BI 8
LECTURE 1

THE GENOME: CODES, ORGANIZATION, EXPRESSION VS. INHERITANCE

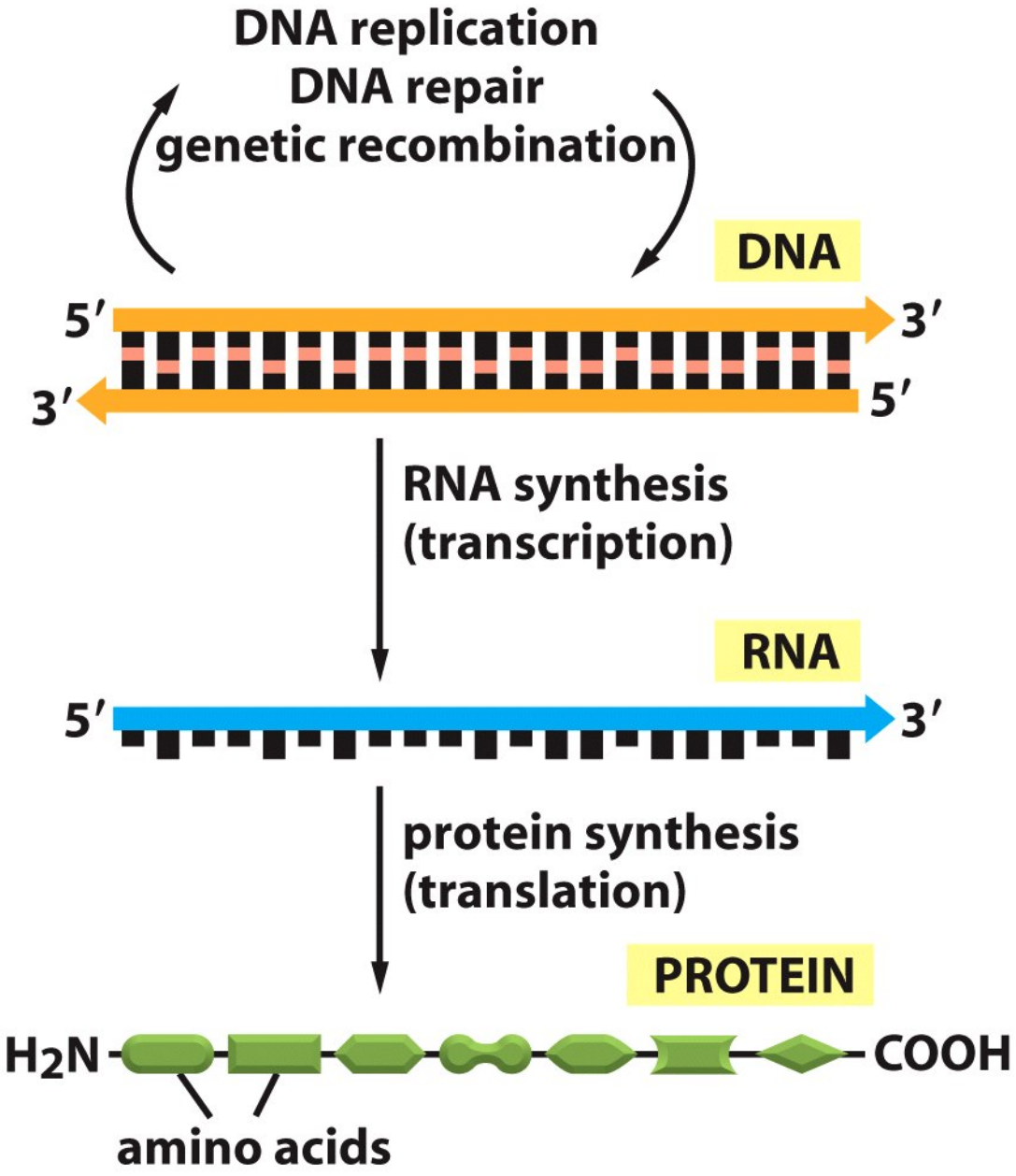
Ellen Rothenberg
5 January 2016

Kinds of codes incorporated in DNA

- Codes as blueprints for proteins (static codes)

- Codes as software to control conditional expression of coding genes (dynamic codes)
  - Conditions needed to express a given gene
  - Conditions needed to express the regulators of a given gene
  - ➢ Logic structure for defining cell type identities and developmental programs to produce them

A starting point…



DNA is the inherited template for protein coding, via an RNA messenger (“the central dogma”)

Figure 6-2 *Molecular Biology of the Cell* (© Garland Science 2008)

Reading to complement lectures

Today's lecture:

Alberts et al 6th edition *(though Alberts et al have a different game plan than this course!)*

- Chapter 1
- Ch.4: pp. 173-186 & 216-236

Thursday's lecture:

Alberts et al 6th edition

- Ch.4: pp. 173-186 & 216-236
- Ch. 6 pp. 299-301
- panel 2-6 (see panels 2-1, 2-2, 2-3 for background)

DNA is made of four subunits that pair in a restricted way: the basis of faithful DNA replication

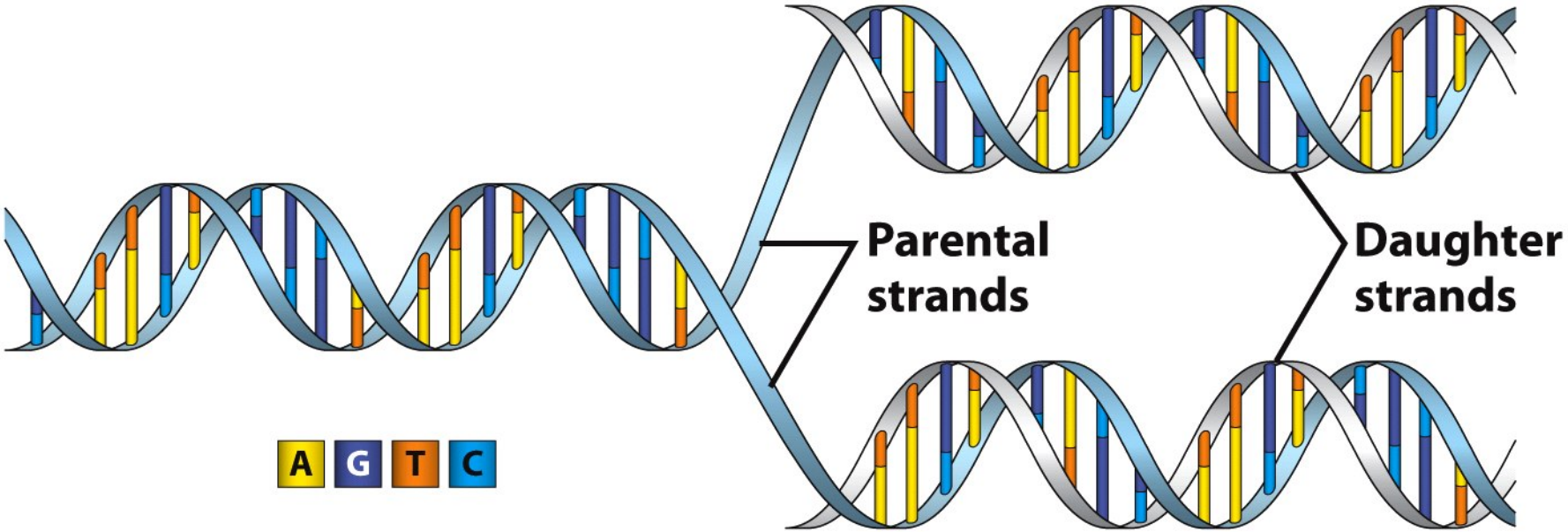


Figure 1-10
Molecular Cell Biology, Sixth Edition
© 2008 W. H. Freeman and Company

Antiparallel backbones and two kinds of complementary base pairs

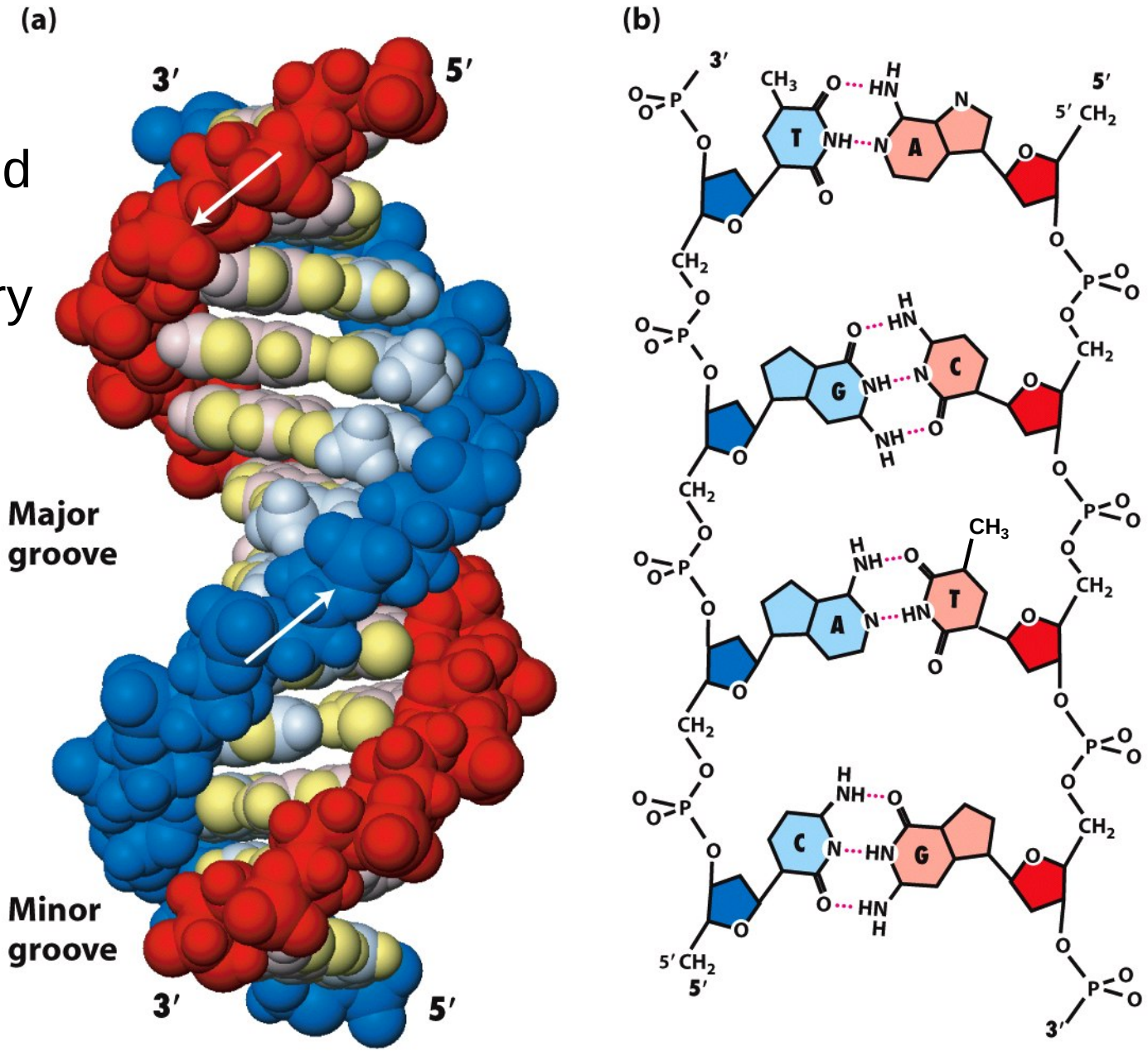


Figure 4-3
Molecular Cell Biology, Sixth Edition
© 2008 W. H. Freeman and Company

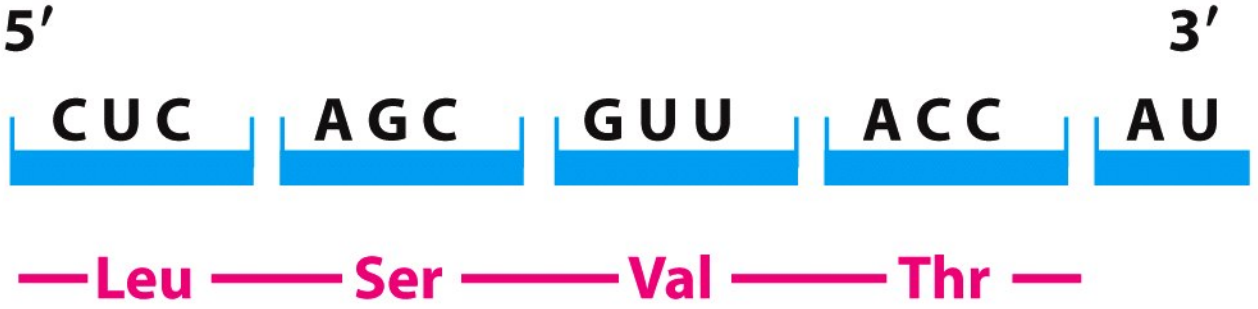
Reading genome sequence 1: the triplet code for protein coding

TABLE 4-1 The Genetic Code (Codons to Amino Acids)*

FIRST POSITION (5′ END)	SECOND POSITION: U	C	A	G	THIRD POSITION (3′ END)
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 (Met)*	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met (Start)	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val (Met)*	Ala	Glu	Gly	G

*AUG is the most common initiator codon; GUG usually codes for valine and CUG for leucine, but, rarely, these codons can also code for methionine to initiate a protein chain.

Table 4-1
Molecular Cell Biology, Sixth Edition
© 2008 W. H. Freeman and Company



An RNA sequence can directly be “read” into protein sequence: by you or by the cell

This code is universal, bacteria to us

Figure 6-51 *Molecular Biology of the Cell* (© Garland Science 2008)

Technological progress has made it possible to sequence massive amounts of DNA quickly

- Combinations of techniques
- Artificial copying of unknown DNA with labeled components
- Generation of large numbers of partial sequences
- Alignment computationally (major role for computation in bioinformatics!)
- “The $1000 personal genome”

But… knowing where to start decoding matters…

same RNA sequence encodes *different* proteins in different "reading frames"

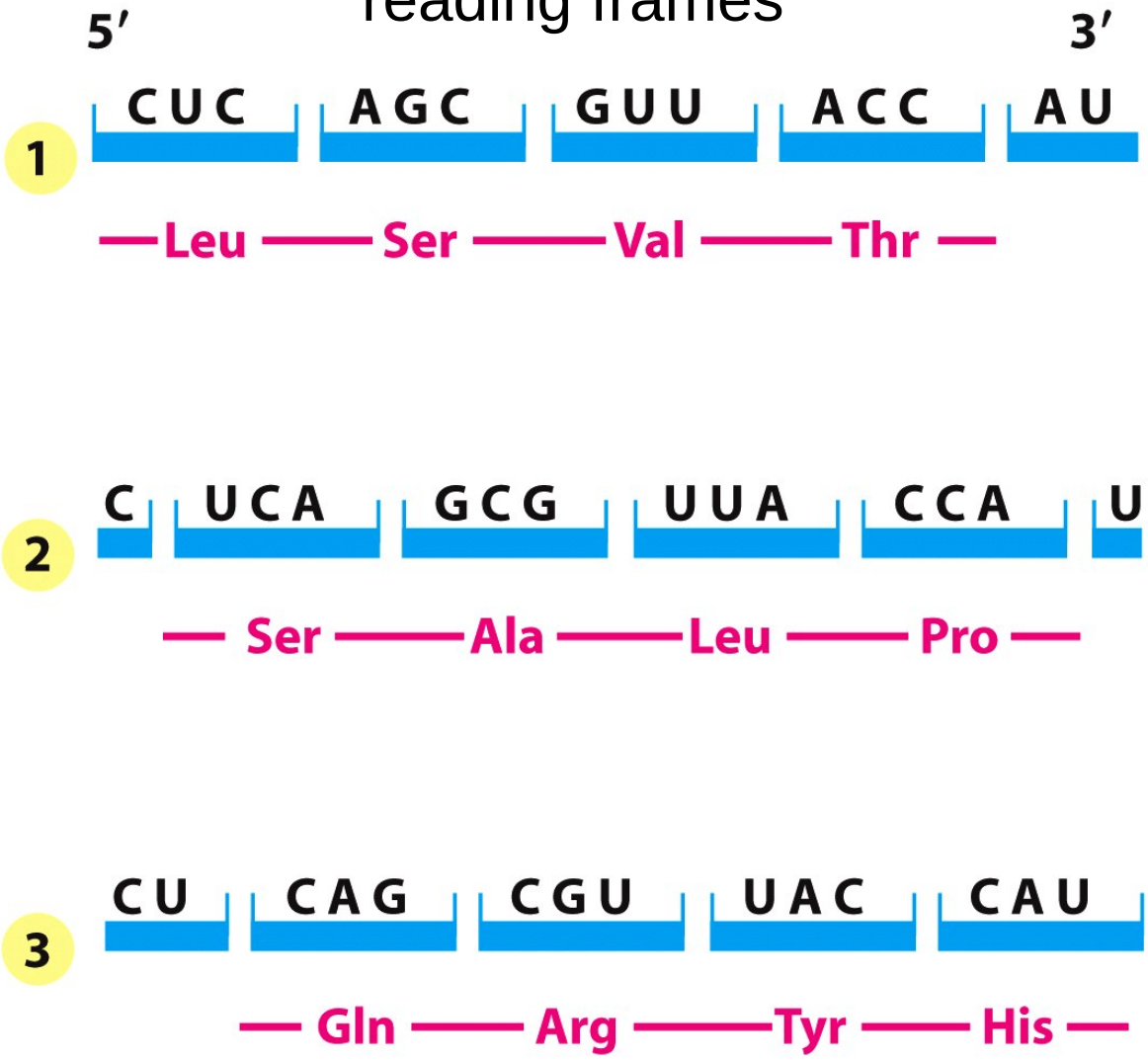


Figure 6-51 *Molecular Biology of the Cell* (© Garland Science 2008)

Protein coding sequence ≠ RNA sequence, and RNA sequence ≠ gene sequence in genomic DNA

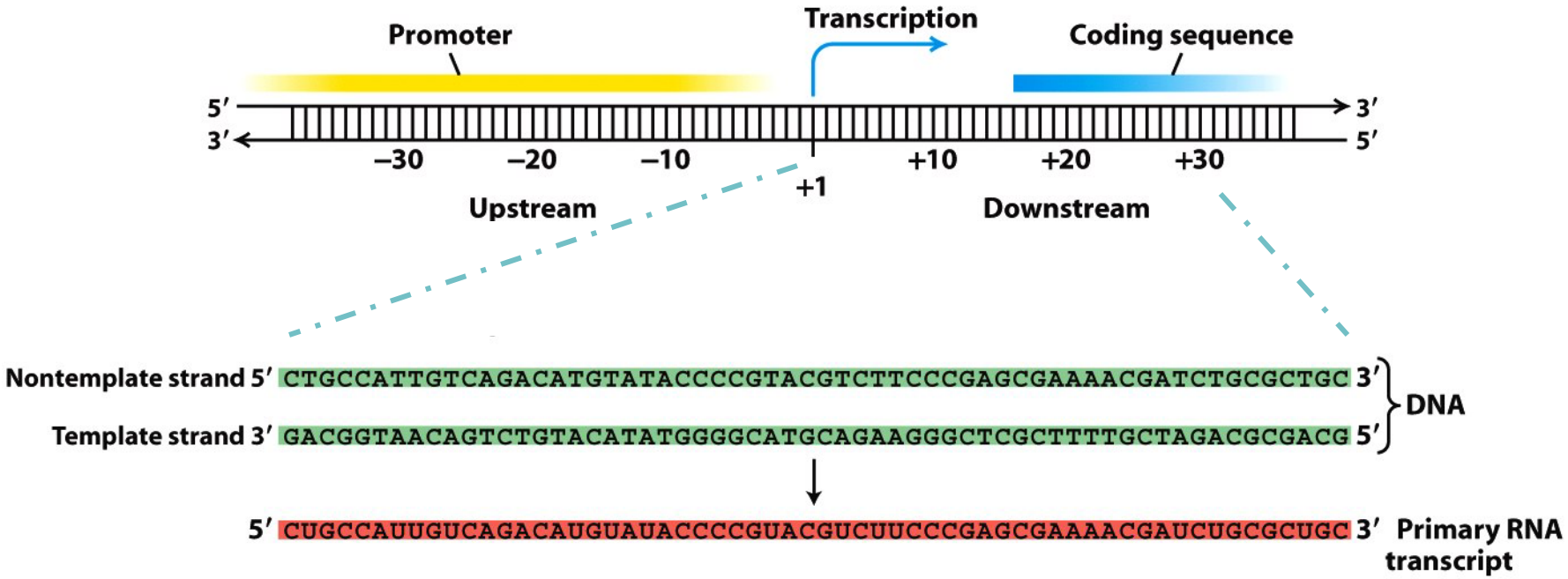


Figure 4-10b
Molecular Cell Biology, Sixth Edition
© 2008 W. H. Freeman and Company

Met Tyr Thr Pro Tyr Val Phe Pro Ser Glu Asn Asp Leu Arg Cys…

- DNA two complementary strands, RNA is just one strand
- Unless you know which strand is used to make RNA, 6 different possible reading frames per DNA sequence!
- DNA of a “gene” also includes start and stop signals for RNA synthesis (and a lot more)
- RNA includes start and stop signals for protein synthesis

Protein coding sequence ≠ RNA sequence, and RNA sequence ≠ gene sequence in genomic DNA

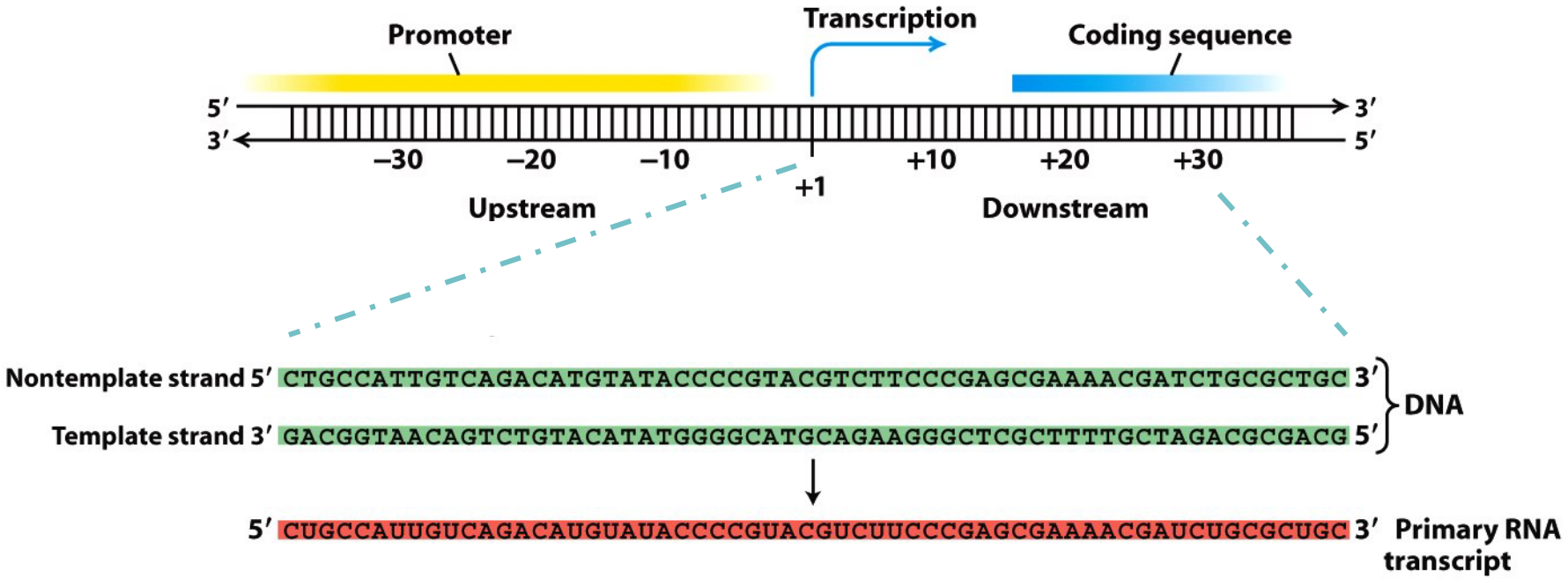


Figure 4-10b
Molecular Cell Biology, Sixth Edition
© 2008 W. H. Freeman and Company

Met Tyr Thr Pro Tyr Val Phe Pro Ser Glu Asn Asp Leu Arg Cys…

- To make a protein-coding RNA from a gene, only one strand needs to be copied from DNA
- The DNA strand with the sequence matching mRNA is “sense” strand… but it is not the template from which RNA is made
- Template is actually the *complement* of RNA: “antisense”

In prokaryotes (bacteria and archaea), this would be all.

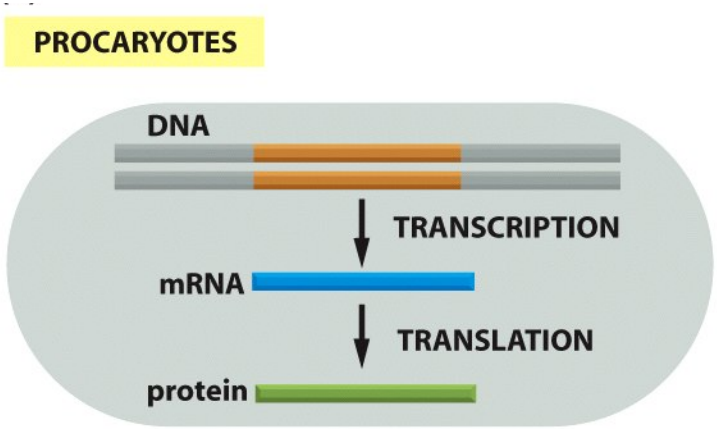
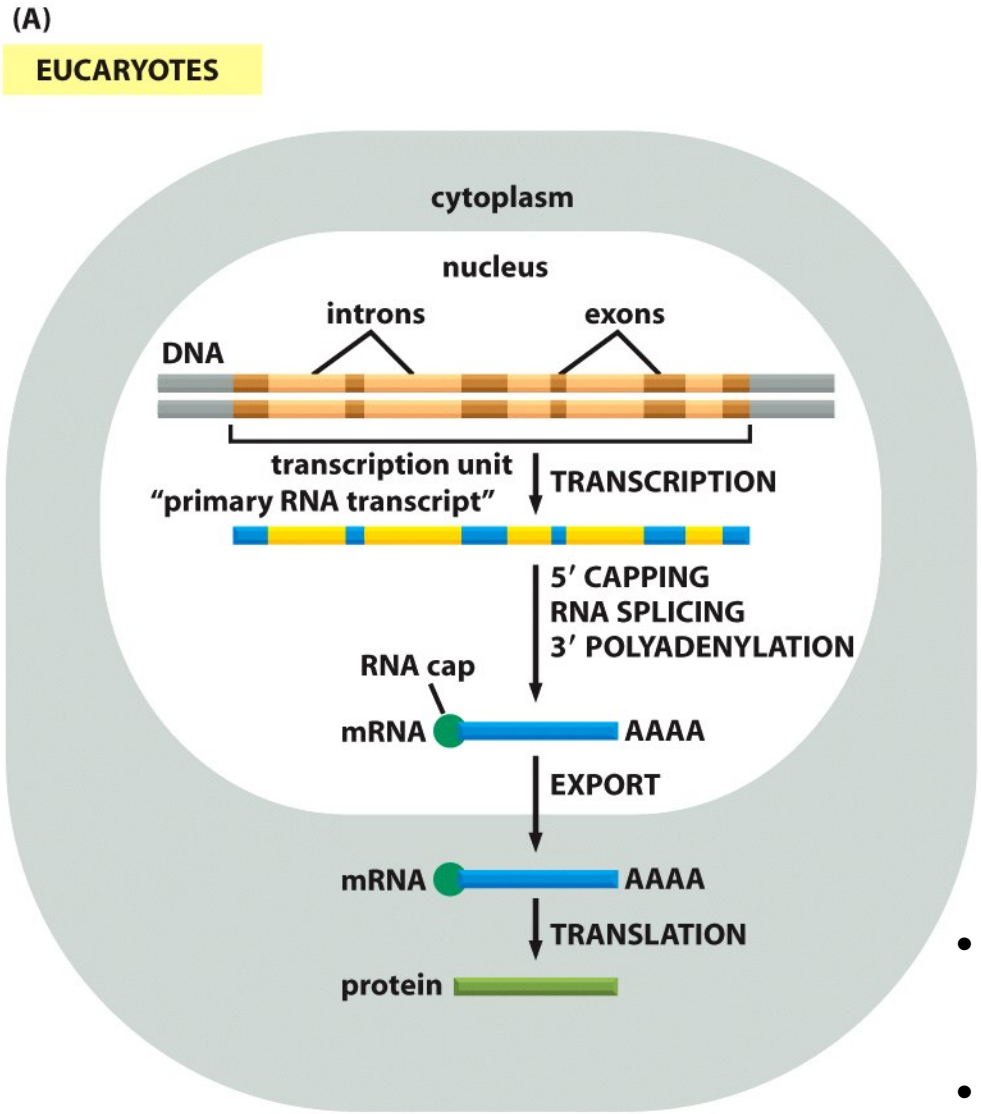
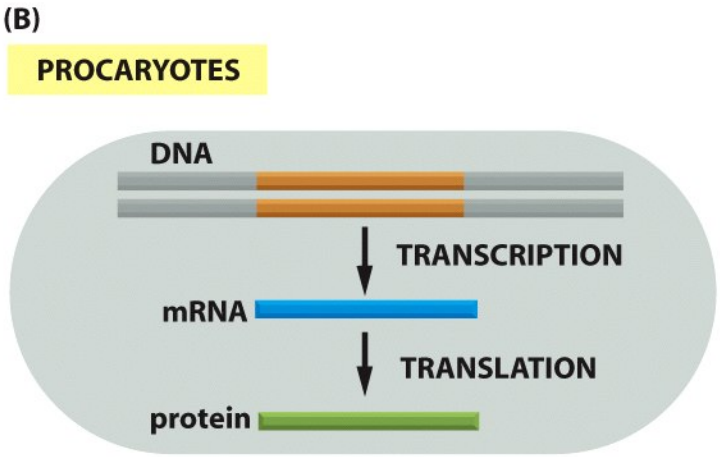


Figure 6-21 *Molecular Biology of the Cell* (© Garland Science 2008)



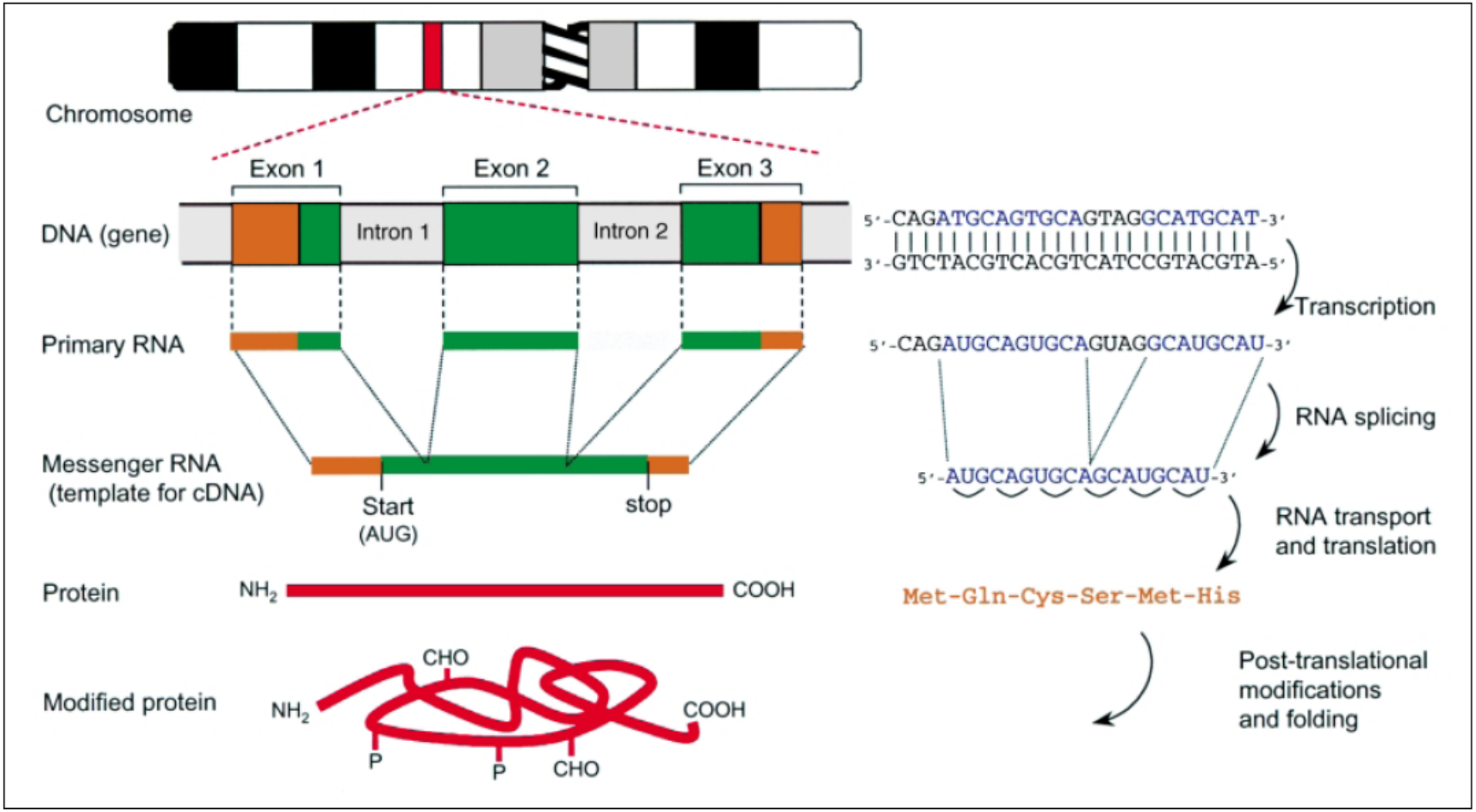


But not in eukaryotes, from us to yeast and all kinds of multicellular organisms

- Transcription and translation are physically separated in eukaryotes
- Genes do not have same sequence as RNA copied from them

Figure 6-21 *Molecular Biology of the Cell* (© Garland Science 2008)

RNAs transcribed from most eukaryotic genes have large fractions of sequence spliced out during synthesis: “introns”



It’s the genetic code of the spliced RNA, not the genomic DNA, that is read into protein

DNA code is like a book written with “nonsense” between the words and no spaces: reading is hard

For protein coding genes, you can only interpret the sequence if you know where intron/exon boundaries lie……. And if there are any mutations, their exact position determines whether or not they change the code

Genome sources part 1

- UCSC Genome Browser
  - http://genome.ucsc.edu/
  - Genomes
  - Blat (sequence search)
  - Easily customized with upload annotations

- Ensembl Genome Browser
  - http://www.ensembl.org/
  - Many genomes
  - “Blast” sequence search tool
  - Annotations
  - Transcripts and protein variants, splice sites and domains
  - Highly curated

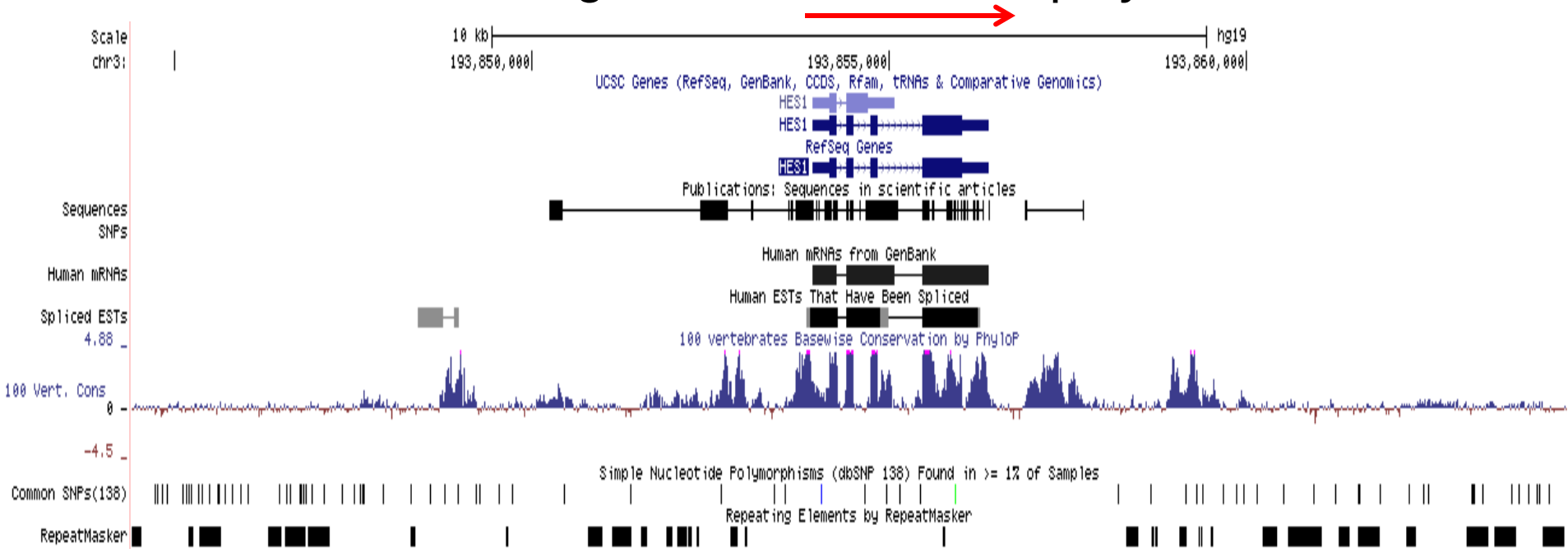
Genome sources part 2

- National Center for Biotechnology Information (NCBI)
  - http://www.ncbi.nlm.nih.gov/Entrez/
  - Cross-database searches
  - PubMed (literature), Genes, Proteins, OMIM (human physiology connections for gene), Genome, Unigene, GEO (Gene Expression Omnibus – microarray and ultra-high throughput sequencing data)
  - Many databases curated and cross referenced

- Specialized databases for particular organisms
  - Mouse genome informatics
  - Flybase
  - Wormbase
  - SpBase
  - Combine genome sequence, RNA sequence, and RNA expression data with annotation about proteins encoded by these genes

An important, small gene: *Hes1*

UCSC genome browser display



- Uses known human genomic sequence as yardstick (3 x 10^9 bp)

Where are the gene boundaries?

→ Existence of previously cloned, sequenced mRNAs in database

- Exons and introns can be seen as sequences that are also found joined in mRNA (exons), or clipped out of it (introns)
- “Fat” bars in “genes” tracks represent protein coding sequences
- “Thin” bars are parts of sequence included in Hes1 RNA, but outside the protein coding part of RNA (to be described in much detail later)

What is all that other space in the genome around the genes?

- Junk?
- Selfish DNA?
- Regulatory elements?

Maybe, but less that we thought;

Yes;

Yes absolutely!

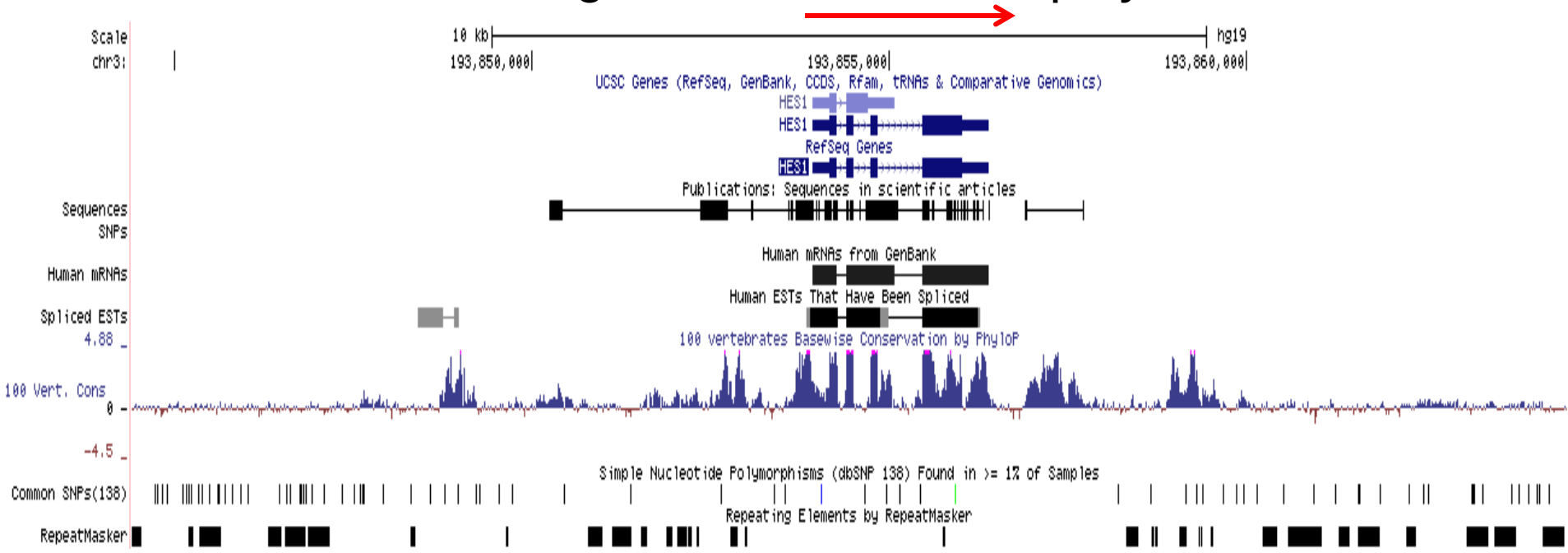
Using evolution to distinguish functional from non-functional DNA sequences

- Mutations will happen *(more about this later)*
- But will the animals with those mutations survive?
- Sequences that are important may not be mutated without harming survival of animals over many generations…

➢ Future generations have unusually low mutation rates in sequences that are important

An important, small gene: *Hes1*

UCSC genome browser display



- Evolution selects for sequences that mediate important function, coding OR regulatory
- → Conservation defines physiologically most-important sequences
- Browser shows sites where sequence is unusually constant between humans (SNPs) and other distantly related organisms (“vert. cons.”)
- What else is there? Repeats, and “empty” space…

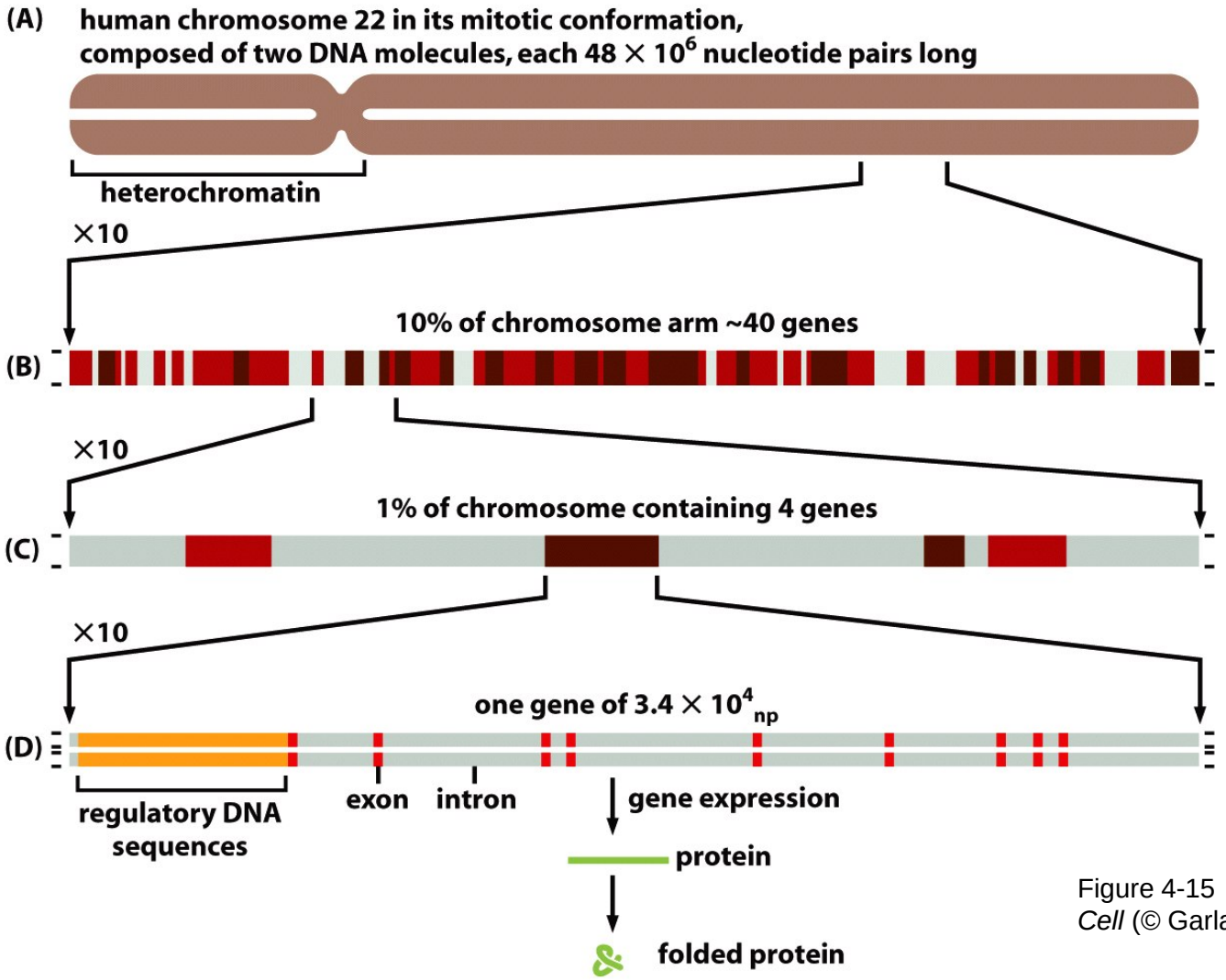


Figure 4-15 *Molecular Biology of the Cell* (© Garland Science 2008)

Total length of human DNA: ~3.2 x 10^9 base pairs; approx 25,000 genes
Gene sizes: avg. 27K bp, made of avg. 104 exons; max known 2.4 x 10^6 bp, up to >175 exons
~50% of DNA in high-copy repeats; about 40% uncharacterized
Just 1.5% protein coding….. ***3.5% highly conserved, noncoding***…
What is the function for which these sequences are conserved??

Cardinal rules of genomic information in metazoan multicellular organisms (i.e. us)

- The DNA is the ***same*** in all cells in the body of a given organism
- The part of the DNA that is transcribed into RNA is ***not*** the same in different cell types
- Different cell types express ***different*** sequences in their RNA, even though they share the sequences for ***all*** genes in their DNA

➢ Differentially regulated gene expression is the key to multicellular life

Overwhelmingly, the dominant mechanism that controls levels of protein expression is control of RNA copy number in cell

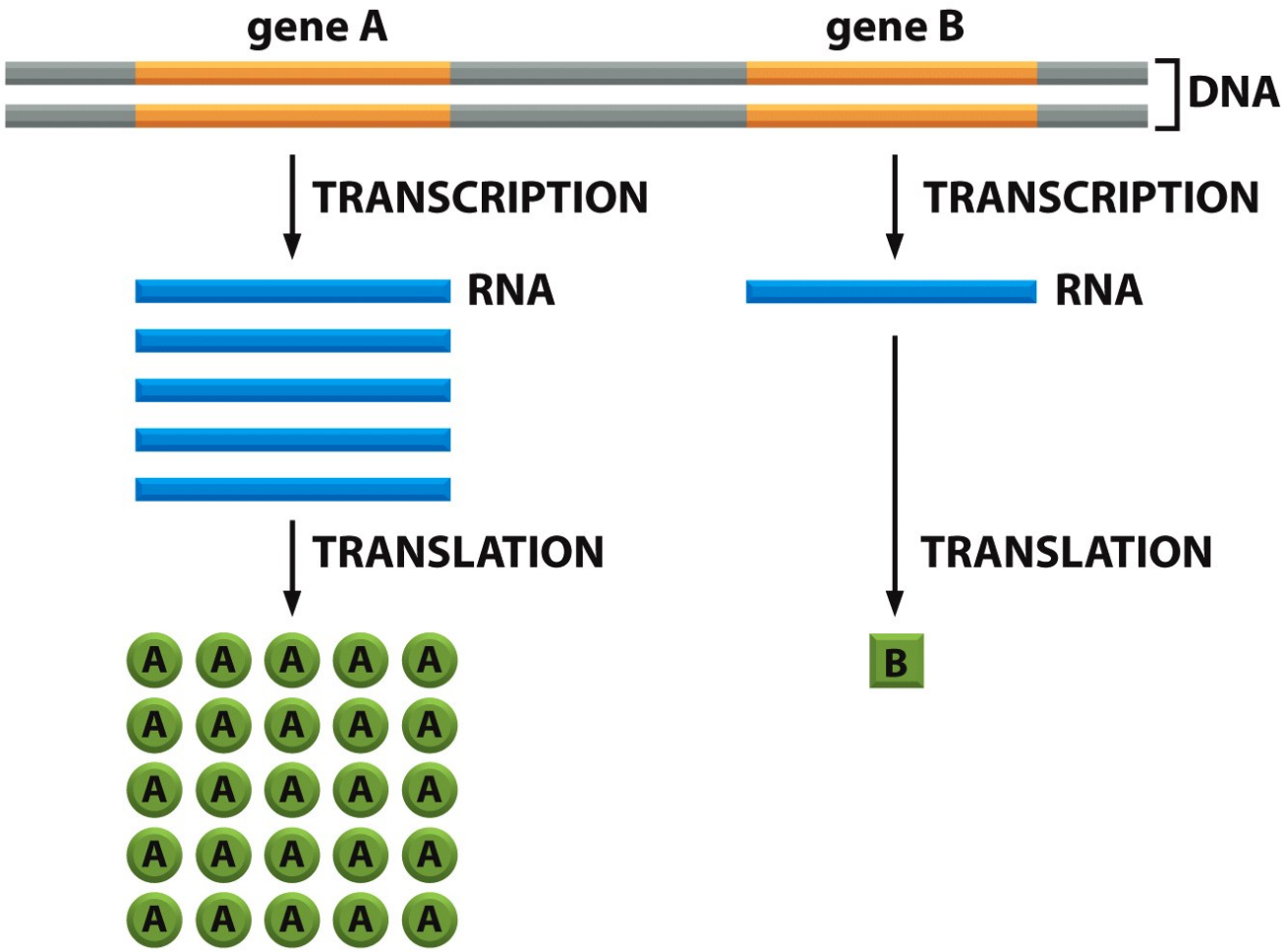


Figure 6-3 *Molecular Biology of the Cell* (© Garland Science 2008)

Gene expression regulation is the central core of development in multicellular organisms

- Information needed to determine where, when, how much a gene is expressed is ITSELF encoded in the genome

- Need to decode sequence content that controls ***RNA expression and processing***, as well as translation into protein, in order to “read” genetic code

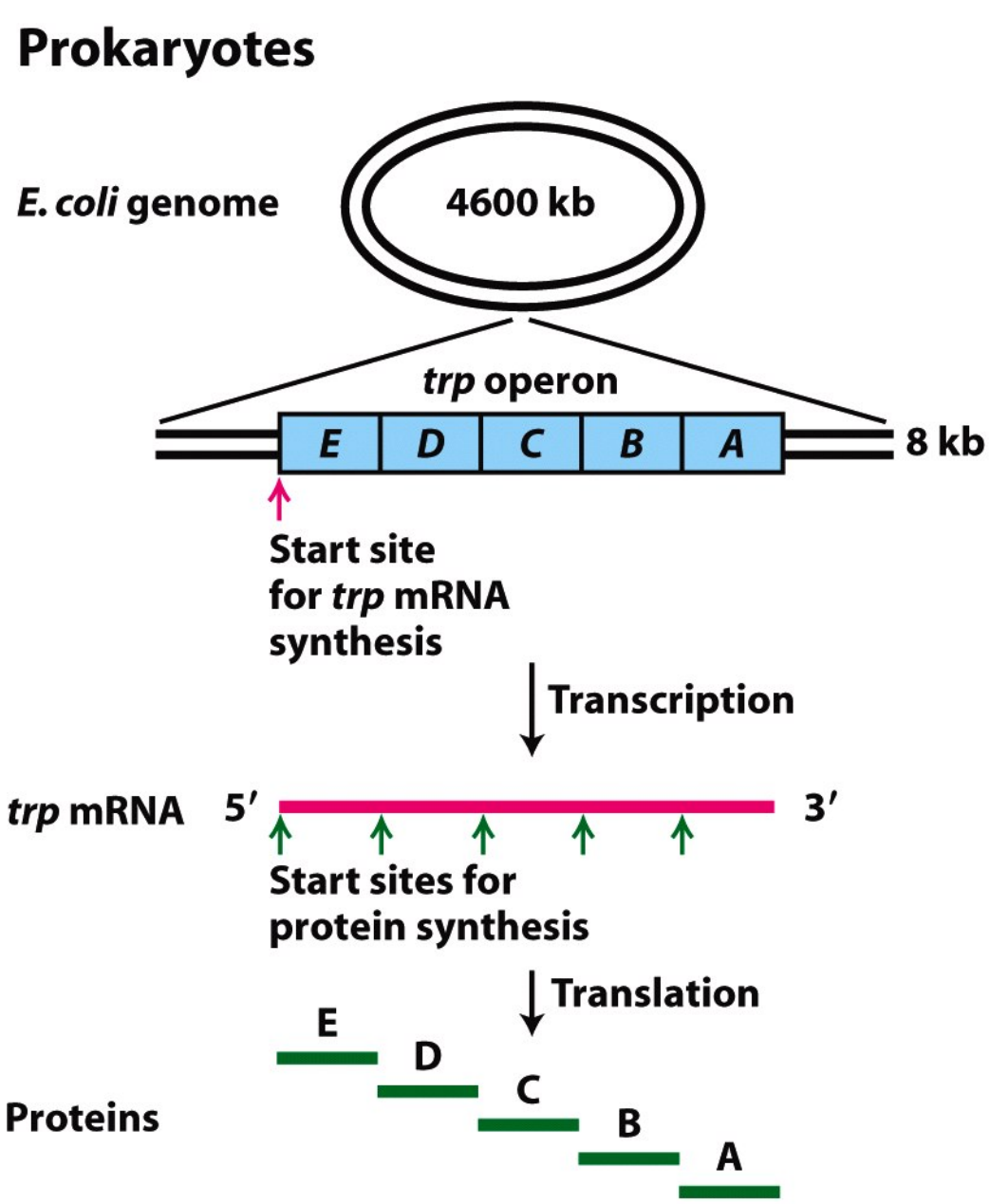


Figure 4-13a
Molecular Cell Biology, Sixth Edition
© 2008 W. H. Freeman and Company

Even prokaryotes have a need for coordinating expression of certain proteins that are only needed occasionally

"Operon" structure allows one RNA to encode a chain of distinct proteins

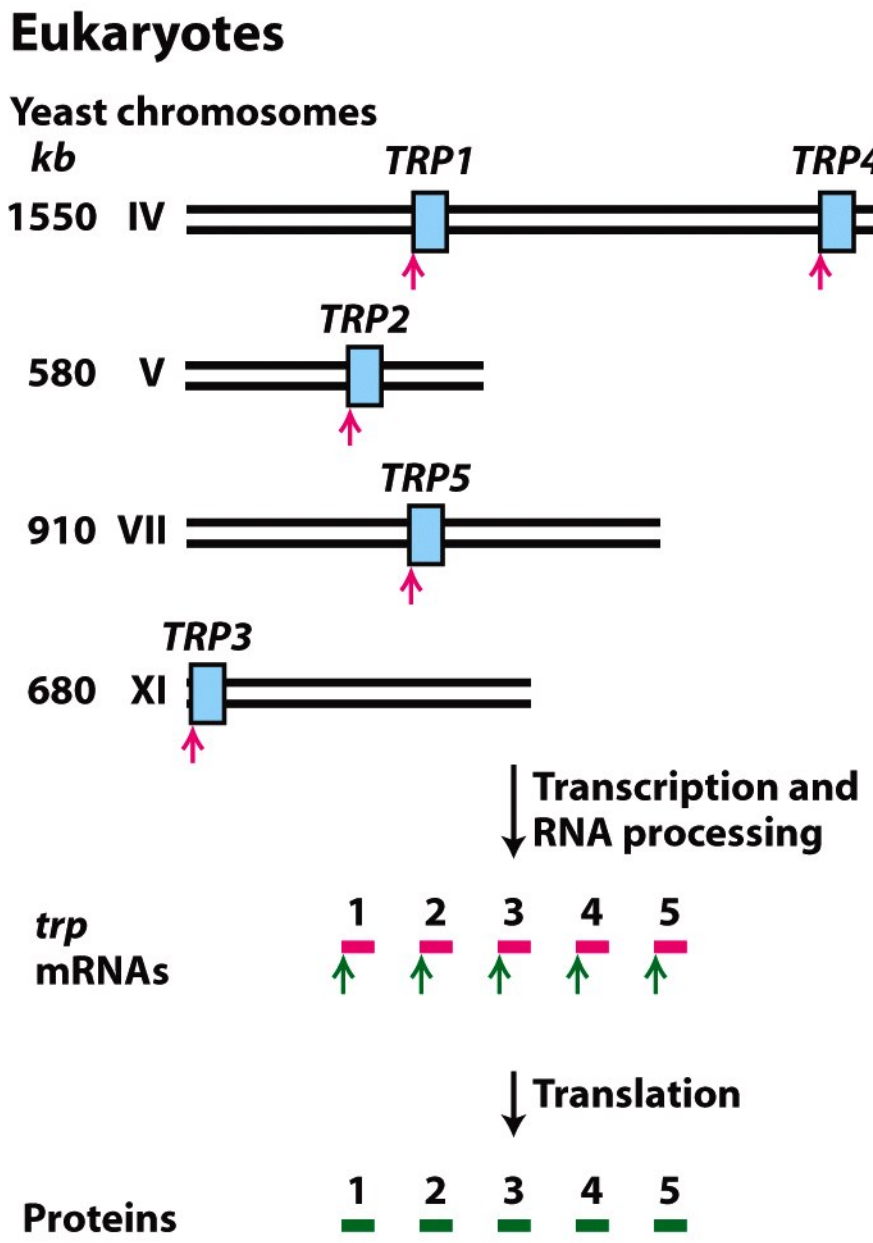


Figure 4-13b
Molecular Cell Biology, Sixth Edition
© 2008 W. H. Freeman and Company

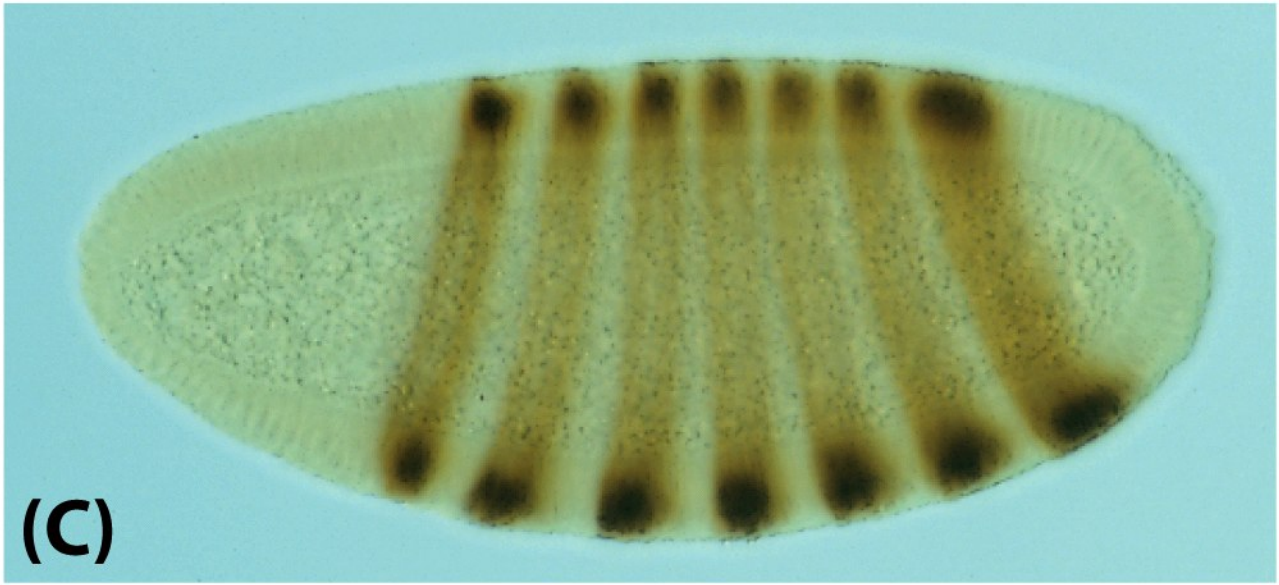
Eukaryotes need to transcribe separate RNAs for each protein

Corollary: Each protein's level depends on control of transcription of a separate RNA…
coordinated synthesis requires coordination of transcription

Elegant demonstrations of the power of gene expression regulation come from embryos

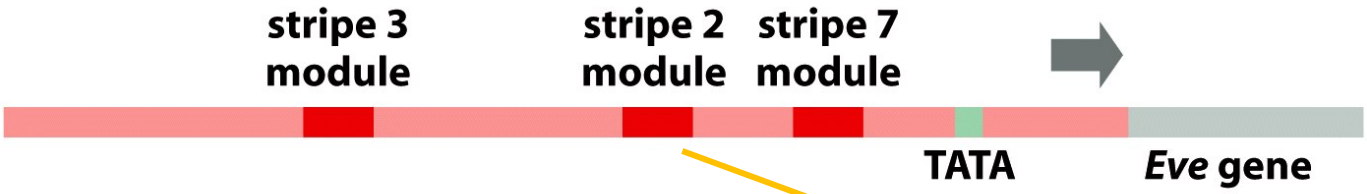
Major subdivisions of the future body plan are laid out initially by differential levels of expression of key genes in different bands or patches of cells

…often in intricate patterns

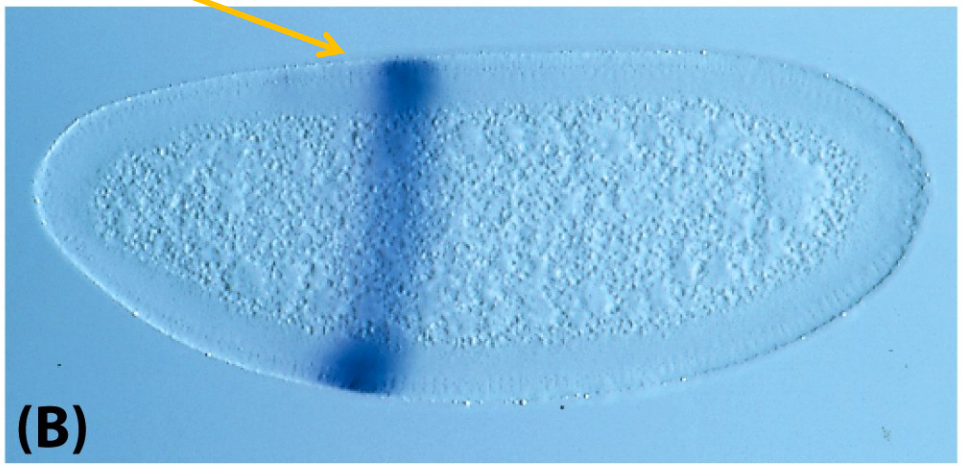


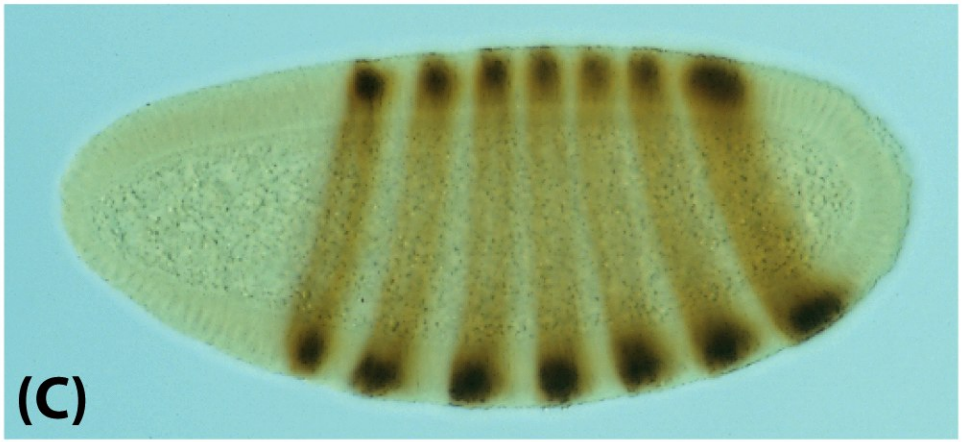
Early fly embryo

Figure 7-55b,c *Molecular Biology of the Cell* (© Garland Science 2008)



Information to control these patterns is inherited in non-protein-coding sequence of the genome

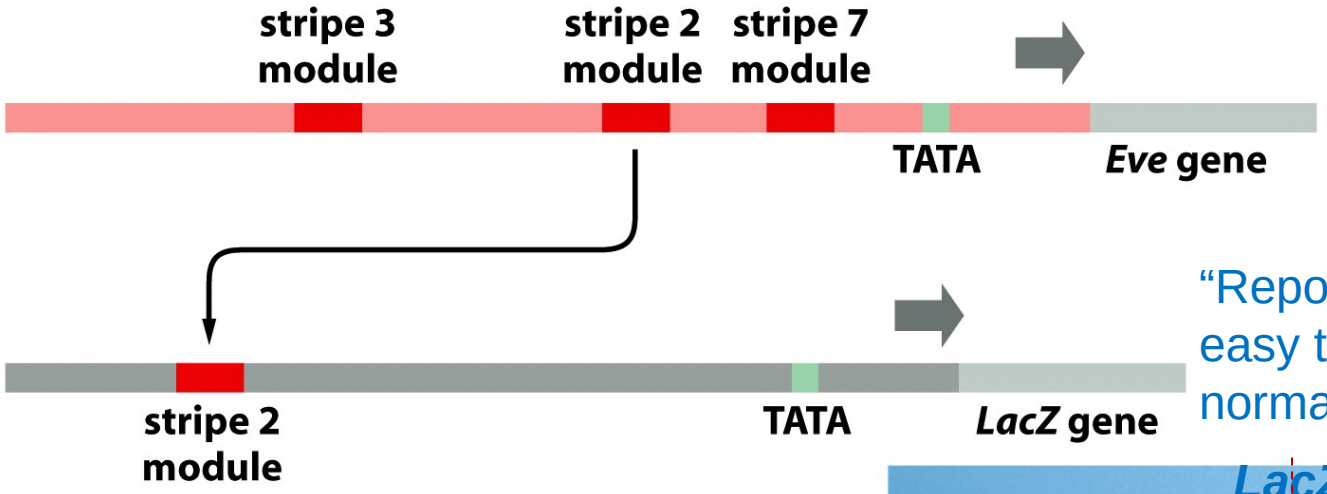




Regulatory sequences can be identified using a “reporter gene expression” assay

- DNA pieces can be isolated by “cloning”
- DNA pieces from one source can be ligated together in a very precise way with DNA from another source
- DNA can be introduced into a cell or an embryo that would not normally contain it – “transfection” or “microinjection”
- Transfected cell uses DNA information obediently, to promote RNA expression and/or protein expression

Candidate regulator sequence drives expression of reporter: results show which sequences work to promote expression in the right pattern…

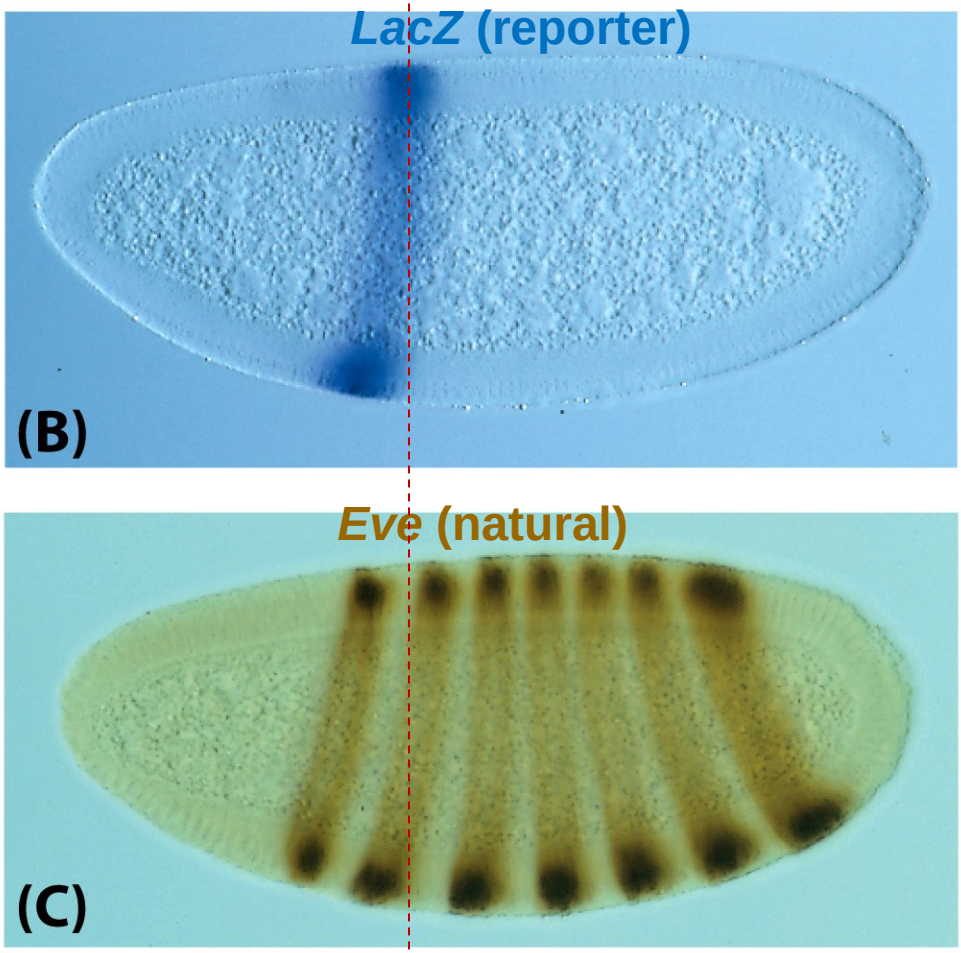


"Reporter" protein that is easy to detect and not in the normal cell

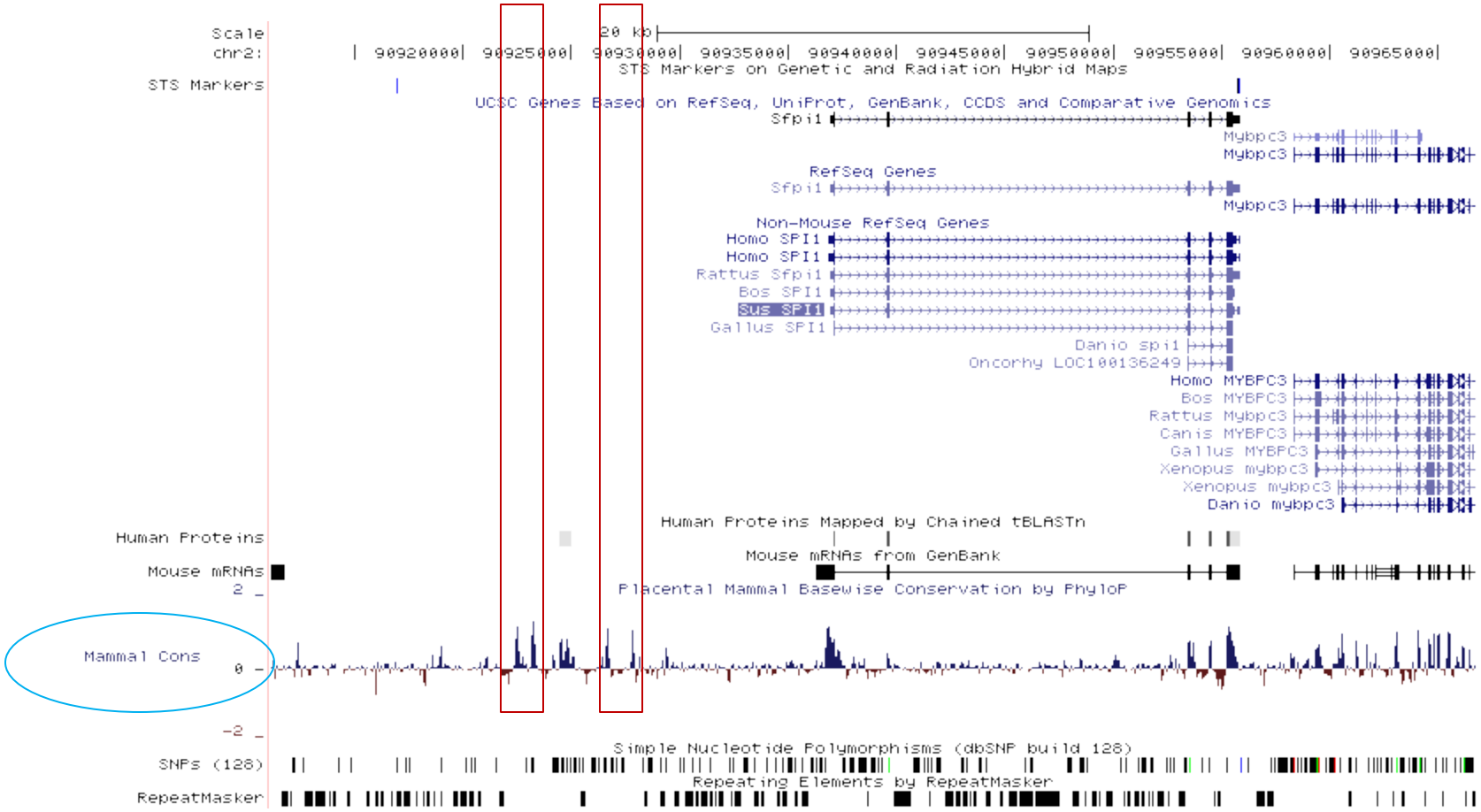
Candidate regulatory sequence (one of many tested)

Figure 7-55a *Molecular Biology of the Cell* (© Garland Science 2008)

Information to control these patterns is inherited in non-protein-coding sequence of the genome

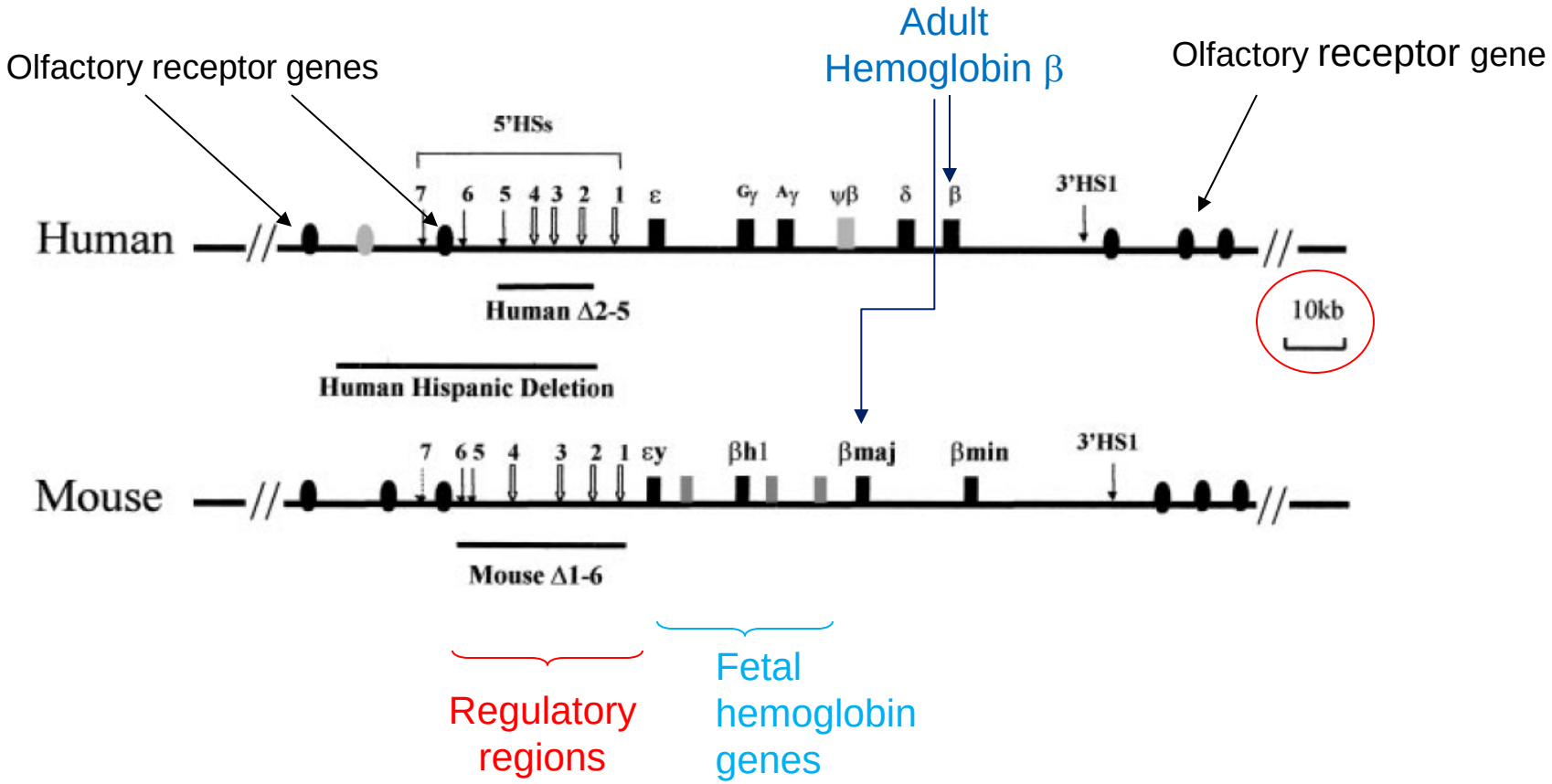


Conservation can help to show where to look to identify crucial regulatory elements



Experimentally validated regulatory sequences can also be downstream… far downstream… even beyond the next gene

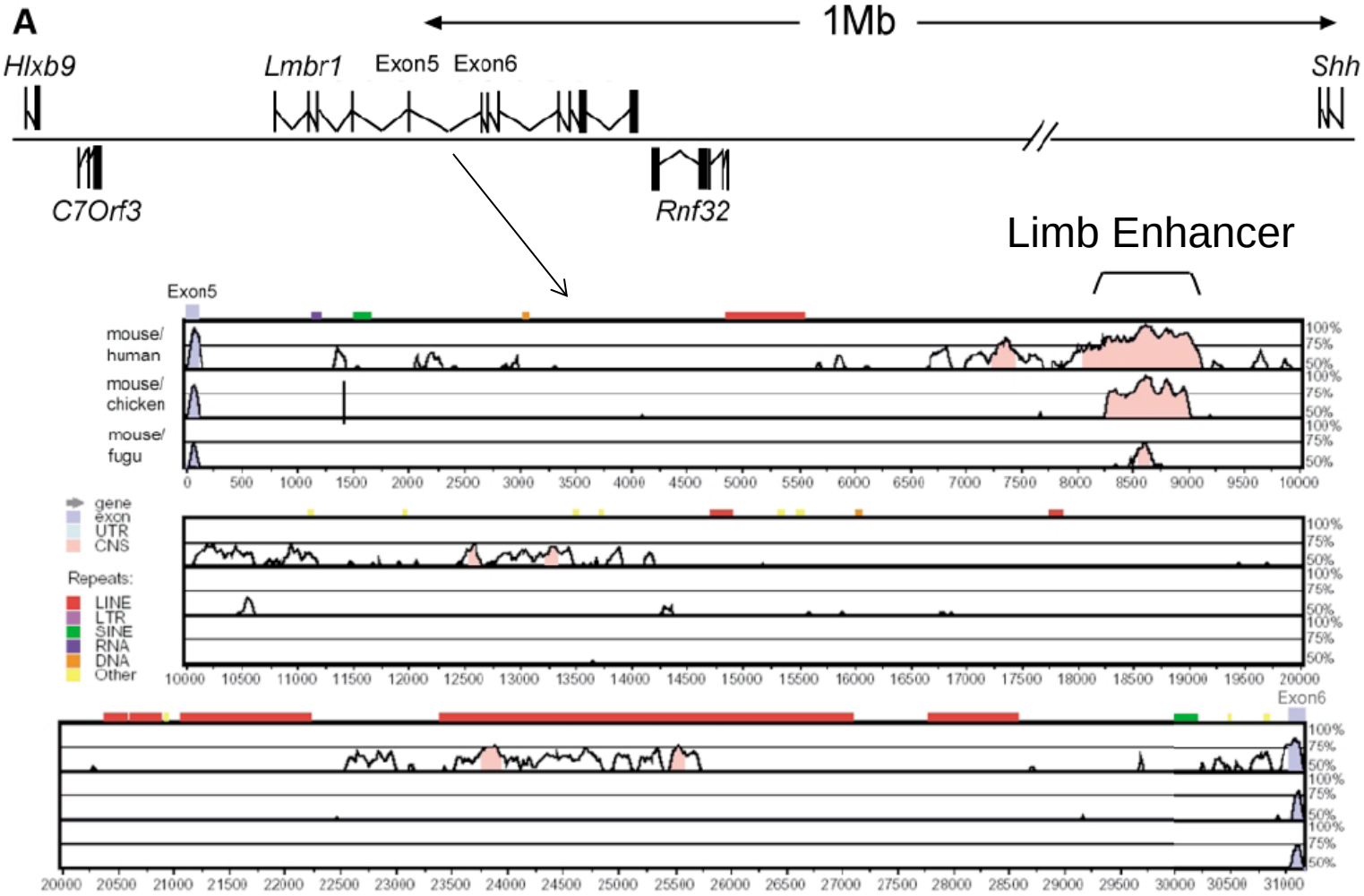
Regulatory regions important for correct expression of mammalian β-globin genes lie far upstream



(Q. Li, K.R. Peterson, X. Fang, & G. Stamatoyannopoulos Blood 100:3077-3086 (2002))

A distant enhancer for the *Shh* gene, more than 10^6 bp away in the intron of another gene, is crucial for vertebrate limb development…. This DNA sequence is conserved between us and chickens, pufferfish

Human Molecular Genetics, 2003, Vol. 12, No. 14 (*Lettice et al.*)



Single base changes in enhancer make Shh expression abnormal in limbs & cause deformity…

But Shh can still be expressed elsewhere

Implications of gene regulation for definition of a "gene"

- The boundaries of genes are ***not*** easily predicted in genomes of multicellular eukaryotes based on DNA sequence alone
- "Genes" are units defined by transcription into RNA
- But… because RNA expression is different among different cells, any one cell or tissue may not contain a given RNA, even though the gene exists and is expressed elsewhere
  - Can't count all the genes in an organism's genome unless all the RNAs expressed in *all* tissues are identified
  - Alternatively, need evidence that the equivalent sequence in a different organism is a real gene
- RNA detection limits based on *experimental measurement* are crucial for assessing what sequences are "genes"

Cardinal rules of genomic information in metazoan multicellular organisms (i.e. us)

- The DNA is the ***same*** in all cells in the body of a given organism
- The part of the DNA that is transcribed into RNA is ***not*** the same in different cell types
- Different cell types express ***different*** sequences in their RNA, even though they share the sequences for ***all*** genes in their DNA

➢ Differentially regulated gene expression is the key to multicellular life