Sekvenování genomů

Human Genome Project:
historie, výsledky a důsledky

MUDr. Jan Plátenník, PhD.

(Prosinec 2020)

Počátky sekvenování

- **1965: přečtena sekvence tRNA kvasinky (80 bp)**
- **1977: vynalezeny Sangerova a Maxam & Gilbertova metoda sekvenování**
- **1981: sekvence lidské mitochondriální DNA (16,5 kbp)**
- **1983: sekvence bakteriofága T7 (40 kbp)**
- **1984: Virus Epsteina a Barrové (170 kbp)**

Homo sapiens

- 1985-1990: diskuse o sekvenování lidského genomu
  - "nebezpečné" - "nesmyslné" - "nemožné"
- 1988-1990: Založen HUMAN GENOME PROJECT
  - Mezinárodní spolupráce: HUGO (Human Genome Organisation)
  - Cíle:
    - genetická mapa lidského genomu
    - fyzická mapa: marker každých 100 kbp
    - sekvenování modelových organismů (E. coli, S. cerevisiae, C. elegans, Drosophila, myš)
    - objevit všechny lidské geny (předpokl. 60-80 tisíc)
    - sekvenování celého lidského genomu (4000 Mbp) do r. 2005

Další genomy

- červenec 1995: Haemophilus influenzae (1,8 Mbp) ... První genom nezávisle žijícího organismu
- říjen 1996: Saccharomyces cerevisiae (12 Mbp) ... První Eukaryota
- prosinec 1998: Caenorhabditis elegans (100 Mbp) ... První Metazoa

květen 1998:

- Craig Venter zakládá soukromou biotechnologickou společnost CELERA GENOMICS, Inc. a vyhlašuje záměr sekvenovat celý lidský genom za 3 roky a 300 mil. USD metodou *whole-genome shotgun*
- V té době výsledek práce HGP: sekvenováno cca 4 % lidského genomu.

březen 2000:

- Celera Genomics & akademičtí spolupracovníci publikují draft genomu Drosophila melanogaster (cca 2/3 z 180 Mbp)
- ... *whole-genome shotgun* lze použít i pro velké genomy
- ... ... Lidský genom: závod mezi Human Genome Project a Celera Genomics

International Human Genome Sequencing Consortium (Human Genome Project, HGP)

- Otevřeno spolupráci z každé země na světě
- 20 laboratoří z USA, Velké Británie, Japonska, Francie, Německa a Číny
- Asi 2800 lidí, vedoucí: Francis Collins, NIH
- Financování z veřejných zdrojů (celkové náklady 3 miliardy USD)
- Metoda: *clone-by-clone*
- Výsledky: „Bermudská pravidla“ …každá sekvence do 24 hodin na Internet, přístup zdarma, stálá aktualizace.

Clone-by-clone

genomová DNA

↓

fragmenty cca 150 000 bp

↓

klonování v BAC (*bacterial artificial chromosome*)

↓

určení pozice klonů v genomu na základě fyzického mapování
(STS - *sequence tagged site*, *fingerprint* - štěpení restriktázami)

↓

digesce každého klonu na krátké fragmenty cca 500 bp

↓

sekvenování

↓

sestavení celé sekvence každého klonu pomocí počítače

Celera Genomics, Inc.

- Soukromá biotechnologická společnost, Rockville, Maryland, USA. Prezident Craig Venter.
- Investice do automatizace a počítačového zpracování dat, pár desítek zaměstnanců
- Metoda: *whole-genome shotgun* + ale také využití zveřejněných dat z HGP.
- Výsledky: hrubá data zpřístupněna na www stránkách firmy, další aktualizace a anotace ale výlučně pro komerční účely.

Whole-genome shotgun

genomová DNA

↓

fragmenty 2, 10, 50 kbp

↓

klonování v plasmidech E.coli

↓

sekvenování

↓

sofistikované počítačové metody k sestavení celé sekvence

Únor 2001:

- **International Human Genome Sequencing Consortium publikuje draft lidského genomu v časopisu Nature 15.2.2001.**
  - Draft: 90 % euchromatinu (2,95 Gbp, celý genom 3,2 Gbp). 25 % definitivní.
- **Celera Genomics, Inc. publikuje svou sekvenci lidského genomu v časopisu Science 16.2.2001.**
  - Sekvence euchromatinu (2,91 Gbp)

Pokrok v sekvenování

1985: 500 bp /laboratoř a den

Stále Sangerova dideoxynukleotidová metoda, ale
- **místo gelu kapilární elektroforesa**
- **místo radioaktivity fluorescence**
- **úplná automatizace a robotizace**
- **computer power**

2000: 175 000 bp /den (Celera)

1000 bp/sec. (HGP)

Sekvenování genomů pokračuje...

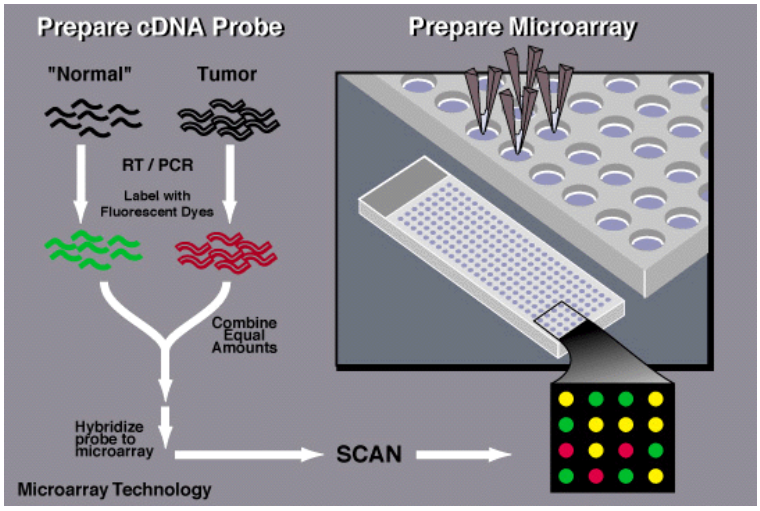
- **Lidský genom nyní:** Definitivní verze publikována 14.4. 2003 …50 let od objevu DNA double helix.
- **Fugu rubripes:** draft genomu v srpnu 2002
- **Myš:**
  - Celera Genomics: draft v červnu 2001
  - Mouse Genome Sequencing Consortium: Nature, prosinec 2002
- **Laboratorní potkan:** draft v březnu 2004
- **Šimpanz:** září 2005

- **… a mnoho dalších genomů:** malárie (původce Plasmodium falciparum a přenašeč Anopheles gambiae), zebrafish, rýže, pes, kráva, ovce, prase, kuře, včela, mamut ad.

Výzkum v "postgenomové" éře

- **Nové přístupy ke studiu genů a proteinů:**
  - **GENOMIKA** ... analýza celého genomu a jeho exprese
  - **PROTEOMIKA** ... analýza celého proteomu, tj. všech proteinů tkáně nebo organismu
  - **BIOINFORMATIKA** ... zpracování, analýza a interpretace velkých souborů dat (NK a AMK sekvencí, gene arrays, 3D struktury proteinů atd. Experimenty ***in silico***
- **Rychlý vývoj nových technologií:**
  - Př. **DNA Microarray** – možnost studovat expresi tisíců genů najednou

DNA Microarray (“ DNA chip”)



Single Nucleotide Polymorphism (SNP)

A G A G T T C T G C T C G

A G G G T T C T G C G C G

SNP se vyskytuje cca 1x na 1000 bp v sekvencích dvou nepříbuzných lidských bytostí (0,1 % genomu)

Asi 10 miliónů SNP s výskytem >1%

Kódující/nekódující

Strukturu proteinu mění/nemění

International HapMap Project

- **Další mezinárodní spolupráce 2002-2009**
- **Sekvenování DNA od 270 lidí ze čtyř různých populací (USA, Nigerie, Japonsko, Čína)**
- **S cílem najít**
  - **Všechny významné lidské SNP (asi 10 000 000)**
  - **Jejich stabilní kombinace (haplotypy)**
  - **Jeden „tag SNP“ typický pro každý haplotyp**
- **Data veřejně přístupná k dalšímu výzkumu a využití**

Lidská genetická variabilita

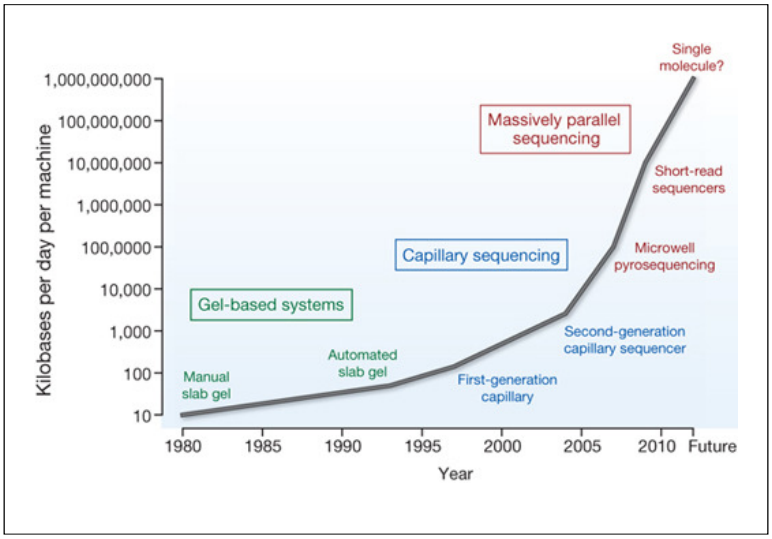
- Dva nepříbuzní lidé mají 99,5% genomu identické
  - Single Nucleotide Polymorphism: 0,1%
  - Copy number variation (inserce, delece, duplikace): 0,4%
- Variace počtu tandemových repetic (…“DNA fingerprinting“)
- Unikátní individuální inserce transpozonů
- Epigenetika (metylace)

Sekvenátory druhé generace

Např. firma Illumina Co., XII/2008:

- Genome Analyzer (Illumina Inc.) udělá za 3 dny to, co by ABI 3730xl (použitý Celera Genomics) trvalo 60 let…
- Náklady na sekvenování jednoho lidského genomu: 40-50 000 $
- ….První sekvenované individuální lidské genomy:
  - 2007: Craig Venter, James Watson – oba genomy zpřístupněny na internetu

… a třetí generace



Graf: Nature 458, 719-724 (2009).
Získáno z http://genome.wellcome.ac.uk

Next-Generation Sequencing (NGS)

- 454 pyrosequencing
- Sequencing by synthesis (Illumina)
- SOLiD sequencing by ligation
- Ion Torrent semiconductor sequencing
- DNA nanoball sequencing
- Heliscope single molecule sequencing
- Single molecule real time (SMRT) sequencing
- Nanopore DNA sequencing

Archon X Prize for Genomics

$ 10 000 000

Vyhlášena v roce 2006.

Pro první tým který osekvenuje 100 lidských genomů za dobu 30 dní nebo kratší v určité požadované kvalitě a s náklady ne více než $1 000 na jeden genom.

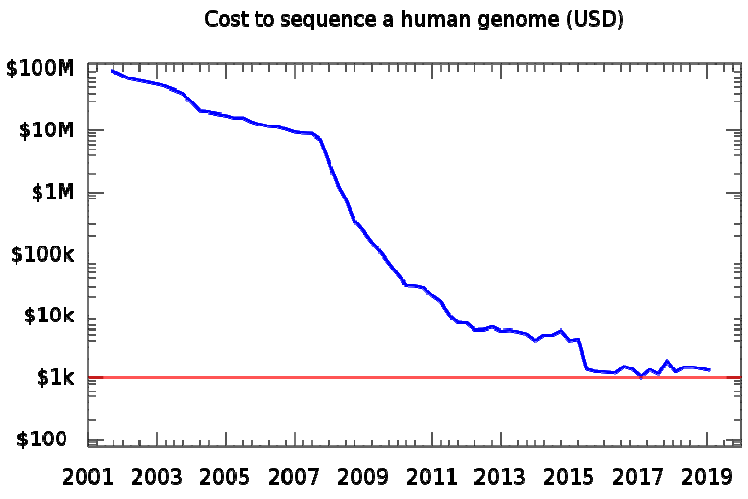
Archon X Prize for Genomics

$ 10 000 000

Vyhlášena v roce 2006.

Pro první tým který s… 100 lidských genom… dní nebo krat… adované kval… ne více než $1 … genom.

Cena zrušena 22.8.2013
„Outpaced by innovation“



(Vývoj ceny sekvenování lidského genomu podle NHGRI, Wikimedia Commons)

Next-Generation Sequencing (NGS)

Současné možnosti, např. Illumina HiSeq 3000 nebo HiSeq 4000:

Sequencing by synthesis (SBS)

Až 400 Gb/den …12 lidských genomů nebo 96 exomů za 3,5 dne

www.illumina.com

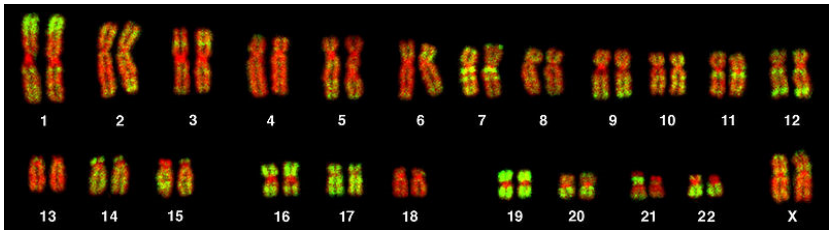
https://www.youtube.com/watch?v=9YxExTSwgPM

EARTH BIOGENOME PROJECT

- Sekvenování všech asi 1,5 milionu známých druhů eukaryotických organismů na Zemi (… veškeré rostliny, živočichové, prvoci a houby, dosud <0.2 % sekvenováno)
- Vyhlášeno 1/11/2018, má trvat 10 let a stát 4,7 miliardy dolarů.

Sekvenování lidského genomu: Výsledky

Lidský genom

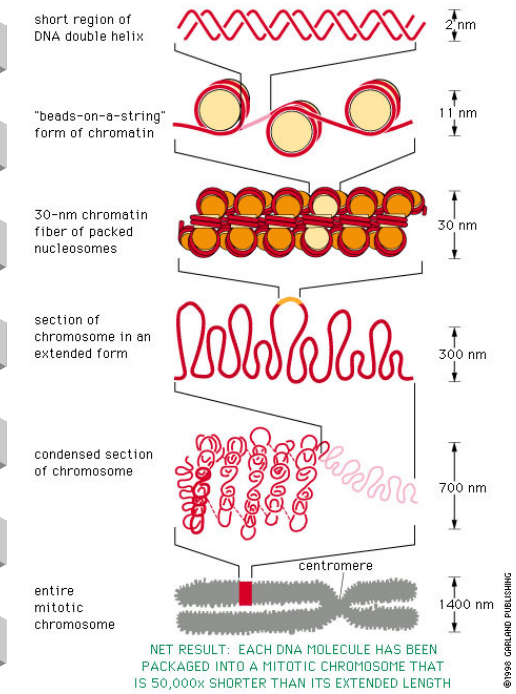


Obr.: Bolzer et al. 2005, PLoS Biol. 3(5): e157 DOI: 10.1371/journal.pbio.0030157

Haploidní genom: 3 miliardy párů bazí rozdělené do 23 chromosomů

- 1 metr DNA při max. roztažení
- 750 Mb (1 CD)
- 2 milióny normostran A4

  (50 úhozů/řádek, 30 řádků/strana)



DNA v buněčném jádře

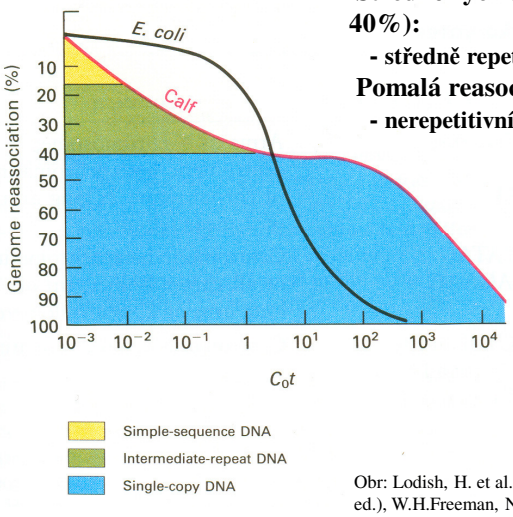
Jádro typické lidské buňky má 5-8 µm v průměru a obsahuje 2 m DNA … odpovídá tenisovému míčku, do kterého se podařilo úhledně poskládat 20 km tenké nitě.

Klasifikace eukaryotické genomové DNA:

- podle "sbalenosti":
  - euchromatin
  - heterochromatin (cca 10%, sekv. obtížné)
- podle opakování:
  - vysoce repetitivní
  - středně repetitivní
  - nerepetitivní
- podle funkce:
  - strukturní (centromery, telomery)
  - kódující proteinové sekvence
  - přepisované do nekódující RNA (introny, rRNA, tRNA, miRNA etc.)
  - transpozony
  - regulační sekvence
  - junk…?

Experimenty s denaturací & reasociací DNA:

Rychlá reasociace (10-15%):
- vysoce repetitivní DNA
Středně rychlá reasociace (25-40%):
- středně repetitivní DNA
Pomalá reasociace (50- 60%):
- nerepetitivní (unikátní) DNA

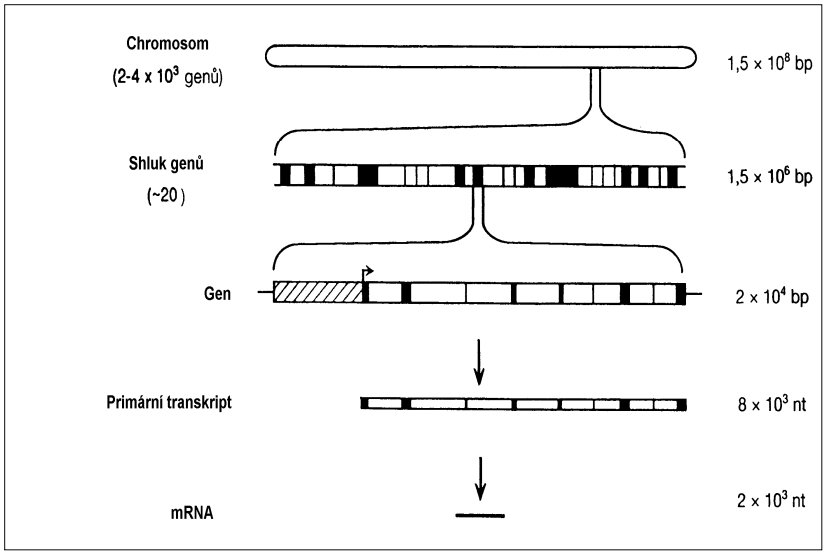


Obr: Lodish, H. et al.: Molecular Cell Biology (3rd ed.), W.H.Freeman, New York 1995.

Klasifikace eukaryotické genomové DNA:

- **Vysoce repetitivní (simple-sequence DNA):**
  - **Veškerý heterochromatin** (centromery, telomery, 8% genomu, stále nesekvenován)
  - **Minisatelity** (3% z euchromatinu)
- **Středně repetitivní:**
  - **Tandemově zmnožené geny kódující rRNA, tRNA a histony** (více stejných kopií genů za sebou, za účelem větší produktivity transkripce, př. geny pro rRNA u eukaryot >100 kopií)
  - **Transpozony**
- **Nerepetitivní:**
  - **Proteinové geny**
  - **Geny pro nekódující RNA**
  - **Regulační sekvence**

Eukaryotický GEN

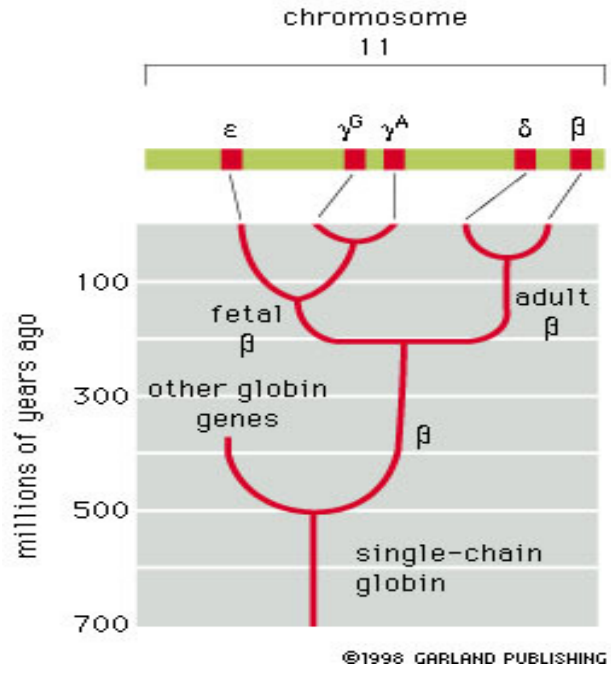


Obr: Murray, R.K. et al.: Harperova biochemie, Appleton & Lange 1993, v češtině nakl. H&H 2002.

Rozmístění genů v genomu není rovnoměrné

- Velké rozdíly mezi chromosomy:
  - chromosom 1: 2057 proteinových genů
  - chromosom Y: 66 proteinových genů
- oblasti bohaté na geny (“města”)
  - více C a G
- oblasti chudé na geny (“pouště”)
  - více A a T, až 3 Mb!
- CpG ostrůvky - “bariéra mezi městy a pouštěmi” ... regulace genové aktivity

- **Solitární gen:**
  - v celém genomu v jediné kopii (asi polovina genů)
- **Tandemově duplikované geny pro histony a rRNA**
- **Genová rodina:**
  - skupina genů evolučně pocházející z jediného genu, vznik duplikací a postupnou diverzifikací sekvence a funkce
- **Pseudogen:**
  - gen který zmutoval natolik že už nemůže být přepisován (…„molekulární fosilie“)
- **Zpracovaný („processed“) pseudogen:**
  - pseudogen vzniklý zpětným přepisem mRNA a integrací do genomu



Počet genů v lidském genomu

- **Kódující geny: 20 448**
- **Nekódující geny: 23 997**
  - **Krátké nekódující geny** (do 200 bp, rRNA, miRNA, ncRNA, snRNA, snoRNA …): **4 867**
  - **Dlouhé nekódující geny** (nad 200 bp, různé nekódující RNA): **16 909**
  - **Různé nekódující geny**: **2 221**
- **Pseudogeny: 15 217**
- **Celkem genové transkripty: 232 186**

Ensembl release 102, Nov. 2020 (www.ensembl.org)

Proteinové geny v lidském genomu

cca 20 400

Asi 25% genomu přepisováno do pre-mRNA,

z toho ale jen 5% jsou exony

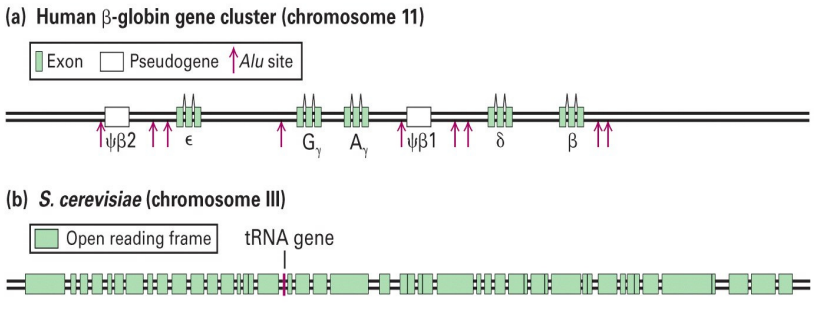
…Lidský EXOM: cca 1.5 % genomu

Počet genů neodpovídá komplexitě organismu?!

Sacch. cerevisiae	6 600 genů
C. elegans	20 191 genů
Drosophila	13 931 genů
Arabidopsis thaliana	27 655 genů

Srovnání genomu člověka/myši s genomy nižších organismů (C.elegans, Drosophila):

- menší hustota genů, delší introny



Obr: Lodish, H. et al.: Molecular Cell Biology (5th ed.), W.H.Freeman, New York 2004.

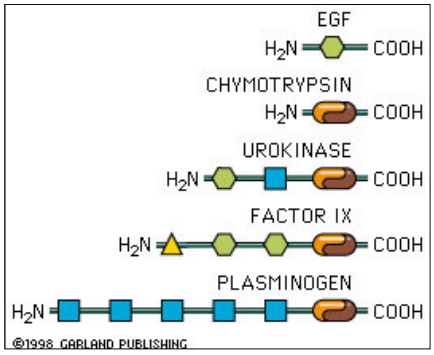
Jak se hledají geny v genomech:

- **Bakterie, kvasinky:**
  - open reading frames (ORFs)
- **Vyšší organismy:**
  - srovnání s transkriptomem (RNA-seq)
  - podobnost se známými geny
  - hledání rozpoznávacích sekvencí pro místa sestřihu
  - podobnost s genomy jiných organismů

Srovnání genomu člověka/myši s genomy nižších organismů (C.elegans, Drosophila):

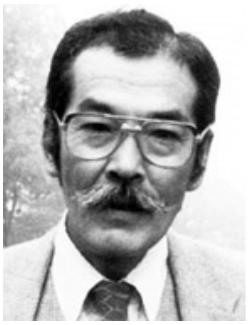
- expanse genů /nové geny se vztahem k:
  - srážení krve
  - získaná (specifická) imunita
  - nervový systém
  - intra- a intercelulární komunikace
  - kontrola genové exprese
  - programová buněčná smrt (apoptosa)

- jen málo proteinových domén zcela nových u obratlovců, ale
  - expanse proteinových rodin
  - složitější architektura proteinů, nové kombinace domén a více domén/ protein



  - více proteinů z jednoho genu - alternativní sestřih až v 95 %

Susumu Ohno, 1972



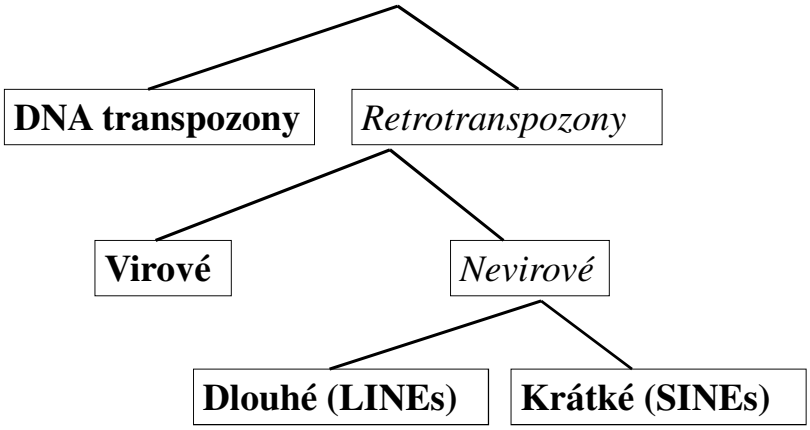
Susumu Ohno
Feb. 1, 1928 - Jan 13, 2000

- Vzhledem k rychlosti vzniku mutací nemůže lidský haploidní genom obsahovat více jak asi 30 000 genů.
- Většina DNA je tedy navíc … junk!

http://www.junkdna.com/ohno.html

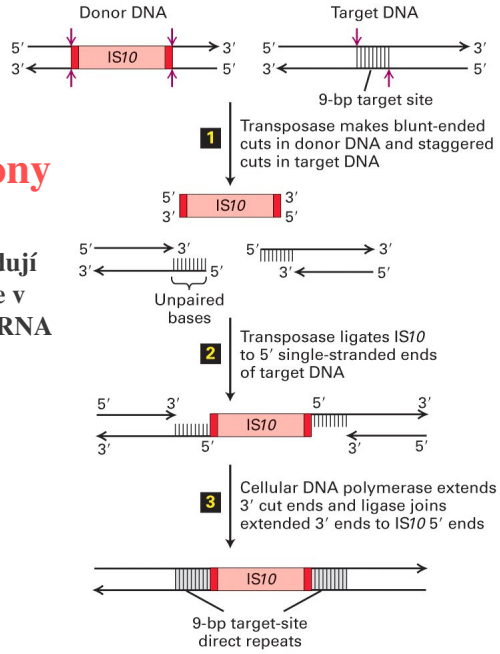
Mobilní DNA elementy (transpozony)

Autonomní DNA sekvence, které se samy množí, představují 44 % genomu



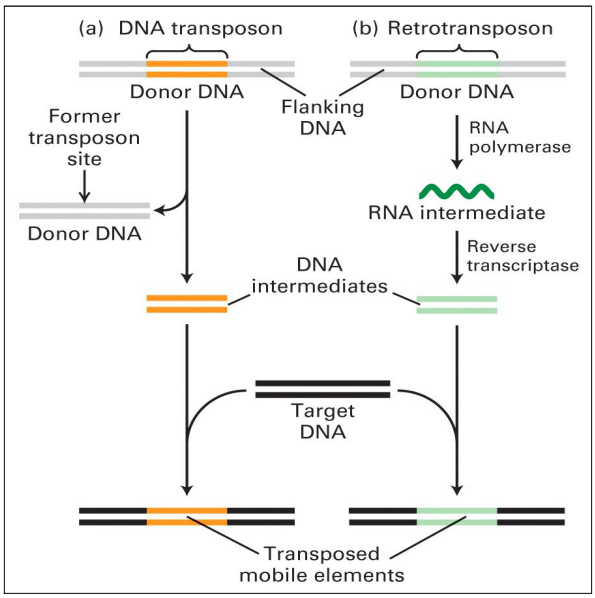
DNA transpozony

2-3 kb (nebo kratší), kódují transposasu, cut & paste v genomu bez přepisu do RNA



Obr: Lodish, H. et al.: Molecular Cell Biology (5th ed.), W.H.Freeman, New York 2004.

Transpozony:



Obr: Lodish, H. et al.: Molecular Cell Biology (5th ed.), W.H.Freeman, New York 2004.

Mobilní (parazitické) elementy v savčím genomu:

- **DNA transpozony**
  - **2-3 kb (nebo kratší), kódují transposasu, cut & paste v genomu bez přepisu do RNA**
- **Virové retrotranspozony**
  - **6-11 kb (nebo kratší), retroviry bez genu pro proteinový obal (env)**
- **LINEs (long-interspersed repeats),**
  - **6-8 kb, př. L1, kódují 2 proteiny (1 je reversní transkriptasa)**
- **SINEs (short-interspersed repeats),**
  - **100-300 bp, př. Alu, nekódují nic, množení závisí na LINEs, původ: z malých nekódujících buněčných RNA**

Census parazitických elementů v lidském genomu:

LINEs:	**850 000x**	**21 % genomu**
SINEs:	**1 500 000x**	**13 % genomu**
Retrovirus-like:	**450 000x**	**8 % genomu**
DNA transposony:	**300 000x**	**3 % genomu**

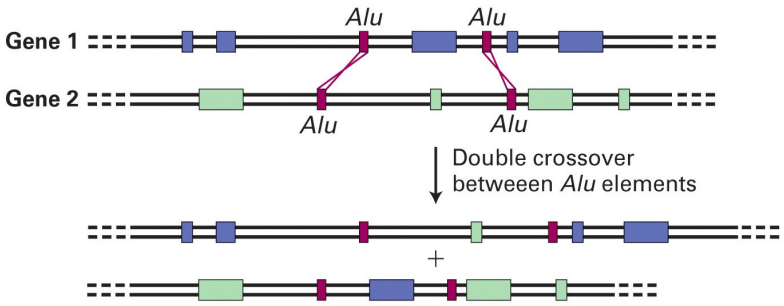
- **V drtivé většině ale mutované/nekompletní kopie, jen malá část (<0,05%) je aktivní:**
  - **LINEs: 80-100 L1**
  - **SINEs: 2000-3000 Alu, <100 SVA**
  - **Retrovirus-like: ?** *(HERV-K…opravdu vyhynul?)*
  - **DNA transpozony: 0**
- **V genomu myši aktivních transposonů mnohem více** *(...proč?)*

Význam transpozonů v lidském genomu

- Transpozice v germinálních buňkách nastává relativně vzácně (cca 1x na 20 živě narozených, většinou Alu)
- I tak významný zdroj lidské genetické variability
- Může vést k inaktivaci genu, dokumentováno jako vzácná příčina vrozených chorob
- V somatických buňkách může být příčinou mosaicismu
  - úloha L1 v neurogenesi?

- Transpozony usnadňují rekombinaci

….hnací síla evoluce !



Obr: Lodish, H. et al.: Molecular Cell Biology (5th ed.), W.H.Freeman, New York 2004.

Neklasifikovaná „spacer“ DNA:
nerepetitivní, nekódující, >1/2 genomu …
zřejmě rovněž mrtvé transpozony, které už mutovaly natolik že nejsou rozpoznány

Projekt ENCODE, 2012: žádná junk DNA!

- Až 80% genomu má biologickou funkci
- Až 75% genomu je aspoň někdy a někde přepisováno do RNA
- Přesto že evolučně konzervováno není více jak 20% genomu

….?????.....

Sekvenování lidského genomu: Důsledky

Přínos sekvenování genomů

- Usnadnění výzkumu molekulární podstaty chorob
- Studium evoluce a migrace lidského druhu
- Co vlastně genom kóduje (“nature vs. nurture”) a jaký je genetický podklad rozdílů mezi lidmi
- Genomická medicína farmakogenomika, personalizovaná medicína….

Genomická medicína

- **1) Diagnostika na úrovni genů**
  - **Vzácné monogenní choroby**
  - **Posun do časnější životní fáze**
    - **Možnost diagnózy dříve než se nemoc objeví**
    - **Novorozenecký screening**
    - **Prenatální diagnostika z fetální DNA v cirkulaci matky**
    - **Prekoncepční testování rodičů, preimplantační testy u IVF**
  - **Genetická analýza nádorů umožňuje racionální volbu cílené biologické léčby**
  - **U komplexních, polygenně podmíněných chorob (srdeční choroby, cukrovka) zatím obtížné**

Personal Genomics: 23andME

- Vzorek sliny zaslaný DHL, genotypizace cca 700 000 SNPs
- DNA relatives
- Ancestry:
  - Ancestry Composition
  - Paternal (Y chromosome haplogroup)
  - Maternal (mitochondrial DNA haplogroup)
  - Per cent Neanderthal DNA
- Health

Personal Genomics: 23andME

- Vzorek sliny zaslaný DHL, genotypizace cca 700 000 SNPs
- DNA relatives
- Ancestry
- Health:
  - Disease risk: 123 (31 high confidence)
  - Drug response: 25 (12 high confidence)
    Inherited conditions: 53 (all high confidence)
  - Traits: 63 (15 high confidence)

Genome-Wide Association Studies (GWAS)

- Fenotyp (vlastnost nebo choroba) + genotypizace SNP
- Statistická analýza: současný výskyt?
- Velké množství účastníků (>10 000) třeba k dosažení signifikance
- Běžné varianty: mnoho, jednotlivě malý vliv, ale celkově většina dědičnosti
- Vzácné varianty: neobvyklé, často de novo, pokud přítomné mají velký vliv

Proč analýza SNP neříká víc?

- Informace o běžných SNP nestačí – třeba najít individuální (vzácné) polymorfismy
- SNP nejsou hlavní příčinou lidské genetické variability – duplikace/delece a inserce transpozonů významnější
- Situace kdy o znaku rozhoduje jeden gen je relativně vzácná – častěji je fenotyp výsledek souhry mnoha genů
- O fenotypu rozhoduje exprese genů!
  - Polymorfismy v regulační nekódující DNA
  - Epigenetika (metylace DNA atd.) – též lze dědit!

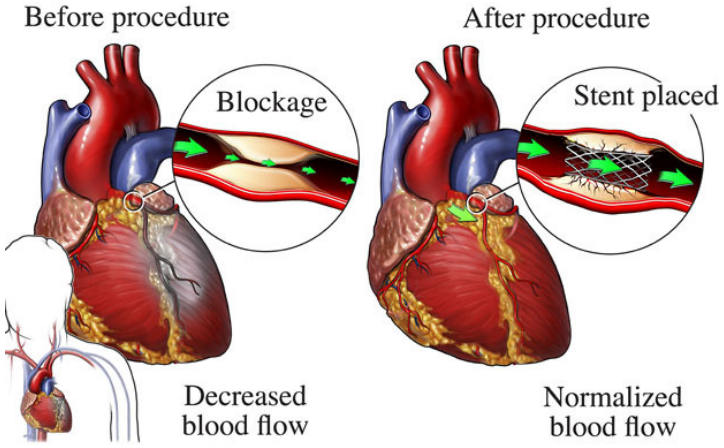
Genomická medicína

- 2) Farmakogenomika
  - Cílená biologická léčba nádorů na základě jejich genetické analýzy
    - Př.: protilátka proti HER-2 jen u nádorů prsu které tento protein exprimují
  - Predikce účinnosti a případných nežádoucích účinků léku na základě markerů v genomu pacienta
    - Př.: léčba chronické hepatitidy C, HIV, možná i dávkování warfarinu

… personalizovaná medicína

Genomická medicína

- 2) Farmakogenomika – příklad:



Obr.: https://www.telegraph.co.uk/news/uknews/theroyalfamily/8977422/Coronary-stenting-how-does-it-work.html

Genomická medicína

- **2) Farmakogenomika – příklad:**
  - **Clopidogrel:**
    - Antikoagulans (blokuje ADP receptor na trombocytech)
    - Účinek vyžaduje metabolickou aktivaci mikrosomálními hydroxylázami (cytochrom P450 2C19)
    - Až 30% má geneticky sníženou/chybějící hladinu aktivujícího enzymu
  - **Studie u pacientů se srdečními stenty:**
    - Volba antikoagulační léčby podle genotypizace CYP2C19: clopidogrel nebo ticagrelol/prasugrel
    - Kontrolní skupina: ticagrelol/prasugrel
    - Výsledek: u genotypizované skupiny stejná účinnost léčby, ale méně krvácivých komplikací

Claasens et al. N. Engl. J. Med. 2019, 381:1621

Genomická medicína

- **3) Microorganismy:**
  - **Patogenní:**
    - **Rychlá diagnostika infekčního onemocnění na základě sekvenování patogenu – významné zejména u nových epidemií (SARS, MRSA…)**
  - **Nepatogenní - Lidský Mikrobiom**
    - **Např. bakterie lidského střeva – metabolická aktivita srovnatelná s játry, individuálně rozdílné spektrum, vztah k střevním zánětům, ateroskleróze, obezitě…**

Etické, legislativní a sociální otázky

- **Gene privacy:**
  - kdo má právo znát něčí genetickou informaci a jak jí smí použít, obava z diskriminace zaměstnavatelem, zdravotní pojišťovnou...
- **Gene testing**
- **Gene therapy**
- **Designer babies**
- **Behavioral genetics:**
  - vztah genů k lidskému chování, možný vývoj ke genetickému determinismu a ztrátě odpovědnosti za vlastní chování
- **GMO**
- **Gene patenting**

Hlavní zdroje:

Alberts, B. et al.: Essential Cell Biology, Garland Publishing, Inc., New York 1998.

Lodish, H. et al.: Molecular Cell Biology, W.H.Freeman, New York 1995, 2004 ("Darnell").

Nature 2001: 409 (6822, 15.2.2001); pp. 813-958.

Science 2001: 291 (5507, 16.2.2001); pp.1177-1351.

Trends in Genetics 2007: 23, pp.183-191.

Nature 2009: 458, 719-724.

FEBS Letters 2011: 585; pp. 1589-1594.

Science Translational Medicine 2013: 5, 189sr4.

PNAS 2014: 111, pp. 6131-6138

N. Engl. J. M. 2019: 381, pp. 1621-1631.

Lecture by Eric Lander, 1.LF UK, 18.2.2020.

www.ncbi.nlm.nih.gov

genomics.energy.gov

en.wikipedia.org

www.ensembl.org

www.illumina.com

www.earthbiogenome.org

www.23andme.com

Obr. "Human and DNA Shadow": U.S. Department of Energy's Joint Genome Institute, Walnut Creek, CA, http://www.jgi.doe.gov.