Detekce biomarkerů z omics experimentů

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

- Podzim 2024

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?

Co je to biomarker?

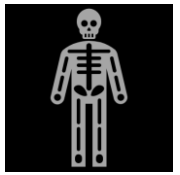
Biologický marker (biomarker):

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

Biomarkers Definitions Working Group (March 2001). "Biomarkers and surrogate endpoints: preferred definitions and conceptual framework". Clin. Pharmacol. Ther. (Review). 69 (3): 89–95.

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 mikroorganizmu

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

….

Využití jednotlivých biomarkerů v rozhodovacím PRAVIDLE (modelu/testu)

Jaké je riziko onemocnění u zdravého jedince?

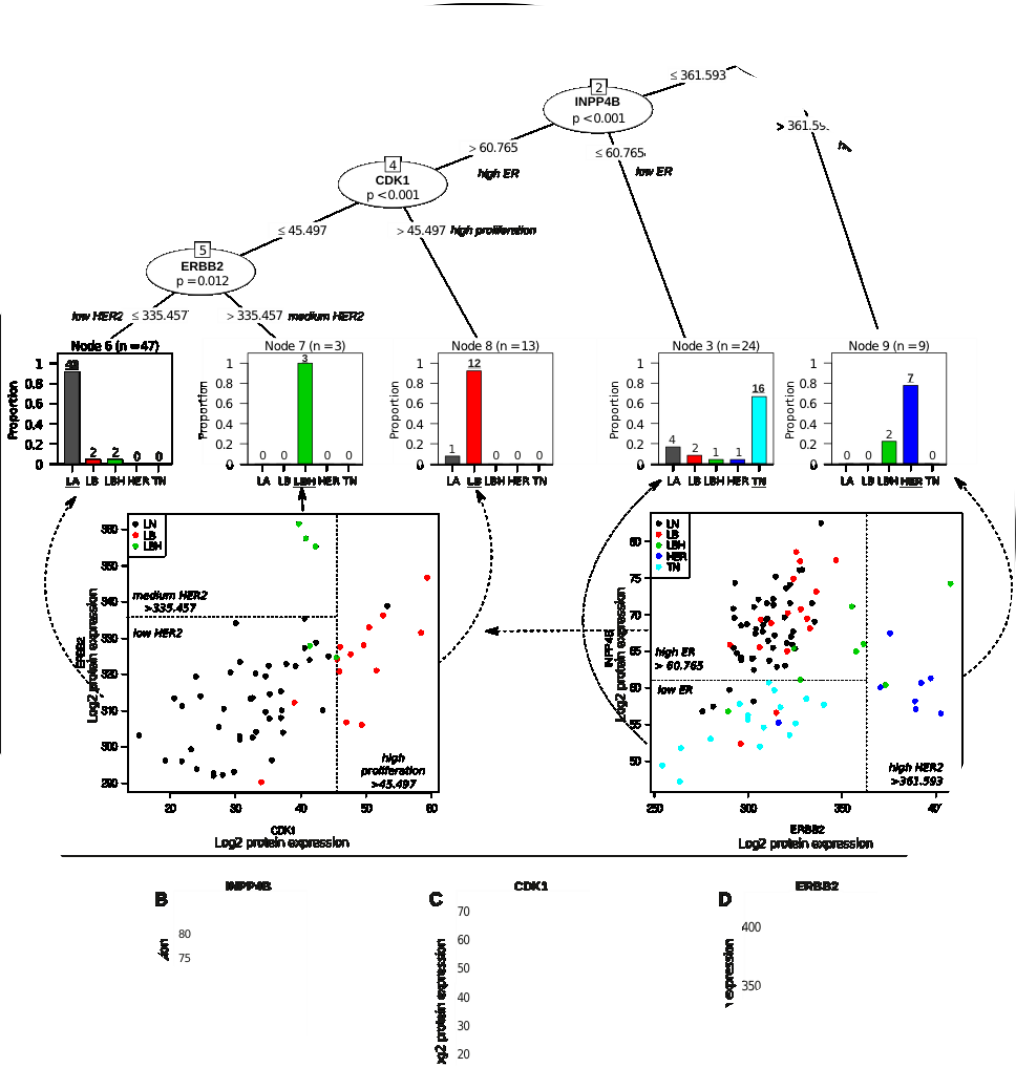
ká je to nemoc?

Zhoršuje se nemoc?

Jak bude nemoc probíhat?

Jaká je šance, že se nemoc vrátí?

Bude terapie úspěšná?



Biomarkery a modely

- Biomarker může být založen na **jediném analytu,** nebo na **jejich kombinaci v modelu** (klasifikátoru)

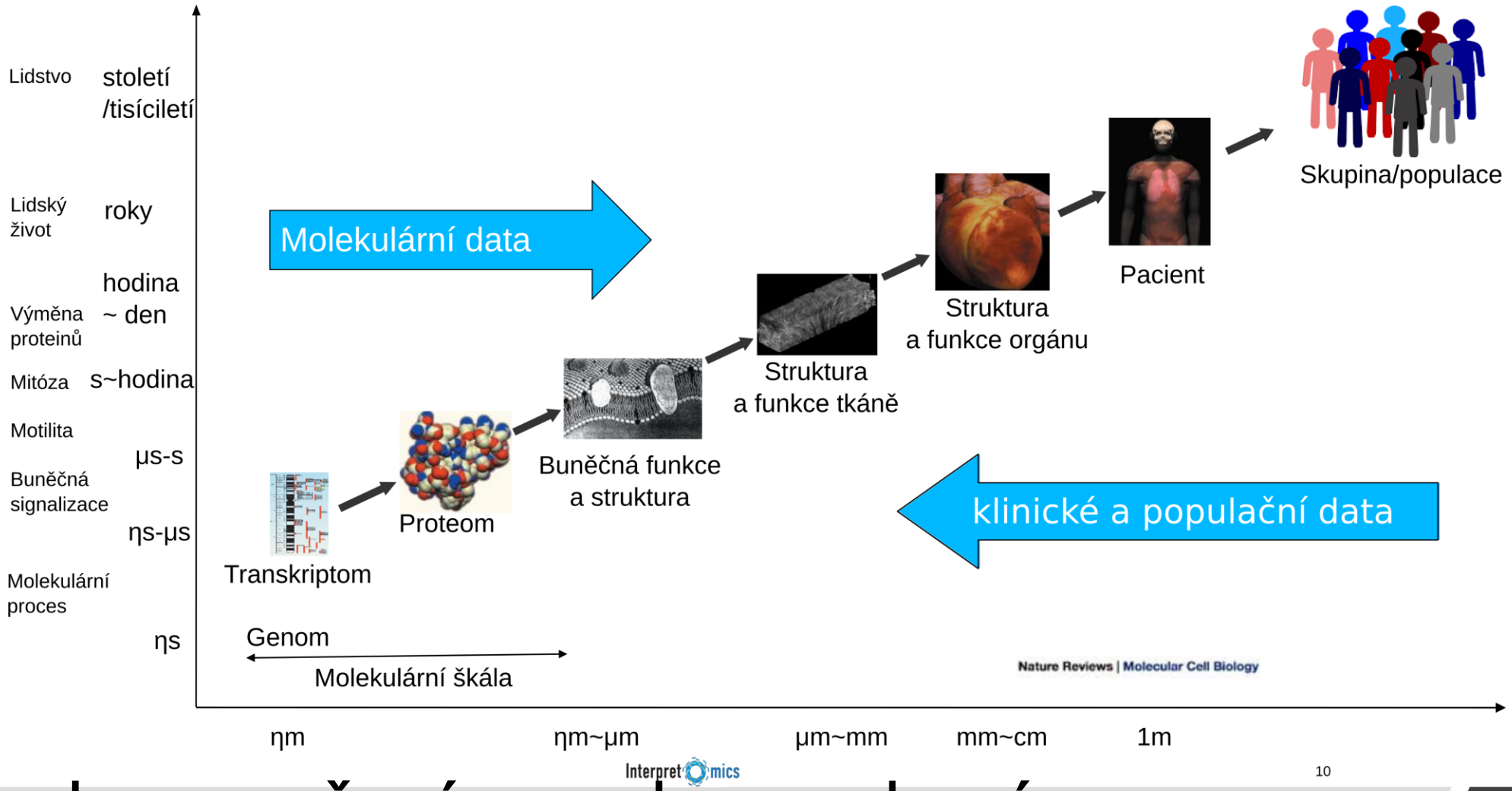
- Je to právě **kombinace více analytů** (genů, proteinů, metabolitů…), která je typická pro biomarkery z omicsových dat

Aktivita

Co musí biomarker (nebo model) splňovat

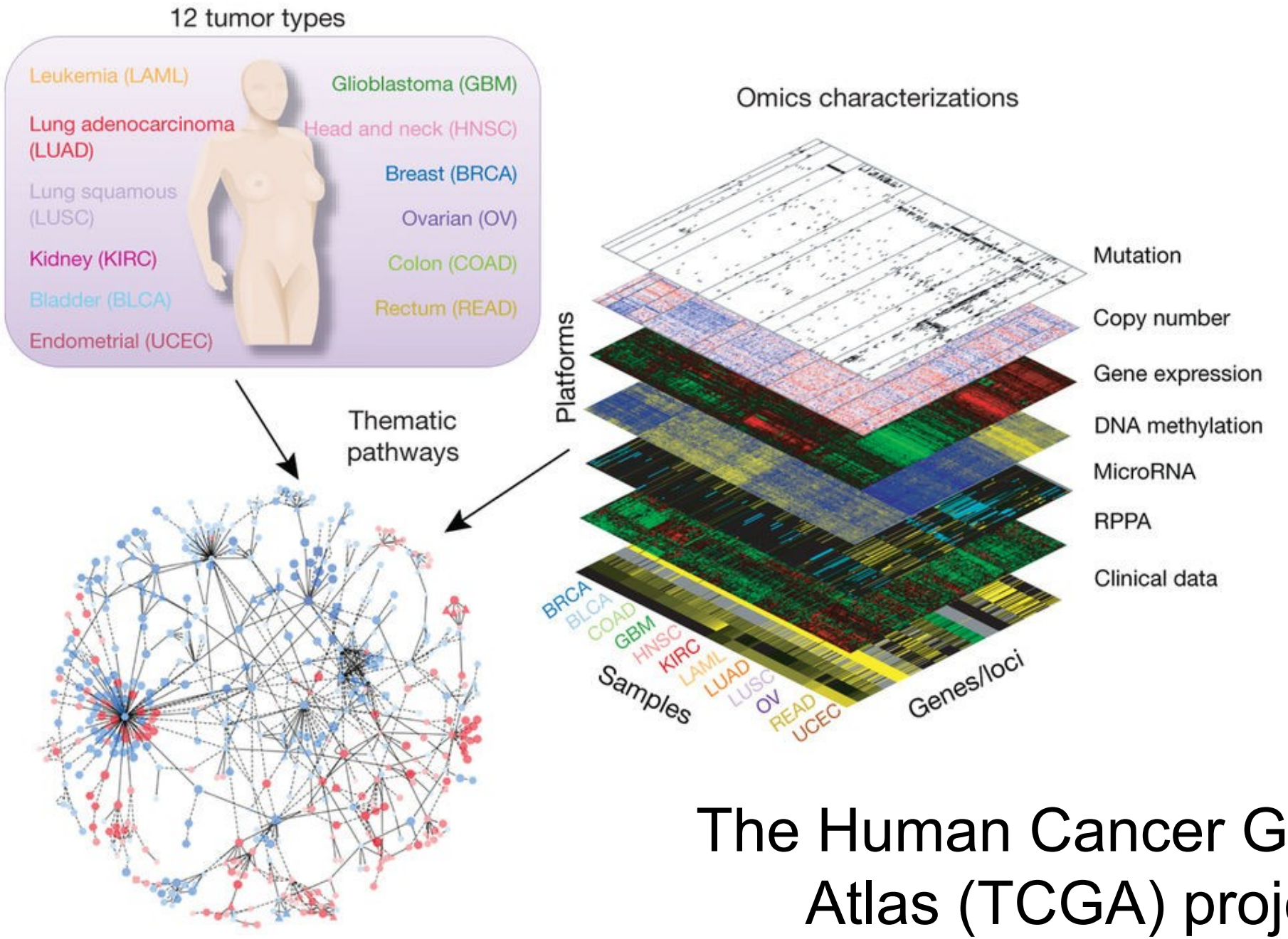
Musí být <u>použitelný rutinně v praxi:</u>

- **přesný** (dostatečně citlivý a dostatečně specifický)
- **robustní** (co nejméně omezen technologií měření)
- **reproducibilní** (obecně platný i na jiné populaci se stejnými charakteristikami než na té na které byl vytvořen)



Mnohorozměrná povaha moderní biomedicíny

Všechny “omics”?

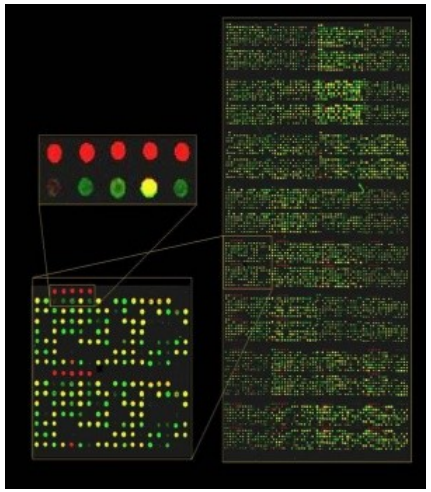


The Human Cancer Genome Atlas (TCGA) projekt

Data z omics experimentů

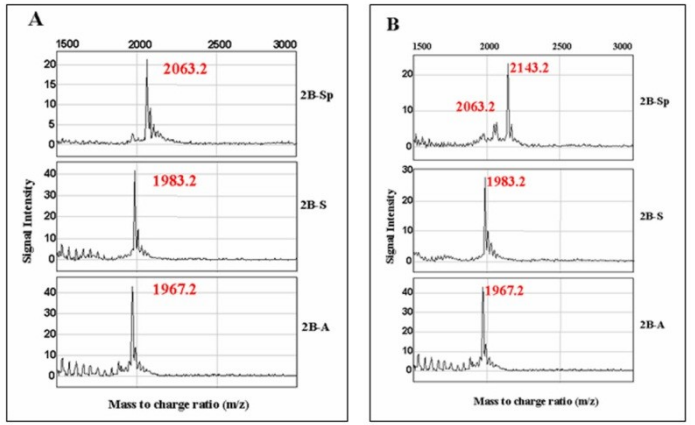
- Moderní vysoce pokryvné molekulární technologie produkují obrovské tabulky komplexních dat

Mikročipy



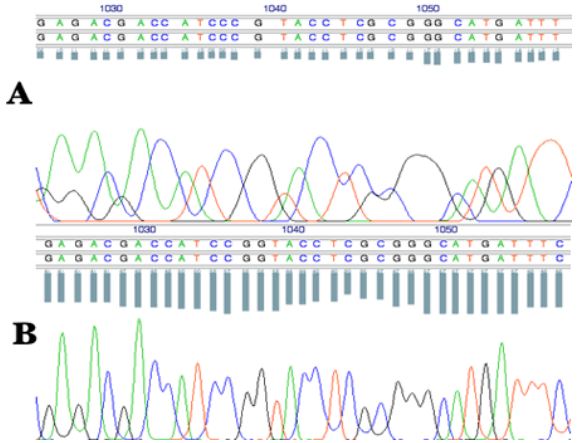
☑ Desítky až tisíce genů nebo transkriptů na vzorek

Hmotnostní spektrometrie



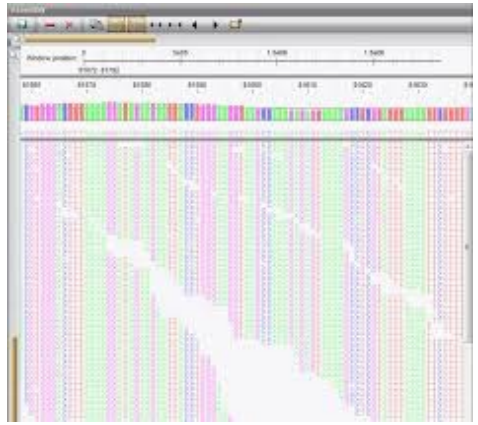
☑ Tisíce spekter proteinů, metabolitů nebo malých molekul na vzorek

Sekvenace DNA



☑Genom s biliony nukleotidů na vzorek

Sekvenace nové generace



☑ Miliony krátkých čtení DNA na vzorek




Data omics experimentů

PHASE TWO : INTERPRETATION

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?

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**

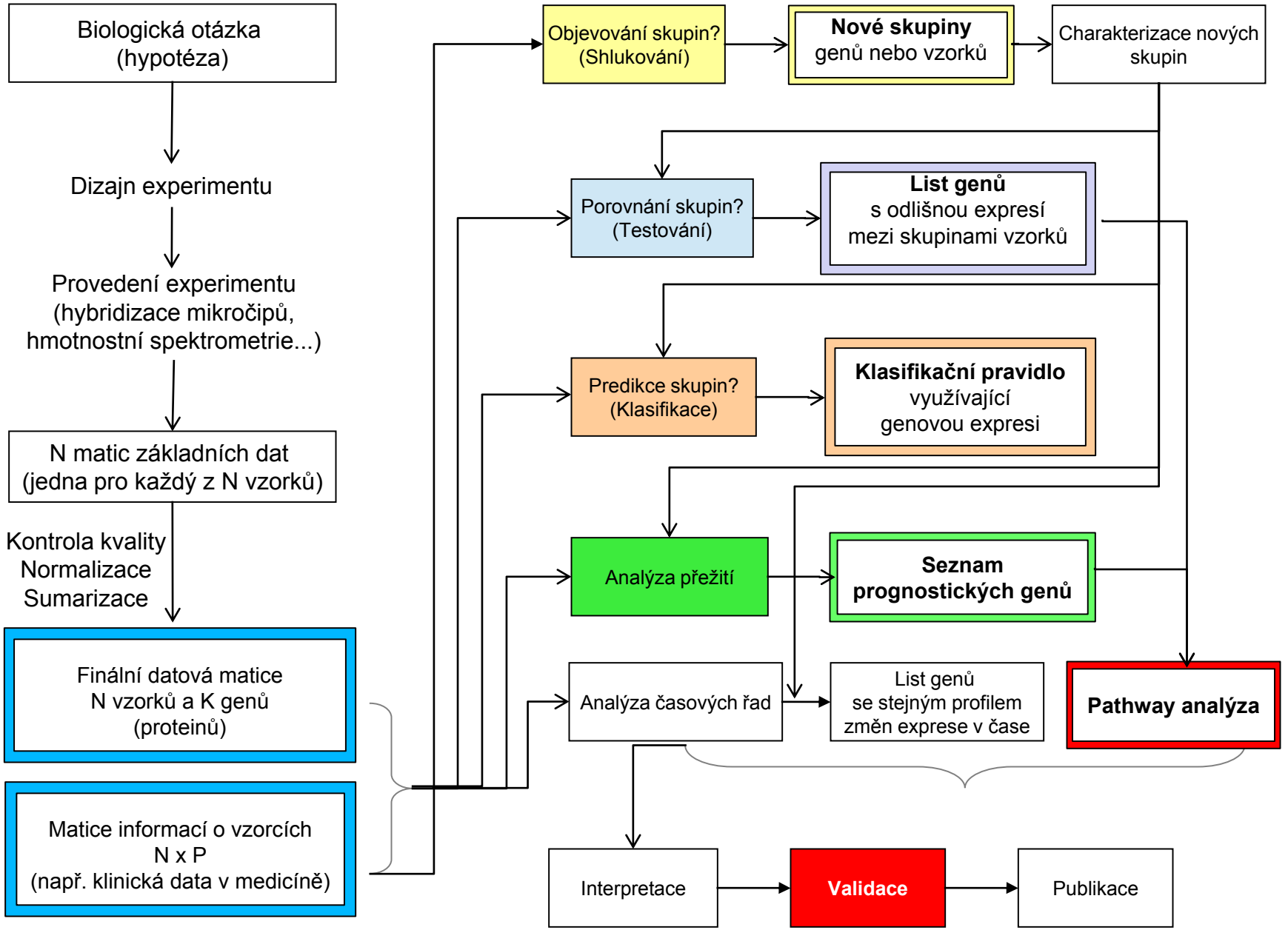
Jejich předzpracování pro statistickou analýzu je **náročné** a **vysoce specifické** pro daný typ platformy

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

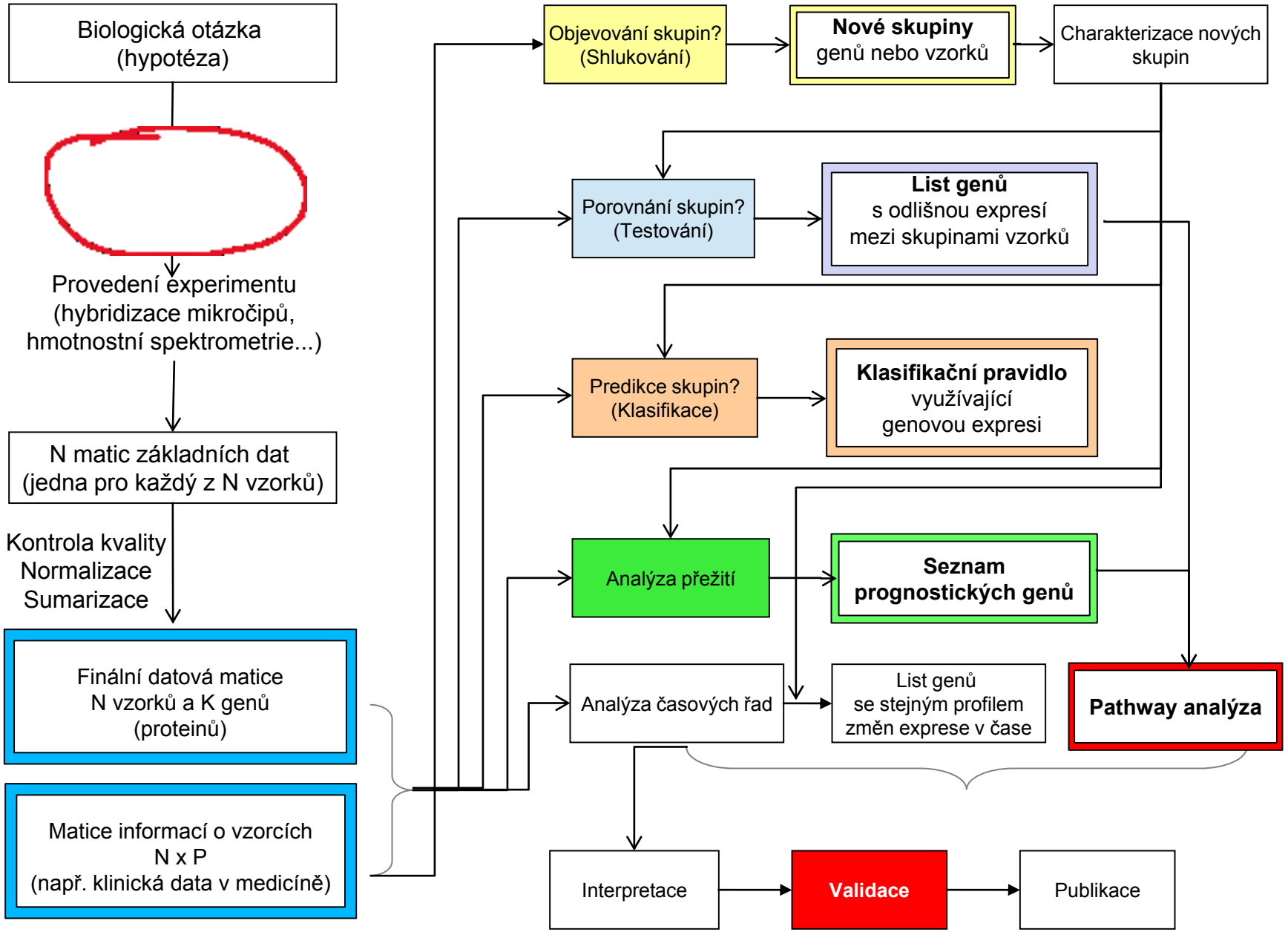
Zkoumané **proměnné jsou často korelované** a mají mezi sebou komplexní vztahy (geny, proteiny…)

Specifika dat z omics experimentů

Jak se hledá potenciální biomarker v omics datech



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)

… analýza těchto da
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”

ARTICLES

nature medicine

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 Chan[4], Robyn Sayer[4], Janiel Cragun[4], Hope Cottrill[4], Michael J Kelley[2], Rebecca Petersen[5], David Harpole[5], Jeffrey Marks[5], Andrew Berchuck[1,6], Geoffrey S Ginsburg[1,2], Phillip Febbo[1–3], Johnathan Lancaster[4] & Joseph R Nevins[1–3]

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!

Více info: https://ori.hhs.gov/content/case-summary-potti-anil

Jak skandál změnil svět omicsového 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.

PERSPECTIVE

OPEN

doi:10.1038/nature12564

Criteria for the use of omics-based predictors in clinical trials

Lisa M. McShane[1], Margaret M. Cavenagh[1], Tracy G. Lively[1], David A. Eberhard[2], William L. Bigbee[3], P. Mickey Williams[4], Jill P. Mesirov[5], Mei-Yin C. Polley[1], Kelly Y. Kim[1], James V. Tricoli[1], Jeremy M. G. Taylor[6], Deborah J. Shuman[1], Richard M. Simon[1], James H. Doroshow[1] & Barbara A. Conley[1]

The US National Canc
'omics'-based test d
omics-based tests for
mathematical model
encouraged to consid
checklist will be use
therapy.

Clinical Chemistry 60:10
1256–1257 (2014)

Perspective

Where Are All the New Omics-Based Tests?

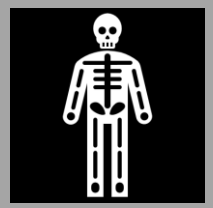
Patrick M. Bossuyt[1*]

"Why?" is the inevitable question. Why have so few biomarkers made it to everyday clinical care? Why, despite billions of dollars worldwide in omics-based research? We have been promised multiple breakthroughs, and numerous biomarker discoveries have been announced, but it is fair to say that, up to this day, clinical medicine has not

issues. A working group then developed a checklist on the basis of the key principles in the IOM report and the results of the NCI workshop *(2)*. A short version appeared in *Nature* last year, and a version with a longer explanation and elaboration was published in *BMC Medicine (3)*.

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 **pochopitelnou 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

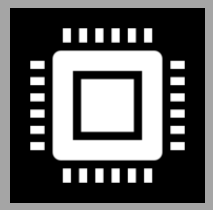
Absence jasného biologického odůvodnění testů omics biomarkerů – proč je to problém

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

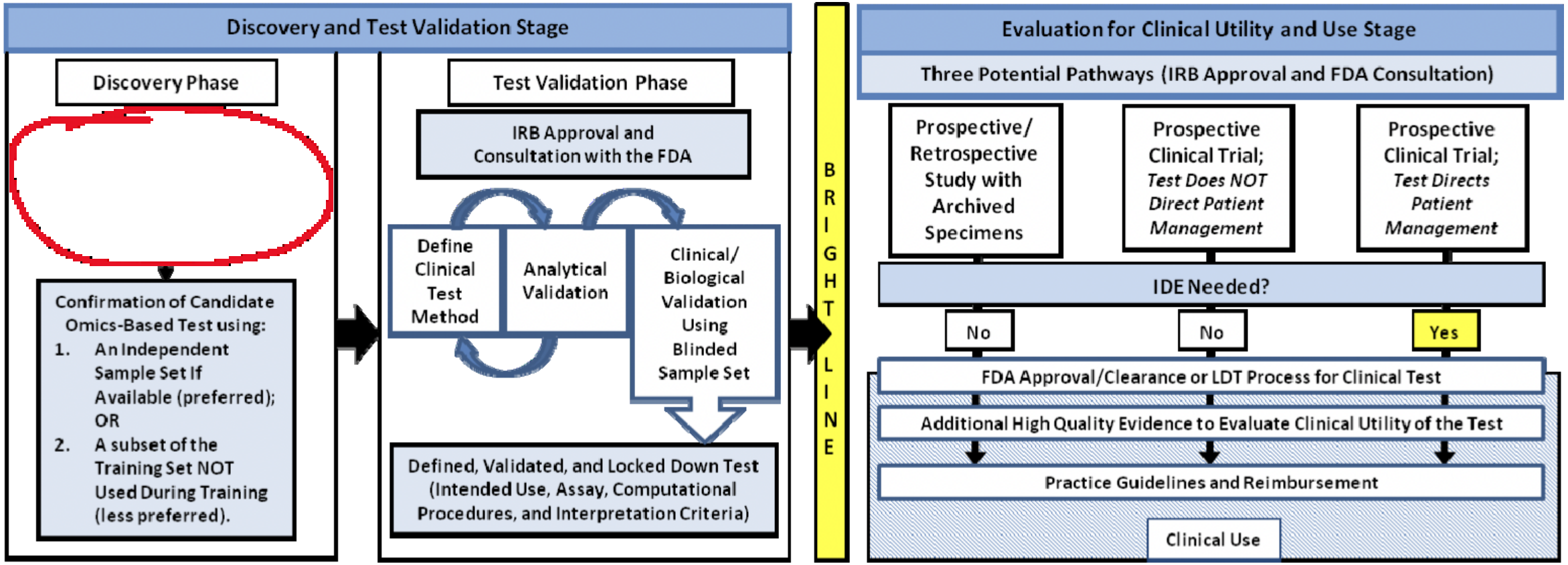
K dispozici jsou databázové úložiště pro soubory omicsových dat, ale sdílení dat není rutinní a bez přístupu k datům a přesně definovanému výpočetnímu modelu je replikace a ověření obtížnější než pro biomarkery založené na jednotlivých analytech.



I když nezávislé validační studie jsou drahé, potřeba replikace v omicsových 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



Published Mar. 26, 2019

RALEIGH, N.C. — Duke University will pay $112 million to settle a whistleblower lawsuit after federal prosecutors said a research technician's fake data landed millions of dollars in federal grants, the school and the government said Monday.

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

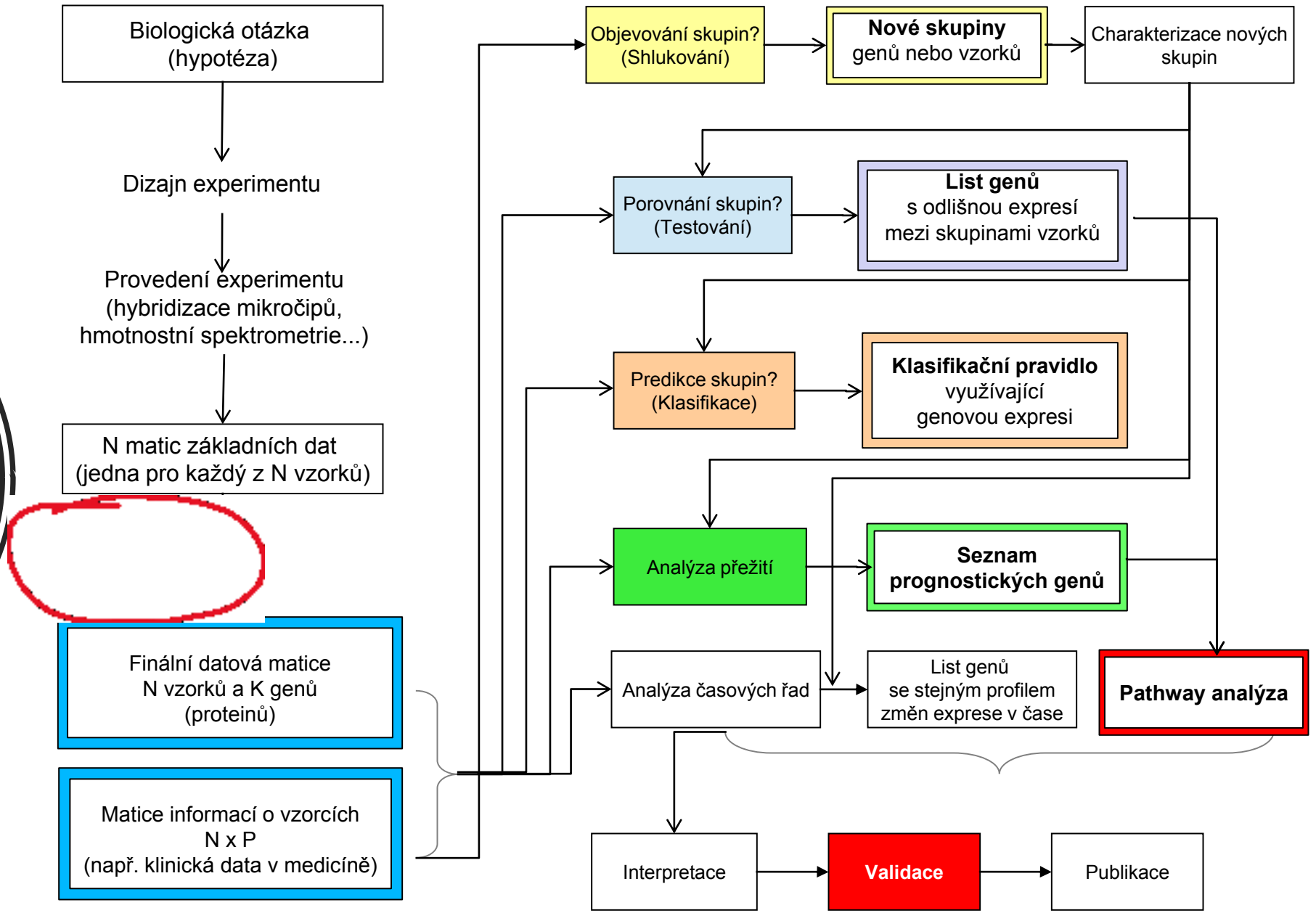
JAN 29 2015 THUR — LAWSUIT SET TO START

Více info: https://ori.hhs.gov/content/case-summary-potti-anil

Anil Potti a další skandály podrobně

- https://www.ncbi.nlm.nih.gov/books/NBK475955/
- https://retractionwatch.com/2015/11/07/its-official-anil-potti-faked-data-say-feds/

Jak se hledá potenciální biomarker v *omics* datech



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

Proč jsou omicsová data problematická?

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**

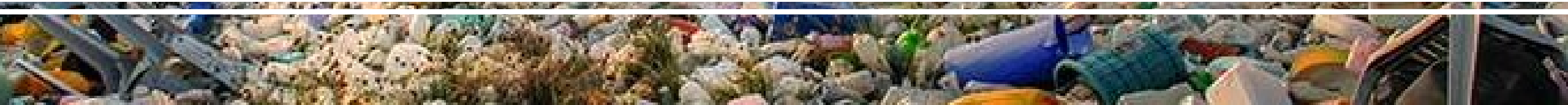
Jejich předzpracování pro statistickou analýzu je **náročné** a **vysoce specifické** pro daný typ platformy

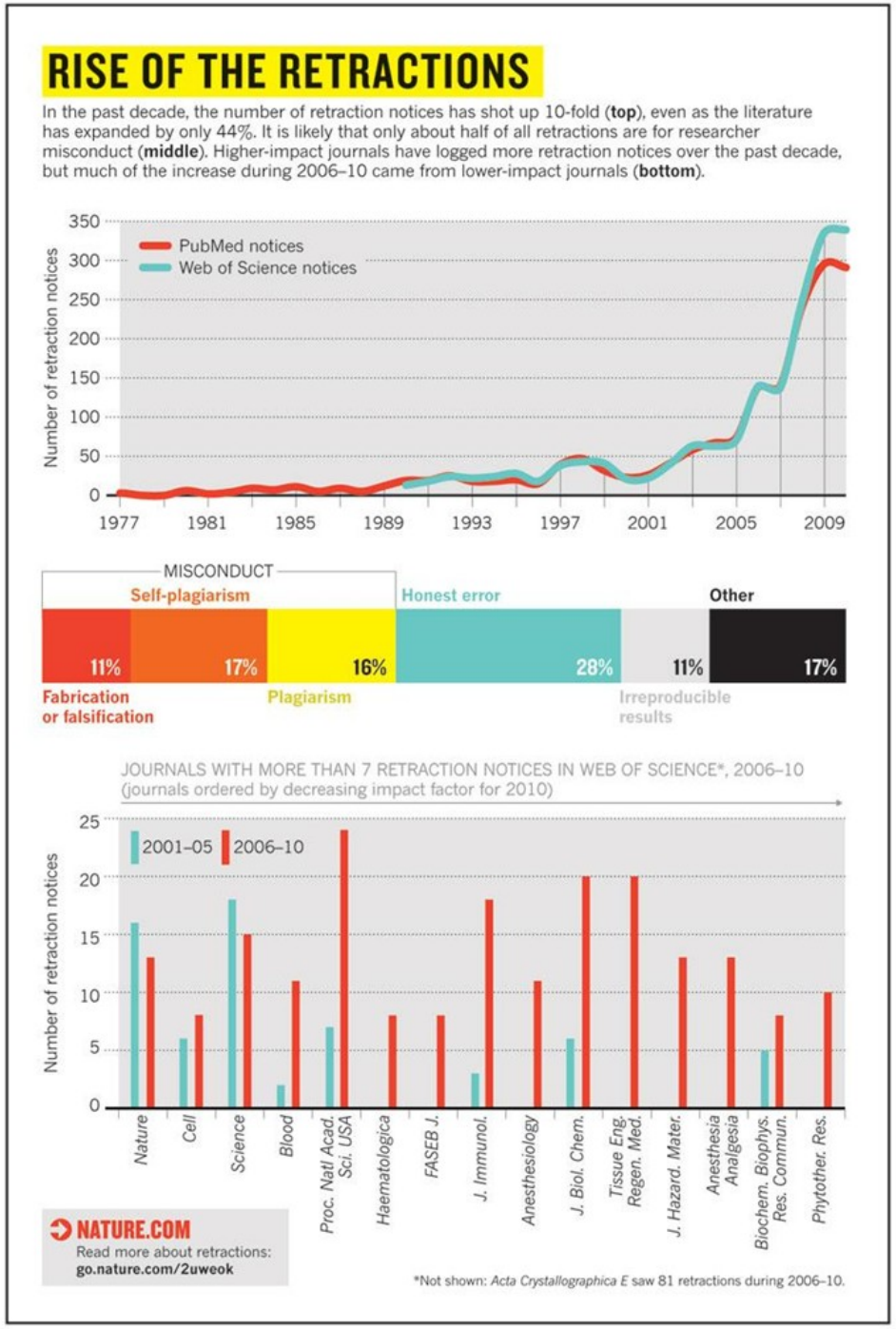
Počet vzorků je mnohem **menší než** počet sledovaných **proměnných**.

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

Specifika dat z omics experimentů

Vědecké časopisy jsou (přesto) plné odpadu





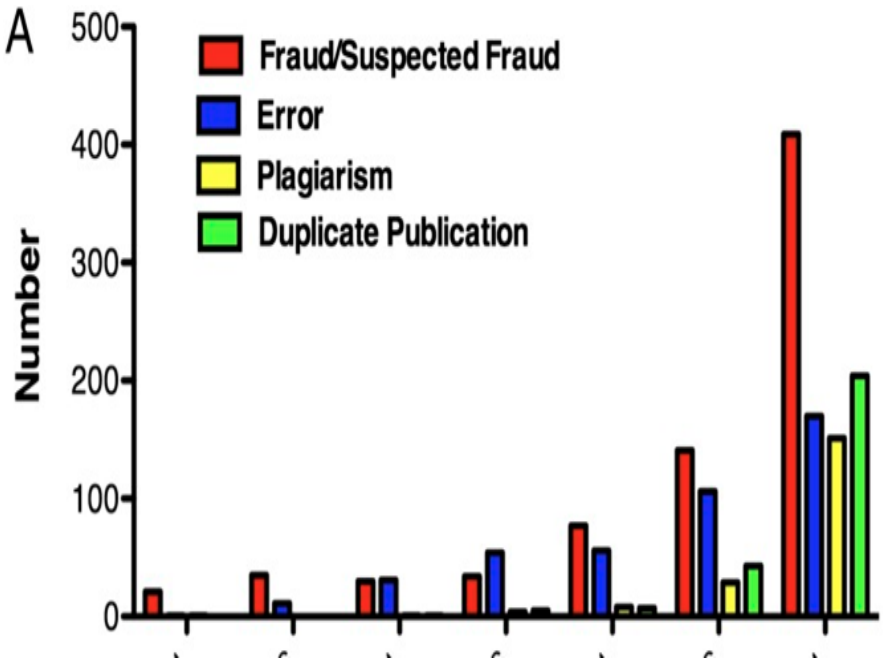
Podíl článků stažených z tisku se zvyšuje

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):17028-1703

DAVIDE BONAZZI

What a massive database of retracted papers reveals about science publishing's 'death penalty'

By Jeffrey Brainard, Jia You | Oct. 25, 2018 , 2:00 PM

Rethinking retractions

Better editorial oversight, not more flawed papers, might explain a flood of retractions

RELATED STORY
A scientist's fraudulent studies put patients at risk

RELATED STORY
Volunteer watchdogs pushed a small country up the rankings

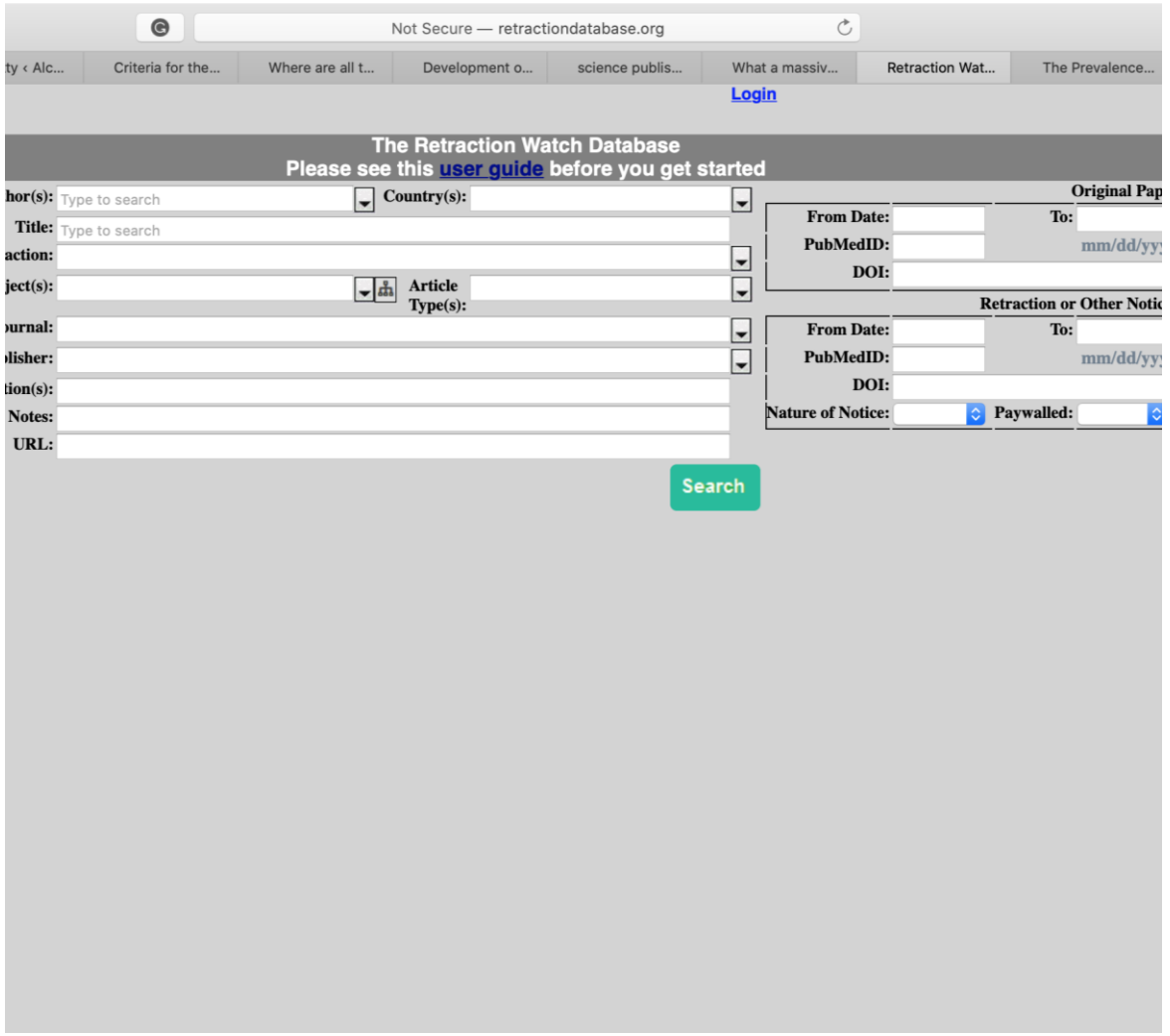
RELATED STORY
One publisher, more than 7000 retractions

RELATED STORY
Fallout for co-authors

METHODOLOGY
About these data

Nearly a decade ago, headlines highlighted a disturbing trend in science: The number of articles

Database of retractions



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.**

Article | Published: 11 March 2020

Microbiome analyses of blood and tissues suggest cancer diagnostic approach

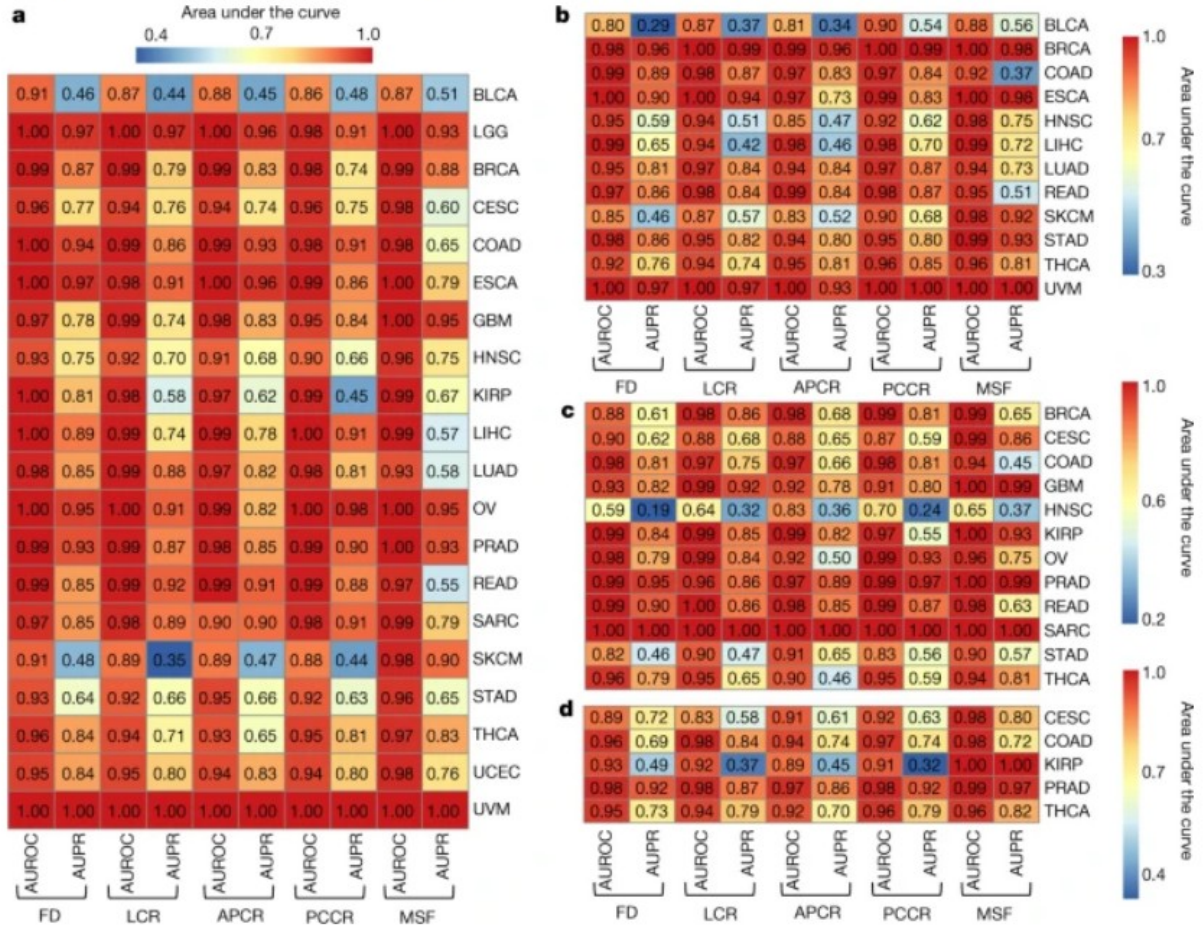
Gregory D. Poore, Evguenia Kopylova, Qiyun Zhu, Carolina Carpenter, Serena Fraraccio, Stephen Wandro, Tomasz Kosciolek, Stefan Janssen, Jessica Metcalf, Se Jin Song, Jad Kanbar, Sandrine Miller-Montgomery, Robert Heaton, Rana Mckay, Sandip Pravin Patel, Austin D. Swafford & Rob Knight

Nature **579**, 567–574 (2020) | Cite this article

Je to pořád náročné...

Systematic characterization of the cancer microbiome provides the opportunity to develop techniques that exploit non-human, microorganism-derived molecules in the diagnosis of a major human disease. Following recent demonstrations that some types of cancer show substantial microbial contributions[1,2,3,4,5,6,7,8,9,10], we re-examined whole-genome and whole-transcriptome sequencing studies in The Cancer Genome Atlas[11] (TCGA) of 33 types of cancer from treatment-naive patients (a total of 18,116 samples) for microbial reads, and found unique microbial signatures in tissue and blood within and between most major types of cancer. These TCGA blood signatures remained predictive when applied to patients with stage Ia–IIc cancer and cancers lacking any genomic alterations currently measured on two commercial-grade cell-free tumour DNA platforms, despite the use of very stringent decontamination analyses that discarded up to 92.3% of total sequence data. In addition, we could discriminate among samples from healthy, cancer-free individuals ($n = 69$) and those from patients with multiple types of cancer (prostate, lung, and melanoma; 100 samples in total) solely using plasma-derived, cell-free microbial nucleic acids. This potential microbiome-based oncology diagnostic tool warrants further exploration.

Fig. 3: Classifier performance for cancer discrimination using mbDNA in blood and as a complementary diagnostic approach for cancer 'liquid' biopsies.



a, Model performance heatmap analogous to Fig. 1f–h to predict one cancer type versus all others using blood mbDNA with TCGA study IDs on the right (Extended Data Fig. 1a); at least 20 samples were

[Microbiome analyses of blood and tissues suggest cancer diagnostic approach | Nature]

MICROBIAL GENOMICS

LETTER

Gihawi *et al.*, *Microbial Genomics* 2023;9:001088

DOI 10.1099/mgen.0.001088

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

Caution regarding the specificities of pan-cancer microbial structure

Abraham Gihawi[1,*], Colin S. Cooper[1] and Daniel S. Brewer[1,2]

Abstract

Results published in an article by Poore *et al.* (*Nature*. 2020;579:567–574) suggested that machine learning models can almost perfectly distinguish between tumour types based on their microbial composition using machine learning models. Whilst we believe that there is the potential for microbial composition to be used in this manner, we have concerns with the paper that make us question the certainty of the conclusions drawn. We believe there are issues in the areas of the contribution of contamination, handling of batch effects, false positive classifications and limitations in the machine learning approaches used. This makes it difficult to identify whether the authors have identified true biological signal and how robust these models would be in use as clinical biomarkers. We commend Poore *et al.* on their approach to open data and reproducibility that has enabled this analysis. We hope that this discourse assists the future development of machine learning models and hypothesis generation in microbiome research.

bioxRiv

Posted January 18, 2023.

Most models do not perform any better than models constructed using no information

Models pronounce nonsensical genera are informative of tumour type

Velarivirus	Cervical Cancer	Grapevine is natural host[8]
Tritimovirus	Colon Cancer	Known to infect cereals[9]
Dinovernavirus	Renal Clear Cell Carcinoma	Contains insect viruses[10]
Bacillarnavirus	Lung Squamous Cell Carcinoma	Infects algae[11]
Rymovirus	Ovarian serous	Infects species of grass[12]
Ignicoccus	Prostate	Identified in marine hydrothermal vents[13]
Salinimicrobium	Testicular Cancer	Halophilic genus identified from marine environments[14]

Potential for read misclassification

Normalization introduces variance and permits modelling

LOGY

DOI 10.1099/mgen.0.001088

OPEN DATA OPEN ACCESS

The models are trained on unbalanced data

l S. Brewer[1,2]

ature. 2020;579:567–574) suggested that machine learning models can almost
ed on their microbial composition using machine learning models. Whilst we
believe that there is the potential for microbial composition to be used in this manner, we have concerns with the paper that
ions drawn. We believe there are issues in the areas of the contribution of con-
sitive classifications and limitations in the machine learning approaches used.
This makes it difficult to identify whether the authors have identified true biological signal and how robust these models would
be in use as clinical biomarkers. We commend Pearce *et al.* on their approach to open data and reproducibility that has enabled
ine learning models and hypothesis genera-

bioxRiv

Posted January 18, 2023.

Cold Spring Harbor Laboratory
bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

Confirmatory Results

Follow this preprin

Reply to: Caution Regarding the Specificities of Pan-Cancer Microbial Structure

Gregory D. Sepich-Poore, Evguenia Kopylova, Qiyun Zhu, Carolina Carpenter, Serena Fraraccio, Stephen Wandro Tomasz Kosciolek, Stefan Janssen, Jessica Metcalf, Se Jin Song, Jad Kanbar, Sandrine Miller-Montgomery, Robert Heaton, Rana Mckay, Sandip Pravin Patel, Austin D Swafford, Rob Knight

doi: https://doi.org/10.1101/2023.02.10.528049

Posted February 13, 2023.

Abstract

The cancer microbiome field tremendously accelerated following the release of our manuscript nearly three years ago[1], including direct validation of our cancer type-specific conclusions in independent, international cohorts[2,3] and the tumor microbiome's adoption into the hallmarks of cancer[4]. Disentangling contamination signals from biological signals is an important consideration for this research field. Therefore, despite numerous, high-impact, peer-reviewed research papers that either validated our conclusions or extended them using data we released[2,5–13], we carefully considered criticism raised by Gihawi *et al.* about potential mishandling of contaminants, batch effects, and machine learning approaches—all of which were central topics in our manuscript. Nonetheless, a close examination of each concern alongside the original manuscript and re-analyses of our published data strongly demonstrates the robustness of the original findings. To remove all doubt, however, we have reproduced all key conclusions from the original manuscript using only overlapping bacterial genera identified in a highly decontaminated, multi-cancer, international cohort (Weizmann Institute of Science, WIS)[2], with or without batch correction, and with multiclass machine learning analyses to mitigate class imbalances. Our published pan-cancer mycobiome manuscript[3] also affirms these findings using updated, state-of-the-art methods. We also note that every analysis shown here was possible using public data and code that we had already provided.

Cold Spring Harbor Laboratory

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

Contradictory Results

Follow this preprint

Major data analysis errors invalidate cancer microbiome findings

Abraham Gihawi, Yuchen Ge, Jennifer Lu, Daniela Puiu, Amanda Xu, Colin S. Cooper, Daniel S. Brewer, Mihaela Pertea, Steven L. Salzberg

doi: https://doi.org/10.1101/2023.07.28.550993

This article is a preprint and has not been certified by peer review [what does this mean?].

1 | 0 | 4 | 0 | 8 | 0 | 1519

Abstract | **Full Text** | Info/History | Metrics | Preview PDF

Posted July 31, 2023.

ASM Journals / mBio / Vol. 14, No. 5 / Major data analysis errors invalidate cancer microbiome findings

Human Microbiome | Research Article | 9 October 2023

Major data analysis errors invalidate cancer microbiome findings

Authors: Abraham Gihawi, Yuchen Ge, Jennifer Lu, Daniela Puiu, Amanda Xu, Colin S. Cooper, Daniel S. Brewer, Mihaela Pertea, Steven L. Salzberg | AUTHORS INFO & AFFILIATIONS

DOI: https://doi.org/10.1128/mbio.01607-23 · Check for updates

34 / 46 990 | CITE | PDF/EPUB

We re-analyzed the data from a recent large-scale study that reported strong correlations between microbial organisms and 33 different cancer types, and that created machine learning predictors with near-perfect accuracy at distinguishing among cancers. We found at least two fundamental flaws in the reported data and in the methods: (1) errors in the genome database and the associated computational methods led to millions of false positive findings of bacterial reads across all samples, largely because most of the sequences identified as bacteria were instead human; and (2) errors in transformation of the raw data created an artificial signature, even for microbes with no reads detected, tagging each tumor type with a distinct signal that the machine learning programs then used to create an apparently accurate classifier. Each of these problems invalidates the results, leading to the conclusion that the microbiome-based classifiers for identifying cancer presented in the study are entirely wrong. These flaws have subsequently affected more than a dozen additional published studies that used the same data and whose results are likely invalid as well.

Zapamatujme si:
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