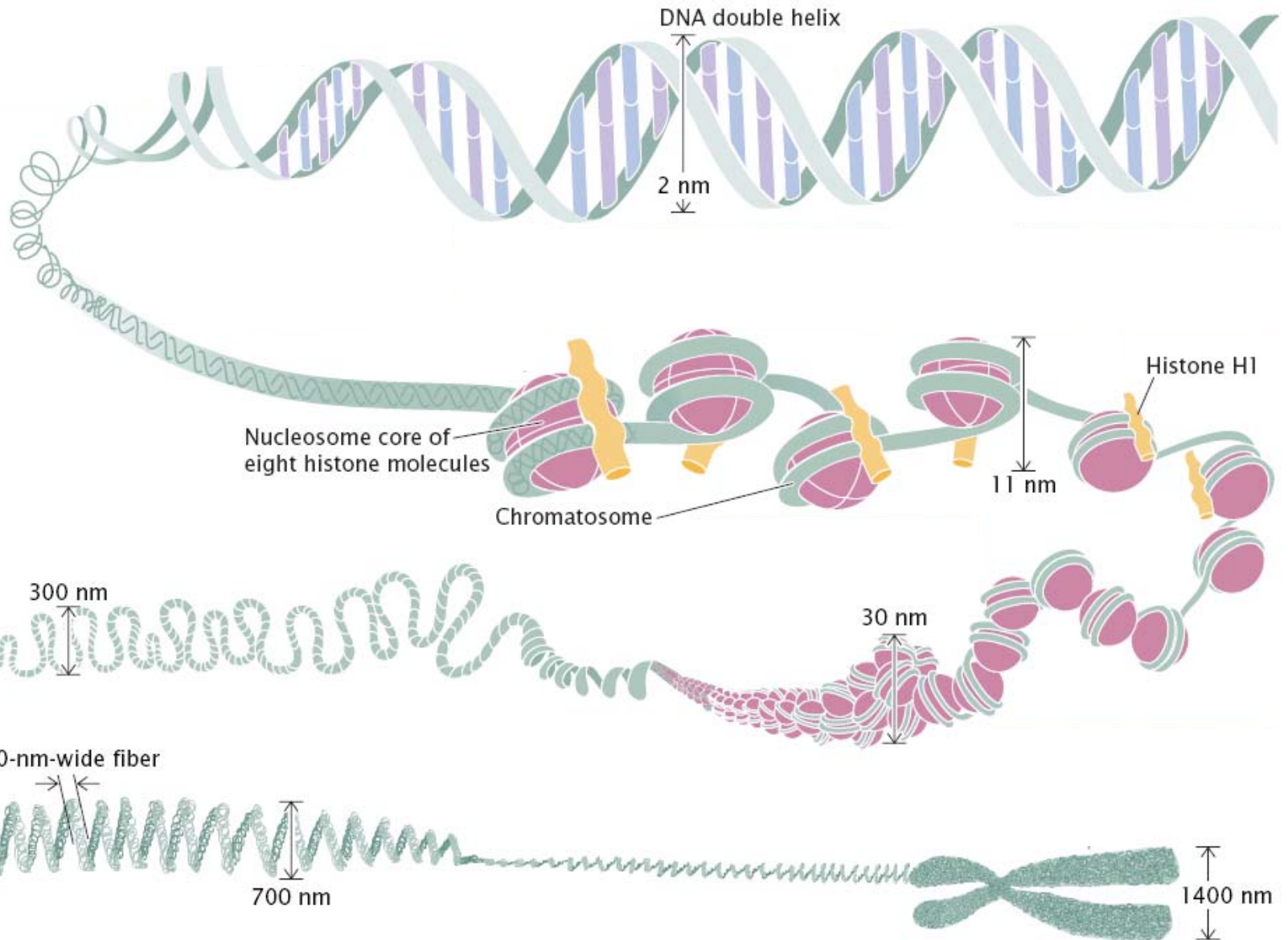


Computational Analysis of Genomes

Shinichi Morishita

The figures, photos and moving images with ♯marks attached belong to their copyright holders. Reusing or reproducing them is prohibited unless permission is obtained directly from such copyright holders.

Chromosomes, Chromatin Structure and Genomes



≠

Figure 1 : Chromatin has highly complex structure with several levels of organization.

Used with permission. © 2005 by W. H. Freeman and Company. All rights reserved.

Ref: Annunziato, A. DNA packaging: Nucleosomes and chromatin. *Nature Education* 1(1), (2008)

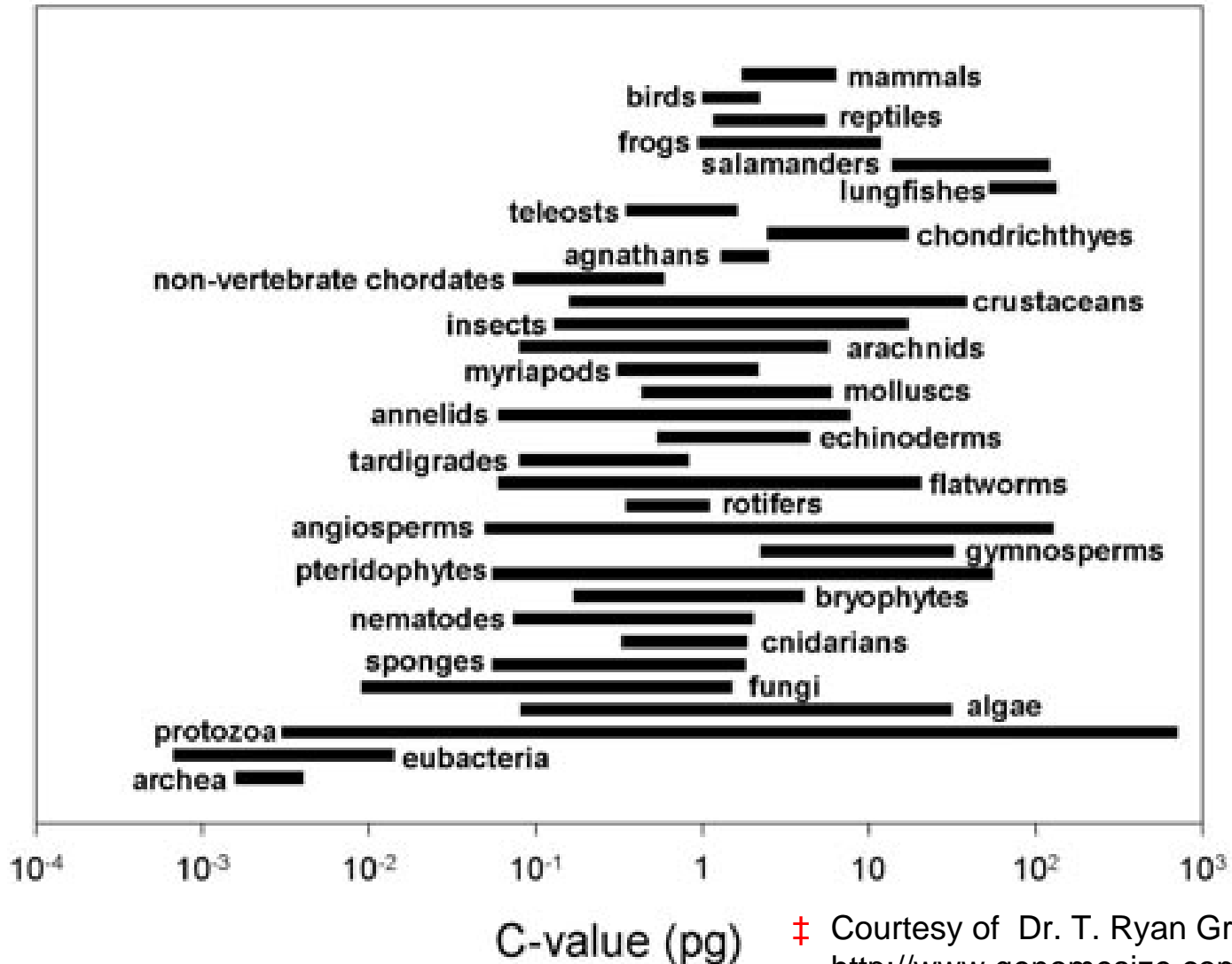
Sequencing a Genome

- Do the genome size and number of chromosomes characterize an organism?
- What could the sequenced genome be used for?
- How to sequence genomes?
- How to identify the gene coding regions?
- Recent revolutionary advances in genome sequencing equipment
- How to infer chromatin structure?

Genome Size

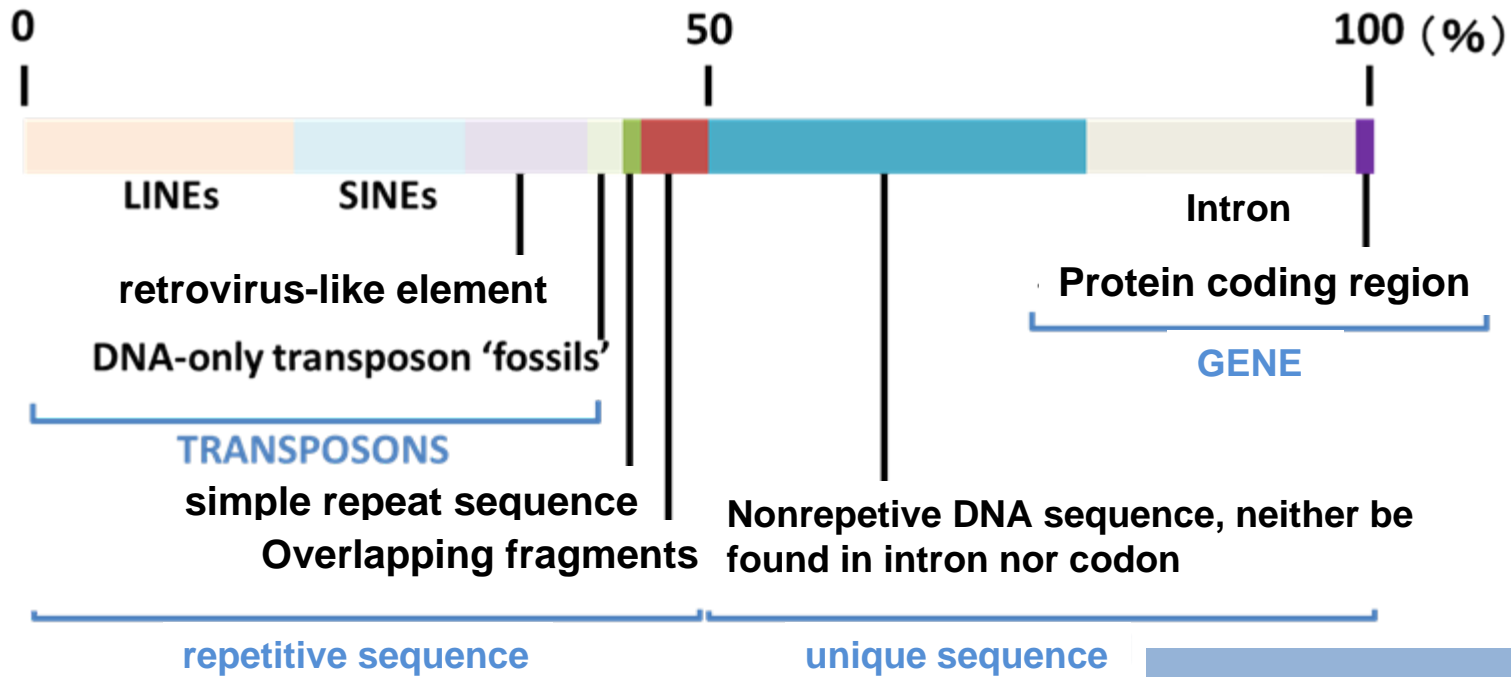
1 pg (10^{-12} g) \cong 1 billion base pairs

(more accurately, 978 million pairs)



† Courtesy of Dr. T. Ryan Gregory
<http://www.genomesize.com/statistics.php>

Why Are Genomes So Different in Size?

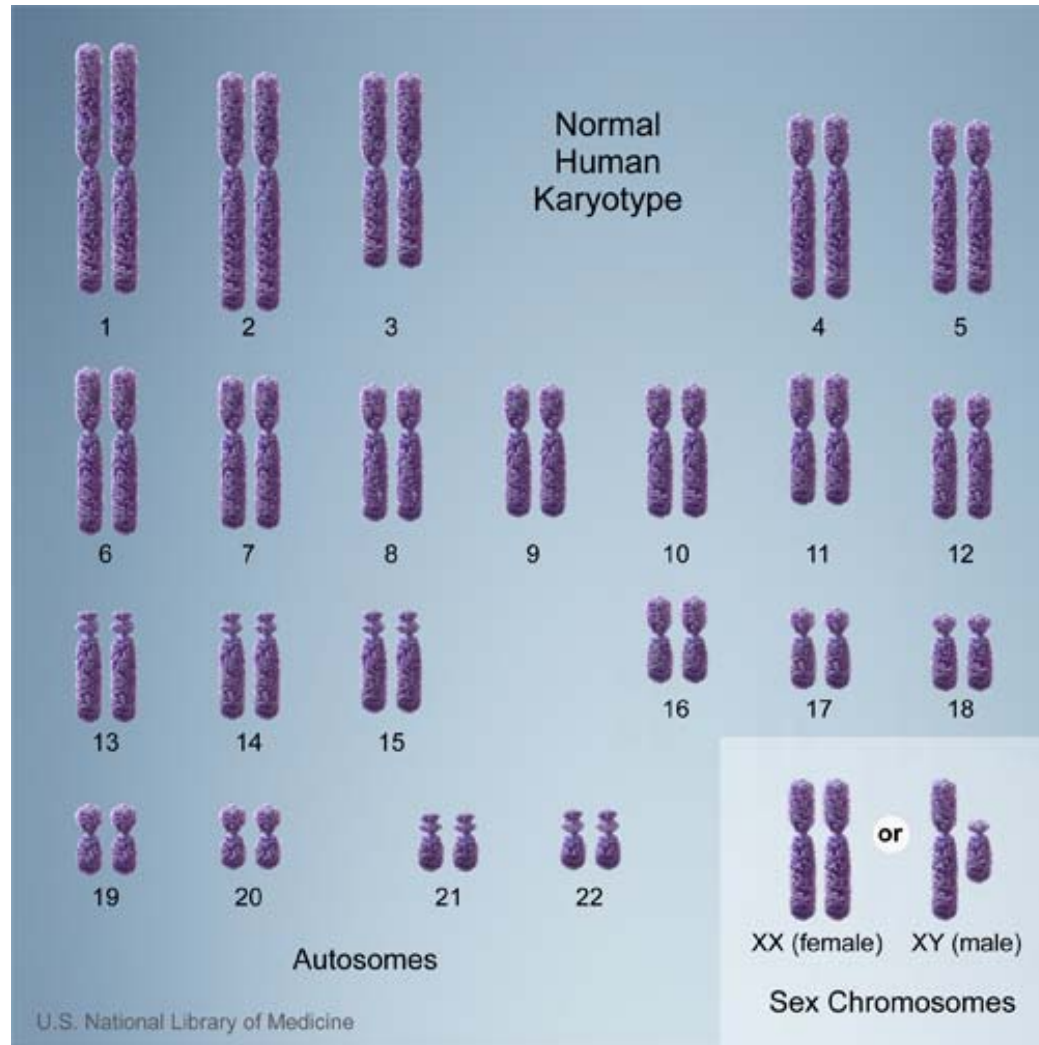


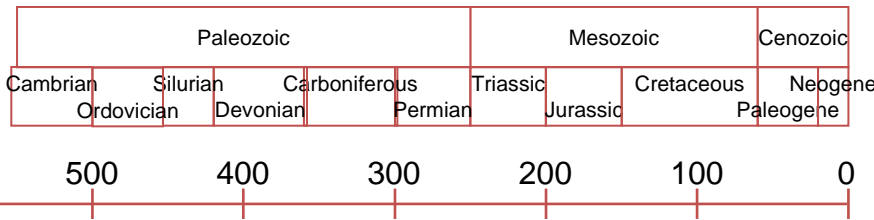
Human genome

Figure removed due to
copyright restrictions

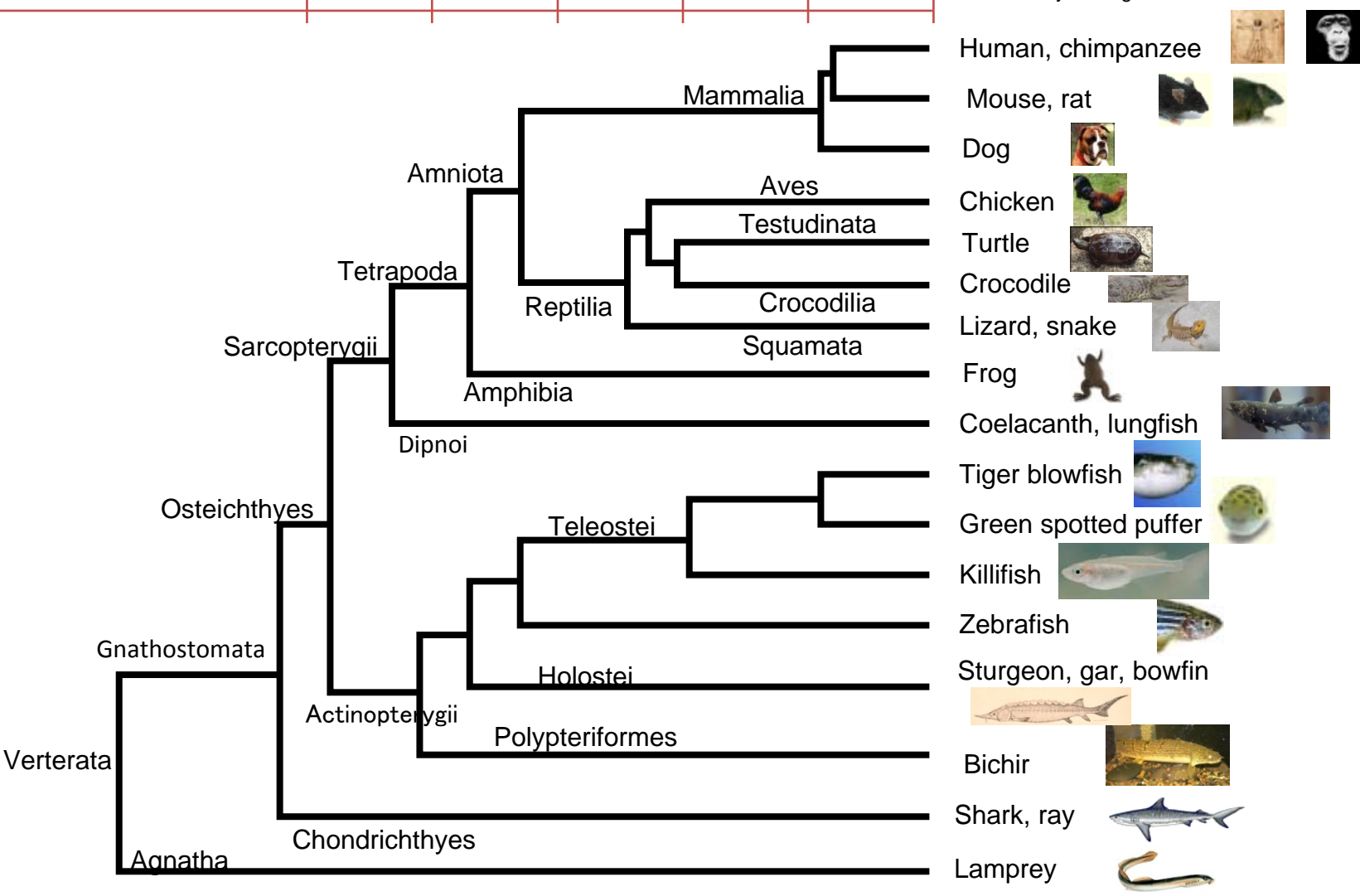
Molecular Biology of the Cell - Fifth Edition
Garland Science (2008)
Figure 5-75

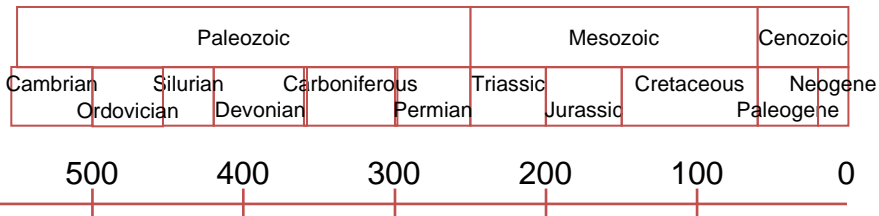
Human Chromosomes





Vertebrate Chromosome Counts





Chromosome Count Distribution
 Vertical axis: No. species,
 Horizontal axis: Chromosome count

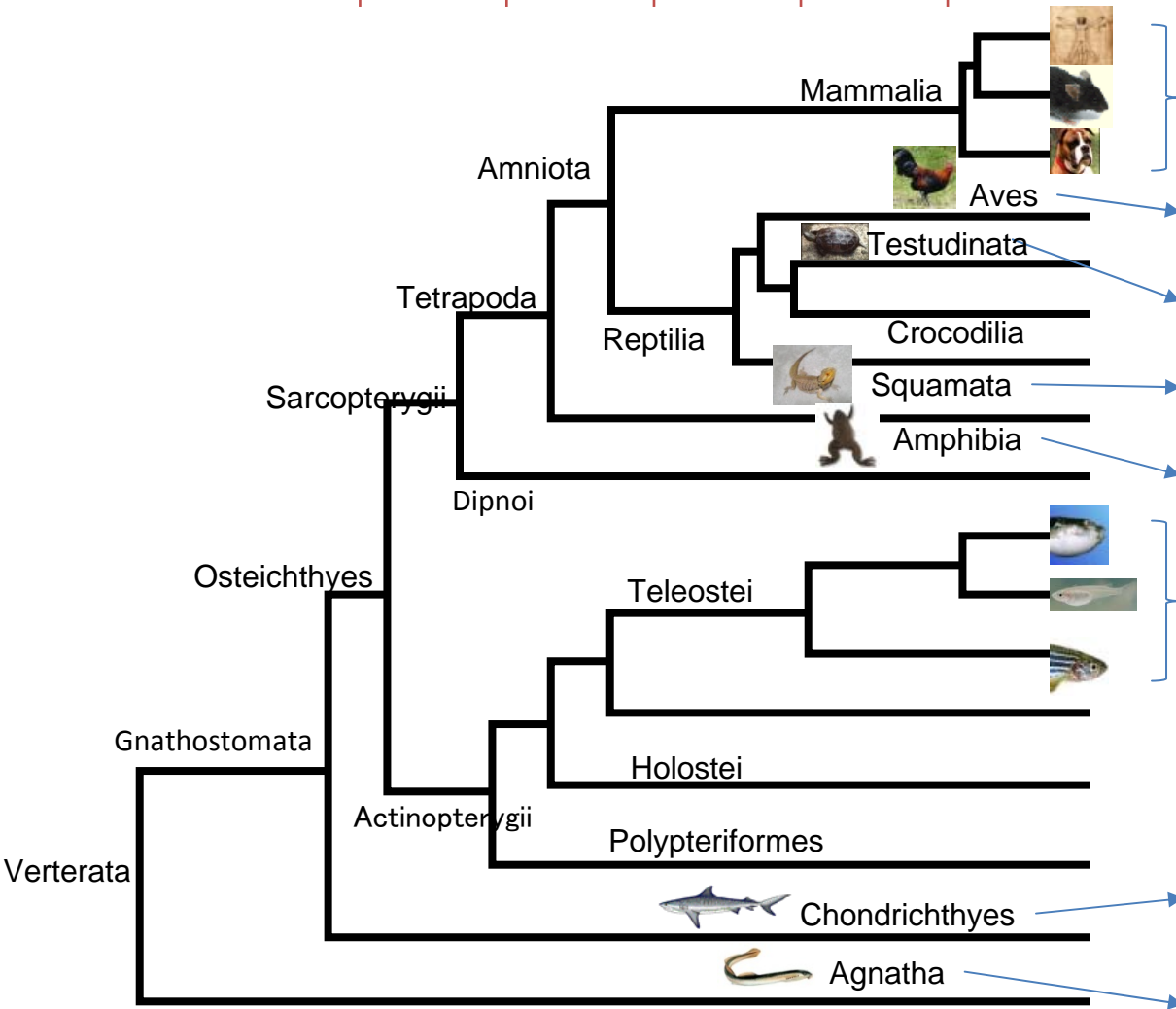


Figure removed due to copyright restrictions

Nakatani et al., 2007, Genome Res., 17, 1254-1265
 Figure6

Figure removed due to copyright restrictions

Molecular Biology of the Cell - Fifth Edition
Garland Science (2008)

Figure 4-14

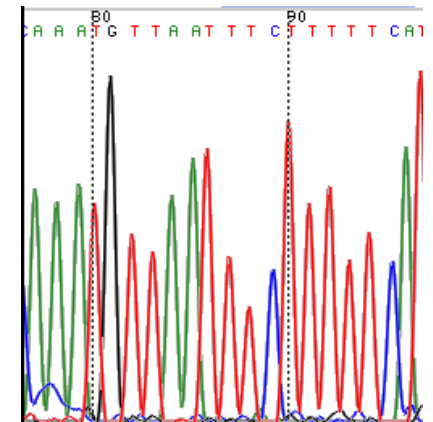
What Uses Has the Genome?

- It tells us whether a given gene is present.
- Chicken genome sequenced Dec 2004
- Do chickens have a poor sense of smell?
- There are calculated to be 218 genes that could be olfactory receptors.
- What happened to the genes for flight?

How to Sequence the Genome?



⌘ Applied Biosystems Japan Co., Ltd.



⌘ Applied Biosystems Japan Co., Ltd.

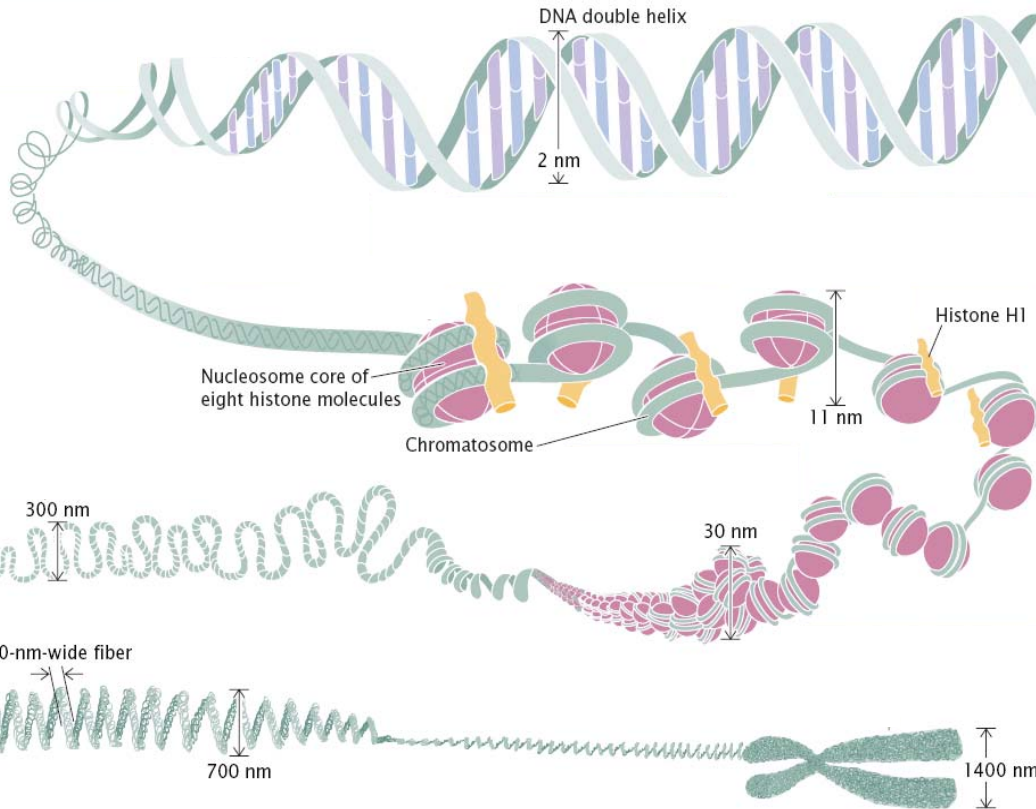
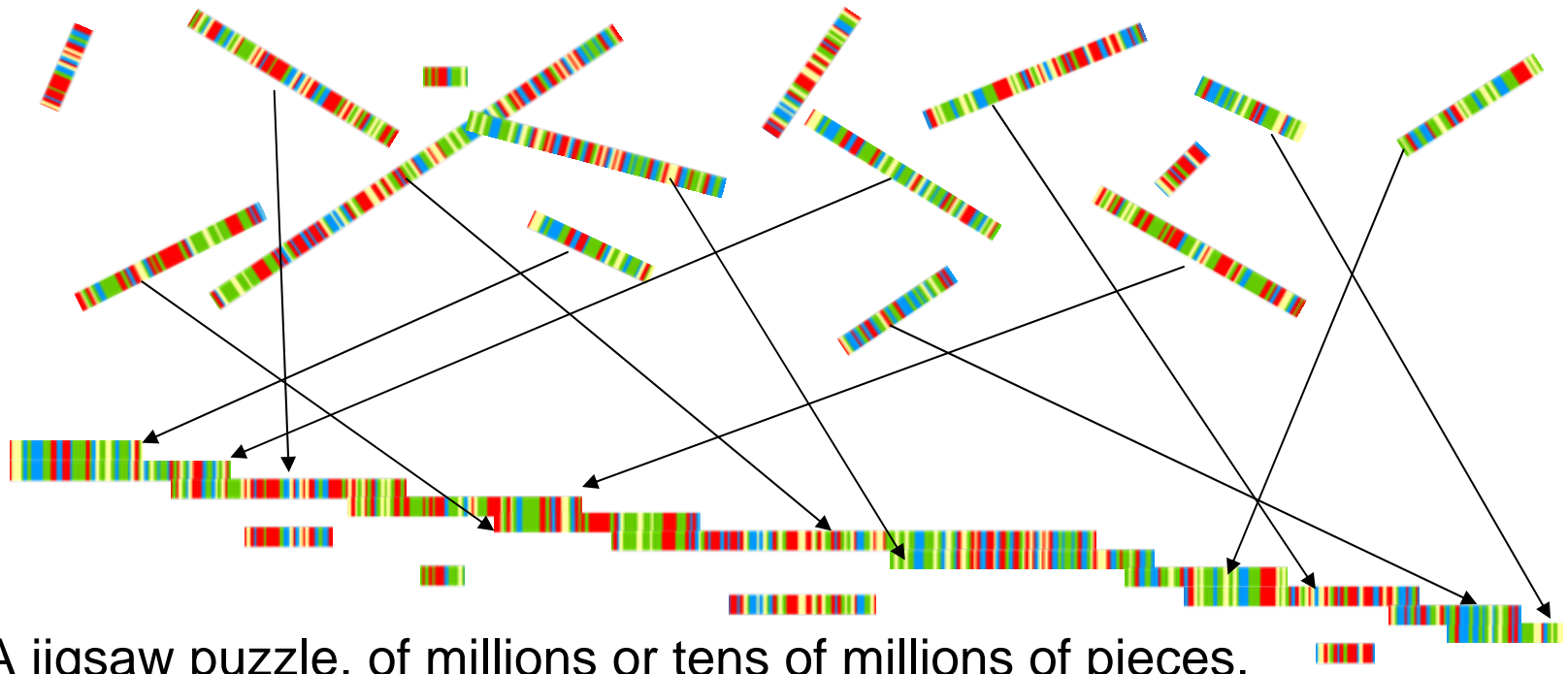
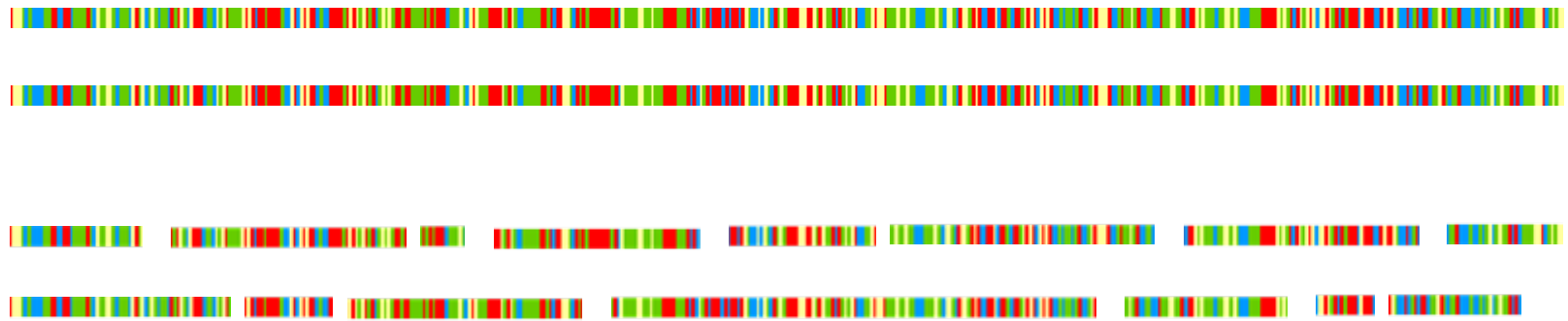
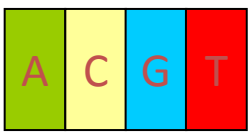


Figure 1 : Chromatin has highly complex structure with several levels of organization.
Used with permission. © 2005 by W. H. Freeman and Company. All rights reserved.

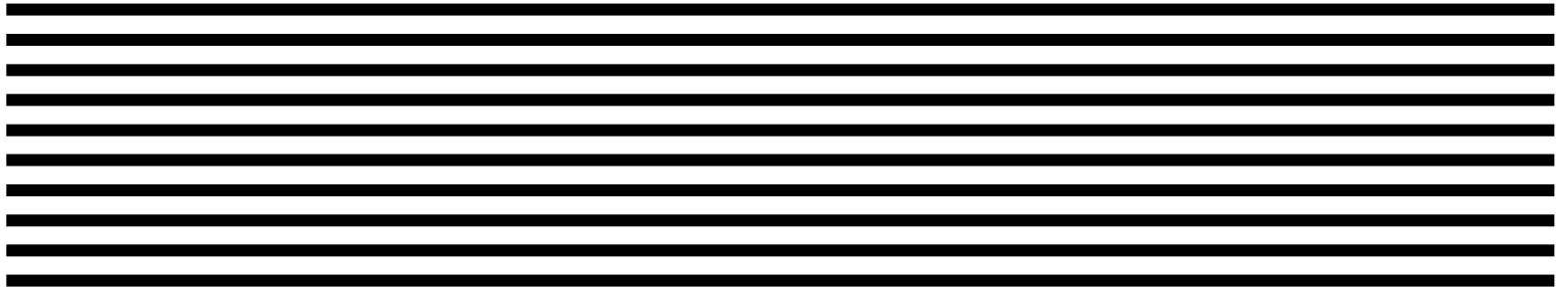
⌘ Ref: Annunziato, A. DNA packaging: *Nucleosomes and chromatin*. *Nature Education* 1(1), (2008)

The Sanger method allows reading of 500 to 800 bases . . .

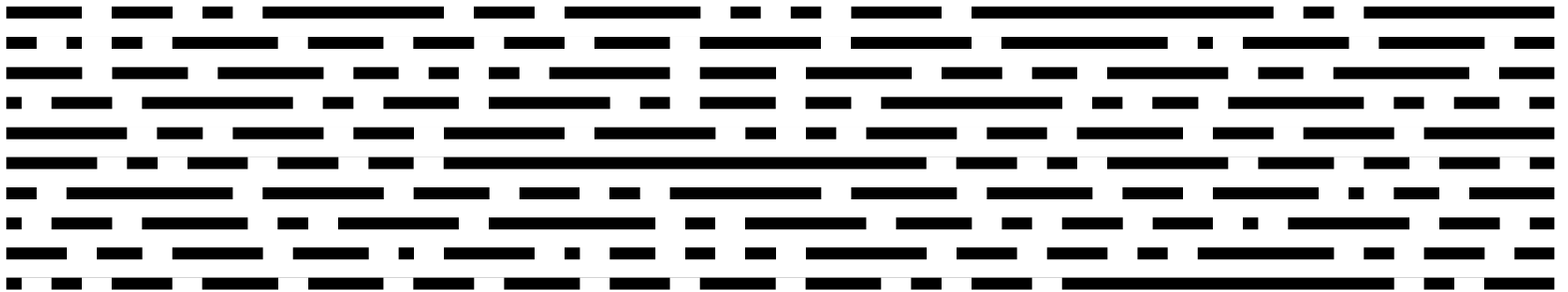


A jigsaw puzzle, of millions or tens of millions of pieces, whose source image is unknown.

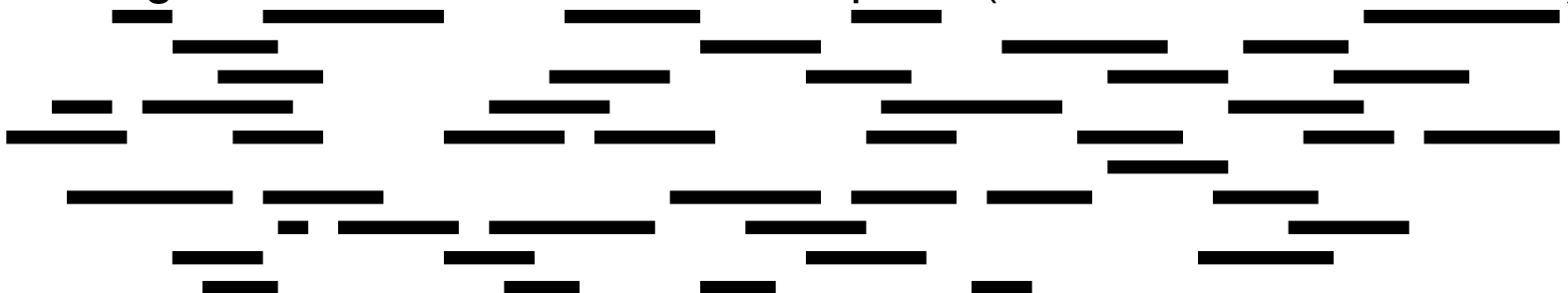
Copy the genome



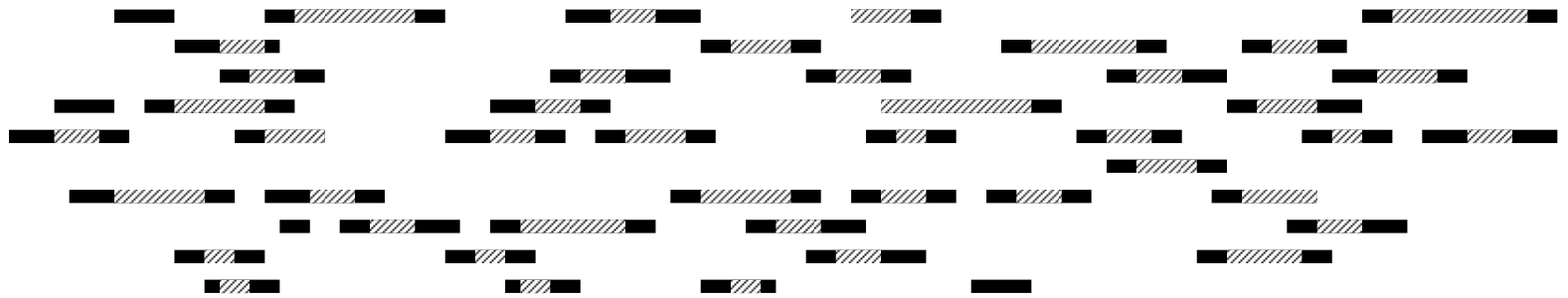
Use a high-velocity water jet to fragment the genome into random pieces



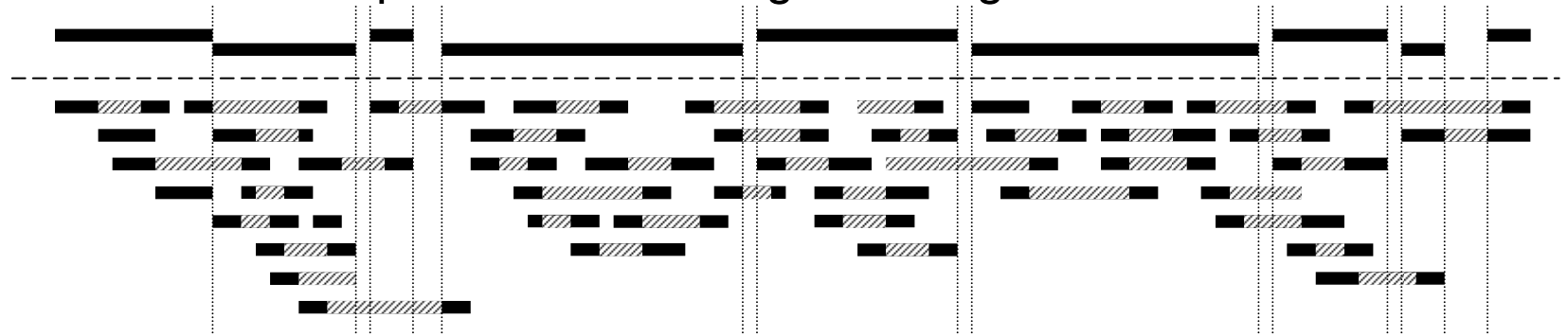
Collect fragments of around 2,000 base pairs (or other suitable volume)



Read 500 to 800 bases at either end of the group



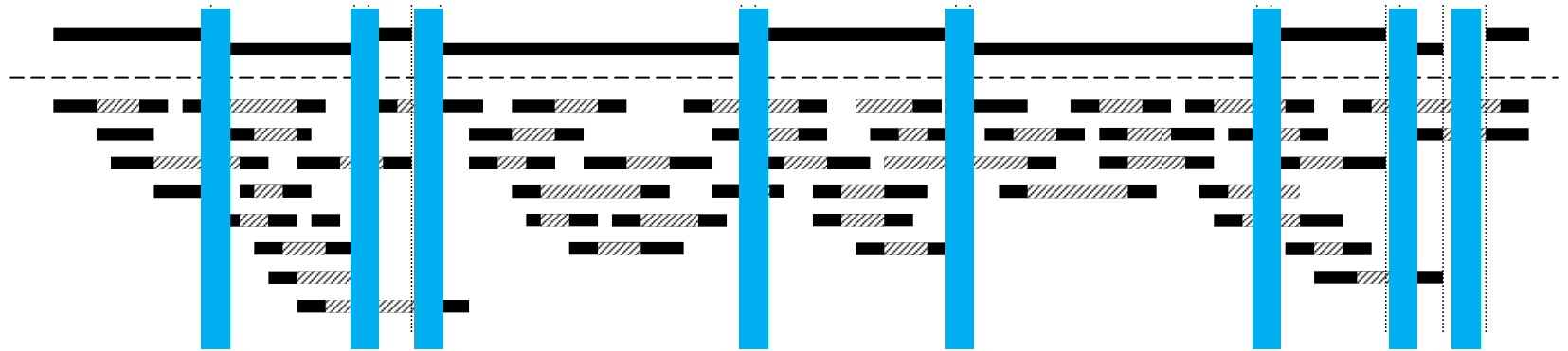
Assemble read sequences into contiguous fragments



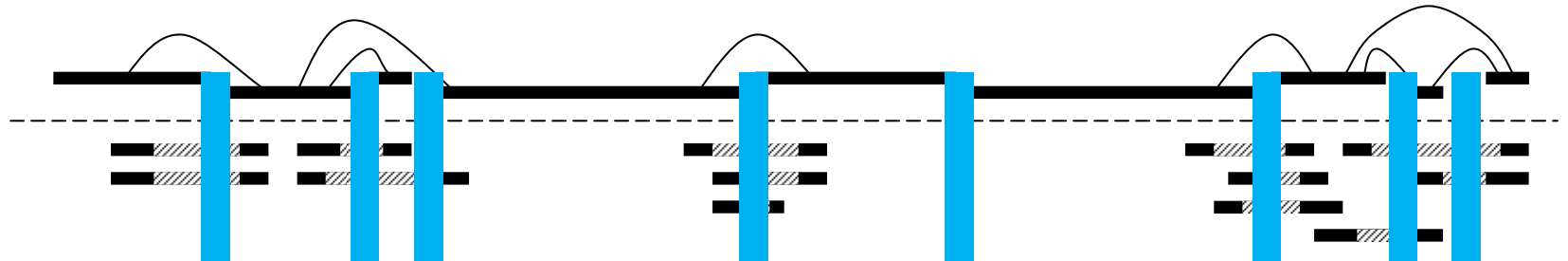
Example where it takes more than one method to assemble read sequences



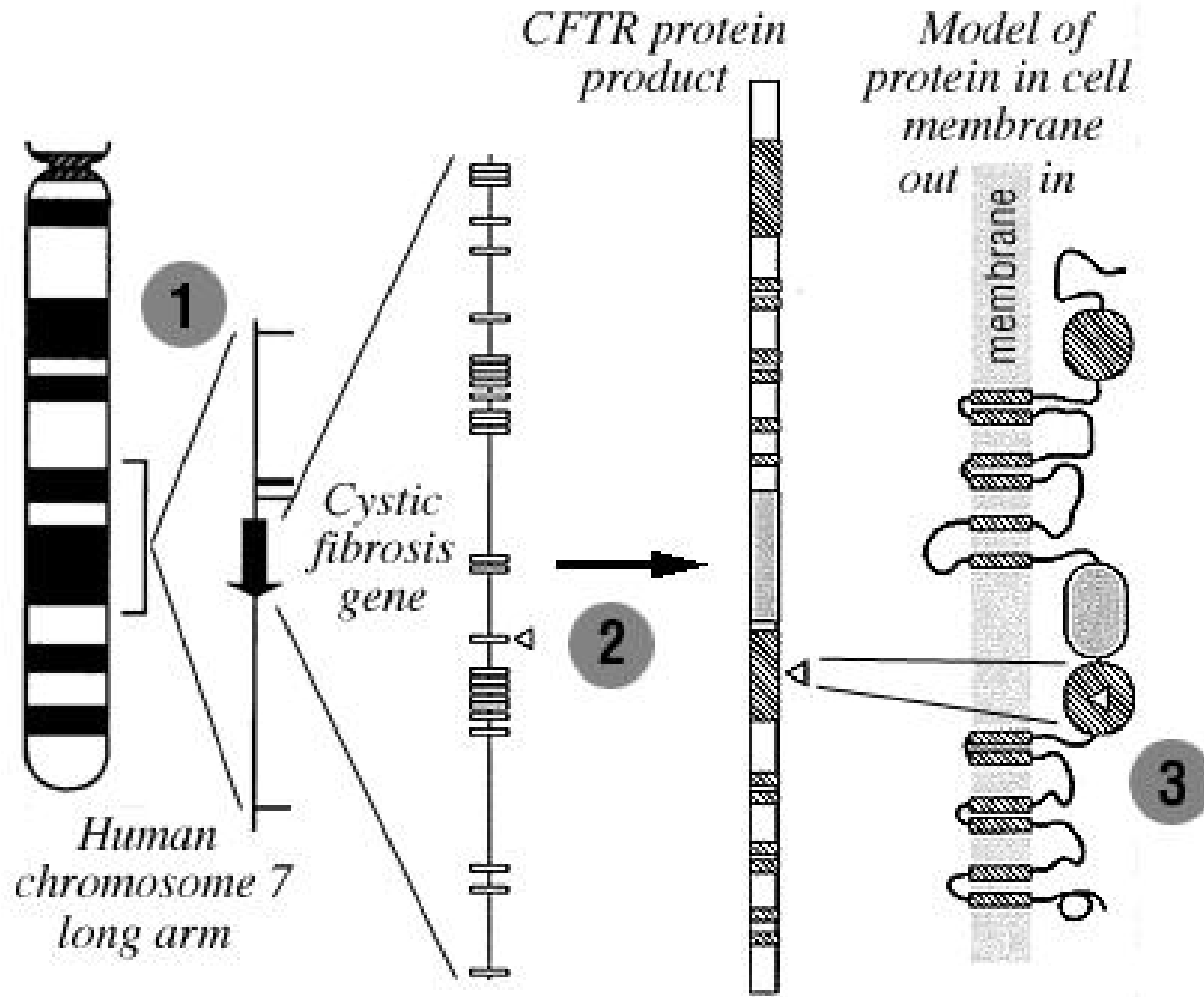
Assemble contiguous fragments



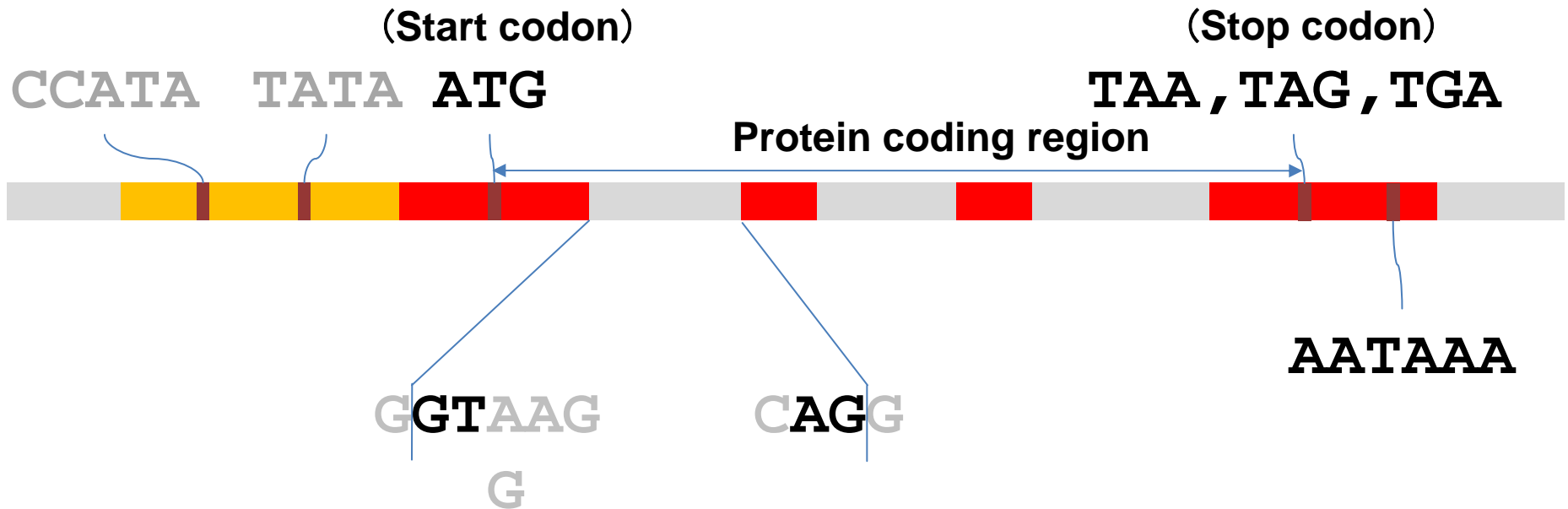
Assemble non-contiguous fragments



Gene Code Regions Within a Genome



Can Gene Coding Regions Be Predicted from a Genome Alone?



Coding potential

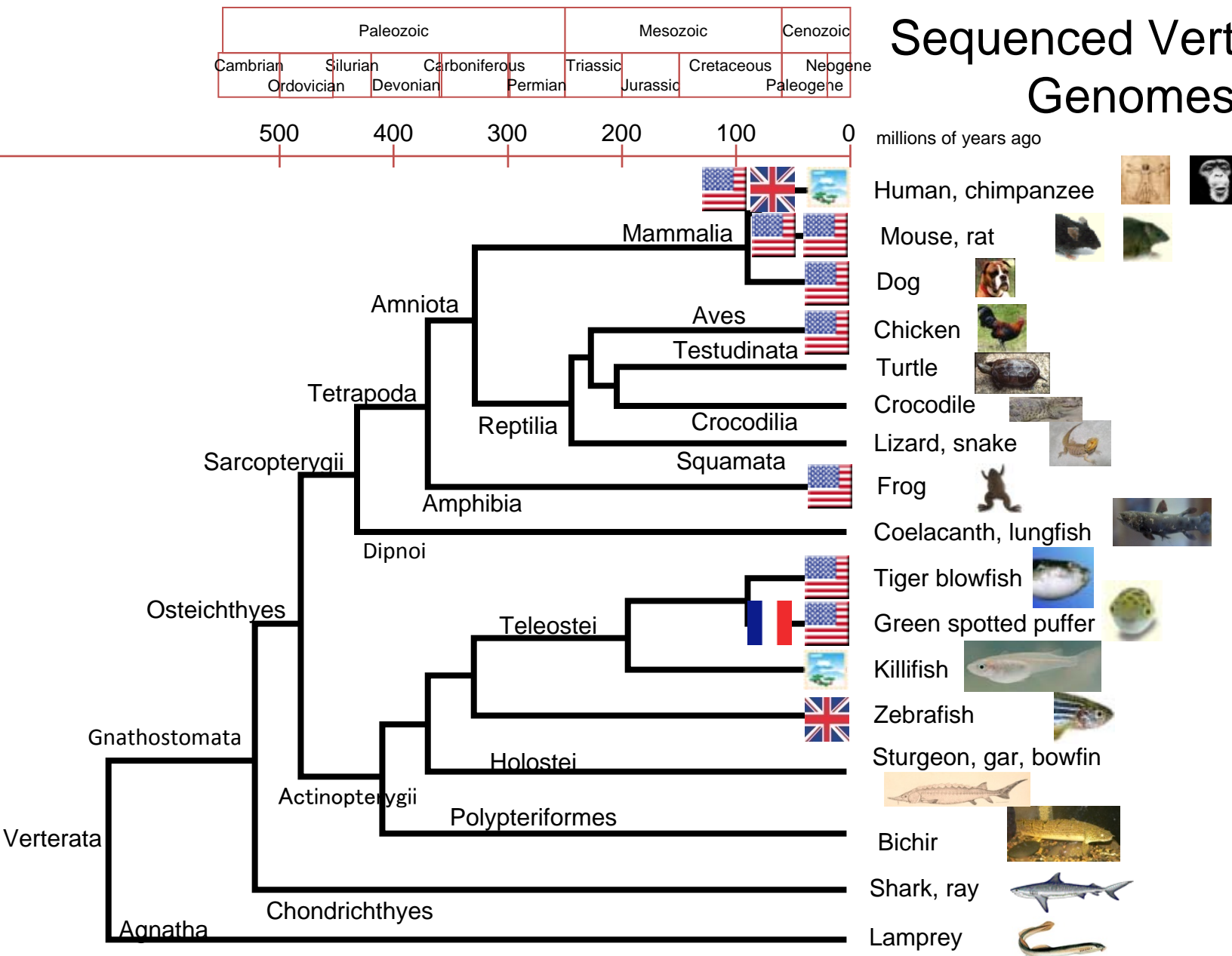
The frequency of codon usage is a bias specific to the organism.

The periodicity of a coding region is the nucleotide triplet.

The standard bias used is that of a six-nucleotide (two-codon) frequency.

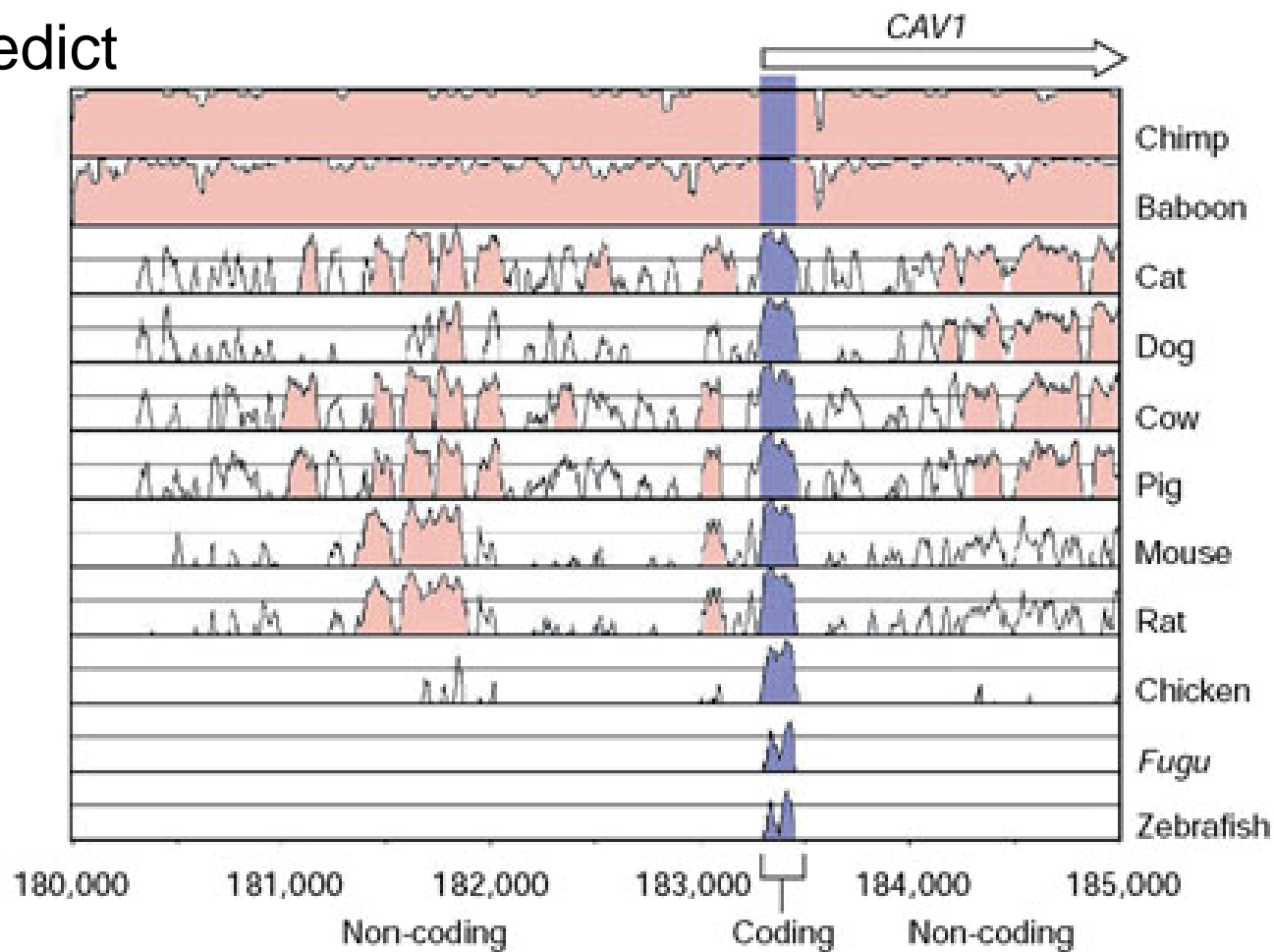
Hidden Markov model

Sequenced Vertebrate Genomes



Compare genomes, identify stored regions, predict genes

Compare
genomes,
identify stored
regions, predict
genes



† *Dubchak and Frazer, 2003, Genome Biology, 4,122*
<http://genomebiology.com/2003/4/12/122>

Acquiring Gene Sequences

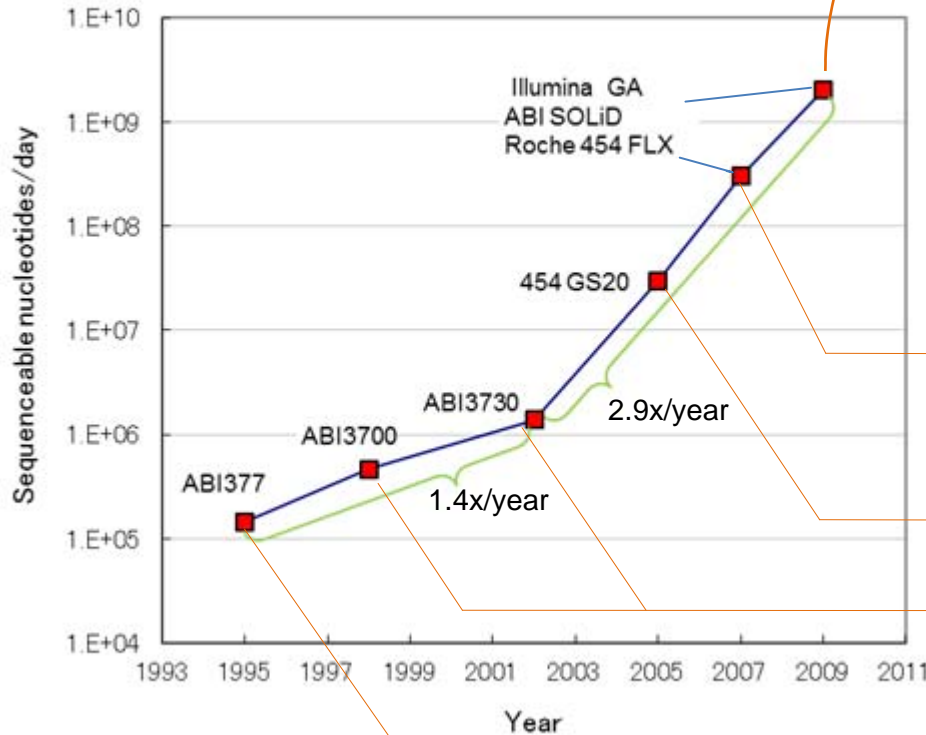
- Synthesize cDNA from mRNA
- Combine cDNA into vectors, propagate and store copies: cDNA library
- Areas where Japan has global prominence:
Sumio Sugano (The University of Tokyo, Institute of Medical Science): Human, other
Yoshihide Hayashizaki (Riken): Mouse
- Extreme difficulty of identifying all mRNA

Figure removed
due to copyright restrictions

Molecular Biology of the Cell - Fifth Edition
Garland Science (2008)
Figure 8-43

Accelerating Genome Sequencing

	Illumina GAIIx	ABI SOLiD 3	Roche 454FLX Titanium
Read length (nucleotides)	75 x 2 = 150	50	500
Reads (hundred mn) / run	0.96~1.2	4	0.01
Days / run	9.5	16	0.4 (10 hr)
Nucleotides per unit of time	15~19	12.5	12
Nucleotides (hundred mn)/day			
Sample volume (µg)	0.1~1	0.01~5	3~5

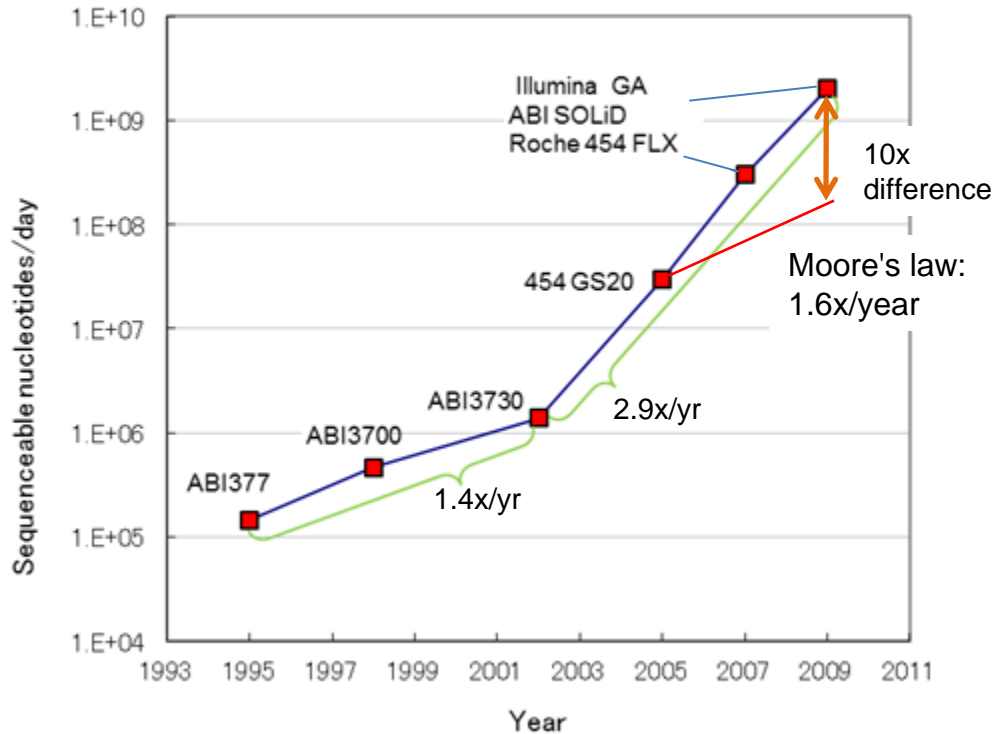


- Genome re-sequencing, transcription start sites, chromatin structures, DNA methylation, RNA sequencing ==> Illumina GA
- De novo sequencing of large genomes, full-length cDNA sequencing, selective splicing ==> Roche 454
- Single-molecule prediction ==> Observations of early development
- Re-sequencing of the Watson genome (454, c.250 bp)
- Re-sequencing of the Asian human genome (Illumina, c.35 bp): mutations, insertions/deletions, inversions, etc.
- DNA methylation (Roche 454, 100-250 bp; Illumina, 36 bp after target capture)
- RNA sequencing (Illumina: 25~35 bp)
- Cover the start points of transcription (Illumina/SOLiD: 25bp)
- Chromatin structures (Illumina/SOLiD, 25 bp)
- Partial sequencing of Neanderthal genome
- Chromatin structures (c.100 bp read length)
- De novo sequencing of human and other large vertebrate genomes, full-length cDNA sequencing (500-800 bp read lengths)

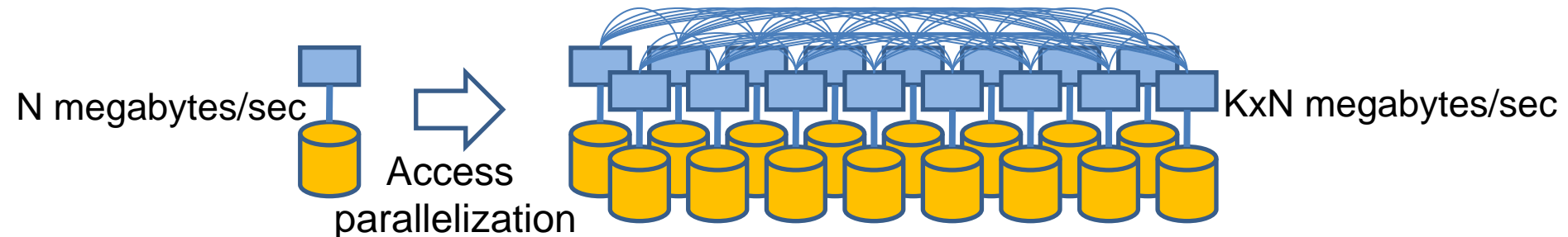
Partial collection of gene sequences

Parallelization of Computational Resources to Improve Performance of Next-Generation Sequencers

Accelerating genome sequencing



- Moore's Law: CPU performance (the number of transistors on a chip) doubles every 1.5 years.
- The performance of next-generation sequencers outstripping Moore's Law will improve some tenfold over four years.
- Parallelize ten-times the number of CPUs to maintain processing speeds.
- Bottleneck simultaneous access to secondary storage devices



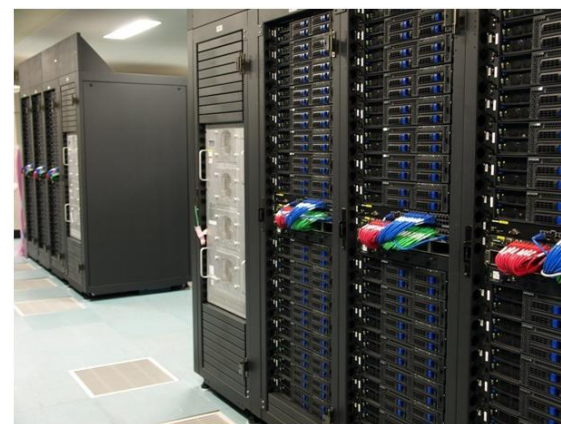
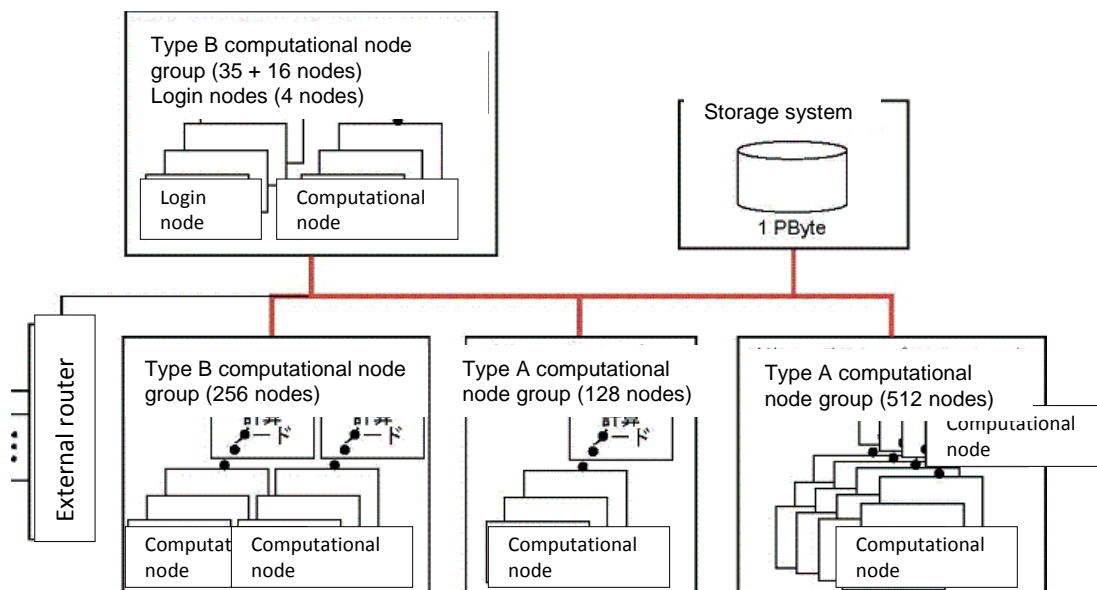
HA8000 Cluster System

at the U. Tokyo Information Technology Center

Nodes	Logical operational performance	147.2 GFLOPS
	Processors (cores)	4(16)
	Main memory	32 GB (936 nodes) 128 GB (16 nodes)
	Local disk capacity	250 GB (incl. RAID 1 OS space)
Processors	Processors (clock speed)	AMD Opteron 8386 (2.3 GHz)
	Cache memory	L2: 512 KB/core L3: 2 MB/processor
	Logical core operational performance	9.2 GFLOPS

Top-performing
supercomputer in Japan

Global Top 500
27th in Nov 2008
16th in Jun 2008



<http://www.cc.u-tokyo.ac.jp/ha8000/>
 † the U. Tokyo Information Technology Center



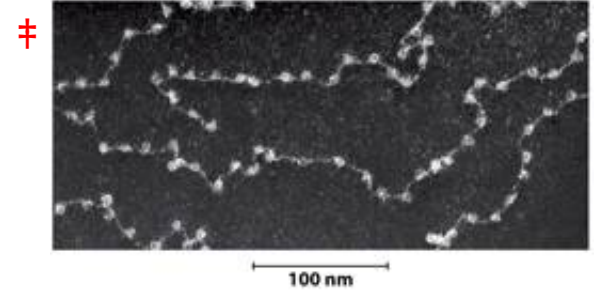
†

Jun Wang (1976 -)

Comprehensive Description of Chromatin Structure

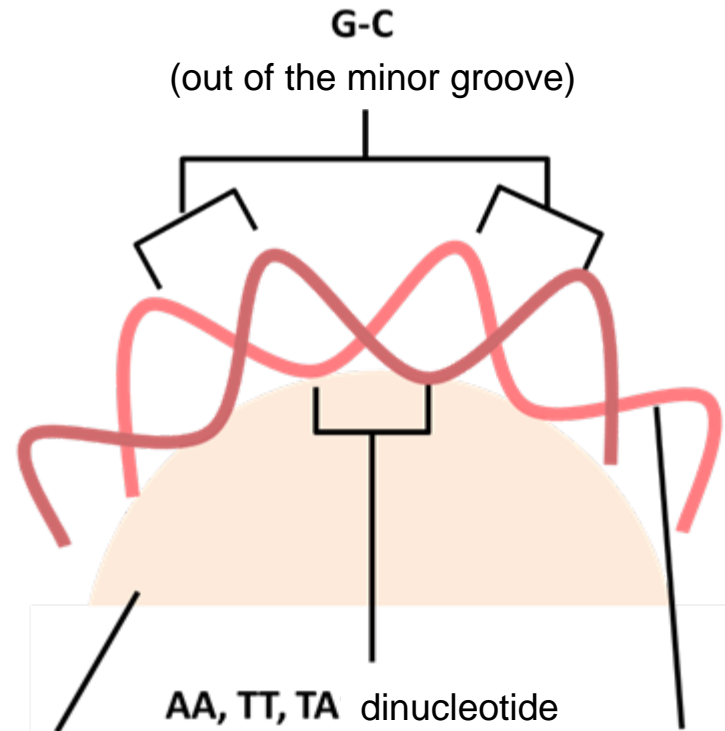
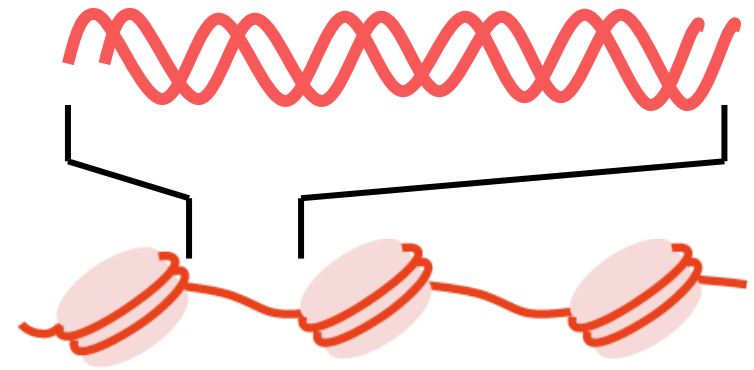
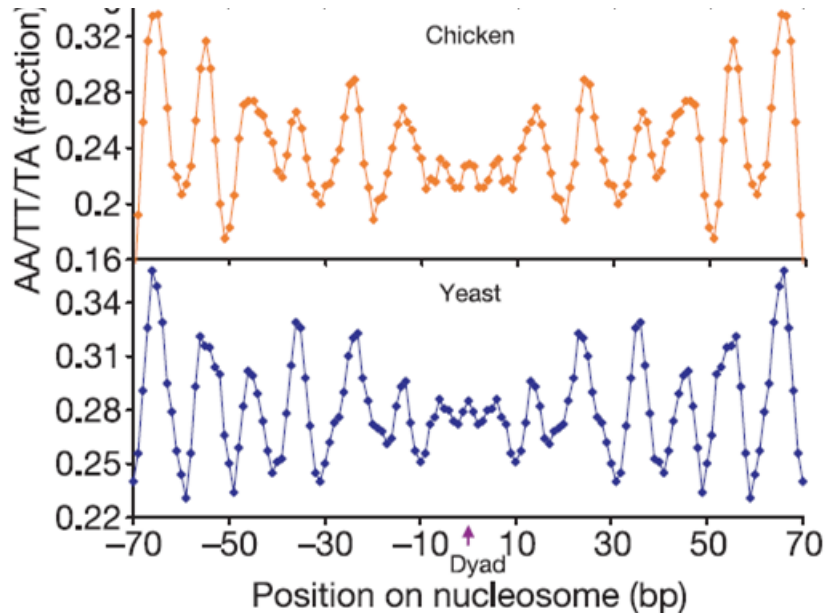
Figure removed
due to copyright restrictions

Molecular Biology of the Cell - Fifth Edition
Garland Science (2008)
Figure 4-72



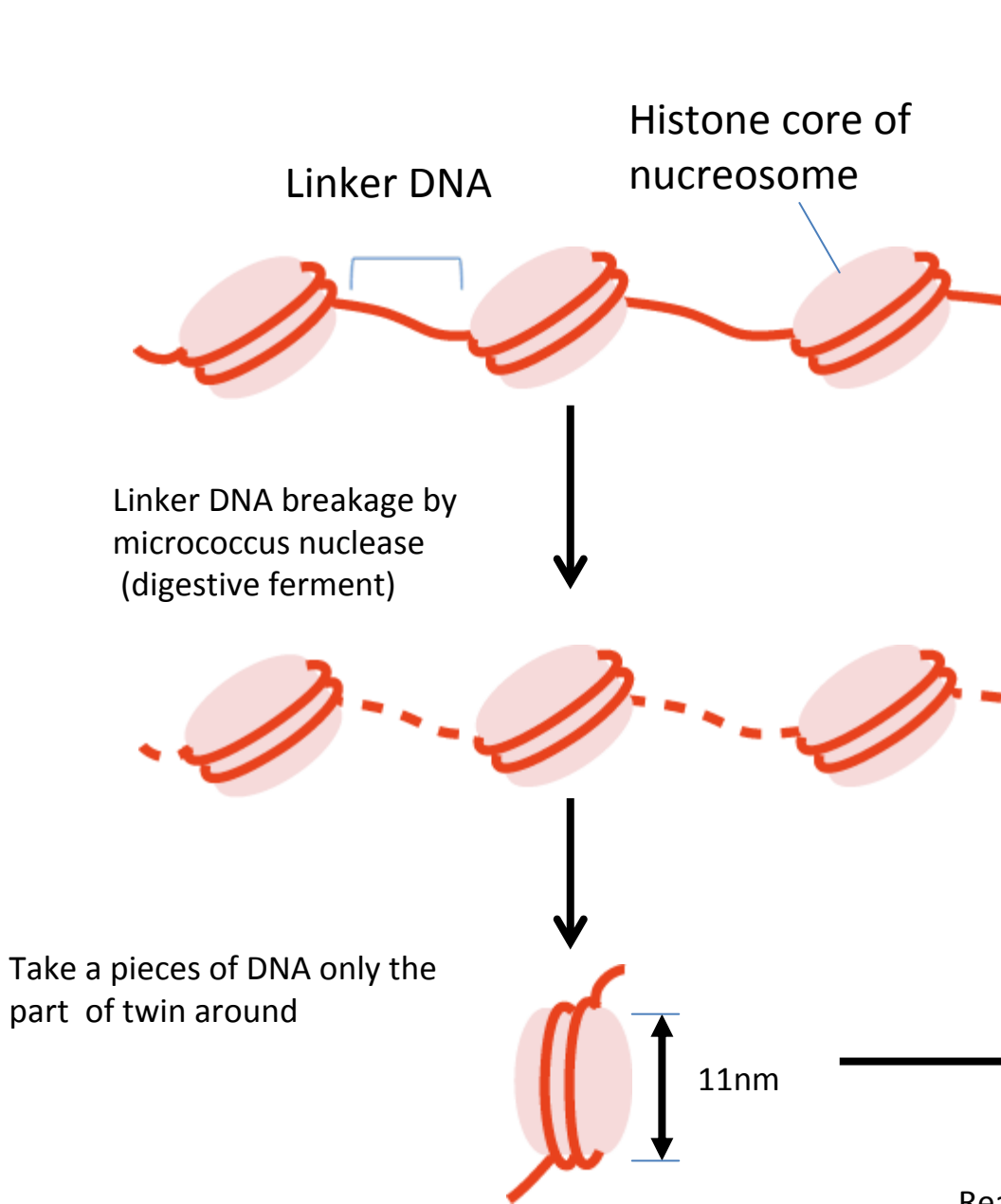
Jeremy M. Berg,
2006,
Biochemistry 6th edition,
W.H. Freeman & Co.

Can nucleosome core positions be predicted from a genome sequence alone?



Histone core (inside of the minor groove) DNA of nucleosome

† Reprinted by permission from Macmillan Publishers Ltd: Segal et al., *Nature* 442(7104):772-8, copyright (2006)



- A nucleosome core is present every 160-200 base pairs
- The human genome has 15 to 20 million nucleosome cores
- c.2,000 sequences/day in 2002 (ABI 3730)
- 10 million sequences/day attainable since 2007 (Illumina GA)

Read both ends out from a gene-sequencing machine

In a population of cells, positions of nucleosome cores are unlikely to be stable.

Figure removed
due to copyright restrictions

Molecular Biology of the Cell - Fifth Edition
Garland Science (2008)
Figure 4-23 (part2of2)

Summary

- Genome size and chromosome count do not necessarily characterize an organism.
- A genome has many different uses.
- Repeated sequences complicate genome sequencing. Computational analysis is essential.
- Prediction, genome comparison and cDNA collection are used in conjunction to infer gene code regions.
- The capability of genome sequencers is making tremendous strides in recent years.
- It is now possible to characterize chromatin structures.