MicroRNA sequencing (miRNAseq),
Robust experimental design
Data analysis using the
CAP-miRSeq: A comprehensive analysis
pipeline for deep microRNA sequencing
E. Starr Hazard
Over view of this first lecture

1) Review of very basic miRNA biology

2) Examination of the CAPmiRSeq bioinformatics pipeline

3) Interface to Systems Biology
miRNA has been identified in all three eukaryotic branches as well as in Protists.

miRNAs came to science in the early 1990s but are in fact ancient biology.
MicroRNA sequencing (miRNAseq)

https://en.wikipedia.org/wiki/MicroRNA

6/7/16
Within the nucleus DROSHA and DGCR8 combine with Primary miRNA. Transcripts regions within the 3’UTR to cleave and release hairpin shaped pre-miRNA molecules. These Pre-miRNA are exported to the cytoplasm.
Two “rulers” within Dicer-1

1. Dicer-1 checks the distance from the 3’ overhang to the terminal loop of the pre-miRNA via its PAZ and helicase domains, respectively.

2. Dicer-1 then cuts the pre-miRNA at a fixed distance (~22 nt) from the 3’ overhang via its RNase III domains.
Guide and Passenger miRNA sequencing (miRNAseq) with Dicer and Helicase.

Guide

Passenger

see Naming Conventions Slide

Argonaute-2 Plus Helicase
Split duplex; Passenger miR* Is released and degrades.
Guide sequence remains bound.

“Charged” Argonaute-2
Can scan 3’UTR for
Regions complementary to Guide

MicroRNA sequencing (miRNAseq)
Argonaute2 with mature miRNA NOT miRNA*  
Argonaute miRNA (mature) aka RISC complex
Charged Argonaute (i.e. with mature miRNA) guide sequence ~ RNA-induced silencing complex (RISC). RISC complex scans 3'UTR for cognate sequences binds and terminates further gene transcription.
miRNA naming conventions

Example: ‘**hsa-mir-528**’; the numbers are sequential by date of accession.
   First three = species; ‘**hsa**’, Homo sapiens.
   Moniker for mature miRNA ‘**miR-528**’; Gene name ‘**mir-528**’,
   Stem-loop structure of primary miRNA ‘**mir528**’

Names such as **hsa-mir-121-1** and **hsa-mir-121-2**
   Refer to identical miRNA from distinct genomic loci.

**hsa-miR-121a** and **hsa-miR-121b** come from similar genes
   **hsa-mir-121a** and **hsa-mir-121b**.

   If two 22nt miRNAs are identified from the same precursor
   Then, the most abundant is **miR-56**; the less abundant **miR-56**
   If its difficult make a call based on abundance, then the miRNAs
   Might be named **miR-141-5p** and **miR-141-3p**

   There is also a unique, stable **db accession number** and **NCBI geneID**.
### Mature sequence hsa-miR-29b-3p

<table>
<thead>
<tr>
<th>Accession number</th>
<th>MIMAT0000100</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>hsa-miR-29b-3p</td>
</tr>
<tr>
<td>Previous IDs</td>
<td>hsa-miR-29b</td>
</tr>
<tr>
<td>Sequence</td>
<td>uagccaguguaacugugu</td>
</tr>
<tr>
<td>Stem-Loop</td>
<td>hsa-mir-29b-1 hsa-mir-29b-2</td>
</tr>
</tbody>
</table>

### References

1. PMID: 11914277  
   "miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs"  
   Mourelatos Z, Dosie J, Paushkin S, Sharma A, Charroux B, Abel L, Rappsilber J, Mann M, Dreyfuss G  

2. PMID: 15183728  
   "Human embryonic stem cells express a unique set of microRNAs"  

3. PMID: 17604727  
   "A mammalian microRNA expression atlas based on small RNA library sequencing"  

4. PMID: 17616659  
   "Patterns of known and novel small RNAs in human cervical cancer"  
   Liu WO, Pourmand N, Patterson BK, Fire A  
We get our data input from Bob Wilson’s Sequencing Facility

![Image of small RNA libraries]

**Purified Small RNA Libraries**

- Human Brain
- Constructs Containing
- 125 bp
- 100 bp
- 93 bp
- 75 bp

**Select Small RNA Libraries**

Below are two gel images representing small RNA libraries generated from human and mouse brain total RNA and a third library made from small RNA fragments purified from 1 μg of human brain total RNA.

Sequencing can be conducted on individual bands or from pooled bands. The 93 nucleotide band primarily contains mature microRNA generated from approximately 22 nucleotide small RNA fragments.

A second band containing piwi interacting RNAs, as well as some microRNAs and other regulatory small RNA molecules, corresponds to 100 nucleotides in length and is generated from approximately 30 nucleotide RNA fragments.

**Preparing Samples for Small RNA Sequencing Using the Alternative v1.5 Protocol**

FOR RESEARCH ONLY

**FASTQ files**
MicroRNA sequencing (miRNAseq)
CAP-miRSeq at MUSC
A comprehensive analysis pipeline for deep microRNA sequencing

Integrates read preprocessing

Alignment

Mature/precursor/novel miRNA qualification

Variant detection in miRNA coding region

Flexible differential expression between experimental conditions
FASTQ format files from Sequencers

The CAP-miRSeq analysis workflow

CAP-miRSeq
- Raw Reads (fastq/fastq.gz)
  - FastQC (pre-trim)
  - FastQC (post-trim)
  - Cutadapt
  - Adapter Trimmed Reads (fastq)
  - Bowtie
  - miRDeep2 mapper
    - Ref Genome
    - miRBase
  - miRDeep2 module
    - Summary report
      - Known/Novel miRNAs
      - Other RNAs
      - miRNA coding region variants
    - edgeR
      - Differential miRNAs

LEGEND
- Program
- File
- Reference

A. Read pre-processing
B. Core module for known and novel miRNA detection
C. Differential miRNA expression
D. Read visualization, SNV detection, all RNA quantification

MicroRNA sequencing (miRNAseq)
MicroRNA sequencing (miRNAseq)
Deep sequencing reads mapped to the genome

Optional: genome annotation

Discard reads that map to many genomic loci
Optional: discard reads that map to rRNAs, tRNAs, etc.

Use sequence reads to excise potential miRNA precursors from the genome

Discard unlikely miRNA precursors

miRDeep core algorithm: probabilistic scoring of structure and signature

Known and new mature and precursor miRNAs

Optional: estimate the number of false positives
Initial filtering
Eliminate potential precursors that did not fold into a hairpin or that had reads inconsistent with Dicer processing.

Probabilistic scoring of the potential miRNA precursors.
Each potential precursor sequence that passed the initial filtering was then scored probabilistically. The score is a log-odds probability of being a genuine miRNA precursor versus the probability that it is a background hairpin, given the evidence from the data:

1. score = \log \left( \frac{P(\text{pre} | \text{data})}{P(\text{bgr} | \text{data})} \right)
   
The probability of the sequence being a precursor is given by Bayes’ theorem:
   
2. P(\text{pre} | \text{data}) = \frac{P(\text{data} | \text{pre}) P(\text{pre})}{P(\text{data})}
3. P(\text{pre} | \text{data}) = \frac{P(\text{abs} | \text{pre}) P(\text{rel} | \text{pre}) P(\text{sig} | \text{pre}) P(\text{star} | \text{pre}) P(\text{nuc} | \text{pre})}{P(\text{pre})} / P(\text{data})
   
The same holds for the probability of the sequence being a background hairpin:
   
4. P(\text{bgr} | \text{data}) = \frac{P(\text{data} | \text{bgr}) P(\text{bgr})}{P(\text{data})}
5. P(\text{bgr} | \text{data}) = \frac{P(\text{abs} | \text{bgr}) P(\text{rel} | \text{bgr}) P(\text{sig} | \text{bgr}) P(\text{star} | \text{bgr}) P(\text{nuc} | \text{bgr})}{P(\text{bgr})} / P(\text{data})
### Survey of mirDeep2 performance for score cut-offs -10 to 10

<table>
<thead>
<tr>
<th>mirDeep2 score</th>
<th>novel miRNAs</th>
<th>known mirBase miRNAs</th>
<th>excision gearing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>predicted by mirDeep2</td>
<td>estimated false positives</td>
<td>true positives</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0 ± 0</td>
<td>0 ± 0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0 ± 0</td>
<td>0 ± 0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0 ± 0</td>
<td>0 ± 0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0 ± 0</td>
<td>0 ± 0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0 ± 0</td>
<td>0 ± 0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0 ± 0</td>
<td>0 ± 0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0 ± 0</td>
<td>0 ± 0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1 ± 1</td>
<td>2 ± 1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>1 ± 1</td>
<td>2 ± 1</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>5 ± 2</td>
<td>3 ± 2</td>
</tr>
<tr>
<td>0</td>
<td>11</td>
<td>19 ± 4</td>
<td>0 ± 0</td>
</tr>
<tr>
<td>-1</td>
<td>14</td>
<td>25 ± 5</td>
<td>0 ± 0</td>
</tr>
<tr>
<td>-2</td>
<td>18</td>
<td>33 ± 5</td>
<td>0 ± 0</td>
</tr>
<tr>
<td>-3</td>
<td>69</td>
<td>45 ± 6</td>
<td>24 ± 6</td>
</tr>
<tr>
<td>-4</td>
<td>112</td>
<td>79 ± 8</td>
<td>33 ± 8</td>
</tr>
<tr>
<td>-5</td>
<td>124</td>
<td>297 ± 15</td>
<td>0 ± 0</td>
</tr>
<tr>
<td>-6</td>
<td>134</td>
<td>382 ± 17</td>
<td>0 ± 0</td>
</tr>
<tr>
<td>-7</td>
<td>209</td>
<td>431 ± 16</td>
<td>0 ± 0</td>
</tr>
<tr>
<td>-8</td>
<td>279</td>
<td>466 ± 16</td>
<td>0 ± 0</td>
</tr>
<tr>
<td>-9</td>
<td>310</td>
<td>486 ± 16</td>
<td>0 ± 0</td>
</tr>
<tr>
<td>-10</td>
<td>337</td>
<td>502 ± 16</td>
<td>0 ± 0</td>
</tr>
</tbody>
</table>
### MicroRNA sequencing (miRNAseq)

### miRNA List

<table>
<thead>
<tr>
<th>Symbol</th>
<th>GeneID</th>
<th>Description</th>
<th>MirBase ID</th>
<th>C</th>
<th>C</th>
<th>T</th>
<th>T</th>
<th>T</th>
<th>logFC</th>
<th>Linear FC</th>
<th>FC</th>
<th>logCPM</th>
<th>LR</th>
<th>PValue</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa-miR-21-5p</td>
<td>406991A 21</td>
<td>microRNA 21</td>
<td>MI0000077859468766794125102218108244</td>
<td>0.20493</td>
<td>1.15263</td>
<td>1.15263</td>
<td>16.5517</td>
<td>9.87278</td>
<td>0.001677</td>
<td>0.045378</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsa-miR-548d-5p</td>
<td>693130</td>
<td>microRNA 548d-1</td>
<td>MI00036680009513</td>
<td>5.64134</td>
<td>2373</td>
<td>49.9129</td>
<td>5362</td>
<td>3.42364</td>
<td>2694</td>
<td>7.00103</td>
<td>3325</td>
<td>0.008146</td>
<td>0.107363</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsa-miR-99a-5p</td>
<td>407055NA 99a 101</td>
<td>microRNA 99a</td>
<td>MI0000101210371425</td>
<td>0.00178</td>
<td>1.00123</td>
<td>1.00123</td>
<td>5.0309</td>
<td>8.997388</td>
<td>0.999142</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

567 miRNA species identified in this Differential expression comparison. Based on FDR value entries are colored: blue (FDR <0.1), red (FDR <0.4) or black (FDR >0.4)
### Report on mRNA variants

<table>
<thead>
<tr>
<th>Chr</th>
<th>Pos</th>
<th>Ref</th>
<th>Alt</th>
<th>Precursor_miR</th>
<th>Mature_miRNA</th>
<th>Seed_Region</th>
<th>Genotype</th>
<th>Allelic_Depth</th>
<th>Read_Depth</th>
<th>Genotype_Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1</td>
<td>567783</td>
<td>T</td>
<td>C</td>
<td>hsa-mir-6723</td>
<td>hsa-miR-6723</td>
<td>5</td>
<td>C/C</td>
<td>0;2</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>chr1</td>
<td>1102563</td>
<td>G</td>
<td>A;T</td>
<td>hsa-mir-200b</td>
<td>.</td>
<td>0</td>
<td>A/T</td>
<td>2;50;186</td>
<td>246</td>
<td>99</td>
</tr>
<tr>
<td>chr1</td>
<td>1103319</td>
<td>C</td>
<td>A;T</td>
<td>hsa-mir-200a</td>
<td>.</td>
<td>0</td>
<td>A/T</td>
<td>1;5;12</td>
<td>18</td>
<td>99</td>
</tr>
<tr>
<td>chr1</td>
<td>1104457</td>
<td>C</td>
<td>A</td>
<td>hsa-mir-429</td>
<td>.</td>
<td>0</td>
<td>A/A</td>
<td>0;5</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>chr1</td>
<td>41220067</td>
<td>G</td>
<td>T</td>
<td>hsa-mir-30e</td>
<td>.</td>
<td>0</td>
<td>G/T</td>
<td>3;21</td>
<td>27</td>
<td>46</td>
</tr>
<tr>
<td>chr1</td>
<td>41222996</td>
<td>G</td>
<td>T</td>
<td>hsa-mir-30c-1</td>
<td>.</td>
<td>0</td>
<td>T/T</td>
<td>0;4</td>
<td>5</td>
<td>12</td>
</tr>
</tbody>
</table>
hsa-mir-30a
Shows both 5p and 3p reads/
MicroRNA sequencing (miRNAseq)
hsa-mir-493
Also shows -5p and 3p reads
MicroRNA sequencing (miRNAseq)
Some miRNA is down regulated

1260 868 416 152

Interpreting the down regulation requires inference from many sources

There are 632 predicted targets for hsa-miR-29b-3p in miRDB.

<table>
<thead>
<tr>
<th>Target Detail</th>
<th>Target Rank</th>
<th>Target Score</th>
<th>miRNA Name</th>
<th>Gene Symbol</th>
<th>Gene Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Details</td>
<td>1</td>
<td>100</td>
<td>hsa-miR-29b-3p</td>
<td>COL3A1</td>
<td>collagen, type III, alpha 1</td>
</tr>
<tr>
<td>Details</td>
<td>2</td>
<td>100</td>
<td>hsa-miR-29b-3p</td>
<td>ERCC6</td>
<td>excision repair cross-complementation group 6</td>
</tr>
<tr>
<td>Details</td>
<td>3</td>
<td>100</td>
<td>hsa-miR-29b-3p</td>
<td>NFIA</td>
<td>nuclear factor I/A</td>
</tr>
<tr>
<td>Details</td>
<td>4</td>
<td>100</td>
<td>hsa-miR-29b-3p</td>
<td>TET3</td>
<td>tet methylcytosine dioxygenase 3</td>
</tr>
<tr>
<td>Details</td>
<td>5</td>
<td>100</td>
<td>hsa-miR-29b-3p</td>
<td>DGKH</td>
<td>diacylglycerol kinase, eta</td>
</tr>
<tr>
<td>Details</td>
<td>6</td>
<td>100</td>
<td>hsa-miR-29b-3p</td>
<td>BRWD3</td>
<td>bromodomain and WD repeat domain containing 3</td>
</tr>
<tr>
<td>Details</td>
<td>7</td>
<td>100</td>
<td>hsa-miR-29b-3p</td>
<td>FBN1</td>
<td>fibrillin 1</td>
</tr>
<tr>
<td>Details</td>
<td>8</td>
<td>99</td>
<td>hsa-miR-29b-3p</td>
<td>TET1</td>
<td>tet methylcytosine dioxygenase 1</td>
</tr>
<tr>
<td>Details</td>
<td>9</td>
<td>99</td>
<td>hsa-miR-29b-3p</td>
<td>ATAD2B</td>
<td>ATPase family, AAA domain containing 2B</td>
</tr>
<tr>
<td>Details</td>
<td>10</td>
<td>99</td>
<td>hsa-miR-29b-3p</td>
<td>RNF19A</td>
<td>ring finger protein 19A, RBR E3 ubiquitin protein ligase</td>
</tr>
<tr>
<td>Details</td>
<td>11</td>
<td>99</td>
<td>hsa-miR-29b-3p</td>
<td>VEGFA</td>
<td>vascular endothelial growth factor A</td>
</tr>
<tr>
<td>Details</td>
<td>12</td>
<td>99</td>
<td>hsa-miR-29b-3p</td>
<td>DNMT3A</td>
<td>DNA (cytosine-5')-methyltransferase 3 alpha</td>
</tr>
<tr>
<td>Details</td>
<td>13</td>
<td>99</td>
<td>hsa-miR-29b-3p</td>
<td>FBXW9</td>
<td>F-box and WD repeat domain containing 9</td>
</tr>
</tbody>
</table>
iPathway Guide can deduce a series of miRNA changes likely to have occurred for a given set of RNA expression changes.

---

MicroRNA sequencing (miRNAsseq)
Figure 4: Inference of perturbed oncogenic miRNA regulatory network in colorectal cancer. (A) ROC curves of predicting CRC-related miRNAs using different methods. (B) The perturbed key miRNA regulatory sub-network in CRC. Genes are colored on the basis of their expression fold change. (C) Enrichment results of the 338 LE target genes in KEGG pathways. The number of LE genes included in each pathway is shown beside the bar.
Conclusions

The Bioinformatics core can provide you with a robust analysis of a microRNA sequencing experiment including:

- Mature/precursor/novel miRNA qualification
- Variant detection in miRNA coding region
- Flexible differential expression between experimental conditions

AND provide you with access to systems biology analysis of your results. Dr. Sean Courtney will review this topic in his talk.