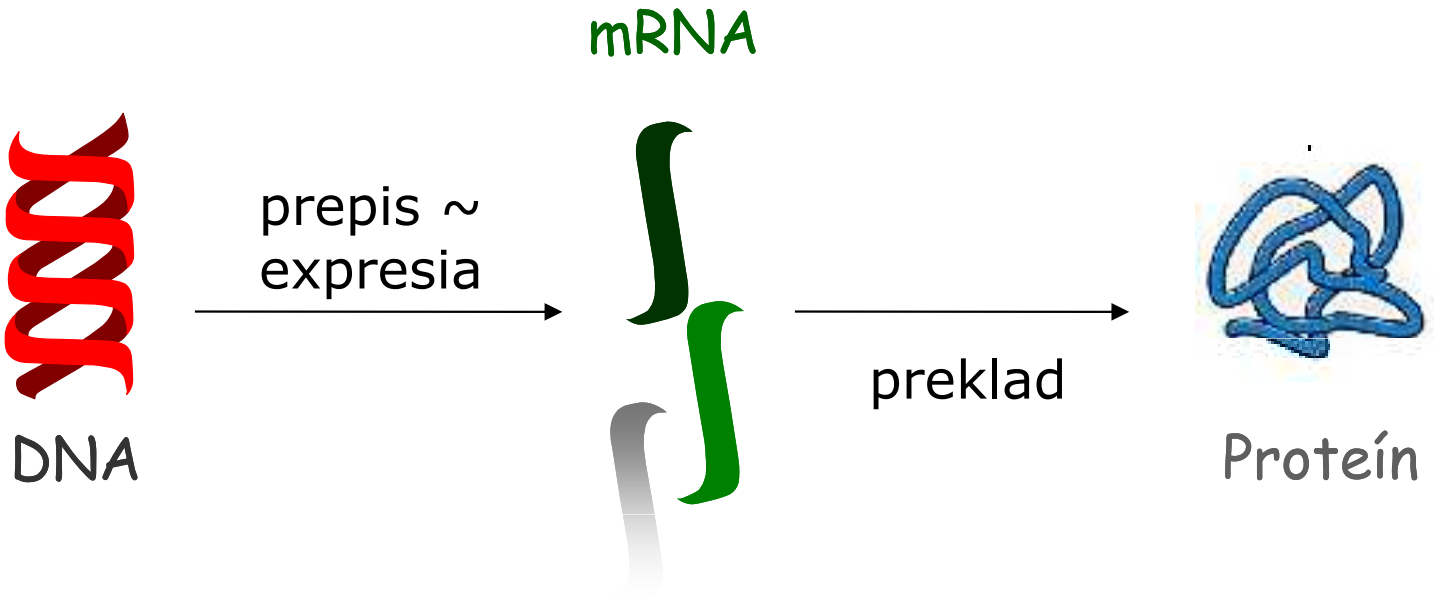
Kapitola III.

Společné principy analýzy genomických a proteomických dat

Génová expresia



- Povieme, že gén je exprimovaný, keď sa *prepisuje* do mRNA
- Ak sa gén prepisuje, znamená to, že je aktívny
- Aktivitu génu môžeme merať meraním množstva príslušnej mRNA v bunke

MU IBA

Tradičné schémy analýzy I.

- Každý experiment má odlišné ciele, v závislosti od typu dát a záujmov výskumníkov, ale existujú tradičné schémy ktoré sa opakujú:
- ***Učenie s učiteľom (supervised learning)***

Známa štruktúra dát musí byť zovšeobecnená na nové dáta

- Porovnávanie skupín (class comparison)
  - hľadáme rozdiely v expresii, v počte kópií či štruktúre génov/proteínov medzi už *definovanými skupinami*
- Predpovedanie skupín (class prediction)
  - *na známych skupinách* sa snažíme sa vytvoriť klasifikátor, ktorý by dokázal *zaradiť nového pacienta* do jednej zo skupín

MU IBA

Tradičné schémy analýzy II.

- ***Učenie bez učiteľa (unsupervised learning)***
  - Objavovanie skupín (class discovery)
    - *Štruktúra v dátach nie je známa,je potrebné ju vytvoriť,objaviť!*
    - Na základe informácií o génoch/proteínoch *hľadáme nové skupiny*
    - Príklady:
      - Existujú nejaké súbory génov ktoré sa exprimujú rovnako vo všetkých podmienkach?
      - Ochorenie X je veľmi heterogénne. Môžeme identifikovať špecifickejšie podtypy,ktoré by mohli byť cieľom cielenej terapie?

MU IBA

Společná schéma analýzy dat

Biologická otázka
(hypotéza)

Dizajn experimentu

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

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

Kontrola kvality
Normalizace
Sumarizace

Finální datová matice
N vzorků a K genů
(proteinů)

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

Objevování skupin?
(Shlukování)

Nové skupiny
genů nebo vzorků

Charakterizace nových skupin

Porovnání skupin?
(Testování)

List genů
s odlišnou expresí
mezi skupinami vzorků

Predikce skupin?
(Klasifikace)

Klasifikační pravidlo
využívající
genovou expresi

Analýza přežití

Seznam prognostických genů

Analýza časových řad

List genů
se stejným profilem
změn exprese v čase

Analýza genov
sad / genových

Interpretace

Validace

Publikace

IBA

Kapitola V.1.
Porovnávanie skupín

Príklady porovnávania skupín

- Ak chceme zistiť
  - aké gény sú aktívne/neaktívne
  - aký je rozdiel v prítomných proteínoch

    medzi dvoma alebo viacerými skupinami:
    - chorí vs. zdraví pacienti
    - pacienti pred vs. po terapii
    - pacienti v čase diagnózy a v čase relapsu
    - baktérie v aerobickom vs anaerobickom prostredí
    - druh 1 vs druh 2
    - porovnávame podtypy chorôb

Základné metódy pre porovnávanie

Môžeme rozdeliť do troch hlavných skupín:

- Metódy študujúce veľkosť efektu zmeny medzi skupinami
- Testovanie hypotéz
- Regresné stratégie

Základné metódy pre porovnávanie

Môžeme rozdeliť do troch hlavných skupín:

- Metódy študujúce veľkosť efektu zmeny medzi skupinami
- Testovanie hypotéz
- Regresné stratégie

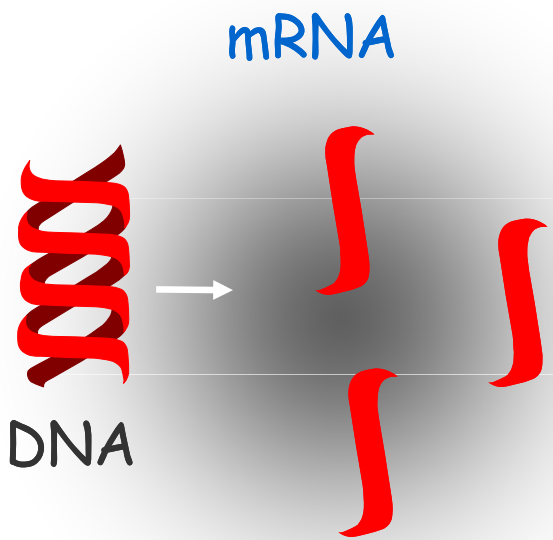
Veľkosť efektu / zmeny II.

1. Porovnáva sa pomer priemerov/mediánov jednej a druhej skupiny: mean(X)/mean(Y).
2. Stanovia sa fixné deliace hranice, ktoré určujú, aká veľkosť efektu je pre nás zaujímavá

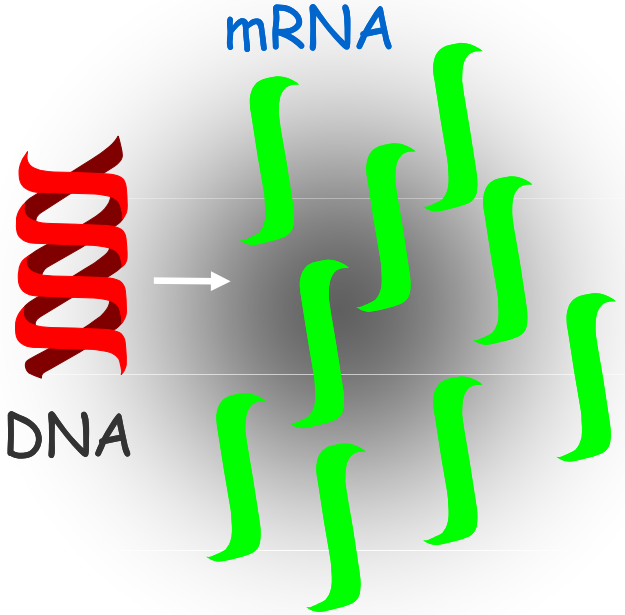
- Príklad: génová expresia, mean(X)/mean(Y), kde X a Y sú génové expresie v skupinách. Použitá hranica: 2!
- Výhody:
  - jednoduché

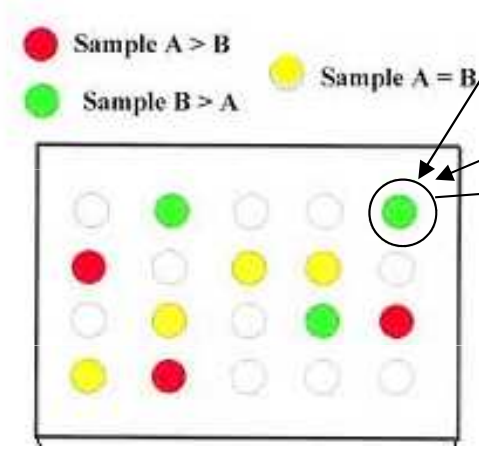
Veľkosť efektu / zmeny III.

Skupina A. Zdravá tkáň



Skupina B. Nádor





9/3 = 3

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

Veľkosť efektu / zmeny IV.

- **Nevýhody:**
  - Aj menšie zmeny môžu byť biologicky významné

  (malý efekt génu/proteínu môže byť znásobený kooperáciou viacerých génov v dráhe)
  - Dáta sú ovplyvnené technickou a biologickou variabilitou:
    - Čo ak máme 1.9?
    - Pomery môžu byť vychýlené smerom k nule (napríklad u nádorov prímesou normálnych buniek vo vzorke)
    - Neberú do úvahy variabilitu!

Testovanie hypotéz

Základné metódy pre porovnávanie

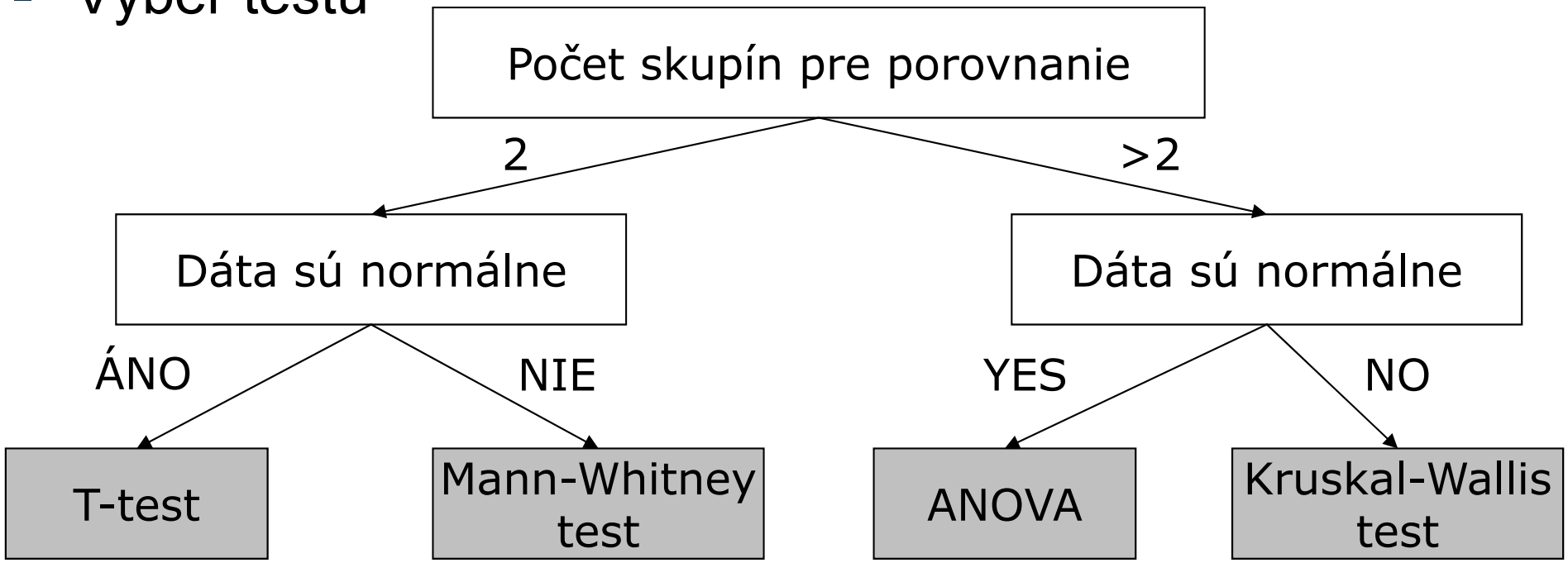
Môžeme rozdeliť do troch hlavných skupín:

- Metódy študujúce veľkosť efektu zmeny medzi skupinami
- Testovanie hypotéz
- Regresné stratégie

Testovanie hypotéz

- *Kladieme si otázku: Je aktivita/množstvo proteínu/génu) v skupine A odlišné od priemernej aktivity/množstva proteínu/génu v skupine B?*

- Na každý proteín/gén aplikujeme štatistický test, ktorým získame T_g štatistiku a príslušné *p-hodnoty*
- Výber testu



Testovanie hypotéz II.

Testuje sa

- *Nulová hypotéza (H_0)*:

Gén / proteín nie je odlišne exprimovaný medzi skupinami

versus

- *Alternatívna hypotéza (H_1)*:

Gén je odlišne exprimovaný medzi skupinami

➔Na základe našich dát musíme rozhodnúť, čo je pravda

- Nulovú hypotézu zamietneme len ak existuje *dostatočne silná evidencia*, že je neplatná
  - Evidencia – štatistika a p-hodnota!

T-štatistika I.

- Aby sme rozhodli, ktorá hypotéza je pravdivá,sumarizujeme dáta do jedného čísla
- V testovaní hypotéz sa toto číslo nazýva *štatistika (T-štatistika,Z-štatistika,F-štatistika...)*

- T-štatistika porovnáva signál so šumom
  - Signál = rozdiel priemerov v skupinách (u microarray dát sa jedná o log(skupina 1)-log(skupina 2) = log(skupina1/skupina2))
  - Šum = smerodatná odchýlka rozdielu (SD)

- T = log(skupina 1/skupina 2)/SD
- T hodnoty ďaleko od nuly indikujú zníženie alebo zvýšenie expresie v jednej zo skupín

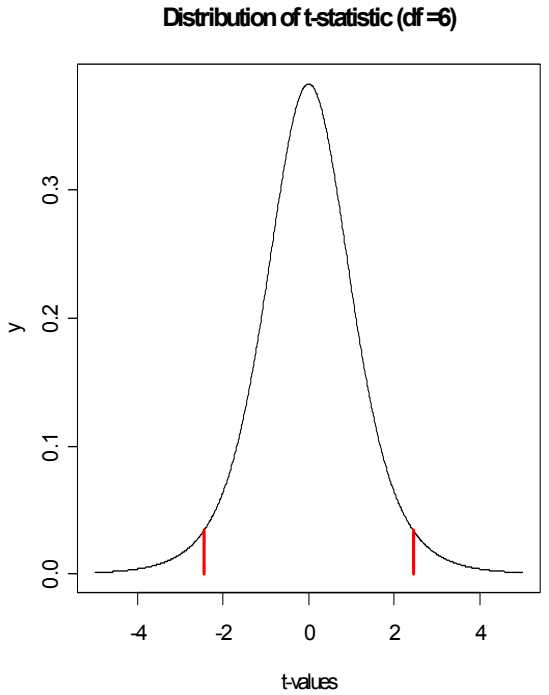
IBA

T-štatistika II.

- Dvojvýberový T-test pre porovnanie rovnosti dvoch priemerov *μ1, μ2*:
  - Priemer expresie génu v skupine 1 vs priemer v skupine 2

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

variabilita



- Ak dáta majú normálne rozloženie a neexistuje rozdiel medzi skupinami, tak T-štatistiky pochádzajú z T-rozloženia.
- p-hodnota = pravdepodobnosť že dostaneme danú hodnotu T-štatistiky alebo väčšiu, v prípade,že neexistuje rozdiel medzi skupinami

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

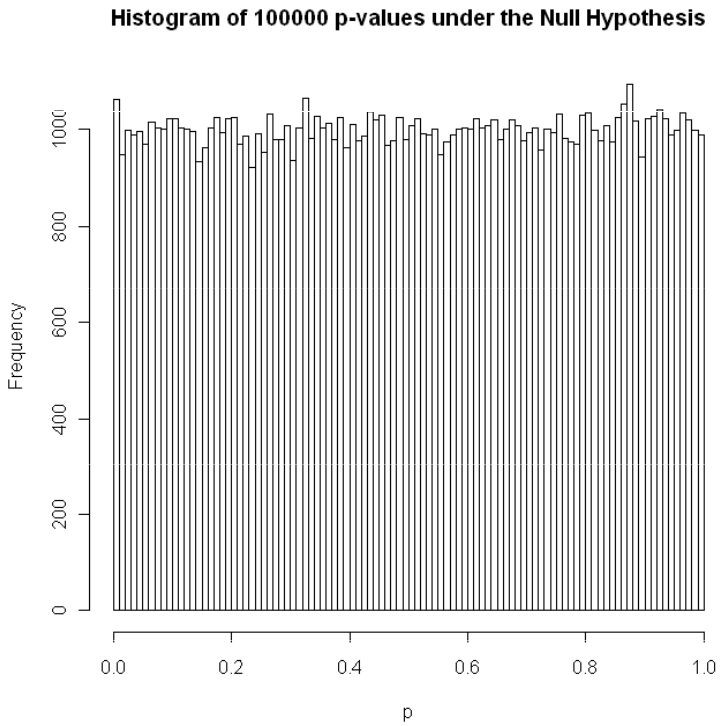
- Dostatočne malá p-hodnota = významný rozdiel (silná evidencia)

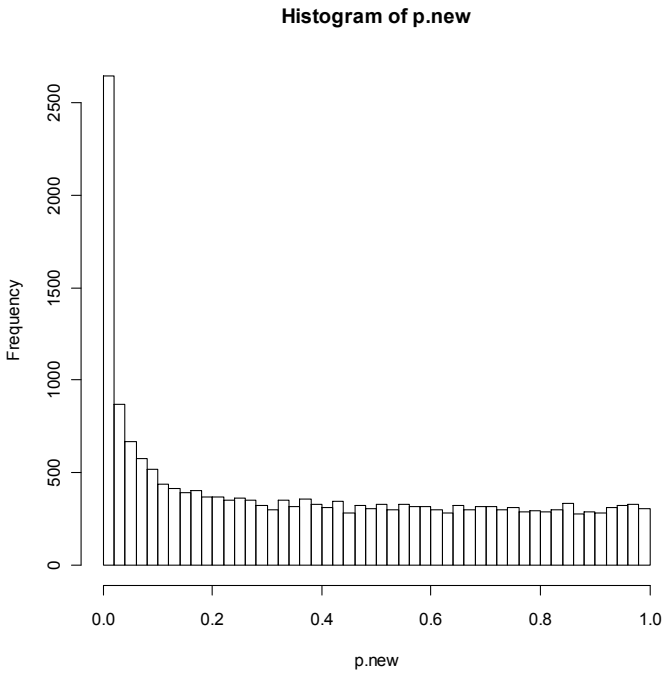
Testovanie hypotéz III.

	H0 nezamietneme	H0 zamietneme
H0 je pravdivá (gén nie je odlišne exprimovaný)	Pravdivá negativita (PN)	Falošná pozitivita (FP) Chyba 1. druhu
H0 nie je pravdivá (gén je odlišne exprimovaný)	Falošná negativita (FN) Chyba II. druhu	Pravdivá pozitivita (PP)

Testovanie hypotéz IV.

- Typické rozhodovacie pravidlo:
  - Výpočet T-štatistiky a p-hodnoty
  - ak p < 5%, gén je označený za odlišne exprimovaný
- Dôležité: V prípade, že platí nulová hypotéza, sú p-hodnoty rovnomerne rozložené (vľavo). V prípade,že je značná časť génov odlišne exprimovaná, rozloženie p-hodnôt už nie je uniformné (vpravo).





Problém mnohonásobného porovnávania

Porovnávame tisíce génov/proteínov medzi skupinami.

⬇

Hypotézu testujeme pre každý gén!

⬇

Máme zvýšenú šancu falošne pozitívnych výsledkov!

Príklad: 10 000 génov, žiadny odlišne exprimovaný medzi skupinami => 0.05 x 10 000 = 500 s $p < 0.05$.

⬇

$p < 0.05$ už negarantuje významnosť výsledku

⬇

Musíme teda spraviť korekciu p-hodnôt na mnohonásobné porovnávanie

Korekcia problému mnohonásobného porovnávania

	# nezamietnuté (NZ)	# zamietnuté (Z)
#bez rozdielu	Pravdivá negativita (PN)	Falošná pozitivita (FP) Chyba 1. druhu
# odlišné gény/proteíny	Falošná negativita (FN) Chyba II. druhu	Pravdivá pozitivita (PP)

Chyby 1. druhu:

1. **Family–wise error rate (FWER)**: Pravdepodobnosť aspoň jednej chyby prvého druhu (falošnej pozitivity): FWER = Pr(FP > 0)

1. **False discovery rate (FDR)**(Benjamini & Hochberg,1995): Očakávaný podiel falošne pozitívnych výsledkov medzi zamietnutými hypotézami

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

Korekcia p-hodnôt

- Kontrolujeme FWER
  - Bonferroniho korekcia (pre nezávislé testy!)

$p < \alpha / m$ (napr. $p < 0.05/10\,000$)

- Kontrolujeme FDR
  - Benjamini/Hochberg procedúra

    FDR = 10% (zo 100 zamietnutých hypotéz očakávame 10 falošne pozitívnych)

Ktorý typ korekcie použiť?

- FWER ak chceme aby VŠETKY vybrané gény/proteíny boli naozaj významné. Na druhú stranu, nevyberieme tak všetky významné gény!

- FDR ak preferujeme vybrať väčšinu významných génov/proteínov, a nevadia nám nejaké falošne pozitívne

- q-hodnota je najmenšia FDR pri ktorej daný gén ešte ostáva na liste pozitívnych

Moderovaná T-štatistika

- Problém v štatistickom testovaní mikročipových dát:

Príliš malé hodnoty expresie (blízke šumu) vykazujú malú variabilitu => vysoké T-štatistiky u biologicky nerelevantných génov!

Príklad:

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

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

$$s_g = 0.02$$

$$\Rightarrow T_g = -25$$

- Aby sa dali štatistiky porovnať, treba zjednotiť variabilitu:
- Moderovaná T-štatistika:

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

Konštanta korigujúca variabilitu

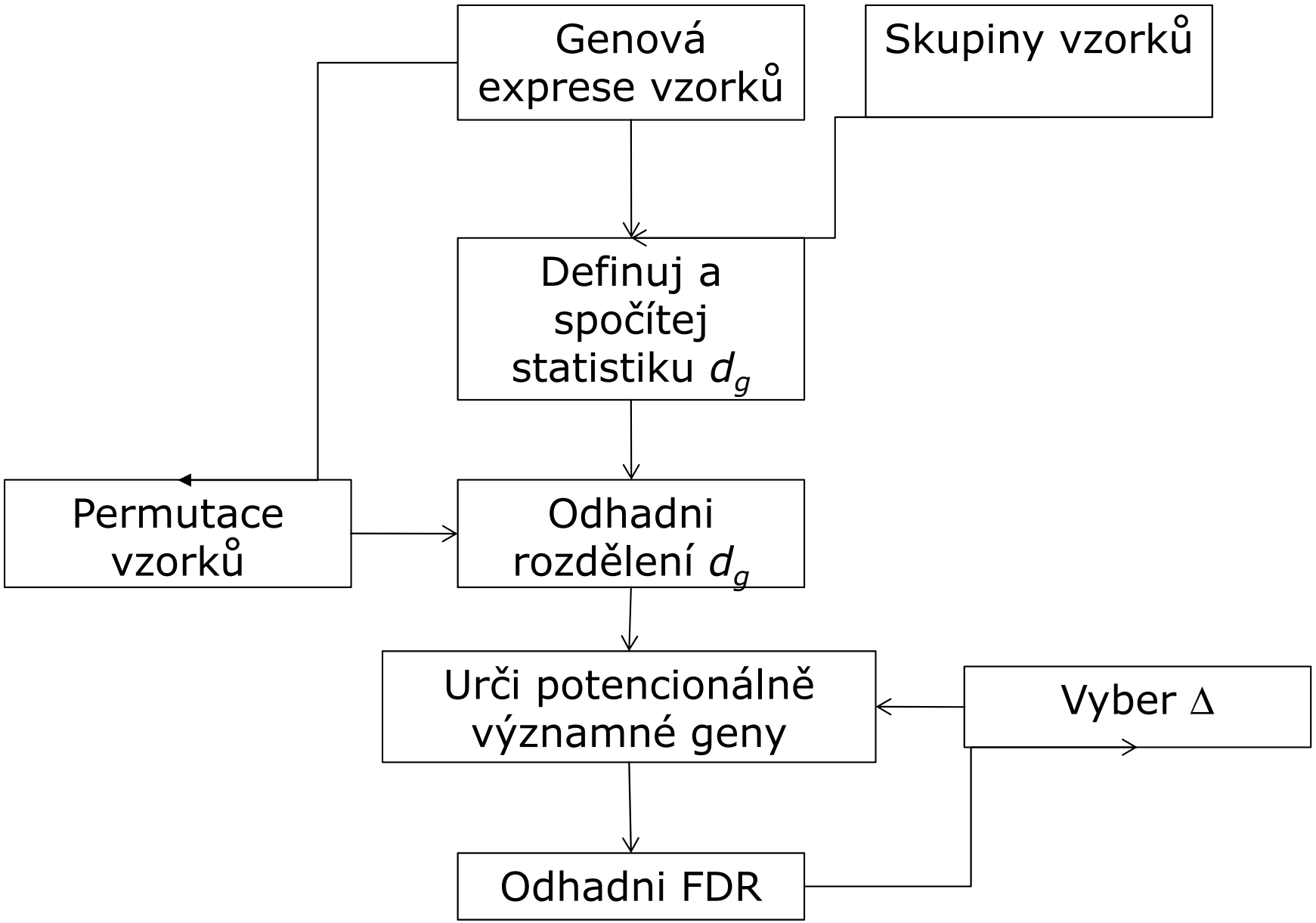
Significance analysis of microarrays (SAM)

- Tusher, Tibshirani a Chu (2001)
- Založená na moderovanej *t*-štatistike (d_g), počíta FDR

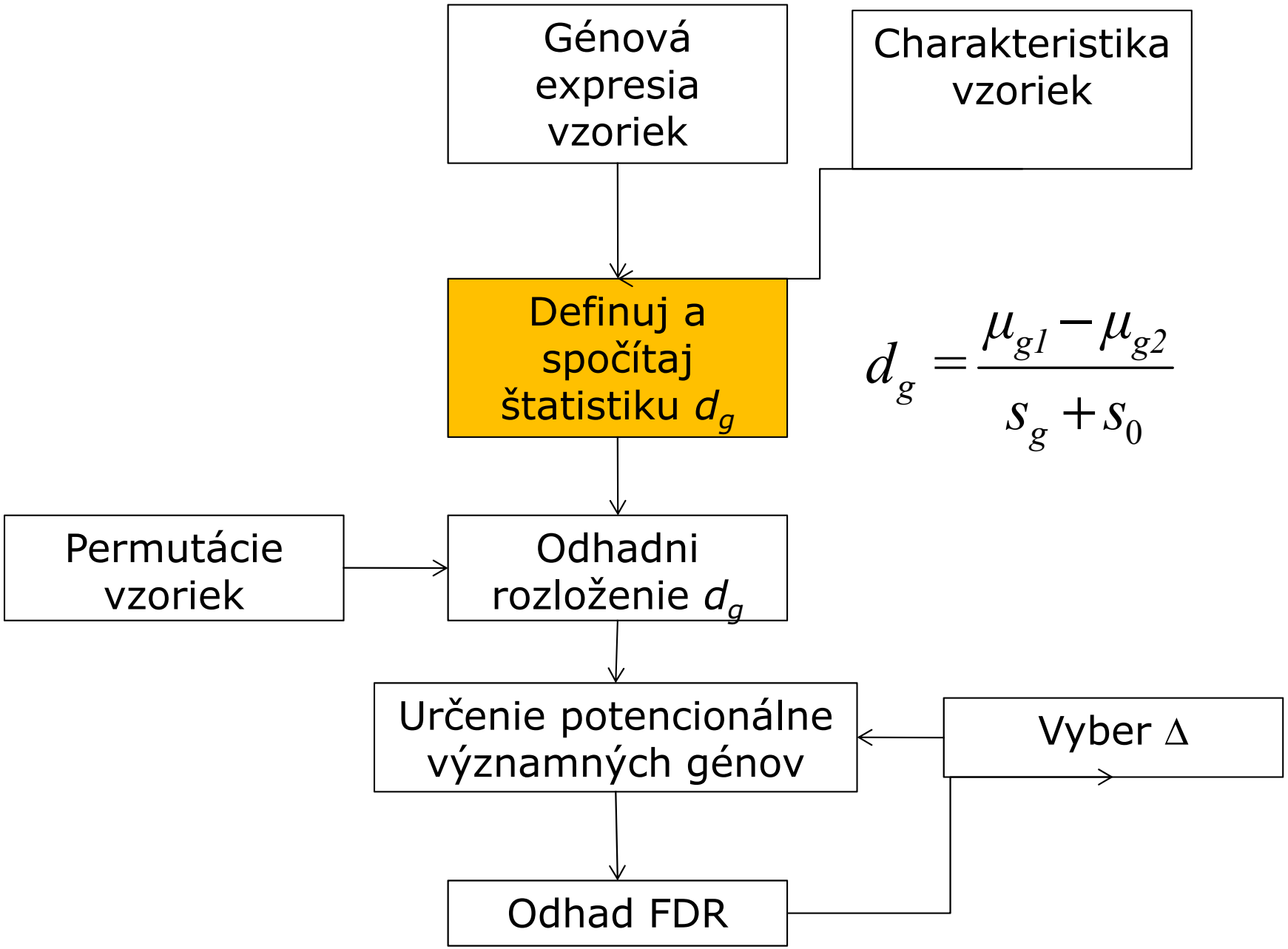
$$d_g = \frac{\mu_{g1} - \mu_{g2}}{s_g + s_0}$$

- Štatistická významnosť d_g je následne stanovená permutáciami pôvodných dát a kalkuláciou očakávaného skóre v prípade, že platí nulová hypotéza (d_e)
- Gén je štatisticky významný, keď splňuje podmienku $|d_g - d_e| > \Delta$.
- Výhody: jednoduché
- Nevýhody: výpočtovo náročné (permutácie)
- Výstup: *q*-hodnoty
- `biocLite("samr")`
- `library(samr)`

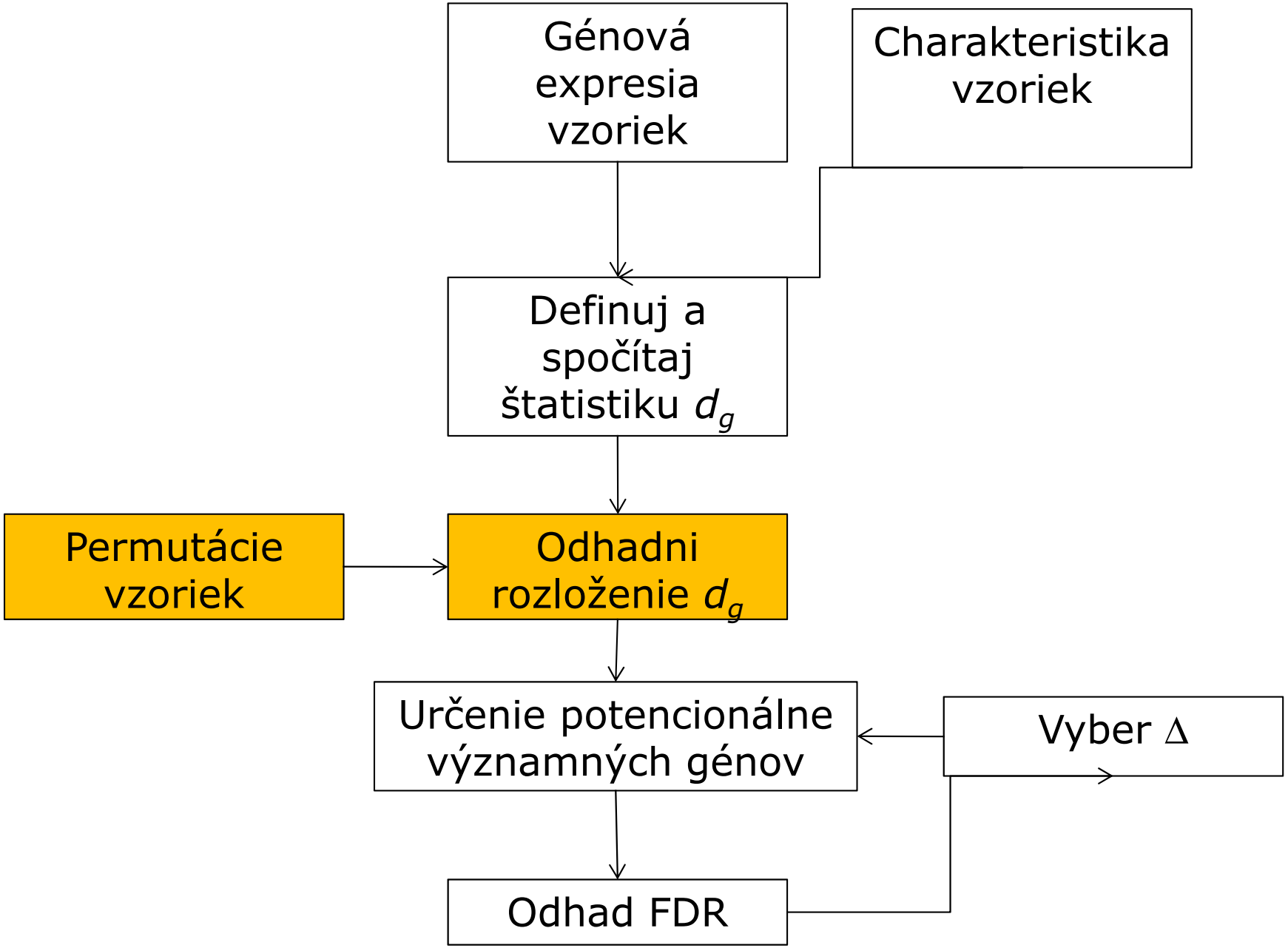
SAM - algoritmus



SAM - algoritmus



SAM - algoritmus



SAM - výpočet očakávaných hodnôt

- Pre každú permutáciu *p* spočítaj d_{gp}

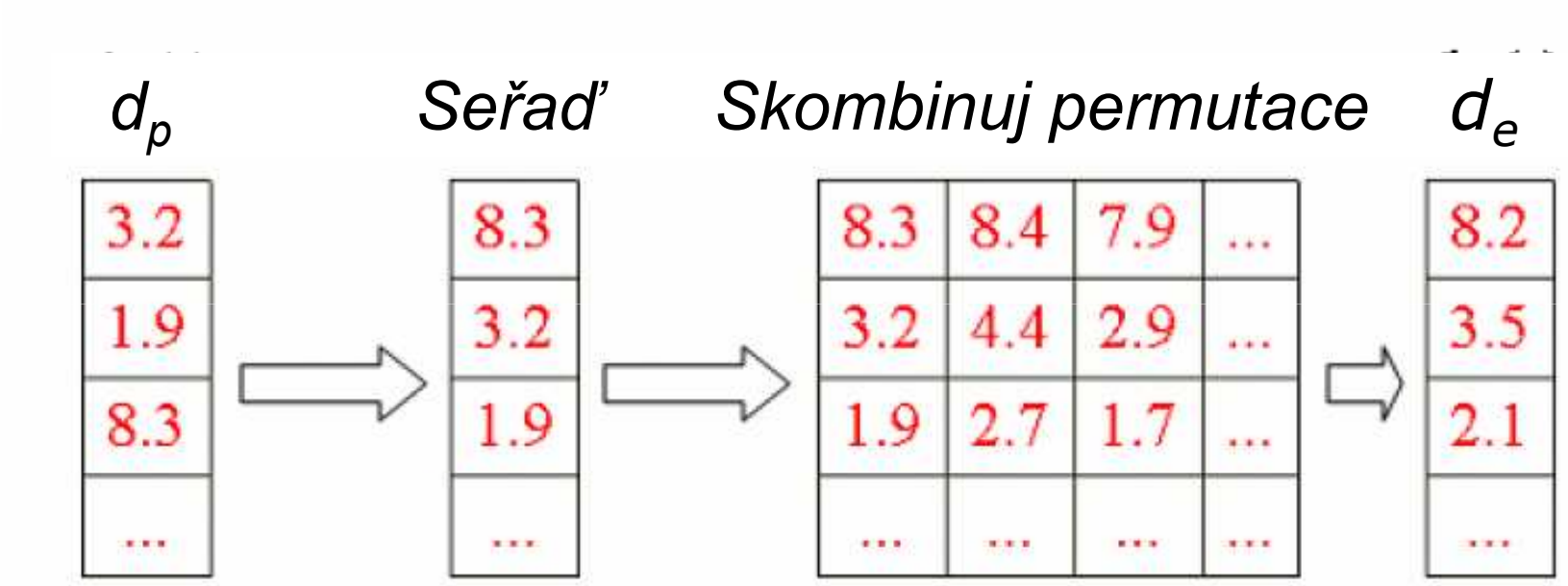
$$d_{gp} = \frac{\mu_{g1} - \mu_{g2}}{s_g + s_0}$$

- Zoraď štatistiky podľa veľkosti

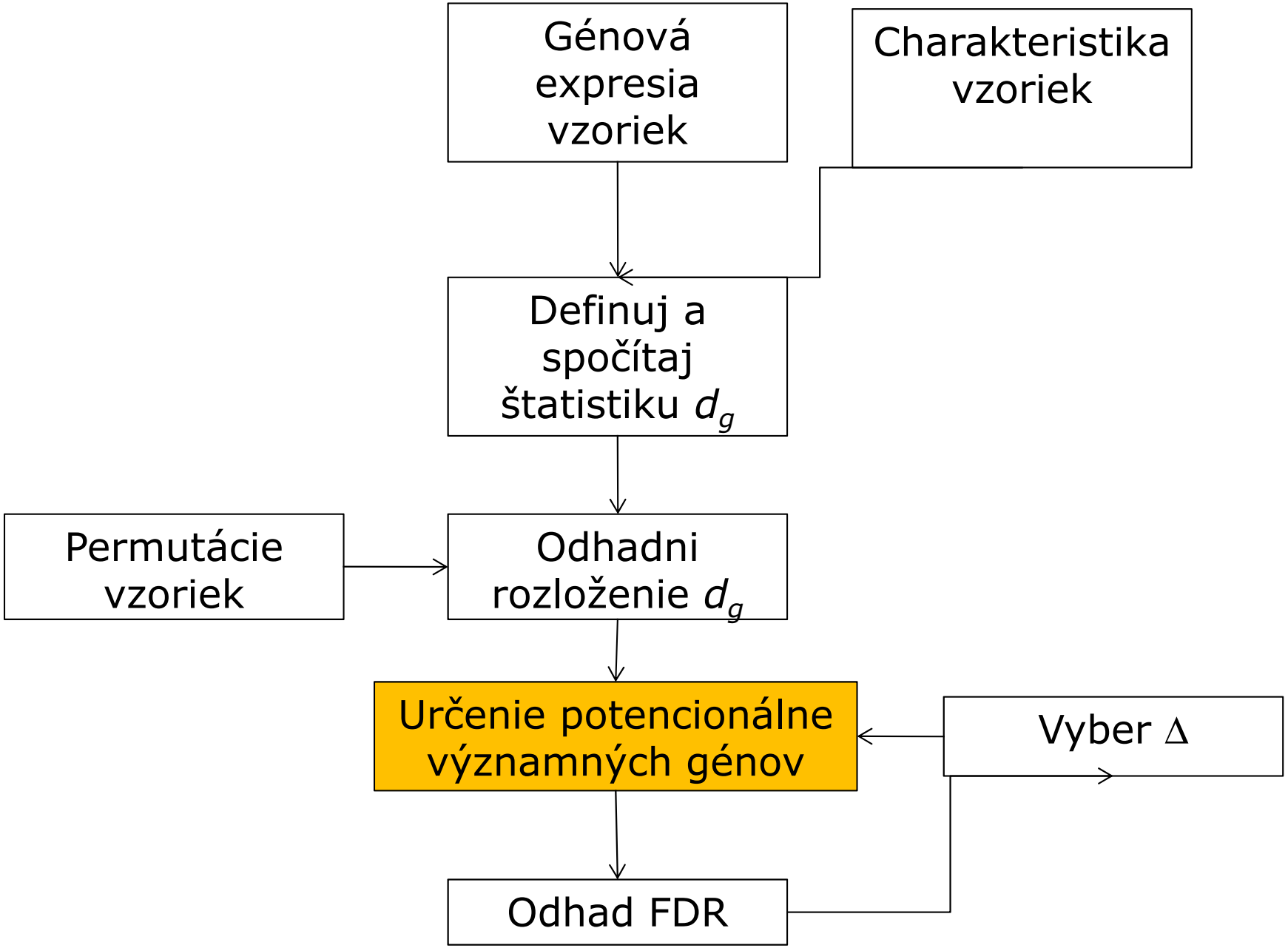
$$d_{1p} \geq d_{2p} \geq d_{3p} \geq \cdots$$

- Definuj *g*-tú očakávanú hodnotu na základe N permutácií

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



SAM - algoritmus

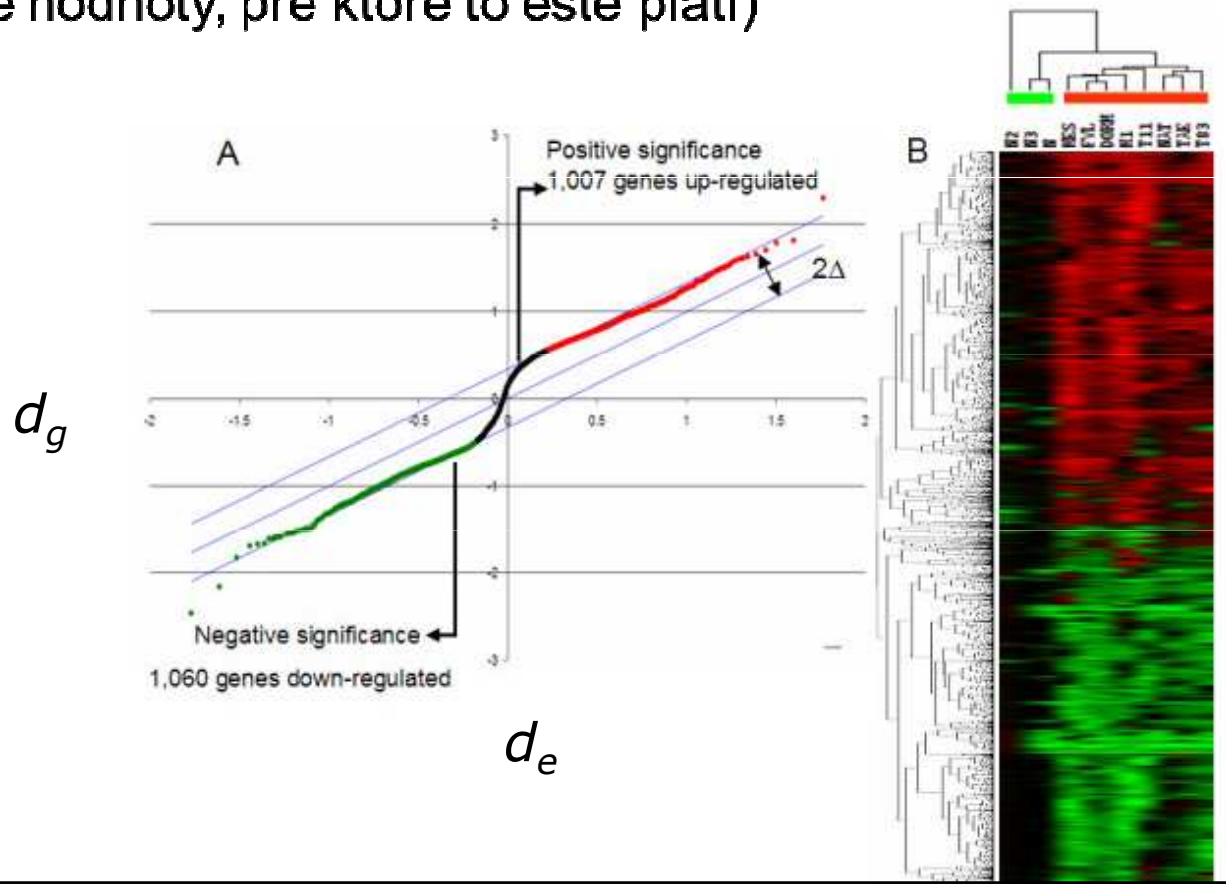


SAM – určenie významných génov I

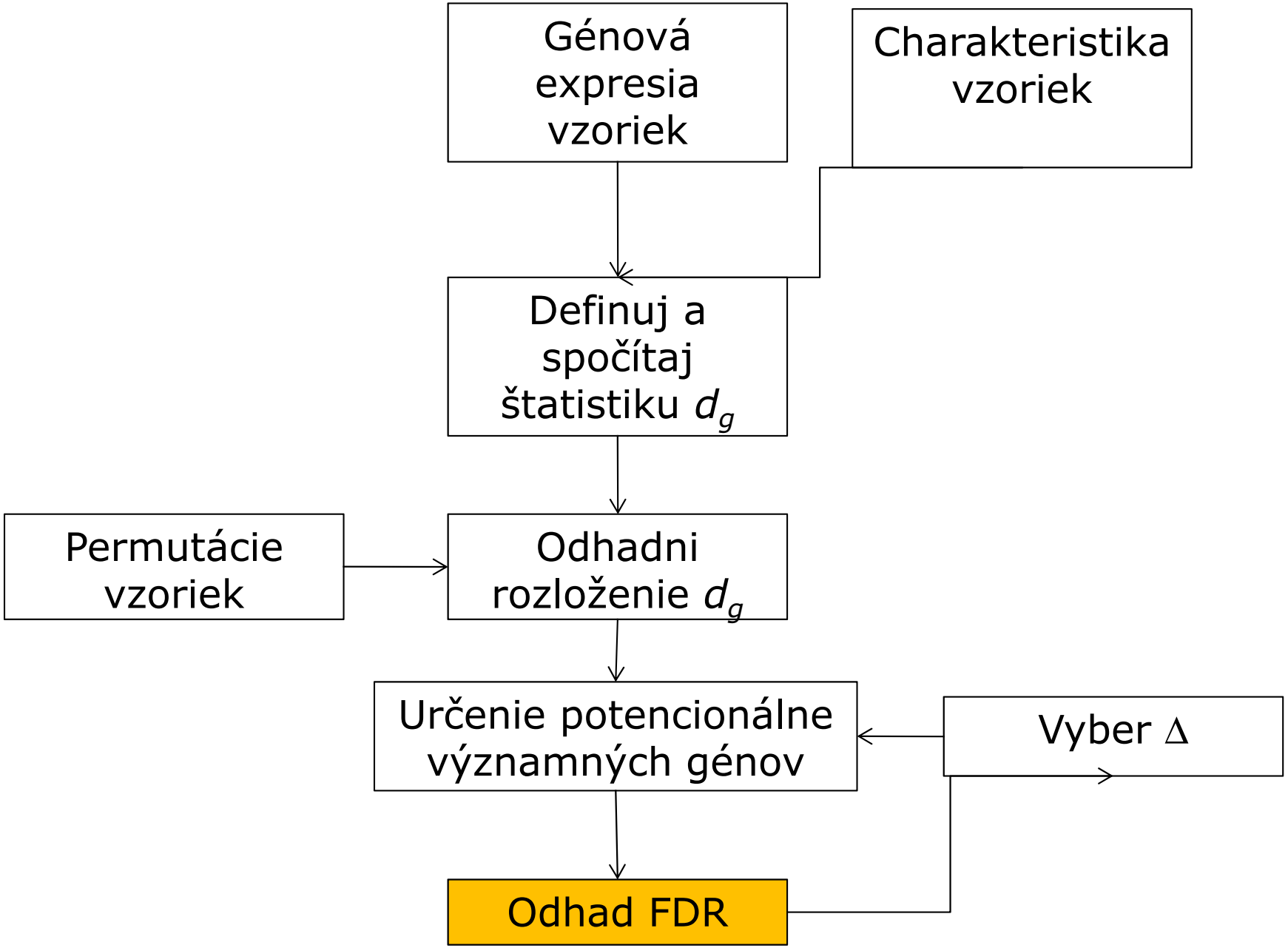
- Zoraď pôvodné štatistiky podľa veľkosti

$$d_1 \geq d_2 \geq d_3 \geq \cdots$$

- Nakresli graf d_g vs d_e a definuj Δ
- Gén je štatisticky významný, keď spĺňa podmienku $|d_g - d_e| > \Delta$ (označme t1 a t2 hraničné hodnoty, pre ktoré to ešte platí)



SAM - algoritmus



SAM – výpočet FDR

- t1 a t2 budú použité ako hranice
- Vypočítaj priemerný počet génov, ktoré v permutáciách tieto hranice prekročili (boli významné)
- Odhadni počet falošne pozitívnych génov v prípade, že platí nulová hypotéza podelením počtom významnách génov v originálnom pozorovaní:

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

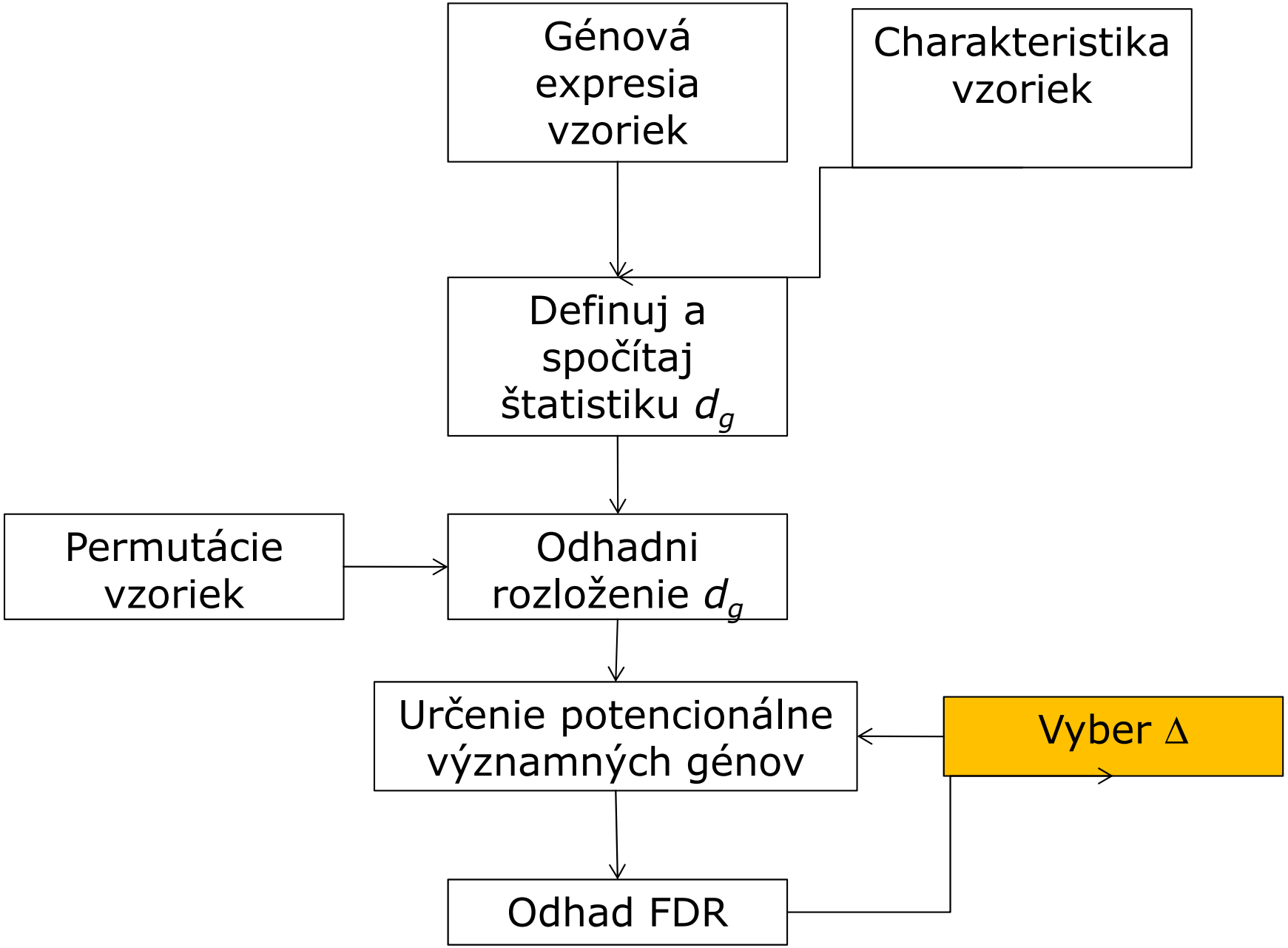
SAM – výpočet FDR, príklad

	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 – ako vybrať Δ

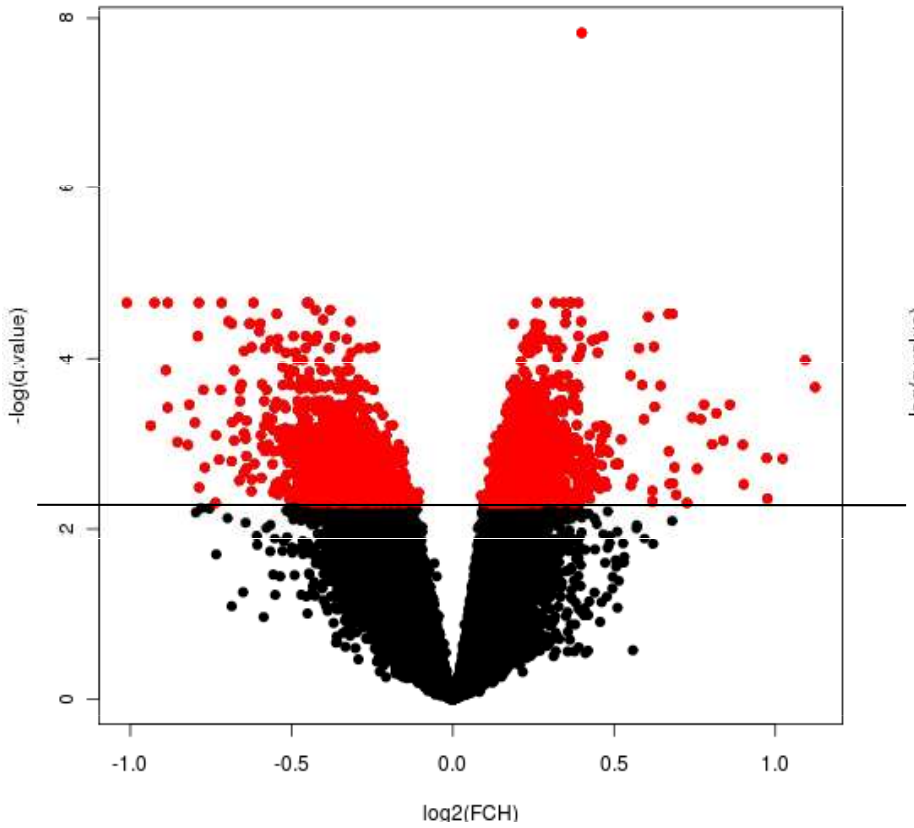
Parameter	Počet falošne pozitívnych (z permutácií)	Počet označených za významné (v orig.)	FDR
SAM			
$\Delta = 0.4$	134.9	288	47%
$\Delta = 0.5$	78.1	192	41%
$\Delta = 0.6$	56.1	162	35%
$\Delta = 0.9$	19.1	80	24%
$\Delta = 1.2$	8.4	46	18%

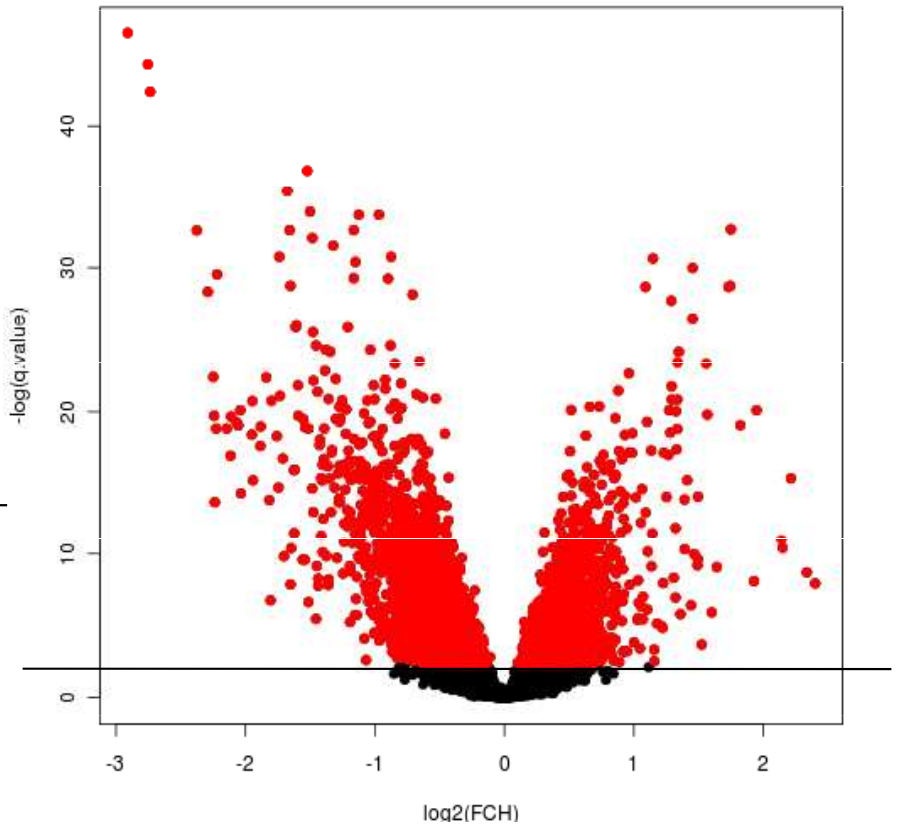
Limma

- 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. http://www.bepress.com/sagmbvol3/iss1/art3
- **Lineárne modely pre stanovenie odlišnej expresie z mikročipových dát**
- Balík so súborom funkcií pre normalizáciu dát a porovnanie expresie medzi skupinami (vrátane časových radov)
- Moderovaná štatistika: variabilita je vyhladená pomocou empirických bayesovských metód
- `biocLite("limma")`
- `library(limma)`

Volcano plots I.

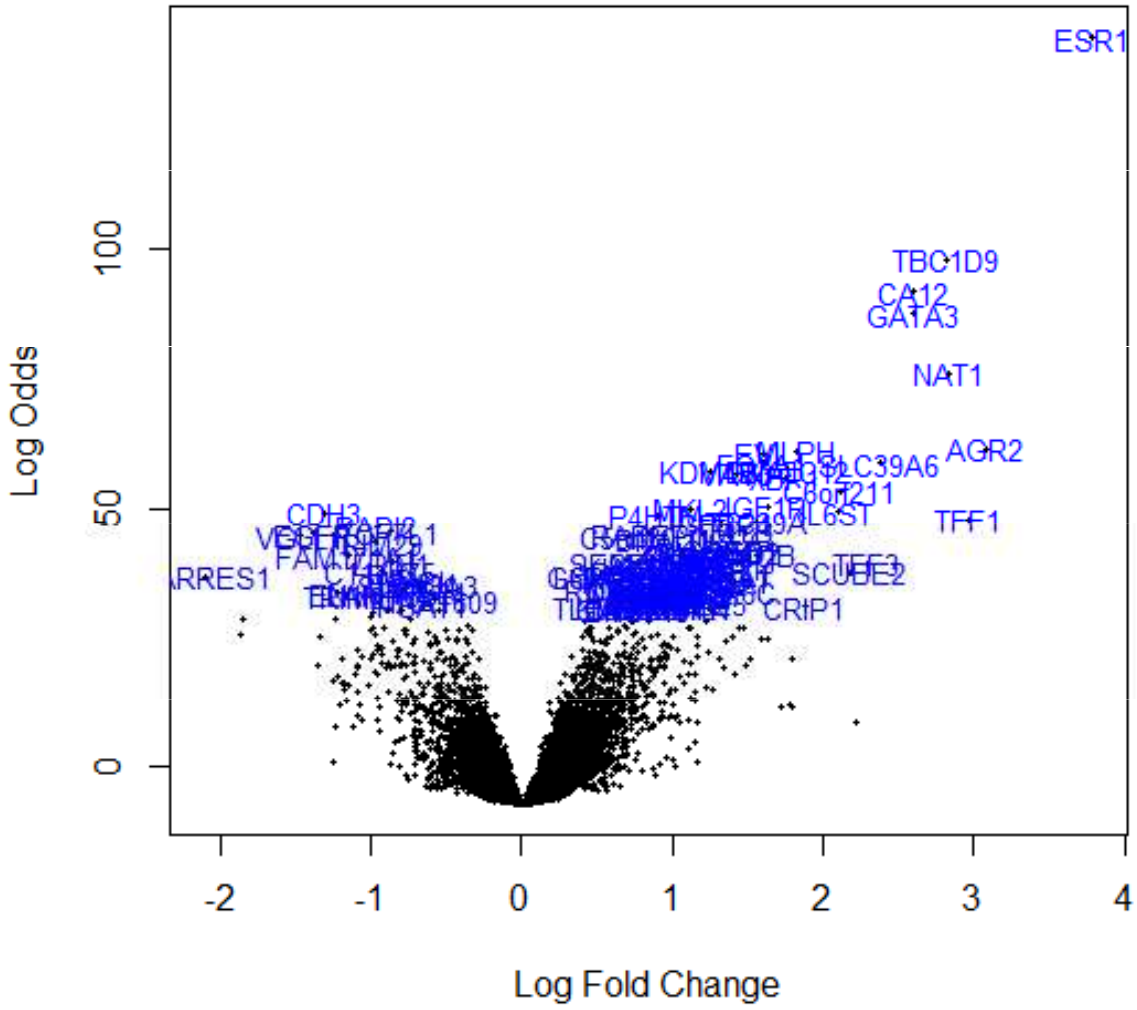
- log10(q-value) ~ -log10(0.1)=2.3





Volcano plots II.

```
library(limma)
volcanoplot(fit2, highlight=100)
```



Základné metódy pre porovnávanie

Môžeme rozdeliť do troch hlavných skupín:

- Metódy študujúce veľkosť efektu zmeny medzi skupinami
- Testovanie hypotéz
- Regresné stratégie

Regresné stratégie

- Ak máme viac ako 1 premennú, ktorá môže ovplyvniť génovú/proteínovú expresiu
  - génová expresia ~ skupina + pohlavie

  Lineárne modelovanie

- Ak sa snažíme zistiť, ako veľmi sa génová expresia zmení, ak sa zmení hodnota nejakej *spojitej premennej*
  - génová expresia ~ prežitie
  - génová expresia ~ vek

  Lineárne modelovanie, Coxov model proporcionálnych rizík

- Chceme nájsť pravdepodobnosť, že vzorka patrí do istej skupiny na základe expresnej hodnoty daného génu

  Logistická regresia

IBA

Porovnanie skupín