UNIVERSITY OF CAMBRIDGE

TRINITY HALL
CAMBRIDGE

Wellcome Trust Advanced Course - RNA Transcriptomics 2023

Anton Enright

Group Leader, RNA Genomics Laboratory

Academic Lead (Genomics),

Dept. Of Pathology

University of Cambridge

Staff Fellow (Pathology), Director of Studies, Graduate Tutor, Trinity Hall (Cambridge)

aje39@cam.ac.uk

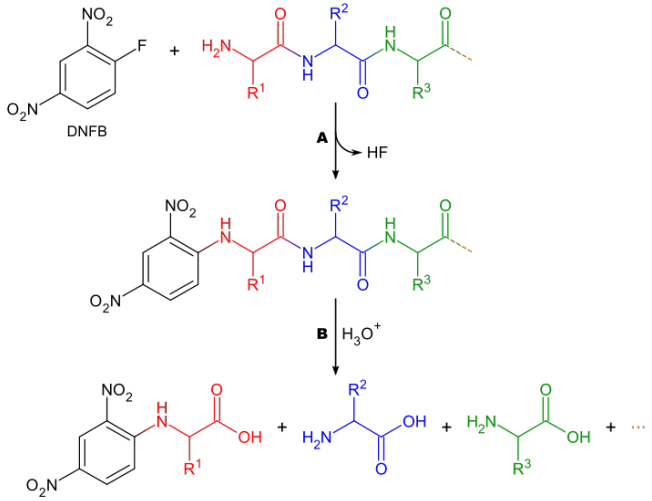
Department of Pathology

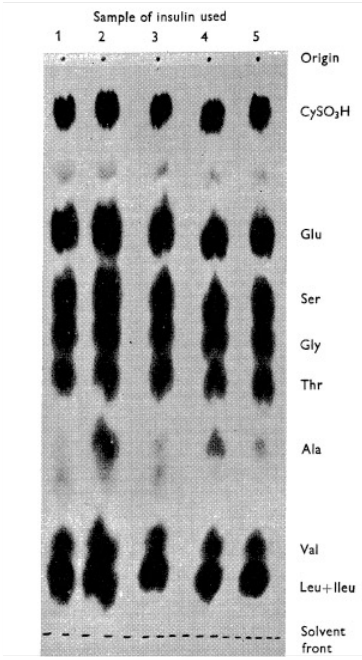
Overview - Computational Analysis

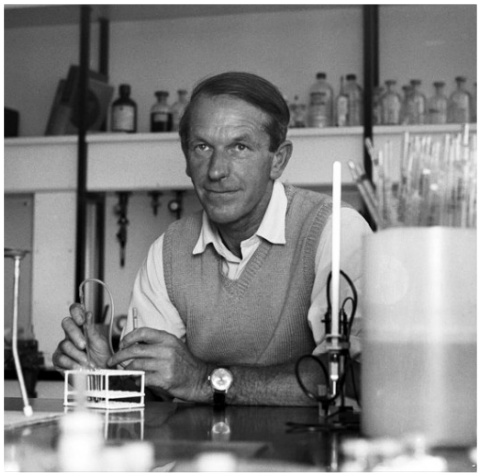
- **Course Modules and Analysis**
  - Background - Sequencing, Genomics, Transcriptomics
  - Main Biological Focus
  - Analysis Framework
  - Different Protocols Used
  - Goals

Sequencing in molecular biology

- Birth of sequencing - Fred Sanger 1951
- Proteins thought to be quite amorphous
- Looking for composition of amino acids in Bovine Insulin
- N-terminal Labelling then hydrolysis
- Chromatographic separation



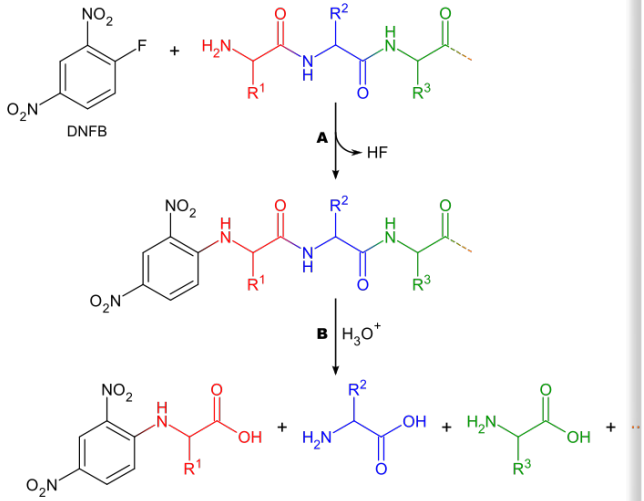


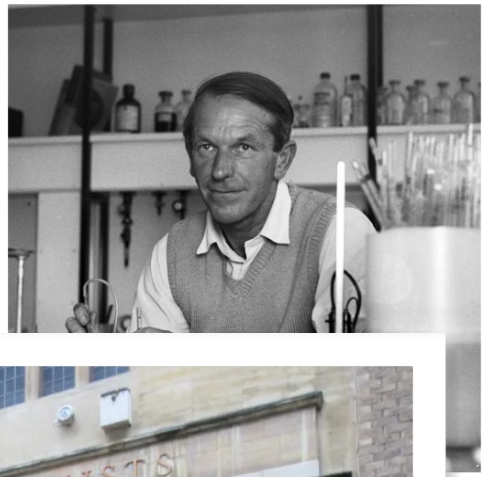


Fred Sanger
Biochemistry Dept circa 1952

Sequencing in molecular biology

- Birth of sequencing - Fred Sanger 1951
- Proteins thought to be quite amorphous
- Looking for composition of amino acids in Bovine Insulin
- N-terminal Labelling then hydrolysis
- Chromatographic separation

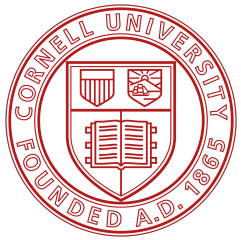


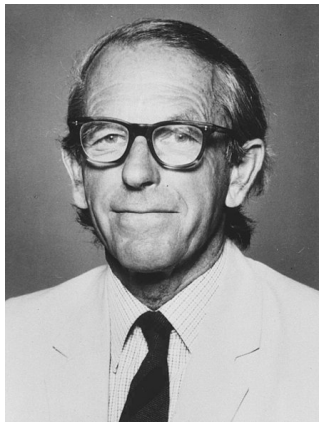


Sequencing Nucleotides - RNA was first

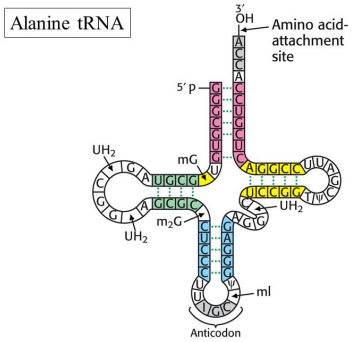
- Double Ribonuclease digestion of RNA molecules - 1965, 1967

Robert Holley
Cornell

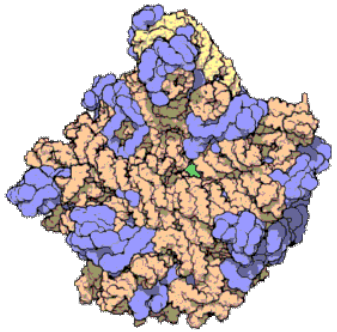




Fred Sanger
LMB

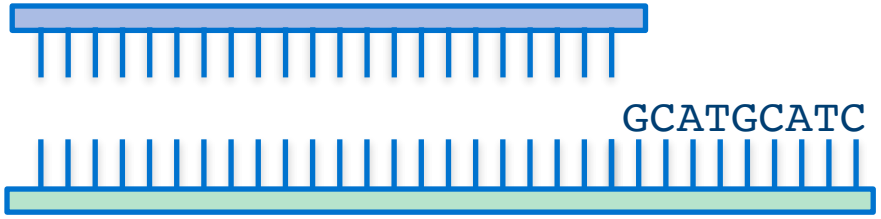


Alanine tRNA



5s Ribosomal RNA

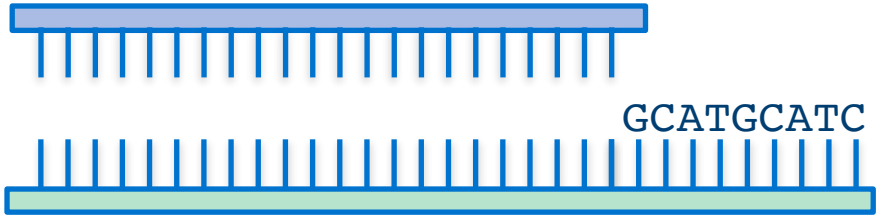
Chain Termination Sequencing - Sanger Sequencing

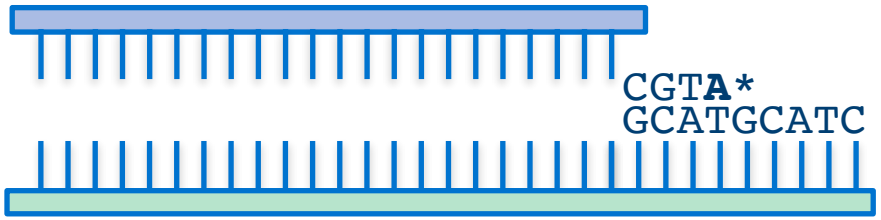


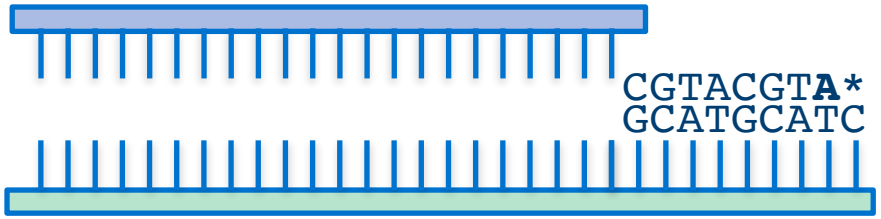
Reaction for ddA

A
T
G
C
A*

Chain Termination Sequencing - Sanger Sequencing

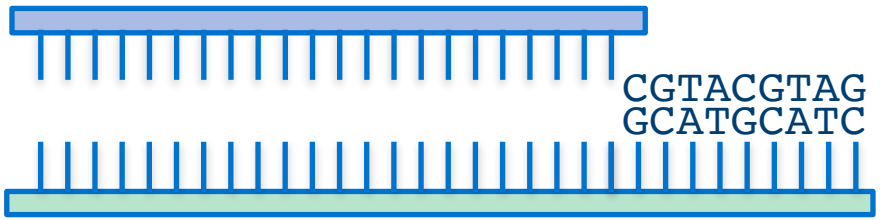




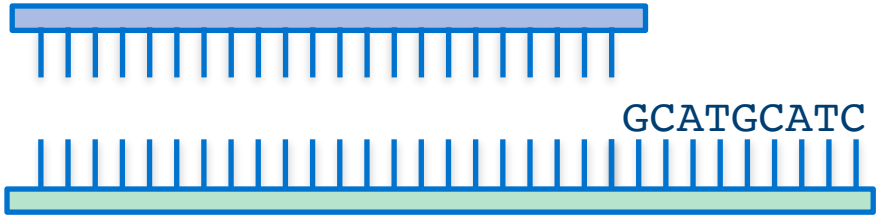


Reaction for ddA

A
T
G
C
A*



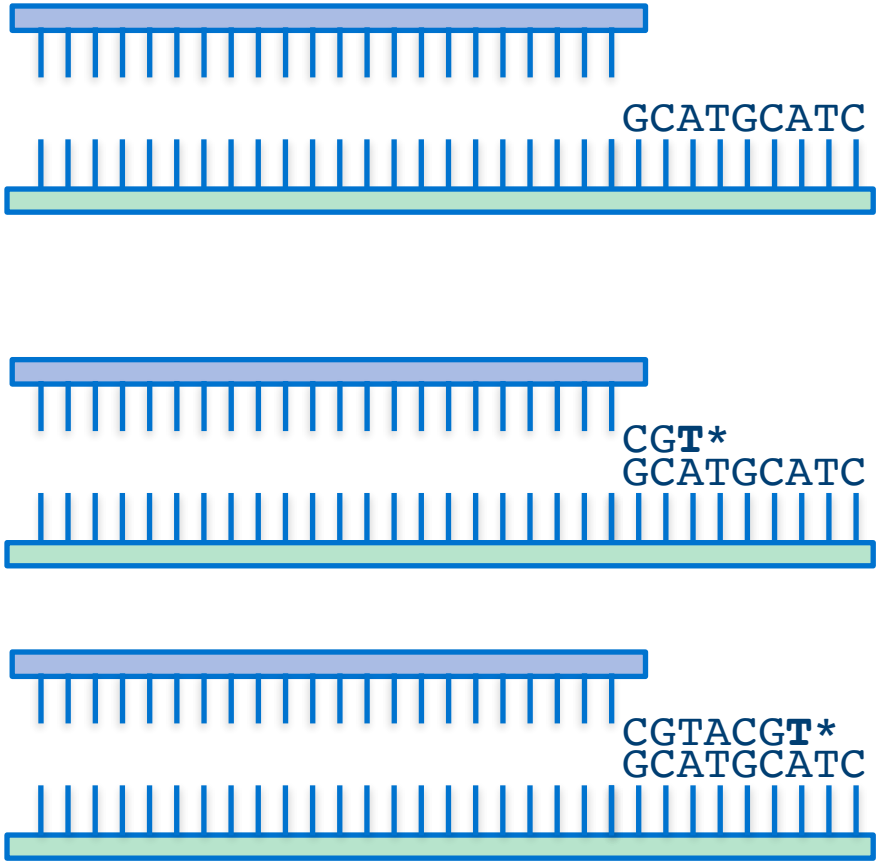
Chain Termination Sequencing - Sanger Sequencing



Reaction for ddT

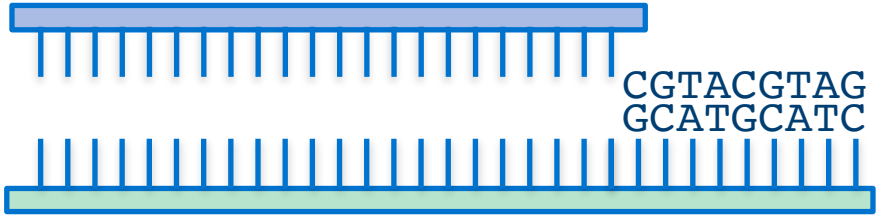
A
T
G
C
T*

Chain Termination Sequencing - Sanger Sequencing

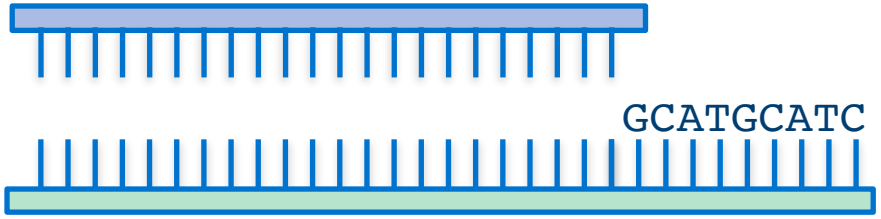


Reaction for ddT

A
T
G
C
T*



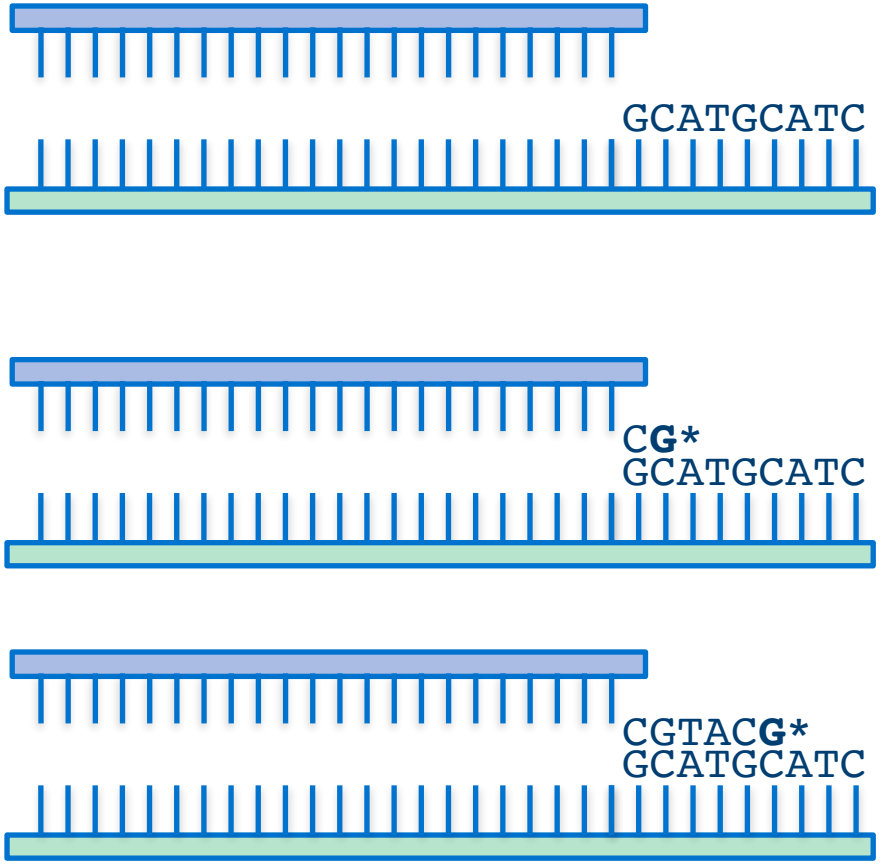
Chain Termination Sequencing - Sanger Sequencing



Reaction for ddG

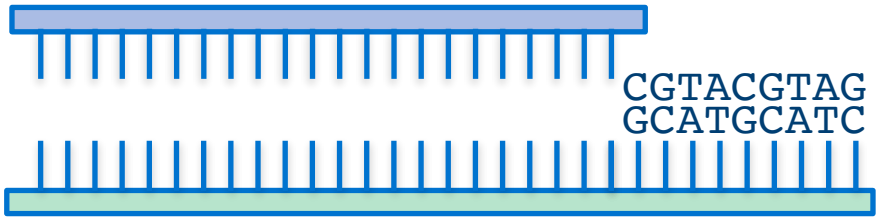
A
T
G
C
G*

Chain Termination Sequencing - Sanger Sequencing

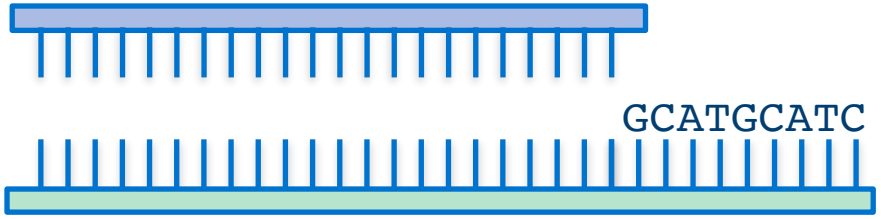


Reaction for ddG

A
T
G
C
G*



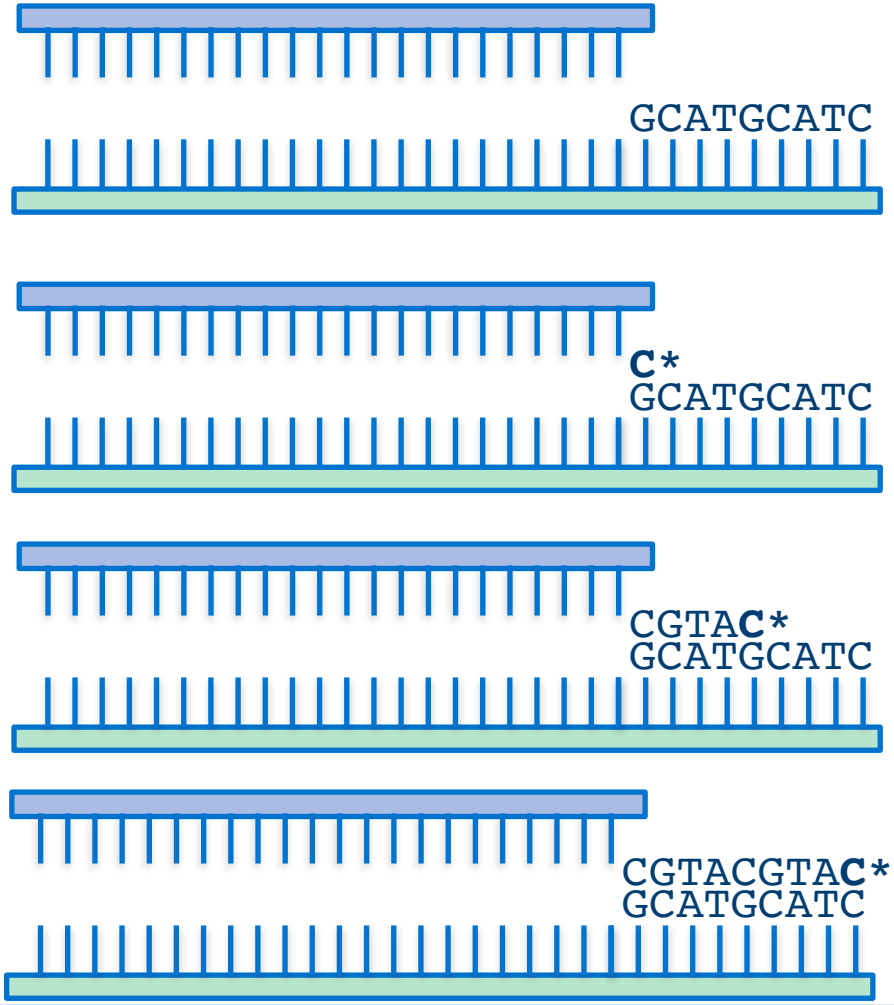
Chain Termination Sequencing - Sanger Sequencing



Reaction for ddC

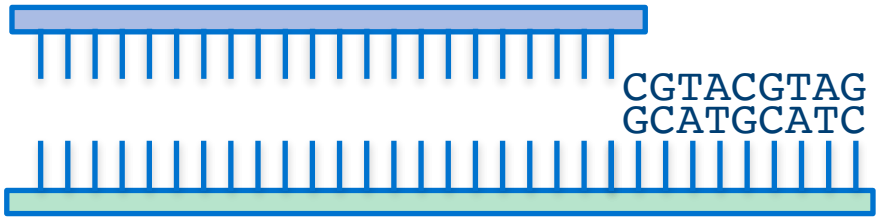
A
T
G
C
C*

Chain Termination Sequencing - Sanger Sequencing



Reaction for ddC

A
T
G
C
C*



Chain Termination Sequencing

CGTA*
CGTACGTA*

CGT*
CGTACGT*

C*
CGTAC*

CG*
CGTACG*
CGTACGTAG*

Chain Termination Sequencing

```
CGTA*
CGTACGTA*

CGT*
CGTACGT*

C*
CGTAC*

CG*
CGTACG*
CGTACGTAG*
```

```
C*
CG*
CGT*
CGTA*
CGTAC*
CGTACG*
CGTACGT*
CGTACGTA*
CGTACGTAG*
```

Chain Termination Sequencing

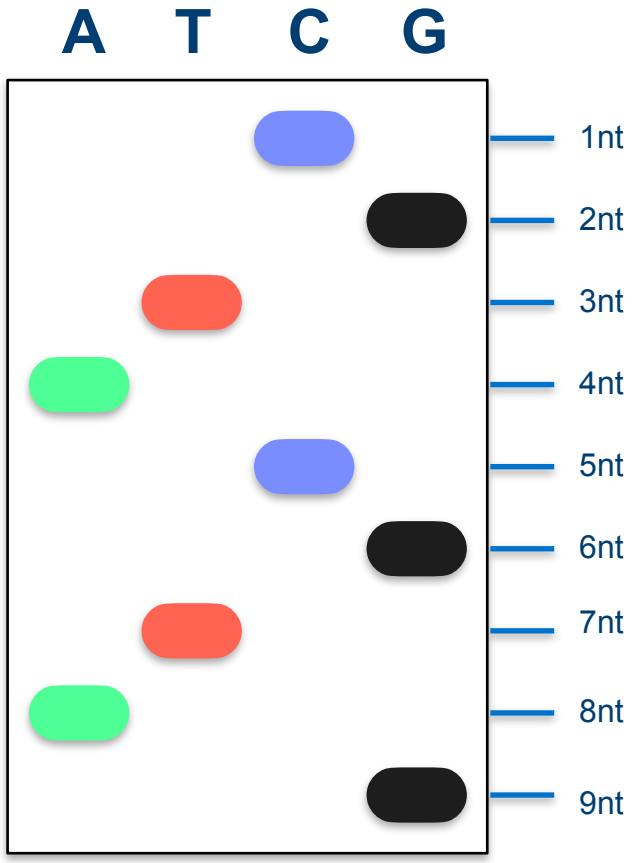
CGTA*
CGTACGTA*

CGT*
CGTACGT*

C*
CGTAC*

CG*
CGTACG*
CGTACGTAG*

C*
CG*
CGT*
CGTA*
CGTAC*
CGTACG*
CGTACGT*
CGTACGTA*
CGTACGTAG*



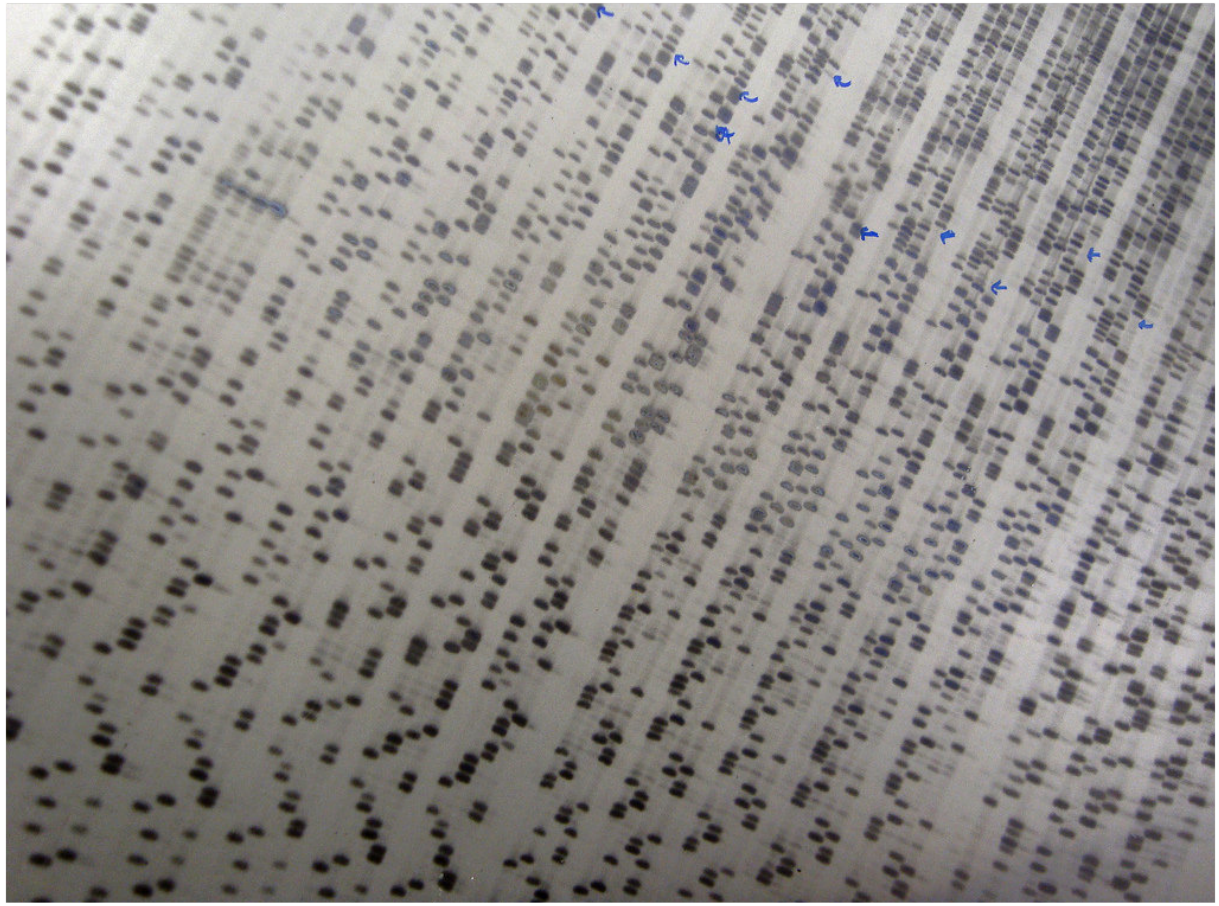
Chain Termination Sequencing

CGTA*
CGTACGTA*

CGT*
CGTACGT*

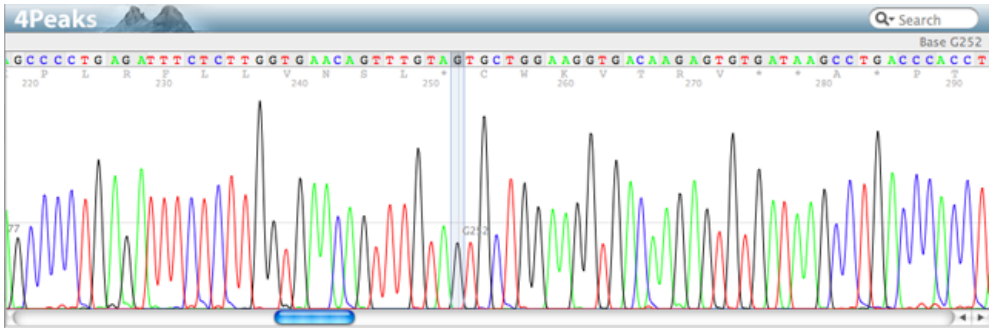
C*
CGTAC*

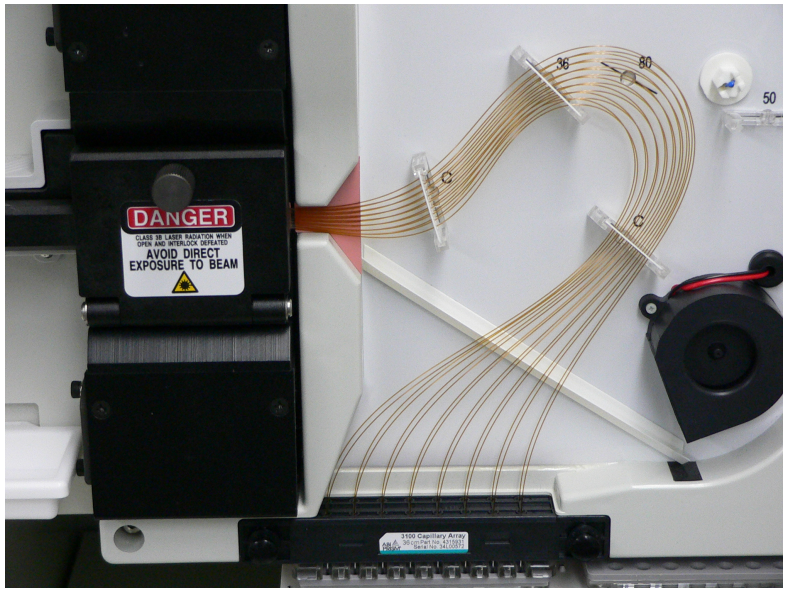
CG*
CGTACG*
CGTACGTAG*



Enabling Genomics: Sanger Sequencing to the next level

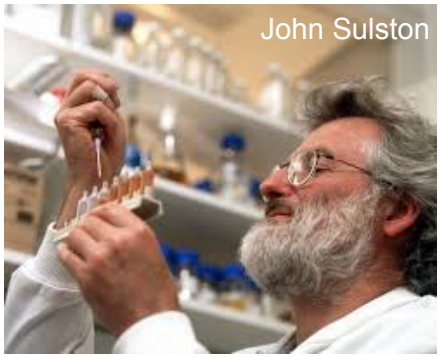
- Cambridge was part of a massive project to sequence the first human genome
- There are 3.2 Billion nucleotides (A,T,G and C) in the sequence
- The first genome took 10 years and cost $ 3 Bn





The Genome Factory in Cambridge

Wellcome Trust Genome Campus in Hinxton



Sanger Centre, opened in 1992
European Bioinformatics Institute shortly after to support the effort to release the data

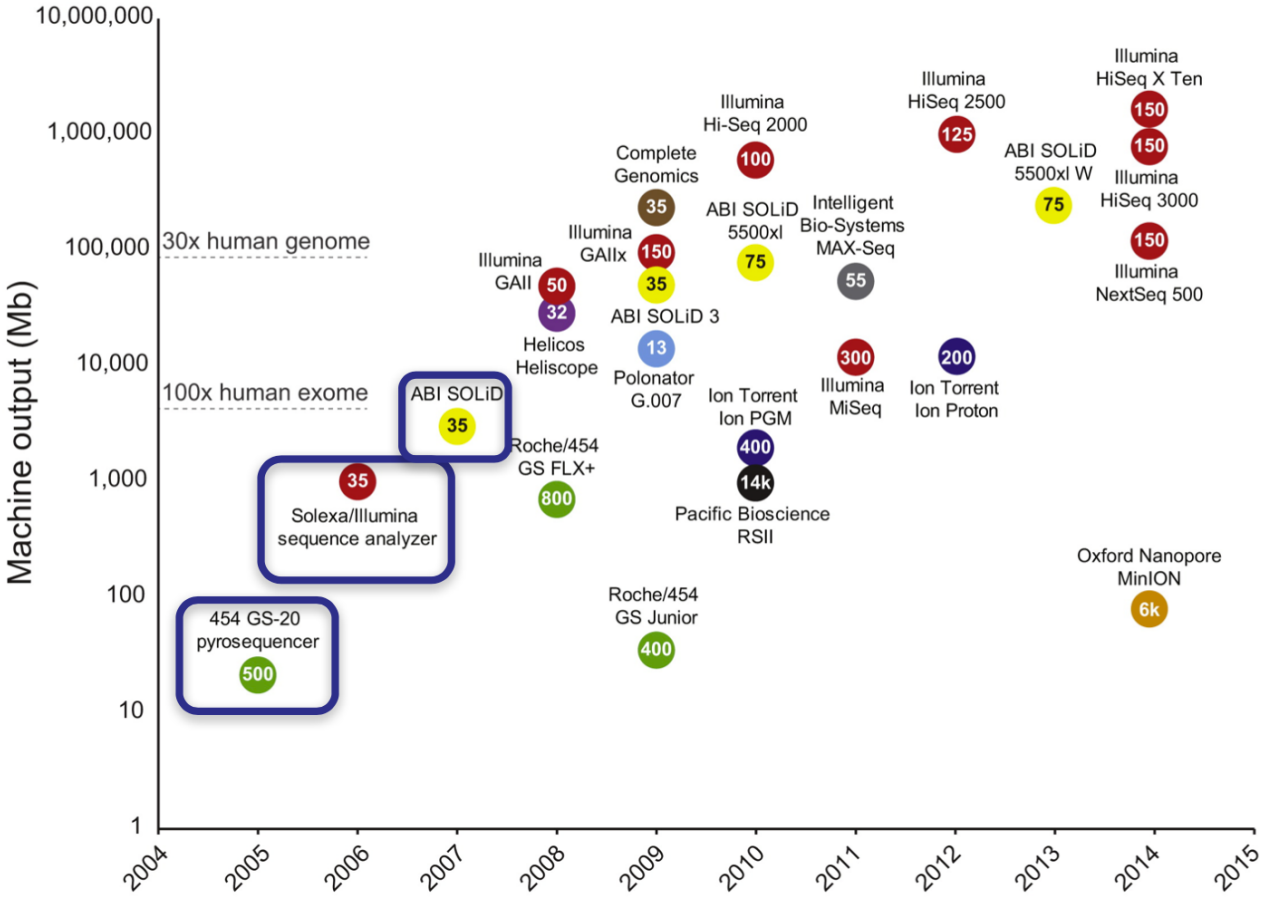
Sanger Centre, opened in 1992

European Bioinformatics Institute shortly after to support the effort to release the data

A paradigm shift

- The original technology to produce the first genome (early 2000s) obsolete
- Since 2010, significantly faster and cheaper approaches available
- The first human genome cost $3Bn and took 10 years
- A full human genome now costs about £600-900 and takes 1-2 days

The Next Generation - High Output, Low Cost

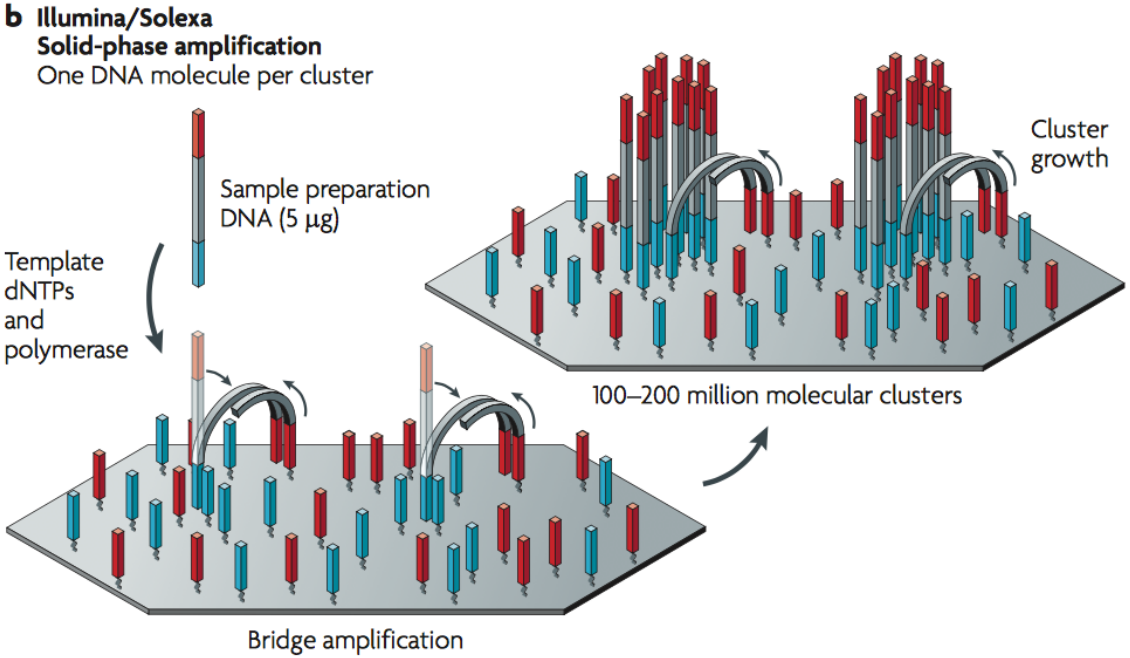


Reuter JA, Spacek DV & Snyder MP, Mol Cell. 2015 May 21;58(4):586-97.

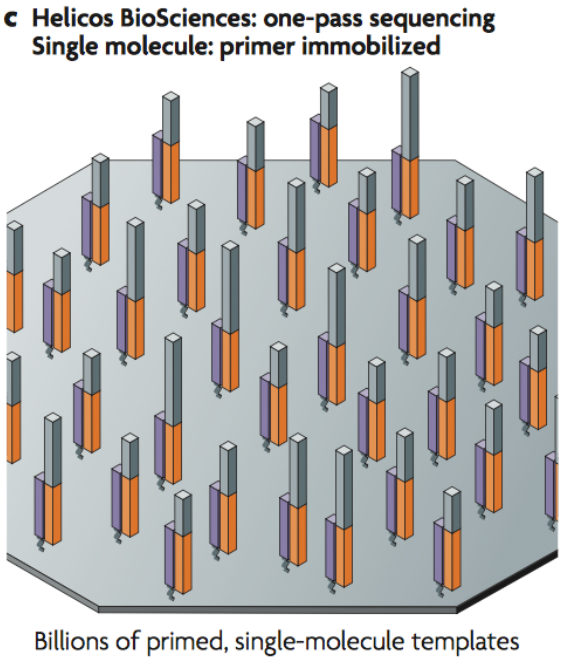
Solid Phase Immobilisation

- Solid Phase methods immobilisation

Images from: Metzker ML, Nat Rev Genet. 2010 Jan;11(1):31-46.



Illumina Sequencing (PCR)

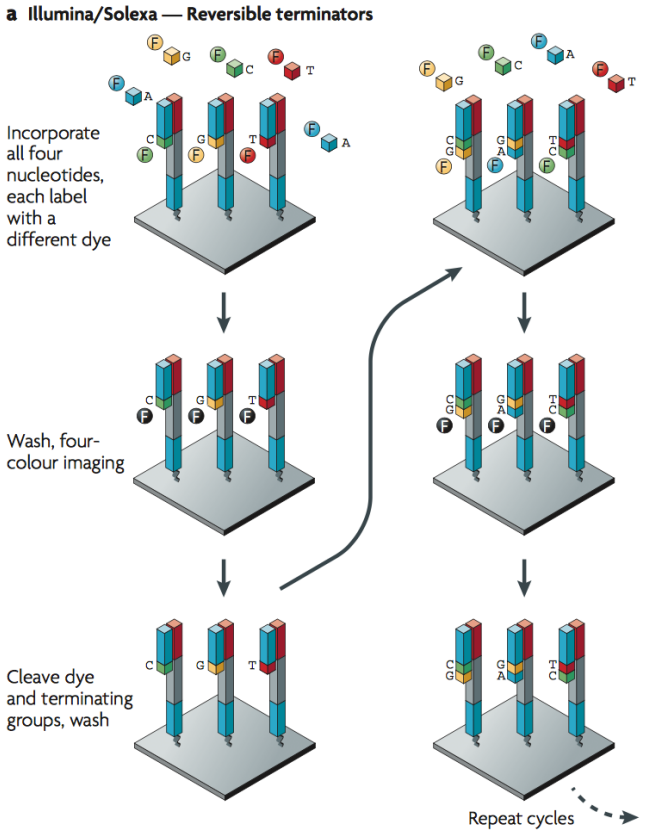


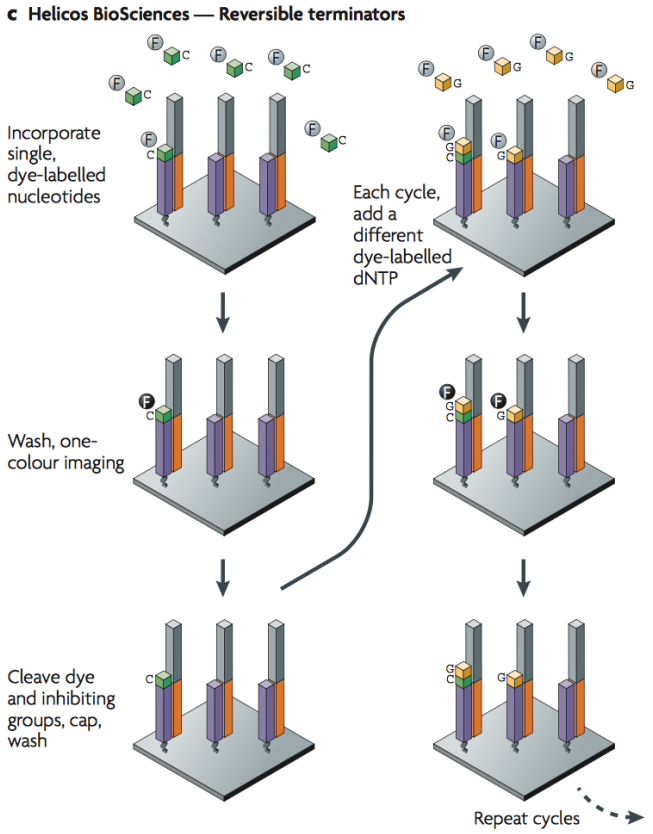
Helicos (Single Molecule)

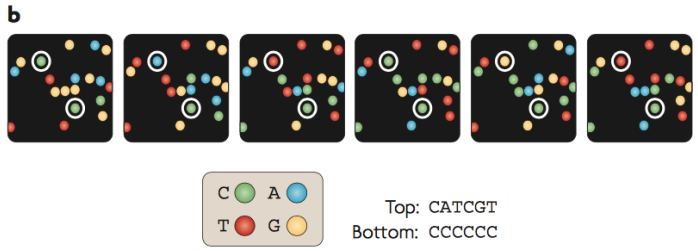
Illumina - Reversible terminator sequencing

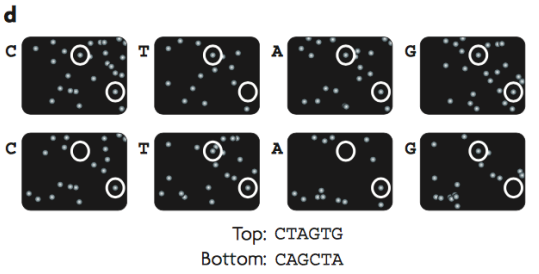
Images from: Metzker ML, Nat Rev Genet. 2010 Jan;11(1):31-46.

- Nucleotides are flourescently labelled
- They contain terminators attached to the flourophore
- After all four bases are incorporated by a polymerase the clusters are laser imaged
- These terminators are removed after a cycle and the next base is added



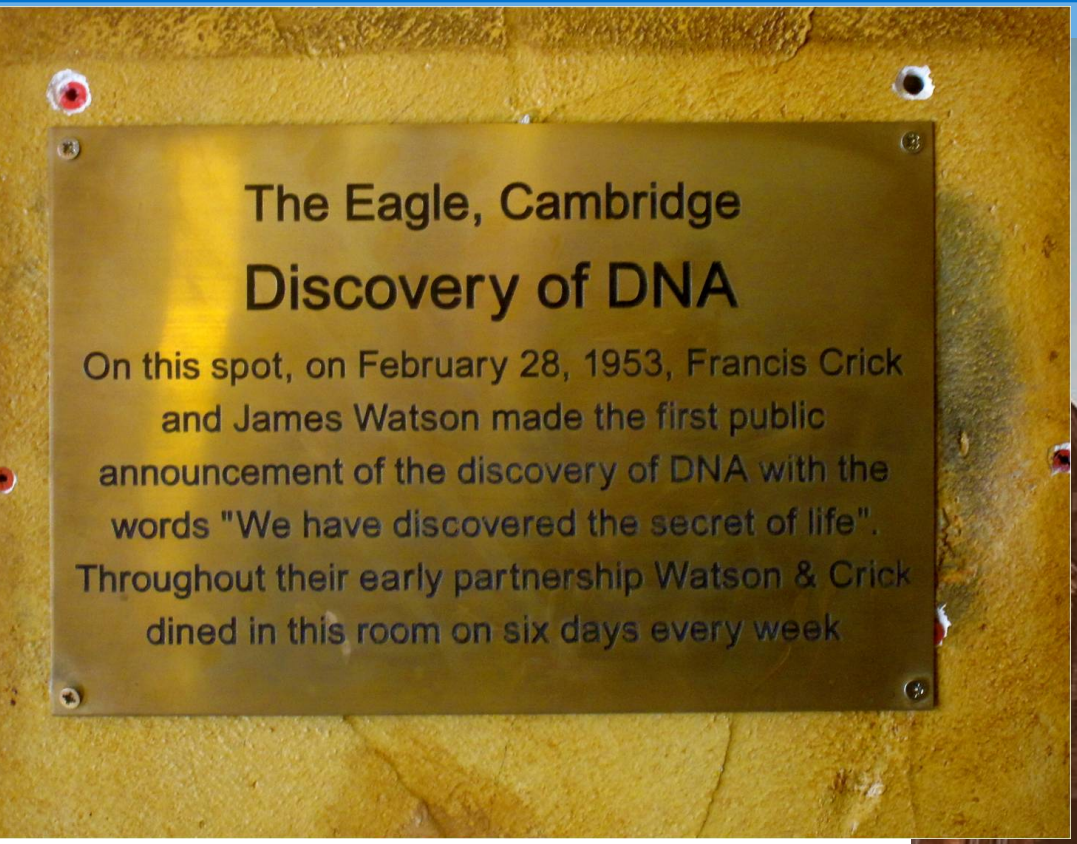






UNIVERSITY OF CAMBRIDGE

Genome Pubs of Cambridge - Eagle and Panton Arms

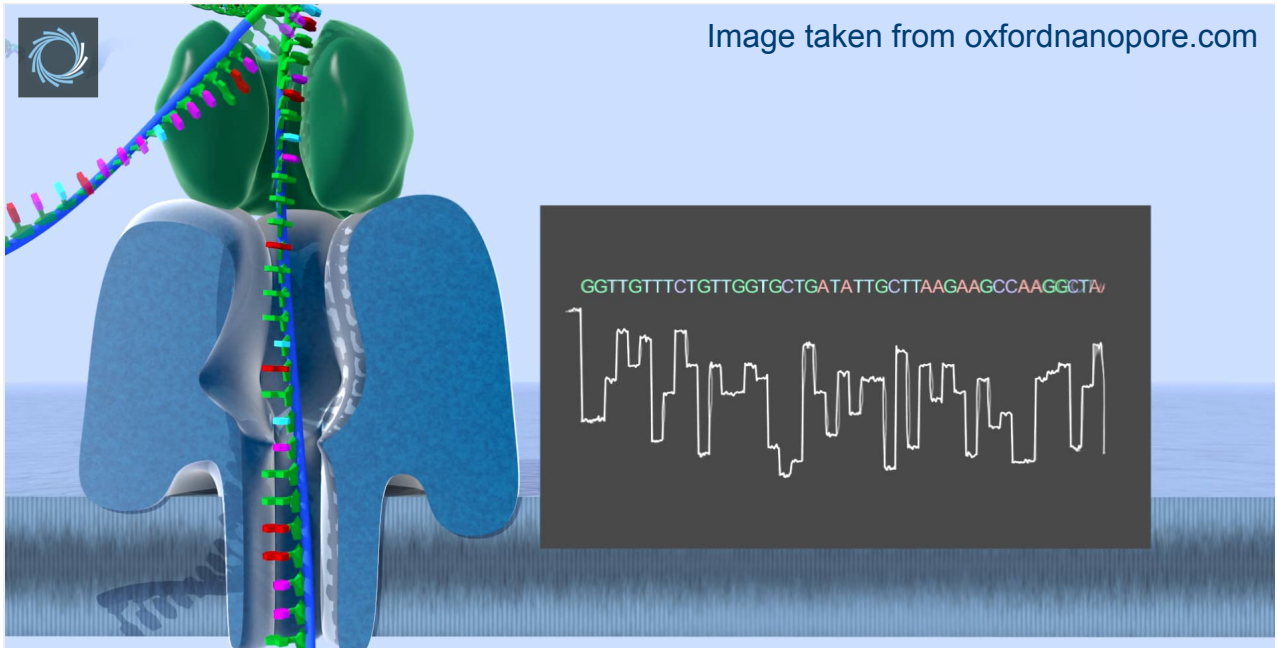


Genome Pubs of Cambridge - Eagle and Panton Arms

4th Gen Approaches to sequencing

Cheap single molecule sequencing with Nanopores

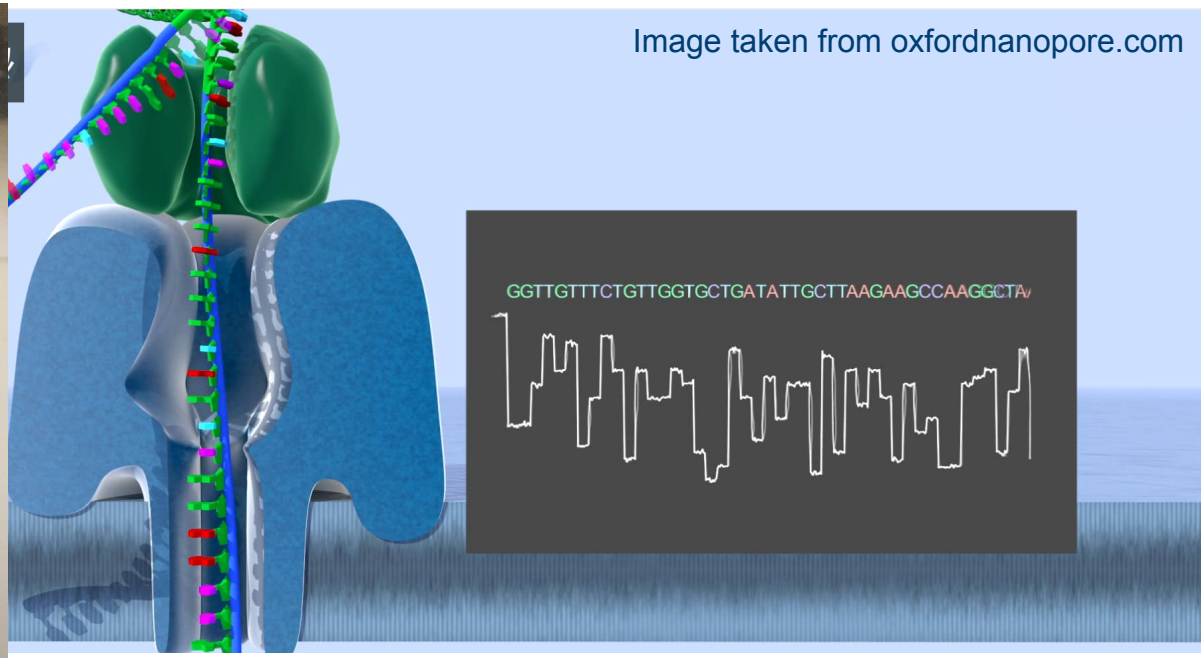
- gDNA fragmented and adapters attached
- gDNA ligated to a molecular motor
- Molecular motor attaches to a protein pore across an artificial membrane
- Voltage across each pore is measured in real-time
- Nucleotides can be called based on current changes



- Very long reads, 50-100kb possible
- Currently relatively high error rates
- Low cost, tiny machine

4th Gen Approaches to sequencing

Cheap single molecule sequencing with Nanopores



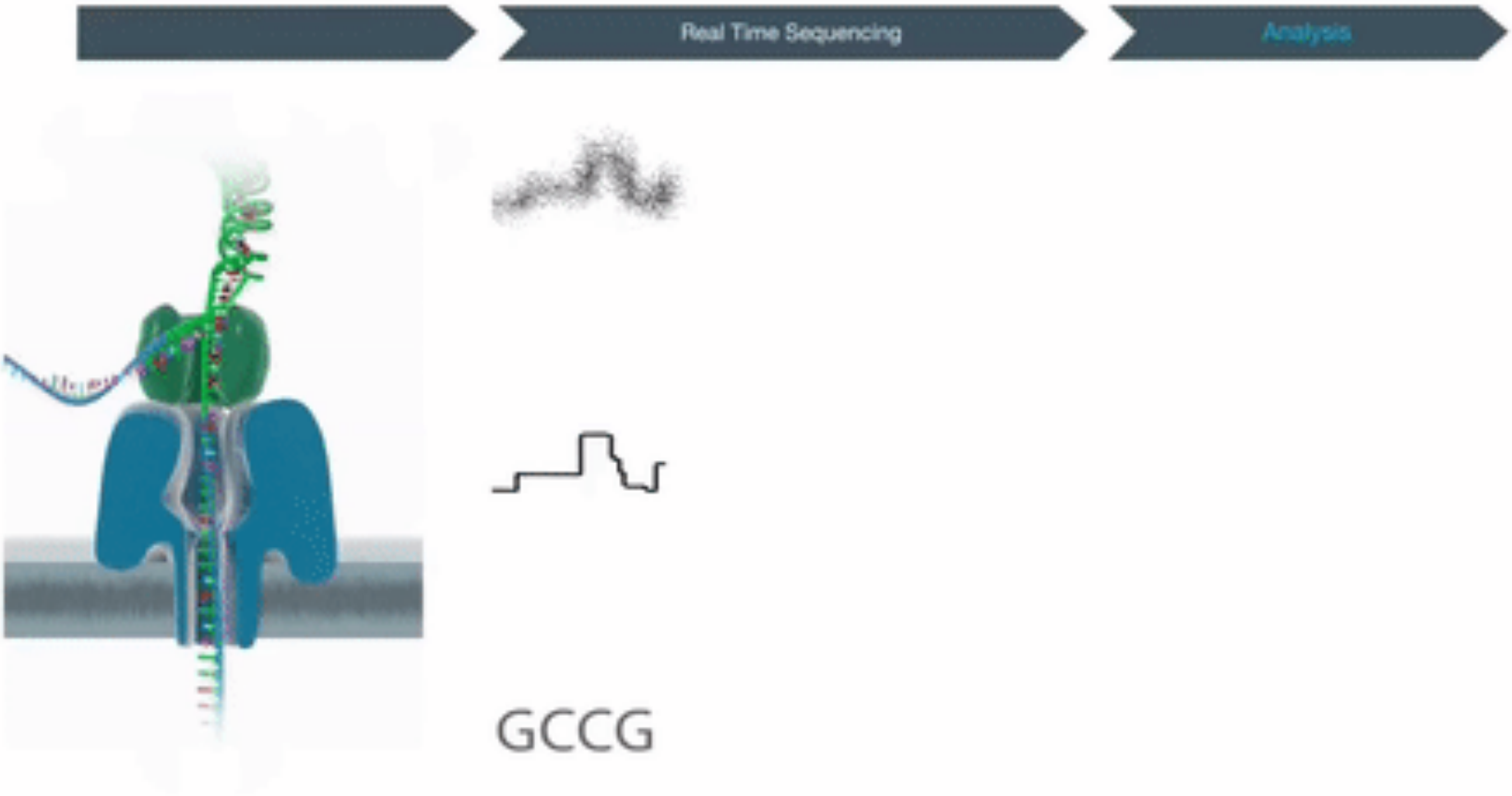
- Very long reads, 50-100kb possible
- Currently relatively high error rates
- Low cost, tiny machine

Oxford Nanopore PromethION

School of Biological Sciences PromethION (CGS, Enright Lab)

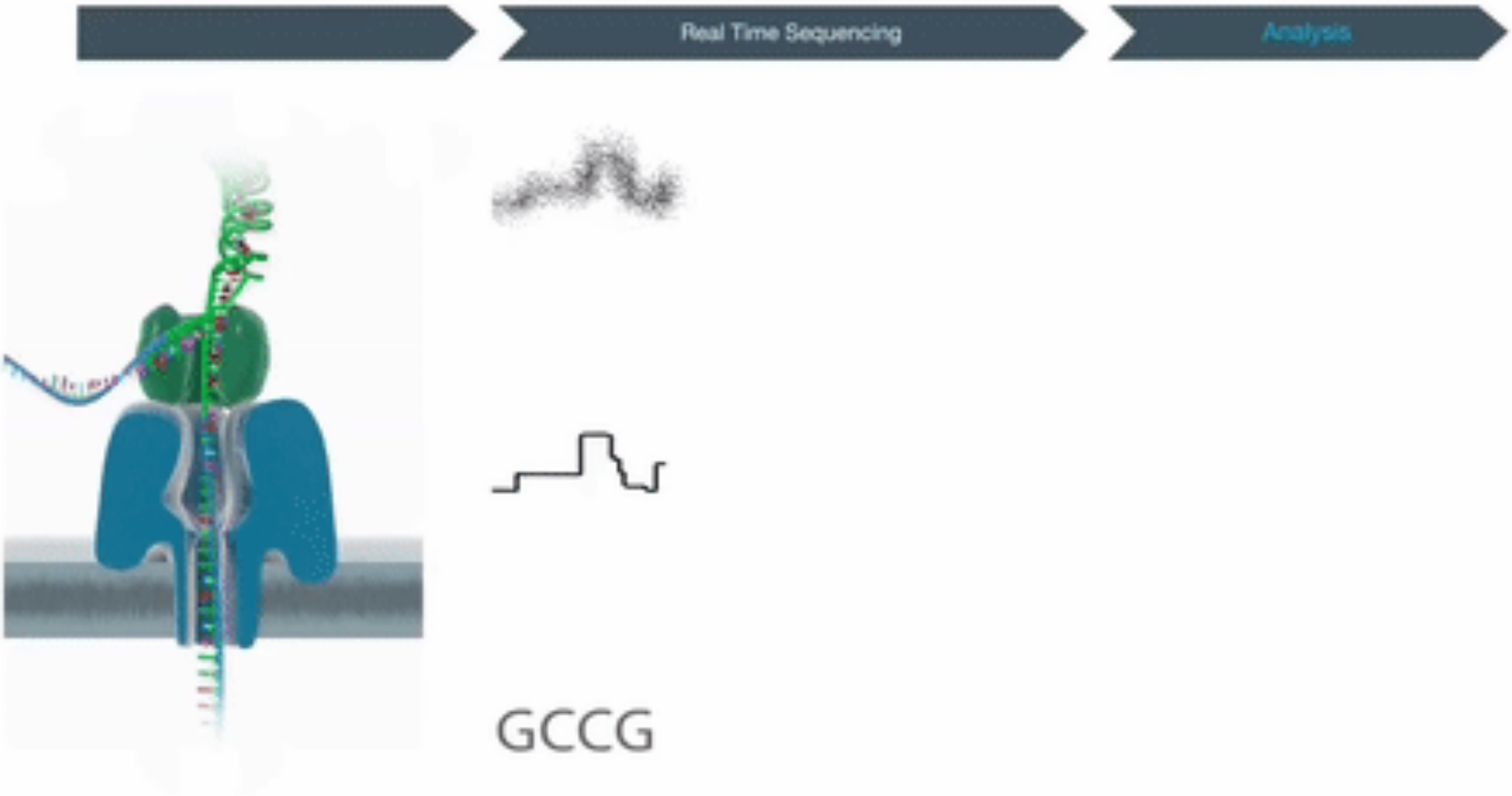
Nanopore in action

This is lies….. Signals look nothing like as clean as this

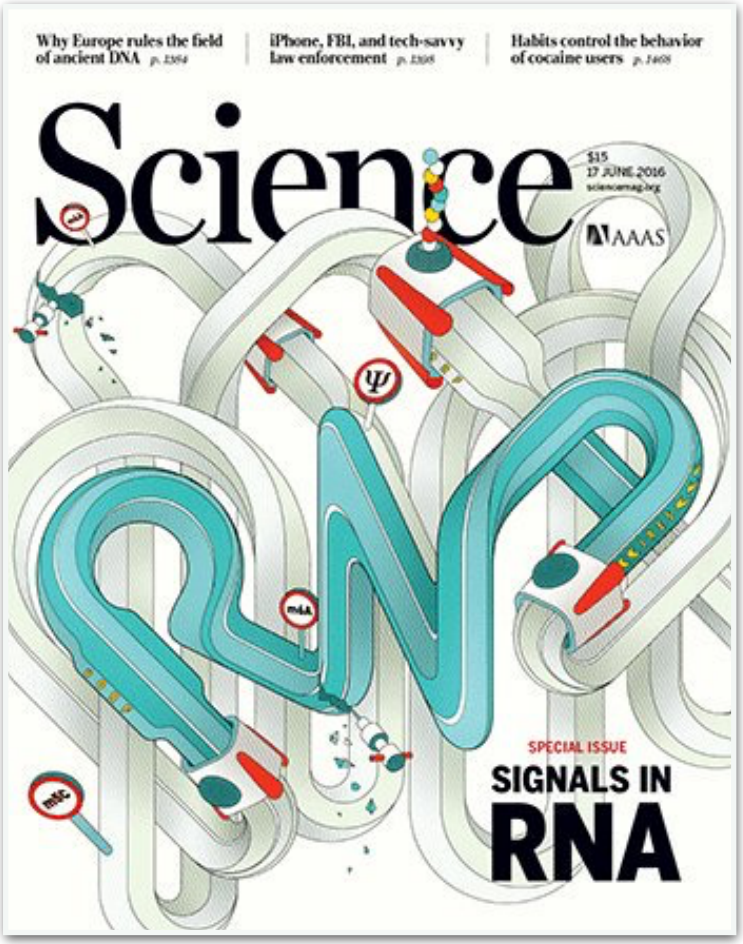


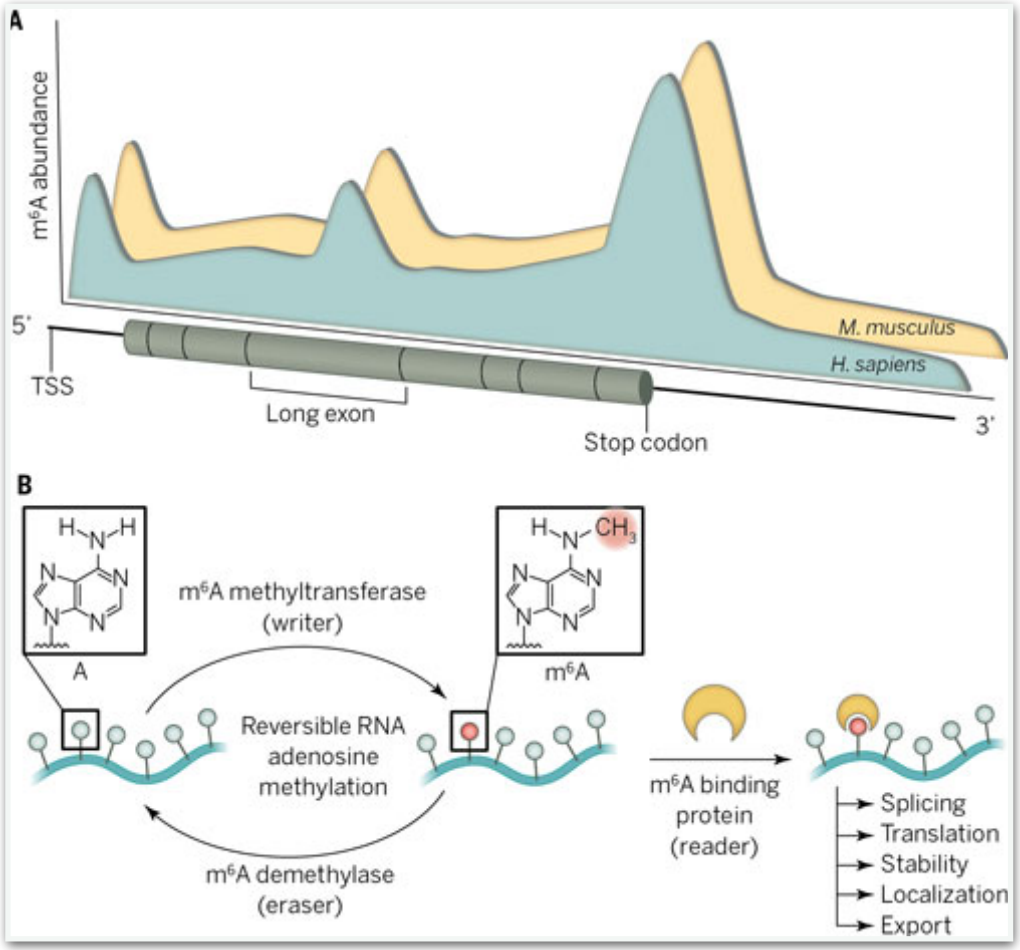
Nanopore in action

This is lies….. Signals look nothing like as clean as this



A New Frontier - RNA Epitranscriptomics






Alzheimers (m6A)
ER- Breast cancer (m6A)
ADHD (m6A)
Melanoma (m6A)
Multiple sclerosis (m6A)
Longevity (m6A)
Mental retardation (m5C)
Protein translation, localisation
Splicing
Structural Scaffolds
Chromatin recruitment
Ribozymes
Environmental Sensing
Post Transcriptional regulation
Gene silencing
Germaine defence
Intercellular signalling
RNA modification

WTAC RNA 2023

- **Experimental samples from Alena Shkumatava**
- Neural progenitor cells cultured from ES cells (mouse)
- miR-29 microRNA binding site mutation versus wildtype

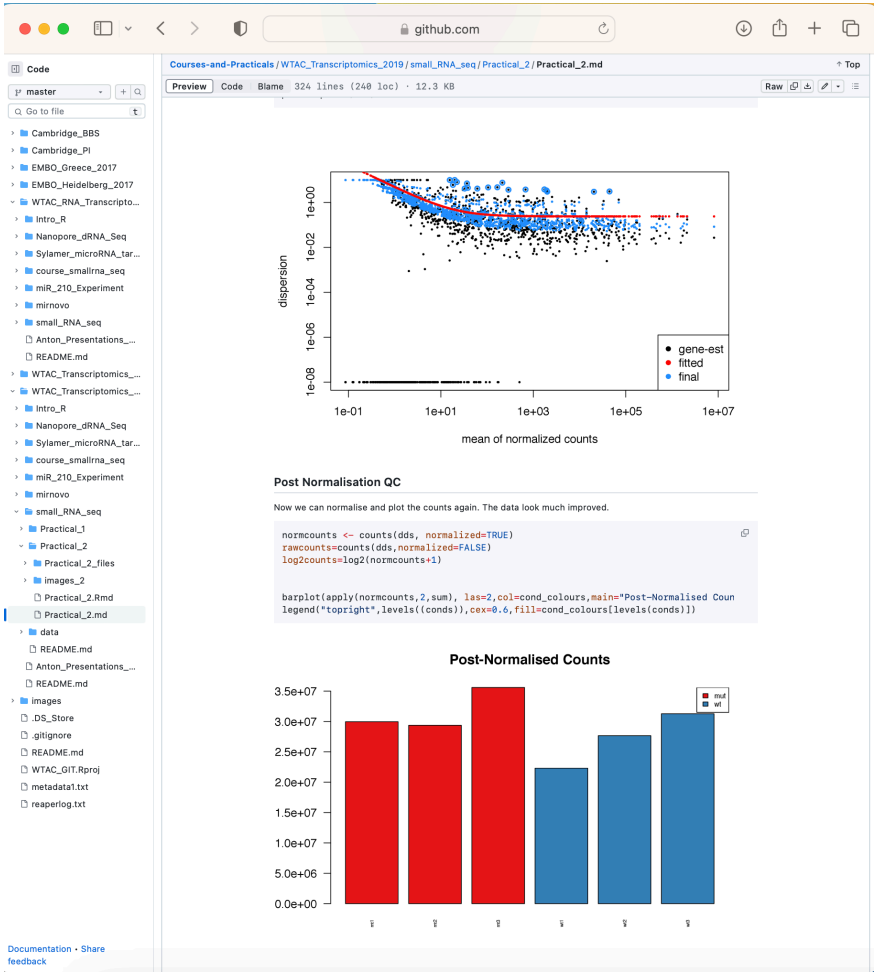
- cDNA Sequencing (Nanopore)
- Direct RNA Sequencing (Nanopore)
- Single Cell Sequencing (10X, illumina)
- Small RNA sequencing (illumina)

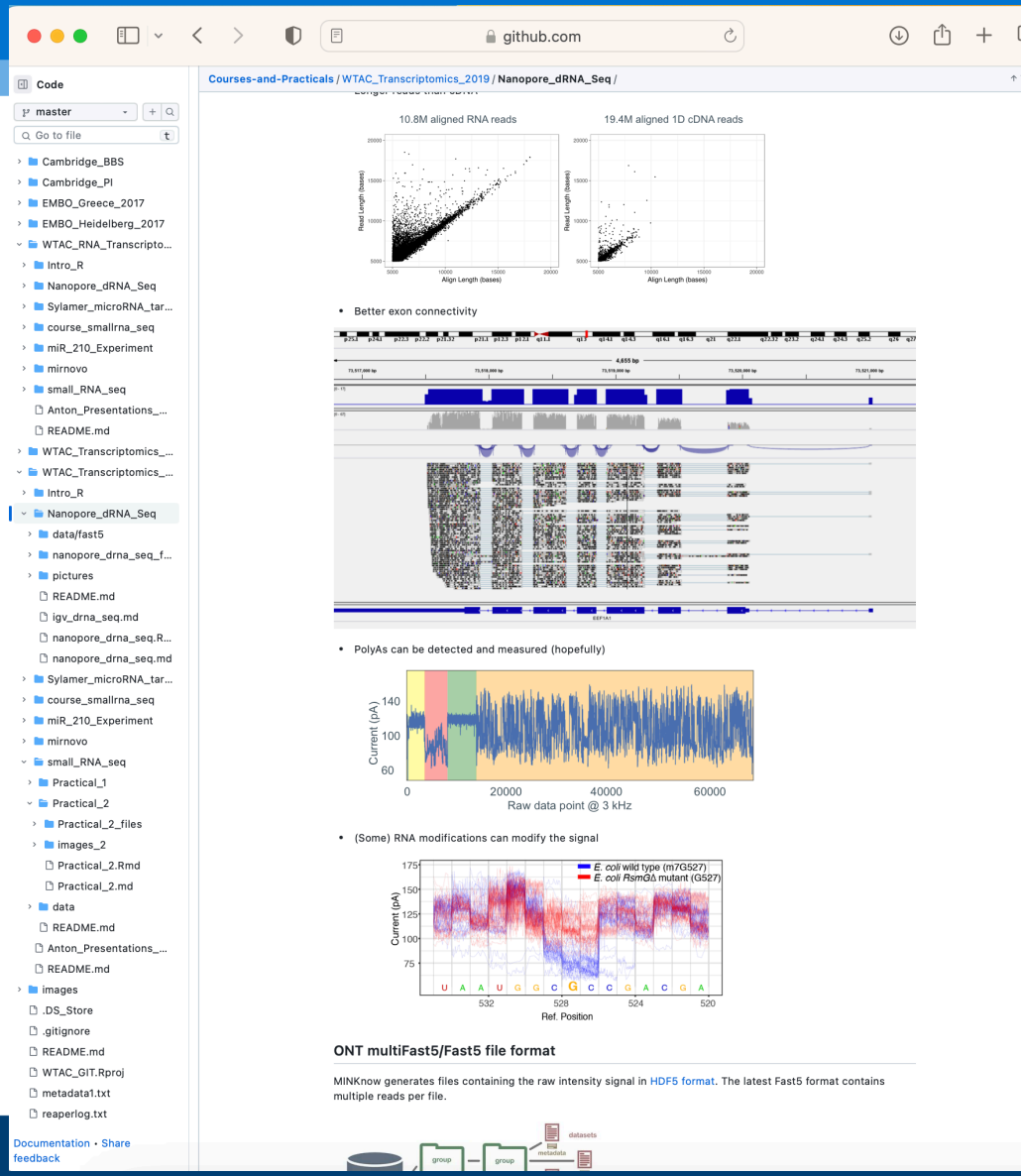
Computational Materials

- Computational Protocols available indefinitely
  - www.tinyurl.com/wtacrna2023
    - GitHub and WWW

- Computational Practicals performed on virtual machines in training room
- Materials, Protocols and VMs available after the course
- You can repeat the analyses yourself at home or adapt them to your own data with ease.

Some examples (2019 data)



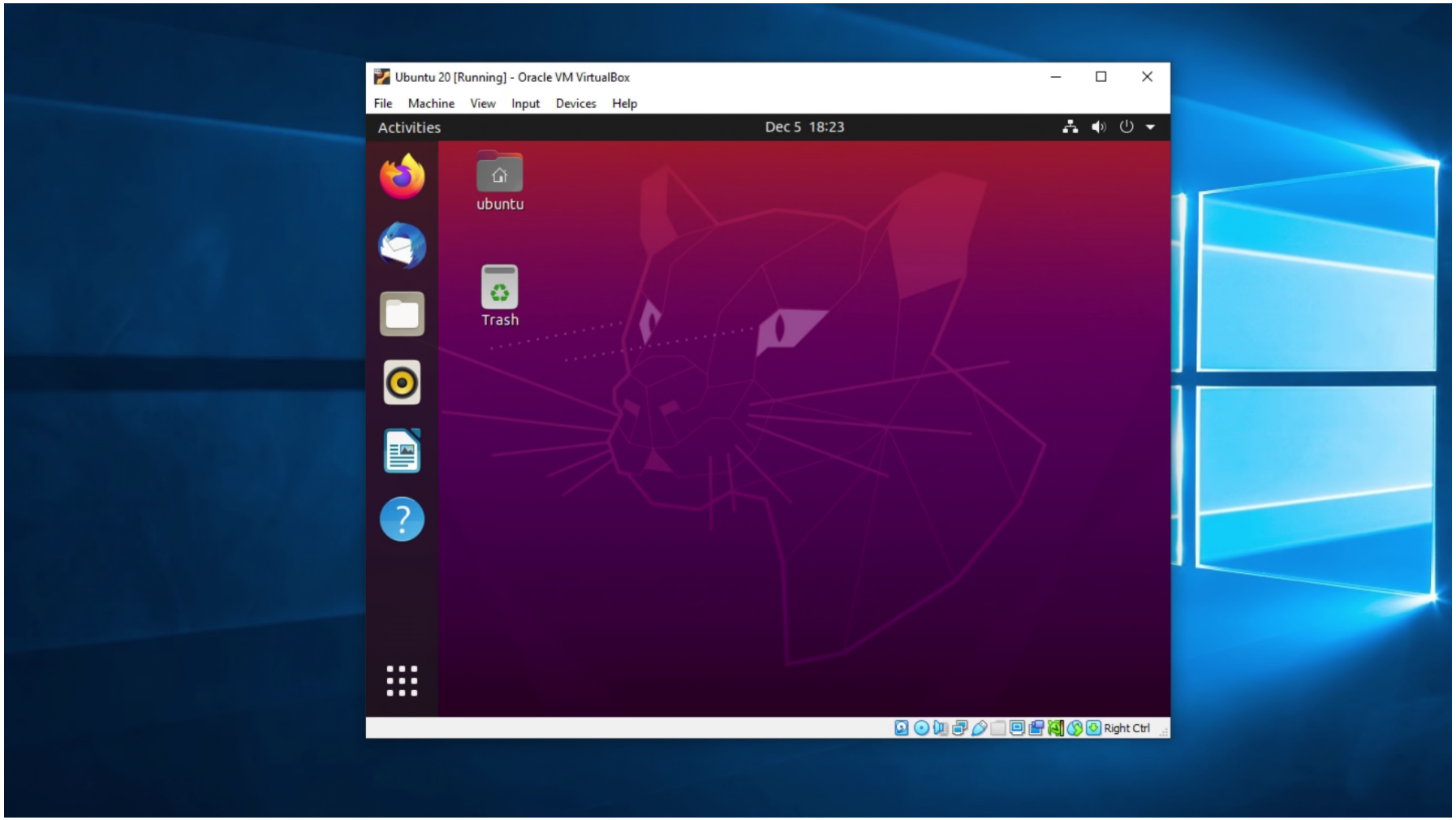


UNIVERSITY OF CAMBRIDGE

VirtualBox

- VM software compatible with Linux, Windows and Mac
- Allows you to run a virtual computer inside your own computer
- Alternatives: conda, docker
- www.virtualbox.org

VM in use


Ubuntu 20 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Activities
Dec 5 18:23
ubuntu
Trash
Right Ctrl

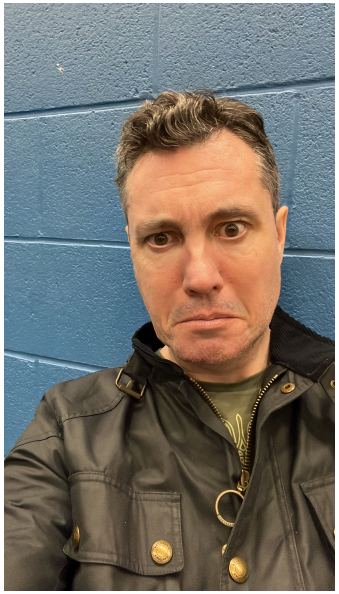
Why R and Bioconductor ?

- **Freely available**
- **Updated constantly**
- Available here:
  - http: //www. r-project. org/
  - http: //www. bioconductor. org/
  - LIMMA user guide and the Springer book series are well worth investigating
  - DESeq2 Vignette and Manual Pages excellent also
- **Full R courses available (e.g. data camp)**

Computational Training

- **Introduction to R/BioConductor**
- Small RNA NGS sequencing data analysis (Anton)
- mRNA NGS sequencing data analysis (Francesca)
- Oxford Nanopore Direct RNA and cDNA analysis (Anton, Spyros)
- Single Cell Data analysis (Francesca)
- microRNA Target analysis and interpretation (Francesca, Anton)
- Functional analyses (Francesca)
- **Bringing everything together**

Overview of my Sections

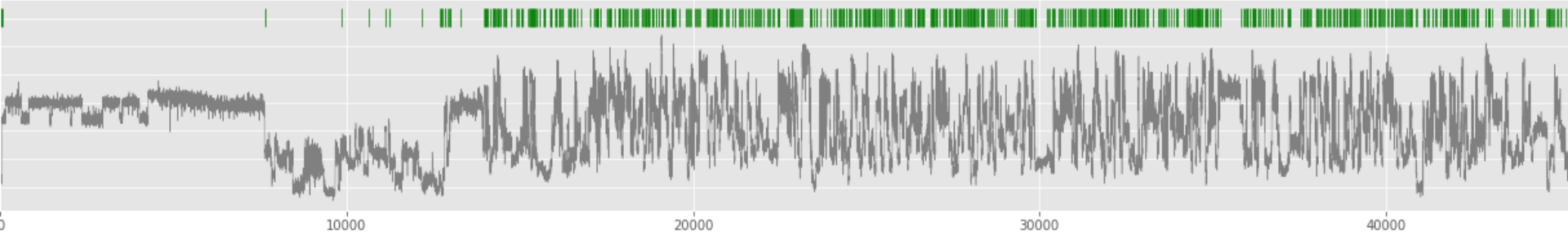


Anton

Steph

Nanopore Analysis

- **Library Prep and Run Setup (MinKnow) for MinION**
- Live Monitoring of Sequencing Run and QC checks during Run
- Basecalling (GPU Accelerated vs CPU) and file formats (FAST5 and FASTQ)
- QC of pores, sequences and samples
- Analysis of detected sequences
    - Mapping of long-read data (minimap2)
    - Quantitation of reads (salmon)
    - Differential expression (DESeq2)
- Comparison between cDNA and RNA
- Visualisation of splicing



- **Library Prep and Run Setup (MinKnow) for MinION**
- Live Monitoring of Sequencing Run and QC checks during Run
- Basecalling (GPU Accelerated vs CPU) and file formats (FAST5 and FASTQ)
- QC of pores, sequences and samples
- Analysis of detected sequences
    - Mapping of long-read data (minimap2)
    - Quantitation of reads (salmon)
    - Differential expression (DESeq2)
- Comparison between cDNA and RNA
- Visualisation of splicing

Small RNA NGS Analysis

- FASTQ file format
- Read cleanup and QC in REAPER (Enrightlab R Module)
- Adapter Removal
    - Contaminant Removal
    - Size Selection 15-28nt
- Read mapping using ChimiRa (Enrightlab)
    - BLAST Based mapping against miRBase precursors
- Differential Expression Analysis (DESeq2)
- microRNA expression and statistics across samples
- Starting point for functional assays

Final Points

- This is a relatively new course
- We are using new technologies
- We are generating a tremendous amount of data in a short time
- We aim to teach the whole experiment from design to analysis
- Things can and will go wrong…….
- **Ask for help!**