Exome Sequencing: A Comparison of Enrichment Technologies

Michael James Clark, Ph.D

August 3rd, 2011
Michael James Clark, Ph.D
Geneticist & Bioinformaticist
Stanford University
Mike Snyder Lab
Likes: *Sequencing Exomes*
Contact: mjclark@stanford.edu
Twitter: @mjcgenetics
Why exome sequencing?
Why exome sequencing?

http://map.seqanswers.com
Exome is highly interpretable

- Exome Variation
  - Missense variations—determinable amino acid changes
  - Nonsense variations—estimable effect on expression
  - Copy number variations—estimable gene dosage effects

- Non-exome variation
  - Enhancer/promoter variations—mostly indeterminate gene expression effect
  - ?
Why exome over whole genome?

Whole genome sequencing requires 50-100x more sequencing, and is therefore more expensive.

http://www.genome.gov/sequencingcosts/
Why exome over whole genome?

Exome-sequencing is <1/10\textsuperscript{th} the cost of WGS.

The cost of whole genome sequencing remains relatively high.
Method: exome enrichment

- Biotinylated oligonucleotide baits complementary to exome sequences

Example images borrowed from Illumina TruSeq exome enrichment workflow
Exome enrichment platform comparisons

- TruSeq Exome Enrichment Kit
- SureSelect All Exon 50M
- SeqCap EZ Exome v2.0
### Targeting strategy design differences

<table>
<thead>
<tr>
<th>Type</th>
<th>Lengths</th>
<th>Quantity</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nimblegen (DNA)</td>
<td>55-105bp</td>
<td>&gt;2,100,000 baits</td>
<td>44,007,233bp</td>
</tr>
<tr>
<td>Agilent (RNA)</td>
<td>114-126bp</td>
<td>655,872 baits</td>
<td>51,542,882bp</td>
</tr>
<tr>
<td>Illumina (DNA)</td>
<td>95bp</td>
<td>340,427 baits</td>
<td>61,884,224bp</td>
</tr>
</tbody>
</table>
Targeting region differences

Total Bases Targeted

- **Nimblegen**: 4.4Mb
- **Agilent**: 9.4Mb
- **Illumina**: 0.9Mb

**RefSeq (coding):**
- **Nimblegen**: 200kb
- **Agilent**: 10kb
- **Illumina**: 10kb

**RefSeq (UTR):**
- **Nimblegen**: 400kb
- **Agilent**: 28Mb
- **Illumina**: 28Mb

**Ensembl (CDS):**
- **Nimblegen**: 1.4Mb
- **Agilent**: 300kb
- **Illumina**: 300kb

**miRBase:**
- **Nimblegen**: 55kb
- **Agilent**: 60kb
- **Illumina**: 28kb
## Assay differences

<table>
<thead>
<tr>
<th></th>
<th>Nimblegen</th>
<th>Agilent</th>
<th>Illumina</th>
</tr>
</thead>
<tbody>
<tr>
<td>Starting DNA (pg)</td>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Duration (days)</td>
<td>7</td>
<td>3.5</td>
<td>3.5</td>
</tr>
<tr>
<td>Hybridization (hours)</td>
<td>72</td>
<td>24</td>
<td>2 x (16-20)</td>
</tr>
<tr>
<td>Pre-hyb PCR</td>
<td>12</td>
<td>4-6</td>
<td>10</td>
</tr>
<tr>
<td>Post-hyb PCR</td>
<td>18</td>
<td>10-12</td>
<td>10</td>
</tr>
<tr>
<td>qPCR validation</td>
<td>Yes</td>
<td>None</td>
<td>Recommended</td>
</tr>
<tr>
<td>Molecule used</td>
<td>DNA</td>
<td>RNA</td>
<td>DNA</td>
</tr>
<tr>
<td>Automation</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Experiment! Compare enrichment assays

Illumina TruSeq Exome
Agilent SureSelect 50M
Nimblegen EZ Exome 2.0

The Fun Part
## Data filtering and normalization

<table>
<thead>
<tr>
<th>Platform</th>
<th>Raw Reads</th>
<th>Mapped</th>
<th>Uniquely Mapped</th>
<th>Not Duplicates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agilent</td>
<td>124,112,466</td>
<td>123,202,356</td>
<td>112,602,746</td>
<td>94,779,030</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(99.2%)</td>
<td>(91.4%)</td>
<td>(84.2%)</td>
</tr>
<tr>
<td>Nimblegen</td>
<td>184,983,780</td>
<td>183,502,451</td>
<td>175,593,794</td>
<td>154,270,343</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(99.2%)</td>
<td>(95.7%)</td>
<td>(87.9%)</td>
</tr>
<tr>
<td>Illumina</td>
<td>112,885,944</td>
<td>110,977,932</td>
<td>100,236,505</td>
<td>88,759,249</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(98.3%)</td>
<td>(90.3%)</td>
<td>(88.5%)</td>
</tr>
</tbody>
</table>

80,000,000 reads
Targeting efficiency

Overall Targeting Efficiency

% of Total Targeted Bases

Depth of Coverage

Legend:
- Red (N)
- Blue (A)
- Yellow (I)
Targeting efficiency

Results:

- Nimblegen = most efficient
- Illumina = captures the most
- Illumina requires more sequencing

Denser baits result in higher efficiency and less sequencing.

Less dense designs take more sequencing, but capture more.
Off-target enrichment

Coverage

Off-target

Exon

Target Region

-500b

+500b
Nimblegen

- Off-target: 9%
- On-target: 91%

Agilent

- Off-target: 13%
- On-target: 87%

Illumina

- Off-target: 36%
- On-target: 64%
Off-target enrichments

% overlapping RepeatMasker

<table>
<thead>
<tr>
<th></th>
<th>On-target</th>
<th>Off-target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agilent</td>
<td>8.0%</td>
<td>38.3%</td>
</tr>
<tr>
<td>Nimblegen</td>
<td>7.7%</td>
<td>32.7%</td>
</tr>
<tr>
<td>Illumina</td>
<td>6.6%</td>
<td>47.0%</td>
</tr>
</tbody>
</table>

% overlapping Segdups

<table>
<thead>
<tr>
<th></th>
<th>On-target</th>
<th>Off-target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agilent</td>
<td>10.9%</td>
<td>19.5%</td>
</tr>
<tr>
<td>Nimblegen</td>
<td>5.0%</td>
<td>10.9%</td>
</tr>
<tr>
<td>Illumina</td>
<td>9.0%</td>
<td>18.9%</td>
</tr>
</tbody>
</table>
Small variant analysis

Comparing sensitivity and trends in small variant detection by exome-seq
Single nucleotide variations

SNVs detected

SNP chip concordance

Reference Bias
SNV detection trends

Total SNVs

# of Reads (millions)

# variants

Platform
- Nimblegen
- Agilent
- Illumina
Platform-specific SNVs
Indel detection trends

Total Indels

![Graph showing Indel detection trends for different platforms (Nimblegen, Agilent, Illumina) across different numbers of reads (in millions) with varying number of variants. The graph highlights the trend and performance of each platform.]
Indel size distributions

**Indel Size Distribution**

- Platform
  - Nimblegen
  - Agilent
  - Illumina

**Indel Size Distribution (RefSeq Coding Exons Only)**
Comparision of model exome-seq and whole genome sequencing experiments
Create model experiments

Whole Genome Sequencing

- Hi-seq 2000 (7 lanes)
- 1,194,622,756 reads
- 35x genome-wide coverage
- 98.5% SNP chip concordance

Exome-Sequencing

- Illumina TruSeq Exome
- 1-lane Hi-seq 2000
- Normalized reads (50M-80M)
- Coverage range from 30x-48x
Quality correlates with coverage

**Mean variant quality score**

- **Read Count**
  - 50M
  - 60M
  - 70M
  - 80M

- **Mean variant quality score**
  - All SNVs
  - Exome-only

**Mean Base Coverage**

- **Read Count**
  - 50M
  - 60M
  - 70M
  - 80M

- **Values**
  - 30
  - 36
  - 42
  - 48
## SNV overlap

<table>
<thead>
<tr>
<th>Runs</th>
<th>WGS</th>
<th>Exome-seq</th>
<th>Exome-seq</th>
</tr>
</thead>
<tbody>
<tr>
<td>50M</td>
<td>6,126</td>
<td>42,633</td>
<td>5,408</td>
</tr>
<tr>
<td>60M</td>
<td>5,634</td>
<td>43,125</td>
<td>5,687</td>
</tr>
<tr>
<td>70M</td>
<td>5,329</td>
<td>43,430</td>
<td>5,881</td>
</tr>
<tr>
<td>80M</td>
<td>5,083</td>
<td>43,676</td>
<td>6,060</td>
</tr>
</tbody>
</table>

- 50M reads / 30x exome-seq, 35x WGS
- 60M reads / 36x exome-seq, 35x WGS
- 70M reads / 42x exome-seq, 35x WGS
- 80M reads / 48x exome-seq, 35x WGS
Novel variant rates

- Whole exome
- Exome-seq-specific
- Whole Genome
- WGS-specific

False-positive rate of experiment-specific variants
Variant summary

60M
- Exome-seq-specific, 1,877
- WGS-Exome-seq-specific, 1,521
- Both, 43,125

70M
- WGS-specific, 1,447
- Both, 43,430

80M
- Exome-seq-specific, 2,020
- WGS-specific, 1,384
- Both, 43,676
Experiment-specific disease variants

Illumina TruSeq Exome
Normalized to 80M Reads

- Both, 43,676
- WGS-specific, 1,383
- Disease-associated, 467
- Exome-seq-specific, 2,020
- Other, 1,553
Conclusions

- Bait length, target choice, PCR cycles all influence exome-seq performance
- Greater coverage requires greater sequencing
- Most important factors to consider:
  - Regions covered
  - Amount of sequencing to be done
- Exome-seq observes important variants missed by a typical WGS experiment
- **These lessons extend to all target enrichment experiments**
Thank you
GC-content bias

- High or low GC-content
- reduces amplification in PCR
- negatively impacts oligonucleotide array hybridization
GC-content bias

GC-Content vs Targeting Efficiency (Agilent)

GC-Content vs Targeting Efficiency (Nimblegen)

GC-Content vs Targeting Efficiency (Illumina)
SNV detection trends
Indel detection trends

RefSeq (Coding) Indels

# of Reads (millions)

# variants