6.874 Recitation

3-7-13

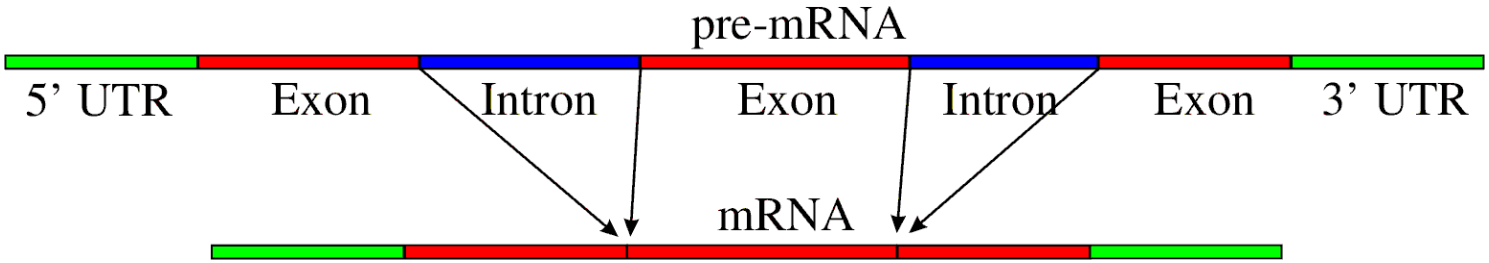
DG Lectures 8 + Topic Models

Announcements

- Project specific aims due Sunday
  - Look at NIH examples
- Pset #2 due in 1 week (03/13)
  - 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
  - Mean dispersion function problem

RNA-Seq Analysis

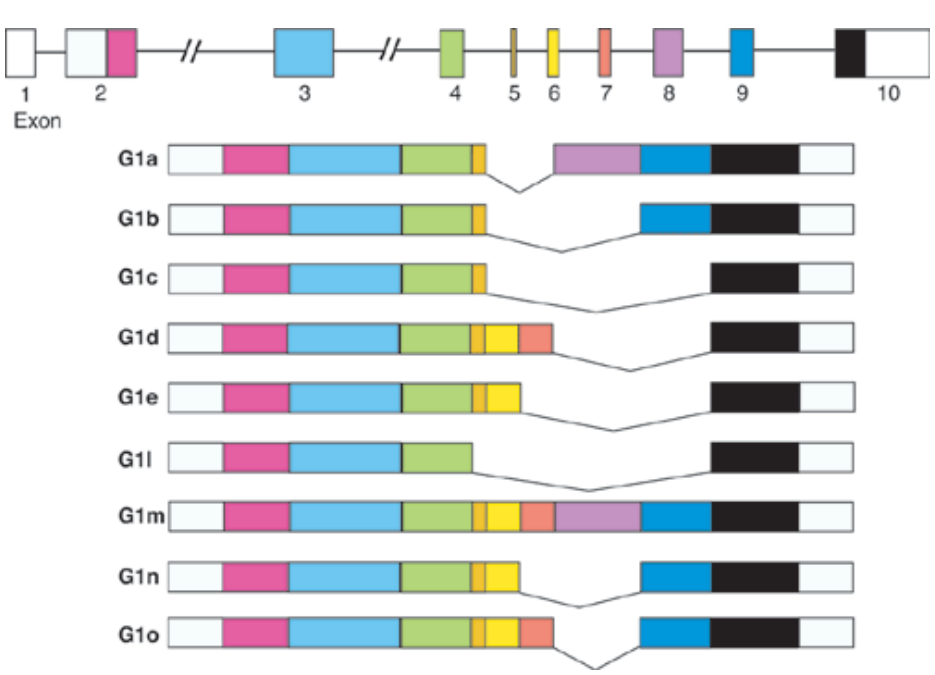
- Central Dogma: DNA → mRNA →protein
  - pre-mRNA contains not only protein coding exons, but non-coding regions: 5'- and 3'-UTR, introns, poly(A) tail
  - Introns must be spliced out to create mature mRNA that can be translated into protein
  - Some exons may also be spliced out (alternative splicing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



- We'd like to know what mRNA isoforms of gene are present in cells

RNA-Seq Analysis – Alternative Splicing

- Central Dogma: DNA ➔ mRNA ➔protein
  - some exons may also be spliced out (alternative splicing to create different mRNA isoforms of the same gene)



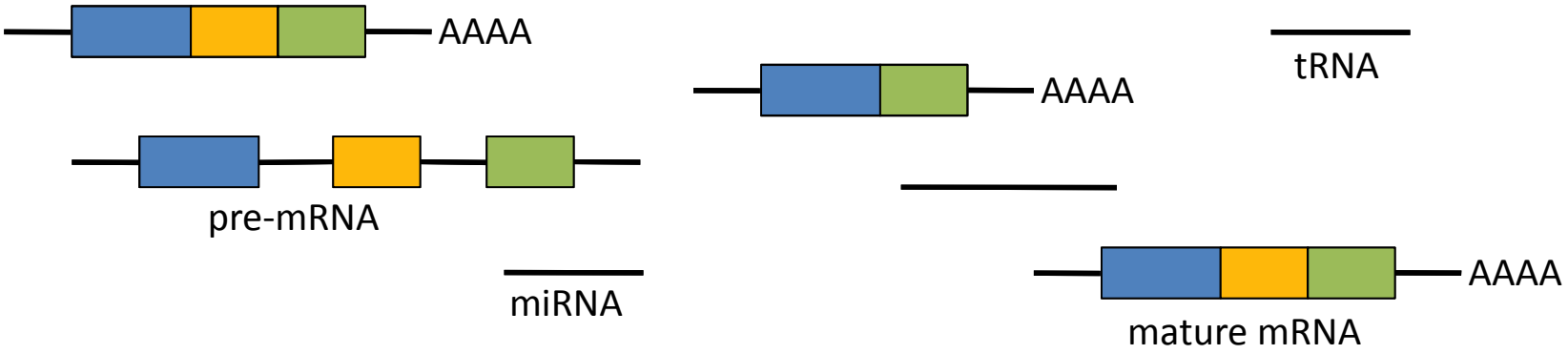
Gene has 10 exons – exons 5-9 are alternatively spliced

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

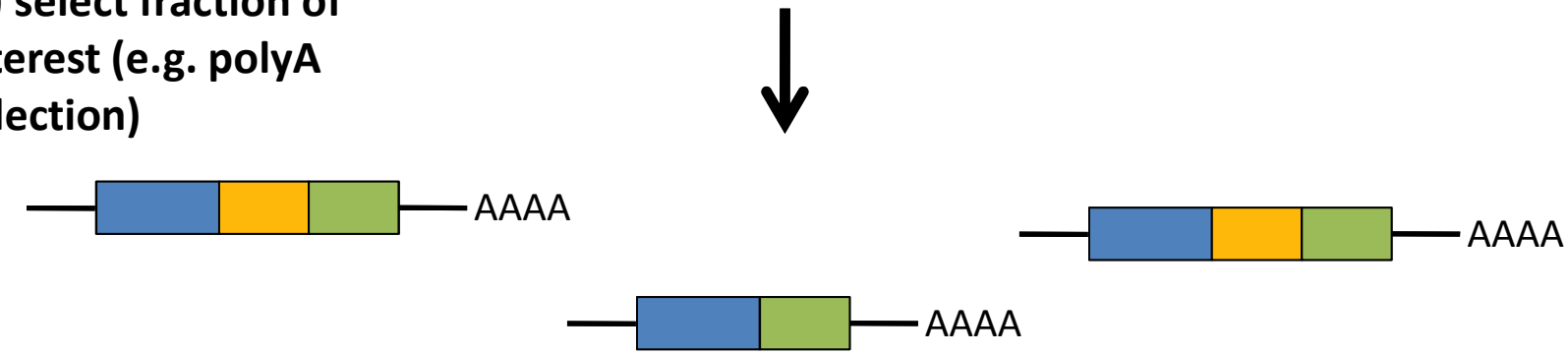
Source: Aoki-Suzuki, Mika, Kazuo Yamada, et al. "A Family-based Association Study and Gene Expression Analyses of Netrin-G1 and-G2 Genes in Schizophrenia." *Biological Psychiatry* 57, no. 4 (2005): 382-93.

RNA-seq Protocol

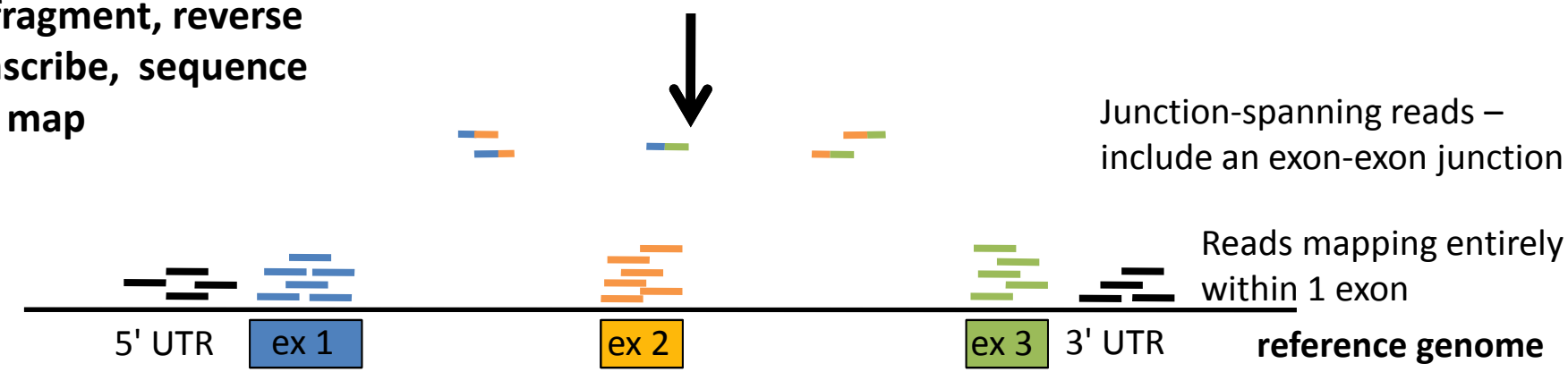
(1) isolate total RNA

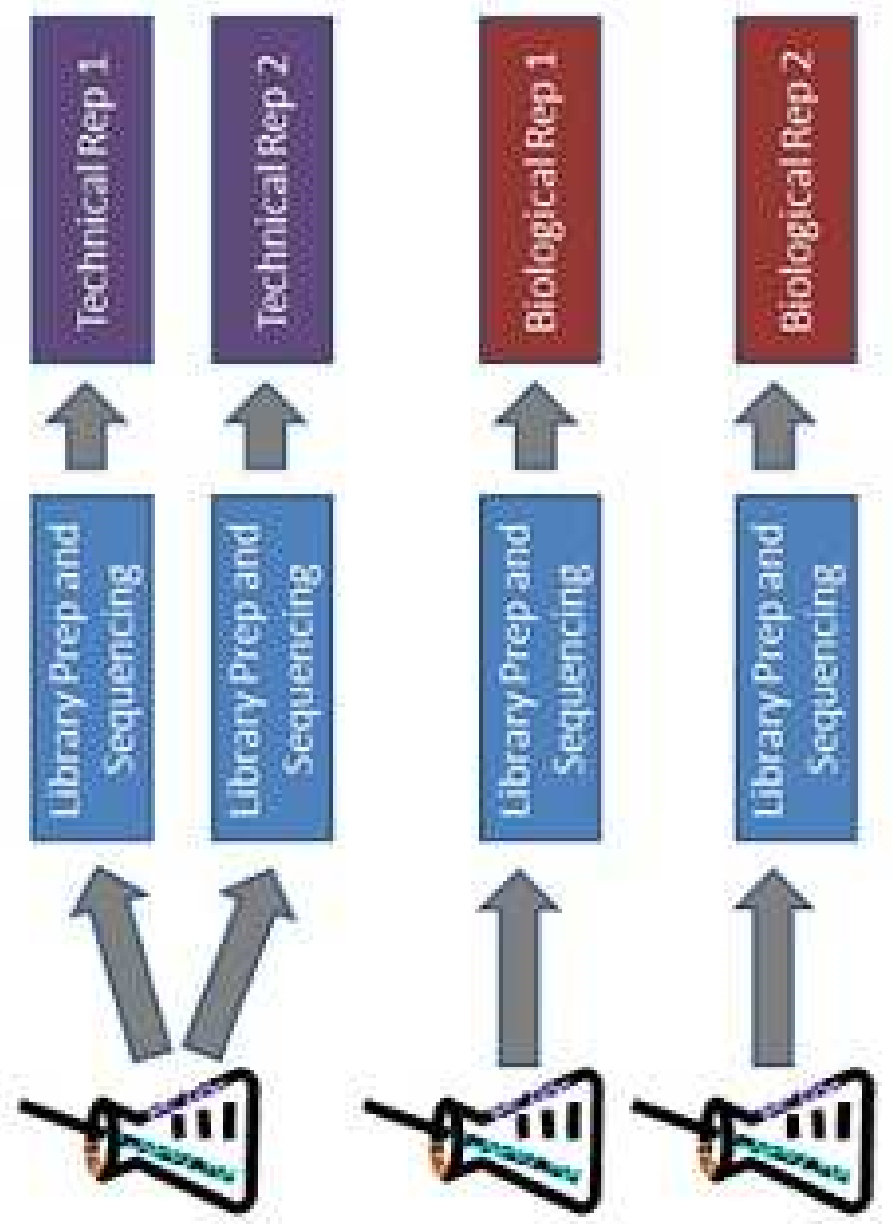


(2) select fraction of interest (e.g. polyA selection)



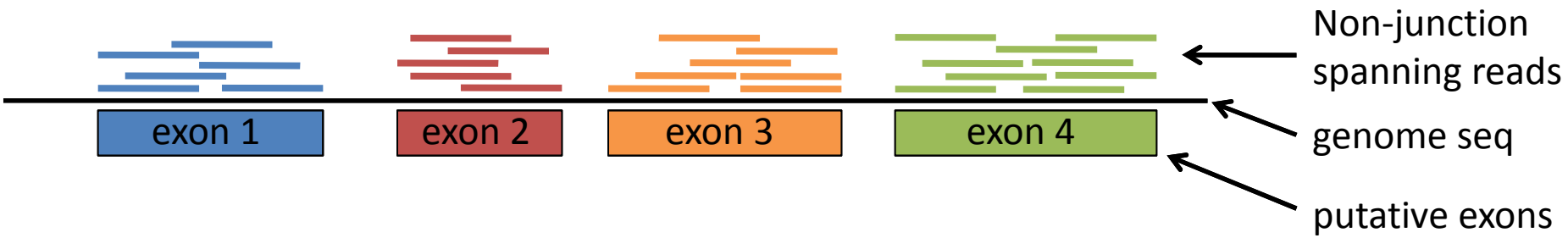
(3) fragment, reverse transcribe, sequence and map





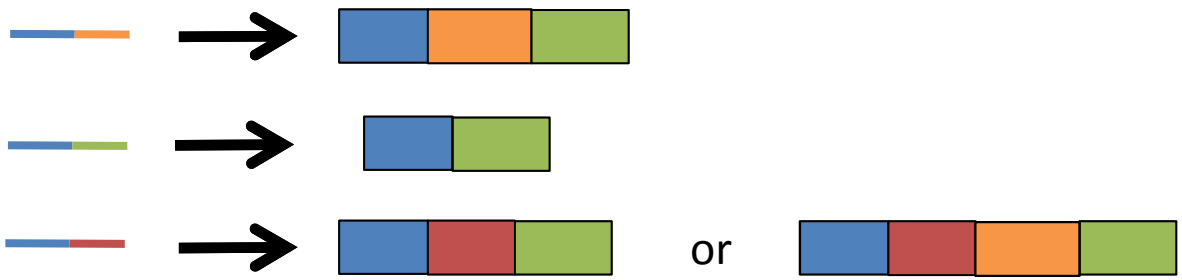
RNA-seq: identifying isoforms

- 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)



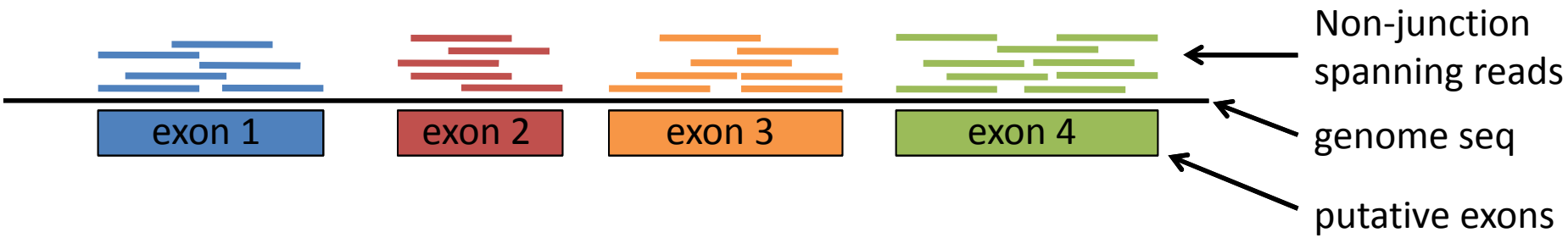
- How do we directly identify the isoforms that generated these reads? Look at junction-spanning reads!

- 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)

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 *i* is approx. distributed according to the negative binomial:

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

- 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
- we then test for significance using a LRT

DEseq

- the Likelihood Ratio Test is the ratio of the probability under the null model and the alternate model
- for example, if we are testing for whether there is significant difference in counts in condition A relative to B, we calculate:

$$T_i = 2log\frac{P(K_{iA}|H_a)P((K_{iB}|H_a)}{P(K_{iA}, K_{iB}|H_0)}$$

- for H_a, we allow the distribution of K_{iA} and K_{iB} to be different, while under H_0 we assume that K_{iA}, K_{iB} are drawn from the same distribution (e.g. isoform *i* is identically expressed under conditions A and B)

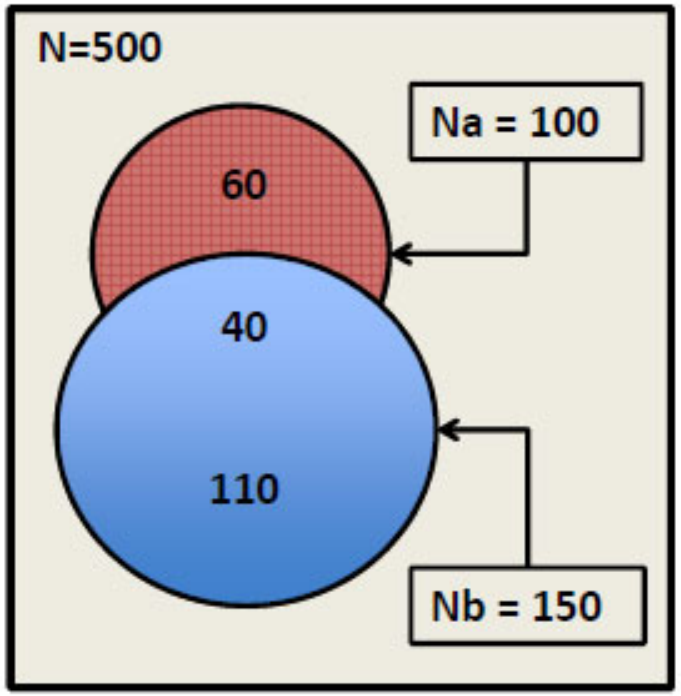
- then *T_i* follows a Chi Square distribution with *df* = 4 – 2 = 2

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**

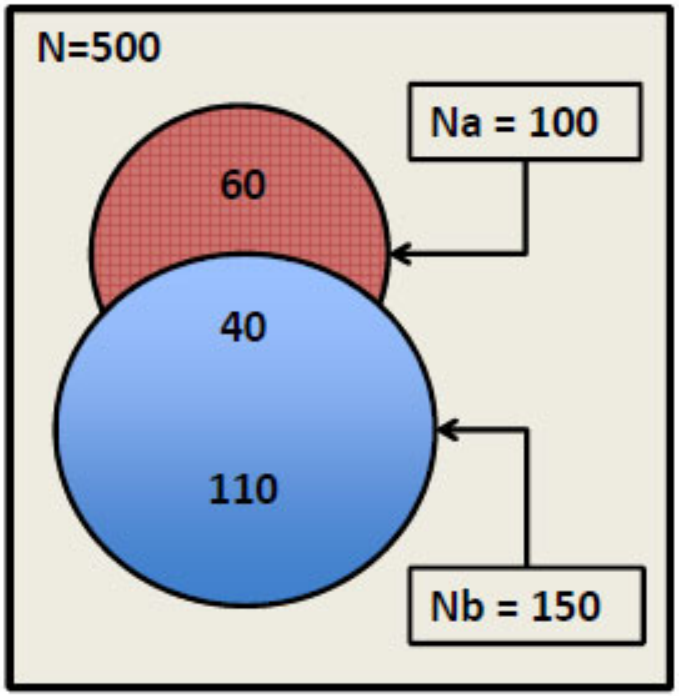
Is this overlap significant (e.g. unlikely by chance)?
-> do a hypergeometric test



Hypergeometric Test

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)

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

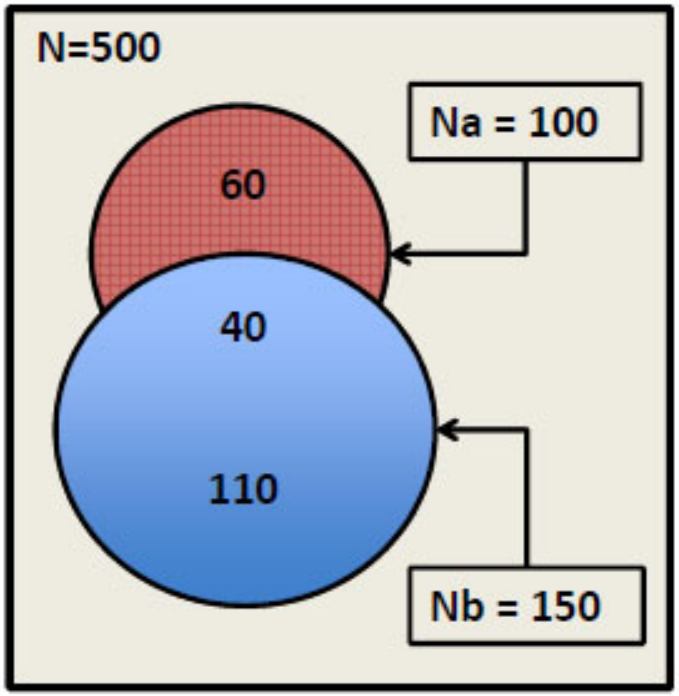
Hypergeometric Test

For this example, we obtain:

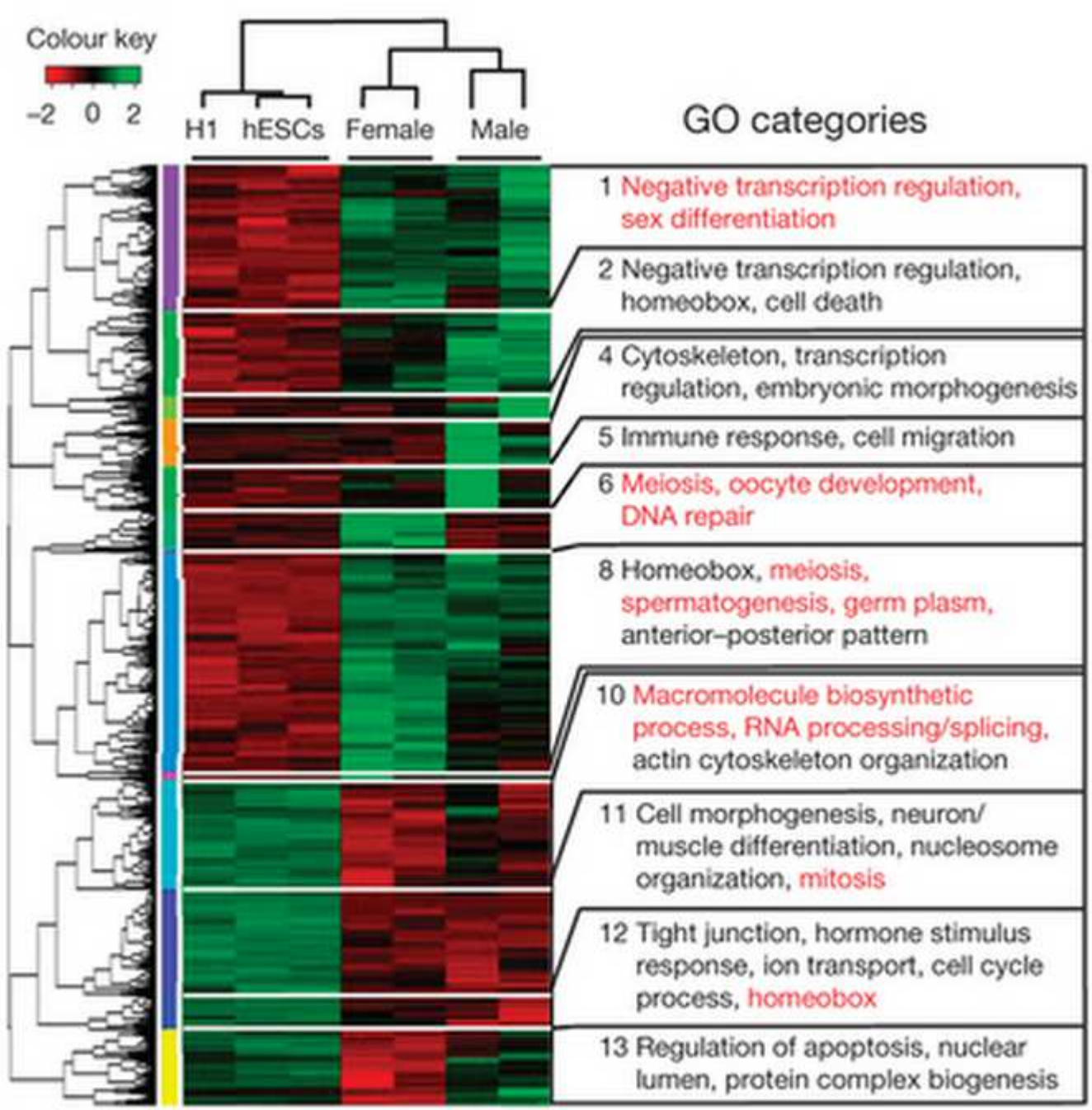
$$P(x \geq 40) = \sum_{i=40}^{100} P(i; 100, 150, 500)$$

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

$$= 0.0112$$

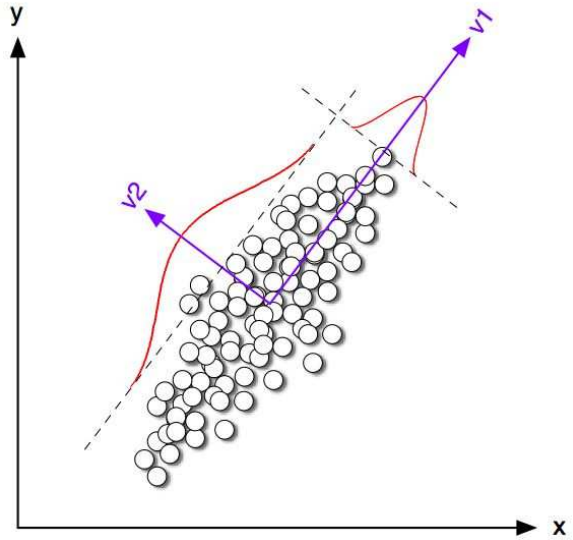


Therefore, with α = 0.05, we reject the null hypothesis that the overlap between conditions A and B are due to random chance, suggesting there is some similarity between gene expression changes caused by heat shock and oxidative stress



Source: Gkountela, Sofia, Ziwei Li, et al. "The Ontogeny of cKIT+ Human Primordial Germ Cells Proves to be a Resource for Human Germ Line Reprogramming, Imprint Erasure and in Vitro Differentiation." *Nature Cell Biology* 15, no. 1 (2013): 113-22.

PCA identifies the directions (PC1 and PC2) along which the data have the largest spread



-1st principal component is the direction of maximal variation among your sample

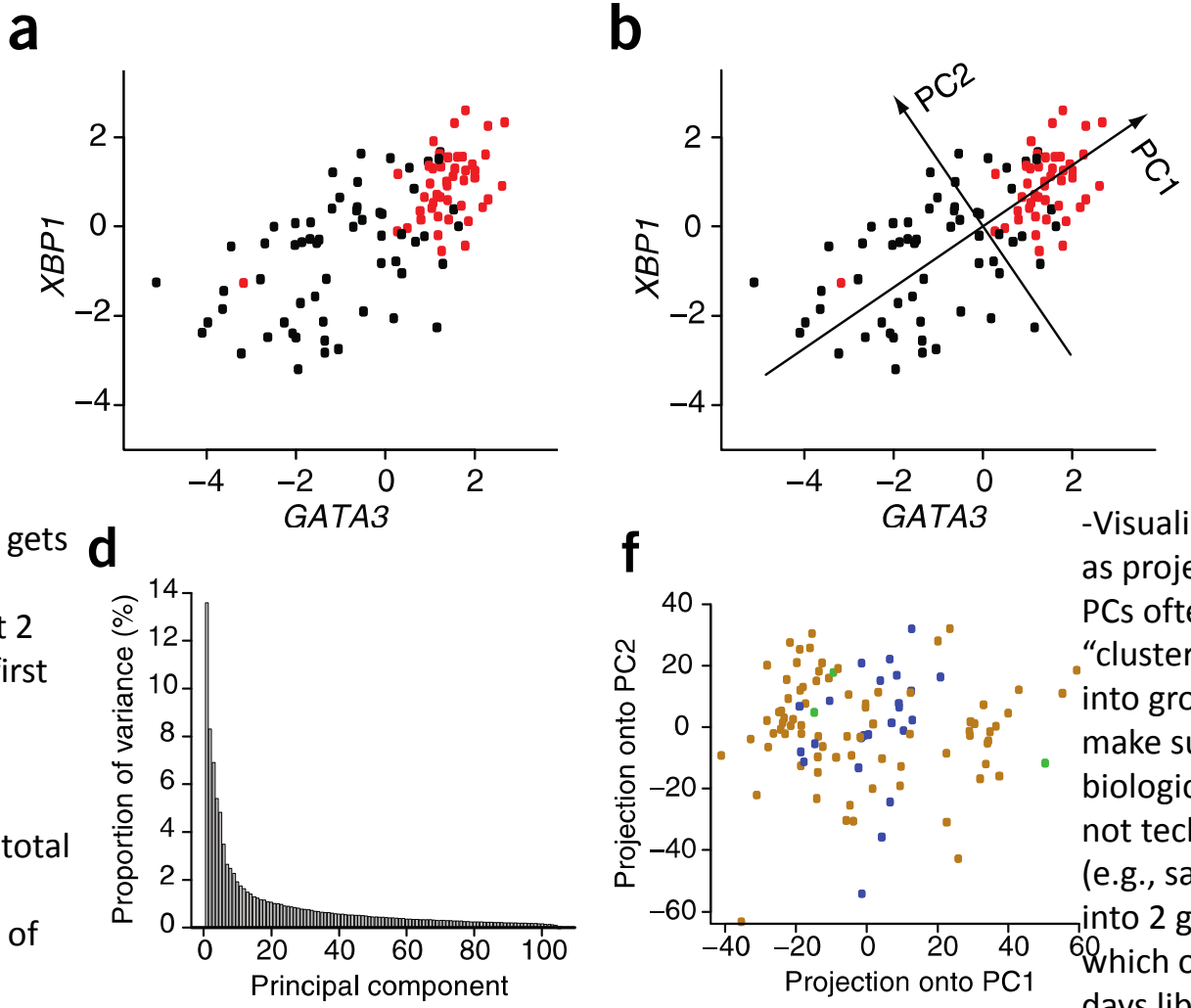
- Magnitude of this component is related to how much variation there is in this direction

-2nd principal component is next direction (orthogonal to 1st direction) of remaining maximal variation in your sample

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

etc.

PCA identifies the directions (PC1 and PC2) along which the data have the largest spread



-Each principal component gets smaller –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

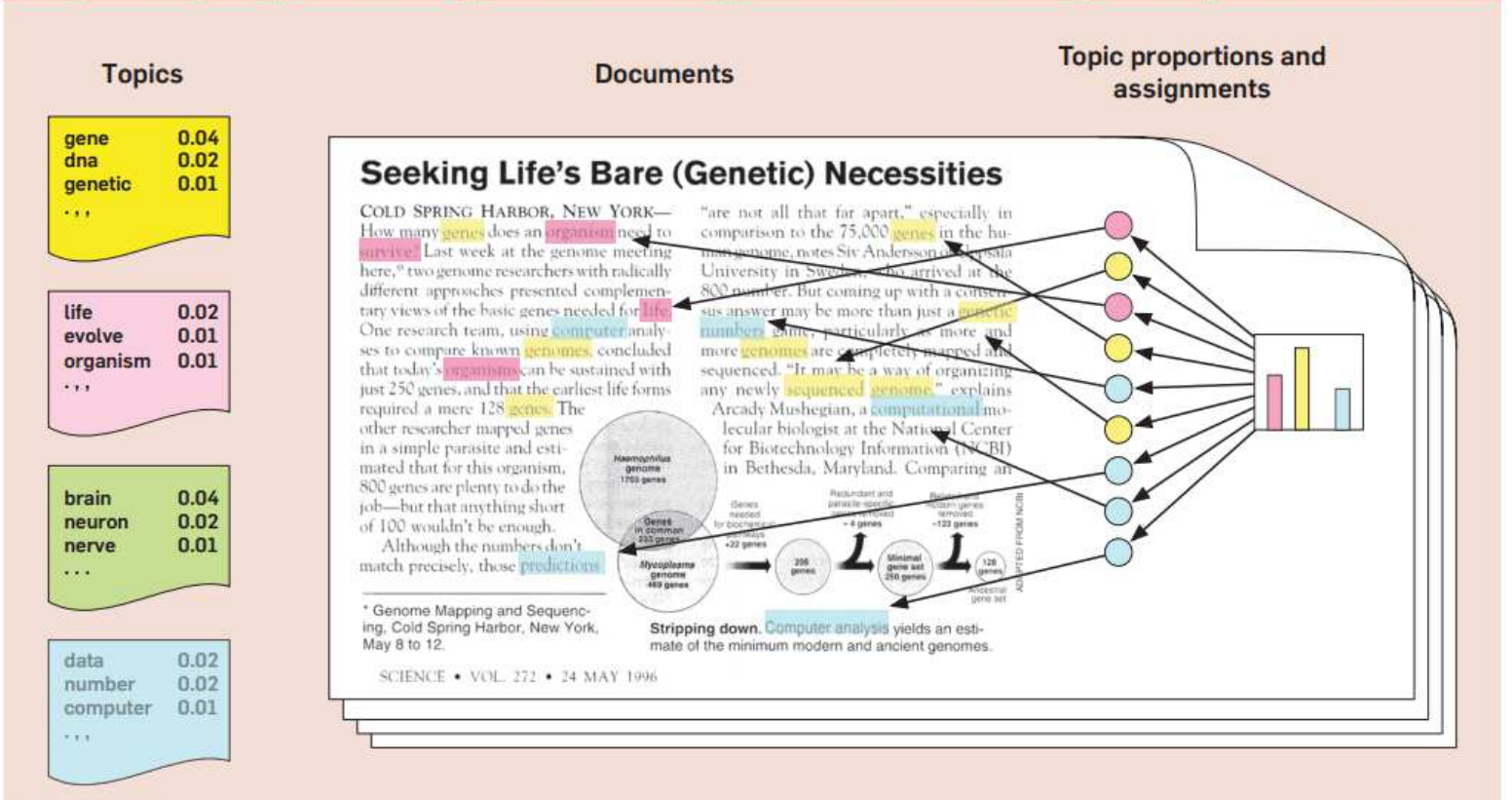
-Visualizing data points as projections onto first PCs often reveals "clustering" of samples into groups – want to make sure these are biologically relevant and not technical artifacts (e.g., samples cluster into 2 groups based on which of two different days libraries were prepared)

See 2 page Nature Biotech Primer:

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

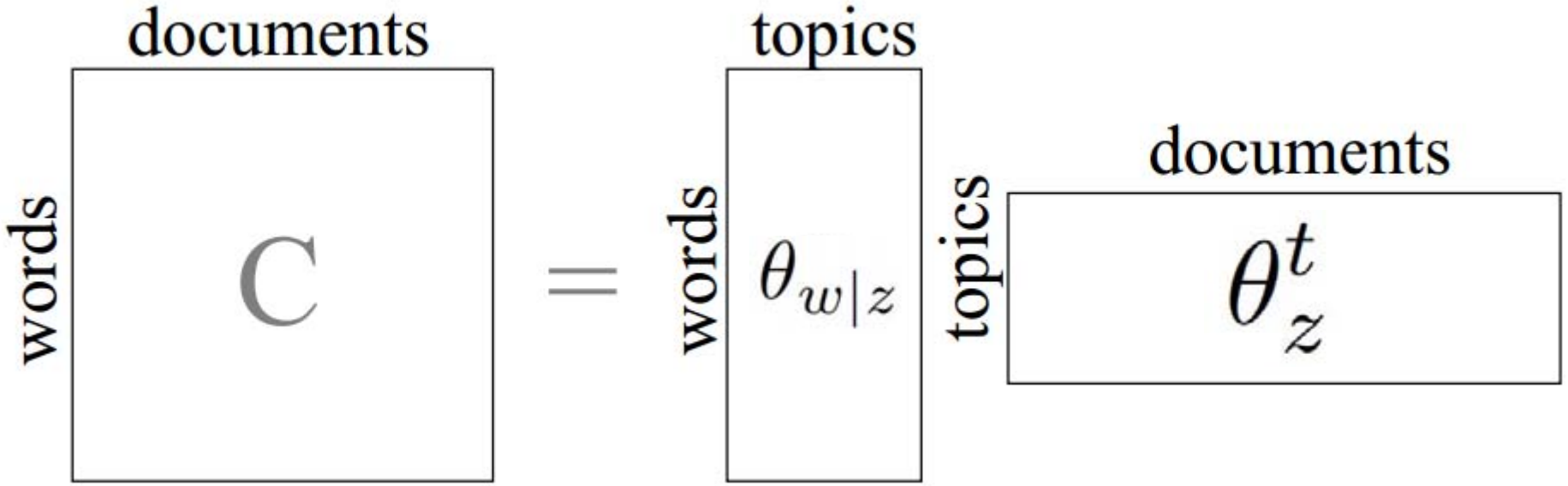
Topic Models

Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.



Source: Blei, David M. "Probabilistic Topic Models." *Communications of the ACM* 55, no. 4 (2012): 77-84.

Topic Models



Source: Landauer, Thomas K., Danielle S. McNamara, eds. "Handbook of Latent Semantic Analysis." *Psychology Press*, 2013.

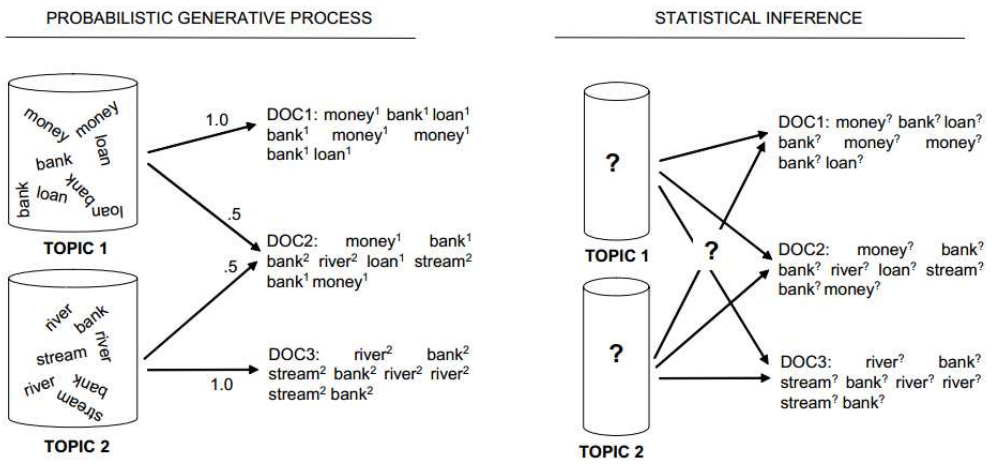
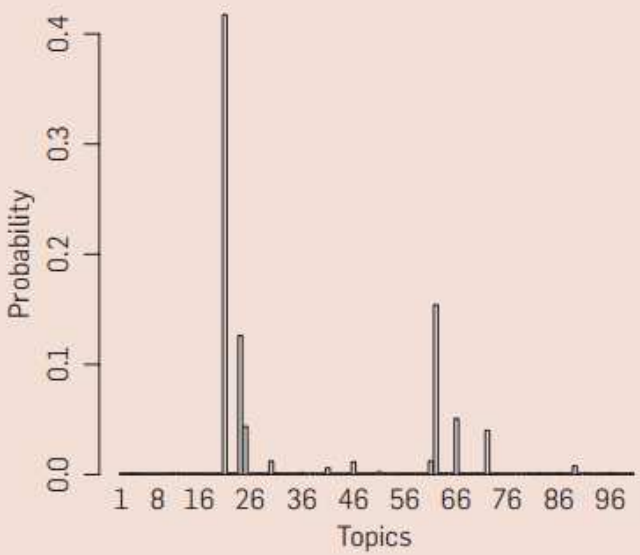


Figure 2. Illustration of the generative process and the problem of statistical inference underlying topic models

Topic Models

Figure 2. Real inference with LDA. We fit a 100-topic LDA model to 17,000 articles from the journal *Science*. At left are the inferred topic proportions for the example article in Figure 1. At right are the top 15 most frequent words from the most frequent topics found in this article.



"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Source: Blei, David M. "Probabilistic Topic Models." *Communications of the ACM* 55, no. 4 (2012): 77-84.

Single Topic Model

- Assume word drawn from a single topic

$$w \sim \text{Multinomial}(\theta), \text{ where } \theta_w \geq 0, \sum_w \theta_w = 1$$

- Probability assigned to a document

$$P(d) = \prod_{w \in d} \theta_w = \prod_w \theta^{n_w(d)}$$

$n_w(d)$ = # of times word *w* occurs in document *d*

Multiple Topics

- Each word is now attributed to a topic, z

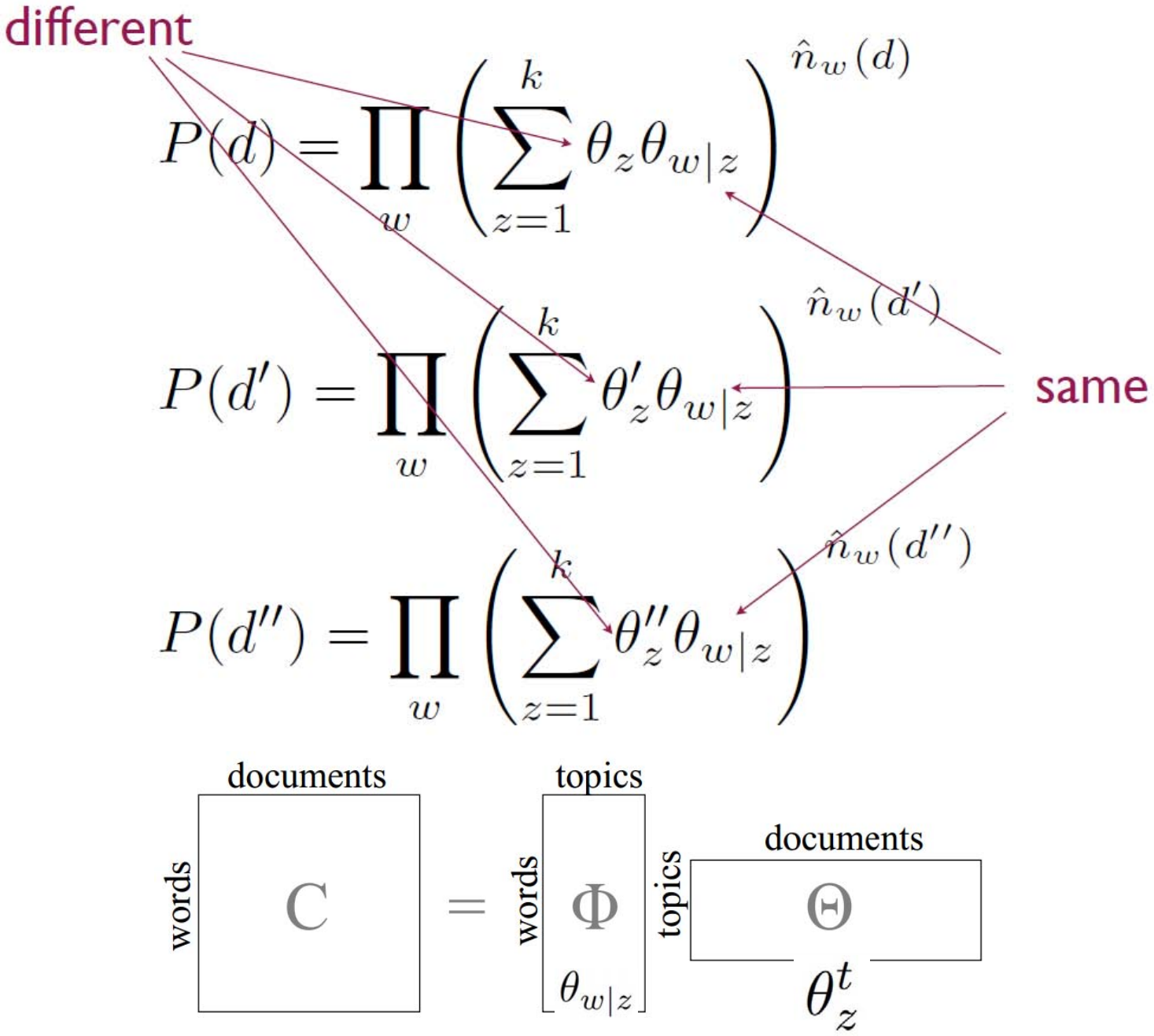
$$z \sim \text{Multinomial}(\theta)$$

- Words are now generated according to a topic specific distribution

$$w \sim \text{Multinomial}(\theta_z)$$

- Probability assigned to a document is now

$$P(d) = \prod_w \left(\sum_{z=1}^{k} \theta_z \theta_{w|z} \right)^{\hat{n}_w(d)}$$



Source: Landauer, Thomas K., Danielle S. McNamara, eds. Handbook of Latent Semantic Analysis. *Psychology Press*, 2013.

Estimation: EM-algorithm

- Estimate the parameters by maximizing the log-likelihood of the observed data

$$\sum_{t=1}^{n} \log P(d^t) = \sum_{t=1}^{n} \log \left[\prod_{w} \left(\sum_{z=1}^{k} \theta_z^t \theta_{w|z} \right)^{\hat{n}_w(d^t)} \right]$$

$$= \sum_{t=1}^{n} \sum_{w} n_w(d^t) \log \left(\sum_{z=1}^{k} \theta_z^t \theta_{w|z} \right)$$

EM-algorithm

- E-step:

$$P(z|w,t) = \frac{\theta_z^t \theta_{w|z}}{\sum_{z'=1}^{k} \theta_{z'}^t \theta_{w|z'}}$$

topic for a word w in document t

$$n(z|t) = \sum_w n_w(d^t) P(z|w,t)$$

topic usage in document t

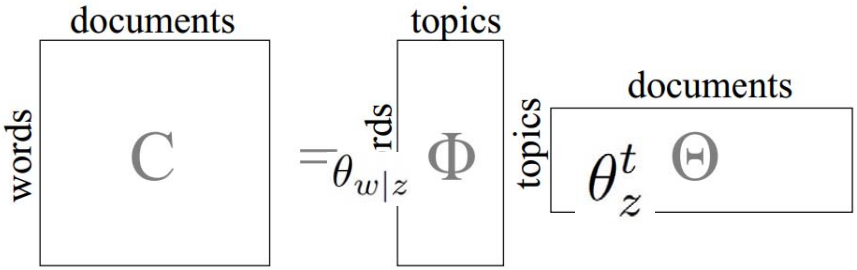
$$n(w,z) = \sum_{t=1}^{n} n_w(d^t) P(z|w,t)$$

how many times topic z is used with word w across documents

- M-step:

$$\theta_z^t = \frac{n(z|t)}{\sum_{z'=1}^{k} n(z'|t)}$$

$$\theta_{w|z} = \frac{n(w,z)}{\sum_{w'} n(w',z)}$$



Source: Landauer, Thomas K., Danielle S. McNamara, eds. Handbook of Latent Semantic Analysis. *Psychology Press*, 2013.

The number of topics

- If the EM algorithm succeeds finding a good solution in each case, the log-likelihood of the good solution should be higher than worse solutions

$$\text{BIC-score} = \text{log-likelihood of data} - \frac{\#\text{ of param}}{2}\log(N)$$

# of independent parameters in the topic model

of "data points"

Topic Models

- *Expression programs* - sets of co-expressed genes orchestrating normal or pathological processes

Text Analysis	Molecular Biology
Word	Expression for a particular gene
Document	Expression for all genes in a particular experiment (cell type)
Topic	Regulatory Program

Length of document = # of genes profiled

The higher the expression of a gene the more times it occurs in the documents

Examples of topics = immune response, stress response, development, apoptosis

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?

7.91J / 20.490J / 20.390J / 7.36J / 6.802 / 6.874 / HST.506 Foundations of Computational and Systems Biology
Spring 2014