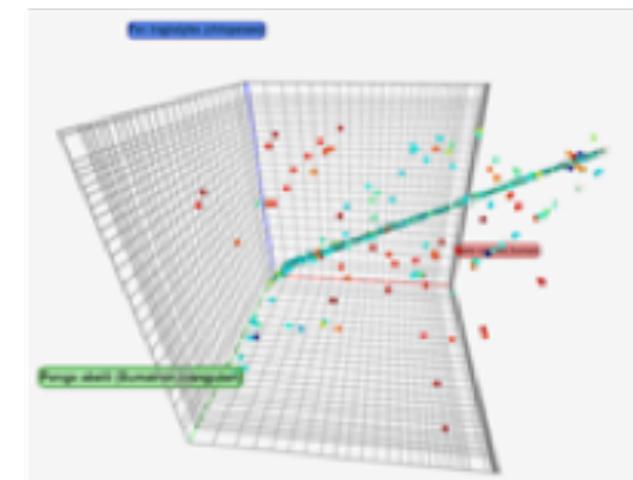
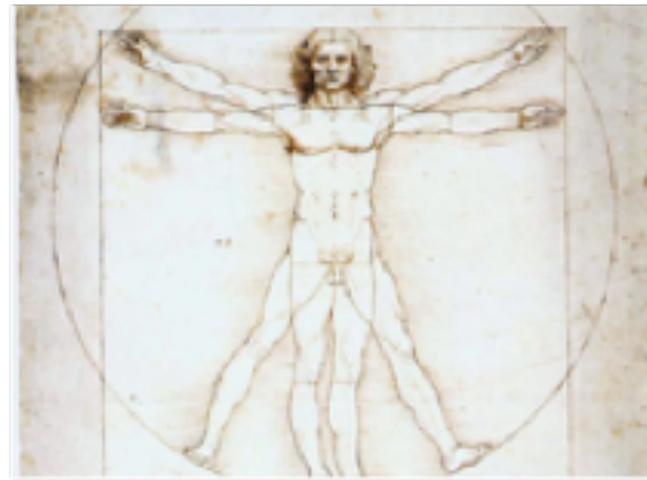
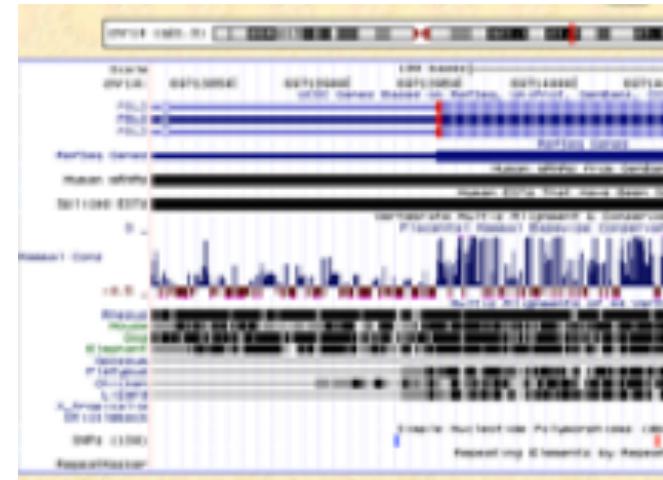
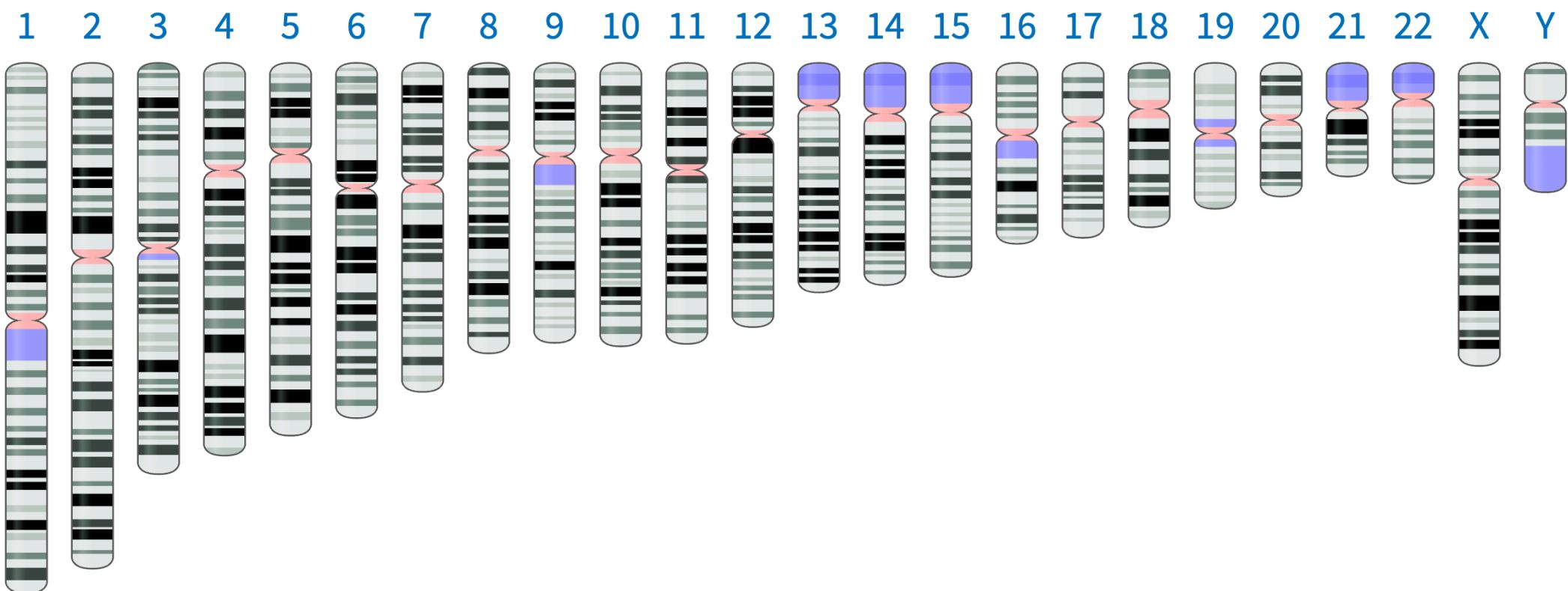


Computational Genomics

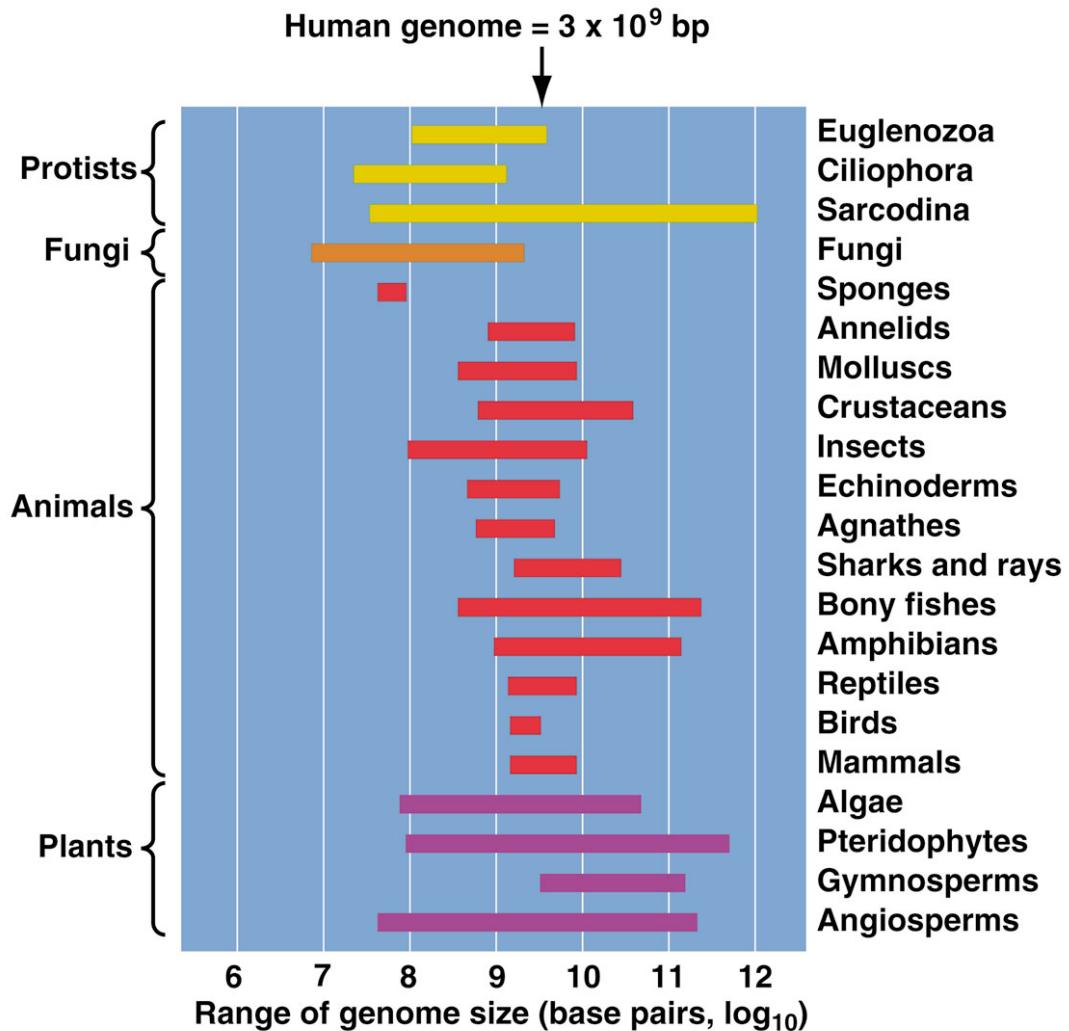
The Human Genome



The Human Genome - Our Genome...

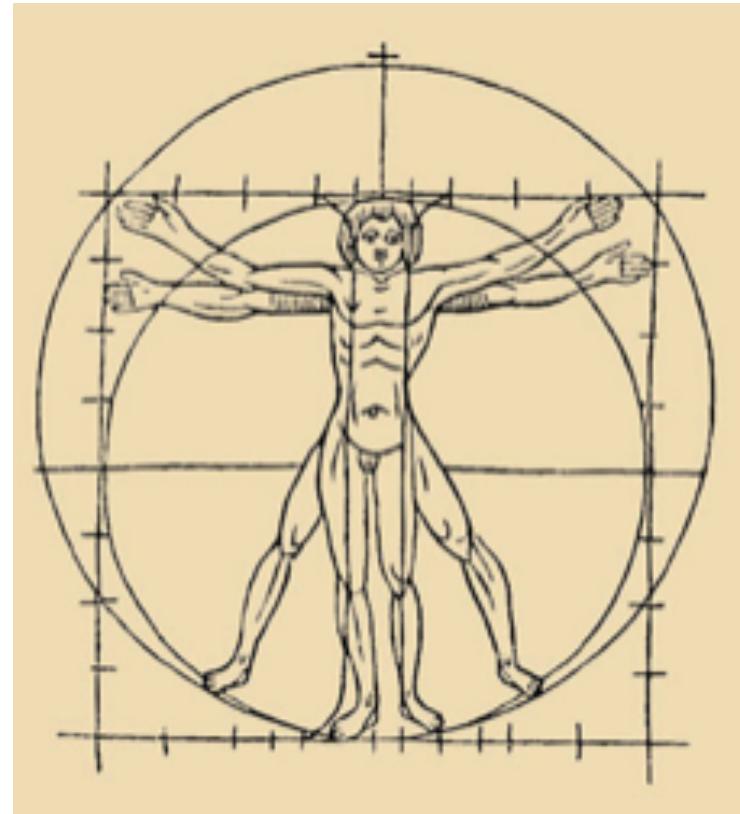


THE SIZE OF A GENOME DOES NOT REFLECT THE COMPLEXITY OF THE ORGANISM



The Human Genome

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



$\sim 3 \times 10^9$ nucleotides

The Human Genome

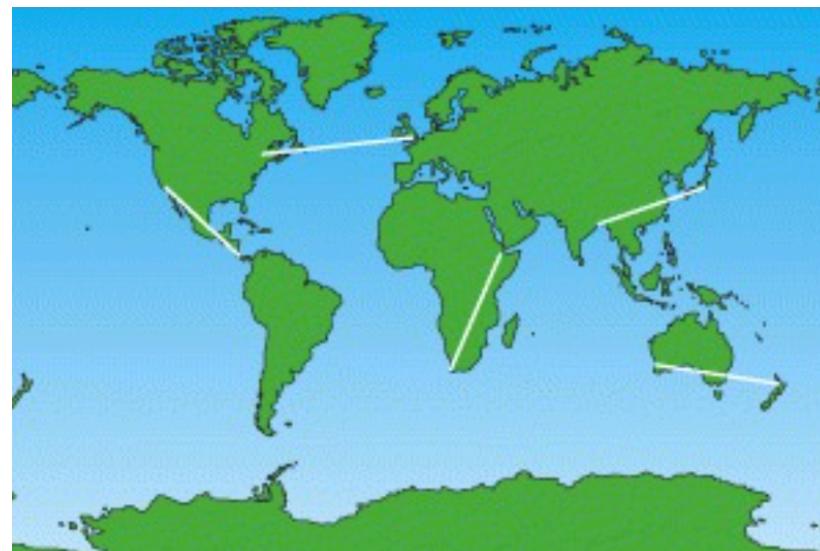
- So..How Large It Is?...



The Human Genome

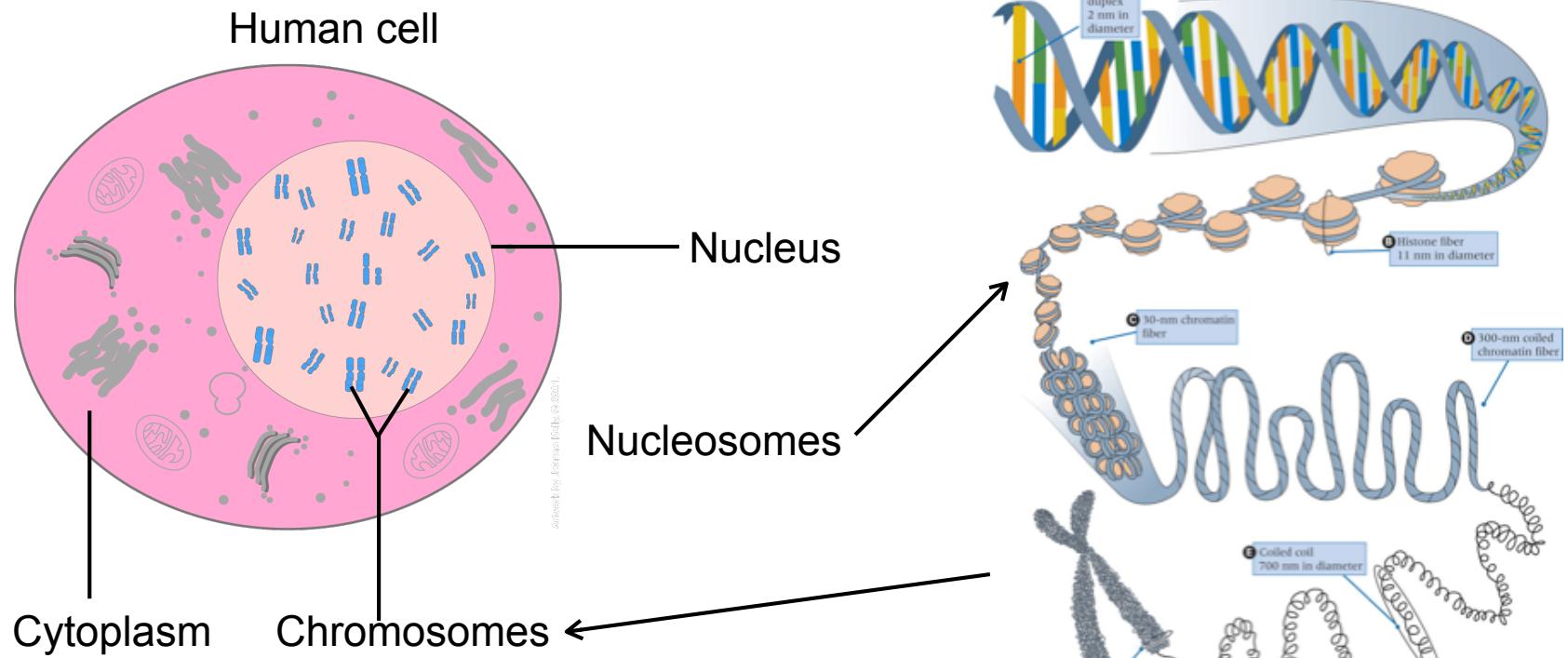
- So..How Large It Is?...

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



The Human Genome

- How is it organized?

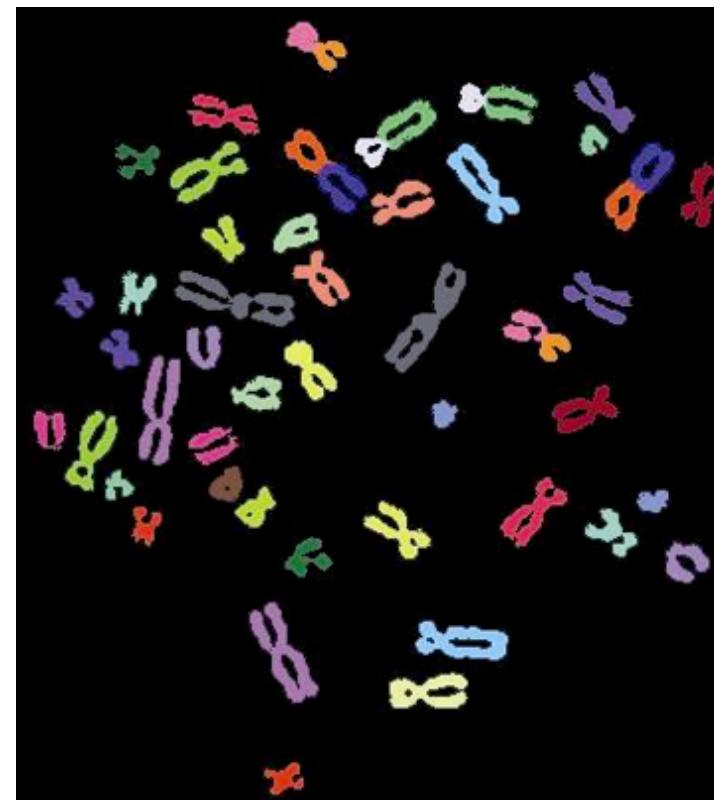
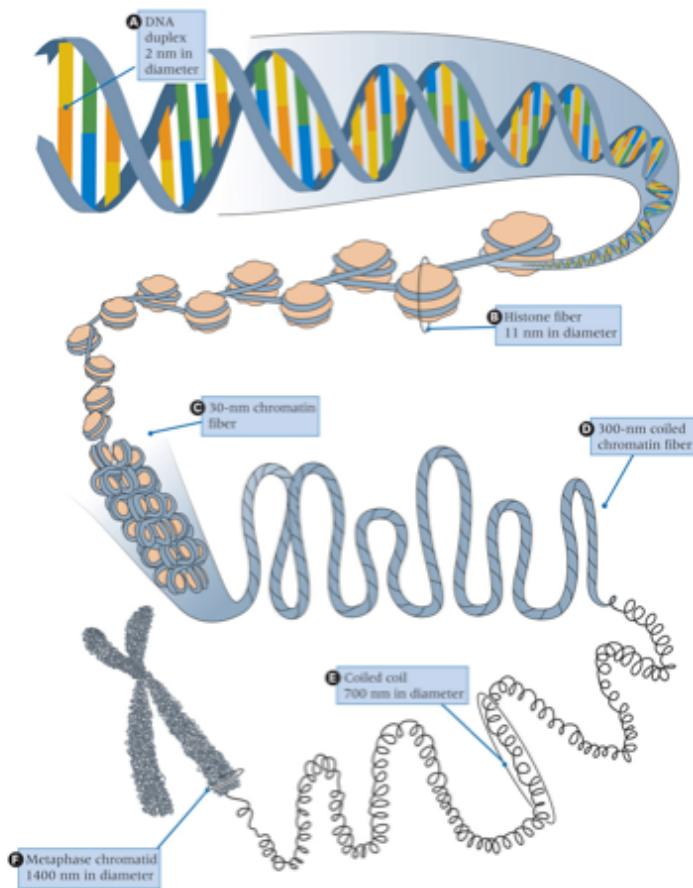


NATIONAL CANCER INSTITUTE -
[http://www.nci.nih.gov/
cancertopics/understandingcancer](http://www.nci.nih.gov/cancertopics/understandingcancer)

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

The Human Genome

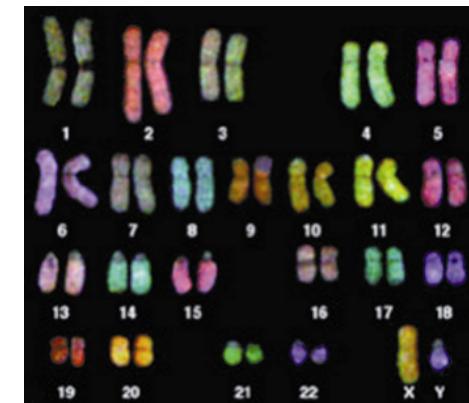
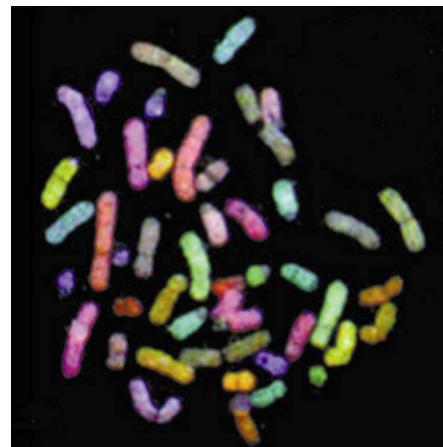
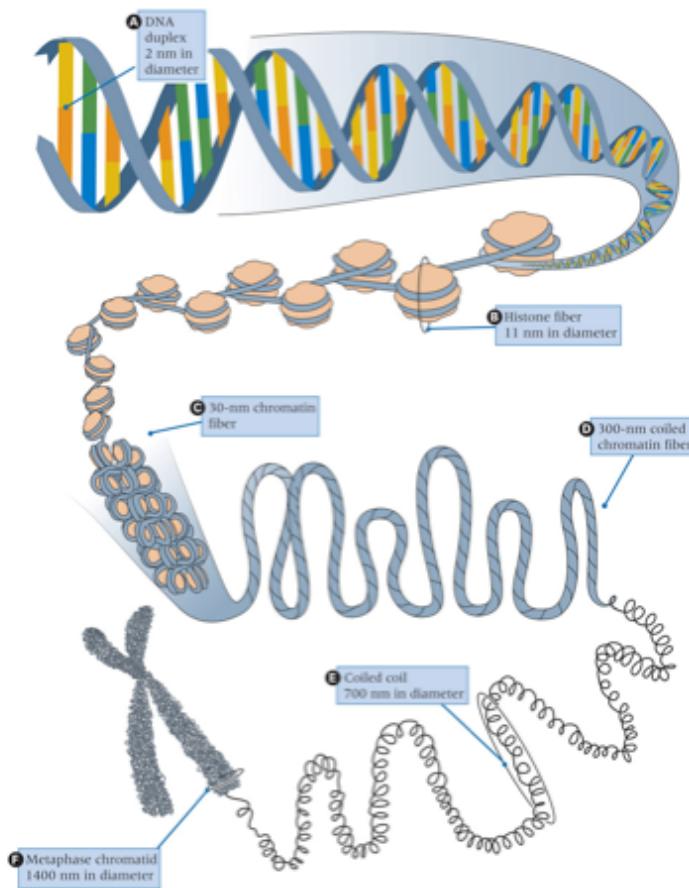
- How is it organized?



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

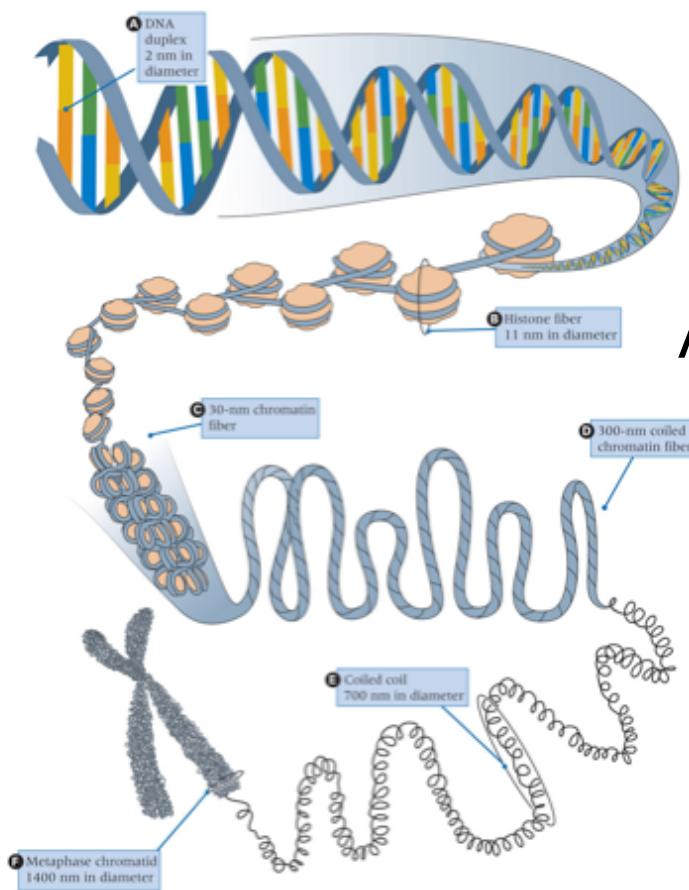
The Human Genome

- How is it organized?

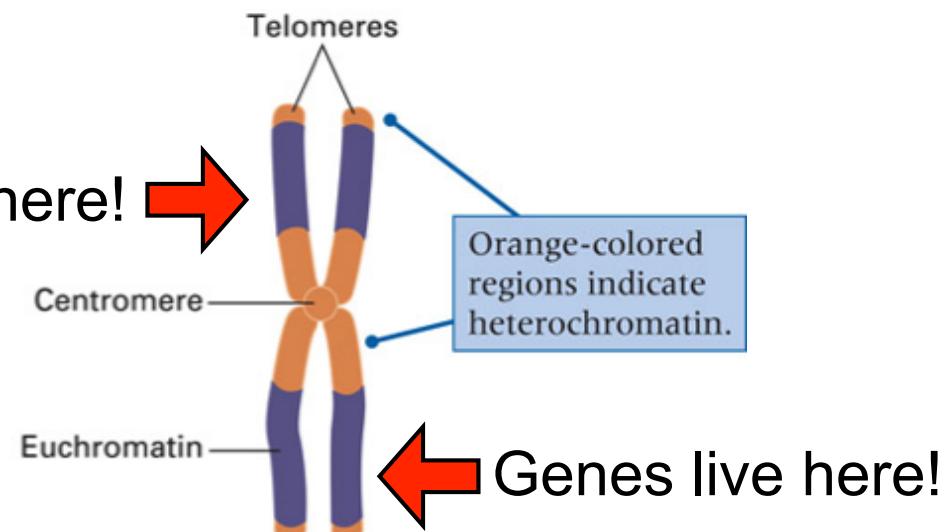


The Human Genome

- How is it organized?

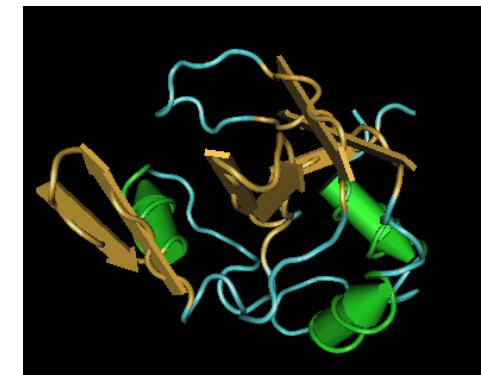
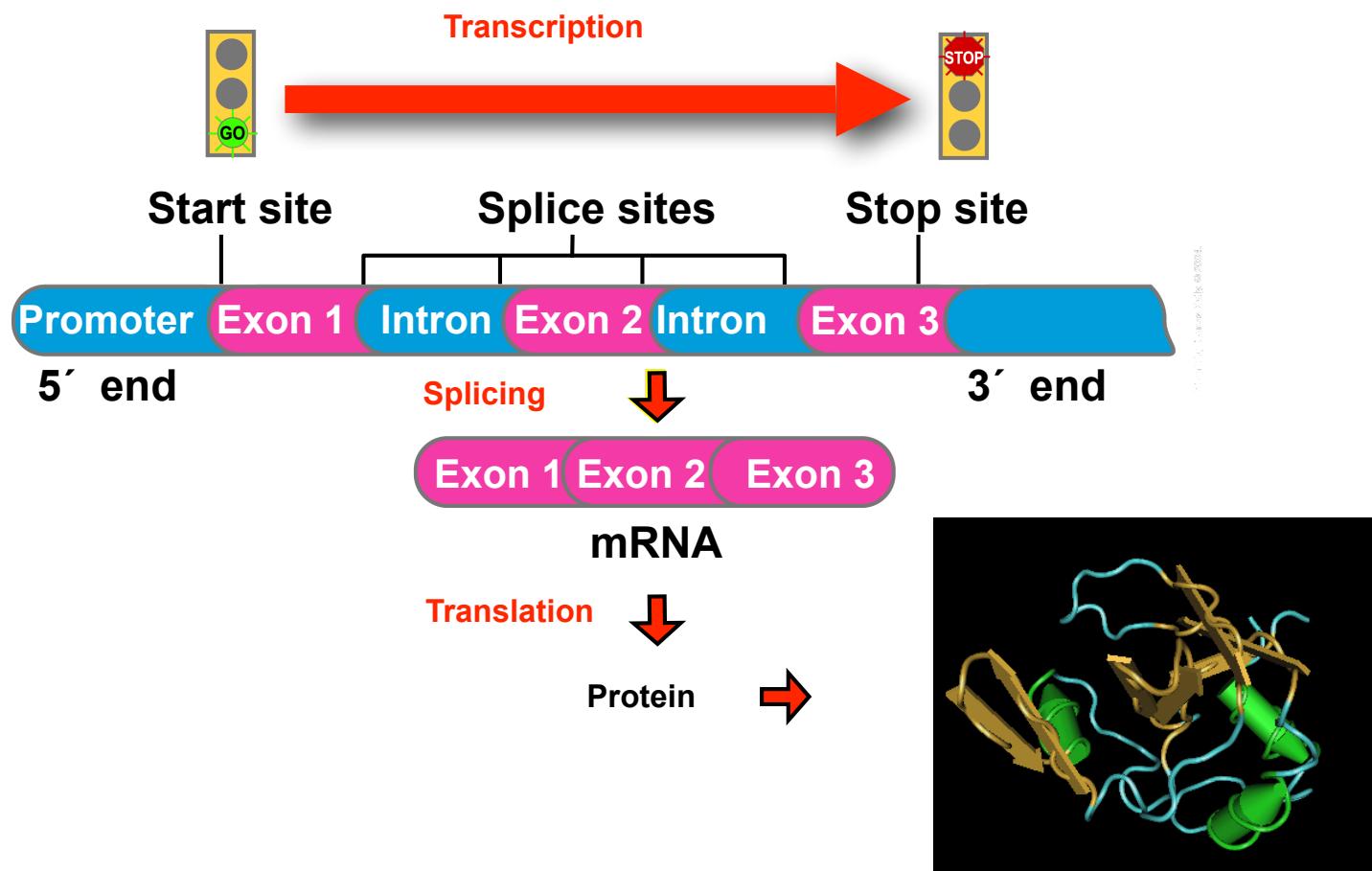


And here!



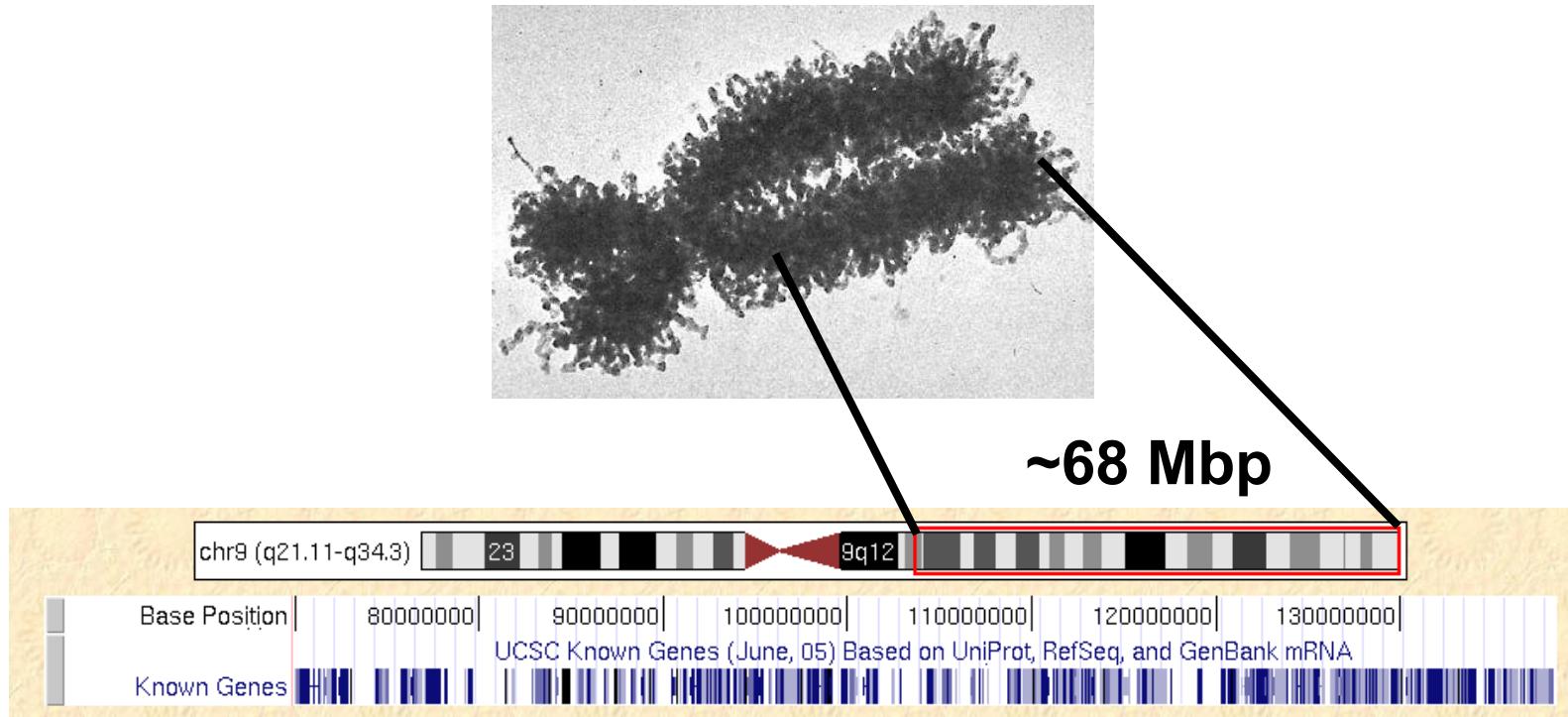
The Human Genome

- What is a (Protein-Coding) **Gene**?



The Human Genome

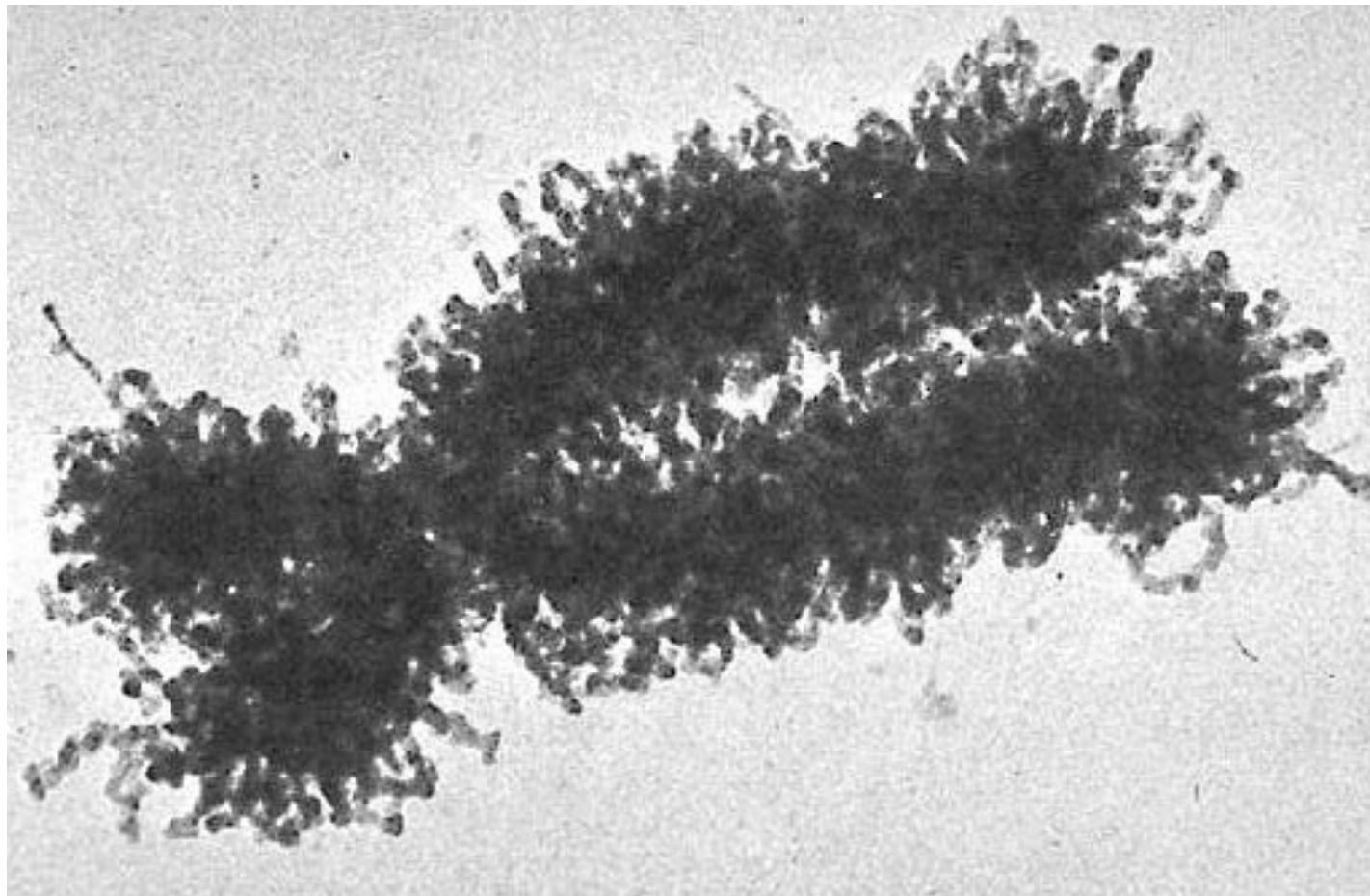
- How are Genes Distributed? Where do they live?



UCSC Genome Browser with VISTA tracks on Human May 2004 Assembly

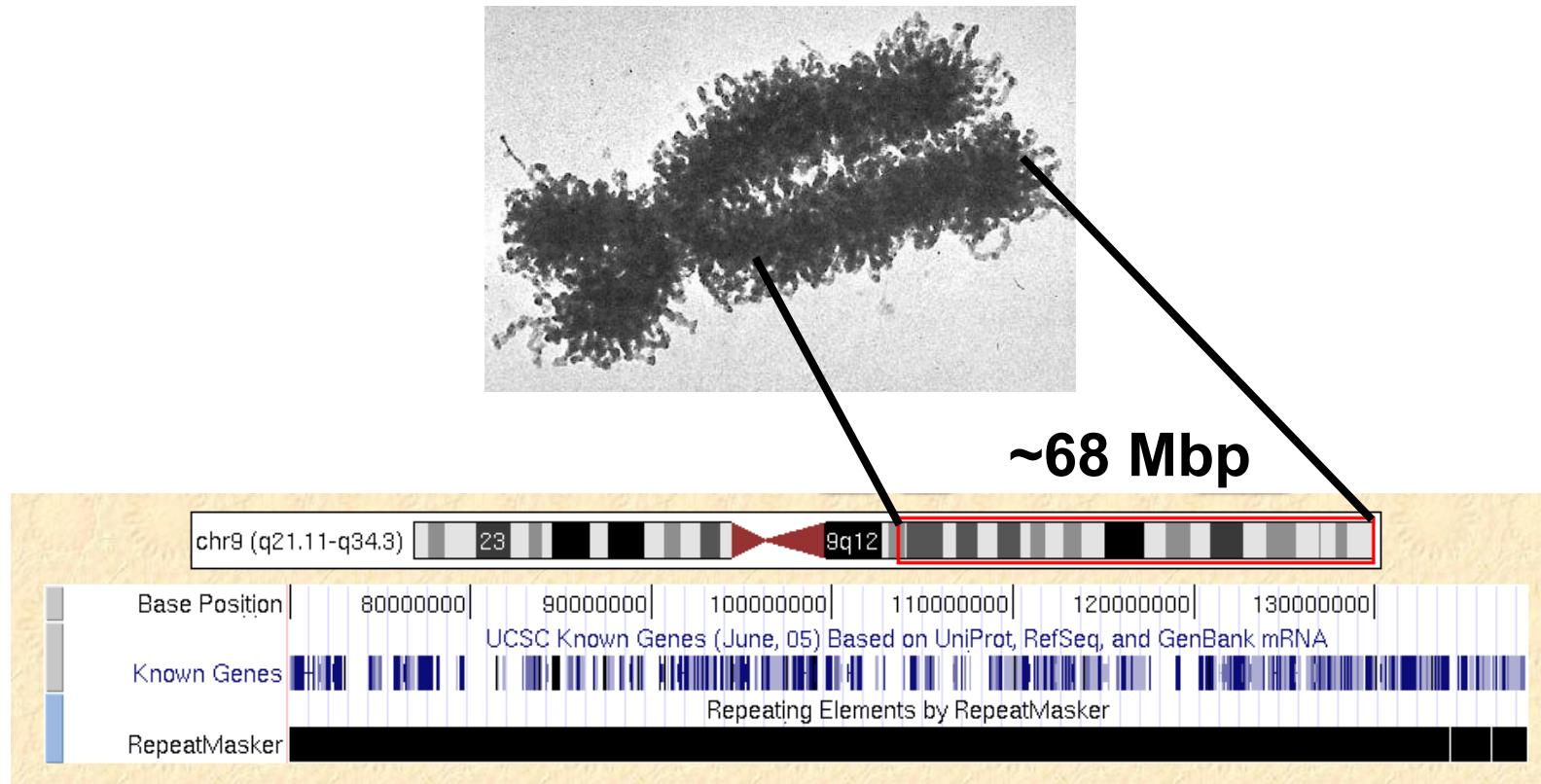
The Human Genome

- The Human Genome contains Tons of Repeated Elements!



The Human Genome

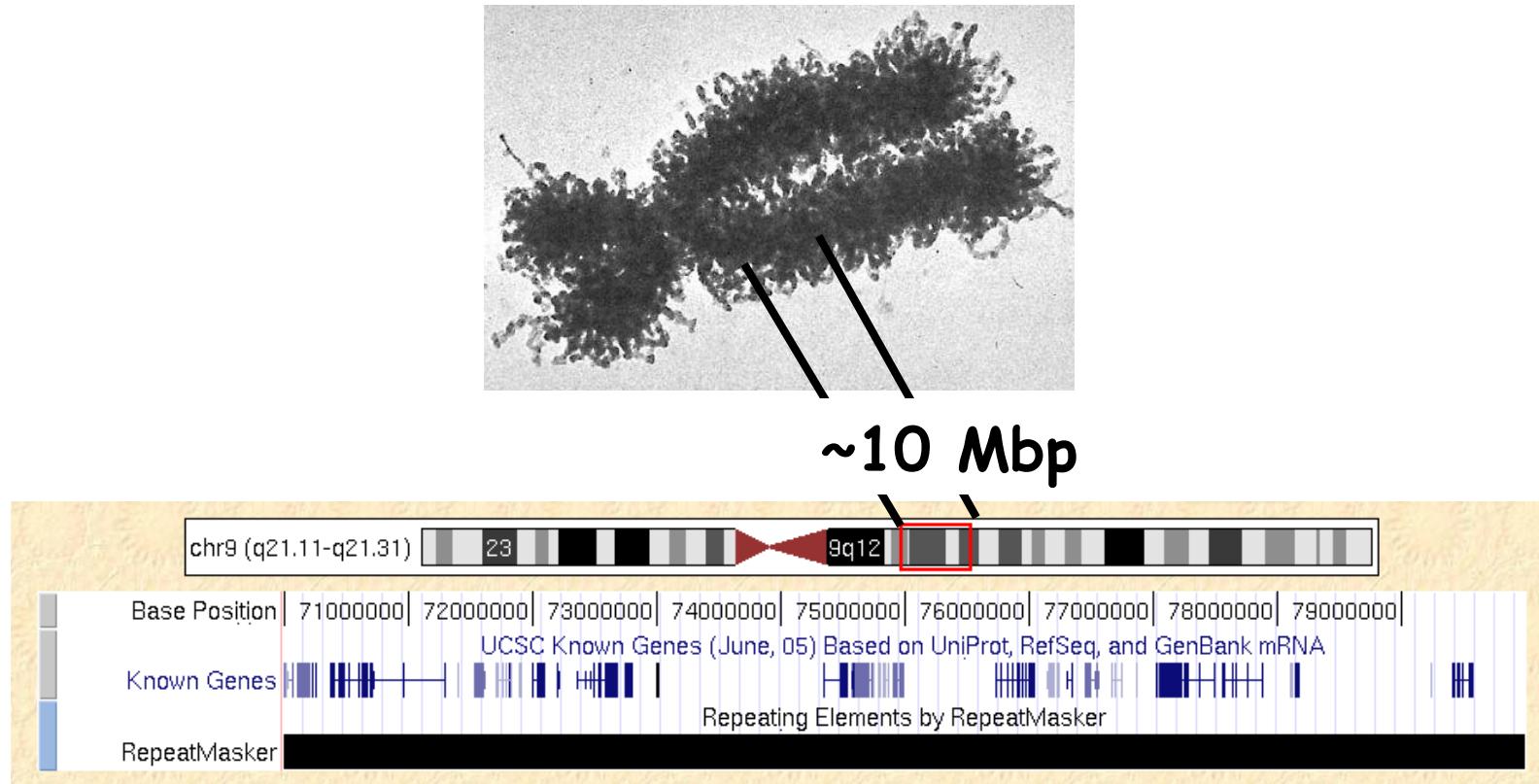
- The Human Genome is Full of Repeats!



UCSC Genome Browser with VISTA tracks on Human May 2004 Assembly

The Human Genome

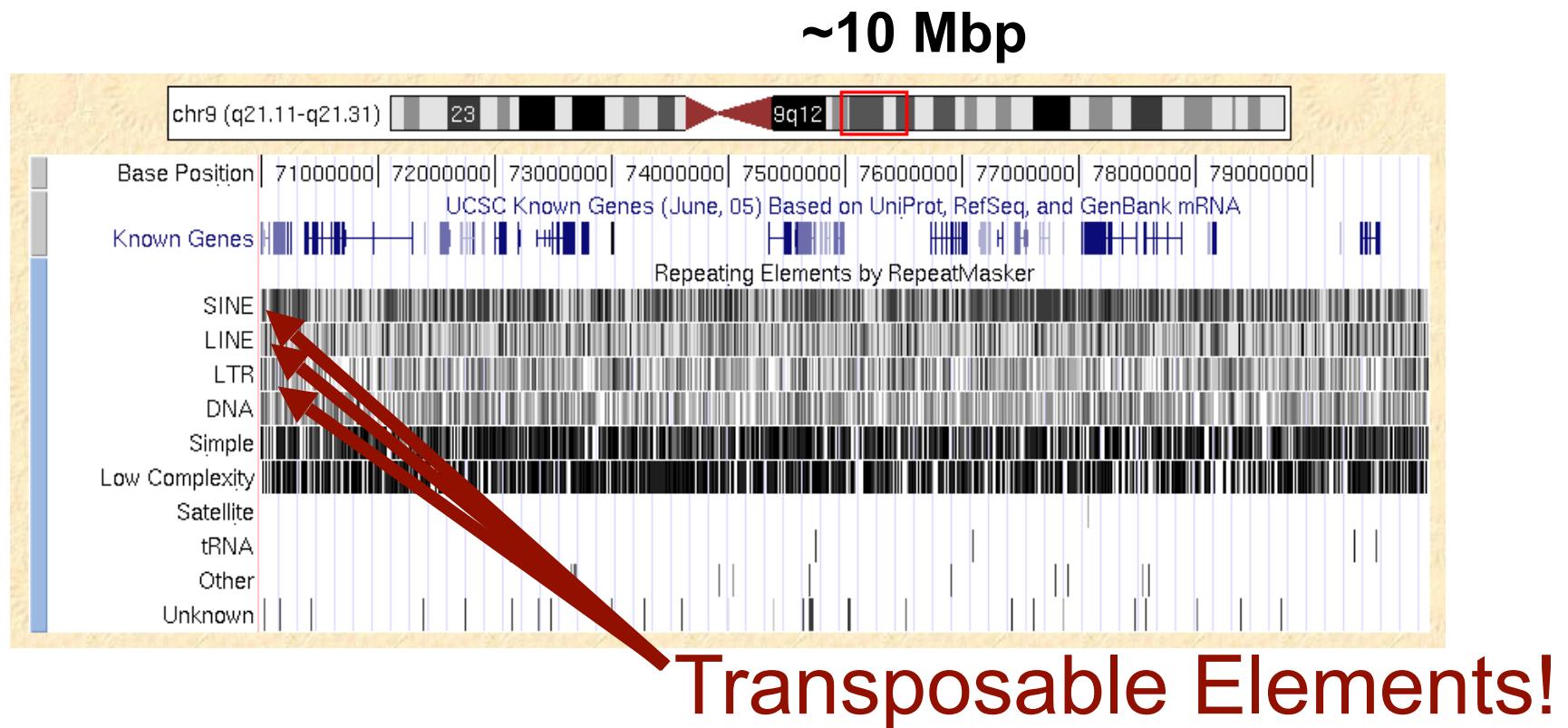
- The Human Genome is Full of Repeats!



UCSC Genome Browser with VISTA tracks on Human May
2004 Assembly

The Human Genome

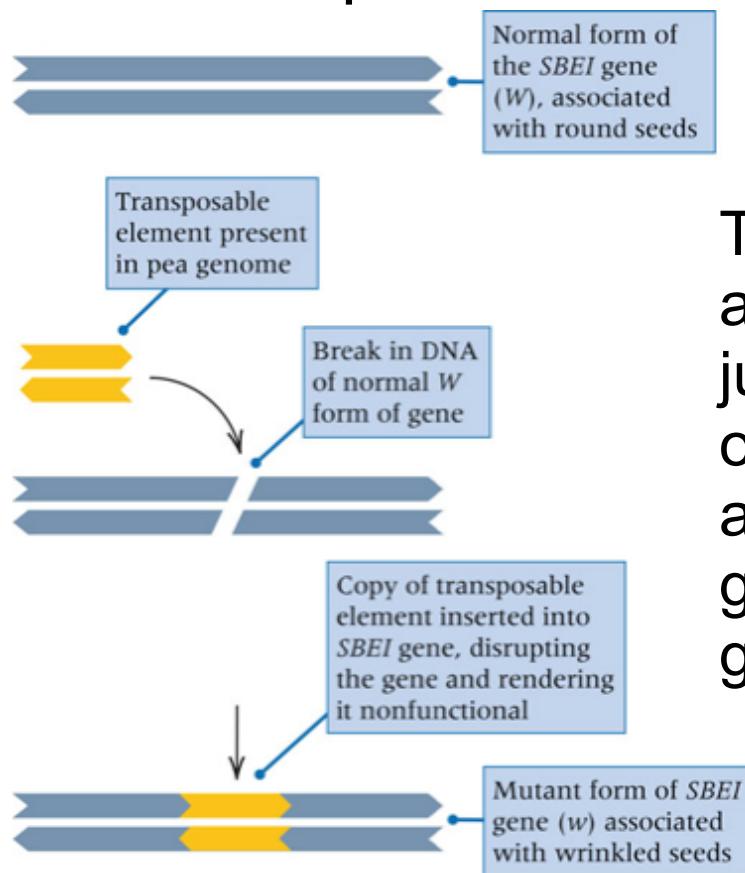
- The Human Genome is Full of Repetitive Elements!



UCSC Genome Browser with VISTA tracks on Human May 2004 Assembly

The Human Genome

- What are Transposable Elements?

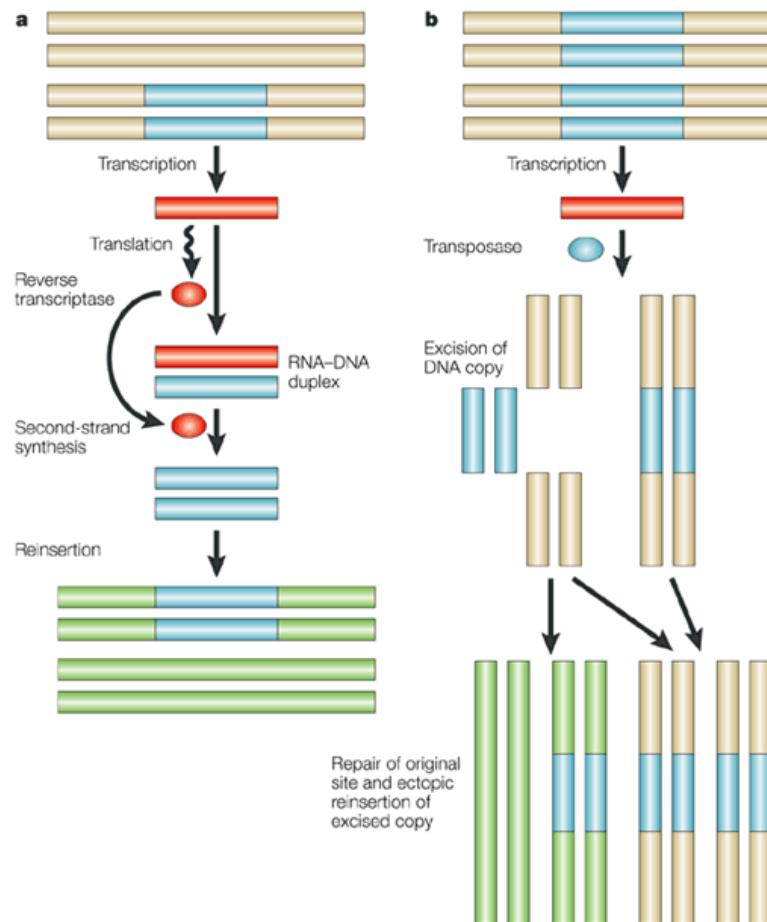


Transposable infectious agents, mobile parasites, that jump (i.e., transpose) from one place of the genome to another, sometimes with grave consequences to the genome!

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

The Human Genome

- How do they Work?



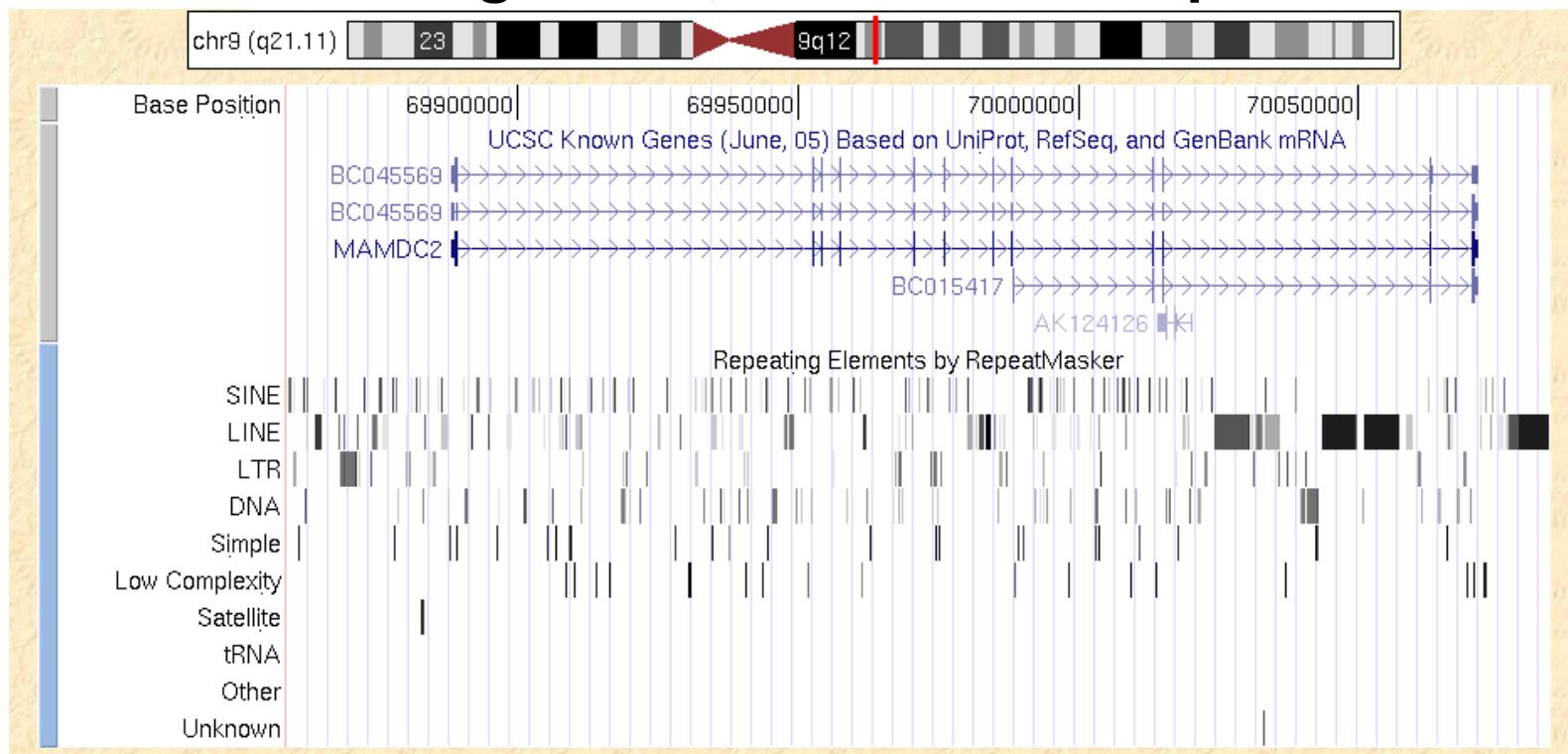
NATURE REVIEWS
GENETICS - Gregory D. D.
Hurst and John H. Werren 2,
597-606 (August 2001)

Nature Reviews | Genetics

Transposons and Us

- Selfish Genetic Elements Live within our own Genes!

Zooming in ~45,000X to ~225 kbp

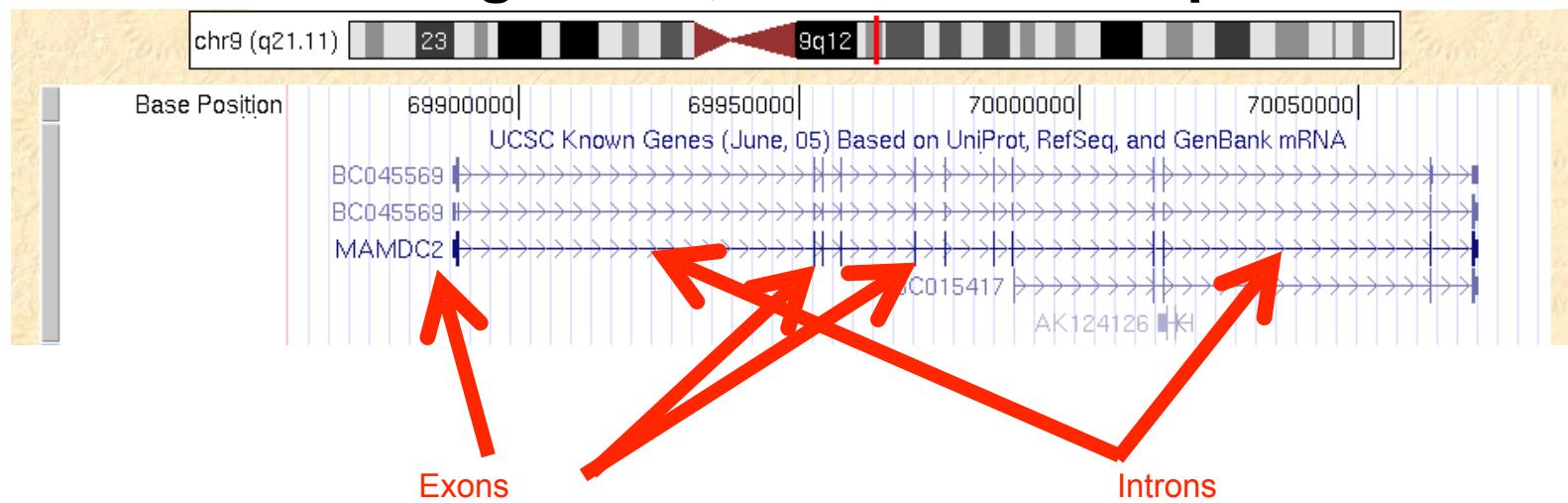


UCSC Genome Browser with VISTA tracks on Human May 2004 Assembly

Transposons and Us

- Selfish Genetic Elements Live within our own Genes!

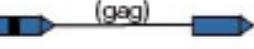
Zooming in ~45,000X to ~225 kbp



UCSC Genome Browser with VISTA tracks on Human May 2004 Assembly

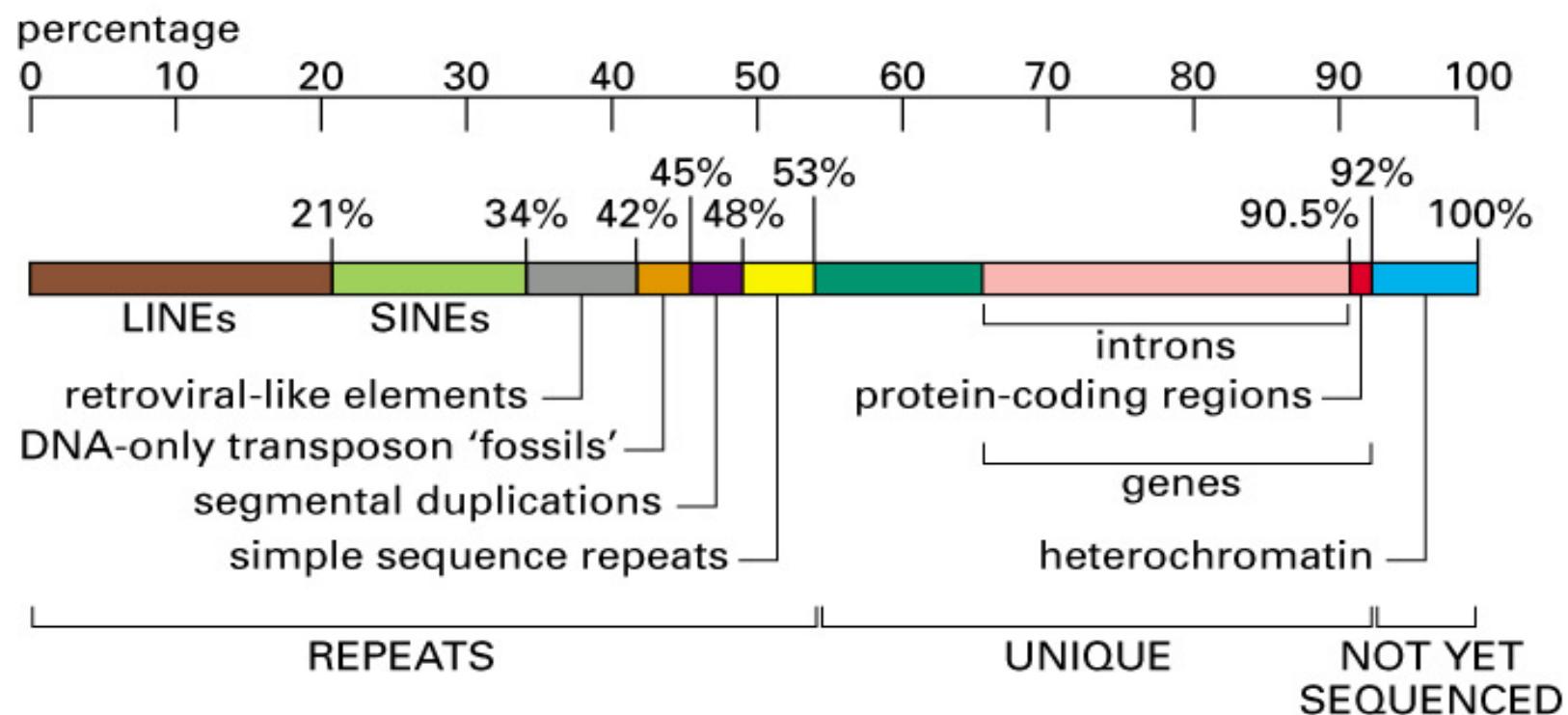
Transposons and Us

- Classes of Interspersed Repeat in Humans

			Length	Copy number	Fraction of genome
LINEs	Autonomous		6–8 kb	850,000	21%
SINEs	Non-autonomous		100–300 bp	1,500,000	13%
Retrovirus-like elements	Autonomous		6–11 kb	450,000	8%
	Non-autonomous		1.5–3 kb		
DNA transposon fossils	Autonomous		2–3 kb	300,000	3%
	Non-autonomous		80–3,000 bp		

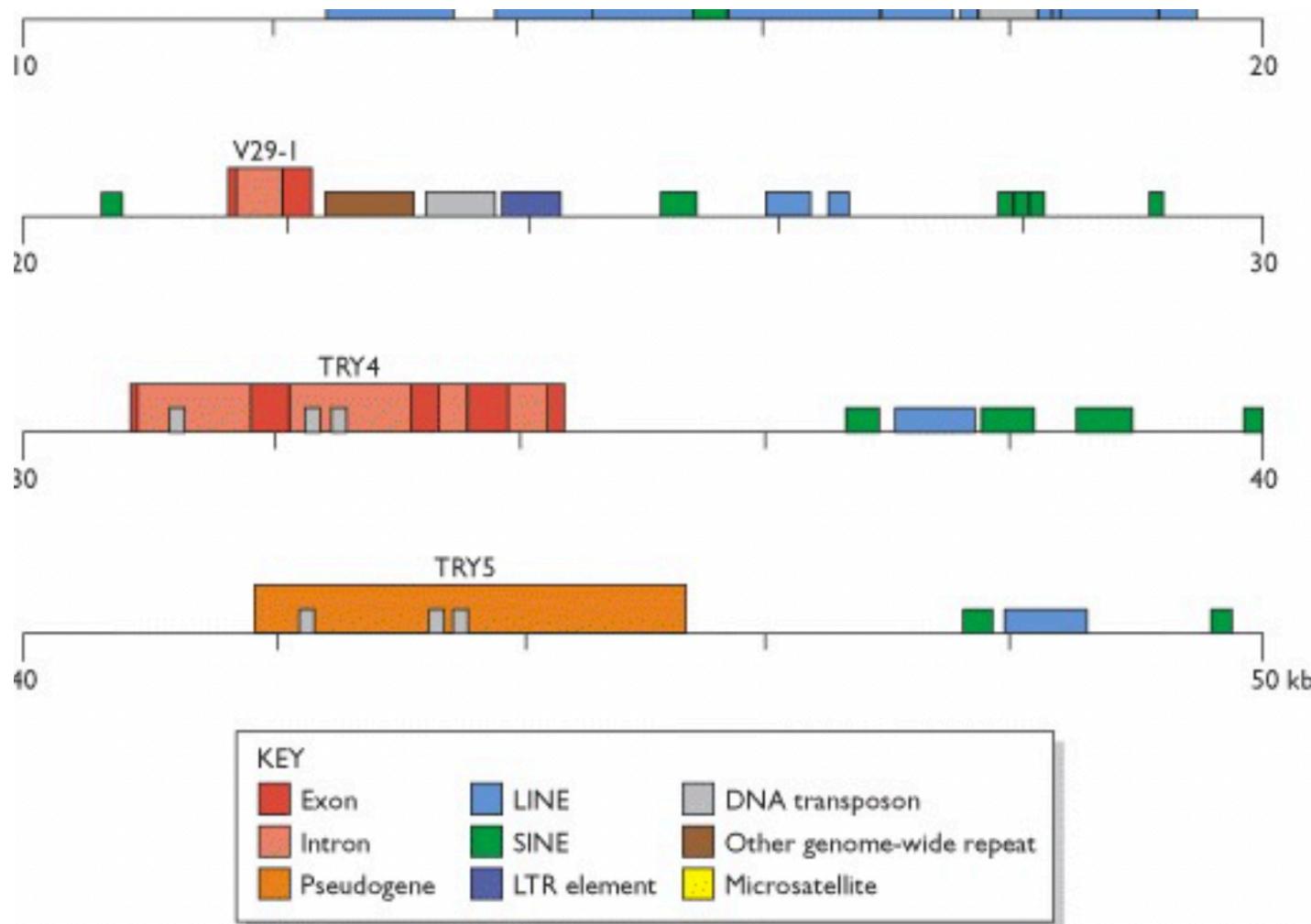
Nature 409, 860921 (15 February 2001)

The Human Genome is a transposon Traffic Jam



The Human Genome

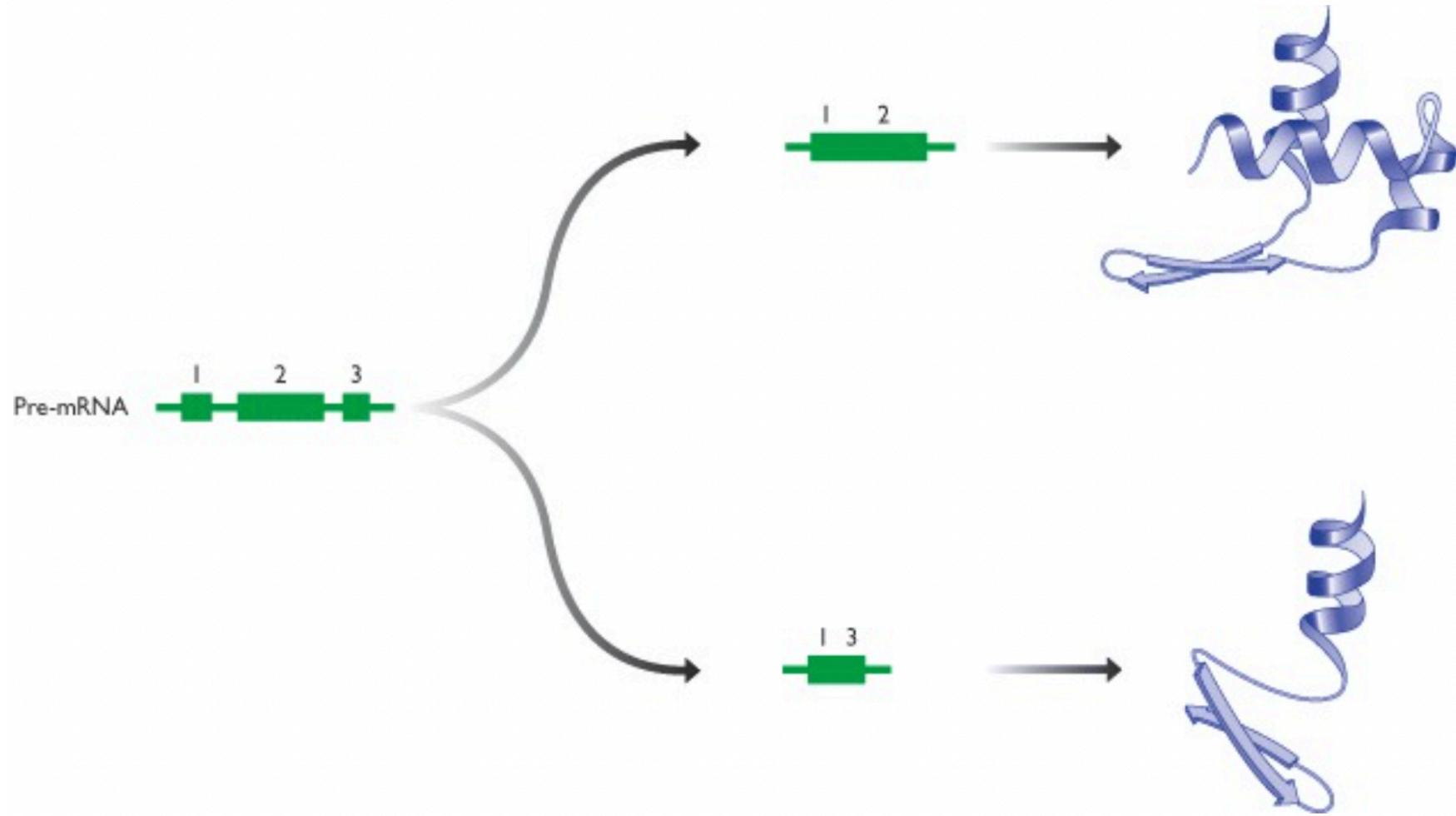
- The Human Genome is Complex!



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).

The Human Genome

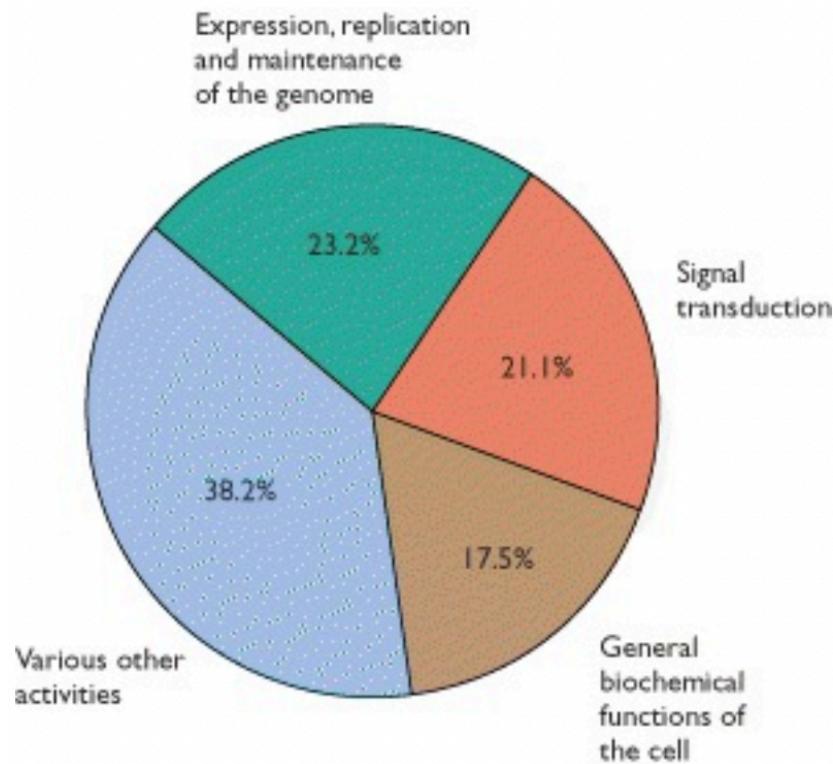
- The Human Genome is Complex!



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

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).

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.

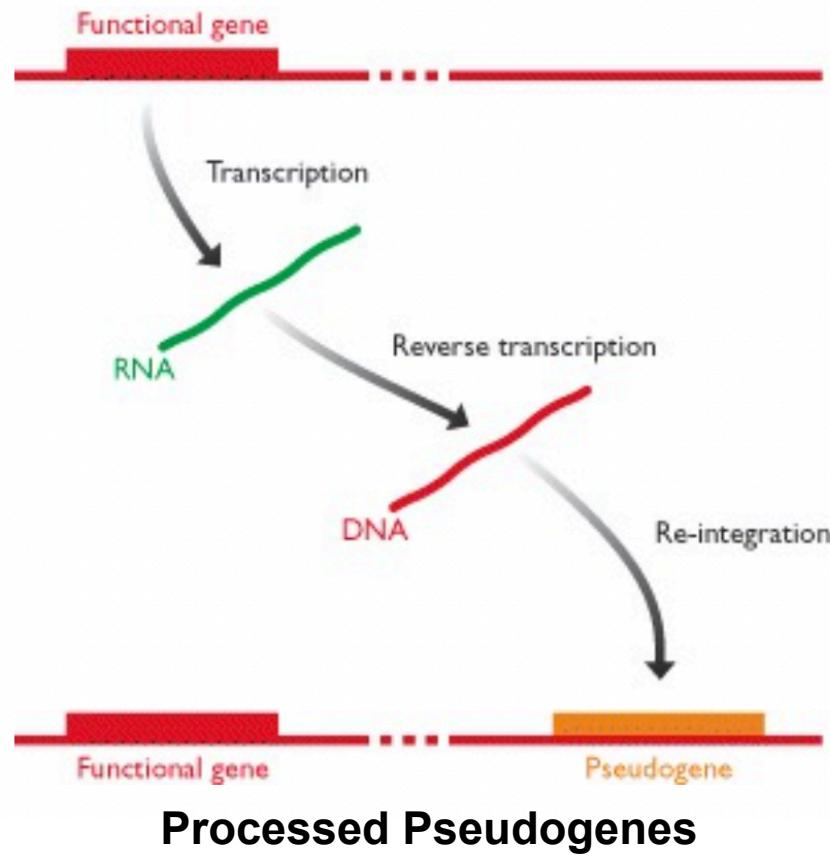
Many mutations have only minor effects on the activity of a gene but some are more important and it quite possible for a single nucleotide change to result in a gene becoming completely non-functional.

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

The Human Genome

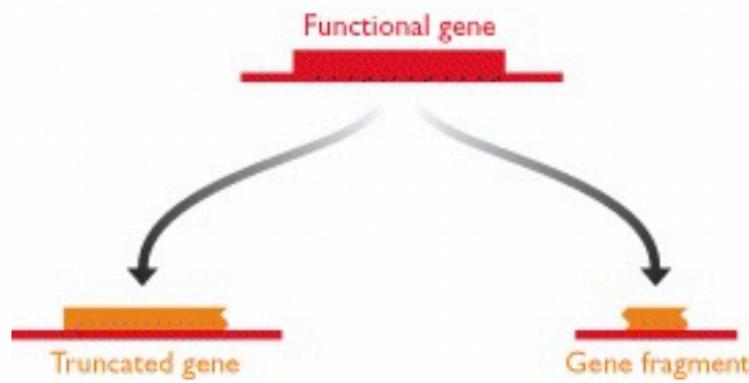
- The Human Genome is Loaded with Evolutionary Relics!



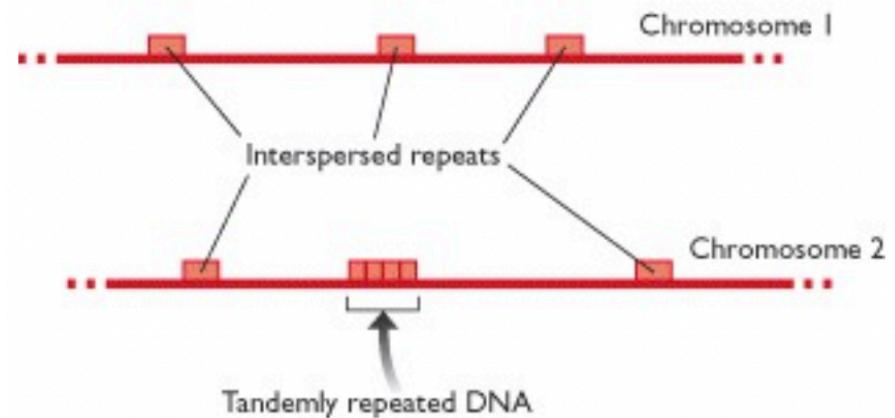
A processed pseudogene is thought to arise by integration into the genome of a copy of the mRNA transcribed from a functional gene. The process by which mRNA is copied into DNA is called reverse transcription and the product is called complementary DNA (cDNA). The cDNA may integrate into the same chromosome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

The Human Genome

- The Human Genome is Loaded with Evolutionary Relics!



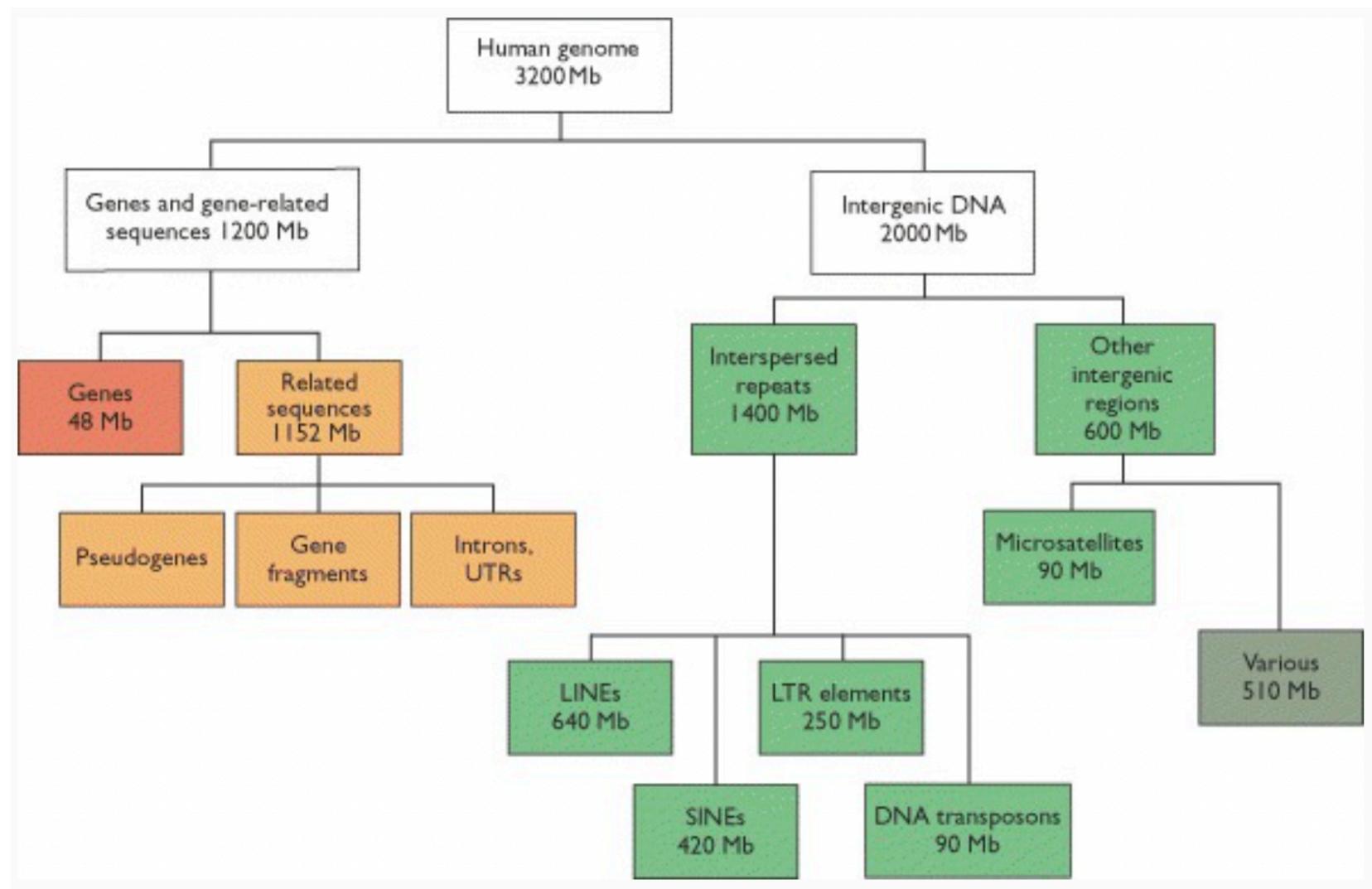
Truncated Genes



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

The Human Genome

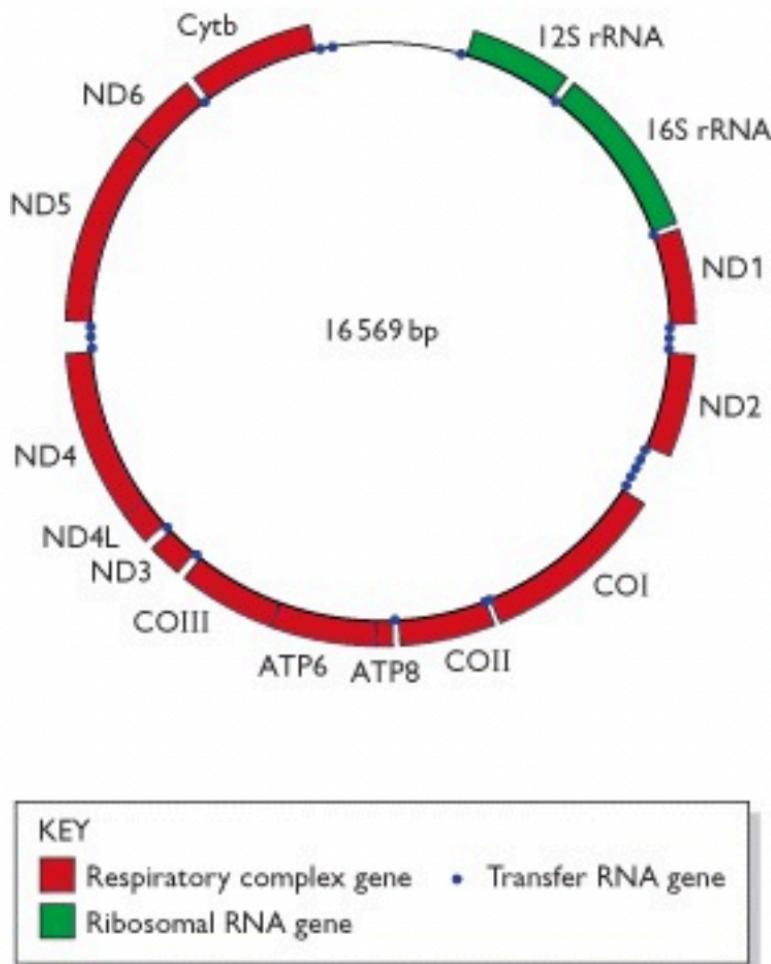
- The Organization of the Human Genome



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

The Human Genome

- The Organization of the Human Genome Mitochondria!



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

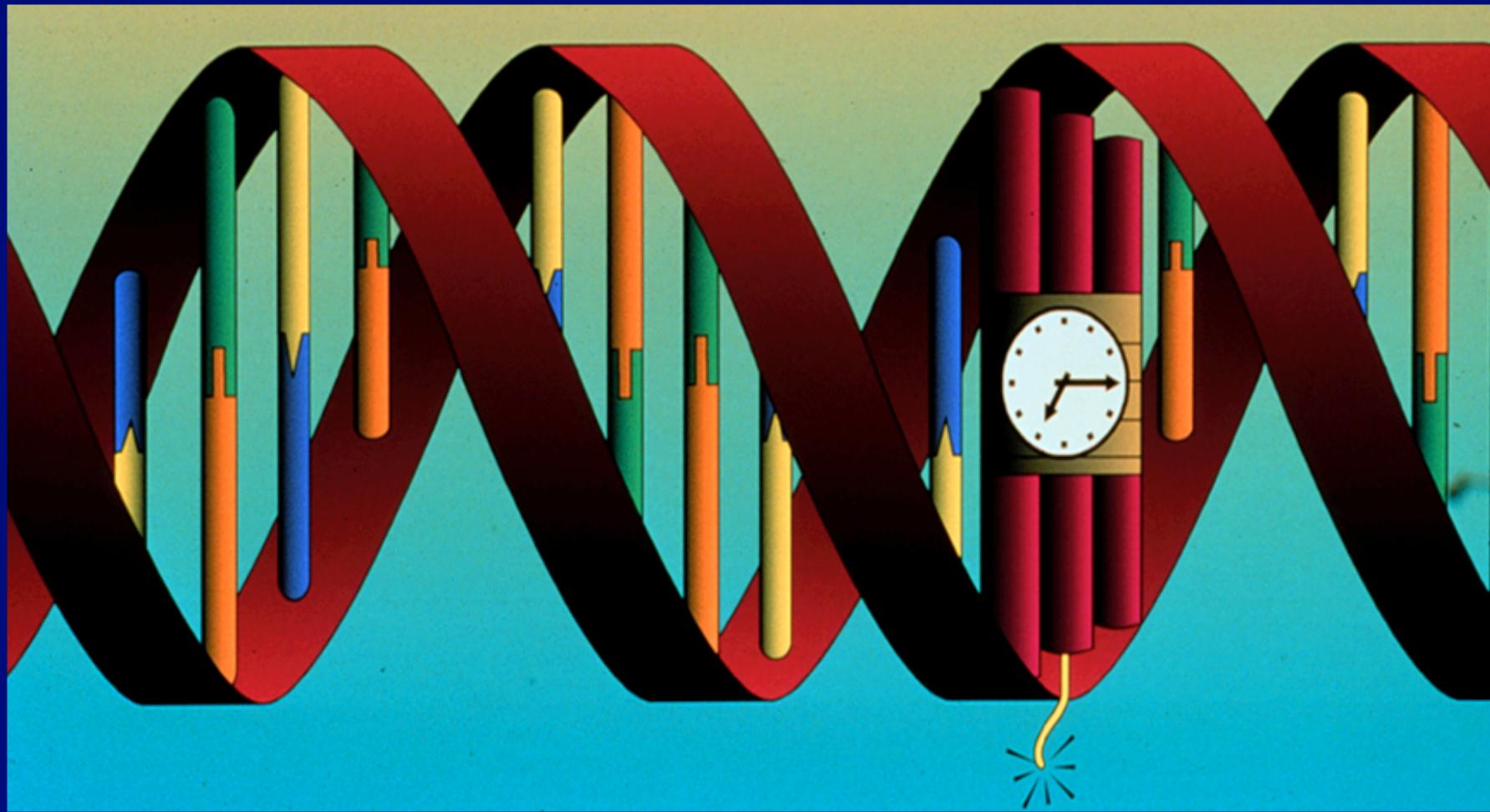
Abbreviations:

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

The human mitochondrial genome

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

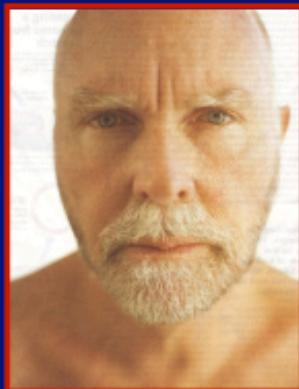


all of us carry a significant number of 'glitches' in our genomes.

Individual Genome Sequences

The Diploid Genome Sequence of an Individual Human

Samuel Levy^{1*}, Granger Sutton¹, Pauline C. Ng¹, Lars Feuk², Aaron L. Halpern¹, Brian P. Walenz³, Nelson Axelrod¹, Jiaqi Huang¹, Ewen F. Kirkness¹, Gennady Denisov¹, Yuan Lin¹, Jeffrey R. MacDonald², Andy Wing Chun Pang², Mary Shago², Timothy B. Stockwell¹, Alexia Tsiamouri¹, Vineet Bafna³, Vikas Bansal³, Saul A. Kravitz¹, Dana A. Busam¹, Karen Y. Beeson¹, Tina C. McIntosh¹, Karin A. Remington¹, Josep F. Abril⁴, John Gill¹, Jon Borman¹, Yu-Hui Rogers³, Marvin E. Frazier¹, Stephen W. Scherer², Robert L. Strausberg¹, J. Craig Venter¹



PLoS Biol (2007)

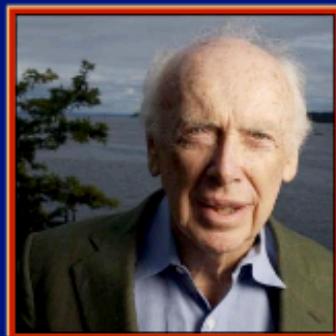
Accurate whole human genome sequencing using reversible terminator chemistry

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

Nature (2008)

The complete genome of an individual by massively parallel DNA sequencing

David A. Wheeler^{1*}, Maithreyan Srinivasan^{2*}, Michael Egholm^{2*}, Yufeng Shen^{1*}, Lei Chen¹, Amy McGuire³, Wen He², Yi-Ju Chen², Vinod Makhijani², G. Thomas Roth², Xavier Gomes², Karrie Tartaro^{2†}, Faheem Niazi², Cynthia L. Turcotte², Gerard P. Irzyk², James R. Lupski^{4,5,6}, Craig Chinault⁴, Xing-zhi Song⁴, Yue Liu¹, Ye Yuan¹, Lynne Nazareth¹, Xiang Qin¹, Donna M. Muzny¹, Marcel Margulies², George M. Weinstock^{1,4}, Richard A. Gibbs^{1,4} & Jonathan M. Rothberg²



Nature (2008)

The diploid genome sequence of an Asian individual

Jun Wang^{1,2,3,4*}, Wei Wang^{1,3,4*}, Ruiqiang Li^{1,3,4*}, Yingrui Li^{1,5,6*}, Geng Tian^{1,7}, Laurie Goodman³, Wei Fan¹, Junqing Zhang¹, Jun Li¹, Juanbin Zhang¹, Yiran Guo^{1,7}, Binxiao Feng¹, Heng Li^{1,8}, Yao Lu¹, Xiaodong Fang¹, Huiqing Liang¹, Zhenglin Du¹, Dong Li¹, Yiqing Zhao^{1,7}, Yujie Hu^{1,7}, Zhenzhen Yang¹, Hancheng Zheng¹, Ines Hellmann², Michael Inouye³, John Pool⁷, Xin Yi^{1,7}, Jing Zhao¹, Jinjie Duan¹, Yan Zhou¹, Junjie Qin^{1,7}, Lijia Ma^{1,7}, Guoqing Li¹, Zhentao Yang¹, Guojie Zhang^{1,7}, Bin Yang¹, Chang Yu¹, Fang Liang^{1,7}, Wenjie Li¹, Shaochuan Li¹, Dawei Li¹, Peixiang Ni¹, Jue Ruan^{1,7}, Qibin Li^{1,7}, Hongmei Zhu¹, Dongyuan Liu¹, Zhike Lu¹, Ning Li^{1,7}, Guangwu Guo^{1,7}, Jianguo Zhang¹, Jia Ye¹, Lin Fang¹, 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 Yang¹, Fang Chen^{1,7}, Li Li¹, Ke Zhou¹, Hongkun Zheng^{1,4}, Yuanyuan Ren¹, Ling Yang¹, Yang Gao^{1,7}, Guohua Yang^{1,2}, Zhuo Li¹, Xiaoli Feng¹, Karsten Kristiansen⁴, Gane Ka-Shu Wong^{1,10}, Rasmus Nielsen⁹, Richard Durbin⁶, Lars Bolund^{3,11}, Xiuqing Zhang^{1,6}, Songgang Li^{1,2,5}, Huanming Yang^{1,2,3} & Jian Wang^{1,2,3}

Nature (2008)

A highly annotated whole-genome sequence of a Korean individual

Jong-II Kim^{1,2,4,5*}, Young Seok Ju^{1,2*}, Hansoo Park^{1,5}, Sheehyun Kim⁴, Seonwook Lee⁴, Jae-Hyuk Yi¹, Joann Mudge⁶, Neil A. Miller⁶, Dongwan Hong¹, Callum J. Bell⁶, Hye-Sun Kim⁴, In-Soon Chung⁴, Woo-Chung Lee⁴, Ji-Sun Lee⁴, Seung-Hyun Seo⁵, Ji-Young Yun⁴, Hyun Nyun Woo⁴, Heewook Lee⁴, 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 Lee⁵, Hyunjun Park¹, Jeong Yeon Kim¹, Omer Gokcumen⁷, Ryan E. Mills⁷, Alexander Wait Zarank⁸, Joseph Thakuria⁸, Xiaodi Wu⁸, Ryan W. Kim⁶, Jim J. Huntley⁷, Shujun Luo⁹, Gary P. Schroth⁷, Thomas D. Wu¹⁰, HyeRan Kim⁶, Kap-Seok Yang⁶, Woong-Yang Park^{1,2,3}, Hyungtae Kim⁶, George M. Church⁸, Charles Lee⁷, Stephen F. Kingsmore⁶ & Jeong-Sun Seo^{1,2,3,4,5}

Nature (2009)

1000 Genomes - Home

1000 Genomes

A Deep Catalog of Human Genetic Variation

Home About Partners Data Contact Wiki

1000 GENOMES PROJECT DATA RELEASE

SNP data downloads and genome browser representing four high coverage individuals

The first set of SNP calls representing the preliminary analysis of four genome sequences are now available to download through the [EBI FTP site](#) and the [NCBI FTP site](#). The README file dealing with the FTP structure will help you find the data you are looking for.

The data can also be viewed directly through the 1000 Genomes browser at <http://browser.1000genomes.org>. Launch the browser and [view a sample region here](#).

More information about the data release can be found in the [data section](#) of this web site.

Download the 1000 Genomes Browser Quick Start Guide

[Quick start \(pdf\)](#)

1000genomes.org



THE CANCER GENOME ATLAS



Search
 GO

[About TCGA](#)[Program Components](#)[Policies](#)[Media Center](#)[Launch Data Portal](#)

I Mission and Goal

The Cancer Genome Atlas (TCGA) is a comprehensive and coordinated effort to accelerate our understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing.

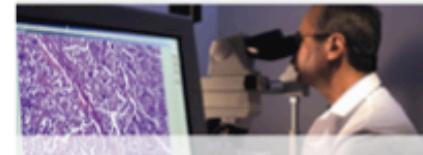
[Learn more >>](#)

I News from the Pilot Project

NEW* TCGA Network Identifies More Than 6,000 Targets for Sequencing

The Cancer Genome Atlas (TCGA) network has selected more than 6,000 gene and miRNA targets for sequencing that represent both protein-coding genes and microRNAs (miRNAs). While not exhaustive, this list represents genes and

I TCGA Data Portal



■ [Access TCGA Data Portal](#)
■ [New* View](#) the list of target genes and miRNAs selected by the TCGA network for sequencing.

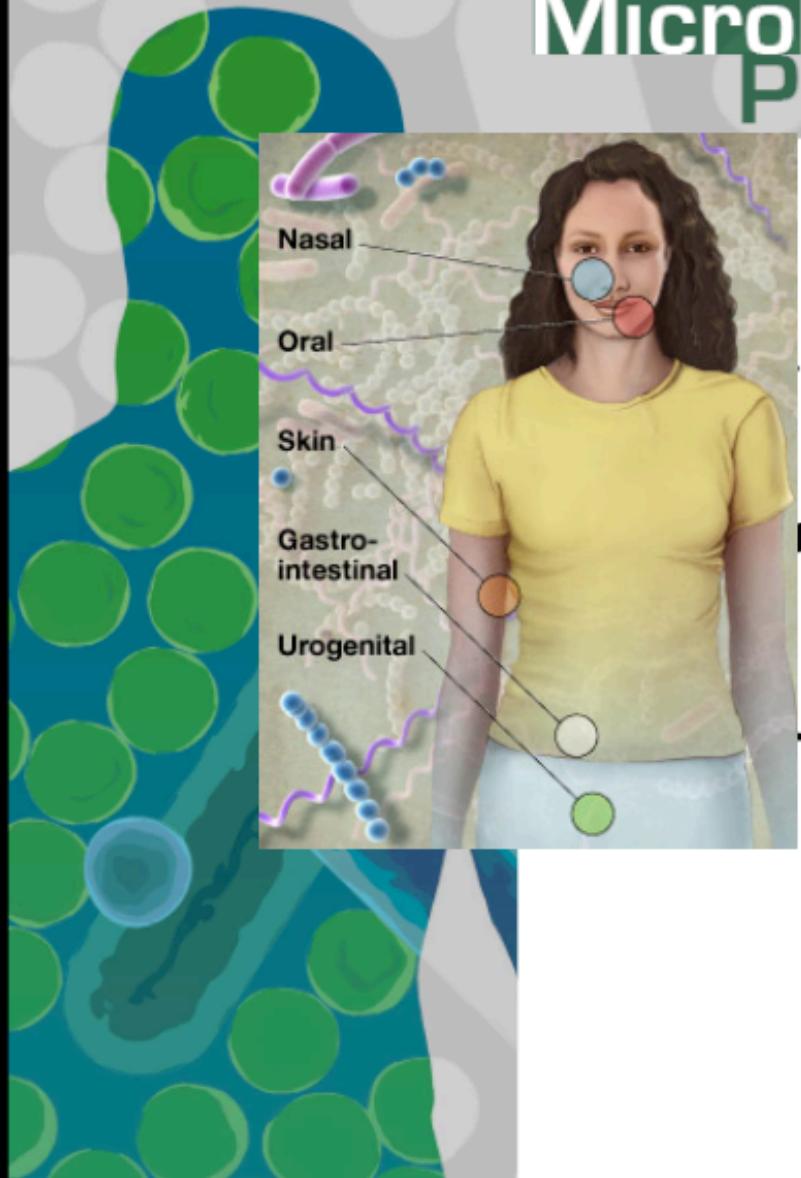
■ [NEW Data Available* View](#) Molecular Characterization Data for Ovarian Cancer.

I TCGA: How Will It Work?



■ [Click here for more information](#)

The Human Microbiome Project



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 \(Pairwise Sequence Comparison\)](#)

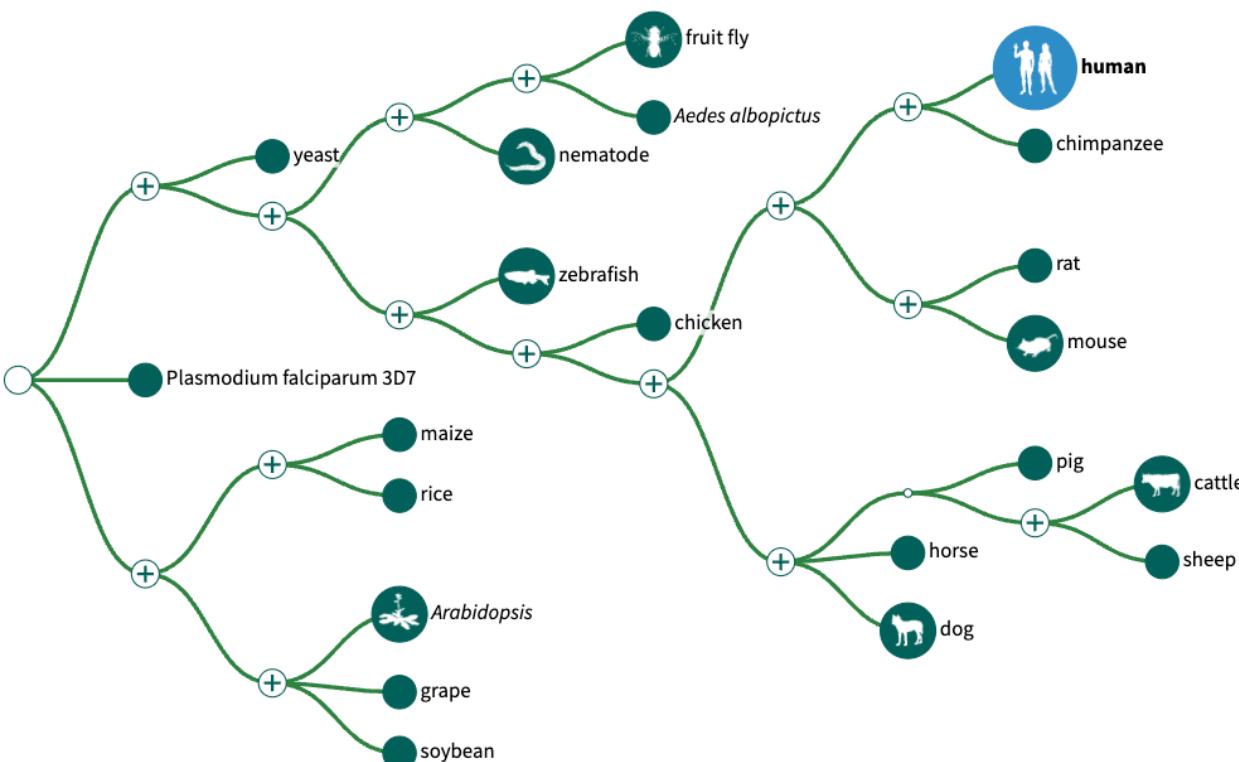
External Resources

[GOLD - Genomes Online Database](#)

[Bacteria Genomes at Sanger](#)

[Ensembl](#)

Genome Data Viewer



Homo sapiens (human)



Search in genome
Location, gene or phenotype Q

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

Assembly
GRCh38.p13

Browse genome BLAST genome

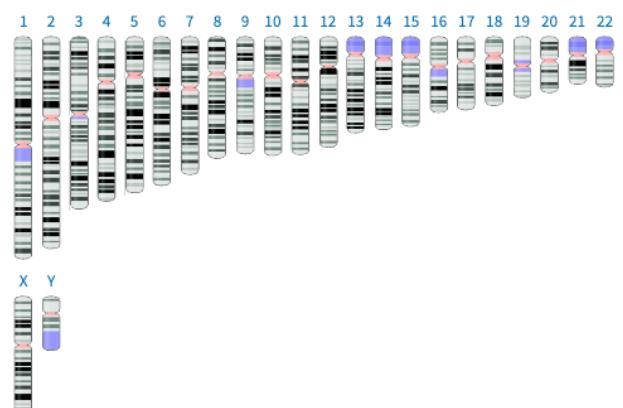
Download via NCBI Datasets

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 i
Release date	Nov 21, 2021



Human Resources at NCBI

Search for Human Genes

Search

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y MT

Select a chromosome to access the [Genome Data Viewer](#)

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

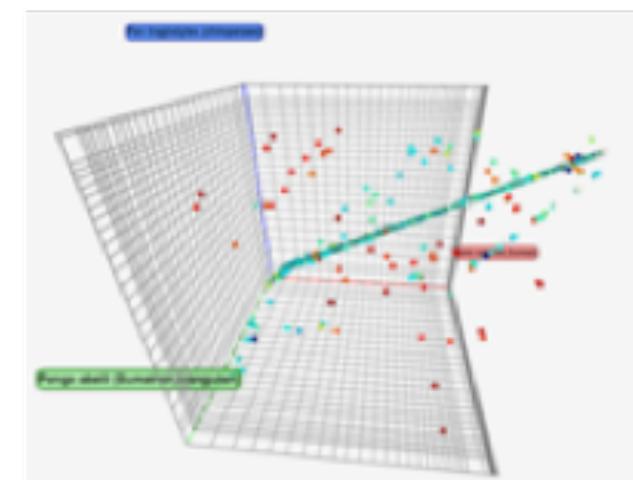
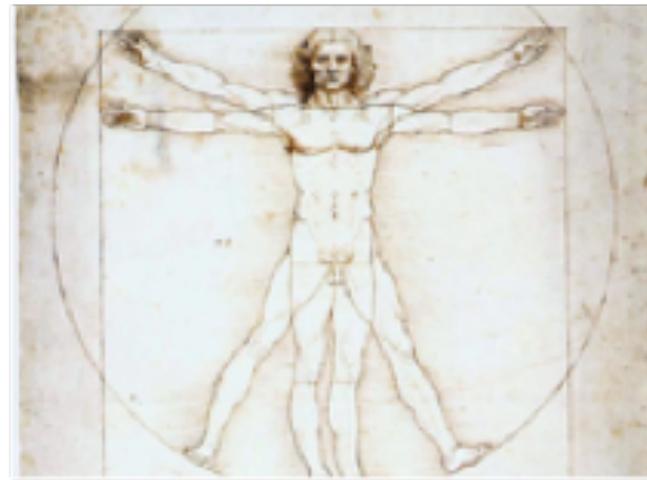
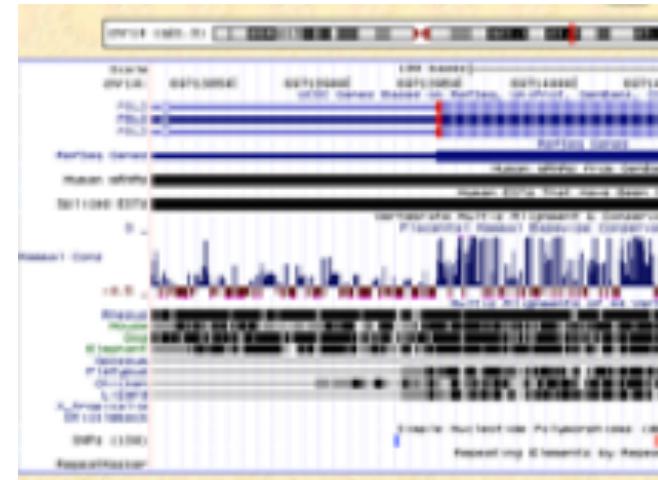
dbVar

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)



THE VERTEBRATE GENOMES PROJECT
IS A PROJECT OF THE G10K CONSORTIUM



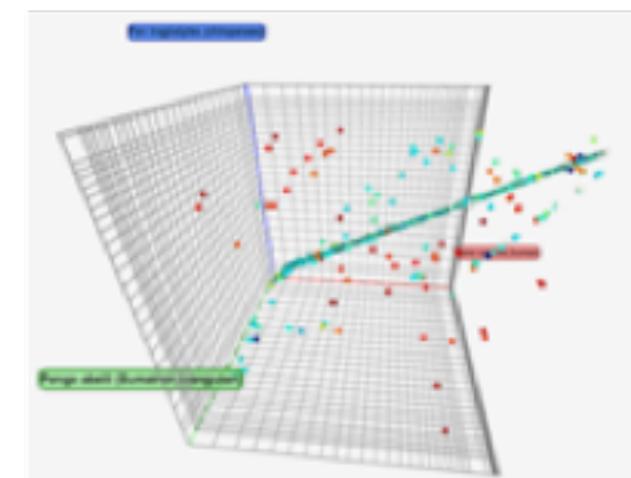
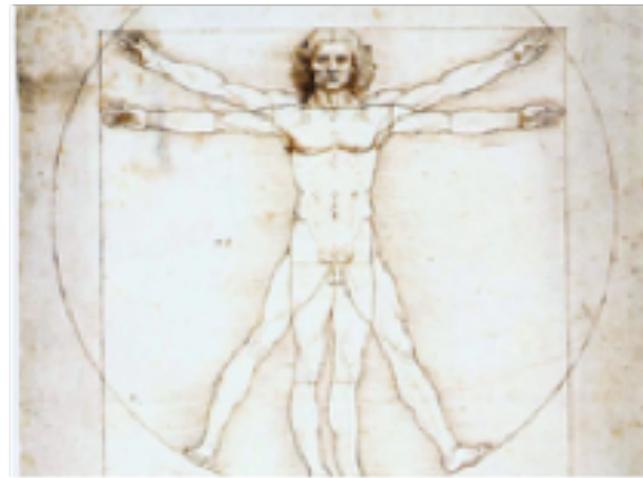
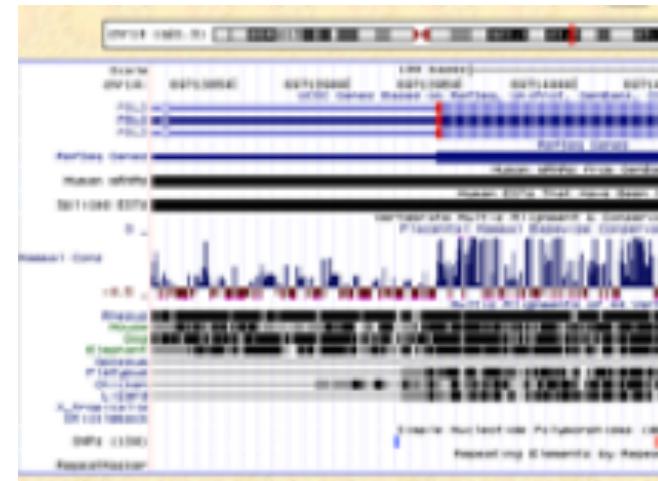
EARTH BIOGENOME PROJECT

sequencing life for the future of life

<https://www.earthbiogenome.org/>

Computational Genomics

A Note About Computational Biology



Realities of New DNA Sequencing Technologies...





62,789

samples

1112

projects

2020

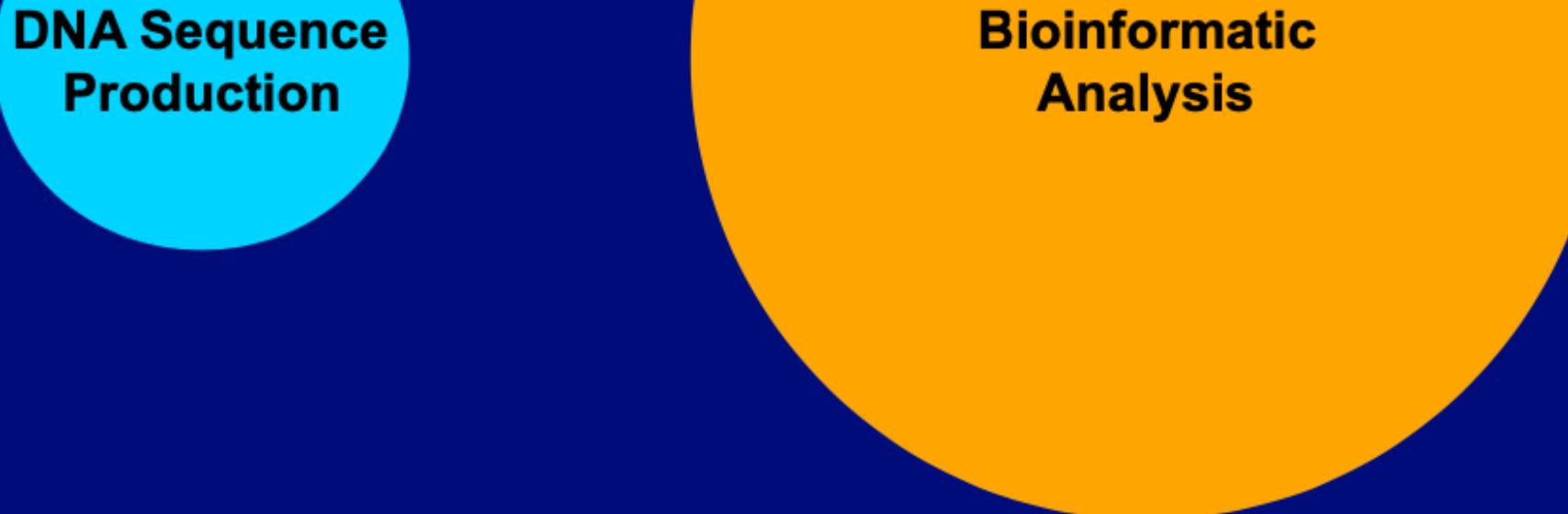
765.3

$\times 10^{12}$ base pairs

1X

Human genome
every 2 minutes

Changing Infrastructure Requirements



DNA Sequence
Production

Bioinformatic
Analysis

The Computational Bottleneck

