Modern Approaches to Sequencing

Dr Konrad Paszkiewicz, Head, Exeter Sequencing Service,
Director Wellcome Trust Biomedical Informatics Hub,
January 2014
Contents

• Review of Sanger Sequencing
• Timeline and impact of human genome project
• Second generation sequencing technologies
• Third generation sequencing technologies
• Nanopore sequencing technologies
USB Nanopore sequencer
1944: Avery, O.T., et al. “Studies on the chemical nature of the substance inducing transformation of Pneumococcal types”


1959 – First homeogenous DNA purified
Timeline

1970 – First discovery of type II restriction enzymes

1972: sequencing of the first gene from RNA by Walter Fiers

1976: sequencing of the first complete genome by Fiers (Bacteriophage MS2 which infects E.coli)

1977: Maxam AM, Gilbert W. "A new method for sequencing DNA".

1977: Sanger F, Nicklen S, Coulson AR. "DNA sequencing with chain-terminating inhibitors"
**Timeline**

1985-86: Leroy Hood use fluorescently labeled ddNTPs, set the stage for automated sequencing

1987: Applied Biosystems markets first automated sequencing machine (ABI 370)

1990: National Institutes of Health (NIH) begins large-scale sequencing trials ($0.75/base)

Human Genome Project (HGP) begins, $3-billion and 15 years

1995: Craig Venter at TIGR published the Haemophilus influenzae genome. First use of whole-genome shotgun sequencing

Timeline

1998: Green & Ewing publish “phred” base caller/scorer

2000: Sydney Brenner and Lynx Therapeutics publishes “MPSS”, parallelized bead-base sequencing tech, launches “Next-Gen”

2001: HGP/Celera draft assembly published in Nature/Science

2003: HGP “complete” genome released. ENCODE project launched

2004: 454 releases pyrosequencer, costs 6-fold less than automated Sanger sequencing
Illumina-era Timeline

1998: Shankar Balasubramanian and David Klenerman patent "A method for reproducing molecular arrays" and found Solexa

2001: £12 million series A funding


2004: Solexa acquires Solid phase DNA amplification method
Illumina-era Timeline

2006: Solexa release Genome Analyser I

2007: Illumina acquires Solexa

2008: Illumina releases GAII

2008: Human microbiome project launched

2010: Illumina HiSeq 2000 released

2011: MiSeq launched

2012: ENCODE publications

2012: Illumina HiSeq 2500 released

2013-2015: RIP 454
Review of Sanger Sequencing
Fred Sanger 1918-2013
Double Nobel laureate and developer of the dideoxy sequencing method, first published in December 1977. [Credit: Wellcome Images]

"Fred Sanger is a quiet giant, whose discoveries and inventions transformed our research world.” (A.Bradley, WTSI.)
The challenge of DNA sequencing

• 1953 – Double helix discovered
• 1971 - First DNA sequence determined (12bp!)
• 1977 – Sanger sequencing method published

• Why did it take so long?
Some possible reasons

• The chemical properties of different DNA molecules were so similar that it appeared difficult to separate them.

• The chain length of naturally occurring DNA molecules is much greater than for proteins and made complete sequencing seem unapproachable.

• The 20 amino acid residues found in proteins have widely varying properties that had proven useful in the separation of peptides.
  – Only four bases in DNA made sequencing a more difficult problem for DNA than for protein.

• No base-specific DNAases were known.
  – Protein sequencing had depended upon proteases that cleave adjacent to certain amino acids.

• DNA was considered boring compared to proteins.
Steps on the road to sequencing

• 1959 – First homeogenous DNA purified
• 1970 – First discovery of type II restriction enzymes
• 1972 – First RNA gene sequence published (lac operon)
• 1975 – Sanger first publishes his plus/minus method of sequencing (unable to distinguish homopolymers)
• 1977 – Maxam & Gilbert publish their method (could distinguish homopolymers)
• 1977 – Sanger publishes Dideoxy sequencing method
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.

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 is performed by chain breakage at specific nucleotides.
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

AGCTGCCCCG

A
- ddATP + four dNTPs
  - ddA
  - dAdGdCdTdGdCdCdCdG

C
- ddCTP + four dNTPs
  - dAdGddC
  - dAdGdCdTdGdCdCdC
  - dAdGdCdTdGdCdCdCdC

G
- ddGTP + four dNTPs
  - dAddG
  - dAdGdCdTddG
  - dAdGdCdTdGdCdCdCdddG

T
- ddTTP + four dNTPs
  - dAdGdCdTdGdCdCdCdG
Sanger di-deoxy method

Longer fragments

Shortest fragments
Sanger Sequencing

- 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, 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.

Chain terminates at ddG.
Dideoxy Method

• Run four separate reactions each with different ddNTPs
• Run on a gel in four separate lanes
• Read the gel from the bottom up
Improvements to Sanger’s original method

• Cycle sequencing is chain termination sequencing performed in a thermal cycler.
  – Requires a heat-stable DNA polymerase.

• Fluorescent dyes are multicyclic molecules that absorb and emit fluorescent light at specific wavelengths.
  – Examples are fluorescein and rhodamine derivatives.
  – For sequencing applications, these molecules can be covalently attached to nucleotides.
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 Version of the Dideoxy Method
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.
# First generation (Sanger) sequencing

<table>
<thead>
<tr>
<th>Throughput</th>
<th>50-100kb, 96 sequences per run</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read length</td>
<td>0.5-2kbp</td>
</tr>
<tr>
<td>Accuracy</td>
<td>high quality bases - 99%: ~900bp</td>
</tr>
<tr>
<td></td>
<td>very high quality bases - 99.9%:</td>
</tr>
<tr>
<td></td>
<td>~600bp</td>
</tr>
<tr>
<td></td>
<td>99.999%: 400-500bp</td>
</tr>
<tr>
<td>Price per raw base</td>
<td>~$200,000/Gb</td>
</tr>
</tbody>
</table>
Sanger Sequencing
Useful videos

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

• http://www.youtube.com/watch?v=bEFLBf5Wetc&feature=fvwrel
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
  - $3 Billion 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
Human Genome Project

This blog post indicates ~2.86Gbase/3.1Gbase of the non-redundant genome has been sequenced in hg18 or ~92% centromeres, telomeres, and highly repetitive regions left.

http://bit.ly/qML5Uq
How it was Accomplished

- Public Project
  - Hierarchical shotgun approach
  - Large segments of DNA were cloned via BACs and located along the chromosome
  - These BACs were shotgun sequenced
- Celera
  - Pure shotgun sequencing
  - Used public data (released daily) to help with assembly
Shotgun Sequencing

- **Celera**
  - Started in Sept 1999, goal was to do in $300M and 3 years what the public project was doing for $3B and 15 years!
  - Whole-genome shotgun sequencing
  - Used both whole-genome assembly and regional chromosome assembly
  - Incorporated data from the public project
  - Raised ethical concerns about the ownership of the human genome and patentability of genes
Hierarchical Sequencing

Hierarchical shotgun sequencing

Genomic DNA

BAC library

Organized mapped large clone contigs

BAC to be sequenced

Shotgun clones

Shotgun sequence

Assembly

Celera Shotgun Sequencing

- Used paired-end strategy with variable insert size: 2, 10, and 50kbp
HGP Data Access

Results in GenBank, UCSC, Ensembl & others
Growth of Genbank

December 2013  156,230,531,562 bases
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

Achievements Since the HGP

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
Second generation sequencing tech
Second generation sequencing definition

Illumina HiSeq
Illumina HiSeq Key Features

• Advantages
  – Large volume of data (300Gb per run)
  – Short run time (< 1 day)
  – Straightforward sample prep
  – Well established open source software community

• Disadvantages
  – Requires pooling of large numbers of samples to achieve lowest costs
  – Short reads (36-150bp)
Illumina Sequence By Synthesis

- Produces approximately 1.6 billion short reads (18bp-150bp) per flowcell
- Each run takes 2-9 days depending on the configuration
- Each flowcell is divided into either 2 or 8 separate lanes (channels)
Illumina HiSeq setup

Automated sample preparation

cBot Cluster generation

HiSeq 2500

Illumina HiSeq Flowcell

6cm
Illumina Sequencing

DNA sample preparation (over-simplified)

1) Extract DNA

2) Randomly shatter and PCR

3) Attach adapter sequence
4) Attach to flow-cell surface

5) PCR-amplify into clusters
Sequence clusters on the flow cell

Cluster 1

Cluster 2

Cluster 3

flow-cell surface
Sequencing cycle 1

add free adapters and dye-labelled bases
Sequencing cycle 1
Sequencing cycle 1

Fire laser

record intensities
Sequencing cycle 1

Wash to remove block
Sequencing cycle 2

add dye-labelled bases
Sequencing cycle 2

Fire laser

record intensities
Sequencing cycle 3

Fire laser

record intensities
Note

• The schema is over-simplified

• In reality
  – Sequencing is done from the top of the strand down towards the bottom of the flowcell
  – Reversible blocking agents are part of the modified dNTP fluorophores
Paired-End Sequencing

• Provides distance relationship between two reads
• Important for many applications
  – Characterise insertions, deletions, copy number variants, rearrangements
  – Required for De novo assembly
  – Enables sequencing across repeats
  – Useful for Di-Tags, cDNA sequencing, etc.
• Enables sample multiplexing (identifier tags)
• Increases output per flowcell
Working with Paired Reads

• Applicable to different fragment size ranges
  - up to ~600 bp for standard libraries
  - 2 - 20kb mate-pair libraries

Enables alignment software to assign unique positions to previously non-unique reads
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
Illumina Sequencing: How it looks

1.6 BILLION CLUSTERS PER FLOW CELL

100 MICRONS

A C G T

20 MICRONS
Base calling from raw data

The identity of each base of a cluster is read off from sequential images.

Current read lengths = 36-150 nt
Total sequence data for 1 paired-end run with 100bp = 300Gb!
HiSeq 2000 vs 2500 flowcells

HiSeq 2000
8 lanes
12 day run time

HiSeq 2500
2 lanes
2 day run time
## Comparison

<table>
<thead>
<tr>
<th>APPLICATION</th>
<th>RAPID RUN MODE</th>
<th>HIGH OUTPUT MODE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChIP-Seq Transcription Factor</td>
<td>40 Samples 7 Hours</td>
<td>200 Samples 2 Days</td>
</tr>
<tr>
<td>1 x 36 bp</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mRNA-Seq</td>
<td>24 Samples 16 Hours</td>
<td>120 Samples 5 Days</td>
</tr>
<tr>
<td>2 x 50 bp</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TruSeq Exome Seq 62 MB Region</td>
<td>15 Samples 27 Hours</td>
<td>85 Samples 12 Days</td>
</tr>
<tr>
<td>100x Coverage 2 x 100 bp</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human Whole Genome &gt;30x Coverage</td>
<td>1 Sample 27 Hours</td>
<td>5 Samples 12 Days</td>
</tr>
<tr>
<td>2 x 100 bp</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rapid run</td>
<td>Slow run</td>
<td></td>
</tr>
<tr>
<td>-----------------------------------------------</td>
<td>-----------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>48 genomes (£250 per sample)</td>
<td>48 genomes/lane (£210 per sample)</td>
<td></td>
</tr>
<tr>
<td>10 genomes (£510 per sample)</td>
<td>10 genomes/lane (£350 per sample)</td>
<td></td>
</tr>
<tr>
<td>8 genomes (£590 per sample)</td>
<td>8 genomes/lane (£400 per sample)</td>
<td></td>
</tr>
<tr>
<td>1 genome (£3400)</td>
<td>1 genome (£4000)</td>
<td></td>
</tr>
</tbody>
</table>
Potential issues with Illumina sequencing

• Low diversity sequences
  – 16S/amplicon sequences
  – Custom adaptors with barcodes at 5’ end

• GC/AT bias
  – GC clusters are smaller than AT
  – (less of a problem post June 2011)

• Specific motifs which are difficult to sequence
  – GGC motif
  – Inverted repeats

Low Diversity samples

[Graph showing data]

http://www.plosone.org/article/info:doi/10.1371/journal.pone.0016607
Low Diversity Example

• Cycle 1
Low diversity workarounds

- Multiplex with diverse samples and sequence across multiple lanes/runs
- Add PhiX control
- If dealing with amplicon, TraDIS or RAD-seq material, design multiple offset primers
Low diversity issue ‘fixed’

• Illumina are bringing out two improvements which are claimed to alleviate the problem

  – Software fix to prevent quality fall off due to incorrect phasing estimates. Available on MiSeq now.
    • Possibility of implementing those same changes on the HiSeq is uncertain

  – Ordered flowcells – base-caller will know apriori where clusters should be
Sequence-specific error motifs (GCC and inverted repeats)

Nakamura K et al. Nucl. Acids Res. 2011;nar.gkr344
Illumina MiSeq

- Same technology and chemistry as HiSeq
- 2X300bp reads
- 15 Gbase/run
- Run 48-72 hours
- $800-$1000 / run
- $100K instrument
- $50k for additional 2 year service contract
- Capable of sequencing 48 5Mb bacterial genomes per run
- Libraries compatible with HiSeq
Future Illumina developments

• 2x250bp reads (HiSeq fast run mode)
• Ordered flowcells
• 2x400bp reads (MiSeq)
• 10kb synthetic reads (approx. 5-6 million per lane)
  – Useful for phasing of haplotypes
  – Formed from short reads so repeat spanning is still problematic
Other equipment (optional)

- Agilent Bravo liquid handling robot £85k
- Agilent Tapestation £30k
- Covaris 96-well sonicator £90k
Roche 454 Key Features

• Advantages
  – Long read lengths (200-1000bp)
  – Multiple samples possible
  – Short run time (< 1 day)

• Disadvantages
  – Relatively expensive (~£8k per run)
  – Low volume of sequence data (100Mb-1Gb)
  – Complex sample prep
  – Roche discontinuing support from 2015
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

water drop:
beads + DNA template + PCR reagents
454 Step 3: Load emPCR products

- 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
SOLiD

• Differs from Illumina and 454
  – No dXTP reagents are used
  – Oligonucleotide primer-based sequencing is used
  – Two bases are read at a time
  – High accuracy

BUT – Only one colour is emitted
  Need several sequencing steps to convert colour to a sequence
Life Technologies SOLiD

• Advantages
  – Two base encoding system
  – Every base read twice
  – Large volume of sequence data (270Gb per run possible)

• Disadvantages
  – Short read lengths (30-80bp)
  – Complex sample prep
  – Bioinformatics support less comprehensive
  – Paired-end reads more complex than Illumina or 454
SOLiD: Step 1 Sample Prep

Sheared genomic DNA

P1 Adapter  P2 Adapter

clonally-PCR amplified DNA fragments

chem. modification: covalent bonding to slide
SOLiD: Step 2 Attach beads

3’-modified beads deposited onto glass slide

Sequential ligation with dye-labeled oligonucleotides
SOLiD: Step 3 Sequencing 1

1. Prime and Ligate

**PRIMER ROUND 1**

Universal seq primer (n) + Ligase

P1 Adapter

Template Sequence

*
SOLiD: Step 3 Sequencing 2

2. Image

3. Cleave off Fluor
4. Repeat steps 1-4 to Extend Sequence

A random primer is ligated to the template only when the labeled nucleotide complements the fifth nucleotide on the template, counting from the end of the previously ligated primer.
SOLiD Step 3 Sequencing 4

5. Primer Reset

Universal seq primer \((n-1)\)

1. Melt off extended sequence
2. Primer reset
SOLiD Step 3 Sequencing 5

1. Prime and Ligate

6. Repeat steps 1-5 with new primer
SoLID Colour space

Possible dinucleotides encoded by each color

1st Base
A
C
G
T

2nd Base
A
C
G
T

Template Sequence
TA
AC
AA
GA
TC
CG
CA
CC
TC
GC
GT
GG
AG
CT
Common features

• Most 2\textsuperscript{nd} generation platforms share the following:
  – Adaptor sequences to fix probes to a surface/bead
  – Amplification
  – Use of fluorescent probes/CCD devices or pH sensors
  – Capable of paired-end reads
  – Post-processing software to determine image quality
  – Shorter read lengths compared to traditional capillary based sequencers
  – Much higher data volumes (~Gb)
  – Sequence a human genome in a matter of days
Common features
Phred Score

- Phred program: http://en.wikipedia.org/wiki/Phred_base_calling
- \( Q = -10 \log_{10}(P) \)
- \( P = 10^{(-Q/10)} \)

<table>
<thead>
<tr>
<th>Phred Quality Score</th>
<th>Probability of Incorrect base call</th>
<th>Base call accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1 in 10</td>
<td>90 %</td>
</tr>
<tr>
<td>20</td>
<td>1 in 100</td>
<td>99 %</td>
</tr>
<tr>
<td>30</td>
<td>1 in 1000</td>
<td>99.9 %</td>
</tr>
<tr>
<td>40</td>
<td>1 in 10000</td>
<td>99.99 %</td>
</tr>
<tr>
<td>50</td>
<td>1 in 100000</td>
<td>99.999 %</td>
</tr>
</tbody>
</table>
Bioinformatics implications

- 100-10,000 fold increase in data volumes
- Tool development
- Data quality is poorer
- Less bioinformatics manpower available per sequencing project
- Finished genomes are usually of poorer quality than Sanger ‘gold-standard’ genomes
- Due to data volume, other applications have become feasible
- E.g. RNA-seq, ChIP-seq, Meth-Seq.
Benchtop sequencers
The NGS Market

- Currently dominated by Illumina (70% instruments)
- Market split into:
  - Low throughput but fast: clinical applications and sequencer for individual labs
  - Very high throughput: genome centers and large-scale projects
- E.g Illumina HiSeq 2000 vs. MiSeq
  - 300Gbase per 10 day run vs 7 Gbase in 48 hours

Benchtop sequencers

• Roche 454 Junior, Illumina Miseq are essentially miniature versions of the 454 and HiSeq
• Life Technologies Ion Torrent and Ion Proton are benchtop sequencers derived from 454 pyrosequencing
• Designed for individual groups
• Typical instrument cost is $150k (inc 3 year service contract)
• Typical run cost in consumables: $1000/run (at maximum output)
Illumina MiSeq

- Same technology and chemistry as HiSeq
- 2X250bp
- 7.5 Gbase/run
- Run 48 hours
- $800 / run
- $100K instrument
- $50k for additional 2 year service contract
- No additional wet-lab equipment required
- Capable of sequencing 20-30 bacterial genomes per run
- RNA-seq of up to 6 samples
- Libraries compatible with HiSeq
Roche 454 Junior

- Same chemistry
- 100K reads, 700bp
- 70 Mbases/run
- Focus on clinical, 510K validated assays
- $1000 per run
- $100K instrument

- Now uncompetitive – Roche reviewing future
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

<table>
<thead>
<tr>
<th>SMALL GENOMES</th>
<th>SETS OF GENES</th>
<th>GENE EXPRESSION</th>
<th>WHOLE TRANSCRIPTOMES</th>
<th>HUMAN GENOMES</th>
<th>HUMAN PROTEOMES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ion PGM™ Sequencer</td>
<td></td>
<td>Ion Proton™ Sequencer</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Ion Torrent
Library prep

- 454 style library using emulsion PCR
Ion Torrent

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

Two bases are incorporated

Two hydrogen ions are released

H⁺ H⁺
Benchtop sequencers

**Ion Proton (P1 chip)**
- 60-80M reads
- Up to single-end 200 base pair runs
- 16Gb/run
- 4 hour run time
- $1500/run
- $150K instrument
- One touch system required

**Illumina MiSeq**
- 30M reads
- 2X300bp
- 15 Gbase/run
- Run 48-72 hours
- $1000 / run
- $100K instrument
- No additional equipment required

**Roche 454 Junior**
- Same chemistry
- 100K reads, 700bp
- 70 Mbases/run
- Focus on clinical, 510K validated assays
- $1000 per run
- $100K instrument
- No additional equipment required
Useful benchtop review paper

Possible problems

• These are common to all platforms
  – Biases introduced by sample preparation
  – Errors in base-calling
  – High GC/AT biases can cause difficulties

• 454 and Ion Torrent have difficulty sequencing homopolymeric tracts accurately

• Latest Ion Torrent reagent upgrades claim to reduce these

• Illumina also has specific motifs which are difficult to sequence

Homopolymer errors

• Different between signal of 1 and signal of 2 = 100%.
• Different between signal of 5 and 6 is 20%
• More difficult to decide if we have AAAAA or AAAAAA
• Is the final sequence: ACTTNAAAAA or ANTTAAAAAAA or ATTAAAA.... Etc
Third generation sequencers
Third generation sequencers

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

• Based on monitoring a single molecule of DNA polymerase within a zero mode waveguide (ZMW)
  – 150,000 ZMWs on a SMRT flowcell on PacBio RSII
• 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
Library prep

**Sample Preparation**

- Genomic DNA (200bp to 10kb+)
  - WGA (10-20 ng)
  - Targeted Enrichment Products
- DNA Sample
- Fragment DNA
- Repair Ends
- Ligate Adapters
- Purify DNA
- Binding

**Building of SMRTbell**

- WGA (10-20 ng)
- cDNA
SMRT Cell
Observing a single polymerase
What it looks like

PacBio ZMWs with single DNA strand
Ordered

Illumina DNA mono-colonal clusters
Unordered
Output statistics

- Approximately 100,000-150,000 sequences per SMRT flowcell
- 300-500Mb output per SMRT flowcell
  - $500 per run
- Library prep required
  - ~$500 per sample
  - ~0.5ug per sample
- Size selection required to get the longest reads
- Read lengths
  - Distribution
  - Mean 8.5kb up to 20-25kb
Read lengths

P4-C2 Chemistry

P5-C3 Chemistry
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
Circular consensus sequencing
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 for rRNA or microsatellites
Issues to be aware of

- PCR chimeras (affects all PCR amplicon methods)
- Chimeric sequences can be generated during library preparation
- Shorter sequences can be loaded preferentially
  - Uniform amplicon size reduces this
  - PacBio Magbead loading system
Circular consensus sequencing (CCS)

- Raw error rates of a single pass read is high (10-15%)
- It is possible to read the same molecule repeatedly using CCS mode sequencing
- Can do this up to seven times to reduce error rates to around 0.1-0.5%
- Disadvantages
  - Reduction in read length proportional to number of passes (e.g. 7 passes – max read length 3kb).
  - Reduction of total number of reads as some ZMW polymerases will fail
Pacific Biosciences

• Advantages
  – Longer reads lengths (median 8.5kb up to 25kb with P5-C3 chemistry)
  – 40 minute run time
  – Cost per run is low ($400 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 required (micrograms needed)
  – If you use PCR based methods – it is NO LONGER single molecule
  – Enzyme based
  – Only 50,000 reads/run. 400-500Mb yield
  – High (10-15%) error rate per run (but CCS can reduce this to <~1%)
  – $750k machine
  – Lab requirements very stringent
Bioinformatics Implications

• Relatively low data and high per base cost limits practical widespread use
• Can obtain useful 20-25kb fragments (C5 chemistry)
• Best used in conjunction with error correction algorithms utilising shorter PacBio reads or Illumina data
• Excellent to help scaffold genomes
• Able to generate complete bacterial genomes

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

Hierarchical genome assembly

HGAP Example - *Meiothermus ruber* (JGI)

10 kb SMRTbell™ library

4 SMRT® Cells - 330 Mb (>100X)

Long seed reads (>5 kb) 92 Mb (30X)

*pre-assemble*

Pre-assembled long reads 61 Mb (20X)

Celera® Assembler

5 contigs

*clean-up (Minimus2)*

1 contig
PacBio training resources

- https://github.com/PacificBiosciences/Bioinformatics-Training/wiki
Nanopore sequencing
What is a nanopore?

- Nanopore = ‘very small hole’
- Electrical current flows through the hole
- Introduce analyte of interest into the hole \(\Rightarrow\) identify “analyte” by the disruption or block to the electrical current
What is a nanopore?

• 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 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
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 capacitative 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

- Company which appears closest to commercialisation
- Two approaches to sequencing
  - Exo-nuclease sequencing (originally part of a co-marketing agreement with Illumina)
  - Strand sequencing
- Both use synthetic membranes compatible with alpha-haemolysin derived pores
- Strand sequencing method is being commercialised
Continuous base identification for single-molecule nanopore DNA sequencing

James Clarke, Hai-Chen Wu, Lakmal Jayasinghe, Alpesh Patel, Stuart Reid, and Hagan Bayley

Nature Nanotechnology

ARTICLES

PUBLISHED ONLINE: XX XX 2009 DOI: 10.1038/NNANO.2009.12

Nucleotide Recognition
Exonuclease sequencing

Alpha-hemolysin protein pore

Exonuclease to chop DNA into constituent nucleotides
• Cyclodextrin inside alpha-hemolysin transiently binds to DNA base
• Interrupts the current through the pore
• Signal is indicative of base
Similar to Genia approach
Strand-sequencing

- Used in the recently advertised GRIDIon and MinIon systems
Novel applications

Adaptable protein nanopore:
- DNA Sequencing
- Proteins
- Polymers
- Small Molecules

Sensor array chip: many nanopores in parallel

Electronic read-out system
Oxford Nanopore Platforms

- **GridION for sequencing centres**
  - Promise of human genome in a few hours for around $1000
  - 2000 and 5000 pore instruments
  - No estimated pricing of instrument

- **MinIon for individual researchers**
  - $900 for 512 pore chip
  - 100Mb-1Gb per MinIon
  - Disposable after 6 hour run
  - 4% error rate in trials (mostly deletions)
MinIon Details

• Uses the strand sequencing technique
• Requires library preparation
• ‘Run –until’ technology
• More of a ‘sequencing sensor’ than a direct competitor to 2\textsuperscript{nd} generation sequencing
• Very useful for detection
• Likely to become as ubiquitous as a PCR-machine
Library preparation

• Transposase-based library preparation is still required
• Enzymatic biases will still be present and may be more difficult to detect with lower number of reads
• Efficiency of transposase may limit maximum read lengths
• Unclear whether system can be washed effectively part-way through a run to load different samples
Library preparation – Step 1

http://erlichya.tumblr.com/
Library preparation – Step 2

http://erlichya.tumblr.com/
Library preparation – Step 3

http://erlichya.tumblr.com/
Caveats
Cost per megabase

![Graph showing cost per megabase over time with different eras: Sanger era, 454 era, Illumina era, PacBio, Ox Nanopore MinIon.](image)
Oxford Nanopore is not single molecule

- The lipid bi-layers contain different types of nanopore
- Each has a different error profile
- It will still be necessary to over-sample and use sequences determined from complementary nanopores to reduce the overall error rate
- Will still likely need minimum of 5-10x coverage per genome (5-10 bacterial genomes per run)
Oxford nanopore

• Potential Advantages
  – Long reads lengths (10s – 100s kb)
  – Protein $\rightarrow$ solid-state upgrades may eliminate reagent costs (3-5 years)
  – Fast turn around
  – Could measure epigenetic modifications and other molecules

• Potential Disadvantages
  – Potentially non-stochastic errors (i.e. some sequences harder to sequence accurately)
  – Library prep required
  – Not single molecule
  – Cost per base is ~$10
Bioinformatics Implications

• Will prove to be yet another step change as with 2\textsuperscript{nd} generation sequencing
• Could obtain $>100$kb fragments
• Denovo assembly and phasing will be made easier
• Low number of reads per run and high per base cost may not make it useful for standard RNA-seq
• Burden will shift even further towards data management and downstream annotation
• ...it will lead to different bottlenecks
Min Ion Access Programme

MinION™ Access Programme

In late November 2013, Oxford Nanopore opened registration for a MinION Access Programme (MAP - product preview). This is a substantial but initially controlled programme designed to give life science researchers access to nanopore sequencing technology at no risk and minimal cost.

MAP participants will be at the forefront of applying a completely novel, long-read, real-time sequencing system to existing and new application areas. MAP participants will gain hands-on understanding of the MinION technology, its capabilities and features. They will also play an active role in assessing and developing the system over time. Oxford Nanopore believes that any life science researcher can and should be able to exploit MinION in their own work. Accordingly, Oxford Nanopore is accepting applications for MAP participation from all1,2.

About the programme

A substantial number of selected participants will receive a MinION Access programme package. This will include:

- At least one complete MinION system (device, flowcells and software tools).
- MAP participants will be asked to pay a refundable $1,000 deposit on the MinION USB device, plus shipping.
- Oxford Nanopore will provide a regular baseline supply of flowcells sufficient to allow frequent usage of the system. MAP participants will ONLY pay shipping costs on these flowcells. Any additional flowcells required at the participants’ discretion may be available for purchase at a MAP-only price of $999 each plus shipping and taxes.
- Oxford Nanopore will provide Sequencing Preparation Kits. MAP participants may choose to develop their own sample preparation and analysis methods; however, at this stage on an unsupported basis.

https://www.nanoporetech.com/technology/the-minion-device-a-miniauturised-sensing-system/map-application-form
Useful papers/videos

Beyond nanopores
General issues with nanopores

• Single base-pair resolution is not available
  – Typically 3-4 nucleotides fit into a nanopore
• Only one detector per DNA strand
• Fast translocation of DNA through pore
• Small signal and high noise
• Unstable lipid bilayers
Nanowire alternatives

- QuantumDx QSEQ
Many others in development

- http://www.allseq.com/knowledgebank/sequencing-platforms
Sequencing – back on the benchtop

1980

2000

2015?
Thanks to:

Audrey Farbos
Karen Moore

Wellcome Trust

Contact me:
k.h.paszkiewicz@exeter.ac.uk