RNA-Seq – measuring gene expression

- Quantifying gene expression \(\rightarrow\) 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 \(\rightarrow\) regulation
Two major domains application

1. Known Transcriptome
   - discover new isoforms
   - transcript variation \(\rightarrow\) 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 often incorrect
  - it is a wild race - with often greatly exaggerated claims
  - it is quite complicated \(\rightarrow\) the rise of the black box \(\rightarrow\) 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 reads per kilo-base exon model per millions of mappable reads
  
  **Kilobase and Millions are the units of measurement!**

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

  \[
  \text{RPK} = \frac{5}{0.2} = 25
  \]

  \[
  \text{RPKM} = \frac{5}{(0.2 \times 0.1)} = 250
  \]

RPKM – simplified formula

Move the magnitudes to the denominator

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

Total number of reads \( N = 100,000 \)
Exon size \( L = 200 \text{ bp} \)
Reads aligned over the exon \( C = 5 \)

FPKM

- 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!
Representing interval relationships

• We have a gene with three splicing variants

How to represent this in data analysis?

GFF/GTF formats

GTF attributes:

- `gene_id` value;
  a globally unique identifier for the genomic source of the transcript
- `transcript_id` value;
  a globally unique identifier for the predicted transcript.

```
gene_id "G1" transcript_id "T1"
```

GFF attributes:

```
ID=exon1; Parent=T1
```

See the GFF3 site for exact specification of these meanings.

Important: More than one parent may be listed!

Data representation

(somewhat ill defined)

• Both BED and GFF files can represent them

• Two common versions of GFF → GTF and GFF3
  (note: tool documentation is often wrong and shows a weird combination of these two formats)

• The content of the ATTRIBUTE (9th) column specifies the relationship between features

Example interval as GTF

A distinct line is entered for each exon, repeated for each transcript
Example interval as GFF 3

The same exon may be part of different transcripts (parents)

Example interval in BED

From the BED format specification

Visualizing in IGV

What if we just align it with a typical aligner?

1. Generate the sequence for each transcript
2. Simulate an experiment from these transcripts
3. Align them with a mapper
Next lecture we will use a tool that takes into account RNA-seq specific properties, biases, and errors.

Important Notes

- We were only able to align reads that reside fully in exonic regions and only if the read length is shorter than the exon.

- To detect junctions reads that span exons we need a special (junction) aligner

- Even so do have a sense of where the data maps and we can assess coverage for many exons

More sophisticated tools use the reads that we were unable to map above to determine the connections between exons.

Homework 26

- Create and visualize in IGV an interval file that contains three splice variants of a 1kb long gene with 5 exons.