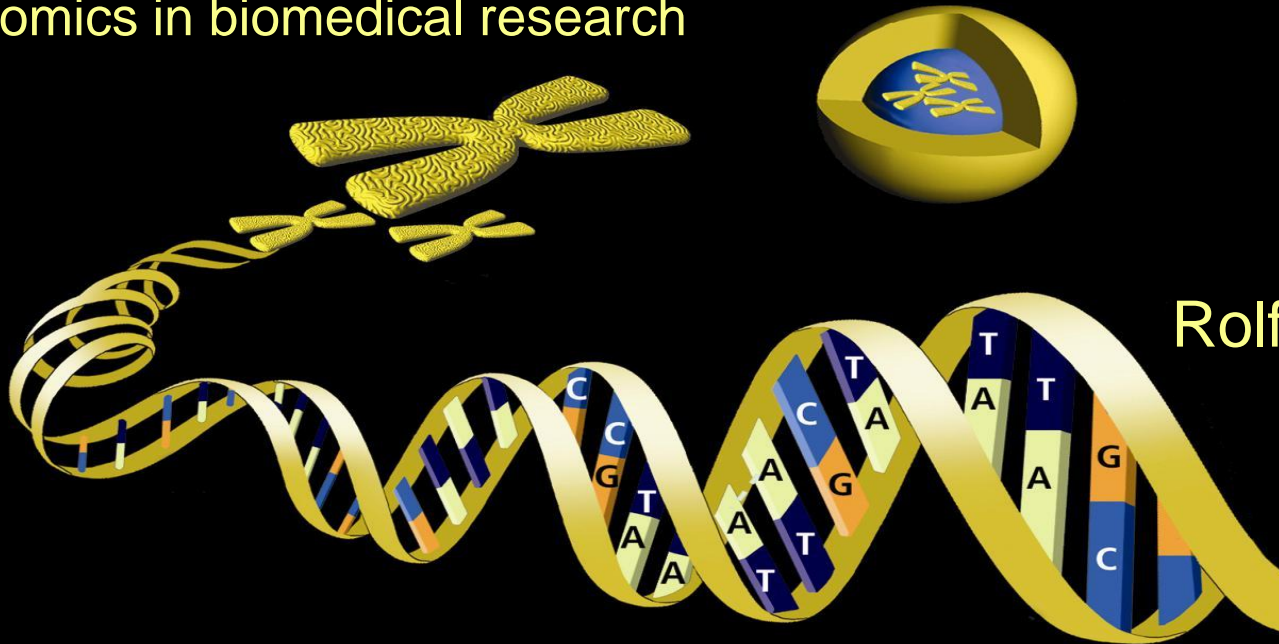


Fundamentals of molecular biology

IN-BIOS 5000/9000

1. A guided tour of the (human) genome
2. From DNA to biological function
3. Genomics in biomedical research



Rolf I. Skotheim
28.10.2024

A guided tour of the (human) genome

Basic biology incl brief history of genetics
and genome sequencing

Early days of genetics (unaware of DNA)

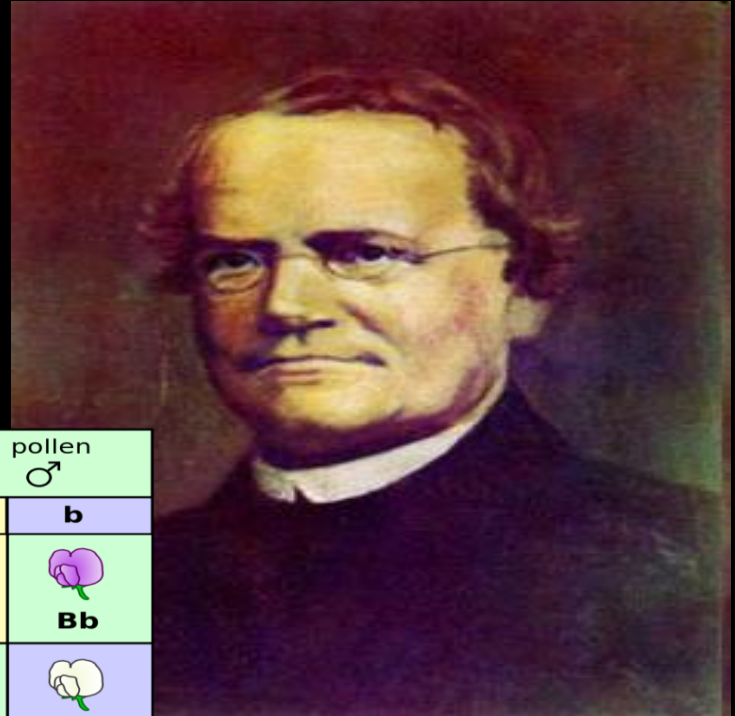
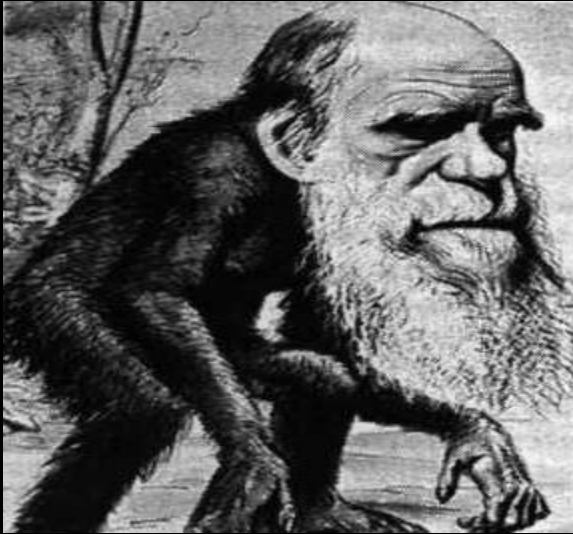
Breeding and selection









Early days of genetics (unaware of DNA)

Gregor Mendel

Charles Darwin

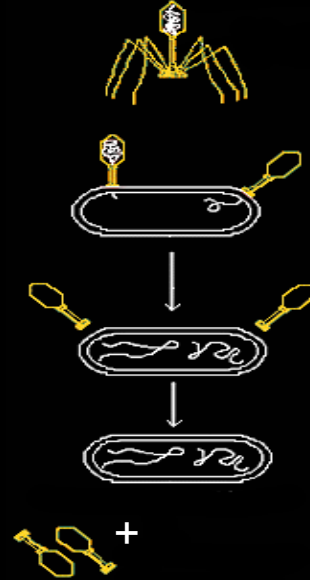


		 pollen ♂	
		B	b
 pistil ♀	B	 BB	 Bb
	b	 Bb	 bb

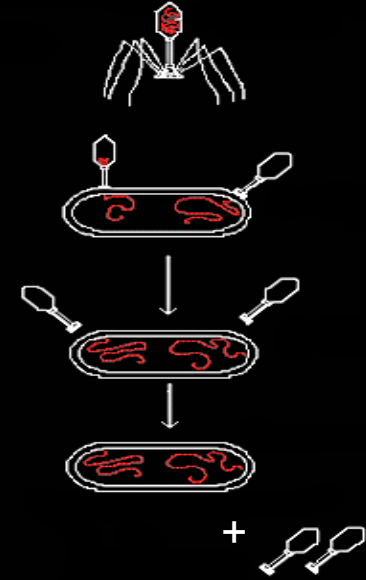
Early days of genetics (unaware of DNA)

Proof that genes are made of DNA

radioactive sulphur-
labelled protein capsule



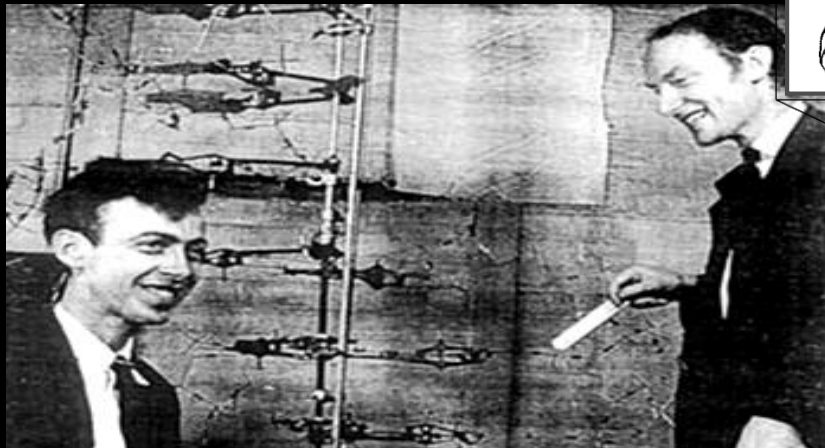
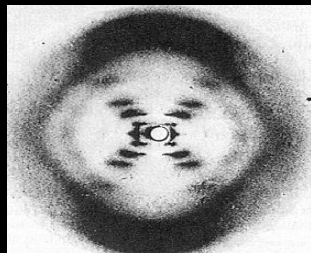
radioactive phosphorus-
labelled DNA core



Hershey & Chase, 1952

1953: The DNA double helix

- Double helix
- Bidirectional



No. 4356 April 25, 1953 NATURE 737

equipment, and to Dr. G. E. R. Deacon and the captain and officers of R.R.S. *Discovery II* for their part in making the observations.

*Young, F. B., Gerard, H., and Jevons, W., *Phil. Mag.*, **46**, 149 (1900).

*Lorentz, H., M. S., *Mon. Not. Roy. Astr. Soc., Geophys. Supp.*, **8**, 266 (1910).

*Van Arman, S., *Wegs Hole Papers in Phys. Oceanogr. Meteor.*, **21** (1910).

*James, V. W., *Arkiv. Mat. Astron. Fysik.* (Stockholm), **2** (11) (1905).

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

We wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey*. They kindly made their manuscript available to us in advance of publication. Their model consists of three intertwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Frazer† (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure is also unsatisfactory, and for this reason we shall not comment on it.

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate di-ester groups joining 5'-deoxy-ribose residues with 3',5' linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furburgh's model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furburgh's 'standard configuration', the sugar being roughly perpendicular to the attached base. There is a residue on each chain every 3.4 Å. in the z-direction. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, the distance of a phosphorus atom from the fibre axis is 10 Å. As the phosphates are on the outside, cations have easy access to them.

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical z-co-ordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrimidine position 6.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms (that is, with the keto rather than the enol configurations) it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine).

In other words, if an adenine forms one member of a pair, on either chain, then on these assumptions the other member must be thymine; similarly for guanine and cytosine. The sequence of bases on a single chain does not appear to be restricted in any way. However, if only specific pairs of bases can be formed, it follows that if the sequence of bases on one chain is given, then the sequence on the other chain is automatically determined.

It has been found experimentally* that the ratio of the amounts of adenine to thymine, and the ratio of guanine to cytosine, are always very close to unity for deoxyribose nucleic acid.

It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact.

The previously published X-ray data* on deoxyribose nucleic acid are insufficient for a rigorous test of our structure. So far as we can tell, it is roughly compatible with the experimental data, but it must be regarded as unproved until it has been checked against more exact measurements. Some of these are given in the following communications. We were not aware of the details of the results presented there when we devised our structure, which rests mainly though not entirely on published experimental data and stereochemical arguments.

It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

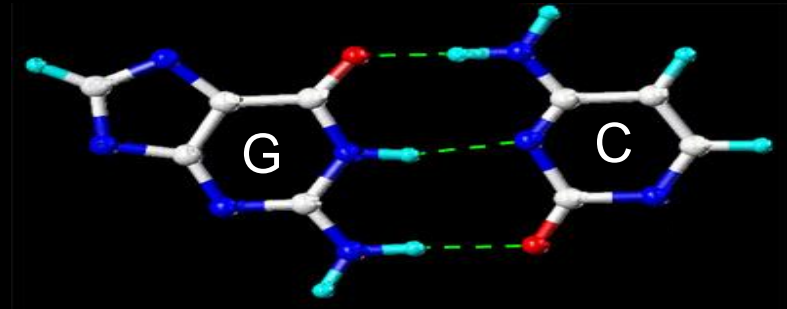
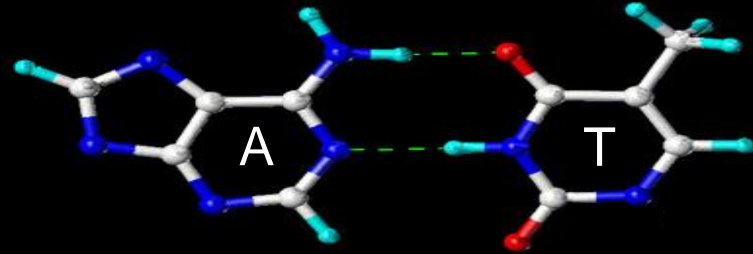
Full details of the structure, including the conditions assumed in building it together with a set of co-ordinates for the atoms, will be published elsewhere.

We are much indebted to Dr. Jerry Donohue for constant advice and criticism, especially on inter-atomic distances. We have also been stimulated by a knowledge of the general nature of the unpublished experimental results and ideas of Dr. M. H. F. Wilkins, Dr. R. E. Franklin and their co-workers at

Watson & Crick
Wilkins, Stokes, & Wilson
Franklin & Gosling

1953: The DNA double helix

- Double helix
- Bidirectional
- Base-specific pairing



Watson & Crick
Wilkins, Stokes, & Wilson
Franklin & Gosling

The central dogma

DNA



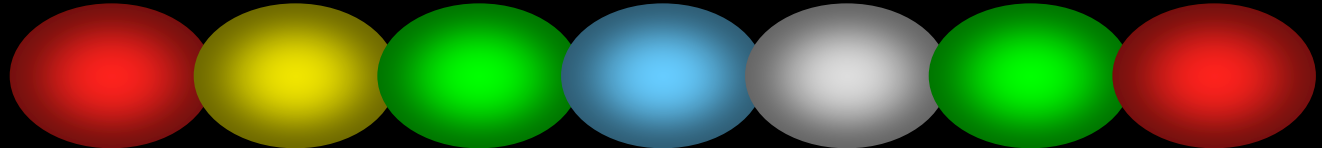
RNA



Protein

ACGTCCATGCAGGATATGACG

ACGUCC AUG CAG GAU AUG ACG



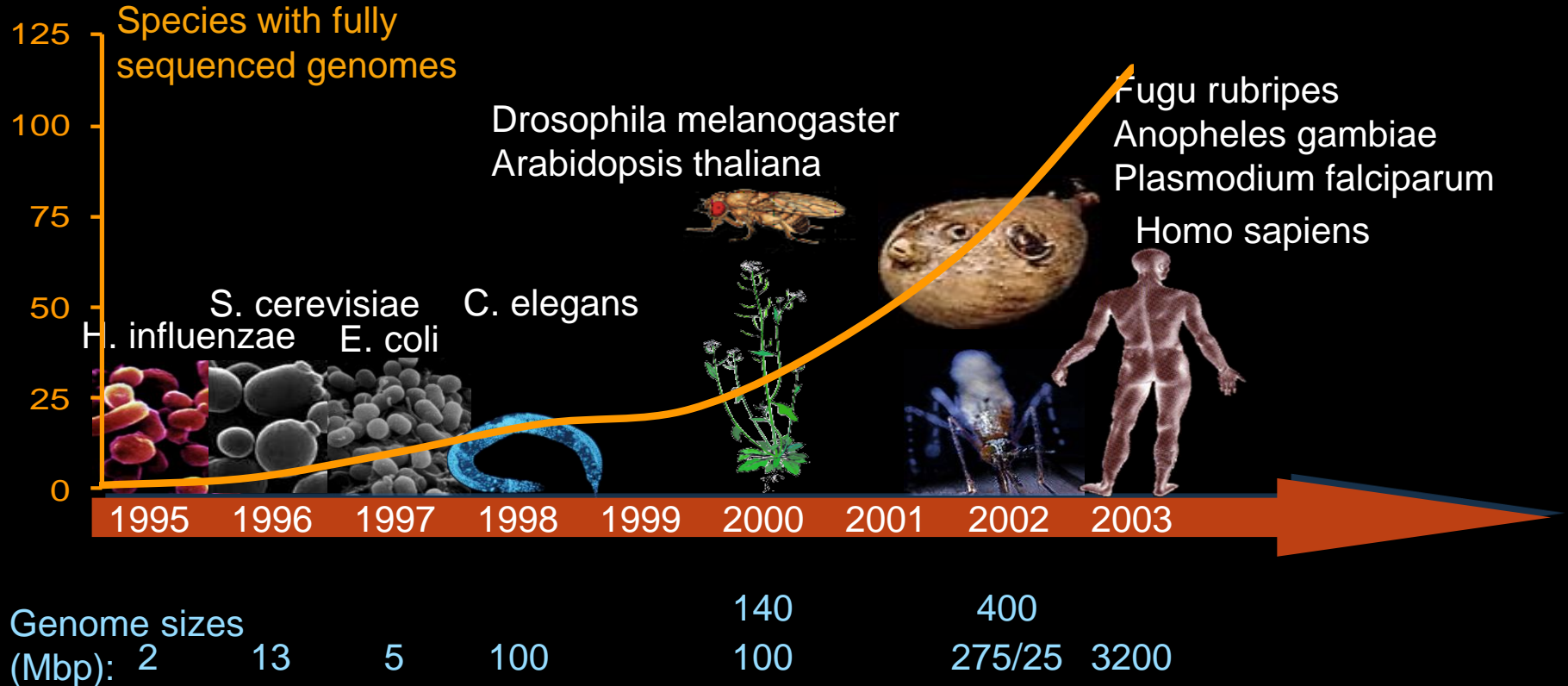
The genetic code

		2					
		U	C	A	G		
1	U	UUU } Phe	UCU } Ser	UAU } Tyr	UGU } Cys	U C A G	3
		UUC } Phe	UCC } Ser	UAC } Tyr	UGC } Cys		
		UUA } Leu	UCA } Ser	UAA } Stop	UGA } Stop		
		UUG } Leu	UCG } Ser	UAG } Stop	UGG } Trp		
	C	CUU } Leu	CCU } Pro	CAU } His	CGU } Arg	U C A G	
		CUC } Leu	CCC } Pro	CAC } His	CGC } Arg		
		CUA } Leu	CCA } Pro	CAA } Gln	CGA } Arg		
		CUG } Leu	CCG } Pro	CAG } Gln	CGG } Arg		
	A	AUU } Ile	ACU } Thr	AAU } Asn	AGU } Ser	U C A G	
		AUC } Ile	ACC } Thr	AAC } Asn	AGC } Ser		
		AUA } Met	ACA } Thr	AAA } Lys	AGA } Arg		
		AUG } Met	ACG } Thr	AAG } Lys	AGG } Arg		
	G	GUU } Val	GCU } Ala	GAU } Asp	GGU } Gly	U C A G	
		GUC } Val	GCC } Ala	GAC } Asp	GGC } Gly		
		GUA } Val	GCA } Ala	GAA } Glu	GGA } Gly		
		GUG } Val	GCG } Ala	GAG } Glu	GGG } Gly		

A guided tour of the human genome

Basic biology incl brief history of
genetics and genome sequencing

Some early sequenced genomes



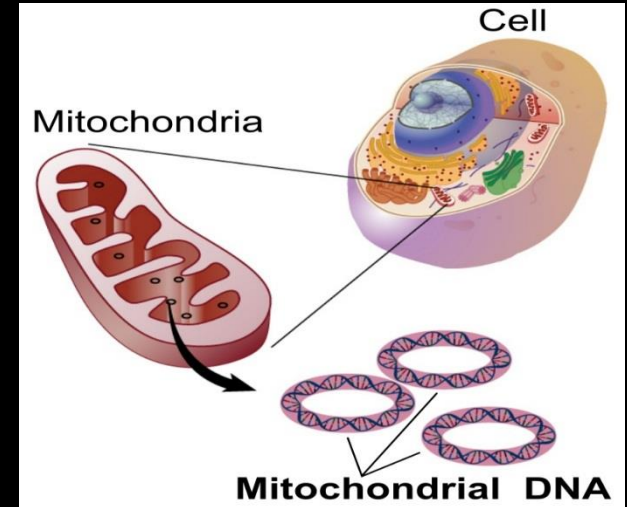
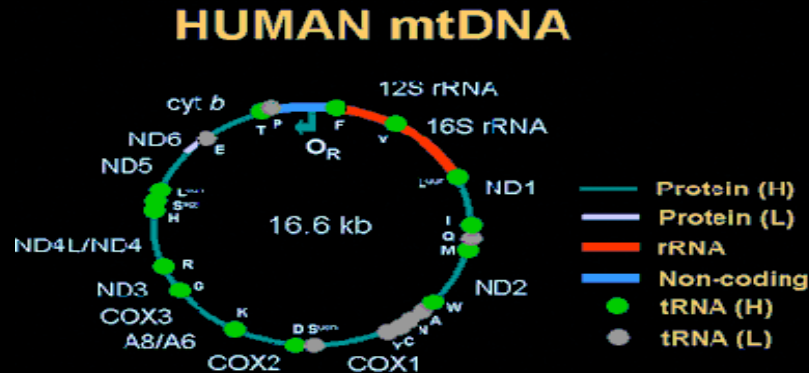
The human genome in numbers

- ~6.4 billion basepairs on 46 linear DNA molecules
 - (2 x 23 chromosomes [22 auto-chromosomes and X, Y sex chromosomes])



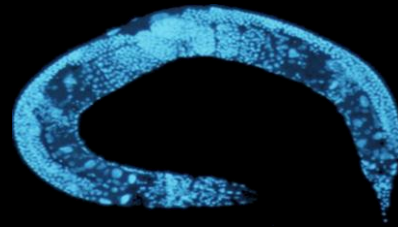
The human genome in numbers

- ~6.4 billion basepairs on 46 linear DNA molecules
- Mitochondrion: circular DNA molecule of 16 569 bp



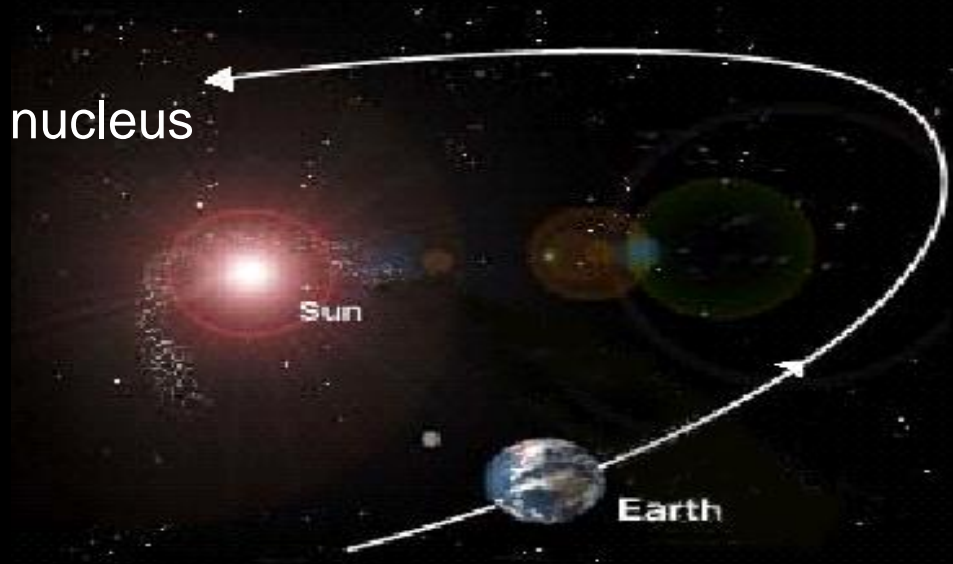
The human genome in numbers

- ~6.4 billion basepairs on 46 linear DNA molecules
- Mitochondrion: circular DNA molecule of 16 569 bp
- 20 000 protein coding genes



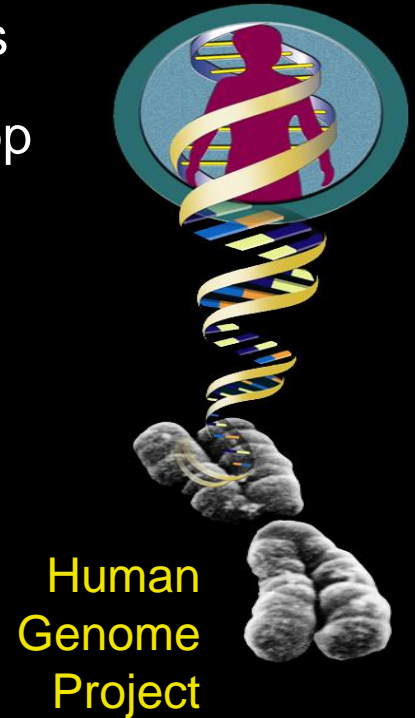
The human genome in numbers

- ~6.4 billion basepairs on 46 linear DNA molecules
- Mitochondrion: circular DNA molecule of 16 569 bp
- 20 000 protein coding genes
- Two meters DNA in each cell nucleus

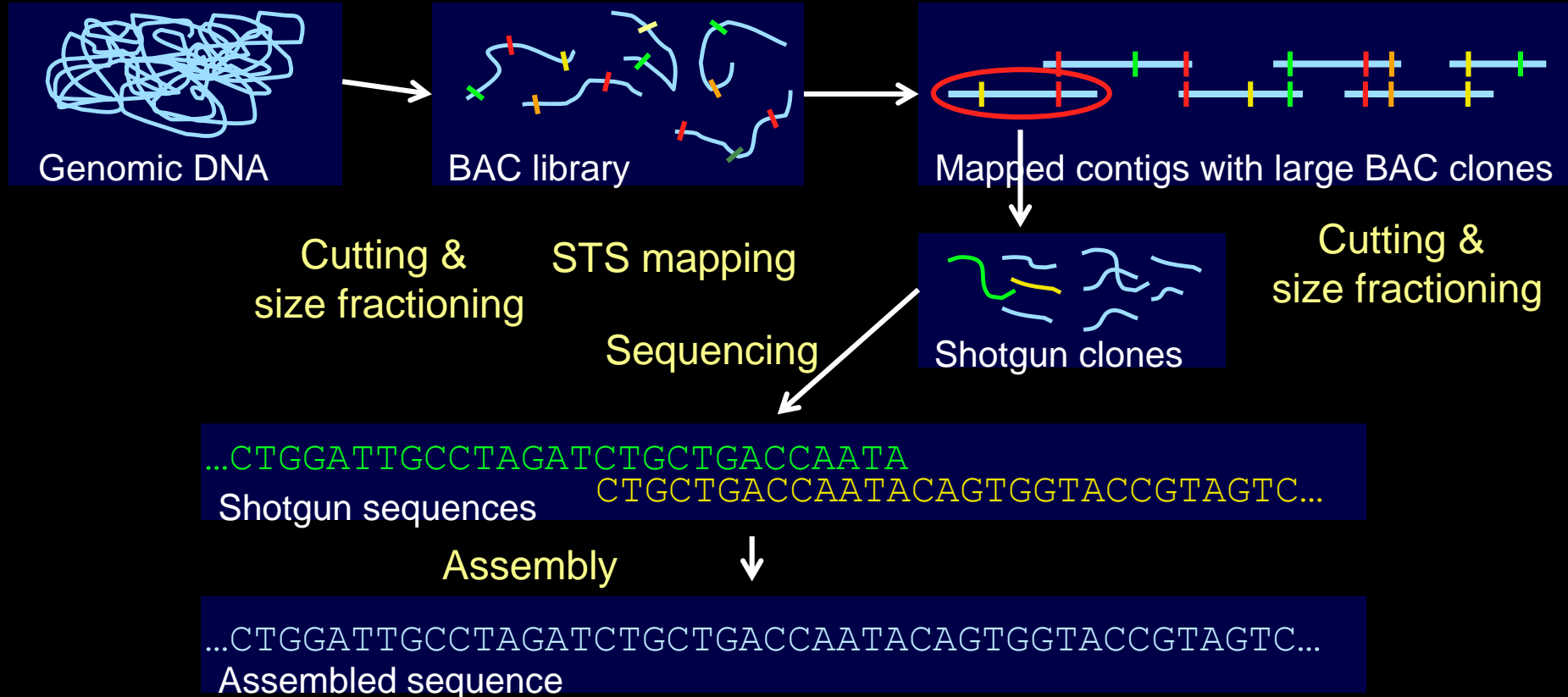


The human genome in numbers

- ~6.4 billion basepairs on 46 linear DNA molecules
- Mitochondrion: circular DNA molecule of 16 569 bp
- 20 000 protein coding genes
- Two meters DNA in each cell nucleus
- 3 billion US \$



How to sequence a genome (historic)



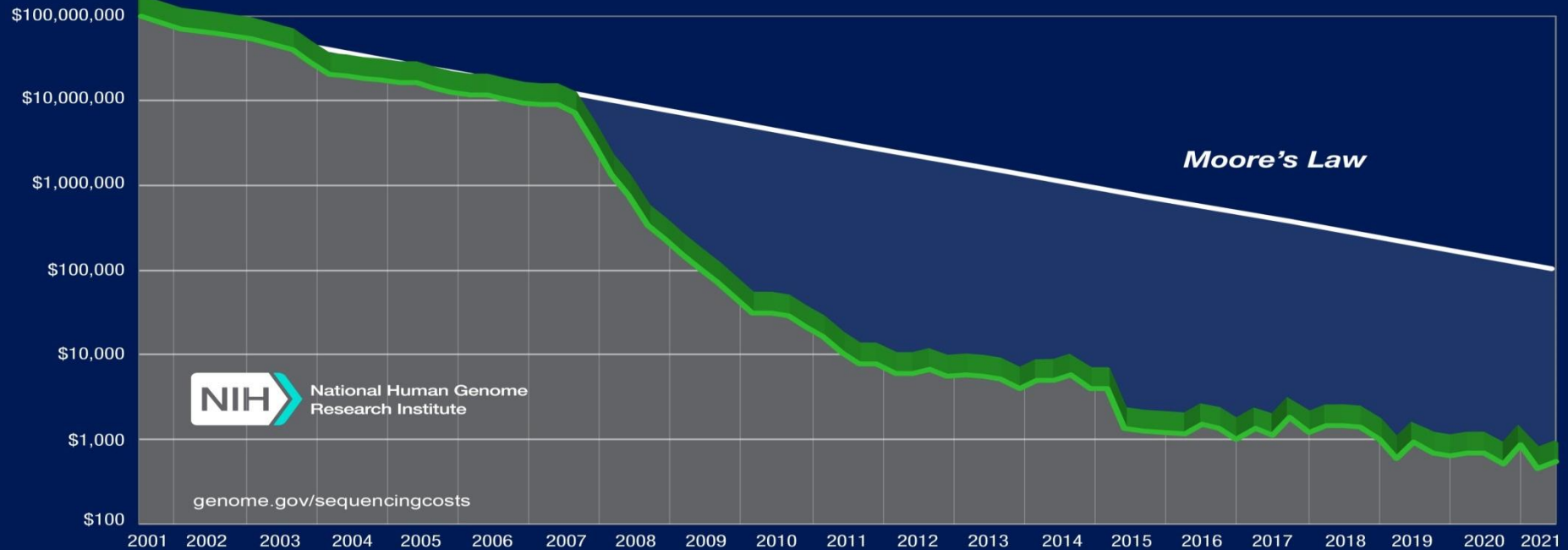
Genome sequencing, then and now

Year 2000, announcement of the first human genome sequence



Genome sequencing, then and now

Cost per human genome



Main current («Next-generation») genome sequencing (NGS) technologies

- Short-read sequencing (Illumina)
- Long-read sequencing (Pacific Biosciences, Oxford Nanopore Technologies)

Sequence annotation



AGGCGTCGAACGTTGCACCACGCTTCAACGAATAGGCGTCGAACGTTGCACCACGCGTTACGAATACGC
GCTACGTCAACGACGACGATACGCGCGCGTCGCGACGACGTCGTGCGACGACGCTACGTGCGAAATACGC
GCGCGTCGCGAACGTACGTGCGGACCACGCTTCAAATAGGCGTCGAACGTACGTGCACCACGCTACGTC
AAATATATAAGGCGTCGACGTTGCACCACGCTTCTCAAGCGCTACCAATAGGCGTCGAACGTTGCACCACG
CTTCAAATATGCGTCGAACGTTGCACCACGCTGAGGTAAGTCGAATAGGCGTCGAACGTTGCACCACGCT
ACGTCAATAGGCGTCGAACGTTGCACCACGCTTCAAATAGGCGTCGAACGTACGTGCACCACGCTTCAA
TAGGCGTCGAACGTTGCACCACGCTTCAAATAGGCGTCGAACGTTGCACCACGCTTCAAAT
AGGCGTCGAACGTTGCACCACGCTTCAACGAATAGGCGTCGAACGTTGCACACGCTTCAAATAGGCGTCG
AACGTACGTGCACCACGCTTCAAGGTAAGTAATAGGCGTCGAACGTTGCACCACGCTACGTCAAATAGGC
GTCGAACGTACGTGCACCACGCTACGTCAACGAATAGGCGTCGAACGTTGCACCACGCTACGTCAATAGG
CGTCGAACGTTGCACCACGCTTCAACGAACGAATAGGCACGGTCGAACGTACGTGCACCACGCTTCAAATA
GGCGTCGAACGTTGCACCACGCTTCAAATAGGCGTCGAACGTACGTGCACCACGCTACGTCAAATCCTTC
ACAGTAGGCGTCGAACGTTGCACCACGCTTCAAATAGGCGTCGAACGTTGCACACGCTTCAAATAAGGCG
TCGAACGTTGCACCACGAGGTAAGTCTACGTCAAACGATAGGCGTCGAACGTACGTGCACCACG

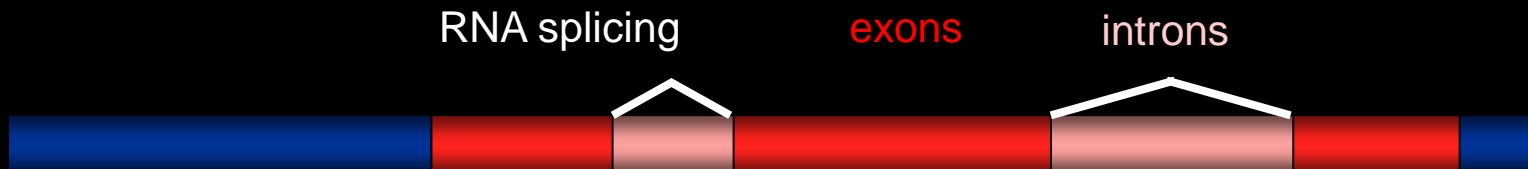
Sequence annotation

The gene (pre-mRNA)



AGGCGTCGAACGTTGCACCACGCTTCAACGAATAGGCGTCGAACGTTGCACCACGCGTTCACGAATACGC
GCTACGTCAACGACGACGATACGCGCGCGTTCGCGACGACGTCGTGCGACGACGCTACGTTCGAAATACGC
GCGCGTCGCGAACGTACGTTCGCGACCACGCTTCAAATAGGCGTCGAACGTACGTGCACCACGCTACGTC
AAATATATAAGGCGTCGACGTTGCACCACGCTTCTCAAGCGCTACCAATAGGCGTCGAACGTTGCACCACG
CTTCAAATATGCGTCGAACGTTGCACCACGCTGAGGTAAGTCGAATAGGCGTCGAACGTTGCACCACGCT
ACGTCAATAGGCGTCGAACGTTGCACCACGCTTCAAATAGGCGTCGAACGTACGTGCACCACGCTTCAA
TAGGCGTCGAACGTTGCACCATCCTTCACAGCGCTTCAAATAGGCGTCGAACGTTGCACCACGCTTCAAAT
AGGCGTCGAACGTTGCACCACGCTTCAACGAATAGGCGTCGAACGTTGCACACGCTTCAAATAGGCGTCG
AACGTACGTGCACCACGCTTCAAGGTAAGTAATAGGCGTCGAACGTTGCACCACGCTACGTCAAATAGGC
GTCGAACGTACGTGCACCACGCTACGTCAACGAATAGGCGTCGAACGTTGCACCACGCTACGTCAATAGG
CGTCGAACGTTGCACCACGCTTCAACGAACGAATAGGCACGGTCGAACGTACGTGCACCACGCTTCAAATA
GGCGTCGAACGTTGCACCACGCTTCAAATAGGCGTCGAACGTACGTGCACCACGCTACGTCAAATCCTTC
ACAGTAGGCGTCGAACGTTGCACCACGCTTCAAATAGGCGTCGAACGTTGCACACGCTTCAAATAAGGCG
TCGAACGTTGCACCACGAGGTAAGTCTACGTCAAACGATAGGCGTCGAACGTACGTGCACCACG

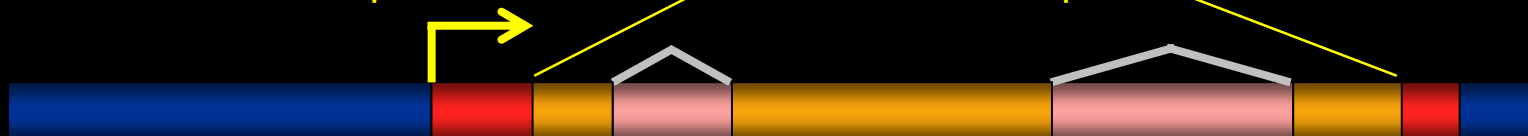
Sequence annotation



AGGCGTCGAACGTTGCACCACGCTTCAACGAATAGGCGTCGAACGTTGCACCACGCGTTACGAATACGC
GCTACGTCAACGACGACGATACGCGCGCGTCGCGACGACGTCGTGCGACGACGCTACGTGAAATACGC
GCGCGTCGCGAACGTACGTGCGGACCACGCTTCAAATAGGCGTCGAACGTACGTGCACCACGCTACGTC
AAATATATAAGGCGTCGACGTTGCACCACGCTTCTCAAGCGCTACCAATAGGCGTCGAACGTTGCACCACG
CTTCAAATATGCGTCGAACGTTGCACCACGCTGAGGTAAGTCGAATAGGCGTCGAACGTTGCACCACGCT
ACGTCAATAGGCGTCGAACGTTGCACCACGCTTCAAATAGGCGTCGAACGTACGTGCACCACGCTTCAA
TAGGCGTCGAACGTTGCACCATCCTTCACAGCGCTTCAAATAGGCGTCGAACGTTGCACCACGCTTCAAAT
AGGCGTCGAACGTTGCACCACGCTTCAACGAATAGGCGTCGAACGTTGCACACGCTTCAAATAGGCGTCG
AACGTACGTGCACCACGCTTCAAGGTAAGTAATAGGCGTCGAACGTTGCACCACGCTACGTCAAATAGGC
GTCGAACGTACGTGCACCACGCTACGTCAACGAATAGGCGTCGAACGTTGCACCACGCTACGTCAATAGG
CGTCGAACGTTGCACCACGCTTACGAACGAATAGGCACGGTCGAACGTACGTGCACCACGCTTCAAATA
GGCGTCGAACGTTGCACCACGCTTCAAATAGGCGTCGAACGTACGTGCACCACGCTACGTCAAATCCTTC
ACAGTAGGCGTCGAACGTTGCACCACGCTTCAAATAGGCGTCGAACGTTGCACACGCTTCAAATAAGGCG
TCGAACGTTGCACCACGAGGTAAGTCTACGTCAAACGATAGGCGTCGAACGTACGTGCACCACG

Sequence annotation

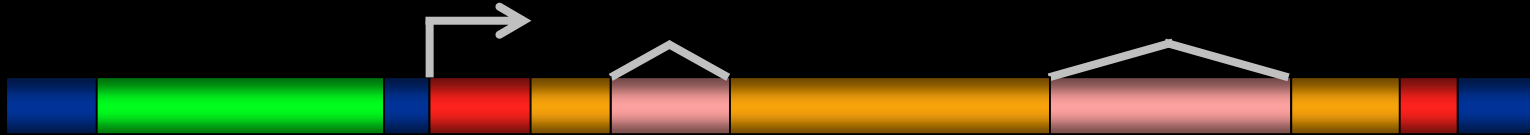
Transcription start site Translation start/stop sites



AGGCGTCGAACGTTGCACCACGCTTCAACGAATAGGCGTCGAACGTTGCACCACGCGTTACGAATACGC
GCTACGTCAACGACGACGATACGCGCGCGTCGCGACGACGTCGTGCGACGACGCTACGTGAAATACGC
GCGCGTCGCGAACGTACGTGCGGACCACGCTTCAAATAGGCGTCGAACGTACGTGCACCACGCTACGTC
AAATATATAAGGCGTCGACGTTGCACCACGCTTCTCAAGCGCTACCAATAGGCGTCGAACGTTGCACCACG
CTTCAAATATGCGTCGAACGTTGCACCACGCTGAGGTAAGTCGAATAGGCGTCGAACGTTGCACCACGCT
ACGTCAATAGGCGTCGAACGTTGCACCACGCTTCAAATAGGCGTCGAACGTACGTGCACCACGCTTCAA
TAGGCGTCGAACGTTGCACCATCCTTCACAGCGCTTCAAATAGGCGTCGAACGTTGCACCACGCTTCAAAT
AGGCGTCGAACGTTGCACCACGCTTCAACGAATAGGCGTCGAACGTTGCACACGCTTCAAATAGGCGTCG
AACGTACGTGCACCACGCTTCAAGGTAAGTAATAGGCGTCGAACGTTGCACCACGCTACGTCAAATAGGC
GTCGAACGTACGTGCACCACGCTACGTCAACGAATAGGCGTCGAACGTTGCACCACGCTACGTCAATAGG
CGTCGAACGTTGCACCACGCTTCAACGAACGAATAGGCACGGTCGAACGTACGTGCACCACGCTTCAAATA
GGCGTCGAACGTTGCACCACGCTTCAAATAGGCGTCGAACGTACGTGCACCACGCTACGTCAAATCCTTC
ACAGTAGGCGTCGAACGTTGCACCACGCTTCAAATAGGCGTCGAACGTTGCACACGCTTCAAATATAGGCG
TCGAACGTTGCACCACGAGGTAAGTCTACGTCAAACGATAGGCGTCGAACGTACGTGCACCACG

Sequence annotation

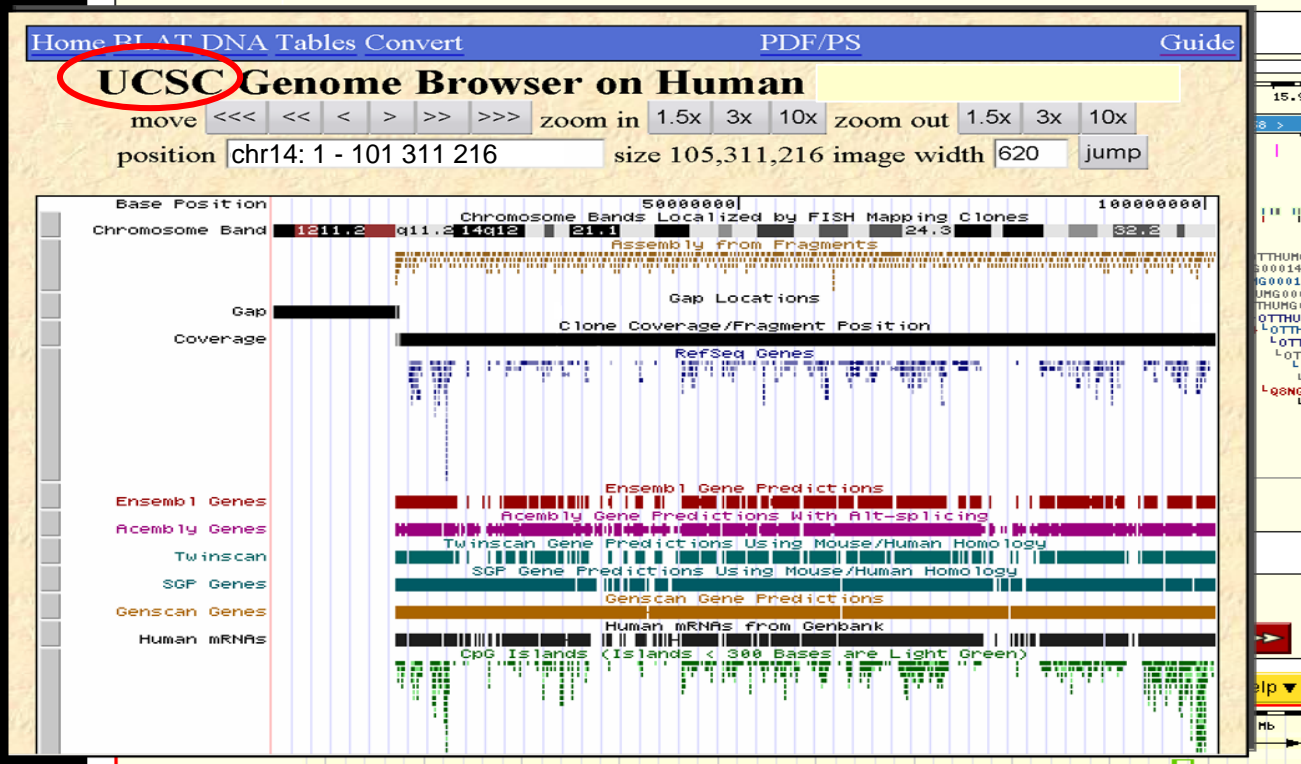
Promoter

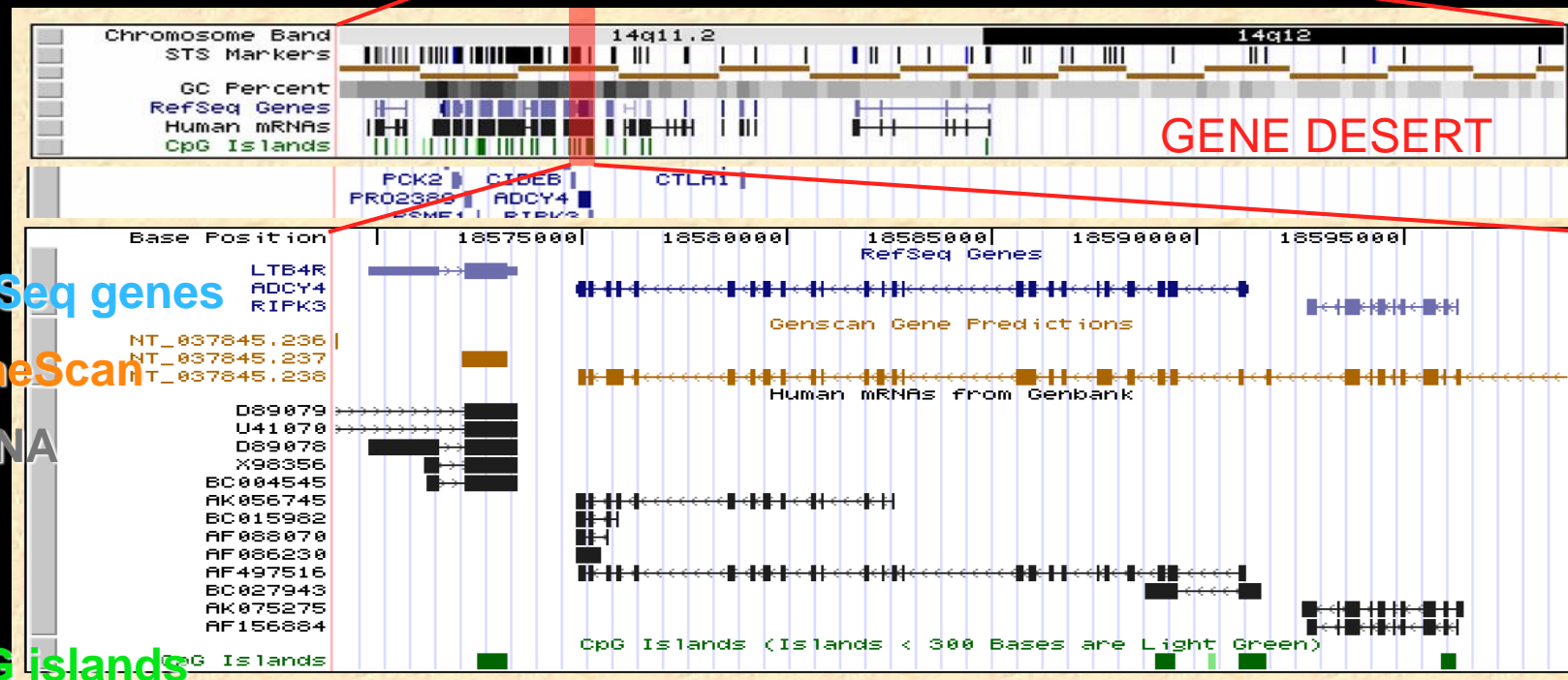
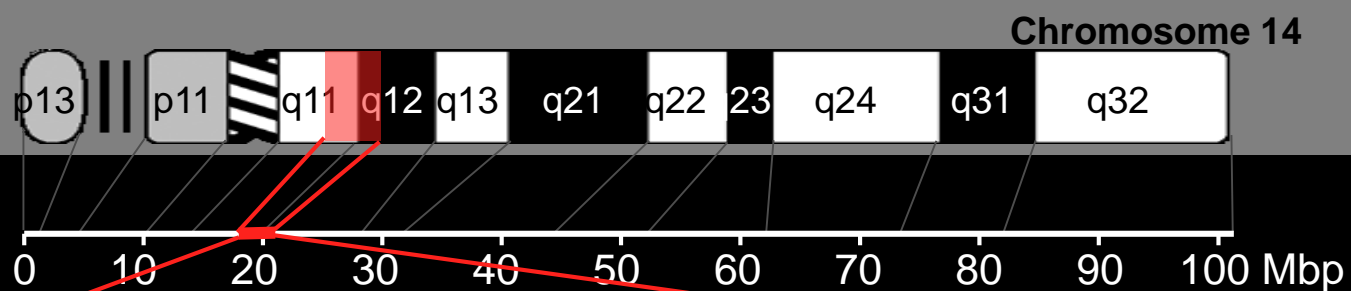
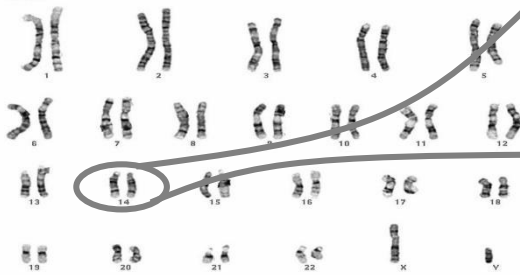


AGGCGTCGAACGTTGCACCACGCTTCAACGAATAGGCGTCGAACGTTGCACCACGCGTTACGAATACGC
GCTACGTCAACGACGACGATACGCGCGCGTCGCGACGACGTCGTGCGACGACGCTACGTCGAAATACGC
GCGCGTCGCGAACGTACGTCGCGACCACGCTTCAAATAGGCGTCGAACGTACGTGCACCACGCTACGTC
AAATATATAAGGCGTCGACGTTGCACCACGCTTCTCAAGCGCTACCAATAGGCGTCGAACGTTGCACCACG
CTTCAATATGCGTCGAACGTTGCACCACGCTGAGGTAAGTCGAATAGGCGTCGAACGTTGCACCACGCT
ACGTCAATAGGCGTCGAACGTTGCACCACGCTTCAAATAGGCGTCGAACGTACGTGCACCACGCTTCAA
TAGGCGTCGAACGTTGCACCATCCTTCACAGCGCTTCAAATAGGCGTCGAACGTTGCACCACGCTTCAAAT
AGGCGTCGAACGTTGCACCACGCTTCAACGAATAGGCGTCGAACGTTGCACACGCTTCAAATAGGCGTCG
AACGTACGTGCACCACGCTTCAAGGTAAGTAATAGGCGTCGAACGTTGCACCACGCTACGTCAAATAGGC
GTCGAACGTACGTGCACCACGCTACGTCAACGAATAGGCGTCGAACGTTGCACCACGCTACGTCAATAGG
CGTCGAACGTTGCACCACGCTTCAACGAACGAATAGGCACGGTCGAACGTACGTGCACCACGCTTCAAATA
GGCGTCGAACGTTGCACCACGCTTCAAATAGGCGTCGAACGTACGTGCACCACGCTACGTCAAATCCTTC
ACAGTAGGCGTCGAACGTTGCACCACGCTTCAAATAGGCGTCGAACGTTGCACACGCTTCAAATAAGGCG
TCGAACGTTGCACCACGAGGTAAGTCTACGTCAAACGATAGGCGTCGAACGTACGTGCACCACG



Genome browsers





3 Mbp

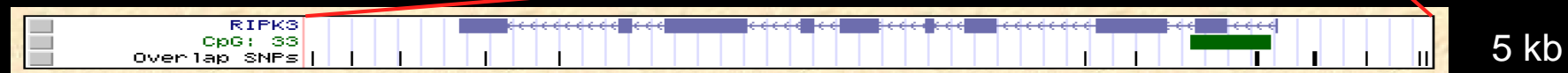
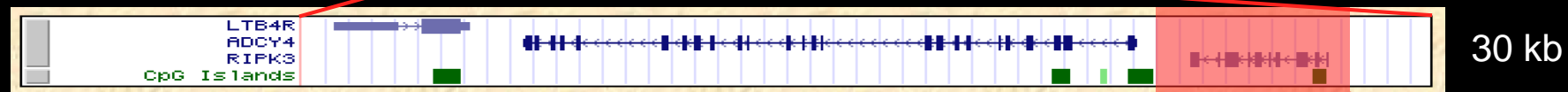
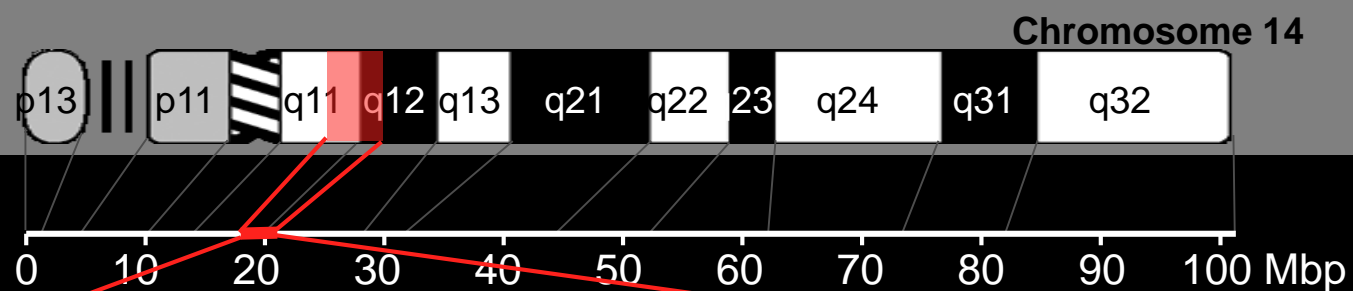
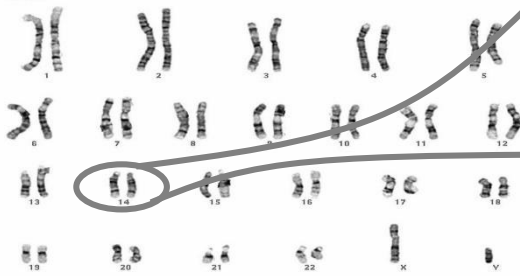
30 kb

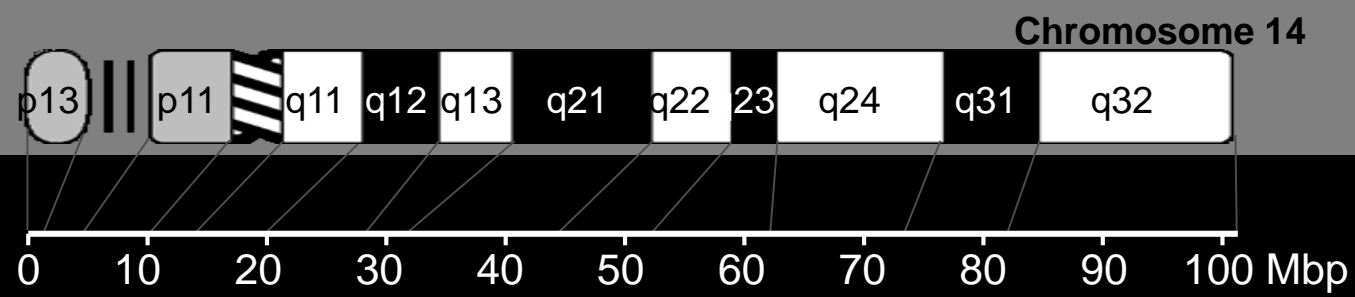
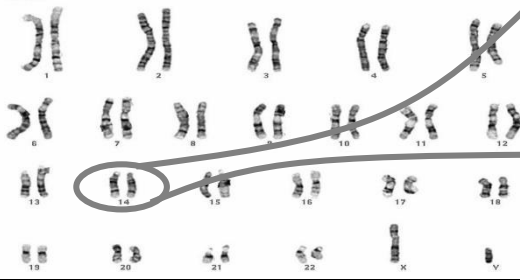
RefSeq genes

GeneScan

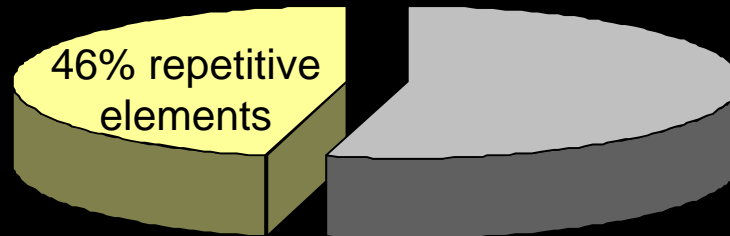
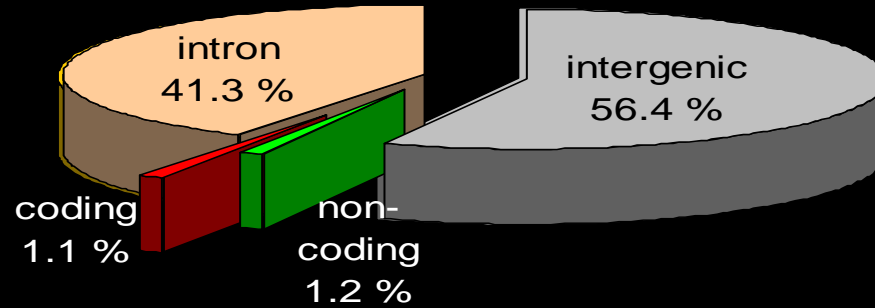
mRNA

CpG islands





- 1050 genes
- 1 gene / 100 kb



Gene ontology

- + Molecular function
- + Biological process
- + Cellular component

Current release 2024-09-08:

43 939 GO terms | 7 894 411 annotations

1 573 444 gene products | 5 426 species

geneontology.org/

Gene ontology

- + Molecular function
- Biological process
 - + behaviour
 - + cellular process
 - + physiological process
 - + viral life cycle
 - + development
- + Cellular component

Current release 2024-09-08:

43 939 GO terms | 7 894 411 annotations

1 573 444 gene products | 5 426 species

geneontology.org/

Gene ontology

Current release 2024-09-08:

43 939 GO terms | 7 894 411 annotations

1 573 444 gene products | 5 426 species

geneontology.org/

- + Molecular function
- Biological process
 - + behaviour
 - cellular process
 - + cell communication
 - + cell death
 - + cell differentiation
 - + cell motility
 - + membrane fusion
 - + physiological process
 - + viral life cycle
 - + development

Gene ontology

Current release 2024-09-08:

43 939 GO terms | 7 894 411 annotations

1 573 444 gene products | 5 426 species

geneontology.org/

- + Molecular function

- Biological process

 - + behaviour

 - cellular process

 - cell communication

 - + cell adhesion

 - + cell invasion

 - + signal transduction

 - + response to extra-cellular stimulus

 - + cell-cell signalling

 - + host-pathogen interaction

 - + cell death

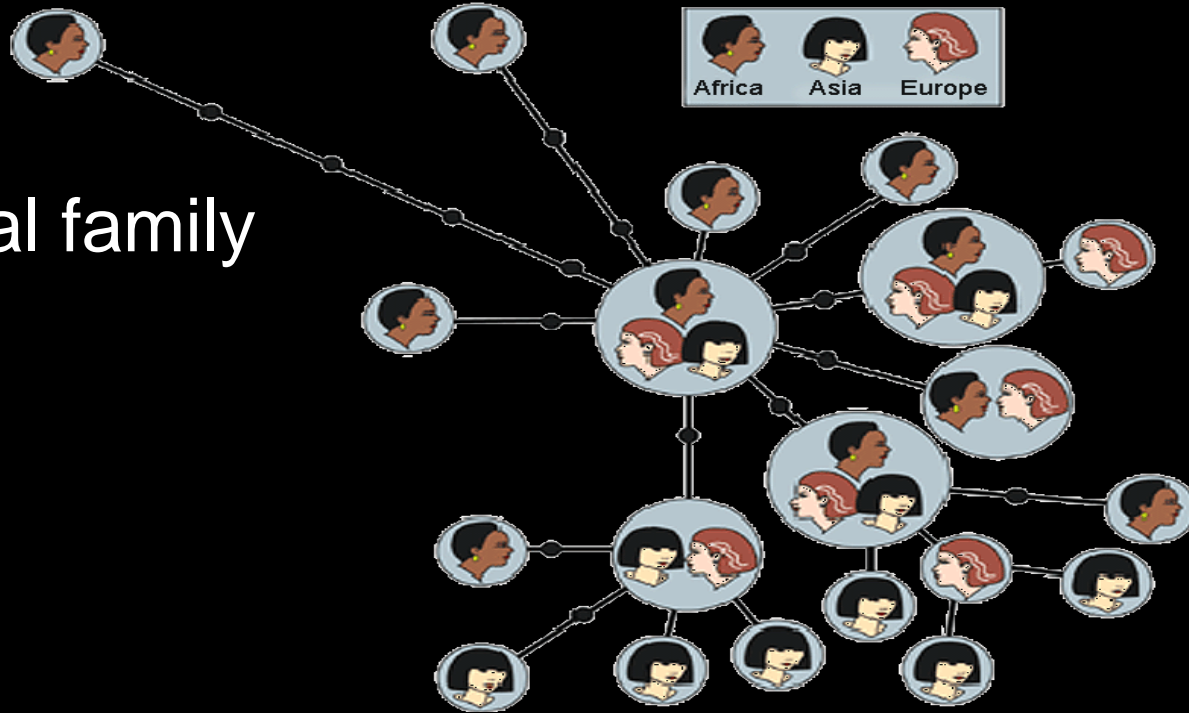
 - + cell differentiation

DNA sequence variation

→ variant protein product?

DNA sequence variation

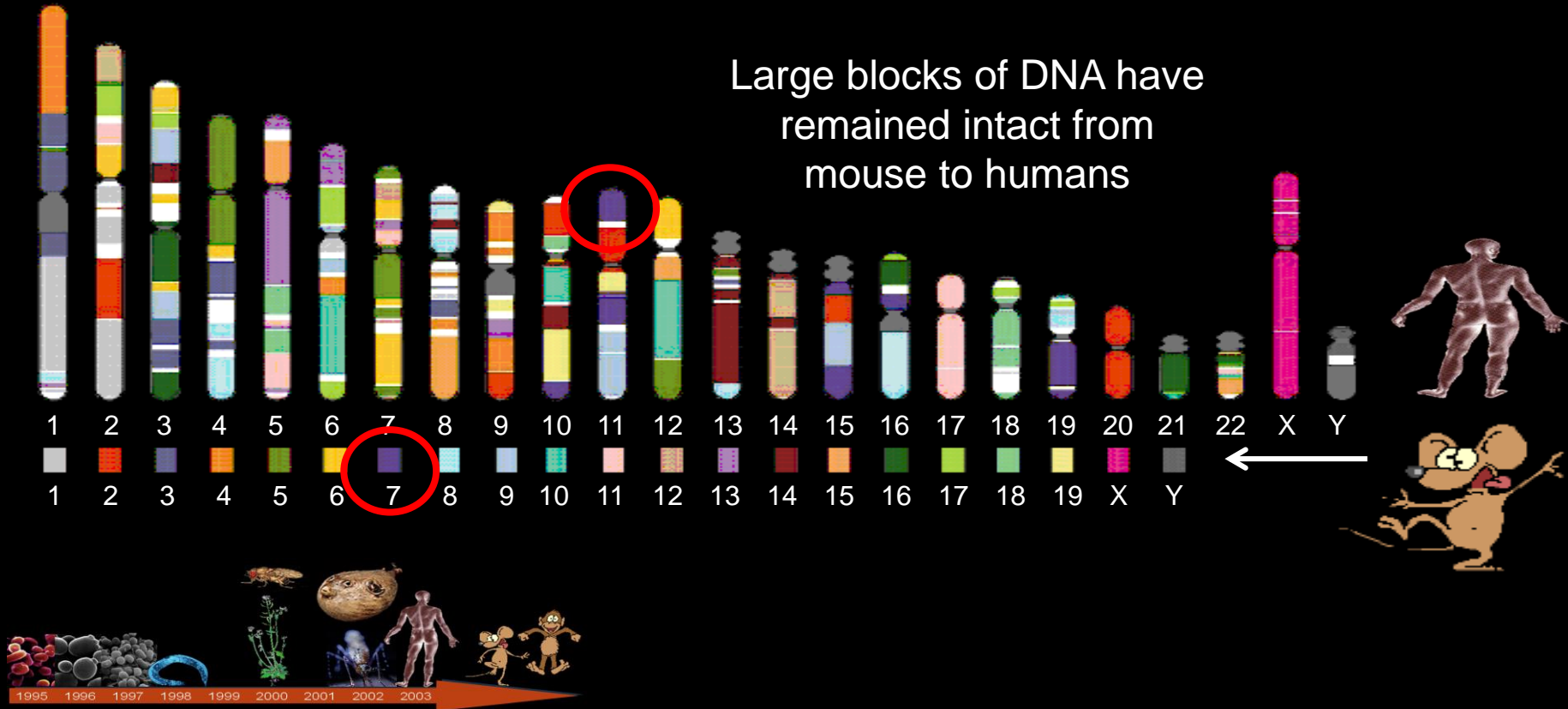
The global family



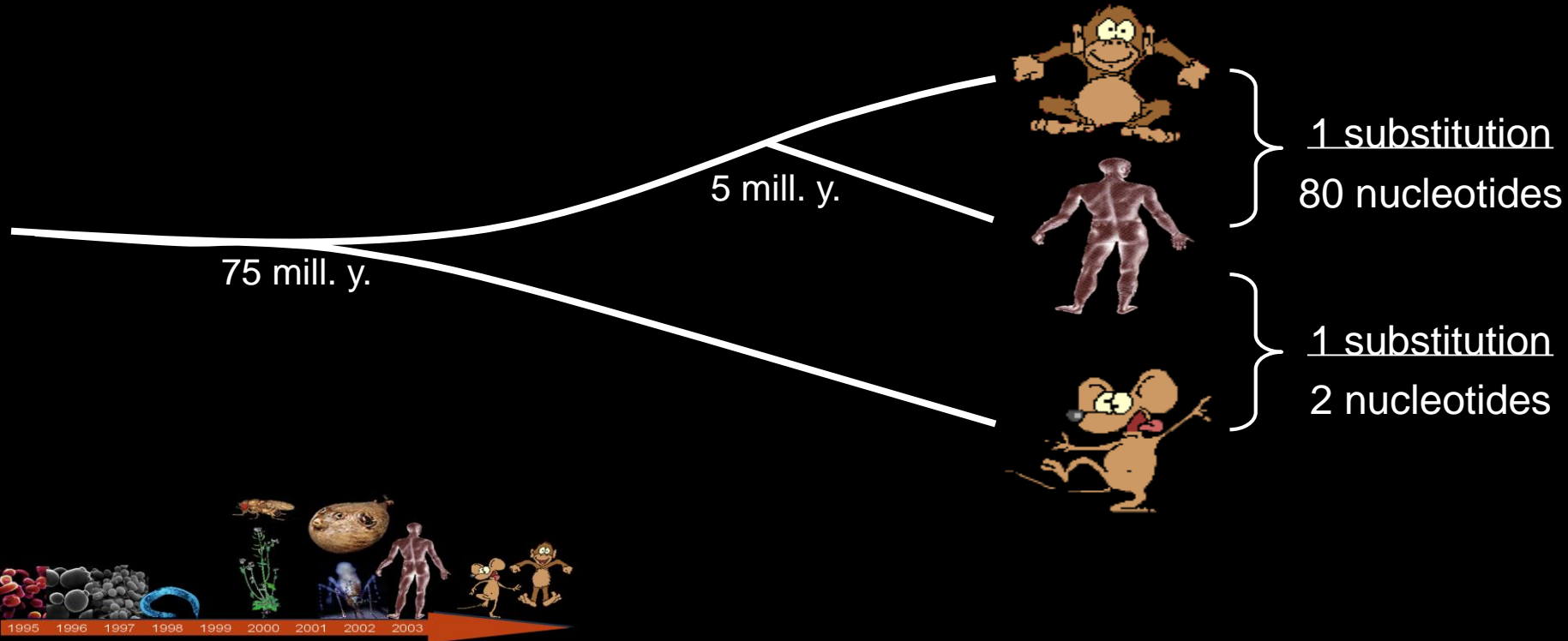
Svante Pääbo, Science 2001
Nobel Prize, 2022

Comparative genomics

Large blocks of DNA have remained intact from mouse to humans

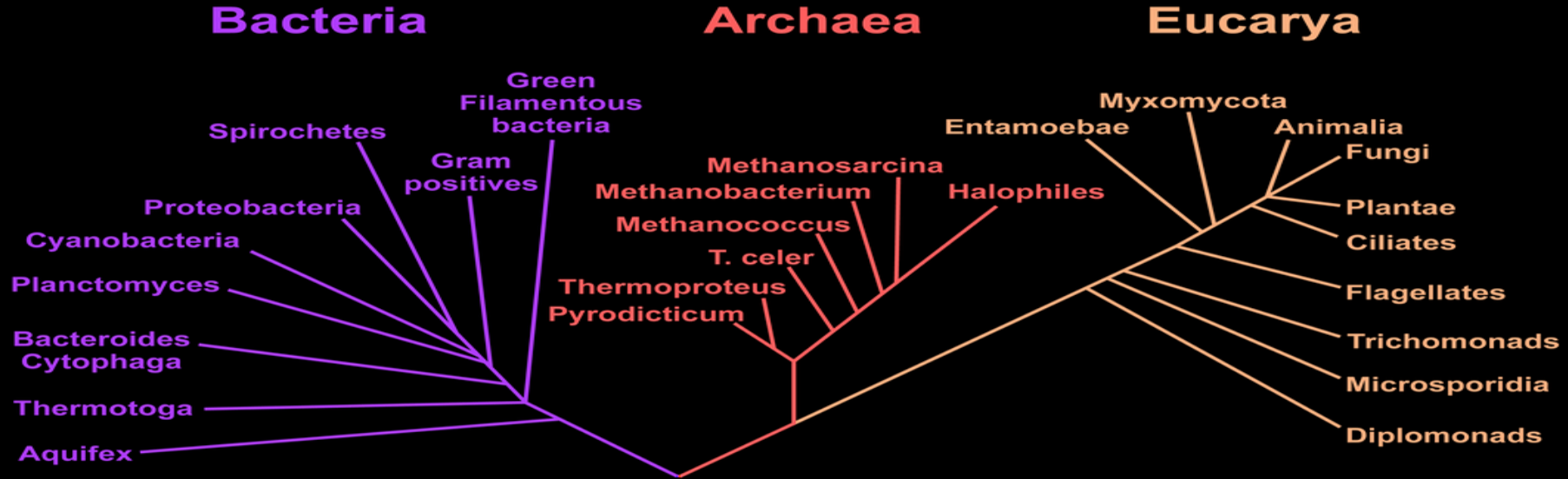


Comparative genomics



The phylogenetic tree of life

Taxonomy in biology. Assigning vectors (from aligned genomics data, commonly ribosomal-RNA) to species, calculating matrix of distances, group them with cluster analysis to obtain a tree or dendrogram



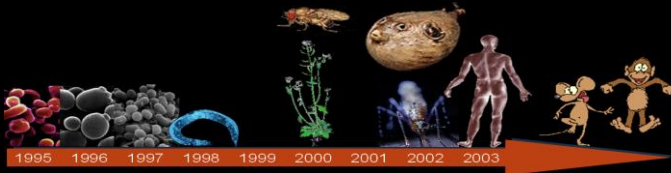
Comparative genomics

Distribution of gene-rich areas

human



non-human (*i.e.* most non-mammalian model organisms)



Comparative genomics

Repeat sequences

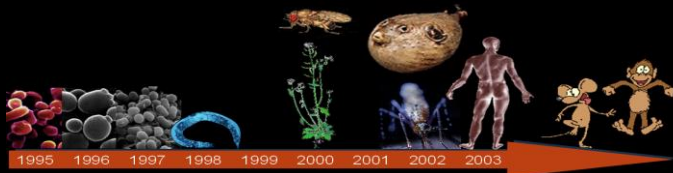
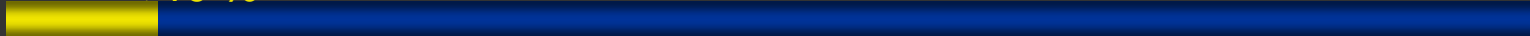
human

~ 50 %



non-human (*i.e.* most non-mammalian model organisms)

< 10 %

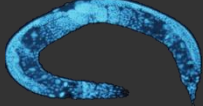


Comparative genomics

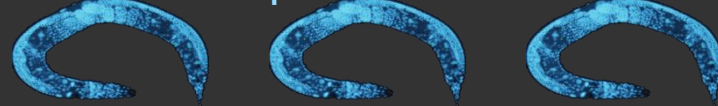
The human genome has many protein variants

human

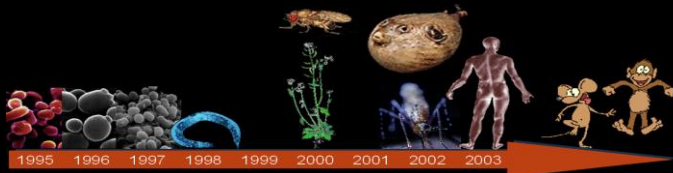
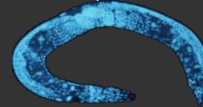
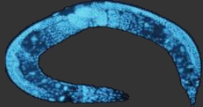
Number of genes



Number of protein variants



non-human (*i.e.* most non-mammalian model organisms; here: *C. elegans*)

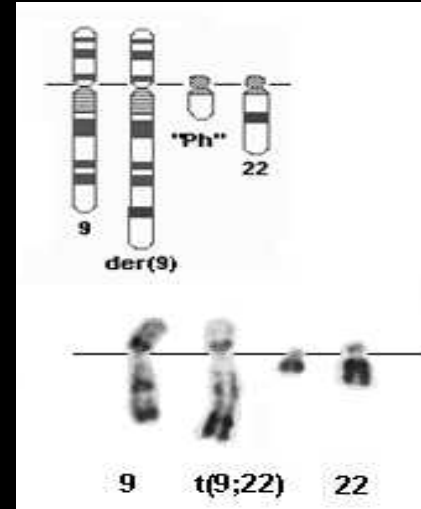


Gene regulation

- Time
- Space
- Level
- Alternative splicing
- Activity

Genomics into clinics

- Individualise treatment
- Targeted and tailored “designer” medicine
 - Gleevec
 - t(9;22): Philadelphia chr.
 - Chronic myeloid leukaemia



Genomics into clinics

- Individualise treatment
- Targeted and tailored “designer” medicine
- High-throughput technologies
 - Primarily sequencing of DNA and RNA
 - DNA mutations/variation: base-level and larger
 - RNA expression: quantitative and qualitative
- Pre-symptomatic diagnosis (and genetic predisposition)
 - Huntington’s disease
 - Cystic fibrosis
 - Breast cancer
- Potential future health

Ethical, legal, and social implications

