

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

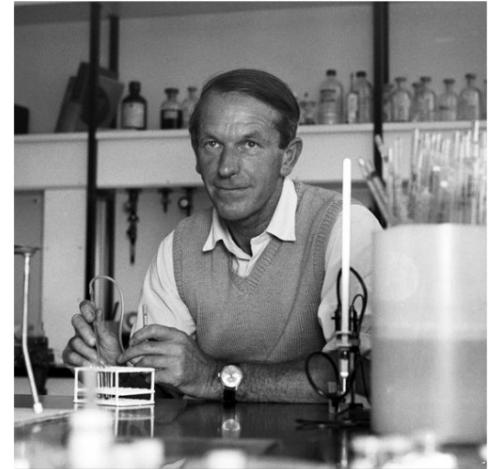
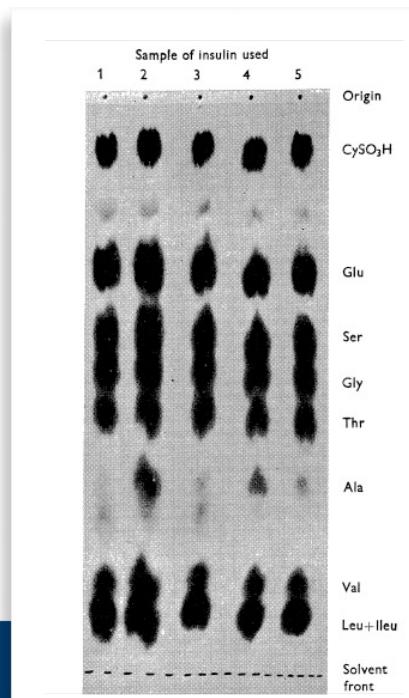
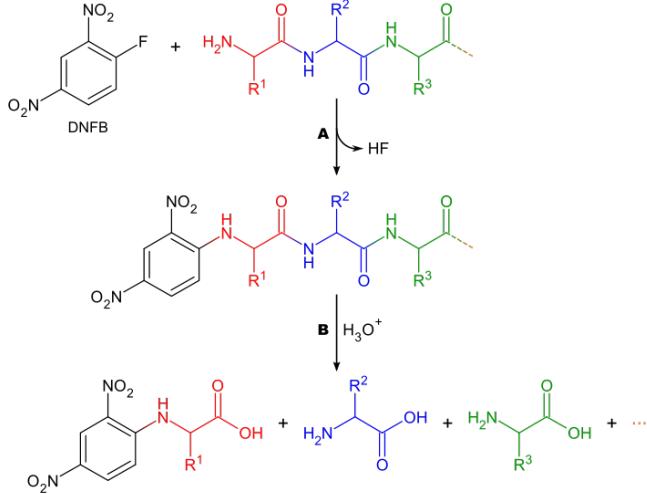
Overview - Computational Analysis

- **Course Modules and Analysis**

- Background - Sequencing, Genomics, Transcriptomics
- Main Biological Focus
- Analysis Framework
- Different Protocols Used
- Goals

Sequencing in molecular biology

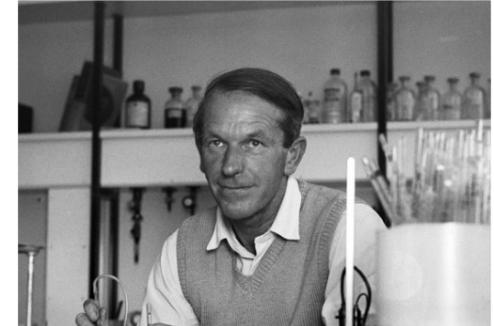
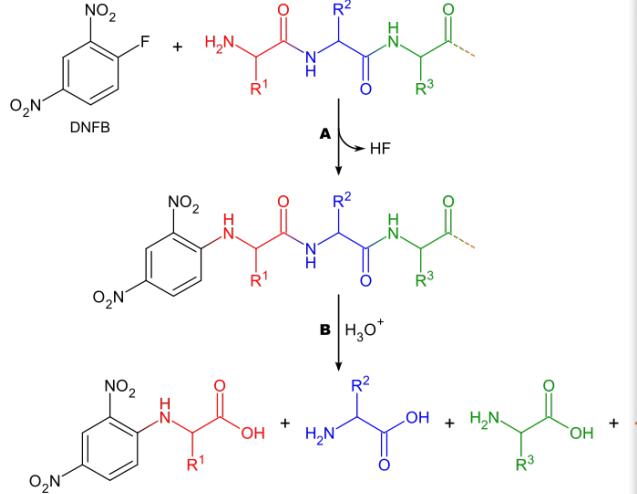
- Birth of sequencing - Fred Sanger 1951
- Proteins thought to be quite amorphous
- Looking for composition of amino acids in Bovine Insulin
- N-terminal Labelling then hydrolysis
- Chromatographic separation



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

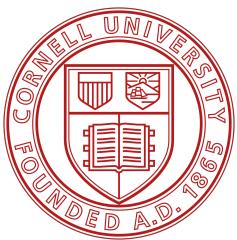


Sequencing Nucleotides - RNA was first

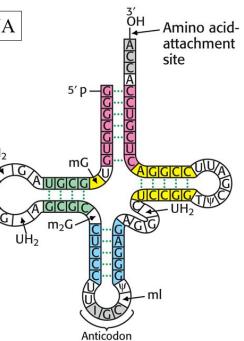
- Double Ribonuclease digestion of RNA molecules - 1965, 1967



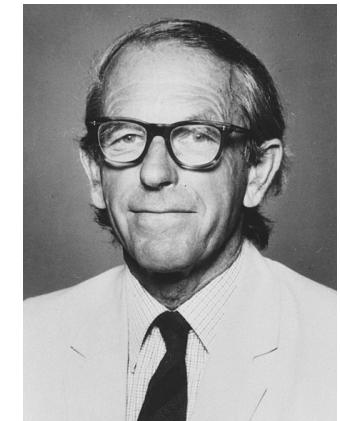
Robert Holley
Cornell



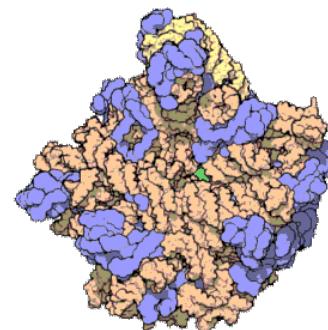
Alanine tRNA



Alanine tRNA

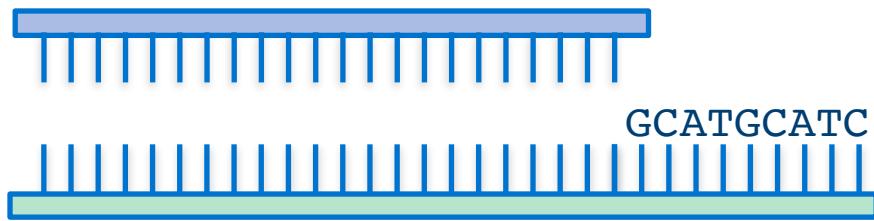


Fred Sanger
LMB

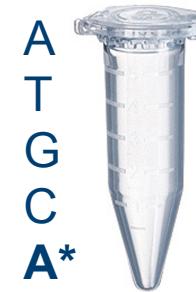


5s Ribosomal RNA

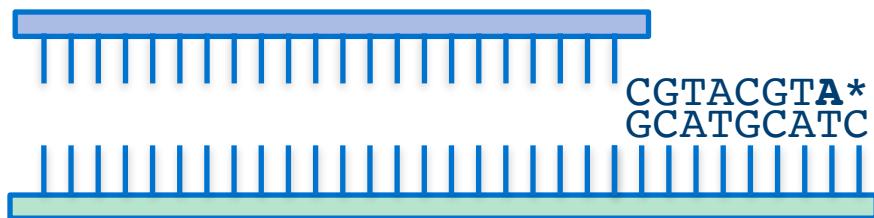
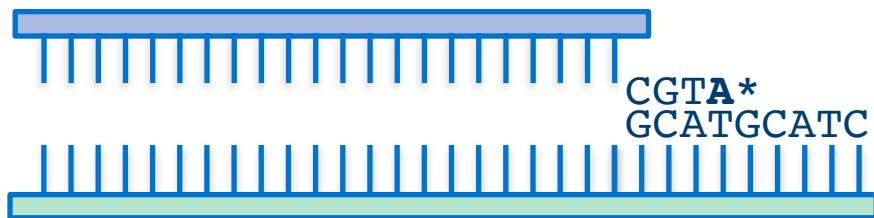
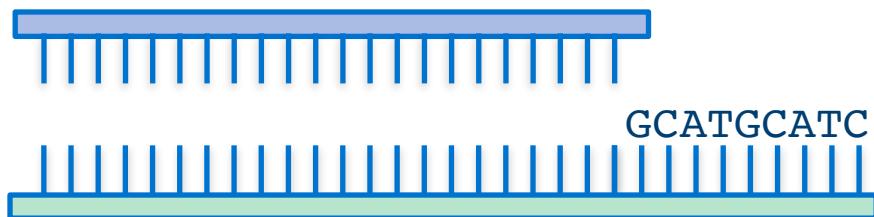
Chain Termination Sequencing - Sanger Sequencing



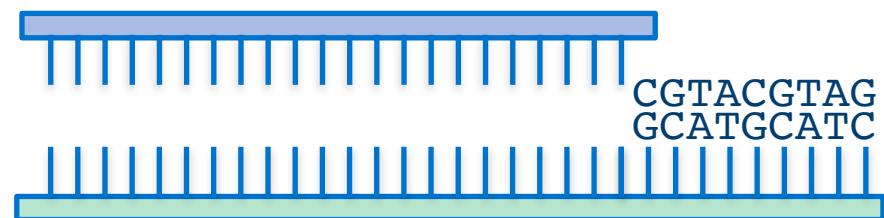
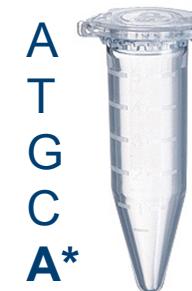
Reaction for ddA



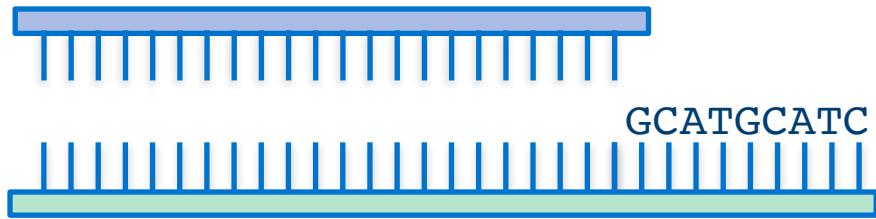
Chain Termination Sequencing - Sanger Sequencing



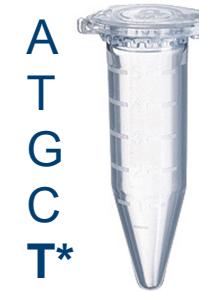
Reaction for ddA



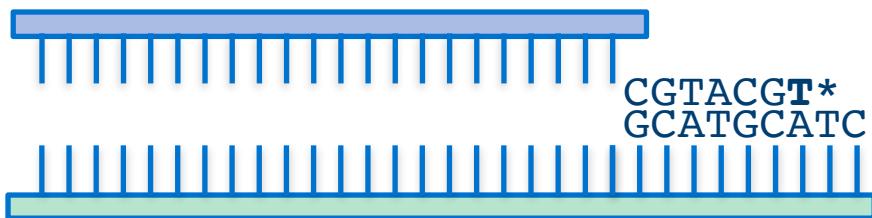
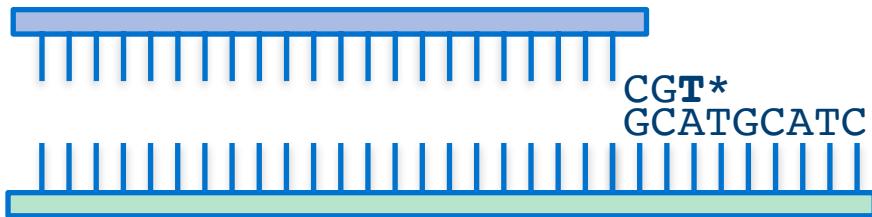
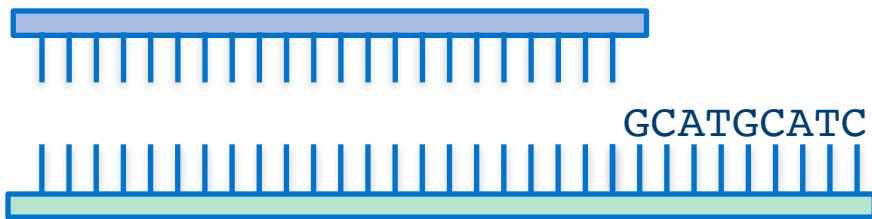
Chain Termination Sequencing - Sanger Sequencing



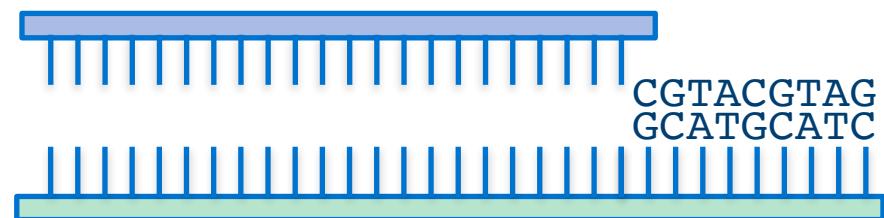
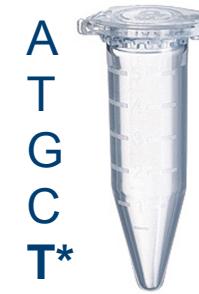
Reaction for ddT



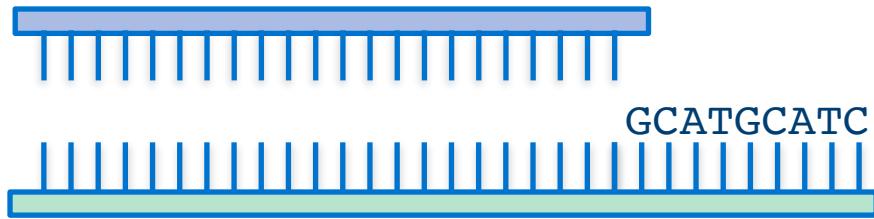
Chain Termination Sequencing - Sanger Sequencing



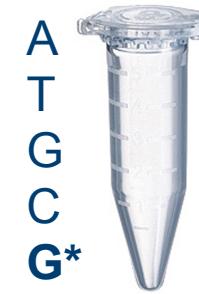
Reaction for ddT



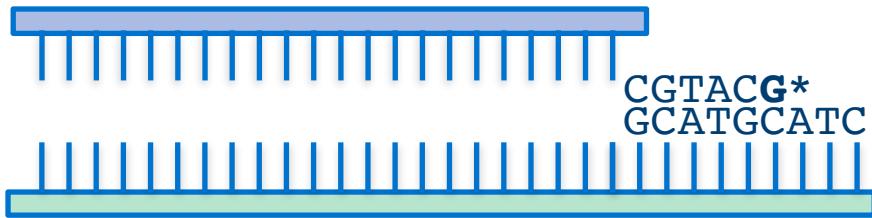
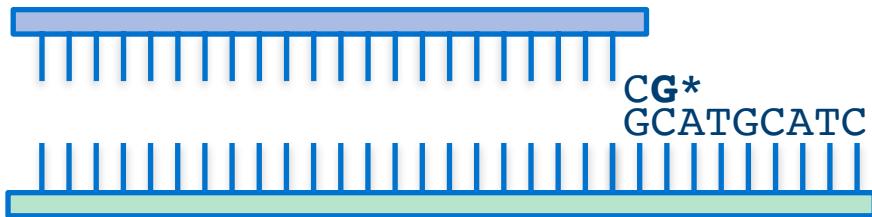
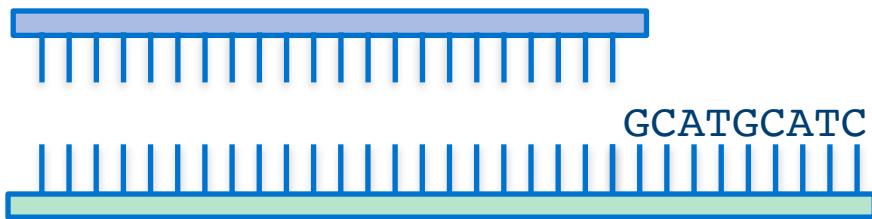
Chain Termination Sequencing - Sanger Sequencing



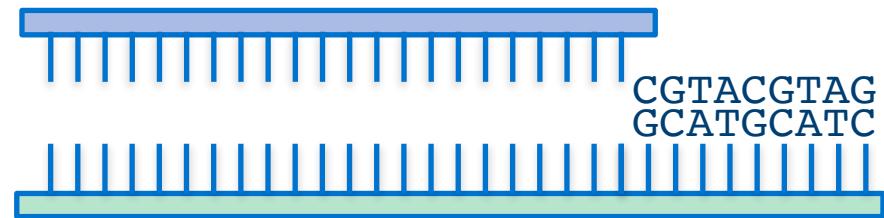
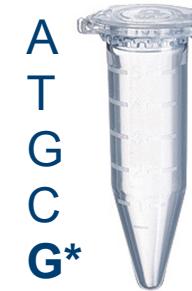
Reaction for ddG



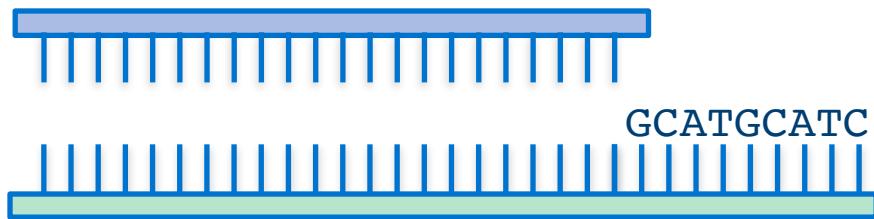
Chain Termination Sequencing - Sanger Sequencing



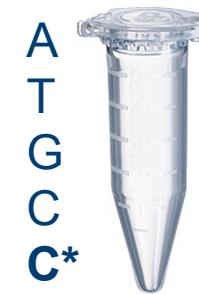
Reaction for ddG



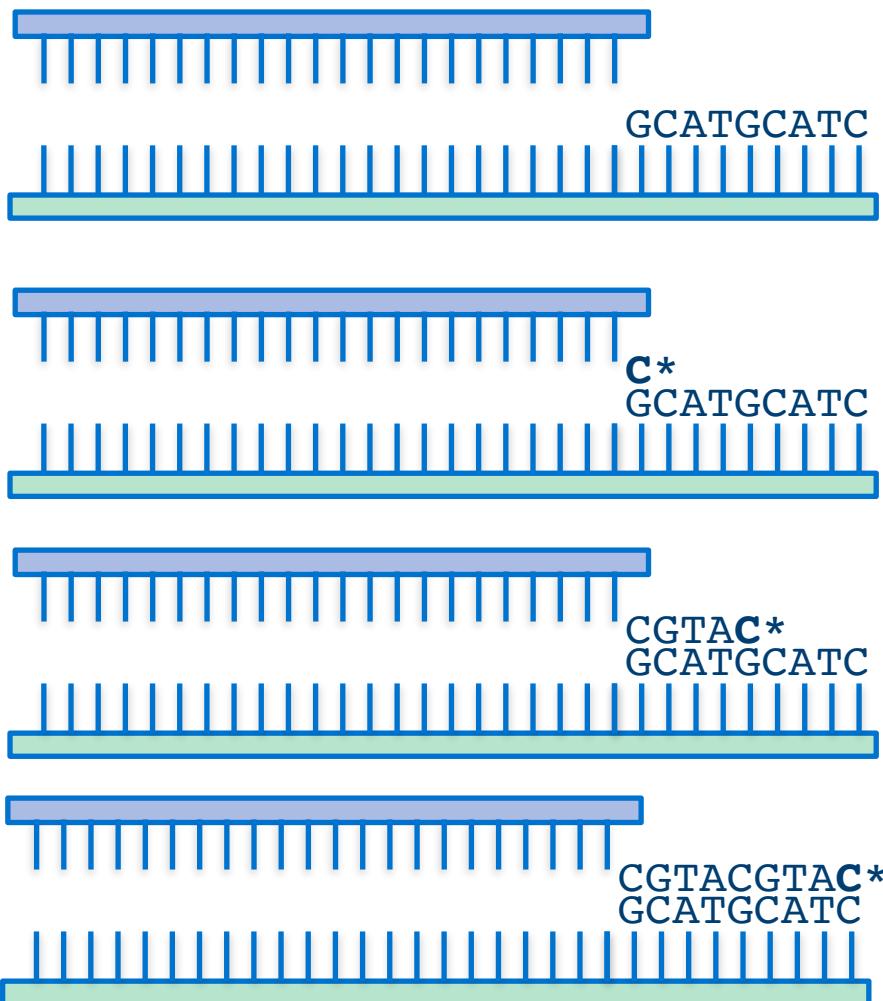
Chain Termination Sequencing - Sanger Sequencing



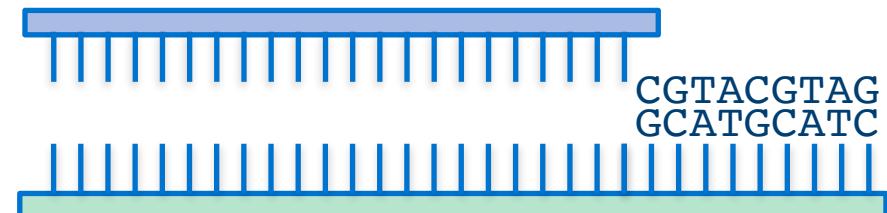
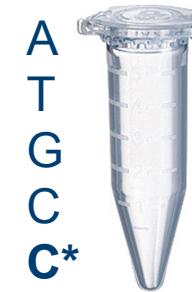
Reaction for ddC



Chain Termination Sequencing - Sanger Sequencing



Reaction for ddC



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*

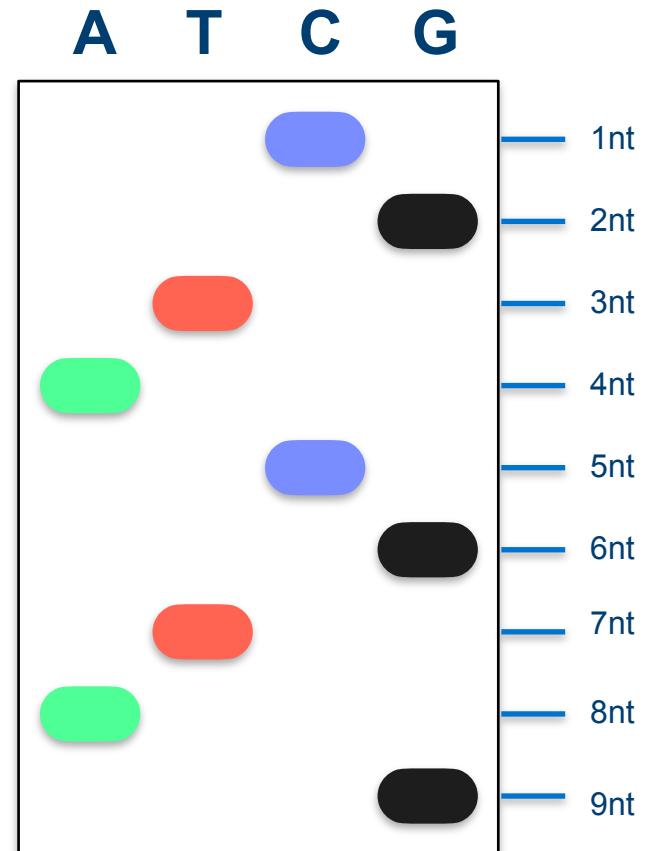
CGTAC*

CGTACG*

CGTACGT*

CGTACGTA*

CGTACGTAG*



Chain Termination Sequencing

CGTA*

CGTACGTA*

CGT*

CGTACGT*

C*

CGTAC*

CG*

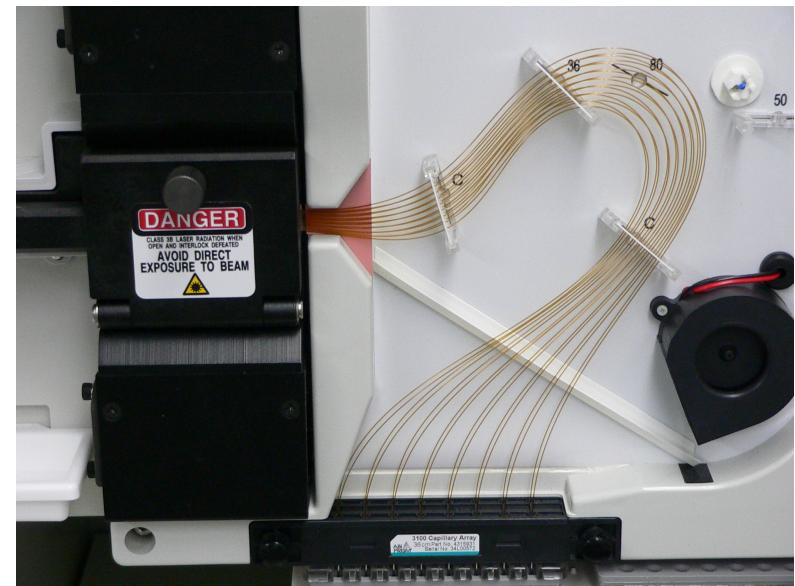
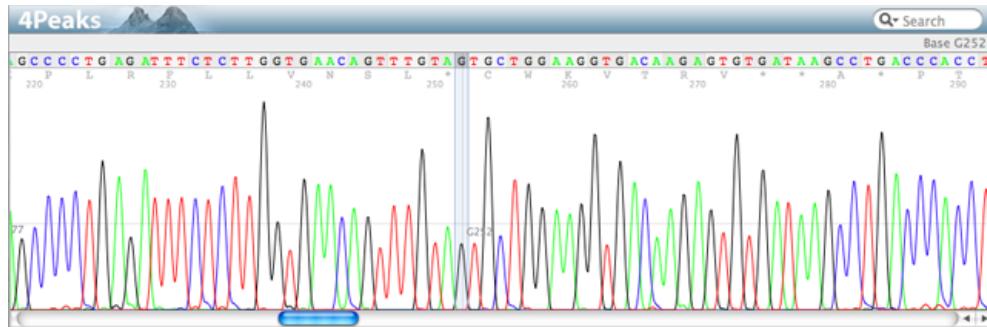
CGTACG*

CGTACGTAG*



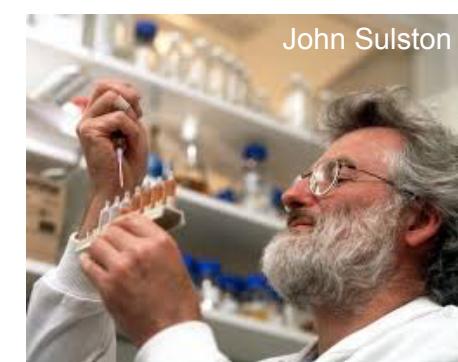
Enabling Genomics: Sanger Sequencing to the next level

- 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



John Sulston



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

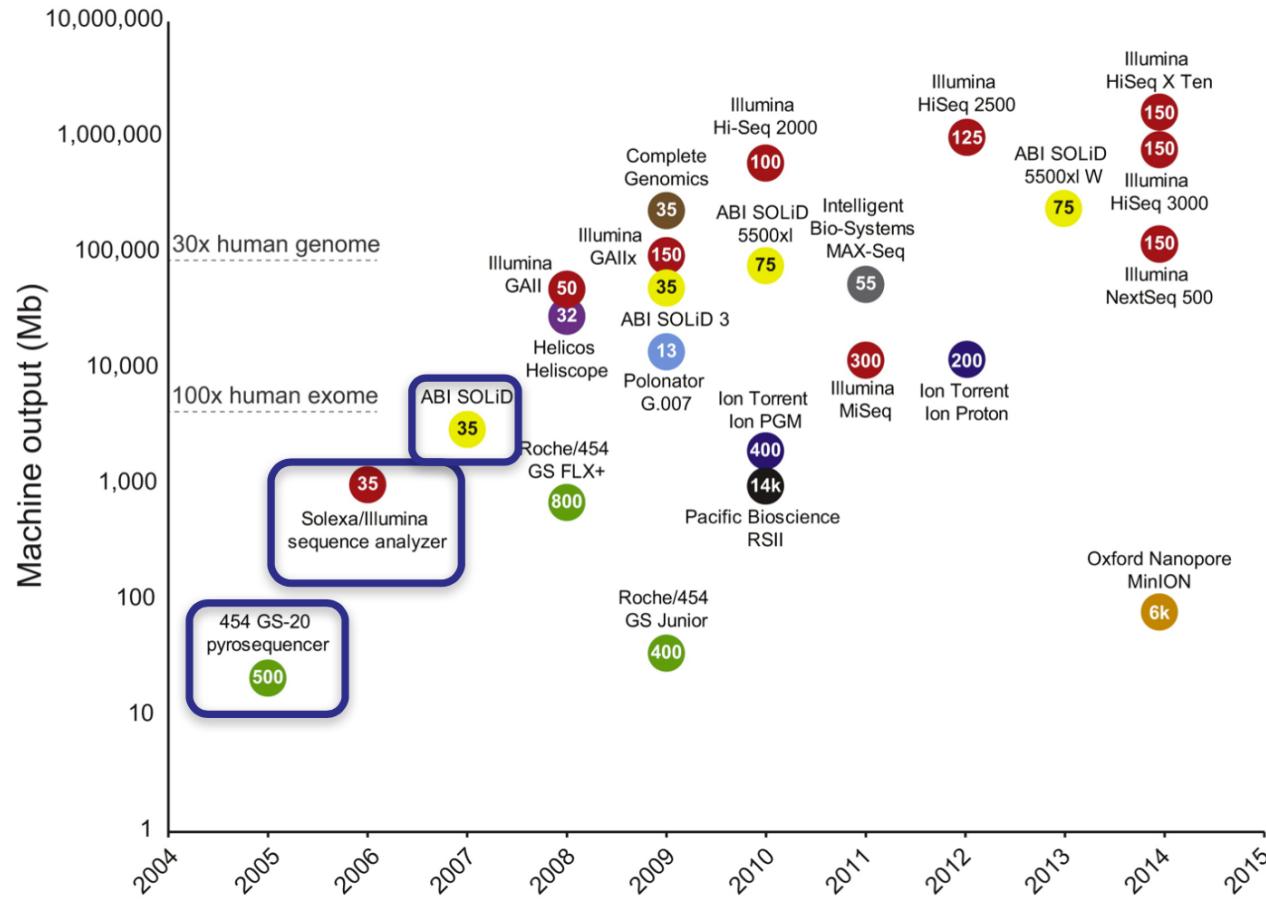


Sanger Centre, opened in 1992
European Bioinformatics Institute shortly after to support the effort to release the data

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



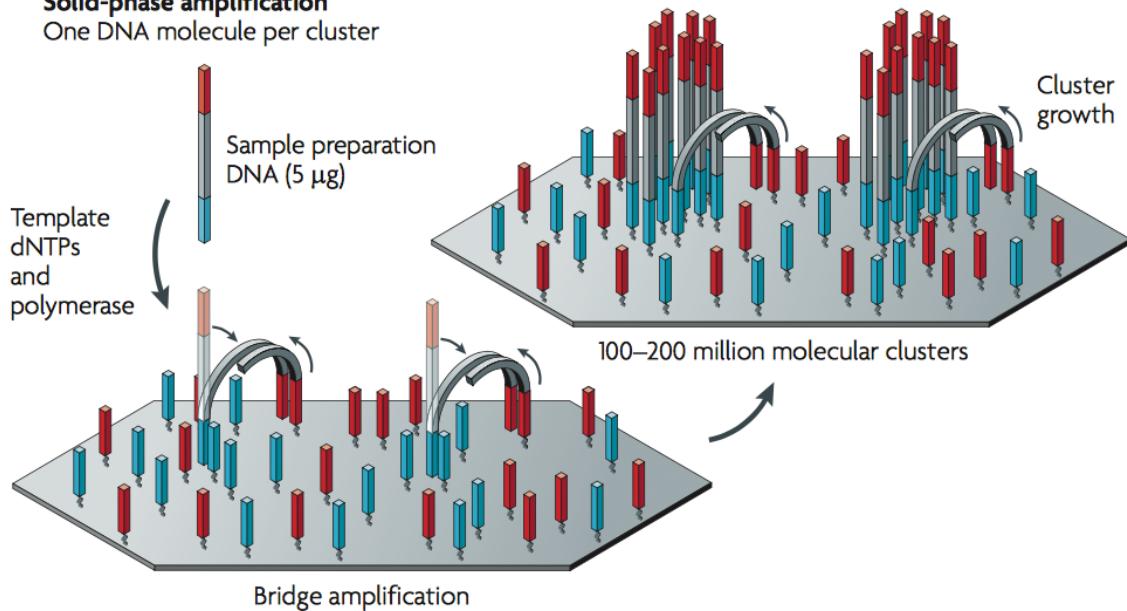
Reuter JA, Spacek DV & Snyder MP, Mol Cell. 2015 May 21;58(4):586-97.

Solid Phase Immobilisation

- Solid Phase methods immobilisation

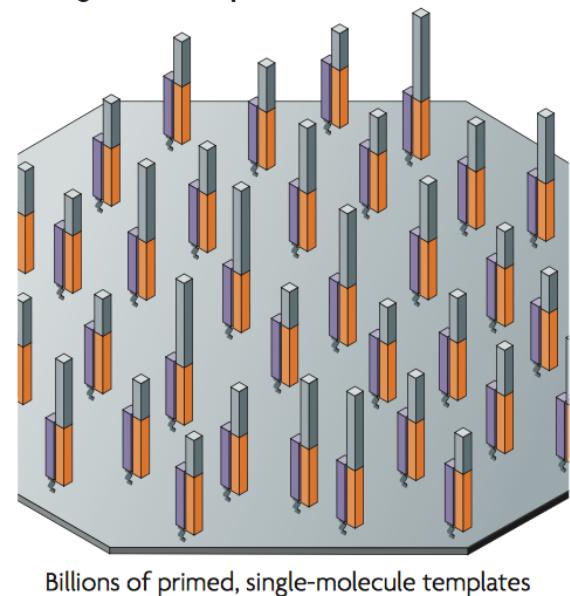
Images from: Metzker ML, Nat Rev Genet. 2010 Jan; 11(1):31-46.

b Illumina/Solexa
Solid-phase amplification
One DNA molecule per cluster



Illumina Sequencing (PCR)

c Helicos BioSciences: one-pass sequencing
Single molecule: primer immobilized

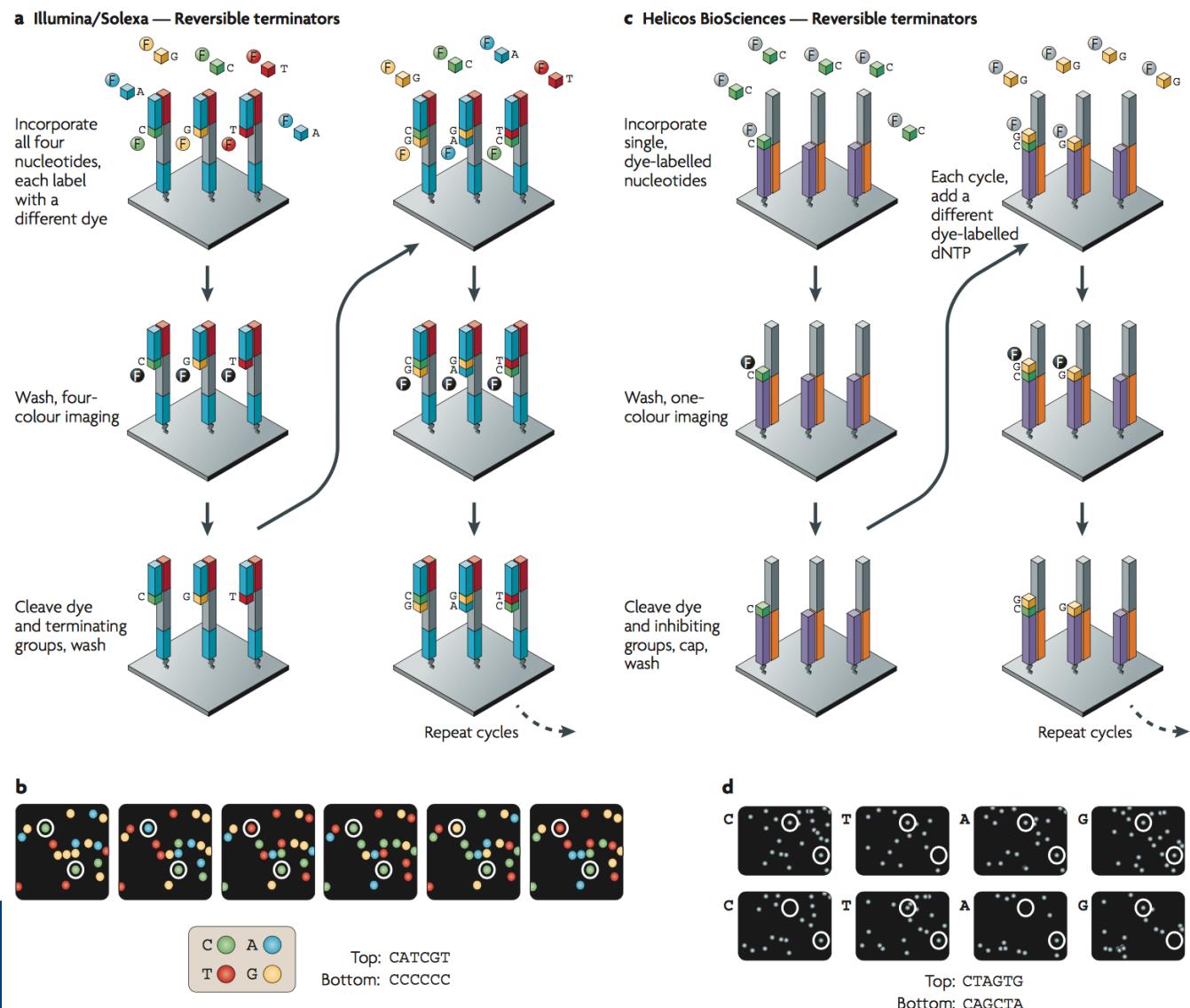


Helicos (Single Molecule)

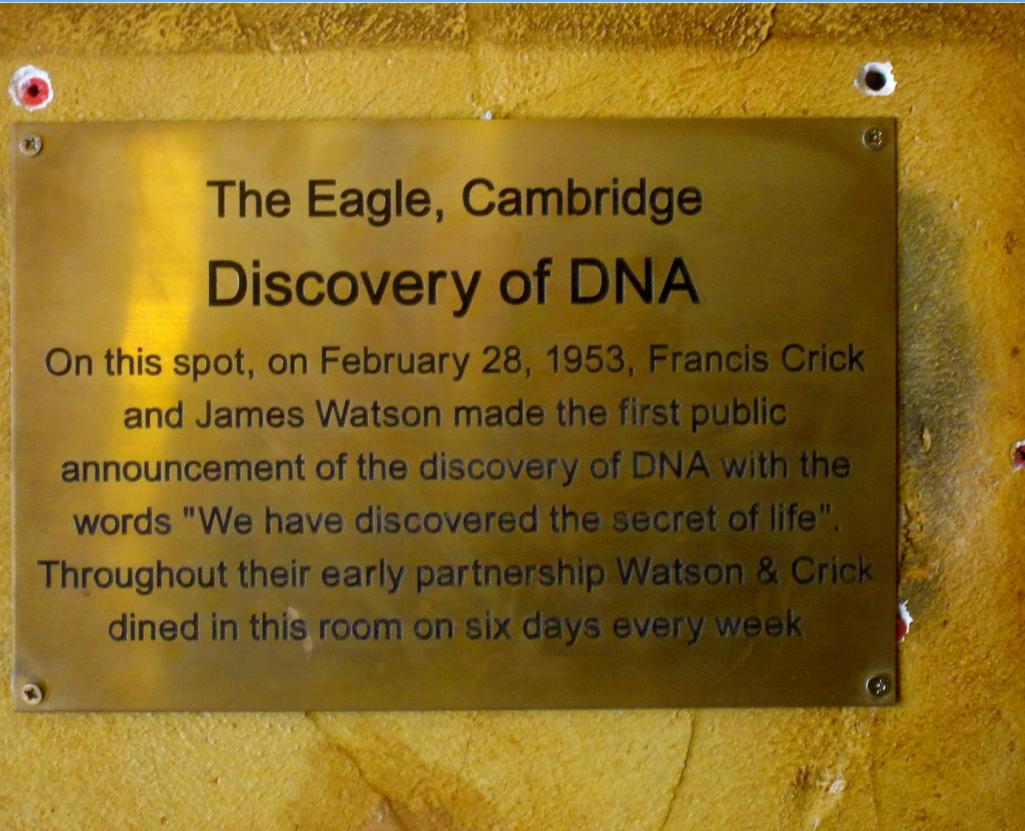
Illumina - Reversible terminator sequencing

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

- Nucleotides are fluorescently labelled
- They contain terminators attached to the fluorophore
- 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



Genome Pubs of Cambridge - Eagle and Panton Arms



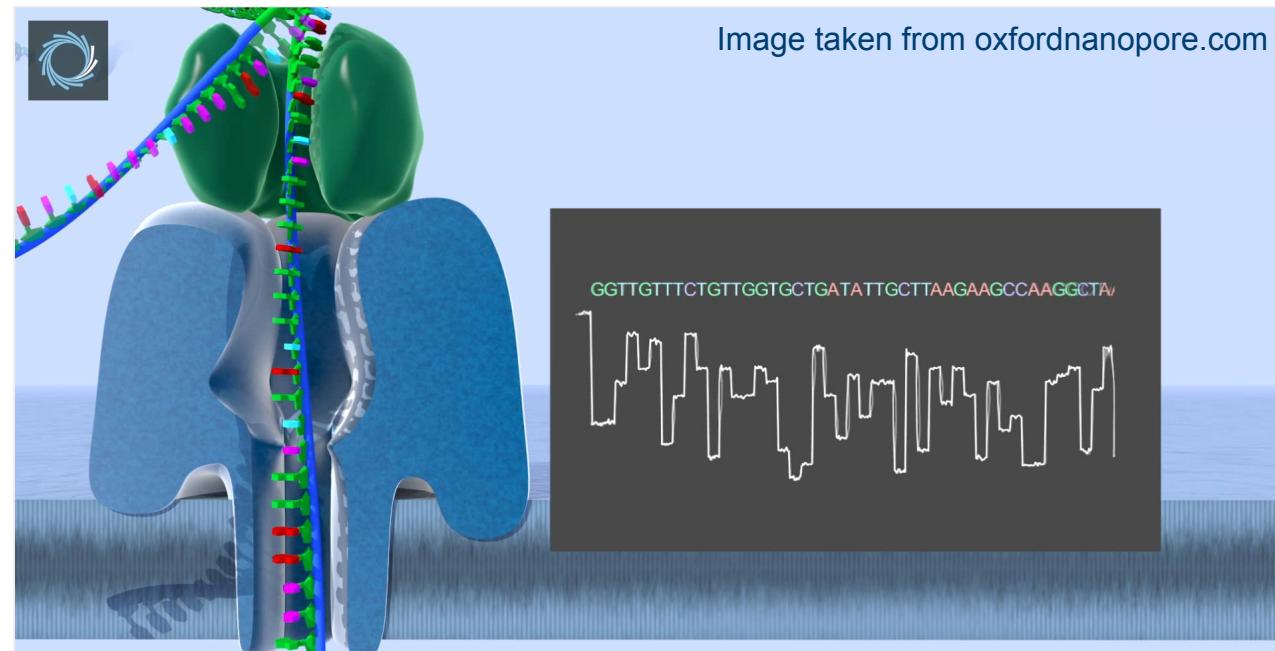
Genome Pubs of Cambridge - Eagle and Panton Arms



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

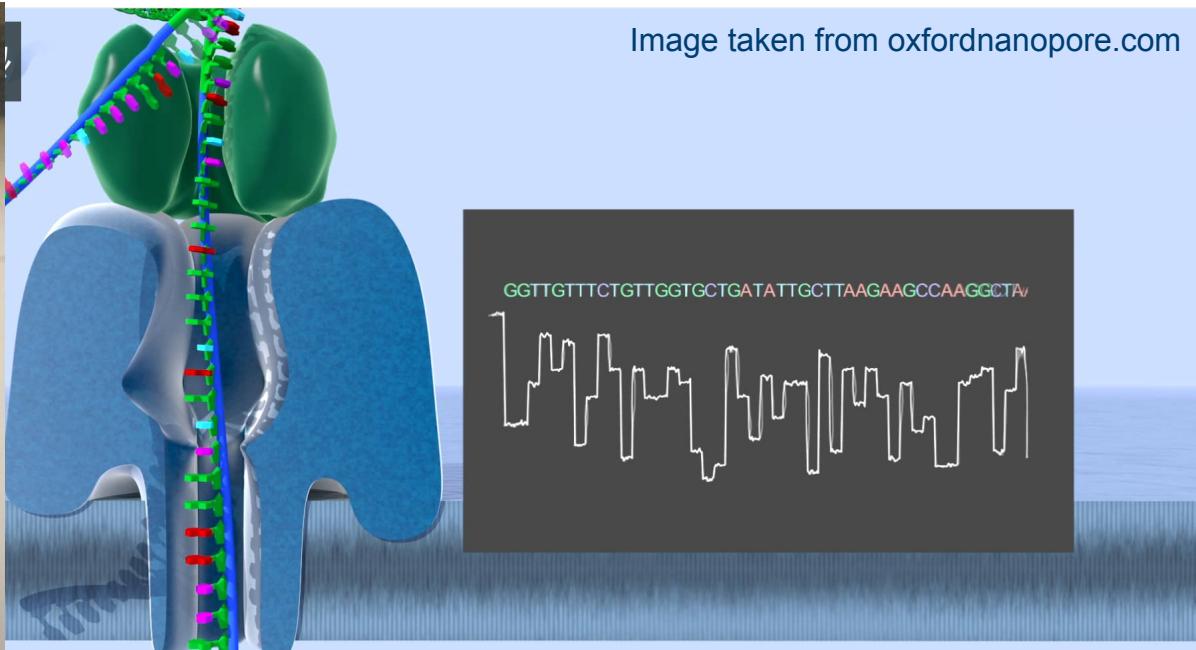


Image taken from oxfordnanopore.com

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

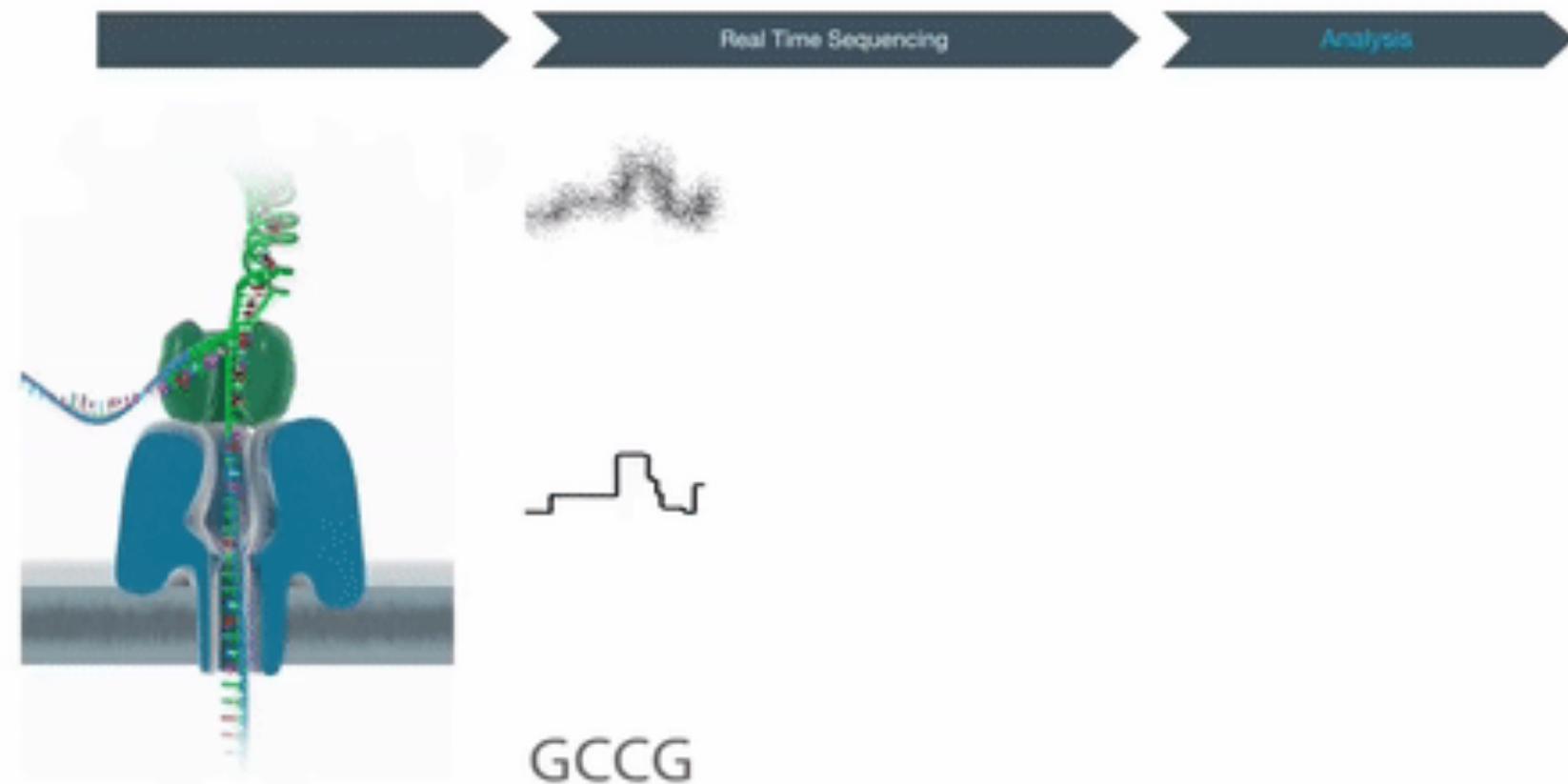
Oxford Nanopore PromethION



School of Biological Sciences PromethION (CGS, Enright Lab)

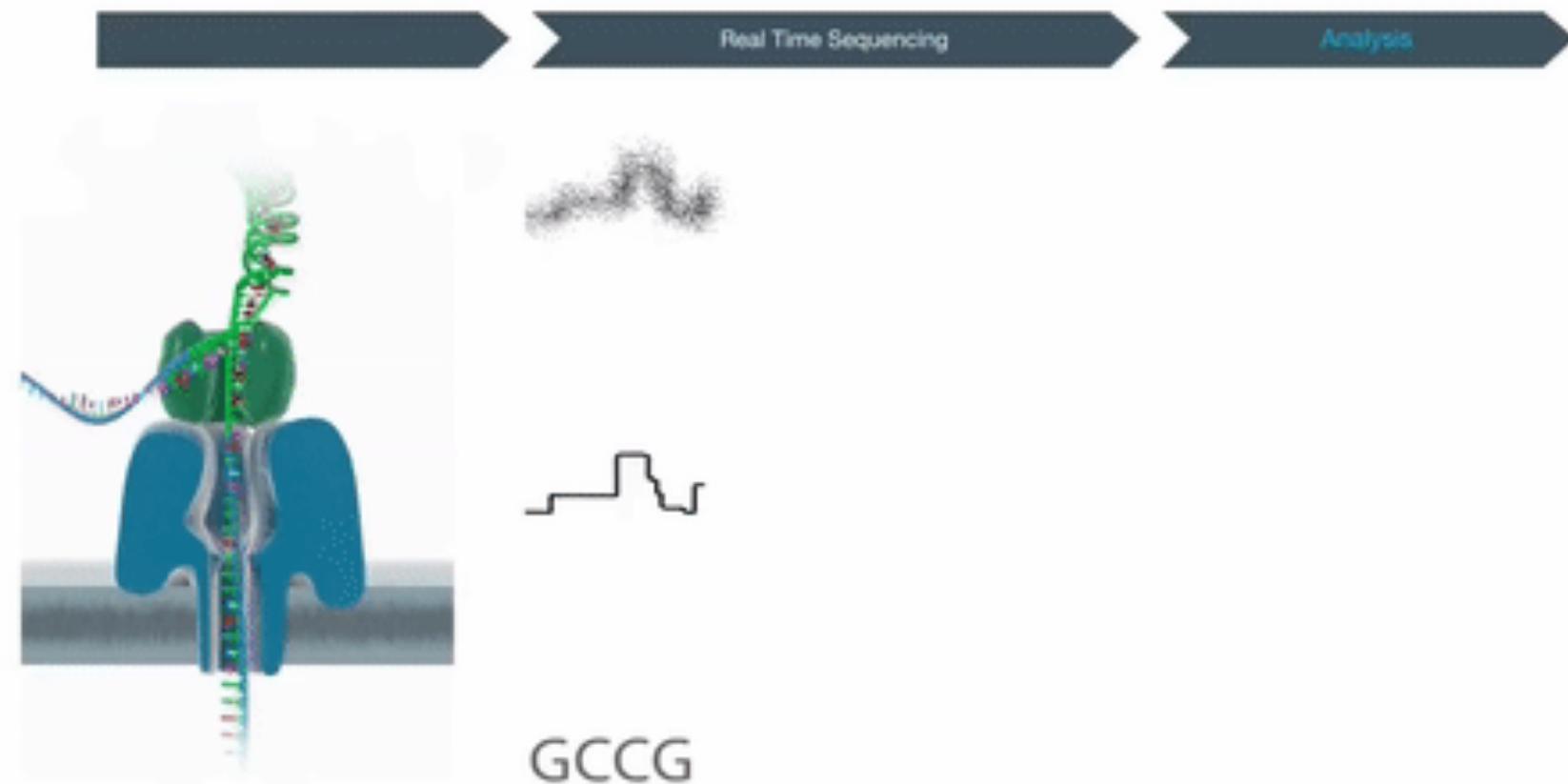
Nanopore in action

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

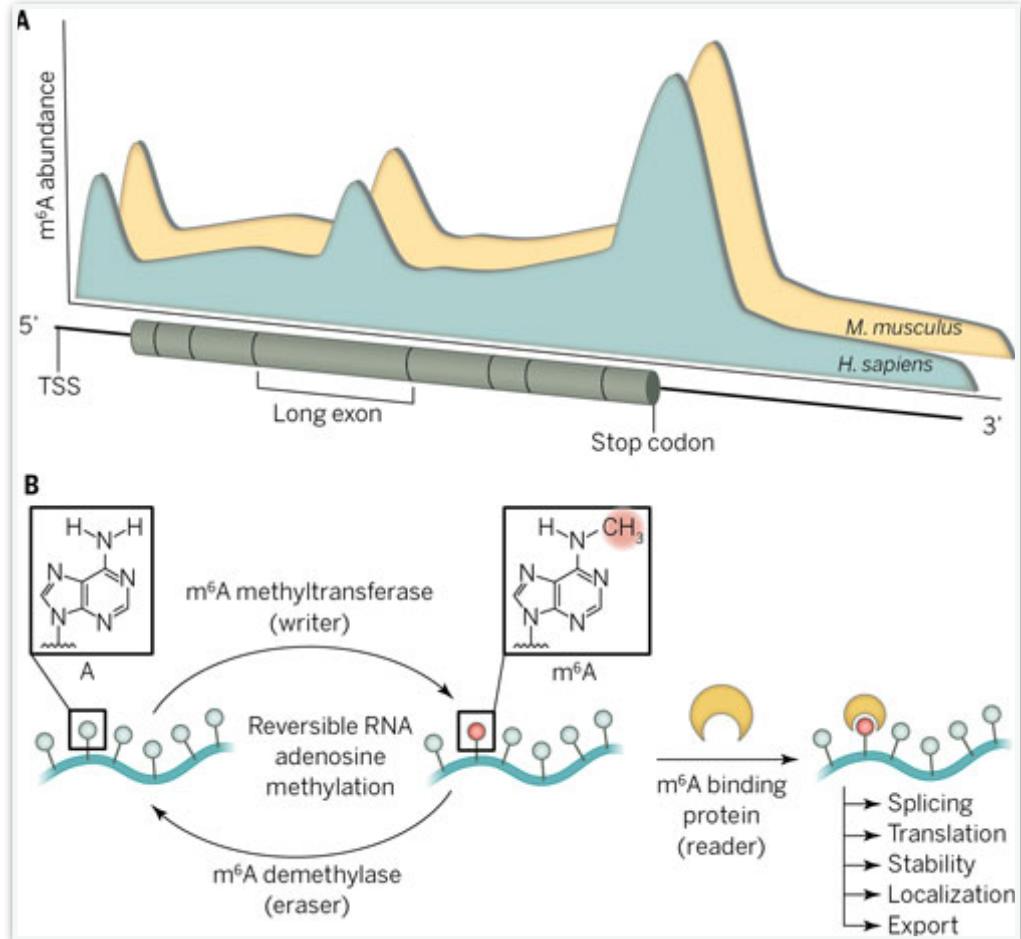
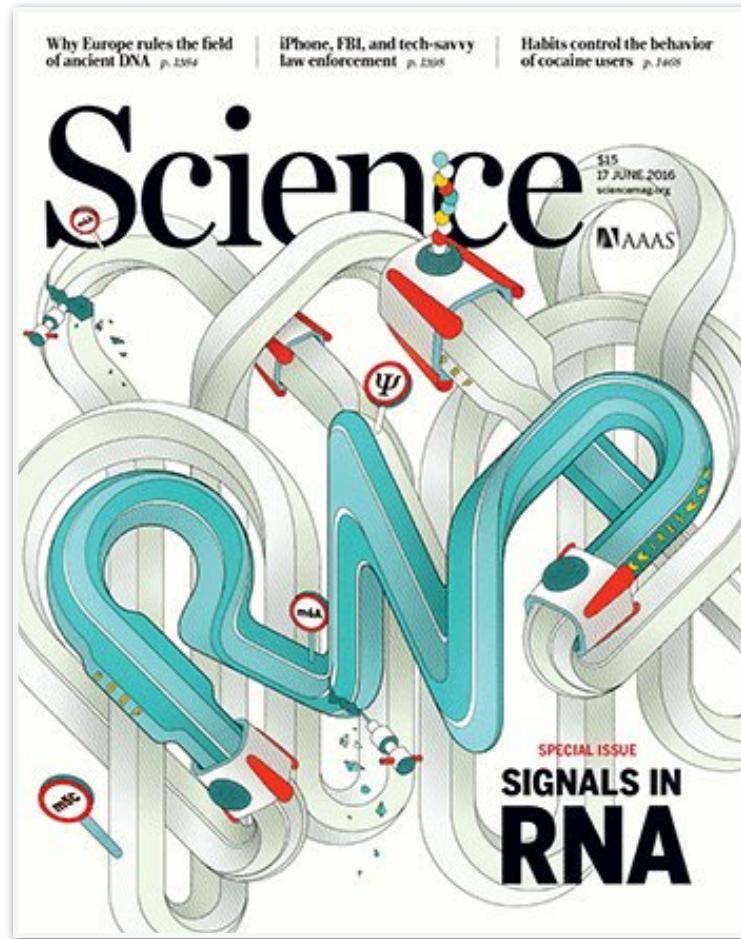


Nanopore in action

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

- **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)

github.com

Courses-and-Practicals / WTAC_Transcriptomics_2019 / small_RNA_seq / Practical_2 / Practical_2.md

Code Blame 324 lines (240 loc) · 12.3 KB

Preview Raw

Code

- master
- + |
- Code
- Blame
- 324 lines (240 loc) · 12.3 KB
- Raw
- +
- ...

dispersion

mean of normalized counts

● gene-est
● fitted
● final

Post Normalisation QC

We can now normalise and plot the counts again. The data look much improved.

```
normcounts <- counts(dds, normalized=TRUE)
rawcounts=counts(dds,normalized=FALSE)
log2counts=log2(normcounts+1)

barplot(apply(normcounts,2,sum), las=2,col=cond_colours,main="Post-Normalised Counts"
legend="topright",levels((cond),cex=0.6,fill=cond_colours[levels(cond)])
```

Post-Normalised Counts

3.5e+07
3.0e+07
2.5e+07
2.0e+07
1.5e+07
1.0e+07
5.0e+06
0.0e+00

II VI II VI VI

■ wt ■ wt ■ wt ■ wt ■ wt ■ wt

github.com

Courses-and-Practicals / WTAC_Transcriptomics_2019 / Nanopore_dRNA_Seq /

Code

- master
- + |
- Code
- Blame
- 324 lines (240 loc) · 12.3 KB
- Raw
- +
- ...

10.8M aligned RNA reads

19.4M aligned 1D cDNA reads

Read Length (bases)

Align Length (bases)

Better exon connectivity

4,855 bp

TLS,912,000 bp TLS,912,000 bp TLS,912,000 bp TLS,912,000 bp

PolyAs can be detected and measured (hopefully)

Current (pA)

Raw data point @ 3 kHz

(Some) RNA modifications can modify the signal

Current (pA)

E. coli wild type (m7G527) E. coli RsmQΔ mutant (G527)

Ref. Position

ONT multiFast5/Fast5 file format

MINKnow generates files containing the raw intensity signal in **HDF5 format**. The latest Fast5 format contains multiple reads per file.

Documentation · Share feedback

group group datasets

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



About
Screenshots
Downloads
Documentation
End-user docs
Technical docs
Contribute
Community

VirtualBox

Welcome to VirtualBox.org!

VirtualBox is a powerful x86 and AMD64/Intel64 [virtualization](#) product for enterprise as well as home use. Not only is VirtualBox an extremely feature rich, high performance product for enterprise customers, it is also the only professional solution that is freely available as Open Source Software under the terms of the GNU General Public License (GPL) version 3. See "[About VirtualBox](#)" for an introduction.

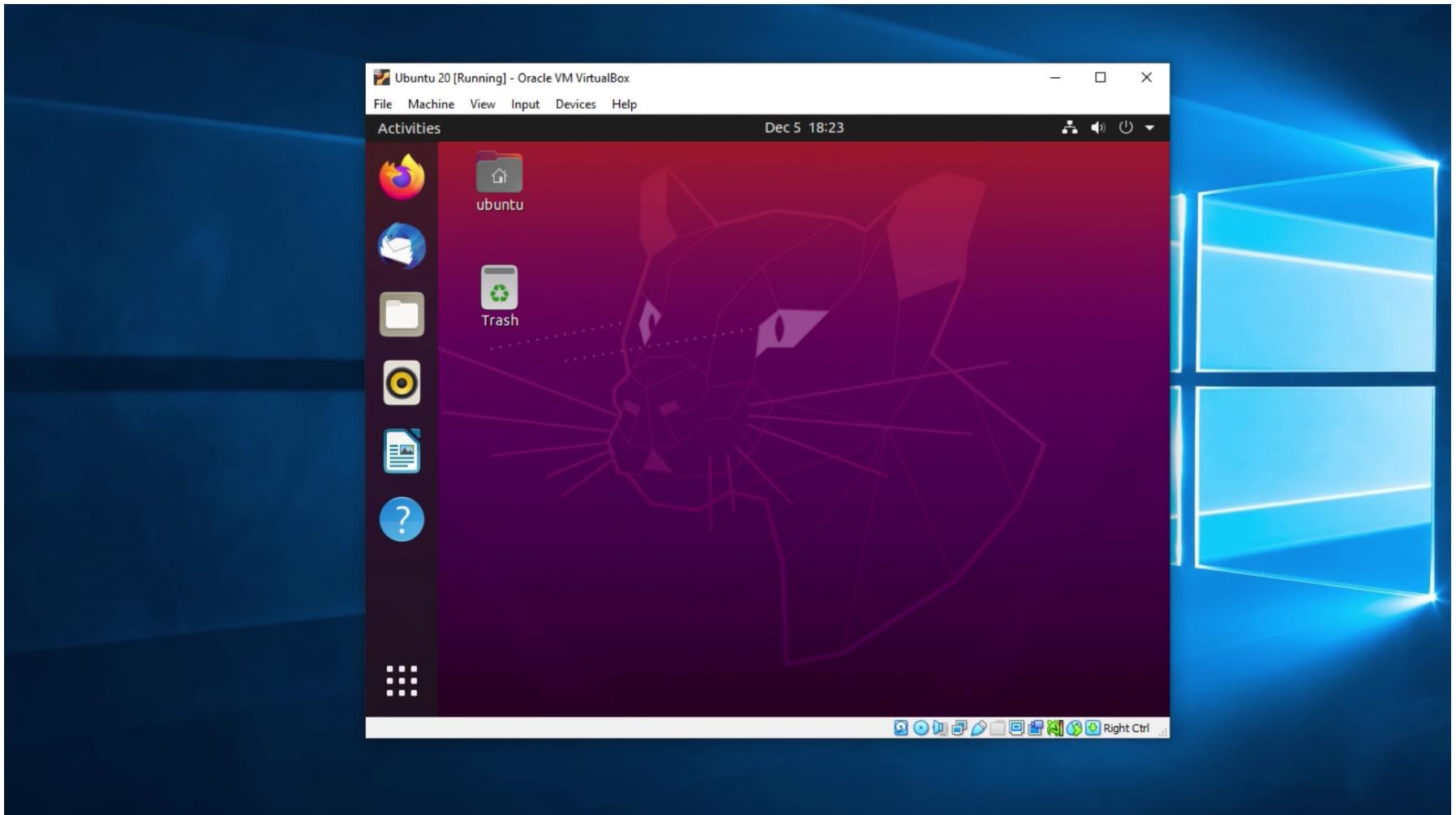
Presently, VirtualBox runs on Windows, Linux, macOS, and Solaris hosts and supports a large number of [guest operating systems](#) including but not limited to Windows (NT 4.0, 2000, XP, Server 2003, Vista, Windows 7, Windows 8, Windows 10), DOS/Windows 3.x, Linux (2.4, 2.6, 3.x and 4.x), Solaris and OpenSolaris, OS/2, and OpenBSD.

VirtualBox is being actively developed with frequent releases and has an ever growing list of features, supported guest operating systems and platforms it runs on. VirtualBox is a community effort backed by a dedicated company: everyone is encouraged to contribute while Oracle ensures the product always meets professional quality criteria.

News Flash

- **New April 18th, 2023**
[VirtualBox 7.0.8 released!](#)
Oracle today released a 7.0 maintenance release which improves stability and fixes regressions. See the [Changelog](#) for details.
- **New April 18th, 2023**
[VirtualBox 6.1.44 released!](#)
Oracle today released a 6.1 maintenance release which improves stability and fixes regressions. See the [Changelog](#) for details.
- **New January 17th, 2023**
[VirtualBox 7.0.6 released!](#)
Oracle today released a 7.0 maintenance release which improves stability and fixes regressions. See the [Changelog](#) for details.
- **New January 17th, 2023**
[VirtualBox 6.1.42 released!](#)
Oracle today released a 6.1 maintenance release which

VM in use



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



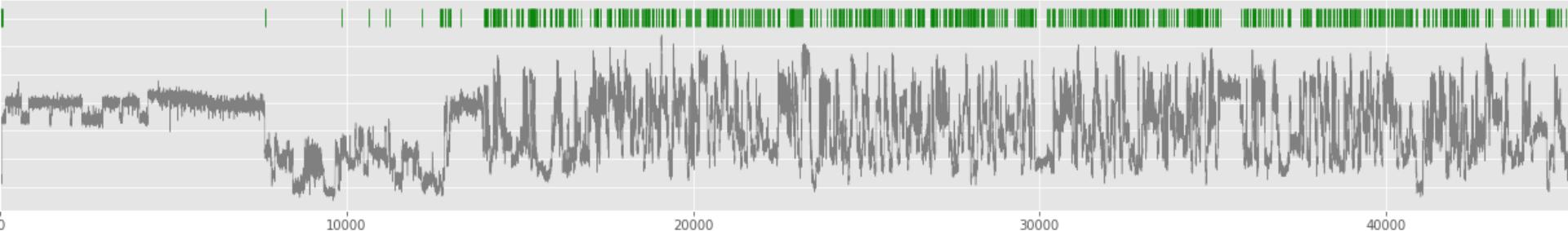
Anton



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!**