

# Lecture 7: The RNA quantification problem



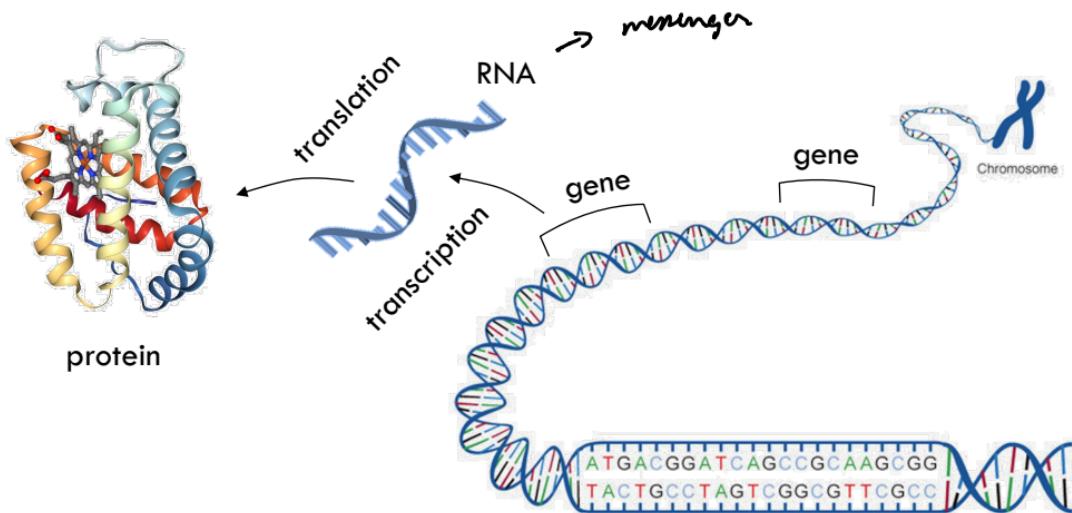
ECE 365

## Announcements:

- Lab 3 deadline moved to Friday 03/26
- Lab 3 session moved to Thursday 03/25
- Quiz 1 tonight at 7pm
  - Material includes everything up to (and including) GWAS
  - You are allowed one double-sided sheet of handwritten notes
  - You are allowed to use a calculator

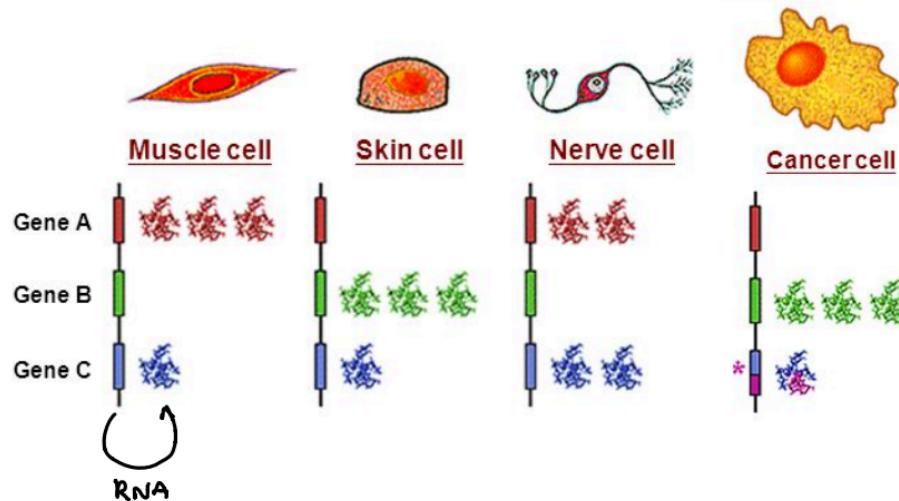
# The central dogma of molecular biology

- DNA → RNA → protein



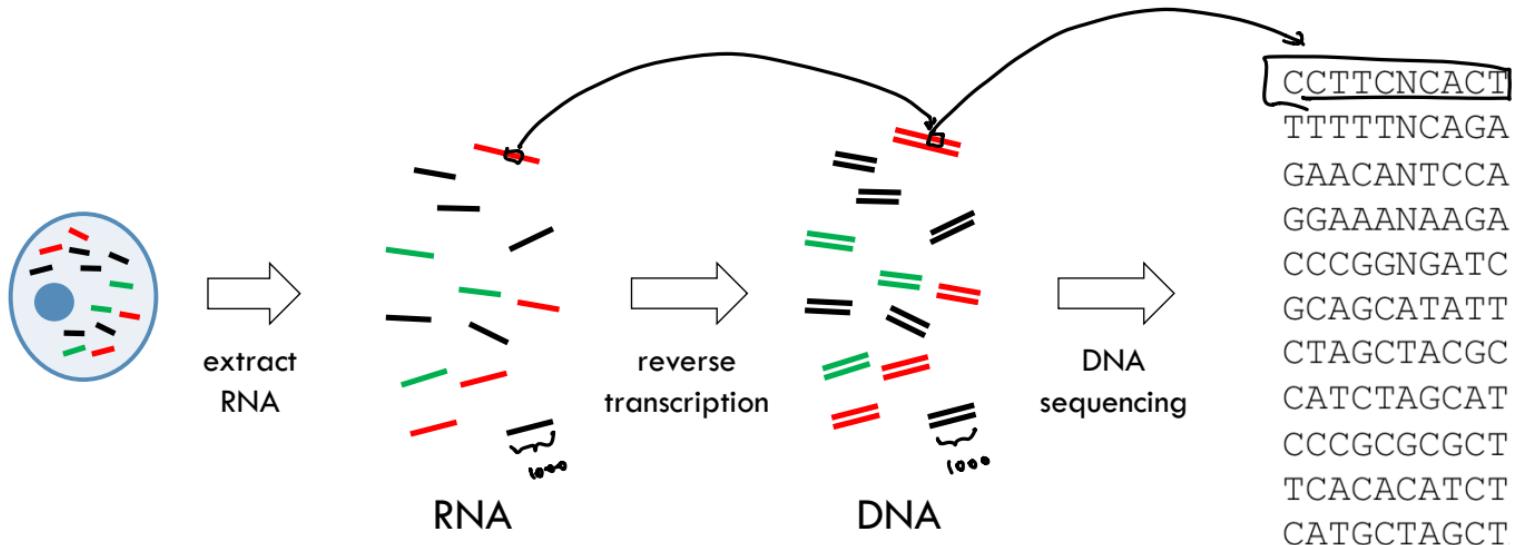
# Why do we care about RNA?

- Different genes are expressed in different cell types



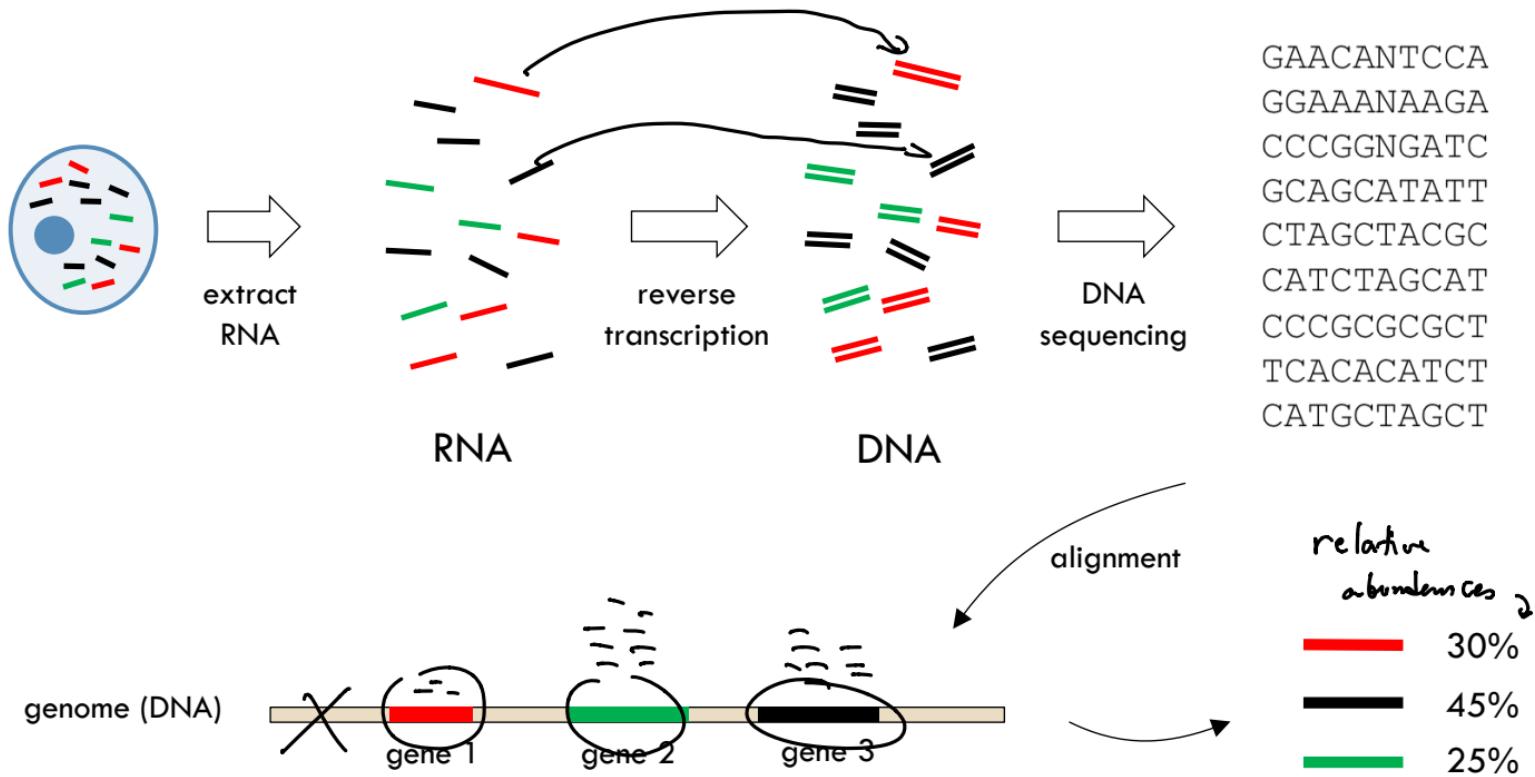
- RNA is an “intermediate step” between genes and proteins
- RNA levels in a cell can tell us which genes are “on/off”

# How do we sequence RNA?

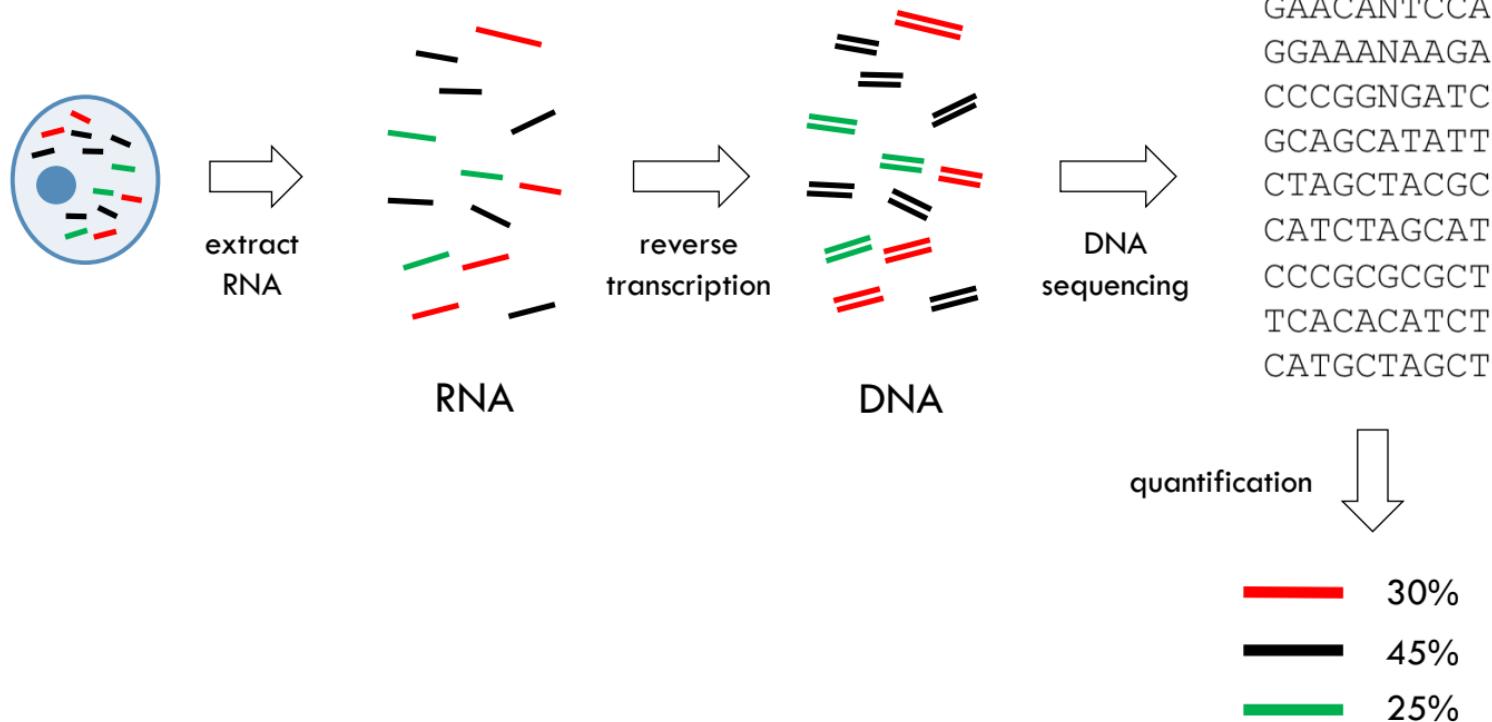


RNA - seq

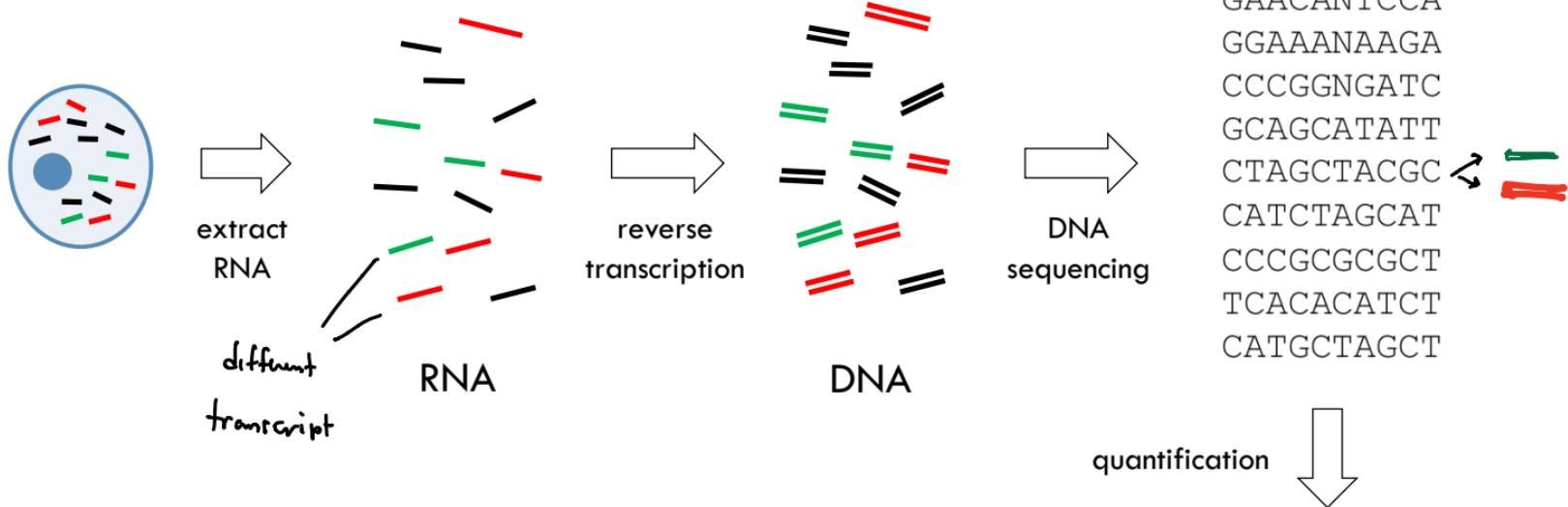
# Our first goal: RNA quantification



# Our first goal: RNA quantification



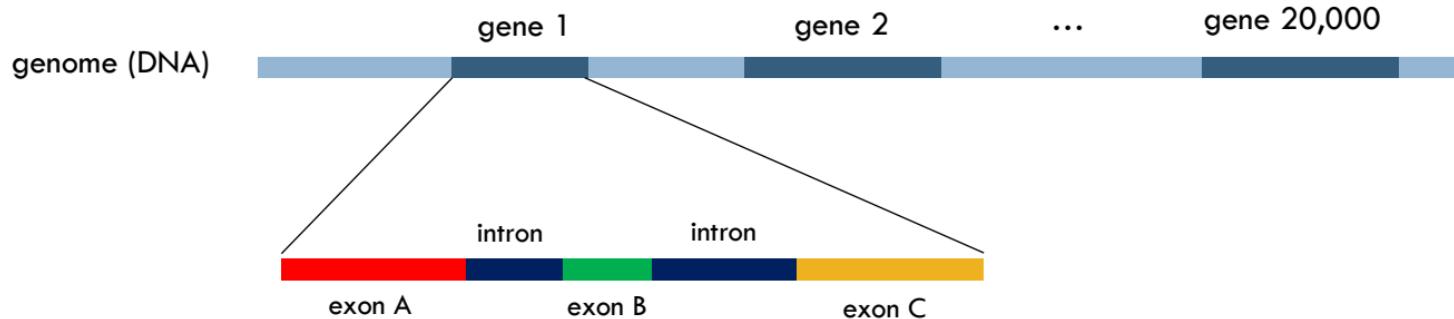
# Our first goal: RNA quantification



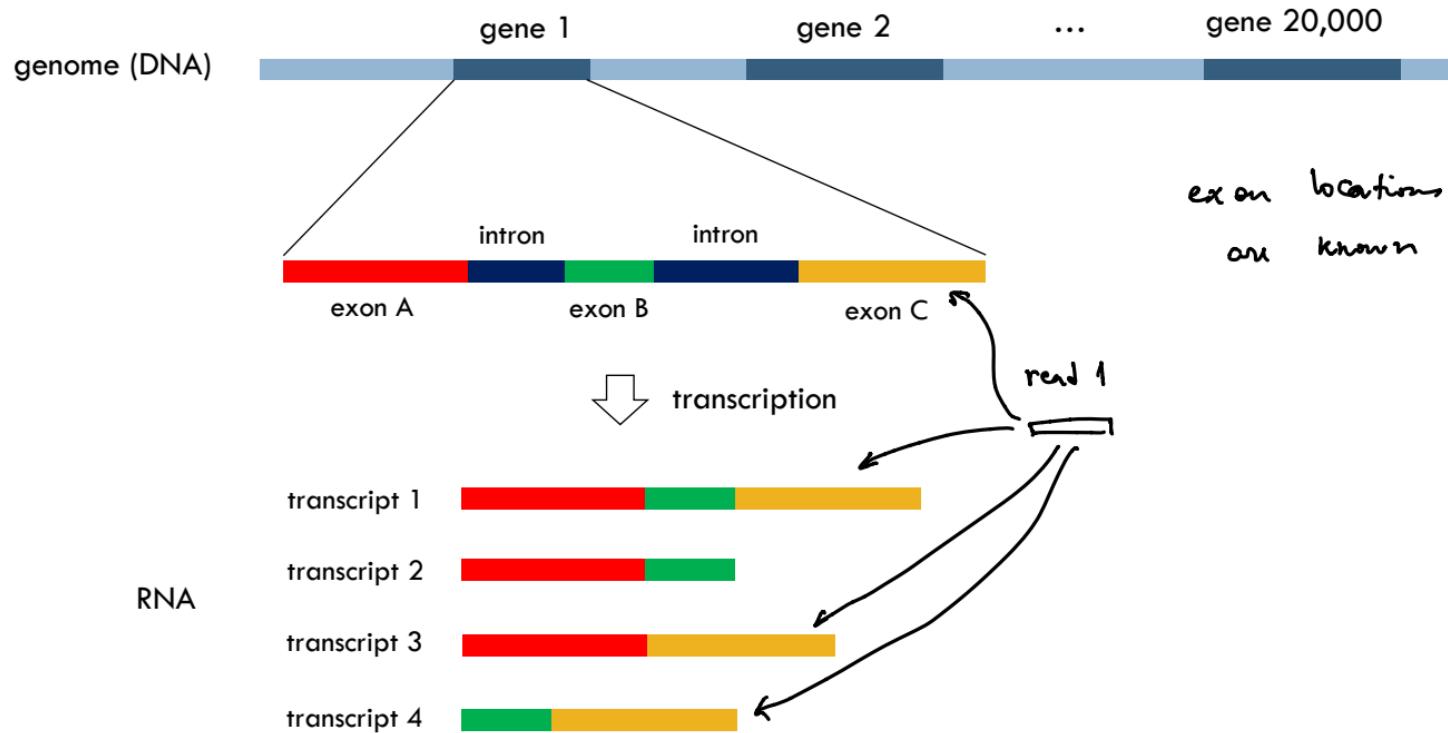
Why is this hard? Because transcripts may “look alike”



# From genes to transcripts



# From genes to transcripts



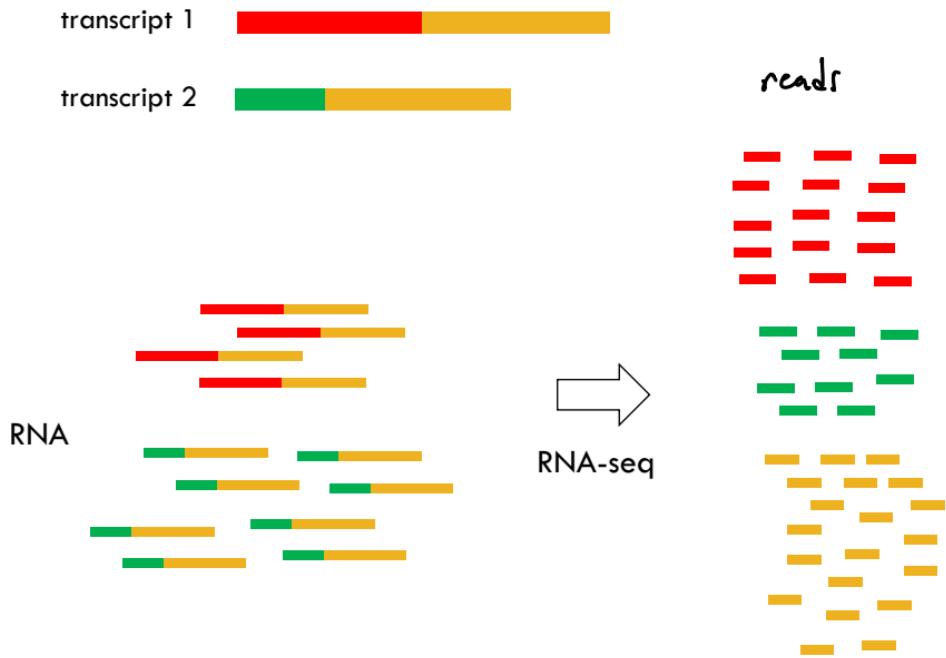
# RNA (transcript) quantification problem

transcript 1      

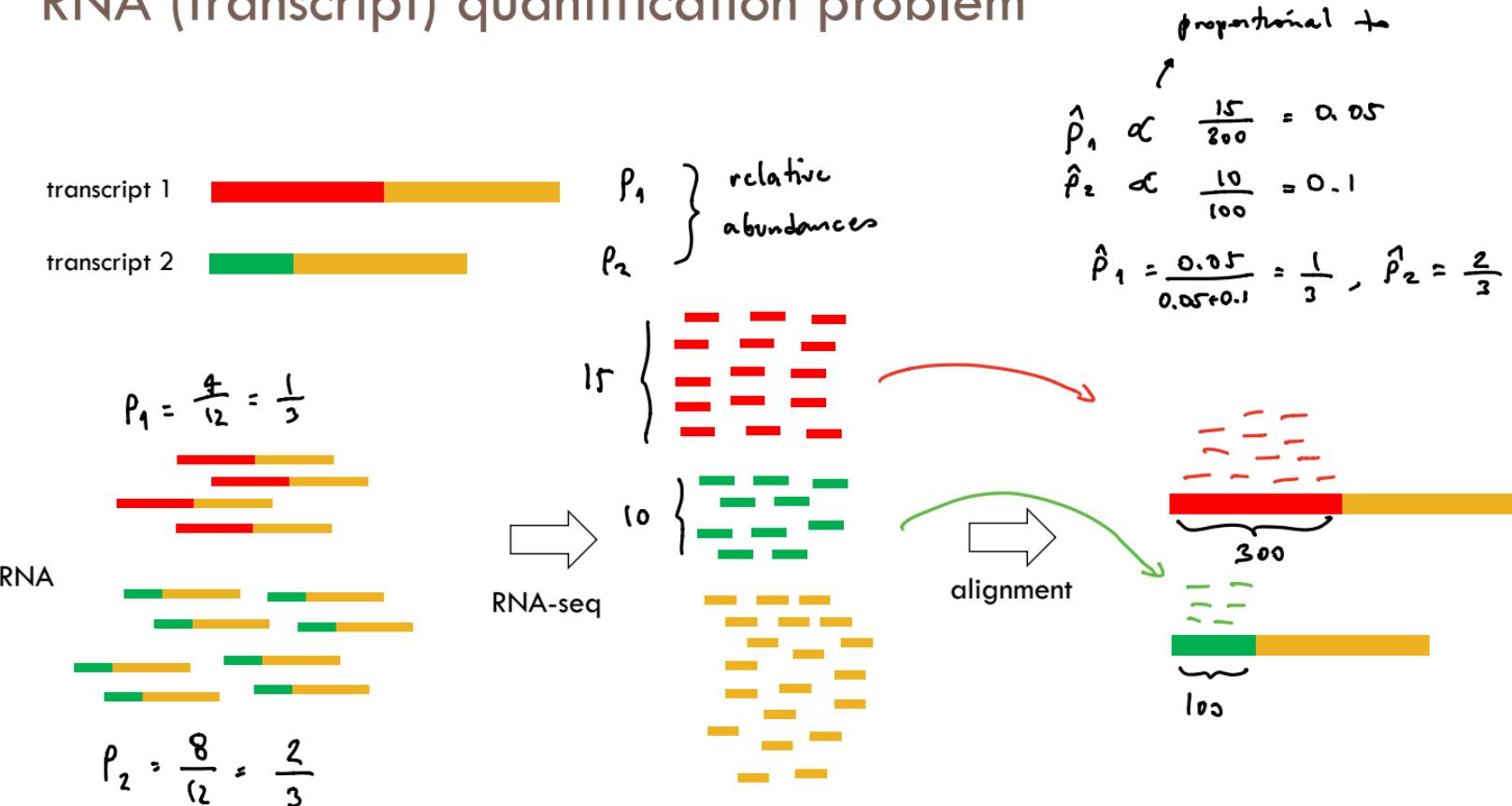
transcript 2      



# RNA (transcript) quantification problem



# RNA (transcript) quantification problem



# RNA (transcript) quantification problem

transcript 1      

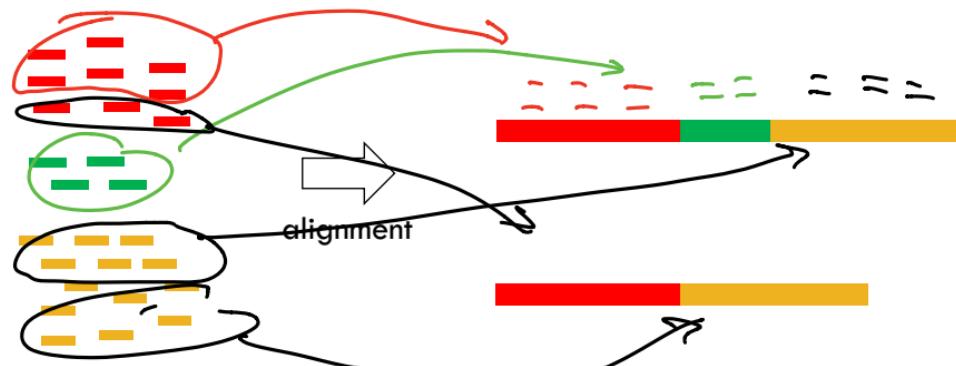
transcript 2      

$$p_1 = \frac{4}{10}$$

RNA

$$l_2 = \frac{6}{10}$$

RNA-seq



# RNA (transcript) quantification problem

transcript 1



transcript 2



transcript 3



RNA

A collection of horizontal bars representing different transcripts. The segments within each bar vary in length and position, illustrating the complexity of the RNA sample.

RNA-seq

A grid of short vertical lines of varying heights, representing the sequencing process where many short DNA or RNA fragments (reads) are sequenced.

A grid of short vertical lines corresponding to the RNA-seq grid, representing the alignment of the short reads to the longer RNA molecules.

alignment

A grid of short vertical lines where multiple lines point to the same position, indicating that a single RNA molecule has been mapped to multiple locations, which is referred to as "no unique alignment".

A single horizontal bar divided into segments, representing the final aligned RNA molecule. The segments correspond to the RNA shown at the top of the diagram.

# General RNA quantification problem

- Input: aligned read data

	tr. 1	tr. 2	...	tr. K	
read 1	0	1	1	0	0
read 2	0	0	0	1	0
	0	1	1	0	0
:	0	0	0	0	1
	1	0	0	0	0
	1	0	0	0	1
read N	0	1	0	0	1

relative abundances

$\hat{p}_1 = 0.01$

$\hat{p}_2 = 0.04$

$\vdots$

$\hat{p}_K = 0.02$

$\Rightarrow$

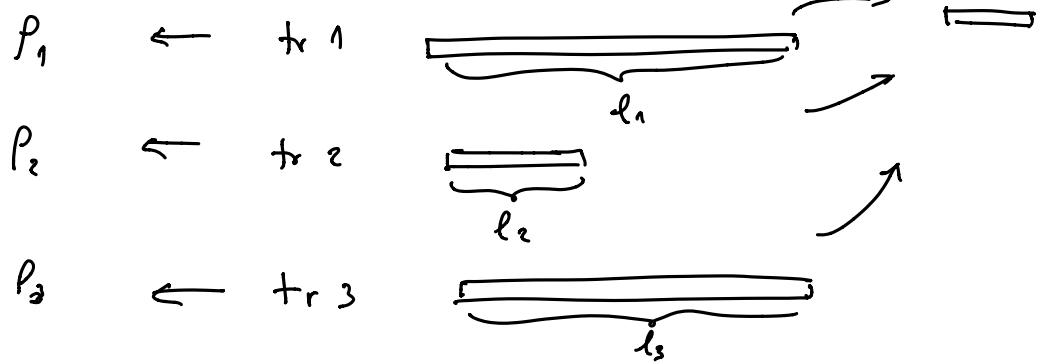
means: read N maps to tr. 2

General approach to solve this:

Expectation - Maximization Algorithm (EM)

Model for read data generation:

relative abundances (unknown)



$$P(\text{read 1 comes from tr. } k) = \frac{P_k l_k}{\sum_{j=1}^k P_j l_j}$$

True read origin/assignment (unknown)

$$z_{ik} = \begin{cases} 1 & \text{if read } i \text{ comes from tr. } k \\ 0 & \text{otherwise} \end{cases}$$

form a  
matrix

$K$  transcripts

$N$  reads

$$\left( \begin{array}{cccc} 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{array} \right)$$

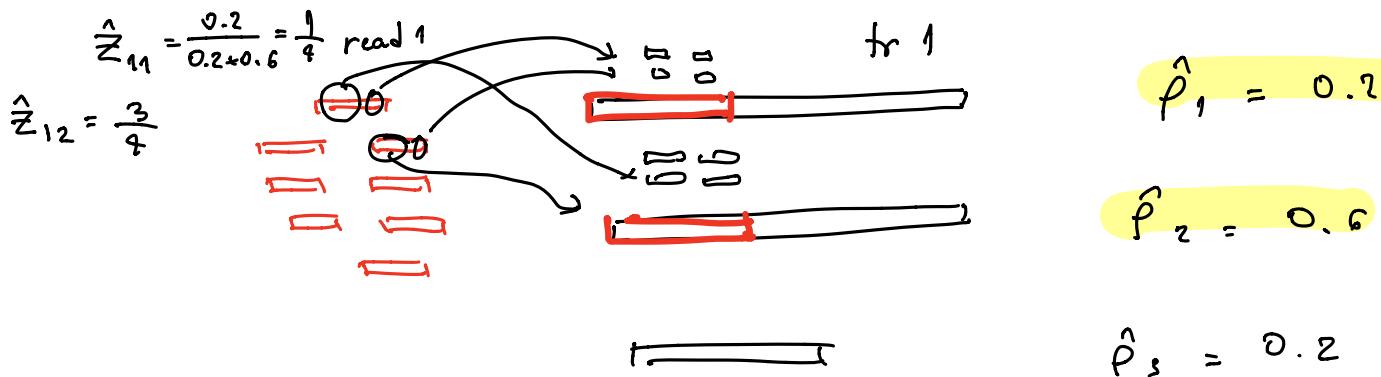
(one 1 per row)

→ "cleaned up" version  
of data matrix

EM strategy: Estimate  $\hat{P}_k$  and  $\hat{z}_{ik}$  in an  
iterating fashion and recursively

Initialize :  $\hat{p}_k = \frac{1}{K}$  for  $k = 1, \dots, K$

Based on  $\hat{p}_k$ 's, estimate  $\hat{z}_{ik}$ 's :



Let  $S_i$  be the set of transcripts read in major to

$$\hat{z}_{ik} = \begin{cases} \frac{\hat{p}_k}{\sum_{j \in S_i} \hat{p}_j} & \text{if } k \in S_i \\ 0 & \text{otherwise} \end{cases}$$