MUNI | RECETOX
SCI

Analýza genomických a proteomických dat

Mgr. Eva Budinská, Ph.D.

Jaro 2024

Osnova přednášek

MUNI | RECETOX

I. Současné výzvy a technologie genomiky a proteomiky
II. Princip a analýza obrazu DNA mikročipů
III. Úprava a normalizace dat cDNA mikročipů
IV. Úprava a normalizace dat oligonukleotidových mikročipů
V. Princip, úprava a normalizace dat dalších mikročipů (Epigenetické mikročipy, Illumina BeadChip, SNP chip...)
VI. Proteomika
VII. Společné principy analýzy genomických a proteomických dat
VIII. Porovnávání tříd
IX. Predikce tříd
X. Objevování tříd
XI. Analýza přežití a další regrese
XII. Analýza genových sad a genových sítí
XIII. Meta-analýza

Požadavky

MUNI | RECETOX

- Individuální projekt (**15 bodů**) - 40% z celkového hodnocení zkoušky

- Písemná zkouška (**20 bodů**) – 50% z celkového hodnocení zkoušky

- Aktivita a přítomnost na cvičeních a prezentaci projektu (**5 bodů**) – 10% z celkového hodnocení zkoušky

- Úspěšné absolvování:
  - **min 21 bodů**, z toho min **8** z projektu a **min 10** ze zkoušky

POZOR – ke zkoušce se lze přihlásit pouze po odevzdání projektu který bude ohodnocen minimálně 8 body!

Hodnocení projektu trvá max 5 pracovních dní - nutno zohlednit pro plánování přihlašování ke zkoušce.

Požadavky k projektu a zadávání projektů

MUNI | RECETOX

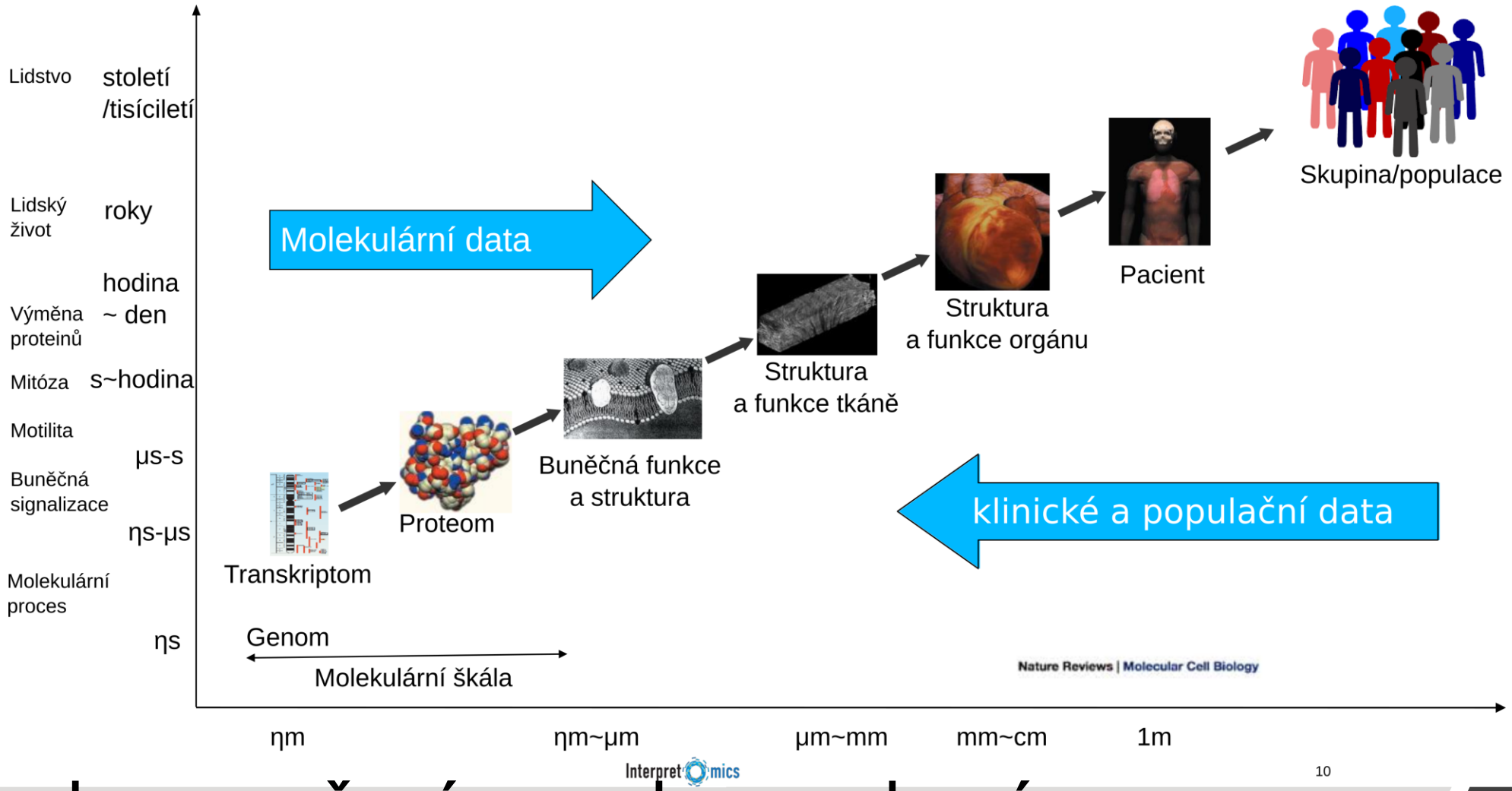
- Zpracovává se samostatně
- Možnost zpracovávat vlastní data nebo data z veřejné dostupných databází
- Výběr ze stanovených projektů, vlastní téma nutno schválit předem – **nejzazší termín výběru projektu**: **13.3.2024**
- Projekt nutno odevzdat před zkouškou, pouze po odevzdání a obdržení **8** bodů z projektu je možné přihlásit se na zkušební termín
- Nejzazší termín odevzdání projektů pro kontrolu počtu bodů: **5 pracovních dní** před zkušebním termínem
- **Student navrhne svůj projekt** v **rozpisu témat na IS,** po zkontrolování/případné úpravě vyučujícím, tento potvrdí téma a přihlášení studenta k projektu

Požadavky vypracování projektu

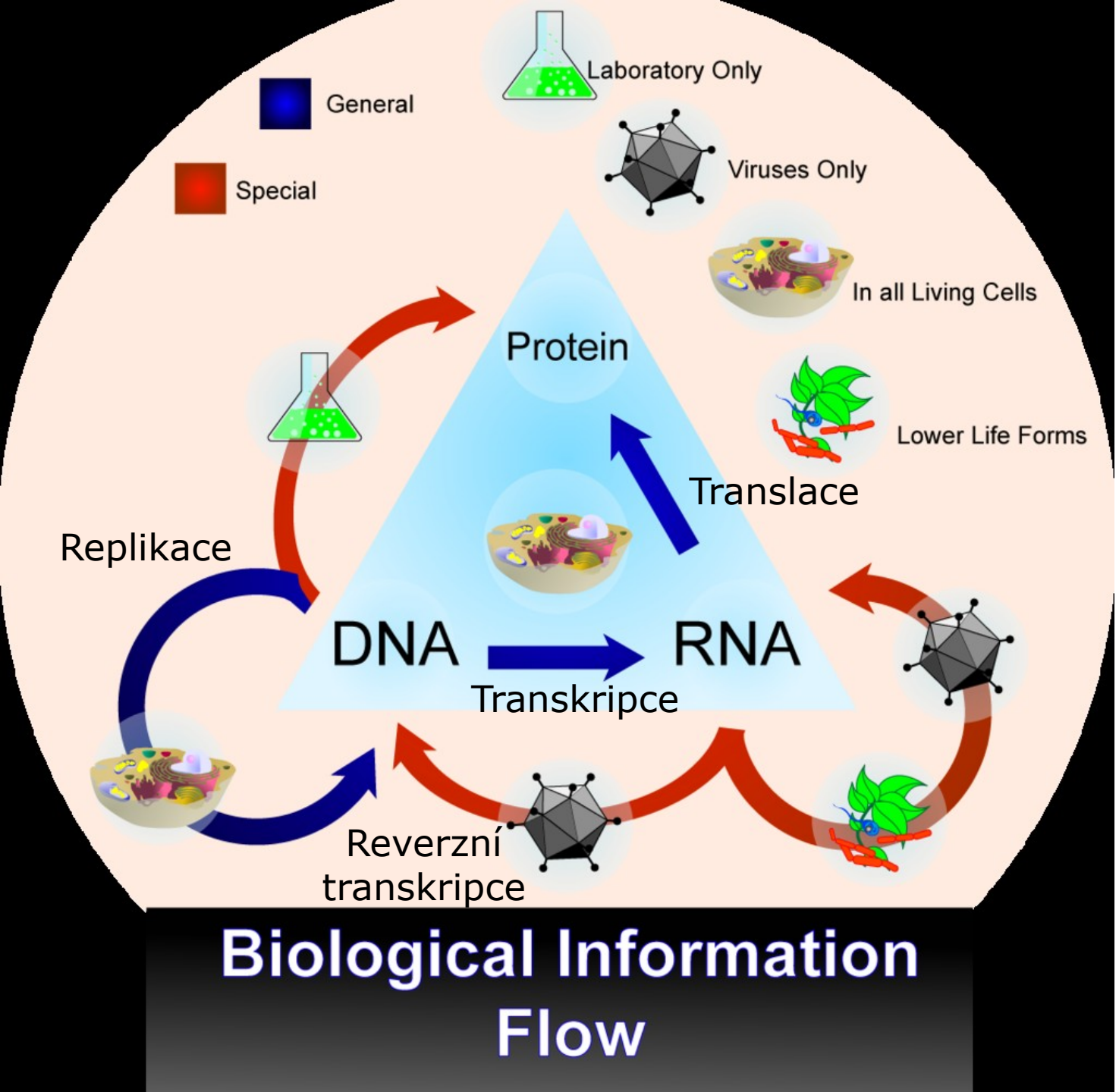
MUNI | RECETOX

- **Student odevzdává 2 soubory**:
  - **popis projekt**u ve formátu .pdf ve struktuře definované níže
  - **.R soubor se skriptem analýzy** od načtení dat po finální grafy
- Struktura **popisu projektu (pdf)**:
  - **Název**
  - **Úvod** – co je cílem projektu, přesně definované hypotézy
  - **Data** – přesně definovaný typ dat, odkaz na stažení dat, počet vzorků, typ platformy ze které byly data získány, kolik bylo na platformě sond, kolik genů reprezentovaly, v případě dvoukanálového experimentu jasná definice vzorků v jednotlivých kanálech, ...
  - **Metodika** – jaké metody zpracování dat od úpravy až po finální interpretaci byly použity a proč
  - **Výsledky** – výsledková část rozdělená na
    - a. Předzpracování a normalizace základních dat (popis, grafy, interpretace výsledků vzhledem k dalším analýzám)
    - b. Statistická analýza a data mining – rozděleno dle typu analýzy, popis nejdůležitějších výsledků a jejich sumarizace, grafy (např. venovy diagramy, heatmapa, volcano plot, forest plot...), sumární tabulky výsledků, odkazy na tabulky s podrobnými výsledky
    - c. Biologická interpretace
- Struktura **.R souboru se skriptem analýzy**:
  - Skript rozdělený do kapitol podle analýzy dat s podrobným komentářem jednotlivých kroků

Současné výzvy genomiky a proteomiky



Mnohorozměrná povaha moderní biomedicíny



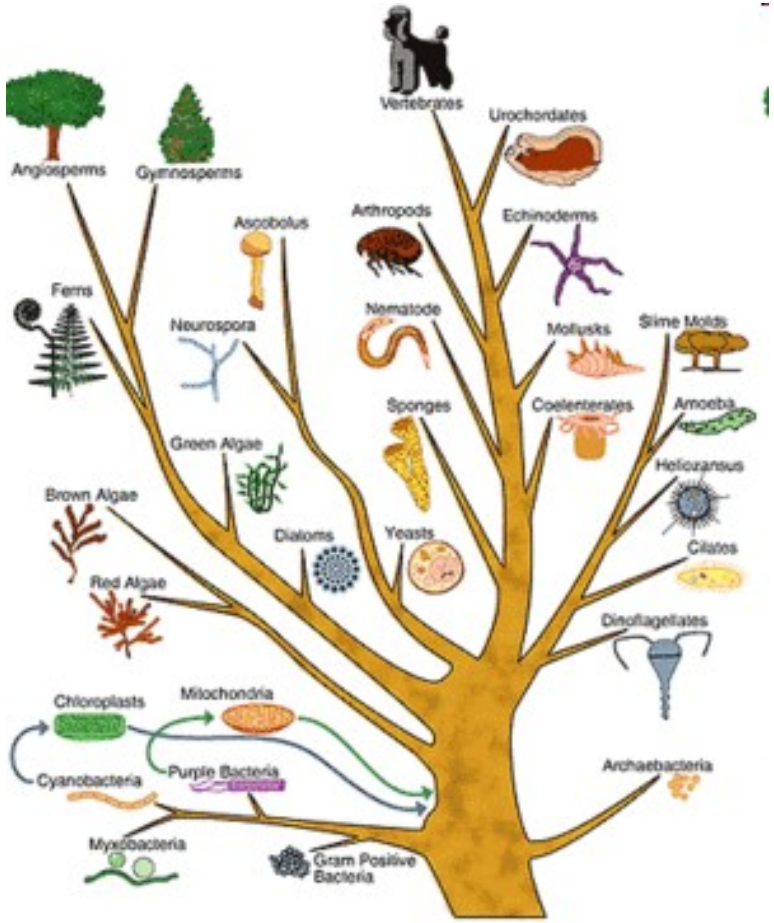
Centrální dogma molekulární biologie

Genomika je věda zabývající se studiem souboru genů v buňce (genom)

Proteomika je věda zabývající se studiem souboru proteinů v buňce (proteom)

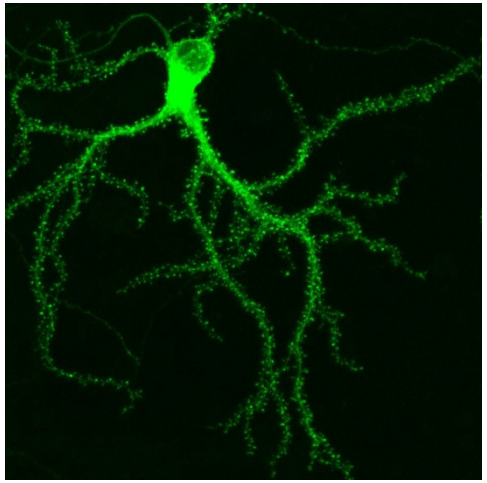
Geny podmiňují fyzický **vzhled** organismu a jeho **schopnost adaptace** na prostředí, ve kterém žije a jeho pomalé i náhlé změny (stres).

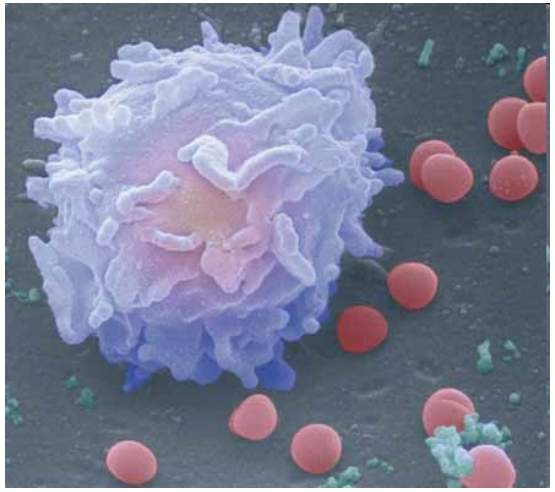
Rozdíly mezi organismy jsou podmíněné rozdíly v genomu.



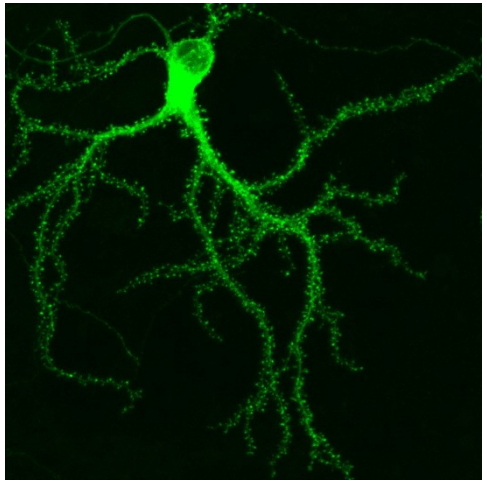
- Odolnost bakterií na antibiotika podmíněná mutacemi.
- Adaptace na extrémní podmínky - život ve vesmíru, v sopce, sirných pramenech, vařících pramenech a mrazech do -70

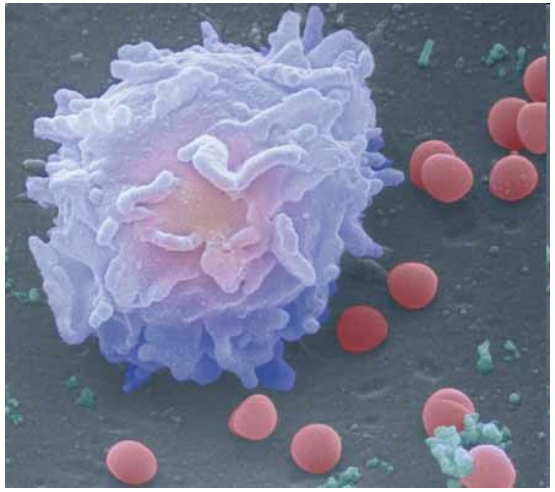
Jak je možné, že se navzájem liší i buňky v rámci jednoho organismu, když mají stejnou sadu genů?





Jak je možné, že se navzájem liší i buňky v rámci jednoho organismu, když mají stejnou sadu genů?





Tyto rozdíly jsou důsledkem odlišné aktivity genů a jejich produktů, proteinů a funkčních RNA molekul.

Genomika a proteomika v BIOLOGII

MUNI | RECETOX

Dekódování genomu u různých druhů

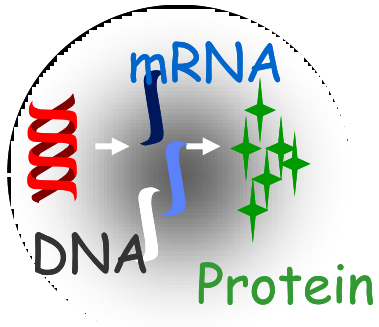
Můžeme studovat

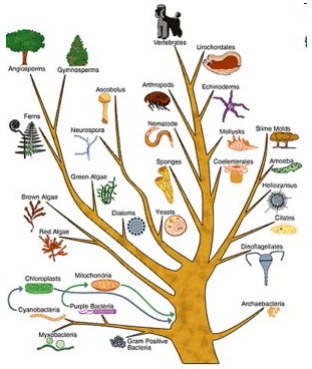
⇩

Rozdíly v genomu/proteomu jednotlivých druhů

⇩

studovat tak evoluční propojení a vytvářet fylogenetické stromy





⇩

aktivitu genů a proteinů organizmů v rozdílných podmínkách

⇩

Můžeme pochopit mechanismy působení parazitů a jejich přizpůsobení se hostiteli, případně studovat bakterie a jejich mechanismy přizpůsobení se extrémním podmínkám …

Genomika a proteomika v MEDICÍNĚ

MUNI | RECETOX

Studium genetické podstaty dědičných i získaných onemocnění

⇩

Můžeme studovat

⇩ ⇩

Genetické mutace, a jiné genetické/genomické aberace způsobující nemoci

Rozdílnou aktivitu genů a proteinů u konkrétních nemocí v porovnání se zdravým organismem

⇩

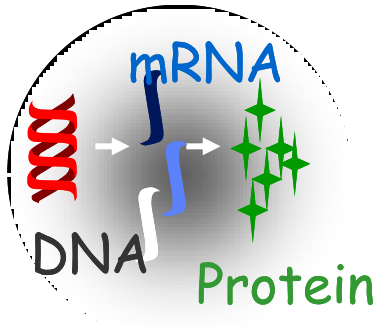
Jsme schopní
korelovat funkci produktů jednotlivých genů s **onemocněním**
NEMOC ⇔ GEN (Y)

⇩

Pochopit **podstatu** onemocnění

⇩

Najít nejvhodnější způsob léčby (cílená léčba),
prevence a diagnostiky onemocnění



Geny a onemocnění I. - příčiny

- Downův syndrom, hemofilie, cystická fibróza, svalová dystrofie, rakovina...
- Dědičné i získané, u některých stačí jediná *mutace* v patřičném genu a vzniká choroba, u jiných je zapotřebí více genetických změn

1. Změny ve struktuře DNA:
   - Mutace ve struktuře jednoho genu (jednonukleotidové polymorfizmy, delece, inzerce, amplifikace nukleotidů)
   - Aberace celého genu a nebo části chromozomu (delece, translokace, inzerce, amplifikace)
   - Aberace celých chromozomů

2. Změny v expresi a aktivitě genů a jejich produktů
3. Změny v posttranslačních úpravách proteinů

Geny a onemocnění II. - mutace

MUNI | RECETOX

- Buňky v organismu se stále obnovují a dělí - při každém dělení replikují celý genom na nukleotid přesně. Tento proces není při velikosti lidského genu (3.2 bilionu nukleotidů) jednoduché.

- Proto existuje mnoho kontrolních mechanismů:
  - na opravu poškozené časti DNA
  - pro správnou distribuci chromozomů v procese mitózy/meiózy
  - pro případnou apoptózu (regulovanou smrt buňky) v případě nezvratných změn
  - apod.…

- Genetické aberace vznikají selháním kontrolních mechanismů

Geny a onemocnění III. – aktivita genů

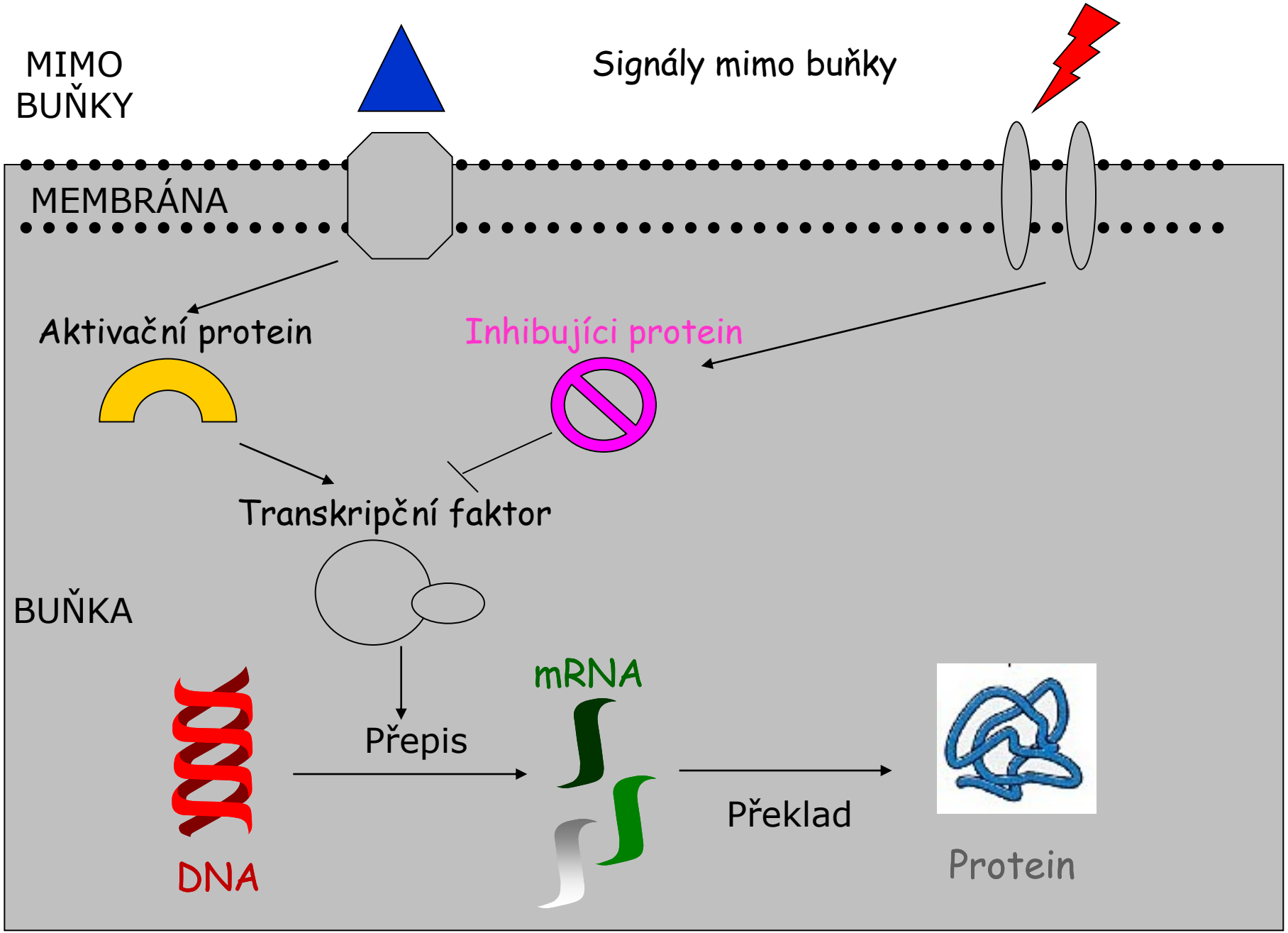
- Nejen mutace, ale i *nesprávná aktivita* genů může vést ke vzniku onemocnění.
- V lidské buňce probíhá každou chvíli obrovské množství procesů, přepisují se stovky genů a neustále se vytvářejí proteiny na základě vnitřních a venkovních podnětů.
- Tyto podněty jsou regulované stovkami regulačních mechanismů, které jsou opět založené na proteinech.
- Chyba v jednom z mechanismů může také skončit vyvinutím onemocnění.

Geny a onemocnění IV. - shrnutí

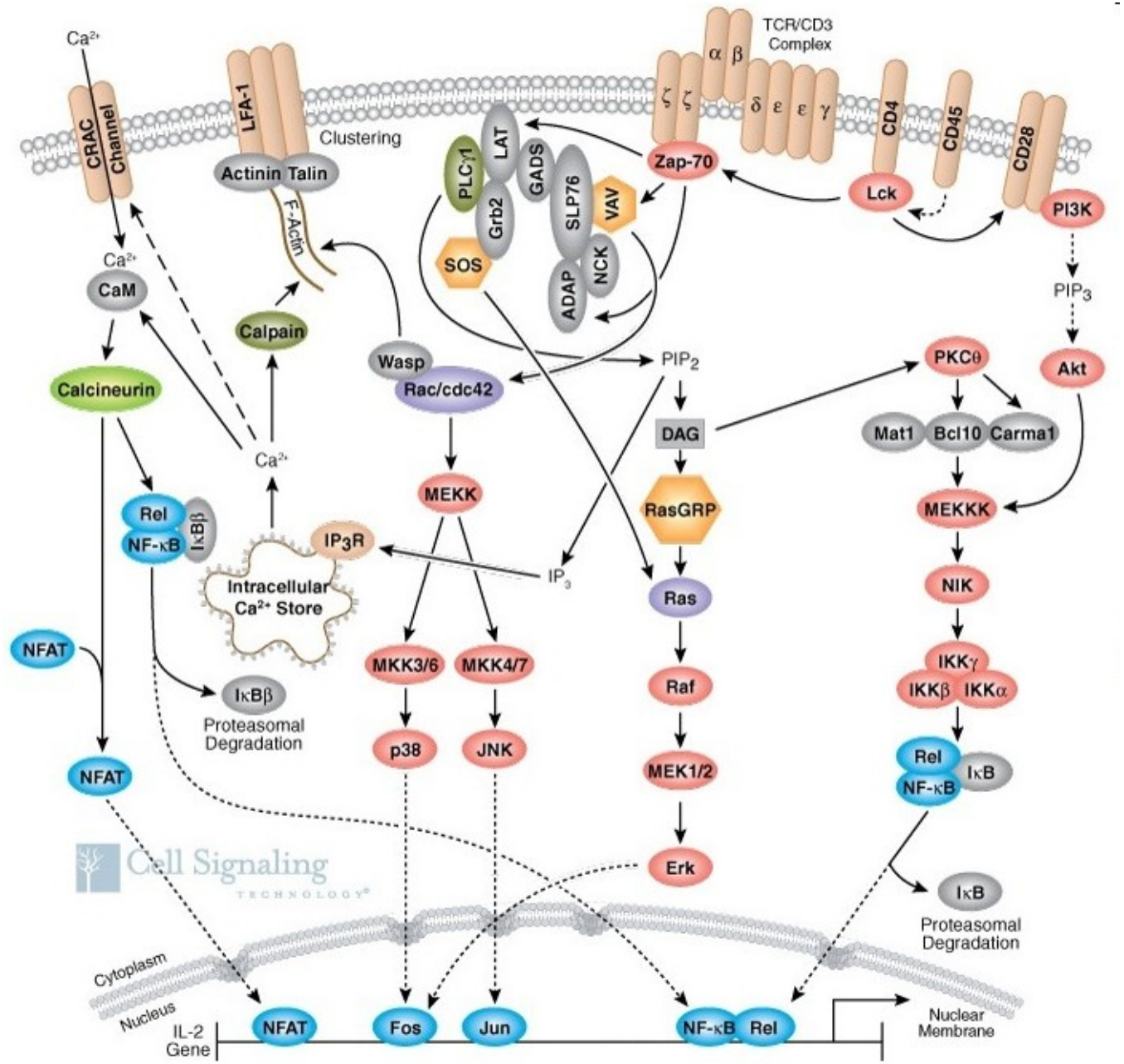
- Co způsobuje onemocnění – **proteiny a jiné funkční molekuly**, které mají změněnou svojí funkčnost, nebo expresi.

- Příčiny nesprávné funkce:
  - Mutace v příslušném genu, způsobující v důsledku změnu v sekvenci aminokyselin proteinu a tím jeho:
    - nefunkčnost
    - nadměrnou aktivitu
  - Změny v mechanismech kontroly exprese daného proteinu, který je následně produkovaný
    - v nedostačujícím množství
    - v nadměrném množství
  - Změny v postranslačních úpravách a sekundární/terciární struktuře proteinu

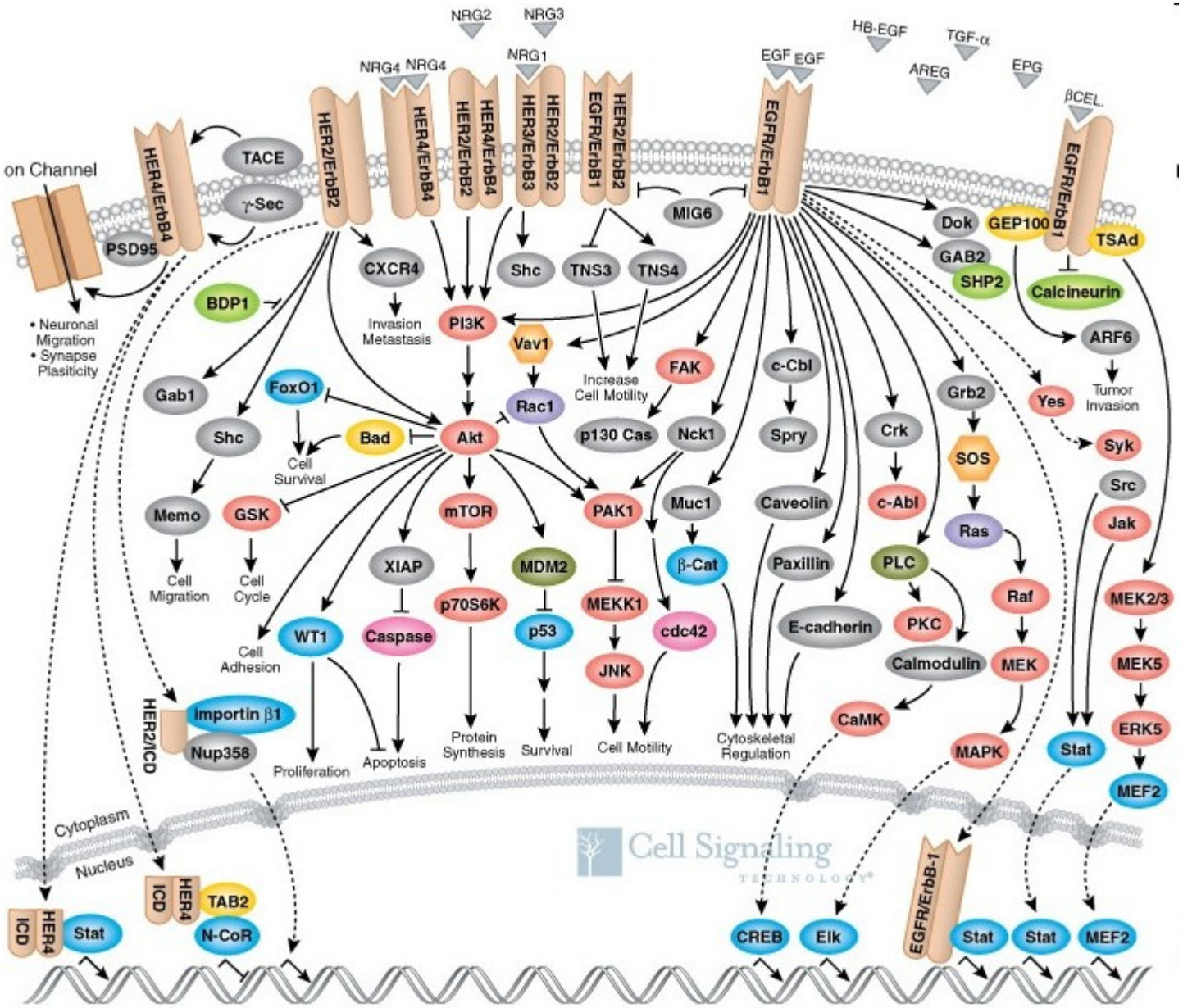
Co ještě víme



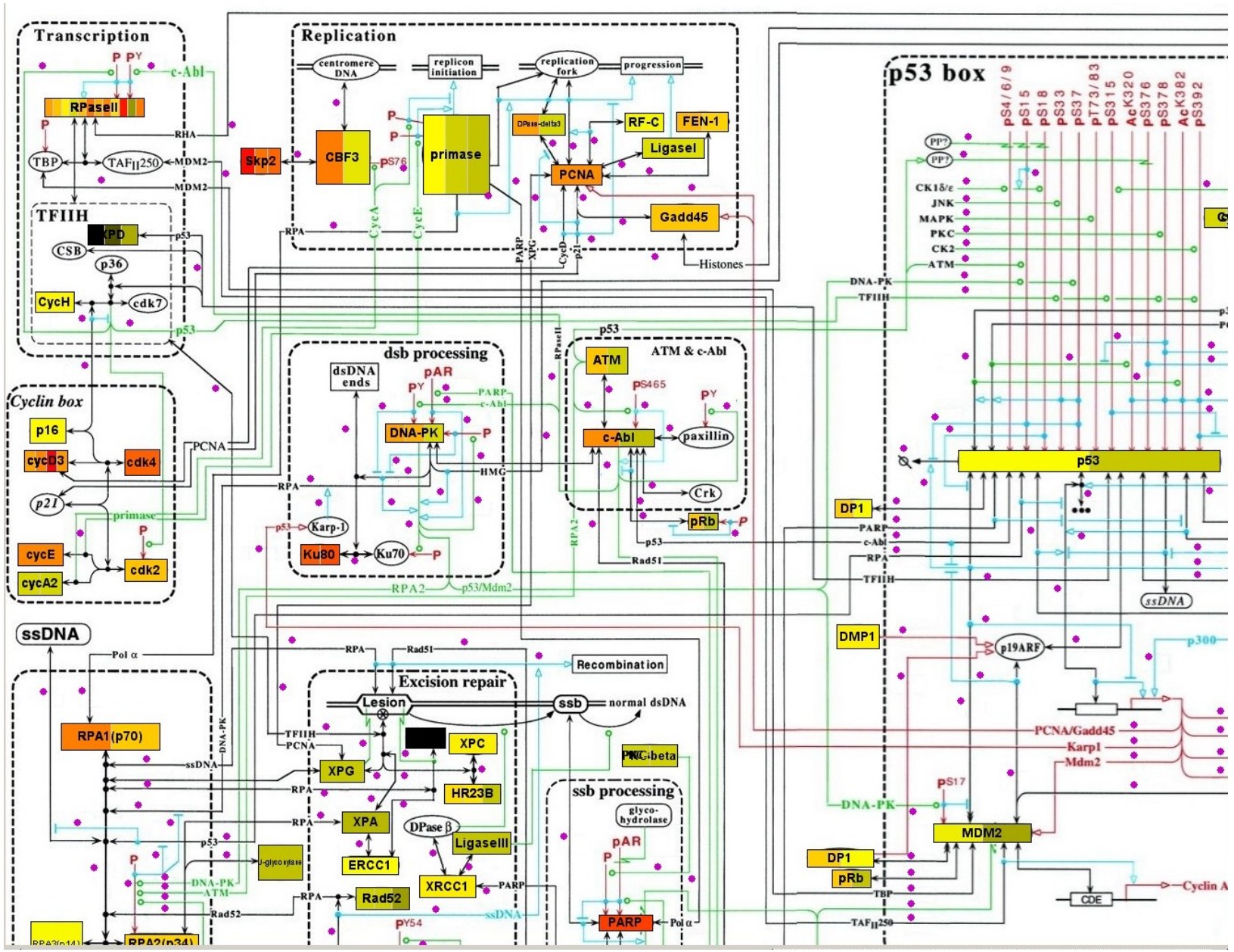
Ale víme ještě víc



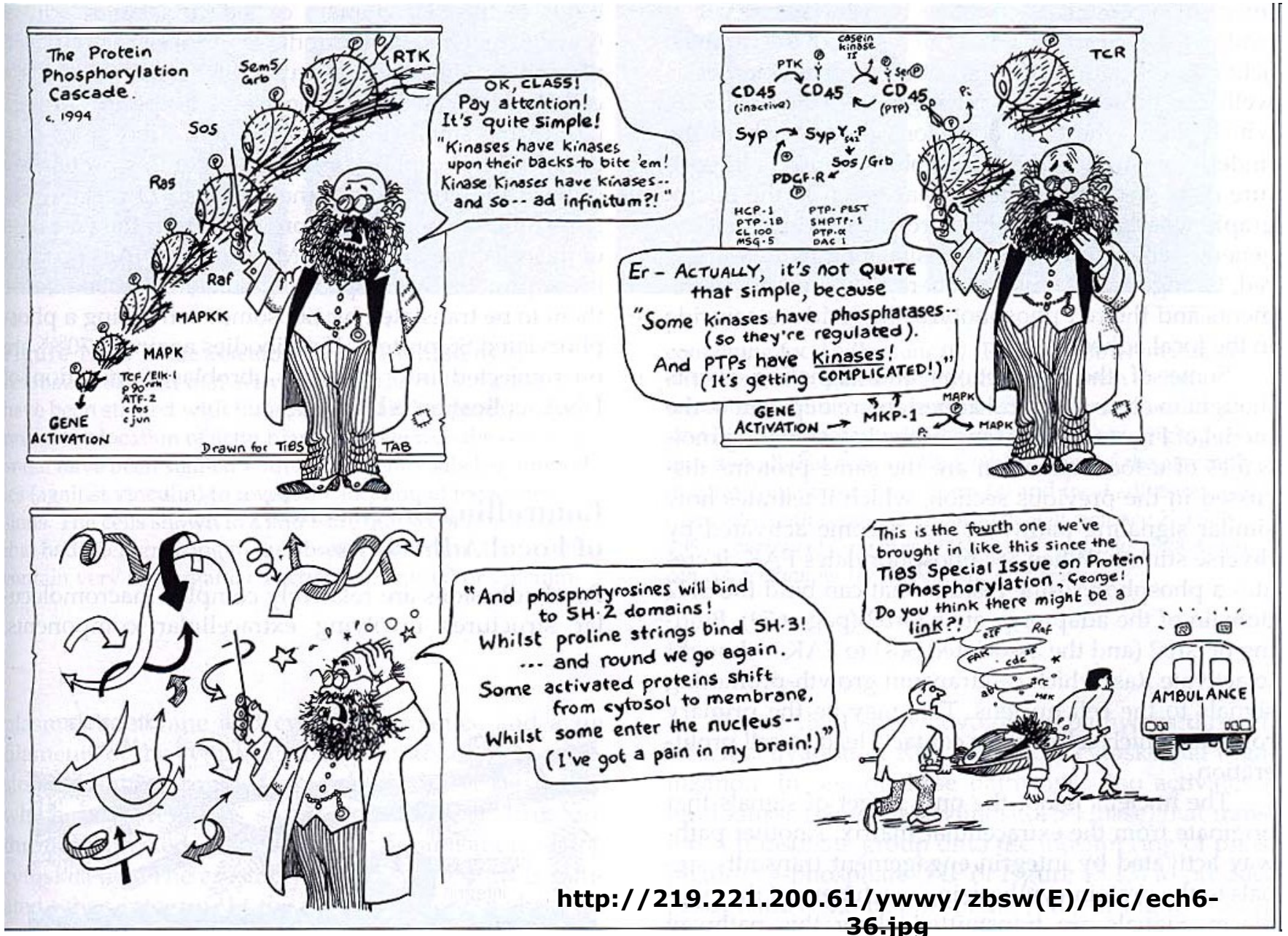
..a ještě víc...


NRG2
NRG3
NRG1
NRG4
NRG4
HB-EGF
TGF-α
AREG
EPG
EGF
EGF
βCEL
HER2/ErbB2
HER4/ErbB4
HER4/ErbB4
HER2/ErbB2
HER4/ErbB4
HER3/ErbB3
HER2/ErbB2
EGFR/ErbB1
HER2/ErbB2
EGFR/ErbB1
EGFR/ErbB1
on Channel
TACE
γ-Sec
PSD95
MIG6
Dok
GEP100
GAB2
SHP2
TSAd
Calcineurin
CXCR4
Shc
TNS3
TNS4
BDP1
PI3K
Vav1
Invasion
Metastasis
• Neuronal Migration
• Synapse Plasiticity
ARF6
Tumor Invasion
Gab1
FoxO1
Rac1
Increase Cell Motility
FAK
c-Cbl
Grb2
Yes
Shc
Bad
Akt
p130 Cas
Nck1
Spry
Crk
SOS
Syk
Cell Survival
Src
Memo
GSK
mTOR
PAK1
Muc1
Caveolin
c-Abl
Ras
Jak
Cell Migration
Cell Cycle
XIAP
MDM2
β-Cat
Paxillin
PLC
Raf
MEK2/3
WT1
Caspase
p70S6K
p53
MEKK1
cdc42
E-cadherin
PKC
MEK
MEK5
Cell Adhesion
JNK
Calmodulin
ERK5
Importin β1
Nup358
HER2/ICD
Proliferation
Apoptosis
Protein Synthesis
Survival
Cell Motility
Cytoskeletal Regulation
CaMK
MAPK
Stat
MEF2
Cytoplasm
Nucleus
Cell Signaling
TECHNOLOGY
ICD
HER4
Stat
ICD
HER4
TAB2
N-CoR
CREB
Elk
EGFR/ErbB-1
Stat
Stat
MEF2

...a ještě víc...


Transcription
Replication
p53 box
Cyclin box
dsb processing
ATM & c-Abl
ssDNA
Excision repair
ssb processing
Recombination
RPasell
c-Abl
TBP
TAFII250
MDM2
TFIIH
XPD
CSB
p36
CycH
cdk7
p53
Skp2
CBF3
primase
RF-C
FEN-1
Ligasel
PCNA
Gadd45
Histones
CycA
CycE
p21
p16
cycD3
cdk4
cycE
cycA2
cdk2
dsDNA ends
DNA-PK
Karp-1
Ku80
Ku70
ATM
c-Abl
paxillin
Crk
pRb
Rad51
RPA
RPA2
p53/Mdm2
HMG
PARP
pAR
Pol α
RPA1(p70)
RPA2(p34)
XPC
XPG
HR23B
XPA
ERCC1
XRCC1
Rad52
DPase β
LigaseIII
PKC-beta
glyco-hydrolase
PARP
Lesion
ssb
normal dsDNA
p53
DP1
DMP1
p19ARF
p300
MDM2
PCNA/Gadd45
Karp1
Mdm2
CDE
Cyclin A
ssDNA
CK1δ/ε
JNK
MAPK
PKC
CK2
ATM
DNA-PK
TFIIH

...ale je velmi obtížné to vše propojit a interpretovat



http://219.221.200.61/ywwy/zbsw(E)/pic/ech6-36.jpg

Co zkoumáme v genomice a proteomice

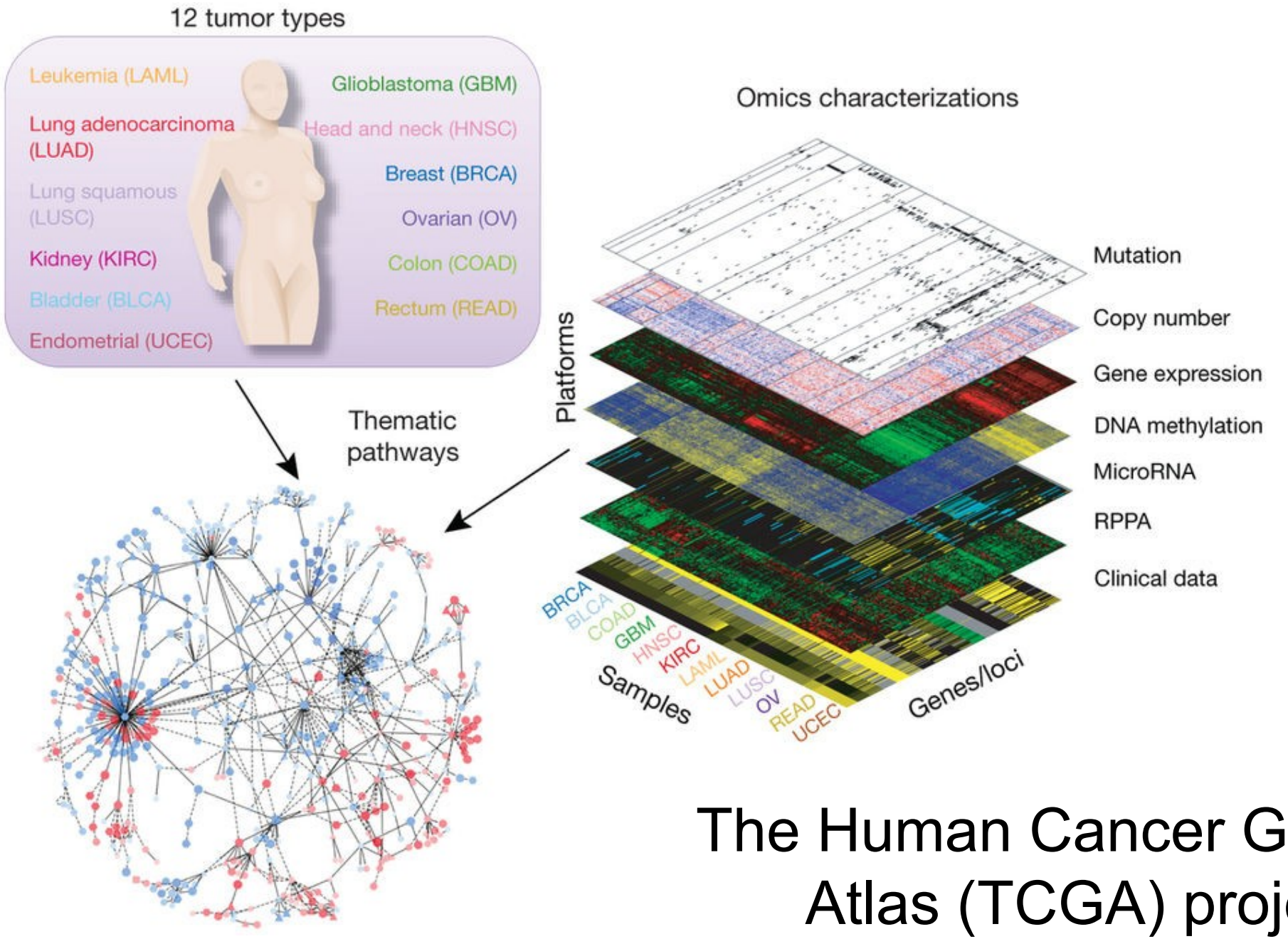
- U **genů** můžeme zkoumat jejich
  - **Strukturu a její změny** – sekvence nukleotidů A, C, G, T
  - **Množství** – zda jsou a nebo nejsou přítomné a v jakém počtu kopií
  - **Aktivitu** – zda se gen přepisuje do mRNA a v jakém množství
- U **proteinů** zkoumáme
  - **Složení** – z jakých aminokyselin
  - **Strukturu** – jak jsou řetězce peptidů uspořádané do 3D struktur
  - **Množství** – zda jsou a nebo nejsou přítomné a v jakém množství
  - **Funkci** – modelování, identifikace aktivních vazebných míst
- Další fáze je **modelování komplexních buněčných systémů** – proteinové interakce, buněčné dráhy, regulační a metabolické sítě …

Metody studia genomu a proteomu

- *Klasické metody* molekulární biologie a cytogenetiky:
  - Metody zkoumající jen jeden nebo několik genů a proteinů v jednom experimentu:
    - PCR, RT-PCR, real-time PCR
    - FISH (fluorescence in-situ hybridization)
    - gelová elektroforéza, ...

- *Vysocepokryvné metody* molekulární biologie:
  - schopné zkoumat tisíce molekul v jednom experimentu....
    - ... jak vznikly?

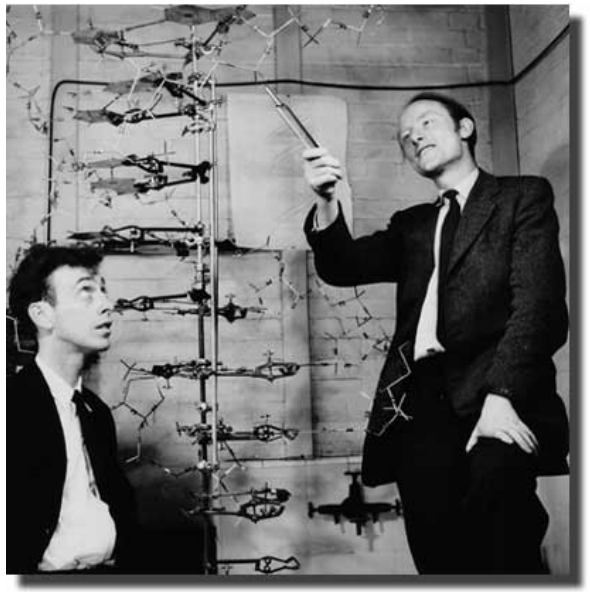
Proteomika a genomika

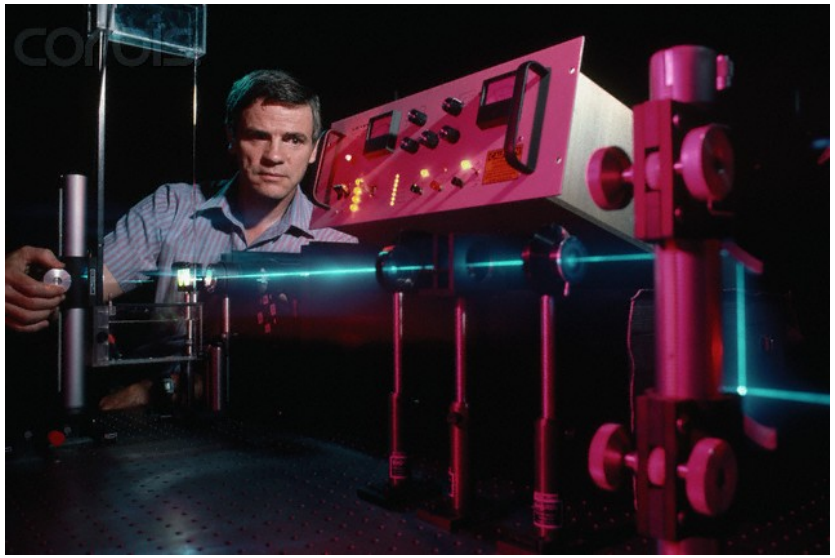


The Human Cancer Genome Atlas (TCGA) projekt

Od Watsona & Cricka po Leroya Hooda

- Na začátku byl dvoušroubovicový model DNA...





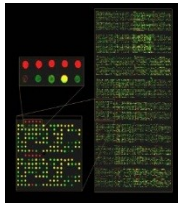
a na konci byly:

- automatické **sekvenátory** DNA a proteinů
- automatické **syntetizátory** DNA a proteinů

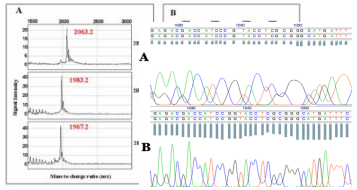
Nové možnosti

Sekvenátory umožnily rychle dekódovat sekvenci genů a proteinů

Znalost přesné sekvence umožnila navrhovat *specifické genové sondy* a syntetizátor umožňoval jejich rychlou a *automatickou výrobu*.

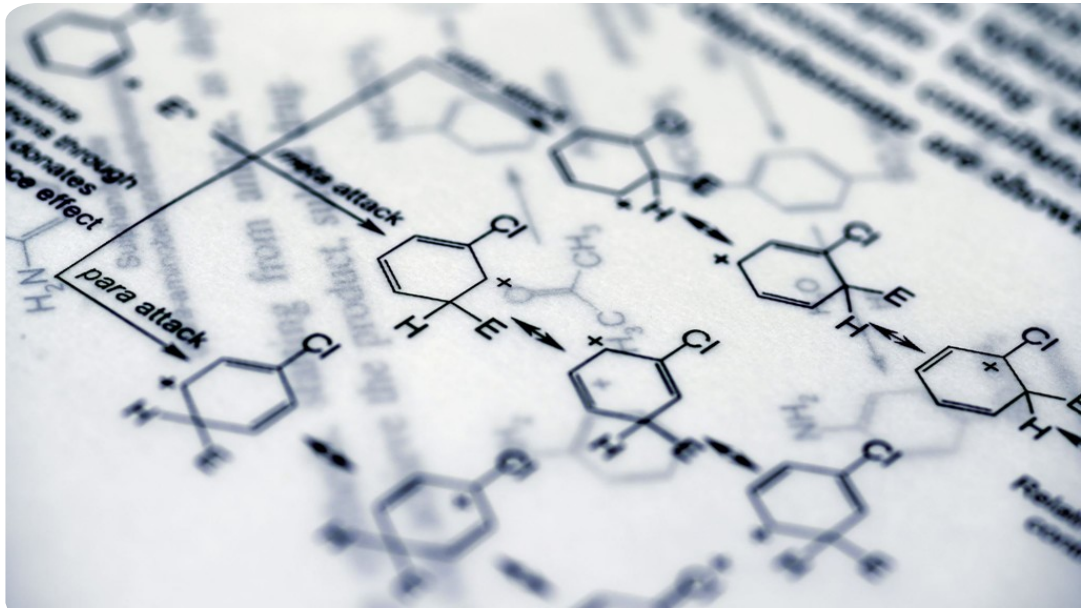


Otevřely se dveře pro nové, vysocepokryvní technologie, schopné analyzovat tisíce genů/proteinů v jednom experimentu!



Analýza genomu

- Od nukleotidových sekvencí po úplně anotovaný genom
- Analýza **struktury**
    - DNA sekvenace, Chip-seq, WES (whole exome sequencing), WGS
    - Srovnávací genomika – aCGH čipy, SNP polymorfismy, alternative splicing arrays, fingerprinting
- Analýza **aktivity** (exprese) – Mikročipy, SAGE, MPSS, Expressed sequence tags (ESTs), RNA-seq, …
- Regulace genomu
    - Chip-on-chip
    - Epigenetika (mikročipy, metylace...)



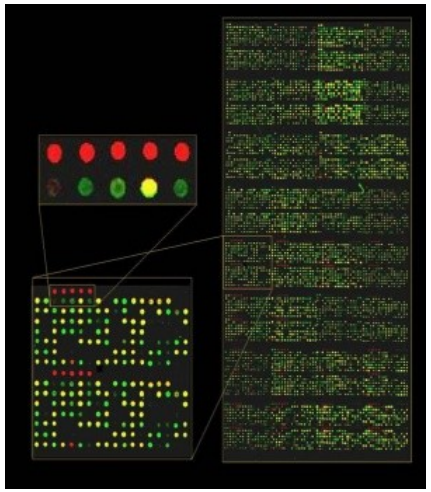
Analýza proteomu

- Od hmostnostních spekter – přes komplexní struktury proteinových shluků - po analýzu funkce proteinů
    - Analýza **struktury**: Proteinová sekvenace
    - Analýza **exprese**: Hmotnostní spektrometrie, 2D gelová elektroforéza, Proteínové mikročipy...
    - Analýza **funkce**: Modelování makromolekulárních systémů – odvození vlastností z atomových interakcí

Data z omics experimentů

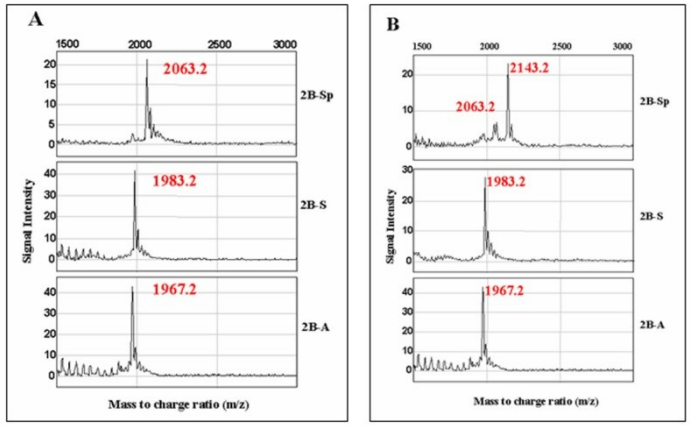
- Moderní vysoce pokryvné molekulární technologie produkují obrovské tabulky komplexních dat

Mikročipy



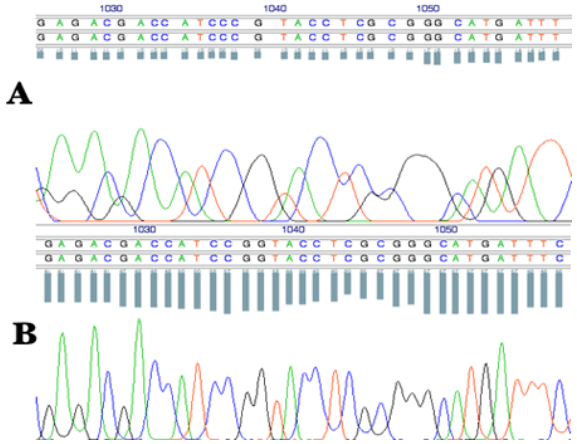
Desítky až tisíce genů nebo transkriptů na vzorek

Hmotnostní spektrometrie



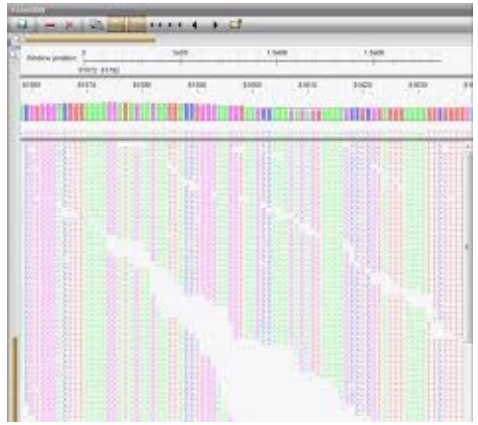
Tisíce spekter proteinů, metabolitů nebo malých molekul na vzorek

Sekvenace DNA



Genom s biliony nukleotidů na vzorek

Sekvenace nové generace



Miliony krátkých čtení DNA na vzorek

Genomická a proteomická data

Proč jsou data high-throughput genomických a proteomických experimentů problematická?

Obsahují **množství šumu** (technická i biologická variabilita)

...

...

...

...

...

Specifika dat z omics experimentů

Obsahují **množství šumu** (technická i biologická variabilita)

Nejsou skutečnými hodnotami (koncentrace, počty) sledovaných molekul

...

..

...

...

Specifika dat z omics experimentů

Obsahují **množství šumu** (technická i biologická variabilita)

Nejsou skutečnými hodnotami (koncentrace, počty) sledovaných molekul

Pocházejí z komplexních technologií, které bývají **velice citlivé na vnější vlivy**

…

…

…

Specifika dat z omics experimentů

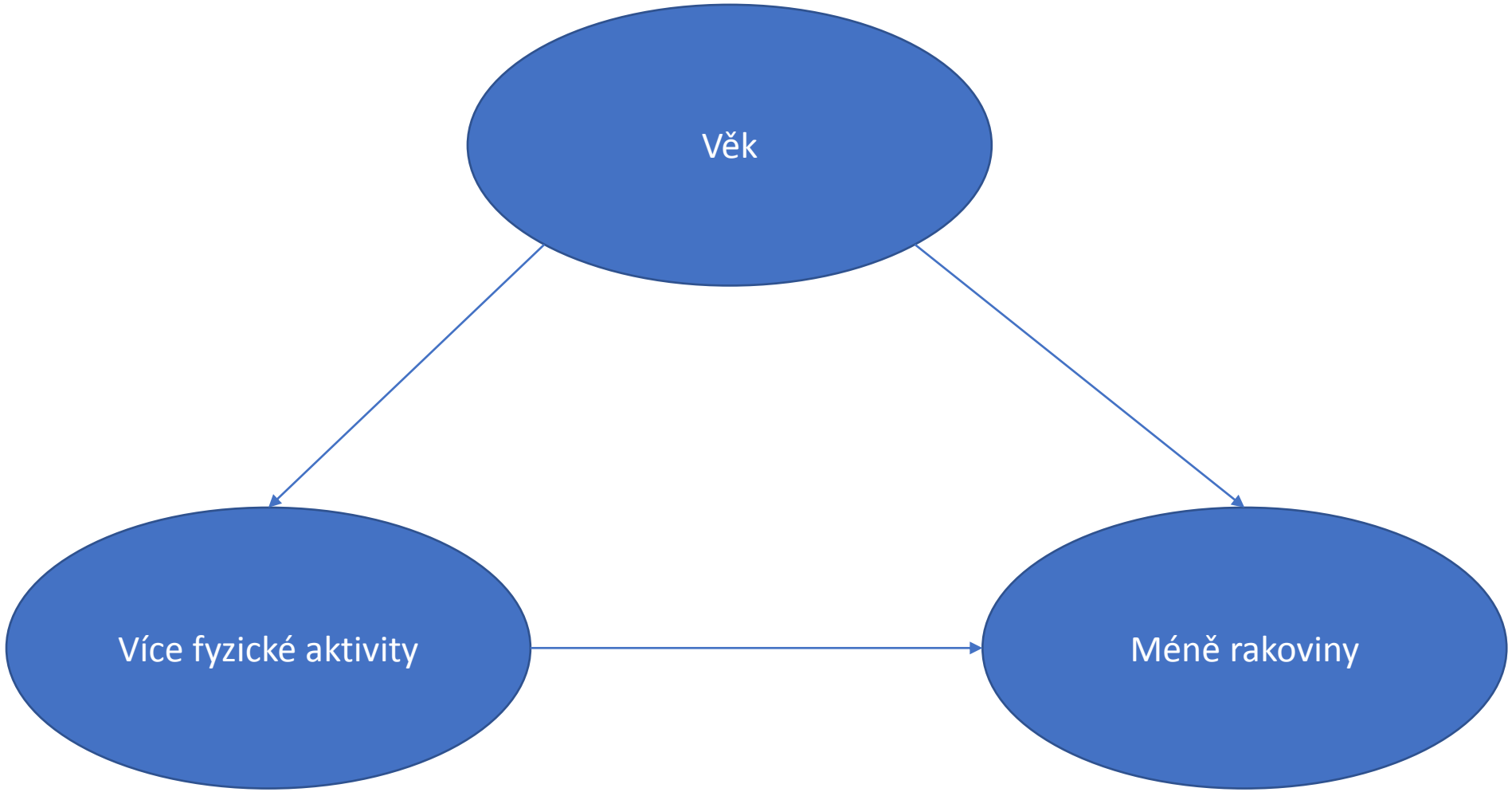
Za všechno mohou matoucí vlivy (confounding effects)?

Co je to matoucí faktor

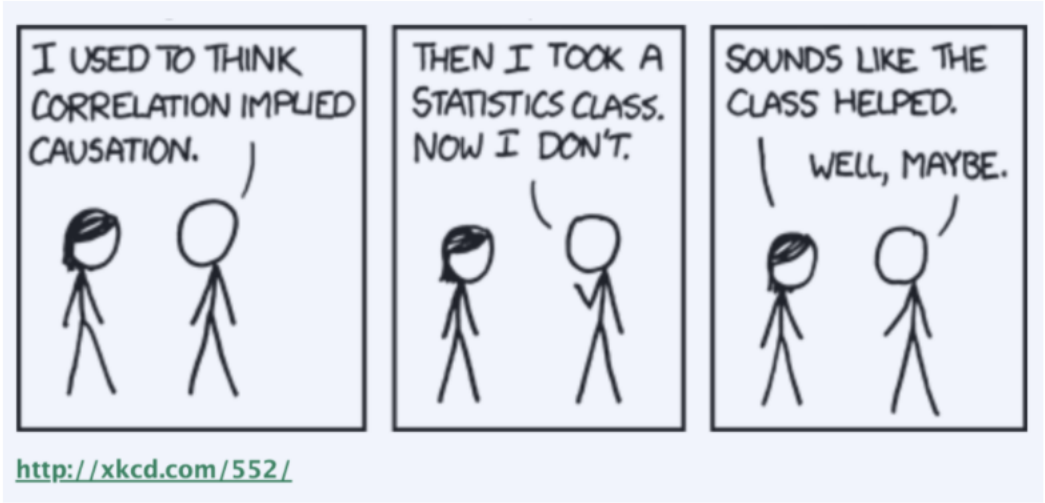
Matoucí faktor (*confounding factor*) je (neznámá) vnější proměnná, která ovlivňuje závislou proměnnou i nezávislou proměnnou v analýze, což způsobuje jejich falešnou asociaci a špatnou interpretaci.

Jiným způsobem, vzniká korelace, která není kauzalita….

Matoucí vliv



Pochybné korelace….

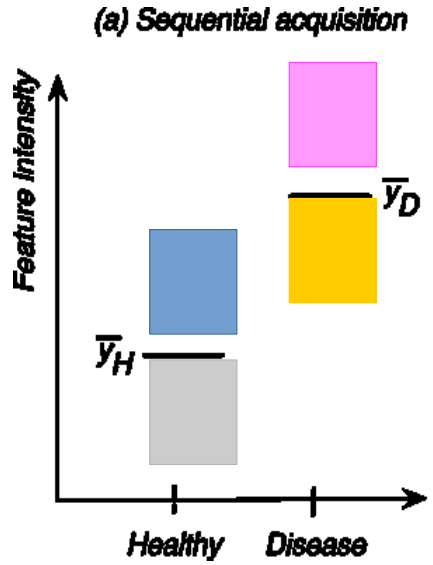


Efekt dávky

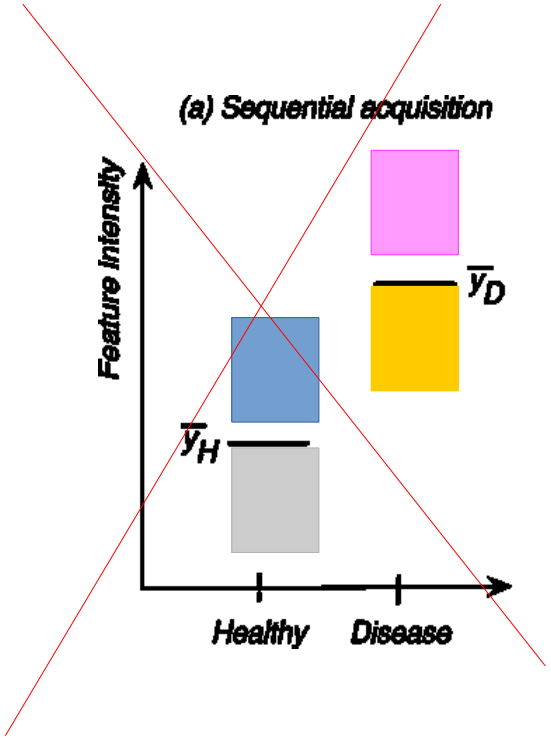
- Efekt dávky (*batch effect*) se objevuje vždy, když externí faktory spojené s laboratorní prací ovlivňují výsledky, které měříte ve studii.

- Efekt dávky je speciální typ matoucího faktoru v případě, že je dávka spojená s proměnnou, kterou sledujeme

Efekt dávky


(a) Sequential acquisition
Feature Intensity
$\overline{y}_D$
$\overline{y}_H$
Healthy
Disease

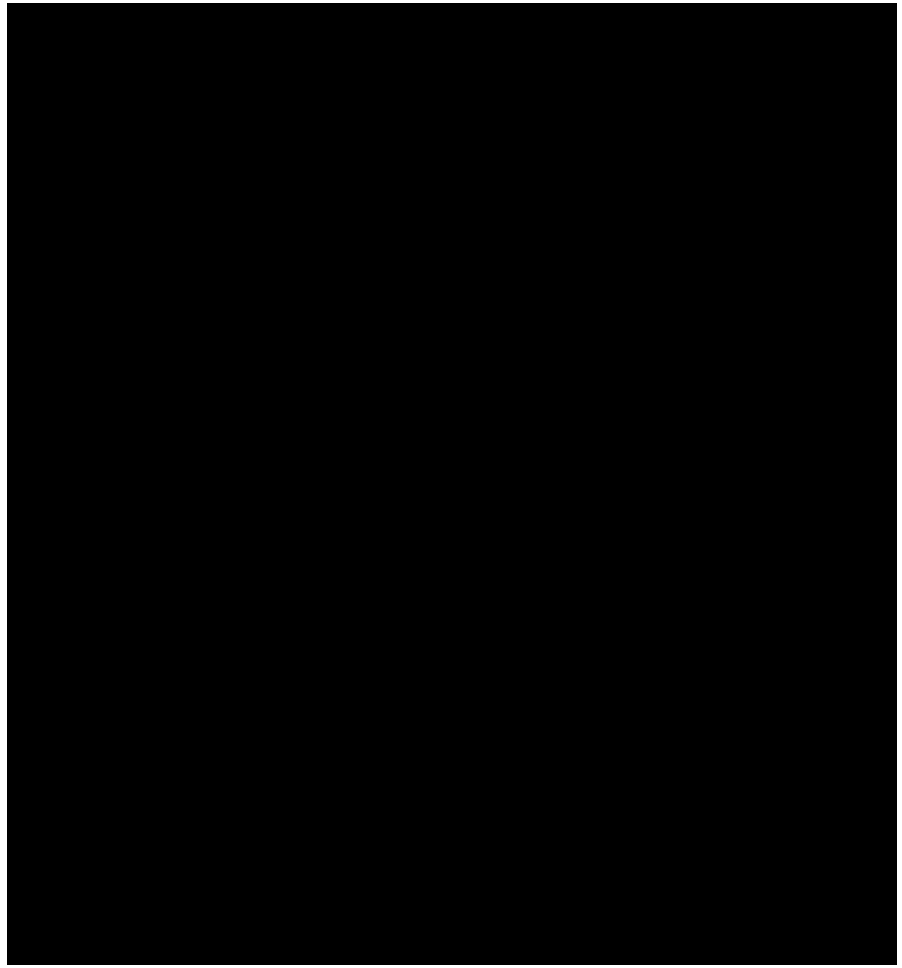
Efekt dávky



NENÍ MOŽNÉ STATISTICKY ODDĚLIT TECHNICKÝ EFEKT OD BIOLOGICKÉHO!!!

Lidé a myši na mikročipech

V článku z roku 2004, mikročipová analýza genové exprese několika různých tkání u lidí a myší vedla autory k závěru, že **„jakákoli lidská tkáň je více podobná jakékoli jiné vyšetřované lidské tkáni než její odpovídající tkáni myší“.**



Yanai I, Graur D, Ophir R. Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. OMICS. 2004 Spring;8(1):15-24.

Lidé a myši na mikročipech

Následují články (2006, 2007, 2010), které dokazují, že tyto rozdíly jsou založeny pouze na faktu, že se jednalo o dva různé mikročipy…:

1. Sondy na mikročipech jsou navrženy odděleně pro lidské a myší ortologické geny a necílí na stejné sekvence. Proto mají lidské sondy a myší sondy různé afinity k jejich cílovým RNA
2. Signál (S) detekovaný mikročipem je přibližně lineární se skutečným množstvím cílové RNA v rozumných rozsazích měření (Affymetrix 2001), hodnoty S transformované log2 mají tendenci přeceňovat rozdíl mezi dvěma nízkými hodnotami exprese, ale podceňují rozdíl mezi dvěma vysokými hodnotami exprese.

Lidé a myši na mikročipech

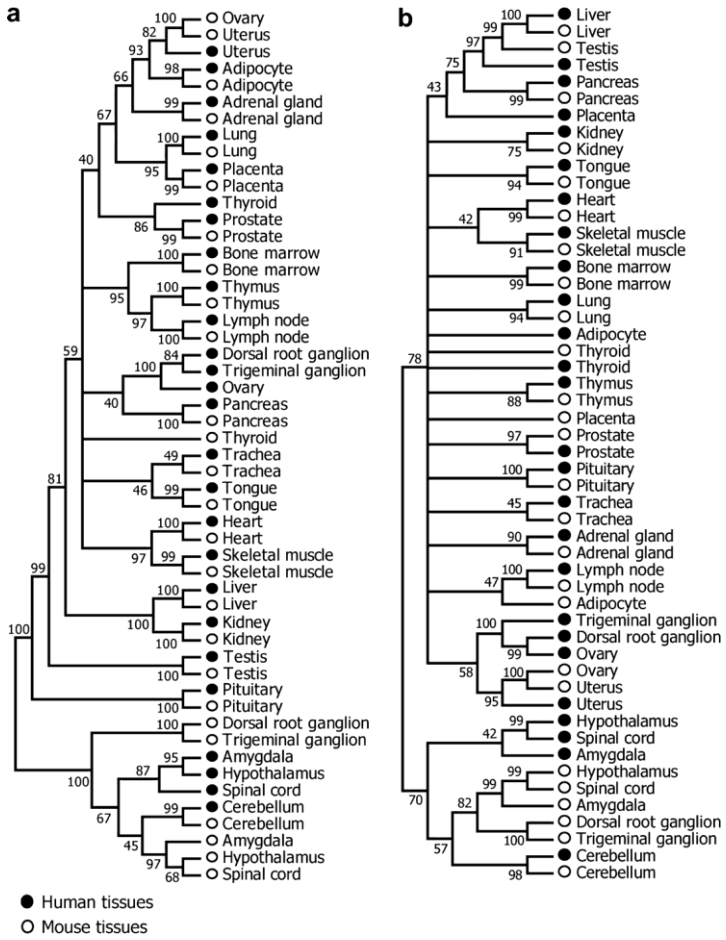


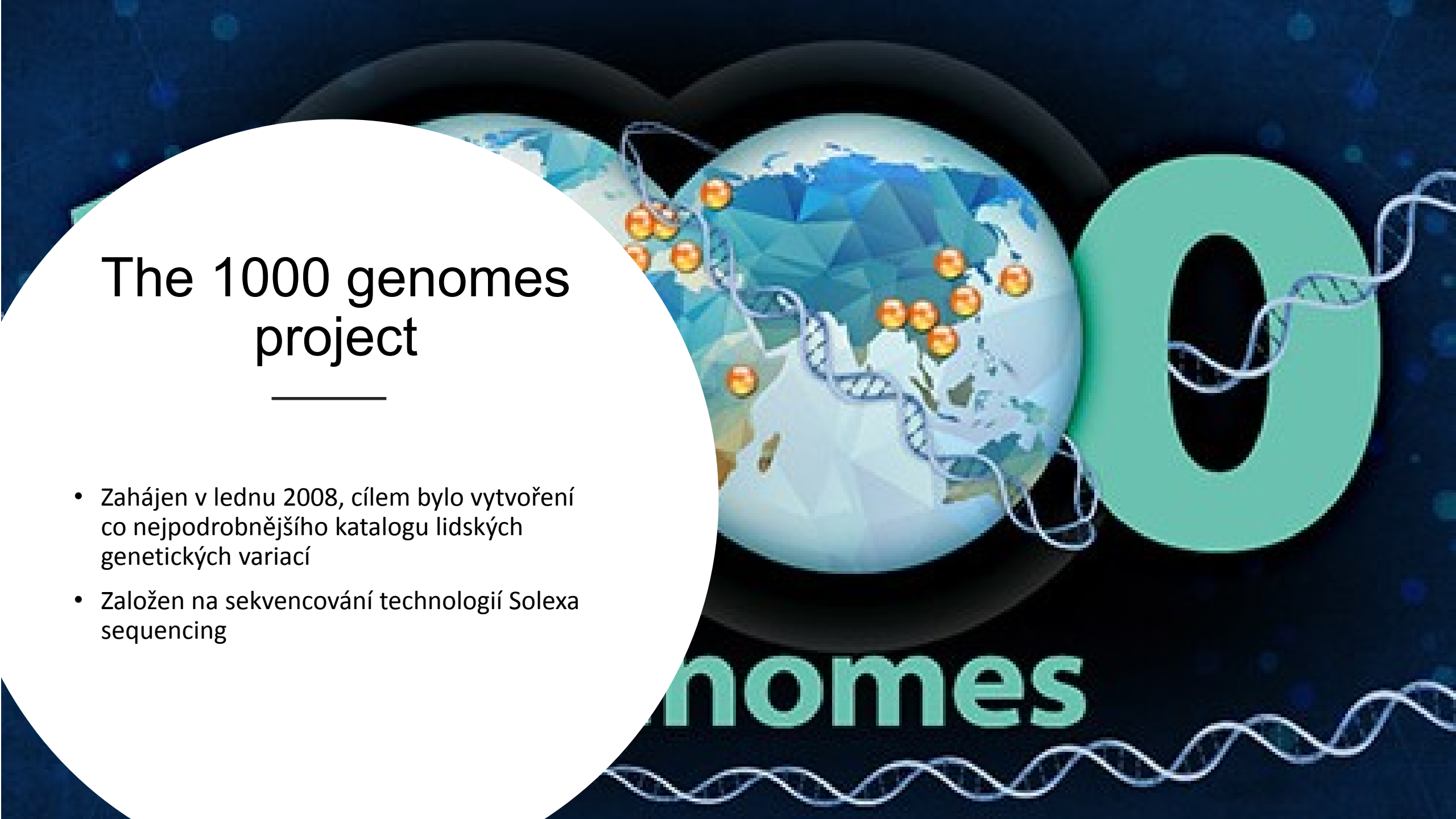
FIG. 5.—
Dendrograms of 26 human and 26 mouse tissues based on (a) 1 − Pearson's correlation coefficient r and (b) Euclidean distance d of tissues..

Ben-Yang Liao, Jianzhi Zhang (2006) Evolutionary Conservation of Expression Profiles Between Human and Mouse Orthologous Genes . Molecular Biology and Evolution, Volume 23, Issue 3, March 2006, Pages 530-540

OXFORD
UNIVERSITY PRESS

The 1000 genomes project

- Zahájen v lednu 2008, cílem bylo vytvoření co nejpodrobnějšího katalogu lidských genetických variací
- Založen na sekvencování technologií Solexa sequencing



Jaký je vliv data sekvencování na genetickou variabilitu mezi sekvencemi?

nature REVIEWS GENETICS

Opinion | Published: 14 September 2010

Tackling the widespread and critical impact of batch effects in high-throughput data

Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly & Rafael A. Irizarry

Nature Reviews Genetics **11**, 733–739 (2010) | Download Citation

5716 Accesses | **732** Citations | **182** Altmetric | Metrics »

Zjistili, že se studovanými biologickými rozdíly bylo spojeno pouze 17% variability sekvencí, zatímco neuvěřitelných 32% bylo možné vysvětlit datem, kdy byly vzorky zpracovány.

Obsahují **množství šumu** (technická i biologická variabilita)

Nejsou skutečnými hodnotami (koncentrace, počty) sledovaných molekul

Pocházejí z komplexních technologií, které bývají **velice citlivé na vnější vlivy**

Jejich předzpracování pro statistickou analýzu je **náročné** a **vysoce specifické** pro daný typ platformy

…

…

Specifika dat z omics experimentů

Obsahují **množství šumu** (technická i biologická variabilita)

Nejsou skutečnými hodnotami (koncentrace, počty) sledovaných molekul

Pocházejí z komplexních technologií, které bývají **velice citlivé na vnější vlivy**

Jejich předzpracování pro statistickou analýzu je **náročné** a **vysoce specifické** pro daný typ platformy

Počet vzorků je mnohem **menší než** počet sledovaných **proměnných**.

...

Specifika dat z omics experimentů

Obsahují **množství šumu** (technická i biologická variabilita)

Nejsou skutečnými hodnotami (koncentrace, počty) sledovaných molekul

Pocházejí z komplexních technologií, které bývají **velice citlivé na vnější vlivy**

Jejich předzpracování pro statistickou analýzu je **náročné** a **vysoce specifické** pro daný typ platformy

Počet vzorků je mnohem **menší než** počet sledovaných **proměnných**.

Zkoumané **proměnné jsou často korelované** a mají mezi sebou komplexní vztahy (geny, proteiny…)

Specifika dat z omics experimentů

Cíle předmětu

Podrobné představení technologií a analýzy jejich dat od předzpracování až po finální biologickou interpretaci.

- Mikročipy: cDNA, Affymetrix, Illumina
- Proteomická hmotnostní spektrometrie a gelová elektroforéza

Související doporučené předměty (povinné pro MBB)

- Analýza NGS dat – E5444 (podzim)
- Analýza non-target MS dat - E5020 (jaro)
- Detekce biomarkerů z omicsových eperimentů - E7527 (podzim)

Po absolvování předmětu student/studentka:

- Zná otázky, na které lze odpovědět studiem genomických a proteomických dat.; • **Zná vybrané technologie, které jsou zdroji vysokohustotních genomických a proteomických dat (typy DNA mikročipů, arrayCGH, hmotnostní spektrometrie);** • Zná typy dat produkovaných genomickými a proteomickými technologiemi a jejich problémy z pohledu analytika; • **Umí vyjmenovat obecné základní kroky analýzy genomických a proteomických dat.; • Zná pro analýzu potřebné technické detaily technologie mikročipů a hmotnostní spektrometrie.;** • Rozumí základním metodám kvantifikace měření těchto technologií a následným úpravám jejich hodnot při vytváření finálního datového souboru.; • **Zná specifické, na technologii závislé zdroje nežádoucího technického šumu v datech.;** • Pomocí grafických nástrojů a statistických metod dokáže identifikovat přítomnost a zdroj tohoto šumu v datech.; • **Umí aplikovat statistické metody používané pro odstranění nežádoucího šumu.;** • Dokáže provést transformaci zdrojových dat (normalizaci); • **Standardizuje měření mezi jednotlivými datovými soubory.;** • Vytvoří finální datový soubor pro další analýzy.; • **Dokáže identifikovat a odstranit efekt dávky (batch effect) v datech.;** • Popíše obecné principy analýzy genomických a proteomických dat.; • **Na základě biologického zadání správně určí typ potřebné analýzy.;** • Navrhne analytický postup pro příkladovou studii. ; • **Na základě biologického zadání správně určí typ hypotézy a metody jejího testování.;** • Rozumí specifickým metodám porovnávání hypotéz, jako je SAM a limma.; • Aplikuje testování hypotéz pro určení odlišně exprimovaných genů/proteinů.; • **Zná statistické metody predikce tříd a umí je aplikovat na příkladová data.;** • Zná metody dolování dat pro predikci tříd a umí je použít v genomických datech.; • **Zná pozitiva a negativa metod pro predikci a umí se rozhodnout, kterou aplikovat.;** • Aplikuje MAQC II standardy pro tvorbu klasifikátorů z microarray dat.; • **Vybere a aplikuje vhodné vícerozměrné regresní metody pro propojení vztahu exprese genů/proteinů s dalšími (např. klinickými) faktory.;** • Na genomických datech provede analýzu přežití pomocí Kaplan-Meierovych odhadů a Coxova modelu proporcionálních rizik.; • **Zná princip a metody analýzy genových sad;** • Zná princip a základní metody konstrukce genových sítí; • **Aplikuje na modelovém příkladu nejběžnější metody analýzy genových sad a pathway analýzy.;** • Vyhledá vysokohustotní genomické a proteomické data ve veřejných databázích.; • **Použije Fisherovu Z-transformaci a další meta-analytické transformace.;** • Aplikuje meta-analytické metody seřazení výsledků analýzy odlišné genové/proteinové exprese z více zdrojů.; • **Pracuje v SW prostředí R a Bioconductor pro úpravy a analýzu genomických a proteomických dat.;** • Aplikuje vybrané balíky R a Bioconductor a jejich funkce v úpravě a analýze dat.;