University of Zurich[UZH]

Statistical Bioinformatics // Institute of Molecular Life Sciences

Intro to STA 426 course structure + Some Molecular Biology + Some GitHub/R/knitR/Bioconductor

Today's structure

9.00-9.30: Ice Breakers + Surveys

9.30-9.45: Tina Siegenthaler - computer overview

10.00-10.45: Course structure, evaluations, Introduction to Molecular Biology (Hubert)

11.00-11.45: Troubleshooting computing/logins; Introduction to R/ Bioconductor exercise

Survey 1: About the students

movo.ch

Token:

BE TU TE JY

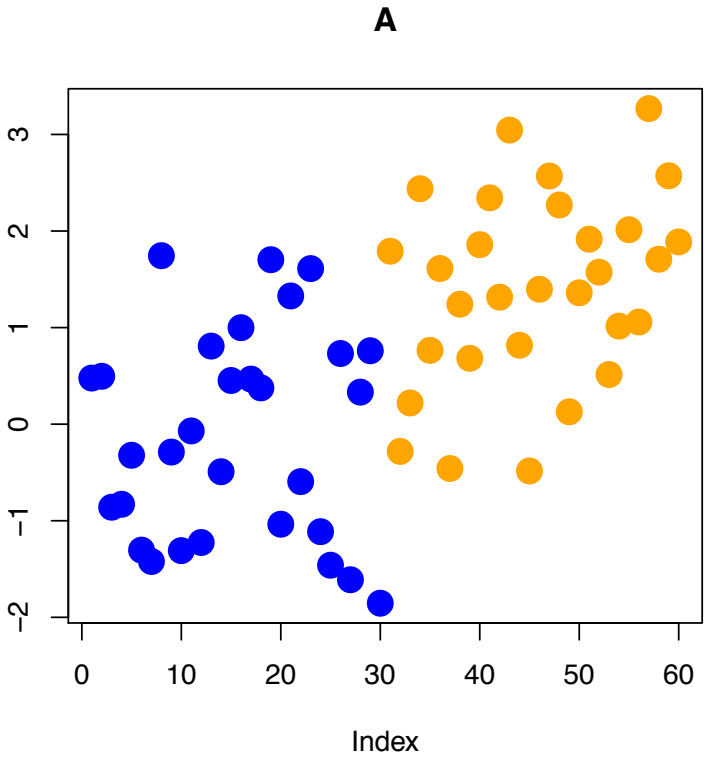
Survey 2: Statistical Insight

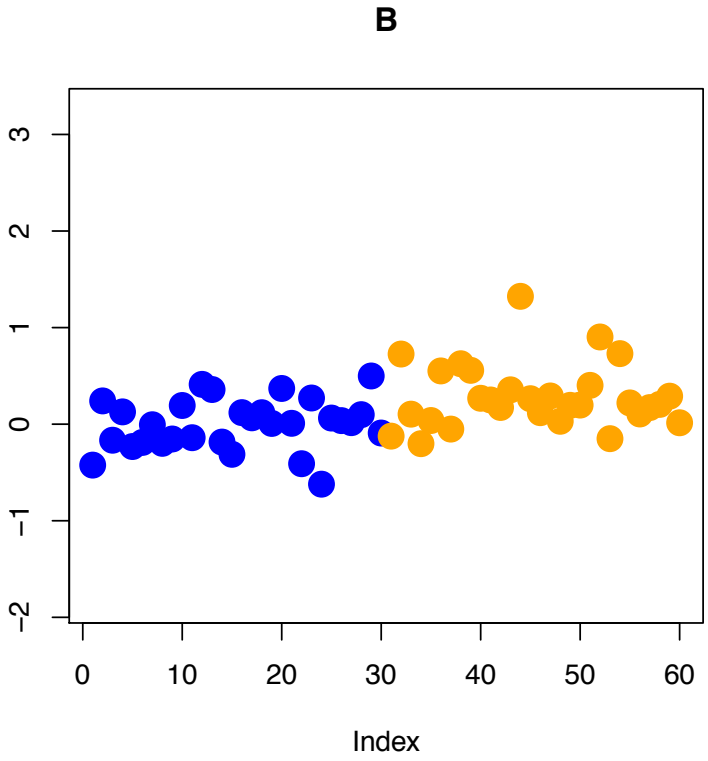
movo.ch

Token:

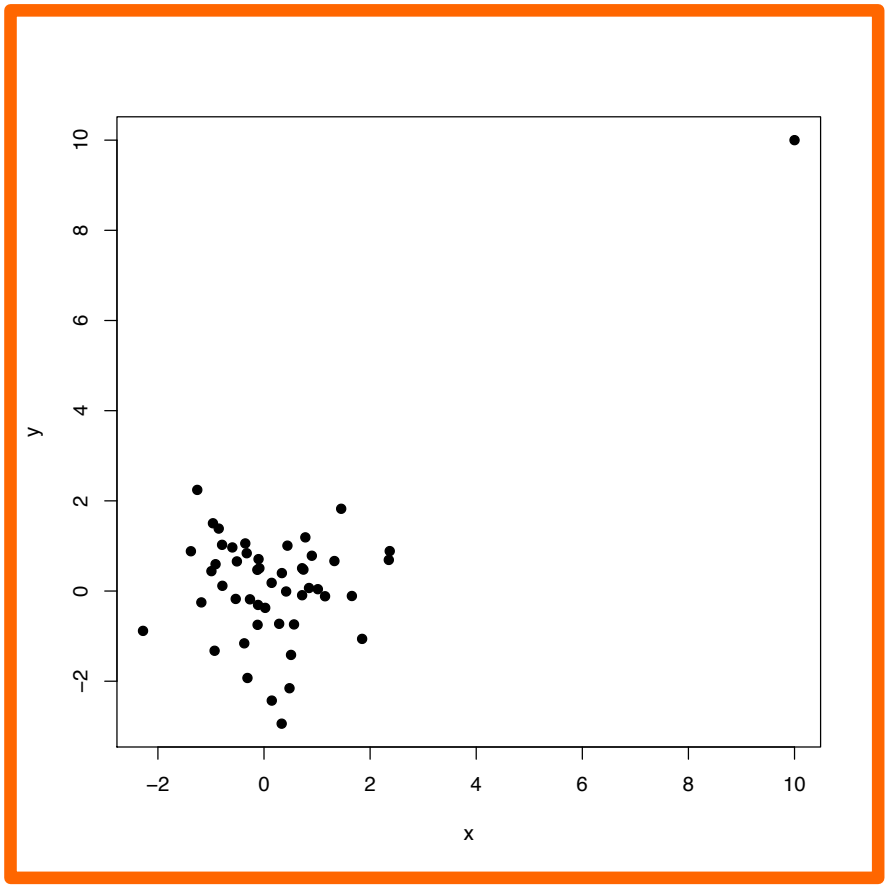
CU PU QO RI

Question 1: Which plot highlights more (statistical) evidence for a change in the population means (between orange and blue)?





Question 2: In your view, what best describes the associations shown in the plot of 'x' and 'y' ?



Question 4: Given this design matrix, describe the experimental design.

$$X = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

http://en.wikipedia.org/wiki/Statistical_hypothesis_testing

Question 6: Of these equations, which one resembles the standard two sample t-test ?

1

$$\frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

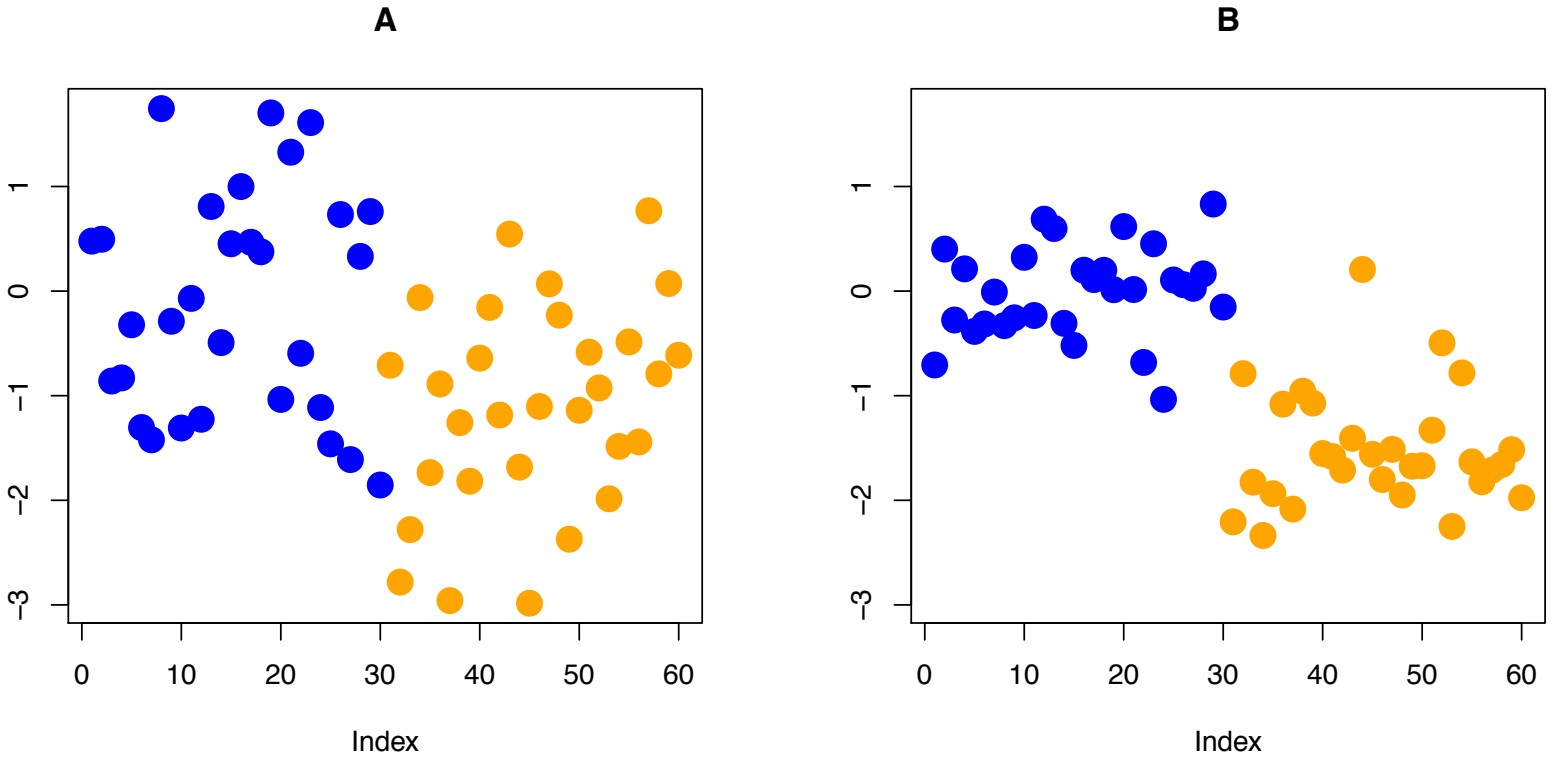
2

$$\sum^{k} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

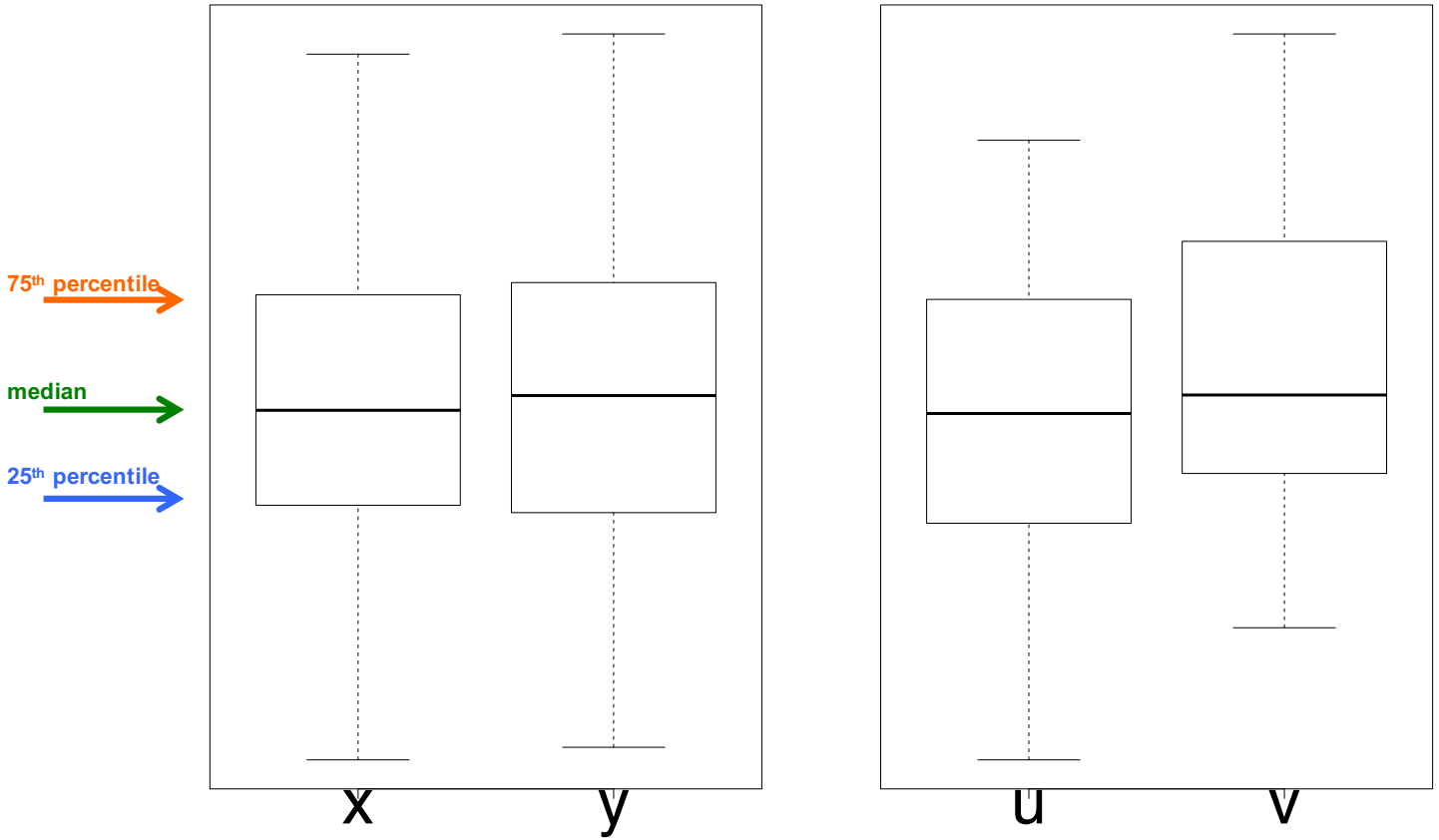
3

$$\frac{(\overline{x}_1 - \overline{x}_2) - d_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

Question 7: Which plot highlights more (statistical) evidence for a change in the population means (between orange and blue)?



Question 8: Given these boxplots, which of two underlying distributions are more similar?



Rough structure of lecture/exercise time

Monday mornings: we will run X.00-X.45; X in {9,10,11}

- 9.00-9.45: Y27-H-46
- 10.00-end: Y11-J-05
- Lecture/journal club presentation (9.00-whenever)
- Remaining time: in the computer lab (Y11-J-05) doing exercises/project

M.Sc. thesis projects

If you are:

- in a M.Sc. programme (ETHZ or UZH)
- have a solid background in mathematics / statistics / computation
- have an interest in research in this field (“statistical bioinformatics”)
- looking for a thesis project

→ Discuss a project in my lab

http://simplystatistics.org/2013/07/03/repost-the-5-most-critical-statistical-concepts/

Critical skills needed by statisticians (Jeffrey Leek's words):

With all the excitement going on around statistics, there is also increasing diversity. It is increasingly hard to define "statistician" since the definition ranges from very mathematical to very applied. An obvious question is: what are the most critical skills needed by statisticians?

So just for fun, I made up my list of the top 5 most critical skills for a statistician by my own definition. They are by necessity very general (I only gave myself 5).

1. **The ability to manipulate/organize/work with data on computers** - whether it is with excel, R, SAS, or Stata, to be a statistician you have to be able to work with data.
2. **A knowledge of exploratory data analysis** - how to make plots, how to discover patterns with visualizations, how to explore assumptions
3. **Scientific/contextual knowledge** - at least enough to be able to abstract and formulate problems. This is what separates statisticians from mathematicians.
4. **Skills to distinguish true from false patterns** - whether with p-values, posterior probabilities, meaningful summary statistics, cross-validation or any other means.
5. **The ability to communicate results to people without math skills** - a key component of being a statistician is knowing how to explain math/plots/analyses.

Learning outcomes (in my words)

- Understand the fundamental “scientific process” in the field of Statistical Bioinformatics
- Be equipped with the skills / tools to preprocess genomic data (Unix, Bioconductor, mapping, etc.) and ensure reproducible research (R / markdown)
- Have a general knowledge of (some) **types** of data and **biological applications** encountered with high throughput genomic data
- Have the general knowledge of the range of statistical methods that get used with microarray and sequencing data
- Gain the ability to apply statistical methods / knowledge / software to a collaborative biological project
- Gain the ability to critical assess the statistical bioinformatics literature
- Write a coherent summary of a bioinformatics problem and it’s solution in statistical terms

Course evaluation

1. Journal club presentation	20%
2. Project	50%
3. Exercises	30%
4. Technology day (participation)	0% or -10%

The semester-long course structure (subject to change)

Schedule

Date	Lecturer	Topic	JC
18.09.2017	Mark	admin, mol. biology basics, R markdown	
25.09.2017	Hubert	exploratory data analysis	
02.10.2017	Mark + Hubert	interactive technology session	
09.10.2017	Hubert	NGS intro; mapping	
16.10.2017	Mark	limma 1	
23.10.2017	Mark	limma 2	
30.10.2017	Hubert	RNA-seq quantification	
06.11.2017	Mark	edgeR+friends 1	
13.11.2017	Charlotte	hands-on session #1: RNA-seq	X
20.11.2017	Mark	edgeR+friends 2	
27.11.2017	Hubert	classification	
04.12.2017	Mark	single-cell	
11.12.2017	Gosia	hands-on session #2: mass cytometry	X
18.12.2017	Mark	epigenomics, DNA methylation, ChIP data, gene set analysis	

Expectations: **journal club** presentation

- 20-25 minutes (+5 minutes discussion)
- MUST:
  - ➡ be a paper about a **statistical** method in genomics
  - ➡ be approved by Mark/Hubert
- Should:
  - ➡ describe the biological context
  - ➡ describe the (new) model used
  - ➡ describe comparisons to existing methods
- Should not:
  - ➡ be one of the papers discussed in detail in lectures: limma, edgeR, DEXSeq, etc.
- (new for 2017) Expectations of observers: fill out feedback form

Expectations: **project**

- ~10-15 page report, with R code in line (e.g. **knitR**)
- Describe the biological setting, statistical analysis, exploratory analysis with publication-quality graphics embedded
- Three possibilities:
  - Comparison of statistical methods (simulation / independent reference data + metrics)
  - Reproduce an analysis from a paper from the raw data
  - Real collaborative project with FGCZ or a local laboratory
- Be strategic: work on something related to your interests!

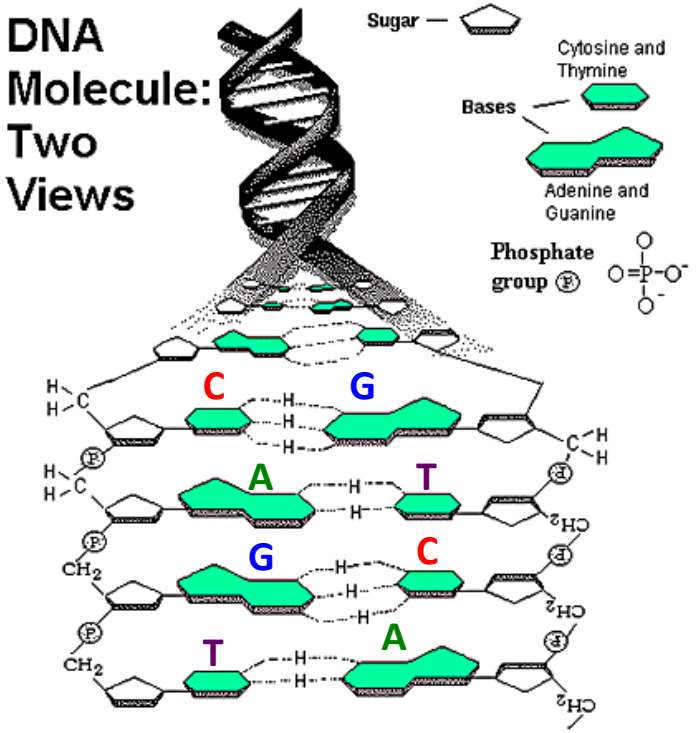
Soft technical skills needed (developed) in this course …

- **Data Science!**
- Use unix-like operating system to run command-line programs
- Options:
  - use your own Linux/MacOSX computer; N.B.: you may be able to do everything from Windows (e.g., cygwin)
  - use the Macs in Y11-J-05
- R: from the command line or R studio; getting help; creating workflows; how to make publication-quality graphics; knitr/Rmarkdown
- Bioconductor – www.bioconductor.org

Some molecular biology concepts (slides from Hubert)

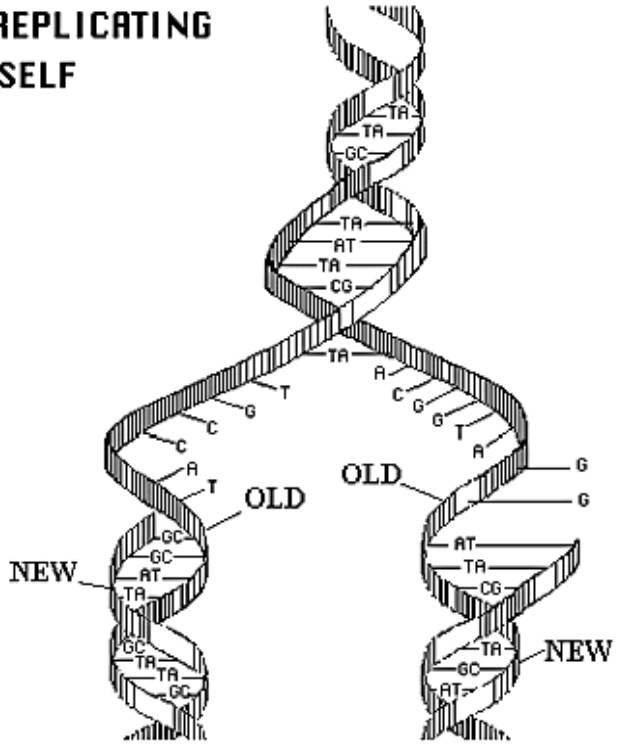
What is DNA?

- A long backbone of sugars with nucleotides attached
  - **Adenine (A)**
  - **Guanine (G)**
  - **Cytosine (C)**
  - **Thymine (T)**
- It can form a self-complementary **double helix**
- In living organisms, the DNA is the carrier of the hereditary information, it is the source code of life



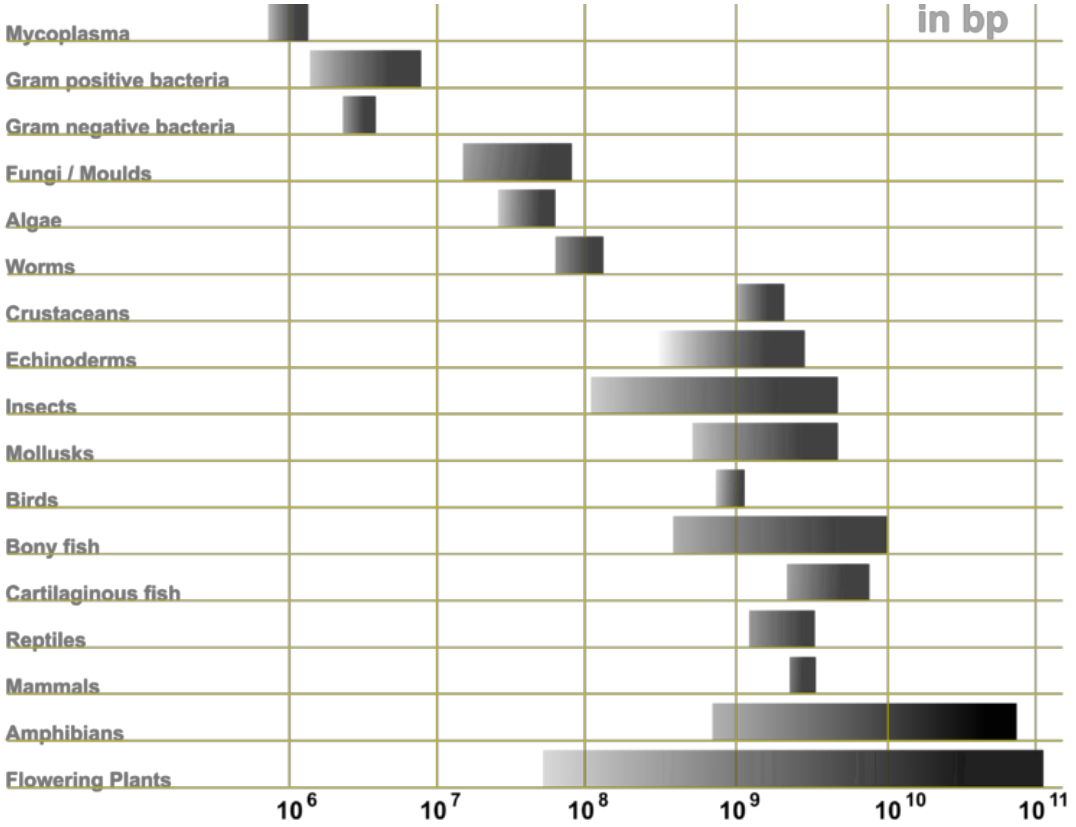
DNA replication

– The helix becomes unzipped and each strand acts as a template for a new complementary strand of DNA



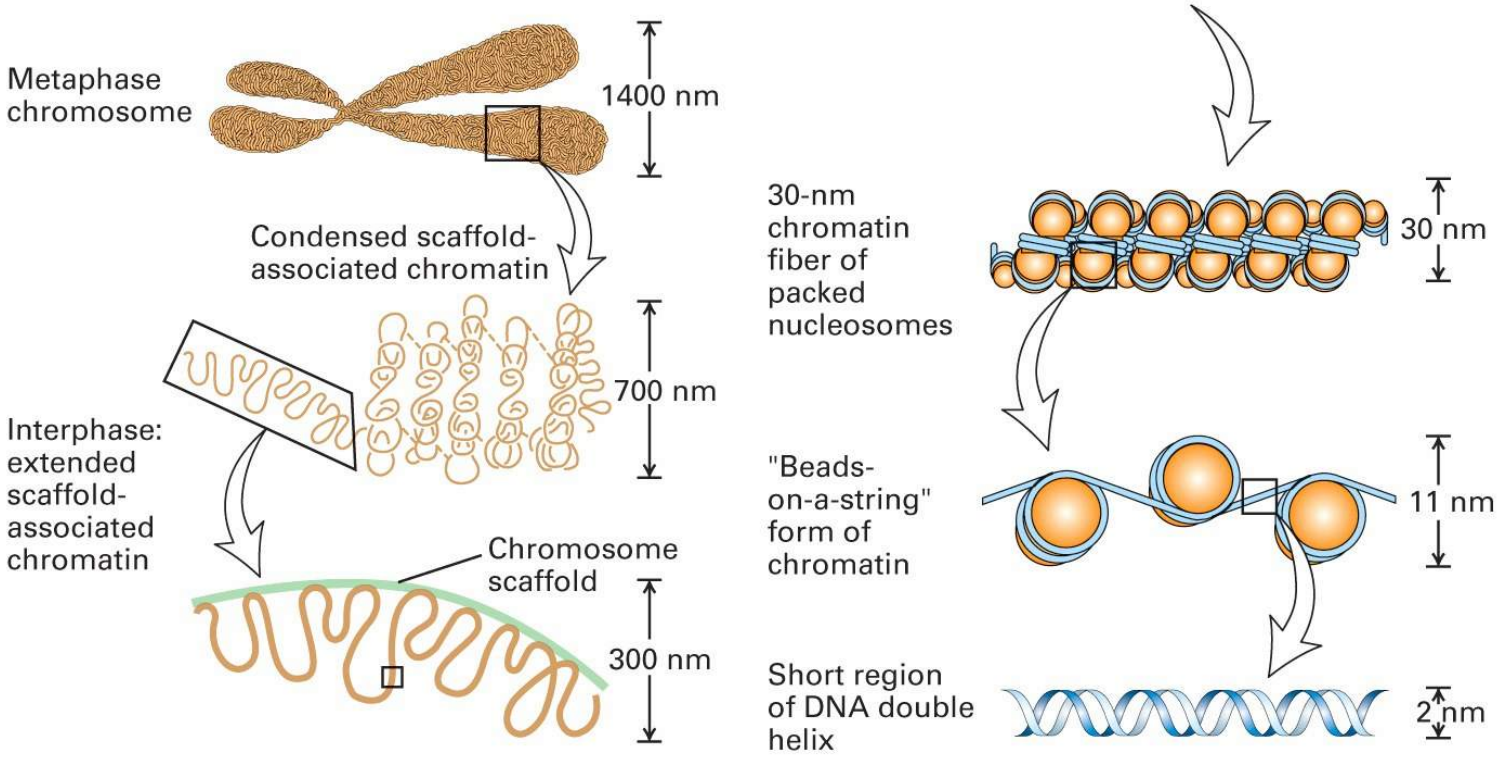
Genome Sizes

The size of the human genome is 3.2 billion base pairs. The length of this DNA string is approx. 2m.



http://en.wikipedia.org/wiki/Genome_size

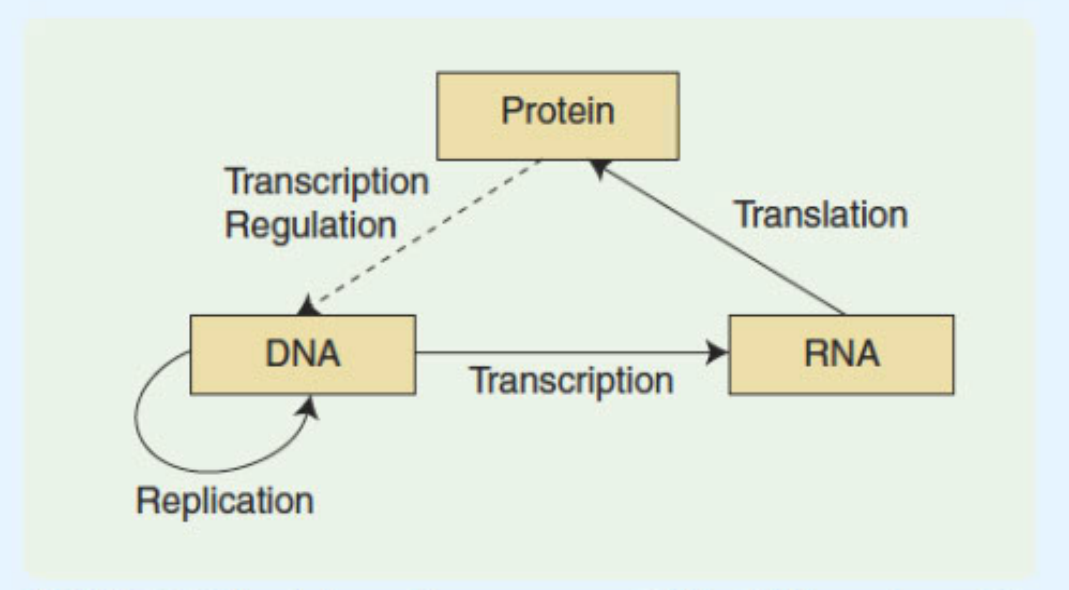
DNA Superstructure



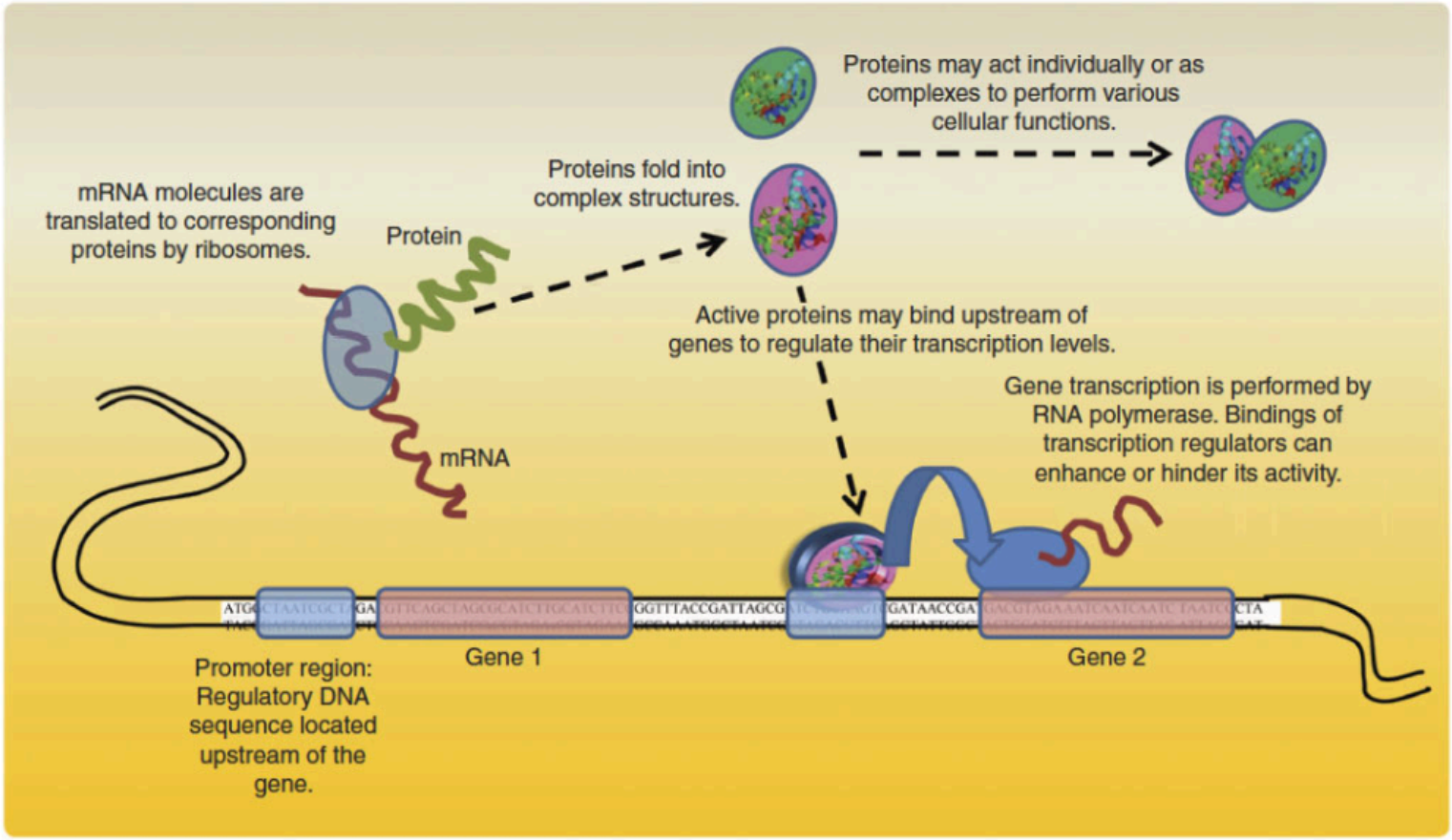
Lodish et al. *Molecular Biology of the Cell* (5th ed.). W.H. Freeman & Co., 2003.

Genes

- A gene is a region of DNA that controls a hereditary characteristic
- Usually a gene is transcribed into a messenger RNA which is then translated into a protein.
- In humans genes constitute only ~3% of the human genome



The Central Dogma



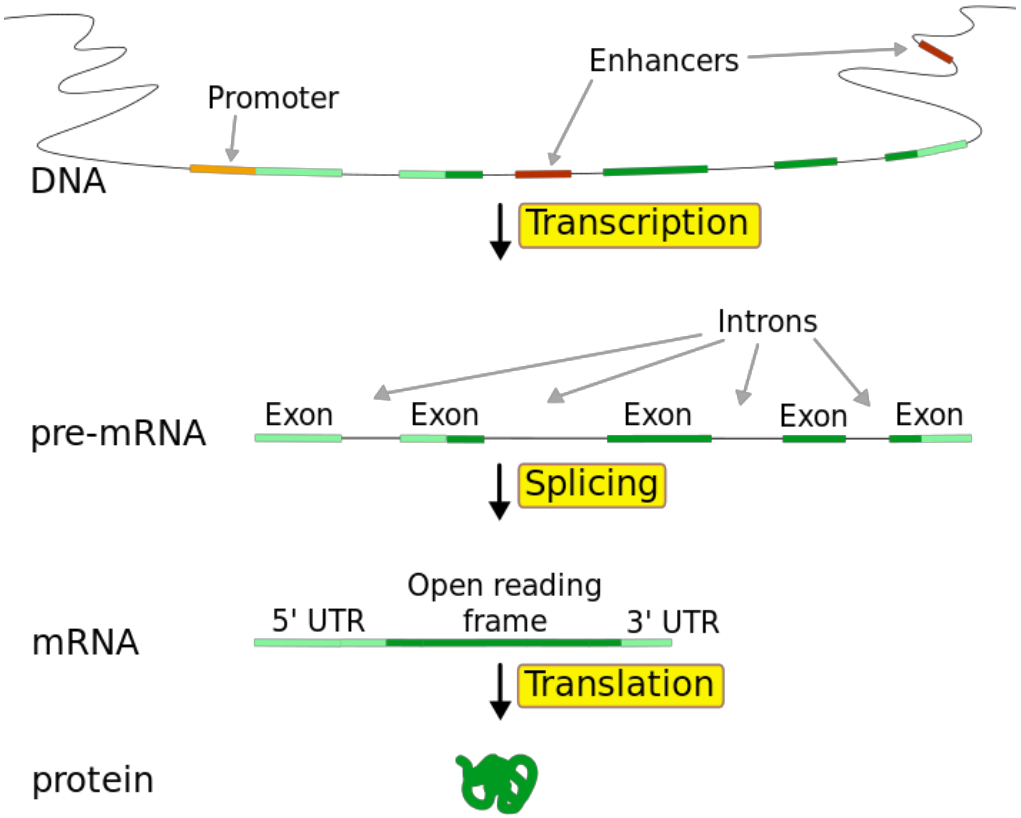
Elvitigala et al. 2010

Transcription

The transcription process generates a messenger RNA molecule from a gene region.
RNA is like DNA but

- the sugar-phosphate is different: ribose instead of deoxyribose
- In all places where the DNA has a T the the RNA has a U (uracil)

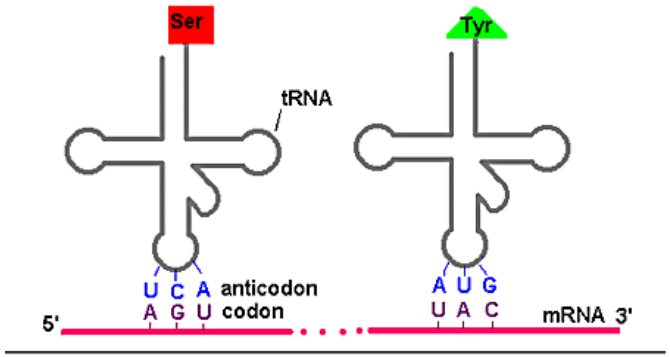
In higher organisms the protein coding sequences (exons) are interspersed by non-coding sequences (introns) which are spiced out.



Translation: The Genetic Code

The translation process generates a protein based on the information in the messenger RNA

- A protein is a linear polymer of amino acids linked together by peptide bonds.
- Proteins are the main functional chemicals in the cell, carrying out many functions, for example catalysis of the reactions involved in metabolism.
- Proteins have a complex spatial structure

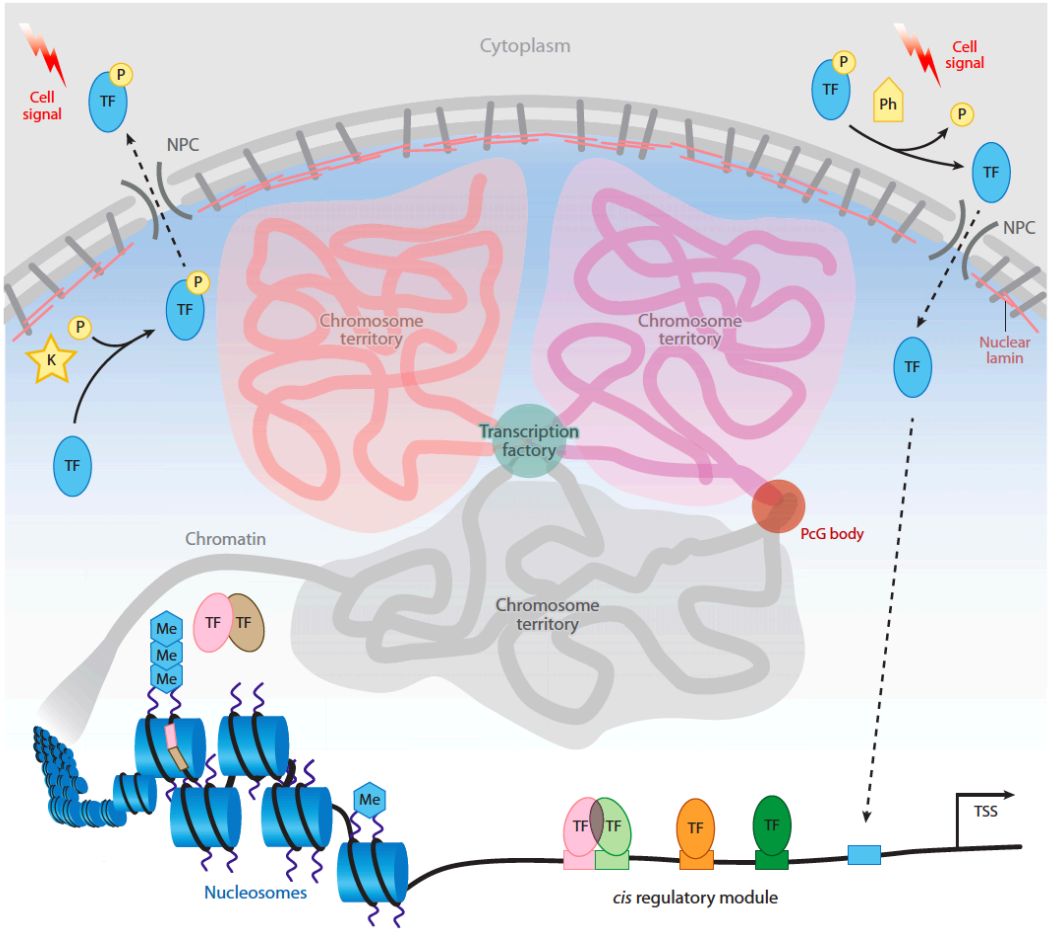


2nd base in codon

1st base in codon	U	C	A	G	3rd base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

The Genetic Code

Transcriptional Regulation



http://www.annualreviews.org/doi/pdf/10.1146/annurev-genet-110711-155437

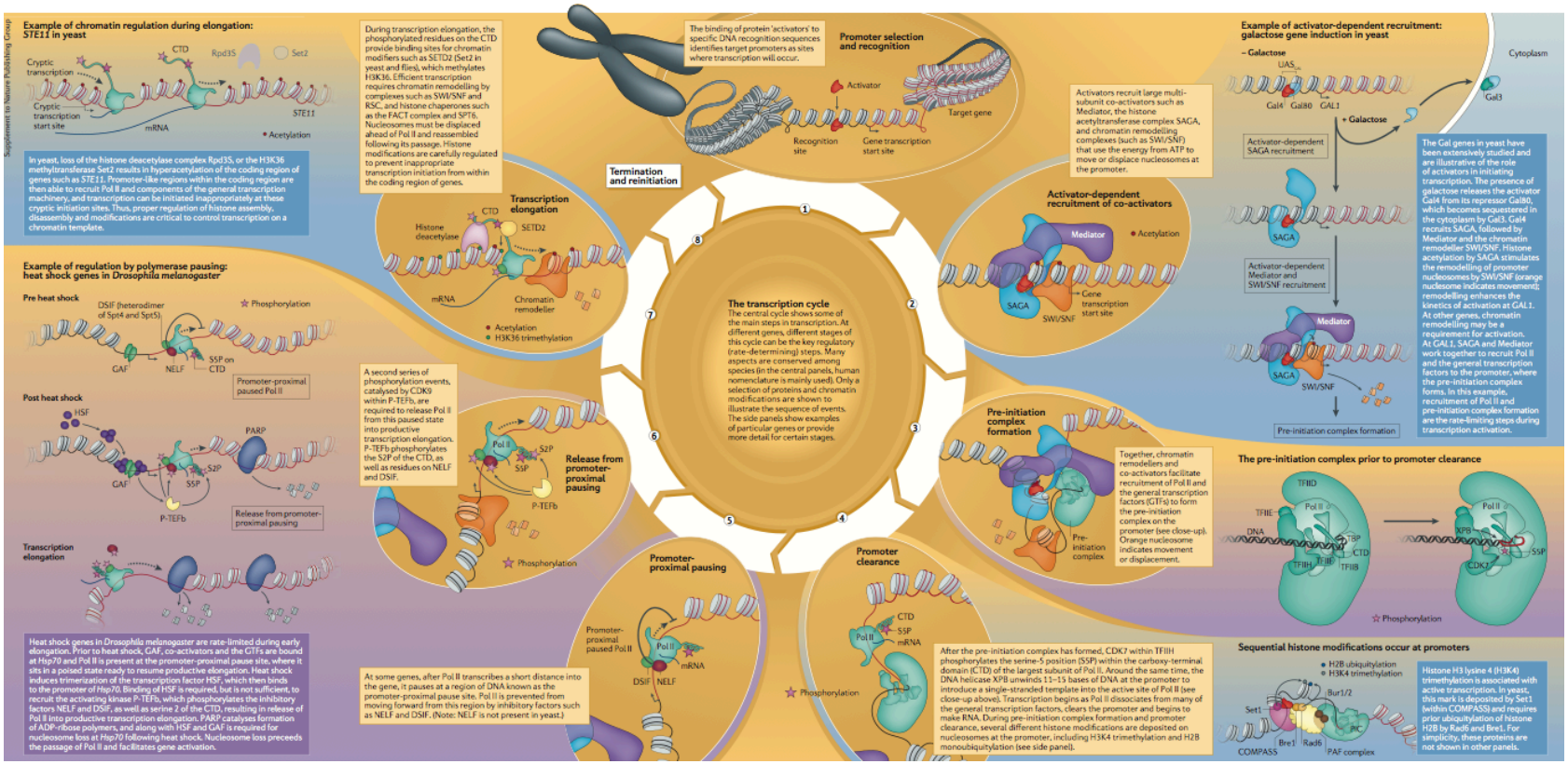
nature REVIEWS GENETICS

Chromatin remodelling and the transcription cycle

Vikki M. Weake and Jerry L. Workman

Transcription by RNA polymerase II (Pol II) occurs in the context of chromatin within a eukaryotic cell. Chromatin is generally inhibitory to transcription, so a variety of mechanisms are required to activate transcription from a nucleosomal template. One of the first steps is that large co-activator complexes interact with small activator proteins to identify gene promoters that are ready to be transcribed. Nucleosome remodelling complexes that use energy from ATP to move or displace nucleosomes from DNA facilitate the recruitment and assembly of these complexes on the promoter and enable rapid gene activation. Even during transcription elongation, nucleosomes must be removed for efficient passage of the polymerase. Furthermore, these same nucleosomes must be reassembled rapidly and modified appropriately following passage of the polymerase to prevent inappropriate initiation of transcription from promoter-like elements within the coding region.

abcam



GitHub + knitR exercise

Main repo for course: https://github.com/sta426hs2017/material

All homework submissions occur via github

Homework for today (part 1):

1. Acquaint yourself with the idea of github [1]
2. Create a github account at github.com
3. Make sure you know how to check in / out files from the command line or from an app [2]
4. Create a new public repository, add a README.md (learn a bit of markdown [3]) and add some content
   - Include an image
   - Include a web link
   - add an issue to the materials repo to let me know that you've done it
   - (you can delete the repo after I've closed the issue, if you want)

[1] https://gist.github.com/andrewpmiller/9668225
[2] https://confluence.atlassian.com/stash/basic-git-commands-278071958.html
[3] http://markdowntutorial.com/

Rmarkdown / knitR for executable documents / reproducibility

Homework for today (part 2):

1. Acquaint yourself with knitR PDF/HTML Rmarkdown documents [1], perhaps both in R studio and from command prompt
2. Create an HTML/PDF document that samples 100 values from a log-normal distribution (say, mu=1, sigma=.25); create a histogram of the distribution and the distribution on the log scale; report the mean and variance of the sample in line in the text.
   - Do not just dump the R code and plots in the HTML/PDF document; add some text and headings and make it into a readable story (i.e., the document should be self-explanatory)