

Projekt lidského genomu

Úkolem je zjistit kompletní genomovou sekvenci.

Problém:

Genom je velký!

3 miliardy párů bází

1000 telefonních seznamů



Genomová sekvenace

6/ 25/ 04

1128 genomových projektů:

199 kompletních (včetně 28 eukaryontních)

508 prokaryotických genomů před dokončením

421 eukaryotických genomů před dokončením

nejmenší: archaeobacterium *Nanoarchaeum equitans* **500 kb**

Bacillus anthracis (anthrax) **5228 kb**

S. cerevisiae (kvasinka) **12,069 kb**

Arabidopsis thaliana **115,428 kb**

Drosophila melanogaster (octomilka) **137,000 kb**

Anopheles gambiae **278,000 kb**

Oryza sativa (rýže) **420,000 kb**

Mus musculus (myš) **2,493,000 kb**

Homo sapiens (člověk) **2,900,000 kb**

S. cerevisiae

200x

H. sapiens

200x

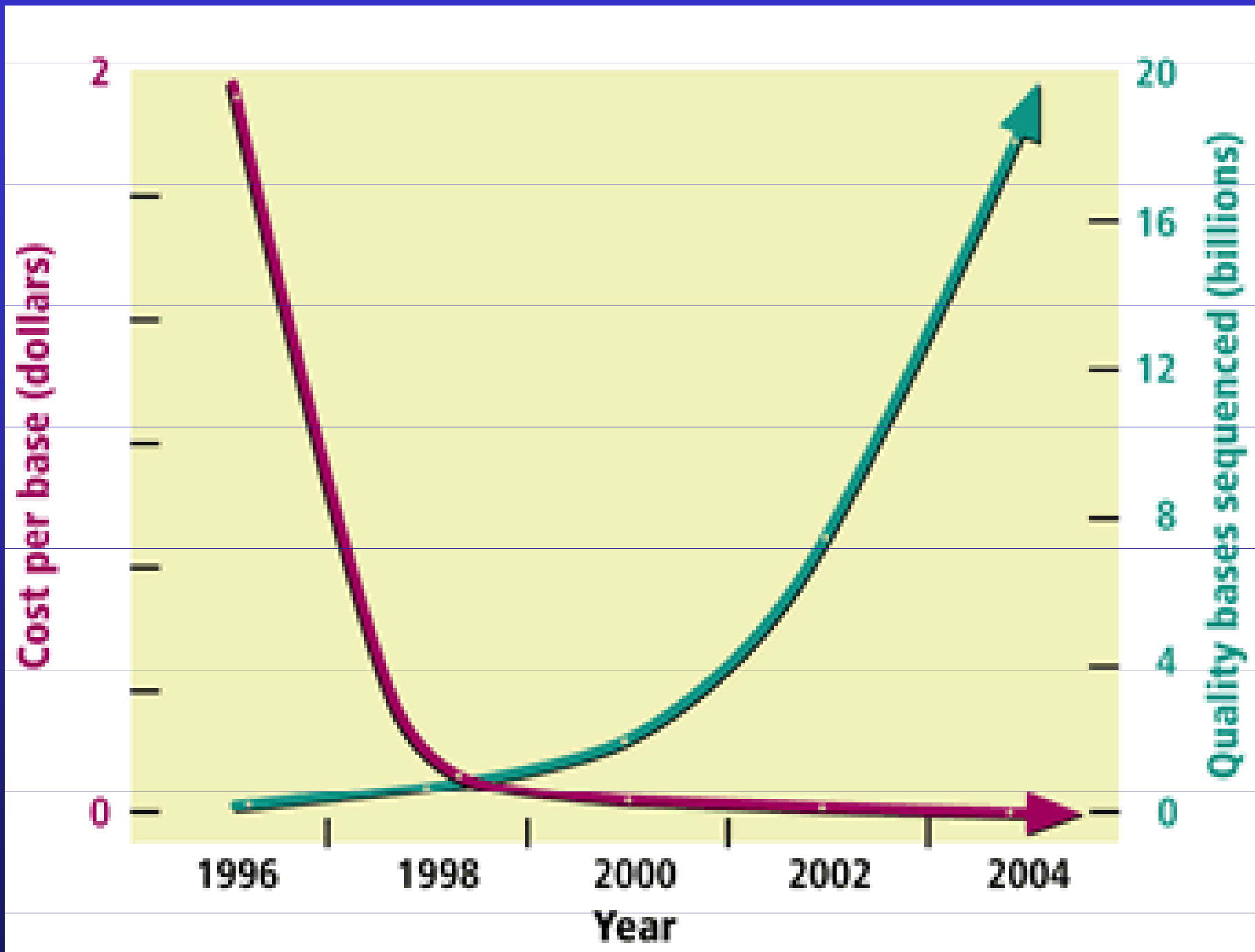
A. dubia

1980 - \$10/bp

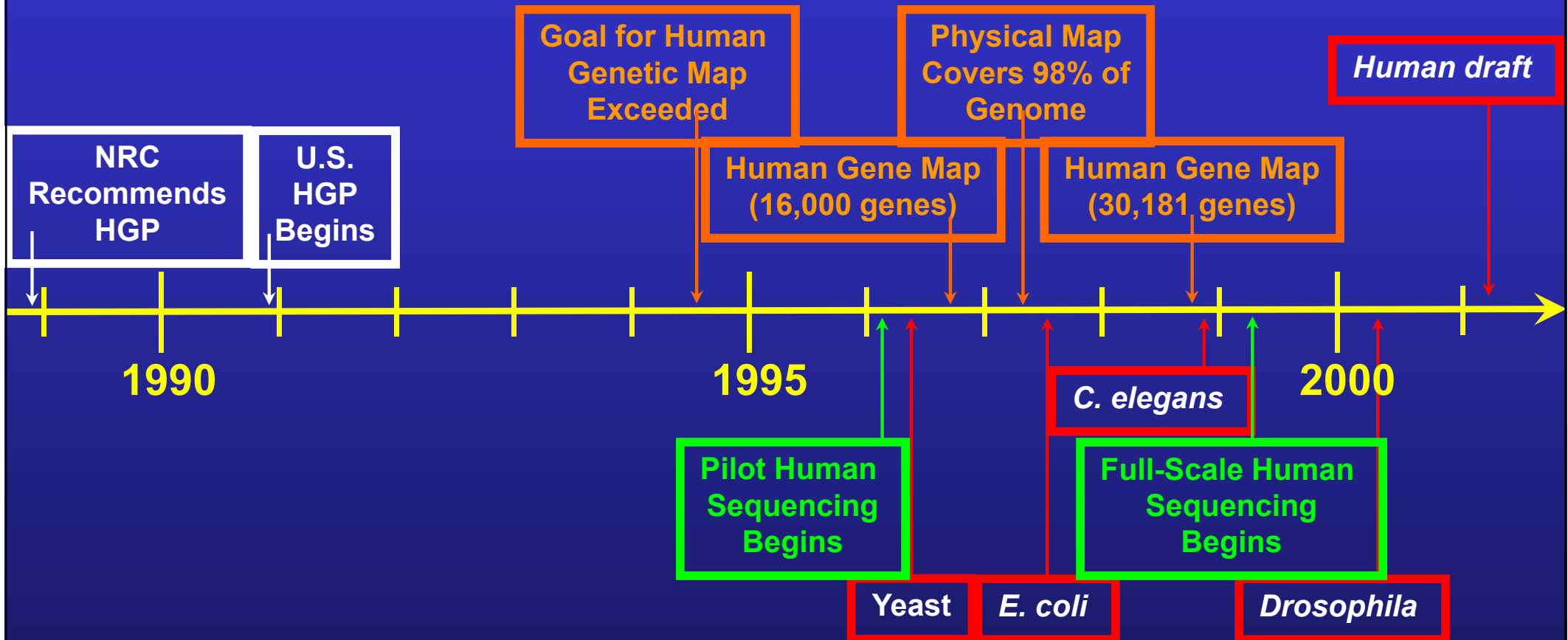
2001 - \$0.1/bp

2006 - \$0.01/bp

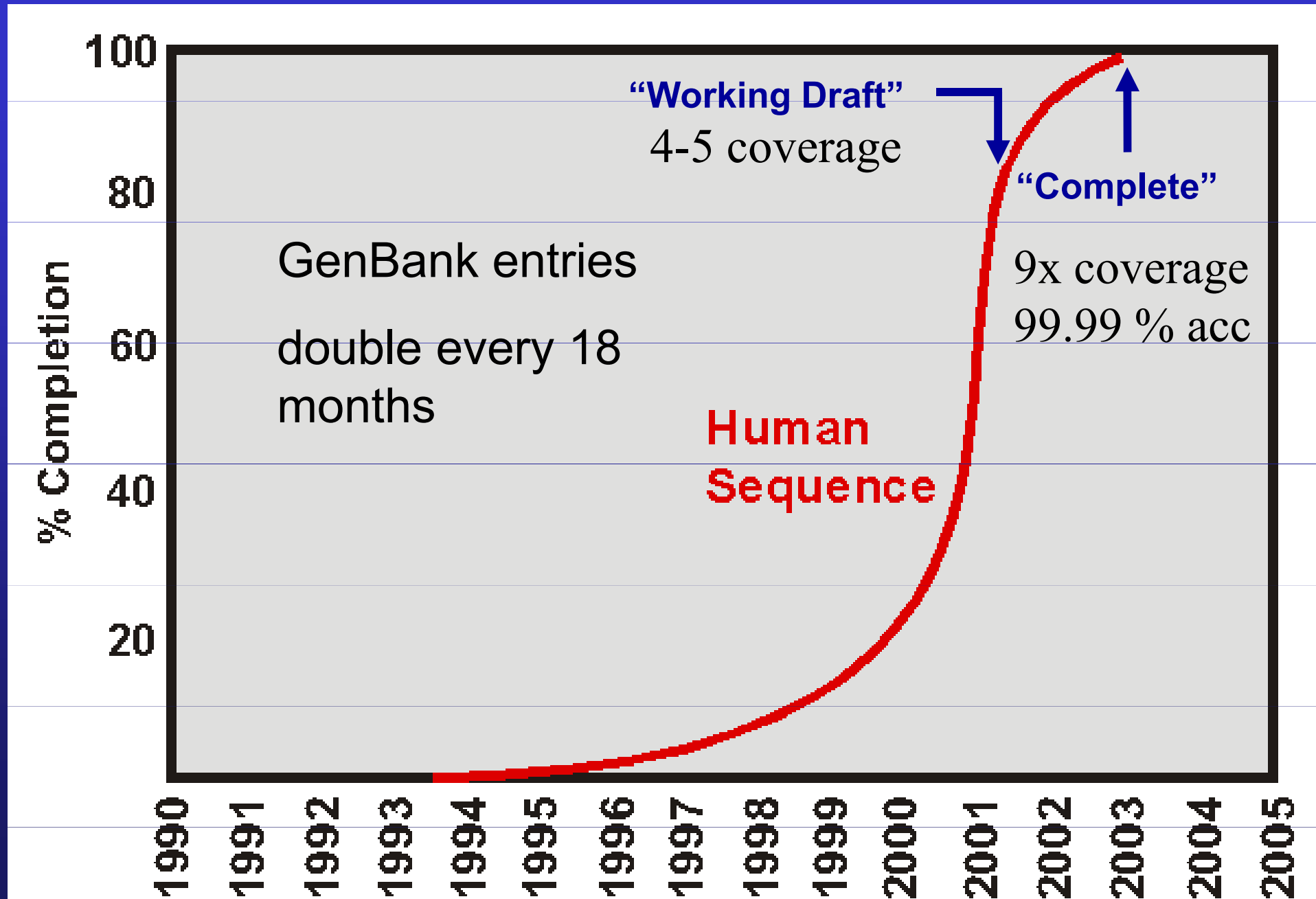
[http:// www. genomesonline. org/](http://www.genomesonline.org/)



Projekt sekvenace lidského genomu - HGP



Kompletace lidské genomové sekvence



Kompletace lidské genomové sekvence

International Human Genome Sequencing Consortium.

Finishing the euchromatic sequence of the human genome.

Nature **2004** Oct 21;431(7011):931-45.

„The current genome sequence (Build 35) contains **2.85** billion nucleotides interrupted by only **341** gaps.

It covers approximately **99%** of the euchromatic genome and is accurate to an error rate of approximately **1 event per 100,000** bases.

Human genome seems to encode only **20,000-25,000** protein-coding genes“

2.85 miliard nt a **341** neosekvencovaných oblastí.

1 chyba na **100 000** nt.

20 000-25 000 genů kódujících proteiny.

Pracoviště, které osekvencovaly 85% genomové sekvence

- 1. Whitehead Institute for Biomedical Research, Center for Genome Research, Cambridge, MA**
- 2. The Sanger Centre, Cambridge, UK**
- 3. Washington University Genome Sequencing Center, St. Louis, MI**
- 4. US Department of Energy, JGI, Walnut Creek, CA**
- 5. Baylor College of Medicine Human Genome Sequencing Center, Houston, TX**

USA, UK, Japan, Germany, China, France

Sekvenace genomu

Genom: 3 Gb



Štěpit genom na větší kusy DNA

Klonovat do BACs: 100 kb

Mapování BAC klonů podél chromosomů



Znovu štěpit



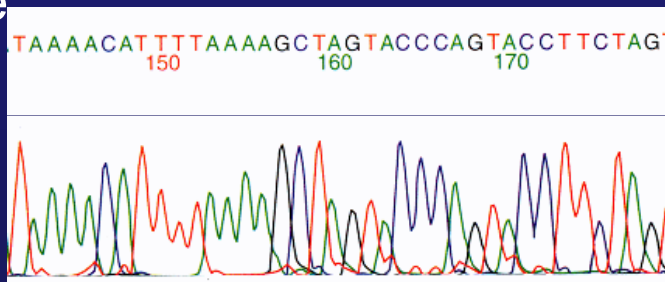
Dokončit sekvenci

...TTGTAAGTGAGAACAGGACGTATGTGGTTTTCTACTCCTGTGTT...

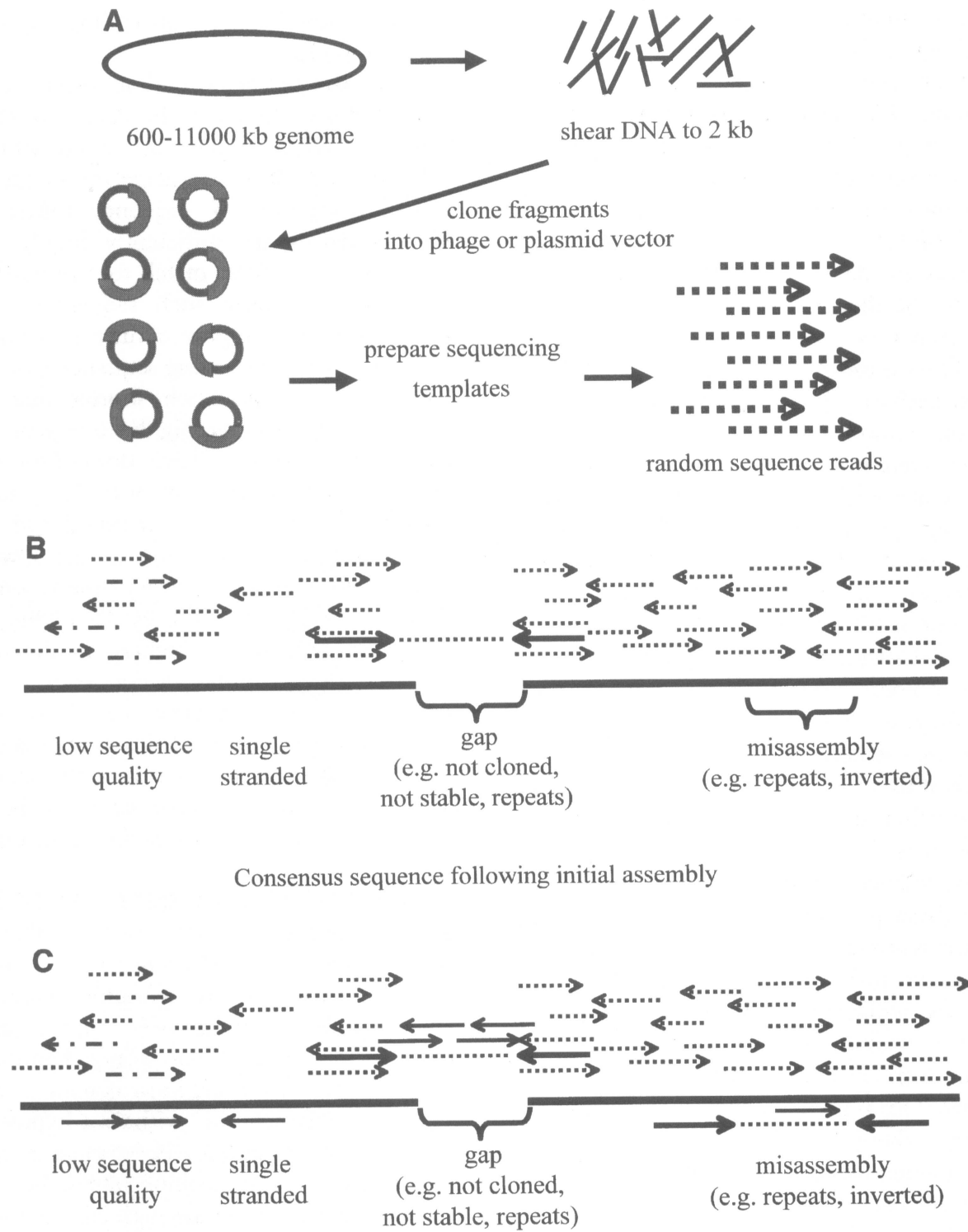
uspořádat BAC sekvence

TTGTAAGTGAGAAC
AGAACAGGACGTATGTGGT
TGTGGTTTTCTACTCC
CTACTCCTGTGTT

Sekvenace



Problémy při sekvenaci



Princip sekvenace

a

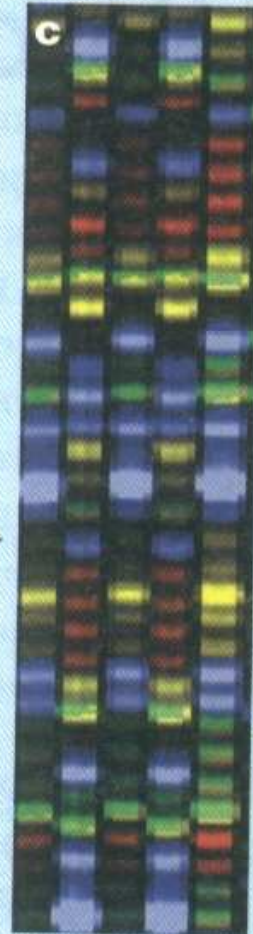
DNA polymerase

5'-TGGGGCTAACAAAGCAAATGATCTGTAGT
3'-ACCCGATTGTTTCGTTTACTAGACATCAATTGTCT-5'

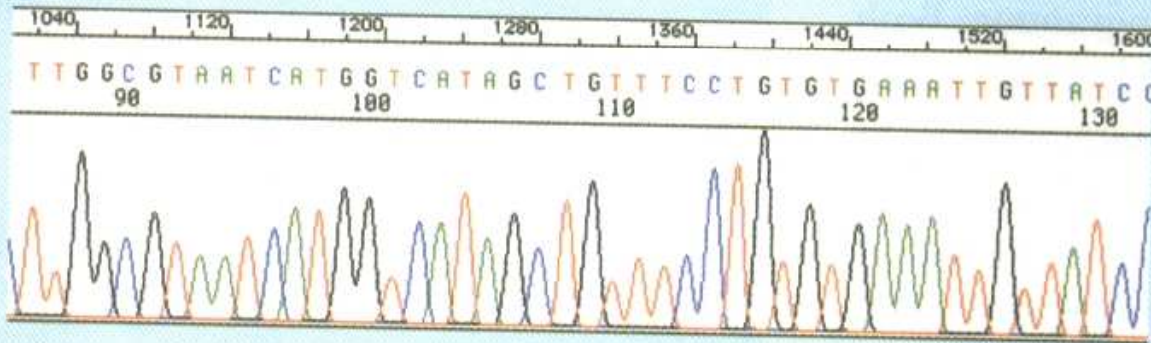


b

TGGGGCTAACAAAGCAAATGATCTGTAGT ●
TGGGGCTAACAAAGCAAATGATCTGTAG ●
TGGGGCTAACAAAGCAAATGATCTGTA ●
TGGGGCTAACAAAGCAAATGATCTGT ●
TGGGGCTAACAAAGCAAATGATCTG ●
TGGGGCTAACAAAGCAAATGATCT ●
TGGGGCTAACAAAGCAAATGATC ●
TGGGGCTAACAAAGCAAATGAT ●
TGGGGCTAACAAAGCAAATGA ●



d



Sekvenační gel



Kousek sekvence chromosomu 21

TCACAATTTAGACATCTAGTCTTCCACTTAAGCATATTTAGATTGTTTTCCAGTTTTCCAGCTTTTATGACTAAATCTTCTAAAATTTGTTTTCCCTAAATGTATATTTAAATTTGTCTCAGGAGTAGAATTTCTGAGTCATAAAGCGGT
CATATGTATAAATTTTAGTGCCTCATAGCTTCCAAATAGTCATCCATTTTATACATCCAGGCAATATATGAGAGTCTTGGTGCTCCACATCTTAGCTAGGATTTGATGTCAACCAGTCTCTTAAATTTAGATATCTAGTACAT
ACAAAATAACCTCAGTGTAACTCTGTTTGTATTTCCCTTGATTAAGTATGCTGAGCAGATCTTCAATGCTTATTGACCATTAATTAGTCTTATTTGTTAAATGTCTCAAATATTTTATACAGTTTTACATTGTGTATTTCATT
TTTTAAAAAATTCATTTTAGGTTATATGTATGTGTGTCAAAGTGTGTGTACATCTATTTGATATATGTATGTCTATATATTTCTGGATACCATCTCTGTTTTCATGCATTGCATATATATTTGCTTATTTAGTGGTTTTATCTTTTCAT
TTTTCTTTTGGTATCTTTTCATTAGAAATGTTATTTATTTTGTAGTAAAGTAAACATTTAATATATTTCTGTAAACATTTAATGAATCATTTTATGTTATGTTTGTATTTAAATTTCTGAAAAATTTCTATGTATTTCTACTAGAATTTGTCATAA
TTTTATCTTTTATATACATTGATATTTTATGTCAAATATGTAGGTATGTGATATTTATGCACATGGTTTTAATTCAGTTAATTTGTTCTTCCAGATGTTTTGTACCATTCCAACATCATTTAAATCATTTAAATGAAAAGCCTTTCCCTTAC
TAGCTAGCCAGCTTTGAAAATCCATTCATAGGGTTTTGTGTTAATATATTTTTGTTCTTTTTTTTCCCTTCTACTGATCTCTTTATATTAATACCTACTGTGGCTTTTATGAAGTCATGGAATAATACGTAGTAAGCCCTCTAACACT
GTTCTGTTACTGTTGTTATTGTTTTCTCAGGGTACTTTGAAAATATTCGAGATTTTATTATTTTTTAGTAGCCTAGATTTCAAGATTGTTTTGACGATCAATTTTTGAAATCAATTTGTCAATATTTTTTAGTAATAAAATGATGATTTTTG
ATTGAAAATACATTTAAATCTATAAGCCAAATTTGGAGATTATTGATATATTAACAAAAATGAGTTTTCCAGTCCATGAATGTATGCACATTTAAAAATCATTTCTTAAGTATGTCAATTTTTTAAGTTTTAGTTTTGAGCAGTATATGTTTT
GTTACATAGGTAACCTCTGTCTATGGGGTTAGTTGTACAGTTATTTTTATCATCCAGGCATAAAGCCAGTACCAGTAGTTATCTTTTTCTGCTCTCCCTCTGTCACCCTCCACTCTCAAGTAGACCCAGTTTTCTGTGTTCT
TCTTCTTTGCAATTAAGTCTCATCATTTAGATTGCACCTGTAAGTGAGAACAGGACGATGTGGTTTTCTACTCTGTGTAGTTGCTAAGTAACACCTCCATCCATGTTCCACAAAAGACATGATCTCCTTTTT
TTATGGCTGCATATTTATCCATGGTATATATGTACCACATTTTCTTTTATCCAATCTGTCATTTGATGGACATTTAGTTGTTTTCCACATCATTTGCCGTTGTAATACTGCTGCAGTGAATATTCGTGTGTATGTCTTTTATGGTAGAATG
ATTTATATTTCTCTGGGTATATTTCCAAGTAATGGGATGGTTGGGTCAAATGGTAATTTCTGCTTTTAGCTTTTTGAGGAATGGCCATATTTGCCCTTCCACAACGGTTGAACATAATTTATACTCCCAAGAGTGTATAAGTTGTTCCCTTTT
TCTCTGCAACCTCGACATCACCTGTTATTTATGACTTTTATATAAAGCCATTTCTGCTGGTCTGAGATGGTATCTCATTATGATTTTGATTTGCATTTCTCTAATGCTCAGTGAATTTGAGCTTTGGCTGCATATATGTCTTTTAA
AAATATCTGTTTATGCTTTTGCCTAATTTATAACGGGGTGTGTTTTGTTTTCTCTTGTAATTTGTTTTAAGTTCCCTTATAGATTTCTAGGTATTTAAACCTTTTTTCCAGAGCGTGGCTTTGCAAAATTTTTCTCCATTTCTATAGGTTGT
CTGTTTTATCTGTTGATAGTTTTCCCTTGCTGTGCAGAGCTCTTAACTTTAATTTAGATCCGACTTGTCAATTTTTGCTTTGGTGCAATTTGCTTTTGGATGTTATTGTGCTGAAATCTTTGCTAGTTCTTAGGTCAGGATGATATTCG
CCAAGTTGCTTCCAGGGCTTTTATAATTTTGGATTTTACATTTAAGTCTTAAATATATTTTATAAATTTGTTAGGGTTTTCAGGATACAAGGACAATATAGCAGCAAAACAATGTAAGAAAGTAAATCTGAAAAATAATAGAAAACAGTTT
AATGAACTTTTACCATTATGTAATGCCCTTCTTTGTCTTTCCCTGATCTTTGTTGGTTTTGAAAGTTCAAAAAAGACAAACTTAATGGTACAATAGGATTTGTAGATTTTCCAGGACTTTCTGTATAAAAATATTTGTATATATGAATAGA
TCATTTTTTATTTTACAGCTTTTAAACATTTTCTTAAACATTTTCTTATTTGCTTCACTCTGCTAGGACTTCCAGTGTGTAACAGAAATTTTGTATATTTTGTATATTTTGTATATTTTGTATATTTTGTATATTTTGTATATTTTGT
TATTTTCAACATAAAGCAGCATGTTAATGGCAAGTTTTAATGCTGTATCAGATTTGAAATTTTTTGTATATTTTGTATATTTTGTATATTTTGTATATTTTGTATATTTTGTATATTTTGTATATTTTGTATATTTTGTATATTTTGT
ATTTTTTTTATTTCTCTTTTACCTTAAAGCATTAACCTGACTACGATTTGAATATAAACA
TTAGACATTTTTGCGTCGATATTTATGAGGGATGTTGATCTGTAAAAGTATTTAAAATGCCCTTTGACAGATTTGATAGCAGTGTATTCTGGCCTAATAAATCAAACCTGAGGTATGATCTTCCCTTTTCTATTTCTTAAATAGCATTTT
TAAAATTTGGTGGTTTTTTCCCTTCTTAGTGAATTTTACCAGCAAAAGTAAACAGGCCTTATATTTCTCTTGTGGAATAATTTAATTTCAAATTAATGGTATTTTGTCTTGTAGGGTGGTAATTTTCTCTGTGTTTGGTCTTAAATGGAC
TCTTAGCTGATCACCGATTACTCAGCGAGGTCTTCTCACTCTGGAAGAGCTGGAACCTCCAGTGTGTTTTAGTGCAGCATGACCAGGGTATTTACCCTTCAACATTTAGGCTTTTATCAGTGAATACTATTTGTCCTCATGGAGTTTTT
GCCGCTGGCCTACACAGTTTTAGGCTTCCAGCTTAGAACACATAATGAATCTTATGACAGATTTCTGCCACCTTTGACCTTTTATGATTTCTCTTCTTGGGTAAGCTGCCTTATTAATCTGATACACTTCCAGCAGTCCAGAACTACA
CTCTTTCCCTTCTCTGCTTGGAGATGACTCTTTTGTCTGAGATTCACCTTTGCTGTGCTGAAAAAGAAAAGTGCCTCAAGGAAGATACCAGGAAAAATCACAGGGCTCATTATGATTTTCTTCTTTTCAAGGACTACAGCTTTGT
GTTGCCATGTTCAATTTCTGAAAAATAATAGAGCATATATACCTGTGTGAGAAGGCAATCCAGACAGTTAGTTTGTATGACTAGAAGCAGAAGTCTACATGGAGAATTTTACTTAACTGTGTTATAGTTTCTTTAATATTTCAA
GATATGTTAATGTTCCACAGATCTCATTTCTATAAATCTTTATCATCTTAGAGCTCTGATACTATTTAGAATTTACTATTTCCCTTCAAATAAGAGATTAGAAACAGGGTTATATTTGGGGTAGGTTGACTTACTTTTCTGGGAACCAAA
GCATATTAATGACCAGTTTTAACACACTTCTATGTATGCACAAAAGATATATTTACATTTCTGCAAAATCATTTTCCCTTTTGAATTTGAAAAGGATCTTTGGTATACAGATATTTCAATAGCCAGCTGAAGATTTTGAAT
TCATTTAATGTTTAGATTTCACTACATGAAATGATCCAGAAAGAGTACTCAAATATAAGTATCTATAACGATGGAATATACATCTCCACTGCCCAAGATGGTAGTCAAGTCAATATTGATCATGTGAGACGTGGCAAGTGTACT
CAGGGTCTCAATATTTAAATGTATTAAGCTTTAATTAATGTAAATTTGAATTTAGCAAAACATGTATAGCTTGTGGTTACTGTTTTATTAGTGCCTAATATAGAACATTTCCATGATTACAGAAAAGTTATCTTAGAATACTCAGTTCT
GGACTATTTTATCTGGCTAAATTTAAATGTTAAAATATTTACAAATTCATCTTCCAGGCTGGCTGTTGAATATTTTTATAGCAAAAGTCAATTTAATAATTTAAAATCAAATAATTTATCTTTTTCAATATGTAATAATATGTCTTTACATAT
TCTACTCCCTTCTTACATACATATTTCTGATGTAACATAGGTATTTCTTATTCAATGCACACTGAAATGACAACATAAATAATTTTACTAAGTGTCAACATATAAAAACTTTGAACAAAATCAGATTATATCACTGTGGATATTTCTA
TTTTGAACTAACTTAGATGATAATTTAATCTATATCTTAGATGAACTTTAAATCAATAAAATCTCTCAATGGTGTATAAATCTCAAGCCATTAGCCACTGATTATCCATTTTTATTCTTTTCAATATTAATTTTATTGCCATGTAT
GAATGCTGTAGCATCCATGTTTAAATACTAGTTAACAATAATGCACCTGGCATCAGATACAATAAGGATGAAATGAGATATAATTTAGGACTCTGGTAACACACATAAAATTTGGAAGATACCCCTGAAATTCAGCAAGAGATATTTAT
CCAGCTTATTTTATTTTGTAGACAGAGTCTTGGCTCTCTCACTCAGGCTGGAGTGCAGTGGACCATTCTAGGCTCGCTCCAACCTCTGTCTCCCAATTTGAAGTAATTTCTGTGCTCAATCTCCGAGTAGCTGGGATTACAGGCATGT
GTCACCAAGCCTGGCTGATTTTTGTAGTTTTAGTAGAGACGGGTTTTCCACATGATGGCCAGGCTGGTCTTGAACCTCTGGCTCAAGTGAACCTCGGCTCCTAAAGTGTGGGATTACAGACGAGAGCCACTGAACAGC
TTTGTACCAACTTATTTGGATGAATGAGTTACATATTTTACATTAATCTGTTATTGTGATAATTTCTCATGTTATTTTCCATGTATAGATTTATATAATAAGTAAATTTAATTTTTTTTCCAGGAGATATAACAACAACAATATTT
TATAAACAGGATAAATAAATAAGCAAAAATTTGTTGAAATGTCTTCAATTTGACTACTAATTTTTTACATGTTTGTACTTTGAAGTGTATCAATACTTGTGATGTATACAATTAAGTAAAGATTTAAAGATGCCATTTTTAACT
TATTATGACACAAAGTCTATAAATTTTATATTTTGTAGATTTGTATTTAAATAACTTGTGAAATTTAATTTTTAAAATAAAATTTCTTCTATGGATTGGTCTTCAATCGAGGCATAAAAAGGAATATAACAGTGTGGCCTATAACTTC
TATATTTGAATTTCTATATTTTAAACACAATTAATAATTTTGTCTAATGAATTTGTAATGTTTTTAAAAGCTAGGTGAATTTTATTTAAATTTTATTTACATGGCGATAACACAGAGAAAAATTTTTGGGATTTCTTTAAAATGGTATGTAC
AAAAGCTTAAAGTTGTTATGTAGTGGCAGAGATAAAAAGTAAAACAAAAAAGCTTAAAAGTTTTGCTTTACTATTTATAGGCTCATAAGTGAAGTGTGCCAGAAAATGAAAAAGAAAGGAGAGAAATATAAATAACTGTGTGG
AAAACACAGATAAAGCATAAAGATAGAATAAAGATAGAAGCATTTAATATGAGGCAGTGTATGGCTTTTTGAAGAATCCCAACTAAGGACCTACTTTTATGTTAATAAATAATATGTTTCTAATCCCTATATTGTCCACAGCAACCT
TTTTAGGACATGGAGCAGTACTATGAGTGCAGAAAGCAAGAGTGAAGCAATTTGAAAAATCATGAACACTAGTTTGTAAAATCTCACTGAGATATAATATCTGTTTGCCTTCACTTAGAATTTAATGTCTTGGGGCTGGGA

PHASE : INTERPRETATION
TWO :

SEEMAN The Star-Ledger



Co lze nalézt v genomu?

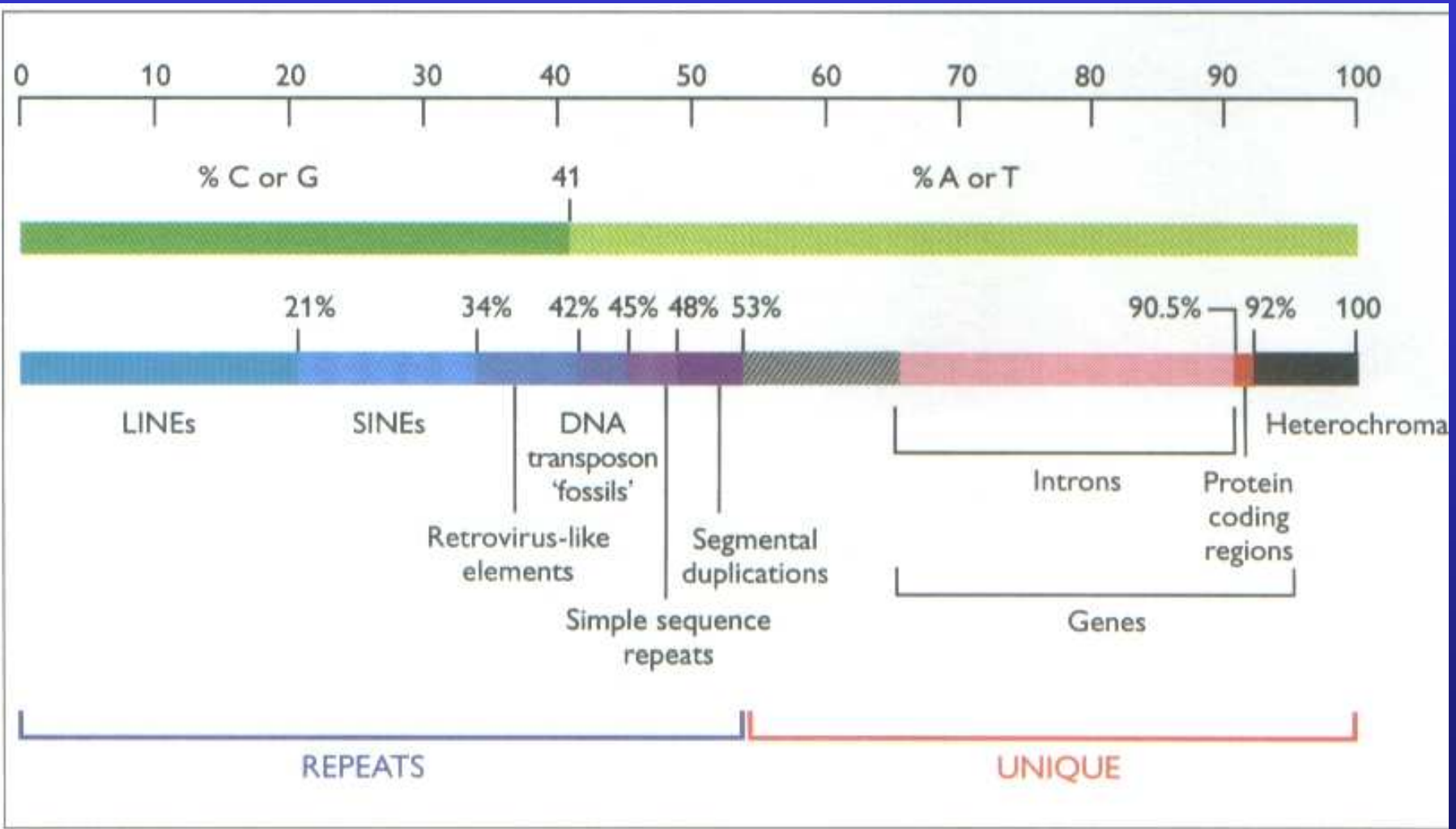
Geny (tj. protein kódující oblasti)

jen <2% genomu kóduje proteiny

- geny pro nekódující RNA (rRNA, tRNA, miRNAs, atp.)
- strukturální sekvence (scaffold attachment regions)

Regulační sekvence



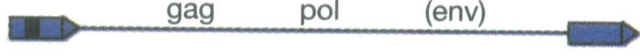



- “junk” (zahrnující transposony, retroviry, atp.)



REPEATS

UNIQUE

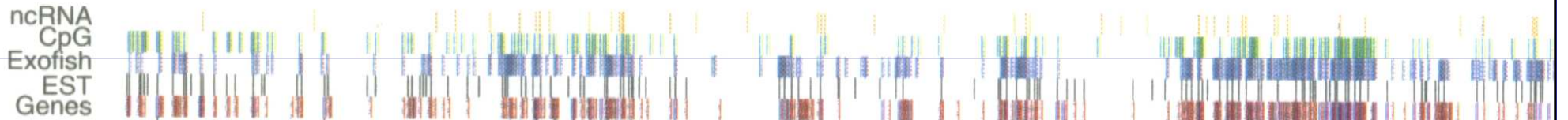
Classes of interspersed repeat in the human genome

			Length	Copy number	Fraction of genome
LINEs	Autonomous		6–8 kb	850,000	21%
	Non-autonomous		100–300 bp		
Retrovirus-like elements	Autonomous		6–11 kb	450,000	8%
	Non-autonomous		1.5–3 kb		
DNA transposon fossils	Autonomous		2–3 kb	300,000	3%
	Non-autonomous		80–3,000 bp		



Chromosome 2

Mb 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32



- Known genes (OMIM)**
- TPO
 - TSSC1
 - ADAM17
 - YWHAO
 - HPCAL1
 - LBP-32
 - P5
 - ODC1
 - ROCK2
 - KCNF1
 - GS39555
 - PRO2958
 - DDX1
 - MYCN
 - VSNL1
 - KCNS3
 - MBNT
 - SDC1
 - ARHB
 - APOB
 - PRO0397
 - FKBP1B
 - PIG3
 - PRO2015
 - SRC1
 - ADCY3
 - POMC
 - DNMT3A
 - DTNB
 - KIF3C
 - HADHB
 - HADHA
 - CENPA
 - KCNK3
 - MAPRE3
 - CRMP5
 - KHK
 - EMILIN
 - PREB
 - CGR11
 - APR-3
 - SLC5A6
 - SLC30A3
 - CAD
 - UCN
 - MPV17.GTF3C2
 - GCKR
 - PPM1G
 - NRBP
 - BRE
 - FOSL2
 - PPP1CB
 - ALX

Shrnutí

- Variabilní distribuce řady parametrů (GC, CpG islands, repetice)
- 20.000-25.000 protein-kódujících genů
- Proteom je mnohem komplexnější než u bezobratlých
- Stovky genů mají původ v horizontálním transferu
- V genomu se vyskytuje asi 10 miliónů SNP

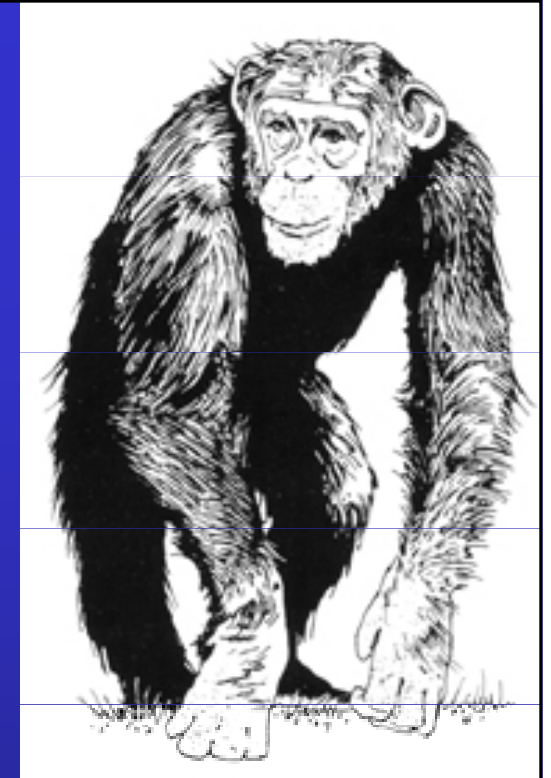
Aplikace znalostí lidského genomu

- **Geny podmiňující gen. choroby** – poziční klonování (30 genů)
- **Paralogní geny** (achromatopsie, CNGA3, **CNGB3**); (971 známých genů => 286 paralogních genů)
- **Cíle zásahu medikamentů** – recentní kompendium = 483 cílů, **18** nově identifikovaných; (Alzheimer's disease, β -amyloid is generated by processing APP by BACE; BACE2 in obligatory Down's syndrom region of chromosome 21)
- **Obecná biologie** – hořká chuť – nová rodina G-proteinových receptorů

Sekvenace genomu šimpanze

Chimpanzee Sequencing and Analysis Consortium

Initial sequence of the chimpanzee genome and comparison with the human genome.



Nature 2005 Sep 1;437(7055):69-87.

Thirty-five million single-nucleotide changes, five million insertion/deletion events, and various chromosomal rearrangements.

98,6 % identity to human genome sequence

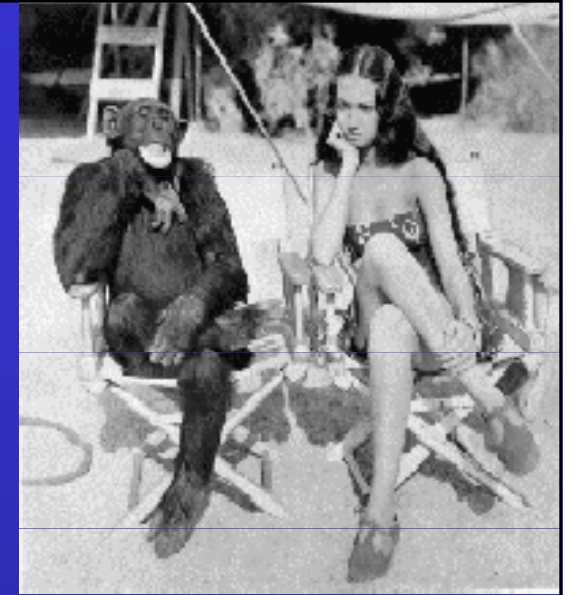
Differences in gene/exon structures

35 miliónů záměn nt, 5 miliónů inzercí, delecí a dalších změn

96% identita s lidskou genomovou sekvencí

Změny ve struktuře genů popř. exonů

Rozdíly mezi lidmi a dalšími primáty



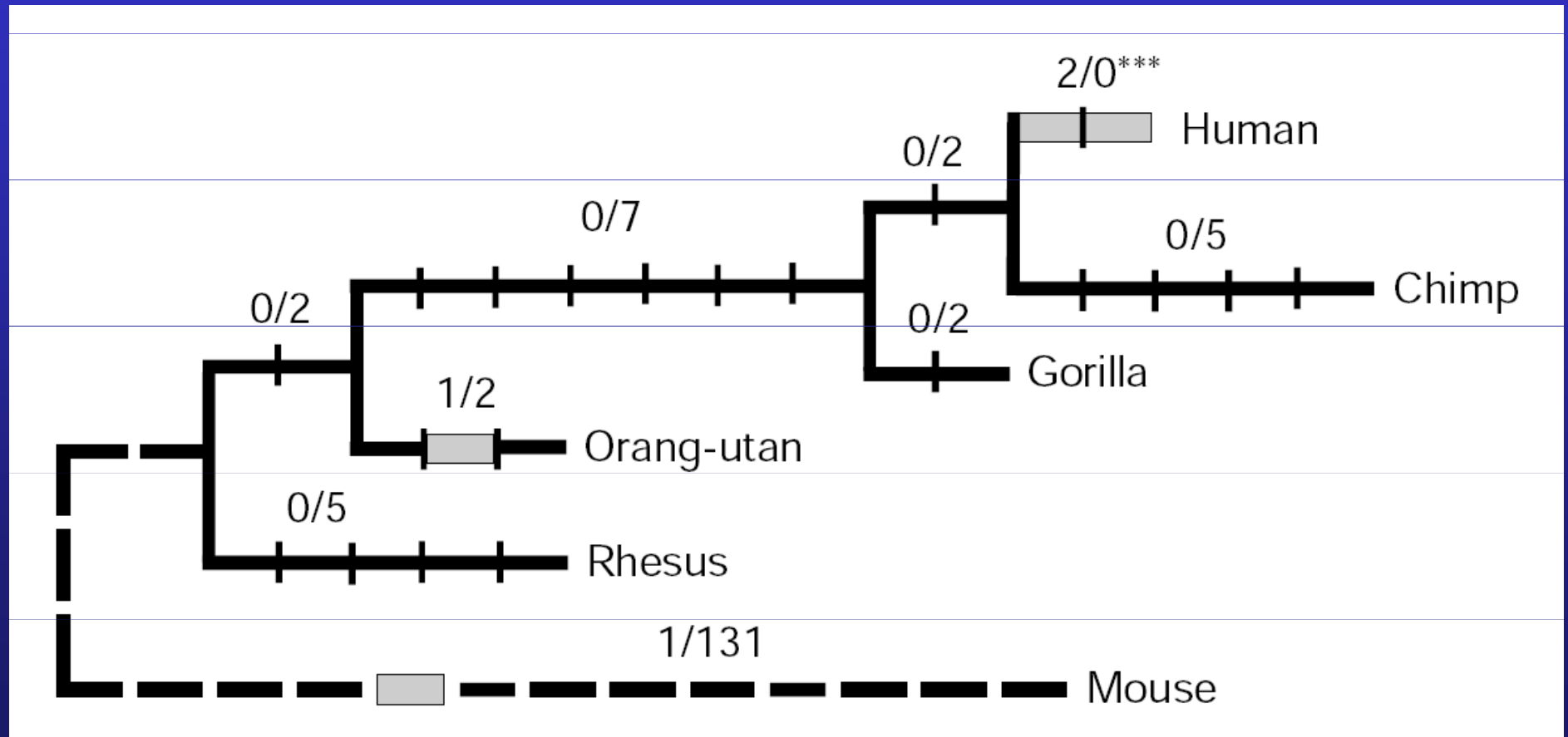
	Lidé	Primáti
<i>Definitivní</i>		
HIV progrese v AIDS	častá	vzácná
Symptomatologie chřipky A	středně těžká až závažná	lehká
Komplikace u hepatitidy B/C	středně těžké až závažné	lehké
Malárie (<i>P. falciparum</i>)	citliví	rezistentní
Menopauza	obligátní	vzácná
<i>Pravděpodobné</i>		
<i>E. coli</i> K99 gastroenteritida	rezistentní	sensitivní?
Rozvoj m. Alzheimer	kompletní	částečný
Koronární ateroskleróza	častá	vzácná
Karcinomy	časté	vzácné

Genetické rozdíly mezi současným člověkem (Člověk moudrý, *Homo sapiens*) a vybranými organismy na celogenomové úrovni a odhadovaná evoluční vzdálenost od posledního společného předka

Druh	Odhadovaná doba uplynulá od výskytu společného předka	Přibližná příbuznost na úrovni DNA
Člověk moudrý (<i>Homo sapiens</i>), kterýkoliv jiný jedinec		~99,9 %
Člověk neandertálský (<i>Homo neandertalensis</i>)	0,5 mil. let	≈99,5 %
Šimpanz učenlivý (<i>Pan troglodytes</i>)	6 mil. let	~96 %
Makak rhesus (<i>Maccaca mulata</i>)	25 mil. let	~93 %
Potkan obecný (<i>Rattus norvegicus</i>)	75 mil. let	~40 % genomu je sekvenčně příbuzných (nikoliv shodných)

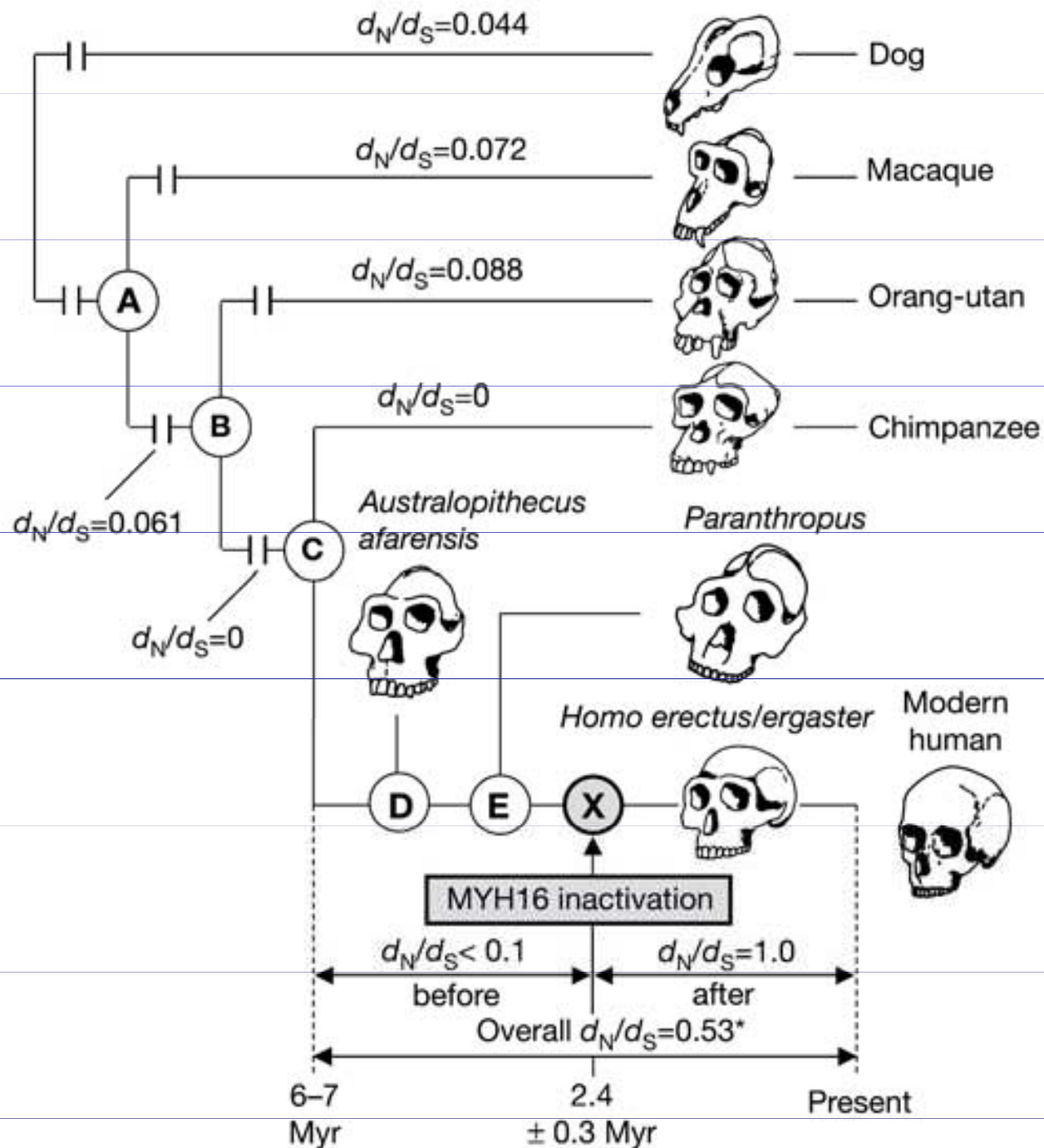


FOXP2 evolution

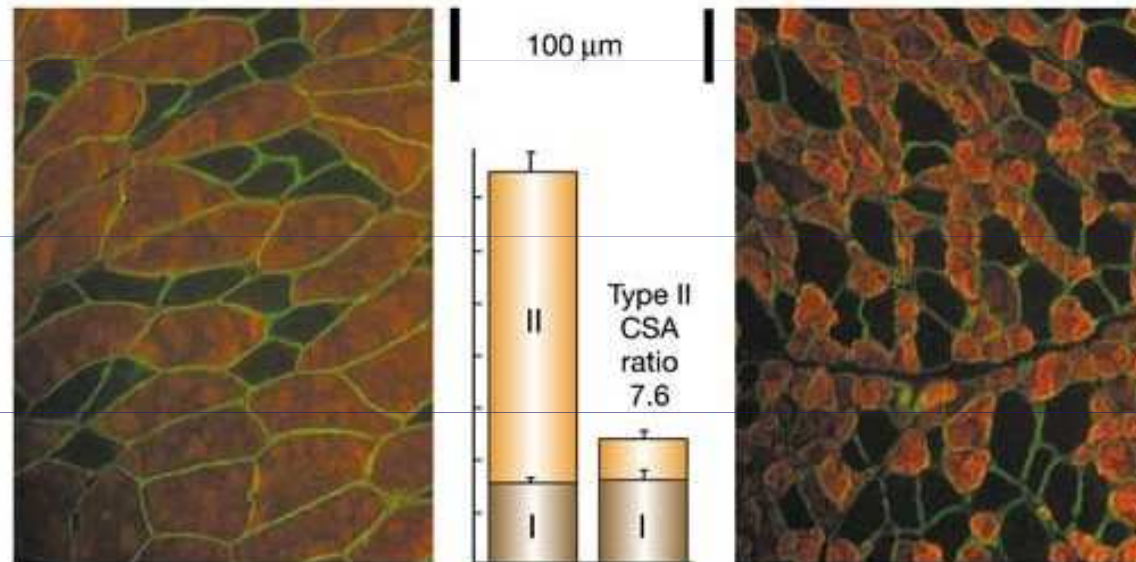
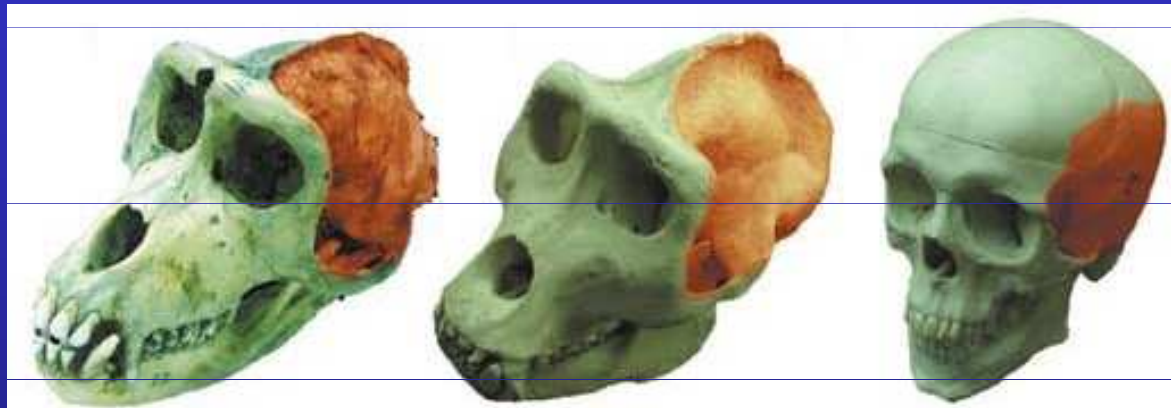


— nucleotide change

▭ aminoacid change

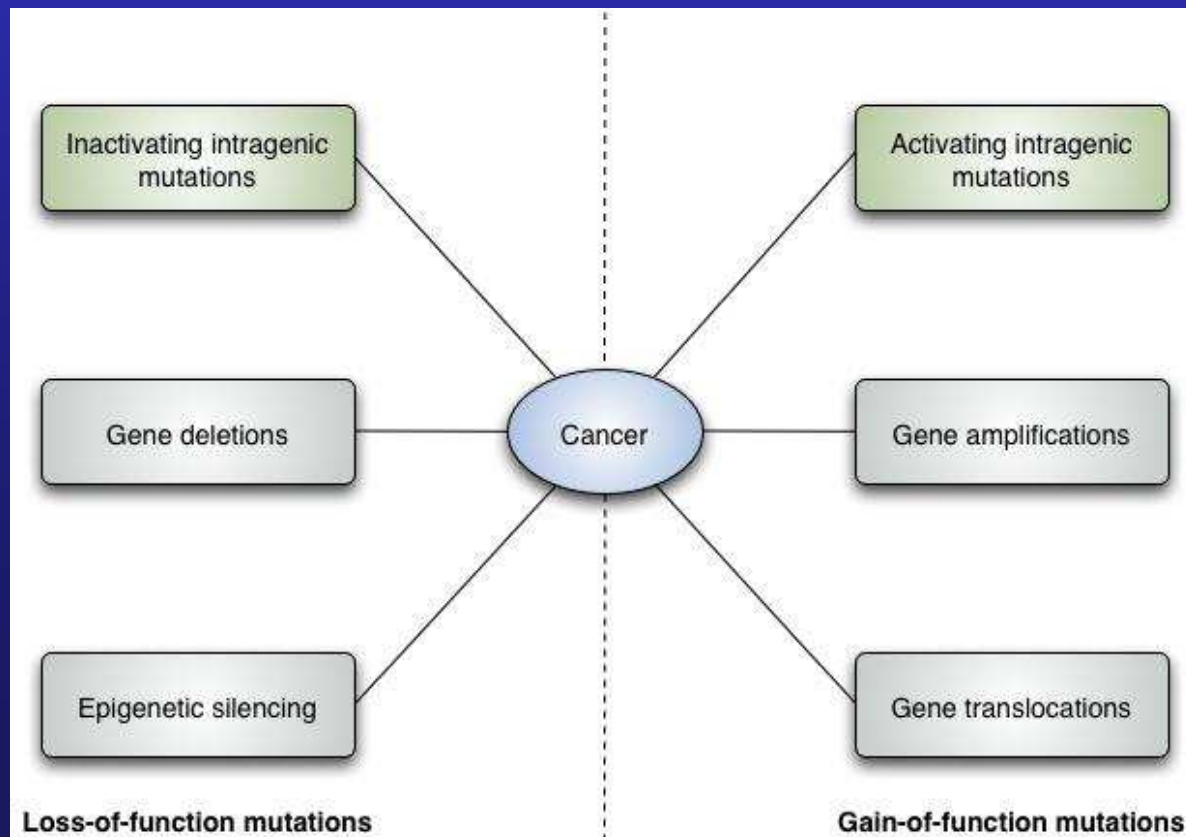
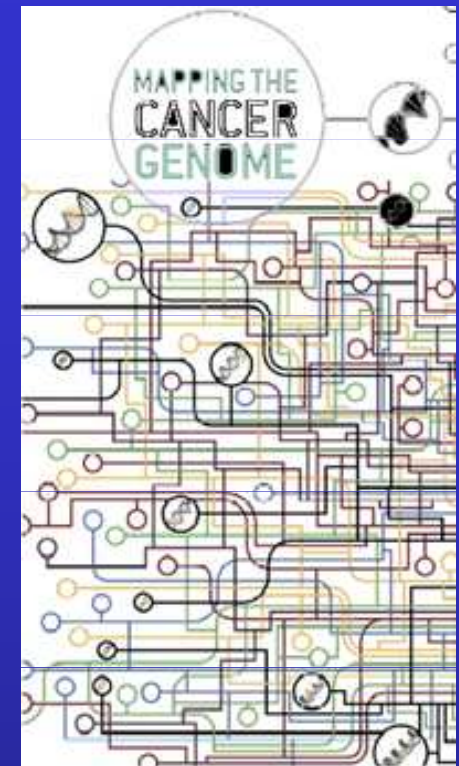


MYH16 inactivation



The Cancer Genome Atlas (TCGA)

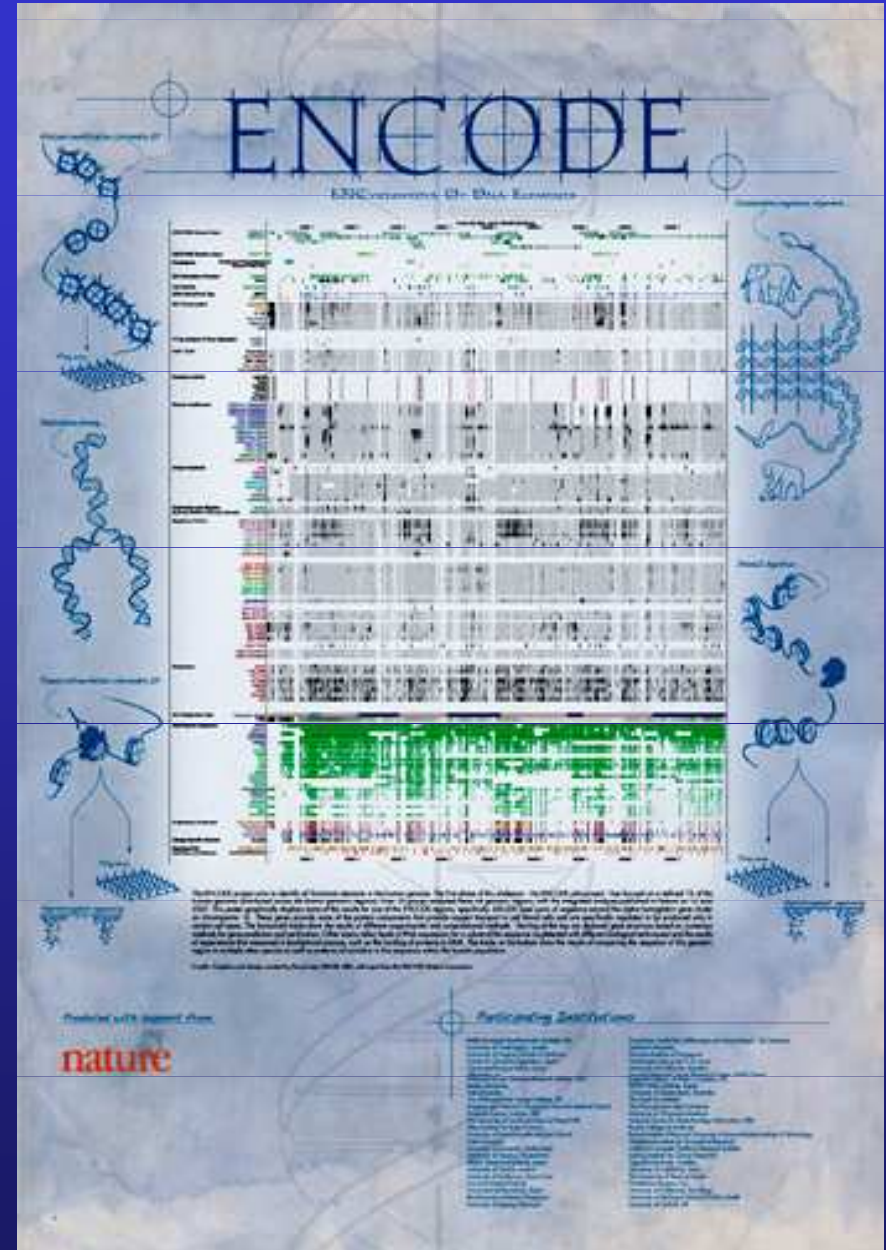
- Jde o vysokokapacitní sekvenaci (tj. sekvenaci téměř celého genomu) mnoha nádorových vzorků jednoho typu nádoru od mnoha různých pacientů.
- Tohle vše je plánováno pro mnoho typů nádorů. Jde tedy o jakýsi frontální útok na odhalení genetického pozadí nádorových onemocnění.



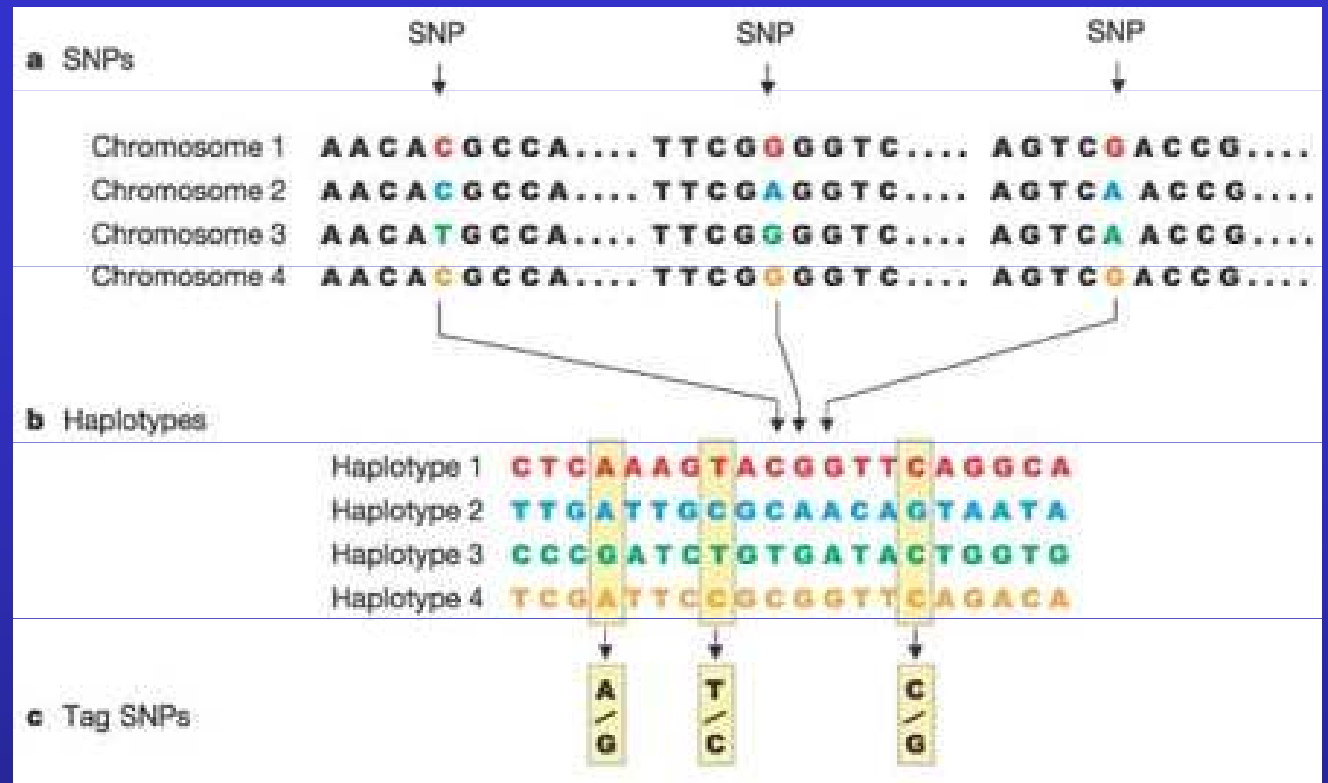
ENCODE

(Encyclopedia of DNA Elements)

- Transkripčně aktivní jsou také podstatné části genomu (molekuly DNA), o kterých se dosud soudilo, že jsou nefunkční a jsou pouhým „balastem“. Přitom tato DNA tvoří okolo 98 % veškeré lidské DNA.
- Znamená to tedy, že i když RNA kódovaná touto „nefunkční“ DNA není přepisována do bílkovin, tvoří se v takovém množství a na tak rozsáhlé části DNA, že nějakou její funkci lze oprávněně očekávat.



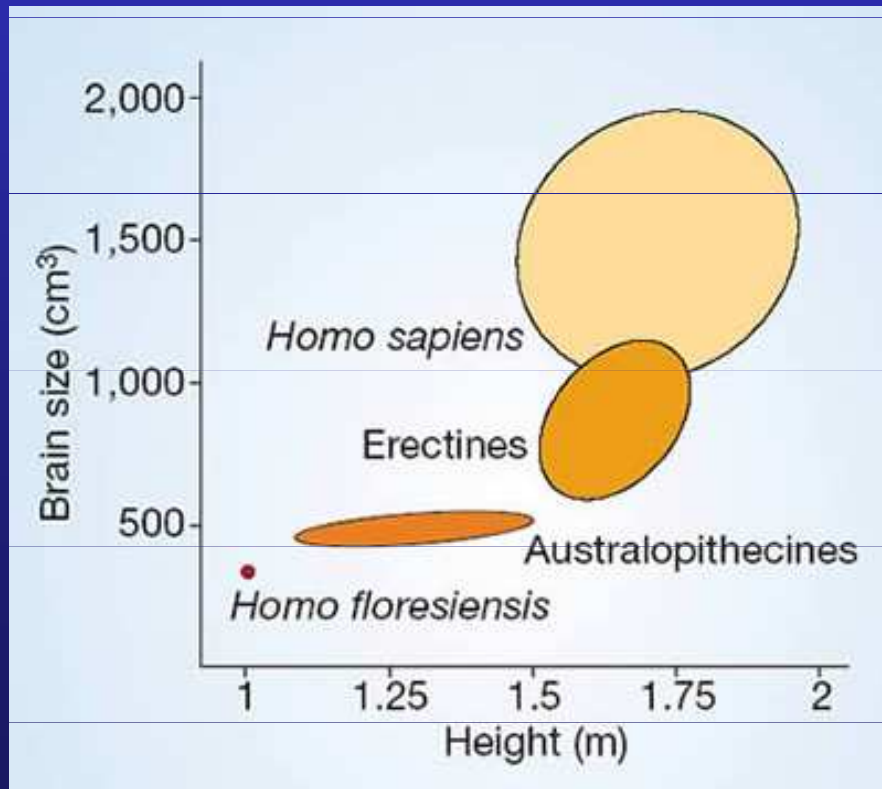
Projekt HapMap



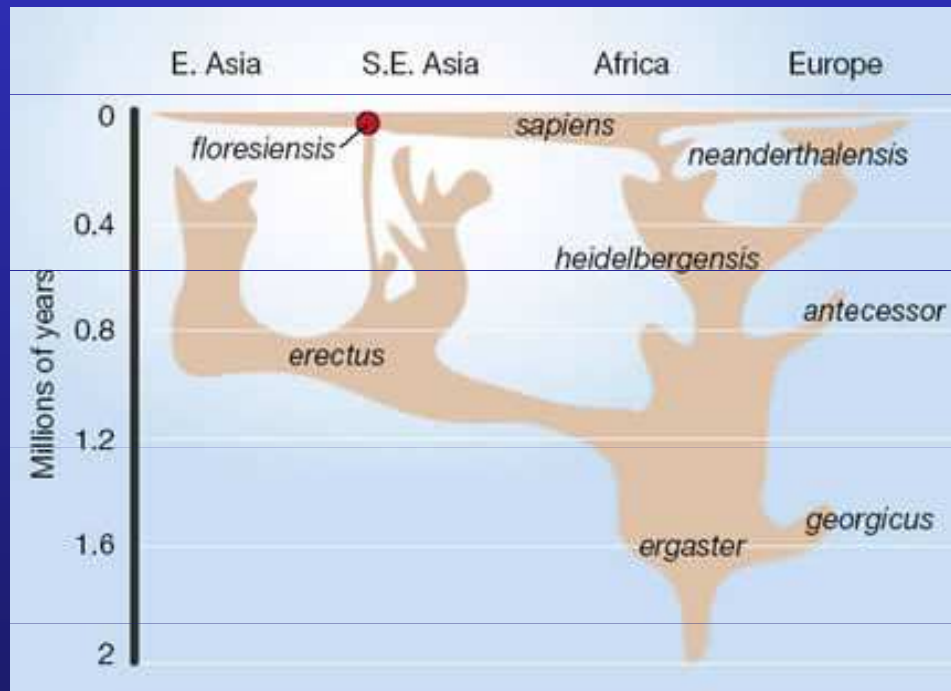
- Vzorky DNA od z 269 lidí z Afriky, Japonska, Číny a USA
- Bylo identifikováno okolo 10 miliónů míst, ve kterých se lidé čtyř různých populací liší nejčastěji. (To přibližně odpovídá shodě 99,9 % mezi kterýmikoliv dvěma osobami.)
- Tato místa se označují jako SNP (jednonukleotidové polymorfismy).
- Platí přitom, že jednotlivé SNP se dědí po určitých blocích (haplotypech). Z toho vyplývá možnost definovat genom individuálního člověka jako kombinaci určitých haplotypů a pro zjištění genotypu konkrétní osoby tedy není třeba mapovat všech 10 miliónů jeho genomových míst, ale stačí genotypovat jen 300.000 – 600.000 klíčových SNP.

Homo floresiensis

Brain size



Homo floresiensis



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



“We think it has something to do with your genome.”



















