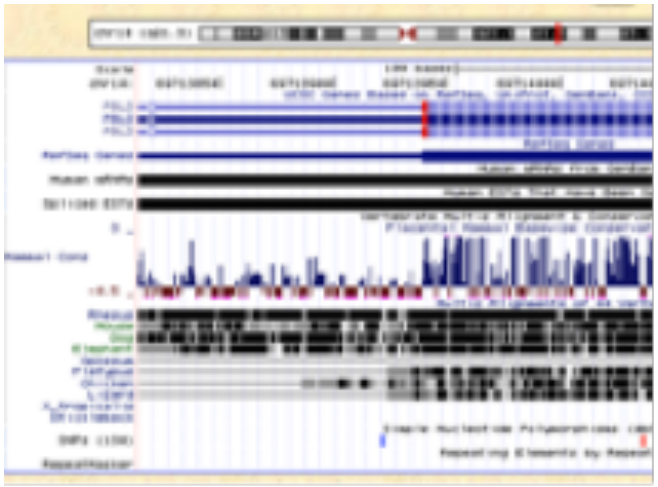
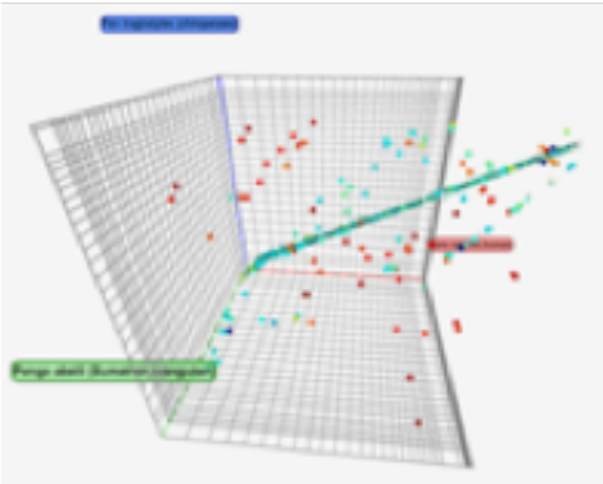
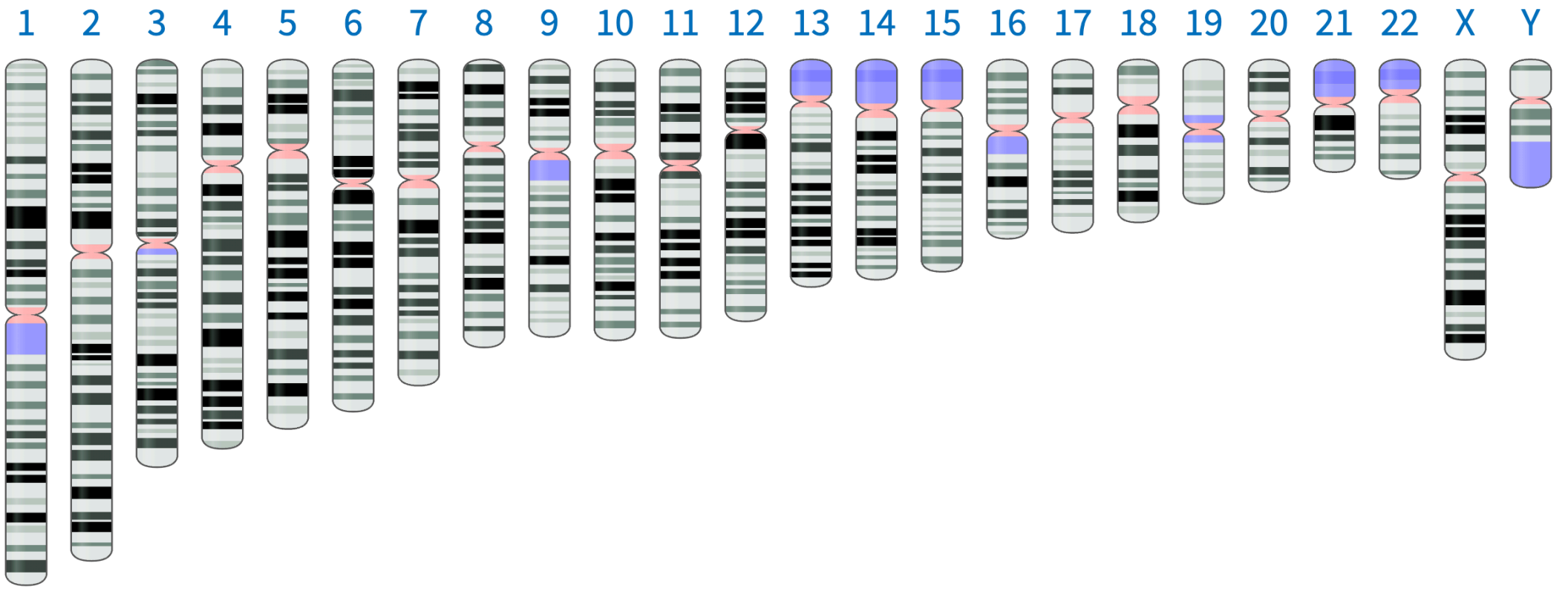
Computational Genomics

The Human Genome

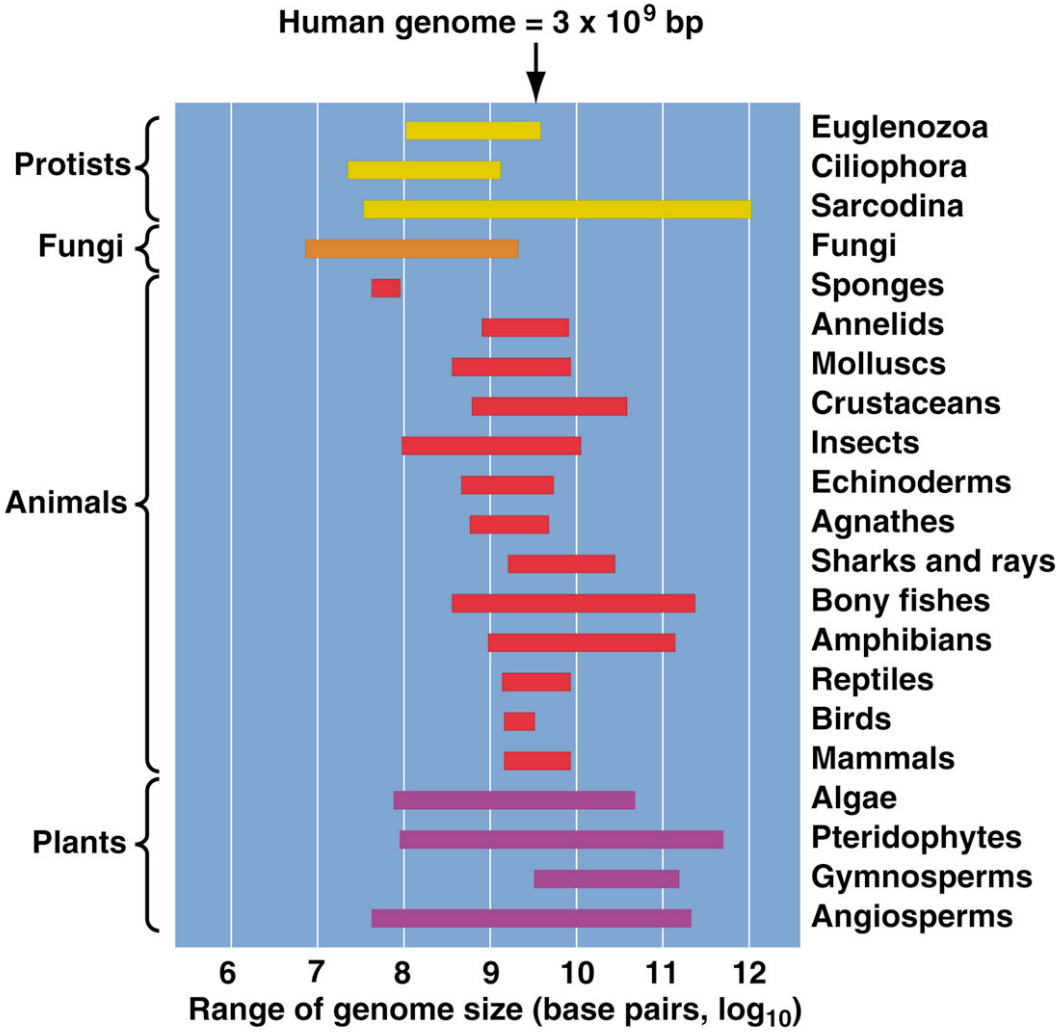




The Human Genome - Our Genome…

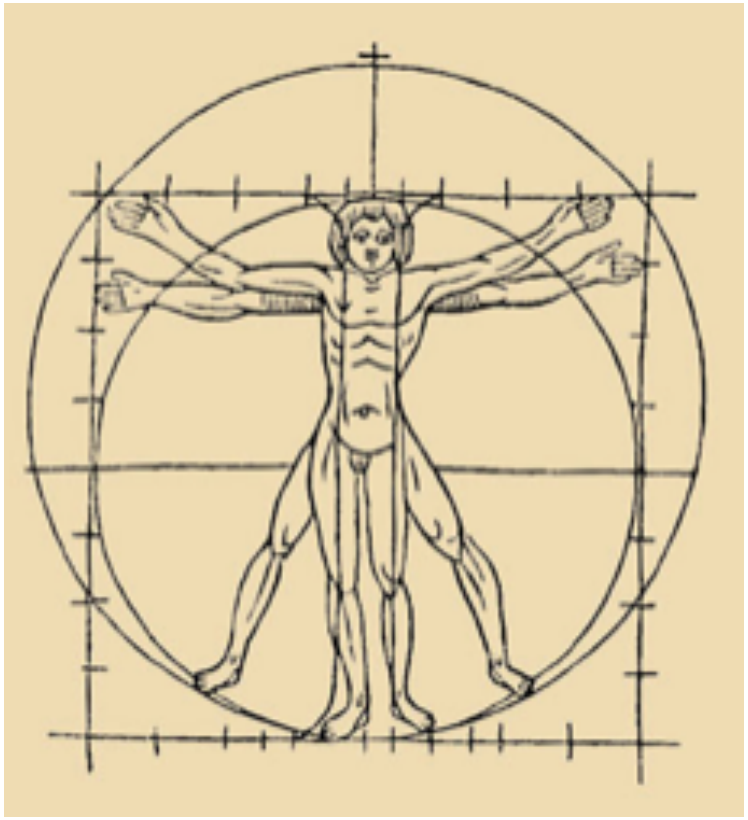


THE SIZE OF A GENOME DOES NOT REFLECT THE COMPLEXITY OF THE ORGANISM



The Human Genome

- So..How Large It Is?...



~3 x 10E9 nucleotides

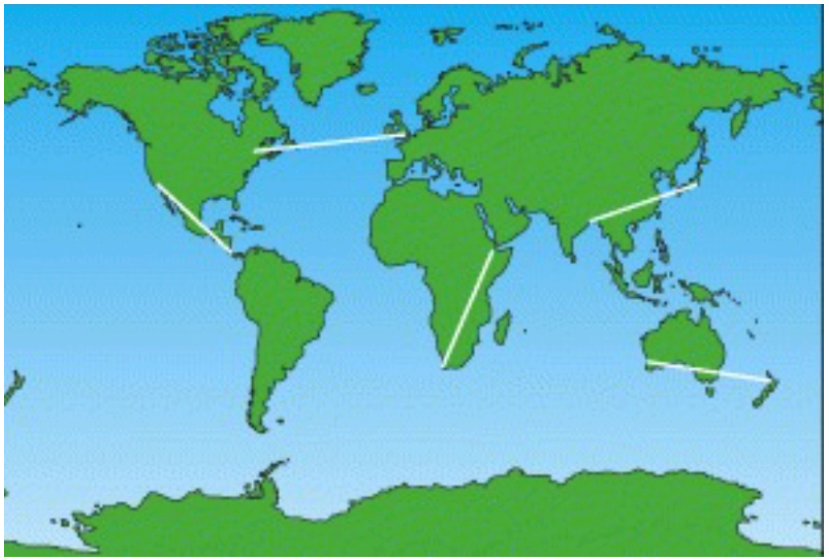
The Human Genome

- So..How Large It Is?...

The Human Genome

- So..How Large It Is?...

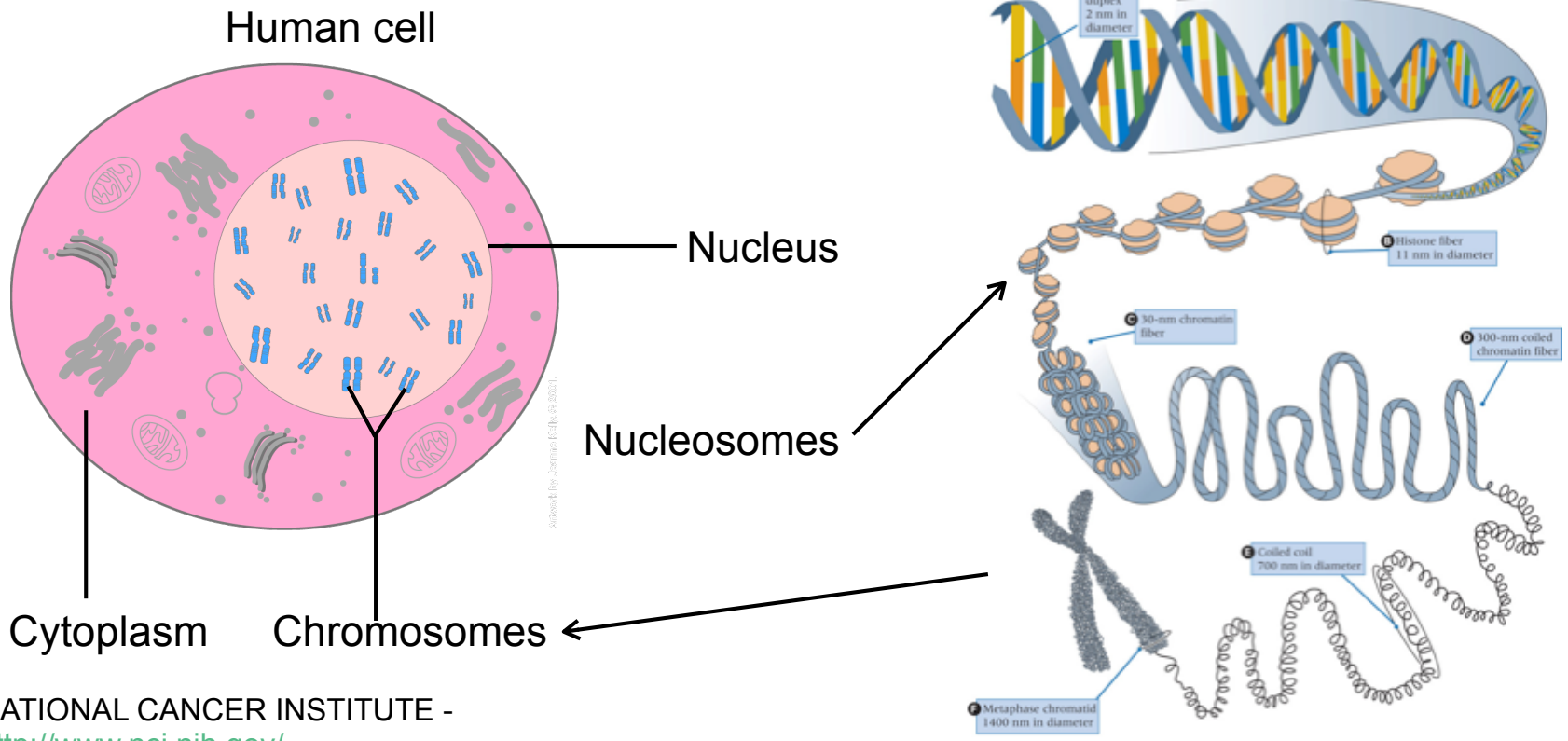
If we were to write 60 nt of DNA in a line
10 cm in length, in this format, the
human genome sequence would stretch for 5000 km,
the distance from Montreal to London, Los Angeles to
Panama, Tokyo to Calcutta, Cape Town to Addis Ababa,
or Auckland to Perth



https://www.ncbi.nlm.nih.gov/books/NBK21134/

The Human Genome

- How is it organized?

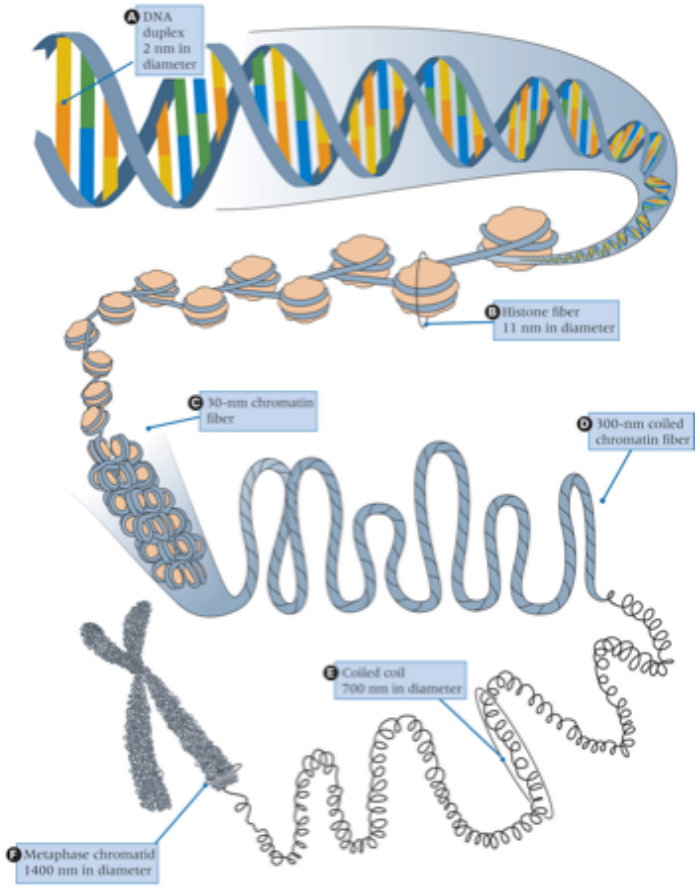


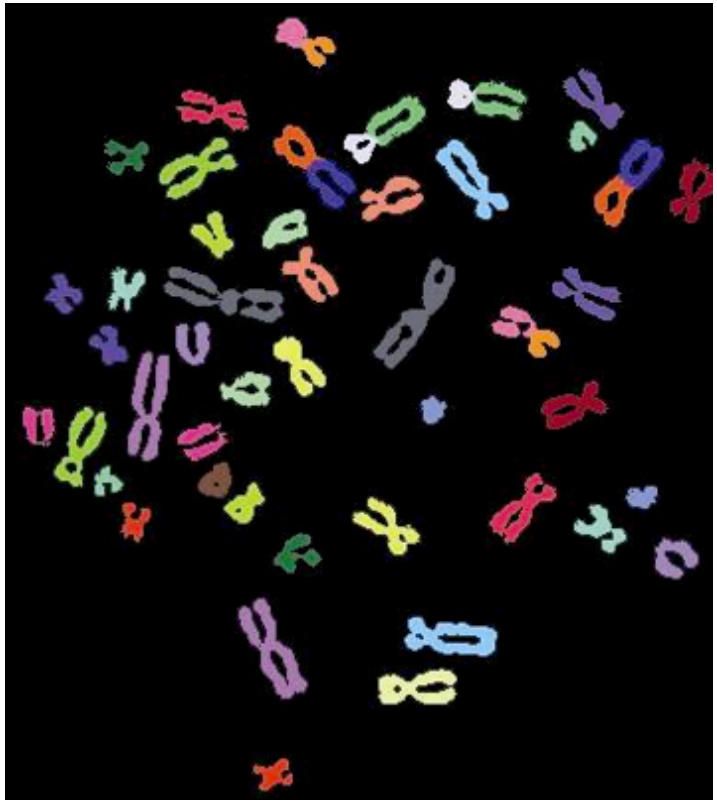
NATIONAL CANCER INSTITUTE - http://www.nci.nih.gov/cancertopics/understandingcancer

GENETICS - Daniel L. Hartl and Elizabeth W. Jones. 6th Edition

The Human Genome

- How is it organized?

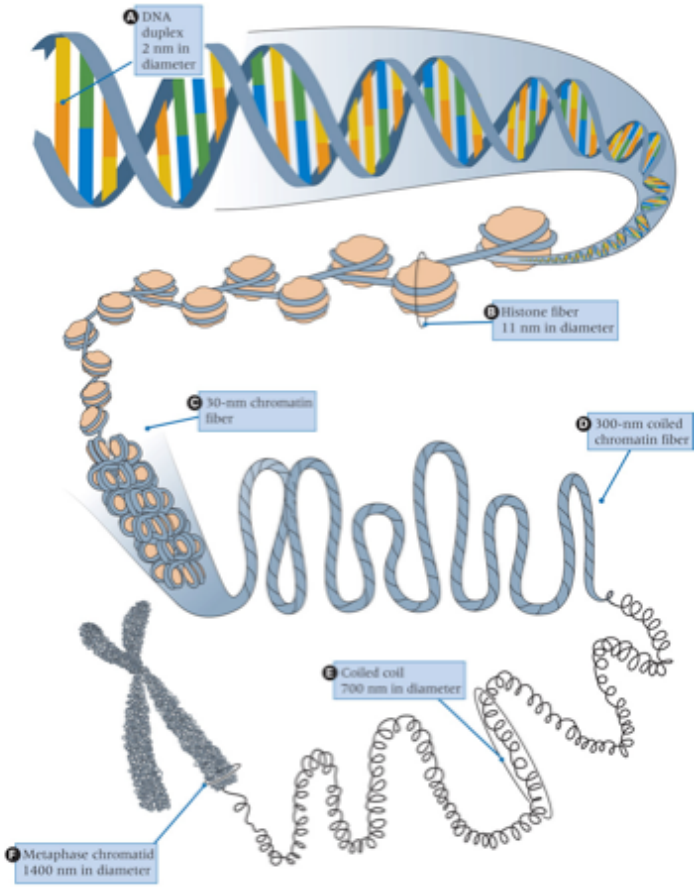


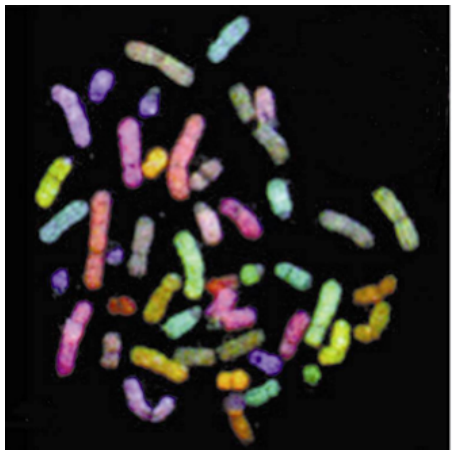


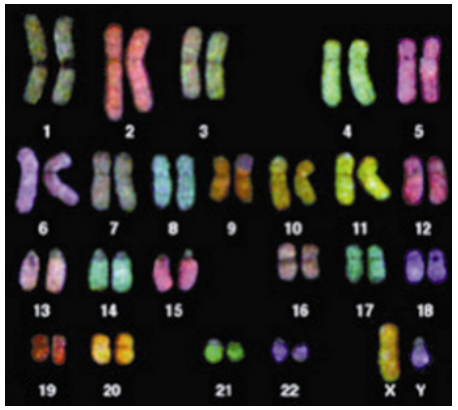
GENETICS - Daniel L. Hartl and Elizabeth W. Jones. 6th Edition

The Human Genome

- How is it organized?



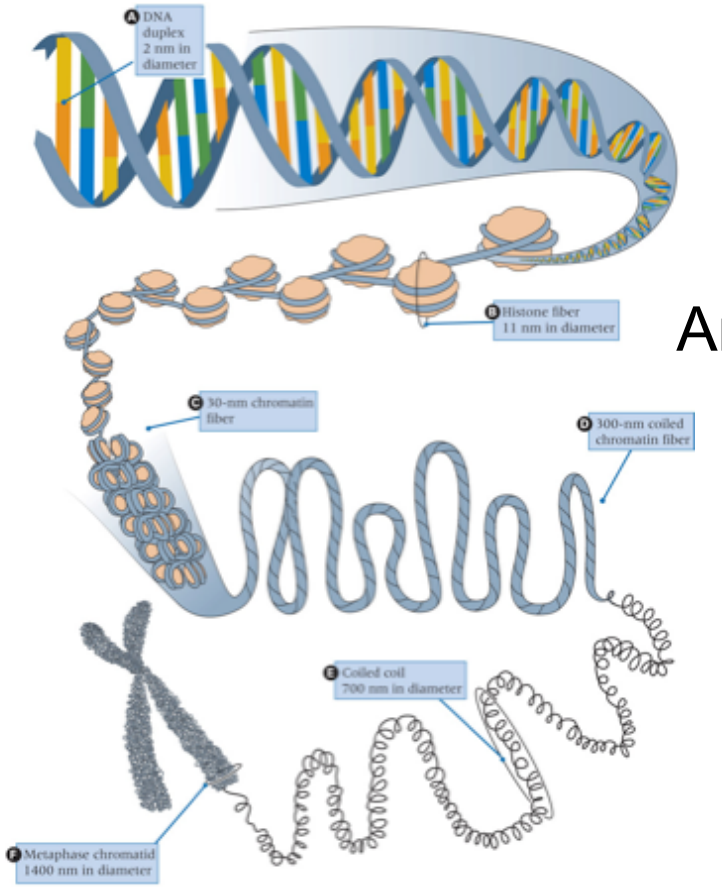




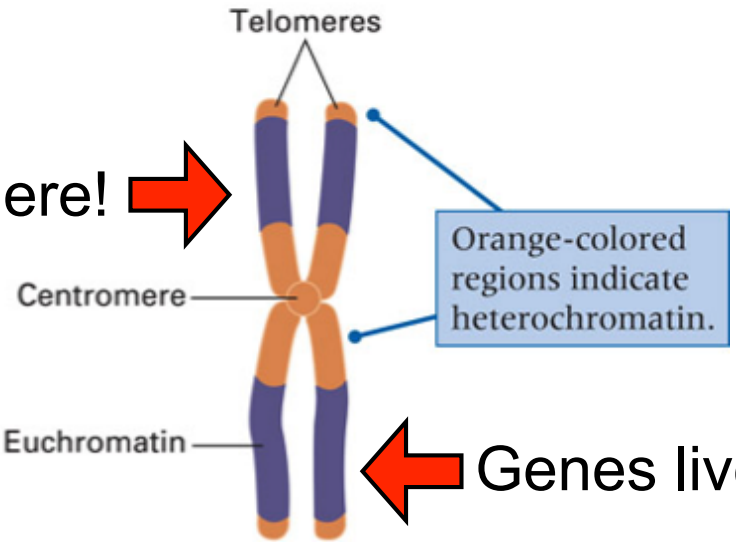
GENETICS - Daniel L. Hartl and Elizabeth W. Jones. 6th Edition

The Human Genome

- How is it organized?



And here!

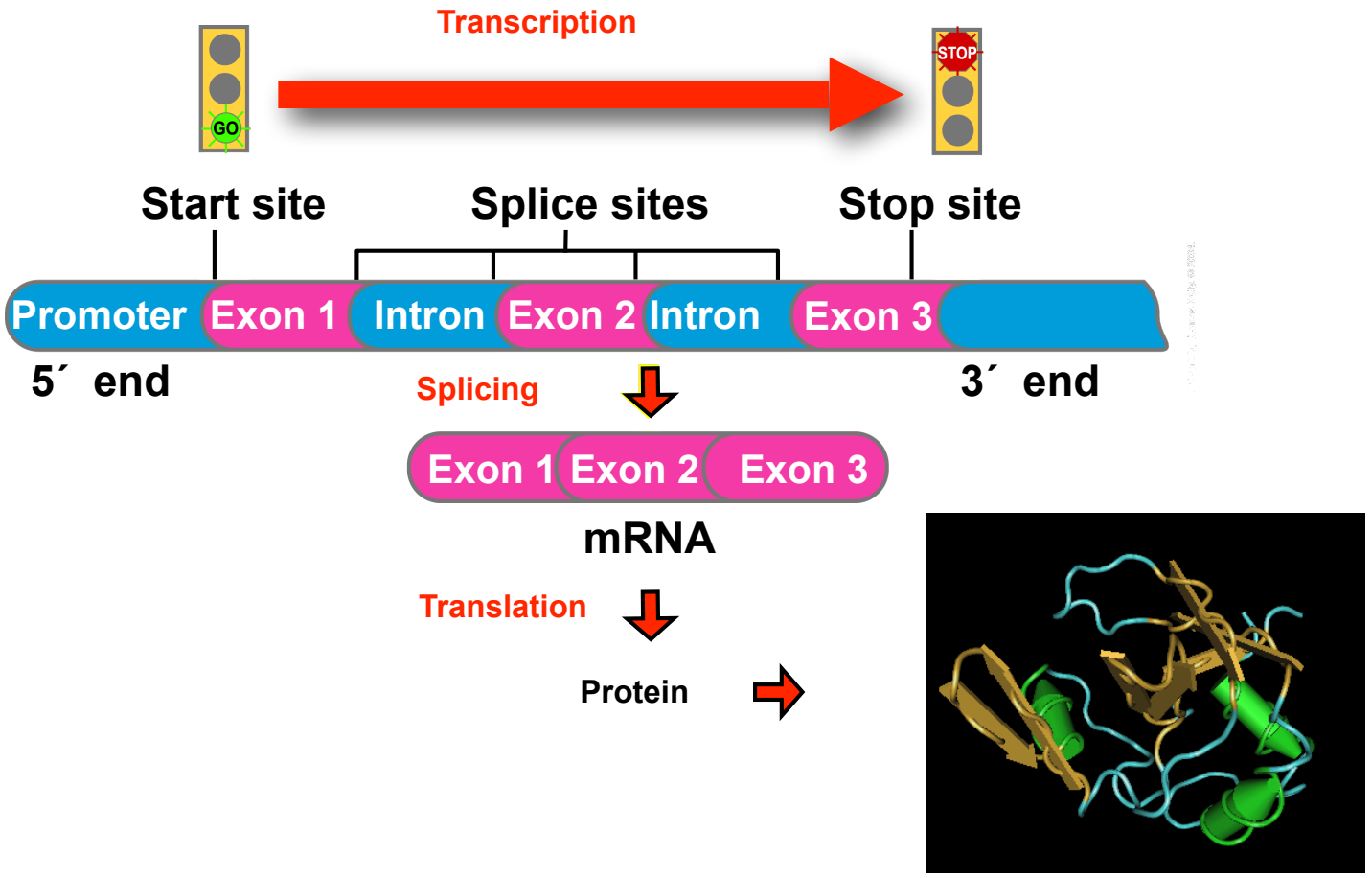


Genes live here!

GENETICS - Daniel L. Hartl and Elizabeth W. Jones. 6th Edition

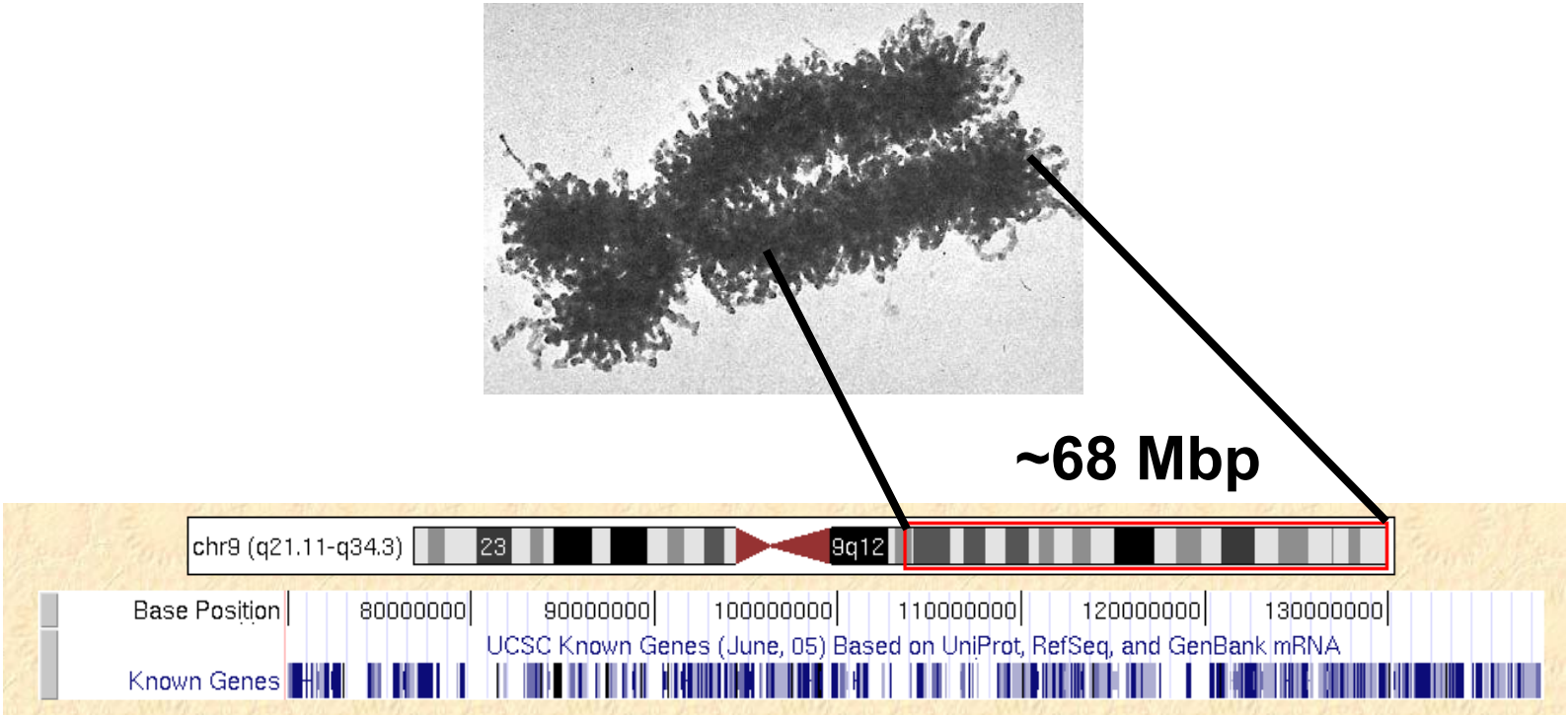
The Human Genome

- What is a (Protein-Coding) Gene?



The Human Genome

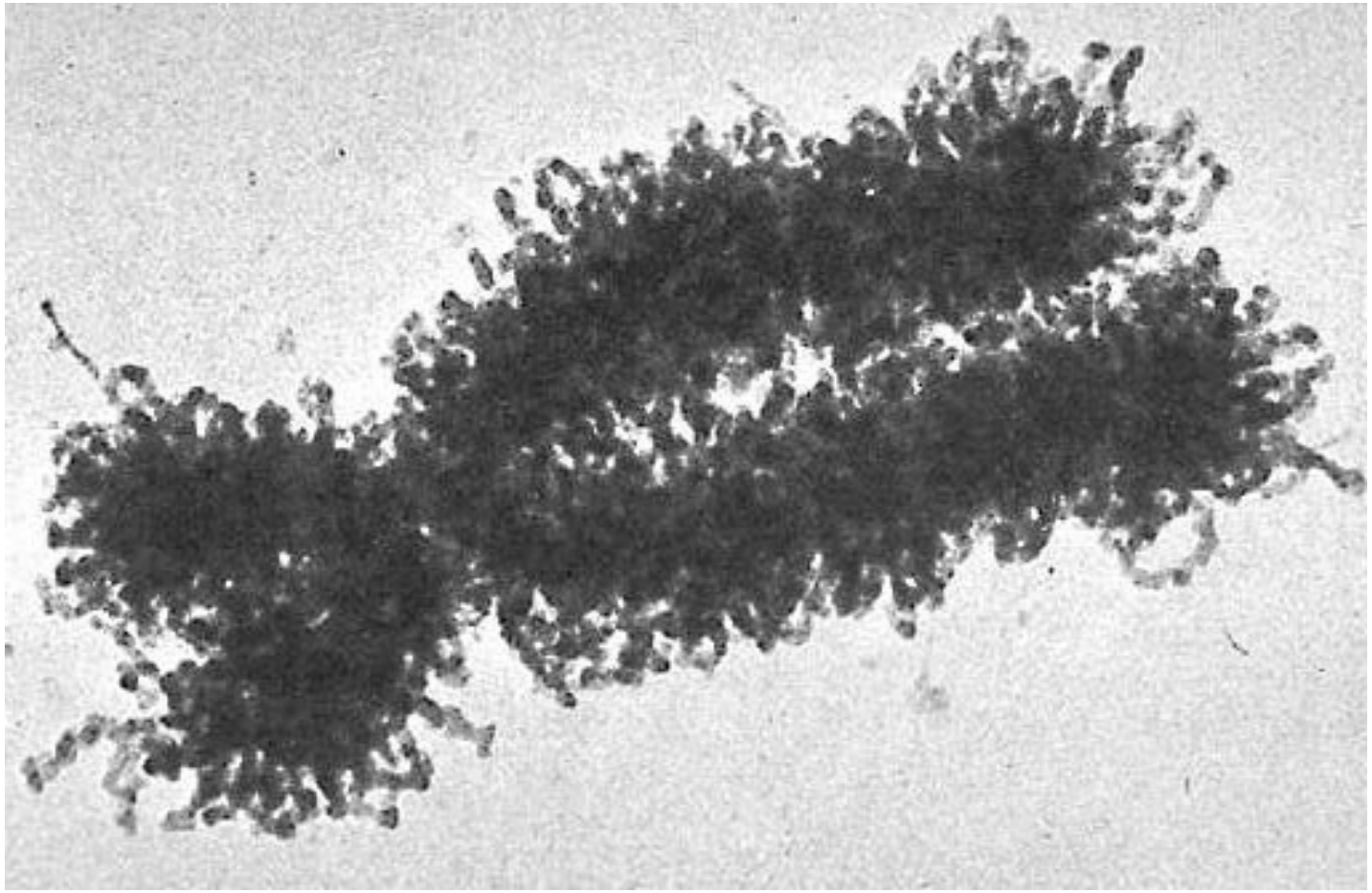
- How are Genes Distributed? Where do they live?



UCSC Genome Browser with VISTA tracks on Human May 2004 Assembly

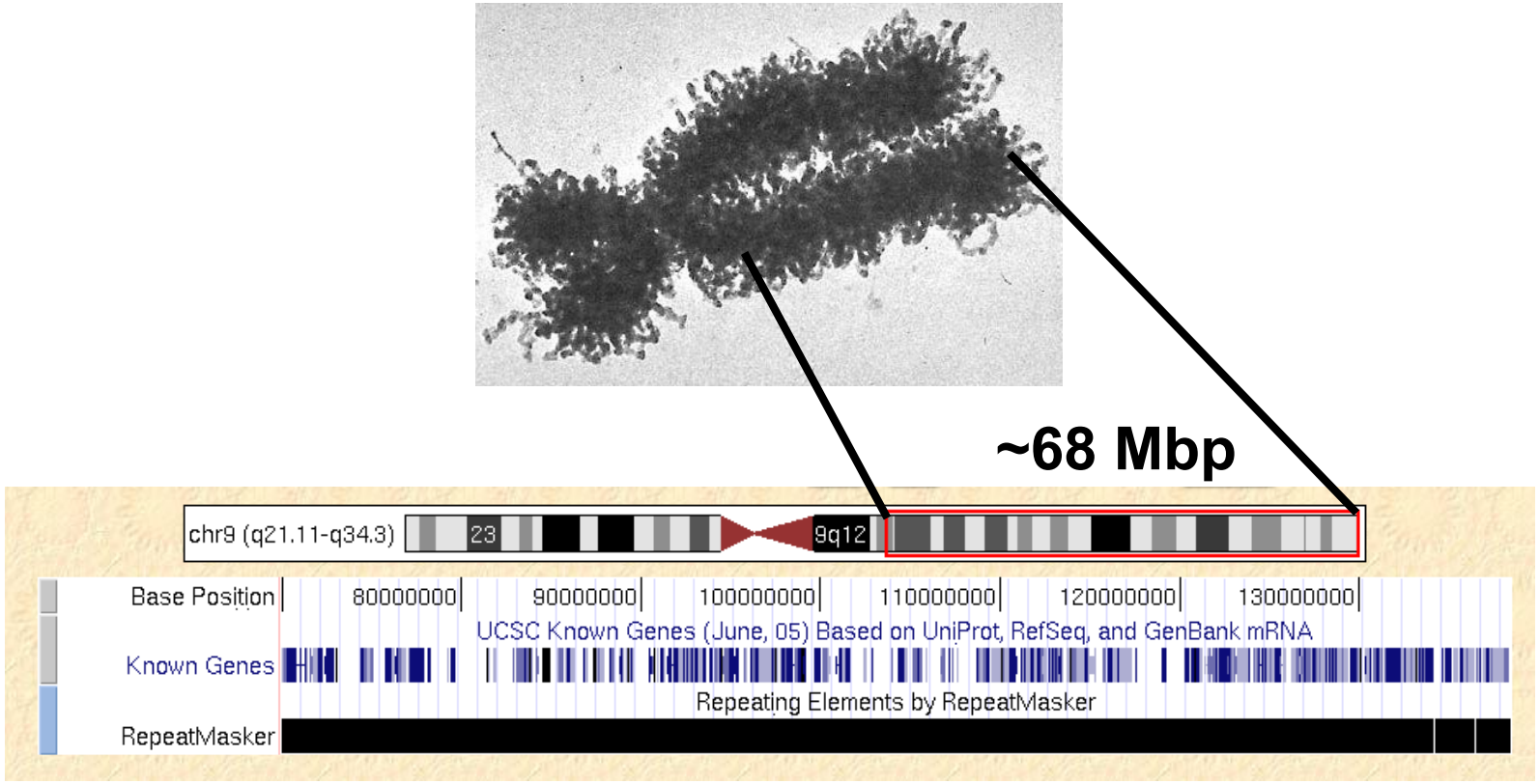
The Human Genome

- The Human Genome contains Tons of Repeated Elements!



The Human Genome

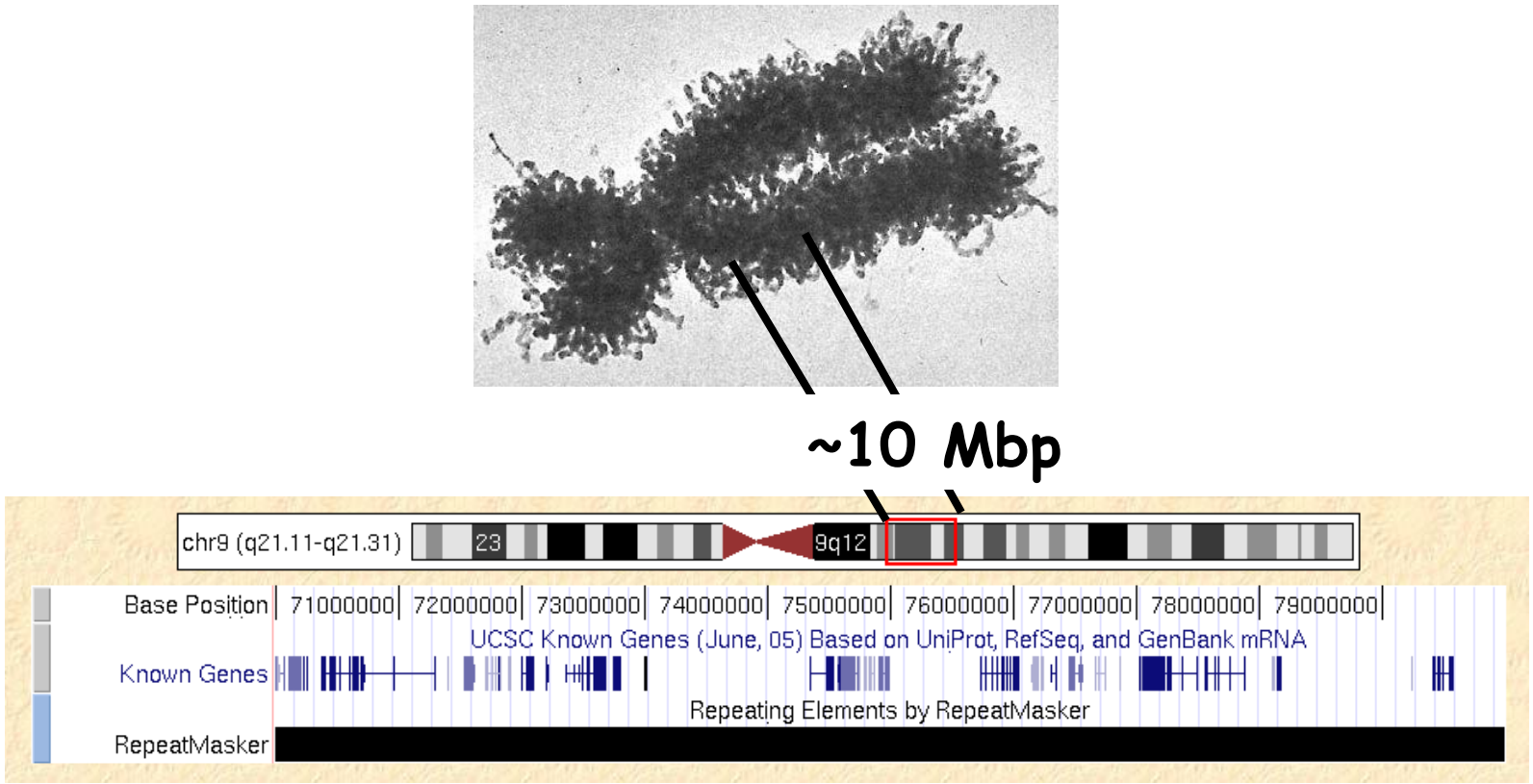
- The Human Genome is Full of Repeats!



UCSC Genome Browser with VISTA tracks on Human May 2004 Assembly

The Human Genome

- The Human Genome is Full of Repeats!

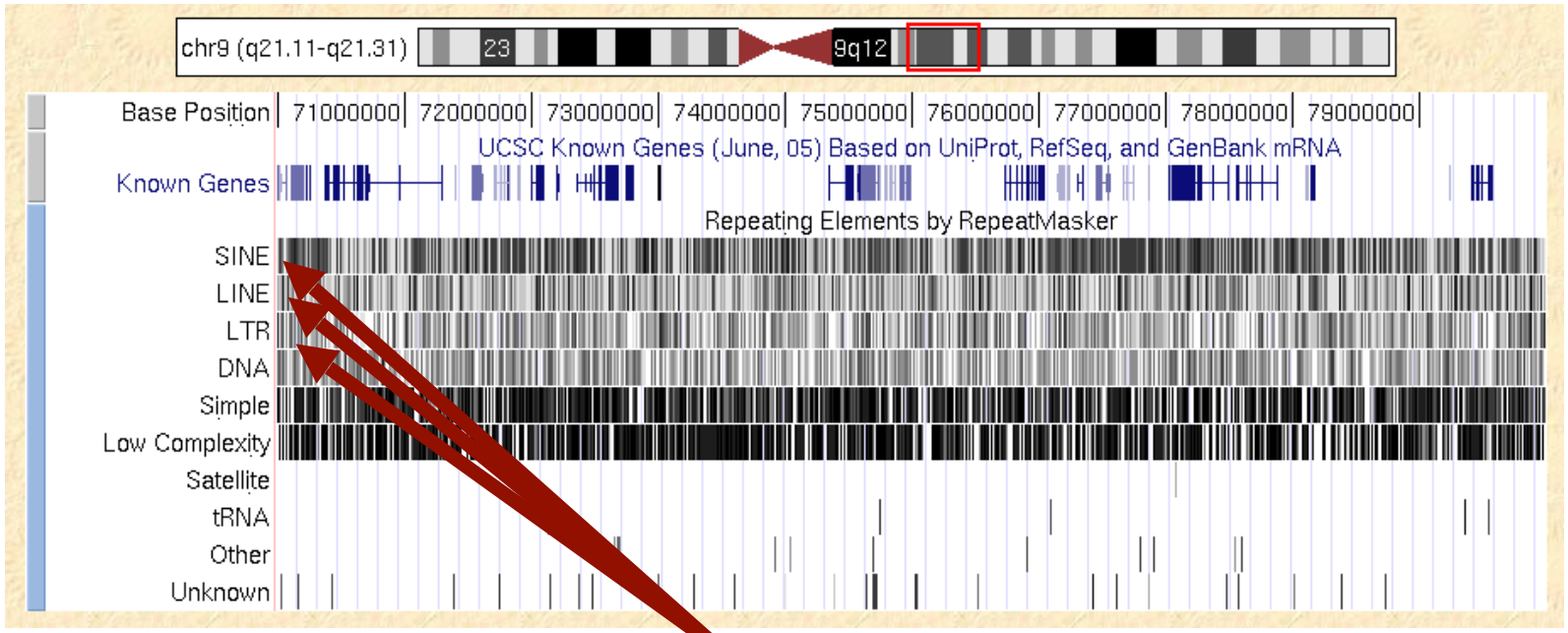


UCSC Genome Browser with VISTA tracks on Human May 2004 Assembly

The Human Genome

- The Human Genome is Full of Repetitive Elements!

~10 Mbp

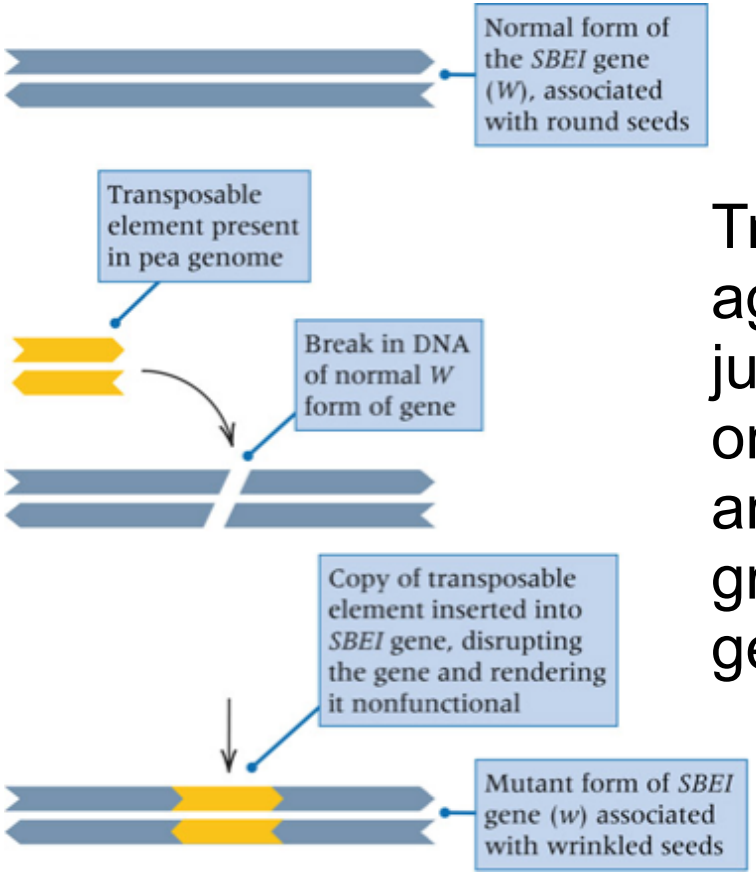


Transposable Elements!

UCSC Genome Browser with VISTA tracks on Human May 2004 Assembly

The Human Genome

- What are Transposable Elements?

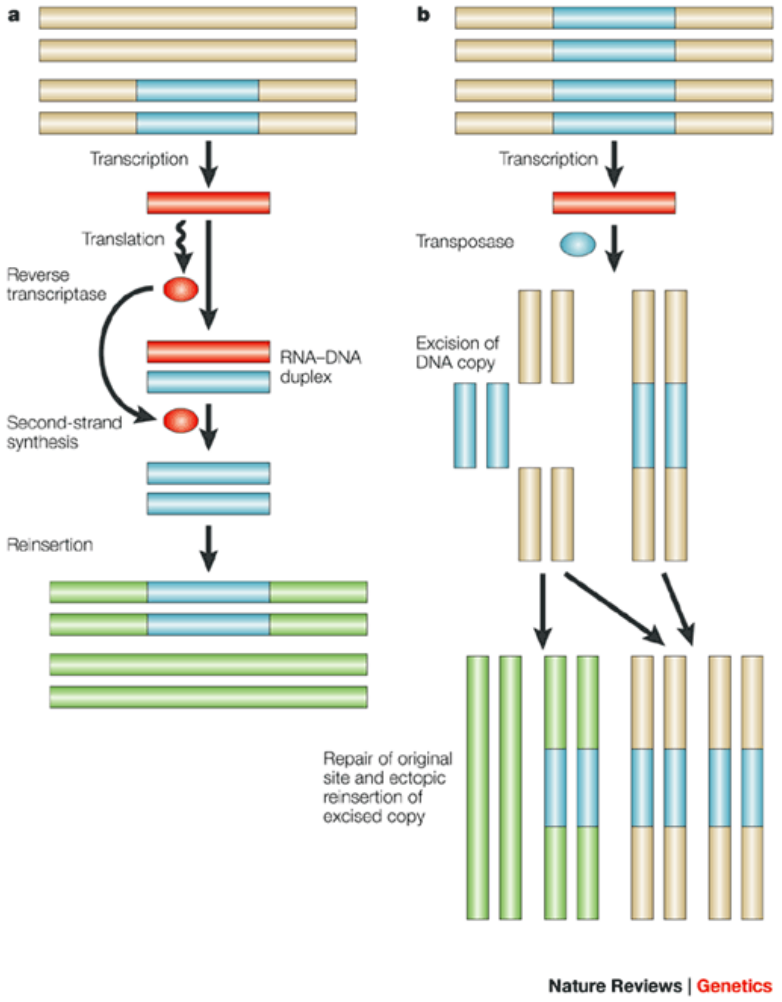


Transposable infectious agents, mobile parasites, that jump (i.e., transpose) from one place of the genome to another, sometimes with grave consequences to the genome!

GENETICS - Daniel L. Hartl and Elizabeth W. Jones. 6th Edition

The Human Genome

- How do they Work?

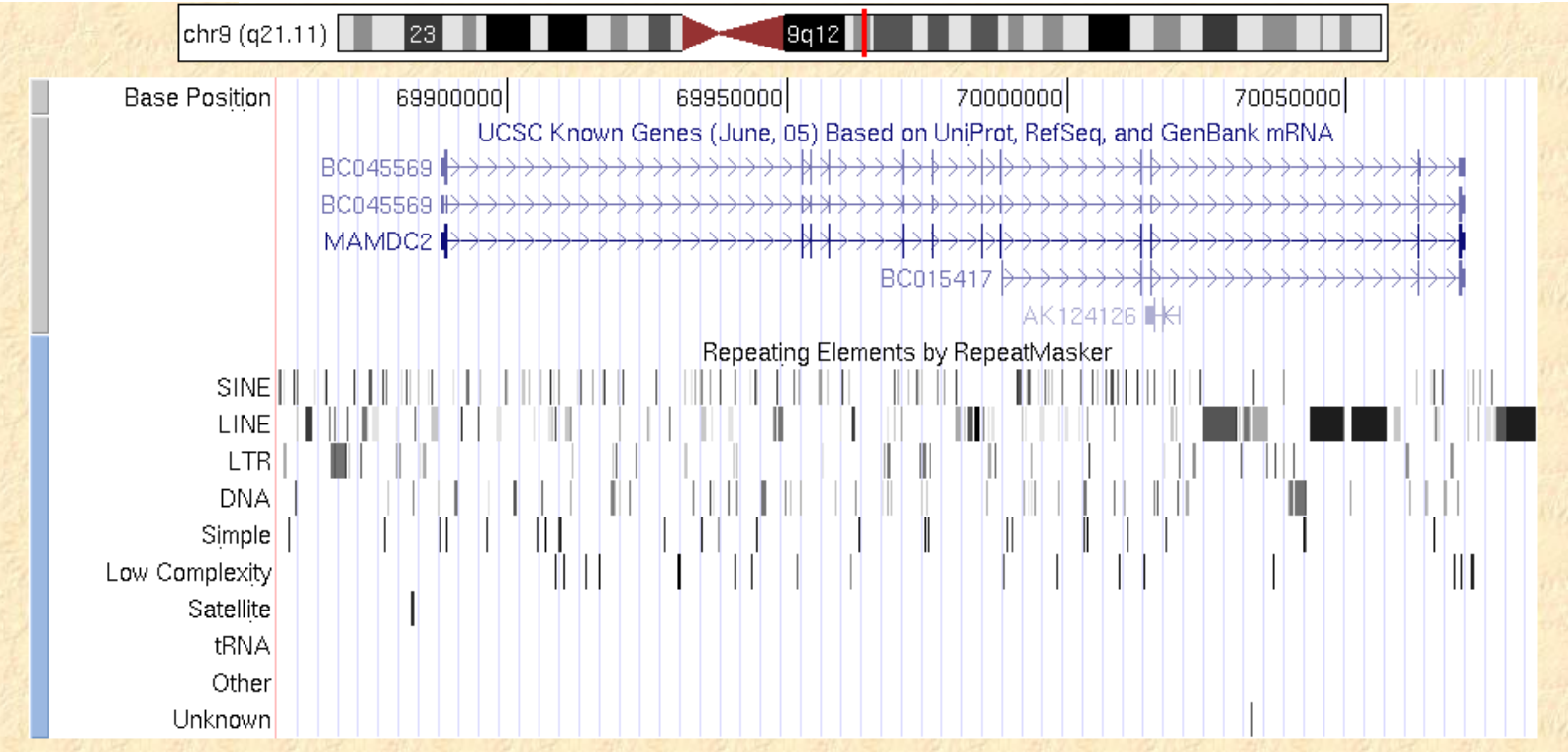


NATURE REVIEWS GENETICS - Gregory D. D. Hurst and John H. Werren 2, 597-606 (August 2001)

Transposons and Us

- Selfish Genetic Elements Live within our own Genes!

Zooming in ~45,000X to ~225 kbp

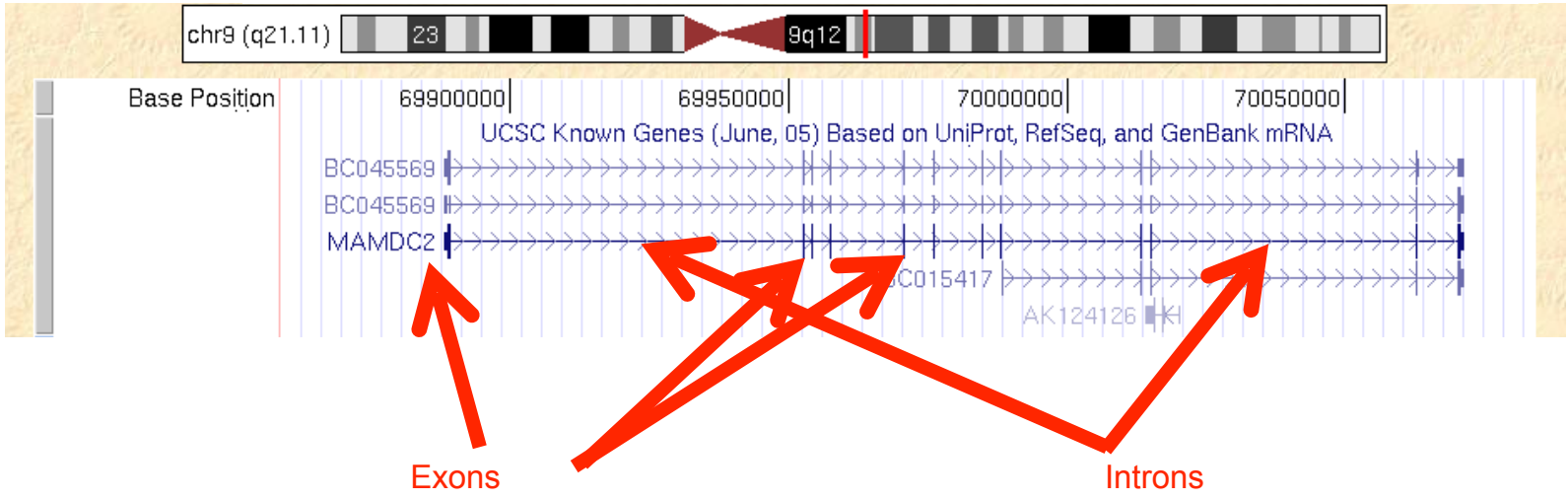


UCSC Genome Browser with VISTA tracks on Human May 2004 Assembly

Transposons and Us

- Selfish Genetic Elements Live within our own Genes!

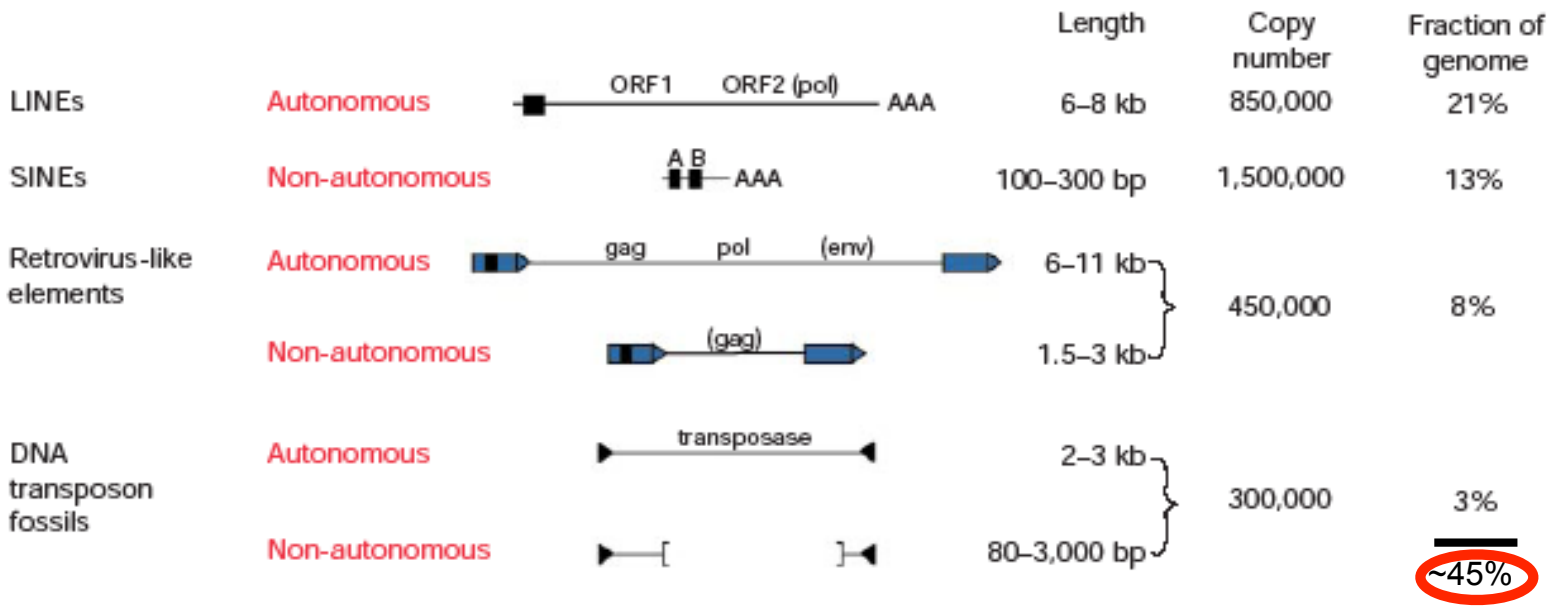
Zooming in ~45,000X to ~225 kbp



UCSC Genome Browser with VISTA tracks on Human May 2004 Assembly

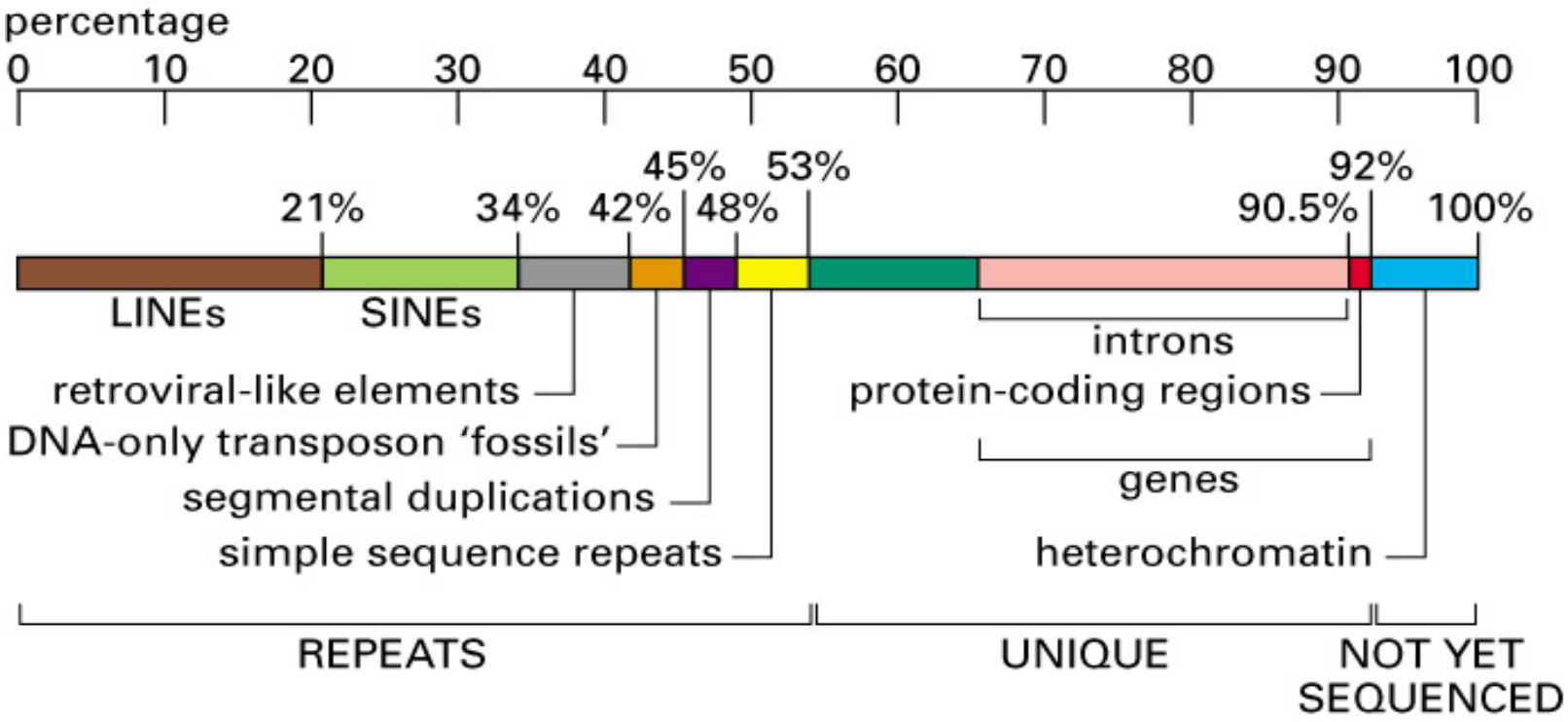
Transposons and Us

- Classes of Interspersed Repeat in Humans



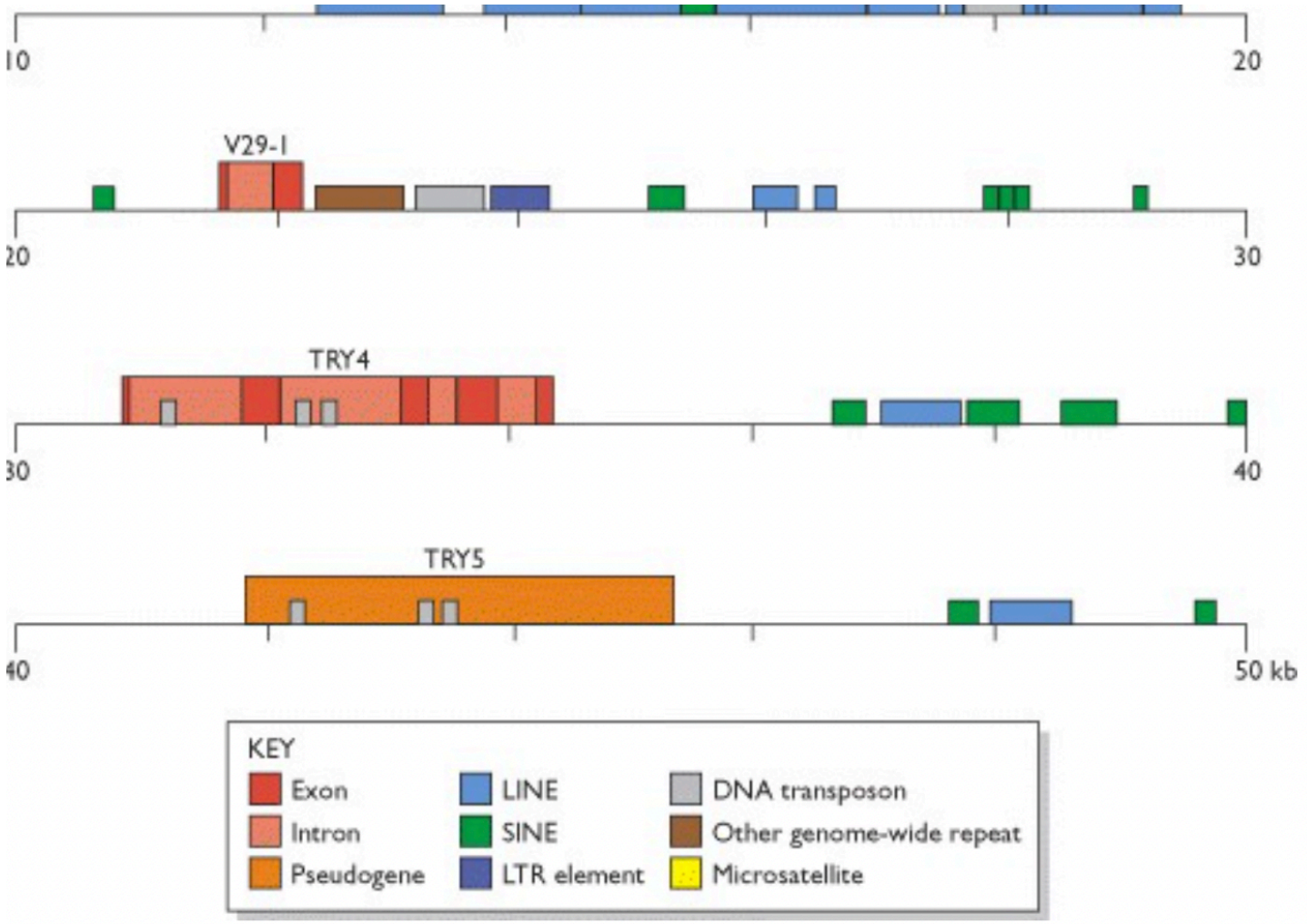
Nature 409, 860921 (15 February 2001)

The Human Genome is a transposon Traffic Jam



The Human Genome

- The Human Genome is Complex!

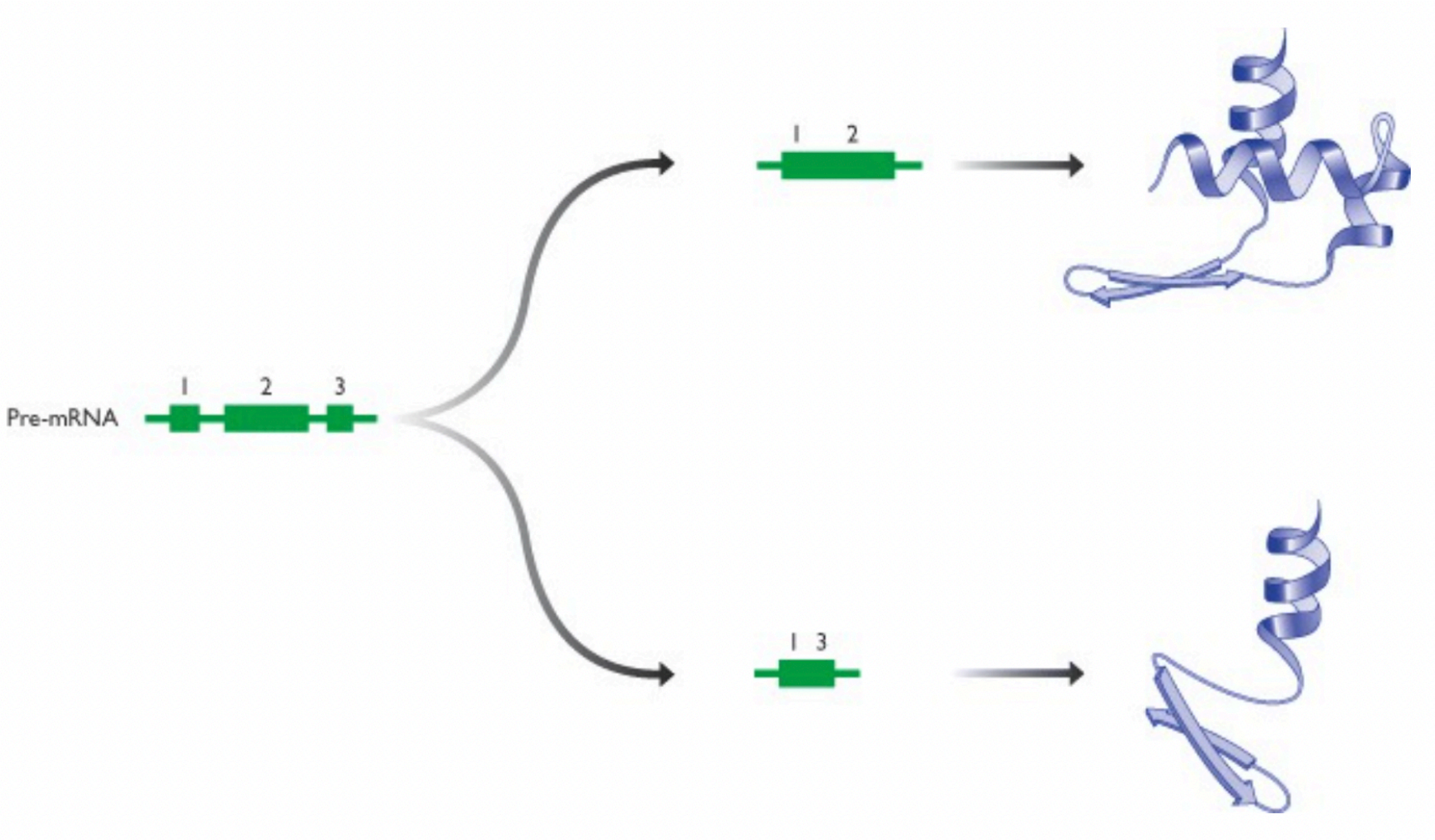


This map shows the location of genes, gene segments, pseudogenes, **genome**-wide repeats and microsatellites in a 50-kb segment of the human β T-cell receptor locus on chromosome 7. Redrawn from Rowen *et al*. (1996).

https://www.ncbi.nlm.nih.gov/books/NBK21134/

The Human Genome

- The Human Genome is Complex!

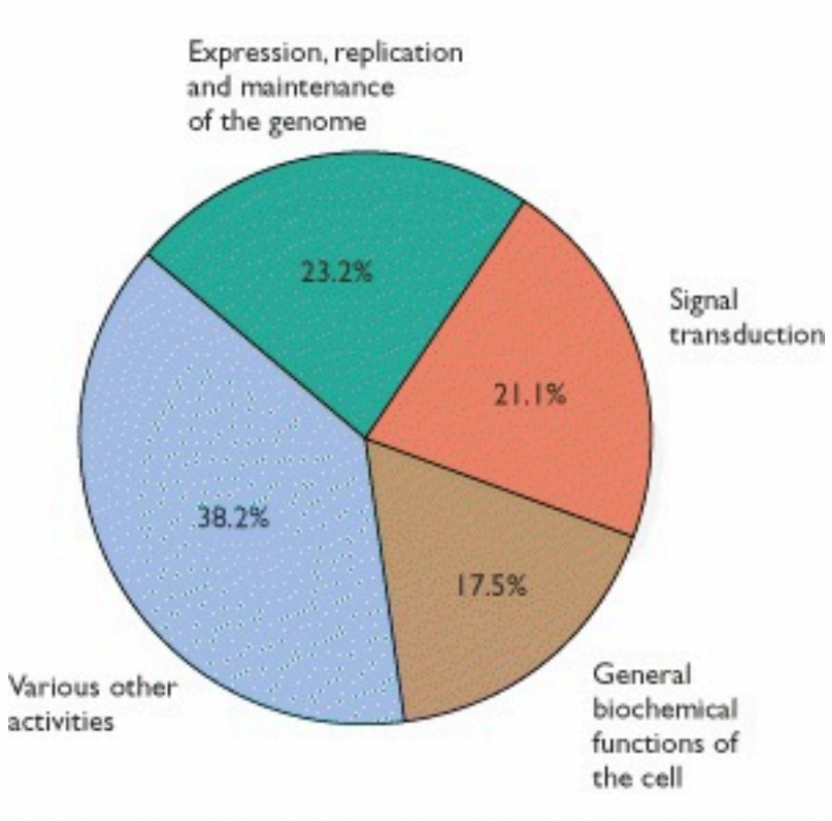


Alternative splicing results in different combinations of exons becoming linked together, resulting in different proteins being synthesized from the same pre-mRNA.

https://www.ncbi.nlm.nih.gov/books/NBK21134/

The Human Genome

- The Human Genome is Complex!



The pie chart shows a categorization of the identified human protein-coding genes. It omits approximately 13 000 genes whose functions are not yet known. The segment labeled ‘various other activities’ includes, among others, proteins involved in biochemical transport processes and protein folding, immunological proteins, and structural proteins. Based on Figure 15 of Venter et al. (2001).

https://www.ncbi.nlm.nih.gov/books/NBK21134/

The Human Genome

- The Human Genome is Loaded with Evolutionary Relics!

Conventional Pseudogenes

A conventional pseudogene is a gene that has been inactivated because its nucleotide sequence has changed by mutation.

Many mutations have only minor effects on the activity of a gene but some are more important and it quite possible for a single nucleotide change to result in a gene becoming completely non-functional.

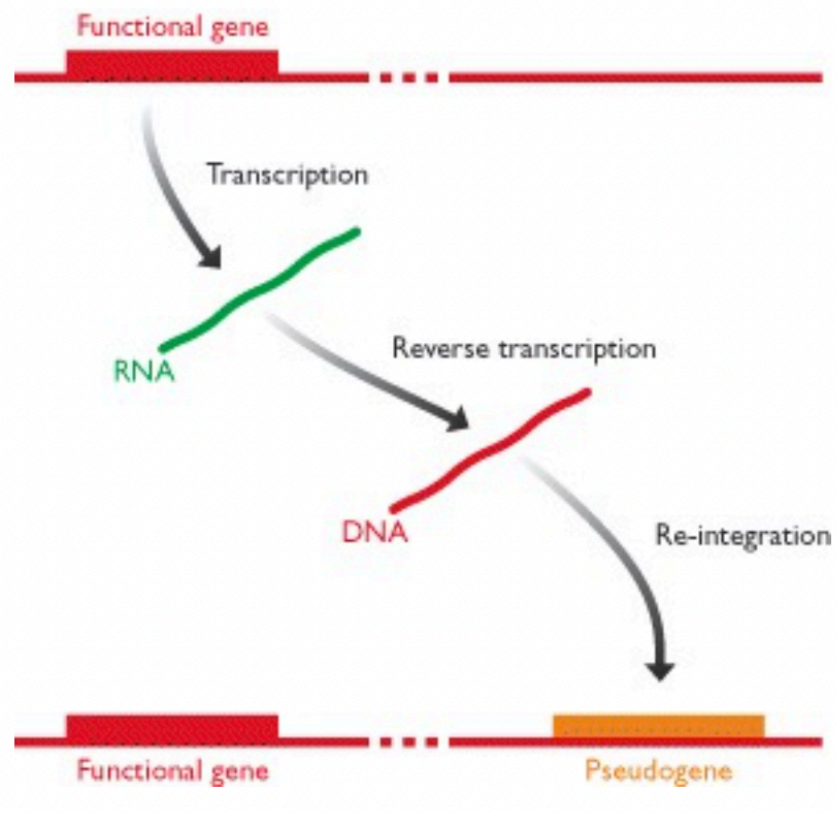
Once a pseudogene has become non-functional it will degrade through accumulation of more mutations and eventually will no longer be recognizable as a gene relic.

TRY5 is an example of a conventional pseudogene

https://www.ncbi.nlm.nih.gov/books/NBK21134/

The Human Genome

- The Human Genome is Loaded with Evolutionary Relics!



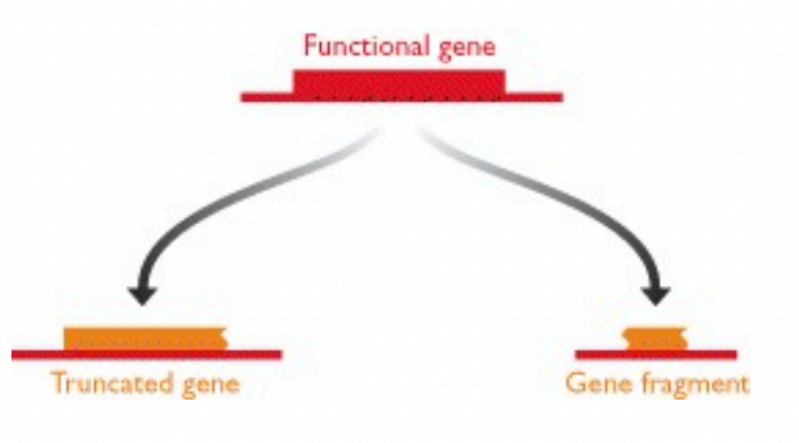
Processed Pseudogenes

A processed pseudogene is thought to arise by integration into the genome of a copy of the mRNA transcribed from a functional gene. The process by which mRNA is copied into DNA is called reverse transcription and the product is called complementary DNA (cDNA). The cDNA may integrate into the same chromosome as its functional parent, or possibly into a different chromosome.

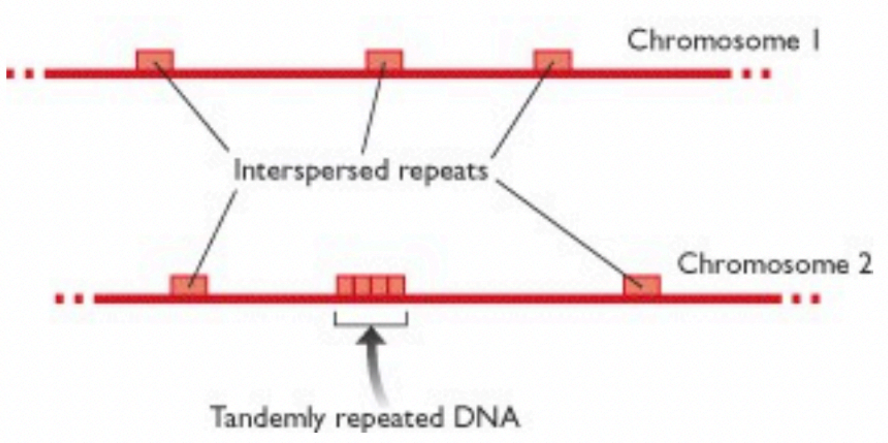
https://www.ncbi.nlm.nih.gov/books/NBK21134/

The Human Genome

- The Human Genome is Loaded with Evolutionary Relics!



Truncated Genes

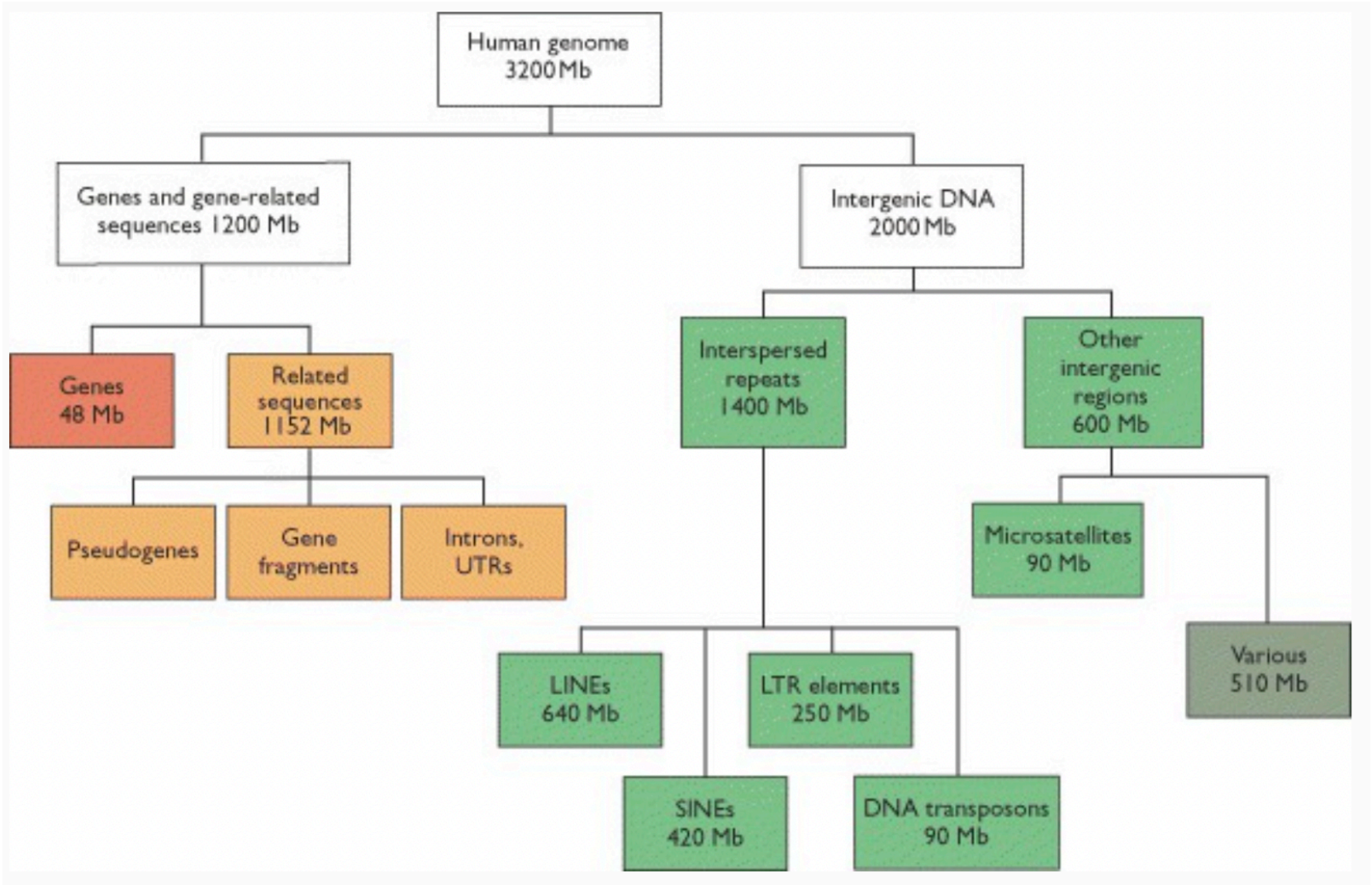


The two types of repetitive DNA: interspersed repeats and tandemly repeated DNA

https://www.ncbi.nlm.nih.gov/books/NBK21134/

The Human Genome

- The Organization of the Human Genome

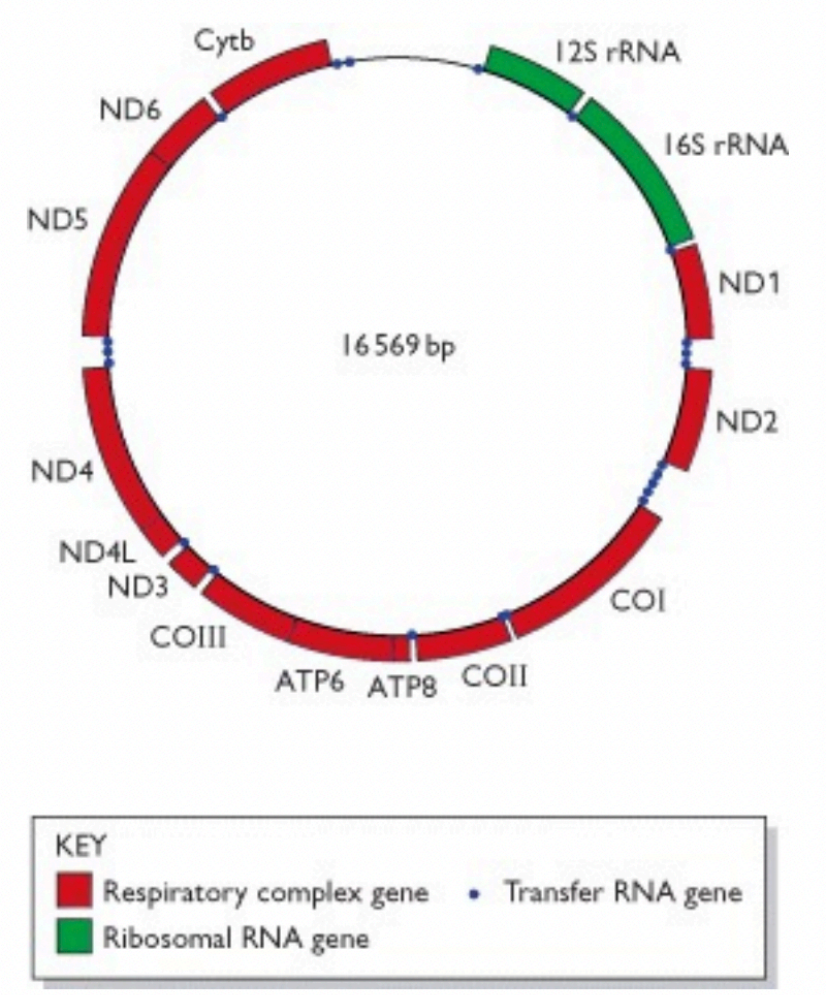


Based on IHGSC (2001) and Venter et al. (2001).

https://www.ncbi.nlm.nih.gov/books/NBK21134/

The Human Genome

- The Organization of the Human Genome Mitochondria!



The human mitochondrial genome

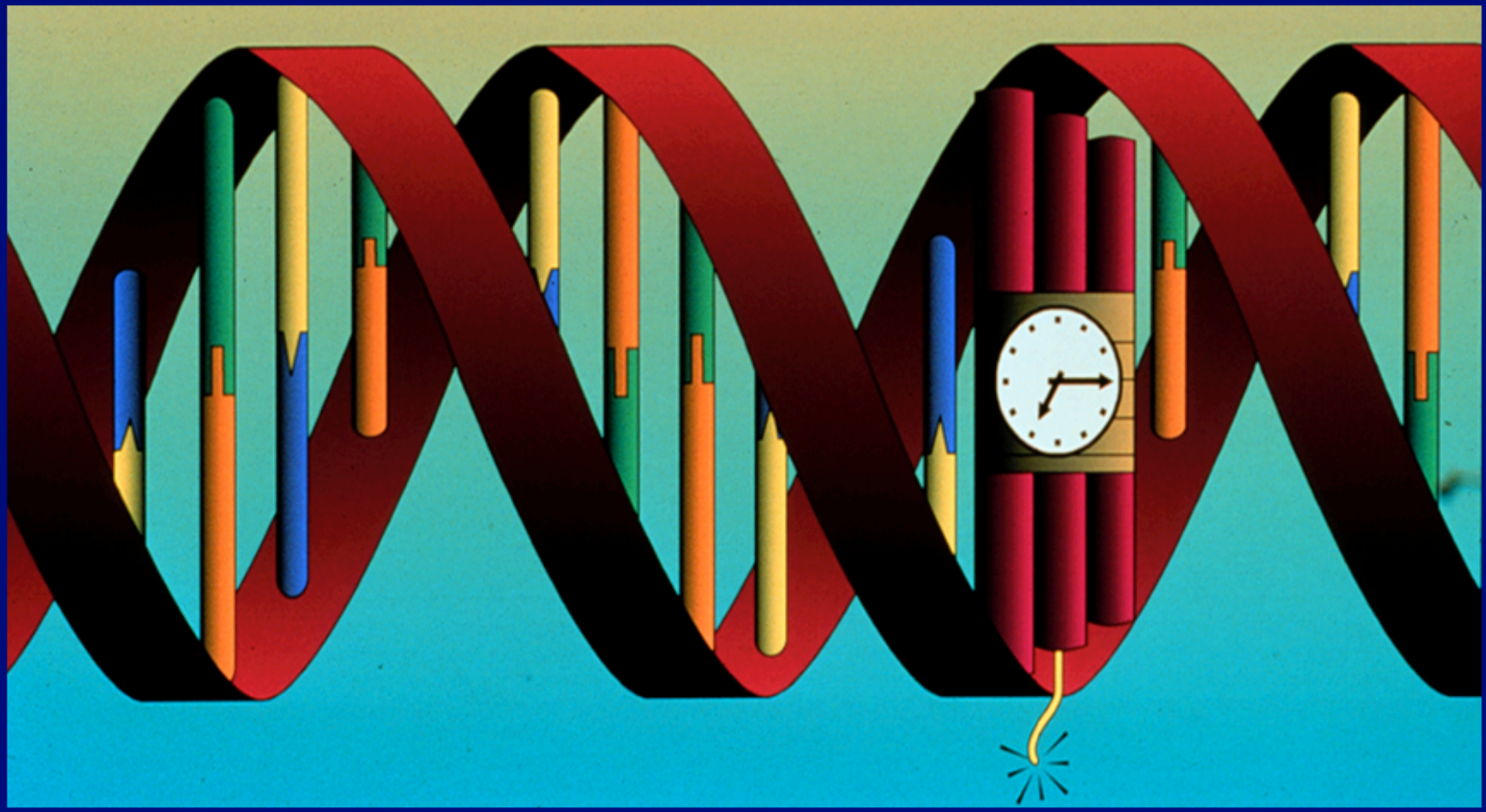
The human mitochondrial genome is small and compact, with little wasted space, so much so that the ATP6 and ATP8 genes overlap.

Abbreviations:
ATP6, ATP8, genes for ATPase subunits 6 and 8;
COI, COII, COIII, genes for cytochrome c oxidase subunits I, II and III;
Cytb, gene for apocytochrome b;
ND1–ND6, genes for NADH hydrogenase subunits 1–6.

Ribosomal RNA and transfer RNA are two types of non-coding RNA

https://www.ncbi.nlm.nih.gov/books/NBK21134/

All humans are ~99.7% identical at the DNA sequence level, and yet…



all of us carry a significant number of ‘glitches’ in our genomes.

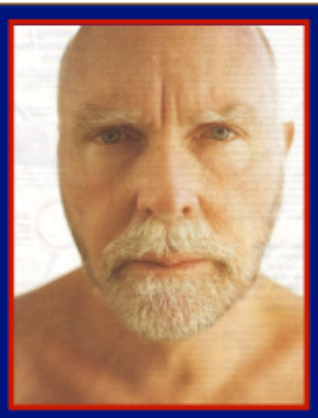
https://www.genome.gov/event-calendar/Current-Topics-in-Genome-Analysis

Individual Genome Sequences

The Diploid Genome Sequence of an Individual Human

Samuel Levy[1*], Granger Sutton[1], Pauline C. Ng[1], Lars Feuk[2], Aaron L. Halpern[1], Brian P. Walenz[1], Nelson Axelrod[1], Jiaqi Huang[1], Ewen F. Kirkness[1], Gennady Denisov[1], Yuan Lin[1], Jeffrey R. MacDonald[2], Andy Wing Chun Pang[2], Mary Shago[2], Timothy B. Stockwell[1], Alexia Tsiamouri[1], Vineet Bafna[3], Vikas Bansal[3], Saul A. Kravitz[1], Dana A. Busam[1], Karen Y. Beeson[1], Tina C. McIntosh[1], Karin A. Remington[1], Josep F. Abril[4], John Gill[1], Jon Borman[1], Yu-Hui Rogers[1], Marvin E. Frazier[1], Stephen W. Scherer[2], Robert L. Strausberg[1], J. Craig Venter[1]

***PLoS Biol* (2007)**



Accurate whole human genome sequencing using reversible terminator chemistry

A list of authors and their affiliations appears at the end of the paper

***Nature* (2008)**

The diploid genome sequence of an Asian individual

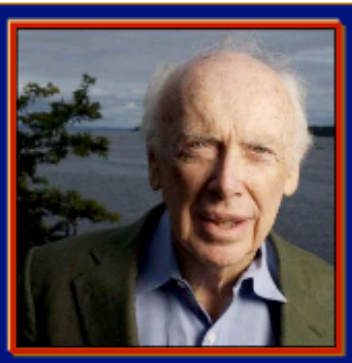
Jun Wang[1,2,3,4*], Wei Wang[1,3*], Ruiqiang Li[1,3,4*], Yingrui Li[1,5,6*], Geng Tian[1,7], Laurie Goodman[1], Wei Fan[1], Junqing Zhang[1], Jun Li[1], Juanbin Zhang[1], Yiran Guo[1,7], Binxiao Feng[1], Heng Li[1,8], Yao Lu[1], Xiaodong Fang[1], Huiqing Liang[1], Zhenglin Du[1], Dong Li[1], Yiqing Zhao[1,7], Yujie Hu[1,7], Zhenzhen Yang[1], Hancheng Zheng[1], Ines Hellmann[9], Michael Inouye[8], John Pool[9], Xin Yi[1,7], Jing Zhao[1], Jinjie Duan[1], Yan Zhou[1], Junjie Qin[1,7], Lijia Ma[1,7], Guoqing Li[1], Zhentao Yang[1], Guojie Zhang[1,7], Bin Yang[1], Chang Yu[1], Fang Liang[1,7], Wenjie Li[1], Shaochuan Li[1], Dawei Li[1], Peixiang Ni[1], Jue Ruan[1,7], Qibin Li[1,7], Hongmei Zhu[1], Dongyuan Liu[1], Zhike Lu[1], Ning Li[1,7], Guangwu Guo[1,7], Jianguo Zhang[1], Jia Ye[1], Lin Fang[1], Qin Hao[1,7], Quan Chen[1,5], Yu Liang[1,7], Yeyang Su[1,7], A. san[1,7], Cuo Ping[1,7], Shuang Yang[1], Fang Chen[1,7], Li Li[1], Ke Zhou[1], Hongkun Zheng[1,4], Yuanyuan Ren[1], Ling Yang[1], Yang Gao[1,6], Guohua Yang[1,2], Zhuo Li[1], Xiaoli Feng[1], Karsten Kristiansen[8], Gane Ka-Shu Wong[1,10], Rasmus Nielsen[9], Richard Durbin[8], Lars Bolund[1,11], Xiuqing Zhang[1,6], Songgang Li[1,2,5], Huanming Yang[1,2,3] & Jian Wang[1,2,3]

***Nature* (2008)**

The complete genome of an individual by massively parallel DNA sequencing

David A. Wheeler[1]*, Maithreyan Srinivasan[2]*, Michael Egholm[2]*, Yufeng Shen[1]*, Lei Chen[1], Amy McGuire[3], Wen He[2], Yi-Ju Chen[2], Vinod Makhijani[2], G. Thomas Roth[2], Xavier Gomes[2], Karrie Tartaro[2]†, Faheem Niazi[2], Cynthia L. Turcotte[2], Gerard P. Irzyk[2], James R. Lupski[4,5,6], Craig Chinault[4], Xing-zhi Song[1], Yue Liu[1], Ye Yuan[1], Lynne Nazareth[1], Xiang Qin[1], Donna M. Muzny[1], Marcel Margulies[2], George M. Weinstock[1,4], Richard A. Gibbs[1,4] & Jonathan M. Rothberg[2]†

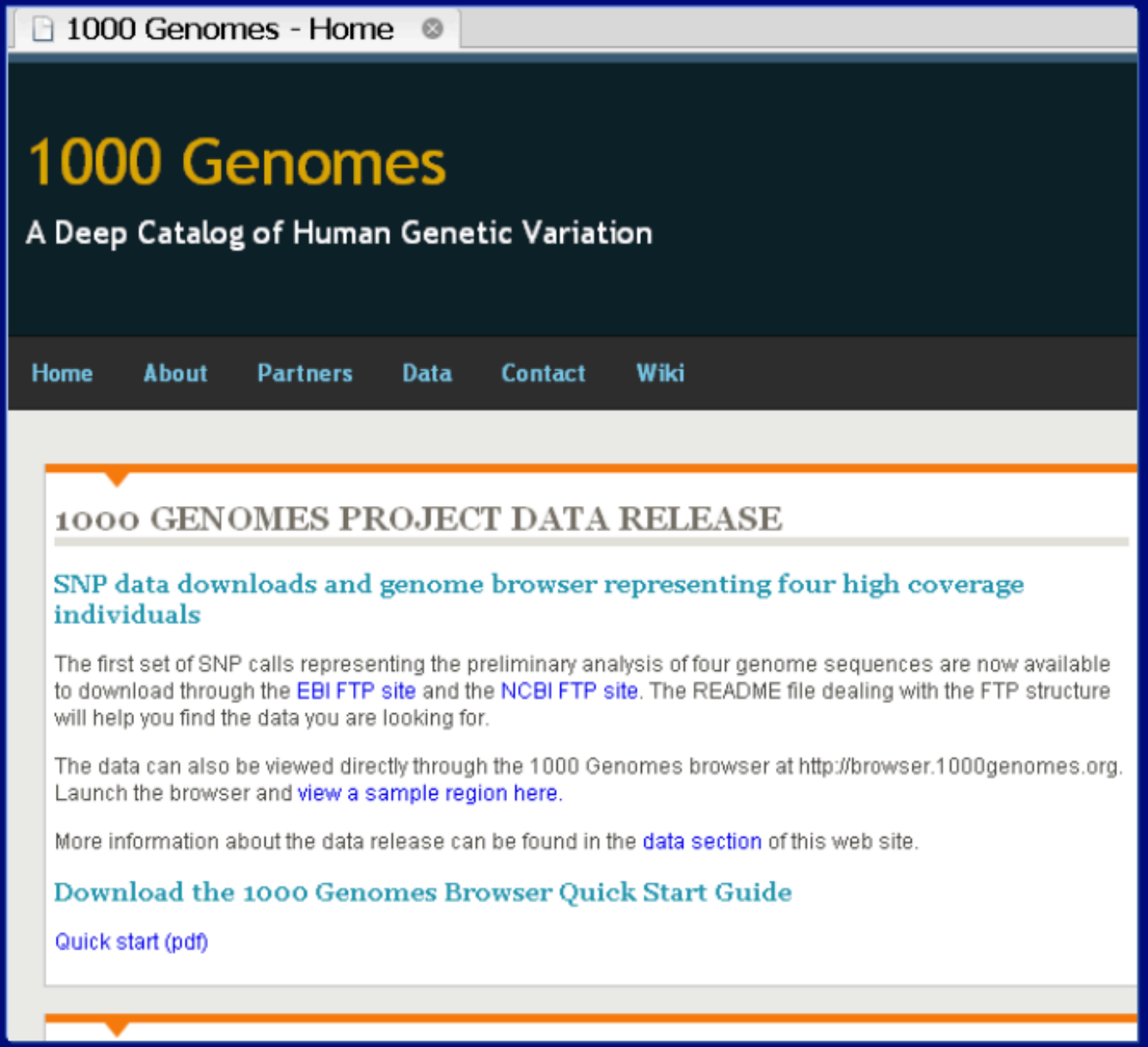
***Nature* (2008)**



A highly annotated whole-genome sequence of a Korean individual

Jong-Il Kim[1,2,4,5*], Young Seok Ju[1,2*], Hansoo Park[1,5], Sheehyun Kim[4], Seonwook Lee[4], Jae-Hyuk Yi[1], Joann Mudge[6], Neil A. Miller[6], Dongwan Hong[3], Callum J. Bell[6], Hye-Sun Kim[4], In-Soon Chung[4], Woo-Chung Lee[4], Ji-Sun Lee[4], Seung-Hyun Seo[5], Ji-Young Yun[5], Hyun Nyun Woo[4], Heewook Lee[4], Dongwhan Suh[1,2,3], Seungbok Lee[1,2,3], Hyun-Jin Kim[1,3], Maryam Yavartanoo[1,2], Minhye Kwak[1,2], Ying Zheng[1,2], Mi Kyeong Lee[3], Hyunjun Park[1], Jeong Yeon Kim[1], Omer Gokcumen[7], Ryan E. Mills[7], Alexander Wait Zaranek[8], Joseph Thakuria[8], Xiaodi Wu[8], Ryan W. Kim[6], Jim J. Huntley[9], Shujun Luo[9], Gary P. Schroth[9], Thomas D. Wu[10], HyeRan Kim[4], Kap-Seok Yang[4], Woong-Yang Park[1,2,3], Hyungtae Kim[4], George M. Church[8], Charles Lee[7], Stephen F. Kingsmore[6] & Jeong-Sun Seo[1,2,3,4,5]

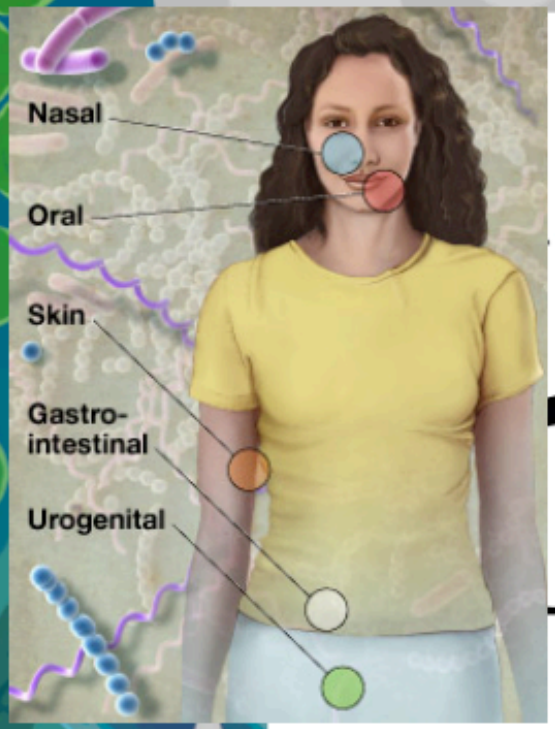
***Nature* (2009)**



1000genomes.org

cancergenome.nih.gov

The Human Microbiome Project



HMP

nihroadmap.nih.gov/hmp

NCBI Genomes Database

Genome

This resource organizes information on genomes including sequences, maps, chromosomes, assemblies, and annotations.

Using Genome

- Help
- Browse by Organism UPDATED
- Download / FTP
- Download FAQ
- Submit a genome

Custom resources

- Human Genome
- Microbes
- Organelles
- Viruses
- Prokaryotic reference genomes

Other Resources

- Assembly
- BioProject
- BioSample
- Genome Data Viewer
- NCBI Datasets NEW

Genome Tools

- BLAST the Human Genome
- Microbial Nucleotide BLAST

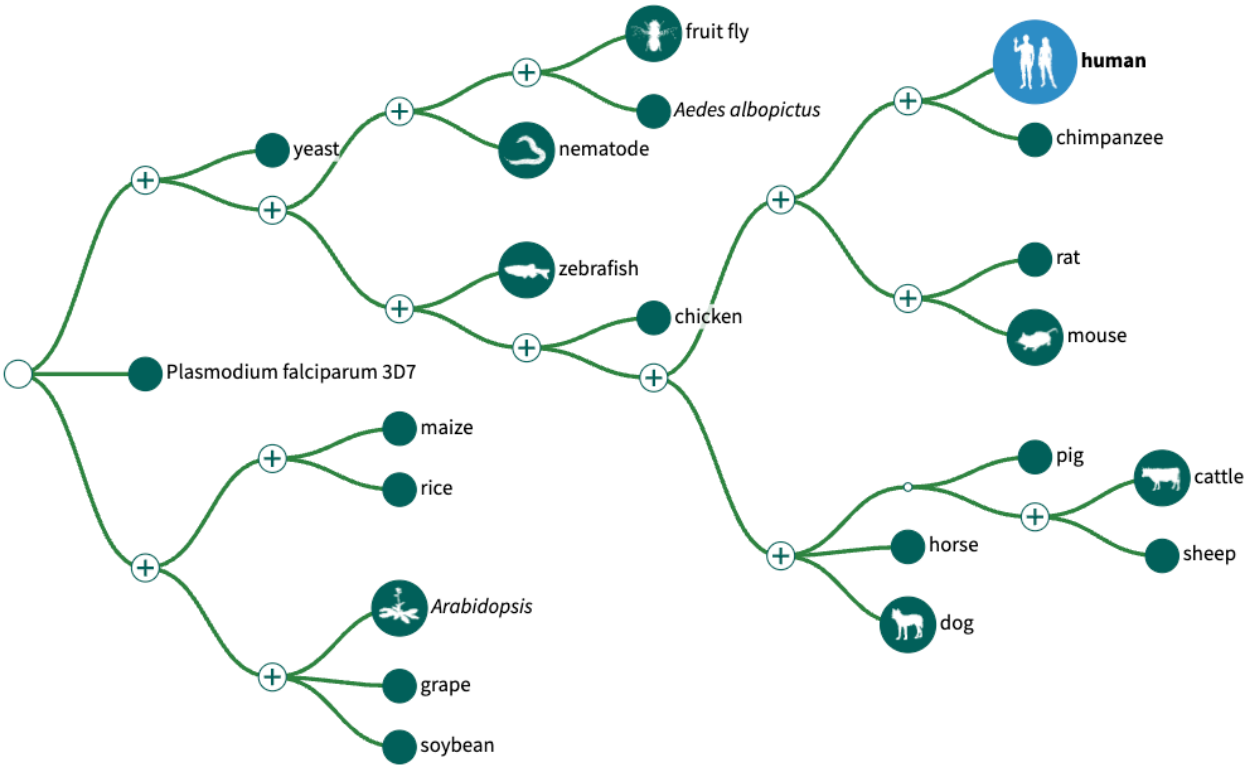
Genome Annotation and Analysis

- Eukaryotic Genome Annotation
- Prokaryotic Genome Annotation
- PASC (Pairwise Sequence Comparison)

External Resources

- GOLD - Genomes Online Database
- Bacteria Genomes at Sanger
- Ensembl

Genome Data Viewer



Homo sapiens (human)

Search in genome

Location, gene or phenotype

Examples: TP53, chr17:7667000-7689000, DNA repair

Assembly

GRCh38.p13

Browse genome **BLAST genome**

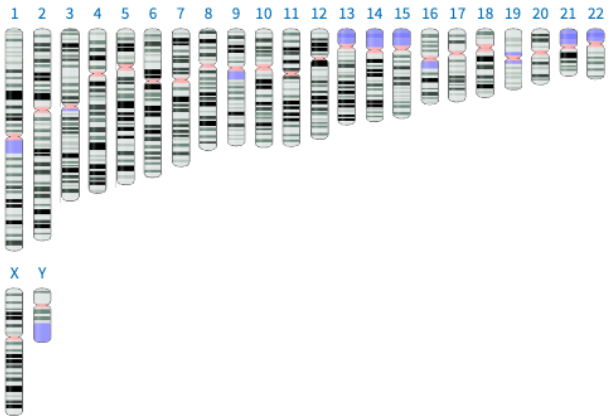
Download via NCBI Datasets

Assembly details

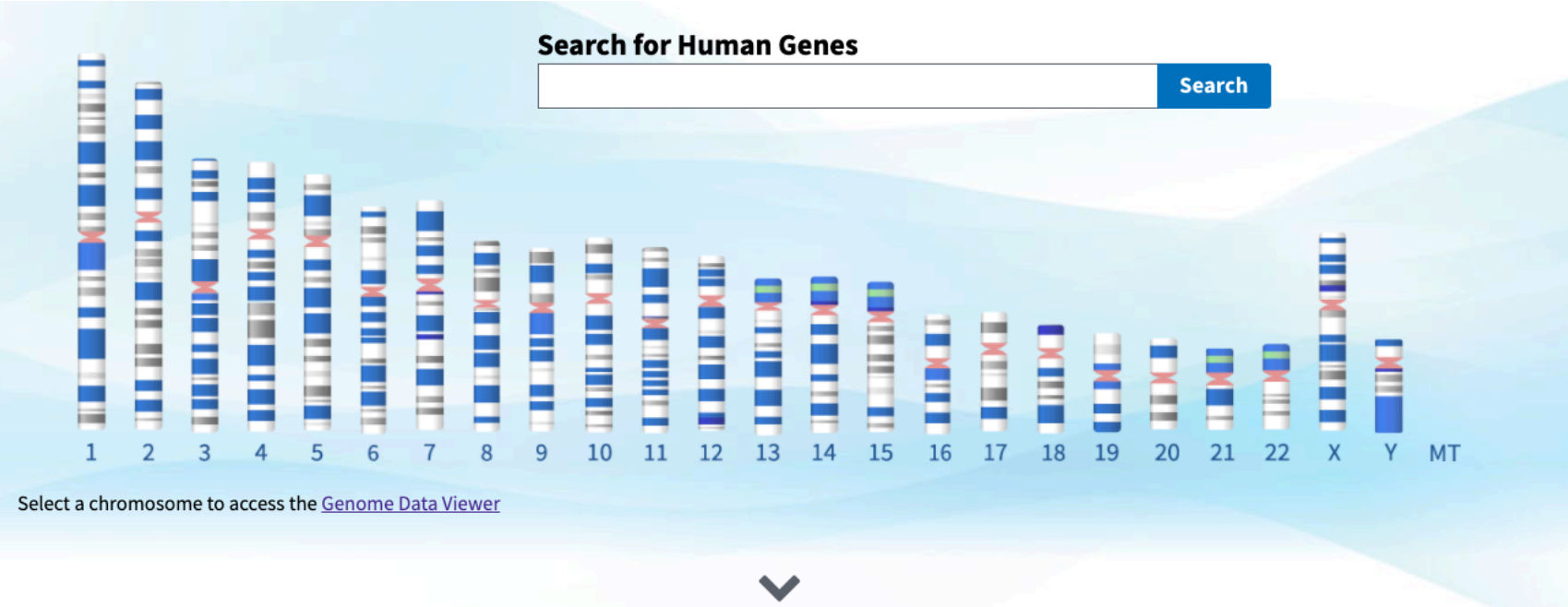
Name	GRCh38.p13
RefSeq accession	GCF_000001405.39
GenBank accession	GCA_000001405.28
Submitter	Genome Reference Consortium
Level	Chromosome
Category	Reference genome
Replaced by	GCF_000001405.25

Annotation details

Annotation Release	109
Release date	Nov 21, 2021



Human Resources at NCBI

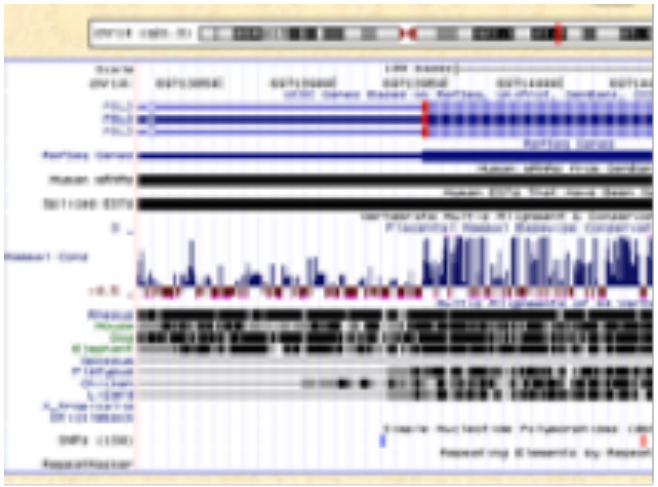


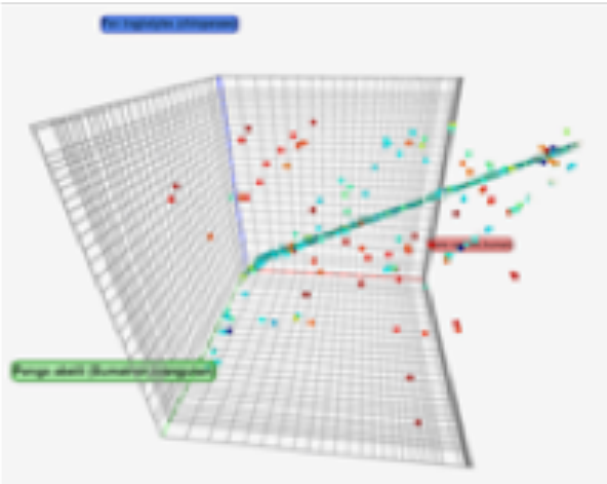
Download

	GRCh38	GRCh37
Reference Genome Sequence	Fasta	Fasta
RefSeq Reference Genome Annotation	gff3	gff3
RefSeq Transcripts	Fasta	Fasta
RefSeq Proteins	Fasta	Fasta
ClinVar	vcf	vcf
dbSNP	vcf	vcf
dbVar	vcf	vcf

Computational Genomics

Other Genome Projects





Earth BioGenome Project: Sequencing life for the future of life

Harris A. Lewin[a,b,c,d,1], Gene E. Robinson[e], W. John Kress[f], William J. Baker[g], Jonathan Coddington[f], Keith A. Crandall[h], Richard Durbin[i,j], Scott V. Edwards[k,l], Félix Forest[g], M. Thomas P. Gilbert[m,n], Melissa M. Goldstein[o], Igor V. Grigoriev[p,q], Kevin J. Hackett[r], David Haussler[s,t], Erich D. Jarvis[u], Warren E. Johnson[v], Aristides Patrinos[w], Stephen Richards[x], Juan Carlos Castilla-Rubio[y,z], Marie-Anne van Sluys[aa,bb], Pamela S. Soltis[cc], Xun Xu[dd], Huanming Yang[ee], and Guojie Zhang[dd,ff,gg]

Edited by John C. Avise, University of California, Irvine, CA, and approved March 15, 2018 (received for review January 6, 2018)

Increasing our understanding of Earth's biodiversity and responsibly stewarding its resources are among the most crucial scientific and social challenges of the new millennium. These challenges require fundamental new knowledge of the organization, evolution, functions, and interactions among millions of the planet's organisms. Herein, we present a perspective on the Earth BioGenome Project (EBP), a moonshot for biology that aims to sequence, catalog, and characterize the genomes of all of Earth's eukaryotic biodiversity over a period of 10 years. The outcomes of the EBP will inform a broad range of major issues facing humanity, such as the impact of climate change on biodiversity, the conservation of endangered species and ecosystems, and the preservation and enhancement of ecosystem services. We describe hurdles that the project faces, including data-sharing policies that ensure a permanent, freely available resource for future scientific discovery while respecting access and benefit sharing guidelines of the Nagoya Protocol. We also describe scientific and organizational challenges in executing such an ambitious project, and the structure proposed to achieve the project's goals. The far-reaching potential benefits of creating an open digital repository of genomic information for life on Earth can be realized only by a coordinated international effort.

Welcome to the Vertebrate Genomes Project (VGP), which aims to generate near error-free reference genome assemblies of all 70,000 extant vertebrate species.

ACCESS OUR VGP GENOMES @GENOMEARK (GITHUB)

GENOME 10K.

THE VERTEBRATE GENOMES PROJECT
IS A PROJECT OF THE G10K CONSORTIUM

NIH National Human Genome Research Institute

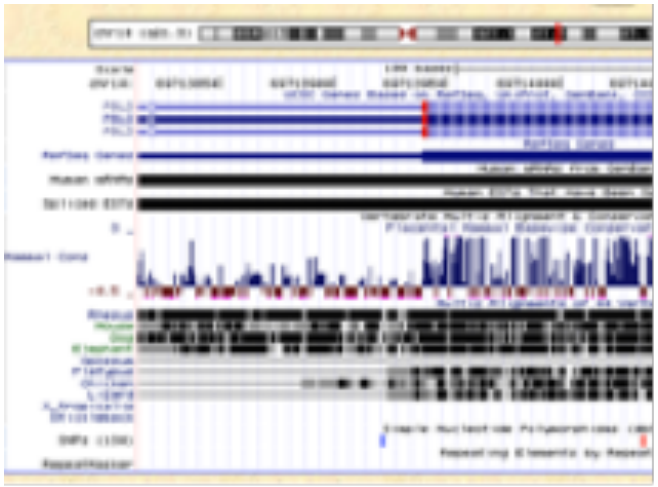
EARTH BIOGENOME PROJECT

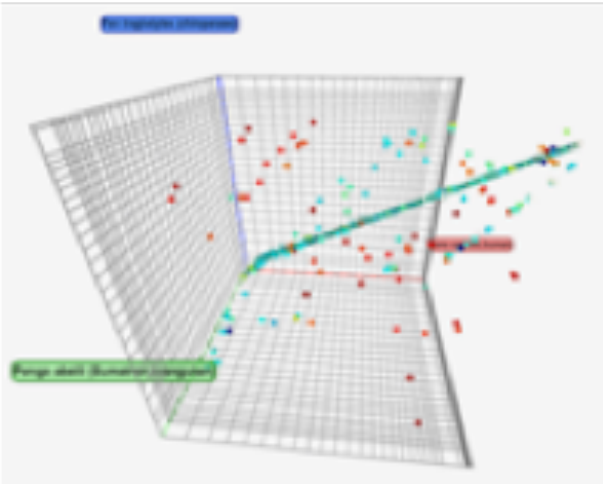
sequencing life for the future of life

https://www.earthbiogenome.org/

Computational Genomics

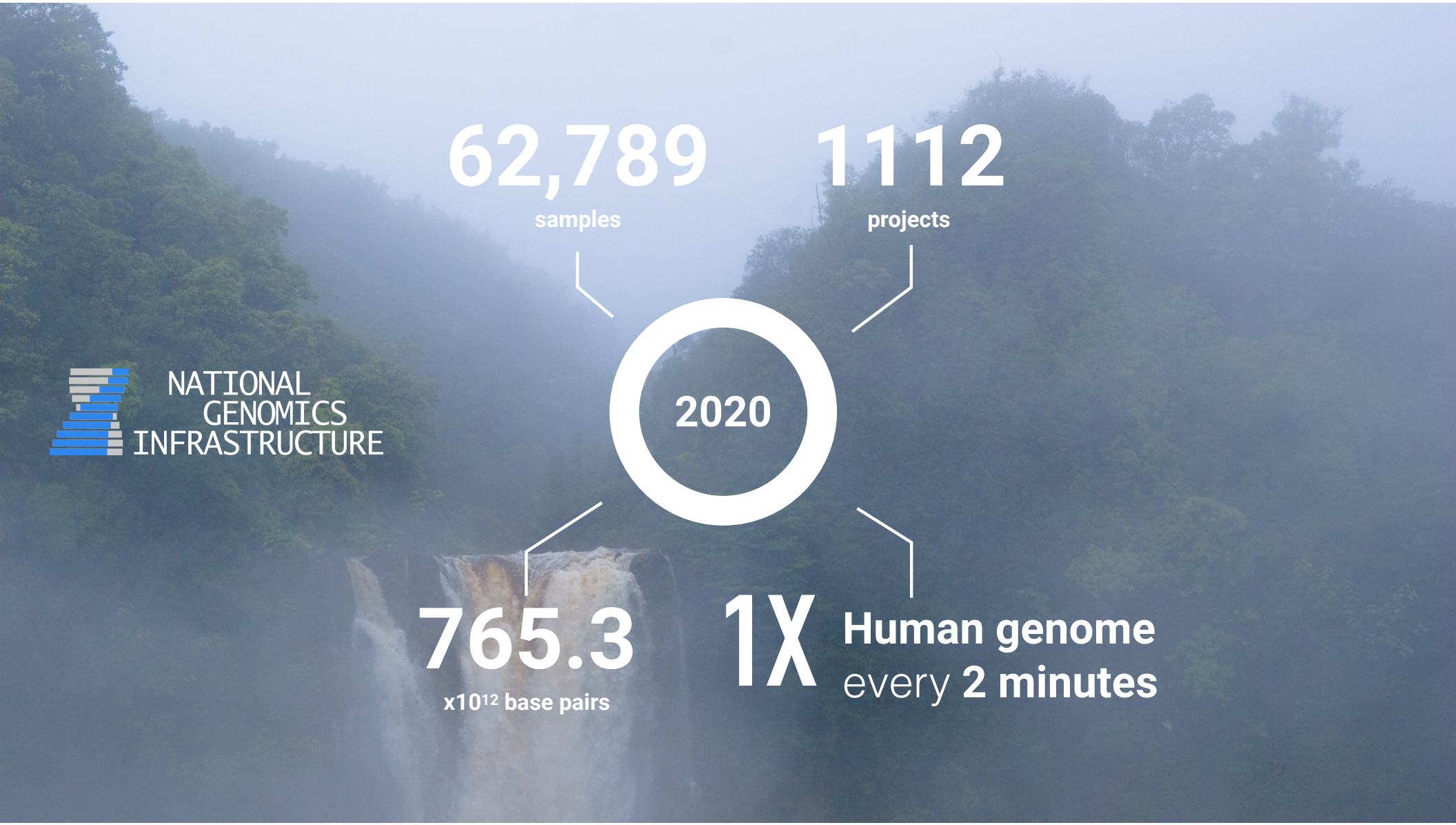
A Note About Computational Biology



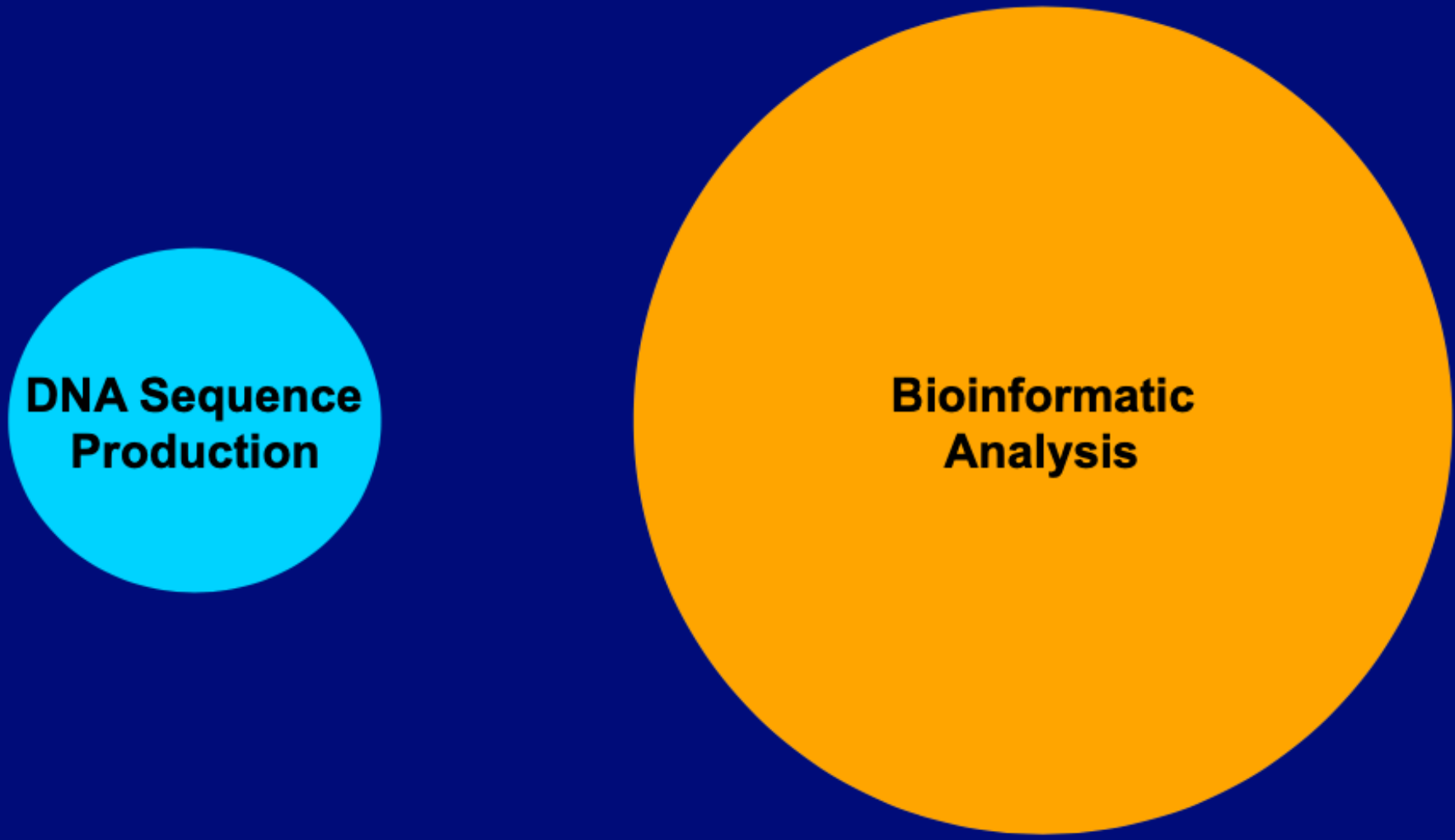


Realities of New DNA Sequencing Technologies…

https://www.genome.gov/event-calendar/Current-Topics-in-Genome-Analysis


62,789
samples
1112
projects
NATIONAL
GENOMICS
INFRASTRUCTURE
2020
765.3
x10¹² base pairs
1X
Human genome
every 2 minutes

Changing Infrastructure Requirements



The Computational Bottleneck