RNA-Seq Module 3
Advanced RNA-Seq Analysis Topics and Trouble-Shooting

Kevin Silverstein PhD, John Garbe PhD and Ying Zhang PhD,
Research Informatics Support System (RISS)
MSI
May 24, 2012
RNA-Seq Tutorials

• Tutorial 1: Introductory (Mar. 28 & Apr. 19)
  – RNA-Seq experiment design and analysis
  – Instruction on individual software will be provided in other tutorials

• Tutorial 2: Introductory (Apr. 3 & Apr 24)
  – Analysis RNA-Seq using TopHat and Cufflinks

• Tutorial 3: Intermediate (May 24)
  – Advanced RNA-Seq analysis topics and troubleshooting

• Hands-on tutorials (Summer 2012)…
RNA-Seq Module 3
Advanced RNA-Seq Analysis Topics and Trouble-Shooting

**Part I:** Review and Considerations for Different Goals and Biological Systems (Kevin Silverstein, PhD)

**Part II:** Read Mapping Statistics and Visualization (John Garbe, PhD)

**Part III:** Post-Analysis Processing – Exploring the Data and Results (Ying Zhang, PhD)
Part I

Review and Considerations for Different Goals and Biological Systems

Kevin Silverstein, PhD
Typical RNA-seq experimental protocol and analysis

Sample

↓

mRNA isolation

↓

RNA -> cDNA

→

Fragmentation

↓

Paired End (PE) Sequencing

→

Map reads

Genome

Reference Transcriptome

A

B
Steps in RNA-Seq data analysis depend on your goals and biological system.

1. **Quality Control**
2. **Data prepping**
3. **Map Reads to Reference Genome/Transcriptome**
4. **Assemble Transcriptome**
   - **Other applications:** De novo Assembly, Refine gene models
   - **Identify Differentially Expressed Genes**
   - “microarray simulation”
   - Discovery mode
Programs used in RNA-Seq data analysis depend on your goals and biological system

**Step 1:** Quality Control
- FastQC

**Step 2:** Data prepping
- Filter/Trimmer/Converter
- TopHat, GSNAP

**Step 3:** Map Reads to Reference Genome/Transcriptome
- Cufflinks, Cuffmerge

**Step 4:** Assemble Transcriptome
- Identify Differentially Expressed Genes
- Cuffdiff
- IGV

Other applications:
- Refine gene models
Programs used in RNA-Seq data analysis depend on your goals and biological system

Step 1: Quality Control
- FastQC

Step 2: Data prepping
- Filter/Trimmer/Converter
- Bowtie, BWA

Step 3: Map Reads to Reference Genome/Transcriptome
- Cufflinks, Cuffmerge
- Assemble Transcriptome

Step 4: Other applications:
- De novo Assembly
- Refine gene models
- Identify Differentially Expressed Genes
- Cuffdiff

Glimmer
Artemis
Trinity, Velvet
Visualizing microbial data in Artemis

Programs used in RNA-Seq data analysis depend on your goals and biological system.

**Step 1:**
- Quality Control
  - FastQC

**Step 2:**
- Data prepping
  - Filter/Trimmer/Converter
  - TopHat, GSNAP

**Step 3:**
- Map Reads to Reference Genome/Transcriptome
  - Cufflinks, Cuffmerge

**Step 4:**
- Assemble Transcriptome
  - GeneMark, FGeneSH
  - Trinity, TransABySS
  - BLAT

Other applications:
- De novo Assembly
- Refine gene models
- Identify Differentially Expressed Genes
- Cuffdiff

**University of Minnesota**
Driven to Discover
Programs used in RNA-Seq data analysis depend on your goals and biological system.

Step 1: Quality Control
- FastQC

Step 2: Data prepping
- Filter/Trimmer/Converter

Step 3: Assemble Transcriptome
- Trinity, TransABySS
- Other applications: De novo Assembly, Refine gene models
- Cufflinks, Cuffmerge
- Identify Differentially Expressed Genes
- Cuffdiff
Programs used in RNA-Seq data analysis depend on your goals and biological system

Step 1: Quality Control

Step 2: Data prepping

Step 3: Map Reads to Reference Genome/Transcriptome

Step 4: Assemble Transcriptome

Other applications:
- De novo Assembly
- Refine gene models

Identify Differentially Expressed Genes

GeneMark, FGeneSH, Trinity, TransABySS, BLAT

FastQC

Filter/Trimmer/Converter

TopHat, GSNAP

Cufflinks, Cuffmerge

Cuffdiff

Consider sequencing and assembling the genome first if not too big…
Library construction and sequencing design decisions
Library type (SE/PE) and insert size

Sample

\[ \text{mRNA isolation} \]

\[ \text{Fragmentation} \]

\[ \text{Library preparation} \]

\[ \text{Size: 200-500 bp} \]

\[ \text{Sequence fragment end(s)} \]

\[ \text{SE sequencing} \]

\[ \text{PE sequencing} \]
Library type (Mate-pair) and insert size

Sample → mRNA isolation → Fragmentation → Library preparation → Circulation → Fragmentation → Sequence fragment end(s) → Mate-Pair sequencing

Size: 2000-8000 bp
Optimal library size depends on goals and organism: *exon size*

**Arabidopsis exon size distribution**

- **Mode:** 100 bp
- **Median:** 150 bp
- **Mean:** 300 bp
Optimal library size depends on goals and organism: \textit{exon size}

Adjacent connectivity \hspace{1cm} Insert size = exon size

Minimal connectivity \hspace{1cm} Insert size << exon size

Long-range connectivity \hspace{1cm} Insert size >> exon size

One size doesn’t fit all: organisms can differ in exon size distribution
How does connectivity play into the analysis?

1. splice-align reads to the genome

2. Build a graph representing alternative splicing events

3. Traverse the graph to assemble variants

4. Assemble isoforms

Some algorithms (e.g., tophat) exhaustively look for candidate splices in a specified distance pegged to the expected intron size distribution (default 70-500,000)
Why not just leave the defaults? (e.g., 70-500,000 bp)

- ~3500 Arabidopsis introns < 70 bp
- Huge increase in computation time
- Will accumulate spurious long-range splice junctions
Many plant genomes have undergone ancient Whole Genome Duplications (WGDs)

- Difficulty mapping uniquely to related gene family members
- Abundance levels (e.g., FPKMs) can become skewed for members of large gene families
- Both PE strategies and longer reads help to distinguish paralogs

http://genomevolution.org
Some genomes are rife with repetitive elements

- 50%, 65% of the human and maize genome are repeat elements, respectively (repbase, Kronmiller et al., Plant Phys 2008; )
- PE, mate-pair strategies and multiple insert sizes help to uniquely map repeats
- Long reads can help for small-scale or simple repeats
Why is PE crucial for repetitive genomes and those with paralogous gene families?

2 x 50 bp is better than 1 X 100 bp for most applications and systems.
Sequencing depth needed depends on transcriptome size and the project goals

- **Sequencing Depth** is the average read coverage of target sequences
  - Sequencing depth = total number of reads $\times$ read length / estimated target sequence length
  - Example, for a 5MB transcriptome, if 1 Million 50 bp reads are produced, the depth is $1 \text{ M} \times 50 \text{ bp} / 5\text{M} \sim 10 \text{X}$

- Average coverage may be misleading, since expression levels can vary more than 5 orders of magnitude!

- Differential expression requires less depth than assembly, gene model refinement and structural variant discovery.
Polyploidy is particularly problematic

- Difficult to distinguish alleles from paralogs
- Genome assembly often intractable
- Need care in design of transcriptome experiment
Certain applications and biological systems will require special design considerations for maximal resolution

- Polyploid genomes may require long reads, multiple insert sizes and custom software to distinguish among highly similar alleles at each locus.
- Ditto for those who wish to interrogate allele-specific differential expression (e.g., maternal or paternal imprinting).
# Genome size characteristics (iGenomes)

<table>
<thead>
<tr>
<th>Species</th>
<th>Number of genes</th>
<th>Transcriptome size (Mbp)</th>
<th>Mode</th>
<th>Avg exon size</th>
<th>Intron size range (1%</th>
<th>99%)</th>
<th>% genome repetitive</th>
<th>% genes in families*</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Homo sapiens</em></td>
<td>29230</td>
<td>70.1</td>
<td>100</td>
<td>300</td>
<td>77</td>
<td>107000</td>
<td>47</td>
<td>20</td>
</tr>
<tr>
<td><em>Mus musculus</em></td>
<td>24080</td>
<td>61.4</td>
<td>100</td>
<td>300</td>
<td>78</td>
<td>100000</td>
<td>44</td>
<td>NA</td>
</tr>
<tr>
<td><em>Gallus gallus</em>*</td>
<td>4906</td>
<td>11.1</td>
<td>100</td>
<td>230</td>
<td>73</td>
<td>120000</td>
<td>10</td>
<td>NA</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>18436</td>
<td>30.1</td>
<td>150</td>
<td>450</td>
<td>30</td>
<td>25000</td>
<td>32</td>
<td>7</td>
</tr>
<tr>
<td><em>Caenorhabditis elegans</em></td>
<td>23933</td>
<td>28.0</td>
<td>110</td>
<td>220</td>
<td>43</td>
<td>8000</td>
<td>4</td>
<td>24</td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>27278</td>
<td>51.1</td>
<td>70</td>
<td>300</td>
<td>46</td>
<td>4900</td>
<td>9</td>
<td>35</td>
</tr>
<tr>
<td><em>Saccharomyces cerevisiae</em></td>
<td>6692</td>
<td>8.9</td>
<td>75</td>
<td>1200</td>
<td>20</td>
<td>2600</td>
<td>1</td>
<td>36</td>
</tr>
<tr>
<td><em>Escherichia coli</em>**</td>
<td>4290</td>
<td>0.6</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>3</td>
<td>52</td>
<td></td>
</tr>
</tbody>
</table>

* % genes with at least one paralog in the COG database ( unicellular) or included in the COG lineage specific expansion (LSE) list. (These percentages are likely systematic underestimates)
** Poor annotation is suspected for iGenomes UCSC-based Gallus gallus (galGal3)
*** http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/E/Esch.coli.html; ecocyc; Gur-Arie, Genome Res 2000;
Summary of Library Construction and Sequencing Decisions

<table>
<thead>
<tr>
<th>Project Goals:</th>
<th>De novo Assembly of transcriptome</th>
<th>Refine gene model</th>
<th>Differential Gene Expression</th>
<th>Identification of structural variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Library Type:</td>
<td>PE, Mated PE</td>
<td>PE, SE</td>
<td>PE</td>
<td>PE, Mated PE</td>
</tr>
<tr>
<td>Sequencing Depth:</td>
<td>Extensive (&gt; 50 X)</td>
<td>Extensive</td>
<td>Moderate (10 X ~ 30 X)</td>
<td>Extensive</td>
</tr>
</tbody>
</table>

- SE may be OK for (3) DGE if you have a good annotation and a simple genome.
- Strand-specific library creation may be necessary for organisms with a large percentage of genes that overlap on opposite strands (e.g. bacteria, yeast), or if you’re interested in antisense regulation.
### Sample Replicates and Pooling Decisions

<table>
<thead>
<tr>
<th>Project Goals</th>
<th>De novo Assembly of transcriptome</th>
<th>Refine gene model</th>
<th>Differential Gene Expression</th>
<th>Identification of structural variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pooling OK?</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes, for discovery</td>
</tr>
<tr>
<td>Biological Replicates?</td>
<td>Yes</td>
<td>Yes, if not pooling</td>
<td>Yes</td>
<td>Yes, if not pooling</td>
</tr>
</tbody>
</table>

- Pooling may be advisable if RNA is limited or if not interested in biological variability.
- As a general rule, the following biological replicates are advisable for DGE:
  - 3+ for cell lines and pooled samples
  - 5+ for inbred lines (e.g., BL6 mice, NILs, RILs)
  - 20+ for human samples
Part II

Read Mapping Statistics and Visualization

John Garbe, PhD
Mapping Statistics

How well did my sequence library align to my reference?
Mapping Statistics

- **Mapping Output**
  - SAM (text) / BAM (binary) alignment files
  - Summary statistics (per read library)
    - % reads with unique alignment
    - % reads with multiple alignments
    - % reads with no alignment
    - % reads properly paired (for paired-end libraries)
    - Mean and standard deviation of insert size

Mapping Statistics

- SAM Tools
- Picard
- Tophatstats
Mapping Statistics – SAMtools

- Galaxy
  - NGS: SAM Tools -> flagstat
- MSI Command line
  - Module load samtools
  - samtools flagstat accepted_hits.bam
Mapping Statistics – SAMtools

• SAMtools output

```bash
% samtools flagstat accepted_hits.bam
31443374 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 duplicates
31443374 + 0 mapped (100.00%:-nan%)
31443374 + 0 paired in sequencing
15771038 + 0 read1
15672336 + 0 read2
15312224 + 0 properly paired (48.70%:-nan%)
29452830 + 0 with itself and mate mapped
1990544 + 0 singletons (6.33%:-nan%)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```
Mapping Statistics – Picard

• Galaxy
  – NGS: Picard (beta) -> SAM/BAM Alignment Summary Metrics

• Command line:
  – module load picard-tools
  – java -Xmx2g -jar CollectAlignmentSummaryMetrics.jar
    INPUT=accepted_hits.bam OUTPUT=stats.txt
Mapping Statistics – Picard

- Picard output

<table>
<thead>
<tr>
<th>CATEGORY</th>
<th>TOTAL_READS</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIRST_OF_PAIR</td>
<td>14739626</td>
</tr>
<tr>
<td>SECOND_OF_PAIR</td>
<td>14653925</td>
</tr>
<tr>
<td>PAIR</td>
<td>29393551</td>
</tr>
</tbody>
</table>
Mapping Statistics – tophatstats

- Galaxy
  - MSI -> tophatstats

- Command line
  - module load tophatstats
Mapping Statistics – tophatstats

- Tophatstats output (paired-end reads)

```bash
% tophatstats.pl accepted_hits.bam L1_R1_sample1.fastq
Input files: accepted_hits.bam L1_R1_sample1.fastq
250000 total read pairs in fastq file
120004 (48.00%) read pairs mapped with correct insert size
   (116869 with unique alignments)
50536 (20.21%) read pairs mapped with wrong insert size
   (49351 with unique alignments)
24368 (9.75%) read pairs with only one read in the pair mapped
   (23544 with unique alignments)
55092 (22.04%) read pairs with no mapping
60.13bp average inner distance between read pairs
```
Mapping Visualization

• Integrative Genomics Viewer (IGV)
  – Fast genome browser
  – Supports array-based and next-generation sequence data, and genomic annotations
  – Free Java program

http://www.broadinstitute.org/igv/home
Mapping Visualization

Bam file viewed with IGV
Causes of poor mapping

- Improper alignment parameters
- Poor quality sequence library
- Contaminated sequence library
- Poor quality reference
- Repetitive genome
- Divergence between sequenced population and reference
- Mislabeled samples
- Corrupted files
- Short read length
- Poor choice of mapping software
- Bug in mapping software
- ...
Poor Quality Library

Quality scores across all bases (Illumina 1.5 encoding)

Good

 Poor quality read library decreases mapping performance

Bad

Trimming needed
Bug in software

<table>
<thead>
<tr>
<th>Tophat 2.0.0</th>
<th>Tophat 2.0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>35% mapped, properly paired</td>
<td>48% mapped, properly paired</td>
</tr>
<tr>
<td>33% mapped, wrong insert size</td>
<td>20%</td>
</tr>
<tr>
<td>10% singleton</td>
<td>9% singleton</td>
</tr>
<tr>
<td>22% no mapping</td>
<td>22% no mapping</td>
</tr>
</tbody>
</table>

New “bugfix” release of Tophat improves mapping performance
## Poor Quality Reference

<table>
<thead>
<tr>
<th>Sus scrofa 9.2</th>
<th>Sus scrofa 10.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>46%</td>
<td>48% mapped, properly paired</td>
</tr>
<tr>
<td>17%</td>
<td>20% mapped, wrong insert size</td>
</tr>
<tr>
<td>9%</td>
<td>9% singleton</td>
</tr>
<tr>
<td>26%</td>
<td>22% no mapping</td>
</tr>
</tbody>
</table>

Mapping performance improves due to improvement in Pig genome build
Large and small sequence divergence between two human samples and the human reference genome
Contaminated sequence library

### Overrepresented sequences

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Count</th>
<th>Percentage</th>
<th>Possible Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTATTACAGA7CGGAGACCCGCATGCAGAATCCCGAGACCAGACCGCTCCG</td>
<td>920428</td>
<td>2.8366639370528275</td>
<td>Illumina Paired End PCR Primer 2 (100% over 43bp)</td>
</tr>
<tr>
<td>GTATCAAGA7CGGAGACCCGCATGCAGAATCCCGAGACCAGACCGCTCCG</td>
<td>749728</td>
<td>2.5922157461699773</td>
<td>Illumina Paired End PCR Primer 2 (100% over 44bp)</td>
</tr>
<tr>
<td>CGTTCAAGGAGA7CCCGAGATCAGAAGAGGCTTACCGAGAATCCCGAGACCAGACCGCTCCG</td>
<td>648852</td>
<td>2.243432780066747</td>
<td>Illumina Paired End Adapter 2 (100% over 31bp)</td>
</tr>
<tr>
<td>CATTCAAGA7CGGAGACCCGCATGCAGAATCCCGAGACCAGACCGCTCCG</td>
<td>176765</td>
<td>0.611172340330748</td>
<td>Illumina Paired End PCR Primer 2 (97% over 36bp)</td>
</tr>
<tr>
<td>ACCTGAGA7CGGAGACCCGCATGCAGAATCCCGAGACCAGACCGCTCCG</td>
<td>143840</td>
<td>0.4973327302615156</td>
<td>Illumina Paired End PCR Primer 2 (100% over 43bp)</td>
</tr>
<tr>
<td>GTATCAAGA7CGGAGACCCGCATGCAGAATCCCGAGACCAGACCGCTCCG</td>
<td>124281</td>
<td>0.4297067271727257</td>
<td>Illumina Paired End PCR Primer 2 (100% over 44bp)</td>
</tr>
<tr>
<td>GTATCAAGA7CGGAGACCCGCATGCAGAATCCCGAGACCAGACCGCTCCG</td>
<td>99207</td>
<td>0.34301232917842867</td>
<td>Illumina Paired End PCR Primer 2 (100% over 45bp)</td>
</tr>
<tr>
<td>ACCTGAGA7CGGAGACCCGCATGCAGAATCCCGAGACCAGACCGCTCCG</td>
<td>96289</td>
<td>0.332923222279941655</td>
<td>Illumina Paired End PCR Primer 2 (100% over 50bp)</td>
</tr>
<tr>
<td>CGGAAGAGATTACGAGAATCCCGAGACCAGACCGCTCCG</td>
<td>93842</td>
<td>0.3244626105124245</td>
<td>Illumina Paired End PCR Primer 2 (96% over 33bp)</td>
</tr>
<tr>
<td>CGTTCAGA7CGGAGACCCGCATGCAGAATCCCGAGACCAGACCGCTCCG</td>
<td>75370</td>
<td>0.26059491013918545</td>
<td>Illumina Paired End PCR Primer 2 (100% over 43bp)</td>
</tr>
<tr>
<td>CGTTCAGA7CGGAGACCCGCATGCAGAATCCCGAGACCAGACCGCTCCG</td>
<td>63691</td>
<td>0.22021428183196043</td>
<td>Illumina Paired End PCR Primer 2 (100% over 44bp)</td>
</tr>
<tr>
<td>ACCTGAGA7CGGAGACCCGCATGCAGAATCCCGAGACCAGACCGCTCCG</td>
<td>56765</td>
<td>0.19626734873359242</td>
<td>Illumina Paired End PCR Primer 2 (100% over 46bp)</td>
</tr>
<tr>
<td>TAATGAGA7CGGAGACCCGCATGCAGAATCCCGAGACCAGACCGCTCCG</td>
<td>42991</td>
<td>0.14064317070139472</td>
<td>Illumina Paired End PCR Primer 2 (100% over 43bp)</td>
</tr>
</tbody>
</table>

FastQC output showing ~10% adapter contamination
Poor choice of mapping software

- BWA (not splice aware)
- GSNAP (splice aware)
## Improper alignment parameters

<table>
<thead>
<tr>
<th>Correct inner distance (60)</th>
<th>Incorrect inner distance (220)</th>
</tr>
</thead>
<tbody>
<tr>
<td>48%</td>
<td>43%</td>
</tr>
<tr>
<td>mapped, properly paired</td>
<td>mapped, wrong insert size</td>
</tr>
<tr>
<td>20%</td>
<td>25%</td>
</tr>
<tr>
<td>singleton</td>
<td>no mapping</td>
</tr>
<tr>
<td>9%</td>
<td>10%</td>
</tr>
<tr>
<td>22%</td>
<td>22%</td>
</tr>
</tbody>
</table>

Incorrect “inner mate pair distance” parameter decreases mapping performance
Corrupted files

<table>
<thead>
<tr>
<th>Correct fastq file</th>
<th>Corrupted fastq file</th>
</tr>
</thead>
<tbody>
<tr>
<td>48%</td>
<td>22% mapped, properly paired</td>
</tr>
<tr>
<td>20%</td>
<td>46% mapped, wrong insert size</td>
</tr>
<tr>
<td>9%</td>
<td>10% singleton</td>
</tr>
<tr>
<td>22%</td>
<td>22% no mapping</td>
</tr>
</tbody>
</table>

Unsynchronized paired-end fastq file decreases percentage of properly-paired reads
Part III

Post-Analysis Processing - Exploring the Data and Results

Ying Zhang, PhD
Workflow of a typical NGS project

Step 1: Design NGS experiment
1. What are my goals?
2. What are the characteristics of my system?

Step 2: Sequencing Process
1. Illumina / SOLiD / 454 / Ion Torrent / Helicos

Step 3: Quality control And Data Prepping
1. Sequence contamination
2. Low quality reads
3. Trimming / Filtering / Synchronization

Step 4: In silico Analysis
1. Which program to use?
2. How to measure the performance?

Step 5: Post-Analysis Processing
1. Explore results using empirical cases
2. Statistical checks for data grouping behavior
3. Pathway analysis and system biology

University of Minnesota
Driven to Discover™
Widely-used Tools in Data Exploring

• Direct visualization of “positive controls”:
  – IGV viewer
  – UCSC Genome Browser

• Statistical checks of data structure:
  – PCA: principle component analysis
  – MDS: multi-dimension scaling
  – Unsupervised clustering and Heatmap

• System-level Analysis:
  – IPA: ingenuity pathway analysis
Integrative Genomics Viewer (IGV)

- Fast genome browser
- Supports array-based and next-generation sequence data, and genomic annotations
- Free Java program

Launch:
- From Galaxy
- From Desktop: allocate enough memory

http://www.broadinstitute.org/igv/home
No. 1 in your Check-List

“Are my data behaving as expected?”
Exploring results using Empirical Cases – Example I: no reads mapped at knock-out site

Data Courtesy of Dr. Mike Farrar and Dr. Lynn Harris (unpublished data)
Example II: Housekeeping genes should behave similarity across multiple samples

Data Courtesy of Dr. David Bernlohr and Dr. Ann Hertzel (unpublished data)
Example III: review of known biomarkers, for example, known SNP and indel

Heterozygous deletion of ‘T’ with 46% penetrance

Data Courtesy of Dr. John Ohlfest and Dr. Flavia Popescu (unpublished data)
Example IV: detect the caveat of programs

Data courtesy of Dr. Steve Gantt and Dr. Karen Tang (unpublished data)
Cufflinks developer: “We don’t recommend assembling bacteria transcripts using Cufflinks at first. If you are working on a new bacteria genome, consider a computational gene finding application such as Glimmer.”

So for bacteria transcriptome:

• If the genome is available, do genome annotation first then reconstruct the transcriptome.

• If the genome is not available, try de novo assembly of the transcriptome, followed by gene annotation.
Explore the global distribution of data

Many genes will have little or no expression.

A set of genes have a high expression.

Very few genes have an usually high expression.
Exclude the highly-expressed genes for highly-unbalanced expression between conditions. Set “yes” to “Perform quartile normalization”.

Example: red cell blood compared to other tissue
Warning: don’t throw the baby with the bathwater…

**Cuffdiff:** “Min Alignment Count” must be satisfied in all samples – too high a value will remove genes not expressed in one condition but strongly expressed in another!

This gene was reported as DE with “Min Alignment Count” = 10, but not with 100.
Biological replicates should show grouping behavior in multi-variable analysis:

- innate consistence between samples

**Statistical Checks of data structure – Multi-Variable Analysis**
Within-group variation: non-biological variations

- **Source of non-biological variation:**
  - **Batch effect**
    - How were the samples collected and processed? Were the samples processed as groups, and if so what was the grouping?
  - **Non-synchronized cell cultures**
    - Were all the cells from the same genetic backgrounds and growth phase?
  - Use technical replicates rather than biological replicates
How to check for data variation?

- **Principle Component Analysis (PCA)**
  - Uses an orthogonal transformation
  - The first principle component has the largest possible variance

- **Multi-Dimensional Scaling (MDS)**
  - Computes euclidean distances among all pairs of samples

- **Unsupervised Clustering / heatmap**
  - Identify the hidden structure in “unlabeled” data

- **Tools:**
  - Galaxy
  - Statistical Package: R, SPSS, MatLab
  - Partek and Genedata Expressionist
Steps in PCA analysis

1. Construct the multiple variable matrix

<table>
<thead>
<tr>
<th>transcript</th>
<th>Sample A</th>
<th>Sample V</th>
<th>Sample O</th>
<th>Sample E</th>
<th>Sample I</th>
<th>Sample U</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene1</td>
<td>6.18</td>
<td>6.64</td>
<td>6.46</td>
<td>6.30</td>
<td>6.58</td>
<td>6.54</td>
</tr>
<tr>
<td>gene2</td>
<td>5.48</td>
<td>0.11</td>
<td>1.00</td>
<td>0.24</td>
<td>0.02</td>
<td>0.68</td>
</tr>
<tr>
<td>gene3</td>
<td>20.53</td>
<td>18.93</td>
<td>18.79</td>
<td>18.51</td>
<td>18.00</td>
<td>18.26</td>
</tr>
<tr>
<td>gene4</td>
<td>55.47</td>
<td>52.71</td>
<td>50.39</td>
<td>54.66</td>
<td>49.15</td>
<td>44.68</td>
</tr>
<tr>
<td>gene5</td>
<td>7.28</td>
<td>8.09</td>
<td>8.57</td>
<td>7.82</td>
<td>8.29</td>
<td>9.38</td>
</tr>
<tr>
<td>gene8</td>
<td>6.17</td>
<td>6.79</td>
<td>7.20</td>
<td>6.70</td>
<td>8.42</td>
<td>7.26</td>
</tr>
<tr>
<td>gene9</td>
<td>25.83</td>
<td>24.24</td>
<td>25.63</td>
<td>27.09</td>
<td>22.18</td>
<td>23.09</td>
</tr>
<tr>
<td>gene10</td>
<td>38.04</td>
<td>30.39</td>
<td>35.53</td>
<td>37.42</td>
<td>28.72</td>
<td>27.28</td>
</tr>
<tr>
<td>gene11</td>
<td>195.06</td>
<td>179.88</td>
<td>178.18</td>
<td>208.25</td>
<td>179.01</td>
<td>155.15</td>
</tr>
<tr>
<td>gene12</td>
<td>32.82</td>
<td>32.04</td>
<td>31.84</td>
<td>33.62</td>
<td>31.06</td>
<td>29.46</td>
</tr>
<tr>
<td>gene13</td>
<td>18.41</td>
<td>16.75</td>
<td>16.72</td>
<td>17.33</td>
<td>16.32</td>
<td>16.87</td>
</tr>
<tr>
<td>gene14</td>
<td>24.00</td>
<td>21.05</td>
<td>22.68</td>
<td>22.72</td>
<td>22.08</td>
<td>22.45</td>
</tr>
</tbody>
</table>

Group 1 (A,V,O)  
Group 2 (E,I,U)

e.g. tables of FPKM values

2. Run PCA analysis and explore the result
Heatmap: Unsupervised clustering

1. Construct the multiple variable matrix
2. Run Unsupervised Clustering and generate Heatmap

e.g. tables of FPKM values

<table>
<thead>
<tr>
<th>Transcript</th>
<th>Sample A</th>
<th>Sample V</th>
<th>Sample O</th>
<th>Sample E</th>
<th>Sample I</th>
<th>Sample U</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene1</td>
<td>6.18</td>
<td>6.64</td>
<td>6.46</td>
<td>6.30</td>
<td>6.58</td>
<td>6.54</td>
</tr>
<tr>
<td>gene2</td>
<td>5.48</td>
<td>0.11</td>
<td>1.00</td>
<td>0.24</td>
<td>0.02</td>
<td>0.68</td>
</tr>
<tr>
<td>gene3</td>
<td>20.53</td>
<td>18.93</td>
<td>18.79</td>
<td>18.51</td>
<td>18.00</td>
<td>18.26</td>
</tr>
<tr>
<td>gene4</td>
<td>55.47</td>
<td>52.71</td>
<td>50.39</td>
<td>54.66</td>
<td>49.15</td>
<td>44.68</td>
</tr>
<tr>
<td>gene5</td>
<td>7.28</td>
<td>8.09</td>
<td>8.57</td>
<td>7.82</td>
<td>8.29</td>
<td>9.38</td>
</tr>
<tr>
<td>gene8</td>
<td>6.17</td>
<td>6.79</td>
<td>7.20</td>
<td>6.70</td>
<td>8.42</td>
<td>7.26</td>
</tr>
<tr>
<td>gene9</td>
<td>25.83</td>
<td>24.24</td>
<td>25.63</td>
<td>27.09</td>
<td>22.18</td>
<td>23.09</td>
</tr>
<tr>
<td>gene10</td>
<td>38.04</td>
<td>30.39</td>
<td>35.53</td>
<td>37.42</td>
<td>28.72</td>
<td>27.28</td>
</tr>
<tr>
<td>gene11</td>
<td>195.06</td>
<td>179.88</td>
<td>178.18</td>
<td>208.25</td>
<td>179.01</td>
<td>155.15</td>
</tr>
<tr>
<td>gene12</td>
<td>32.82</td>
<td>32.04</td>
<td>31.84</td>
<td>33.62</td>
<td>31.06</td>
<td>29.46</td>
</tr>
<tr>
<td>gene13</td>
<td>18.41</td>
<td>16.75</td>
<td>16.72</td>
<td>17.33</td>
<td>16.32</td>
<td>16.87</td>
</tr>
<tr>
<td>gene14</td>
<td>24.00</td>
<td>21.05</td>
<td>22.68</td>
<td>22.72</td>
<td>22.08</td>
<td>22.45</td>
</tr>
</tbody>
</table>

Group 1 (A, V, O)
Group 2 (E, I, U)
Exploring data at system-level: Ingenuity Pathway analysis

• Using the differentially expressed genes
• Connecting the genes with known knowledge
• Testing for the significance of the identified network
• Check the details at:
Pro-angiogenic Genes and microRNA deregulated in Ovarian Cancer

- microRNA up/down regulated in ovarian cancer
- Genes overexpressed in ovarian cancer
- Drugs
Discussion and Questions?

• Get Support at MSI:
  – Email: help@msi.umn.edu
  – General Questions:
    • Subject line: “RISS:…”
  – Galaxy Questions:
    • Subject line: “Galaxy:…”