cDNA CaptureSeq
Reveals unappreciated diversity of the transcriptome

J.A. Jeddeloh, Ph.D.
Director, Business Development, Roche NimbleGen
jeffrey.jeddeloh@roche.com

For Life Science Research Only. Not for Use in Diagnostic Procedures.
Core Technology: Maskless Array Synthesis (MAS)

Differs from earlier approaches to light-directed synthesis in two critical respects:

1. **Novel way to direct light**
   - High info content (density)
   - Design speed and flexibility

2. **Proprietary high-yield chemistry**
   - Highest synthesis yields
   - Fast deprotection/coupling
   - High-fidelity long oligo probes
Thank you to the following contributors

- Thomas J. Albert
- Todd Richmond
- Tracy Millard
- Daniel Gerhardt
- Mark D’Ascenzo

- Faheem Niazi
- Brian Dessany
- Brian Godwin
- G. Ferreri
- Todd Arnold
- Jason Affourtit
- Thomas Jarvie

**cDNA Capture/Splicing**

- John Rinn
- Mitch Guttman
- Manuel Garber
- Joanna Crawford
- Cole Trapnell

- Tim Mercer
- Marcel Dinger
- John Mattick
NimbleGen SeqCap EZ Choice Workflow

- **Target Regions**
- **Prepare with Next-Gen Sequencing Adaptors**
- **SeqCap EZ Choice (Solution Capture)**
- **Amplify DNA and Enrichment QC**
- **Sequence DNA on a Next-Gen Sequencer**

**2.1M Sequence Capture Developer (Array Capture)**

<table>
<thead>
<tr>
<th>Genomic DNA</th>
<th>Library Preparation</th>
<th>Hybridization</th>
<th>Capture and Washing</th>
<th>Amplification and QC</th>
<th>Sequencing</th>
</tr>
</thead>
</table>
HOXA expression correlates with point of origin

J.L. Rinn et al 2007

In this study active chromatin was demonstrated to be demarcated by Noncoding RNA. Locus regulating HOX was discovered and named HOTAIR. PC2 partner

Discovered to be a class of ncRNA subsequently renamed lincRNA
lincRNA - Characterized by expression and tiling array

*If we capture them can we better map their exons*

**GOAL**
- Translate expression tiling information about lincRNA into transcript maps
  - **RNAseq**
    - Problem... about 20% of the reads go to most highly expressed mRNA. Some lincRNA are rare.
    - Splicing (even with PE) is imputed
  - **Targeted capture of cDNA**
    - Allow deep diving of molecular population
    - May complement RNAseq by measuring abundance of lowest expressed proportion
    - 454 AND Illumina -> Mixed read assembly?
      » Long reads should help directly characterize diversity of splicing by focusing read depth against a defined target list from the RNA population
      » Large “N” will help understand Quant

- **Open questions**
  - Will it work? (need careful controls to guide interpretation)
  - Will it be quantitative?
Capture cDNA but use gDNA as a control

Establish probe response and evaluate expression normalization

• Array Capture Design (385K array)
  – ~2,000 exons or short regions targeted for capture
    • Putative Linc (Long Intergenic Non-Coding) RNA
    • Protein coding controls
  – 0.8Mb total

• 454 Titanium AND Illumina optimized capture library created and hybridized to capture array
  – Do RNASeq on Foot and Lung – compare to Capture of cDNA
    – 300bp insert libraries for PE sequencing
  – Make sequencing library from cDNA and capture from it
  – Samples
    • Foot fibroblast cell line (c45) cDNA and gDNA control
    • Lung fibroblast cell line (FL) cDNA and gDNA control
      – cDNA made with and without a size selecting gel slice
        » “Cut” and standard
      – gDNA was a ~700bp average fragment Ti library
Mapping Statistics (Shallow 454 Coverage)
gDNA vs. cDNA

<table>
<thead>
<tr>
<th></th>
<th>cut gDNA 45</th>
<th>cut cDNA 45</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial targets</td>
<td>2092</td>
<td>2092</td>
</tr>
<tr>
<td>Initial target bases</td>
<td>830202</td>
<td>830202</td>
</tr>
<tr>
<td>Final targets</td>
<td>1846</td>
<td>1846</td>
</tr>
<tr>
<td>Final target bases</td>
<td>811370</td>
<td>811370</td>
</tr>
<tr>
<td>Target bases covered</td>
<td>773185</td>
<td>223146</td>
</tr>
<tr>
<td>Percent target bases covered</td>
<td>95.3</td>
<td>27.5</td>
</tr>
<tr>
<td>Target bases not covered</td>
<td>38185</td>
<td>588224</td>
</tr>
<tr>
<td>Percent target bases not covered</td>
<td>4.7</td>
<td>72.5</td>
</tr>
<tr>
<td>Total reads</td>
<td>22564</td>
<td>28945</td>
</tr>
<tr>
<td>Number of reads in target regions</td>
<td>15259</td>
<td>18746</td>
</tr>
<tr>
<td>Percent of reads in target regions</td>
<td>67.6</td>
<td>64.8</td>
</tr>
<tr>
<td>Average coverage</td>
<td>5.2</td>
<td>5.8</td>
</tr>
<tr>
<td>Median coverage</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>

Shallow coverage:
- ~10Mb per sample
- Capture highly specific for both gDNA and cDNA
  - gDNA covered well
  - cDNA covered less well
- Expected due to non-uniform representation of transcribed regions
Uniformity Statistics – gDNA vs. cDNA

Cumulative Coverage (Normalized)

Uniformity:
- gDNA normal coverage distribution
- >80% of target exonic bases have 50% of the expected coverage
- cDNA coverage very non-uniform, as expected
Mapping Statistics (Deeper 454 Coverage)

gDNA vs. cDNA

<table>
<thead>
<tr>
<th></th>
<th>Lung gDNA</th>
<th>Foot gDNA</th>
<th>Lung cDNA</th>
<th>Foot cDNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOTAL_READS</td>
<td>185437</td>
<td>186637</td>
<td>262103</td>
<td>268573</td>
</tr>
<tr>
<td>TOTAL_BASEPAIRS</td>
<td>56102185</td>
<td>55133762</td>
<td>73425525</td>
<td>81247544</td>
</tr>
<tr>
<td>TOTAL_MATCHES</td>
<td>2826924</td>
<td>3786172</td>
<td>946025</td>
<td>1140888</td>
</tr>
<tr>
<td>PERCENT_READS_NO_MATCH</td>
<td>9.80%</td>
<td>10.60%</td>
<td>2.80%</td>
<td>3.10%</td>
</tr>
<tr>
<td>NUM_BASEPAIRS_NO_MATCH</td>
<td>2788580</td>
<td>2993466</td>
<td>820234</td>
<td>986838</td>
</tr>
<tr>
<td>PERCENT_BASEPAIRS_NO_MATCH</td>
<td>5.00%</td>
<td>5.40%</td>
<td>1.10%</td>
<td>1.20%</td>
</tr>
<tr>
<td>NUM_READS_UNIQUELY_MAPPED</td>
<td>165993</td>
<td>165063</td>
<td>250832</td>
<td>256681</td>
</tr>
<tr>
<td>PERCENT_READS_UNIQUELY_MAPPED</td>
<td>89.50%</td>
<td>88.40%</td>
<td>95.70%</td>
<td>95.60%</td>
</tr>
<tr>
<td>NUM_BASEPAIRS_UNIQUELY_MAPPED</td>
<td>51915060</td>
<td>50543591</td>
<td>57643720</td>
<td>61507568</td>
</tr>
<tr>
<td>PERCENT_BASEPAIRS_UNIQUELY_MAPPED</td>
<td>92.50%</td>
<td>91.70%</td>
<td>78.50%</td>
<td>75.70%</td>
</tr>
<tr>
<td>NUM_BASEPAIRS_HSP_TRIMMED</td>
<td>1079817</td>
<td>1134159</td>
<td>1408614</td>
<td>17958647</td>
</tr>
<tr>
<td>PERCENT_BASEPAIRS_HSP_TRIMMED</td>
<td>1.90%</td>
<td>2.10%</td>
<td>19.20%</td>
<td>22.10%</td>
</tr>
<tr>
<td>Initial targets</td>
<td>2092</td>
<td>2092</td>
<td>2092</td>
<td>2092</td>
</tr>
<tr>
<td>Initial target bases</td>
<td>830202</td>
<td>830202</td>
<td>830202</td>
<td>830202</td>
</tr>
<tr>
<td>Final targets</td>
<td>1846</td>
<td>1846</td>
<td>1846</td>
<td>1846</td>
</tr>
<tr>
<td>Final target bases</td>
<td>811370</td>
<td>811370</td>
<td>811370</td>
<td>811370</td>
</tr>
<tr>
<td>Target bases covered</td>
<td>807557</td>
<td>806437</td>
<td>404035</td>
<td>334221</td>
</tr>
<tr>
<td>Percent target bases covered</td>
<td>99.5</td>
<td>99.4</td>
<td>49.8</td>
<td>41.2</td>
</tr>
<tr>
<td>Target bases not covered</td>
<td>3813</td>
<td>4933</td>
<td>407335</td>
<td>477149</td>
</tr>
<tr>
<td>Percent target bases not covered</td>
<td>0.5</td>
<td>0.6</td>
<td>50.2</td>
<td>58.8</td>
</tr>
<tr>
<td>Total reads</td>
<td>165993</td>
<td>165063</td>
<td>250832</td>
<td>256681</td>
</tr>
<tr>
<td>Number of reads in target regions</td>
<td>120481</td>
<td>109857</td>
<td>158006</td>
<td>154725</td>
</tr>
<tr>
<td>Percent of reads in target regions</td>
<td>72.6</td>
<td>66.6</td>
<td>63</td>
<td>60.3</td>
</tr>
<tr>
<td>Average coverage</td>
<td>36</td>
<td>32.5</td>
<td>40.4</td>
<td>42.3</td>
</tr>
<tr>
<td>Median coverage</td>
<td>35</td>
<td>32</td>
<td>40</td>
<td>42.3</td>
</tr>
</tbody>
</table>

- cDNA maps better than gDNA (no repeats, no rRNA)
- Exon junctions cause trimming
- All targets from gDNA present
- Less than 1/2 from cDNA
- Highly specific capture

~73% reads from 0.024% of genome = >3000 Fold Enrichment
Correlation analysis confirms quantitative nature

$qRT$-PCR vs RNAseq (il) vs CaptureSeq vs Replicates (454)
Capture cDNA provides unparalleled resolution

Confirmed previous expression tiling data

Probes

Foot cDNA Reads (bp)

Foot cDNA coverage (X)

Foot tiling array data (2007)

Lung tiling array data (2007)

Lung cDNA coverage (X)

Lung cDNA Reads (bp)

Foot gDNA coverage (X)

Lung gDNA coverage (X)
Capture of cDNA provides unparalleled resolution

*Splicing diversity amazing with long reads*

HOTAIR
• The full length UTR in the reads suggests that despite where the probes are spaced the coverage along the UTR is appropriate.
• The missing bases in the center of the UTR have nothing to do with the probes
• ....And everything to do with processing of the UTR. ~15 different isoforms of this message’s UTR

Capture of cDNA provides unparalleled resolution
Splicing diversity amazing with long reads
Mixed read length assembly splicing analysis

(a) Annotated gene loci
(b) Intergenic loci

Tracks (from outer edge):
- Annotated gene
- RNAseq
- Assembled exon
- Splice junction
- Probed region

- RNAseq
- Novel assembled exon
- Novel splice junction
HOX-A differential expression compare

*Illumina analysis*
P53– Coverage compare Capture vs. RNAseq

*a* RNAseq

*b* Capture Seq

**Known alternative exons/splicing**

**Novel alternative exons/splicing**

*Illumina*
P53 – novel alternative splicing and expression

Expression RPKM

- p53
- i
- ii
- iii
- iv

Known: Exon, Intron
Novel: Exon

Domain:
- Transactivation
- DNA-binding
- Nuclear local signal
- Oligomerisation

Variant

WT

Foot
Lung
P53- RT-PCR validation of 4/4 NOVEL EXONS
Conclusions

cDNA CaptureSeq

- Supplements information available from RNAseq by affording a targeted zoom on data analysis
- DOES NOT NORMALIZE expression but quantitatively samples
- Preserves known differential expression patterns
  - Reproducibly recapitulated known truth - Foot vs. Lung
- Mixed read analysis afforded expanded validation
  - Long-reads provided support despite only ~300K attempts
Conclusions

cDNA CaptureSeq

- Discovered 745 novel intergenic transcripts
  - Exp - subset of cells, perhaps less than 1 in 1000 cells

- Established the baseline of the transcriptome is far below the ability of a single HiSeq lane to resolve
  - >10 B attempts of RNAseq would be required to garner same coverage level as that achieved by CaptureSeq

- Enhanced zoom afforded discovery of four RARE transcripts of P53
Maize translocation visualization
Chr1 – 6FAM, Chr 5 – Tx Red, CentC - White
image from Danilova, Alberts and Birchler

Unless explicitly stated otherwise, all Roche NimbleGen and 454 products and services referenced in this presentation are intended for the following use:

For Life Science Research Only.

Not for Use in Diagnostic Procedures.

454, 454 SEQUENCING, NIMBLEGEN and SEQCAP are trademarks of Roche.

Other brands or product names are trademarks of their respective holders.

© 2012 Roche NimbleGen, Inc.
We Innovate Healthcare