The McGraw-Hill Companies

Lecture PowerPoint to accompany

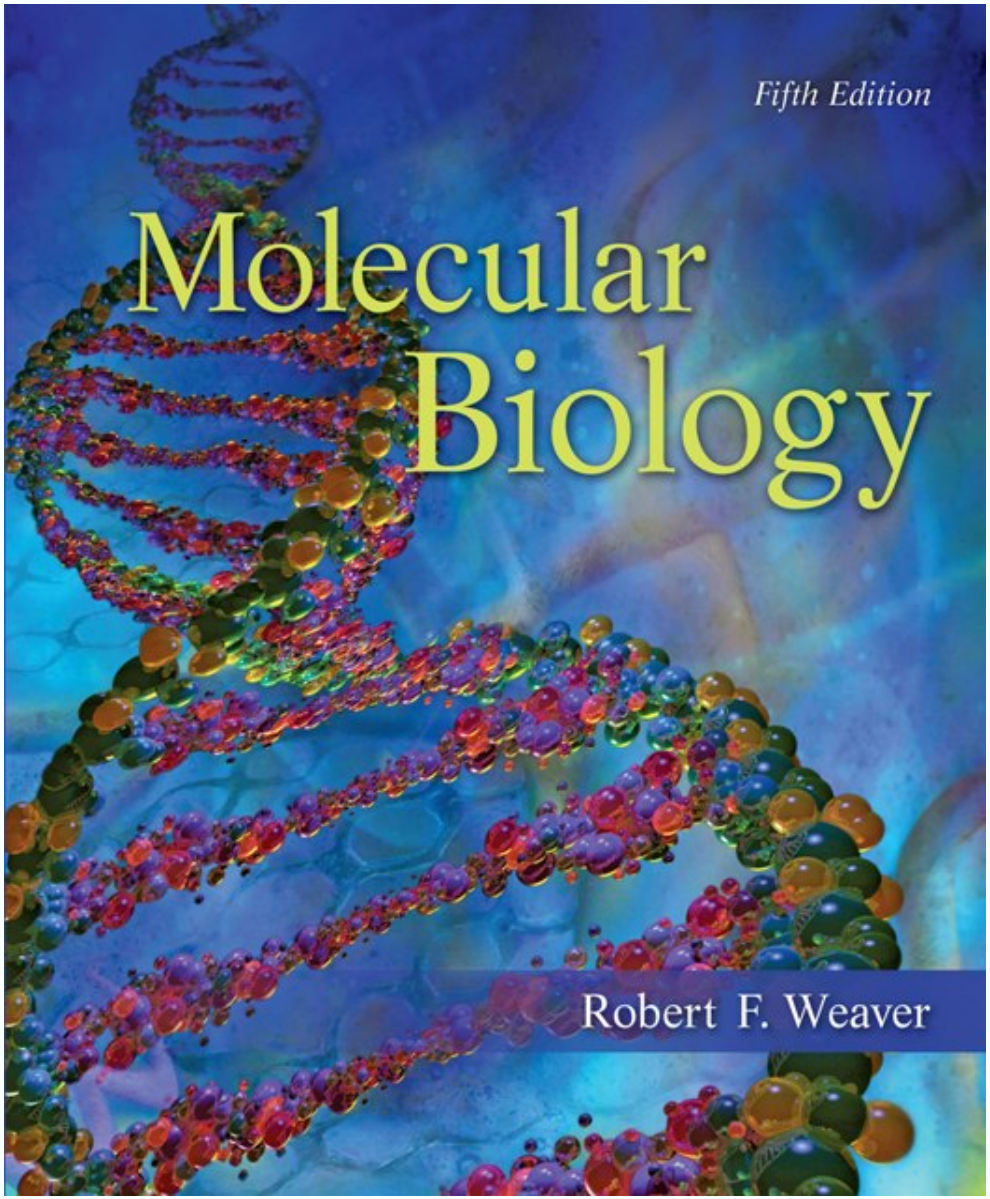
Molecular Biology

Fifth Edition

Robert F. Weaver

Chapter 24

Introduction to Genomics: DNA Sequencing on a Genomic Scale



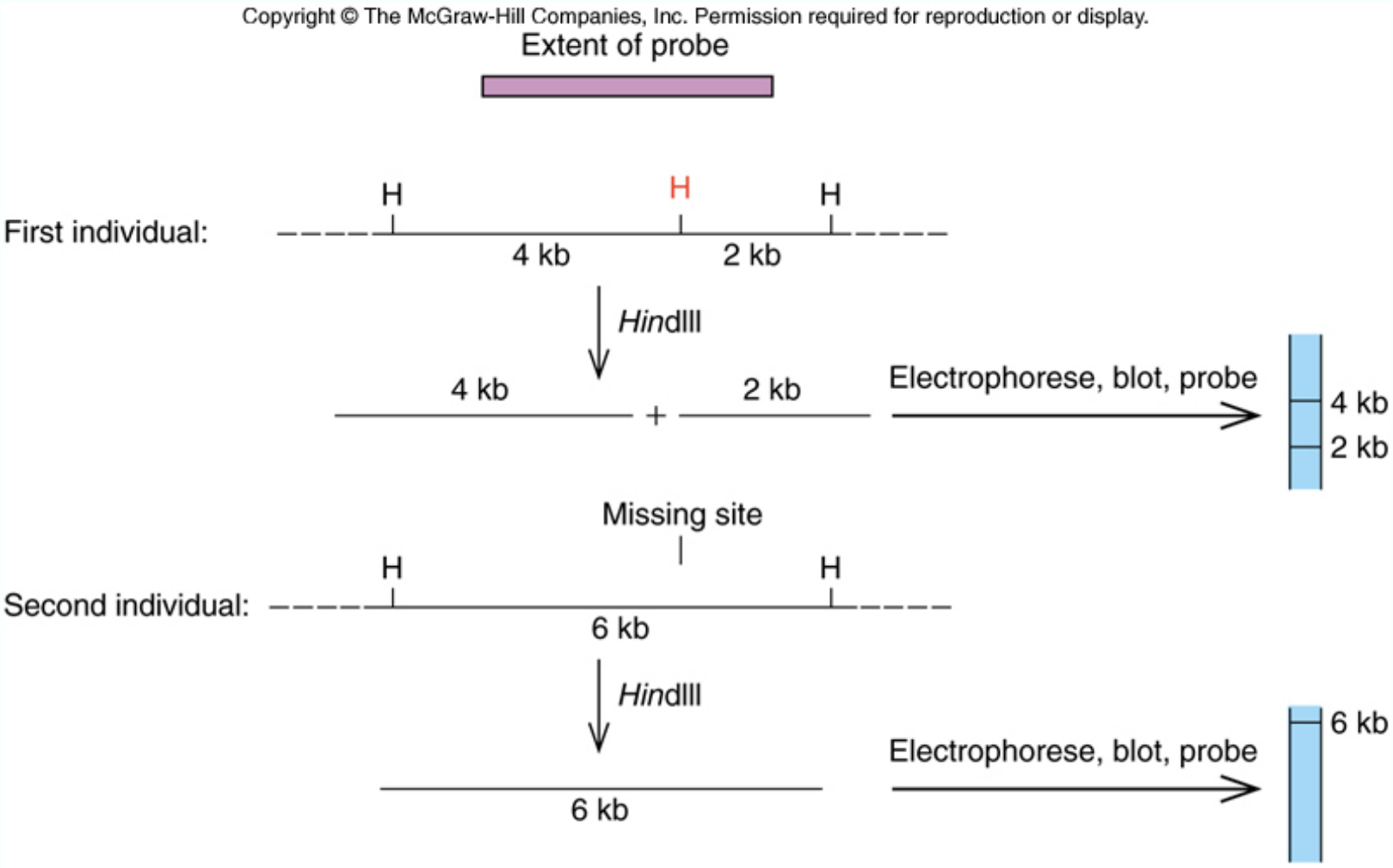
24.1 Positional Cloning

- Positional cloning is a method for the discovery of genes involved in genetic traits
- Positional cloning was very difficult in the absence of genomic information
- Begins with mapping studies to pin down the location of the gene of interest to a relatively small region of DNA

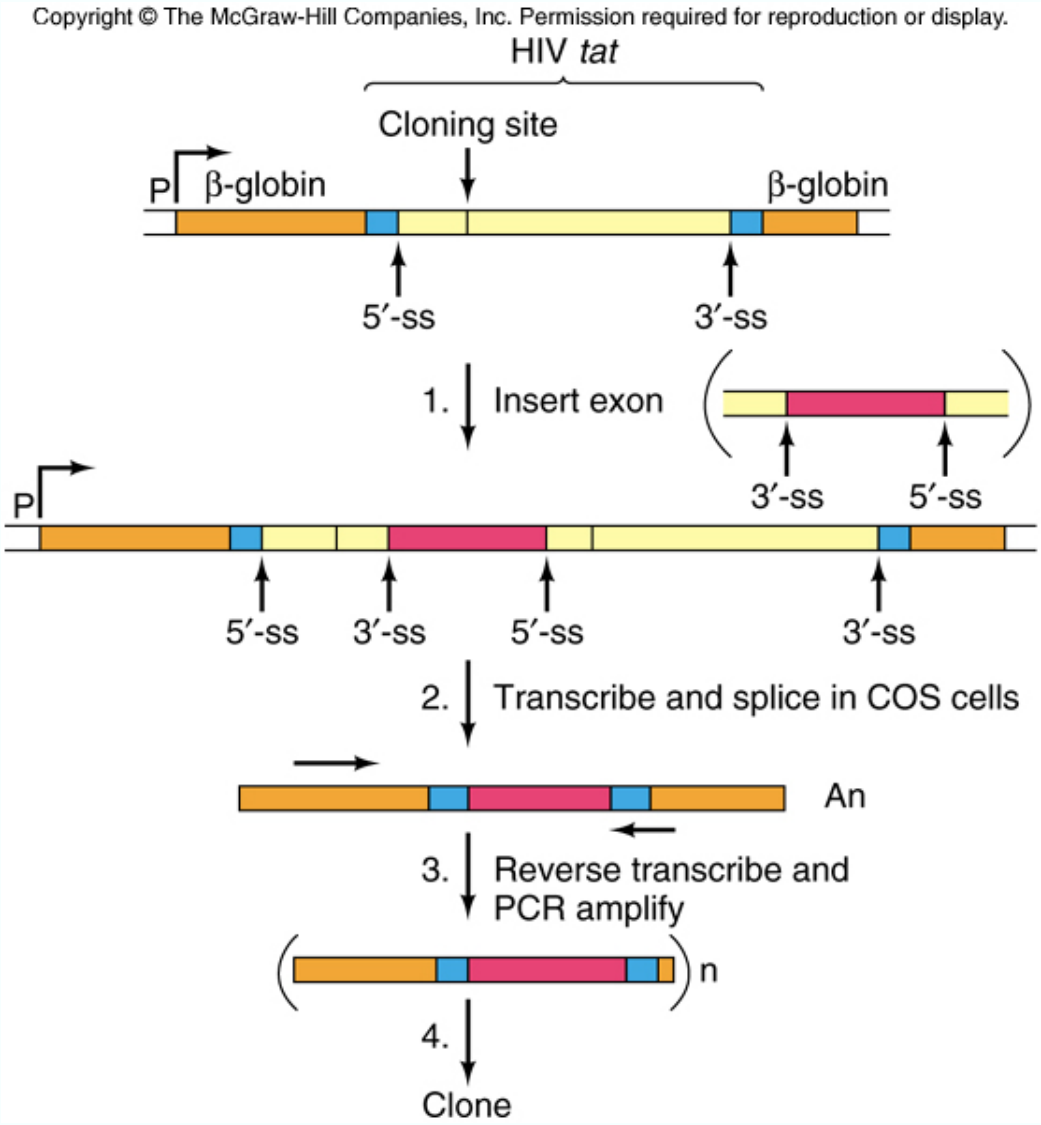
Classical Tools of Positional Cloning

- Mapping depends on a set of landmarks to which gene position can be related
- Restriction Fragment Length Polymorphisms (RFLP) are landmarks with lengths of restriction fragments given by a specific enzyme that vary from one individual to another
- Exon Traps use a special vector to help clone exons only
- CpG Islands are DNA regions containing unmethylated CpG sequences

Detecting RFLPs



Exon Trapping

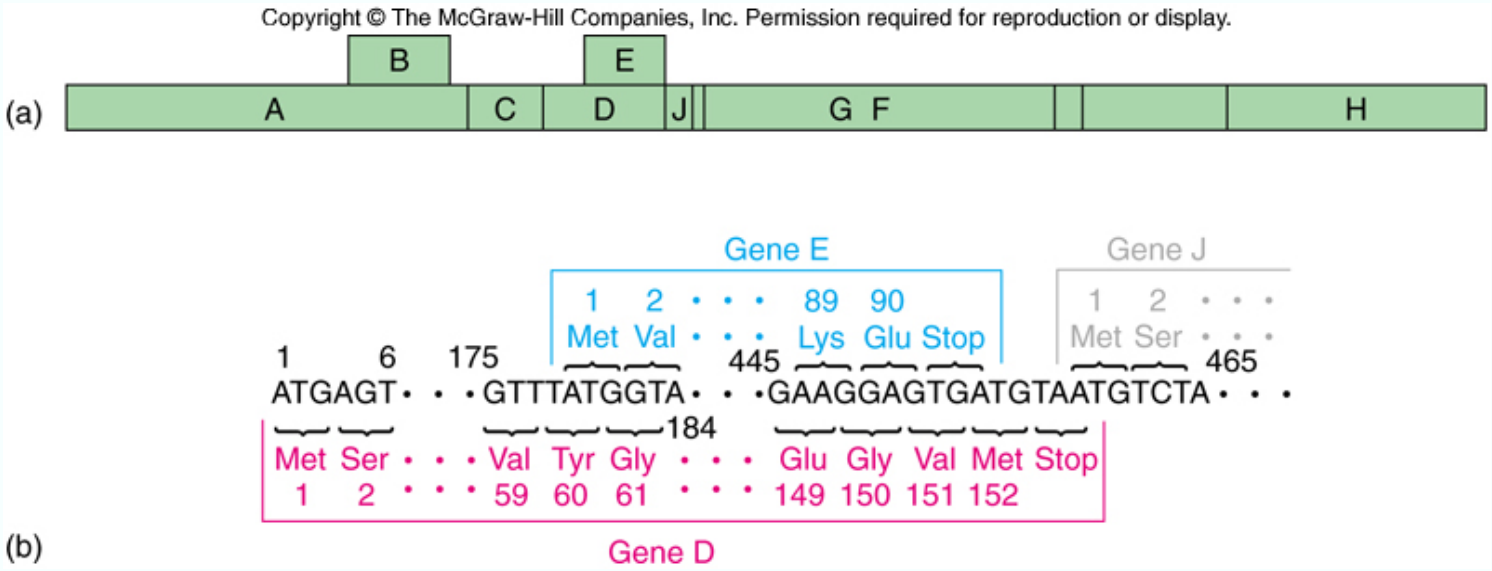


Identifying the Gene Mutated in a Human Disease

- Using RFLPs, geneticists mapped the Huntington disease gene (*HD*) to a region near the end of chromosome 4
- Used an exon trap to identify the gene itself
- Mutation causing the disease is an expansion of a CAG repeat from the normal range of 11-34 copies to abnormal range of at least 38 copies
- Extra repeats cause extra Glu inserted into huntingtin, the product of the *HD* gene

Phage ϕX174 Genome

- First genome sequenced was a very simple one, phage ϕX174
  - Completed by Sanger in 1977
  - 5375-nucleotides
- Note that some of these phage genes overlap



24.2 Techniques in Genomic Sequencing

- What information can be gleaned from genome sequence?
  - Location of exact coding regions for all the genes
  - Spatial relationships among all the genes and exact distances between them
- How is a coding region recognized?
  - Contains an ORF long enough to code for a phage protein
  - ORF must
    - Start with ATG triplet
    - End with stop codon
  - Phage or bacterial ORF is the same as a gene's coding region

Genome Results

- The base sequences of viruses and organisms that have been obtained range from:
  - Phages
  - Bacteria
  - Animals
  - Plants
- A rough draft and finished versions of the human genome have also been obtained
- Comparison of the genomes of closely related and more distantly related organisms can shed light on the evolution of these species

Sequencing Milestones

Table.24.1

Genome (Importance)	Size (bp)	Year
Phage ϕX174 (first genome)	5375	1977
Phage λ (large-DNA phage)	48,513	1983
Herpes simplex virus I (large-DNA eukaryotic virus)	152,260	1988
Haemophilus influenzae (bacterium, first organism)	1,830,000	1995
Mycoplasma genitalium (smallest bacterial genome)	580,000	1995
Saccharomyces cerevisiae (yeast, first eukaryote)	12,068,000	1996
Methanococcus jannaschii (first archaeon)	1,660,000	1996
Escherichia coli (best studied bacterium)	4,639,221	1997
Caenorhabditis elegans (first animal, roundworm)	97,000,000	1998
Human chromosome 22 (first human chromosome)	53,000,000	1999
Arabidopsis thaliana (first plant, mustard family)	120,000,000	2000
Drosophila melanogaster (a favorite genetic model)	180,000,000	2000
Human (working draft of the "holy grail" of genomics)	3,200,000,000	2001
Plasmodium falciparum (the malaria parasite)	23,000,000	2002
Anopheles gambiae (the major mosquito malaria carrier)	278,000,000	2002
Fugu rubripes (tiger pufferfish)	365,000,000	2002
Mus musculus (house mouse)	2,500,000,000	2002
Ciona intestinalis (sea squirt, a primitive chordate)	117,000,000	2002
Canis lupus familiaris (dog, working draft)	~2,400,000,000	2003
Gallus gallus (chicken, first farm animal)	1,050,000,000	2004
Human (finished sequence)	3,200,000,000	2004
Oryza sativa (rice, first cereal grain)	489,000,000	2005
Pan troglodytes (chimpanzee, our closest relative, working draft)	~3,000,000,000	2005
Three trypanosomatids (*Trypanosoma cruzi, T. brucei,* and *Leishmania major,* parasites that cause severe human illness)	25–55,000,000	2005
Populus trichocarpa (black cottonwood, first tree)	~485,000,000	2006
First individual humans (two Caucasians, one African, and one Han Chinese)	3,200,000,000	2007 and 2008
Homo Neanderthalensis (our closest evolutionary relative, working draft)	~3,000,000,000	2010

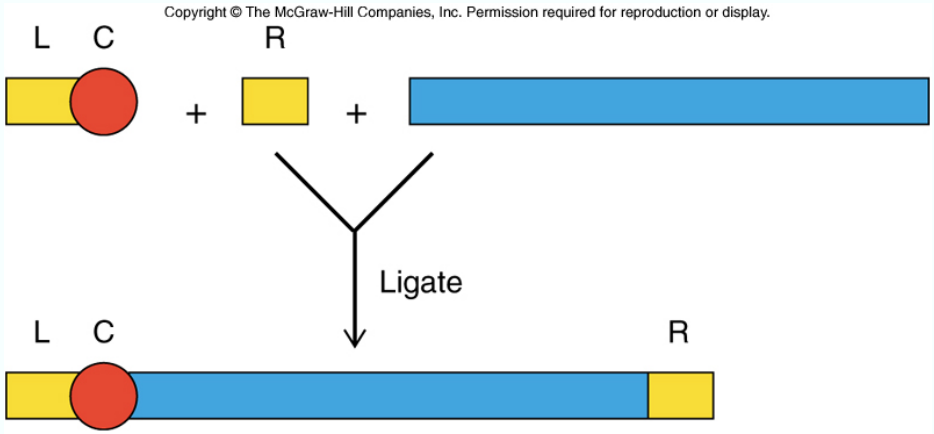
The Human Genome Project

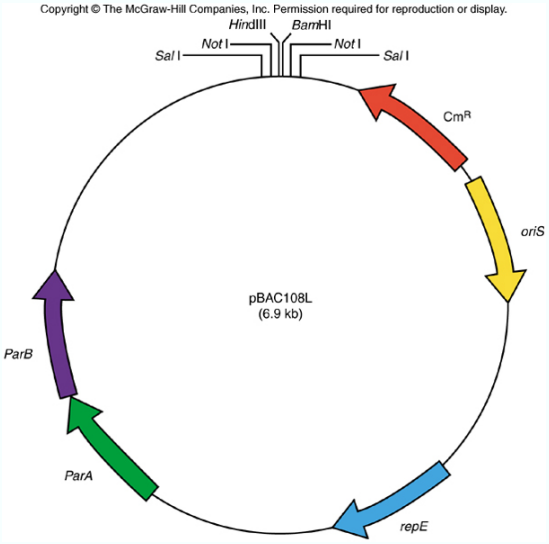
- In 1990, geneticists started to map and ultimately sequence the entire human genome
- Original plan was systematic and conservative
  - Prepare genetic and physical maps of genome with markers to allow piecing DNA sequences together in proper order
  - Most sequencing would be done only after mapping was complete

1998 – Human Genome Project

- Celera, a private, for-profit company, shocked genomic community by announcing Celera would complete a rough draft of human genome by 2000
- Method that would be used was shotgun sequencing, whole human genome would be chopped up and cloned
  - Clones sequenced randomly
  - Sequences would be pieced together using computer programs

Vectors for Large-Scale Genome Projects





- Two high-capacity vectors have been used extensively in the Human Genome Project
  - Mapping was done mostly using the yeast artificial chromosome, accepts million base pairs
  - Sequencing with bacterial artificial chromosomes accepting about 300,000 bp
- BACs are more stable, easier to work with than YACs

The Clone-by-Clone Strategy

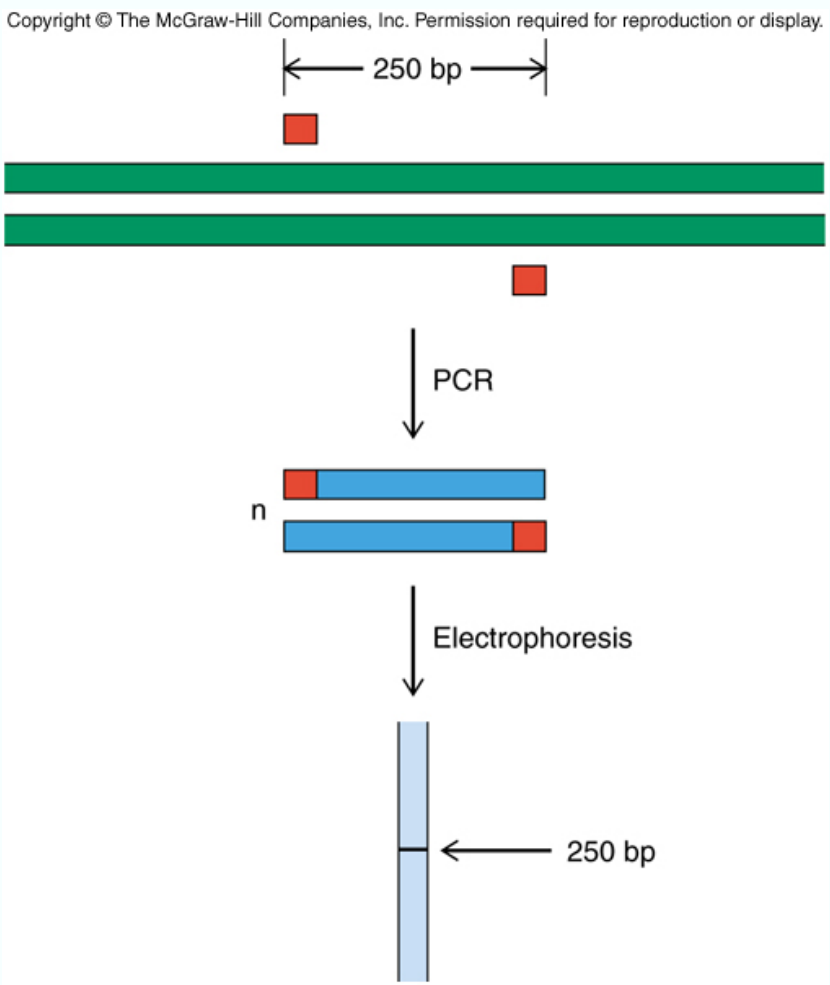
- Mapping the human genome requires a set of landmarks to which we can relate the positions of genes
- Some of these markers are genes, many more are nameless stretches of DNA
  - RFLPs
  - VNTRs, variable number tandem repeats
  - STSs, sequence-tagged sites, expressed-sequence tags (ESTs) and microsatellites

Variable Number Tandem Repeats (VNTRs)

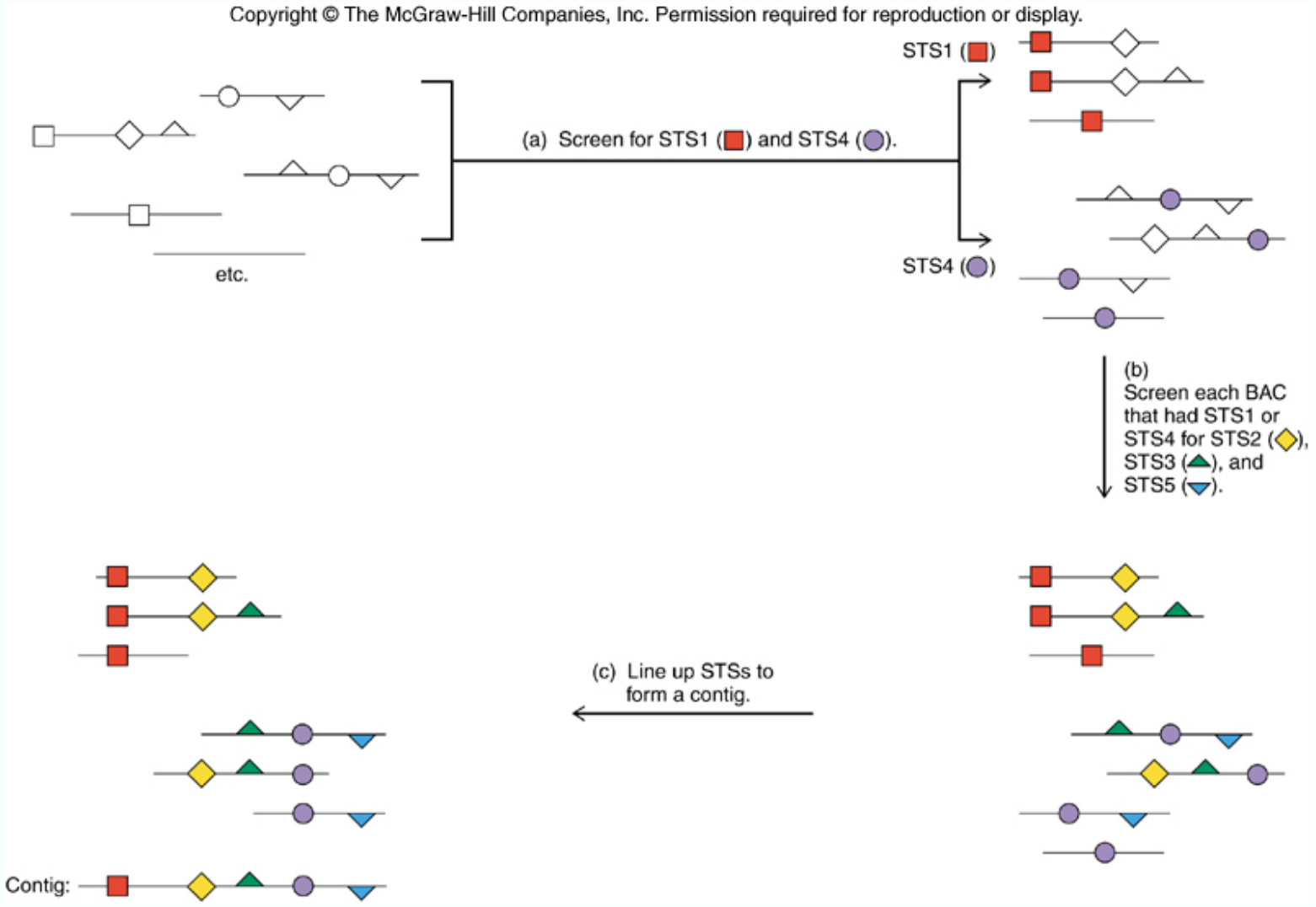
- VNTRs derive from minisatellites, stretches of DNA that contain a short core sequence repeated over and over in tandem (head to tail)
- The number of repeats of the core sequence in a VNTR is likely to be different from one individual to another
  - So VNTRs are highly polymorphic
  - This makes them relatively easy to map
  - Disadvantage as genetic markers as they tend to bunch together at chromosome ends

Sequence-Tagged Sites (STSs)

- STSs are short sequences
  - 60-1000 bp long
  - Detectable by PCR
- Can design short primers
  - Hybridize few hundred bp apart
  - Amplify a predictable length of DNA



Sequence-Tagged Sites Mapping



Microsatellites

- STSs are very useful in physical mapping or locating specific sequences in the genome
  - Worthless as markers in traditional genetic mapping unless polymorphic
- Microsatellites are a class of STSs that are highly polymorphic
  - Similar to minisatellites
  - Consist of a core sequence repeated over and over many times in a row
  - Core here is 2-4 bp long, much shorter

Contig

- A set of clones used by geneticists in physically mapping or sequencing a given region is called a contig
  - Contains contiguous (or overlapping) DNAs spanning long distances
  - Used like putting together a jigsaw puzzle
  - Easier to complete with bigger pieces
  - Helpful to assemble in overlapping fashion

Shotgun Sequencing

Massive sequencing projects can take two forms:

1. Map-then-sequence strategy
   - Produces physical map of genome including STSs
   - Sequences clones (mostly BACs) used in mapping
   - Places sequences in order to be pieced together
2. In the shotgun approach
   - Assembles libraries of clones with different size inserts
   - Sequences the inserts at random
   - Relies on computer program to find areas of overlap among sequences and piece them together

Sequencing Standards

- A “working draft” may be:
  - Only 90% complete
  - Error rate of up to 1%
- A “final draft” (less consensus):
  - Error rate of less than 0.01%
  - Should have as few gaps as possible
- Some researchers require a “final draft” is not completely sequenced until every last gap is completed

24.3 Studying and Comparing Genomic Sequences

- Once a genomic sequence is in hand, scientists can mine it for the wealth of information it contains and compare it to the sequences of other genomes to shed light on the evolution of the species

The Human Genome

- First chromosome completed in the Human Genome Project was chromosome 22 in late 1999
- In February 2001, the Venter group and the public consortium each published their versions of a working draft of the whole human genome

Chromosome 22

- Only the long arm (22q) was sequenced
- Short arm (22p) is composed of pure heterochromatin, likely devoid of genes
- 11 gaps remained in the sequence
  - 10 are gaps between contigs likely due to “unclonable” DNA
  - Other a 1.5-kb region of cloned DNA that resisted sequencing

Findings from Chromosome 22

1. We must learn to live with gaps in our sequence
2. 679 annotated genes categorized as:
   - 274 Known genes, previously identified
   - 150 Related genes, homologous to known genes
   - 148 Predicted genes, sequence homology to ESTs
   - 134 Pseudogenes, sequences are homologous to known genes, but contain defects that preclude proper expression

Chromosome 22 contigs and gaps

Table 24.2 Chromosome 22 Contigs and Gaps

Contig	Gap	Size (kb)	
1		234	
	1		1.9
2		406	
	2		≈150
3		1394	
	3		≈150
4		1790	
	4		≈100
5		23,006	
	5		≈50
6		767	
	6		≈50–100
7		1528	
	7		≈150
8		2485	
	8		≈50
9		190	
	9		≈100
10		993	
	10		≈100
11		291	
	11		≈100
12		380	
Total sequence length		33,464	
Total length of 22q		34,491	

(*Source:* Adapted from Dunham, I., N. Shimizu, B.A. Roe, S. Chissoe, A.R. Hunt, J.E. Collins, et al., The DNA sequence of human chromosome 22. *Nature* 402:491, 1999.)

More From Chromosome 22

3. Coding regions of genes account for only tiny fraction of length of the chromosome
   - Annotated genes are 39% of total length
   - Exons are only 3%
   - Repeat sequences (Alu, LINEs, etc) are 41%
4. Rate of recombination varies across the chromosome
   - Long regions of low recombination interspersed with short regions where it is relatively frequent

Repetitive DNA content of chromosome 22

Table 24.3 Repetitive DNA Content of Human Chromosome 22

Type	Number	Total base pairs	% of chromosome
Alu	20,188	5,621,998	16.80
HERV	255	160,697	0.48
LINE 1	8043	3,256,913	9.73
LINE 2	6381	1,273,571	3.81
LTR	848	256,412	0.77
MER	3757	763,390	2.28
MIR	8426	1,063,419	3.18
MLT	2483	605,813	1.81
THE	304	93,159	0.28
Other	2313	625,562	1.87
Dinucleotide	1775	133,765	0.40
Trinucleotide	166	18,410	0.06
Tetranucleotide	404	47,691	0.14
Pentanucleotide	16	1612	0.0048
Other tandem repeats	305	102,245	0.31
Total	55,664	14,024,657	41.91

(*Source:* Adapted from Dunham, I., N. Shimizu, B.A. Roe, S. Chissoe, A.R. Hunt, J.E. Collins, et al., The DNA sequence of human chromosome 22. *Nature* 402:491, 1999.)

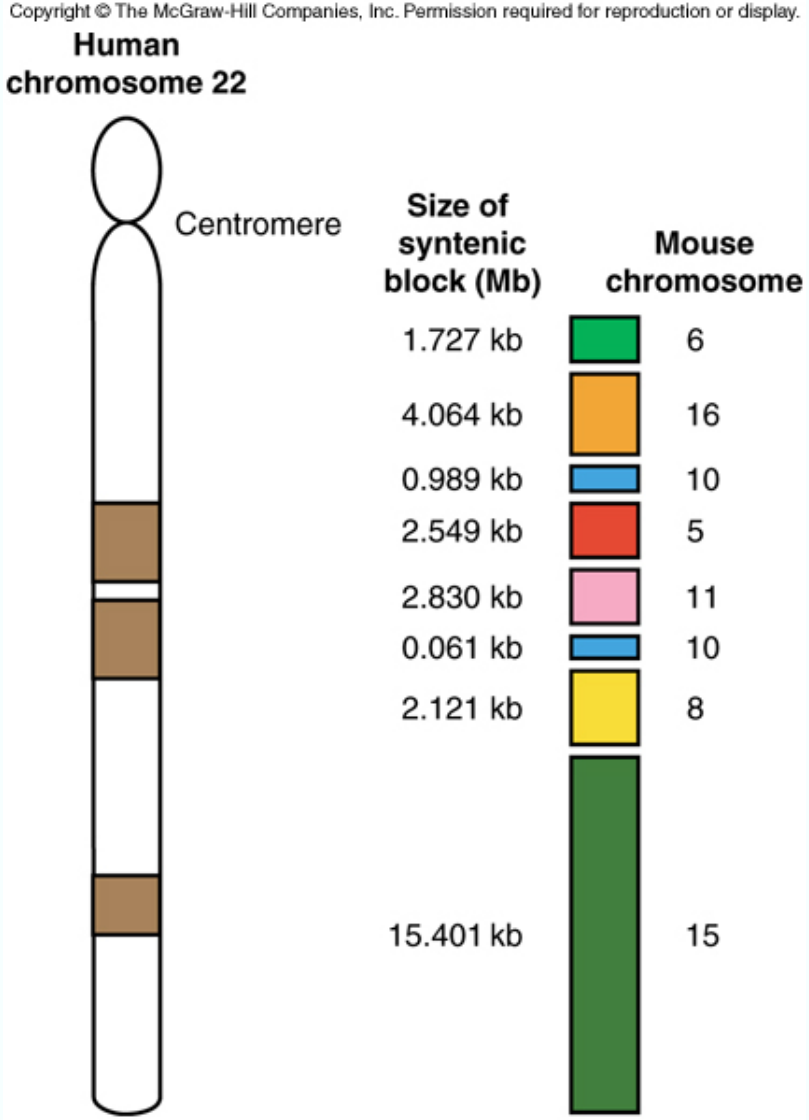
More From Chromosome 22

5. There are local and long-range duplications
   - Immunoglobin λ locus
   - 36 gene segments are clustered together that can encode variable regions
     - 60-kb region is duplicated with greater than 90% fidelity almost 12 Mb away
     - Duplications found in few copies, low-copy repeats
6. Large chunks of human chromosome 22q are conserved in several different mouse chromosomes
   - 113 human genes with mouse orthologs mapped to mouse chromosomes

Homologs

- Orthologs are homologous genes in different species that evolved from a common ancestor
  - 8 regions on 7 mouse chromosomes
- Paralogs are homologous genes that evolved by gene duplication within a species
- Homologs are any kind of homologous genes, both orthologs and paralogs

Regions of conservation between human and mouse chromosome 22



Human Genome Project Status

- Working draft of human genome reported by 2 groups allowed estimates that genome contains fewer genes than anticipated – 25,000 to 40,000
- About half the genome has derived from the action of transposons
- Transposons themselves have contributed dozens of genes to the genome
- Bacteria also have donated dozens of genes
- Finished draft is much more accurate than working draft, but there are still gaps
- Information also about gene birth and death during human evolution

Other Vertebrate Genomes

- Comparing human genome with that of other vertebrates has taught us much about similarities and differences among genomes
  - Comparison has also helped to identify many human genes
  - In future, will likely help identify defective genes involved in human genetic diseases
- Closely related species like mouse can be used to find when and where genes are expressed to predict when and where human genes are likely expressed

Other Vertebrate Genomes

- Comparison of the genomes of human and our closest living relative, the chimpanzee, have identified a few DNA regions that have changed rapidly since the two species diverged
- These are good candidates for the DNA sequences that set humans and chimpanzees apart, yet very few of them are in protein-encoding genes
- Thus, the thing that really sets us apart may be the control of genes, rather than the genes themselves

The Minimal Genome

- It is possible to define the essential gene set of a simple organism
  - Mutate one gene at a time
  - See which genes are required for life
- In theory, also possible to define the minimal genome= set of genes that is minimum required for life
  - Minimum genome likely larger than the essential gene set
- In principle, possible to place minimal genome into a cell lacking genes of its own, create a new life form that can live and reproduce under lab conditions

“Synthetic biology”

- In 2007, Venter and colleagues had reported progress in the realm of “synthetic biology”
- They transplanted the genome of *Mycoplasma mycoides* to another bacterium, *Mycoplasma capricolum*, and through creative manipulations that made the transplant work, the resulting cell thrived

The Barcode of Life

- There is a movement which has begun to create a barcode to identify any species of life on earth
- The first such barcode will consist of the sequence of a 648-bp piece of mitochondrial COI gene from each organism
- This sequence is sufficient to identify uniquely almost any organism
- Other sequences will be worked out for plants and perhaps later for bacteria