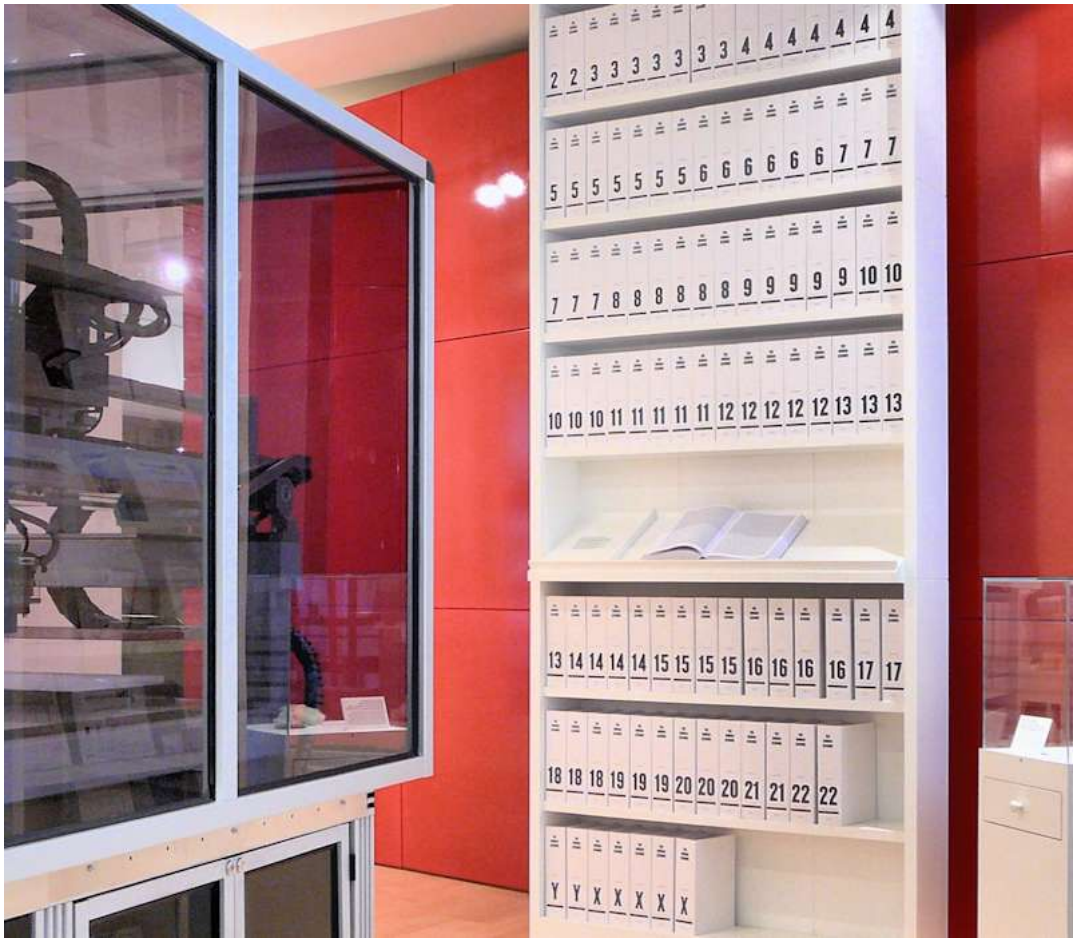
Lecture 17 – The Eukaryotic Genome



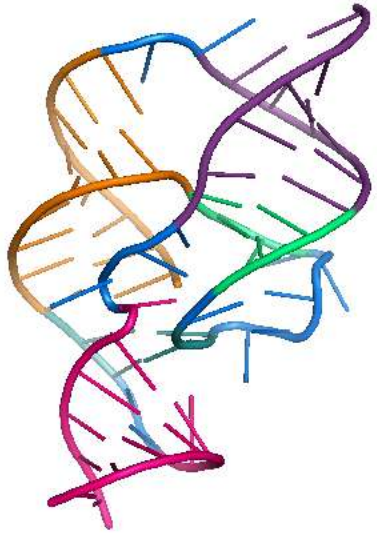
In this lecture…

- Origin and evolution of genomes
  - Molecular Clocks
  - Mitochondrial Eve
  - C-value paradox
- Sequencing the human genome
  - HGP vs. Celera
- Pharmacogenomics
- Revisiting the definition of a gene
  - Introns and exons
- The noncoding genome
  - Repetitive sequences
  - Pseudogenes
  - Transposons

Origins of genomes

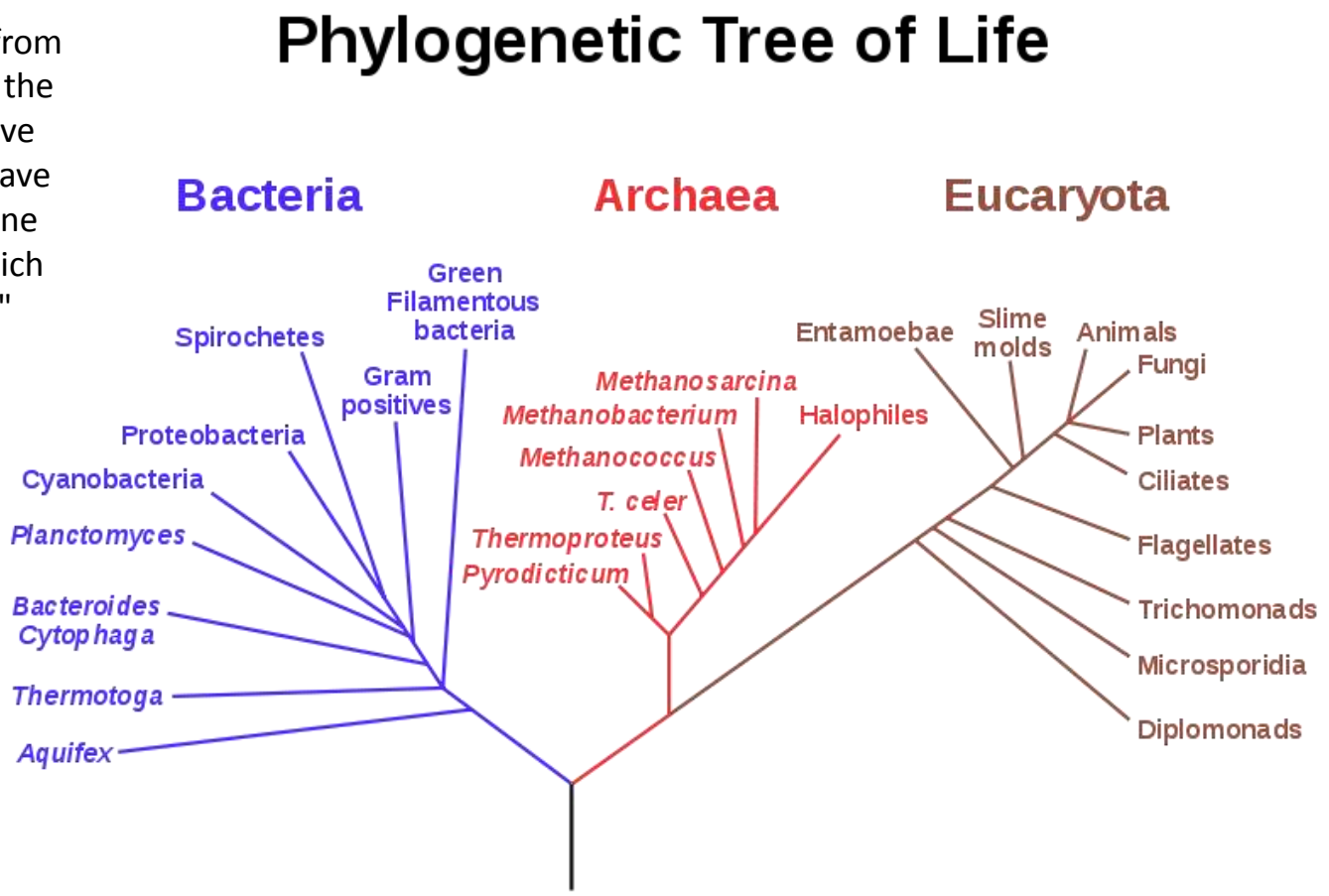
- The RNA world hypothesis
- Three pieces of evidence:
  - RNA can store genetic information
  - Ribozymes can act as enzymes
  - Suggests RNA may be an evolutionary remnant

Ribozymes are still in use today



LUCA: Last Universal Common Ancestor

"Therefore I should infer from analogy that probably all the organic beings which have ever lived on this earth have descended from some one primordial form, into which life was first breathed."



LUCA lived 3.5 to 3.8 billion years ago

Molecular ‘Clocks’

- Uses fossils and rate of mutations to deduce when a species diverged
- Nucleotide or amino acid sequences are compared among different species to date when they last shared common ancestor
- Molecular clocks ‘tick’ at different rates depending on the gene

Below: Alignment of human and rat cytochrome c amino acid sequences using BLAST.
Sequence similarity between the two proteins is 91%.

```
Human Cyt C: MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGIIW  61
Alignment:   MGDVEKGKKIF+ KC+QCHTVEKGGKHKTGPNLHGLFGRKTGQA G+SYT ANKNKGI W
Rat Cyt C:   MGDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAAGFSYTDANKNKGITW  61

Human Cyt C: GEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE               105
Alignment:   GEDTLMEYLENPKKYIPGTKMIF GIKKK ERADLIAYLKKATNE
Rat Cyt C:   GEDTLMEYLENPKKYIPGTKMIFAGIKKKGERADLIAYLKKATNE               105
```

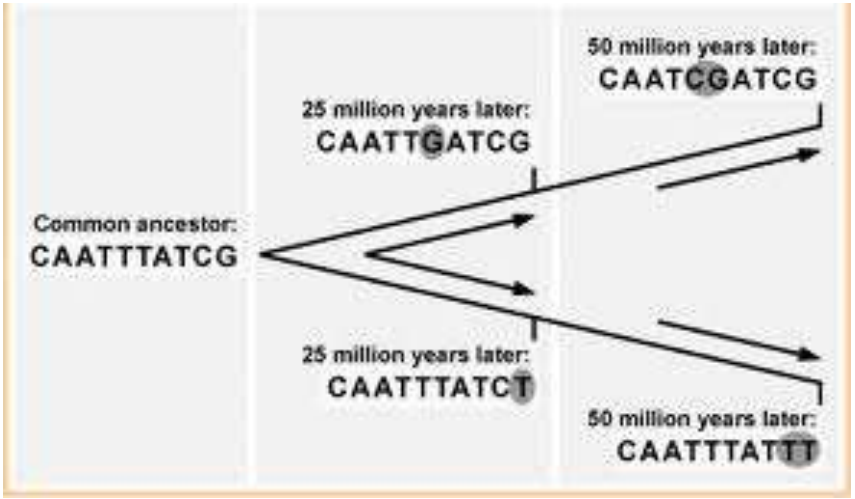
Below: Alignment of human and yeast cytochrome c amino acid sequences using BLAST.
Sequence similarity between the two proteins is 64%.

```
Human Cyt C: GDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGIIWG  61
Alignment:   G  +KG  +F  +C QCHTVEKGG HK GPNLHG+FGR +GQA GYSYT AN  K ++W
Yeast Cyt C: GSAKKGATLFKTRCLQCHTVEKGGPHKVGPNLHGIFGRHSGQAEGYSYTDANIKKNVLWD  66

Human Cyt C: EDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKA                   102
Alignment:   E+ + EYL NPKKYIPGTKM F G+KK+++R DLI YLKKA
Yeast Cyt C: ENNMSEYLTNPKKYIPGTKMAPGGLKKEKDRNDLITYLKKA                   107
```

Molecular Clocks

- Changes in nucleotide sequence of a gene are assumed to occur at a constant rate
- However, that rate may differ from gene to gene
- Genes that are responsible for basic functions accrue mutations much more slowly
- Compare gene mutation rate to the fossil record



Recent evolution in the human genome

http://www.plosgenetics.org/article/info:doi/10.1371/journal.pgen.0030090

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1762380/?tool=pmcentrez

- Strong selection in the past 10,000 years
  - After the emergence of agriculture
- Lowest amount of change in populations in the humid tropics
- Highest change in those living in cold/polar regions
  - Tibetans have evolved to cope with low oxygen levels 3,000 years ago
  - Caucasians have evolved lighter skin to be able to produce vitamin D (25 genes involved in skin color)

Mitochondrial Eve

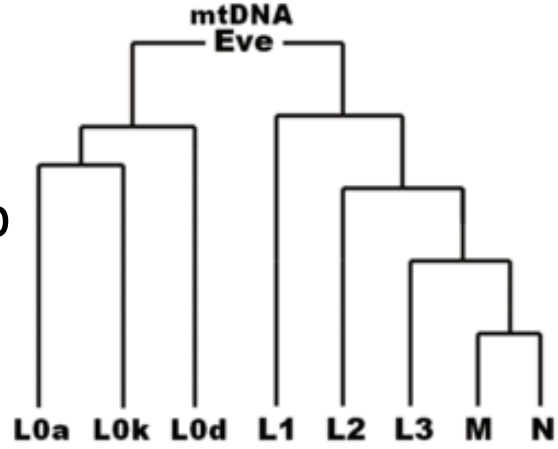
- The woman from whom all living women are descended from
  - All mtDNA is directly passed on from mother to offspring without recombination
  - mtDNA evolves quickly since there is no system to check for errors in DNA, so it's possible to calculate a 'clock'
- Lived about 200,000 years ago in West Africa
- Helped to boost the "Out of Africa" hypothesis in 1987
  - Vs. Multiregional origin of modern humans

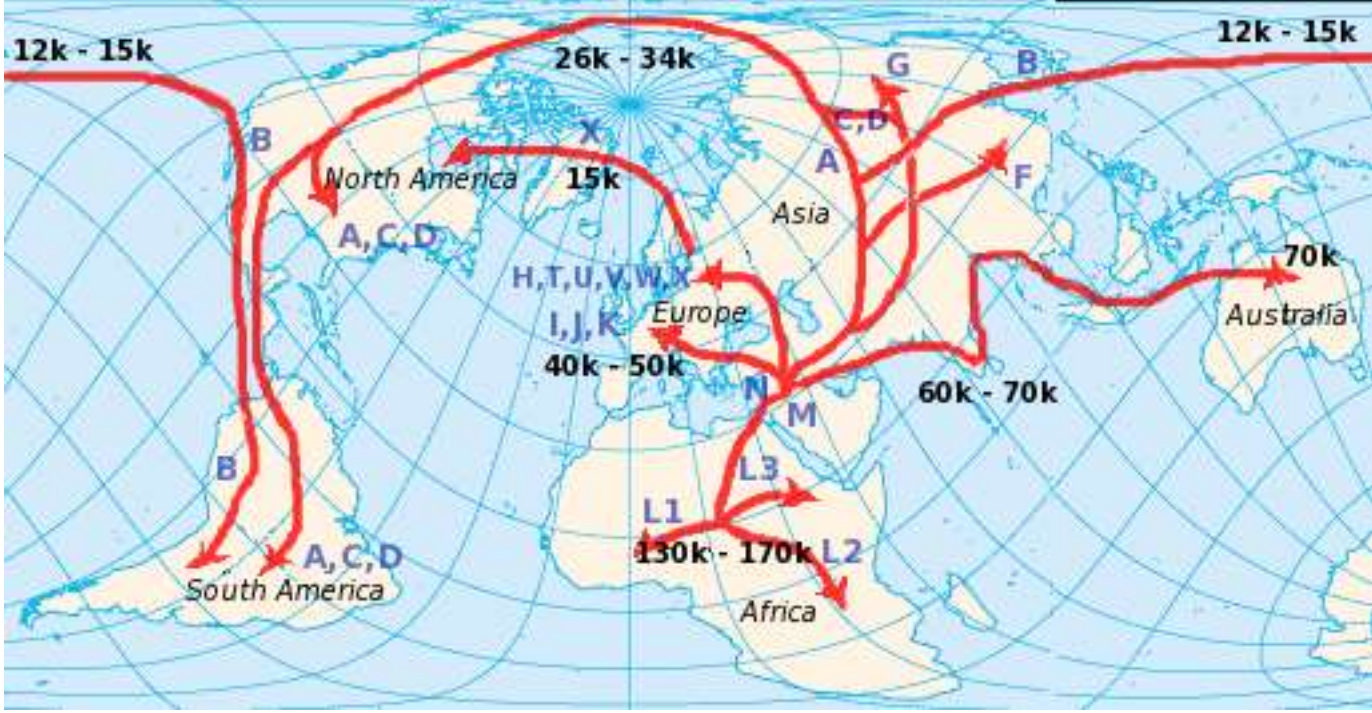
Mitochondrial Haplogroups

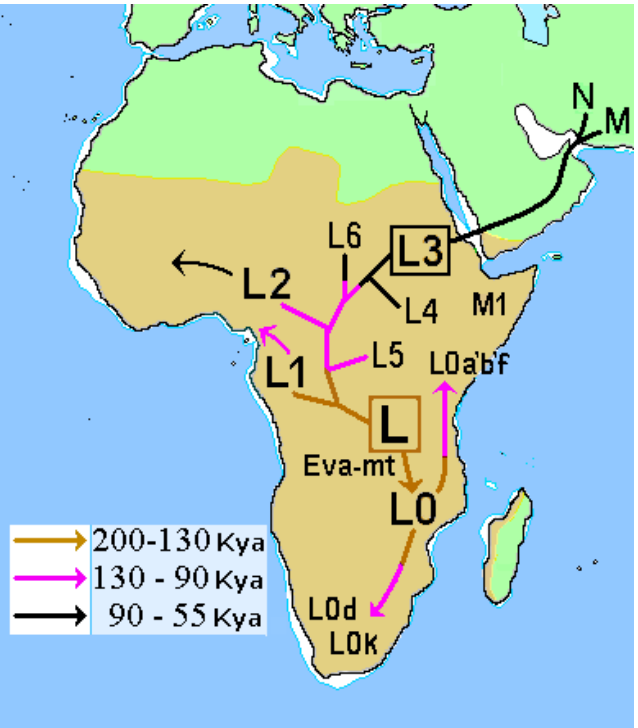
- Mitochondrial genomes are separated into many **haplogroups**
- Several mitochondrial **haplogroups**
  - Group of similar **haplotypes**
  - L0-L6
    - Two further subhaplogroups M and N. which are both descended from L3
    - Lo-L6 are African
    - M, N and their subhaplogroup are Asian and Caucasian

Haplotype = a set of linked mutations

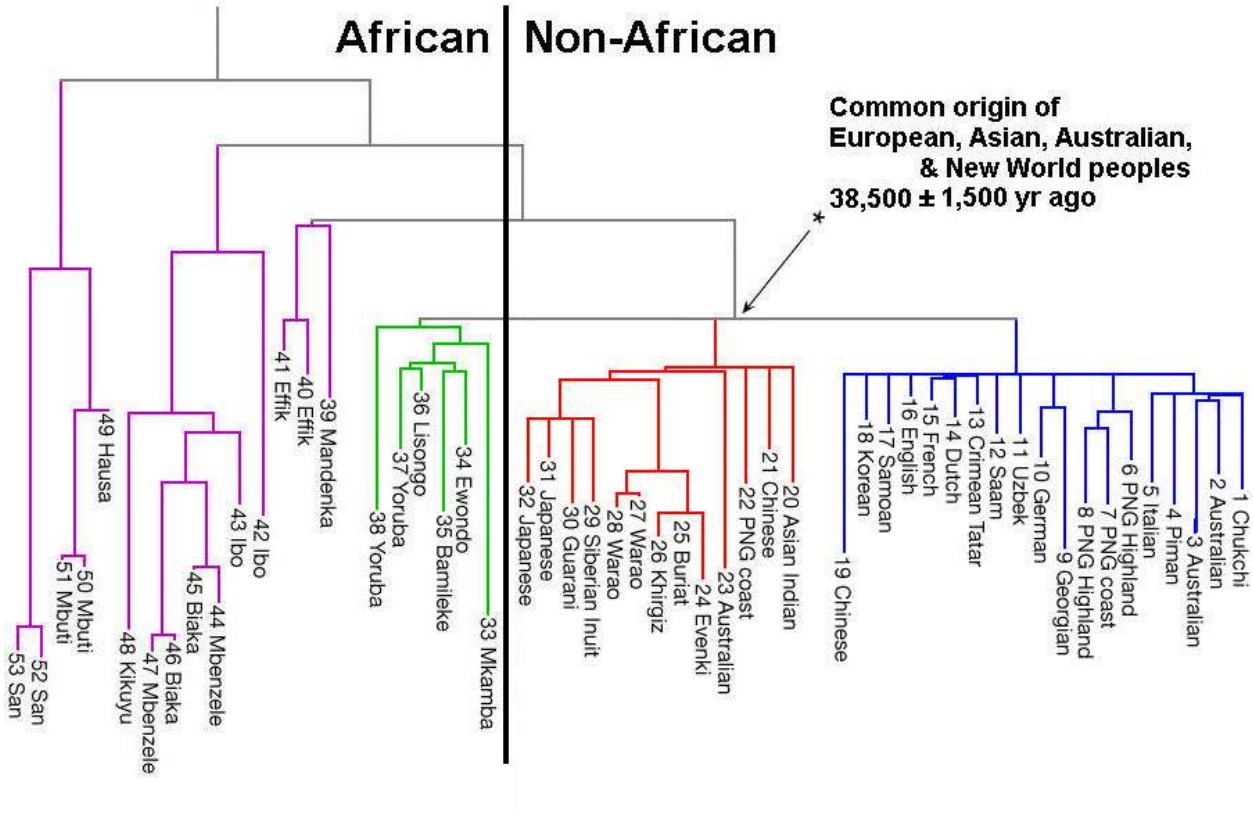
Haplotype = combinations of alleles at adjacent loci







We can trace the migration of the human species through where and when we find which haplogroups

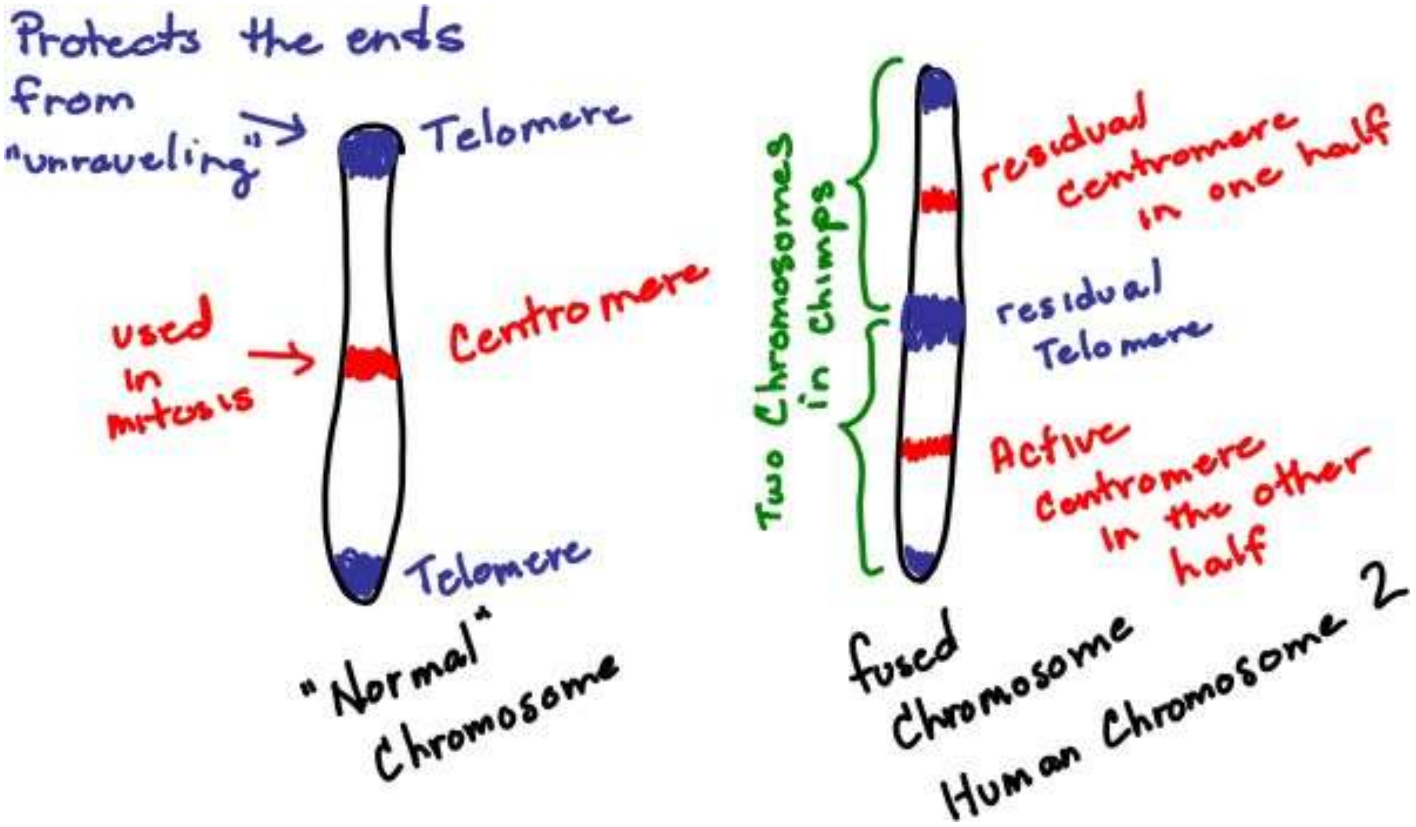


- There is far more genetic diversity between two randomly selected Africans than there are between a Japanese and Caucasian person

Humans vs. Chimps

- 1% divergence between genes shared between humans and chimpanzees
- However, 6% of genes are not shared between humans and chimps
  - Large amounts of loss and gain of genes since evolutionary split
- Human chromosome 2 is a result of the fusion of the chimp chromosomes 2A and 2B (formerly 12 and 13)
- Humans have lost many of their olfactory genes!

Chimp chromosomal fusion

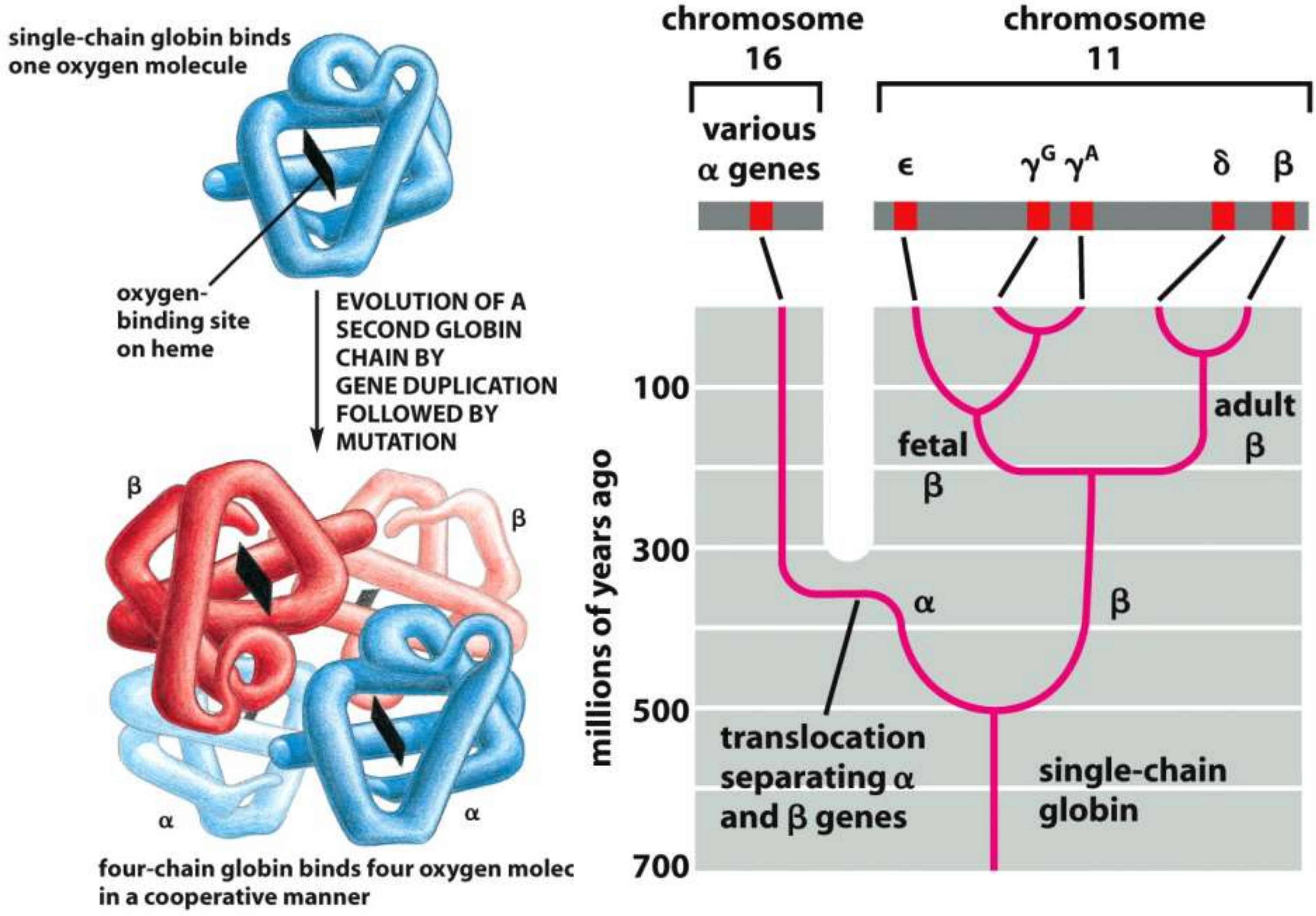


Universal Genes

- What genes do you expect to see in all organisms? In all mammals? In just humans?
- Why might these genes be universal and virtually unchanged in all organisms?
  - Ribosomal RNA
  - Cytochromes
  - RuBisCo
  - Histones
    - On average, one mutation every billion years

Sources of genomic diversity

- Mutations
- Chromosome translocations, inversions, deletions
- Homologous recombination
- Gene duplications
- Genetic drift


single-chain globin binds one oxygen molecule
oxygen-binding site on heme
EVOLUTION OF A SECOND GLOBIN CHAIN BY GENE DUPLICATION FOLLOWED BY MUTATION
β
β
α
α
four-chain globin binds four oxygen molec
in a cooperative manner
chromosome 16
chromosome 11
various α genes
ϵ
γG γA
δ
β
100
300
500
700
millions of years ago
fetal β
adult β
α
β
translocation separating α and β genes
single-chain globin

Genetic Drift

- Evolution caused by chance, not by natural selection
- Population bottlenecking, random chance, bad luck
- Argued to be a bigger contributor to speciation than natural selection

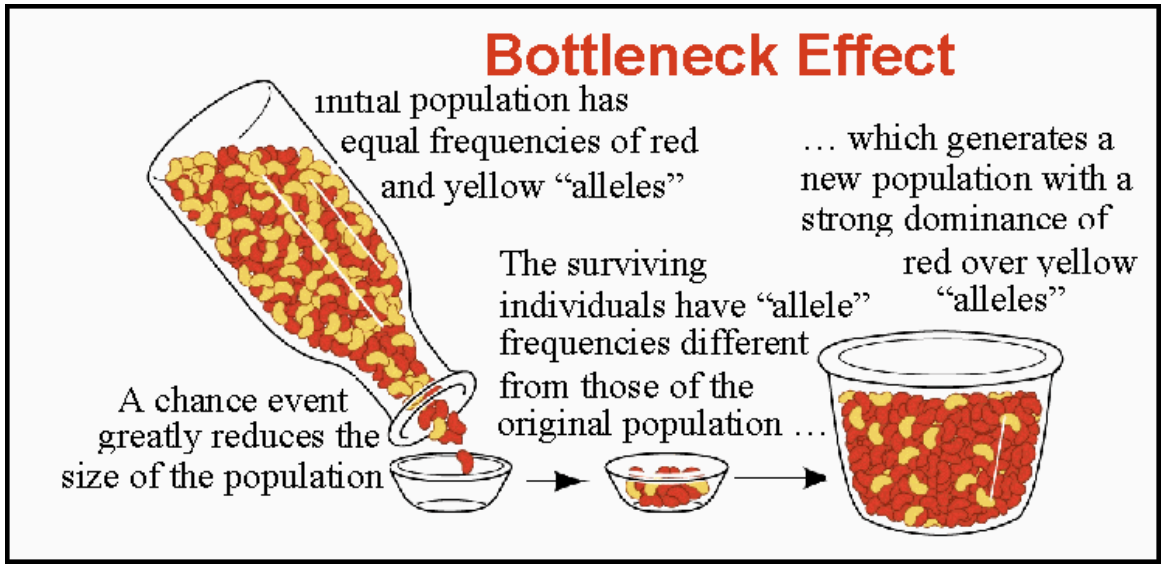


Figure 38-12 AN AMISH CHILD WITH ELLIS–VAN CREVELD SYNDROME.
The child has shortened limbs and six fingers on each hand. All the Amish with this syndrome are descendants of a single couple that helped found the Amish community in Lancaster County, Pennsylvania, in 1744. Because of inbreeding in the isolated community, the recessive trait is now common.

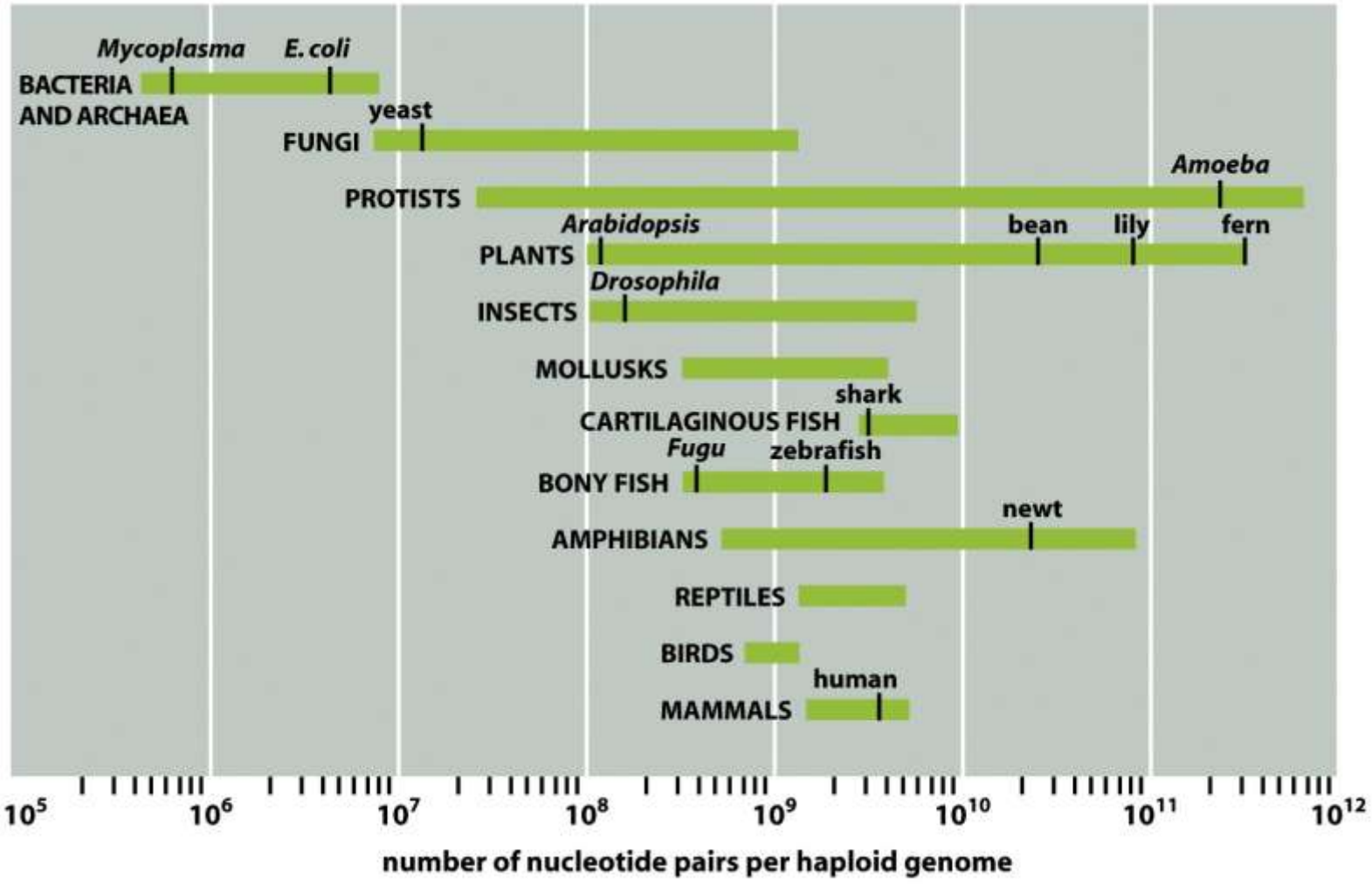
Size of the genome doesn't indicate complexity!

- Eukaryotic genomes are *usually* bigger and more complex than prokaryotic genomes
- Human genes seem to be clustered together with wide expanses of 'junk' DNA in between
  - Other organisms' genes are more spaced out

Size of the genome doesn't indicate complexity!

- Free-living bacteria and archaea have 1,500 to 7,500 genes
- Unicellular fungi have from about 5,000 genes and multicellular eukaryotes from 40,000 genes
- Number of genes is not correlated to genome size
  - Nematode *C. elegans* has 100 Mb and 20,000 genes, while Drosophila has 165 Mb and 13,700 genes
  - Humans and other mammals have the lowest gene density, or number of genes, in a given length of DNA

C-Value Paradox



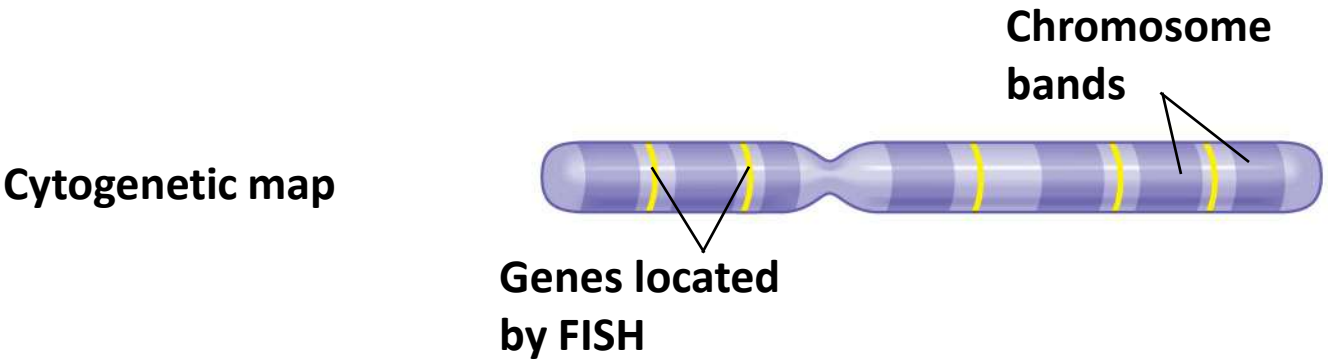
Sequencing genomes

- DNA was first sequenced in 1977 by Fred Sanger
  - The **Sanger method** is VERY tedious and time-consuming
- Complete genome sequences exist for a human, chimpanzee, *E. coli*, brewer's yeast, corn, fruit fly, house mouse, rhesus macaque, and other organisms
- The scientific community strongly favored sequencing the human genome
- Many benefits:
  - More powerful, safer drugs
  - Advanced screening for disease
  - Decreased health care costs
  - Better understanding of human evolution

HGP: Human Genome Project

- Government-funded, international effort
  - NIH, Dept. of Energy and Wellcome Trust
- Started in 1989, expected to take $3 billion, 15 years
- Originally headed by James Watson until Francis Collins took over

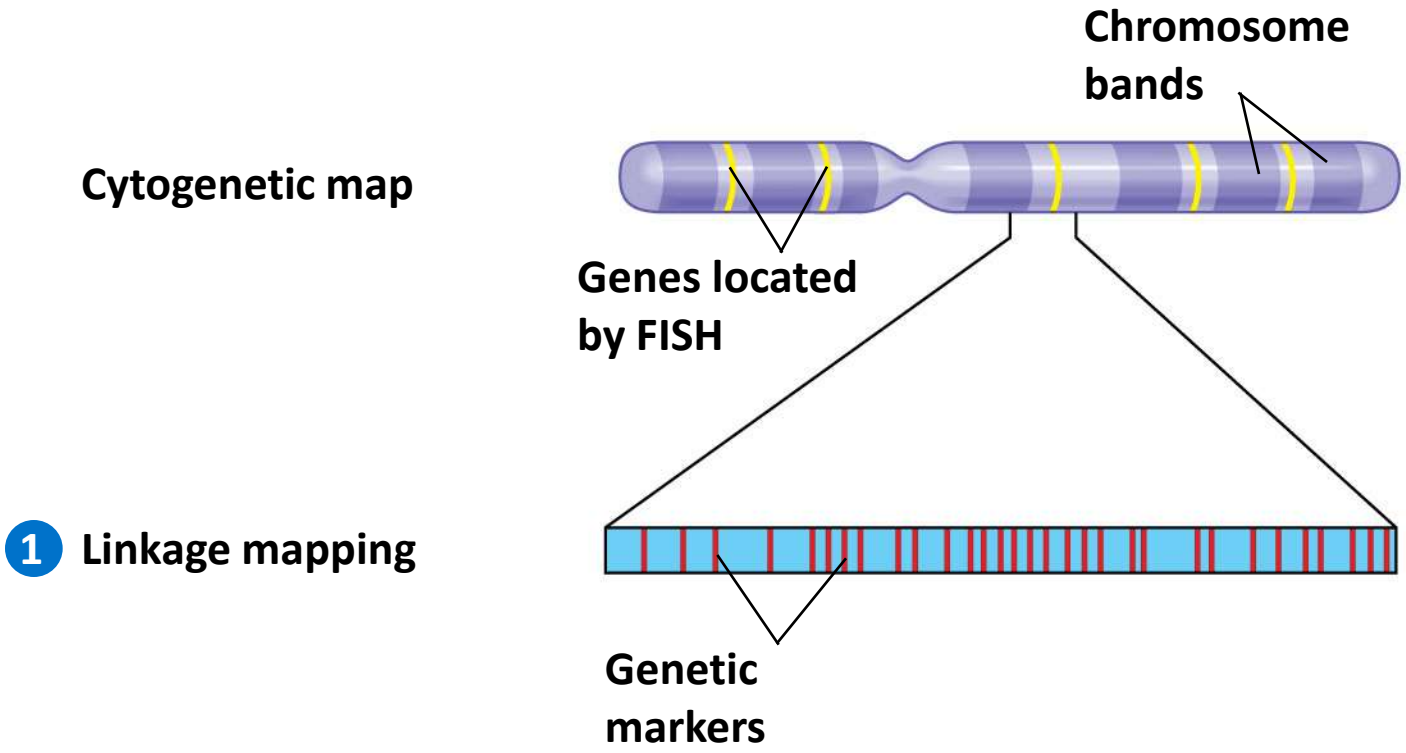
Fig



- Technology limited how many nucleotides we could read at one time
- Researchers began by building very detailed linkage maps of human chromosomes, identifying thousands of “landmarks”

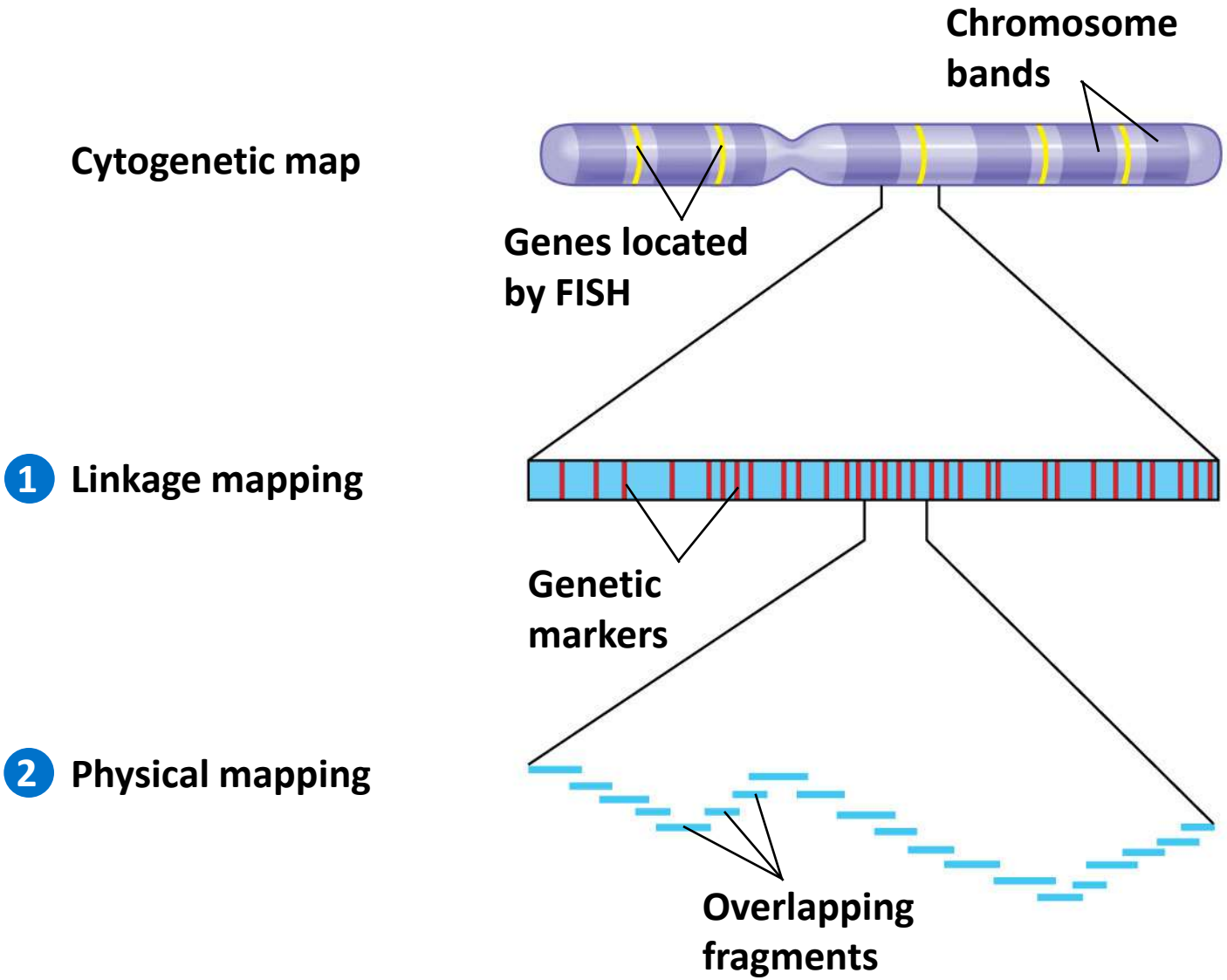
http://www.genome.gov/25019885

Fig



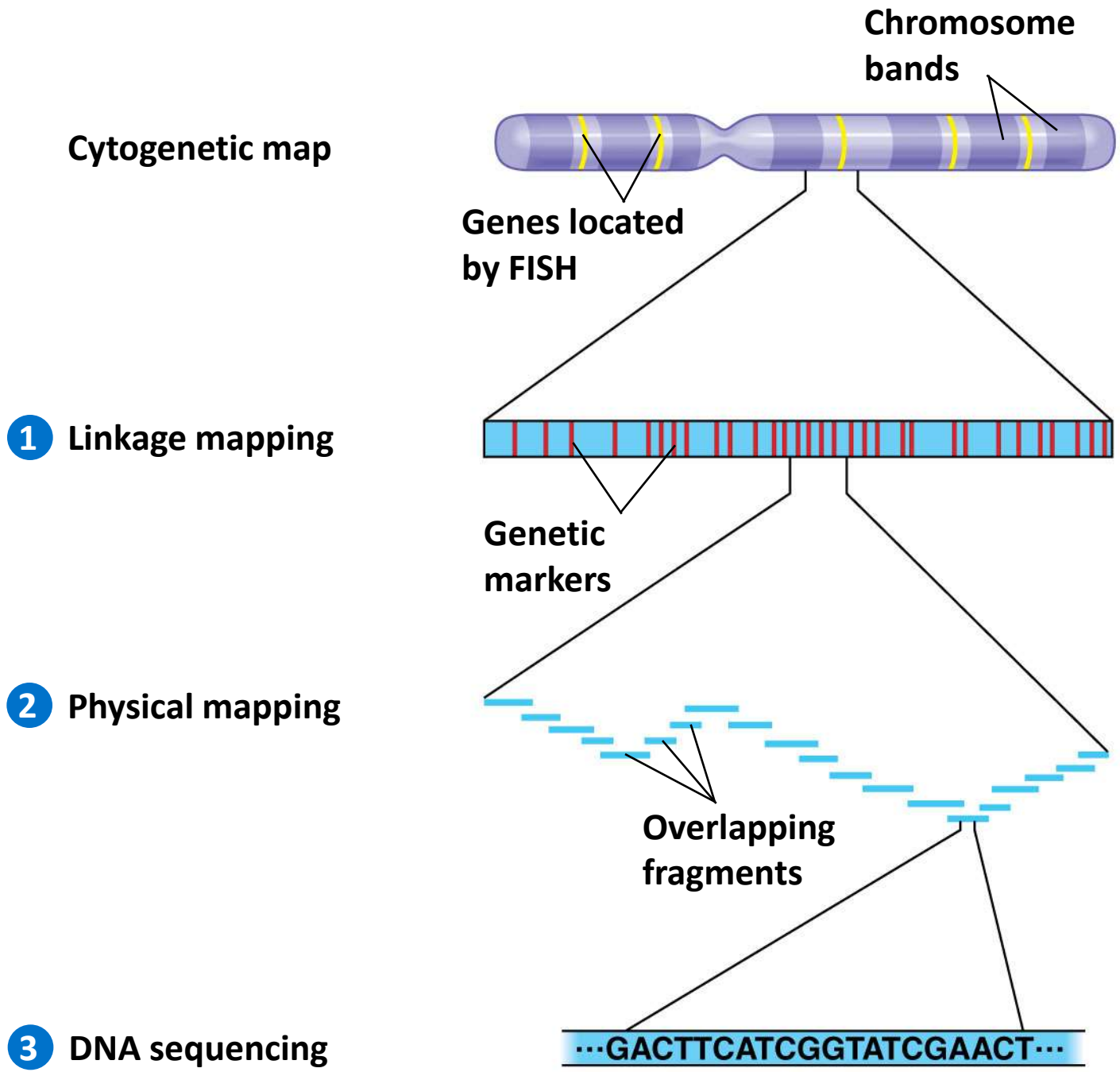
- Researchers established millions of landmarks throughout the genome

Fig



- The genome was broken into overlapping fragments called “contigs”
- Contigs were sequenced, then their overlapping ends matched up with each other to produce the whole sequence

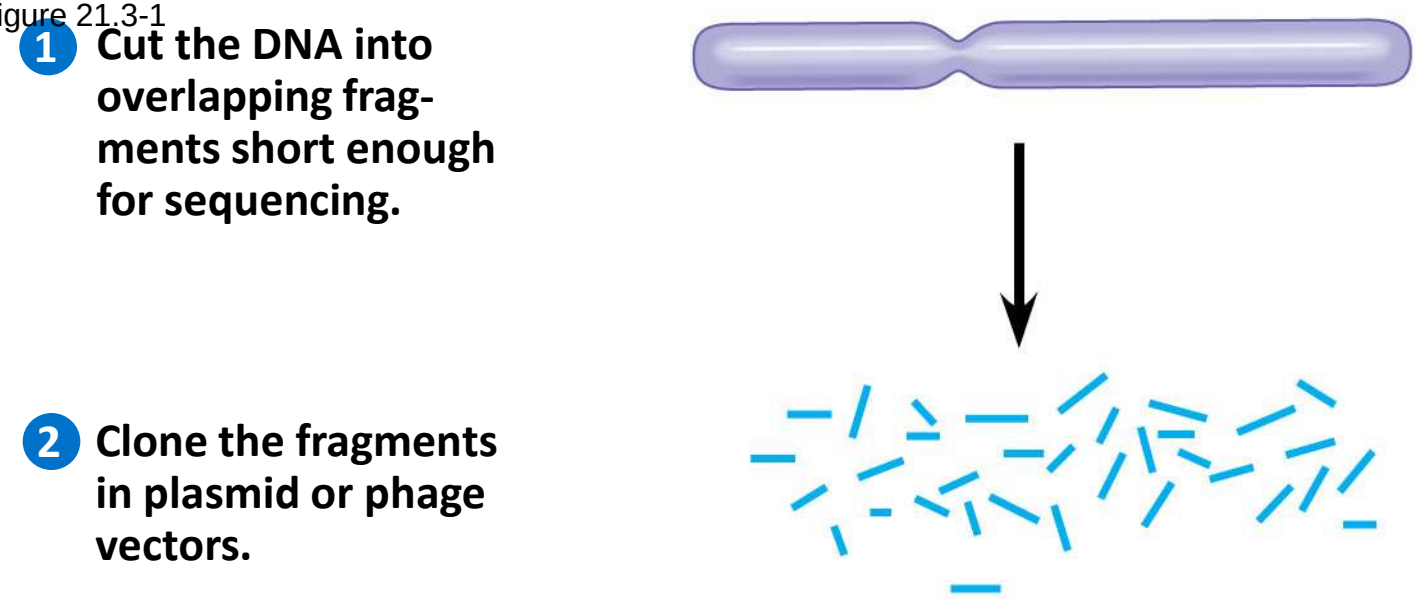
Fig



Craig Venter and Celera

- Venter was a scientist at NIH
- Decided he could do it better than the HGP using a technique called whole-genome shotgun sequencing
  - $300,000 and 3 years
  - This approach skips genetic and physical mapping and sequences random DNA fragments directly
- Started a company and began a private venture to sequence the human genome
- Sought out the help of venture capitalists
  - Promised the patenting of disease genes
  - Also sequenced *Drosophila* and *Haemophilus influenzae*

Figure 21.3-1



- No landmarks were used
- Researchers counted on the overlapping ends to be able to piece the sequences together

Figure 21.3-2

1 **Cut the DNA into overlapping fragments short enough for sequencing.**

2 **Clone the fragments in plasmid or phage vectors.**

3 **Sequence each fragment.**

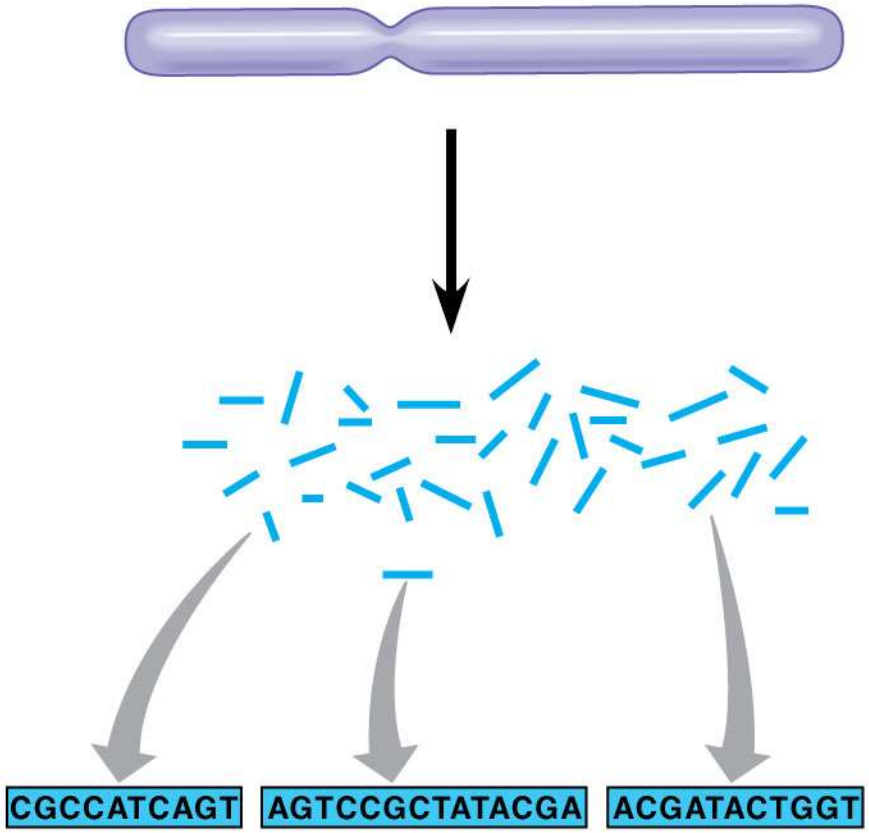
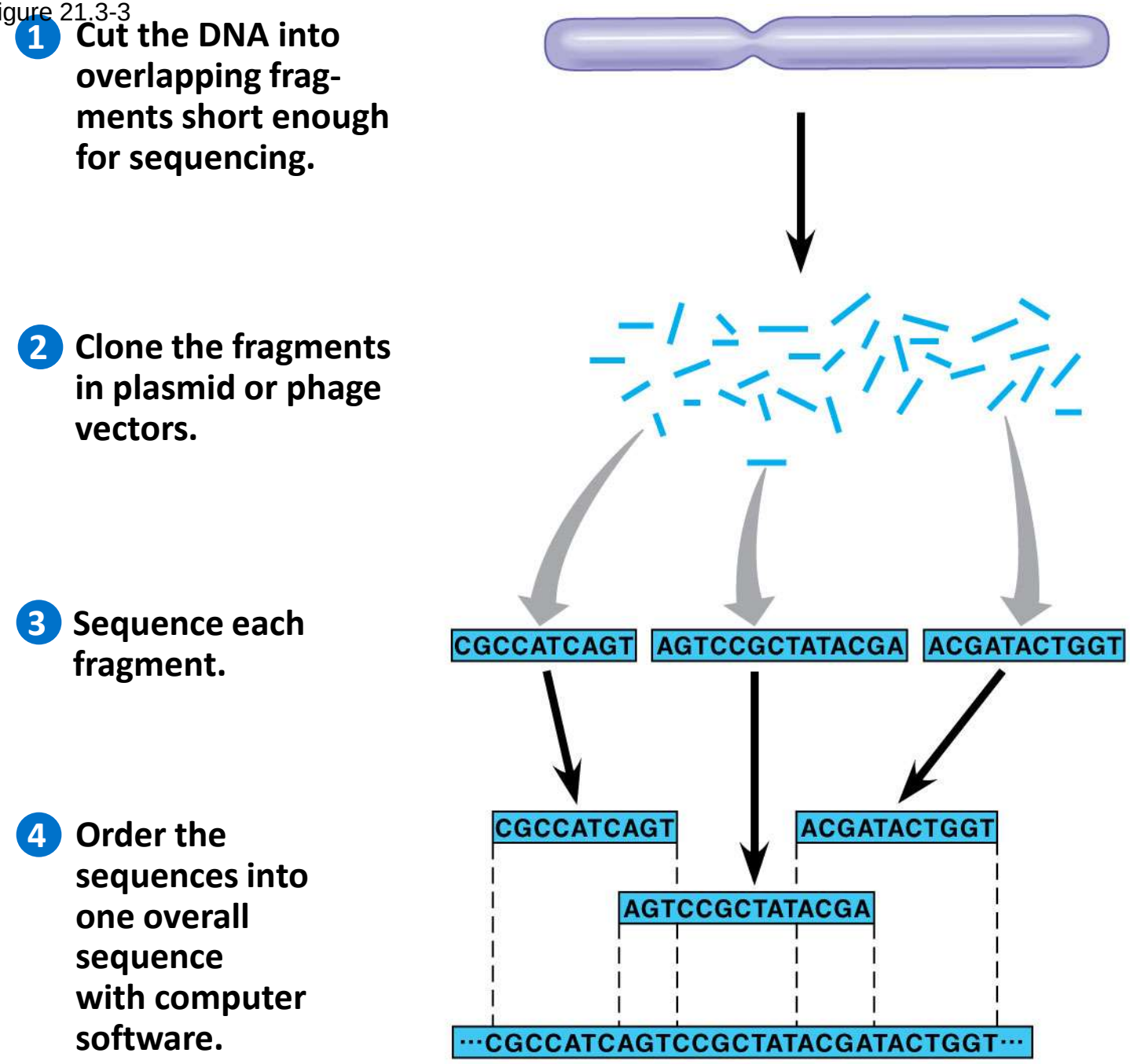


Figure 21.3-3



Venter and Collins: Acrimonious Rivalry

- “Darth Venter”
- The HGP published their data as they sequenced online for everyone
- Celera used the HGP data to piece together their sequences
- The HGP started using the sequencing machines Celera used

The end?

- ‘Rough draft’ finished in 2000 by both public and private groups
- ‘Final draft’ published 2003
- 8% still unsequenced (heterochromatin)

http://uswest.ensembl.org/index.html

Human Genome Statistics

- 3.3 billion base pairs in the haploid genome
- 20,000ish genes
- Average of 3000 nucleotides/gene, with largest dystrophin at 2.4 million nucleotides
- The human genome's gene-dense "urban centers" are predominantly composed of the DNA building blocks G and C
- The gene-poor "deserts" are rich in the DNA building blocks A and T.
- Chromosome 1 has the most genes (2968), and the Y chromosome has the fewest (231)
- Almost all (99.9%) nucleotide bases are exactly the same in all people

Implications

- We have the nucleotide sequences, but we don't necessarily know what they mean
  - Functions unknown for over 50% of genes
- Lots of initial skepticism about the whole-genome shotgun approach, but it is now widely used as the sequencing method of choice
- The development of newer sequencing techniques has resulted in massive increases in speed and decreases in cost
- The sequencing of the human genome has spawned many other efforts to understand the genome
  - HapMap
  - SNP database
  - Cancer Genome Atlas

$1000 Genome

- The ability to sequence an individual's genome for $1000 or less could be a huge boon to pharmacogenomics and personalized medicine
- We just don't know enough about how our genotype produces phenotype for this to be very useful

"Despite all the hype about new genetic knowledge, some of it is going to fall flat because the recommendations are going to involve simple lifestyle changes you could do anyway."

http://www.nytimes.com/2011/01/05/health/05gene.html?_r=2&pagewanted=1&ref=hospitals

“Why did people think there were so many human genes? It's because they thought there was going to be one gene for each human trait. And if you want to cure greed, you change the greed gene, right? Or the envy gene, which is probably far more dangerous. But it turns out that we're pretty complex. If you want to find out why someone gets Alzheimer's or cancer, then it is not enough to look at one gene. To do so, we have to have the whole picture. It's like saying you want to explore Valencia and the only thing you can see is this table. You see a little rust, but that tells you nothing about Valencia other than that the air is maybe salty. That's where we are with the genome. We know nothing.”

-- Venter, 2010

Pharmacogenomics: CYP450

- ‘Personalized medicine’ - how an individual’s genotype affects their reactions to drugs
- 160,000 deaths and 2.2 million ADRs due to improper/mistaken administration of prescription drugs
- **CYP450** (cytochrome P450) is an enzyme responsible for breaking down 80% of current prescription drugs
  - Less active or completely inactive forms of CYP450 metabolize drugs slower
  - These patients are more susceptible to a drug overdose

Pharmacogenomics: TPMT

- TPMT, thiopurine methyltransferase, metabolizes two thiopurine drugs used as immunosuppressants
- People with a deficiency in TPMT need only 6-10% of the standard dose of drug
  - At risk of severe bone marrow suppression if given too much
  - Homozygous recessive genotype
  - Heterozygotes can tolerate a higher dosage, dominant homozygotes an even higher one

Pharmacogenomics: KRAS

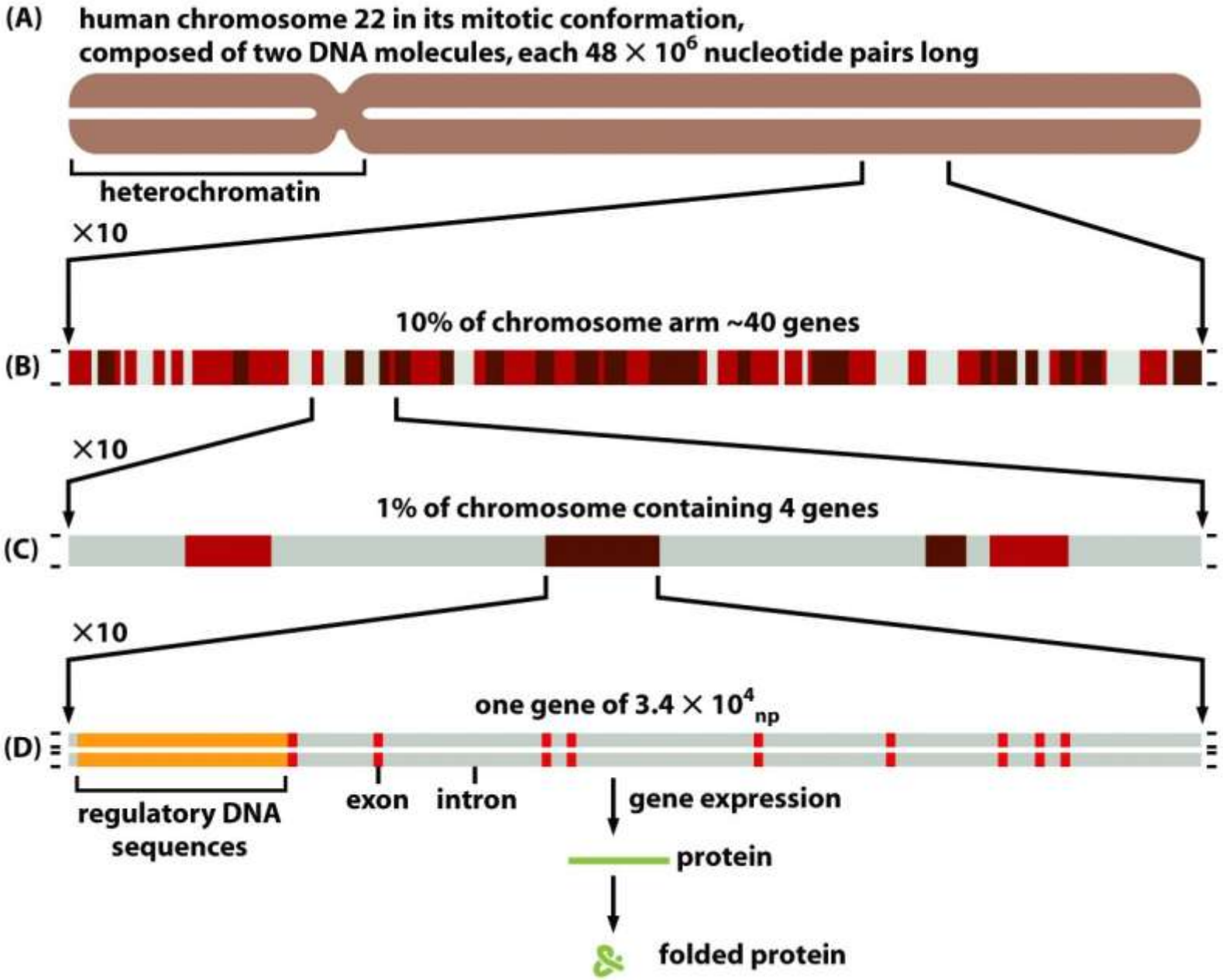
- The drugs Vectibix and Erbitux are used to treat metastatic colorectal cancer
- KRAS gene encodes for a G protein in the EGFR pathway. If KRAS is mutated, neither of these drugs will work
  - Epidermal growth factor receptor, a GPCR
- 70% of patients have wild-type KRAS

Pharmacogenomics: Limitations

- In limited use today
- Sequencing the human genome didn't lead to understanding of CYP450, TPMT, and KRAS
- Many (most?) doctors still don't use or understand the tools pharmacogenomics offers
- Drug companies don't have incentive to create multiple versions of a drug to target all genotypes
- Limited drug alternatives for those who can't take current drugs due to genotype
- Genomics is hard!

Gene stucture

- What is a gene?
  - Exact definition is in flux
  - At one time we considered a gene to be anything that produces a protein
    - However, things like ribozymes, siRNA, transposons, etc. complicate the definition
  - If you define a gene as anything that codes for RNA, most genes don't encode for protein
- Contains:
  - Promoter – directs when and where a gene is expressed
  - Exons and introns

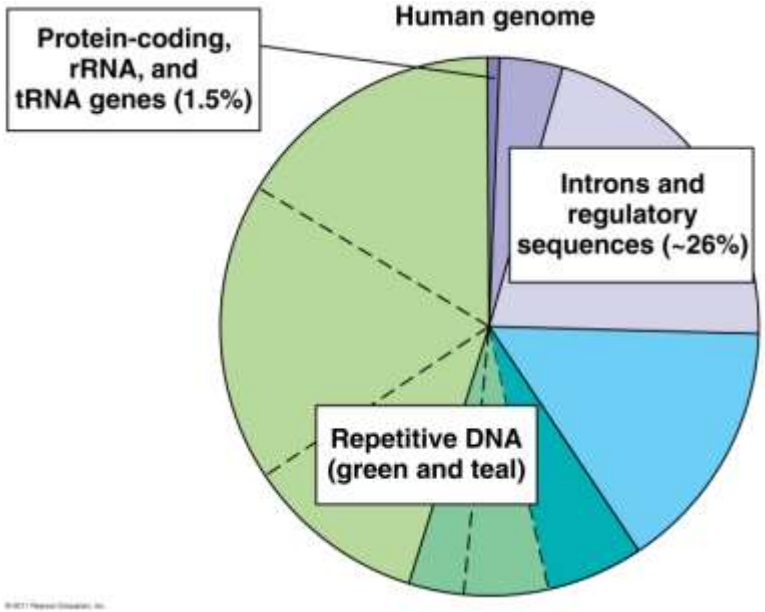

(A) human chromosome 22 in its mitotic conformation,
composed of two DNA molecules, each 48 × 10^6 nucleotide pairs long
heterochromatin
×10
10% of chromosome arm ~40 genes
(B)
×10
1% of chromosome containing 4 genes
(C)
×10
one gene of 3.4 × 10^4 np
(D)
regulatory DNA sequences
exon
intron
gene expression
protein
folded protein

What happens to those introns?

- Recycled back into nucleotide monomers?
- Probably further processed to create noncoding RNA molecules like ribozymes
- Might be transposons

Only 1% genes?!

- 1.5% figure includes things other than genes
- The bulk of most eukaryotic genomes neither encodes proteins nor functional RNAs
- Much evidence indicates that noncoding DNA (previously called “junk DNA”) plays important roles in the cell
- For example, genomes of humans, rats, and mice show high sequence conservation for about 500 noncoding regions



What is the stuff that is not genes?

- Pseudogenes
  - HERVs
- Repetitive DNA
  - LINES/SINES
    - Transposons and retrotransposons
    - ALU elements
  - Microsatellites

Pseudogenes

- **Pseudogenes** are former genes that have accumulated mutations and are no longer expressed in the cell
- Every pseudogene has a sequence similar to a functional gene
- Types of pseudogenes:
  - Disabled genes – frameshift mutation, premature stop signal, mutation in promoter
  - Gene duplications – the duplicate copy is not subject to evolutionary pressure and so undergoes rapid mutation
  - Processed/retrotransposed genes – after exons are spliced out of an mRNA transcript, genes are reverse transcribed back into DNA then inserted into the chromosome

HERVs: Human Endogenous Retroviruses

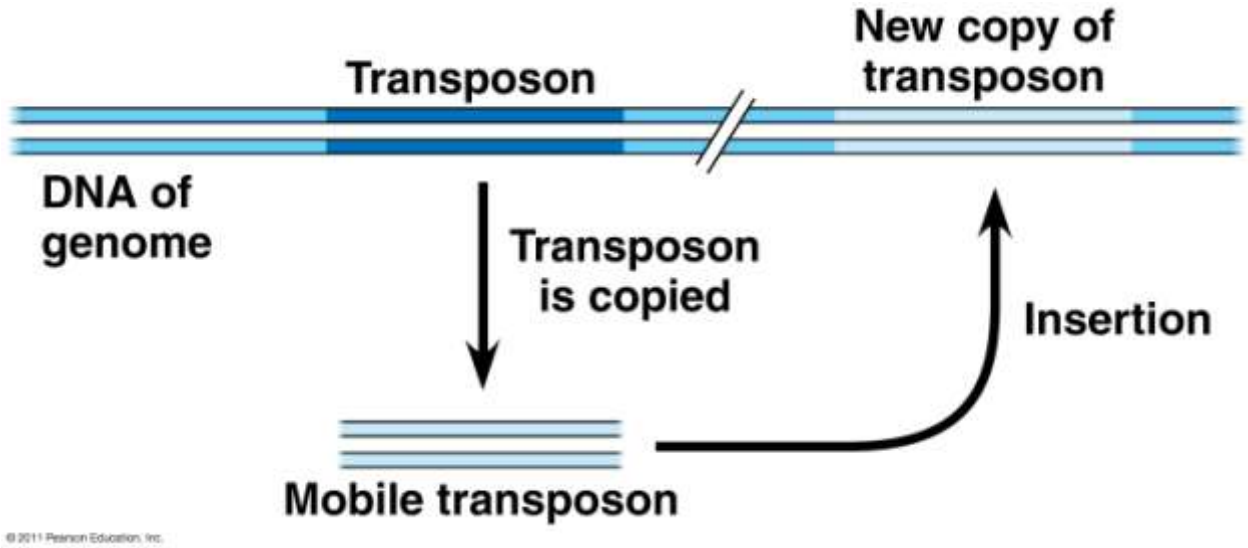
- ~8% of the genome
- Viral sequences in our genome that may have come from ancient viral infections
- The viral sequences somehow got mutated and became nonfunctional viral ‘fossils’ in our genome
  - These fossils are replicated during S phase along with everything else
  - One particular family (HERV-K) has been particularly active in the last couple hundred years
- HERVs are suspected to be a cause behind some autoimmune disorders, particularly multiple sclerosis

LINEs and SINEs

- ~60% of the genome
- Long INterspersed Elements
  - Repeated chunks of DNA
  - Code for the enzyme reverse transcriptase
  - Can duplicate themselves
- Short INterspersed Elements
  - Don’t have an RT gene, and rely on other elements to duplicate

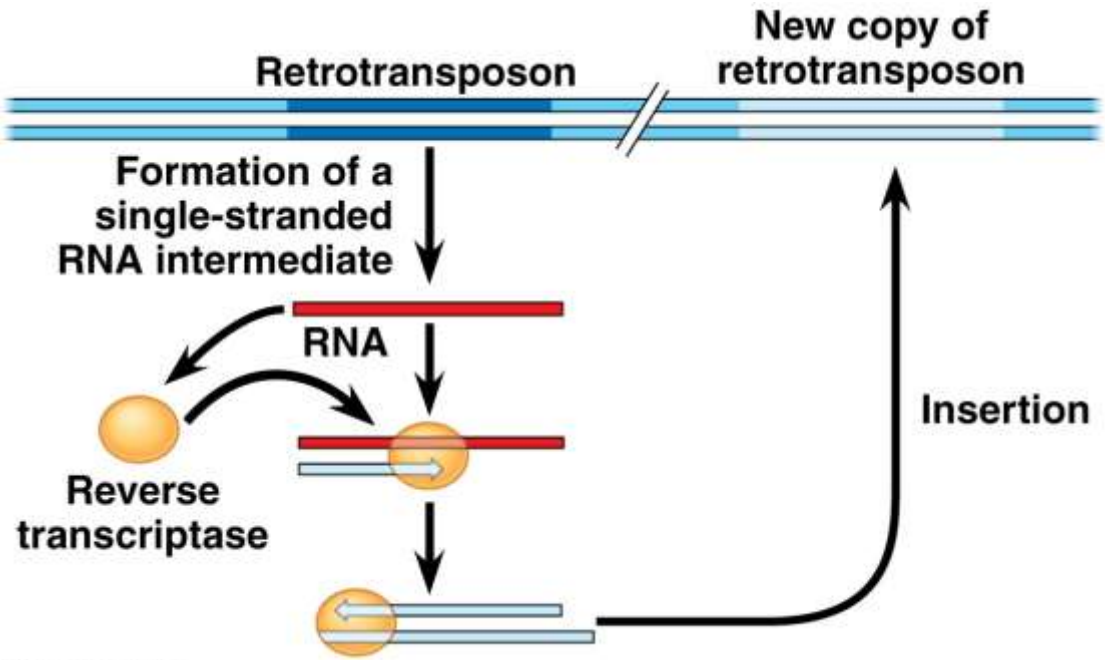
Transposons

- ~44% of the genome together with retrotransposons
- Mobile genetic elements that can replicate themselves – “jumping genes”
- Will randomly remove or copy themselves then reinsert in another locus



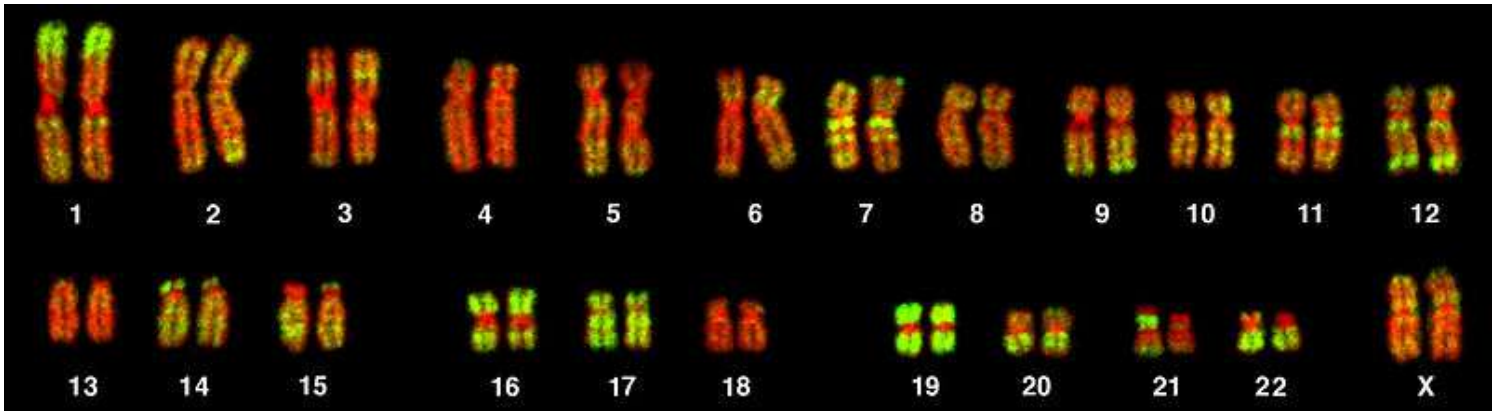
Retrotransposons

- Also mobile genetic elements
- "Jump" by transcribing themselves to RNA, then retrotranscribing back to DNA and inserting into the genome



ALU Elements

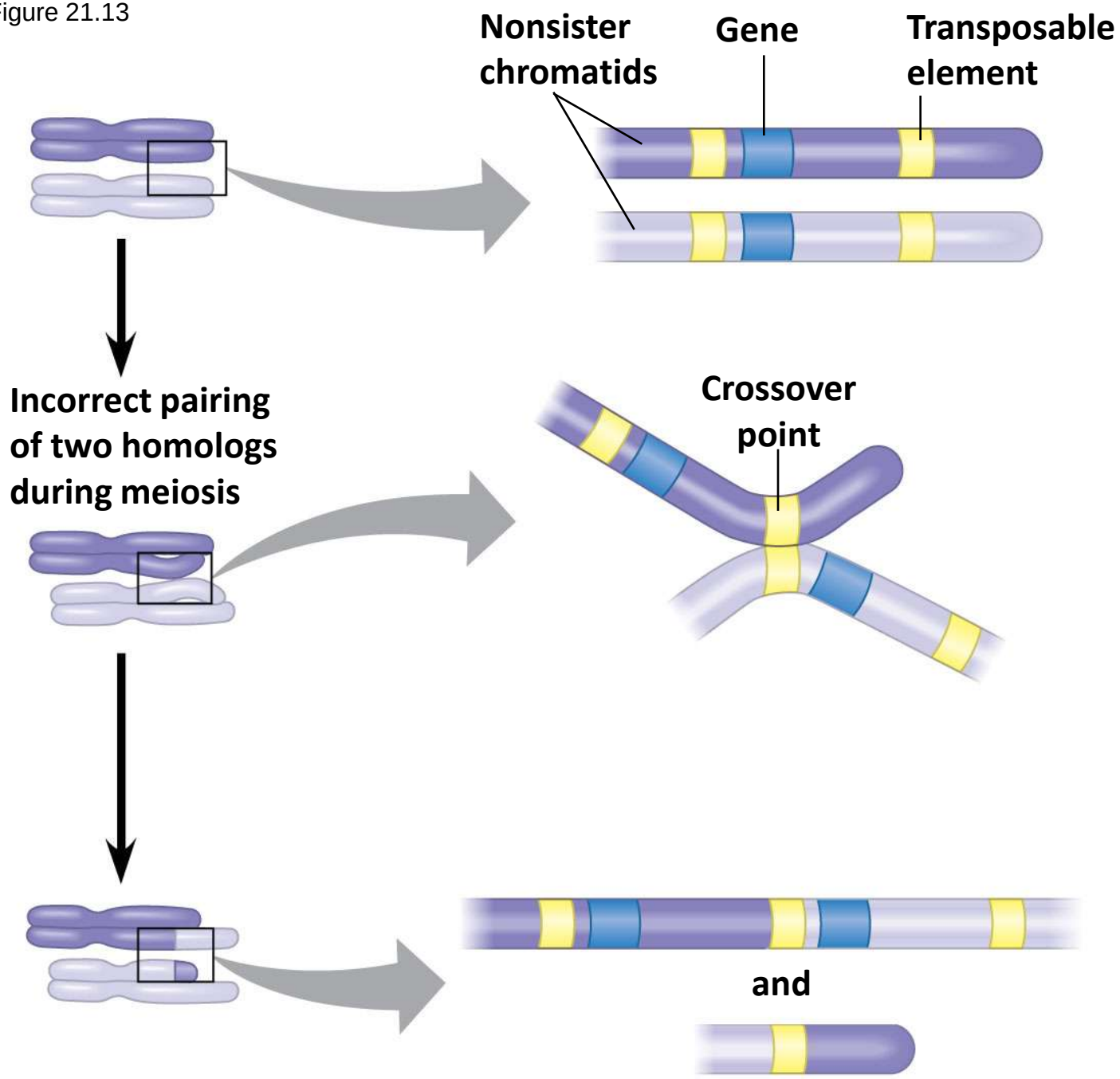
- 10% of the genome
- A family of repetitive sequences
- Most abundant retrotransposon in the human genome
- 300bp long, classified as SINEs
- Many *Alu* elements are transcribed into RNA molecules; however their function, if any, is unknown



Transposons and homologous recombination

- Synapse in homologous recombination occurs between two sites that have similar nucleotide sequences
- Unequal crossing over during prophase I of meiosis can result in one chromosome with a deletion and another with a duplication of a particular region
- Transposable elements (particularly Alu since it's present in such high copy numbers) can provide sites for crossover between nonsister chromatids

Figure 21.13



How transposons muck with proteins

- Multiple copies of similar transposable elements may facilitate recombination, or crossing over, between different chromosomes
- Insertion of transposable elements within a protein-coding sequence may block protein production
- Insertion of transposable elements within a regulatory sequence may increase or decrease protein production

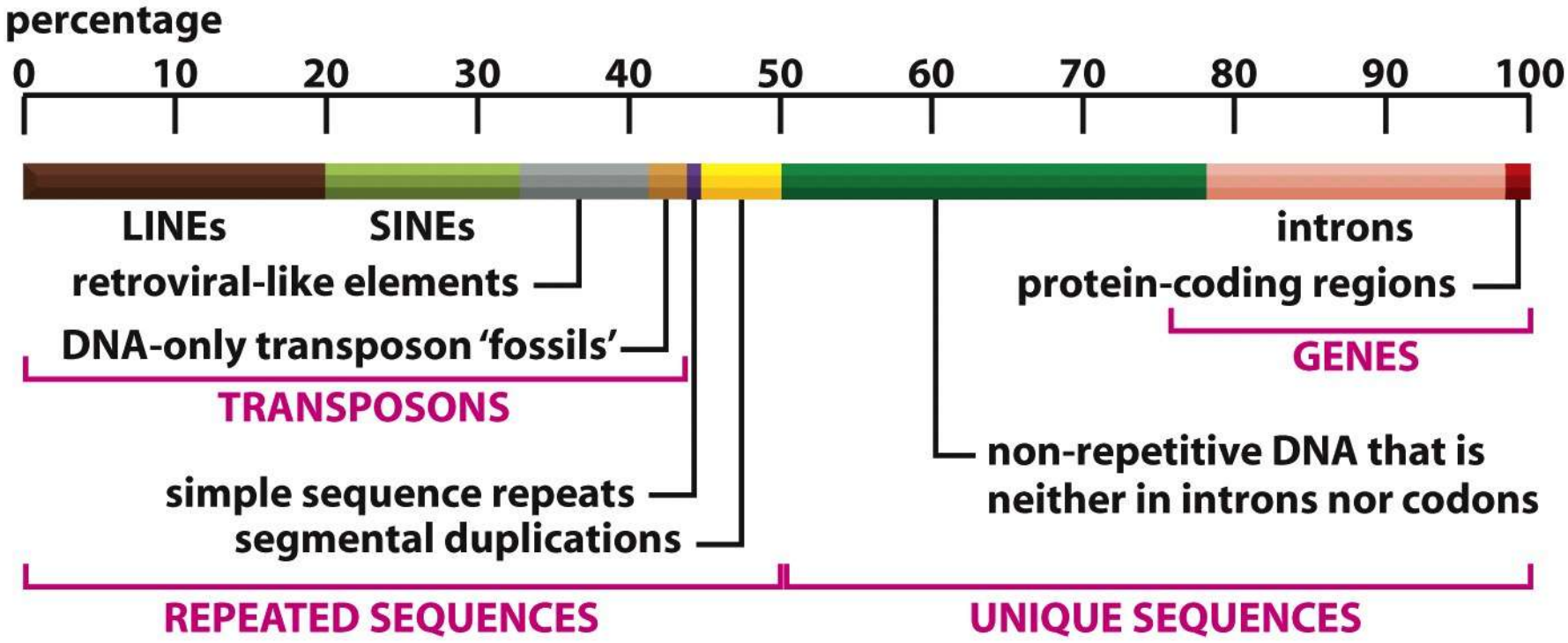
Transposons also help evolution

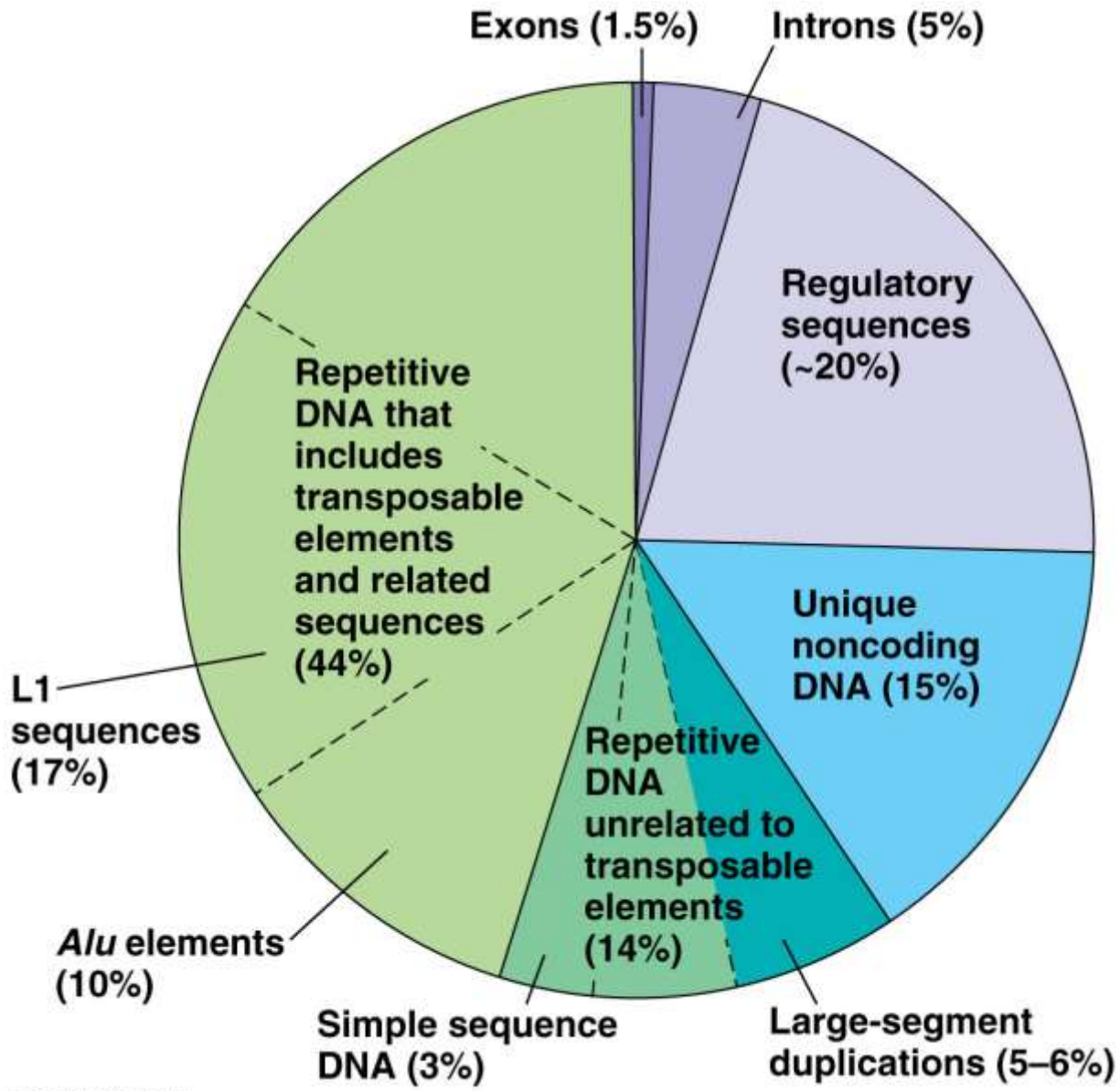
- Transposable elements may carry a gene or groups of genes to a new position
- They may also create new sites for alternative splicing in an RNA transcript
- Chromosome translocation is also facilitated by the presence of transposons

Microsatellites

- Short, 2-5bp repeats in the genome
- Common in centromere regions
- Each person has a different number of microsatellites
- Microsatellites are used as molecular markers in genetic studies, paternity tests, and crime scene analysis

Repeats in the genome





Additional Links

- Human genome shows proof of recent evolution
  - http://www.nytimes.com/2010/07/20/science/20adapt.html?pagewanted=all
- More on molecular clocks and how they're calculated
  - http://www.pbs.org/wgbh/evolution/library/05/1/pdf/l_051_06.pdf
- Venter's "We have learned nothing from the genome" interview
  - http://www.spiegel.de/international/world/0,1518,709174,00.html
- Full free publication of Celera's human genome data
  - http://www.sciencemag.org/content/291/5507/1304.full
- Free full publication of the HGP's human genome data
  - http://www.nature.com/nature/journal/v409/n6822/
- Free full publication of the synthetic life paper
  - http://www.sciencemag.org/content/329/5987/52

Vocabulary

- RNA world
- Ribozyme
- Out of Africa hypothesis
- Mitochondrial Eve
- Haplogroup, haplotype
- One gene, one protein hypothesis
- Exons, introns
- HERVs
- Pseudogenes
- LINEs, SINEs
- Transposons, retrotransposons
- ALU elements
- Microsatellites