Next-Generation Sequencing Technologies

Elaine R. Mardis, Ph.D.
Co-director, The Genome Institute
Robert E. and Louise F. Dunn
Distinguished Professor of Medicine

No Relevant Financial Relationships with Commercial Interests
Next-generation Sequencer basics

How massively parallel sequencing works

Next-generation DNA sequencing instruments

- All NGS platforms require a library obtained either by amplification or ligation with custom linkers (adapters)
- Each library fragment is amplified on a solid surface (either bead or flat Si-derived surface) with covalently attached adapters that hybridize the library adapters
- Direct step-by-step detection of the nucleotide base incorporated by each amplified library fragment set
- Hundreds of thousands to hundreds of millions of reactions detected per instrument run = “massively parallel sequencing”
- A “digital” read type that enables direct quantitative comparisons
- Shorter read lengths than capillary sequencers
Library Construction and Amplification

- Shear high molecular weight DNA with sonication
- Polish ends
- Ligate synthetic DNA adapters (PCR*)
- Produce size fractions (PCR*)
- Quantitate
- Amplify library fragments on flow cell surface (PCR*)
- Denature clusters to single-stranded
- Hybridize sequencing primer to linearized ss cluster DNAs
- Proceed to sequencing or hybrid capture

PCR-related Problems in NGS

- PCR is an effective vehicle for amplifying DNA, however...
- In NGS library construction, PCR can introduce preferential amplification (“jackpotting”) of certain fragments
  - Duplicate reads with exact start/stop alignments
  - Need to “de-duplicate” after alignment and keep only one pair
  - Low input DNA amounts favor jackpotting due to lack of complexity in the fragment population
- PCR also introduces false positive artifacts due to substitution errors by the polymerase
  - If substitution occurs in early PCR cycles, error appears as a true variant
  - If substitution occurs in later cycles, error typically is drowned out by correctly copied fragments in the cluster
- Cluster formation is a type of PCR (“bridge amplification”)
  - Introduces bias in amplifying high and low G+C fragments
  - Reduced coverage at these loci is a result
Hybrid Capture

- **Hybrid capture** - fragments from a whole genome library are selected by combining with probes that correspond to most (not all) human exons or gene targets.
- The probe DNAs are biotinylated, making selection from solution with streptavidin magnetic beads an effective means of purification.
- An "exome" by definition, is the exons of all genes annotated in the reference genome.
- Custom capture reagents can be synthesized to target specific loci that may be of clinical interest.

Multiplex PCR Amplification of Targets

1. Design amplification primer pairs for exons of genes of interest; tile primers to overlap fragments in larger exons
2. Group primer pairs according to G+C content, Tm and reaction condition specifics
3. Amplify genomic DNA to generate multiple products from each primer set; pool products from each set
4. Create library by ligation or tail platform adaptors on the primer ends
5. Sequence
Massively Parallel Sequencing by Synthesis

Incorporate
Detect
De-block
Cleave fluor

Platforms: Illumina

<table>
<thead>
<tr>
<th>Key applications</th>
<th>MiSeq</th>
<th>NextSeq 500</th>
<th>HiSeq 2500</th>
<th>HiSeq X*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small genome, amplicon, and targeted gene panel sequencing</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Exome, transcriptome sequencing, and more</td>
<td>1 or 2</td>
<td>1 or 2</td>
<td>1 or 2</td>
<td>N/A</td>
</tr>
<tr>
<td>Run mode</td>
<td>N/A</td>
<td>Mid-Output</td>
<td>High-Output</td>
<td>High-Output</td>
</tr>
<tr>
<td>Flow cells processed per run</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1 or 2</td>
</tr>
<tr>
<td>Run time</td>
<td>5-45 hours</td>
<td>15-26 hours</td>
<td>12-30 hours</td>
<td>&lt; 1 day - 6 days</td>
</tr>
<tr>
<td>Output range</td>
<td>0.5-15 Gb</td>
<td>20-59 Gb</td>
<td>30-120 Gb</td>
<td>10-180 Gb</td>
</tr>
<tr>
<td>Reads per flow cell</td>
<td>25 Million^2</td>
<td>130 Million</td>
<td>400 Million</td>
<td>300 Million</td>
</tr>
<tr>
<td>Maximum read length</td>
<td>2 x 300 bp</td>
<td>2 x 150 bp</td>
<td>2 x 150 bp</td>
<td>2 x 150 bp</td>
</tr>
</tbody>
</table>

- High accuracy, range of capacity and throughput
- Longer read lengths on some platforms (MiSeq)
- Improved kits, improved software pipeline and capabilities, cloud compute
ION Torrent-pH Sensing of Base Incorporation

- dNTP
- dNTPs
- Sensing Layer
- Sensor Plate
- Bulk
- Drain
- Source
- Silicon Substrate
- Template

Platforms: Ion Torrent

- PGM
  - Three sequencing chips available:
    - 314 = up to 100 Mb
    - 316 = up to 1 Gb
    - 318 = up to 2 Gb
  - 2-7 hour/run
  - up to 400 bp read length
  - 400kreads up to 5 Mreads

- Proton
  - Two human exomes (Proton 1 chip) or one genome (@20X-Proton 2 chip) per run
  - Ion One Touch or Ion Chef preparatory modules
  - 2-4 hour/run
  - ~200 bp average read length
  - Proton 1 produces 60-80 Mreads ≥50 bp

- Low substitution error rate, in/dels problematic, no paired end reads
- Inexpensive and fast turn-around for data production
- Improved computational workflows for analysis
Post Data Generation Analyses
Bioinformatic and computational approaches to NGS

The Goal?

Sequence data alignment is the crucial first step!
Short Read Alignment...

Is like a jigsaw puzzle...

...where they give you the cover on the box

Some pieces are easier to place than others...

pieces that look like each other...

...pieces with unique features
Repetitive Sequences Result in Multiple Read Alignments

Reads are Aligned, Now What?

• Data calibration and cleanup:
  • Mark proper pairs (if applicable)
  • Mark duplicate reads!
  • Correct local misalignments
  • Recalculate quality scores

• Call SNPs

• Evaluate Coverage
  • Compare SNPs from NGS to SNPs from array data
  • Integrated Genome Viewer
  • RefCov and others

• Analyze the data
Integrated Genomics Viewer (IGV)

Whole chromosome view

Zoom in to see more detail

Bases that do not match the reference sequence are highlighted by color

http://www.broadinstitute.org/igv

IGV: Somatic Single Nucleotide Variant QC

Low-quality base calls are faint, semi-transparent.
**RefCov: Coverage Depth and Breadth from Hybrid Capture**

http://gmt.genome.wustl.edu/genome-shipit/gmt-refcov/0.3/index.html

---

**Somatic Variant Discovery Pipeline**

**INPUTS**
- Sequence Reads (Tumor)
- Sequence Reads (Normal)

**ALGORITHMS**
- Read alignment
- Somatic SNV/indel discovery
- Somatic structural variant discovery
- Copy Number Alterations
- Visual QC/Local assembly
- Score/QC by filters

**RESULTS**
- Annotate Variants
### False Negativity/Positivity

- Most false negatives are due to lack of coverage
- False positives are due to multiple reasons, including:
  - Variant is only called on one strand
  - Variant is only called at the end of the read
  - Coverage of the matched normal at that locus is poor
  - Gene has a pseudogene/paralog and the reads are mis-mapped
  - High sensitivity variant calling algorithms have elevated false positive rates to achieve detection of subclonal variants and low false negative rates
- Data that verifies or refutes variant calls can help to define bioinformatic filters to remove them

### Third Generation Sequencers

Variations on a theme
### Real Time Sequencing of Single DNA Molecules

DNA:polymerase complex is immobilized at the ZMW bottom.

Fluorescent nucleotides are introduced to the ZMW.

A light pulse is produced as each fluorescent base is held in the polymerase active site. The phosphate is cleaved during incorporation, releasing the fluorophore.

The process occurs in parallel in all the loaded ZMWs.

### Pacific Biosciences RS

**Sample Prep**
- Shearing (Covaris/Hydroshear)
- Polish ends
- SMRTbell™ ligation
- Sequencing primer annealing

**Library/Polymerase Complex**
- DNA polymerase binding
- Load library/polymerase complex onto SMRT cell

**Sequencing**
- Movie
- Raw reads
- Post-filter reads
- Mapped reads
- SMRTbell™ template
- Polymerase Read
- Subreads
- Read of insert
PacBio: 20 kb Library Preparation and Sequencing

- Covaris g-Tube 20 kb shear
- Pacific Biosciences 20 kb library prep
- Sage Science BluePippin size fractionation
  - 8 - 50 kb
  - 15 - 50 kb
- Pacific Biosciences RSII sequencing
  - Polymerase: P5
  - Sequencing chemistry: C3
  - MagBead loading
  - Per SMRT Cell
    - 180 minute collection time
    - “Stage start”

PacBio: Improvements in Polymerase and Chemistry

![Graphs showing PacBio P4-C2 and P5-C3 Chemistry](image)

- Average: ~5.5 kb
- Maximum: ~24 kb
- Top 5% of reads: ~11 kb
- Half of data in reads: ~8 kb
- Data per SMRT Cell: ~275 Mb

- Average: ~8.5 kb
- Maximum: ~39 kb
- Top 5% of reads: ~18 kb
- Half of data in reads: ~10 kb
- Data per SMRT Cell: ~375 Mb

Based on data from a 20 kb size-selected E. coli library using a 180-minute movie. Each SMRT Cell yields ~50,000 reads.
Chicken 20 kb – BluePippin 8-50 kb

Chicken 20 kb – BluePippin 15-50 kb
### Human BAC/fosmid clones sequenced by PacBio platform

<table>
<thead>
<tr>
<th>clone name</th>
<th>Clone size (bp)</th>
<th>library size</th>
<th>SMRT cell</th>
<th>Number of mapped Subreads</th>
<th>Error Corrected Coverage Post-Vector/E Cell Screened</th>
<th>Number of contigs after de novo assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC11-4734100BCC4</td>
<td>39755</td>
<td>10 kb</td>
<td>1</td>
<td>50884</td>
<td>121.1X</td>
<td>1</td>
</tr>
<tr>
<td>ABC11-47399300K2</td>
<td>38934</td>
<td>10 kb</td>
<td>1</td>
<td>56899</td>
<td>311X</td>
<td>3</td>
</tr>
<tr>
<td>ABC11-4939500A2</td>
<td>41423</td>
<td>10 kb</td>
<td>1</td>
<td>63180</td>
<td>162X</td>
<td>5</td>
</tr>
<tr>
<td>ABC11-46674300M3</td>
<td>39380</td>
<td>10 kb</td>
<td>1</td>
<td>57265</td>
<td>157X</td>
<td>2</td>
</tr>
<tr>
<td>ABC12-47036800M8</td>
<td>40000**</td>
<td>10 kb</td>
<td>1</td>
<td>59535</td>
<td>317.2X</td>
<td>1</td>
</tr>
<tr>
<td>ABC14-50418300F1</td>
<td>40000***</td>
<td>10 kb</td>
<td>1</td>
<td>66469</td>
<td>140X</td>
<td>1</td>
</tr>
<tr>
<td>ABC7C-42060100K1</td>
<td>36886</td>
<td>10 kb</td>
<td>1</td>
<td>42220</td>
<td>109.3X</td>
<td>1</td>
</tr>
<tr>
<td>ABC9-41267000F2</td>
<td>40000***</td>
<td>10 kb</td>
<td>1</td>
<td>53298</td>
<td>337X</td>
<td>7</td>
</tr>
<tr>
<td>ABC9-43178800M9</td>
<td>40000***</td>
<td>10 kb</td>
<td>1</td>
<td>33745</td>
<td>151.4X</td>
<td>1</td>
</tr>
<tr>
<td>ABC9-40109000K1</td>
<td>42398</td>
<td>10 kb</td>
<td>1</td>
<td>47414</td>
<td>117.3X</td>
<td>1</td>
</tr>
<tr>
<td>CH17-176934</td>
<td>20744</td>
<td>10 kb</td>
<td>1</td>
<td>78003</td>
<td>41X</td>
<td>1</td>
</tr>
<tr>
<td>CH17-179417</td>
<td>170000***</td>
<td>10 kb</td>
<td>1</td>
<td>24274</td>
<td>78.6X</td>
<td>1***</td>
</tr>
<tr>
<td>CH17-179112</td>
<td>176000***</td>
<td>10 kb</td>
<td>1</td>
<td>55588</td>
<td>60X</td>
<td>1</td>
</tr>
<tr>
<td>CH17-275114</td>
<td>226651</td>
<td>10 kb</td>
<td>1</td>
<td>84211</td>
<td>77X</td>
<td>2</td>
</tr>
<tr>
<td>CH17-355822</td>
<td>320000***</td>
<td>10 kb</td>
<td>1</td>
<td>35245</td>
<td>109.9X</td>
<td>1</td>
</tr>
<tr>
<td>CH17-390012</td>
<td>177000***</td>
<td>10 kb</td>
<td>1</td>
<td>32540</td>
<td>41X</td>
<td>2</td>
</tr>
<tr>
<td>CH17-445113</td>
<td>150000***</td>
<td>10 kb</td>
<td>1</td>
<td>56444</td>
<td>32X</td>
<td>8</td>
</tr>
<tr>
<td>CH17P-30K13</td>
<td>228074</td>
<td>10 kb</td>
<td>1</td>
<td>54090</td>
<td>35X</td>
<td>3</td>
</tr>
<tr>
<td>RP11-84A4</td>
<td>189483</td>
<td>10 kb</td>
<td>1</td>
<td>45524</td>
<td>44X</td>
<td>4</td>
</tr>
<tr>
<td>WI2-202540C2</td>
<td>37272</td>
<td>10 kb</td>
<td>1</td>
<td>76585</td>
<td>41X</td>
<td>1</td>
</tr>
<tr>
<td>WI2-3067P5</td>
<td>39143</td>
<td>10 kb</td>
<td>1</td>
<td>27716</td>
<td>88X</td>
<td>1</td>
</tr>
</tbody>
</table>

*** Estimated clone size based on restriction enzyme digests and/or type of clone (fosmid/BAC)
** This assembly contains 1 human contig plus contaminated bacterial contigs
* This assembly produces 1 short contig due to collapsed repeat within the contig.

### Comparative assemblies with Illumina or PacBio

<table>
<thead>
<tr>
<th>clone name</th>
<th>Illumina assembly coverage</th>
<th>PacBio ProAssembled read Coverage</th>
<th>Illumina total contig #</th>
<th>PacBio total contig #</th>
<th>Illumina total contig bases (bp)</th>
<th>PacBio total contig bases (bp)</th>
<th>Illumina N50 contig bases</th>
<th>PacBio N50 contig bases (bp)</th>
<th>% GC</th>
</tr>
</thead>
</table>
| H_GD-281P19 | 64X | 83.0X | 93 | 1 | 398861 | 217805 | 13790 | 217805 | 40%
| H_GD-280P20 | 73X | 119.7X | 20 | 1 | 398255 | 197966 | 17306 | 197966 | 34%
| H_GD-381B03 | 70X | 108.3X | 66 | 1 | 172074 | 196563 | 12435 | 196563 | 43%
| H_GD-432G21 | 70X | 101.8X | 90 | 1 | 220679 | 222522 | 5371 | 222522 | 35%
| H_GD-106M31 | 65X | 82.3X | 53 | 7 | 331252 | 197654 | 11085 | 26921 | 39%
| H_GD-219D13 | 74X | 119.5X | 95 | 2 | 107454 | 147058 | 6761 | 122727 | 42%
| H_GD-380L19 | 73X | 97.0X | 20 | 8 | 137328 | 239670 | 13262 | 47406 | 42%
| H_GD-266C19 | 76X | 106.3X | 19 | 1 | 394736 | 194593 | 17995 | 194593 | 36% |
Pac Bio: Long reads improve the Human Reference Genome sequence

Since the HRG finished sequence was announced and published in 2004, our group has continued to improve the reference:

- Addition of new content, including novel content from other human genomes
- Improvement of previously poorly finished regions
- Finishing of regions between segmental duplications

Our new approach to HRG improvement will include sequencing haploid human genomes (hydatidiform mole) with Pacific Biosciences long read sequencing:

- One such genome (CHM1) already has 60X coverage from PacBio
- An assembly of CHM1 is now being compared to the HRG (grCH38)
Nanopore Sequencing

- Exonuclease-aided sequencing
- Pore translocation sequencing

Oxford Nanopore Sequencing

- Exonuclease-aided sequencing
- Variable read lengths
- Electrical current-based detection of triplet nucleotides in pore
Translating the Cancer Genome

Therapeutic Options via NGS and analysis

Cancer is a Disease of the Genome

In the early 1970’s, Janet Rowley’s microscopy studies of leukemia cell chromosomes suggested that specific alterations led to cancer, laying the foundation for cancer genomics.
TGI: Cancer Cases by WGS (March 2014)

- **WGS:** 2,700 from ~1,100 cancer patients
- **Exomes:** 6,200 from ~3,000 cancer patients
- **WU-SJ PCGP:** 750 patients

Pan-Cancer Analyses from TCGA

12 tumor types

- Leukemia (LAML)
- Lung adenocarcinoma (LUAD)
- Lung squamous (LUSC)
- Kidney (KIRC)
- Bladder (BLCA)
- Endometrial (UCEC)
- Glioblastoma (GBM)
- Head and neck (HNSC)
- Breast (BRCA)
- Ovarian (OV)
- Colon (COAD)
- Rectum (READ)

**Oncogenic characteristics**
- Mutation
- Copy number
- Gene expression
- DNA methylation
- MicroRNA
- RIPPA

**Platforms**
- Samples
- Genes

**Thematic pathways**

Nature Genetics 45: 1113-1120 (2013)
Integrated WGS/Exome/RNA-Seq

- **WGS** analysis yields:
  - SNVs (single nucleotide variants)
  - CNVs (amplification/deletion)
  - SVs (translocations, inversions)
  - Indels (focused insertions/deletions)
- **Exome**: validates WGS discoveries, integrated coverage depth allows clonality analysis
- **RNA-Seq**: over-expression metrics, expressed SNVs, gene fusions
- **Clinical Action**: identifying druggable targets

## Linking Somatic Variants to Therapies

Obi Griffith, Ph.D. and Malachi Griffith, Ph.D.
Therapeutic Interpretation of Variants

Somatic/Germline Cancer Events (DNA+RNA)

- Single Nucleotide Variants
- Insertion/deletions
- Structural Variants
- Copy Number Variations
- Expressed variants
- SV-predicted gene fusions
- Differentially Expressed Genes
- Differentially Expressed Isoforms

Clinical prioritization and reporting

- Functional annotation
- Filtered (activating/drivers)
- Candidate genes/pathways
- Clinically actionable events (aka "The Report")

Drug Gene Interaction database (>50 database sources)

- Literature
- dGene
- DrugBank
- TTD
- clinicaltrials.gov
- PharmGKB
- TEND
- TALC
- MyCancerGenome

DoCM: A Database of Canonical Cancer Mutations

- Highly curated database of mutations having a demonstrated association with cancer
- General information about each somatic variant
  - Chromosomal Location
  - Strand
  - Gene
  - Protein impact of variant (annotation)
  - PubMed ID evidence cited, linked
- Easy to access from the web and programmatically through an API
DGIdb: Drug Gene Interaction database

dgidb.org


Use of Whole-Genome Sequencing to Diagnose a Cryptic Fusion Oncogene

Welch et al., JAMA 2011: 305(15): 1577-1584.
Lukas Wartman, M.D. is Patient “ALL1”

**The New York Times**

*In Treatment for Leukemia, Glimpses of the Future*

FLT3 Over-expression in ALL1

- FLT3 was within the top 1% of all expressed genes.
- Absent a normal comparator, the literature report from Marston identified FLT3 over-expression in pre-B-ALL.
- Based on wt FLT3 over-expression by the tumor cells, we predicted the cancer would be sensitive to the FLT3 inhibitor Sunitinib (Sutent) [DrugBank].

**Patient biopsied metastatic melanoma lesions**

Mardis, Schreiber et al., *Nature* 2012

**Tumor and germline DNA sequenced, somatic mutations identified; RNA capture verifies expressed mutations and expression level; netMHC algorithm identifies immunoepitopes**

**Apheresis samples from patient used to verify the algorithmically-identified immunoepitopes that elicit T cell memory**

---

**Sequencing to identify tumor-specific immunoepitopes:**

*Genome-driven cancer immunotherapy*

---

**Dendritic Cell Vaccine Platform**

A dendritic cell-based approach is currently being tested in an FDA approved protocol for metastatic melanoma patients:

- Patient 1 has received all three doses of vaccine, and is being monitored
- Patient 2 has received three doses of vaccine, this patient has measurable disease and will be monitored for progression, stability or regression
- Patient 3 has measurable disease, has completed her vaccine infusions early March
- Patients 4 and 5 have genomic analysis completed, in vitro assays completed, GMP peptides underway
## Acknowledgements

**The Genome Institute**
- Malachi Griffith, Ph.D.
- Obi Griffith, Ph.D.
- Ben Ainscough
- Zach Skidmore
- Avinash Ramu
- Allison Regier
- Lee Trani
- Nick Spies
- Vincent Magrini, Ph.D.
- Sean McGrath
- Ryan Demeter
- Jasreet Hundal, M.S.
- Jason Walker
- David Larson, Ph.D.
- Lucinda Fulton
- Robert Fulton
- Richard K. Wilson, Ph.D.

**WUSM/Siteman Cancer Center**
- Timothy J. Ley, M.D.
- Lukas Wartman, M.D.
- Peter Westervelt, M.D.
- John DiPersio, M.D.
- Gerry Linette, M.D.
- Beatriz Carreno, M.D., Ph.D.

**Thanks also to:**
- Aaron Quinlan
- Gabor Marth
- Michael Zody

**Our patients and their families**