RNA sequencing

Paul Bertone
Solexa transcriptome sequencing

- **Solexa data analysis and associated software development**
  - Unbiased expression profiling
  - Tandem identification of expressed non-coding RNAs
  - MicroRNA identification and expression analysis

- **Advantages over microarrays**
  - Gene expression arrays don’t capture unannotated transcripts
  - Tiling arrays are still expensive for large genomes (e.g. mammals)
  - Small RNAs are too short for stable hybridization
  - No fluorescence correction to account for, essentially zero background

- **Current disadvantages**
  - More expensive than standard expression arrays
  - More time consuming than any microarray technology
  - Some data analysis issues
    - No strand orientation information – sequencing a double-stranded product
    - Computing accurate transcript models, mapping reads to splice junctions
    - Contribution of high-abundance RNAs (e.g. ribosomal) could dilute the remaining transcript population; sequencing depth is important
Transcriptome sequencing methods

Method 1: variant of the LongSAGE protocol
- Poly-A RNA selection
- Double strand cDNA synthesis on beads
- NlaIII digestion to remove 5' portion of cDNAs
- Ligation to 5' adapters containing a MmeI recognition site
- MmeI digestion to remove the 3' portion of cDNA
- This generates a 17nt tag (not including CATG)
- Tags are ligated to a 3' adapter
- The construct is PCR-amplified using primers homologous to 5' and 3' adapters
- PCR products are purified and quantitated (e.g. with Agilent Bioanalyzer)
- Load tag–adapter hybrids into flow cell lanes and sequence

- No concatenation of SAGE tags
- One tag is amplified and sequenced per flow cell cluster
- Read (tag) alignment is performed against a library of virtual tags
**SAGE sequencing output**

<table>
<thead>
<tr>
<th>Command</th>
<th>Arguments</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. <strong>sage</strong></td>
<td></td>
<td>Generates SAGE sequencing output for the specified sample.</td>
</tr>
<tr>
<td>2. <strong>bowtie2</strong></td>
<td></td>
<td>Aligns reads using Bowtie2.</td>
</tr>
<tr>
<td>3. <strong>samtools</strong></td>
<td></td>
<td>Performs variant calling and annotation.</td>
</tr>
<tr>
<td>4. <strong>snpEff</strong></td>
<td></td>
<td>Integrates variant annotations with genome annotations.</td>
</tr>
<tr>
<td>5. <strong>bedtools</strong></td>
<td></td>
<td>Performs interval operations and genome analysis.</td>
</tr>
</tbody>
</table>

*Note: The figures and tables are placeholders and are not actual output from SAGE sequencing.*
Solexa transcriptome sequencing
Solexa transcriptome sequencing

Differential read counts allow us to discern which transcript isoforms are expressed
Features of SAGE analysis

- Complicated library construction
- Good at gene expression analysis
- Short reads (17nt), therefore low rate of unique alignments to reference genome
  - Reads are mapped to virtual tags instead
- Mostly limited to annotated genes
- Can get some information on novel transcripts (limited)
mRNA sequencing

- Similar to SAGE analysis in terms of gene expression

- Simpler library construction

- Not limited to 17nt reads
  - Utilize full read length for alignment
  - Much better genome mapping

- Results are analogous to tiling array profiling
  - Reads map to individual transcript components
  - Ascertain splice variation as well as gene expression
  - Refine existing annotation of exons and UTRs
  - Identify non-coding RNAs
mRNA-seq protocol
Reads mapped to the human genome
Alignment to exon splice junctions

Alignment reference should consider mature transcripts or exon junctions
Sequencing vs. tiling array hybridization

- Example comparison between Solexa WTSS and tiling array hybridization data (S. pombe, Bahler lab Sanger)

- Top image = sense strand; bottom image = antisense strand
- Light blue = annotated genes; Dark blue = new non-coding transcript; Green = intron

Red: tiling array hybridization signal (log2)
Black: sequencing reads (log)
Novel Transcribed Regions: Possibilities

- Many areas of active transcription are observed outside annotated genes
  - Rare or low-abundance protein-coding transcripts
  - Unannotated exons from alternate splice products
  - Previously under-represented 3’ and 5’ UTRs
  - Noncoding RNAs
Splice variation, refinement of existing exon annotation
Detection of microRNA precursors
Protocol variations

- Fragmentation methods
  - RNA: nebulization, hydrolysis
  - cDNA: sonication, Dnase I treatment
- Depletion of highly abundant transcripts
  - e.g. RiboMinus – others?
- Oligo–dT selection for poly(A)+ transcripts vs total RNA
- Coverage issues
  - What is the sequencing depth required?
- Strand specificity
  - Most RNA sequencing is not strand–specific
  - Currently working with Vladimir Benes and Lars Steinmetz on new protocols for this
Specialized RNA-seq applications

• Small RNA sequencing
  • microRNAs
  • piRNAs
  • endo-siRNAs

• Identification of RNAs associated with protein complexes (e.g. Ago2)
  • Immunoprecipitation of RNA-bound protein complexes
  • Proteinase K digestion, purification of nucleic acids for sequencing
Growing number of non-coding RNA classes categorized by many different features (e.g. function, length, secondary structures, expression tissues, species, etc.)

For my projects I am focusing on short regulatory non-coding RNAs.

..paying particular attention to the microRNA and piwiRNA classes
microRNAs and piwiRNAs

Differences

- miRNAs are generally shorter (~21–23nt) than piRNAs (~24–30nt)
- miRNAs are Dicer-dependent
- miRNAs are processed from a dsRNA precursor with a known secondary structure (piRNAs?)
- Expression of piRNAs is thought to be restricted to the germline
- miRNAs bind to Argonaute clade whilst piRNAs to the Piwi clade of the Argonaute protein family

Similarities

- Both show a 5’Up preference
- Both show a 2’O–methyl modification at their 3’ end (plant microRNAs only)
Differences in small RNA sequencing

- Size exclusion of total RNA
  - Selected to target particular species
  - e.g. 17–23nt for microRNAs, 25–32nt for piRNAs
  - 17–32nt can encompass both populations

- Direct ligation of adapters to RNA molecules

- Transcripts are typically shorter than the reads
  - Sequence into the adapters
  - Reveals strand specificity
Adapter masking, low-complexity filtering
Aligned RNA reads from RNA-seq

<table>
<thead>
<tr>
<th>Sample</th>
<th>KS35</th>
<th>KS45</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reads</td>
<td>3,559,384</td>
<td>5,861,316</td>
</tr>
<tr>
<td>Eland placement (total)</td>
<td>2,429,078 (68%)</td>
<td>4,109,776 (70%)</td>
</tr>
<tr>
<td>Unique, no mismatch</td>
<td>1,806,384</td>
<td>3,445,856</td>
</tr>
<tr>
<td>Unique, 1 mismatch</td>
<td>418,151</td>
<td>440,111</td>
</tr>
<tr>
<td>Unique, 2 mismatches</td>
<td>204,543</td>
<td>223,809</td>
</tr>
</tbody>
</table>
Read depth varies across different loci.
Transcriptional units from RNA-seq
Transcriptional units from RNA-seq
Annotating small RNA-seq libraries

- Identify expressed transcripts from trace read alignments to the target genome
- Determine what small RNA components are present
  - Screen for well known structural RNAs (e.g. ribosomal RNA, tRNAs, snoRNAs, etc)
  - Align transcripts to current version of miRbase to identify expressed microRNAs
  - Align transcripts to our own piRNA database built from recently published candidate piRNA sequences
- Set remaining unknown transcript population aside, examine for potentially novel RNAs
High-ranking miRbase alignments

<table>
<thead>
<tr>
<th>KS35 (Spermatocytes)</th>
<th>KS45 (Round Spermatids)</th>
<th>KS35 (Spermatocytes)</th>
<th>KS45 (Round Spermatids)</th>
</tr>
</thead>
<tbody>
<tr>
<td>microRNA</td>
<td>Depth</td>
<td>microRNA</td>
<td>Depth</td>
</tr>
<tr>
<td>mmu-miR-805</td>
<td>1124</td>
<td>mmu-miR-184</td>
<td>5026</td>
</tr>
<tr>
<td>mmu-miR-191</td>
<td>472</td>
<td>mmu-miR-28</td>
<td>2799</td>
</tr>
<tr>
<td>mmu-miR-298</td>
<td>307</td>
<td>mmu-miR-423-5p</td>
<td>2083</td>
</tr>
<tr>
<td>mmu-miR-107</td>
<td>295</td>
<td>mmu-miR-470</td>
<td>494</td>
</tr>
<tr>
<td>mmu-miR-99b</td>
<td>290</td>
<td>mmu-miR-191</td>
<td>462</td>
</tr>
<tr>
<td>mmu-miR-28</td>
<td>255</td>
<td>mmu-miR-10b</td>
<td>411</td>
</tr>
<tr>
<td>mmu-miR-470</td>
<td>247</td>
<td>mmu-miR-34c</td>
<td>342</td>
</tr>
<tr>
<td>mmu-miR-151-3p</td>
<td>245</td>
<td>mmu-miR-182</td>
<td>320</td>
</tr>
<tr>
<td>mmu-miR-423-5p</td>
<td>220</td>
<td>mmu-miR-16</td>
<td>302</td>
</tr>
<tr>
<td>mmu-miR-881</td>
<td>220</td>
<td>mmu-miR-881</td>
<td>272</td>
</tr>
<tr>
<td>mmu-miR-184</td>
<td>195</td>
<td>mmu-miR-195</td>
<td>256</td>
</tr>
<tr>
<td>mmu-let-7d</td>
<td>108</td>
<td>mmu-miR-465c-5p</td>
<td>255</td>
</tr>
<tr>
<td>mmu-miR-34c</td>
<td>107</td>
<td>mmu-miR-743b-3p</td>
<td>161</td>
</tr>
<tr>
<td>mmu-miR-103</td>
<td>103</td>
<td>mmu-miR-151-3p</td>
<td>153</td>
</tr>
<tr>
<td>mmu-miR-743b-3p</td>
<td>87</td>
<td>mmu-miR-298</td>
<td>132</td>
</tr>
<tr>
<td>mmu-miR-202-5p</td>
<td>82</td>
<td>mmu-miR-107</td>
<td>130</td>
</tr>
<tr>
<td>mmu-miR-1196</td>
<td>81</td>
<td>mmu-miR-1195</td>
<td>130</td>
</tr>
</tbody>
</table>
**Composition of piRNA Database**

**Total candidate sequences: 1,524,007**

<table>
<thead>
<tr>
<th>95%, 25nt</th>
<th>Number of piRNAs per publication (Tot=210,576)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-----------</td>
<td>-------------</td>
</tr>
<tr>
<td>16751777</td>
<td>3,538</td>
</tr>
<tr>
<td>17446352</td>
<td>594</td>
</tr>
<tr>
<td>16751776</td>
<td>150</td>
</tr>
<tr>
<td>16766680</td>
<td>0</td>
</tr>
<tr>
<td>16778019</td>
<td>159</td>
</tr>
<tr>
<td>16766679</td>
<td>18922463</td>
</tr>
</tbody>
</table>

**Data came from Table S4. After sorting the table and taking only the uncharacterized sequences there were 40,102 piRNA candidates.**


**Aravin A, Gaidatzis D, Pfeffer S, Lagos-Quintana M, Landgraf P, et al. (2006) A novel class of small RNAs bind to MILI protein in mouse testes. Nature 442: 203-207. PMID 16751777. The piRNA candidate sequences are in an Excel table. It’s supposed to be one of the files in the supplementary data, but is mislabeled on the website as S3 instead of S4. Removing the known sequences left 3,638 piRNA candidates from this study.**


**Grivna ST, Beyret E, Wang Z, Lin H (2006) A novel class of small RNAs in mouse spermatogenic cells. Genes Dev 20: 1709-1714. PMID 16766680. Data came from Table S1 (pdf). The sequences were only 40 of them.**


Composition of piRNA Database
Karyogram (1)
Currently annotated piRNA clusters in mouse genome
Karyogram (2)

Expressed transcripts within piRNA clusters (all levels)

Left of chromosomes:
- KS35 (Spermatocytes)

Right of chromosomes:
- KS45 (Round Spermatids)

Expression observed
Differential expression observed
Karyogram (3)

Transcript abundance at mid-range levels

- Left of chromosomes: KS35 (Spermatocytes)
- Right of chromosomes: KS45 (Round Spermatids)

Expression observed
Differential expression observed
Karyogram (3)

Transcript abundance at high levels

- Left of chromosomes: KS35 (Spermatocytes)
- Right of chromosomes: KS45 (Round Spermatids)

Expression observed
Differential expression observed
Karyogram (4)

Transcript abundance at very high levels

Left of chromosomes:
KS35 (Spermatocytes)

Right of chromosomes:
KS45 (Round Spermatids)

Expression observed
Differential expression observed
Analysis of novel RNA transcripts

- Transcribed regions fall into several categories
  - Correlate well with annotated (coding) gene loci
  - Correlate with existing non-coding RNAs
  - Novel transcripts

- Novel RNAs
  - To further characterize these, we perform RNA secondary structure prediction on thousands of candidate sequences
  - Look for favorable energy conformations
    - RNAfold (Vienna package), Mfold (Zucker lab)
  - Visualization of putative secondary structures
    - RNAplot (Vienna), StructureLab (Shapiro lab)
  - Homology across multiple species
Prediction of RNA Secondary Structure
Prediction of RNA Secondary Structure
Novel microRNA candidates conserved across species

Stable hairpin consensus structures
Stem sequences are highly conserved
Loop sequences are divergent (variable)
Structural features of piRNAs

- As piRNAs are such a new class of regulatory non-coding RNA, their secondary structural properties are unknown.

- Precursor transcripts are processed by a quasi-random mechanism:
  - Weak sequence preference near the 5' U.
Structural features of piRNAs

- Some structures can be identified based on features typically associated with microRNA hairpins
- It remains to be seen whether these will be characteristic of piRNAs as well
Summary

• Wide variety of RNA sequencing applications

• Library construction protocols differ according to the source material and aims of the experiment

• Open questions about strand specificity, level of coverage required for comprehensive transcriptome analysis

• Single- versus paired-end RNA sequencing
  • As read length increases, sequencing more single-end reads may be more informative