Discussion 3 - RNA-seq normalization overview

Albert Xue

January 2023

1 Introduction

Normalization is the heart of almost all the analyses we do as computational biologists; before we do any kind of differential analysis between conditions, we first must construct a valid comparison between conditions.[1] Personally, I also that if you understand the normalization processes we use, and why we use them, then it admits understanding of the data itself.

These notes will essentially be a series of thought experiments. We'll start with length biases, and then move through some thought experiments on the compositional nature of RNA-seq.[2]

2 Length Biases

Question 1. *Let's say two genes, gene A and gene B, have the same observed number of RNA fragments. For argument's sake, let's say we observed all of these fragments to the correct number. Which gene is more highly expressed?*

As it turns out, with the given information anything is possible. A could be more highly expressed than B; it could be less highly expressed than B; it could even be equally expressed to B. This is because I failed to give you the lengths of each gene, and that has fundamental consequences for downstream quantification.

The argument is entirely physical. Recall that we never directly observe a single mRNA transcript directly; instead, we observe a noisy picture after each mRNA transcript has been broken down into fragments. When we break mRNA transcripts into fragments, long mRNA transcripts will therefore have more fragments in expectation, especially sequence-able fragments (see Figure 1).

Even if we observe the exact same number of fragments for each gene, the answer is entirely dependent on the lengths of A and B. If $\text{len}(A) < \text{len}(B)$, the exact same number of RNA fragments between the two would actually mean that A is more highly expressed than B. On the other hand, if $\text{len}(A) > \text{len}(B)$, then B is suddenly actually more highly expressed than A.

Again, in these scenarios none of our data has changed. We've observed the exact same actual number of fragments or reads. If we observed transcripts directly, we wouldn't have length biases; but in sequencing we never directly observe anything, and so we need to normalize our data correctly to avoid false interpretations.

[1] For example, imagine that we have an RNA-seq sample from a healthy individual and an RNA-seq sample from a diseased individual. We might ask if we can identify genes whose expression levels change from the healthy condition to the diseased condition.

[2] By the way, Harold is probably one of the best people in the world to ask about these questions. Also if you ask him enough questions he forgets to give me work to do.

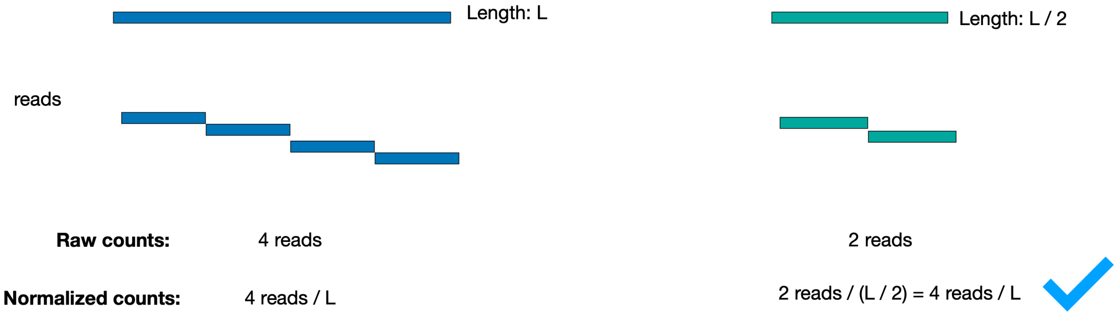

Figure 1: An example of length biasing in fragmentation.

3 Normalization

In addition, RNA-seq is a compositional technology. Let's ignore length bias for a second. For any given gene A, we never observe the true number of RNA fragments of A in a sample.[3] Instead, we observe a proportion of fragments/reads of A in a sample **relative** to the total number of fragments/reads. This means that if in our experiment we see that 10 reads come from A, we have no certainty that there are exactly 10 mRNA molecules transcribed from A. All we can say is that relative to the total number of reads r, a proportion $\frac{10}{r}$ of the reads come from gene A.[4]

Let's return to the example in the slides (Figure 2), and assume that these numbers are controlled for length bias. Can we say that G1 is differentially expressed between control and treatment? After all, G1 has 2 reads in the control, and 6 reads in the treatment. Does that mean it is upregulated? The answer here is not necessarily. The control sample here is sequenced to a depth of 10 counts, and the treatment sample is sequenced to a depth of 100 reads. That means that looking at differences in raw numbers is misleading.

Imagine, for example, that I was interested in whether taking CM122 affects whether students prefer dogs or cats. To do this, I ask 10 random people on campus whether they prefer dogs or cats. Of these, there's an even split; 5 prefer dogs, and 5 prefer cats. Next, I ask my discussion (theoretically 50 people, often less) whether they prefer dogs or cats. In my discussion, 26 people have dogs, and 23 have cats.[5] From these two samples, I conclude that taking CM122 encourages people to prefer dogs, because $26 > 5$.

Does this conclusion make sense? Absolutely not. I might even make the same conclusion about preferring cats, because $23 > 5$. The combination of these two possibilites makes no sense! Because I asked ten times more people in my discussion, of course I'm going to see that more people, in raw numbers, prefer both dogs and cats. But the composition of people who prefer dogs, $\frac{26}{50}$, in my discussion is almost exactly the same as the rate at which they prefer dogs in my random campus sampling, $\frac{5}{10}$. There is essentially no effect.

But wait - compositionally, in the control G1 is $\frac{1}{5}$ of the total sample, while under treatment G1 is $\frac{6}{100}$ of the total sample. Does that mean that G1 is differentially expressed downwards? In turn, does that mean that every gene except FG is downregulated, and FG is upregulated?

[3] Generally this is the theme of these notes: we never directly observe anything, and that has consequences for anything we do.

[4] This is even ignoring the possibility that these reads might map to the same transcript.

[5] One contrarian prefers bunnies.

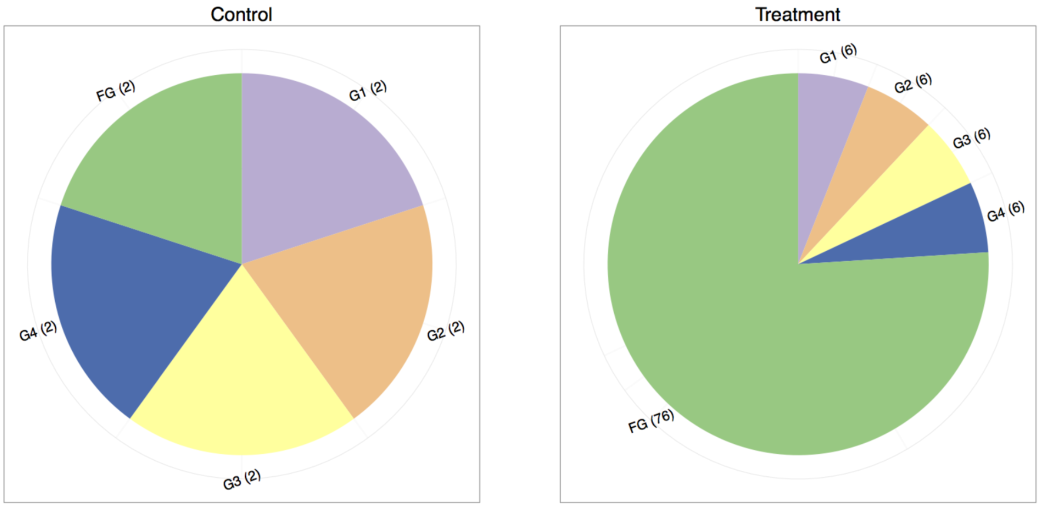

Figure 2: A control and treatment "sample", where FG dominates the composition of the treatment sample.

Intuitively...probably not. A good rule of thumb is that biology is pretty robust; the majority of genes are not going to be differentially expressed between two conditions. Also, we see that the relative proportions of G1 to G2 to G3 to G4 remain relatively constant; it's likely just FG that's screwing up all our inferences.

Here is generally where I provide a solution to our dilemma. But in normalization, there's really no solution.[6] We can't simply remove FG from our samples; that would remove most of our signal, in this case the fact that FG is significantly upregulated in the treatment sample. But the very fact that FG is upregulated means, accordingly, that our analysis of the rest of the data is necessarily skewed. Of course, there are statistical techniques that can ameliorate these issues.[7] But without direct measurement, there's no way to obtain a perfect quantification of the mRNA content of a sample. On some level, this is just how our data looks, and we have to deal with it. Like much of statistical modeling, we have to make judgement calls, and decide which compromises we're willing to make given our data.

3.1 Thought Exercises

Question 2. *Let's say we have the following "true" gene expression table, where "true" means that this is the actual number of mRNA molecules floating around (not just observed reads). Assume that we actually then observe each single mRNA molecule as one read. The table is then both "true" gene expression as well as the reads we actually observe. Also assume that each gene is the same length.*

What can we say about the genes after observing reads from these samples? What about at the (unobserved) molecule level?

As it turns out, the answer to the read-level question is exactly nothing. In Question 2, from control to treatment the expression of all genes increases by two-fold. On the molecule level, if we

[6] Harold likes to say you're fucked whatever you do.

[7] If you're really curious, a good reference would be Robinson et al 2010, or Love et al 2014.

Gene	Control	Treatment
A	6	12
B	10	20
C	15	30
D	9	18
E	20	40

knew that this was the ground truth on the molecule level we could definitively say that every gene is upregulated. But as long as this change is constant across all genes, we won't be able to detect this change at all, even if the change is ten-fold.

Question 3. *What about this table?*

Gene	Control	Treatment
A	6	12
B	10	20
C	15	30
D	9	18
E	20	20

What happens here is that every gene's "true" expression level has increased by two-fold but gene E. The ground-truth, then, is that genes A-D are upregulated.

However, our data will not reflect this. Again, because of the compositional nature of RNA-seq, gene E will be appear to be downregulated, and everything else will appear upregulated. Try it yourself by taking the percentages!

Question 4. *What about this table?*

Gene	Control	Treatment
A	6	10
B	10	14
C	15	19
D	9	13
E	20	24

In this table, every gene's expression has increased by 4. Which genes will appear to be upregulated/downregulated in this case, based on the observed read-level data? What about the unobserved molecule-level data?