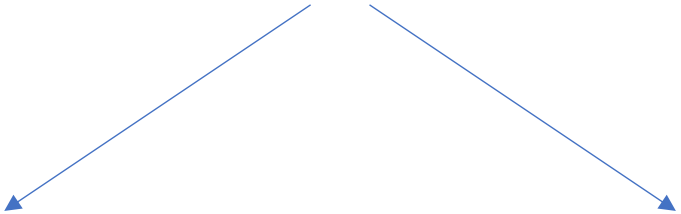
Gene Expression regulation in Eucaryotes



Module I

Prof. Mariangela Morlando

Istituto di Fisiologia Generale
Piano seminterrato stanza 31
Tel: 0649912341

Student office hours: Tuesday
10:00-11:00

mariangela.morlando@uniroma1.it

Module II

Prof. Gaia Di Timoteo

Gene Expression regulation in Eucaryotes module I and II

Genome complexity
Epigenetics
Transcription, pervasive transcription
RNA maturation: transcription initiation and capping
RNA maturation: splicing /Alternative splicing/ sex determination in Drosophila
RNA maturation: polyadenylation, alternative polyadenilation
Transcription termination
RNA transport and localization
RNA Decay and Quality check
Translation and its regulation
Small non-coding RNAs and their functions
Long non-coding RNAs and their functions
Circular RNAs and their functions
Membranless organelles
The epitranscriptome
Genomic imprinting

Methodologies: PCR and its applications, Nucleic acid labelling, RNA/DNA-protein interaction, RNA/RNA/DNA inteaction, Model systems, Imaging, Bioinformatic basis
Genome editing approaches

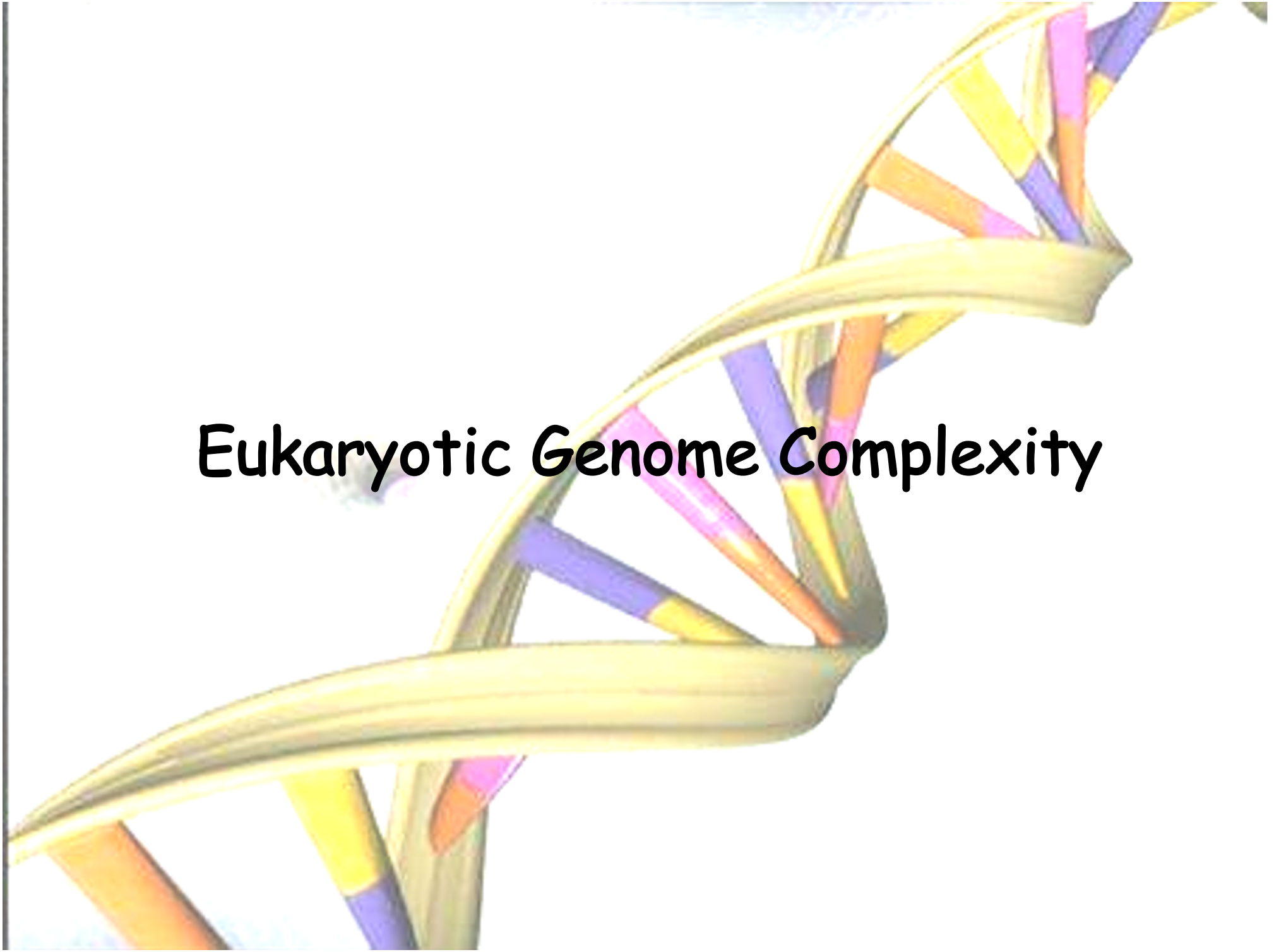
Recommended textbooks

Jordanka Zlatanova Kensal E. van Holde
Biologia molecolare
Struttura e dinamica di genomi e proteomi
Edizione italiana a cura di Vito De Pinto
2018

PREMIO NOBEL

James D Watson Tania A Baker Stephen P Bell Alexander Gann Michael Levine Richard Losick
Biologia molecolare del gene
Ottava edizione italiana a cura di Paolo Plevani
2022

Teaching materials are available here: https://elearning.uniroma1.it/



Eukaryotic Genome Complexity

Catalyzing discoveries in biological regulation

1953 Resolution of DNA structure (Nobel price Watson/Crick)

Implications regarding the mechanisms of DNA replication and gene expression

1973 DNA cloning and DNA sequencing (Nobel price Gilbert/Sanger)

Definition of gene structure
- molecular definition of several pathologies

1987 PCR (Nobel price Mullis)

Huge improvement in DNA and gene expression analysis

1990-2001 Genome sequencing

Identification of **complex functions** and analysis of multifactorial diseases.

GENOME: The total genetic information of a cell or organism *(Lodish – Molecular Cell Biology)*

GENE: The physical and functional unit of inheritance, which carries information from one generation to the next

IN MOLECULAR TERMS, IT IS THE ENTIRE DNA SEQUENCE (INCLUDING EXONS, INTRONS, AND NON-CODING TRANSCRIPTIONAL CONTROL REGIONS) NECESSARY FOR THE PRODUCTION OF A FUNCTIONAL PROTEIN OR RNA. ***(Da Lodish – Molecular Cell Biology)***

Genome

- The genome is all the DNA in a cell.
  - All the DNA on all the chromosomes
  - Includes genes, intergenic sequences, repeats

- Eukaryotes can have 2-3 genomes
  - Nuclear genome
  - Mitochondrial genome
  - Plastid genome

- If not specified, "genome" usually refers to the nuclear genome.

The Genomic Era

- ***Genomics*** is the study of genomes, including large chromosomal segments containing many genes.

- The *initial phase of genomics* aims to map and sequence an initial set of entire genomes.

- *Functional genomics* aims to deduce information about the function of DNA sequences.

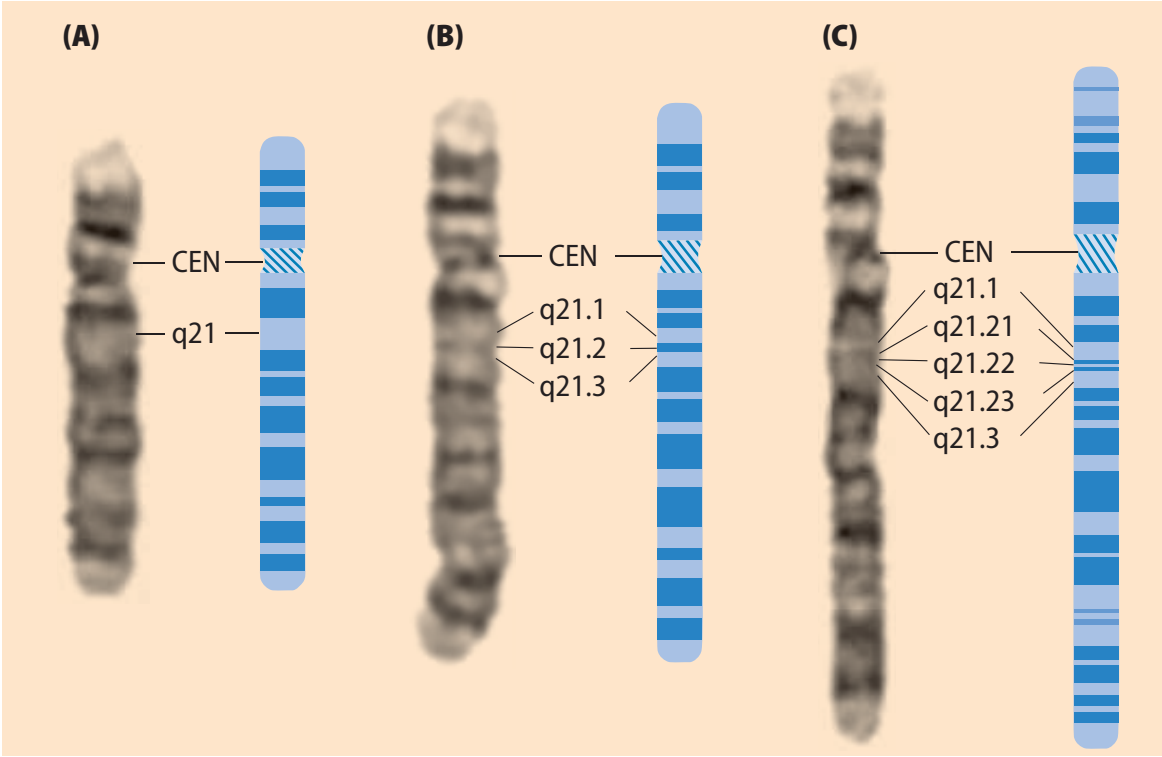
- *Comparative genomics* aims to compare genomic sequences to determine functional or evolutionary relationships between genomes from different organisms

one of the major question was:
how many genes are present in the genome?

For decades, the only available map of the nuclear genome was a low-resolution physical map that was based on chromosome banding.

The most commonly used method in human chromosome banding is **G-banding**.

The chromosomes are treated with trypsin and stained with Giemsa, which preferentially binds AT-rich regions, producing alternating dark bands (Giemsa-positive; AT-rich) and light bands (Giemsa-negative; GC-rich). Because genes are preferentially associated with GC-rich regions, dark bands in G-banding are gene-poor;



Milestones in genome research:

- 1953 – Watson and Crick discover the structure of DNA.
- 1961 – The genetic code for protein synthesis is understood.
- 1977 – Development of the Sanger sequencing method.
- 1990 – Launch of the Human Genome Project.
- 1995 – Sequencing of the first bacterial genome (Haemophilus influenzae).
- 2000 – Sequencing of the Drosophila genome.
- 2001 – Release of the first draft of the human genome.
- 2003 – Completion of the Human Genome Project.

the Human Genome Project (HGP)

International Human Genome Sequencing Consortium:

20 Institutions involved worldwide from 6 countries (China, france, genrmay, jaoan, UK and USA)

The total cost of the project was 3 billion dollars.

The clickable genome

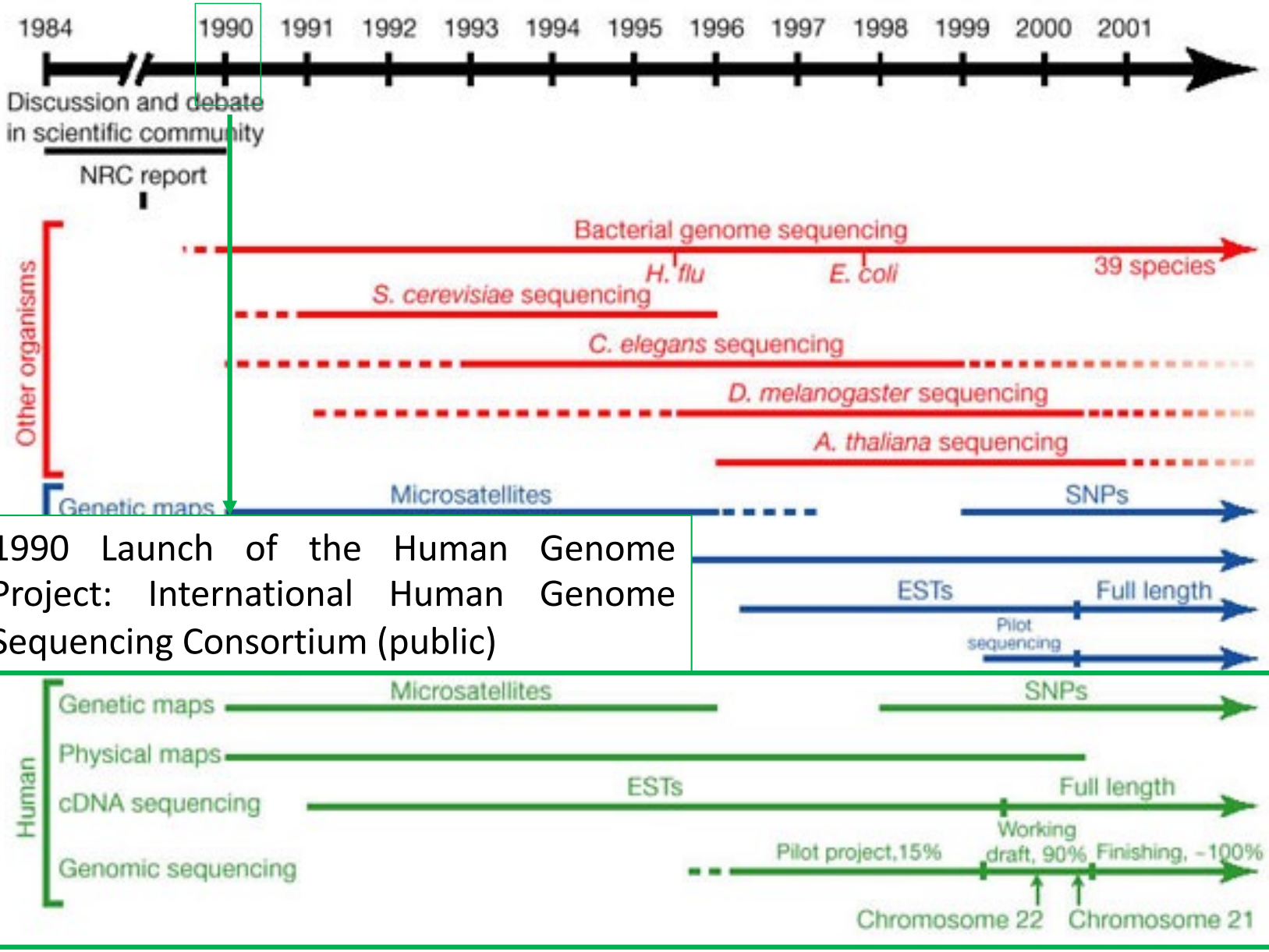
When the assembled draft sequence of the human genome was published in February 2001, it was made available through three public portals: Ensembl, the University of California, Santa Cruz Genome Browser (UCSC) and the NCBI Map Viewer.

Actually, the genomic sequence is continuously refined, both in terms of closing "gaps" and improving its structure (we are currently at the 20th release, known as GRCh38).

Objectives of the Human Genome Project:

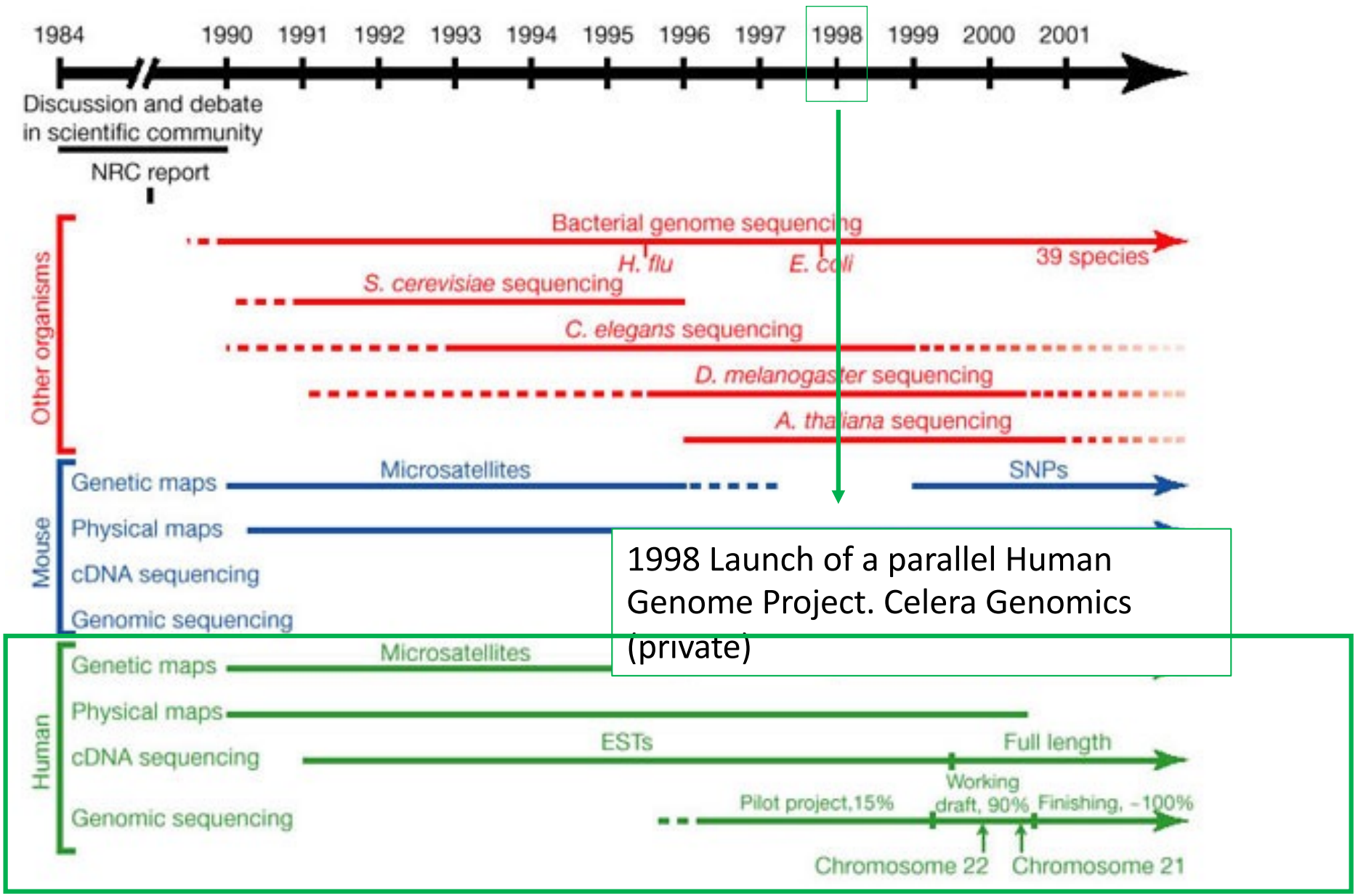
- Generate the sequence of the human genome.
- Identify the genes contained within it.
- Store this information in a database.
- Improve tools for data analysis.
- Transfer technologies to the private sector.
- Address the ethical, legal, and social issues that may arise from the project.

Timeline of large-scale genomic analyses

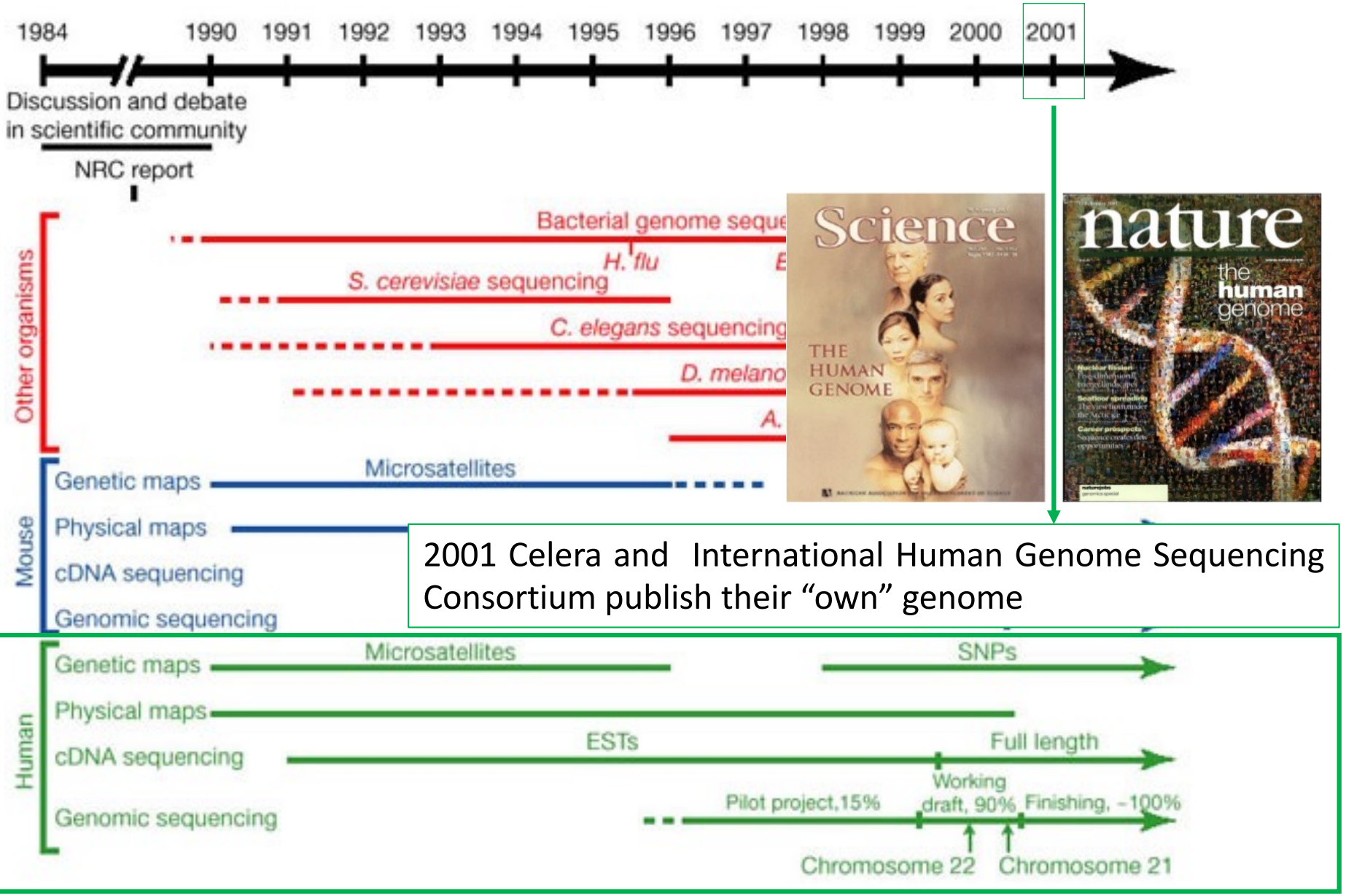


1990 Launch of the Human Genome Project: International Human Genome Sequencing Consortium (public)

Timeline of large-scale genomic analyses

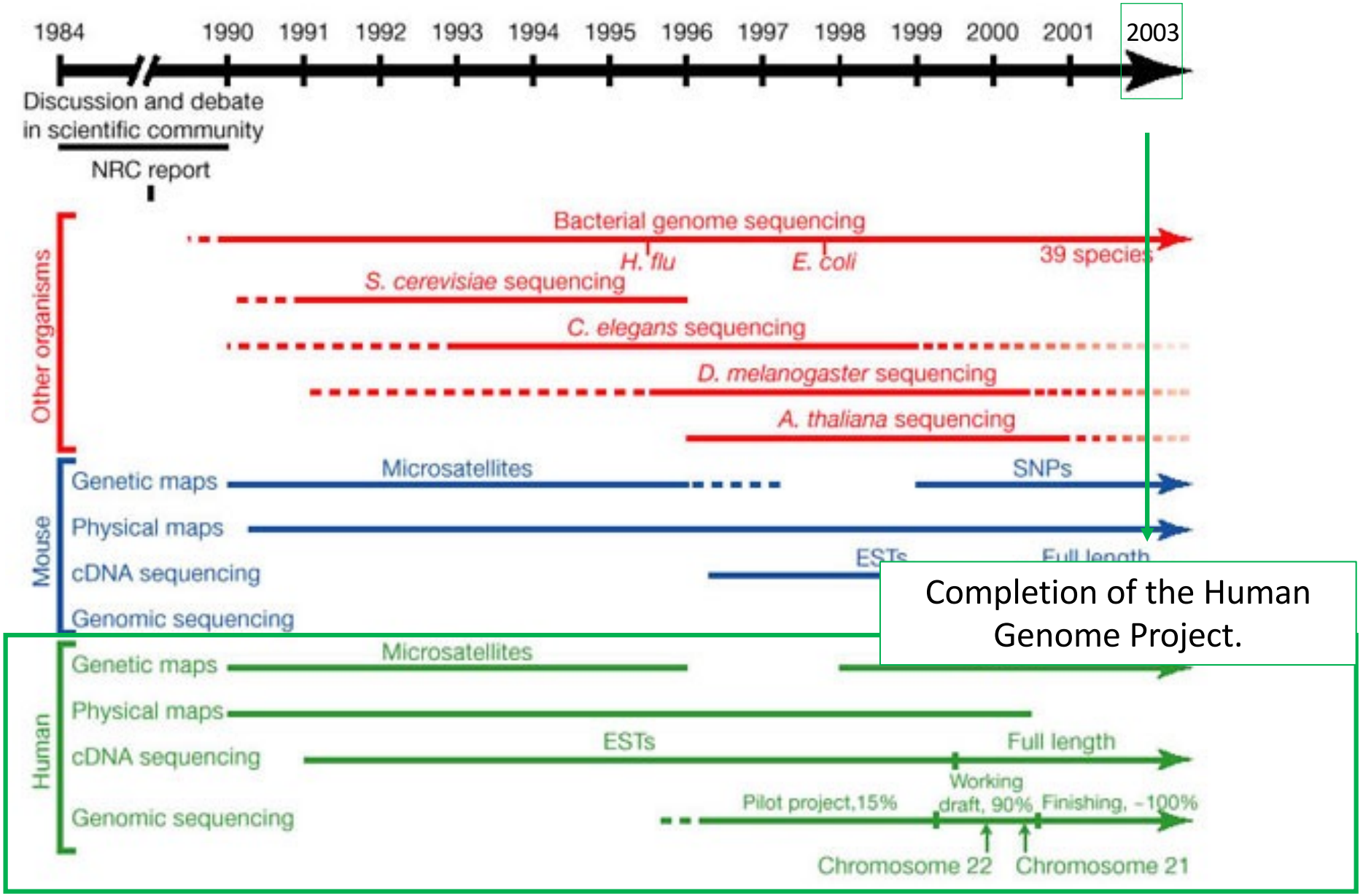


Timeline of large-scale genomic analyses



2001 Celera and International Human Genome Sequencing Consortium publish their "own" genome

Timeline of large-scale genomic analyses



Genomes

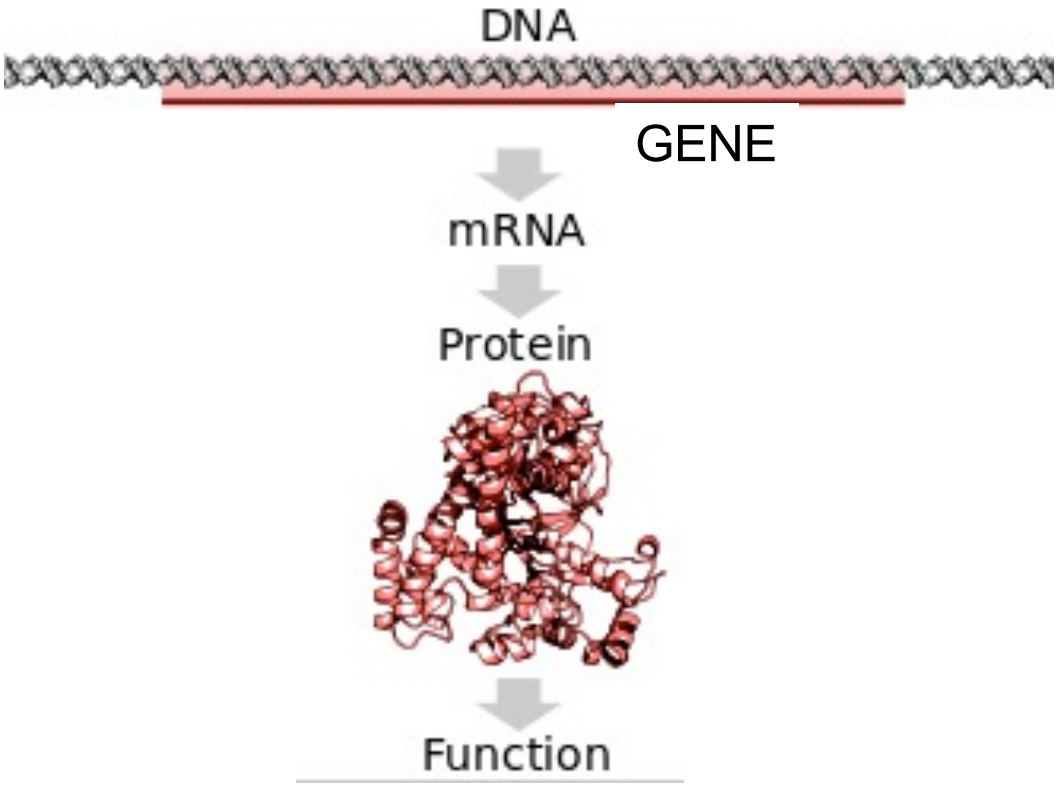
Genome sequencing is often compared to "decoding," but a sequence is still very much in code. In a sense, a genome sequence is simply a very long string of letters in a mysterious language.

lgvrdlmnqvtthequickababcmfxlqbrownfoxjulrvsmped
overthelazyyyzplfdogjjiurttiythedoglayhhbeldquietly
dreaminghwwiqldnsofdinnerplwosiucnd

lgvrdlmnqvt**thequick**ababcmfxlq**brownfoxju**lrvs**mped**
overthelazyyyzplf**dog**jjiurttiy**thedoglay**hhbeld**quietly**
dreaminghwwiqldns**ofdinner**plwosiucnd

Initial definition of GENE

gene: the basic physical unit of heredity; a linear sequence of nucleotides along a segment of DNA that provides the coded instructions for **synthesis of RNA**, which, when translated into protein, leads to the expression of hereditary character.



After completing the human genome we faced 3 Gigabytes of this

Not immediately apparent where the genes are…

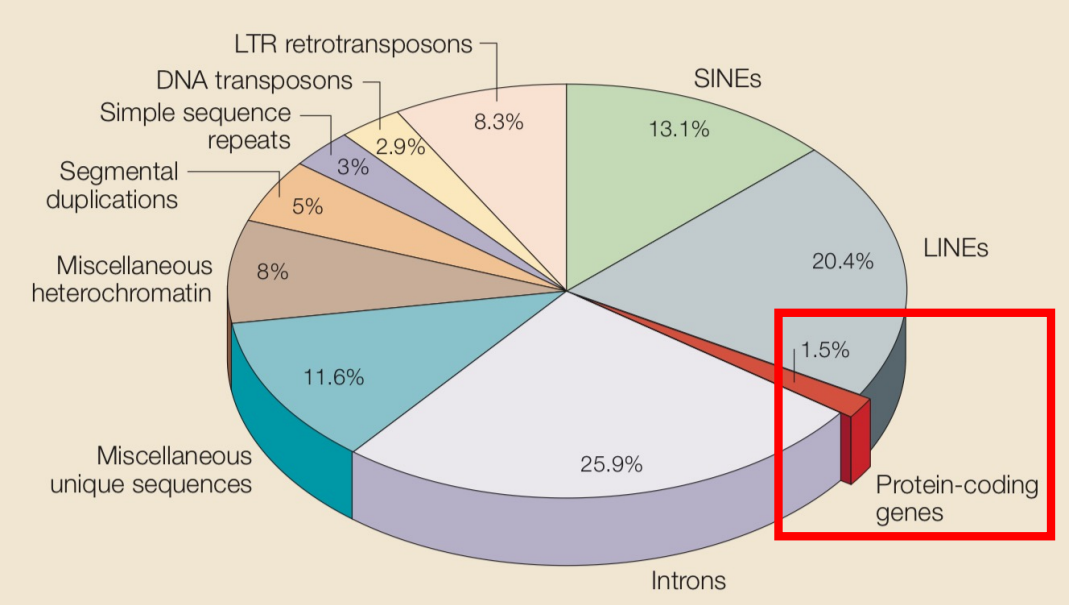
The main elements of the human genome:

- 3.2 billion bases for aploid genome

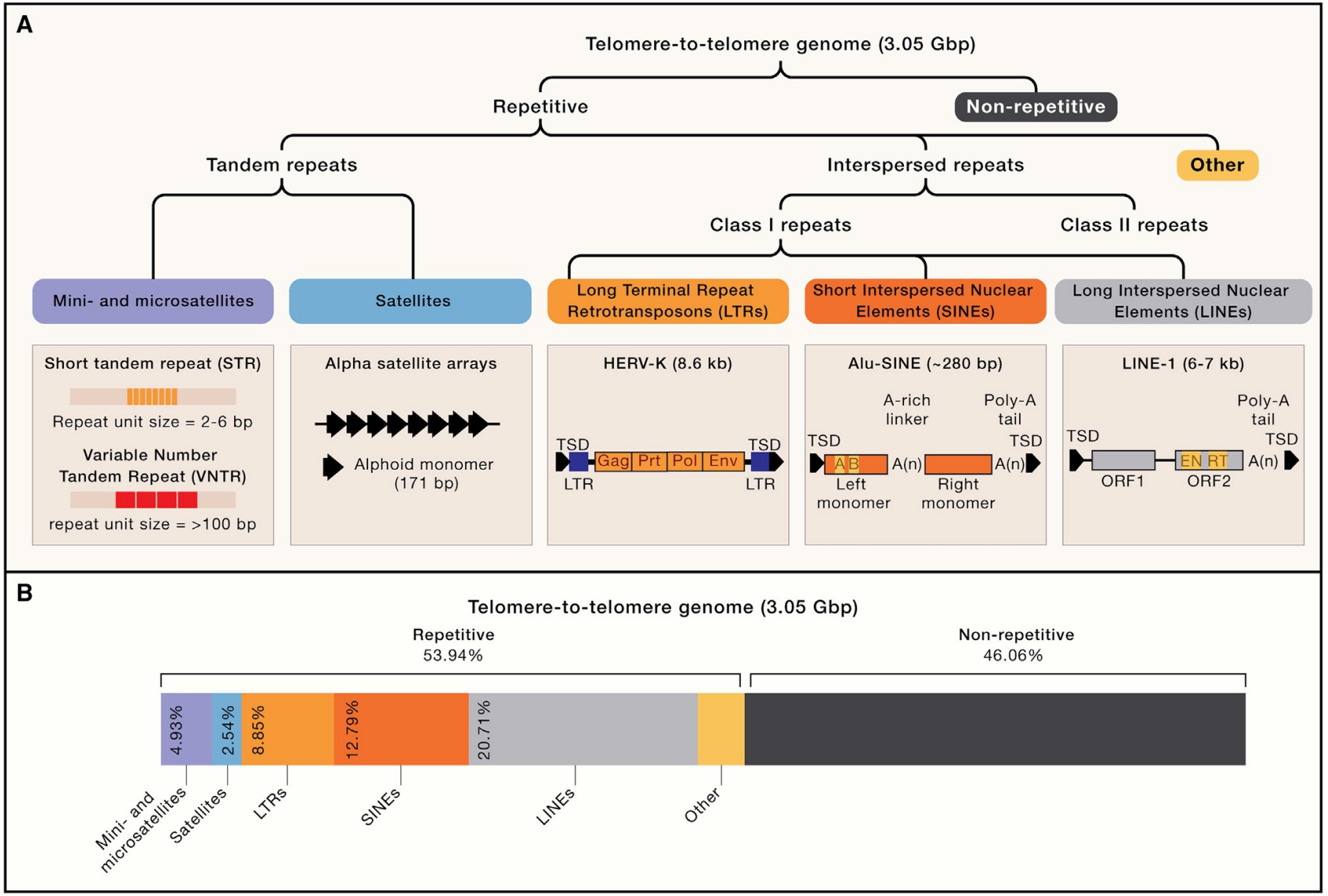
- Distributed across 22 autosomes plus the sex chromosomes

- Protein-coding sequences constitute only about 1-2% of the human genome (approximately 20,000-26,000 genes)

- Repeated sequences make up more than 50% of the genome



Repeated sequences



How many genes in human genome?

- 2000: must be at least 100000 (Rice has ~40,000, *C. elegans* has ~19,000)
- 2003: human genome sequenced completed: 20687 genes

There is a remarkable lack of correspondence between genome size and organism complexity, especially among eukaryotes

Species and Common Name	Estimated Total Size of Genome (bp)*	Estimated Number of Protein-Encoding Genes*
Saccharomyces cerevisiae (unicellular budding yeast)	12 million	6,000
Trichomonas vaginalis	160 million	60,000
Plasmodium falciparum (unicellular malaria parasite)	23 million	5,000
Caenorhabditis elegans (nematode)	95.5 million	18,000
Drosophila melanogaster (fruit fly)	170 million	14,000
Arabidopsis thaliana (mustard; thale cress)	125 million	25,000
Oryza sativa (rice)	470 million	51,000
Gallus gallus (chicken)	1 billion	20,000–23,000
Canis familiaris (domestic dog)	2.4 billion	19,000
Mus musculus (laboratory mouse)	2.5 billion	30,000
Homo sapiens (human)	2.9 billion	20,000–25,000

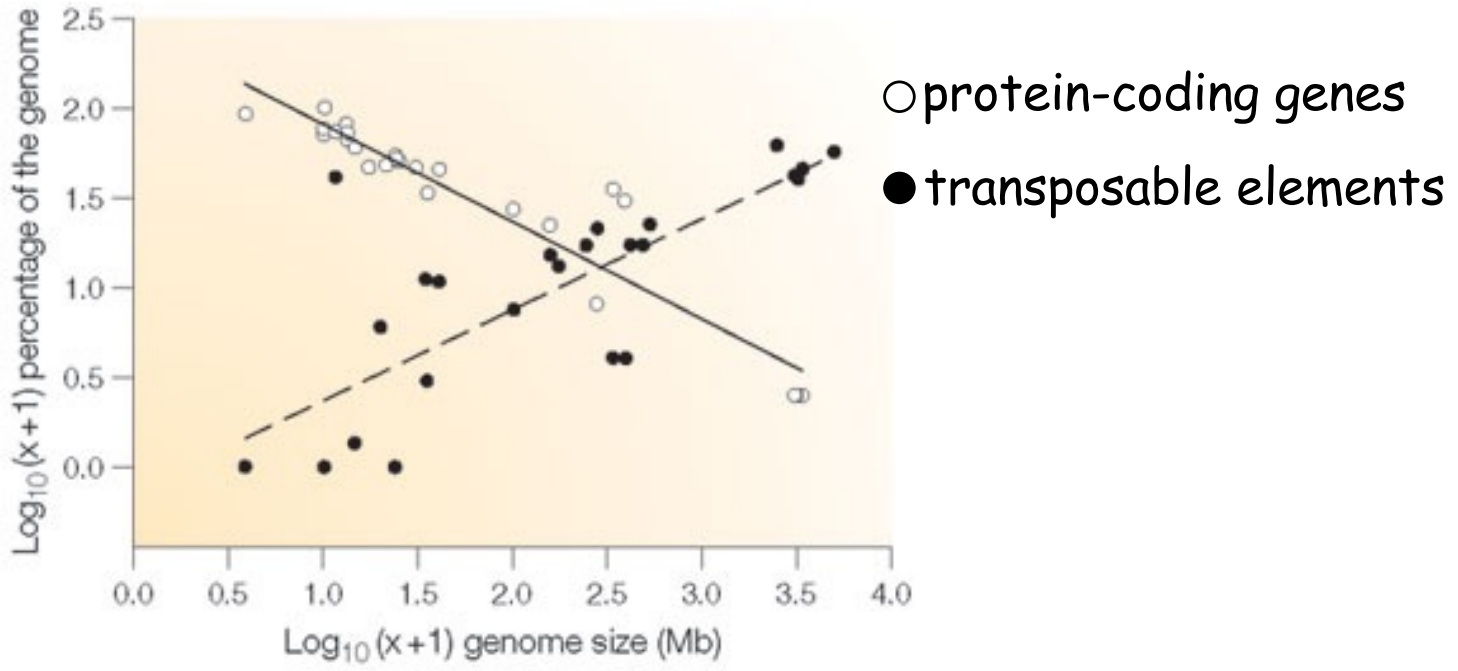
Where is the information that programs our complexity?

- For many years it was believed that the complexity of a given organism was defined solely by the number of proteins encoded in its genome.

- The biggest surprise of the genome sequencing projects was the discovery that the **number of protein-coding genes does not scale strongly or consistently with complexity**.

Genome size and total transposable-element content are strongly correlated

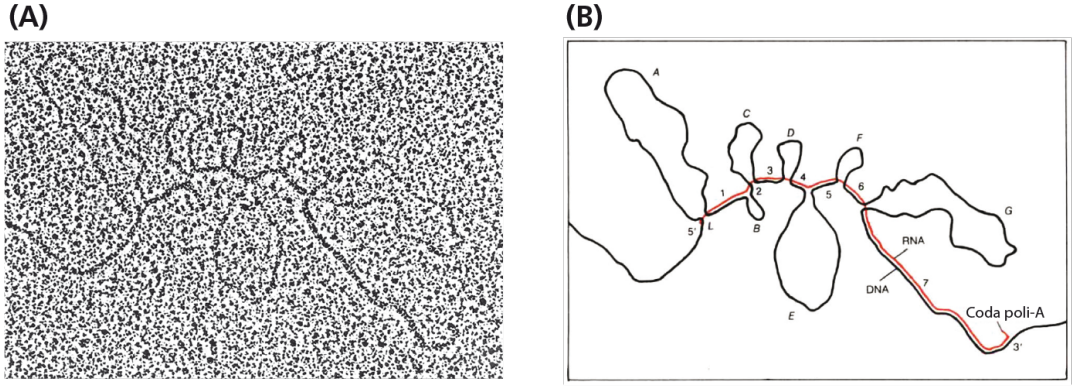
The relationships between haploid genome size and the percentage of the genome that consists of protein-coding genes (white circles) and transposable elements (black circles) are shown



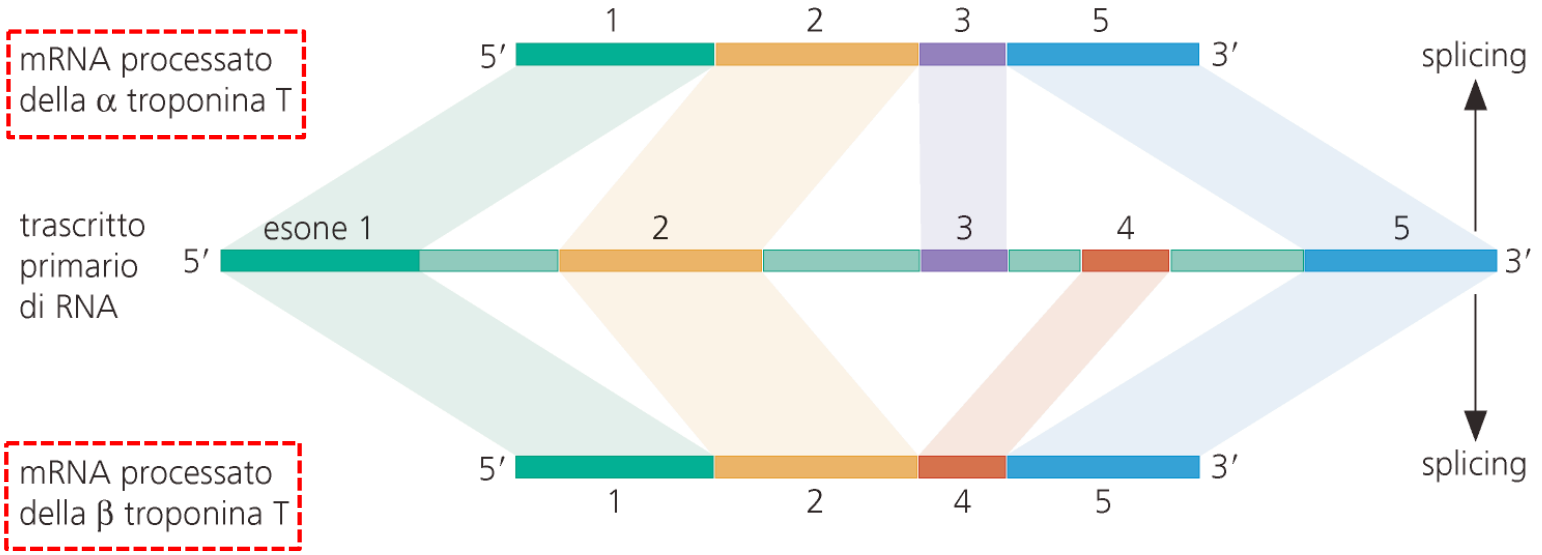
Larger genomes contain proportionately fewer genes and more transposable elements than small genomes

Mammalian Coding genes are complex

Genes contain introns that brake that exons continuity



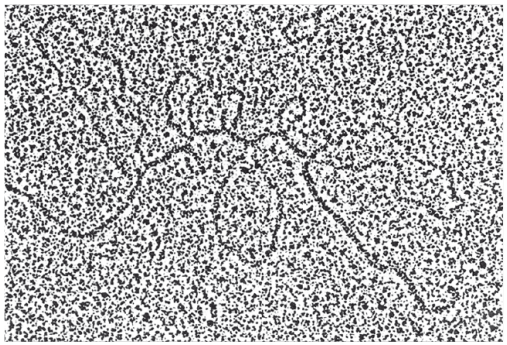
90% of pre-mRNA is alternatively spliced



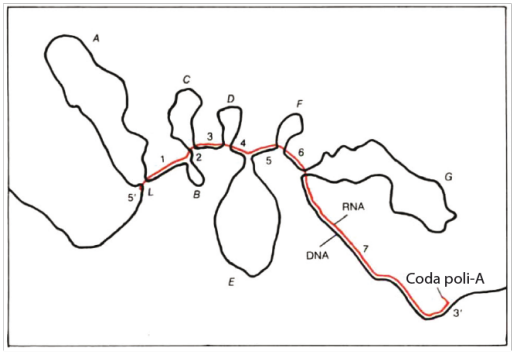
Mammalian Coding genes are complex

Genes contain introns that brake that exons continuity

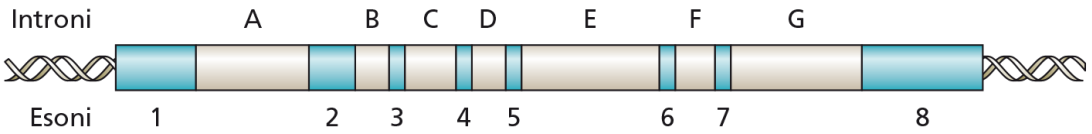
(A)



(B)



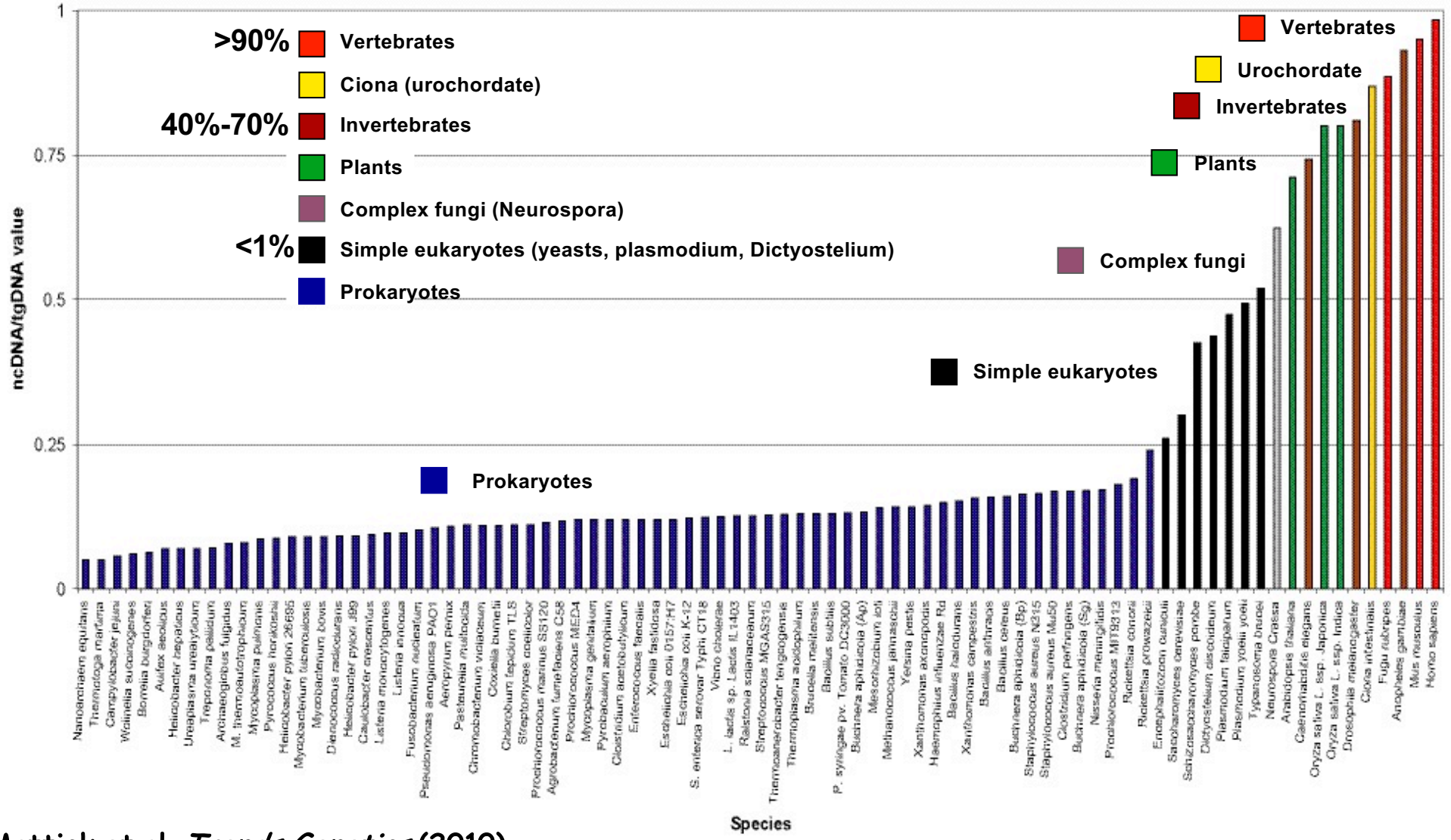
(C)



90% of pre-mRNA is alternatively spliced

Ensembl 98 release, September 2019, is 20000 coding genes - but with alternative splicing these produce likely *~ 200000 coding protein transcripts.*

The proportion of non-coding DNA broadly increases with developmental complexity but….

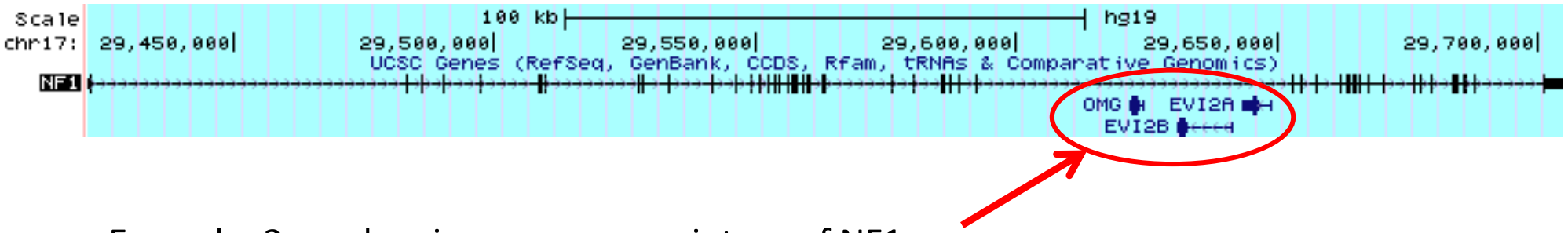


Mattick et al. *Trends Genetics* (2010).

Partially Overlapping Genes

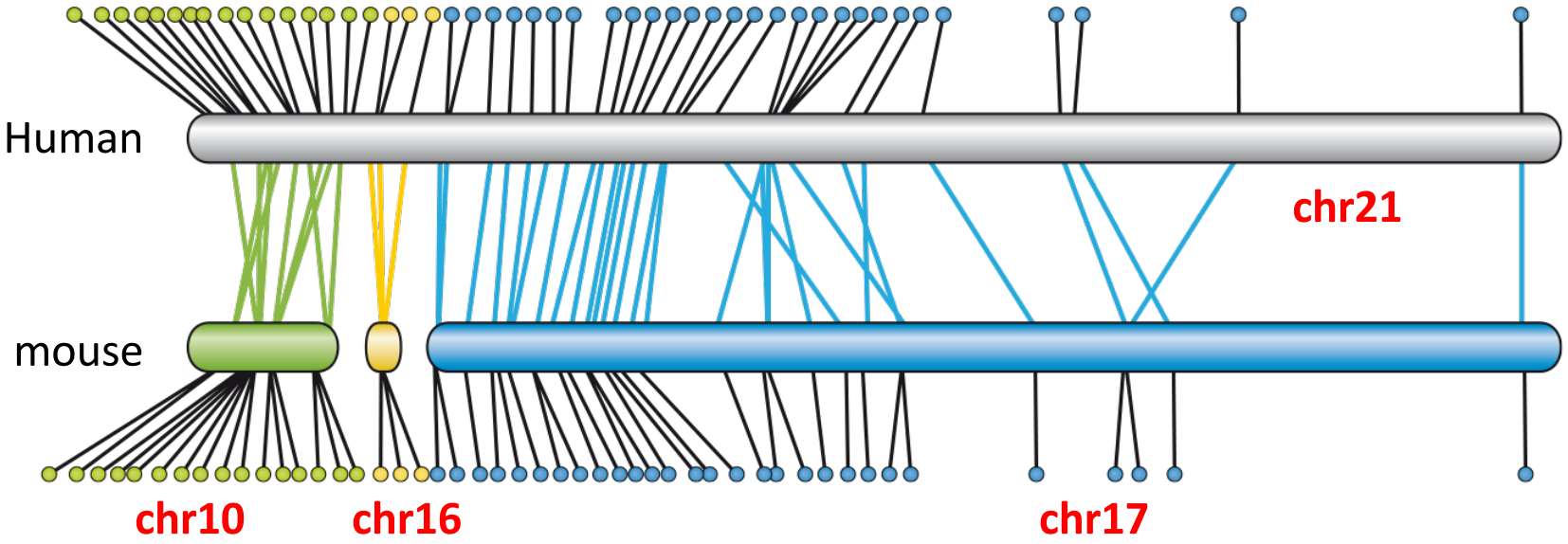
- Gene density varies widely from chromosome to chromosome and between different regions of the same chromosome.

- In high-density regions, genes can be partially overlapping and are generally transcribed in the opposite direction (9% of genes).

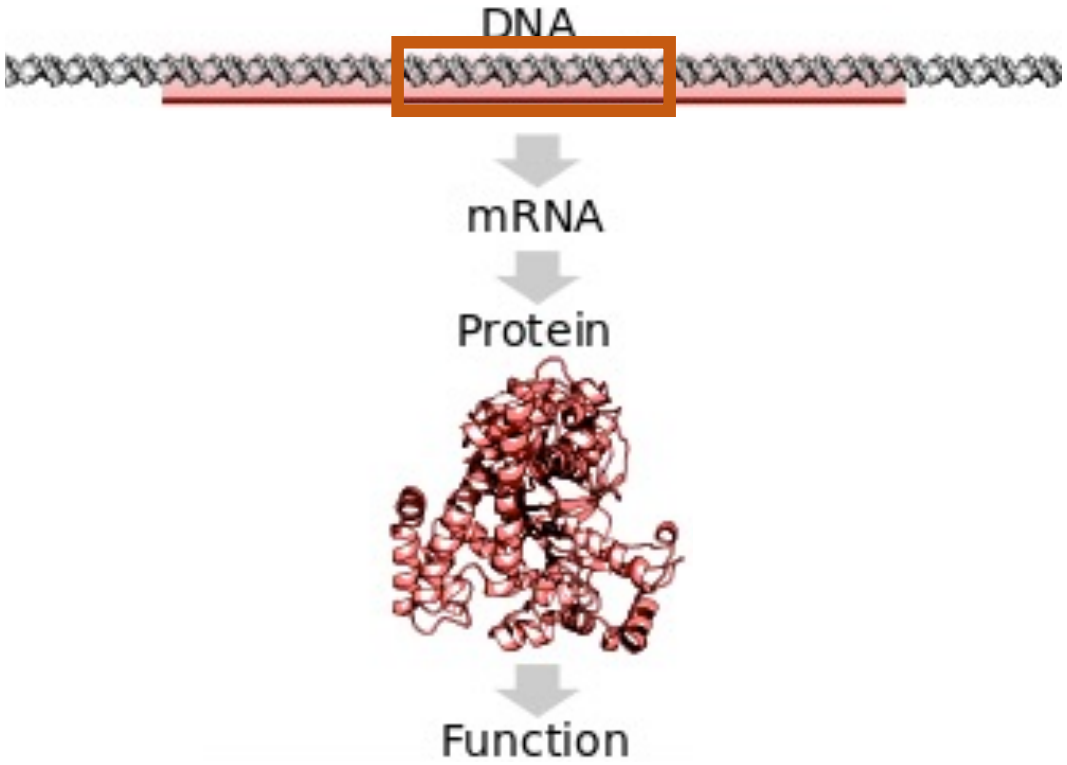


Example: 3 overlapping genes on an intron of NF1

- Conservation in Gene Order



Central dogma of Molecular Biology



Initial definition of GENE

gene: the basic physical unit of heredity; a linear sequence of nucleotides along a segment of DNA that provides the coded instructions for synthesis of RNA, which, when translated into protein, leads to the expression of hereditary character.

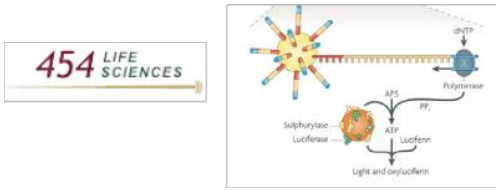
Why it is important to know the genome sequence

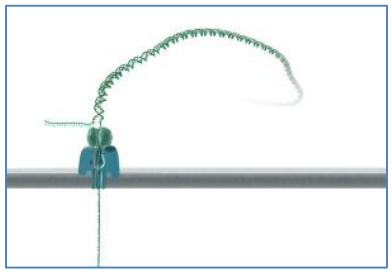
- To study human variability
- To study gene expression
- To study evolutionary relationships between humans and other organisms
- To identify correlations between the information contained in the genome and susceptibility and predisposition to diseases, as well as responses to drugs (pharmacogenomics)

Next Generation Sequencing (NGS)

- 2005 – 454 Life Sciences acquired by Roche Diagnostics; based on emPCR and pyrosequencing. Massive parallel Sequencing by Synthesis.
- 2007 – SOLiD (Applied Biosystems): based on Sequencing By Ligation technology.
- 2007 – Solexa acquired by Illumina:

  Based on Sequencing by Synthesis with cleavable fluorescent didesoxy-nucleotides.
- 2011 – Pacific Biosciences: single molecule real time sequencing (SMRT)
- 2015(founded) – Oxford Nanopore: nanopore based sequencing of single DNA molecules.

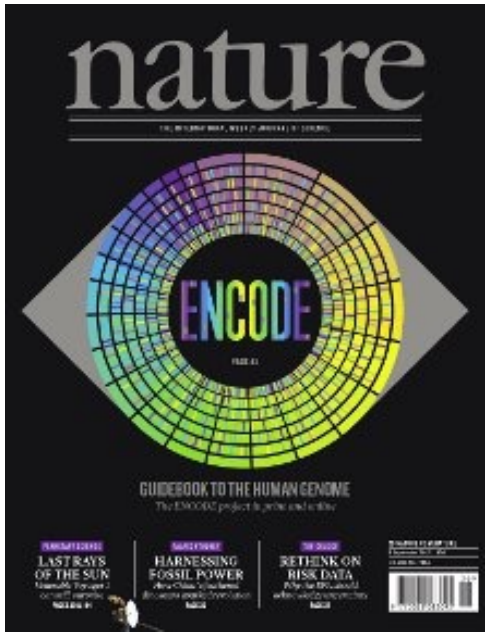




The progress in recent years of sequencing technologies allows us to reach the goal of a 1,000 dollar genome, thus making the sequencing of each individual possible …… the bottleneck is data interpretation.

The ENCODE project

The National Human Genome Research Institute (NHGRI) launched a public research consortium called ENCODE, **Encyclopedia Of DNA Elements**, in September 2003.



The goal of the ENCODE project was to create a catalog of functional elements in the human genome, including transcribed regions, regulatory elements, and modifications of DNA and DNA-binding proteins

Even though only 1% encodes for proteins, the project highlighted that more than **80%** of the genome is functionally "active."

The FANTOM project

FANTOM is an international research consortium founded by Dr. Hayashizaki and his colleagues in 2000 to functionally annotate all the full-length cDNAs that had been collected by the Mouse Encyclopedia Project at the Japanese research institute RIKEN

Functional annotation of a full-length mouse cDNA collection

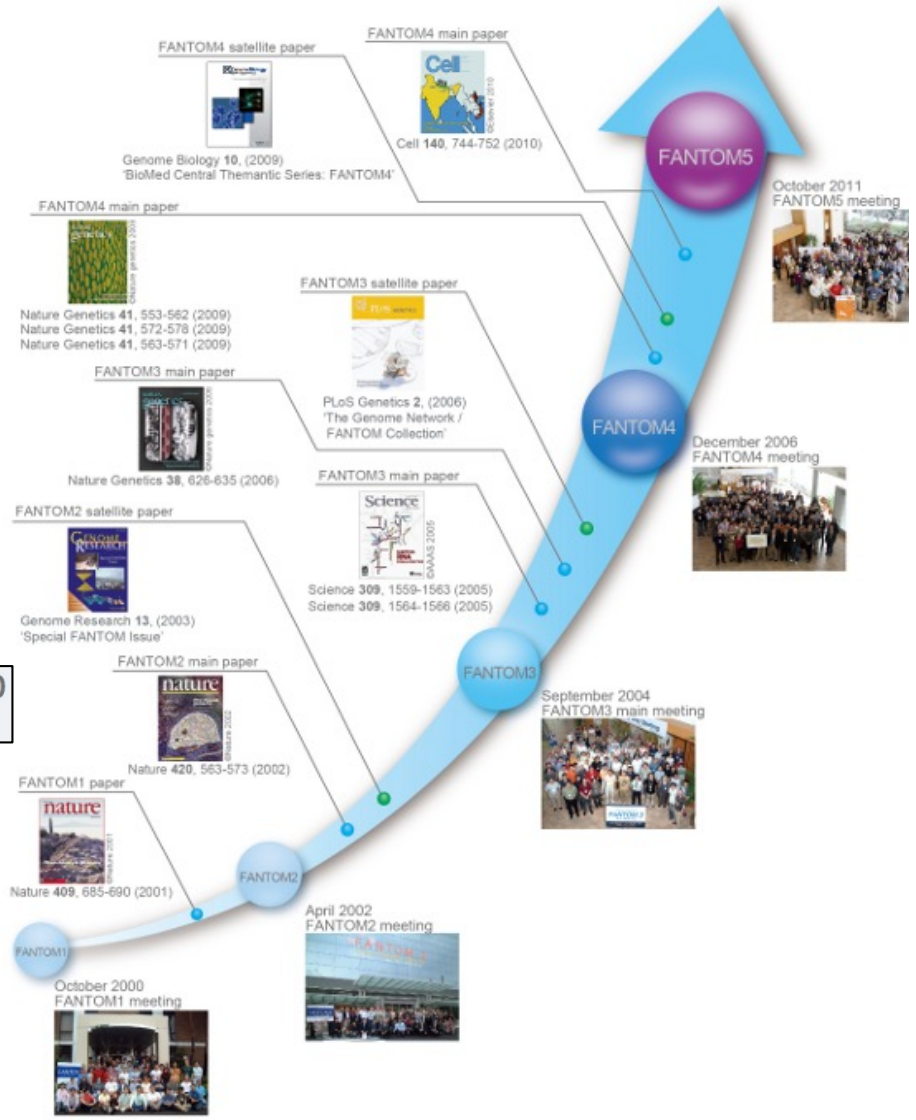
Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs

The transcriptional landscape of the mammalian genome

An atlas of combinatorial transcriptional regulation in mouse and man

An atlas of active enhancers across human cell types and tissues.

A promoter level mammalian expression atlas.



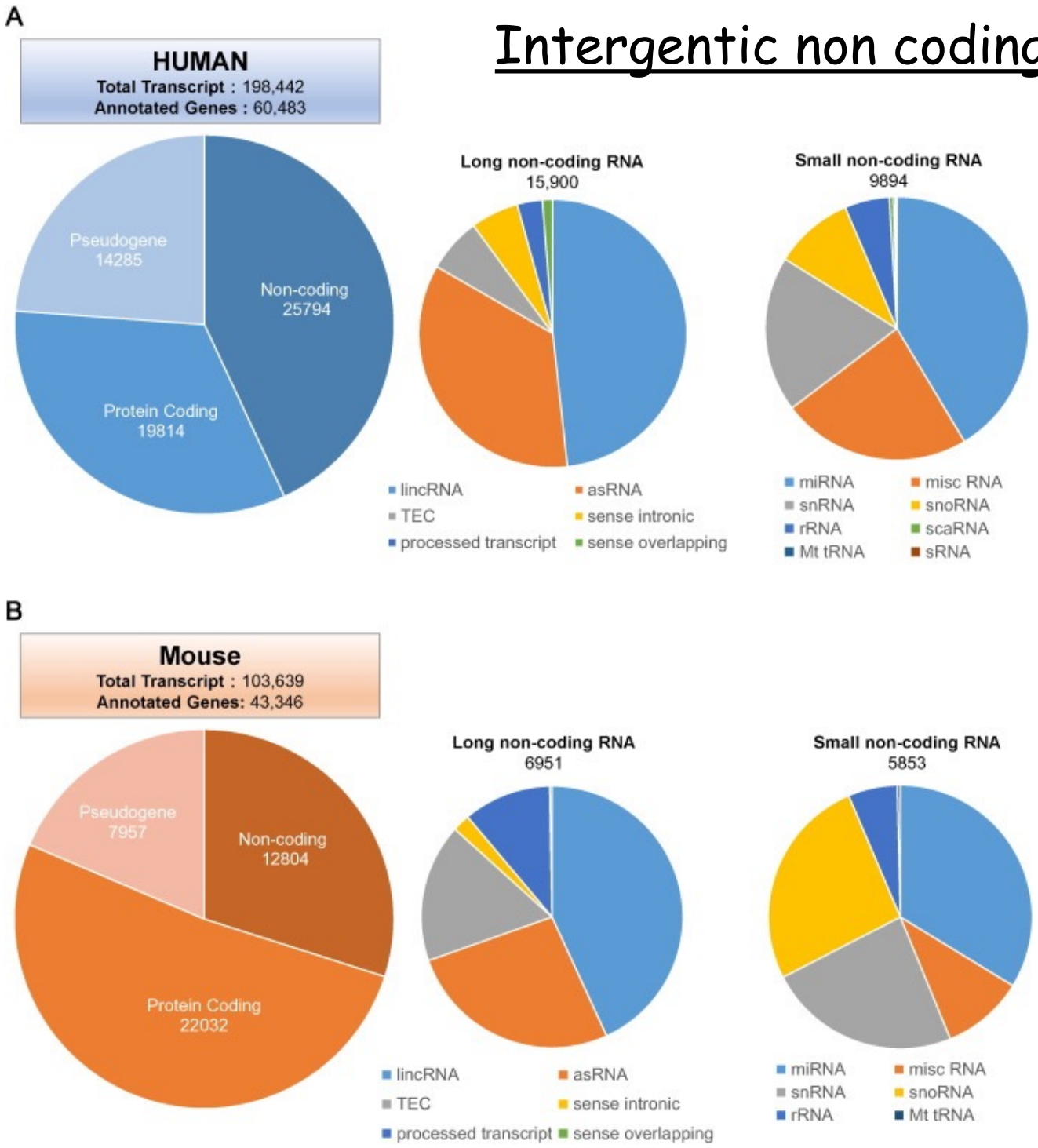
Non coding DNA in Complex genomes

Complex genomes have about 10 to 30 times more DNA than is required to produce all the RNAs or proteins of an organism.

The non-coding DNA includes:

- Introns
- Regulatory elements of genes
- Multiple copies of genes, including pseudogenes
- Intergenic sequences
- Interspersed repeats
- Small and long non coding genes

Intergentic non coding sequences



The complexity of an organism, therefore, arises from how the expression of different genes is regulated

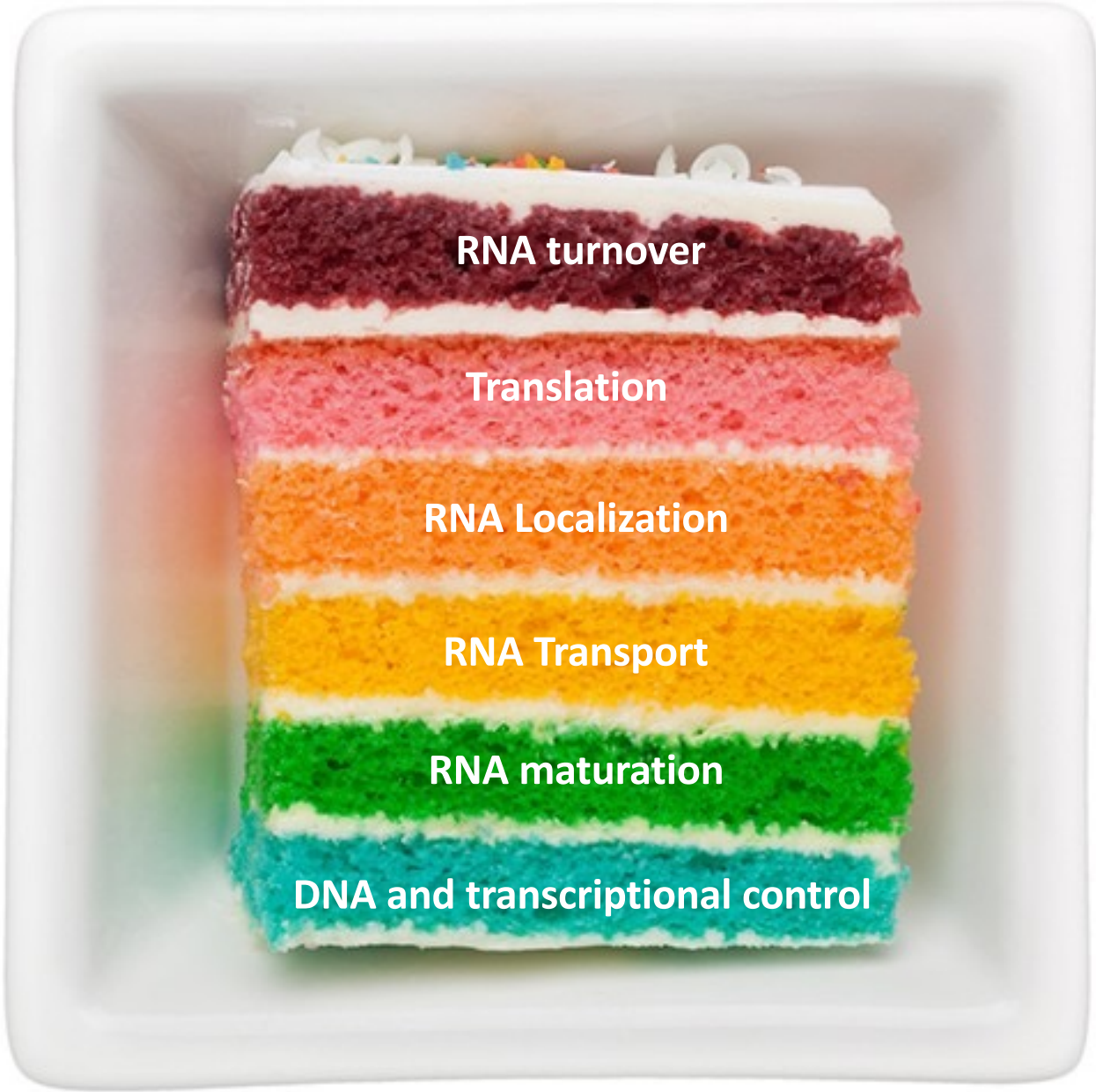
The complexity of an organism is the result of more than just the number of nucleotides that make up the genome and more than just the number of coding genes:

1. A coding sequence can produce a large number of protein products thanks to alternative splicing.

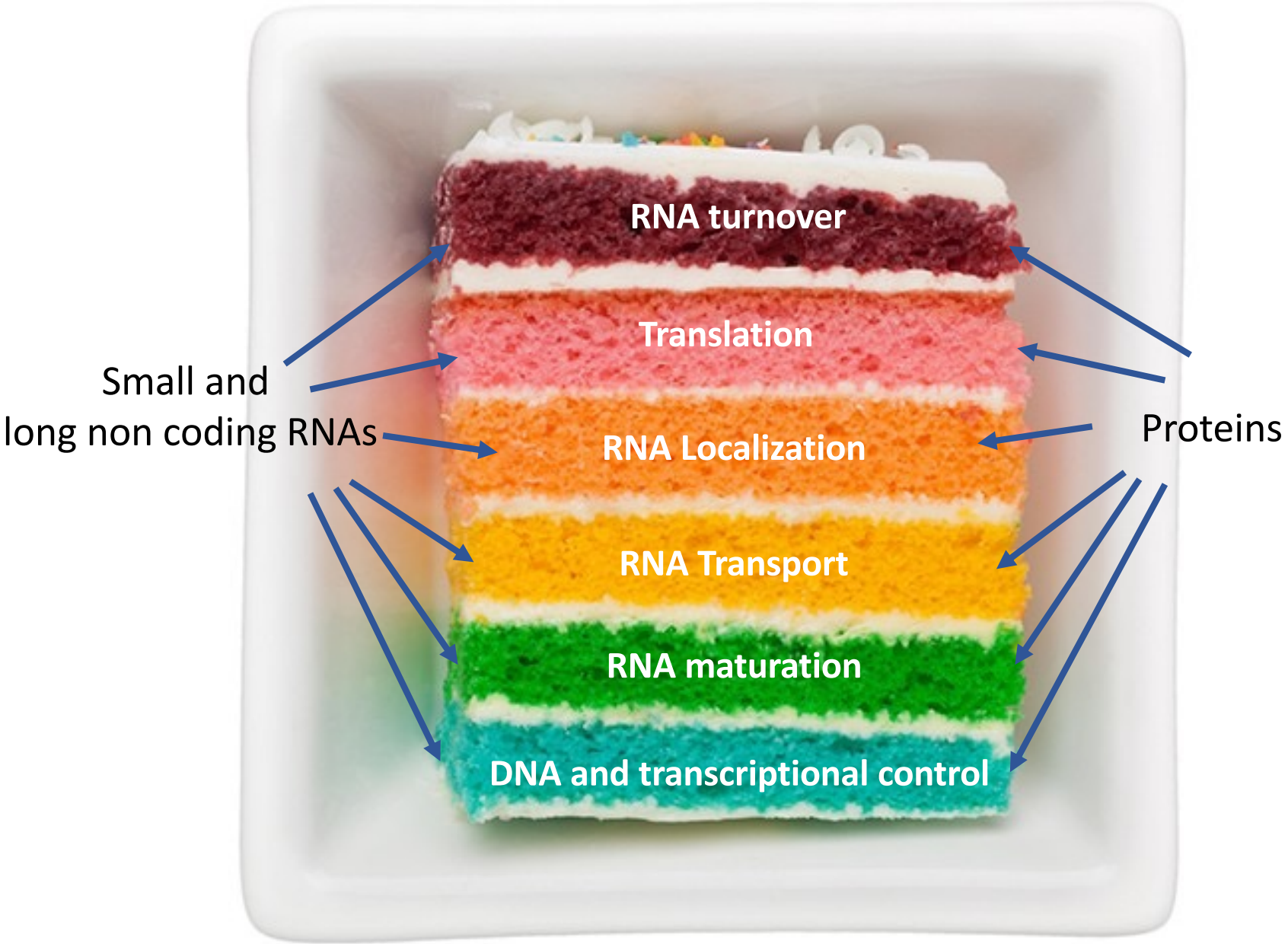
2. Transcribed sequences in non-coding RNA coordinate gene expression (at the transcriptional and post-transcriptional levels!).

The combination of 1. and 2. with other regulatory elements, such as enhancers and promoters, makes it clear that while the size of the genome is a component of complex organisms, **its contribution to the complexity of such organisms is minimal.**

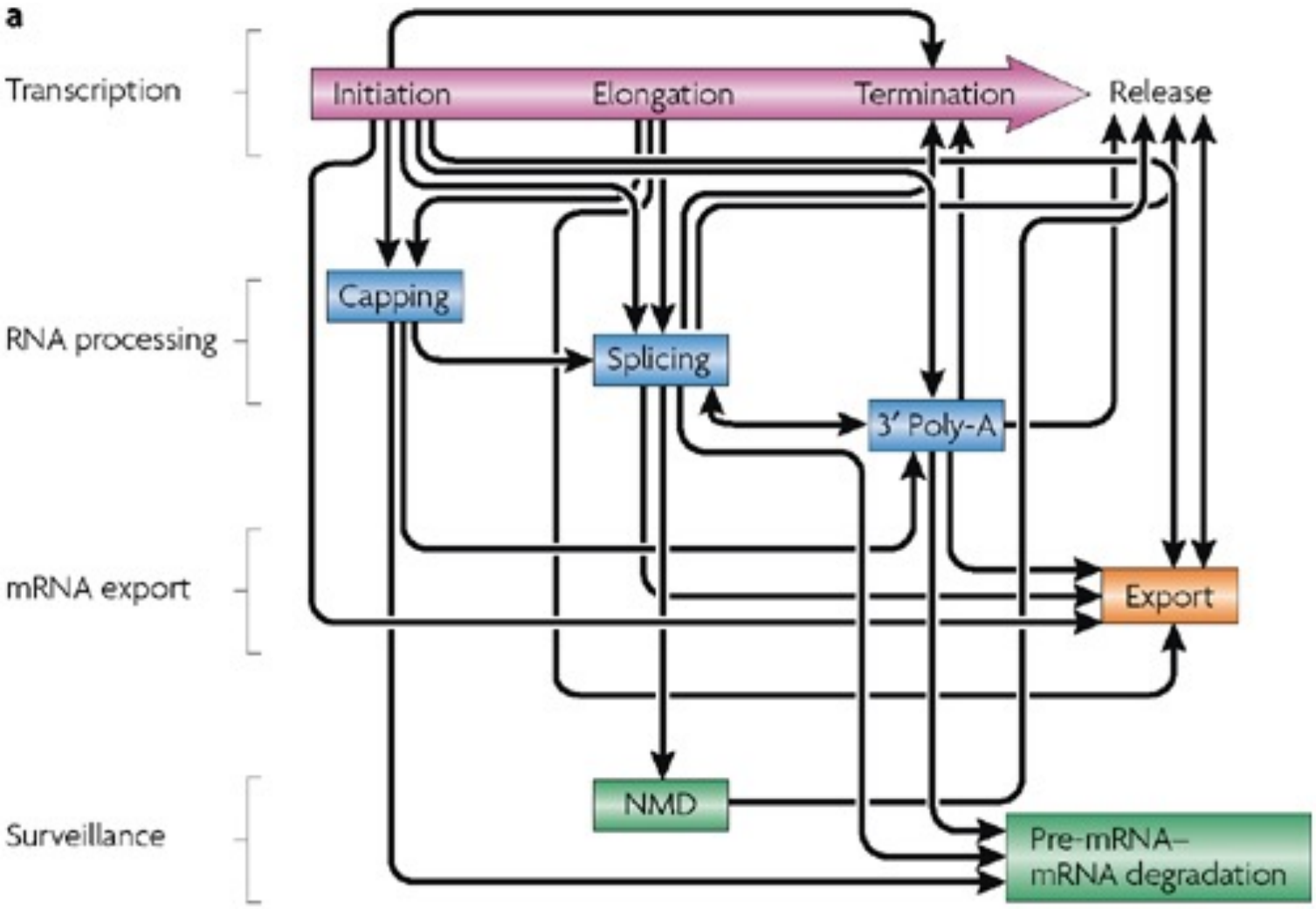
Gene Expression regulation in Eucaryotes



Gene Expression regulation in Eucaryotes



The different stages of gene expression are interconnected at various levels



Questions about the genome

Obtaining the genome sequence and its functional elements is only one step toward understanding biological processes.

Questions:

- What is transcribed from the genome?
- Which proteins and RNAs interact with the functional elements of the genome?
- What is the epigenetic state?

In other words, how does our genome function?