



Bioinformatika

Jak velké genetické rozdíly jsou důležité –
srovnání DNA člověka a šimpanze

Jan Pačes

jan.paces@img.cas.cz

Ústav molekulární genetiky AVČR

<http://www.img.cas.cz>



CZECH FOBIA

Úvod do terminologie

- Bioinformatika: jak se informace ukládá a šíří v živé přírodě
- DNA: deoxyribonukleová kyselina
- Genomika
 - Strukturní a funkční
- Transkriptomika
- Proteomika

Proč čteme DNA?

Fundamental struggle of evolution takes place not among individuals or species but at the level of the chromosome. Organisms serve genes, rather than the other way around: We are machines for propagating DNA.

Richard Dawkins

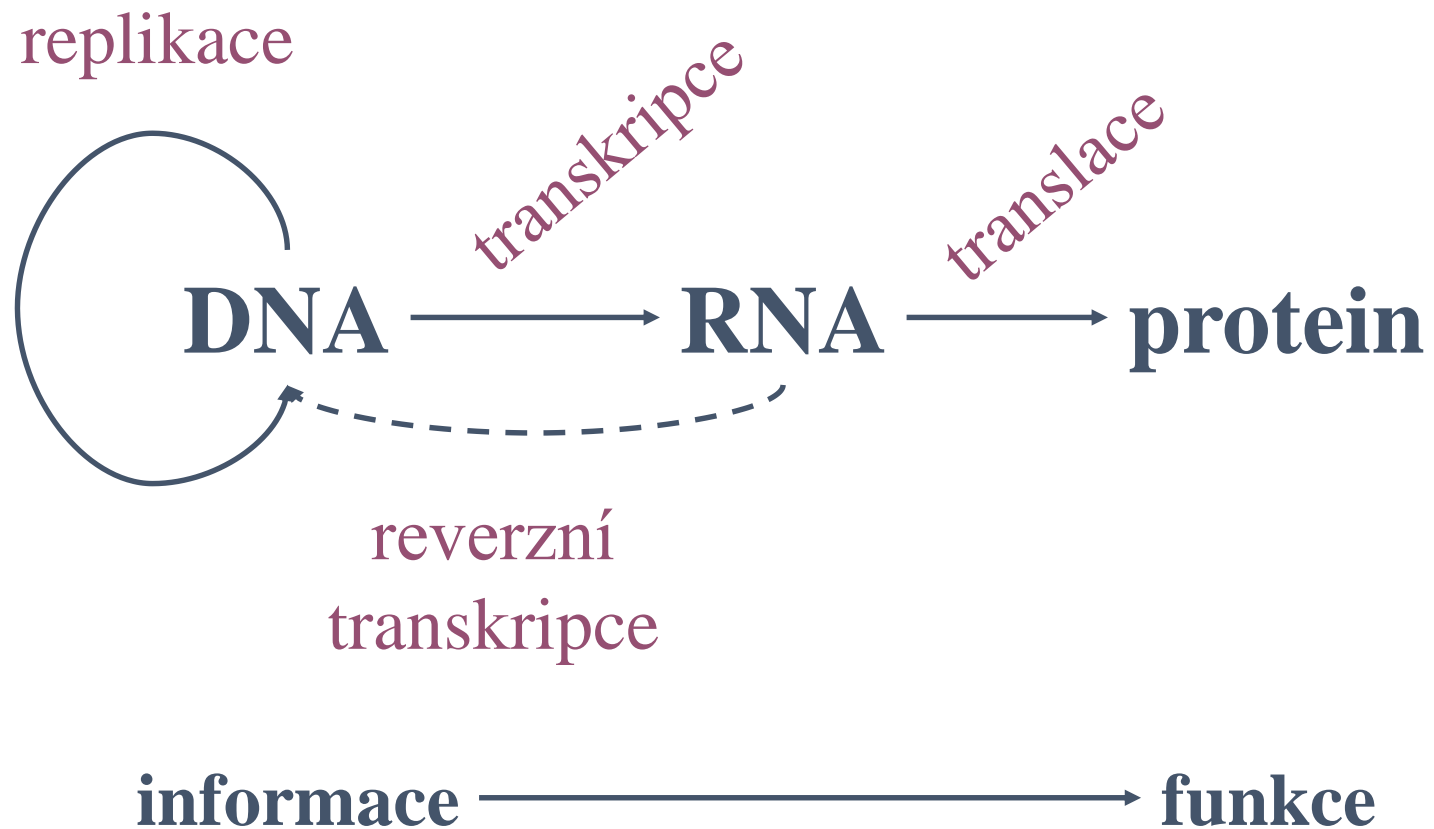
DNA zde není proto, aby sloužila organizmům, naopak, organizmy jsou zde proto, aby sloužily DNA

Richard Dawkins

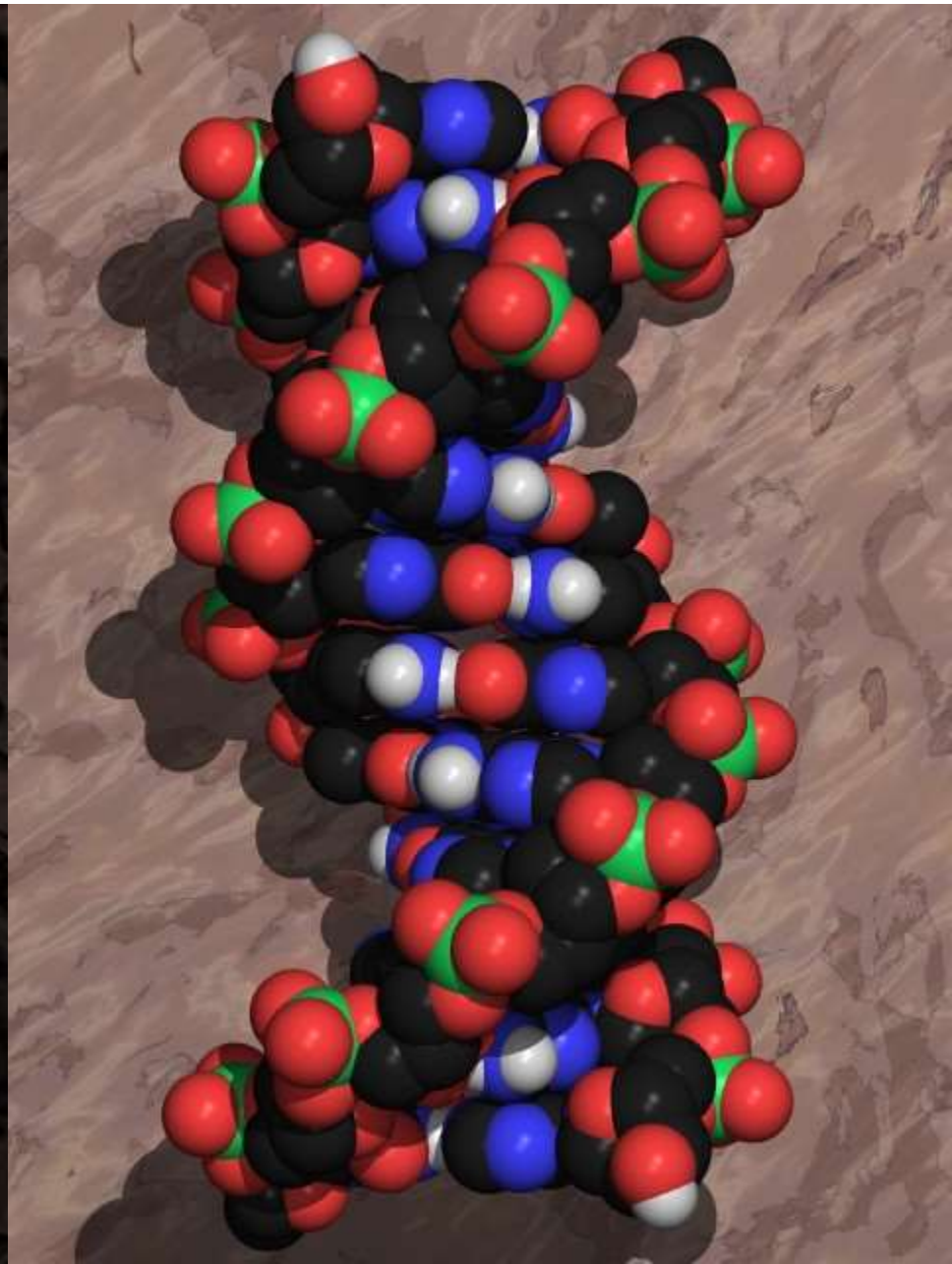
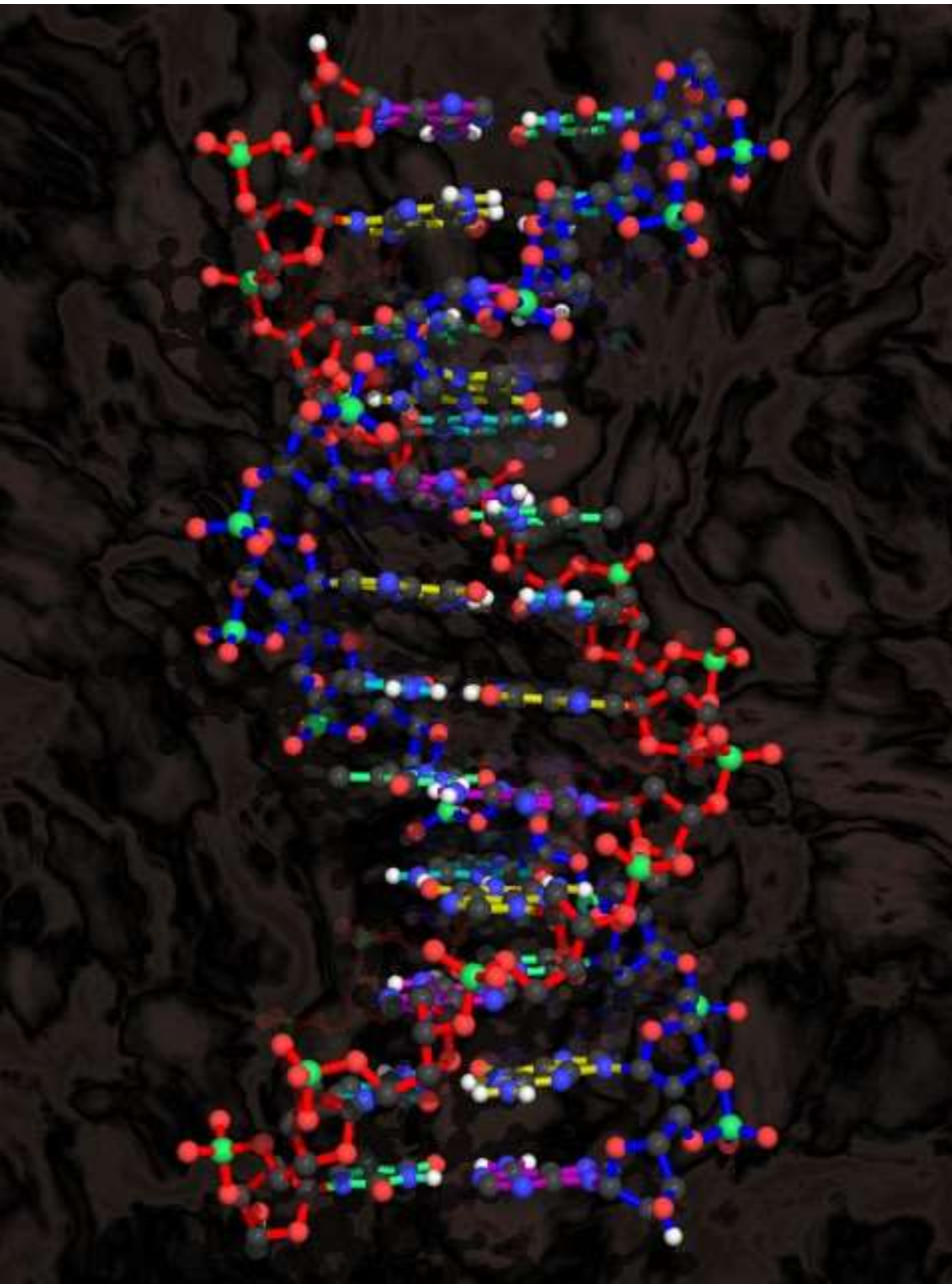
Komparativní genetika se zabývá podobnostmi. Ale v případě genomu šimpanze hledáme především rozdíly.

Svante Pääbo

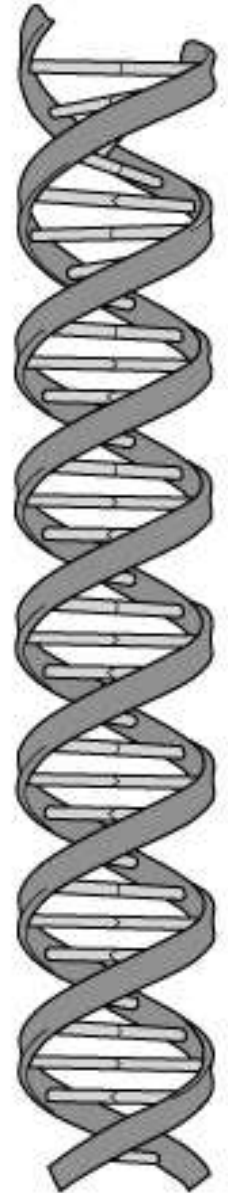
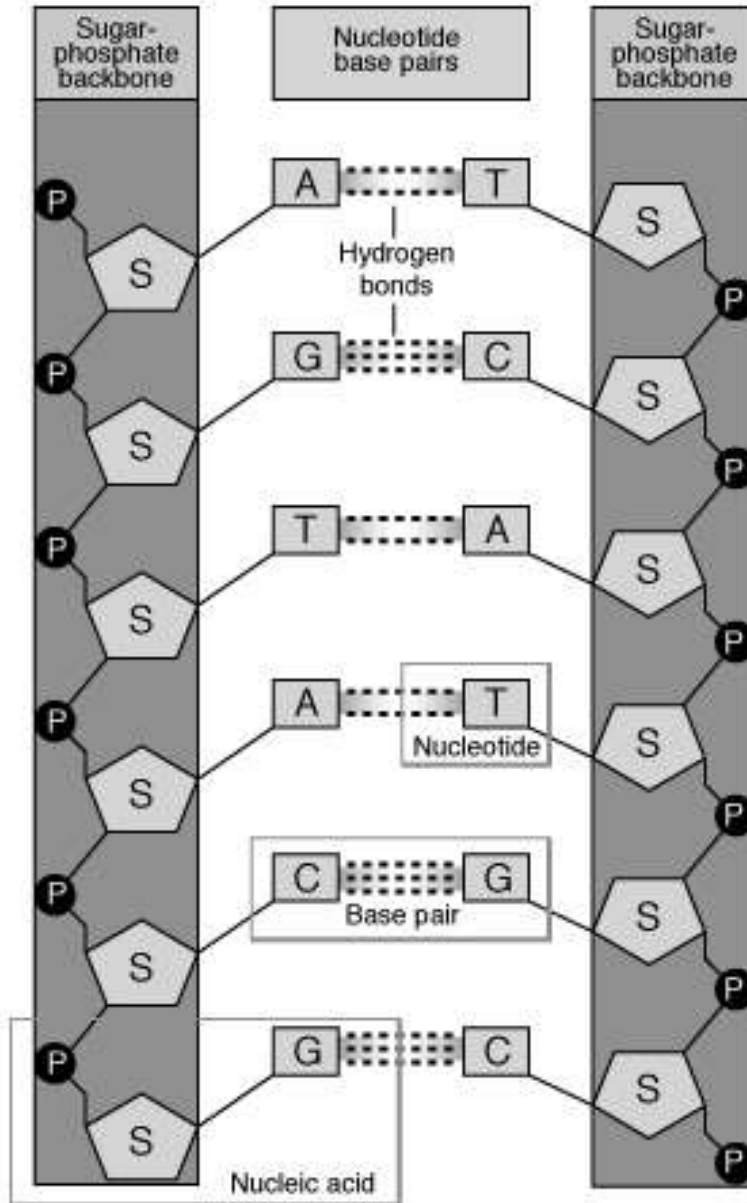
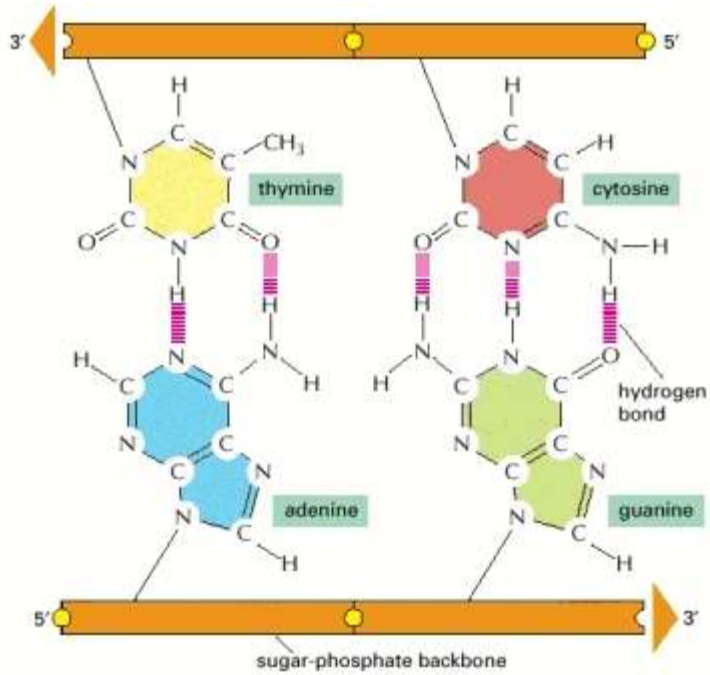
Centrální dogma molekulární genetiky



DNA



DNA



Transkripce a translace

DNA 5' > ATGAAGCCGAGTCAT 3'
 3' TACTTCGGCTCAGTA <5'

transkripce

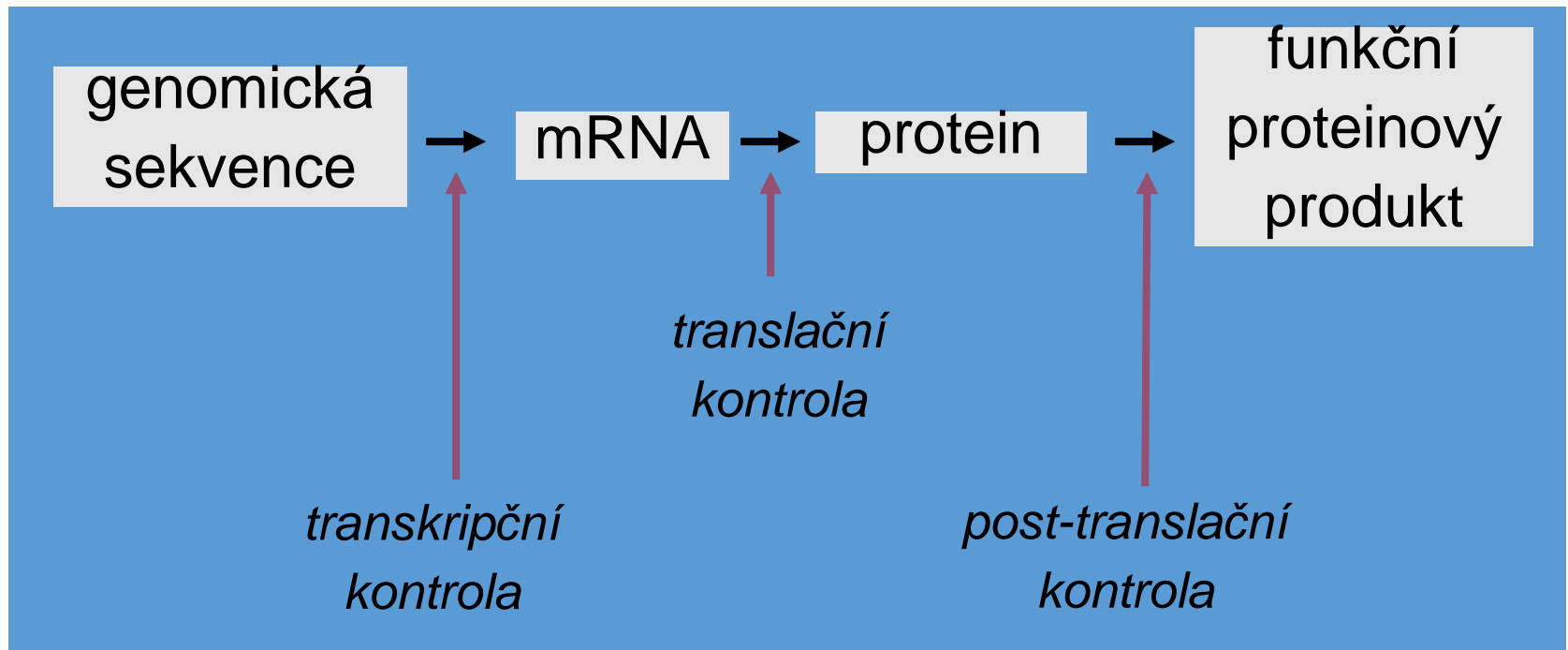
mRNA 5' > AUGAAGCCGAGUCAU 3'

translace

Protein N> MetLysProSerHis C

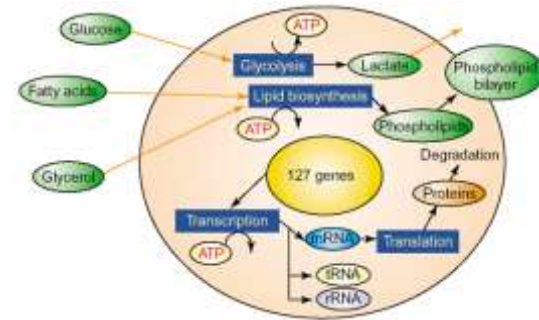
Úrovně kontroly

počet genů \neq úroveň mRNA \neq úroveň genové exprese \neq
množství a efektivní účinnost proteinu



Jak dobře dnes rozumíme DNA

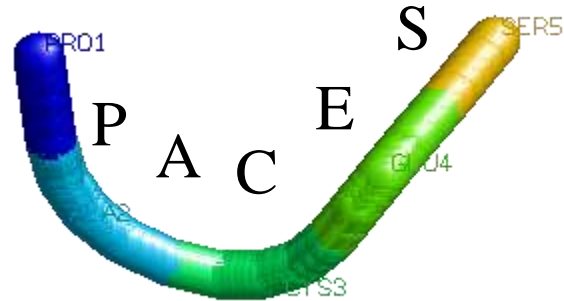
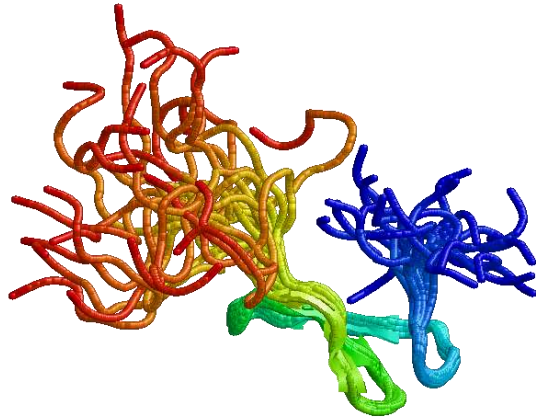
- E-cell



- Syntetická biologie:

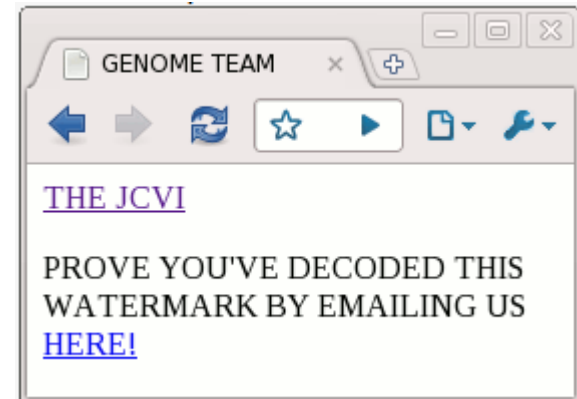
- *Mycoplasma laboratorium* Gibson D, et al. (2008): Complete Chemical Synthesis, Assembly, and Cloning of a *Mycoplasma genitalium* Genome. Science. DOI: 10.1126/science.1151721
- *Synthia*: 1. syntetický organizmus Gibson D, et al. (2010): Creation of a bacterial cell controlled by a chemically synthesized genome. Science. DOI: 10.1126/science.1190719

Ukládání informace v DNA



Watermarks:




- VENTERINSTITVTE CRAIGVENTER
HAMSMITH CINDIANDCLYDE
GLASSANDCLYDE
- Html code in *synthia*



Ukládání informace v DNA

STORAGE LIMITS

Estimates based on bacterial genetics suggest that digital DNA could one day rival or exceed today's storage technology.

	 Hard disk	 Flash memory	 Bacterial DNA
Read-write speed (μs per bit)	> ~3,000–5,000	> ~100	> <100
Data retention (years)	> >10	> >10	> >100
Power usage (watts per gigabyte)	> ~0.04	> ~0.01–0.04	> <10 ⁻¹⁰
Data density (bits per cm ³)	> ~10 ¹³	> ~10 ¹⁶	> ~10 ¹⁹

WEIGHT OF DNA NEEDED TO STORE WORLD'S DATA



©nature

Zakódování Shakespeareových sonetů do DNA

Thou art more lovely ...

text do ASCII

0101010101000111000101001 ...

ASCII do „trits“ (0,1,2)

20112 20200 02110 10002 ...

„trits“ do DNA

TAGAT GTGTA CAGAC TAGCG ...

aby se každé písmenko lišilo od předcházejícího



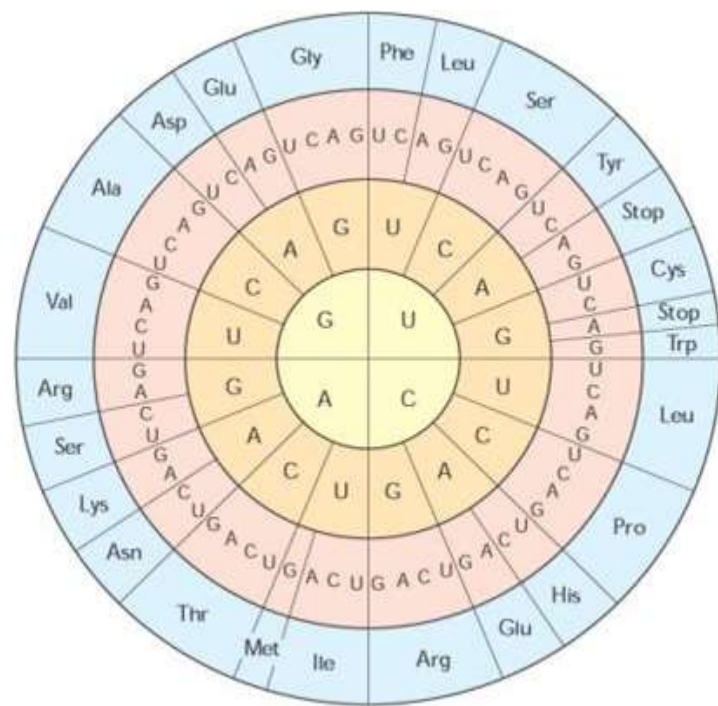
DNA fragmenty

překrývající se a s unikátním indexem

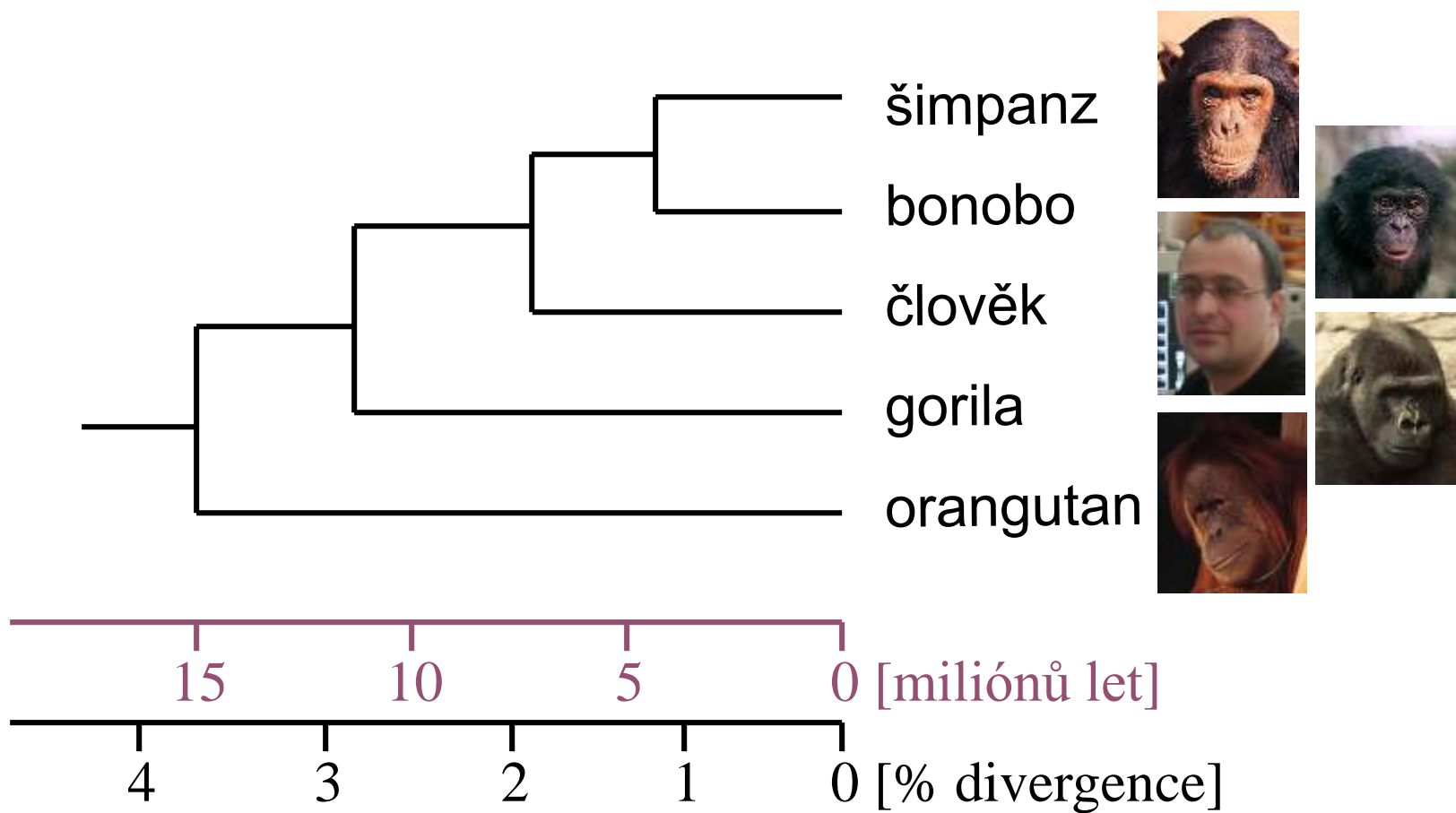
Jak sledovat (měřit) evoluci

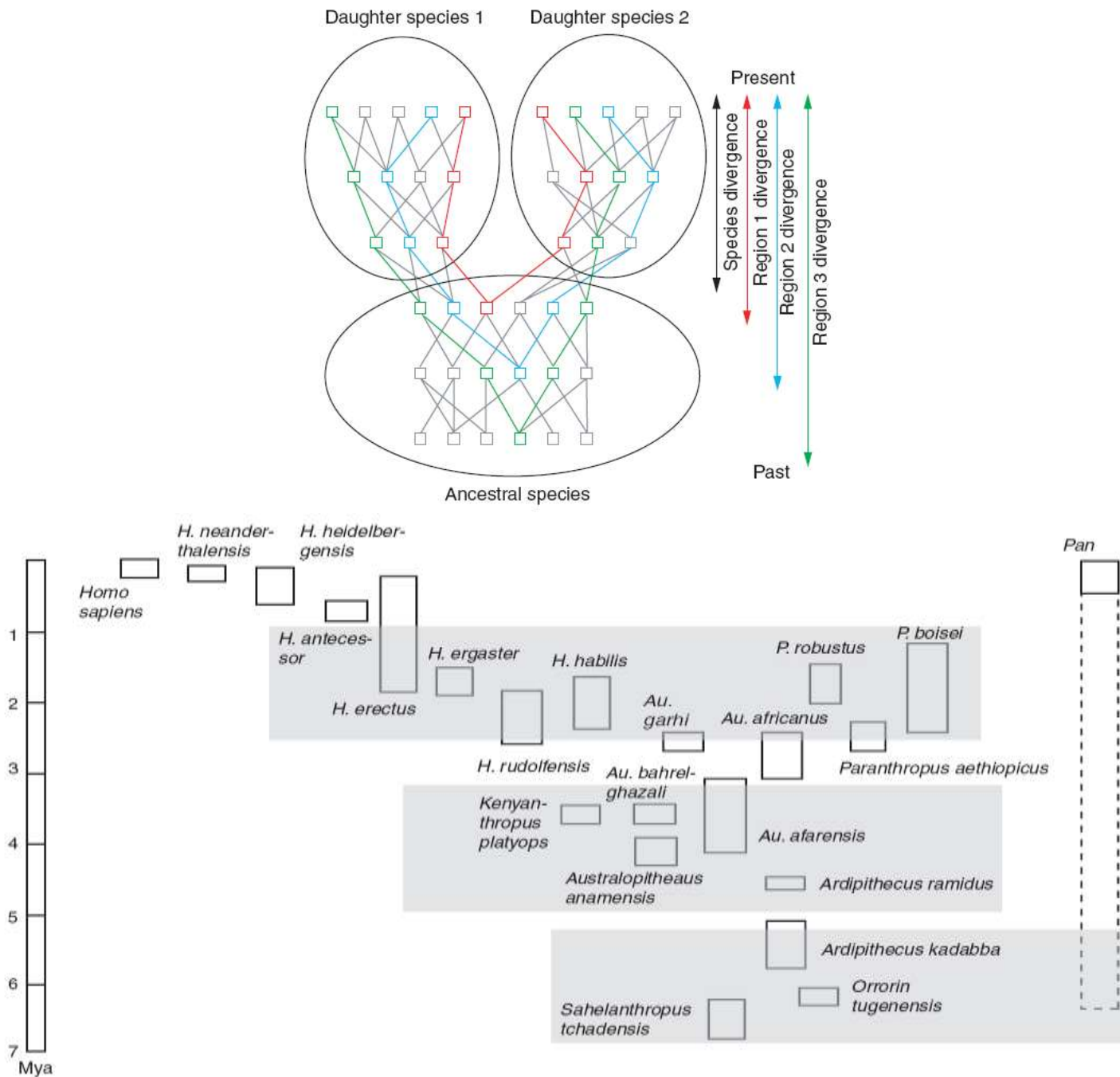
Ka/Ks (dn/ds) ratio:
poměr nesynonymních
a synonymních substitucí

$\ll 1$: negativní (purifying) selekce
 ~ 1 : neutrální
 $\gg 1$: pozitivní selekce



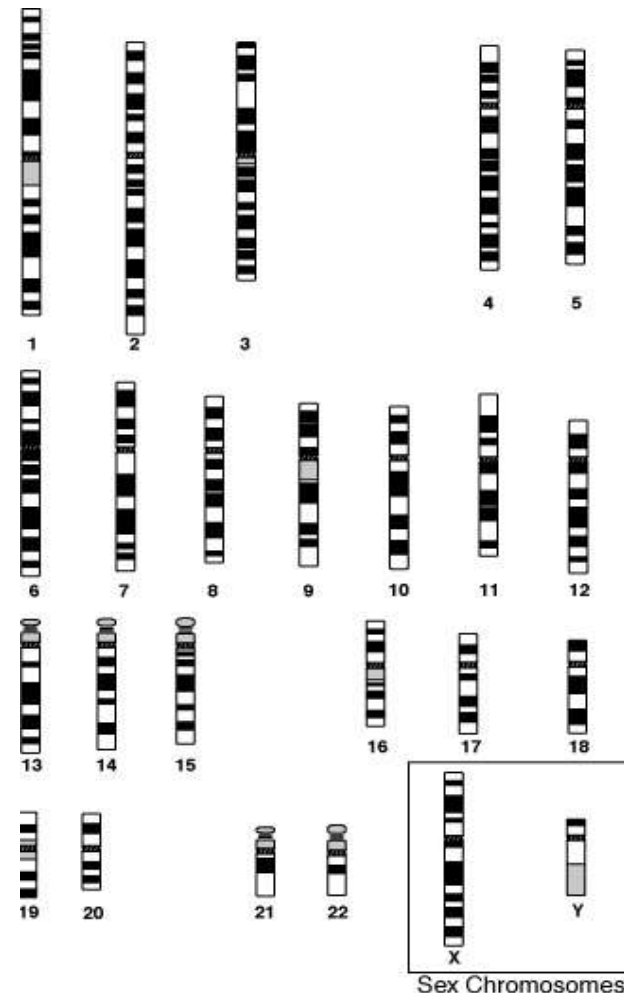
Evolve primátů





Lidský genom

22 chromozómových párů
1 autozóm (sex chromozóm)
~3,3 miliard párů bází
(~10% heterochromatin)
~25 tisíc genů



Koho jsme přečetli?

HGC: 9 neznámých lidí

- 5x mužská krev
- 3x spermie
- 1x 987SK buňky

Celera: 2 muži, 3 ženy

- Afroameričan
- Asiat - Číňan
- 2 Zakavkazané
- Hispánec - Mexičan

šimpanz: Clint (Yerkes National Primate Research Center)

Clint

He's tall, dark, and handsome, with a grin that turns heads, especially those of older women. Smart, playful, and a flirt, he's happiest when someone's scratching his back.



January 8th, 2005: Clint was put down at the Yerkes National Primate Research Center in Atlanta. The cause of death was not immediately known. Clint, 24, was the living reference point for \$18 million worth of genetic code.

Základní rozdíly hs x pt

1.44% rozdíl mezi DNA

68000 indels mezi hs chr. 21 a pt chr. 22

15% všech CpG je mutováno (23x více transicí a 7x více transverzí)

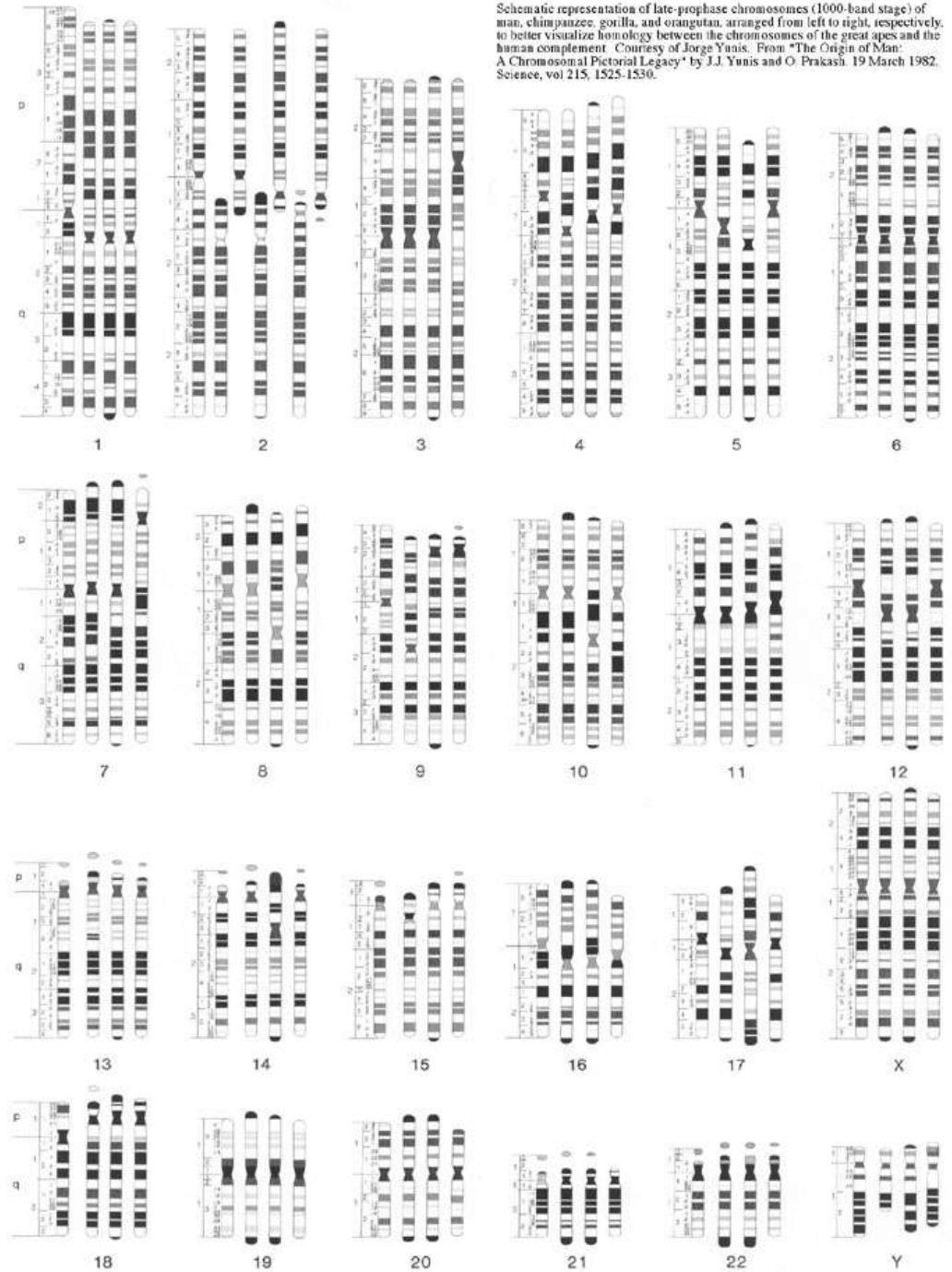
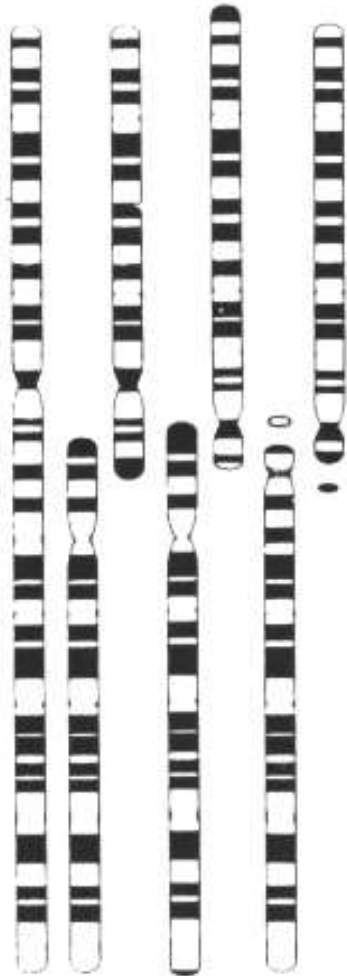
83% genů obsahuje rozdíl na úrovni aminokyselin

nejvíce jsou mutovány U3' oblasti



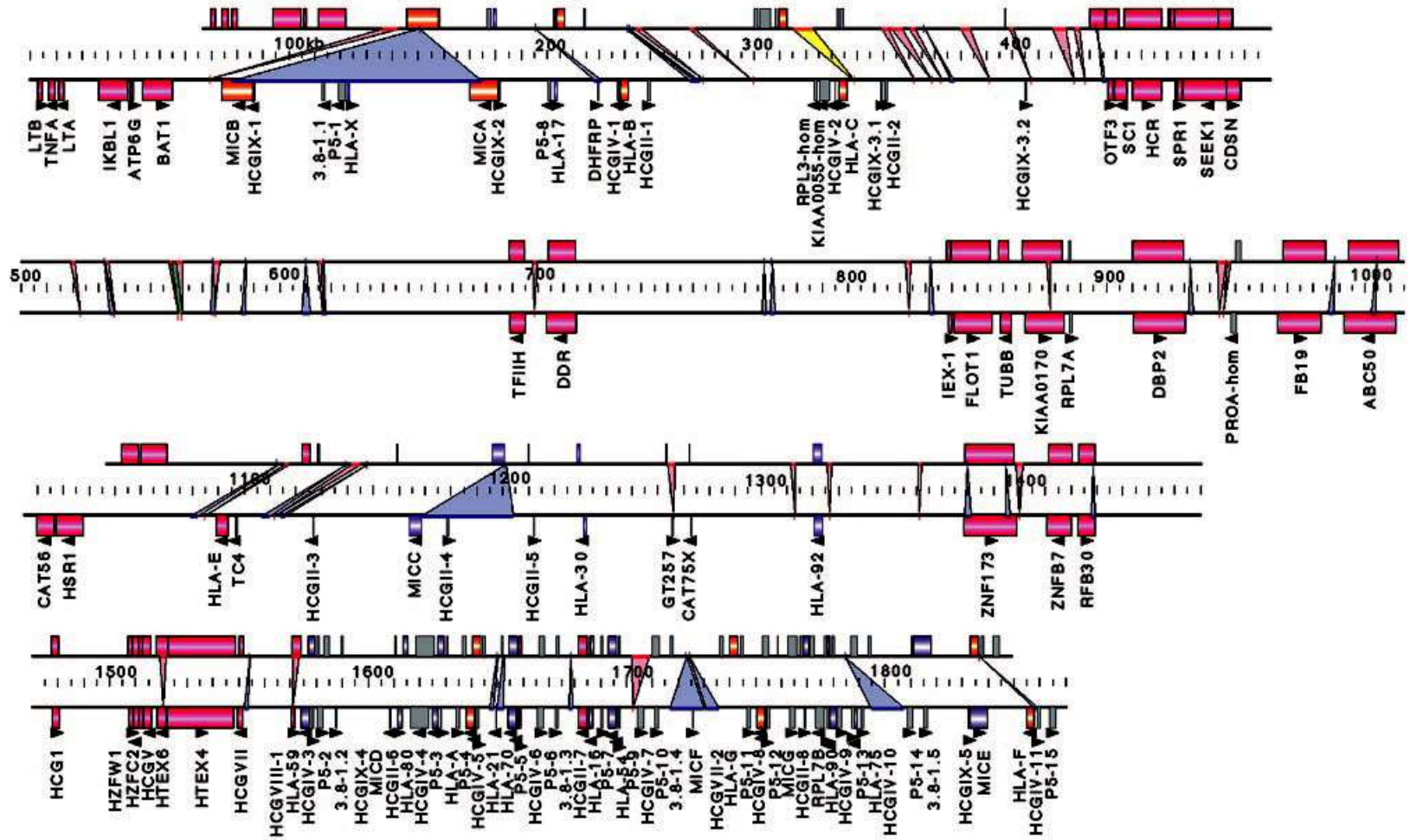
Hledáme něco navíc

člověk
šimpanz
gorila
orangutan

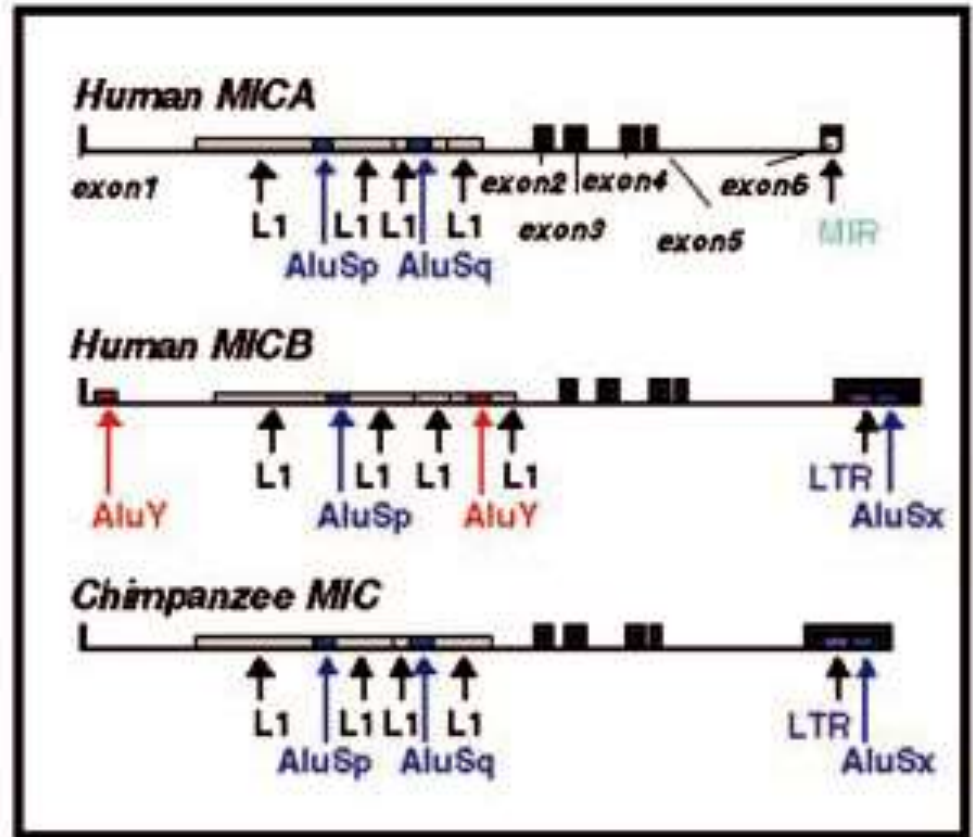
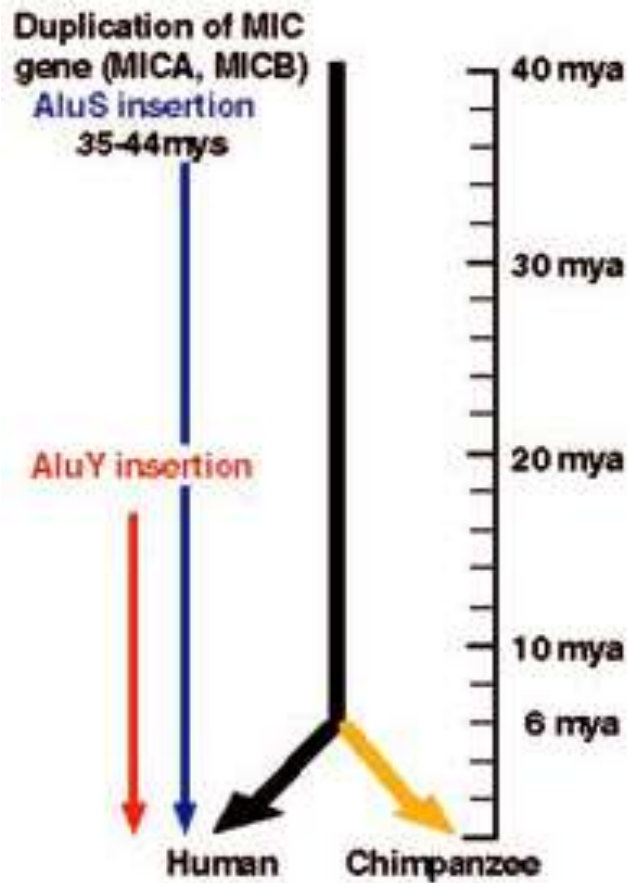


Schematic representation of late-prophase chromosomes (1000-band stage) of man, chimpanzee, gorilla, and orangutan, arranged from left to right, respectively, to better visualize homology between the chromosomes of the great apes and the human complement. Courtesy of Jorge Yunis. From "The Origin of Man: A Chromosomal Pictorial Legacy" by J.J. Yunis and O. Prakash, 19 March 1982, Science, vol 215, 1525-1530.

MHC I



MIC A/B



Hledáme změnu

FOXP2

one major hope is that the differences between the sequences will reveal the genetic basis for our mental and linguistic capacities

- obsahuje forkhead doménu
- poškození způsobuje abnormality v řeči a chápání jazyka (KE rodina)
- silně konzervován - pouze 3 aminokyselinové změny mezi člověkem a myší
- dvě specifické záměny se rozšířily před cca 100 000 – 200 000 lety (vznik moderního člověka)
thr -> asp (233) a asp -> ser (325)

KE family

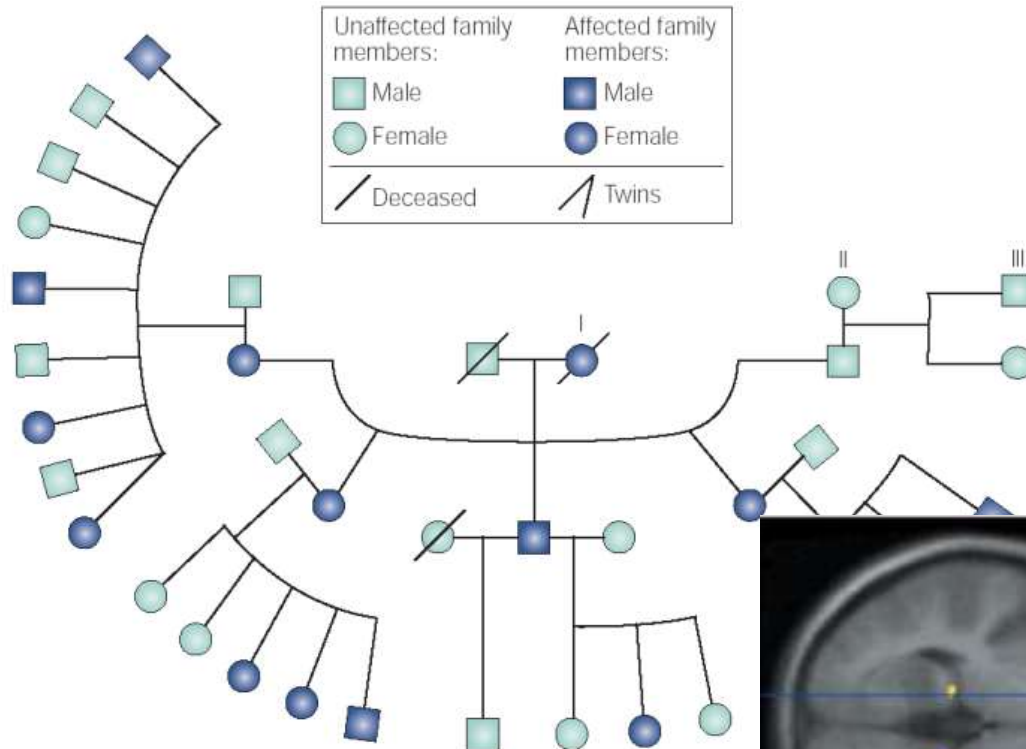
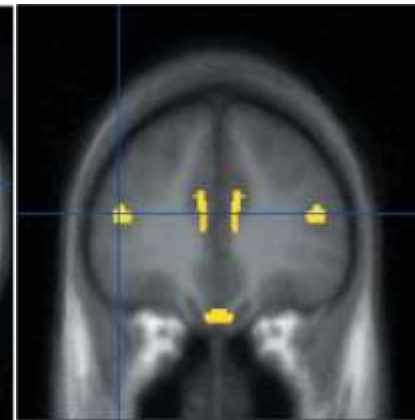
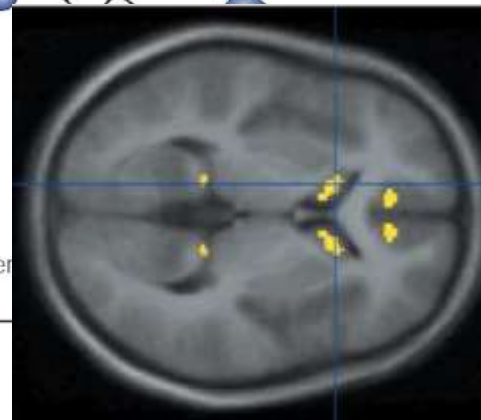
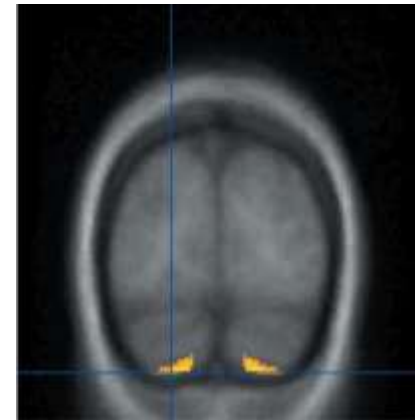
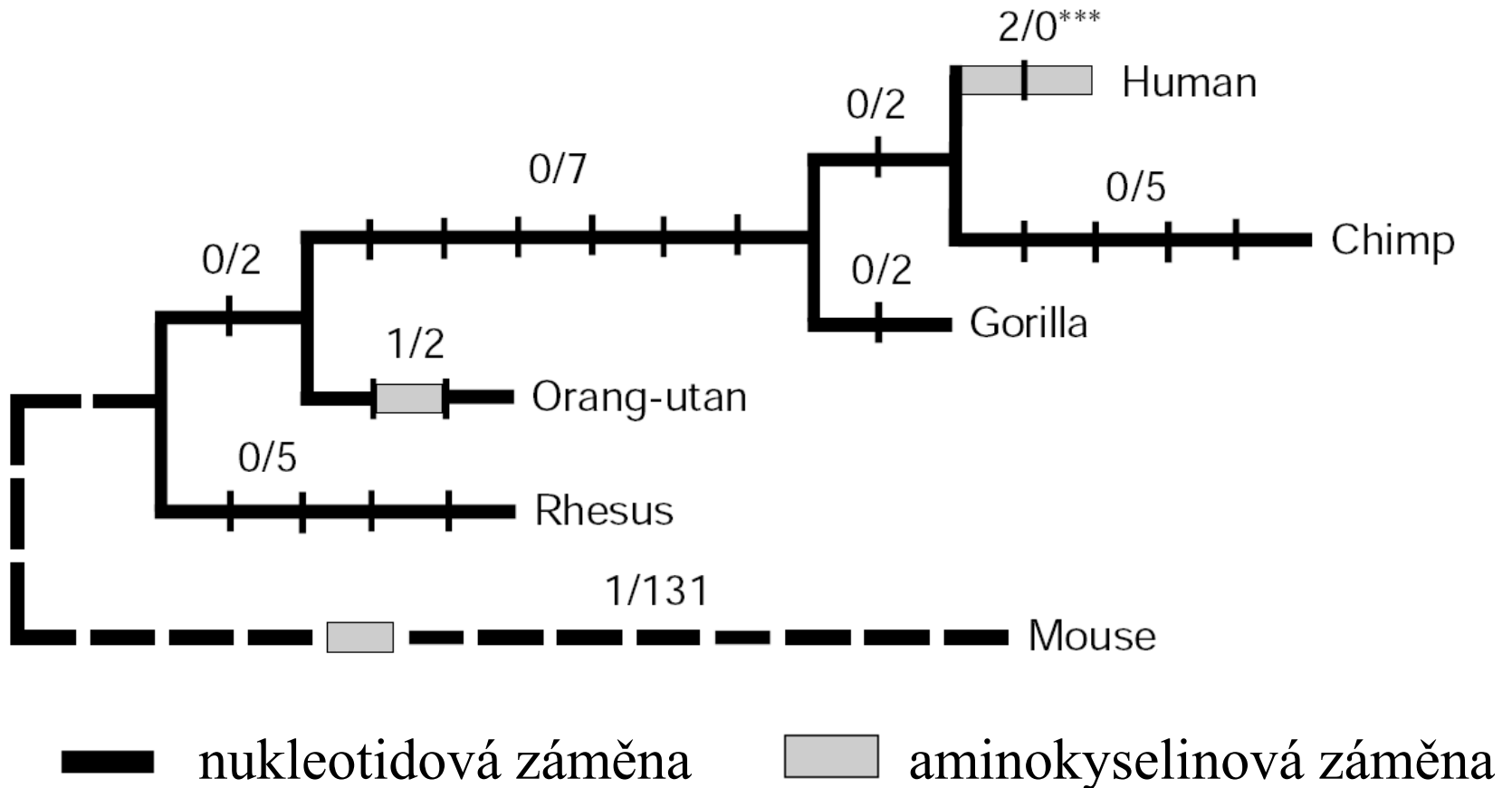


Figure 1 | **Pedigree of the KE family.** I, II and III represent the generation, from REF. 14 © (2002) Oxford University Press.

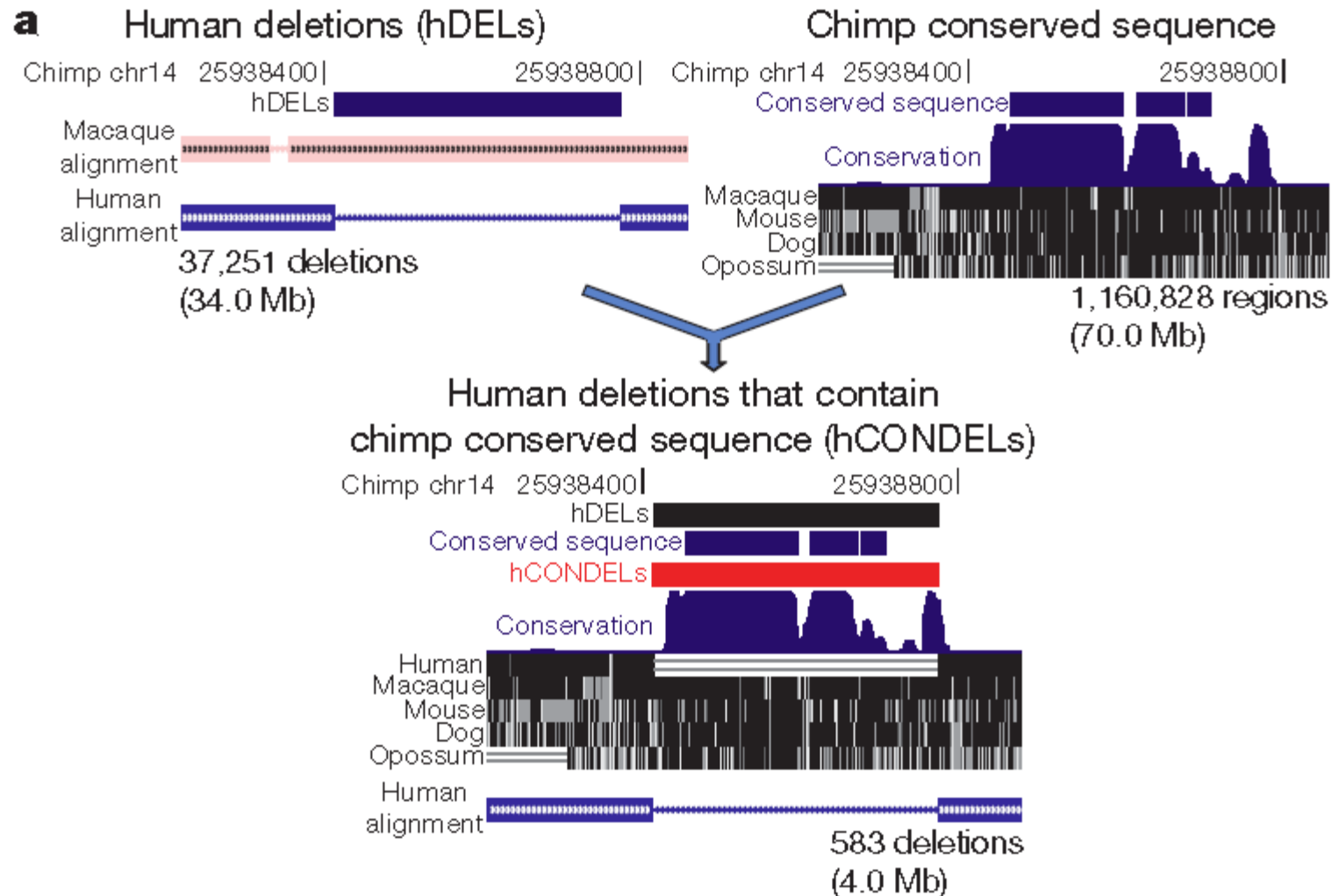


evoluce FOXP2

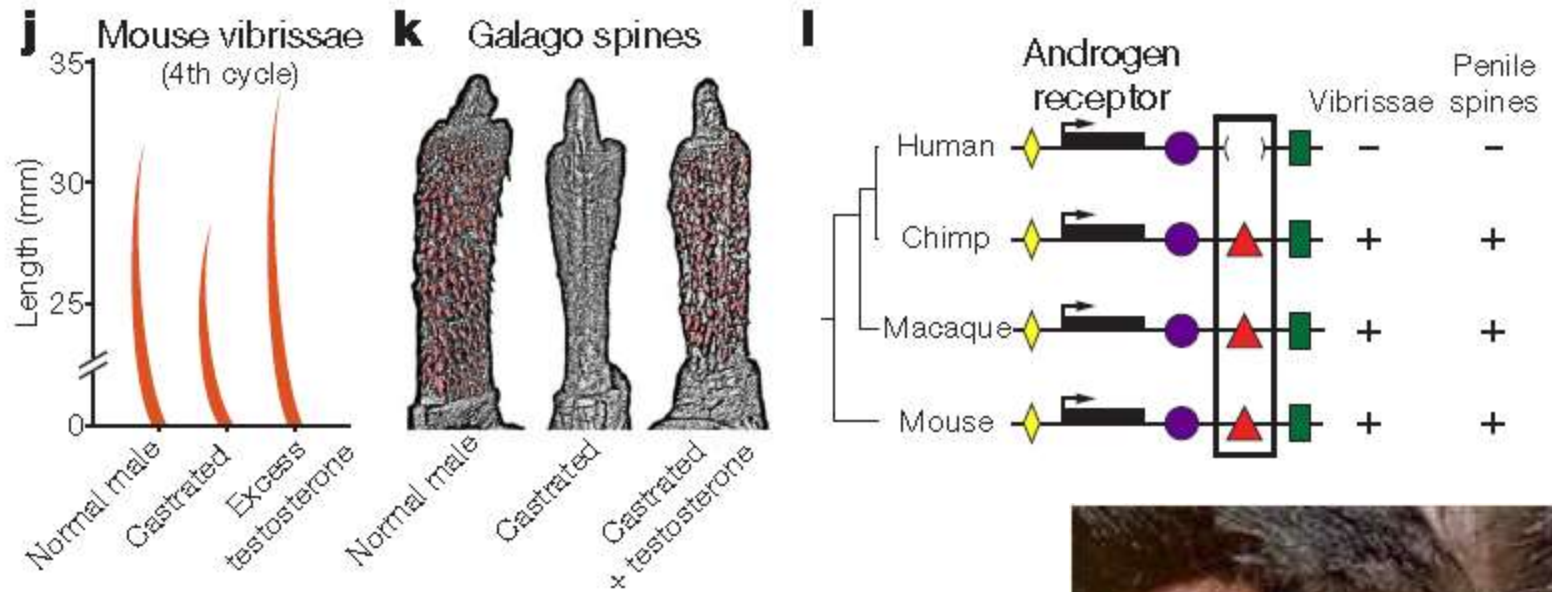


Hledáme co chybí

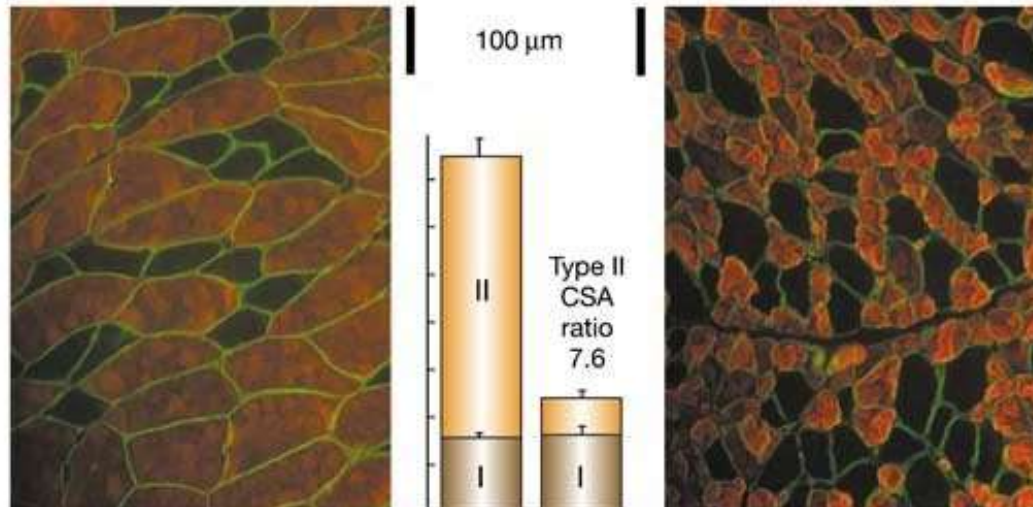
hCONDELS

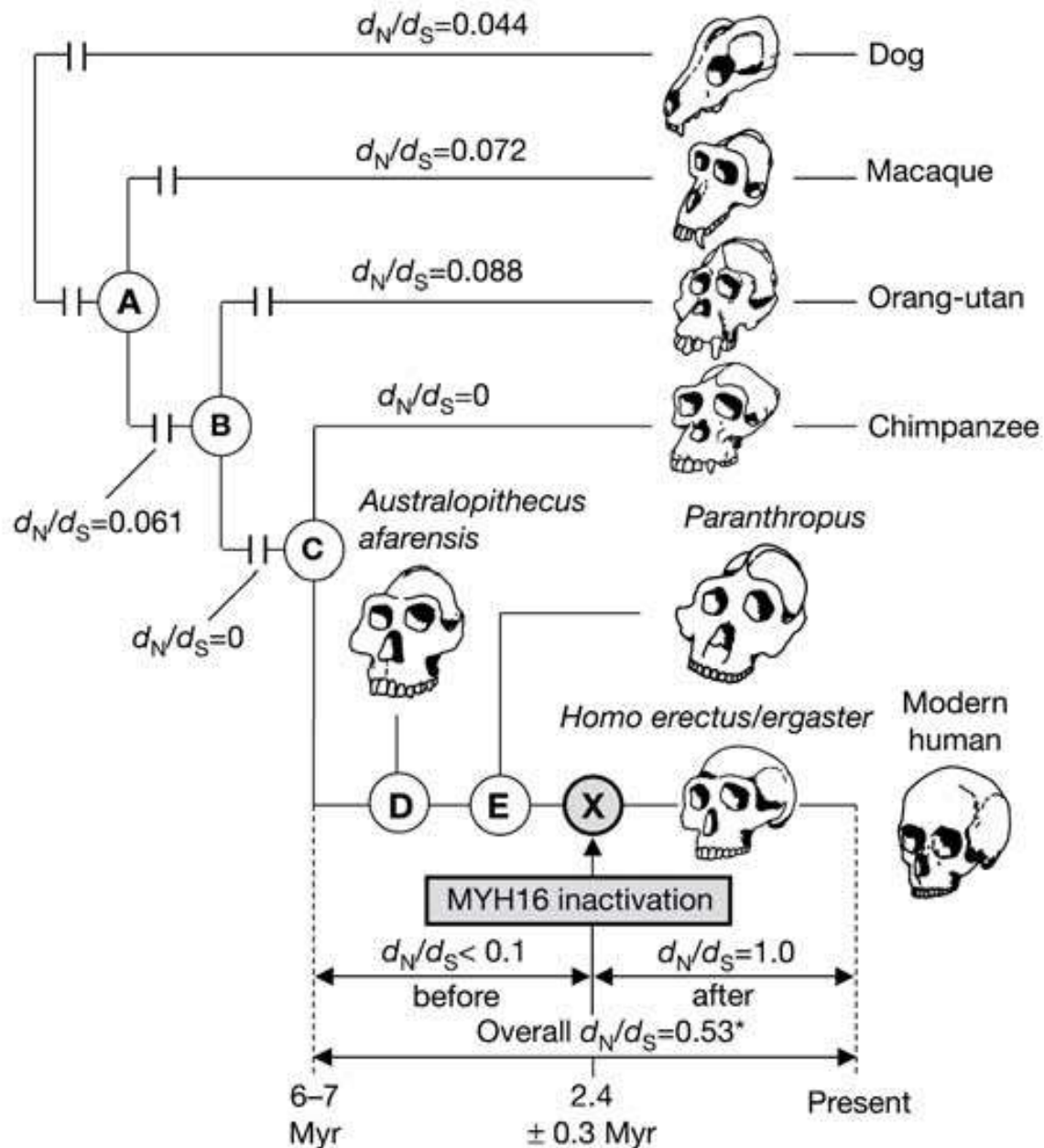


Regulace androgenního receptoru



MYH16 inactivation

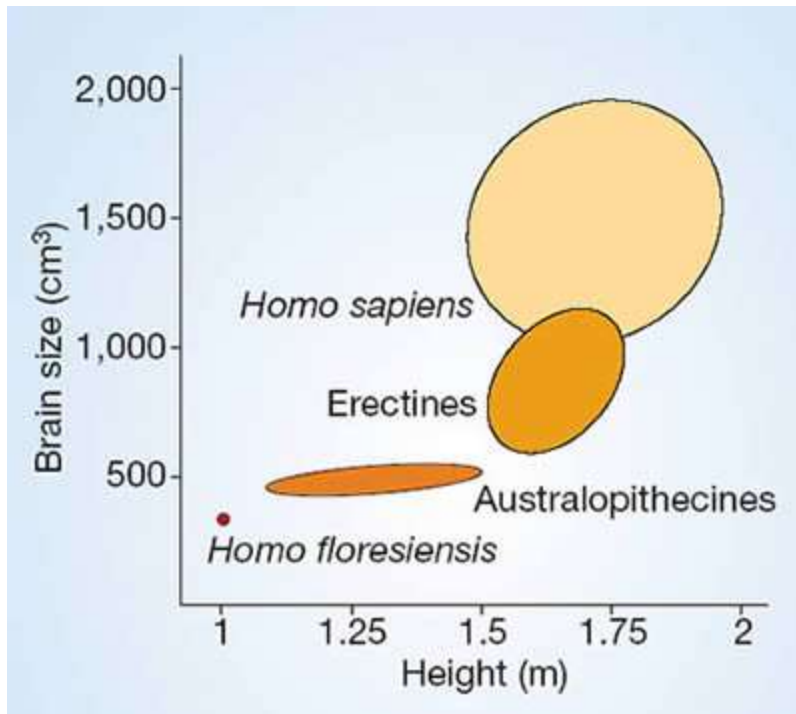




Stedman HH, Kozyak BW, Nelson A, Thesier DM, Su LT, Low DW, Bridges CR, Shrager JB, Minugh-Purvis N, Mitchell MA. Myosin gene mutation correlates with anatomical changes in the human lineage. *Nature*. 2004 Mar 25;428(6981):415-8.

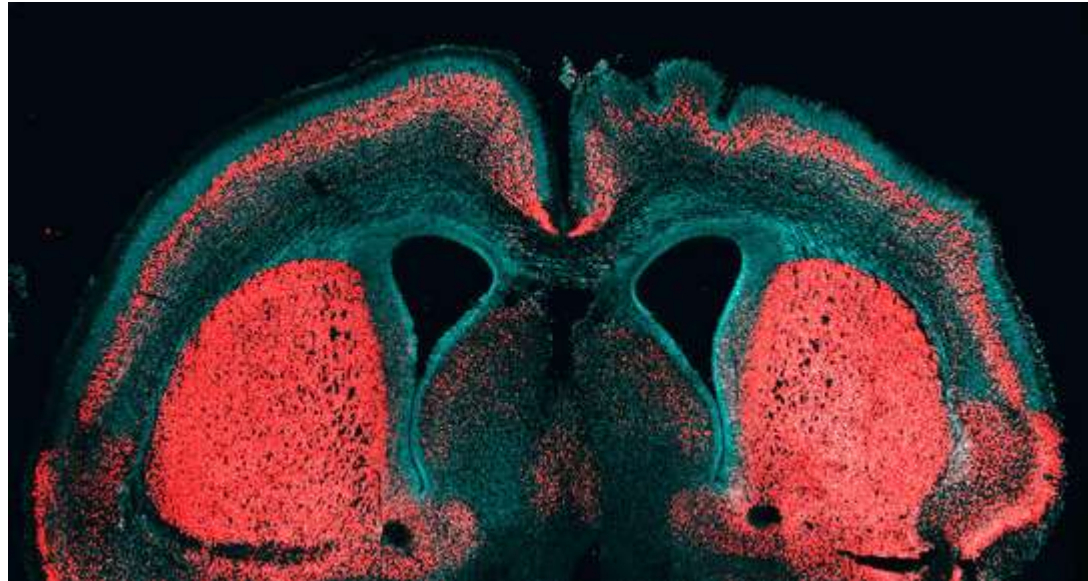
Homo floresiensis

Velikost lebky (mozku)



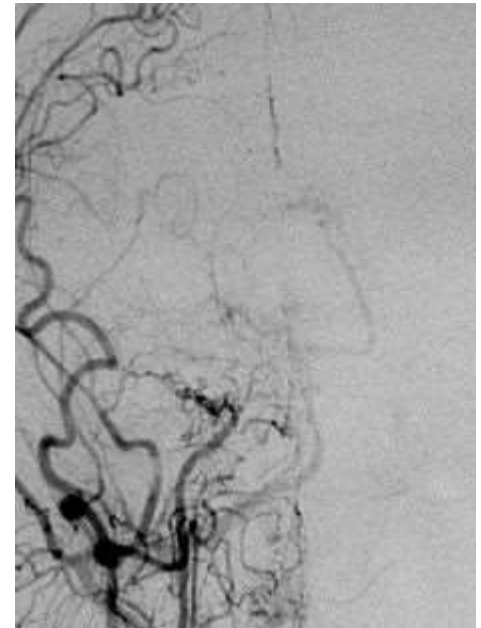
Přeci jen něco navíc

- **ARHGAP11B** (Rho GTPase-activating-protein)
 - Pouze u člověka, ne u šimpanze ani myši
 - Vznikl částečnou duplikací genu ARHGAP11



Pozitivně selektované geny

- **RNF213**
 - **Ring finger protein 213**
 - Mutace způsobuje Moyamoya syndrom
 - Ovlivňuje velikost artérií v mozku.
 - Pozitivní selekce u primátů



kontakt

Jan Pačes

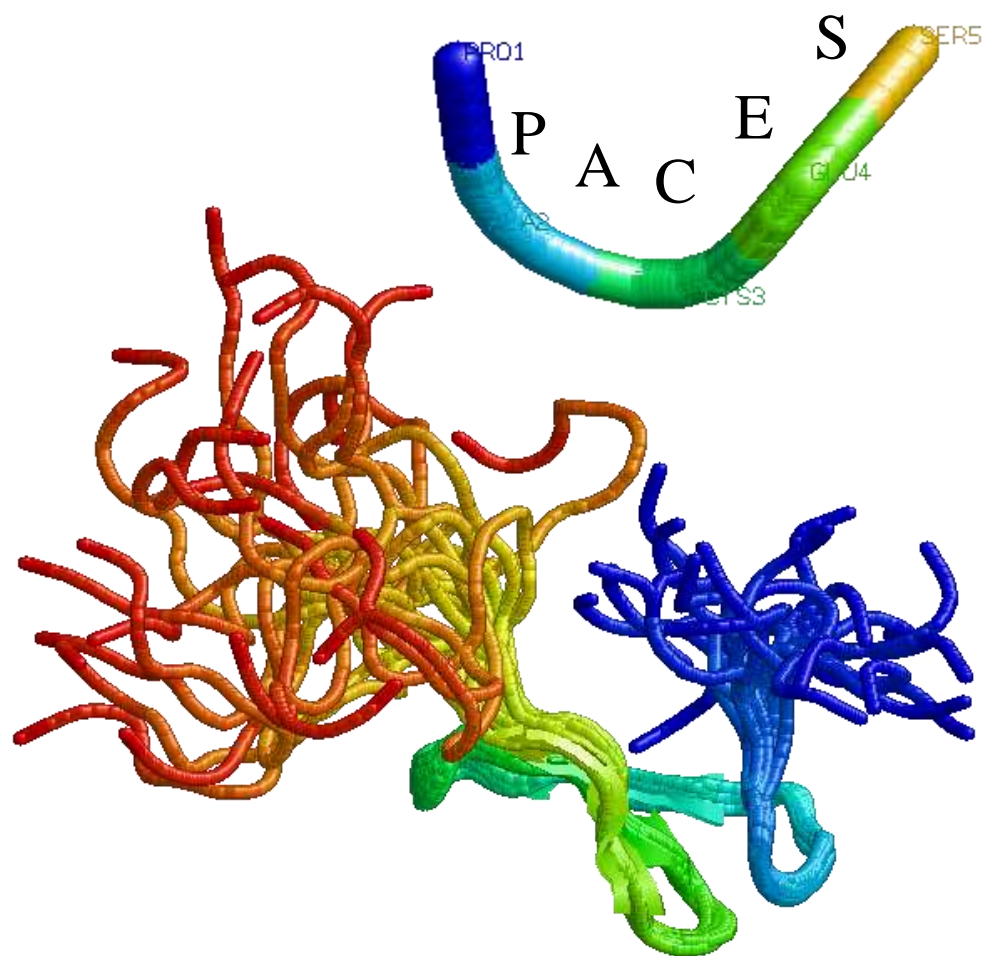
www: <http://bio.img.cas.cz>

email: hpaces@img.cas.cz

icq: #110872370

irc: efnet #hpaces

tel: +420 220183446



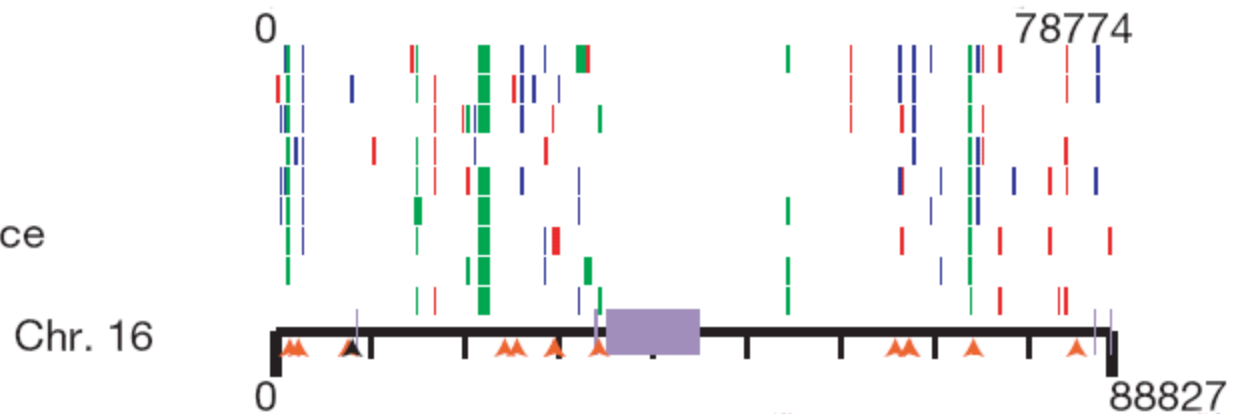


Proč zrovna genom šimpanze?

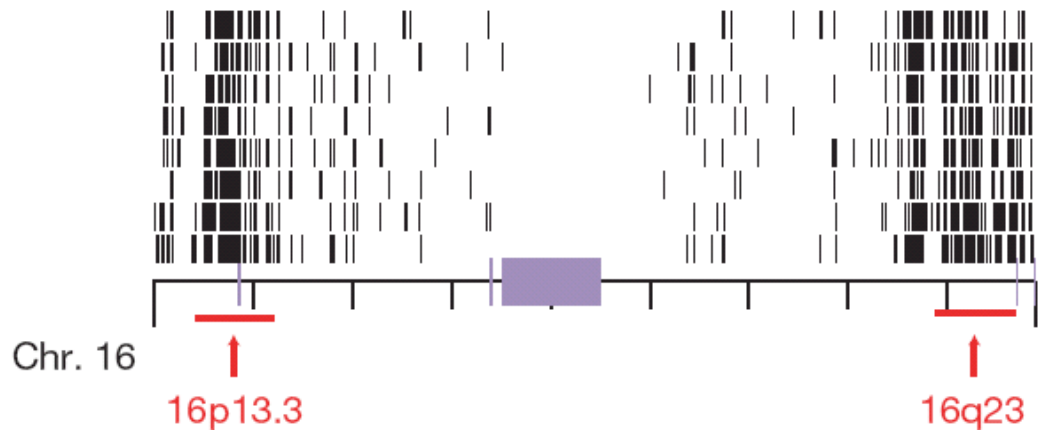
Medical Condition	Humans	Great Apes
<i>Definite</i>		
HIV progression to AIDS	Common	Very rare
Influenza A symptomatology	Moderate to severe	Mild
Hepatitis B/C late complications	Moderate to severe	Mild
<i>P. falciparum</i> malaria	Susceptible	Resistant
Menopause	Universal	Rare
<i>Likely</i>		
<i>E. coli</i> K99 gastroenteritis	Resistant	Sensitive?
Alzheimer's disease pathology	Complete	Incomplete
Coronary atherosclerosis	Common	Uncommon
Epithelial cancers	Common	Rare
<i>Speculative</i>		
Menstrual blood loss	Variable	Lower amount?
Early fetal wastage	High	Low?

8 human genomes comparison

- Deletion
- Insertion
- Inversion
- ▲▲ Novel sequence



SNP frequency



Hledáme rozdíl

kde začít na internetu

enPath

me.ucsc.edu

links

.cas.cz/links

Ensembl

http://www.ensembl.org/Homo_sapiens

n/Entrez/hum_srch

vybrané zdroje dat na internetu

SwissProt <http://www>
Entrez <http://www>

The screenshot shows the NCBI Entrez Nucleotide search interface. The browser window title is "Netscape: Entrez - Nucleotide". The address bar shows the URL: <http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Nucleotide>. The page features a search bar with a dropdown menu set to "Nucleotide" and a "Go" button. Below the search bar, there are links for "Limits", "Preview/Index", "History", and "Clipboard". A yellow highlighted box contains the text: "Entrez Nucleotides, part of the Entrez search and retrieval system, is a collection of nucleotide entries from GenBank. The number of bases in GenBank grows at an exponential rate. Today's total is: 10527520791". Below this, there are two sections: "Decoding the human genome" and "VecScreen". The "Decoding the human genome" section mentions the publication of complete DNA sequences for human chromosomes 22 and 21, along with draft versions for chromosomes 5, 16, and 19. The "VecScreen" section mentions a new vector contamination detection tool. At the bottom, there is a horizontal bar chart titled "Top 10 organisms of the month" showing the number of new nucleotide sequences processed in GenBank for May, 2000. The chart shows human as the top organism with approximately 900 sequences, followed by HIV type 1, pig, house mouse, Drosophila simulans, Botryllus schlosseri (sea squirts), rhesus monkey, Theileria annulata, Norway rat, and Brachionus plicatilis (rotifers).

10527520791

Decoding the human genome

Thanks to a multinational effort, the complete DNA sequences for human chromosomes 22 and 21 have been published, along with draft versions for chromosomes 5, 16 and 19.

VecScreen

NCBI now has a vector contamination detection tool. [VecScreen](#) your sequence.

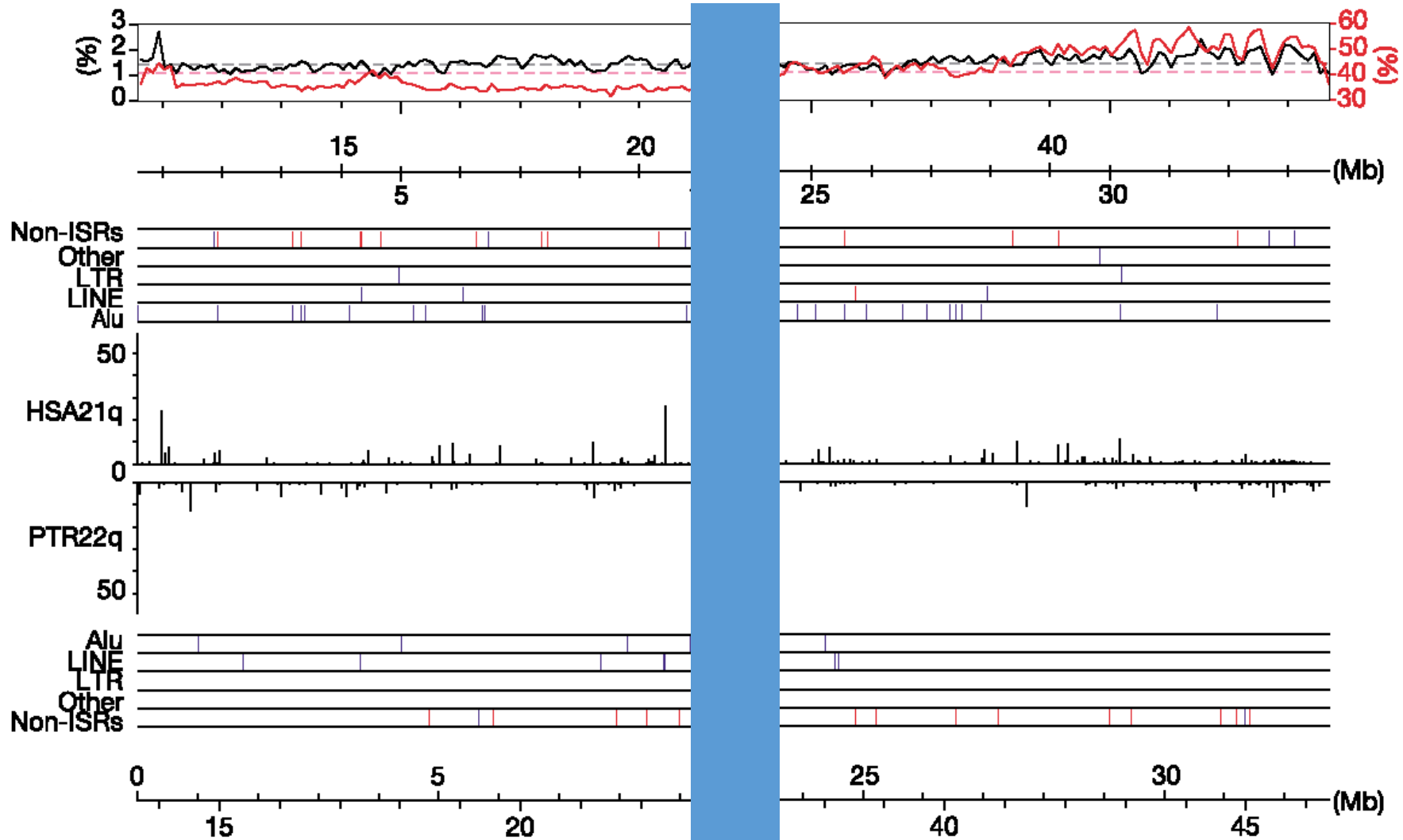
Top 10 organisms of the month

Organism	Number of new nucleotide sequences
human	~900
HIV type 1	~750
pig	~450
house mouse	~350
Drosophila simulans	~250
Botryllus schlosseri (sea squirts)	~200
rhesus monkey	~150
Theileria annulata	~100
Norway rat	~80
Brachionus plicatilis (rotifers)	~50

Number of new nucleotide sequences processed in GenBank for May, 2000.

CZEFCH
FOBIA

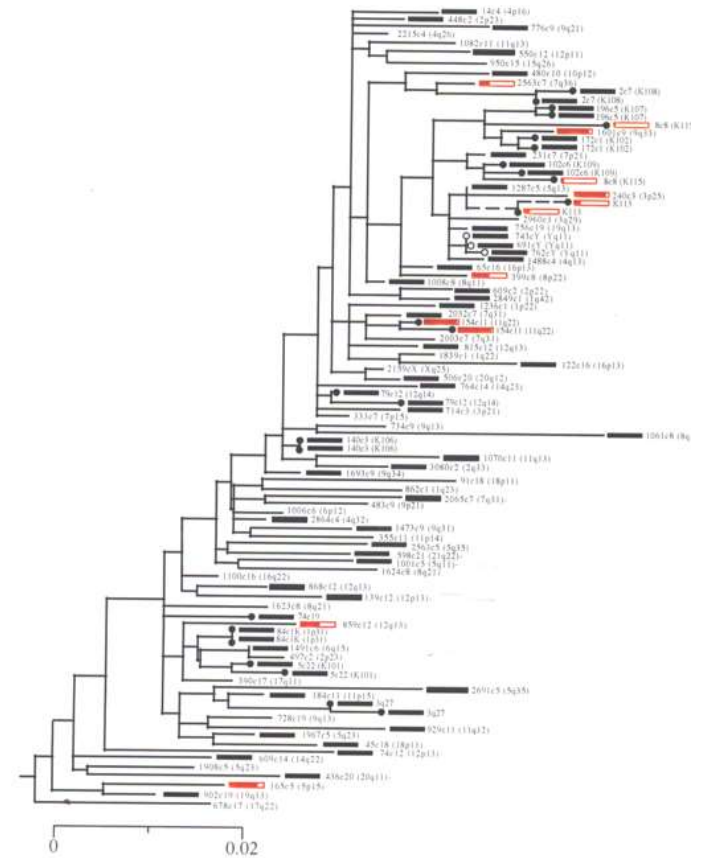
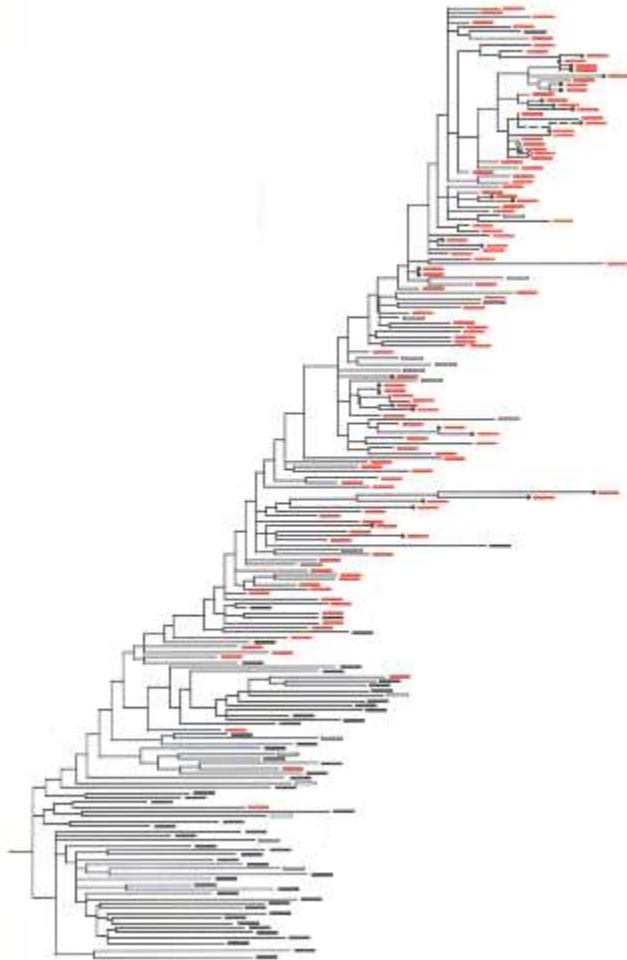
hs ch21 vs pt ch22



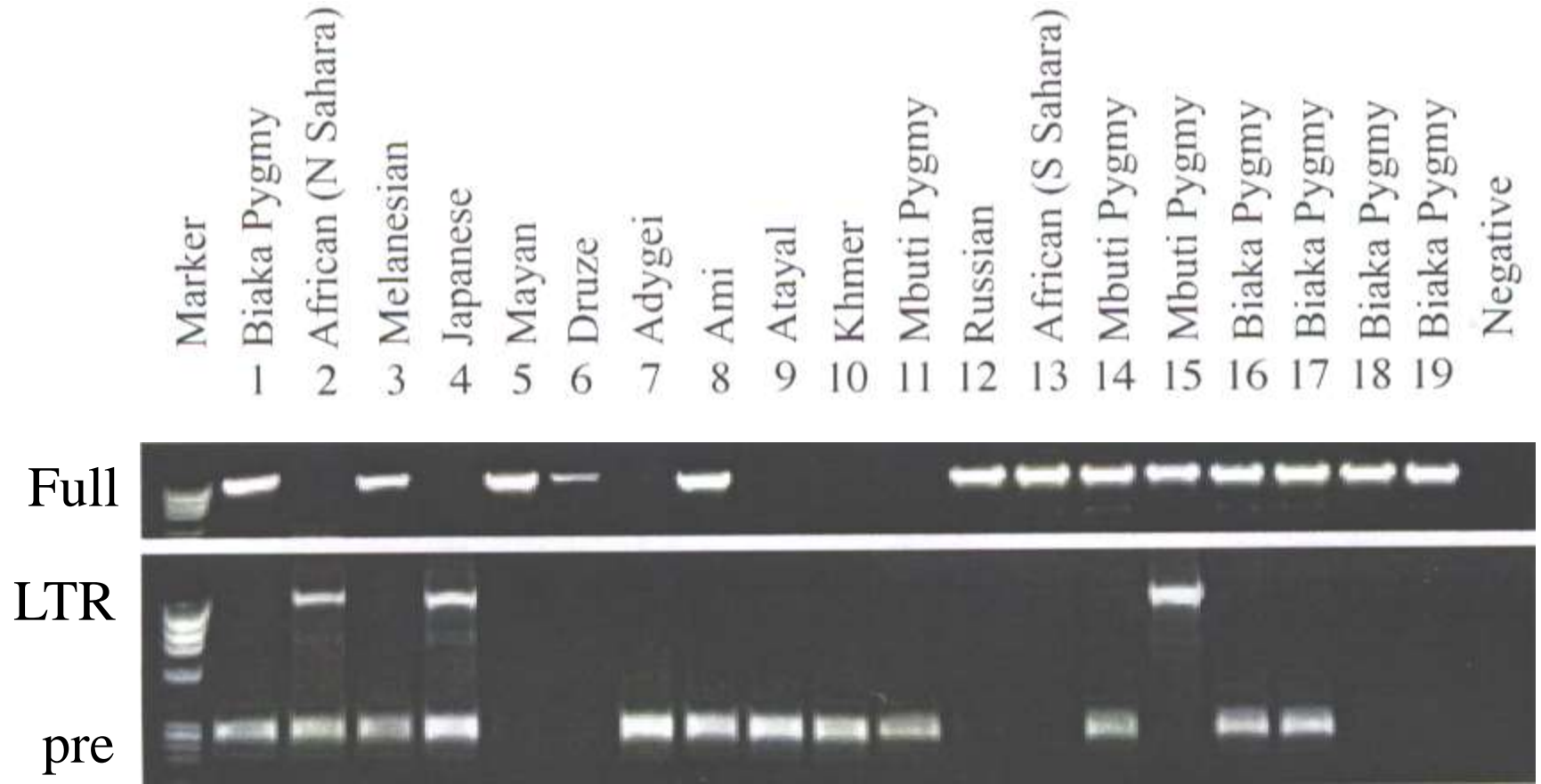
human vs chimp ERV analýza



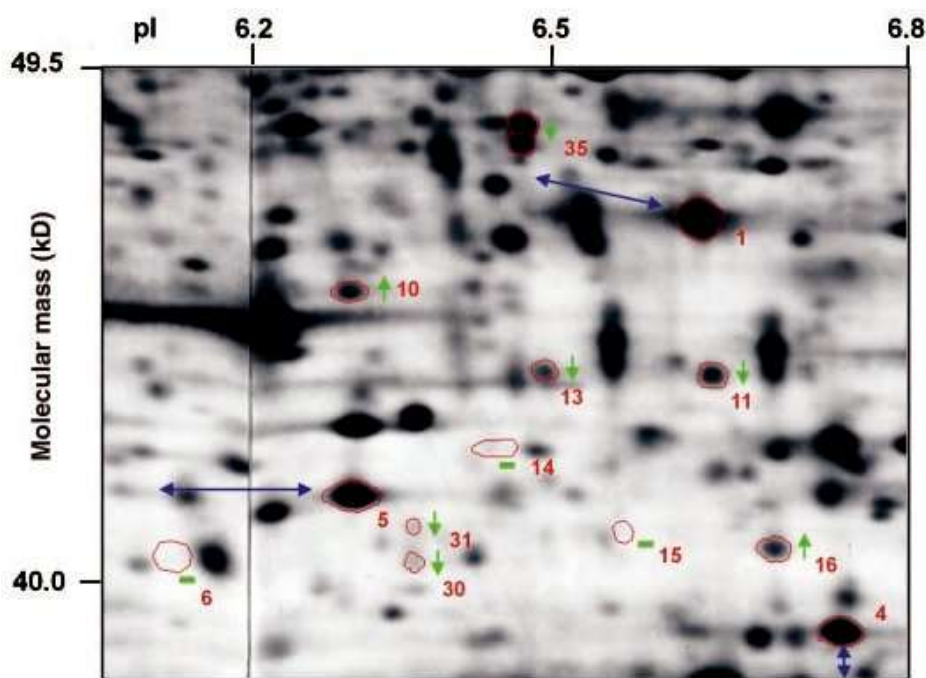
HERVK (HML2)



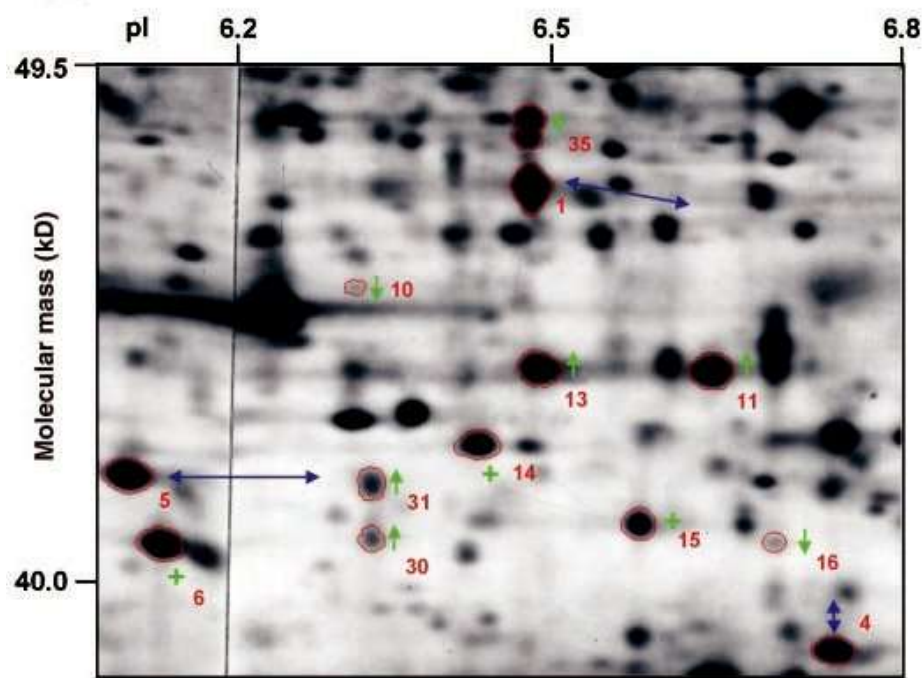
Polymorfní herv 259c12



proteom



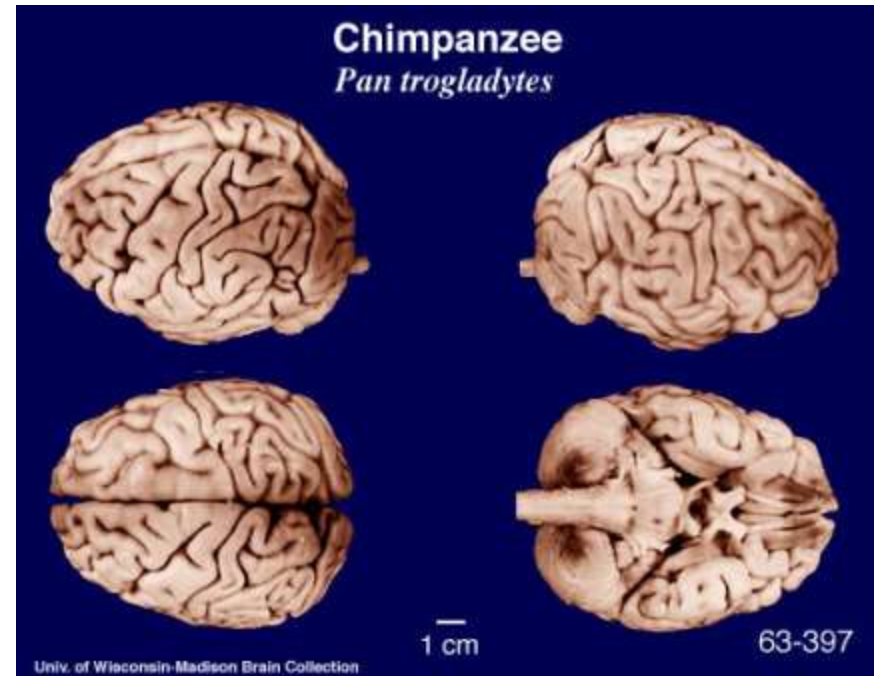
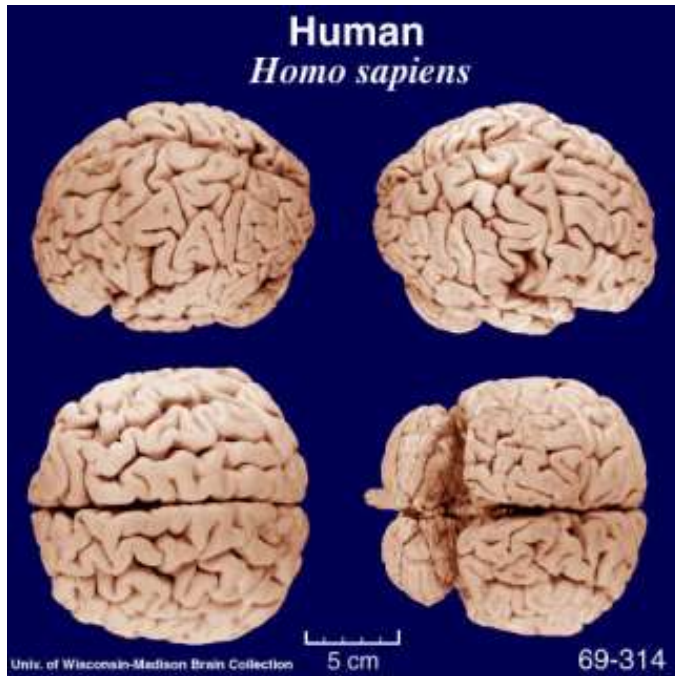
člověk



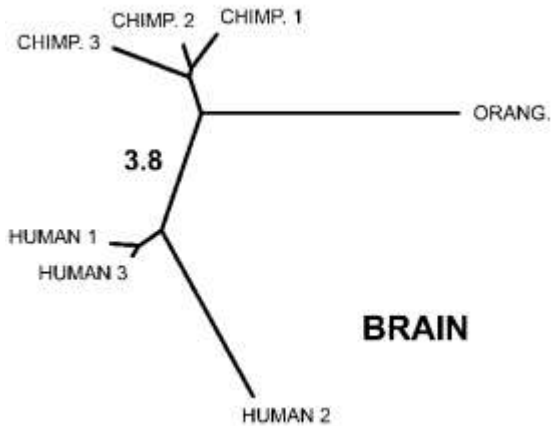
šimpanz

Enard W, Khaitovich P, Klose J, Zollner S, Heissig F, Giavalisco P, Nieselt-Struwe K, Muchmore E, Varki A, Ravid R, Doxiadis GM, Bontrop RE, Paabo S. Intra- and interspecific variation in primate gene expression patterns. *Science*. 2002 Apr 12;296(5566):340-3.

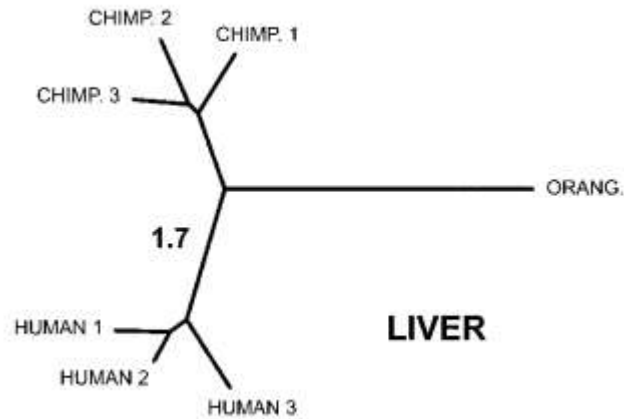
mozek



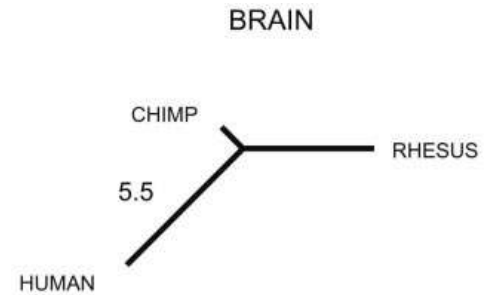
akcelerace vývoje mozku



BRAIN



LIVER



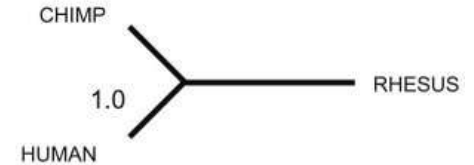
BRAIN

HUMAN

5.5
CHIMP

RHESUS

BLOOD

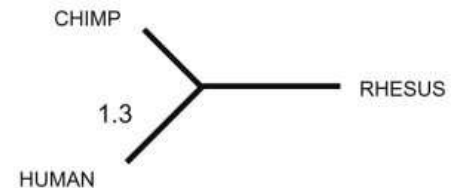


HUMAN

1.0
CHIMP

RHESUS

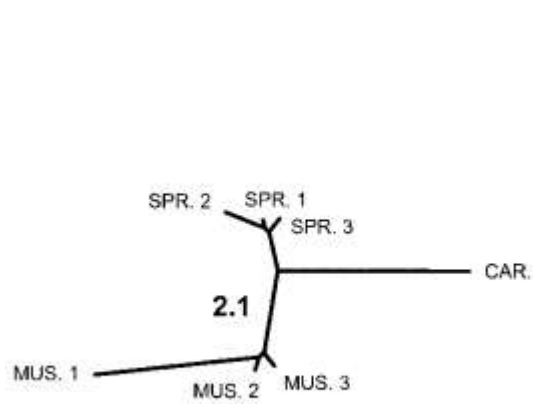
LIVER



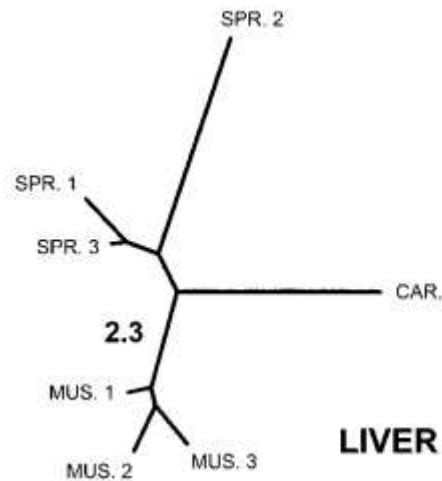
HUMAN

1.3
CHIMP

RHESUS

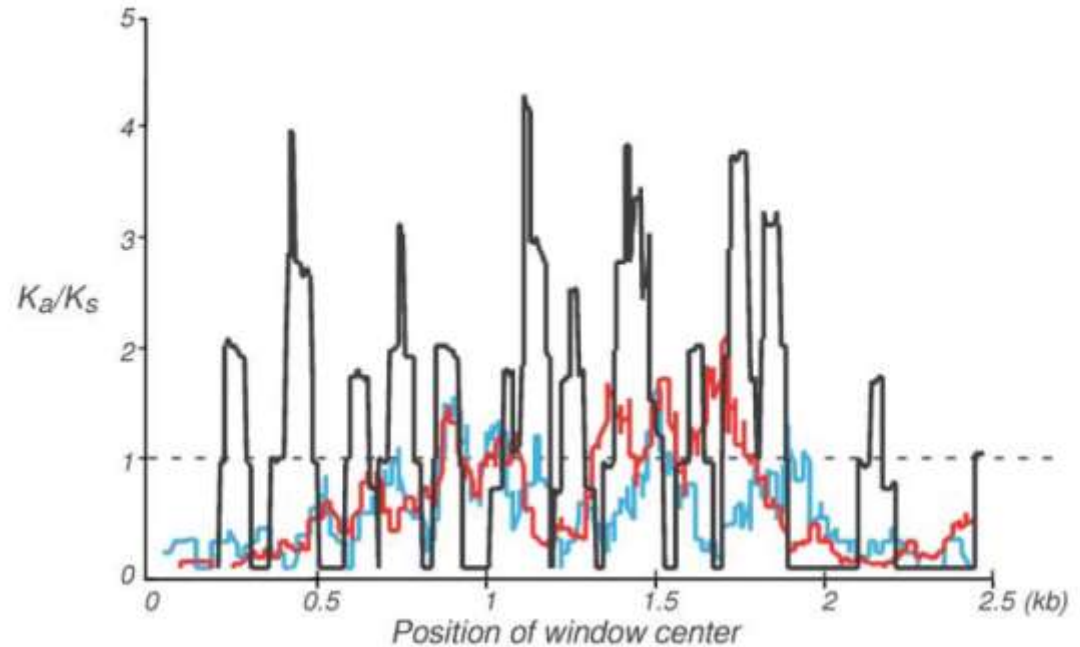


BRAIN

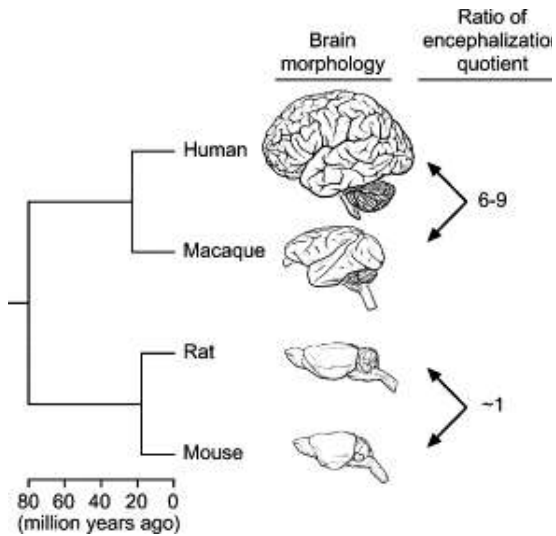


LIVER

microcephalin



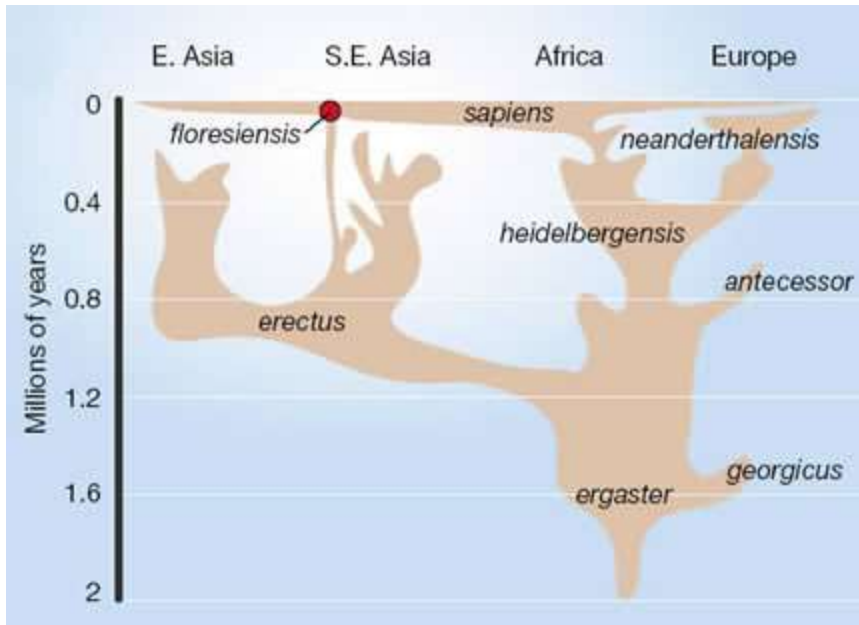
- Catarrhine ancestors to humans
- Carnivores (dog vs. cat)
- Rodents (rat vs. mouse)



pozitivně selektované typy genů

host – patogen interakce (MHC, CMAH)
reprodukce
adaptace na potravu
vzhled (MC1R)
smysly (čich, sluch)
chování
mozek

Homo floresiensis



H. floresiensis was part of the Asian dispersals of the descendants of *H. ergaster* and *H. erectus*.

Homo sapiens

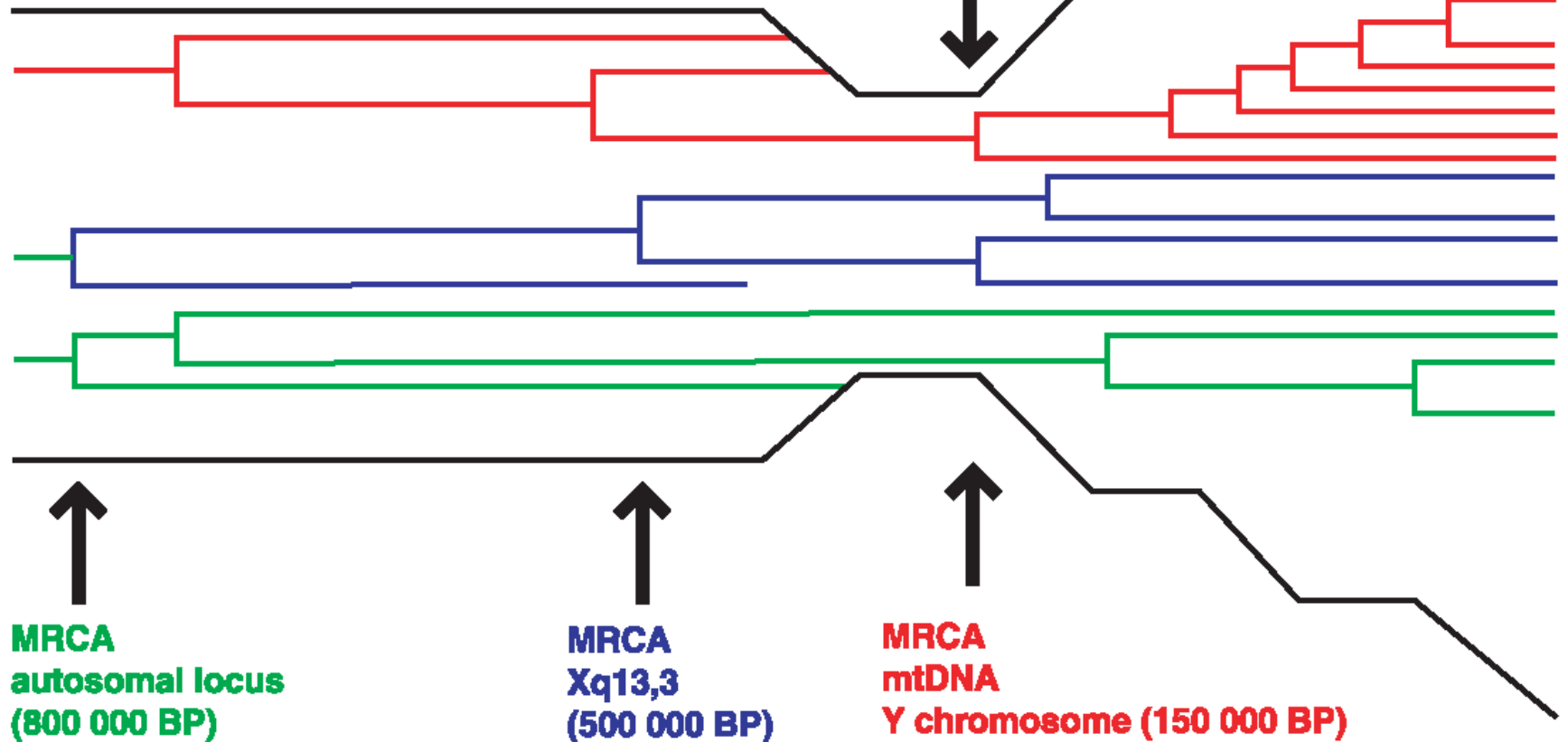
Present:

10 000 BP:
"Neolithic
expansion"

50 000 BP:
"mtDNA
expansion"

100 000–200 000 BP:
"Xq13,3
expansion"

Homo erectus



Cizorodé elementy

DNA, která se do genomu dostala jinak než vertikálním transferem, tj z předků na potomky

principy detekce:

- „cizí“ DNA je odlišná od průměrné „vlastní“ DNA
- experimentální data

Repetitivní elementy

- Transpozóny:** transposon-derived repeats, interspersed repeats; 45% genomu
- Mikro a minisatelitey:** simple sequence repeats, opakování krátkých přímých repetitivních; 3% genomu
- Duplikace:** duplikace různě dlouhých (10-300 kb) genomových segmentů - inter i intrachromosomové; 3,3% genomu
- Jiné typy repetitivních:** centromerické a telomerické repeaty

DNA transpozóny



2-3 kb

terminální reverzní repetice (50 - 100 b)

cut-and-paste mechanismus kopírování

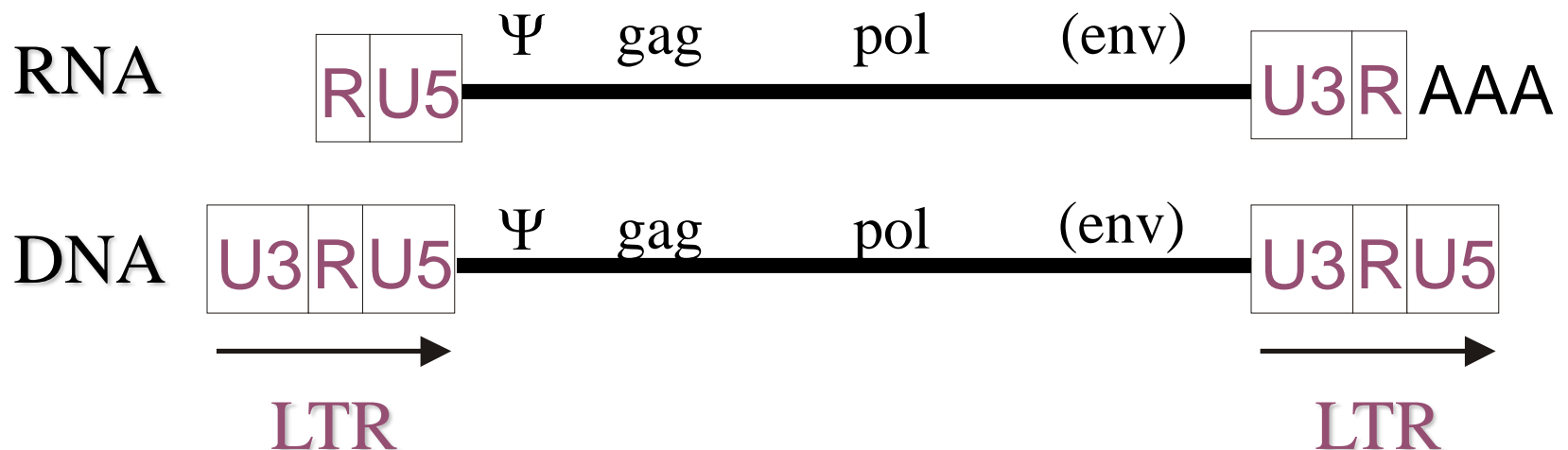
3% genomu

minimálně 7 tříd, které nejsou (blízce) příbuzné

LTR retrotranspozóny

HERV:

1. 6 - 8 % lidského genomu
2. 100 000 elementů
3. desítky rodin



Transpozóny

DNA transpozóny

retrotranspozóny

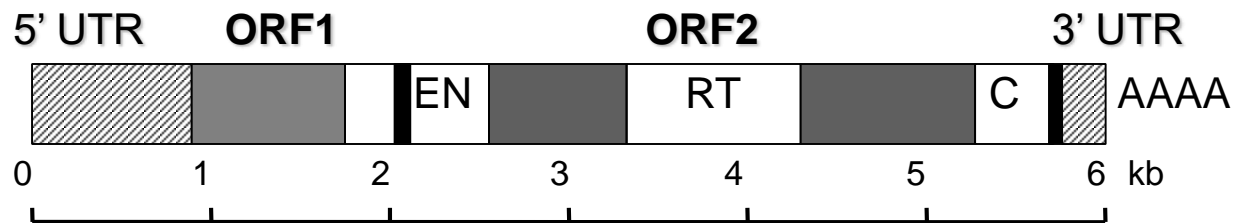
(RNA intermediát, reverzní transkripce)

LTR transpozóny (podobné retrovirům)

polyA (non LTR) retrotranspozóny

(kolineární s mRNA, mají polyA)

non-LTR (LINE1 nebo L1 elementy)



LINE – long interspersed elements

poly A (non-LTR) retrotranspozóny

RNA intermediát (interní promotor pro RNA pol. II); polyA

krátká inzerční duplikace (5-15 bp)

inzerční preference (TT|AAAA)

17 % genomu

500 000 elementů, často zkrácených na 5' konci

30-60 aktivních LINE1 elementů v genomu

Neautonomní elementy

nekódují enzymy pro svou vlastní transpozici

pro každou třídu autonomních elementů existuje neautonomní element, který používá mechanismus replikace „svého“ autonomního elementu

DNA transpozóny

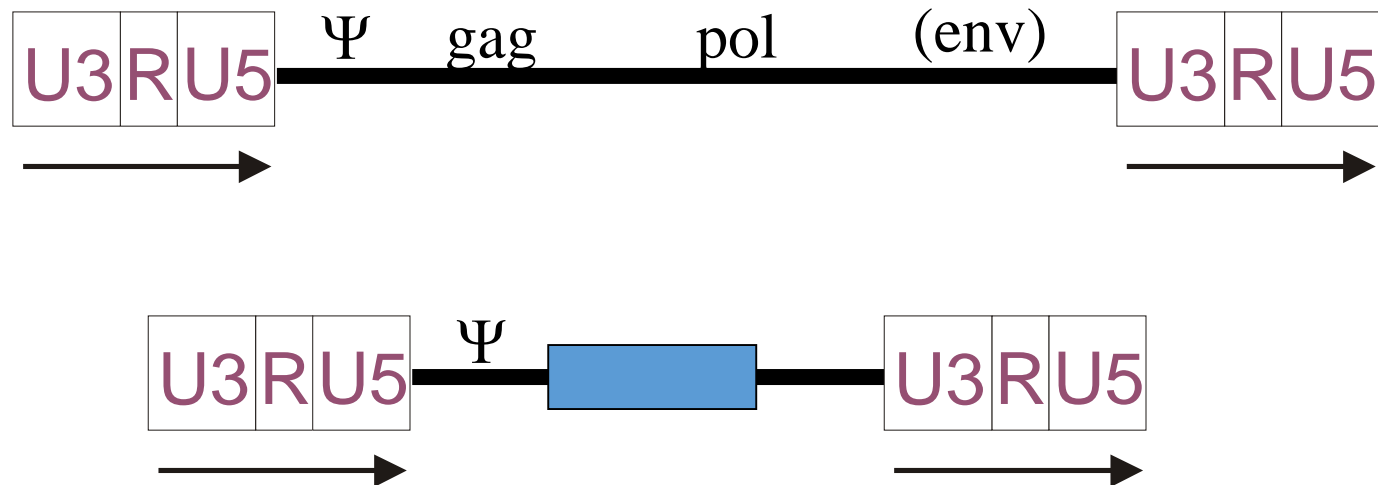


2-3 kb; terminální reverzní repetice



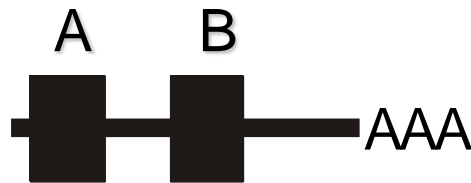
80-3000 bp; terminální reverzní repetice

Lidské endogenní retroviry (HERVs)



LTR; krátké inzerční duplikace; primer binding site

SINE (Alu) elements



SINE – short interspersed elements

poly A (non-LTR) retrotranspozóny

interní promotor pro RNA pol. III; polyA

inzerční duplikace (5-15 bp)

inzerční preference (TT|AAAA)

10 % genomu

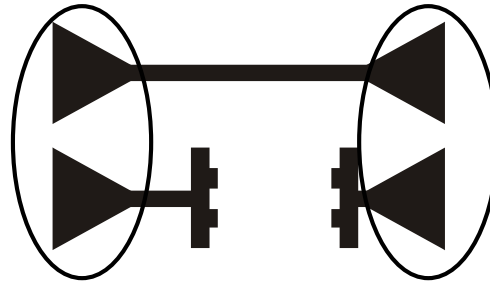
1 000 000 elementů, často zkrácených na 5' konci

Procesované pseudogeny

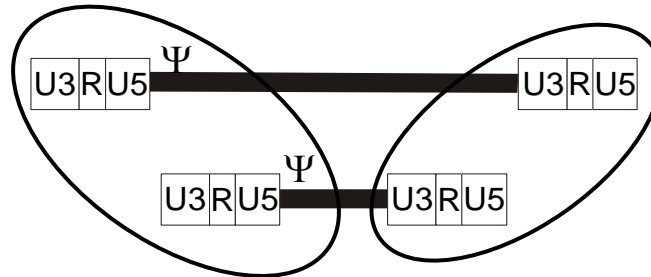
kolineární s mRNA, chybějí introny a promotory
poly A
často zkrácené na 5' konci
krátké inzerční duplikace

Koevoluce parazitů

DNA



LTR



polyA



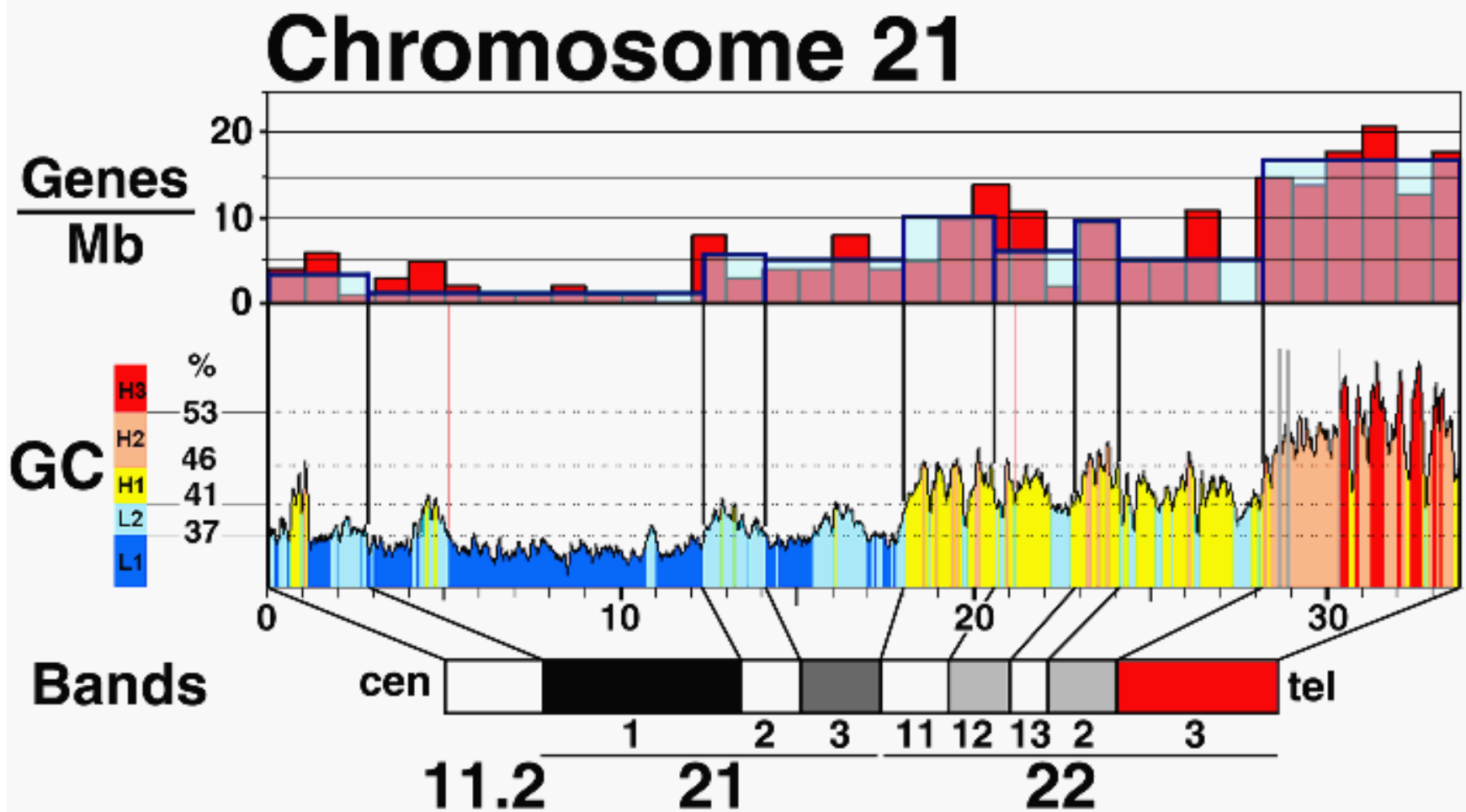
čím to začalo?

International Human Genome Sequencing

Consortium: Initial sequencing and analysis of the human genome. *Nature*. 2001 Feb 15; 409 (6822): 860-921.

Celera: The Sequence of the Human Genome. *Science*. 2001 Feb 16; 291 (5507): 1304-1351.

rozmístění genů



CZECH

FOBIA