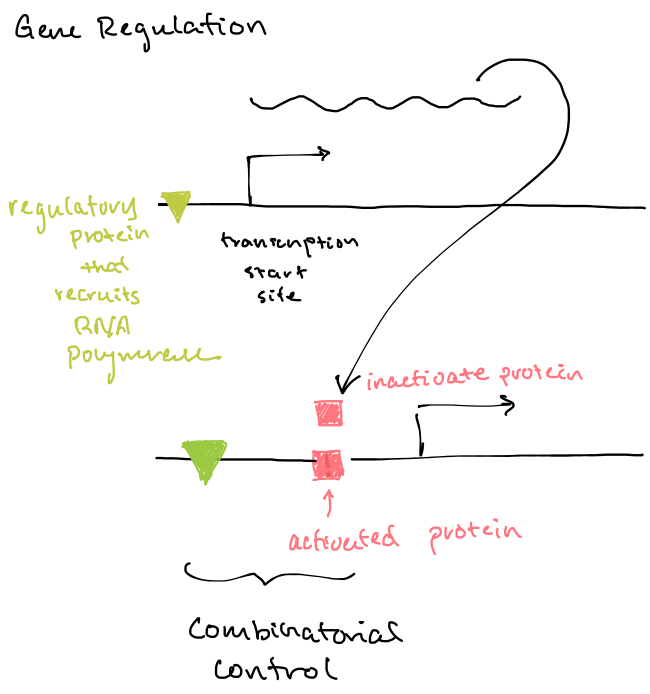
7 ChIP-seq Analysis; DNA-protein interactions

1. Overview

- How do we study gene regulation?
- Transcription factors regulate gene expression.

2. Gene Regulation



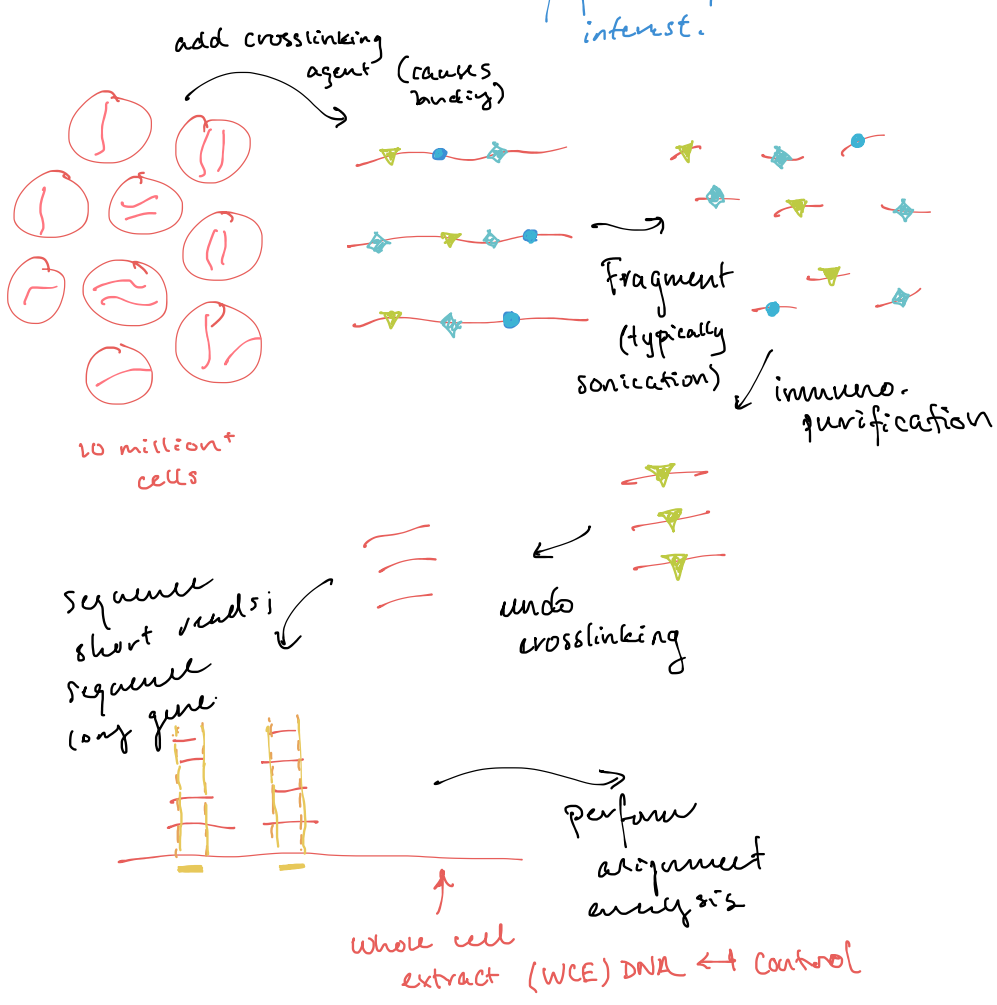
REGULATORY CODE

- currently, we can only consider 1 protein at a time.
- we can also look at more general epigenetic marks.

3. ChIP-seq [Assumes we know what the regulator is]

- chromatin immunoprecipitin, sequencing.
- we can see where proteins bind within 10 bp.
- needs
  - good antibody or,
  - epitopic tag

} for immunopurification of proteins of interest.



- We need to get to the point where we can discover regulatory proteins de novo.

- construct mixture model, map of spatial distribution of reads.

- GPS [genome positioning system]

$$P(r_i \mid b_j) = P(r_i \mid z_{ij} = 1) = \text{emp}\left(-1^{s_i}(r_i - b_j)\right)$$

0 for forward, 1 for reverse.

b_j: binding event at position b_j

- $\mathbb{P}(R \mid \pi) = \prod_{i=1}^{N} \sum_{m=1}^{M} P(r_i \mid m)\pi_m$ subject to $\sum_{m=1}^{M} \pi_m = 1$

N: # of reads; M: # bases in genome.

R: read set (all reads)

- Goal: we want to find

$$\pi^* = \operatorname{argmax}_{\pi} \mathbb{P}(R \mid \pi).$$

- insight:

$$g(z_n = m) = \begin{cases} 1 & \text{if read } n \text{ came from } m \\ 0 & \text{otherwise} \end{cases}$$

$$N_m = \sum_{i=1}^{N} g(z_i = m) \Rightarrow \pi_m = \frac{N_m}{\sum_{m=1}^{M} N_m}$$

N_m: number of reads caused by binding event m.

(g is a latent variable)

- estimate g → estimate π_m → update estimate for g

E step

$$\gamma(z_n = m) = \frac{\pi_m P(r_n \mid m)}{\sum_{m'=1}^{M} \pi_{m'} P(r_n \mid m')}$$

fraction of read n assigned to event m.

$$\hat{\pi}_m^{(t)} = \frac{N_m}{\sum_{m'=1}^{M} N_{m'}}$$

$$N_m = \sum_{n=1}^{N} \gamma(z_n = m).$$

N_m: effective # reads assigned to event m

Assuming punctate (single point) binding proteins.

finds MLE for assumed $\gamma \approx g$

Add a sparse prior on π, the binding events.

$$p(\pi) \propto \prod_{m=1}^{M} \frac{1}{(\pi_m)^{\alpha}}, \quad \alpha > 0.$$ ⇝ now use MAP, & now need α reads

negative Dirichlet prior

$$\hat{\pi}_m^{(t)} = \frac{\max(0, N_m - \alpha)}{\sum_{m'=1}^{M} \max(0, N_{m'} - \alpha)}$$

α or eliminated (component elimination)

- Benjamin-Hotchberg Correction

$P_1, P_2, P_3, P_4, P_5, \ldots$ ~ p-values. (smallest → largest)

$$P_i \leq \frac{i}{N} \cdot \alpha$$

α: desired FALSE DISCOVERY RATE

(replicates)

- Run experiment twice (at least) — are the results concordant?

- $\psi_n(t)$ ≐ fraction of events paired in the top $n \times t$ events

n: total # of events.

t: fraction of events we are considering

↑
- rank in order of significance
- use rank correlation (gets rid of numerical dependencies, like # of reads).

used to find where results are no longer concordant

IDR (irreproducibility discovery rate)