Week 15, Lecture 29

István Albert

Biochemistry and Molecular Biology and Bioinformatics Consulting Center
Penn State

Final Project

- Will account for 50% of your the grade.
- It will consist of many datasets and a number of hypotheses that need to be tested.
- Project given out Thursday, Dec 10th and is due 5pm Saturday, Dec 14th.
- You may email your project as a document.
- Office hour Wed 1-3pm and Friday 1-3pm.
- Plus on the StudyRoom webpage.

Data and scripts for this lecture

- Data, code and scripts are packaged in lec29.tar.gz.
- RNA-Seq one has to master many tools and it is easy to get stuck.
- Use it as a guide - don’t just copy/paste.
- Develop your own mini libraries of tools, shell scripts and methods.

RNA – Ribonucleic Acid
RNA-Seq – measuring gene expression

- Quantifying gene expression → one of the “holy grails” of molecular biology

- The underlying (often tacit) assumption is that most phenotype can be explained by gene expression levels

- There is also a more DNA centric view where most phenotype is thought to be determined by the DNA sequence

Goals of RNA-Seq

- Quantify RNA abundance/ quantify changes in abundance

- Identify full transcripts: variants and isoforms

- Identify transcriptions start/end (TSS, TSE) sites → regulation

Two major domains application

1. Known Transcriptome
   - discover new isoforms
   - transcript variation → protein variation
   - differential expression

2. Unknown Transcriptome
   - identify transcripts and variants

RNA-SEQ is a new, emerging field

- With that come the challenges:
  - the majority of tools techniques are immature and make use of empirical observations
  - it is a wild race - with often greatly exaggerated claims
  - it is quite complicated → the rise of the black box → a complex tool with a fancy name that needs to be followed rigorously step by step, in return it promises an easy answer
  - Historically speaking new releases of these black boxes produce results that are only partially concordant (say 50% of genes are identical)
Important publications
Seminal papers


RNA-Seq: gene expression levels
Sample composition

1. Genes have various levels of expression → higher expression levels produce more reads for that gene

2. Genes of various lengths → longer lengths produce more reads for that gene

3. Sequencing coverage determines the rarest transcript that can be detected
   (there is an analogy here to meta-genomic -> genome size, abundances)

Measure of gene expression levels RPKM

Number of mapped reads per kilo-base exon model per millions of mappable reads

Kilobase and Millions are the units of measurement!

Reads aligned over the exon \( R = 5 \)
Total number of reads \( N = 100,000 \rightarrow 0.1 \text{ million} \) reads
Feature size \( L = 200 \text{ bp} \rightarrow 0.2 \text{ kilobase} \) of exons

\[
\text{RPKM} = \frac{R}{(0.2 \times 0.1)} = 250
\]
RPKM – simplified formula
Move the magnitudes to the denominator

\[ \text{RPKM} = 10^9 \times \frac{R}{NL} \]

Total number of reads \( N = 100,000 \)
Feature size \( L = 200 \) bp
Reads aligned over the exon \( R = 5 \)

(note how the units are 1/bp)

Fundamental problems

- Most scientists use RPKM without understanding what it is
- Most scientists were mislead into believing that RPKM is a way to quantify gene expression levels and they use RPKM to compare genes within and across samples
- Tools readily produce RPKM and thus is very convenient to use the values

RPKM measure is inconsistent among samples

http://blog.nextgenetics.net/?e=51

Attempt to fix it FPKM

(it is conceptually worse)

Number of fragments per kilo-base exon model per million mappable reads

- Computed from paired end reads where the fragment size is known (column 9 in SAM format)

fragment size not equal to transcript size!
New suggestion: TMP

Transcripts per million

\[ RPKM_i \sim \frac{R_i}{NL_i} \]

\[ TPM_i \sim \frac{R_i}{TL_i} \]

N = Total mapped reads
T = Average normalized transcript length
T = \( \sum (R_i/L_i) \)
L_i = Feature length

Tuxedo Suite

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: cuffmerge

cuffmerge — it can merge different cufflinks assemblies into one file

1. Assembles tophat output into transcripts, transcripts.gtf, isoforms.fpkm_tracking, and genes.fpkm_tracking

2. Output files: transcripts.gtf, isoforms.fpkm_tracking, and genes.fpkm_tracking

3. May be reference guided or unguided

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 a full time job.

It has lots of documentation but also immense amount of tacit assumptions.