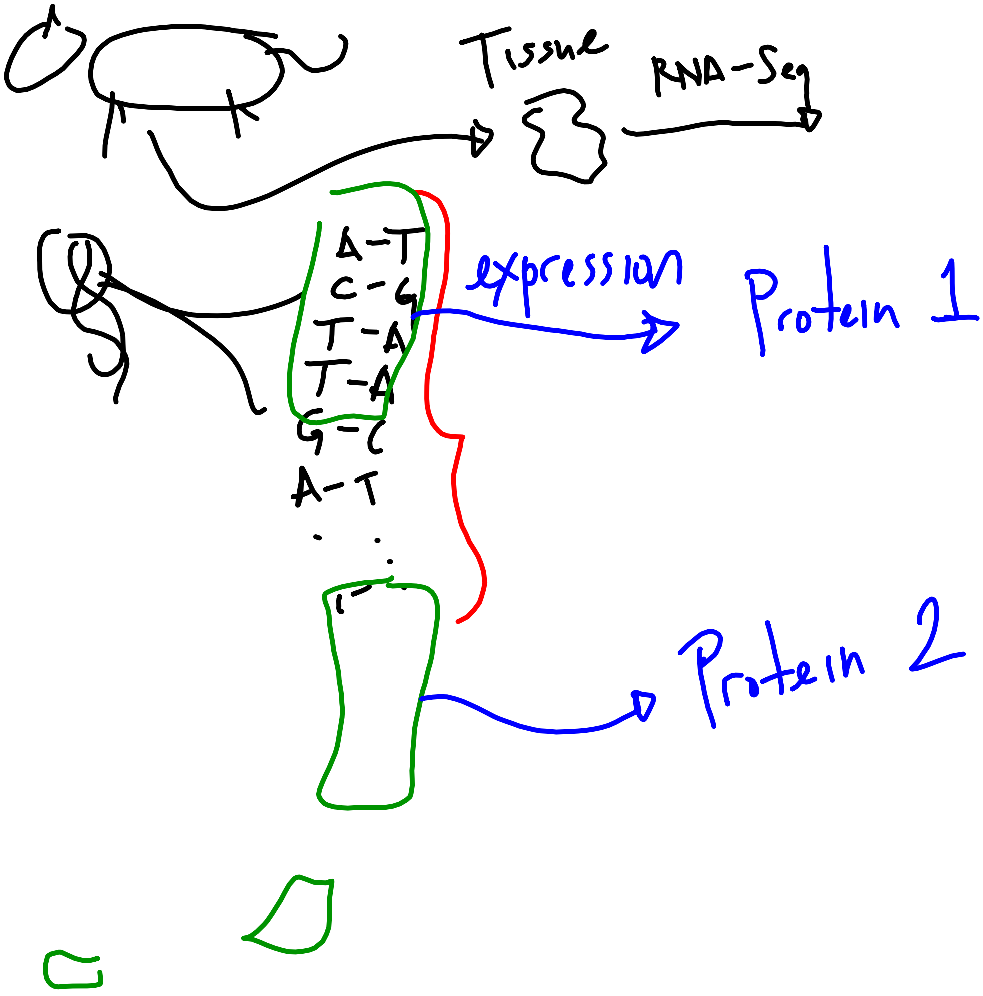
Tissue
RNA-Seq
A-T
C-G
T-A
T-A
G-C
A-T
expression
Protein 1
Protein 2

Data: $X$

$n \times p$

individuals

variables ("genes") "reads"

RNA-Seq:

Quantification done by "read counts"

$X_{ij}$ : # of times read $j$ was observed in sample $i$.

Counts

What we have currently:

- Not $X$
- Raw data files for each individual:
  - FASTQ

Step 1:

- Process the FASTQ files into $X$

Reference R docs | sent.

Step 1b: Create $Z_{n\times 4}$ matrix of "meta data".

## Step 2:

$\underset{n\times p}{X}$ : Which of the "genes" differ between diseased & healthy, adjusted for the meta data variables.

$$\left.\begin{array}{l}\text{Gene 1: p-value}\\ \vdots\\ \\ \end{array}\right\}\longrightarrow \text{List of "differentially expressed" genes}$$

Statistical analysis:

Gene-by-gene:

$$y_{ij} = \beta_0 + \beta_1 TX_j + \beta_2 Act_j + \cdots$$

$$H_0: \beta_1 = 0 \quad \text{vs.} \quad H_a: \beta_1 \neq 0$$

In R: edgeR, DESeq2, limma

---

Instead of p-values, we'll actually use alternative significance measure called False Discovery Rate (FDR)

## Next Steps:

- Create Github user account.
  - I'll create repository
- Do some background reading
  - R/Bioconductor tutorial
  - Khan Academy for basic (2 hrs) genetics & cell/molecular biology
  - You guys web conference to brainstorm, assign tasks & due dates.

R Resources:

- R Cookbook
- Codeschool.com:
  - "Try R" tutorial