

Analýza genomických a proteomických dat

Mgr. Eva Budinská, Ph.D.

Jaro 2022

- 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. Úprava dat proteomické hmotnostní spektrometrie
- 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

- 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

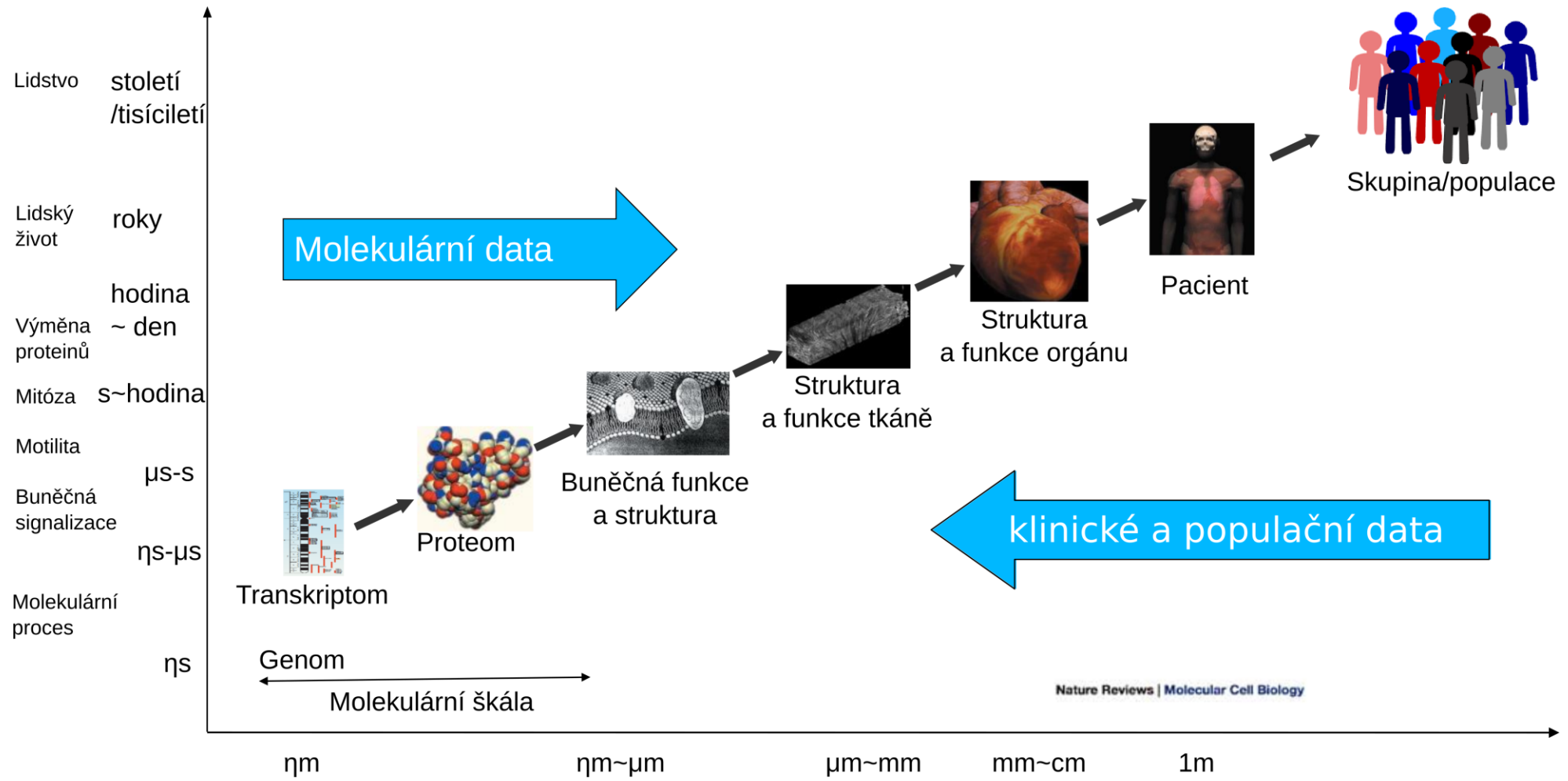
POZOR – ke zkoušce se lze přihlásit pouze po odevzdání projektu který bude ohodnocen minimálně 8 body!

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

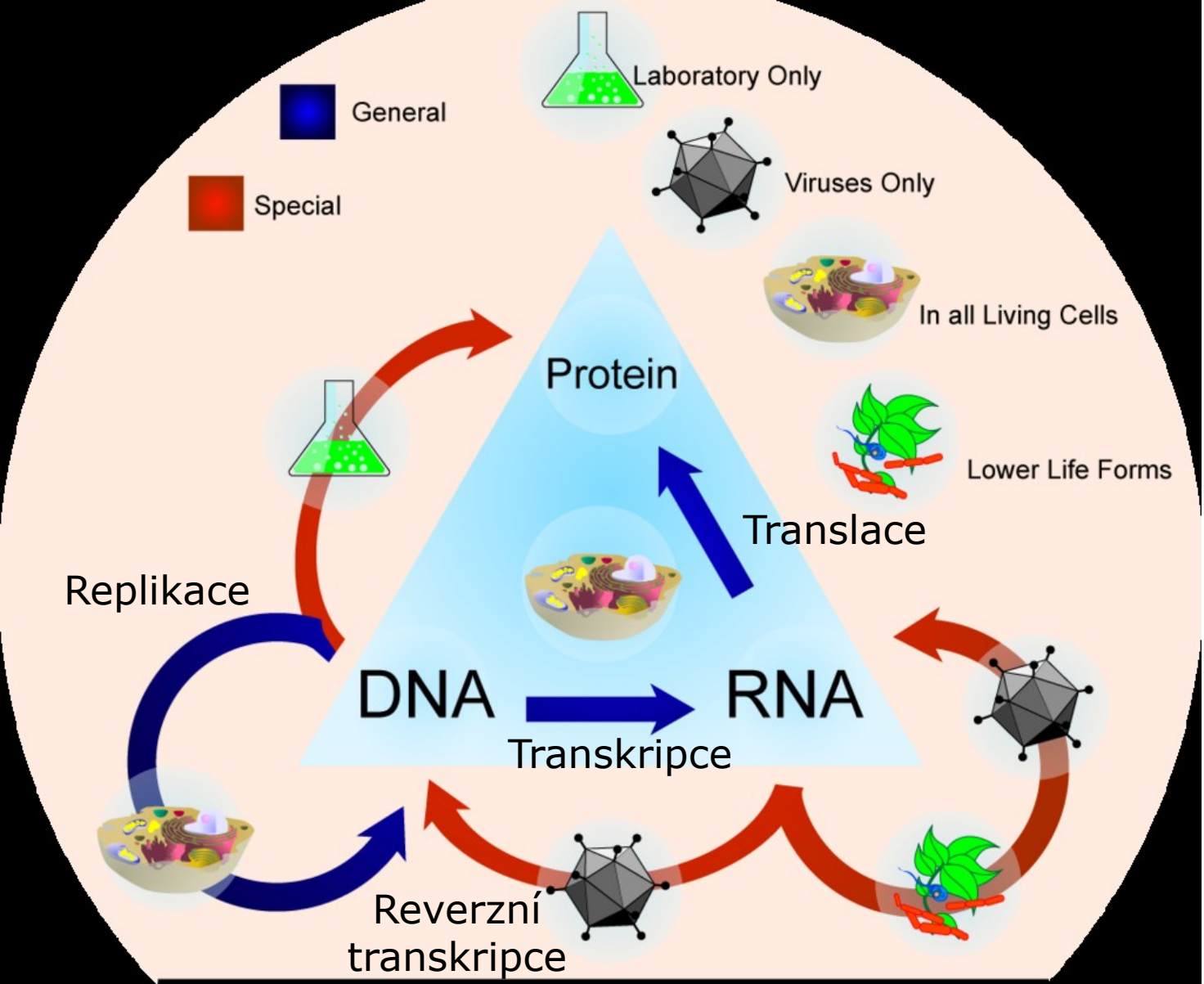
- Zpracovává se samostatně
- Možnost zpracovávat vlastní data nebo data z veřejně dostupných databází
- Výběr ze stanovených projektů, vlastní téma nutno schválit předem – **nejzazší termín výběru projektu: 8.3.2022**
- Projekt nutno odevzdat před zkouškou, pouze po odevzdání a obdržení **8** bodů z projektu je možné přihlásit se na zkušební termín
- Nejzazší termín odevzdání projektů pro kontrolu počtu bodů: **6 dní** před zkušebním termínem
- **1x** možnost opravy projektu dle připomínek

- **Student odevzdává 2 soubory:**
 - **popis projektu** ve formátu pdf ve struktuře definované níže
 - **.R soubor se skriptem analýzy** od načtení dat po finální grafy
- **Struktura popisu projektu:**
 - **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



**Biological Information
Flow**

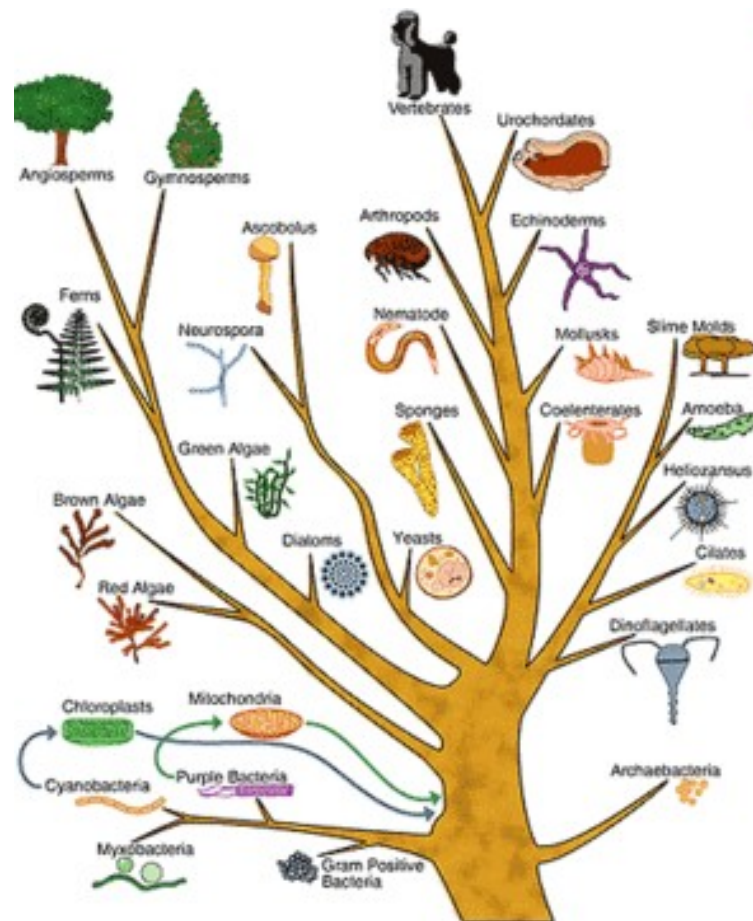
Centrální dogma molekulární biologie

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

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

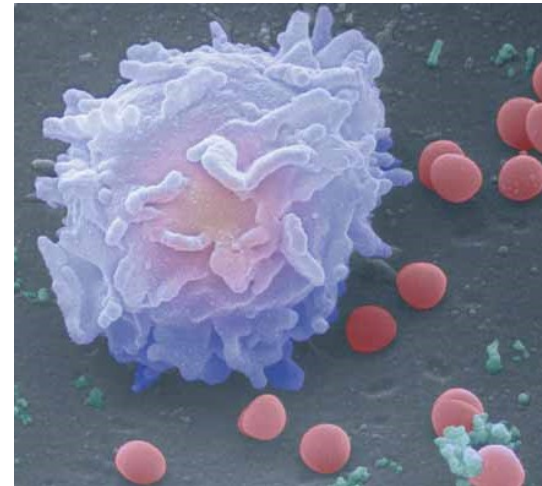
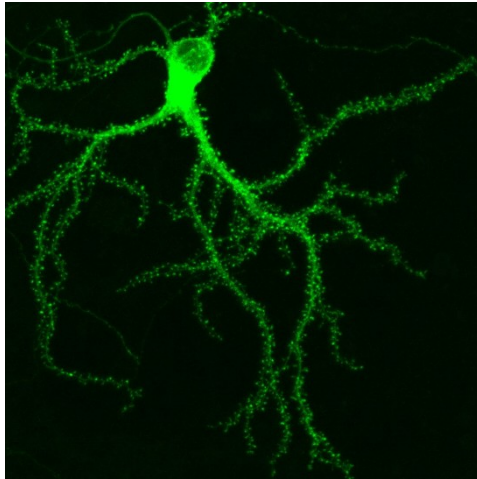
Geny podmiňují fyzický **vzhled** organismu a jeho **schopnost adaptace** na prostředí, ve kterém žije a jeho pomalé i náhlé změny (stres).

Rozdíly mezi organismy jsou podmíněné rozdíly v genomu.

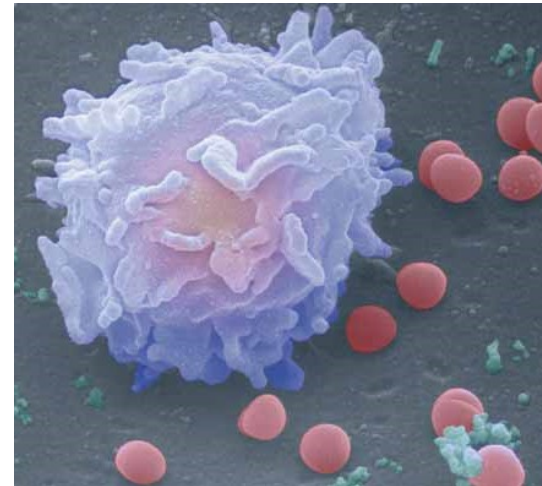
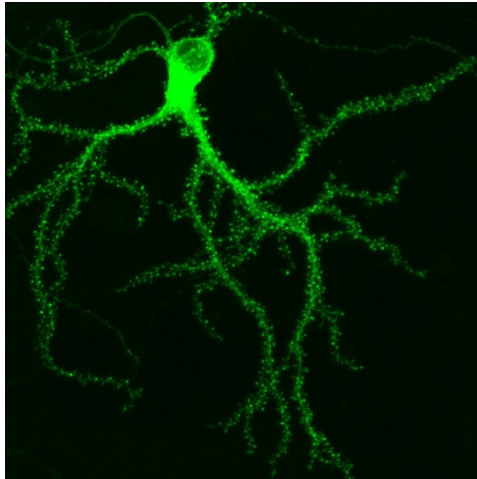


- Odolnost bakterií na antibiotika podmíněná mutacemi.
- Adaptace na extrémní podmínky - život ve vesmíru, v sopce, sirných pramenech, vařících pramenech a mrazech do -70

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**.

Dekódování genomu u různých druhů

Můžeme studovat



Rozdíly v genomu/proteomu
jednotlivých druhů



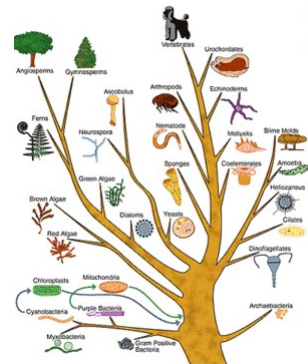
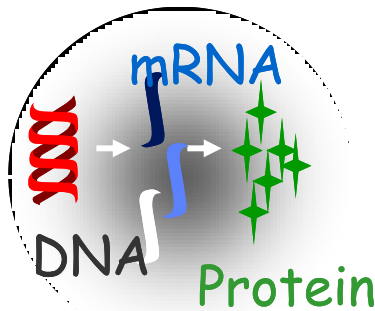
studovat tak evoluční propojení a
vytvářet fylogenetické stromy



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 ...



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**

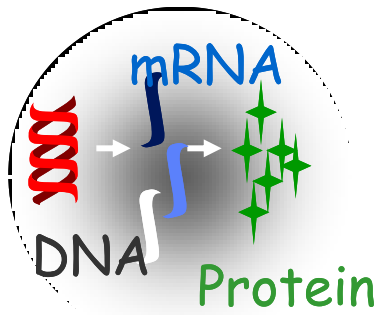
NEMOC ↔ GEN (Y)



Pochopit **podstatu** onemocnění



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



- 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, inserce, amplifikace nukleotidů)
- Aberace celého genu a nebo části chromozomu (delece, translokace, inserce, amplifikace)
- Aberace celých chromozomů

2. Změny v expresi a aktivitě genů a jejich produktů

3. Změny v posttranslačních úpravách proteinů

- 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 části 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

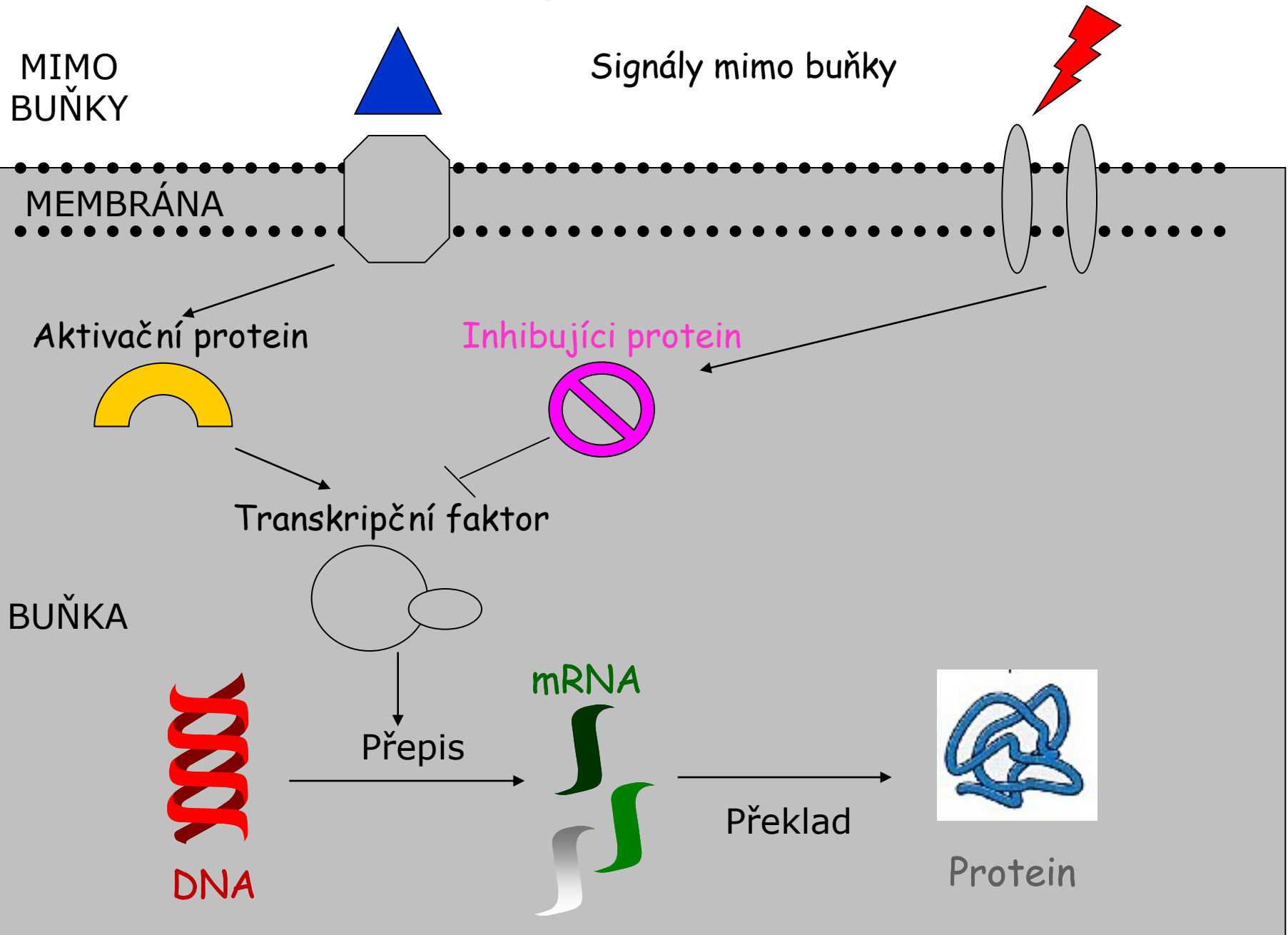
Geny a onemocnění III. – aktivita genů

- 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.

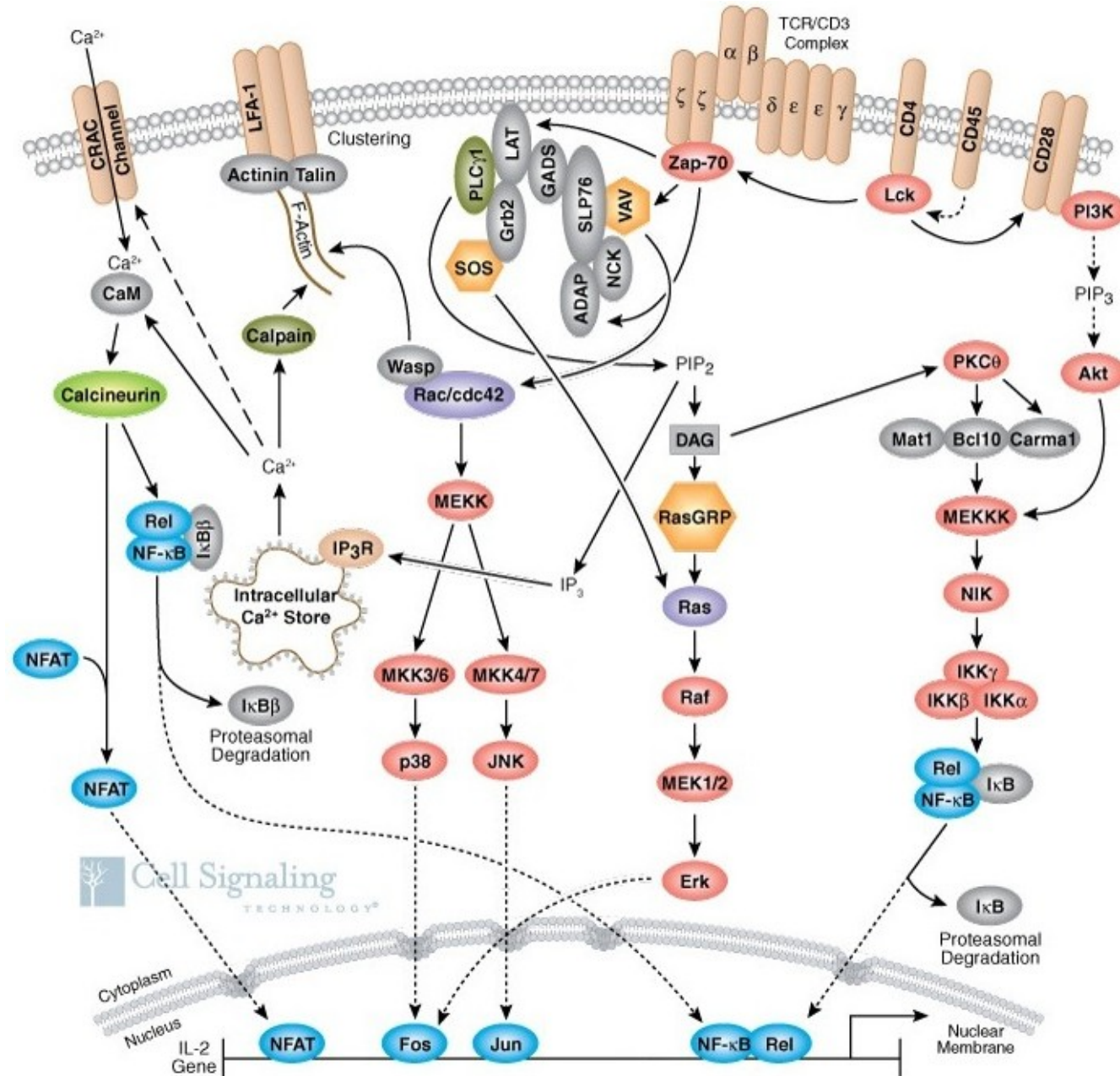
Geny a onemocnění IV. - shrnutí

- 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 produkován
 - v nedostačujícím množství
 - v nadměrném množství
 - **Změny v postranlačních úpravách** a sekundární/terciární struktury **proteinu**

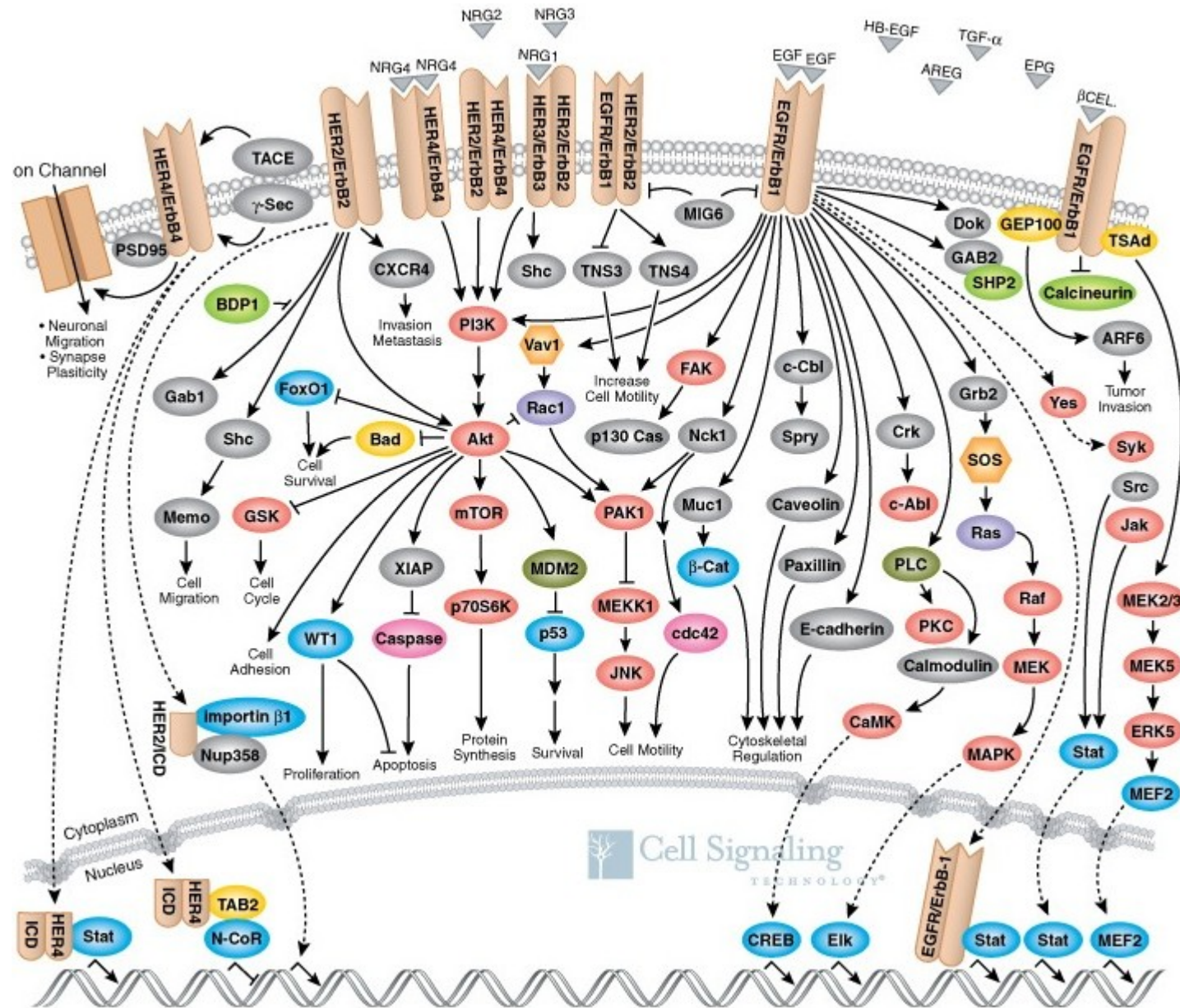
Co ještě víme



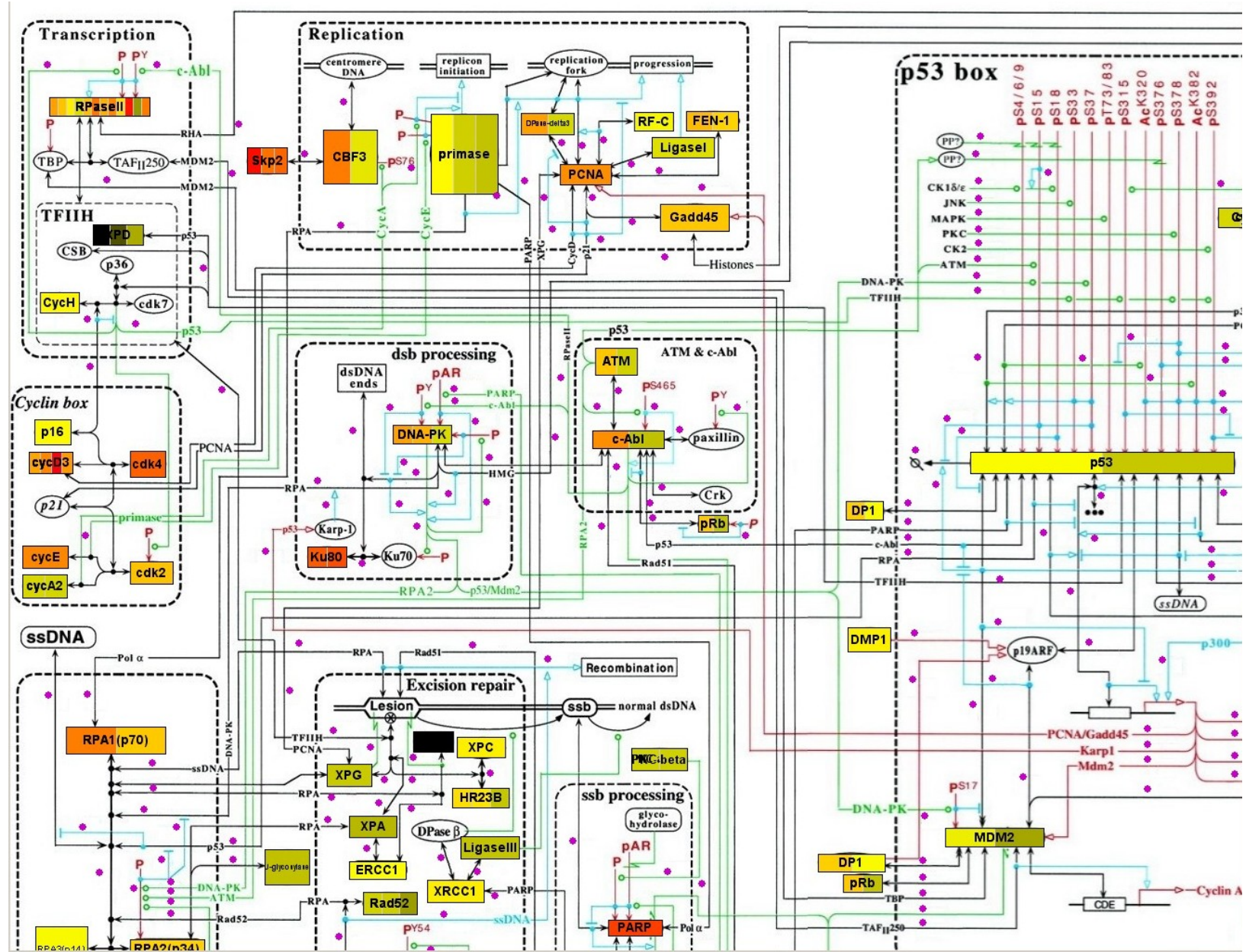
Ale víme ještě víc



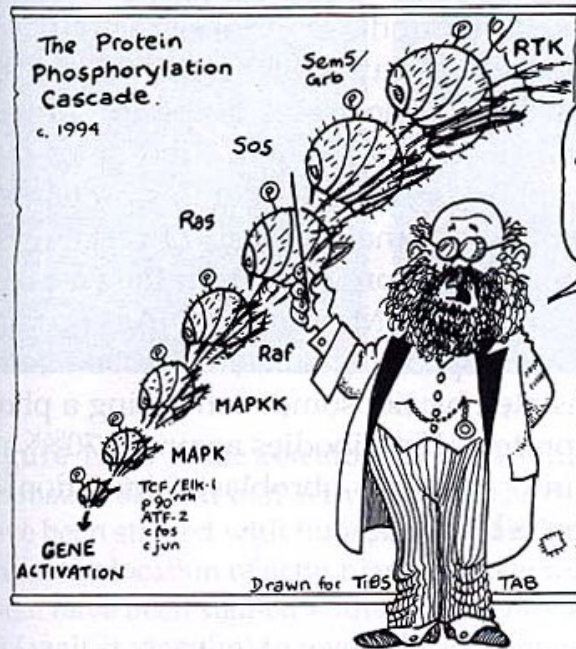
..a ještě víc...



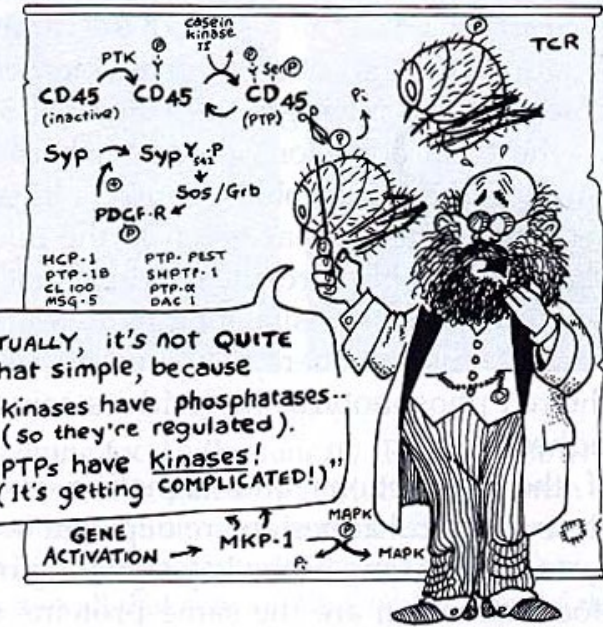
...a ještě víc...



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



OK, CLASS!
Pay attention!
It's quite simple!
"Kinases have kinases
upon their backs to bite 'em!
Kinase Kinases have kinases--
and so-- ad infinitum?!"



Er - ACTUALLY, it's not QUITE
that simple, because
"Some kinases have phosphatases--
(so they're regulated).
And PTPs have Kinases!
(It's getting COMPLICATED!)"



"And phosphotyrosines will bind
to SH-2 domains!
Whilst proline strings bind SH-3!
... and round we go again.
Some activated proteins shift
from cytosol to membrane,
Whilst some enter the nucleus--
(I've got a pain in my brain!)"



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

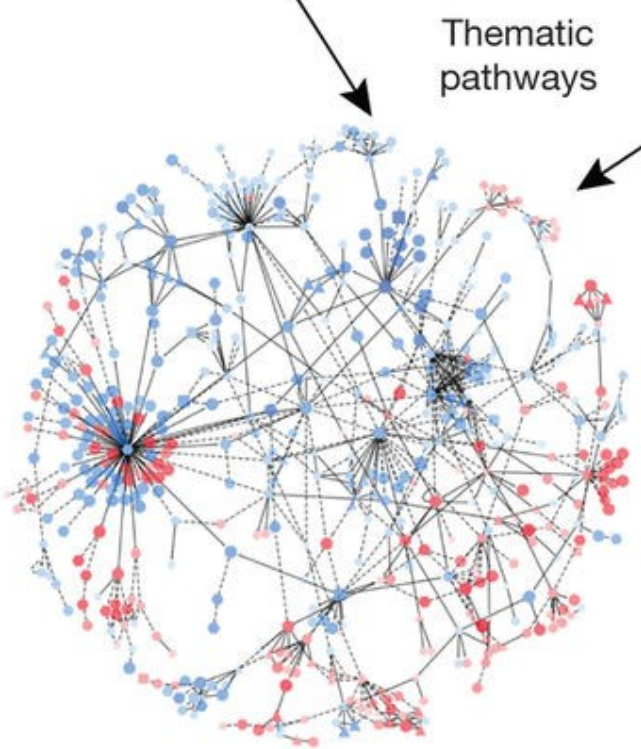
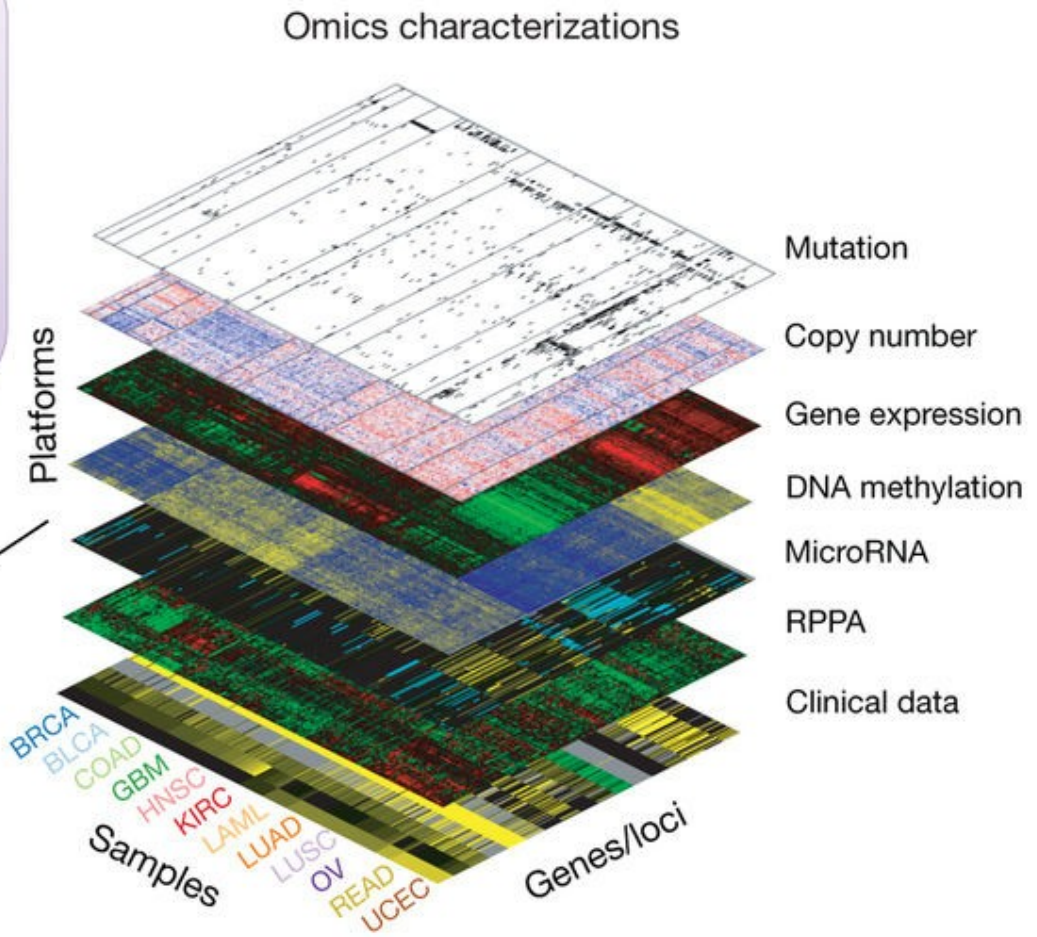
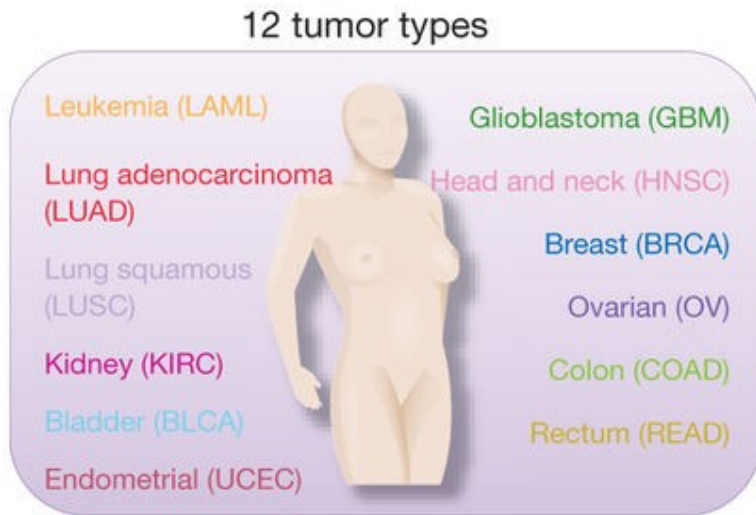
Další fáze je **modelování komplexních buněčných systémů** – proteinové interakce, buněčné dráhy, regulační a metabolické sítě ...

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?



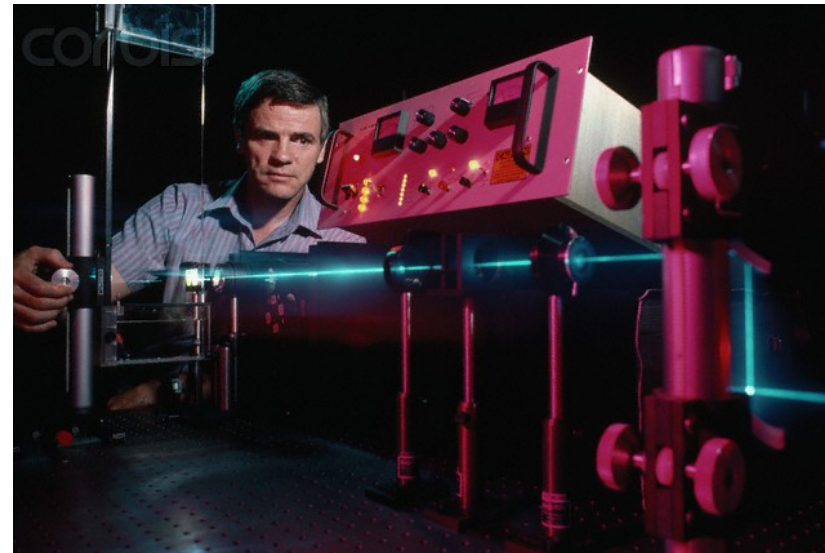
Proteomika a genomika



The Human Cancer Genome Atlas (TCGA) projekt

Od Watsona & Cricka po Leroya Hooda

- Na začátku byl dvoušroubovicový model DNA...



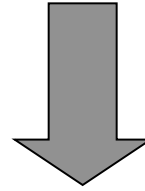
a na konci byly:

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

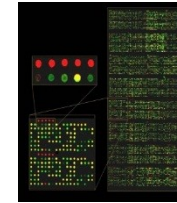
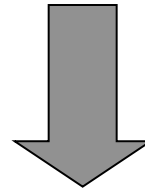
Nové možnosti



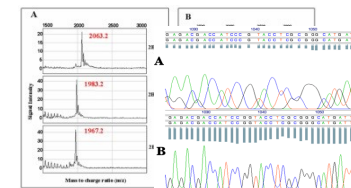
Sekvenátory umožnily rychle dekodovat sekvenci genů a proteinů



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

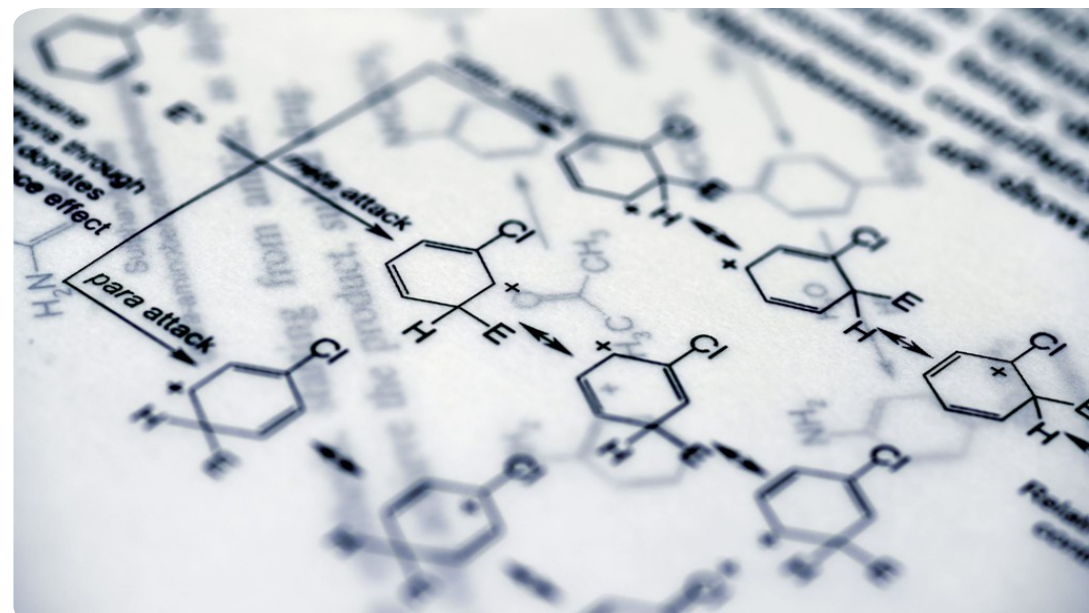


Otevřely se dveře pro nové, vysocepokryvní technologie, schopné analyzovat tisíce genů/proteinů v jednom experimentu!



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

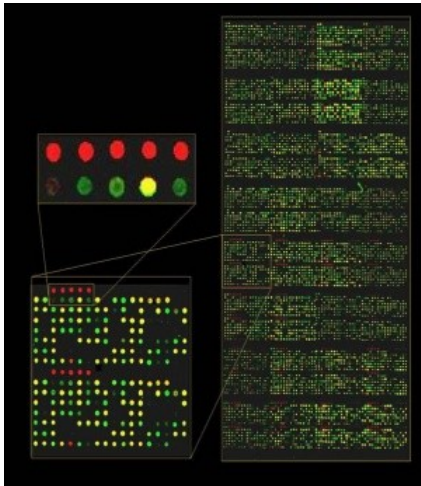
- Od hmotnostních spekter – přes komplexní struktury proteinových shluků - po analýzu funkce proteinů
 - Analýza **struktury**: Proteinová sekvenace
 - Analýza **exprese**: Hmotnostní spektrometrie, 2D gelová elektroforéza, Proteínové mikročipy...
 - Analýza **funkce**: Modelování makromolekulárních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



Data z omics experimentů

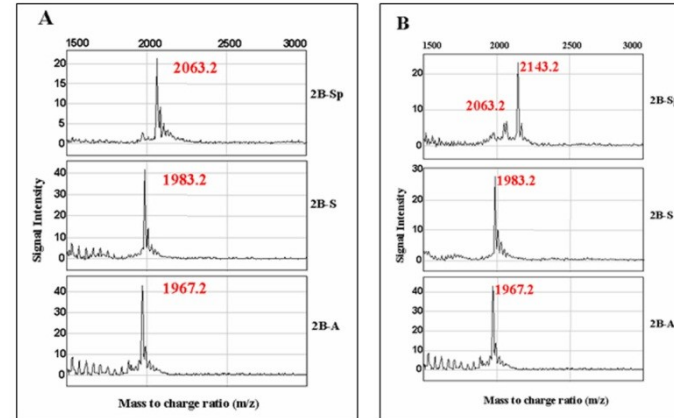
- Moderní vysoce pokryvné molekulární technologie produkují obrovské tabulky komplexních dat

Mikročipy



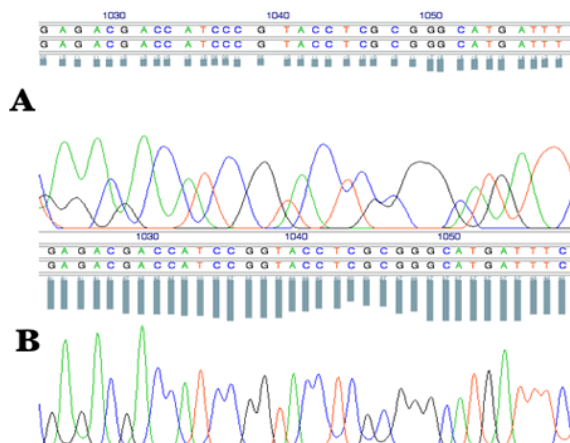
Desítky až tisíce genů
nebo transkriptů na
vzorek

Hmotnostní spektrometrie



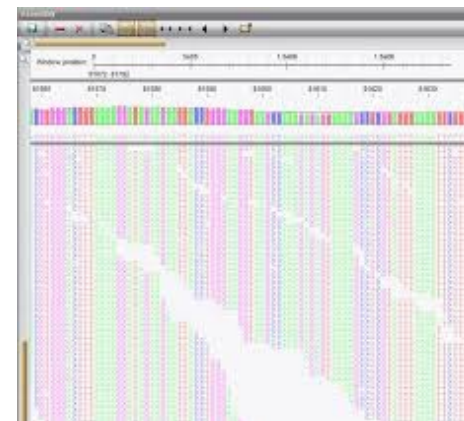
Tisíce spekter
proteinů,
metabolitů nebo
malých molekul na
vzorek

Sekvence DNA



Genom s biliony
nukleotidů na
vzorek

Sekvence 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

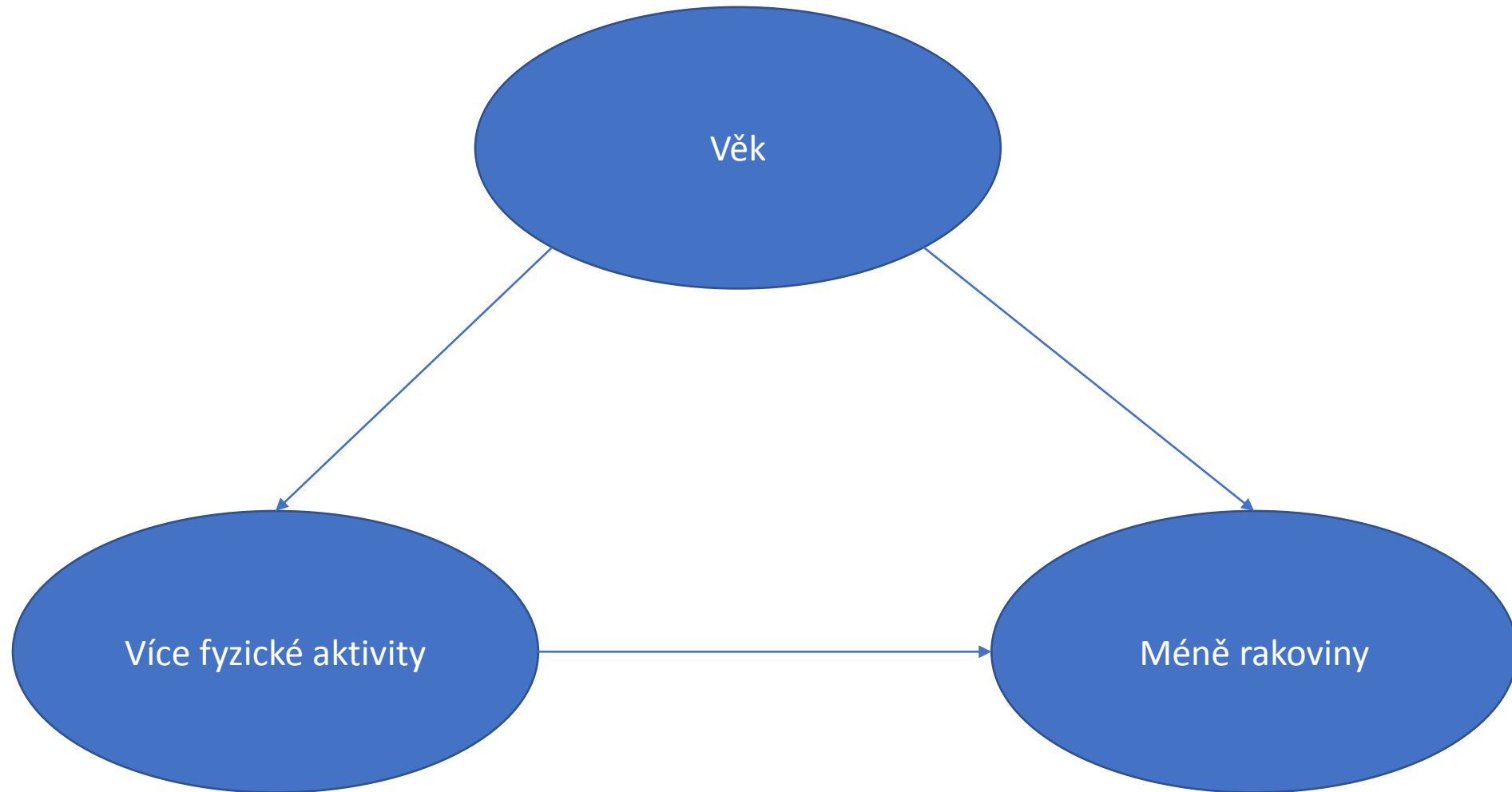
Za všechno mohou
matoucí vlivy
(confounding effects)?

Co je to matoucí faktor

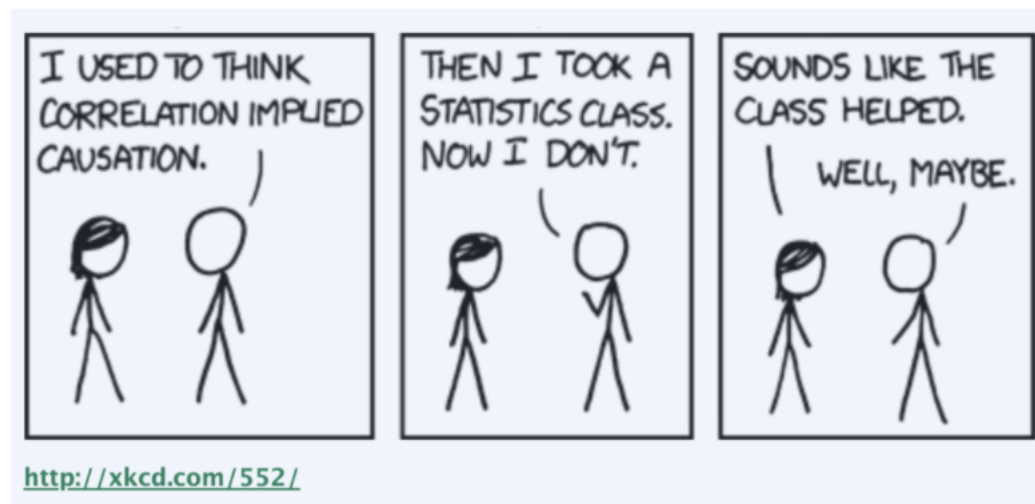
Matoucí faktor (*confounding factor*) je (neznámá) vnější proměnná, která ovlivňuje závislou proměnnou i nezávislou proměnnou v analýze, což způsobuje jejich falešnou asociaci a špatnou interpretaci.

Jiným způsobem, vzniká korelace, která není kauzalita....

Matoucí vliv



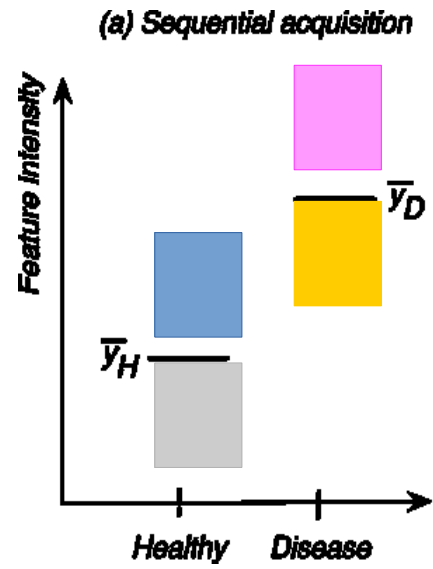
Pochybné korelace....



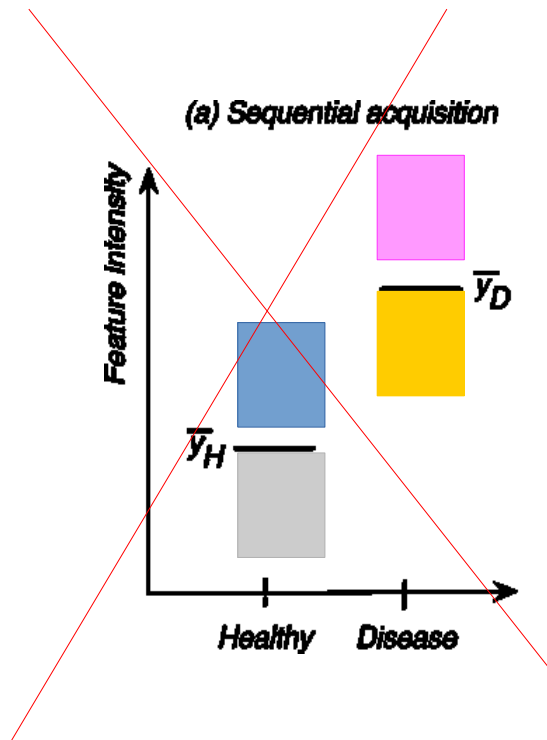
Efekt dávky

- 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



Efekt dávky

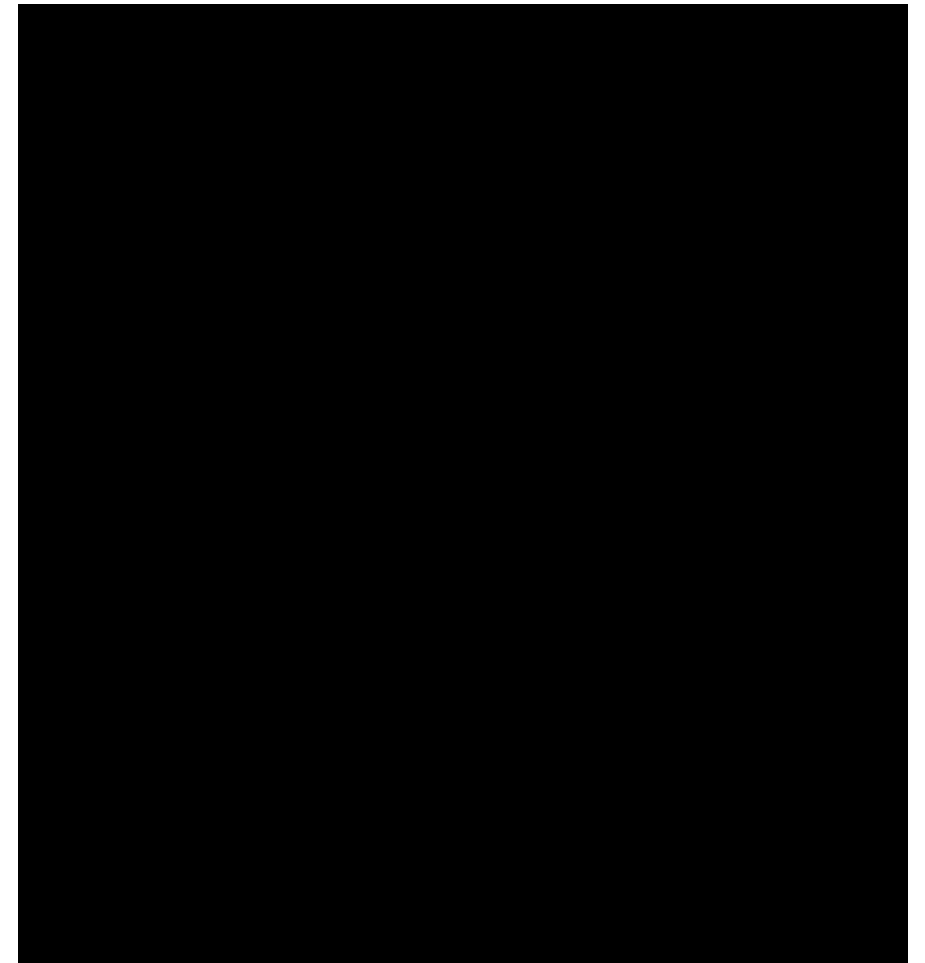


NENÍ MOŽNÉ STATISTICKY ODDĚLIT TECHNICKÝ EFEKT OD BIOLOGICKÉHO!!!

Lidé a myši na mikročipech

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

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 \log_2 mají tendenci přeceňovat rozdíl mezi dvěma nízkými hodnotami exprese, ale podceňují rozdíl mezi dvěma vysokými hodnotami exprese.

Lidé a myši na mikročipech

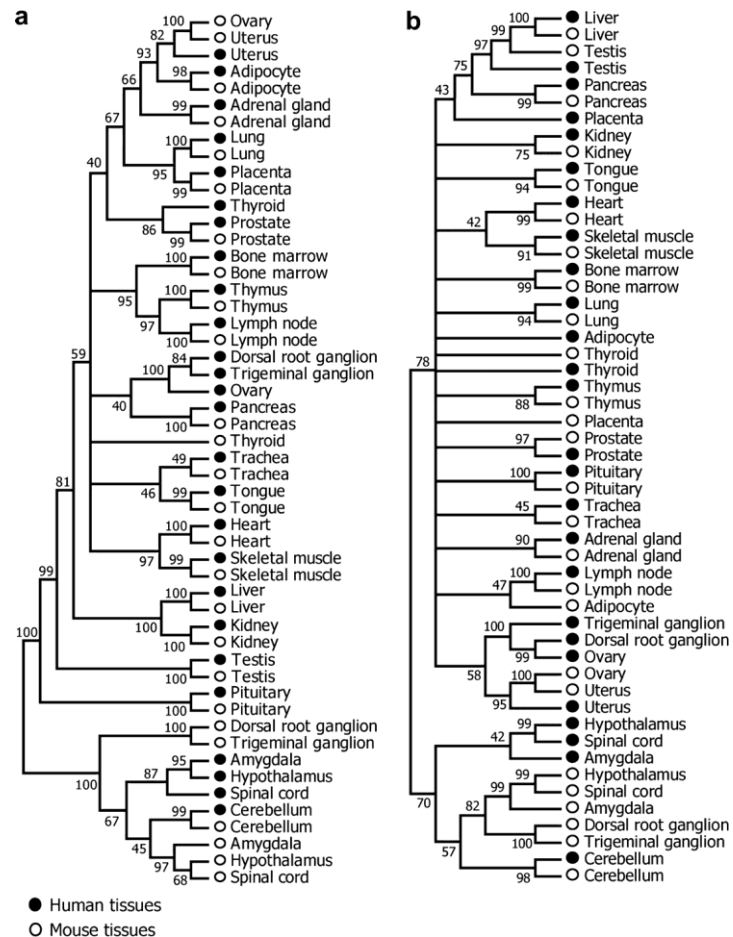


FIG. 5.—
Dendrograms of 26 human and 26 mouse tissues based on (a) 1 – Pearson's correlation coefficient r and (b) Euclidean distance d of tissues..

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

The 1000 genomes project


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

1000 genomes

Jaký je vliv data sekvencování na genetickou variabilitu mezi sekvencemi?

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**

Jeich 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**

Jeich 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**

Jeich 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**
- Analýza NGS dat – samostatný předmět E5444 (podzim)
- Analýza non-target MS dat - samostatný předmět E5020 (jaro)
- Detekce biomarkerů z omicsových experimentů