

Genome Sequences

Sequenced libraries of cDNA clones: ESTs
Genomic DNA sequences

Abundance and complexity of mRNA

- Kinetics of hybridization of labeled cDNA to an excess of mRNA allows the determination of complexity and abundance of mRNA.
- Analogous to strategy for determining complexity and repetition frequency of genomic DNA
- First-order kinetics since the mRNA is in large excess over the labeled cDNA

R_0 = original [RNA],

will not change measurably during renaturation

$$k = \frac{\ln 2}{R_0 t_{1/2}}$$

$$R_0 t_{1/2} \propto N \text{ of RNA}$$

Example of mRNA from chick oviduct,

Compo- Nent	frac- tion	$R_0 t_{1/2}^{\text{mix}}$	$R_0 t_{1/2}^{\text{pure}}$	N (nt)	# mRNAs	Abundance
1 st	0.50	0.0015	0.00075	2,000	1	120,000
2 nd	0.15	0.04	0.006	15,000	7-8	4,800
3 rd	0.35	30	10.5	2.6×10^7	13,000	6-7

Normalized cDNA libraries

- Goal: obtain cDNA libraries with roughly comparable representation of *every* mRNA from a tissue, including the *rare* mRNAs.
- Hybridize the cDNA back to the template mRNA to a sufficiently high $R_0 t$
 - Most of the abundant cDNA is in duplex with the mRNA
 - Essentially all the rare cDNA is single-stranded
- Collect the single-stranded cDNA and clone into a vector.

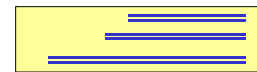
ESTs from normalized cDNA libraries

- EST = Expressed Sequence Tag
- A short DNA sequence (a "tag") from a cDNA clone (hence it is expressed)
- Large-scale projects: sequence one or both ends from each clone in the normalized libraries
- Have generated 2,274,459 ESTs (as of Sept. 08, 2000).
- The database of ESTs provides information on most (?) mammalian genes - even the unidentified ones!

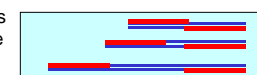
cDNA clones and ESTs

mRNA 5' UTR Protein coding 3' UTR
5' AAAAA 3'

Duplex inserts in cDNA clones



ESTs are sequences from each end of the cDNA inserts



Unigene cluster is a group of overlapping ESTs, likely from one gene



Genome sequences available

- >28 eubacteria
- 6 archaea
- 1 fungus: yeast *Saccharomyces cerevisiae*
- 1 protozoan: *Plasmodium falciparum*
- 1 worm, nematode *Caenorhabditis elegans*
- 1 insect: *Drosophila melanogaster*
- 2 mammals: *Homo sapiens*, *Mus domesticus*
- 2 plants: *Arabidopsis*, rice

Genome sequencing after mapping

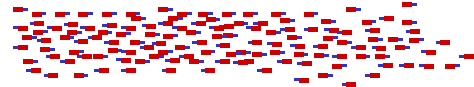
- Libraries of BACs have been screened and mapped to find overlapping arrays of contiguous clones (contigs)
 - E.g. find common restriction fragments in collections of clones
- Ends of the BACs are sequenced to provide markers through the genome
- Mapped contigs are then sequenced, using a combination of shotgun sequencing and directed sequencing

Shotgun sequencing of whole genomes

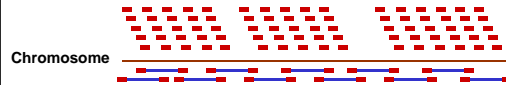
- Break total genomic DNA into small pieces (around 1000 bp in size) and clone into plasmids
- Sequence about 500 bp from each end.
- Use sequence alignments to assemble a final sequence.
- Requires that each bp be determined multiple times
 - about 3x coverage for small genomes (1-5 million bp)
 - about 10x coverage for large genomes (> 1 billion bp)

Shotgun sequencing and assembly

Sequence the ends of a huge number of small insert plasmids:



Align the sequences into contiguous assemblies (contigs):

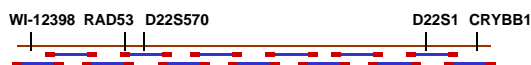


The end sequences from mapped BAC contigs are used to assemble longer sequences from complex genomes. Gaps must be filled by directed sequencing.

Directed sequencing of BAC contigs

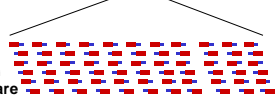
Chromosome 22 (part)

Anonymous markers and known genes mapped:



BAC contig, ends sequenced

Mapped BACs are broken into small pieces, which are shot-gun sequenced and assembled.



Gaps must be filled by alternate approaches, e.g. directed PCR.

Identifying genes in genomic DNA sequences

- Identical to a known gene in the same species
- Highly significant match to a known gene in another species.
- Highly significant match to a spliced EST from the same or related species
- Parts of a gene may match portions of known genes at lower % identity
 - Assign potential functional domains by conserved motifs, e.g. protein kinase, ATPase, transmembrane domain
- Use **sequence alignment programs**

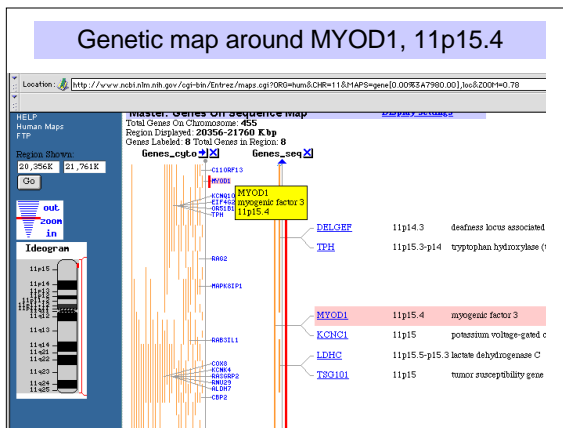
Computational tools for predicting genes and important sequences

- Gene prediction
 - Properties of coding regions (e.g. Genscan)
 - Open reading frames
 - Splice sites, regulatory signals
 - Codon usage characteristic of a particular organism
 - Alignments
 - Interspecies (human vs. mouse or fish)
 - Align to cDNAs
 - Both: e.g. Twinscan
- Regulatory elements
 - Interspecies alignments
 - Matches to transcription factor binding sites

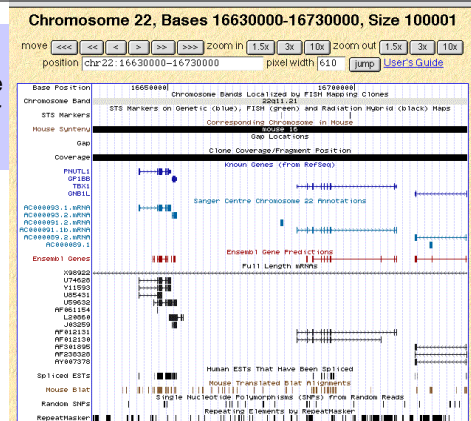
Databases for genomic analysis

- Nucleic acid sequences
 - genomic and mRNA, including ESTs
- Protein sequences
- Protein structures
- Genetic and physical maps
- Organism-specific databases
- MedLine (PubMed)
- Online Mendelian Inheritance in Man (OMIM)

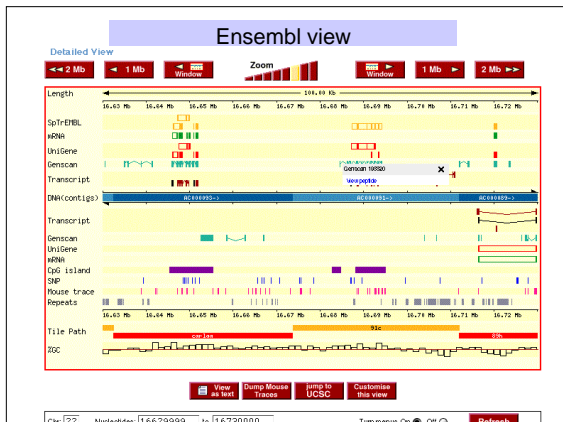
Genetic map around MYOD1, 11p15.4



Human Genome Browser view



Ensembl view

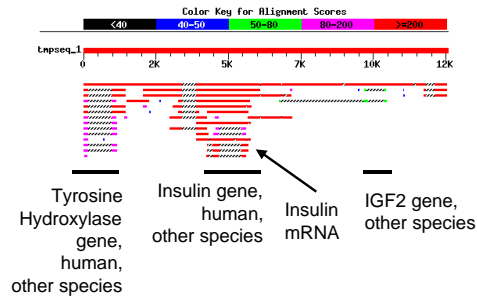


Programs for sequence analysis

- BLAST to search rapidly through sequence databases
- PipMaker (to align 2 genomic DNA sequences)
- Gene finding by ab initio methods (GenScan, GRAIL, etc.)
- RepeatMasker

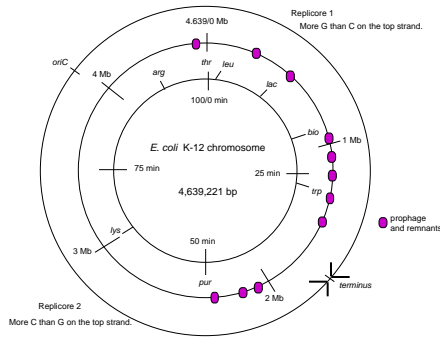
Results of BLAST search, *INS* vs. nr

L15440 (*INS* and flanking genes) vs. nr database



Large scale genome organization

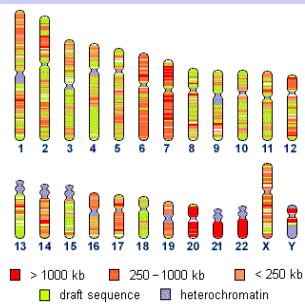
E. coli genome with sequence features



New insights for *E. coli*

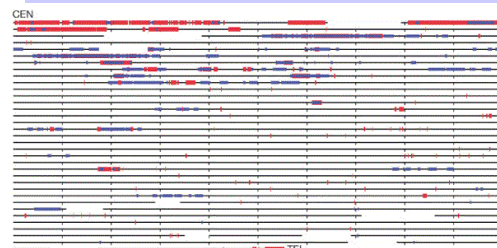
- Organization with respect to direction of replication:
 - Transcription of most genes
 - G>C on “top” strand (leading strand in replication)
 - Recombination hotspot Chi: more abundant on leading strand
- At least 18 families of repeated DNA
 - Long *Rhs* elements: 5.7 to 9.6 kb, 5 copies
 - Short REP elements: 0.04 kb, 581 copies
 - Prophage; transposable elements

Human chromosomes sequenced



<http://www.ncbi.nlm.nih.gov/genome/seq/>

Segmental duplications are common



The size and location of intrachromosomal (blue) and interchromosomal (red) duplications are depicted for chromosome 22q, using the PARASIGHT computer program (Bailey and Eichler, unpublished). Each horizontal line represents 1 Mb (ticks, 100-kb intervals). Pairwise alignments with > 90% nucleotide identity and > 1 kb long are shown.

Comparative Genomics

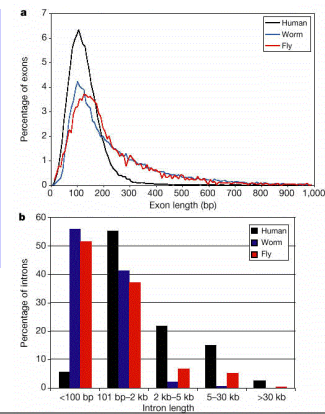
Genome size

- Bacterial genome size range:
 - 0.58 million bp (Mb), 467 genes (*Mycoplasma genitalium*)
 - 4.64 Mb, 4289 genes (*Escherichia coli*)
- Yeast *S. cerevisiae*: 12 Mb, 6241 genes
 - Only 2.6 X that of *E. coli*.
- *Caenorhabditis elegans*: 97 Mb; 18,424 genes
- *Drosophila melanogaster*: 180 Mb; 13,601 genes
 - ~120 Mb euchromatic (sequenced)
- *Homo sapiens*: ~3200 Mb; ~30,000 genes

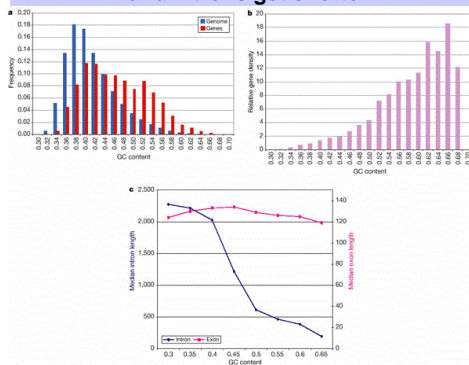
Gene size and number

- Average gene size:
 - Bacteria: 1100 bp
 - Yeast: ~1200 bp
 - Worm: ~5000 bp
 - Human: ~27,000 bp (range up to 2.4 Mb)
- Distance between genes:
 - Bacteria: 118 bp
 - Yeast: ~700 bp
 - Human: range from overlapping to ~1 Mb
- Exons sizes similar for worm, fly, human
 - Exons commonly ~125 bp
 - Typical length of coding seq for gene: 1300-1400 bp
- Intron sizes differ
 - Humans have substantially more very long introns > 5 kb

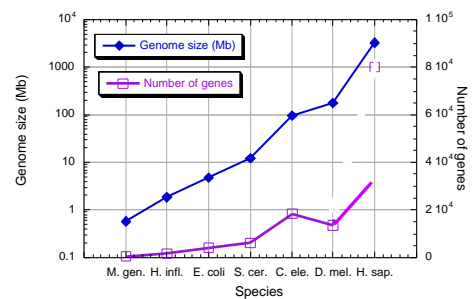
Compared to worm and fly, human has shorter exons and longer introns on the extremes of the distribution



As G+C increases, gene density increases and introns get shorter



Genome size increases exponentially, but not number of genes



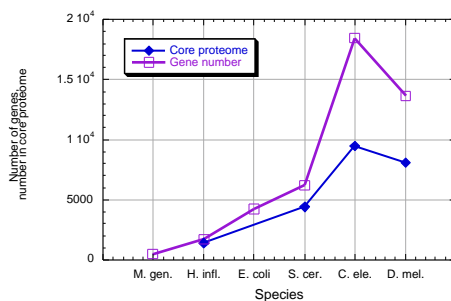
Paralogous genes

- Genes that are similar because of descent from a common ancestor are **homologous**.
- Homologous genes that have diverged after speciation are **orthologous**.
- Homologous genes that have diverged after duplication are **paralogous**.
- One can identify **paralogous groups** of genes encoding proteins of similar but not identical function in a species
 - E.g. ABC transporters: 80 members in *E. coli*

Core proteomes vary little in size

- Proteome: all the proteins encoded in a genome
- Core proteome
 - Count each group of paralogous proteins only once
 - Number of distinct protein families in each organism
- | Species | Number of genes | Core proteome |
|---------------|-----------------|---------------|
| – Haemophilus | 1709 | 1425 |
| – Yeast | 6241 | 4383 |
| – Worm | 18424 | 9453 |
| – Fly | 13601 | 8065 |

Little change in core proteome size in eukaryotes



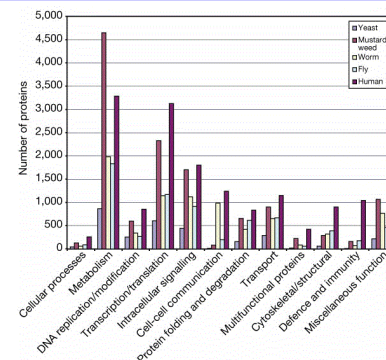
Core proteomes are conserved

- Many of the proteins in the core proteomes are shared among eukaryotes
 - 30% of fly genes have orthologs in worm
 - 20% of fly genes have orthologs in both worm and yeast
 - 50% of fly genes have likely orthologs in mammals
- Function of proteins in flies (and worms and yeast) provides strong indicators of function in humans
- Flies have orthologs to 177 of the 289 human disease genes
- Rubin et al. (2000) Science 287: 2204.

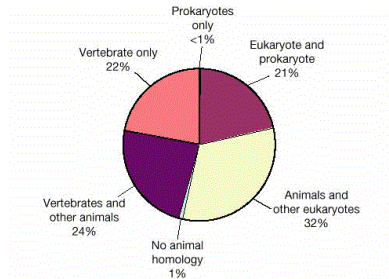
Types of information one can get

- **Sequences** of all the genes
- **Functions** of many/all the genes
- Sequences **regulating** gene expression
 - Promoters, enhancers, etc.
- Sequences needed for genome **maintenance** (?)
 - Regulation of the replicon, telomere maintenance, etc.
- **Large-scale structure** of the genome

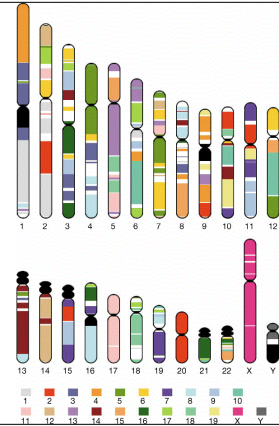
Functional categories in eukaryotic proteomes



Distribution of the homologues of the predicted human proteins



Conserved segments in the human and mouse genome



OTC problems illustrate use the Web resources from genome sequencing

We used arginine biosynthesis to illustrate complementation analysis and construction of a pathway. The steps involved in arginine synthesis are also part of the urea cycle. One of the enzymes catalyzes the formation of citrulline from carbamoyl phosphate and ornithine. Let's find out more about this enzyme, called ornithine transcarbamoylase, or OTC.

Use your favorite Web browser to go to the URL for NCBI (National Center for Biotechnology Information).

<http://www.ncbi.nlm.nih.gov/>

Click on the Entrez button. Entrez provides a portal to many types of information at this server. Let's start with DNA and protein sequences.

Click on the Nucleotides button. Enter "X00210" and press the Search button.