CSEP590A
Computational Biology

http://www.cs.washington.edu/csep590a

Larry Ruzzo

Spring 2013

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

-- Chinese Proverb

University of Washington

Computer Science & Engineering

CSE P590A, Sp '13: Computational Biology (Professional Masters Program)

CSE Home | About Us | Search | Contact Info

Administrative
- Schedule & Reading
- HW0: Background Survey

Course Email/BBoard
- Subscription Options
- Class List Archive

Lecture Notes

Previous Versions
- CSEP 590B, 2011
- CSEP 590A, 2008
- CSEP 590A, 2006
- CSE 590TV, 2003

Resources
- Pubmed
- NCBI Science Primer
- NHGRI Talking Glossary
- ORNL Genome Glossary
- A Molecular Biology Glossary
- BLAST
- Swiss-Prot
- PDB

Lecture: MGH 231 (schematic) Th 6:30-9:20

		Office Hours	Location	Phone
Instructor:	Larry Ruzzo, ruzzo@cs	By appt.	CSE 554	(206) 543-6298
TA:	Daniel Jones, dcjones@cs	By appt.		

Course Email: csep590a_sp13@uw.edu. Staff announcements and general interest student/staff [illegible] etc. The instructor and TA are subscribed to this list. Enrolled students are as well, but probably should ch[illegible]ions. Messages are automatically archived.

Discussion Board: Also feel free to use Catalyst GoPost to discuss h[illegible]

Catalog Description: An introduction to the use of comput[illegible]ing of biological systems at the molecular level.

Prerequisite: None

Credits: 4

Learning Objectives: The a[illegible] sequences of humans and other organisms is one of the landmark achievements of science. Understandi[illegible]s a problem that will challenge scientists for decades to come, and the nature and scope of the problem mea[illegible]y a vital role. The primary objective of the course is for students to understand the variety of com[illegible]at arise in this interdisciplinary field. Students will learn enough of the basic concepts of molecular biology to [illegible]omputational problems presented in the rest of the course. They will learn how some of the computational methods they [illegible]er courses can be applied to solve problems in modern molecular biology. An important component is to learn the nature and [illegible]me of the key public databases available for the solution of these problems, as well as publicly available computational analysis tools [illegible]algorithmic principles underlying them.

Textbook: Richard Durbin, Sean R. Eddy, Anders Krogh and Graeme Mitchison, *Biological Sequence Analysis: Probabilistic models of proteins and nucleic acids*, Cambridge, 1998. (Available from U Book Store, Amazon, etc.) Errata.

References: See Schedule & Reading

Computer Science & Engineering
University of Washington
Box 352350
Seattle, WA 98195-2350

Please do this now

http://www.cs.washington.edu/csep590a

Tonight

Admin

Why Comp Bio?

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

Admin Stuff

Course Mechanics & Grading

Web http://www.cs.washington.edu/csep590a

Reading

In class discussion

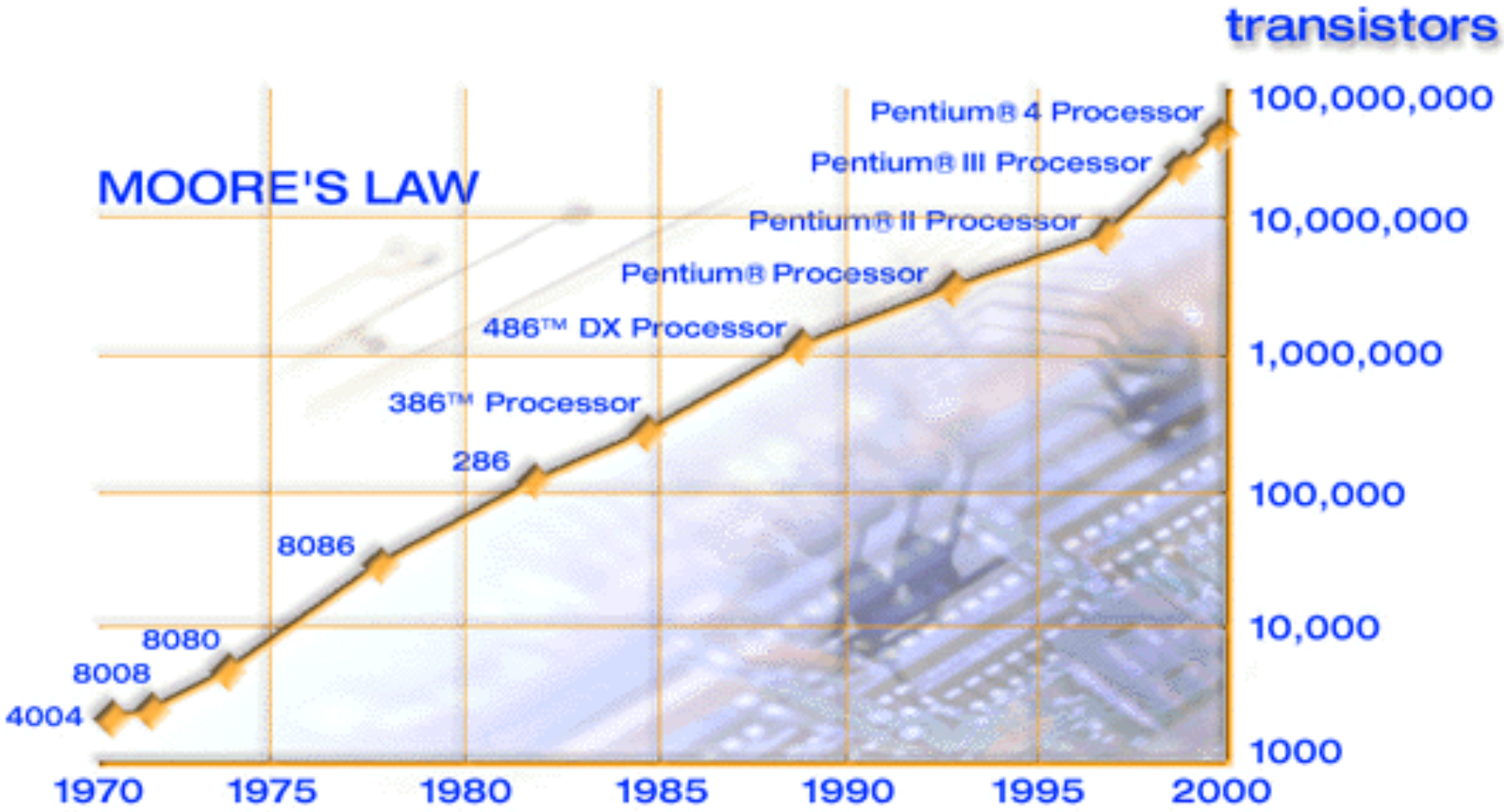
Homeworks

- reading blogs
- paper exercises
- programming

Check web for 1st, ~~soon~~ now

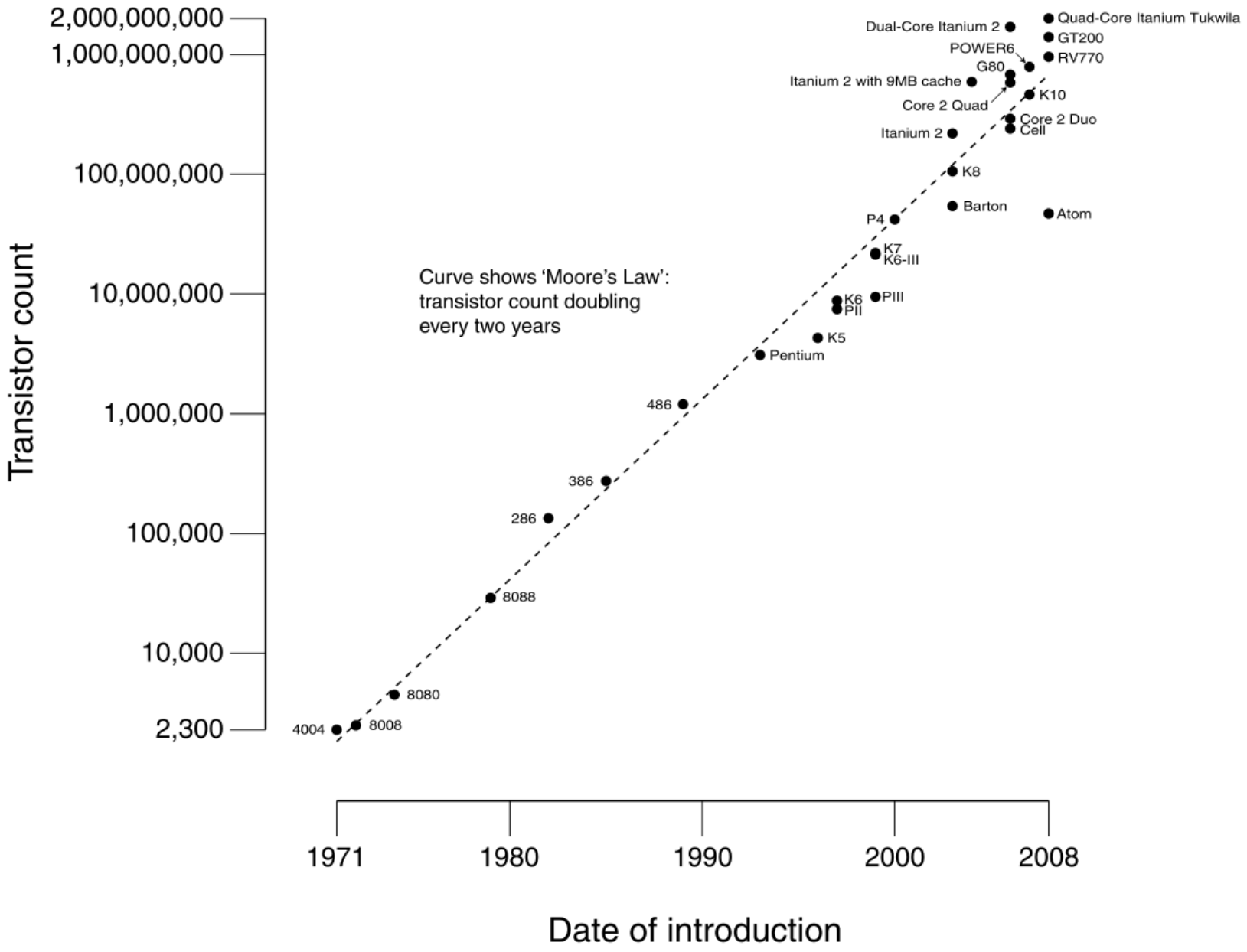
No exams, but possible oversized last homework in lieu of final

Background & Motivation

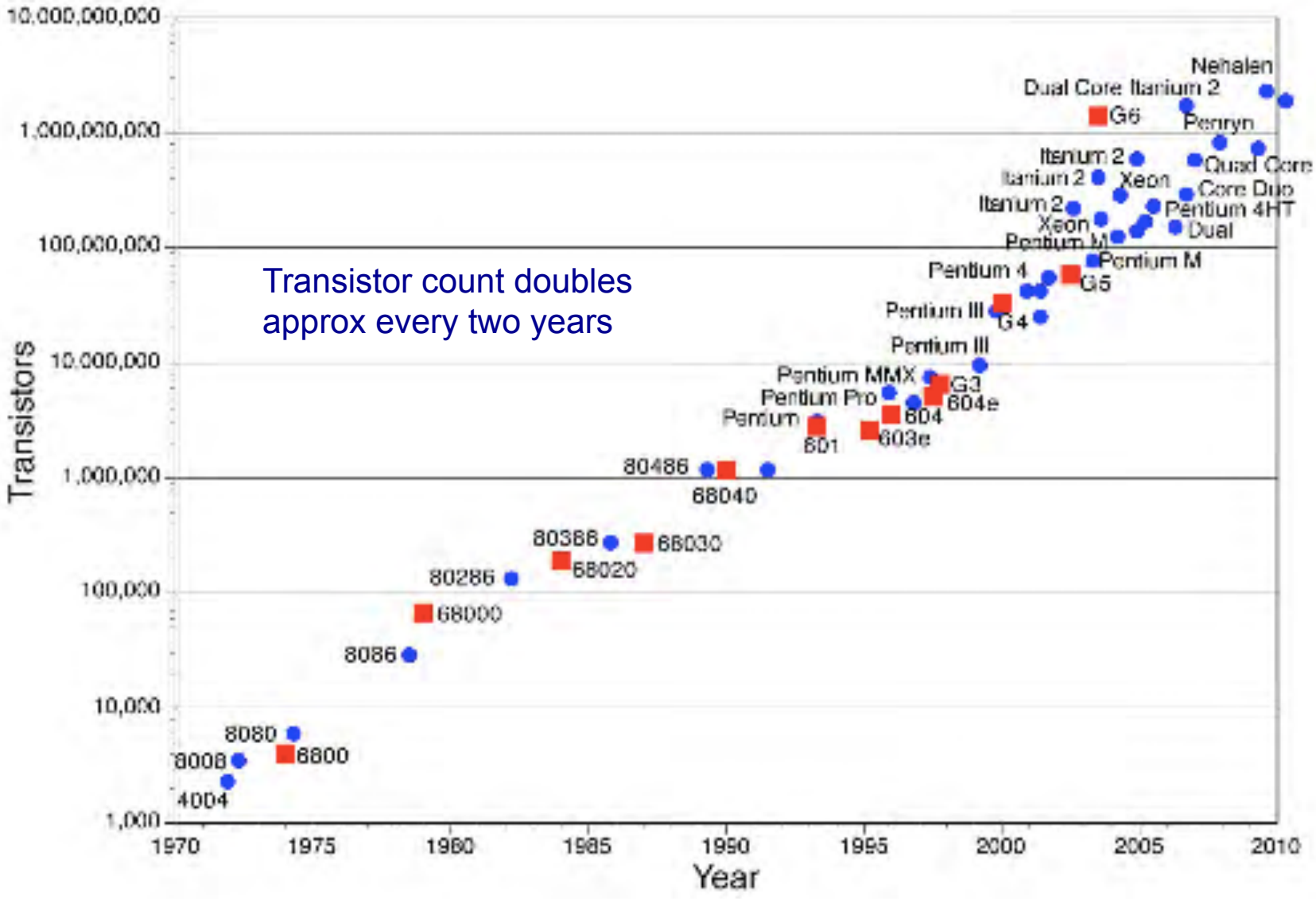


Source: http://www.intel.com/research/silicon/mooreslaw.htm

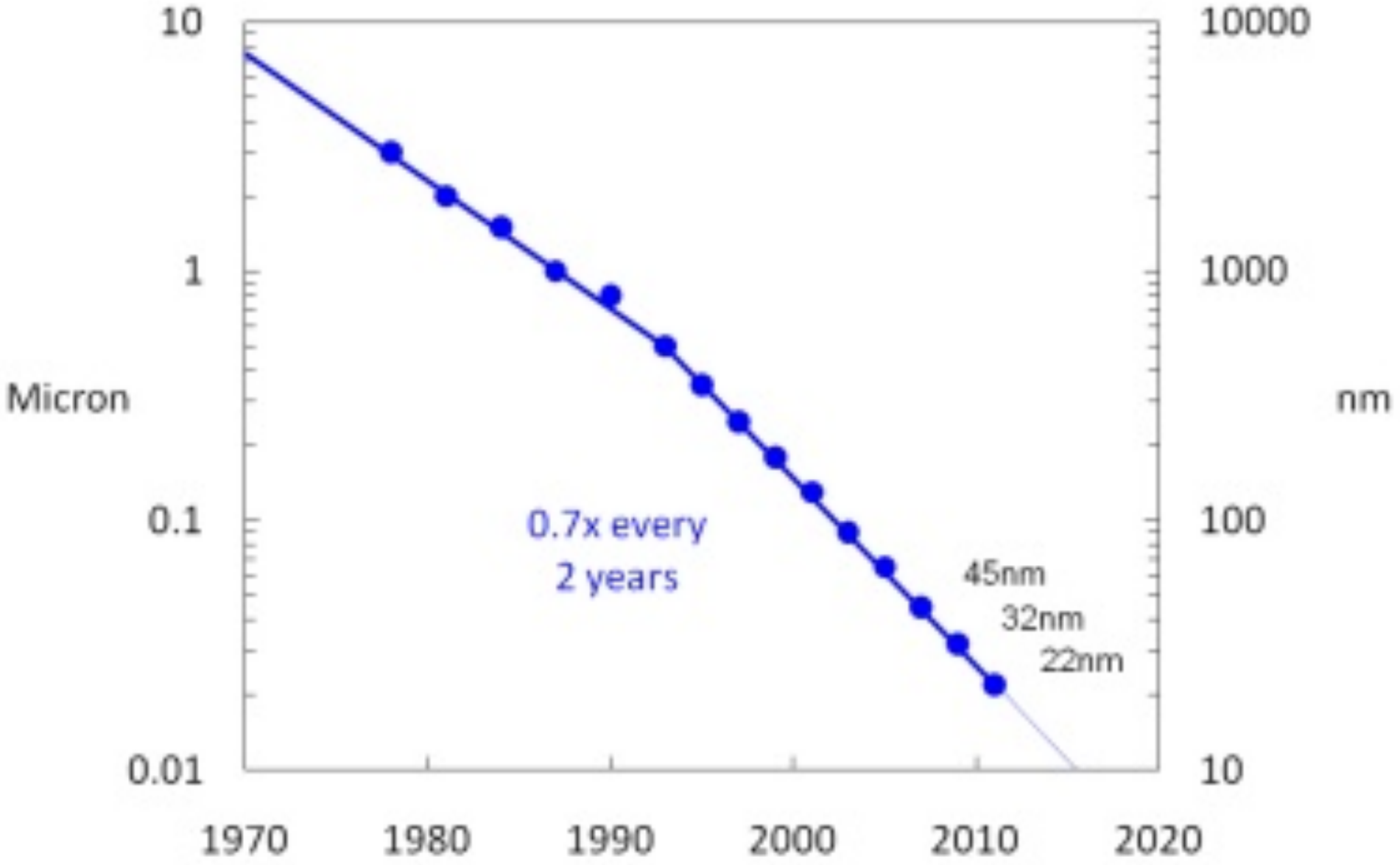
CPU Transistor Counts 1971-2008 & Moore's Law



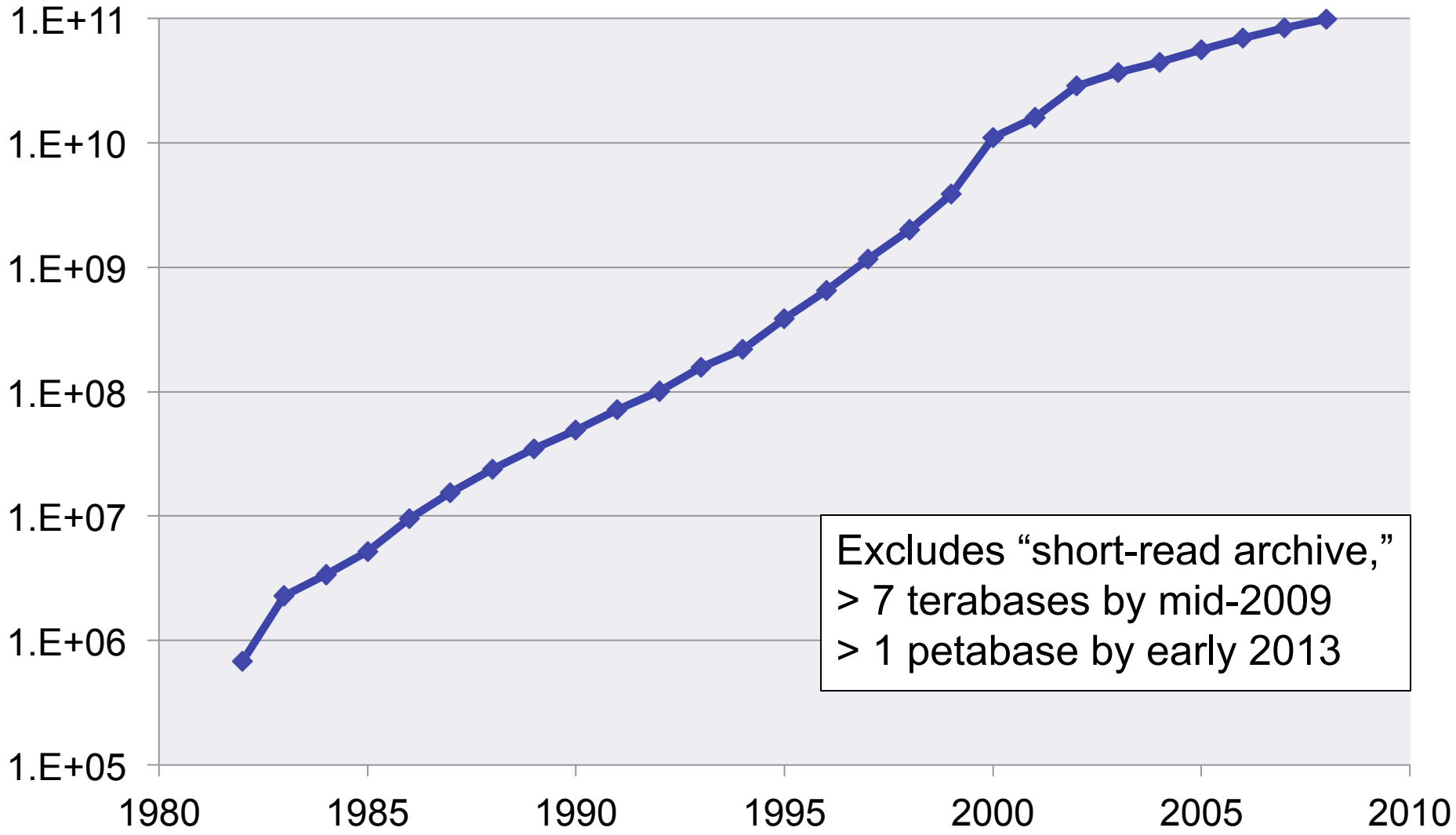
Moore's Law


Transistor count doubles
approx every two years
Transistors
Year
10,000,000,000
1,000,000,000
100,000,000
10,000,000
1,000,000
100,000
10,000
1,000
1970
1975
1980
1985
1990
1995
2000
2005
2010
Nehalem
Dual Core Itanium 2
G6
Penryn
Itanium 2
Quad Core
Itanium 2
Xeon
Core Duo
Itanium 2
Pentium 4HT
Xeon
Dual
Pentium M
Pentium M
Pentium 4
G5
Pentium III
G4
Pentium III
Pentium MMX
G3
Pentium Pro
604e
Pentium
604
601
603e
80486
68040
80386
68030
68020
80286
68000
8086
8080
8008
6800
4004

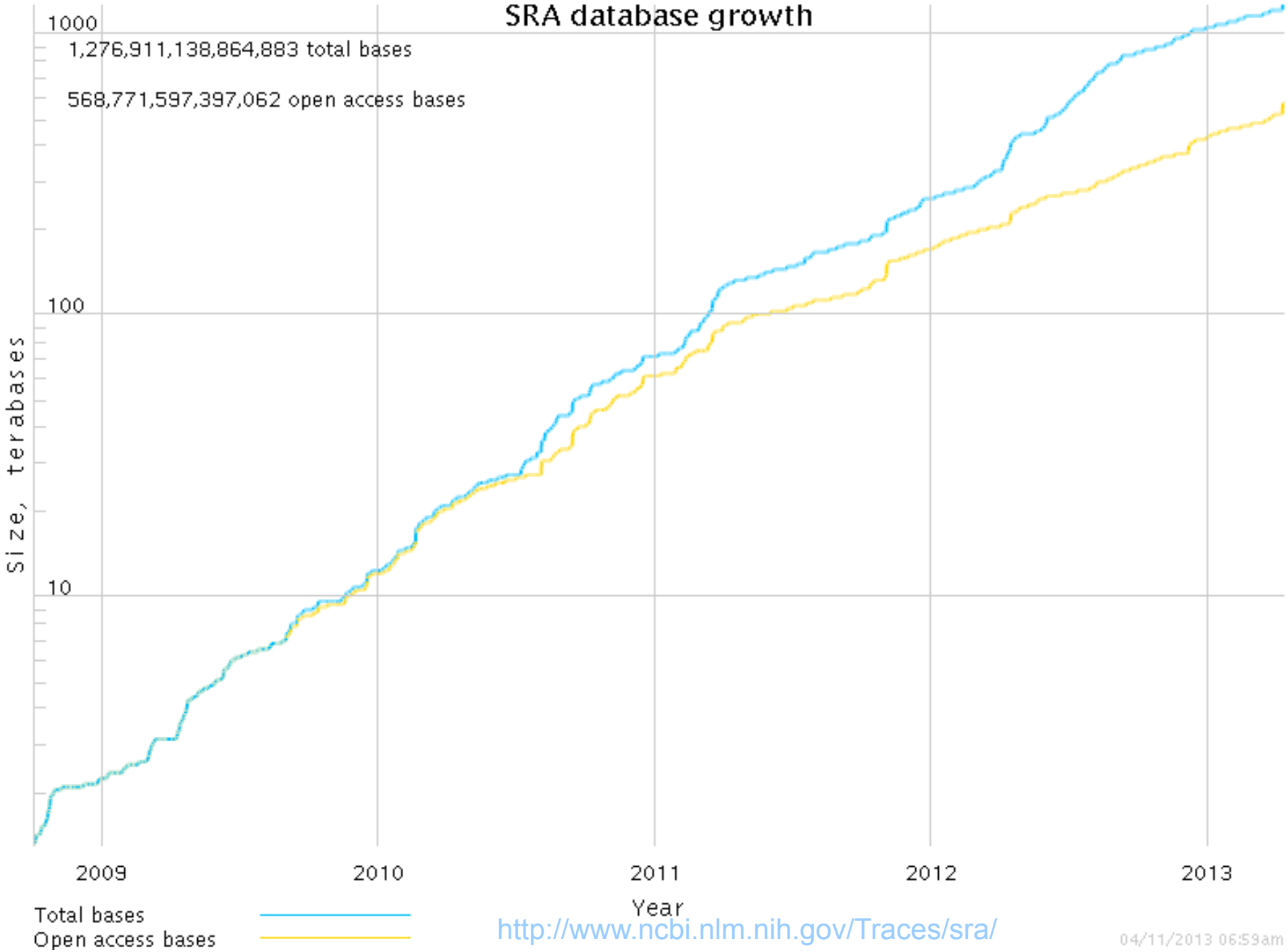
Feature Size



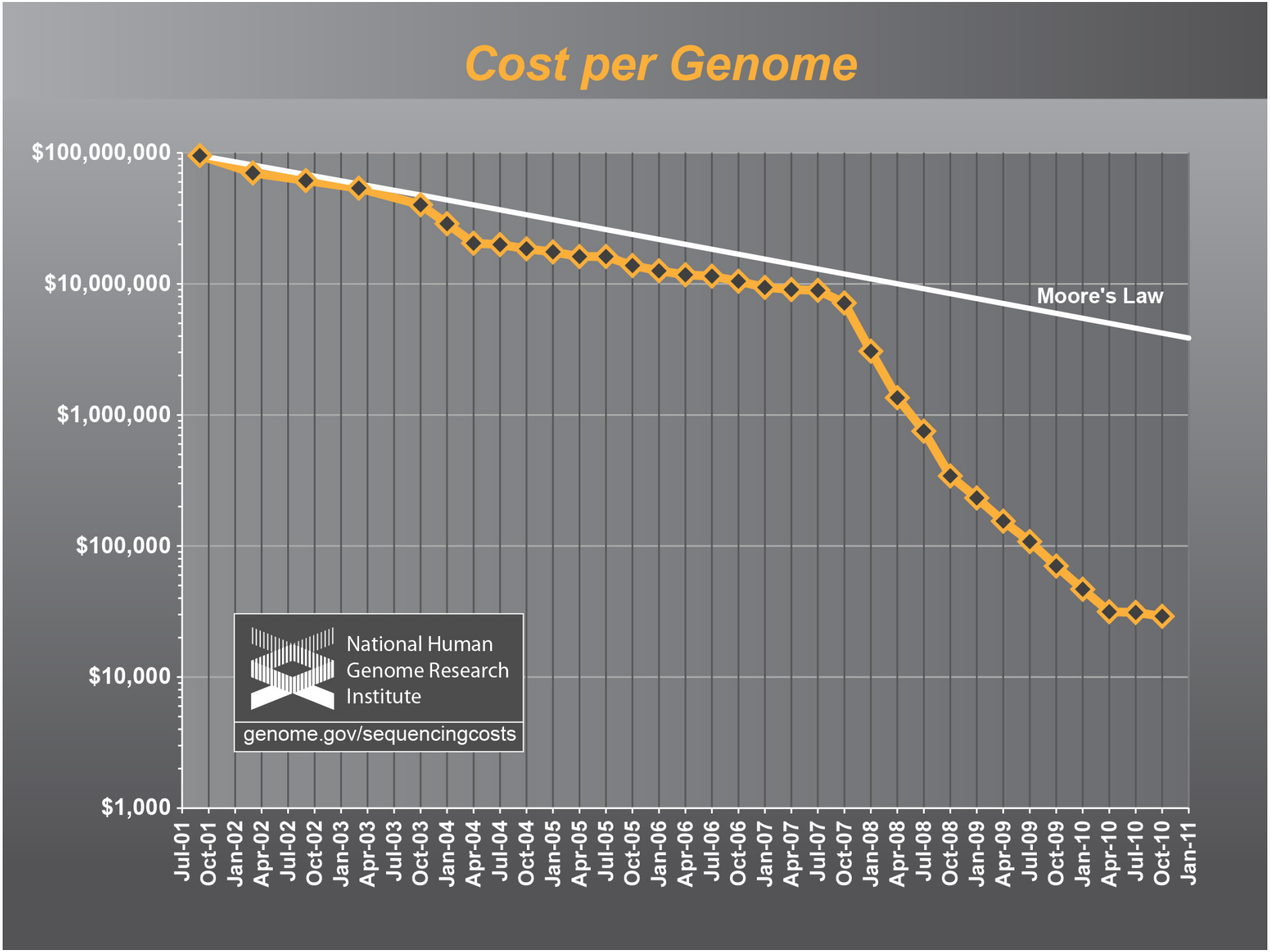
Growth of GenBank (Base Pairs)



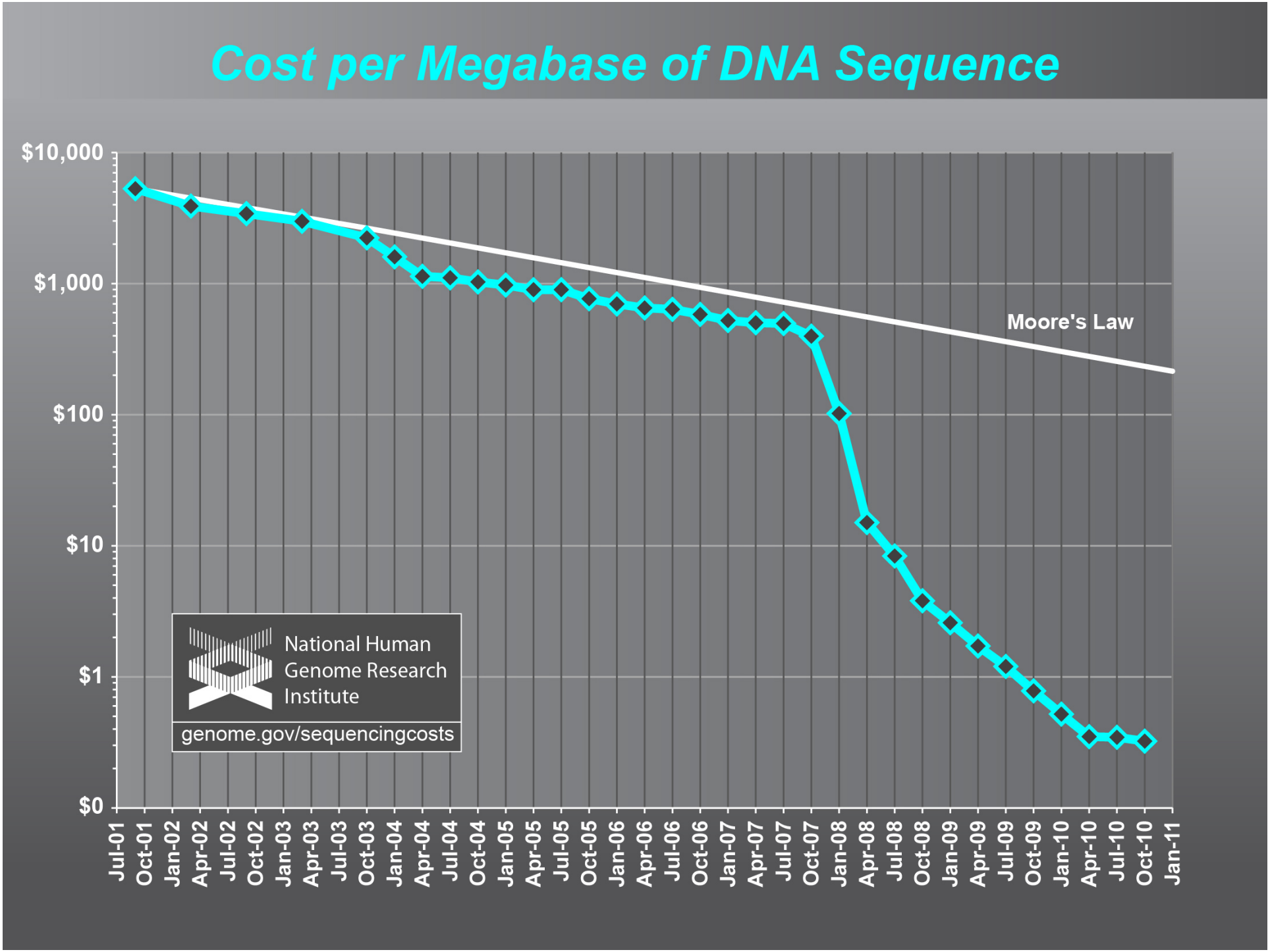
Source: http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html

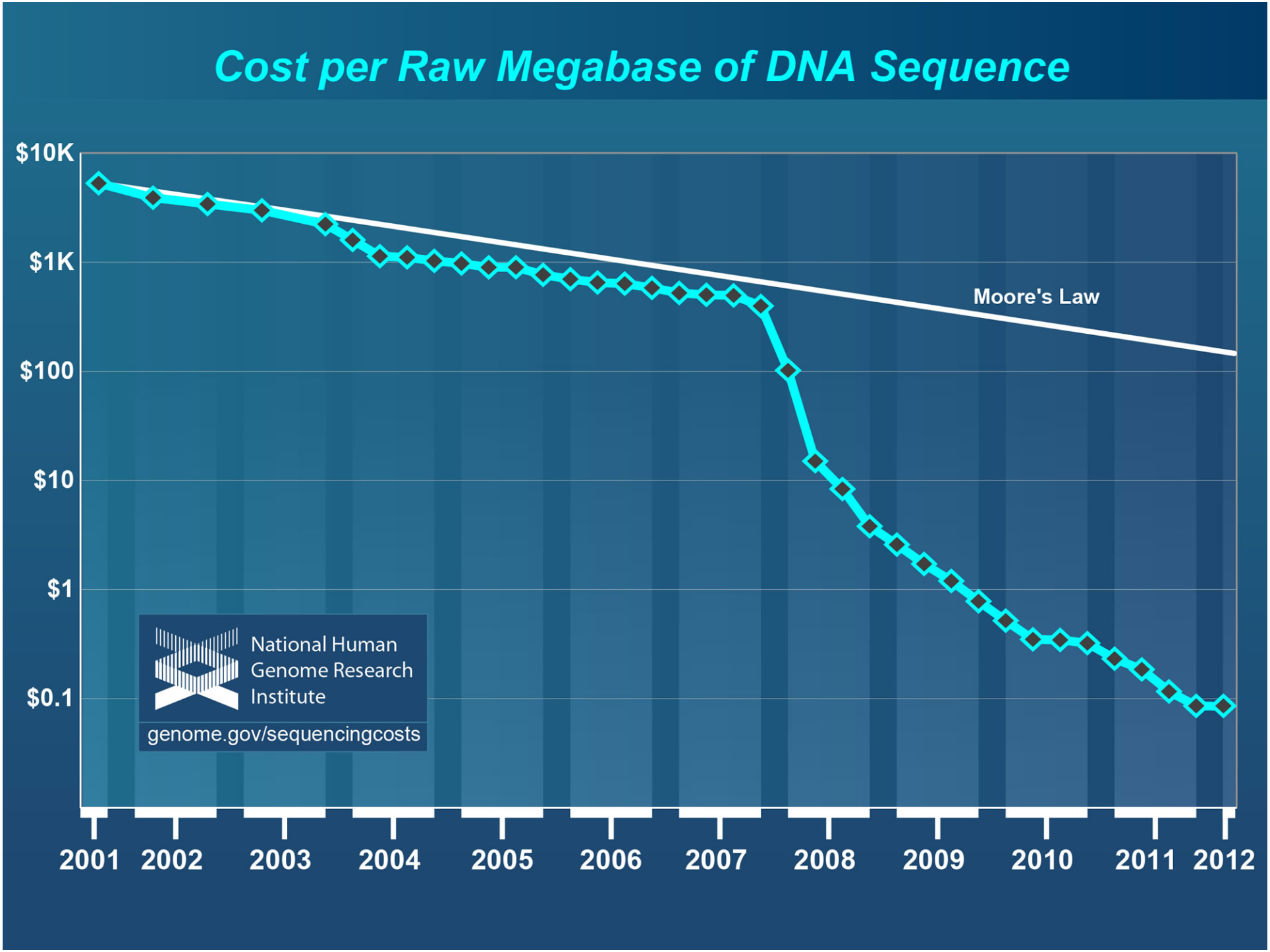

SRA database growth
1000
1,276,911,138,864,883 total bases
568,771,597,397,062 open access bases
100
10
Size, terabases
2009
2010
2011
2012
2013
Year
Total bases
Open access bases
http://www.ncbi.nlm.nih.gov/Traces/sra/
04/11/2013 06:59am

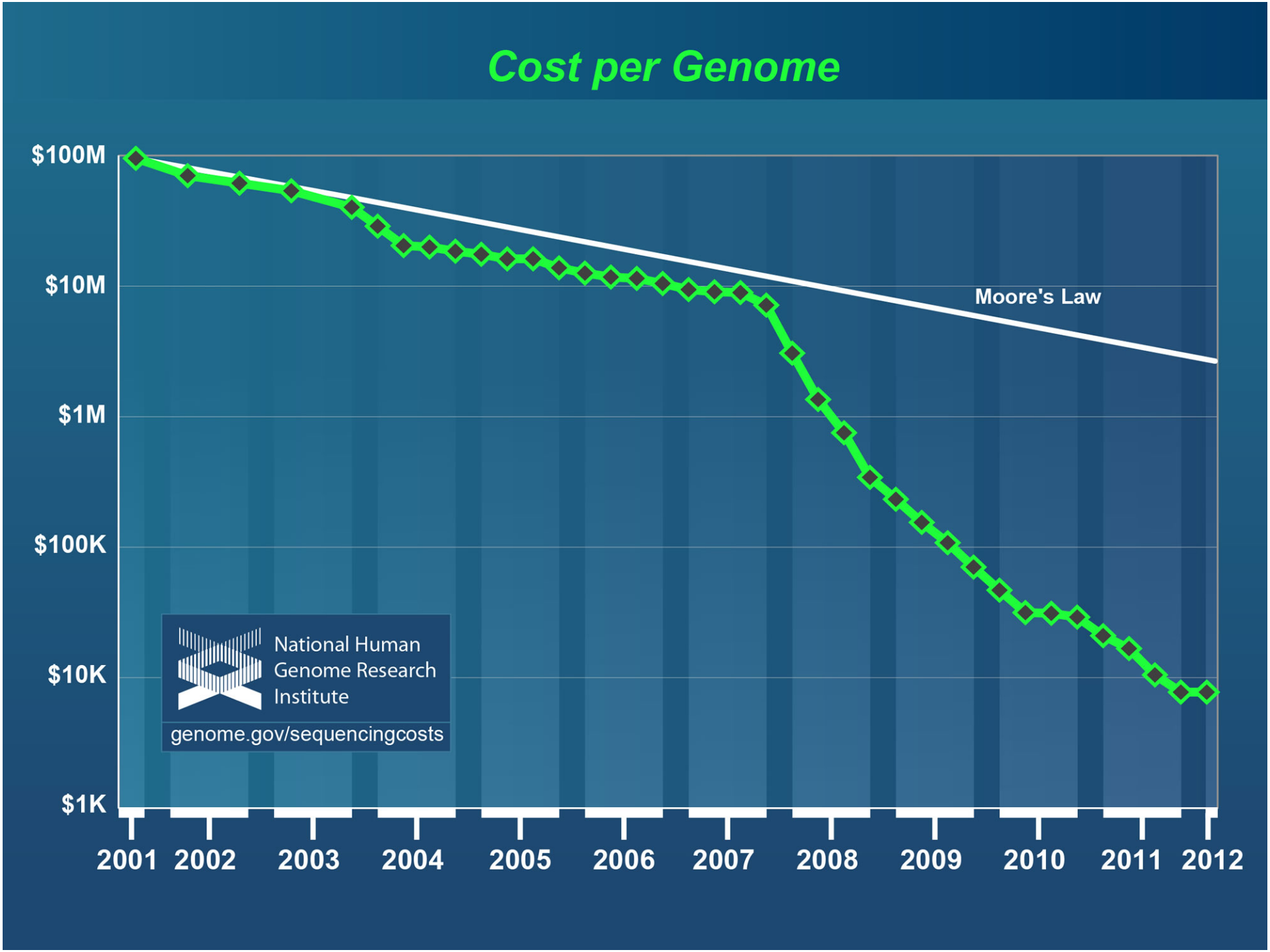
Cost per Genome



Cost per Megabase of DNA Sequence

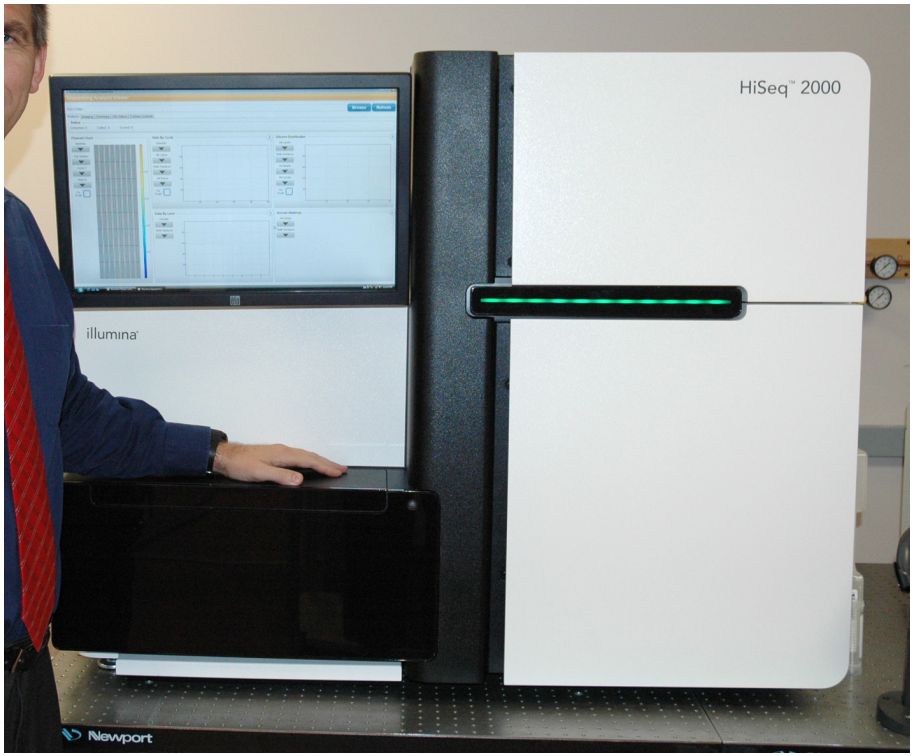



Cost per Raw Megabase of DNA Sequence
$10K
$1K
$100
$10
$1
$0.1
Moore's Law
National Human Genome Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
National Human Genome Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012

Modern DNA Sequencing

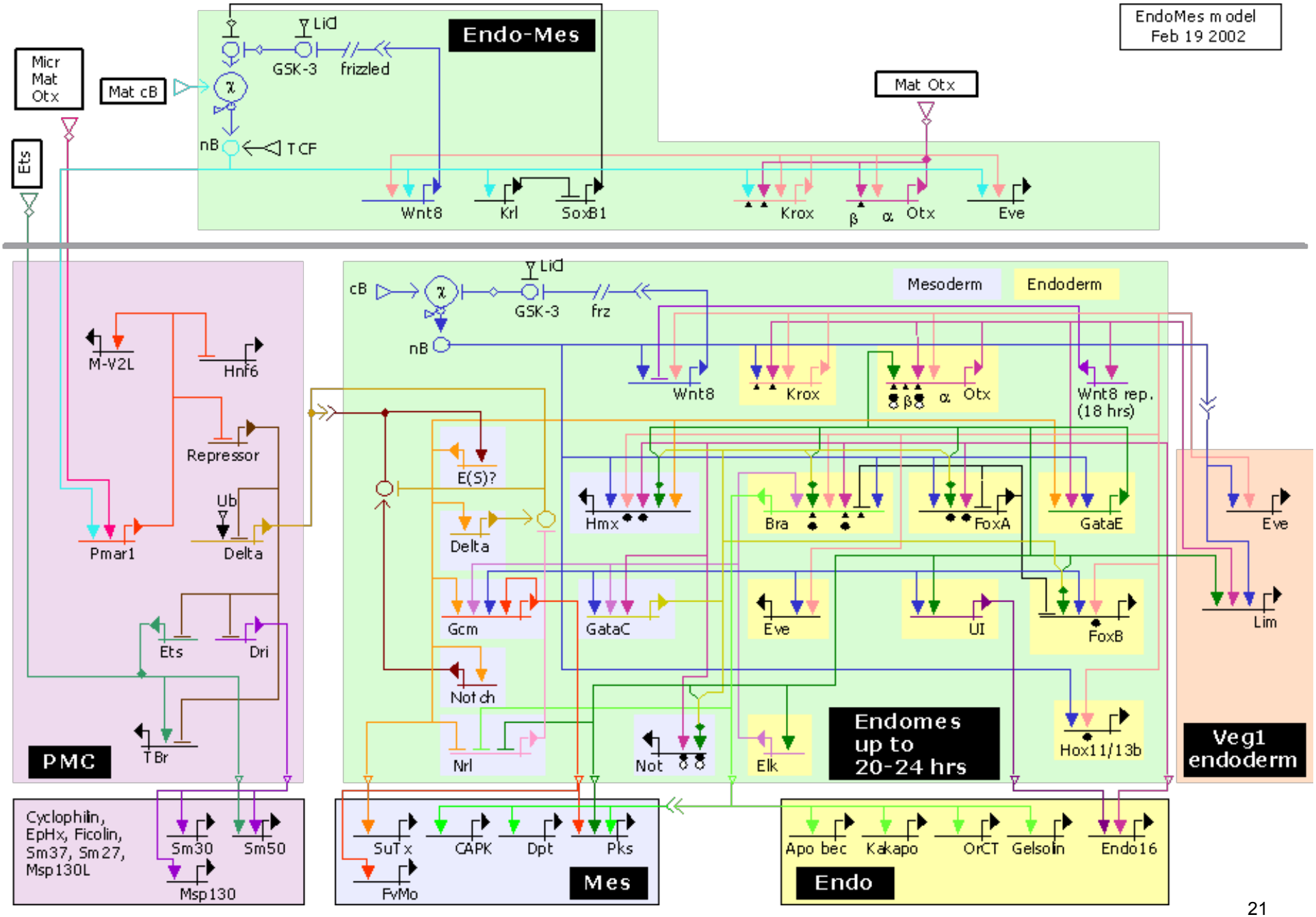
A table-top box the size of your oven
(but costs a bit more … ;-)
can generate
~100 billion BP of DNA seq/day; i.e.
= 2008 genbank,
= 30x your genome



The Human Genome Project

```
   1 gagcccggcc cgggggacgg gcggcgggat agcgggaccc cggcgcggcg gtgcgcttca
  61 gggcgcagcg gcggccgcag accgagcccc gggcgcggca agaggcggcg ggagccggtg
 121 gcggctcggc atcatgcgtc gagggcgtct gctggagatc gccctgggat ttaccgtgct
 181 tttagcgtcc tacacgagcc atggggcgga cgccaatttg gaggctggga acgtgaagga
 241 aaccagagcc agtcgggcca agagaagagg cggtggagga 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 ctcactgtgg 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 ctgtcaggga ggaaattgca ttaatactgt tgggtctttt gagtgcaaat gccctgctgg
 961 acacaaactt aatgaagtgt cacaaaaatg tgaagatatt gatgaatgca gcaccattcc
1021 ...
```

The sea urchin *Strongylocentrotus purpuratus*


EndoMes model
Feb 19 2002
Endo-Mes
Micr
Mat
Otx
Mat cB
Ets
LiCl
GSK-3
frizzled
nB
TCF
Wnt8
Krl
SoxB1
Mat Otx
Krox
β
α
Otx
Eve
PMC
M-V2L
Hnf6
Repressor
Ub
Pmar1
Delta
Ets
Dri
TBr
Cyclophilin,
EpHx, Ficolin,
Sm37, Sm27,
Msp130L
Sm30
Sm50
Msp130
cB
LiCl
GSK-3
frz
nB
Mesoderm
Endoderm
Wnt8
Krox
α
Otx
Wnt8 rep.
(18 hrs)
E(S)?
Delta
Notch
Nrl
Hmx
Bra
FoxA
GataE
Gcm
GataC
Eve
UI
FoxB
Not
Elk
Hox11/13b
Endomes
up to
20-24 hrs
Eve
Lim
Veg1
endoderm
SuTx
CAPK
Dpt
Pks
FvMo
Mes
Apo bec
Kakapo
OrCT
Gelsolin
Endo16
Endo

Goals

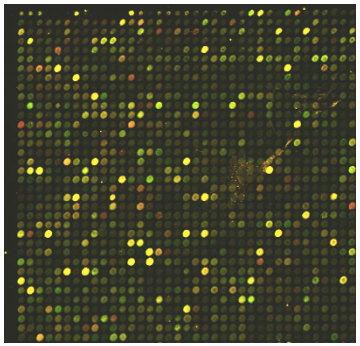
Basic biology

Disease diagnosis/prognosis/treatment

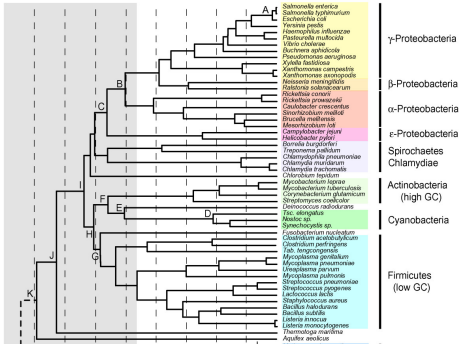
Drug discovery, validation & development

Individualized medicine

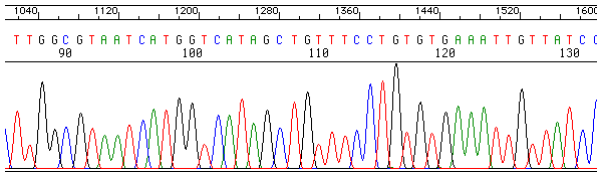
…

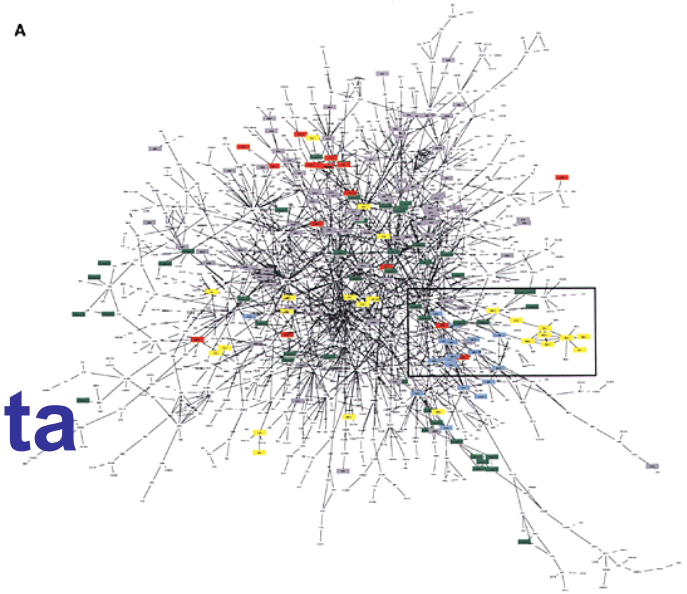


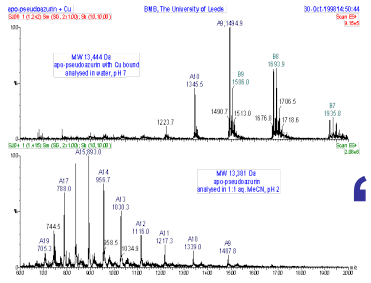
“High-Throughput BioTech”



- Sensors
  - DNA sequencing
  - Microarrays/Gene expression
  - Mass Spectrometry/Proteomics
  - Protein/protein & DNA/protein interaction
- Controls
  - Cloning
  - Gene knock out/knock in
  - RNAi







***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

ACGGGTAA

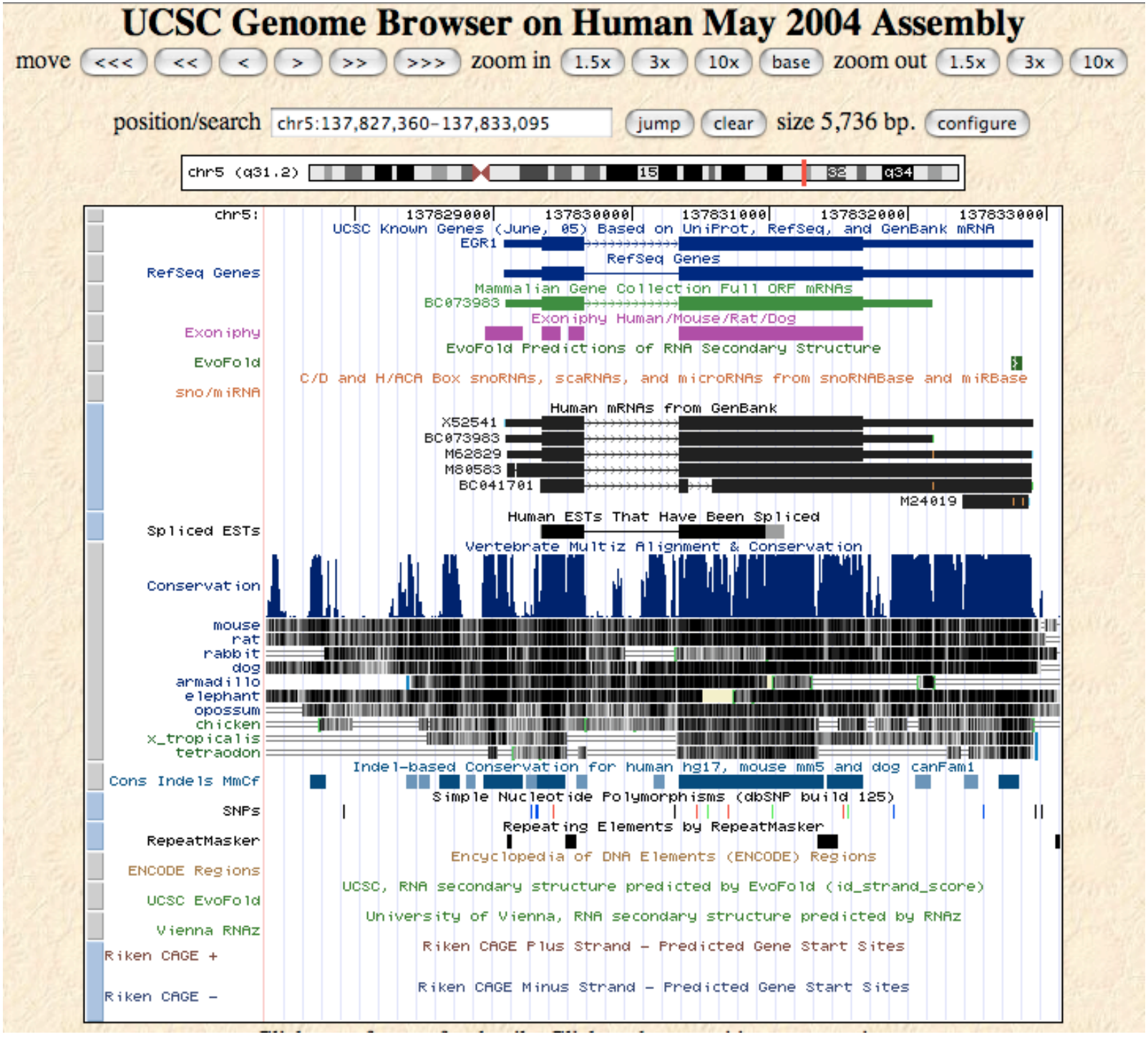
AC GGTAA

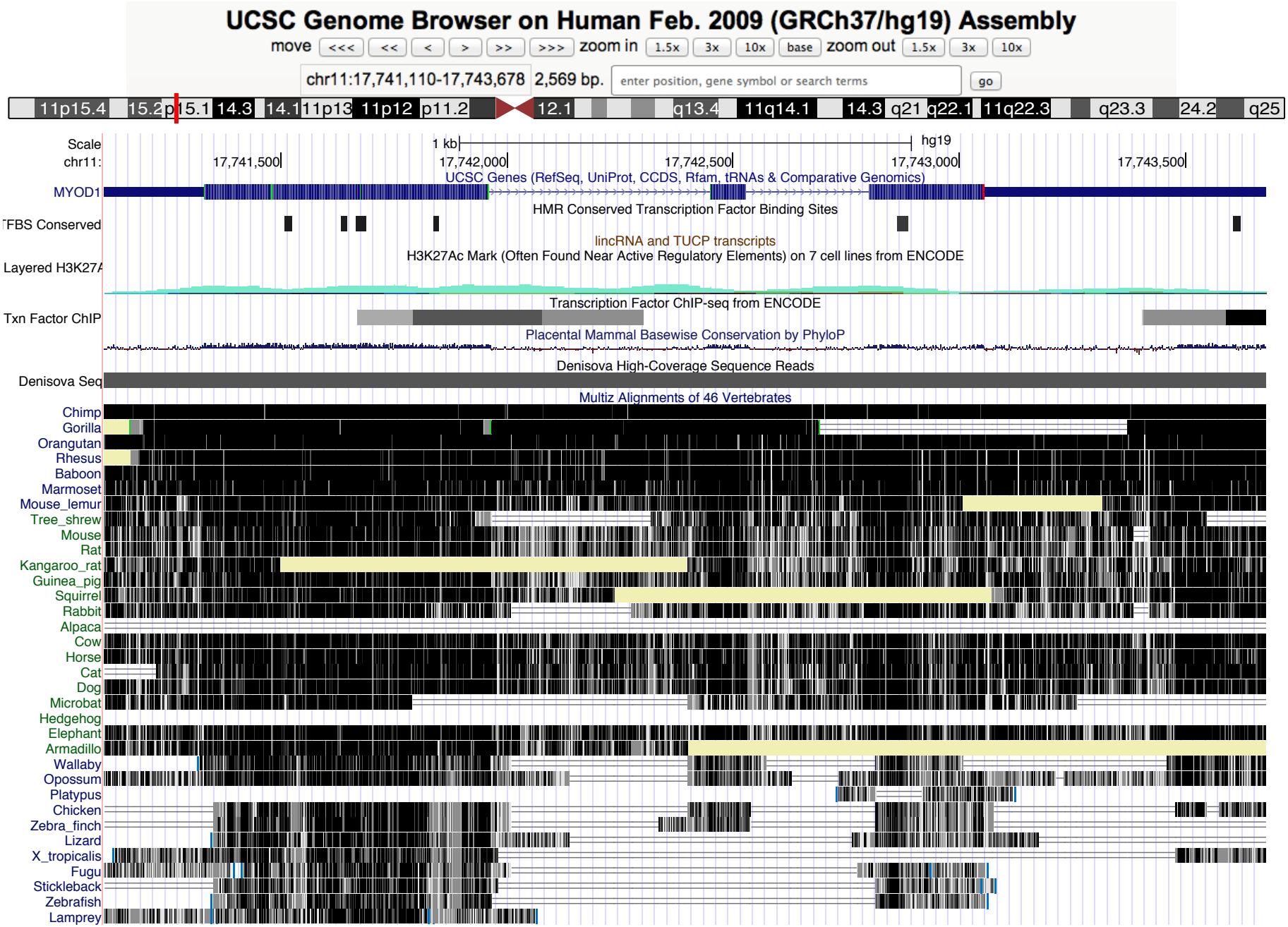
Microfile

Sequence alignment by word processor

D. Ross Boswell

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


UCSC Genome Browser on Human May 2004 Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
position/search chr5:137,827,360-137,833,095 jump clear size 5,736 bp. configure
chr5 (q31.2)
chr5:
137829000 137830000 137831000 137832000 137833000
UCSC Known Genes (June, 05) Based on UniProt, RefSeq, and GenBank mRNA
EGR1
RefSeq Genes
RefSeq Genes
Mammalian Gene Collection Full ORF mRNAs
BC073983
Exoniphy
Exoniphy Human/Mouse/Rat/Dog
EvoFold
EvoFold Predictions of RNA Secondary Structure
sno/miRNA
C/D and H/ACA Box snoRNAs, scaRNAs, and microRNAs from snoRNABase and miRBase
Human mRNAs from GenBank
X52541
BC073983
M62829
M80583
BC041701
M24019
Spliced ESTs
Human ESTs That Have Been Spliced
Conservation
Vertebrate Multiz Alignment & Conservation
mouse
rat
rabbit
dog
armadillo
elephant
opossum
chicken
x_tropicalis
tetraodon
Cons Indels MmCf
Indel-based Conservation for human hg17, mouse mm5 and dog canFam1
SNPs
Simple Nucleotide Polymorphisms (dbSNP build 125)
RepeatMasker
Repeating Elements by RepeatMasker
ENCODE Regions
Encyclopedia of DNA Elements (ENCODE) Regions
UCSC EvoFold
UCSC, RNA secondary structure predicted by EvoFold (id_strand_score)
Vienna RNAz
University of Vienna, RNA secondary structure predicted by RNAz
Riken CAGE +
Riken CAGE Plus Strand - Predicted Gene Start Sites
Riken CAGE -
Riken CAGE Minus Strand - Predicted Gene Start Sites


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.
enter position, gene symbol or search terms
go
11p15.4 15.2 p15.1 14.3 14.1 11p13 11p12 p11.2 12.1 q13.4 11q14.1 14.3 q21 q22.1 11q22.3 q23.3 24.2 q25
Scale
chr11:
1 kb
hg19
17,741,500
17,742,000
17,742,500
17,743,000
17,743,500
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
MYOD1
TFBS Conserved
HMR Conserved Transcription Factor Binding Sites
lincRNA and TUCP transcripts
Layered H3K27A
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Txn Factor ChIP
Transcription Factor ChIP-seq from ENCODE
Placental Mammal Basewise Conservation by PhyloP
Denisova Seq
Denisova High-Coverage Sequence Reads
Multiz Alignments of 46 Vertebrates
Chimp
Gorilla
Orangutan
Rhesus
Baboon
Marmoset
Mouse_lemur
Tree_shrew
Mouse
Rat
Kangaroo_rat
Guinea_pig
Squirrel
Rabbit
Alpaca
Cow
Horse
Cat
Dog
Microbat
Hedgehog
Elephant
Armadillo
Wallaby
Opossum
Platypus
Chicken
Zebra_finch
Lizard
X_tropicalis
Fugu
Stickleback
Zebrafish
Lamprey

An Algorithm Example: ncRNAs

The “Central Dogma”:

- DNA -> messenger RNA -> Protein

Last ~5 years:

- 100s – 1000s of examples of functionally important ncRNAs

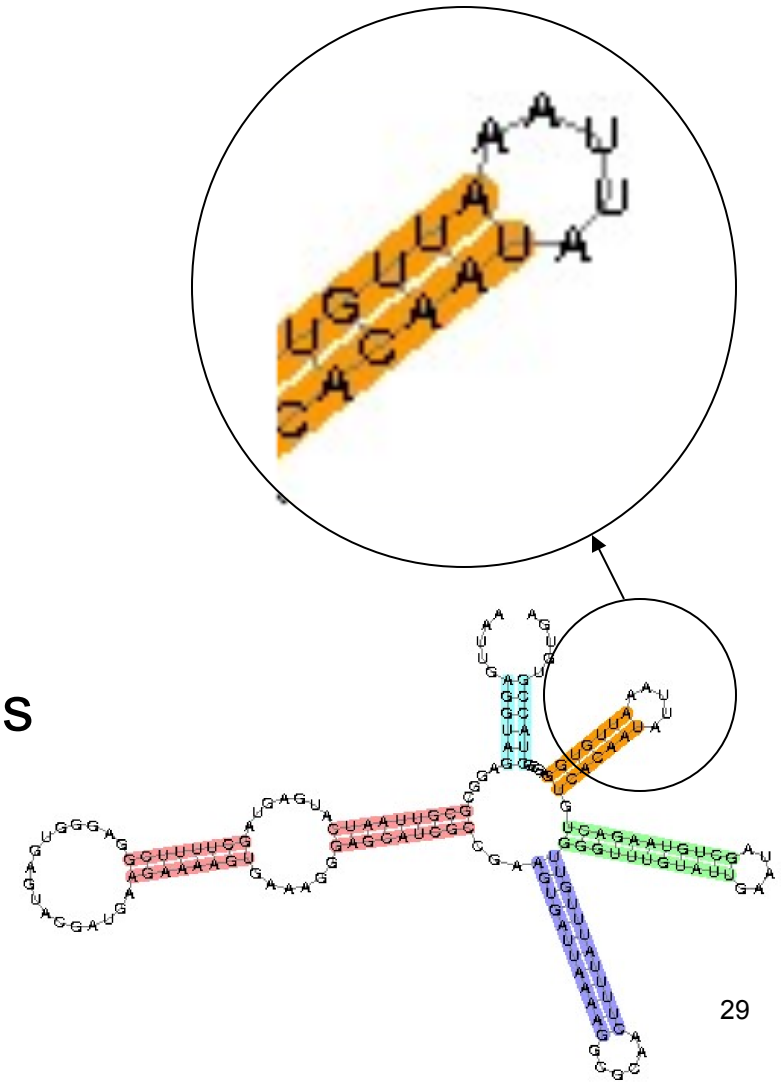
Much harder to find than protein-coding genes

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* ≥ 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,…)

CENSORED
Details
(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

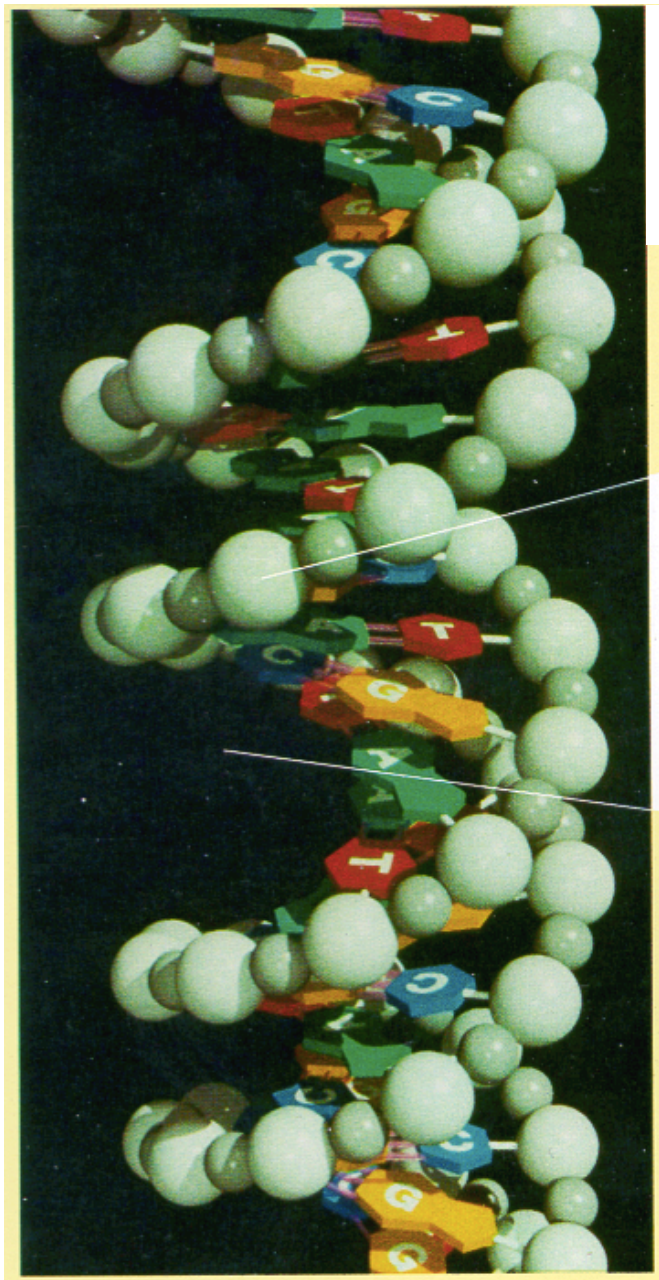
The hereditary info present in every cell

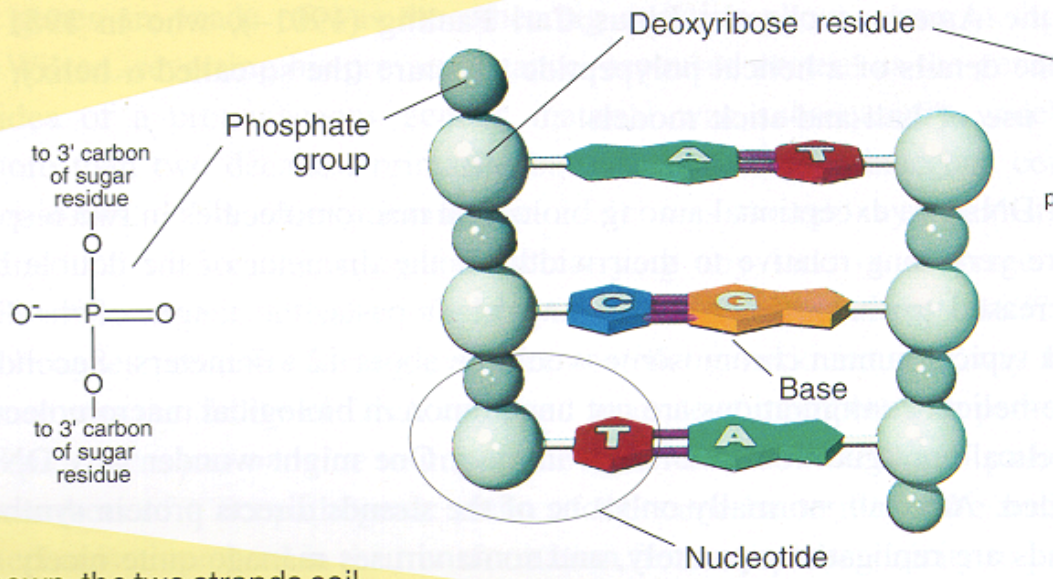
DNA molecule -- a long sequence of *nucleotides* (A, C, T, G)

Human genome -- about 3 x 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 pai chemist's viewpoint, each stra 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 ⟷ T C ⟷ 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

most

Eukaryotes - ~~all~~ 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”

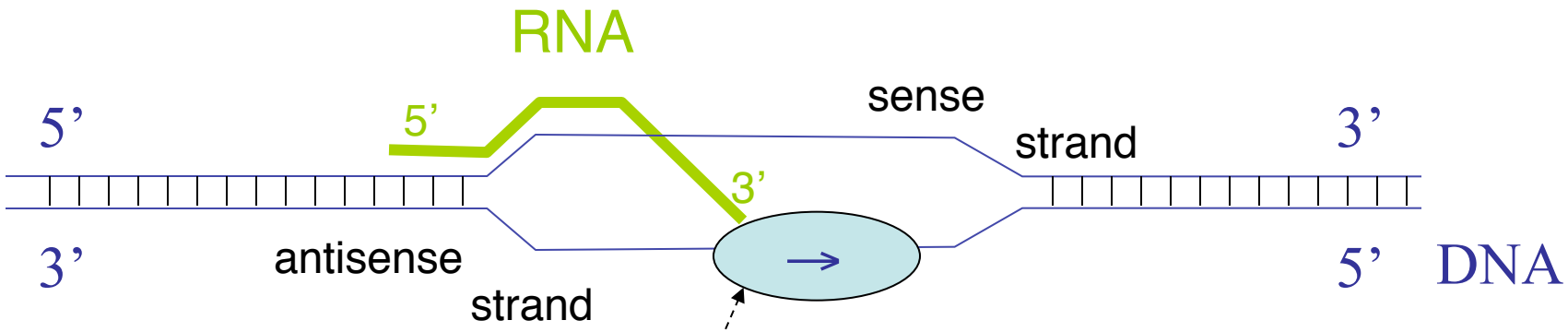
Genes encode proteins

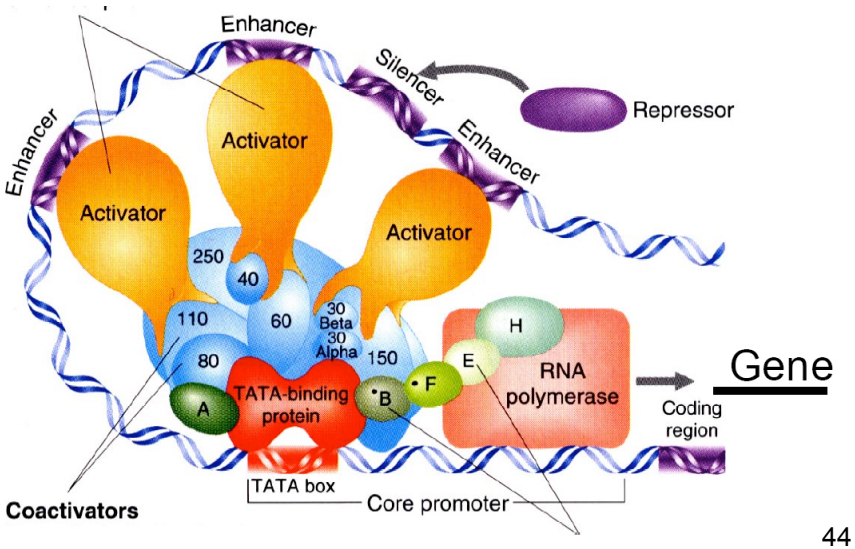
DNA transcribed into messenger RNA

mRNA translated into proteins

Triplet code (codons)

Transcription: DNA → RNA



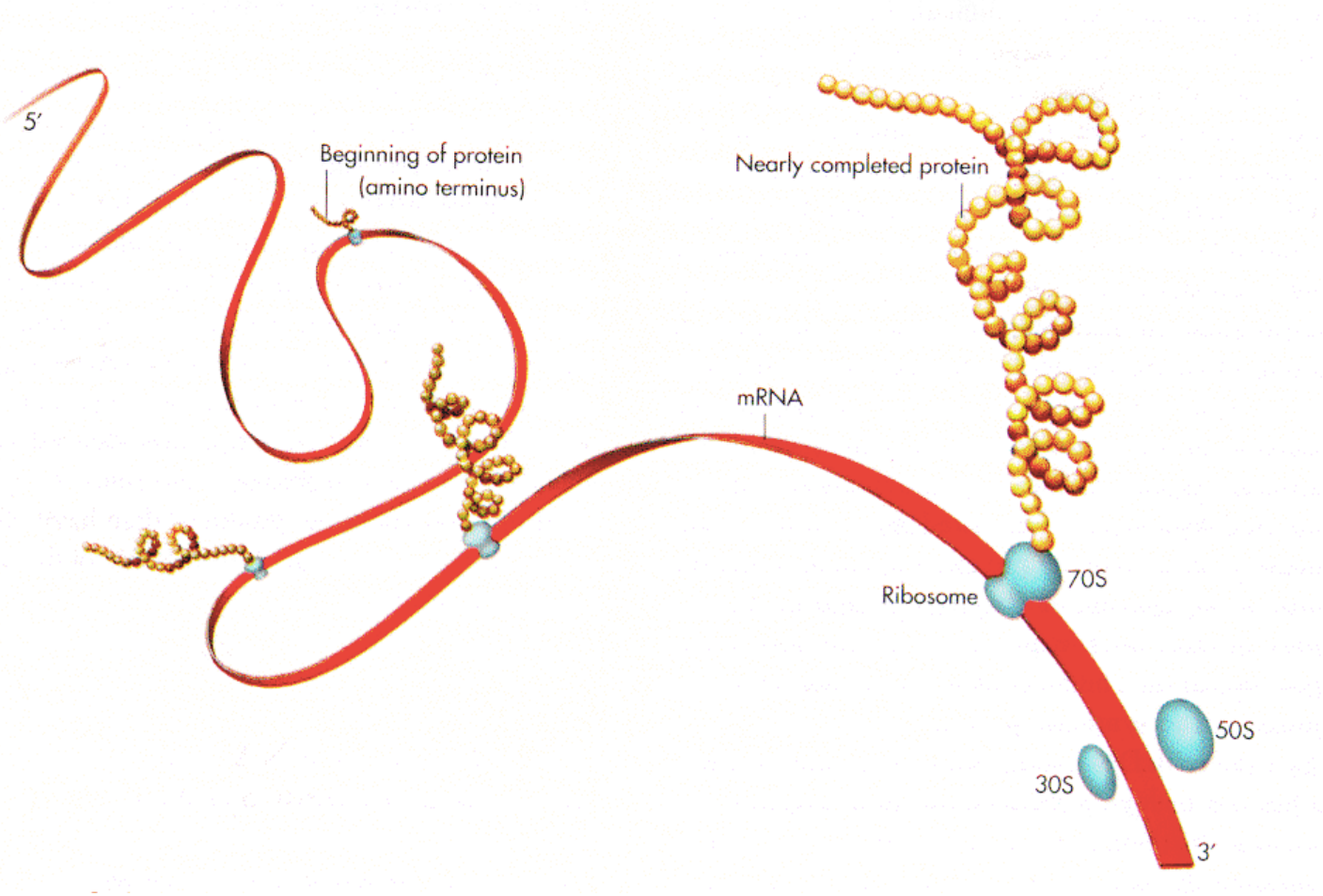


Codons & The Genetic Code

First Base	Second Base U	Second Base C	Second Base A	Second Base G	Third Base
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	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

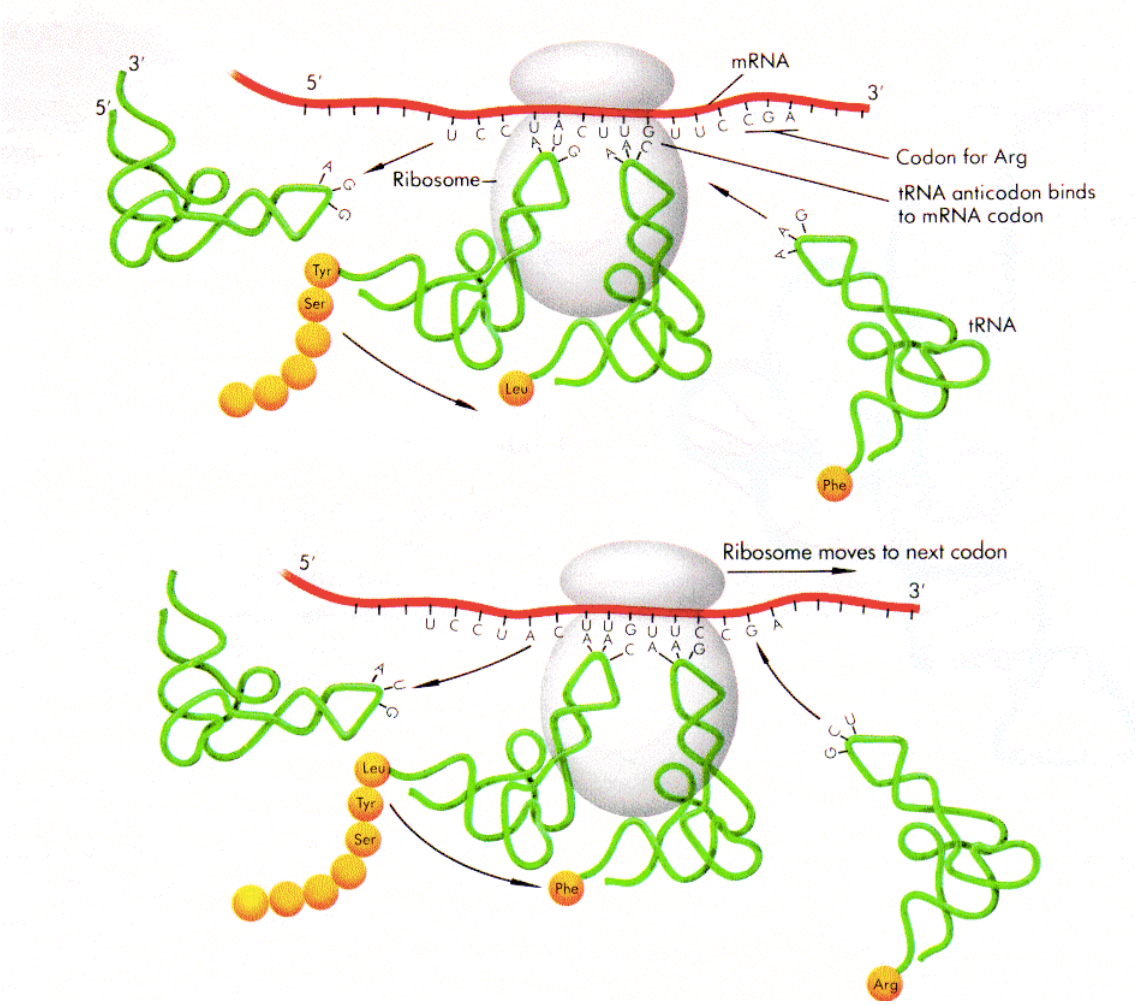
Ala : Alanine
Arg : Arginine
Asn : Asparagine
Asp : Aspartic acid
Cys : Cysteine
Gln : Glutamine
Glu : Glutamic acid
Gly : Glycine
His : Histidine
Ile : Isoleucine
Leu : Leucine
Lys : Lysine
Met : Methionine
Phe : Phenylalanine
Pro : Proline
Ser : Serine
Thr : Threonine
Trp : Tryptophane
Tyr : Tyrosine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

Translation: mRNA → Protein



Watson, Gilman, Witkowski, & Zoller, 1992

Ribosomes



Watson, Gilman, Witkowski, & Zoller, 1992

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
MimiVirus	1,200,000	1,260
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	3.3×10^9	~25,000

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 >> 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

Homework #1 (partial)

Read Hunter's "bio for cs" primer;

Find & read another

Post a few sentences saying

- What you read (give me a link or citation)
- Critique it for your meeting your needs
- Who would it have been good for, if not you

See class web (coming soon) for more details

Digression: Evolution & scientific literacy

"human beings, as we know them, developed from earlier species of animals"

(avoiding the now politically charged word "evolution")

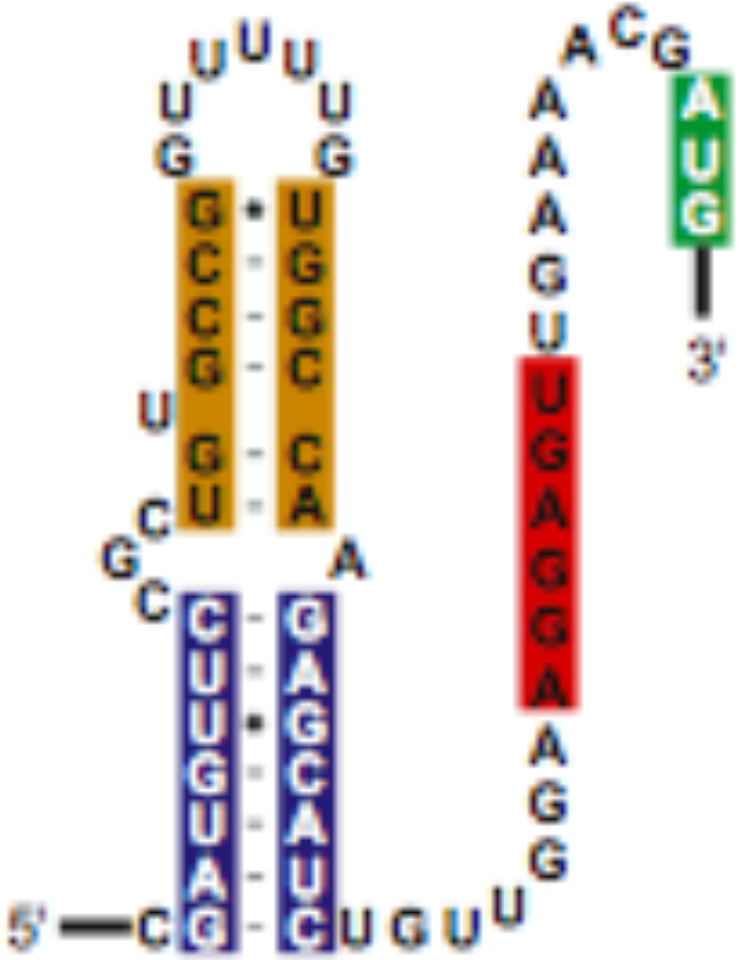
from 1985 to 2005, the % of Americans

- rejecting: declined from 48% to 39%
- accepting: also declined 45% to 40
- uncertain: increased 7% to 21%

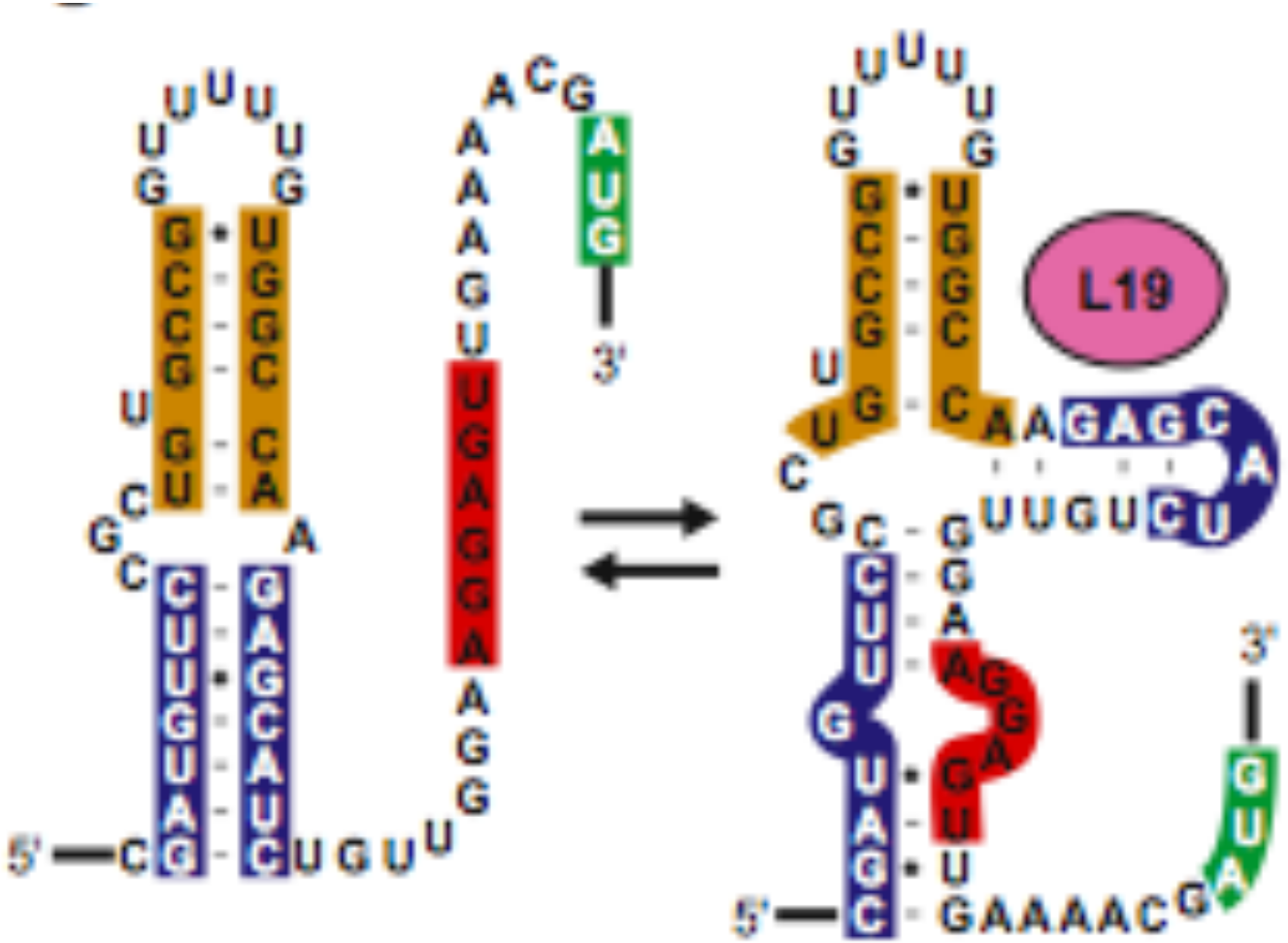
In a 2005 survey,the proportion of adults who accept evolution in 34 European countries and Japan, the United States ranked 33rd, just above Turkey.

http://biology.plosjournals.org/perlserv/?request=get-document&doi=10.1371/journal.pbio.0040167

An RNA Structure

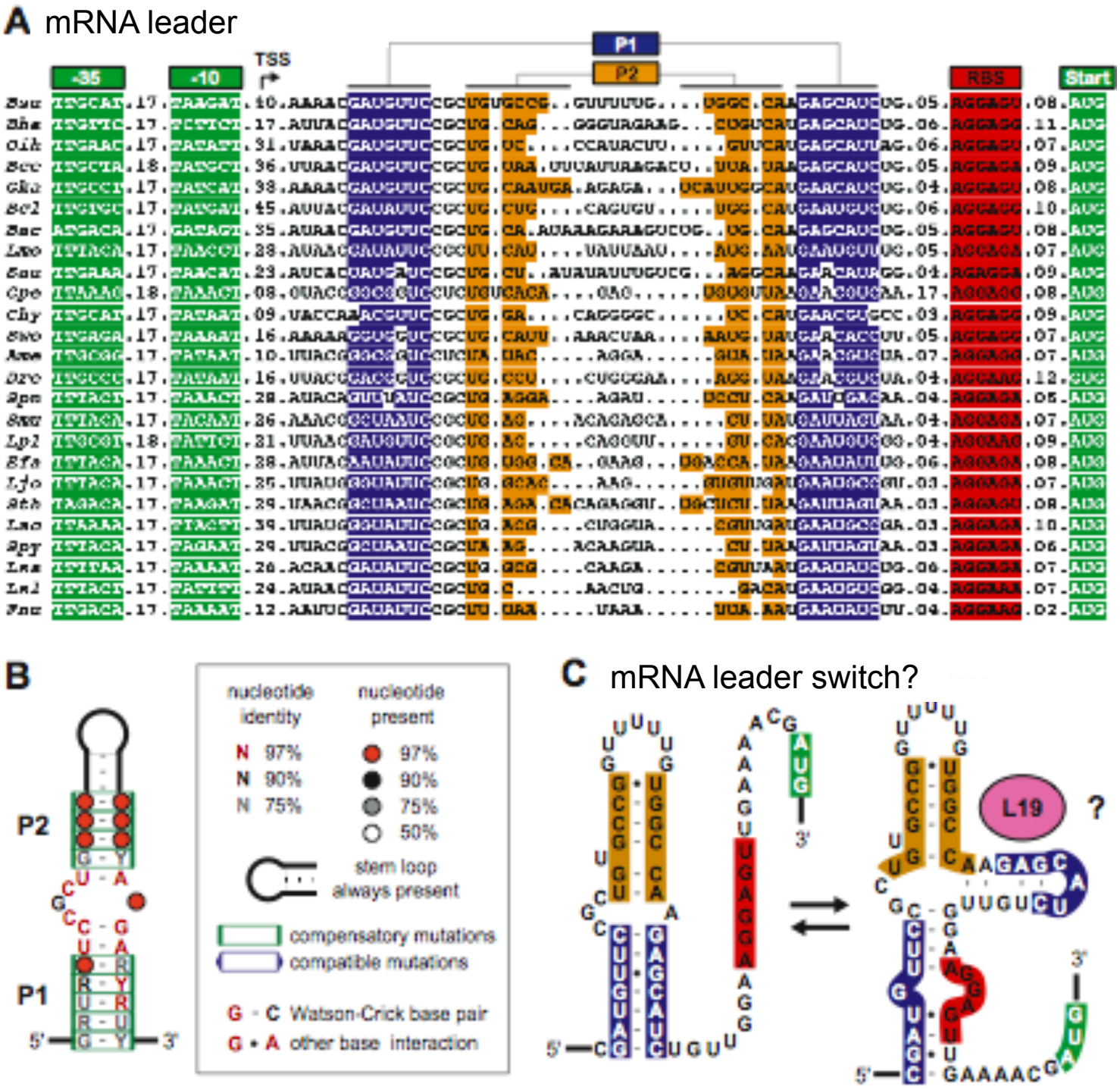


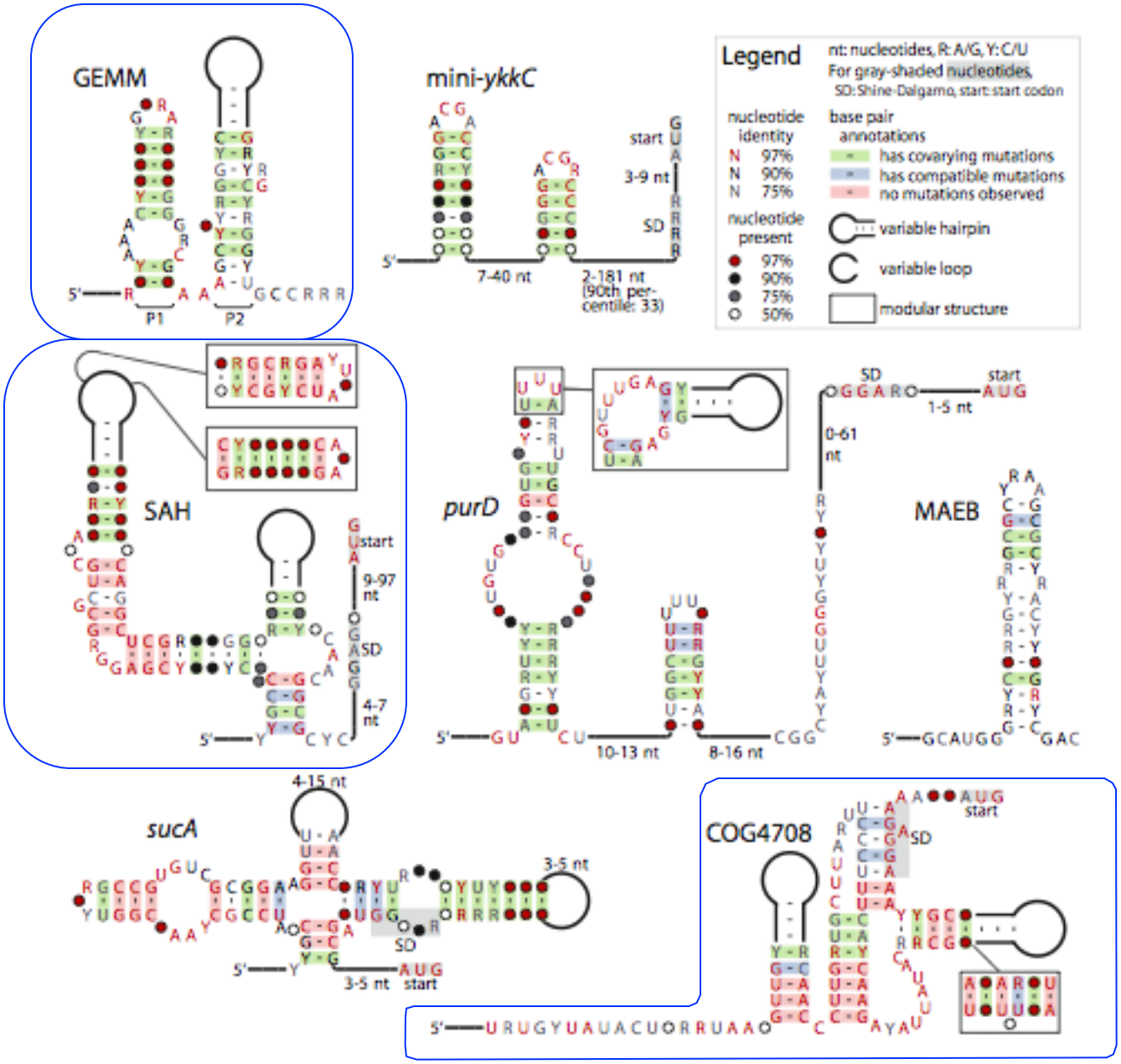
An RNA Sensor & On/Off Switch



L19 absent: Gene On

L19 present: Gene Off


A mRNA leader
P1
P2
TSS
-35
-10
RBS
Start
B
P2
P1
5'
3'
nucleotide identity
nucleotide present
N 97%
N 90%
N 75%
97%
90%
75%
50%
stem loop always present
compensatory mutations
compatible mutations
G - C Watson-Crick base pair
G • A other base interaction
C mRNA leader switch?
L19
?
5'
3'



boxed = confirmed riboswitch (+2 more)

Bottom Line

CFG technology is a key tool for RNA description, discovery and search

A very active research area. (Some call RNA the “dark matter” of the genome.)

Huge compute hog: results above represent hundreds of CPU-years, and smart algorithms can have a big impact