Sekvenování genomů

Human Genome Project:
historie, výsledky a důsledky

MUDr. Jan Pláteník, PhD.

(Prosinec 2010)

Počátky sekvenování

- **1965: přečtena sekvence tRNA kvasinky (80 bp)**
- **1977: vynalezena Sangerova a Maxam & Gilbertova metoda**
- **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

únor 2001:

- International Human Genome Sequencing Consortium publikuje draft lidského genomu v časopisu Nature 15.2.2001.
- Celera Genomics, Inc. publikuje svou sekvenci lidského genomu v časopisu Science 16.2.2001.

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
- Metoda: *clone-by-clone*
- Výsledky: každá sekvence do 24 hodin na Internet, přístup zdarma, stálá aktualizace.
- Draft (Nature 15.2.2001): 90 % euchromatinu (2,95 Gbp, celý genom 3,2 Gbp). 25 % definitivní.
- Definitivní verze: 14.4. 2003 (50 let od objevu DNA double helix).

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.
- Publikace v Science 16.2.2001: sekvence euchromatinu (2,91 Gbp)
- 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

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:** dokončen (99 % euchromatinu)
- **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.

Veřejně přístupné databáze DNA/RNA sekvencí

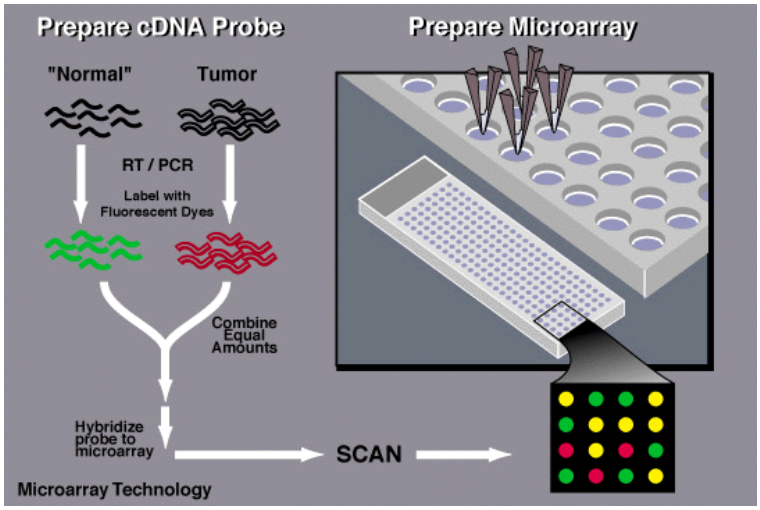
- GenBank, National Center for Biotechnology Information (NCBI), Bethesda, Maryland, USA
- EMBL-Bank, EMBL's European Bioinformatics Institute, Hinxton, UK
- DNA Data Bank of Japan, National Institute of Genetics, Mishima, Japan

Obsah všech tří databází v srpnu 2005 překročil 100 000 000 000 párů basí (100 Gb)
... z genů/genomů 165 000 různých druhů organismů

Výzkum v "postgenomové" éře

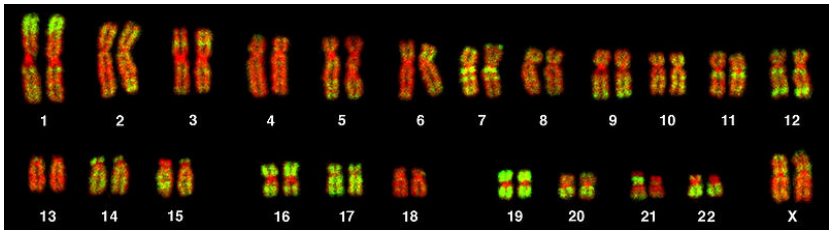
- **Nové přístupy ke studiu genů a proteinů:**
  - **GENOMIKA** ... analýza celého genomu a jeho exprese
  - **PROTEOMIKA** ... analýza celého proteomu, tj. všech proteinů tkáně nebo organismu
  - **BIOINFORMATIKA** ... zpracování, analýza a interpretace velkých souborů dat (NK a AMK sekvencí, gene arrays,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

DNA Microarray (" DNA chip")



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)

Klasifikace eukaryotické genomové DNA:

- podle “sbalenosti”:
  - euchromatin
  - heterochromatin
- podle opakování:
  - vysoce repetitivní
  - středně repetitivní
  - nerepetitivní
- podle funkce:
  - strukturní (centromery, telomery)
  - kódující (protein nebo RNA)
  - nekódující
    - regulační oblasti
    - spacer, “junk DNA”, “selfish DNA”

Experimenty s denaturací & reasociací DNA:

- **Rychlá reasociace (10-15%):**
  - **simple-sequence DNA:**
    - **heterochromatin v okolí telomer a centromery**
    - **satelity, minisatelity.**
- **Středně rychlá reasociace (25- 40%):**
  - **intermediate-repeat DNA (mobilní elementy, transpozony)**
  - **tandemově zmnožené geny pro rRNA, histony**
- **Pomalá reasociace (50- 60%):**
  - **single-copy DNA**
    - **geny**
    - **ostatní neklasifikovaná spacer DNA**

Klasifikace eukaryotické genomové DNA:

- Geny kódující proteiny
- Tandemově zmnožené geny kódující rRNA, tRNA a histony
- Repetitivní DNA
- Neklasifikovaná *spacer* DNA

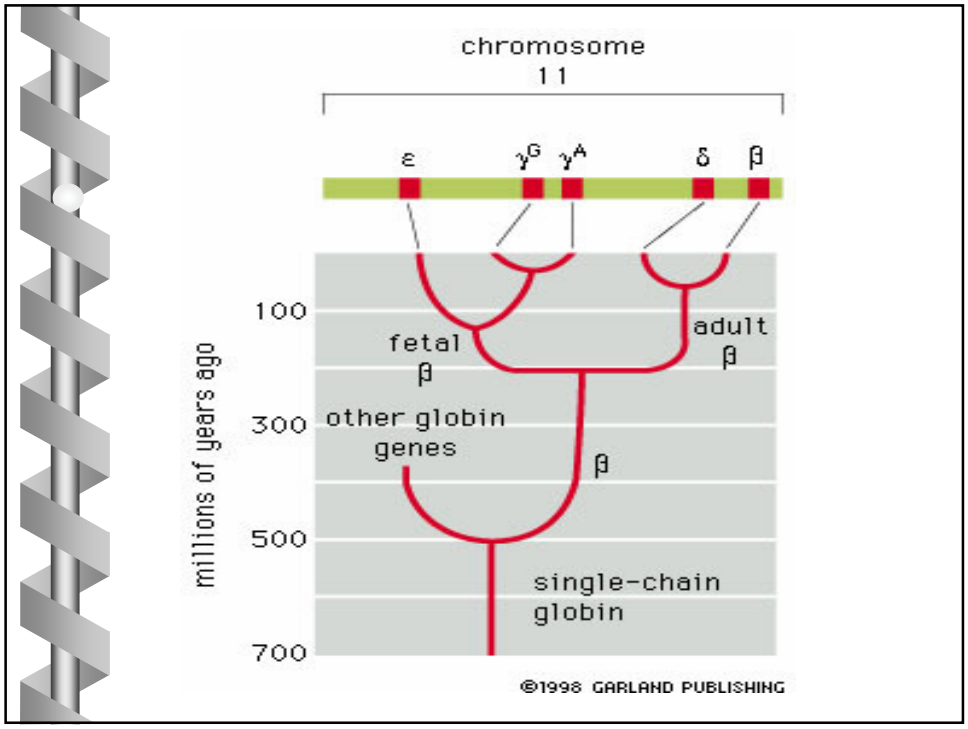
Klasifikace eukaryotické genomové DNA:

- Geny kódující proteiny
  - asi 25 % genomu, z toho ale jen 5 % (1,4 % genomu) připadá na exony
- Tandemově zmnožené geny kódující rRNA, tRNA a histony
- Repetitivní DNA
- Neklasifikovaná *spacer* DNA

Rozmístění genů v genomu není rovnoměrné

- Velké rozdíly mezi chromosomy:
  - chromosom 1: 2968 genů
  - chromosom Y: 231 genů
- oblasti bohaté na geny (“města”)
  - více C a G
- oblasti chudé na geny (“pouště”)
  - více A a T
- CpG ostrůvky - “bariéra mezi městy a pouštěmi” ... regulace genové aktivity

- **Solitární gen:**
  - **v celém genomu v jediné kopii (asi polovina genů)**
- **Genová rodina:**
  - **skupina genů evolučně pocházející z jediného genu, v evoluci postupná diverzifikace sekvence a funkce**
- **Pseudogen:**
  - **gen který zmutoval natolik že nemůže být přepisován (v celém genomu > 20 000 !)**
- **Zpracovaný (“processed”) pseudogen:**
  - **pseudogen vzniklý zpětným přepisem mRNA a integrací do genomu**



Počet genů v lidském genomu

20 000 - 25 000

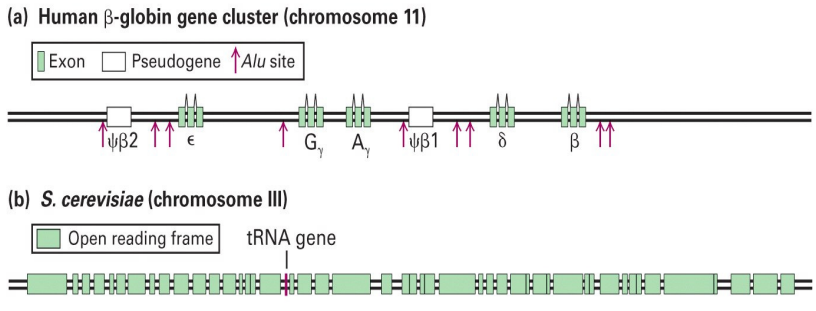
Mnohem méně než se čekalo!

... neodpovídá komplexitě organismu:

Sacch. cerevisiae	6 000 genů
C. elegans	18 000 genů
Drosophila	13 000 genů
Arabidopsis thaliana	26 000 genů

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

- menší hustota genů, delší introny

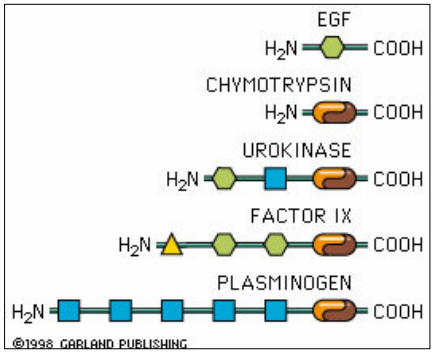


Obr: Lodish, H. et al.: Molecular Cell Biology (5th ed.), W.H.Freeman, New York 2004.

Jak se hledají geny v genomech:

- **Bakterie, kvasinky:**
  - **open reading frames (ORFs)**
- **Vyšší organismy:**
  - **hybridizace/srovnání s cDNA nebo EST (expressed sequence tag = část cDNA)**
  - **podobnost se známými geny**
  - **hledání rozpoznávacích sekvencí pro místa sestřihu**
  - **podobnost s genomy jiných organismů**

- **jen asi 7 % proteinových domén zcela nových u obratlovců, ale**
  - **expanse proteinových rodin**
  - **složitější architektura proteinů, nové kombinace domén a více domén/ protein**



  - **více proteinů z jednoho genů - alternativní sestřih až v 60 %**

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)**

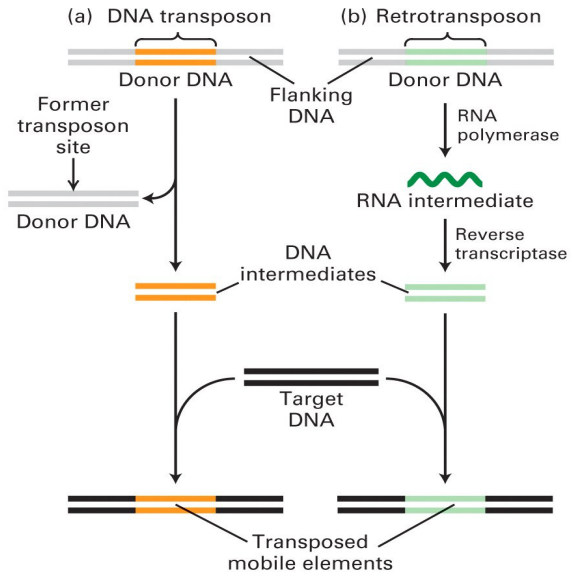
Klasifikace eukaryotické genomové DNA:

- Geny kódující proteiny
- Tandemově zmnožené geny kódující rRNA, tRNA a histony
  - lze též řadit mezi středně repetitivní DNA
  - více stejných kopií genů za sebou, za účelem větší produktivity transkripce
  - geny pro ribosomální RNA u eukaryotů: >100 kopií
- Repetitivní DNA
- Neklasifikovaná *spacer* DNA

Klasifikace eukaryotické genomové DNA:

- Geny kódující proteiny
- Tandemově zmnožené geny kódující rRNA, tRNA a histony
- Repetitivní DNA
  - simple-sequence DNA: vysoce repetitivní, 3% z euchromatinu (minisatelity) + veškerý heterochromatin (centromery, telomery, 8 % genomu, stále nesekvenován)
  - interspersed repeats (středně repetitivní DNA) = mobilní elementy (transpozony), 45 % z euchromatinu
- Neklasifikovaná *spacer* DNA

Mobilní elementy (transpozony):



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

Mobilní (parazitické) elementy v savčím genomu:

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

Census parazitických elementů v lidském genomu:

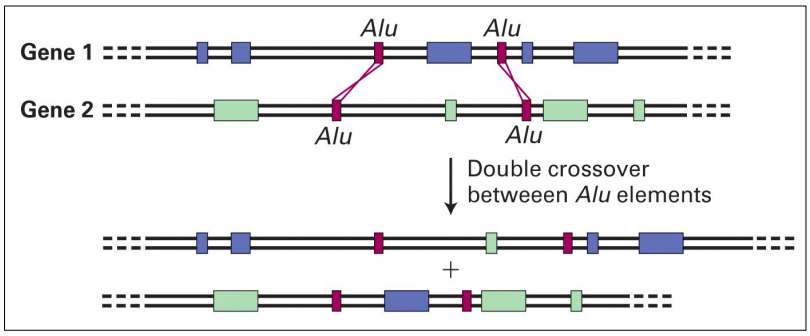
LINEs:	850 000x	21 % genomu
SINEs:	1 500 000x	13 % genomu
Retrovirus-like:	450 000x	8 % genomu
DNA transpozony:	300 000x	3 % genomu

- V celém genomu jen malý počet aktivních LINEs a endogenních retrovirů.
- Vše ostatní: staré, mutované a neaktivní "molekulární fossilie".
- V genomu myši aktivních transpozonů mnohem více (...proč?).

Význam transpozonů v lidském genomu

- Namnožení těchto elementů v genomu: ve více vlnách dávno v evoluci, možná ještě před radiací savců na začátku třetihor
- Zodpovídají za část spontánních mutací (1:600 u člověka)
- Pro jedince je aktivita transpozonů negativní (možná inaktivace genů)
  - ... transposice L1 (LINEs) dokumentována jako vzácná příčina genetických chorob u člověka

- **Ale pro vývoj druhu zřejmě pozitivní**
  - **Repetitivní elementy -> více rekombinací -> urychlení evoluce**



  - **Zpětný přepis mRNA - "processed pseudogenes"** (někdy vznikne funkční gen)
  - **Alu (SINEs) - vyšší výskyt v oblastech bohatých na geny, možná kódují RNA molekulu s funkcí v buňce** ....Alu sekvence podobná RNA ze SRP (signal recognition particle)

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

Klasifikace eukaryotické genomové DNA:

- **Geny kódující proteiny**
- **Tandemově zmnožené geny kódující rRNA, tRNA a histony**
- **Repetitivní DNA**
- **Neklasifikovaná *spacer* DNA:**
  - **nerepetitivní, nekódující, cca čtvrtina genomu. Zřejmě mrtvé transpozony tak mutované že nejsou počítačem rozpoznány**

Naprostá většina naší DNA vznikla reversní transkripcí retrotranspozonů

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 CG

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

Lidská genetická variabilita

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

Sekvenování lidského genomu: Důsledky

Přínos sekvenování genomů

- **Milník v historii lidstva**
- **Usnadnění výzkumu molekulární podstaty chorob**
  - "virtuální klonování" in silico místo zdlouhavého pozičního klonování a sekvenování
- **Molekulární medicína**
  - dg. na úrovni genů, časná detekce predispozice k určitým chorobám
  - genová terapie
  - možnost stanovení individuální senzitivity k rizikovým faktorům v prostředí (radiace, mutageny)

Přínos sekvenování genomů

- **Nové cíle pro farmakoterapii**
  - všechny dosud známé léky - interakce jen asi s 500 proteinovými molekulami !!!
- **Farmakogenomika**
  - jak individuální genetická informace ovlivňuje reakci na léčbu ... možnost "terapie ušité na míru" pro každého pacienta
- **Studium evoluce a migrace lidského druhu**
- **Co vlastně genom kóduje ("nature vs. nurture") a jaké SNPs zodpovídají za rozdíly mezi lidmi**

Přínos sekvenování genomů: Kde jsme teď?

- **Překotný vývoj technologií pro dostupné sekvenování individuálních genomů**
- **Výzkum a počátky využití SNP**

Sekvenování se musí stát ještě rychlejší a lacinější aby mohlo být využíváno v běžné zdravotní péči...

- **J. Craig Venter Science Foundation, 2003:**
  - **$ 500 000 Genomic Technology Prize** ... pro vynálezce nové technologie sekvenování schopné přečíst savčí genom <$ 1000
- **X Prize Foundation, 2006:**

  $ 10 000 000 Archon X Prize for Genomics pro

  "první tým který postaví zařízení a úspěšně použije k sekvenování 100 lidských genomů za dobu 10 dní nebo kratší, s přesností maximálně jedné chyby na každých 100 000 sekvenovaných bazí, tak aby výsledné sekvence pokrývaly nejméně 98% genomu, a s náklady ne více než $10 000 na jeden genom."

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 $ (v roce 2010 <20 000 $)

První sekvenované individuální lidské genomy:

2007: Craig Venter, James Watson – oba genomy zpřístupněny na internetu

(od té doby několik dalších)

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 $ (v roce 2010 <20 000 $)

Závody ve vývoji sekvenování genomu <1000 $:

- Illumina, Pacific Biosciences, Complete Genomics, Helicos Biosciences, IBM a několik dalších usilují o to stát se prvním výrobcem dostupného osobního sekvenátoru … možná během 5 let?

Přínos sekvenování genomů: Kde jsme teď?

- **Překotný vývoj technologií pro dostupné sekvenování individuálních genomů**
- **Výzkum a počátky využití SNP**

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

Personal/Recreational Genomics

- Analýza osobních SNPs (původ předků, vrozené vlohy, náchylnost k chorobám…)
- V současnosti 4 firmy:
  - **Personal Genome Project**
  - **deCODE ME**
  - **Navigenics**
  - **23andME**

23andME

- Vzorek sliny zaslaný DHL, genotypizace 580 000 SNPs
- Ancestry:
  - Paternal (Y chromosome haplogroup)
  - Maternal (mitochondrial DNA haplogroup)
  - Finding relatives, global similarity search
- Health:
  - Disease risk: 92 (25 high confidence)
  - Carrier status: 24
  - Response to drugs: 18 (8 high confidence)
  - Traits: 44 (13 high confidence)

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**
- **Behavioral genetics:**
  - vztah genů k lidskému chování, možný vývoj ke genetickému determinismu a ztrátě odpovědnosti za vlastní chování
- **GM potraviny**
- **Gene patenting:**
  - odhadem asi pětina lidských genů nebo jejich fragmentů je patentována, většinou biotechnologickými firmami

References:

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

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

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

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

Lecture by dr. M. Lebl: Next generation DNA sequencing (1.LF UK, 1.12.2008)

http://www.ncbi.nlm.nih.gov

http://genomics.energy.gov

http://en.wikipedia.org

http://hapmap.ncbi.nlm.nih.gov

https://www.23andme.com

and internet in general…

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