Lecture 6: Finish GWAS + Intro to RNA sequencing

ECE 365

Announcements:

- Lab 3 (GWAS) released (due 03/25)
- Quiz 1 – 03/23
  - Material includes everything up to (and including) GWAS

Revisiting Logistic Regression

- Can we use Logistic Regression for GWAS?

- Problem: Number of SNPs can be $\sim 10^6$

GWAS via *univariate* logistic regression

- Run a separate logistic regression for each SNP

	SNP 1			x_i	SNP m		phenotype
person 1	0	1	1	0	0	1	0
	1	0	2	1	0	2	0
	1	1	1	0	0	0	1
	0	0	0	2	1	1	1
	1	1	1	0	0	0	1
	0	2	0	0	1	0	0
person n	1	1	0	1	0	2	1

$$p(1 \mid x_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_i x_i)}}$$

captures association between i^{th} SNP and phenotype

- Use this to identify small subset of SNPs associated with phenotype
- Let's look at some examples on a Jupyter notebook

GWAS via *univariate* logistic regression

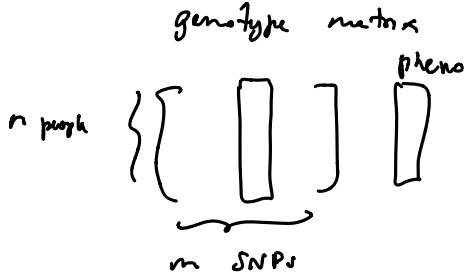
- Idea: combine all beta coefficients into a single model:

$$\underbrace{\ln\left(\frac{p(1|x)}{1-p(1|x)}\right)}_{\text{LOR}} = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$

- Problems with this approach:
  - Since m is very large, some β_is will be large **by chance**
  - Some x_is are correlated

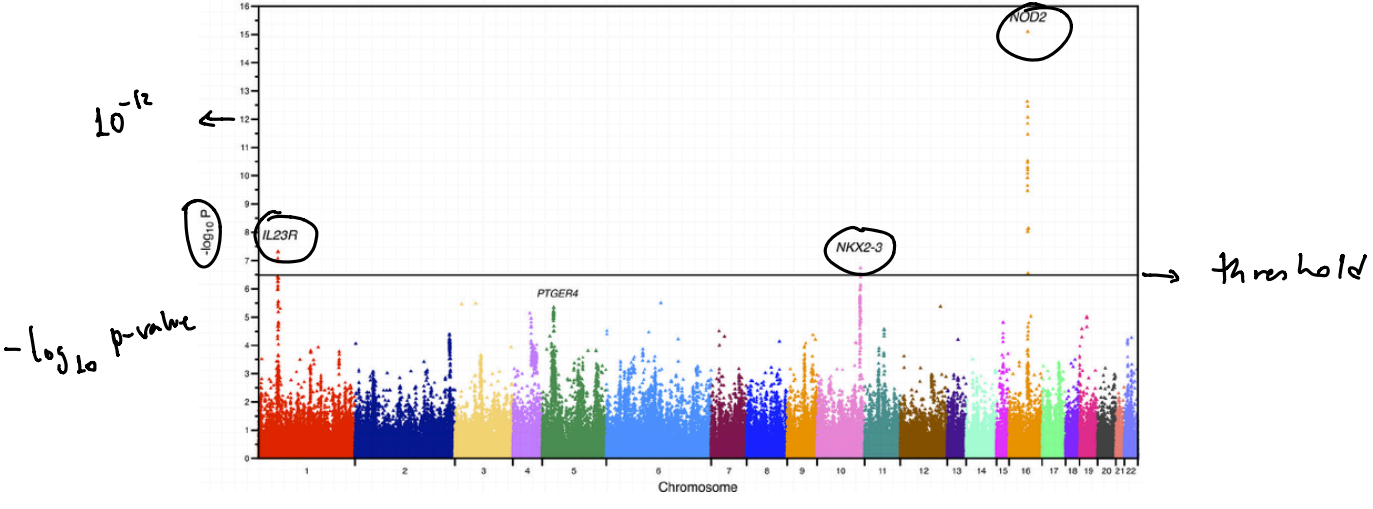
(e.g., anyone with $x_3 = 1$ has $x_4 = 1$)

Manhattan plots



- Allow us to see the significance of all SNPs in the genome
- We plot $-\log_{10}(p-\text{value})$

Manhattan plot of SNP association with Irritable Bowel Disease



GWAS via *univariate* logistic regression

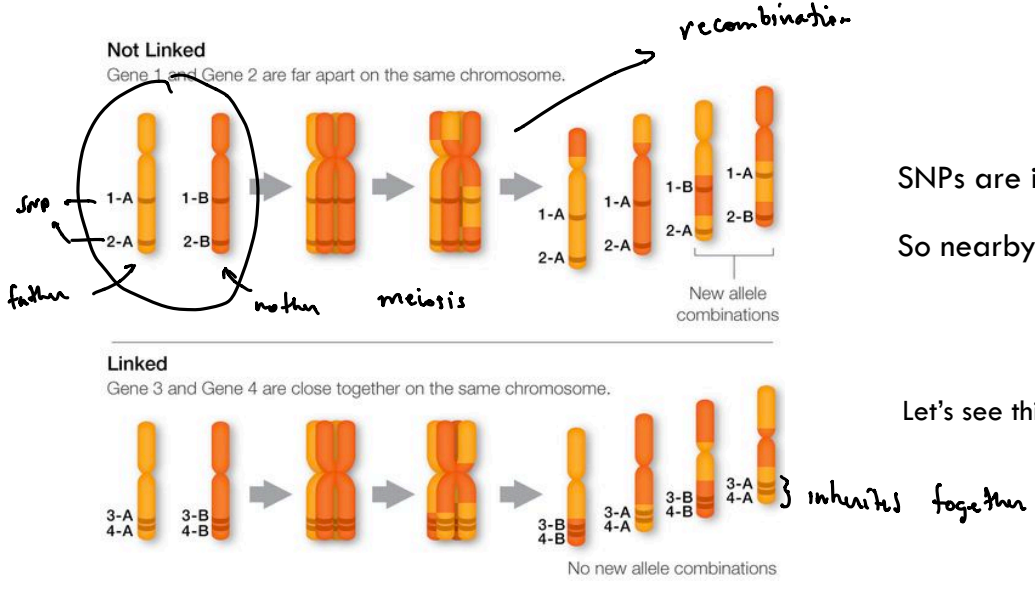
- Idea: combine all beta coefficients into a single model:

$$\ln\left(\frac{p(1|x)}{1-p(1|x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$

- Problems with this approach:
  - Since m is very large, some β_is will be large **by chance** (use p-values to select β_is)
  - Some x_is are correlated

Correlation between SNPs

- SNPs can be correlated due to genetic linkage



SNPs are inherited in blocks.

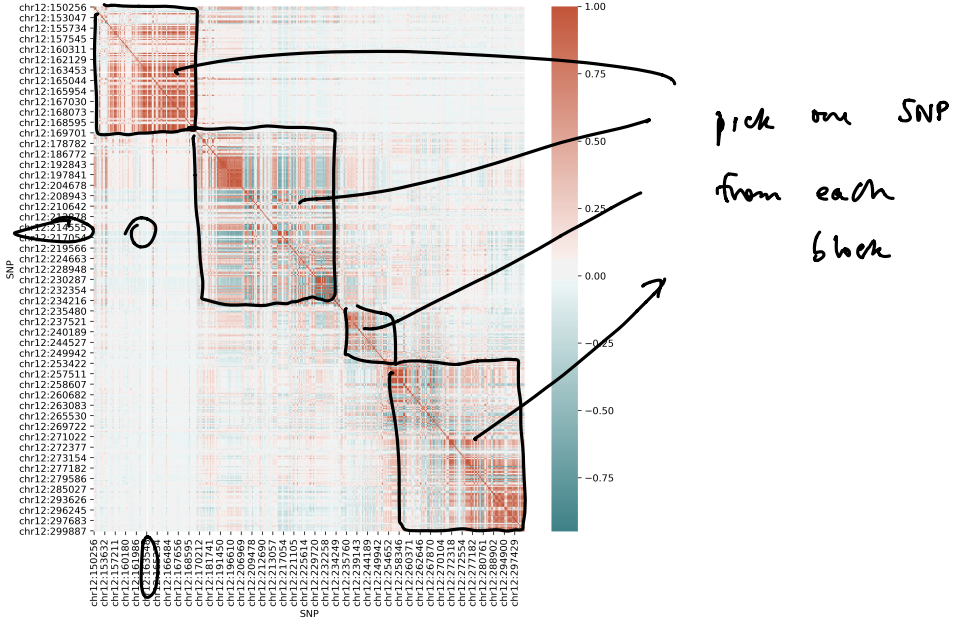
So nearby SNPs are more correlated.

Let's see this in the data!

https://learn.genetics.utah.edu/content/pigeons/geneticlinkage/

Correlation between SNPs

- In practice, SNPs that are "close" to each other are discarded



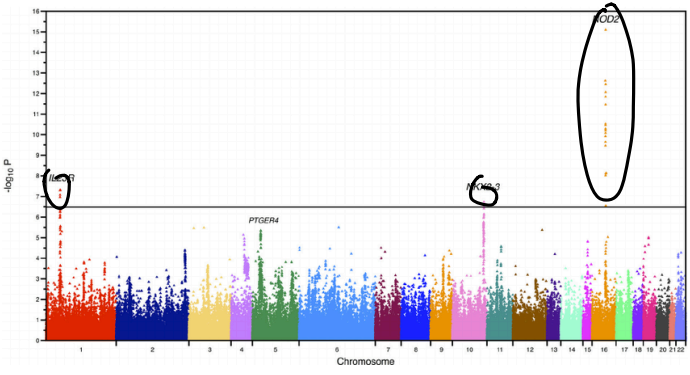
Final predictive model from GWAS

$$\ln\left(\frac{p(1|x)}{1-p(1|x)}\right) = \beta_0 + \beta_5 x_5 + \beta_{981} x_{981} + \cdots + \beta_{150127} x_{150127}$$

Final predictive model from GWAS

$$\ln\left(\frac{p(1|x)}{1-p(1|x)}\right) = \beta_0 + \beta_5 x_5 + \beta_{981} x_{981} + \cdots + \beta_{150127} x_{150127}$$

Only SNPs with high statistical significance

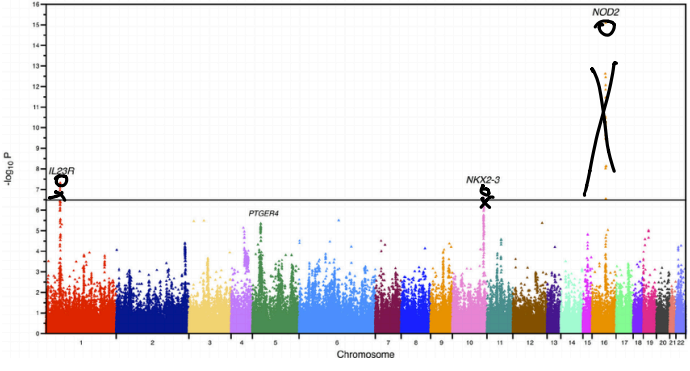


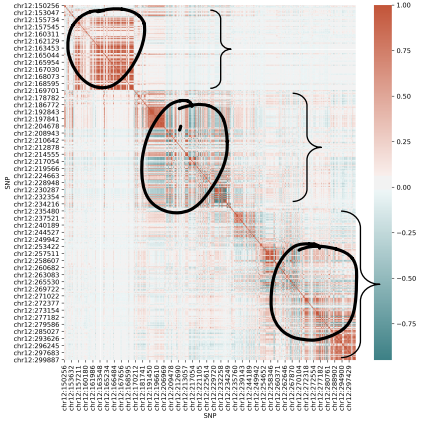
Final predictive model from GWAS

$$\ln\left(\frac{p(1|x)}{1-p(1|x)}\right) = \beta_0 + \beta_5 x_5 + \beta_{981} x_{981} + \cdots + \beta_{150127} x_{150127}$$

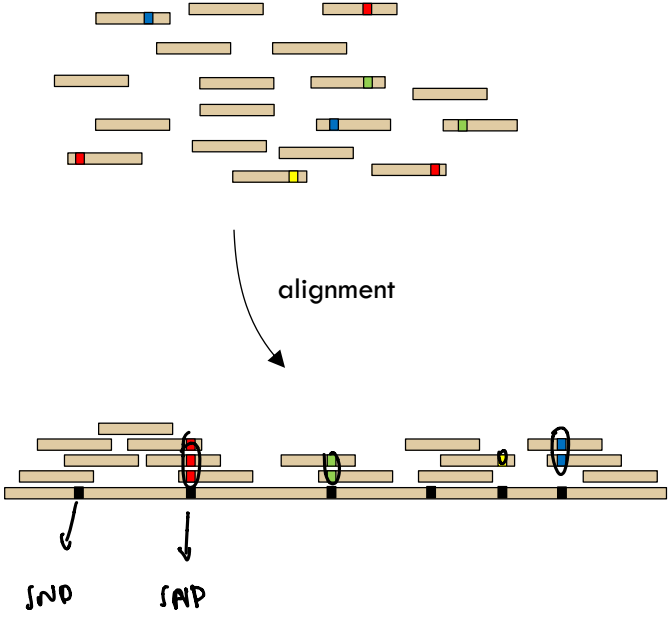
Only SNPs with high statistical significance

SNPs that are distant from each other (small correlation)

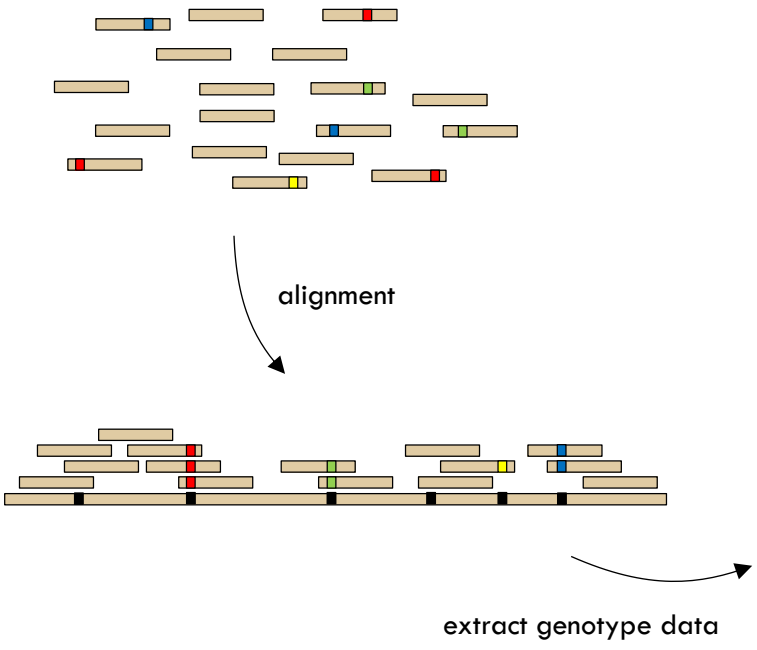




Big picture:



Big picture:



person 1

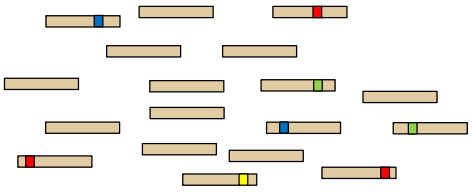
person n

$$\begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$

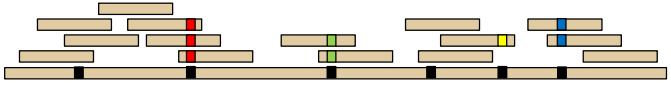
genotype

phenotype

Big picture:



alignment



extract genotype data

$$\ln\left(\frac{p(1|x)}{1-p(1|x)}\right) = \beta_0 + \beta_5 x_5 + \cdots + \beta_{981} x_{981}$$

GWAS

person 1

$$\begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$

person n

genotype

phenotype

Part II: RNA sequencing data analysis

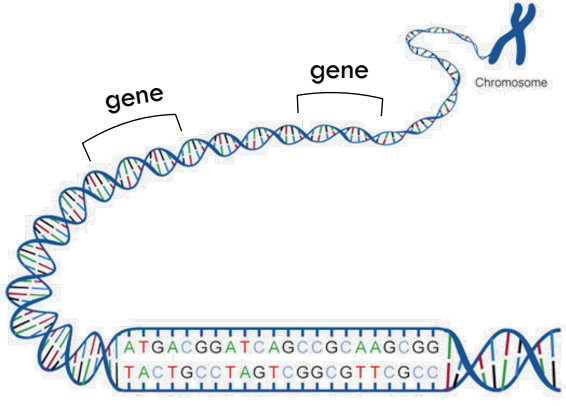
ECE 365

The central dogma of molecular biology

- DNA → RNA → protein

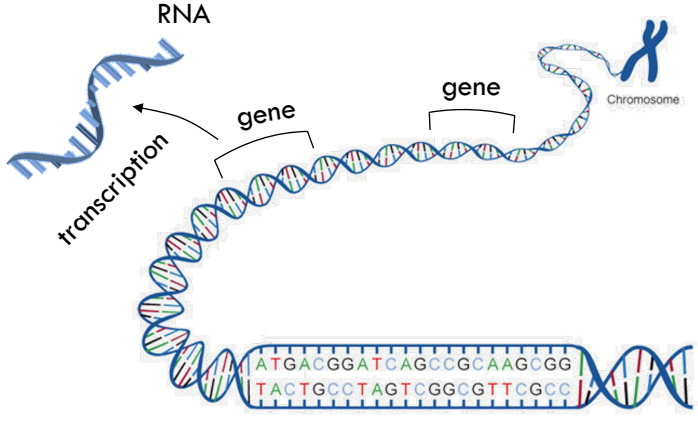
The central dogma of molecular biology

- DNA → RNA → protein



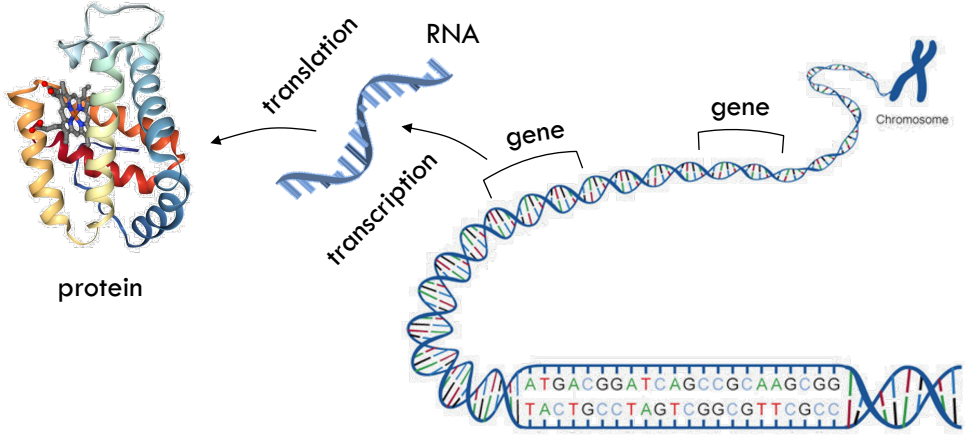
The central dogma of molecular biology

- DNA → RNA → protein

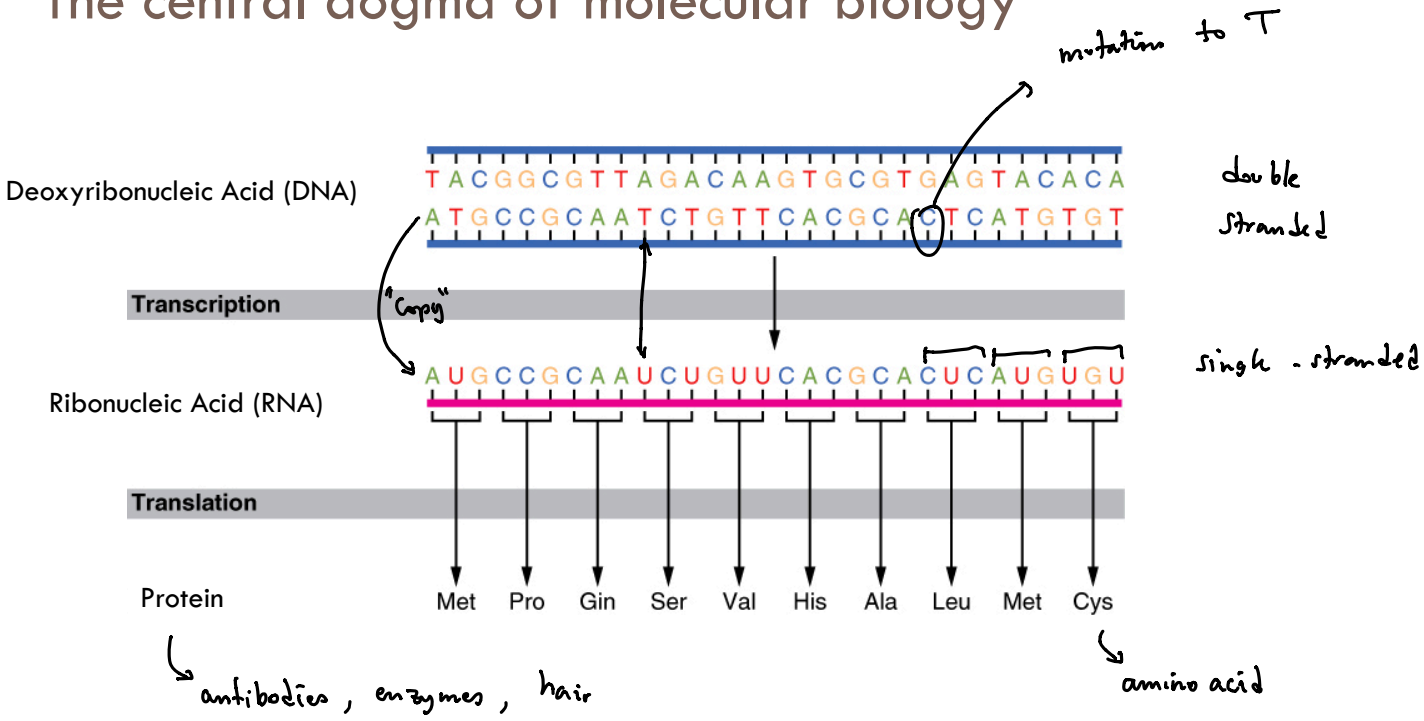


The central dogma of molecular biology

- DNA → RNA → protein



The central dogma of molecular biology

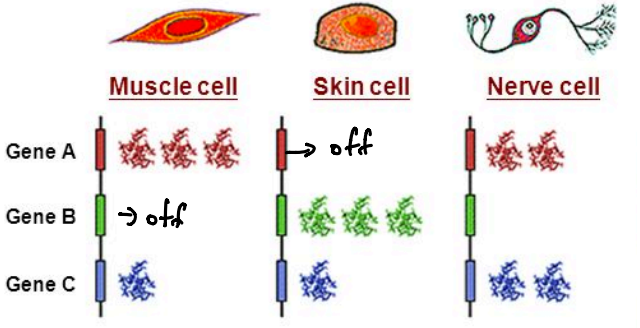


Why do we care about RNA?

- Different genes are *expressed* in different cell types

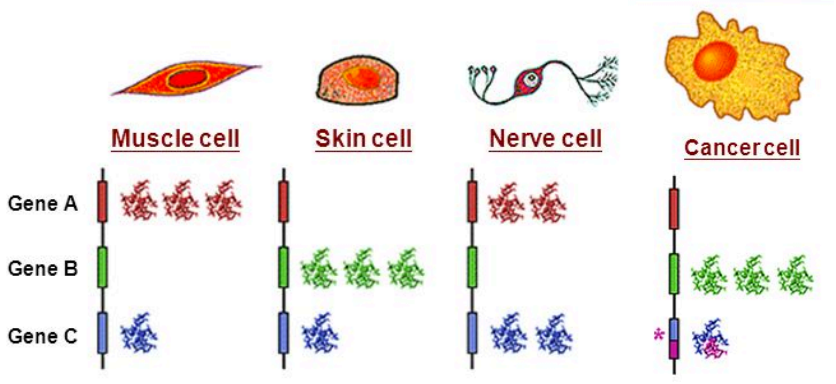
Why do we care about RNA?

- Different genes are *expressed* in different cell types



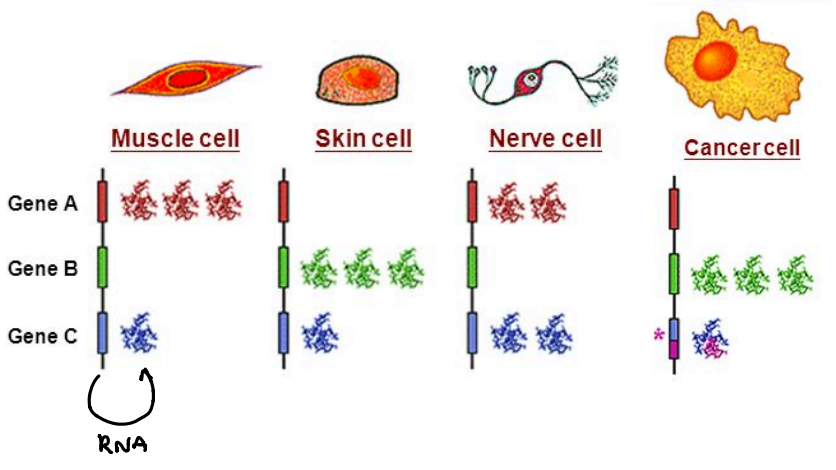
Why do we care about RNA?

- Different genes are *expressed* in different cell types



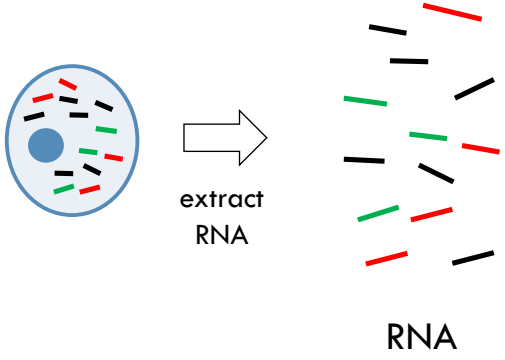
Why do we care about RNA?

- Different genes are *expressed* in different cell types

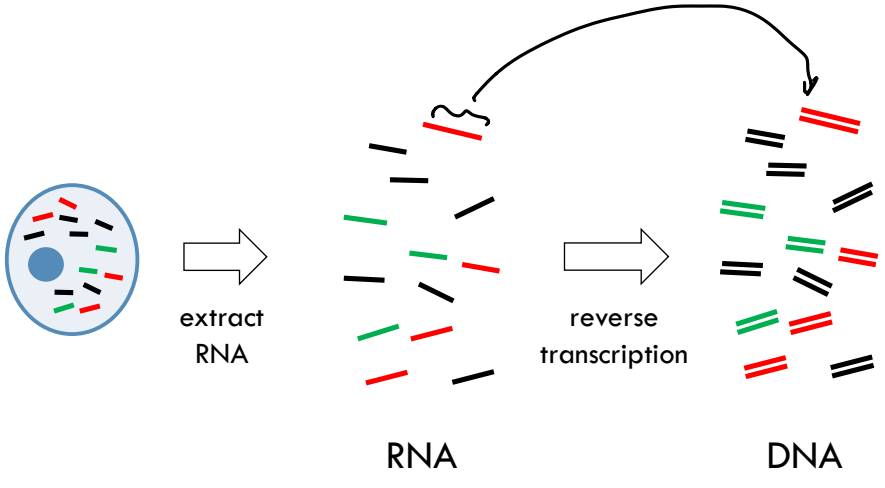


- RNA is an "intermediate step" between genes and proteins
- RNA levels in a cell can tell us which genes are "on/off"

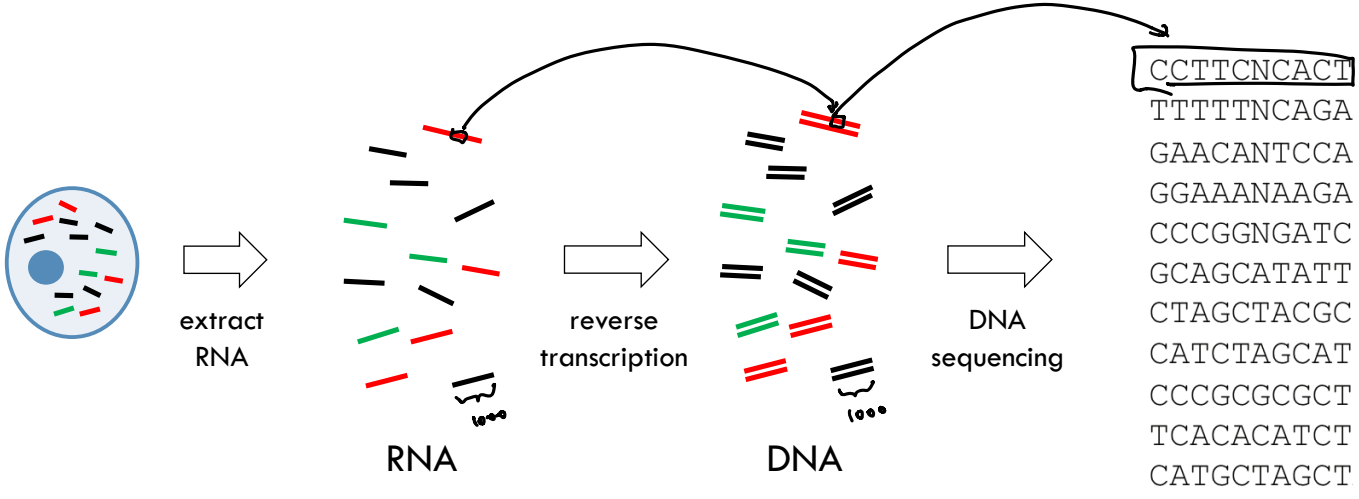
How do we sequence RNA?



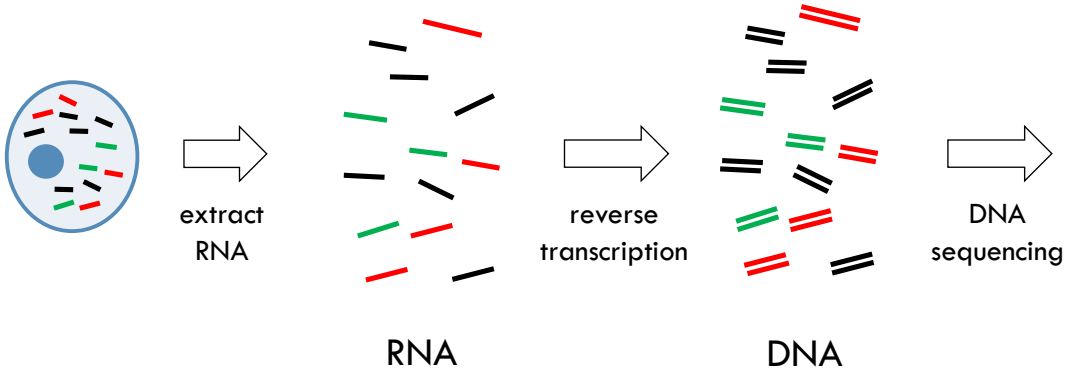
How do we sequence RNA?



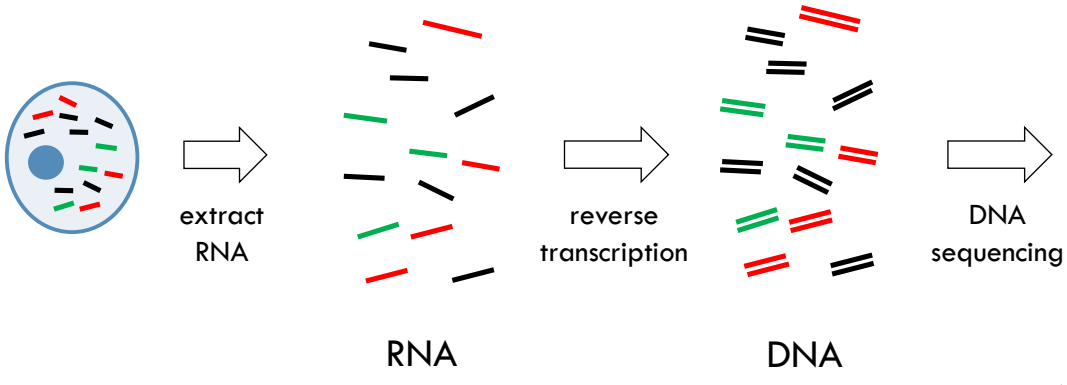
How do we sequence RNA?



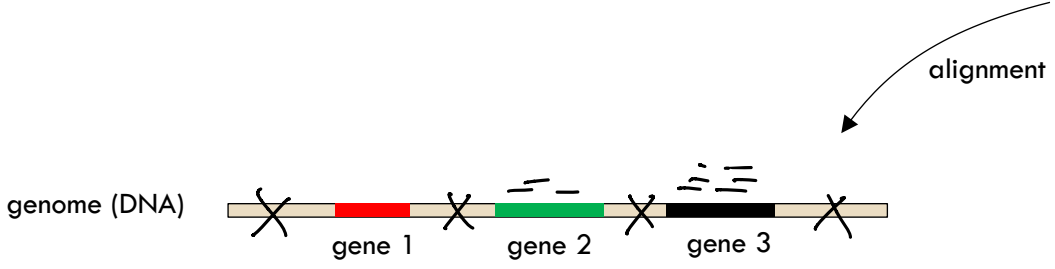
How do we sequence RNA?



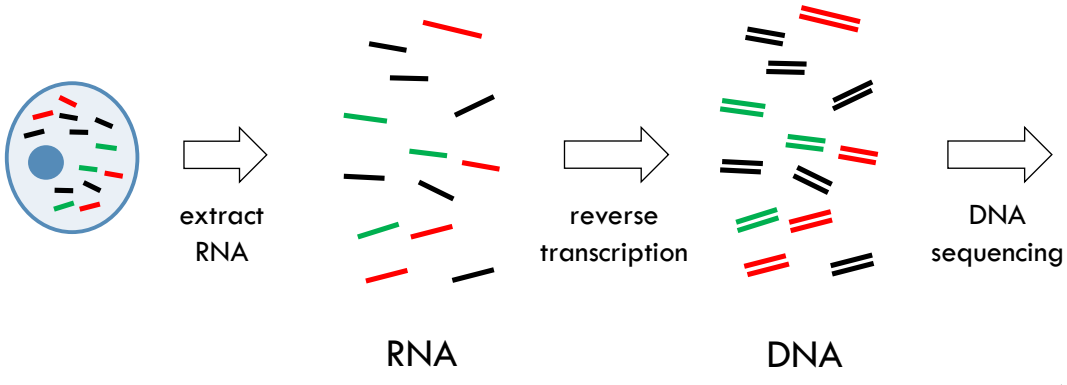
How do we sequence RNA?



GAACANTCCA
GGAAANAAGA
CCCGGNGATC
GCAGCATATT
CTAGCTACGC
CATCTAGCAT
CCCGCGCGCT
TCACACATCT
CATGCTAGCT



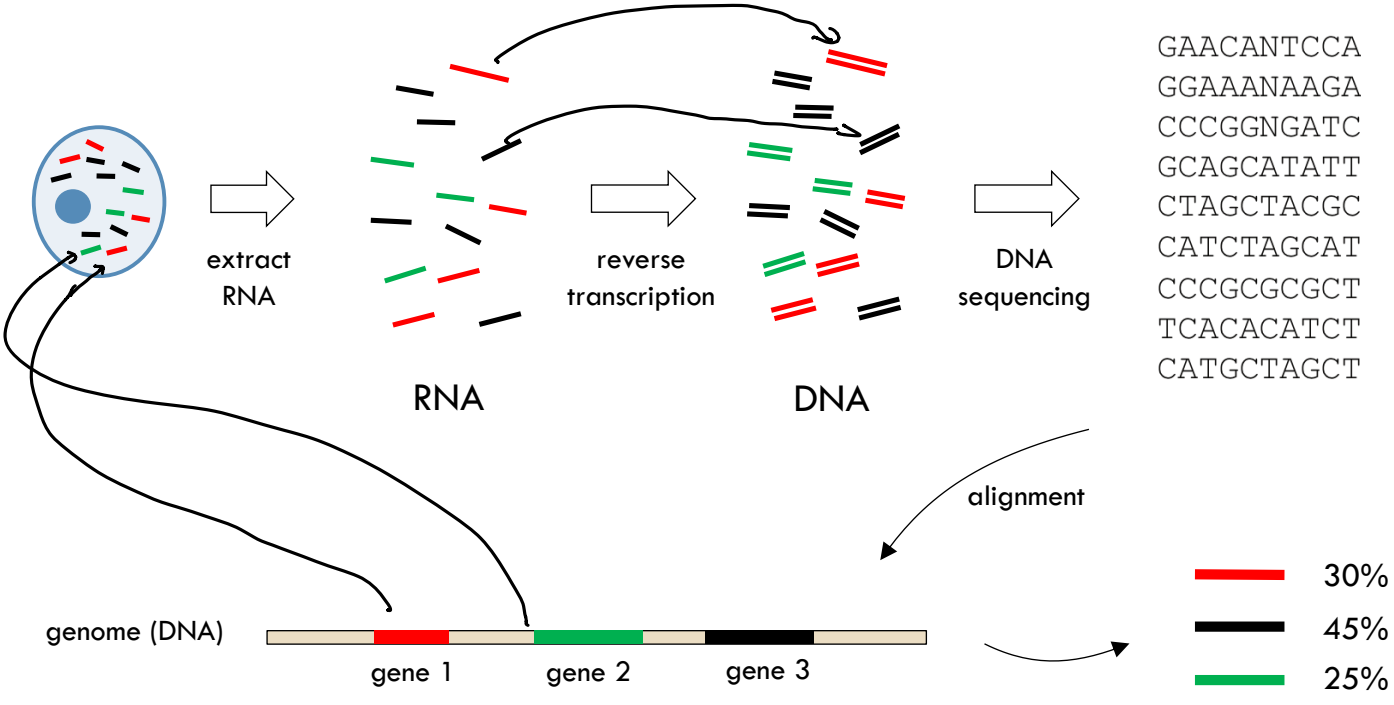
Our first goal: RNA quantification



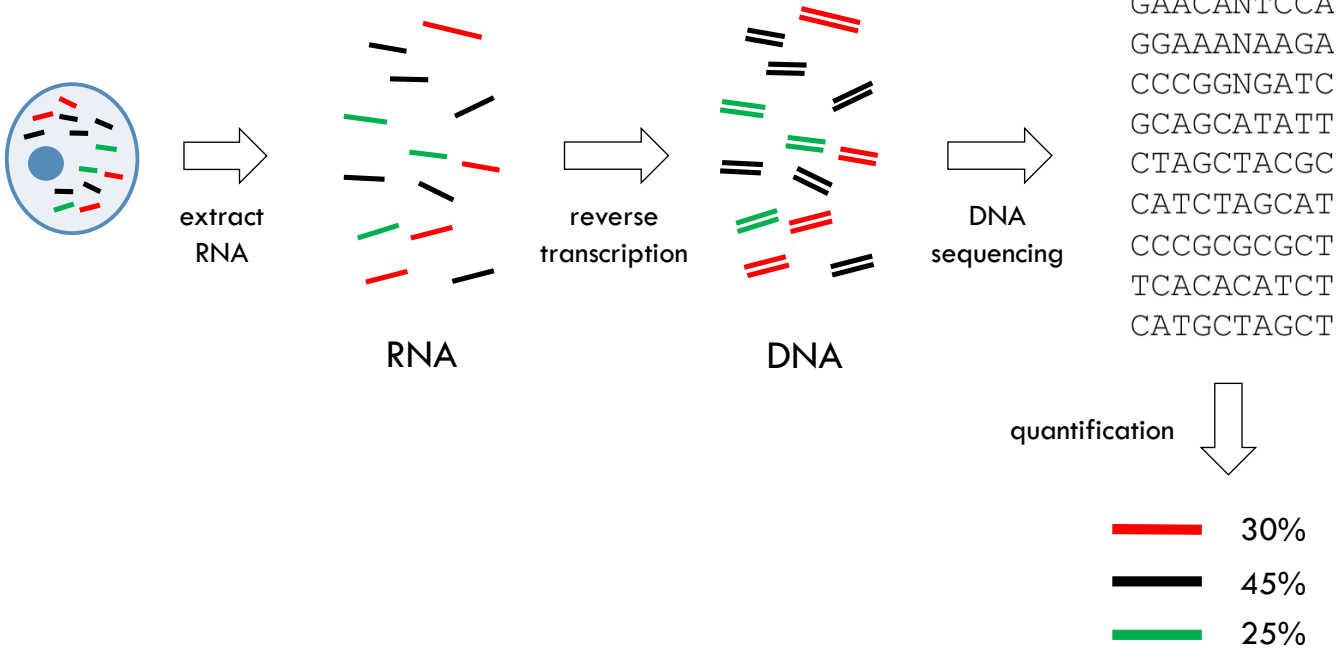
GAACANTCCA
GGAAANAAGA
CCCGGNGATC
GCAGCATATT
CTAGCTACGC
CATCTAGCAT
CCCGCGCGCT
TCACACATCT
CATGCTAGCT

alignment

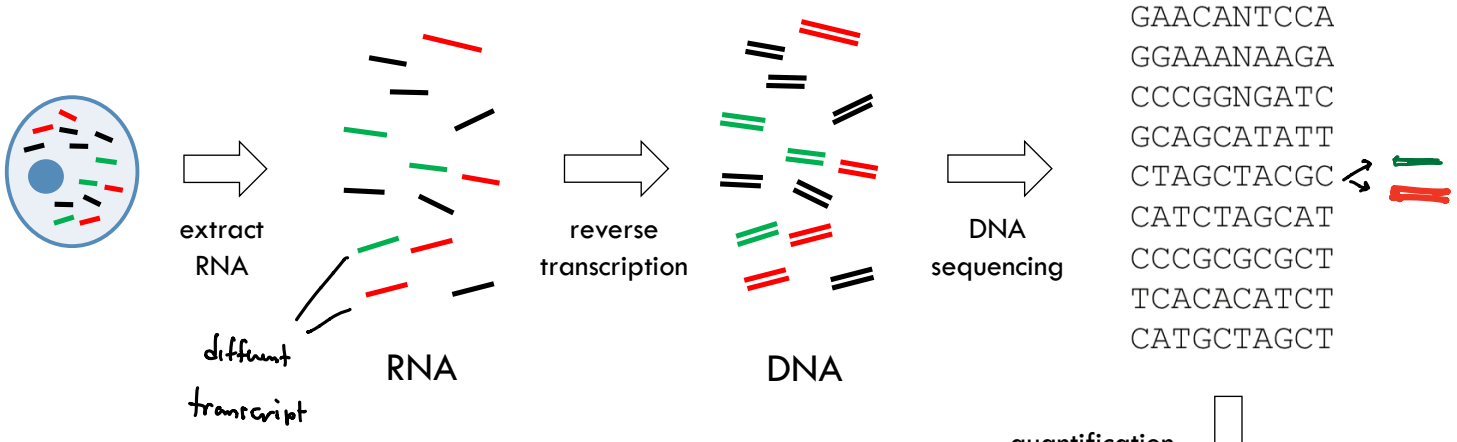
Our first goal: RNA quantification



Our first goal: RNA quantification



Our first goal: RNA quantification



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