Visualizing Whole Genomes

Agenda

- Molecular biology/single gene questions to bioinformatic/whole genome questions
- How high throughput sequence data is generated
- Visualizing whole genome sequence data
- Hands-on Exercise: Using Genome Browsers

Timeline

1865 Mendel's experiments reveal that, in peas, there are two different copies (alleles) of each gene and these copies segregate independently during meiosis and reunite during fertilization.

1953 DNA structure revealed by X-ray diffraction.

1966 The genetic code deciphered.

1977 Sanger sequencing invented.

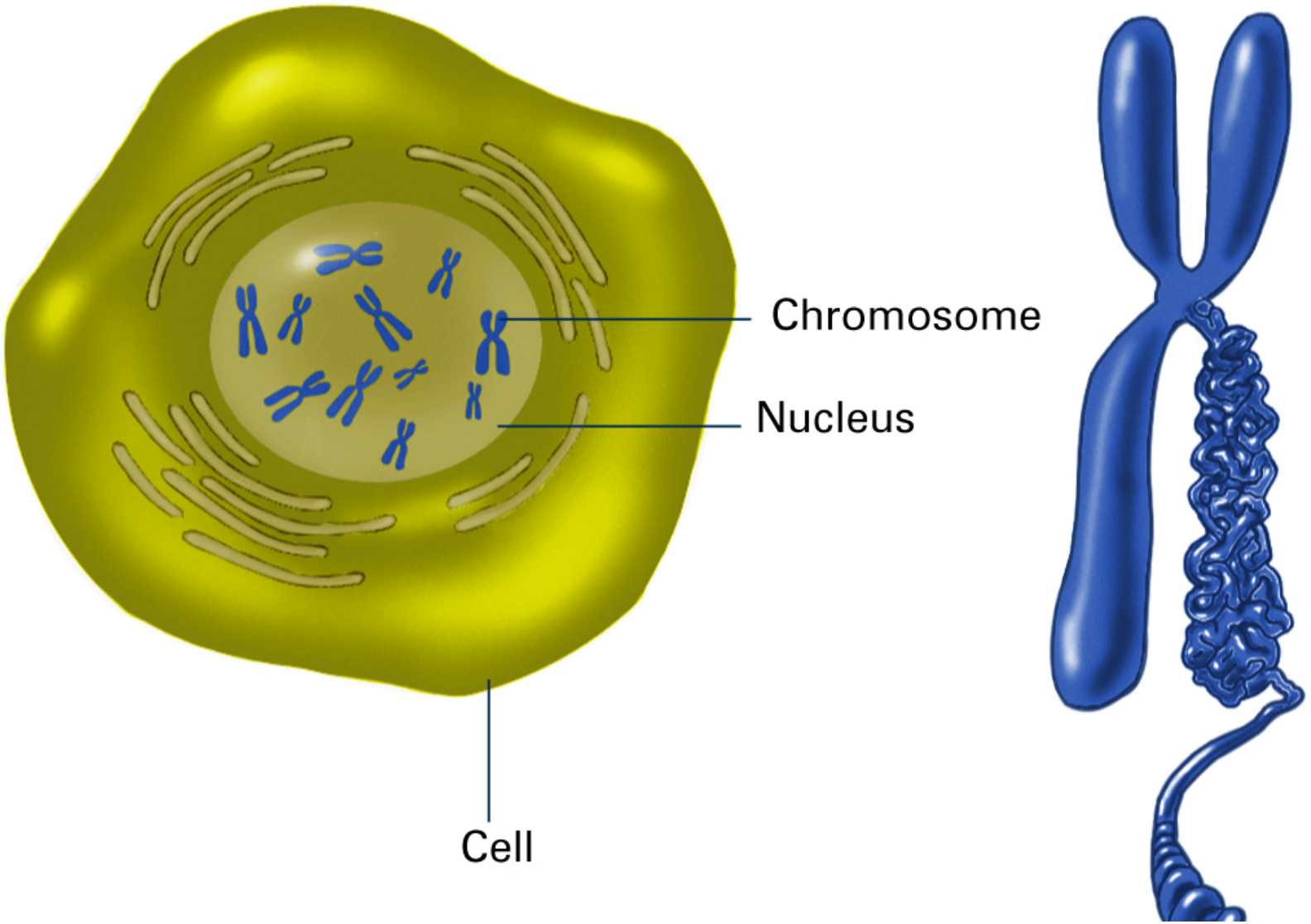
1985 PCR invented.

1990 Human genome project started.

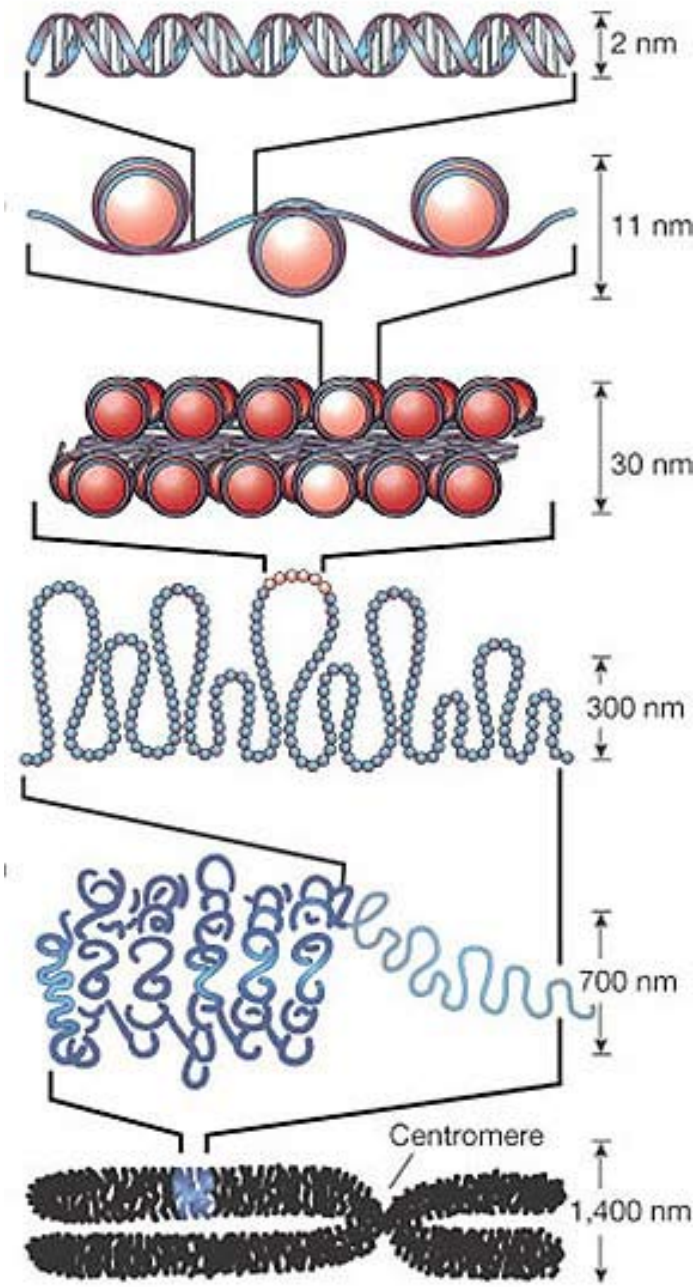
1995 First bacterial genome sequenced (H. influenzae).
1996 First eukaryote sequenced (S. cerevisae).
1998 First multicellular eukaryote sequenced (C. elegans).
2001 "Draft" human genome sequenced "finished".

2012 First photo illustrating DNA double-helical structure obtained.

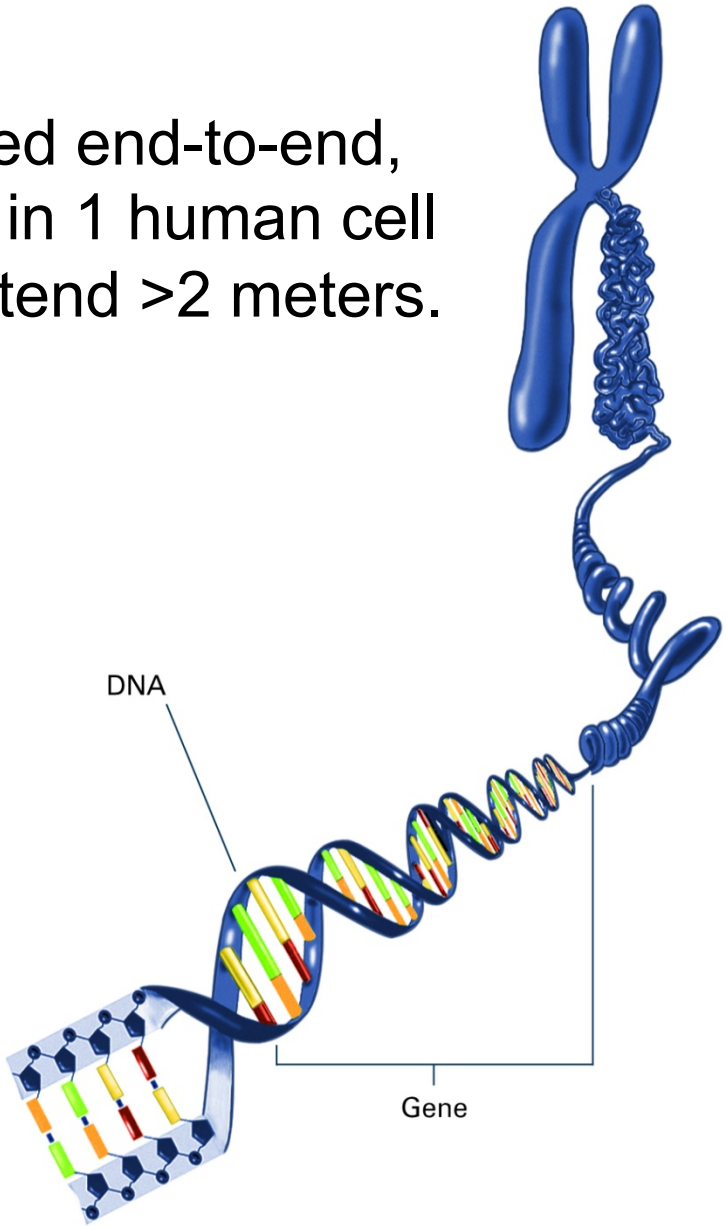
Cells and DNA



DNA

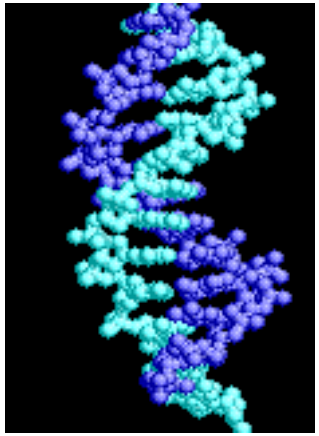


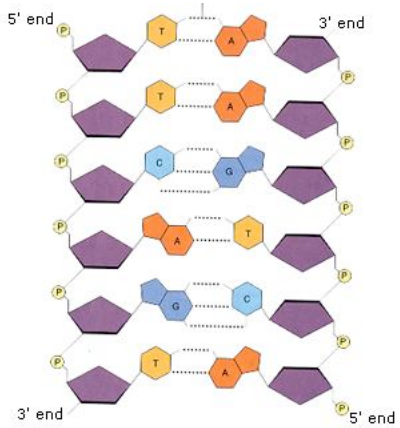
Stretched end-to-end, the DNA in 1 human cell would extend >2 meters.

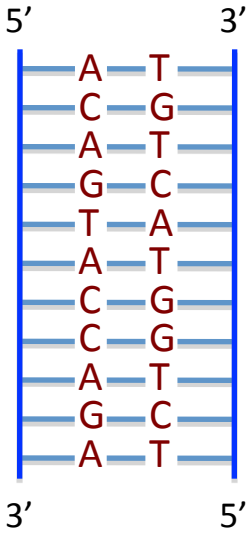


A, T, G, C

DNA







5’ ACAGTACCAGACCAGACCATACATACCATC 3’
3’ TGTCATGGTCTGGTCTGGTATGTATGGTAG 5’

ACAGTACCAGACCAGACCATACATACCATC

Genomic Gene structure of a typical Eukaryotic gene

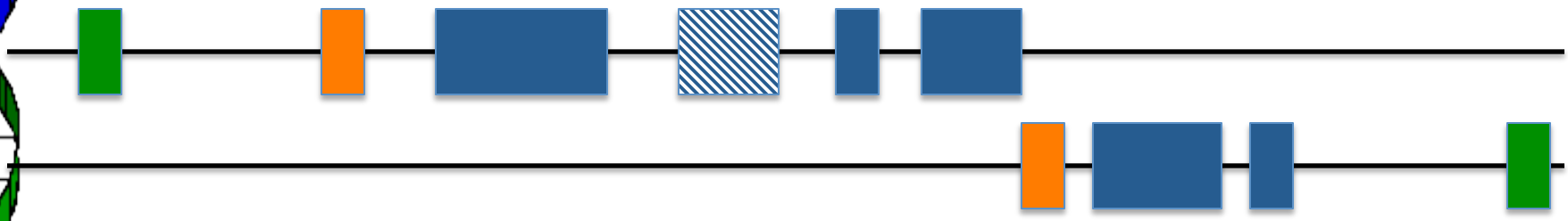
- Coding vs. template strand
- Coding regions-Exons or Open Reading Frames (ORF)
- Introns
  - Splice site coding region
    - GU….CACUGAC…. AG

```
5′ ACAGTACCAGACCAGACCATACATACCATC 3′
3′ TGTCATGGTCTGGTCTGGTATGTATGGTAG 5′
```

Genomic Gene structure of a typical Eukaryotic gene

- Promoters
  - transcription factor binding sites
  - Transcription start sites
- Enhancers
- Transcriptional and translational start/stop sites

```
5′ ACAGTACCAGACCAGACCATACATACCATC 3′
3′ TGTCATGGTCTGGTCTGGTATGTATGGTAG 5′
```



Direct Visualization of the DNA Double Helical Structure *(in 2012!)*

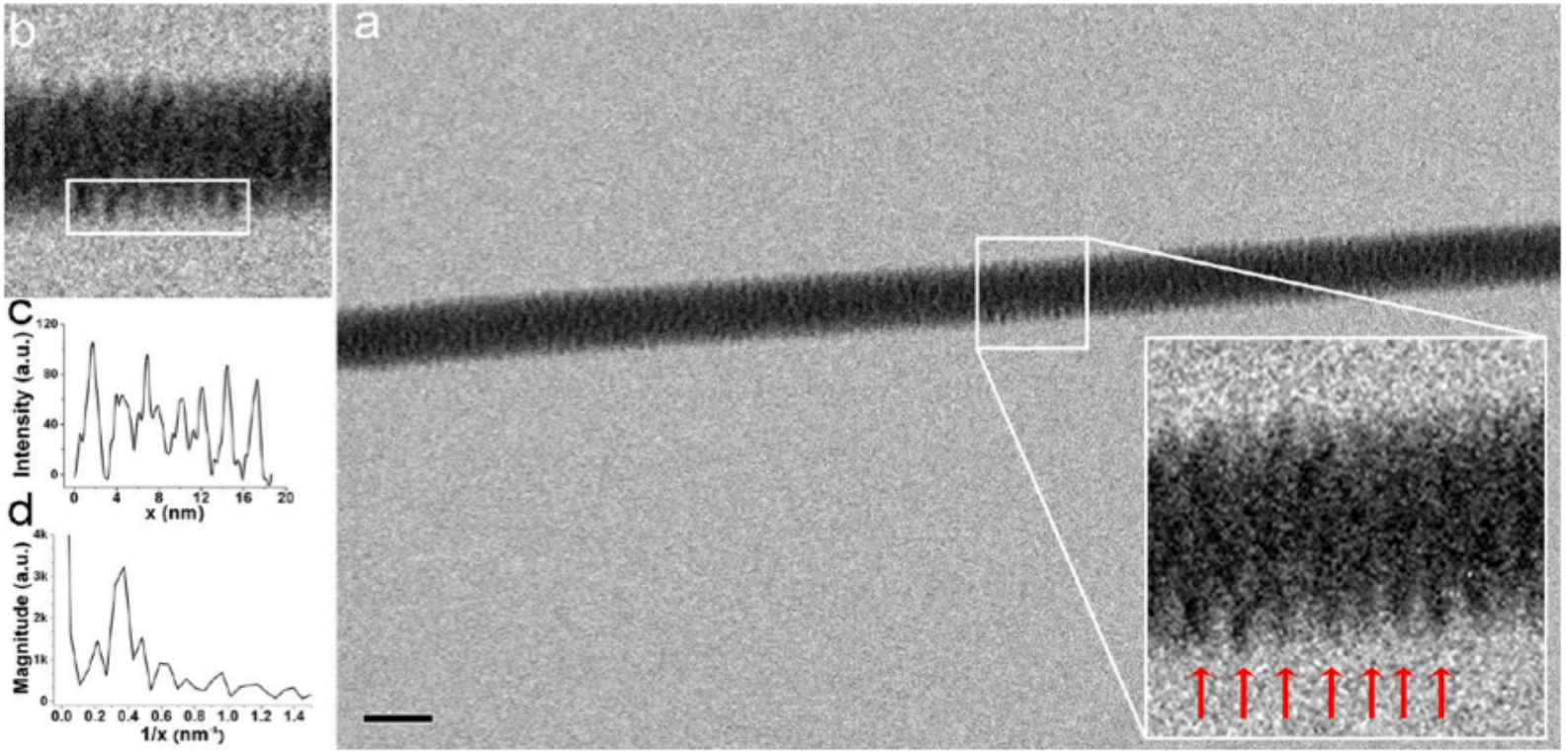
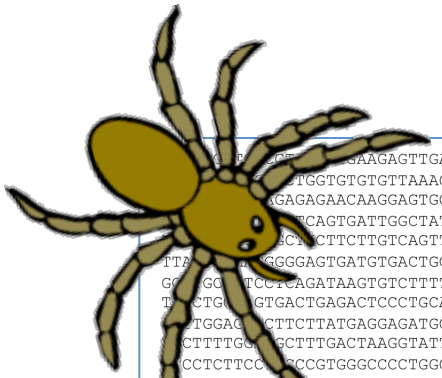
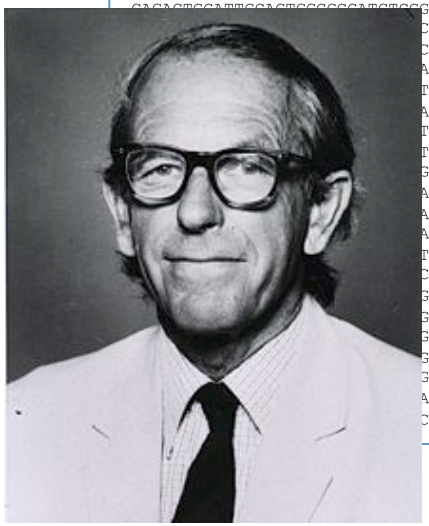


Figure 2. TEM image with intensity profile and corresponding FFT pitch calculation of λ-DNA fibers. (a) DNA fiber TEM image. The inset shows higher magnification DNA fiber details; the red arrows point out the 2.7 nm pitch of A double helix. The scale bar corresponds to a length of 20 nm. In panel b, a white rectangle is superimposed, showing where the intensity profile was measured. The peaks in plot c correspond to the alternation of bright and dark bands in the original image (b): plot c displays a two-dimensional graph where the *Y*-axis reports the pixel intensity integrated along the height of the rectangle and the *X*-axis represents the distance measured on the rectangle. Plot d shows the FFT of the signal displayed in plot c: a well-defined maximum is observed at 0.37 ± 0.02 1/nm, corresponding to a frequency of 2.7 ± 0.2 nm.

Gentile et al. 2012

Single Gene Analysis





Frederick Sanger

Won the Nobel Prize in Chemistry.

Twice!

GENSCAN 1.0 Date run: 11-Oct-110 Time: 18:59:17

Predicted genes/exons:

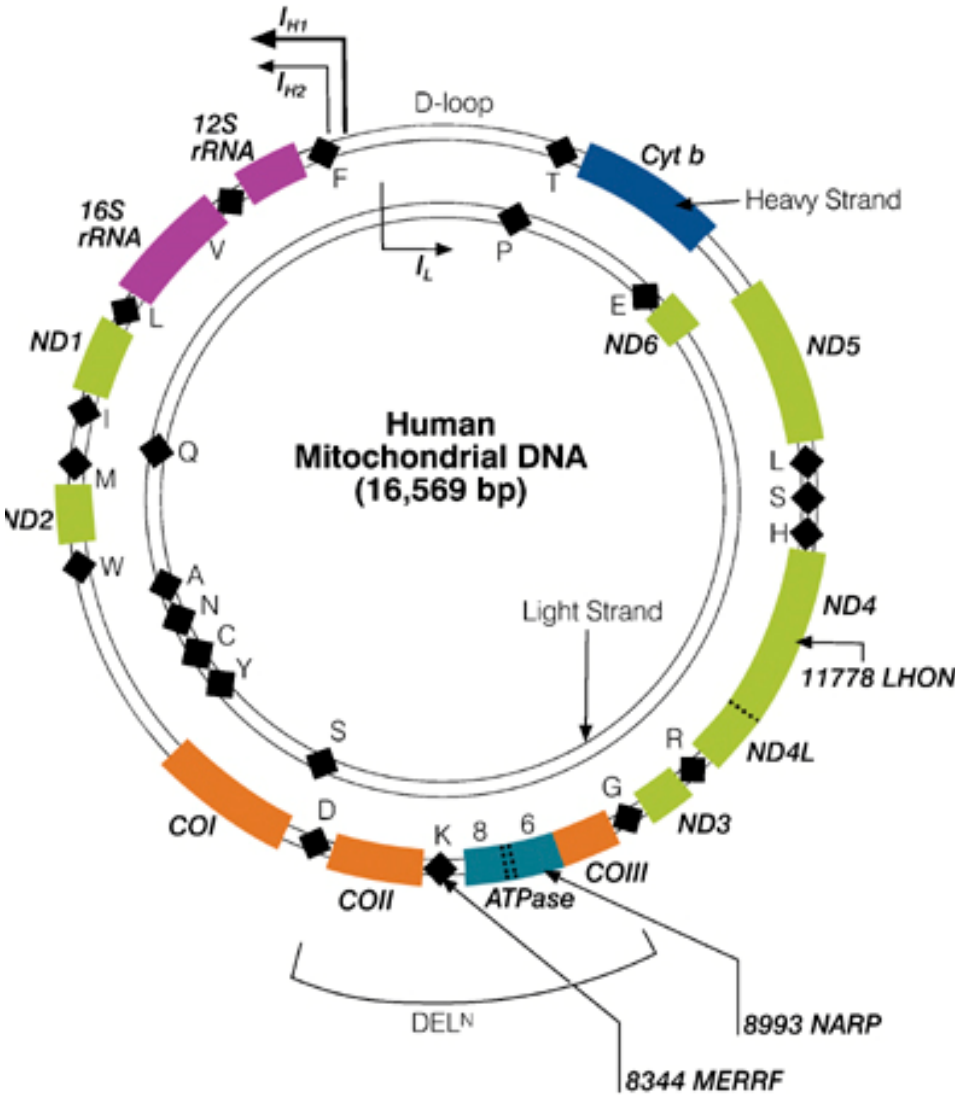
Gn.Ex	Type	S	.Begin	...End	.Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr..
2.19	PlyA	-	7211	7206	6							1.05
2.18	Term	-	13843	13590	254	2	2	108	42	118	0.968	4.02
2.17	Intr	-	15438	15265	174	1	0	84	101	177	0.992	17.59
2.16	Intr	-	15685	15554	132	2	0	52	88	81	0.948	4.20
2.15	Intr	-	15898	15772	127	1	1	36	90	145	0.995	8.83
2.14	Intr	-	16113	15992	122	1	2	89	103	97	0.999	10.59
2.13	Intr	-	16751	16527	225	0	0	107	110	117	0.999	12.83
2.12	Intr	-	17118	16989	130	0	1	70	108	64	0.994	5.95
2.11	Intr	-	21198	21070	129	0	0	96	76	138	0.711	13.37
2.10	Intr	-	22694	22569	126	2	0	68	88	31	0.654	1.06
2.09	Intr	-	23217	23044	174	0	0	40	116	72	0.846	4.41
2.08	Intr	-	24430	24355	76	0	1	70	80	80	0.886	4.00
2.07	Intr	-	24727	24558	170	1	2	72	89	66	0.983	2.92
2.06	Intr	-	32478	32216	263	1	2	13	76	214	0.570	8.98
2.05	Intr	-	32660	32508	153	0	0	70	90	172	0.522	14.72
2.04	Intr	-	37066	36834	233	0	2	27	31	206	0.117	5.29
2.03	Intr	-	37751	37488	264	1	0	65	45	139	0.133	3.00
2.02	Intr	-	42503	42384	120	1	0	55	48	89	0.086	0.29
2.01	Init	-	48661	48513	149	1	2	90	42	133	0.361	8.51
2.00	Prom	-	56908	56869	40							-5.85

Predicted peptide sequence(s):

>GENSCAN_predicted_peptide_2|1006_aa
MSKLVKIIRRPLKLDSNQALFLVVSRDSLVSISMLIYEVYESEKDEDEFLGSLEFGSPEH
VWQLKGHWELDLFYGTSRIILGVSPHGLKCASSAWAPTLAALEEPFSPRLHCGSLFLGWP
RPEPAPSACGEVWRERRGWEPGLPMALEGQREFRVGMGSACPALRAASRPAIPAGPGQVC
ECTNGHSVSGYSGGDLENLCVDTLYLANLVVTWRTFVSRSGIVNAPISILSKRTNQLFVK
WTNRLSVKWTNQQDVVSTDWPLVLLLDRSRLPPPSGAFRWADSRPLPPASHRSSRNECFH
LFRDEPDEGDGAARIREEGMAPADGKRVGGAEAVSSSGEAAPSGRRRPLGREARAALGPR
SSAAMVAKLRWRPRAGAGRLGERGGWSPFATLARVCGHEMLAHRAVLACCSPYLFEIFNS
DSDPHGISHVKFDDLNPEAVEVLLNYAYTAQLKADKELVKDVYSAAKKLKMDRVKQVCGD
YLLSRMDVTSCISYRNFASCMGDSRLLNKVDAYIQEHLLQISEEEEFLKLPRLKLEVMLE
DNVCLPSNGKLYTKVINWVQRSIWENGDSLEELMEEVQTLYYSADHKLLDGNLLDGQAEV
FGSDDDHIQFVQVHIAQSEKKPPRENGHKQISSSSTGCLSSPNATVQSPKHEWKIVASEK
TSNNTYLCLAVLDGIFCVIFLHGRNSPQSSPTSTPKLSKSLSFEMQQDELIEKPMSPMQY
ARSGLGTAEMNGKLIAAGGYNREECLRTVECYNPHTDHWSFLAPMRTPRARFQMAVLMGQ
LYVVGGSNGHSDDLSCGEMYDSNIDDWIPVPELRTNRCNAGVCALNGKLYIVGGSDPYGQ
KGLKNCDVFDPVTKLWTSCAPLNIRRHQSAVCELGGYLYIIGGAESWNCLNTVERYNPEN
NTWTLIAPMNVARRGAGVAVLNGKLFVCGGFDGSHAISCVEMYDPTRNEWKMMGNMTSPR
SNAGIATVGNTIYAVGGFDGNEFLNTVEVYNLESNEWSPYTKIFOF

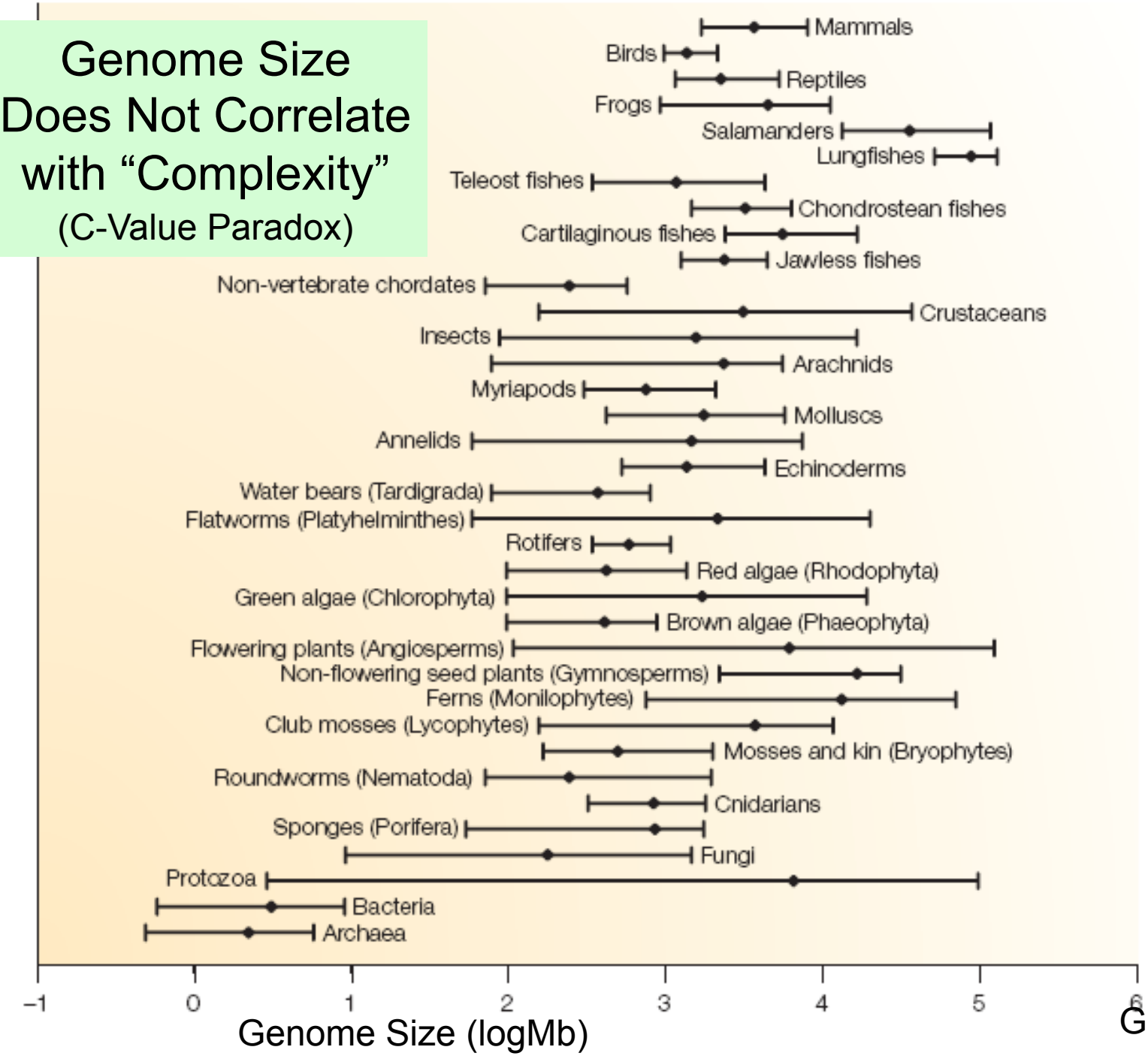
The WHOLE GENOME SEQUENCING Era
Genome Content, Size, and Organization Varies

- Genome Content
  - Number of genes
  - Number of repeats
- Genome size
  - Usually just refers to the nuclear genome
    - mtDNA
    - Chloroplast
    - Others!
- Ploidy



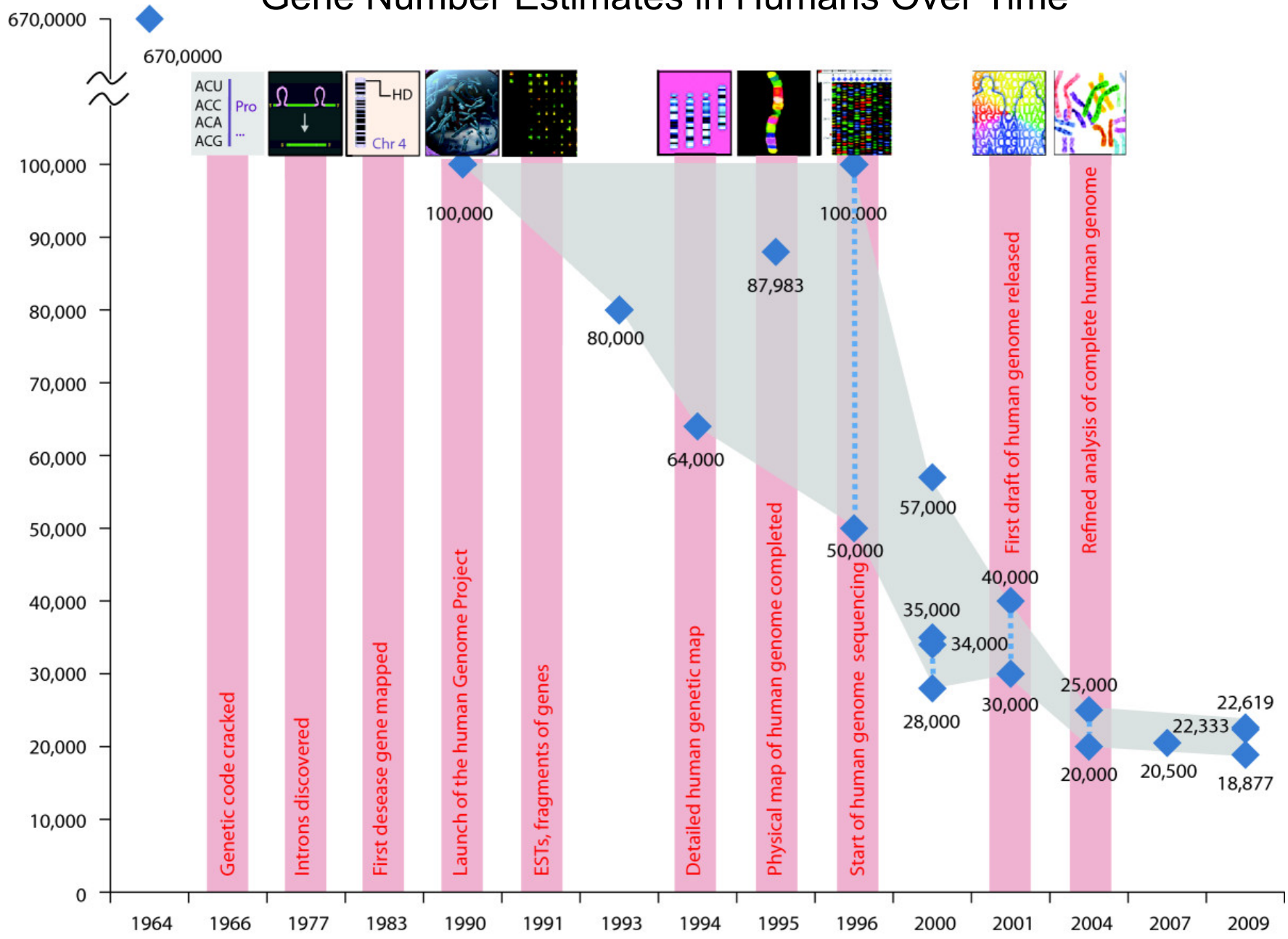
Genome Size Does Not Correlate with "Complexity"

(C-Value Paradox)



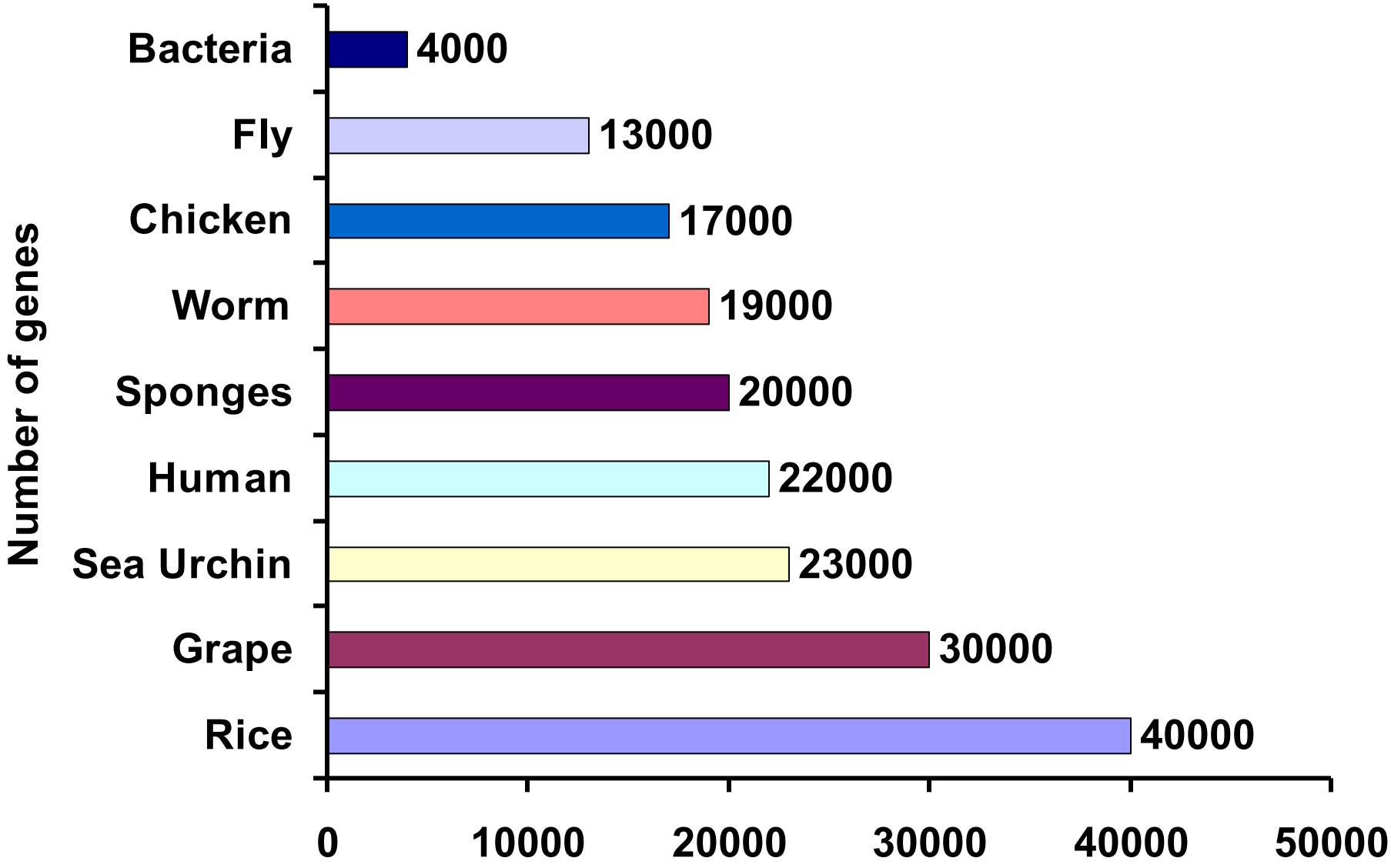
Gregory 2005

Gene Number Estimates in Humans Over Time

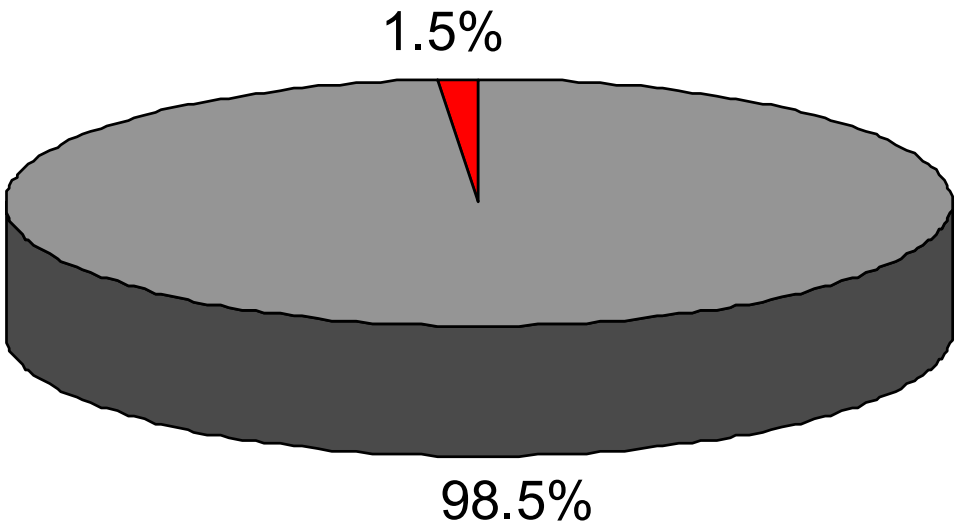


Pertea and Salzberg 2010

Approximate Number of Genes in the Whole Genome Sequence of Various Organisms



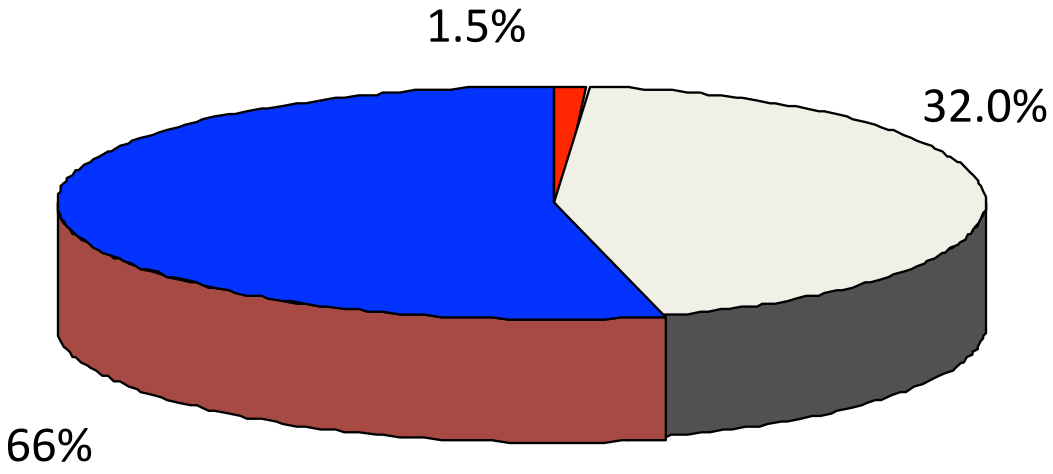
Percent Coding DNA



Humans, as an example

- Non-Coding DNA
- Coding DNA

What is the genome composed of, if not genes?



- Coding DNA
- Transposable Elements
- Non-coding DNA (which may also include many unrecognizable TEs)

Lander et al. 2001, De Koning et al. 2012

What sequencing methods exist?

- Sanger
- Shotgun
- High Throughput
- 3rd Generation

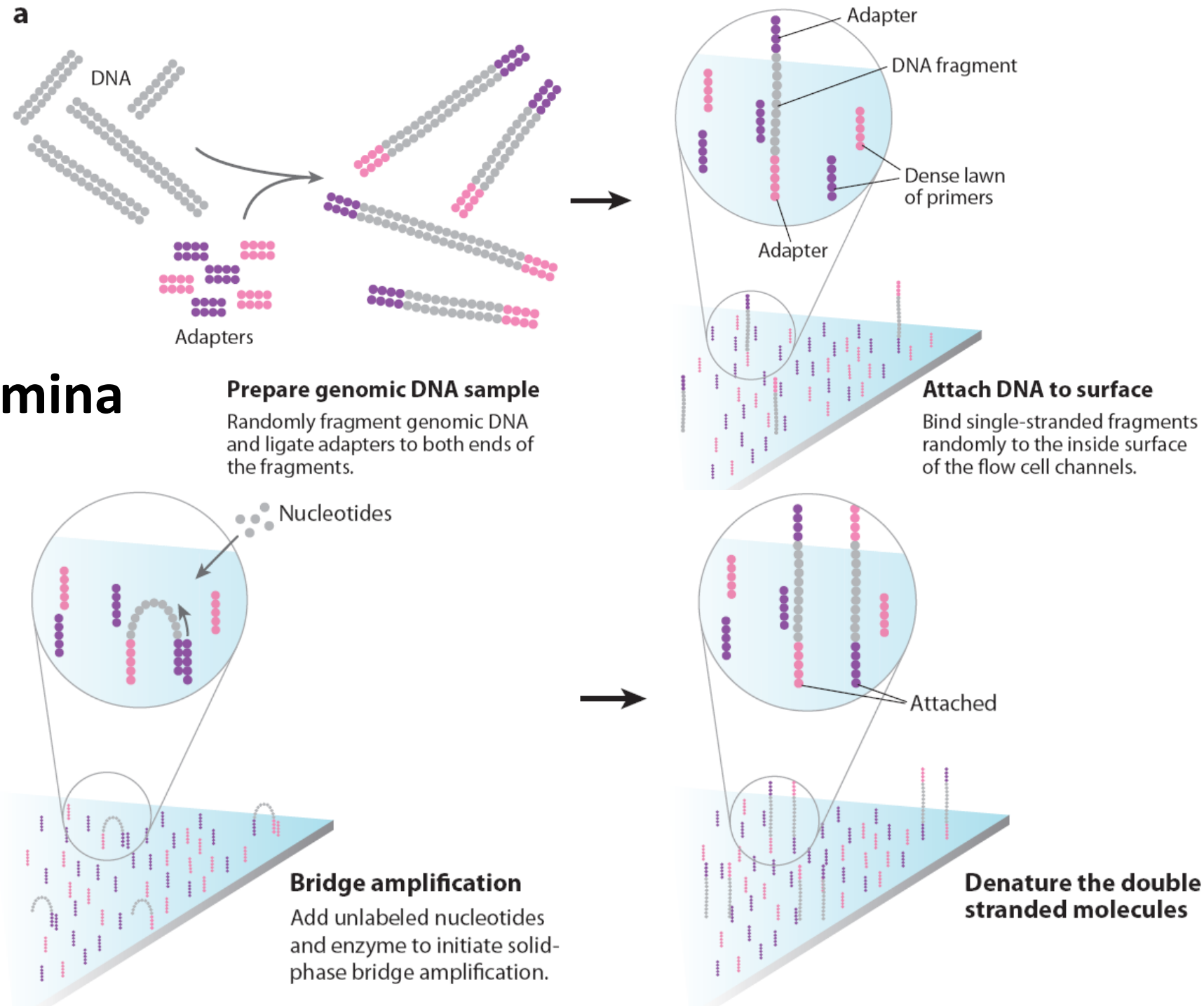
Single molecule sequencing

- Machines
- Core Facilities
- Genome Centers
- Cloud Computers

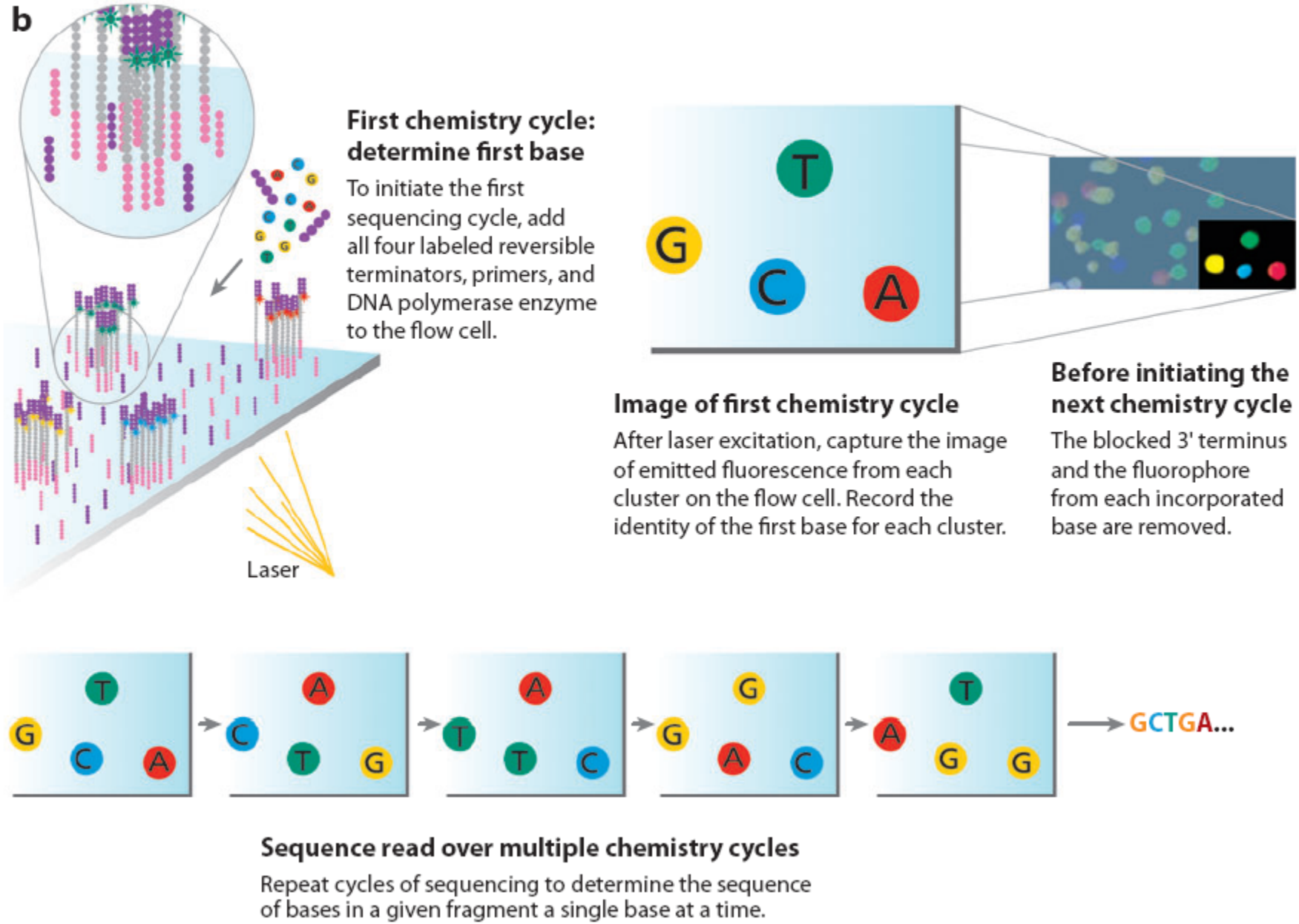
Considerations

Time (sample prep, machine time)
Cost
Error rate
Assembly

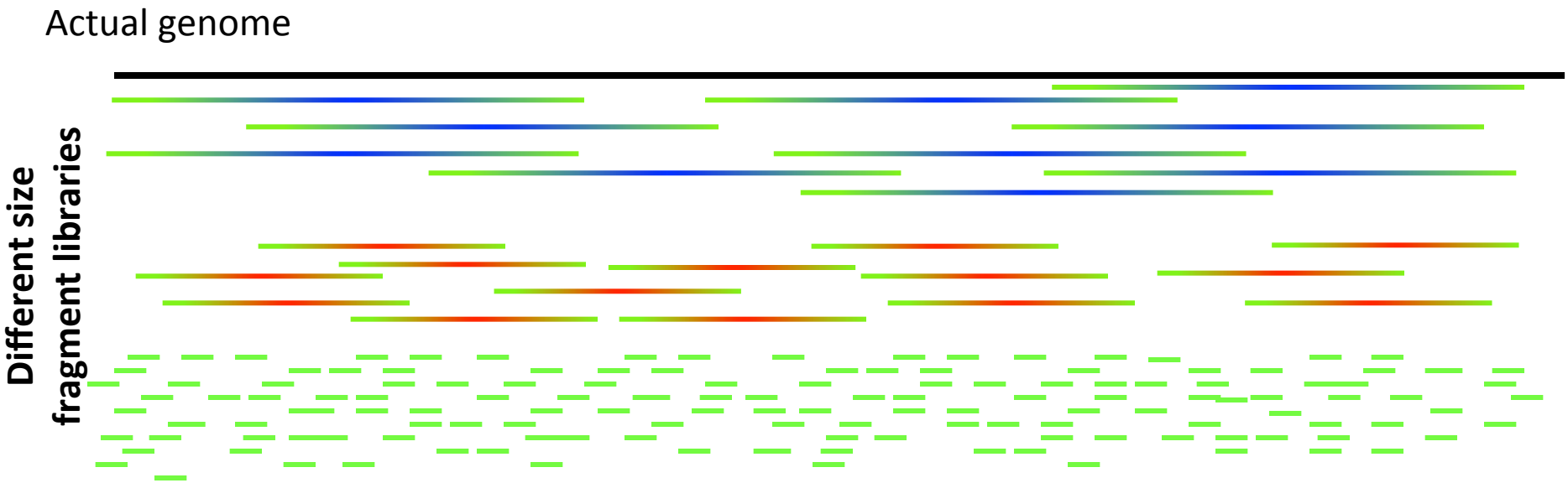
Illumina


a
DNA
Adapters
Prepare genomic DNA sample
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.
Adapter
DNA fragment
Dense lawn of primers
Adapter
Attach DNA to surface
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.
Nucleotides
Bridge amplification
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.
Attached
Denature the double stranded molecules

Illumina, cont.



Shotgun Sequencing and Assembly



DNA is sheared and only the ends are sequenced.

These are paired, and the assembly algorithm makes use of this information in order to assemble the short reads into contigs and scaffolds.

Problems:

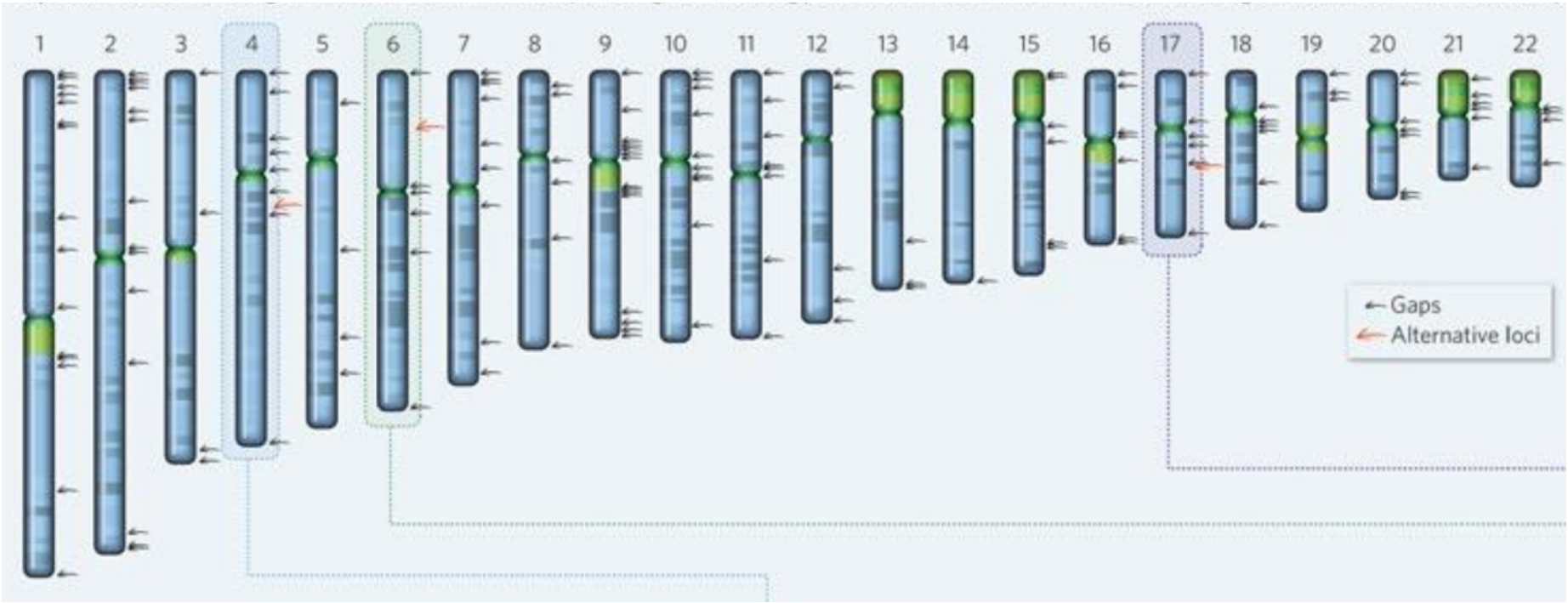
Non-random shearing

Amplification bias

Repeats

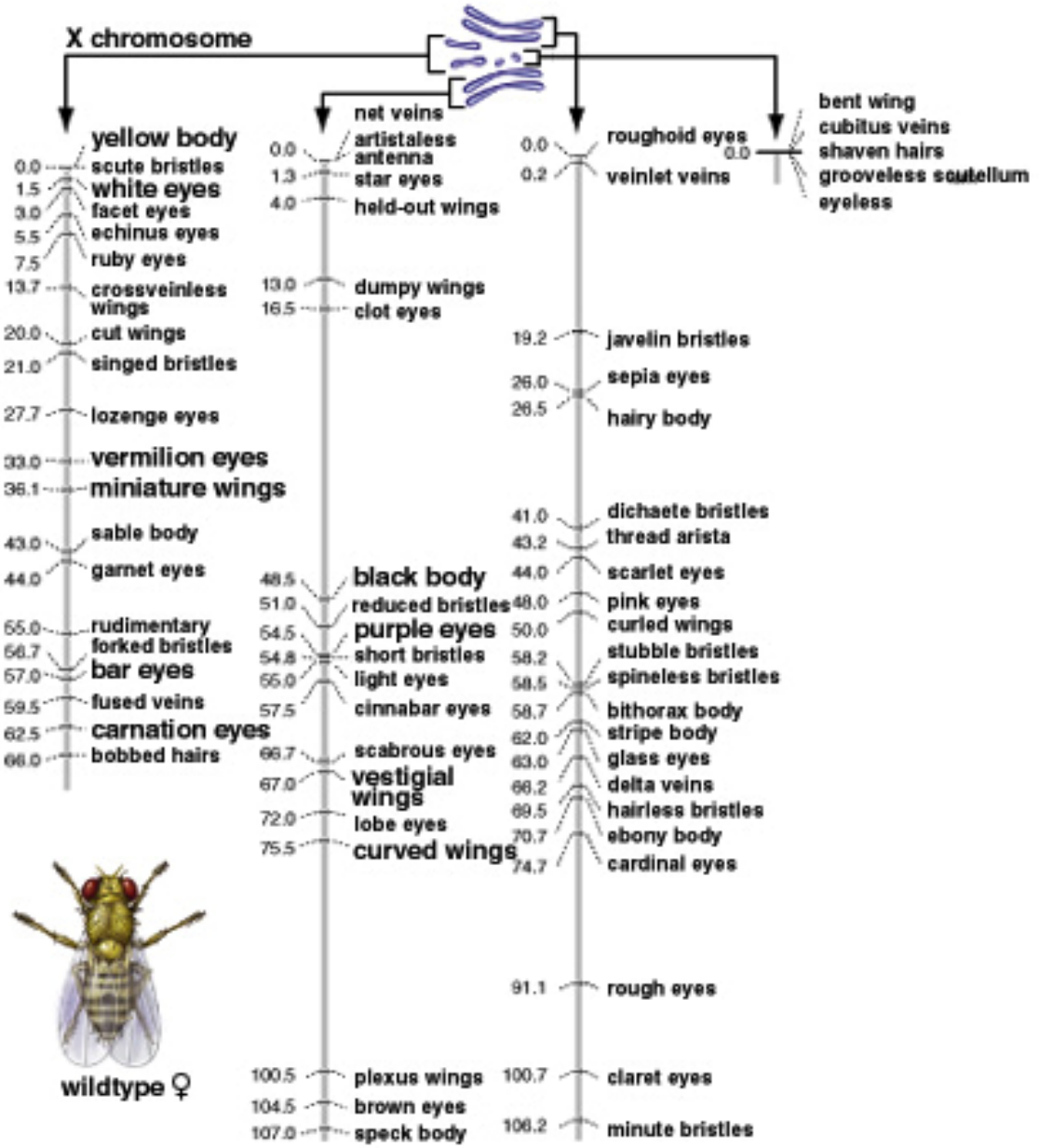
Finishing a WGS

Additional paired end sequencing, manual sequencing across gaps, genetic maps, $, time.



- Fix (mis)assemblies, closing gaps, linkage group=chromosome
- Takes many years and $$$
- Less momentum, interest, lower publication value
- Certain regions “can’t” be finished

Drosophila genetic map



Ongoing Issues

- Error rates
- Problems sequencing heterochromatin
- Variable loci
  - paralogs versus alleles (heterozygosity)
  - heteroplasmy (organellar genomes)
- Assembly of short reads
  - repetitive DNA

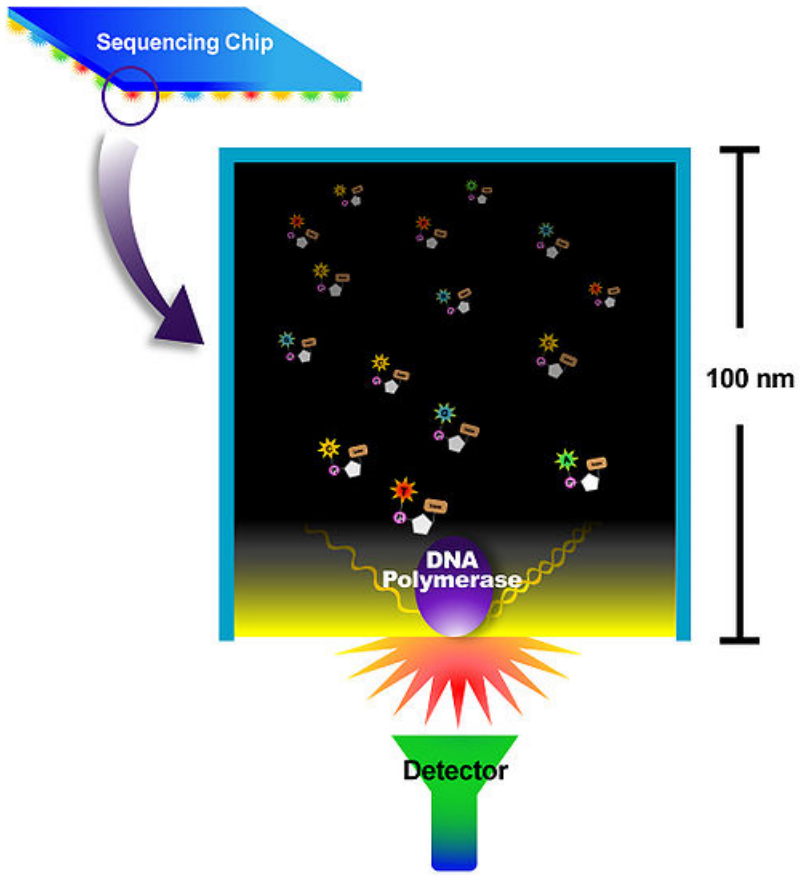
3rd Gen Advances

Single Molecule Real Time Sequencing

Less reagent and sample preparation
no PCR, no amplification bias

Longer readlengths
(average = 1000 bps, but up to 10,000 bps) < coverage required to assemble or detect rearrangements

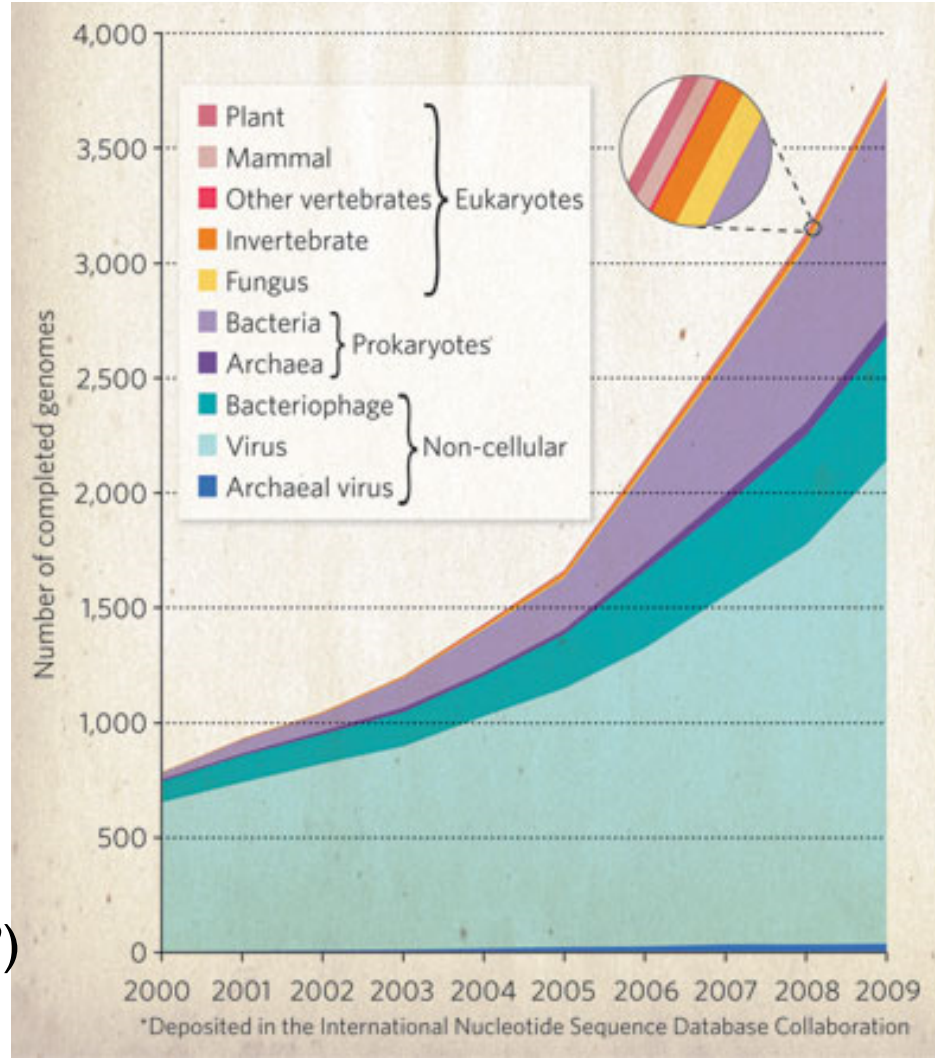
Faster
no flushing, scanning and washing steps



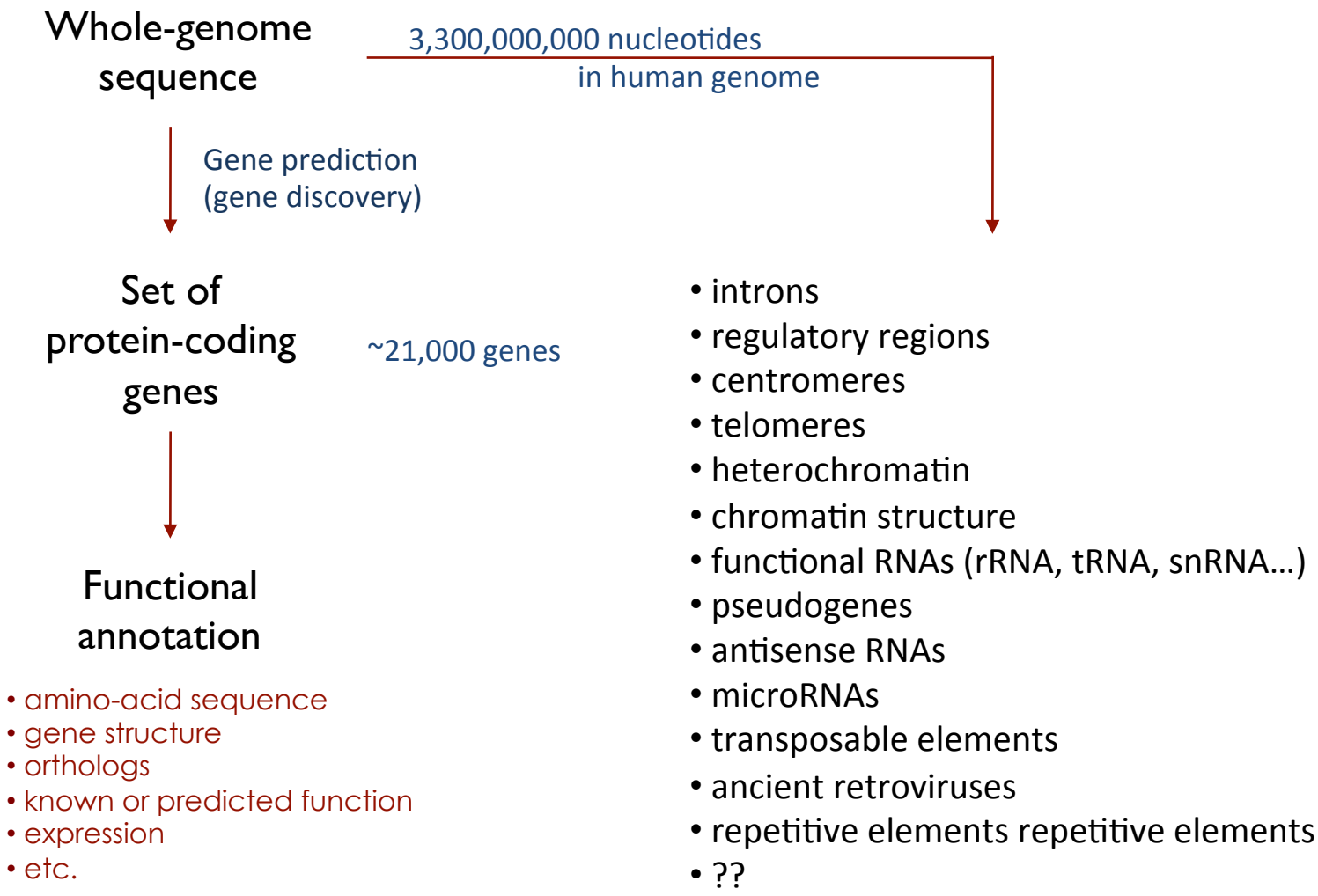
What's Done With Next Gen Sequencing?

- Whole genomes
- Transcriptomes
- Targeted resequencing
  - Tissue comparisons
  - Small RNAs
  - Population genomics
  - Epigenomics
  - ChIP-seq
- Metagenomics

1st human genome (13 yrs, $3B)
human genome today ($10K, 8 days)
human genome in 3 years ($1K, 15 minutes?)



Whole Genome Analysis

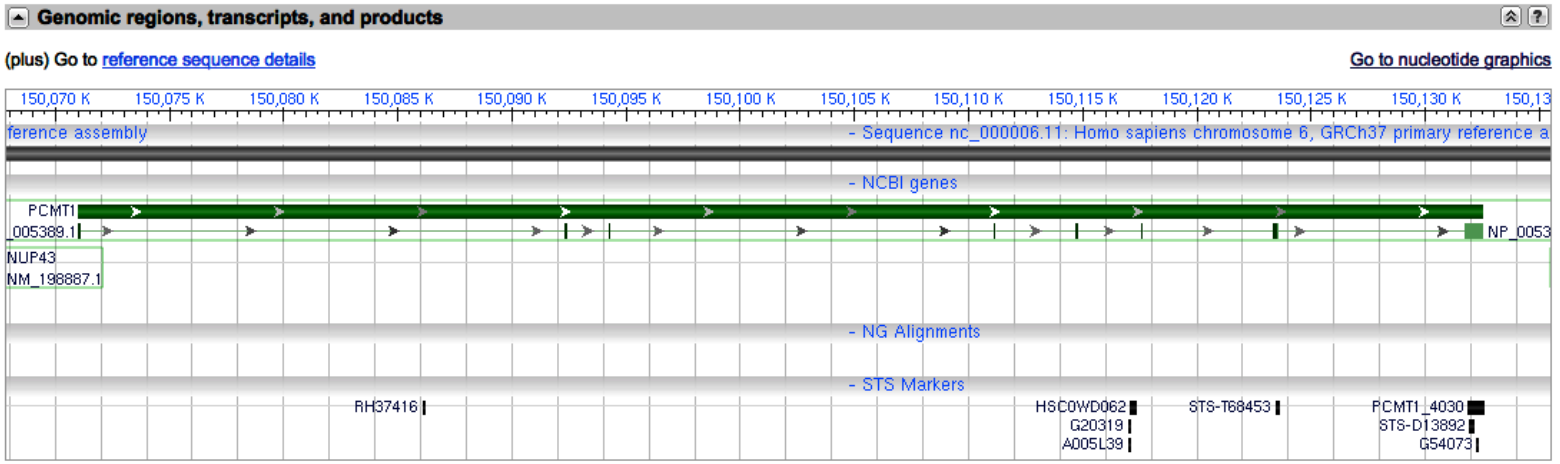


Many Types of Whole Genome Analysis

- Gene/protein prediction
- Transcription levels (RNA seq)
- Identify orthologs
- Microarray/chip hybridization studies
- ChIP
- Phylogenomics
- Protein interactions
- Structure prediction/comparison
- Splicing prediction
- Network analysis
- Variant analysis
- Epigenetic analysis

Also known as PCMT1

Summary Three classes of protein carboxyl methyltransferases, distinguished by their methyl-acceptor substrate specificity, have been found in prokaryotic and eukaryotic cells. The type II enzyme catalyzes the transfer of a methyl group from S-adenosyl-L-methionine to the free carboxyl groups of D-aspartyl and L-isoaspartyl residues. These methyl-accepting residues result from the spontaneous deamidation, isomerization, and racemization of normal L-aspartyl and L-asparaginyl residues and represent sites of covalent damage to aging proteins PCMT1 (EC 2.1.1.77) is a protein repair enzyme that initiates the conversion of abnormal D-aspartyl and L-isoaspartyl residues to the normal L-aspartyl form.[supplied by OMIM]



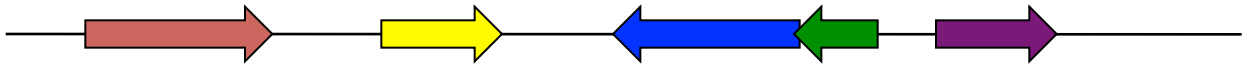
Links

- Order cDNA clone
- BioAssay, by Gene target
- CCDS
- Conserved Domains
- EST
- Full text in PMC
- GEO Profiles
- Genome
- HomoloGene
- Map Viewer
- Nucleotide
- OMIM
- Peptidome
- Probe
- Protein
- PubChem Compound
- PubChem Substance
- PubMed
- PubMed (GeneRIF)
- PubMed (OMIM)
- RefSeq Proteins
- RefSeq RNAs
- SNP

Genome Content

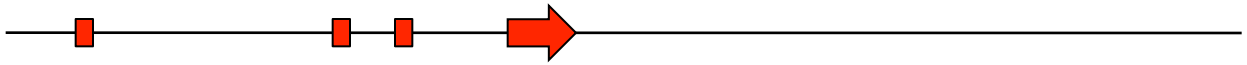
E. coli genome:

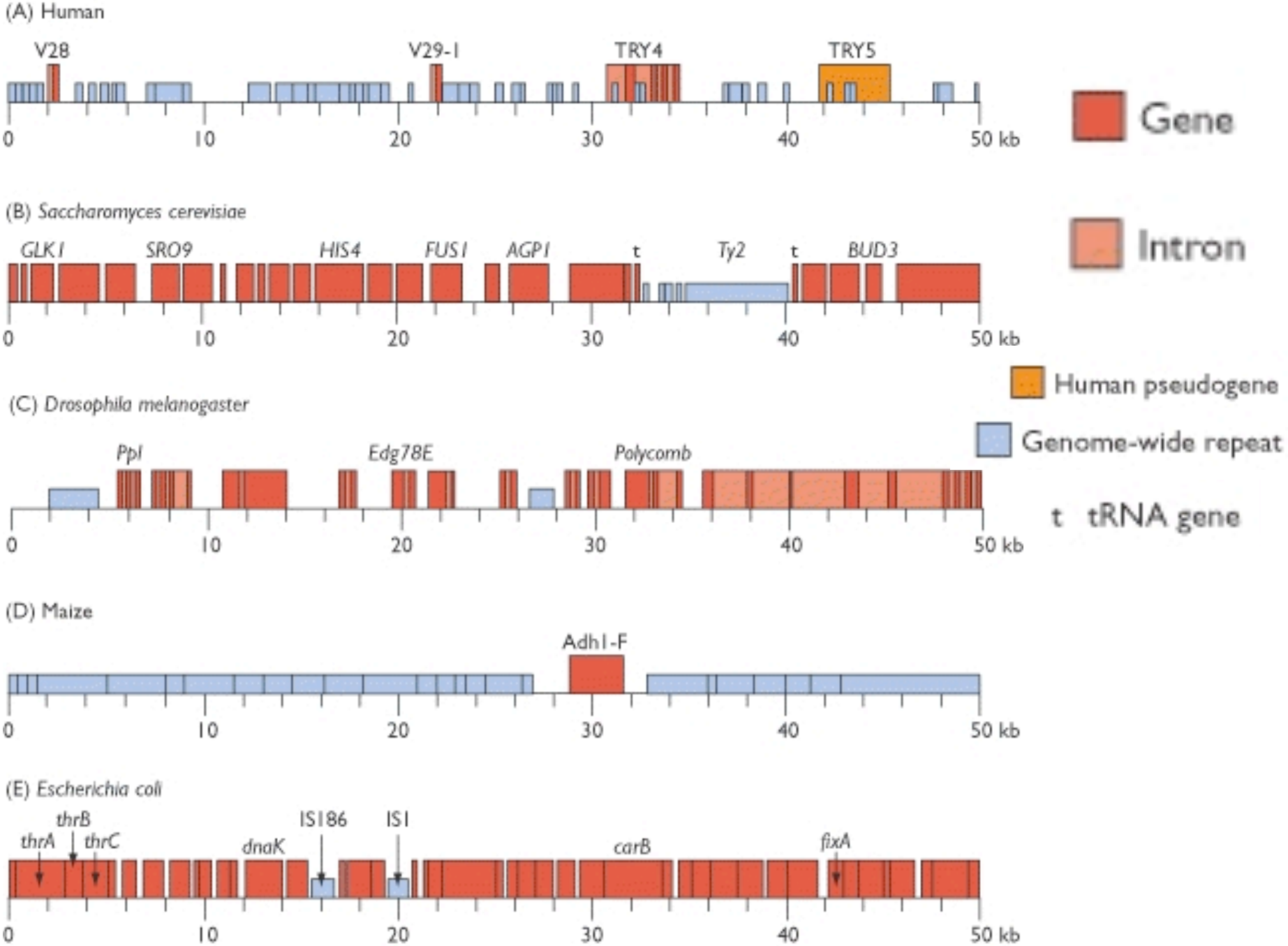
- 4,639,000 bp of DNA
- 4,377 genes (1 gene per 1000 nt; 89% coding)



Human genome:

- 3,200,000,000 bp of DNA
- 21,500 genes (1 gene per 153,000 nt; <2% coding)



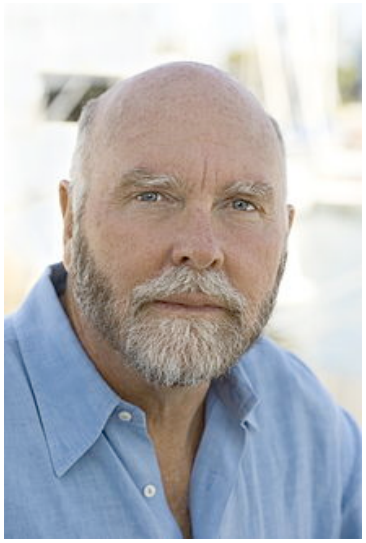

(A) Human
V28
V29-1
TRY4
TRY5
0
10
20
30
40
50 kb
(B) Saccharomyces cerevisiae
GLK1
SRO9
HIS4
FUS1
AGP1
t
Ty2
t
BUD3
0
10
20
30
40
50 kb
(C) Drosophila melanogaster
Ppl
Edg78E
Polycomb
0
10
20
30
40
50 kb
(D) Maize
Adh1-F
0
10
20
30
40
50 kb
(E) Escherichia coli
thrA
thrB
thrC
dnaK
IS186
IS1
carB
fixA
0
10
20
30
40
50 kb
Gene
Intron
Human pseudogene
Genome-wide repeat
t tRNA gene

The Human Genome Project

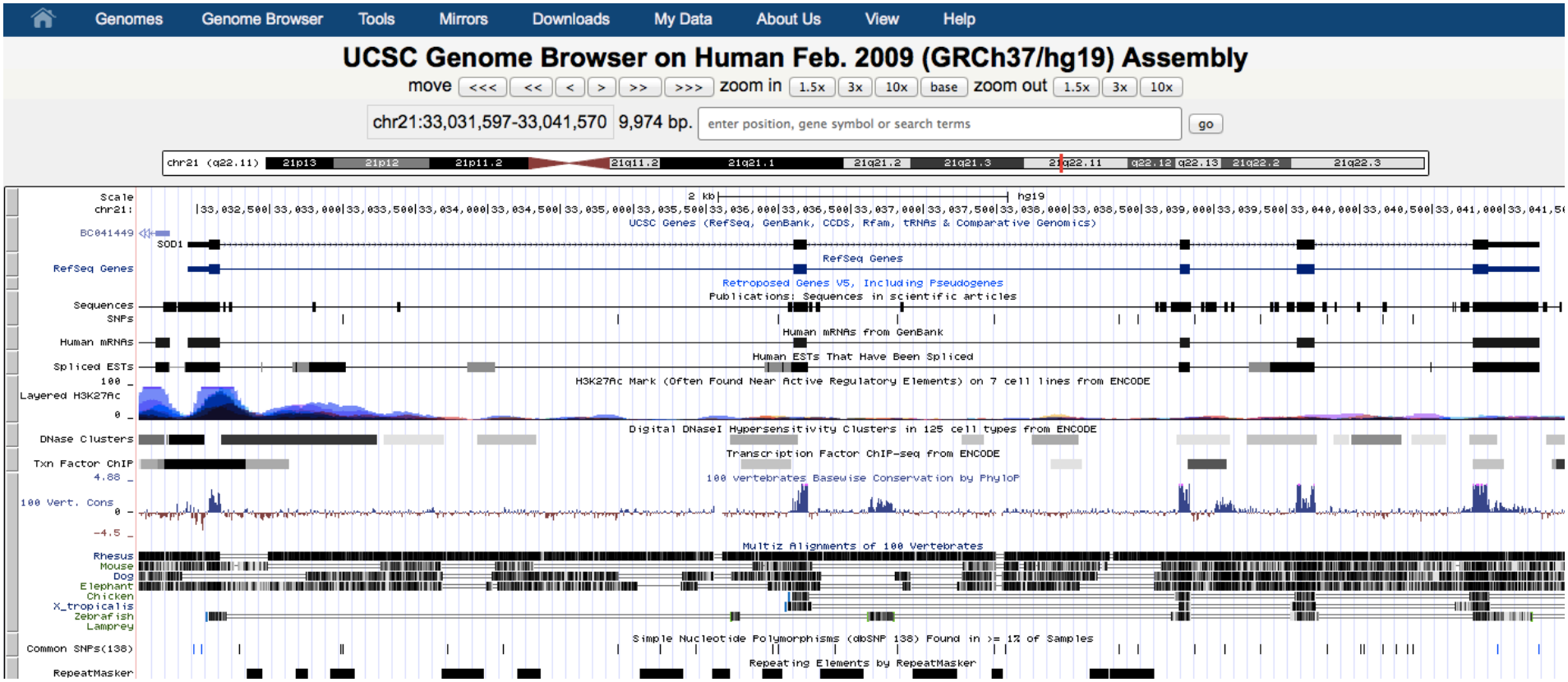
Genome “completed” in 2000

UCSC Genome Browser built in 2000

- Different versions of the human genome
- Many other genomes
- Many other datasets (as different “tracks”)
  - new mRNA data added each night
  - new EST data added each week
- Third party tracks
- Your tracks!

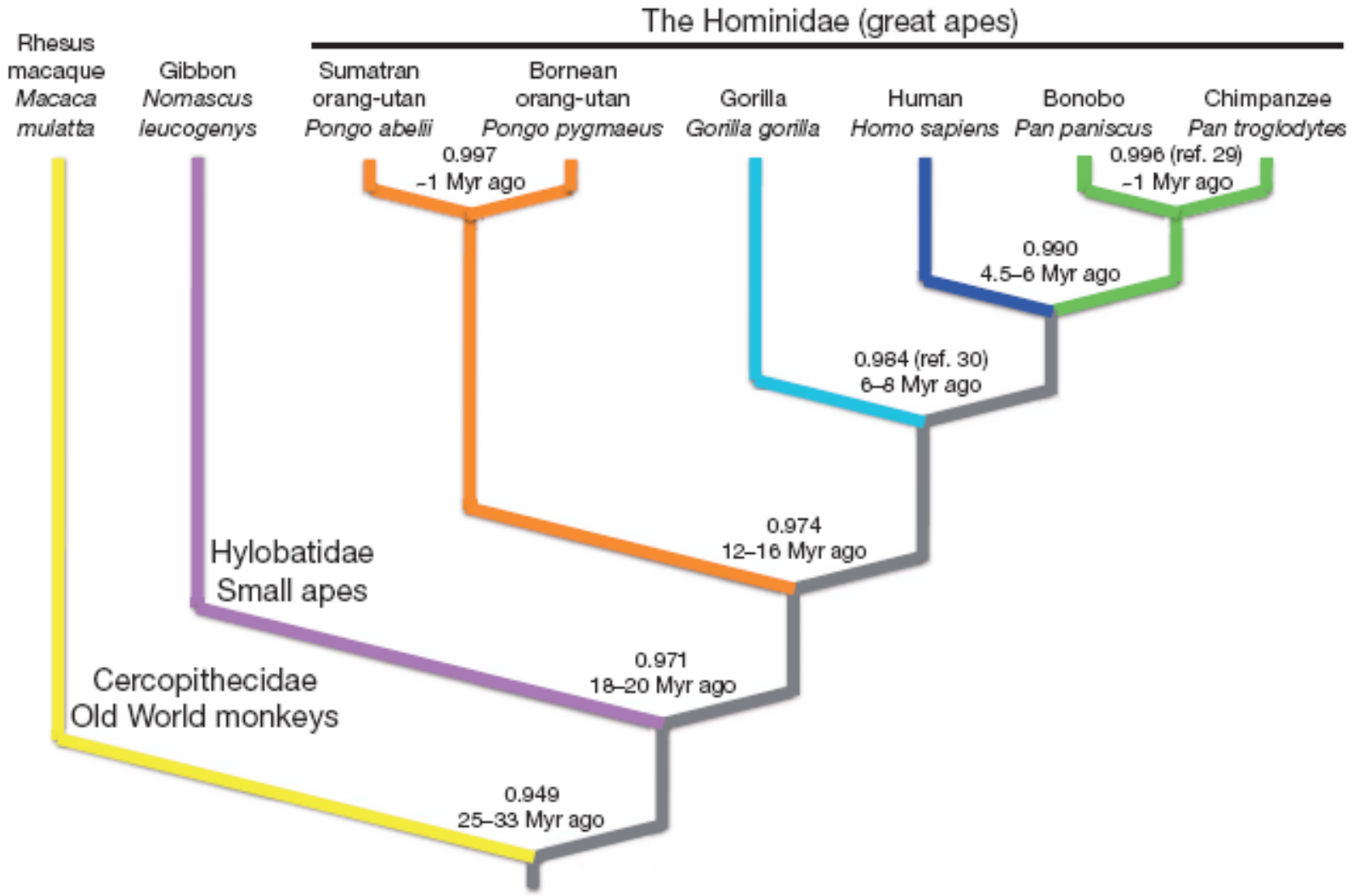


Pioneer of shotgun sequencing, personal genomics, synthetic biology.

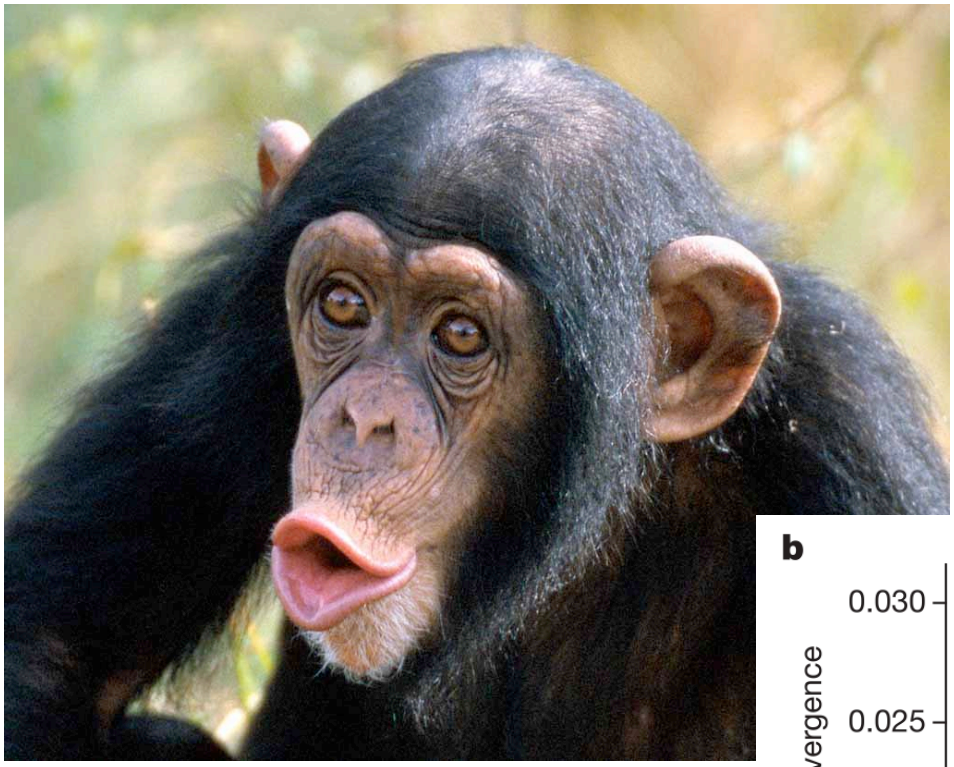

Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
About Us
View
Help
UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
chr21:33,031,597-33,041,570 9,974 bp.
enter position, gene symbol or search terms
go
chr21 (q22.11) 21p13 21p12 21p11.2 21q11.2 21q21.1 21q21.2 21q21.3 21q22.11 q22.12 q22.13 21q22.2 21q22.3
Scale
chr21:
2 kb
hg19
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
BC041449
SOD1
RefSeq Genes
Retroposed Genes V5, Including Pseudogenes
Publications: Sequences in scientific articles
Sequences
SNPs
Human mRNAs
Human mRNAs from GenBank
Spliced ESTs
Human ESTs That Have Been Spliced
Layered H3K27Ac
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
DNase Clusters
Digital DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE
Txn Factor ChIP
Transcription Factor ChIP-seq from ENCODE
100 Vert. Cons
100 vertebrates Basewise Conservation by PhyloP
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
Common SNPs(138)
Simple Nucleotide Polymorphisms (dbSNP 138) Found in >= 1% of Samples
RepeatMasker
Repeating Elements by RepeatMasker

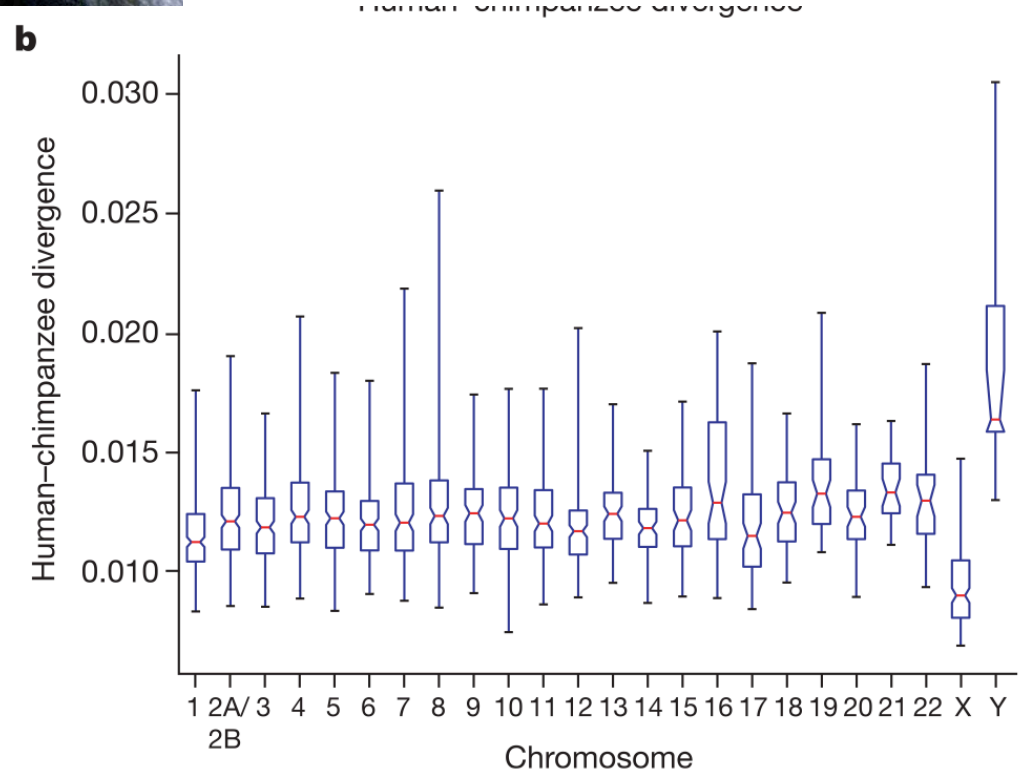

move <<< << < > >> >>> zoom in 1.5x
chr21:33,031,597-33,041,570 9,974 bp.
enter position, gene
chr21 (q22.11) 21p13 21p12 21p11.2 21q11.2 21q21.1
Scale
chr21:
2 kb
33,032,500 33,033,000 33,033,500 33,034,000 33,034,500 33,035,000 33,035,500 33,036,000 33,036,
UCSC Genes (RefSeq, GenBank, C
BC041449
SOD1
RefSeq Genes
Retroposed Genes
Publications: Sequ
Sequences
SNPs
Human m
Human mRNAs
Human ESTs
Spliced ESTs
100
Layered H3K27Ac
H3K27Ac Mark (Often Found Near Active
0
Digital DNaseI Hypersensitivit
DNase Clusters
Transcription F
Txn Factor ChIP
4.88
100 vertebrates Bas
100 Vert. Cons
0
-4.5
Multiz Alignm
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
Simple Nucleotide Polymorphis
Common SNPs(138)
Repeating El
RepeatMasker

Comparative Genomics

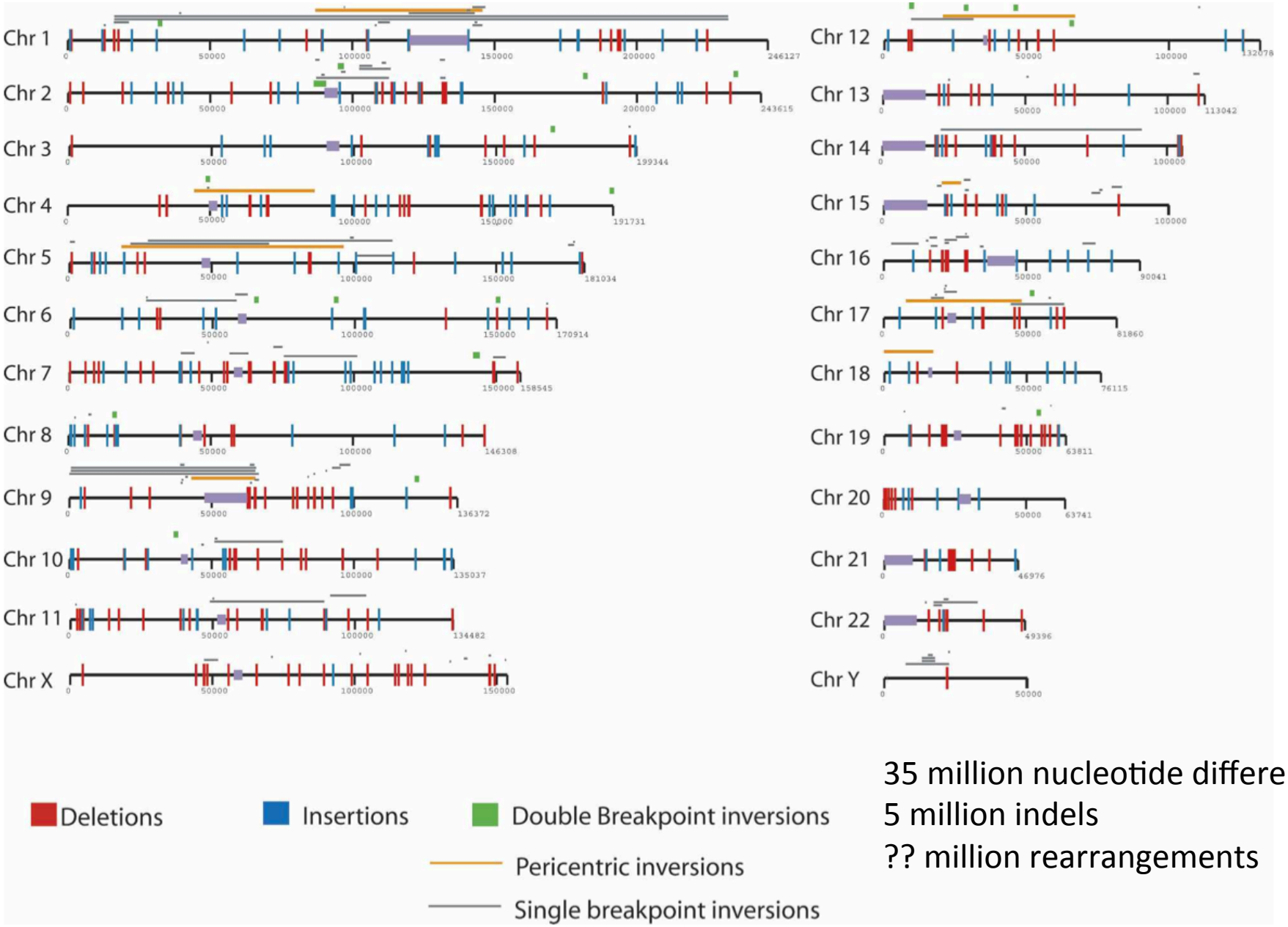


Locke et al. 2011




b
Human–chimpanzee divergence
0.030
0.025
0.020
0.015
0.010
1 2A/2B 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
Chromosome

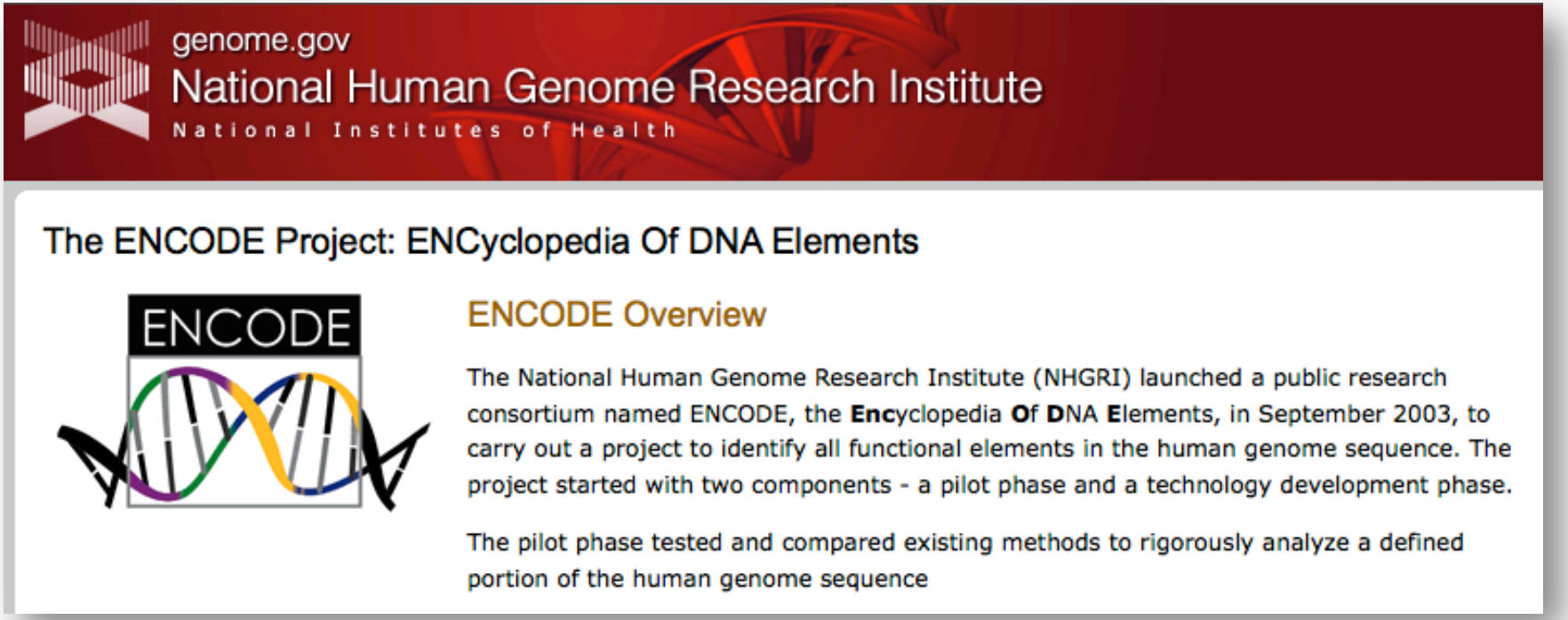
CSAC 2005

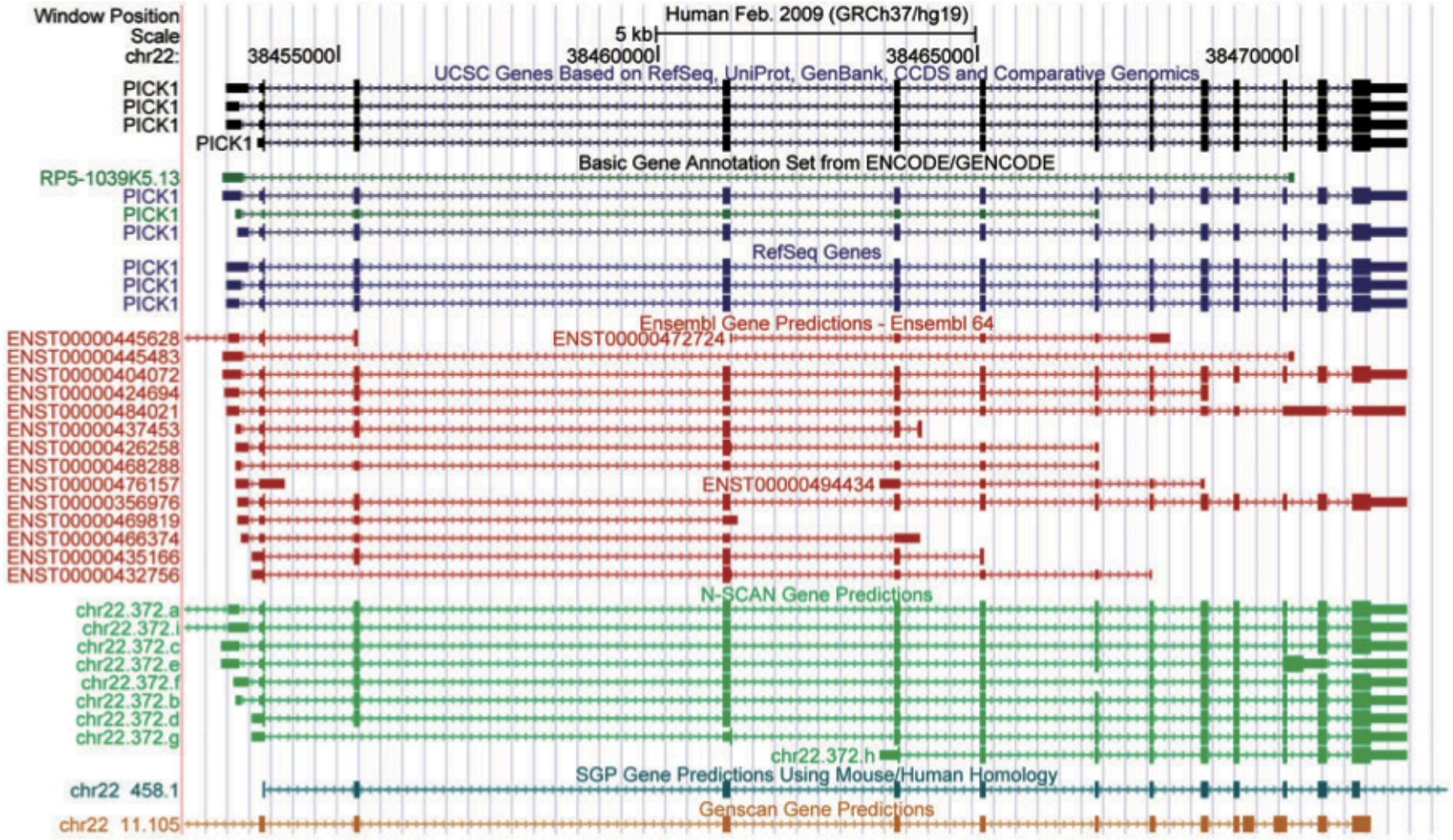


35 million nucleotide differences
5 million indels
?? million rearrangements

Perry et al. 2012

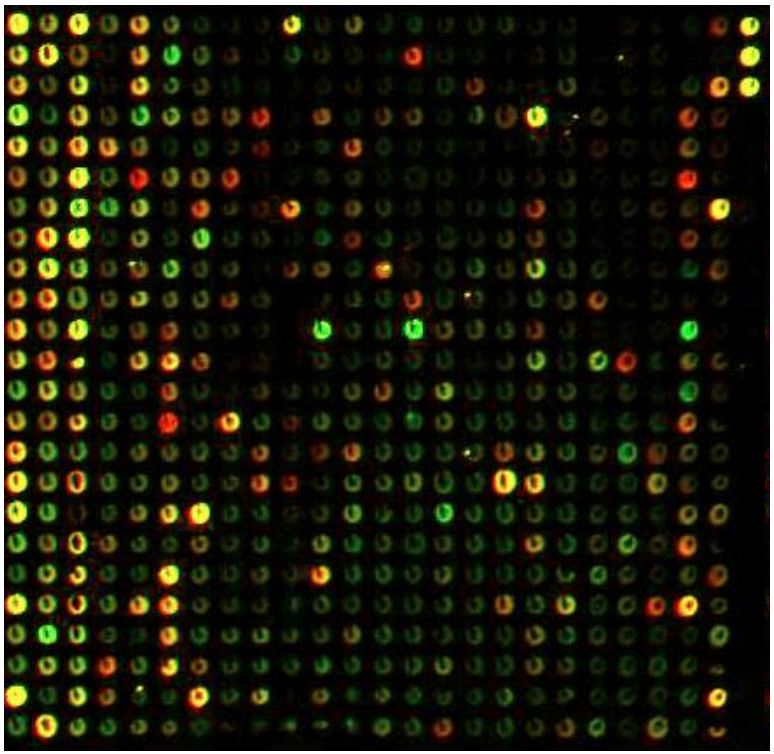
Many tracks in the UCSC Genome Browser come from data gathered as part of the ENCODE project




Window Position
Scale
chr22:
Human Feb. 2009 (GRCh37/hg19)
5 kb
38455000
38460000
38465000
38470000
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
PICK1
PICK1
PICK1
PICK1
Basic Gene Annotation Set from ENCODE/GENCODE
RP5-1039K5.13
PICK1
PICK1
PICK1
RefSeq Genes
PICK1
PICK1
PICK1
Ensembl Gene Predictions - Ensembl 64
ENST00000445628
ENST00000472724
ENST00000445483
ENST00000404072
ENST00000424694
ENST00000484021
ENST00000437453
ENST00000426258
ENST00000468288
ENST00000476157
ENST00000494434
ENST00000356976
ENST00000469819
ENST00000466374
ENST00000435166
ENST00000432756
N-SCAN Gene Predictions
chr22.372.a
chr22.372.i
chr22.372.c
chr22.372.e
chr22.372.f
chr22.372.b
chr22.372.d
chr22.372.g
chr22.372.h
SGP Gene Predictions Using Mouse/Human Homology
chr22 458.1
Genscan Gene Predictions
chr22 11.105

Non-sequence data are also visible in genome browsers

- Microarrays, first published in 1995
- Measure transcription (not REALLY gene expression) across the genome instead of by gene
- Now used for comparative genomics, DNA capture



Pat Brown. Stanford

Keep in mind….

- Gene duplication is common!
- Gene duplicates can be retained if
  - They are not too harmful, then they can be retained by *chance*
  - They are helpful, then they can be retained by *selection*
    - because more protein product is beneficial
    - because copies diversify (neo- or subfunctionalization)
- Pseudogenes are common in the genome
  - Can result from
    - the retrotransciption of an mRNA
    - a “defunctionalizing” mutation in a gene duplicate

Hands-on Exercise

- Using genome browsers to visualize data and develop hypotheses
- UCSC Genome Browser