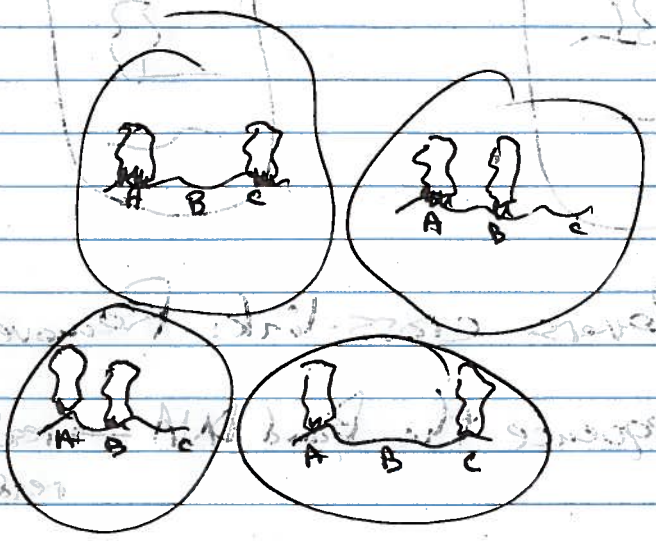
ChIP-Seq + Motif discovery

Goal of ChIP-Seq: Identify regions of genome bound by a given TF in a given type of cells.

ChIP-Seq: Chromatin Immunoprecipitation followed by Seq.



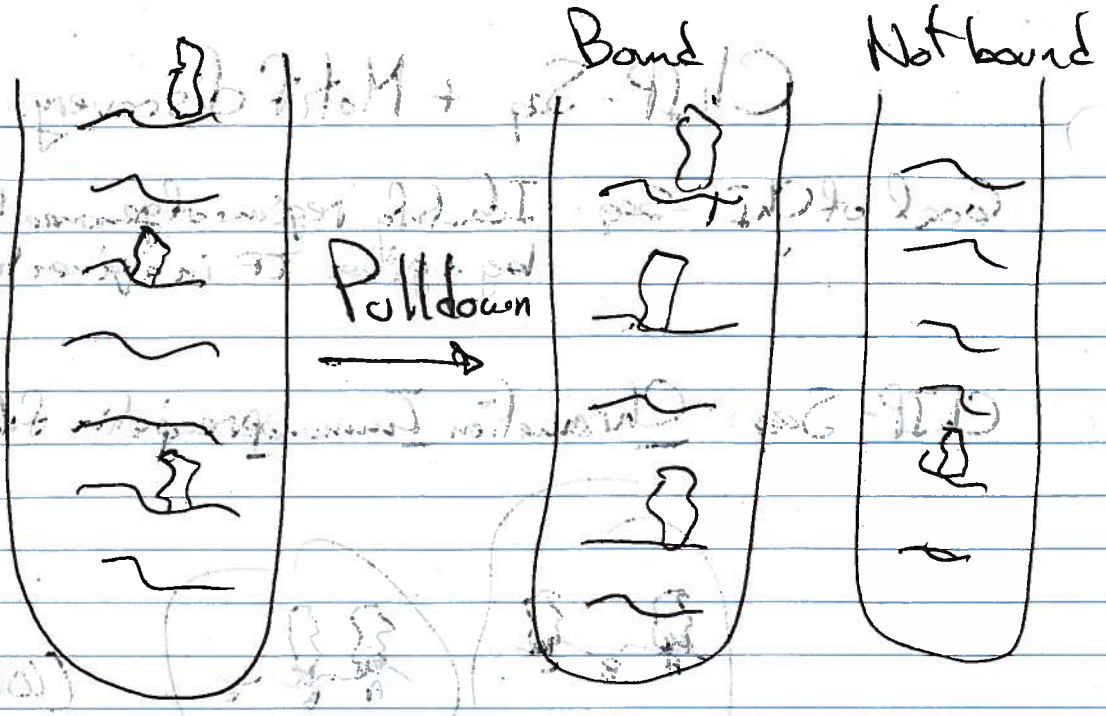
(10 Million cells)

(Typically: 100 - 100,000 sites occupied by TF)

① Cross-linking proteins to DNA: Strengthen bonds btw prot. and DNA

② Extract DNA with proteins attached to it

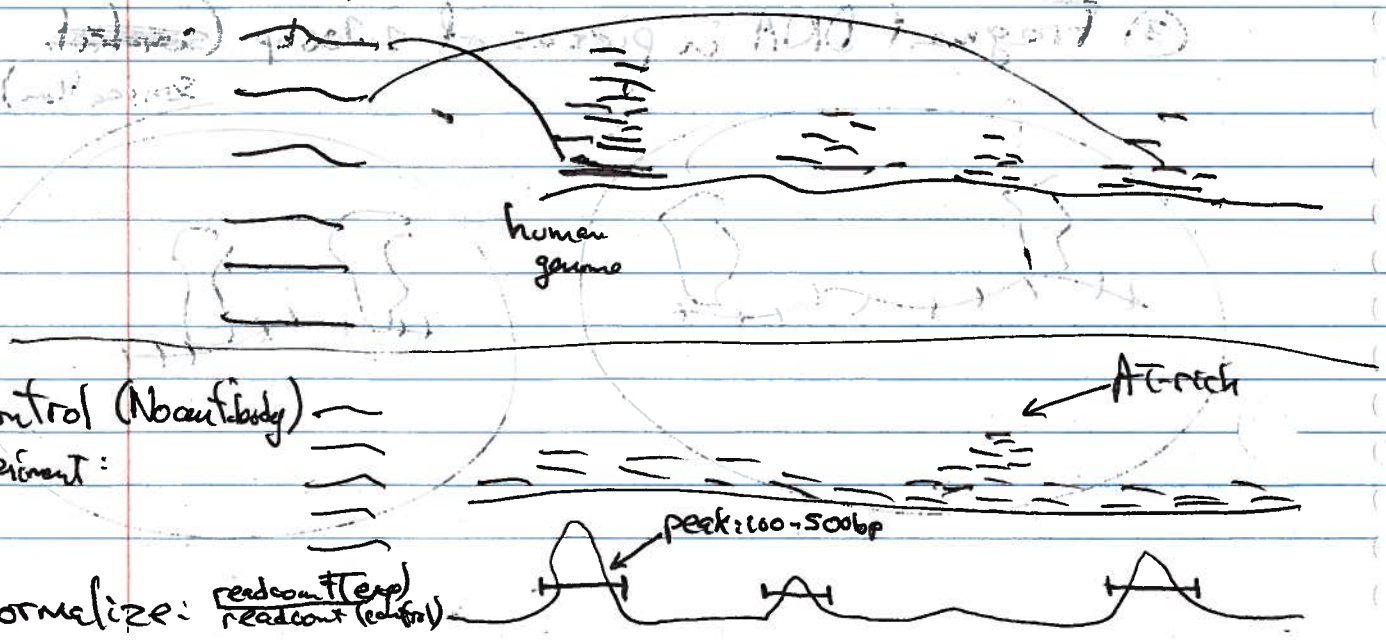
③ Fragment DNA in pieces of ~200bp (~~sonicate~~ sonication)



⑤: Reverse cross-link: Remove the TF from DNA

⑥: Sequence the Bound DNA → read1: ATTAGGTC (200bp)
read2: TGTGCTA...
⋮
read 10^7: TAATGAT...

⑦ ~~Map~~ Read Mapping
Local alignment btw reads and the human genome



100 - 100,000

ChIP-seq outcome: List of genomic regions (100-500bp that are believed to be bound by TF

Problem: Doesn't tell us which 6-15bp region is bound

⇒ Can't build consensus or PWM for TF

Motif Discovery Problem

Given: Set of sequences $S_1, S_2, \ldots S_n$ (of length 100-500) each believed to contain binding site for TF

Find: Consensus sequence for TF

OR

PWM for TF

Consensus sequence: Regular expression

$w = \{A\}\{C\}\{A\}\{T\}\{A\}$
$\quad\quad\{T\}\quad\{C\}$
$\quad\quad\{G\}$

How should we score a candidate cons. seq $w = w_1 \ldots w_k$

Criteria:

- w should occur in each of $S_1 \ldots S_n$
  - → too strict: Some S_i might be false-positives
  - → $\{A\}\{A\}\ldots\{A\}$ matches everywhere

Motif enrichment approach → consensus sequence

Let N_w = # of matches for w in $S_1, S_2, \ldots S_n$

E_w = Expected # of matches of w in a set of random sequences $R_1, R_2, \ldots R_n$, where R_i has same length S_i

→ each nucleotide is chosen independently with $P_A: 0.3$, $P_C: 0.2$, $P_G: 0.2$, $P_T: 0.3$

How to compute E_w? $w = w_1 w_2 \ldots w_k$ where $w_i \in \{A, C, G, T\}$

Ex: $w = \{A\}\{C\}\{A, T\}\{G\}\{T\}$

$\Pr[w$ has a match starting at position p in random seq$] = ?$

R: . . . . . p ... p+k-1

$$= \prod_{i=1}^{k} \Pr[w_i \text{ has match at position } p+i-1]$$

$$P_{match}(w) = \prod_{i=1}^{k} \left(\sum_{a \in w_i} P_a \right)$$

$$E_w = (\#\text{positions eligible for match}) \cdot P_{match}(w)$$

$$= \sum_{i=1}^{n} (\text{length of } S_i - k + 1) \cdot P_{match}(w)$$

Finally, from N_w, E_w

We want to find w where N_w and E_w are the most different, i.e. $N_w >> E_w$

Z-score approach

$$Z_w = \frac{N_w - E_w}{\sqrt{E_w}}$$

Complete algo.

For each possible consensus seq. w

Calculate N_w, E_w, Z_w

Report word w with highest z_w