Next-Generation Sequencing: an overview of technologies and applications

Matthew Tinning
Australian Genome Research Facility

July 2013
A QUICK HISTORY OF SEQUENCING
A quick history of sequencing

1869 – Discovery of DNA
1909 – Chemical characterisation
1953 – Structure of DNA solved
1977 – Sanger sequencing invented
   – First genome sequenced – ΦX174 (5 kb)
1986 – First automated sequencing machine
1990 – Human Genome Project started
1992 – First “sequencing factory” at TIGR
A quick history of sequencing

1995 – First bacterial genome – *H. influenzae* (1.8 Mb)
1998 – First animal genome – *C. elegans* (97 Mb)
2003 – Completion of Human Genome Project (3 Gb)
   – 13 years, $2.7 bn
2005 – First “next-generation” sequencing instrument
2013– >10,000 genome sequences in NCBI database
A quick history of sequencing

• 1977
  - First genome (ΦΧ174)
  - Sequencing by synthesis (Sanger)
  - Sequencing by degradation (Maxam-Gilbert)
Sanger sequencing: chain termination method

• Uses DNA polymerase
• All four nucleotides, plus one dideoxynucleotide (ddNTP)
• Random termination at specific bases
• Separate by gel electrophoresis
Sanger sequencing: chain termination method

Incorporation of di-deoxynucleotides terminates DNA elongation

Individual reactions for each base
Sanger sequencing: chain termination method

TCTGATGCAT*
TCTGATGCATGAACT*
TCTGATGCATGAACTGCT*
TCTGATGCATGAACTGCTCAT*
AGACTACGTACTTGACGAGTAC . . . . . .
Sanger sequencing: chain termination method

Separation of fragments by gel electrophoresis
Sanger sequencing: dye-terminator sequencing

1986: 4 Reactions to 1 Lane fluorescently labelled ddNTPs
Sanger sequencing: dye-terminator sequencing

Automated DNA Sequencers

ABI 377 Plate Electrophoresis  
ABI 3730 xl Capillary Electrophoresis
Sanger sequencing: dye-termination sequencing
Sanger sequencing: dye-termination sequencing

- Maximum read length: ~900 base
- Maximum yield/day: < 2.1 million bases (rapid mode, 500 bp reads)

< 0.1% of the human genome
> 1000 days of sequencing for a 1 fold coverage...
Sanger sequencing: shotgun library preparation

DNA extraction

DNA fragmentation

Clone into Vectors

Transform bacteria, grow, isolate vector DNA

Sequence the library

Assemble contiguous fragments
Human Genome Project

• Launched in 1989 – expected to take 15 years
  – Competing Celera project launched in 1998

• Genome estimated to be 92% complete
  – 1st Draft released in 2000
  – “Complete” genome released in 2003
  – Sequence of last chromosome published in 2006

• Cost: ~$3 billion
  – Celera ~$300 million
Human Genome Project
NEXT GENERATION SEQUENCING
Next-gen sequencing technologies

- Four main technologies
- All massively parallel sequencing
  - Sequencing by synthesis
  - Sequencing by ligation
- Mostly produce short reads - from <400bp
- Read numbers vary from ~ 1 million to ~ 1 billion per run
Next-gen sequencing technologies

• With massively parallel sequencing new methods for sequencing template preparation is required

• Current NGS platforms utilize clonal amplification on solid supports via two main methods:
  – *emulsion PCR (emPCR)*
  – *bridge amplification (DNA cluster generation)*
Next-gen sequencing technologies

Roche

Applied Biosystems
an Applera Corporation Business

illumina®

agraf
Next-gen sequencing technologies

Roche GS-FLX

Life Technologies SOLiD

Illumina HiSeq

Life Technologies Ion Torrent/Proton
Roche GS-FLX
Next-gen sequencing: shotgun library preparation
emPCR

Emulsion PCR is a method of clonal amplification which allows for millions of unique PCRs to be performed at once through the generation of micro-reactors.
emPCR

The Water-in-Oil-Emulsion
Pyrosequencing
Massively Parallel Sequencing
454: Data Processing

Example of a Flowgram

Key sequence = TCAG for identifying wells and calibration

T Base Flow → A Base Flow → C Base Flow → G Base Flow

Raw Image Files → Image Processing → Base-calling → Quality Filtering → SFF File
# 454 Platform Updates

<table>
<thead>
<tr>
<th>Platform</th>
<th>Read Length</th>
<th>Mbp per Run</th>
<th>Run Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>GS20</td>
<td>100bp</td>
<td>~20Mbp</td>
<td></td>
</tr>
<tr>
<td>GS-FLX</td>
<td>250bp</td>
<td>~100 Mbp</td>
<td>7.5 hrs</td>
</tr>
<tr>
<td>GS-FLX Titanium</td>
<td>400bp</td>
<td>~400 Mbp</td>
<td>10 hrs</td>
</tr>
<tr>
<td>GS-FLX Titanium Plus</td>
<td>700bp</td>
<td>~700 Mbp/Run</td>
<td>18 hrs</td>
</tr>
<tr>
<td>GS Junior</td>
<td>400 bp</td>
<td>~35Mbp</td>
<td>10 hrs</td>
</tr>
</tbody>
</table>
454 Sequencing Output

- *.sff (standard flowgram format)
- *.fna (fasta)
- *.qual (Phred quality scores)
Illumina HiSeq
Illumina Sequencing Technology

Robust Reversible Terminator Chemistry Foundation

DNA (0.1-1.0 ug)

Sample preparation

Cluster growth

Sequencing

1 2 3 4 5 6 7 8 9

Image acquisition

Base calling
Illumina: Data Processing

1. Nucleotide Flows
2. Raw Images
3. Image Processing
4. Base-calling
5. Quality Filtering
6. .bcl
## Platform Updates

<table>
<thead>
<tr>
<th>Platform</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solexa 1G</td>
<td>18bp reads, ~1Gbp / run</td>
</tr>
<tr>
<td>Illumina GA</td>
<td>36bp reads ~3Gbp / run</td>
</tr>
<tr>
<td>Illumina GAIi</td>
<td>75bp paired ends ~10Gbp / run (8 days)</td>
</tr>
<tr>
<td>Illumina GAIiX</td>
<td>75bp paired end reads ~40Gbp / run (8 days)</td>
</tr>
<tr>
<td>Illumina HiSeq 2000</td>
<td>100 bp paired end reads ~200 Gbp/ run (10 days)</td>
</tr>
<tr>
<td>Illumina HiSeq, v3 SBS</td>
<td>100bp paired end reads ~600Gbp / run (12 days)</td>
</tr>
<tr>
<td>Illumina HiSeq 2500 (Rapid)</td>
<td>150 bp paired end reads ~ 180 Gbp/ run (2 days)</td>
</tr>
<tr>
<td>MiSeq</td>
<td>250 bp paired end reads ~8 Gb/run (2 days)</td>
</tr>
</tbody>
</table>

Maximum yield / day 50,Gbp
~16x the human genome
Illumina Sequencing Output

- *.fastq (sequence and corresponding quality score encoded with an ASCII character, phred-like quality score + 33)
1. unique instrument ID and run ID
2. Flow cell ID and lane
3. tile number within the flow cell lane
4. 'x'-coordinate of the cluster within the tile
5. 'y'-coordinate of the cluster within the tile
6. the member of a pair, /1 or /2 (paired-end or mate-pair reads only)
7. N if the read passes filter, Y if read fails filter otherwise
8. Index sequence
Applied Biosystems SOLiD
Sequencing by Ligation
Base Interrogations
2 Base encoding

Possible Dinucleotides Encoded By Each Color

Template Sequence

AT
AC
AA
GA
CG
CA
CC
TC
GC
GT
GG
AG
TA
TG
TT
CT

Double Interrogation

With 2 base encoding each base is defined twice
emPCR and Enrichment

3’ Modification allows covalent bonding to the slide surface
## Platform Updates

<table>
<thead>
<tr>
<th>Platform</th>
<th>Reads Type</th>
<th>Yield</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOLiD 3</td>
<td>50bp Paired</td>
<td>50 Gbp / run</td>
<td>12 days</td>
</tr>
<tr>
<td>SOLiD 4</td>
<td>50bp Paired</td>
<td>100 Gbp / run</td>
<td>12 days</td>
</tr>
<tr>
<td>5500xl</td>
<td>75bp Paired</td>
<td>300 Gbp / run</td>
<td>14 days</td>
</tr>
</tbody>
</table>

Maximum yield / day 21,000,000,000bp
7x the human genome
3.5 hours of sequencing for a 1 fold coverage.....
SOLiD Colour Space Reads

- *.csfasta (colour space fasta)
- *.qual (Phred quality scores)

>853_17_1660_F3
T32111011201320102312......

<table>
<thead>
<tr>
<th>AA</th>
<th>CC</th>
<th>GG</th>
<th>TT</th>
<th>Colour</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC</td>
<td>CA</td>
<td>GT</td>
<td>TG</td>
<td>Blue</td>
</tr>
<tr>
<td>AG</td>
<td>CT</td>
<td>GA</td>
<td>TC</td>
<td>Green</td>
</tr>
<tr>
<td>AT</td>
<td>CG</td>
<td>GC</td>
<td>TA</td>
<td>Yellow</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Red</td>
</tr>
</tbody>
</table>
Applied Biosystems: Ion Torrent PGM
Ion Torrent

- Ion Semiconductor Sequencing
- Detection of hydrogen ions during the polymerization DNA
- Sequencing occurs in microwells with ion sensors
- No modified nucleotides
- No optics
Ion Torrent

DNA → Ions → Sequence
- Nucleotides flow sequentially over Ion semiconductor chip
- One sensor per well per sequencing reaction
- Direct detection of natural DNA extension
- Millions of sequencing reactions per chip
- Fast cycle time, real time detection
## Ion Torrent: System Updates

<table>
<thead>
<tr>
<th>Chip</th>
<th>100bp reads ~10 Mb/run (1.5 hrs)</th>
<th>100 bp reads ~100 Mbp / run (2 hrs)</th>
<th>200 bp reads ~200 Mbp/run (3 hrs)</th>
<th>200 bp reads ~1 Gbp / run (4.5 hrs)</th>
<th>100 bp reads ~8 Gbp/run</th>
</tr>
</thead>
<tbody>
<tr>
<td>314 Chip</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
</tr>
<tr>
<td>316 Chip</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
</tr>
<tr>
<td>318 Chip</td>
<td>•</td>
<td></td>
<td>•</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P1 Chip</td>
<td>•</td>
<td></td>
<td>•</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Ion Torrent Reads

- *.sff (standard flowgram format)
- *.fastq (sequence and corresponding quality score encoded with an ASCII character, phred-like quality score + 33)
Rapid Innovation Driving Cost Down

Evolution of NGS system output

Cost per Human Genome

Throughput (GB)

300GB

Cost per Genome

National Human Genome Research Institute

Cost per sequence reads

Moore's Law

$100,000,000

$10,000,000

$1,000,000

$100,000

$10,000

$1,000

$100

$10

$1

2007 2008 2009 2010
Summary of NGS Platforms

- Clonal amplification of sequencing template
  - emPCR (454, SOLiD and Ion Torrent)
  - Bridge amplification (Illumina)
- Sequencing by Synthesis
  - 454 Pyrosequencing
  - Illumina Reversible Terminator Chemistry
  - Ion Torrent Ion Semiconductor Sequencing
- Sequencing by ligation
  - SOLiD – 2 base encoding
- Dramatic reduction in cost of sequencing
  - GS-FLX provides > 100x decrease in costs compared to Sanger Sequencing
  - HiSeq and SOLiD > 100x decrease in costs over GS-FLX
NEXT GENERATION SEQUENCING APPLICATIONS
Applications

- DNA
  - whole genome
    - Shotgun & Mate Pair
  - targeted re-sequencing
    - hybrid capture
    - amplicon
  - ChIP-seq

- RNA
  - mRNA
  - whole transcriptome
  - small RNA
Sample preparation

- mRNA
  - Chemical Fragmentation
  - cDNA Synthesis

- DNA
  - Mechanical Fragmentation
  - Ligation of Amplification/Sequencing Adaptors

- Library Fragment Size Selection
Next-gen sequencing: shotgun library preparation

**Shotgun libraries**
- Whole genome sequencing
  - Input: 100-1,000 ng of DNA
  - shear DNA (<1,000 bp)
  - *End repair*
  - *A-tailing*
  - *Ligation of sequencing adapters*
Next-gen sequencing: shotgun library preparation

**Mate pair libraries**
- scaffolding and structural variation
  - Input: 5-20 µg of DNA
  - Shear DNA to 3kb, 8kb and 20Kb fragments
  - Ligation of biotinylated circularization adapters
  - Shear circularized DNA
  - Isolate biotinylated mate pair junction
  - Ligate sequencing adapters
Whole Genome Sequencing

- *de novo* assembly
- Reference Mapping
  - SNVs, rearrangements
- Comparative genomics

*E. coli* assembly from MiSeq Data
Illumina application notes
RNA-seq (cDNA libraries)

- **Shotgun cDNA library of**
  - Isolation of Poly(A) RNA or removal of rRNA
  - (100 ng – 4 ug of total RNA)
  - Chemical fragmentation of RNA
  - Random primed cDNA Synthesis & 2\textsuperscript{nd} strand Synthesis
  - Follows standard “DNA” library protocol

- **Stranded cDNA libraries**
  - 2\textsuperscript{nd} Strand “Marking” incorporation of dUTP in place of dTTP during second strand synthesis.
  - Selective enrichment for non-uracil containing 1\textsuperscript{st} cDNA strand by
    - Use of a polymerase that cannot amplify uracil containing templates

- **Small RNA Sample Preparation**
  - RNA-adaptor ligation before cDNA synthesis
  - Small RNA size selection via PAGE
    - Library fragment ~145-160bp (insert 20-33 nucleotides)
RNA-seq applications

- Gene Expression
- Alternative Splicing & Allele Specific Expression
- Transcriptome Assembly
Targeted re-sequencing: hybrid capture

- Enrichment for specific targets via capture with oligonucleotide baits
  - Exome Capture
    - Capture 40-70 Mb of annotated exons and UTRs
  - Custom Capture
    - up to 50 Mb of target sequences
Targeted re-sequencing: amplicons

- Preparation of amplicons tagged with sequencing adapters
  - Well suited for 454 and bench top sequencers
  - Deep sequencing for detection of somatic mutations
  - 16S Sequencing for microbial diversity
SUMMARY
Summary

- Next generation sequencing (NGS) is massively parallel sequencing of clonally amplified templates on a solid surface.
- NGS platforms generate millions of reads and billions of base calls each run.
- There are four main sequencing methods:
  - Pyrosequencing (454)
  - Reversible terminator sequencing (Illumina)
  - Sequencing by ligation (SOLiD)
  - Semiconductor sequencing (Ion Torrent)
- NGS reads are typically short (<400 bp).
- Next generation sequencing is used for a range of applications including:
  - Sequencing whole genomes
  - Sequencing specific genes or genomic regions
  - Gene expression analysis
  - Study of epigenetics