CSE 427
Computational Biology

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

Larry Ruzzo

Autumn 2021

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

University of Washington
Computer Science & Engineering

CSE 427, Au '21: Computational Biology

Administrative
FAQ
Schedule & Reading
Course Email/BBoard
Subscription Options
Class List Archive
E-mail Course Staff
Lecture Slides
Lecture Recordings
Here
Resources
Pubmed
BLAST
PDB
NHGRI Talking Glossary

Lecture: CSE2 G10, TuTh 11:30-12:50

		Office Hours Lo
Instructor:	Larry Ruzzo, ruzzo@cs	TBA
TA :	Cailin Winston, cailinw@cs	TBA
TA :	Zoey Shi, shiz27@cs	TR

Course Email: cse427a_au21@uw.e ...icements and general interest student/staff Q&A about homework, lectures, etc. The ins... are subscribed to this list. Enrolled students are as well, but probably should change t... ...scription options. Messages are automatically archived.

For fastest response, que... ...neral interest should be directed to the instructor and TAs collectively via the "... ...K at left. Individual email addresses (above) may also be used, if needed.

Discussion P... ...l free to use the course-specific Ed Discussion Board to discuss lectures, homewo... ...g else (loosely) course-related.

Catal... ...n: Algorithmic and analytic techniques underlying analysis of large-scale biological ... ...s DNA, RNA, and protein sequences or structures, expression and proteomic profiling. ...perience with databases, analysis tools, and genome markers. Applications such as sequence ...t, BLAST, phylogenetics, and Markov models.

...quisites: CSE 312; CSE 332

...redits: 3

Learning Objectives: The availability of the complete genome sequences of humans and other organisms is one of the landmark achievements of science. Understanding this enormous volume of data is a problem that will challenge scientists for decades to come, and the nature and scope of the problem means that computer scientists will play a vital role. The primary objective of the course is for students to understand the variety of computational problems and solutions that arise in this interdisciplinary field. Students will

Course Mechanics & Grading

Hybrid in-class/livestream format

Web:

http://courses.cs.washington.edu/courses/cse427/21au

Reading

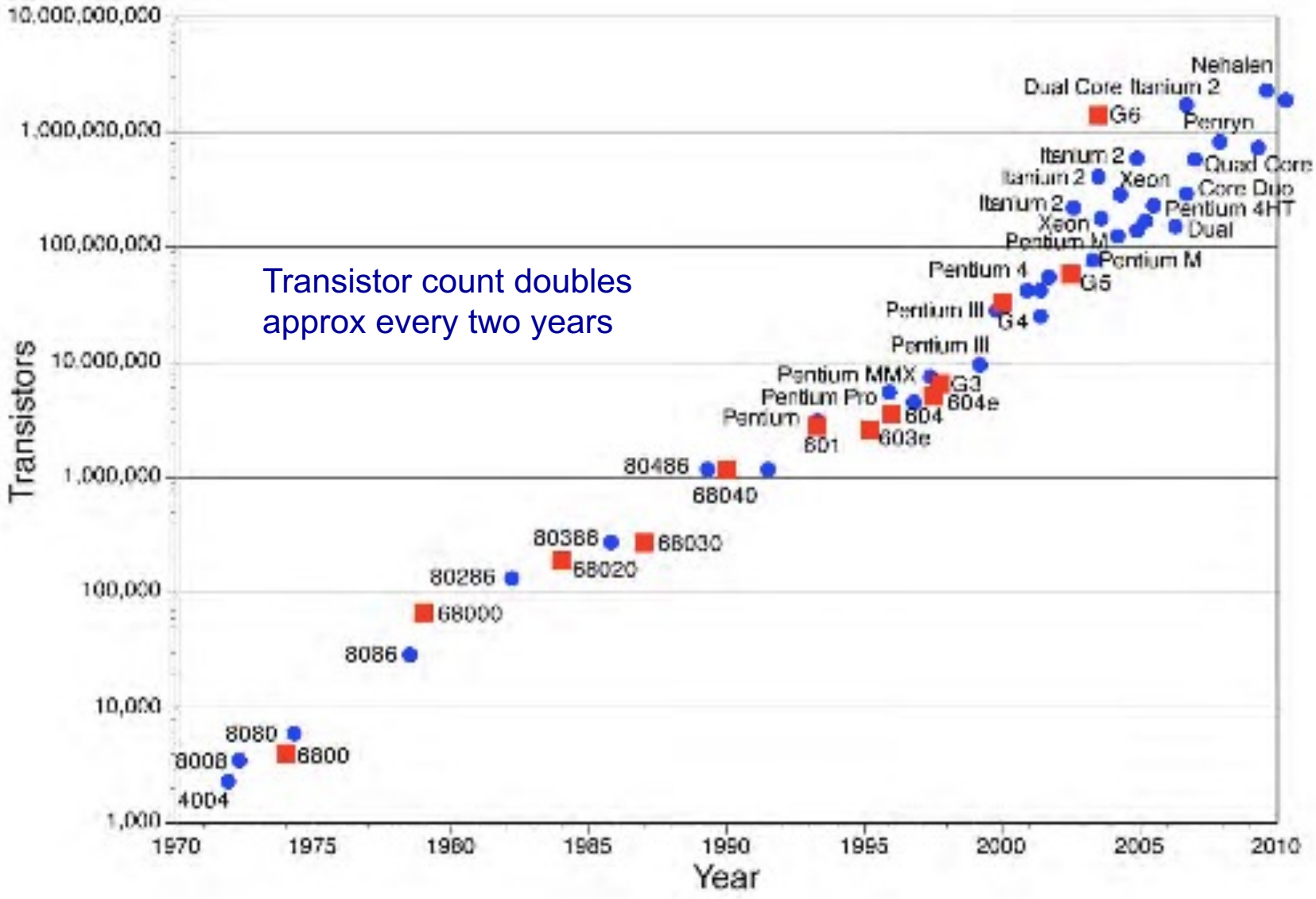
In class discussion

Homeworks: paper exercises & programming

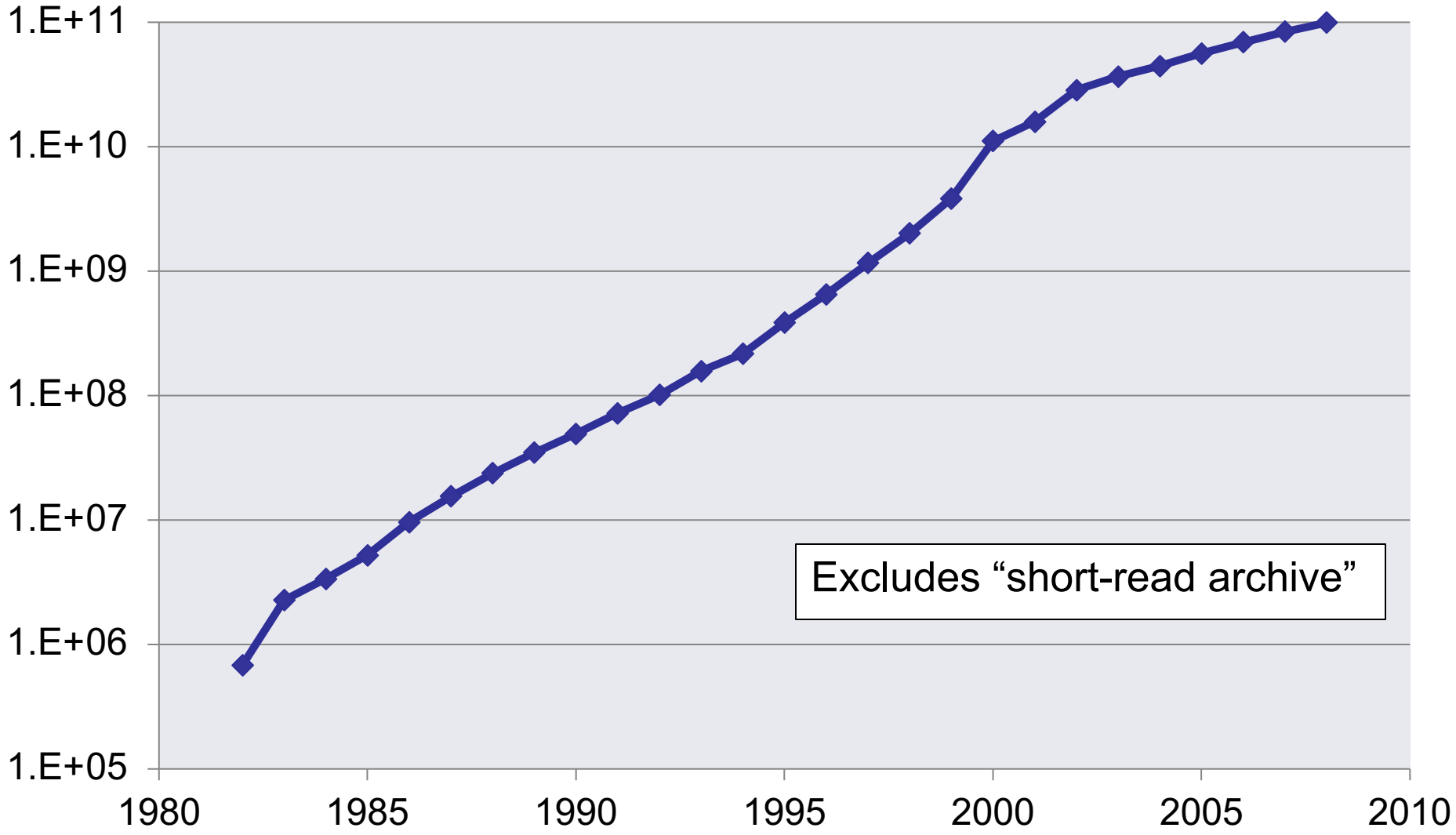
No exams: maybe oversized last homework in lieu of final

Background & Motivation

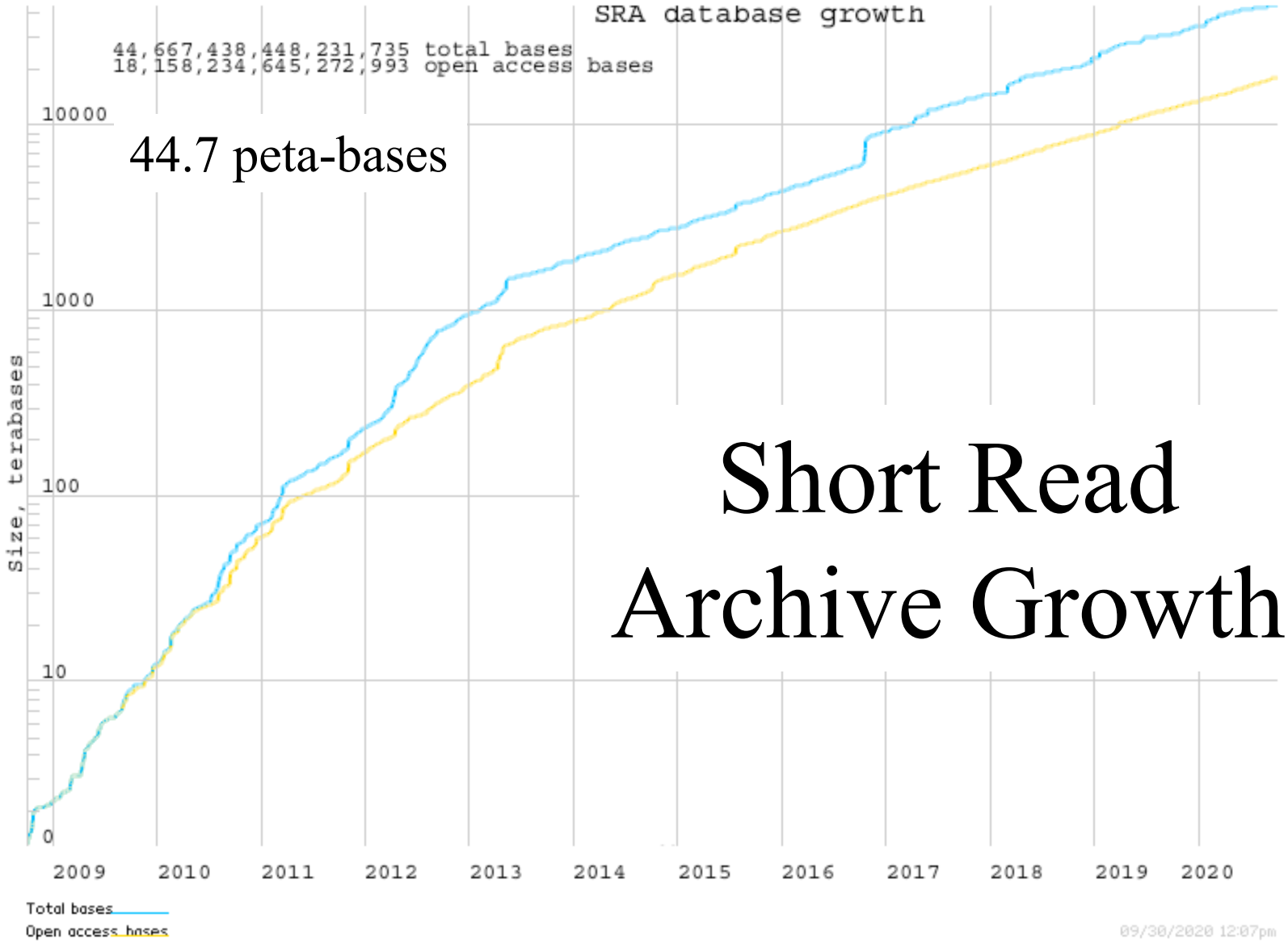
Moore's Law


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

Growth of GenBank (Base Pairs)

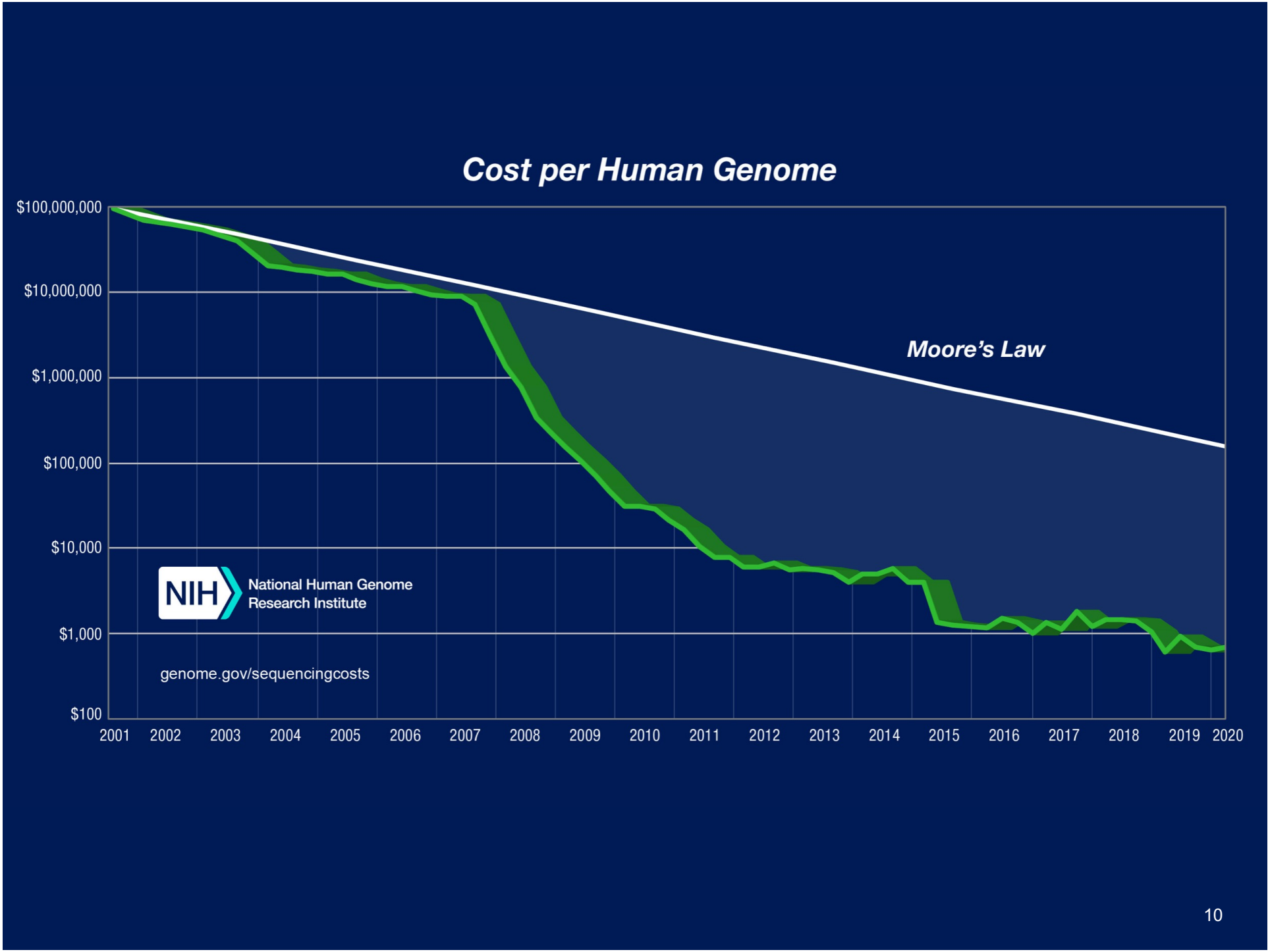


Source: http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html



Short Read Archive Growth

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


Cost per Human Genome
$100,000,000
$10,000,000
$1,000,000
$100,000
$10,000
$1,000
$100
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts

Modern DNA Sequencing

A box the size of a double oven

(but costs a bit more … ;-)

can generate ~3 x 10^{12} BP of DNA seq/day; i.e.,

1^{st} 30 yrs of genbank

1000 x your genome

Big Data: Astronomical or Genomical?

Stephens, et al. (2015). PLoS Biol 13(7): e1002195. doi:10.1371/journal.pbio.1002195

Table 1. Four domains of Big Data in 2025.

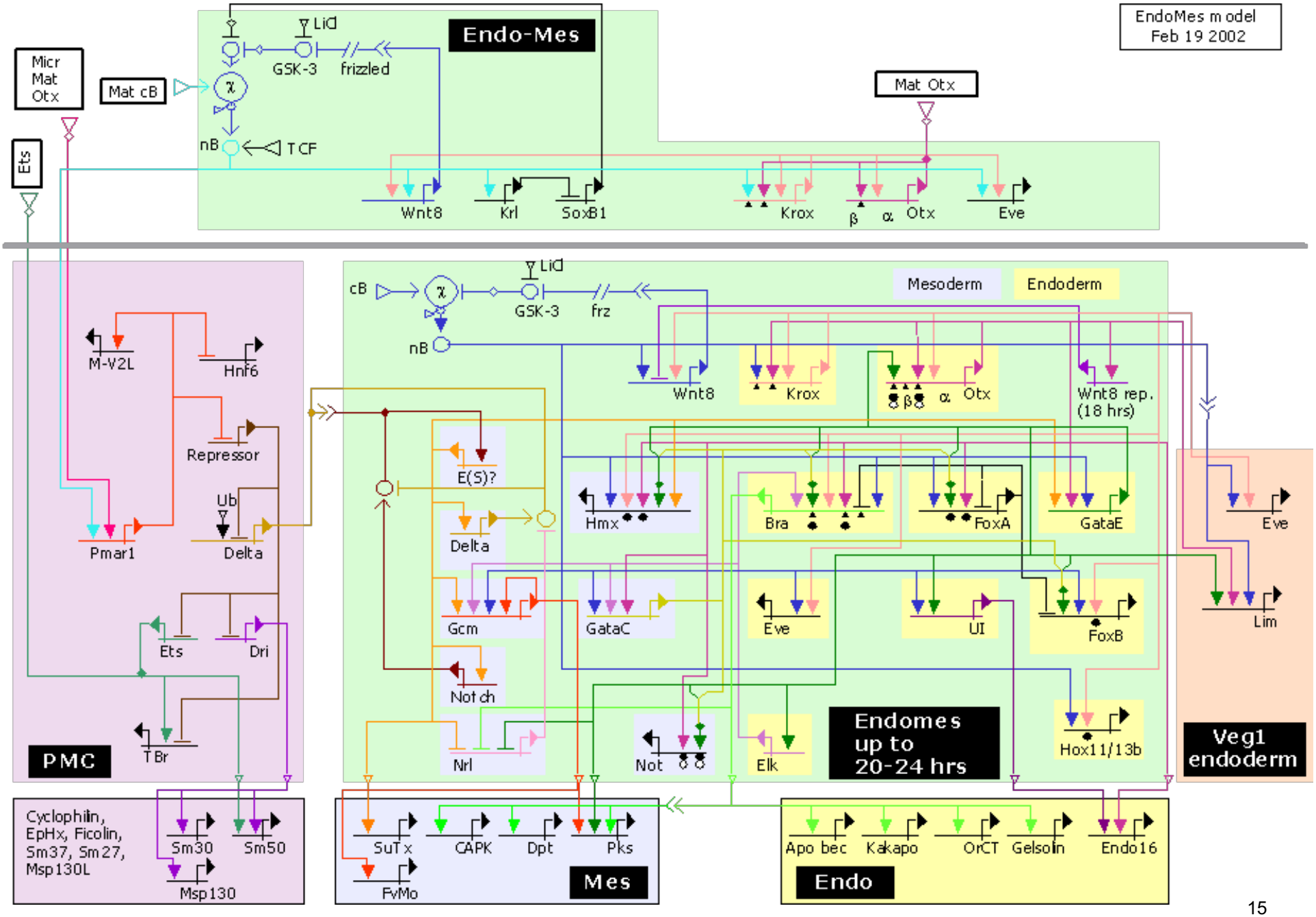
In each of the four domains, the projected annual storage and computing needs are presented across the data lifecycle.

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

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

Goals

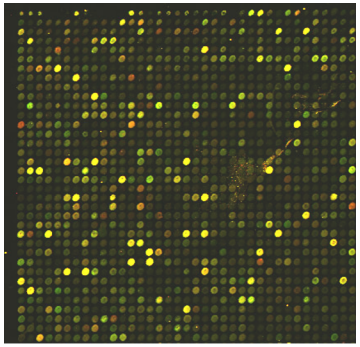
Basic biology

Drug discovery, validation & development

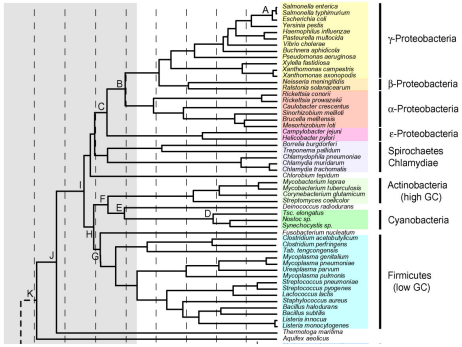
Disease diagnosis/prognosis/treatment

Individualized/precision medicine

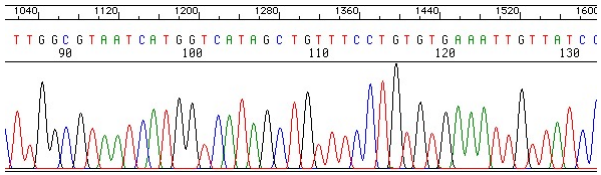
…

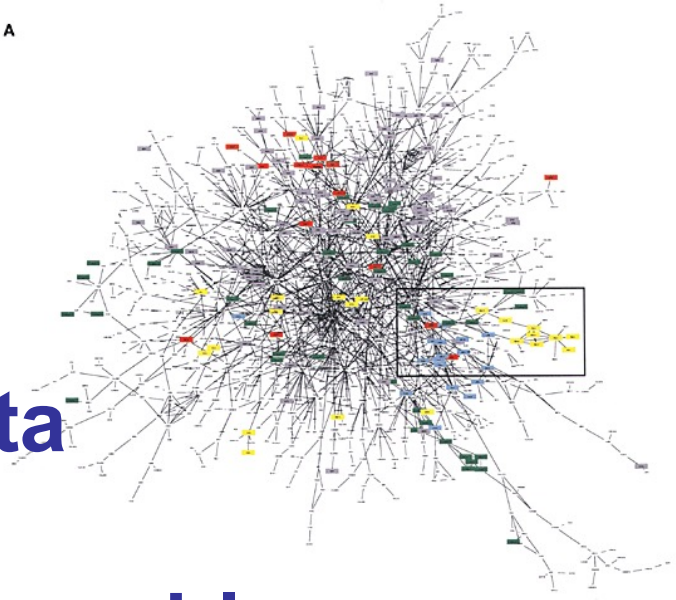


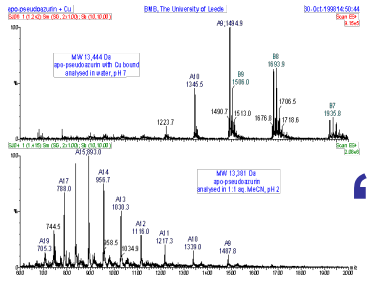
“High-Throughput BioTech”



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







***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, development, immune response, …
- 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
- ...

Computers in biology: Then & now

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

ACGGGTAA

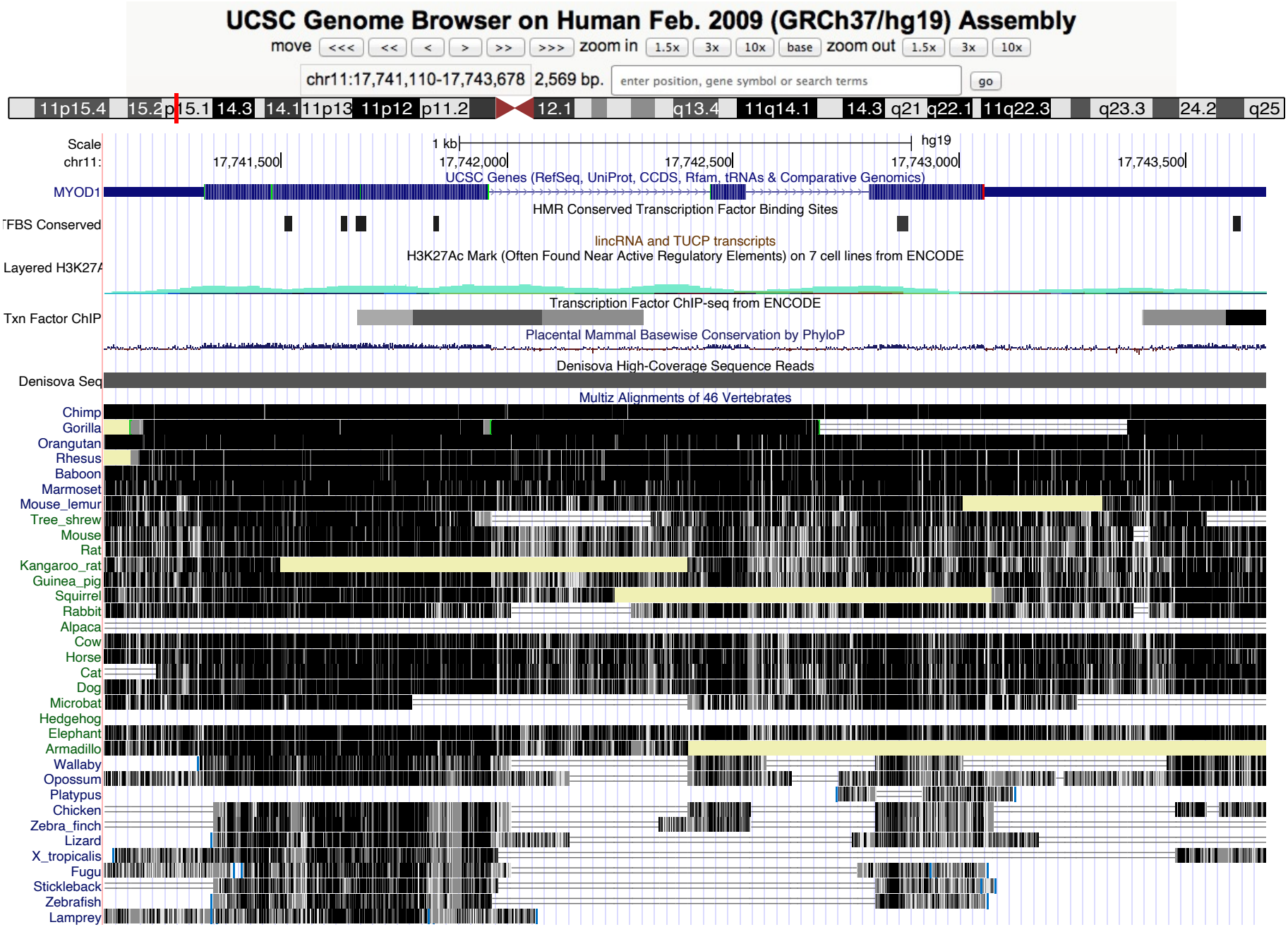
AC GGTAA

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 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
1 kb
hg19
chr11:
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

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

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

A *VERY* Quick Intro To Molecular Biology

The Genome

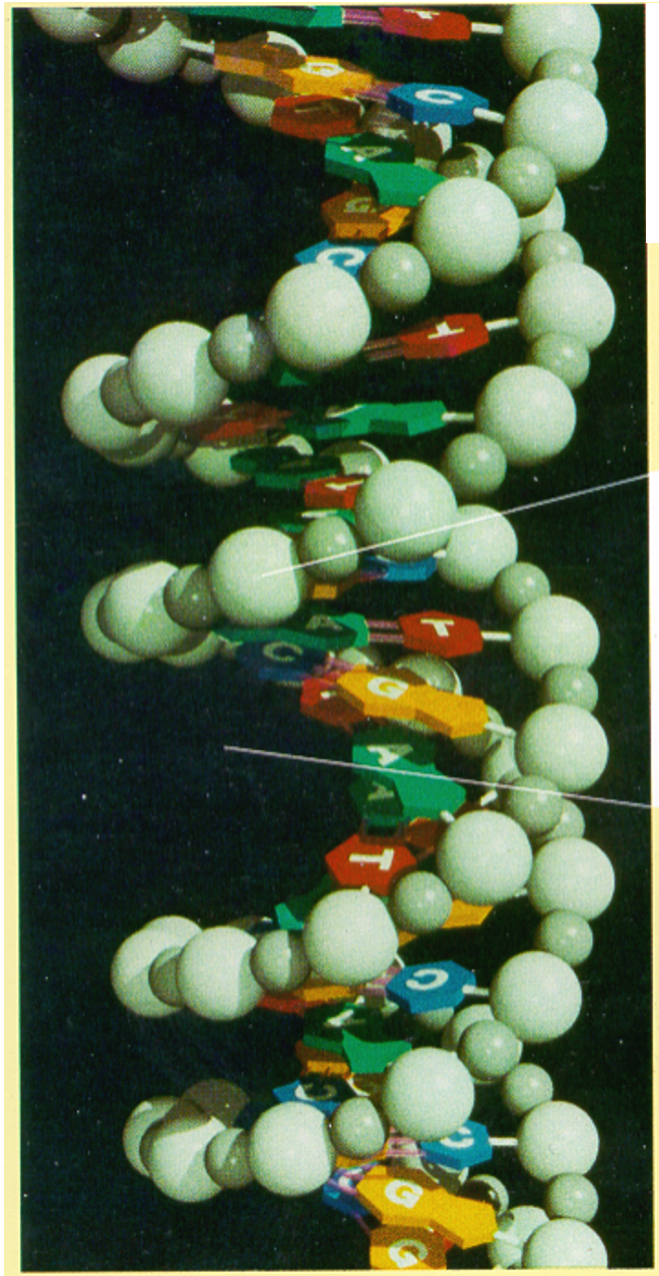
The hereditary info present in every cell

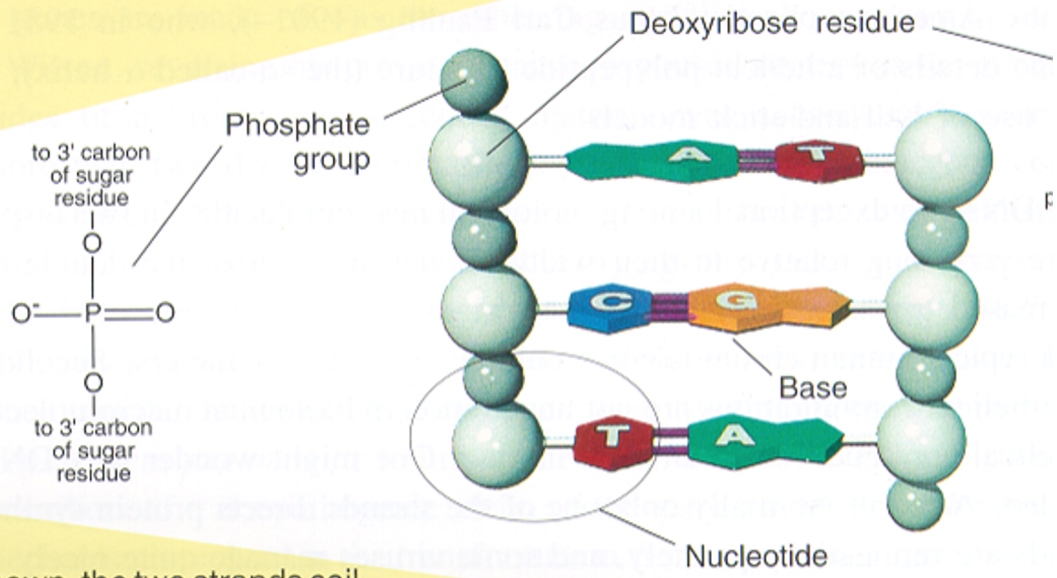
DNA molecule -- a long sequence of *nucleotides* (A, C, T, G)

Human genome -- about 3×10^9 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 $\leftrightarrow$ T C $\leftrightarrow$ 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 has 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, like photosynthesis

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 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 specialized 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”

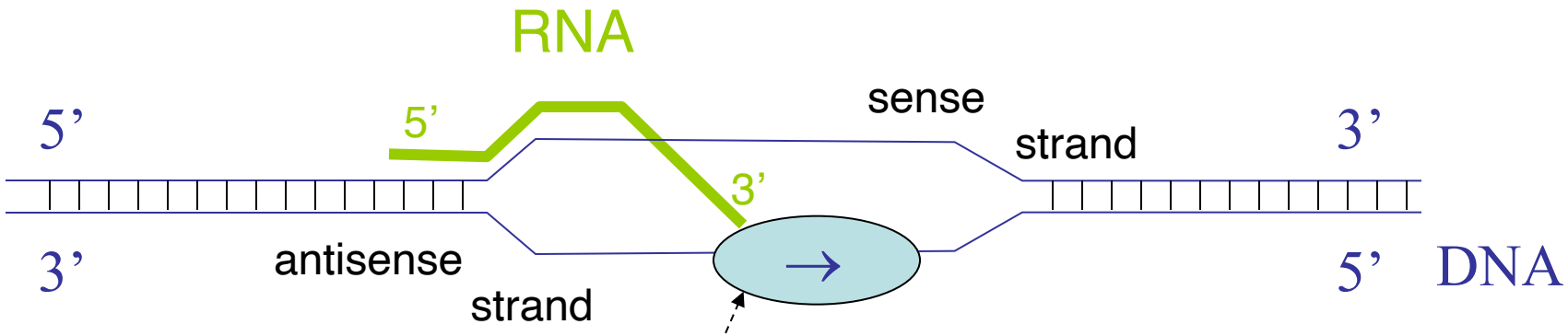
Genes encode proteins

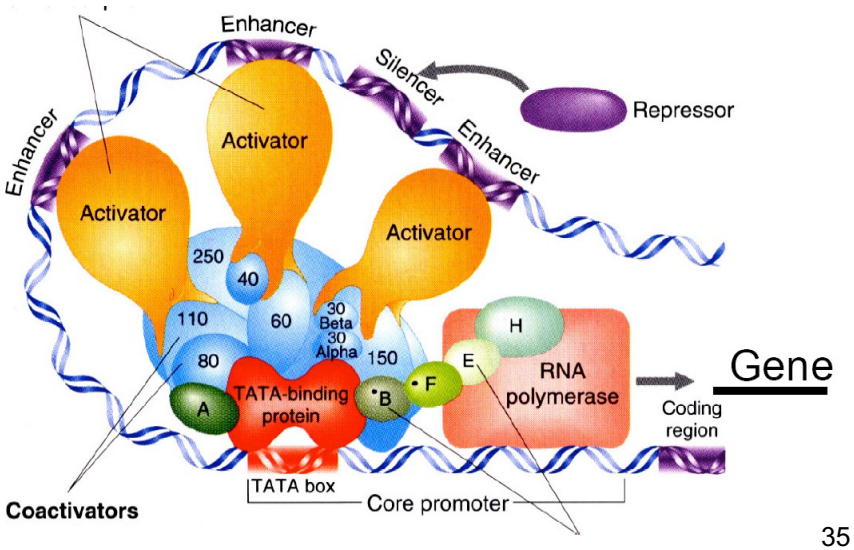
DNA transcribed into messenger RNA

mRNA translated into proteins

Triplet code (codons)

Transcription: DNA → RNA



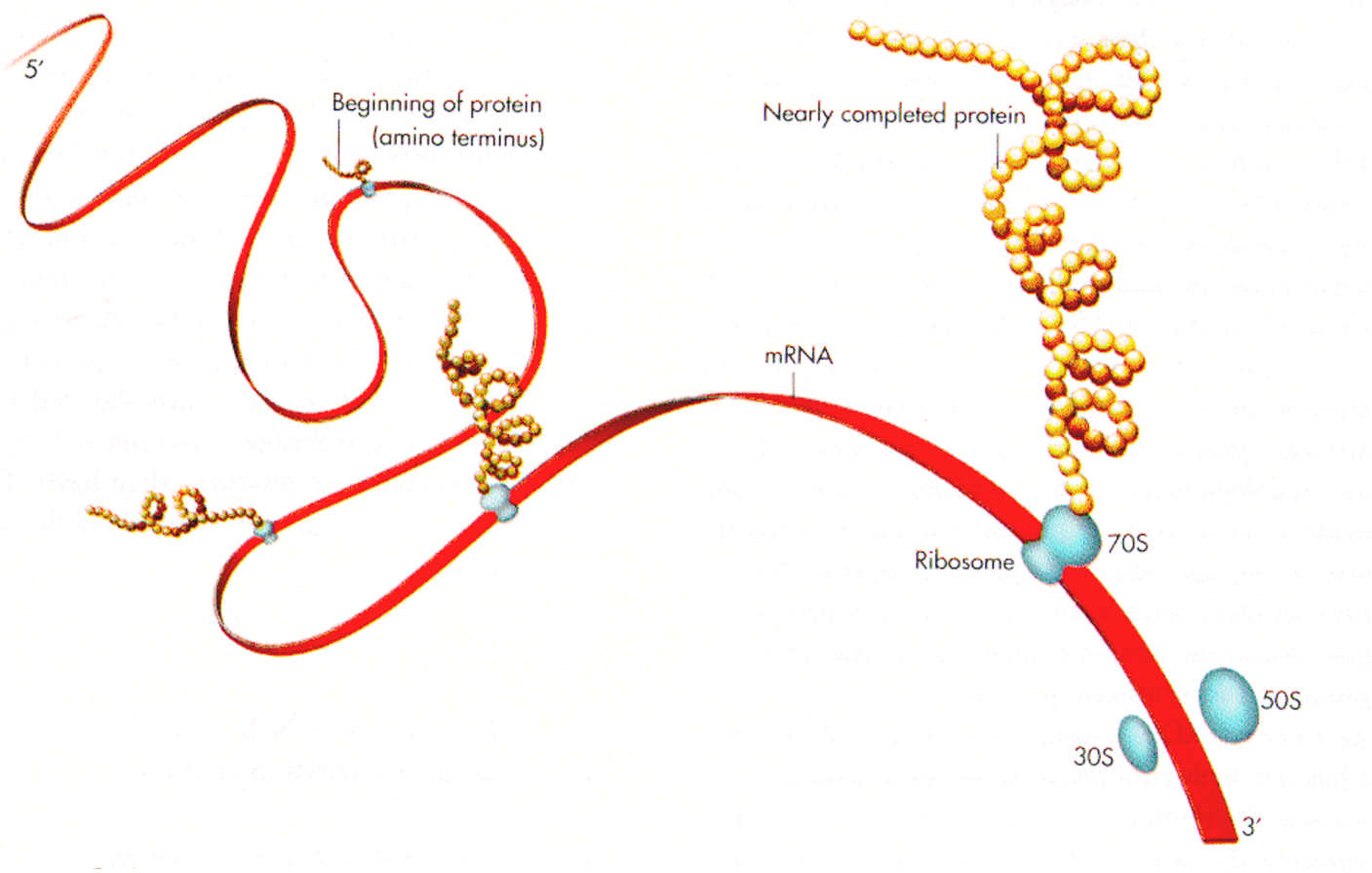


Codons & The Genetic Code

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

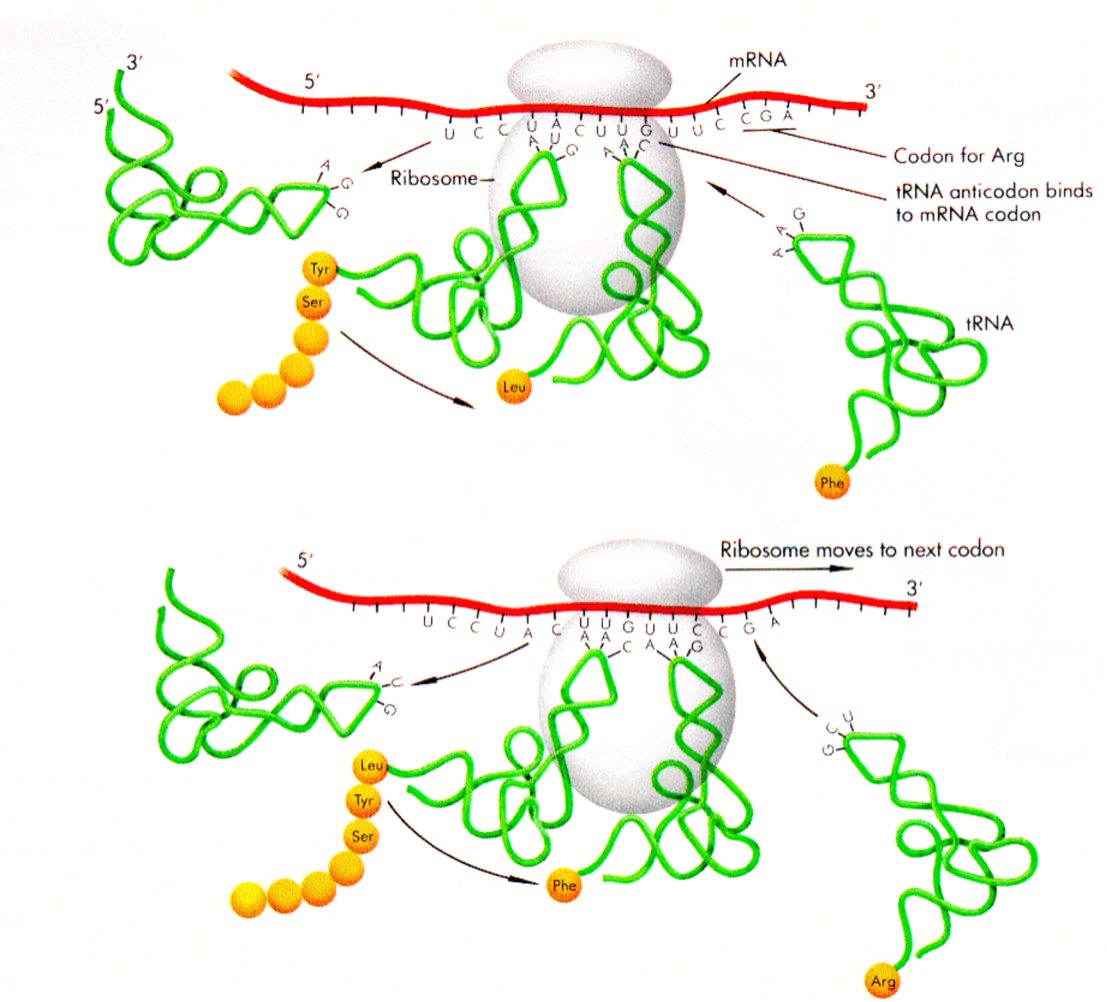
Ala : Alanine
Arg : Arginine
Asn : Asparagine
Asp : Aspartic acid
Cys : Cysteine
Gln : Glutamine
Glu : Glutamic acid
Gly : Glycine
His : Histidine
Ile : Isoleucine
Leu : Leucine
Lys : Lysine
Met : Methionine
Phe : Phenylalanine
Pro : Proline
Ser : Serine
Thr : Threonine
Trp : Tryptophane
Tyr : Tyrosine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

Translation: mRNA → Protein



Watson, Gilman, Witkowski, & Zoller, 1992

Ribosomes



Watson, Gilman, Witkowski, & Zoller, 1992

Gene Structure

mRNA built 5’ to 3’

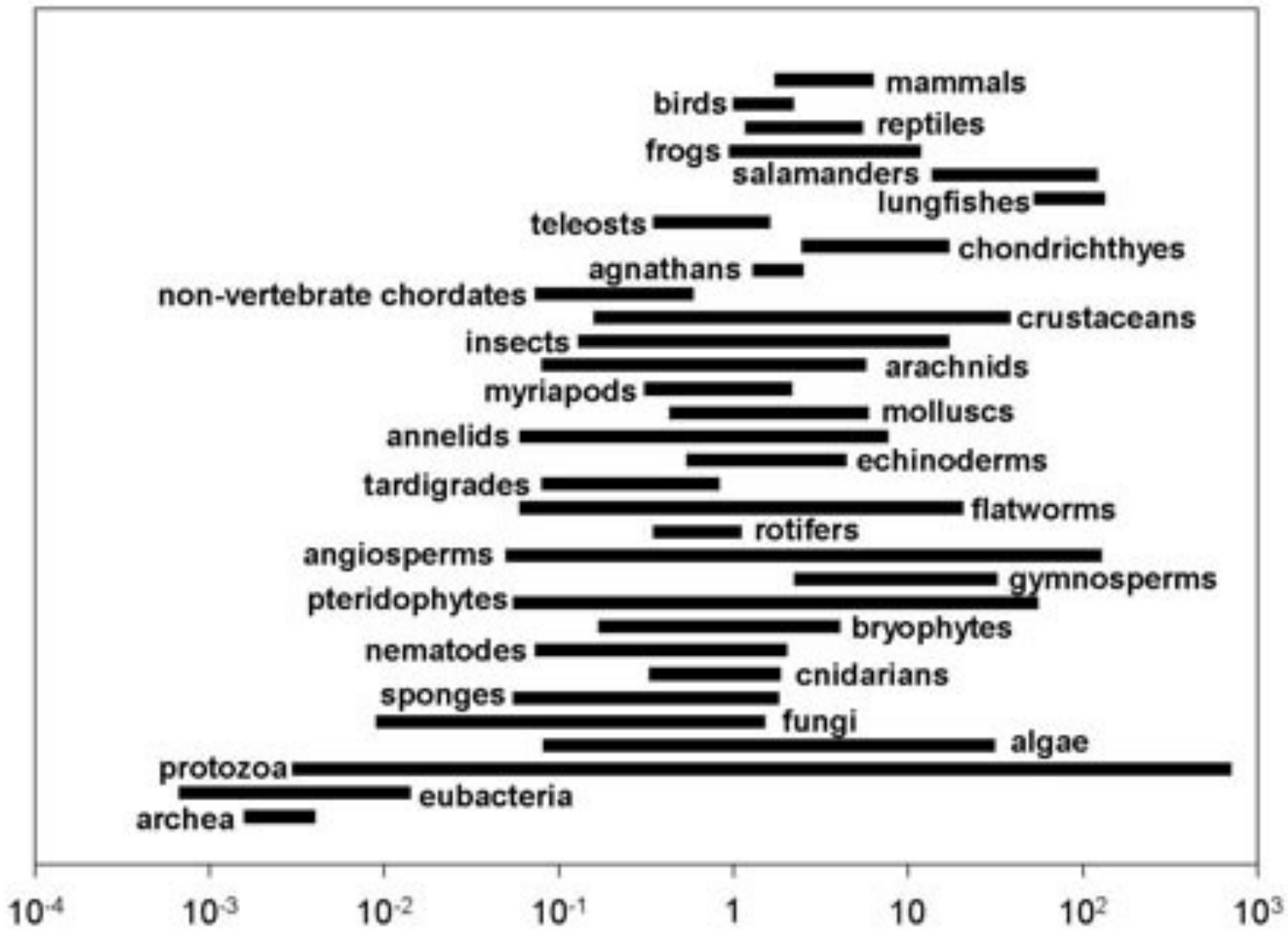
Promoter region and transcription factor binding sites (usually) precede 5’ end

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

	Bases	Genes
SARS-CoV-2	29,903	12
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	~21,000
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

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

Subset of DNA being transcribed is >> 2% coding, giving many *non-coding RNAs* -- more than protein-coding genes, by some estimates

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