

# CSE 427

# Computational Biology

<http://courses.cs.washington.edu/courses/cse427>

Larry Ruzzo  
Winter 2014



UW CSE Computational Biology Group

He who asks is a fool for five minutes, but he who does not ask remains a fool forever.

-- Chinese Proverb



# Today

Admin

Why Comp Bio?

The world's shortest Intro. to Mol. Bio.



**Admin Stuff**

# Course Mechanics & Grading

Web:

<http://courses.cs.washington.edu/courses/cse427>

Reading

In class discussion

Homeworks

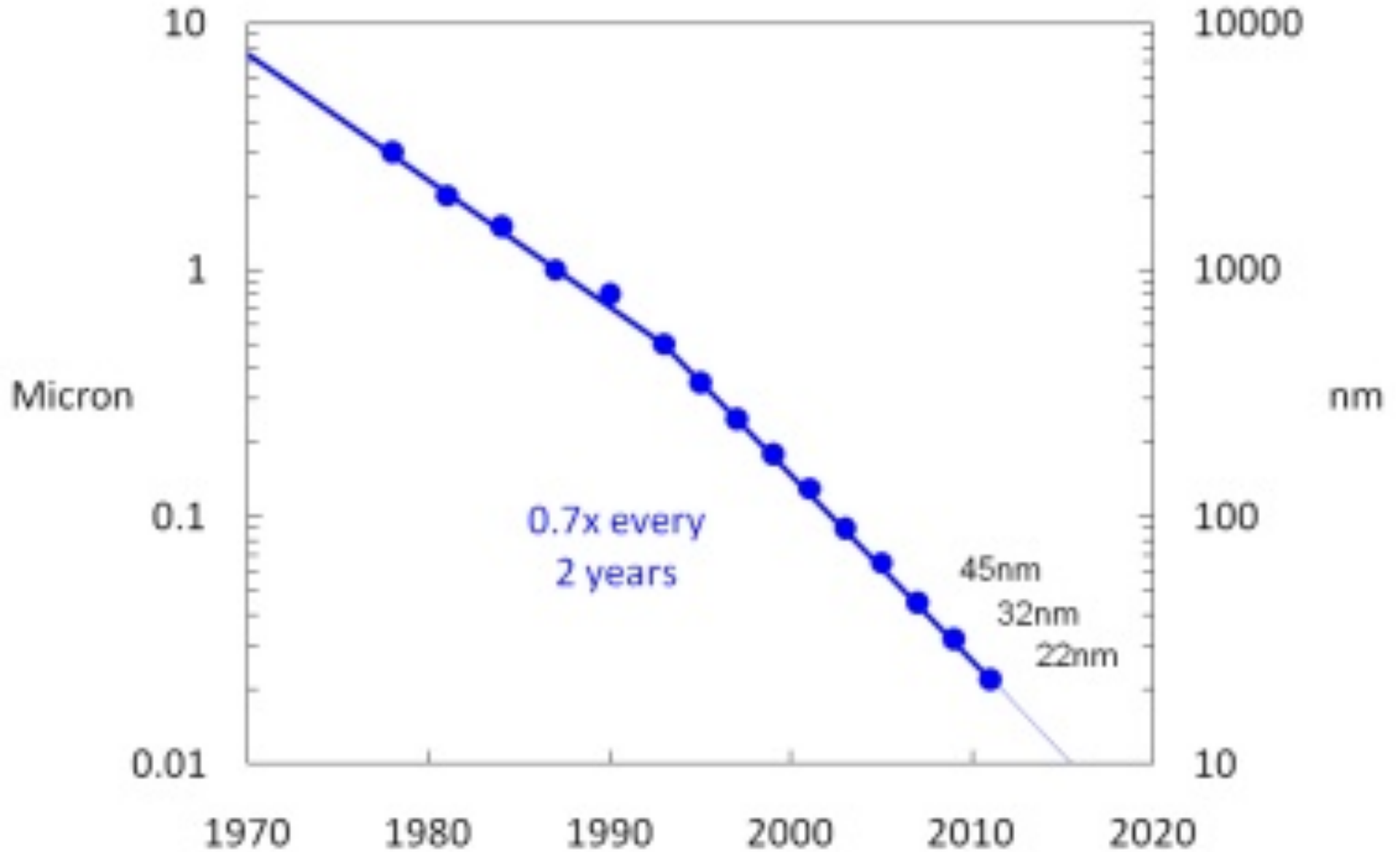
paper exercises & programming

No exams, but possible oversized last homework in lieu of final

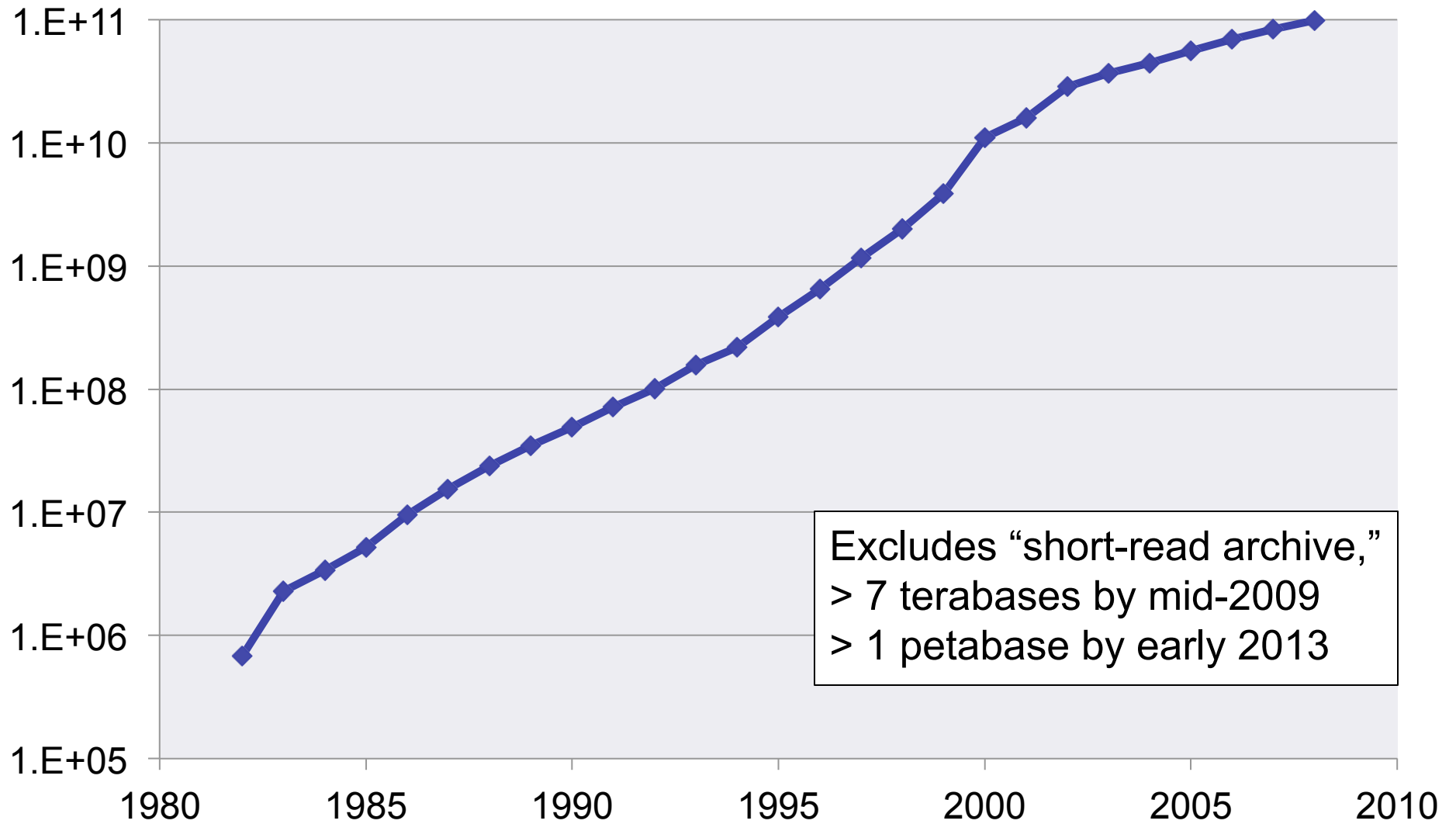
# Background & Motivation



# Feature Size

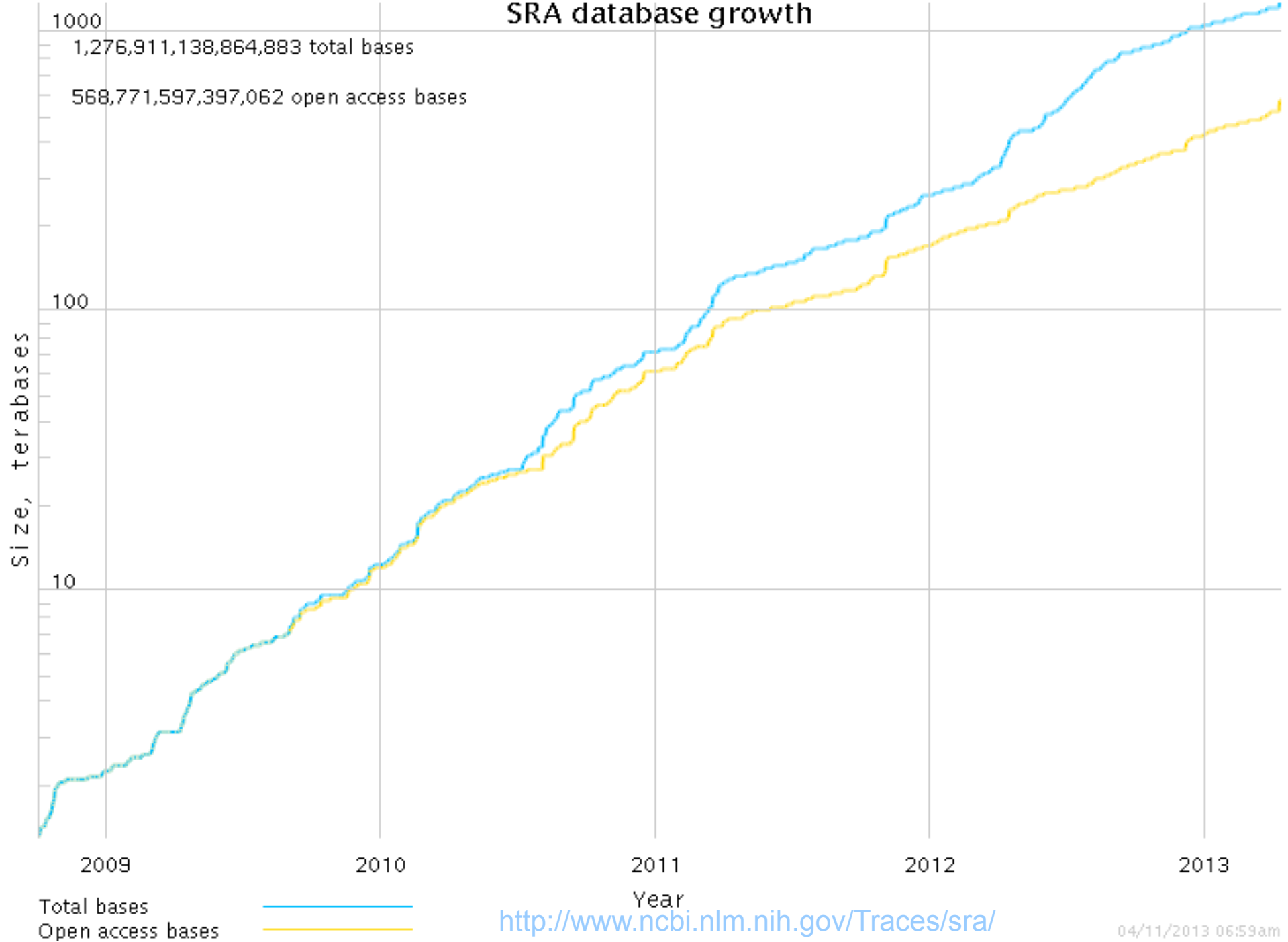


## Growth of GenBank (Base Pairs)

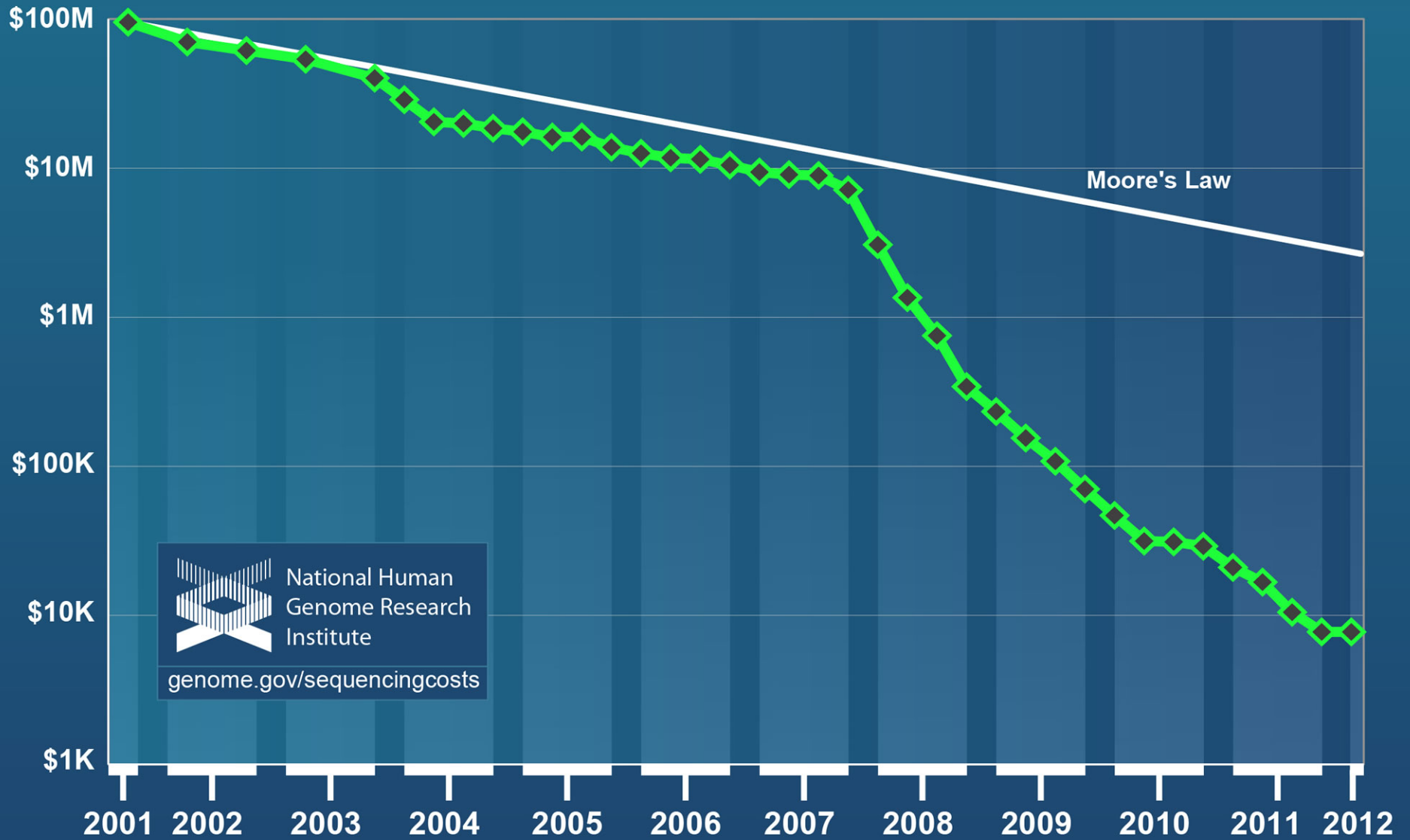


Source: <http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html>

# SRA database growth



# Cost per Genome

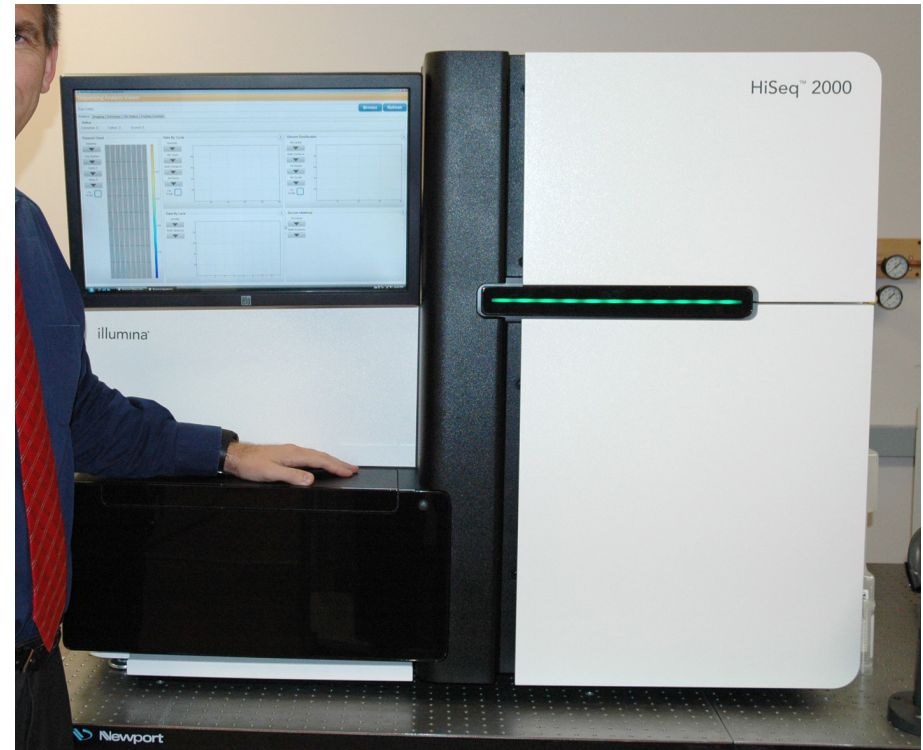


 National Human  
Genome Research  
Institute  
[genome.gov/sequencingcosts](http://genome.gov/sequencingcosts)



# Modern DNA Sequencing

A table-top box the size of your oven (but costs a bit more ... ;-)  
can generate  
~100 billion BP of DNA seq/day; i.e.  
= 2008 genbank,  
= 30x your genome



# The Human Genome Project

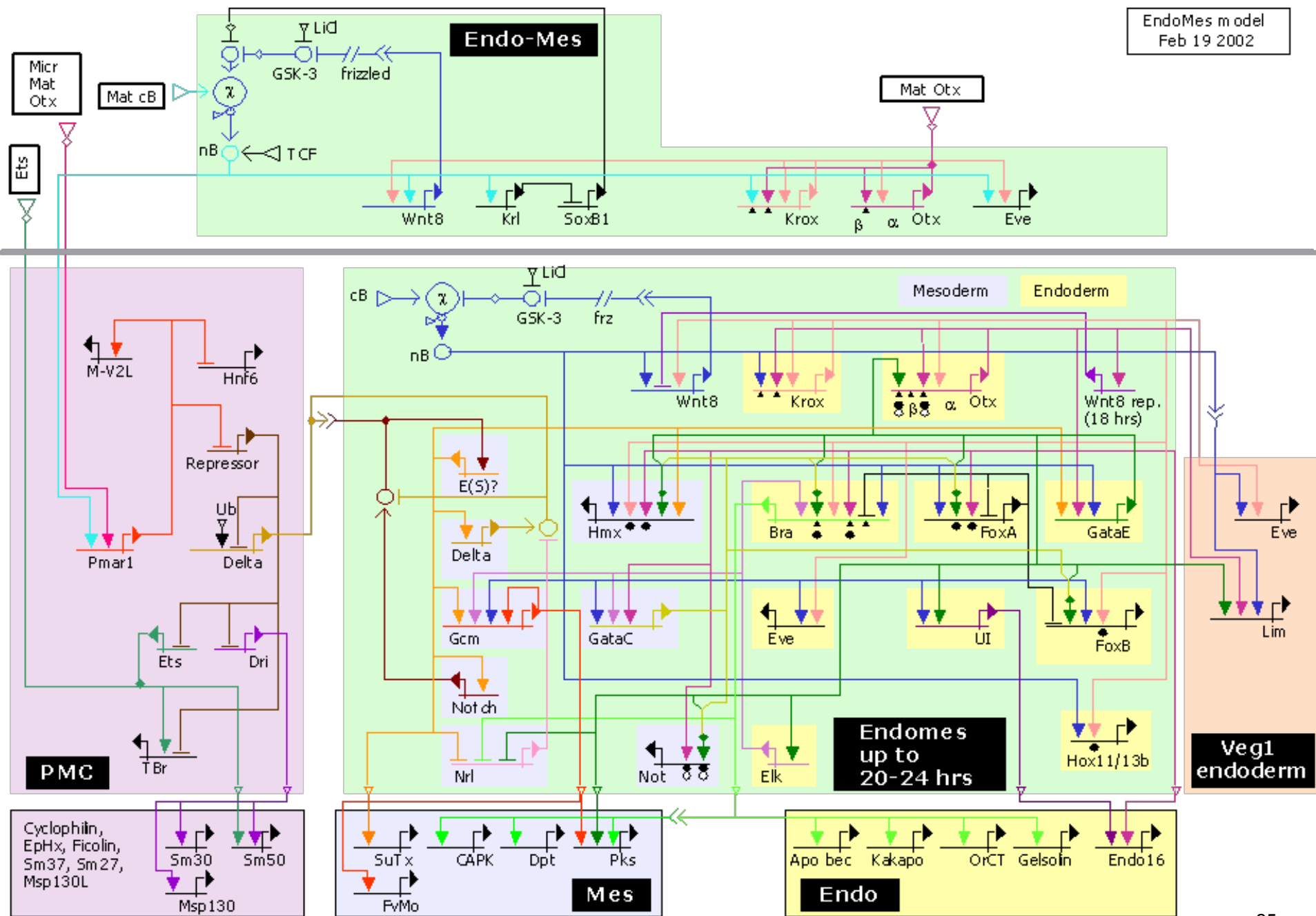
```
1 gagcccggcc cgggggacgg gcggcgggat agcgggaccc cggcgcggcg gtgcgcttca
61 gggcgcagcg gcggccgcag accgagcccc gggcgcggca agaggcggcg ggagccggtg
121 gcggctcggc atcatgctc gagggcgtct gctggagatc gccctgggat ttaccgtgct
181 tttagcgtcc tacacgagcc atggggcgga cgccaatttg gaggctggga acgtgaagga
241 aaccagagcc agtcgggcca agagaagagg cgggtggagga cacgacgcgc ttaaaggacc
301 caatgtctgt ggatcacgtt ataatgctta ctgttgccct ggatggaaaa ccttacctgg
361 cggaaatcag tgtattgtcc ccatttgccg gcattcctgt ggggatggat tttgttcgag
421 gccaaatatg tgcacttgcc catctggtca gatagctcct tcctgtggct ccagatccat
481 acaacactgc aatattcgct gtatgaatgg aggtagctgc agtgacgatc actgtctatg
541 ccagaaagga tacataggga ctactgtgg acaacctgtt tgtgaaagtg gctgtctcaa
601 tggaggaagg tgtgtggccc caaatcgatg tgcatgcact tacggattta ctggacccca
661 gtgtgaaaga gattacagga caggcccatg ttttactgtg atcagcaacc agatgtgcca
721 gggacaactc agcgggattg tctgcacaaa acagctctgc tgtgccacag tcggccgagc
781 ctggggccac ccctgtgaga tgtgtcctgc ccagcctcac ccctgccgcc gtggcttcat
841 tccaaatatc cgcacgggag cttgtcaaga tgtggatgaa tgccaggcca tccccgggct
901 ctgtcagggg gaaattgca ttaatactgt tgggtctttt gagtgcaaat gcctgtctgg
961 acacaaactt aatgaagtgt cacaaaaatg tgaagatatt gatgaatgca gcaccattcc
1021 ...
```





The sea urchin *Strongylocentrotus purpuratus*





# Goals

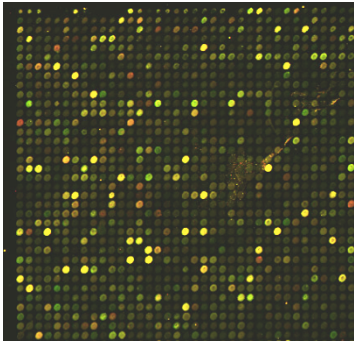
Basic biology

Disease diagnosis/prognosis/treatment

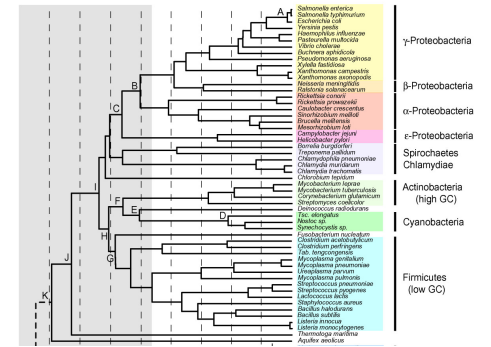
Drug discovery, validation & development

Individualized medicine

...



# “High-Throughput BioTech”

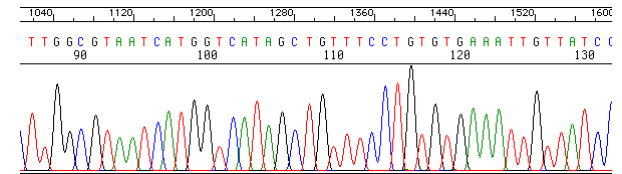
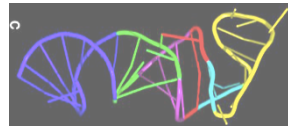


## Sensors

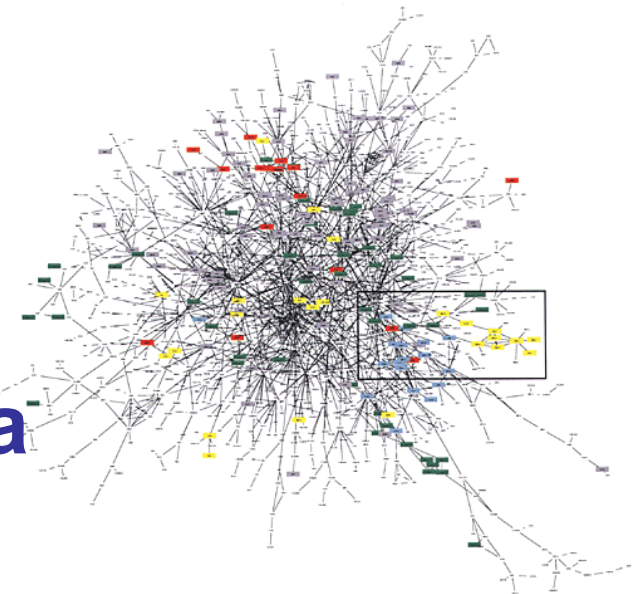
- DNA sequencing
- Microarrays/Gene expression
- Mass Spectrometry/Proteomics
- Protein/protein & DNA/protein interaction

## Controls

- Cloning
- Gene knock out/knock in
- RNAi

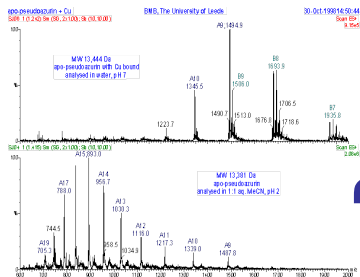


A



*Floods of data*

“Grand Challenge” problems



# What's all the fuss?

The human genome is “finished” ...  
Even if it were, that's only the beginning  
Explosive growth in biological data is  
revolutionizing biology & medicine

“All pre-genomic lab  
techniques are obsolete”

(and computation and mathematics are  
crucial to post-genomic analysis)

# CS Points of Contact & Opportunities

## Scientific visualization

- Gene expression patterns

## Databases

- Integration of disparate, overlapping data sources

- Distributed genome annotation in face of shifting underlying genomic coordinates, individual variation, ...

## AI/NLP/Text Mining

- Information extraction from text with inconsistent nomenclature, indirect interactions, incomplete/inaccurate models, ...

## Machine learning

- System level synthesis of cell behavior from low-level heterogeneous data (DNA seq, gene expression, protein interaction, mass spec,...)

...

## Algorithms



# Computers in biology: Then & now

**Trends in Biochemical Sciences**  
Volume 12 , 1987, Pages 279-280

doi: 10.1016/0960-0804(87)90105-6  
Copyright © 1987 Published by Elsevier Science Ltd.

**Microfile**

## Sequence alignment by word processor

**D. Ross Boswell**

Department of Haematological Medicine, University of Cambridge School of Clinical Medicine, Addenbrooke's Hospital, Cambridge CB2 2QL, UK

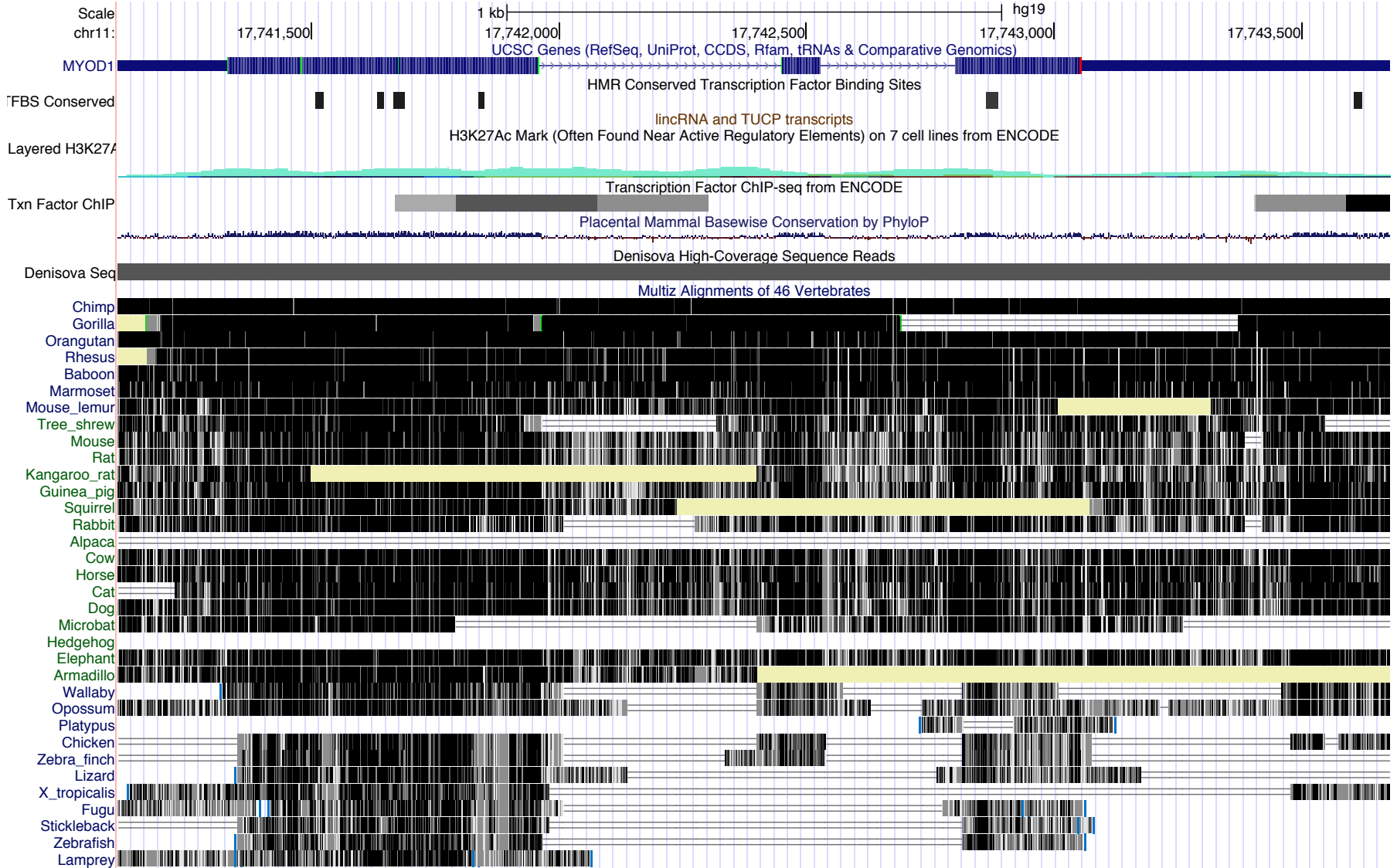
ACGGGTAA

← AC GGTA

# UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

chr11:17,741,110-17,743,678 2,569 bp.



# An Algorithm Example: ncRNAs

The “Central Dogma”:

DNA → messenger RNA → Protein

Last ~5 years:

100s – 1000s of examples of functionally important ncRNAs

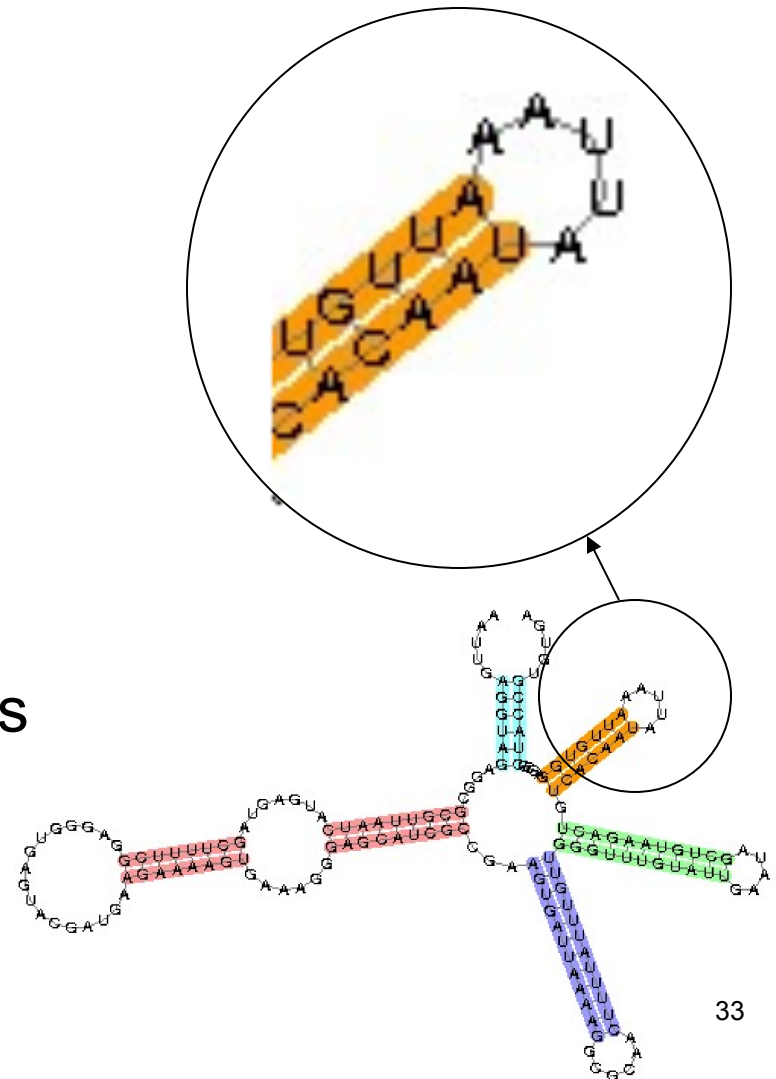
Much harder to find than protein-coding genes

Main method - Covariance Models

≈ stochastic context free grammars

Main problem - Sloooow

$O(nm^4)$



# “Rigorous Filtering” - Z. Weinberg

Convert CM to HMM

(AKA: stochastic CFG to stochastic regular grammar)

Do it so HMM score *always*  $\geq$  CM score

Optimize for most aggressive filtering subject to constraint that score bound maintained

A large convex optimization problem

Filter genome sequence with (fast) HMM, run (slow) CM only on sequences above desired CM threshold; guaranteed not to miss anything

Newer, more elaborate techniques pulling in key secondary structure features for better searching (uses automata theory, dynamic programming, Dijkstra, more optimization stuff,...)

Details  
**CENSORED**  
(but stay tuned...)  
Plenty of CS here

# Results

Typically 200-fold speedup or more

Finding dozens to hundreds of new ncRNA genes in many families

The *computational* advance has enabled new *biological* discoveries

Newer, more elaborate techniques pulling in key secondary structure features for better searching (uses automata theory, dynamic programming, Dijkstra, more optimization stuff,...)

**More Admin**

# Course Focus & Goals

Mainly sequence analysis

Algorithms for alignment, search, & discovery

Specific sequences, general types (“genes”, etc.)

Single sequence and comparative analysis

Techniques: HMMs, EM, MLE, Gibbs, Viterbi...

Enough bio to motivate these problems

including very light intro to modern biotech supporting them

Math/stats/cs underpinnings thereof

Applied to real data

A *VERY* Quick Intro To  
Molecular Biology



# The Genome

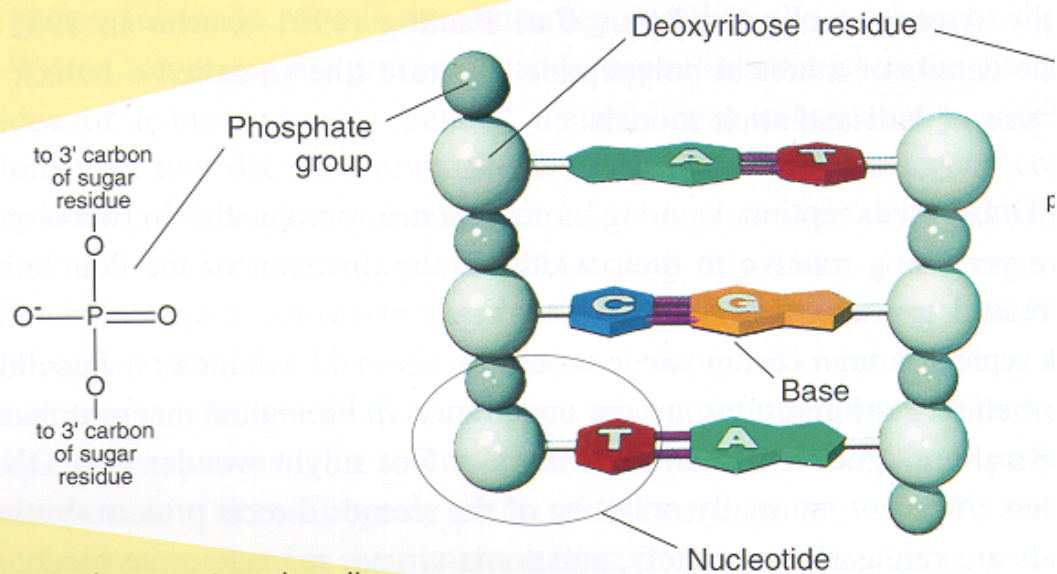
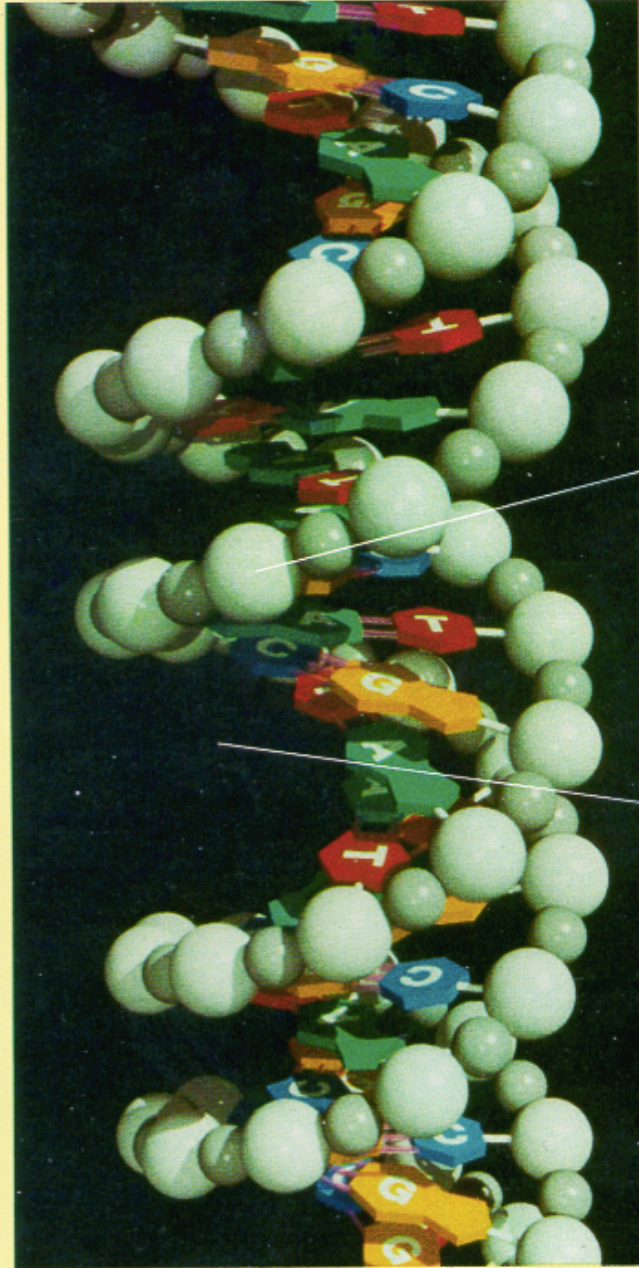
The hereditary info present in every cell

DNA molecule -- a long sequence of  
*nucleotides* (A, C, T, G)

Human genome -- about  $3 \times 10^9$  nucleotides

The genome project -- extract & interpret  
genomic information, apply to genetics of  
disease, better understand evolution, ...

# The Double Helix



As shown, the two strands coil about each other in a fashion such that all the bases project inward toward the helix axis. The two strands are held together by hydrogen bonds (pink rods) linking each base projecting from one backbone to its so-called complementary base projecting from the other backbone. The base A always bonds to T (A and T are comple-

Shown in (b) is an uncoiled fragment of (a three complementary base pair chemist's viewpoint, each strand a polymer made up of four re-called deoxyribonucleotides

# DNA

Discovered 1869

Role as carrier of genetic information - much later

4 “bases”:

adenine (A), cytosine (C), guanine (G), thymine (T)

The Double Helix - Watson & Crick (& Franklin) 1953

Complementarity

$A \longleftrightarrow T$      $C \longleftrightarrow G$

Visualization:

<http://www.rcsb.org/pdb/explore.do?structureId=123D>

# Genetics - the study of heredity

A *gene* -- classically, an abstract heritable attribute existing in variant forms (*alleles*)

ABO blood type—1 gene, 3 alleles

Mendel

Each individual two copies of each gene

Each parent contributes one (randomly)

Independent assortment (approx, but useful)

*Genotype vs phenotype*

I.e., genes vs their outward manifestation

AA or AO genotype → “type A” phenotype

# Cells

Chemicals inside a sac - a fatty layer called the *plasma membrane*

*Prokaryotes* (bacteria, archaea) - little recognizable substructure

*Eukaryotes* (all multicellular organisms, and many single celled ones, like yeast) - genetic material in nucleus, other organelles for other specialized functions

# Chromosomes

1 pair of (complementary) DNA molecules  
(+ protein wrapper)

Most prokaryotes: just 1 chromosome

Eukaryotes - ~~all~~<sup>most</sup> cells have same number  
of chromosomes, e.g. fruit flies 8, humans  
& bats 46, rhinoceros 84, ...

# Mitosis/Meiosis

Most “higher” eukaryotes are *diploid* - have homologous pairs of chromosomes, one maternal, other paternal (exception: sex chromosomes)

*Mitosis* - cell division, duplicate each chromosome, 1 copy to each daughter cell

*Meiosis* - 2 divisions form 4 *haploid* gametes (egg/sperm)

*Recombination/crossover* -- exchange maternal/paternal segments

# Proteins

Chain of amino acids, of 20 kinds

Proteins: the major functional elements in cells

- Structural/mechanical

- Enzymes (catalyze chemical reactions)

- Receptors (for hormones, other signaling molecules, odorants,...)

- Transcription factors

- ...

3-D Structure is crucial: the protein folding problem



# The “Central Dogma”

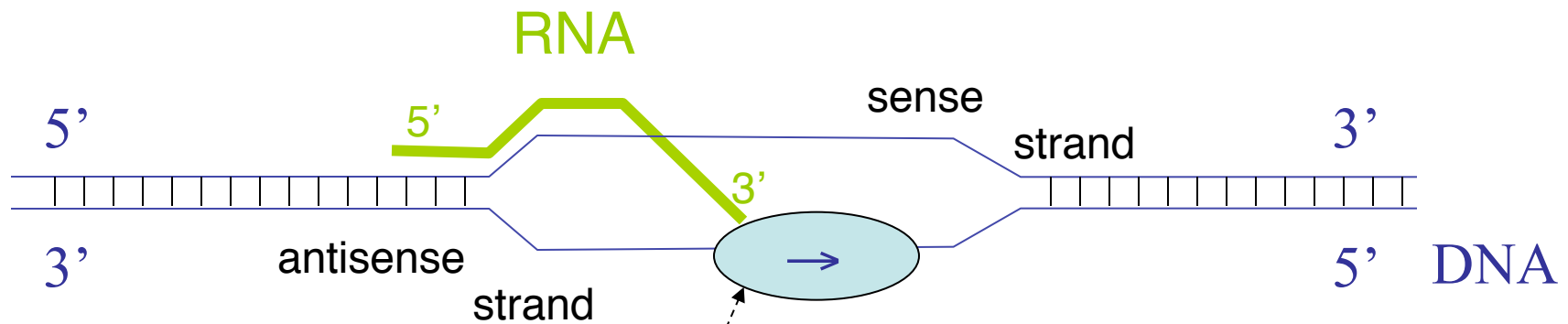
Genes encode proteins

DNA transcribed into messenger RNA

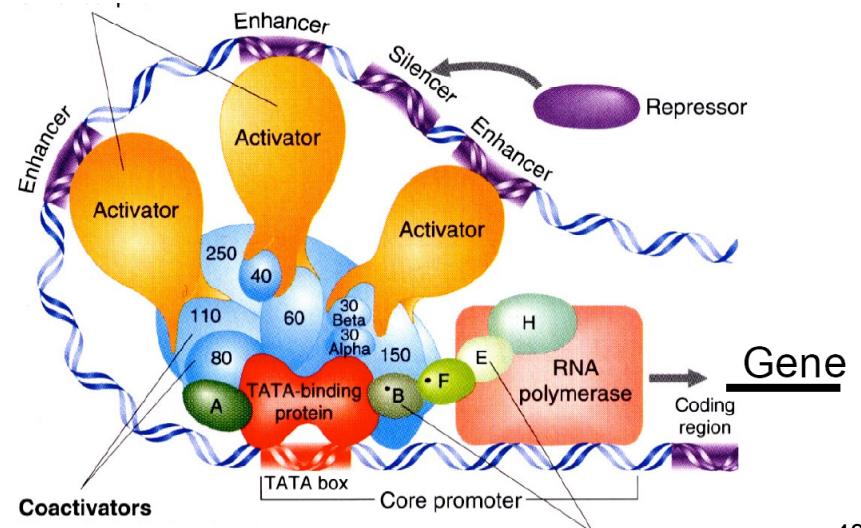
mRNA translated into proteins

Triplet code (codons)

# Transcription: DNA → RNA



RNA polymerase

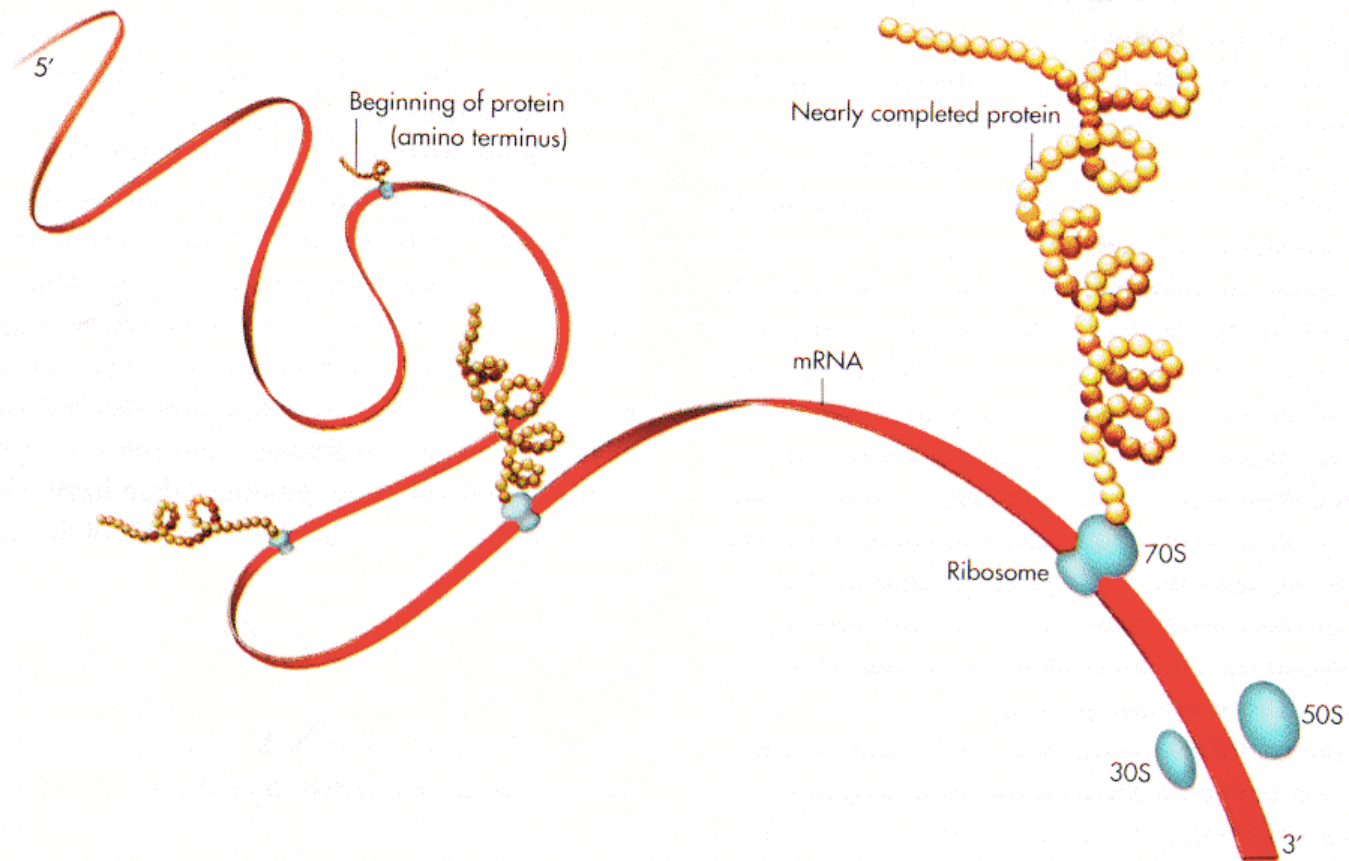


# Codons & The Genetic Code

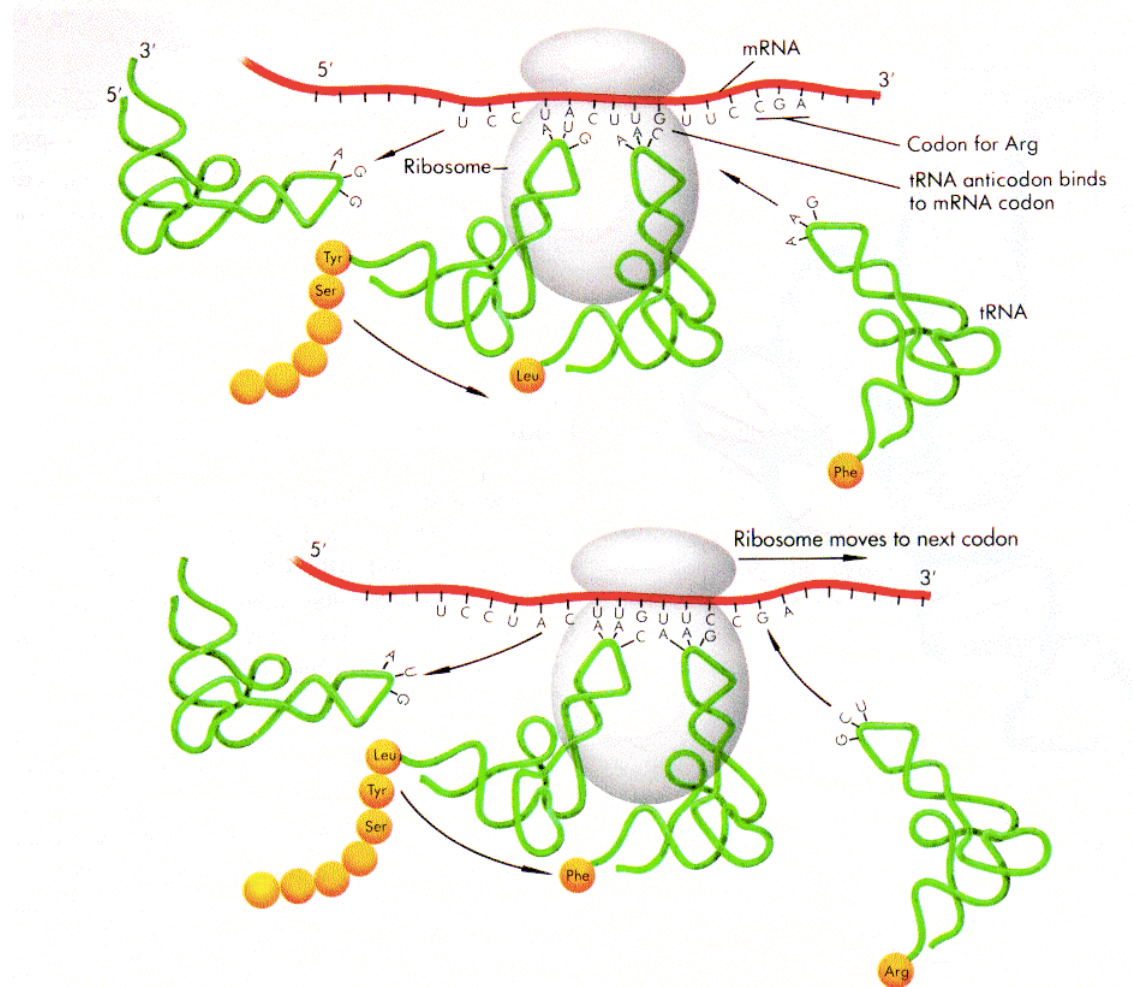
		Second Base					
		U	C	A	G		
First Base	U	Phe	Ser	Tyr	Cys	Third Base	U
		Phe	Ser	Tyr	Cys		C
		Leu	Ser	Stop	Stop		A
		Leu	Ser	Stop	Trp		G
	C	Leu	Pro	His	Arg		U
		Leu	Pro	His	Arg		C
		Leu	Pro	Gln	Arg		A
		Leu	Pro	Gln	Arg		G
	A	Ile	Thr	Asn	Ser		U
		Ile	Thr	Asn	Ser		C
		Ile	Thr	Lys	Arg		A
		Met/Start	Thr	Lys	Arg		G
	G	Val	Ala	Asp	Gly		U
		Val	Ala	Asp	Gly		C
		Val	Ala	Glu	Gly		A
		Val	Ala	Glu	Gly		G

Ala : Alanine  
 Arg : Arginine  
 Asn : Asparagine  
 Asp : Aspartic acid  
 Cys : Cysteine  
 Gln : Glutamine  
 Glu : Glutamic acid  
 Gly : Glycine  
 His : Histidine  
 Ile : Isoleucine  
 Leu : Leucine  
 Lys : Lysine  
 Met : Methionine  
 Phe : Phenylalanine  
 Pro : Proline  
 Ser : Serine  
 Thr : Threonine  
 Trp : Tryptophane  
 Tyr : Tyrosine  
 Val : Valine

# Translation: mRNA → Protein



# Ribosomes



# Gene Structure

mRNA built 5' to 3'

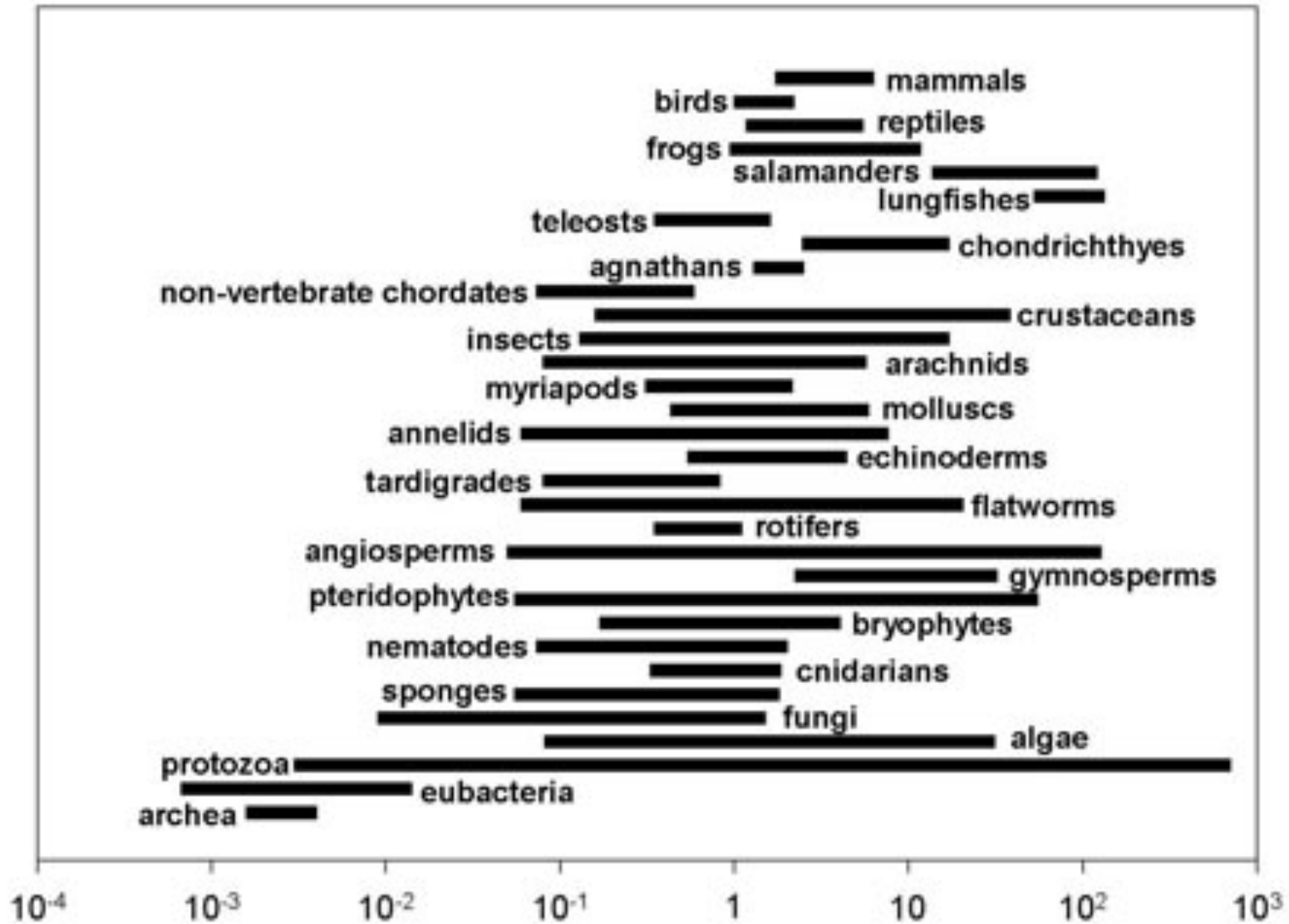
Promoter region and transcription factor binding sites (usually) precede 5' end

Transcribed region includes 5' and 3' untranslated regions

In eukaryotes, most genes also include *introns*, spliced out before export from nucleus, hence before translation

# Genome Sizes

	Base Pairs	Genes
Mycoplasma genitalium	580,073	483
Pandora Virus	2,900,000	2,500
E. coli	4,639,221	4,290
Saccharomyces cerevisiae	12,495,682	5,726
Caenorhabditis elegans	95,500,000	19,820
Arabidopsis thaliana	115,409,949	25,498
Drosophila melanogaster	122,653,977	13,472
Humans	$3.3 \times 10^9$	~25,000
Amoeba dubia	~ 200 x human	



DNA content (picograms)

<http://www.genomesize.com/statistics.php>



# Genome Surprises

Humans have  $< 1/3$  as many genes as expected

But perhaps more proteins than expected, due to *alternative splicing, alt start, alt end*

Protein-wise, all mammals are just about the same

But more individual variation than expected

And many more *non-coding RNAs* -- more than protein-coding genes, by some estimates

Many other non-coding regions are highly conserved, e.g., across all vertebrates

Subset of DNA being transcribed is  $\gg 2\%$  coding

Complex, subtle “epigenetic” information

... and much more ...

Read one of the many intro surveys or books for much more info.

# Bio Concept Summary

cells

DNA

base pairing

genome

replication, transcription, translation