of the sugar ring.

The dye molecule on a dideoxy nucleotide is linked to the nitrogenous base.

Deoxynucleotide (dNTP)  

Dideoxynucleotide (ddNTP)
The recipe for Sanger sequencing

• Mix DNA sample with primer, DNA polymerase, and regular nucleotides. The four dye-labeled, chain-terminating dideoxy nucleotides are also added, but in **much smaller amounts**.
• Heat mixture to denature the template DNA (separate the strands), then cool to anneal the primer to the single-stranded template.
• Once primer has bound, raise temperature, allowing DNA polymerase to synthesize new DNA starting from the primer. DNA polymerase continues adding nucleotides to the chain until it happens to add a **dideoxy nucleotide** instead of a regular one. At this point, **no further nucleotides** can be added, and the strand **ends** with the dideoxy nucleotide.
• Repeat many times so that it’s guaranteed that a dideoxy nucleotide will have been incorporated at every single position of the target DNA in at least one reaction.
• Tube will contain many fragments of different lengths, ending at each of the nucleotide positions in the original DNA. The ends of the fragments will be labelled with dyes that indicate their final nucleotide.
Are all fragments labelled?

• No - not all fragments are labelled as it’s chance whether a dideoxy nucleotide gets incorporated in a particular polymerization reaction. Some synthesized pieces of DNA will consist only of normal, unlabeled nucleotides, and will simply end when the polymerase falls off the template, not due to the addition of a chain-terminating nucleotide. These unlabeled DNA molecules do not interfere with the sequencing reaction, as they are "invisible" in the detection step due to their lack of a dye label.
How to read the strands?

• The DNA fragments are separated using **capillary electrophoresis (CE)**
• The fluorescent signal from the four difference colour nucleotides is read with a laser
Capillary Electrophoresis

DNA has constant charge/mass ratio therefore difficult to separate on size alone.

A gel is used into which the molecule are entangled. All the DNA migrates to the +ve electrode with smaller fragments appearing first.
SUMMARY

1. Primer for replication

2. Prepare four reaction mixtures; include in each a different replication-stopping nucleotide

3. Replication products of "C" reaction

4. Separate products by gel electrophoresis

Sequence of interest

Read sequence as complement of bands containing labeled strands
Uses and limitations

• Sanger sequencing gives high-quality sequence for relatively long stretches of DNA (up to about 900 base pairs). It's typically used to sequence individual pieces of DNA, such as bacterial plasmids or DNA copied in PCR.

• Expensive and inefficient for larger-scale projects, such as the sequencing of an entire genome or metagenome.
Next Generation Sequencing (NGS)

In 2001, the cost of sequencing a human genome was $100M. Now it’s less than $1000

- **Highly parallel**: many reactions take place at the same time
- **Micro scale**: reactions are tiny
- **Fast**: reactions done in parallel
- **Low-cost**: sequencing is much cheaper than Sanger
- **Shorter length**: reads typically range from 50 - 700 nucleotides in length

NGS like running a very large number of tiny Sanger sequencing reactions in parallel.
Illumina Sequencing-by-synthesis

- Developed by Shankar Balasubramanian and David Klenerman at Cambridge.
- Illumina is the market lead and now worth £31 Billion.
“Illumina” method

- DNA sequencing libraries are amplified by PCR.
- DNA sequenced by synthesis. Sequence is determined by the addition of nucleotides to the complementary strand rather through chain termination chemistry (Sanger).
- Spatially segregated, amplified DNA templates are sequenced simultaneously in a massively parallel fashion without the requirement for a physical separation step.
How it works

• Three components:
  – Amplify
  – Sequence
  – Analyse - informatics
• Chop up DNA into small pieces and modify with molecules that act as reference points during sequencing and adaptor sequences that bind to small oligos on a chip surface
• Load modified DNA onto the chip and amplify in-situ
• Chip contains hundreds of thousands of short oligos anchored to the bottom
• They grab the many different DNA fragments with adaptor sequences
Only ONE nucleotide is incorporated each cycle. But there are 100s of millions of simultaneous reactions.
Image of a chip surface

https://www.youtube.com/watch?v=77r5p8IbwJk
Sequencing by Synthesis: Ion Torrent

- An *electrical* signal is produced when each dNTP incorporates into a strand.
- The four nucleotides are added *sequentially* and therefore no need for a blocked dNTP.
- Ion Torrent reads the signal as *change in pH* each time a nucleotide is incorporated.
1st step is **bead-based template enrichment**
Then the beads are arrayed onto a CMOS chip along with primers and different beads that contain the enzymes.
Each well contains a **single transistor** and a **single bead**.
During the first cycle, a single nucleotide species is added and each complementary base is incorporated into a newly synthesized strand by a DNA polymerase.
The by-product of this reaction is a **pyrophosphate molecule (PPi)** and a **H⁺**.
Bead based template enrichment: emulsion PCR

Fragmented DNA templates (with adapters) are captured in an aqueous droplet (micelle) in oil, along with a bead covered with complementary adapters, deoxynucleotides (dNTPs), primers and DNA polymerase.

Multiple cycles of PCR cover each bead with thousands of copies of the same DNA sequence.
For the geometry and bead size used in the device, the $\text{H}^+$ release results in a 0.02 unit change in pH.

This is measured with CMOS ion-sensitive field-effect transistor (ISFET).

**Include figure of hydrolysis of dNTP to produce Pi and $\text{H}^+$**
dNTPs washed away after each cycle and the change in pH (if any) is recorded per well to give the sequence.
Chips are made on a 0.35\(\mu\)m CMOS node. Well is 3.5\(\mu\)m diameter and includes a single bead. Proton sensing layer (Tantalum Oxide) gives a sensitivity of 58mV per pH.
ASIC and Fluidics
WORKFLOW

1. Prepare Genomic Library
2. Prepare Template on Bead
3. Sequence on Ion Chip
4. Signal Processing and Base Calling

DNA → Fragmented DNA → Size-Selected DNA → Adapted Library DNA

Thermocycle PCR
Sequencing with Nanochannels
3rd generation sequencing
Stochastic sensing using molecular pores

Protein pores - $\alpha$-haemolysin

Current fluctuates as ion passage is blocked by molecules.

The molecule to be detected is a transient blocker. It interacts with protein - binds and dissociates, hence stochastic current events.
DNA sequencing - principle

Signals from individual bases cannot be detected
12 nucleotides lie within the lumen of the haemolysin pore

Strand-sequencing using ionic current blockage. A typical trace of the ionic current amplitude through α-haemolysin pore clearly differentiates between an open pore (top right) and one blocked by a strand of DNA (bottom right).

It cannot distinguish between the ~12 nucleotides that simultaneously block the narrow transmembrane channel domain (red bracket)

[Graph showing current amplitude over time with open and blocked pores]
(a) Cross section of α-hemolysin heptamer in a lipid bilayer. 
(b) Capture of a single strand of poly A. 
(c) Ionic current baseline, and blockage caused by movement of oligomer.

120 mV across the open pore causes an ~120 pA ionic current (in 1 M KCl).

DNA translocating through the pore causes an abrupt reduction in current to about 14 pA. Each nucleotide traverses the channel in about 2 μs.
Concept for sequencing DNA by using a single protein pore. A single-stranded DNA (or RNA) molecule moves through the pore in the electric field. As it passes a “contact site”, each base produces a characteristic modulation of the amplitude in the single channel current.

The amount of current which can pass through the nanopore at any given moment varies depending on whether the nanopore is blocked by an A, a C, a G or a T (in theory).
DNA detection by modification of pore

A - unmodified pore
B - current fluctuations with oligo A bound in pore, as oligo B passes and partly binds.
S events are transients from B not binding to A
C – flooding solution with oligo A mops up oligo B. Current events cease
Issues:

Capturing the DNA molecules in solution into the pore

Single nucleotide resolution requires a 0.3nm long pore, ion-current blockades are ~10–15 nucleotides

Single-strands are translocated at ~1 nucleotide/μs with ~150 mV bias.

Resolving single bases with small sub-pA currents means slowing the translocation so each base occupies the nanopore detector for ≥1 msec.
Exonuclease-sequencing by modulation of the ionic current.

An exonuclease is attached to the top of the α-hemolysin pore through a genetically encoded (deep blue), or chemical linker. The enzyme sequentially cleaves nucleotides (gold) off the end of a DNA strand (in this case, one strand of a double-stranded DNA).

The identity (A, T, G or C) is determined by the level of the current blockade it causes when driven into an cyclodextrin adaptor (red) lodged within the pore. After a few milliseconds, the nucleotide is released and exits on the opposite side of the bilayer.
Modified pore together with the exonuclease has recently been shown to enable nucleotide discrimination.

For practical sequencing the exonuclease must be in close proximity to the nanopore sensor; (chemical attachment or a genetic fusion of nanopore and exonuclease).

Location of the exonuclease active site must be manipulated to align the trajectory of the released nucleotide with nanopore entrance.

It may be necessary to vary the rate of the enzyme, either by mutation or changes to physical conditions such as temperature, to ensure that a nucleotide is recorded before the next base enters the nanopore.
Pores are set in an electrically insulating polymer membrane, not a BLM.

Each membrane patch sits on top of an ASIC that records the current spikes (pA).

Illumina working on similar technology.
Other approaches

Si nanopore

Graphene membranes

Al2O3 pore
Summary

- NGS - high-throughput, parallel, rapid DNA sequencing
- 3rd Generation – single molecule, real time, reduced chemistry, portable

Basic workflow:

Applications – *de novo* DNA seq, RNA seq. Transcriptomics, Single Cell…
Further Reading
Step 1 Template Binding

Template DNA fragments (green and pink) first ligated to adaptor molecules (orange and red) in a test tube.

These are then denatured to form single stranded DNA.

This is hybridized to complementary capture oligonucleotides stuck on the surface of the flow chip.
Step 2 Bridge Amplification

A special kind of PCR amplification called BRIDGE AMPLIFICATION generates clusters of multiple clonal copies of the template on the surface of the flow cell. This process results in the generation of 100s of millions of individual clusters containing over one thousand copies of amplified DNA molecules on the surface of the chip.

**Bridge amplification:** The two ends of the hybridised template DNA interacts with nearby primers and amplification then takes place.

**Cluster generation:** After several round of PCR 100-200 million clonal cluster are formed.
A mixture of primers, DNA polymerase and **modified nucleotides** are added to the flow cell.

Each nucleotide is **blocked** (like in Sanger) and labelled with a **cleavable** fluorescent molecule (F).

During each cycle, the fragments in each cluster will incorporate just **one nucleotide** as the blocked 3’ group prevents additional incorporations.
Imaging:
Unincorporated bases are washed away and the slide imaged with a laser.
Colour identifies which base was incorporated in each cluster.

Cleavage:
Dye is cleaved and the 3’-OH is regenerated.
The cycle of nucleotide addition, elongation and cleavage is repeated.

Step 4 Image and repeat