Lecture 7: The RNA quantification problem

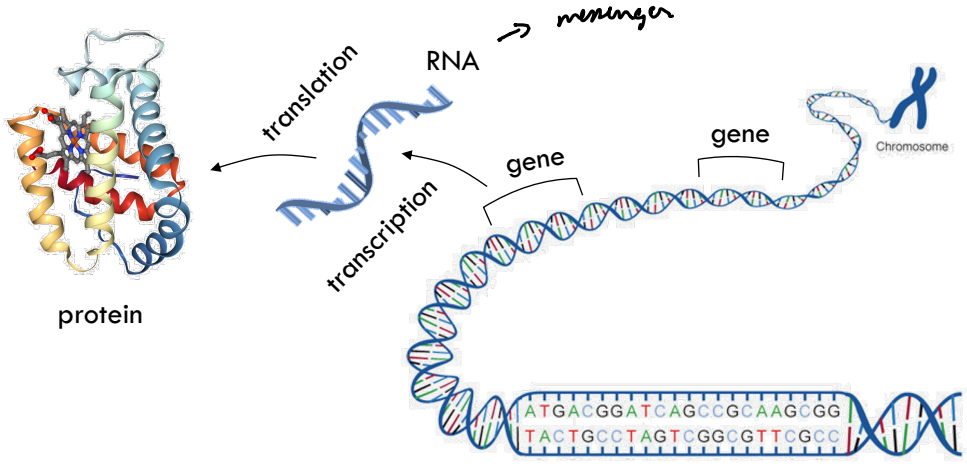
ECE 365

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

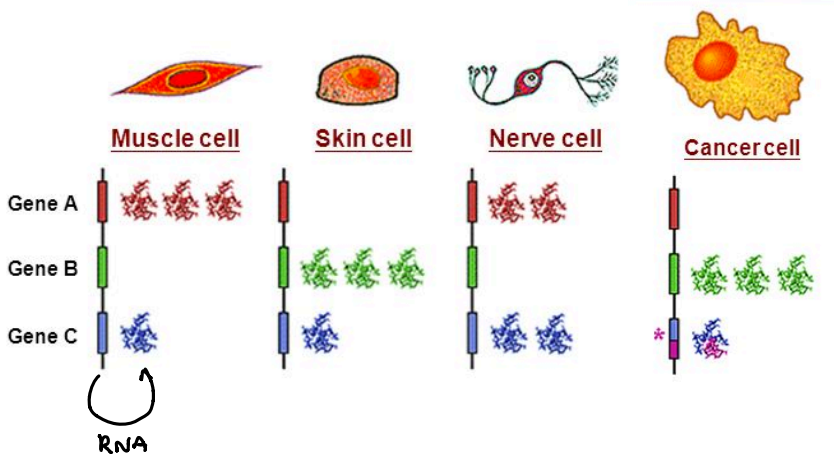
The central dogma of molecular biology

- DNA → RNA → protein



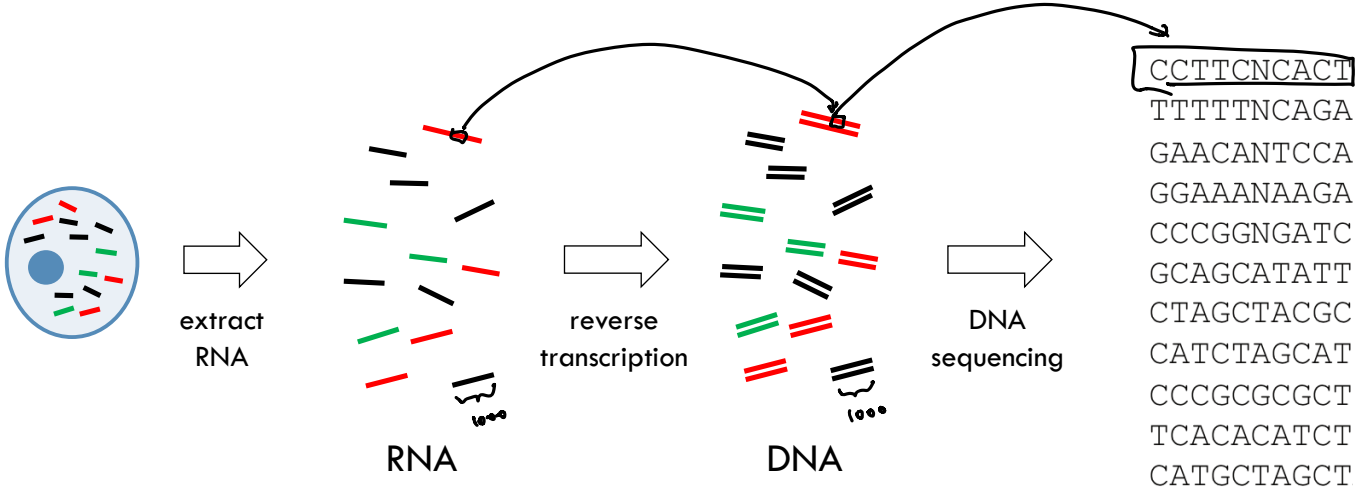
Why do we care about RNA?

- Different genes are *expressed* in different cell types



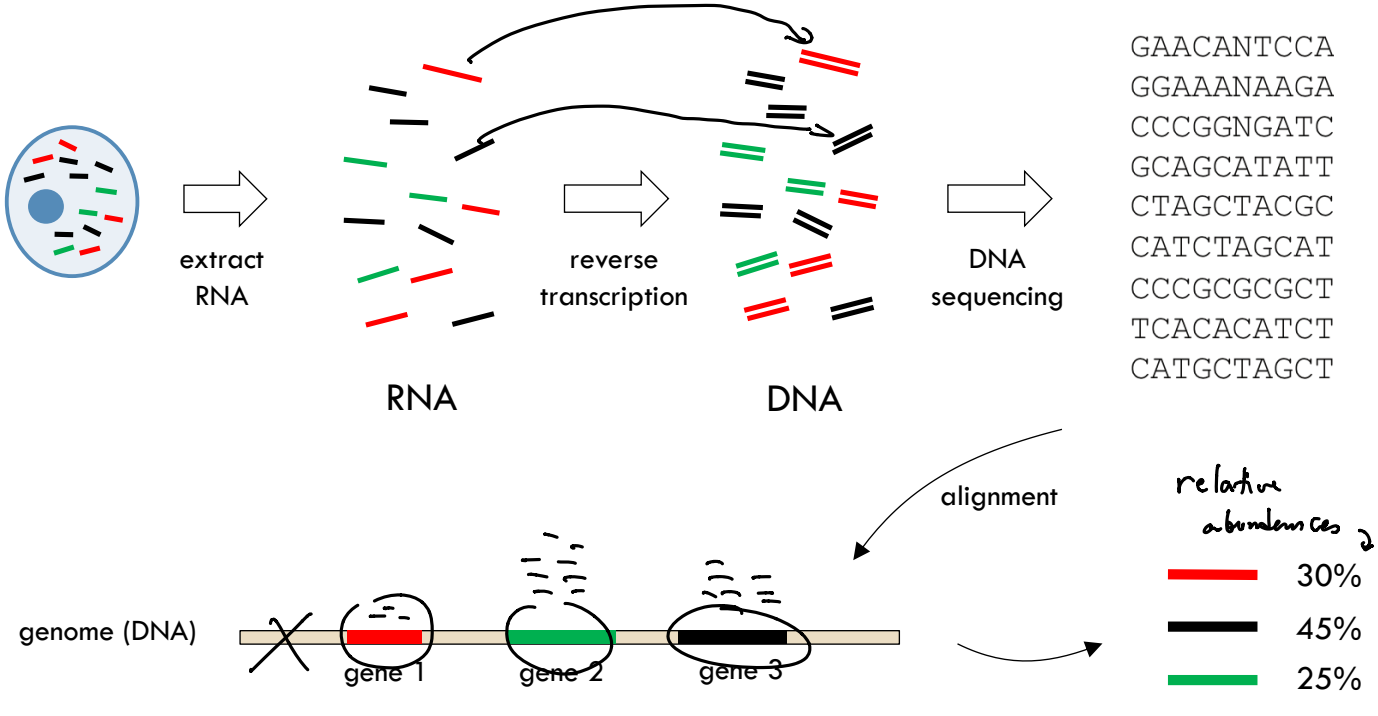
- RNA is an "intermediate step" between genes and proteins
- RNA levels in a cell can tell us which genes are "on/off"

How do we sequence RNA?

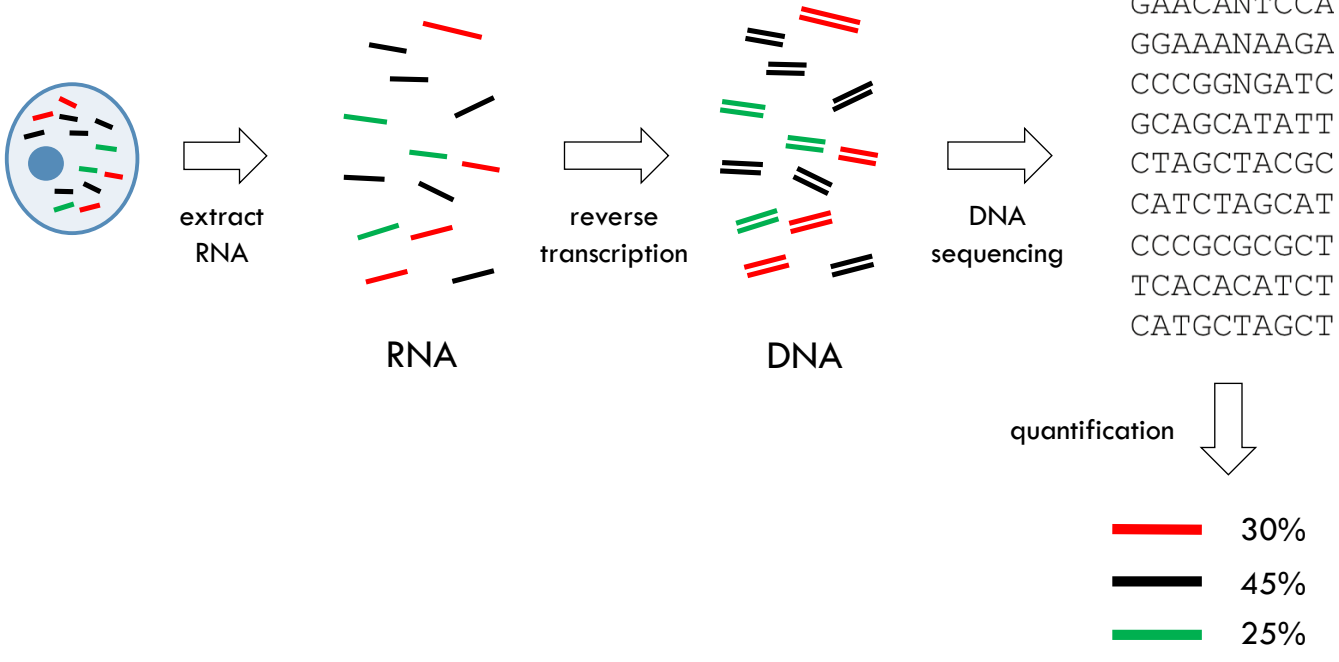


RNA-seq

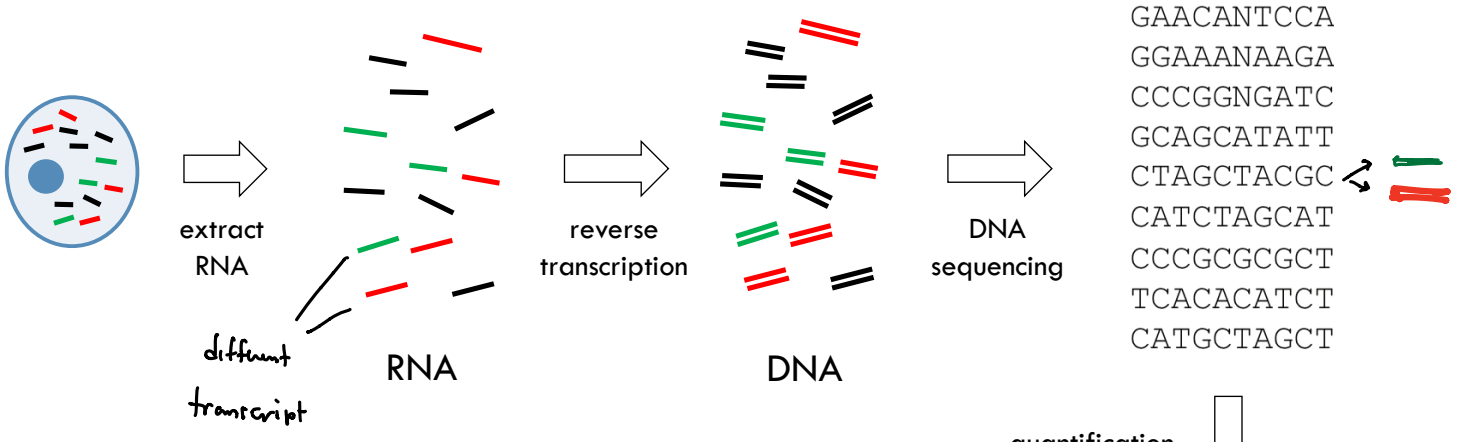
Our first goal: RNA quantification



Our first goal: RNA quantification

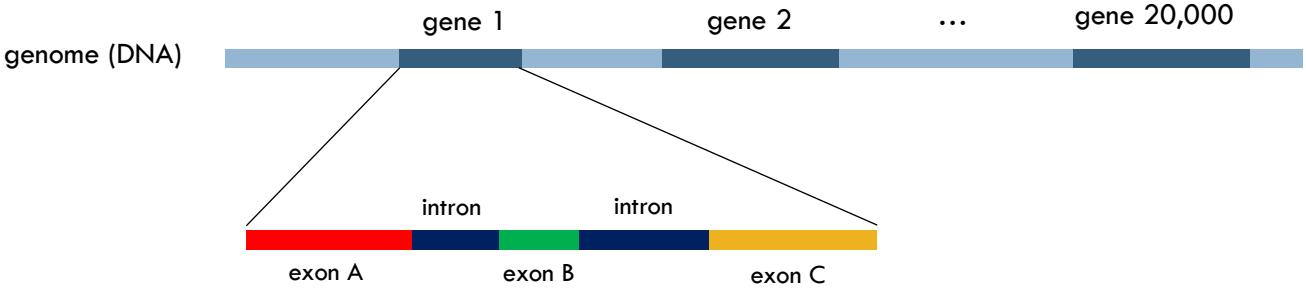


Our first goal: RNA quantification

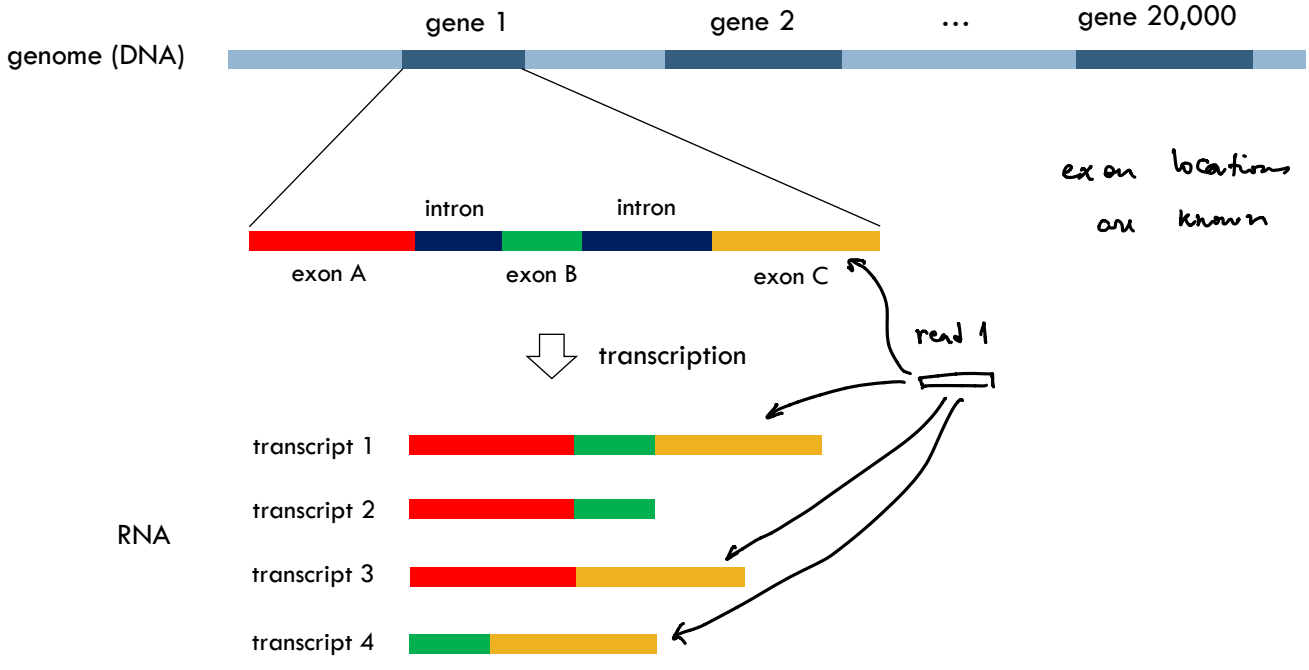


Why is this hard? Because transcripts may "look alike"

From genes to transcripts

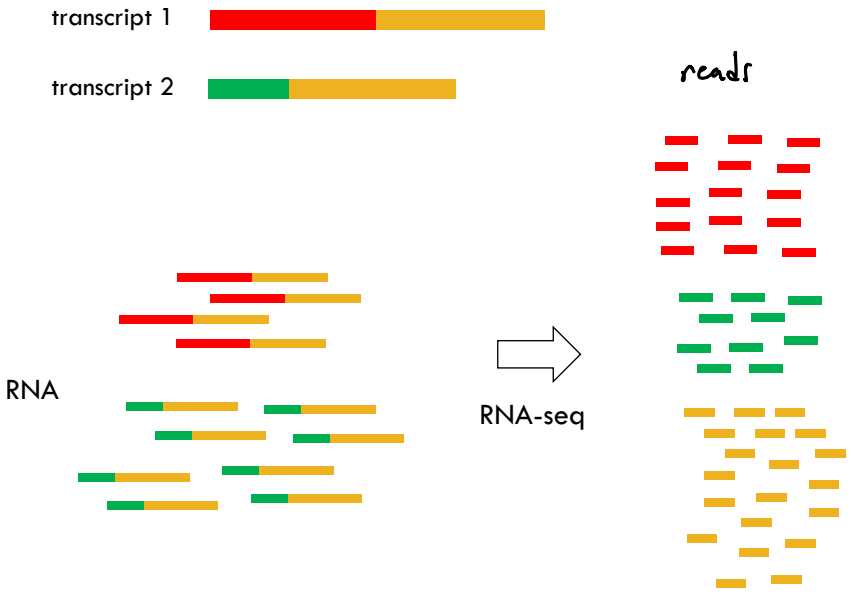


From genes to transcripts

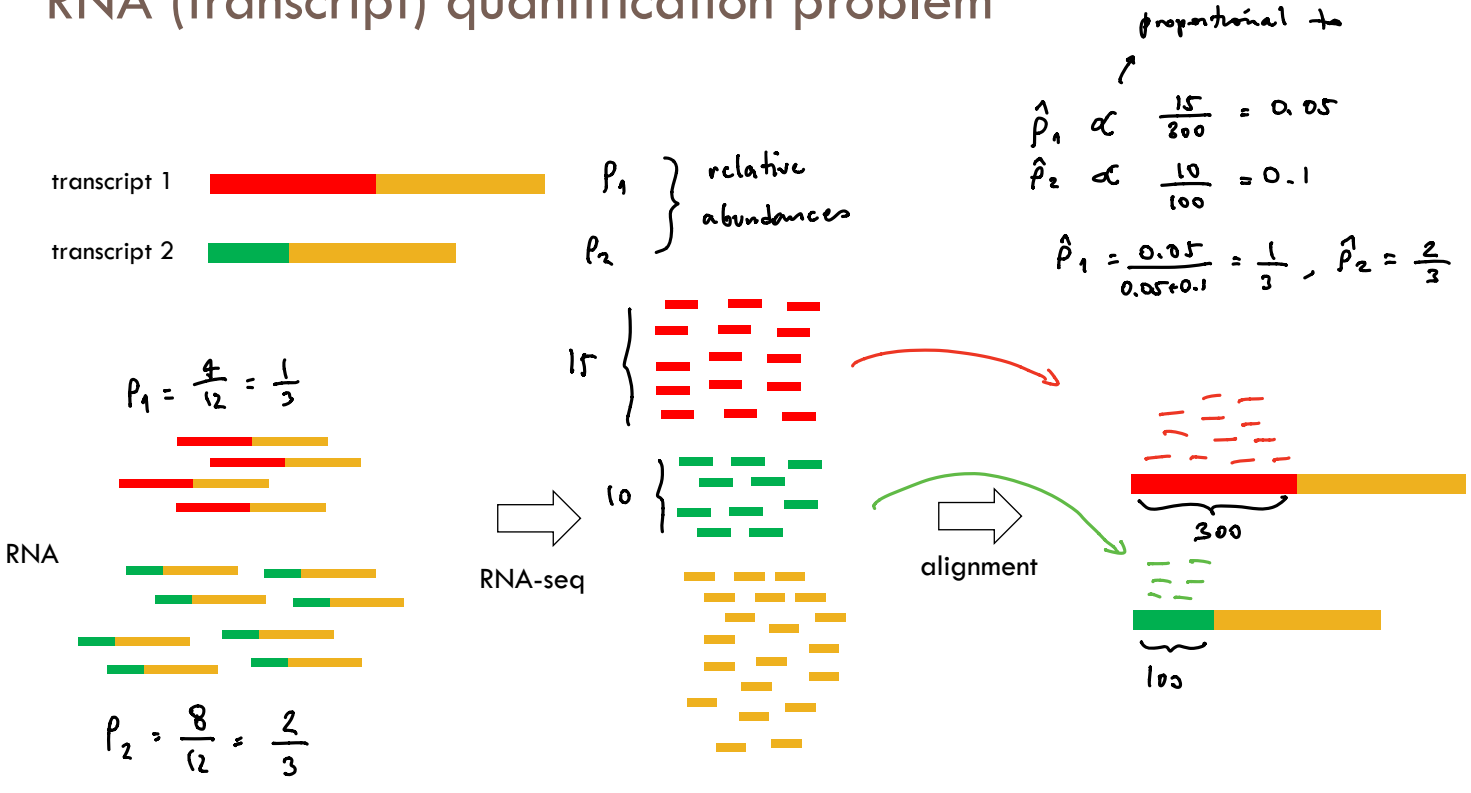


RNA (transcript) quantification problem

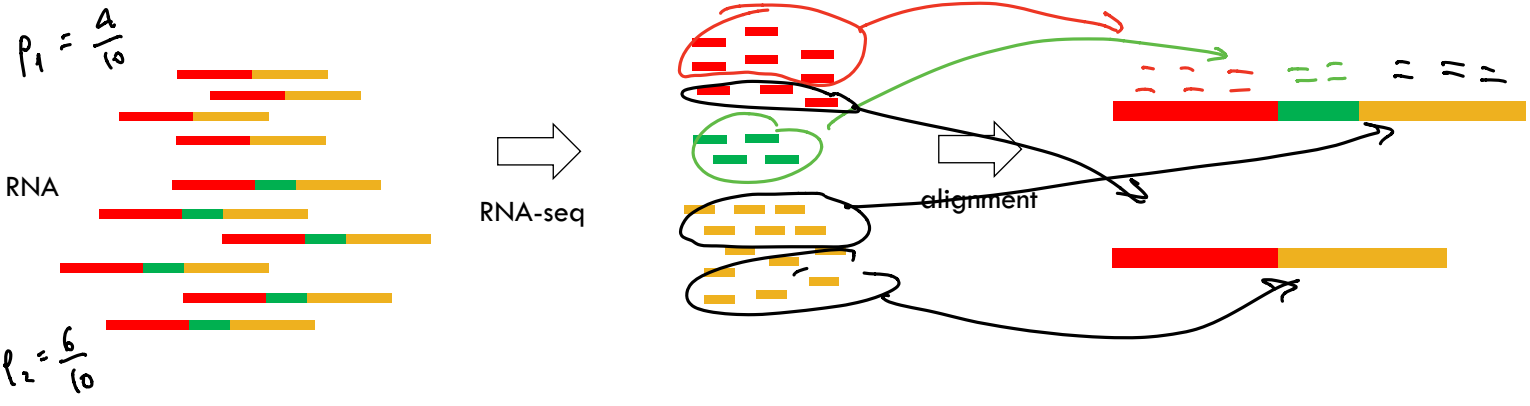
RNA (transcript) quantification problem



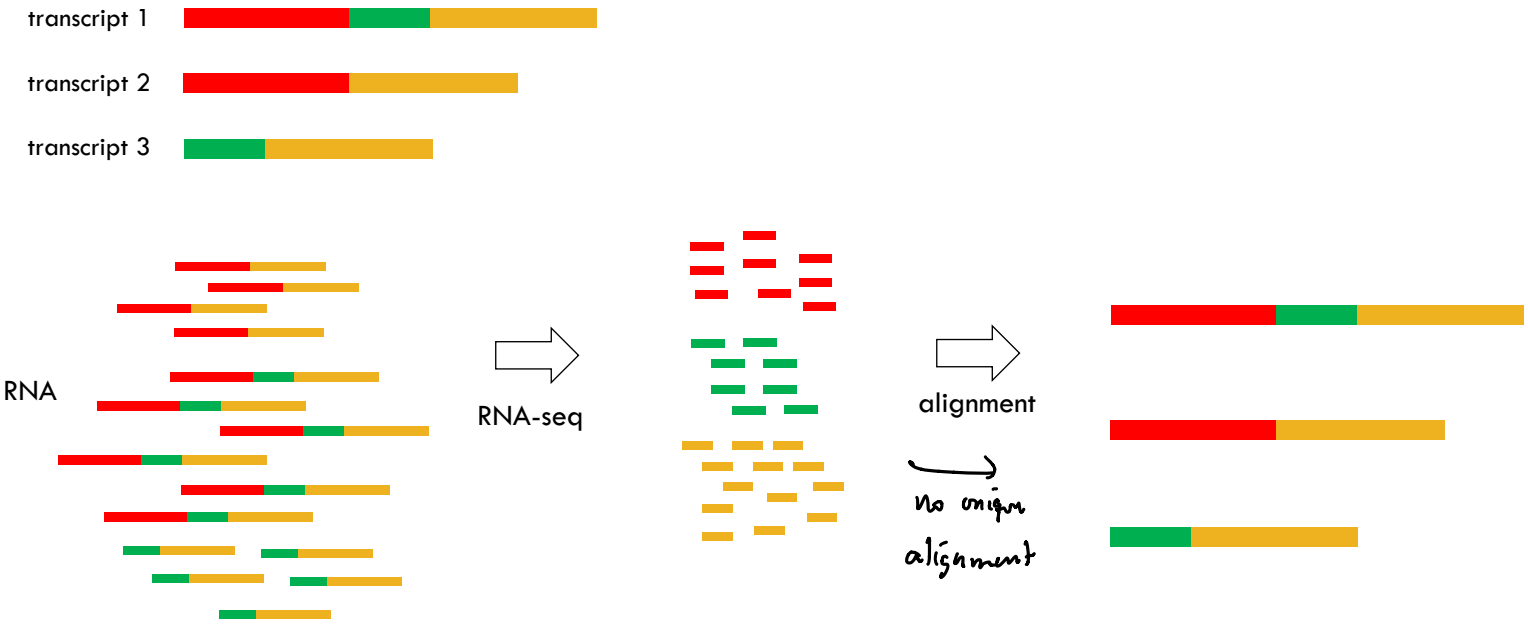
RNA (transcript) quantification problem



RNA (transcript) quantification problem



RNA (transcript) quantification problem



General RNA quantification problem

- Input: **aligned read data**

relative abundances

	tr. 1	tr 2	⋯		tr. K
read 1	0	1	1	0	0	0
read 2	0	0	0	1	0	0
	0	1	1	0	0	0
⋮	0	0	0	0	1	0
	1	0	0	0	0	0
	1	0	0	0	1	0
read N	0	1	0	0	0	1

$\Rightarrow$

$\hat{\rho}_1 = 0.01$

$\hat{\rho}_2 = 0.04$

⋮

$\hat{\rho}_K = 0.02$

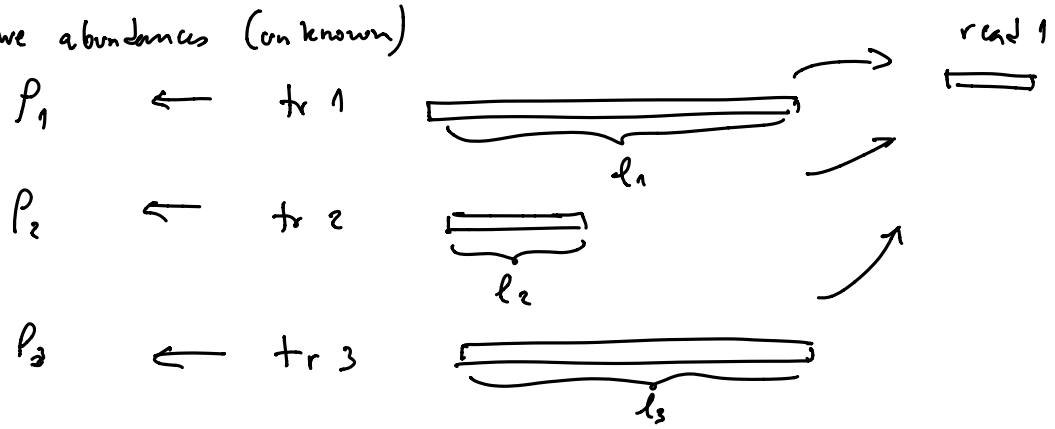
means: read N maps to tr. 2

General approach to solve this:

Expectation - Maximization Algorithm (EM)

Model for read data generation:

relative abundances (unknown)



$$P(\text{read 1 comes from tr. } k) = \frac{\rho_k \ell_k}{\sum_{j=1}^{K} \rho_j \ell_j}$$

True read origins/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

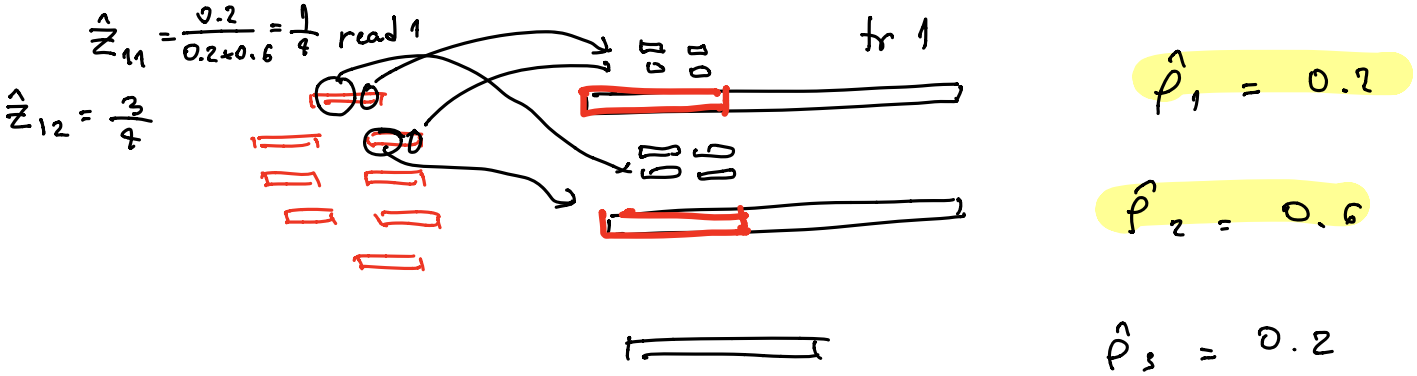
N reads $\left\{ \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \right.$ (one 1 per row)

→ "cleaned up" version of data matrix

EM strategy: Estimate $\hat{\rho}_k$ and $\hat{Z}_{ik}$ in an alternating fashion and recursively

Initialize : $\hat{\rho}_k = \frac{1}{K}$ for $k = 1, \ldots, K$

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



Let S_i be the set of transcripts read i maps to

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