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

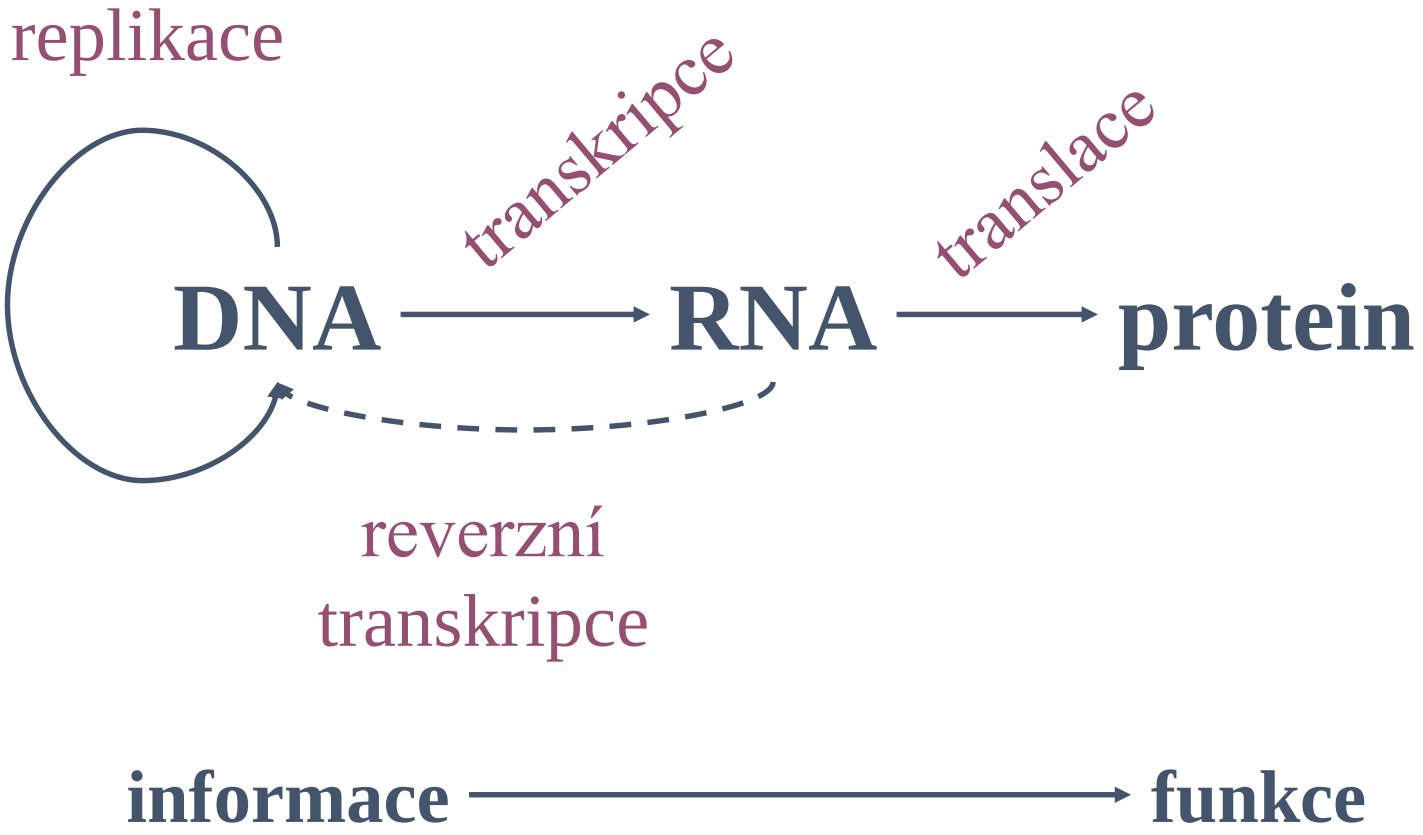
DNA zde není proto, aby sloužila organizmům, naopak, organizmy jsou zde proto, aby sloužily DNA

Richard Dawkins

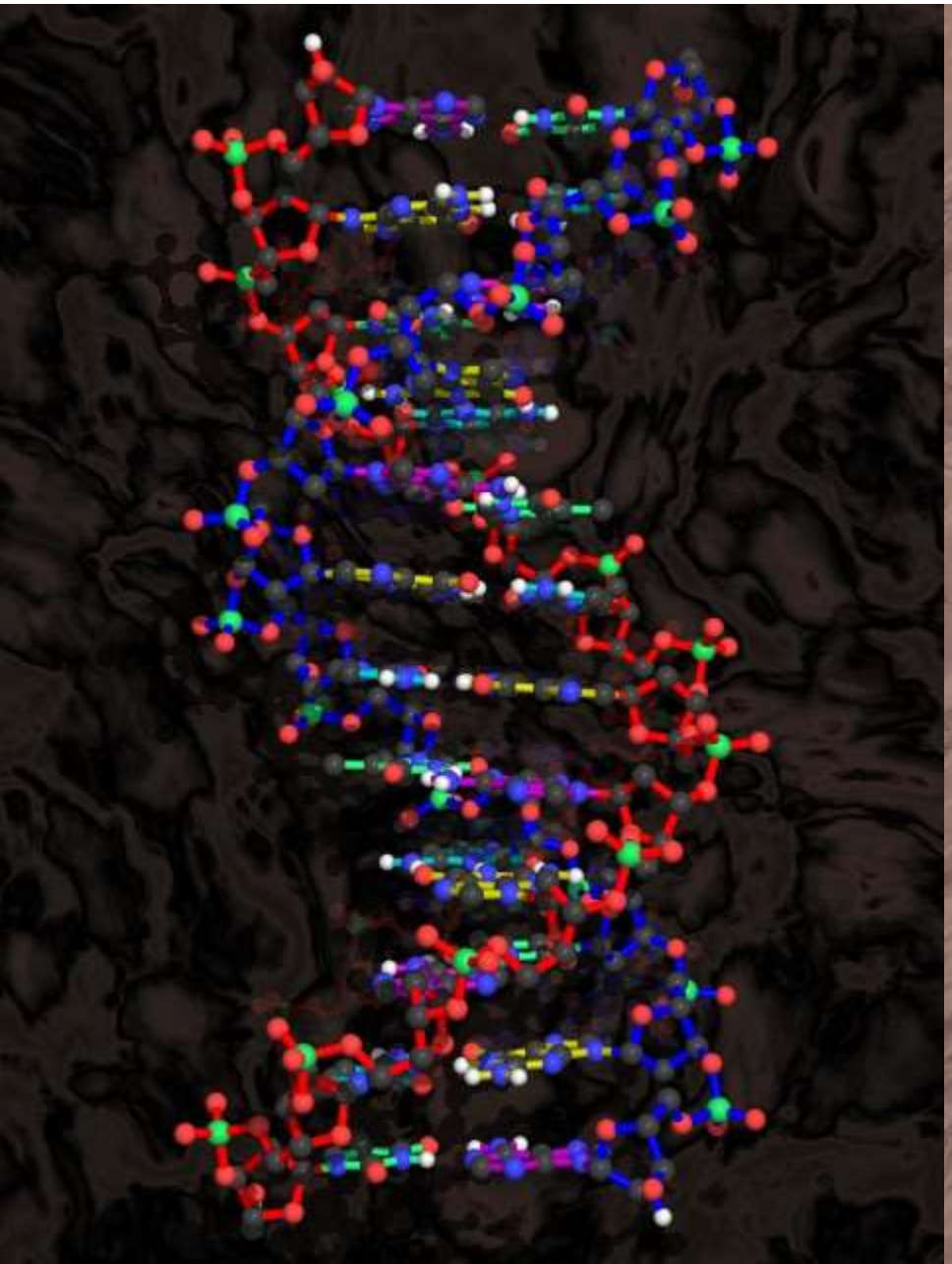
Komparativní genetika se zabývá podobnostmi. Ale v případě genomu šimpanze hledáme především rozdíly.

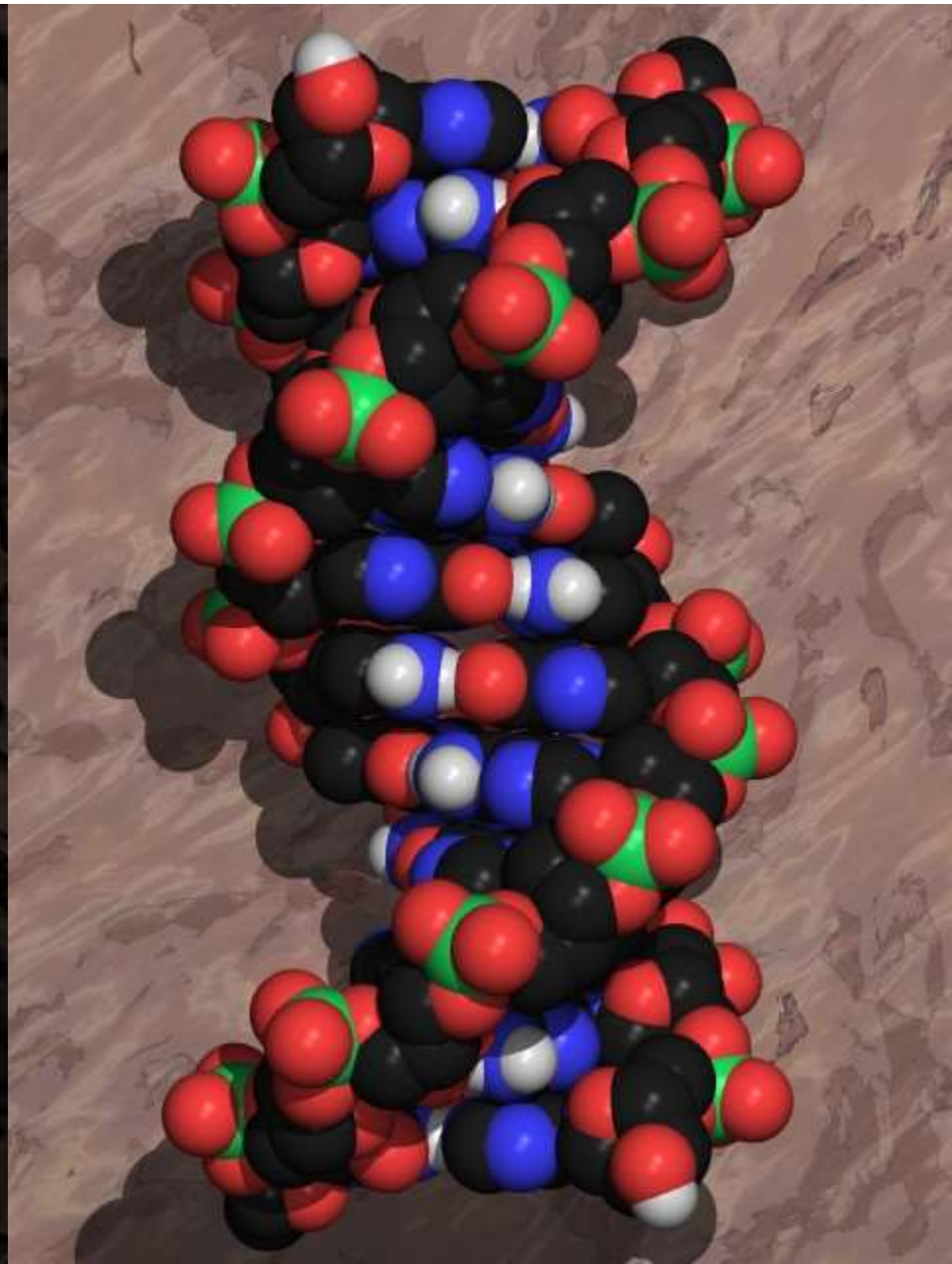
Svante Pääbo

Centrální dogma molekulární genetiky

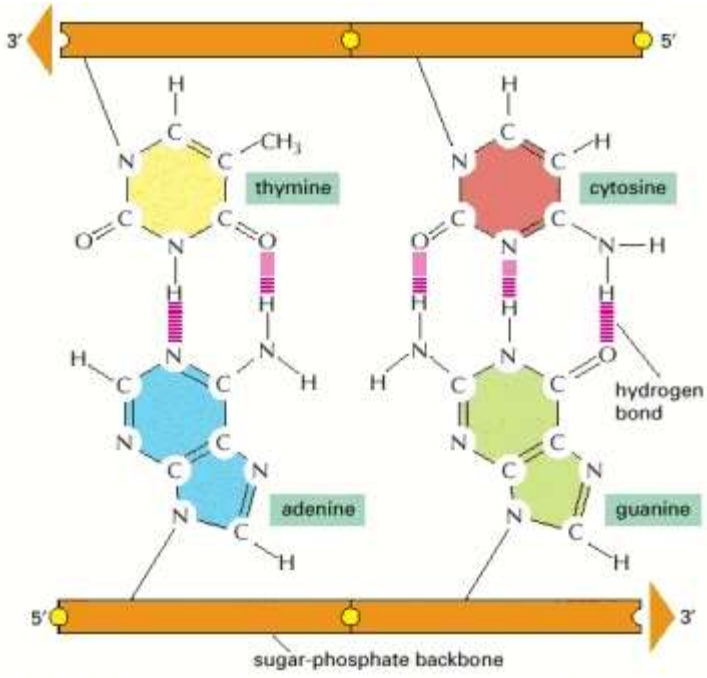


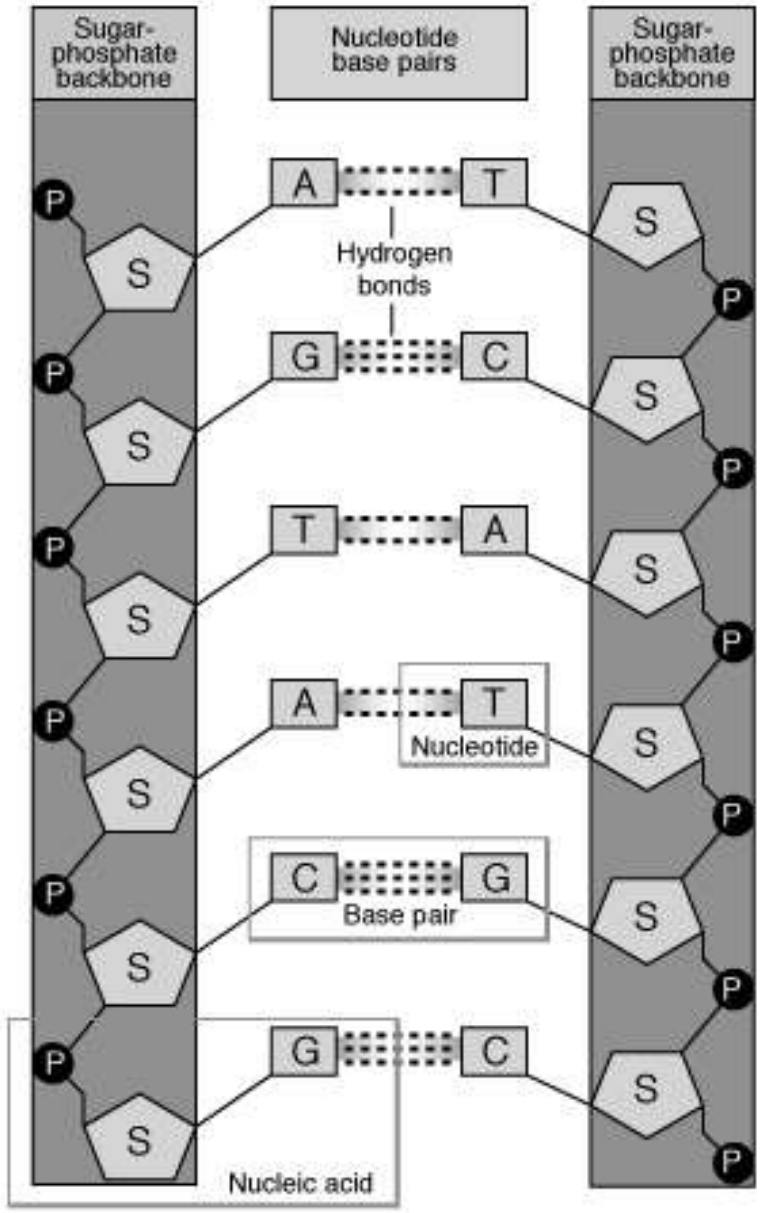
DNA





DNA


3′
5′
thymine
cytosine
hydrogen bond
adenine
guanine
5′
3′
sugar-phosphate backbone


Sugar-phosphate backbone
Nucleotide base pairs
Sugar-phosphate backbone
Hydrogen bonds
Nucleotide
Base pair
Nucleic acid

Transkripce a translace

DNA **5′> ATGAAGCCGAGTCAT 3′**
3′ TACTTCGGCTCAGTA <5′

transkripce

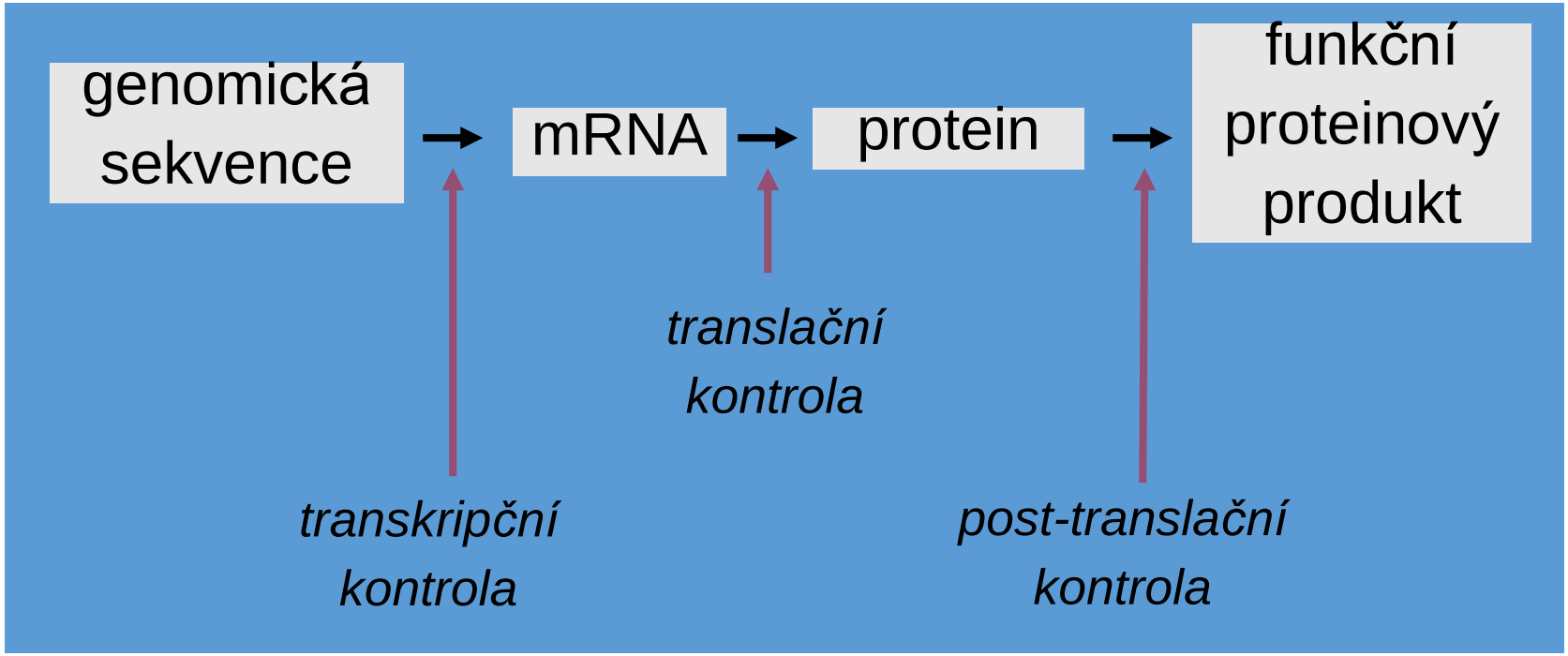
mRNA **5′> AUGAAGCCGACUGAT 3′**

translace

Protein **N> MetLysProSerVal C**

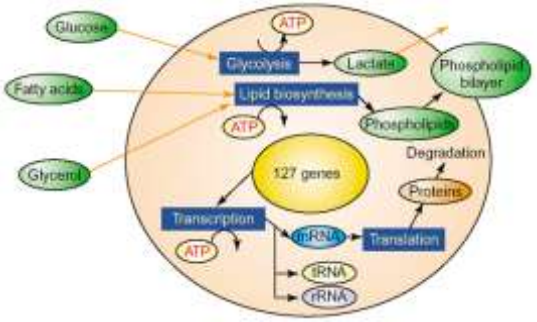
Úrovně kontroly

počet genů ≠ úroveň mRNA ≠ úroveň genové exprese ≠ množství a efektivní účinnost proteinu



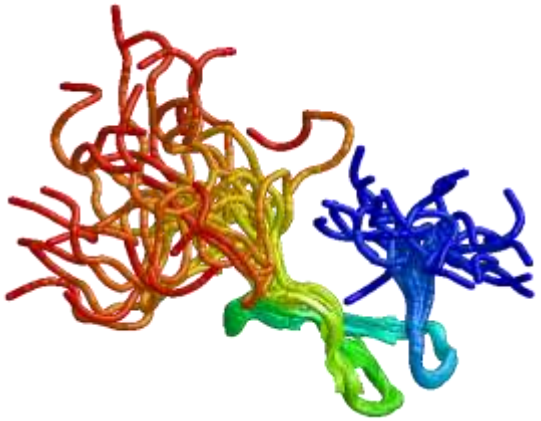
Jak dobře dnes rozumíme DNA

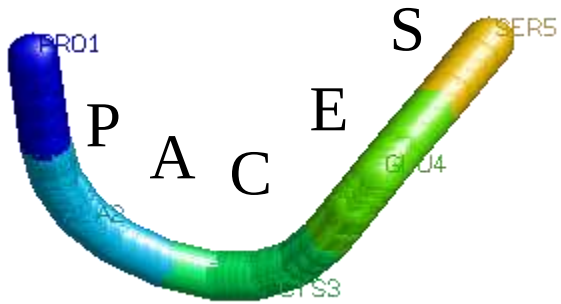
- E-cell



- 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

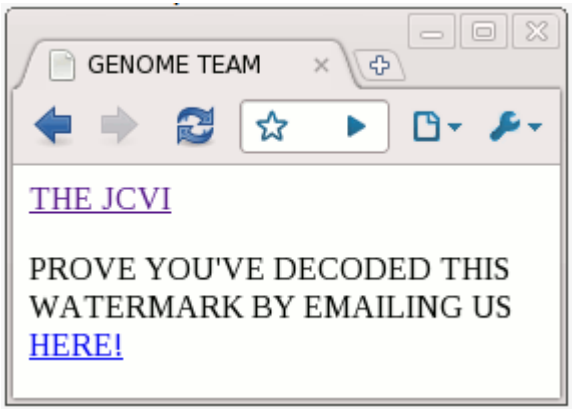
Ukládání informace v DNA



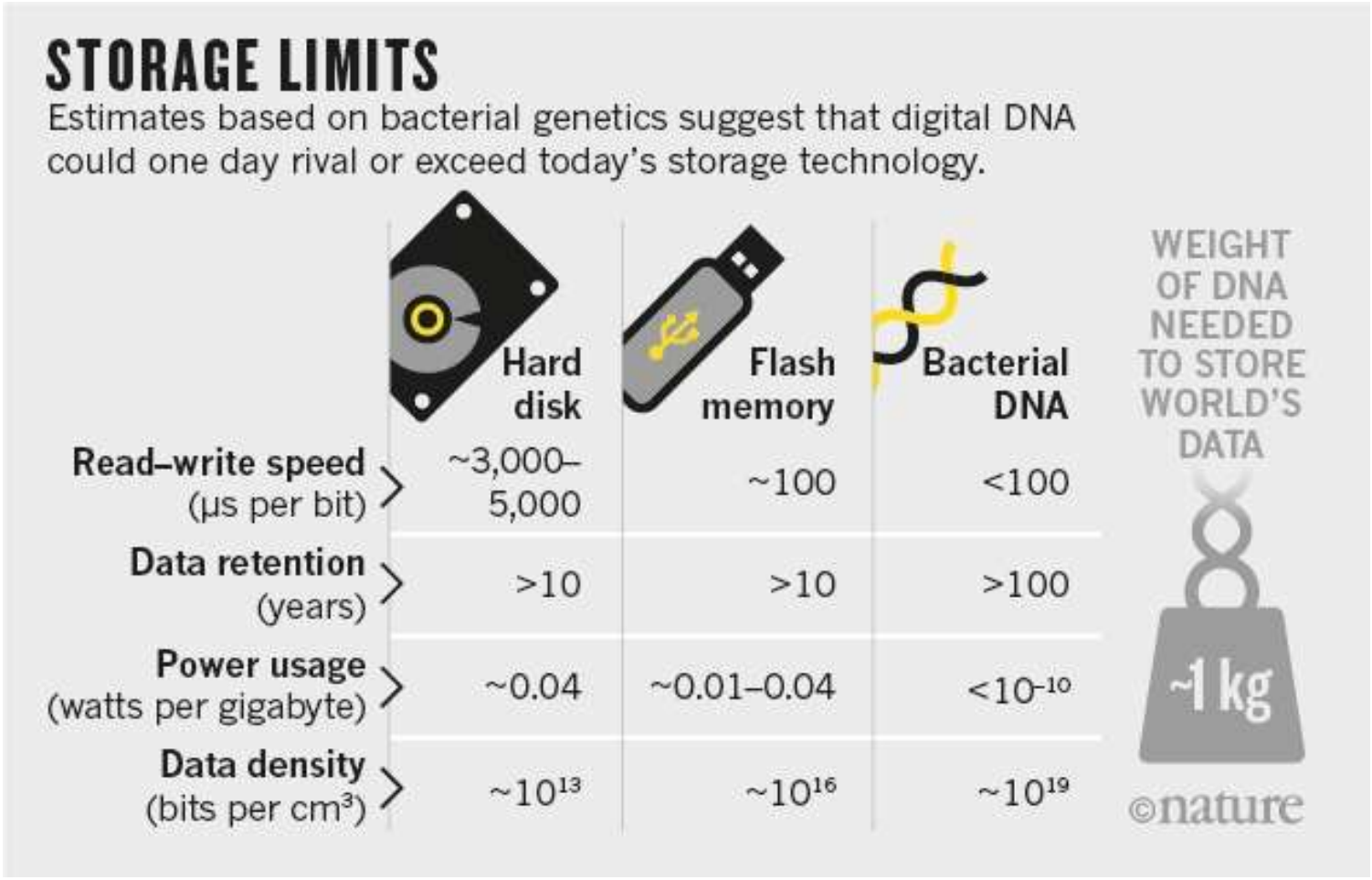


Watermarks:

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



Ukládání informace v DNA



Zakódování Shakespearových sonetů do DNA

Thou art more lovely …

text do ASCI

01010101010001110001010 01 …

ASCII do „trits“ (0,1,2)

20112 20200 02110 10002 …

„trits“ do DNA

aby se každé písmenko lišilo od předcházejícího

TAGAT GTGTA CAGAC TAGCG …

DNA fragmenty

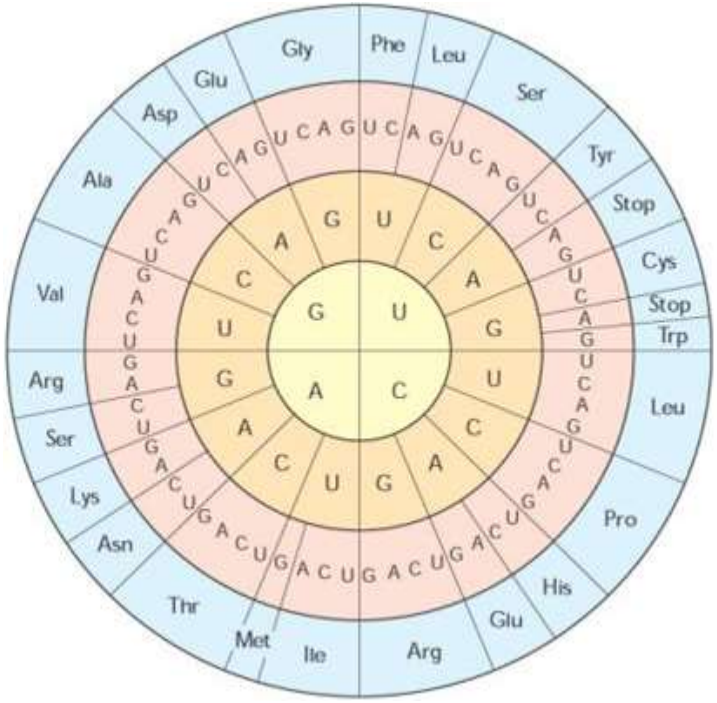
překrývající se a s unikátním indexem

Jak sledovat (měřit) evoluci

Ka/Ks (dn/ds) ratio:

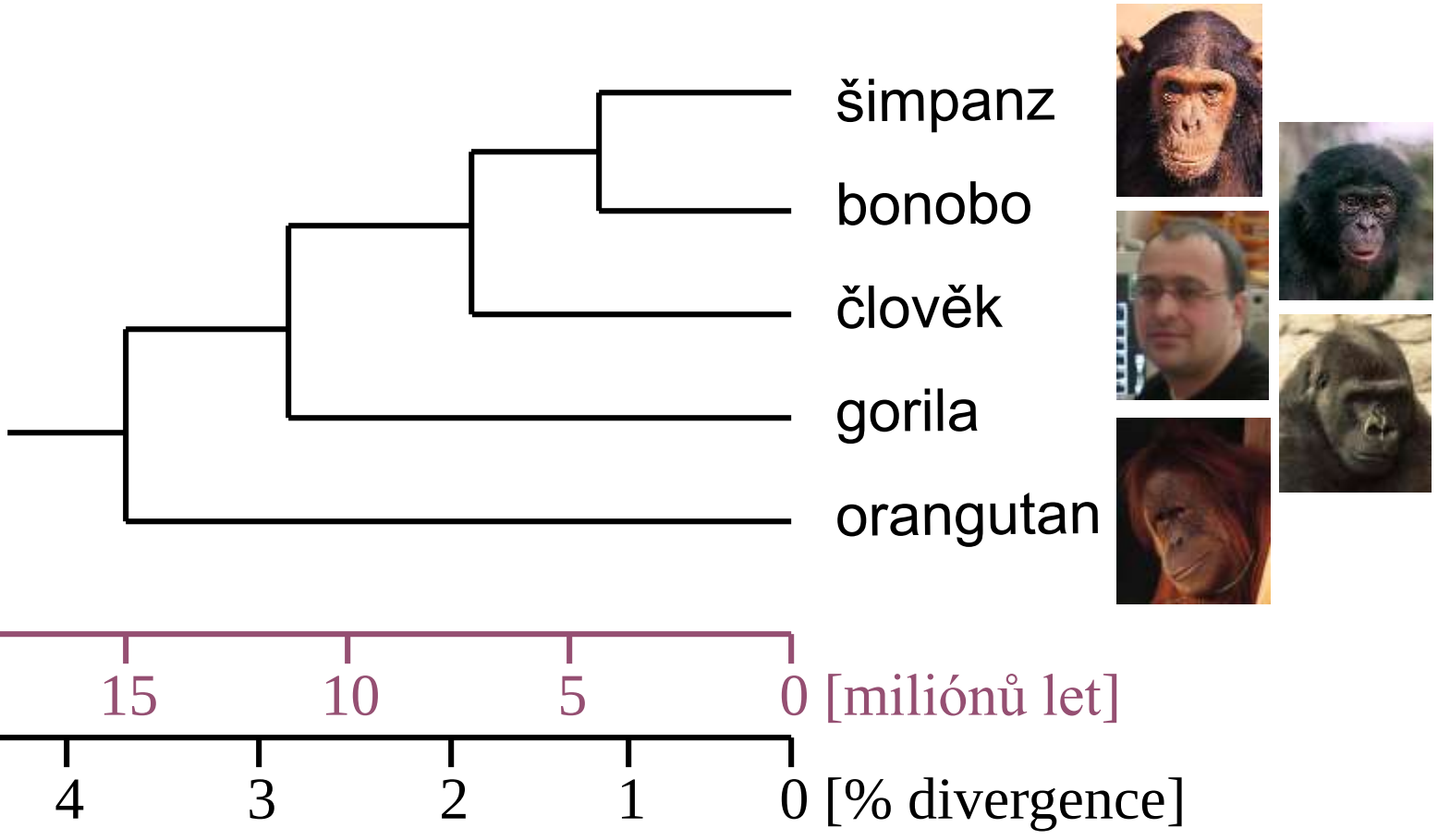
poměr nesynonymních
a synonymních substitucí

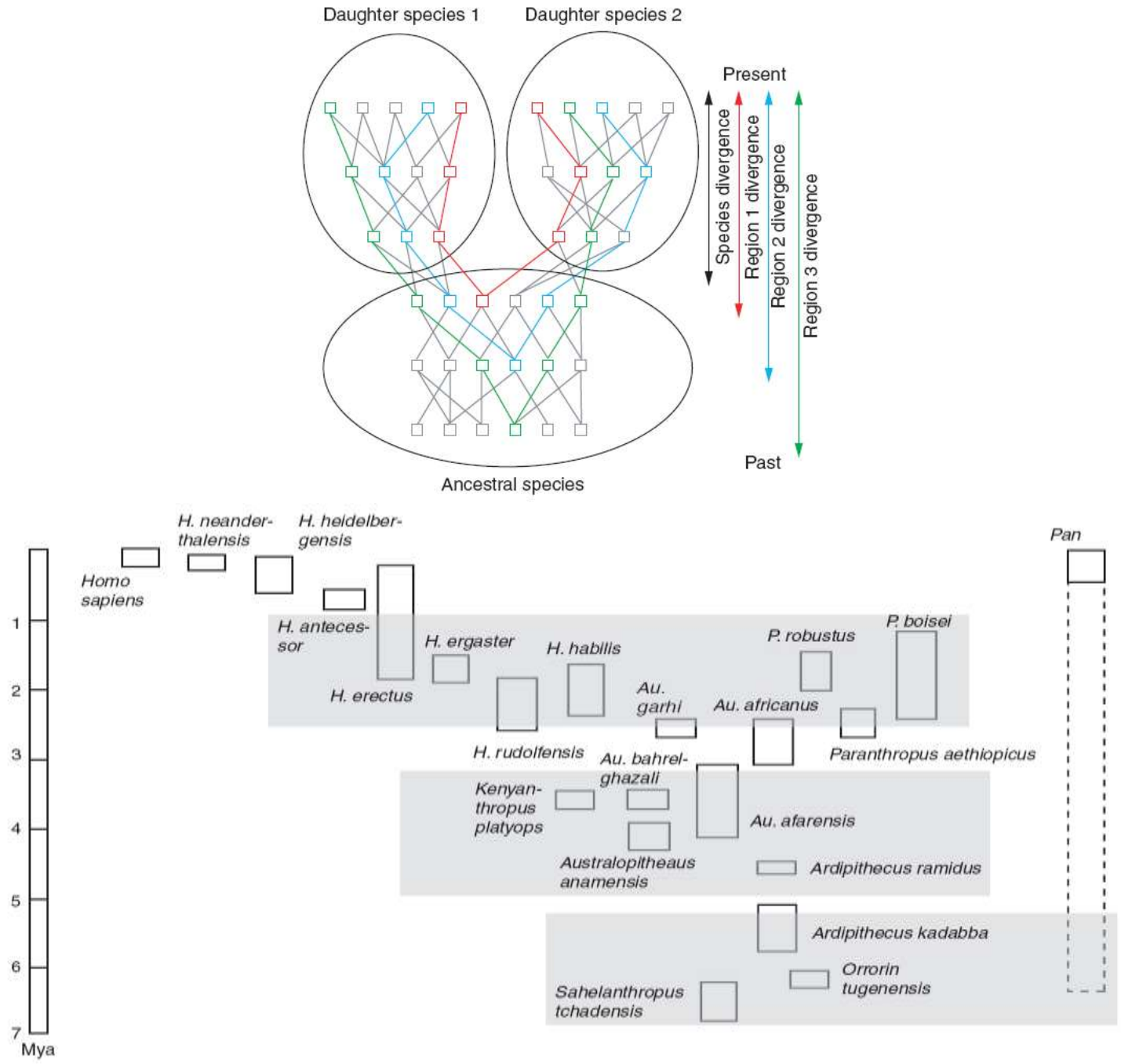
<<1 : negativní (purifying) selekce
~1 : neutrální
>>1 : pozitivní selekce



Hurst LD: The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet. 2002 Sep;18(9):486.

Evoluce primátů

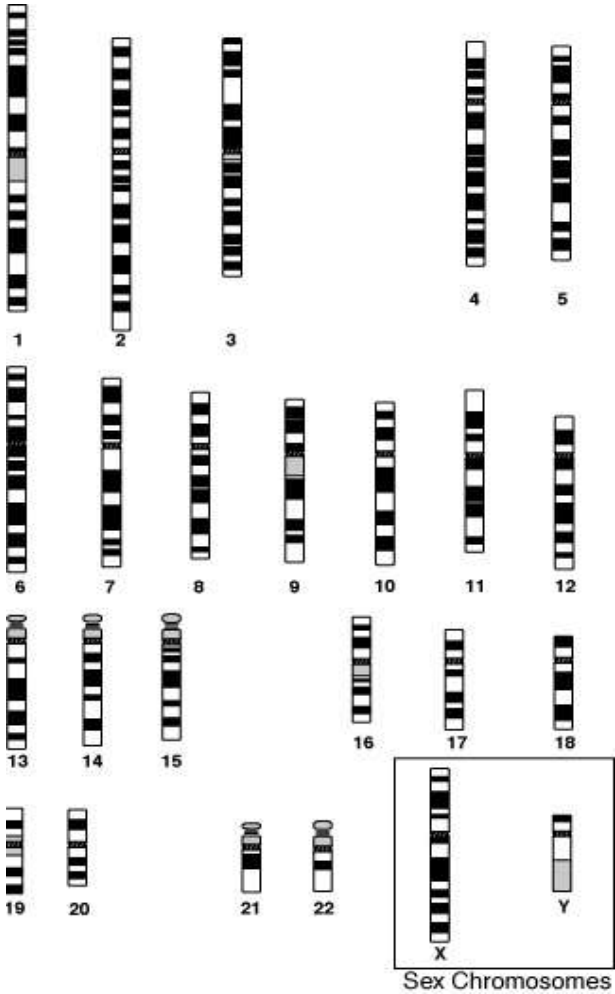




Disotell TR: 'Chumanzee' evolution: the urge to diverge and merge Genome Biology, Vol. 7 (24 November 2006), 240

Lidský genom

22 chromozómových párů
1 autozóm (sex chromozóm)
~3,3 miliard párů bazí
(~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)

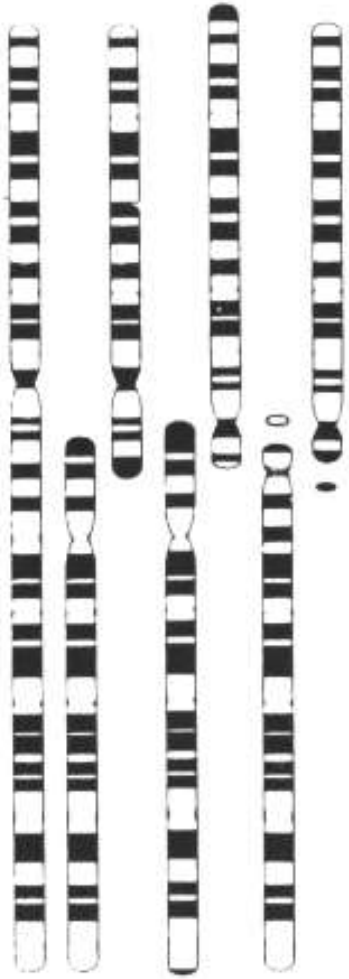
83% genů obsahuje rozdíl na úrovni aminokyselin

nejvíce jsou mutovány U3‘ oblasti

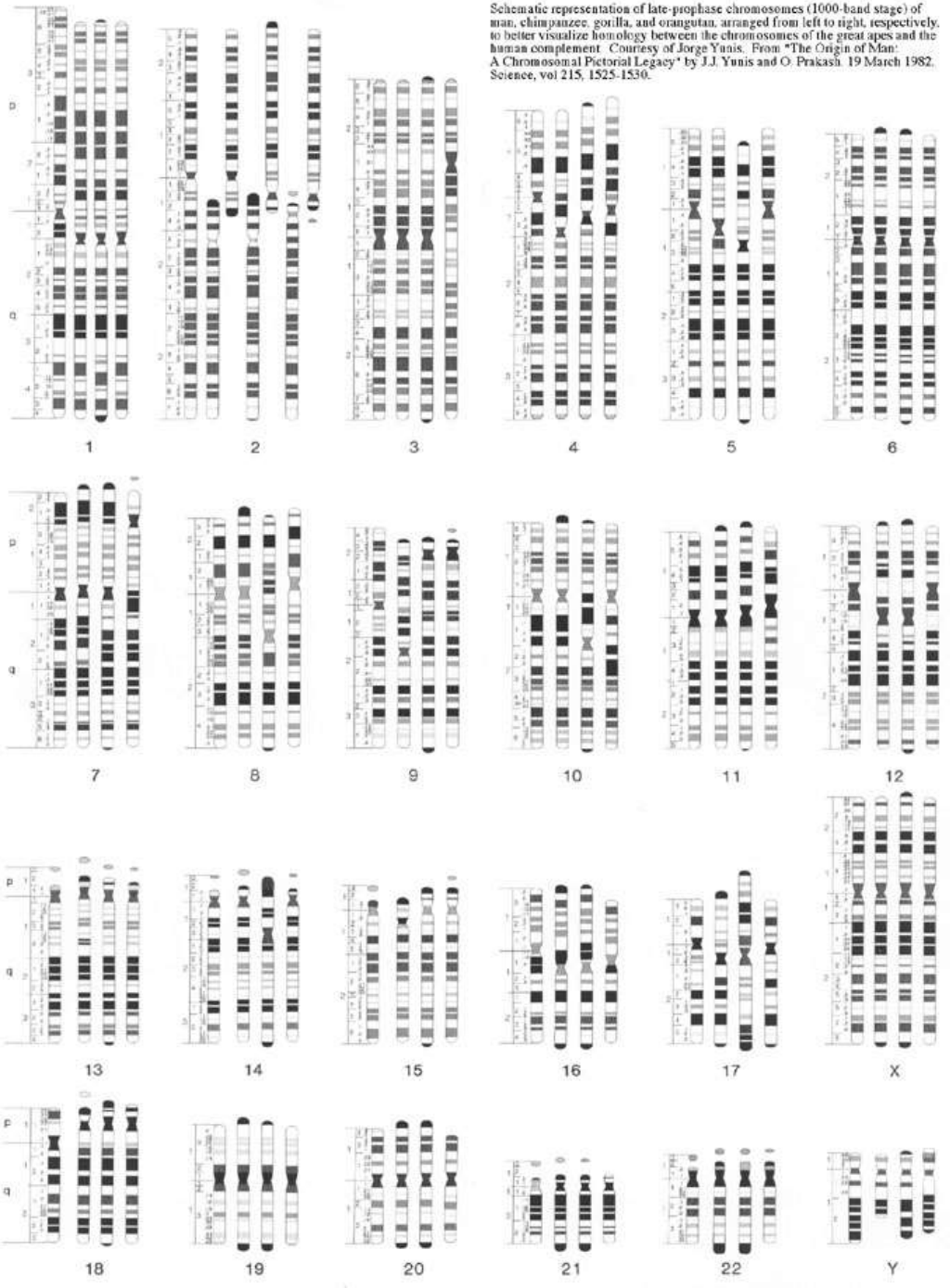
Watanabe H, et. al. DNA sequence and comparative analysis of chimpanzee chromosome 22. Nature. 2004 May 27;429(6990):382-8.
Ingo Ebersberger, Dirk Metzler, Carsten Schwarz and Svante Paabo: Am. J. Hum. Genet. 70:1490 1497, 2002 Genomewide Comparison of DNA Sequences between Humans and Chimpanzees

Hledáme něco navíc

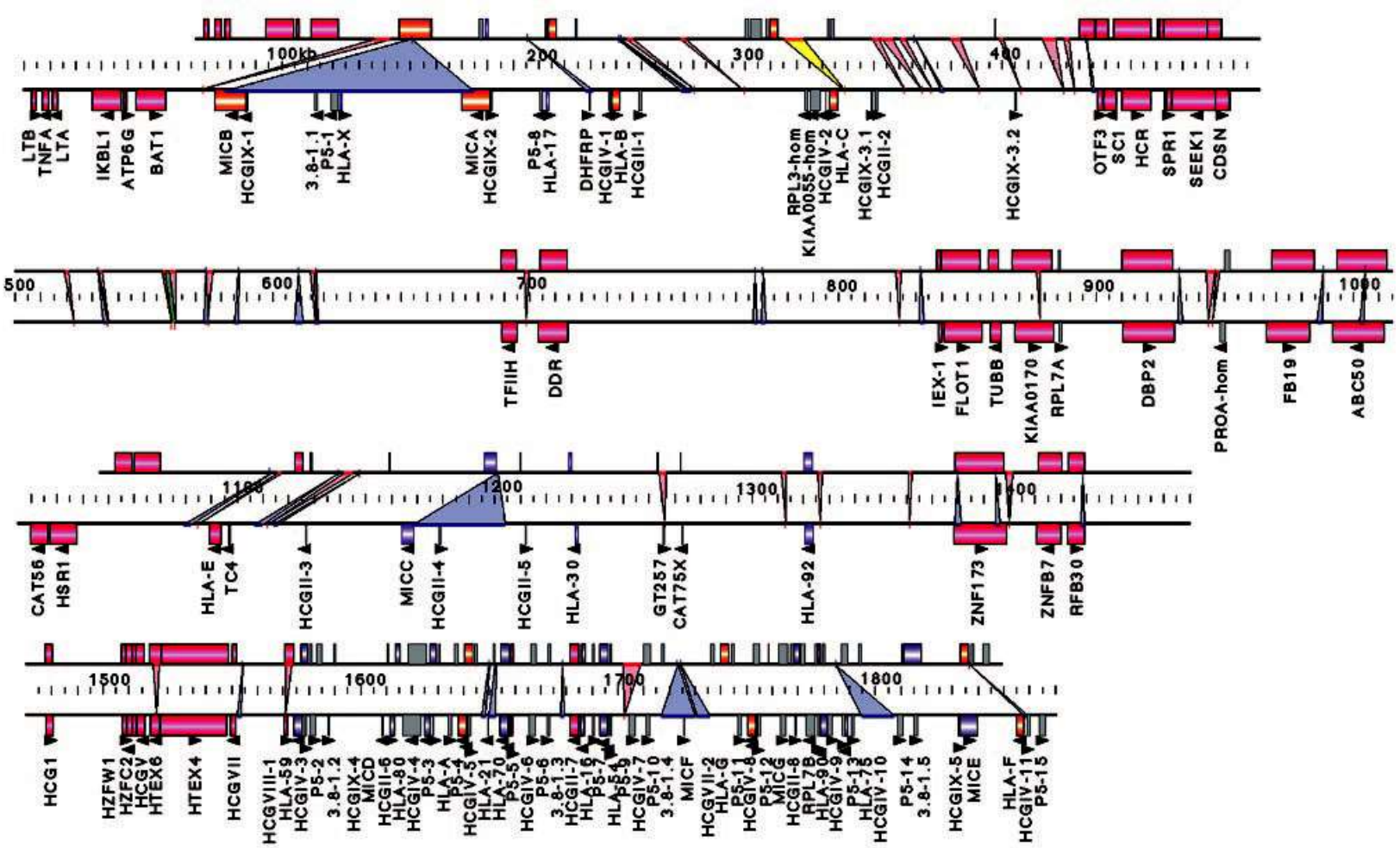
člověk
šimpanz
gorila
orangutan



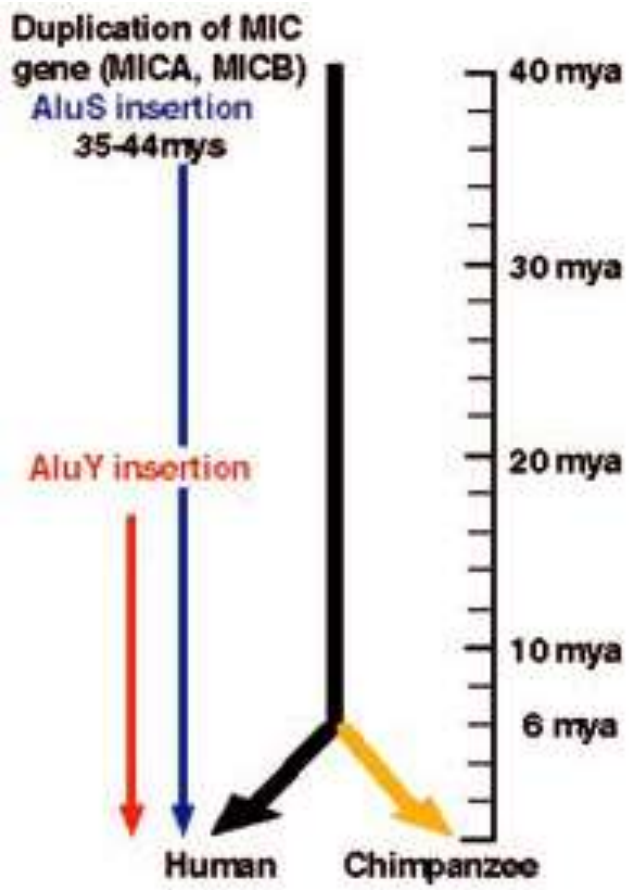
Schematic representation of late-prophase chromosomes (1000-band stage) of man, chimpanzee, gorilla, and orangutan, arranged from left to right, respectively, to better visualize homology between the chromosomes of the great apes and the human complement. Courtesy of Jorge Yunis. From "The Origin of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

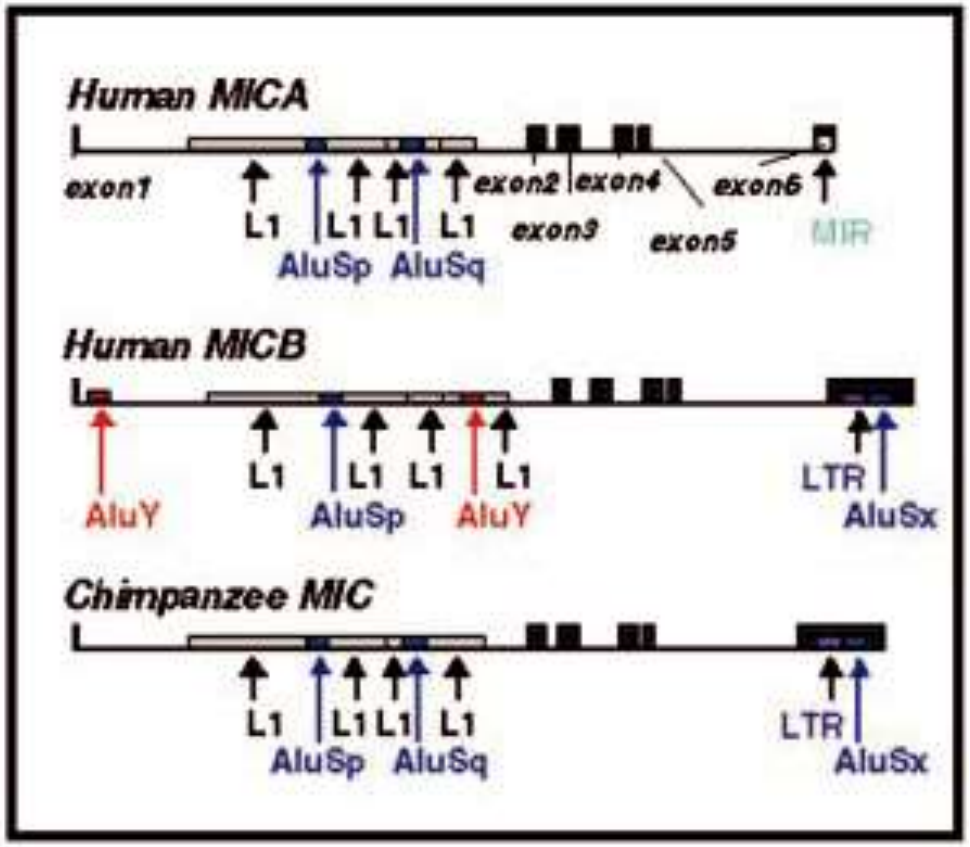


MHC I


100kb
200
300
400
LTB
TNFA
LTA
IKBL1
ATP6G
BAT1
MICB
HCGIX-1
3.8-1.1
P5-1
HLA-X
MICA
HCGIX-2
P5-8
HLA-17
DHFRP
HCGIV-1
HLA-B
HCGII-1
RPL3-hom
KIAA0055-hom
HCGIV-2
HLA-C
HCGIX-3.1
HCGII-2
HCGIX-3.2
OTF3
SC1
HCR
SPR1
SEEK1
CDSN
500
600
700
800
900
1000
TFIIH
DDR
IEX-1
FLOT1
TUBB
KIAA0170
RPL7A
DBP2
PROA-hom
FB19
ABC50
1100
1200
1300
1400
CAT56
HSR1
HLA-E
TC4
HCGII-3
MICC
HCGII-4
HCGII-5
HLA-30
GT257
CAT75X
HLA-92
ZNF173
ZNFB7
RFB30
1500
1600
1700
1800
HCG1
HZFW1
HZFC2
HCGV
HTEX6
HTEX4
HCGVII
HCGVIII-1
HLA-59
HCGIV-3
P5-2
3.8-1.2
HCGIX-4
MICD
HCGII-6
HLA-80
HCGIV-4
P5-3
HLA-A
P5-4
HCGIV-5
HLA-21
HLA-70
P5-5
HCGIV-6
P5-6
3.8-1.3
HCGII-7
HLA-16
P5-7
HLA-54
P5-9
HCGIV-7
P5-10
3.8-1.4
MICF
HCGVII-2
HLA-G
P5-11
HCGIV-8
P5-12
MICG
HCGII-8
RPL7B
HLA-90
HCGIV-9
P5-13
HLA-75
HCGIV-10
P5-14
3.8-1.5
HCGIX-5
MICE
HLA-F
HCGIV-11
P5-15

MIC A/B


Duplication of MIC gene (MICA, MICB)
AluS insertion
35-44mys
AluY insertion
40 mya
30 mya
20 mya
10 mya
6 mya
Human
Chimpanzee


Human MICA
exon1
L1
L1
L1
L1
exon2
exon4
exon6
exon3
exon5
MIR
AluSp
AluSq
Human MICB
AluY
L1
L1
L1
AluY
L1
LTR
AluSp
AluSx
Chimpanzee MIC
L1
L1
L1
L1
LTR
AluSp
AluSq
AluSx

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

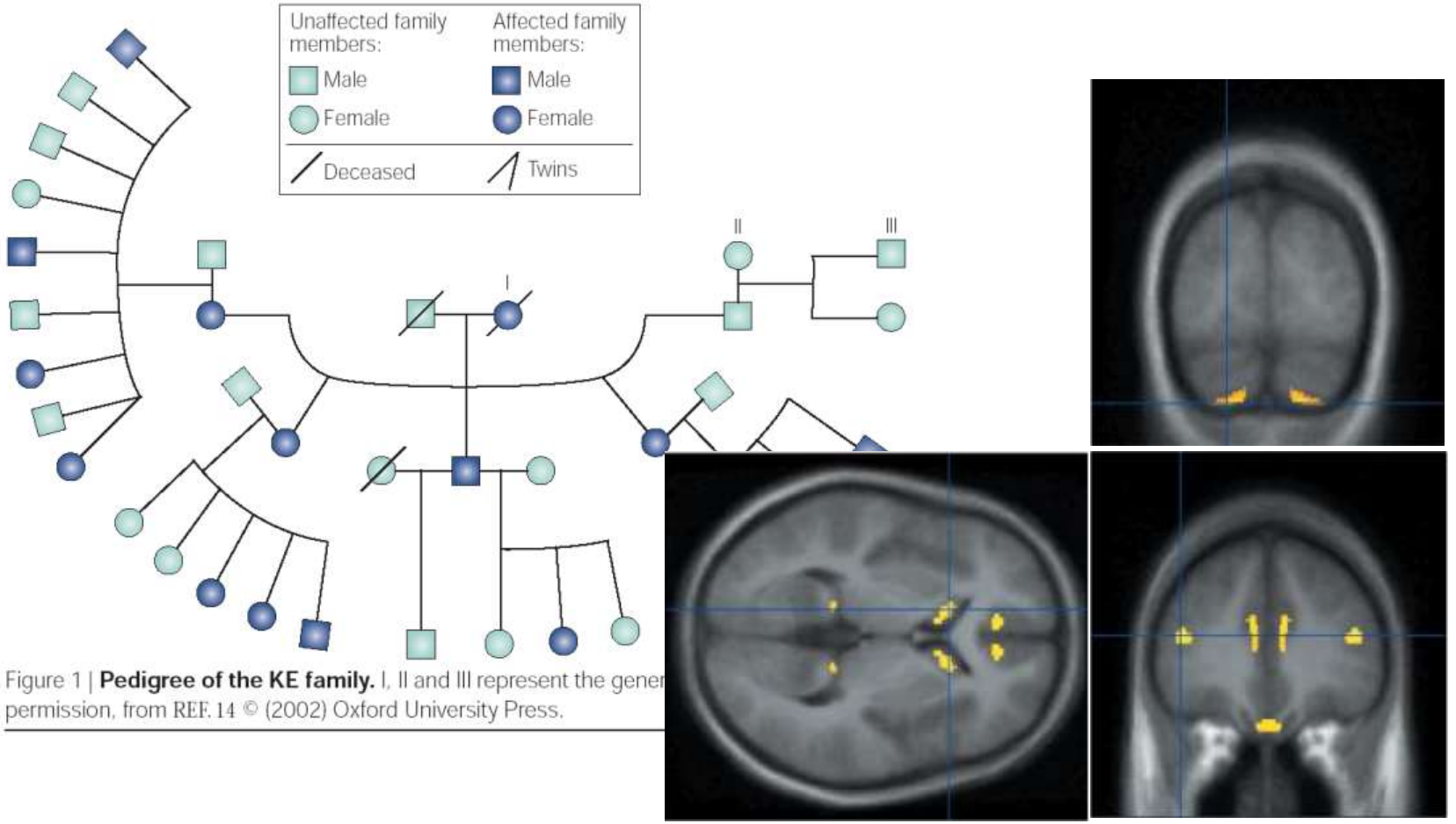
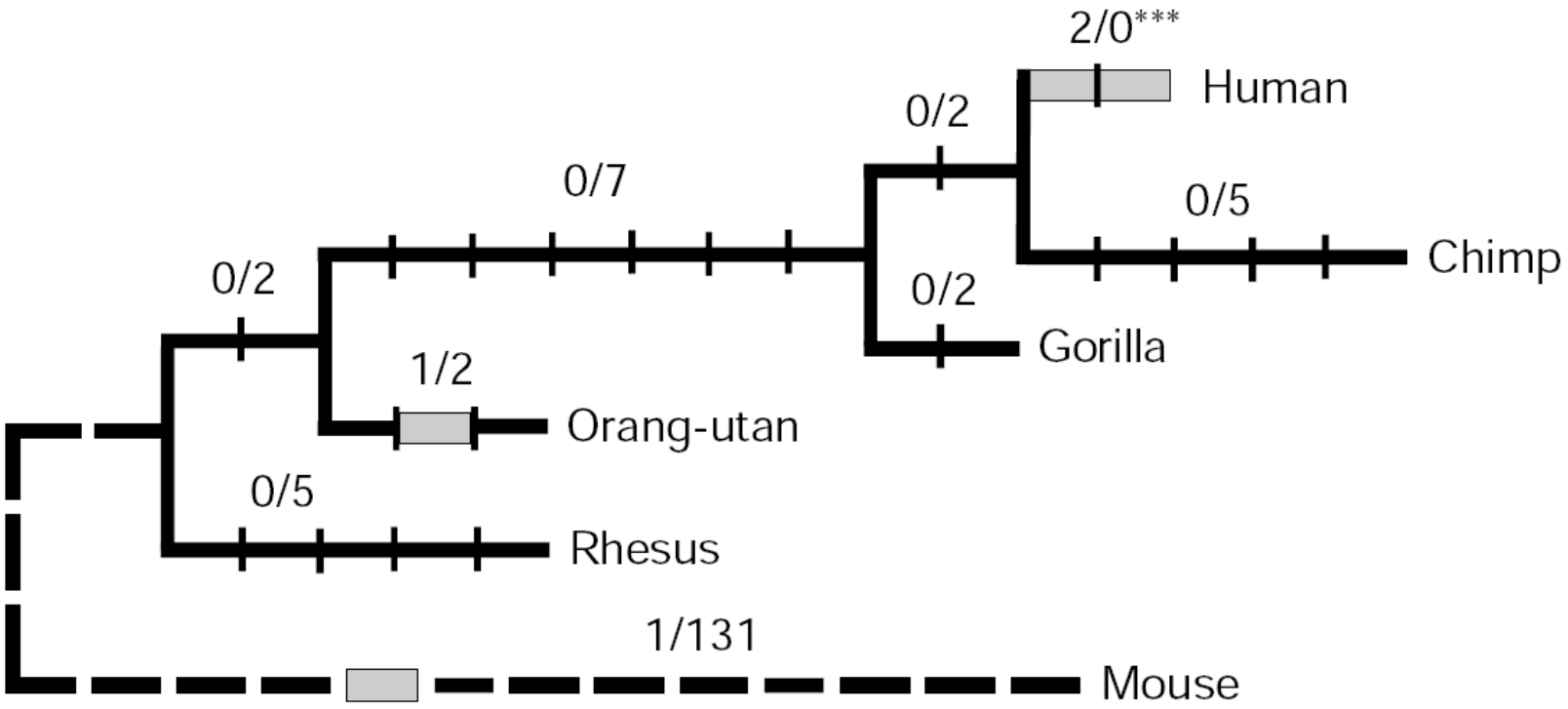


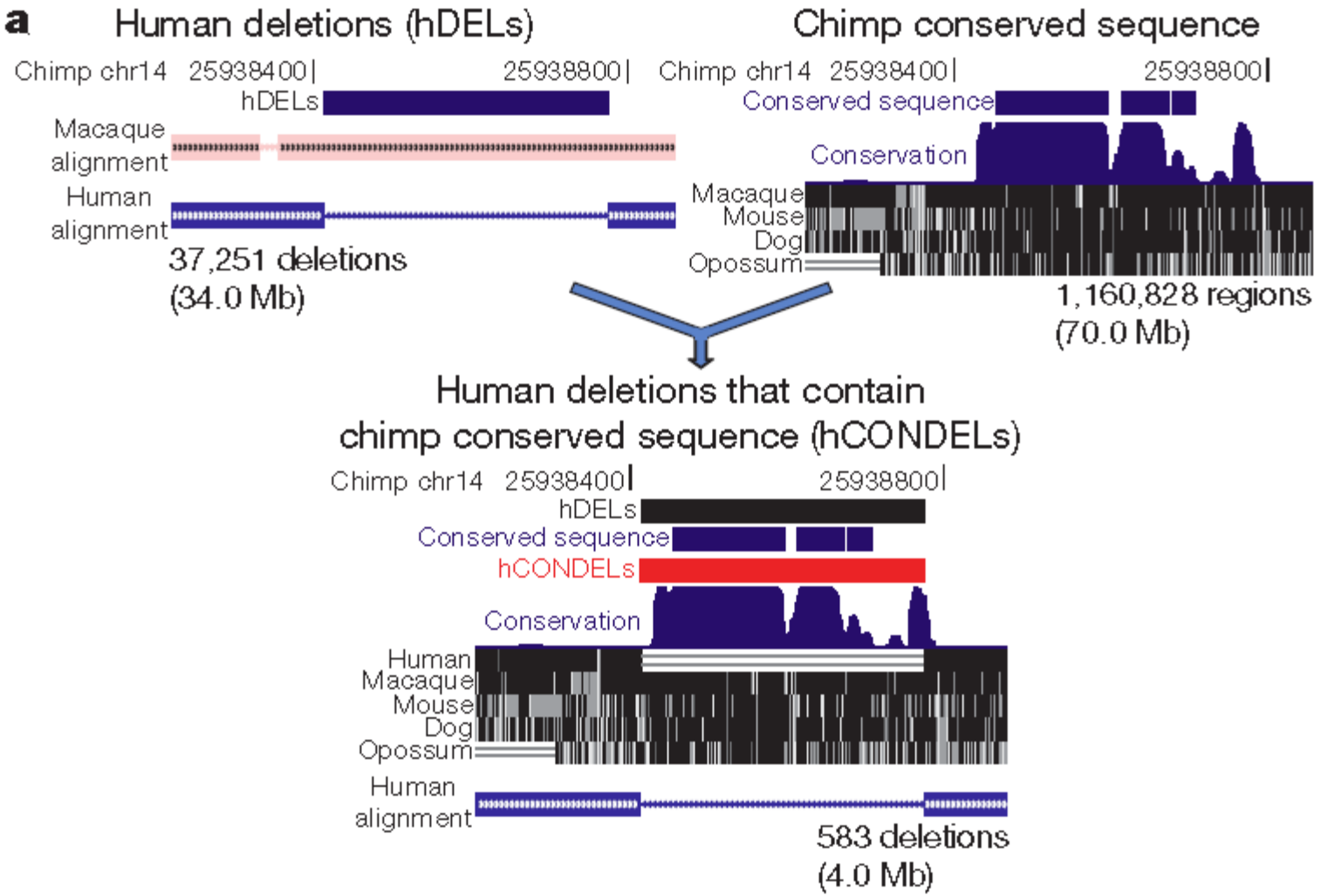
Figure 1 | **Pedigree of the KE family.** I, II and III represent the gener
permission, from REF. 14 © (2002) Oxford University Press.

evoluce FOXP2

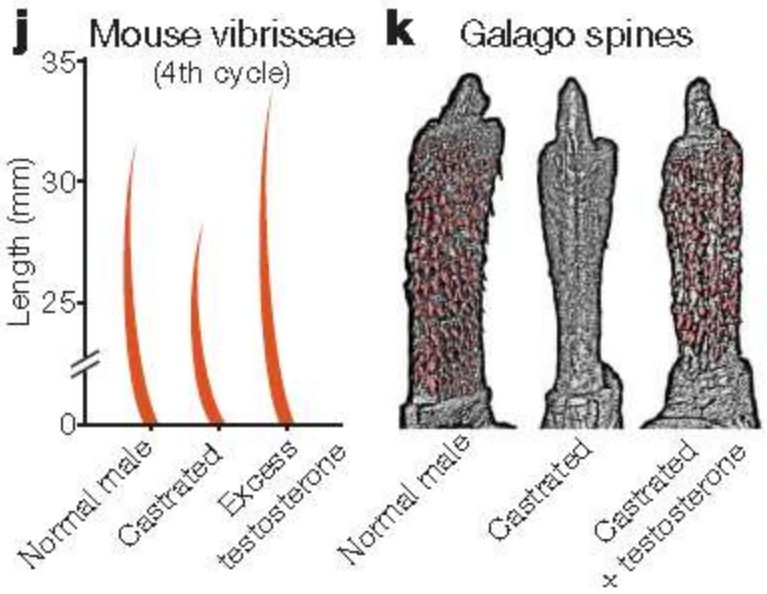


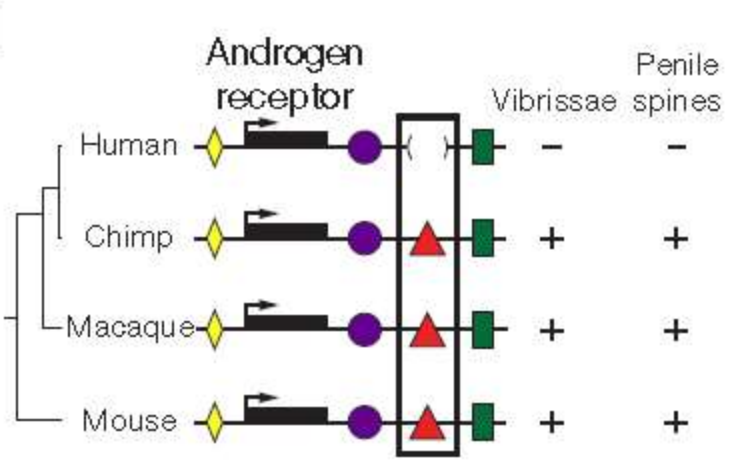
Hledáme co chybí

hCONDELS

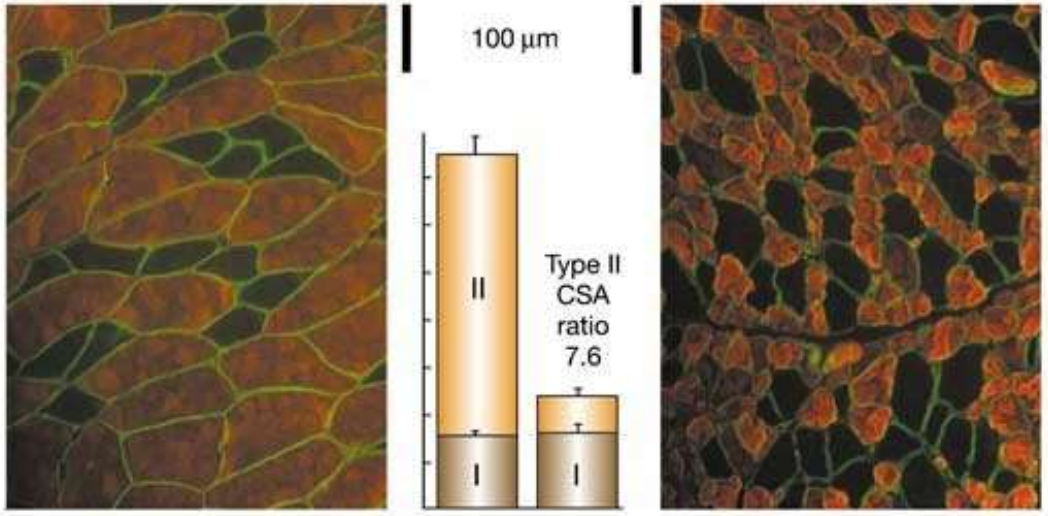

a
Human deletions (hDELs)
Chimp chr14 25938400| 25938800|
hDELs
Macaque alignment
Human alignment
37,251 deletions
(34.0 Mb)
Chimp conserved sequence
Chimp chr14 25938400| 25938800|
Conserved sequence
Conservation
Macaque
Mouse
Dog
Opossum
1,160,828 regions
(70.0 Mb)
Human deletions that contain chimp conserved sequence (hCONDELs)
Chimp chr14 25938400| 25938800|
hDELs
Conserved sequence
hCONDELs
Conservation
Human
Macaque
Mouse
Dog
Opossum
Human alignment
583 deletions
(4.0 Mb)

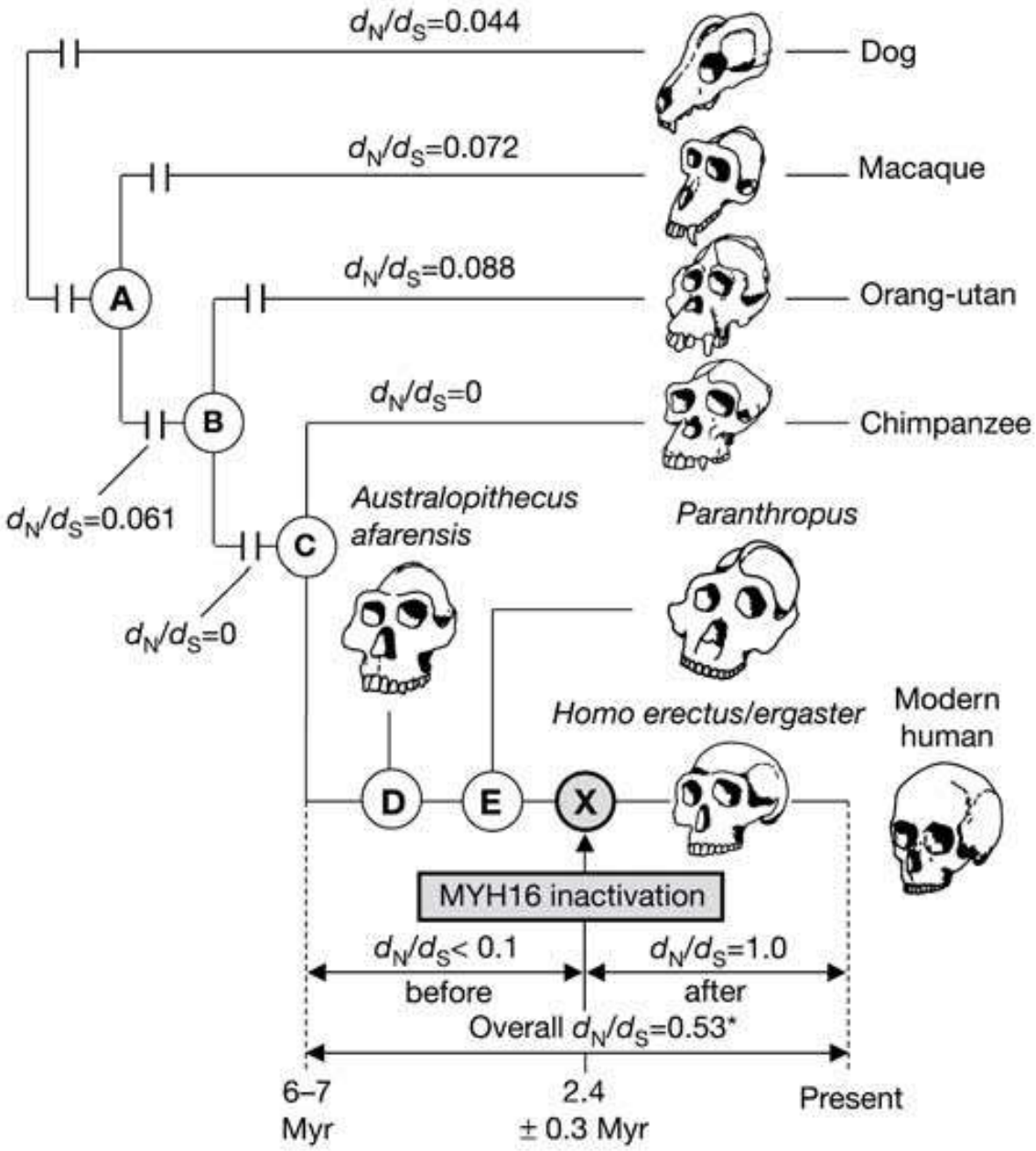
Regulace androgenního receptoru





MYH16 inactivation

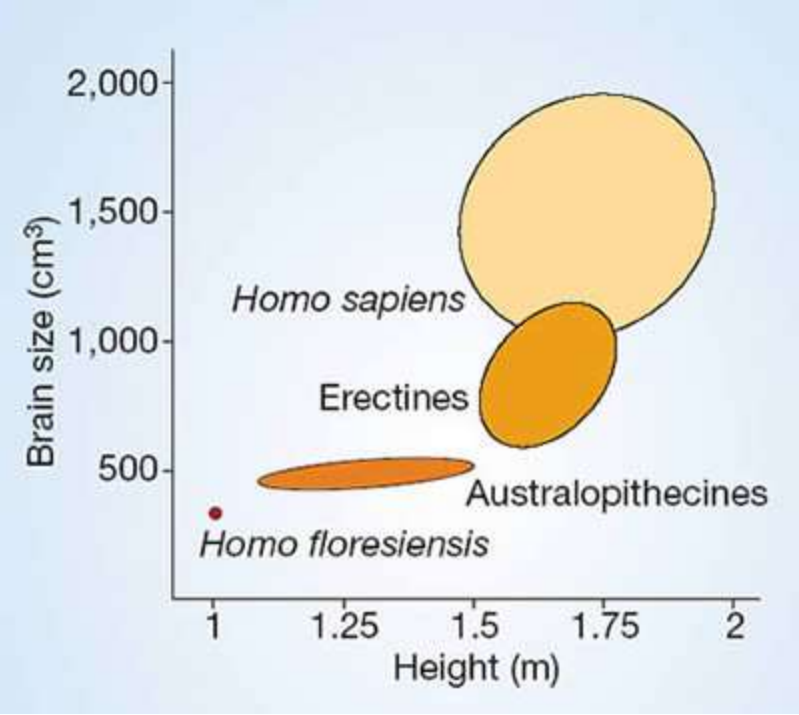




Stedman HH, Kozyak BW, Nelson A, Thesier DM, Su LT, Low DW, Bridges CR, Shrager JB, Minugh-Purvis N, Mitchell MA. Myosin gene mutation correlates with anatomical changes in the human lineage. Nature. 2004 Mar 25;428(6981):415-8.

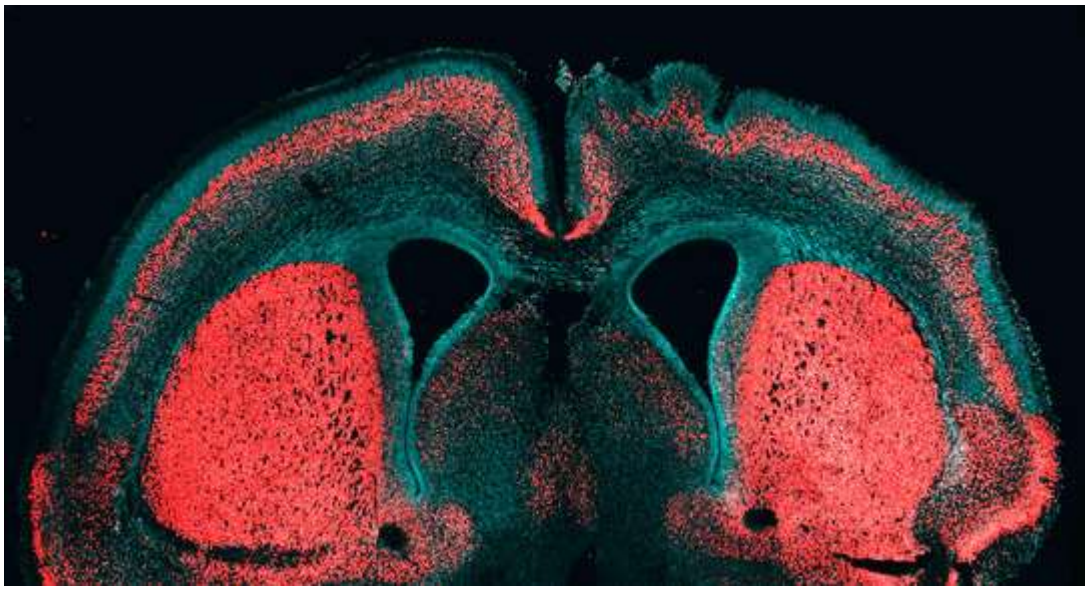
Homo floresiensis

Velikost lebky (mozku)



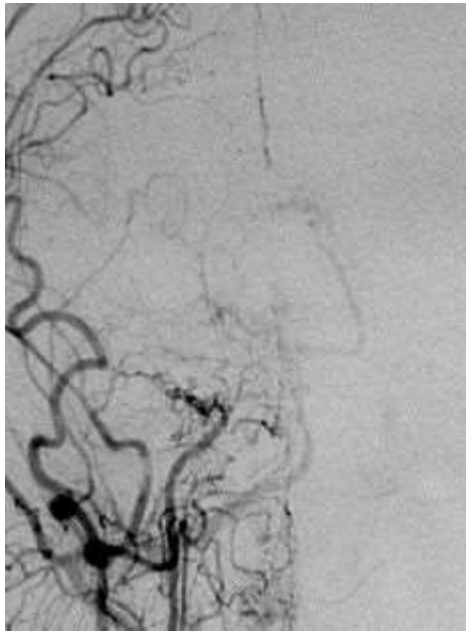
Přeci jen něco navíc

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



Pozitivně selektované geny

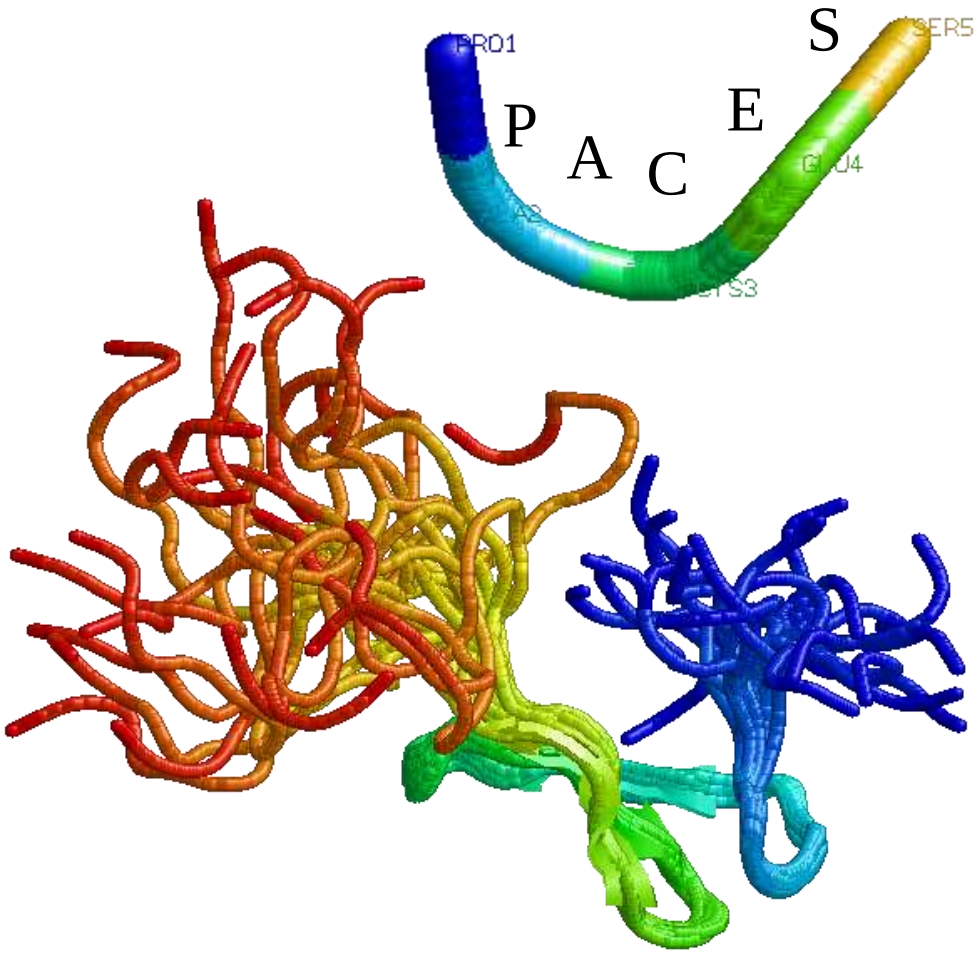
- **RNF213**
  - **Ring finger protein 213**
  - Mutace způsobuje Moyamoya syndrom
  - Ovlivňuje velikost artérií v mozku.
  - Pozitivní selekce u primátů



kontakt

Jan Pačes

www: http://bio.img.cas.cz
email: hpaces@img.cas.cz
icq: #110872370
irc: efnet #hpaces
tel: +420 220183446

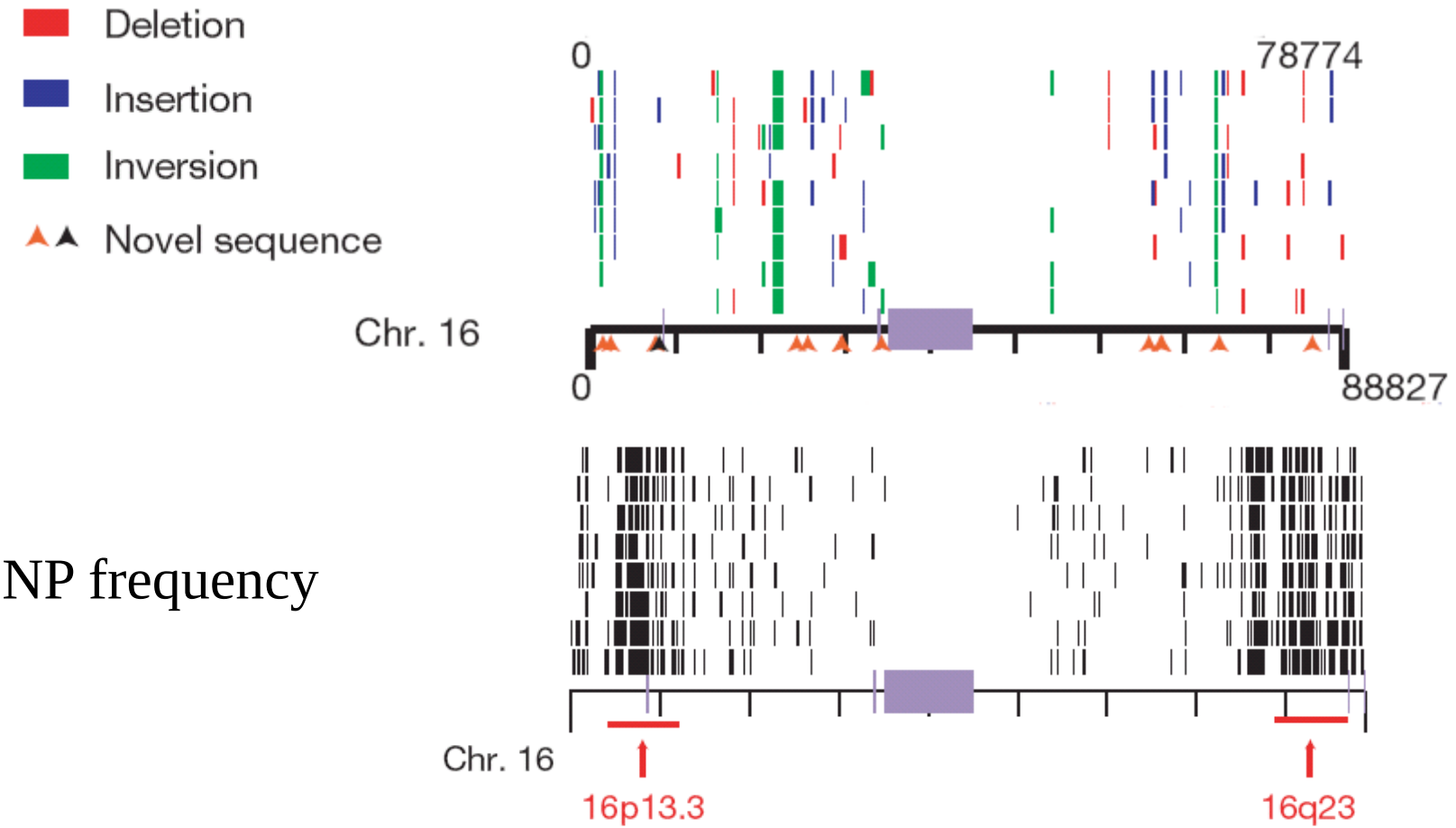


Proč zrovna genom šimpanze?

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

A White Paper Advocating Complete Sequencing of the Genome of the Common Chimpanzee, *Pan Troglodytes*
Maynard V. Olson,a,b Evan E. Eichler,c Ajit Varki,d Richard M. Myers,e Joseph M. Erwin,f and Edwin H. McConkeyg

8 human genomes comparison

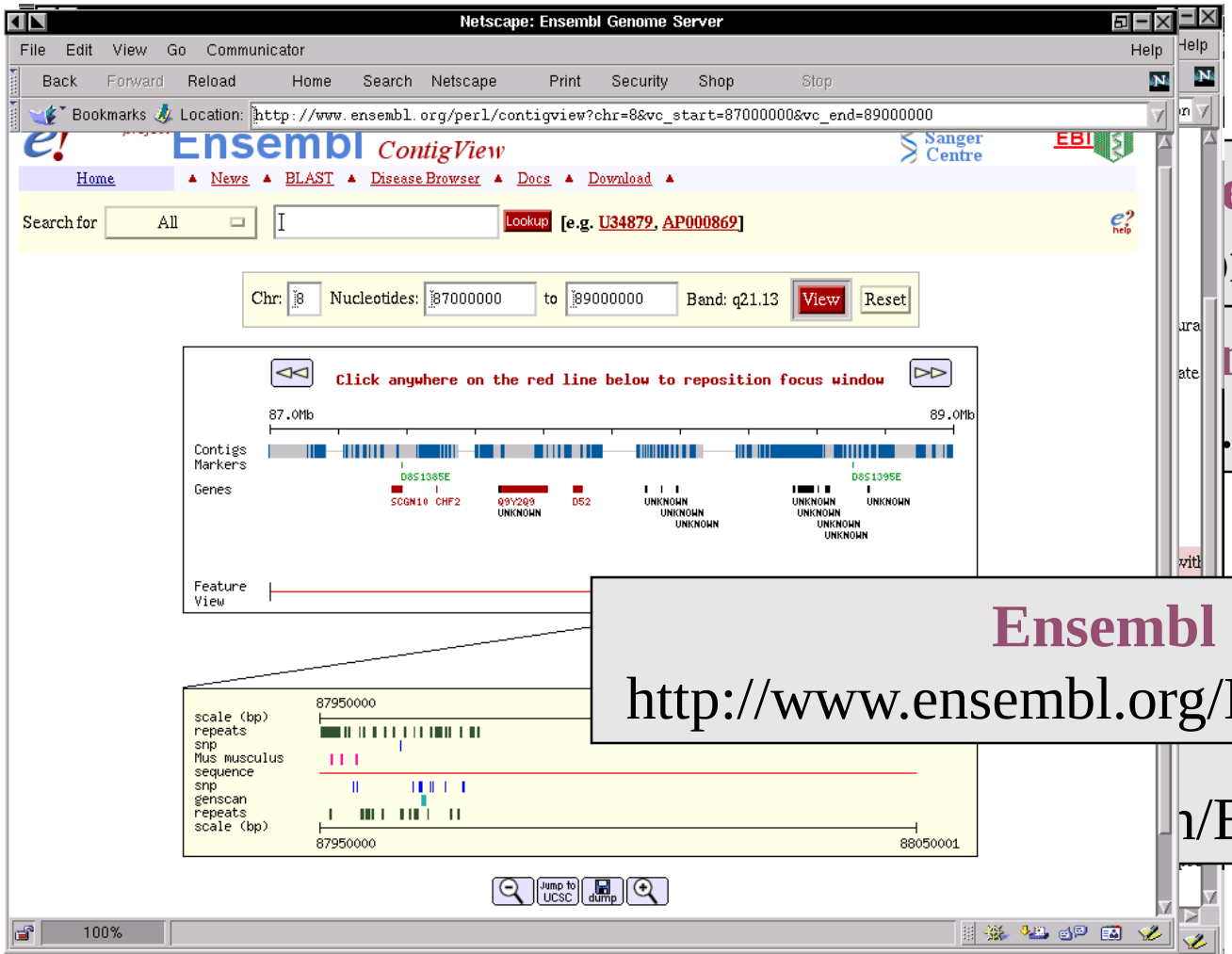


FOXP2


Human MMQESATETI SNSSMNQNGM STLSSQLDAG SRDGRSSGDT SSEVSTVELL HLQQQQALQA ARQLLLQQQT SGLKSPKSSD KQRPLQVPVS VAMMTPQVIT
Chimp .....
Gorilla .....
Orang .....V.....
Rhesus .....
Mouse .....E.....
Human PQQMQQILQQ QVLSPQQLQA LLQQQQAVML QQQQLQEFYK KQQEQLHLQL LQQQQQQQQQ QQQQQQQQQQ QQQQ-QQQQQ QQQQQQQQQQ QQHPGKQAKE
Chimp .....Q..... .....
Gorilla .....-..... .....--..... .....
Orang .....-..... .....-..... .....
Rhesus .....-..... .....-..... .....
Mouse .....-..... .....Q..... .....
Human QQQQQQQQQQ LAAQQLVFQQ QLLQMQQLQQ QQHLLSLQRQ GLISIPPGQA ALPVQSLPQA GLSPAEIQQL WKEVTGVHSM EDNGIKHGGL DLTTNNSSST
Chimp .....
Gorilla .....
Orang .....-.....
Rhesus .....
Mouse .....-.....
Human TSSNTSKASP PITHHSIVNG QSSVLSARRD SSSHEETGAS HTLYGHGVCK WPGCESICED FGQFLKHLNN EHALDDRSTA QCRVQMQVVQ QLEIQLSKER
Chimp ...T.....N.....
Gorilla ...T.....N.....
Orang ...T.....N.....
Rhesus ...T.....N.....
Mouse ...T.....N.....

Hledáme rozdíl

kde začít na internetu



enPath

me.ucsc.edu

nks

.cas.cz/links

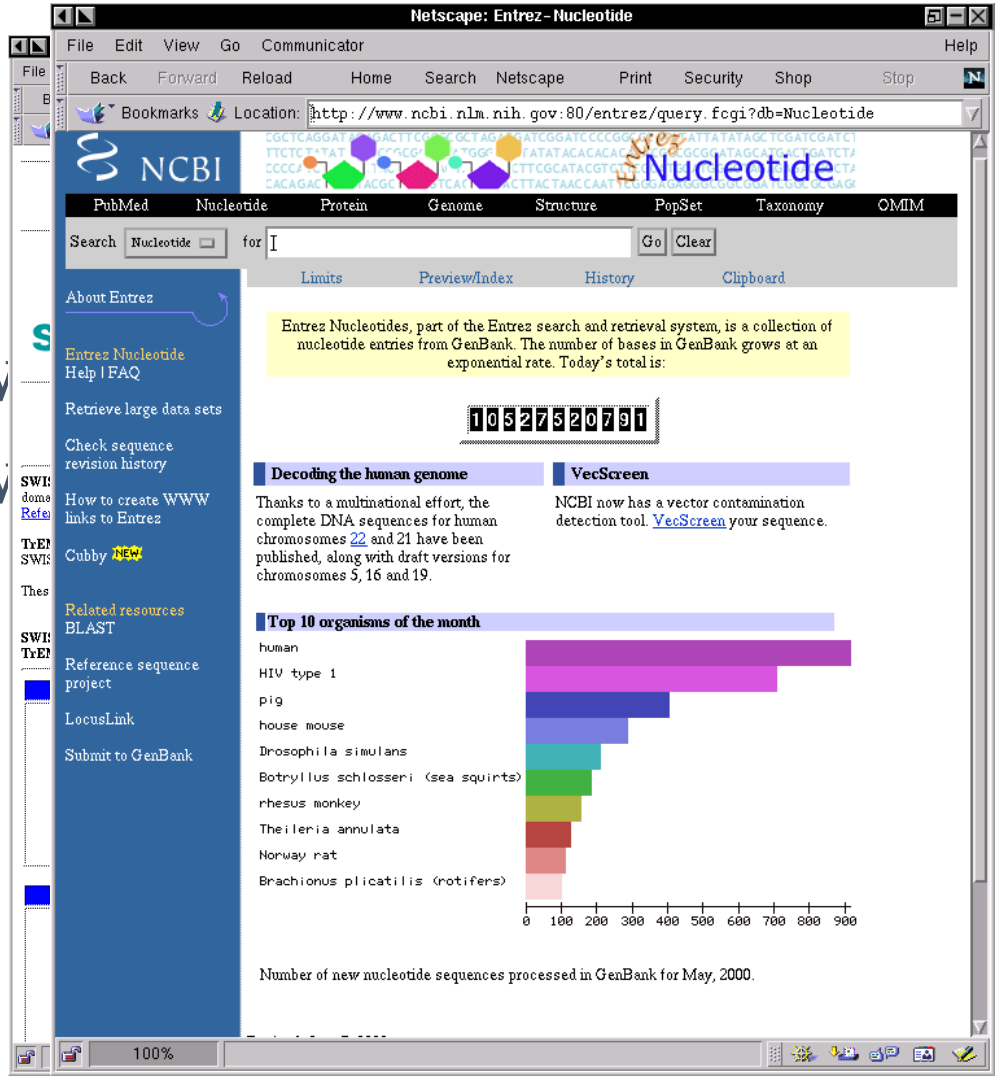
Ensembl

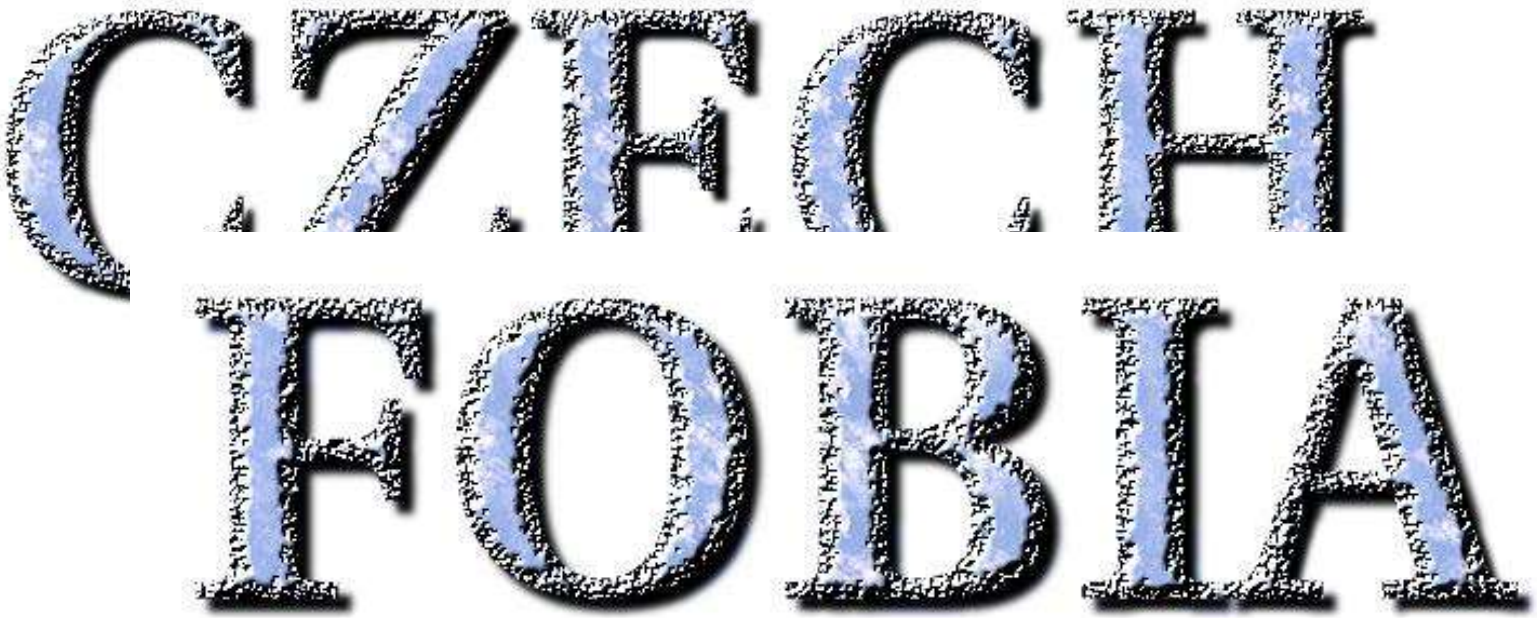
http://www.ensembl.org/Homo_sapiens

n/Entrez/hum_srch

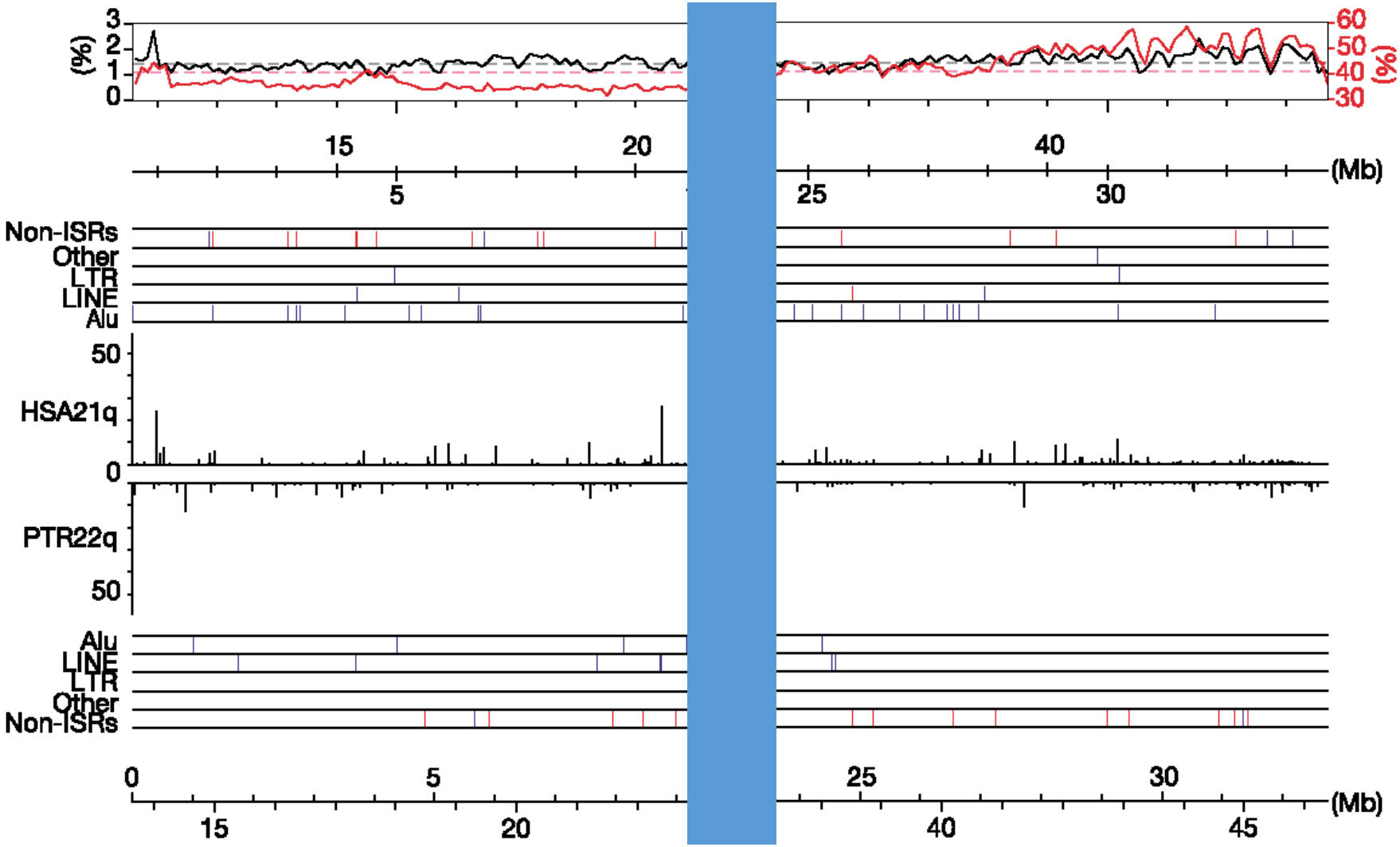
vybrané zdroje dat na internetu

SwissProt http://www
Entrez http://www



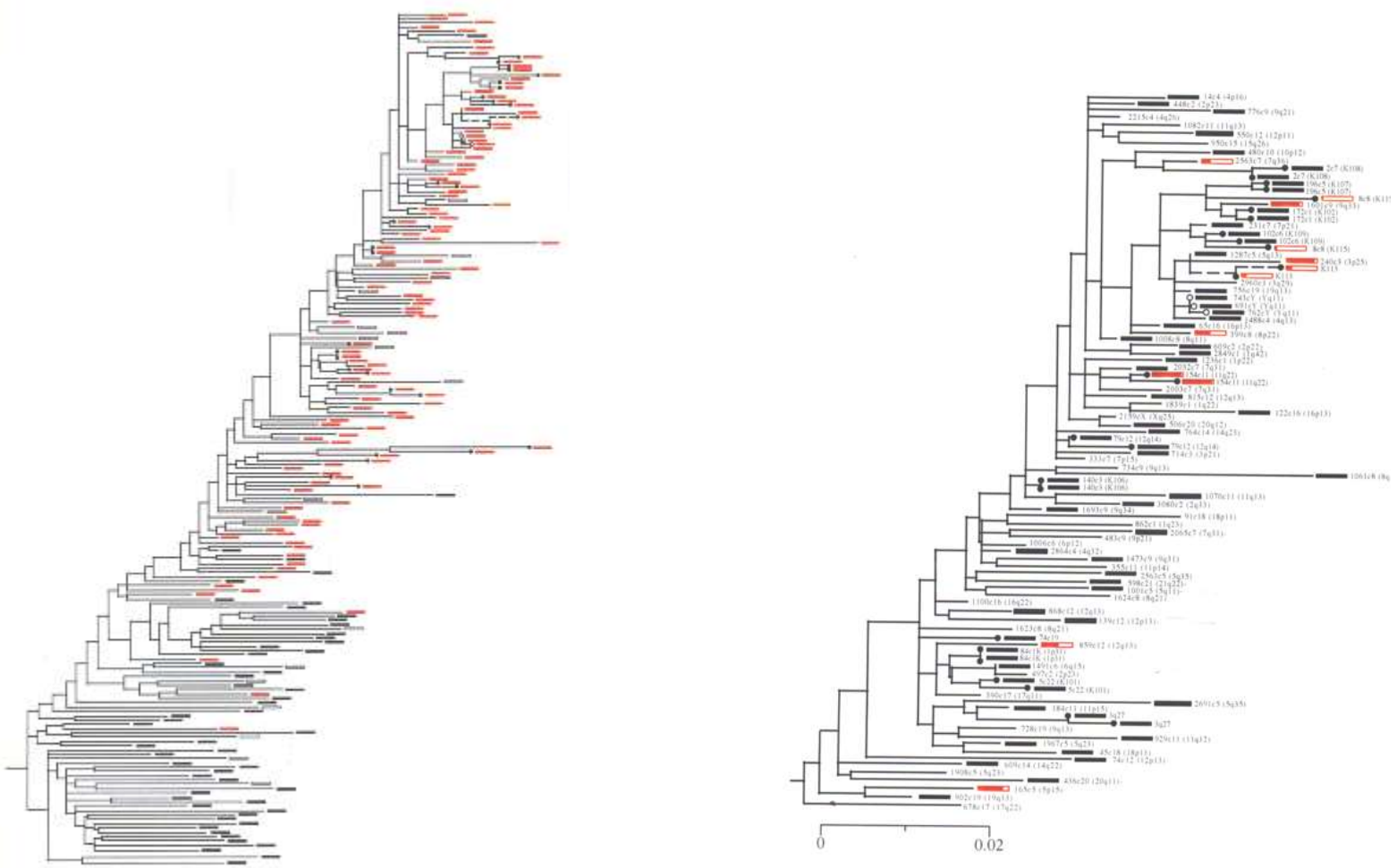

CZECH
FOBIA

hs ch21 vs pt ch22

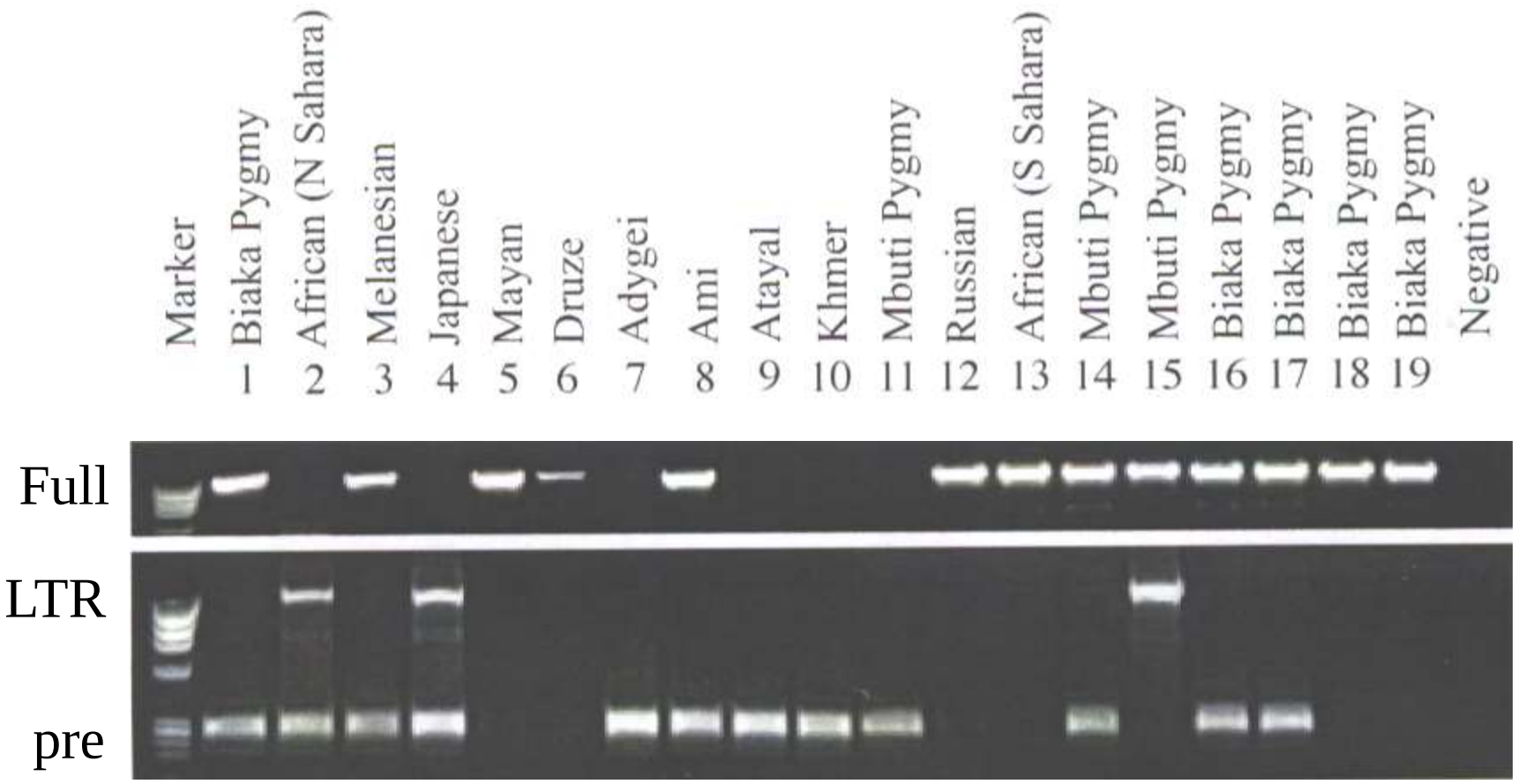


human vs *chimp* ERV analýza

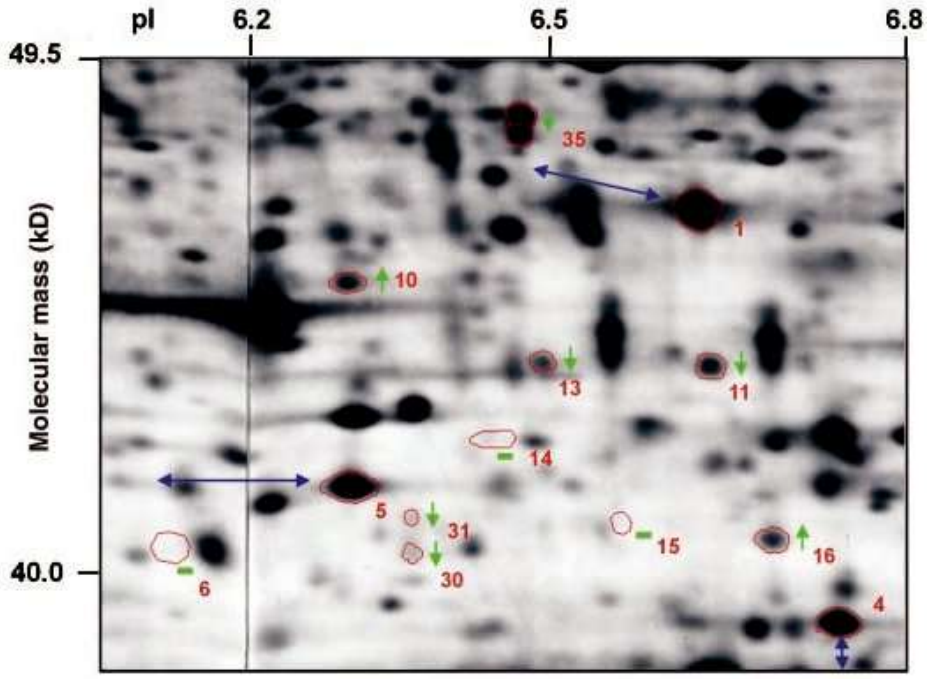
HERVK (HML2)



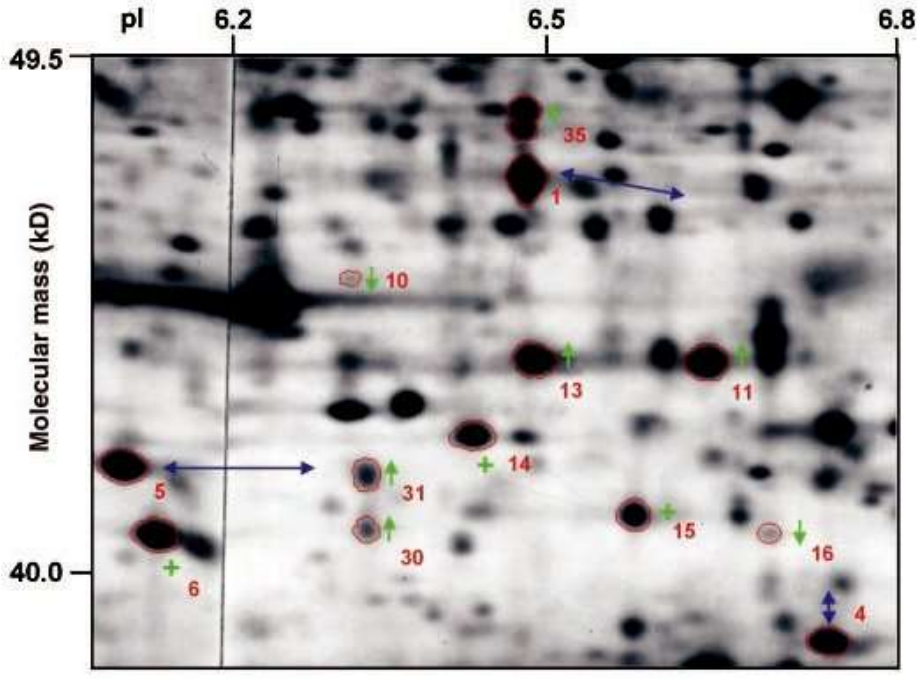
Polymorfní herv 259c12



proteom



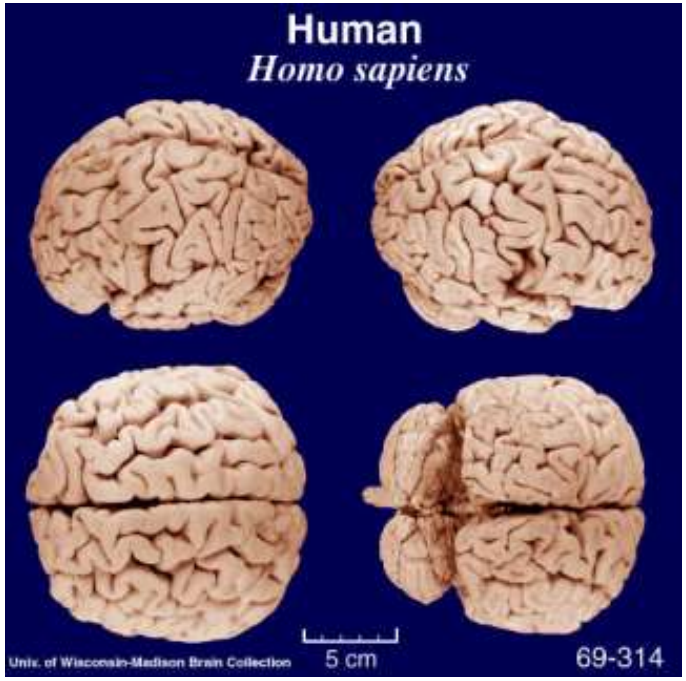
člověk

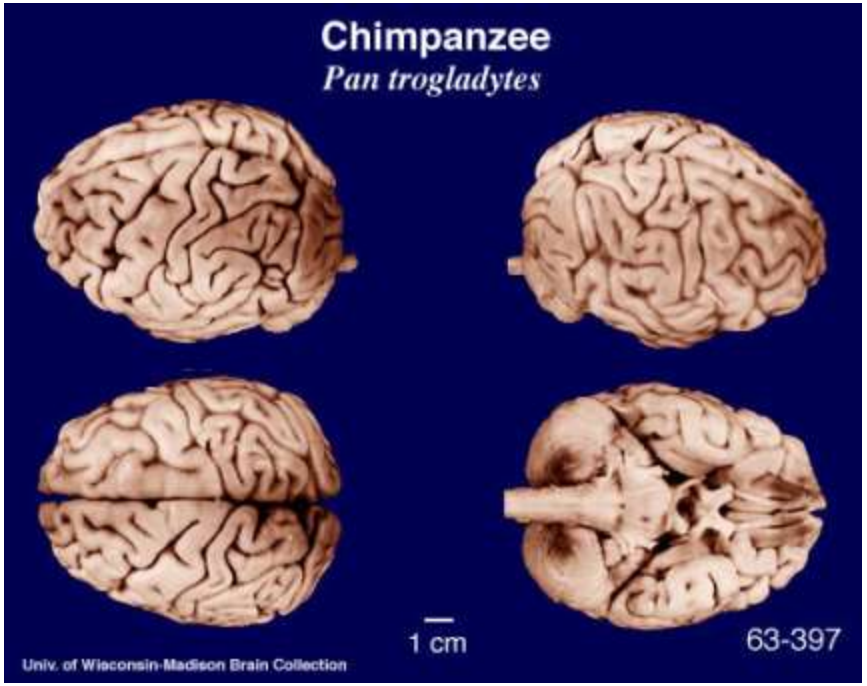


šimpanz

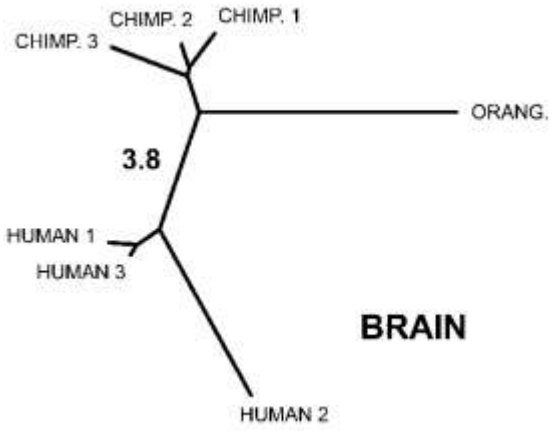
Enard W, Khaitovich P, Klose J, Zollner S, Heissig F, Giavalisco P, Nieselt-Struwe K, Muchmore E, Varki A, Ravid R, Doxiadis GM, Bontrop RE, Paabo S. Intra- and interspecific variation in primate gene expression patterns. Science. 2002 Apr 12;296(5566):340-3.

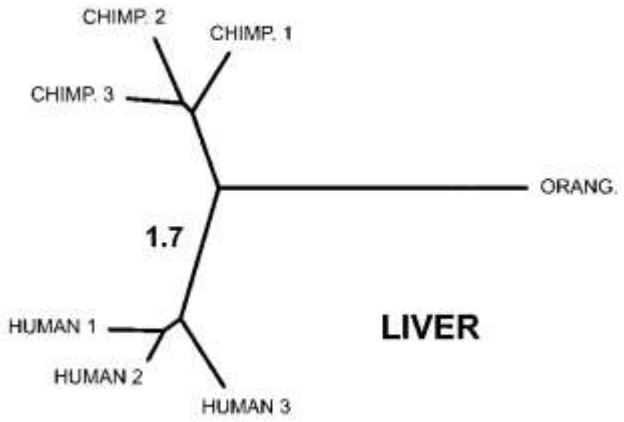
mozek

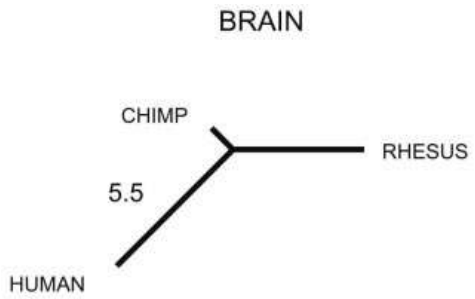

Human
Homo sapiens
Univ. of Wisconsin-Madison Brain Collection
5 cm
69-314

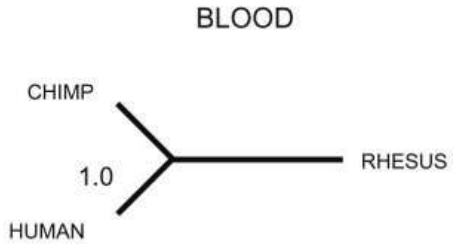

Chimpanzee
Pan trogladytes
Univ. of Wisconsin-Madison Brain Collection
1 cm
63-397

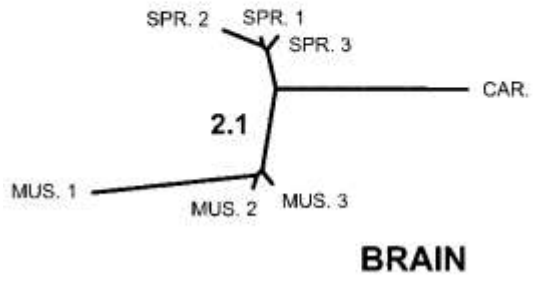
akcelerace vývoje mozku

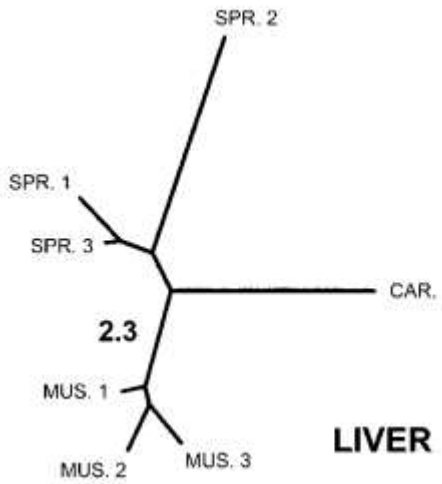


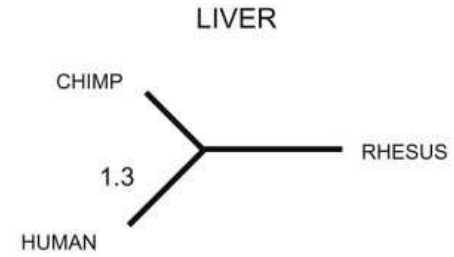




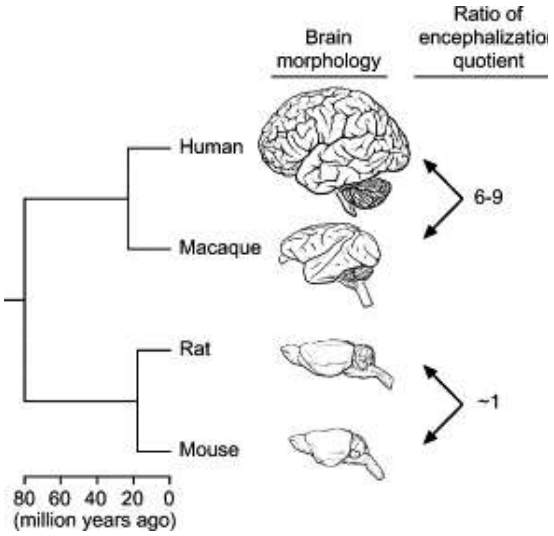


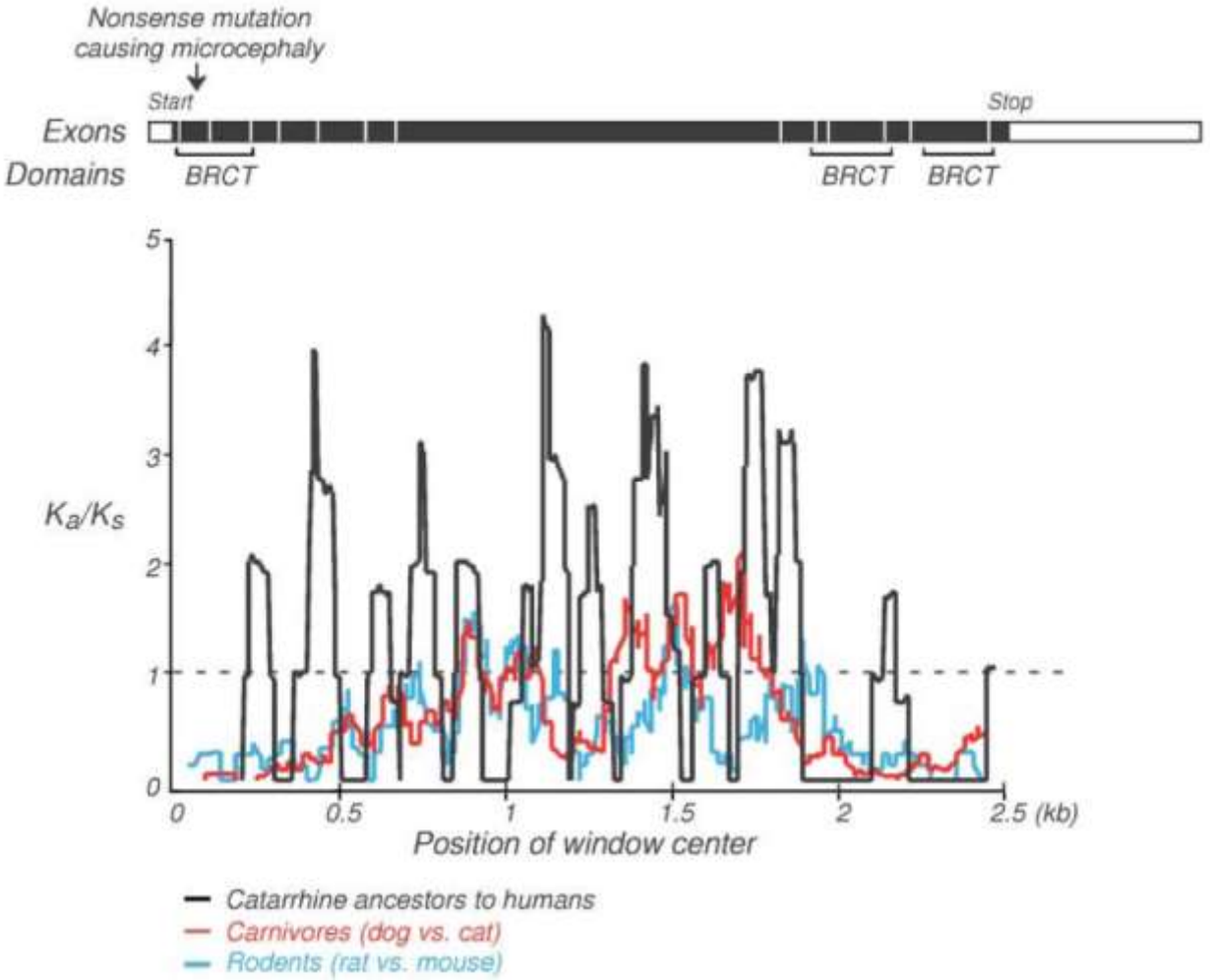






microcephalin



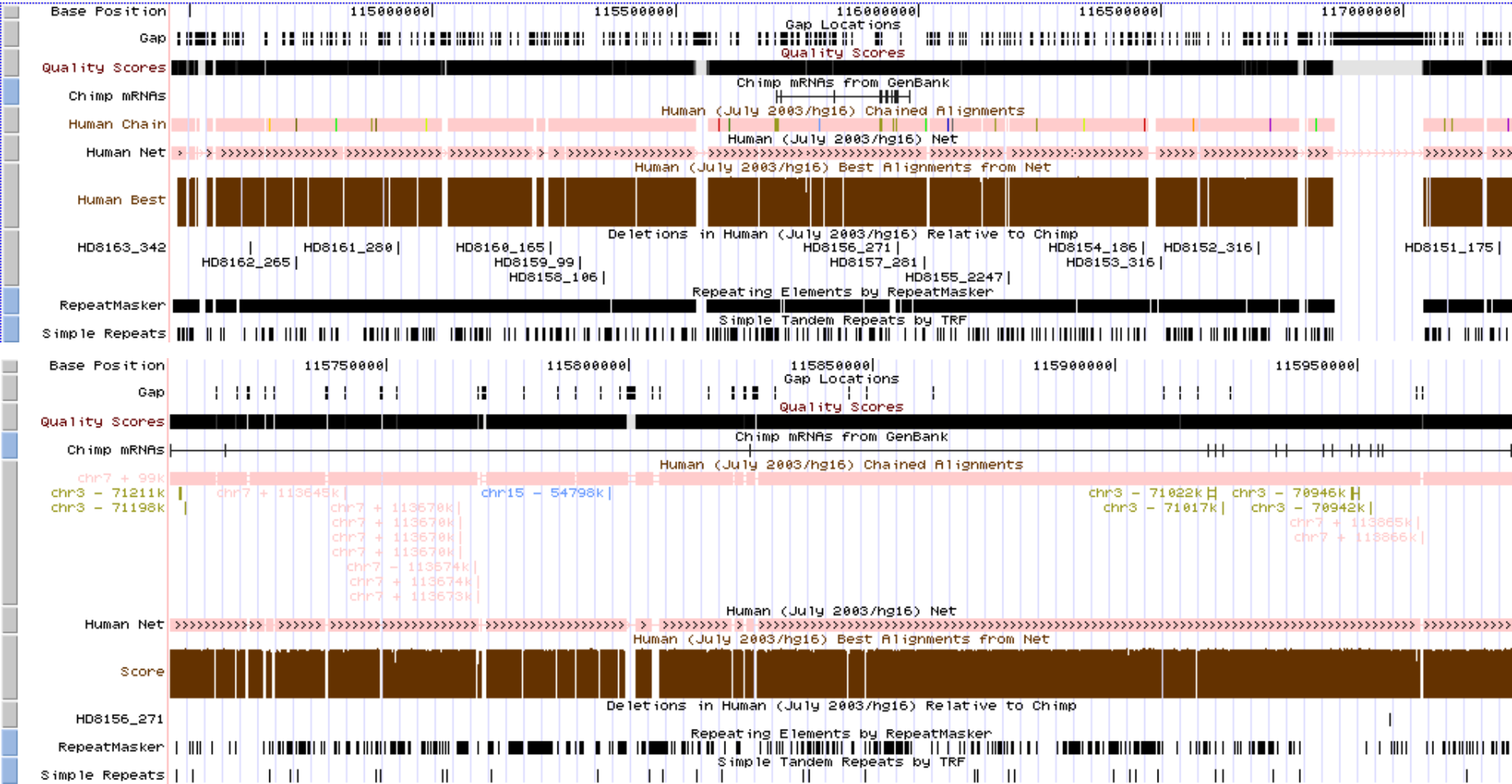


pozitivně selektované typy genů

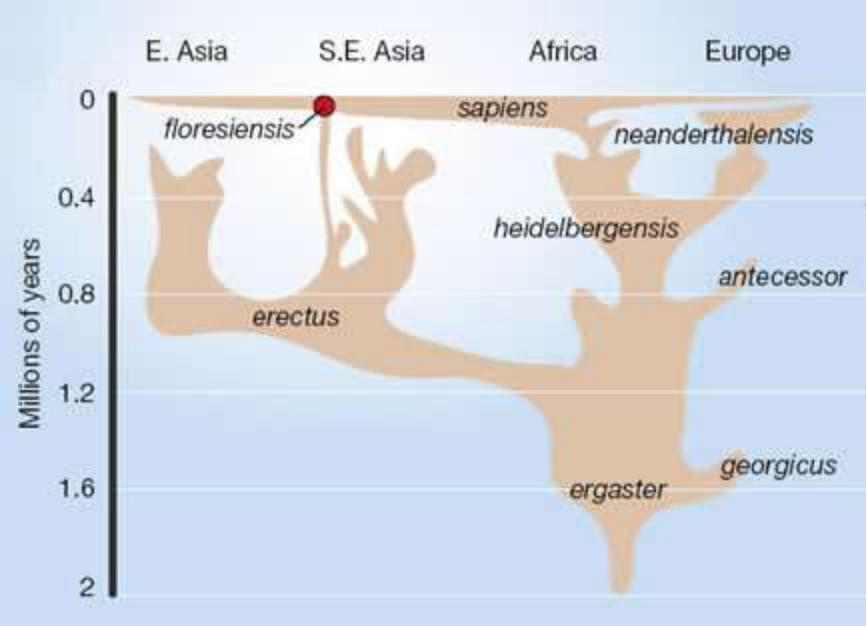
host – patogen interakce (MHC, CMAH)
reprodukce
adaptace na potravu
vzhled (MC1R)
smysly (čich, sluch)
chování
mozek

Vallender EJ, Lahn BT. Positive selection on the human genome. Hum Mol Genet. 2004 Oct 1;13 Spec No 2:R245-54.
Hellmann I, Zollner S, Enard W, Ebersberger I, Nickel B, Paabo S. Selection on human genes as revealed by comparisons to chimpanzee cDNA. Genome Res. 2003 May;13(5):831-7.

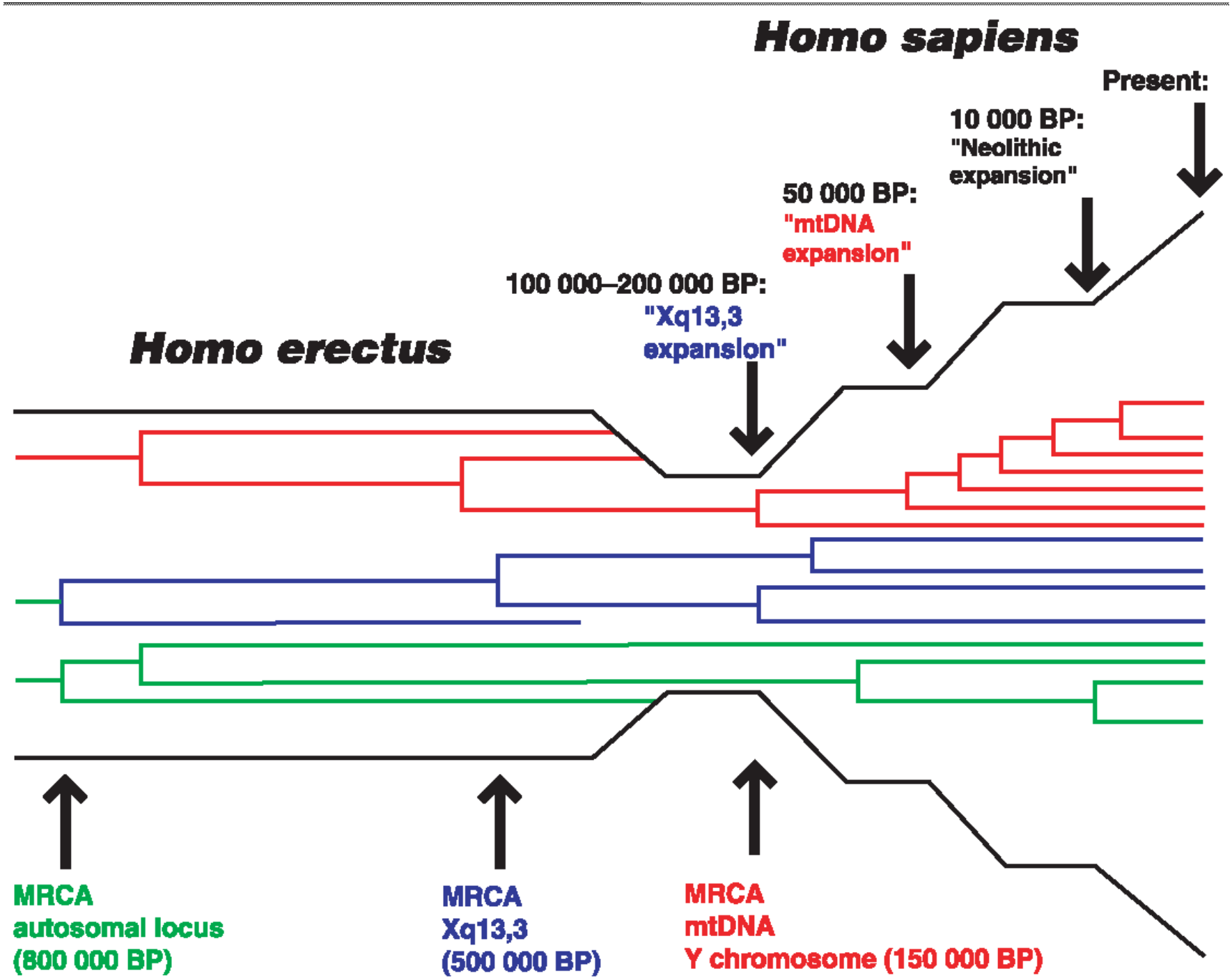
Chimp @ UCSC


Base Position
115000000
115500000
116000000
116500000
117000000
Gap
Gap Locations
Quality Scores
Chimp mRNAs
Chimp mRNAs from GenBank
Human Chain
Human (July 2003/hg16) Chained Alignments
Human Net
Human (July 2003/hg16) Net
Human Best
Human (July 2003/hg16) Best Alignments from Net
Deletions in Human (July 2003/hg16) Relative to Chimp
HD8163_342
HD8161_280
HD8160_165
HD8156_271
HD8154_186
HD8152_316
HD8151_175
HD8162_265
HD8159_99
HD8157_281
HD8153_316
HD8158_106
HD8155_2247
RepeatMasker
Repeating Elements by RepeatMasker
Simple Repeats
Simple Tandem Repeats by TRF
115750000
115800000
115850000
115900000
115950000
chr7 + 99k
chr3 - 71211k
chr3 - 71198k
chr7 + 113645k
chr7 + 113670k
chr15 - 54798k
chr3 - 71022k
chr3 - 70946k
chr3 - 71017k
chr3 - 70942k
chr7 + 113865k
chr7 + 113866k
chr7 - 113574k
chr7 + 113674k
chr7 + 113673k
Score
HD8156_271

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
MRCA
autosomal locus
(800 000 BP)
MRCA
Xq13,3
(500 000 BP)
MRCA
mtDNA
Y chromosome (150 000 BP)

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

Repetetivní elementy

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

HGSC (2001) Nature 409: 860-921

DNA transpozóny

2-3 kb

terminální reverzní repetice (50 - 100 b)

cut-and-paste mechanismus kopírování

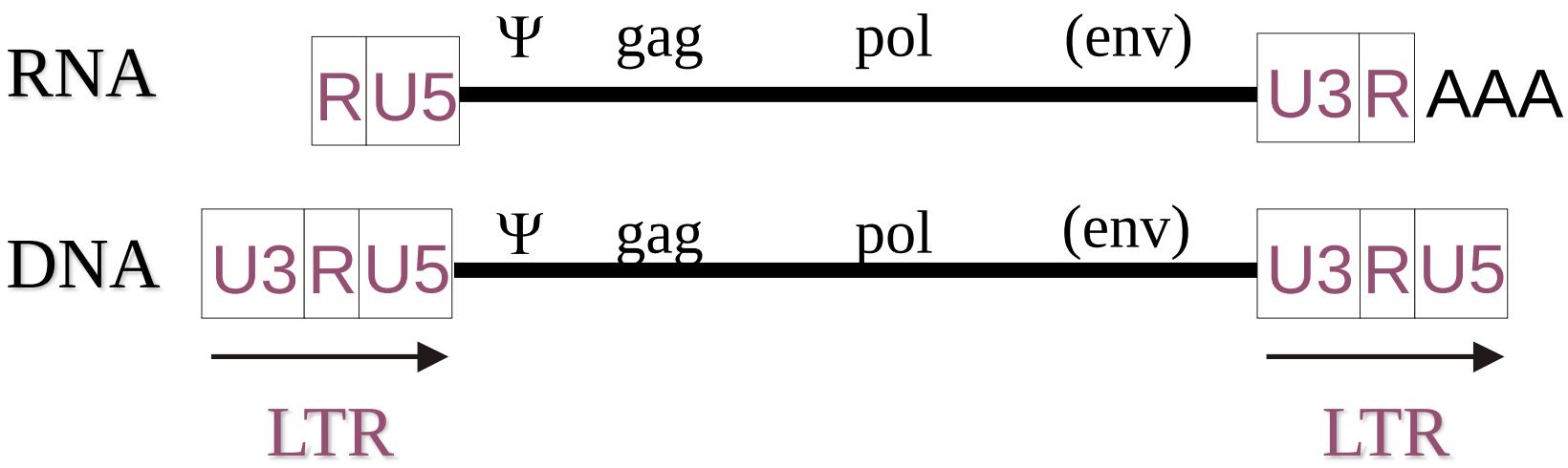
3% genomu

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

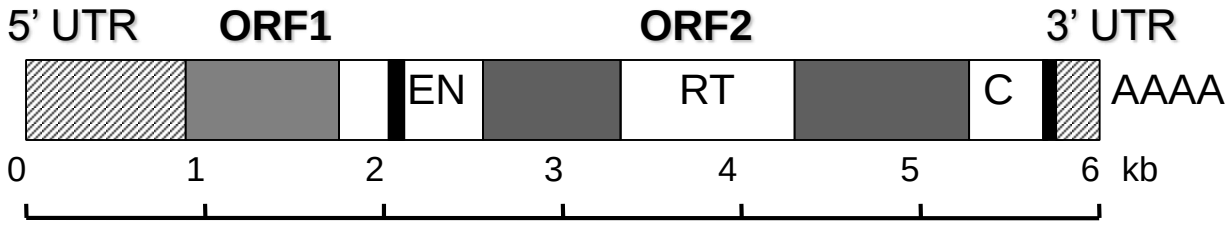
retrotranspozóny

(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

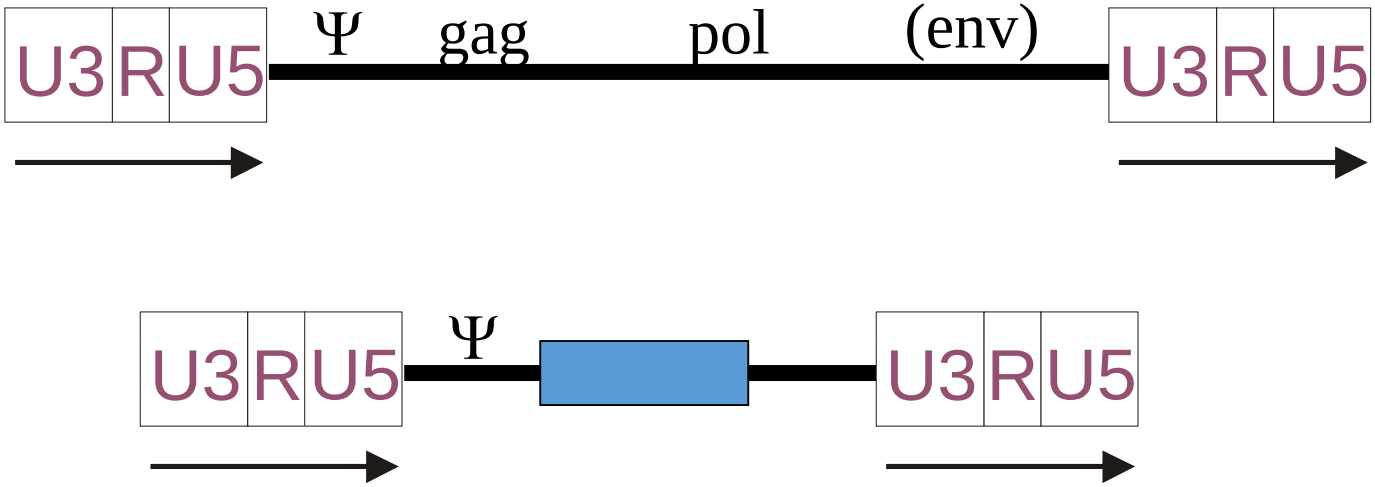
pro každou třídu autonomních elementů existuje neautonomní element, který používá mechanismus replikace „svého“ autonomního elementu

DNA transpozóny

2-3 kb; terminální reverzní repetice

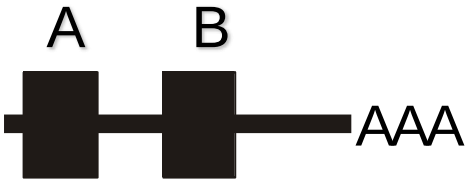
80-3000 bp; terminální reverzní repetice

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
inserční duplikace (5-15 bp)
inserční preference (TT|AAAA)
10 % genomu
1 000 000 elementů, často zkrácených na 5' konci

Procesované pseudogeny

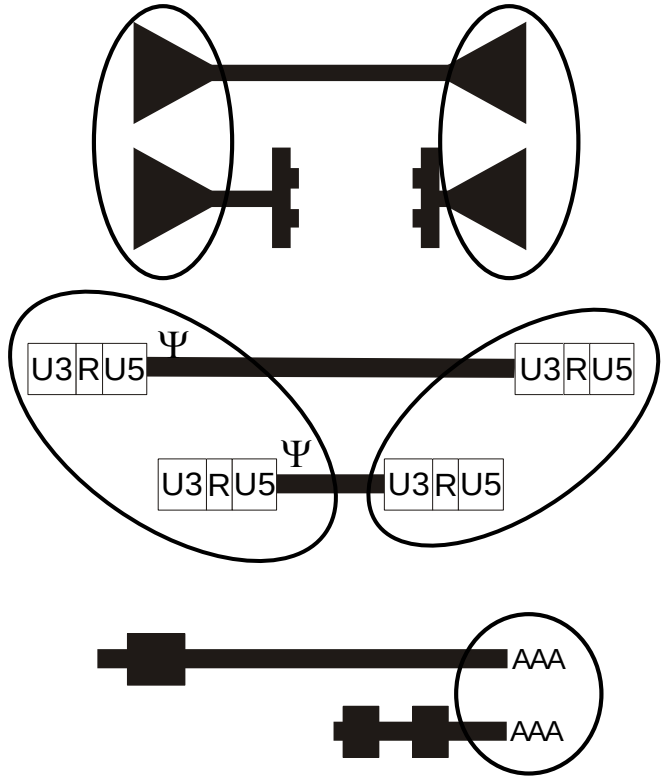
kolineární s mRNA, chybějí introny a promotory
poly A
často zkrácené na 5’ konci
krátké inzerční duplikace

Koevoluce parazitů

DNA

LTR

polyA

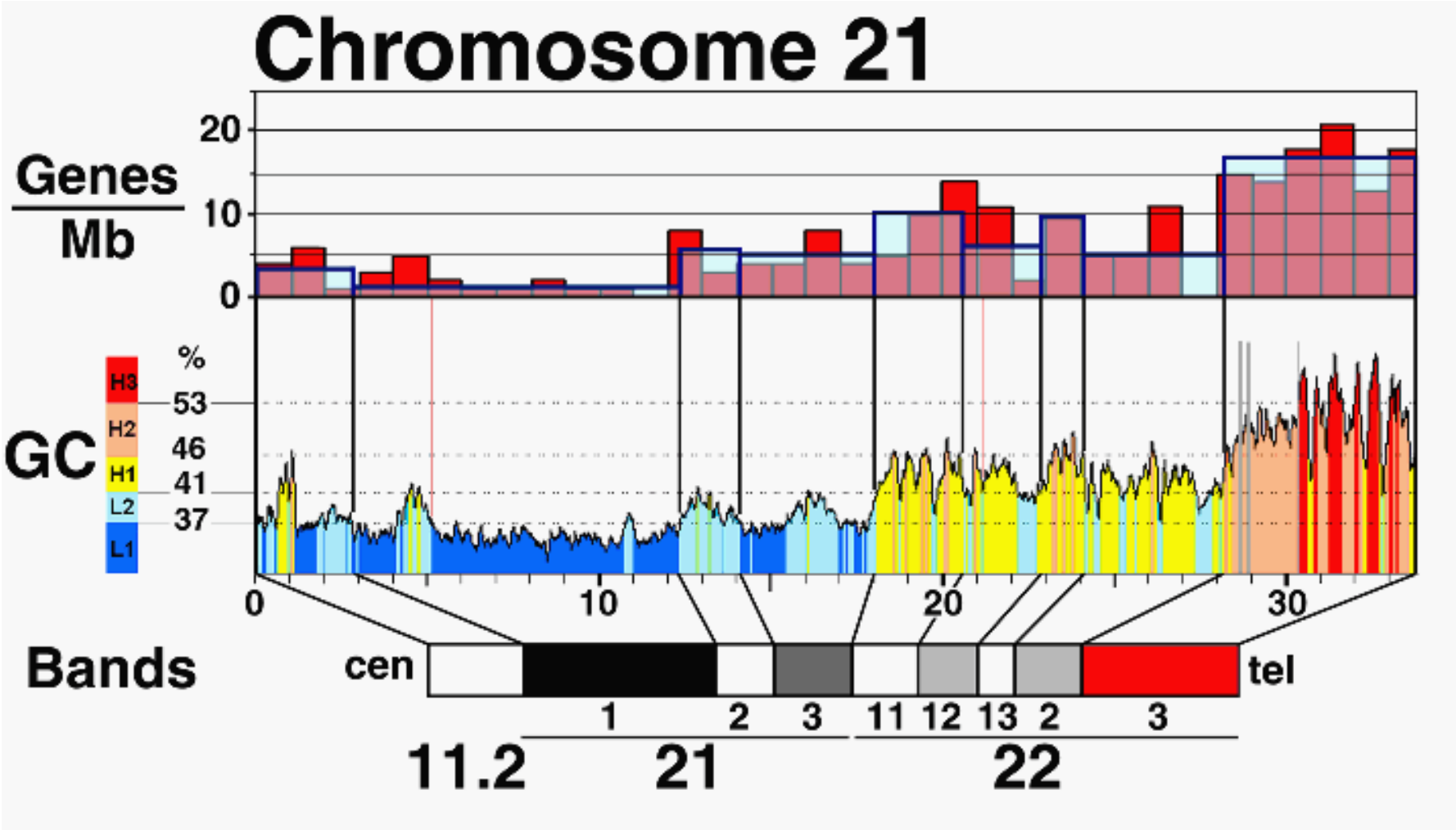


čím to začalo?

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