Next-Generation Sequencing Technologies

Elaine R. Mardis, Ph.D.
Co-director, McDonnell Genome Institute
Robert E. and Louise F. Dunn Distinguished Professor of Medicine
Current Topics in Genome Analysis 2016

Elaine Mardis

Qiagen NV

Paid Member of Board
Massively Parallel Sequencing basics

How massively parallel sequencing works
NGS has transformed biomedical inquiry

Massively Parallel DNA sequencing instruments

- All MPS 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 for MPS

- Shear high molecular weight DNA with sonication
- Enzymatic treatments to blunt ends
- Ligate synthetic DNA adapters (each with a DNA barcode), PCR amplify
- Quantitate library
- Proceed to WGS, or perform exome or specific gene hybrid capture
PCR-related Problems in MPS

- PCR is an effective vehicle for amplifying DNA, however...
- In MPS 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

1. OH OH → → → →  
   P7 PS
2. 1st cycle denaturation → 1st cycle annealing → 1st cycle extension → 2nd cycle denaturation
3. 2nd cycle denaturation → 2nd cycle annealing → 2nd cycle extension → 3'st total
4. Cluster Amplification → PS Linearization → Block with ddNTPS → Denaturation and Sequencing Primer Hybridization

Emission
Incorporate
Detect
De-block
Cleave fluor

Excitation

OH free 3’ end

3’ 5’
Illumina Patterned Flow Cell

Patterned flow cell advantages
- Higher data output
- More sequencing reads
- Faster run times
Platforms: Illumina

- 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
Platforms: Ion Torrent

- 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

- 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 Human Genome Reference enables MPS Genomics

- The human genome reference sequence is the keystone interpreting MPS sequencing read data
- **Alignment** of reads to the human reference sequence is the first step to identify variation of all types
- Mis-aligning sequences identify structural alterations
- Alignment and analysis of RNA sequence data provides information about gene expression changes
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

Lander et al. 2001
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 OR
  - Compare SNPs from tumor to normal (Which are shared? Number?)
  - 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.

Low-quality base calls are faint, semi-transparent.

http://www.broadinstitute.org/igv
IGV: Somatic Single Nucleotide Variant QC

Normal
Normal bam coverage

Tumor
Tumor bam coverage

Key:
- C
- A
- G
- T
- Y
- S
- R
- B
- V
- D
- H
- K
- M
- N

NHGRI Current Topics in Genome Analysis 2016
Week 14: Next-Generation Sequencing Technologies

May 25, 2016
Elaine Mardis, Ph.D
RefCov: Coverage Depth and Breadth from Hybrid Capture

http://gmt.genome.wustl.edu/genome-shipit/gmt-refcov/0.3/index.html
## Data and coverage characteristics

### Sequencing Data

<table>
<thead>
<tr>
<th></th>
<th>Exome</th>
<th>Whole Genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data generated</td>
<td>6 Gbp</td>
<td>118 Gbp</td>
</tr>
<tr>
<td>Target space</td>
<td>37 Mbp</td>
<td>3.2 Gbp</td>
</tr>
<tr>
<td>Typical map rate</td>
<td>98.5%</td>
<td>95.0%</td>
</tr>
<tr>
<td>Typical dup rate</td>
<td>8.8%</td>
<td>3.4%</td>
</tr>
<tr>
<td>Avg. CDS sequence depth</td>
<td>65x</td>
<td>30x</td>
</tr>
<tr>
<td>% of CDS covered &gt;10x</td>
<td>90-95%</td>
<td>95-99%</td>
</tr>
</tbody>
</table>

### Variant Detection

<table>
<thead>
<tr>
<th></th>
<th>Exome</th>
<th>Whole Genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNV calling</td>
<td>Good</td>
<td>Good</td>
</tr>
<tr>
<td>Small indel calling</td>
<td>Good</td>
<td>Good</td>
</tr>
<tr>
<td>CNV calling</td>
<td>Poor</td>
<td>Good</td>
</tr>
<tr>
<td>SV calling</td>
<td>Poor</td>
<td>OK</td>
</tr>
<tr>
<td>Typical SNVs called</td>
<td>~50,000</td>
<td>~3 million</td>
</tr>
</tbody>
</table>
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
Integrated Systems: Qiagen GeneReader

- Modules for Sample Prep, Library Construction/Amplification, SBS sequencing
- Onboard analysis and interpretation software for mutation detection and interpretation based on Ingenuity Variant and Pathway Analysis
Third Generation Sequencers

Variations on a theme
Real Time Sequencing of Single DNA Molecules

DNA:polymerase complex is immobilized at the ZMW bottom. The process occurs in parallel in all the loaded ZMWs. Fluorescent nucleotides are introduced to the ZMW. A light pulse is produced as each fluorescent base is held in the polymerase active site. Phosphate is cleaved during incorporation, releasing the fluorophore.
Pacific Biosciences RS

**Sample Prep**
- DNA Shear
  - AMPure PB
- DNA Damage Repair
- Ligation/Exonuclease
  - 3x AMPure PB
- BluePippin Sizing
  - Rinse wells
  - AMPure PB
- DNA Damage Repair
  - AMPure PB

**Sequencing Prep**
- Seq. Primer
  - Denature primer prior to use
- P6 Polymerase Bind
  - 30 minutes or 4 hours
- MagBead Bind
  - 20 minutes to 2 hours

**Sequencing**
- Raw reads
- Post-filter reads
- Mapped reads

**Movie**
- SMRTbell™ Template
- Polymerase Read
- Subreads
- Read of Insert

**4 to 6 hour collection time**
Considerations for PacBio WGS

- High molecular weight genomic DNA
  - DNA must be of sufficient quality to allow for >30 kb shearing to produce PacBio Continuous Long Reads (CLR)

- Consistent shearing >30 kb
  - Shearing genomic DNA >30 kb is challenging and requires a consistent technology
  - Preferred method: Diagenode Megaruptor
  - Alternate method: Covaris g-Tube

- Sufficient DNA for PacBio sample prep
  - A single PacBio sample prep reaction requires 5 µg sheared DNA
  - One library is composed of 8-10 sample prep reactions
  - At least 2-4 libraries are required for 60x coverage
Read Length Comparisons

Library #1 18-50kb
- Mean Subread Length 13,720 bp
- N50 Length 19,411 bp

Library #3 30-80 kb
- Mean Subread Length 15,008 bp
- N50 Length 24,136 bp
The Human Reference is a Work in Progress!

• The current reference - GRCh38 - is not optimal for some regions of the genome and/or some individuals/ancestries.

• GRCh38 is comprised of DNA from several individual humans.

• Allelic diversity and structural variation present major challenges when assembling a representative diploid genome.

• New technologies, methods, and resources since 2003 have allowed for substantial improvements in the reference genome.

• Additional high-quality reference sequences are needed to represent the full range of genetic diversity in humans.
Improving the Human Reference Genome(s)

The MGI Reference Genomes Improvement Project

Funded by the NIH, the MGI Reference Genomes Improvement Project aims to increase the quality and diversity of existing scientific resources. We will sequence and assemble at least 5 diploid genomes from individuals selected to maximize human genetic diversity (right). All sources have BAC libraries available and whenever possible, we will use samples from a trio (two parents and child). We will sequence the parents within the trio at a lower depth of coverage to enable haplotype phasing of the proband sequence. Other independent efforts to sequence and assemble new reference genomes include two Japanese, one Malaysian, a Han Chinese and an Ashkenazim trio (as part of the Genome in a Bottle Effort).

http://genome.wustl.edu/projects/detail/reference-genomes-improvement/
Sequencing Plan

- PacBio Reads
- de Novo Assembly
- Scaffolded using 10X, BioNano Dovetail and aligned to reference
- PacBio Sequenced BACs to fill targeted regions
- Improved Reference “Gold Genome”
First Gold Genome - NA19240

- NA19240 - Yoruban sample
- Generated >70X raw PacBio data

<table>
<thead>
<tr>
<th></th>
<th>Initial Assembly Stats</th>
</tr>
</thead>
<tbody>
<tr>
<td># Seq Contigs</td>
<td>3569</td>
</tr>
<tr>
<td>Max Contig Length</td>
<td>20,393,869 bp</td>
</tr>
<tr>
<td>Total Assembly Size</td>
<td>2,745,634,789 bp</td>
</tr>
<tr>
<td>N50</td>
<td>6,003,115 bp</td>
</tr>
<tr>
<td>N90</td>
<td>848,151 bp</td>
</tr>
<tr>
<td>N95</td>
<td>345,457 bp</td>
</tr>
</tbody>
</table>
Alignment of NA19240 to BioNano map

Conflict identified by BioNano data
Finished BACs Resolve This Region

PB Assembly

BAC Alignments

Seg Dup

Duplicates, EcoNIB Lab, total 78 features shown
## Platinum/Gold Genome Status

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Origin</th>
<th>Level of Coverage</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHM1</td>
<td>NA</td>
<td>Platinum</td>
<td>Assembly Improvement</td>
</tr>
<tr>
<td>CHM13</td>
<td>NA</td>
<td>Platinum</td>
<td>Data Generation</td>
</tr>
<tr>
<td>NA19240</td>
<td>Yoruban</td>
<td>Gold</td>
<td>Analysis Underway</td>
</tr>
<tr>
<td>HG00733</td>
<td>Puerto Rican</td>
<td>Gold</td>
<td>Assembly Assessment</td>
</tr>
<tr>
<td>HG00514</td>
<td>Han Chinese</td>
<td>Gold</td>
<td>Assembly QC</td>
</tr>
<tr>
<td>NA12878</td>
<td>European</td>
<td>Gold</td>
<td>Assembly QC</td>
</tr>
<tr>
<td>HG01352</td>
<td>Columbian</td>
<td>Gold</td>
<td>Assembly Underway</td>
</tr>
<tr>
<td>HG02818</td>
<td>Gambian</td>
<td>Gold</td>
<td>Not Started</td>
</tr>
<tr>
<td>HG02059</td>
<td>Kinh Vietnamese</td>
<td>Gold</td>
<td>Not Started</td>
</tr>
</tbody>
</table>
10X Genomics: a novel twist

1. Molecular Barcoding in GEMs

2. Pool, Library Prep

End-repair, A-tail, Ligate, SI-PCR

Final Library Construct

3. Sequence and Analyze

P5 Read 1 gDNA insert Sample Index P7

Chromium Barcode Read 2
10X Genomics: Linked Reads

- Long-range information from short reads
  - Partition long input molecules into GEMs (*Gelbead-in-Emulsion*)
  - Gelbeads carry barcode oligos that are incorporated in sequencing library
  - Use barcodes to link short reads back to original long input molecules

- Resulting barcoded reads are called *Linked-Reads*
10X Genomics: Power of Linked Reads

Multi-Megabase Diploid De Novo Assembly

Using the Supernova™ Assembler, reconstruct multi-megabase diploid assemblies, preserving phasing information for small variants, structural rearrangements and novel sequence without the need for a reference.
Oxford Nanopore Sequencing

Protein nanopore set in an electrically resistant polymer membrane
Each nanopore is controlled and measured by a matched ASIC (Application Specific Integrated Circuit)
Post-run data analysis compares pore current changes to a model of all possible multimers

- Variable read lengths
- Electrical current-based detection of multiplex nucleotides in pore
- Error rate is around 10-20% with newest pore/software
Nanopore Sequencing Devices
Genome-guided Immunotherapy Decision-making
Identifying “non-self” Neoantigen Landscapes
Early work from the labs of Thierry Boon and Hans Schreiber (among others) suggested that tumor specific mutations can sometimes function as tumor specific antigens.

James Allison and Bert Vogelstein predicted that many/most tumors should express mutational antigens based on their genomic repertoire of coding variants, and these might be the ideal tumor-specific targets for cancer immunotherapy.

In the past, identifying the tumor mutation landscape and its most immunogenic peptides has been hampered by technical obstacles, most of which have now been overcome by NGS and bioinformatic approaches to epitope prediction.
Genome-guided Immunotherapy

- In using genomic data to predict neoantigen load for a specific tumor, we utilize:
  - massively parallel sequencing and analysis to compare cancer and normal exomes and identify cancer-unique peptides
  - the HLA haplotypes of the individual patient
  - RNA sequencing data from the cancer cells to identify genes that are mutated and expressed
- These input data are considered by algorithms that model the binding of peptides to the MHC and calculate binding energies, producing a list of tumor specific mutated antigens (TSMAs) or neoantigens
- This information can describe the cancer’s neoantigen load
pVac-Seq Pipeline: Open Source Neoantigen Prediction

**PREPARE INPUTS**

- **Whole Genome Sequencing**
- **Exome Sequencing**
- **Custom Capture**

**PERFORM EPITOPE PREDICTION**

**INTEGRATE SEQUENCE INFO**

**FILTER CANDIDATES**

**Hundal et al., Genome Medicine 2016**
Mutation/Neoantigen Load and Checkpoint Blockade Response Potential

Tumors with a high mutational (neoantigen) load tend to respond to checkpoint blockade immunotherapy.
Germline DNA Repair Defects and Checkpoint Blockade Response Potential

B Radiographic Response

- Mismatch repair–proficient colorectal cancer
- Mismatch repair–deficient colorectal cancer
- Mismatch repair–deficient noncolorectal cancer

Change from Baseline in the Sum of Longest Diameters (%)

- 20% increase (progressive disease)
- 30% decrease (partial response)

Dung et al., NEJM 2015
GBM27: Rapid Progression in CNS

- Male patient, early 30’s, prior history of colon polyps
- GBM removed by craniotomy 8 months ago
- Post surgery temozolomide (TMZ) therapy
- Spinal metastasis resected 3 months ago, FMI test indicated high mutation load/pol E mutated germline status: treatment with Pebrolizumab
- Second spinal metastasis identified upon complications, removed 2 months ago
- All tumors studied by high coverage exome sequencing compared to PBMC normal, and by IHC
GBM27: Clonal Evolution

[Diagram showing clonal evolution across different samples and clusters]
GBM27: Evolving Immune Response

- Patient remains NED on Pembrolizumab (Keytruda)
- Recent MRI indicated response of remaining untreated right frontal horn lesion in the brain
## Acknowledgements

<table>
<thead>
<tr>
<th>McDonnell Genome Institute</th>
<th>WUSM/Siteman Cancer Center</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malachi Griffith, PhD</td>
<td>Gavin Dunn, MD, PhD</td>
</tr>
<tr>
<td>Obi Griffith, PhD</td>
<td>Joshua Rubin, MD, PhD</td>
</tr>
<tr>
<td>Vincent Magrini, PhD</td>
<td>Robert Schreiber, PhD</td>
</tr>
<tr>
<td>Sean McGrath</td>
<td></td>
</tr>
<tr>
<td>Ryan Demeter</td>
<td></td>
</tr>
<tr>
<td>Tina Graves-Lindsay</td>
<td></td>
</tr>
<tr>
<td>Bob Fulton</td>
<td></td>
</tr>
<tr>
<td>Chris Markovic</td>
<td></td>
</tr>
<tr>
<td>Richard K. Wilson, PhD</td>
<td></td>
</tr>
</tbody>
</table>

Thanks also to:

- Aaron Quinlan (Utah)
- Gabor Marth (Utah)
- Michael Zody (NYGC)
- Evan Eichler (U Wash)
- Pui Yan Kwok (UCSF)
- Valerie Schneider (NCBI)
- Jason Chin (Pac Bio)
- Adam Phillippy (NHGRI)
- Sergey Koren (NHGRI)

Our patients and their families