7.36/7.91 Recitation

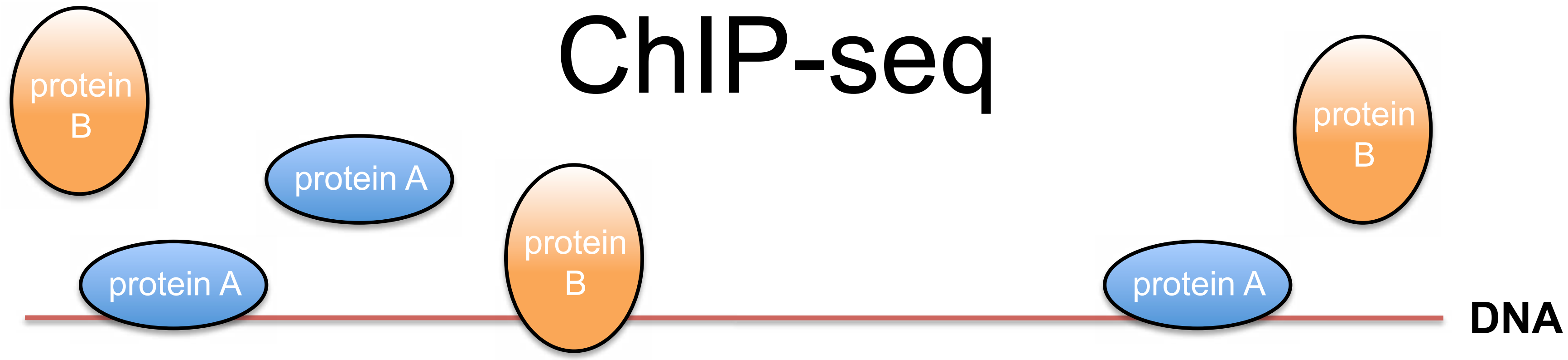
3-5-13

DG Lectures 7 & 8

Announcements

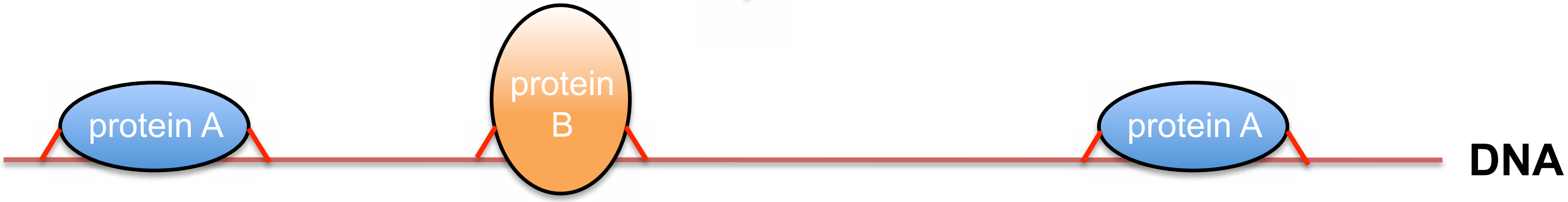
- Project specific aims due this ~~Friday (2/07)~~
  - examples will be posted this evening
- Pset #2 due in 1 week (02/13)
  - For problem 2B, Matlab and Mathematica use a (1-p) parameterization in contrast to lecture slides (p):
    - R or N = 1/k (same as in lecture slides)
    - P = $\frac{\frac{1}{k}}{\lambda + \frac{1}{k}}$ for Matlab/Mathematica vs. $\frac{\lambda}{\lambda + \frac{1}{k}}$ in lecture slides

ChIP-seq

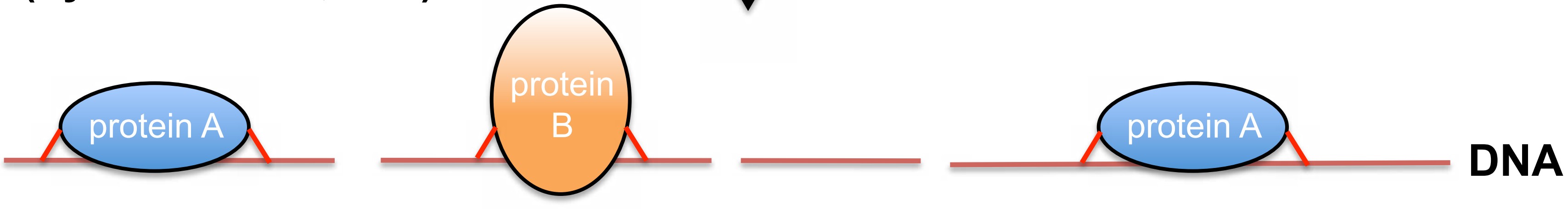


(1) crosslink proteins and DNA ↓

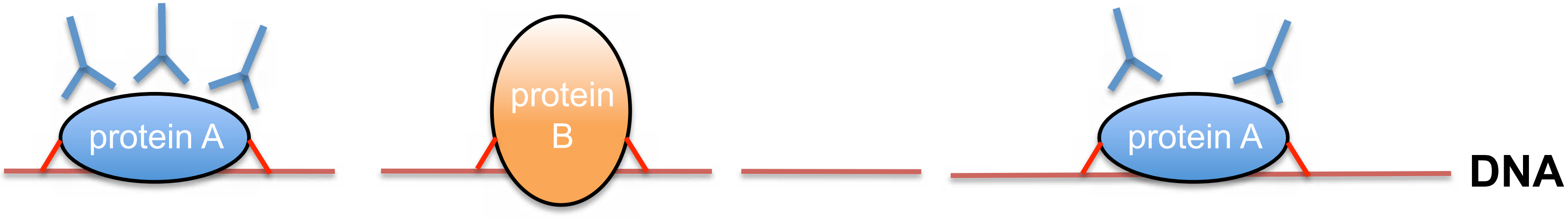
tightly but reversibly binds proteins to nearby DNA



(2) fragment DNA
(by sonication, etc.) ↓

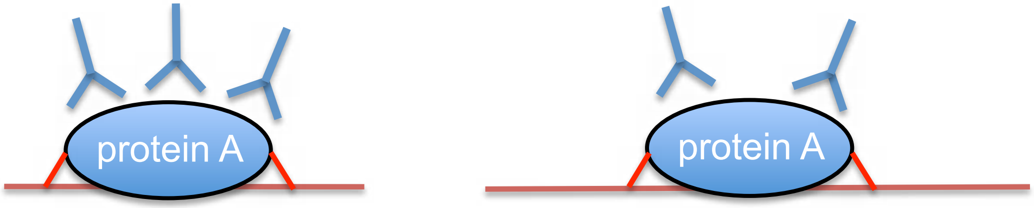


ChIP-seq



(3) immunoprecipitate

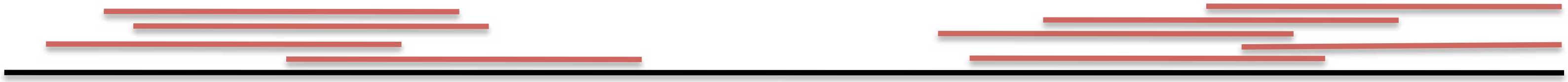
use antibodies against protein of interest (here protein A)



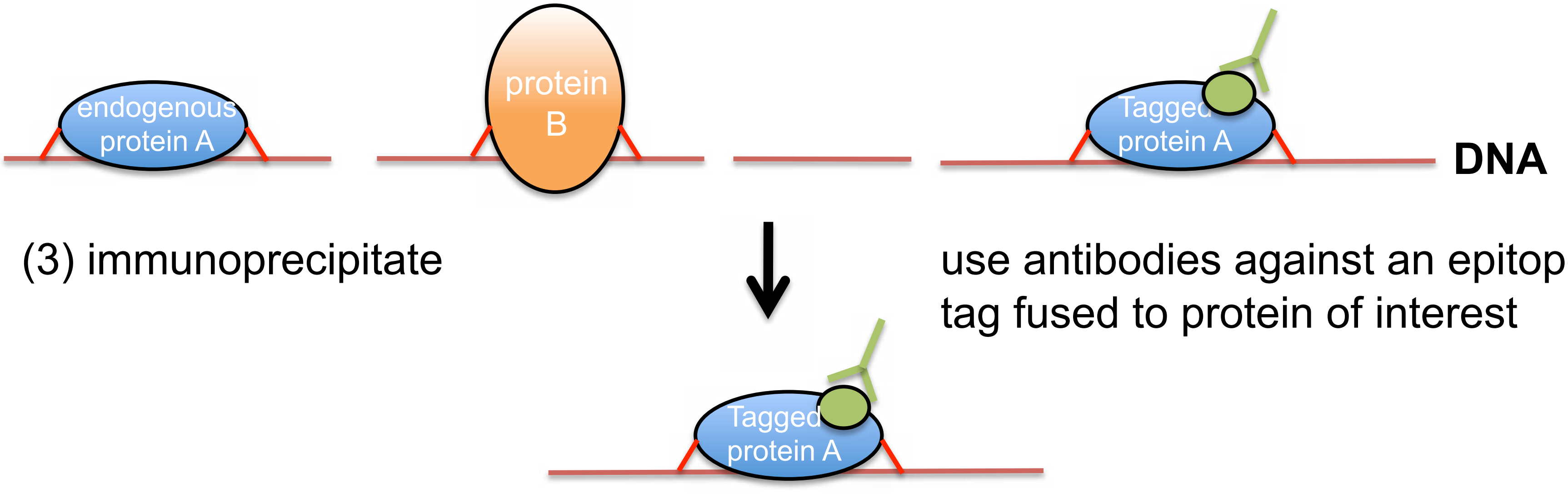
(4) reverse crosslinks, purify DNA

(5) sequence purified DNA, align reads to genome

reference genome in black
reads in maroon



ChIP-seq



(3) immunoprecipitate

use antibodies against an epitope tag fused to protein of interest

-If a high-quality antibody is not available for your protein of interest, an alternative is to introduce a construct of the protein with an epitope tag into cells and use antibody to epitope

- Epitope is a short (5-10) amino acid sequence with antibody available (e.g. HA, His)
- Some caveats to keep in mind:
    - Expression of your construct may not be at WT levels
    - Epitope tag could alter function and/or localization of the protein, which could lead to non-native binding locations

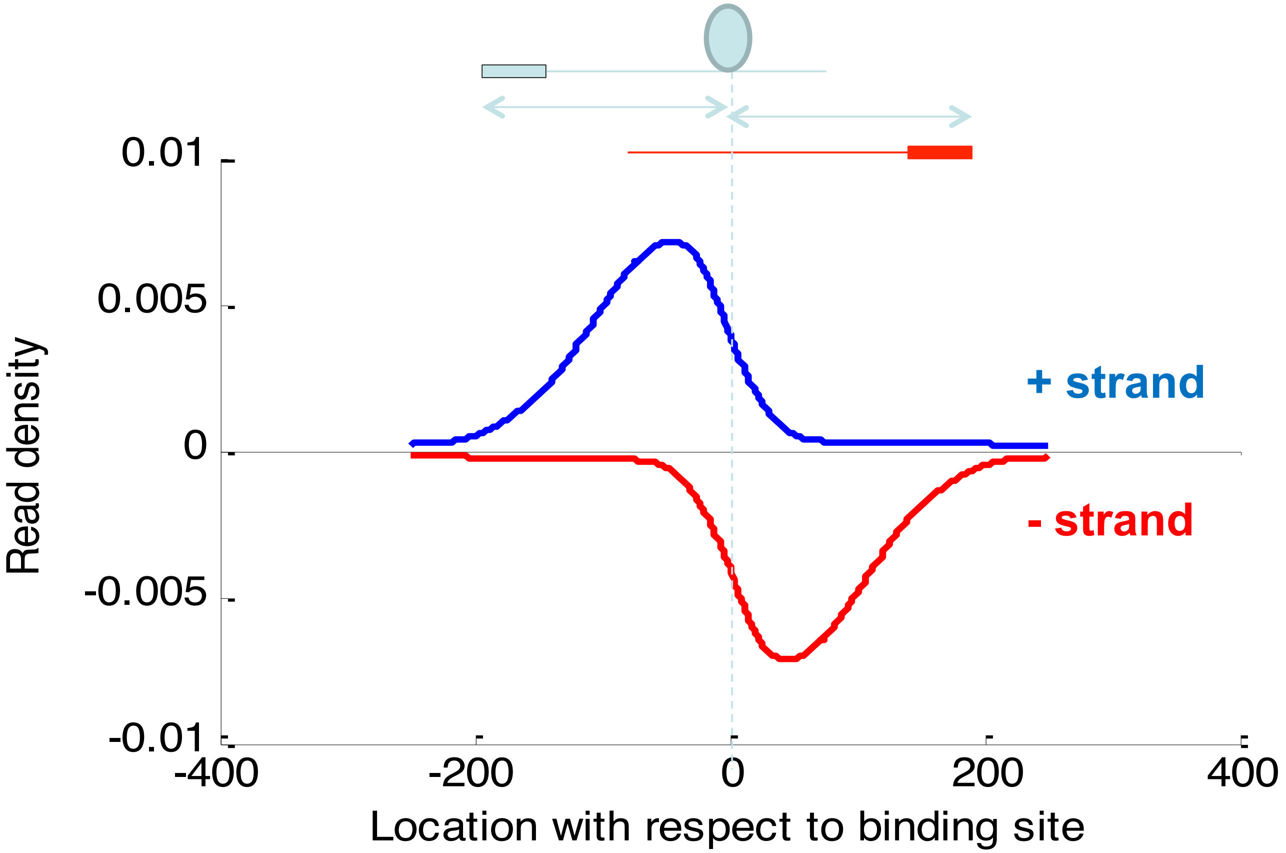
Peak-calling applications

- ChIP-Seq
- MeDIP-Seq (Methylated DNA IP)
  - Finding methylated regions of DNA
- CLIP-Seq (CrossLinking and ImmunoPrecipitation)
  - Conceptually similar to ChIP-Seq, but for RNA instead of DNA
- MeRIP-Seq (aka M6A-Seq)
  - Post-transcriptional methylation of N6 position of adenosine in RNA

GPS – genome positioning system

- Find binding sites of DNA binding proteins (e.g. transcription factors) that have sequence specificity (6-8 bp of preferred motifs)
  - ChIP-seq experiments help us understand how gene expression programs are regulated and how they change in development or in response to a stimulus
- In addition to ChIP with your antibody, perform a parallel set of experiments with everything the same except either with no antibody or a nonspecific (e.g. IgG) antibody
  - Whole cell extract
  - Negative control that represents the experimental background – only binding stronger than background is specifically due to your protein of interest binding

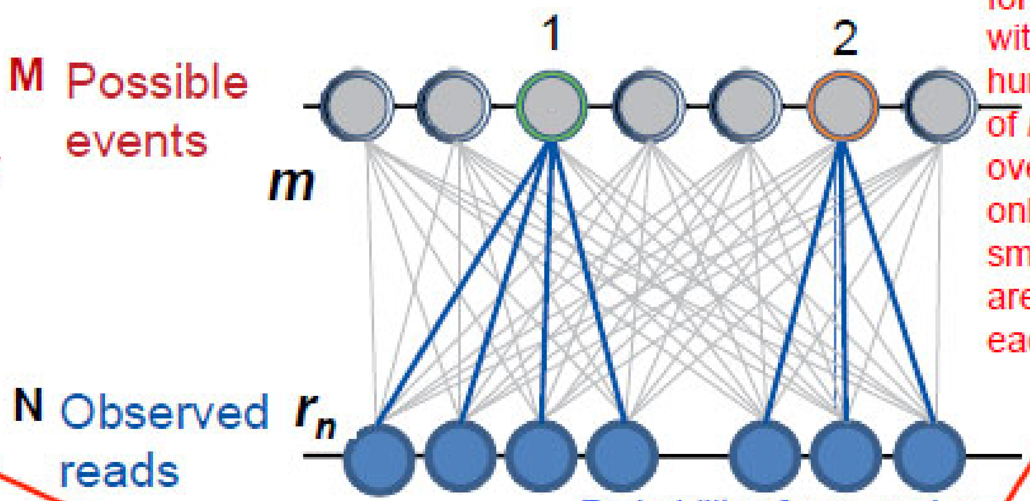
GPS empirically estimates read distribution profiles



-Each protein, read length, sequencing depth, etc. will have a slightly different binding profile, so it's important to estimate this for each experiment

GPS models every base *m* in genome as a possible binding event

We want to maximize this quantity, the probability of observing the set of R reads from our ChIP seq experiment, given that the protein is bound at the nonzero positions in π



This probability is nonnegligible for a read *r* only within a few hundred bases of *m*. We sum over all *m*, but only one or a small few terms are nonzero for each read.

π is a genome-size (e.g.~3 billion for human) dimensional vector – want to get it to have only a few thousand nonzero elements

Likelihood of observed reads

$$p(R\,|\,\pi)=\prod_{n=1}^{N}\overbrace{\sum_{m=1}^{M}\pi_m p(r_n\,|\,m)}^{\text{Probability of one read } n},\qquad \sum_{m=1}^{M}\pi_m=1$$

-On average, there are on the order of ~10,000 binding events. But genome size M is ~3 billion, so we want to force most to zero and only ~10,000 to be nonzero.

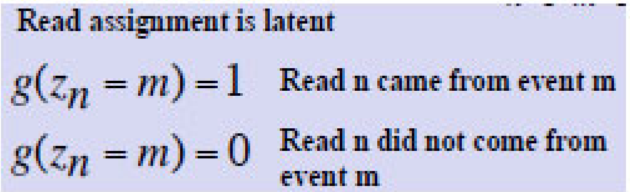
Likelihood of observed reads

$$p(R \mid \pi) = \prod_{n=1}^{N} \sum_{m=1}^{M} \pi_m p(r_n \mid m), \qquad \sum_{m=1}^{M} \pi_m = 1$$

- We want to maximize the probability of observing our reads
  - We can maximize this if there are binding events near the clusters of ChIP-seq reads, since each of these reads will then individually have a relatively large probability of having been observed.
  - So maximizing this likelihood of observed reads is coupled to learning the distribution of binding events, π
  - We start out with no knowledge of π, and want to force most of its terms to 0 (component elimination) and learn the ~10,000 that are nonzero
- The EM (Expectation-Maximization) algorithm is an iterative method for finding the maximum likelihood (ML) estimates of parameters in statistical models, where the model depends on unobserved latent variables
  - What is the parameter to be estimated?
    - Choose π to be that which (arg max) maximizes the likelihood of the ChIP-seq read set R

$$\pi = \arg\max_{\pi} p(R \mid \pi)$$

  - What are the unobserved latent variables?
    - We don't know which binding event each read came from



Read assignment is latent

$g(z_n = m) = 1$ Read n came from event m

$g(z_n = m) = 0$ Read n did not come from event m

EM algorithm

Expectation-Maximization (EM) algorithm with component elimination

E step

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

$\gamma(z_n=m)$: the fraction of read *n* assigned to event *m*

M step

$$\hat{\pi}_m^{(i)} = \frac{N_m}{\sum_{m'=1}^{M} N_{m'}}$$

← Reinforcing constraint that π must sum to 1

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

N_m : the effective number of reads assigned to event *m*

- Although z_n truly have *hard* (i.e. 0/1) values – either the read did or didn't come from a particular binding event – throughout the iterative algorithm we will allow them to take on *soft* values - $\gamma(z_n=m)$ - anywhere between 0 and 1, reflecting our uncertainty as to exactly which π element the read came from
- For all binding events *m'*, calculate the probability that a read came from a binding event ($p(r_n|m')$), weighted by the strength of the binding event ($\pi_{m'}$). The probability that a read came from any particular binding event *m* is just that event's fraction of the total sum.

- N_m is the number of reads assigned to binding event m, where the sum is over the soft counts $\gamma(z_n=m)$
- $\pi^{(i)}_m$: The strength of binding event *m* in the *i*th iteration of the EM algorithm.
  - Note that the denominator sum equals *N*, the total number of reads in the ChIP-seq data set

EM algorithm

Expectation-Maximization (EM) algorithm with component elimination

E step

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

γ (z_n=m)* : the fraction of read *n* assigned to event *m

M step

$$\hat{\pi}_m^{(i)} = \frac{N_m}{\sum_{m'=1}^{M} N_{m'}}$$

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

N_m* : the effective number of reads assigned to event *m

E (Expectation) step:
- Calculate the expected value of the likelihood of the observed reads. We must assign reads to binding events in order to do this.

M (Maximization) step:
Choose values of the parameters (π, the locations of the binding events) that maximize the likelihood of the observed reads – these formulas give the values of those parameters.

Initialize the binding events to be equally likely at every nucleotide ($\pi_j = \frac{1}{M}$), then run the EM algorithm (iterate between the E and M steps – assigning reads to events and then updating the location/strength of events) until convergence

EM algorithm

-The previous formulation works pretty well, but distributes the binding events π to *too many* nonzero components

- For example, it might assign equal binding weight of 0.0033 to three adjacent nucleotides when we know there was a binding event starting at only one of them which should have weight 0.01 with the other two being 0.

- Solution: introduce a sparse prior, which penalizes the a nucleotide for having nonzero π component - the likelihood is multiplied by this prior to get calculate a posterior likelihood. If α is chosen within an appropriate range, this can force nearby components to 0 while allowing one nucleotide in the neighborhood to have nonzero component in the π that maximizes the posterior likelihood.

Likelihood of observed reads

$$p(R \mid \pi) = \prod_{n=1}^{N} \sum_{m=1}^{M} \pi_m p(r_n \mid m), \quad \sum_{m=1}^{M} \pi_m = 1$$

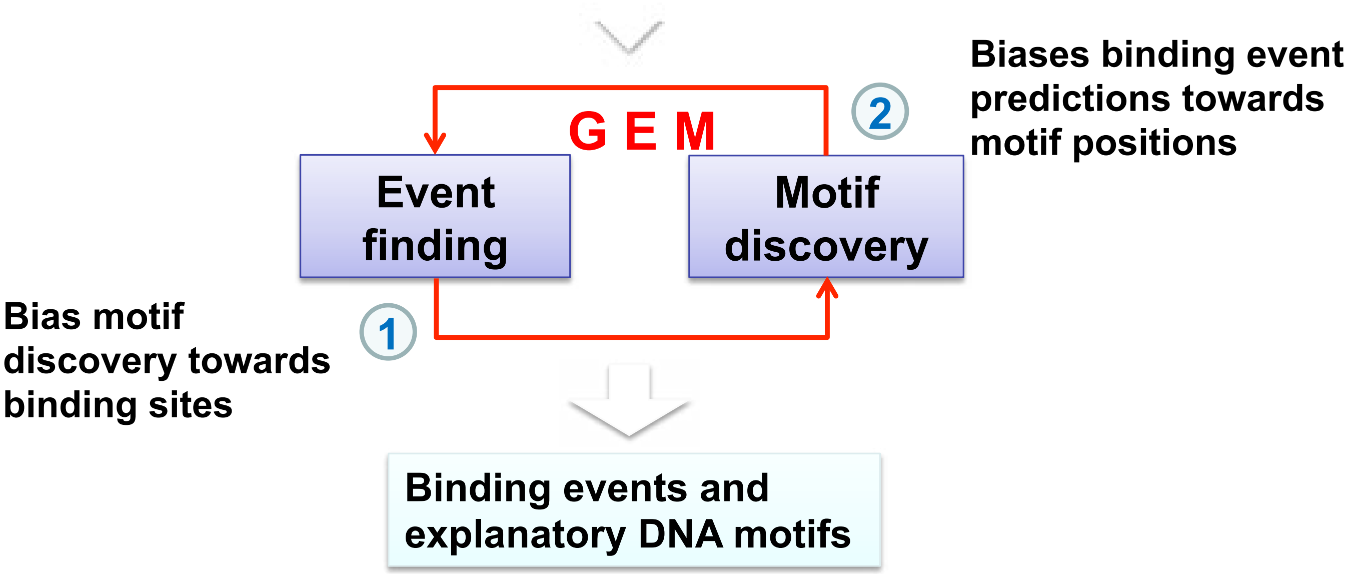
A sparse prior on mixture components (binding events)

$$p(\pi) \propto \prod_{m=1}^{M} \frac{1}{(\pi_m)^{\alpha}}, \alpha > 0$$

(Figueiredo and Jain, 2002)

GEM: Genome-wide Event-finding and Motif discovery

- If we know the motif(s) that the protein binds to, incorporating this information can inform where binding sites are, which can strongly improve spatial resolution
  - Even if unknown ahead of time, we can try to learn the motif(s) from overrepresented *k*mers around the predicted binding sites


GEM
Event finding
Motif discovery
2
Biases binding event predictions towards motif positions
1
Bias motif discovery towards binding sites
Binding events and explanatory DNA motifs

How significant are resulting peaks?

- Compute a scaling factor between control and IP reads using the read counts from the non-peak regions
- Then, for each binding event (nonzero π element), calculate the P-value for observing the number of ChIP reads under the null hypothesis that there was no binding event (i.e., reads equally likely in control and IP since both are background):

Under null hypothesis, reads should be equally distributed between control and IP. In reality, we observe more reads in IP and fewer in control.

P-value: if the *n* reads near the binding site were randomly distributed between the control and IP, what's the probability of observing *k* or fewer reads in the control?

$$\longrightarrow F(k,n,P) = \sum_{l=0}^{\lceil k \rceil} \binom{n}{l} P^{l} (1-P)^{(n-l)}$$

k: scaled control read count

n: total count of IP and scaled control reads

P: probability that reads occur from IP data, $P = 0.5$.

Multiple Hypothesis Correction

- When testing many events or hypotheses simultaneously, some will pass significance by chance (i.e., perform an experiment 100 times, and about 5 times you will get P≤0.05 even if the null hypothesis is true)

- To try to mitigate this when testing multiple hypotheses, we make the P-value threshold for significance even lower when testing

Multiple Hypothesis Correction

Common corrections:

- **Bonferroni**: Divide desired P-value cutoff by # of hypotheses
  - Conservative correction, lowers P-value more than necessary and may produce more false negatives than desired
  - example: if for a single test you would reject if p-value ≤ 0.05, and you're running 100 tests, reject those individually if p-value ≤ 0.05/100 = 0.0005
    - the probability that your 100 tests include at least one false positive due to chance is 0.05
- **Benjamini-Hochberg** (less conservative):

$$Q - value = P - value \times \frac{Count}{Rank}$$

Count: total number of binding events tested.

Rank: Rank of event in list list of p-values, from most significant (rank = 1) to least (rank = Count)

  - The Q-value is the False Discovery Rate (FDR) analog of the P-value
  - Choose a FDR that you're willing to tolerate (e.g. 0.05 – an expected 5% of times you reject the null will be false positives)
    - Then accept events as significant (i.e., reject the null hypothesis) for events of rank 1 through all those for which the Q-value is less than the desired FDR
    - Extra 6.874 problem on Pset2 – see answer key when released.

Irreproducible Discovery Rate (IDR)

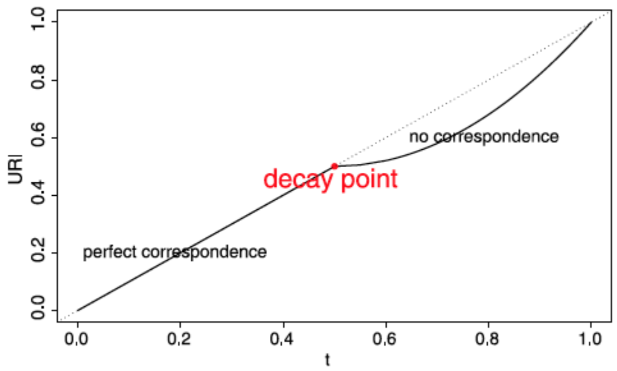
- For two replicates of an experiment, how do we choose which events are reproducible between the two replicates?

  - In each of the replicates, order the events by their significance (p-value)

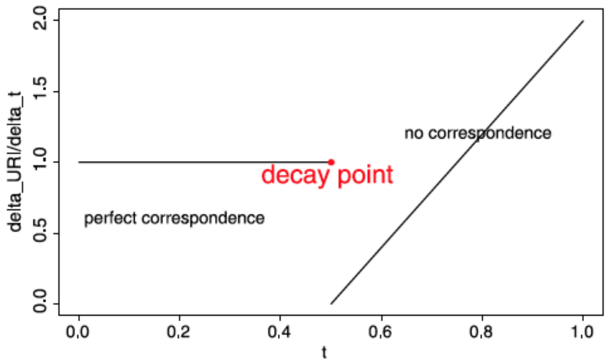
  - The idea is that a reproducible event should have approximately the same rank in the two events – the most significant effect should have rank 1 in both replicates, etc.

Irreproducible Discovery Rate (IDR)

- Two replicates: X and Y, each of which has *n* events ranked by significance
  - Let *t* vary from 0 to 1 (the proportion of the n total events we're considering, starting with the most significant).
  - Ψ_n(t) = the fraction the *n* events that are in the top (n*t) events of both replicates.



(a) Ψ_n



(b) Ψ'_n

- Note that Ψ_n(t) = (n*t) / n = t if the top *t* of the events are the same in both of the replicates
- Ψ_n(t) < t if some events are in the top *t* of one of the replicates but not of the other

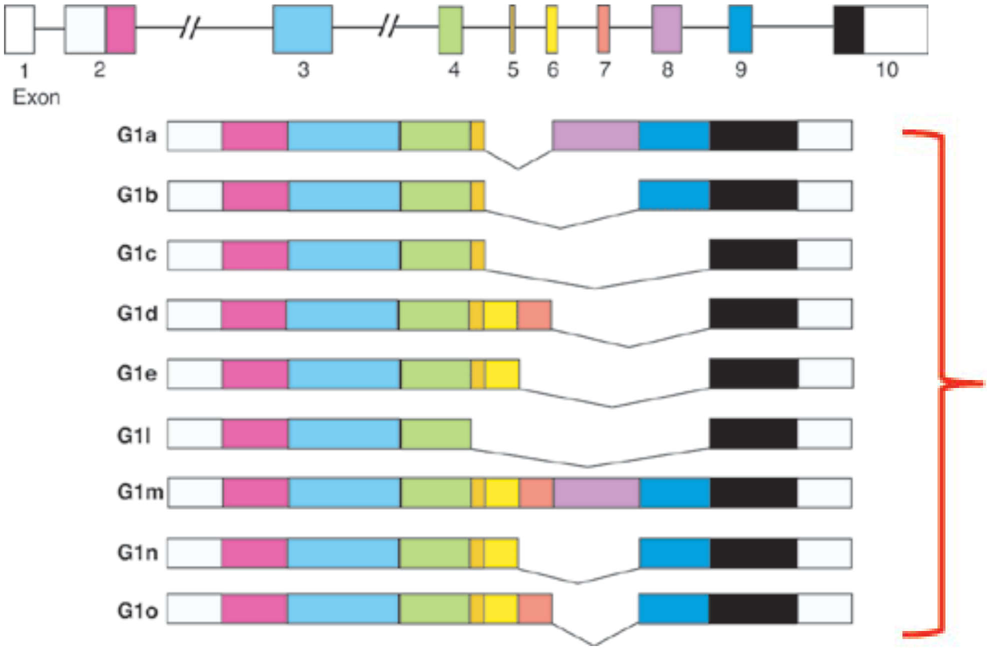
- Derivative Ψ'_n(t) = 1; these events are "reproducible"
- Ψ'_n(t) drops to below 1 beginning at the irreproducible events

RNA-Seq Analysis

- Central Dogma: DNA ➔ mRNA ➔protein
  - pre-mRNA contains not only protein coding exons, but non-coding regions: 5’- and 3’-UTR, introns, poly(A) tail
  - In metazoans, introns must be spliced out to create mature mRNA that can be translated into protein; some exons may also be spliced out (alternative splicing to create different mRNA isoforms of the same gene)

RNA-Seq Analysis

- Central Dogma: DNA ➔ mRNA ➔protein
  - pre-mRNA contains not only protein coding exons, but non-coding regions: 5’- and 3’-UTR, introns, poly(A) tail
  - In metazoans, introns must be spliced out to create mature mRNA that can be translated into protein; some exons may also be spliced out (alternative splicing to create different mRNA isoforms of the same gene)



Gene has 10 exons – exons 5-9 are alternatively spliced

Different mRNA isoforms (mature mRNAs) of this gene that create different proteins

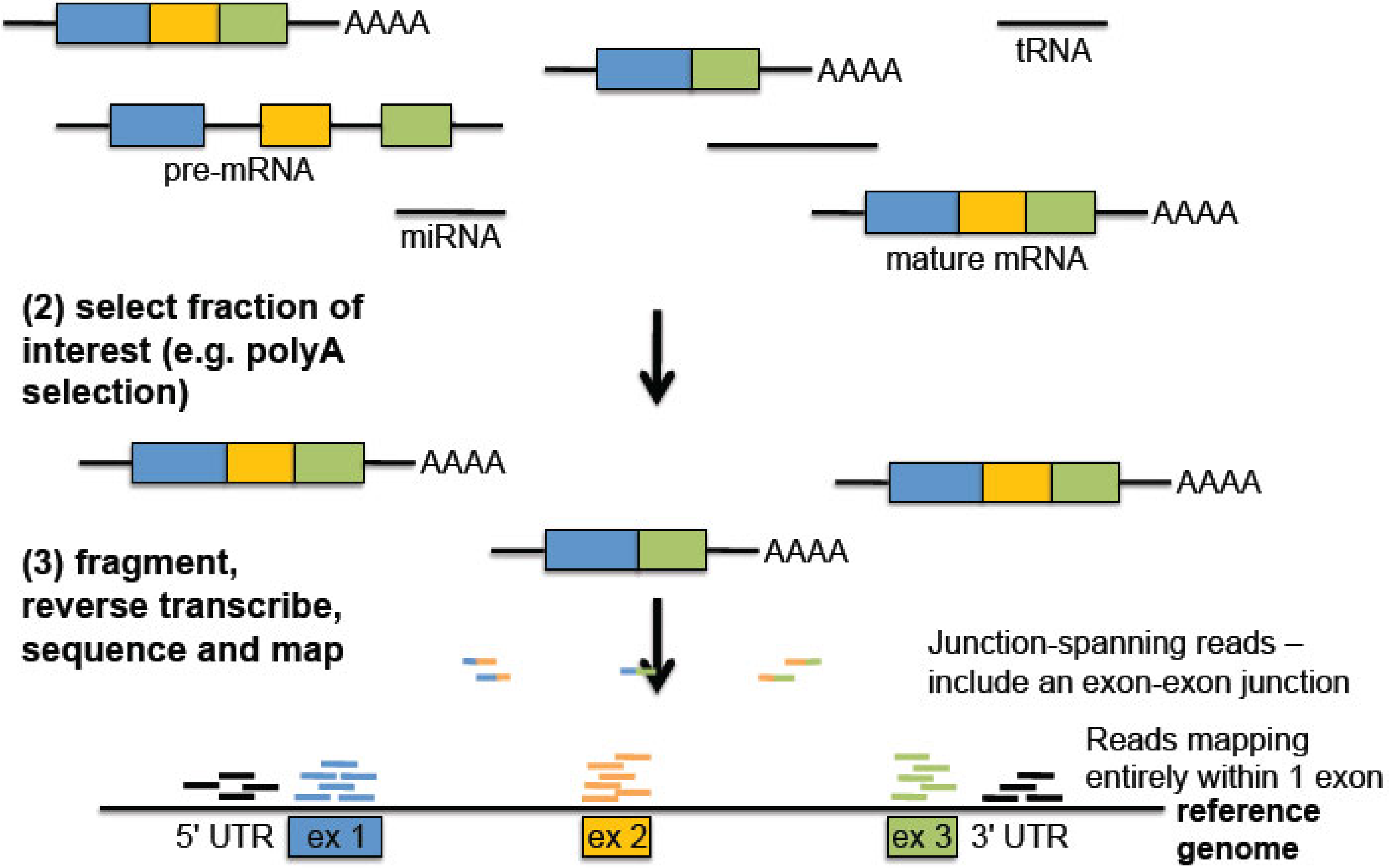
brain.riken.jp

RNA-Seq Analysis

- Central Dogma: DNA ➔ mRNA ➔protein
  - pre-mRNA contains not only protein coding exons, but non-coding regions: 5'- and 3'-UTR, introns, poly(A) tail
  - In metazoans, introns must be spliced out to create mature mRNA that can be translated into protein; some exons may also be spliced out (alternative splicing to create different mRNA isoforms of the same gene)
    - Alternative splicing allows great diversity from only ~20,000 genes: different alternatively spliced transcripts (and proteins) are regulated throughout development and in different cell types to create complexity required for higher organisms
    - Most genes (~95%) are alternatively spliced in humans
    - Disease relevant: ~15% of disease SNPs lie in splice sites and another ~20% lie in regulatory elements affect splicing
- We'd like to know what mRNA isoforms of gene are present in cells

RNA-seq Protocol

(1) isolate total RNA



(2) select fraction of interest (e.g. polyA selection)

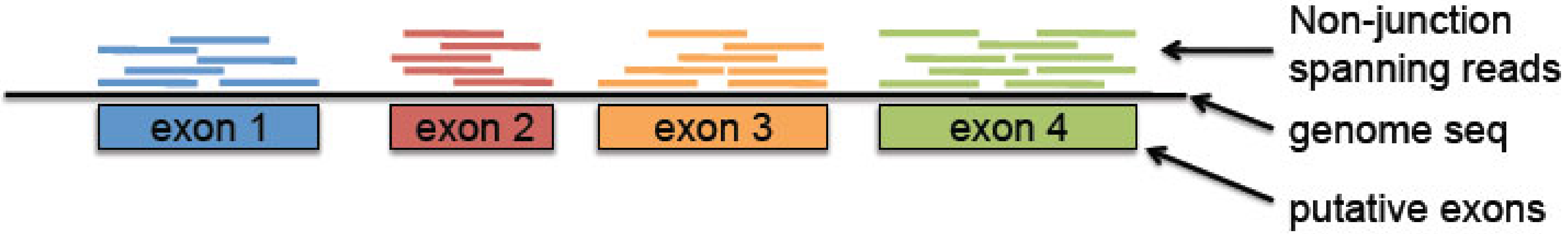
(3) fragment, reverse transcribe, sequence and map

Junction-spanning reads – include an exon-exon junction

Reads mapping entirely within 1 exon

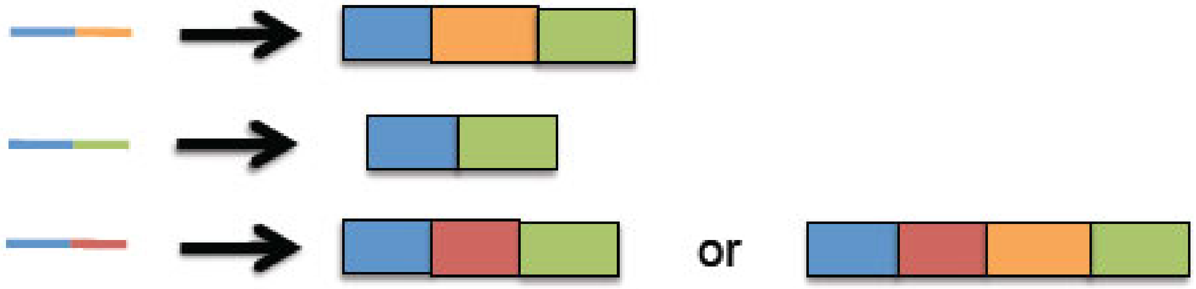
RNA-seq: identifying isoforms

- Some reads map completely within a single exon – don’t directly tell us which isoforms are present, although expression levels of different exons can be helpful (e.g. twice as many exon 1 reads compared to exon 4 – probably some isoforms that include exon 1 but not exon 4)



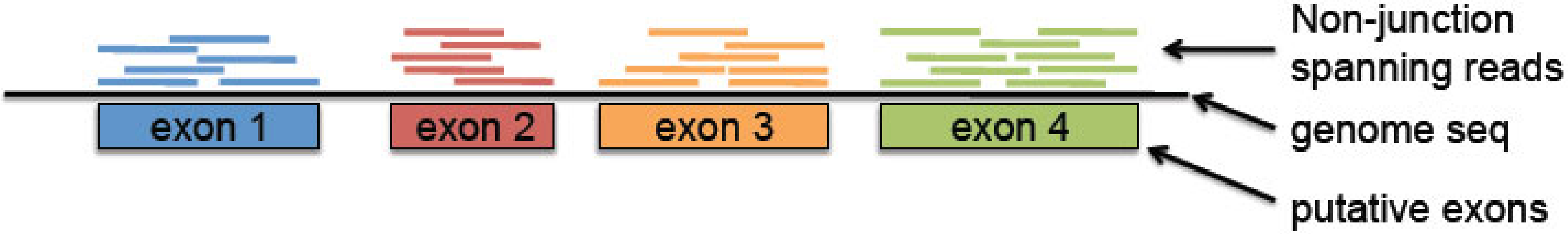
- How do we directly identify the isoforms that generated these reads? Look at junction-spanning reads!

- Assuming exons 1 and 4 must be included, which isoform(s) are consistent with the following reads?



RNA-seq: identifying isoforms

- Some reads map completely within a single exon – don't directly tell us which isoforms are present, although expression levels of different exons can be helpful (e.g. twice as many exon 1 reads compared to exon 4 – probably some isoforms that include exon 1 but not exon 4)



- How do we directly identify the isoforms that generated these reads? Look at junction-spanning reads!

- Since reads are generally 100bp or shorter, most reads only span 1 junction to give adjacent exons present in isoforms – assembling the full isoforms of 5-10+ exons and estimating their expression levels from only adjacent exon pairs is difficult
    - Promise in longer read (kb) technologies (e.g. Pacific Biosciences, Oxford Nanopore sequencing)

RNA-seq: quantifying isoforms

- we want to find the proportions $\Psi_1, ..., \Psi_n$ of each of the *n* isoforms $T_1, ..., T_n$ after observing *m* reads $R_1, ..., R_m$
- In other words, find abundances Ψ that maximize the likelihood of observing our reads *R:*

$$\Psi = \underset{\Psi}{argmax} L(R|\Psi)$$

- You can estimate the expression of each isoform using the total # of reads that map to the gene and Ψ

-for example, see Cufflinks (http://cole-trapnell-lab.github.io/cufflinks/) for a common program that does this

Likelihood ratio test

-Used to compare the fit of two models (the *null* model H_0 and the *alternative* model H_1) to the data

-The likelihood ratio Λ tells us how many times more likely our observed data x are under one model than the other:

$$\Lambda = \frac{\mathcal{L}(H_1|x)}{\mathcal{L}(H_0|x)} = \frac{P(x|H_1)}{P(x|H_0)}$$

-since the model with more parameters (usually H_1) will always fit at least as well as the other (therefore $\mathcal{L}(D|H_1) \geq \mathcal{L}(D|H_0)$), we can't just choose the model with the greatest likelihood – we can only choose H_1 if it is "significantly" better – how to decide?

- compute the test statistic *T* : $T = 2ln\Lambda$

- *T* is asymptotically χ^2 distributed with *df* equal to the difference in the # of free parameters between the two hypotheses

DEseq

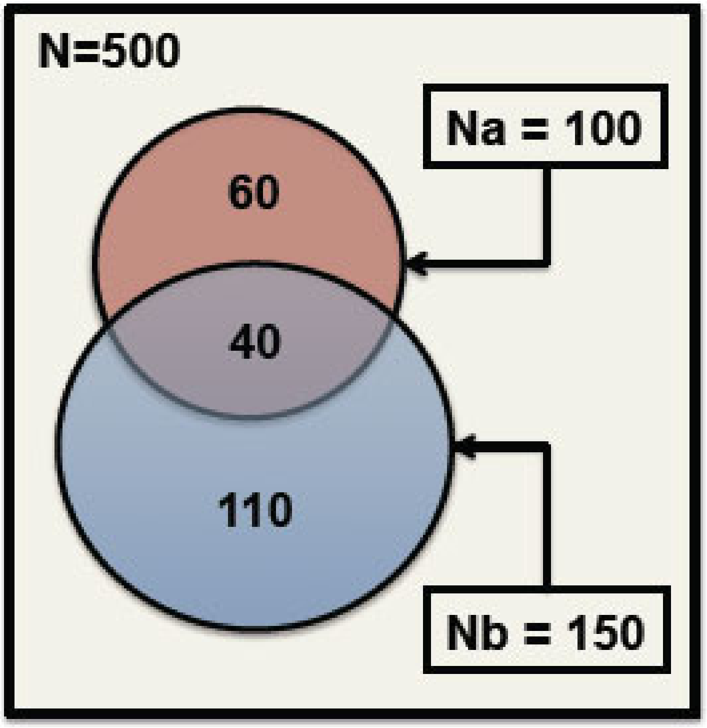
- we would like to know whether, for a given *region* (e.g. gene, TF binding site, etc.), an observed difference in read counts between different biological conditions is significant
- assume the number of reads in sample *j* that are assigned to region (gene) *i* is approx. distributed according to the negative binomial:

$$K_{ij} \sim NB(\mu_{ij}, \sigma_{ij}^2)$$

- the NB has two parameters, which we need to estimate from the data, but typically the # of replicates is too small to get good estimates, particularly for the variance for region i
- if we don't have enough replicates to get a good estimate of the variance for region i under condition $\rho(j)$, DEseq will pool the data from regions with similar expression strength to try to get a better estimate
- From these estimates of the background variability among samples in the same condition, we can determine if the variance between different conditions is significantly greater than that observed between samples within the same condition

Hypergeometric Test: when you want to know if overlap between two subsets is significant

- From DESeq, we identified genes differentially expressed between control and treatment after treatment with two different stress conditions: (A) heat shock and (B) oxidative stress
- We propose that the pathways involved in the responses to A and B are similar, so the genes affected by A might overlap with the genes affected by B
- We observe the following:

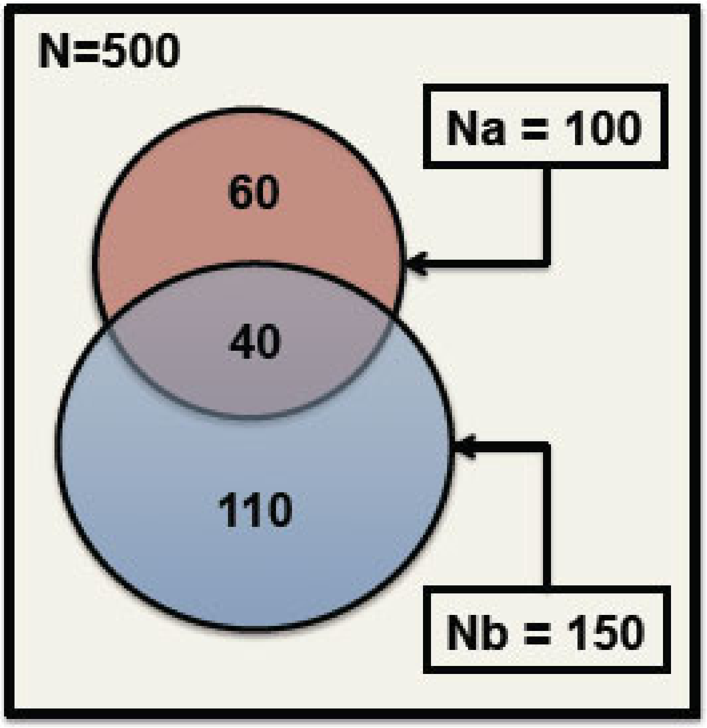


N = total # of genes measured = **500**
Na = total # genes changed in A = **100**
Nb = total # genes changed in B = **150**
k = genes changed in both A and B = **40**

Is this overlap significant (e.g. unlikely by chance)?
-> do a hypergeometric test

Hypergeometric Test

- Say you have already chosen the Na items of set A – now you just need to choose set B. Then probability of observing exactly *k* items overlapping among *Na* and *Nb* size groups drawn from *N* total items is:



$$\frac{\text{\# total ways of choosing B, given } k \text{ must overlap with A}}{\text{\# total ways of choosing B w/o any constraints}}$$

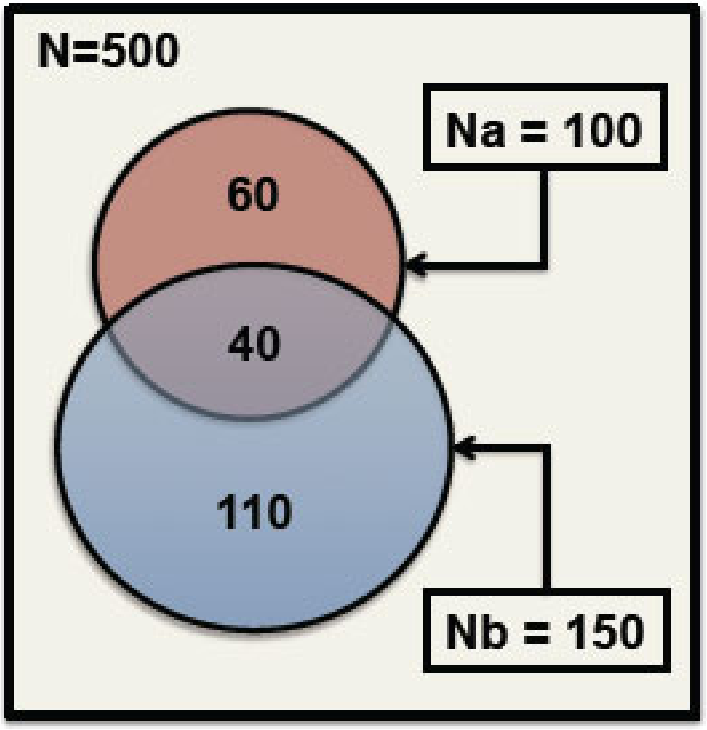
$$\frac{\text{(choose } k \text{ from A to be in B's overlap)(choose } n_b \quad k \text{ of B that don't overlap, from pool of } N - n_a \text{ non-A's)}}{\text{(choose any } n_b \text{ for B)}}$$

$$\implies P(k; n_a, n_b, N) = \frac{\binom{n_a}{k}\binom{N-n_a}{n_b-k}}{\binom{N}{n_b}}$$

Hypergeometric Test

The probability of observing exactly *k* items overlapping among *Na* and *Nb* size groups drawn from *N* total items is

$$P(k;n_a,n_b,N) = \frac{\binom{n_a}{k}\binom{N-n_a}{n_b-k}}{\binom{N}{n_b}}$$



Our p-value is the probability of observing an overlap *at least as extreme* as the overlap we observed (which is *k*):

max value for *k* (smaller set completely contained within larger set)

$$P(x \geq k) = \sum_{i=k}^{min(n_a,n_b)} P(i;n_a,n_b,N)$$

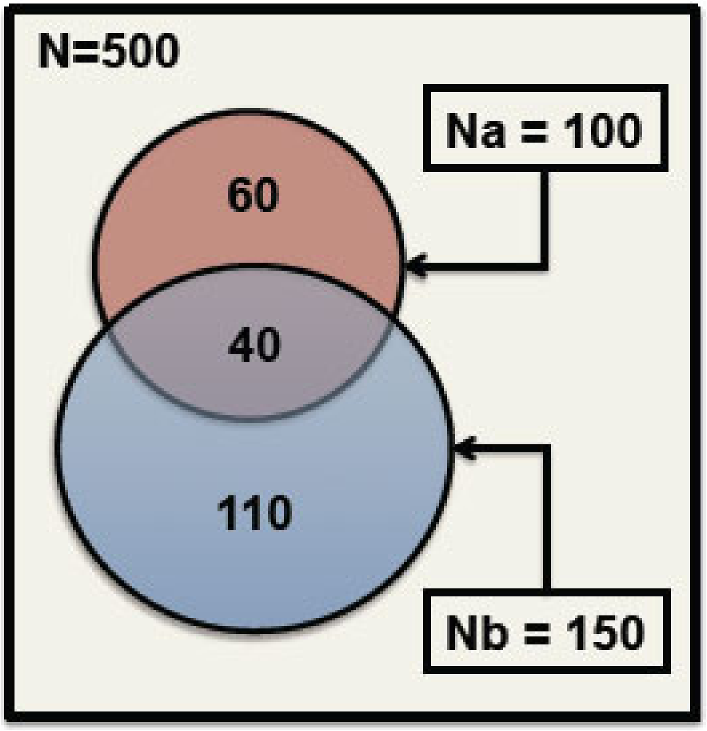
Hypergeometric Test

For this example, we obtain:

$$P(x \quad 40) = \sum_{i=40}^{100} P(i; 100, 150, 500)$$

$$= \sum_{i=40}^{100} \frac{\binom{100}{i}\binom{500-100}{150-i}}{\binom{500}{150}}$$

$$= 0.0112$$

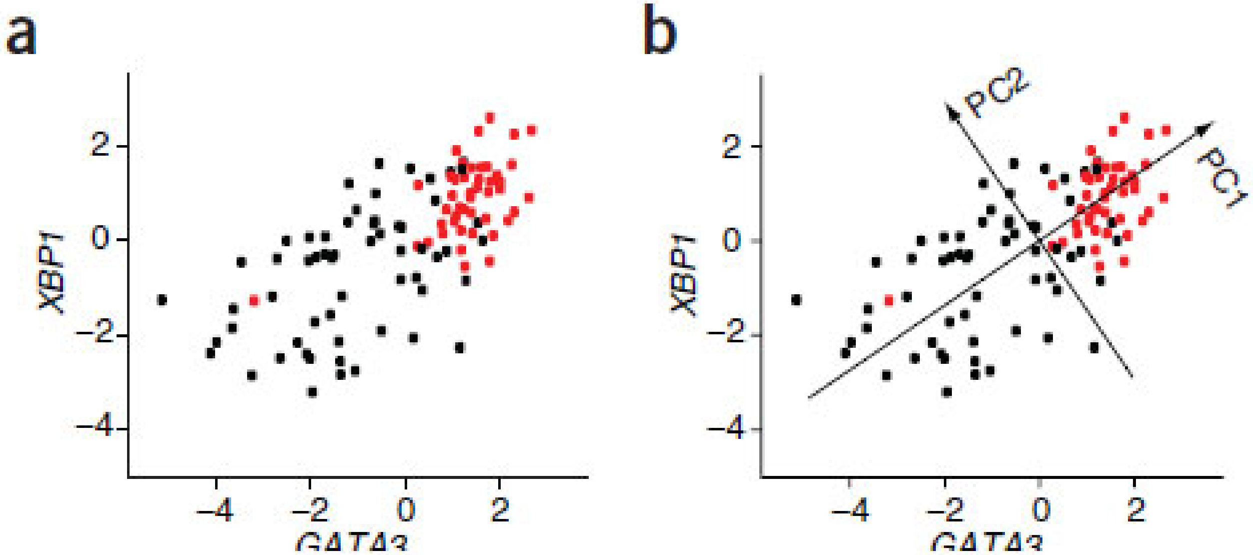


Therefore, with α = 0.05, we reject the null hypothesis that the overlap between conditions A and B are due to random chance, suggesting there is some similarity between gene expression changes caused by heat shock and oxidative stress

Principal Component Analysis (PCA)

- Typical high-throughput biological experiment: thousands of measurements (e.g. 20,000 gene expression levels)
  - High dimensional data are hard to visualize and interpret
- Can use Principal Component Analysis (PCA): mathematical algorithm for reducing the dimensionality of the data while retaining most of the variation in the data set
  - Accomplishes this reduction by identifying directions, called principal components, along which the variation in the data is maximal
  - By using a few components, each sample can be represented by relatively few numbers instead of thousands values for thousands of variables
  - Can then plot samples in 2D or 3D space by using the top 2 or 3 principal components

PCA identifies the directions along which the data have the largest spread



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Ringnér, Markus. "What is Principal Component Analysis?"
Nature Biotechnology 26, no. 3 (2008): 303-4.

-1st **principal component (PC1)** is the direction of maximal variation among your data

- Magnitude of this component is related to how much variation there is in this direction

-2nd **principal component (PC2)** is next direction of remaining maximal variation in your sample that is perpendicular to PC1

- Magnitude of this component will be smaller than that of 1st

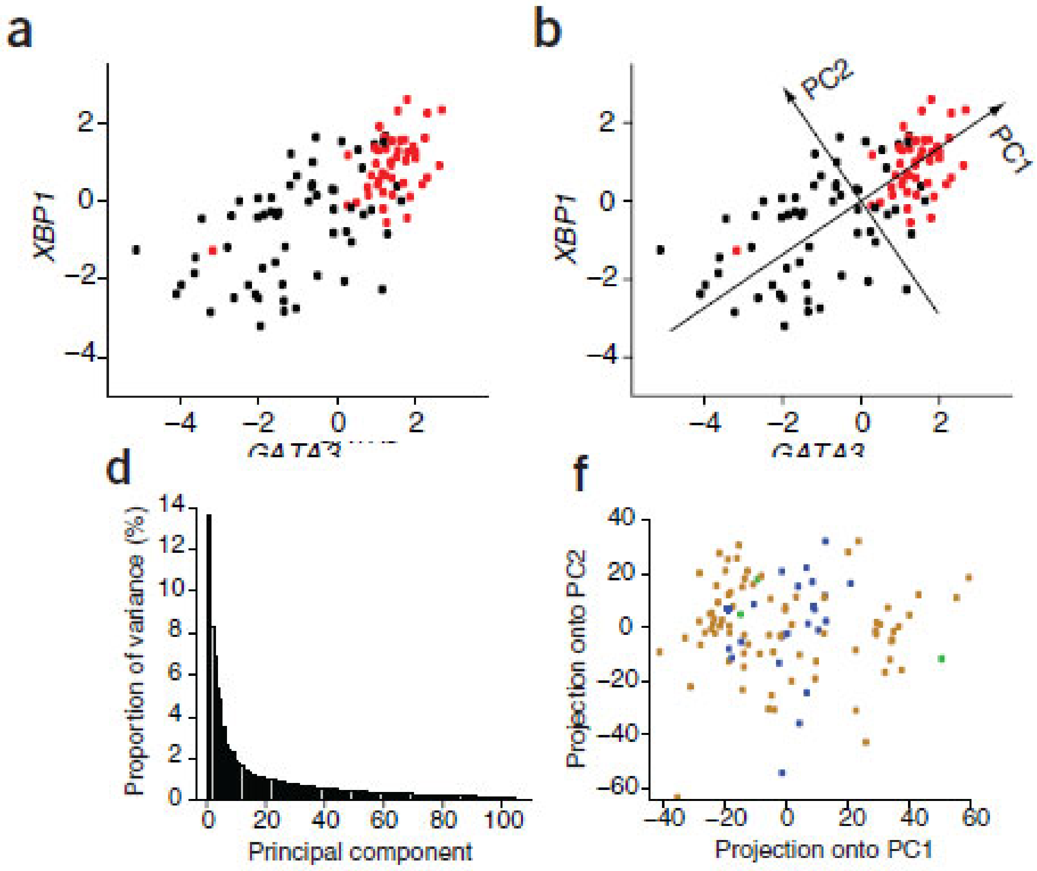
- and so on for PC3, PC4, etc. (each PC must be orthogonal to all previous PCs)

See 2 page Nature Biotech Primer:
http://www.nature.com/nbt/journal/v26/n3/pdf/nbt0308-303.pdf

PCA identifies the directions along which the data have the largest spread

-Succssive principal components explain smaller and smaller amounts of variance in the data –this is why summarizing data with first 2 or 3 components is an OK first approximation of data

- This example: first 2 components retain 22% of total variance;
63 components retain 90% of variance



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Ringnér, Markus. "What is Principal Component Analysis?" *Nature Biotechnology* 26, no. 3 (2008): 303-4.

-Visualizing data points as projections onto first PCs often reveals “clustering” of samples into groups – but make sure these are biologically relevant and not technical artifacts (e.g., samples cluster into 2 groups based on which of two different days libraries were prepared)

See 2 page Nature Biotech Primer:

http://www.nature.com/nbt/journal/v26/n3/pdf/nbt0308-303.pdf

Single-cell RNA-seq

- Bulk cell RNA-seq only captures average behavior of millions of cells, but individual cells in the population can have different behavior
- Solution: single-cell RNA-seq
  - Instead of taking an aliquot of millions of cells to prepare a library, first sort single cells into wells and then do each library prep on each individual cell
  - Each cell has its own 6nt barcode in adapter; can then pool libraries from multiple cells together to sequence on one flow cell

- Caveats:
  - Library prep with such little starting RNA from 1 cell is technically very challenging
  - Much more likely that random sampling during library prep will produce strong biases, further amplified by PCR
  - For example, if a transcript is very lowly expressed, you might have only one or a few molecules in your single cell RNA sample - easily lost due to stochastic sampling during library prep
  - Hard to interpret "negative" results of a gene or isoform not being expressed – is it actually not expressed in the cell, or did you just lose it during library prep?