Advanced RNA-Seq course
October 15-17, 2012

Introduction

Peter-Bram ’t Hoen
Expression profiling

• DNA $\rightarrow$ mRNA $\rightarrow$ protein

• Comprehensive RNA profiling possible: determine the abundance of all mRNA molecules in a cell / tissue
Expression profiling: applications

• Qualitative: which part of the genome is expressed, in which cells, which mRNA isoforms

• Quantitative: compare across conditions, understand biological processes / mechanisms
  • Tumor vs. Normal tissue
  • Knock-out vs. wild-type mouse
  • Changing nutrient conditions in yeast
  • Etc.
Transcriptome analysis

• Genome-wide expression profiling
  • Serial Analysis of Gene Expression (SAGE)
  • Expression Microarrays
  • Digital Gene Expression (DeepSAGE)
  • Shotgun RNA sequencing
Expression microarray

- Relative abundance
- Limited by content
Serial analysis of gene expression (SAGE)

- Sequence and count short tags representative for a transcript
- Absolute abundance of transcript
NGS-based sequencing vs. microarray

• RNA-seq
  • Counting
  • Absolute abundance of transcript
  • All transcripts present

• Expression microarray
  • Recording hybridization signal to complementary probe
  • Relative abundance
  • Cross-hybridization possible
  • Content limited
Main RNA-seq applications

- Quantification of gene (transcript) expression
- Gene / transcript identification in species without reference genome
- Detection of novel (non-coding) transcripts
- Detection and quantification of alternative splicing, alternative promoter and alternative polyadenylation site usage
- Quantification of allele-specific expression
- eQTL (splice eQTL) analysis
- Fusion gene detection (cancer)
RNA-seq platforms

• Transcript structure (long reads / paired-end / mate-pair)
  → Illumina paired-end reads, PacBio or hybrid approaches

• Expression differences (millions of reads)
  → Illumina, SOLiD, Helicos
Number of reads required
Deep sequencing-based expression profiling

• Tag-based: one read per transcript
  • DeepSAGE → most 3’ CATG
  • DeepCAGE → 5’-end
  • PolyA -> ultimate 3’-end

• RNA-Seq: multiple reads per transcript
  • Whole mRNA sequencing after fragmentation

• miRNA (short RNA) sequencing
1) oligo(dT) bead capture
2) single strand cDNA synthesis
3) double stranded cDNA synthesis
4) Restriction Digest 1
5) Add linker 1 & Restriction Digest 2
6) Add linker 2
7) Sequencing
Example gene: Gapd

14542

12555
Example gene: alternative polyadenylation
CAGE (Cap analysis of gene expression)

1) Capture 5' CAP
2) single strand cDNA synthesis (random priming)
3) Add Linker 1
4) double stranded cDNA synthesis
5) Restriction Digest
6) Add Linker 2
7) Sequencing
Example CAGE
More new transcription start sites (CAGE)

Better annotation of promoter regions
**General RNA-seq sample prep**

1. Isolation of polyA+ mRNA with oligo-dT
2. Fragmentation by heating for 8 min at 94°C
3. Random-primed first and second strand cDNA synthesis
4. End repair
5. Fragmentation
6. Adenylation of 3’-ends
7. Ligation of adapters (containing barcodes)
8. PCR amplification (15 cycles)
9. Clean-up
Example RNA-Seq
Alternative splicing
**Ovation: not so random-primed**

- No polyA+ selection
- No fragmentation
Protocol for strand-specific RNAseq

First-strand synthesis with normal dNTP's
Second-strand synthesis with dTTP -> dUTP
Y-adaptor ligation
UNG treatment
Preamplification and sequencing from #Ad1 side

Helicos single molecule sequencing

Helicos tSMS™
Sequencing by Synthesis

1. Synthesize
2. Wash
3. Image
4. Cleave

Sequencing by Synthesis
Example RNA-Seq (Helicos)

**ADAMTS8**

**ADAMTS15**

**NOV**

Peter Henneman
Example polyA profiling on Helicos

1. mRNA capture

2. First- and Second-Strand synthesis on the beads

3. NlaIII Digestion

4. Denaturation and isolation of the Second-Strand

5. PolyA extension and block

6. Hybridization on oligo(dT50)-coated flowcell surface

7. "Fill and lock" and sequencing-by-synthesis

---

de Klerk et al., Nucleic Acids Research 40:9089-9101 (2012)
Example polyA profiling

de Klerk et al., Nucleic Acids Research 40:9089-9101 (2012)
Direct RNA sequencing

Fatih Ozsolak\textsuperscript{1}, Adam R. Platt\textsuperscript{1}, Dan R. Jones\textsuperscript{1}, Jeffrey G. Reifenberger\textsuperscript{1}, Lauryn E. Sass\textsuperscript{1}, Peter McInerney\textsuperscript{1}, John F. Thompson\textsuperscript{1}, Jayson Bowers\textsuperscript{1}, Mirna Jarosz\textsuperscript{1} & Patrice M. Milos\textsuperscript{1}
Example PacBio whole transcript sequencing

Henk Buermans, unpublished
Small RNA profiling

- SOLiD® Total RNA-Seq (Invitrogen)
- Strand-specific
- Ligation-dependent

Modification for Illumina:
Comparison to microarrays

Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms

Peter A. C. ’t Hoen¹,*; Yavuz Ariyurek¹, Helene H. Thygesen¹, Emo Vreugdenhil², Rolf H. A. M. Vossen¹, Renée X. de Menezes¹, Judith M. Boer¹, Gert-Jan B. van Ommen¹ and Johan T. den Dunnen¹

¹The Center for Human and Clinical Genetics and the Leiden Genome Technology Center, Leiden University Medical Center and ²The Department of Medical Pharmacology from the Leiden/Amsterdam Center for Drug Research, Leiden, The Netherlands
Illumina features: Excellent reproducibility

Raw data

Square root-transformed and scaled data
Excellent reproducibility between labs

\[ y = 0.9768x + 0.3347 \]

\[ R^2 = 0.9579 \]
Analysis of replicate samples

- Pooling: small contaminations can have large effect on outcome

<table>
<thead>
<tr>
<th>GENE</th>
<th>Name</th>
<th>Pool WT</th>
<th>Pool dC</th>
<th>WT1</th>
<th>WT3</th>
<th>WT4</th>
<th>WT6</th>
<th>dC1</th>
<th>dC2</th>
<th>dC3</th>
<th>dC4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exosc8</td>
<td>Exosome component 8</td>
<td>14</td>
<td>0</td>
<td>28</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Fgg</td>
<td>Fibrinogen, gamma polypeptide</td>
<td>60</td>
<td>0</td>
<td>72</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Gc</td>
<td>Group specific component</td>
<td>22</td>
<td>0</td>
<td>41</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Itih4</td>
<td>Inter alpha-trypsin inhibitor, heavy chain 4</td>
<td>26</td>
<td>0</td>
<td>51</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Mug1</td>
<td>Murinoglobulin 1</td>
<td>20</td>
<td>0</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mup1</td>
<td>Major urinary protein 1</td>
<td>14</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mup1</td>
<td>Major urinary protein 1</td>
<td>18</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Orn1</td>
<td>Orosomucoid 1</td>
<td>11</td>
<td>0</td>
<td>22</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Rdh7</td>
<td>Retinol dehydrogenase 7</td>
<td>17</td>
<td>0</td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Serpina1a</td>
<td>Serine (or cysteine) peptidase inhibitor, clade A, member 1a</td>
<td>35</td>
<td>0</td>
<td>71</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Serpina3k</td>
<td>Serine (or cysteine) peptidase inhibitor, clade A, member 3K</td>
<td>87</td>
<td>0</td>
<td>143</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- Technical replicates: not really necessary when sufficient sequencing depth is reached
- Biological replicates important for determination of biological variation
Power comparison

Van Iterson, BMC Genomics, 2009
Power comparison (2)

Intensity range

Power

Number of samples

- Affymetrix
- Agilent
- Solexa
CAGE vs. SAGE

Tissue-specific transcript annotation and expression profiling with complementary next-generation sequencing technologies

Matthew S. Hestand¹,², Andreas Klingenhoff³, Matthias Scherf³, Yavuz Ariyurek², Yolande Ramos⁴, Wilbert van Workum⁵, Makoto Suzuki⁶, Thomas Werner³, Gert-Jan B. van Ommen¹, Johan T. den Dunnen¹,², Matthias Harbers⁶ and Peter A.C. ’t Hoen¹,*
Correlation CAGE vs. SAGE (gene level)

Logratio differentiated vs. proliferating

Differentially expressed genes

<table>
<thead>
<tr>
<th></th>
<th>SAGE</th>
<th>CAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Differentially expressed (*)</td>
<td>Differentially expressed (*)</td>
</tr>
<tr>
<td></td>
<td>2144</td>
<td>2160</td>
</tr>
<tr>
<td></td>
<td>Unchanged</td>
<td>Unchanged</td>
</tr>
<tr>
<td></td>
<td>1702</td>
<td>3234</td>
</tr>
</tbody>
</table>

Only detected with CAGE: 1169
Only detected with SAGE: 1747

* Bayesian error rate < 0.05
### Top genes SAGE and CAGE

<table>
<thead>
<tr>
<th>CAGE gene</th>
<th>Ratio</th>
<th>Microarray</th>
<th>SAGE gene</th>
<th>Ratio</th>
<th>Microarray</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hfe2</td>
<td>4,073</td>
<td>NA</td>
<td>RP23-36P22.5</td>
<td>576</td>
<td>NA</td>
</tr>
<tr>
<td>Myom3</td>
<td>1,624</td>
<td>NA</td>
<td>Neb</td>
<td>525</td>
<td>NA</td>
</tr>
<tr>
<td>Lmod2</td>
<td>1,305</td>
<td>NA</td>
<td>Mylpf</td>
<td>504</td>
<td>Yes</td>
</tr>
<tr>
<td>Myh7</td>
<td>1,124</td>
<td>Yes</td>
<td>Ttn</td>
<td>380</td>
<td>NA</td>
</tr>
<tr>
<td>Mb</td>
<td>908</td>
<td>Yes</td>
<td>Myh3</td>
<td>368</td>
<td>Yes</td>
</tr>
<tr>
<td>RP23-36P22.5</td>
<td>735</td>
<td>NA</td>
<td>Xirp1</td>
<td>306</td>
<td>Yes</td>
</tr>
<tr>
<td>Pygm</td>
<td>717</td>
<td>Yes</td>
<td>1110002H13Ri</td>
<td>263</td>
<td>NA</td>
</tr>
<tr>
<td>Myl4</td>
<td>614</td>
<td>Yes</td>
<td>Tnnnc1</td>
<td>232</td>
<td>Yes</td>
</tr>
<tr>
<td>Synpo21</td>
<td>595</td>
<td>NA</td>
<td>Cav3</td>
<td>150</td>
<td>Yes</td>
</tr>
<tr>
<td>Myh1</td>
<td>561</td>
<td>Yes</td>
<td>Cbfa2t3</td>
<td>133</td>
<td>Yes</td>
</tr>
<tr>
<td>……</td>
<td>……</td>
<td>……</td>
<td>……</td>
<td>……</td>
<td>……</td>
</tr>
</tbody>
</table>

13 out of 30 not found by microarray

10 out of 30 not found by microarray
## Most significant pathways

<table>
<thead>
<tr>
<th>CAGE GO</th>
<th>SAGE GO</th>
<th>Microarray GO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regulation of striated muscle contraction</td>
<td>Regulation of muscle contraction</td>
<td><em>Cycline-dependent protein kinase inhibitor activity</em></td>
</tr>
<tr>
<td>Cardiac muscle contraction</td>
<td>Cardiac muscle contraction</td>
<td>Myogenesis</td>
</tr>
<tr>
<td>Myogenesis</td>
<td>Myogenesis</td>
<td>Skeletal muscle development</td>
</tr>
<tr>
<td>Regulation of muscle contraction</td>
<td>Regulation of striated muscle contraction</td>
<td>Myoblast differentiation</td>
</tr>
<tr>
<td>Skeletal muscle development</td>
<td>Skeletal muscle development</td>
<td><em>6-phosphofructokinase activity</em></td>
</tr>
<tr>
<td>Muscle development</td>
<td>Myofibril assembly</td>
<td>Muscle development</td>
</tr>
<tr>
<td>Striated muscle contraction</td>
<td>Muscle development</td>
<td>Muscle cell differentiation</td>
</tr>
<tr>
<td>Myoblast differentiation</td>
<td>Myoblast fusion</td>
<td><em>Tumor suppressor activity</em></td>
</tr>
<tr>
<td>Muscle cell differentiation</td>
<td>Striated muscle contraction</td>
<td>Myofibril assembly</td>
</tr>
<tr>
<td>Sarcomere organization</td>
<td>Muscle cell differentiation</td>
<td>Heart development</td>
</tr>
<tr>
<td>10/10 muscle related</td>
<td>10/10 muscle related</td>
<td>7/10 muscle related</td>
</tr>
</tbody>
</table>
Conclusions

• Next generation sequencing provides higher power, sensitivity and reproducibility than expression microarrays

• RNAseq offers more than microarrays
  • Alternative transcription start site usage
  • Alternative splicing
  • Alternative polyadenylation
  • Allele-specific expression
  • Small RNAs other than miRNAs
Acknowledgements

Henk Buermans
Eleonora de Klerk
Yavuz Ariyurek
Matt Hestand
Johan den Dunnen
Gertjan van Ommen

Matthias Harbers
Makoto Suzuki
Andreas Klinghoff
Matthias Scherf
Thomas Werner

Wilbert van Workum