MUNI | RECETOX

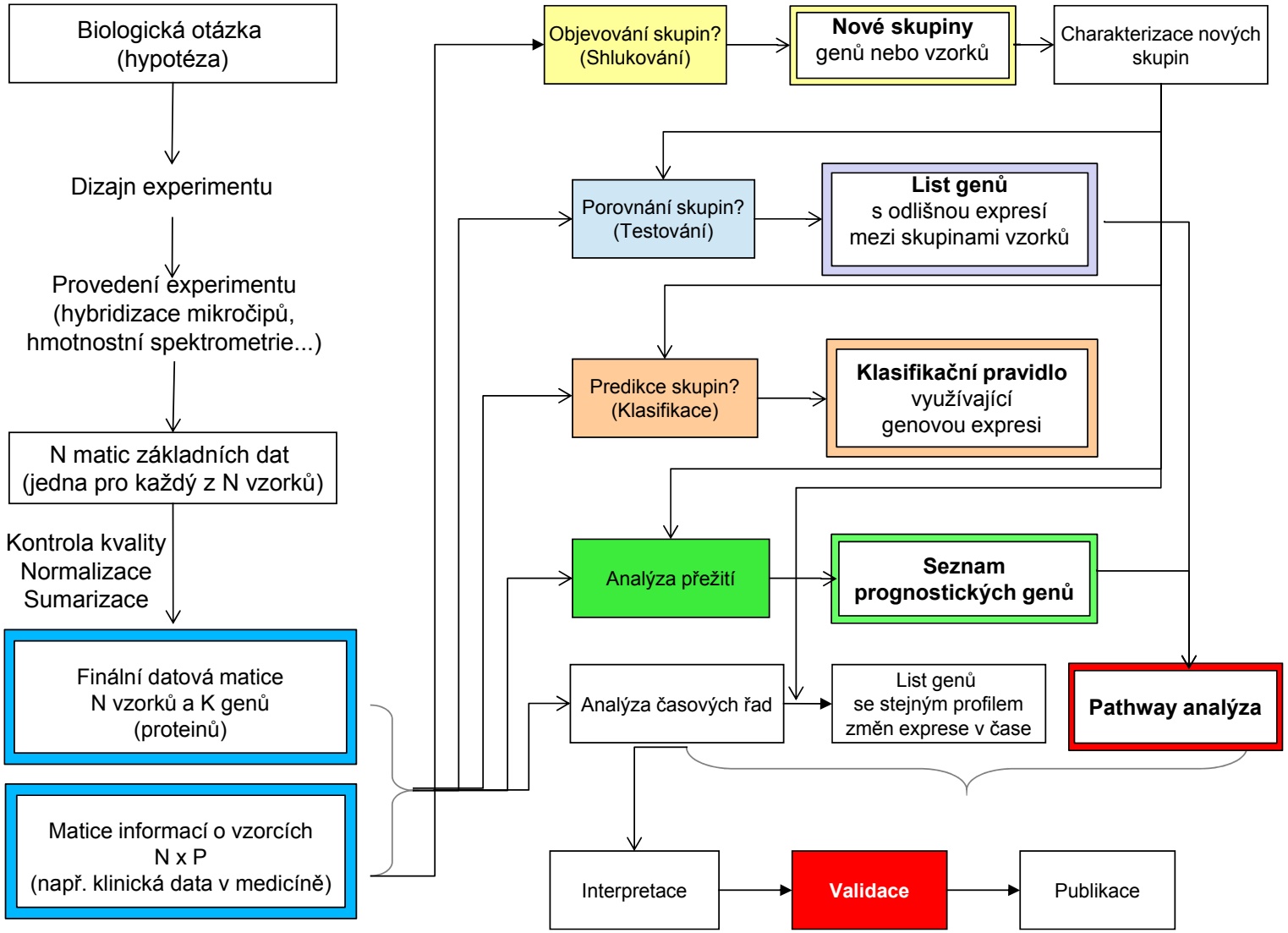
Analýza genomických a proteomických dat

Základní schéma analýzy
Porovnání skupin

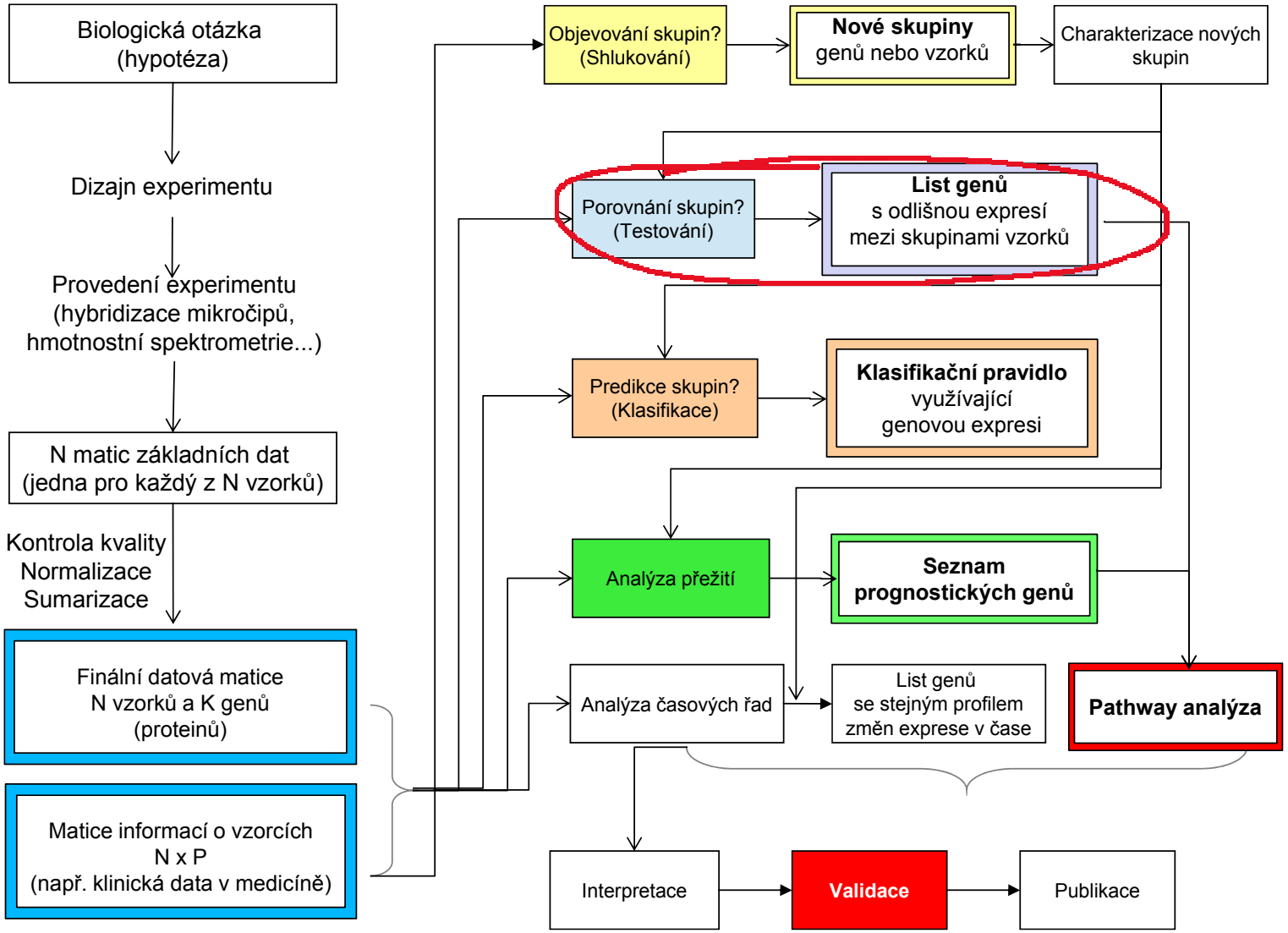
Jaro 2024
17. duben 2024
Eva Budinská (eva.budinska@recetox.muni.cz)

MUNI | RECETOX

Základní schéma analýzy genomických a proteomických dat



Základní schéma analýzy genomických a proteomických dat



MUNI | RECETOX

Porovnávání skupin

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

MUNI | RECETOX

Příklady porovnáván 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 dělící hranice velikosti efektu/změny mezi skupinami
- Testování hypotéz
- Regresní strategie

Základní metody pro porovnávání

Můžeme rozdělit do tří hlavních skupin:

Metoda dělící hranice velikosti efektu/změny mezi skupinami

Testování hypotéz

Regresní strategie

Metoda dělící hranice velikosti efektu / změny

Princip:

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

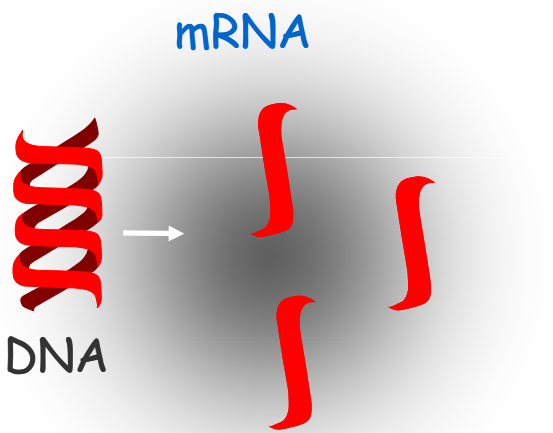
Příklad:

- genová exprese, průměr(X)/průměr(Y), kde X a Y jsou genové exprese ve skupinách, použitá dělící hranice: 2

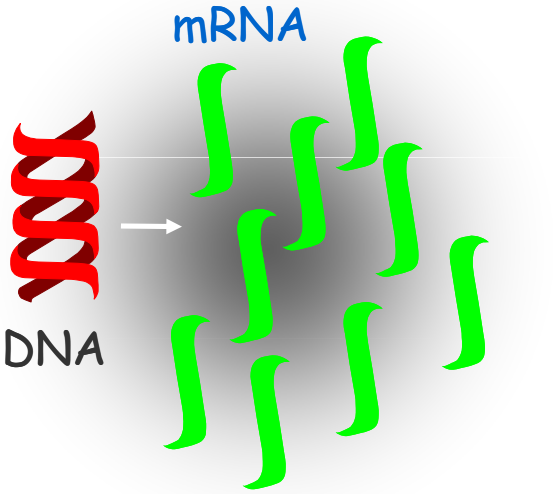
Výhoda: jednoduché

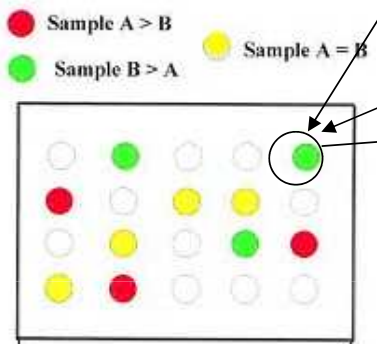
Metoda dělící hranice velikosti efektu / změny

Skupina A. Zdravá tkáň



Skupina B. Nádor





9/3 = 3

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

Metoda dělící 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 porovnávání

Můžeme rozdělit do tří hlavních skupin:

Metoda dělící hranice velikosti efektu/změny mezi skupinami

Testování hypotéz

Regresní strategie

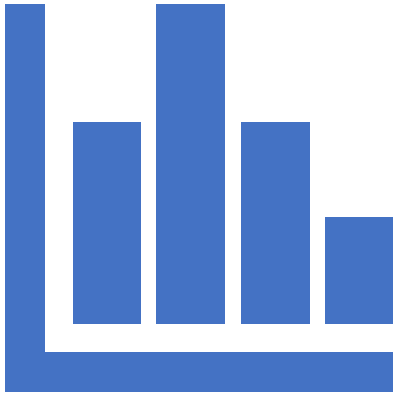
Testování hypotéz

- Testuje se:
  - *Nulová hypotéza (H_0)*:
    Gen / protein není odlišně exprimovaný mezi skupinami
- proti
  - *Alternativní hypotéza (H_1)*:
    Gen je odlišně exprimovaný mezi skupinami

- Na základě dat musíme rozhodnout, co je pravda

- Nulovou hypotézu **zamítneme** jen pokud existuje ***dostatečně silná evidence***, že je neplatná
  - Evidence – statistika a p-hodnota!

Co je to *statistika*



- Abychom rozhodli, která hypotéza je pravdivá, sumarizujeme data do **jednoho čísla**
- V testovaní hypotéz se toto číslo nazývá ***statistika*** *(T-statistika, Z-statistika, F-statistika...)*

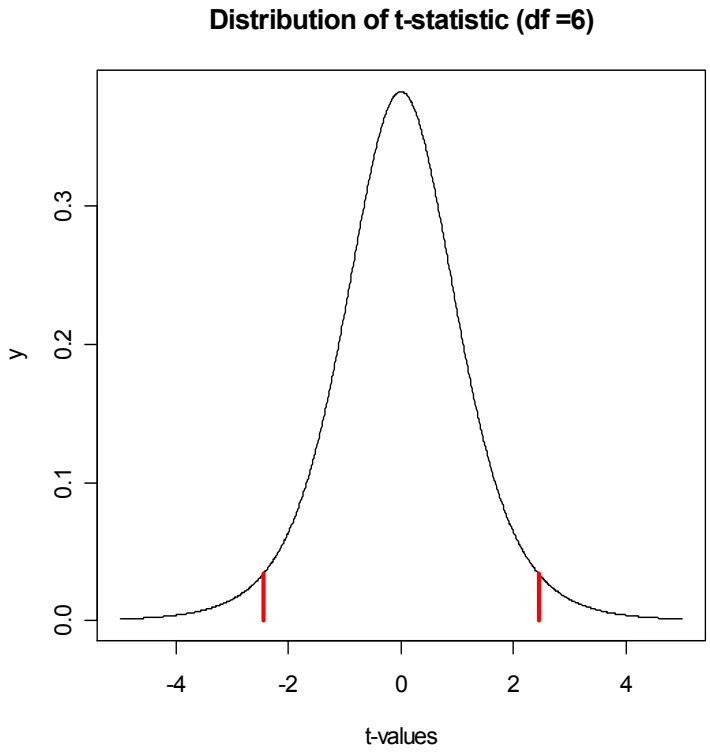
- Statistiky jsou definovány různě a mají různé předpoklady.
- Například T-statistika porovnává signál se šumem a předpokládá normalitu dat.

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

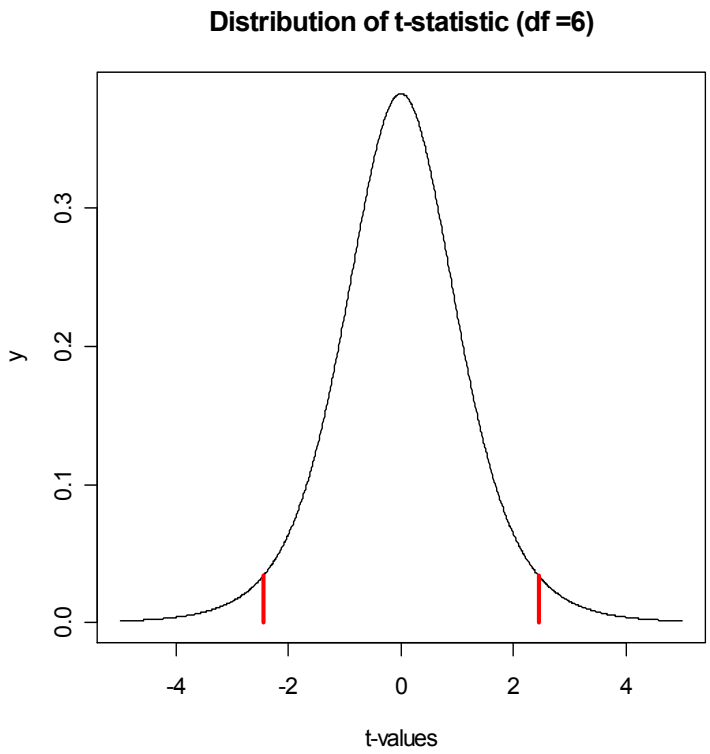


- Dvouvýběrový T-test pro porovnání rovnosti dvou průměrů *μ1, μ2*:
  - Průměr exprese genu ve skupině 1 vs. průměr ve skupině 2

$$T_g = \frac{\mu_{g1} - \mu_{g2}}{s_g \sqrt{\frac{1}{n_1} - \frac{1}{n_2}}}$$

Variabilita (vyjádřená jako směrodatná odchylka)

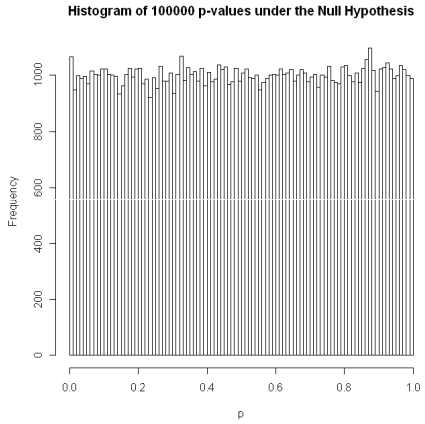
T-test a T-statistika

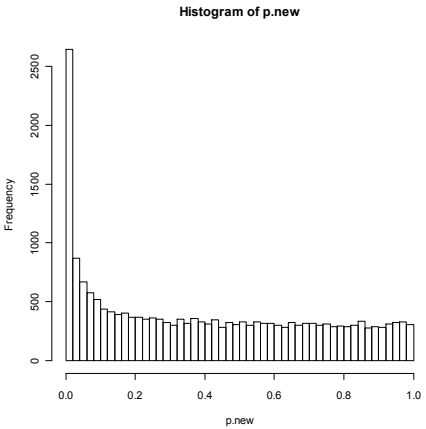


- Pokud data mají **normální rozložení a neexistuje rozdíl mezi skupinami**, tak T-statistiky pocházejí z **T-rozložení**.
- **p-hodnota** = pravděpodobnost že dostaneme danou hodnotu T-statistiky nebo hodnotu větší, v případě, že neexistuje rozdíl mezi skupinami

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

- 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 mnohonásobného 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 x 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 pozitivity): FWER = Pr(FP > 0)

1. **False discovery rate (FDR)**(Benjamini & Hochberg,1995): Očakávaný podíl falešně pozitivních výsledků mezi zamítnutými hypotézami

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

Korekce p-hodnot při mnohonásobném testování

Kontrolujeme FWER

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

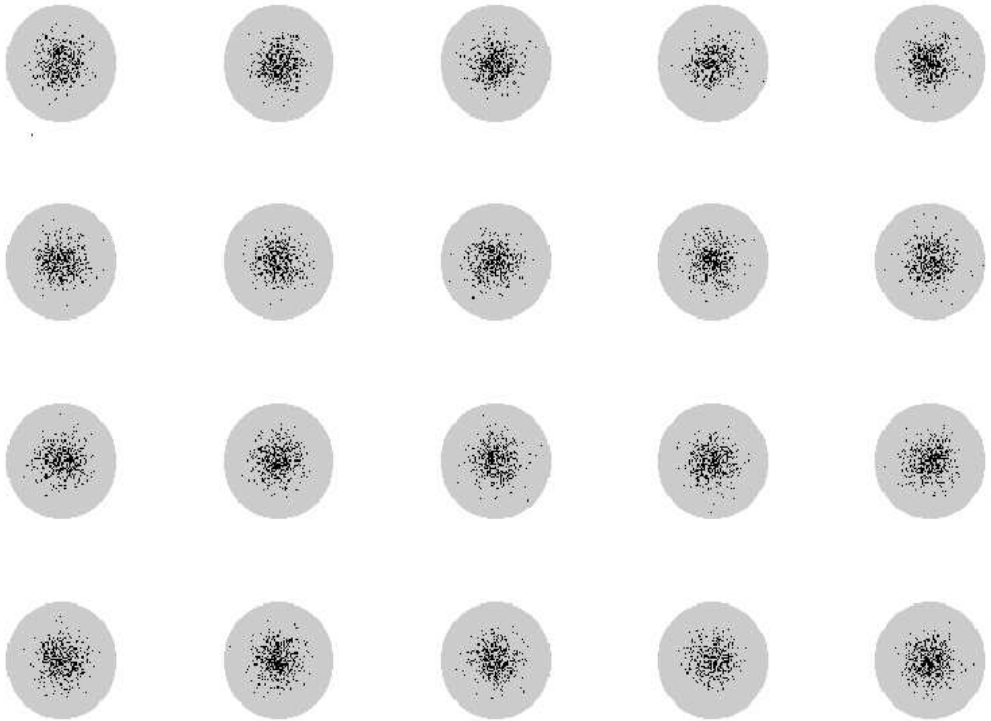
Kontrolujeme FDR

- 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)

! Existuje více druhů metod pro kontrolu FDR!

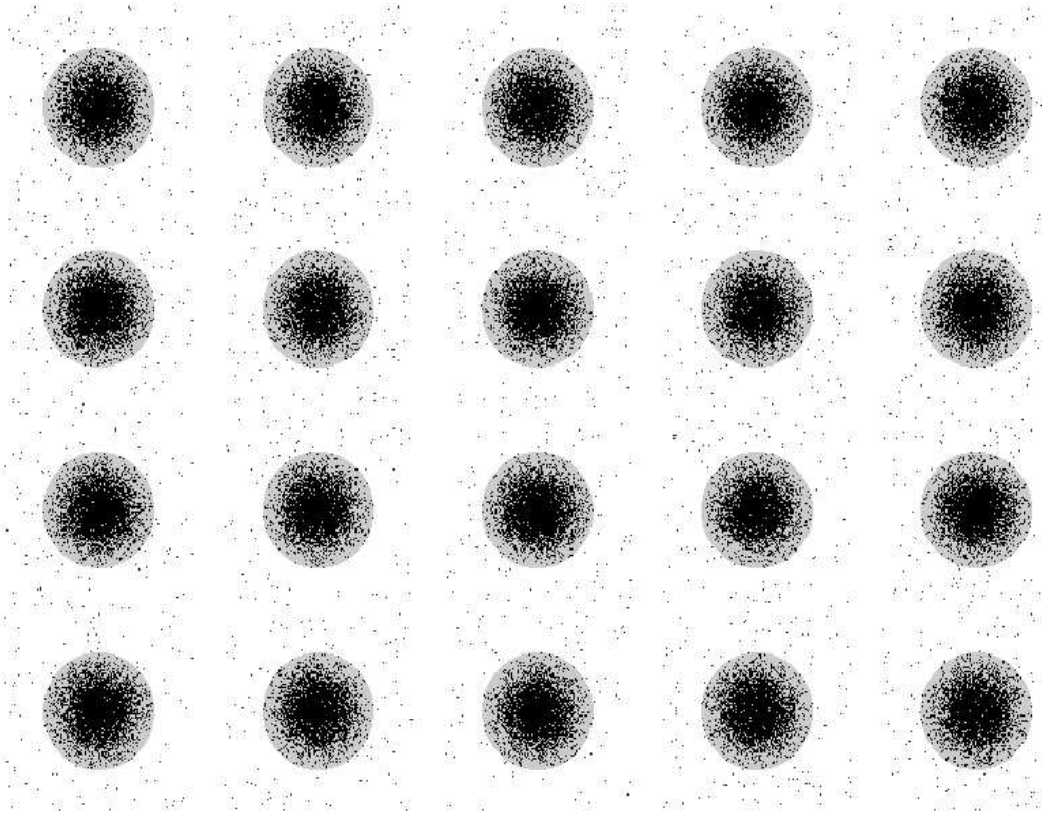
Který typ korekce použít?

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

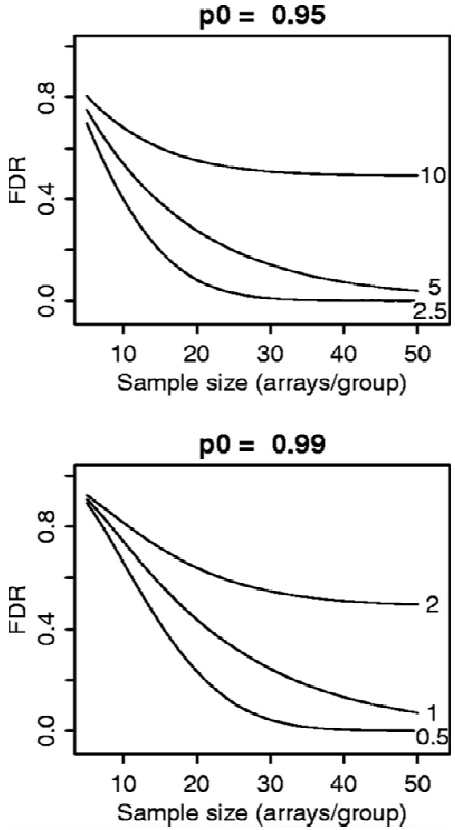


Který typ korekce použít?

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



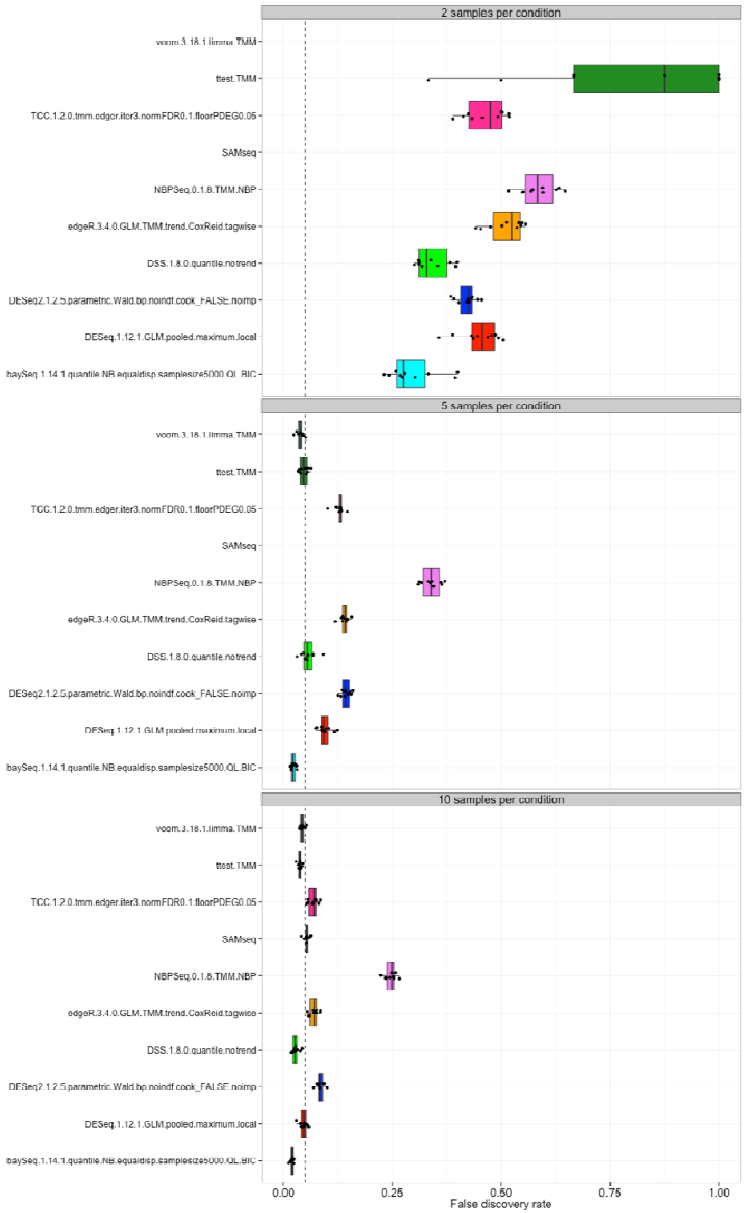
Vliv počtu vzorků na falešně pozitivní výsledky



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

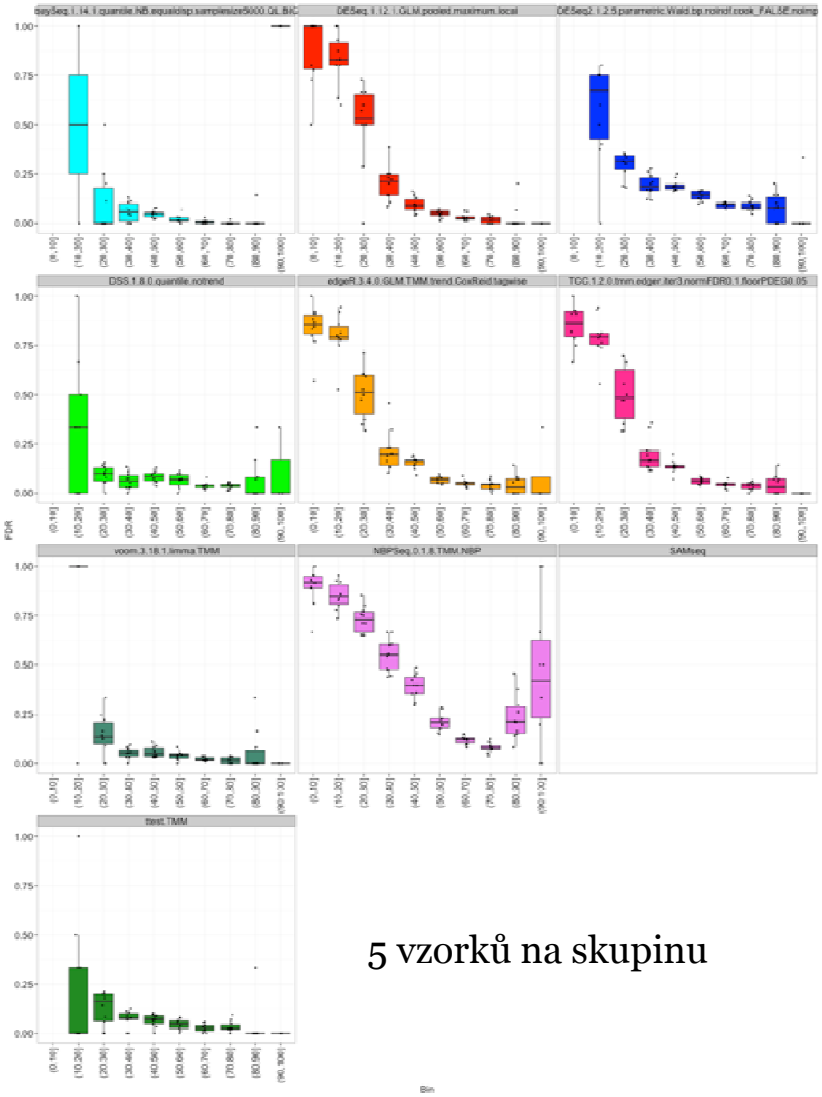
FDR (false discovery rate) jako funkce velikosti vzorku a percenta významných výsledků.
Každá křivka představuje fixní percento genů označených jako významných.

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

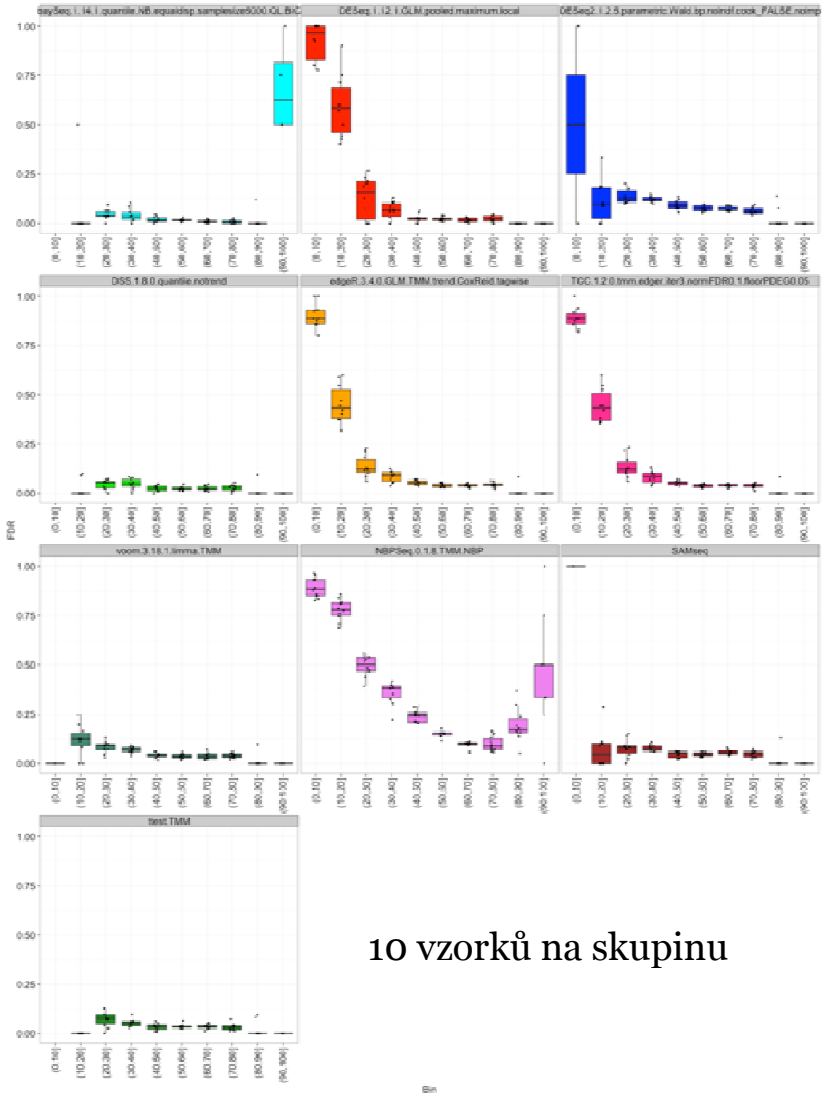


FDR (False discovery rate) jako funkce počtu vzorků na skupinu a metody použité pro normalizaci sekvenačních dat a testování hypotéz

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

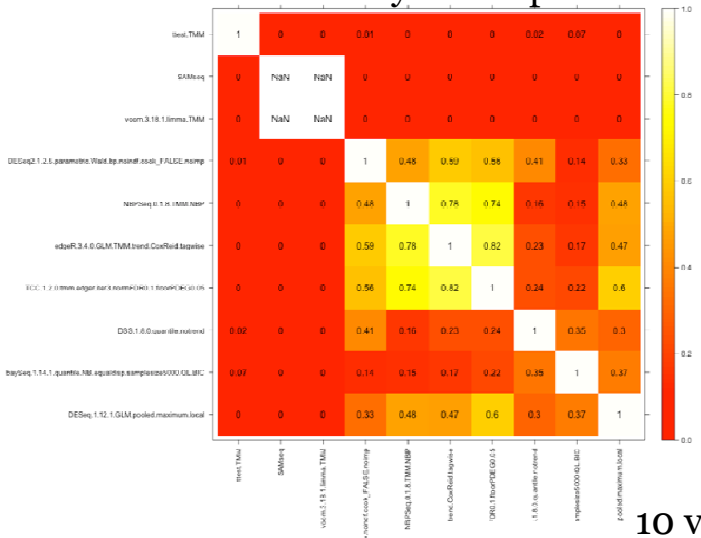


5 vzorků na skupinu

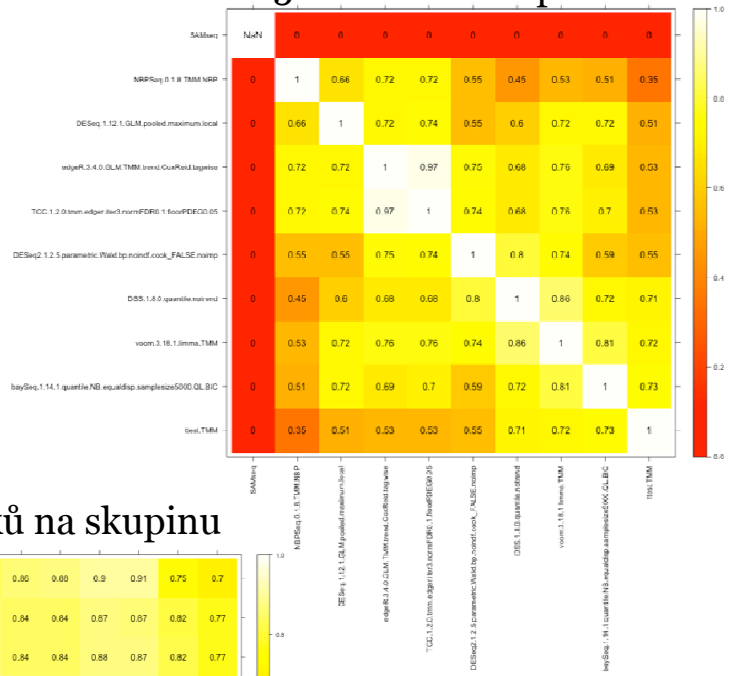


10 vzorků na skupinu

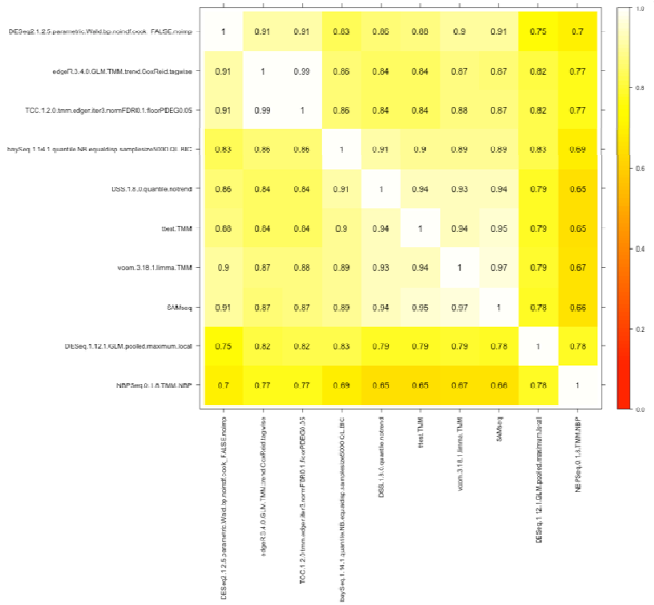
2 vzorky na skupinu



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 dělící 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 ~ skupina + pohlaví
  - *Lineární modelování (limma)*

- Pokud se snažíme zjistit, jak velmi se genová exprese změní, pokud se změní hodnota nějaké *spojité proměnné*
  - genová exprese ~ prežití
  - genová exprese ~ 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

$$T_g = \frac{\mu_{g1} - \mu_{g2}}{s_g}$$

$$\mu_{g1} = 2, \mu_{g2} = 2.5,$$

$$s_g = 0.02$$

$$=> T_g = -25$$

Problém ve statistickém testovaní 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:

$$d_g = \frac{\mu_{g1} - \mu_{g2}}{s_g + s_0}$$

Konstanta korigující variabilitu

Significance analysis of microarrays (SAM)

- Tusher, Tibshirani a Chu (2001)
- Založená na moderované *t*-statistice (d_g), počítá FDR

$$d_g = \frac{\mu_{g1} - \mu_{g2}}{s_g + s_0}$$

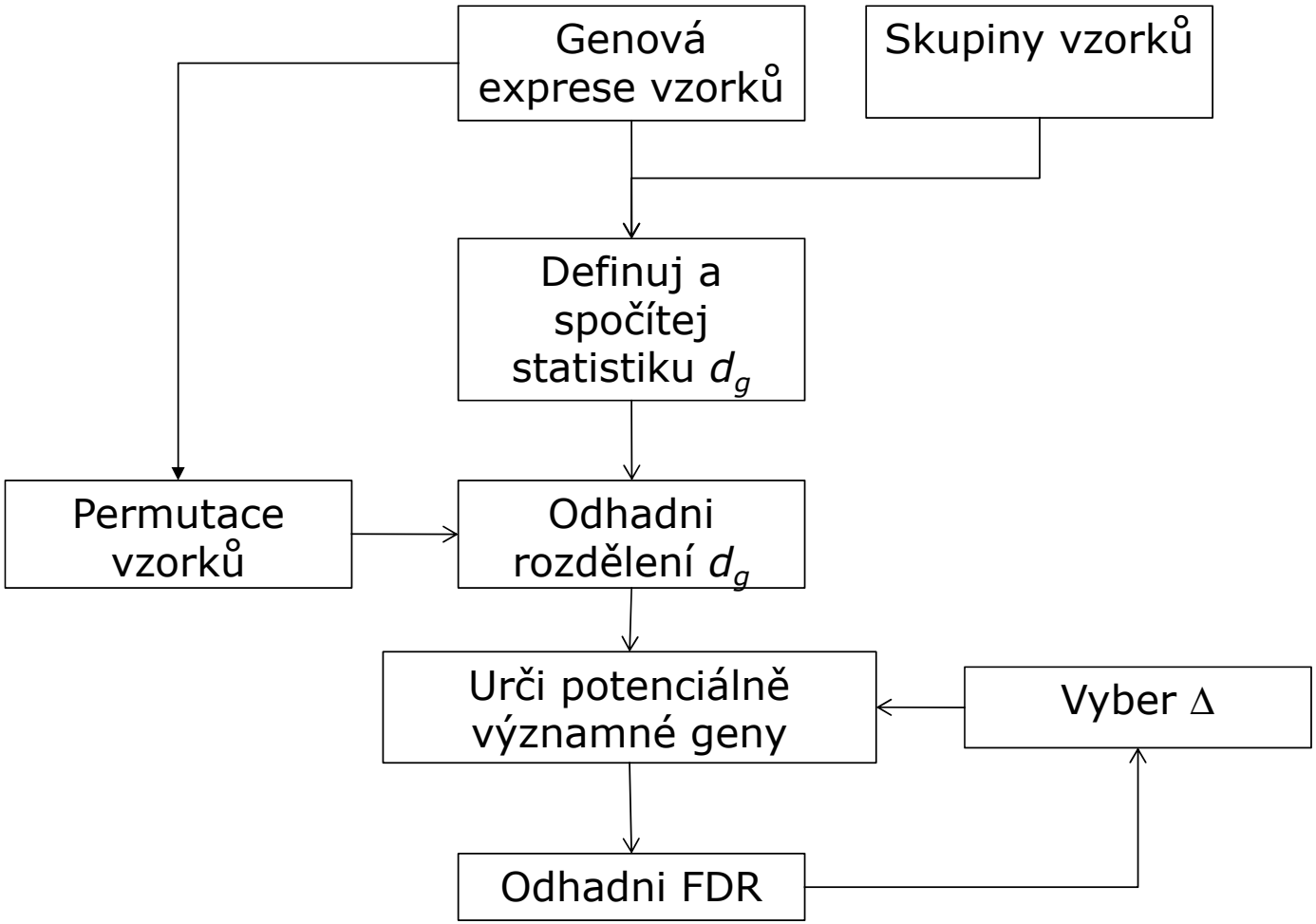
- Statistická významnost d_g je následně stanovená permutacemi původních dat a kalkulací očekávaného skóre v případě, že platí nulová hypotéza (d_e)
- Gen je statisticky významný, pokud splňuje podmínku $|d_g - d_e| > \Delta$.
- Výhody: jednoduché

  Nevýhody: výpočtově náročné (permutace)

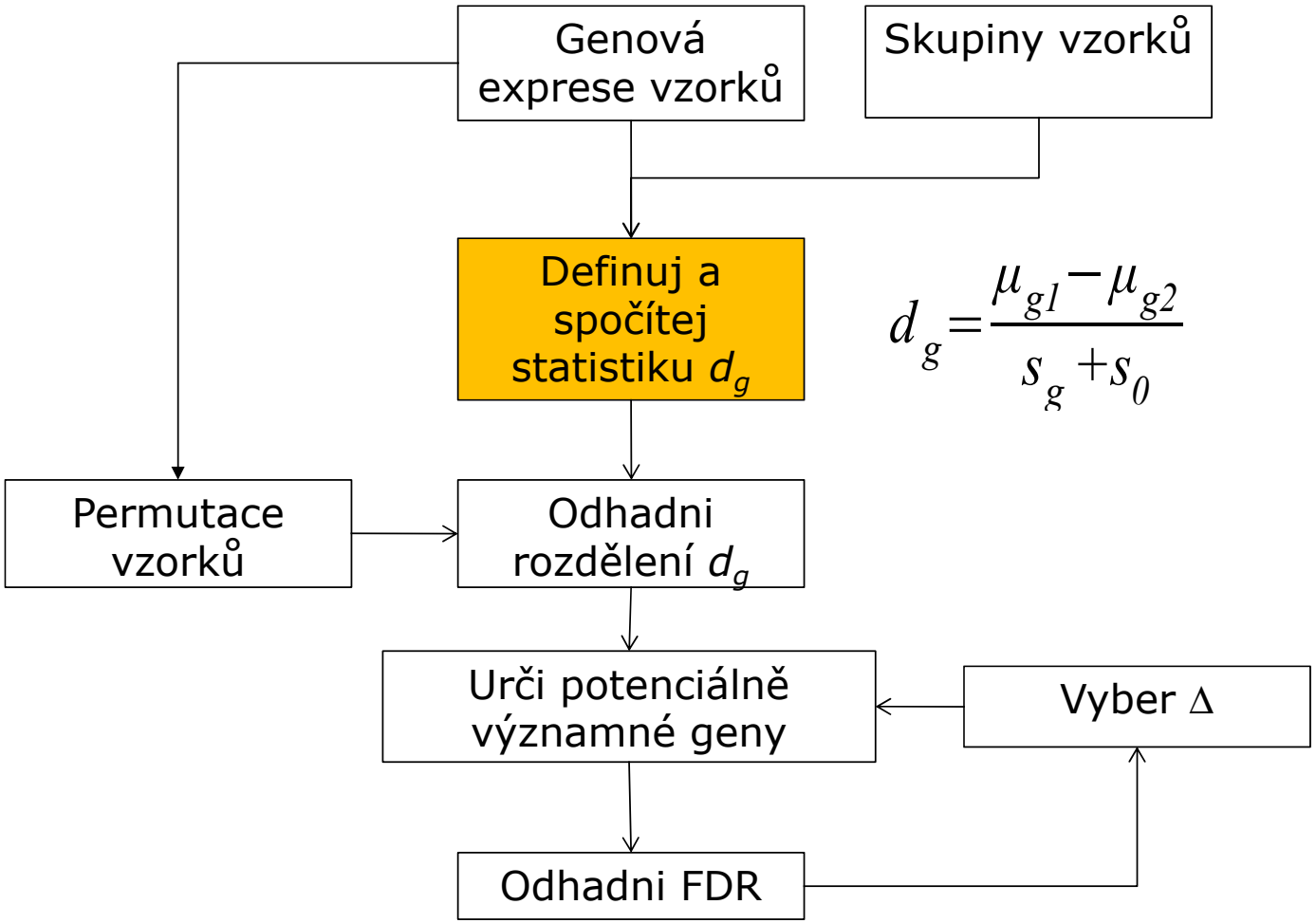
  Výstup: *q*-hodnoty

```
biocLite("samr")
library(samr)
```

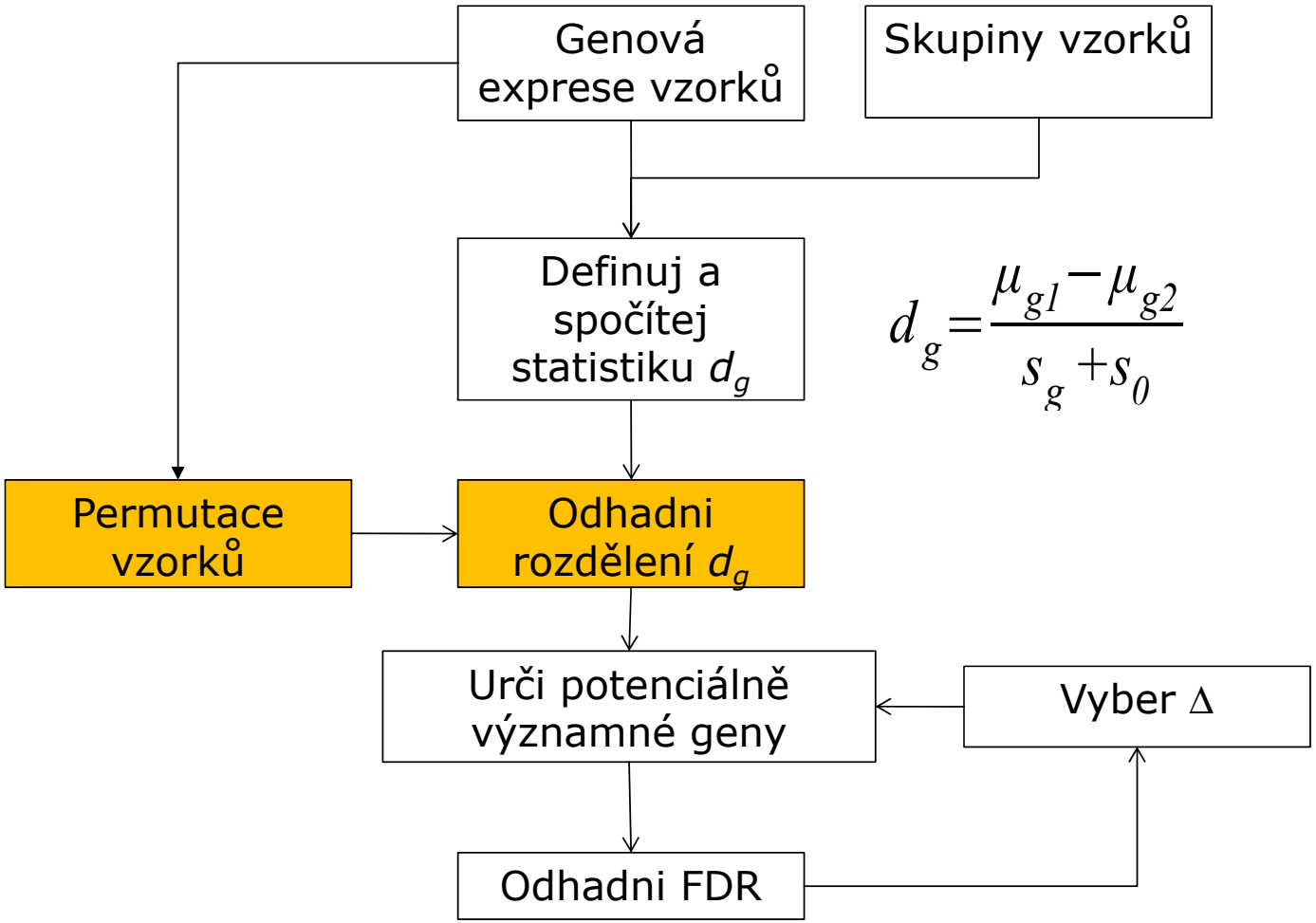
SAM - algoritmus



SAM - algoritmus



SAM - algoritmus



SAM - výpočet očekávaných hodnot

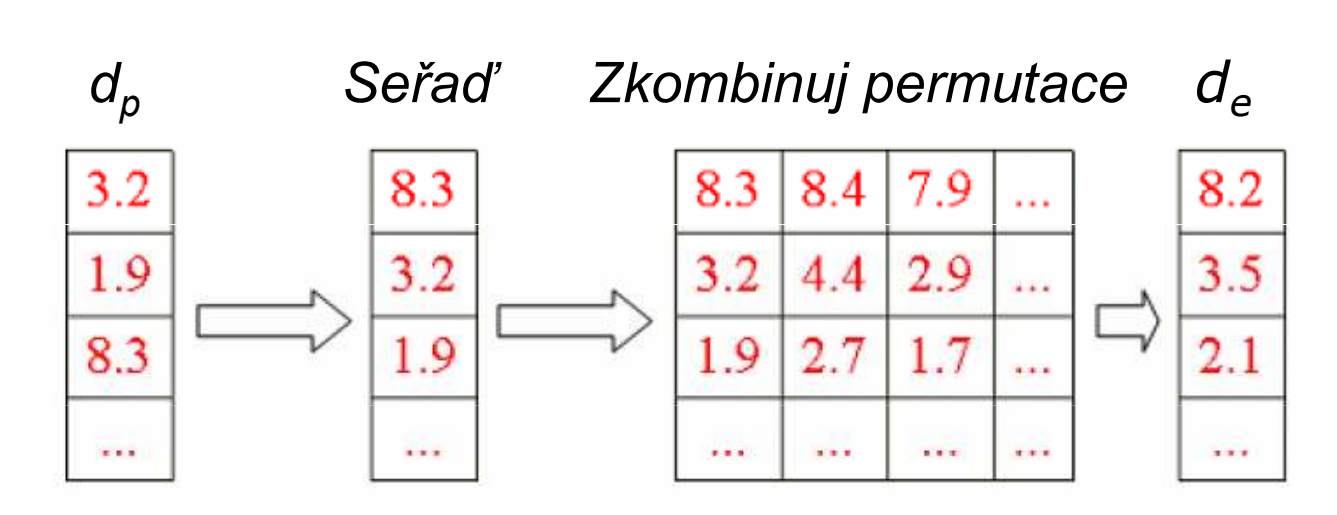
- Pro každou permutaci p spočítej d_{gp}

$$d_{gp} = \frac{\mu_{g1} - \mu_{g2}}{s_g + s_0}$$

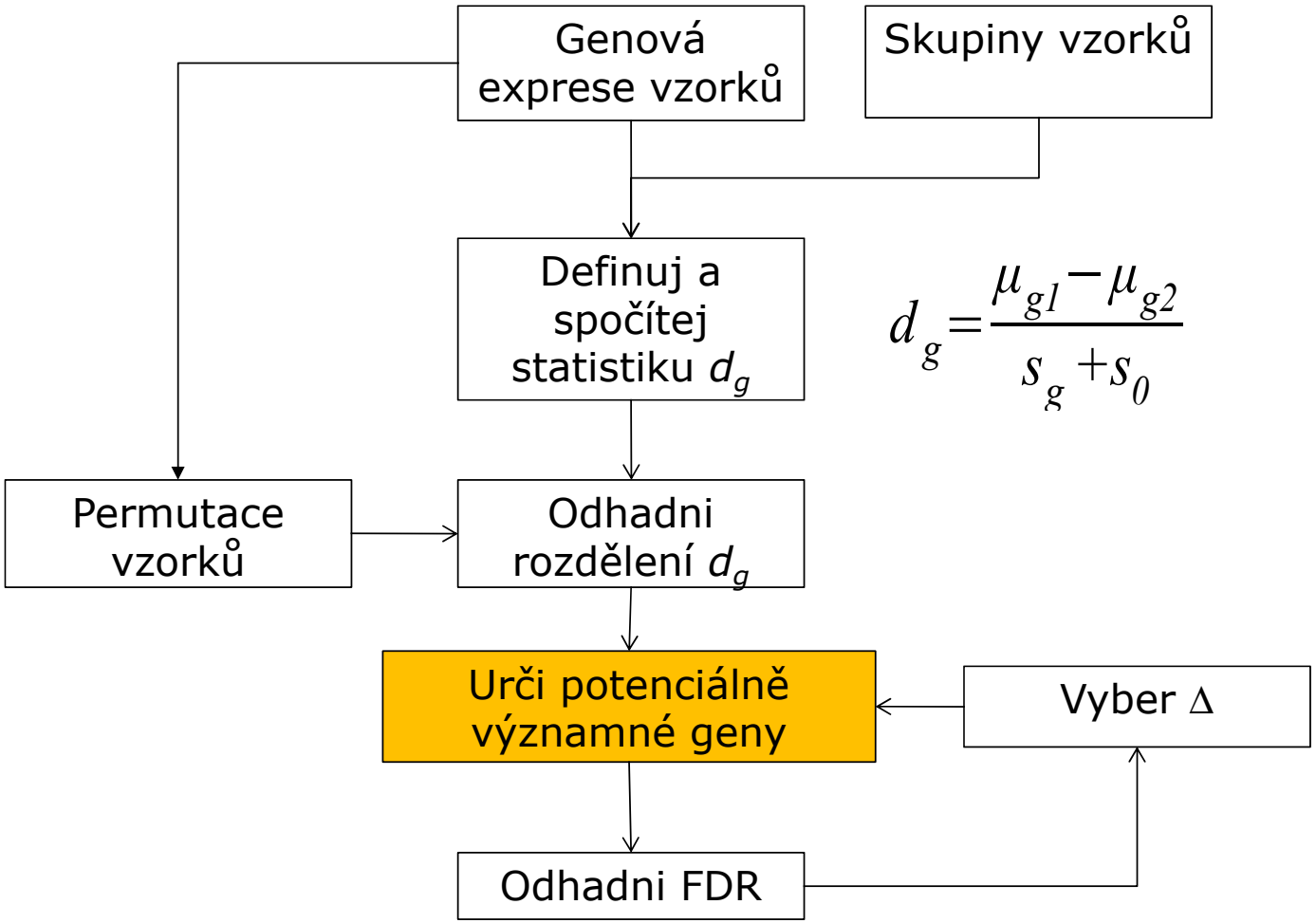
- Seřaď statistiky podle velikosti

- Definuj g-tou očekávanou hodnotu na základě N permutací

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

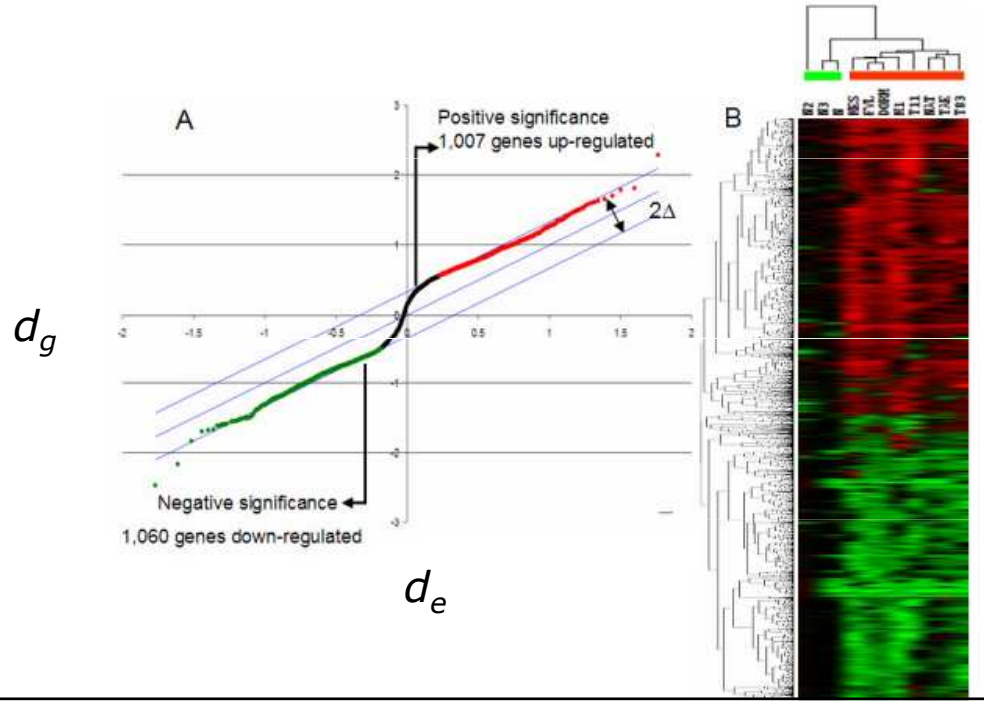


SAM - algoritmus

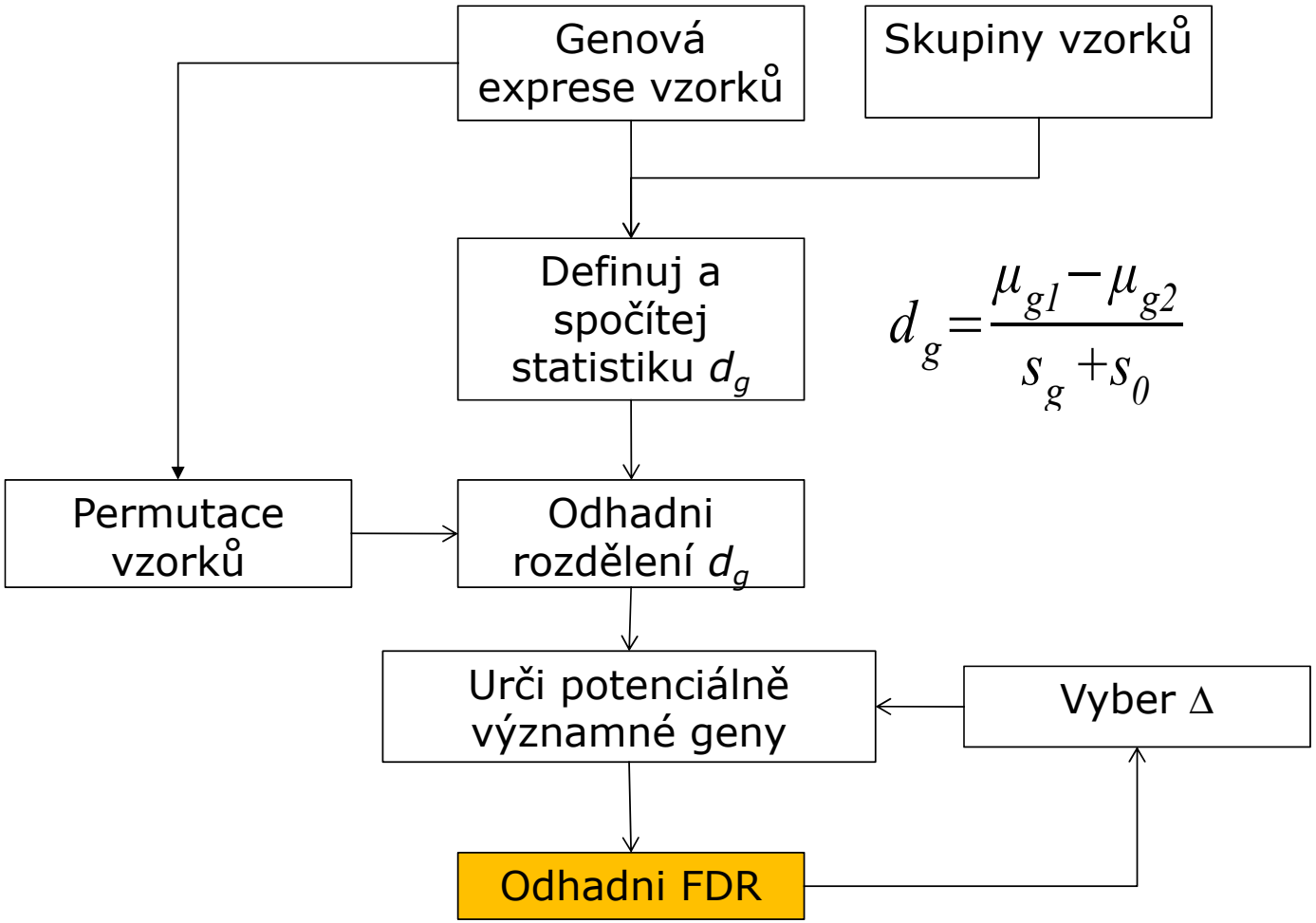


SAM – určení významných genů I

- Seřaď původní statistiky podle velikosti $d_1 \geq d_2 \geq d_3 \geq \ldots$
- 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

- t1 a t2 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:

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

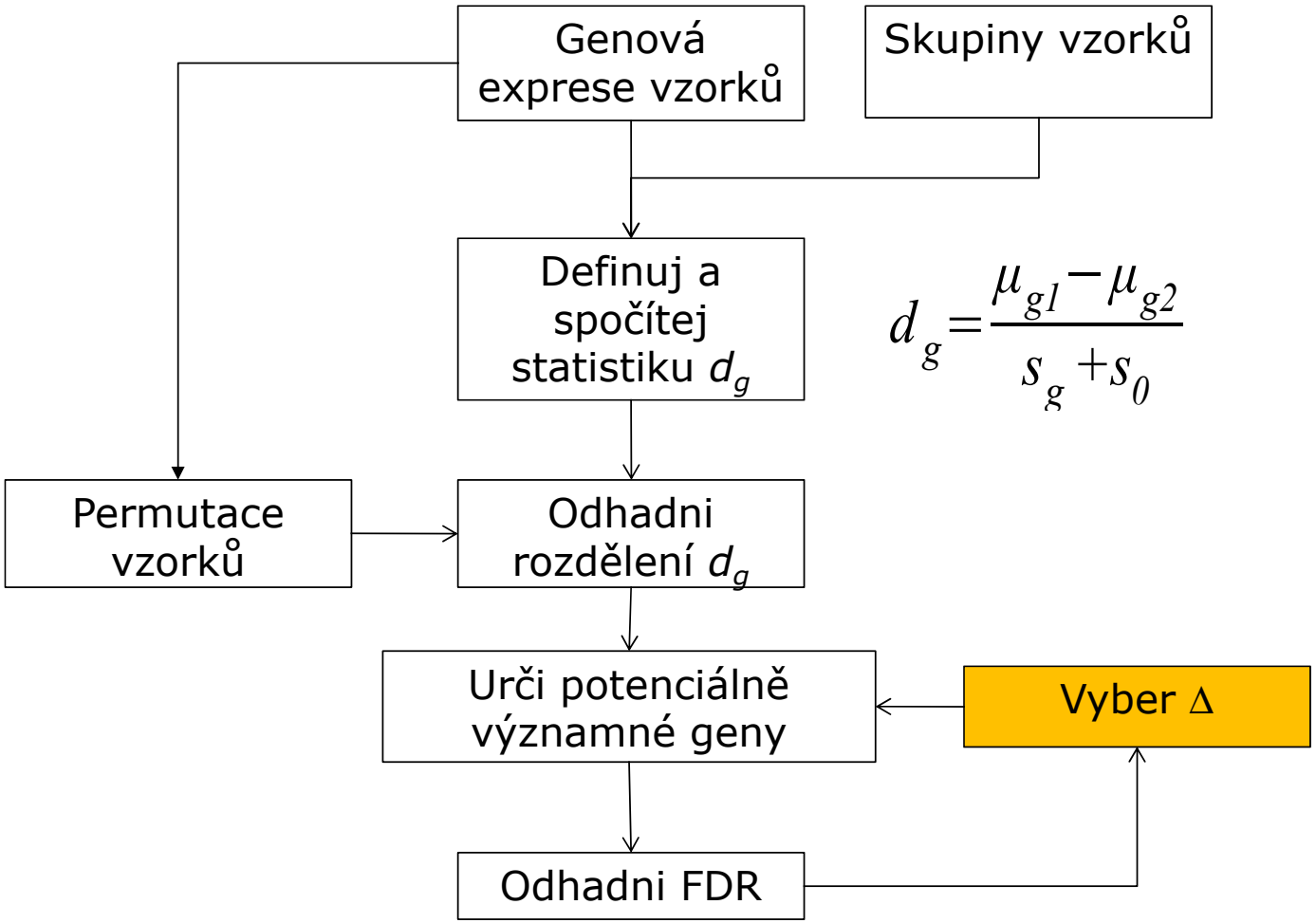
SAM – výpočet FDR, příklad

	d_g
	8.3
t_1	4.2
	2.9
t_2	−0.5

d_p			
8.3	8.4	7.9	8.1
3.2	4.4	2.5	1.6
1.9	2.7	1.7	0.1
0.3	−0.6	1.0	−2.1

$$FDR \approx \frac{\frac{7}{4}}{3} = 0.5833$$

SAM - algoritmus



SAM – jak vybrat Δ

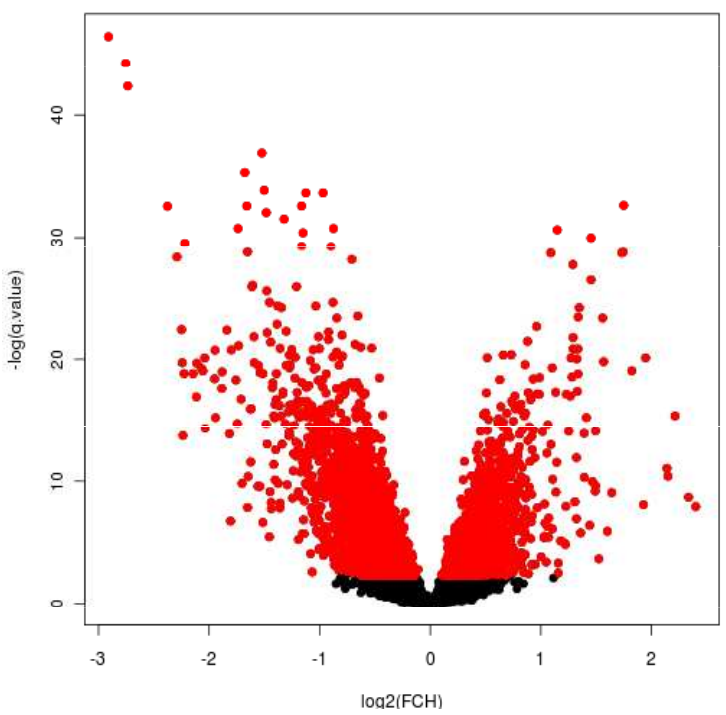
Parametr	Počet falešně pozitivních (z permutací)	Počet označených za významné (v orig.)	FDR
SAM			
$\Delta = 0.4$	134.9	288	47%
$\Delta = 0.5$	78.1	192	41%
$\Delta = 0.6$	56.1	162	35%
$\Delta = 0.9$	19.1	80	24%
$\Delta = 1.2$	8.4	46	18%

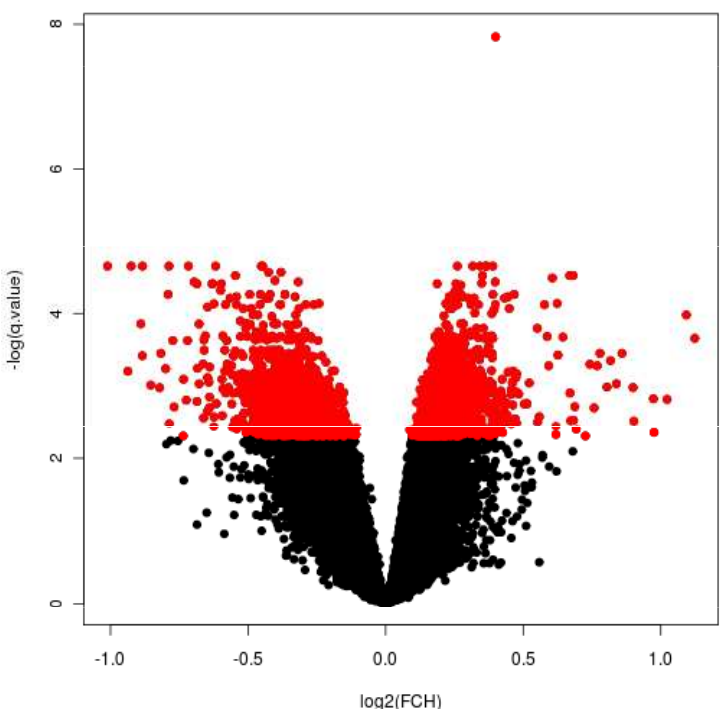
Limma

- Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology, Volume 3, Article 3.
- **Lineární modely pro stanovení odlišné exprese z mikročipových dat**
- Balík se souborem funkcí pro normalizaci dat a porovnání exprese mezi skupinami (včetně časových řad)
- Moderovaná statistika: variabilita je vyhlazená pomocí empirických bayesovských metod

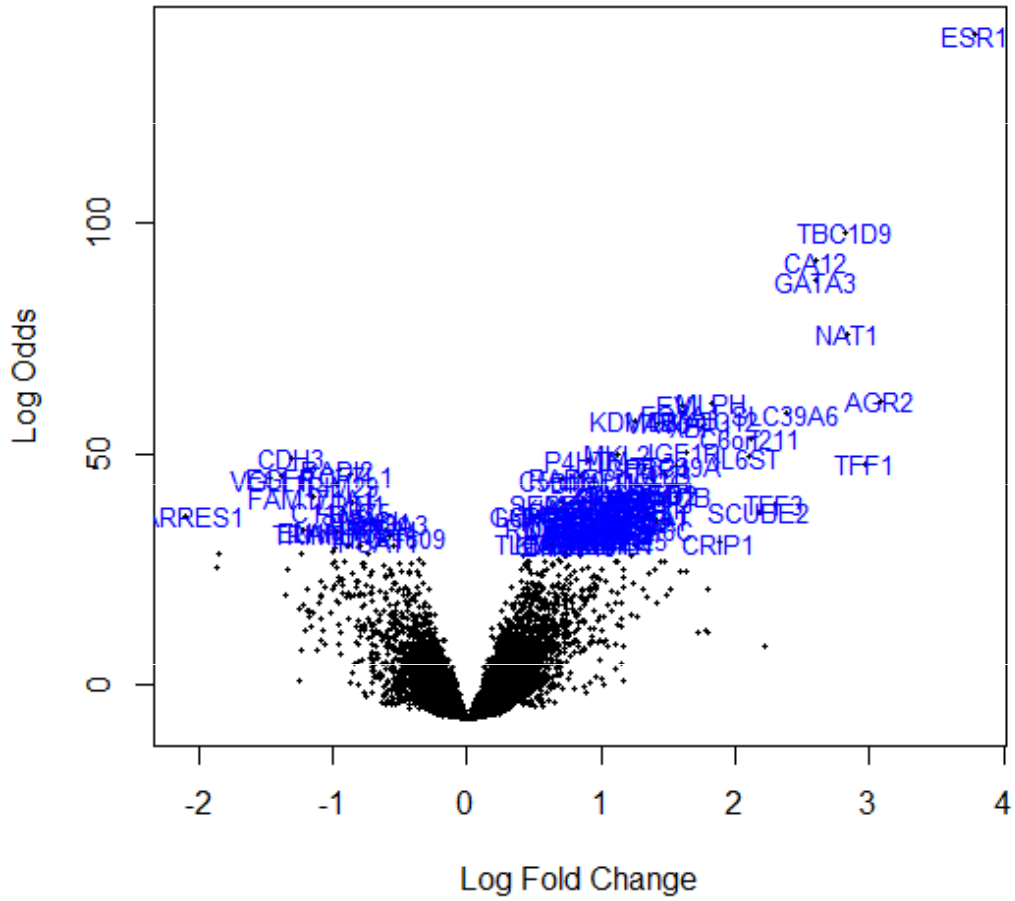
Typické zobrazení významnosti genů
Volcano plot

- log10(q-value) ~ -log10(0.1)=2.3

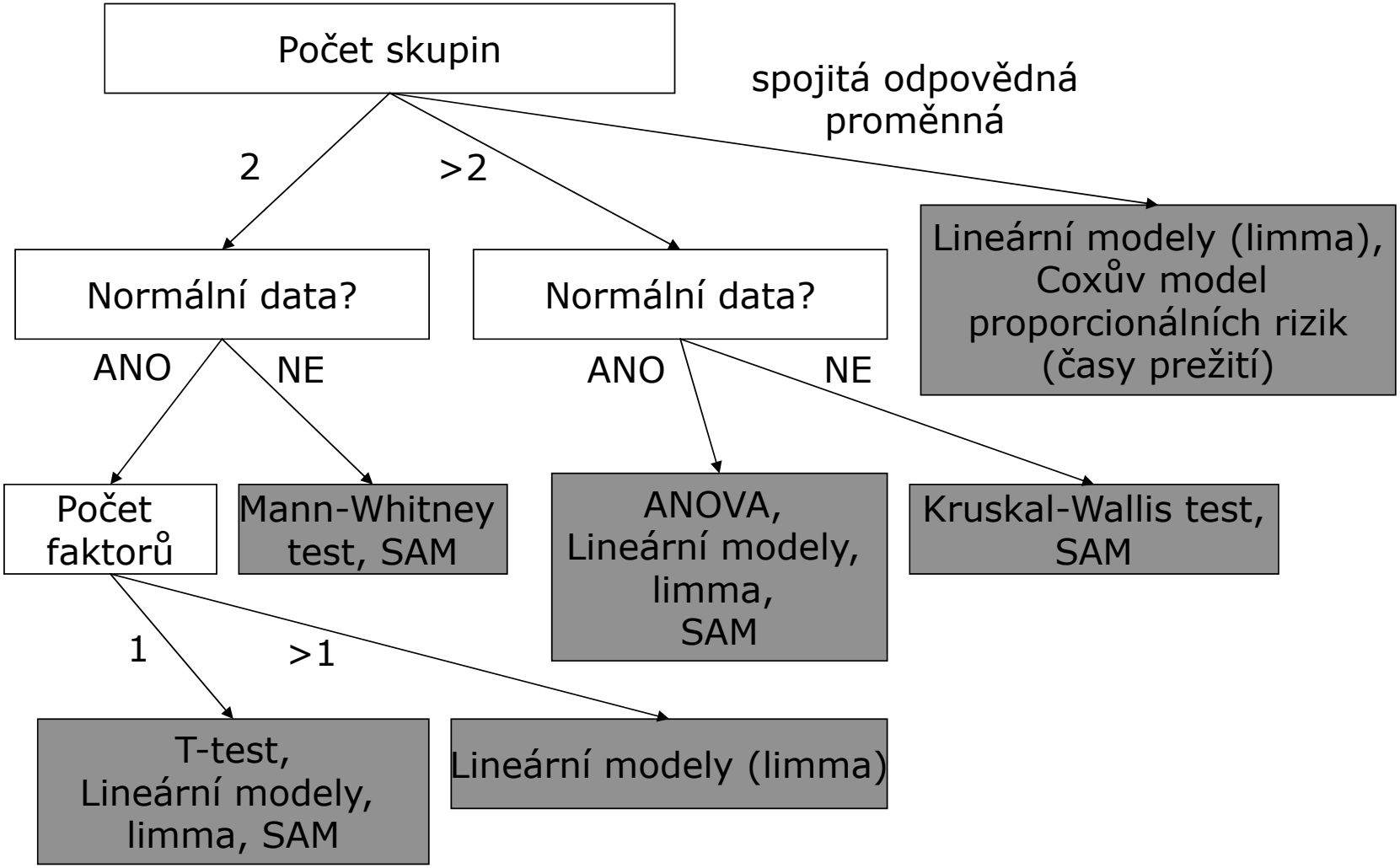




Volcano plot



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