Genomics Era Sequencing Technologies and Analysis

Spring 2017

People you need to know

- Instructor – Me

- Grader/TA - Haseeb

New(ish) Course

- Trying new things (lab)
- You are the learning experts
  - May ask for feedback frequently
- Exciting field others want to teach too
  - Writing a paper on our experiences
  - Some things you'll be asked to do will be to help others teach this too

What is Genomics?

- Articles you were asked to bring:
  - What is is about?
  - What do you find interesting about it?
  - How is it related to genomics?

- What are genomes, what's in them, and how are they structured
- How DNA sequences are measured
- How to identify differences between individuals' genomes
- The strengths and limitations of different technologies
- How to assemble genomes from the DNA sequences we can generate
- Identify what a particular genome sequence does for its organism
- How to correlate differences in genotypes to differences in phenotype
- How to correlate differences in gene expression to phenotypes
- How to determine all the genomes in an environment

Boring Class Info

- M/W – here Parrington Hall 106
- F labs – MGH 044 (Mac Lab)

Key Syllabus Points

- What percentage of your final grade is determined by the different categories of evaluation?
- Do you have to buy the textbook?
- Where will you be turning in assignments?
- How will your midterm be taken?
- What are the options for the final project?
- What is the late policy for quizzes? Why is it set up that way?
- What is the late policy for lab reports?
- If you, or a friend, needs accommodations to succeed, what should you do?
- What are the rules around collaborating?
- Which learning objective are you most interested in?

Notice: The University has a license agreement with VeriCite, an educational tool that helps prevent or identify plagiarism from Internet resources. Your instructor may use the service in this class by requiring that assignments are submitted electronically to be checked by VeriCite. The VeriCite Report will indicate the amount of original text in your work and whether all material that you quoted, paraphrased, summarized, or used from another source is appropriately referenced.

Academic Integrity

- As future engineers and scientists, you are held to a high standard of honesty and integrity.
- All instances of academic misconduct will be reported and may result in removal from and failure of this course.
- Plagiarism is unacceptable. If you have difficulty rephrasing information you have read, see me for advice and referrals to writing assistance

Technology in the classroom

- MANY studies indicate it is bad for learning to have it in the classroom unless you are specifically using it
- We are all distracted by screens – so anyone behind you is also being distracted by your computer/phone/text conversation/Pokemon

Please avoid in class unless you are asked to do something with it

- If you have a reason to need it, be reasonable and respectful

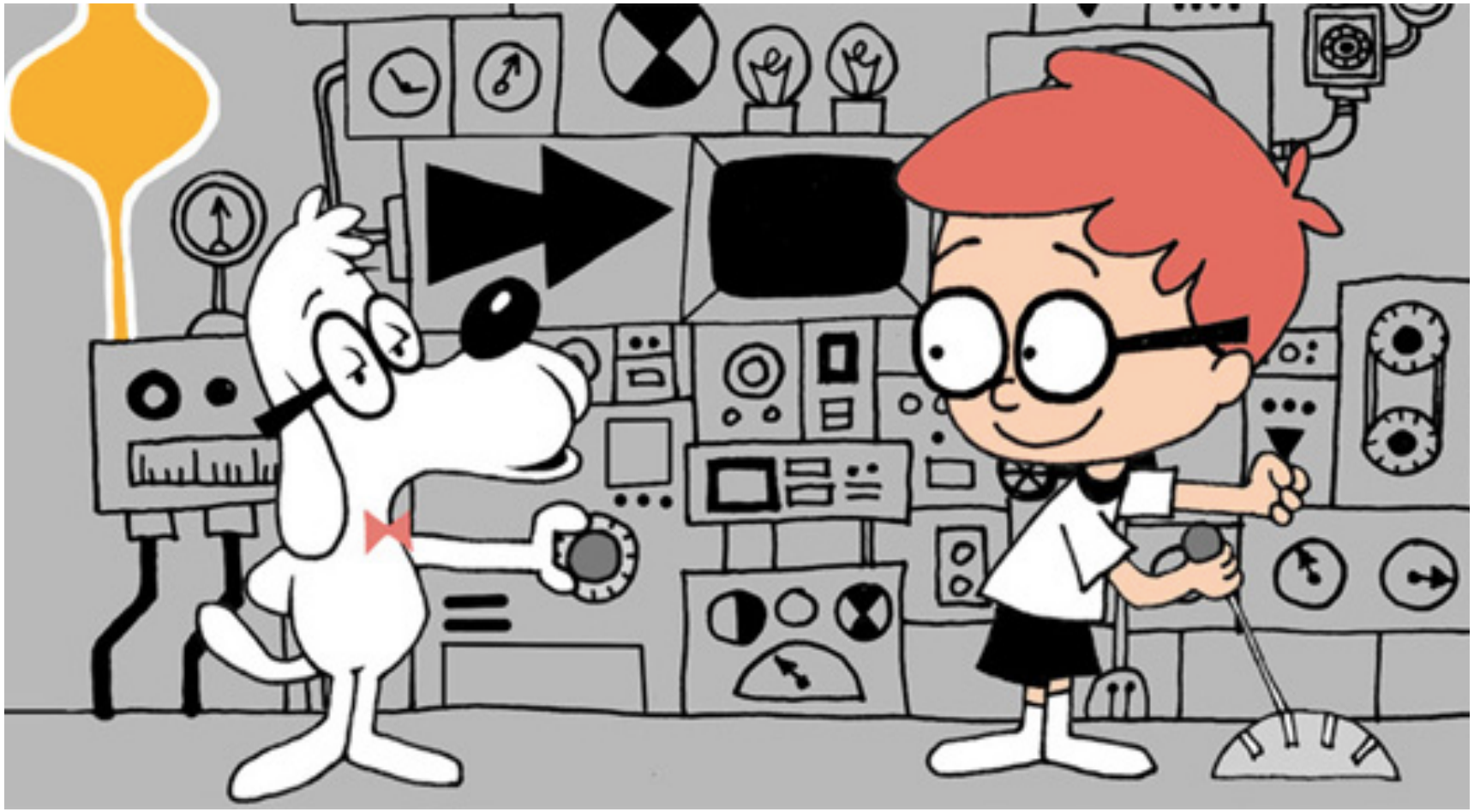
Week	Topic
1	Genes, chromosomes, genetics and jargon
2	Libraries and how we make them
3	Illumina, Ion Torrent, and NanoString (guest speaker)
4	Pacific Biosciences and Oxford Nanopore Introduction to Genome Assembly
5	NorthShore Bio Nanopores (guest speaker) and imagining better sequencers
	Online Midterm
6	Gene expression and RNA-Seq
7	Single cell sequencing
8	Population genetics and ancestry analysis
9	Additional uses and research
10	Grad student presentations

Lab Date	Topic
Mar 31	Genome browsers, genome references, online resources
Apr 7	R and sequence alignment
Apr 14	Assessing Quality of Reads
Apr 21	Genome assembly 1
Apr 28	Genome assembly 2
May 5	RNA Seq alignment
May 12	RNA Seq quantification
May 19	In class time to work on final projects: Go to Rushmer Lecture
May 26	Hap Map/Ancestry analysis
June 2	Metagenomics

Questions?

Let’s get started with the fun stuff!

The Wayback Machine



The Human Genome Project

- Early ‘80s advances in molecular biology led to ability to sequence large sections of DNA
- Mid-80’s DOE was trying to better understand potential risks of radiation exposure
- 1988 Congress approved $ for DOE and NIH to sequence the entire human genome
- Anticipated it taking a few decades

Human Genome Project

- First started in 1990
- First five or so years was mostly technology development
  - How to get large sections of DNA
  - How to sequence large sections quickly and efficiently
  - How to assemble genomes
  - How to visualize genomes

Human Genome Project

- NHGRI under Francis Collins was sequencing, slowly, expensively, and with lots of caution
- Craig Venter decided he could do better – and make money doing it – founded Celera in 1998
- Wellcome Trust jumped in to increase funding to Sanger Centre in Britain

Pace of sequencing increased for everyone

Completed! (sort of)

- 2000 Bill Clinton, Tony Blair, Francis Collins and Craig Venter jointly announced the release of the Human Genome!

Not really completed.

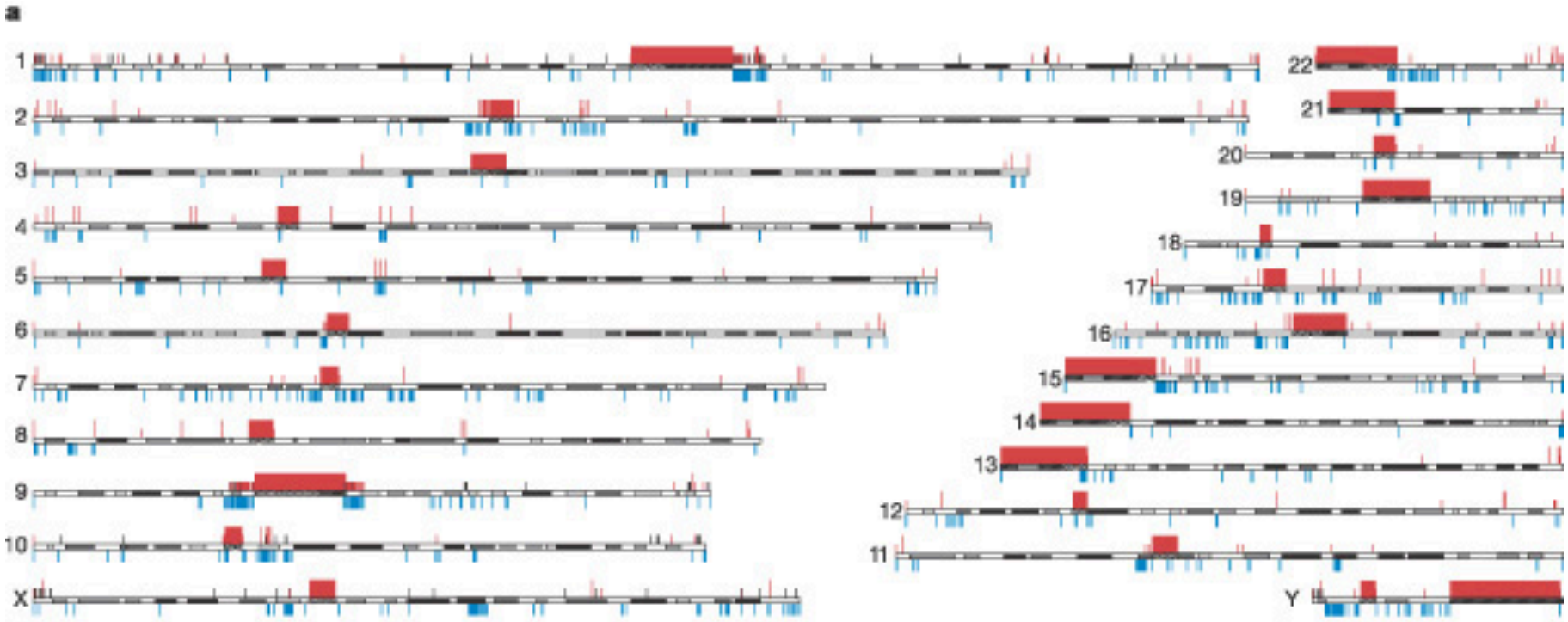
90% completed. Lots of hard to sequence sections still unsequenced.

- In 2004 final gold standard reference released
- Still has 300+ gaps

Regions missing

- Centromeres
  - Central region of chromosome important for segregation during mitosis
- Large repeat regions, like rRNA regions
  - Large repeats are hard to determine exact number of
  - Seem to be different between people

Large duplications in the genome that result in repeats that are hard to sequence well



Ethics and Genomes

- During this process NIH was very concerned about ethical uses of the data
  - 5% of the HGP funds went to fund ethical research in this area
- We will be discussing ethical dilemmas and concerns along the way

Whose genome is it exactly?

- Unknown precisely
- HGP researchers took blood from many people
- DNA was extracted from white blood cells of 2 males and 2 females
- However, post-hoc, most appears to be from a single anonymous male donor from Buffalo, NY

What did the HGP find?

What did the HGP find?

- ~30,000 genes initially; since revised to ~20,000
  - 4000 genes in *E. coli*
  - ~17,000 in fruit flies
  - 21,733 in *C. elegans*
- Sites of duplication
  - Many medical conditions arising from duplications
  - ~5.3% of human genome is in a segmental duplication
  - Y chromosome is >25% duplications
- Birth of genes
  - Duplication of genes that seem to have evolved to be different
- Death of genes
  - Recent mutations that seem to make the gene inactive/ unprocessable

How did this change medicine/ research?

How did this change medicine/ research?

- Now compare each individual to a reference genome
- Compare to other organisms
- Identify genes in a region involved in disease/ condition
- Predict function of genes from sequence
- Understand evolution of humans from evidence in genome

What was still to do?

- All the cool stuff!

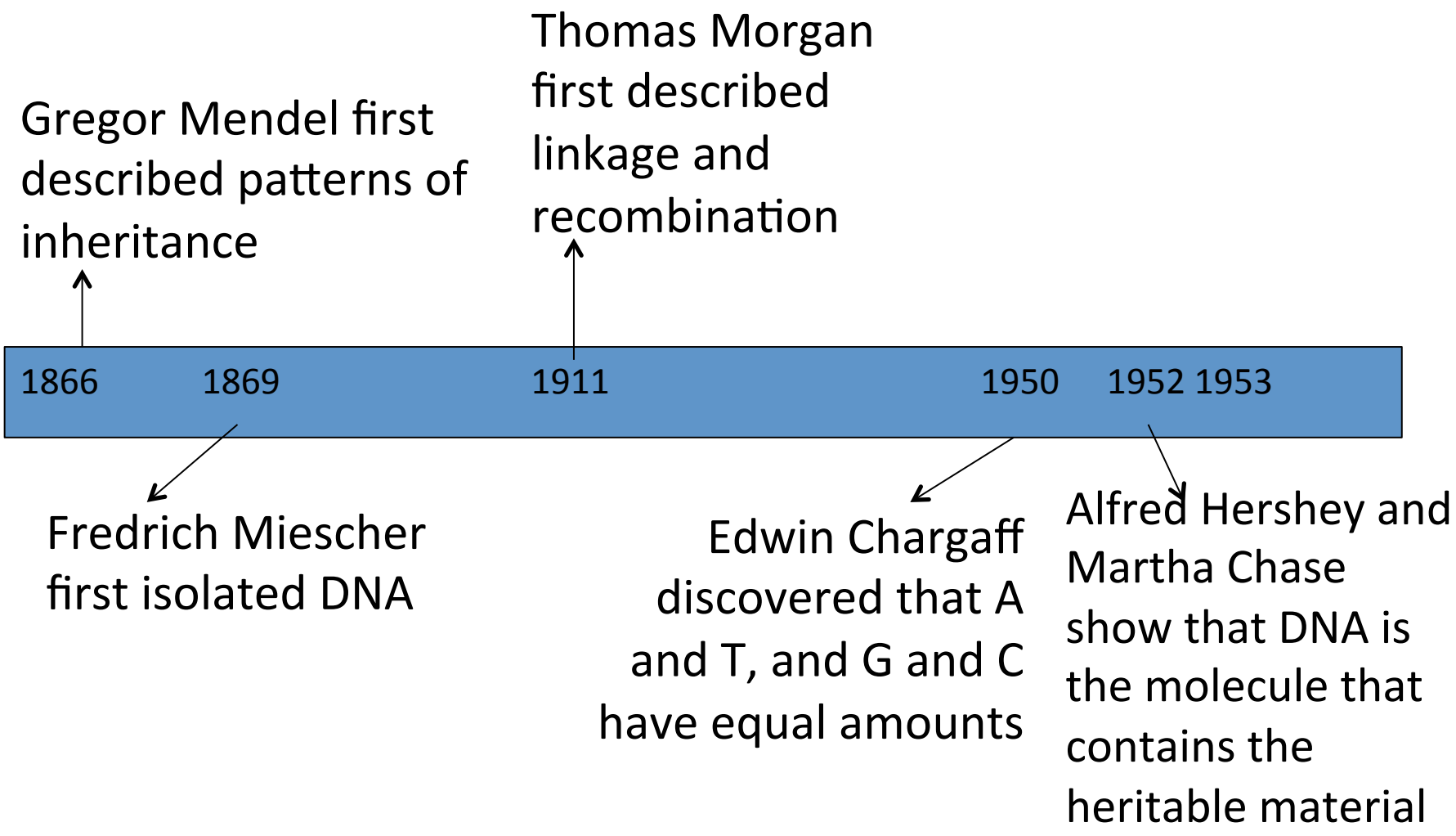
What was still to do?

- Fill in the gaps
- How do we sequence an individual genome instead of a collection?
- How are individuals different from the reference?
- How are populations different from each other based on ancestry?
- How do we relate small genetic changes to phenotypic changes?

HGP project

Genomes & Genetics Review

History of DNA



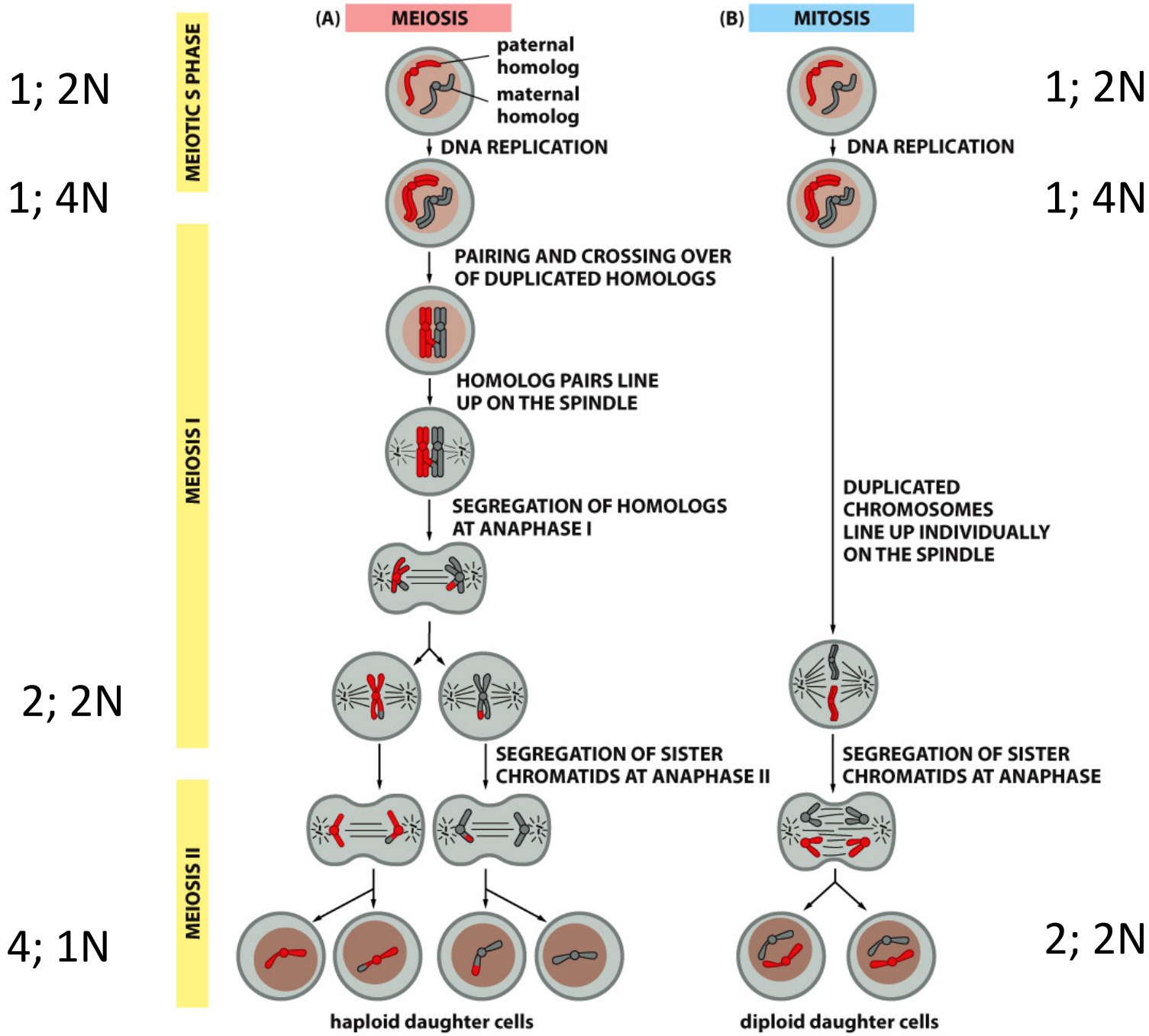
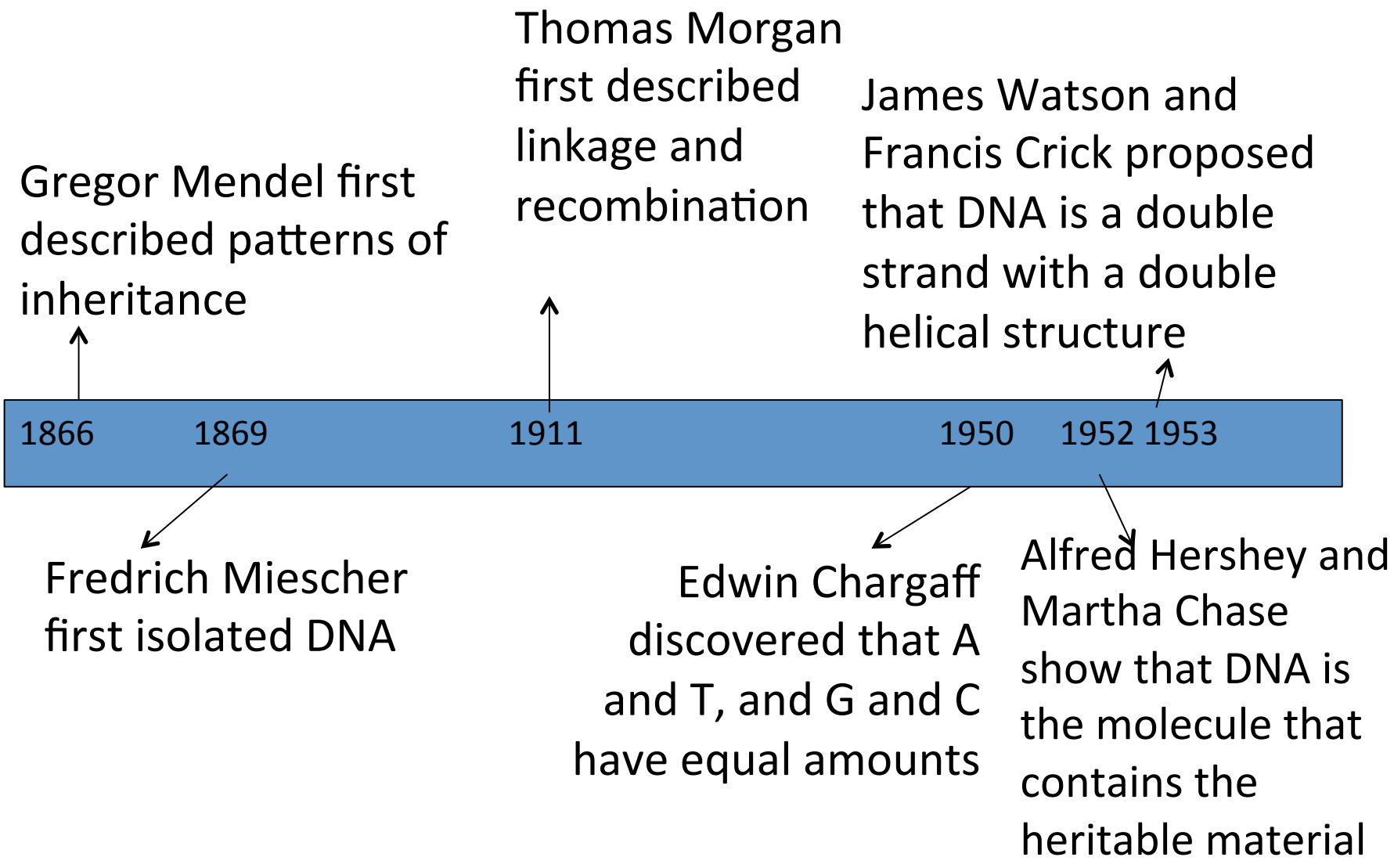
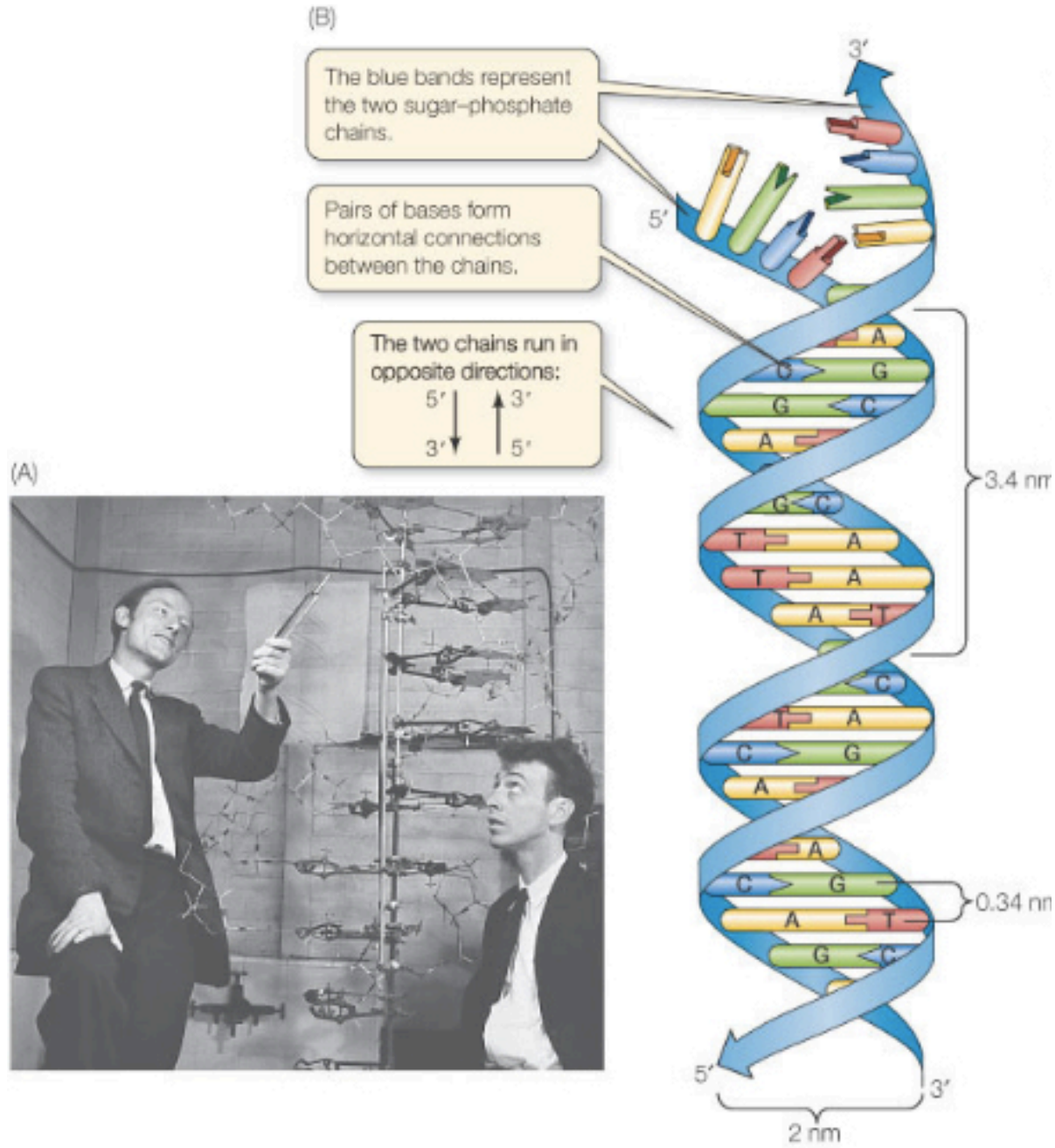


Figure 17-53 Molecular Biology of the Cell 6e (© Garland Science 2015)

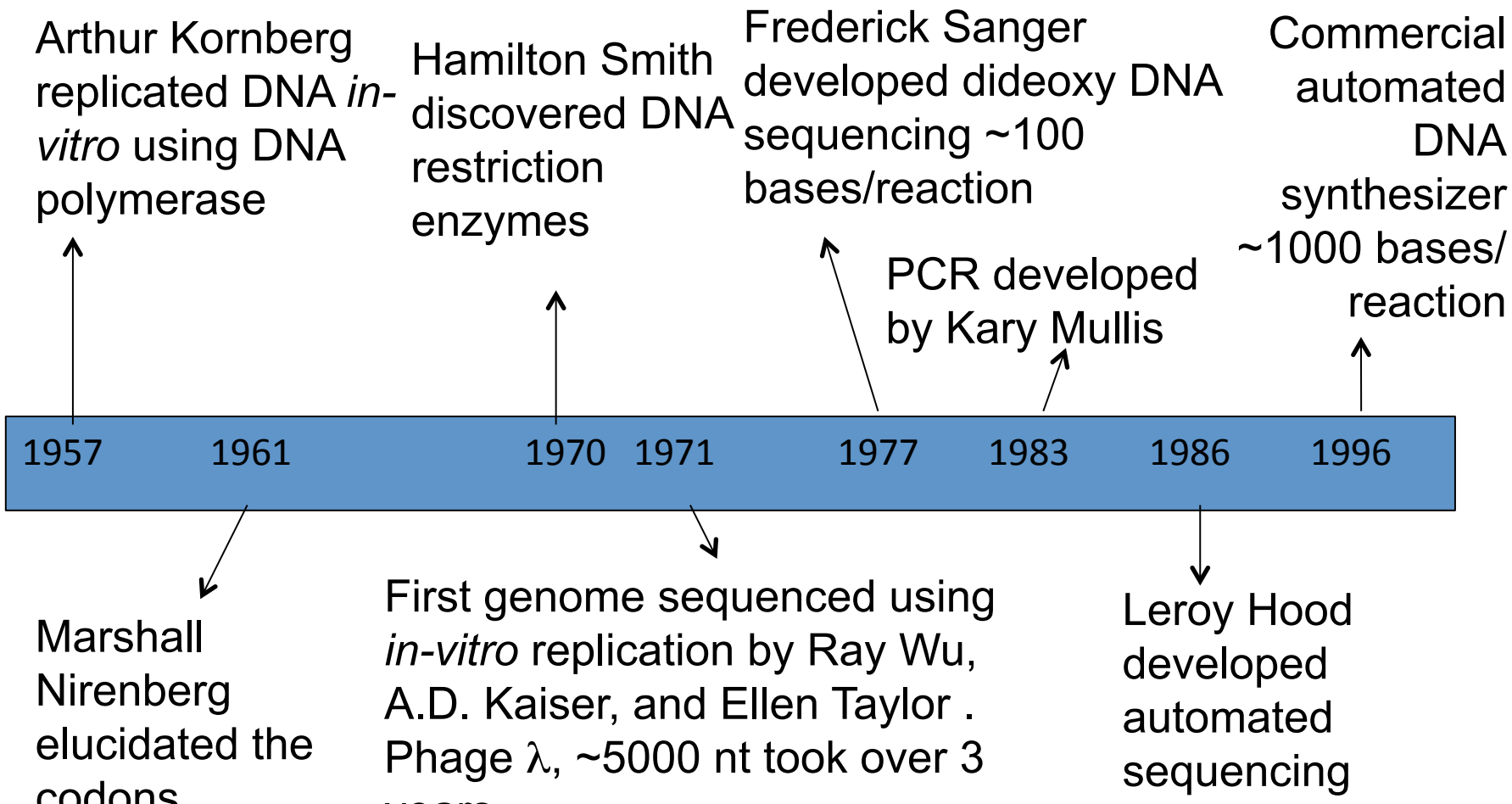
History of DNA





http://www.nature.com/scitable/content/dna-is-a-double-helix-24263

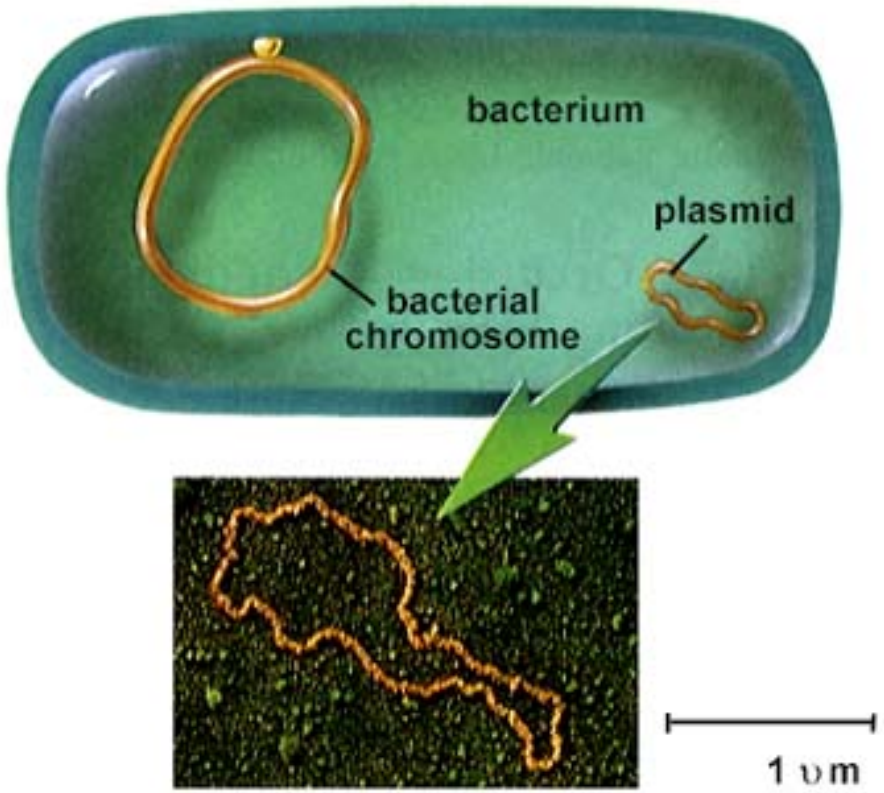
History of DNA



Bacteria have two types of DNA

Genomic:
Chromosome

Extragenomic:
Plasmids



http://www.aliciapatterson.org/APF2202/Sachs/Sachs02.jpg

A drawing in the sand of a replicon as presented in Jacob *et al*. (1963). The main circle represents the replicon, the square box represents the replicator, and the arrow indicates where the initiator protein is encoded, which, when synthesized, binds to the replicator. (Figure provided by M. Méchali.)

Circular chromosomes require 1 structural element: Origin of replication

Some also have a structure that allows for segregation of chromosomes during mitosis (centromere-like)

Human Chromosomes

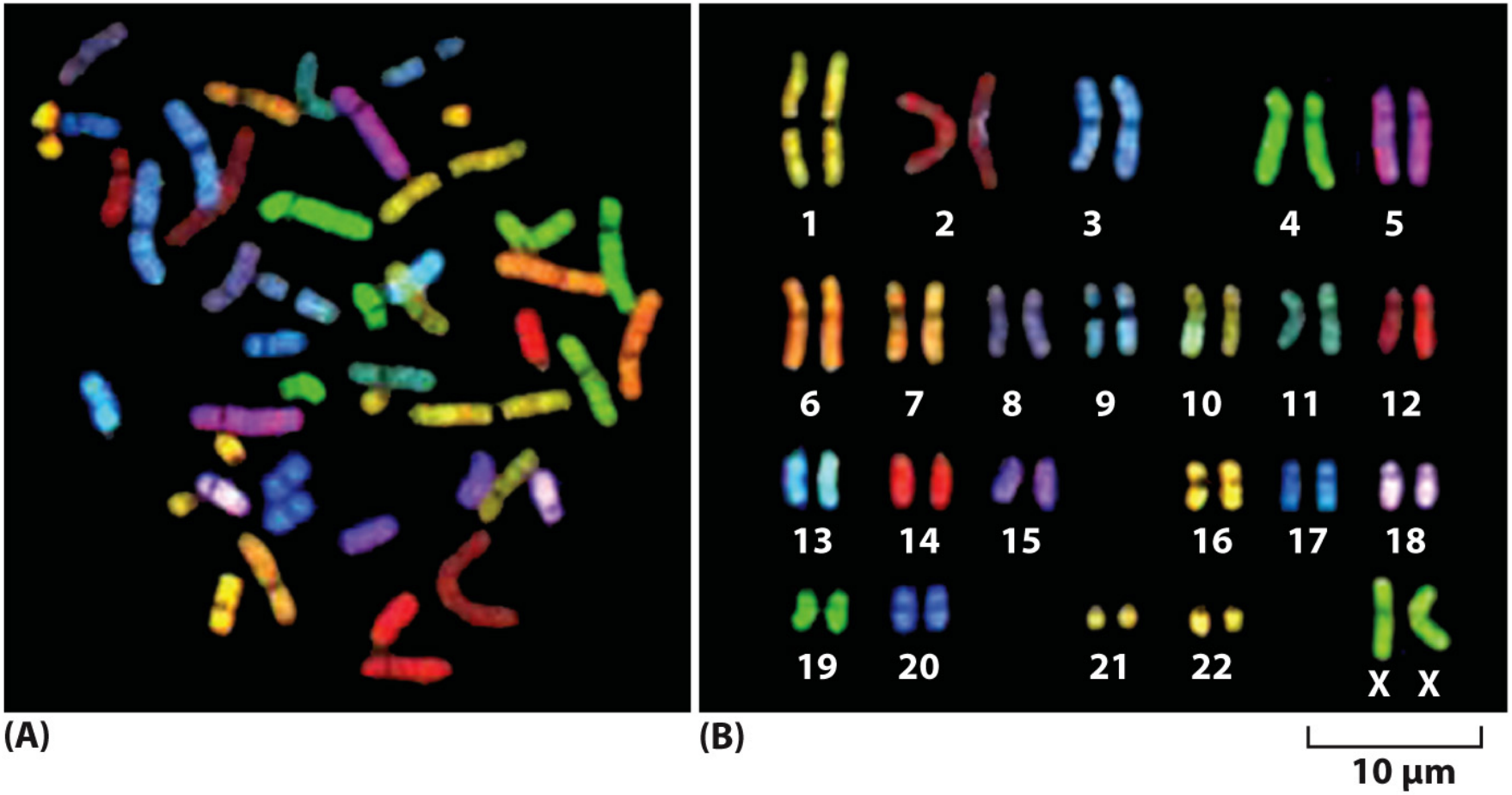
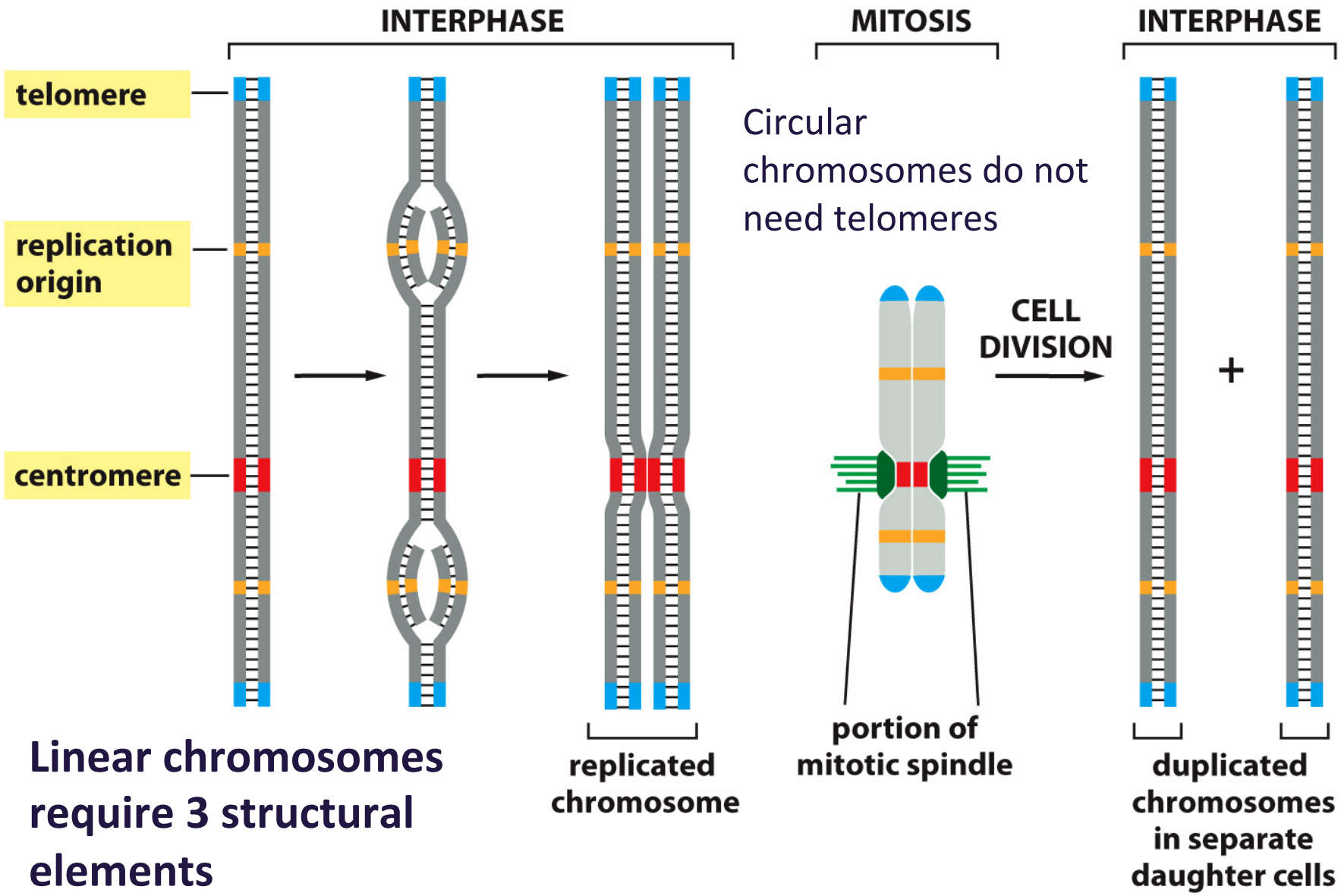


Figure 4-10 Molecular Biology of the Cell 6e (© Garland Science 2015)



Linear chromosomes require 3 structural elements

Figure 4-19 Molecular Biology of the Cell 6e (© Garland Science 2015)

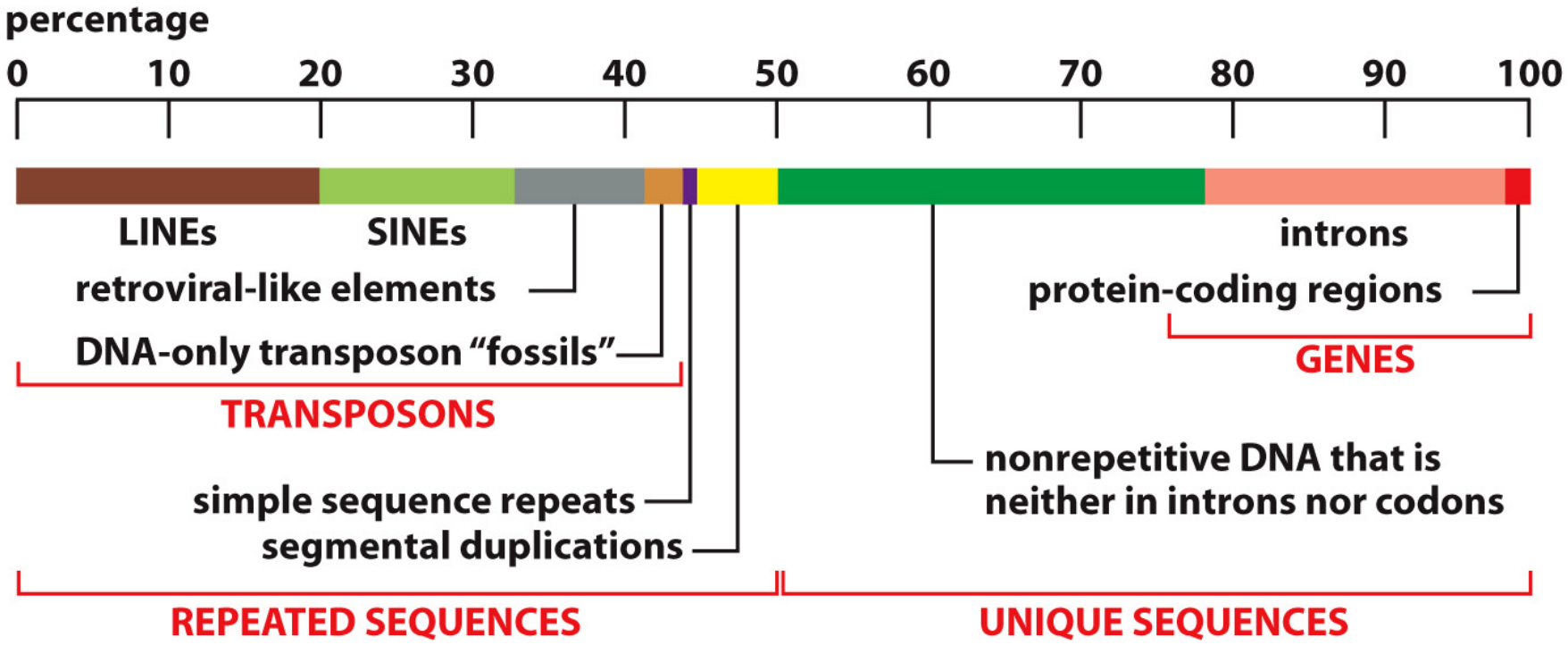


Figure 4-62 Molecular Biology of the Cell 6e (© Garland Science 2015)

Central Dogma of Biology

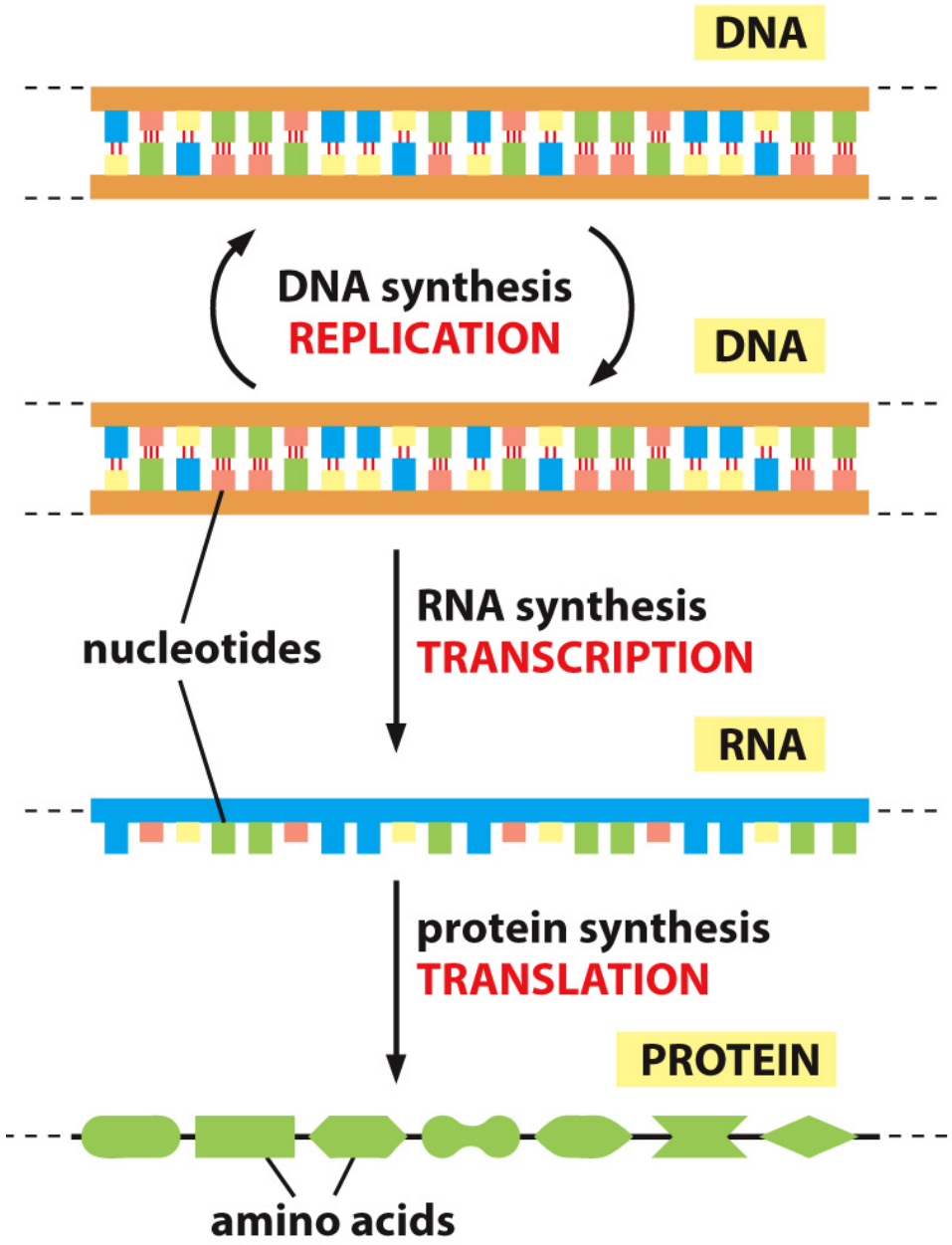


Figure 1-4 Molecular Biology of the Cell 6e (© Garland Science 2015)

What is a gene?

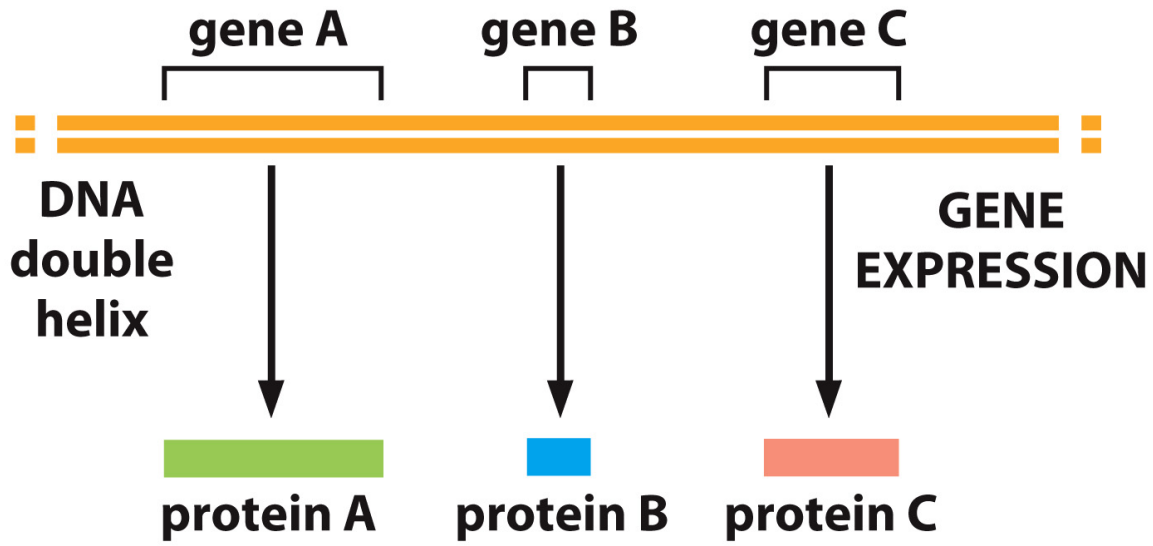


Figure 4-7 Molecular Biology of the Cell 6e (© Garland Science 2015)

Always write genes 5' → 3'

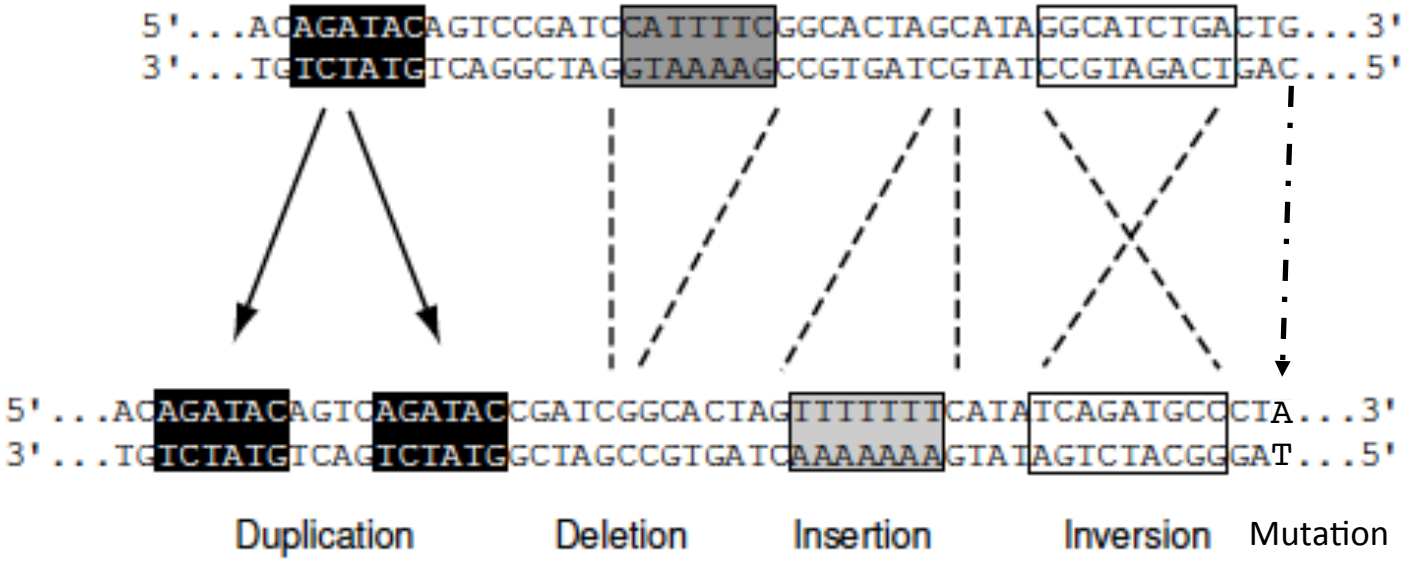
Eukaryotic genes almost all have introns

Genes can be on either strand of DNA

but always are 5'→3' since that's how they are "read" RNA polymerase and ribosome

Figure 4-8 Molecular Biology

DNA is not stable over time, it changes – leading to new sequences



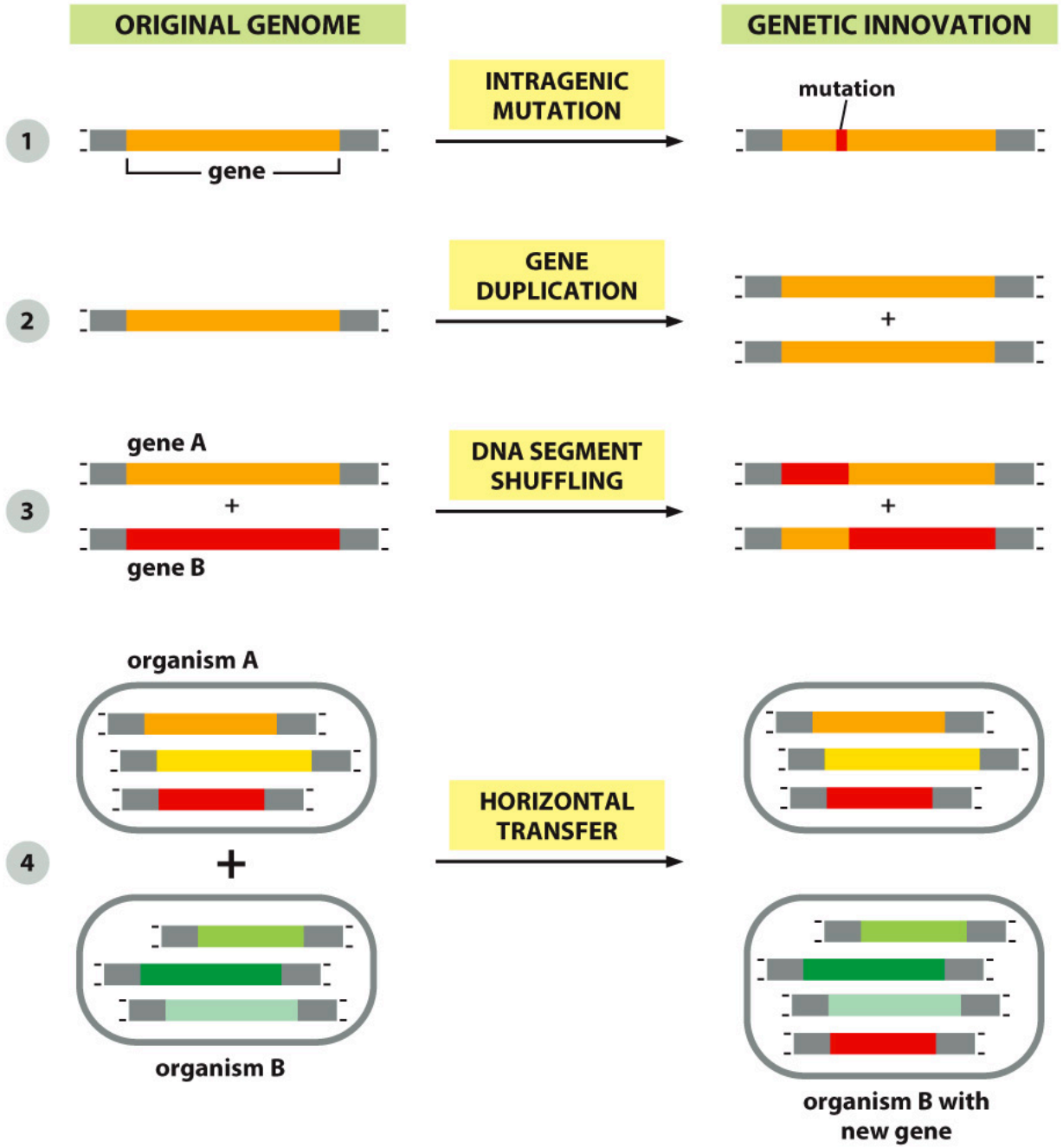


Figure 1-19 Molecular Biology of the Cell 6e (© Garland Science 2015)

These can eventually lead to new genes, or pseudogenes

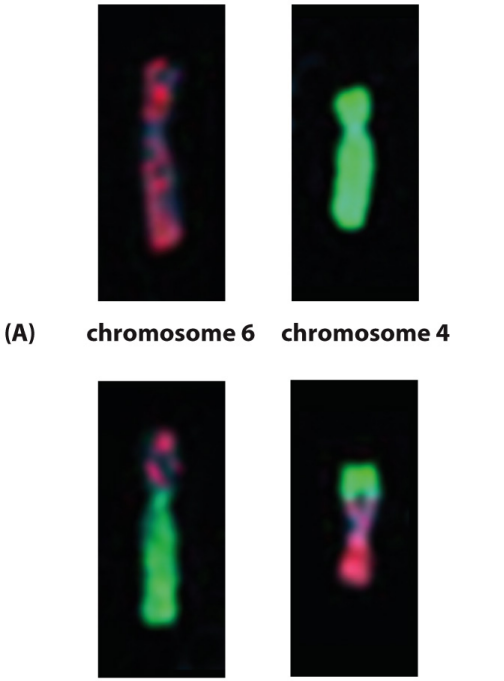


Figure 4-12 Molecular Biology of the Cell 6e (© Garland Science 2015)

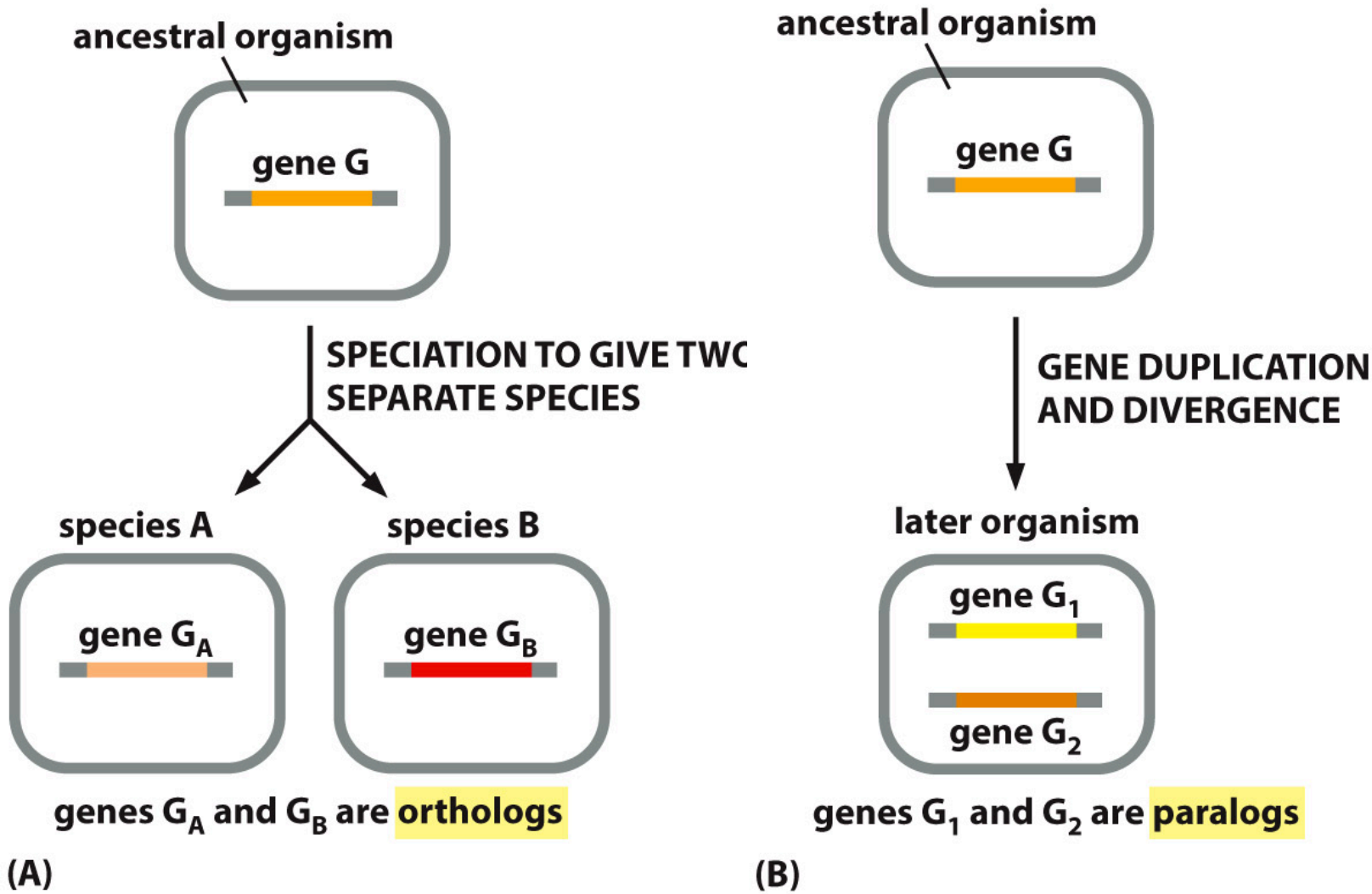
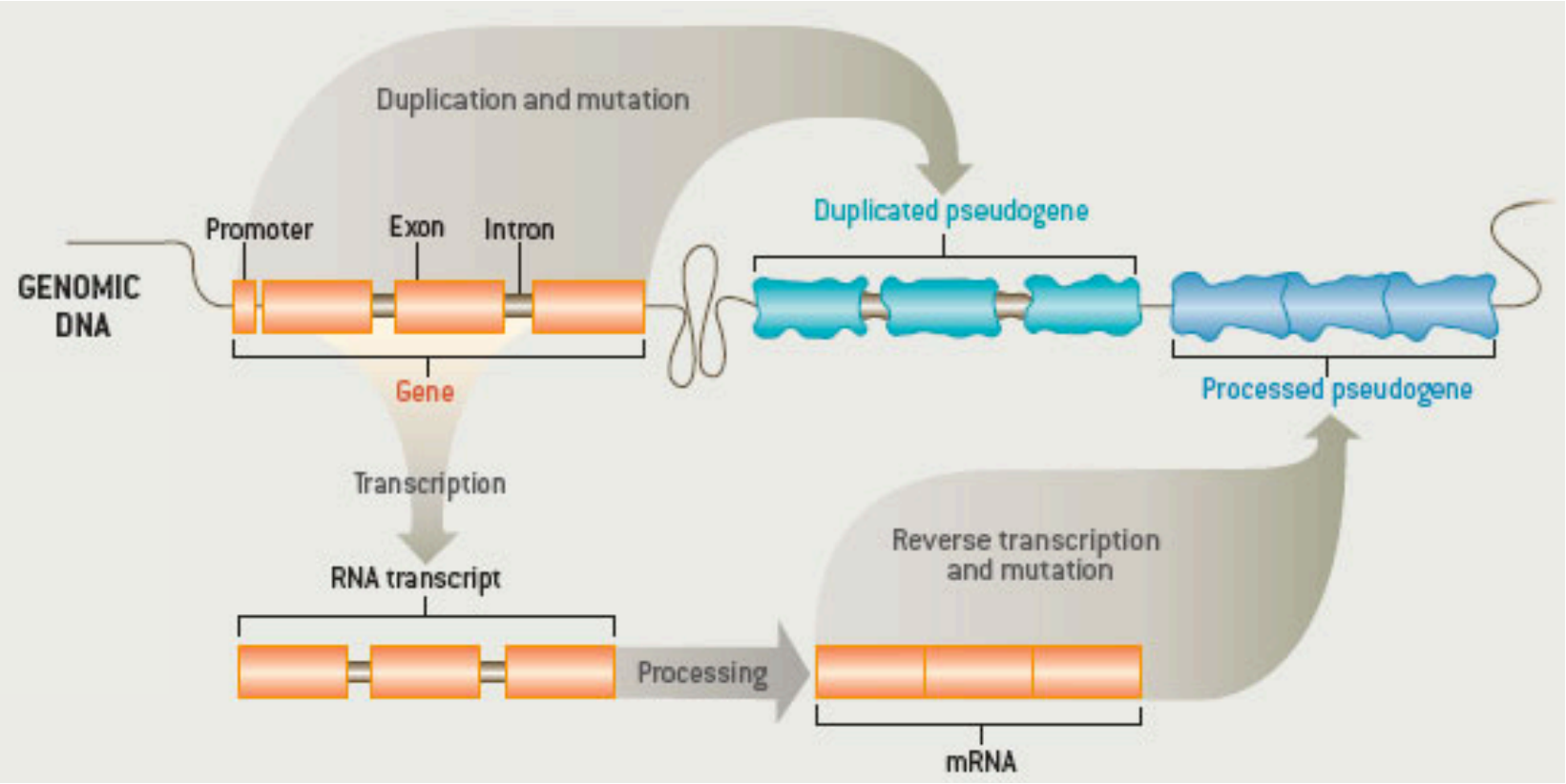
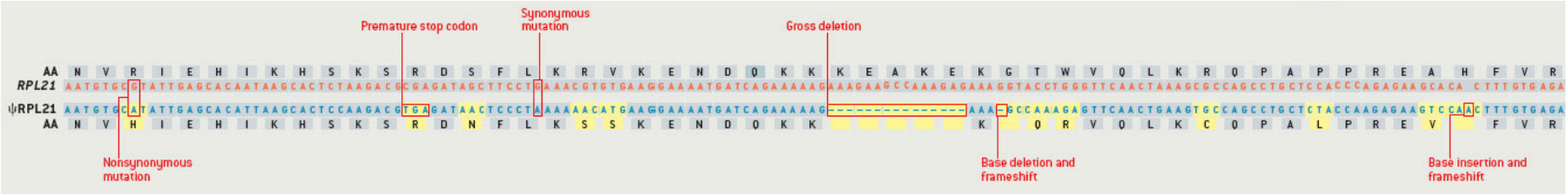


Figure 1-21 Molecular Biology of the Cell 6e (© Garland Science 2015)



Mark Gerstein and Deyou Zheng, *Scientific American* **295**, 48 - 55 (2006), doi:10.1038/scientificamerican0806-48

~20,000 pseudogenes in the human genome (more than functional genes)



Mark Gerstein and Deyou Zheng, *Scientific American* **295**, 48 - 55 (2006), doi:10.1038/scientificamerican0806-48

Mutations make the gene no longer able to produce protein

No longer under conservation pressure, what does that mean for mutation rate?

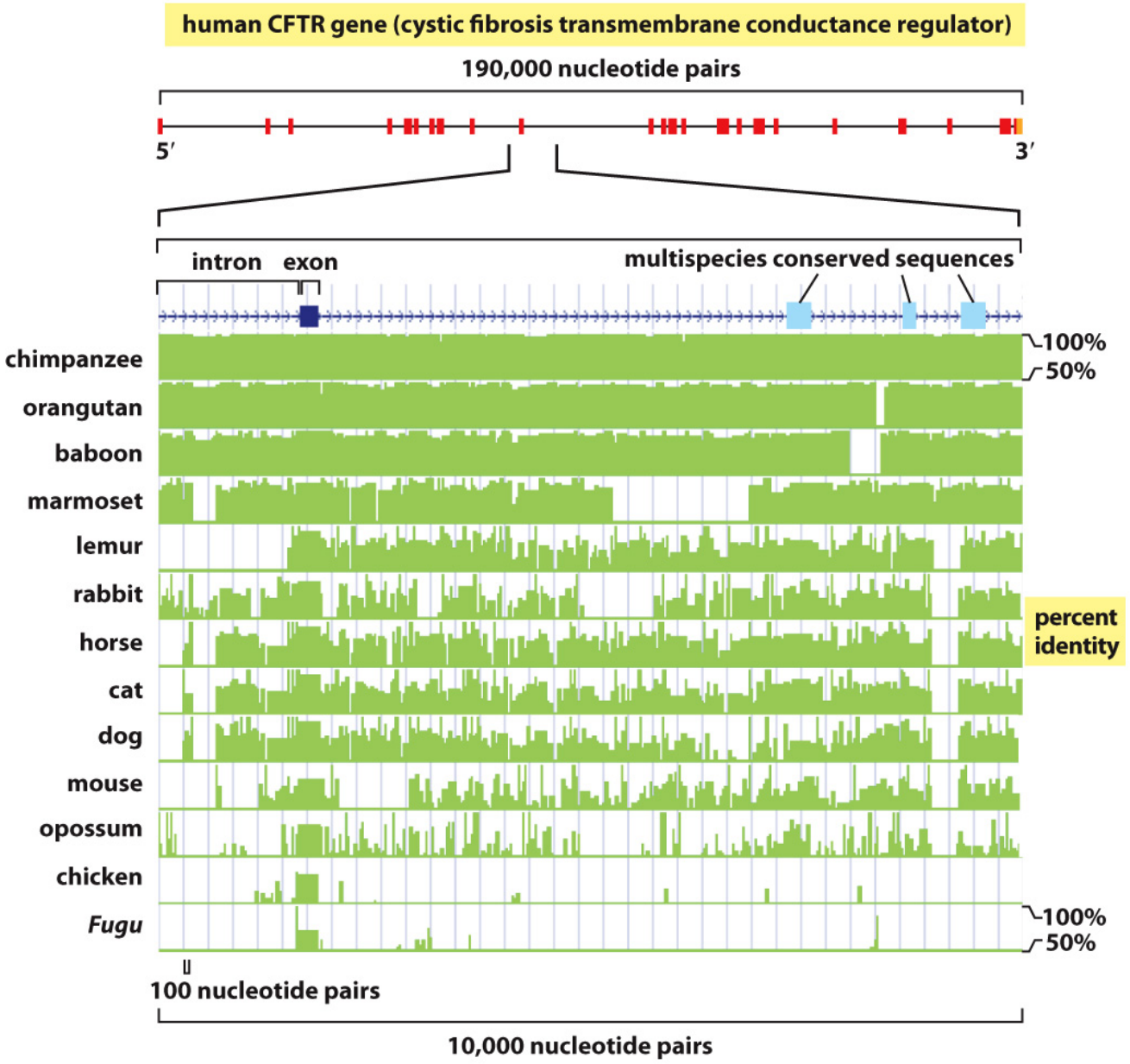


Figure 4-73 Molecular Biology of the Cell 6e (© Garland Science 2015)

Multi-species conservation can point to functional importance