RNA-Seq concepts

An RNA-Seq analysis has three separate yet equally important steps

1. Identify/assign reads to locations

2. Estimating abundances per transcript

3. Comparing abundances → differential expression

There is enormous disagreement about how each these steps should be performed → hence very large number of options

There are workflows that do all three. There are workflows that mix and match from methods.
Identify/assign reads to locations

Splicing aware alignment tools for genome. Transcript aware alignment for transcriptome.

- TopHat (one of the first tool that worked RNA-Seq, Tuxedo suite)
- HiSat (same developers aligns much faster)
- Subread
- GEM ...

There are a very large number of tools – many of them are suspect.
Estimate Abundance

• Cuffquant/cuffdiff (part of the Tuxedo suite)

• Featurecounts (from the Subread developers)

• Htseq-count
Differential expression

- Cuffdiff (part of the Tuxedo suite)
- DESeq and DESeq2 (R-package)
- EdgeR (R-Package)
Tuxedo Suite

Tophat (splice mapping)

accepted_hits.bam

Cufflinks (assemble transcripts)

transcript.gtf

Cuffcompare

Cuffmerge (merge transcripts)

merged.gtf

Cuffdiff (differential expression)

differential expression with other software
Tuxedo Suite: tophat

**Tophat** – splice mapper

1. Uses Bowtie to map reads to genome
2. Takes unmapped reads and maps them to junctions
3. Identifies junctions from an external file or by finding potential junction sites (GT-AG) splice sites
Tophat output

creates a new output directory with files

• accepted_hits.bam

• various bed files: junctions, insertions, deletions
Tuxedo suite: cuffdiff, cuffcompare

- Cuffcompare: track transcripts across multiple experiments
- Cuffdiff: calculate expression levels → differential expression

Each tool generates a very large number of output files
Understanding the Tuxedo suite is almost like a **full time job**.

It has lots of documentation but also makes substantial tacit assumptions.
Homework 29

Due by midnight, Sunday Dec 13th, 2015

Find the 10 genes that show the most fold change in the dataset that comes with the tutorial:

“Informatics for RNA-seq: A web resource for analysis on the cloud”

See course webpage for link.