“The double helix is indeed a remarkable molecule. Modern man is perhaps 50,000 years old, civilization has existed for scarcely 10,000 years and the United States for only just over 200 years; but DNA and RNA have been around for at least several billion years.

All that time the double helix has been there, and active, and yet we are the first creatures on Earth to become aware of its existence.”

Francis Crick (1916–2004)
Contents

• A short history of DNA

• Review of first generation sequencing techniques

• Short-read second generation sequencing technology
  – Illumina
  – Life Tech Ion Torrent

• Third generation single molecule sequencing
  – PacBio
  – Oxford Nanopore
A short history of DNA
“The double helix is indeed a remarkable molecule. Modern man is perhaps 50,000 years old, civilization has existed for scarcely 10,000 years and the United States for only just over 200 years; but DNA and RNA have been around for at least several billion years.

All that time the double helix has been there, and active, and yet we are the first creatures on Earth to become aware of its existence.”

Francis Crick (1916–2004)
The first person to isolate DNA

• Friedrich Miescher
  – Born with poor hearing
  – Father was a doctor and refused to allow Freidrich to become a priest

• Graduated as a doctor in 1868
  – Persuaded by his uncle not to become a practising doctor and instead pursue natural science
  – But he was reluctant...
Biology PhD angst in the 1800s

“I already had cause to regret that I had so little experience with mathematics and physics... For this reason many facts still remained obscure to me.”

*His uncle counselled:*

“I believe you overestimate the importance of special training...”

Friedrich Miescher
1869 - First isolation of DNA

• Worked in Felix Hoppe-Seyler’s laboratory in Tubingen, Germany
  – The founding father of biochemistry

• The lab was one of the first to crystallise haemoglobin and describe the interaction between haemoglobin and oxygen

• Friedrich was interested in the chemistry of the nucleus

• Friedrich extracted ‘nuclein’ on cold winter nights
  – Initially from human leukocytes extracted from bandage pus from the local hospital
  – Later from salmon sperm
Meisher’s isolation technique

- Cells from surgical bandages or salmon sperm
- Alcohol to remove outer cell membrane
- Pepsin from pig stomachs
- Basic solution to dissolve nuclein in the nucleus
- Acid solution to precipitate the nuclein
- Difficult to do without also precipitating bound protein

Biology PhD angst in the 1800s

His student remembered:

“Friedrich failed to turn up for his own wedding. We went off to look for him. We found him quietly working in his laboratory.”

“I go at 5am to the laboratory and work in an unheated room. No solution can be left standing for more than 5 minutes... Often it goes on until late into the night.”

Friedrich Miescher
1874 - First hints to composition

• By 1874 Meischer had determined that nuclein was
  – A basic acid
  – High molecular weight
  – Nuclein was bound to ‘protamin’
• Came close to guessing its function
  – “If one wants to assume that a single substance is the specific cause of fertilisation, the one should undoubtedly first and foremost consider nuclein”
  – Discarded the idea because he thought it unlikely that nuclein could encode sufficient information
1881 - Discovering the composition of nuclein

- Kossel worked in the same lab as Freidrich Miescher
- Wanted to relate chemical composition to biological function

- Discovered fundamental building blocks of nuclein
  - Adenine, Cytosine, Guanine, Thymine, and Uracil
  - Identified histone proteins and that nuclein was bound to histone in the nucleus
  - Inferred that nuclein was not used for energy storage but was linked to cell growth
1890s - Molecular basis of heredity

- How are characteristics transmitted between generations?

- Lots of theories
  - Stereo-isomers
  - Asymmetric atoms
  - Complex molecules

- Realisation that hereditary information is transmitted by one or more molecules

- 1893 August Weismann – germ plasm theory

- 1894 Eduard Strasburger- “nuclei from nuclei”
# 1900 - What we knew

<table>
<thead>
<tr>
<th>Known</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Distinction between proteins and nucleic acids</td>
<td>• Mendel’s lost laws</td>
</tr>
<tr>
<td>• Somehow nuclein was involved in cell growth</td>
<td>• Base composition of nucleic acids</td>
</tr>
<tr>
<td>• Somehow the nucleus was involved in cell division</td>
<td>• Role of the nucleus</td>
</tr>
<tr>
<td>• Rediscovery of Mendel’s laws</td>
<td>• Distinction between RNA and DNA</td>
</tr>
<tr>
<td></td>
<td>• Significance of chromosomes</td>
</tr>
<tr>
<td></td>
<td>• That enzymes were proteins</td>
</tr>
<tr>
<td></td>
<td>• Most of biochemistry</td>
</tr>
</tbody>
</table>
1902 – Linking chromosome count to somatic and gametic cells

• Walter Sutton using grasshopper gametes
• Theodore Boveri using sea urchins

“I may finally call attention to the probability that the association of paternal and maternal chromosomes in pairs and their subsequent separation during the reducing division as indicated above may constitute the physical basis of the Mendelian law of heredity.”

1910s - More on the composition of DNA

- Determined relative composition of sugars, phosphate and sugars by hydrolysis of nucleic acid
- Enabled the discovery of DNA and RNA bases
- Unfortunately, this method can destroy bases and bias results
- Made it impossible to compare composition between species
- Phoebus Levene proposed the tetranucleotide hypothesis
  - DNA consisted of repeating units of thymine, guanine adenine and cytosine
  - E.g. GACT GACT GACT
  - Convinced many that DNA could not be a carrier of hereditary information
  - Led to the assumption that DNA was just a structural component of cells
1910-30s - Chromosome theory of heredity

- Chromosome as a unit of heritability confirmed by Thomas Morgan and his student Alfred Sturtevant creates the first genetic map in 1913-15 in Drosophila.

- Hypothesized that crossing-over during meiosis could explain variations in progeny phenotype after crossing over.

- Genetic recombination shown to be caused by physical recombination of chromosomes by Barbara McClintock & Harriet Creighton in 1930.
1928 - Inheritance of virulence

• Established that non-virulent pneumococci bacteria could be converted be made virulent by exposure to lysed virulent bacteria

Frederick Griffiths

“Could do more with a kerosene tin and a primus stove than most men could do with a palace”

Hedley Wright

• What was the ‘transforming principle’ which underlay this observation?

http://mic.sgmjournals.org/content/73/1/1.full.pdf
1944 – What is life?

- An ‘aperiodic solid crystal’ could code for an organism
- “A well-ordered association of atoms endowed with sufficient resistivity to keep its order permanently”
- Also placed living systems into a thermodynamic framework
- Served as inspiration for Watson & Crick

Erwin Schrödinger
1944 – Establishing DNA as the transforming principle

- Separated cellular components and repeated Griffiths experiments
- Enabled by new ‘ultra-centrifugation’ technology
- Extended Griffiths work to prove that nucleic acids were the ‘transforming principle’
- Also demonstrated that DNA, not RNA was the genetic material
- Incredibly small amounts – 1 in 600 million were sufficient to induce transformation

[Image of Oswald Avery]

[Link: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2135445/]

Oswald Avery
1945 – 1952 Critique

- Alfred Mirksy was a pioneer of molecular biology
- Isolated chromatin from a wide variety of cells
- He was concerned that Avery’s results could be the result of protein contamination
- Fought a battle against Avery
- Convinced the Nobel panel not to award a prize to Avery
- Later, Mirsky would actually demonstrate the ‘Constancy’ of DNA throughout somatic cells
1950 – Base composition between organisms

- Erwin Chargaff hit back at Mirsky and developed the base complementarity hypothesis with Masson Gulland
- Determined that the molar ratio of A/T and G/C were always very close to 1
- Relative proportions of bases varied between species but was the same within species
- Refuted Levene’s 30 year-old tetranucleotide hypothesis
1952- Confirmation of Avery’s experiment

- Grow bacteriophage using radioactive substrates
  - Protein with radioactive sulphur
  - DNA with radioactive phosphorous

- Bacteriophages infected bacteria by injecting DNA, not protein

- Was taken as confirmation of the role of DNA as genetic material

- Yet there was still the possibility of protein contamination here

Hershey Chase experiment

http://en.wikipedia.org/wiki/Hershey%E2%80%93Chase_experiment
1952 – X-ray diffraction patterns of DNA

• Wilkins, Franklin and Gosling
• Much improved X-ray diffraction patterns of the B-form of DNA
• Wilkins developed a method to obtain improved diffraction patterns using sodium thymonucleate to draw out long thin strands of DNA

1953 – Watson & Crick obtain a structure for DNA

- B-model of DNA
- Relied upon data from Maurice Wilkins and Rosalind Franklin via Maz Perutz
- "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material."
- Broad acceptance of the structure did not occur until around 1960
1958 – Evidence for the mechanism of DNA replication

- Meselson & Stahl
- Supported Watson & Crick’s hypothesis of semi-conservative DNA replication

Three postulated methods of DNA Replication

- Semi-Conservative
- Conservative*
- Dispersive*

* not found to be biologically significant
1958 – Evidence for the semi-conservative mechanism of DNA replication
Other developments in molecular biology

- 1954 - George Gamow proposed a 3-letter code
- 1955 – Polynucleotide phosphorylase discovered
  - Enabled synthesis of homogeneous nucleotide polymers
- 1957 – Crick lays out ‘central dogma’
- 1957-1963
  - RNA structure
  - Work on DNA-RNA hybridization
- 1960s
  - Crystal structures of tRNAs
  - Role in protein synthesis
  - Role of ribosomes
- Set the stage for...
1961 - Deciphering the genetic code

• How did DNA code for proteins?

• Nirenberg and Matthaei

• Used polynucleotide phophorylase to construct a poly-uracil polymer

• Added to a cell-free system containing ribosomes, nucleotides, amino acids, energy

• This produced an amino acid chain of phenylalanine

• Completed in mid 1960s by Har Gobind Khohrana
Other key figures

• Max Delbruck
  – Physicist who helped found molecular biology

• Salvador Luria
  – James Watson’s PhD supervisor
  – Demonstrated with Delbruck that inheritance in bacteria was Darwinian and not Lamarkian

• Linus Pauling
  – Proposed triple helix model for DNA

• Lawrence Bragg
  – Hosted Watson & Crick
  – Rival of Pauling’s

• Jerry Donohue, William Astbury, Raymond Gosling, John Randall, Fred Neufield, Herbert Wilson...
1962

• Nobel Prize awarded for Physiology or Medicine to Watson, Crick and Wilkins
• Rosalind Franklin died in 1958 of suspected radiation induced cancer
Oswald Avery

- Avery died in 1955
- It is unknown whether he learned of Watson & Crick’s structure of DNA
- He never received a Nobel for his work
- However his 1944 paper is cited around 40 times a year and has cited over 2000 times in the past 20 years
Further reading

• Eighth day of creation – Horace Freeland Judson

• Life’s Greatest Secret – Matthew Cobb

First generation sequencing
The development of sequencing methodologies

• What do we mean by ‘sequencing’?

• Determining the order and identity of chemical units in a polymer chain
  – Amino acids in the case of proteins
  – Nucleotides in the case of RNA and DNA

• Why do we do it?
  – 3D structure and function is dependent on sequence
1949 – Amino acids

- Sequenced bovine insulin
- Developed a method to label N-terminal amino acids
  - Enabled him to count four polypeptide chains
- Used hydrolysis and chromatography to identify fragments

Fred Sanger
1965 - RNA sequencing and structure

- Sequenced transfer RNA of alanine
- Used 2 ribonuclease enzymes to cleave the enzyme at specific motifs
- Chromatography
- 1968 Nobel prize

http://www.sciencemag.org/content/147/3664/1462
1975 - The dawn DNA sequencing

- Between 1975-1977 three methods of DNA sequencing were published
- Fred Sanger’s Plus/Minus method
- Maxam-Gilbert
- Fred Sanger’s chain termination method
Maxam-Gilbert sequencing is performed by chain breakage at specific nucleotides.
Maxam-Gilbert Sequencing

Proc. Natl. Acad. Sci. USA
Vol. 74, No. 2, pp. 560–564, February 1977
Biochemistry

A new method for sequencing DNA
(DNA chemistry/dimethyl sulfate cleavage/hydrazine/piperidine)

ALLAN M. MAXAM AND WALTER GILBERT
Department of Biochemistry and Molecular Biology, Harvard University, Cambridge, Massachusetts 02138
Contributed by Walter Gilbert, December 9, 1976

ABSTRACT DNA can be sequenced by a chemical procedure that breaks a terminally labeled DNA molecule partially at each repetition of a base. The lengths of the labeled fragments then identify the positions of that base. We describe reactions that cleave DNA preferentially at guanines, at adenines, at cytosines and thymines equally, and at cytosines alone. When the products of these four reactions are resolved by size, by electrophoresis on a polyacrylamide gel, the DNA sequence can be read from the pattern of radioactive bands. The technique will permit sequencing of at least 100 bases from the point of labeling.

We have developed a new technique for sequencing DNA molecules. The procedure determines the nucleotide sequence of a terminally labeled DNA molecule by breaking it at adenine, guanine, cytosine, or thymine with chemical agents. Partial cleavage at each base produces a nested set of radioactive

THE SPECIFIC CHEMISTRY

A Guanine/Adenine Cleavage (2). Dimethyl sulfate methylates the guanines in DNA at the N7 position and the adenines at the N3 (3). The glycosidic bond of a methylated purine is unstable (3, 4) and breaks easily on heating at neutral pH, leaving the sugar free. Treatment with 0.1 M alkali at 90° then will cleave the sugar from the neighboring phosphate groups. When the resulting end-labeled fragments are resolved on a polyacrylamide gel, the autoradiograph contains a pattern of dark and light bands. The dark bands arise from breakage at guanines, which methylate 5-fold faster than adenines (3).

This strong guanine/weak adenine pattern contains almost half the information necessary for sequencing; however, ambiguities can arise in the interpretation of this pattern because the intensity of isolated bands is not easy to assess. To determine
Maxam-Gilbert Sequencing

Sequencing gels are read from **bottom to top** (5’ to 3’).
DNA sequencing with chain-terminating inhibitors
(DNA polymerase/nucleotide sequences/bacteriophage φX174)

F. SANGER, S. NICKLEN, AND A. R. COULSON

Medical Research Council Laboratory of Molecular Biology, Cambridge CB2 2QH, England

Contributed by F. Sanger, October 3, 1977

ABSTRACT A new method for determining nucleotide sequences in DNA is described. It is similar to the “plus and minus” method [Sanger, F. & Coulson, A. R. (1975) J. Mol. Biol. 94, 441-448] but makes use of the 2',3'-dideoxy and arabinonucleoside analogues of the normal deoxynucleoside triphosphates, which act as specific chain-terminating inhibitors of DNA polymerase. The technique has been applied to the DNA of bacteriophage φX174 and is more rapid and more accurate than either the plus or the minus method.

The “plus and minus” method (1) is a relatively rapid and simple technique that has made possible the determination of the sequence of the genome of bacteriophage φX174 (2). It depends on the use of DNA polymerase to transcribe specific regions of the DNA under controlled conditions. Although the method is considerably more rapid and simple than other

a stereoisomer of ribose in which the 3'-hydroxyl group is oriented in trans position with respect to the 2'-hydroxyl group. The arabinosyl (ara) nucleotides act as chain terminating inhibitors of Escherichia coli DNA polymerase I in a manner comparable to ddT (4), although synthesized chains ending in 3' araC can be further extended by some mammalian DNA polymerases (5). In order to obtain a suitable pattern of bands from which an extensive sequence can be read it is necessary to have a ratio of terminating triphosphate to normal triphosphate such that only partial incorporation of the terminator occurs. For the dideoxy derivatives this ratio is about 100, and for the arabinosyl derivatives about 5000.

METHODS
Sanger Sequencing

- Uses a mixture of radio-labelled di-deoxy (ddNTP) and deoxy (dNTP) nucleotides to terminate base incorporation as soon as a ddNTP is encountered.
- With addition of enzyme (DNA polymerase), the primer is extended until a ddNTP is encountered.
- The chain will end with the incorporation of the ddNTP.
- With the proper dNTP:ddNTP ratio (about 100:1), the chain will terminate throughout the length of the template.
- All terminated chains will end in the ddNTP added to that reaction.
How is sequencing terminated at each of the 4 bases?

The 3’-OH group necessary for formation of the phosphodiester bond is missing in ddNTPs.
Sanger sequencing

AGCTGCCCCG

**Possible fragment lengths**

2

4

3

2

A

- ddATP + ddA
- four dNTPs
- dAdGdCdTdGdCdCdCdG

C

- ddCTP + dAdGddC
- four dNTPs
- dAdGdCdTdGdCdCdC
- dAdGdCdTdGdCdCdCdC

G

- ddGTP + dAddG
- four dNTPs
- dAdGdCdTddG
- dAdGdCdTdGdCdCdCdddG

T

- ddTTP + dAdGdCddT
- four dNTPs
- dAdGdCdTdGdCdCdCdCdG
Sanger di-deoxy method

AGCTGCCC

Longer fragments

Shortest fragments

G
A

3'

G
C
C
C

ddG ddA ddT ddC

C
G
T
C
G
A

5'
1985: Automating Sanger Sequencing

• Disadvantages of Sanger sequencing
  – Labour intensive
  – Used radioactive labels
  – Interpretation/analysis was subjective
• Difficult to scale up
• Leroy Hood, Michael Hunkapiller developed an automated method utilising:
  – Fluorescent labels instead of radioactivity
  – Utilise computerised algorithms to analyse data
  – Robotics
• Development of PCR by Kary Mullis (NGS would be impossible without it)
Dye Terminator Sequencing

• A distinct dye or “color” is used for each of the four ddNTP.

• Since the terminating nucleotides can be distinguished by color, all four reactions can be performed in a single tube.
Dye Terminator Sequencing

The DNA ladder is resolved in one gel lane or in a capillary
Dye Terminator Sequencing

- The DNA ladder is read on an electropherogram.
Automated Sequencing

• Dye primer or dye terminator sequencing on capillary instruments.
• Sequence analysis software provides analyzed sequence in text and electropherogram form.
• Peak patterns reflect mutations or sequence changes.
Sanger Sequencing

Useful videos

• http://www.youtube.com/watch?v=91294ZAG2hg&feature=related

• http://www.youtube.com/watch?v=bEFLBf5WEtc&feature=fvwrel
Features of Sanger Sequencing

- 96-384 sequences per run
- 500bp-1kb read lengths
- $100 per megabase
- Accuracy decreases with length (99.999% at 500bp down to 99% at 900bp)
- Still the most accurate technique for sequencing
Limitations of Sanger Sequencing

• Cloning/Sub-cloning
  – DNA must be compatible with biological machinery of host cells and can introduce bias
  – Labour and/or machines to prepare clones requires significant capital
• Difficult to distinguish allele frequency
  – Especially if not in 1:1 ratio
• Cost
  – $10,000,000 to sequence a 1Gbase genome to 10x coverage
Human genome project
Human Genome Project

- One of the largest scientific endeavors
  - Target accuracy 1:10,000 bases
  - Started in 1990 by DoE and NIH
  - $3Billion and 15 years
  - Goal was to identify 25K genes and 3 billion bases
- Used the Sanger sequencing method
- Draft assembly done in 2000, complete genome by 2003, last chromosome published in 2006
- Still being improved
Human Genome Project

How it was Accomplished

- Public Project
  - Hierarchical shotgun approach
  - Large segments of DNA were cloned via BACs and located along the chromosome
  - These BACs where shotgun sequenced

- Celera
  - Pure shotgun sequencing
  - Used public data (released daily) to help with assembly
Method 1: Hierarchical Sequencing

1. Sonication or enzymatic cut
2. Sizing to 2-3 kb
3. Cut and ligation
4. Infection of host cells
5. Cell growth and colony picking
6. Purified DNA
7. Massive sequencing (6-8 → redundancy)
8. Sequence assembly programs

Using Bacterial artificial chromosomes (BACs) to aid assembly
Using optical mapping approaches to aid assembly
Method 2: Celera Shotgun Sequencing

- Used paired-end strategy with variable insert size: 2, 10, and 50kbp
Paired-end sequencing

Genomic DNA fragments or cDNAs

- Cloning based
  - Vector
  - Linker
  - Restriction digest
  - Release ditags

- Cloning-free based
  - Self-circulation
  - Ligate to adaptors

PET Library

- Concatenated PETs
  - Dimerized PETs
  - Single PETs

Sequencing the PETs and mapping to the genome

"PETworkflow" by Jimhuang02 - Own work. Licensed under Public Domain via Commons - https://commons.wikimedia.org/wiki/File:PETworkflow.png#/media/File:PETworkflow.png
Paired-end reads are important

- Known Distance
- Read 1
- Read 2

- Repetitive DNA
- Unique DNA

Paired read maps uniquely

Single read maps to multiple positions
Using genetic maps to aid assembly

• Calculate recombination frequencies for microsatellites or sequence-tag-sites
• Can help to order assembled sequence scaffolds

https://sites.google.com/site/geneticmappingwiki/linkage_maps
Chromonomomer – Improving genetic maps for the NGS era

- Takes RAD and/or other marker data
- Assembly AGP file
- Alignment of markers to assembly (SAM/BAM)

- Re-order and if necessary break scaffolds to ensure assemblies are consistent with genetic map

Outcome of the HGP

- Spurred the sequencing of other organisms
  - 36 “complete” eukaryotes (~250 in various stages)
  - 1704 “complete” microbial genomes
  - 2685 “complete” viral genomes
- Enabled a multitude of related projects:
  - Encode, modEncode
  - HapMap, dbGAP, dbSNP, 1000 Genomes
  - Genome-Wide Association Studies, WTCCC
  - Medical testing, GeneTests, 23AndMe, personal genomes
  - Cancer sequencing, COSMIC, TCGA, ICGC
- Provided a context to organize diverse datasets
HGP Data Access

Results in GenBank, UCSC, Ensembl & others
Achievements Since the HGP

Genomic achievements since the Human Genome Project

Economic Impact of the Project

- Battelle Technology Partnership Practice released a study in May 2011 that quantifies the economic impact of the HGP was $796 billion!

- Genomics supports:
  - >51,000 jobs
  - Indirectly, 310,000 jobs
  - Adds at least $67 billion to the US economy

http://www.genome.gov/27544383
2004 onwards:
Beyond 1 species, 1 genome

• Cost of producing a single genome could vary from $100,000s to $10s of millions using capillary sequencers
• Labour intensive methodology

• New methods were required to lower the overall cost per genome
Second generation short read technologies
Large scale sequencing 2006

PRODUCTION

Rooms of equipment
Subcloning > picking > prepping
35 FTEs
3-4 weeks

SEQUENCING

74x Capillary Sequencers
10 FTEs
15-40 runs per day
1-2Mb per instrument per day
120Mb total capacity per day
Key advantages over Sanger Sequencing

- Hugely reduced labour requirements
  - No need to perform cloning
- Reduced Cost per sequence
- Reduced time to result
- Decentralisation
Worldwide sequencers

Fun fact

• Clive Brown
• Formerly director of Computational Biology at Solexa (Illumina)
• Chief Technology Officer at Oxford Nanopore
Illumina Sequencing By Synthesis

Illumina HiSeq

Illumina NextSeq

Illumina MiSeq
Step 1: Sample Preparation

1. Prepare genomic DNA sample
   Randomly fragment genomic DNA and ligate adapters to both ends of the fragments

   - **Active Chromatin**
   - **mRNA**
   - **0.5µg total RNA**
   - **ChIP-sequencing**
   - **10 ng**

2. **Genomic DNA**
   min 1ng
   - **Small RNA**
   - 1 ug total RNA
   - **Other Apps**

Step 2: Clonal Single Molecule Arrays

~1000 molecules per ~ 1 um cluster
~2 billion clusters per flowcell
1 cluster = 1 sequence

Random array of clusters

Attach single molecules to surface
Amplify to form clusters
Step 3: Sequencing By Synthesis (SBS)

Cycle 1:
- Add sequencing reagents
- First base incorporated
- Remove unincorporated bases
- Detect signal
- Deblock and defluor

Cycle 2-n: Add sequencing reagents and repeat
Illumina video
Under the hood:
Illumina Sequencing: How it looks

- 2 billion clusters per flow cell
- 100 microns
- 20 microns
Base calling from raw data

The identity of each base of a cluster is read off from sequential images.
Illumina Paired-End Sequencing

1. **Cluster amplification**
2. **Linearize DNA**
3. **Sequence 1st strand**
4. **Strand re-synthesis**
5. **Linearize DNA**
6. **Sequence 2nd strand**
7. **Read 1**
8. **Read 2**
Working with Paired Reads

- Applicable to different fragment size ranges
  - up to ~800 bp for standard paired-libraries
  - 2 - 20kb mate-pair libraries

Enables alignment software to assign unique positions to previously *non-unique* reads
Mate pair vs paired-end reads

- Paired end reads

Read 1  ~500bp  Read 2
Mate pair vs paired-end reads on

- Mate pair libraries
  - Purify fragments containing biotin moiety using Streptavidin beads
  - Create a standard Illumina library and sequence using paired-end reads
  - Physical fragment size is 500
  - Genomic distance between read 1 and read 2 is 19.5kb

= Biotin moiety
Illumina platforms

**Illumina HiSeq**
- 500Gbase/flowcell
- 8 human genomes
- 6 day run time
- High output or rapid run mode
- Read lengths up to 250bp
- Requires large numbers of samples (or large genomes) to obtain lowest cost
- 4 colour chemistry
- £750,000 incl 3 year servicing

**Illumina NextSeq 500**
- 90Gbase/flowcell
- 1 human genome
- 2 day run time
- High output or rapid run mode
- Read lengths up to 150bp
- 2-colour chemistry
- £250,000 incl 3 year servicing

**Illumina MiSeq**
- 15Gbase/flowcell
- 2 day run time
- Read lengths up to 300bp
- 4 colour chemistry
- £90,000 incl 3 year servicing
## Illumina Developments

<table>
<thead>
<tr>
<th>Run Mode</th>
<th>HiSeq 2500</th>
<th>HiSeq 3000</th>
<th>HiSeq 4000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run Mode</td>
<td>Rapid Run</td>
<td>High-Output</td>
<td>N/A</td>
</tr>
<tr>
<td>Flow Cells per Run</td>
<td>1 or 2</td>
<td>1 or 2</td>
<td>1</td>
</tr>
<tr>
<td>Output Range</td>
<td>10-300 Gb</td>
<td>50-1000 Gb</td>
<td>125-750 Gb</td>
</tr>
<tr>
<td>Run Time</td>
<td>7-60 hours</td>
<td>&lt;1-6 days</td>
<td>&lt;1-3.5 days</td>
</tr>
<tr>
<td>Reads per Flow Cell †</td>
<td>300 million</td>
<td>2 billion</td>
<td>2.5 billion</td>
</tr>
<tr>
<td>Maximum Read Length</td>
<td>2 x 250 bp</td>
<td>2 x 125 bp</td>
<td>2 x 150 bp</td>
</tr>
</tbody>
</table>
Illumina Miniseq

- $50,000
- 25 million reads per run
- 2 colour SBS chemistry
Increasing throughput

• NextSeq
  – Utilises 2-colour instead of 4-colour chemistry to reduce sequencing time

• HiSeq 2500, NextSeq and MiSeq
  – Clusters formed randomly on the surface of the flowcell

• HiSeq 3000, 4000
  – Clusters only form within nanowell
  – Patterned flowcells
2-colour chemistry

• Instead of using 4 different dyes for each nucleotide, use 2
• Label T as green and C as red
• Label A as green and red
• Label G with no dye
• Rely on cluster position to call G bases
2-colour chemistry

Green channel

Red channel

http://core-genomics.blogspot.cz/2014/01/nextseq-500s-new-chemistry-described.html
Advantages/disadvantages

• Advantages
  – Speed
  – Only two pictures need to be taken each cycle instead of four

• Disadvantages
  – Higher likelihood of errors
  – Difficult to calibrate guanine quality scores
  – With fragments shorter than read length, tendency is to call G with high quality scores
Patterned flowcells

Randomly clustered flowcell
(2500)

Patterned flowcell
(3000/4000)
Comparison
Advantages

• Removes the need to detect cluster location during first 4 cycles of sequencing

• Lower sensitivity to over-clustering
Advantages

• Allows for exclusion amplification to reduce the number of polyclonal clusters
• Utilises an electric field to transport labelled dNTPs to wells faster than amplicons can diffuse between wells
• 1 sequence per well
• Whichever sequence starts replicating first within a well will rapidly out-compete other sequences
• Removing the upper poisson-limit of random flowcell clustering (~37%)

Disadvantages

• Possibility to obtain large number of duplicated sequences across wells
  – Caused by seeding of adjacent wells
  – Can be caused by under-clustering
  – Still important to load correct concentrations

• Limits on DNA fragment length
Potential issues with Illumina sequencing

- Specific motifs which are difficult to sequence
  - GGC motif
  - Inverted repeats

- Now mostly resolved
  - Low diversity sequences
    - 16S/amplicon sequences
    - Custom adaptors with barcodes at 5’ end
    - Now a much reduced problem thanks to software updates
  - GC/AT bias
    - GC clusters are smaller than AT
    - (less of a problem post June 2011)

Why do quality scores drop towards the end of a read?
3 main factors

Schematic representation of main Illumina noise factors.
(a–d) A DNA cluster comprises identical DNA templates (colored boxes) that are attached to the flow cell. Nascent strands (black boxes) and DNA polymerase (black ovals) are depicted.
(a) In the ideal situation, after several cycles the signal (green arrows) is strong, coherent and corresponds to the interrogated position.
(b) Phasing noise introduces lagging (blue arrows) and leading (red arrow) nascent strands, which transmit a mixture of signals.
(c) Fading is attributed to loss of material that reduces the signal intensity (c).
(d) Changes in the fluorophore cross-talk cause misinterpretation of the received signal (blue arrows; d). For simplicity, the noise factors are presented separately from each other.

http://arep.med.harvard.edu/pdf/Fuller_09.pdf

Limits to Illumina technology

• Limitations:
  – Reagent degradation
  – Dephasing
    • Leads to higher error rates
    • A 1% loss of signal or polymerase error every cycle leads to only 35% correct signal after 100 cycles
  – Sequencing time is always governed by the cyclic nature of the instrument (one base at a time)
    • Ideally dispense with incorporate, image, wash cycles
  – Size of fragments which can be clustered on the flowcell
    • Read lengths beyond the size of the DNA fragment are useless
    • Inefficient clustering >800bp
  – PCR steps during library preparation or clustering will result in bias
Features of Illumina Sequencing

• 1 – 300 million sequences per run/lane (depending on platform and configuration)
• 36-300bp read lengths
• $0.01 - $0.1 per megabase
• Accuracy decreases along read length but ~0.5%
454 and Ion Torrent
Fun fact

• Jonathan Rothberg
• Set up 454 in late 90s
• Sold to Roche in 2007
• Founded Ion Torrent in 2007
• Superseded 454
• Sold to Life Tech in 2010
454 Step 1: Sample preparation

1. Genomic DNA is isolated and fragmented.
2. Adaptors are ligated to single stranded DNA
3. This forms a library
4. The single stranded DNA library is immobilised onto proprietary DNA capture beads
454 Step 2: Amplification

Water-based emulsion PCR
454 Step 3: Load emPCR products

Picotitre plate

- enrich for DNA + beads
- diameter of the wells allows for only 1 bead/well

Smaller beads (red) carrying immobilized enzymes required for pyrophosphate sequencing are deposited into each well.
454 Step 4: Pyro-sequencing

1. Nucleotides are pumped sequentially across the plate
2. ~ 1 million reads obtained during 1 run
3. Addition of nucleotides to DNA on a particular bead generates a light signal
454 Chemistry
Life Technology Ion Torrent

454-like chemistry without dye-labelled nucleotides
- No optics, CMOS chip sensor
- Up to 400bp reads (single-end)
- 2 hour run-time (+5 hours on One Touch)
- Output is dependent on chip type (314, 316 or 318)
- 318 (11M wells) >1Gbase in 3 hours
- $700 per run
- $50K for the instrument, plus $75k for additional One Touch station and Server
- Libraries not compatible with Ion Proton
Life Technology Ion Proton

- 454-like chemistry without dye-labelled nucleotides
- No optics, CMOS chip sensor
- Up to 200bp reads (single-end)
- 2 hour run-time (+8 hours on One Touch)
- Output is dependent on chip type (P1 or P2 coming soon)
- 60-80 million reads (P1)
- $1500 per run
- $150K for the instrument, plus $75k for additional One Touch station and Server
- Libraries not compatible with Ion Torrent
# Ion Torrent vs Ion Proton

## Comparison Table

<table>
<thead>
<tr>
<th>Feature</th>
<th>Ion PGM™ Sequencer</th>
<th>Ion Proton™ Sequencer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small Genomes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sets of Genes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene Expression</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whole Transcriptomes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human Genomes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human Genomes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

### Image Description

- **Ion PGM™ Sequencer**
  - ![Ion PGM™ Sequencer Image]
  - Features: Small Genomes, Sets of Genes, Gene Expression

- **Ion Proton™ Sequencer**
  - ![Ion Proton™ Sequencer Image]
  - Features: Whole Transcriptomes, Human Genomes, Human Genomes
Ion Torrent
Library prep

• 454 style library using emulsion PCR

water drop:
beads + DNA template + PCR reagents
- enrich for DNA + beads
- diameter of the wells allows for only 1 bead/well
Ion System
Homopolymer issues

• Unlike Illumina platforms 454/Ion Torrent does not contain a blocking agent
• At each cycle multiple bases of the same type can be incorporated
• E.g.
  – Sequence is AGTCCCCCT
  – The CCCC will be incorporated in a single cycle
Homopolymer issues
Limits to Ion technology

- Limitations:
  - Reagent degradation
  - Dephasing
    - Leads to higher error rates
    - A 1% loss of signal or polymerase error every cycle leads to only 35% correct signal after 100 cycles
  - Sequencing time is always governed by the cyclic nature of the instrument
    - Ideally dispense with incorporate, image, wash cycles
  - Size of fragments which can be amplified on beads
    - Read lengths beyond the size of the DNA fragment are useless
  - PCR steps during library preparation or clustering will result in bias
  - Inaccurate homopolymer calls
Features of Ion Torrent Sequencing

- 5-80 million sequences per run (depending on platform and configuration)
- 100-400 bp read lengths
- ~$0.1 per megabase
- Accuracy decreases along read length but ~0.5%-1%
- Note increased propensity for homopolymers (although much work has gone in to reduce the impact)
Third generation sequencers

• My definition: single-molecule DNA sequencing
• Currently only PacBio RS II is commercially available
Pacific Biosciences RS II
Introduction

• Based on monitoring a single molecule of DNA polymerase within a zero mode waveguide (ZMW)

• Nucleotides with fluorophore attached to phosphate (rather than base) diffuse in and out of ZMW (microseconds)

• As polymerase attaches complementary nucleotide, fluorescent label is cleaved off

• Incorporation excites fluorescent label for milliseconds -> nucleotide recorded
PacBio video
Zero mode waveguide

Free nucleotides

Laser and detector

Immobilised DNA polymerase
Reducing noise in the Zero Mode Waveguide (ZMW)

• Sequencing takes place in the ZMW
• Sequencing can only take place if one and only one DNA/polymerase complex is present in each ZMW
• Each ZMW is just 70nm wide
• Wavelength of laser light used to illuminate ZMW ~500nm
• Therefore light incident on ZMW will act as an *evanescent wave* and only penetrate the first ~30nm
• This reduces the amount of noise from fluorescence of non-incorporated fluorophores
Zero mode waveguide

Free nucleotides

Immobilised DNA polymerase

Laser and detector
Observing a single polymerase
What it looks like

PacBio ZMWs with single DNA strand
Ordered

Illumina DNA mono-colonal clusters
Unordered
Output statistics

• Approximately 50,000 sequences per SMRT flowcell
• 500Mb-1Gbase output per SMRT flowcell
  – $200 per run
• Library prep required
  – ~$500 per sample
  – ~1ug per sample for short 1-2kb reads
  – 20-50ug for 20-30kb reads
• Size selection required to get the longest reads
• Read lengths
  – Distribution
  – Mean 12kb up to 20-40kb
Polymerase read lengths

P6-C4 Chemistry

Data per SMRT® Cell: 500 Mb – 1 Gb

Half of data in reads: > 14 kb

Top 5% of reads: > 24 kb

Maximum read length: > 40 kb

Based on data from a 20 kb size-selected E. coli library using a 4-hour movie.
Each SMRT Cell yields ~50,000 reads.
Circular consensus sequencing

1. generate amplicon
2. ligate adaptors
3. sequence
4. data analysis
   - raw long read
   - processed long read
   - single-molecule fragments
   - circular consensus sequence (ccs)

Circular consensus sequencing

Standard Sequencing for Continuous Long Reads (CLR)

Large Insert Sizes

Generates one pass on each molecule sequenced

Circular Consensus Sequencing (CCS)

Small Insert Sizes

Generates multiple passes on each molecule sequenced
Note on read length and accuracy

- Error rate of a single read has a phred quality score of approximately 9 (~12% error rate)
- Therefore to obtain a phred quality above 30 we need to have at least 4 passes of the molecule
- With a median polymerase read length of 12kb, this means fragments with a size of approximately 3kb will be high quality

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4152597/
Note on read length and accuracy

• In short
  – Shorter reads will tend to be higher quality
  – Longer reads will tend to be lower quality

• Depending on the application this may or may not be a problem
  – Long, error prone fragments are OK for scaffolding genome assemblies
  – May not be so good if you have a 20kb amplicon
Issues to be aware of

• PCR chimeras and artefacts (affects all PCR amplicon methods)
• Partial adaptor ligation
Issues to be aware of

• Shorter sequences will be loaded preferentially
  – Need to enrich for long fragments using gel extraction (hence large starting material)
Issues to be aware of

- DNA quality
  - This is absolutely crucial as we are sequencing native DNA
  - You need to be aware of several things
  - What is the source of your DNA (plants, fungi, etc)
    - Fungi often have carbohydrates bound to DNA which are hard to remove
  - Method of DNA isolation
    - Any carry over of CTAB, phenol/chloroform etc will interfere with sequencing
  - Quantification of contaminants
    - Dye-based (e.g. Qubit)
    - Nanodrop absorbance at 260/230 (2.0-2.2) and 260/280 (1.8)
  - Run a to look at your material and check for degradation
    - Don’t be tempted to run degraded material if you want long and representative reads
Library prep notes from the field  
(Courtesy Jeremie Poschmann)
Fragment DNA

Shearing: hydro-shear, g-tube, needle, sonication etc.
Repair DNA

- Exonuclease treatment to remove single stranded DNA
- Make blunt ends for ligation
- Beware of:
  - DNA damage: alkylation, oxydation, UV-crosslinks, AP-sites, intercalating agents
  - DNA binders: polyphenols, secondary metabolites, pigments, polysaccharides
  - Polymerase inhibitors: salts (EDTA), phenol, alcohols
Example nanodrop curves

FIGURE 2. Spectra of purified DNA without contamination (A), and of the same DNA sample contaminated with guanidine (B) and phenol (C).

Estimating the quality of DNA

Different methods will often give different results – particularly when DNA is damaged or contaminated.
Ligate adapters

• Blunt end ligation:

Beware of chimeras:
Purify DNA

• Do size selection and multiple clean ups in order to remove smaller fragments and adapter chimeras
• Remove non circular DNA using Exo VII and III
• Size select with Blue Pippin
Take home message 1
You want nice and clean DNA

0.5Gb output

1.5Gb output
Take home message 2

You need a lot of it

Minimum requirements is 0.1 ug
Pacific Biosciences

• Advantages
  – Longer reads lengths (median 12kb up to 40kb with P6-C4 chemistry)
  – Cost per run is low ($250 per run plus $400 per library prep)
  – Same molecule can be sequenced repeatedly
  – Epigenetic modifications can be detected
  – Long reads enable haplotype resolution

• Disadvantages
  – Library prep still required (micrograms needed)
  – If you use PCR based methods – you are no longer sequencing true single molecules
  – Still enzyme based
  – Only 50,000 reads/run. 400-500Mb yield
  – High (12%) error rate per read (but consensus can reduce <1%)
  – $800k machine
  – Run time is anywhere between 40 mins-6 hours per SMRT cell
Bioinformatics Implications

- Relatively low data and high per base cost limits widespread use
- Can obtain useful 20-40kb fragments (P4-C6 chemistry)
- Best used in conjunction with error correction algorithms utilising shorter PacBio reads (or Illumina data)— e.g. Wheat D genome
- Excellent to assist scaffolding of genomes
- Able to generate complete bacterial genomes
- Has been used to generate higher eukaryote genomes (e.g. Drosophila, Human) but cost can be prohibitive
- Ability to sequence entire


Koren, Sergey; Schatz, Michael C; Walenz, Brian P; Martin, Jeffrey; Howard, Jason T et al. (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads
Nature biotechnology vol. 30 (7) p. 693-700

Genome assembly methods

Novel applications

• Epigenetic changes (e.g. Methylation) affect the amount of time a fluorophore is held by the polymerase

• Circularise each DNA fragment and sequence continuously
Epigenetic changes
Types of modifications

**Eukaryotes:**

- $6\text{-mA}$
- $5\text{-mC}$
- $5\text{-hmC}$
- $5\text{-fC}$
- $5\text{-eaC}$
- $U$
- Base J

**Prokaryotes:**

- $6\text{-mA}$
- $4\text{-mC}$
- $5\text{-mC}$
- PT

http://www.pacb.com/applications/epigenetics/
Full length transcript sequencing

- Requires polyA RNA
- Uses SMRTer approach
- Ability to sequence full length transcripts with no need for assembly
Circular consensus sequencing for rRNA or microsatellites

http://www.microbiomejournal.com/content/1/1/10
Long reads to resolve haplotypes
PacBio developments

- Sequel system due to ship in the coming months
- Generate approximately 7Gbase data per cell
- $350,000 (half the cost of an RSII)
- Chemistry and analysis should be similar to that described above
PacBio training resources

- [https://github.com/PacificBiosciences/Bioinformatics-Training/wiki](https://github.com/PacificBiosciences/Bioinformatics-Training/wiki)
Useful PacBio papers

• Resolving the complexity of the human genome using single-molecule sequencing
• Defining a personal, allele-specific, and single-molecule long-read transcriptome
Features of PacBio sequencing

- ~50,000 sequences per SMRT-cell
- Up to 40kb read lengths
- ~$5 per megabase
- Accuracy depends on DNA length ~0.5%-12% error rate
- Ability to sequence native DNA – relieving PCR biases
- Claimed random error profile
Nanopore sequencing
What is a nanopore?

- Nanopore = ‘very small hole’
- Electrical current flows through the hole
- Introduce analyte of interest into the hole ➞ identify “analyte” by the disruption or block to the electrical current

https://nanoporetech.com/science-technology/introduction-to-nanopore-sensing/introduction-to-nanopore-sensing
Detection

Types of pore

• Either biological or synthetic

• Biological
  – Lipid bilayers with alpha-haemolysin pores
  – Best developed
  – Pores are stable but bilayers are difficult to maintain

• Synthetic
  – Graphene, or titanium nitride layer with solid-state pores
  – Less developed
  – Theoretically much more robust
Nanopore DNA sequencing

• Theory is quite simple
• Feed a 4nm wide DNA molecule through a 5nm wide hole
• As DNA passes through the hole, measure some property to determine which base is present
• Holds the promise of no library prep and enormously parallel sequencing
• In practice this is not easy to achieve

http://thenerdyvet.com/category/tech/
Nanopore sequencing

• In practice, it is much harder
• Problems:
  – DNA moves through the pore quickly
  – Holes are difficult/impossible to design to be thin enough so that only one base is physically located within the hole
  – DNA bases are difficult to distinguish from each other without some form of labelling
  – Electrical noise and quantum effects make signal to noise ratios very low
  – Search space for DNA to find a pore is large

http://omicfrontiers.com/2014/04/10/nanopore
Nanopore sequencing

http://www.nature.com/nbt/journal/v30/n4/full/nbt.2181.html
Continuous base identification for single-molecule nanopore DNA sequencing

James Clarke¹, Hai-Chen Wu², Lakmal Jayasinghe³, Alpesh Patel¹, Stuart Reid³ and Hagan Bayley²

![Diagram of nanopore structure](image.png)

**Panel a**
- Residual pore current (pA) vs. time (s)
- Nucleotides identified: T T T A T AA C G C T C A C C C C T C T G C A T C T C

**Panel b**
- Event count vs. residual pore current (pA)
- Peaks for different nucleotides: dGMP, dTMP, dAMP, dCMP
Approaches to simplify nanopore sequencing

• Slow down movement of bases through nanopore
  – Use an enzyme to chop DNA up and sequence individual bases as they pass through a pore
  – And/or use an enzyme to slow the progress of DNA through a pore
  – Monitor capacitive changes in the bilayer

• Hybridize labels to single stranded DNA
  – Force the labels to disassociate as they pass through the pore
  – Detect the labels

Companies involved

- ONT is closest to commercialisation
Oxford Nanopore MAP programme

• MinIon Access Programme

• Provides access to several flowcells and reagents for library preparation

• Beta testing programme – open to all

• Registration fee of $1000
  – Access to a MinIon system and starter pack of reagents

• Additional flowcells for $500-$900 each
Oxford Nanopore
Oxford Nanopore platforms

MinION Mk 1
Up to 1-2Gbases/run

PromethION
Up to 20-30Gbase/run
DNA binding to membrane

Double stranded DNA fragment

Pore
Oxford Nanopore principle
DNA can be sequenced by threading it through a microscopic pore in a membrane. Bases are identified by the way they affect ions flowing through the pore from one side of the membrane to the other.

1. One protein unzips the DNA helix into two strands.
2. A second protein creates a pore in the membrane and holds an “adapter” molecule.
3. A flow of ions through the pore creates a current. Each base blocks the flow to a different degree, altering the current.
4. The adapter molecule keeps bases in place long enough for them to be identified electronically.
MinIon features

- 512 pores
- Library preparation is required
- Read lengths up to 40-50kb
  - Limited by input DNA
- Up to 2Gbase output
  - Complex flowcell manufacturing
  - Improved in the past few months
Library preparation

• Input requirements
  – Depends on fragment length required
  – Ideally upwards of 1ug of DNA
  – Low input option available – 20ng

• Issues – keeping long DNA fragments

• New kits attempting to minimise input requirements
Challenge of basecalling

- Hidden Markov model
- Only four options per transition
- Pore type = distinct kmer length

- Form probabilistic path through measured states currents and transitions
  - e.g. Viterbi algorithm
Challenge of 5-mer basecalling

- TTTTTATTTTTT vs TTTTTCTTTTT
Challenge of 5-mer basecalling

- CGTATTTCGAA vs CGTATCTCGAA
Basecalling 1D vs 2D reads

- Both the template and complementary strand can be sequenced
- This doesn’t always work
- If it does, the base-calling can be improved
- Different kmers at the same locus can improve basecalling
- Up to 40% of reads using latest chemistry are 2D
- Attempts are being made to modify the library preparation to increase the proportion of 2D reads
Lengths of 2D vs 1D reads
Read lengths

- Highly dependent on input DNA length
- Difficult to preserve DNA lengths
Longest perfect stretches

Best Perfect kmer distribution

Frequency

best perfect kmer length
Yield: 35947 reads and 71203856 base pairs.
Hard to read motifs

Perfect kmers K=5

Hard Kmers
AAAAA
ACCTA
ACCTC
AGCGC
AGCTA
AGGTC

Easy Kmers
AAGAA
ACGAA
CGTTC
CTTTC
GAACG
Coverage of an E.coli genome
MinIon Analysis Reference Consortium (MARC)

- Group of 20 labs evaluating MinIon performance using E.coli
- http://f1000research.com/articles/4-1075/v1
Yield of MinIon flowcells

Percentage of 2D pass reads produced over time.
Error rates of BWA-MEM EM-corrected alignments of target 2D base-calls.

Improvements

- FASTQ – per base quality values don’t make much sense
- Move towards 6mer basecalling
- Other types of model taking into account effects of bases sitting outside the pore
- Improvements to pore types
  - Utilise multiple pore types on a single flowcell
  - This would not make it a true single molecule sequencer since we would rely upon consensus (probably does not matter)
- Library preparation
- Flowcell reliability
- Access to basecalling API
Cost per megabase

http://wwwnc.cdc.gov/eid/article/22/2/15-1796_article

Sanger era

454 era

Illumina era

Ox. Nanopore Minlon

PacBio
Minlon for denovo assembly and variant calling

• *De novo* sequencing and variant calling with nanopores using PoreSeq

• Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome
Novel applications

• Portable in-situ sequencing

• Nanopore sequencing as a rapidly deployable Ebola outbreak tool

• Nanopore sequencing in microgravity

Now available... minipcr

Small, simple, accessible PCR
Portable unit you can take anywhere
Fits in the palm of your hands
Easy to use
Control via PC, Laptop or any Android device

Only £499
Novel applications

• Direct RNA sequencing
Novel applications

- Rapid pathogen detection in clinical settings
  - Rapid identification of viral pathogens
  - Rapid draft sequencing of Salmonella during a hospital outbreak
Novel applications

• Teaching aids

• Integration of mobile sequencers into a classroom
Novel applications

• Direct RNA sequencing
Software packages

• Tracking and managing MinIon data
  – MinoTour - http://minotour.github.io/minoTour

• Processing ONT data
  – Poretools https://github.com/arq5x/poretools
  – poRe http://sourceforge.net/projects/rpore/
  – NanoCorr http://schatzlab.cshl.edu/data/nanocorr/
  – Nanopolish https://github.com/jts/nanopolish/
  – PoreSeq - http://www.nature.com/nbt/journal/v33/n10/full/nbt.3360.html
  – MarginAlign - https://github.com/benedictpaten/marginAlign
  – Lordec https://www.gatb.fr/software/lordec/

• Alignment
  Sensitive but slow aligners
  • BLAST
  • BLAT
  • LAST
  • BWA with the right parameters

• Assembly
  – Possible to obtain ONT-only assembly
  – Error correction with other data types is also possible
  – offer an alternative to PacBio or Illumina synthetic long reads
Summary

- Very encouraging
- Portable applications have been demonstrated
- Error rate is ~12-15% for R8 2D reads
  - Some success with completing bacterial genomes using this data
- Probably 12 months before full commercial launch (I said this last year as well...)
- Fast-mode promises 2-3x increase in data output (at the probable cost of accuracy)
- Highly recommended review paper by MARC consortium
  - [http://f1000research.com/articles/4-1075/v1](http://f1000research.com/articles/4-1075/v1)
Opportunities

• Converting reference genomes to ‘wiggle-space’
• Developing ‘wiggle-space’ aligners
• Online ‘streaming’ bioinformatics
  – Analytics - one read at a time
• Developing complex sample -> sequencer ready protocols for use in the field
• Identifying sources of bias
General issues with nanopores

• Single base-pair resolution is not available
  – Typically 4-5 nucleotides fit into a nanopore
• Only one detector per DNA strand
  – One or two shots at detection
• Fast translocation of DNA through pore
• Small signal and high noise
• Bilayer stability
Nanowire alternatives

- QuantumDx QSEQ
Many others in development

- http://www.allseq.com/knowledgebank/sequencing-platforms
In conclusion

• We are mastering reading DNA (at least some of it)
• Now we are in a position to precisely edit and engineer biological systems
Thanks to:

Karen Moore
Jeremie Poschmann
Audrey Farbos
Paul O’Neill

Wellcome Trust

Contact me:

k.h.paszkiewicz@exeter.ac.uk
http://sequencing.exeter.ac.uk

Supported by Wellcome Trust