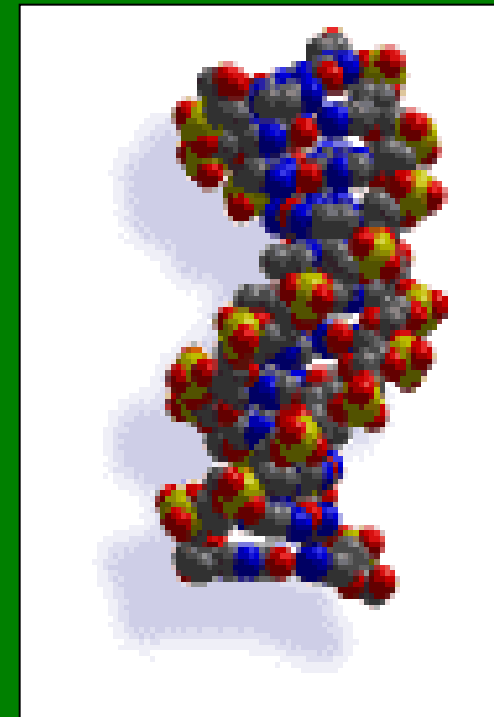


Genomes: What we know ... and what we don't know



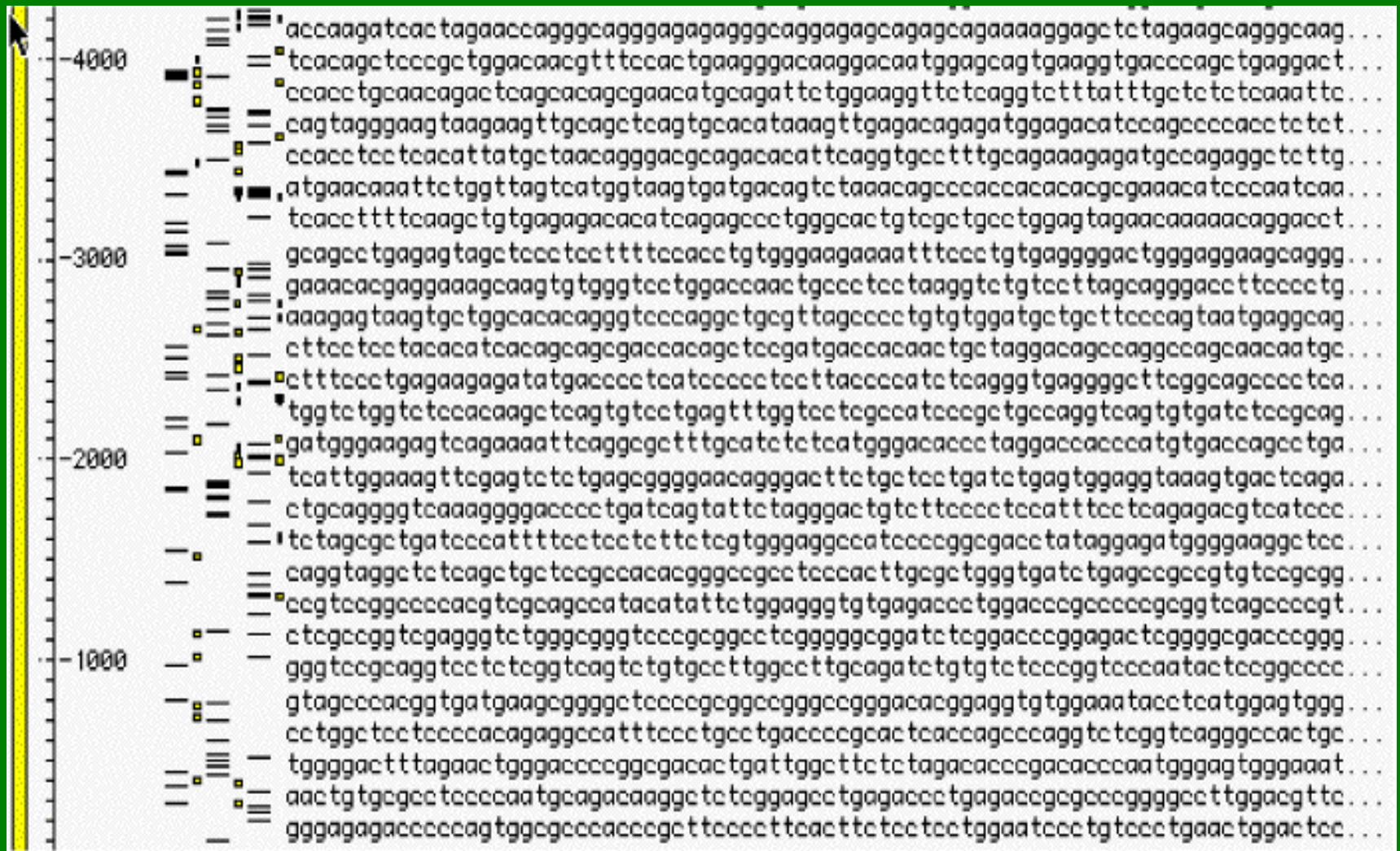
Complete draft
sequence 2001



November 11, 2009

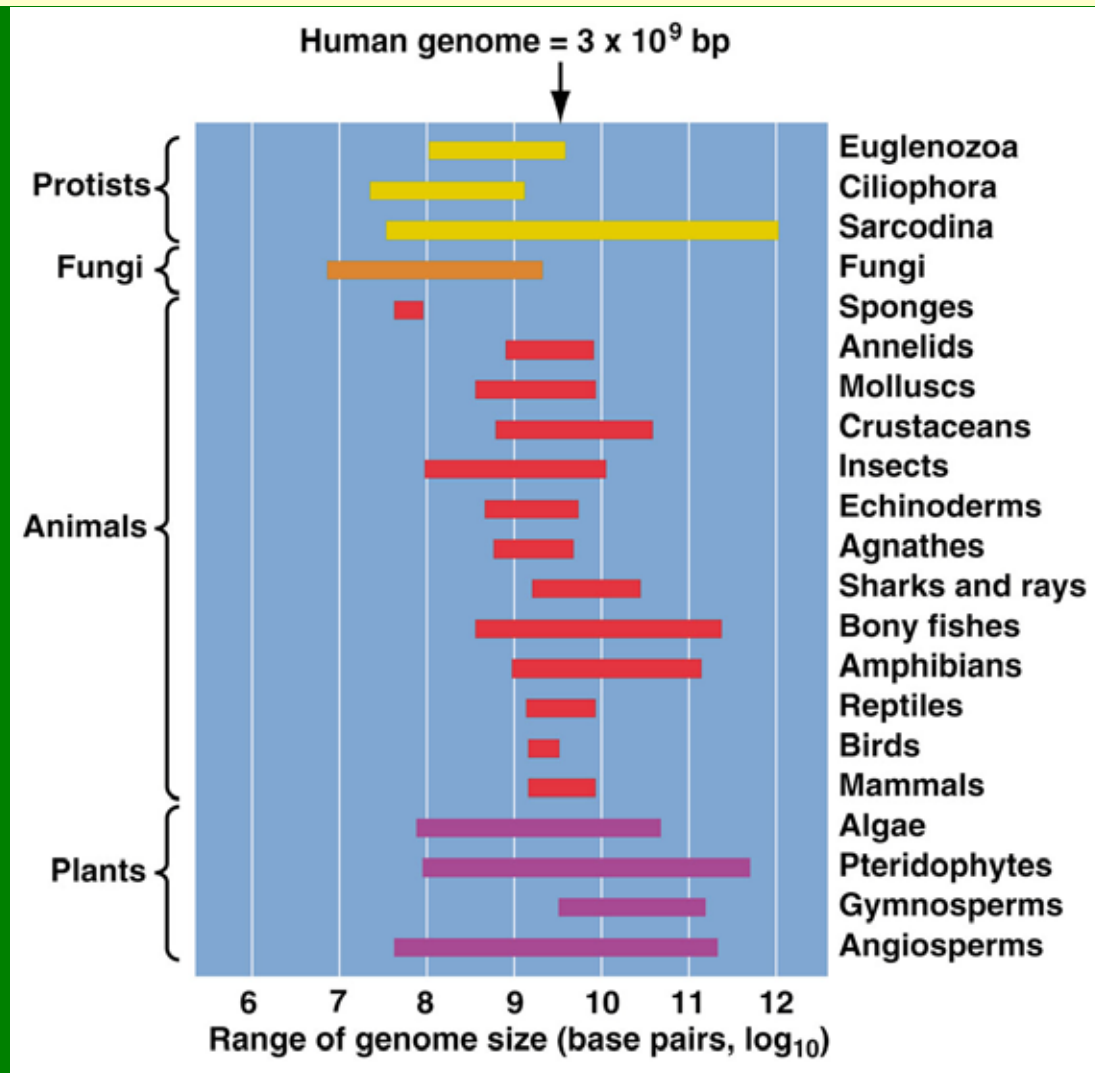
Dr. Stefan Maas, BioS Lehigh U.

What we know



Raw genome data

The range of genome sizes in the animal & plant kingdoms



⇒ No correlation between genome size and complexity

What accounts for the often massive and seemingly arbitrary differences in genome size observed among eukaryotic organisms?

The fruit fly
Drosophila melanogaster



180 Mb

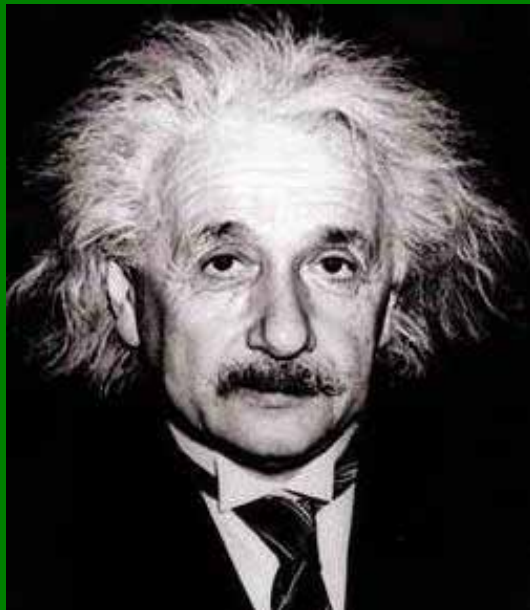
The mountain grasshopper
Podisma pedestris



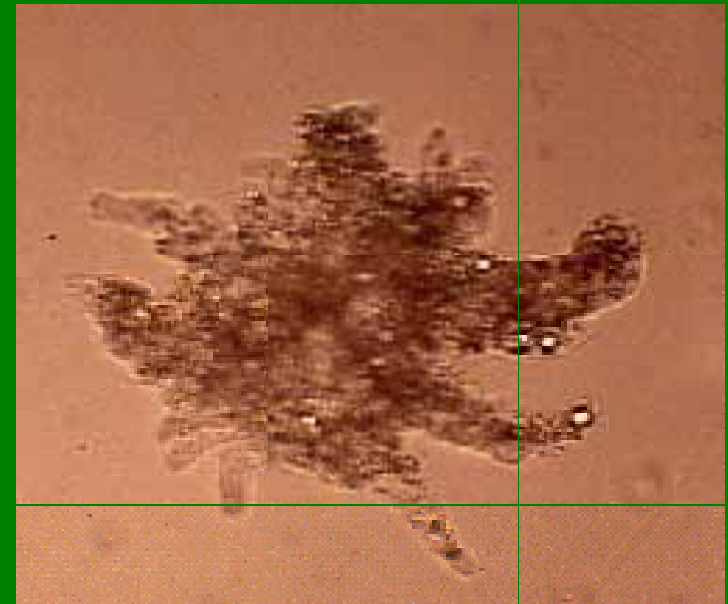
18,000 Mb

The difference in genome size of a factor of 100 is difficult to explain in view of the apparently similar levels of evolutionary, developmental and behavioral complexity of these organisms.

Complexity does not correlate with genome size

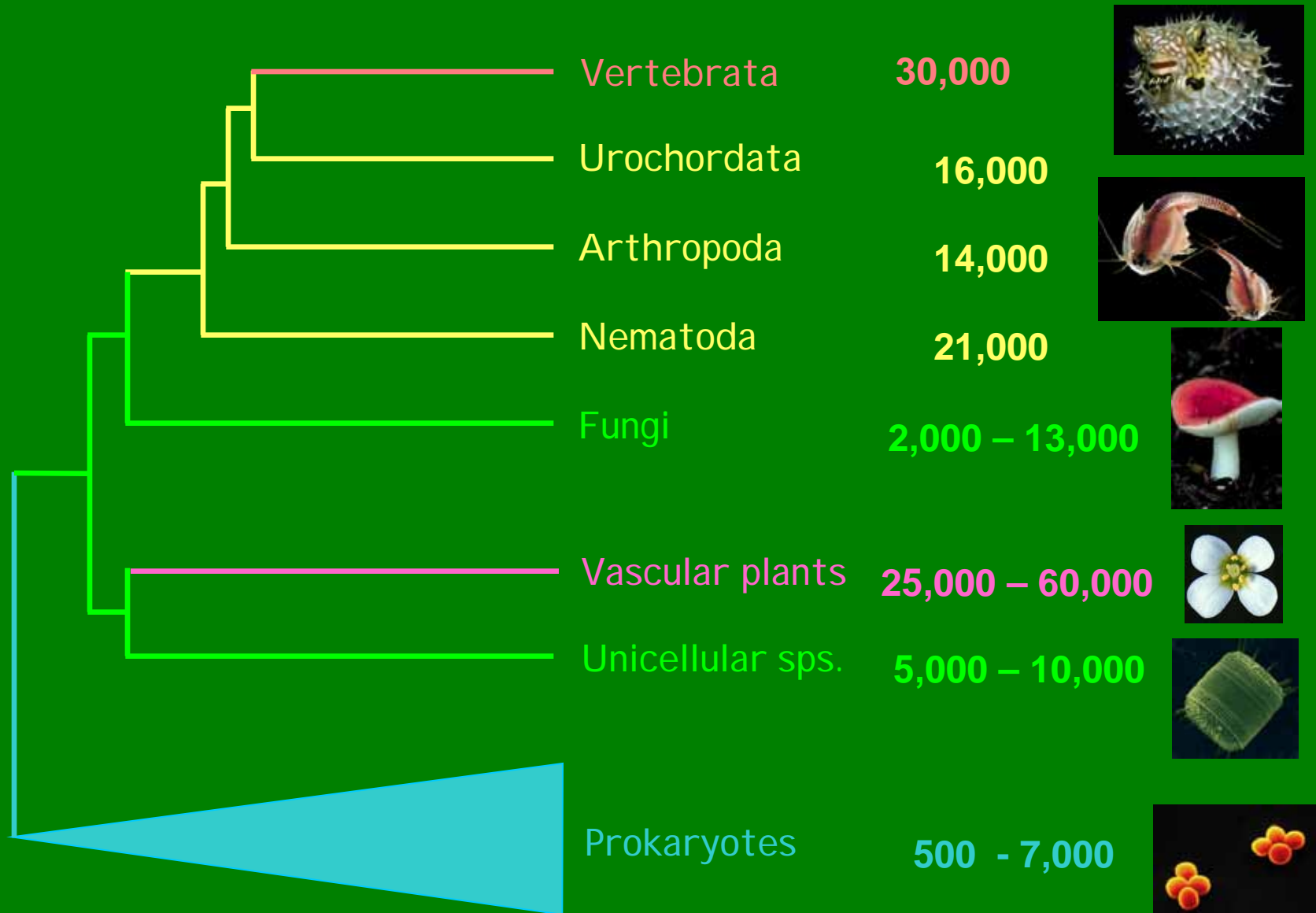


$3.4 \cdot 10^9$ bp
Homo sapiens

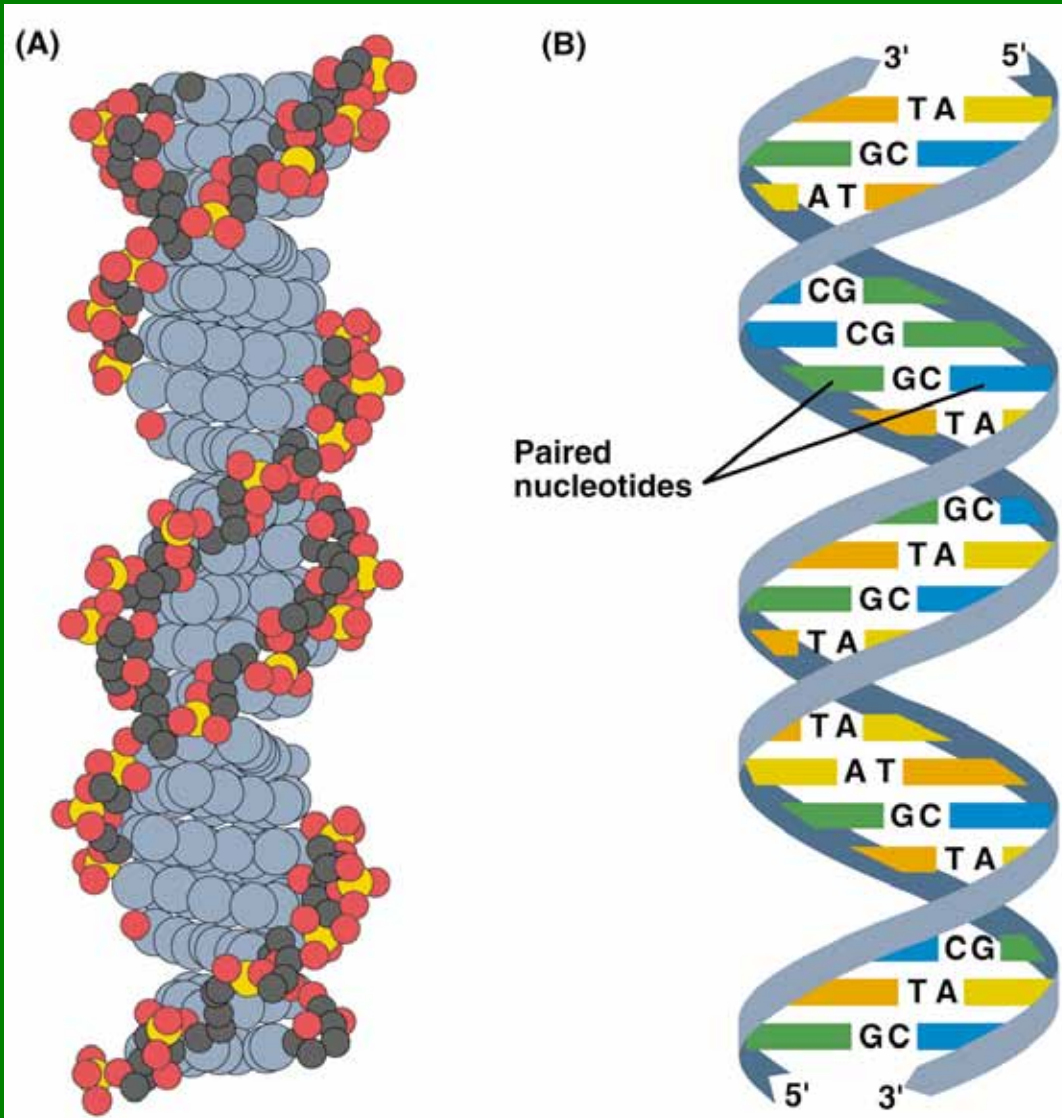


$6.7 \cdot 10^{11}$ bp
Amoeba dubia

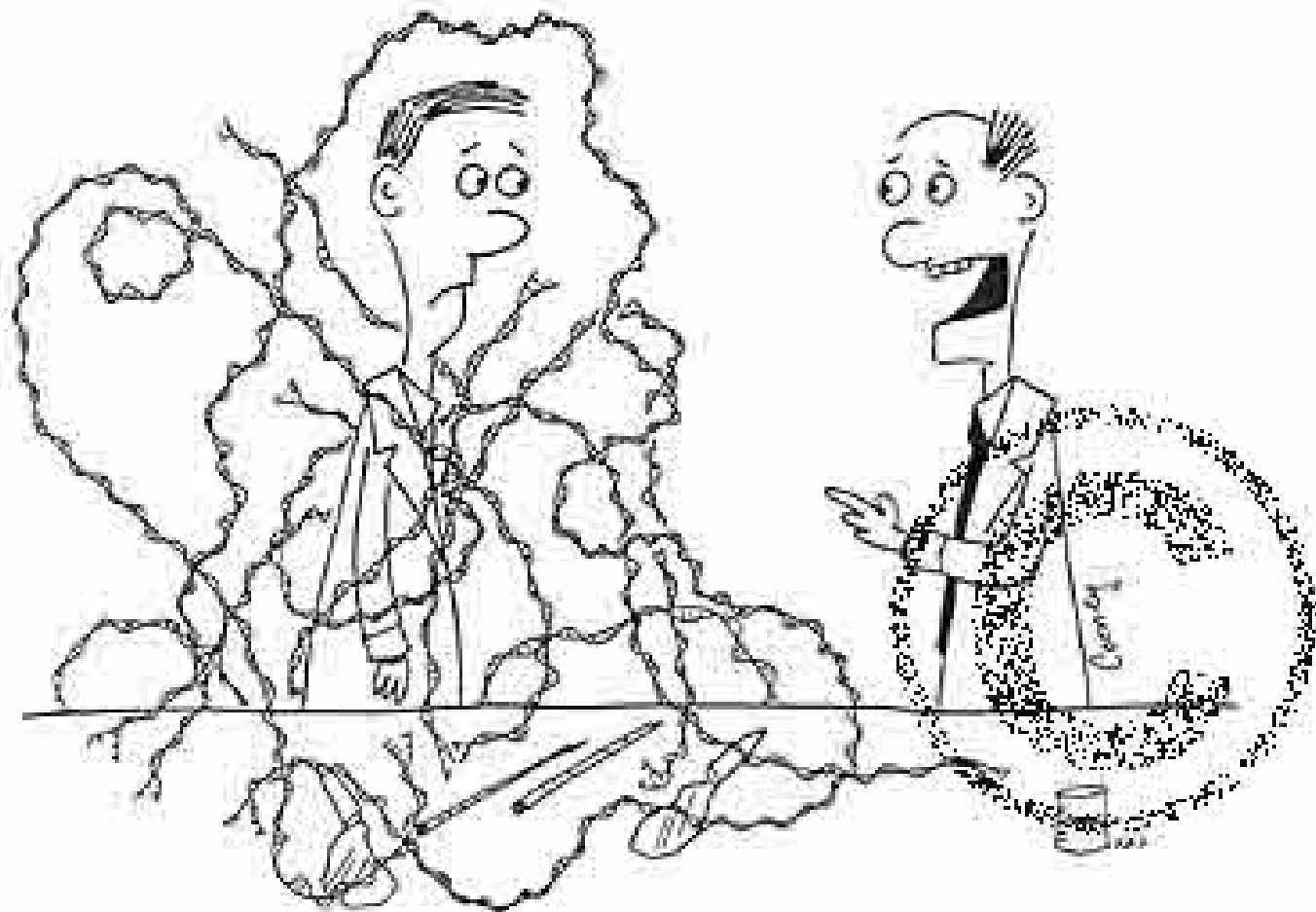
Is an Expansion in Gene Number driving Evolution of Higher Organisms?



Structure of DNA

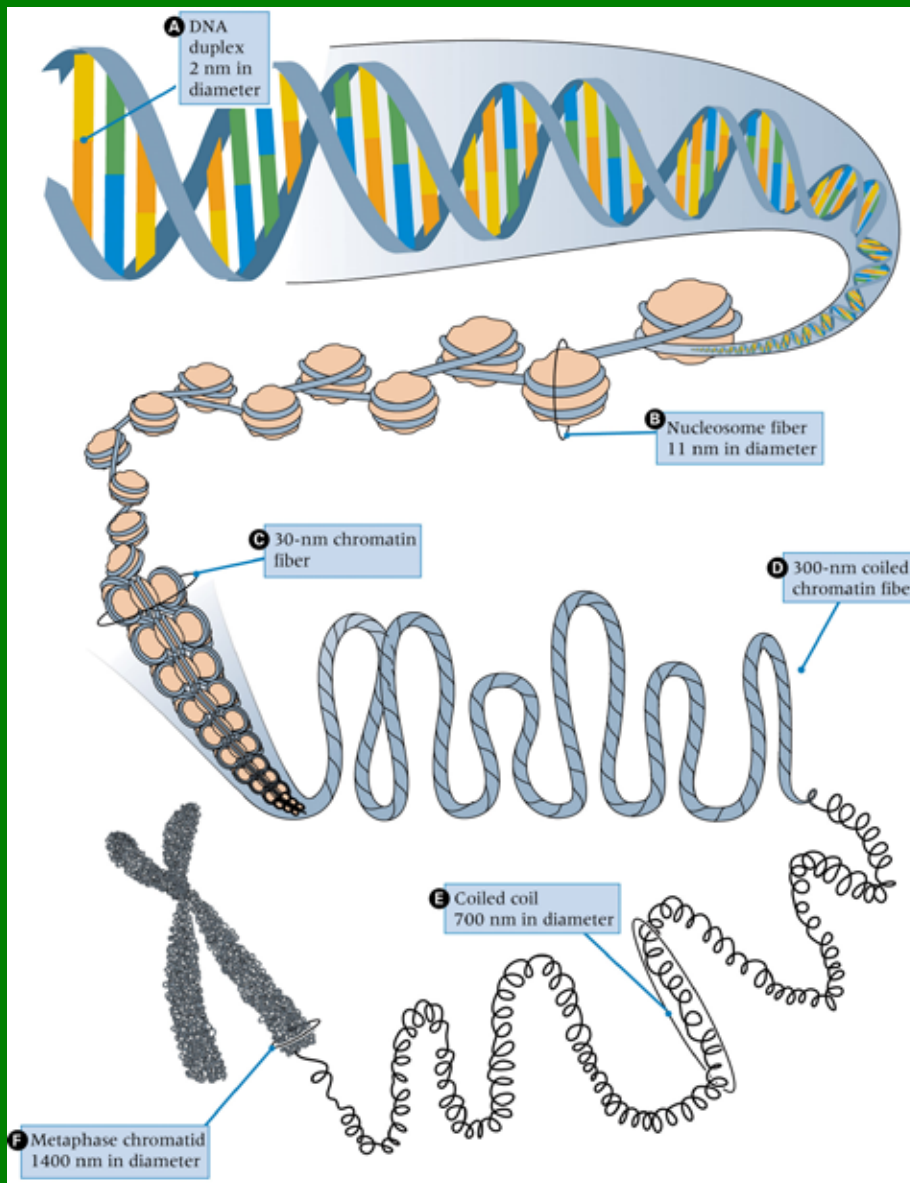


Watson and Crick in 1953 proposed that DNA is a double helix in which the 4 bases are base paired, Adenine (A) with Thymine (T) and Guanine (G) with Cytosine (C).



"From now on, take the Human Genome
outside before you unravel it."

Steps in the folding of DNA to create an eukaryotic chromosome



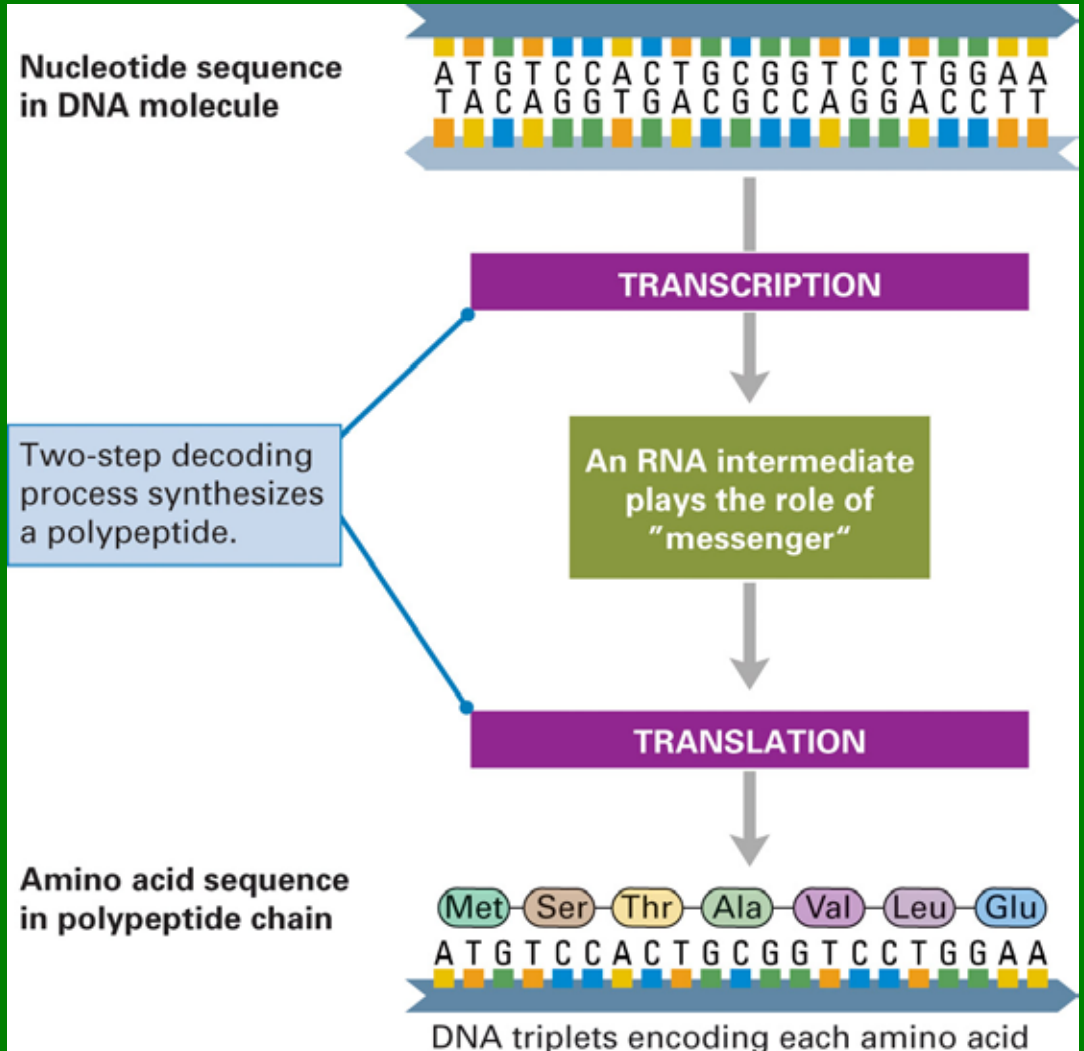
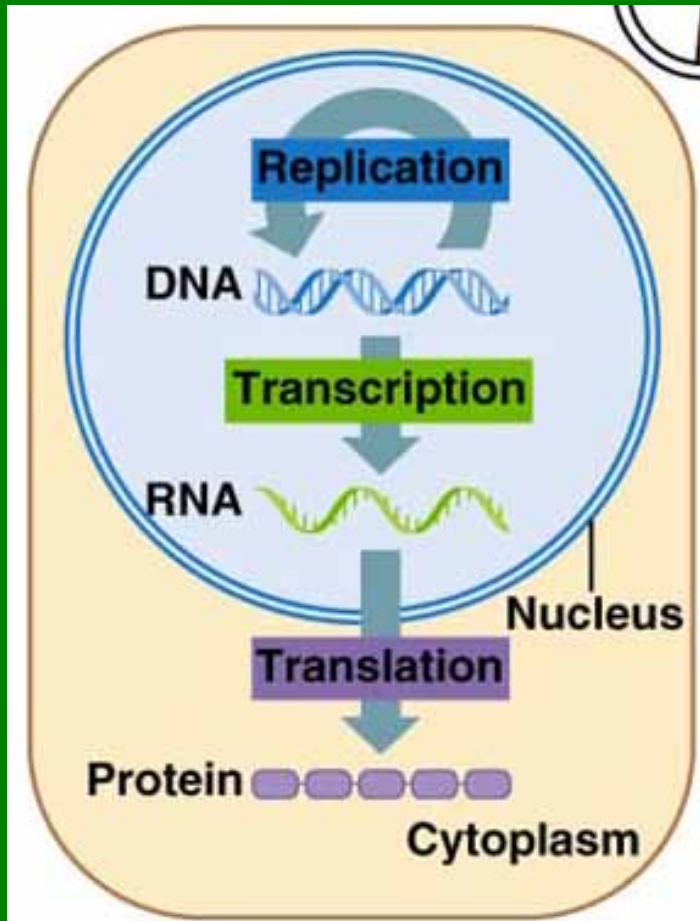
30 nm fiber
(6 nucleosomes per turn)

**Factor of condensation:
Ca. 10,000 fold**

Why?

- è Facilitates movement during cell division
- è Decreases error rate

Genes code for Proteins with RNA as an intermediate



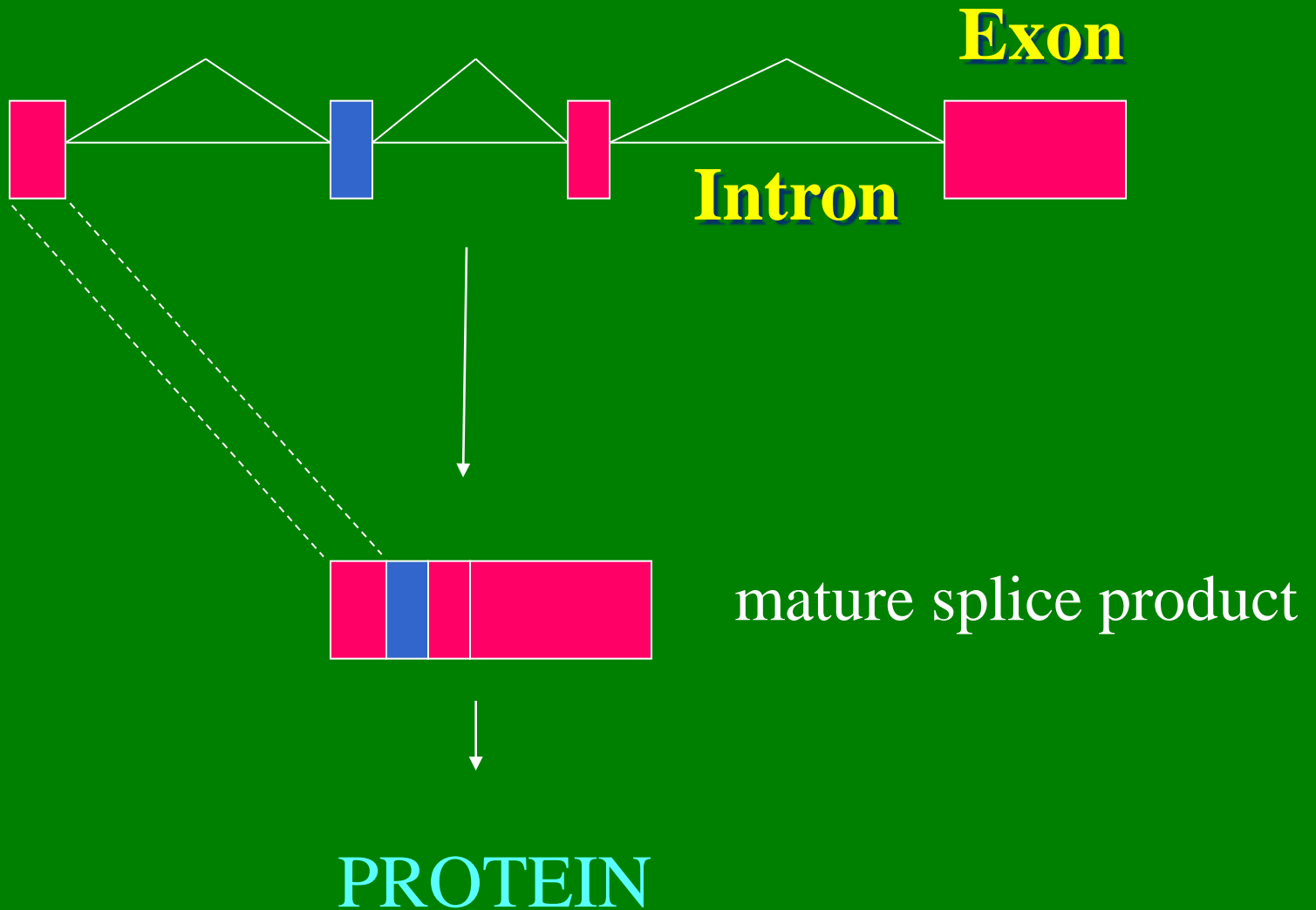
The Genetic Code

Table 1.1 The standard genetic code

		Second nucleotide in codon														
		U			C			A			G					
U	UUU	Phe	F	<i>Phenylalanine</i>	UCU	Ser	S	<i>Serine</i>	UAU	Tyr	Y	<i>Tyrosine</i>	UGU	Cys	C	<i>Cysteine</i>
	UUC	Phe	F	<i>Phenylalanine</i>	UCC	Ser	S	<i>Serine</i>	UAC	Tyr	Y	<i>Tyrosine</i>	UGC	Cys	C	<i>Cysteine</i>
	UUA	Leu	L	<i>Leucine</i>	UCA	Ser	S	<i>Serine</i>	UAA	Termination			UGA	Termination		
	UUG	Leu	L	<i>Leucine</i>	UCG	Ser	S	<i>Serine</i>	UAG	Termination			UGG	Trp	W	<i>Tryptophan</i>
C	CUU	Leu	L	<i>Leucine</i>	CCU	Pro	P	<i>Proline</i>	CAU	His	H	<i>Histidine</i>	CGU	Arg	R	<i>Arginine</i>
	CUC	Leu	L	<i>Leucine</i>	CCC	Pro	P	<i>Proline</i>	CAC	His	H	<i>Histidine</i>	CGC	Arg	R	<i>Arginine</i>
	CUA	Leu	L	<i>Leucine</i>	CCA	Pro	P	<i>Proline</i>	CAA	Gln	Q	<i>Glutamine</i>	CGA	Arg	R	<i>Arginine</i>
	CUG	Leu	L	<i>Leucine</i>	CCG	Pro	P	<i>Proline</i>	CAG	Gln	Q	<i>Glutamine</i>	CGG	Arg	R	<i>Arginine</i>
A	AUU	Ile	I	<i>Isoleucine</i>	ACU	Thr	T	<i>Threonine</i>	AAU	Asn	N	<i>Asparagine</i>	AGU	Ser	S	<i>Serine</i>
	AUC	Ile	I	<i>Isoleucine</i>	ACC	Thr	T	<i>Threonine</i>	AAC	Asn	N	<i>Asparagine</i>	AGC	Ser	S	<i>Serine</i>
	AUA	Ile	I	<i>Isoleucine</i>	ACA	Thr	T	<i>Threonine</i>	AAA	Lys	K	<i>Lysine</i>	AGA	Arg	R	<i>Arginine</i>
	AUG	Met	M	<i>Methionine</i>	ACG	Thr	T	<i>Threonine</i>	AAG	Lys	K	<i>Lysine</i>	AGG	Arg	R	<i>Arginine</i>
G	GUU	Val	V	<i>Valine</i>	GCU	Ala	A	<i>Alanine</i>	GAU	Asp	D	<i>Aspartic acid</i>	GGU	Gly	G	<i>Glycine</i>
	GUC	Val	V	<i>Valine</i>	GCC	Ala	A	<i>Alanine</i>	GAC	Asp	D	<i>Aspartic acid</i>	GGC	Gly	G	<i>Glycine</i>
	GUA	Val	V	<i>Valine</i>	GCA	Ala	A	<i>Alanine</i>	GAA	Glu	E	<i>Glutamic acid</i>	GGA	Gly	G	<i>Glycine</i>
	GUG	Val	V	<i>Valine</i>	GCG	Ala	A	<i>Alanine</i>	GAG	Glu	E	<i>Glutamic acid</i>	GGG	Gly	G	<i>Glycine</i>

Codon Three-letter and single-letter abbreviations

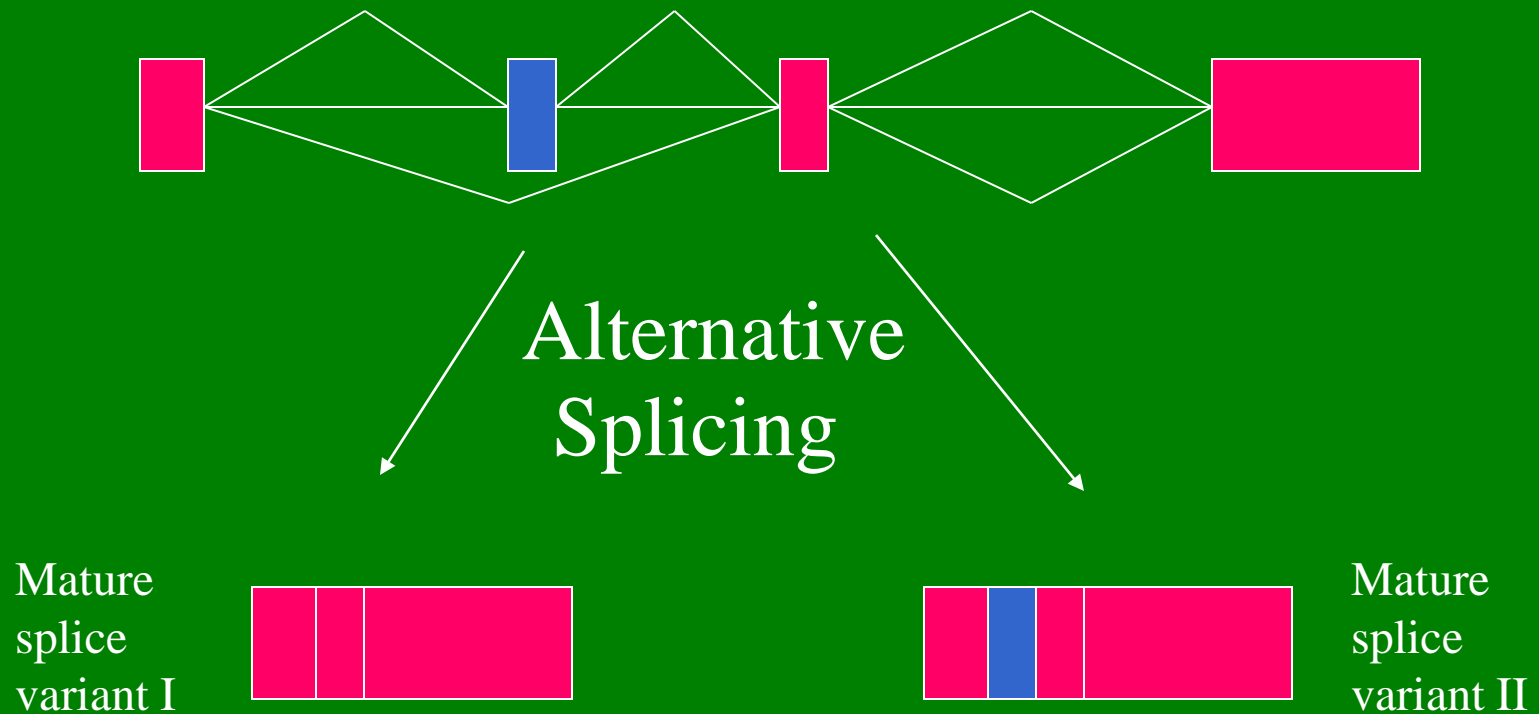
Gene splicing: Removal of non-coding introns



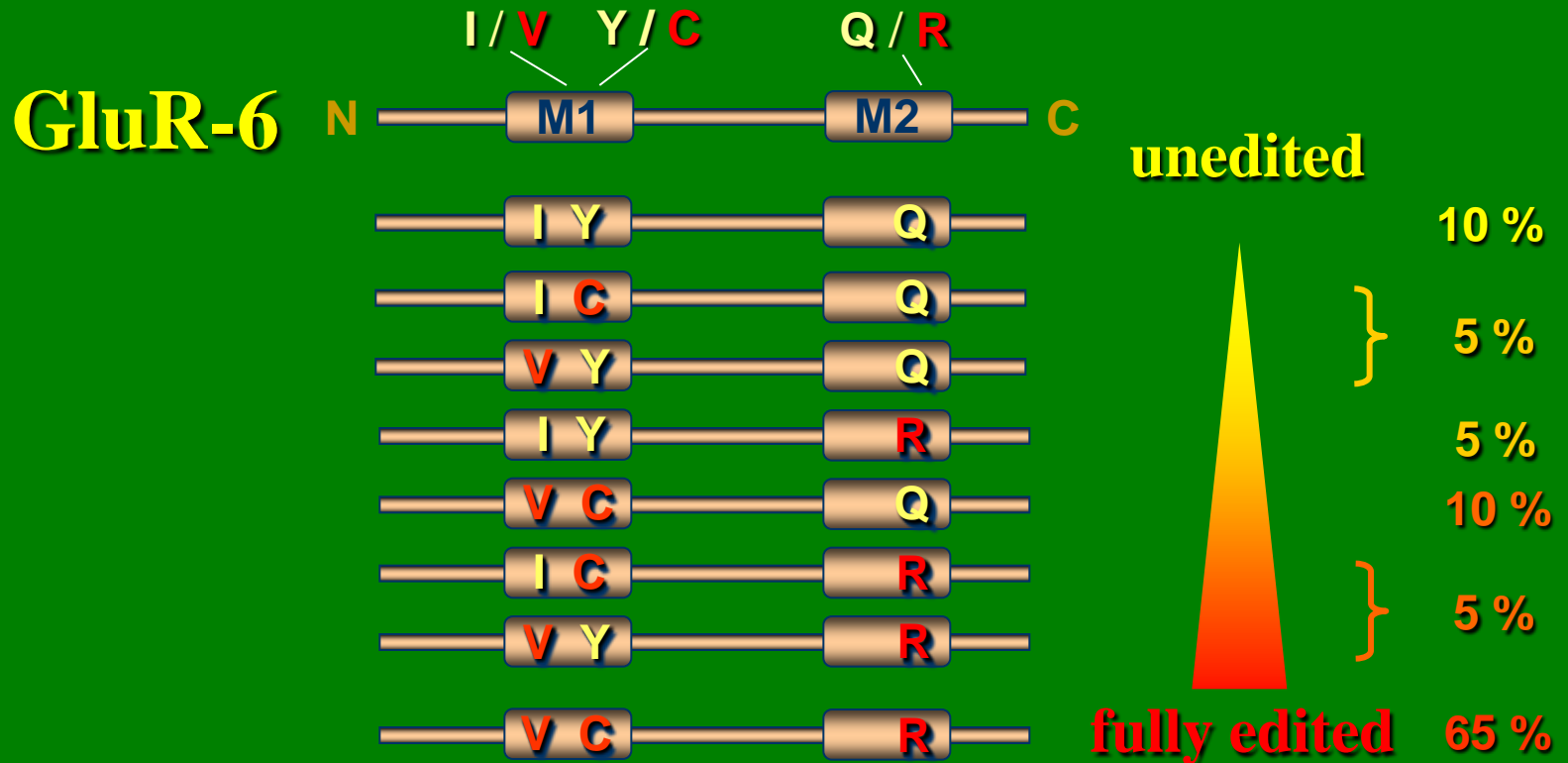
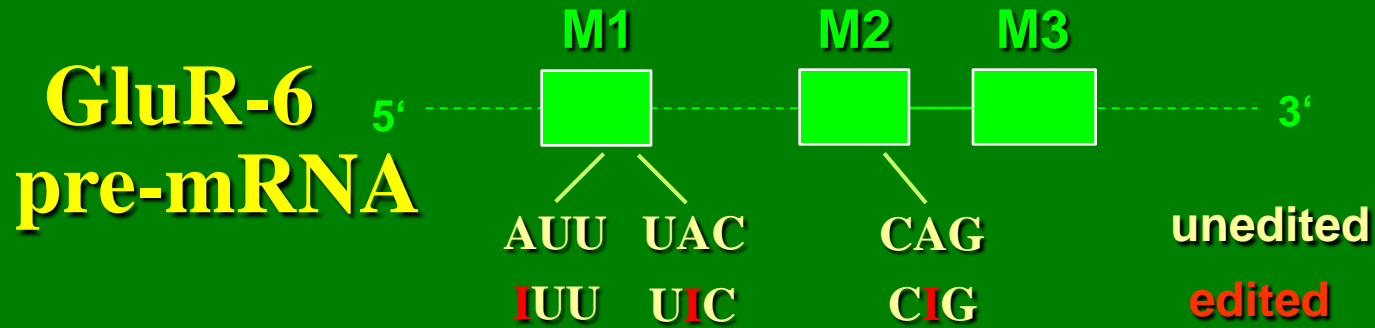
Human genome facts

- ~ 30,000 genes
- On average:
 - Coding length: 1.4 kb
 - Gene extent: 30 kb
 - 8 Exons (135 bp), 7 introns (2,200 bp)
 - Gene density: 11.5/1 Mb

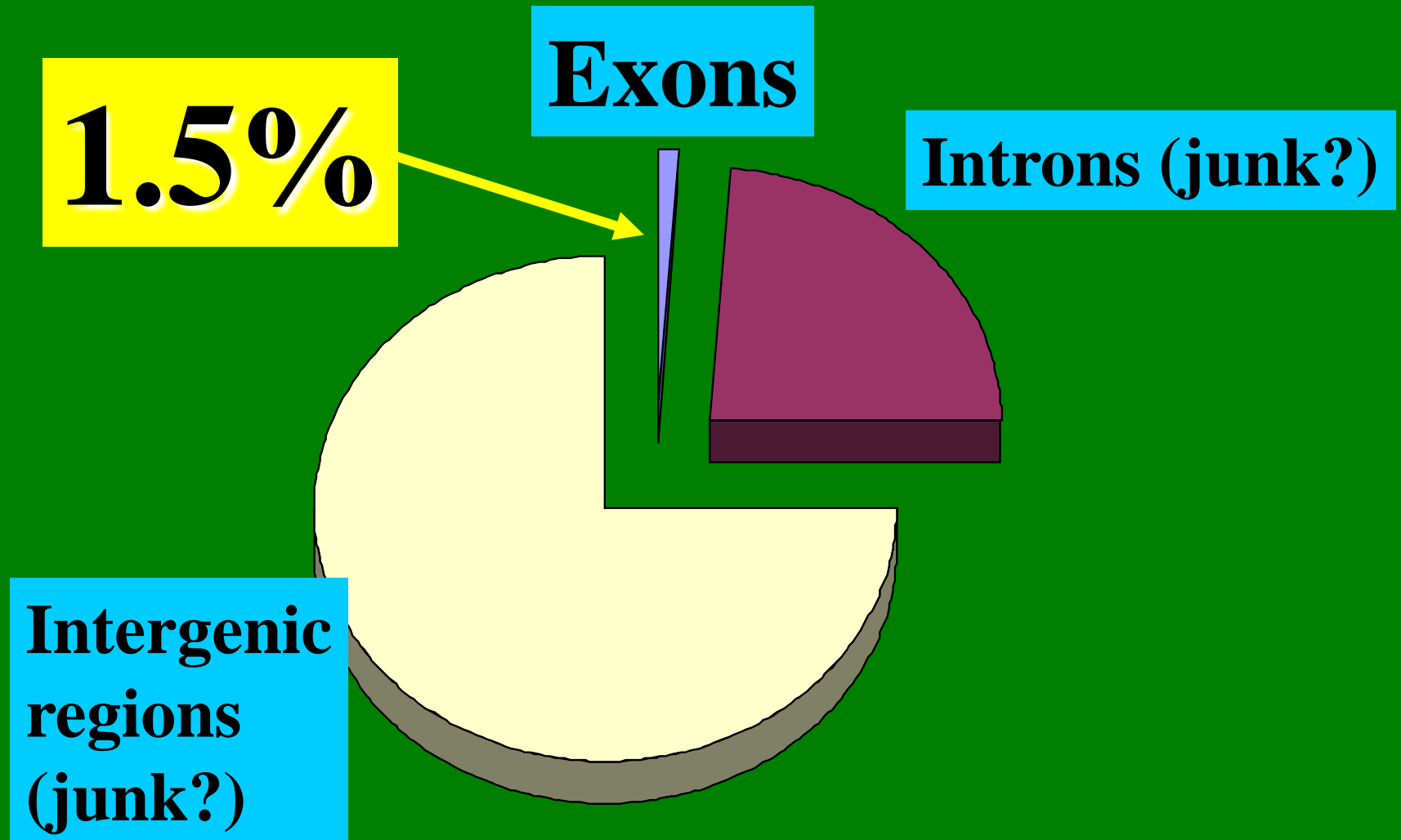
Alternative splicing: One gene, several proteins!



Diversity through RNA editing

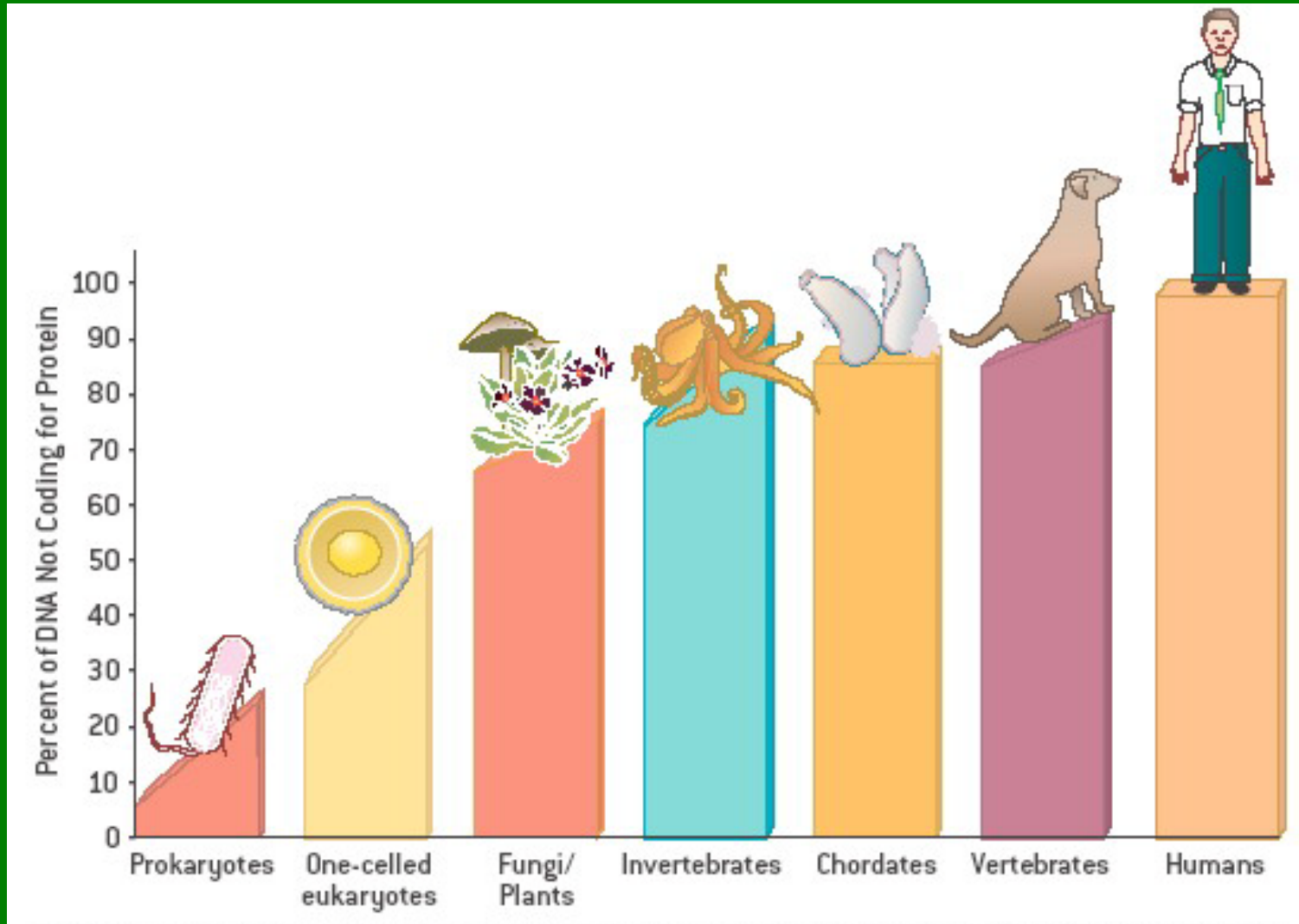


Is the genome empty ?



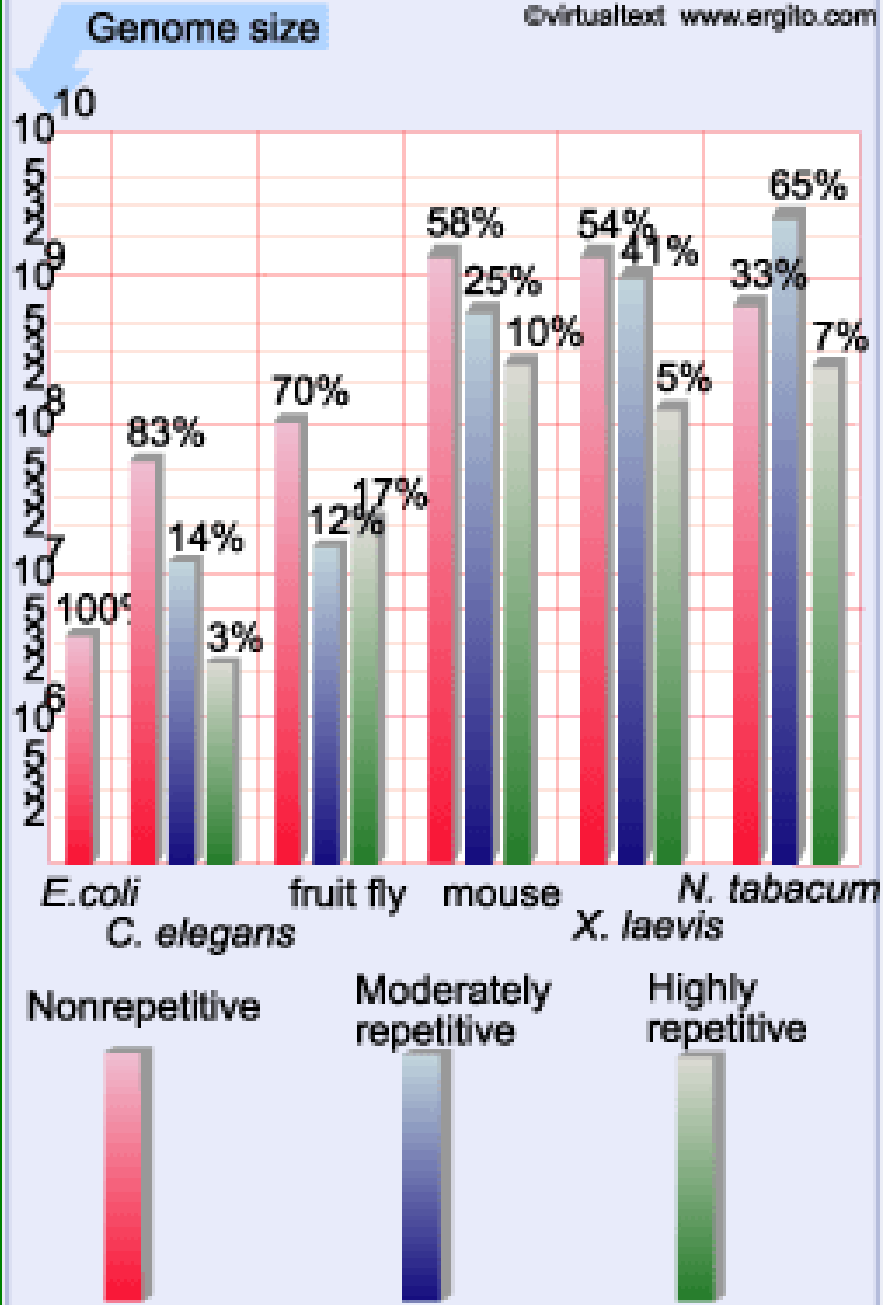
Correlation between complexity and amount of non-coding DNA

1.0

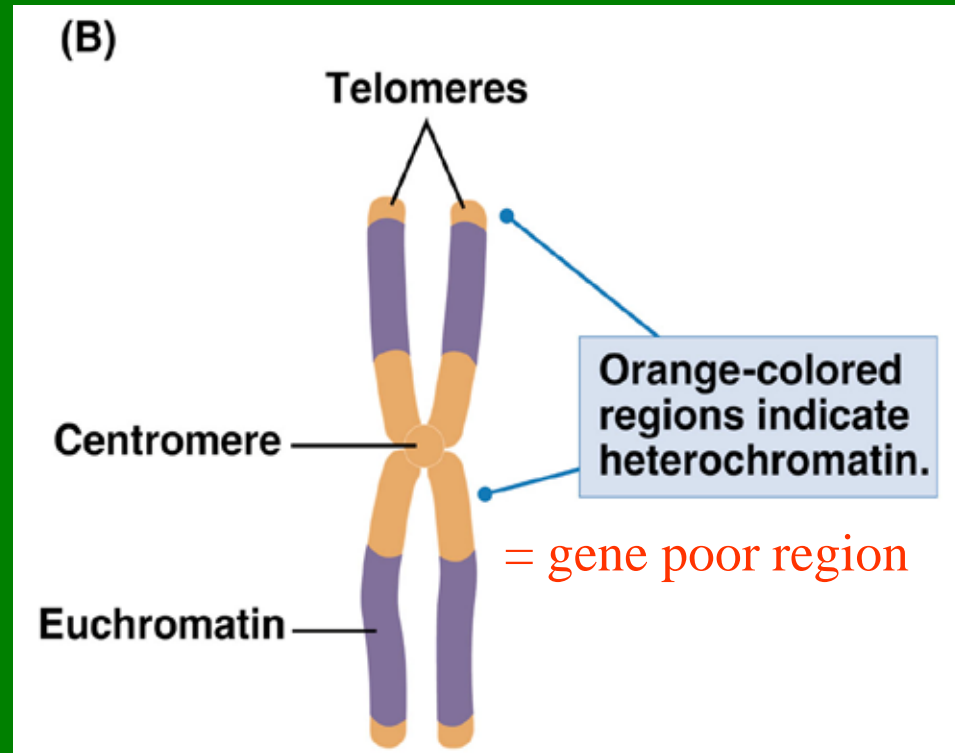
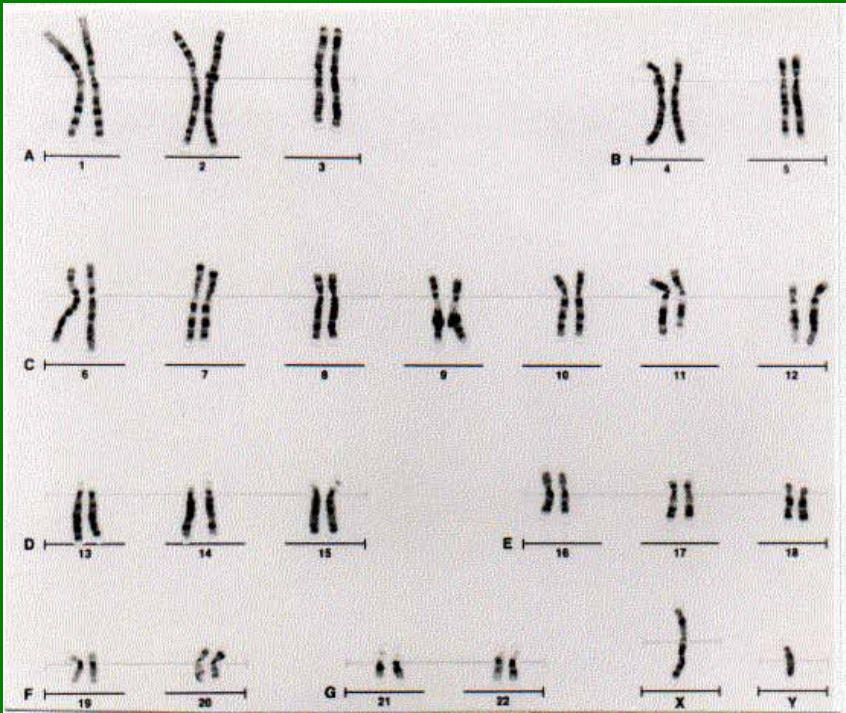


Nonrepetitive DNA is only part of the genome

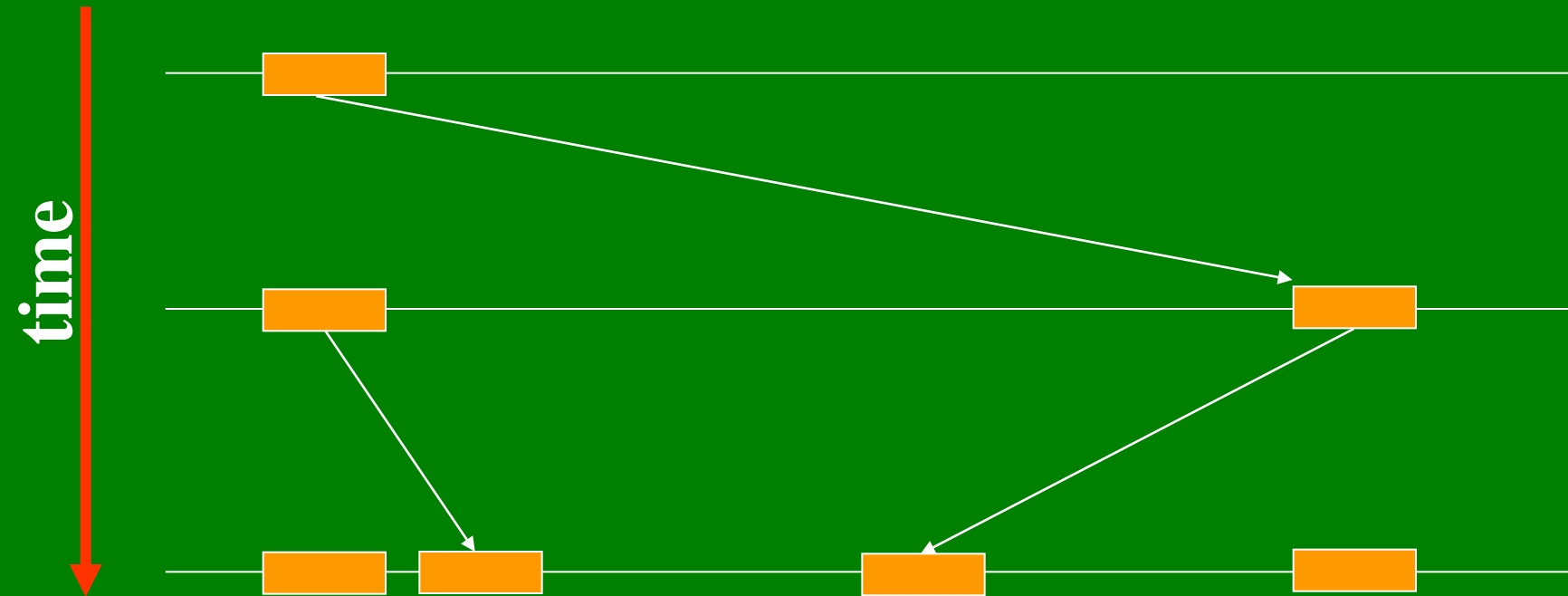
©virtualtext www.ergito.com



DNA with low complexity is located in the middle and at the ends of a chromosome

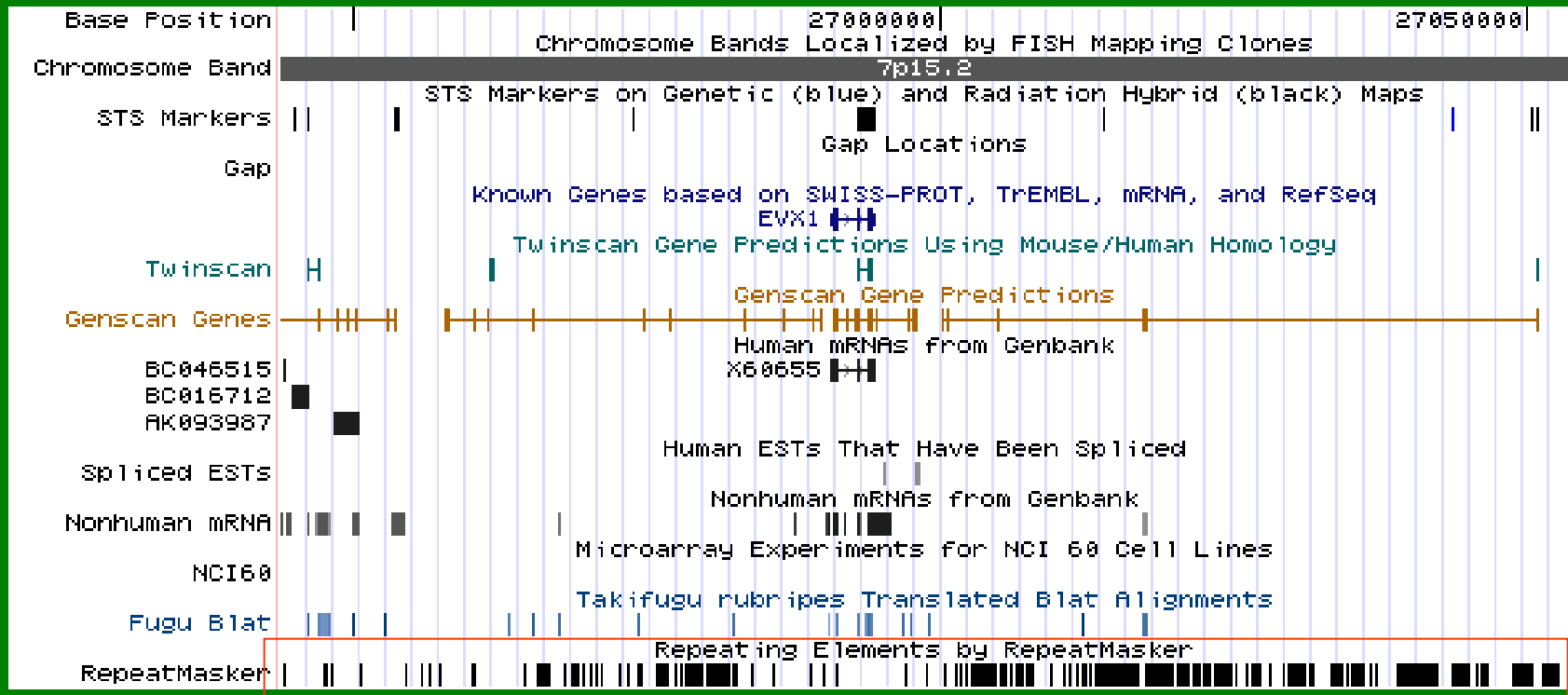


Imagine a gene that could duplicate copies of itself within the genome



Repeats dominate the human genome

One megabase from chromosome 7



Interspersed repeats

UCSC Genome Browser
<http://genome.cse.ucsc.edu>

Human genome: ~ 50% repetitive sequences

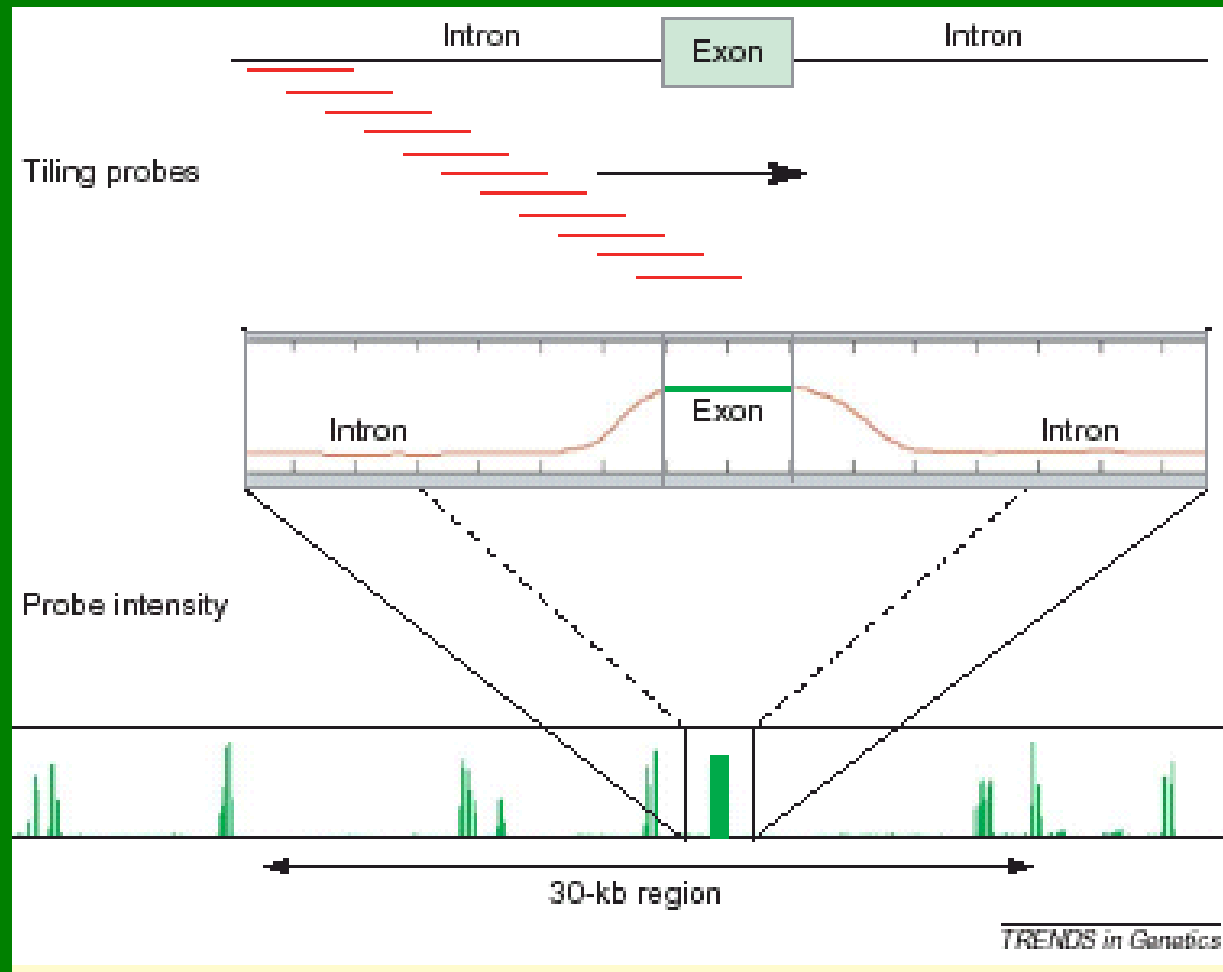
Transposable elements in the human genome

Type	Number of copies	Percentage of total genome
SINEs	1,558,000	13.1
<i>Alu</i>	1,090,000	10.6
LINEs	868,000	20.4
<i>LINE1</i>	516,000	16.9
LTR elements	443,000	8.3
DNA elements	294,000	2.8
<i>mariner</i>	14,000	0.1
Unclassified	3,000	0.1
Total of all types		44.7

Source: Data from E. S. Lander et al. 2001. *Nature* 409: 860.

Genome vs Transcriptome

How much of the genomic DNA is converted into RNA sequences?



Transcriptome facts

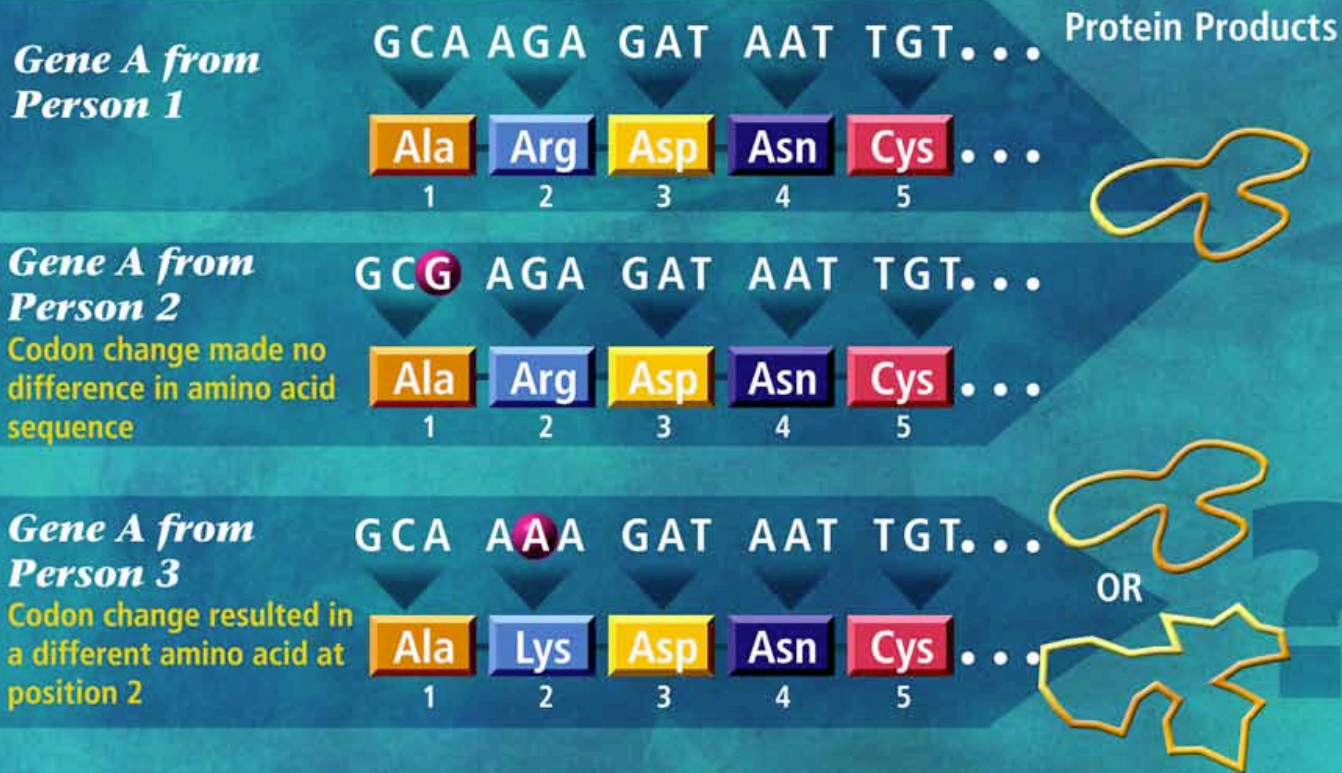
- **97-98% of transcriptional output is non coding RNA (ncRNA)**
- **Only 1.5% of genome are protein coding**
- **But: including introns, 30% of genome is transcribed**
- **Adding ncRNA genes: >50% of genome is transcribed**

Genotyping



Sequence variations between individuals

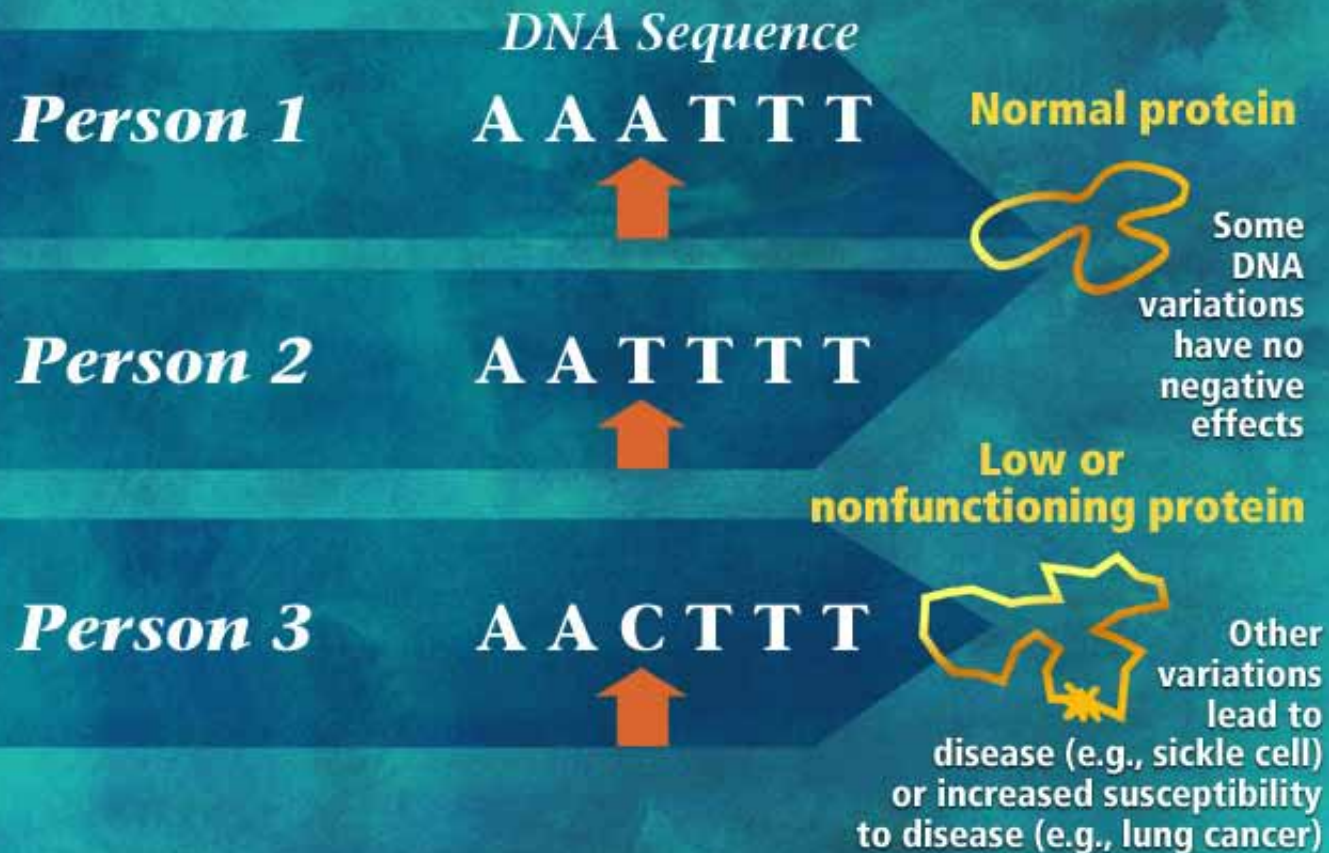
DNA Sequence Variation in a Gene Can Change the Protein Produced by the Genetic Code



Y-GA 98-649

Sequence variations between individuals

Health or Disease?



YGG-00-0480



Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution

Parallel deep sequencing of genomes and transcriptomes

Primary tumor and secondary tumor (metastases 9 years later) from same patient

Total depth of sequencing:

Genome:

2.9 billion reads
141,000 Gigabases
(43x coverage)

Transcriptome:

180 Million reads
7,000 Gigabases

Comparison of all mutations present in primary tumor versus genome mutations present in metastases.

In addition: epigenetic changes in primary vs secondary tumor

The Genome Sequence -- an open book

(written in well-known language but poorly understood grammar)

How make sense of whole genome sequences?

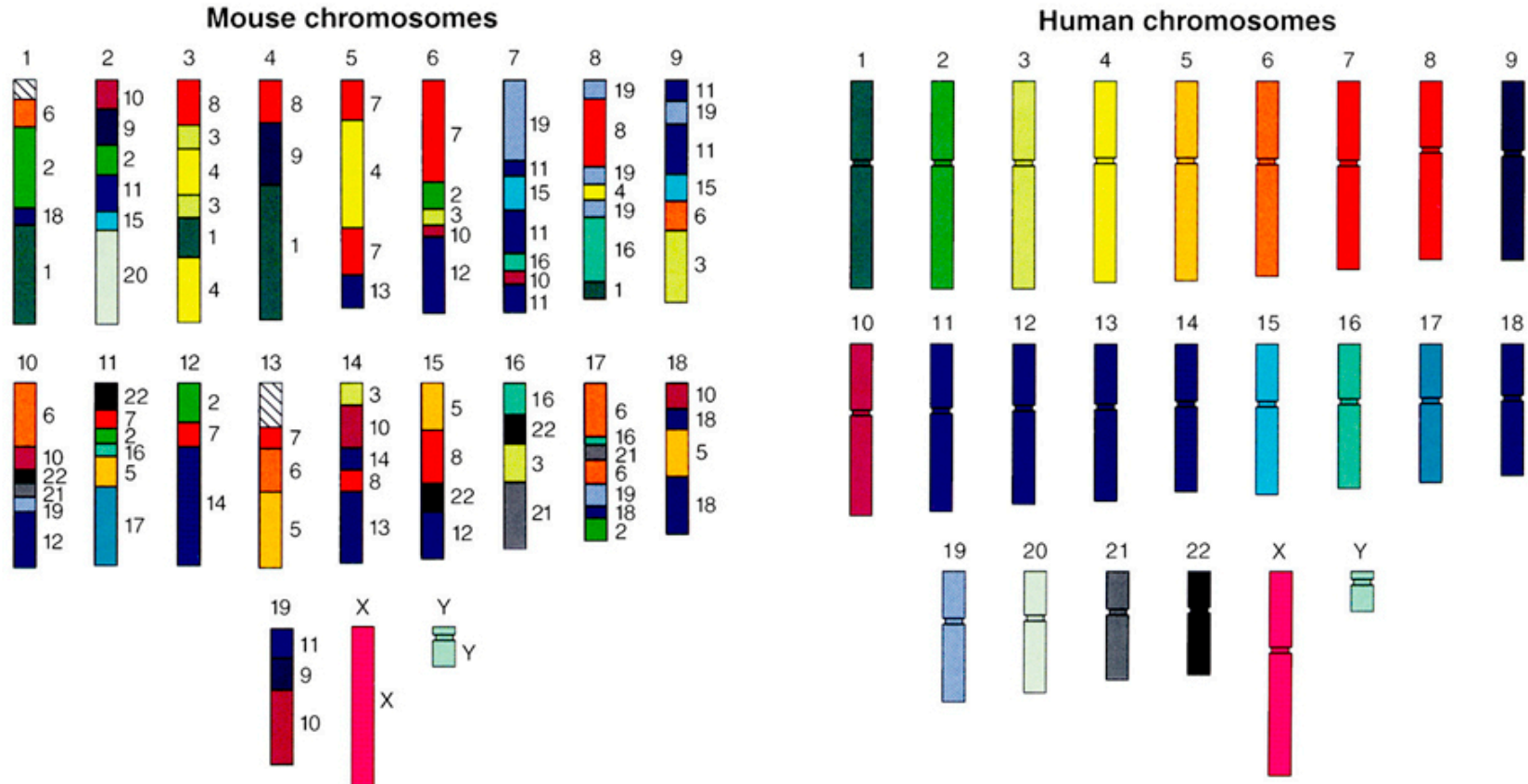
Bioinformatics helps to:

- ∅ Find regulatory sequences
- ∅ Find protein coding sequences
- ∅ Find related sequences
- ∅ Compare sequences across species

Main aim:

understand what are the functions of all genes and how they work together

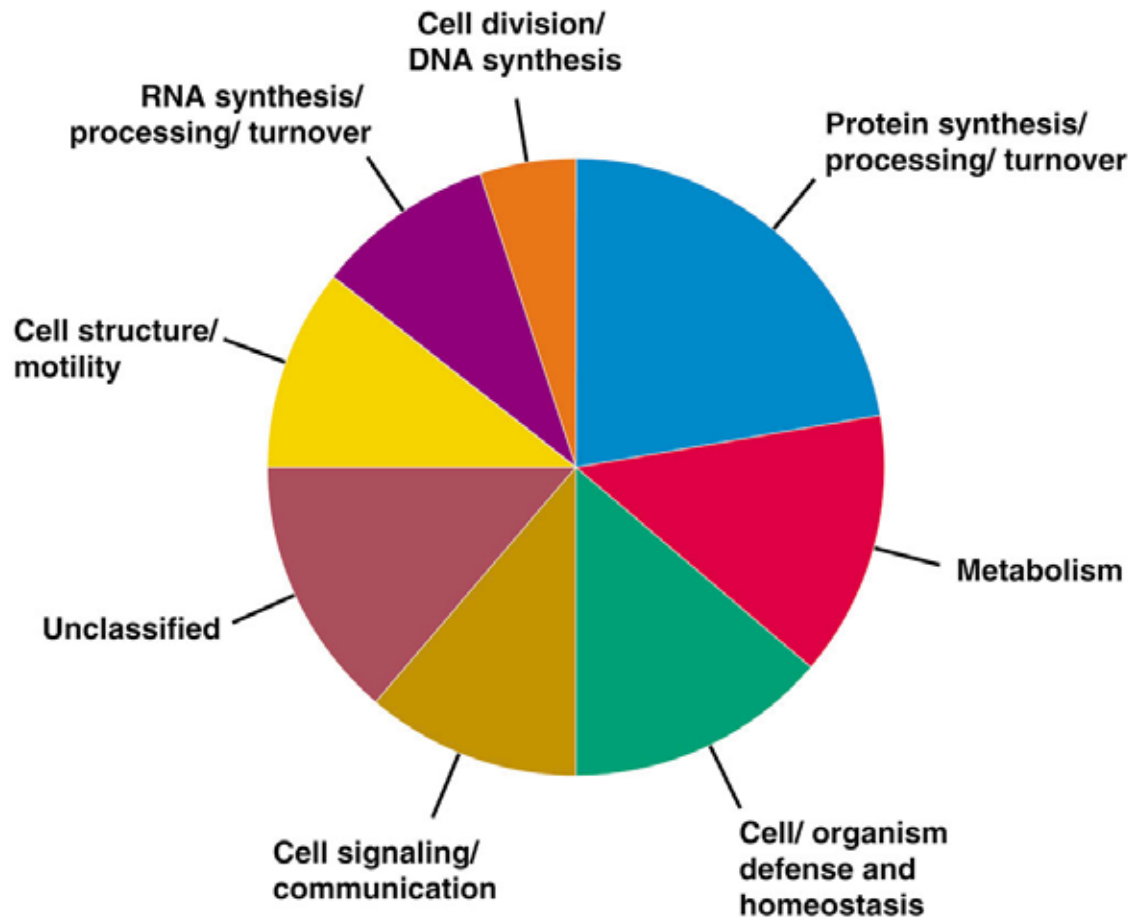
Mouse and Human Genetic Similarities



Courtesy Lisa Stubbs
Oak Ridge National Laboratory

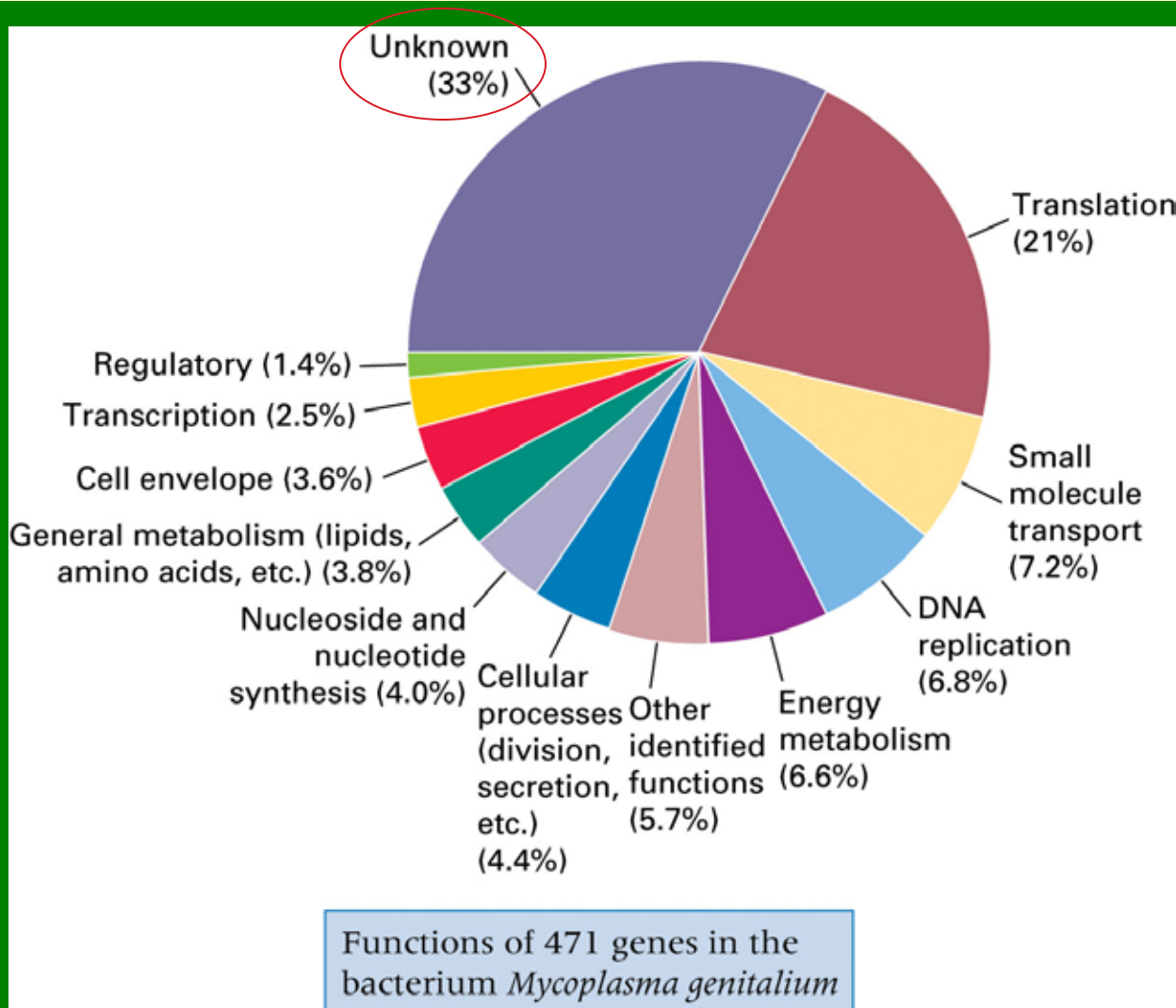
Functional classification of expressed genes

Collection of expressed genes



Randomly
selected
sequences from
human cells
grouped by
function

Genes in the *Mycoplasma genitalium* classified by function



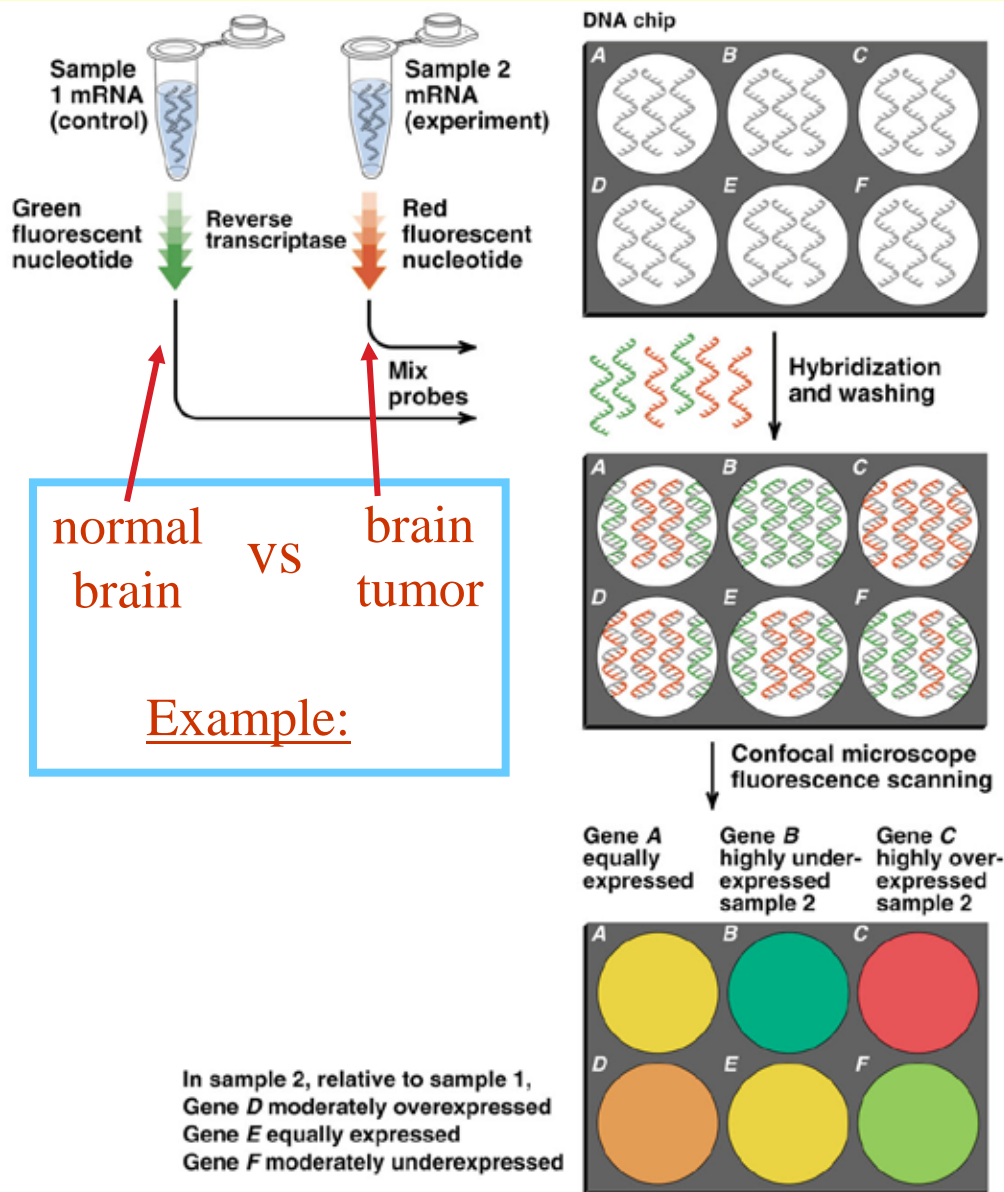
Smallest genome of any known free living organism

è What is the smallest number of genes needed for survival?

580 kb genome \leftrightarrow 471 genes

Functional genomics I

Use of DNA microarrays (chips)

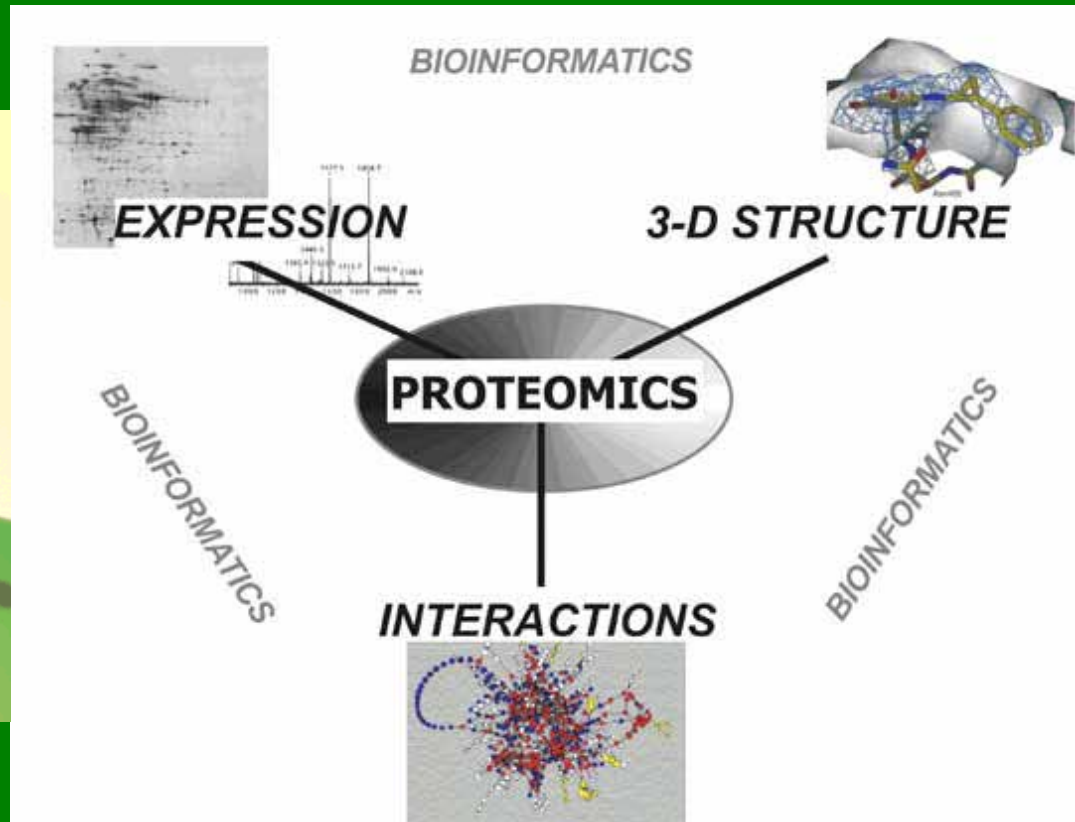
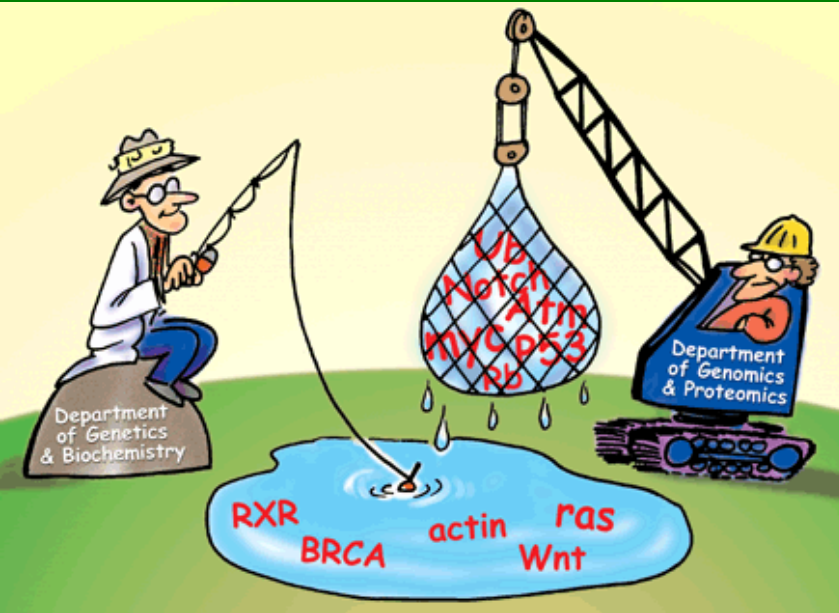


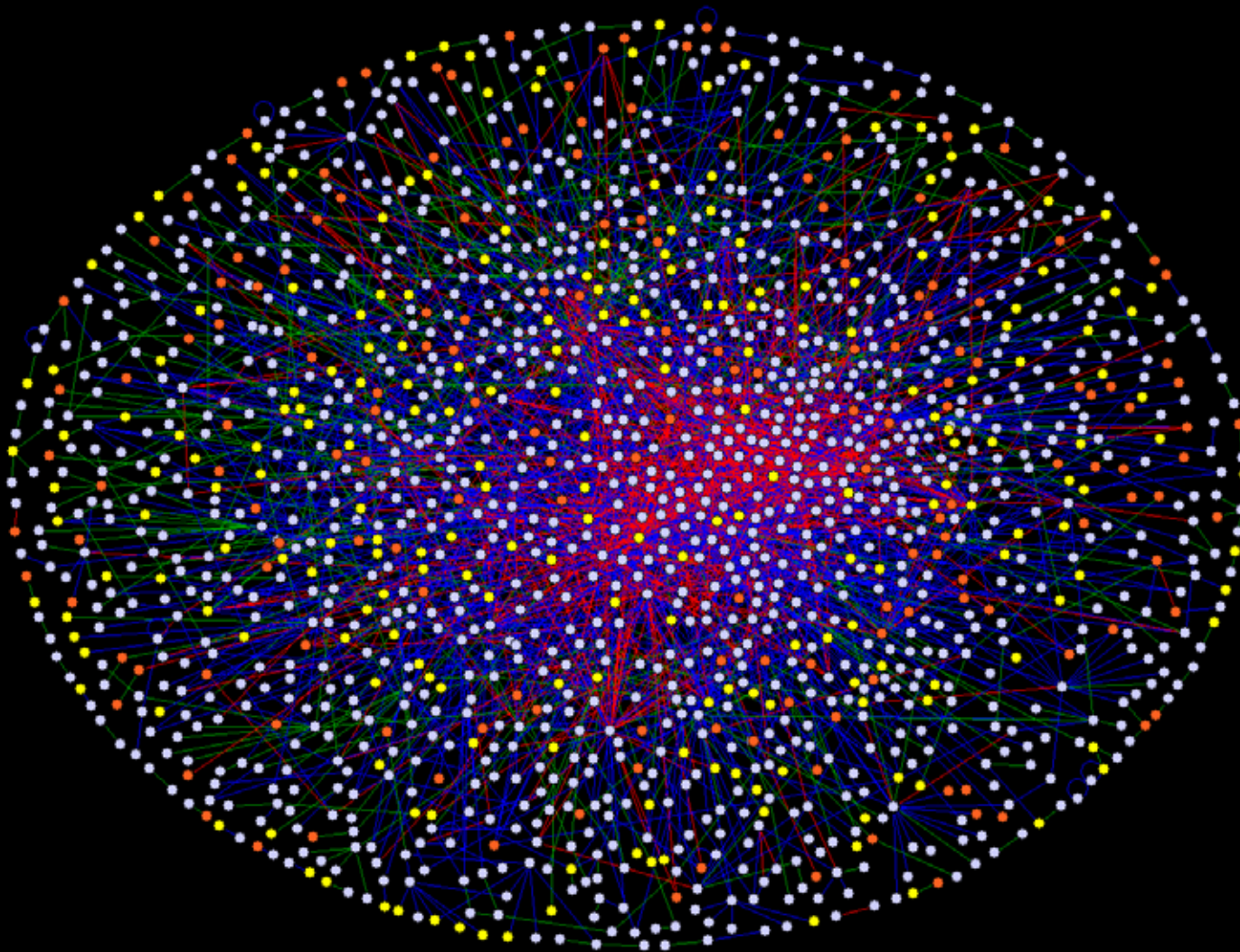
normal brain vs brain tumor

Example:

Fluorescently tagged cDNA probes are hybridized to DNA spots in the microarray for studying differential expression of thousands of genes at a time in two mRNA samples

What is Proteomics





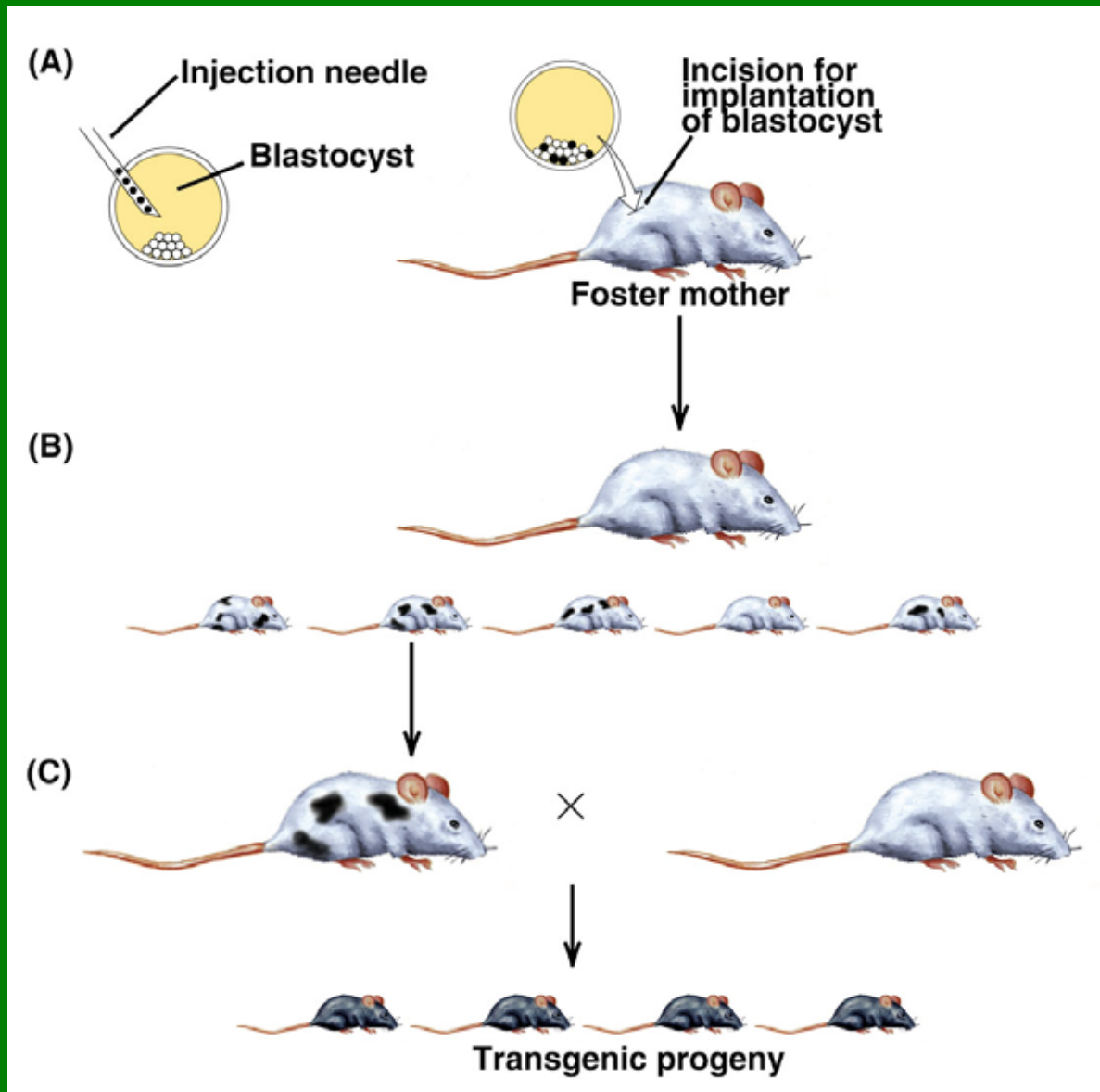
Colored circles represent proteins (nodes):

- Light blue, known proteins
- Orange, disease proteins
- Yellow, uncharacterized proteins

Interactions (links) are represented by color-coded lines:

- Red, high confidence (HC) interactions
- Blue, medium confidence (MC) interactions
- Green, low confidence (LC) interactions

Steps in the creation of a transgenic mouse



Summary - Conclusions

We do know:

- Complete human genome sequence
- Function of ca. 50% of protein-coding genes
- Some interactions and regulation of genes
- > 2700 disease-causing single gene mutations

We don't know:

- Function of at least 50% of genes
- How do gene products work together ? (within cells and within organism)
- What makes the human so special ? (its not the gene number or genome size)