Next-Generation Sequencing Overview: Resources Available At The DNA Facility

Kevin L. Knudtson, PhD
Director, IIHG DNA Research Facility
Iowa Institute of Human Genetics
Carver College of Medicine

IIHG Bioinformatics Course 2013
July 29th 2013

http://www.medicine.uiowa.edu/humangenetics
# Evolution of DNA Sequencers

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>S35 ddNTPs</strong></td>
<td></td>
<td>Fluorescent ddNTPs</td>
<td>Fluorescent ddNTPs</td>
<td>Fluorescent ddNTPs</td>
</tr>
<tr>
<td><strong>Gels</strong></td>
<td></td>
<td>Gels</td>
<td>Capillaries*</td>
<td>Capillaries*</td>
</tr>
<tr>
<td><strong>Manual loading</strong></td>
<td></td>
<td>Manual loading</td>
<td>Robotic loading*</td>
<td>Robotic loading*</td>
</tr>
<tr>
<td><strong>Manual base calling</strong></td>
<td></td>
<td>Automated base calling*</td>
<td>Automated base calling*</td>
<td>Automated base calling*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Breaks down frequently</td>
<td>Reliable*</td>
</tr>
</tbody>
</table>

**1st Generation DNA Sequencers**

Sanger dideoxy-based
Evolution of DNA Sequencers

1st Generation DNA Sequencers
Sanger dideoxy-based

2005
Roche GS20
Aka “the 454”

2nd (Next) Generation Sequencers
“Single Molecule Detection After Amplification”

Early 2007
Solexa 1G

Late 2007
ABI SOLiD

“Long-Read” Instrument
300 base reads
400K reads/run
70 million bases/run

“Short-Read” Instrument
36 Base reads
40M reads/run
1 billion bases/run

“Short-Read” Instrument
35 Base reads
40M reads/run
1 billion bases/run

1992-1999
1999
2003

“old fashioned way”

Pre-1992
ABI 373/377
ABI 3700
ABI 3730XL

Fluorescent ddNTPs
Capillaries
Robotic loading
Automated baseline calibration

Reliable*

Fluorescent ddNTPs*
Capillaries*
Robotic loading*
Automated baseline calibration*
Evolution of DNA Sequencers

1st Generation DNA Sequencers
Sanger dideoxy-based

2nd (Next) Generation Sequencers
“Single Molecule Detection After Amplification”

3rd (Next-Next) Generation Sequencers
“Single Molecule Detection”

1992-1999

1999

2003

“old fashioned way”

Pre-1992

ABI 373/377

ABI 3700

ABI 3730XL

Fluorescent ddNTPs

Capillaries

Robotic loading

Automated baseline calling

Reliable*

Fluorescent ddNTPs

Capillaries*

Robotic loading*

Automated baseline calling

Breaks down frequently

1st Generation DNA Sequencers
Sanger dideoxy-based

2nd (Next) Generation Sequencers
“Single Molecule Detection After Amplification”

3rd (Next-Next) Generation Sequencers
“Single Molecule Detection”

2008

Helicos HeliScope

2010

PacBio RS

2012

Oxford Nanopore GridION/MinION

tSMS
True Single-Molecule Sequencing
Short-read system
35 base average read
~25 Gb/run
Up to 4,800 samples/run

SMRT Sequencing
Single-molecule Real-time
Long-Read System
3 kb average read, up to 10 kb

Nanopore Sensing
Scalable for long- or short-read
Evolution of NGS Platforms

2nd (Next) Generation Sequencers
“Single Molecule Detection After Amplification”

Roche GS-FLX

Illumina GAIIx

ABI SOLiDv4

GS-FLX Titanium

Illumina HiSeq2000

ABI 5500xl

Junior

Roche GS-FLX+

Illumina MiSeq

Illumina HiSeq2500

ABI Ion Torrent

PGM

Proton
Advantages and Challenges of NGS

**Advantages**

- No sub-cloning, no need for bacterial host
  - Less cloning bias
  - Easier to create more libraries
- Large number of individual sequence reads generated
  - Permits quantification by counting # of reads
  - Enhanced dynamic range
  - Permit detection of rare variants
- Readily adaptable to a variety of applications
  - Whole genomes (DNA-Seq)
  - Whole transcriptomes (RNA-Seq)
    - Novel Transcripts
    - Novel Splice Variants
    - Strand-Specific expression
  - DNA-protein binding interactions (ChIP-Seq)
  - Epigenomes (Methyl-Seq)
  - Small RNA (miRNA-Seq)
- Landscape is changing rapidly
  - Dramatic decrease in cost and speed of sequencing data
  - Permitted experiments that were previously cost prohibitive
  - Revolutionized the way we do sequencing

**Challenges**

- Landscape is changing rapidly
  - Platforms
  - Analysis software
  - Sequencing chemistries
  - Difficult to have continuity within projects
- Vendors overstate performance
  - Fail to meet performance expectation deadlines
  - Fragile instruments
- Data storage
  - Large data files expensive to store
- Data analysis
  - High power computing
  - Not do-it-yourself
Store the Data or Repeat the Experiment?

**Moore’s Law**: number of components in integrated circuits had doubled every year from the invention of the integrated circuit.

**Kryder’s Law**: Magnetic disk areal storage density doubles annually.

Lincoln Stein, *Genome Biology* 2010, **11**:207
DNA Facility Resources & Services

Services

• Synthetic Oligonucleotide

• DNA Sequencing/Fragment Analysis
  o Capillary-based Sanger dideoxy sequencing

• DNA Microarrays
  o Affymetrix GeneChip System
  o Illumina BeadArray System

• Quantitative “Real-time” PCR
  o Microvolume Systems (e.g., ABI 7900HT)
  o Nanovolume System (Fluidigm EP1)

• Molecular Biology Computing
  o Metaminer GeneGo Pathway Analysis
  o DNAStar LaserGene

• Genome sequencing with
  o Roche GS-FLX+ (aka ‘the 454’)
  o Illumina HiSeq 2000 & MiSeq
  o LifeTech/Ion Torrent-Personal Genome Machine (PGM)

Resources

• Nanodrop Spectrophotometer & Qubit Fluorometer
• Bioanalyzer-Assess nucleic acid quality
• Covaris S2 & E220 Sonicator-Nucleic acid fragmentation
• 2 Caliper SciClone & Beckman Biomek FX liquid handling robots
• Pippen Prep-Automated DNA fragment extraction from gel
Roche Genome Sequencer FLX System
(Long-Read System)

Standard FLX Chemistry
- 100 million bases/full plate run
- 400,000 sequence reads
- 200–400 bases per read
- 99.9% accuracy by consensus

Titanium Upgrade (Fall 2008)
- ~400 million bases/full plate run
- ~1 million sequence reads
- 350-500 bases per read

FLX+ Upgrade (Fall 2011)
- ~800 million bases/full plate run
- ~1 million reads
- 700 to 1,000 bases per read

Applications
- De novo whole Genome Sequencing (virus, bacteria, eukaryotes)
- Ultra Wide Sequencing (microRNA, SAGE, expression)
- Ultra Deep Sequencing (Oncology, population biology, HIV)
Ion Torrent Personal Genome Machine

Pairs semiconductor technology with sequencing chemistry

High-density array of micro-machined wells to perform reactions in a massively parallel way and beneath the wells is an ion-sensitive layer and beneath that a proprietary ion sensor.

When a nucleotide is incorporated into a strand of DNA by polymerase, a hydrogen ion is released as a byproduct.

Cross-Section of Flow Cell

Addition of a Base Gives Signal
Ion Torrent Personal Genome Machine

Applications:
- Small Genome Sequencing
- Targeted (Amplicon) Re-sequencing
- Target Capture
- Copy Number
- Chip-Seq
- RNA-Seq
- Bar Codes, Paired End Reads

314 Chip give ~1 million reads/run
316 & 318 Chips give ~2 million reads/run

<table>
<thead>
<tr>
<th>Ion Semiconductor Sequencing Chip</th>
<th>Output</th>
<th>Read Length</th>
<th>Total Sequencing Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>314</td>
<td>&gt; 10Mb</td>
<td>2011</td>
<td>2012</td>
</tr>
<tr>
<td>316</td>
<td>&gt; 100Mb</td>
<td>&gt; 200bp</td>
<td>&gt; 400bp</td>
</tr>
<tr>
<td>318</td>
<td>&gt; 16b</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Accuracy: >99.99% consensus accuracy and >99.5% raw accuracy.
Illumina HiSEQ 2000 (short-read system)

2 Flowcell System
• 8 lanes/flowcell
• Read Length: 35-100 bp
• No. of Reads/Lane: >150M
• Accuracy
  >85% bases higher than Q30 at 2 x 50 nt
  >80% bases higher than Q30 at 2 x 100 nt
• Total: Up to 3 Billion SE or 6 Billion PE reads
• Run Times: days to weeks
  1 x 35 nt = ~1.5 days
  2 x 100 nt = ~8 - 11 days
Illumina HiSEQ 2000

• Applications
  – Resequencing
  – ChipSeq
  – miRNA identification & quantification
  – Gene expression
  – Methylation profiling
  – Target capture

Sequence Clusters

Flow Cell

Cluster Generation
Illumina MiSeq v2

<table>
<thead>
<tr>
<th>Read Length</th>
<th>Total Time</th>
<th>MiSeq v2 Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 x 36</td>
<td>~4 hrs</td>
<td>540-610 Mb</td>
</tr>
<tr>
<td>2 x 25</td>
<td>~5.5 hrs</td>
<td>750-850 Mb</td>
</tr>
<tr>
<td>2 x 100</td>
<td>~16.5 hrs</td>
<td>3.0-3.4 Gb</td>
</tr>
<tr>
<td>2 x 150</td>
<td>~27 hrs</td>
<td>4.5-5.1 Gb</td>
</tr>
<tr>
<td>2 x 250</td>
<td>~39 hrs</td>
<td>7.5-8.5 Gb</td>
</tr>
</tbody>
</table>

Applications:
- Small targeted panels
- Metagenomics
- Small genome seq
- Test runs for HiSeq
- Paired-end runs
NGS Applications Supported by the Core

Genome Sequencing - *de novo* & re-sequencing
- Small Genomes - 454, Ion Torrent, and MiSeq
- Large Genomes – HiSeq

DNA Amplicons

Target Capture
- Agilent SureSelect Custom
- Agilent SureSelect Whole Exome Catpure

RNA-Seq
- Illumina TruSeq RNA-Seq v2
- Illumina TruSeq Stranded

ChIP-Seq
- Illumina TruSeq ChIP-Seq
- Nugen Ovation Ultralow Library System

smRNA Seq - Illumina TruSeq smRNA-seq

Methyl-Seq - Agilent SureSelect Methyl Seq
How Deep Do I Need To Go?

Discovery vs. Known Targets

5-10 million SE reads (35-50 bp) for Counting/Profiling applications such as ChipSeq and small RNA Profiling of known RefSeqs or small RNAs.

30-50 million SE reads (35-50 bp) for Discovery & Counting/Profiling applications such as ChipSeq and small RNA profiling.

30-50 million PE reads (50-100 bp) for whole transcriptome profiling (RNAseq) to quantify known RefSeqs.

>100 million PE reads (50-100 bp) for whole transcriptome profiling (RNAseq) to discover and quantify known RefSeqs and non-coding RNAs.
The Encode Project
(ENCyclopedia Of DNA Elements)

http://encodeproject.org/ENCODE/

International collaboration of research groups funded by the National Human Genome Research Institute (NHGRI)

The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

Guidelines for Experiments

Current Guidelines:


ChIP-seq, ChIP-chip, DNase-seq, FAIRE-seq and DNAme Standards v2.0 (July 2011)

RNA Standards v1.0 (May 2011)

RIP Standards v2.0 (Jan 2012)

http://genome.ucsc.edu/ENCODE/experiment_guidelines.html
RNA-Seq
(Whole Transcriptome Shotgun Sequencing)

Uses massively parallel DNA sequencing to sequence cDNA to acquire information about a sample’s RNA content.

Depending on depth of coverage, RNA-Seq can be used to measure:

• Differential expression of genes
  • Gene alleles
  • Differently spliced transcripts
• Non-coding RNAs
• Post-transcriptional mutations/editing
  • Gene fusions

RNA-Seq v Microarray

• Discovery
• Sensitivity
• Cost
• Ease of Analysis
ChIP-Seq

Used to analyze protein interactions with DNA

Combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins.
Sequence Target Capture

A process for the enrichment of selected genomic regions from the full complexity of an entire genome

Why Target?
- Can focus on only regions or genes of interest
- It still costs too much to sequence the entire genome
- Can process more samples per sequencing run

Potential Issues
- Poor or no capture of some targets
- Sequencing whole genome may soon be cost effective
- For Exome products, not all genes are targeted
Amplicon Sequencing

Highly targeted approach to sequence a region of interest
  • Usually smaller target size than capture approaches
  • Able to run more samples (up to 1,536 sample workflows)

Unlike target capture methods, you generally have good coverage of all target region.

How many samples can I analyze per run? Depends!
  • Coverage
  • Target size
  • Number of reads/lane (instrument)

Amplicon balance can be tricky to achieve
  • Amplification efficiencies differ between primer sets
  • Amplicon size differences can be problematic
  • Some targets have more secondary structure

Applications
  • Gene panel (e.g., Cancer gene panels)
  • Metagenomics-16s rRNA sequencing
  • Small target-many sample sequencing projects
Illumina Amplicon Sequencing

Step 1: PCR to amplify regions of interest
- Overhang adapter sequence used in Step 2
- Locus-specific sequence
- Genomic DNA

Step 2: 2nd round of PCR to add ILMN indices and sequencing adapters
- Index adapter oligos from ILMN: contain P5/P7 adapters to make template compatible with flow cell, also contains a unique sample index
- P5, Index, Insert to be sequenced, Index, P7

Example project of 5 amplicons per sample, 96 sample project

Step 3: Sequence on MiSeq
- Pool
- Index 2
- Read 1
- Read 2 (optional)
- Index 1
Sequencing Low-Complexity Amplicons

Illumina sequencers require sequence complexity

PhiX control DNA can be added to introduce complexity. Sacrifice read depth.

Use “stuffer” sequence in the primers used in the “First Step” to introduce complexity.

Append to 5’ end of forward PCR primer:
5’ TCGTCGGGCAGCGTCAGATGTGTATAAGAGACAG-[locus specific sequence]
5’ TCGTCGGGCAGCGTCAGATGTGTATAAGAGACAGN-[locus specific sequence]
5’ TCGTCGGGCAGCGTCAGATGTGTATAAGAGACAGNN-[locus specific sequence]
5’ TCGTCGGGCAGCGTCAGATGTGTATAAGAGACAGNNN-[locus specific sequence]

Append to 5’ end of reverse PCR primers:
5’ GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG-[locus specific sequence]
5’ GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGN-[locus specific sequence]
5’ GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNN-[locus specific sequence]
5’ GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNN-[locus specific sequence]
Caliper SciClone Liquid Handling Robot for NGS Applications

Applications/Methods
- SureSelect Exome Capture
- SureSelect Custom Capture
- RNA-Seq (Illumina TruSeq)
- DNA-Seq (Illumina TruSeq)

Coming Soon
- smRNA-Seq (Illumina TruSeq)

Features
- Integrates with LIMS
- Core has 2 robots
- Up to 96 libraries/week/robot
- Process 8-96 samples/run
Covaris Sonicator

Model S2 (single sample)  
Model E220 (96-samples)

Adaptive Focused Acoustics™
Isothermal Process
No Contact-Closed Cuvette
Very Reproducible

Genomics - Controlled DNA shearing for next generation sequencing and library constructions. Chromatin shearing for ChIP-Seq. RNA extraction from various biological fluids and tissues, especially difficult tissues, such as skin, bones, corneal, muscle, etc.

Proteomics - Protein & Biomarker Extraction, Protein Digestion

Cell Biology - Tissue disruption & Homogenization, Cell Lysis, Plasma Extraction
Covaris Sonicator Workflow?

S2 (single cuvette)

Optimize

Confirm

E220 (96 cuvette)

Mass Produce

Able to reproducibly generate the same fragment size following optimization.
Pippin Prep

Preparative electrophoresis platform used to separate and extract size-specific DNA fragments

Quicker than manual gel purification

Samples run on individual channels to eliminate sample cross-contamination

NGS Applications

- smRNA-Seq
- ChIP-Seq
- emPCR beads for Ion Torrent

Minimal low MW contamination Highly Reproducible & Discriminating
I was struck by the number of talks that described the use of whole-genome sequencing and analysis to reveal the genetic basis of disease in patients. … patients included a child with irritable bowel disease, a child with severe combined immunodeficiency, two siblings affected with Miller syndrome, and several with cancers of different types.

…each presenter emphasized the rapidity with which these data can now be generated using next-generation sequencing instruments, they also listed the large number of people involved in the analysis of these datasets. The required expertise to ‘solve’ each case included molecular and computational biologists, geneticists, pathologists and physicians with exquisite knowledge of the disease and of treatment modalities, research nurses, genetic counselors, and IT and systems support specialists, among others.

…although the idea of clinical whole-genome sequencing for diagnosis is exciting and potentially life-changing … one does wonder how … such a ‘dream team’ of specialists would be assembled for each case. In other words, even if the cost and speed of generating sequencing data continue their precipitous decreases, the cost of ‘team’ analysis seems unlikely to immediately follow suit.”
Personnel

Director: Kevin Knudtson, Ph.D., 323 EMRB, 5-7251

DNA and Genome Sequencing, 300H EMRB, 5-6736
  – Jennifer Bair
  – Angela Crock
  – Cory Hammitt
  – Einat Snir

DNA Microarray and Real-Time PCR, 321 EMRB, 5-7928
  – Mary Boes
  – Garry Hauser

Oligonucleotide Ordering: Jan Fuller, 329 EMRB, 5-7923