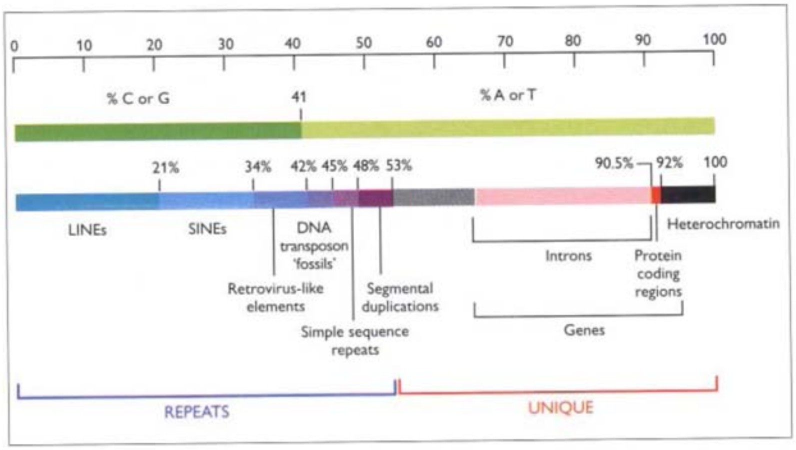
cDNA libraries, EST clusters, gene prediction and functional annotation

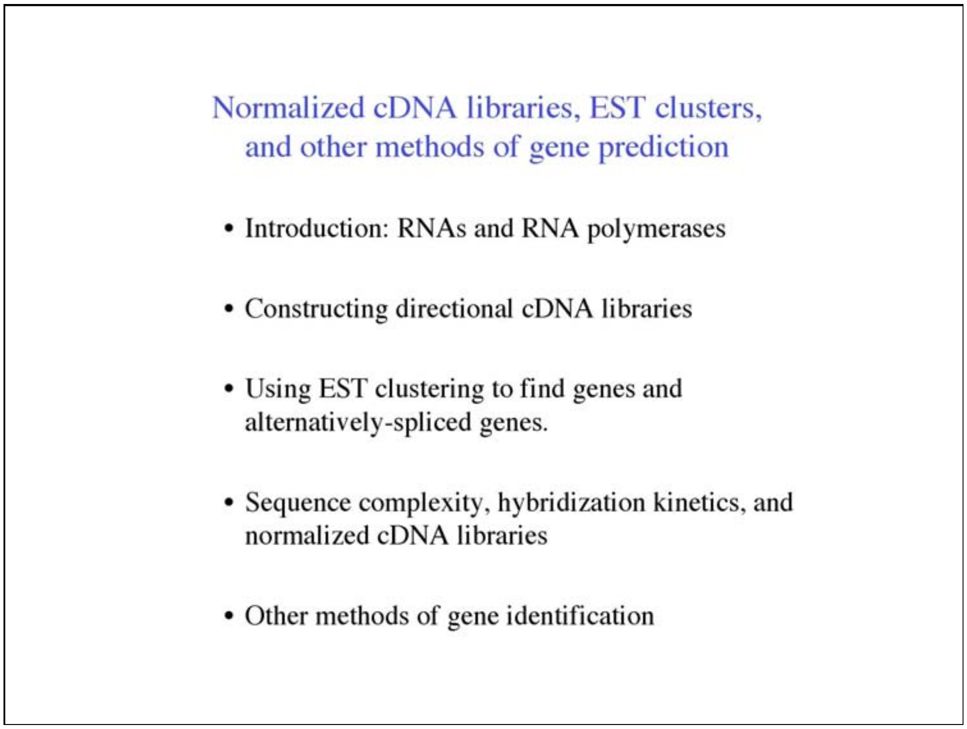
Biosciences 741: Genomics

Fall, 2013

Week 3

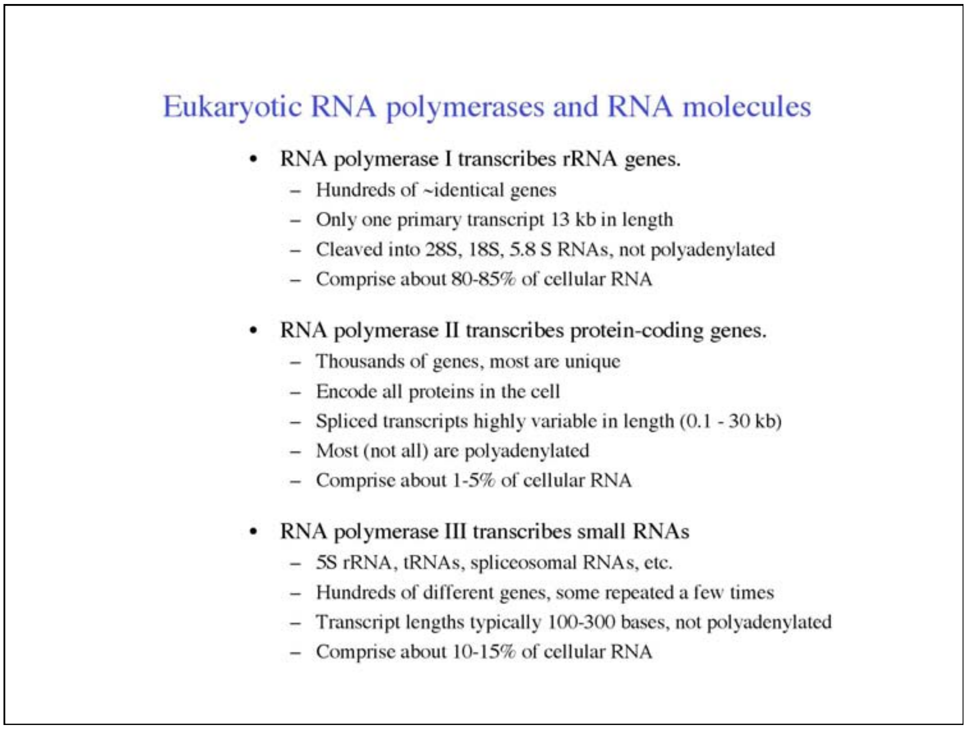
Content of the human genome





Normalized cDNA libraries, EST clusters, and other methods of gene prediction

- Introduction: RNAs and RNA polymerases
- Constructing directional cDNA libraries
- Using EST clustering to find genes and alternatively-spliced genes.
- Sequence complexity, hybridization kinetics, and normalized cDNA libraries
- Other methods of gene identification



Eukaryotic RNA polymerases and RNA molecules

- RNA polymerase I transcribes rRNA genes.
  - Hundreds of ~identical genes
  - Only one primary transcript 13 kb in length
  - Cleaved into 28S, 18S, 5.8 S RNAs, not polyadenylated
  - Comprise about 80-85% of cellular RNA
- RNA polymerase II transcribes protein-coding genes.
  - Thousands of genes, most are unique
  - Encode all proteins in the cell
  - Spliced transcripts highly variable in length (0.1 - 30 kb)
  - Most (not all) are polyadenylated
  - Comprise about 1-5% of cellular RNA
- RNA polymerase III transcribes small RNAs
  - 5S rRNA, tRNAs, spliceosomal RNAs, etc.
  - Hundreds of different genes, some repeated a few times
  - Transcript lengths typically 100-300 bases, not polyadenylated
  - Comprise about 10-15% of cellular RNA

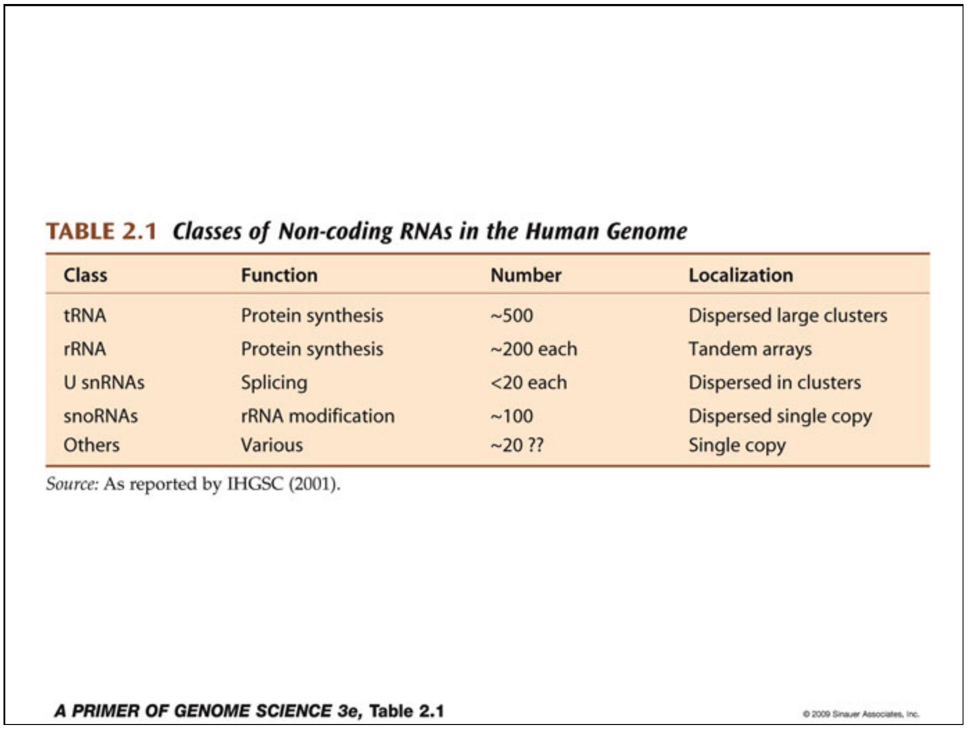


TABLE 2.1 ***Classes of Non-coding RNAs in the Human Genome***

Class	Function	Number	Localization
tRNA	Protein synthesis	~500	Dispersed large clusters
rRNA	Protein synthesis	~200 each	Tandem arrays
U snRNAs	Splicing	<20 each	Dispersed in clusters
snoRNAs	rRNA modification	~100	Dispersed single copy
Others	Various	~20 ??	Single copy

Source: As reported by IHGSC (2001).

***A PRIMER OF GENOME SCIENCE 3e*, Table 2.1**

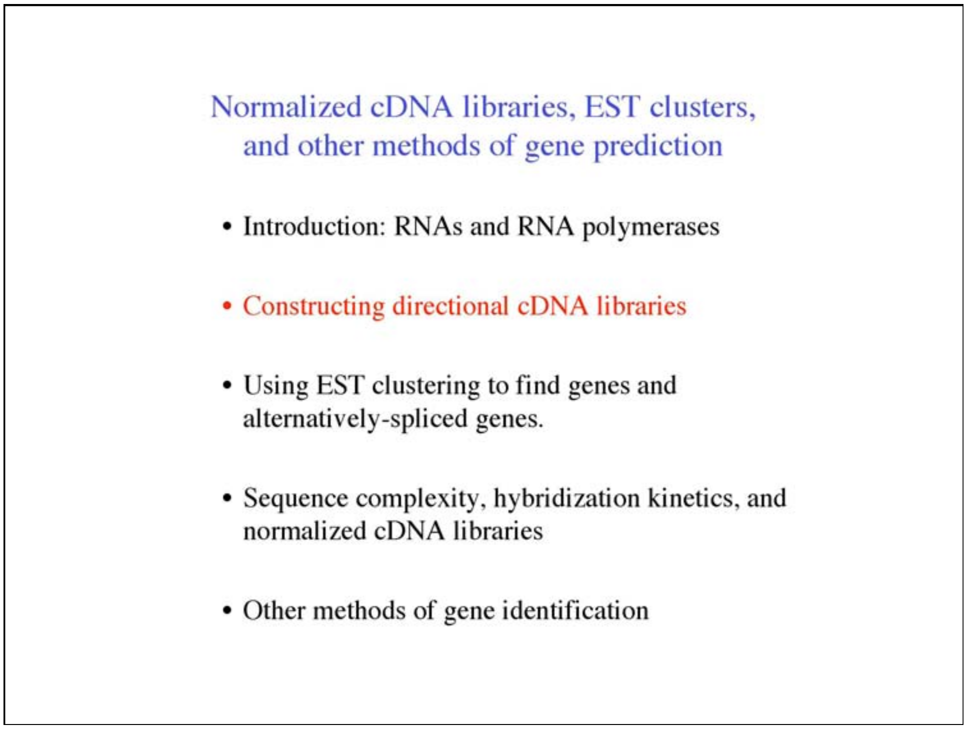
Comparison of eukaryotic, prokaryotic, and bacteriophage RNA polymerases

- Many bacteriophage RNA polymerases (T3, T7, Sp6, etc. consist of a single polypeptide chain that can autonomously initiate RNA synthesis and absolutely requires a single, specific promoter sequence about 15-20 bp in length.

- Prokaryotic RNA polymerase has little ability to initiate transcription by itself. The active form is a holoenzyme complex that contain multiple subunits and recognize multiple promoter sequences.

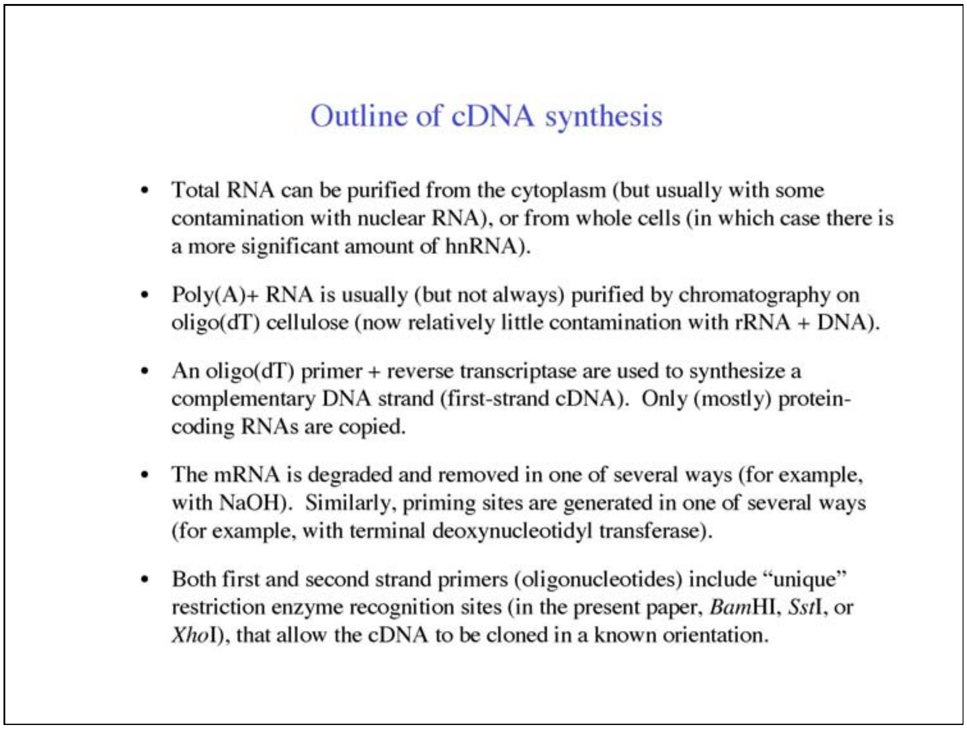
- These trends have continued in eukaryotes, in which RNA polymerase is unable to initiate transcription by itself, the active transcription complex involves dozens of proteins, and very large number of very different promoters are recognized. Along with this complexity has come a significant background of nonspecific transcription, plus significant variability in the 5' ends of many genes.

- Eukaryotic mRNA transcript termination is also a somewhat imprecise process - sometimes polymerase reads through a valid poly(A) signal and keeps on going.



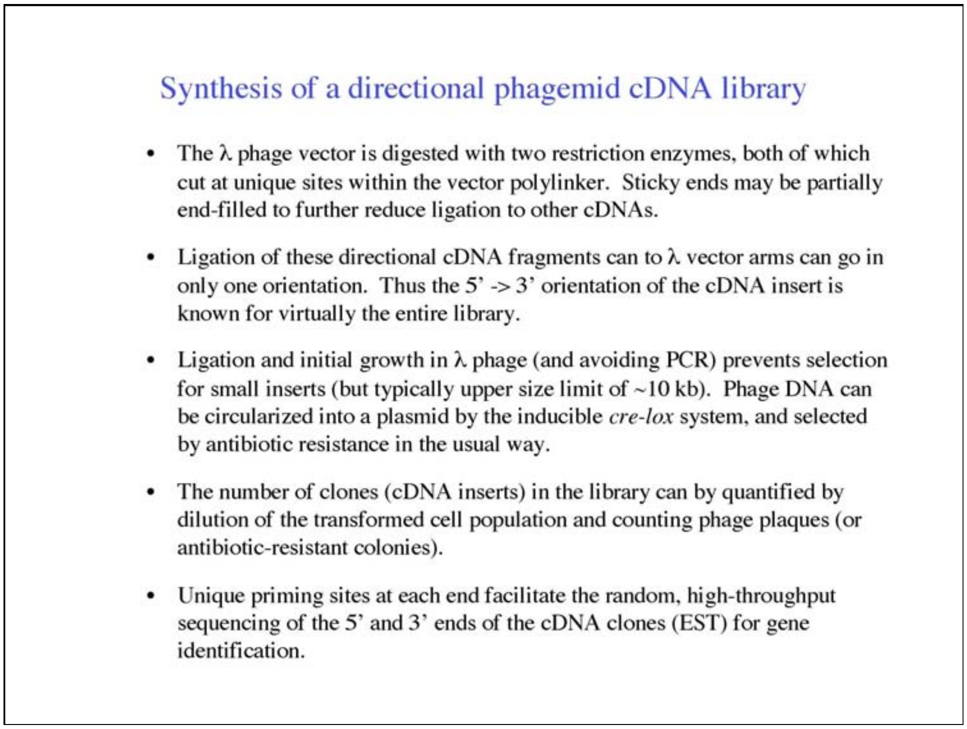
Normalized cDNA libraries, EST clusters, and other methods of gene prediction

- Introduction: RNAs and RNA polymerases
- Constructing directional cDNA libraries
- Using EST clustering to find genes and alternatively-spliced genes.
- Sequence complexity, hybridization kinetics, and normalized cDNA libraries
- Other methods of gene identification



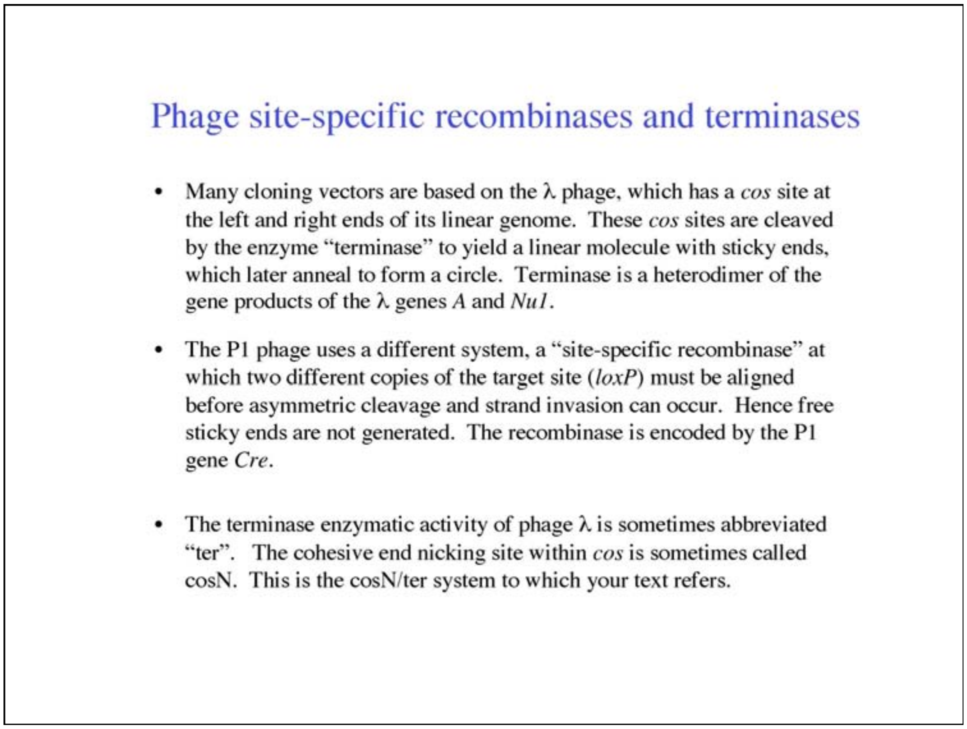
Outline of cDNA synthesis

- Total RNA can be purified from the cytoplasm (but usually with some contamination with nuclear RNA), or from whole cells (in which case there is a more significant amount of hnRNA).
- Poly(A)+ RNA is usually (but not always) purified by chromatography on oligo(dT) cellulose (now relatively little contamination with rRNA + DNA).
- An oligo(dT) primer + reverse transcriptase are used to synthesize a complementary DNA strand (first-strand cDNA). Only (mostly) protein-coding RNAs are copied.
- The mRNA is degraded and removed in one of several ways (for example, with NaOH). Similarly, priming sites are generated in one of several ways (for example, with terminal deoxynucleotidyl transferase).
- Both first and second strand primers (oligonucleotides) include "unique" restriction enzyme recognition sites (in the present paper, *Bam*HI, *Sst*I, or *Xho*I), that allow the cDNA to be cloned in a known orientation.



Synthesis of a directional phagemid cDNA library

- The λ phage vector is digested with two restriction enzymes, both of which cut at unique sites within the vector polylinker. Sticky ends may be partially end-filled to further reduce ligation to other cDNAs.
- Ligation of these directional cDNA fragments can to λ vector arms can go in only one orientation. Thus the 5' -> 3' orientation of the cDNA insert is known for virtually the entire library.
- Ligation and initial growth in λ phage (and avoiding PCR) prevents selection for small inserts (but typically upper size limit of ~10 kb). Phage DNA can be circularized into a plasmid by the inducible *cre-lox* system, and selected by antibiotic resistance in the usual way.
- The number of clones (cDNA inserts) in the library can by quantified by dilution of the transformed cell population and counting phage plaques (or antibiotic-resistant colonies).
- Unique priming sites at each end facilitate the random, high-throughput sequencing of the 5' and 3' ends of the cDNA clones (EST) for gene identification.



Phage site-specific recombinases and terminases

- Many cloning vectors are based on the λ phage, which has a *cos* site at the left and right ends of its linear genome. These *cos* sites are cleaved by the enzyme "terminase" to yield a linear molecule with sticky ends, which later anneal to form a circle. Terminase is a heterodimer of the gene products of the λ genes *A* and *Nu1*.
- The P1 phage uses a different system, a "site-specific recombinase" at which two different copies of the target site (*loxP*) must be aligned before asymmetric cleavage and strand invasion can occur. Hence free sticky ends are not generated. The recombinase is encoded by the P1 gene *Cre*.
- The terminase enzymatic activity of phage λ is sometimes abbreviated "ter". The cohesive end nicking site within *cos* is sometimes called cosN. This is the cosN/ter system to which your text refers.

Normalized cDNA libraries, EST clusters, and other methods of gene prediction

- Introduction: RNAs and RNA polymerases
- Constructing directional cDNA libraries
- Using EST clustering to find genes and alternatively-spliced genes.
- Sequence complexity, hybridization kinetics, and normalized cDNA libraries
- Other methods of gene identification

Directional cDNA libraries can be searched by EST sequencing and sequence alignments

- Cycle sequencing can be used to generate sequence data from either end of the insert in a double-stranded DNA plasmid (cycle sequencing is essentially just dideoxy sequencing, as presented last time, plus *Taq* polymerase and a temperature cycle).
- Clones are typically sequenced from the 3' end first, because:
  - the end of the poly(A) tract provides a reproducible reference point
  - neutral evolution causes closely related genes to have highly diverged 3' untranslated regions.
- The other steps (select many clones at random, perform only one sequencing run per clone, automated sequence alignment into contigs) are similar to shotgun sequencing. But in this case the contigs are called "unigene clusters", each of which corresponds (more or less) to a gene.
- Documenting the 5' end of a gene is more difficult, because reverse transcriptase often falls off the mRNA before completely copying it.

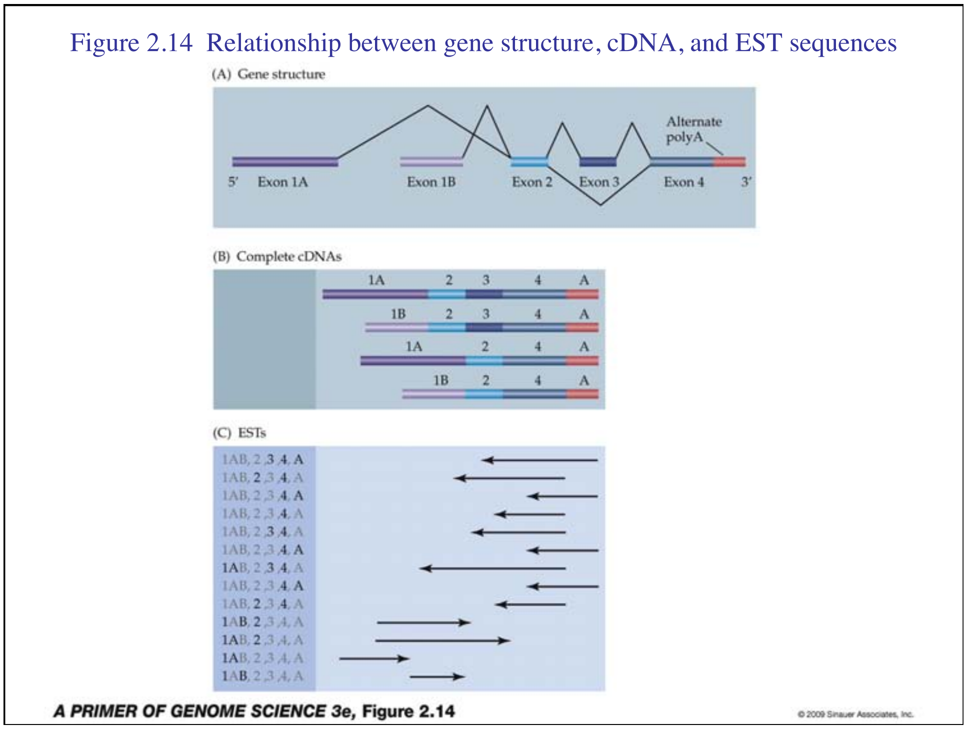


Figure 2.14 Relationship between gene structure, cDNA, and EST sequences

***A PRIMER OF GENOME SCIENCE 3e*, Figure 2.14**

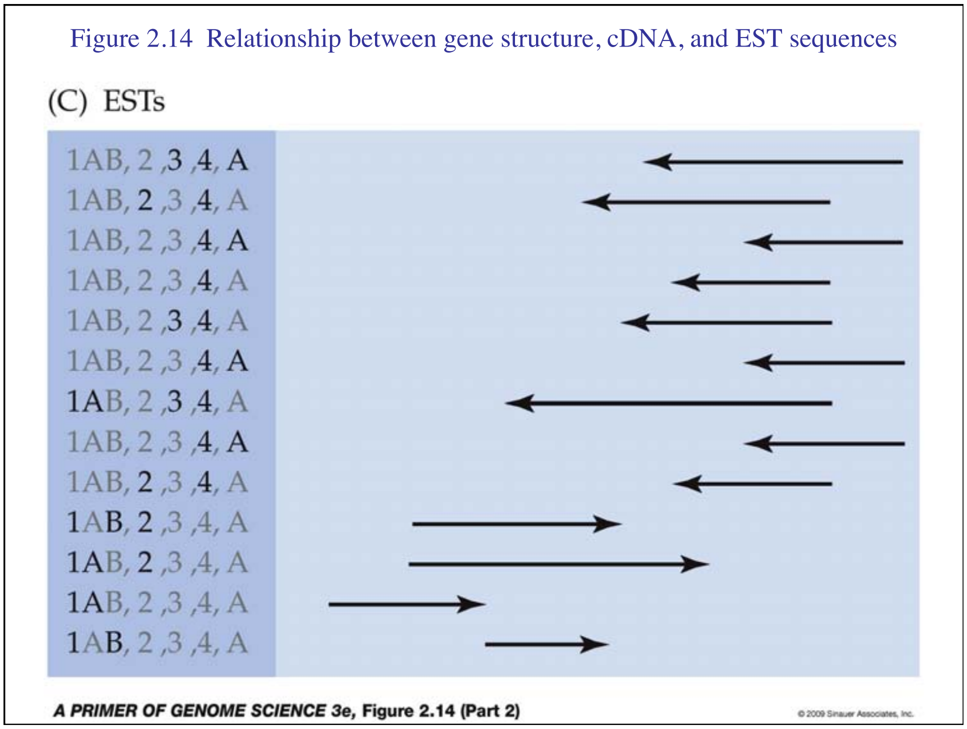
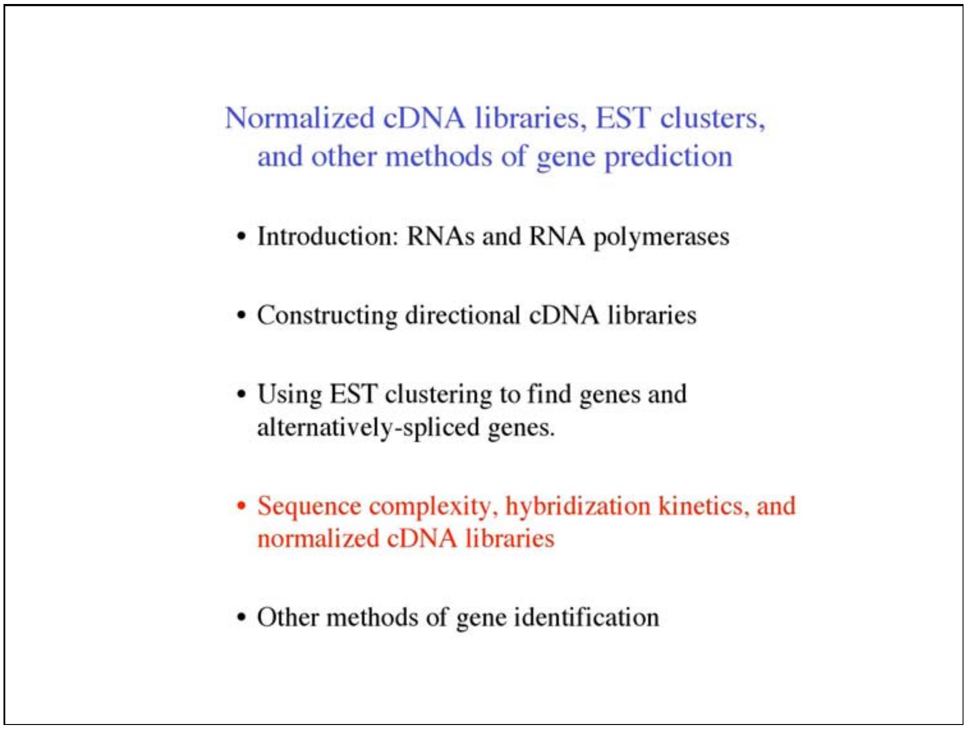


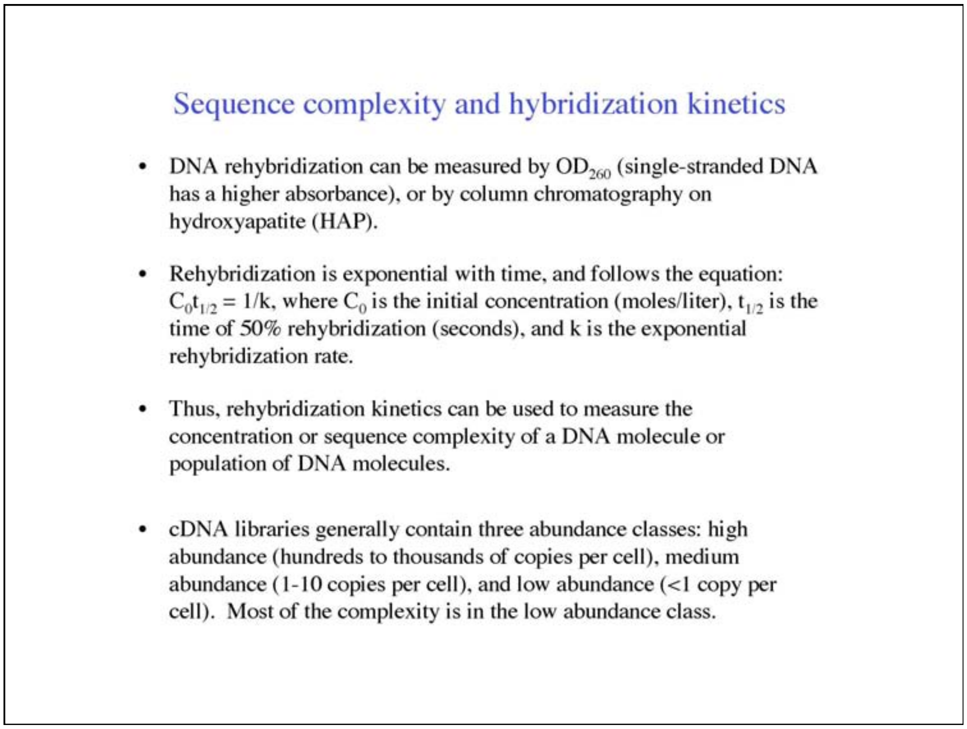
Figure 2.14 Relationship between gene structure, cDNA, and EST sequences

***A PRIMER OF GENOME SCIENCE 3e*, Figure 2.14 (Part 2)**



Normalized cDNA libraries, EST clusters, and other methods of gene prediction

- Introduction: RNAs and RNA polymerases
- Constructing directional cDNA libraries
- Using EST clustering to find genes and alternatively-spliced genes.
- Sequence complexity, hybridization kinetics, and normalized cDNA libraries
- Other methods of gene identification



Sequence complexity and hybridization kinetics

- DNA rehybridization can be measured by OD_{260} (single-stranded DNA has a higher absorbance), or by column chromatography on hydroxyapatite (HAP).
- Rehybridization is exponential with time, and follows the equation: $C_0t_{1/2} = 1/k$, where C_0 is the initial concentration (moles/liter), $t_{1/2}$ is the time of 50% rehybridization (seconds), and k is the exponential rehybridization rate.
- Thus, rehybridization kinetics can be used to measure the concentration or sequence complexity of a DNA molecule or population of DNA molecules.
- cDNA libraries generally contain three abundance classes: high abundance (hundreds to thousands of copies per cell), medium abundance (1-10 copies per cell), and low abundance (<1 copy per cell). Most of the complexity is in the low abundance class.

Subtracted and/or normalized cDNA libraries

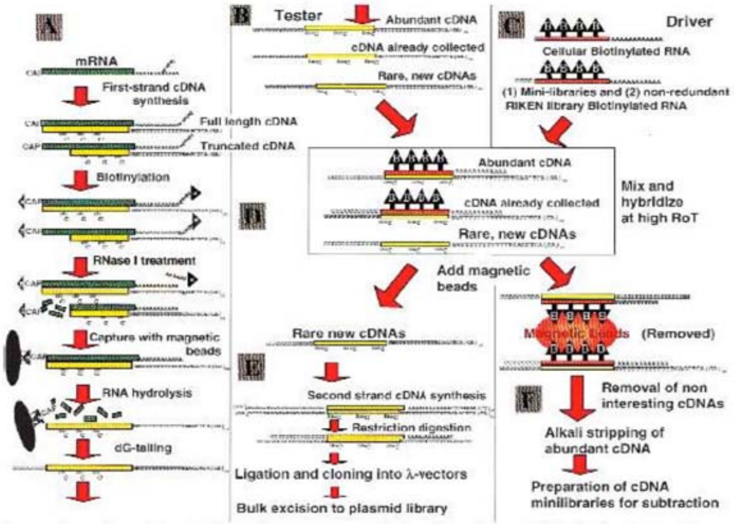

A
mRNA
CAP
First-strand cDNA synthesis
Full length cDNA
Truncated cDNA
Biotinylation
RNase I treatment
Capture with magnetic beads
RNA hydrolysis
dG-tailing
B
Tester
Abundant cDNA
cDNA already collected
Rare, new cDNAs
C
Driver
Cellular Biotinylated RNA
(1) Mini-libraries and (2) non-redundant RIKEN library Biotinylated RNA
D
Abundant cDNA
cDNA already collected
Rare, new cDNAs
Mix and hybridize at high RoT
Add magnetic beads
Rare new cDNAs
Magnetic beads
(Removed)
E
Second strand cDNA synthesis
Restriction digestion
Ligation and cloning into λ-vectors
Bulk excision to plasmid library
F
Removal of non interesting cDNAs
Alkali stripping of abundant cDNA
Preparation of cDNA minilibraries for subtraction

Figure 1 Schematic diagram of the normalized-subtracted cDNA preparation protocol. (*A*) general scheme for preparing full-length single-strand cDNA; (*B*) representation of various populations of tester cDNAs; (*C*) normalizing driver (cellular mRNA) and subtracting drivers (run-off transcripts); (*D*) hybridization; (*E*) rare/new cDNAs are used for second-strand cDNA preparation (normalized/subtracted cDNA library); (*F*) abundant cDNAs/unwanted cDNAs are removed and may be used for the preparation of minilibraries to implement subtraction.

Carninci et al. (2000) Genome Res. 10, 1617-1630

Subtracted and/or normalized cDNA libraries greatly reduce sequencing redundancy in EST genome projects

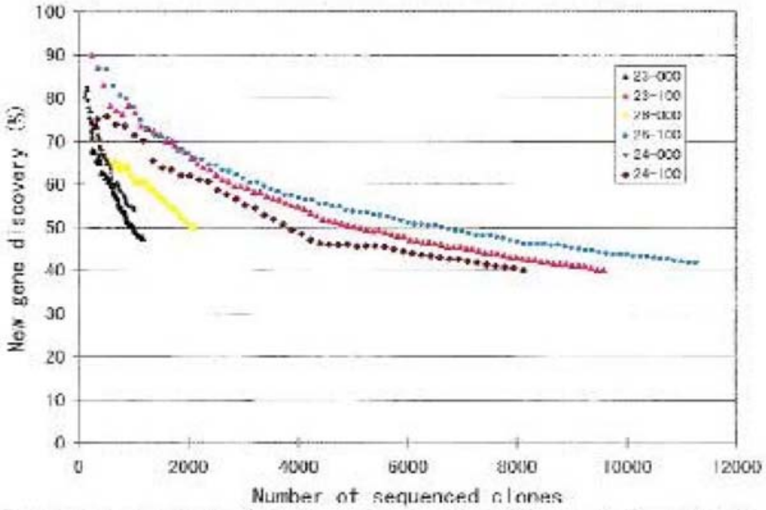
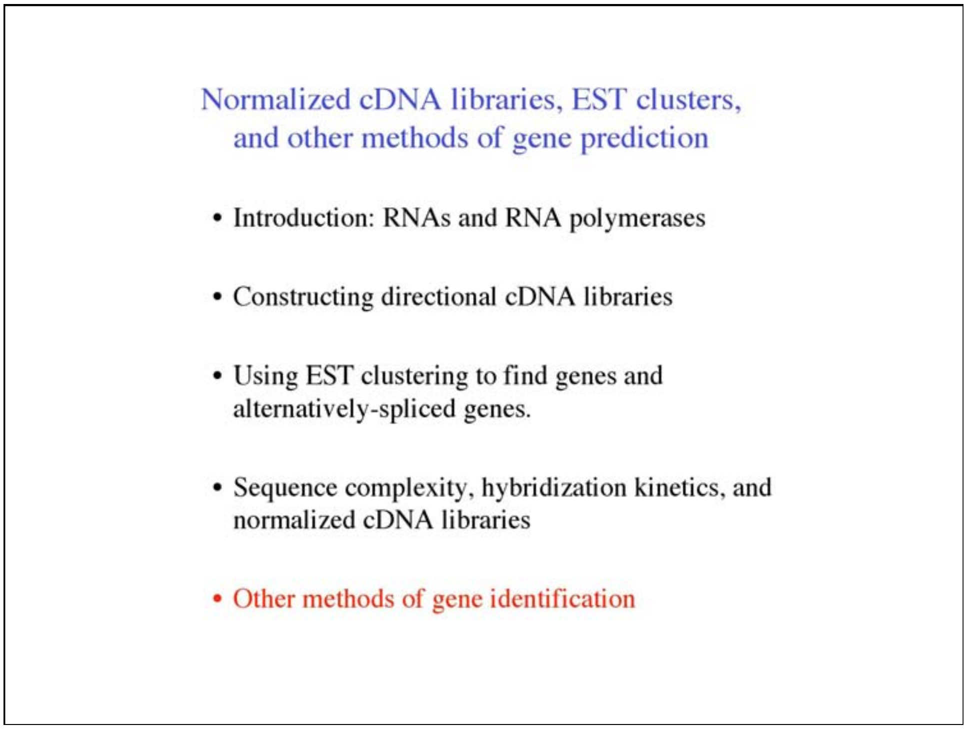


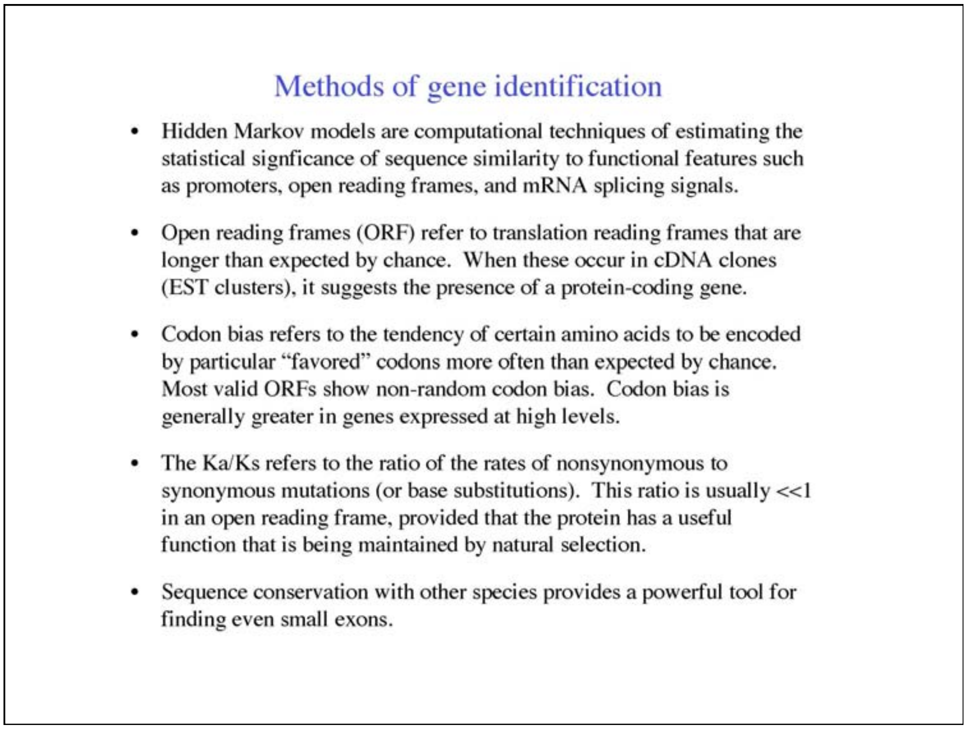
Figure 4 Sequencing redundancy (or the decrease in new gene discovery) increases sharply in standard cDNA libraries (-000 libraries), but in normalized/subtracted full-length cDNA libraries (-100 libraries), redundancy increases much more slowly. New genes (%) are referred as singleton (%) within a given cDNA library.

Carninci et al. (2000) Genome Res. 10, 1617-1630



Normalized cDNA libraries, EST clusters, and other methods of gene prediction

- Introduction: RNAs and RNA polymerases
- Constructing directional cDNA libraries
- Using EST clustering to find genes and alternatively-spliced genes.
- Sequence complexity, hybridization kinetics, and normalized cDNA libraries
- Other methods of gene identification



Methods of gene identification

- Hidden Markov models are computational techniques of estimating the statistical signficance of sequence similarity to functional features such as promoters, open reading frames, and mRNA splicing signals.
- Open reading frames (ORF) refer to translation reading frames that are longer than expected by chance. When these occur in cDNA clones (EST clusters), it suggests the presence of a protein-coding gene.
- Codon bias refers to the tendency of certain amino acids to be encoded by particular "favored" codons more often than expected by chance. Most valid ORFs show non-random codon bias. Codon bias is generally greater in genes expressed at high levels.
- The Ka/Ks refers to the ratio of the rates of nonsynonymous to synonymous mutations (or base substitutions). This ratio is usually <<1 in an open reading frame, provided that the protein has a useful function that is being maintained by natural selection.
- Sequence conservation with other species provides a powerful tool for finding even small exons.

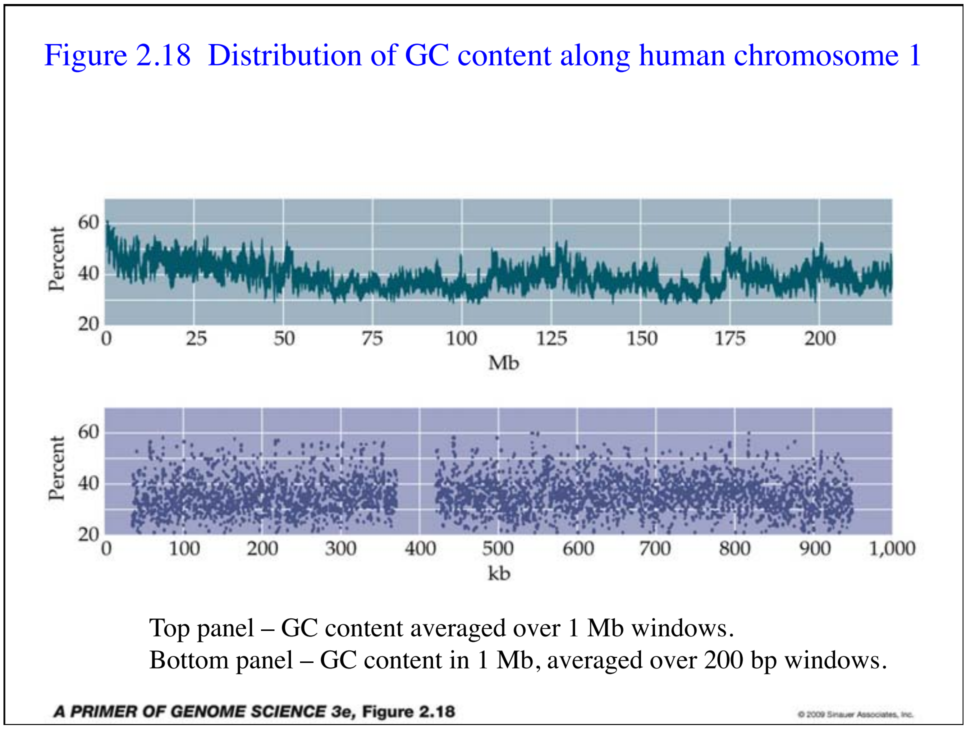
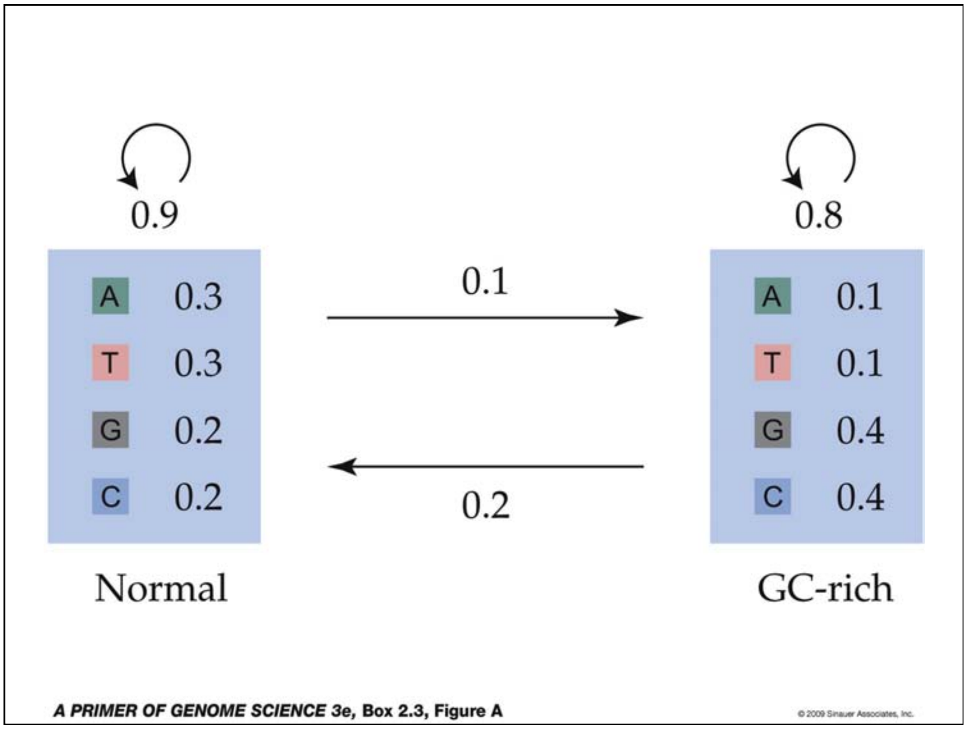


Figure 2.18 Distribution of GC content along human chromosome 1

Top panel – GC content averaged over 1 Mb windows.
Bottom panel – GC content in 1 Mb, averaged over 200 bp windows.

A PRIMER OF GENOME SCIENCE 3e, Figure 2.18



0.9

	Normal
A	0.3
T	0.3
G	0.2
C	0.2

Normal

0.1

0.2

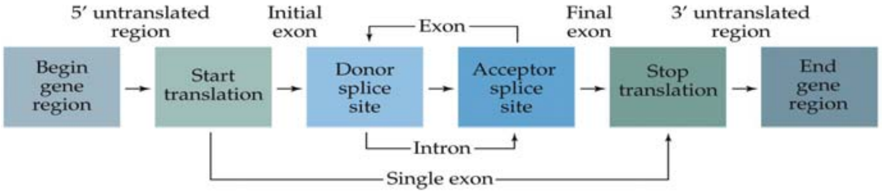
0.8

	GC-rich
A	0.1
T	0.1
G	0.4
C	0.4

GC-rich

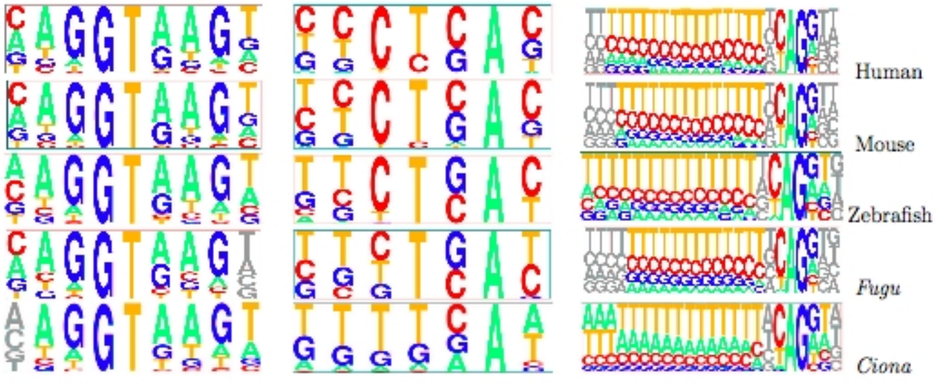
A PRIMER OF GENOME SCIENCE 3e, Box 2.3, Figure A

Box 2.3, Figure B Schematic of the hidden states included in an HMM



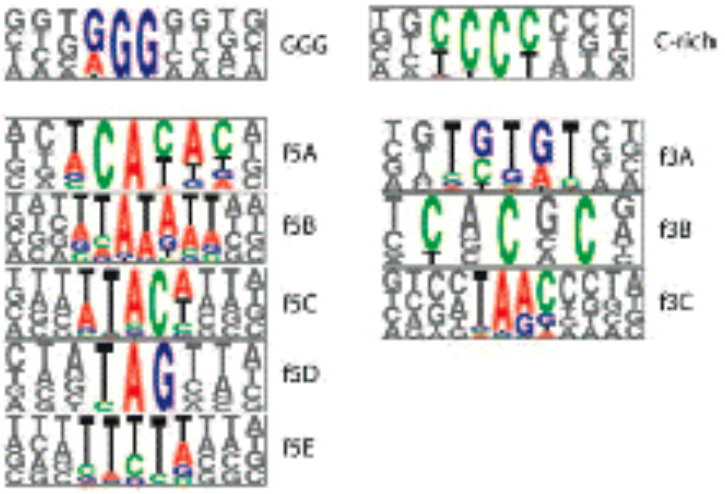
A PRIMER OF GENOME SCIENCE 3e, Box 2.3, Figure B

RNA splicing signals near the splice and lariat sites are moderately conserved in vertebrates, but difficult to identify from sequence information alone



Yeo, G. *et al*. (2004) *Proc. Natl. Acad. Sci. USA* **101**, 15700-15705.

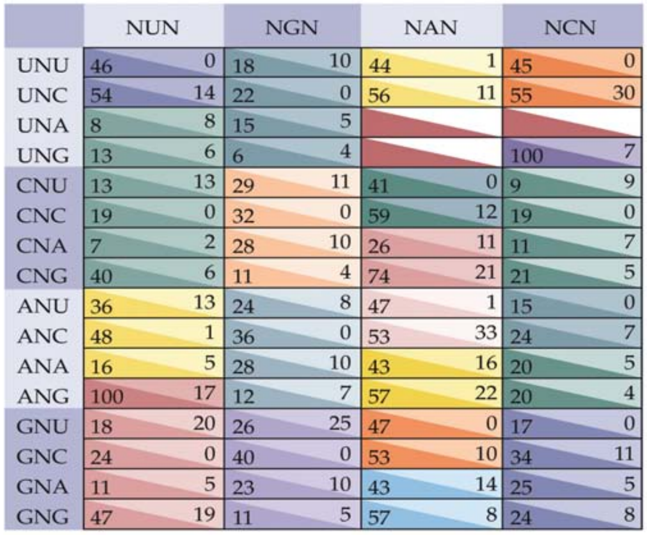
Splice enhancing signals are not conserved between fish and mammals



Yeo, G. *et al*. (2004) *Proc. Natl. Acad. Sci. USA* **101**, 15700-15705.

Figure 2.16 Transfer RNA content in the human genome

(A)

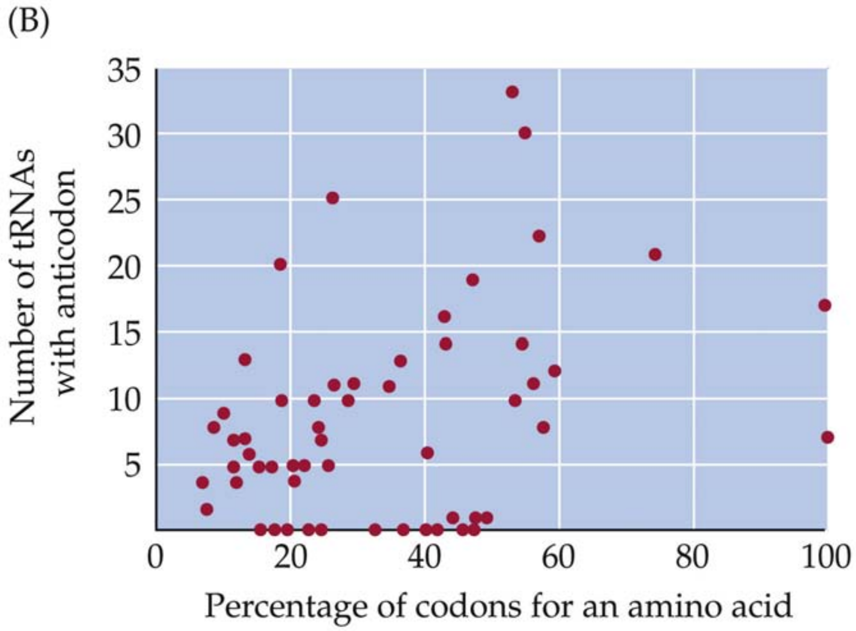


	NUN		NGN		NAN		NCN	
UNU	46	0	18	10	44	1	45	0
UNC	54	14	22	0	56	11	55	30
UNA	8	8	15	5				
UNG	13	6	6	4			100	7
CNU	13	13	29	11	41	0	9	9
CNC	19	0	32	0	59	12	19	0
CNA	7	2	28	10	26	11	11	7
CNG	40	6	11	4	74	21	21	5
ANU	36	13	24	8	47	1	15	0
ANC	48	1	36	0	53	33	24	7
ANA	16	5	28	10	43	16	20	5
ANG	100	17	12	7	57	22	20	4
GNU	18	20	26	25	47	0	17	0
GNC	24	0	40	0	53	10	34	11
GNA	11	5	23	10	43	14	25	5
GNG	47	19	11	5	57	8	24	8

***A PRIMER OF GENOME SCIENCE 3e*, Figure 2.16 (Part 1)**

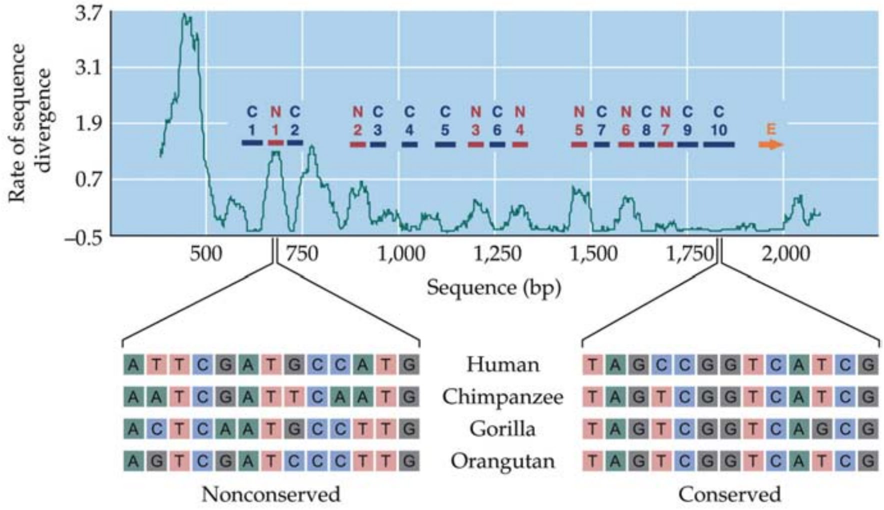
(left) percent codons for each amino acid (color). (right) percent of tRNAs in genome.

Figure 2.16 Transfer RNA content in the human genome



***A PRIMER OF GENOME SCIENCE 3e*, Figure 2.16 (Part 2)**

Multiple sequence alignment ("phylogenetic shadowing") can be very helpful in gene finding.



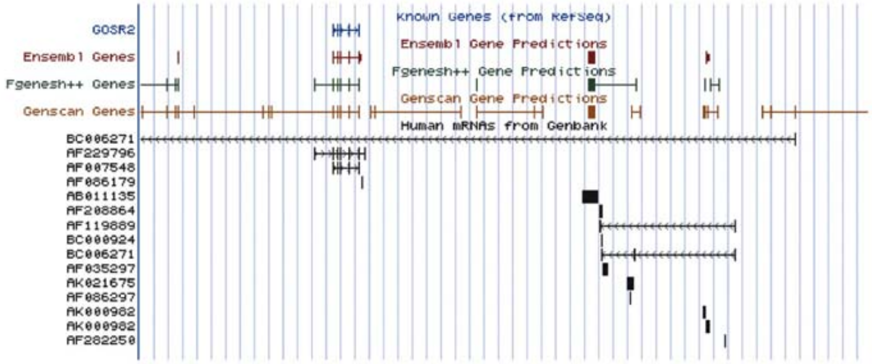
A	T	T	C	G	A	T	G	C	C	A	T	G	Human	T	A	G	C	C	G	G	T	C	A	T	C	G
A	A	T	C	G	A	T	T	C	A	A	T	G	Chimpanzee	T	A	G	T	C	G	G	T	C	A	T	C	G
A	C	T	C	A	A	T	G	C	C	T	T	G	Gorilla	T	A	G	T	C	G	G	T	C	A	G	C	G
A	G	T	C	G	A	T	C	C	C	T	T	G	Orangutan	T	A	G	T	C	G	G	T	C	A	T	C	G

Nonconserved

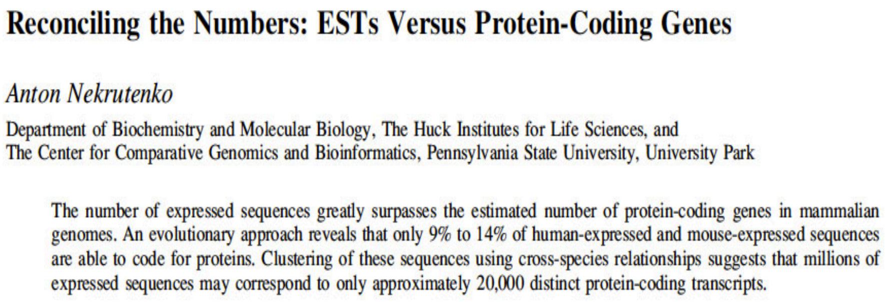
Conserved

***A PRIMER OF GENOME SCIENCE 3e*, Figure 2.15**

Three different gene prediction methods (Ensembl, Fgenesh, and Genescan) were used on a region of chromosome 17 that includes The GOSR2 gene. The black images below indicate matching cDNA/EST sequences.



A PRIMER OF GENOME SCIENCE 3e, Box 2.3, Figure C



Reconciling the Numbers: ESTs Versus Protein-Coding Genes

Anton Nekrutenko

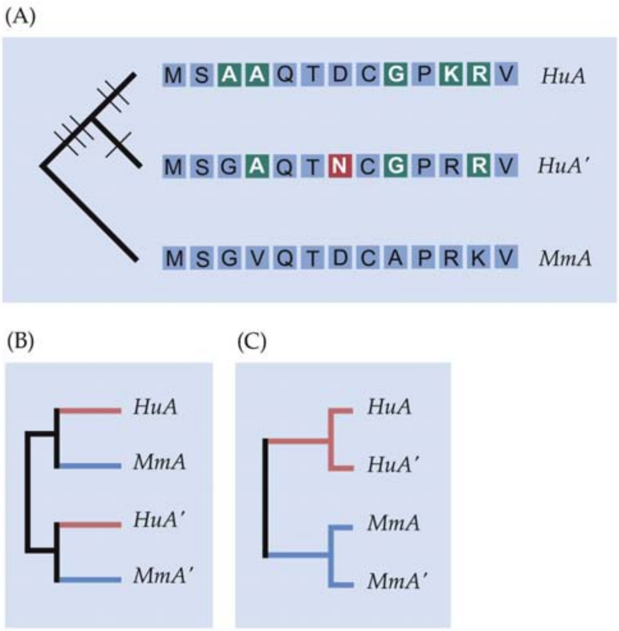
Department of Biochemistry and Molecular Biology, The Huck Institutes for Life Sciences, and
The Center for Comparative Genomics and Bioinformatics, Pennsylvania State University, University Park

The number of expressed sequences greatly surpasses the estimated number of protein-coding genes in mammalian genomes. An evolutionary approach reveals that only 9% to 14% of human-expressed and mouse-expressed sequences are able to code for proteins. Clustering of these sequences using cross-species relationships suggests that millions of expressed sequences may correspond to only approximately 20,000 distinct protein-coding transcripts.

Quality control

- The majority of EST clusters, particularly singletons, do not appear to correspond to valid protein-coding genes, by several criteria of evolutionary conservation (Nekrutenko 2004).
- ORF comparison has been very helpful, but may have nearly run its course for the human genome (Brent).
- Additional refinements may involve genes expressed at very low levels, or particular tissues/stages of development.
- Those may require RT-PCR, which ironically requires some previous knowledge of the transcript structure and sequence.

Orthologs versus paralogs

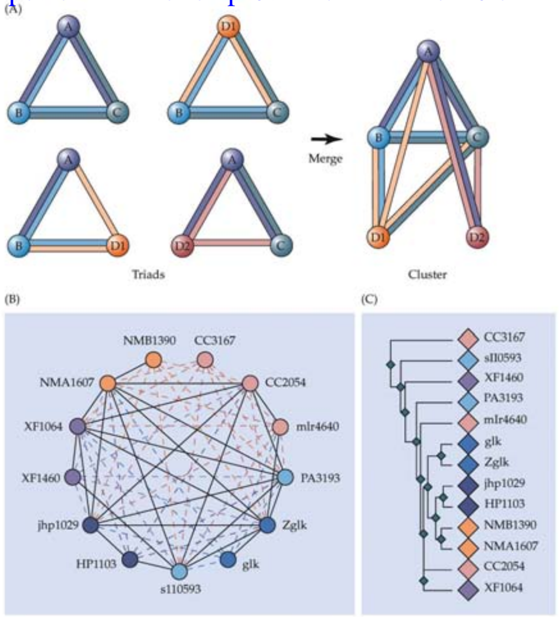

(A)
M S A A Q T D C G P K R V HuA
M S G A Q T N C G P R R V HuA′
M S G V Q T D C A P R K V MmA
(B)
HuA
MmA
HuA′
MmA′
(C)
HuA
HuA′
MmA
MmA′

A PRIMER OF GENOME SCIENCE 3e, **Figure 2.21**

Clusters of orthologous genes, sequence motifs, and gene families

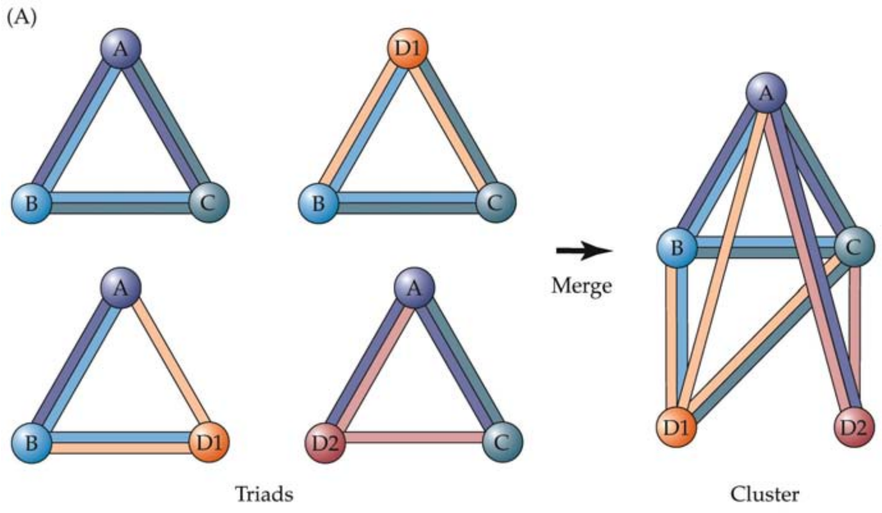
- Gene families are groups of genes that share descent from a common ancestor gene. A gene family may include several genes in one species (paralogs) or corresponding genes in different species (orthologs). Gene families (particularly orthologs) tend to maintain similar *biochemical* functions over long periods of evolutionary time, although details (such as expression patterns) may differ (particularly in paralogs).
- Clusters of orthologous genes (COGs) are discovered by automated sequence alignment between species. This procedure is supposed to facilitate the discovery of orthologs, and hence the inference of gene function. COGs are not necessarily reliable, but can be quite valuable in suggesting putative gene functions.
- Protein sequence motifs are relatively short conserved protein sequences that occur in many gene families. Although the secondary structure and basic function of sequence motifs are conserved, they can be deployed in different gene families (as a result of recombination events in evolutionary history).

Clusters of orthologous genes are assembled by merging triads of genes from different species that are reciprocal best matches to each other.

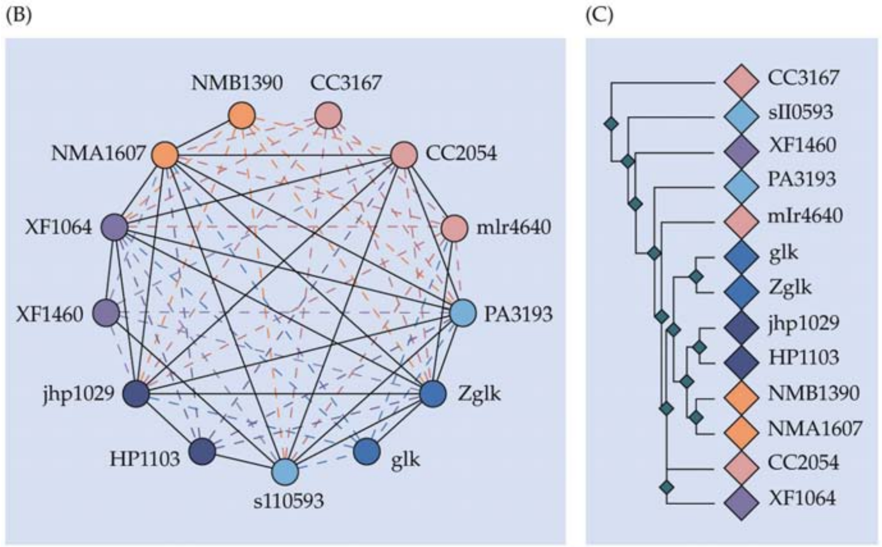


***A PRIMER OF GENOME SCIENCE* 3e, Figure 2.20**

© 2009 Sinauer Associates, Inc.

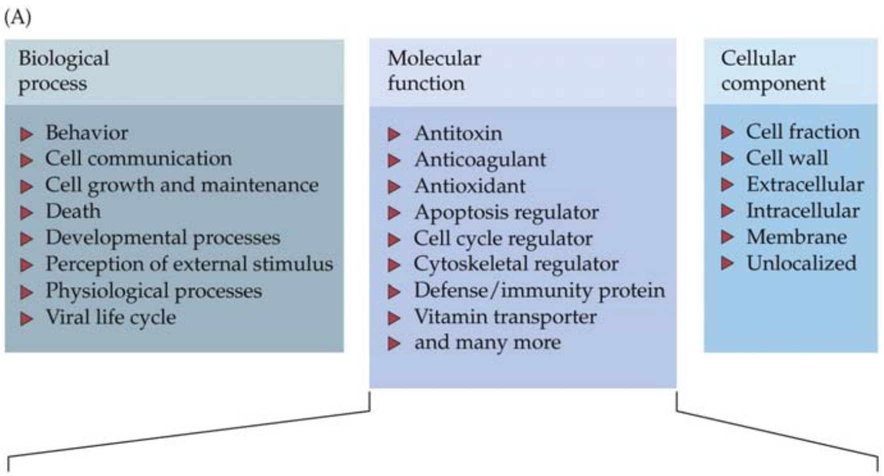


***A PRIMER OF GENOME SCIENCE 3e*, Figure 2.20 (Part 1)**



***A PRIMER OF GENOME SCIENCE 3e*, Figure 2.20 (Part 2)**

Gene ontologies are assembled from various lines of evidence in multiple species.



(A)

Biological process

- ▶ Behavior
- ▶ Cell communication
- ▶ Cell growth and maintenance
- ▶ Death
- ▶ Developmental processes
- ▶ Perception of external stimulus
- ▶ Physiological processes
- ▶ Viral life cycle

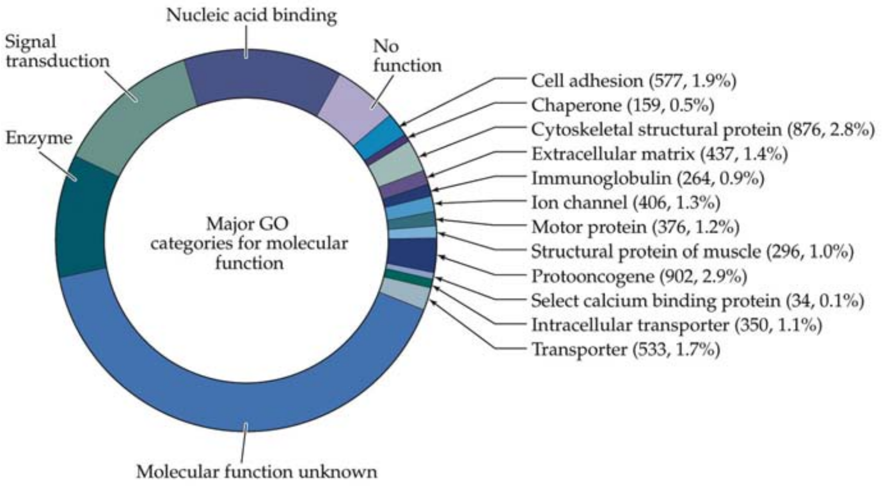
Molecular function

- ▶ Antitoxin
- ▶ Anticoagulant
- ▶ Antioxidant
- ▶ Apoptosis regulator
- ▶ Cell cycle regulator
- ▶ Cytoskeletal regulator
- ▶ Defense/immunity protein
- ▶ Vitamin transporter
- ▶ and many more

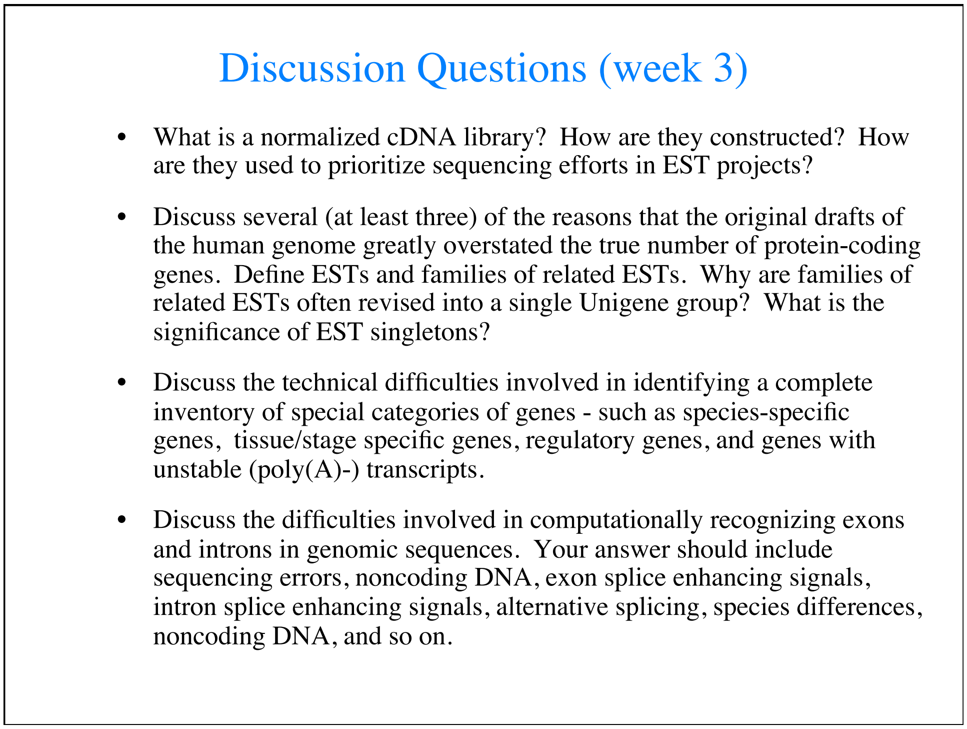
Cellular component

- ▶ Cell fraction
- ▶ Cell wall
- ▶ Extracellular
- ▶ Intracellular
- ▶ Membrane
- ▶ Unlocalized

***A PRIMER OF GENOME SCIENCE 3e*, Figure 2.23 (Part 1)**



***A PRIMER OF GENOME SCIENCE 3e*, Figure 2.23 (Part 2)**



Discussion Questions (week 3)

- What is a normalized cDNA library? How are they constructed? How are they used to prioritize sequencing efforts in EST projects?
- Discuss several (at least three) of the reasons that the original drafts of the human genome greatly overstated the true number of protein-coding genes. Define ESTs and families of related ESTs. Why are families of related ESTs often revised into a single Unigene group? What is the significance of EST singletons?
- Discuss the technical difficulties involved in identifying a complete inventory of special categories of genes - such as species-specific genes, tissue/stage specific genes, regulatory genes, and genes with unstable (poly(A)-) transcripts.
- Discuss the difficulties involved in computationally recognizing exons and introns in genomic sequences. Your answer should include sequencing errors, noncoding DNA, exon splice enhancing signals, intron splice enhancing signals, alternative splicing, species differences, noncoding DNA, and so on.