



Detekce biomarkerů z omics experimentů

- Mgr. Eva Budinská, PhD
- RECETOX
- budinska@recetox.muni.cz
- Podzim 2019

Cíl kurzu



...podrobně seznámit posluchače s **hlavními principy** analýzy dat z molekulárních 'omics experimentů (mikročipy, hmotnostní spektrometrie, NGS,...), **se zvláštním důrazem** na **plánování experimentů** a **validaci výsledků** při detekci **biomarkerů**.... a z nich odvozených **modelů**.

Co je to
biomarker?

Biologický marker (biomarker):

Charakteristika, která je objektivně měřena a hodnocena jako indikátor normálních biologických procesů, patogenních procesů nebo farmakologických odpovědí na terapeutický zásah.

Biomarkerem
může být



Molekula a její stav
(mutace DNA,
hodnota exprese
miRNA, zvýšená
hladina proteinu...)



Aktivita buněk v
konkrétních
oblastech (lymfocyty
v invazivním frontu
nádoru)



**Přítomnost
mikroorganismu**



Proces (zvýšená
proliferace,
přítomnost stromální
reakce v nádoru, ...)

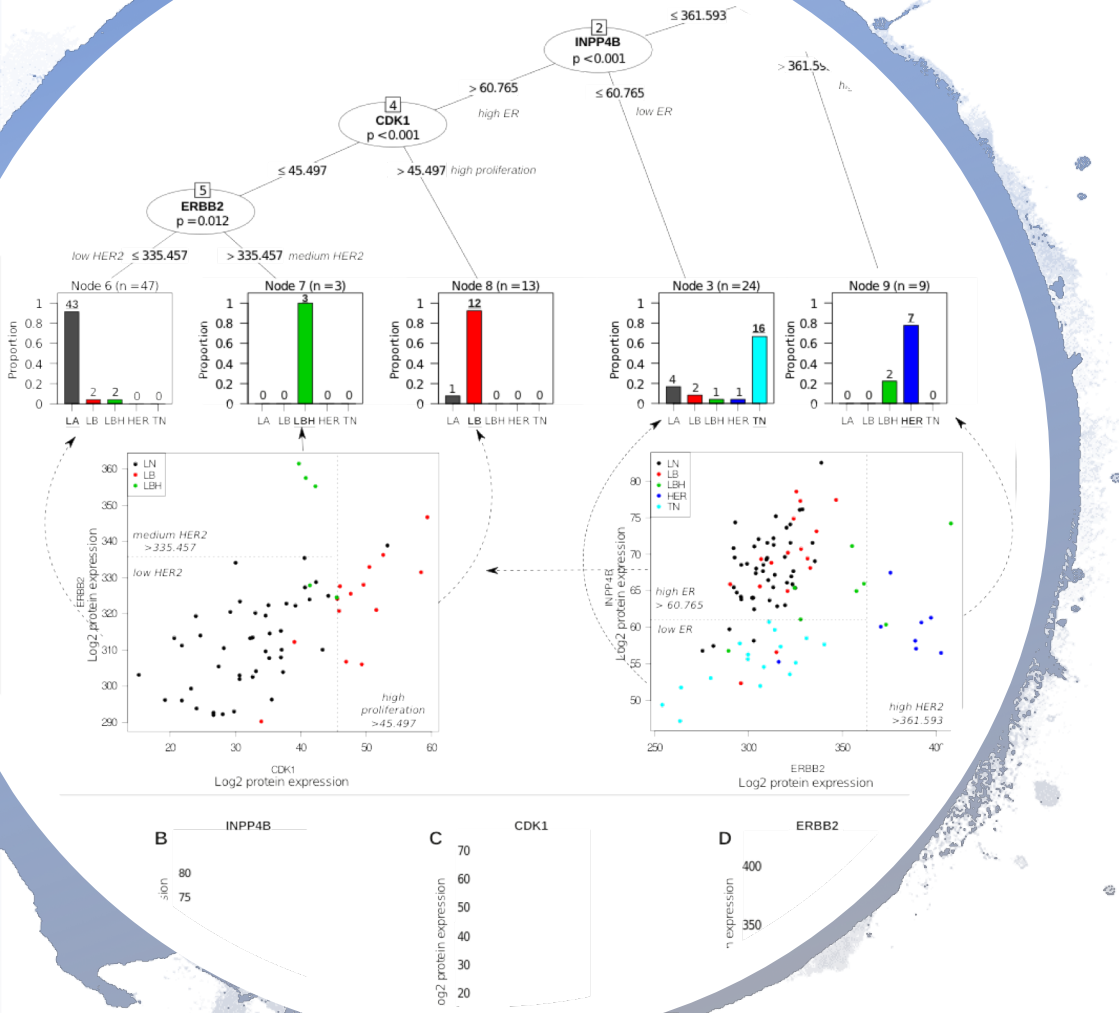


....

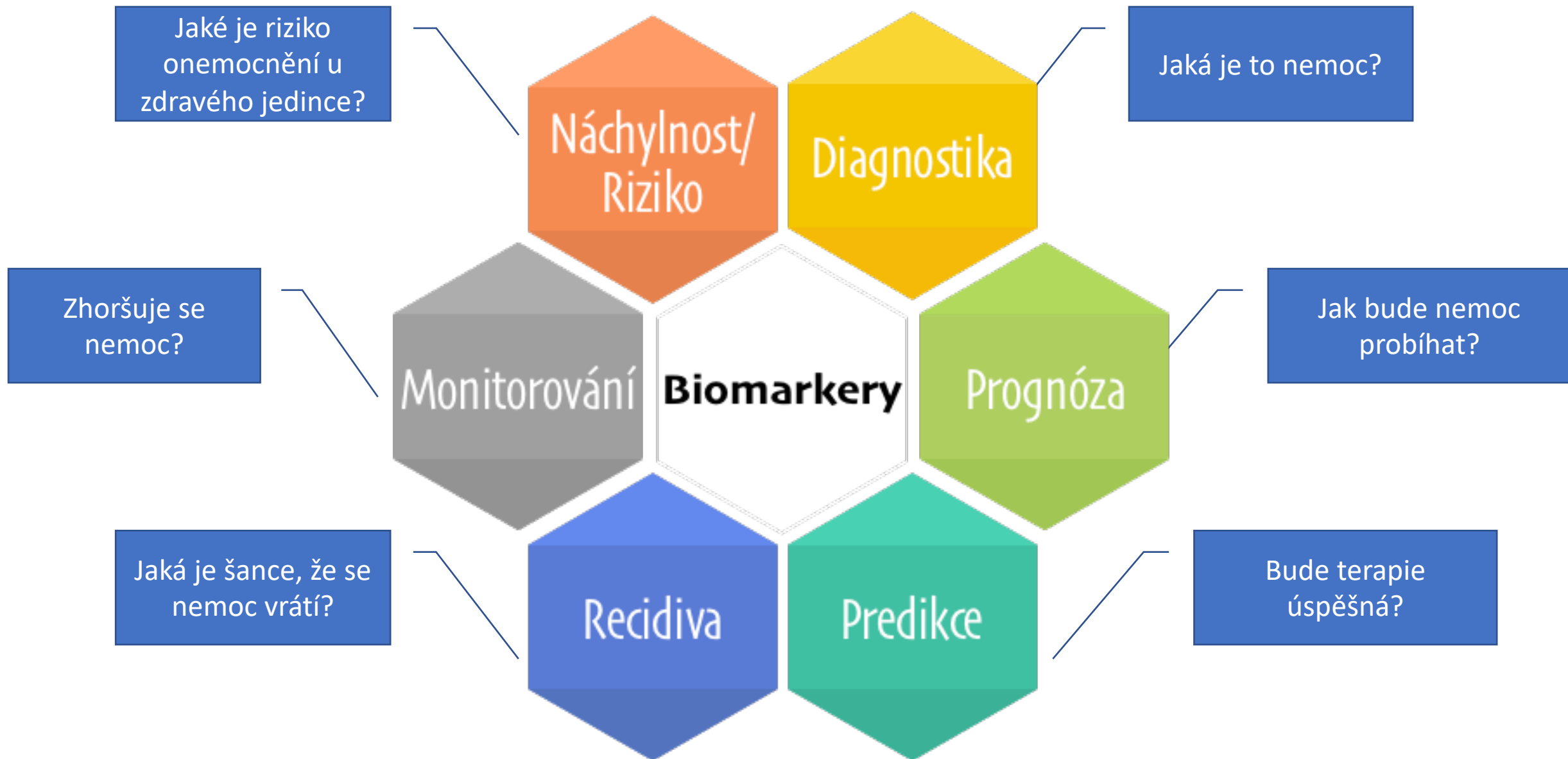


**Využití jednotlivých
biomarkerů v
rozhodovacím
PRAVIDLE
(modelu/testu)**

Biomarkery a modely



- Biomarker může být založen na **jediném analytu**, nebo na **jejich kombinaci v modelu** (klasifikátoru)
- Je to právě **kombinace více analytů** (genů, proteinů, metabolitů...), která je typická pro biomarkery z omicsových dat

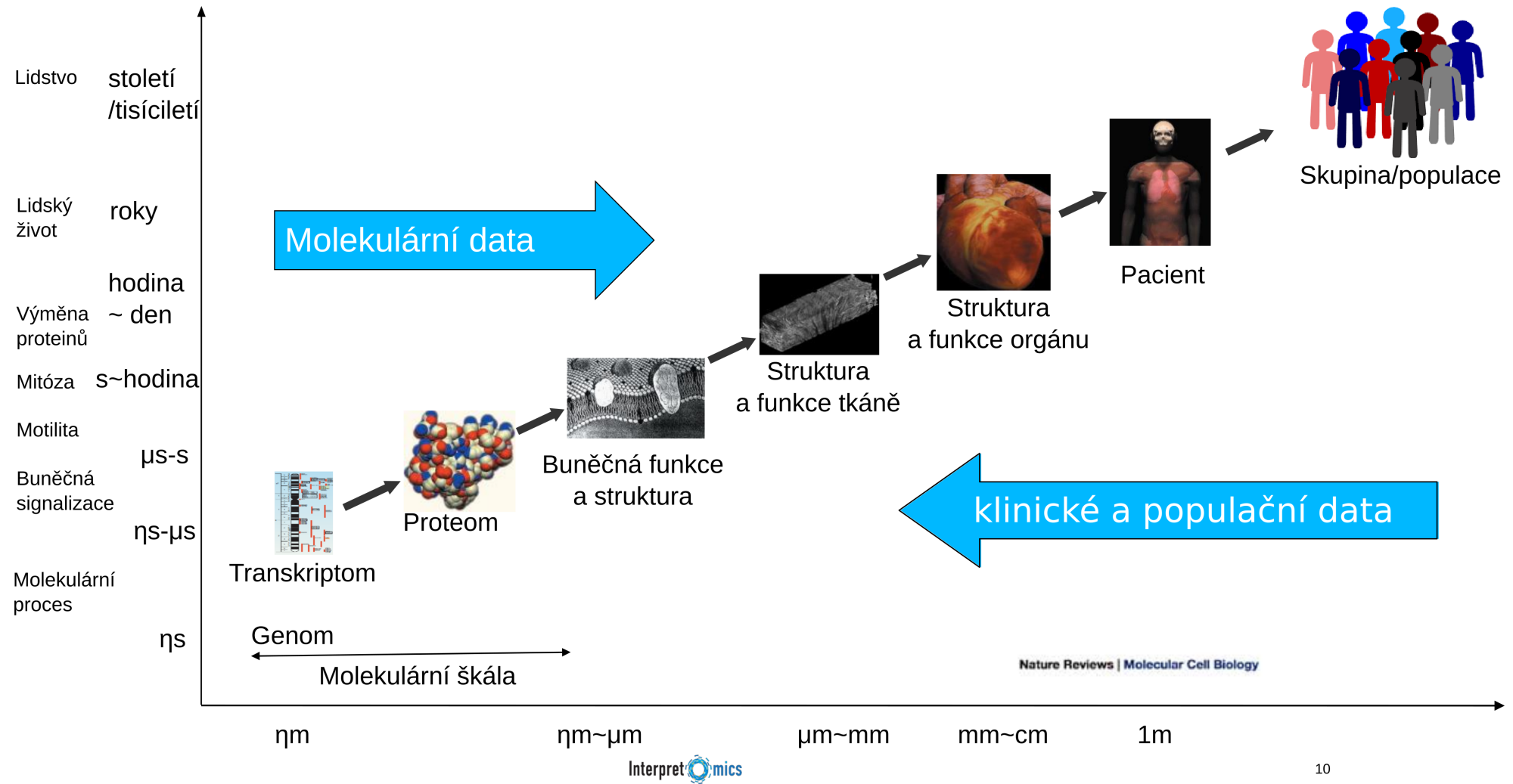




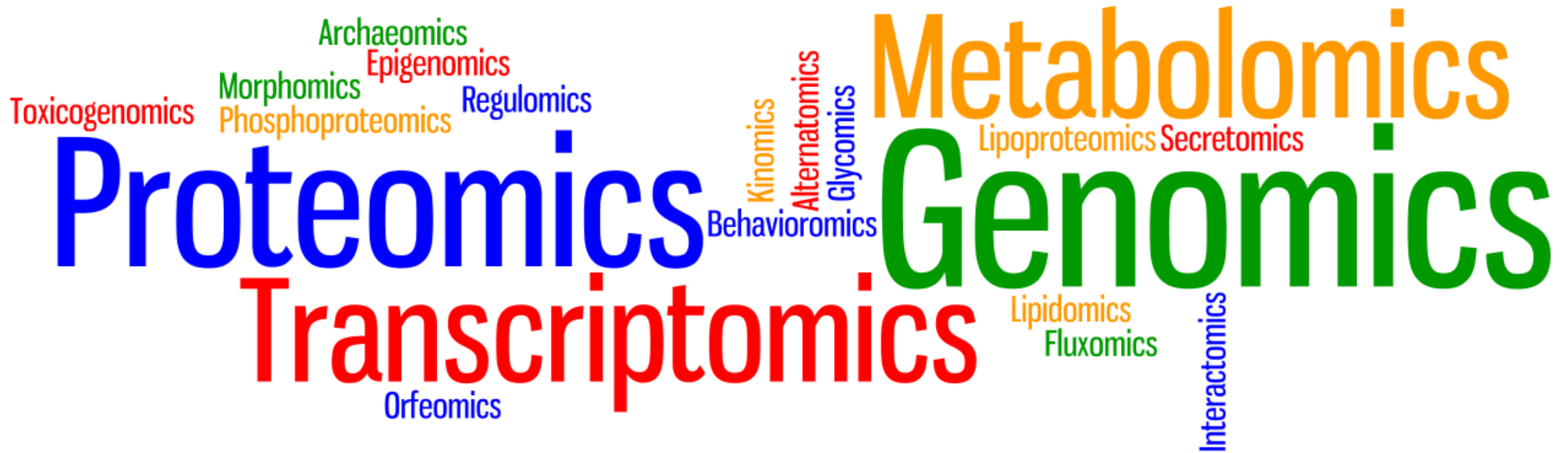
Co musí biomarker (nebo model) splňovat

Musí být použitelný rutinně v praxi:

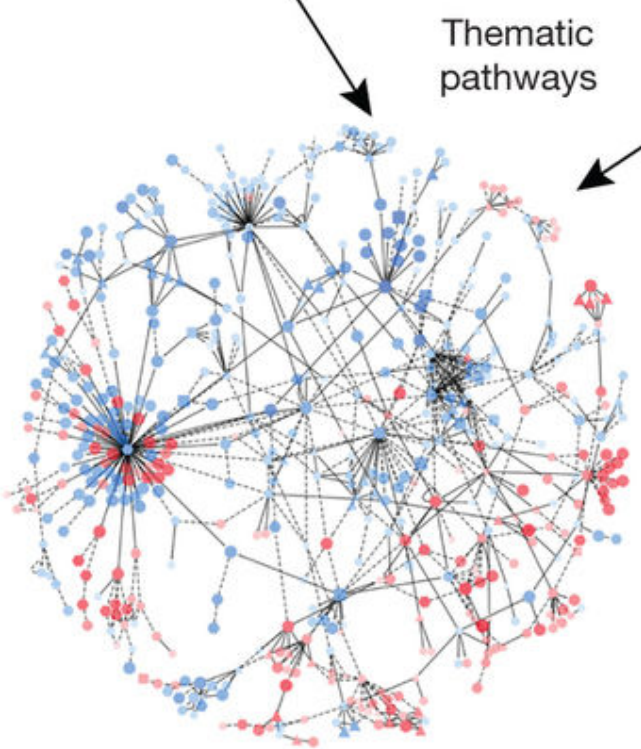
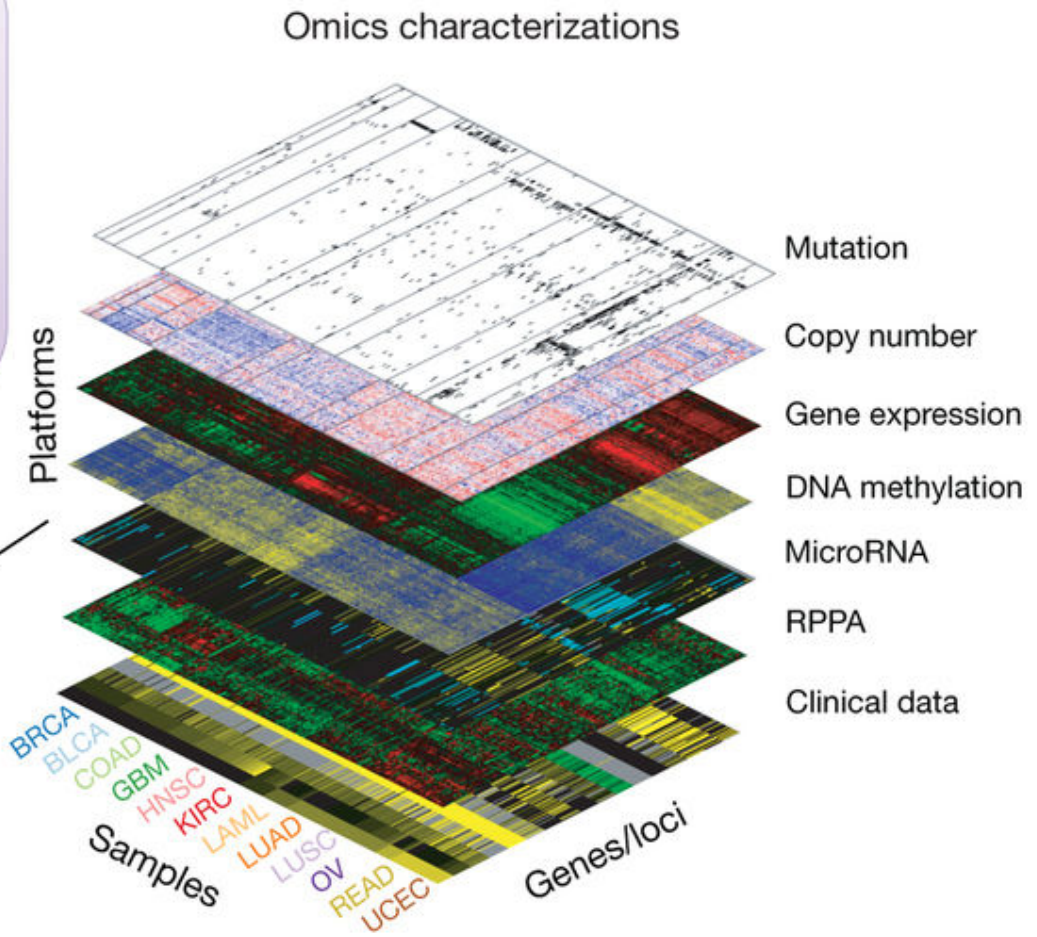
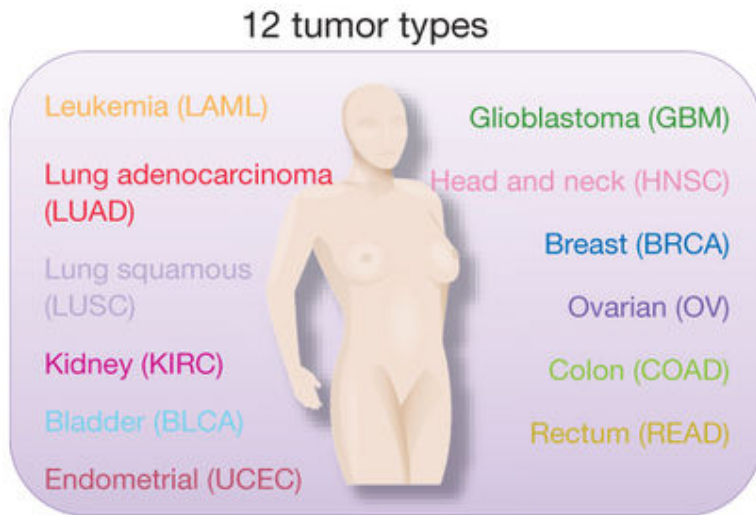
- **přesný** (dostatečně citlivý a dostatečně specifický)
- **robustní** (co nejméně omezen technologií měření)
- **reproducibilní** (obecně platný)



Mnohorozměrná povaha moderní biomedicíny



All the omics?

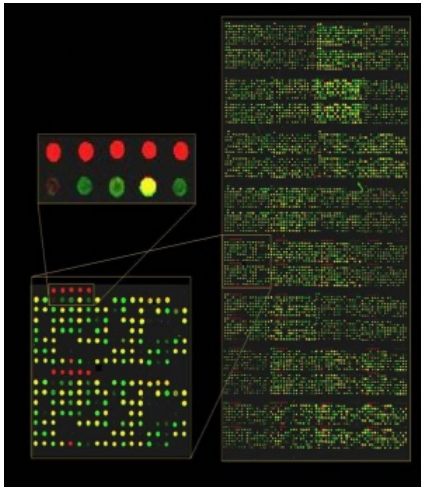


The Human Cancer Genome Atlas (TCGA) projekt

Data z omics experimentů

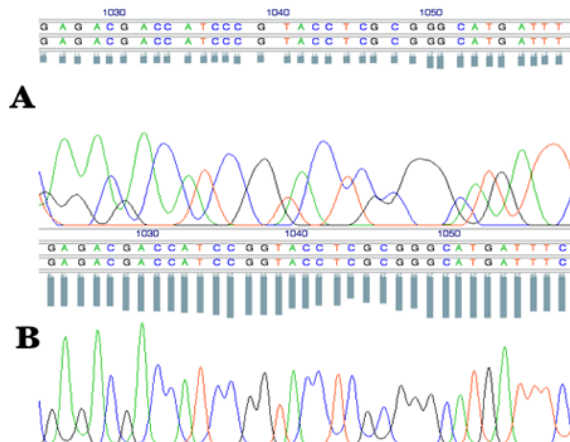
- Moderní vysocepokryvné molekulární technologie produkují obrovské tabulky komplexních dat

Mikročipy



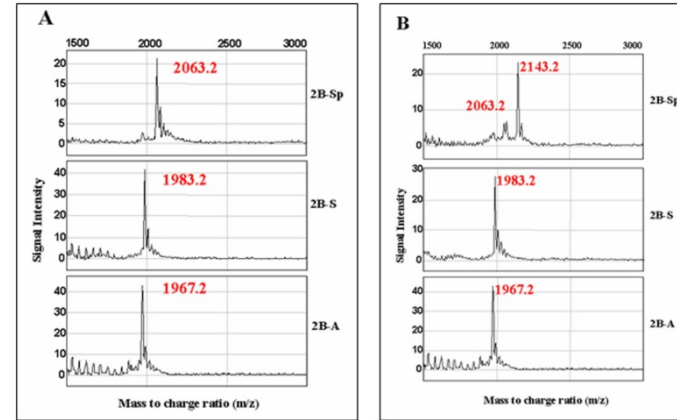
- ☑ Desítky až tisíce genů nebo transkriptů na vzorek

Sekvence DNA



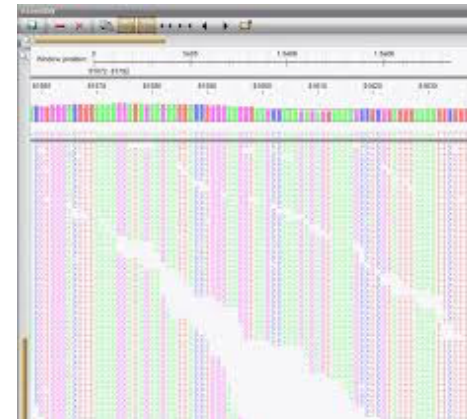
- ☑ Genom s biliony nukleotidů na vzorek

Hmotnostní spektrometrie



- ☑ Tisíce spekter proteinů, metabolitů nebo malých molekul na vzorek

Sekvence nové generace



- ☑ Miliony krátkých čtení DNA na vzorek



Data omics experimentů

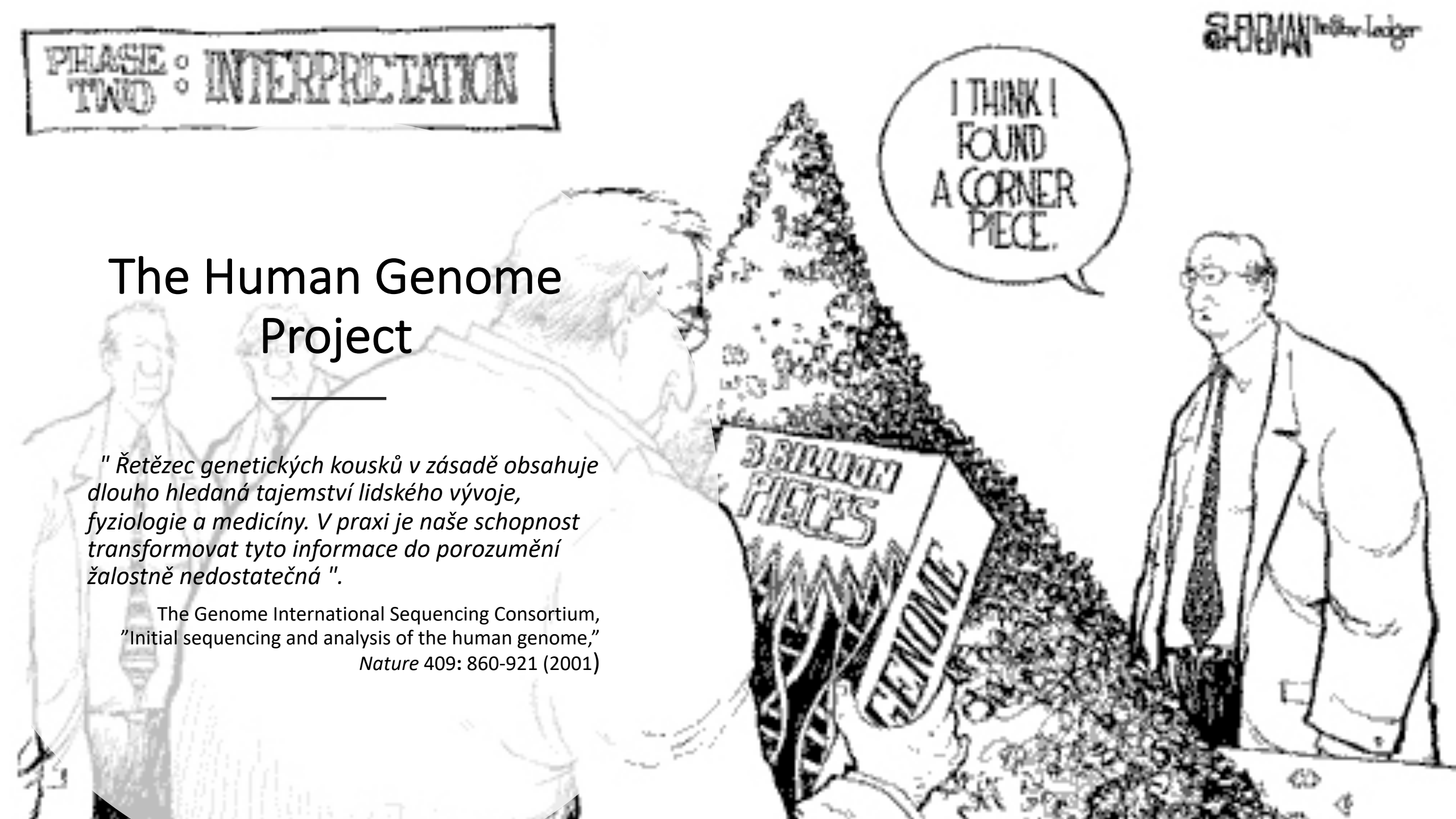
PHASE TWO: INTERPRETATION

SEIDMAN The Star Ledger

The Human Genome Project

"Řetězec genetických kousků v zásadě obsahuje dlouho hledaná tajemství lidského vývoje, fyziologie a medicíny. V praxi je naše schopnost transformovat tyto informace do porozumění žalostně nedostatečná".

The Genome International Sequencing Consortium,
"Initial sequencing and analysis of the human genome,"
Nature 409: 860-921 (2001)





Hledání jehly v kupě
sena?

Jak se hledá potenciální biomarker v omics datech

Biologická otázka (hypotéza)

Dizajn experimentu

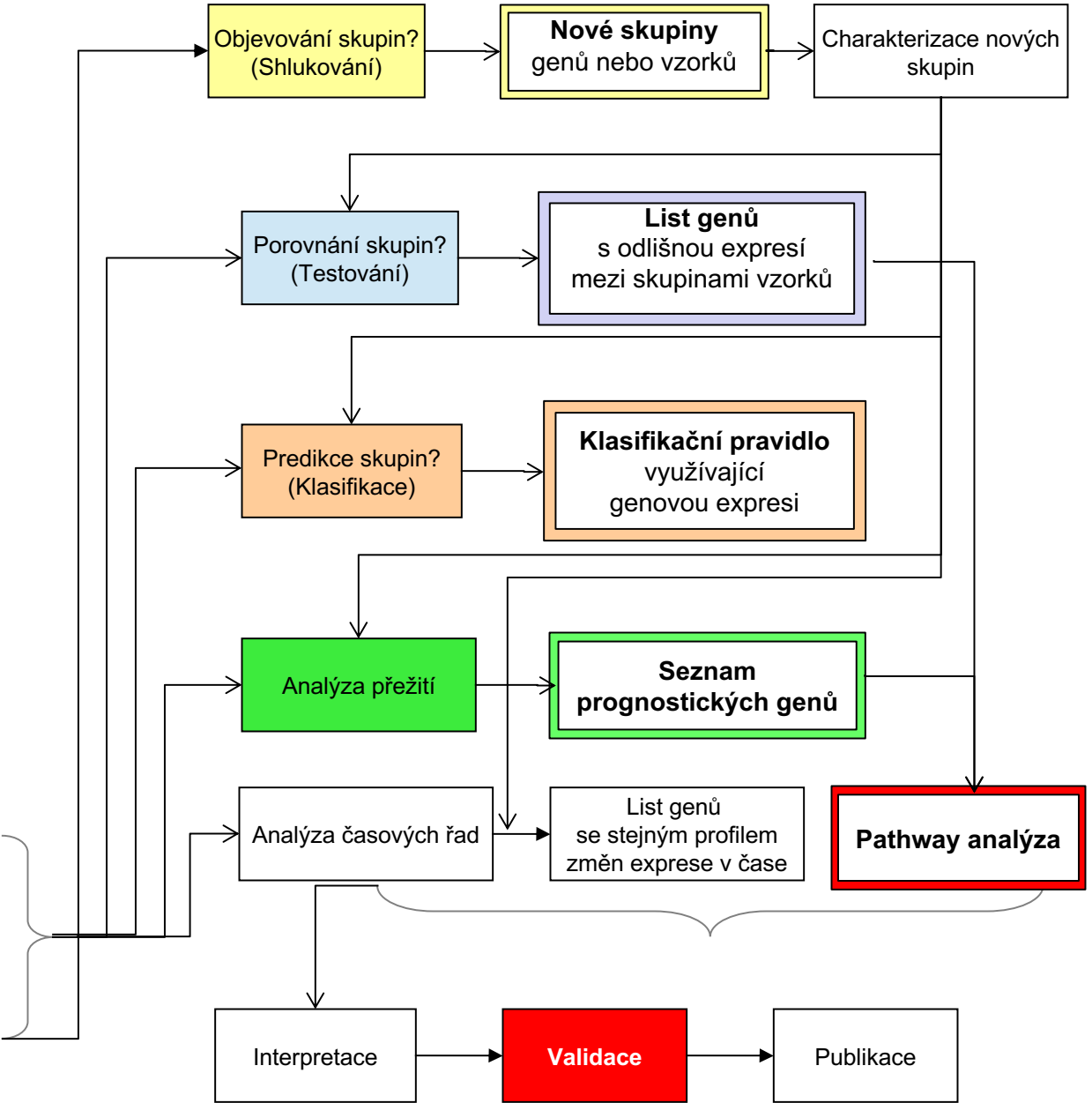
Provedení experimentu (hybridizace mikročipů, hmotnostní spektrometrie...)

N matic základních dat (jedna pro každý z N vzorků)

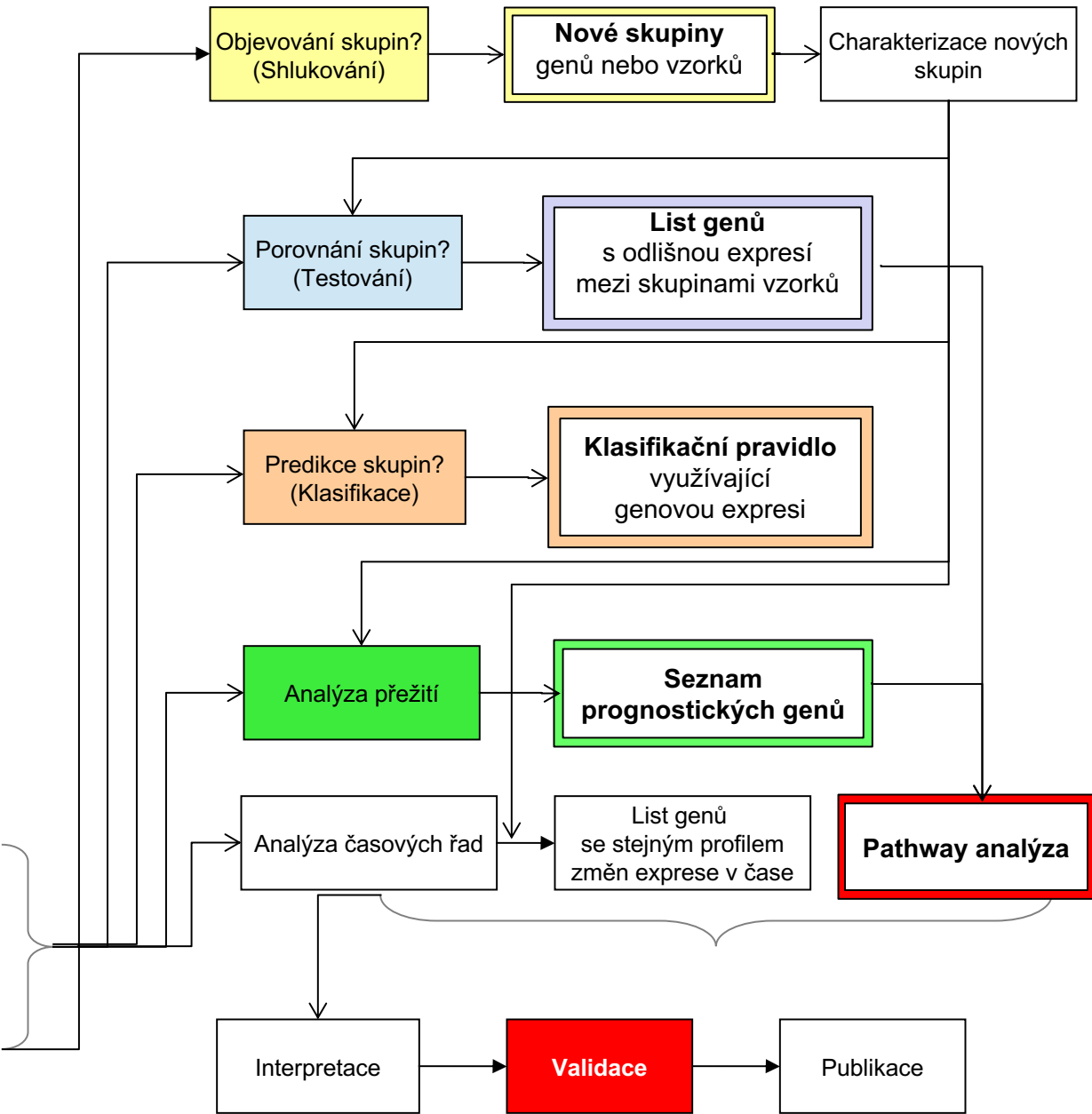
Kontrola kvality
Normalizace
Sumarizace

Finální datová matice N vzorků a K genů (proteinů)

Matice informací o vzorcích N x P (např. klinická data v medicíně)



Jak se hledá potenciální biomarker v omics datech



“Přijít za statistikem po dokončení experimentu je často to samé jako požádat ho aby provedl posmrtné vyšetření. Možná bude schopen říct, na co experiment zemřel.”

(Ronald Fisher)



Proč jsou
omicsová data
problematická?



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

Nerepresentují skutečné hodnoty (koncentrace, počty) sledovaných molekul

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

Jejich předzpracování 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



... ale i analýza těchto dat
a vytváření omics
biomarkerových modelů
má svá specifika!





Skandál na Duke university

Severní Karolína, USA





2006 – Anil Potti, nadějný vědec z Duke University publikuje v Nature Medicine s kolegy článek o biomarkerech rezistence na chemoterapeutika v onkologii.



Genomické signatury byly odvozeny z analýzy exprese (mikročipy) senzitivních a rezistentních buněčných linií, výsledky validovány na pacientech.



Obrovský ohlas, v roce 2006 článek zařazen mezi “The Top 6 Genetic Stories of 2006”

Genomic signatures to guide the use of chemotherapeutics

Anil Potti^{1,2}, Holly K Dressman^{1,3}, Andrea Bild^{1,3}, Richard F Riedel^{1,2}, Gina Chan⁴, Robyn Sayer⁴,
Janiel Cragun⁴, Hope Cottrill⁴, Michael J Kelley², Rebecca Petersen⁵, David Harpole⁵, Jeffrey Marks⁵,
Andrew Berchuck^{1,6}, Geoffrey S Ginsburg^{1,2}, Phillip Febbo¹⁻³, Johnathan Lancaster⁴ &
Joseph R Nevins¹⁻³

Using *in vitro* drug sensitivity data coupled with Affymetrix microarray data, we developed gene expression signatures that predict sensitivity to individual chemotherapeutic drugs. Each signature was validated with response data from an independent set of cell



2006 – Biostatistici K. Coombes, J. Wang and K.A. Baggerly se snaží o aplikaci signatur na data výzkumníků z jejich univerzity, ovšem bez úspěchu.



Aktivně konzultují s autory článku.



Čím více se noří do dat, tím více mají pochybností o validitě závěrů a správnosti samotných dat!

2007 – Coombes a kol. publikují v Nature Medicine dopis zpochybňující Pottiho výzkum

(Coombes, Wang, Baggerly. Microarrays: retracing steps, Nature Medicine, 2007)



Reportují tyto chyby:

označení senzitivních a rezistentních buněčných linií nesedí!

tabulka se seznamem významných genů a jejich sond obsahuje systematickou chybu (posun o políčko) – geny nesedí se sondami, po korekci tabulky se podařilo reprodukovat pouze 3 ze 7 seznamů a výsledků senzitivity

Model rezistence na doxacel – podařilo se zreprodukovat pouze 31 z 50 genů publikovaných v článku, ostatních 19 bylo zřejmě přidáno ručně “aby byla validace úspěšná”

Autorský SW (algoritmus), který Potti používá, pracuje s validačními a testovacími daty společně. Po korekci této chyby jsou výsledky validace klasifikátorů špatné – na validačních datech téměř rovné náhodě.

Mezitím vycházejí další články:

Blood (2006), NEJM (2006), JCO (2007), Lancet Oncology (2007), JAMA (2008), PLOS (2008), PNAS (2008), Clin Can Res (2009)

V roce 2009 již **212 citací**, několik klinických studií, stovky **léčených pacientů**

V roce 2010 – Anil Potti obviněn z falzifikace výsledků a vyšetřován

Trvá 4 roky a mnoho úsilí, než jsou chyby uznány a články staženy!



ANIL POTTI

CASE PROGRESSION

**JULY
2010**

Potti is accused of falsifying information on his resume, and Duke launches an investigation into his work

**NOV
2010**

Potti resigns

**OCT
2011**

Patients in Potti's clinical trials file a lawsuit against the University

**JAN 29 THUR
2015**

**LAWSUIT SET
TO START**

Jak skandál změnil svět omics výzkumu

Červenec 2010 – ředitel National Cancer Institute (NCI) Harold Varmus obdržel **dopis** od více než **30 statistiků** a bioinformatiků, ve kterém vyjádřili své obavy nad použitím několika testů založených na genové expresi, které se používali v již probíhajících klinických studiích na Duke University k predikci odpovědi na chemoterapii.

V důsledku vznikla komise Institutu medicíny (IOM), cílem které bylo sepsání doporučení pro vývoj testů z omicsových studií



Evolution of Translational Omics: Lessons Learned and the Path Forward

ISBN
978-0-309-22418-5

300 pages
6 x 9
PAPERBACK (2012)

Christine M. Micheel, Sharly J. Nass, and Gilbert S. Omenn, Editors;
Committee on the Review of Omics-Based Tests for Predicting Patient
Outcomes in Clinical Trials; Board on Health Care Services; Board on
Health Sciences Policy; Institute of Medicine

IOM (Institute of Medicine). 2012. *Evolution of Translational Omics: Lessons Learned and the Path Forward*. Washington, DC: The National Academies Press.

IOM komise: Specifika testů založených na omics



Testy na bázi omics a ve skutečnosti všechny klinické laboratorní testy podléhají odlišnému regulačnímu rámci než léky



Absence jasného biologického zdůvodnění na rozdíl od většiny ostatních klinických laboratorních testů založených na jediném analytu



Složitost omicsového výzkumu ztěžuje sdílení komplexních datových souborů a výpočetních modelů, což omezuje schopnost ostatních vědců replikovat a ověřovat zjištění a závěry těchto studií

Absence jasného
biologického
odůvodnění testů
omics biomarkerů

Biologické zdůvodnění **testu s jedním analytem** je často zcela zřejmé: Test je užitečný, protože gen, RNA, protein nebo metabolit hraje **pochoitelnou roli** v patologii onemocnění nebo jiném vyšetřovaném biologickém procesu.

Příklady:

Testování karcinomu prsu lidským
epidermálním růstovým faktorem 2
(HER2)

Měření hladiny cholesterolu
lipoproteinů s nízkou hustotou (LDL)
pro hodnocení srdečního rizika

Absence
jasného
biologického
odůvodnění
testů omics
biomarkerů –
proč je to
problém

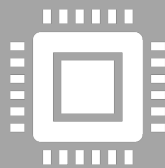
Když se nedá test založený na omics biomarkerech biologicky odůvodnit, je o to důležitější ho správně VYTVOŘIT a poté správně VALIDOVAT, aby byla zajištěna vědecká spolehlivost!

Z důvodů vyššího rizika „přetrénování“ těchto testů je potřeba přísných kritérií, validace a odpovědnosti ještě vyšší než u samostatných testů založených na biomarkerech.

Problém (ne) sdílení komplexních datových souborů a výpočetních modelů



K dispozici jsou databázové úložiště pro soubory omicových dat, ale sdílení dat není rutinní a bez přístupu k datům a přesně definovanému výpočetnímu modelu je replikace a ověření obtížnější než pro biomarkery založené na jednotlivých analytech.

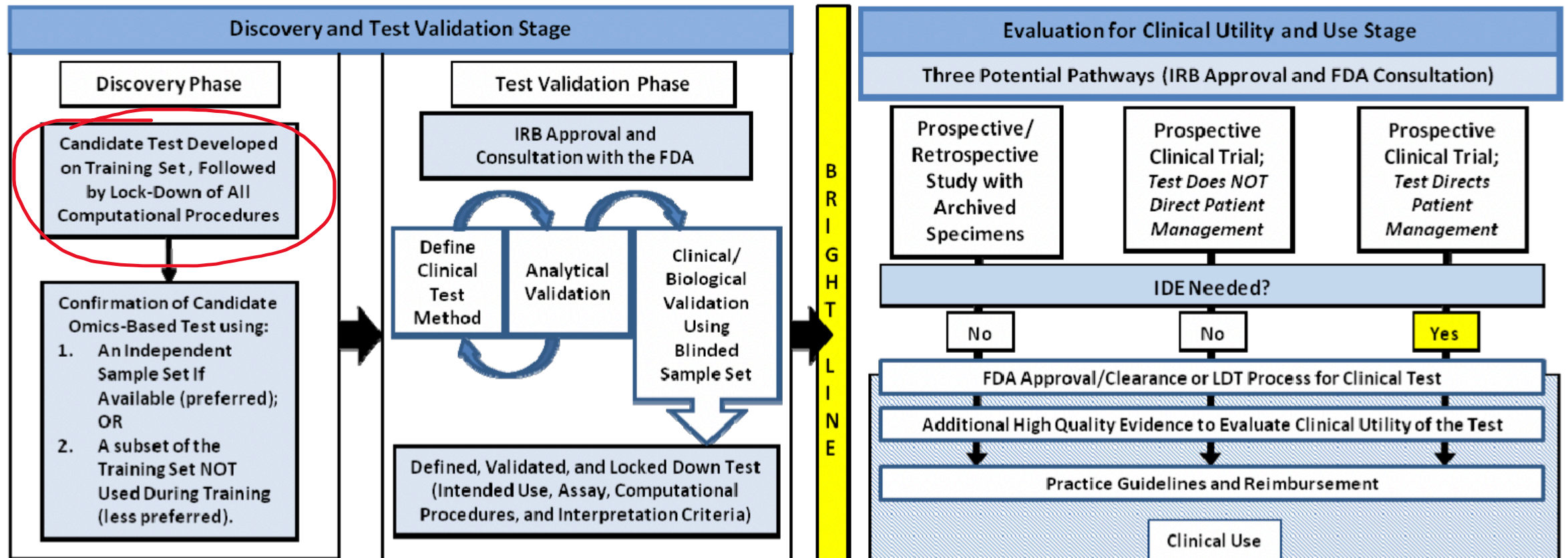


I když nezávislé validační studie jsou drahé, potřeba replikace v omicových studiích je nutná vzhledem ke složitosti dat, které mohou vést k chybám (od jednoduchých chyb správy dat až po nesprávně navržené výpočetní modely).



Tato úroveň složitosti neexistuje pro výzkum, vývoj a validaci testů s jedním biomarkerem.

Doporučení IOM komise pro vývoj testů založených na omicsových datech



Jak se hledá potenciální biomarker v omics datech

Biologická otázka (hypotéza)

Dizajn experimentu

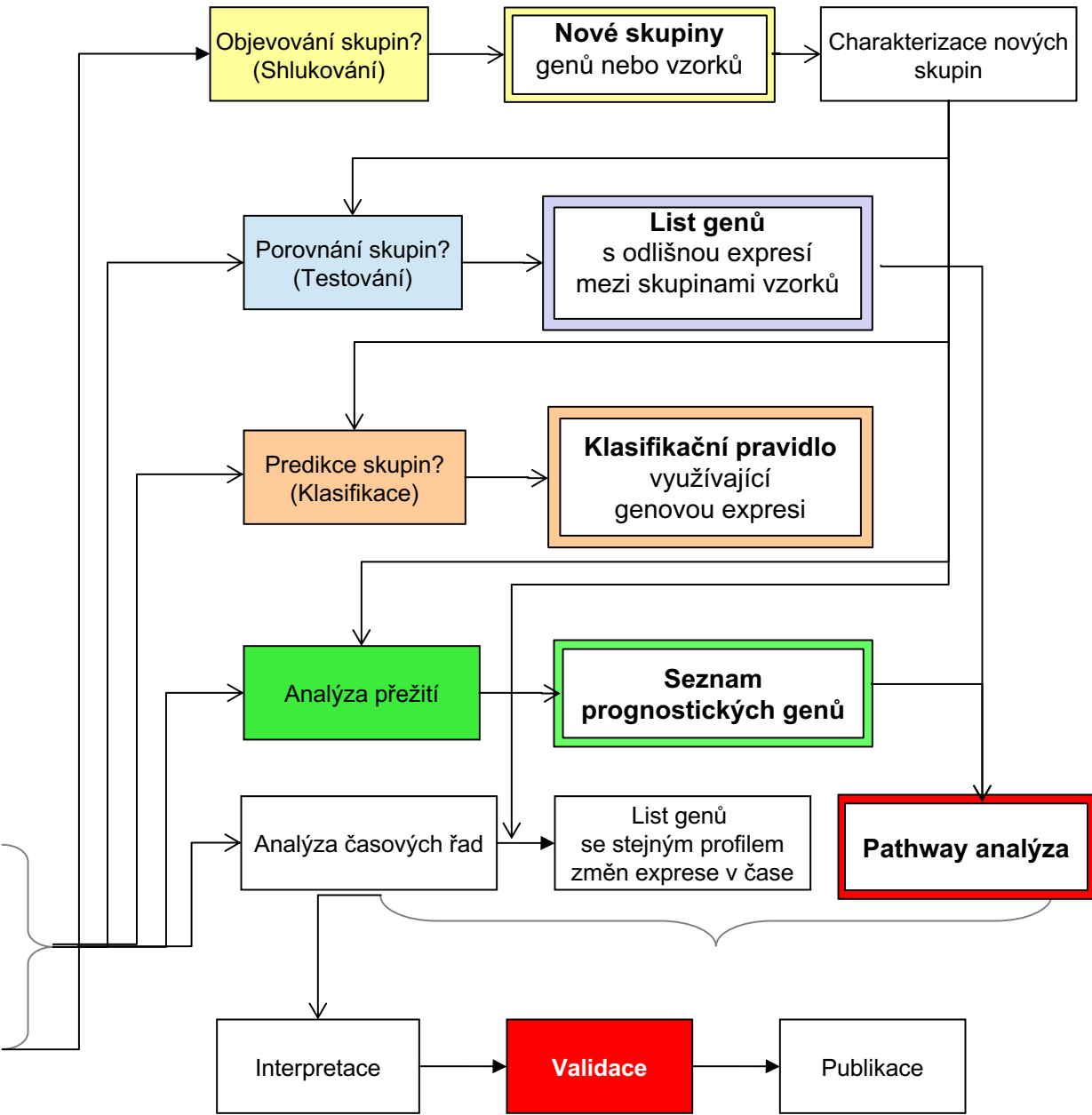
Provedení experimentu (hybridizace mikročipů, hmotnostní spektrometrie...)

N matic základních dat (jedna pro každý z N vzorků)

Kontrola kvality
Normalizace
Sumarizace

Finální datová matice N vzorků a K genů (proteinů)

Matice informací o vzorcích N x P (např. klinická data v medicíně)



Úprava
omicsových
dat do podoby,
kdy je možná
derivace
biomarkerů
trvá podstatně
déle než u
jiných dat

Data obsahují velké množství technického i biologického šumu, který je nutné odstranit

Protože jedno spuštění přístroje obvykle není schopno analyzovat všechny vzorky, vytváří se nežádoucí matoucí efekty (efekty dávky), které je nutno odstranit

Technologie jsou velice nové (a vznikají stále!) a algoritmy pro optimální zpracování jejich dat se vytvářejí a testují i 5-10 let - neexistují zlaté standardy a mnohé implementace jsou plné chyb

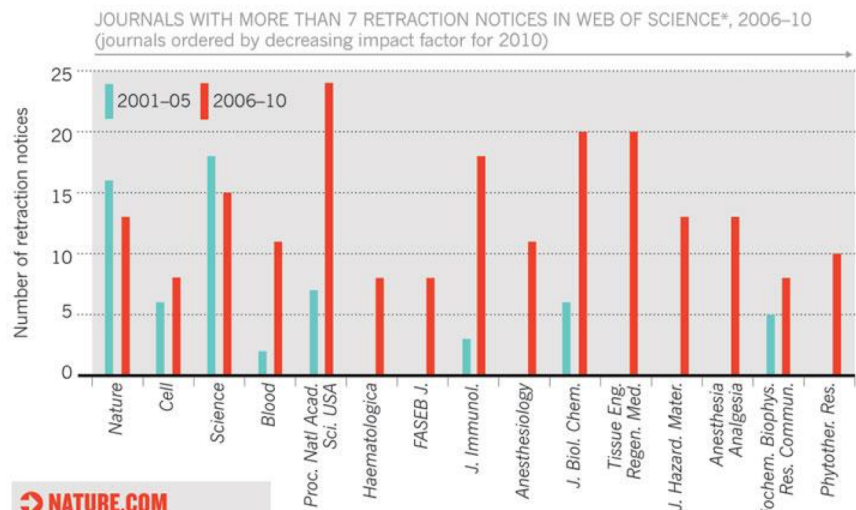
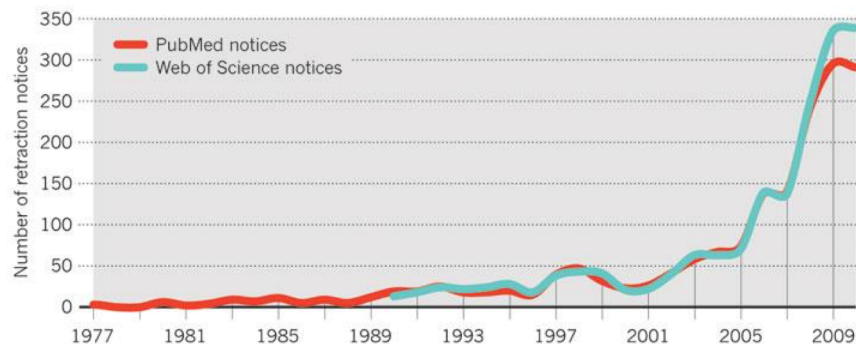


Vědecké časopisy jsou (přesto) plné odpadu



RISE OF THE RETRACTIONS

In the past decade, the number of retraction notices has shot up 10-fold (**top**), even as the literature has expanded by only 44%. It is likely that only about half of all retractions are for researcher misconduct (**middle**). Higher-impact journals have logged more retraction notices over the past decade, but much of the increase during 2006–10 came from lower-impact journals (**bottom**).



NATURE.COM
Read more about retractions:
go.nature.com/2uweek

*Not shown: *Acta Crystallographica E* saw 81 retractions during 2006–10.

Podíl článků stažených z tisku se zvyšuje

Za analyzovanou dekádu vzrostl počet článků pouze o 44%, počet retrakcí článků se zvýšil desetinásobně!

Pouze **0.02%** článků je staženo z tisku!

Van Noorden, R. (2011) Science publishing: The trouble with retractions. *Nature*. 2011 Oct 5;478(7367):26-8.

Důvody stažení publikací

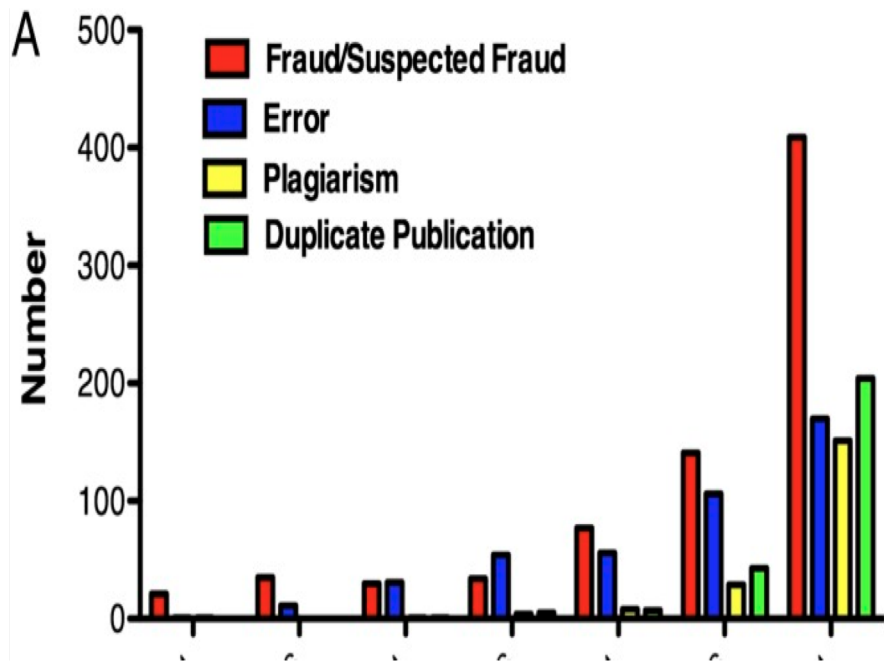
Analýza **2,047** stažených biomedicínských a přírodovědných vědeckých článků

Nejčastější důvod: **podvod** (nebo podezření z podvodu): **43.4%**

21.3% článků bylo staženo kvůli **čestné chybě (honest error)**

Až **31.8%** článků, které byly staženy zůstali neoznačeny

Většina autorů své stažené články stále cituje...



Adapted Figure 1 from Fang et al. (2012) Misconduct accounts for the majority of retracted scientific publications. PNAS 2012 Oct 16; 109(42):17028-17033

Retractions in the medical literature: how many patients are put at risk by flawed research?

R Grant Steen

Bylo analyzováno **180 primárních a 851 odvozených klinických studií**, které byly provedeny na základě výzkumu ze stažených publikací.

U 180 primárních studií bylo léčeno **9189** pacientů (z více než 28 tisíc)

U 851 odvozených studií bylo léčeno **70 501** pacientů (z více než 400 tisíc)

Studie, které byly staženy pro **podvod**, léčily statisticky významně více pacientů, než studie, které byly staženy pro **chybu**.

Biomarkery z omicsových dat

Jsou často komplexní:

- **Složené z více charakteristik** (více genů, proteinů...)
- **Bez jasně definovaného biologického zdůvodnění**

Pocházejí z dat:

- zatížených **významným technickým šumem** z různých zdrojů
- analyzovaných **metodami, které nejsou standardizované**
- které jsou pouze **korelované** s měřenou proměnnou (např. nejsou koncentrace ani počty molekul)
- které jsou **komplexní a obtížně se sdílejí**

Co dál?

- V průběhu semestru si ukážeme hlavní principy hledání jednotlivých biomarkerů a na nich založených modelů (testů), s důrazem na **reproducibilitu, robustnost a validaci**
- Vše budeme ilustrovat na konkrétních příkladech z praxe
- Budete mít možnost konzultovat vlastní experimenty