RNA Quantitation from RNAseq Data

June 16, 2017

1 A Model for RNAseq-based Quantitation

An important use of RNAseq data is to quantify gene expression.

- Suppose we have an experimentally derived sample of RNA, e.g. all transcripts from a population of cells.
- Intuitively, the more frequently some RNA transcript is expressed in our sample, the more RNAseq reads should arise from that transcript.
- But how can we make this intuition quantitative?
- Can we infer things like:
  - "Gene g is expressed at a level of k copies per cell in this sample."
  - "Gene g is expressed at twice the level of gene g' in this sample."
  - "Gene g is expressed twice as much in sample 2 as in sample 1."
- To see which of these inferences is feasible, we need to think carefully about how the read counts we see are related to the underlying population of RNA transcripts.

Let's construct a simple model relating RNA transcript abundance to what we actually observe in our RNAseq data.

- Assume our sample contains RNA molecules drawn from n different transcripts.
- Transcript i has length ℓ_i bases and occurs c_i times in the sample.
- We perform an RNAseq experiment on our sample, generating M (unpaired) reads of uniform length s. Each read maps unambiguously to one of the n transcripts.
- The last $s-1$ positions of a transcript cannot be the start of a mapped read, because a sequence of length s cannot begin there! Hence, define the *effective length* $\overline{\ell}_i$ of transcript i to be
$$\overline{\ell}_i = \ell_i - s + 1,$$
which is the number of places a mapped read of length s could start.
- We assume that any possible start position of any RNA molecule in the sample is equally likely to generate a mapped read.

- The total number of possible start positions C across all RNA molecules in the sample is given by
$$C = \sum_j c_j \bar{\ell}_j.$$

2 Estimates of Transcript Abundance I: RPKM

How does RNAseq data inform our beliefs about RNA abundance?

- Let f_i be the fraction of all possible starting positions for a mapped read in our sample that lie inside a copy of transcript i.
- How is f_i related to c_i, the number of copies of transcript i in the sample?
- By definition of f_i, we have that
$$f_i = \frac{c_i \bar{\ell}_i}{\sum_j c_j \bar{\ell}_j} = \frac{c_i \bar{\ell}_i}{C}.$$
- Hence, the quantity $f_i/\bar{\ell}_i$ is proportional to c_i, with the same constant of proportionality $(1/C)$ for every transcript i in the sample.
- Now RNAseq tells us a number m_i of reads observed to map to each transcript i.
- Because our model assumes that each read is sampled randomly from among all feasible starting positions, we expect reads from transcript i to constitute an f_i fraction of all reads observed.
- More precisely, in the limit of many observed reads M, we have by the Law of Large Numbers that
$$E[m_i/M] \to f_i.$$
That is, m_i/M is a good estimate of f_i.
- Conclude that we can estimate c_i/C from the RNAseq data as $R_i = m_i/M\bar{\ell}_i$.
- In practice, a scaled version of R_i is usually reported, namely
$$m_i/(M/10^6)(\bar{\ell}_i/10^3),$$
which has units of "Reads Per Kilobase of transcript i per Million reads sampled" (abbreviated RPKM).

What can and cannot be done with RPKM?

- Given two transcripts i and j in the *same* sample, their ratio of abundances c_i/c_j can be estimated as R_i/R_j.
- However, the same R_i value could correspond to different counts c_i in two different samples!
- In particular, the constant of proportionality C for each sample may be different, depending on which *other* RNAs are present in each sample and in what quantities.
- Hence, RPKM values cannot be used to compare the abundance of a single transcript across two samples.

3 Estimates of Transcript Abundance II: TPM

Can we obtain a measure of transcript abundance that is more meaningful across samples?

- Let t_i be the fraction of all RNA *molecules* in a sample that are copies of transcript i.
- t_i is proportional to $f_i/\bar{\ell}_i$ – the number of positions at which a mapped read could start that lie in copies of transcript i, divided by the (effective) length of i.
- However, we know by definition that

$$\sum_i t_i = 1.$$

- Hence, we can estimate t_i directly by the quantity

$$T_i = \frac{m_i/M\bar{\ell}_i}{\sum_j m_j/M\bar{\ell}_j}.$$

- T_i is usually multiplied by 10^6 for reporting, which gives it the units of "copies of Transcript i Per Million RNA molecules" (abbreviated TPM).

Is TPM better than RPKM?

- Within a single sample, TPM is no better than RPKM for comparing abundances.
- However, TPM values are comparable, in a limited sense, across samples.
- One can show that transcript i forms a larger or smaller *fraction* of all transcripts in sample 2 compared to sample 1.
- This kind of inference is not possible with RPKM values.
- Even TPM values cannot tell us the *absolute number* of copies c_i of a transcript in a sample.
- To estimate the ratio of abundances of a transcript in two samples, we need either to know (or infer) the relative numbers of RNA molecules in each sample, or to use an internal standard of known absolute abundance, to relate T_i to c_i in each sample.

Software for comparing the absolute expression of a gene across samples using RNAseq (e.g. DESeq, edgeR) uses neither RPKM nor TPM but instead applies a quite different modeling approach that compares raw read counts, normalized by the estimated sequencing depth per sample. But that's a topic for another day...

4 Caveats

Our simple model for estimating mRNA abundance from mapped reads is *too* simple. Most simplifications lead to errors in sample-to-sample abundance comparisons!

- *Assumption*: reads are unpaired.
- If we can identify a read pair as mapping to the same transcript, we should only count one "hit" to the transcript, corresponding to the entire RNA fragment from which the two reads were obtained.
- Adjusting RPKM to account for paired reads yields FPKM (R: "reads", F: "fragments").
- Normalization to TPM is similar if all fragments have the same length, but in practice, the fragment length distribution must be considered (s is now variable and large relative to transcript lengths ℓ_i).
- *Assumption*: reads are sampled uniformly from all RNA molecules.
- Sampling may be biased due to limitations of library construction and sequencing protocols.
- We can model such biases explicitly as part of abundance estimation.
- *Assumption*: each read maps to only one transcript.
- Different isoforms of a gene, and often different genes, share sequence, leading to multiply mapped reads.
- Various models try to account for (known) multiple mappings when estimating abundance.
- *Assumption*: every read maps to some transcript.
- Reads may fail to map due to sequencing errors, sample contamination, or the presence of real transcripts not known from the reference genome / transcriptome.
- Normalization to TPM when not all transcripts in the sample are known will result in incorrect estimates of t_i.
- Unmapped reads caused by unknown transcripts can be assembled to discover new transcripts, or simply assigned to a "garbage" transcript as part of TPM normalization.
- Unmapped reads caused by sequencing error or polymorphism can often be assigned to a transcript using approximate mapping algorithms.

Various RNAseq quantitation tools try to address one or more of these limitations. The model used to estimate abundance may become quite complex; See, e.g., RSEM (Li and Dewey 2011).