Detekce biomarkerů z omics experimentů

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

- Podzim 2023

Jak probíhá předzpracová ní omicsových dat

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**

Jak vznikají čestné chyby?

Práce ve skupinách:

Napište tři příklady ke každému bodu

(10 min)

Nejčastější zdroje “čestných chyb” (honest errors)

- Chyby v měření a v laboratorních postupech
- Nesprávně zvolena statistická metodologie
- Manuální práce s daty

Jak můžeme tyto chyby minimalizovat?

Nedostatek kontroly

Nedostatek financí

Nedostatek znalostí

Nedbalost

Nedostatek času

Čestná chyba (honest error) – jak ji minimalizovat

Vhodný **návrh experimentu** (výběr analytické metody, počet a typ vzorků, randomizace….)

Reálný časový odhad

Minimalizace chyb v laboratoři

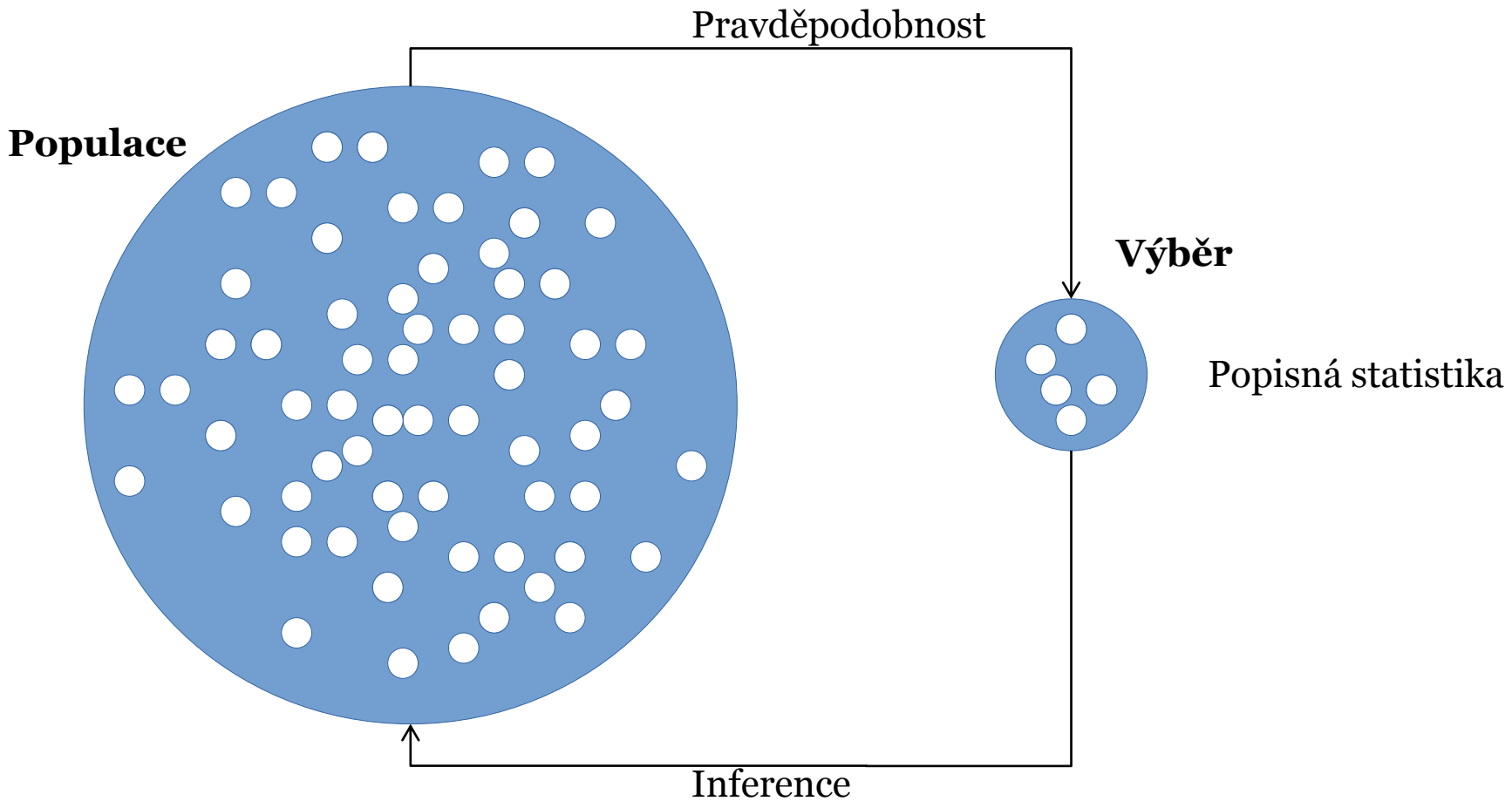
Vedení kompletních **záznamů**

Výběr vhodných metod pro statistickou analýzu dat

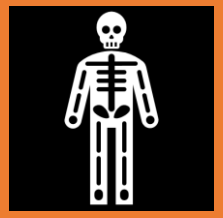
Správná **validace** výsledků

Návrh experimentu

Centrální dogma statistiky



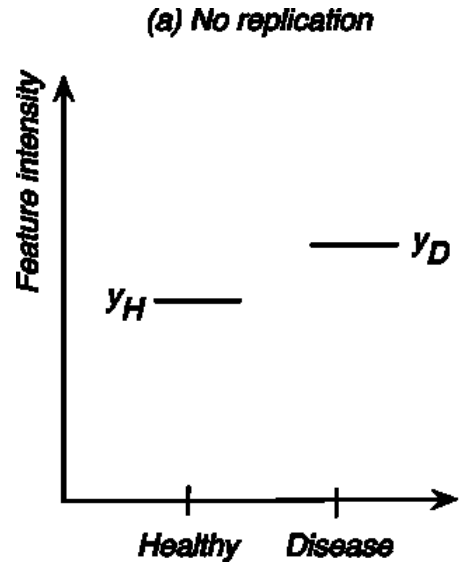
Kolik vzorků???



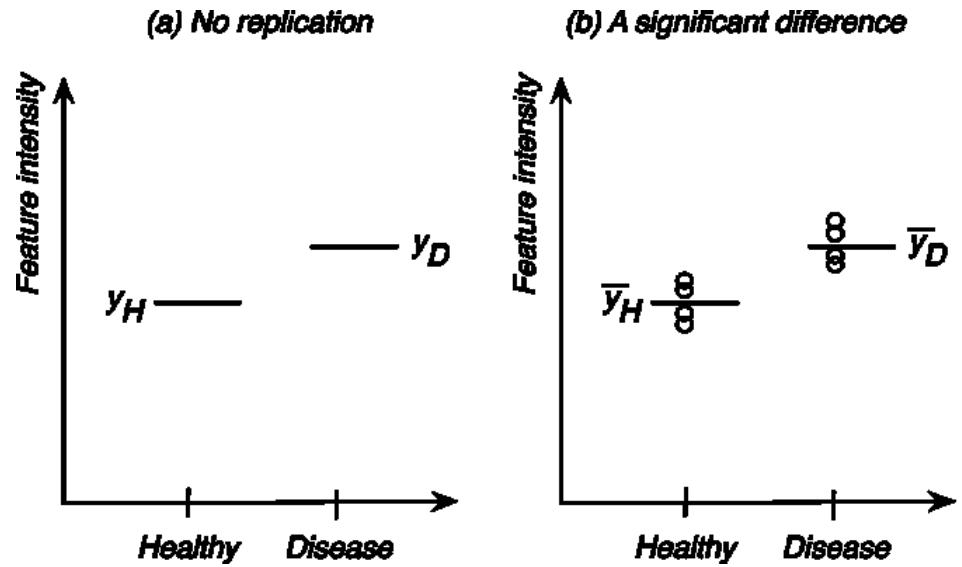
Čím variabilnější populace, tím více vzorků je potřeba na její dostatečný popis!

POČET VZORKŮ JE TAKÉ ZÁVISLÝ NA POUŽITÝCH STATISTICKÝCH METODÁCH!

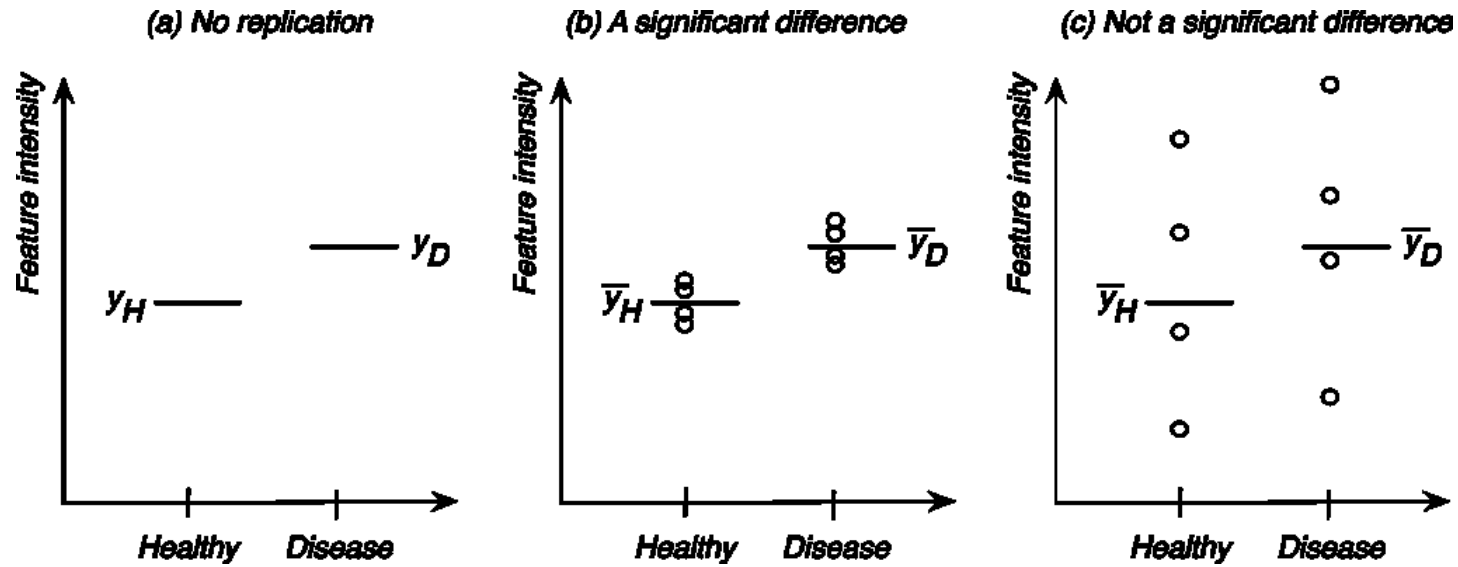
Replikáty



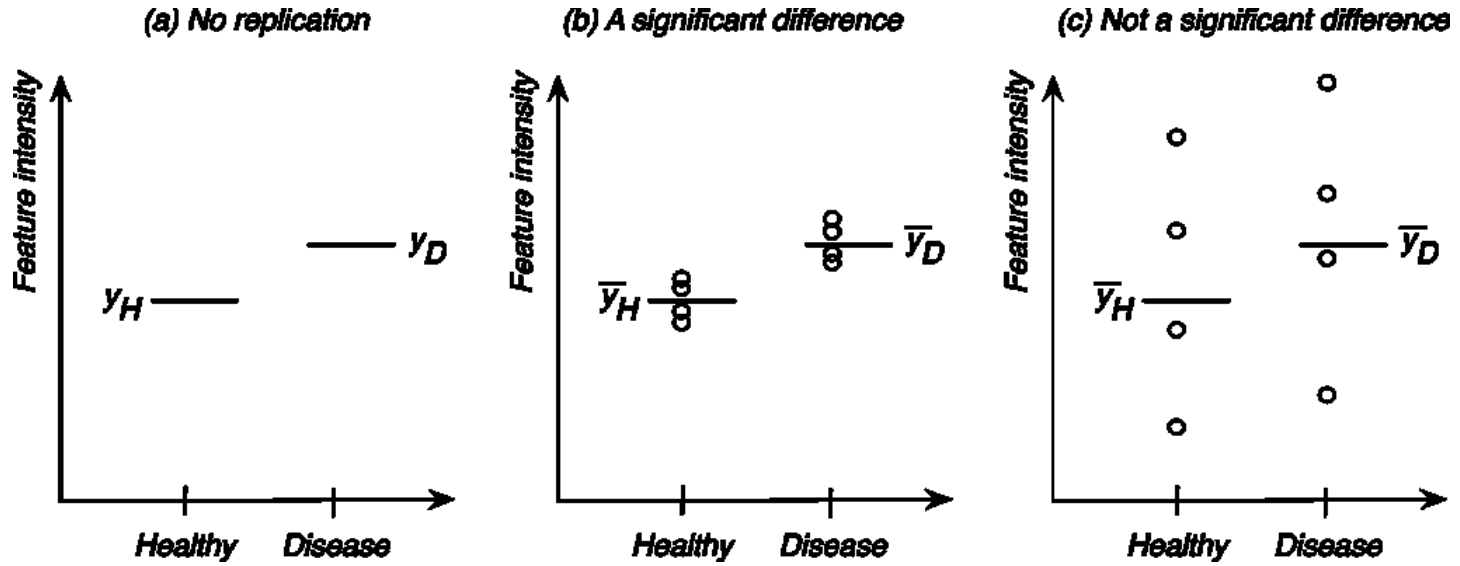
Replikáty


(a) No replication
(b) A significant difference
Feature intensity
Feature intensity
y_H
y_D
$\bar{y}_H$
$\bar{y}_D$
Healthy
Disease
Healthy
Disease

Replikáty


(a) No replication
Feature intensity
y_H
y_D
Healthy
Disease
(b) A significant difference
Feature intensity
$\overline{y}_H$
$\overline{y}_D$
Healthy
Disease
(c) Not a significant difference
Feature intensity
$\overline{y}_H$
$\overline{y}_D$
Healthy
Disease

Replikáty



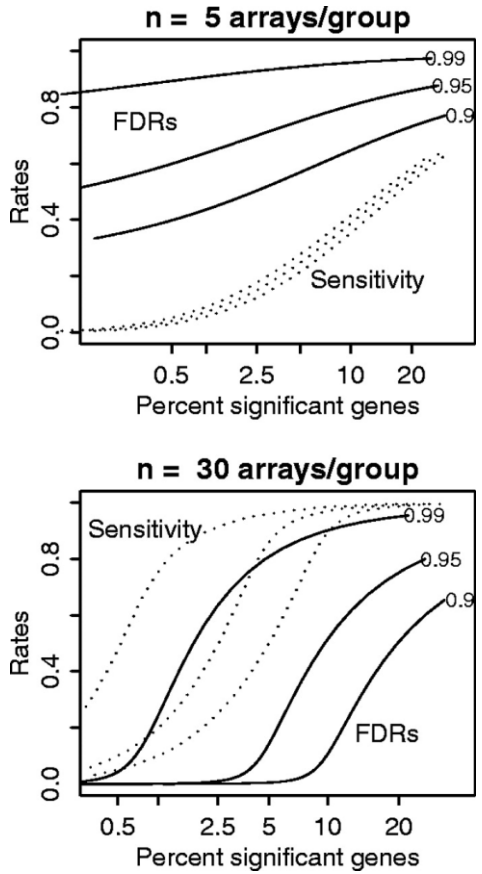
Replikáty jsou nutné pro odhad variability a statistické významnosti

Replikáty rozlišujeme: **technické** a **biologické**

Technické replikáty nezastupují replikáty biologické!!!

Technické replikáty pouze popisují přesnost postupu a techniky, ne však variabilitu v cílové populaci.

Vliv počtu vzorků a podílu významných genů na falešně pozitivní výsledky



FDR – false discovery rate - (plné křivky) a citlivost (tečkované křivky) jako funkce procenta významných genů. Každá křivka FDR je označena podílem skutečně neodlišně exprimovaných genů (p0). Křivky sensitivity jsou ve stejném pořadí jako křivky FDR, například horní křivka odpovídá p0 = 0,99.

From: False discovery rate, sensitivity and sample size for microarray studies
Bioinformatics. 2005;21(13):3017-3024. doi:10.1093/bioinformatics/bti448

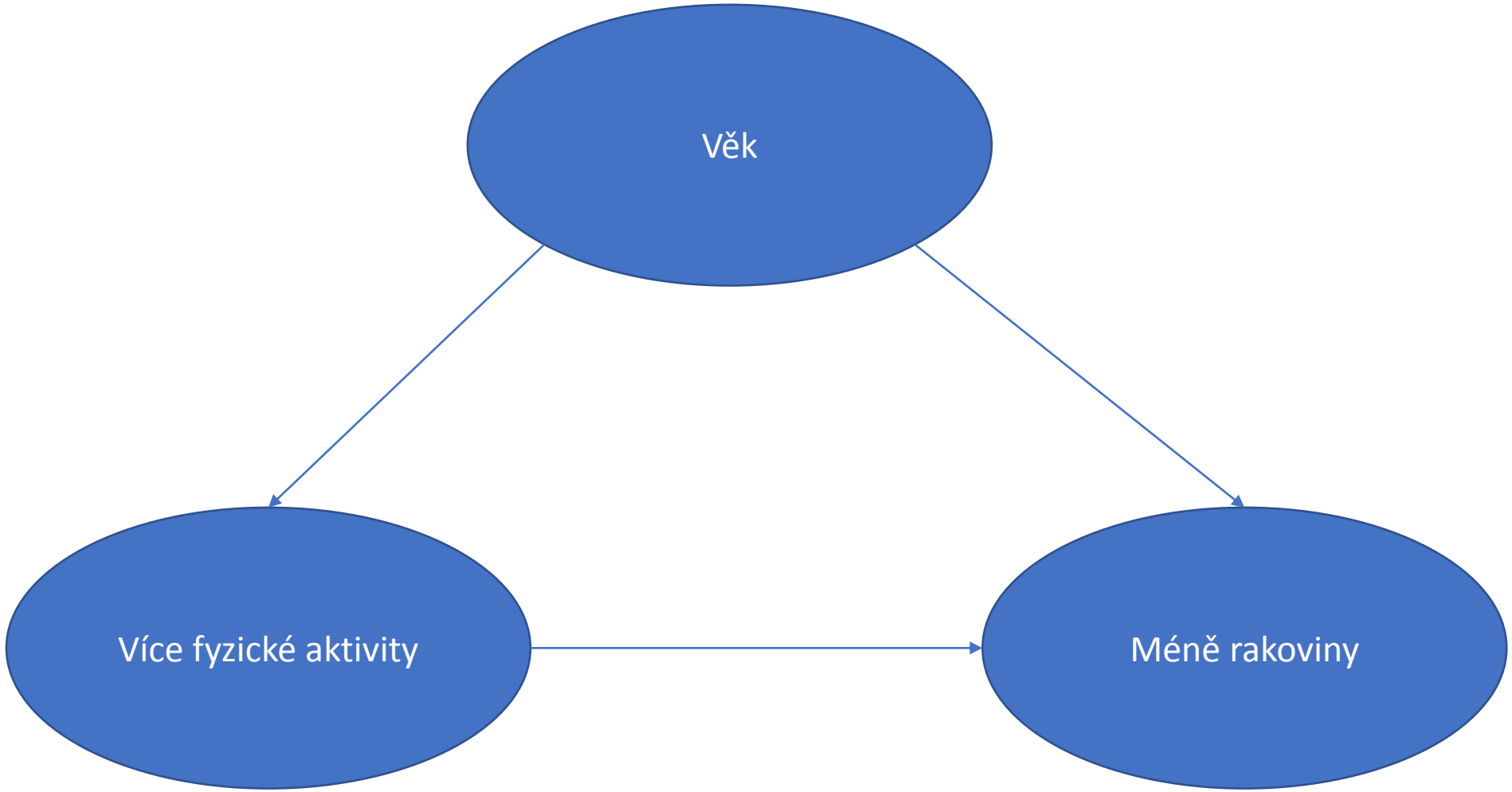
Za všechno mohou matoucí vlivy (confounding effects)?

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.

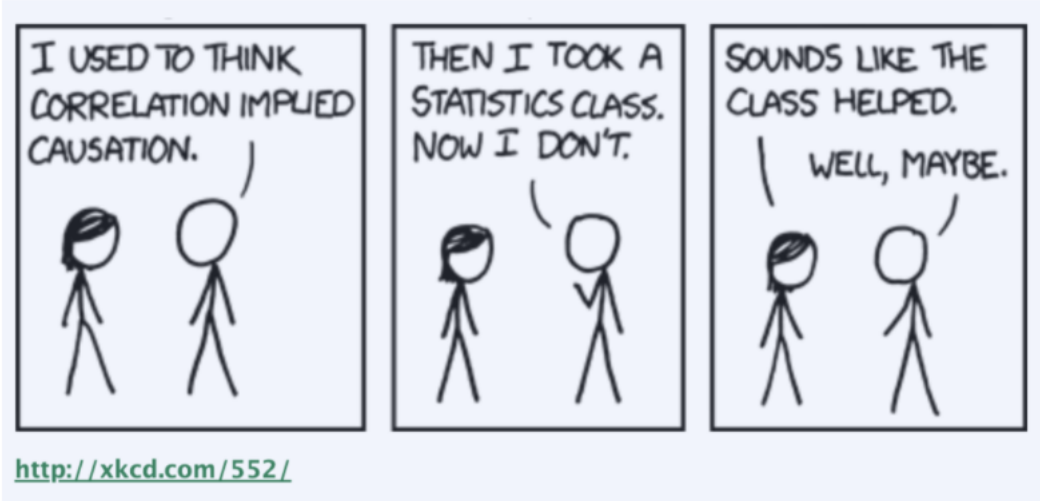
Jiným způsobem, vzniká korelace, která není kauzalita….

Matoucí vliv



Pochybné korelace….

https://www.tylervigen.com/spurious-correlations

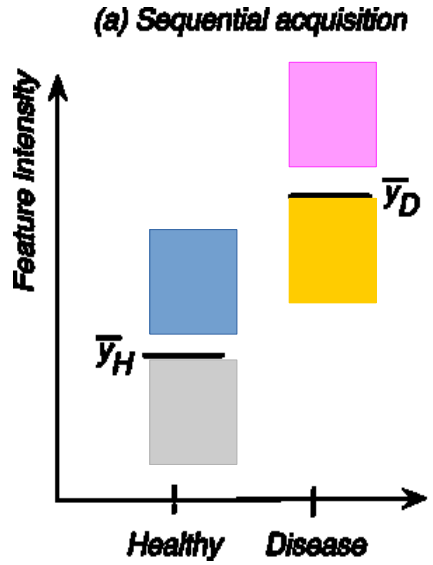


Efekt dávky

- 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



Pozorovaná proměnná (zdraví vs nemoc)
se překrývá s jinou technickou proměnnou, např:

1. a 2. den analýza zdravé tkáně
3. a 4. den analýza nádorové tkáně

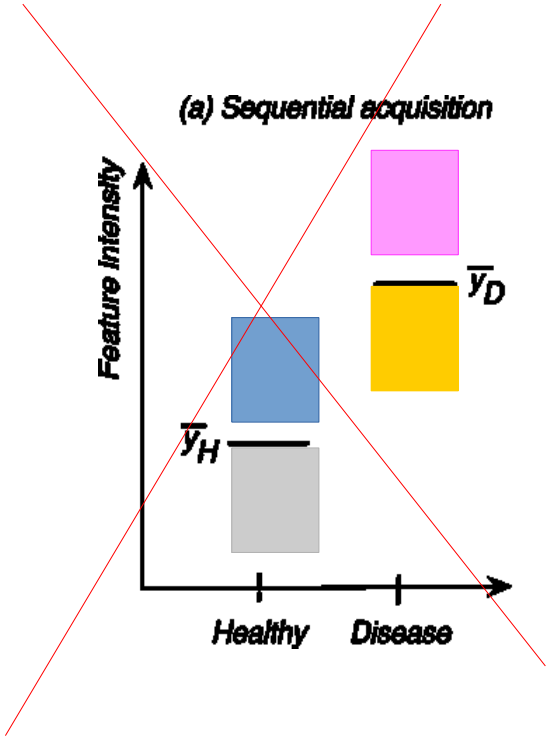
Nebo

Laborant 1 – zdravá tkáň, laborant 2 – nádorová tkáň

Nebo

Illumina primery (pro sekvenaci) 1-6 pro vzorky stolice,
Illumina primery 7-12 pro bukální stěry

Efekt dávky



Pozorovaná proměnná (zdraví vs nemoc)
se překrývá s jinou technickou proměnnou, např:

1. a 2. den analýza zdravé tkáně
3. a 4. den analýza nádorové tkáně

Nebo

Laborant 1 – zdravá tkáň, laborant 2 – nádorová tkáň

Nebo

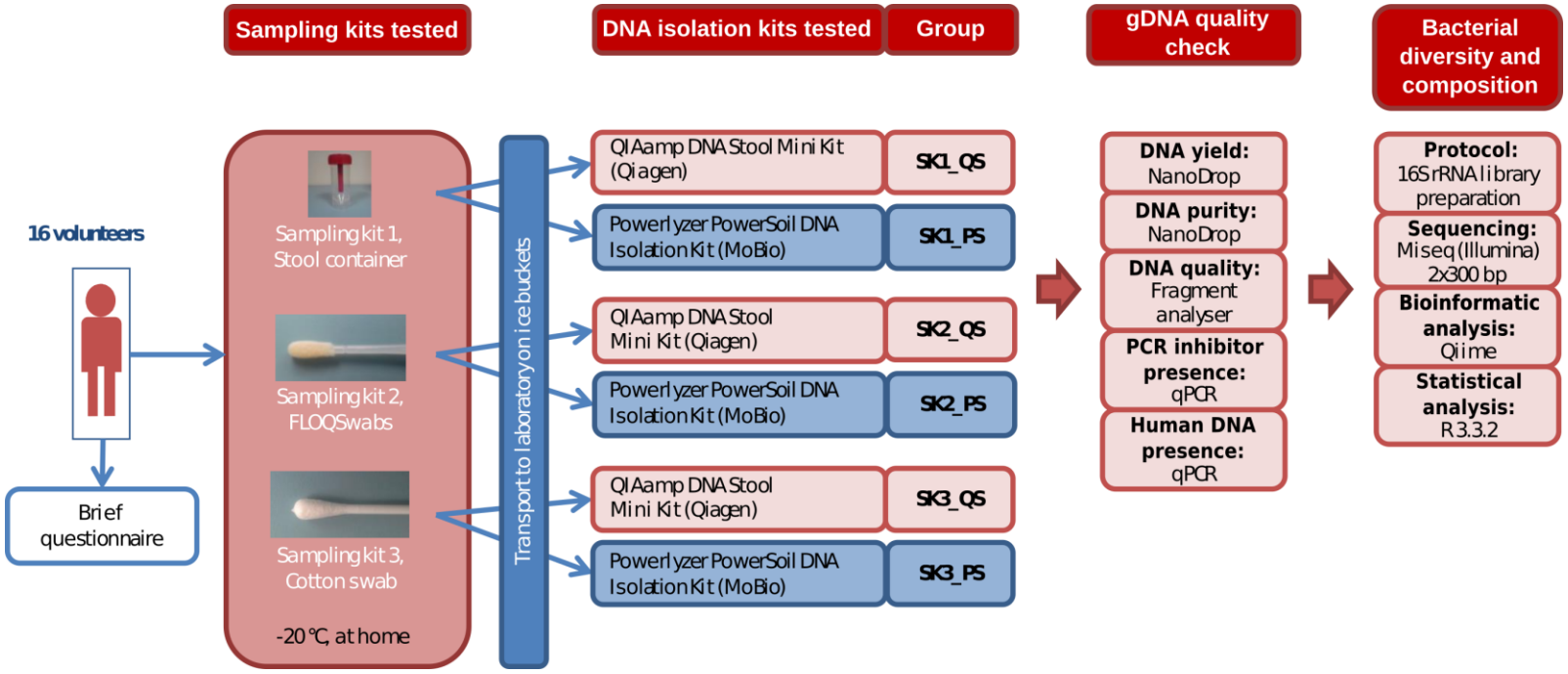
Illumina primery (pro sekvenaci) 1-6 pro vzorky stolice,
Illumina primery 7-12 pro bukální stěry

NENÍ MOŽNÉ STATISTICKY ODDĚLIT TECHNICKÝ EFEKT OD BIOLOGICKÉHO!!!

Příklady efektu dávky z praxe

Sekvencování mikrobiomu – efekt primeru Illumina

Experiment:
Sekvenace genu pro 16S rRNA
Cíl: Porovnat **vliv odběrových a izolačních kitů** na složení mikrobiomu ve stolici

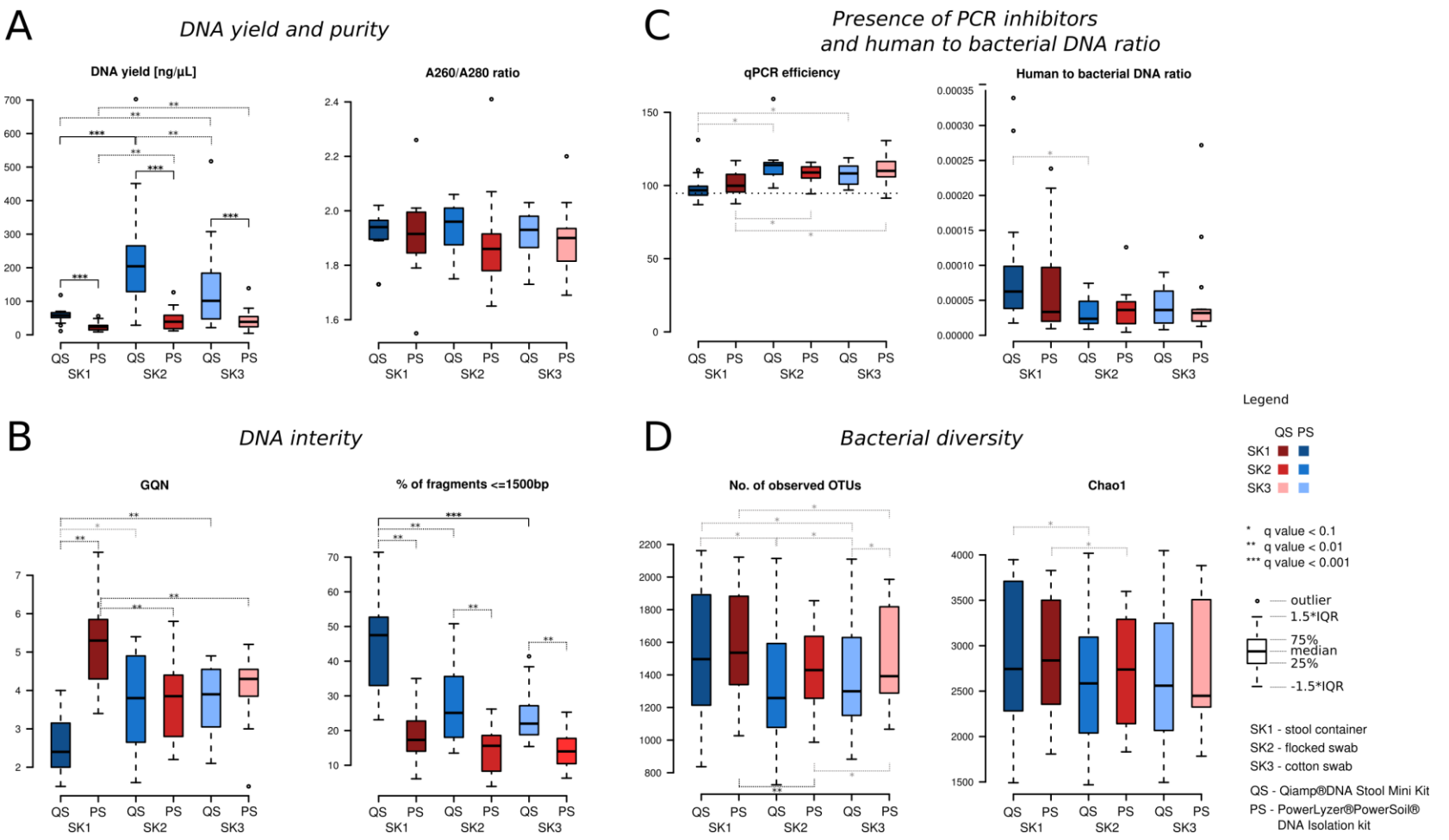


Porovnání 3 odběrových kitů (S1, S2, S3) a 2 DNA izolačních kitů (1,2)

16 dobrovolníků použilo všechny odběrové kity na odběr stolice,
z každého odběru izolace DNA dvěma kity

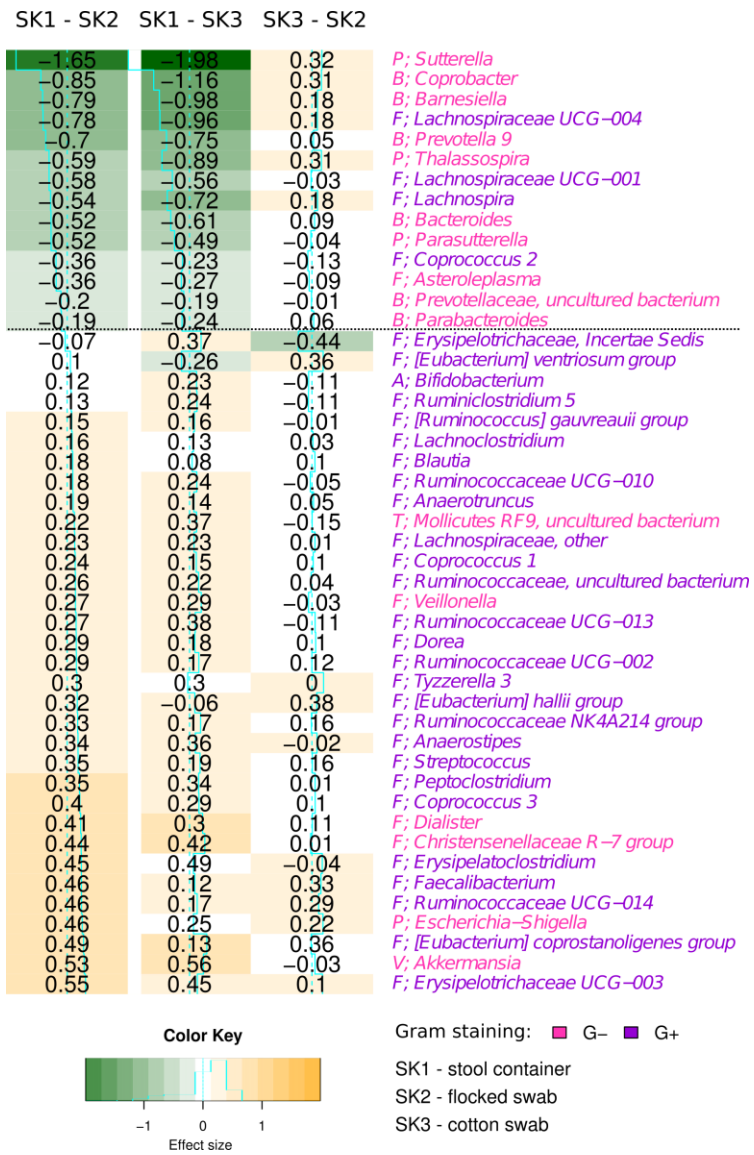
=> sekvenační analýza genu pro 16S rRNA

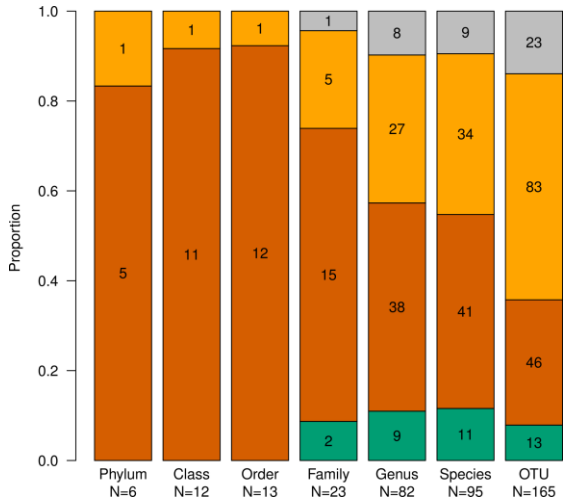
Experiment:
Sekvenace genu pro 16S rRNA
Cíl: Porovnat **vliv odběrových a izolačních kitů** na složení mikrobiomu ve stolici



Nalezen vliv odběrového a izolačního kitu na kvalitu a kvantitu DNA a také na složení mikrobiomu!

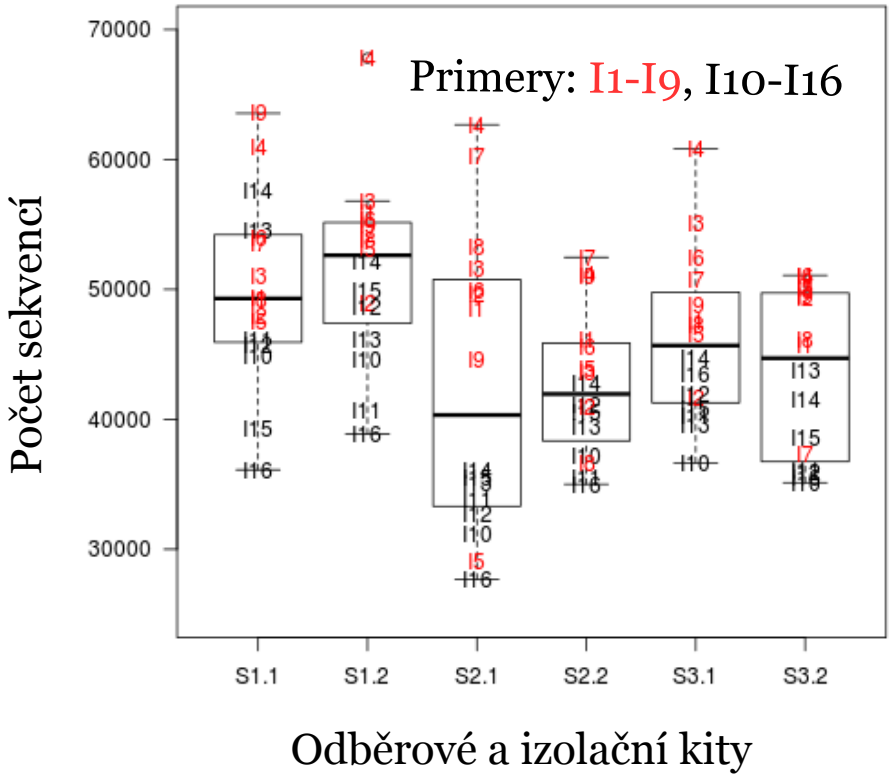
Experiment:
Sekvenace genu pro 16S rRNA
Cíl: Porovnat **vliv odběrových a izolačních kitů** na složení mikrobiomu ve stolici





Nalezen vliv odběrového a izolačního kitu na kvalitu a kvantitu DNA a také na složení mikrobiomu!

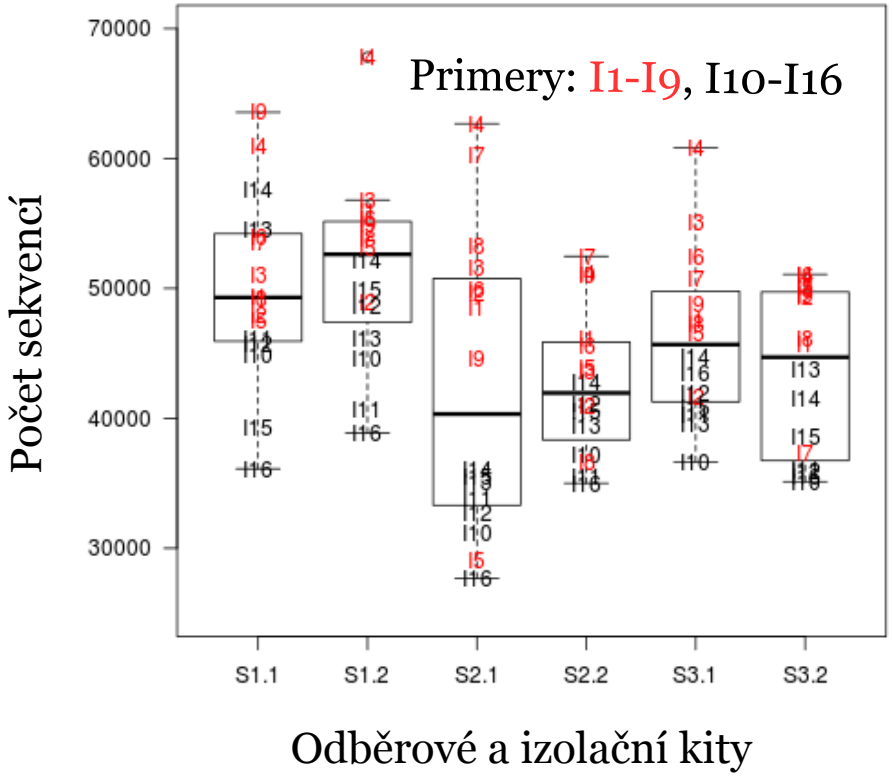
Experiment:
Sekvenace genu pro 16S rRNA
Cíl: Porovnat **vliv odběrových a izolačních kitů** na složení mikrobiomu ve stolici



Každý účastník měl vždy stejný primer.

Počet sekvencí je statisticky významně vyšší u primerů I1-I9 v porovnání s primery I10-I16!!!

Experiment:
Sekvenace genu pro 16S rRNA
Cíl: Porovnat **vliv odběrových a izolačních kitů** na složení mikrobiomu ve stolici

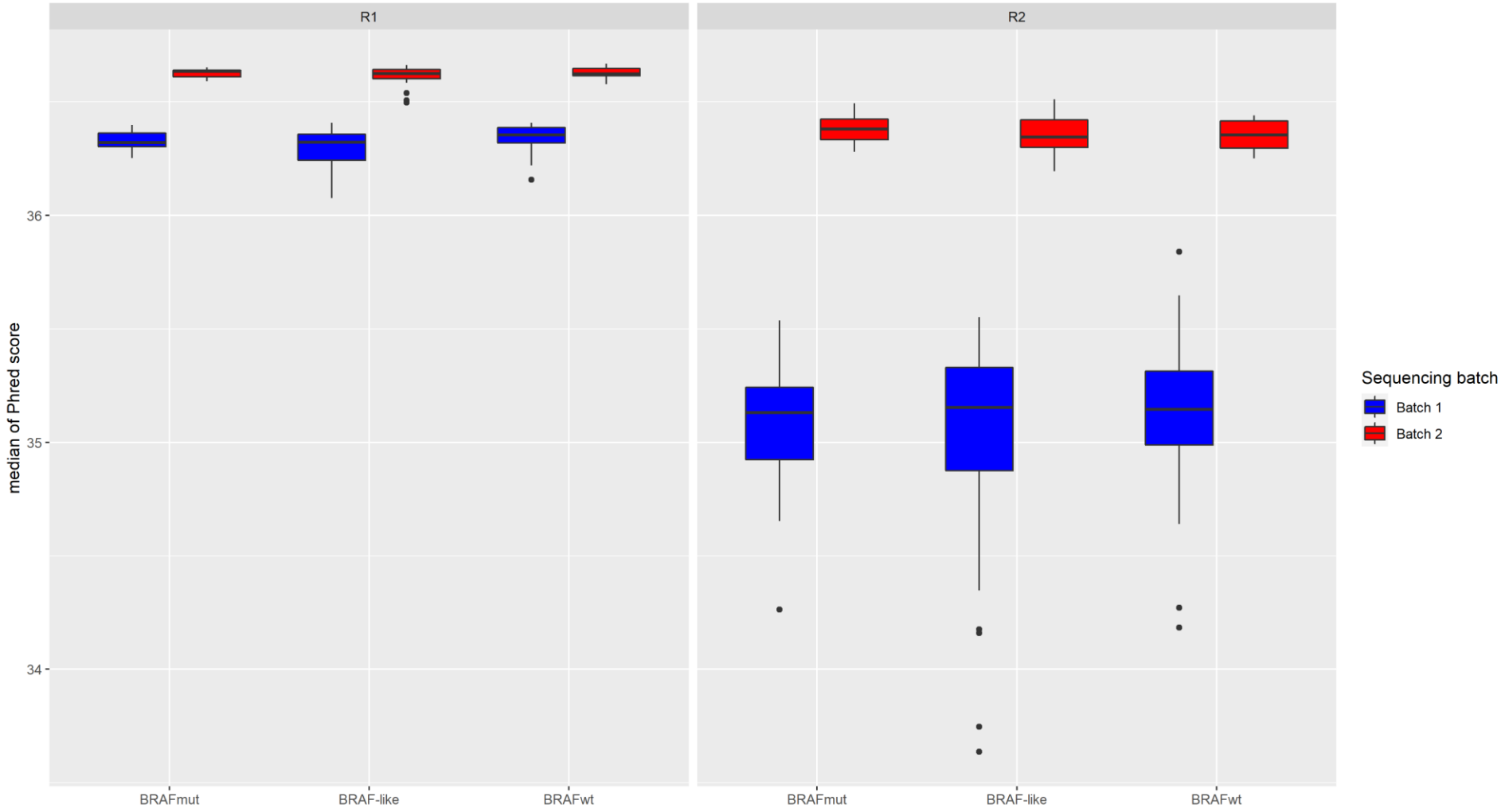


PROBLÉM: primer může mít efekt na složení mikrobiomu

ŘEŠENÍ: primer (nebo lépe řečeno skupina I1-I9 vs I10-I16) jako nová proměnná ve statistické analýze, odhad efektu skupiny primerů:

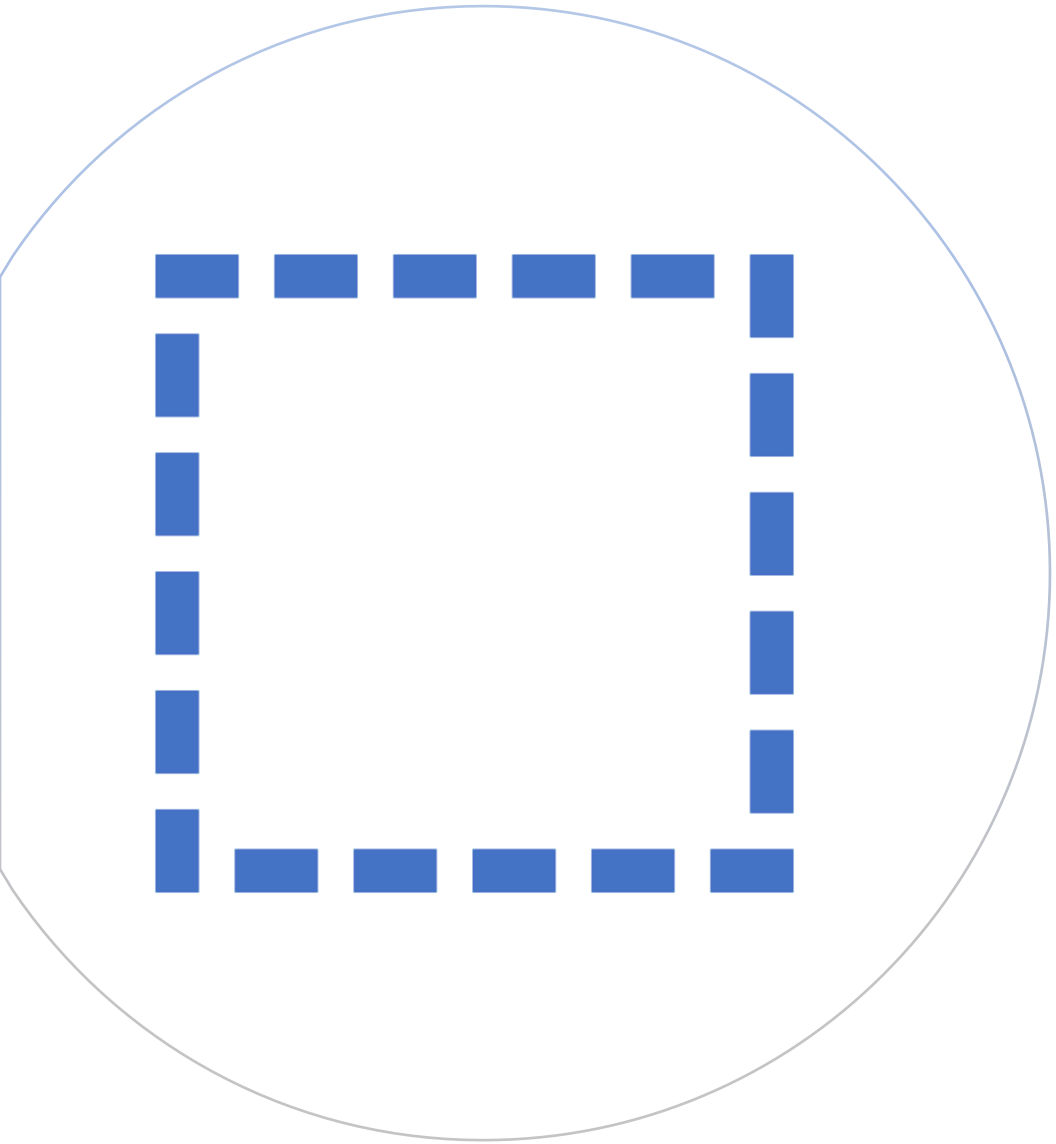
VÝSLEDEK: zdá se, že primer ovlivňuje pouze počet sekvencí, ne složení mikrobiomu (?).

Illumina sekvencování RNAseq kolorektálního karcinomu ve 2 dávkách



- Kvalita čtení se výrazně liší mezi dávkami

Mikrobiální kontaminace v NGS



Mikrobiální kontaminace

- Velký problém zejména u metagenomických studií a u vzorků s nízkým obsahem bakteriální DNA

Figure 1. The contents of non-aligning reads from 57 human whole genome sequencing runs.

Baylor College of Medicine (BCM),
the Broad Institute (BI),
Illumina (ILLUM),
the Max Planck Institute for Molecular Genetics (MPIMG),
the Sanger Center (SC),
Washington University Genome Sequencing Center (WUGSC).

Center	Run	All read pairs									
		Primate, EBV, phage or discarded due to low quality/ duplicate/ low entropy	Contamination candidates								
			Known contaminants (high homology - ≥90%)								Low homology (<90%)
			Eukaryote	Prokaryote					Viral	Dual homology	
				Ultrapure water system contaminants			Enterobacteriaceae	Others & dual homology			
				Bradyrhizobium		Others & dual homology					
				DFCI-1	Others & dual homology						
BCM	SRR385754	50112167	9	438	433	431	0	172	2	61	543
BCM	SRR385755	43233137	1	227	179	234	0	88	3	23	278
BCM	SRR385758	48859661	12	0	0	1	6	32	0	66	61
BCM	SRR385759	79360966	9	1153	1126	1514	0	517	6	190	1736
BCM	SRR385761	52277896	8	0	0	2	4	14	0	84	209
BCM	SRR385762	103886262	164	0	0	1	1	127	0	5	278
BCM	SRR385763	53807301	9	321	282	363	0	120	1	53	411
BCM	SRR385764	46462143	6	149	145	169	0	65	0	20	277
BCM	SRR385765	43367799	18	625	623	792	171	298	4	4015	1055
BCM	SRR385767	63753008	44	0	0	3	0	27	0	131	747
BCM	SRR385768	46642864	28	0	0	2	1	18	0	62	470
BCM	SRR385769	92608130	21	0	0	22	0	222	0	112	361
BCM	SRR385770	36818041	7	1	4	6	0	2	0	1	4
BCM	SRR385772	60254401	9	7	10	4	16	21	1	460	23
BCM	SRR385773	52061042	0	30	76	18	0	10	7	16	38
BCM	SRR385774	42439279	2	28	41	14	14	22	1	372	39
BCM	SRR385776	42169205	3	35	61	27	0	11	0	3	50
BCM	SRR385777	48462601	4	34	54	24	0	24	0	0	43
BCM	SRR393988	58072454	1	145	102	156	0	57	1	14	211
BCM	SRR393989	49568344	1	67	65	63	0	35	0	13	102
BCM	SRR393990	51017564	34	0	0	4	0	34	0	7	561
BCM	SRR393993	48369009	82	0	0	0	5	13	0	90	67
BCM	SRR393994	56927949	25	2	0	2	0	26	0	5	436
BCM	SRR400037	47847795	16	0	0	21	8	87	0	153	92
BCM	SRR741366	74950213	22	56	40	1	15	36	0	267	56
BCM	SRR768303	69790675	9	18	19	0	2	27	0	23	12
BCM	SRR768304	51864089	15	12	11	2	8	15	2	88	14
BCM	SRR768309	80919926	13	24	23	4	5	18	0	127	24
BI	SRR067576	18153122	21	0	0	15	35	7	0	639	280
BI	SRR067577	46251353	37	0	0	0	0	5	0	0	40
BI	SRR067578	46490939	49	0	0	0	0	4	0	1	53
BI	SRR067579	46015083	27	0	0	0	0	6	0	1	51
BI	SRR068130	125872967	19	0	2	1	0	139	0	106	47
BI	SRR075005	29392952	7	0	1	1	0	43	0	44	17
BI	SRR075006	62188209	496	5	5	0	0	3	0	5	78
ILLUM	ERR091571	211437693	95	0	0	2	1	58	0	6	64
ILLUM	ERR091575	205185449	5	0	0	42	1	21	0	20	103
MPIMG	ERR233225	104620504	0	0	0	4	1	1	0	1	8
MPIMG	ERR233227	106377916	638	0	0	2	5	7	1	62	27
MPIMG	ERR233301	119475893	28	0	0	0	0	3	0	3	18
MPIMG	ERR233302	111852187	63	0	0	54	0	154	0	10	219
MPIMG	ERR234321	108573882	725	0	0	1	0	0	2	4	14
MPIMG	ERR234322	113157820	473	0	0	1	1	12	0	5	30
MPIMG	ERR234323	114214873	13	0	0	0	0	17	0	1	9
MPIMG	ERR234324	118549617	1235	0	0	0	0	2	0	0	21
MPIMG	ERR234325	111531797	386	0	0	0	0	4	0	0	32
MPIMG	ERR234327	112766941	530	0	0	1	7	12	0	16	75
MPIMG	ERR234328	114857252	2388	0	0	1	1	2	0	0	128
MPIMG	ERR234329	111235200	188	0	0	0	0	4	0	2	67
MPIMG	ERR239333	116722187	892	0	0	3	0	3	0	9	50
MPIMG	ERR239334	105779573	577	0	1	0	0	10	0	8	362
SC	ERR050082	42242519	89	0	1	7	0	20	0	6	2432
SC	ERR050083	66388415	61	1	0	4	0	18	0	1422	1146
WUGSC	SRR211275	43009432	50	0	0	0	0	93	2	4	59
WUGSC	SRR211278	55193260	23	0	0	0	1	42	0	7	39
WUGSC	SRR407429	41694818	6	0	0	1	1	23	4	27	13
WUGSC	SRR407508	42586738	240	0	0	1	2	87	2	42	33

"Abychom posoudili rozsah a rozmanitost sekvenačních kontaminantů, provedli jsme mapování 57 sekvenčních běhů z projektu „1000 Genomes Project“ ze šesti center proti čtyřem největším databázím NCBI BLAST. Detekovali jsme čtení různých druhů kontaminantů ve všech bězích a identifikovali nejběžnější z těchto rodů kontaminantů (Bradyrhizobium) v sestavených genomech z databáze genomu NCBI. "

Laurence M, Hatzis C, Brash DE (2014) Common Contaminants in Next-Generation Sequencing That Hinder Discovery of Low-Abundance Microbes. PLOS ONE 9(5): e97876. https://doi.org/10.1371/journal.pone.0097876
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0097876

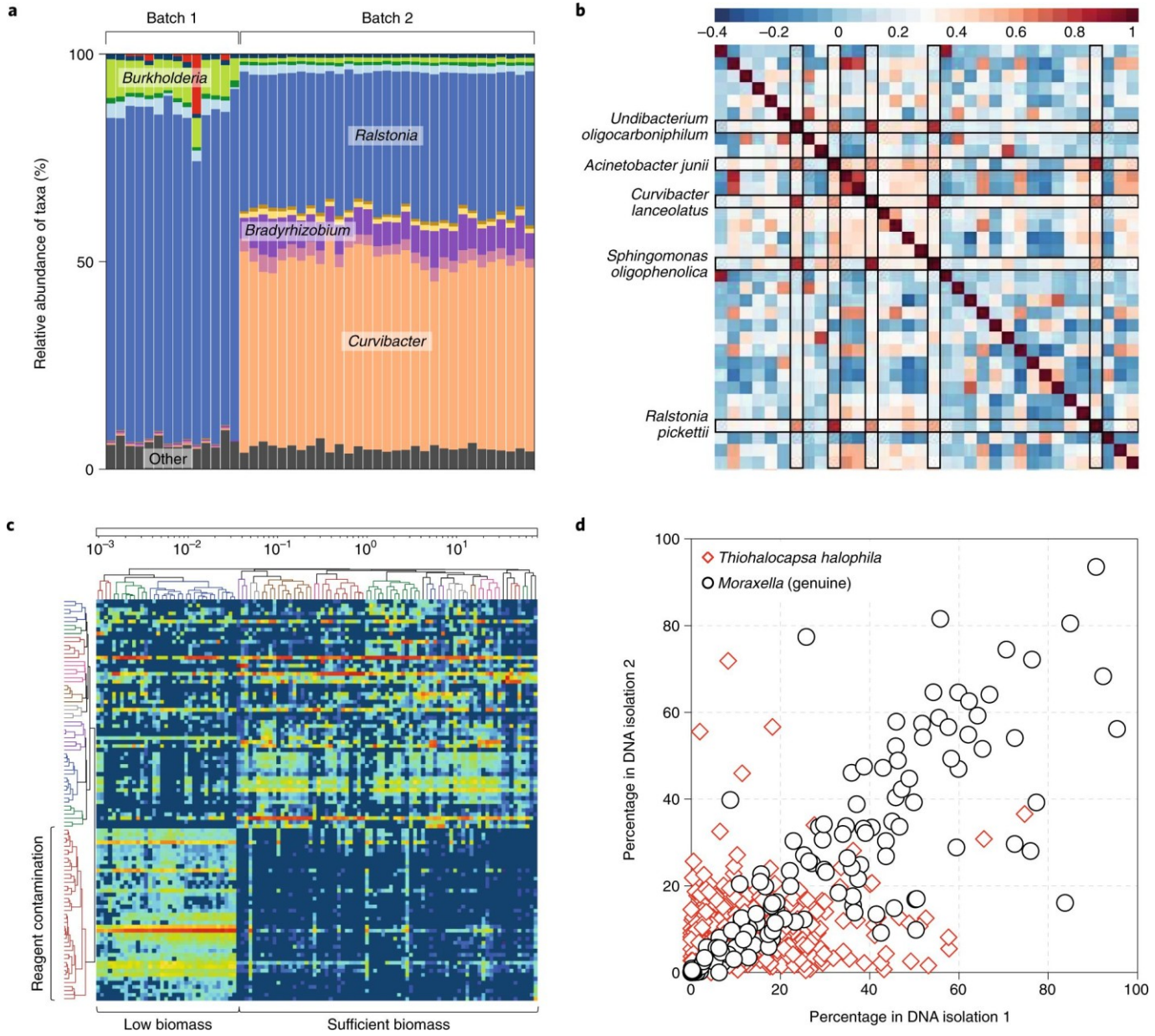


Fig. 1: Reagent contamination recognition strategies.

a, Between-batch variation allows for rapid identification of reagent contamination. This example is from a 16S analysis of placental tissues… FastDNA SPIN kits with different lot numbers were used for batches 1 and 2.

…

b, Spearman's rho correlation coefficient heatmap of a subset of the most common species detected (*x*- and *y*-axes) during a study of necrotizing enterocolitis in pre-term infants

…

c, Reagent contaminants are especially abundant in samples with low biomass that failed 16S amplification, and in negative controls; both of which cluster together in the lower left corner. This dataset is from a study where bacterial DNA was enriched from nasal swabs and sequenced with an ILLUMINA HiSeq v4 sequencing kit.

…

d, **Genuine signals are reproducible** and separate measurements from the same sample using different DNA isolation kits should correlate with one another while reagent contamination signals do not. The genuine *Moraxella* signal is from a reanalysis of the 16S data of Salter et al.[1], whereas the reagent contamination example, *Thiohalocapsa halophila*, is from an analysis of placental tissues.

de Goffau, MC; Lager, S; Salter, SJ; Wagner, J; Kronbichler, A; Charnock-Jones, DS; Peacock, SJ; Smith, GCS; Parkhill, J; (2018) Recognizing the reagent microbiome. Nature microbiology, 3 (8). pp. 851-853. ISSN 2058-5276 DOI: https://doi.org/10.1038/s41564-018-0202-y

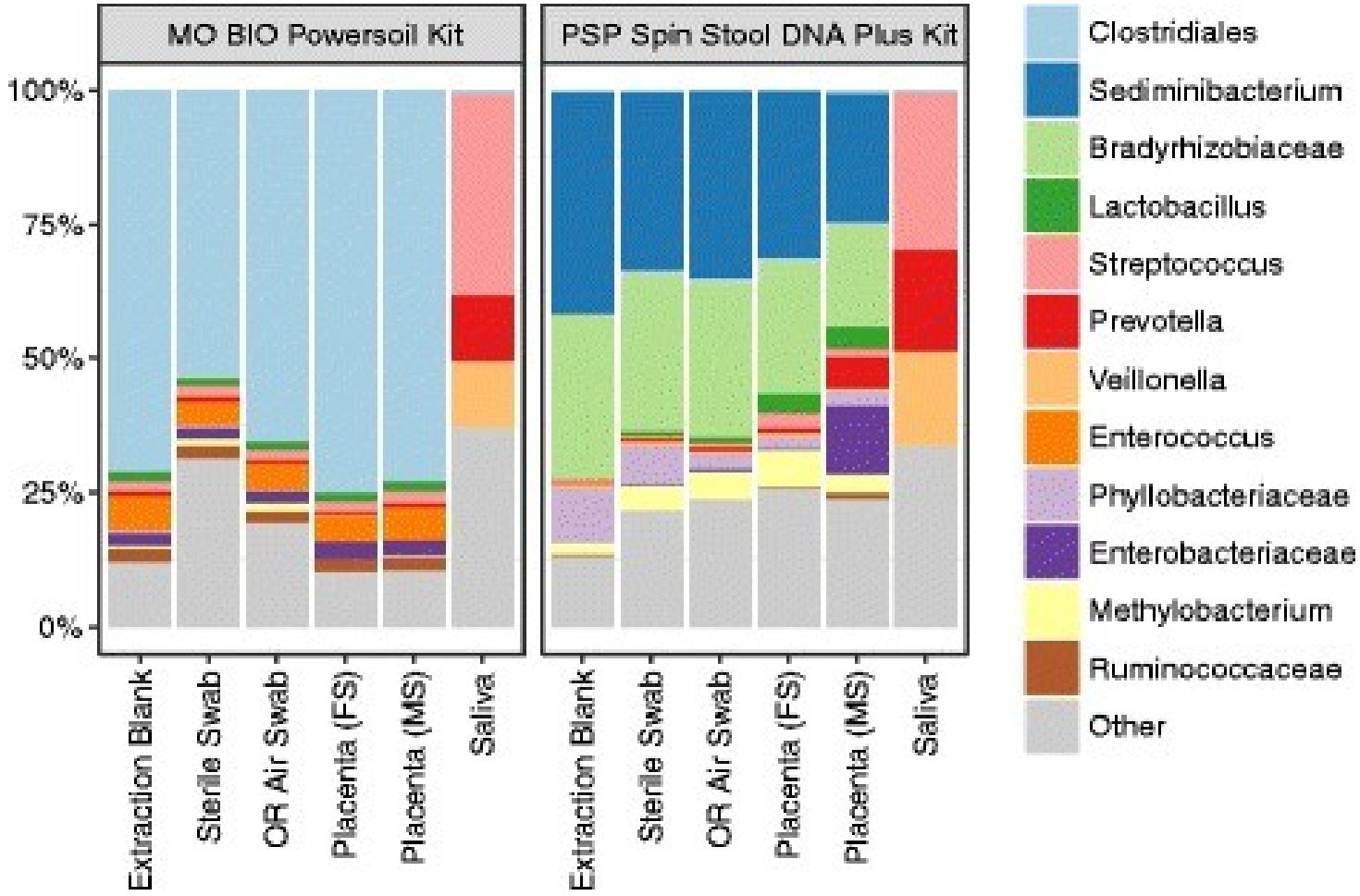


Fig. 3 Wrestling with kit contamination—similar bacterial composition in placental samples and negative controls.

Kim D, Hofstaedter CE, Zhao C, Mattei L, Tanes C, Clarke E, Lauder A, Sherrill-Mix S, Chehoud C, Kelsen J, et al. Optimizing methods and dodging pitfalls in microbiome research. Microbiome. 2017;5(1):52.

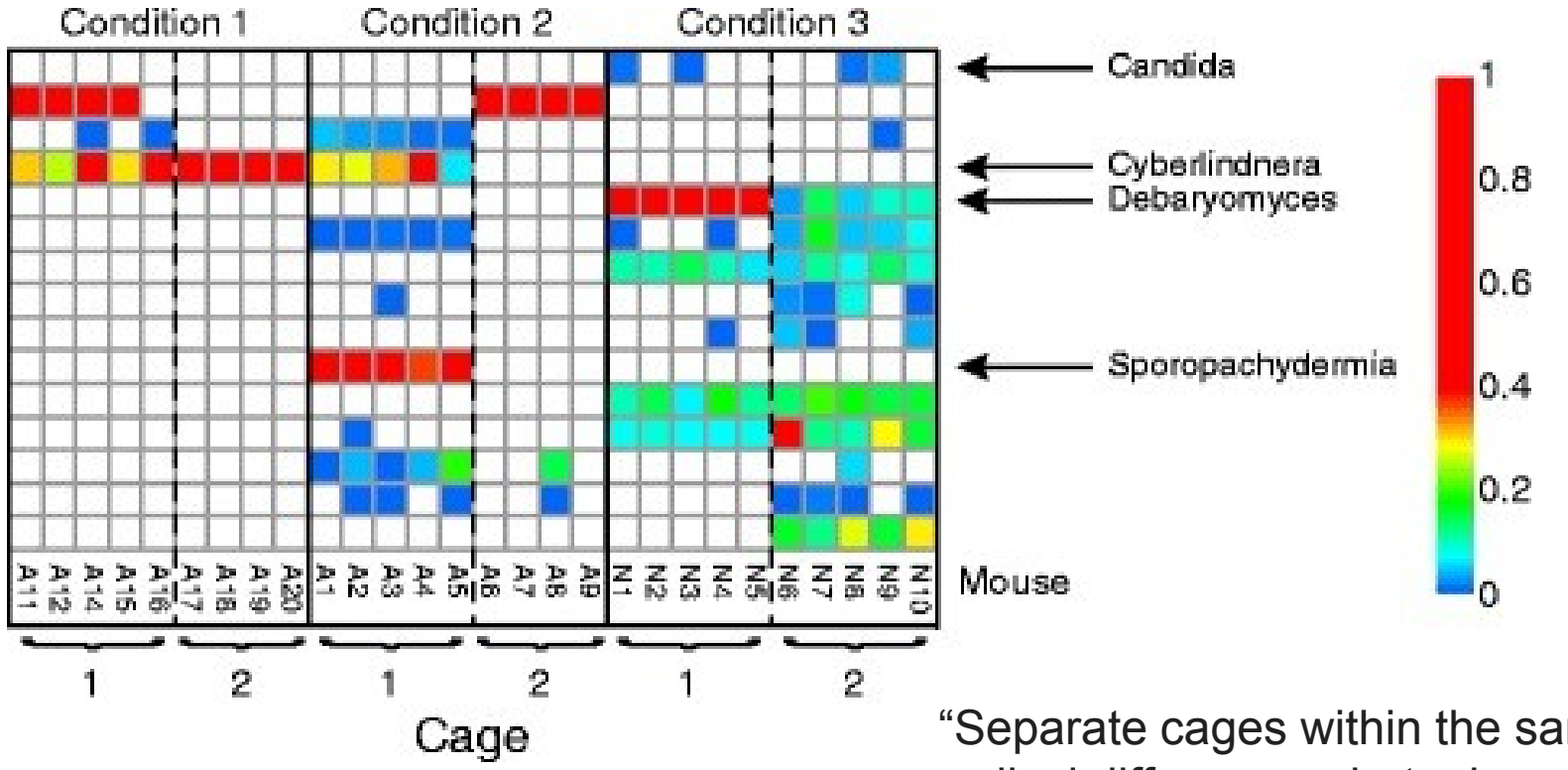


Fig. 1 Example of cage effects dominating a mouse study of fungal communities.

...

The three conditions studied were continuous exposure to antibiotics (*Condition 1*), short-term exposure to antibiotics (*Condition 2*), and no exposure to antibiotics (*Condition 3*).

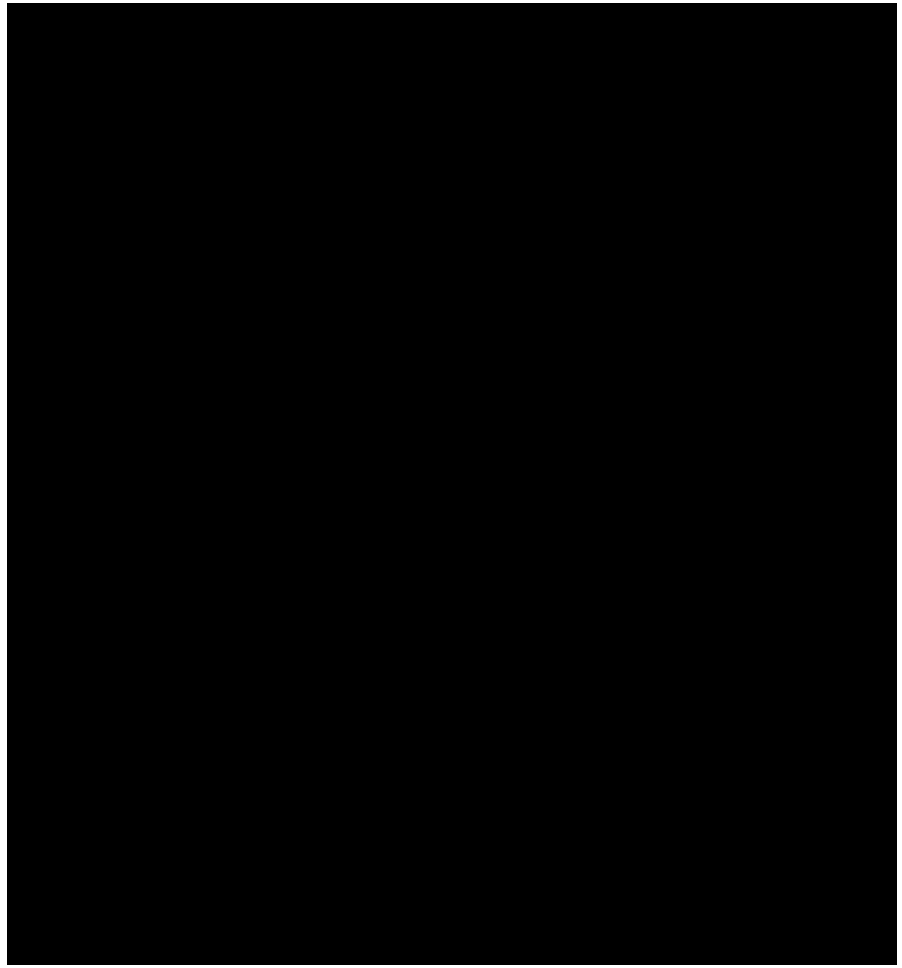
"Separate cages within the same treatment group showed radical differences, but mice within a cage generally behaved similarly"

Dollive S, Chen YY, Grunberg S, Bittinger K, Hoffmann C, Vandivier L, Bushman FD. Fungi of the murine gut: episodic variation and proliferation during antibiotic treatment. PLoS One. 2013;8(8):e71806.

Efekt dávky - platforma

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

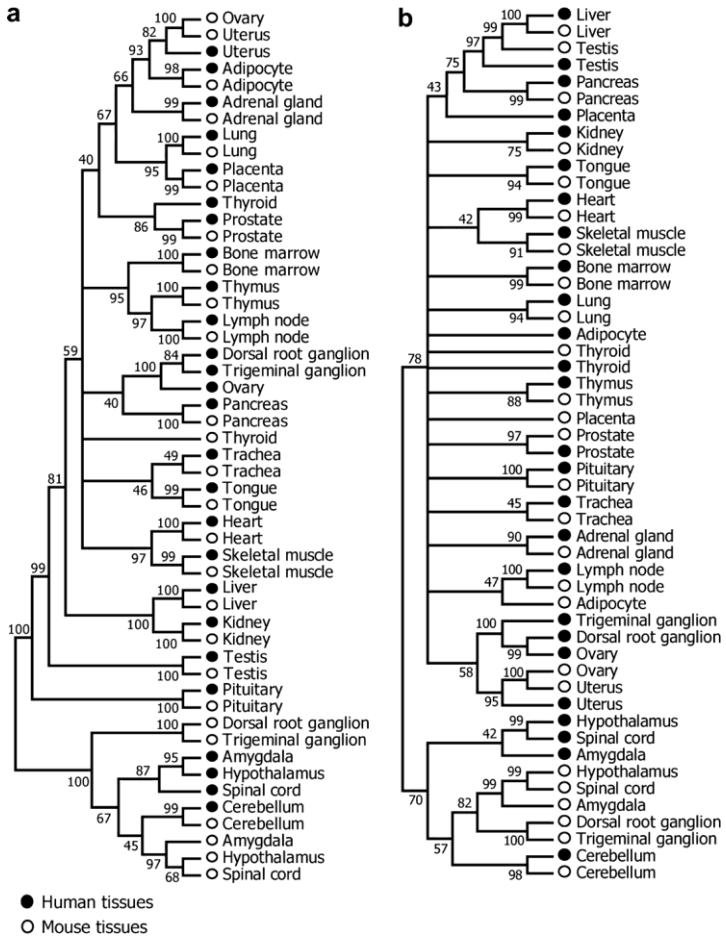


FIG. 5.—
Dendrograms of 26 human and 26 mouse tissues based on (a) 1 − Pearson's correlation coefficient r and (b) Euclidean distance d of tissues..

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

Lidé a myši na RNAseq

Navzdory tomu se problém v roce 2014 opakuje!!

PNAS

Comparison of the transcriptional landscapes between human and mouse tissues

Shin Lin[a,b,1], Yiing Lin[c,1], Joseph R. Nery[d], Mark A. Urich[d], Alessandra Breschi[e,f], Carrie A. Davis[g], Alexander Dobin[g], Christopher Zaleski[g], Michael A. Beer[h], William C. Chapman[c], Thomas R. Gingeras[g,i], Joseph R. Ecker[d,j,2], and Michael P. Snyder[a,2]

[a]Department of Genetics, Stanford University, Stanford, CA 94305; [b]Division of Cardiovascular Medicine, Stanford University, Stanford, CA 94305; [c]Department of Surgery, Washington University School of Medicine, St. Louis, MO 63110; [d]Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037; [e]Centre for Genomic Regulation and UPF, Catalonia, 08003 Barcelona, Spain; [f]Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, 08003 Barcelona, Spain; [g]Functional Genomics, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11742; [h]McKusick-Nathans Institute of Genetic Medicine and the Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21205; [i]Affymetrix, Inc., Santa Clara, CA 95051; and [j]Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, CA 92037

Contributed by Joseph R. Ecker, July 23, 2014 (sent for review May 23, 2014)

Lidé a myši na RNAseq

„V této studii velkého počtu tkání mezi lidmi a myšmi odhalila vysoce výkonná transkriptomická a epigenomická sekvenace, že obecně dominují rozdíly mezi těmito dvěma druhy.“

Tentokrát byla RNAseq použita pro oba druhy, a proto to vypadalo, že není žádný problém s rozdílnou platformou….

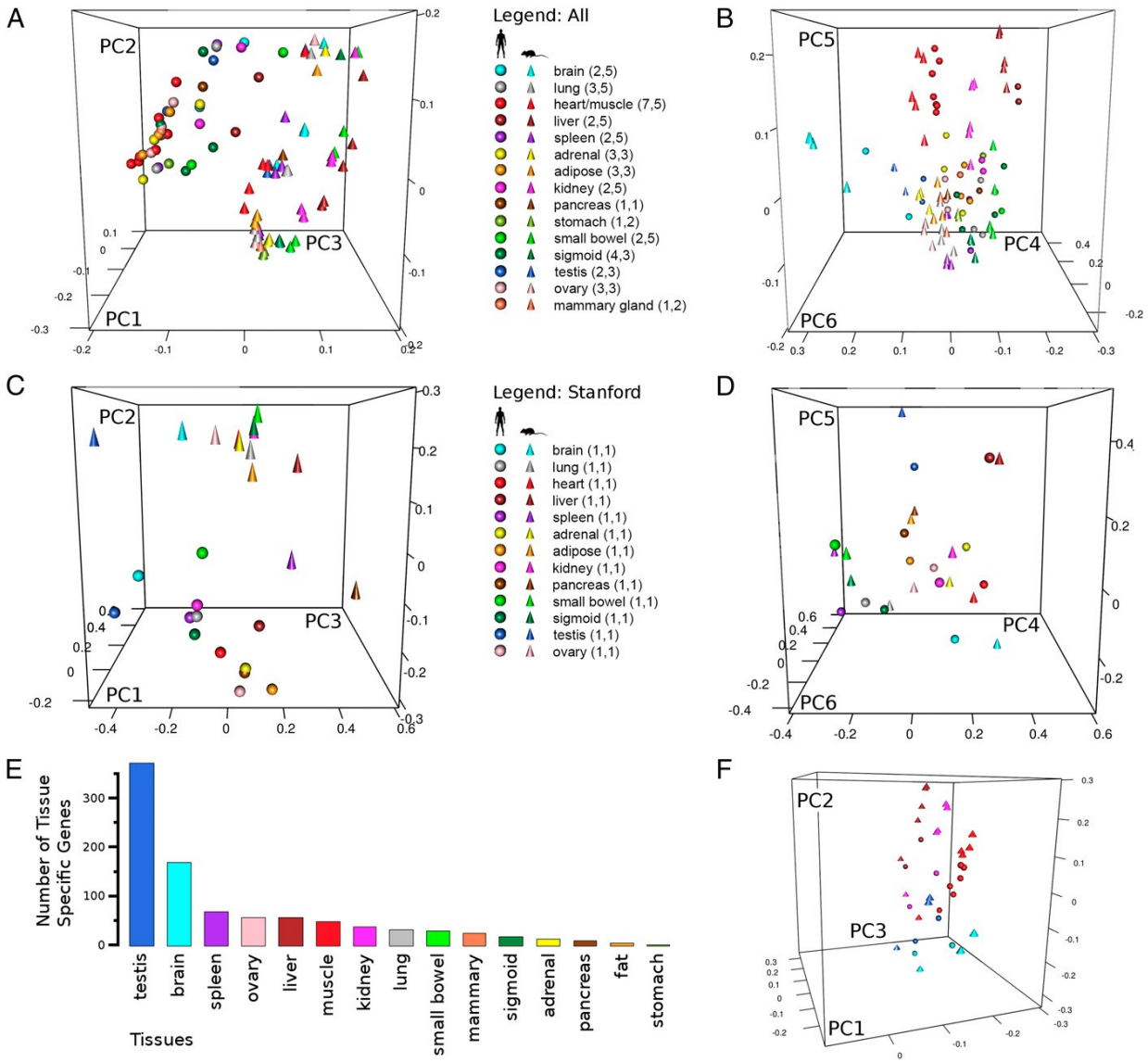


Fig. 1. Loading plots from PCA on human and mouse gene expression data.

Lidé a myši na RNAseq

- Následná reanalýza z roku 2015 ukázala, že rozdíly jsou pravděpodobně způsobeny efektem dávky flow cell a ranu!

F1000Research

F1000Research 2015, 4:121 Last updated: 16 MAY 2019

Check for updates

RESEARCH ARTICLE

A reanalysis of mouse ENCODE comparative gene expression data [version 1; peer review: 3 approved, 1 approved with reservations]

Yoav Gilad, Orna Mizrahi-Man

Department of Human Genetics, University of Chicago, Chicago, IL, 60637, USA

D87PMJN1 (run 253, flow cell D2GUAACXX, lane 7)	D87PMJN1 (run 253, flow cell D2GUAACXX , lane 8)	D4LHBFN1 (run 276, flow cell C2HKJACXX , lane 4)	MONK (run 312, flow cell C2GR3ACXX , lane 6)	HWI-ST373 (run 375, flow cell C3172ACXX , lane 7)
heart	adipose	adipose	heart	brain
kidney	adrenal	adrenal	kidney	pancreas
liver	sigmoid colon	sigmoid colon	liver	brain
small bowel	lung	lung	small bowel	spleen
spleen	ovary	ovary	testis	Human
testis		pancreas		Mouse

Figure 1. Study design.
Sequencing batches as inferred based on the sequence identifiers of the RNA-Seq reads

Lidé a myši na RNAseq

… po korekci efektu dávky to vypadá tak jak má

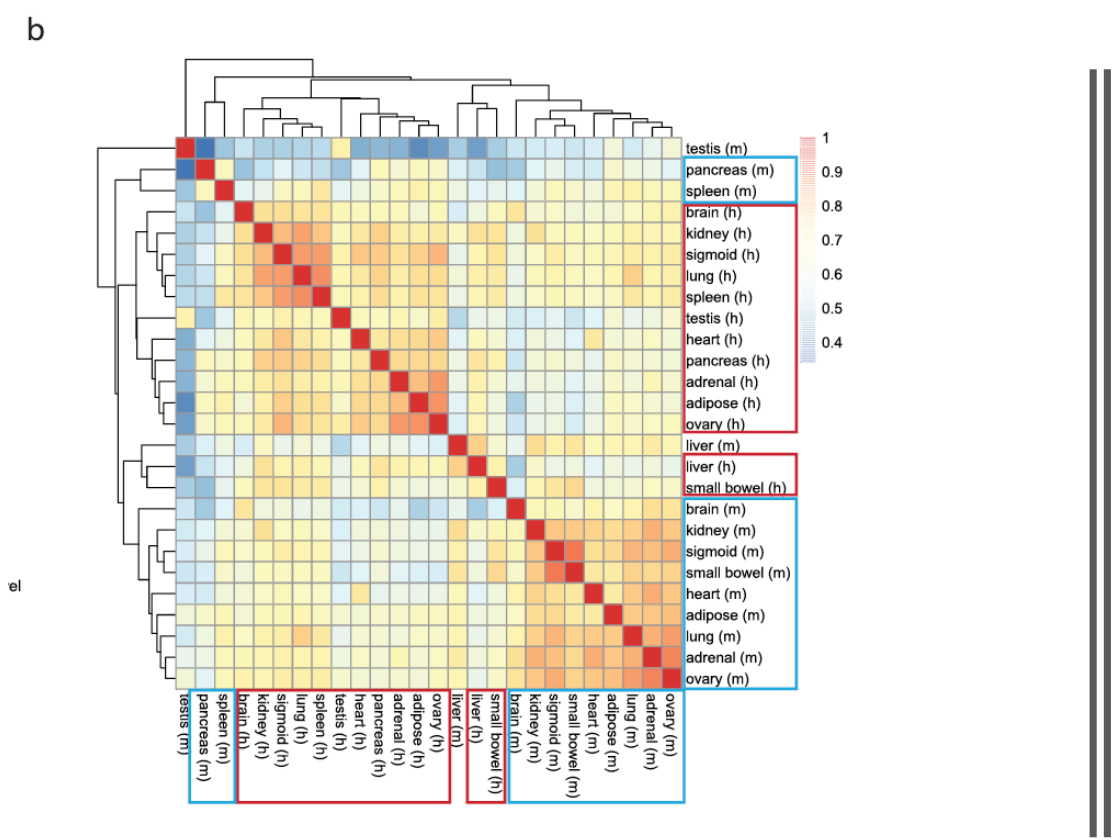


Figure 2. Recapitulating the patterns reported by the mouse ENCODE papers.

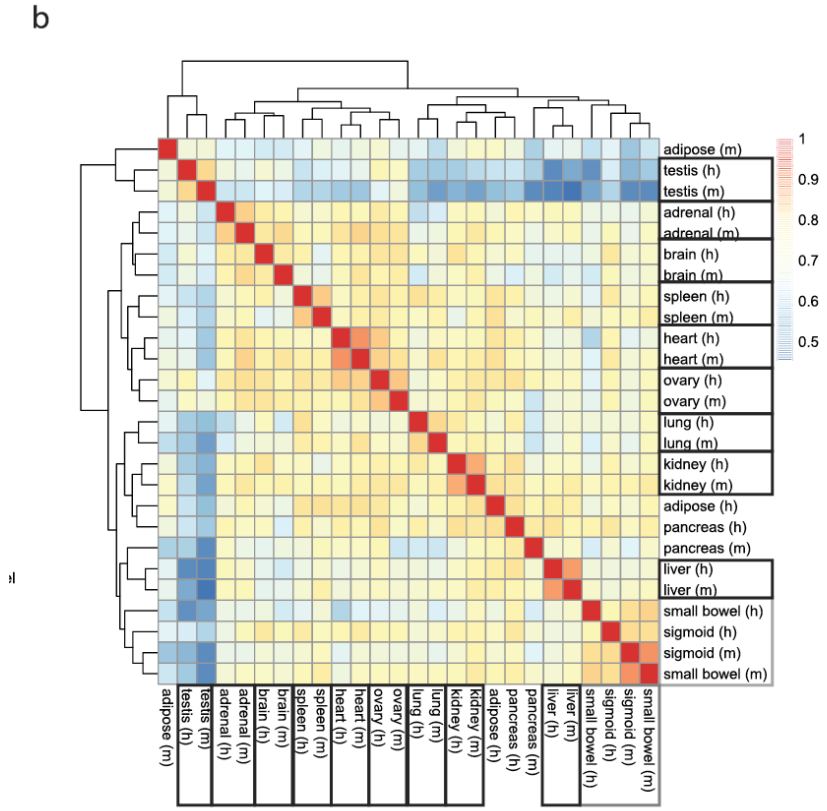


Figure 3. Clustering of data once batch effects are accounted for

Lidé a myši na RNAseq

Ovšem pozor, v čem je problém?

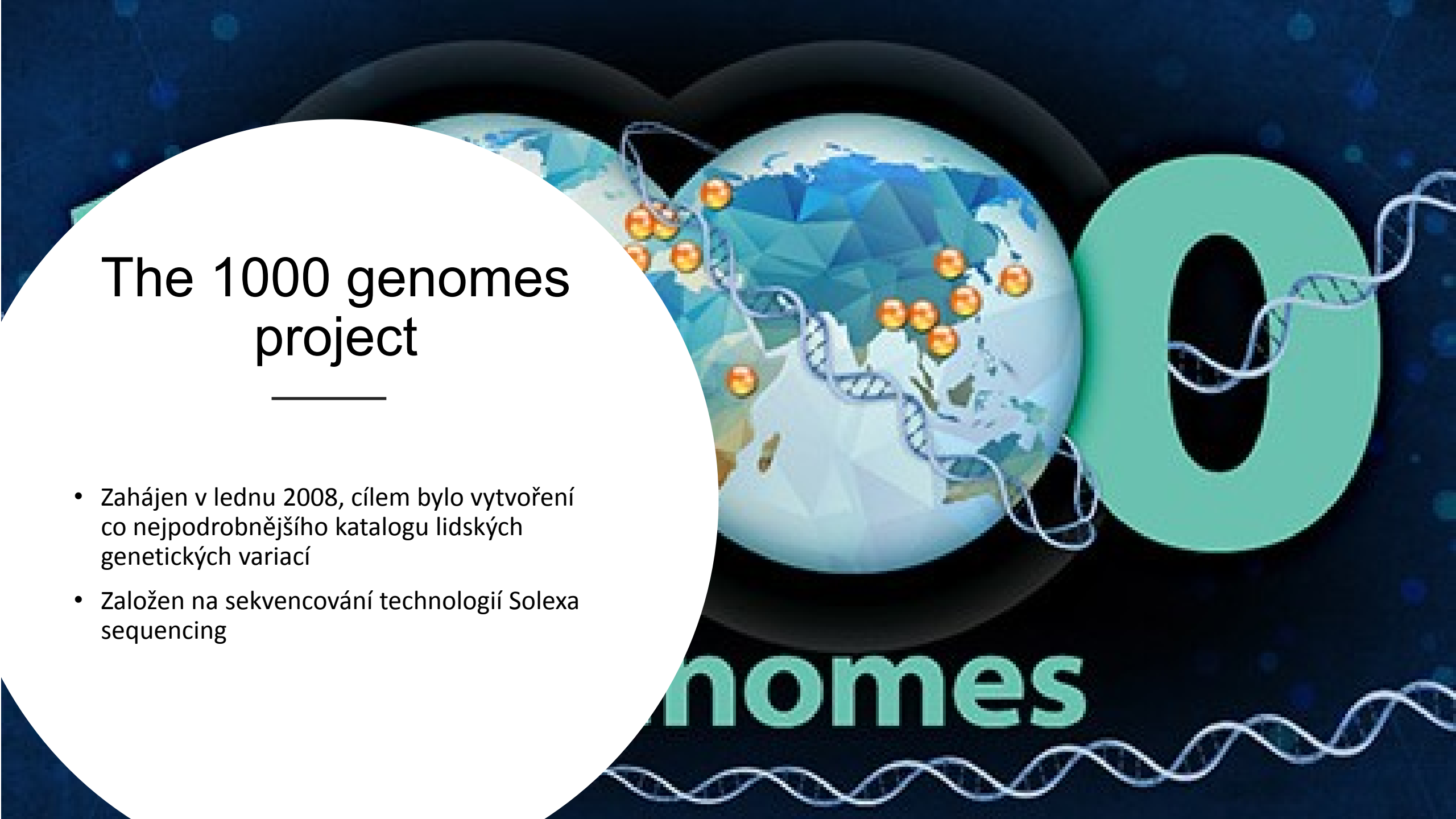
Protože šlo v tomto případě o téměř perfektní batch efekt – tedy téměř 100% překryv efektu lane a ranu vs organizmus, odstraněné rozdíly batch efektu mohou být také ty biologické.

Jinak řečeno - tyto data nemohou odpovědět na otázku která byla položena.

Doporučuji diskuzi pod článkem z F1000research…

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.

Ani jeden z těchto článků nebyl stažen z tisku….

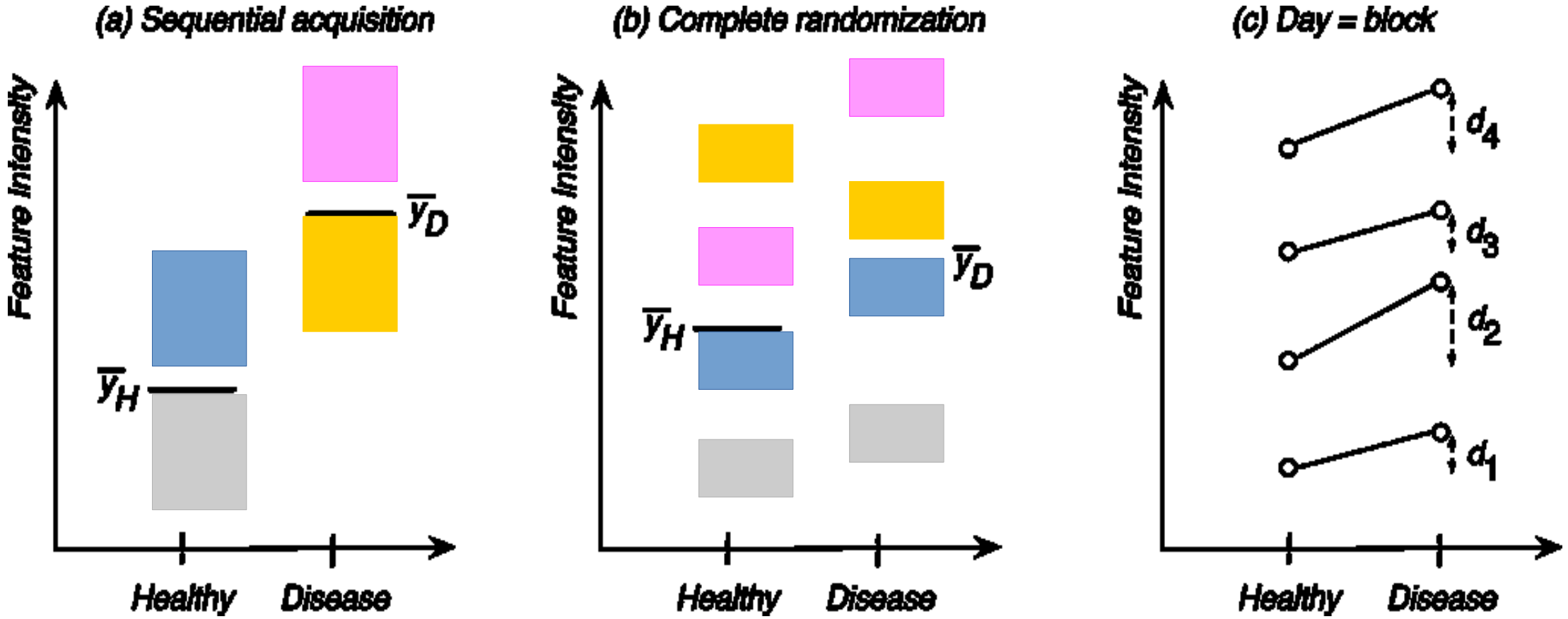
Preventivně: Stratifikací a randomizací

Ad-hoc: Regresními strategiemi

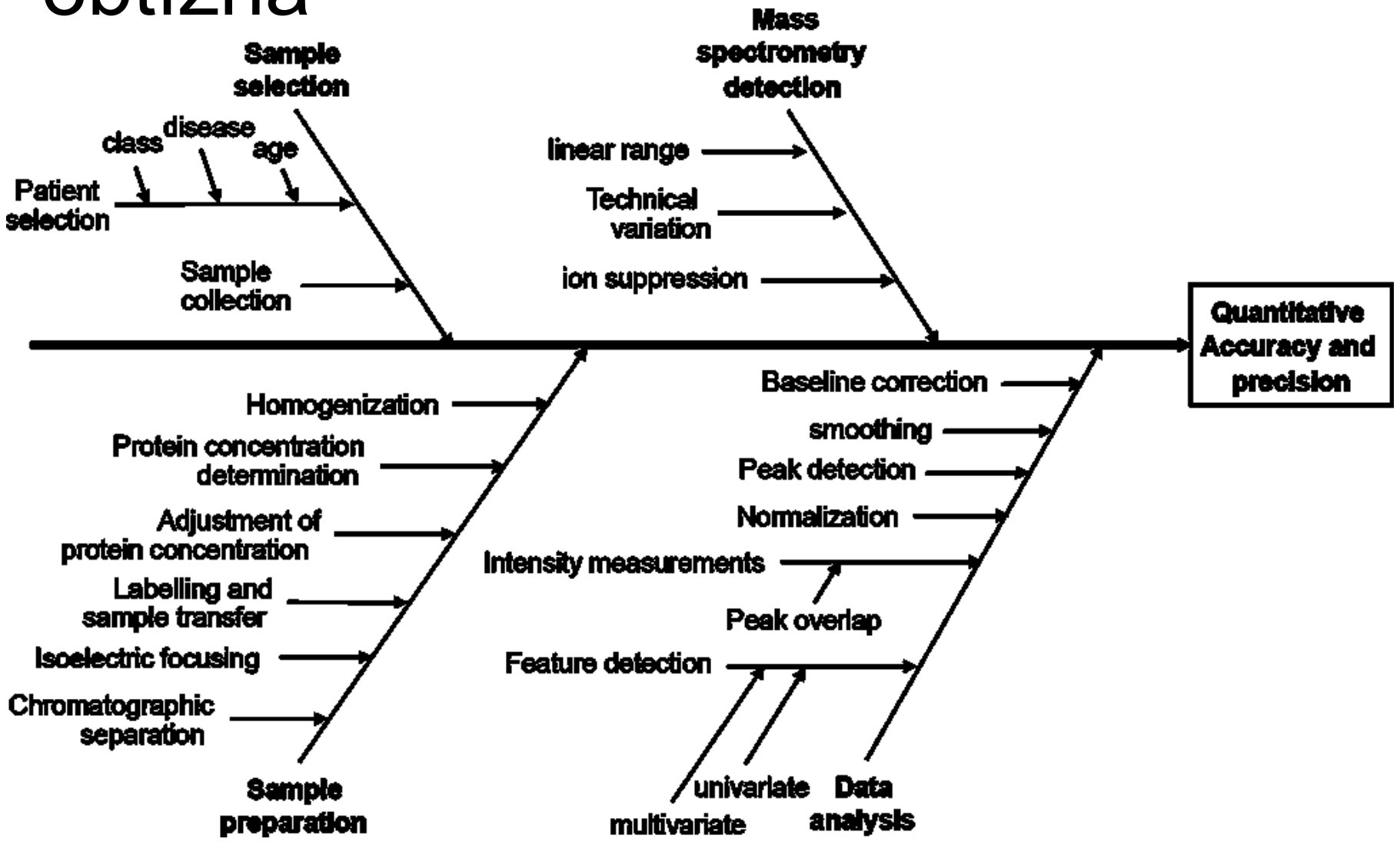
Preventivně: Stratifikací a randomizací

Ad-hoc: Regresními strategiemi

Randomizace pomáhá minimalizovat efekt dávky



U ‘omics dat je randomizace obtížná



Forshed J. (2017) Experimental design in clinical ‘omics biomarker discovery. Journal of Proteome Research 16, 3954-3960

Co když je randomizace nemožná (nebo ohrožena)

- Někdy všechno nejde naplánovat a něco se změní – experimenty můžou být dlouhodobé a spolupracovat může více stran, laboratoří, každá s vlastními postupy.

- Spolupráce více laboratoří – možnost randomizace na všech úrovních.

- Problematické bývá znovuoživení experimentu, který byl “u ledu” kvůli nedostatku financí (mezitím se změnili postupy).

- Další změny běžně ohrožující plánovanou randomizaci.
  - výměna laboranta…
  - pokazení stroje a nutná oprava nebo výměna
  - staré kity pro izolaci DNA už nevyrábějí, nutno použít jiné
  - ...

Preventivní minimalizace chyb

1. Protože vždy nevíme, co všechno může mít vliv, je důležité vést **PODROBNÉ ZÁZNAMY** – všechno co nás napadne!

- přesný záznam postupu, včetně uskladnění vzorku a jeho pozice v lednici
- kdo prováděl který typ analýzy a KDY
- každá změna v protokolu
- zaznamenáme všechny identifikační čísla jednotlivých kitů, primerů, čehokoliv
- všechny změny v kalibraci přístrojů, nebo informace o jejich čištění
- změny v teplotách
- způsob odběru vzorku (ležel materiál někde několik hodin mimo mrazák?)
- …

2. Provádíme po konzultaci se statistikem – **randomizaci a dizajn experimentu.**

3. V případě změn znovu konzultujeme další postup.

Co když je randomizace nemožná (nebo ohrožena)

- **KAŽDOU ZMĚNU KONZULTUJTE SE STATISTIKEM!**
- **ŘEŠENÍ (OBVYKLE) EXISTUJE !**
- **Efekt dávky se dá odstranit, máme-li dostatek stejných vzorků analyzovaných před i po změně –** vhodnými metodami se odhadne efekt a ten se pak z dat odstraní.
- **POZOR – je to nákladné a není to dokonalé, takže lépe je tyto efekty minimalizovat.**

Preventivně: Stratifikací a randomizací

Ad-hoc: Regresními strategiemi

Regresní strategie

Základní myšlenka je modelovat efekt dávky jako jednu z proměnných kterých vliv sledujeme

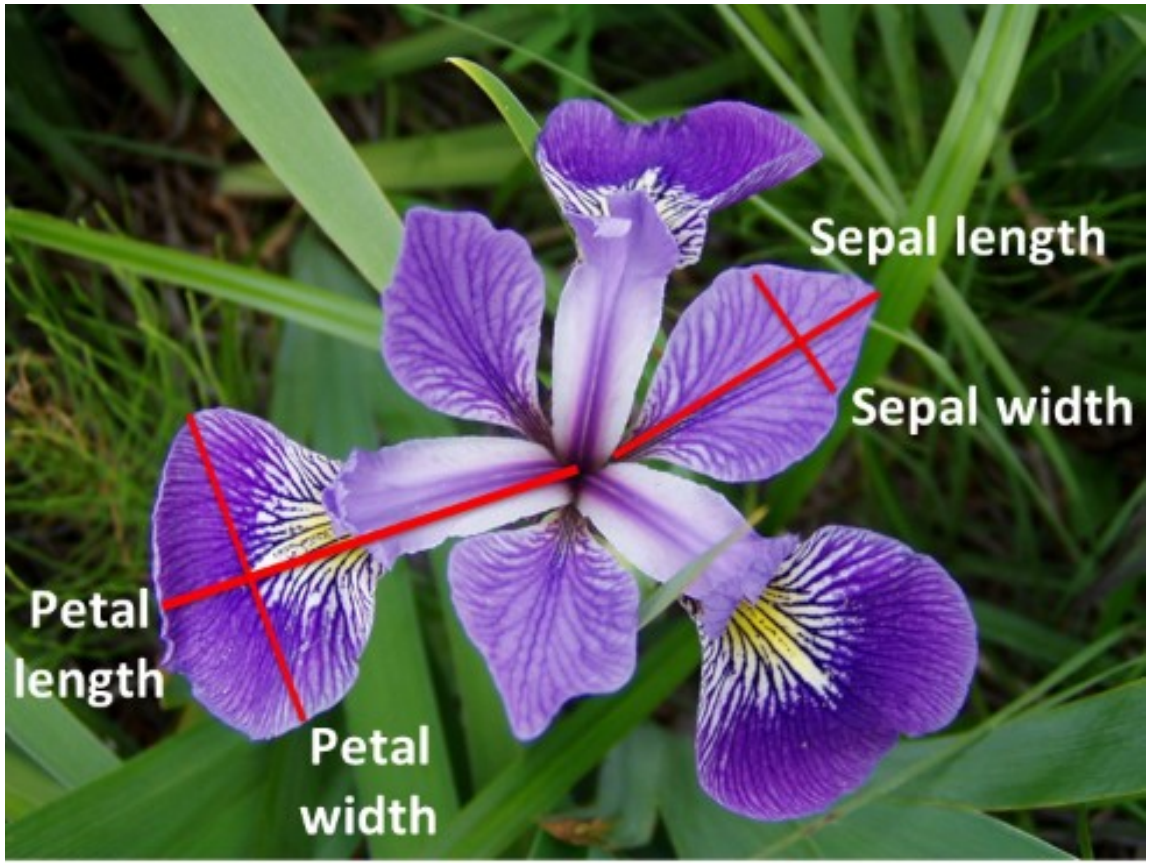
Odhadnutý efekt pak můžeme odstranit

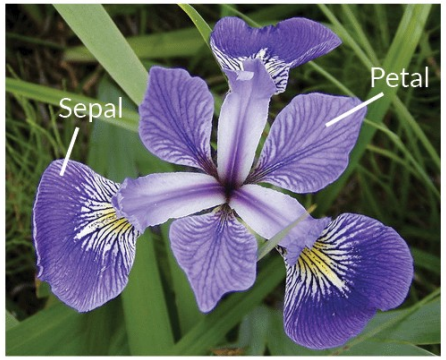
Nejčastěji regresní strategie

ComBat (R)

Odstranění batch efektu – jednoduchý příklad

3 druhy kosatců se liší na základě **šířky** a **délky** kališních (sepal) a okvětních (petal) lístků





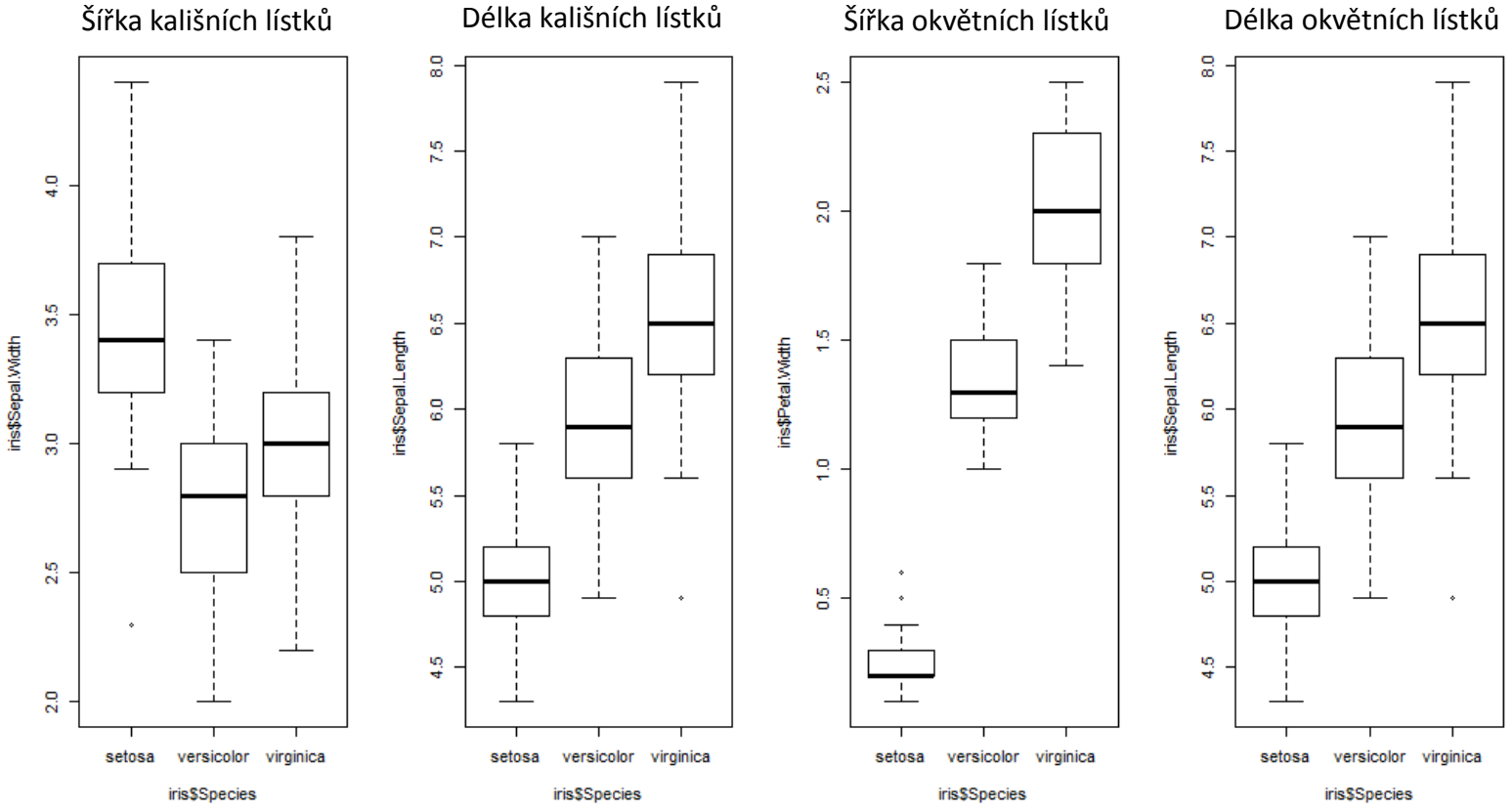
Iris Versicolor

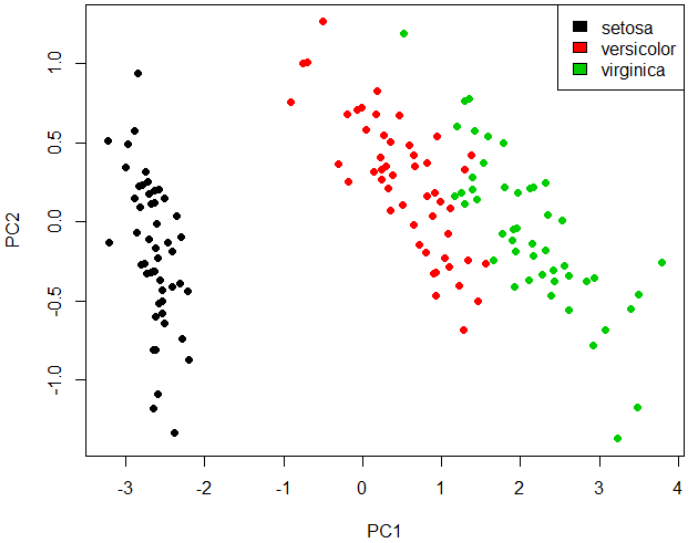
Iris Setosa

Iris Virginica

Odstranění batch efektu – jednoduchý příklad

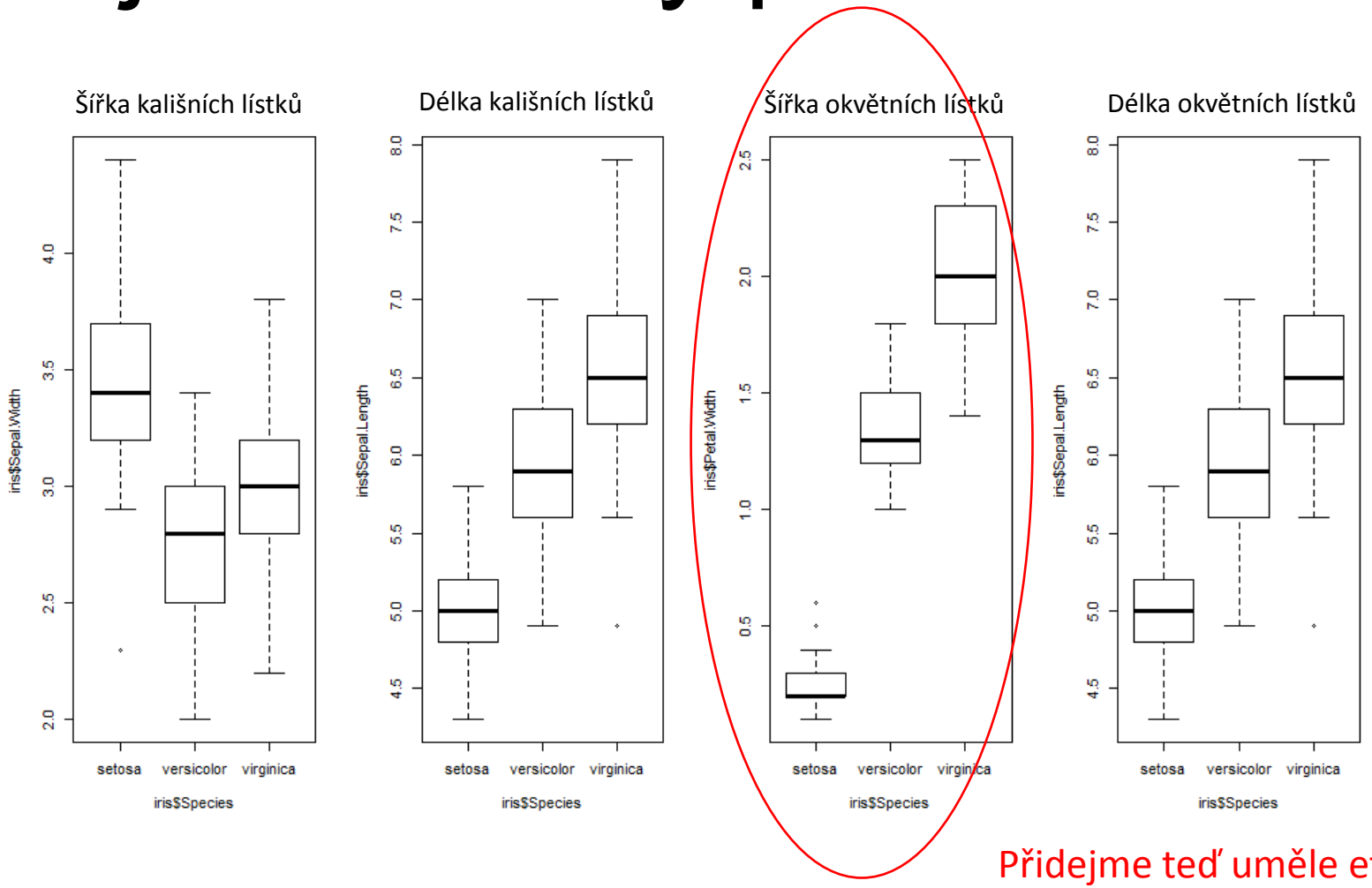
- 3 druhy kosatců se liší na základě **šířky** a **délky** kališních (sepal) a okvětních (petal) lístků

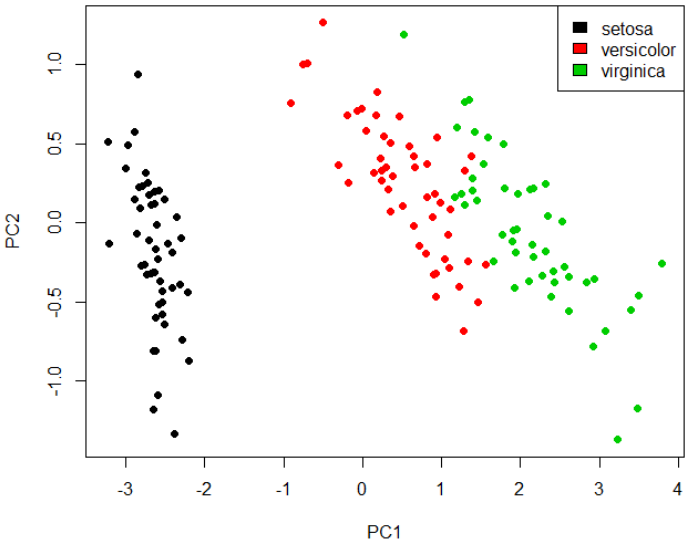




Odstranění batch efektu – jednoduchý příklad

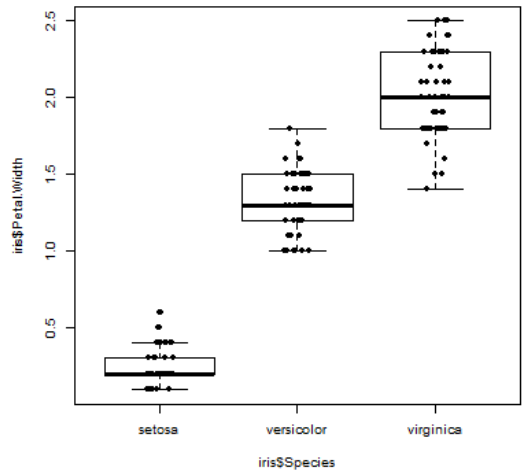
- 3 druhy kosatců se liší na základě **šířky** a **délky** kališních (sepal) a okvětních (petal) lístků

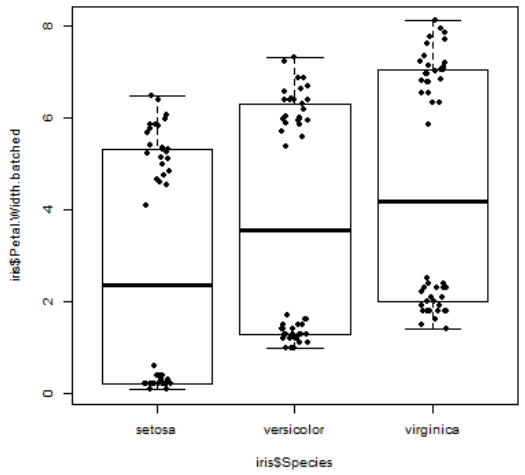




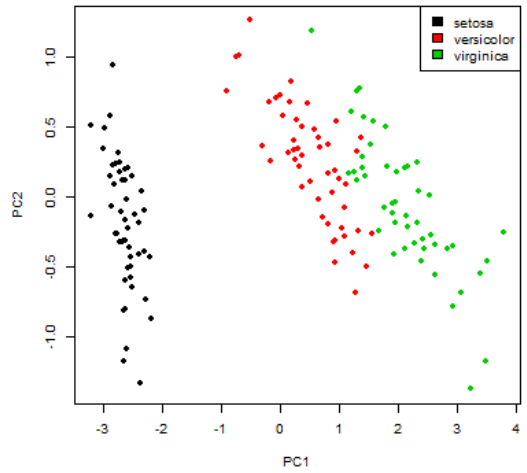
Přidejme teď uměle efekt dávky u této proměnné.

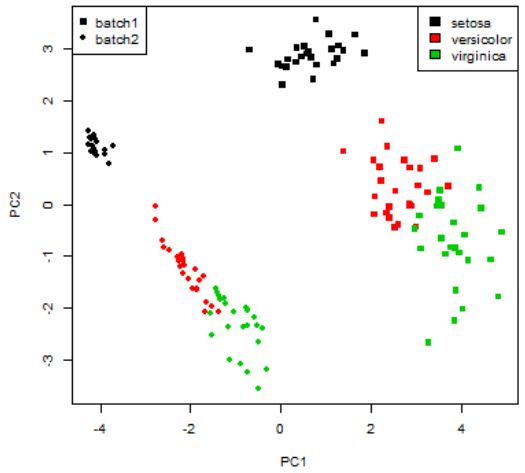
Odstranění batch efektu – jednoduchý příklad





PCA na původním souboru



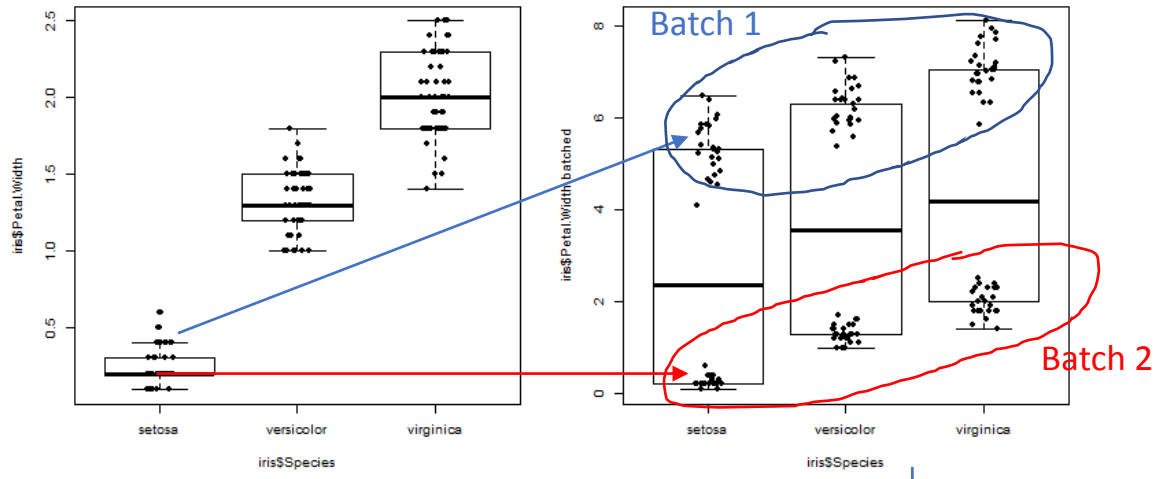


PCA na souboru s přidaným batch efektem

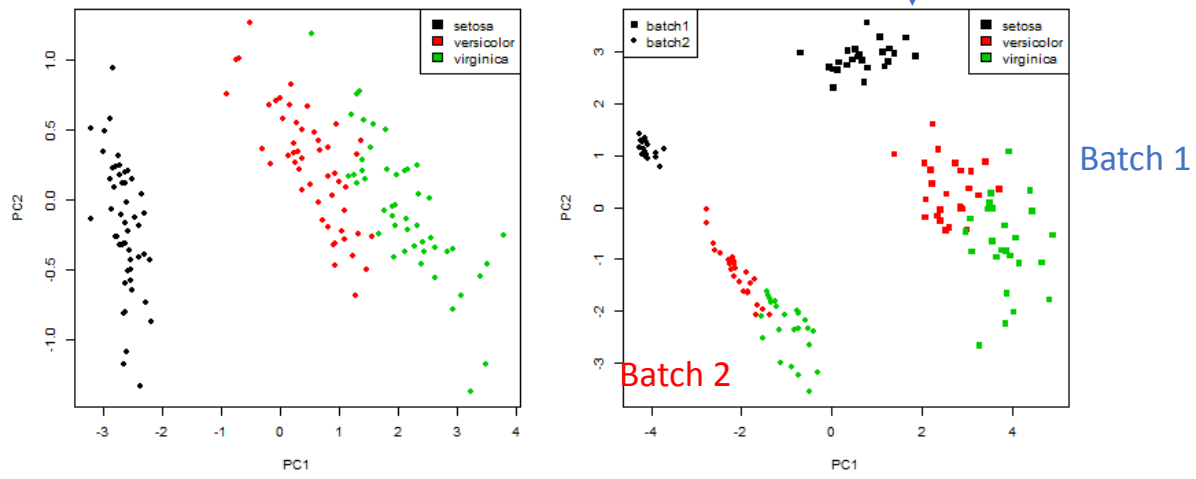
- 3 druhy kosatců se liší na základě **šířky** a **délky** kališních (sepal) a okvětních (petal) lístků

- K šířce okvětních lístků byl přidán batch efekt: k polovině hodnot u každého z druhů kosatce jsem připočítala hodnoty z normálního rozložení o **průměru 5** a standardní odchýlce 0,5

Odstranění batch efektu – jednoduchý příklad



PCA na původním souboru

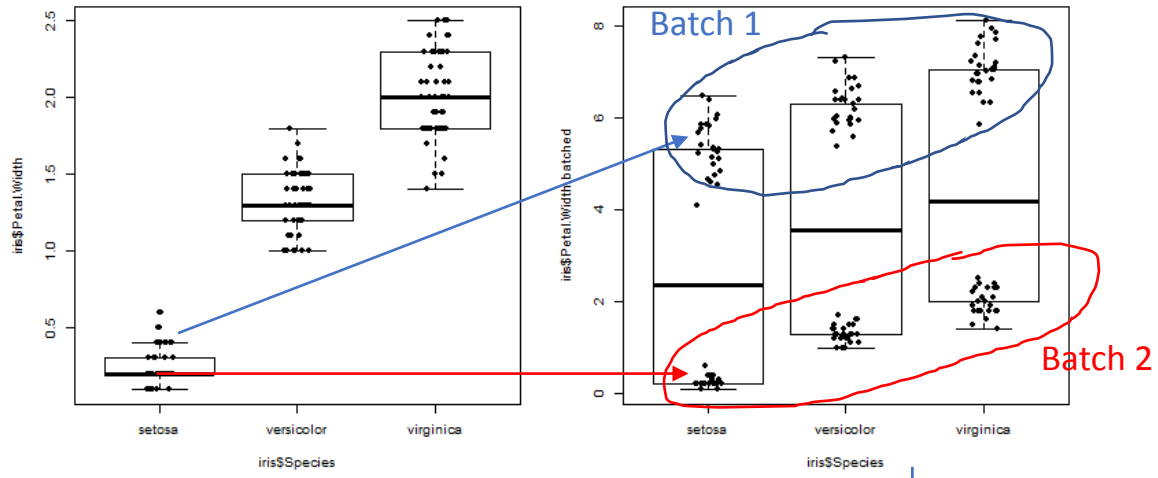


PCA na souboru s přidaným batch efektem

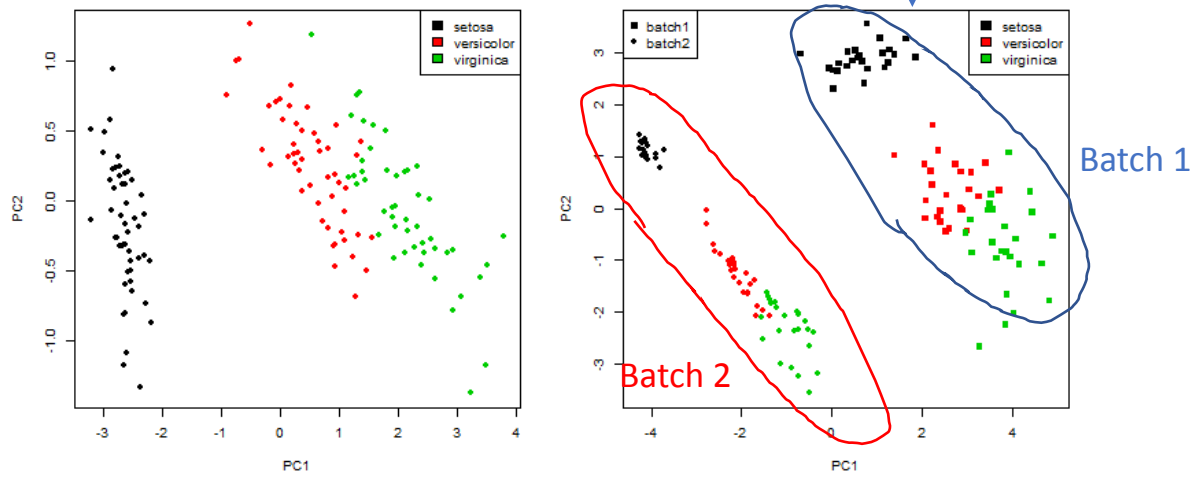
- 3 druhy kosatců se liší na základě **šířky** a **délky** kališních (sepal) a okvětních (petal) lístků

- K šířce okvětních lístků byl přidán batch efekt: k polovině hodnot u každého z druhů kosatce jsem připočítala hodnoty z normálního rozložení o **průměru 5** a standardní odchýlce 0,5

Odstranění batch efektu – jednoduchý příklad



PCA na původním souboru



PCA na souboru s přidaným batch efektem

- 3 druhy kosatců se liší na základě **šířky** a **délky** kališních (sepal) a okvětních (petal) lístků

- K šířce okvětních lístků byl přidán batch efekt: k polovině hodnot u každého z druhů kosatce jsem připočítala hodnoty z normálního rozložení o **průměru 5** a standardní odchýlce 0,5

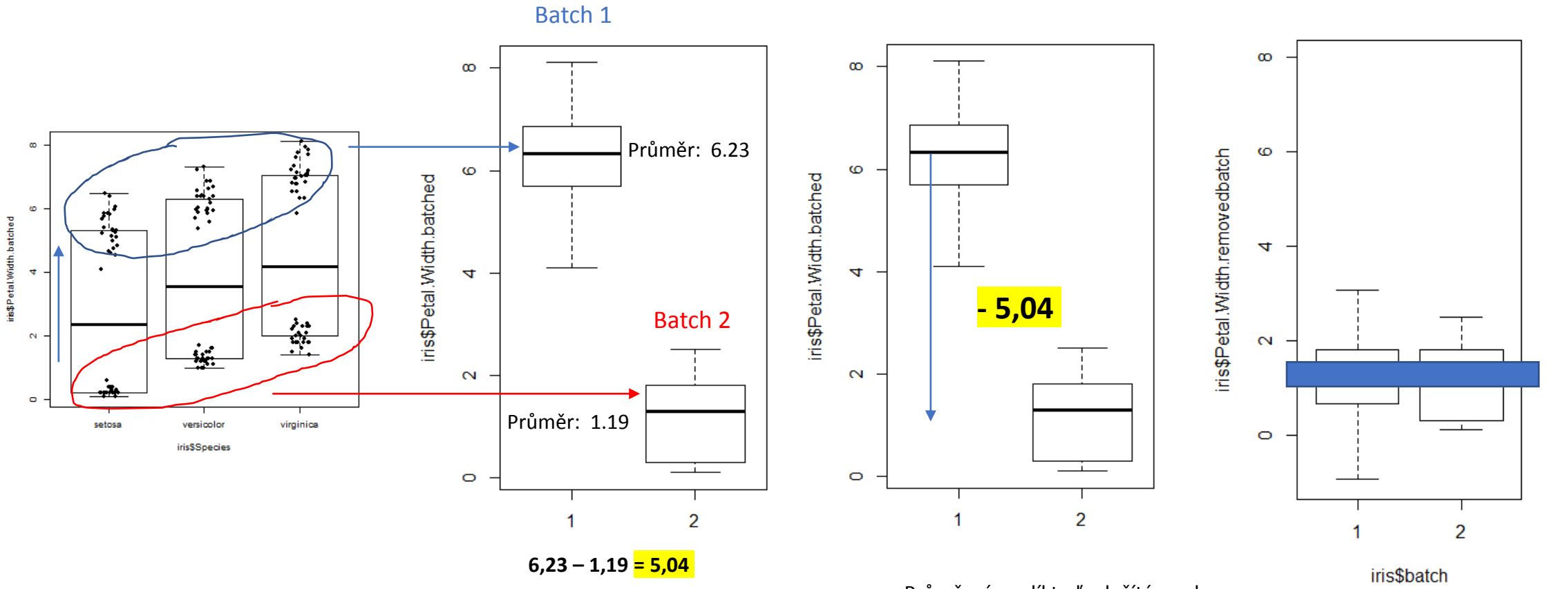
Odstranění batch efektu – jednoduchý příklad

- Jak odstranit batch efekt?

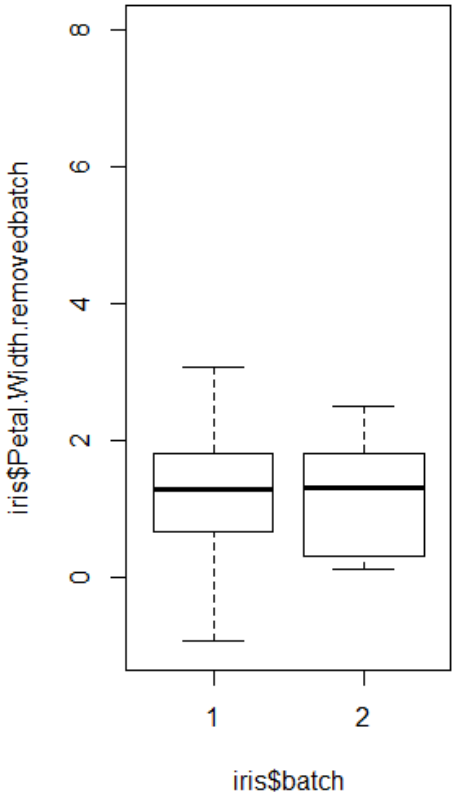
1. Nejdříve efekt odhadneme - je to posun v průměrné nebo mediánové hodnotě? Nebo je rozdíl i ve variabilitě?
   - Použijeme regresní modelování, testování hypotéz, stanovíme fold change a změnu variability
2. Tyto efekty pak odstraníme tak, že je odečteme (například průměr), nebo provedeme škálovou normalizaci a podobně

Odstranění batch efektu – jednoduchý příklad

- V našem příkladě: polovici hodnot jsem zvýšila v průměru o 5

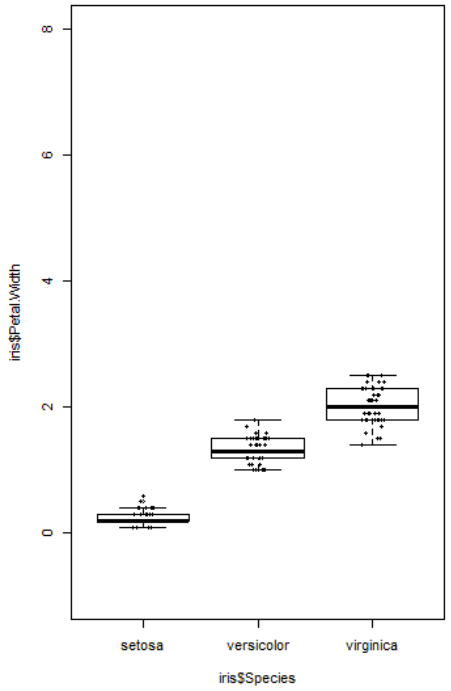


Odstranění batch efektu – jednoduchý příklad

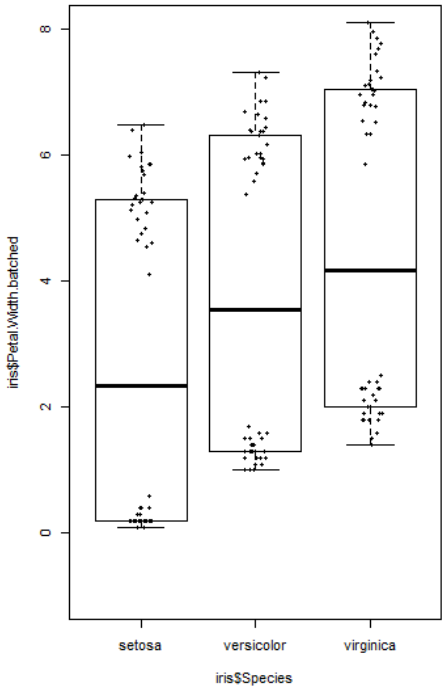


Odečteno, průměry jsou stejné

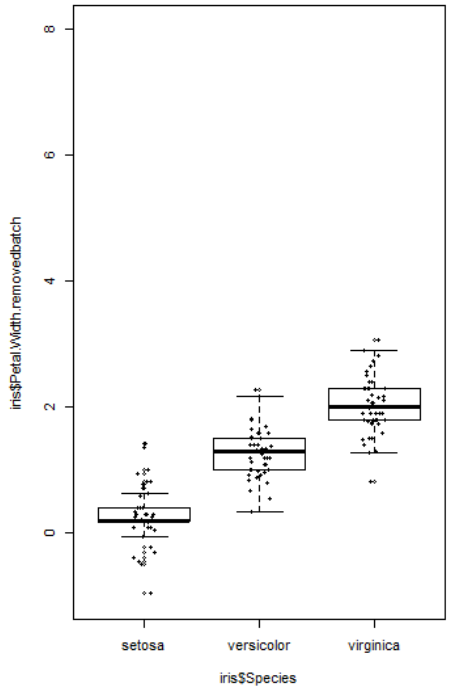
Původní

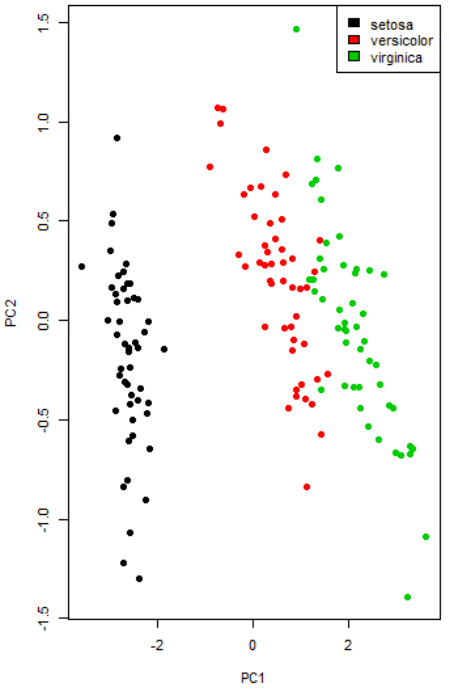


S batch efektem



Po odstranění





Doporučená literatura a další zdroje

THE UNIVERSITY OF TEXAS
MD Anderson Cancer Center
Making Cancer History®

Legal Notice

Batch Effects Viewer 2019-07-31-1100

Data Browser

Query Form

Index File: TCGA DCC 2016_12_12_1530 (histo

Disease: coad

Data Type: transcriptome

Platform: agilentg4502a-07-3 gene

Level: Level 3

Collection: Tumor-original

Algorithm: PCA

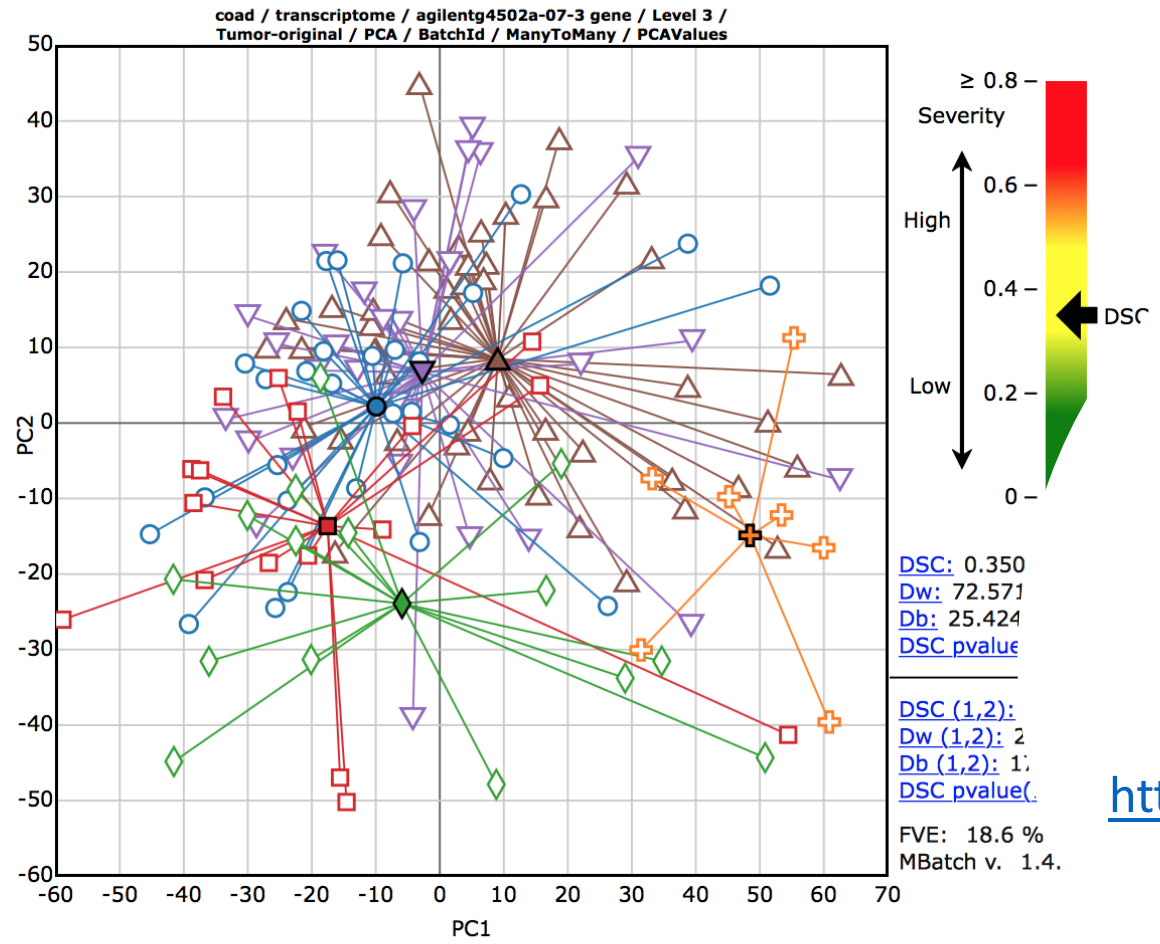
Diagram Type: BatchId

Sub-Type: ManyToMany

Diagram: PCAValues

coad: transcriptome, BatchId

View In New Window | Download Archive | Bookmark | Toggle Tooltips | Interpretation | Capture Diagram



Severity
≥ 0.8
0.6
0.4
0.2
0
High
Low
DSC

DSC: 0.350
Dw: 72.571
Db: 25.424
DSC pvalue

DSC (1,2):
Dw (1,2): 2
Db (1,2): 1
DSC pvalue(

FVE: 18.6 %
MBatch v. 1.4.

Legend
(n: 150)
28 (30)
29 (7)

Reset Zoom X-axis: PC1 Y-axis: PC2

TCGA Batch Effects Viewer

https://bioinformatics.mdanderson.org/BatchEffectsViewer/

Ann. Appl. Stat.
Volume 3, Number 4 (2009), 1309-1334.

Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology

Keith A. Baggerly and Kevin R. Coombes

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

Oytam *et al. BMC Bioinformatics* (2016) 17:332
DOI 10.1186/s12859-016-1212-5

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Risk-conscious correction of batch effects: maximising information extraction from high-throughput genomic datasets

Yalchin Oytam[1,2*], Fariborz Sobhanmanesh[1], Konsta Duesing[1], Joshua C. Bowden[3], Megan Osmond-McLeod[2] and Jason Ross[1]

Trends in Biotechnology CellPress

Go to Trends in Biotechnology on ScienceDirect

Vo es 498-507

Opinion

Special Issue: Computation and Modeling

Why Batch Effects Matter in Omics Data, and How to Avoid Them

Wilson Wen Bin Goh [1, 2], Wei Wang [1], Limsoon Wong [2, 3]

Show more

https://doi.org/10.1016/j.tibtech.2017.02.012

Journal of proteome research

Tutorial

Cite This: J. Proteome Res. 2017, 16, 3954-3960

pubs.acs.org/jpr

Experimental Design in Clinical 'Omics Biomarker Discovery

Jenny Forshed*

Department of Oncology-Pathology, Karolinska Institutet, BOX 1031, SE-171 21, Stockholm, Sweden