Expression Quantification and Differential Expression Analysis
How do we analyze RNA-Seq data?

• **STEP 1**: EVALUATE AND MANIPULATE RAW DATA
• **STEP 2**: MAP TO REFERENCE, ASSESS RESULTS
• **STEP 3**: ASSEMBLE TRANSCRIPTS
• **STEP 4**: QUANTIFY EXPRESSION
• **STEP 5**: TEST FOR DIFFERENTIAL EXPRESSION
• **STEP 6**: VISUALIZE AND PERFORM OTHER DOWNSTREAM ANALYSIS
STEP 4- Quantify Expression

• Mapping tells us where every read came from.

• How do we go from that to gene expression?

• What genes are expressed?

• What is the expression level for each gene/gene isoform?
What is gene expression?

A gene is expressed when it’s corresponding DNA sequence is transcribed into mRNA (for translation into protein).

What is gene expression level?

The amount of mRNA detected in a sample.

- Read depth = mRNA amount = expression level of gene
STEP 4: Quantify Expression

• Bedtools
  – **Bedtools multicov**: Takes a feature file (GFF/GTF) and counts how many reads in the mapped output file (BAM) overlap the features.

  – Remember that the chromosome names in your gff file should match the chromosome names in the reference fasta file used in the mapping step.
STEP 4 : Quantify Expression

HTSeq –
- Gives you fine grained control over how to count genes, especially when a read overlaps more than one gene/feature.
STEP 4- Quantify Expression

- Quantifying a gene is simpler than quantifying its different isoforms/transcripts.
- Tools: kallisto, stringtie, and cufflinks

What is a gene? What is a transcript?
A gene can have multiple transcripts!
STEP 4- Quantify Expression

Why quantifying all transcripts of the gene may be important?
STEP 5- Test for Differential Expression

- After mapping and quantifying the genes for each sample:
  - compare gene counts across samples/conditions.

- But first, **normalize**!
  - Normalization evens out the technical variations so that any variation you see between samples is “hopefully” due to real biological reasons.
  - Normalize for **read depth** differences
  - Normalize for **gene/transcript length** differences

- RPKM = Reads Per **Kilobase of transcript** per Million mapped reads

- Other normalization methods: upper quartile, median read count and more complicated scaling factors (DESeq2 R package)

<table>
<thead>
<tr>
<th>Gene</th>
<th>Read Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABAR1</td>
<td>1200</td>
</tr>
<tr>
<td>ATXN1</td>
<td>1345</td>
</tr>
<tr>
<td>ATXN2</td>
<td>2</td>
</tr>
<tr>
<td>BRAT2</td>
<td>0</td>
</tr>
<tr>
<td>GABA</td>
<td>24</td>
</tr>
<tr>
<td>GABRA2</td>
<td>456</td>
</tr>
<tr>
<td>GABRA4</td>
<td>45345</td>
</tr>
</tbody>
</table>

Figure: doi:10.1038/nmeth.1613
STEP 5- Test for Differential Expression

- Input: Gene Expression Matrix

![Gene Expression Matrix Table]

- Outputs like:

![Downregulated and Upregulated Gene Expression Graphs]

Figure: doi:10.1038/nn.4065