RNA-Seq Analysis Pipeline

This pipeline uses an annotated genome to identify differential expressed genes/transcripts. 10 hour minimum ($470 internal, $600 external) per project.

1. Quality Assessment

Quality of data assessed by FastQC; results of quality assessment will be evaluated prior to downstream analysis.

- **Deliverables:**
  - reports generated by FastQC
- **Tools used:**
  - FastQC: (Andrews 2010) used to generate quality summaries of data:
    - Per base sequence quality report: useful for deciding if trimming necessary.
    - Sequence duplication levels: evaluation of library complexity. Higher levels of sequence duplication may be expected for high coverage RNAseq data.
    - Overrepresented sequences: evaluation of adapter contamination.

2. Fastq Preprocessing

Quality assessment used to decide if any preprocessing of the raw data is required and if so, preprocessing is performed.

- **Deliverables:**
  - Trimmed/filtered fastq files.
- **Tools Used:**
  - Fastx-toolkit: Used to preprocess fastq files.
    - Fastq quality trimmer: Trimming reads based on quality.
  - Cutadapt: Used to remove adaptor from reads.

3. Mapping

Mapping to genome reference performed using BWA-mem or Tophat.

- **Deliverables:**
  - Mapping results, as bam files and mapping statistics.
- **Tools Used:**
  - BWA-mem: (Li 2013) primary aligner used to generate read alignments.
  - Tophat: (Kim 2011) aligner used to generate read alignments in a splice-aware manner and identify novel junctions.
  - Samtools: (Li 2009) used to generate mapping statistics.

4. Gene/Transcript Counting

Counting the number of reads mapping to annotated intervals to obtain abundance of genes/transcripts.

- **Deliverables:**
  - Raw gene/transcript counts
- **Tools Used:**
  - HTSeq-count: (Anders 2014) used to count reads overlapping gene intervals.

5. DEG Identification

Normalization and statistical testing to identify differentially expressed genes.

- **Deliverables:**
  - DEG Summary and master file containing fold changes and p values for every gene, MA Plots.
- **Tools Used:**
  - DESeq2: (Love 2014) used to perform normalization and test for differential expression using the negative binomial distribution.