COM SCI C121 Week 3

Aidan Jan

April 18, 2024

RNA-seq

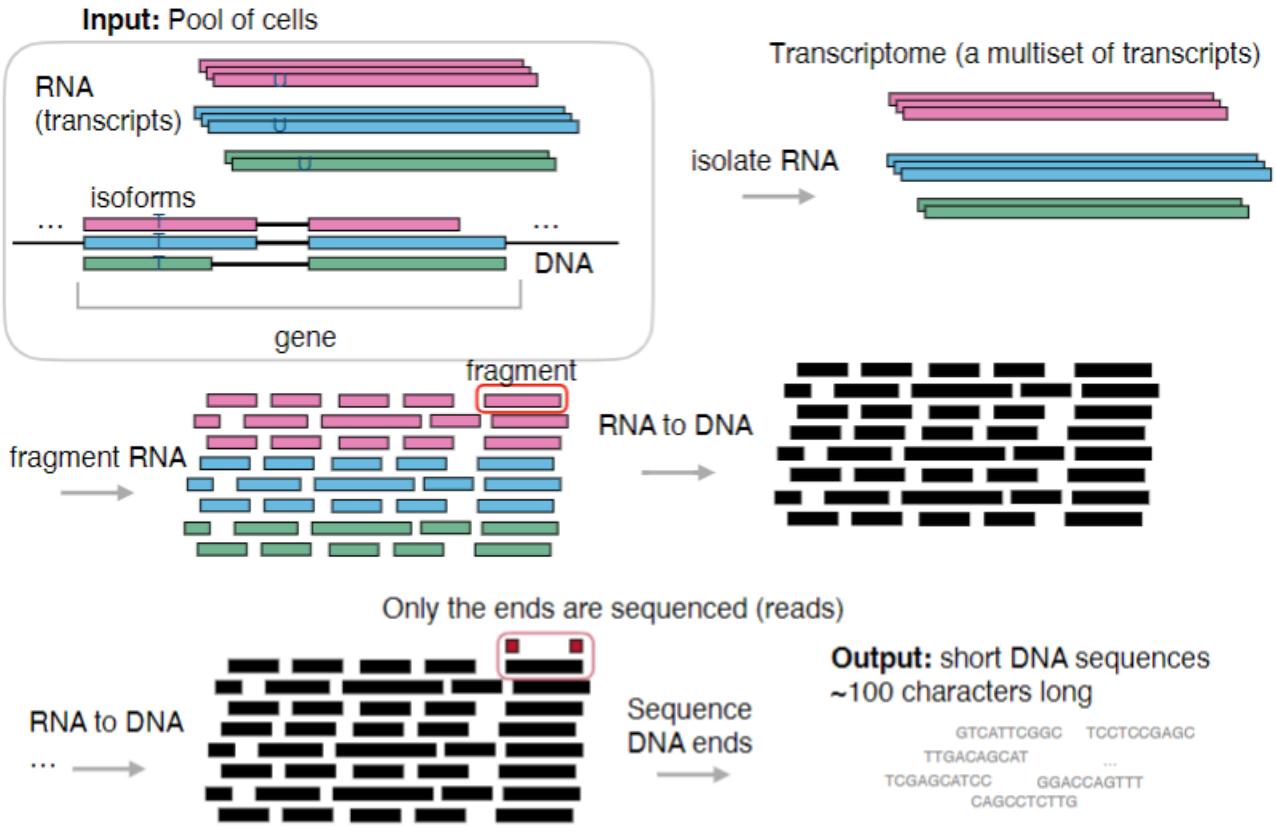
- "-seq": probing the molecular biology of the cell

A lot of times, when we want to measure some trait, the easiest way is to reduce it to sequencing, then sequence the DNA, count occurrences, and analyze.
In this case, we want to measure RNA abundance, so we use the following pipeline.

RNA abundance → cDNA Library Prep → Sequence → Estimate Abundances → Differential Analysis

The below image depicts cDNA library prep, where the RNA is copied, isolated, fragmented, then converted to DNA.

- When converting to DNA, the information to where the fragment came from on the original pool is lost. (This is why the DNA is black in the image.)



- Only the ends of each fragment is sequenced because the sequencer does not give good output when sequencing long sections.
- Also, the middle is not necessarily needed since the sequenced ends have enough base pairs for us to figure out which other fragments it connects to.

Image of Converting RNA to DNA

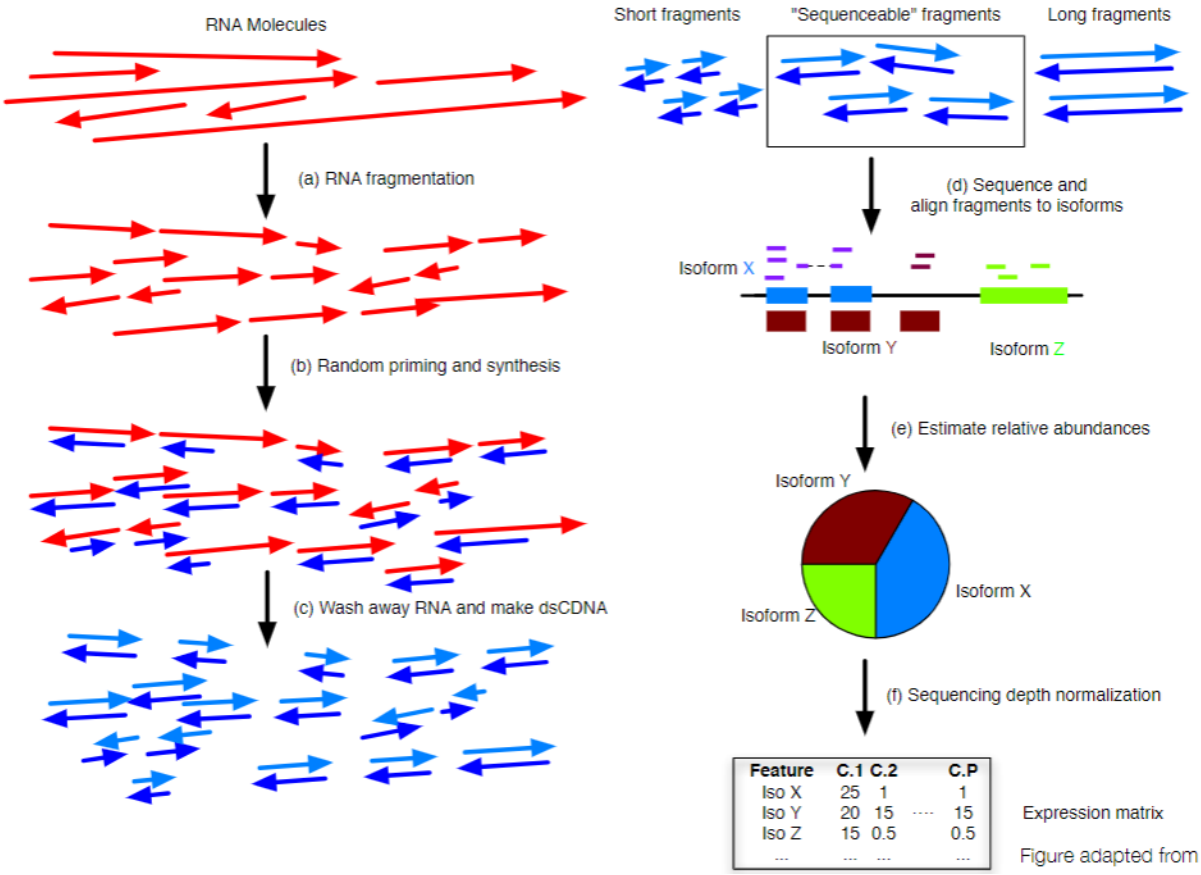
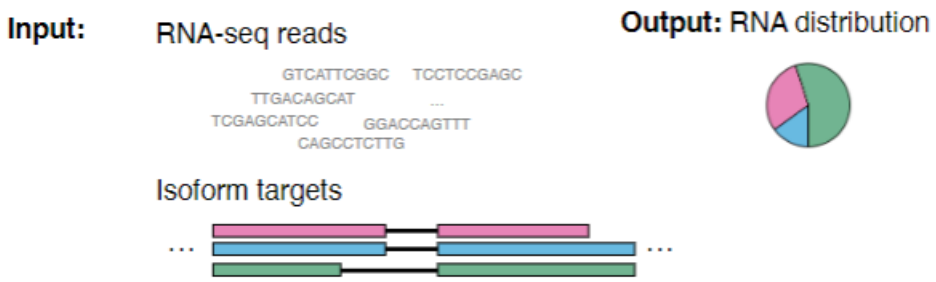


Figure adapted from Roberts et. al. 2011

- Notice that Isoform X and Isoform Y share the same DNA coding region in the image. This is referred to as an *ambiguous read* - the limit of RNA-seq. We want to know how much each X and Y there are, but since the two cover the same DNA region, sequencing cannot give you that information.
- What makes this worse is that in real life (where you are not the oracle), you don't know that the isoforms are overlapping.

RNA-seq quantification: a computational problem

Goal: given a known set on isoform targets (genes) and RNA-seq fragments, recover the distribution of RNA molecules.



All we want is that output pie chart that describes how common each isoform is.

Unlike DNA reads, where we can assume that all fragments appear at a relatively constant frequency, this is not true for RNA reads, where some isoforms may be more common than others. This makes solving the probabilities and thus the genome incredibly hard. (This is an unsolved problem.)

What is the "RNA Distribution"?

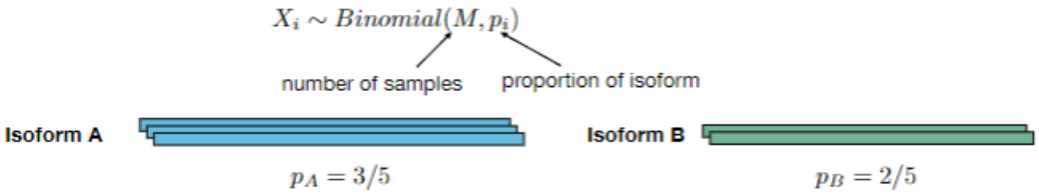
- In reality, there are a *finite* number of RNA molecules in each cell
- By nature of sequencing, we cannot directly sequence every molecule
- Instead, we mix a ton of cells together, isolate their RNA, then get a **relative measurement**.

Use cases for RNA-seq

- Tissue specific gene expression in *D. melanogaster*
- Cancer specific gene expression
- Genetic variation effects on gene expression and their relationships to tissues and complex traits

Binomial Sampling

- I'm going to sample M transcripts at random. Given the proportions below, what is a good model?



Now, suppose we have a very large M and many isoforms, where the proportion of each isoform is close to zero. What is a good model now?

- It turns out that the Poisson distribution is a good model.

$$X_i \sim Poisson(Mp_i)$$

  - where M is the number of reads (samples) and p_i is the original isoform proportion.
- This makes a strong assumption about sampling, that all the isoform lengths are the same. This is not true in reality.
- We have to normalize the number of counts for each isoform based on the length of the isoform.
  - For example, if we have isoform A with length L, and isoform B with length $\frac{L}{2}$, then we would expect half as many reads in B than A. Therefore, to normalize the number of reads, we need to scale the raw count of reads of B by a factor of 2.

Transcript per Million (TPM)

$$\text{TPM}_i = \frac{X_i}{\tilde{l_i}} \cdot \left(\frac{1}{\sum_j \frac{X_j}{\tilde{l_j}}} \right) \cdot 10^6$$

where

- X_i is the number of counts
- $\tilde{l_i}$ is the length
- $\sum_j \frac{X_j}{\tilde{l_j}}$ is the normalization constant
- 10^6 is a big number (the 'Million' part in TPM)

Assuming every site has equal probability of being sampled, what should the expectation of L squiggle be? Remember, not all fragments are of the same length. There's a fragment length probability in the expectation.

Suppose we have a transcript of length 5. Then:

- if length of fragment (F) = 3, then there are three different sites. (e.g., [0, 2], [1, 3], [2, 4])
- if length(F) = 2, then there are 4 sites.
- if length(F) = 1, then there are 5 sites.
- In general, (number of sites) = $l - length(F) + 1$

A simple model for RNA-seq

Conceptually: (let n represent the read nunmber)

1. Randomly select an isoform $I_n|p \sim \text{Categorical}(p)$
2. Randomly select a fragment length

   $$L_n|I_n = i_n \sim \text{Fragment length distribution}(\text{Length}(I_n))$$
3. Randomly select a position to generate a fragment from

   $$R_n|L_n = l_n \sim \text{Uniform}(1, \text{Length}(I_n) - l_n + 1)$$
4. Observe and repeat

What is the probability of a particular arrangement $P(r_n, l_n, i_n)$? Hint: use the Bayes Theorem.

Answer:

$$P(r_n, l_n, i_n) = P(r_n|l_n, i_n) \cdot P(l_n|i_n) \cdot P(i_n)$$

A brief aside on plate notation and graphical models

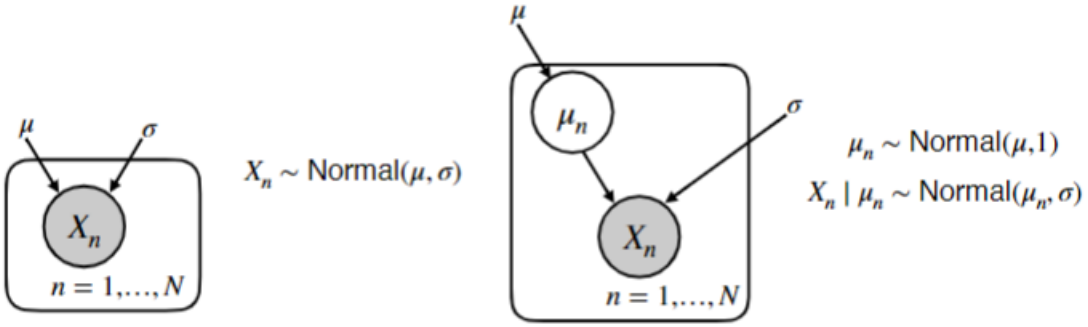
- A graphical model is a way to encode conditional dependencies in a generative model
- Circles denote random variables
- No circle means the value is fixed (a hyperparameter)
- Rectangles denote repetitions of samples
- Arrows denote conditional dependencies
  - If you know the parents, you know how to draw from that corresponding distribution

For a normal distribution, $X_n \sim \text{Normal}(\mu, \sigma)$, the likelihood function is defined as:

$$L_j(\mu, \sigma) = \prod_{n=1}^{N} P(X_n = x_n | \mu, \sigma)$$

If we create another normal distribution such that the mean is a random number chosen from another normal distribution with standard deviation 1 (see image below), we get a **random effects model**.

$$L_n(\mu, \sigma) = \prod_{n=1}^{N} P(X_n = x_n | \mu_n = u_n, \sigma) P(\mu_n = u_n | \mu, 1)$$

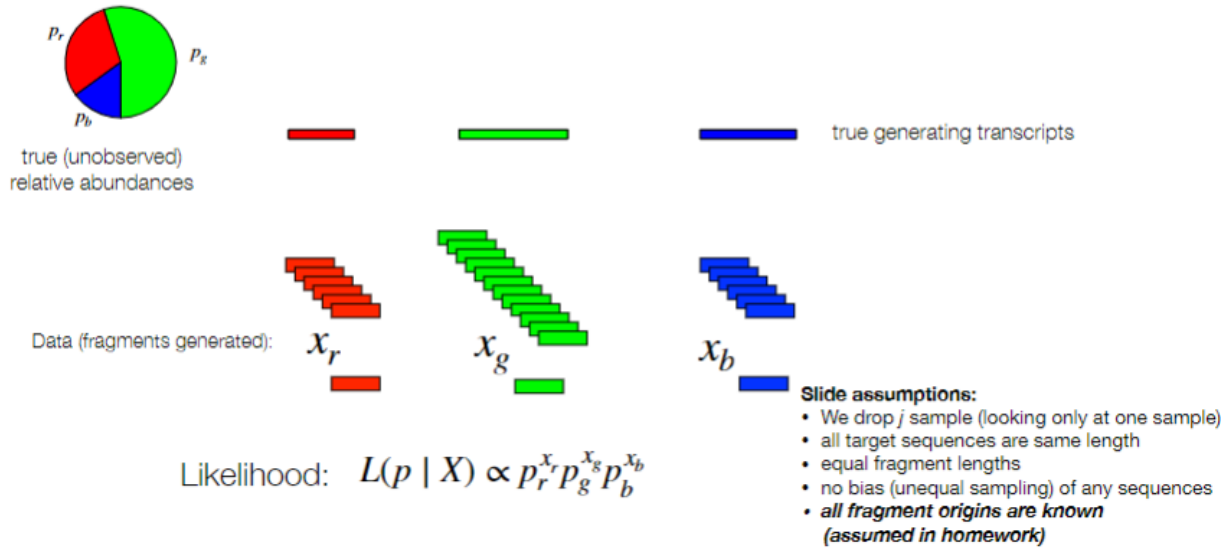


Multinomial Models and RNA-seq

- The operator generates a sequencing library as a "true" distribution (p_j)
- The sequencing *fragments* with their true, *unobserved* labeling represent X_{jg}, the counts
  - For reference, in "the real world" we plug-in E[X] rather than use X directly (because it is unobserved!)

Variables:

- j: sample index
- g: gene index
- X_{jg}: the number of fragments in gene g, sample j



Aside on Goodness of Fit

[FILL]

A heuristic argument for Goodness Of Fit (GOF)

$$Df = Z_i \sim \mathrm{N}(0,1)$$

Then,

$$\sum_{i=1}^{S} Z_i^2 \sim \chi_S^2$$

This is a Chi-squared variable. For large λ_1, $X_i \sim \text{Poisson}(\lambda)$ can be roughly approximated by $\text{Normal}(x, (\sqrt{\lambda})^2)$. To "standardize" a Gaussian, $y_i \sim \mathrm{N}(\mu, \sigma^2)$

$$Z_i = \frac{y_i - \mu}{\sigma} \sim \mathrm{N}(0,1)$$

By substitution,

$$\frac{X_i - \lambda}{\sqrt{\lambda}} \sim \mathrm{N}(0,1)$$

If we sum the square of this over all the samples, we get

$$\sum_{i=1}^{S} \frac{(X_i - \lambda)^2}{\lambda} \sim \chi^2$$

which is equivalent to chi-squared. How is this useful?

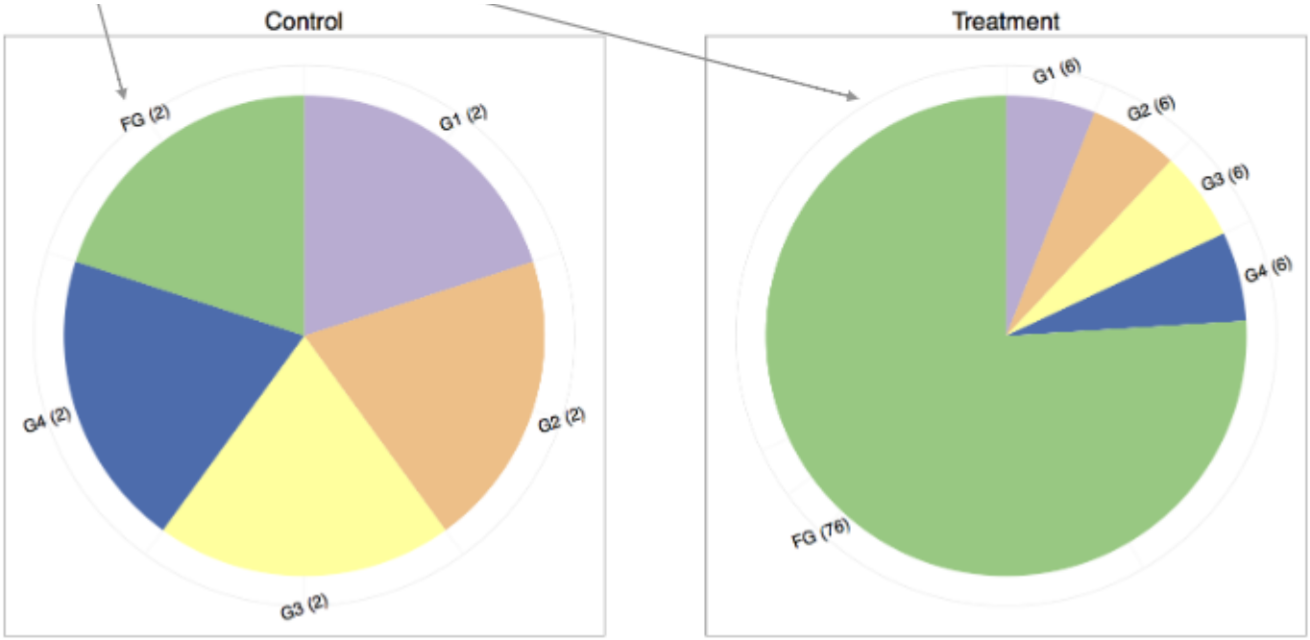
Problem with Sampling 1

- **Problem:** no matter how hard you try, it is very hard to get the same number of reads.

Problem with Sampling 2

- Sequencing is a *composition*.
  - All measurements are relative
  - They will always depend on the denominator
  - One way to think about it: you are sampling from a simplex.

One set of reads making up the bulk of the data screws everything up.



We need to normalize the data somehow.

Normalizing by GOF

Let $X_{ij} \sim \text{Poisson}(\mu)$, where

- i is the sample
- j is the isoform
- $\mu = \mathbb{E}[X_{ij}] = s_i \beta_j$
  - β_j is the proportion to true proportion
  - s_i is the size factor

Then,

$$\hat{s}_i = \frac{\sum_j x_{ij}}{\sum_k \sum_l x_{kl}}$$

where s_i represents the "total count normalization". We can fix $\hat{s}_i$ by conditioning on a set S of gels/isoforms that don't "change". To do this, we need to find some S that does not change between samples. In other words, $\mu_i = \mu_j$.

$$X_{ij} \sim \text{Poisson}(s_i \beta_j)$$

and from before, $\hat{\beta}_j = \sum_n^N x_{nj}$.
Recall that $\mathbb{E}[X_{ij}] = \hat{s}_i \hat{\beta}_j$. Squaring and summing over the samples would give:

$$\frac{(O_i - E_i)^2}{E_i} \Rightarrow \frac{\sum_{i=1}^N (X_{ij} - \hat{s}_i \hat{\beta}_j)^2}{\hat{s}_i \hat{\beta}_j} = T_j$$

Note that T_j is always positive, and is very small when there is close to no error. We can then use T_j values calculated from the estimated β_j values to create a distribution, which can be used to estimate the actual data.

- The tails of the distribution are chopped off because they are not as accurate.