CSE 427
Computational Biology

https://courses.cs.washington.edu/courses/cse427

Larry Ruzzo, Daniel Jones

Autumn 2016

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

-- Chinese Proverb

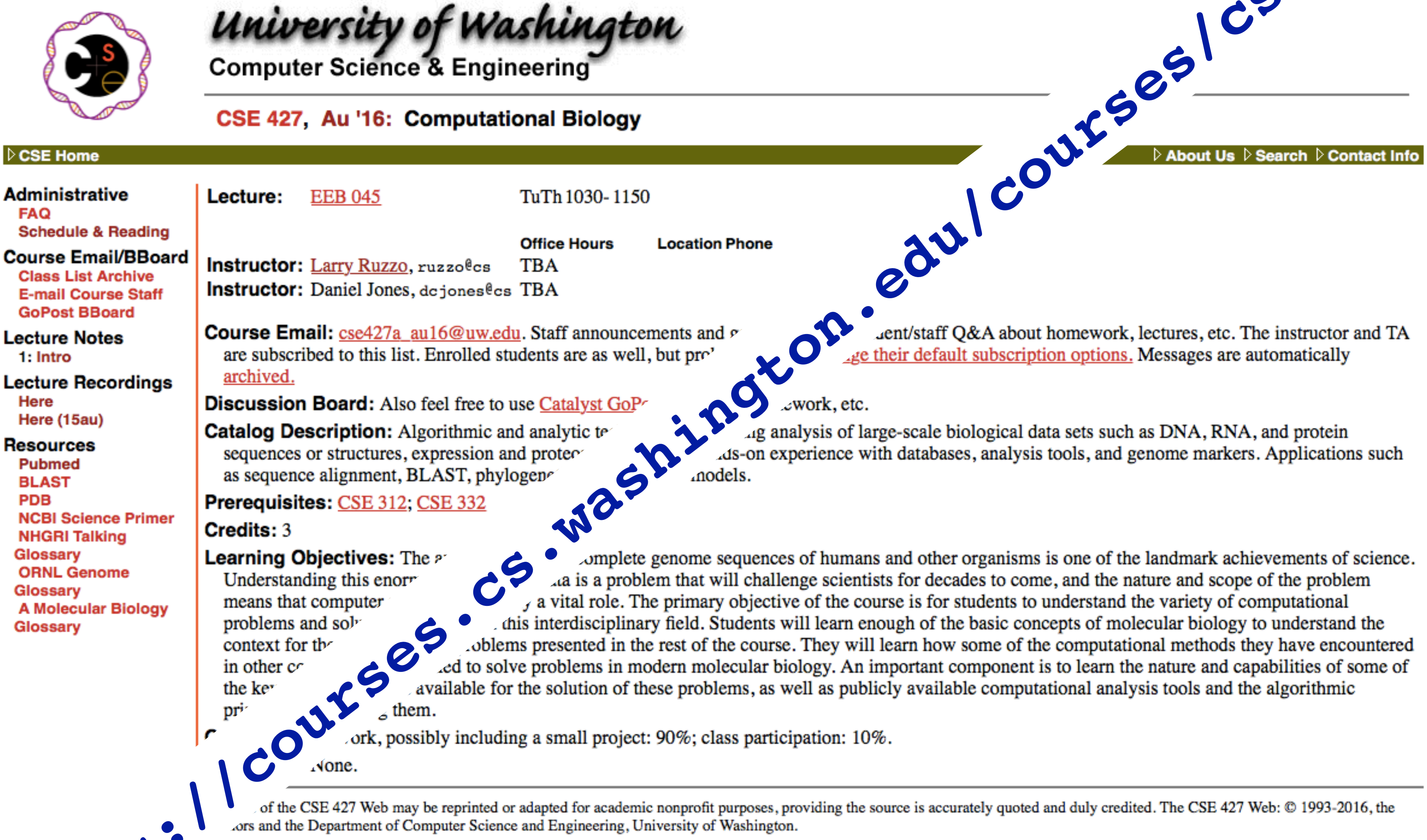
Today

Admin

Why Comp Bio?

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

Admin Stuff



Administrative
- FAQ
- Schedule & Reading

Course Email/BBoard
- Class List Archive
- E-mail Course Staff
- GoPost BBoard

Lecture Notes
- 1: Intro

Lecture Recordings
- Here
- Here (15au)

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

		Office Hours	Location Phone
Lecture:	EEB 045	TuTh 1030- 1150	
Instructor:	Larry Ruzzo, `ruzzo@cs`	TBA	
Instructor:	Daniel Jones, `dcjones@cs`	TBA	

Course Email: cse427a_au16@uw.edu. Staff announcements and ent/staff Q&A about homework, lectures, etc. The instructor and TA are subscribed to this list. Enrolled students are as well, but pr ge their default subscription options. Messages are automatically archived.

Discussion Board: Also feel free to use Catalyst GoP ework, etc.

Catalog Description: Algorithmic and analytic te ng analysis of large-scale biological data sets such as DNA, RNA, and protein sequences or structures, expression and prote ds-on experience with databases, analysis tools, and genome markers. Applications such as sequence alignment, BLAST, phylogen models.

Prerequisites: CSE 312; CSE 332

Credits: 3

Learning Objectives: The omplete genome sequences of humans and other organisms is one of the landmark achievements of science. Understanding this enor a is a problem that will challenge scientists for decades to come, and the nature and scope of the problem means that computer a vital role. The primary objective of the course is for students to understand the variety of computational problems and sol this interdisciplinary field. Students will learn enough of the basic concepts of molecular biology to understand the context for th oblems presented in the rest of the course. They will learn how some of the computational methods they have encountered in other c ed to solve problems in modern molecular biology. An important component is to learn the nature and capabilities of some of the ke available for the solution of these problems, as well as publicly available computational analysis tools and the algorithmic pri them.

ork, possibly including a small project: 90%; class participation: 10%.

None.

of the CSE 427 Web may be reprinted or adapted for academic nonprofit purposes, providing the source is accurately quoted and duly credited. The CSE 427 Web: © 1993-2016, the ors and the Department of Computer Science and Engineering, University of Washington.

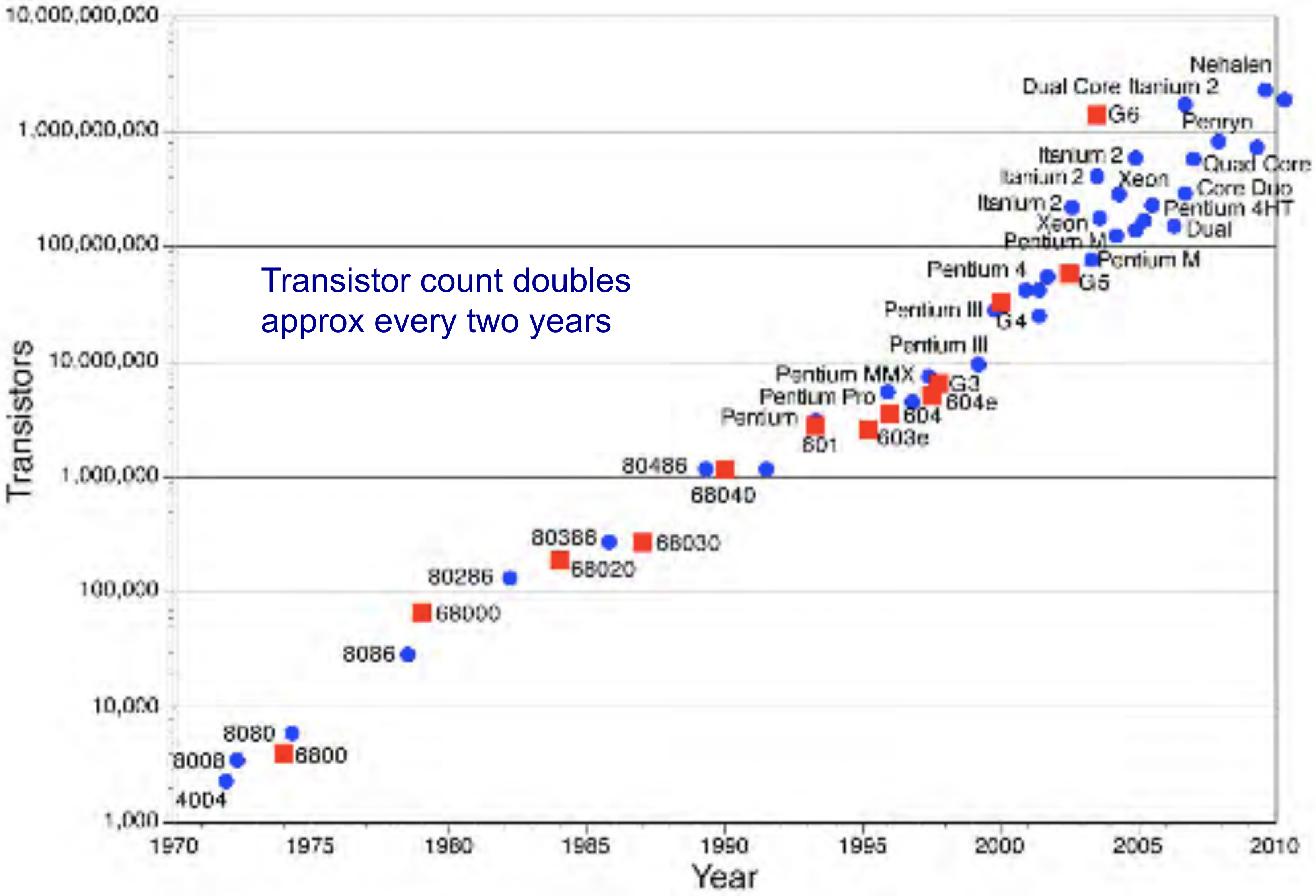
https://courses.cs.washington.edu/courses/cse427

Course Mechanics & Grading

- Web:
  - https://courses.cs.washington.edu/courses/cse427
- Reading
- In class discussion
- Homeworks
  - paper exercises & programming
- No exams, but probable oversized last homework in lieu of final

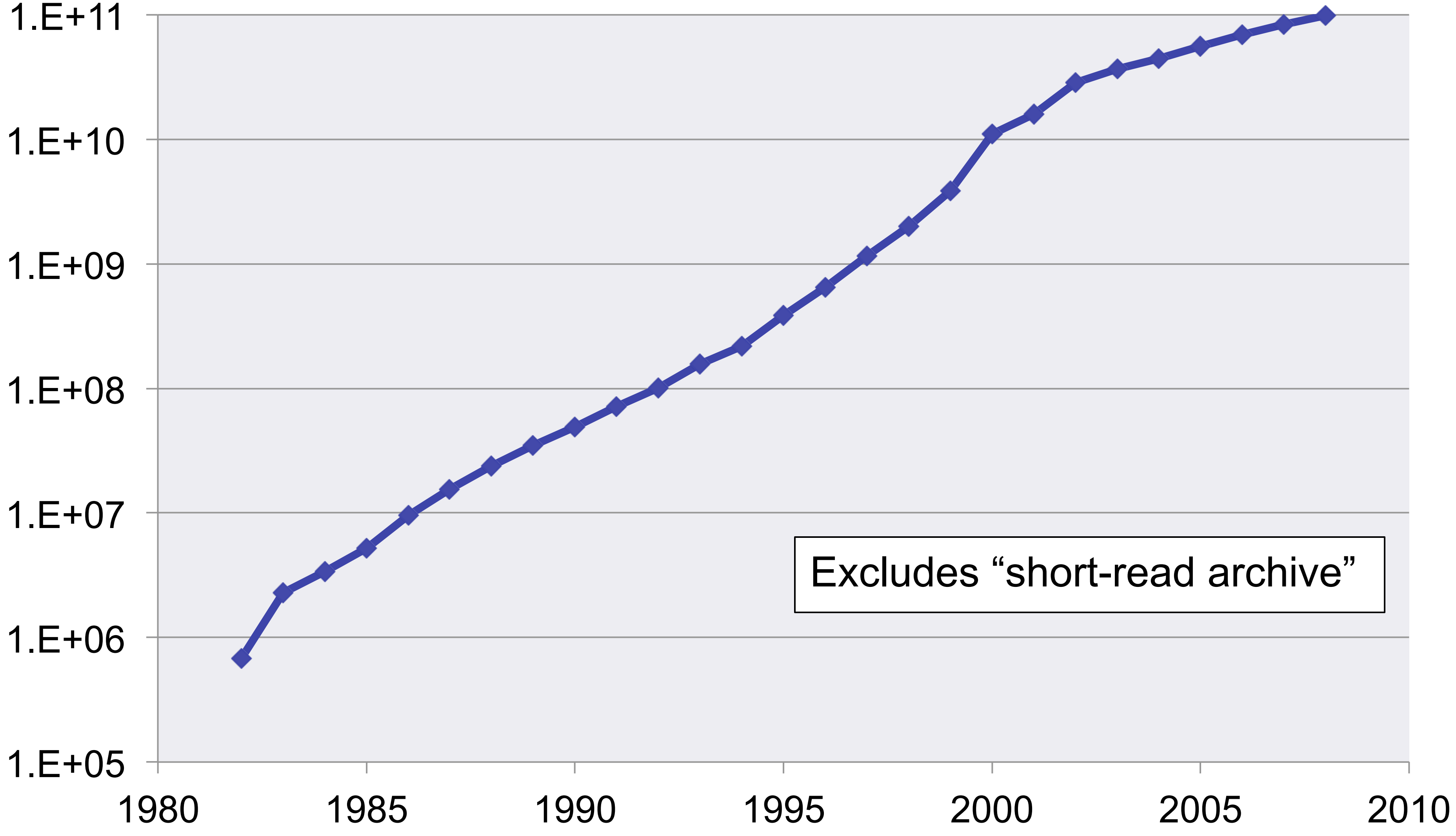
Background & Motivation

Moore’s Law



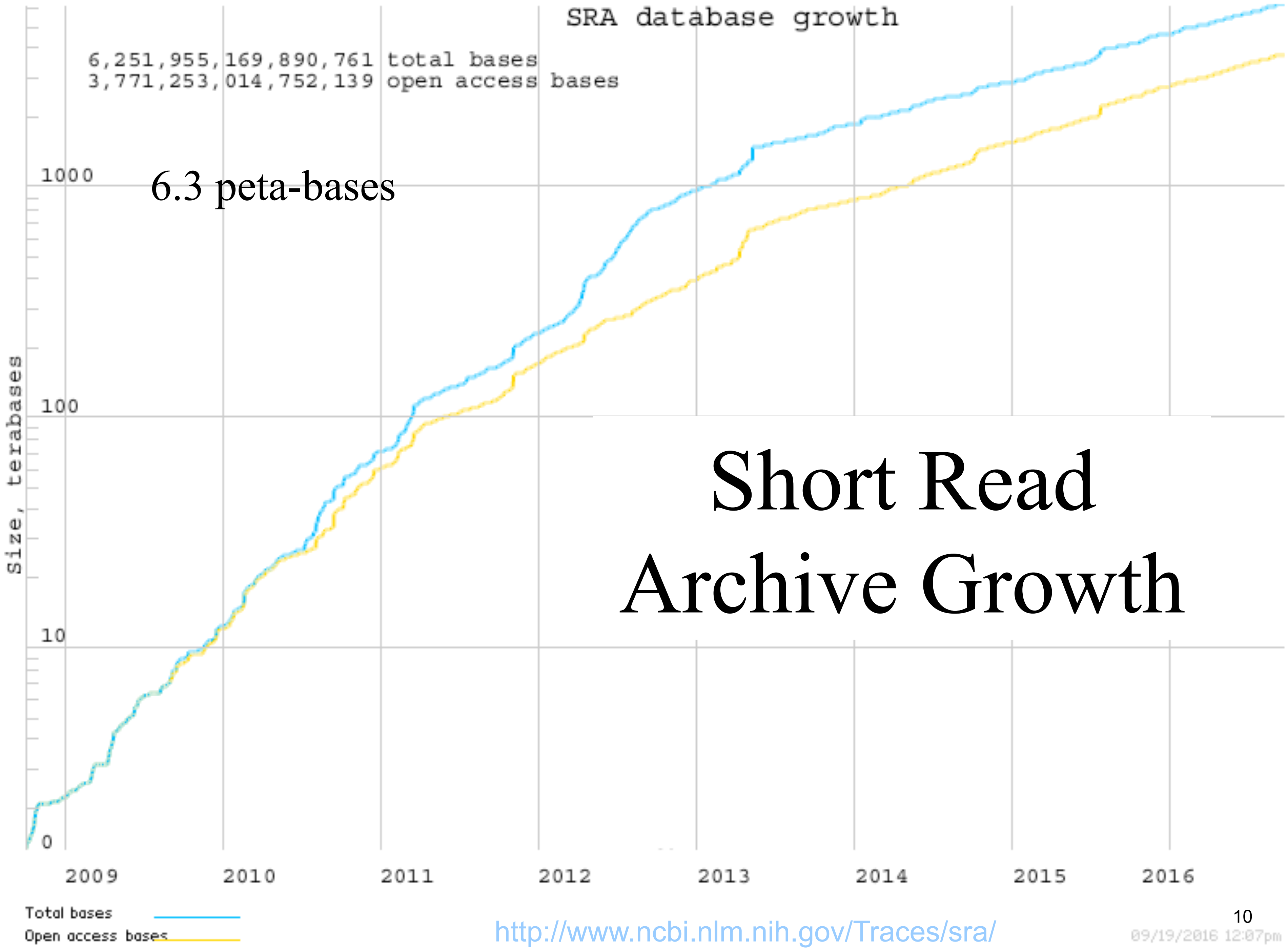
Transistor count doubles approx every two years

Growth of GenBank (Base Pairs)



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

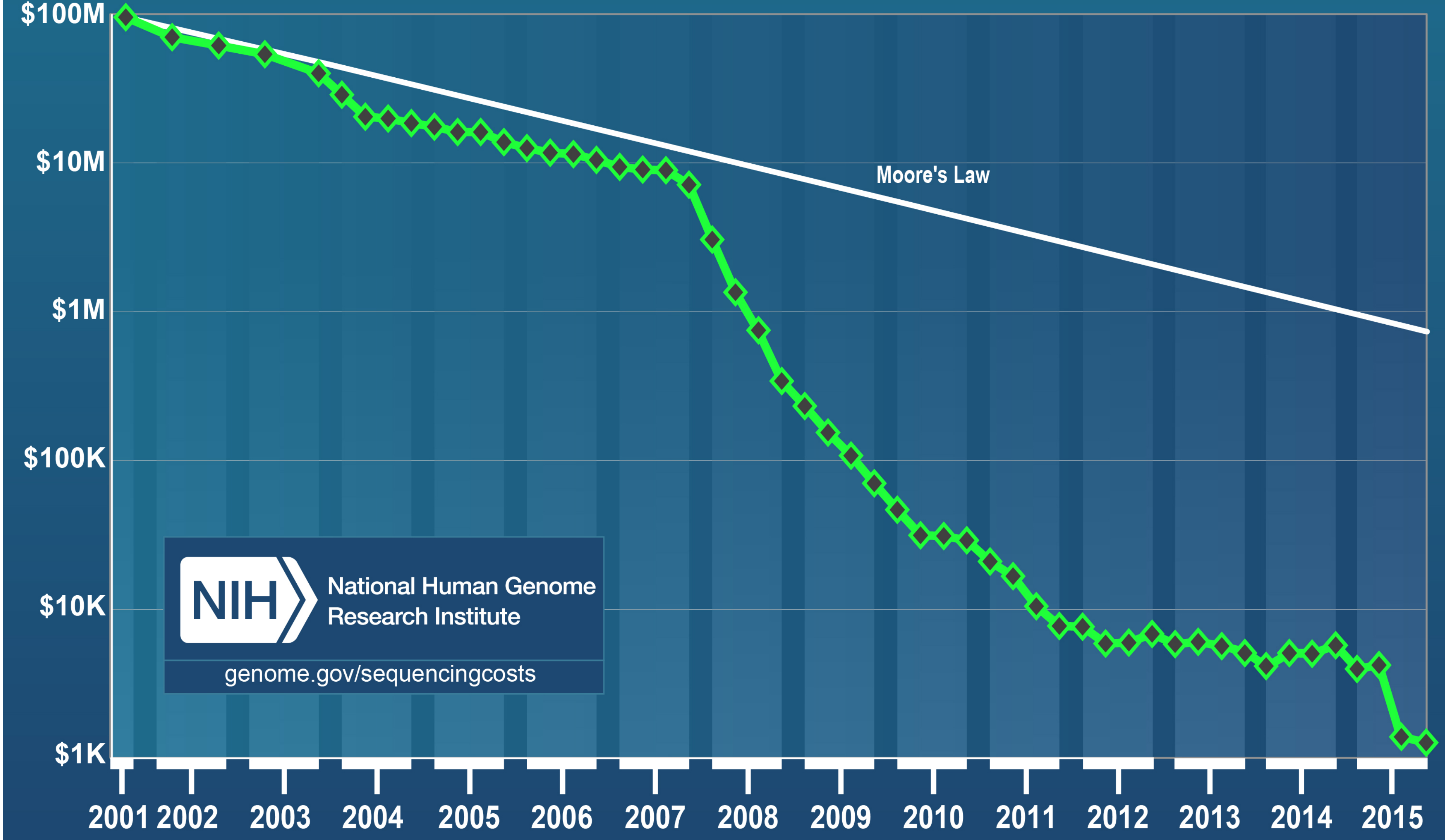
Short Read Archive Growth



6.3 peta-bases

http://www.ncbi.nlm.nih.gov/Traces/sra/

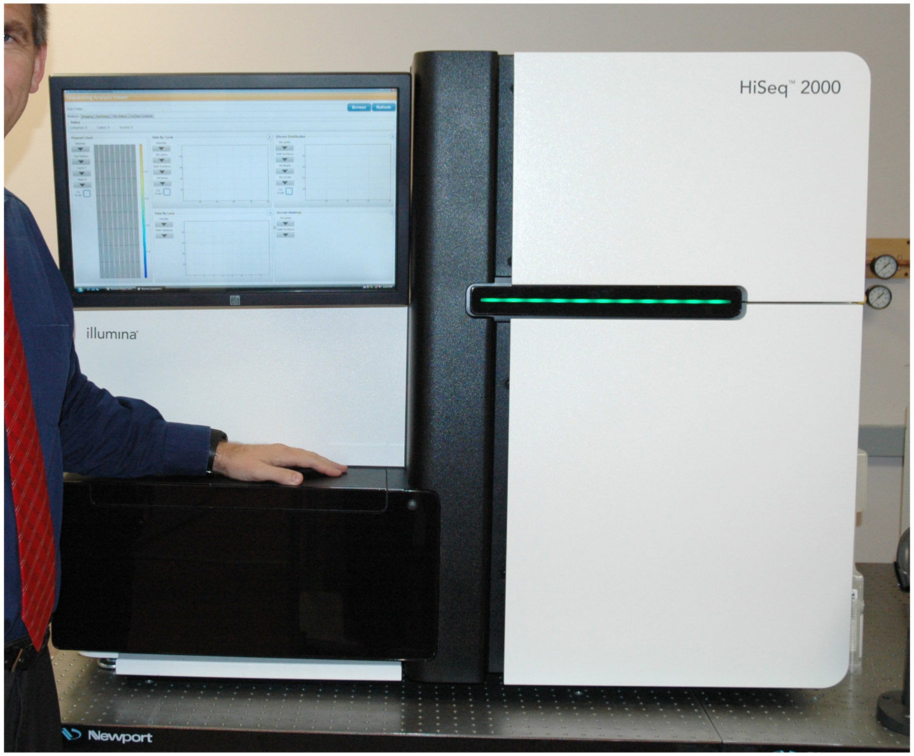
Cost per Genome



https://www.genome.gov/sequencingcostsdata/

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



PLOS | BIOLOGY

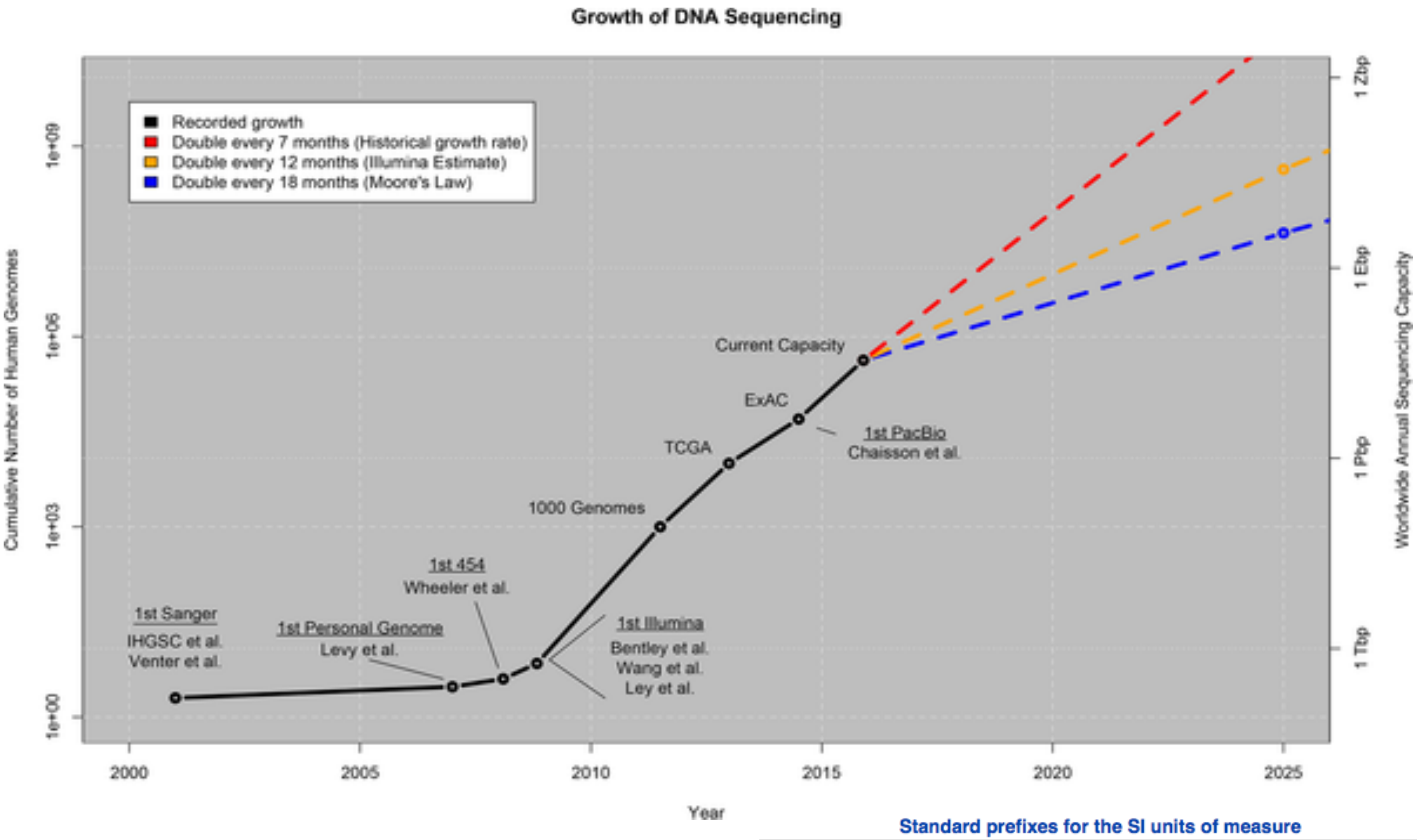
PERSPECTIVE

Big Data: Astronomical or Genomical?

Zachary D. Stephens[1], Skylar Y. Lee[1], Faraz Faghri[2], Roy H. Campbell[2], Chengxiang Zhai[3], Miles J. Efron[4], Ravishankar Iyer[1], Michael C. Schatz[5]*, Saurabh Sinha[3]*, Gene E. Robinson[6]*

PLoS Biol 13(7): e1002195. doi:10.1371/journal.pbio.1002195
http://127.0.0.1:8081/plosbiology/article?id=info:doi/10.1371/journal.pbio.1002195

Fig 1. Growth of DNA sequencing.



Standard prefixes for the SI units of measure

Prefix name		deca	hecto	kilo	mega	giga	tera	peta	exa	zetta
Prefix symbol		da	h	k	M	G	T	P	E	Z
Factor	10^0	10^1	10^2	10^3	10^6	10^9	10^{12}	10^{15}	10^{18}	10^{21}

Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, et al. (2015) Big Data: Astronomical or Genomical?. PLoS Biol 13(7): e1002195. doi:10.1371/journal.pbio.1002195
http://127.0.0.1:8081/plosbiology/article?id=info:doi/10.1371/journal.pbio.1002195

Table 1. Four domains of Big Data in 2025.

In each of the four domains, the projected annual storage and computing needs are presented across the data lifecycle.

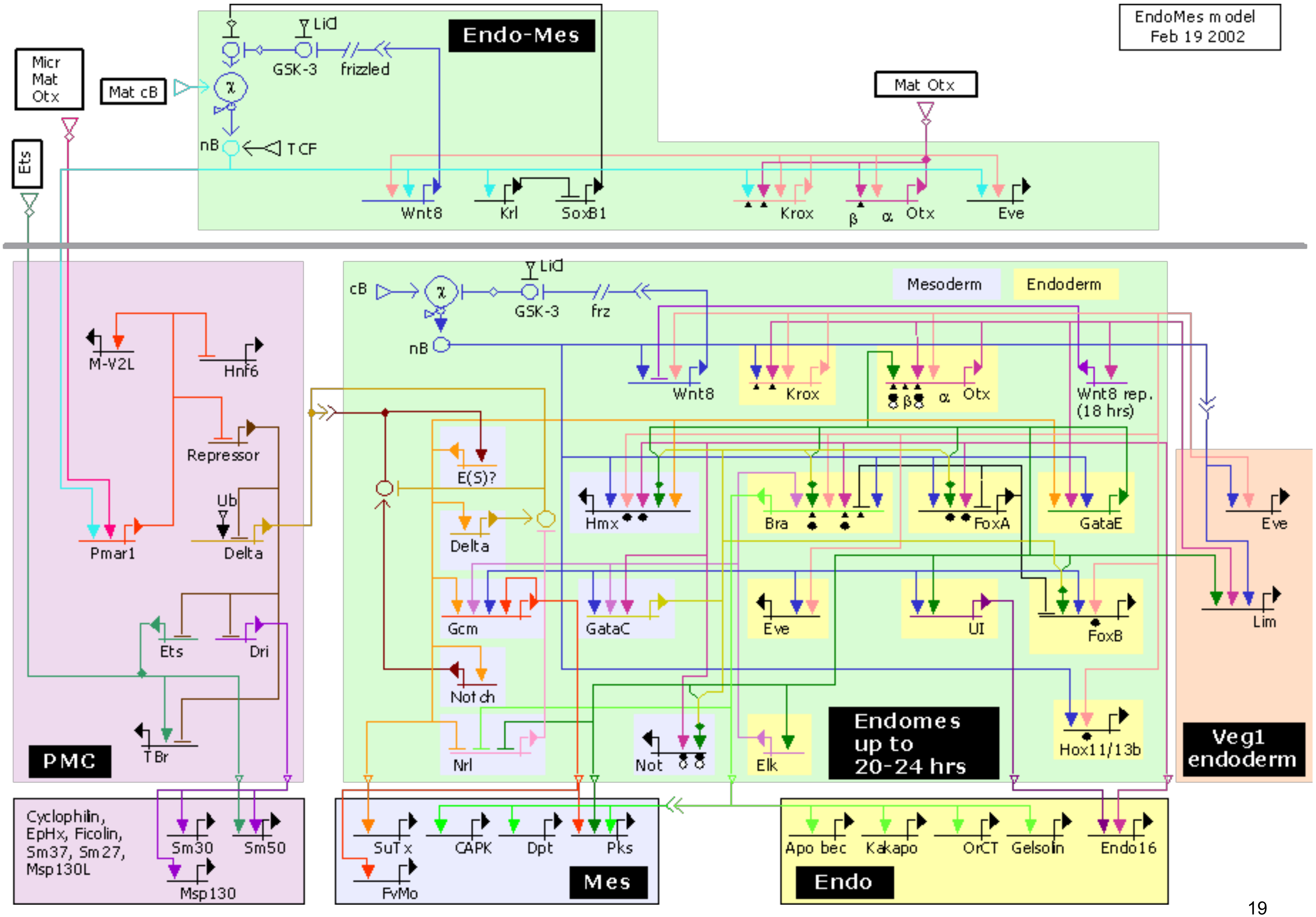
Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion CPU hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movements

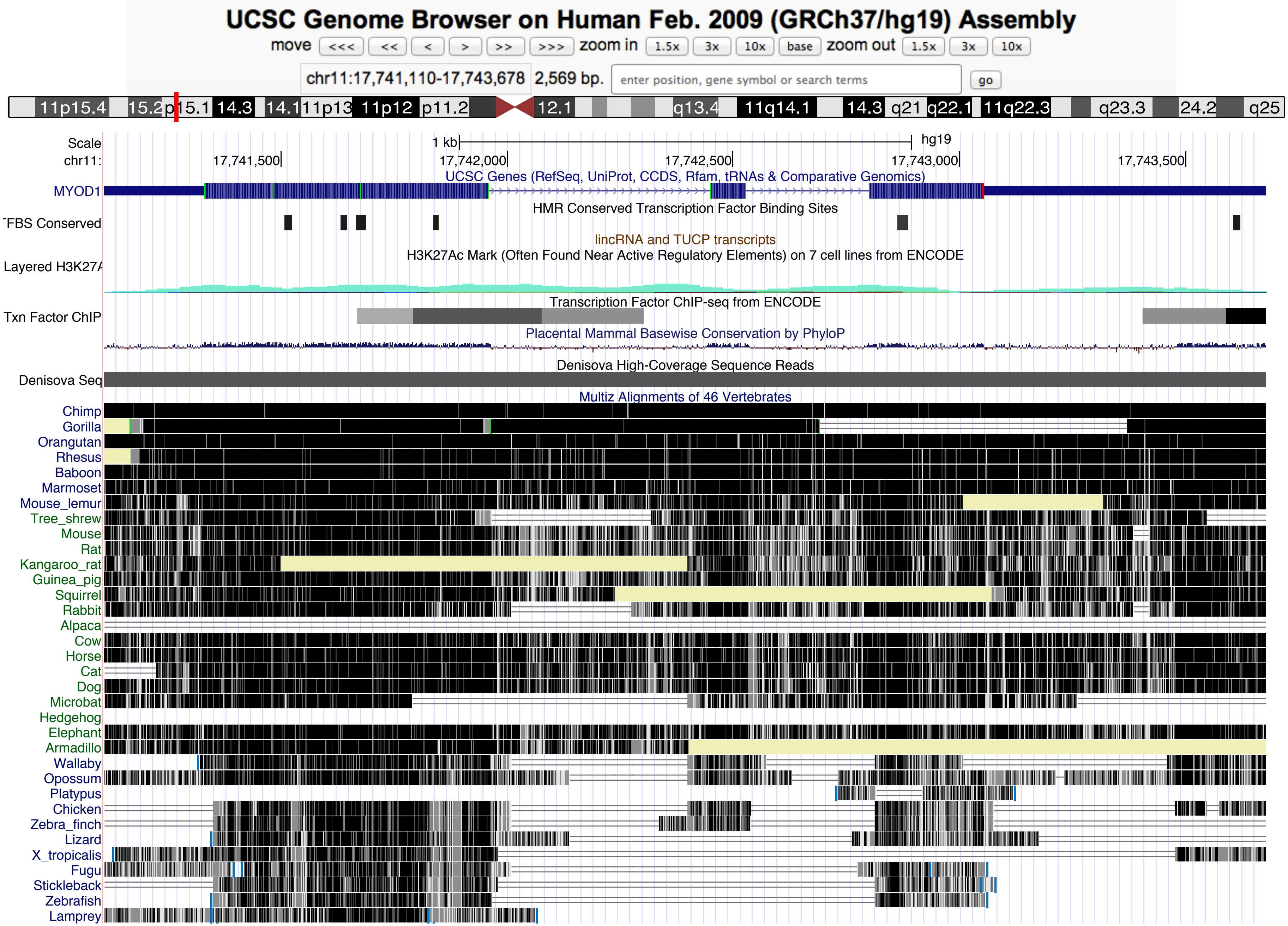
Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, et al. (2015) Big Data: Astronomical or Genomical?. PLoS Biol 13(7): e1002195. doi: 10.1371/journal.pbio.1002195
http://127.0.0.1:8081/plosbiology/article?id=info:doi/10.1371/journal.pbio.1002195

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
cB
LiCl
GSK-3
frz
nB
Mesoderm
Endoderm
Wnt8
Krox
β
α
Otx
Wnt8 rep.
(18 hrs)
E(S)?
Hmx
Bra
FoxA
GataE
Eve
Lim
Delta
Gcm
GataC
Eve
Ul
FoxB
Notch
Nrl
Not
Elk
Endomes
up to
20-24 hrs
Hox11/13b
Veg1
endoderm
Cyclophilin,
EpHx, Ficolin,
Sm37, Sm27,
Msp130L
Sm30
Sm50
Msp130
SuTx
CAPK
Dpt
Pks
FvMo
Mes
Apo bec
Kakapo
OrCT
Gelsolin
Endo16
Endo


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
HMR Conserved Transcription Factor Binding Sites
TFBS Conserved
lincRNA and TUCP transcripts
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27A
Transcription Factor ChIP-seq from ENCODE
Txn Factor ChIP
Placental Mammal Basewise Conservation by PhyloP
Denisova High-Coverage Sequence Reads
Denisova Seq
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

Goals

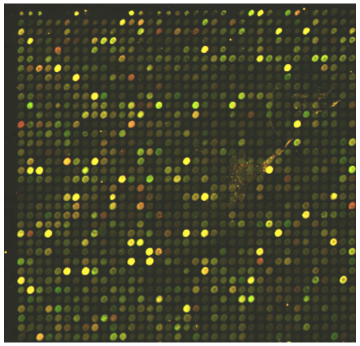
Basic biology

Disease diagnosis/prognosis/treatment

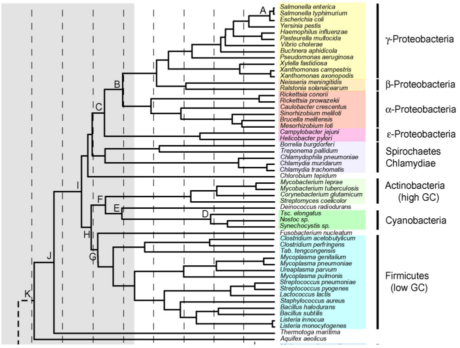
Drug discovery, validation & development

Individualized medicine

…

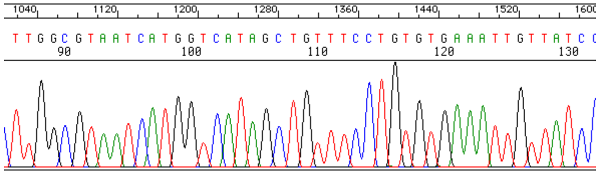


“High-Throughput BioTech”



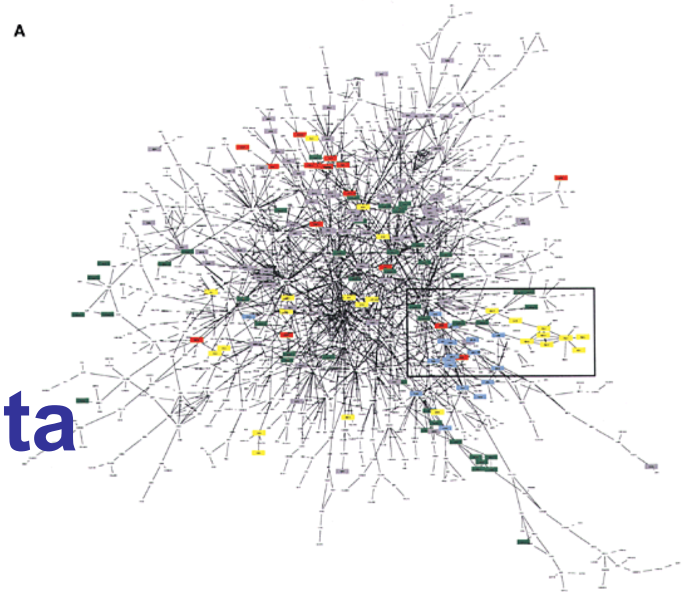
Sensors

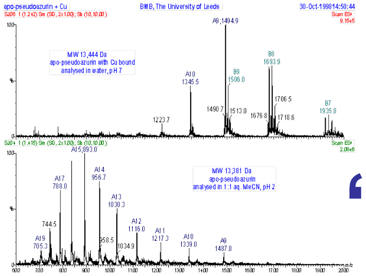
- DNA sequencing
- Microarrays/Gene expression
- Mass Spectrometry/Proteomics
- Protein/protein & DNA/protein interaction



Controls

- Cloning
- Gene knock out/knock in
- RNAi





Floods of data

“Grand Challenge” problems

What’s all the fuss?

The human genome is “finished”…
But that’s only the beginning
Explosive growth in 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 complex, 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

More Admin

Why Take This Course?

IT and Genomics are, and probably will remain, the 2 most explosively transformative technologies of your lifetimes

Even if you don’t choose to work at that interface, having some knowledge of it will be valuable

Hopefully, you will learn useful alg, ML, stats techniques and ideas for how to apply them in novel domains

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

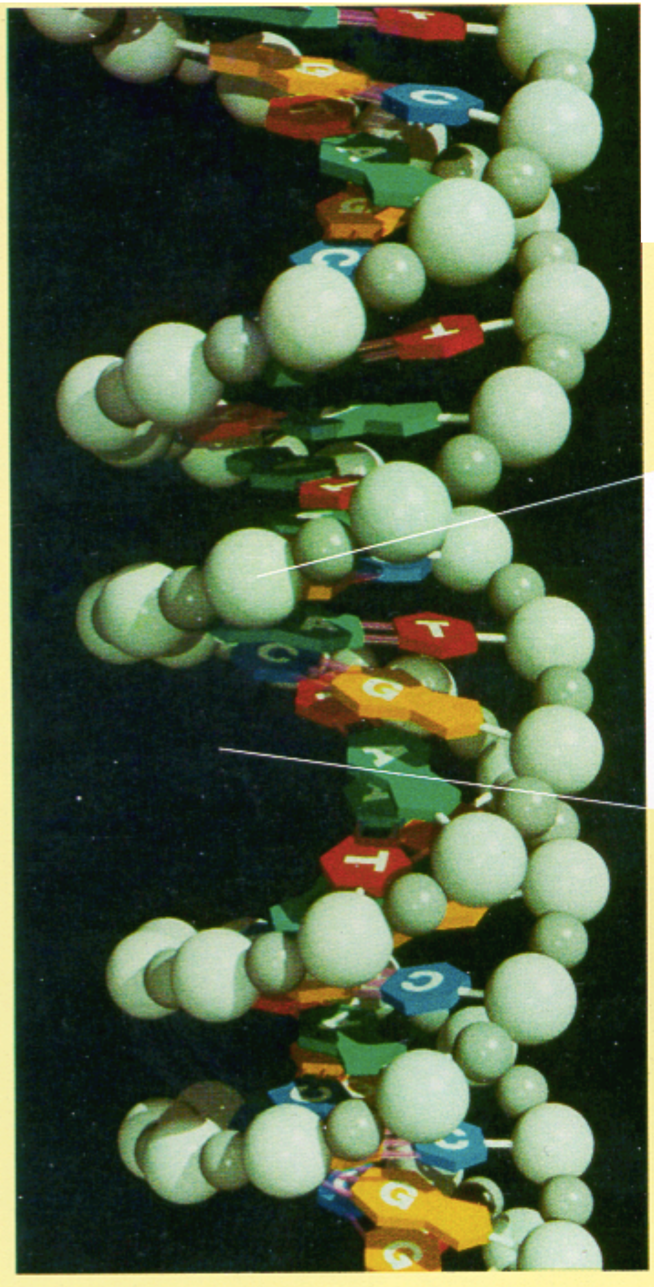
The hereditary info present in every cell

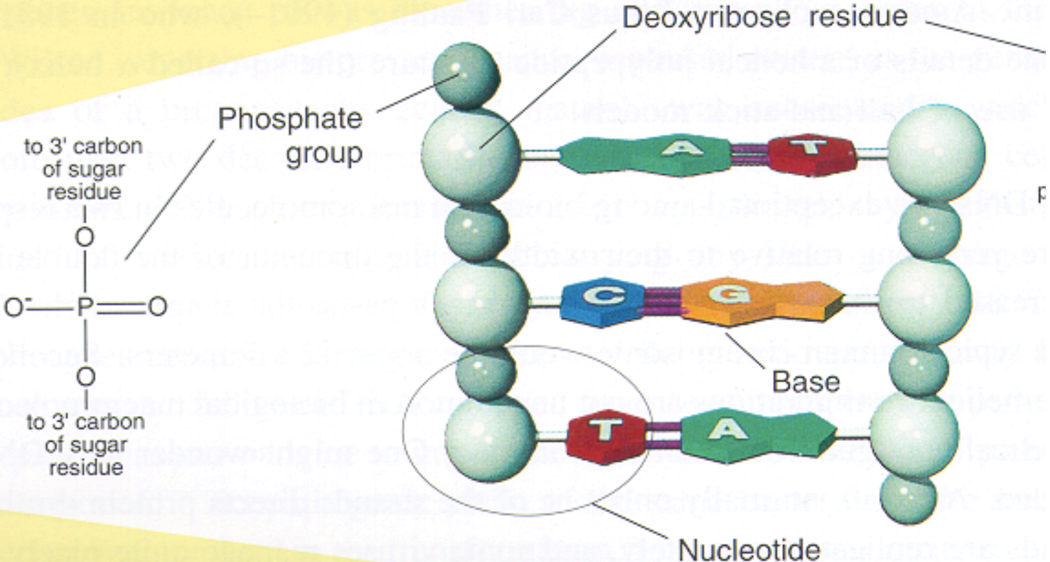
DNA molecule -- a long sequence of *nucleotides* (A, C, T, G)

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 – 1940’s

4 “bases”:

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

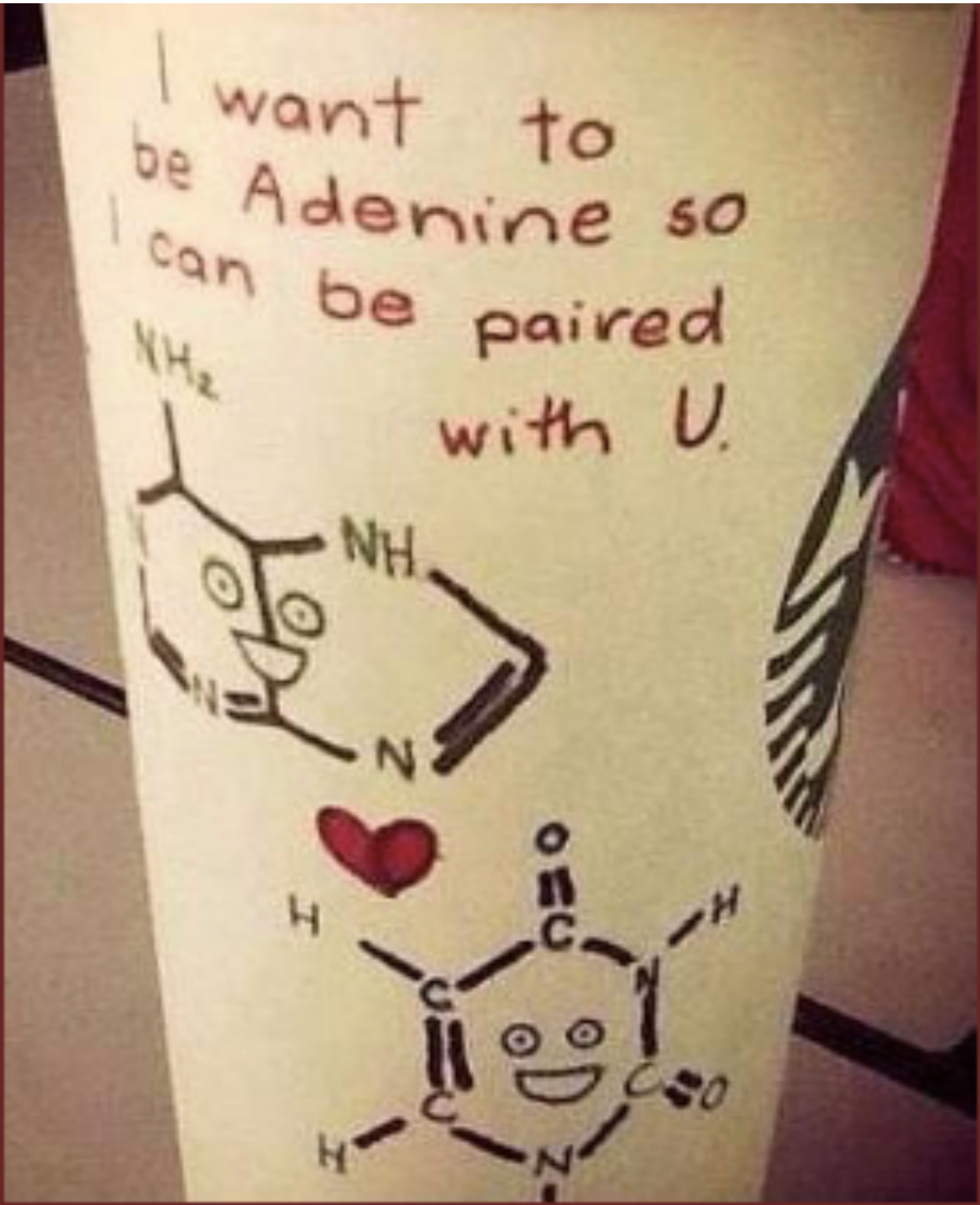
The Double Helix - Watson & Crick (& Franklin) 1953

Complementarity

- A ←→ T   C ←→ G

Visualization:

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



chemical pick-up lines

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 <u>geno</u>type →“type A” <u>pheno</u>type

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~~ most 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”

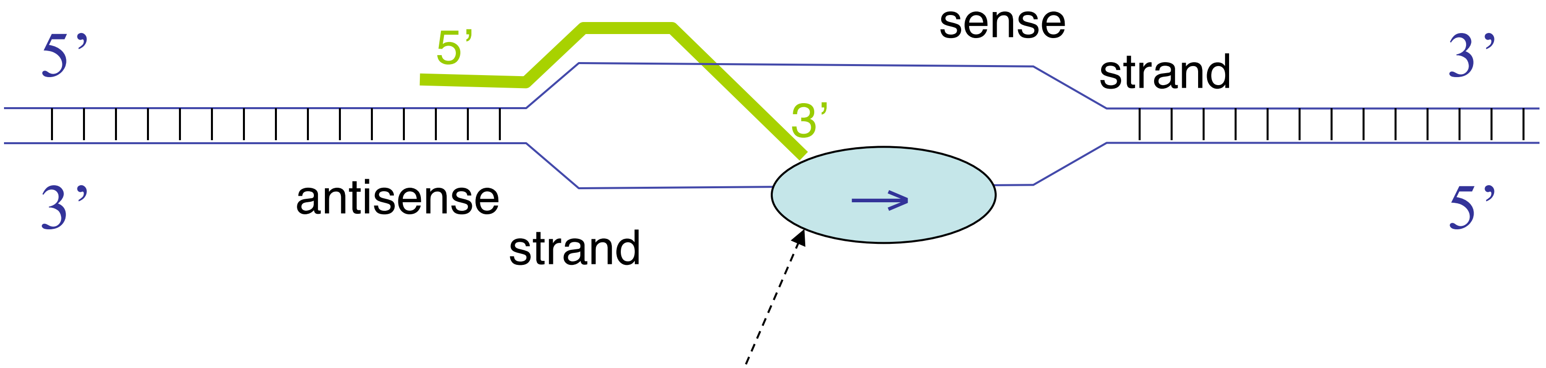
Genes encode proteins

DNA transcribed into messenger RNA

mRNA translated into proteins

Triplet code (codons)

Transcription: DNA → RNA



RNA

5’

3’

sense

strand

3’

5’

3’

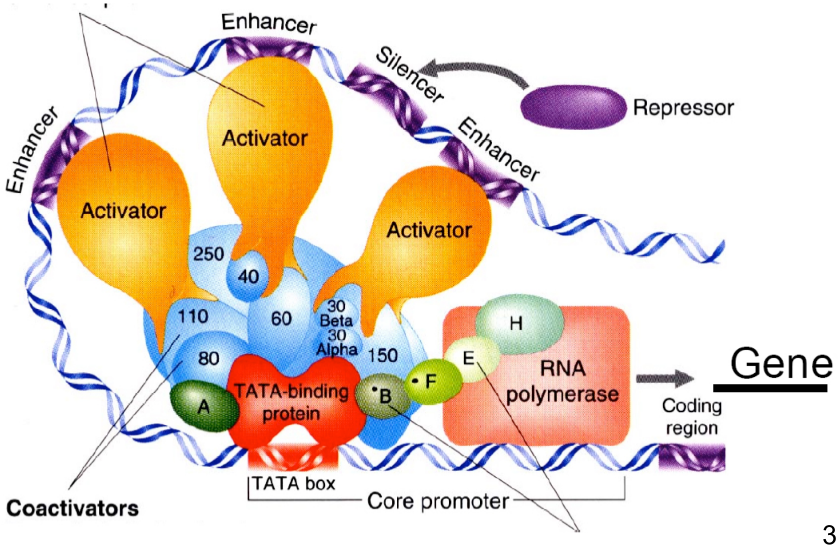
antisense

strand

5’

DNA

RNA polymerase

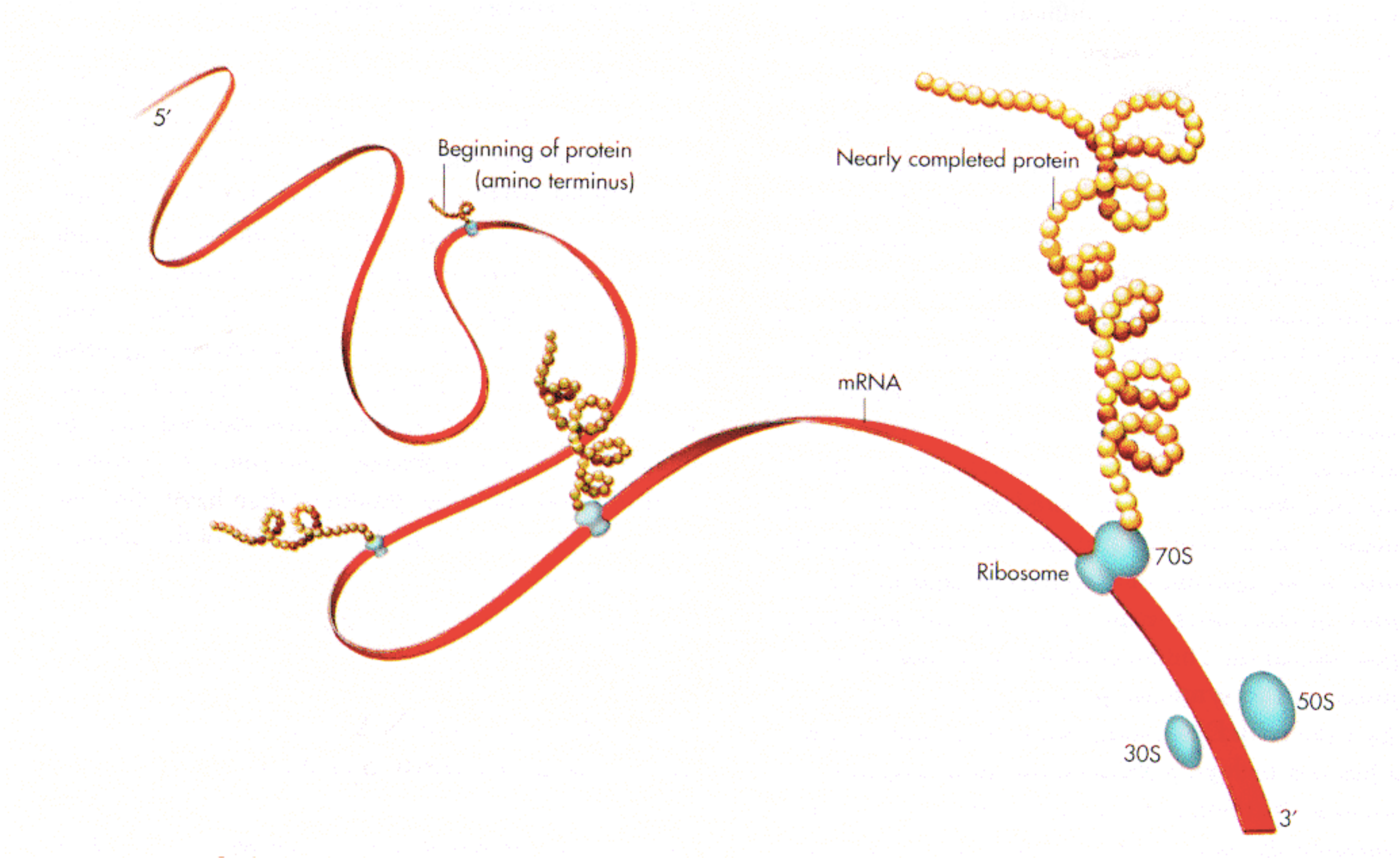


Codons & The Genetic Code

		Second Base					
		U	**C**	**A**	**G**		
First Base	**U**	Phe	Ser	Tyr	Cys	**U**	**Third Base**
		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**	

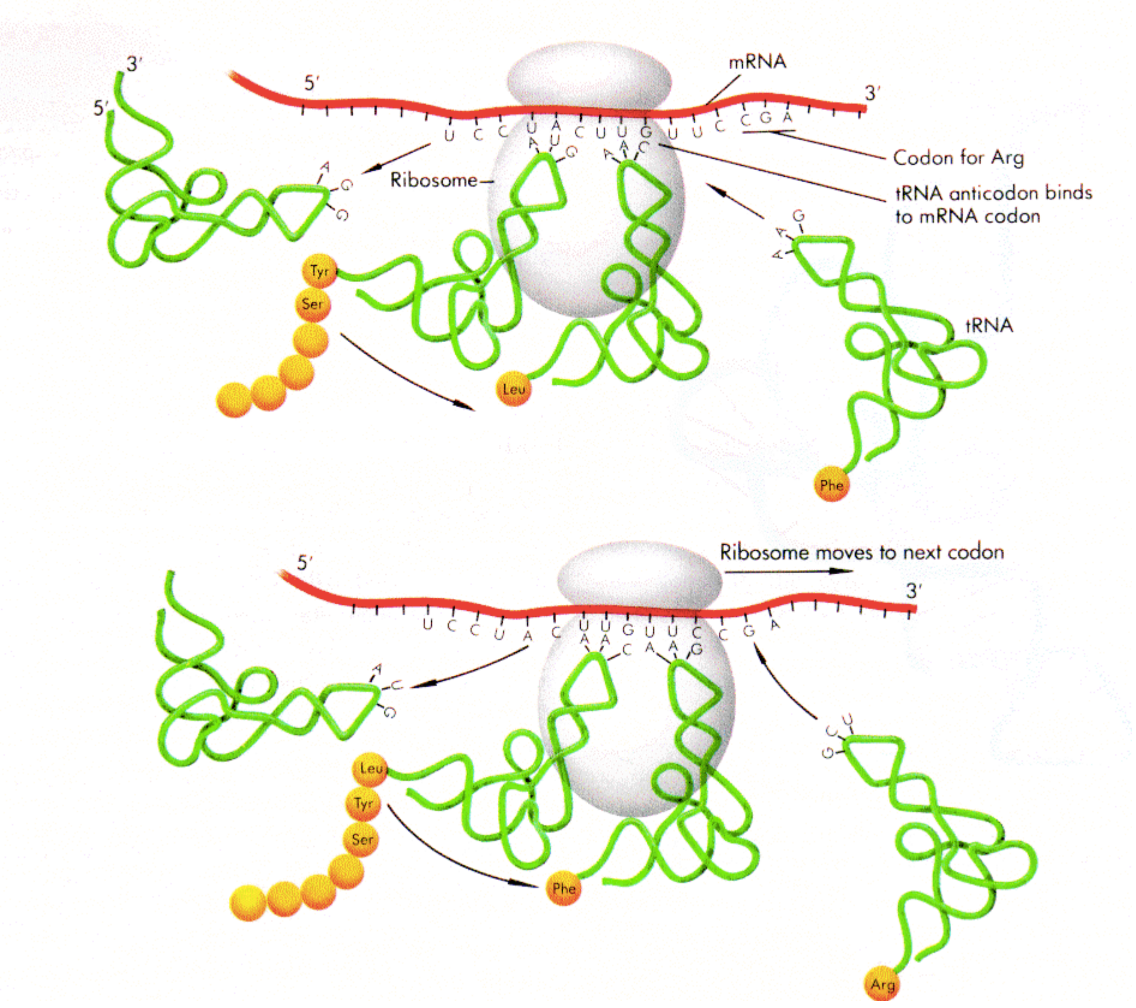
Ala : Alanine
Arg : Arginine
Asn : Asparagine
Asp : Aspartic acid
Cys : Cysteine
Gln : Glutamine
Glu : Glutamic acid
Gly : Glycine
His : Histidine
Ile : Isoleucine
Leu : Leucine
Lys : Lysine
Met : Methionine
Phe : Phenylalanine
Pro : Proline
Ser : Serine
Thr : Threonine
Trp : Tryptophane
Tyr : Tyrosine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

Translation: mRNA → Protein



Watson, Gilman, Witkowski, & Zoller, 1992

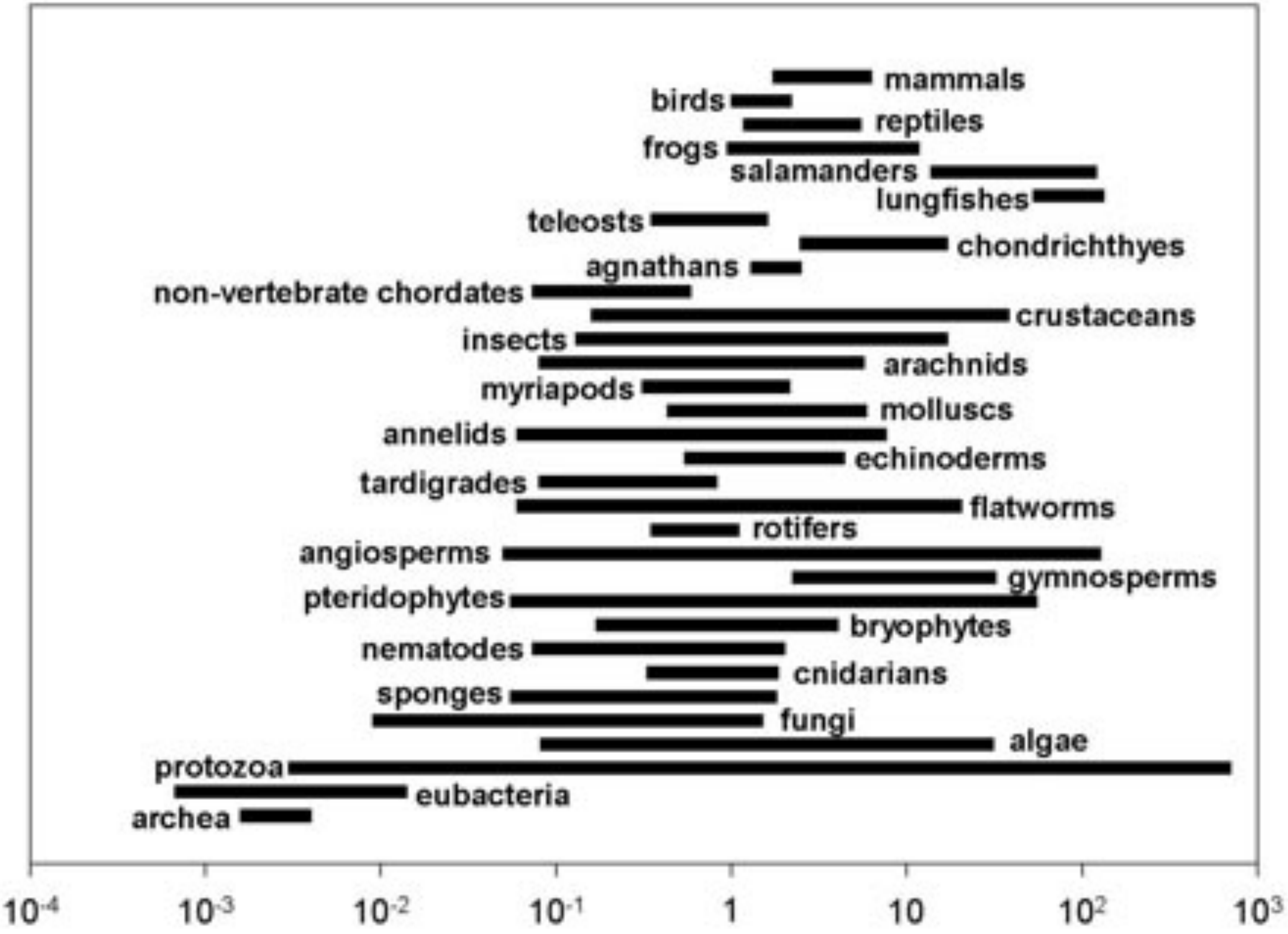
Ribosomes



Watson, Gilman, Witkowski, & Zoller, 1992

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	3.3×10^9	~20,000
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 unexpectedly many proteins, due to *alternative processing*

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