Gene Expression

Genomics: Bio5488
Spring 2020
Recap ...

• Low-throughput methods
• High-throughput methods
• Experimental Design
  • What is your question?
  • Replicates?
  • Power?
• RNA isolation
• Quality Control
• Normalization
• Clustering
Lecture 1:
• Measuring Gene Expression
  • Targeted
    • Northern
    • rtPCR
    • *In situ* Hybridization
    • Reporter Assay
  • High-throughput
    • Microarray

Lecture 2:
• Transcriptomics
  • RNAseq
Measuring Gene Expression: RNAseq

Measuring Gene Expression: RNAseq

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Very direct and quantitative</td>
<td>• Amplification steps can offset balance between high/low abundance transcripts</td>
</tr>
<tr>
<td>• No prior knowledge of genome required</td>
<td>• Analysis is non-trivial - software still evolving</td>
</tr>
<tr>
<td>• Discriminates among regions with high sequence identity</td>
<td></td>
</tr>
</tbody>
</table>
Experimental Design

• “Seventy percent of whether your experiment will work is determined before you touch the first test tube ... “

Experimental Design: Replication!

- “Almost 70% of all the human RNAseq samples in GEO do not have biological replicates”


Experimental Design: Replication!

<table>
<thead>
<tr>
<th>Replicates per group</th>
<th>3</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effect size (fold change)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.25</td>
<td>17 %</td>
<td>25 %</td>
<td>44 %</td>
</tr>
<tr>
<td>1.5</td>
<td>43 %</td>
<td>64 %</td>
<td>91 %</td>
</tr>
<tr>
<td>2</td>
<td>87 %</td>
<td>98 %</td>
<td>100 %</td>
</tr>
<tr>
<td>Sequencing depth (millions of reads)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>19 %</td>
<td>29 %</td>
<td>52 %</td>
</tr>
<tr>
<td>10</td>
<td>33 %</td>
<td>51 %</td>
<td>80 %</td>
</tr>
<tr>
<td>15</td>
<td>38 %</td>
<td>57 %</td>
<td>85 %</td>
</tr>
</tbody>
</table>

Statistical power to detect differential expression varies with effect size, sequencing depth and number of replicates.
Experimental Design: Read Depth (Library Size)

- Different sequencing goals require different read depths
- More genes are detected with higher sequencing depth
- Can always increase depth by additional sequencing

Tarazona et al., Genome Research (2011)
Experimental Design: Power!

An unbiased way to reduce uncertainty is to add more information ...
Genes Detected as a Function of Sequencing Depth

When 10 million reads were sequenced, about 12K genes were measured by at least 10 reads in the sample in Male 3.
Poisson Variance

Poisson distributions show the uncertainty present in a 2X fold change using count data at different counts.
Biological, Technical and Poisson Variance

- **Biological Variance** is variance that naturally occurs within the samples under investigation.

- **Technical Variance** is measurement imprecision that stems from the inability of RNA-Seq measurements to measure expression perfectly.

- **Poisson Variance** is the inherent uncertainty that is present in any measurement made where something is sampled and counted.
RNAseq: Basic Data Generation Steps

1. Isolate total RNA
2. Remove DNA contamination
3. RNA is broken into short fragments
4. RNA fragments are reverse transcribed
5. Sequencing adaptors are ligated
6. Fragment size selection
7. Ends of the cDNAs are sequenced to produce many short reads

RNA Isolation: Quality Matters!

Sigurgeirsson et al., *PLOS One* (2014)
RNA Isolation: Quality Matters!

Gene body coverage

Number of reads (millions)

0 1 2 3

0 20 40 60 80 100

5' --- gene body --- 3'

(%)

- RIN 10 (49.1%)
- RIN 8 (54.3%)
- RIN 6 (58.1%)
- RIN 4 (62.2%)
- RIN 2 (80.1%)
- RiboMinus (52.3%)

Sigurgeirsson et al., PLOS One (2014)
Remove DNA Contamination

- Minimize genomic DNA aligning back to the genome
How to Remove rRNA?

• Ribosomal RNA typically constitutes >90% of total RNA
  • 1-2% mRNA

• Enrich for mRNA?
  • Poly-A selection
    • Requires higher amount of starting material
    • Requires minimal degradation (RIN)

• Deplete rRNA?
  • Ribo-minus
    • When you cannot obtain great enough quantity or good enough integrity
Capture poly-A RNA with poly-T oligo attached beads

- RNA quality must be high – degradation produces 3’ bias
- Non-poly-A RNAs are not recovered

Fragment mRNA

RNA fragments

Synthesize ds cDNA
Ligate adapters
Amplify

Generate clusters and sequence
Ribosomal RNA Subtraction

1. Total RNA
2. Hybridize with Biotin-LNA probe set
3. Capture RNA/probe Complexes
4. Remove RNA/probe Complexes w/ RiboMinus™ beads

Whole Transcriptome
• Single-end sequencing?
  • Less expensive
  • Sufficient for well-annotated organisms

• Paired-end?
  • Better for poorly annotated transcriptomes
  • Preferable for de novo transcript discovery and isoform analysis

Zhernakova et al., PLOS Genetics (2013)
Experimental Design: Read Length

- Longer read length
  - Better ability to assemble unknown transcripts
  - Higher accuracy in complex regions (e.g. repetitive, polymorphic)
  - Splice junction detection
  - Gives minimal to no advantage for differential expression analysis
Sequencing Design

• Spike-In?
  • Technical control
    • Spiking in a known amount of phiX should yield a known amount of phiX clusters
    • Alignment to the phiX reference can be used to calculate the sequencing error rate

• Randomization?
  • Library prep
  • Sequencing run
Each read contains four lines:

1. @SEQ_ID
2. GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGT TCAACTCACAAGTTTT
3. +
4. !"*(((***+))%%%++)(%%%%).1***- +**)))**55CCF>>>>>>>>CCCCCCCC65
Quality Control: FastQC

Per base sequence quality

Good

Not so Good
Quality Control: FastQC

**Per sequence quality scores**

- **Good**
- **Not so Good**
Quality Control: FastQC

Per base sequence content

Good

Not so Good
Quality Control: FastQC

Per sequence GC content

Good

Not so Good
Quality Control: Duplicates

- Some amount of duplication is to be expected in RNAseq
- Indicates very high fold coverage
- Highly expressed transcripts can be oversequenced in order to be able to see lowly expressed transcripts
- A concern when they occur in libraries where you expect equal coverage
- A badly PCR duplicated library might have levels above 90%

**FastQC Failure: Duplicate Sequences**

![Graph showing sequence duplication level](image)

- Sequence Duplication Level > 6.72x
- Duplicate relative to unique
Mapping RNAseq Reads

- Annotated reference is required
- To map junctions the algorithm needs to divide the sequencing reads and map portions independently
- Much more complex algorithms are required to identify alternative transcripts
Which Mapper to Use?

• Alignment algorithm must:
  • Be fast
  • Able to handle polymorphisms and sequencing errors
  • Accurately quantify
  • Detect spliced alignment

• Burrows Wheeler Transform Mappers (BWT)
  • Fast
  • < 3 mismatches allowed
  • Limited indel detection
    • Large or conserved genomes and transcriptomes
      • E.g. Bowtie2, BWA, Tophat

• Hash Table Mappers
  • Indexing requires a lot of RAM
  • Allows more mismatches
  • Indels detect is good
    • Small or highly variable genomes and transcriptomes
      • E.g. SHRiMP, STAR, GSNAP
Alignment Files

• SAM is the standard alignment file format generated from all mappers
  • Sequence Alignment/ Map format

• Alignments are stored in a BAM file (binary version of SAM)
  • Indexed to be read by other tools and genome browsers

• SAMtools is used to convert between SAM and BAM
  • http://samtools.sourceforge.net/

Example:

```
@HD VN:1.5 SN:coordinate
@SQ SN:ref LN:45
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACGG *
     0 ref 9 30 3S6M1P1I4M = 0 0 AAAGATAAAGGATA *
     0 ref 9 30 5S6M = 0 0 GCCTAGCTAA *
     0 ref 16 30 6M1A5N5M = 0 0 ATAGCTTCAGC *
     0 ref 29 17 6H5M = 0 0 TAGGC *
     0 ref 37 30 9M = 7 39 CAGCGGCAT *
     S:1:1
```
Counting Reads: General Rules

• Count mapped reads, not base pairs

• Count each read at most once

• Discard a read if
  • It cannot be uniquely mapped
  • Multimapping can arise if:
    • Primarily repetitive sequence (genome alignment)
    • Maps to multiple isoforms that share the same exon (transcript alignment)
  • Alignment overlaps several genes
  • Quality score is bad
  • Mates of paired-end reads do not map to the same gene

• Do not discard read duplicates
  • But if something looks fishy investigate!

• **ALWAYS** keep a good lab notebook
  • Document all parameters
Counting Reads: Direct Comparisons of Read Counts Can be Biased

- Estimates of gene and transcript expression are based on the number of reads that map to each transcript sequence.
- Raw read counts alone are not sufficient.

More sequence reads mapped to a transcript if it is

a) Long

Read Counts = 12, Depth = 3X, Read Counts = 5, Depth = 3X

b) At higher depth of Coverage

Read Counts = 11, Depth = 5X, Read Counts = 5, Depth = 3X
## Interpreting Read Counts

<table>
<thead>
<tr>
<th>Gene</th>
<th>Rep1</th>
<th>Rep2</th>
<th>Rep3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene A</td>
<td>5</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>Gene B</td>
<td>16</td>
<td>25</td>
<td>35</td>
</tr>
<tr>
<td>Gene C</td>
<td>10</td>
<td>15</td>
<td>3</td>
</tr>
<tr>
<td>Gene D</td>
<td>750</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td>Gene E</td>
<td>1504</td>
<td>1005</td>
<td>1030</td>
</tr>
</tbody>
</table>
Identify the Noise Threshold and Filter

Filter Point = 8.6
Normalizing RNAseq Count Data

• ALWAYS required to compare one sequencing result to another

• ALWAYS required to bring count data from different experiments to the same scale for comparison

• Adjust the data so that genes with different lengths can be compared
  • Necessary for ranking gene expression levels within a sample
  • Not necessary when comparing changes in gene expression of the same gene across samples

• Adjust the data so that total sequence counts (library sizes) are considered
Normalizing RNAseq Count Data: (FPKM)

\[ \text{FPKM}_i = \frac{X_i}{\left( \frac{\tilde{l}_i}{10^3} \right) \left( \frac{N}{10^6} \right)} \]

Where:
- \( X_i \) = counts in feature of interest
- \( l_i \) = length of feature
- \( N \) = total number of reads

- FPKM is fragments per kilobase of exon, per million reads mapped (also see RPKM)
- A within-sample normalization method that will remove feature length and library size effects
Normalizing RNAseq Count Data: Counts Per Million (CPM)

\[ \text{cpm} = \frac{x_{ij}}{N_i} \times 10^6 \]

Where:
- \( j \) = gene
- \( i \) = sample
- \( X_{ij} \) = raw count of the jth gene in the ith sample
- \( N_i \) = total counts of all genes in the ith sample

- Dynamic range of expression detected by RNAseq is large
- Probability that highly expressed genes will be detected as differentially expressed is greater than for lowly expressed genes
- Removing genes with little to no information (e.g. < 1cpm) reduces the number of statistical tests you perform
  - Reduce severity of multiple tests correction
  - Increase power
Normalization Factors (UQ)

\[ UQ_{\text{norm}}_{ij} = \frac{N_{ij}}{D_i} \times \frac{\sum_{i=1}^{n} D_i}{n} \]

where:
- j = gene
- i = sample

- A basic source of variation between samples is the difference in library size
- RNA samples may be sequenced to different depths

Upper Quartile = raw counts are divided by the upper quartile of library size and multiplied by the mean upper quartile across all samples
Normalization Factors (TMM)

- TMM – Trimmed mean of M values
- A scaling normalization that accounts for differences in RNA composition between samples
- Suppose two samples A and B, where every gene expressed in B is expressed in A with the same number of transcripts.
- If A also contains a set of highly expressed genes that are not as highly expressed in B, and the samples are sequenced to equal depth, there will be more RNA production in A and the reads will be spread over more genes.
- The TMM normalization adjusts sample A by an appropriate factor to account for the difference in RNA composition.
Variation in Library Sizes Before Normalization
Variation in Library Sizes After Normalization
Comparison of Common Normalization Factors

Dillies et al., Briefings in Bioinformatic (2012)
Which Normalization Method?

• R/FPKM
  • Corrects for differences in sequencing depth and transcript length
    • Compare a gene across samples and different genes within sample

• TMM
  • Corrects for differences in RNA composition and extreme outliers
    • Provides better across-sample comparability
      • Recommended for differential expression estimates

• CPM/logCPM
  • Stabilize variance and remove the dependence of the variance on the mean
Explore Your Data: Distances Among Samples

Euclidian Distances

\[ d_{(p,q)} = \sqrt{(p_1 - q_1)^2 + \ldots + (p_n - q_n)^2} \]

where:

- \( p_n \) = normalized count for gene \( n \) in first sample of pair
- \( q_n \) = normalized count for gene \( n \) in second sample of pair

Distance Matrix

```r
sampleDists <- dist(t(assay(rld)))
sampleDists
```

```
           SRR1039508  SRR1039509  SRR1039512  SRR1039513  SRR1039516  SRR1039517  SRR1039518
SRR1039508  46.25524
SRR1039509  39.94940  55.67572
SRR1039512  63.36642  45.19462  49.30097
SRR1039513  45.28129  59.89304  44.32383  64.54450
SRR1039516  65.34730  52.25475  60.05523  50.64861  48.05714
SRR1039517  40.20215  58.19904  37.35413  59.19401  47.15396  64.44641
SRR1039520  64.09339  45.70177  58.59277  37.10803  53.09669  50.72
```

HeatMap of Distances
Explore Your Data: How Do Your Samples Cluster?

- What pattern(s) do you see?
- What expectations do you have given your experimental design?

Principle Components Analysis plot of normalized counts
Clustering: Principal Components Analysis

- Identifies new variables (the principle components) that are linear combinations of the original variables
- Standardize data
  - Each gene is centered to average zero expression level
- The PCs are normalized eigenvectors of the covariance matrix of the genes
  - Ordered by how much variance they explain
<table>
<thead>
<tr>
<th>Method</th>
<th>Normalization</th>
<th>Need Replicates?</th>
<th>Input</th>
<th>DE Statistic</th>
<th>Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>edgeR</td>
<td>Library size</td>
<td>Yes</td>
<td>Raw counts</td>
<td>Empirical Bayesian estimation based on Negative binomial distribution</td>
<td>R/Bioconductor</td>
</tr>
<tr>
<td>DESeq</td>
<td>Library size</td>
<td>No</td>
<td>Raw counts</td>
<td>Negative binomial distribution</td>
<td>R/Bioconductor</td>
</tr>
<tr>
<td>baySeq</td>
<td>Library size</td>
<td>Yes</td>
<td>Raw counts</td>
<td>Empirical Bayesian estimation based on Negative binomial distribution</td>
<td>R/Bioconductor</td>
</tr>
<tr>
<td>LIMMA</td>
<td>Library size</td>
<td>Yes</td>
<td>Raw counts</td>
<td>Empirical Bayesian estimation</td>
<td>R/Bioconductor</td>
</tr>
<tr>
<td>CuffDiff</td>
<td>RPKM</td>
<td>No</td>
<td>RPKM</td>
<td>Log ratio</td>
<td>Standalone</td>
</tr>
</tbody>
</table>
Differential Expression Analysis Results

<table>
<thead>
<tr>
<th>Gene or Transcript</th>
<th>Mean Expression Values</th>
<th>Significance: use adjusted pvalue rather than raw pvalue</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>baseMean</td>
<td>baseMeanA</td>
</tr>
<tr>
<td>S100A8</td>
<td>1155.68</td>
<td>313.37</td>
</tr>
<tr>
<td>S100A2</td>
<td>936.33</td>
<td>273.45</td>
</tr>
<tr>
<td>NES</td>
<td>151.28</td>
<td>12.97</td>
</tr>
<tr>
<td>PSCA</td>
<td>1032.32</td>
<td>373.25</td>
</tr>
<tr>
<td>IFI6</td>
<td>9349.03</td>
<td>4582.81</td>
</tr>
<tr>
<td>IFI44L</td>
<td>1096.73</td>
<td>482.03</td>
</tr>
<tr>
<td>KRT6A</td>
<td>932.56</td>
<td>404.19</td>
</tr>
<tr>
<td>SBSN</td>
<td>195.29</td>
<td>50.90</td>
</tr>
<tr>
<td>KLHDC7B</td>
<td>12198.92</td>
<td>6986.00</td>
</tr>
</tbody>
</table>

**Fold Change:** measurement of the changing magnitude (effect size)

\[ FC = \frac{\text{baseMeanB}}{\text{baseMeanA}} \]

Typically log2(FC) is reported
Can You Identify Patterns?

- Functional Profiling
- Pathway analysis
- Interaction analysis
- Gene Ontology analysis
Advanced Analyses: Other RNAseq Applications

- Isoform analysis
  - Characterizing transcriptome complexity

- Small and other non-coding RNAs
  - siRNAs, microRNAs, enhancer associated RNAs

- Gene fusion discovery
  - Hybrid genes formed by two previously separate genes

- Long-read sequencing
  - Amplification-free single molecule sequencing of cDNAs enable recovery of full length transcripts
    - No assembly!

- Single-cell analysis
  - Work from very small amounts of starting mRNA that can be obtained from a single cell with proper amplification
How to integrate next-generation sequence data to attain a more comprehensive and realistic understanding of biological processes?

The Epigenome Roadmap!
Summing It All Up

• Study design and RNA extraction
  • Replicates?
  • Power?
  • What technology is most appropriate?

• Data QC
  • Are there technical issues?
  • How correlated are your replicates?

• Normalization
  • Choose the method that is appropriate for your experiment.

• Explore your data
  • Calculate distances among your samples. Do the patterns fit your expectations?

• Calculate differential expression and look for patterns
  • Choose the method that is most appropriate for your experiment.

• Interpret the results in terms of biological or physiological knowledge
  • YOU must believe in your data ...