

Analýza genomických a proteomických dat

Základní schéma analýzy Porovnání skupin

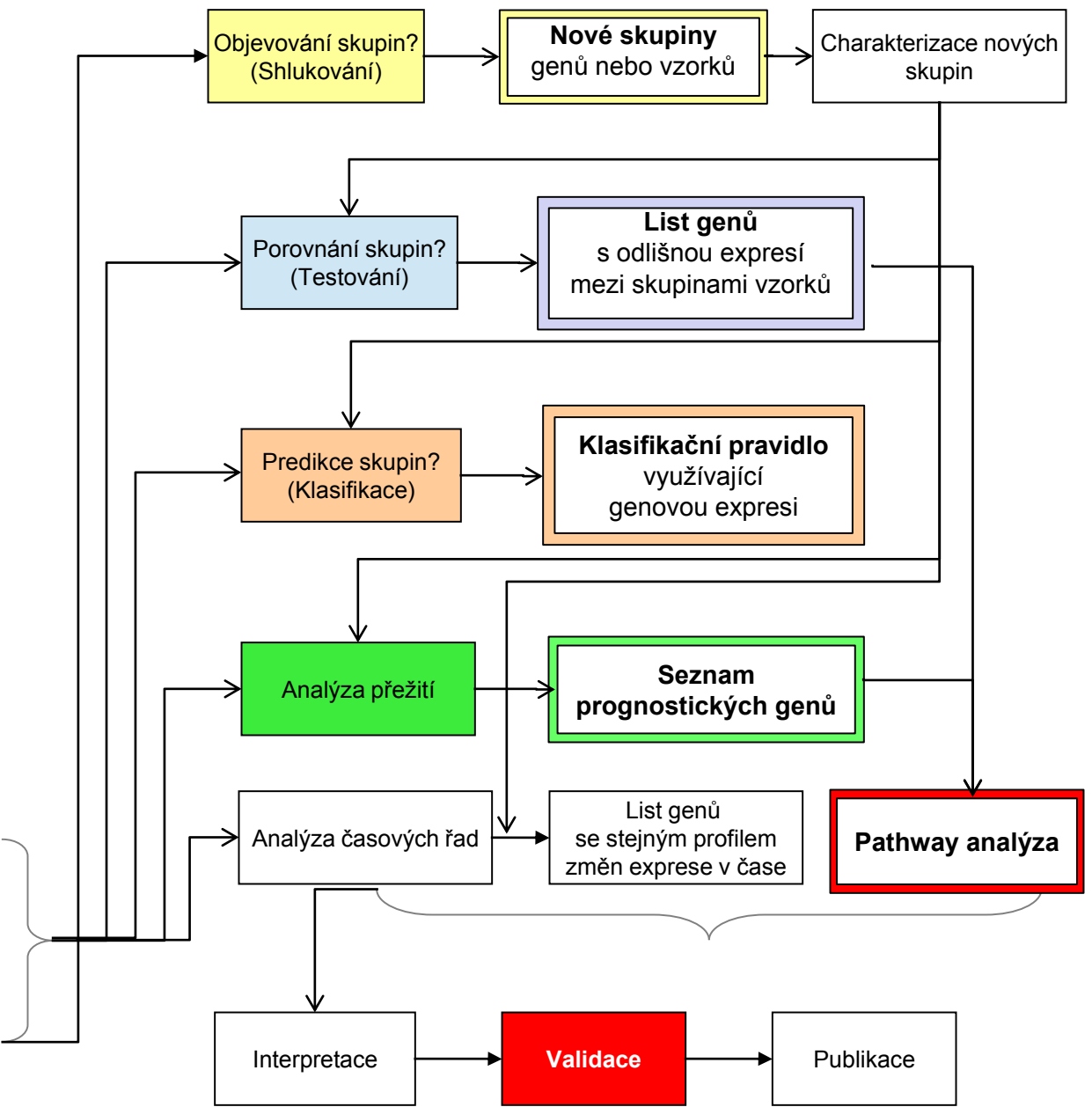
Jaro 2023

29. březen 2023

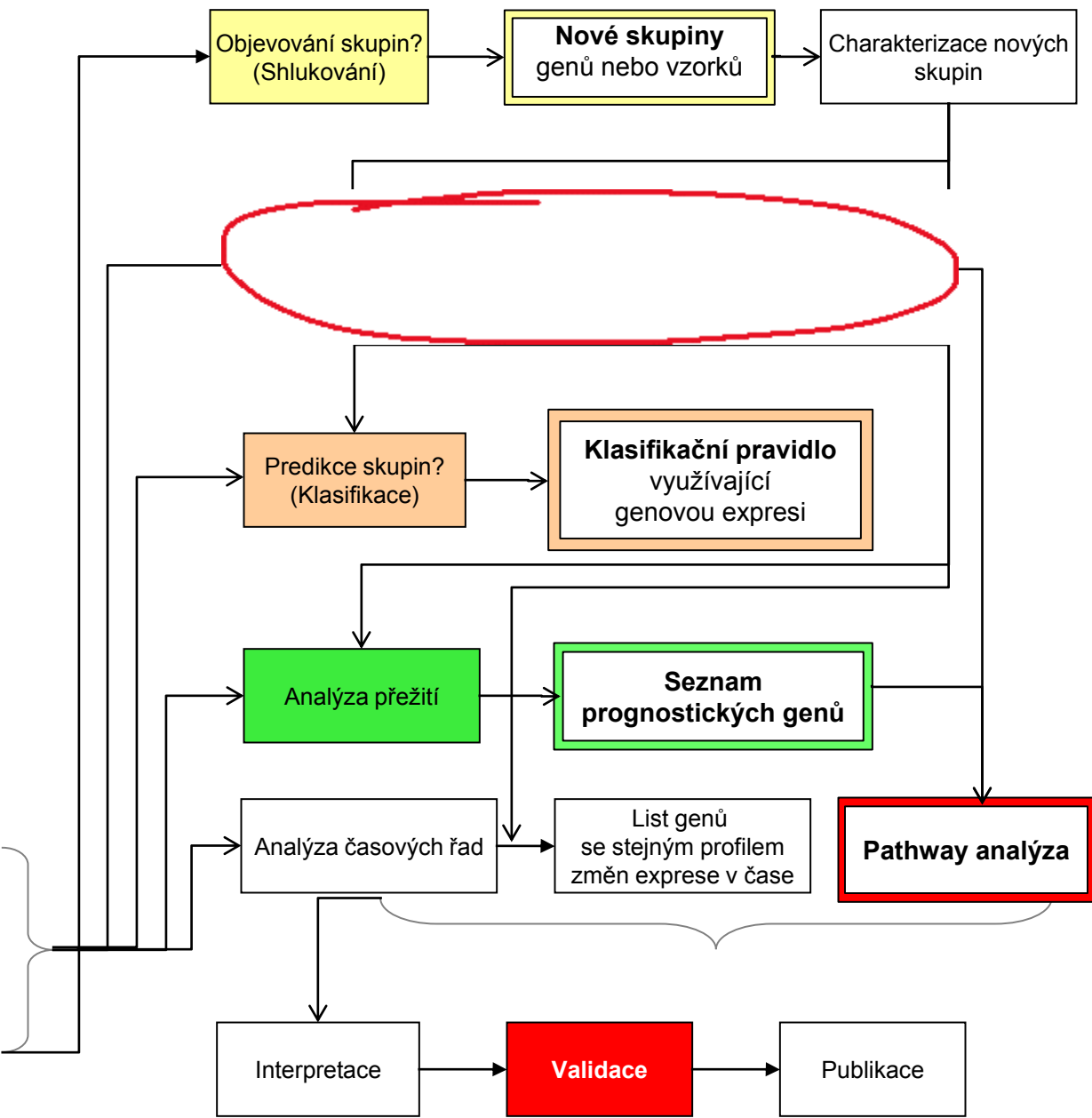
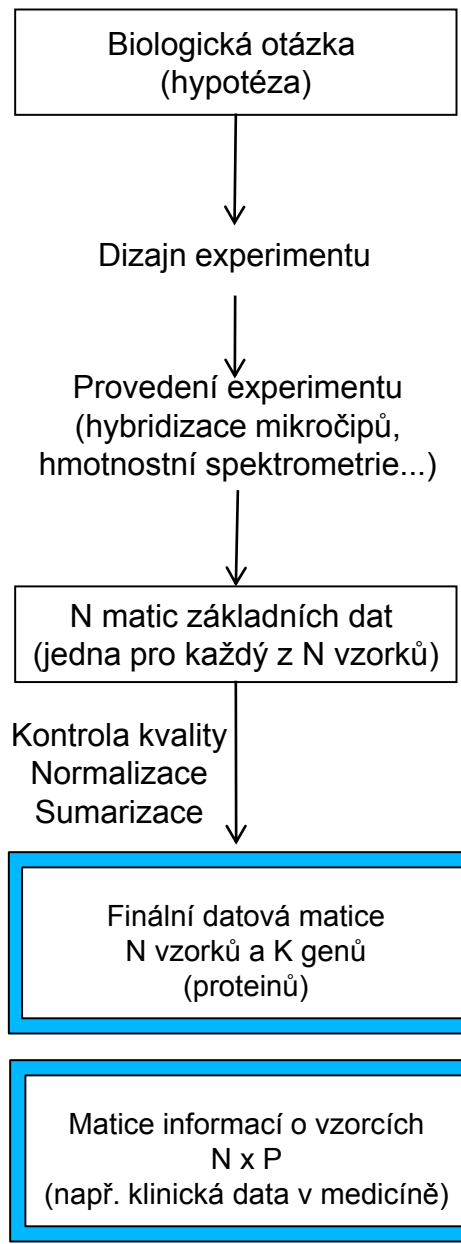
Eva Budinská (eva.budinska@recetox.muni.cz)



Základní schéma analýzy genomických a proteomických dat



Základní schéma analýzy genomických a proteomických dat



Porovnávání skupin

Jaký je rozdíl v přítomných genech/proteinech mezi dvěma nebo více skupinami vzorků?



Příklady porovnávání í skupin

nemocní vs. zdraví pacienti

pacienti před vs. po terapii

pacienti v čase diagnózy a v čase relapsu

bakterie v aerobním vs. anaerobním prostředí

druh 1 vs. druh 2

porovnáváme podtypy onemocnění

Základní metody pro porovnávání

Můžeme
rozdělit
do tří
hlavních
skupin:

Metoda dělicí hranice
velikosti efektu/změny
mezi skupinami

Testování hypotéz

Regresní strategie

Základní metody pro porovnávání

Můžeme
rozdělit
do tří
hlavních
skupin:

Metoda dělicí hranice
velikosti efektu/změny
mezi skupinami

Testování hypotéz

Regresní strategie

Metoda dělicí hranice velikosti efektu / změny

Princip:

- Porovnává se poměr průměrů/mediánů jedné a druhé skupiny: $\text{mean}(X)/\text{mean}(Y)$.
- Stanoví se **fixní dělicí hranice**, které určují, jaká velikost efektu je pro nás zajímavá

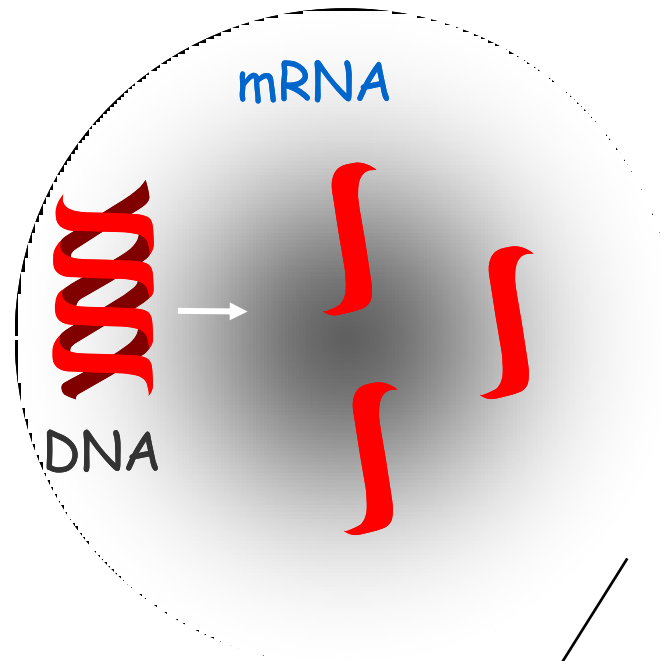
Příklad:

- genová exprese, $\text{průměr}(X)/\text{průměr}(Y)$, kde X a Y jsou genové exprese ve skupinách, použitá dělicí hranice: 2

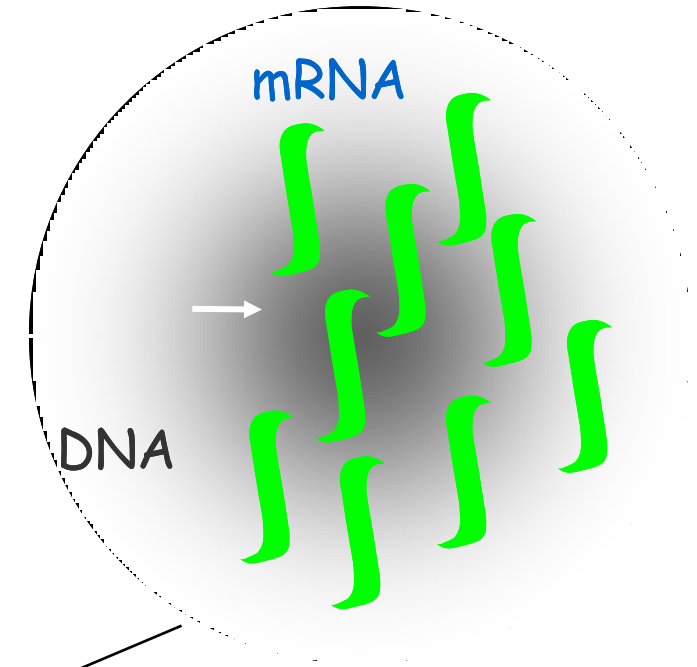
Výhoda: jednoduché

Metoda dělicí hranice velikosti efektu / změny

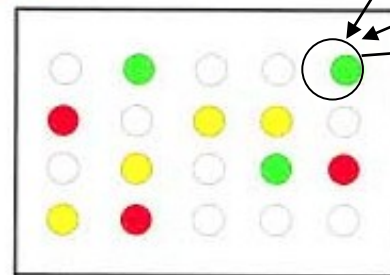
Skupina A. Zdravá tkáň



Skupina B. Nádor



● Sample A > B
● Sample B > A
● Sample A = B



$9/3 = 3$

Gen g_1 je 3x více exprimován v nádoru, než ve zdravé tkáni

Metoda dělicí hranice velikosti efektu / změny

Nevýhody:

- **I menší změny mohou být biologicky významné** (malý efekt genu/proteinu může být znásobený kooperací více genů v dráze)
- Data jsou ovlivněné **technickou a biologickou** variabilitou:
 - Co s hodnotou 1.9999 ?
 - Hodnoty mohou být vychýlené směrem k nule (například u nádorů s příměsí normálních buněk ve vzorku)
- **Neberou do úvahy variabilitu!**

Základní metody pro porovnáván í

Můžeme
rozdělit
do tří
hlavních
skupin:

Metoda dělicí hranice
velikosti efektu/změny
mezi skupinami

Testování hypotéz

Regresní strategie

Testování hypotéz

•

•

•

•

Co je to *statistika*



- Abychom rozhodli, která hypotéza je pravdivá, sumarizujeme data do **jednoho čísla**
- V testování hypotéz se toto číslo nazývá ***statistika*** (*T-statistika, Z-statistika, F-statistika...*)
- Statistiky jsou definovány různě a mají různé předpoklady.
- Například T-statistika porovnává signál se šumem a předpokládá normalitu dat.

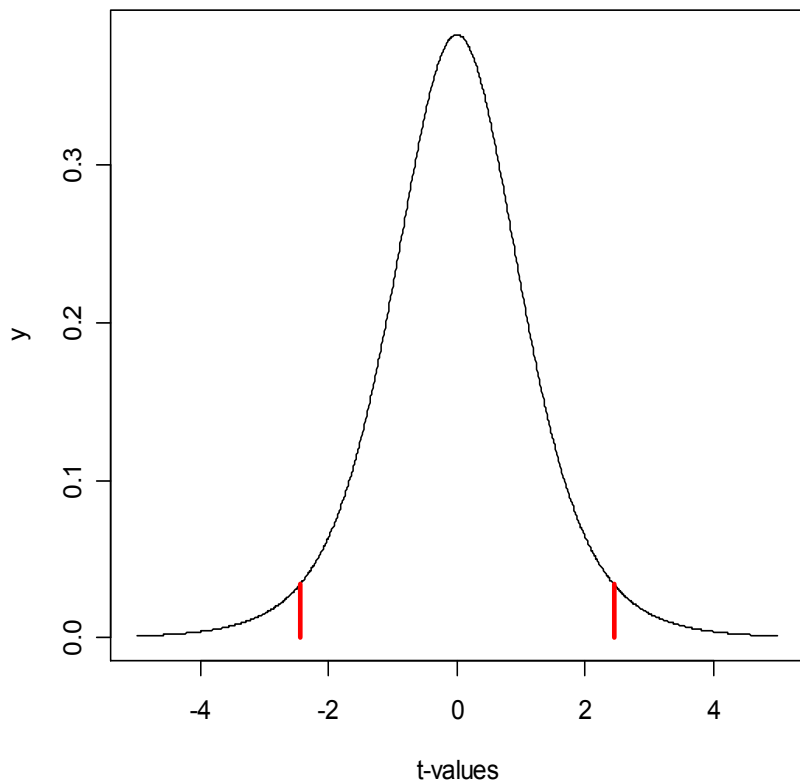
T-test

Klademe si otázku: Je aktivita/množství proteinu/genu ve skupině A odlišné od průměrné aktivity/množství proteinu/genu ve skupině B?

Na každý protein/gen g aplikujeme statistický test, kterým získáme T_g statistiku a příslušné p -hodnoty

T-test a T-statistika

Distribution of t-statistic (df =6)

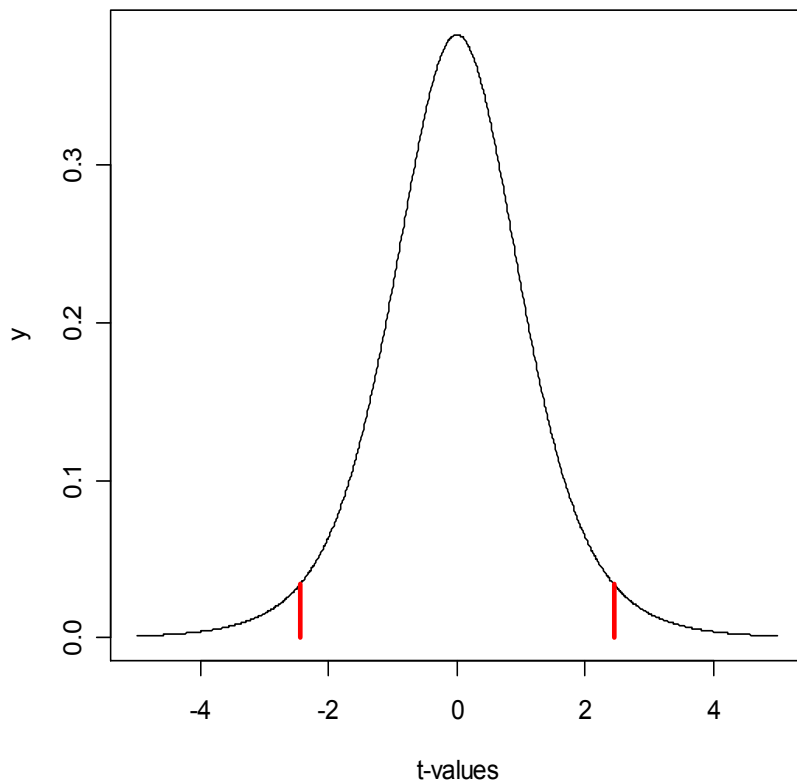


- Dvouvýběrový T-test pro porovnání rovnosti dvou průměrů μ_1, μ_2 :
 - Průměr exprese genu ve skupině 1 vs. průměr ve skupině 2

↑
Variabilita (vyjádřená jako
směrodatná odchylka)

T-test a T-statistika

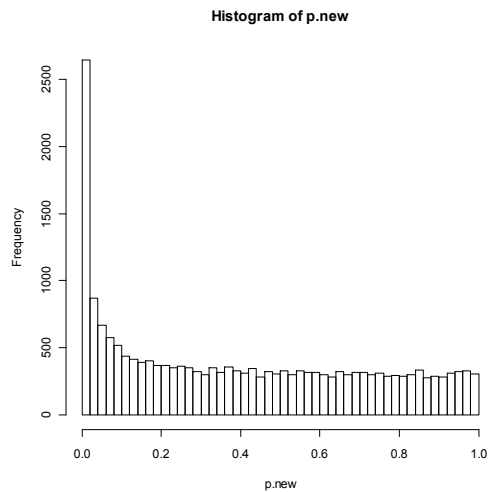
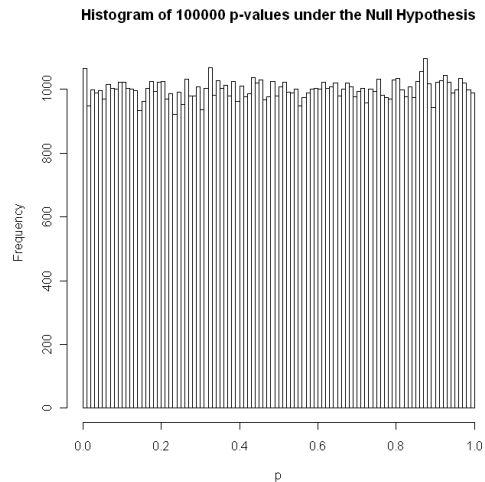
Distribution of t-statistic (df =6)



- Pokud data mají **normální rozložení a neexistuje rozdíl mezi skupinami**, tak T-statistiky pocházejí z **T-rozložení**.
- **p-hodnota** = pravděpodobnost že dostaneme danou hodnotu T-statistiky nebo hodnotu větší, v případě, že neexistuje rozdíl mezi skupinami

$$p_g = \Pr(T_g \leq T)$$

- Dostatečně malá p-hodnota = významný rozdíl (silná evidence)



Testování hypotéz

- Typické rozhodovací pravidlo:
 - Výpočet T-statistiky a p-hodnoty
 - Pokud $p < 5\%$, gen je označený za odlišně exprimovaný

Důležité:

- V případě, že platí nulová hypotéza, jsou **p-hodnoty všech testovaných hypotéz (genů) rovnoměrně rozloženy**.
- V případě, že je značná část genů odlišně exprimovaná, rozložení p-hodnot už není uniformní.

Možné výsledky testování

	H_0 nezamítneme	H_0 zamítneme
H_0 je pravdivá (gen není odlišně exprimovaný)	Pravdivá negativita (PN)	Falešná pozitivita (FP) Chyba I. druhu
H_0 není pravdivá (gen je odlišně exprimovaný)	Falešná negativita (FN) Chyba II. druhu	Pravdivá pozitivita (PP)

Problém mnohonásobné ho porovnávání

Porovnáváme tisíce genů/proteinů mezi skupinami.

Hypotézu testujeme pro každý gen!

Máme zvýšenou šanci falešně pozitivních výsledků!

Příklad: 10 000 genů, žádný odlišně exprimovaný mezi skupinami => $0.05 \times 10\,000 = 500$ s $p < 0.05$.

$p < 0.05$ už negarantuje významnost výsledku

Musíme tedy udělat korekci p-hodnot na mnohonásobné porovnání

Korekce problému mnohonásobného porovnávání

	# nezamítnuté (NZ)	# zamítnuté (Z)
#bez rozdílu	Pravdivá negativita (PN)	Falešná pozitivita (FP) Chyba I. druhu
# odlišné geny/proteiny	Falešná negativita (FN) Chyba II. druhu	Pravdivá pozitivita (PP)

Chyby 1. druhu:

- 1. Family-wise error rate (FWER):** Pravděpodobnost alespoň jedné chyby prvního druhu (falešné positivity): $FWER = Pr(FP > 0)$
- 1. False discovery rate (FDR)**(Benjamini & Hochberg, 1995):
Očekávaný podíl falešně pozitivních výsledků mezi zamítnutými hypotézami

$$FDR = E[FP/Z]$$

Korekce p-hodnot při mnohonásobn ém testování

! Existuje více druhů
metod pro kontrolu FDR!

Kontrolujeme FWER

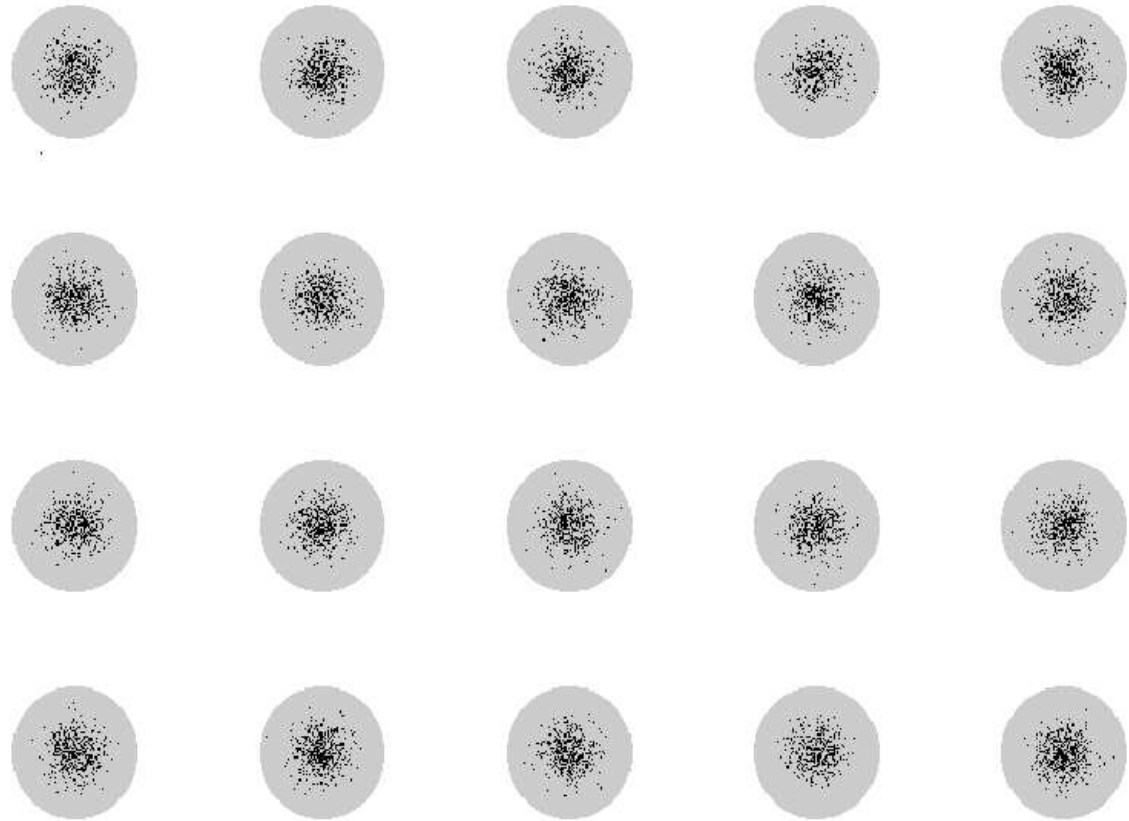
- Bonferroniho korekce (pro nezávislé testy!)
- $p < \alpha / m$ (napr. $p < 0.05/10\ 000$)

Kontrolujeme FDR

- Benjamini/Hochbergova procedura
 - FDR = 10% (ze 100 zamítnutých hypotéz očekáváme 10 falešně pozitivních)
 - (q-hodnota je nejmenší FDR při které daný gen ještě zůstává na listu pozitivních)

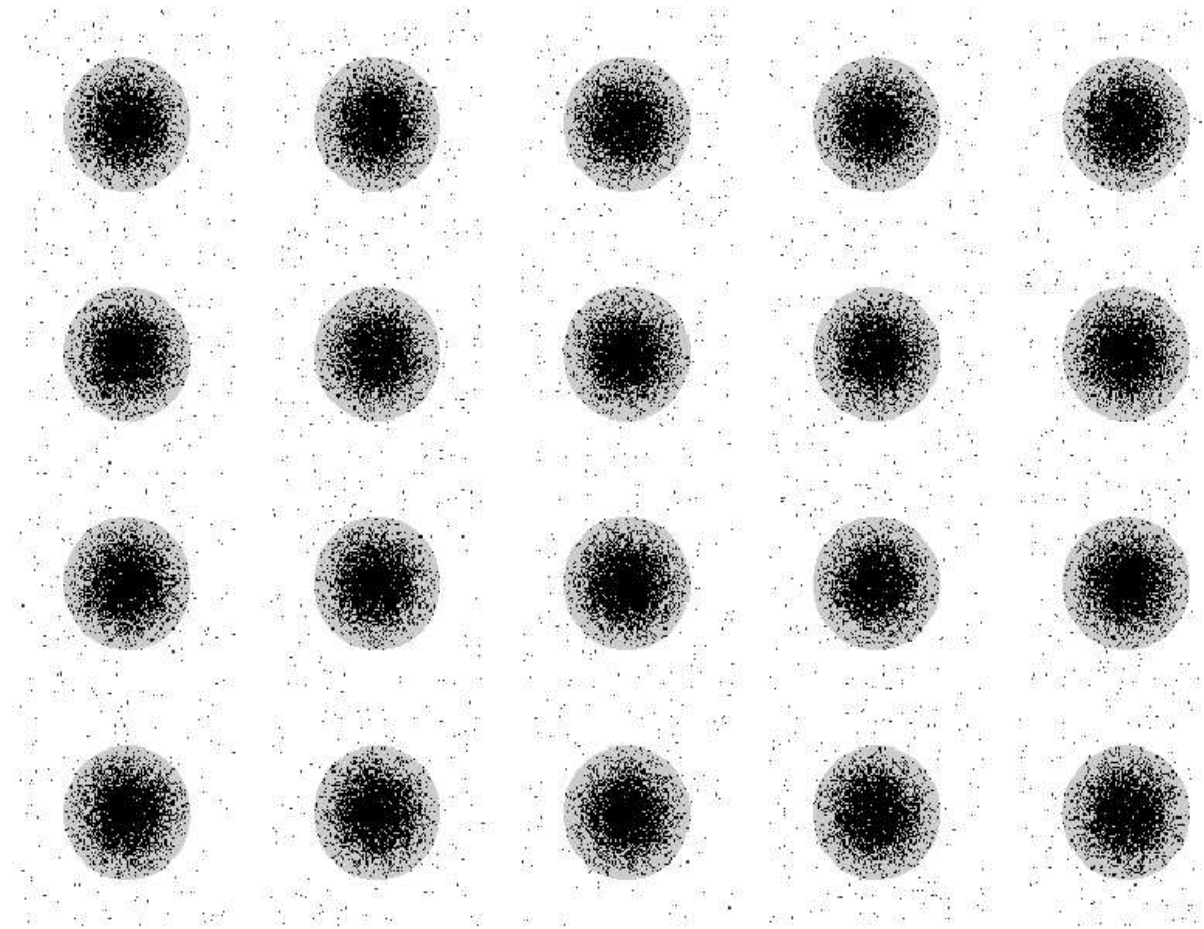
Který typ
korekce
použít?

FWER pokud chceme aby **VŠECHNY**
vybrané geny/proteiny byly opravdu
významné. Na druhou stranu, nevybereme

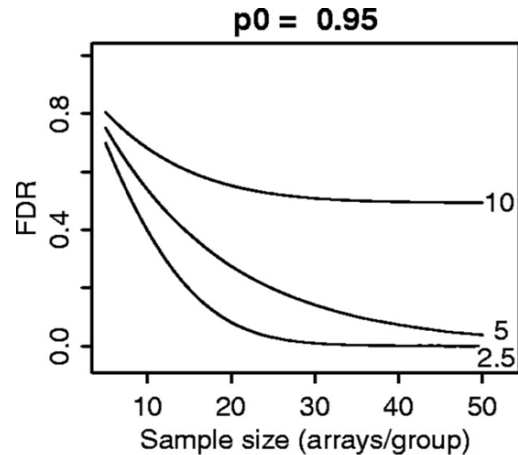


Který typ
korekce
použít?

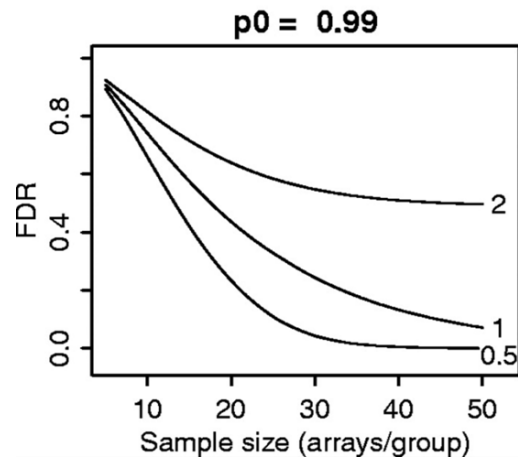
FDR pokud preferujeme vybrat většinu významných genů/proteinů, a nevadí nám nějaké falešně pozitivní



Vliv počtu vzorků na falešně pozitivní výsledky



p_0 : skutečný podíl genů beze změny exprese mezi skupinami (false negative rate)

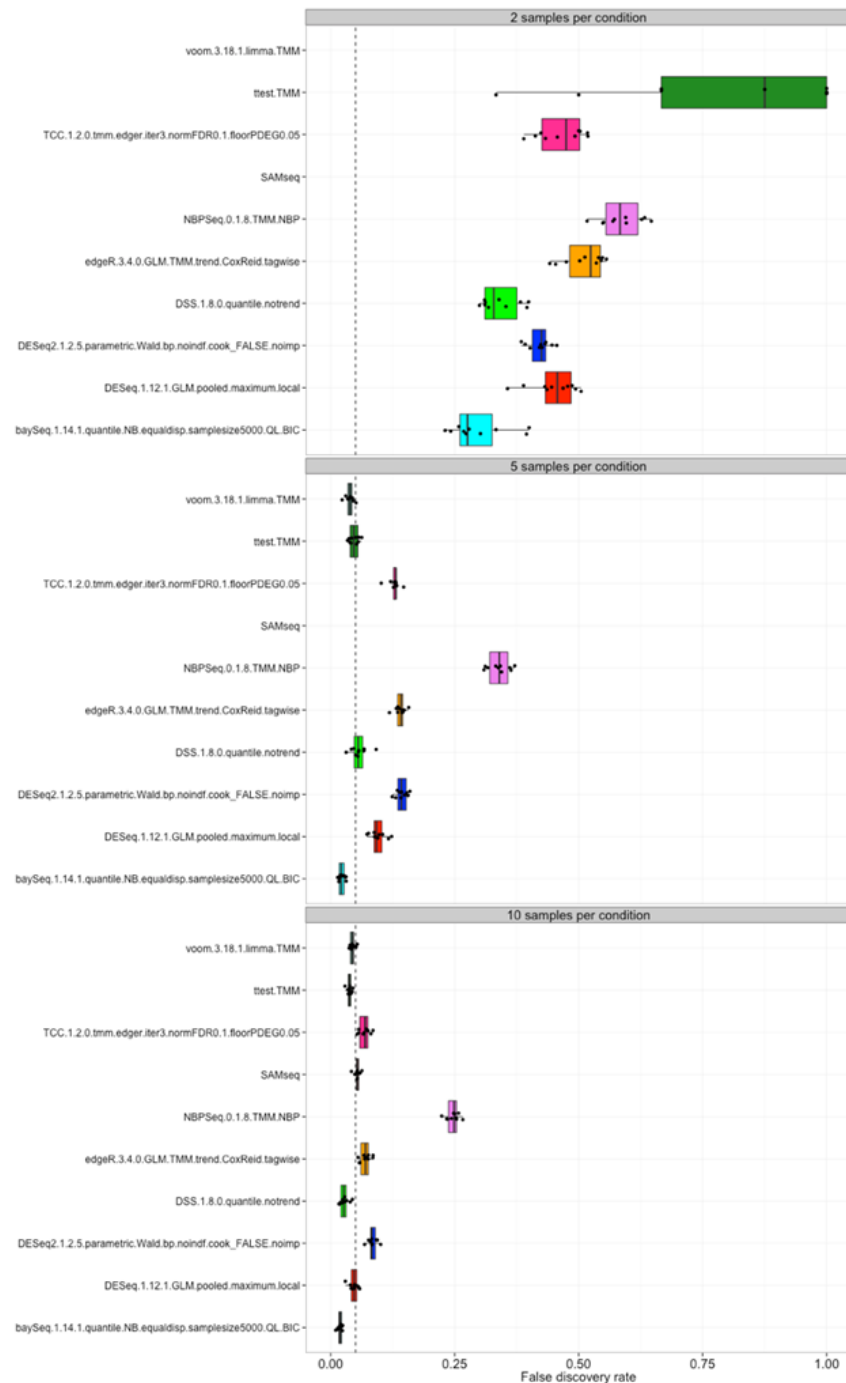


FDR (false discovery rate) jako funkce velikosti vzorku a procenta významných výsledků. Každá křivka představuje fixní procento genů označených jako významných.

From: False discovery rate, sensitivity and sample size for microarray studies

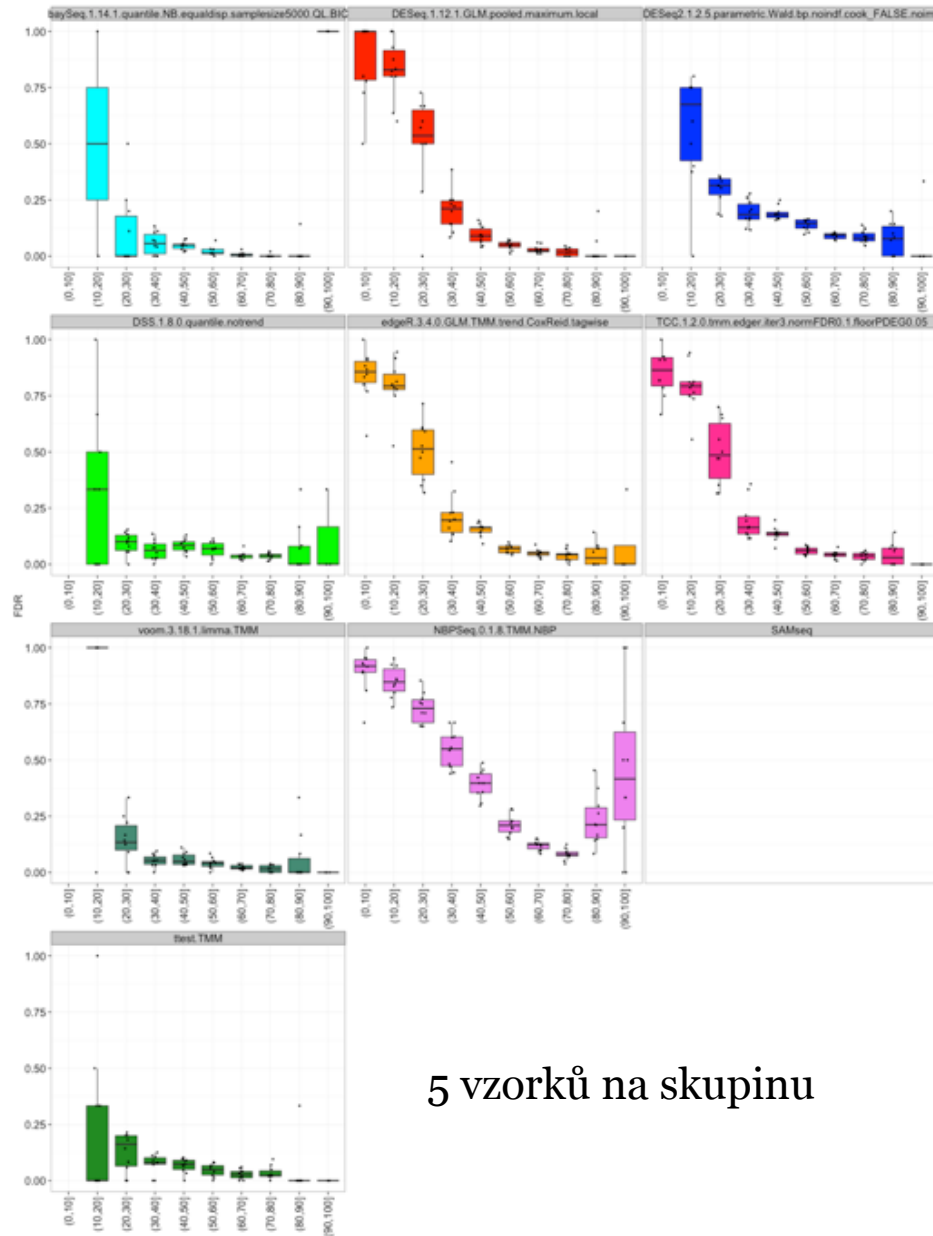
Bioinformatics. 2005;21(13):3017-3024. doi:10.1093/bioinformatics/bti448

Bioinformatics | © The Author 2005. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oupjournals.org

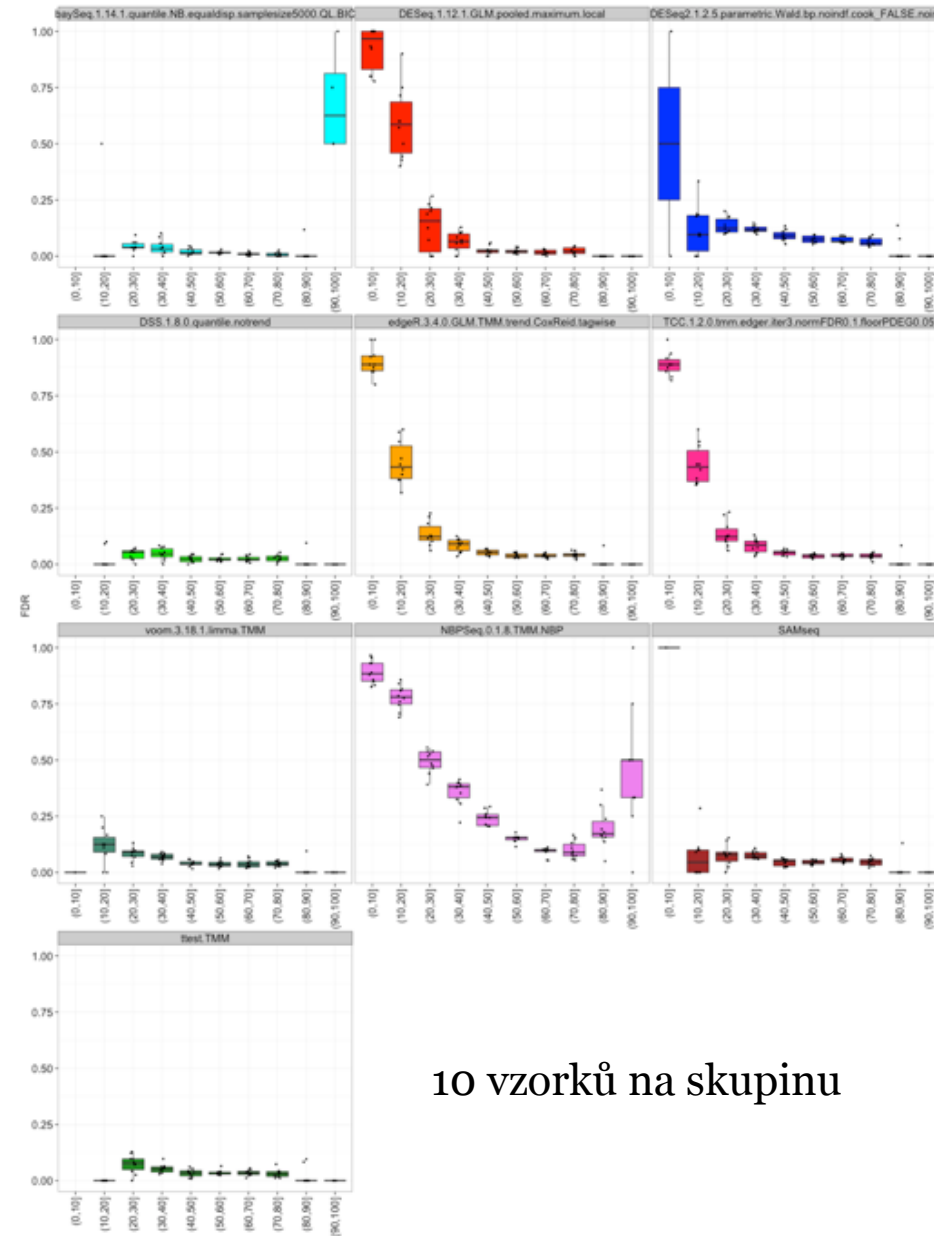


FDR (False discovery rate) jako funkce počtu vzorků na skupinu a metody použité pro normalizaci sekvenačních dat a testování hypotéz

FDR (False discovery rate) jako funkce genové exprese a použité metody pro normalizaci dat a testování

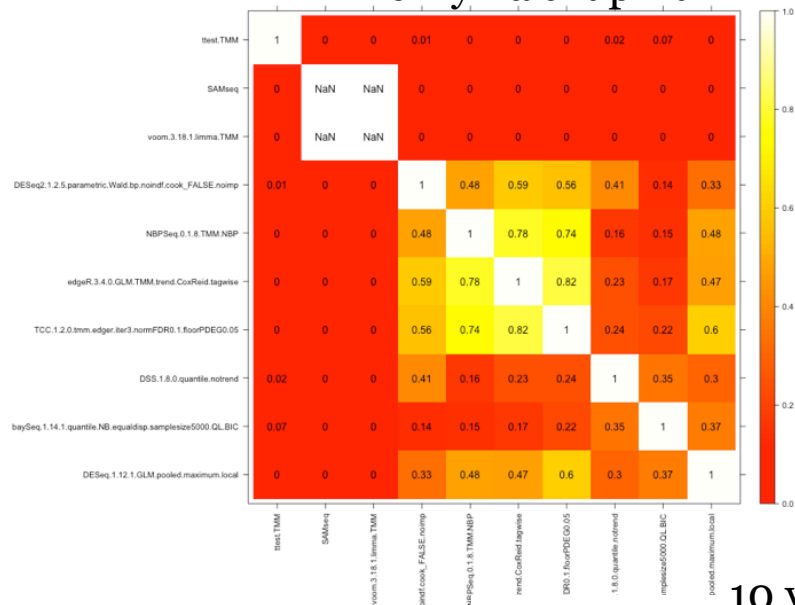


5 vzorků na skupinu

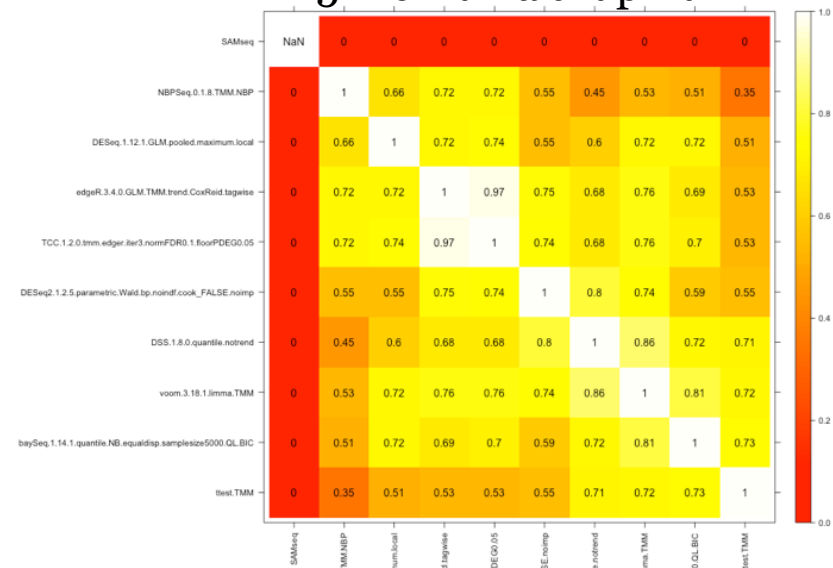


10 vzorků na skupinu

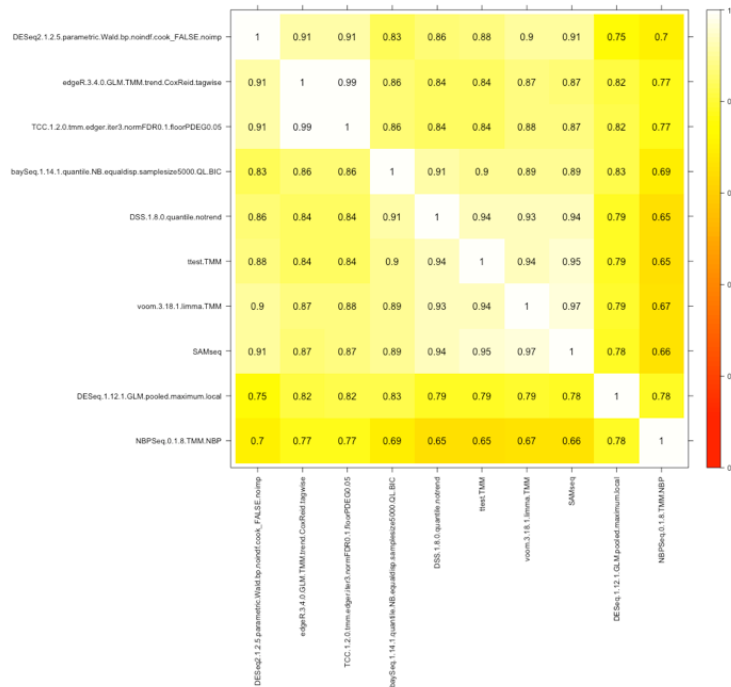
2 vzorky na skupinu



5 vzorků na skupinu



10 vzorků na skupinu



Similarita mezi seznamy odlišně exprimovaných genů mezi metodami u N=2,5 a 10

Doporučená literatura na tému FDR

- <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-450>



Základní metody pro porovnávání

Můžeme
rozdělit
do tří
hlavních
skupin:

Metoda dělicí hranice
velikosti efektu/změny
mezi skupinami

Testování hypotéz

Regresní strategie

Regresní strategie

- Pokud máme víc jak 1 proměnnou, která může ovlivnit genovou/proteinovou expresi
 - genová exprese \sim skupina + pohlaví
 - *Lineární modelování (limma)*
- Pokud se snažíme zjistit, jak velmi se genová exprese změní, pokud se změní hodnota nějaké *spojité proměnné*
 - genová exprese \sim přežití
 - genová exprese \sim věk
 - *Lineární modelování (limma), Coxův model proporcionálních rizik*
- Chceme najít pravděpodobnost, že vzorek patří do určité skupiny na základě expresní hodnoty daného genu
 - *Logistická regrese*

Můžeme používat klasické statistiky u omicsových dat?

Moderovaná T-statistika

Problém ve statistickém testování omicsových dat:

**Příliš malé hodnoty exprese (blízké šumu)
vykazují malou variabilitu**

=>

**vysoké T-statistiky u biologicky nerelevantních
genů!**

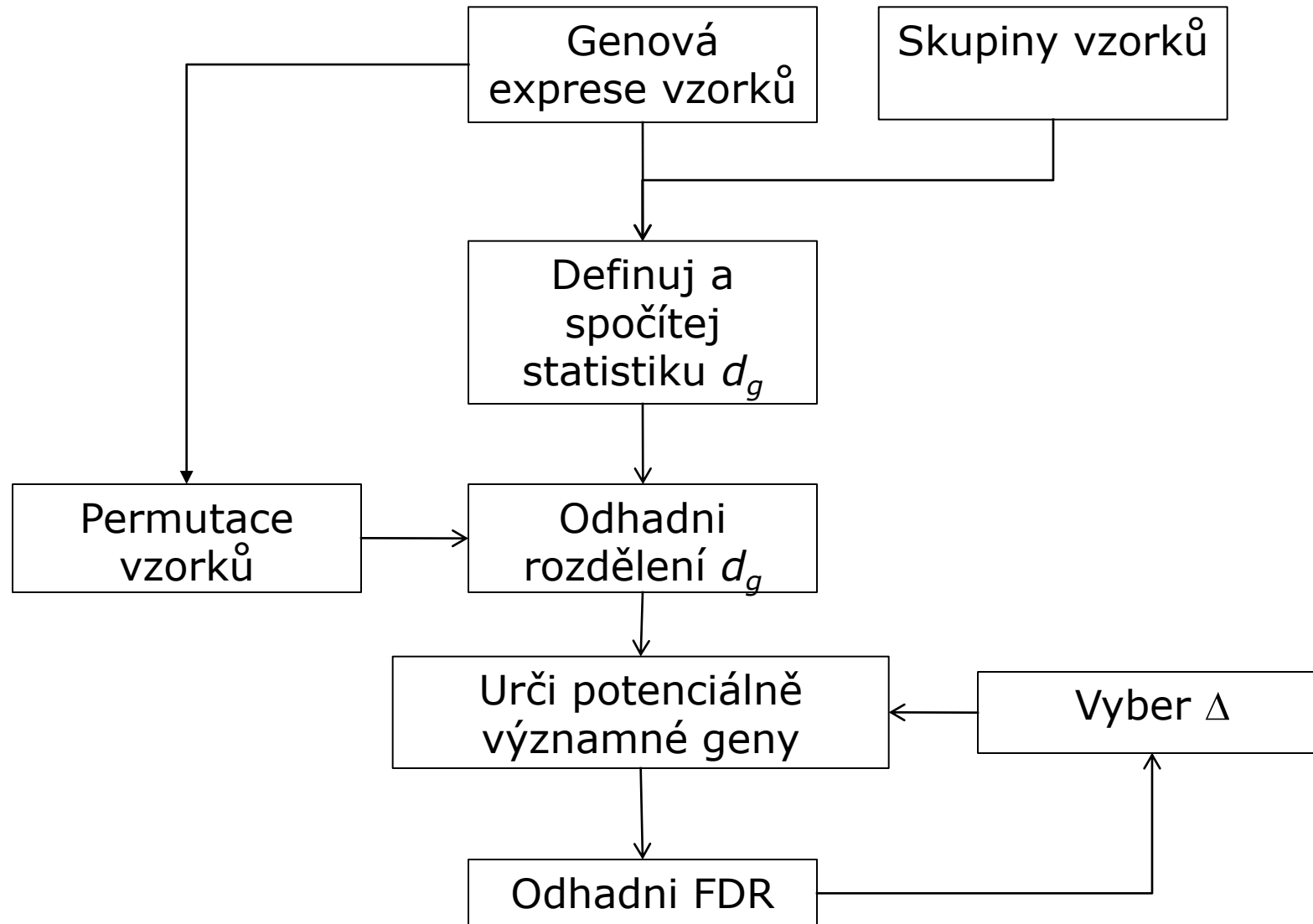
Aby se daly statistiky porovnat, je potřeba sjednotit
variabilitu:

 **Konstanta korigující
variabilitu**

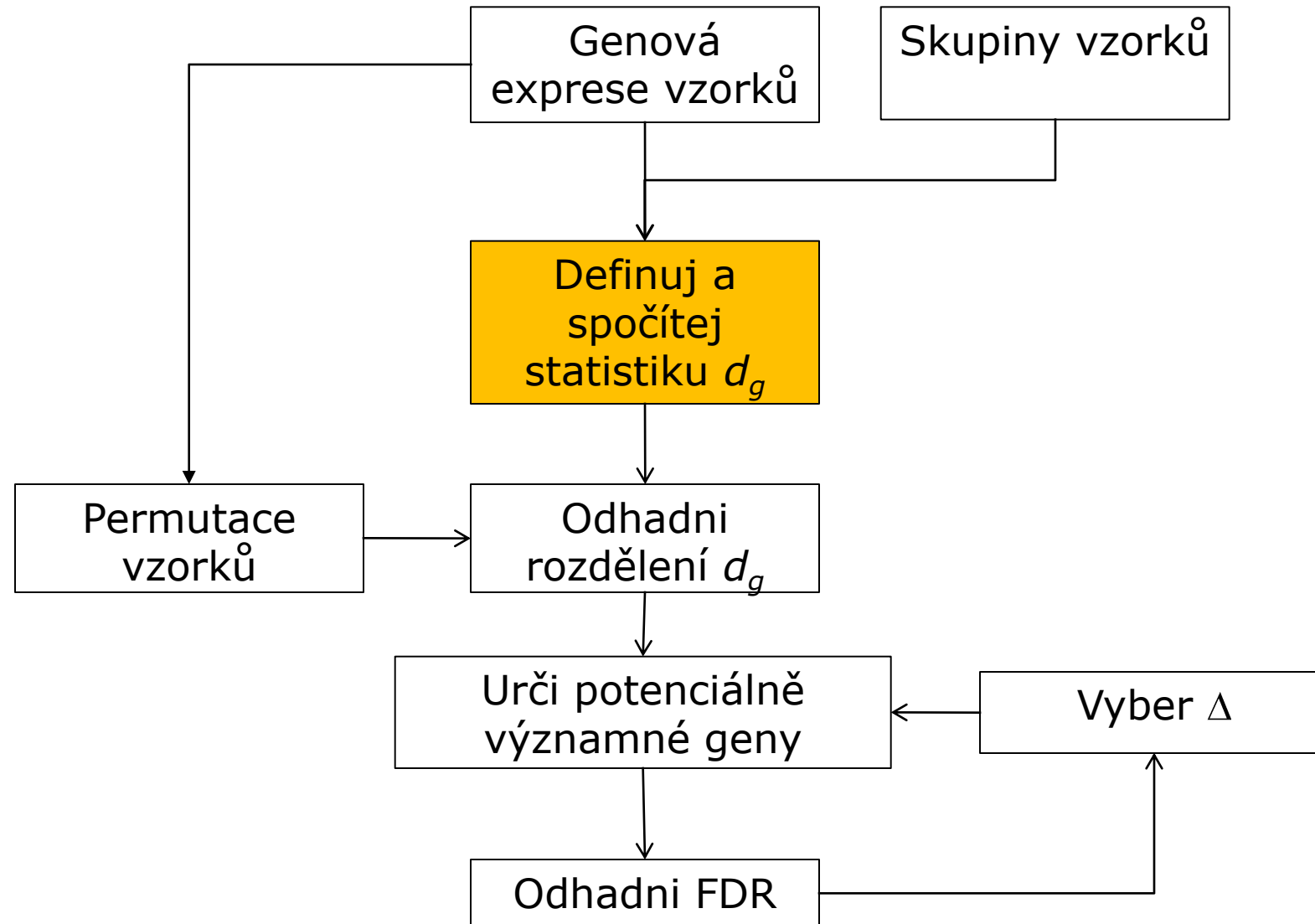
Significance analysis of microarrays (SAM)

-
-
-
-

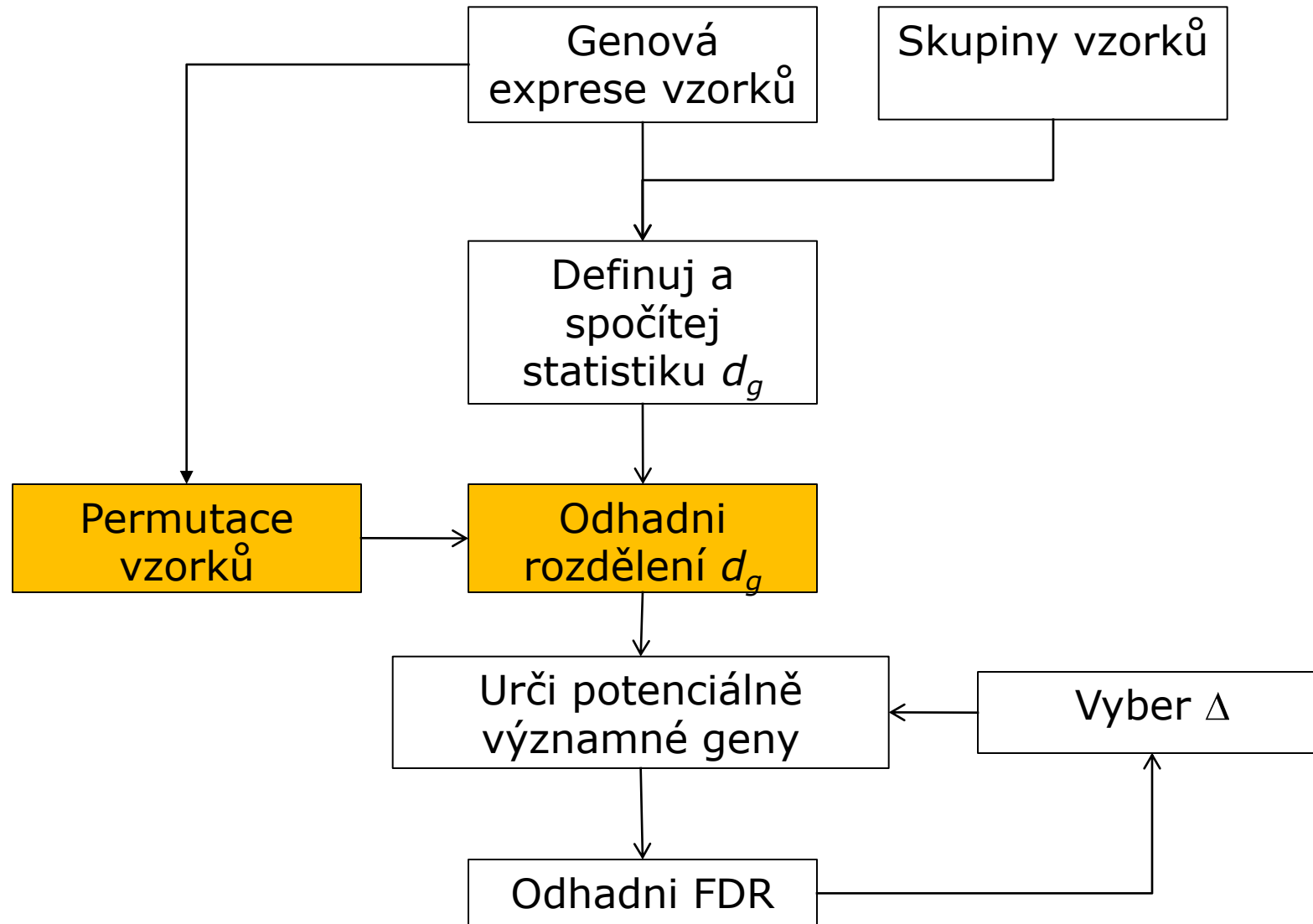
SAM - algoritmus



SAM - algoritmus



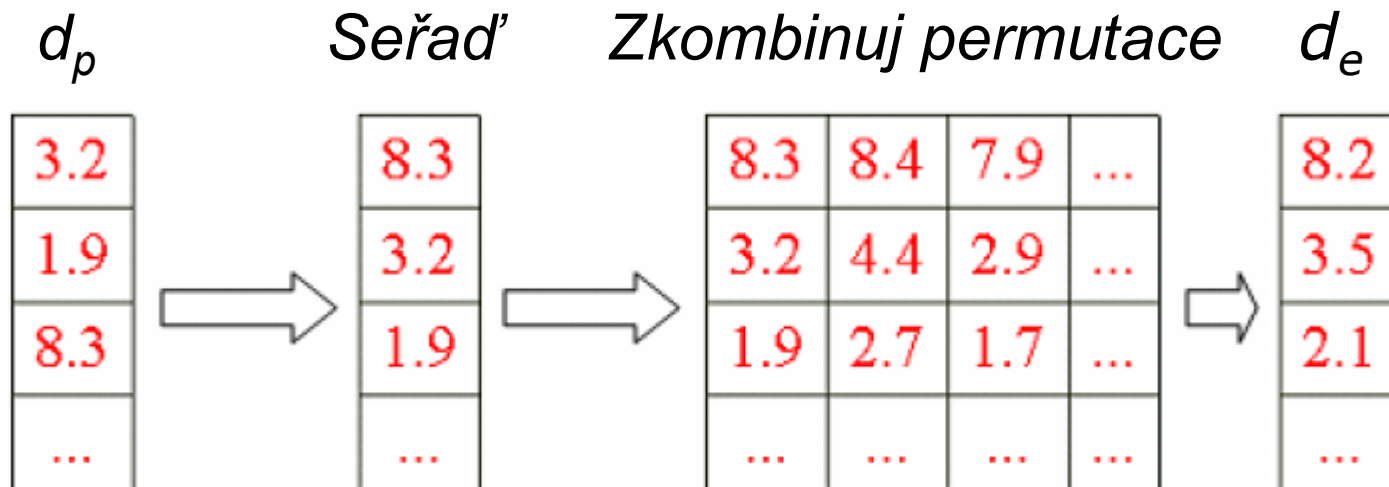
SAM - algoritmus



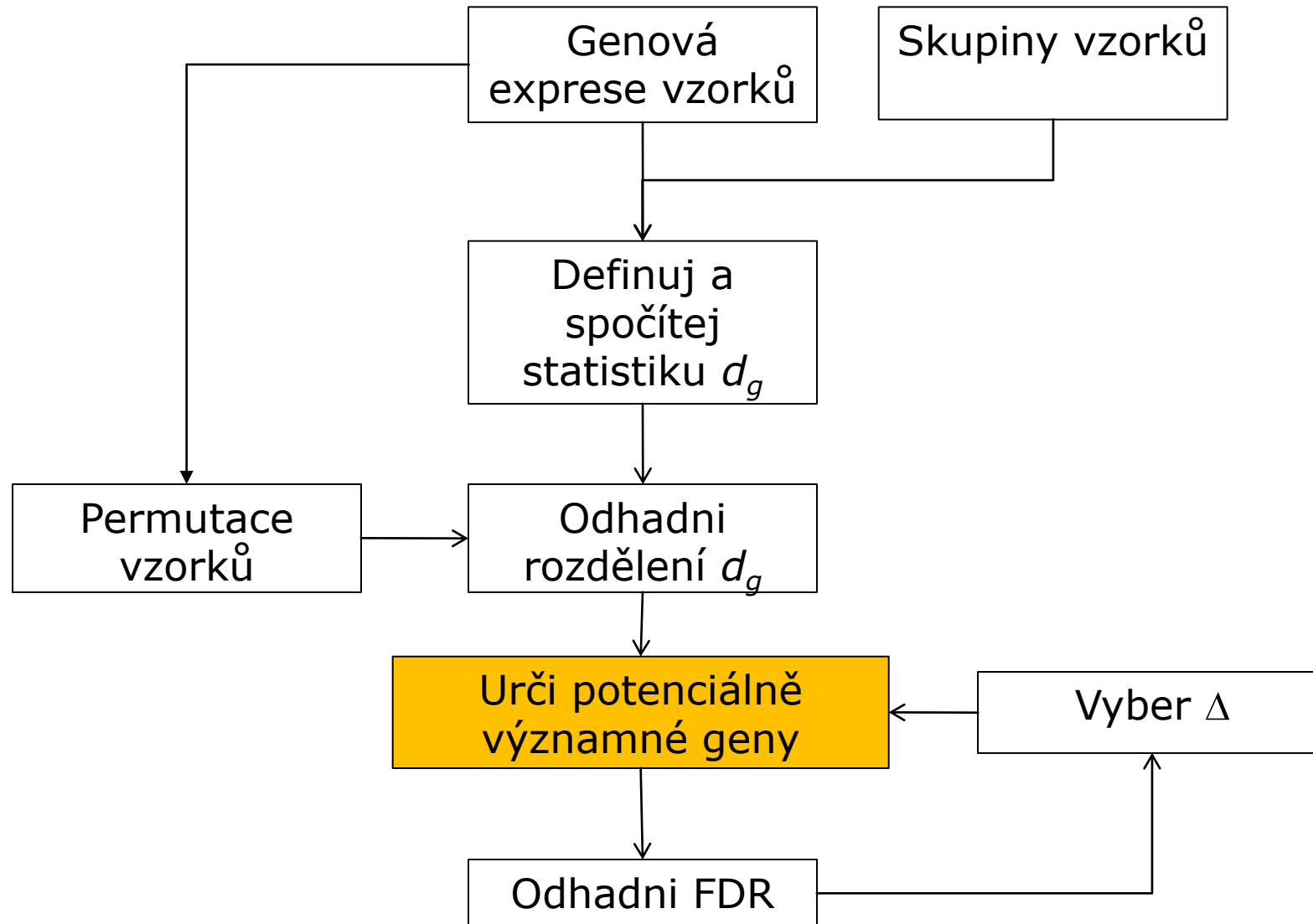
SAM - výpočet očekávaných hodnot

- Pro každou permutaci p spočítej d_{gp}
- Seřad' statistiky podle velikosti
- Definuj g -tou očekávanou hodnotu na základě N permutací

$$d_{ge} = \frac{\sum_{p=1}^N d_{gp}}{N}$$

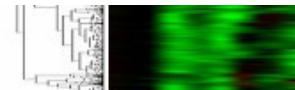


SAM - algoritmus

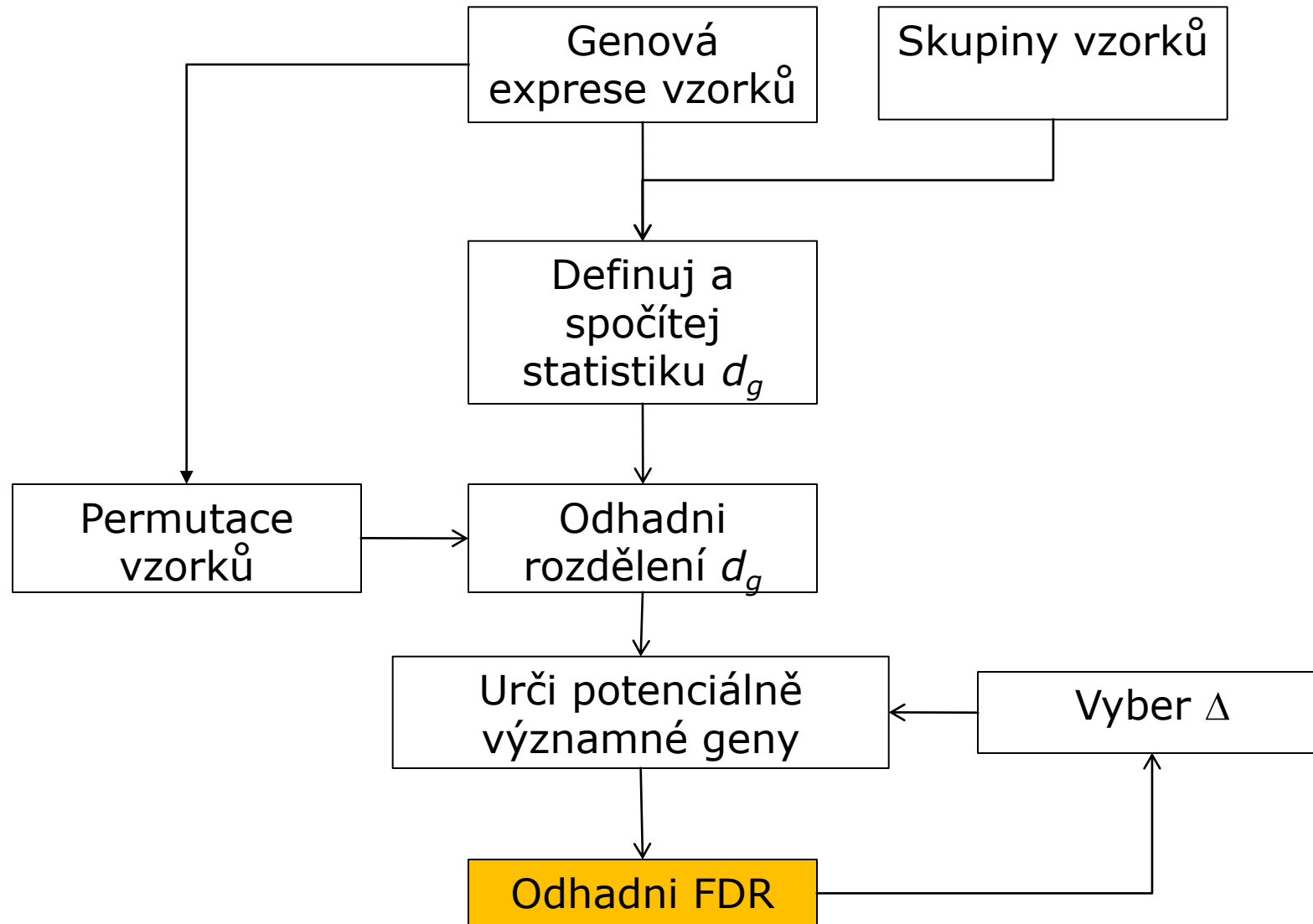


SAM – určení významných genů I

- Seřad' původní statistiky podle velikosti $d_1 \geq d_2 \geq d_3 \geq \dots$
- Nakresli graf d_g vs. d_e a definuj Δ
- Gen je statisticky významný, pokud splňuje podmínku $|d_g - d_e| > \Delta$ (označme t1 a t2 hraniční hodnoty, pro které to ještě platí)



SAM - algoritmus



SAM – výpočet FDR

- t_1 a t_2 budou použité jako hranice
- Vypočítej průměrný počet genů, které v permutacích tyto hranice překročily (byly významné)
- Odhadni počet falešně pozitivních genů v případě, že platí nulová hypotéza podělením počtem významných genů v originálním pozorování:

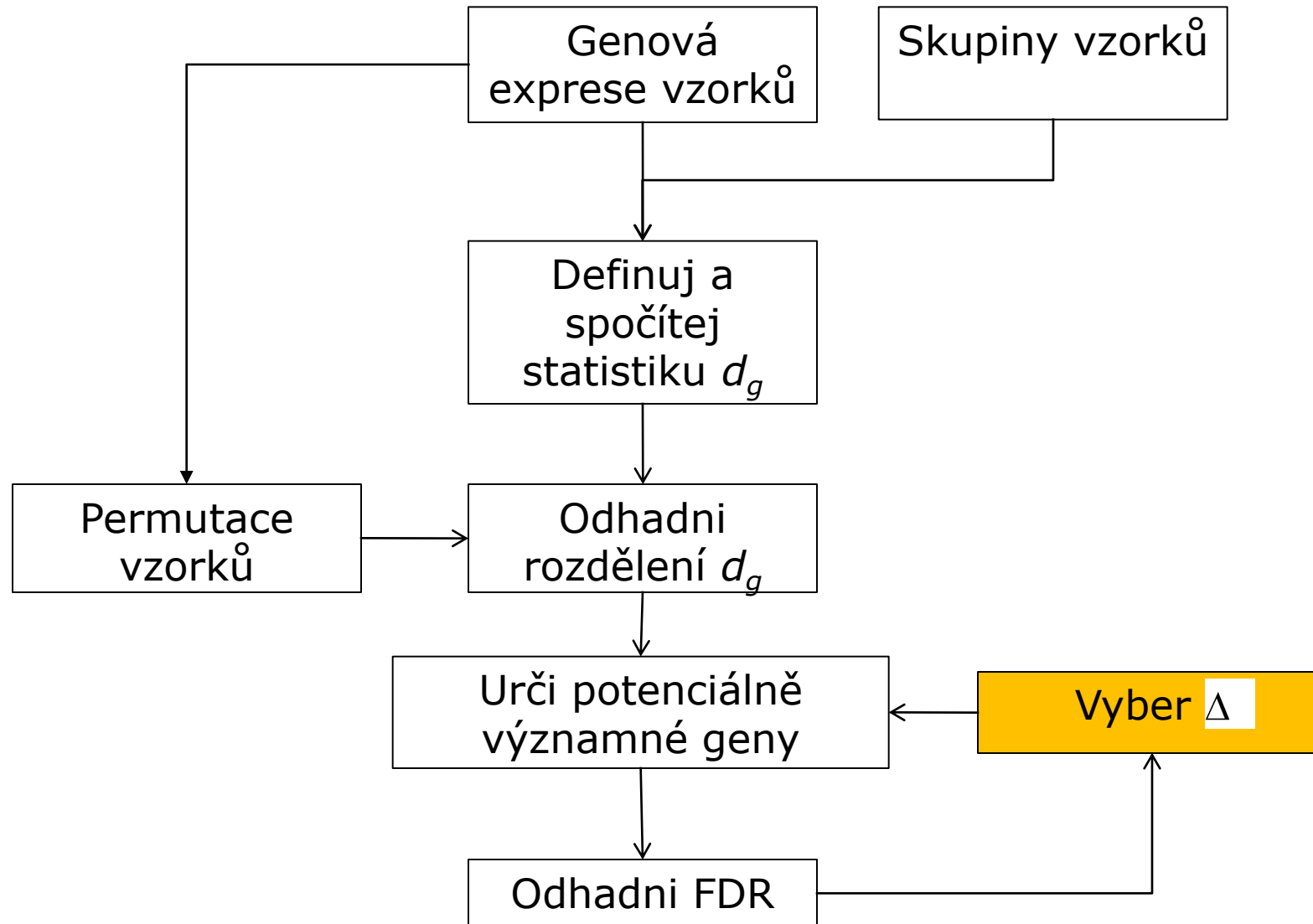
$$\text{FDR} \approx \frac{\frac{1}{N} \sum_{p=1}^N \#\{g | d_{gp} \geq t_1 \vee d_{gp} \leq t_2\}}{\#\{g | d_g \geq t_1 \vee d_g \leq t_2\}}$$

SAM – výpočet FDR, příklad

	d_g	d_p			
t_1	8.3 4.2 2.9	8.3	8.4	7.9	8.1
t_2	-0.5	3.2	4.4	2.5	1.6
		1.9	2.7	1.7	0.1
		0.3	-0.6	1.0	-2.1

$$FDR \approx \frac{7}{4} = 0.5833$$

SAM - algoritmus



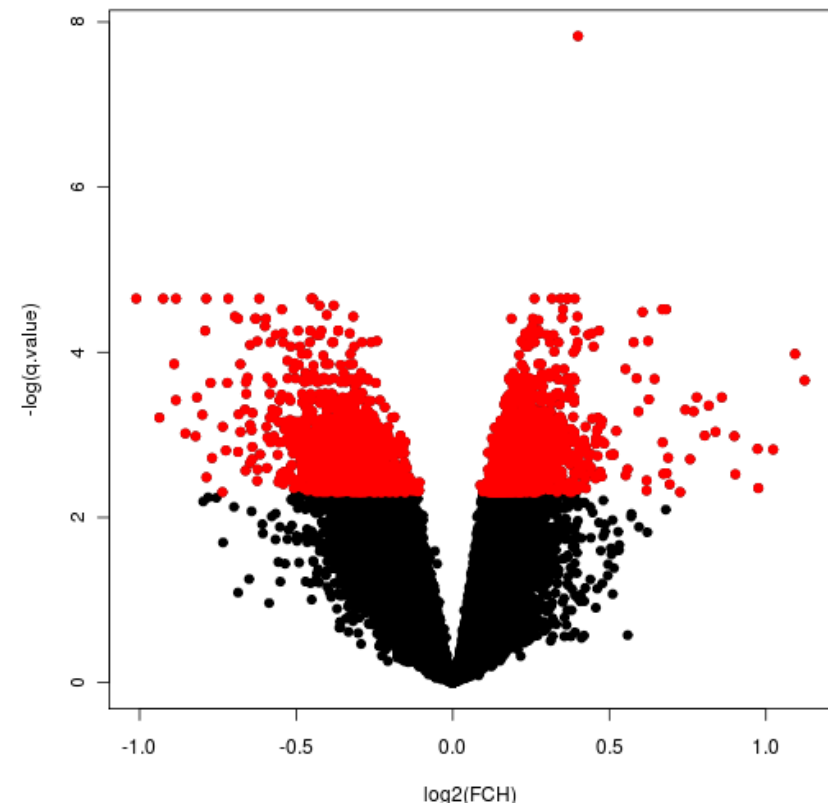
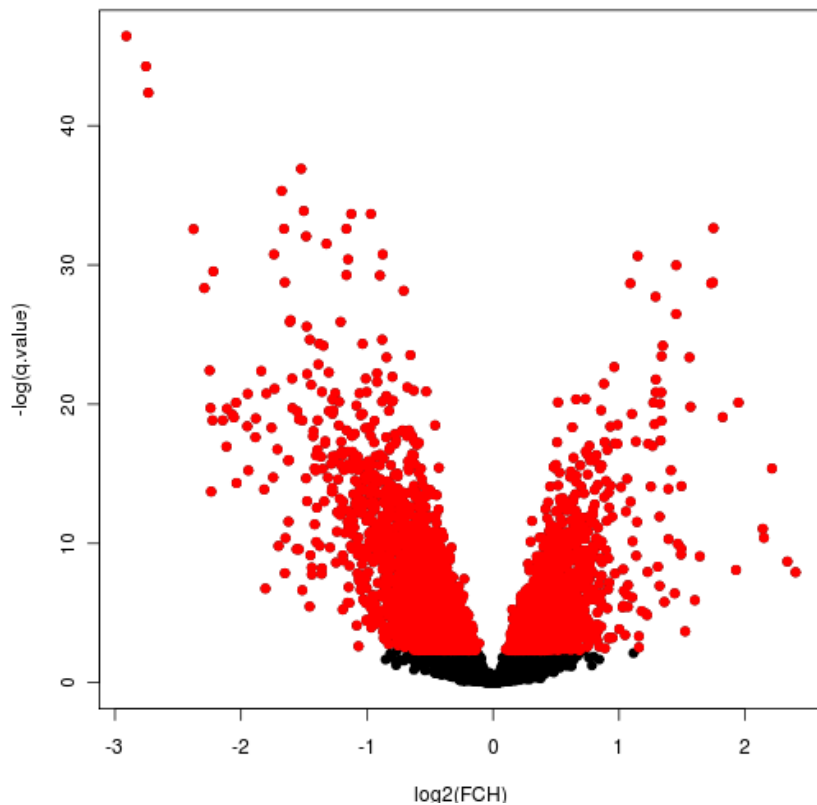
SAM – jak vybrat Δ

Parametr	Počet falešně pozitivních (z permutací)	Počet označených za významné (v orig.)	FDR
SAM			
$\Delta = 0.4$	134.9	288	47%
$\Delta = 0.5$	78.1	192	41%
$\Delta = 0.6$	56.1	162	35%
$\Delta = 0.9$	19.1	80	24%
$\Delta = 1.2$	8.4	46	18%

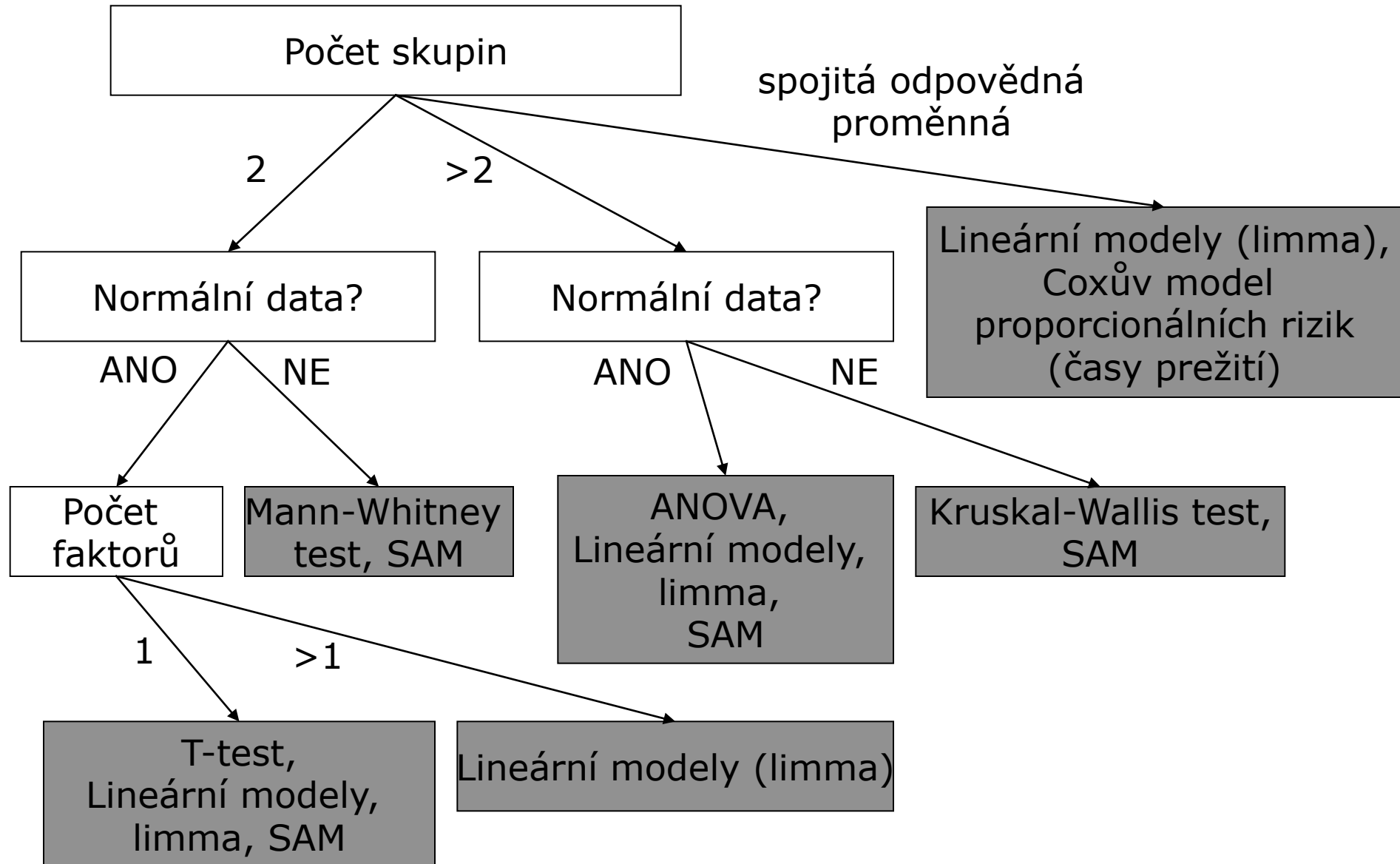
Limma

Typické zobrazení významnosti genů Volcano plot

$$-\log_{10}(\text{q-value}) \sim -\log_{10}(0.1) = 2.3$$



Porovnání skupin – schéma výběru metodiky





Cvičení

- Ve studijních materiálech k předmětu najdete soubor *CviceniPorovnaniSkupin.zip*
- Podívejte se na názvy všech souborů - dokážete určit o jaký datový soubor jde? - identifikujte zdroj dat a typ mikročipu
- Soubor odzipujte a otevřete v RStudio soubor PorovnaniSkupin.R
- Postupujte podle pokynů