Introduction to the MiSeq System
MiSeq Sequencing Workflow

1. Library Preparation
2. Cluster Generation
3. Sequencing
4. Data Analysis
Sample Prep is Critical for Successful Sequencing

The aim of the sample prep step is to obtain nucleic acid fragments with adapters attached on both ends.
Total RNA isolation

mRNA Isolation using Oligo(dT) Magnetic Beads

First-Strand cDNA Synthesis with Random Primers

Second-Strand cDNA Synthesis

Double-Stranded cDNA

End Repair of double-stranded cDNA

Adenylation (A-Tailing)

Blunt/TA Adaptor Ligation

USER enzyme Excision

PCR Amplification using a Universal primer and index primer

Barcode

Purify and size select cDNA Library using AMPure Beads

Total RNA isolation

mRNA Isolation using Oligo(dT) Magnetic Beads

First-Strand cDNA Synthesis with Random Primers

Second-Strand cDNA Synthesis

Double-Stranded cDNA

End Repair of double-stranded cDNA

Adenylation (A-Tailing)

Blunt/TA Adaptor Ligation

USER enzyme Excision

PCR Amplification using a Universal primer and index primer

Barcode

Purify and size select cDNA Library using AMPure Beads
MiSeq Sequencing Workflow

1. Library Preparation

2. Cluster Generation

3. Sequencing

4. Data Analysis
Cluster Generation

Bind single DNA molecules to surface

Amplify on surface

~1000 molecules per ~1 µm cluster
Hybridize Fragment & Extend

Single DNA libraries are hybridized to primer lawn

Bound libraries then extended by polymerases

Surface of flow cell coated with a lawn of oligo pairs

Adapter sequence

3' extension
Hybridize Fragment & Extend

NOTE:
Single molecules bind to flow cell in a random pattern
Denature Double-Stranded DNA

- Double-stranded molecule is denatured
- Original template washed away
- Newly synthesized strand is covalently attached to flow cell surface

Original template
Newly synthesized strand
discard
Bridge Amplification

Single-stranded molecule flips over and forms a bridge by hybridizing to adjacent, complementary primer

Hybridized primer is extended by polymerases
Bridge Amplification

Double-stranded bridge is formed
Denature Double-Stranded Bridge

Double-stranded bridge is denatured

Result:
Two copies of covalently bound single-stranded templates
Bridge Amplification

Single-stranded molecules flip over to hybridize to adjacent primers

Hybridized primer is extended by polymerase
Bridge Amplification

Bridge amplification cycle repeated until multiple bridges are formed
Linearization

dsDNA bridges are denatured
Reverse Strand Cleavage

Reverse strands cleaved and washed away, leaving a cluster with forward strands only.
Blocking

Free 3’ ends are blocked to prevent unwanted DNA priming
Read 1 Primer Hybridization

Sequencing primer is hybridized to adapter sequence

Sequencing primer
MiSeq Sequencing Workflow

1. Library Preparation
2. Cluster Generation
3. Sequencing
4. Data Analysis
Sequencing by Synthesis

1. Add 4 Fi-NTP's + Polymerase
2. Incorporated Fi-NTP imaged
3. Terminator & fluorescent dye cleaved from Fi-NTP

X 36 - 251
Paired-End Flow Cell

Region complementary to P5 grafting primer

Index 2 (i5)

P5 primer

SBS Sequencing Primer (HP10)

DNA insert

Sequence (Cycle 1)

P7 primer

Index 1 (i7)

P7 grafting primer

Flow cell surface

P5 grafting primer
Clustering are images using LED and filter combinations specific for each fluorescently-labeled nucleotide.

After imaging is complete for one section (tile), the flow cell is moved to the next tile and the process is repeated.
Experimental Workflow

• Calculate coverage you want for your genome/transcriptome, and, using the coverage calculator, calculate how many samples will fit on your run.

• Isolate your RNA as usual, and get rid of rRNA if not needed

• Quantitate on the Bioanalyzer (for quality) and Qubit (for quantity) (may want to isolate a couple of extra and take the best ones).

• Prepare each library.

• Quantitate one more time with bioanalyzer and Qubit, then make calculation for dilutions to 4nM which is where the protocol starts on sequencing day.

• On day of sequencing you will denature your 4nM libraries and dilute to desired molarity (generally 10-12pM for these libraries). You will do the same for the PhiX library purchased from Illumina, which you will put on every run as a control.

• Finally, pool all libraries and PhiX together in 650μl, with PhiX being 1-5% of pool and the rest of the libraries evenly represented. Load 600μl into cartridge and run.
## MiSeq Output Calculations

<table>
<thead>
<tr>
<th></th>
<th>MiSeq with:</th>
<th>MiSeq with:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Upgraded hardware, or from September 2012 and later</td>
<td>Upgraded hardware, or from September 2012 and later</td>
</tr>
<tr>
<td></td>
<td>MCS v2.3 or later</td>
<td>MCS v2.0 or later</td>
</tr>
<tr>
<td></td>
<td>MiSeq Reagent Kit v3</td>
<td>MiSeq Reagent Kit v2</td>
</tr>
<tr>
<td>Reads/flow cell</td>
<td>25,000,000</td>
<td>16,000,000</td>
</tr>
<tr>
<td>Genome or region size (in bases)</td>
<td>600,000</td>
<td>600,000</td>
</tr>
<tr>
<td>Coverage</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Total number of cycles (e.g. 300 for 2x150)</td>
<td>150</td>
<td>50</td>
</tr>
<tr>
<td>Total output required (in bases)</td>
<td>60,000,000</td>
<td>60,000,000</td>
</tr>
<tr>
<td>Output/flow cell (bases/flow cell)</td>
<td>3,750,000,000</td>
<td>800,000,000</td>
</tr>
<tr>
<td>Number of flow cells</td>
<td>0.02</td>
<td>0.08</td>
</tr>
<tr>
<td>Number of samples per flow cell</td>
<td>62.50</td>
<td>13.33</td>
</tr>
</tbody>
</table>
## Reads Mapped to Index ID

<table>
<thead>
<tr>
<th>Index Number</th>
<th>Sample ID</th>
<th>Project</th>
<th>Index 1 (17)</th>
<th>Index 2 (18)</th>
<th>% Reads Identified (PF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>NA</td>
<td>ATCACG</td>
<td></td>
<td>6.4687</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>NA</td>
<td>CGATGT</td>
<td></td>
<td>3.7882</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>NA</td>
<td>TTAGGC</td>
<td></td>
<td>3.04</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>NA</td>
<td>TGACCA</td>
<td></td>
<td>4.4433</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>NA</td>
<td>ACAGTG</td>
<td></td>
<td>3.5587</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>NA</td>
<td>GCCAAT</td>
<td></td>
<td>3.7034</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>NA</td>
<td>CAGATC</td>
<td></td>
<td>3.8523</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>NA</td>
<td>ACTTGA</td>
<td></td>
<td>3.5304</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>NA</td>
<td>GATCAG</td>
<td></td>
<td>1.6697</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>NA</td>
<td>TAGCTT</td>
<td></td>
<td>3.5399</td>
</tr>
<tr>
<td>11</td>
<td>11</td>
<td>NA</td>
<td>GGCTAC</td>
<td></td>
<td>2.2518</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
<td>NA</td>
<td>CTGTGT</td>
<td></td>
<td>2.5841</td>
</tr>
<tr>
<td>13</td>
<td>13</td>
<td>NA</td>
<td>AGTCAA</td>
<td></td>
<td>6.5291</td>
</tr>
<tr>
<td>14</td>
<td>14</td>
<td>NA</td>
<td>AGTCC</td>
<td></td>
<td>4.4278</td>
</tr>
<tr>
<td>15</td>
<td>15</td>
<td>NA</td>
<td>ATGCTA</td>
<td></td>
<td>6.6197</td>
</tr>
<tr>
<td>16</td>
<td>16</td>
<td>NA</td>
<td>CCGTCC</td>
<td></td>
<td>0.8886</td>
</tr>
<tr>
<td>17</td>
<td>17</td>
<td>NA</td>
<td>GTCCG</td>
<td></td>
<td>4.7116</td>
</tr>
<tr>
<td>18</td>
<td>18</td>
<td>NA</td>
<td>GTGAAA</td>
<td></td>
<td>2.6488</td>
</tr>
</tbody>
</table>
Bioinformatics on large datasets

Start by coming up with a workflow, then try to find a primary publication that did the same.

**Sample workflow**

QC Data -> Trim Data -> Assemble Data (if no reference de novo) -> Annotate Transcripts

Transcripts -> Map/Align reads to reference (either published or your new annotated de novo assembly) -> Count Reads that mapped to each transcript -> Normalize and Compare Read Counts
Bioinformatics on large datasets

Software Options For RNA-Seq

**QC Data** – FastQC – in Galaxy or Commandline (limits in Galaxy – data size, goes down)

**Trim Data** – FastQCTrimmer vs Quality Filter in Galaxy or Commandline

**Assemble Data** – Trinity in Galaxy or Commandline (limits in Galaxy – control – if on commandline, can normalize (eliminate repeated sequence (called digital normalisation)) to give program less noise -> less memory needed, more speed for assembly)

SoapDeNovoTrans – make sure for RNA otherwise multiples reads of the same sequence will be viewed as repeats.

**Annotate Transcripts** – Trinotate, JAMg

Some websites:  
http://dnasubway.iplantcollaborative.org/
http://goblinx.soic.indiana.edu/src/yrGATE

**Map/Align reads to reference** – Tophat either in Galaxy or Commandline (only works if aligning to a reference genome), Bowtie, RSEM, **BWA**

**Count Reads** – Cufflinks either in Galaxy or Commandline, **HTSeq**, RSEM

**Compare Read Counts** – Use CuffMerge to bring together all data sets you want to compare and then run CuffDiff, **DESeq2**, eXpress, RSEM
From here you will likely want to do:
Gene Ontology – DAVID and Cytoscape (networks), BioCyc or other functional genomic resources depending on organism (eg. EcoCyc is BioCyc for *E. coli*).

SNP detection, variant calling

There is also a very fun website called KeggAnime that gives you animation of your genes of interest in their respective pathways, [http://biit.cs.ut.ee/kegganim/index.cgi](http://biit.cs.ut.ee/kegganim/index.cgi)
A note on normalization:
RPKM switched to FPKM with the advent of paired end reads, making Reads per Kilobase per Million is no longer accurate as there were 2 reads per fragment. Thus, FPKM or Fragment per Kilobase per Million is now used.

HOWEVER, FPKM is a within sample normalization: it allows you to compare relative expression of genes or transcripts within a single sample to for instance plot expression of several genes in a single sample. FPKM is NOT comparable between samples because in addition to normalizing by transcript length and library size, it includes a sample-specific normalization constant. In addition, if you are comparing the expression of a gene across samples, that gene/transcript size is going to be the same across conditions (if comparing different species to each other, then FPKM may still be in play).

Thus, now it is recommended to use TPM (Transcripts per Million reads) as an equivalent metric that is comparable between samples. The math is explained beyond my understanding in a paper by Wagner:

Cufflinks in Galaxy only lets you do FPKM – at the very least you should try one other method of analysis and compare your results.

Don’t use sequencing to draw conclusions. Use sequencing to direct hypotheses that you then explore experimentally.
References and Places to Go

- Watch some training videos that pertain to you at Illumina.com http://support.illumina.com/training/sequencing_training.ilmn (Chemistry overview is wicked helpful).

- Register with Illumina.com to get your BaseSpace account and info on past and future webinars.

Integrated Genome Browsers
- UC Santa Cruz: hosts many genomes, including E.coli
- IGV at Broad Institute: recommended by Illumina

Aligner
- Mauve

Mapping Viewer
- Tablet
  - Tview in Samtools

Gene Ontology
- DAVID
- BioCyc and EcoCyc
- Cytoscape
- KEGG and KeggAnim

When all else fails, Google “How to” or “Manual” or “Tutorial” your problem or get an account at SeqAnswers and BioStar and ask the forums.
Some Illumina Tech Notes to Peruse


And many more at: [http://support.illumina.com/sequencing/literature.ilmn](http://support.illumina.com/sequencing/literature.ilmn)