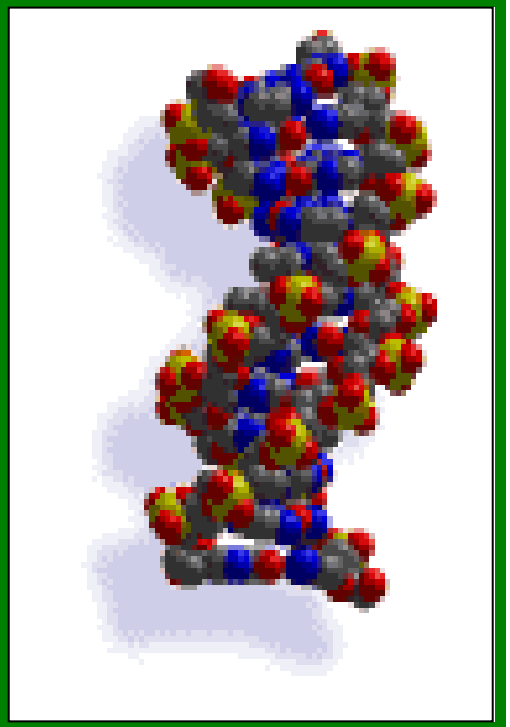
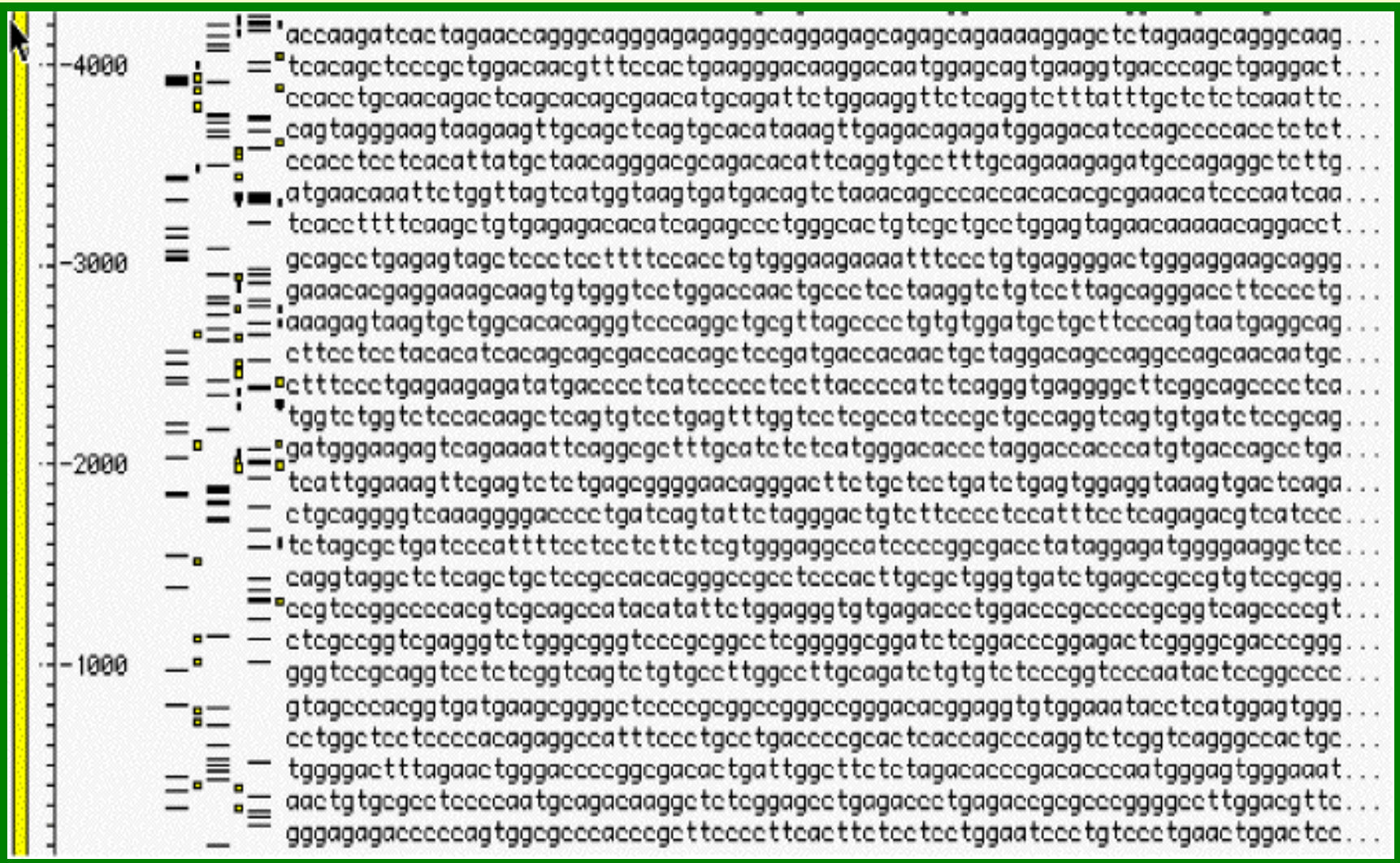
Genomes: What we know … and what we don't know

Complete draft sequence 2001



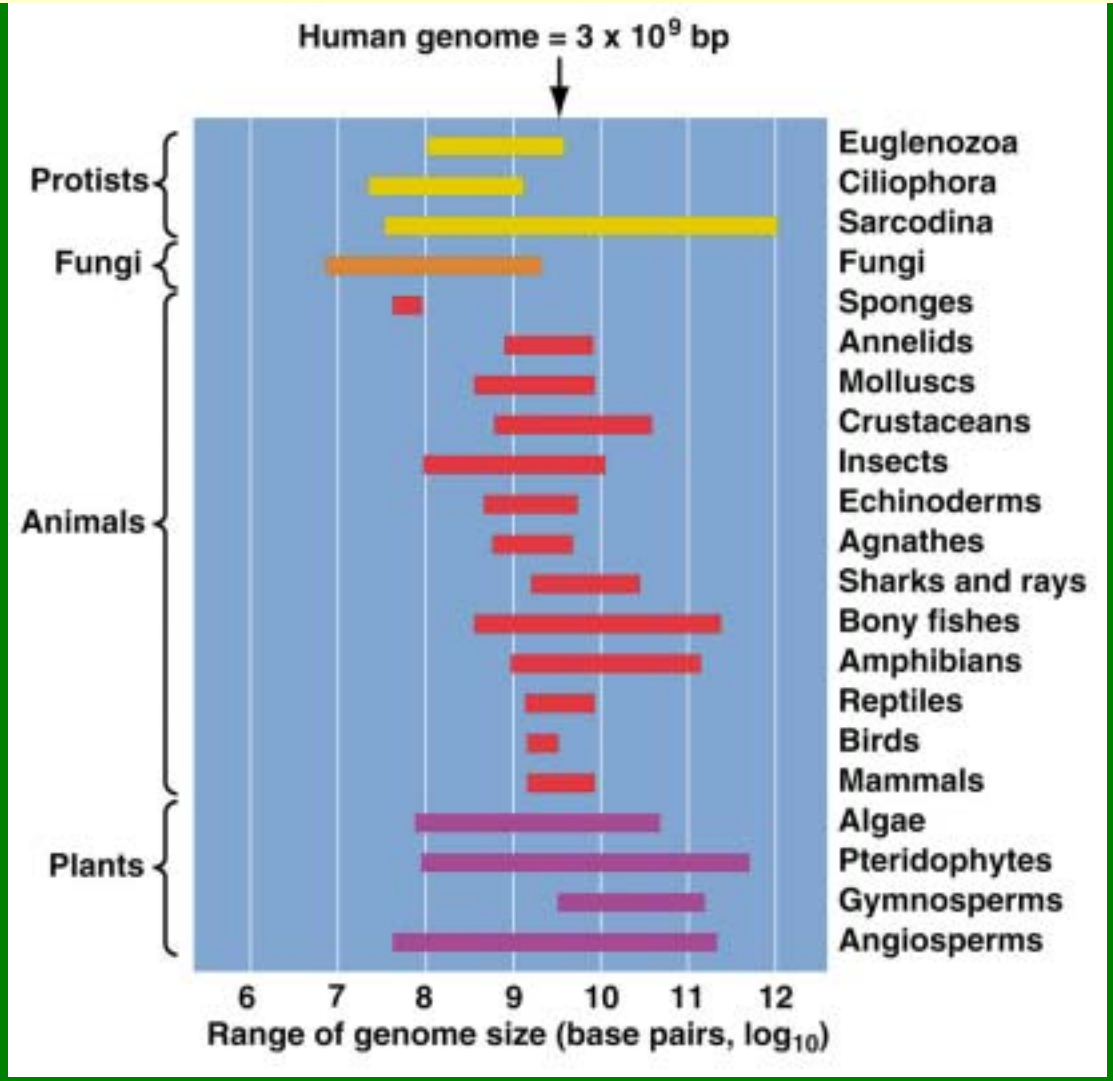
October 15, 2007 *Dr. Stefan Maas, BioS Lehigh U.*

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

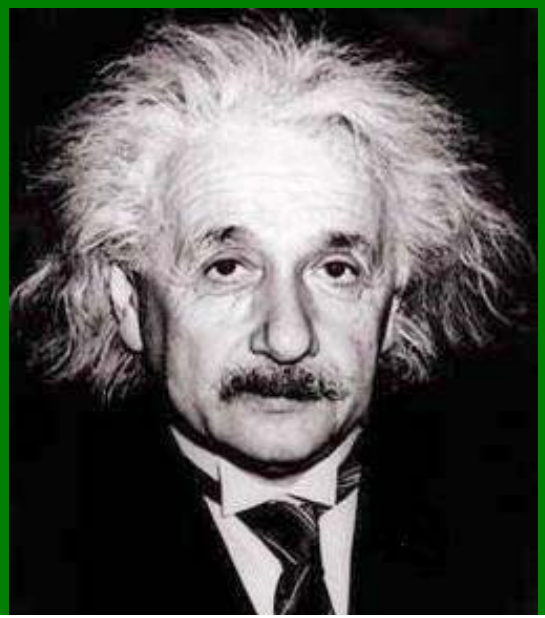
180 Mb

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



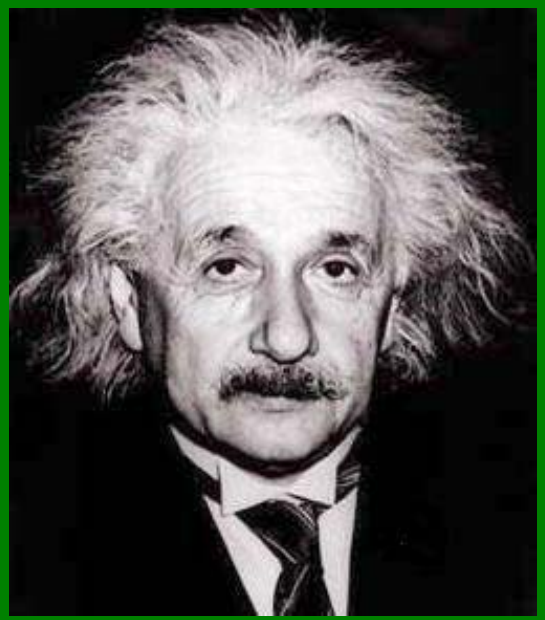
3.4×10^{9} bp

Homo sapiens

6.7×10^{11} bp

Amoeba dubia

Complexity does not correlate with gene number

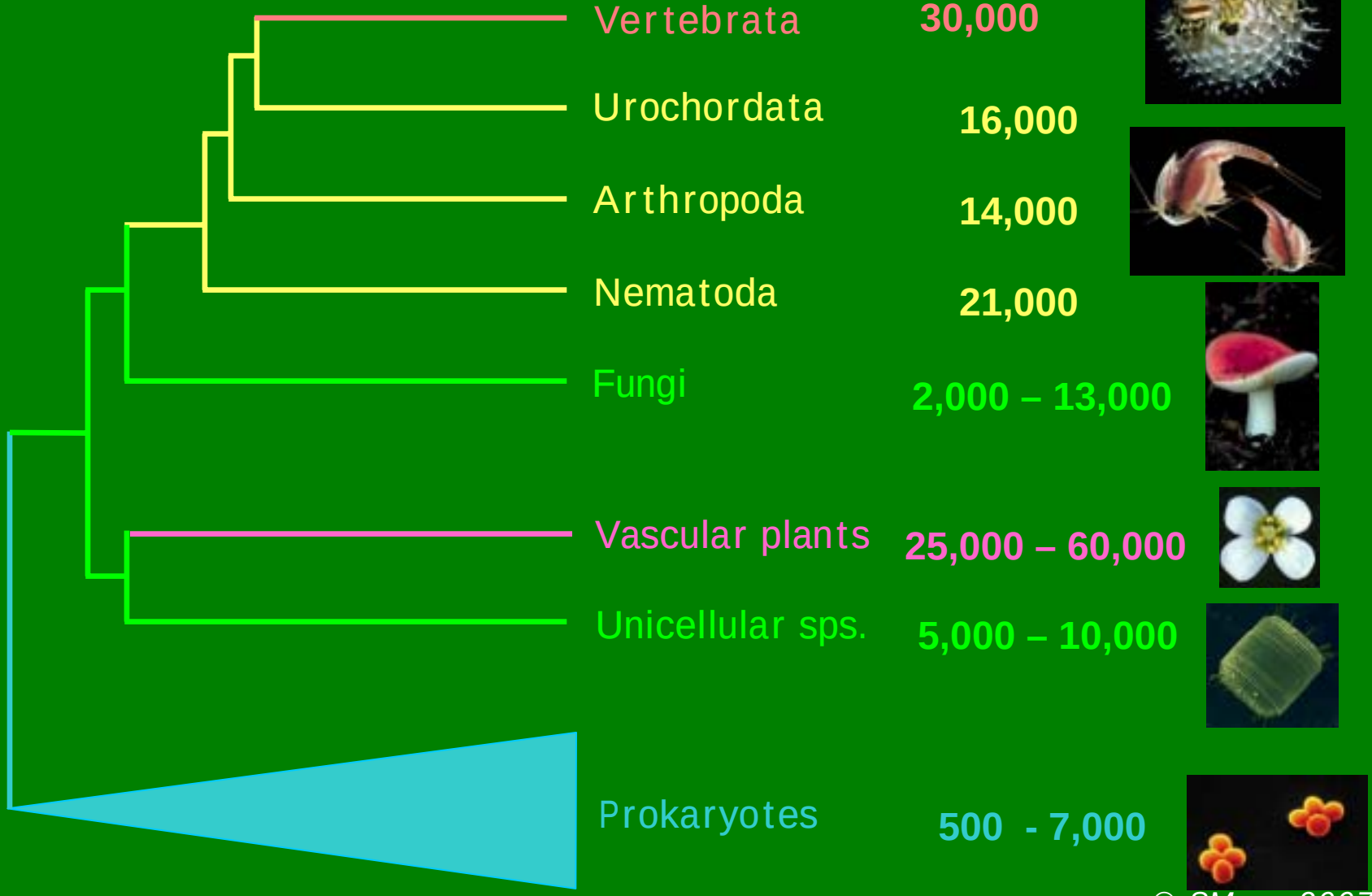


~31,000 genes

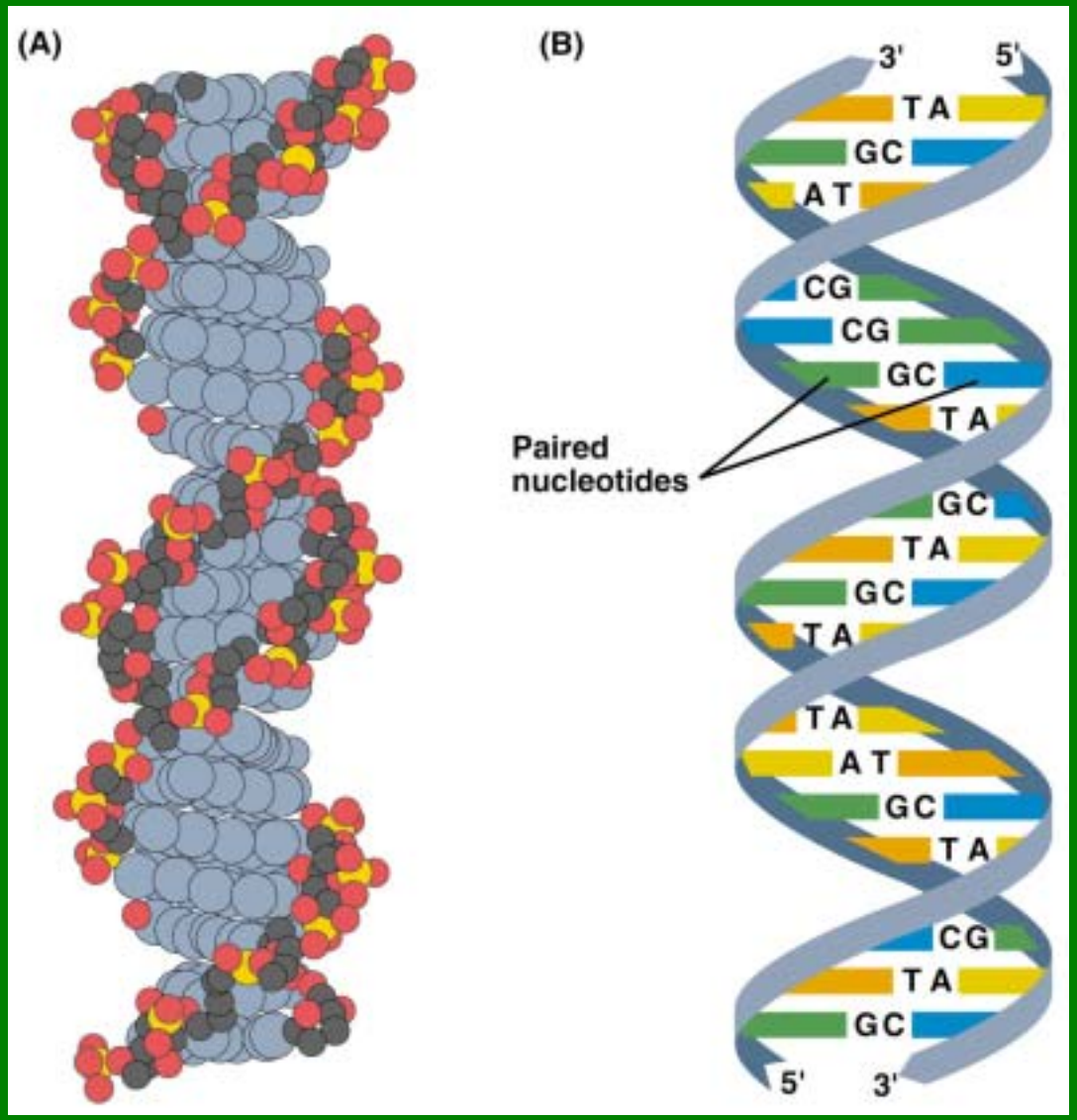
~26,000 genes

~50,000 genes

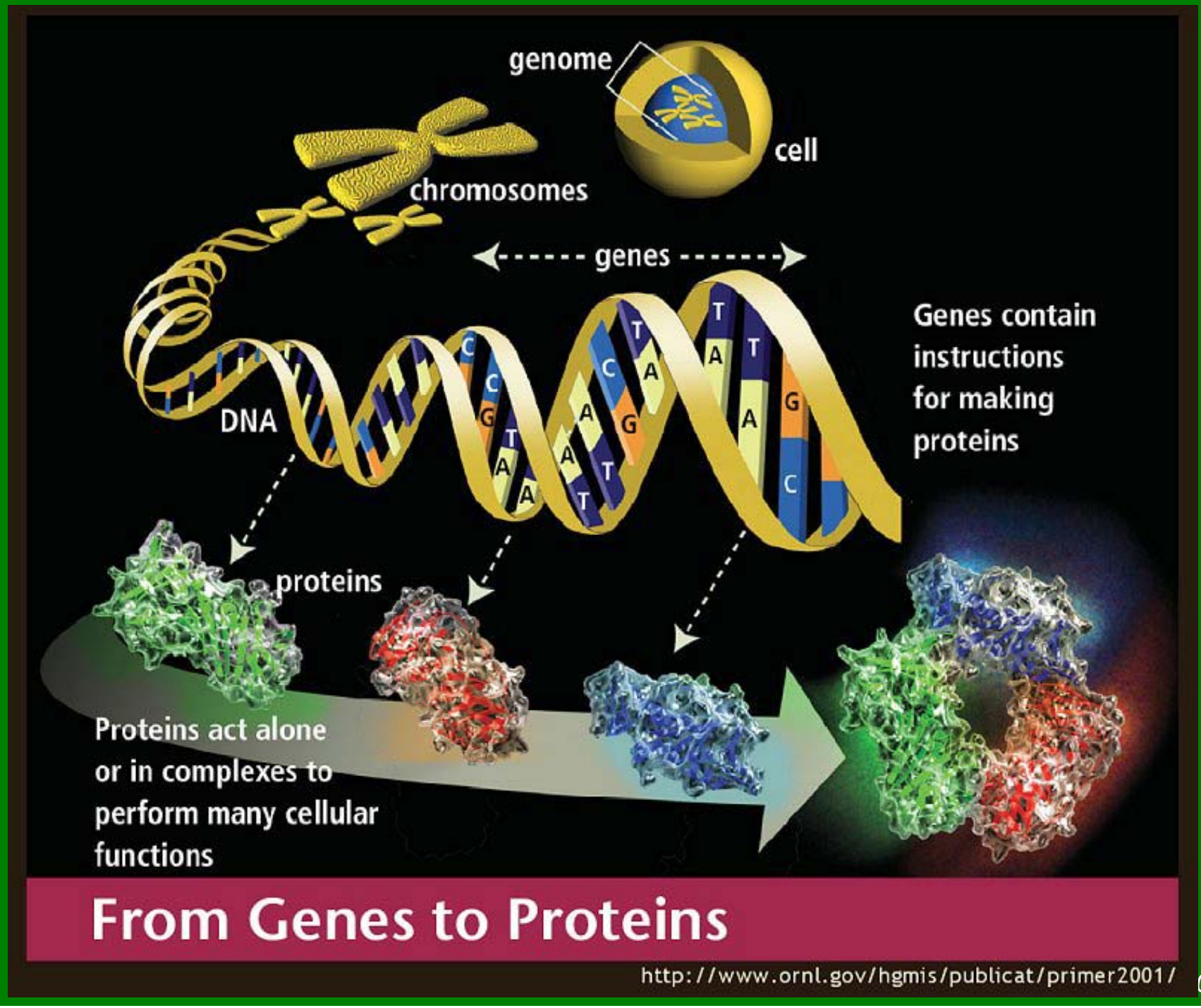
Is an Expansion in Gene Number driving Evolution of Higher Organisms?

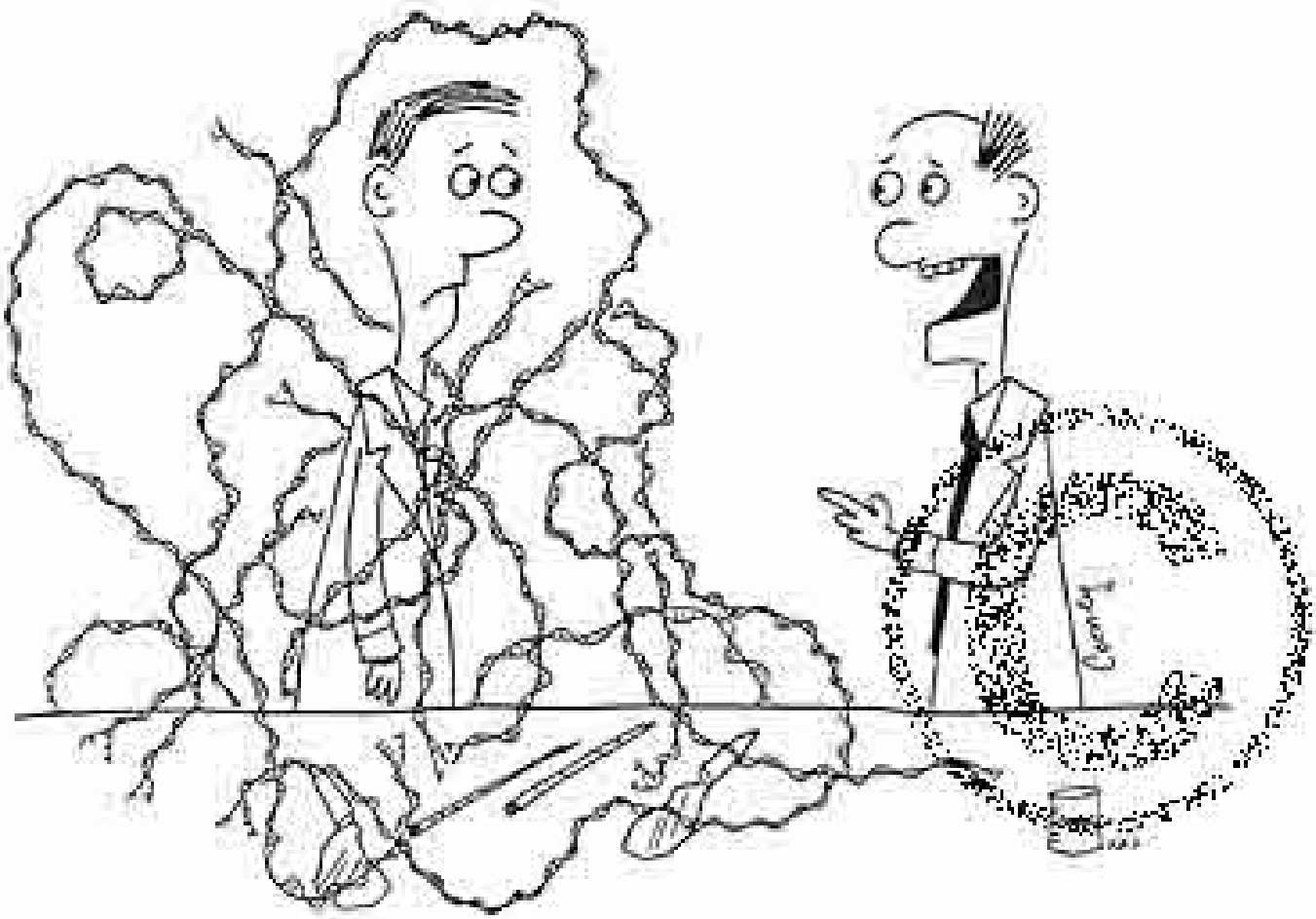


Structure of DNA



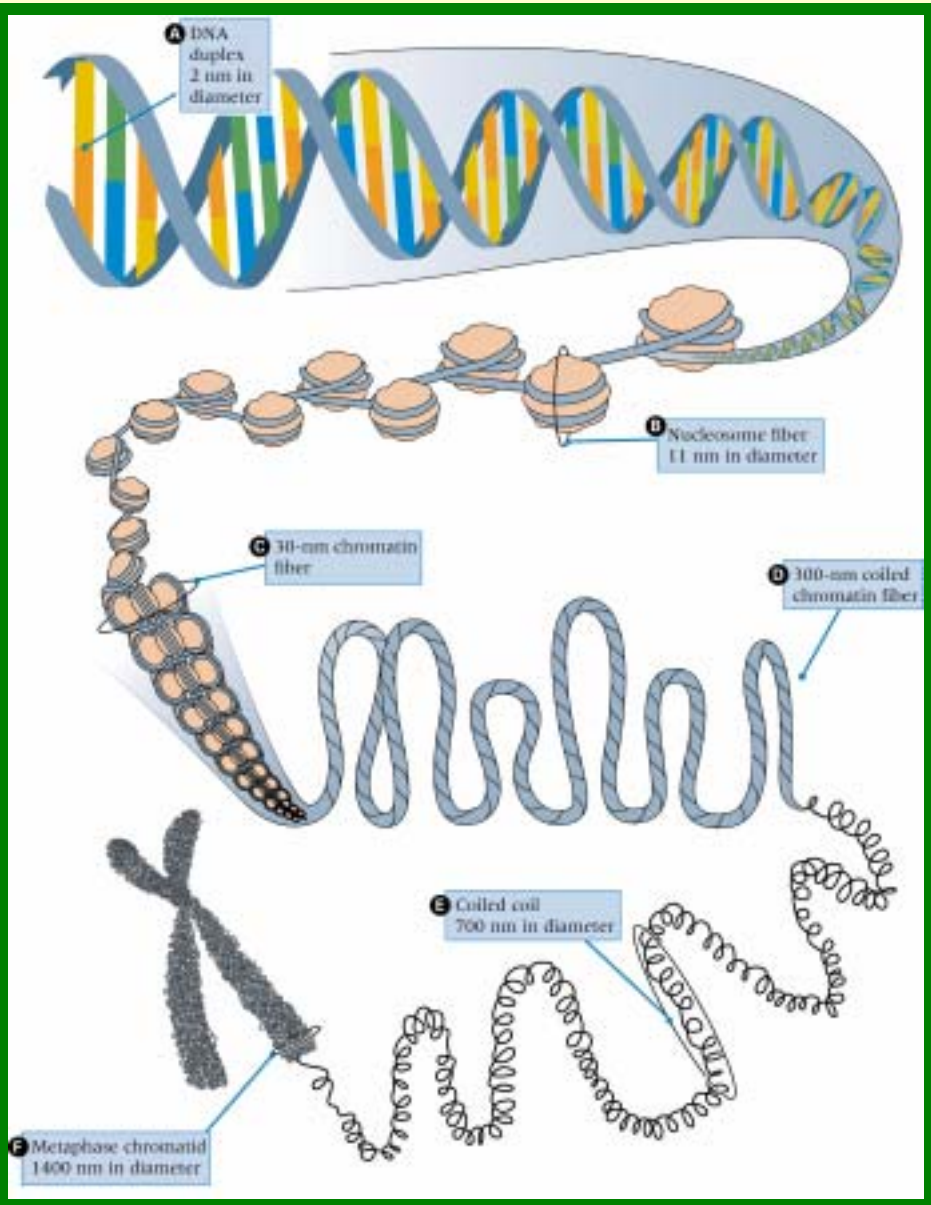
Watson and Crick in 1953 proposed that DNA is a double helix in which the 4 bases are base paired, Adenine (A) with Thymine (T) and Guanine (G) with Cytosine (C).


genome
cell
chromosomes
genes
DNA
Genes contain instructions for making proteins
proteins
Proteins act alone or in complexes to perform many cellular functions
From Genes to Proteins
http://www.ornl.gov/hgmis/publicat/primer2001/



"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)

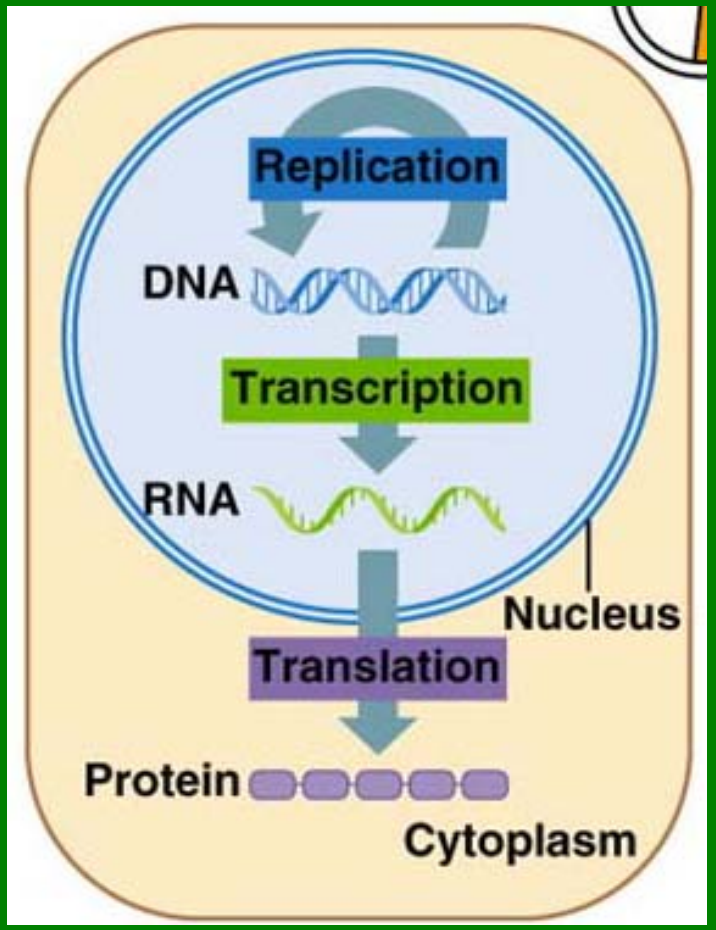
Factor of condensation:
Ca. 10,000 fold

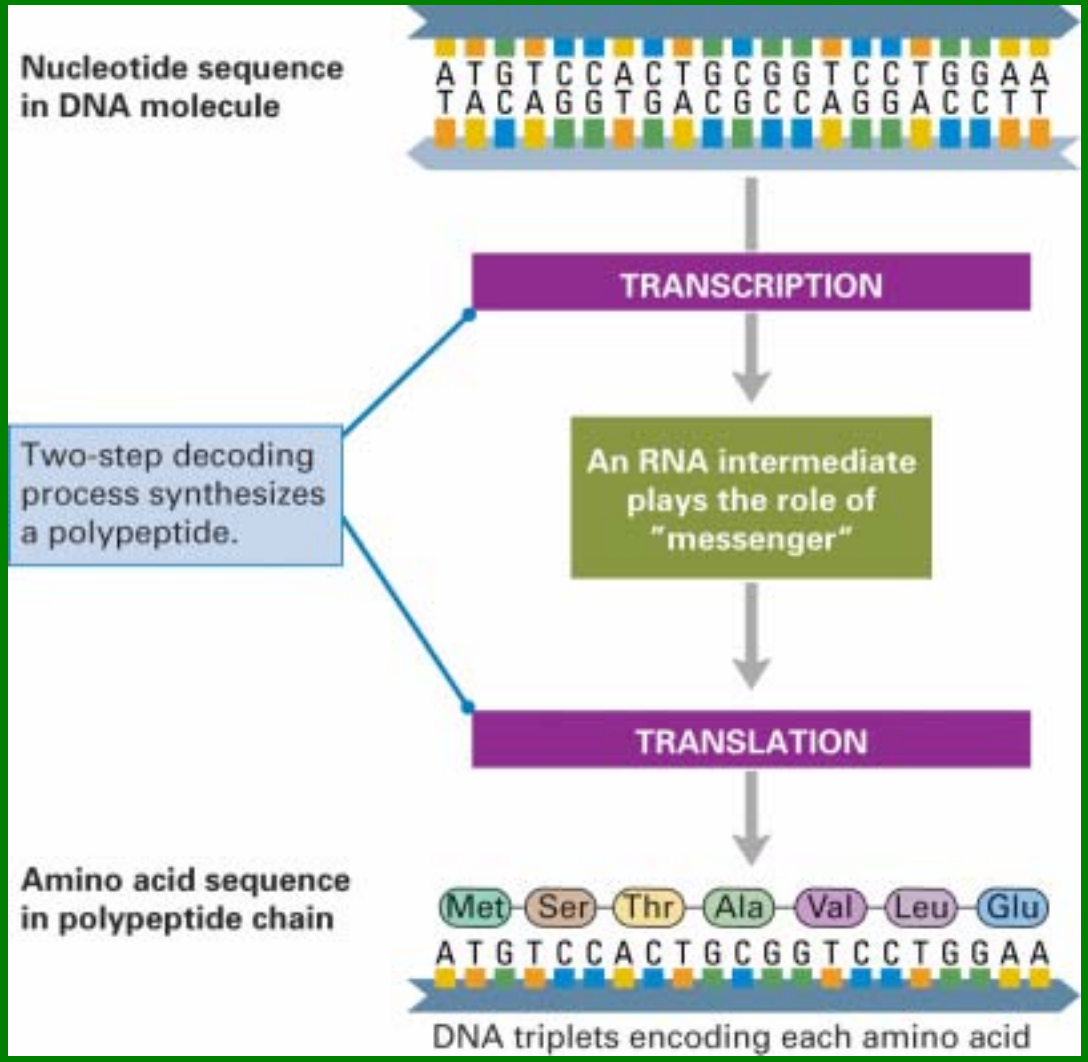
Why?

➔ Facilitates movement during cell division

➔ Decreases error rate

Genes code for Proteins with RNA as an intermediate





The Genetic Code

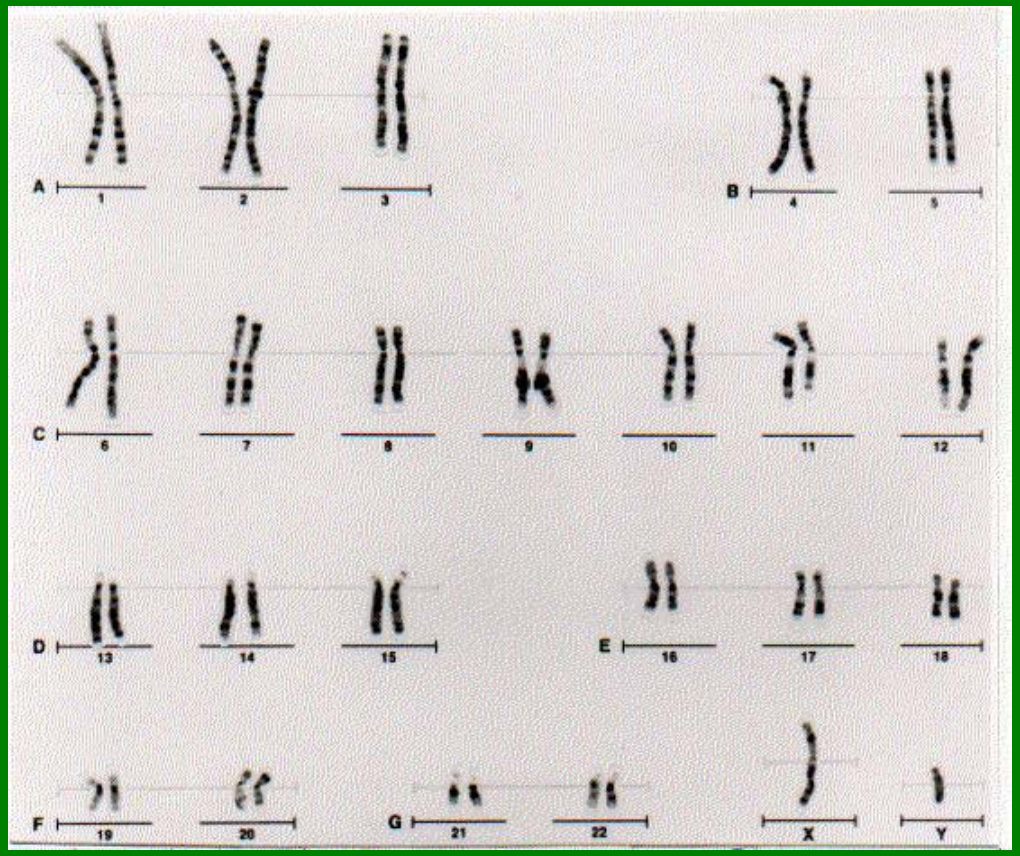
Table 1.1 The standard genetic code

First nucleotide in codon (5′ end)	Second nucleotide in codon: U				Second nucleotide in codon: C				Second nucleotide in codon: A				Second nucleotide in codon: G				Third nucleotide in codon (3′ end)
U	UUU	Phe	F	*Phenylalanine*	UCU	Ser	S	*Serine*	UAU	Tyr	Y	*Tyrosine*	UGU	Cys	C	*Cysteine*	U
	UUC	Phe	F	*Phenylalanine*	UCC	Ser	S	*Serine*	UAC	Tyr	Y	*Tyrosine*	UGC	Cys	C	*Cysteine*	C
	UUA	Leu	L		UCA	Ser	S	*Serine*	UAA	Termination			UGA	Termination			A
	UUG	Leu	L		UCG	Ser	S	*Serine*	UAG	Termination			UGG	Trp	W	*Tryptophan*	G
C	CUU	Leu	L		CCU	Pro	P	*Proline*	CAU	His	H	*Histidine*	CGU	Arg	R	*Arginine*	U
	CUC	Leu	L		CCC	Pro	P	*Proline*	CAC	His	H	*Histidine*	CGC	Arg	R	*Arginine*	C
	CUA	Leu	L		CCA	Pro	P	*Proline*	CAA	Gln	Q	*Glutamine*	CGA	Arg	R	*Arginine*	A
	CUG	Leu	L		CCG	Pro	P	*Proline*	CAG	Gln	Q	*Glutamine*	CGG	Arg	R	*Arginine*	G
A	AUU	Ile	I	*Isoleucine*	ACU	Thr	T	*Threonine*	AAU	Asn	N	*Asparagine*	AGU	Ser	S	*Serine*	U
	AUC	Ile	I	*Isoleucine*	ACC	Thr	T	*Threonine*	AAC	Asn	N	*Asparagine*	AGC	Ser	S	*Serine*	C
	AUA	Ile	I	*Isoleucine*	ACA	Thr	T	*Threonine*	AAA	Lys	K	*Lysine*	AGA	Arg	R	*Arginine*	A
	AUG	Met	M	*Methionine*	ACG	Thr	T	*Threonine*	AAG	Lys	K	*Lysine*	AGG	Arg	R	*Arginine*	G
G	GUU	Val	V	*Valine*	GCU	Ala	A	*Alanine*	GAU	Asp	D	*Aspartic acid*	GGU	Gly	G	*Glycine*	U
	GUC	Val	V	*Valine*	GCC	Ala	A	*Alanine*	GAC	Asp	D	*Aspartic acid*	GGC	Gly	G	*Glycine*	C
	GUA	Val	V	*Valine*	GCA	Ala	A	*Alanine*	GAA	Glu	E	*Glutamic acid*	GGA	Gly	G	*Glycine*	A
	GUG	Val	V	*Valine*	GCG	Ala	A	*Alanine*	GAG	Glu	E	*Glutamic acid*	GGG	Gly	G	*Glycine*	G

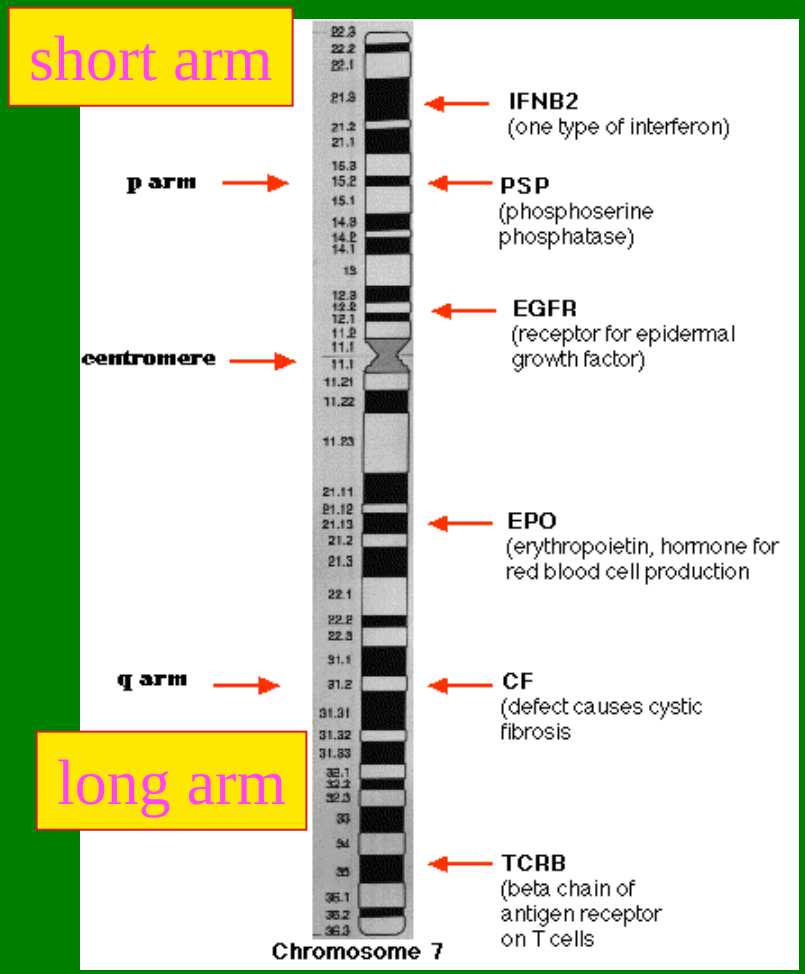
Codon

Three-letter and single-letter abbreviations

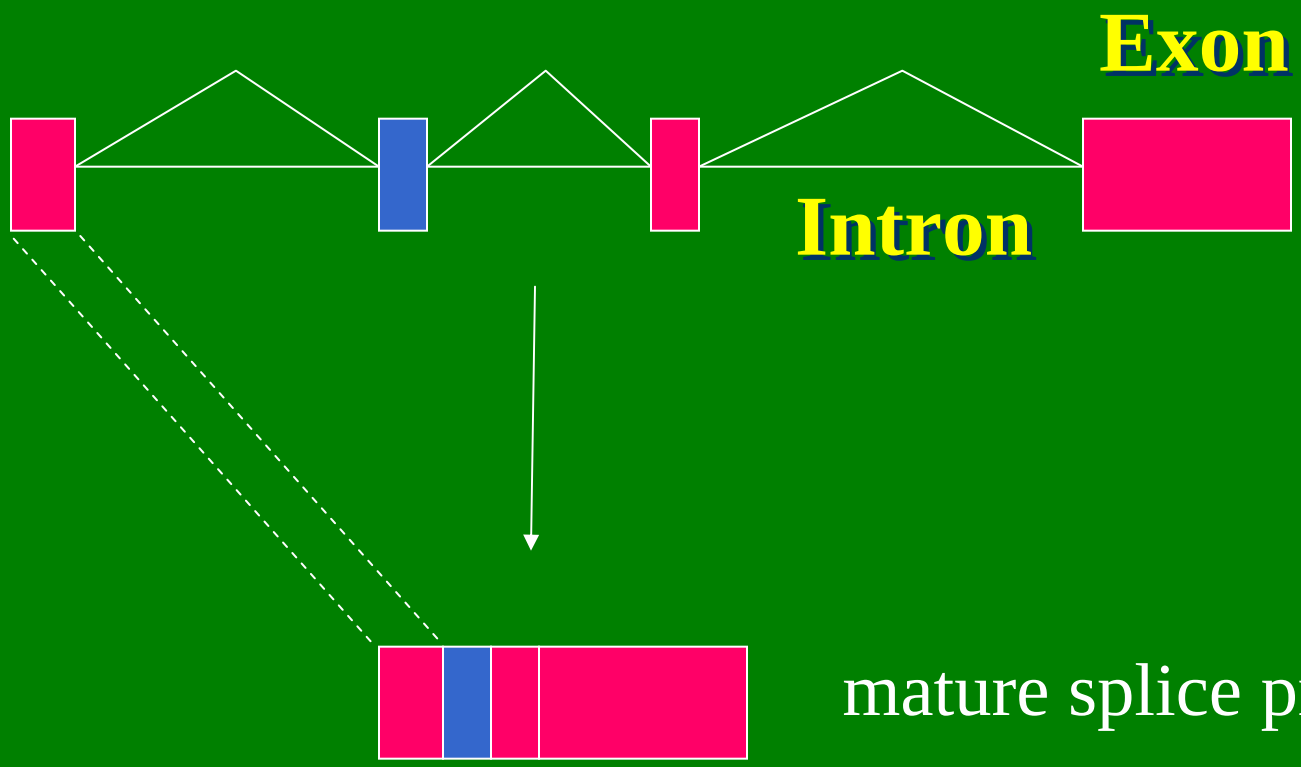
The human karyotype



23 pairs of chromosomes



Gene splicing:
Removal of non-coding introns



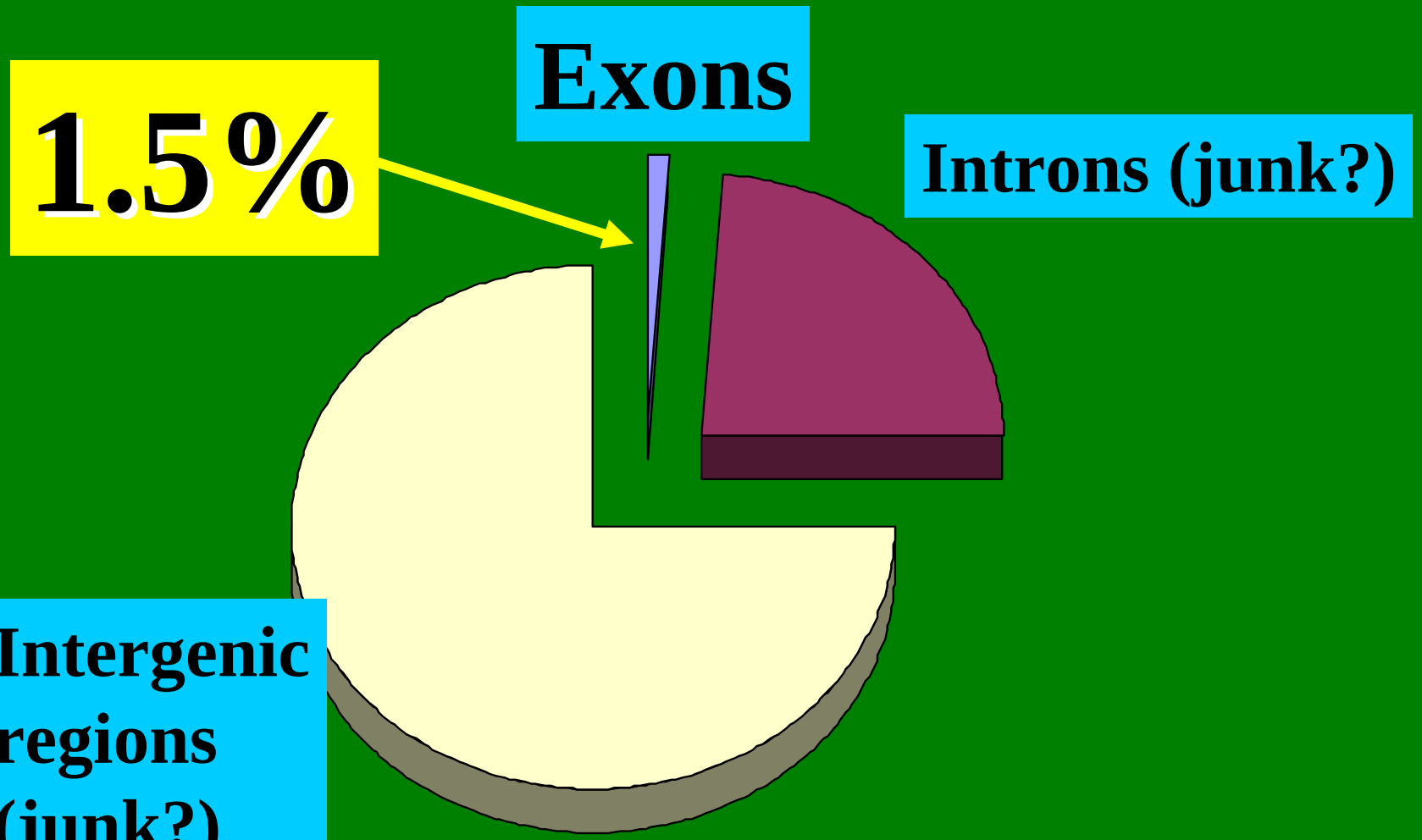
mature splice product

PROTEIN

Human genome facts

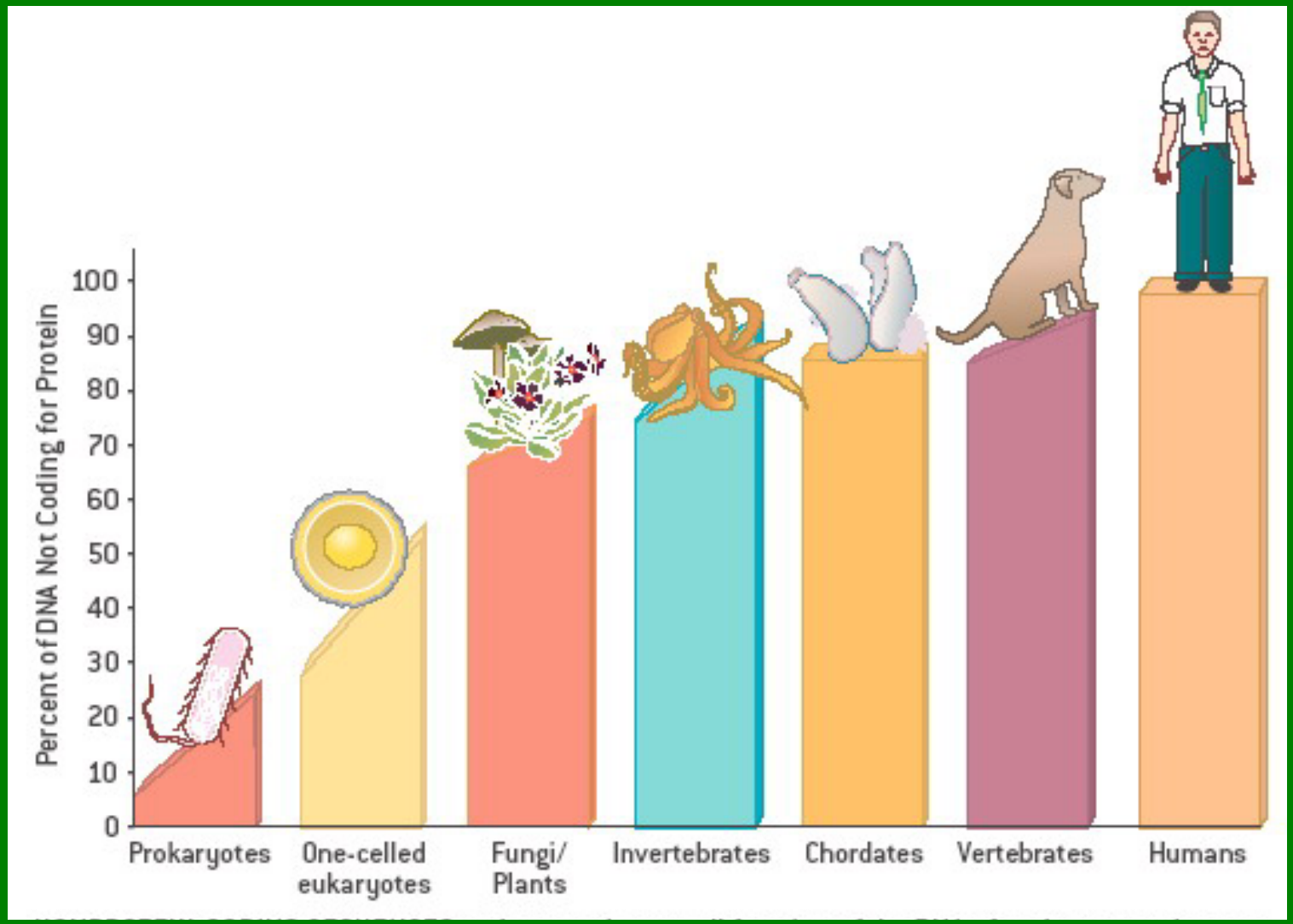
- ~ 30,000 genes
- On average:
  - Coding length: 1.4 kb
  - Gene extent: 30 kb
  - 8 Exons (135 bp), 7 introns (2,200 bp)
  - Gene density: 11.5/1 Mb

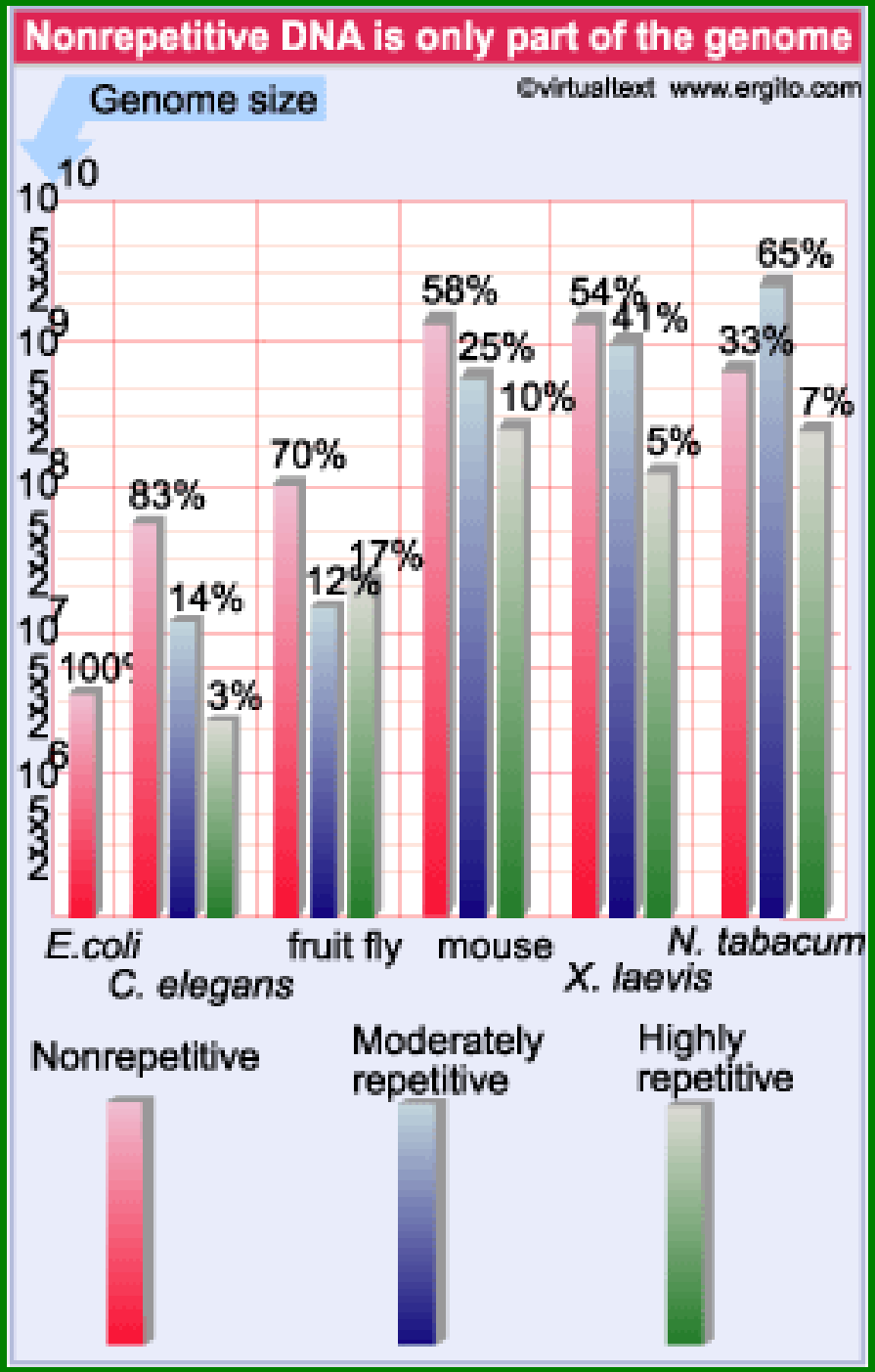
Is the genome empty ?



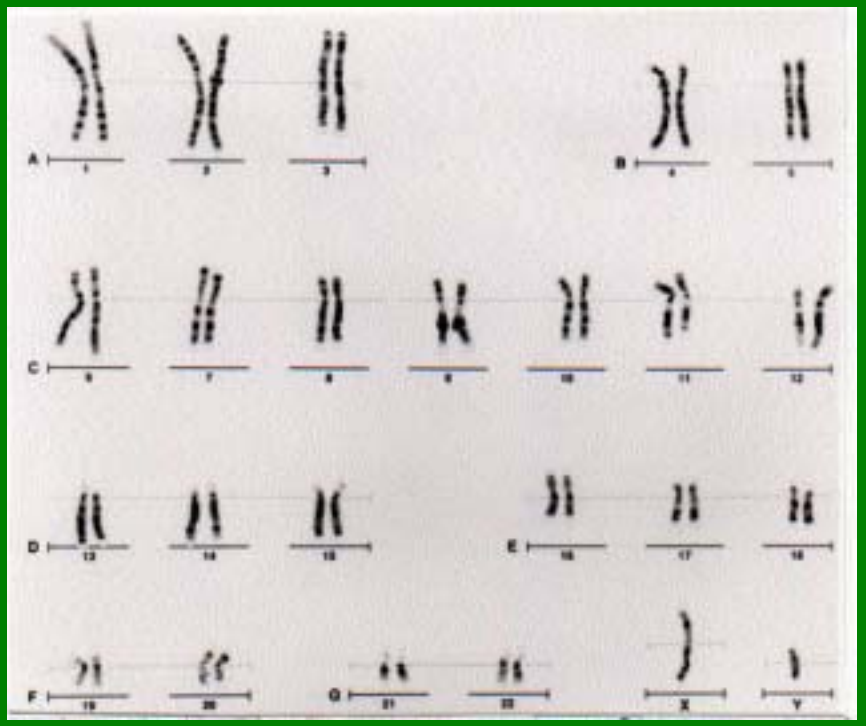
Correlation between complexity and amount of non-coding DNA

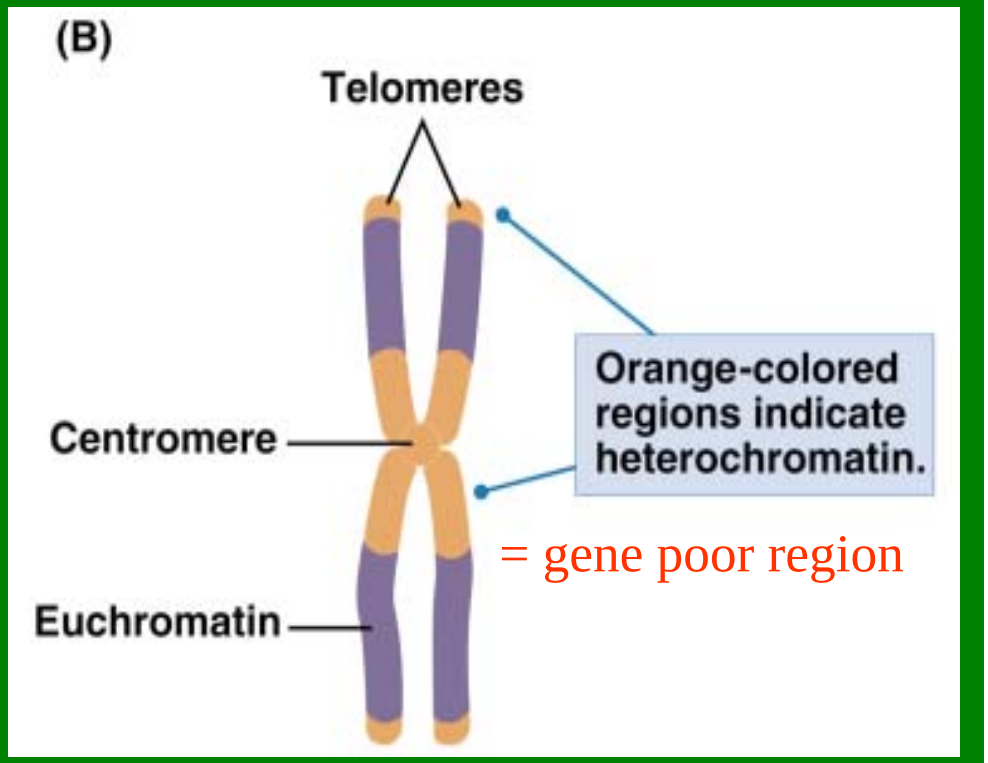
1.0



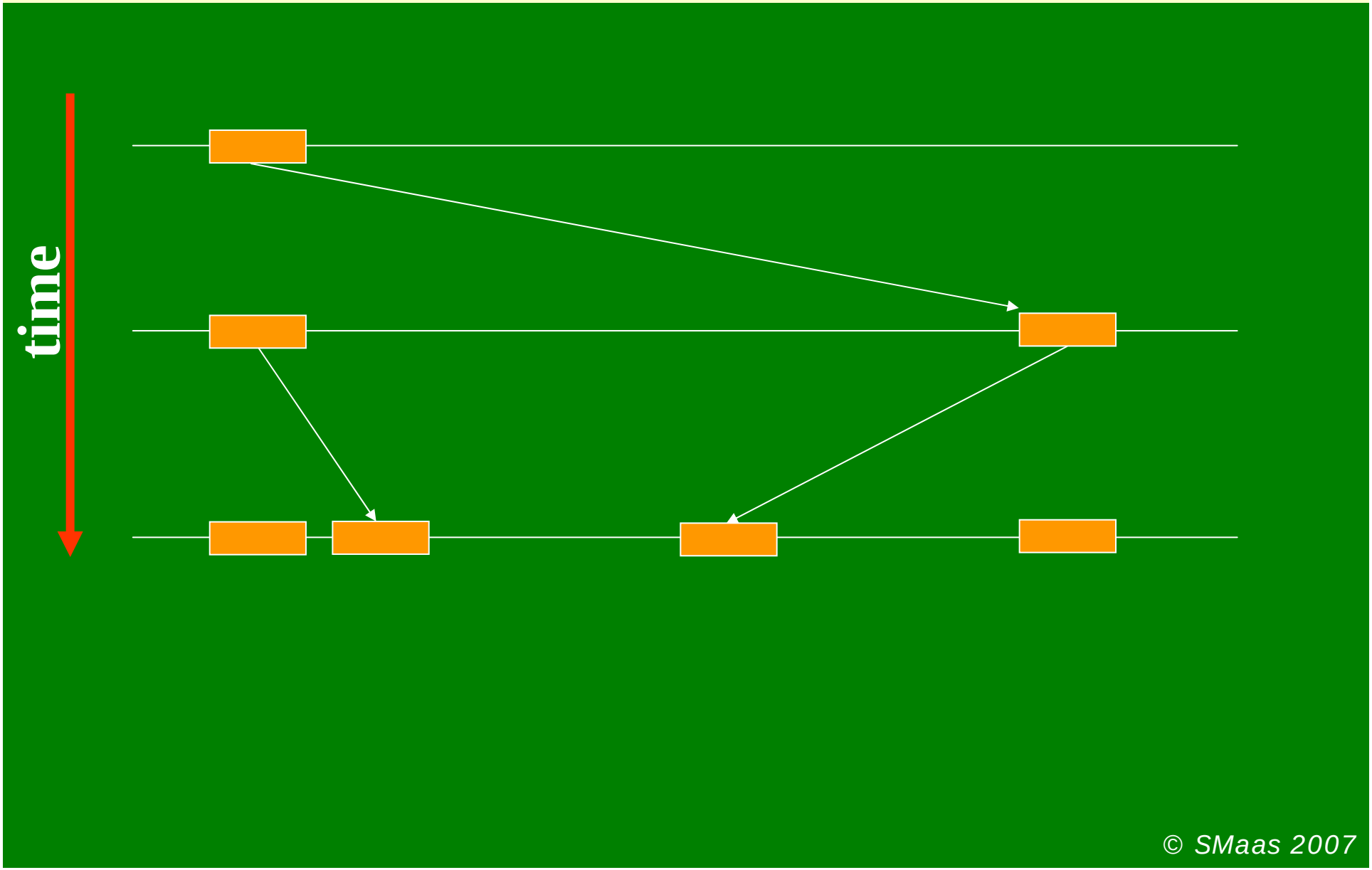

Nonrepetitive DNA is only part of the genome
Genome size
©virtualtext www.ergito.com
10
10
10
10
10
10
10
100%
83%
14%
3%
70%
12%
17%
58%
25%
10%
54%
41%
5%
33%
65%
7%
E.coli
C. elegans
fruit fly
mouse
X. laevis
N. tabacum
Nonrepetitive
Moderately repetitive
Highly repetitive

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



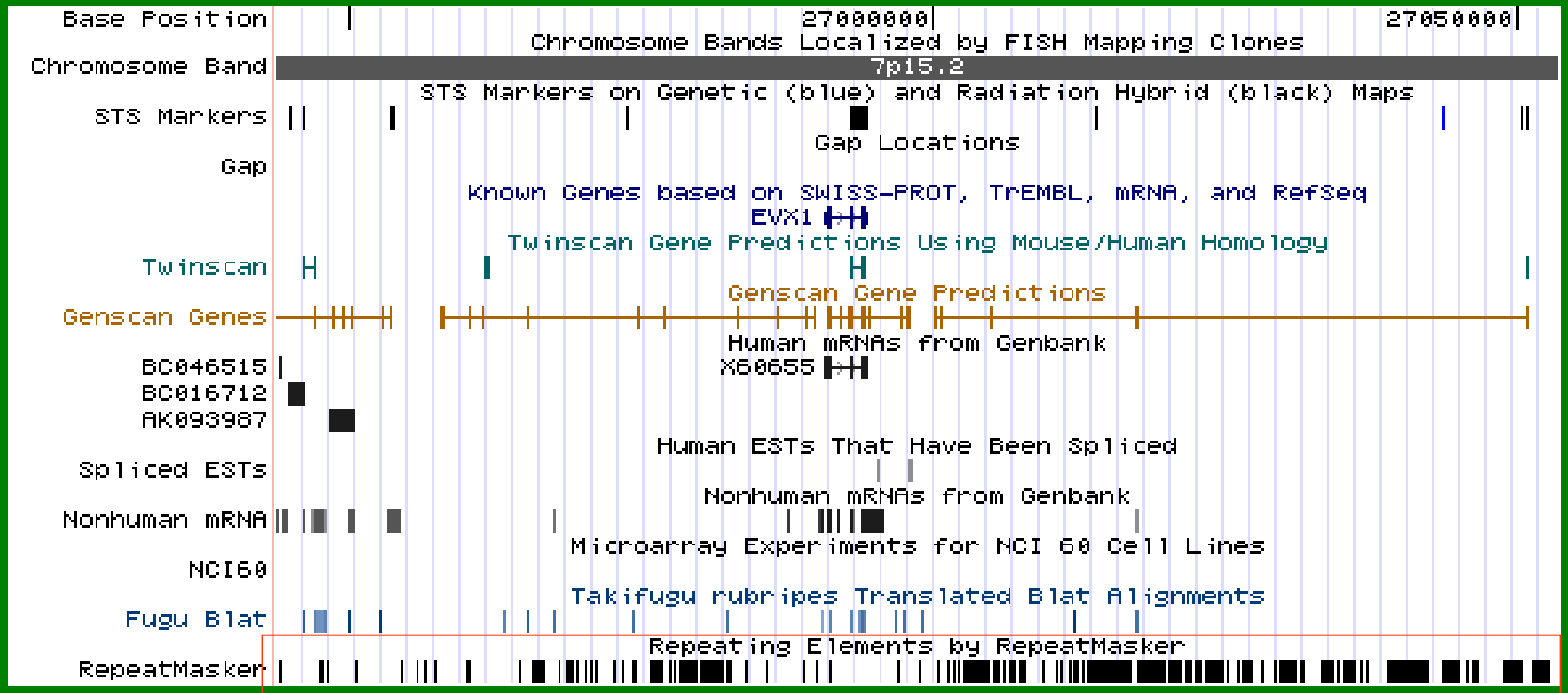


Imagine a gene that could duplicate copies of itself within the genome



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

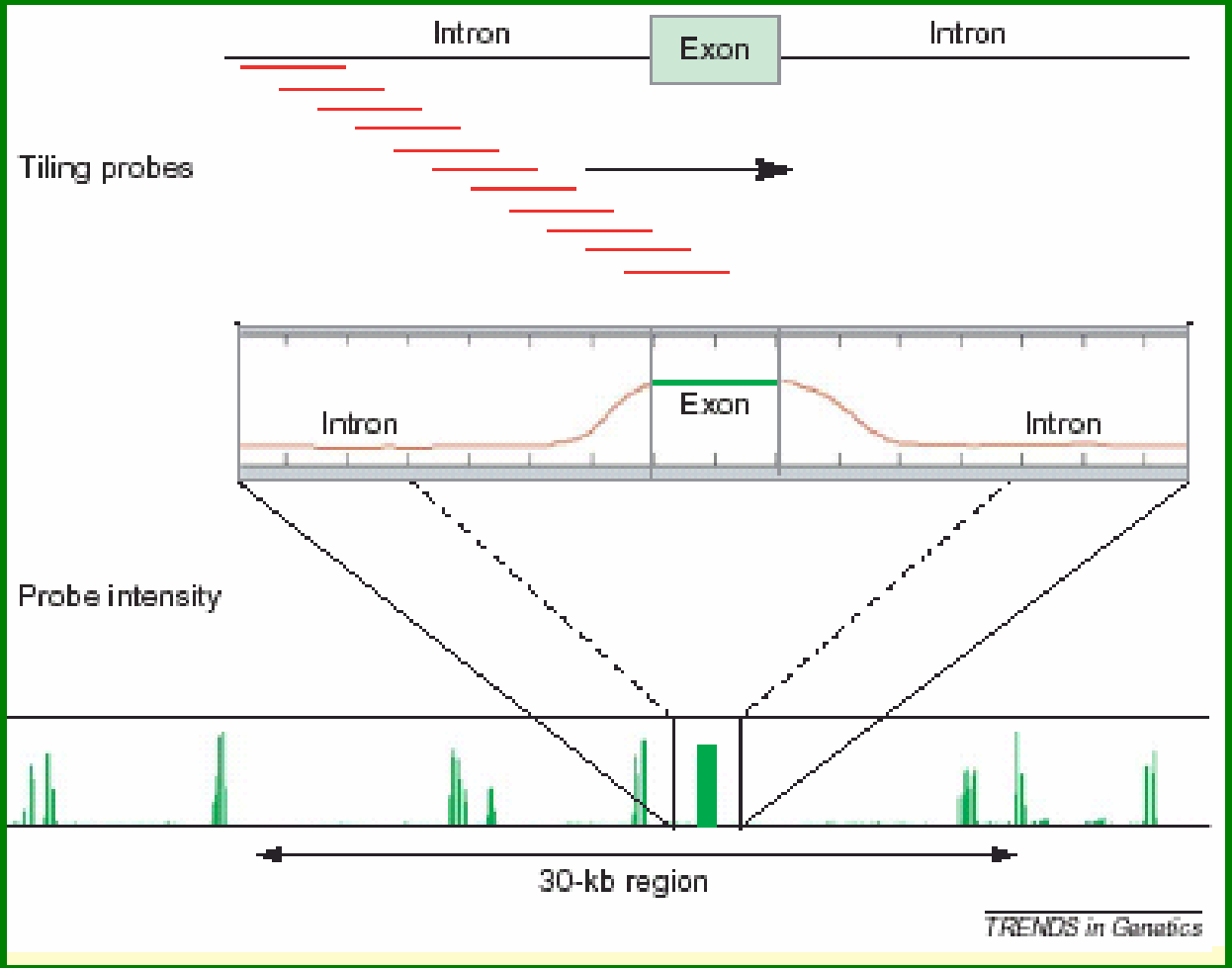
Transposable elements in the human genome

Type	Number of copies	Percentage of total genome
SINEs	1,558,000	13.1
AluI	1,090,000	10.6
LINEs	868,000	20.4
LINE1	516,000	16.9
LTR elements	443,000	8.3
DNA elements	294,000	2.8
mariner	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

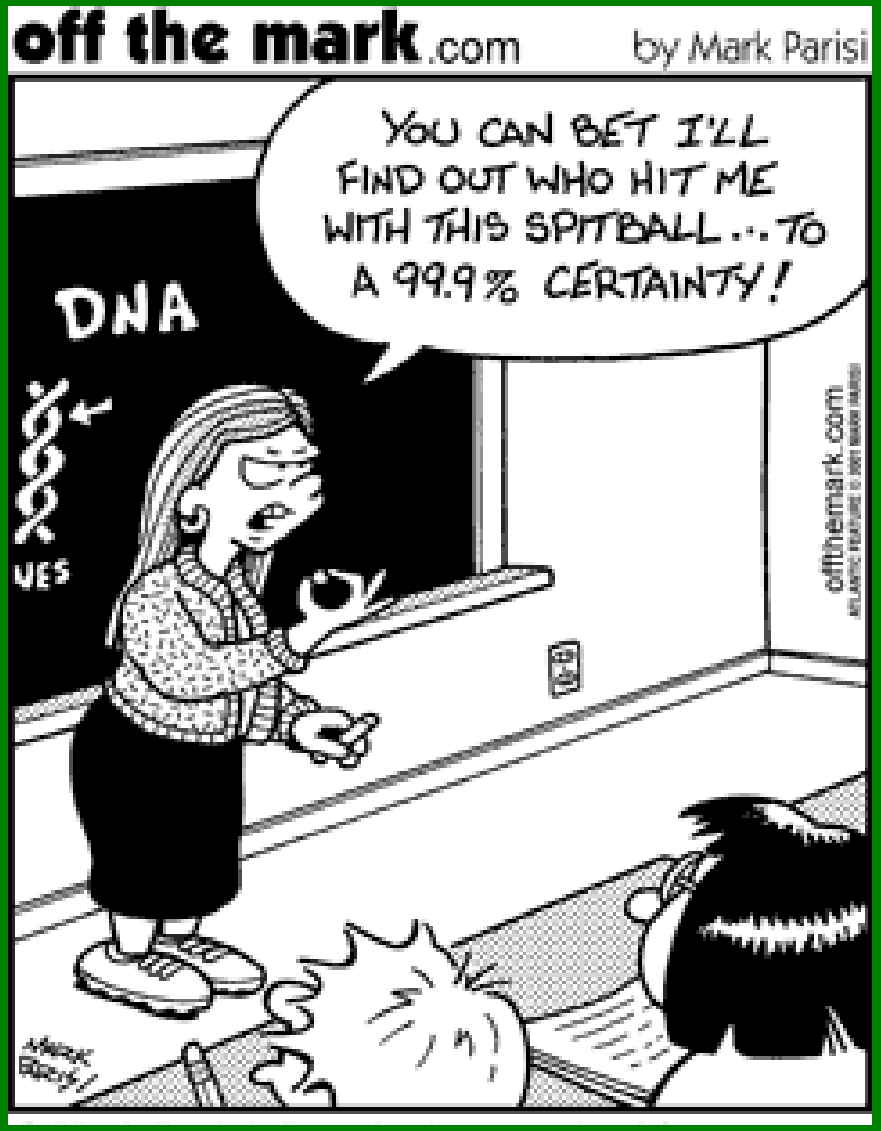
How much of the genomic DNA is converted into RNA sequences?



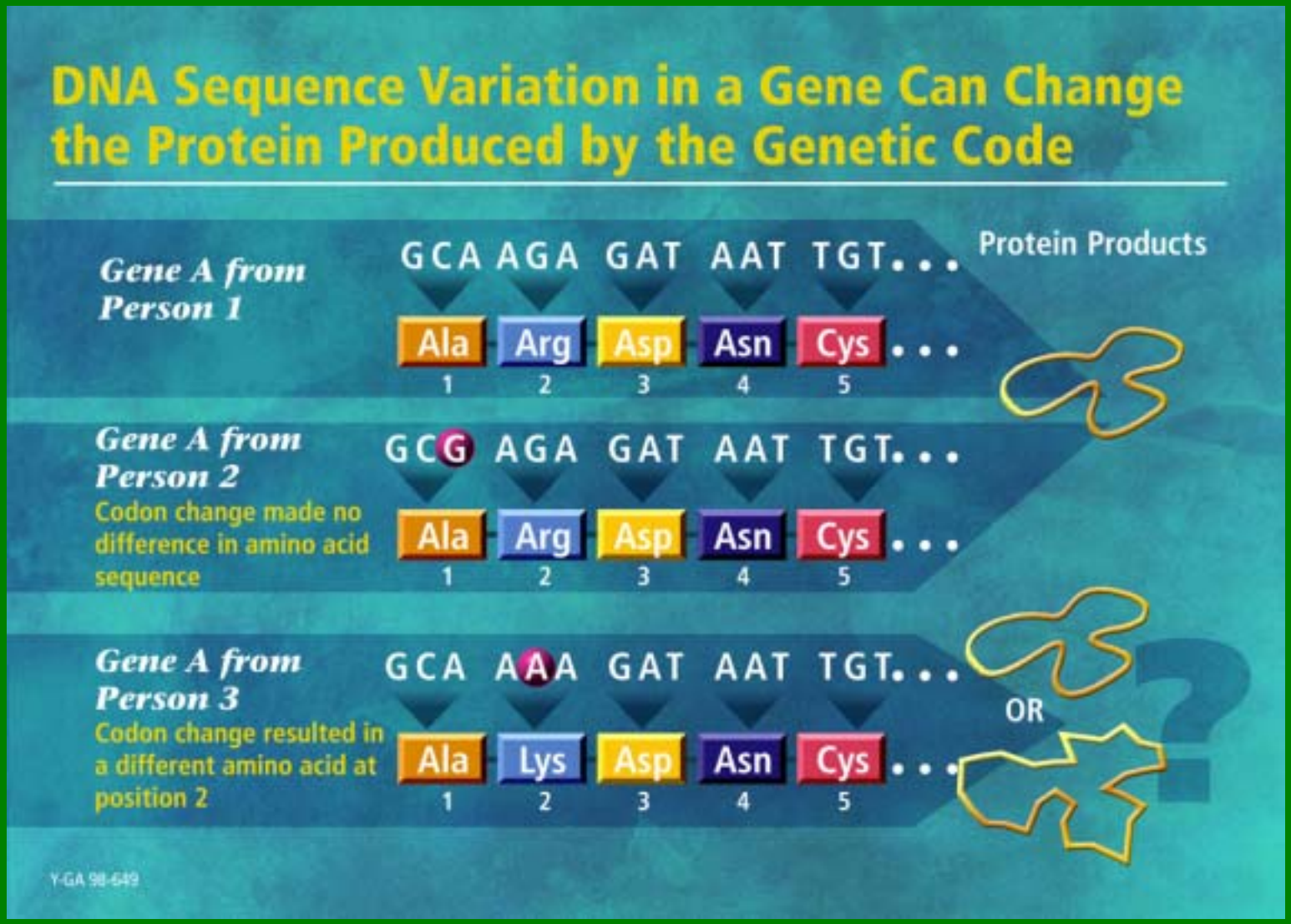
Transcriptome facts

- 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

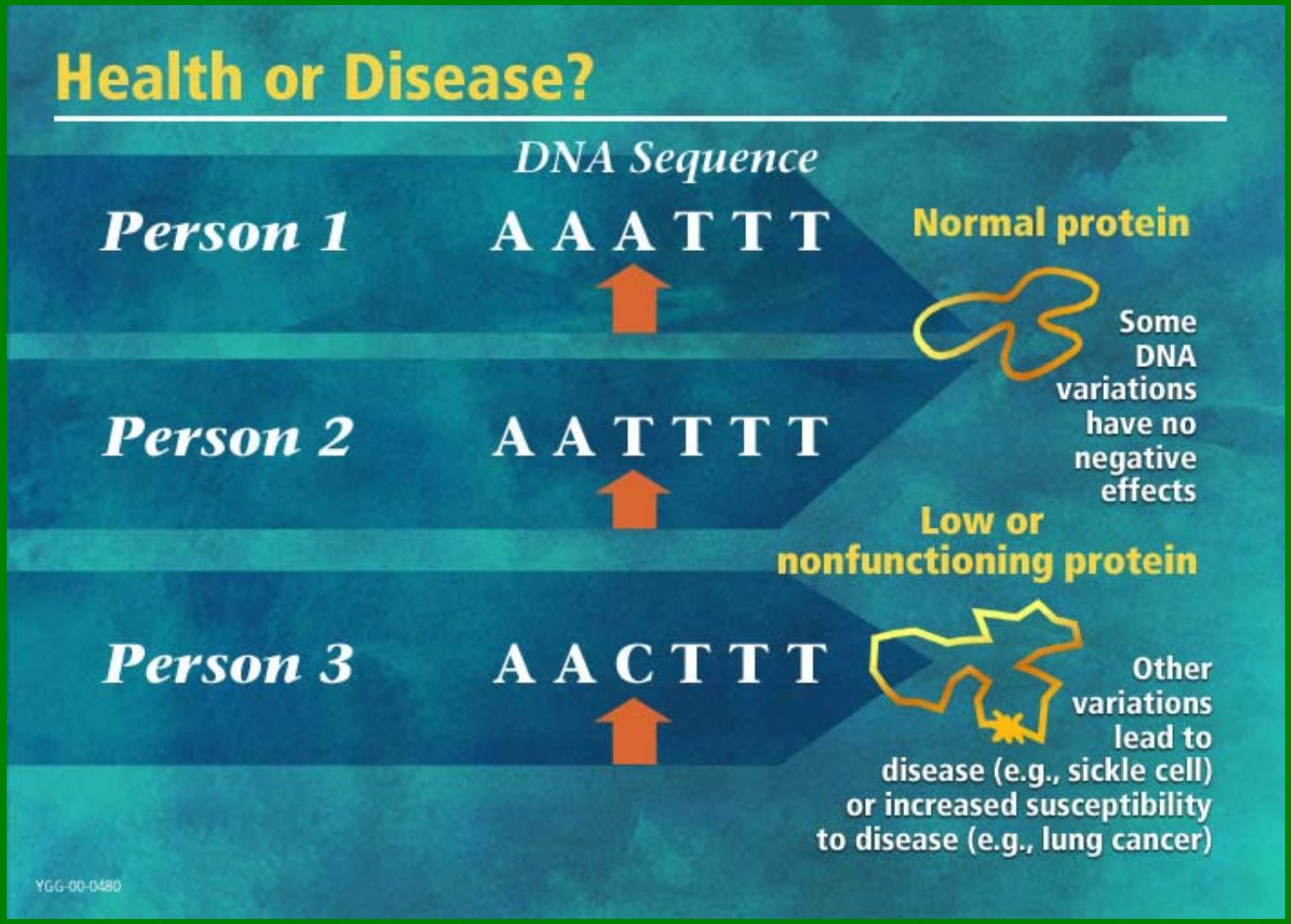
Genotyping



Sequence variations between individuals



Sequence variations between individuals



The Genome Sequence -- an open book ….

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

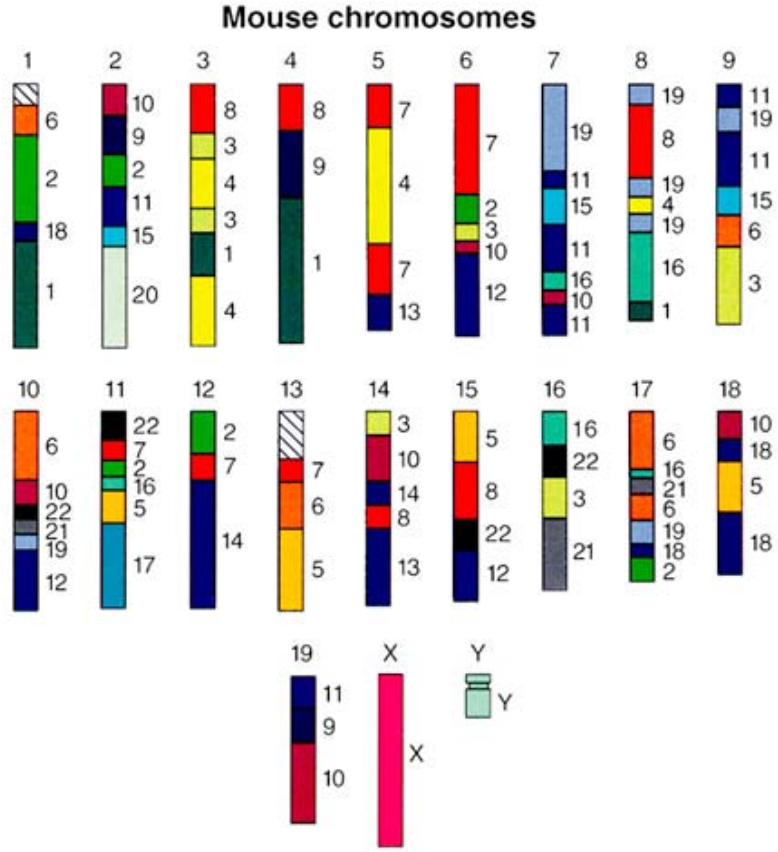
How make sense of whole genome sequences?

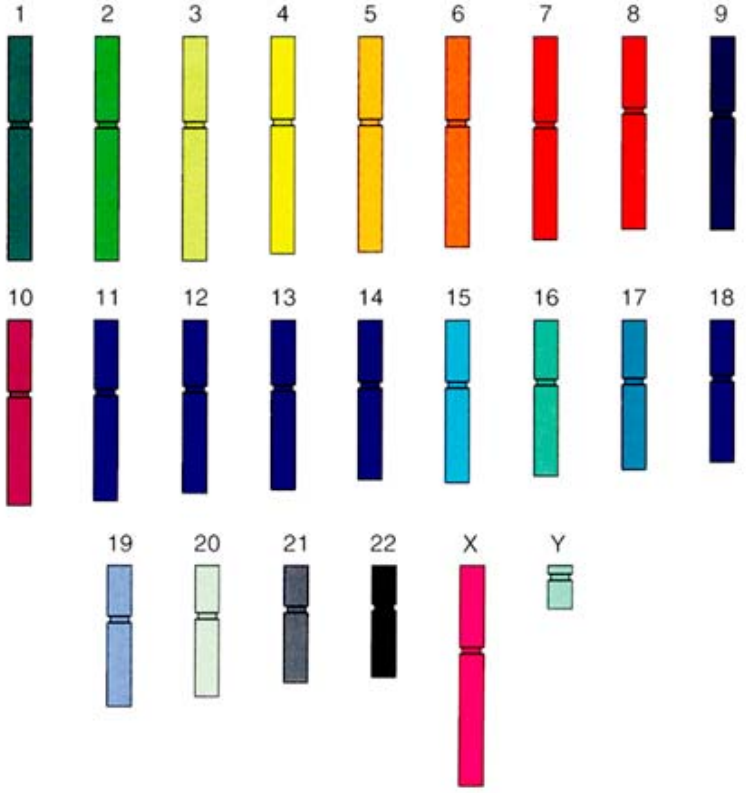
Bioinformatics helps to:

- 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



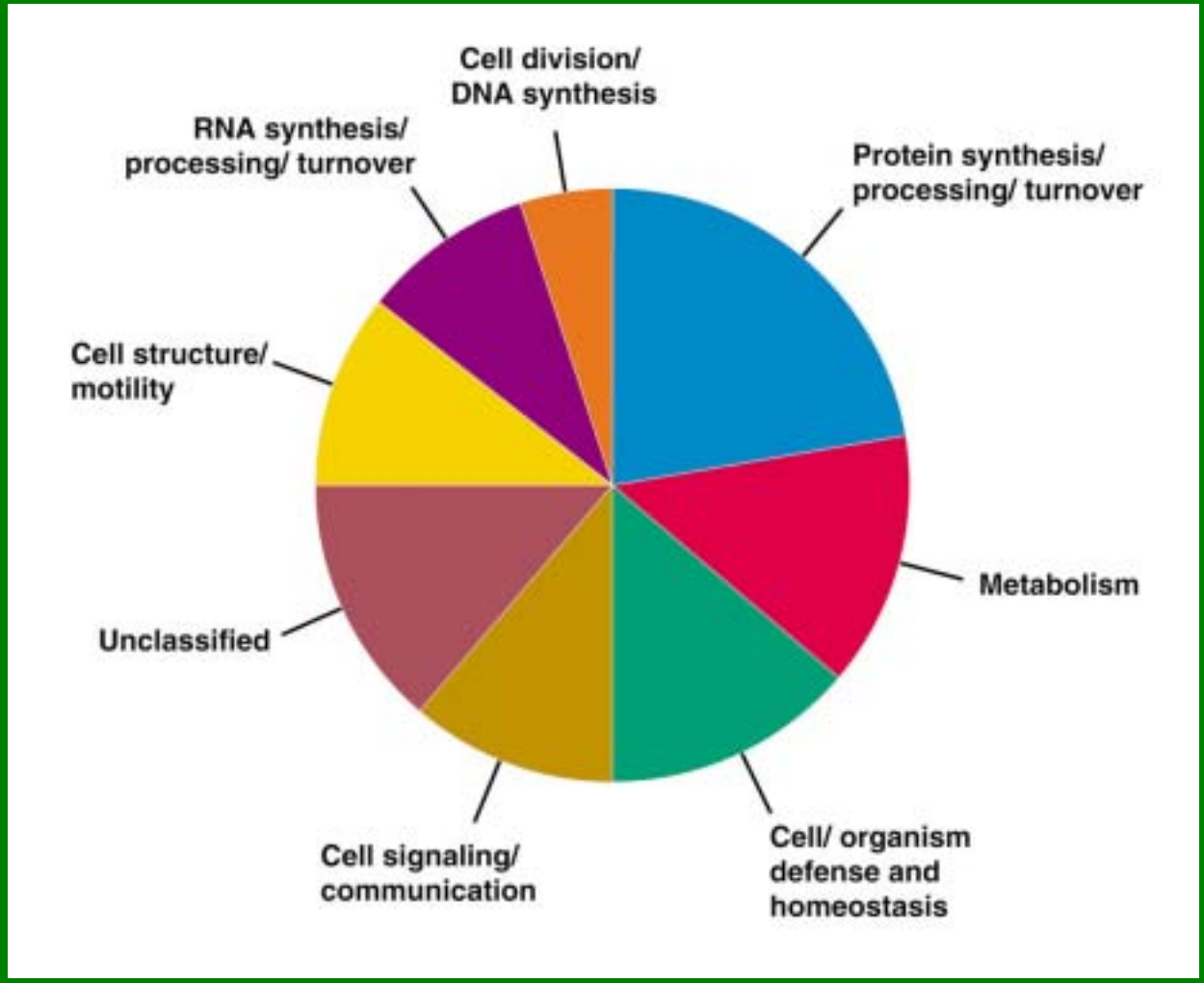


Courtesy Lisa Stubbs
Oak Ridge National Laboratory

YGA 98-075R2

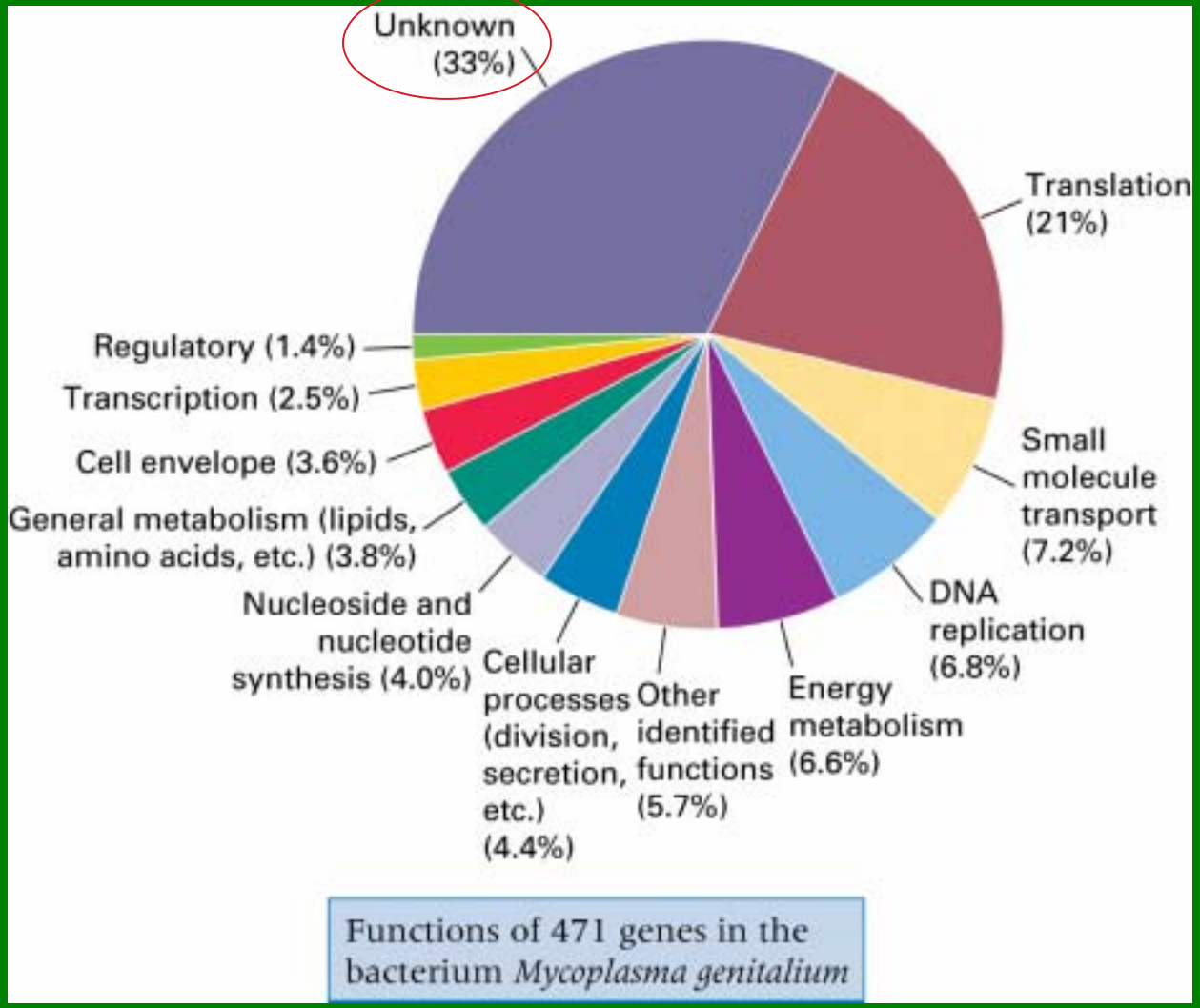
Functional classification of expressed genes

Collection of expressed genes



Randomly selected sequences from human cells grouped by function

Genes in the *Mycoplasma genitalium* classified by function



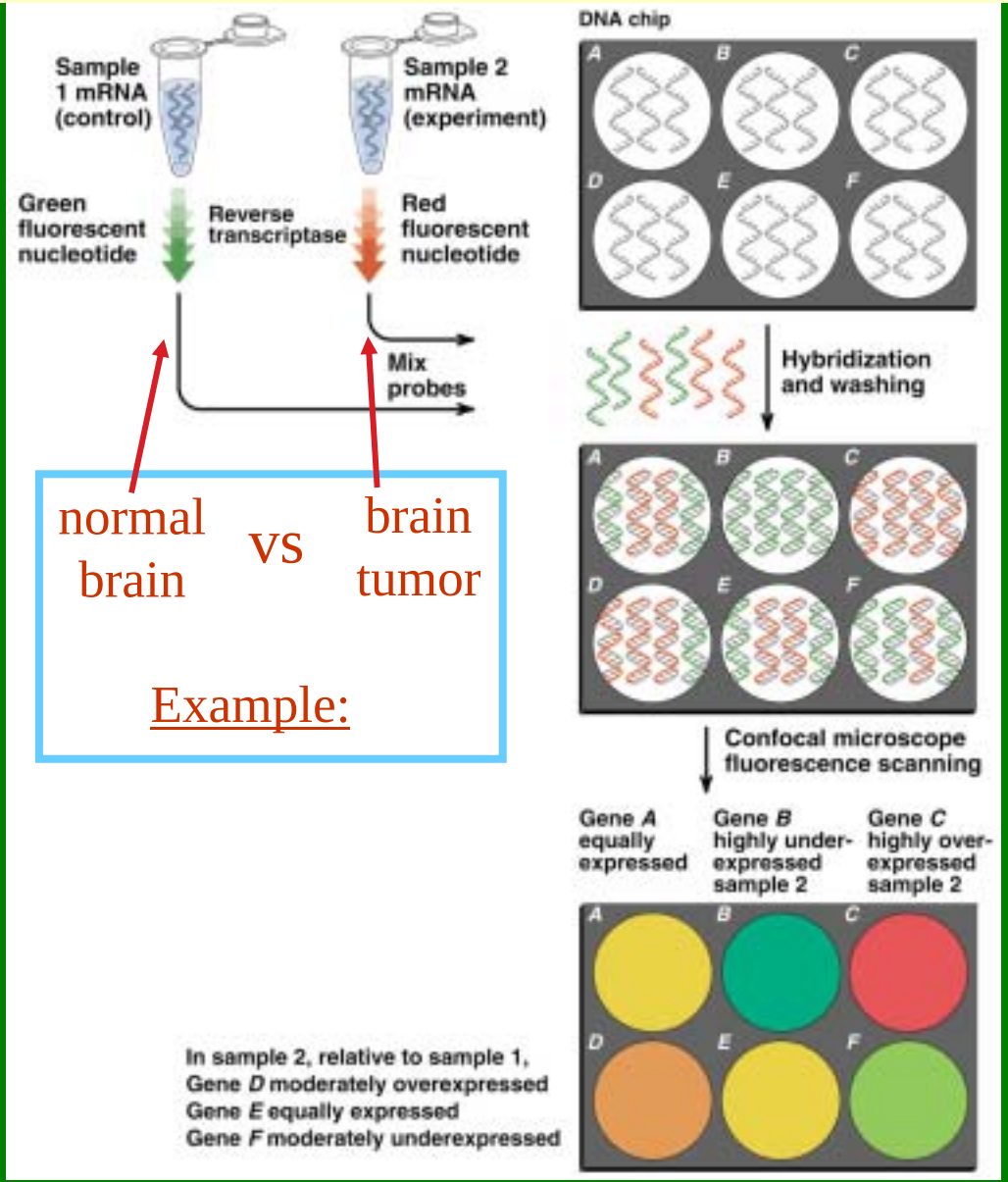
Functions of 471 genes in the bacterium *Mycoplasma genitalium*

Smallest genome of any known free living organism

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

580 kb genome <> 471 genes

Functional genomics I
Use of DNA microarrays (chips)

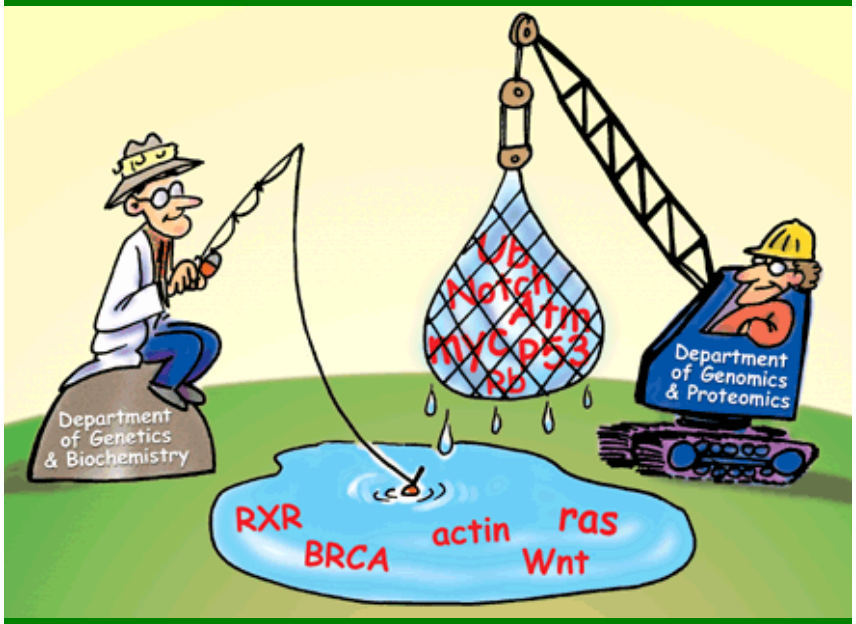


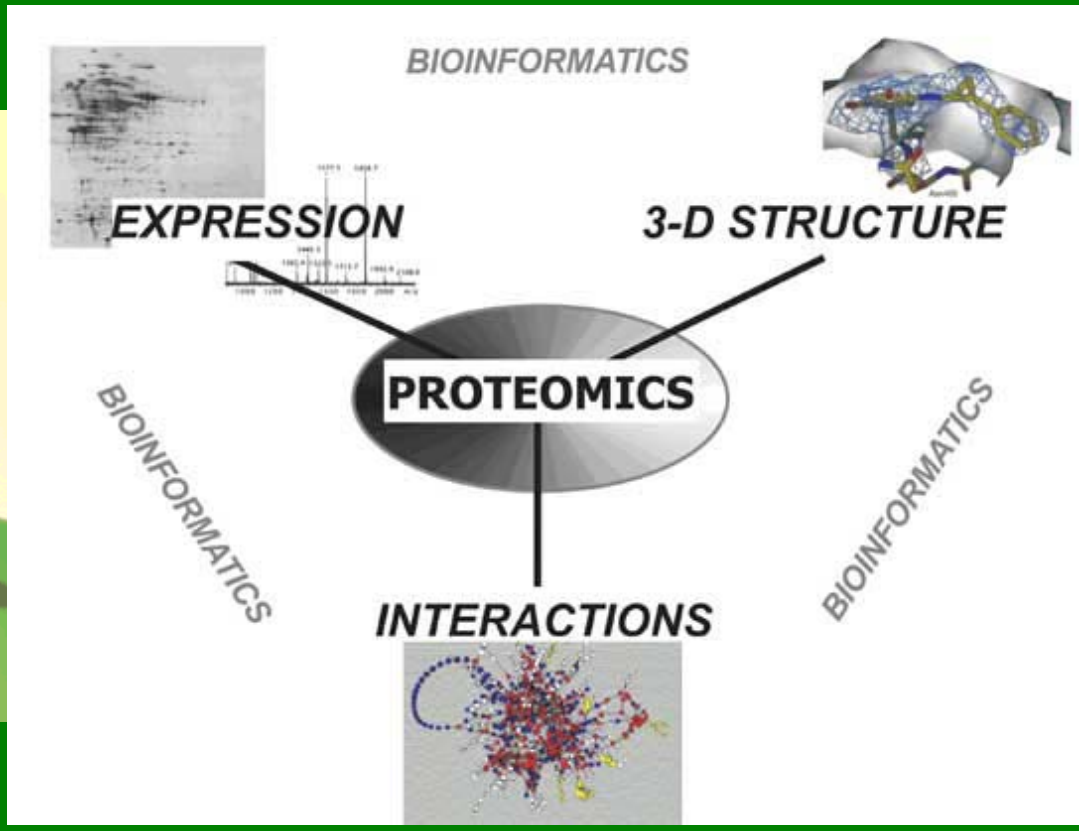
normal brain vs brain tumor

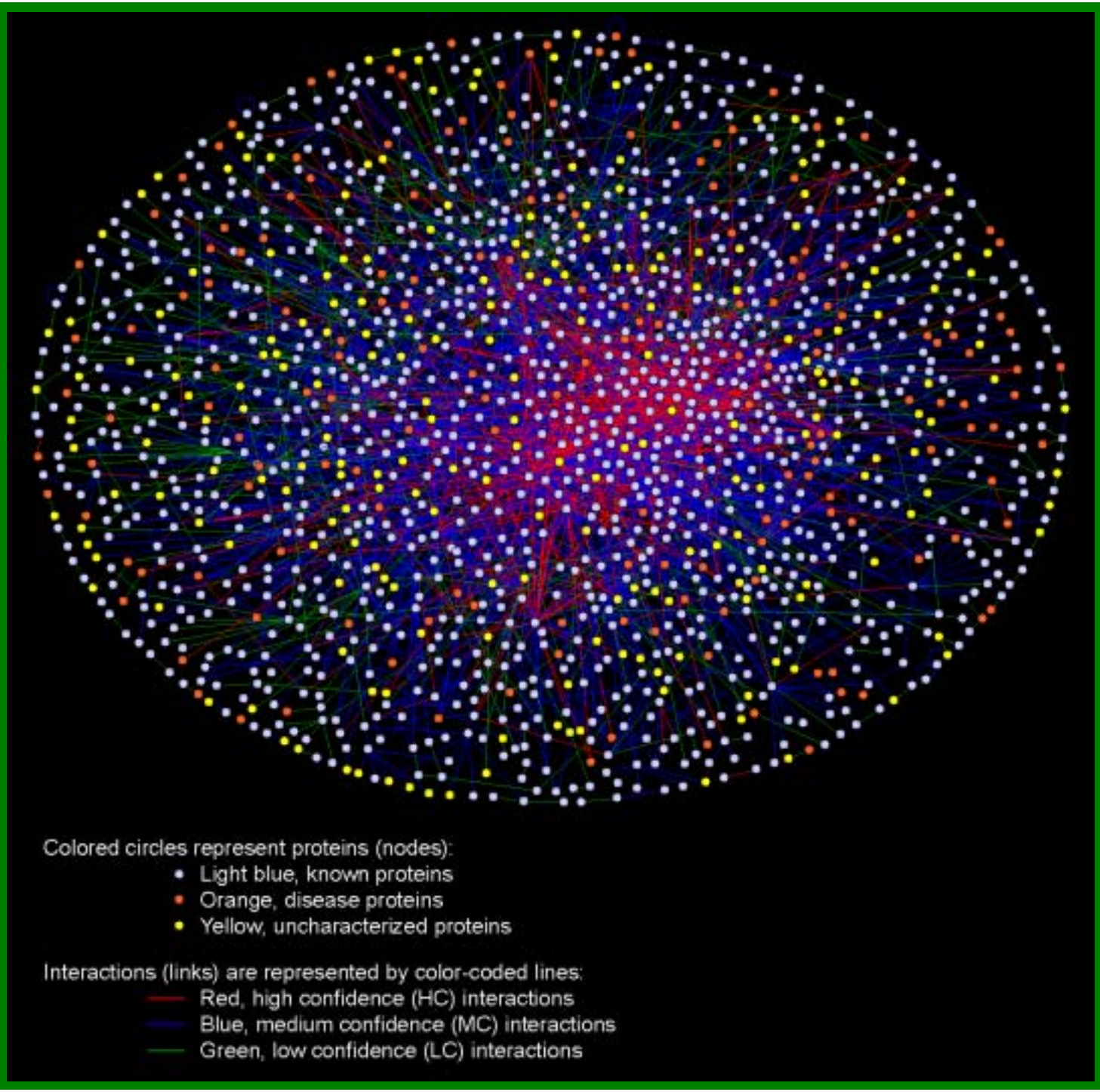
Example:

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

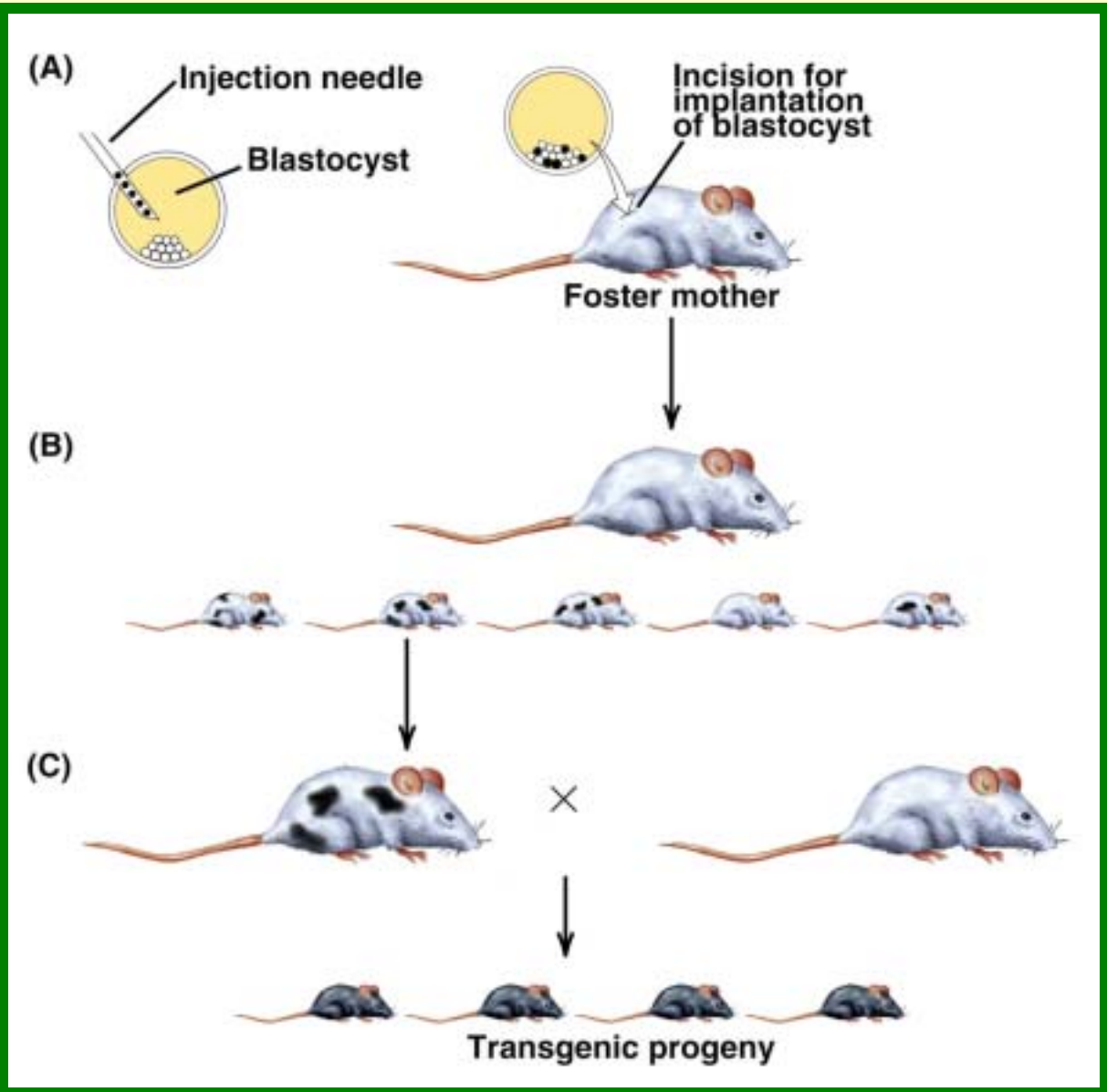
What is Proteomics

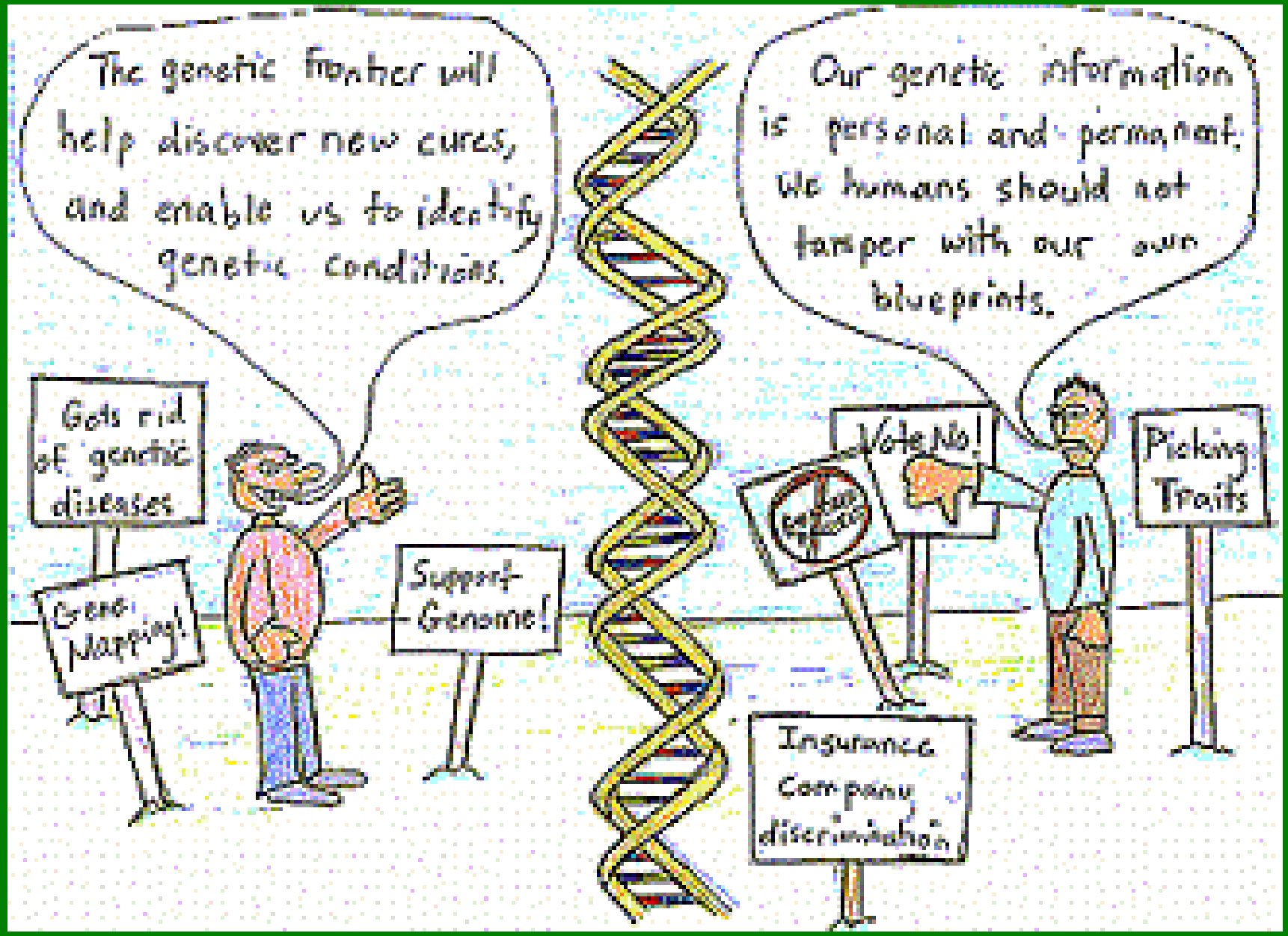





Colored circles represent proteins (nodes):
Light blue, known proteins
Orange, disease proteins
Yellow, uncharacterized proteins
Interactions (links) are represented by color-coded lines:
Red, high confidence (HC) interactions
Blue, medium confidence (MC) interactions
Green, low confidence (LC) interactions

Steps in the creation of a transgenic mouse




The genetic frontier will help discover new cures, and enable us to identify genetic conditions.
Our genetic information is personal and permanent. We humans should not tamper with our own blueprints.
Gets rid of genetic diseases
Gene Mapping!
Support Genome!
Vote No!
Picking Traits
Insurance Company discrimination

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)